University of Cape Town

STA5079W

Data Science Minor Dissertation

Analysis of gender wage gap using mixed effects models

Author:
Magnolia M.Chikanya

Student Number:
CHKMAG002

November 28, 2024

Acknowledgements

I am thankful to my main supervisor, Dr. Sebnem Er, for her boundless patience, guidance and support have been the cornerstone of my journey through this master's program. I also appreciate my co-supervisor, Prof. Sheetal Silal, for her insightful contributions that shaped and refined this thesis. Together, my supervisors have encouraged me to think beyond conventional boundaries, and I am grateful for their mentorship and wisdom.

Furthermore, I extend my appreciation to Clinton Health Access Initiative, (CHAI) organisation for granting me the opportunity to conduct my study and providing resources and support that has been instrumental in the successful completion of this thesis. I would also like to express special thanks to Nikhil Khanna, for actively participating in meetings and providing contributions that enriched this thesis.

To my cherished family, your unwavering encouragement sustained me throughout this journey. This thesis stands as a tribute to the unrelenting backing I have received from those around me.

Analysis of gender wage gap using mixed effects models

Magnolia M.Chikanya

November 28, 2024

Abstract

Despite government interventions, the gender wage gap persists in workplaces. While reports on whether the gap is widening or narrowing vary, addressing this issue remains crucial. Traditionally, researchers have employed methods like the Blinder-Oaxaca decomposition and quantile regression to estimate the gender wage gap. However, these approaches often leave a high unexplained variance attributed to discrimination.

In existing studies, gender wage gap estimates have typically been aggregated, and attempts to disaggregate the analysis have focused on broader levels such as occupations and salary bands. To delve deeper, human resource data from the National Department of Health in South Africa Eastern Cape province was leveraged. The goal was to analyze the gender wage gap for each job title using a novel approach: linear mixed effects regression.

The linear mixed effects model captures both systematic trends and unexplained variability simultaneously to provide a more comprehensive understanding of the gender wage gap. Here are the key findings:

1. The unexplained variance in gender wage gap was remarkably low, accounting for only 3% of total variance. This indicates that the model captures most of the variability in the data as a result there is minimal unexplained variation.
2. Job titles emerged very significant by explaining 83% of the total random variance. This highlights the significance of considering specific roles when analyzing gender wage gap.
3. Over time, interesting patterns were observed. From 2010, the gender wage gap narrowed, but starting around 2015, it gradually widened again.
4. Encouragingly, 42% of the job title groups showed a gender wage gap in favor of women. Additionally, a substantial proportion of females occupied managerial and highly skilled positions.

Therefore, incorporating random effects techniques through linear mixed effects regression enriched the analysis of gender wage gap. By examining job titles individually, detailed insights into this complex issue were gained. These findings underscore the importance of considering both fixed and random effects when studying wage disparities.

Contents

List of Tables

List of Figures

1 Introduction

The elimination of all forms of inequality against women does not only translate to self-actualisation but it contributes to the improved well-being of the families they support as well as the optimal growth and development of the country. Wodon and De La Briere (2018) reported that reduction of gender inequality is not just the right thing to do but it also makes economic sense as global wealth loss due to gender wage gap is equated to be $\$23,620$ per capita. In addition, Adelekan and Bussin (2018) emphasised that the promotion of equal pay for work of equal value is something on the agenda of UN and other countries in the world. Bezuidenhout et al. (2019) claimed that developing countries would benefit the most from gender wage equality because that is where the gender wage gap is more pronounced.

In South Africa, the majority of women are primary breadwinners and the number of households that depend on income from women is soaring (Casale et al., 2021). StatsSA (2022) reported that 42.1% homes in South Africa are female headed and if the issue of gender wage gap is not dealt with then it means those female headed households have a greater chance of suffering from poverty compared to male headed families (Mosomi, 2019).

Gender inequality at the work place can be traced back to the apartheid regime system where it was the core issue against women's development (Sinden, 2017). The apartheid regime contributed to a dysfunctional labour market in South Africa, and left the country with not just large racial inequalities but gender inequality as well (Gradín, 2021). Following the apartheid era, in order to address issues of gender inequality, Casale et al. (2021) reported that:

> "the government introduced different policies such as the Basic Conditions of Employment Act (1997), the Employment Equity Act (1998) and minimum wage legislation in low-wage employment (for example, for contract cleaners in 1999, domestic workers in 2002, agricultural workers in 2003 and nationally in 2018)." (p.2)

In addition, Fisher et al. (2021) acknowledged that since the inception of democracy in 1994, some work has been done to deal with gender inequality originating from the past apartheid policies. This has resulted in significant improvement in employment and labour force being realised in the first decade after apartheid (Mosomi, 2018). Fisher et al. (2021) also reported that though gender gap remains high in South Africa, it has narrowed from the period between 1997 to 2015 and affirmative action policy might have played a role by increasing the productive characteristics of women.

Besides the consequences of apartheid, gender gap goes back to religion beliefs that females are weaker vessels and from birth a girl child is disadvantaged (Nadler and Stockdale, 2012). Most females as well, have perceptions that their primary role is to run the home and bear children and this motivates them to look for work based on convenience and proximity resulting in females not being able to negotiate a better salary thereby accepting low paying jobs (Wekwete, 2014; Adelekan and Bussin, 2018).

However, on the other hand, Wekwete (2014) reported significant changes in most of these socio-cultural beliefs and practices except for stereotype. Nowadays young girls have access to education and are encouraged to take up previously male dominated professions in science, technology, engineering and mathematics (Xu, 2015). Though women have made progress in education, Toczek et al. (2021) argued that those investments in qualifications were not yet rewarding compared to those of men as income differences were still evident even with the same investment in human capital. Thus, despite government intervention and emancipation of women, gender wage gap is still there against females and it is more pronounced at the lower tail of the income distribution

(Steyn and Jackson, 2015).

Adelekan and Bussin (2018) reported varying results regarding the gender wage gap in South Africa. The discrepancies could be attributed to researchers relying on household survey data for analyzing post-apartheid earnings. In addition, Pleace et al. (2023) argued that household surveys have limitations, including reporting bias and under-sampling of certain groups and recommended using administrative data, which provides a more comprehensive and accurate representation of each employee. However, administrative data, particularly from the South African Revenue Service (SARS), has become available more recently for analysis of earnings (Kerr and Wittenberg, 2019). This study utilizes administrative data collected between 2010 and 2021 from the National Department of Health's human resource data in the Eastern Cape province.

Pleace et al. (2023) claimed that the gender wage gap widened in administrative tax data covering the period from 2008 to 2021. However, other studies presented contrasting results. Fisher et al. (2021) and Mosomi (2019) both used the Post Apartheid Labour Market Series (PALMS) dataset and reported a narrowing of the gender wage gap. Despite these differing reports, it remains concerning that the gender wage gap persists in the workplace (Musetsho et al., 2021).

Ansel (2017) suggested that it is crucial to take into account the gender wage inequality in both occupations and in industries in order to get at the bottom of sceptics of whether gender wage is increasing or decreasing. Some studies have analysed gender wage gap across different occupations, industries and pay bands and their findings revealed that gender wage gap was evident in those different groupings (Fisher et al., 2021; Adelekan and Bussin, 2018). However, Casale et al. (2021) argued that these broad occupational categories do not account for the fact that women are taking lower positions in those occupations resulting in them being paid less. Therefore there is need for studies that capture the fine-grained details in the occupations in order to fairly determine the gender wage gap (Casale and Posel, 2011; Casale et al., 2021). This study not only disaggregates the data by occupation but also goes further to analyze the gender wage gap in each job title, which is the narrowest and most specific level.

In order to analyse gender wage gap for each of the many job titles, this study applies linear mixed effect regression models. These models have an advantage of examining gender wage gap while also capturing variability within and across employees and job titles simultaneously (Brown, 2021). Since the dataset constitute more than half a million observations, for easier data management and interpretability, the data was divided into six occupational groupings. Furthermore, given that the data was collected over a period of 12 years, they are repeated measures. Thus, unique employees and job titles are modeled as random intercepts, and the gender variable as a random slope to the job title random intercept.

Blinder-Oaxaca decomposition has been widely used in the analysis of gender wage gap. This method split observed gender earnings disparities into explained and unexplained parts. However, there have been criticisms on the unexplained portion which is taken to be due to discrimination (Bhorat and Goga, 2013). In addition, Strittmatter and Wunsch (2021) argued that the unexplained part cannot be entirely attributed to discrimination, it is possible that other factors not included in the analysis may have effect on the observed differences between men and women.

Furthermore, consistent findings from literature on gender wage gap in South Africa revealed that though the explained portion of gender wage gap has changed, the unexplained portion has remained constantly high (Casale and Posel, 2011; Mosomi, 2018; Casale et al., 2021). This implies that a large proportion of gender wage gap remains unaccounted for by the explanatory variables included in the analysis. Though this unexplained variance could be as a result of inherent randomness or omitted variables, it is also possible that the differences between employees and

occupational groups can also contribute to this residual variance.

This study utilizes linear mixed effect regression, which has the advantage of accounting for variability specific to certain groups or levels within the data, thereby reducing the overall residual. Furthermore, the addition of random slopes provides a more refined understanding of how independent variables improve model performance and recognize the presence of heterogeneity.

Overall, the aim of this study is therefore to perform a fine grained analysis of gender wage gap for each job title using linear mixed effect regression models.

The structure of the study will be as follows: Chapter 2 provides a review of the related literature, covering analyses of gender wage groups, the dataset employed, research methods, and any disaggregations performed. Chapter 3 explains the methodology used, with a primary focus on linear mixed-effects models, their estimation, model selection, and diagnostic procedures. Chapter 4 explores the dataset used by analyzing the response variable and its effect on various explanatory factors. Chapter 5 performs the data analysis by fitting a linear mixed-effects model, and the subsequent results are presented. Finally, Chapter 6 concludes, summarizing key findings and implications based on the analyses conducted throughout the study.

2 Literature Review

2.1 Administrative datasets

Following the apartheid era, data from household surveys such as October Household Survey (OHS), Labour Force Surveys (LFS) and Quarterly Employment Survey (QES) has been popular in gender wage gap studies. Over the recent years, Kerr and Wittenberg (2019) introduced Post-Apartheid Labour Market Series (PALMS) dataset which a number of studies have utilised to analyse gender wage gap in South Africa (Mosomi, 2019; Fisher et al., 2021). Household surveys have advantages of being representative of the nation and they cover a wide set of questions which are useful in analysing gender wage gap. However, these surveys have some drawbacks of not accurately representing the income distribution due to reporting bias and under sampling of some groups. Thus, Pleace et al. (2023) suggested that administrative data which is more comprehensive and accurately represents each employee would be a better alternative to household surveys data.

Furthermore, the gap between administrative data and household surveys was emphasised in a study of inequality in Uruguay when household survey data for the period between 2008-2016 was used and a decrease in inequality was reported. Then another study was conducted using administrative data for the same period but the results were different showing an increase in inequality (Burdín et al., 2022).

In South Africa, Pleace et al. (2023) conducted a study on wage differential by using data from South African Revenue Service administrative tax covering a period between 2008–2021. Quantile via moments regression was used to analyse the income disparities by gender within the formal economy in South Africa. Overall, findings showed an increase in gender wage gap over the period under study. After analysing the highly skilled industry exclusively, the 90th percentile of the conditional income distribution had the highest gender wage gap and also females were more likely to occupy lower-paid positions. Furthermore, males earned much higher than females at the mean of the income distribution.

Bezuidenhout et al. (2019), also utilised administrative tax data based on matched employer to employee data set panel from 2011 to 2016. This data was obtained from South African Revenue Service (SARS). The objective of the study was to examine gender wage gap between trading firms and domestic firms using Mincerian wage equations. Findings revealed that trading firms have higher gender wage gap compared to that of domestic firms. Consistent with findings from Pleace et al. (2023) there was an overall picture of high gender wage gap despite the results being grouped into either trading or domestics firms.

Though administrative data is not commonly used in the developing countries, Fortin et al. (2017) conducted a study across Canada, Sweden, and the United Kingdom. Administrative annual earnings data covering a period between 1983 to 2015 was utilised. The objective was to analyse the gender wage gap in the top 10% income distribution. The authors concluded that the gender wage gap reduced for the majority of the female excluding those in the richest 0.1% of the income distribution. Their findings also suggested that the gender pay gap in all three countries under study was significantly influenced by the lack of women in high-earning groups.

In France, Bargain et al. (2018) used administrative records to find gender wage gap between public and private sector workers. The study revealed that the gender wage gap was most pronounced among those at the top end of the income distribution in the private sector.

Despite that administrative data provides broad coverage and is not affected by response bias, it may lack some specific information which may be directly related to the study purpose. According

to Pleace et al. (2023), tax records data often lack detailed information on factors like education and experience. However, self-reported data, while more comprehensive, do have significant limitations such as misreporting and under-representation of higher-income individuals. This study will utilise the administrative data from human resource records for National Department of Health in the Eastern Cape province.

2.2 Gender wage gap by occupation

Fisher et al. (2021) argued that though a substantial number of studies have focused on post-apartheid wage discrimination across gender and racial groups, only very few have analysed gender wage gap by occupation. Toczek et al. (2021) added that gender wage gap is likely as a result of occupational discrimination of women thus occupational position is a relevant factor in analysing gender wage gap.

Adeleken and Bussin (2022) argued that occupational gender segregation is key in increased labour force participation of females thereby reducing the gender wage disparity. This was also reported by Steyn and Jackson (2015) that narrowing of gender wage gap is associated with the presence of skilled women in high paying occupations. There is a need to disaggregate earnings data at the level of occupation and industry in order to comprehend the drivers of the gender wage gap (Paterson, 2010). This issue has not been identified as important only in South Africa but in many other countries as well (Gradín, 2021).

Adelekan and Bussin (2018) conducted a study to examine the status of gender wage gap when both females and males have similar work using salary bands. The salary bands had six categories which were determined by the complexity and number of decisions associated with the job. Gender wage gap was then calculated within each salary band. The data utilised for this study was secondary data as it was collected in a routine survey by 21st Century (Pty) Ltd for other purposes. The data was analyzed using descriptive statistics, linear regression, the chi-square test, and the Mann-Whitney rank-sum test. It was concluded that government interventions to empower women are working, as women were represented in all salary bands. However, the gender wage gap was present in all salary bands and statistically significant except for senior management and top management.

Later on, another study was conducted by Adeleken and Bussin (2022) to investigate the dynamics between gender wage inequality and occupational segregation using remuneration data obtained from a survey conducted by 21st Century (Pty) Ltd across 700 companies predominantly in the private sector. The cross-sectional data constituted 10 job families comprising of occupations with similar jobs across six industries. The occupation types reported were defined by 21st Century (Pty) Ltd and these were human capital, compliance and risk, secretarial, executive management, information technology, financial and accounting, logistics and procurement, operational, marketing and sales, and technical and specialist. Analysis was performed using Stata statistical software and the gender wage gap in occupation and industries was analysed at five percentile points and inference was conducted using the Mann–Whitney rank-sum test. Findings revealed that in the male occupations such as executive management, compliance and risk and information technology, males were found to be earning higher while in the female occupation such as secretarial, females were earning higher than males. Furthermore, an increase in the number of females in female occupation was resulting in a decrease in the income.

Another study was conducted where data was disaggregated by occupation using the PALMS dataset covering the period between 1997 to 2015 to examine the gender pay gap by occupation before and after the introduction of the impact of affirmative action policies in South Africa (Fisher et al., 2021). The occupational groups were managerial, professional, technical and clerks. Econo-

metric techniques employed included kernel density estimation, Ordinary Least Squares, and the Blinder-Oaxaca decomposition method. Results of the study showed that the gender wage gap had reduced and to some extent due to the affirmative action policy which might have increased the productive characteristics of women. Furthermore, regardless of the occupation males were earning more than females.

On the other hand, (Mosomi, 2019) analysed gender wage gap at different age groups. Post-Apartheid Labour Market Series (PALMS) data was employed to create cohort data from repeated cross section to examine trends in labour market outcomes and gender wage gap in South Africa. The data covered a period between 1993 to 2015. The cohort data was constructed as a group of individuals with the same birth year and the data was trimmed to only cover individuals between ages of 25 to 50 with the youngest having being born in 1989 and the eldest in 1950. The average earnings for all the people born in the same year was calculated and analysed over the reporting years. The results of the study showed that gender wage gap was smaller for the youngest cohorts. Mosomi (2019) attributed this improvement to enhanced human capital characteristics and opportunities for younger generations.

Looking at gender wage gap at an aggregate level is not enough, some form of disaggregation can play a significant part in explaining gender wage for the individual groupings. The studies above looked at gender wage gap disaggregated by different occupational groupings. However, even after disaggregating data by occupations, Casale et al. (2021) argued that within similar job categories, women were employed at lower grades for instance, in the same occupation women are likely to be teachers while men are head masters. It is also reflected in the dataset for this study that males are more likely to be dentists whilst females are dentist assistants. This issue is addressed by conducting an analysis for each job title, grouping employees in the same grade level. Thus, analysis in this study is disaggregated based on job titles, which represent the narrowest and most specific occupation level.

2.3 Blinder-Oaxaca decomposition

The Blinder-Oaxaca decomposition method has been widely used to measure gender wage gap. This method decomposes the gender wage gap at the mean into explained and unexplained components. It calculates how much women would be paid in a case that they have similar productivity characteristics with men. The explained part is due to explanatory variables in the model such as differences in education and experience while the unexplained part is often taken to be a measure of discrimination. Though traditional studies attributed unexplained component to discrimination, recent studies have begun to point out factors such as leadership, ambition and income expectation in explaining the gender wage gap (Bhorat and Goga, 2013).

A study was conducted to explore the gender wage disparity for part-time and full-time salaried employees after the apartheid era in South Africa (Muller et al., 2009). The analysis was performed using the October Household Surveys (OHSs) covering a period between 1995 to 1999 and from the Labour Force Surveys (LFSs) from September 2001 to 2006. The Blinder-Oaxaca decomposition technique was employed alongside multivariate analysis. It was concluded in the study that gender wage gap existed for both part-time and full-time workers. In addition it was reported that although the extent of the total gender wage gap had decreased over the years, this gap was more pronounced for the part time workers.

The Blinder–Oaxaca decomposition method was utilised again by Bhorat and Goga (2013) to evaluate the gender wage gap for Africans in post-apartheid South Africa between 2001 to 2007 using the Labour Force Surveys (LFSs) data. Findings using the Blinder–Oaxaca decomposition,

revealed that there was no substantial reduction in the gender wage gap for the African cohort for the period between 2001-2007. In addition, the explained part of the gender wage gap has been decreasing, while the unexplained part has not shown a significant increase and accounted for 71% of the wage gap in 2007.

In another study, Kollamparambil and Razak (2016) examined the existence of gender wage gaps across races in South Africa and again the Blinder–Oaxaca decomposition was used to analyse the data. The response variable was log of hourly wage and control variables covered individual specific characteristics, for instance, race, union membership, level of education attained, age, province and occupation. Labour Force Surveys (LFS) data was utilised covering a period between March 2001 to March 2007. It was concluded that gender discrimination is still present though the gender wage gap had decreased between 2001-2007 due to improvements in the observed and unobserved characteristics of females compared to males.

The Blinder–Oaxaca decomposition technique has not been only popular in measuring gender wage gap in South Africa but in other developing countries as well. Si et al. (2021) applied Blinder-Oaxaca decomposition and quantile regression to examine the factors that influence gender wage gap across 12 selected developing countries, using the Skills Toward Employment and Productivity (STEP) dataset collected by World bank. Findings of the study revealed that gender wage gap was evident and it was generally decreasing with earnings. Consistent with other studies the results indicated that advancement in women's education and training, together with child friendly labour policies were the key factors in narrowing down the gender gap.

While traditional method for measuring gender wage gap particularly in South Africa has been the Blinder–Oaxaca decomposition technique, there are criticisms around this method. Muller et al. (2009) reported that when decomposing gender wage gap using Blinder–Oaxaca decomposition, many researchers have been concerned about the high unexplained component which in some studies has been found to be higher than the explained portion. Yu and Fredericks (2017) also found that the unexplained component of the Blinder-Oaxaca decomposition estimates was high in a study to analyse gender employment discrimination. Similarly Grün (2004) employed the Blinder-Oaxaca decomposition method and reported a high unexplained gender wage gap and recommended further exploration of this unexplained gender wage gap in order to give appropriate advice to the policy makers. Furthermore, this unexplained proportion is assumed to be a measure of discrimination. Thus, Bhorat and Goga (2013) argued that the unexplained proportion of gender gap cannot be accredited to discrimination only. It could be differences in characteristics of individuals that are unobserved, such as talent, quality of education or family background. This suggests that the explained factors included in the model are not enough there is need to incorporate the differences in individuals to reduce the unexplained variance.

However, on the other hand, quantile regression is the other method which is being commonly used to measure gender wage gap. Some studies are using this method alongside the Blinder–Oaxaca decomposition. Quantile regression is superior to the ordinary least squares regression in that it estimates the relationship between predictor variables at the median or specific percentiles or quantiles of a response variable while the OLS models use the conditional mean of the response variable. Therefore this method has an advantage of analysing gender wage gap across different points in the wage distribution by estimating specific quantiles. Nevertheless, both quantile regression and the Blinder-Oaxaca method fall short in adequately explaining the substantial unexplained variance observed in gender wage inequality.

2.4 Random effects models

Casale et al. (2021) reported that gender wage gap studies in South Africa revealed a consistent finding that the observable characteristics have failed to fully explain the gender gap in earnings and discrimination in occupations could reflect some of the unexplained gap. Given the criticism around the unexplained part of the Blinder-Oaxaca decomposition, there is a need for a model that accounts for the unexplained portion in the gender wage gap. Therefore, random effects can be employed which account for the heterogeneity observed across employees and other groupings as well as the dependencies that are likely to be in the data (Singmann and Kellen, 2019). In addition, Nagle (2019) emphasised that models that include random effects have an advantage of accounting for random variance in different dimensions of the data.

Looking at different studies on gender wage gap in South Africa, it appears like there is not yet a study which has incorporated random effects to explain the unexplained or discrimination effect. However, in the United States, Xu (2015) analysed the gender based gap in earnings of women in Science, Technology, Engineering and Mathematics (STEM) using linear mixed effects approach and multi-group path analysis. Linear mixed effect approach was chosen over ordinary least squares methods in order to cater for clustered data and repeated measures. The dataset had a hierarchical structure where students were nested in institutional clusters and had repeated measures as it covered a ten year time frame. The dataset was collected from tracking students' education and work experience after completing their bachelor's degree in 1993 in the United States. A significant gap between earnings for males and females in the first ten years was noted.

In another study in Germany, Toczek et al. (2021) analysed gender pay differences over a period of 24 years using a data set based on the German cohort study lidA (living at work). A multilevel analysis was employed using growth curve analysis with time of measurements at level 1 being nested within individuals at level 2. Other statistical analysis methods used were Cramer's V and paired two sample t-test. Findings of the study revealed that gender wage gap still persisted even after controlling for other factors such as occupational status, education, work experience and unemployment episodes.

However, even though these two studies employed mixed models in analysing gender wage gap, the random effects were meant to account for the dependence due to repeated measures. Both studies had nested models which did not include occupational grouping but institutional and individual groupings. On the contrary, this analysis employs a cross-model approach, where observations are not strictly nested, allowing employees to belong to any job title.

At the time of writing this thesis, it appears that a study has not been published that implemented linear mixed effects regression with the addition of random slopes in analysing gender wage gap. For the purposes of this study, persal number (which uniquely identifies each employee) was used as a random intercept to capture the dependence due to repeated measures. In order to address the issue of using broad occupations in analysing gender wage gap, job titles were added to the models as another random intercept. Furthermore, considering the context of this study the gender variable was added to the job title random intercept so as to assess gender wage gap for each job title. In addition, all analysis has been performed using lmerTest package which is an extension of lme4 package (Brown, 2021).

2.5 Summary

From the literature reviewed above it is evident that gender wage gap still persist though some studies reported a reduction. In addition, while the gender wage gap is particularly noticeable

among lower-income earners, it remains significant even for those in the top decile of income distribution.

There are three issues observed from the literature review that this study will try to address;

Firstly, households survey data tends to suffer from response bias though it is widely used in South Africa to measure gender wage gap. In contrast, the dataset used for this analysis primarily consists of administrative records from the human resource data from National Department of Health in Eastern Cape.

Secondly, Gradín (2021) highlighted that the impact of occupations on gender wage gap has not received sufficient attention in research in South Africa. Similarly, Casale and Posel (2011) found that studies based on national micro-data failed to account the fact that within broad categories women were doing different jobs to men and therefore paid differently. However, this study aims to delve into more detailed levels of occupation to address this gap by analysing gender wage gap for each job title.

Thirdly, Blinder-Oaxaca decomposition is being criticised for higher proportions of unexplained variance which has been attributed to be due to discrimination. This study will use random effects which account for the variability not explained by the predictor variables. Therefore the unexplained variable which is under debate for the Blinder-Oaxaca decomposition will be explained by both random intercepts and random slopes.

At the time of this study, there exists limited research on the gender wage gap analysis that covers the concepts of random intercepts and more so random slopes. This study will apply both random intercept and random slopes to analyse gender wage gap at the most granulated occupation level. This is supported by Schielzeth and Forstmeier (2009) as they demonstrated that mixed effects models with random slopes are more superior to models with random intercepts only. This is because random slopes tend to account for the between-individual variation thereby reducing the residual variance.

Compared to traditional methods of analysing gender wage gap, the linear mixed effect regression applied in this study provides more accurate estimates of the effects, improves the statistical power and reduces Type I error and deals with unmeasured factors not explicitly included in the model but affect the outcome variable, (Singmann and Kellen, 2019).

3 Methodology

In many areas of research including the medical, biological, physical, and social sciences, there is an increased collection of data from each subject more than once over time or under different conditions. This is mainly done to improve statistical power and precision. However, such repeated measures are likely to be correlated resulting in violations of the independence assumption which is a very important assumption in many traditional statistical techniques such as multiple linear regression. The shortcomings of traditional methods in analysing this type of data can be avoided by using mixed models or multilevel models. According to Brown (2021), mixed models assess the condition of interest while also capturing the within-subject variation and between-subject variation. If the model is also linear then it is known as a linear mixed model (LMM) and this can be extended to handle non-linear mixed models via Generalized Linear Mixed Models (GLMM).

This study will make use of linear mixed models which cater for models with a continuous outcome variable. The response variable for this study is the earnings variable which is a continuous variable which captures both salaries and allowances for employees. The data used in this analysis constitutes a longitudinal data whereby multiple observations were made on the same employees over a period of 12 years. It is important to note that the dataset is not balanced as the employees were getting in and out of the system at any point.

Mixed models are called "mixed" because they incorporate both fixed-effects parameters which make up the systematic part and random effects which make up the stochastic part (Bates, 2010). A fixed variable is a categorical or continuous variable that includes all the levels of the variable that are of interest. Fixed effects represent the overall population-level average effect which is assumed to be consistent across samples (Winter, 2019; Schweinberger, 2021). On the other hand, a random variable has many possible levels but only a random sample is included in the data.

Random effects capture the stochastic variability that arises from grouping variables. Winter (2019); Singmann and Kellen (2019) reported that these grouping factors cannot be continuous variables but they are inherently categorical. Specifying random effects in mixed models helps to deal with the non-independence issue which often affects multiple regression by accounting for random deviations from the mean of the population and linear associations determined by fixed effects (West et al., 2022; Salinas Ruíz et al., 2023; Bates, 2010).

Generally random effects have many different levels, for instance, each subject or participant can be hundreds or thousands or more depending on the sample size. On the other hand, fixed effects normally have only a few levels. Zuur et al. (2013) argued that if a variable has more than 10 levels, it should be considered as a random variable; if it has fewer than 5 levels, it should be considered a fixed variable; and if it is between 5 to 10 levels, it can be considered both.

3.1 Specification of Linear mixed models

When creating linear mixed models there are different components of the model that have to be specified. In this sub-section, a model with fixed effects only is explained first, then a model with random intercepts only, followed by a model with both random intercepts and random slopes. Afterwards, a decision has to be made whether the random effects will be nested or crossed. It is important to note the key variables for this analysis which are;

- Response variable - total earnings.
- Random intercepts - job title and persal number (uniquely identifies each employee).
- Key fixed variables - gender and years in the system.

3.1.1 Fixed effects only model

Firstly, a fixed effect only model is considered which is equivalent to the ordinary regression model that does not cater for the non-independence in the data. Below is an example given by (Singmann and Kellen, 2019, p.3);

$$\begin{aligned} y_{i,j,k} &= \beta_0 + \beta_\delta X_{i,j,k} + \epsilon_{i,j,k}, \\ & i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \quad k = 1, 2, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right). \end{aligned} \tag{1}$$

Level 1: employee (i)= 1,2...n_j.

Level 2: job title (j)= 1,2,...n.

where $y_{i,j,k}$ represents the measure of the continuous dependent variable (the earnings) for the i^{th} employee (persal_number) and j^{th} job title in the k^{th} gender condition where k is either male or female.

Each β parameter represents the fixed effect of a one-unit change in the corresponding X covariate on the mean value of the response variable, Y, assuming that the other covariates remain constant at some value. These β parameters are fixed effects to be estimated, and their linear combination with the X covariates defines the fixed portion of the model.

- β_0 is the fixed intercept,
- β_δ is the fixed slope,
- $\epsilon_{i,j,k}$ is the residual error.

The third row states that the vector of all residual errors is assumed to follow a normal (Gaussian:$\mathcal{N}$) distribution with a mean of 0 and residual variance σ_ϵ^2.

However, Singmann and Kellen (2019) argued that generally, the fixed-effects model does not account for any dependencies in the model thereby poorly fits the model.

3.1.2 Random intercept model

The fixed effects only model has some weaknesses of violating the independence assumption as it only assumes a single intercept β_0 for all participants. In this study, this implies that model estimates are the same for every employee and also estimates are the same for all job title groups.

However this problem is addressed by adding a vector of the idiosyncratic effects to the fixed effect model, which is represented by S_0 as shown in equation (2). S_0, follows a normal distribution with mean zero and variance $\sigma_{s_0}^2$. This random intercept model allows individual employees and individual job title groups to have their own means. It accounts for correlations across data points that are brought by differences in different employees and job title groups. For instance, this model considers the fact that some employees and job title groups tend to earn more than others. The equations below were given by (Singmann and Kellen, 2019, p.3);

$$
\begin{aligned}
y_{i,j,k} &= \beta_0 + S_{0,i} + \beta_\delta X_{i,j,k} + \epsilon_{i,j,k}, \\
i &= 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \quad k = 1, 2, \\
\boldsymbol{\epsilon} &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right), \\
\boldsymbol{S}_0 &\sim \mathcal{N}\left(0, \sigma_{S_0}^2\right).
\end{aligned} \tag{2}
$$

Level 1: $Y_{ij} = \beta_i + S_i + \beta_\delta X_{ij} + \epsilon_{ij}$.

Level 2: $\beta_i = \beta_0 + U_i$.

3.1.3 Random intercept and slope model

The random intercept model in equation 2, performs better in predicting earnings response for a given employee since it allows each employee to have a different intercept. However, it has shortcomings of not accounting for all potential dependencies brought by different employees. Therefore to capture such dependencies at the level of a given factor, random slopes (S_0) are included which allow the relationship between a fixed variable and the earnings to vary across employees. Barr et al. (2013) argued that omitting random slopes from a mixed model results in exaggerated Type I error rates. Thus random slopes results in a reduction in the residual error. The equations below were given by (Singmann and Kellen, 2019, p.5);

$$
\begin{aligned}
y_{i,j,k} &= \beta_0 + S_{0,i} + (\beta_\delta + S_{\delta,i})\, X_{i,j,k} + \epsilon_{i,j,k}, \\
i &= 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \quad k = 1, 2, \\
\boldsymbol{\epsilon} &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right), \\
\begin{pmatrix} \boldsymbol{S}_0 \\ \boldsymbol{S}_\delta \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{S_0}^2 & \rho_{S_0,S_\delta}\sigma_{S_0}\sigma_{S_\delta} \\ \rho_{S_\delta,S_0}\sigma_{S_0}\sigma_{S_\delta} & \sigma_{S_\delta}^2 \end{bmatrix}\right).
\end{aligned} \tag{3}
$$

where S_{0_i} is the displacement of employee i from β_0, and $S_{\delta_{0\,i}}$ is displacement of the same employee i from the fixed effect.

Now the two random effect vectors are estimated, the random intercept S_0 and a random effects term added to the condition effect, S_δ. In addition, Brauer and Curtin (2018) assumed that the two random effects come from a normal distribution with mean zero, variance σ^2 and a correlation, $\rho S_\delta, S_0 = \rho S, S_\delta$.

3.1.4 Crossed linear mixed model

Mixed effects models have an advantage of allowing multiple random effects groupings. This allows to have additional random intercepts besides the groupings by employees. The objective of the study is to assess gender wage gap in the data set. One way to observe this gap is to put employees who have more similar characteristics in a cluster. So, employees have been grouped according to their job title. This job title variable can be added to the model as an additional random intercept. Depending on the relationship between these random effects and the research question, they can be either crossed or nested.

Brown (2021) defined crossed mixed effects when every subject responds to every item and nested when every subject responds to a different set of items. Nested random effects are sometimes

referred to as multilevel or hierarchical linear models. However, this study will analyse employees and job titles as crossed effects, to be more specific partially crossed and not fully crossed. This is because each employee can be in one or more job title group but not every employee is in every job title group.

3.1.5 Final linear mixed effect regression model

Below is the full mixed model represented by employee and by job title random intercepts as well as random slopes for the selected fixed variables and correlation among employee random effects and correlation among job title random effects. (Singmann and Kellen, 2019, p.6) formulated the equations below;

$$
\begin{aligned}
y_{i,j,k} &= \beta_0 + S_{0,i} + I_{0,j} + (\beta_\delta + S_{\delta,i} + I_{\delta,j})\, X_{i,j,k} + \epsilon_{i,j,k}, \\
& i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \quad k = 1, 2, \\
\boldsymbol{\epsilon} &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right), \\
\begin{pmatrix} \boldsymbol{S_0} \\ \boldsymbol{S_\delta} \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{S_0} & \rho_{S_0,S_\delta}\sigma_{S_0}\sigma_{S_\delta} \\ \rho_{S_\delta,S_0}\sigma_{S_0}\sigma_{S_\delta} & \sigma^2_{S_\delta} \end{bmatrix} \right), \\
\begin{pmatrix} \boldsymbol{I_0} \\ \boldsymbol{I_d} \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{I_0} & \rho_{I_0,I_\delta}\sigma_{I_0}\sigma_{I_\delta} \\ \rho_{I_\delta,I_0}\sigma_{I_0}\sigma_{I_\delta} & \sigma^2_{\boldsymbol{I_\delta}} \end{bmatrix} \right).
\end{aligned}
\tag{4}
$$

where I_0 is the by job title group random intercept and I_δ the by job title random slope for the fixed effects for the J different fixed variable levels. Similar to above equations, for each by employee random effect the corresponding variance is estimated, here σ^2_{I0} and $\sigma^2_{I\delta}$, as well as their correlation $\rho I_0, I_\delta, \sigma I_0, \sigma I_\sigma$.

3.1.6 Correlations among random effects

When random intercepts and slopes are specified, the model will also calculate the correlations associated with those random intercepts and slopes. Brown (2021) claimed these correlations provide key insights about individual differences in condition effects and examining the correlation helps to have a deeper understanding of the data.

Suppose that the correlation between the job title group and the gender slope is negative. Assuming the reference group for gender variable is female, then this would suggest that job title groups with males who have higher random intercepts are more likely to have lower slopes. It is important to check whether the fixed coefficient for the gender variable is negative or positive. If it is positive, then lower slopes indicate less positivity. This means that job titles with higher earnings are less affected by the gender effect.

Negative gender coefficients means job titles with higher earnings than average are affected more by the gender condition and if the correlation is negative then it means it is more affected by the gender effect. Therefore in the context of this study, a negative correlation between job title intercept and gender slope when the gender coefficient is positive imply that job title groups with males earning higher than average tend to have females in the same groups earning less.

3.2 Estimation in mixed models

Having defined the random intercepts and slopes in the models, the next step would be estimating the fixed-effect parameters and the covariance parameters. According to West et al. (2022) there

are two standard methods in mixed models to estimate the fixed-effect and covariance parameters which are Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML).

For any given model and data set, maximum likelihood estimation looks for parameter values that make the model's predicted values more like the observed values. Maximum likelihood estimation can be defined as an estimation method where a likelihood function is optimised to get the estimates of unknown parameters. The first thing in the application of ML estimates is to create the likelihood as a function of the parameters in the given model with the consideration of assumption effects. The values of the arguments that maximize the likelihood function are then referred to as Maximum Likelihood estimates. In the context of mixed models, the likelihood function of fixed effects and covariance is constructed by referring to the marginal distribution of the response variable.

REML estimation often referred to as residual maximum likelihood estimation is also based on the maximum likelihood method. (Salinas Ruíz et al., 2023, p.29) stated that;

> "Instead of maximizing the likelihood function of the original data, it maximizes the likelihood function over a set of errors obtained by removing the variables from the original response to fixed effects, which are assumed to be known."

Furthermore, unlike ML, REML takes into account the loss of degrees of freedom due to estimating fixed effects, thereby giving unbiased estimates of covariance parameters. Generally, in most cases REML is selected over ML estimation. However, it is safest to refit all the models using the maximum likelihood criterion when comparing models varying in fixed effects because REML assumes that the fixed effects structure is correct. Though the standard deviation of random effects are under estimated with the ML method, it is the best option when comparing models with varying fixed effects.

The likelihood ratio test (LRT) is a statistical method used to compare the fit of models. In the context of linear mixed-effects models, it helps assess whether adding or removing specific fixed or random effects significantly improves or worsens the model fit. This method results in model-fit statistics such as Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), the log-likelihood (logLik) and the deviance (negative twice the log-likelihood) as the parameter estimates. All these statistics are useful tools in model selection, as they compare the fit of different models to the same data.

Two widely used criteria for model selection that consider both complexity and model fit are AIC and BIC. Both criteria give a quantitative measure which allows evaluation and comparison of different models. Lower values for both measures indicate a better fit. The aim of AIC is to look for a model that maximises the likelihood of the data and at the same time penalises the number of parameters used. The formula for calculating AIC is as follows;

$$\text{AIC} = -2^* \log(\text{L}) + 2^*\text{k}.$$

L is the maximised likelihood which evaluates how well the model fits the data. The number of parameters are represented by k and this includes the intercept and any additional independent variables. AIC incorporates both the number of parameters and the likelihood to avoid overfitting and excessive complexity, thereby reducing the odds of capturing unnecessary features and noise in the data.

Similarly, BIC incorporates both likelihood and the number of parameters. However, it is based on Bayesian principles and penalises for complexity more than AIC. The formula for calculating

BIC is below;

$$BIC = -2 * \log(L) + k^* \log(n).$$

All the terms are the same as for AIC except for $\log(n)$ which represent the logarithm of the sample size n. This term scales the penalty term by the natural logarithm of the number of observations utilised when estimating the model. Therefore, this introduces a stronger penalty for models with more parameters compared to what AIC does.

3.3 Model selection strategies

Considering the dataset for this study, there are choices to be made with regards to which fixed and random effects to include in the final models. West et al. (2022) claimed that this decision includes striking a balance between statistical considerations as well as subject matter concerns in order to select a model that best matches the observed data. Two strategies suggested by West et al. (2022) are the top down strategy and the step up strategy which will be briefly introduced below.

3.3.1 Top-down strategy

The top-down strategy involves selecting the models by starting with a model that contains the highest of fixed effects under consideration. The first step is to include to the model all the possible fixed effects under consideration. Therefore, it is important at this stage to only select fixed effects and not random effects. Thus the argument for REML will be set to FALSE, as the ML estimate is used for model comparisons of varying fixed effects. After this, the model is reduced by removing one variable at a time with replacement to see the variables which impact the model the most. In this case the most important variables would be identified by very high information criteria when that variable is omitted from the model.

3.3.2 Step up strategy

The other strategy is the step up strategy which involves starting with a simple model and then adding more levels of complexity. It involves models with varying random effects. For this study the initial model can be the best model selected from the top down strategy. This model will have all the fixed variables needed and then vary on the random effects. The persal number intercept only is added first to the model, then the job title intercept. Based on the context of this study, the gender slope is added to job title intercept followed by the years in the system slope and then the nature of appointment slope. The estimations will be done using REML as these models only involve a change in random effects and not in fixed effects so as to obtain the unbiased estimates for the variances of random intercept and slope. Again the information criteria and other model fit statistics are used to determine the best model. The model with least values for information criteria will be considered as the final model.

3.3.3 Variance Inflation Factors (VIFs)

In mixed effect models, correlated predictors can affect the accuracy of estimation and interpretation of the coefficients. This occurs when two or more independent variables in a regression model are linearly dependent resulting in inflated standard errors of coefficients. However, the study

objectives and the extent of the correlation between variables determines how much of a problem it is. In order to identify and address multicollinearity there are different steps and methods that can be used which include correlation matrix and the Pearson correlation coefficient. However, Yu et al. (2015) argued that in a large dimensional data set with many explanatory variables, using bivariate relationship between any two independent variables may not be enough as underlying multivariate relationships may not be captured and suggested using variance inflation factor.

VIF is a more robust method for assessing and quantifying the multicollinearity among explanatory variables in a multiple regression model. The equation for calculating VIF is a general form represented below:

$$VIF = \frac{1}{1 - R_p^2}.$$

Where the subscript p indicates the independent variable and R^2 value indicating the percentage of the variance for a response variable that is explained by an exploratory variable in a regression model. As R-squared increases, the denominator decreases, resulting in an increase in the VIF. Deciding on which threshold to set on the VIF values has been problematic. Akinwande et al. (2015) interpreted VIF equal to 1 as indicating collinearity among regressors. Then, VIF between 1 and 5 indicates moderate multicollineariaty, while between 5 and 10 implies high correlation and above 10 indicates very high correlation that can be a cause of concern. However, Yu et al. (2015) reported that a common rule of thumb among researchers is that any VIF value greater than 5 implies a large multicollinearity problem.

When collinearity has been detected the next step is finding a way of dealing with highly correlated independent variables from the model because they might render other significant variables redundant. The other methods can include shrinking the coefficients of the correlated factors to reduce their variance through regularization methods such as ridge regression or lasso regression. Other methods can include using Principal Component Analysis (PCA), factor analysis or Bayesian methods. However in this study, a comparison of different models with and without the correlated variables will be done using likelihood ratio test. In addition, much consideration will be given to the meaning and relevance of the correlated variables in relation to the context of the study.

3.4 Model diagnostics

Loy et al. (2017) pointed out that model diagnostics plays a vital role in statistical modelling by making sure that the assumptions needed for valid inference are not violated. This approach involves evaluating how much the model accounts for the characteristics of a given dataset.

Just like any other statistical methods, there are specific set of assumptions that should be satisfied when employing linear mixed effect regression models. Due to the complexity of these models, the assumptions extent those of the Ordinary Least Squares (OLS) in order to handle clustered data or repeated measurements. For instance for both OLS and mixed models, residuals form the core of diagnostics though the residual analysis is more complex for mixed models. Loy et al. (2017) suggested that this could be because mixed models include residuals at different levels which may be associated with different aspects of the model.

There are many ways that can be used to do this assessment which includes numerical methods and graphical diagnostics. Numerical methods usually depend on t-statistics and p-values in assessing the strength of the evidence and they show the degree of the problem in the model. However, on

the other hand the graphical methods go beyond assessing the degree of the problem and give an idea of what may be the source of the problem. Moreover, they are also useful in complex models and provide insights to the nature of the violations which on the contrary cannot be done using numerical methods.

However, Cho et al. (2023) argued that graphical diagnostic plots can have a weakness of being subjective. In that situation one would opt for numerical or formal statistical tests such as the Levene tests, Shapiro-Wilk test and Kolmogorov-Smirnov test. But these tests have their weakness of being so stringent that in most cases it is difficult to pass the tests. Venables et al. (2002) supported this argument and suggested that in such cases then using graphical diagnostic plots would be better.

Following this, graphical diagnostics will be used to assess if the model assumptions are being upheld and get insights on the nature of the problems that might be in the models.

3.4.1 Linear mixed models assumptions

Different studies mention different linear mixed model assumptions but this study will focus on the three most common assumptions;

Assumption 1 - Linearity

A regression analysis aims to fit a linear equation that best explains the observed data. Just like OLS, linear mixed models assume that the relationship between predictors and the dependent variable is linear. In linear mixed models, this means that the fixed effects should have a linear impact on the response variable. This can be plotted using the residuals (differences between observed and predicted values) against the fitted values (predicted values from the model). Deviations from linearity can be observed as forming a particular shape but if they are random then it means linearity assumption is satisfied.

Assumption 2 - Homogeneity of variance

Heteroscedasticity points the pattern in which the variability of a variable is not equal over a number of values of another variable that explains it (Cho et al., 2022). This assumption implies that predictors at one level are not related to errors at another level. A plot of predicted values vs residual values can be used to check the homogeneity of the variance. As stated by Pinheiro and Bates (2000) a non-random pattern in the plot can be a suggestion of heteroscedasticity in the model. There should be an even spread around the centred line for this assumption to hold.

Assumption 3 - Normality of Residuals

Though the OLS model does not require the outcome variables to be normally distributed, the linear mixed models assume that the residuals of the analysis are normally distributed. Residuals for this assumption can be checked using a histogram or a quantile-quantile (QQ) plot. This study makes use of the QQ plots to assess the normality assumption of the residuals. If the normality assumption holds, then the residuals should fall along a straight line on the QQ plots. However, deviations from this straight line implies that the normality assumption is not met.

Assumption 4 - Normality of random effects

Linear mixed effect regression includes random effects that is random intercepts and/or slopes to capture variability across different groups or levels. Therefore, normality assumption in mixed effect models does not apply only to residuals but random effects as well. Thus, it is important for random intercepts and random slopes to also meet this assumption. In this study, the random

effects to be tested would be the subject (each employee) random intercept, grouping random intercept (job title) and random slopes (gender and years in the system). Furthermore, these plots will be used to identify outliers especially in the random effects that are of more interest in the context of this study (job title and gender).

3.4.2 Violation of the assumptions

Mixed-effects models involve complex assumptions about the distribution of residual and random effects. In real datasets, violations of these assumptions are common, although there is no concrete understanding of how these violations affect the accuracy of the estimates. In a study by Schielzeth et al. (2020) to assess the impact of violating distributional assumptions in mixed effects models, the results showed exceptional robustness. Thus, they concluded that researchers should have some leeway to use mixed effect models even if the assumptions are violated. However, Knief and Forstmeier (2021) argued that though violation of the normality assumption has limited and manageable risk, researchers should be more cautious not to violate the other assumptions as their violations will be more riskier.

Furthermore, Knief and Forstmeier (2021) also cautioned against using small samples as they are more likely to compromise the robustness of the estimates. In a study conducted by Knief and Forstmeier (2021), the study parameters were unbiased and precise only if sample sizes were not small implying that normality tests matter less for large samples. Considering the dataset of this study which is quite a large sample there are not much concerns except for sample sizes for the specific groupings.

In order to fit the linear mixed effects models, lme4 and lmerTest packages will be used for this study (Kuznetsova et al., 2017).

4 Data Exploration

The consensus from the studies on gender wage gap in South Africa shows that gender wage gap still persists (Oyenubi and Mosomi, 2024). Thus, Clinton Health Access Initiative (CHAI) partnered with National Department of Health to assess gender wage gap in the health workforce. The dataset used for this study was retrieved from human resource data from the National Department of Health in the Eastern Cape province. This dataset includes basic personal information like demographic information, occupational level, employment history and compensation.

This chapter will summarise the full dataset used in the study, followed by exploration of earnings by gender for different independent variables, descriptive analysis of the gender promotion gap and then preparation of the dataset for model building.

4.1 Description of the full dataset

The dataset under analysis was obtained via Clinton Health Access Initiative, (CHAI) comprising 604 448 observations. Notably, 75% of these observations correspond to females. Since the data was collected over 12 years, it is possible that employees enter and leave the system at any period. Looking at the unique identifier which is the persal number, it can be noted that over the reporting period there were 91217 unique employees of which 73% were females. Furthermore, about 24% (21 755) of these employees have been in the system for the 12 year reporting period.

Although there are 58 variables in the dataset, this study will mainly utilise the following variables;

- persal number,
- facility name,
- district,
- rural urban,
- race,
- age,
- occupational group,
- occupational classification,
- job title,
- reporting year,
- years in system,
- gender,
- citizenship,
- nature of appointment,
- total allowance,
- notch value.

Table 1 shows summary statistics for the numeric variables of interest. The annual average salary represented by notch value is R171 371 while the annual average allowance is R17 774. On average,

employees are approximately 42 years old, and the majority have been part of the system for a duration of six years.

Table 1: Summary statistics for numeric variables.

variable name	mean	standard deviation	median	min	max
notch value	171,371	182,451	128,535	0	2,471,334
allowance total	17,774	19,093	12,738	0	653,640
age	42	11	42	16	92
years in system	6	5	5	1	20

4.2 Description of the response variable

The goal of this study is to analyse gender wage gap, this implies that the response variable should be earnings. Looking at the data set, there are two earnings variables; the notch value and total allowance. The notch value in the context of South Africa determines government employee salary levels categorized according to job levels and this is set by the Department of Service and Public Administration (DSPA). On the other hand total allowance is the total of all monetary benefit given to the employee on top of the basic salary and this include overtime, travelling expenses, medical and house rent. For the purpose of further analysis, the notch value and total allowance for each observation are summed up to give the earnings variable which is the dependent variable. It is important at this point to clarify that the unit of analysis is the observation which is data for each employee for each year.

Since the data was collected over a 12 year period, it is crucial to account for the impact of inflation and cost of living over those years. In most cases, the Consumer Price Index (CPI) which measures the change in consumer prices over time is used to adjust the nominal values to real values. In order to calculate the real earnings from 2010 to 2021, the average CPI values for each year provided by StatsSA (2024) were used. In South Africa, the CPI is calculated using a base year as a reference point and as of now the base year is currently set at 2021 which is what was used for adjustments in this study. This therefore means that the CPI reflects price changes relative to the cost of goods and services in December 2021. Thus, for the earnings to be comparable over the 12 year period, the notch value and allowance values were divided by the average CPI for that year and then multiplied by 100.

$$\text{Real Value} = \text{Nominal Value} \ / \ \text{Price Index x } 100.$$

Before delving into data exploration, it is important to note that all the earnings values were calculated over an annual period. The average earnings for the full dataset as observed from Table 2 is R242 439 which is higher than the overall median (182 069), implying that the distribution of earnings is skewed to the right.

The few outlying data points on the high end reflected by a very high maximum value (R2 952 748) is pulling the mean up resulting in the mean having a higher value than the median. Further analysis of this data revealed that those top earners were mainly the Heads of clinical department and the majority of these top earners were males. This is confirmed by a higher maximum value for males (R2 952 748) compared to females (R2 759 962. Considering the mean of the income distribution, men had notably higher incomes than women (R272 879 vs R232 323).

Surprisingly, looking at the minimum earnings, there is a minimum value of zero for both females and males. This suggests that there are employees with zero earnings in the data set.

Table 2: Earnings summary statistics.

Gender	Min	Q1	Median	Mean	Q3	Max
Female	0	115,979	184,824	232,323	303,317	2,759,962
Male	0	122,262	175,206	272,879	294,263	2,952,748
Overall	0	118,871	182,069	242,439	301,657	2,952,748

The DPRU (2024) reported that the National Minimum Wage (NMW) is R25.42 per hour as of March 2023. Considering the minimum wage, it appears unusual to encounter observations where earnings are zero. Further analysis revealed that 1.7% (9988) of all the observations in the data set had zero earnings and this is about 7% of the total unique employees. In addition, all the zero earners had different occupational groups, facilities, citizenship and 72% of them were females. However, about 95% of these observations had an abnormal nature of appointment and the remaining 5% received periodical remunerations. This implies that it is conceivable that, due to the nature of their appointment, some employees may not have worked during that specific year, resulting in zero earnings.

A closer analysis of the data also confirmed that most of these employees were getting earnings for other years, only about 9% had never earned anything over the reporting period and though some have been in the system for more than a year. As a result, all observations with zero earnings are excluded from the analysis.

Considering the remainder of the employees, the distribution of earnings for males and females indicates some skewness for female employees concentrated on the lower end while only a few having high earnings compared to male earnings indicated by orange colour. Therefore, Figure 1 suggests that a greater proportion of females have lower earnings compared to males, while a larger proportion of males dominate the high earning range.

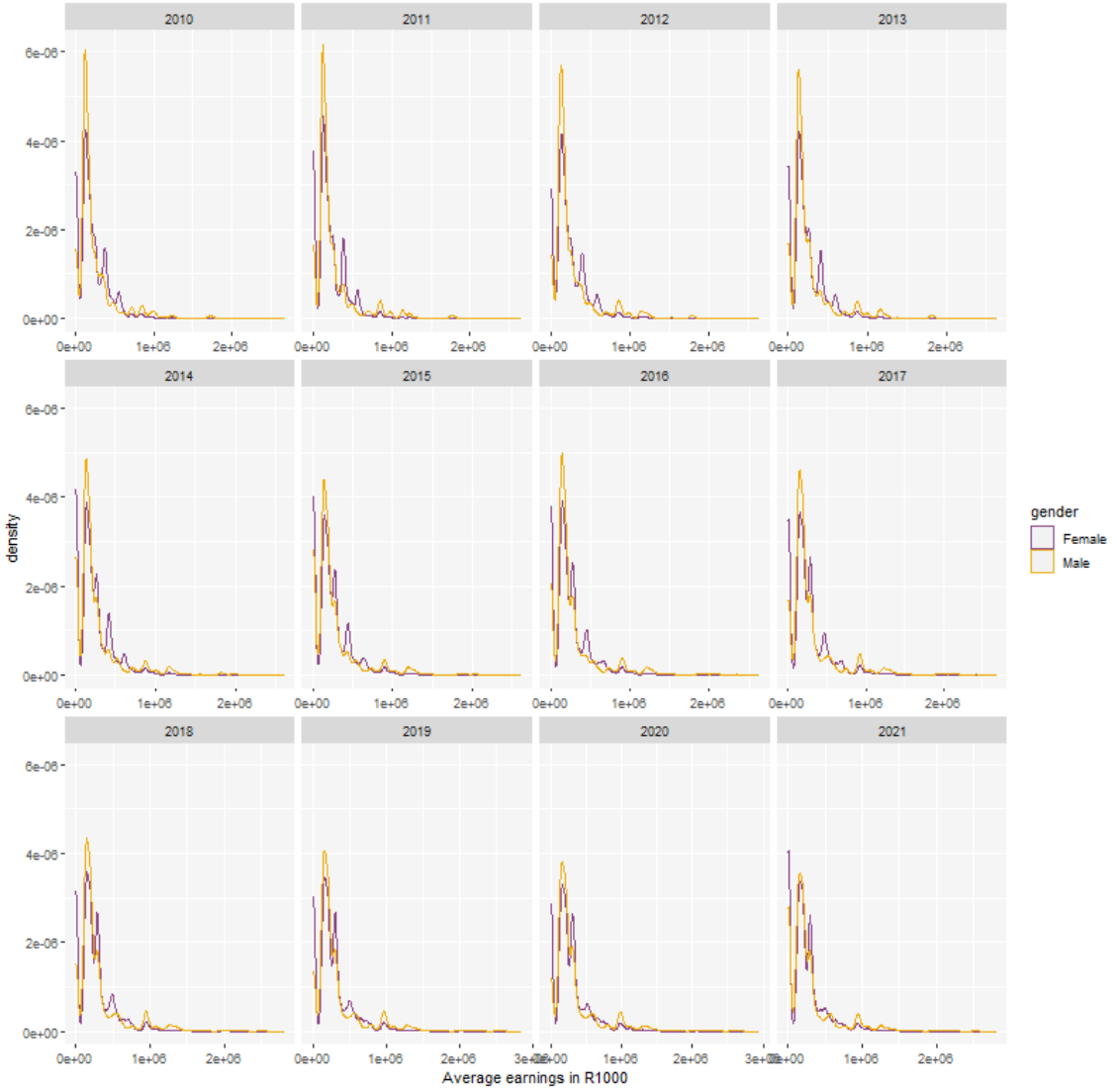

Figure 1: Distribution of earnings by gender for each year.

The gender wage gap can be broken down further by considering the difference below and above the 10th and 90th percentiles as well as between the 10th and 90th percentiles. Looking at the overall average wage analysis in Figure 2, there is a significant gender wage gap against females over the 12 year period. This gender wage gap was more or less the same from 2010 to 2014 but it started to widen from 2015 to 2021. This is consistent with findings reported by Pleace et al. (2023) who covered almost the same period between 2008 to 2021 and reported widening of the gender wage gap.

Looking at the plot for employees earning above the 90th percentile, gender wage gap against

females is more pronounced. This confirms what was observed earlier on that a larger proportion of males are among the highly paid employees. In a study conducted by Pleace et al. (2023) findings revealed that gender wage gap was greatest in the 90th percentile category. Similarly, out of all the companies listed on the Johannesburg Stock Exchange as at April 2019, female executive directors were earning an average of 74.5% of their male counterparts earnings (PricewaterhouseCoopers, 2019). This shows that though women are finding their way up to higher echelons in the corporate industry, gender wage gap still persists even at higher executive levels.

However, on the other end there is no distinct gender wage gap for observations below the 10th percentile. Of note, is a significantly lower average wages for reporting year 2018 for males compared to females. Further analysis shows that 99,98% of observations below the 10th percentile were under abnormal nature of appointment. This type of appointment is irregular and employees receive ad hoc employment and usually there is no structure of payment nor duration of work. Looking specifically for year 2018, averages wages for males are lower compared to females in the same year because a large proportion of females are employed as learnership nurses who are paid more than internship carpentry and plumbing where the majority of males belong.

Furthermore, it is interesting to note that females earn significantly higher than males between the 10th and 90th percentiles. This is probably due to the dataset being predominantly composed of females in the nursing occupation, who typically have lower earnings. Overall, females tend to dominate lower-earning occupations, while males are more prevalent among high earners.

It can also be observed from the majority of the plots in Figure 2 that earnings for year 2021 are very low. A closer analysis revealed that the reporting date for all other years was 31 March of that year but for the 2021 data it was 31 January implying that two months data for that reporting year was missing. Thus, in order to have data which is comparable for all the years, observations for 2021 were excluded from the regression analysis. This was covering about 9.4% of the data set.

Figure 2: Average wages over the years for different percentiles.

4.3 Gender wage gap for independent variables

Given the observed gender wage gap between females and males, it is important to analyse which factors play a significant role in this wage gap. This subsection will show how gender wage gap is distributed for the following independent variables;

- race,
- citizenship,
- years in the system,
- age,
- facility,
- district,
- nature of appointment.

4.3.1 Gender wage gap by race

Though Africans are constituting 88% of the population, they are the least paid race. Looking at Figure 3, the wage gap is more pronounced for higher earnings groups which are whites and Asians. Consistent with previous observations, higher earnings seem to exhibit a stronger positive correlation with the gender wage gap.

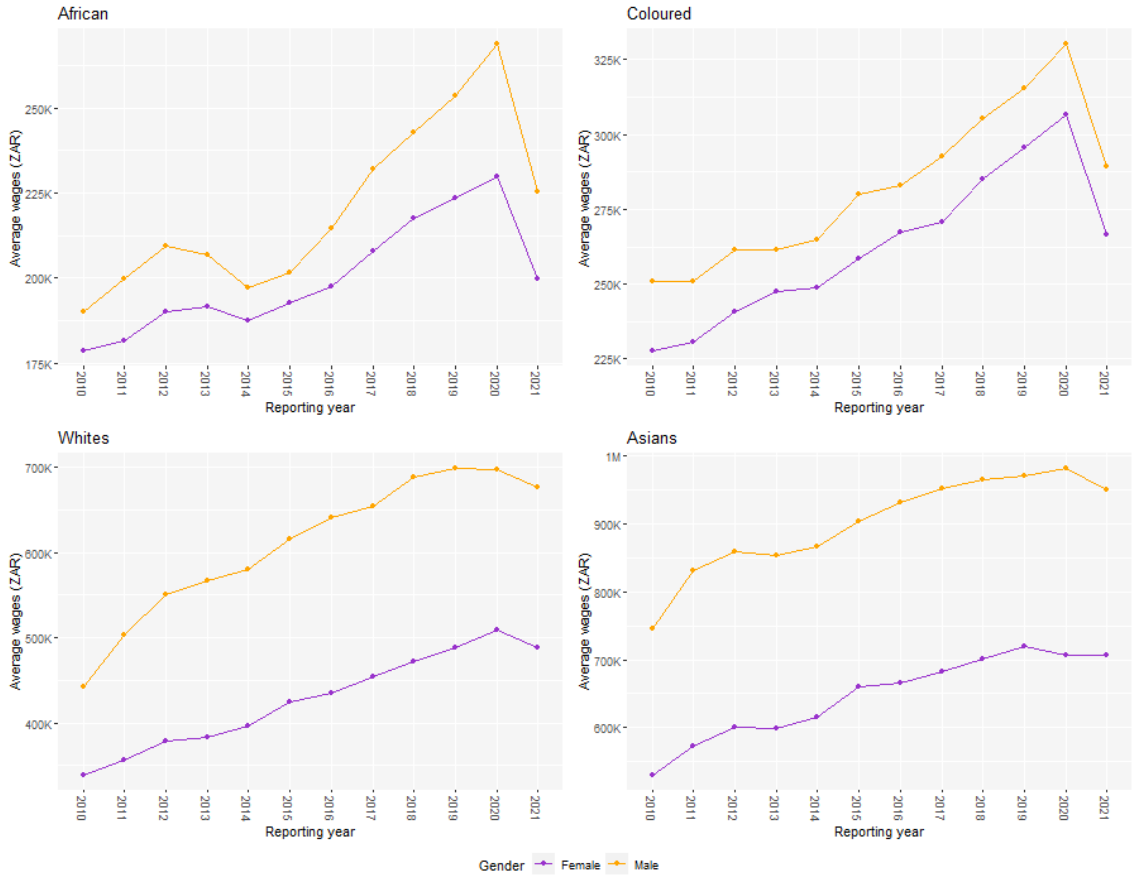

Figure 3: Gender wage gap by race.

4.3.2 Gender wage gap by citizenship

It can be observed from Figure 4 that average earnings for South African citizens are much lower than for Non-South Africans. Upon closer analysis of the data, it becomes evident that a significant proportion of foreign workers hold highly skilled positions such as medical specialists which are well compensated. This aligns with previous observations indicating that gender wage gap is more pronounced in the upper decile.

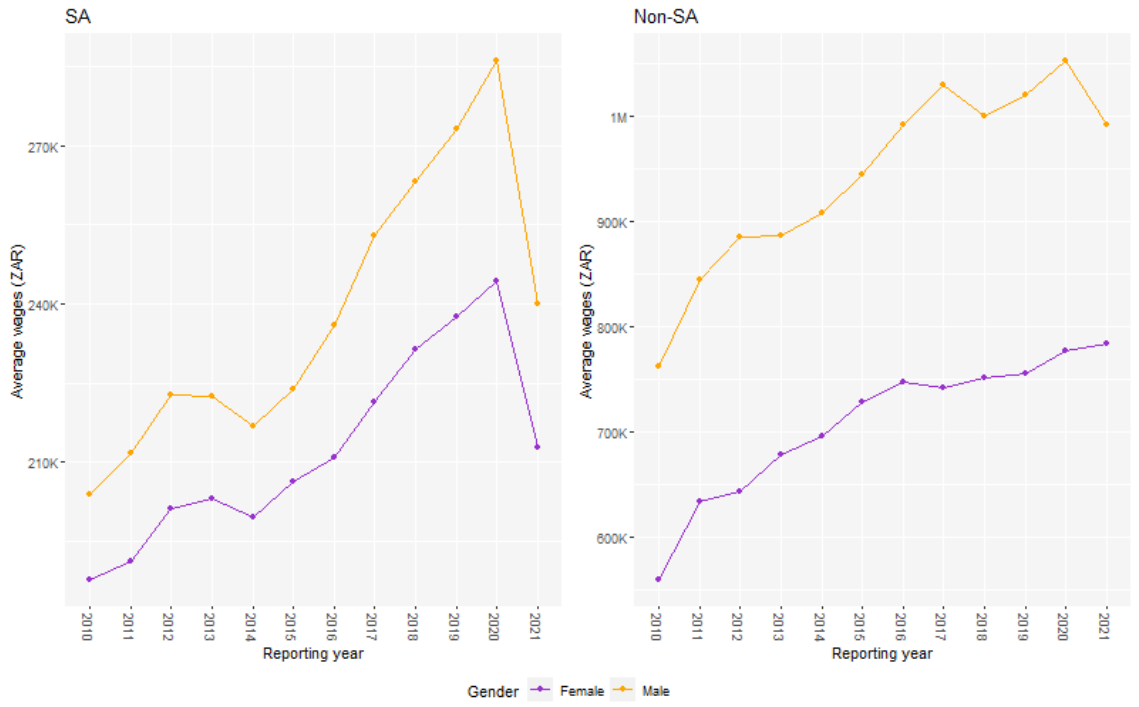

Figure 4: Gender wage gap by citizenship.

4.3.3 Gender wage gap by years in the system

Figure 5 shows that gender wage gap against females is gradually improving as the number of years in the system is increasing. This is reflected by a significant gender wage gap in favor of females for employees who have been in the system for more than 15 years. From a similar perspective, generally females tend to exhibit greater job tenure and less frequent job changes compared to men. Consequently, this longevity within the system may contribute to higher earnings for females relative to males, particularly among those who have remained in the system for an extended duration.

Figure 5: Gender wage gap by years in the system.

4.3.4 Gender wage gap by age

It can be observed from Figure 6 that gender wage gap is more noticeable for ages above 35 years. The gender wage gap for those below 35 years is relatively small, with females earning slightly higher than males for ages between 16 to 25 years. This could be because young workers are more likely to share similar educational background and work experience. For instance, younger employees often participate in internships, and the remuneration for such internships tends to be relatively consistent.

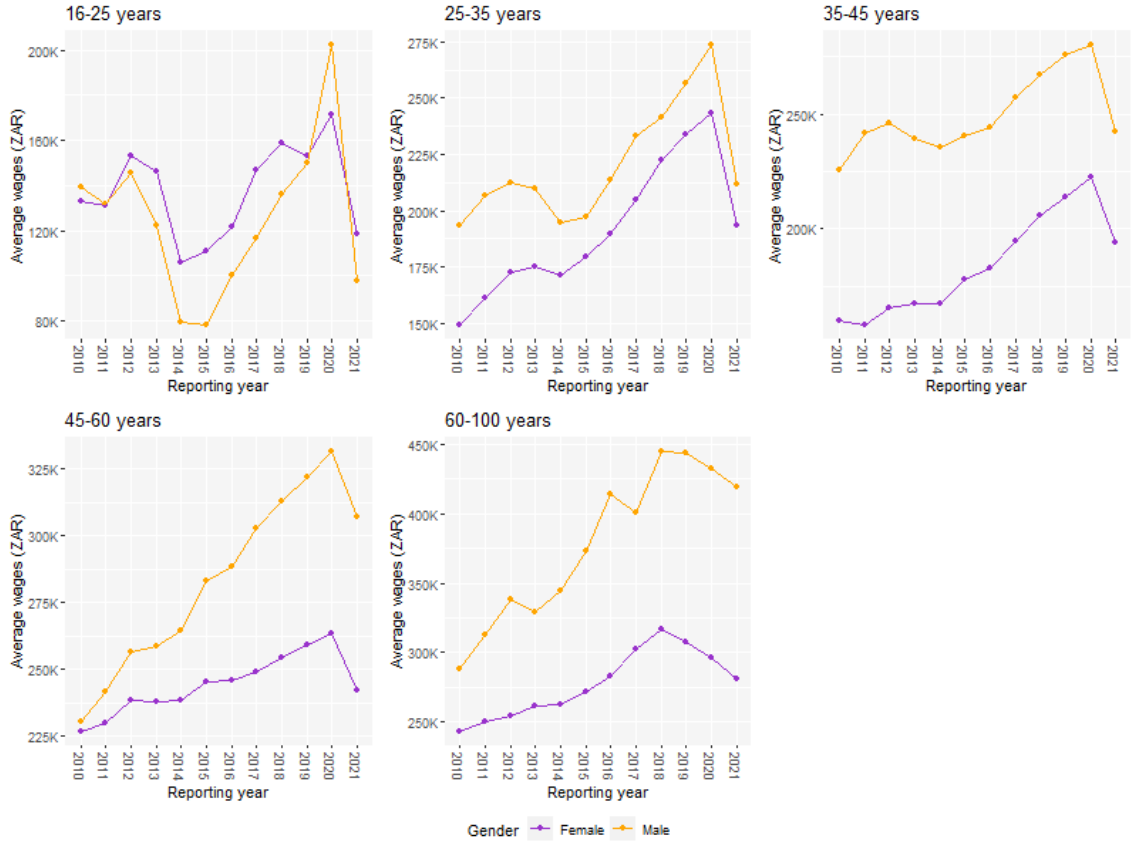

Figure 6: Gender wage gap by age.

4.3.5 Gender wage gap by facility

Looking at Figure 7, it appears that generally males are earning more than females for most of the facilities. In particular, gender wage gap against females is predominant in the district office and mortuary facilities. The district office tends to constitute high earners such as directors and managers while the mortuary is typically male-dominated, with females often occupying lower positions such as receptionists and cleaners. As a result, there is a possibility that this contributes to the disparity in earnings, where males in these facilities earn more than females.

On the other hand, between 2010 and 2017, female earnings at the clinic facility were notably higher than those of males. Given that a significant proportion of the dataset consists of females and nurses, it is common for clinics to be predominantly staffed by nurses while more highly skilled professionals such as medical practitioners are employed at a hospital level. Consequently, this dynamic is likely to result in higher earnings for females compared to males at clinic facilities.

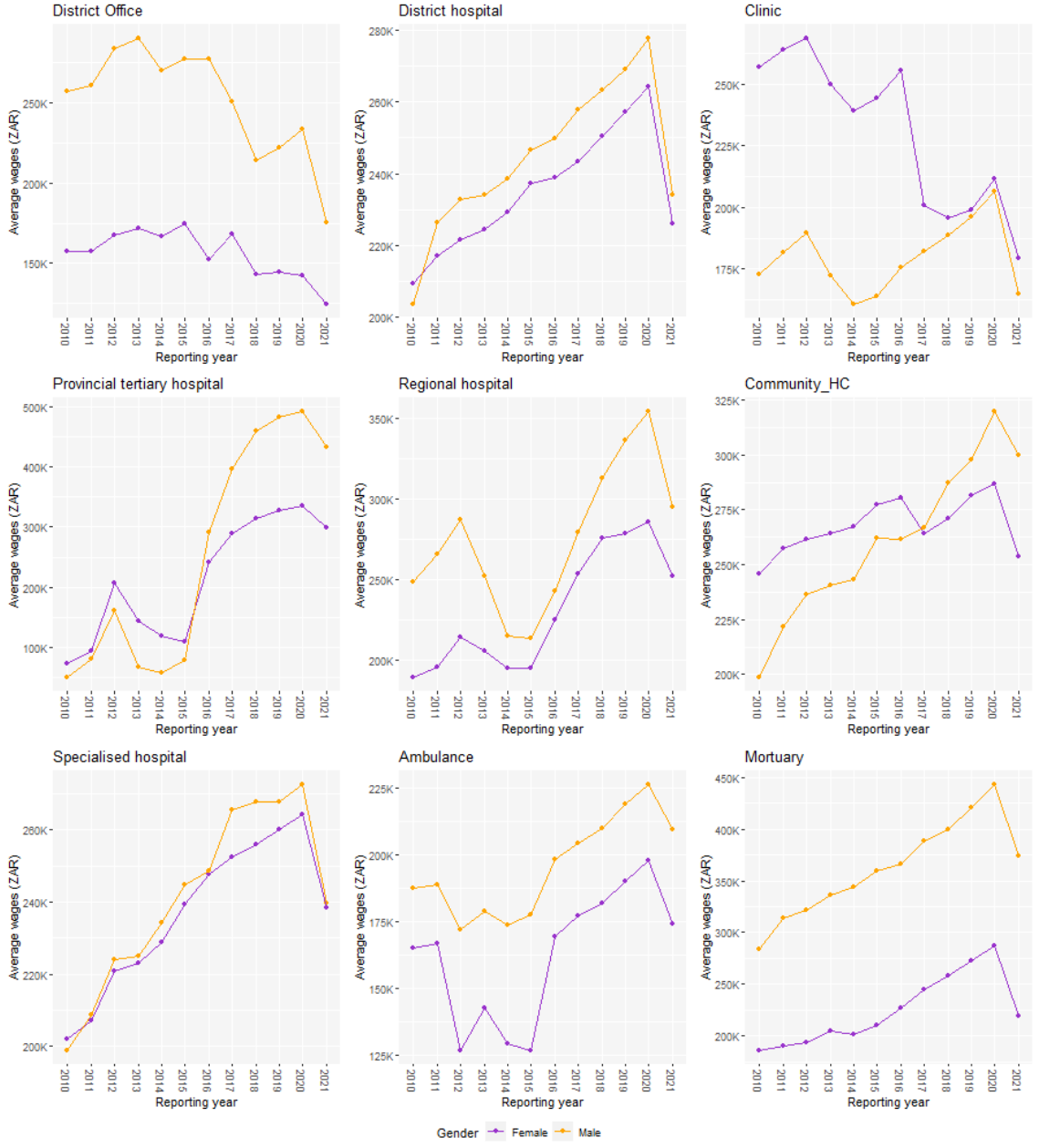

Figure 7: Gender wage gap by facility.

4.3.6 Gender wage gap by district

It can be observed from Figure 8 that considering all the nine districts, male employees are dominating the earnings. This is more evident for the Mandela and Sarah Baartman districts where the gender wage gap is significant throughout the reporting period.

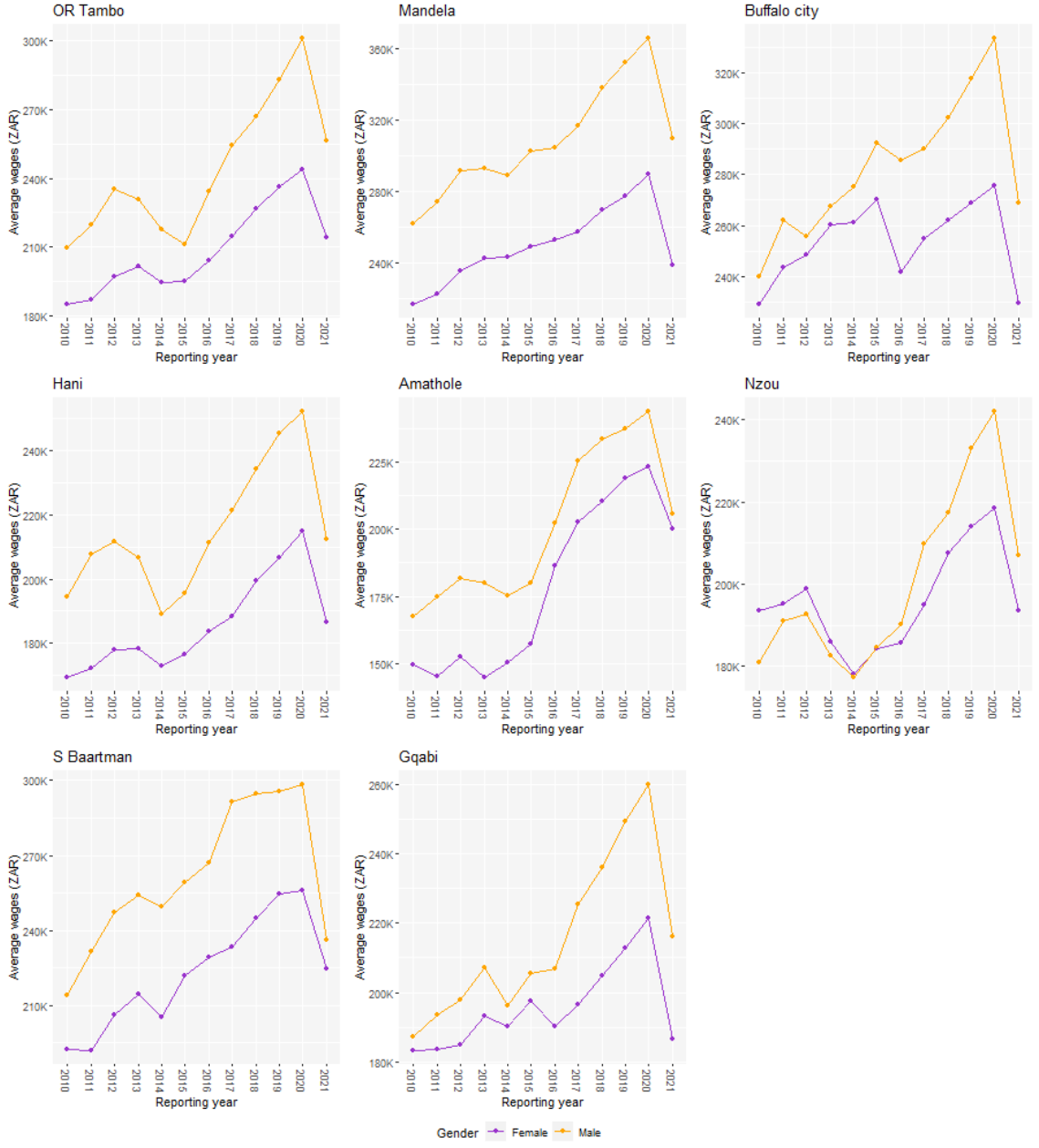

Figure 8: Gender wage gap by district.

4.3.7 Gender wage gap by nature of appointment

The nature of appointment variable involves the type and duration of the job an individual is assigned to. In this dataset there are ten different appointments, with the most populated being the permanent position (65%), followed by abnormal appointment (18%) as shown in Table 3.

Pleace et al. (2023) claimed that women are more likely to struggle to get employment once actively searching compared to males. Furthermore, women often prioritize proximity and convenience (Steyn and Jackson, 2015; Blau and Kahn, 2017). All these factors lead females to remain in their jobs for longer durations. This tendency of women staying longer in their roles allows them to accumulate experience and tenure, over time this can lead to incremental salary increases. Therefore, this is likely to create a gender wage gap that favors women for permanent positions as seen in Figure 9. Furthermore, this aligns with the findings related to the years in the system that females tend to earn more than their male counterparts when they remain in the system for an extended period.

Looking at other nature of appointment plots in Figure 9 such as probation, contract, session, temporary and part-time, gender wage gap against females is evident. However, other nature of appointment groups such as casual and politicians are not relevant to the study since they consists of individuals of the same gender, which inhibits gender comparison. In addition, abnormal and periodic appointments have significantly lower earnings compared to the rest of the groups. It is likely that these appointment groups have irregular earnings since they are employed on ad hoc basis, thus their earnings might not represent annual income. A closer analysis of the abnormal appointment revealed that 83% of its observations are nursing and support personnel such as home-based personal care workers.

Following the discussions above, further analysis will only consider permanent, probation, contract and temporary categories as the rest of the appointments are not comparable.

Table 3: Gender wage gap by nature of appointment.

	Between percentages (frequencies)			**Within percentages**	
Nature of appointment	**Female**	**Male**	**Total**	**Female**	**Male**
Permanent	64.76% (289151)	66.87% (98947)	65.29% (388098)	74.50%	25.50%
Abnormal	19.82% (88490)	11.97% (17712)	17.87% (106202)	83.32%	16.68%
Probation	10.98% (49018)	13.64% (20181)	11.64% (69199)	70.84%	29.16%
Contract	4.13% (18419)	5.78% (8548)	4.54% (26967)	68.30%	31.70%
Session	0.26% (1183)	1.60% (2369)	0.60% (3552)	33.31%	66.69%
Temporary	0.02% (90)	0.11% (166)	0.04% (256)	35.16%	64.84%
Part-time	0.03% (136)	0.01% (12)	0.02% (148)	91.89%	8.11%
Casual	0.00% (0)	0.01% (13)	0.00% (13)	0.00%	100.00%
Periodical	0.00% (3)	0.01% (10)	0.00% (13)	23.08%	76.92%
Politicians	0.00% (7)	0.00% (5)	0.00% (12)	58.33%	41.67%
Total	**100.00% (446497)**	**100.00% (147963)**	**100.00% (594460)**	**75.11%**	**24.89%**

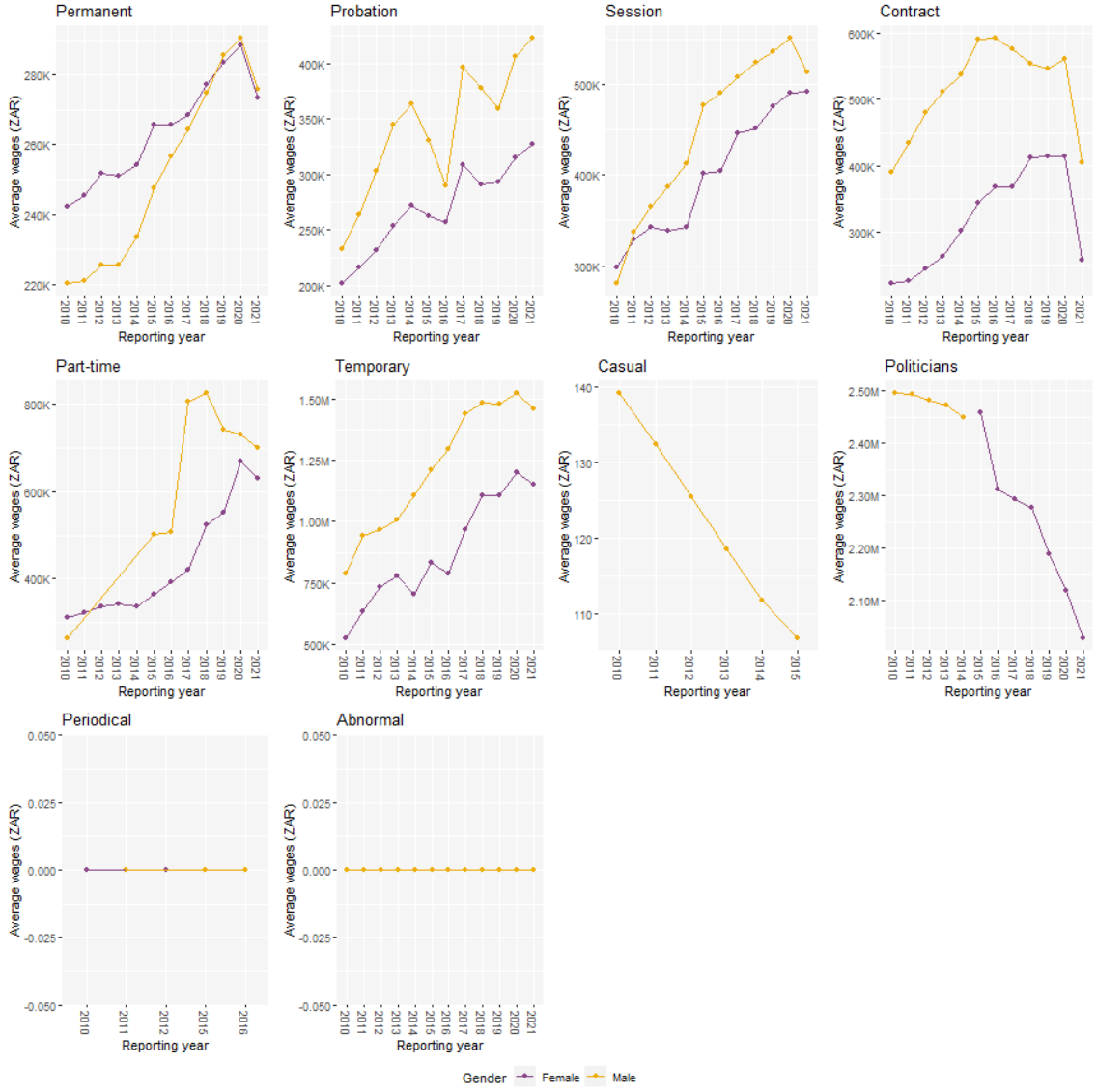

Figure 9: Gender wage gap by nature of appointment.

4.4 Gender wage gap by occupational groups

One of the key variables that contribute to the gender wage gap disparity is the differences in the types of jobs and industries that men and women work. According to Blau and Kahn (2017), females are often concentrated in industries or occupations that have lower pay. Therefore, it is crucial to investigate the variables related to the different industries and occupations in this dataset. The following variables can help to point out the occupations and industries employees are in;

- occupational group,
- occupational classification,
- unit field,
- position title,
- job title.

All these variables are categorical variables, with the occupational group having 25 levels and the job title having up to 1084 levels. This subsection will only consider the variables with the least number of levels which is the occupational group. This variable classifies employees in the data set according to the kind of work they generally perform or their professions and related activities.

Table 4 shows that the most populated occupational group is professional nurse, constituting about 22% of the total population with 89% females. It can be noted that other nursing groups are also highly populated and are mostly dominated by females. For example, the study population consists of four distinct groups of nurses: professional nurses, nursing and support, nursing assistant and staff nurse. Collectively, these groups account for the majority, comprising approximately 56% of the total population. Notably, 87% of these individuals are female.

Looking at Figure 10, gender wage gap in favour of females is evident for a number of occupational groups and this includes, professional nurses, safety related, staff nurse and health related. Notably, the gender wage gap is most distinct within the professional nurses category. This huge gap might be likely due to male nurses gravitating towards promotions and leadership positions, which tend to pay more thereby contributing to the gender wage gap in favour of females.

However, there are other occupational groups which are male dominated to the extent that they do not have any females such as occasional employee, ship's support and economic advisory. These occupational groups with single gender are not relevant to the context of this study. In addition, a closer analysis of the political science occupation showed 12 observations, that is one employee for each year. From 2010 to 2014 there was a male employee who was then replaced by a female employee from 2015 to 2021.

Table 4: Distribution of occupational groups by gender.

	Between percentages (frequencies)			Within percentages	
Occupational group	**Female**	**Male**	**Total**	**Female**	**Male**
professional nurse	26.2% (116891)	10.0% (14869)	22.2% (131760)	88.7%	11.3%
health associated sciences and support personnel	15.9% (70838)	18.5% (27305)	16.5% (98143)	72.2%	27.8%
nursing and support personnel	17.9% (79940)	9.3% (13688)	15.8% (93628)	85.4%	14.6%
nursing assistant	12.8% (57325)	6.3% (9372)	11.2% (66697)	85.9%	14.1%
management and general support personnel	7.6% (33737)	14.5% (21427)	9.3% (55164)	61.2%	38.8%
staff nurse	7.7% (34527)	3.8% (5590)	6.7% (40117)	86.1%	13.9%
medical sciences and support personnel	4.1% (18337)	12.1% (17895)	6.1% (36232)	50.6%	49.4%
emergency service and related personnel	2.0% (8819)	13.2% (19464)	4.8% (28283)	31.2%	68.8%
administrative line function and support personnel	4.3% (19151)	5.1% (7499)	4.5% (26650)	71.9%	28.1%
artisan and support personnel	0.2% (983)	3.7% (5500)	1.1% (6483)	15.2%	84.8%
agricultural related and support personnel	0.3% (1165)	1.7% (2517)	0.6% (3682)	31.6%	68.4%
social services and support personnel	0.5% (2193)	0.1% (215)	0.4% (2408)	91.1%	8.9%
information technology and related personnel	0.1% (614)	0.4% (615)	0.2% (1229)	50.0%	50.0%
communication and information related personnel	0.2% (957)	0.2% (262)	0.2% (1219)	78.5%	21.5%
medical technology and support personnel	0.1% (436)	0.4% (624)	0.2% (1060)	41.1%	58.9%
natural sciences related and support personnel	0.0% (223)	0.4% (634)	0.1% (857)	26.0%	74.0%
engineering related and support personnel	0.0% (75)	0.2% (334)	0.1% (409)	18.3%	81.7%
safety and related personnel	0.0% (190)	0.0% (60)	0.0% (250)	76.0%	24.0%
human resource and support personnel	0.0% (76)	0.0% (35)	0.0% (111)	68.5%	31.5%
legal and support personnel	0.0% (9)	0.0% (23)	0.0% (32)	28.1%	71.9%
occasional employee	0.0% (0)	0.0% (13)	0.0% (13)	0.0%	100.0%
political office- bearers	0.0% (7)	0.0% (5)	0.0% (12)	58.3%	41.7%
ship's and support personnel	0.0% (0)	0.0% (12)	0.0% (12)	0.0%	100.0%
economic advisory and support personnel	0.0% (0)	0.0% (5)	0.0% (5)	0.0%	100.0%
general worker	0.0% (4)	0.0% (0)	0.0% (4)	100.0%	0.0%
total	**100.0% (446497)**	**100.0% (147963)**	**100.0% (594460)**	**75.1%**	**24.9%**

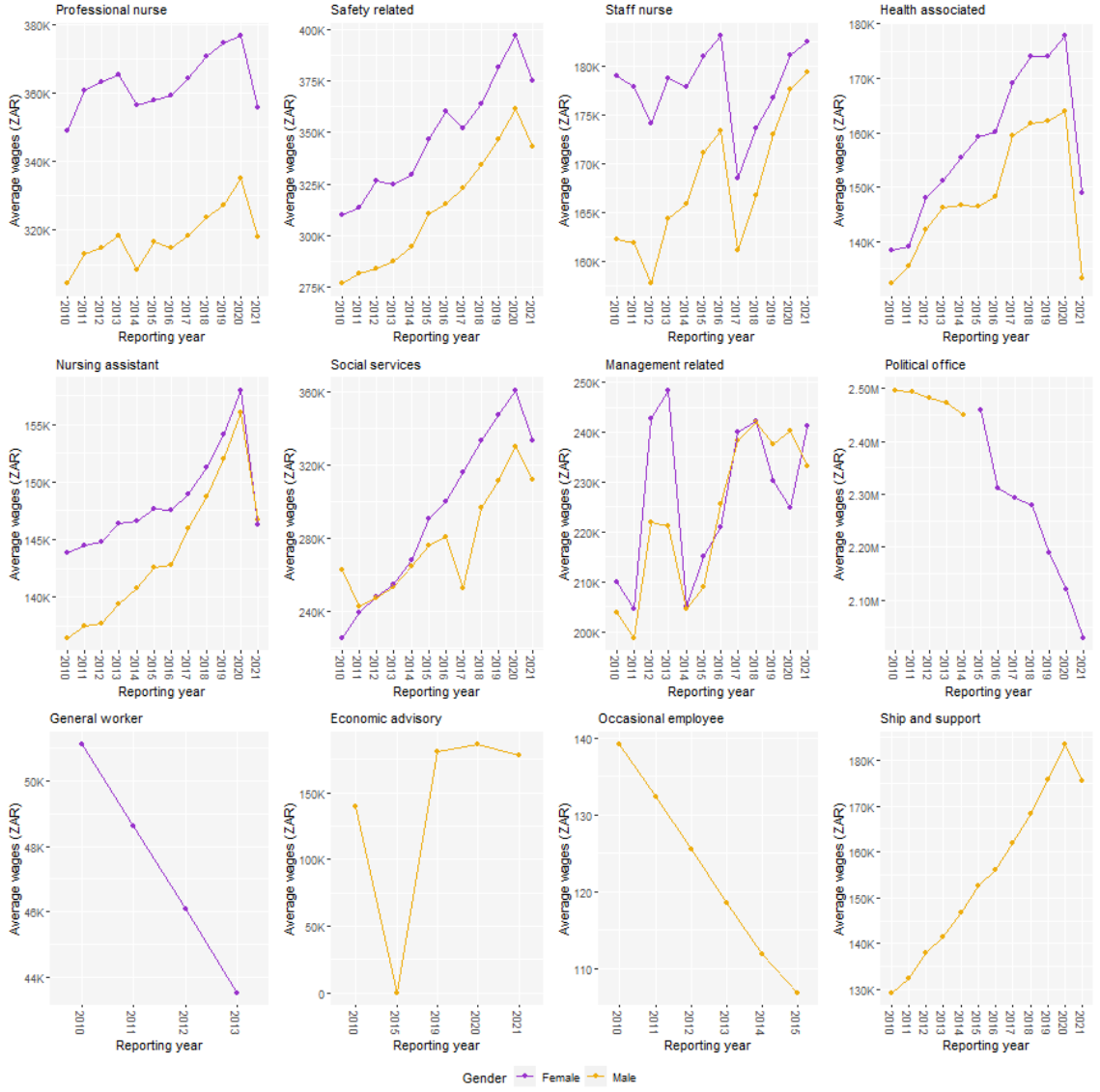

Figure 10: Gender wage gap by occupational groups I.

Further analysis of the occupational groups reveals that gender wage gap against females seem to be more prominent for artisan and support, medical technology, medical sciences and natural sciences as shown in Figure 11. Looking at theses occupational groups, most of them are male dominated. In particular, only 15% of artisan and support, 18% of engineering related and 31% of natural sciences are females. Therefore in comparison to Figure 10, female dominated occupations tend to have gender wage in favour of females while male dominated occupational groups are more likely to have gender wage gap against females.

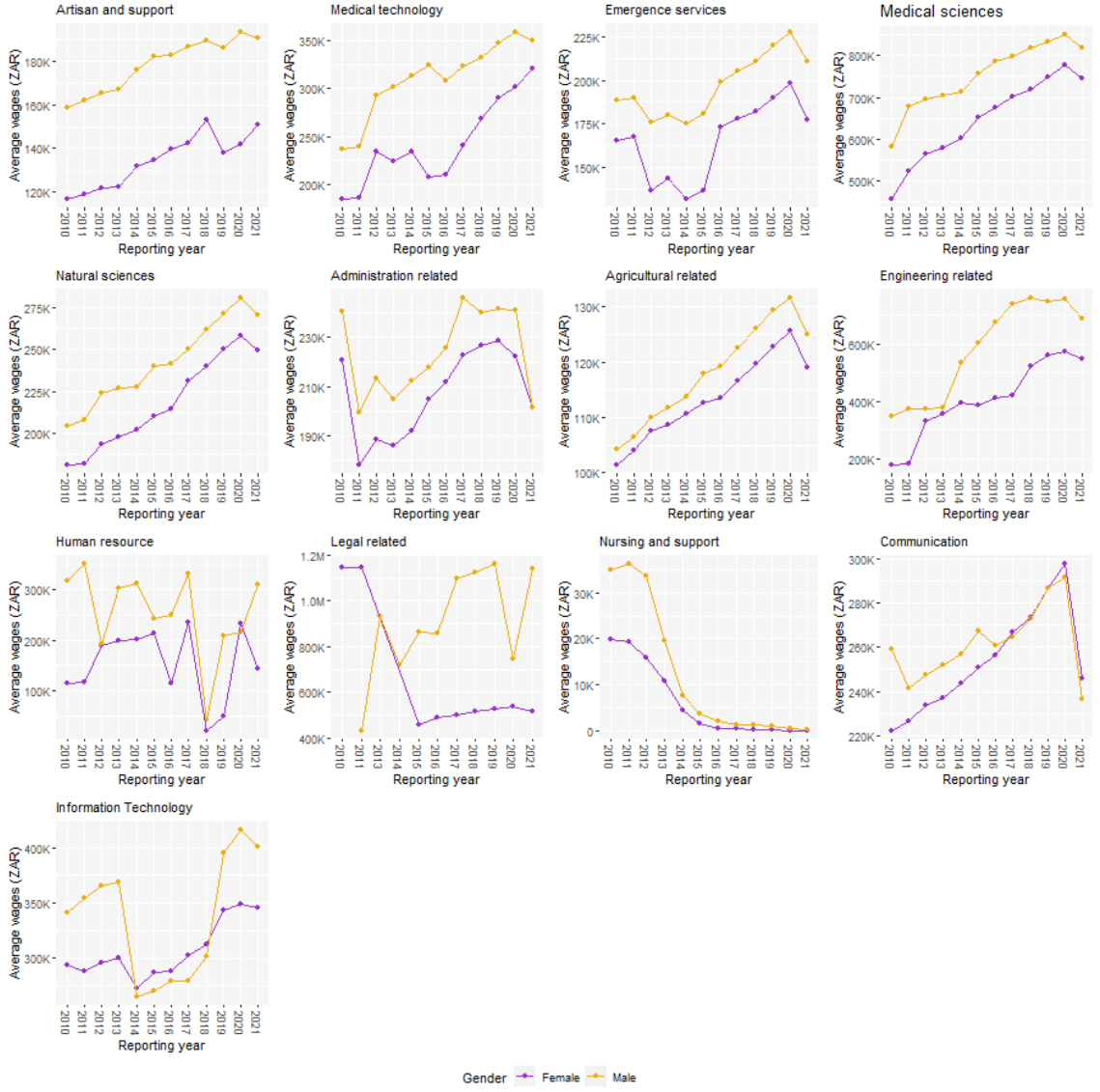

Figure 11: Gender wage gap by occupational groups II.

However, in order to effectively assess the gender wage gap, it is not enough to look at occupational groups since they are broad. It is worthwhile to explore the job title since this is the most specific group compared to other occupation variables though it has the highest number of levels. For instance, under the medical science occupational group there are other occupational classification groups such as medical practitioner, medical specialists and pharmacists. Considering only the medical practitioner occupational classification group there are different job titles such as medical officer grade 1, medical registrar and medical officer (intern). Hence, it is crucial to examine the gender wage gap within these narrower subgroups, as observations in these groups tend to be more closely related and thus more comparable.

Therefore the next subsection will explore the effect of narrowing down the occupation groups into more fine-grained groups.

4.4.1 Medical science group

In the previous section gender wage gap for each occupational group was analysed and there is a chance that this variable might be too broad and it may be difficult to get the true picture of the gender wage gap. The other variables that group employees according to their field of work are unit field, occupational classification, position title and job title. The unit field variable has 42 levels but this variable is also broad and more related to the occupational group variable. The next in line is the occupational classification group which has 114 levels and the other variables are position title and job title which are a bit narrower with the job title being the most specific variable.

In this section, the medical science occupational group which shows a significant gender wage gap against females is analysed further. Table 5 shows that there are different occupational classification groups under the medical science group which includes medical specialists, cleaners and administrators. In such a situation the employees are not comparable as there are very highly skilled workers with high earnings mixed with low wage laborers. Therefore the medical practitioners group will be broken down further.

Table 5: Occupational classification for Medical science.

	Between percentages (frequencies)			Within percentages	
Occupational classification	**Female**	**Male**	**Total**	**Female**	**Male**
Medical Practitioners	54.80% (10049)	66.46% (11893)	60.56% (21942)	45.80%	54.20%
Pharmacists	15.92% (2919)	8.10% (1449)	12.06% (4368)	66.83%	33.17%
Auxiliary And Related Workers	14.13% (2591)	8.48% (1518)	11.34% (4109)	63.06%	36.94%
Medical Specialists	4.49% (823)	9.61% (1720)	7.02% (2543)	32.36%	67.64%
Dental Practitioners	4.67% (857)	4.27% (765)	4.48% (1622)	52.84%	47.16%
Psychologists And Vocational Counsellors	3.43% (629)	0.94% (168)	2.20% (797)	78.92%	21.08%
Health Sciences Related	0.66% (121)	0.65% (117)	0.66% (238)	50.84%	49.16%
Other Occupations	0.82% (150)	0.20% (36)	0.51% (186)	80.65%	19.35%
Pharmaceutical Assistants	0.35% (64)	0.63% (112)	0.49% (176)	36.36%	63.64%
Physicists	0.15% (28)	0.18% (33)	0.17% (61)	45.90%	54.10%
Social Sciences Related	0.17% (31)	0.06% (11)	0.12% (42)	73.81%	26.19%
Senior Managers	0.05% (9)	0.15% (27)	0.10% (36)	25.00%	75.00%
Cleaners In Offices Workshops Hospitals Etc.	0.11% (21)	0.03% (5)	0.07% (26)	80.77%	19.23%
Pharmacologists Pathologists & Related Professiona	0.12% (22)	0.01% (2)	0.07% (24)	91.67%	8.33%
Administrative Related	0.03% (6)	0.03% (5)	0.03% (11)	54.55%	45.45%

4.4.2 Medical practitioner group

The medical practitioner occupational classification group is disaggregated further by the job title variables to get the specific titles. This level of disaggregation represents the narrowest segmentation, where employees within each group tend to exhibit similar characteristics, making them more comparable. Table 6 shows different job titles for medical practitioners and the Medical officer grade 1 group which is the most populated will be explored further.

Table 6: Job title for Medical practitioners.

	Between percentages (frequencies)			Within percentages	
Job title	**Female**	**Male**	**Total**	**Female**	**Male**
medical officer grade 1	33.08% (3326)	29.52% (3517)	31.15% (6843)	48.60%	51.40%
registrar (medical)	4.02% (404)	3.22% (384)	3.59% (788)	51.27%	48.73%
medical officer grade 2	9.07% (912)	8.47% (1009)	8.74% (1921)	47.48%	52.52%
medical officer grade 3	8.65% (870)	13.98% (1665)	11.54% (2535)	34.32%	65.68%
clinical manager (medical) grade 1	1.83% (184)	4.22% (503)	3.13% (687)	26.78%	73.22%
clinical manager (medical) grade 2	0.36% (36)	1.01% (120)	0.71% (156)	23.08%	76.92%
clinical manager (medical) senior grade 1	0.00% (0)	0.04% (5)	0.02% (5)	0.00%	100.00%
head clinical department (medical) grade 1	0.08% (8)	0.22% (26)	0.15% (34)	23.53%	76.47%
head clinical department (medical) grade 2	0.00% (0)	0.08% (9)	0.04% (9)	0.00%	100.00%

Earlier on, medical science group showed significant gender wage gap against females mainly because a broader group with different employees from different skills sets and remuneration was considered. Conversely, the medical officer grade 1 group is more specific, encompassing only those who hold the grade 1 medical officer position. However, this group presents an interesting contrast, as female earnings are surpassing those of males as reflected by Figure 12. Following this, it is important to assess gender wage gap not just from a broad point of view but from employees who share the most similar characteristics.

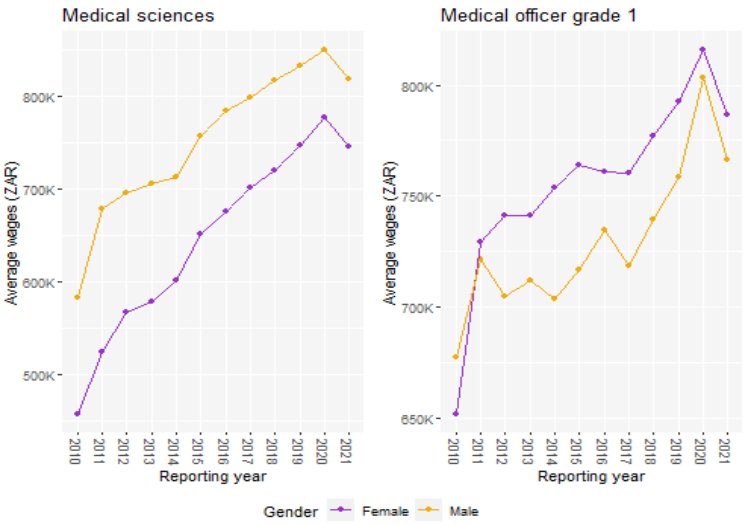

Figure 12: Medical science vs Medical officer grade 1.

4.4.3 Impact of additional independent variables

The medical officer grade 1 group showed gender wage gap against males. While individuals within this group are more likely to have similar characteristics, the scenarios where employees with identical job titles but in different facilities or other factors should be considered.

This dataset has many variables that can be used to disaggregate the job title group further. Figure 13 shows the medical science group disaggregated by district, facility, race and years in the system. It can be observed that other variables also play part in the gender wage gap, this is reflected by varying gender wage gap for each category.

Figure 13: Medical officer grade 1 by other variables.

Overall, it is important to analyze the gender wage gap at the job title level. Other factors that also affect the gender wage gap include years in the system, facility, district, and race. Therefore, the other factors will be incorporated in the linear mixed effects regression model as fixed variables.

4.5 Gender promotions and career trajectories

Previous study results revealed that women tend to earn more than males when they have been in the system for an extended period. Additionally, gender wage gap in favor of females was observed among permanent employees. Considering existing literature, it is evident that females prioritize

convenience and proximity due to their traditional caregiving roles and household responsibilities hence, some women voluntarily position themselves at lower occupational levels with reduced responsibilities (Blau and Kahn, 2017). Studies also indicate that females are less likely to find employment compared to their male counterparts hence the motive to stay longer in the system (Pleace et al., 2023).

These observations collectively highlight that females are more inclined to remain within the system for longer duration and are less eager than men to seek promotions. This section will explore whether females and males experience similar promotion timelines. This assessment is crucial because it impacts the gender wage gap: favoring females for lower positions within an occupation but potentially working against them as males advance to higher paying roles.

4.5.1 Gender wage gap at senior levels

In previous paragraphs, it was noted that female employees were earning higher than their male counterparts for medical officer grade 1. The medical officer grade 1 job title serves as the starting point for a career as a medical practitioner, therefore, other senior levels will be explored starting from this baseline. This will be done to determine whether the gender wage gap persists in favor of females as individuals progress to higher positions.

In Figure 14, it can be observed that there is a significant gender wage gap in favor of female employees within the medical officer grade 1 position. However, examining more senior levels, males begin to dominate the earnings while on the other hand, female representation diminishes, with some years even lacking any female presence. Notably, no females are represented in the highest paid level which is head clinical department (medical) grade 2. This aligns with assertions made by Gradín (2021), highlighting that although women now have greater access to highly skilled positions due to educational improvements, males continue to hold sway in managerial and senior roles.

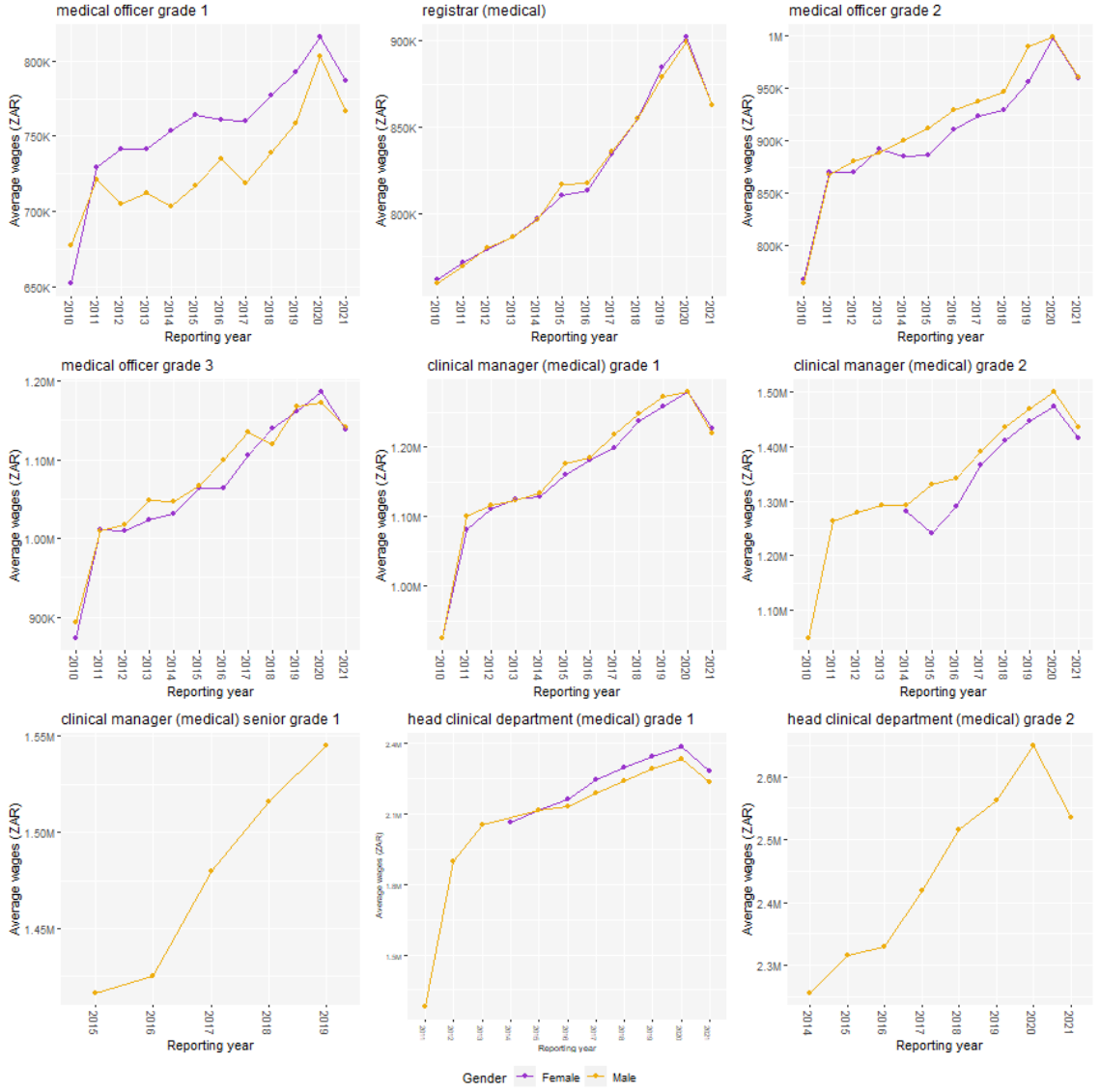

Figure 14: Gender wage gap across senior job titles.

4.5.2 Gender promotions gap

To conduct this analysis, the focus will be on a specific sample from the dataset. Notably, the medical officer grade 1 position since it serves as the baseline in the medical profession and has nearly equal distribution between females and males, with females constituting approximately 49% of the workforce. By delving deeper into this position, the gender promotional gap can be explored.

In this part of the analysis, promotion is when an employee moves from the medical officer grade 1 position to any better job title. It is important to note that for this part of analysis, the unit of measure is the employee and not the observation which is being used throughout the study.

The data from Figure 15 reveals a notable disparity in promotion rates between male and female employees. On average, males advance more swiftly, taking approximately 3.2 years to transition from medical officer grade 1 position to higher titles. In contrast, their female counterparts require an average of 3.9 years for the same progression. Several factors may contribute to this discrepancy. One plausible explanation is that female employees often self-select into lower positions due to family commitments. Furthermore, the additional responsibilities associated with promotions may deter females from pursuing higher roles (Casale et al., 2021; Bhorat and Goga, 2013).

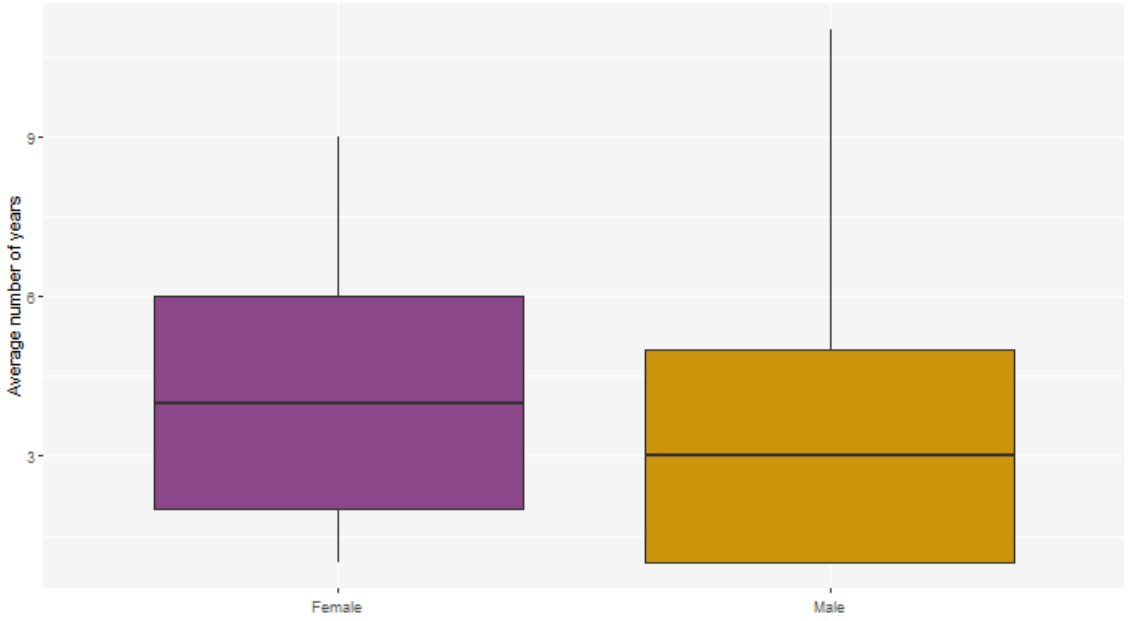

Figure 15: Number of years to get promoted by gender.

4.5.3 Promotions related to changes in facilities

After observing that females experience longer promotion timelines compared to males, it becomes intriguing to explore whether these promotions correlate with facility changes. Existing literature suggests that females often prioritize employment based on proximity and convenience, which may influence their willingness to relocate for promotions (Blau and Kahn, 2017; Steyn and Jackson, 2015; Adelekan and Bussin, 2018).

Out of 576 employees who were promoted from the medical officer grade 1 position, about 27% had these promotions associated with a change in facility. The study reveals that a higher percentage of males (53%) compared to females (47%) grabbed these promotions which involved relocation This discrepancy could be attributed to the fact that females often face family-related constraints, potentially causing them to miss out on promotions that involve relocation. Thus, females in the workforce are constantly faced with a challenge of balancing career aspirations with personal responsibilities.

4.6 Data preparation

In the next chapter, linear mixed effect regression will be employed on the dataset. Therefore in this section, the data is being cleaned and transformed in such a way that it fits in the models and also gives accurate estimates.

The first step is to determine the variables that will be used to fit the linear mixed effect models. Considering observations identified from the data exploration and also the context of the study, the combined wages and allowance (wages-allowance variable) will be the response variable. Since the data set has repeated measures as the data was collected annually for a period of 12 years, the persal number variable which uniquely identifies each employee will be added to the model as a random intercept. Furthermore, job title variable will also be added as another random intercept as gender wage gap is better assessed at the narrowest occupation grouping.

Looking at the explanatory variables, the most important variable is gender since gender wage gap is being analysed. The other independent variables were selected based on a number of reasons which include the following;

- Variables relevant to the context of the study.
- Variables that were appearing to be useful in data exploration above.
- Variables that are likely not to be related with the other independent variables.
- A rule of thumb in regression is to use an independent categorical variable with a maximum of 20 categories. This allows for better handling and interpretation of the results. Thus, variables with not more than 20 levels were considered.

After considering these four points, the following variables will be used in the regression model; gender, years in the system, nature of appointment, district, facility type, rural urban setting, race, citizenship, age and year reported.

4.6.1 Data cleaning

Based on observations from data exploration, it was observed that average earnings in the year 2021 were lower compared to the preceding years due to the absence of data for the last two months of that year. Thus in order to have data which is comparable, observations from the year 2021 were excluded from the analysis.

The nature of appointment variable consists of ten categories, of which some had very low average earnings. The low average earnings could be attributed to the fact that earnings were not consistent and were not reported on an annual basis. This inconsistency was particularly evident in the abnormal and periodic categories. It is important to highlight that earnings were reported annually, and any observations that did not accurately reflect annual earnings introduced distortions in the data. As a result, observations with non-annual earnings were excluded from the analysis.

Through data exploration it was also observed that years in the system variable was not consistent thus this variable was cleaned. Similarly, there were employees who changed gender in the 12 year period and this could potentially be a mistake. Therefore, the first gender was considered as the gender for the employee for the rest of the reporting period. In addition, 58 employees who appeared more than once in a year were excluded since the unit of analysis in this study considers each employee for each reporting period. These observations were about 0.2% and mostly medical practitioners.

4.6.2 Job title

The job title variable plays a crucial role in analyzing gender wage gap. Gender wage gap will be analysed at each job title level by adding gender random slope to the job title intercept to observe how employees at different job title levels differ in their earnings by gender. The initial step in cleaning this variable is to remove all observations with single gender (9538) since it is not possible to have gender comparison.

It was concluded from a simulation study conducted by Maas and Hox (2005) that a sample size of 50 or less for level 2 in multilevel modeling results in biased estimates of the standard errors. Following this, all the job titles with a sample size of less than 50 either for females or males were excluded.

In addition, job titles with a skewed class proportion for gender variable were also excluded. Generally if the minority class constitutes about 1%-20% of the dataset then the degree of imbalance is considered moderate. Therefore all the occupational groups with a proportion of 10% or less for either females or males were excluded from the dataset. As a result, the final dataset has 142 job titles.

4.7 Chapter summary

In this chapter, the dataset was cleaned and prepared for analysis within the framework of linear mixed effect models. Overall, the data exploration revealed that there is gender wage gap against females. However, when smaller occupational subgroups were examined, a more detailed perspective emerged. This highlighted the pivotal role of the job title variable in understanding the gender wage gap. Incorporating these small and several groups into the traditional methods such as ordinary least squares regression and Blinder-Oaxaca decomposition is challenging. Therefore, linear mixed effects models which allows analysis of gender wage gap for each job title will be employed in the next section.

Overall, the data exploration indicates that women tend to gravitate toward lower positions within occupational groups due to factors like proximity and convenience, often influenced by their caregiving responsibilities in families. This trend is supported by the following findings from the data exploration:

1. Females who remained in the system for extended periods and held permanent positions earned higher salaries than their male counterparts.
2. Females took longer to be promoted and were less likely to secure promotion opportunities associated with a change in facility, unlike males.
3. While females earned more in baseline positions within an occupation, the proportion of females decreased as positions became more senior, with males dominating these higher roles.
4. Notably, the gender wage gap was more pronounced against females, particularly among employees in the upper decile.
5. Females tend to earn higher in female-dominated occupations, whereas males earn higher in male-dominated occupations.

5 Data analysis

5.1 Introduction

The Clinton Health Access Initiative, (CHAI) began supporting the South African government health programs in 2003. In this study, CHAI has partnered with the National Department of Health to help the government to deal with gender inequality at the work place. This has been a serious and persistent issue in the country despite the government's efforts to eliminate gender inequality through implementing different anti-discriminatory laws (Musetsho et al., 2021). Furthermore, StatsSA (2022) reported that South African women are earning $23\% - 35\%$ lower than their male counterparts in the same role. This study thus aims to evaluate gender wage gap for employees in the National Health Department in the Eastern Cape province.

To enhance the reliability and validity of the results in this analysis, cleaned data from the Data Exploration chapter was utilized. Initially the data constituted 604 448 observations and 91217 unique employees. About 75% of these observations were females mostly dominating the nursing occupational groups such as professional nurse, staff nurse, nursing assistant and nursing support. However after removing the outliers and irrelevant observations, the data used for this analysis comprises 319,944 observations, of which 71% are females.

It is important to note that the data was divided into six occupational groupings for easier handling and interpretability. Table 7 shows the six different groupings, with nurses comprising the largest portion at 42%. Additionally, half of the female employees held nursing positions (50%). It is interesting to note a significant proportion of females occupying managerial roles (60%) and highly skilled positions in the medical science group (49%).

Table 7: Occupational groupings by gender.

	Between job title percentages (frequencies)			Within job title percentages			
Groupings	Female	Male	Total	Female	Male	Average earnings	Job titles
administration	13.8% (31182)	12.9% (12124)	13.5% (43306)	72.0%	28.0%	230.5	34
general workers	22.8% (51672)	31.0% (29069)	25.2% (80741)	64.0%	36.0%	127.5	31
health related	7.0% (15731)	20.4% (19092)	10.9% (34823)	45.2%	54.8%	264.1	33
managers/directors	2.4% (5493)	3.8% (3596)	2.8% (9089)	60.4%	39.6%	828.4	24
medical science	3.7% (8379)	9.2% (8615)	5.3% (16994)	49.3%	50.7%	894.0	12
nurses	50.3% (113754)	22.7% (21237)	42.2% (134991)	84.3%	15.7%	212.5	8
Total	100.0% (226211)	100.0% (93733)	100.0% (319944)	70.7%	29.3%	252.8	142

Earnings are in 1000 Rands

Since the dataset has repeated measures, there is need to account for individual variability. In addition, there is a possibility that employees in the same occupation group earn higher or lower than other employees in that same group. Therefore, this analysis will employ mixed models,

particularly linear mixed effects regression, which examines the condition of interest while simultaneously accounting for variability within and across employees and occupational groups (Brown, 2021). Furthermore, this study will not only include random and fixed effects in a single model but will also analyse random slopes which allow the slope of the model to vary randomly across the levels of the grouping factor. Brown (2021), also acknowledged that not incorporating random slopes would imply assuming that all participants and groups exhibit identical responses, which is an unjustifiable assumption.

The lmerTest package in R which provides functions to fit and analyze linear mixed models will be employed in this analysis (Kuznetsova et al., 2017). Since the goal of the study is to determine the gender wage gap, the response variable will be the earnings variable. The persal number variable which uniquely identifies each employee will be added in the model as a random intercept to capture dependency among data points due to repeated measures of the employees. Similarly, the job title group will be added to the model as a random intercept to enable the model to estimate each job title's deviation from the fixed intercept. In addition to the random intercepts, gender and years in the system variables are added to the regression to capture the variation in the effect of these variables on the earnings between different job titles.

There are 142 job titles in the cleaned data set which have been divided into six groupings according to the skills sets of the employees. Therefore this analysis will fit six models for the six groupings and an additional model which combines all the models so as to have an overall view of the analysis.

It is important to note that the unit of analysis throughout the analysis is the observation which constitute data for an individual employee over a one year period. In addition, throughout the analysis, earnings have been scaled down by R1000 for the sake of manageability and simplification.

This chapter is structured in such a way that the initial step is to assess the importance of the independent variables selected in Data Exploration chapter. This is followed by model selection which focuses on selecting the random effects that can be added to the model. Thereafter, the Variance Inflation Factor (VIF) test will be performed to diagnose any collinearity or multicollinearity in the final model. Having finalized the best model, seven models will be fitted, and model diagnostic plots will be created to assess the model assumptions. Lastly, the regression output will be discussed for all the models.

5.2 Model specification

5.2.1 Significance of each fixed variable

Before fitting the final model, an analysis was done to compare the relative effect of each predictor on the earnings. This was done by using the top down strategy mentioned by (West et al., 2022). This strategy starts with including the maximum number of fixed effects into the model and one variable dropped at a time with replacement. Table 8 shows different metrics which assess the effect of omitting one fixed variable at a time from the model. It can be observed that the model performs poorly when the appointment variable is removed as reflected by high values for AIC, BIC and Deviance. On the contrary it seems that the rural urban variable is not as important as other predictors as reflected by lower metrics. However, taking into perspective the context of this research and the objectives to be achieved, all the fixed variables under consideration will be included in the analysis.

Table 8: Variable importance.

Ommitted variable	AIC	BIC	Deviance
appointment	2,831,987	2,832,279	2,831,931
race	2,803,394	2,803,708	2,803,334
reporting year	2,802,418	2,802,710	2,802,362
facility	2,800,418	2,800,658	2,800,372
district	2,798,809	2,799,050	2,798,763
gender	2,798,577	2,798,891	2,798,517
age	2,798,374	2,798,688	2,798,314
citizen	2,798,122	2,798,435	2,798,062
years in system	2,797,780	2,798,093	2,797,720
rural urban	2,797,767	2,798,081	2,797,707
All variables	**2,797,767**	**2,798,091**	**2,797,705**

5.2.2 Model selection

Following the decision to include all the fixed variables which were under consideration, the next step is to select the random effects. The goal is to choose the best simple model which considers both statistical and subject matters. For this case, the step-up strategy also mentioned by West et al. (2022) which implies starting with the simplest model was utilised. The random effects that are under consideration are the persal number and job title as the intercepts and also an addition of random slopes. To evaluate the variability across different job titles (rather than individual employees), random slopes are incorporated into the job title intercepts. The three fixed variables to consider for random slopes are gender, years in the system and nature of appointment. These random variables are added one at a time in order of their relevance to the context of the study starting with persal number.

From Table 9, it can be observed that the model including all the random variables under consideration had the lowest values for AIC, BIC and Deviance implying that this model provides the best fit to the data set of this study. However, a closer analysis on the study shows that there is no big difference between this model and the model with gender and years in the system as the random slopes. Thus considering also the complexity and interpretability of the model, Model 5 without nature of appointment random slope will be used for the rest of the analysis.

Table 9: Selection of random variables.

Models	Variables	AIC	BIC	Deviance
Model 1	wages_allowances ~ Gender + Appointment + yrs_system + reporting_year + race + Facility + age + district + Citizen + rural_urban + (1 | persal_number)	2,797,767	2,798,091	2,797,705
Model 2	wages_allowances ~ Gender + Facility + yrs_system + district + race + Appointment + age + Citizen + reporting_year + rural_urban +(1 | persal_number) + (1 | job_title)	2,262,391	2,262,725	2,262,327
Model 3	wages_allowances ~ Gender + Facility + yrs_system + district + race + Appointment + age + Citizen + reporting_year + rural_urban +(1 | persal_number) +(1 + Gender | job_title)	2,261,520	2,261,875	2,261,452
Model 4	wages_allowances ~ Gender + Facility + yrs_system + district + race + Appointment + age + Citizen + reporting_year + rural_urban +(1 | persal_number) +(1 + Gender + Appointment | job_title)	2,124,858	2,125,245	2,124,784
Model 5	wages_allowances ~ Gender + Facility + yrs_system + district + race + Appointment + age + Citizen + reporting_year + rural_urban +(1 | persal_number) +(1 + Gender + yrs_system | job_title)	2,252,423	2,252,903	2,252,331
Model 6	wages_allowances ~ Gender + Facility + yrs_system + district + race + Appointment + age + Citizen + reporting_year + rural_urban +(1 | persal_number) +(1 + Gender + yrs_system + Appointment | job_title)	2,122,261	2,122,804	2,122,157

5.2.3 Multicollinearity in the final model

Following the model selection process, the subsequent step involves assessing whether there is any collinearity among the explanatory variables. As detailed in the Methods section, Variance Inflation Factor (VIF) test is the preferred option for mixed effects because it accounts for the combined effect of multiple predictors and provides a more comprehensive assessment unlike other methods such as Pearson correlation which only captures pairwise relationships. The general rule is that if VIF is greater than 5 then it is an indication of the presence of multicollinearity which can be problematic. The VIF was computed for the predictors in the final model, and the results, as depicted in Table 10, indicate that all variables exhibit a VIF value of less than 5. Facility variable had the highest VIF equal to 1.565. Therefore, the VIF results suggest that there is no significant multicollinearity among the different variables. This finding supports the use of a multifactor correlation regression model.

Table 10: Variance Inflation Factor (VIF).

Fixed variables	VIF	Standard error	VIF 95% CI
facility	1.565	1.251	[1.556;1.573]
district	1.539	1.241	[1.531;1.548]
rural urban	1.312	1.145	[1.306;1.318]
race	1.075	1.037	[1.071;1.08]
reporting year	1.058	1.029	[1.054;1.062]
age	1.041	1.020	[1.037;1.046]
years in system	1.031	1.015	[1.027;1.036]
gender	1.029	1.014	[1.025;1.033]
appointment	1.024	1.012	[1.021;1.029]
citizenship	1.021	1.011	[1.017;1.026]

5.3 Model analysis

This section utilises linear mixed effect regression models which is implemented using the lmer function in R (Kuznetsova et al., 2017). The mixed effects models are useful tool for analyzing data with both fixed and random effects. In standard linear regression, the assumption is that observations are independent. However, in this dataset with repeated measures, the independence assumption is violated. Nevertheless, linear mixed models extend linear regression by providing a flexible framework for handling correlated data and capturing both fixed and random effects. As mentioned earlier, the best model selected has earnings as the response variable. This variable sums up the annual notch value and annual total allowance for each observation. It is important to note that these earnings have been CPI adjusted as detailed in Data exploration chapter. The independent variables are gender, reporting year, district, facility type, age, rural urban setting, nature of appointment, race and citizenship. On the other hand, the random intercepts are persal number and job titles variables. Furthermore, the random slopes for gender and years in the system were added to the job title intercept.

5.3.1 Key highlights of the results

This subsection provides an overview of the results obtained from the data analysis described in the subsequent subsections. Overall, gender wage gap against females is present in the analysed dataset as evidenced by a positive statistically significant fixed gender coefficient of 3.88 for the combined models. Considering all the six groupings, health related grouping had a positive statistically significant fixed gender coefficient (5.38). This suggests that gender wage gap against females is

more pronounced for employees in the health related occupation. Furthermore, the gender wage gap against females is also present in the medical science (1.64) and nursing (2.08) groupings, although the coefficients are not statistically significant. A closer analysis of individual job titles shows that the gender wage gap against females is more prominent for Deputy Director: Administration L11, with a deviation from the overall mean of 57.73, followed by Assistant Director: Logistic Support L9 (52.02) and Deputy Director: Hospitals L11 (38.24).

Despite the overall gender wage gap disadvantaging females, it is noteworthy that there were instances where females actually had a wage advantage within certain groupings. These groupings were managers/ directors with fixed gender coefficient of -2.72, general workers (-0.44) and -0.42 for administration groupings. Furthermore, there was a significant representation of females in highly skilled occupations and managerial positions. For instance, 49% of medical science professionals were females as well as 60% of managers and directors. Examining the job titles, the gender wage gap in favor of women was more distinct for Deputy Director: Finance L11, with a gender random slope estimate of -46.35, followed by Head Clinical Unit (Medical) Grade 1 (-25.98) and Deputy Director-Senior: Administration L12 (-23.93).

While the analysis indicates that females are taking up higher positions, it is also evident that within the same occupational grouping, males tend to dominate the highest positions, while females often hold the lowest positions. In particular, males in the medical science grouping were dominating the most paid positions such as medical specialists. This was evidenced by a positive gender random slope estimate of 1.84 for medical specialist grade 1 and 3.27 for medical specialist grade 2. However, within the same medical science grouping, females were earning more than males in lower positions, such as medical officer intern (-1.16), medical officer community service (-1.31), and dentist (community service) (-1.84).

In addition, considering the nurses grouping, though 84% of the population were females, males were earning higher than females. This was indicated by a positive fixed gender coefficient of 2.37. Furthermore, males were earning higher for top positions such as pnd1 lecturer nursing grade 1 and pna2 professional nurse grade 1 (general nursing) while females were dominating lower positions such as na1 nursing assistant grade 1 and na2 nursing assistant grade 2.

It is important to note that the proportion of unexplained variation was very small about 3.16%. This indicates that most of the variability has been accounted for by the factors in the model. In addition, the job title intercept had a standard deviation of 298.16 for the combined models. This implies that the job title was explaining a significant portion of total variance which is about 83%. Thus, disaggregating the gender wage gap by job title plays a significant role in explaining the gender wage gap.

Examining the combined model estimates, it was observed that gender wage gap manifests more prominently in high-earning positions, both in favor of and against females. This implies that the disparity is more noticeable among employees with substantial incomes, as opposed to those in lower earning brackets. Notably, the disproportionate impact of high earnings may contribute to amplifying the wage gap, leading to wider margins between genders.

The subsequent subsections provide an overview of the fitted models, starting with the administration grouping followed by general workers, health related, medical science, managers/directors, nurses and then the combined groupings will be modeled separately.

Table 11: Descriptive statistics-administration.

	Between job title percentages (frequencies)			Within percentages		Average earnings		
Job title	**Female**	**Male**	**Total**	**Female**	**Male**	**Female**	**Male**	**Overall**
human resource practitioner senior l8	1.0% (318)	1.0% (119)	1.0% (437)	72.8%	27.2%	354.9	362.3	356.9
financial practitioner senior l8	1.1% (347)	1.4% (170)	1.2% (517)	67.1%	32.9%	350.3	355.0	351.8
administrative officer senior l8	8.5% (2663)	10.2% (1233)	9.0% (3896)	68.4%	31.6%	351.8	346.9	350.2
logistic support officer senior l8	0.5% (141)	0.7% (84)	0.5% (225)	62.7%	37.3%	342.6	345.0	343.5
information officer senior l8	0.4% (114)	0.5% (64)	0.4% (178)	64.0%	36.0%	347.8	332.7	342.4
client information clerk chief l7	0.6% (185)	0.6% (75)	0.6% (260)	71.2%	28.8%	299.0	307.7	301.5
information officer l7	0.5% (169)	0.9% (106)	0.6% (275)	61.5%	38.5%	297.8	291.7	295.4
administrative officer l7	5.4% (1678)	5.8% (698)	5.5% (2376)	70.6%	29.4%	293.6	293.0	293.4
financial practitioner l7	4.3% (1349)	2.9% (357)	3.9% (1706)	79.1%	20.9%	290.1	287.7	289.6
administration clerk chief l7	0.6% (200)	1.0% (125)	0.8% (325)	61.5%	38.5%	287.3	291.9	289.0
logistic support officer l7	1.5% (455)	2.3% (279)	1.7% (734)	62.0%	38.0%	285.9	294.2	289.0
human resource practitioner l7	3.2% (991)	3.1% (373)	3.1% (1364)	72.7%	27.3%	288.6	286.2	287.9
administration clerk l6	6.0% (1878)	5.5% (672)	5.9% (2550)	73.6%	26.4%	242.4	241.3	242.1
client information clerk principal l6	1.0% (322)	1.6% (197)	1.2% (519)	62.0%	38.0%	238.6	244.6	240.9
registry clerk principal l6	0.5% (146)	0.4% (52)	0.5% (198)	73.7%	26.3%	242.2	227.0	238.2
community liaison officer l6	1.5% (454)	0.7% (90)	1.3% (544)	83.5%	16.5%	236.7	243.6	237.9
human resource clerk l6	3.2% (1000)	2.8% (345)	3.1% (1345)	74.3%	25.7%	238.5	233.3	237.2
financial clerk l6	2.5% (774)	1.7% (212)	2.3% (986)	78.5%	21.5%	238.2	232.6	237.0
senior provisioning admin clerk l6	2.0% (635)	3.5% (423)	2.4% (1058)	60.0%	40.0%	231.7	237.6	234.1
provisioning admin clerk l5	0.6% (197)	1.2% (140)	0.8% (337)	58.5%	41.5%	204.0	205.5	204.6
opd clerk level5	0.3% (80)	0.5% (62)	0.3% (142)	56.3%	43.7%	204.3	203.1	203.7
registry clerk senior l5	1.1% (332)	0.8% (98)	1.0% (430)	77.2%	22.8%	203.1	199.5	202.3
administration clerk senior l5	14.3% (4464)	11.7% (1418)	13.6% (5882)	75.9%	24.1%	198.3	197.5	198.1
financial clerk l5	2.5% (781)	2.7% (324)	2.6% (1105)	70.7%	29.3%	198.3	196.4	197.8
client information clerk senior l5	1.5% (469)	1.7% (204)	1.6% (673)	69.7%	30.3%	198.5	192.9	196.8
human resource clerk senior l5	1.6% (512)	1.5% (185)	1.6% (697)	73.5%	26.5%	195.7	195.7	195.7
data capturer senior l5	11.1% (3464)	11.1% (1349)	11.1% (4813)	72.0%	28.0%	195.3	195.2	195.3
personnel officer l5	1.5% (469)	1.1% (129)	1.4% (598)	78.4%	21.6%	193.1	193.1	193.1
logistic support clerk senior l5	0.3% (97)	0.7% (79)	0.4% (176)	55.1%	44.9%	181.0	190.0	185.0
client information clerk l4	2.8% (862)	3.1% (378)	2.9% (1240)	69.5%	30.5%	165.3	164.2	164.9
administration clerk l4	14.0% (4363)	12.9% (1564)	13.7% (5927)	73.6%	26.4%	160.7	162.0	161.0
logistic support clerk l4	0.3% (103)	0.4% (52)	0.4% (155)	66.5%	33.5%	164.9	153.1	160.9
data capturer l4	3.1% (965)	3.0% (362)	3.1% (1327)	72.7%	27.3%	160.8	159.0	160.3
registry clerk l4	0.7% (205)	0.9% (106)	0.7% (311)	65.9%	34.1%	162.3	155.7	160.0
Total	**100.0% (31182)**	**100.0% (12124)**	**100.0% (43306)**	**72.0%**	**28.0%**	**229.6**	**232.9**	**230.5**

Earnings are in 1000 Rands

5.3.2 Model 1 administration grouping

There are 34 different types of job titles in this grouping. The average earnings for this grouping is R230 500, with a population constituting 14% of the overall population. It is interesting to note that the highest paid group is human resource practitioner senior l8 with an average of R356 900 while the least paid is registry clerk l4 with average earnings of R160 000 as shown in Table 11. In addition, there is a significantly higher proportion of females compared to males accounting for about 72% of the population.

Random effects-administration

The averages in earnings between females and males observed in Table 11 represent the overall trends of mean effect of gender on the earnings. Comparing the average values of males and females, essentially provides insight into their central tendencies. However, this approach alone may not capture the full complexity of the situation. To gain a more comprehensive understanding, other critical factors, such as variability and context should be considered. This is when random effects come into picture.

Random effects capture individual variability and provide insights into individual deviations from the population average. Thus, random effects capture individual differences that average values might overlook. Table 12 shows random intercept estimates for job titles and random slope estimates for the gender variable. While the job title intercept estimates tells how earnings for each job title differ from the overall population trend. The gender slope estimates indicate how earnings for males differ from earnings for females for each job title.

It is important to note that throughout this analysis the reference group for gender variable is female. In this analysis, for instance, the positive estimate for the gender random slope indicates that, after accounting for the overall effect of job title, captured by the random intercept, there is an additional positive effect associated with being male within that specific job title. In addition, evaluating the statistical significance of these estimates is crucial as it helps ascertain whether the estimates exhibit statistical significance or not. As a result, the standard errors and confidence intervals are also presented in Table 12.

It can be observed that the gender slope estimate for administration clerk chief l7 is notably higher than the rest of the job titles with a gender random slope of 17.24. This implies that in this group, there exists a noticeable gender wage gap, with males earning higher than females. The uncertainty associated with this estimate is 3.04. In addition, there is 95% confidence that the true value is at least 11.29 and no more than 23.19. The non-overlapping confidence intervals imply that the effect is statistically significant.

Table 12: Random effects-administration.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
administration clerk chief l7	21.15	2.99	[15.29;27.01]	**17.24**	3.04	[11.29;23.19]
human resource clerk senior l5	-35.99	1.91	[-39.73;-32.26]	**11.42**	2.32	[6.88;15.97]
client information clerk chief l7	16.69	3.77	[9.31;24.08]	**9.13**	3.34	[2.58;15.68]
human resource clerk l6	-3.50	1.45	[-6.35;-0.65]	**5.72**	1.45	[2.88;8.56]
logistic support officer l7	25.64	2.03	[21.67;29.62]	**5.25**	1.85	[1.61;8.88]
senior provisioning admin clerk l6	-7.01	1.67	[-10.29;-3.74]	**5.20**	1.96	[1.36;9.04]
registry clerk senior l5	-34.17	2.30	[-38.68;-29.66]	**3.33**	2.51	[-1.59;8.25]
administrative officer l7	38.56	1.30	[36.02;41.11]	**3.27**	1.10	[1.11;5.44]
logistic support clerk senior l5	-37.69	2.60	[-42.79;-32.6]	**3.25**	2.75	[-2.13;8.63]
registry clerk l4	-45.74	1.88	[-49.42;-42.07]	**2.32**	2.38	[-2.34;6.98]
logistic support officer senior l8	67.03	3.54	[60.09;73.96]	**2.00**	3.14	[-4.14;8.15]
administration clerk senior l5	-34.89	0.75	[-36.36;-33.43]	**1.97**	0.90	[0.2;3.74]
personnel officer l5	-31.21	1.79	[-34.72;-27.7]	**1.90**	2.44	[-2.89;6.68]
administration clerk l4	-51.54	0.56	[-52.64;-50.43]	**1.31**	0.87	[-0.4;3.01]
opd clerk level5	-39.60	3.96	[-47.36;-31.84]	**0.85**	4.60	[-8.17;9.88]
client information clerk principal l6	-9.23	2.88	[-14.88;-3.58]	**0.71**	2.96	[-5.09;6.51]
client information clerk senior l5	-29.14	2.09	[-33.24;-25.04]	**0.34**	2.96	[-5.46;6.15]
data capturer senior l5	-37.66	0.76	[-39.14;-36.17]	**0.19**	1.06	[-1.89;2.27]
data capturer l4	-62.49	0.87	[-64.2;-60.79]	**0.10**	1.33	[-2.5;2.69]
administration clerk l6	-6.35	1.20	[-8.71;-3.99]	**0.02**	1.18	[-2.29;2.32]
community liaison officer l6	-3.01	3.06	[-9;2.98]	**-0.07**	4.57	[-9.03;8.9]
provisioning admin clerk l5	-41.38	2.81	[-46.89;-35.87]	**-0.49**	2.60	[-5.59;4.61]
client information clerk l4	-47.19	1.20	[-49.55;-44.83]	**-1.23**	2.01	[-5.17;2.71]
human resource practitioner l7	32.21	1.49	[29.28;35.13]	**-1.77**	1.45	[-4.62;1.07]
administrative officer senior l8	109.73	0.97	[107.82;111.63]	**-1.98**	0.89	[-3.73;-0.24]
logistic support clerk l4	-43.42	3.07	[-49.44;-37.39]	**-2.87**	3.61	[-9.95;4.21]
financial clerk l5	-26.48	1.38	[-29.18;-23.77]	**-3.19**	2.00	[-7.11;0.72]
information officer l7	41.26	2.93	[35.52;47]	**-3.50**	3.70	[-10.75;3.76]
financial clerk l6	5.37	1.55	[2.33;8.41]	**-7.07**	2.03	[-11.05;-3.08]
financial practitioner l7	29.38	1.31	[26.81;31.94]	**-7.31**	1.69	[-10.62;-4]
financial practitioner senior l8	77.56	2.27	[73.11;82.01]	**-8.33**	2.45	[-13.13;-3.53]
human resource practitioner senior l8	73.04	3.05	[67.07;79.01]	**-9.97**	2.46	[-14.79;-5.15]
registry clerk principal l6	-8.45	3.33	[-14.98;-1.93]	**-12.85**	3.23	[-19.18;-6.53]
information officer senior l8	98.51	3.90	[90.87;106.15]	**-14.90**	4.47	[-23.66;-6.14]

Visualization of random effects

Figure 16 is a graphical representation of the variation among levels of random effects from the overall model estimates. It can be observed that employees in the administrative officer senior l8 tend to have higher earnings compared to other groups. However, the difference in earnings between males and females within this group is small. This implies that females in this job title are likely to earn slightly higher than males. Additionally, the short vertical line extending from the estimate point confirms the certainty of these results, indicating statistical significance.

Figure 16: Departure from overall estimates-administration.

Normality of random effects

While random effects are not typically checked for normality, the normality plots are important to check whether there are potentially problematic data points. In most cases those values that deviate from normality are likely to be outliers and should be individually checked. Figure 17 shows the plots for random effects of the administration grouping. This study is mainly concerned with interpreting the job title intercept and gender slopes. It can be observed from the job title intercept plot that all the groups seem to lie on the confidence band. However for the gender slope there are some deviations at the tails. The positive deviation on the gender slope is the administration clerk chief l7 as noticed in Figure 17. Closer analysis of this job title reveals that there is one female in 2011 with very low earnings and this could have contributed to some extent to a higher deviation from average.

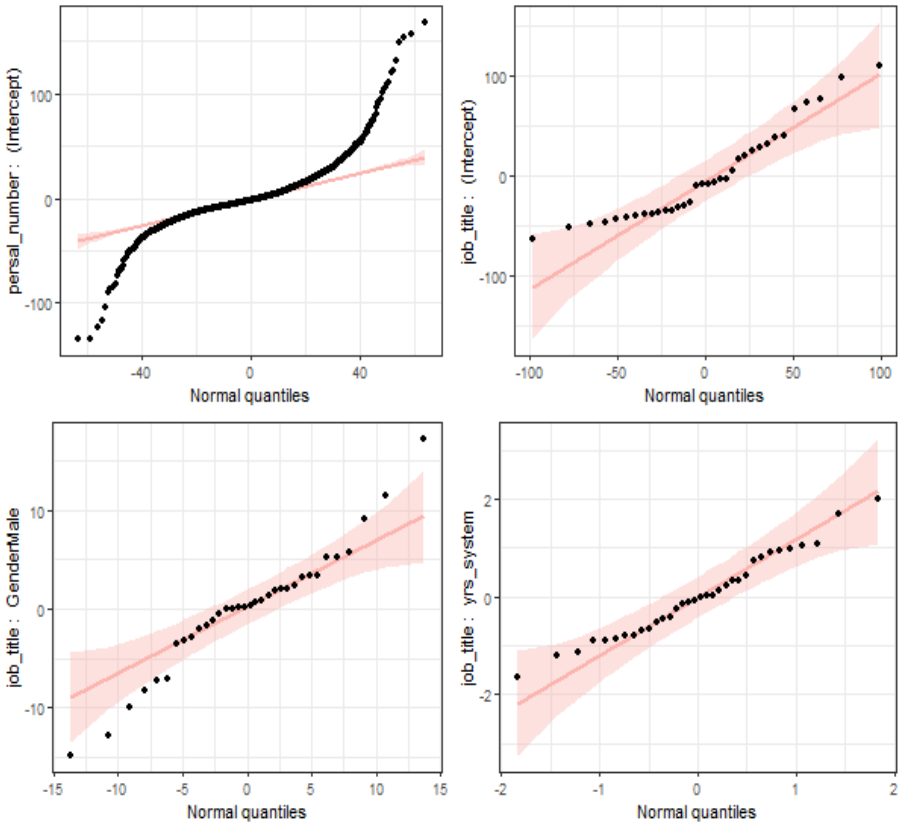

Figure 17: Normality of random effects-administration.

5.3.3 Model 2 general workers

The general worker grouping constitutes the low skilled workforce which includes cleaners, drivers, auxiliary workers and messengers. Among the six groupings, this particular grouping accounts for 25% of all observations and has the lowest average earnings, amounting to R127 500. There are 31 job titles in this grouping with the most populated being general worker grii constituting about 41% of the population and females accounting for the majority of this population (72%) as shown in Table 13.

Table 13: Descriptive statistics-general workers.

	Between job title percentages (frequencies)			Within percentages		Average earnings		
Job title	**Female**	**Male**	**Total**	**Female**	**Male**	**Female**	**Male**	**Overall**
auxiliary worker principal l6	1.5% (785)	0.7% (206)	1.2% (991)	79.2%	20.8%	239.7	237.2	239.2
auxiliary worker(pharmacy) senior l5	0.3% (162)	0.3% (80)	0.3% (242)	66.9%	33.1%	184.2	196.2	188.2
porter chief l4	0.1% (51)	0.5% (152)	0.3% (203)	25.1%	74.9%	157.9	166.2	164.1
food service supervisor l4	1.9% (965)	0.4% (123)	1.3% (1088)	88.7%	11.3%	163.0	160.1	162.6
laundry supervisor l4	1.2% (619)	0.5% (157)	1.0% (776)	79.8%	20.2%	161.4	165.1	162.1
property care taker griv	0.3% (151)	1.3% (371)	0.6% (522)	28.9%	71.1%	149.5	165.1	160.6
auxiliary worker l4	1.9% (956)	1.4% (395)	1.7% (1351)	70.8%	29.2%	159.6	161.8	160.3
auxiliary worker (lay couns) l3	0.2% (102)	0.2% (52)	0.2% (154)	66.2%	33.8%	161.1	153.2	158.4
general worker griv	0.6% (308)	0.4% (130)	0.5% (438)	70.3%	29.7%	155.6	158.6	156.5
auxiliary worker (pharmacy) l4	1.5% (762)	1.5% (439)	1.5% (1201)	63.4%	36.6%	155.8	156.9	156.2
operator l3	0.1% (59)	0.4% (102)	0.2% (161)	36.6%	63.4%	146.7	150.1	148.9
cleaner supervisor l3	0.8% (435)	0.3% (81)	0.6% (516)	84.3%	15.7%	144.1	139.8	143.5
food service aid griii	1.6% (832)	1.0% (284)	1.4% (1116)	74.6%	25.4%	140.9	140.6	140.8
laundry worker l3	1.4% (742)	1.1% (318)	1.3% (1060)	70.0%	30.0%	141.4	138.2	140.4
general worker griii	5.3% (2742)	2.9% (857)	4.5% (3599)	76.2%	23.8%	139.2	137.8	138.9
household worker l3	0.9% (472)	0.2% (61)	0.7% (533)	88.6%	11.4%	139.6	129.5	138.4
porter senior l3	0.4% (207)	2.5% (714)	1.1% (921)	22.5%	77.5%	136.5	138.0	137.6
messenger principal l3	0.1% (55)	0.4% (125)	0.2% (180)	30.6%	69.4%	135.9	135.8	135.8
property care taker griii	0.4% (212)	1.6% (467)	0.8% (679)	31.2%	68.8%	127.7	137.5	134.4
operator l2	0.2% (119)	2.8% (814)	1.2% (933)	12.8%	87.2%	117.4	126.9	125.7
household worker l2	1.5% (784)	0.5% (153)	1.2% (937)	83.7%	16.3%	124.2	121.5	123.8
property care taker gri	0.3% (159)	0.3% (91)	0.3% (250)	63.6%	36.4%	125.2	120.8	123.6
food service aid grii	6.6% (3430)	3.9% (1136)	5.7% (4566)	75.1%	24.9%	122.0	121.0	121.8
general worker gri	1.7% (883)	1.2% (360)	1.5% (1243)	71.0%	29.0%	122.7	118.9	121.6
messenger l2	0.4% (219)	0.6% (178)	0.5% (397)	55.2%	44.8%	122.7	120.2	121.6
cleaner l2	3.9% (2025)	2.1% (612)	3.3% (2637)	76.8%	23.2%	121.8	120.3	121.4
general worker grii	46.2% (23855)	31.4% (9142)	40.9% (32997)	72.3%	27.7%	121.3	121.3	121.3
laundry worker l2	5.2% (2665)	4.6% (1344)	5.0% (4009)	66.5%	33.5%	121.5	120.5	121.2
trade labourer l2	0.4% (213)	4.4% (1292)	1.9% (1505)	14.2%	85.8%	119.7	121.2	121.0
porter l2	4.1% (2117)	11.0% (3192)	6.6% (5309)	39.9%	60.1%	120.1	120.8	120.5
property care taker grii	8.9% (4586)	19.4% (5641)	12.7% (10227)	44.8%	55.2%	119.0	120.6	119.9
Total	**100.0% (51672)**	**100.0% (29069)**	**100.0% (80741)**	**64.0%**	**36.0%**	**128.1**	**126.4**	**127.5**

Earnings are in 1000 Rands

Random effects - general workers

It can be observed from Table 14 that more job titles appear to have positive estimates for the gender slope indicating that more groups are likely to have males earning more than their female counterparts. Conversely the property care taker gri group appears to deviate from the average, exhibiting larger negative margins. This suggests that females within this grouping likely earn more than males, with a greater wage gap.

Table 14: Random effects-general workers.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
auxiliary worker principal l6	48.15	1.05	[46.1;50.21]	**9.49**	1.83	[5.9;13.07]
auxiliary worker l4	10.45	0.79	[8.91;11.99]	**7.61**	1.30	[5.05;10.16]
auxiliary worker (pharmacy) l4	13.91	0.91	[12.13;15.69]	**6.97**	1.42	[4.19;9.74]
laundry supervisor l4	8.99	0.79	[7.44;10.54]	**6.17**	0.96	[4.29;8.06]
household worker l3	-9.00	0.99	[-10.95;-7.05]	**5.22**	1.78	[1.73;8.72]
messenger principal l3	-13.81	2.32	[-18.37;-9.25]	**4.81**	2.38	[0.13;9.48]
messenger l2	-13.10	1.11	[-15.27;-10.94]	**4.67**	1.73	[1.29;8.06]
property care taker griv	-2.09	1.87	[-5.76;1.59]	**3.08**	2.00	[-0.84;7]
property care taker griii	-11.54	0.99	[-13.47;-9.61]	**2.26**	0.82	[0.66;3.86]
porter l2	-11.05	0.50	[-12.03;-10.07]	**1.89**	0.62	[0.68;3.09]
trade labourer l2	-13.28	1.08	[-15.39;-11.16]	**1.36**	1.17	[-0.93;3.66]
porter chief l4	6.31	1.96	[2.47;10.15]	**1.26**	2.16	[-2.98;5.51]
porter senior l3	-7.89	0.86	[-9.57;-6.21]	**0.92**	0.75	[-0.54;2.39]
operator l2	-9.16	2.22	[-13.51;-4.82]	**0.85**	2.32	[-3.68;5.39]
property care taker grii	-9.57	0.33	[-10.22;-8.93]	**0.82**	0.43	[-0.02;1.67]
general worker griii	-9.30	0.41	[-10.11;-8.49]	**0.53**	0.40	[-0.27;1.32]
cleaner l2	-8.76	0.44	[-9.63;-7.89]	**-0.39**	0.95	[-2.26;1.48]
general worker grii	-9.91	0.17	[-10.25;-9.58]	**-0.40**	0.29	[-0.96;0.17]
cleaner supervisor l3	-7.11	0.74	[-8.57;-5.66]	**-0.61**	1.53	[-3.61;2.39]
general worker gri	-9.73	0.55	[-10.8;-8.66]	**-1.20**	1.05	[-3.25;0.86]
general worker griv	9.45	1.17	[7.16;11.75]	**-1.38**	1.64	[-4.6;1.84]
food service aid grii	-9.25	0.36	[-9.95;-8.55]	**-1.69**	0.66	[-2.98;-0.41]
auxiliary worker(pharmacy) senior l5	37.46	1.54	[34.44;40.48]	**-1.98**	2.68	[-7.23;3.26]
laundry worker l2	-8.89	0.38	[-9.64;-8.13]	**-2.33**	0.58	[-3.47;-1.19]
operator l3	-6.89	2.76	[-12.29;-1.48]	**-2.45**	2.56	[-7.46;2.56]
food service supervisor l4	14.54	0.68	[13.2;15.88]	**-2.76**	1.19	[-5.09;-0.44]
food service aid griii	-10.06	0.68	[-11.4;-8.73]	**-2.99**	0.88	[-4.72;-1.25]
laundry worker l3	-3.70	0.73	[-5.13;-2.28]	**-3.20**	0.82	[-4.8;-1.6]
household worker l2	-3.02	0.79	[-4.56;-1.48]	**-6.94**	2.10	[-11.05;-2.83]
auxiliary worker (lay couns) l3	28.74	2.73	[23.39;34.09]	**-12.01**	3.63	[-19.13;-4.89]
property care taker gri	9.11	1.58	[6.02;12.2]	**-17.58**	1.74	[-20.98;-14.17]

Visualization of random effects

Figure 18 shows that the auxiliary worker principal l6, auxiliary worker(pharmacy) senior l5 and auxiliary worker (lay couns) l3 job titles are more likely to earn more than the rest of the groups. On the other hand, as noticed in Table 13, the property care taker gri group and auxiliary worker (lay couns) l3 job titles seem to have females earning significantly higher than their male counterparts. Overall, job titles earning higher than average by greater margins are also more likely to have gender wage gap which is more pronounced. However, there are other job titles which are more likely to have females and males earning the same such as general worker gii and cleaner 12.

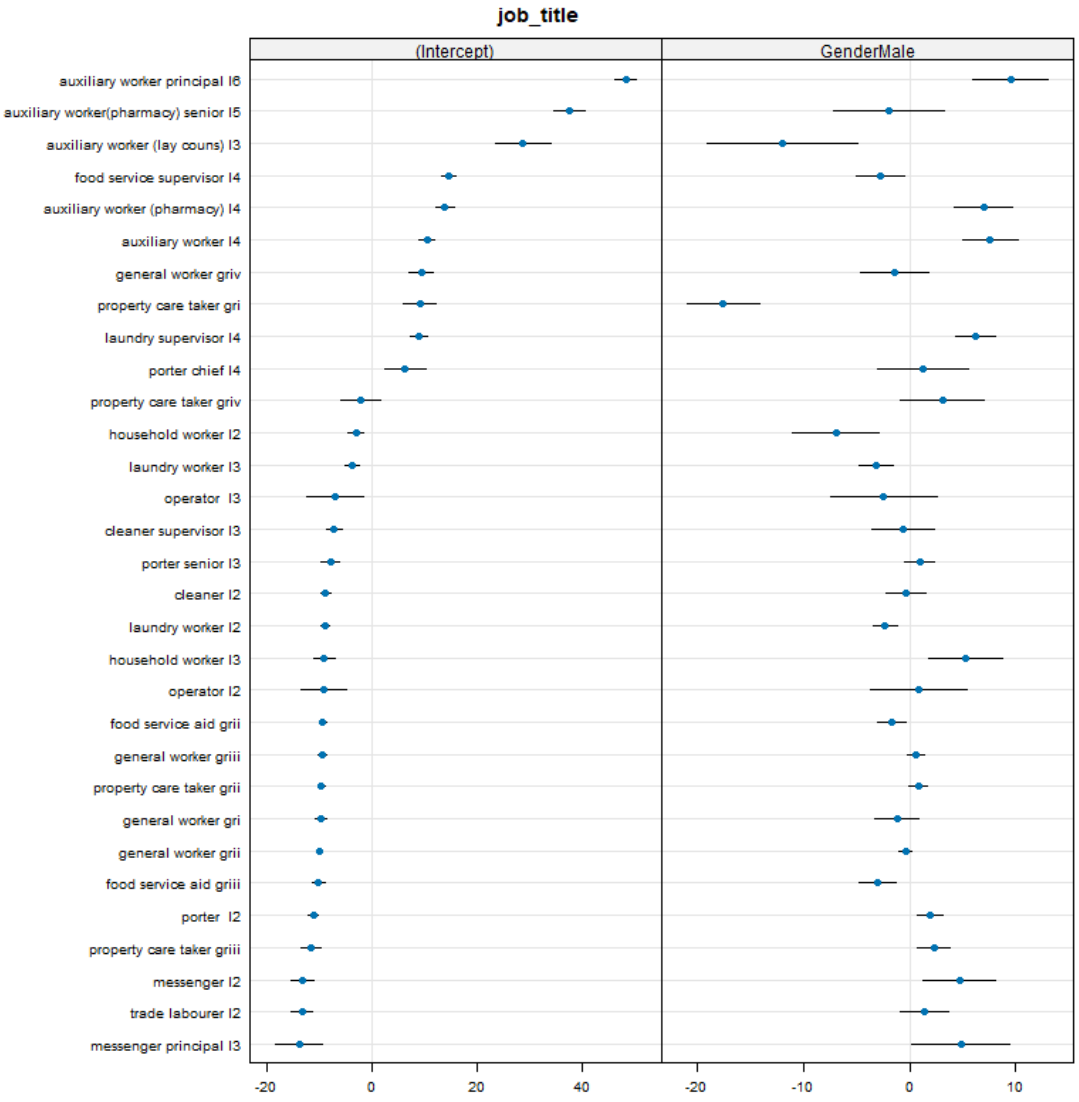

Figure 18: Departure from overall estimates-general workers.

Normality of random effects

As observed from Figure 19, the majority of data points for the job title intercept plot fall within the confidence interval bands, suggesting that the data follows a normal distribution. However, there are a few data points that deviate slightly from these intervals. In the preceding paragraph, these specific job titles (auxiliary worker principal l6, auxiliary worker(pharmacy) senior l5 and auxiliary worker (lay couns) l3) were highlighted for their notably higher earnings compared to other job titles.

In the gender slope plot, most data points align closely with the straight line. However, there are two outliers on the negative side. These outliers correspond to the auxiliary worker (lay couns) l3 and property care taker gri job titles job and it appears that females in these job titles were earning more than their male counterparts.

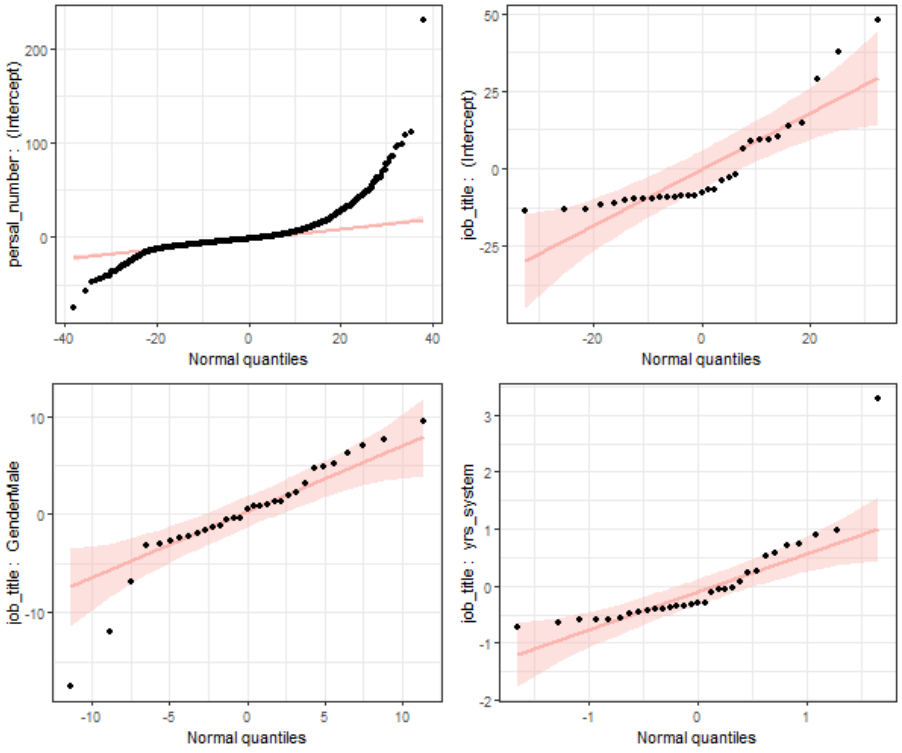

Figure 19: Normality of random effects-general workers.

5.3.4 Model 3 health related

This grouping includes pharmacists, radiographers, forensic officers and mortuary attendants. It is interesting to note an almost equal distribution of females and males exists in this grouping, with a slightly higher proportion for males (55%). The average earnings is R264 100 with females earning higher (R297 500) than their males counterparts (R236 600). The most populated job title is emergency care officer grade 2 constituting 29% of the population with males accounting for the majority (69%) of employees in this group as shown in Table 15.

Table 15: Descriptive statistics-health related.

	Between job title percentages (frequencies)			Within percentages		Average earnings		
Job title	**Female**	**Male**	**Total**	**Female**	**Male**	**Female**	**Male**	**Overall**
pharmacist grade 3	1.5% (239)	0.7% (133)	1.1% (372)	64.2%	35.8%	898.0	880.5	891.7
pharmacist grade 2	1.8% (290)	0.6% (122)	1.2% (412)	70.4%	29.6%	823.1	835.0	826.6
psychologist grade 1	1.8% (285)	0.3% (54)	1.0% (339)	84.1%	15.9%	767.9	756.4	766.1
pharmacist grade 1	5.6% (881)	1.9% (356)	3.6% (1237)	71.2%	28.8%	745.3	744.5	745.1
pharmacist (community service)	2.3% (356)	1.0% (183)	1.5% (539)	66.0%	34.0%	495.4	497.2	496.0
radiographer grade 2	1.5% (231)	0.4% (79)	0.9% (310)	74.5%	25.5%	435.1	435.0	435.1
pharmacist (intern)	1.5% (243)	0.6% (114)	1.0% (357)	68.1%	31.9%	372.1	374.5	372.9
radiographer grade 1	7.8% (1221)	1.8% (351)	4.5% (1572)	77.7%	22.3%	358.2	358.5	358.3
physiotherapist grade 1	4.0% (631)	0.7% (141)	2.2% (772)	81.7%	18.3%	354.1	352.0	353.7
emergency care officer grade 5	1.0% (152)	4.7% (904)	3.0% (1056)	14.4%	85.6%	309.6	317.6	316.5
ems shift leader grade 3	1.7% (264)	2.9% (555)	2.4% (819)	32.2%	67.8%	303.0	306.3	305.3
pharmacist assistant (post-basic) grade 3	1.7% (270)	0.4% (80)	1.0% (350)	77.1%	22.9%	292.3	294.0	292.7
clinical associates	2.4% (374)	1.9% (354)	2.1% (728)	51.4%	48.6%	277.6	278.7	278.1
emergency care officer grade 4	1.1% (176)	4.3% (812)	2.8% (988)	17.8%	82.2%	272.7	274.3	274.0
pharmacist assistant (post-basic) grade 2	1.9% (305)	0.4% (79)	1.1% (384)	79.4%	20.6%	273.1	272.9	273.1
supplementary diagnostic radiographer senior l5	0.5% (84)	2.3% (439)	1.5% (523)	16.1%	83.9%	215.7	275.3	265.7
physiotherapist (community service)	1.9% (293)	0.4% (82)	1.1% (375)	78.1%	21.9%	255.7	255.7	255.7
diagnostic radiographer (community service)	1.5% (239)	0.4% (72)	0.9% (311)	76.8%	23.2%	255.4	256.2	255.6
forensic pathology officer grade 2	0.3% (55)	0.7% (134)	0.5% (189)	29.1%	70.9%	234.5	237.0	236.3
forensic officer grade ii l6	0.4% (58)	0.8% (162)	0.6% (220)	26.4%	73.6%	236.9	233.5	234.4
pharmacist assistant (post-basic) grade 1	6.9% (1086)	1.5% (292)	4.0% (1378)	78.8%	21.2%	232.5	232.6	232.5
emergency care officer grade 3	6.0% (941)	8.9% (1696)	7.6% (2637)	35.7%	64.3%	207.7	209.7	209.0
forensic officer grade i lev 5	0.6% (87)	0.6% (112)	0.6% (199)	43.7%	56.3%	190.8	196.8	194.1
emergency care practitioner (intermediate) l5	1.1% (169)	4.4% (844)	2.9% (1013)	16.7%	83.3%	188.9	194.0	193.2
forensic pathology officer grade 1	0.6% (87)	0.7% (125)	0.6% (212)	41.0%	59.0%	191.1	191.7	191.5
emergency care officer grade 2	19.9% (3128)	36.6% (6986)	29.0% (10114)	30.9%	69.1%	190.5	191.4	191.2
pharmacist assistant (basic) grade 2	3.7% (580)	0.8% (146)	2.1% (726)	79.9%	20.1%	184.7	185.2	184.8
dental assistant grade 1	2.2% (342)	0.5% (86)	1.2% (428)	79.9%	20.1%	183.7	185.8	184.1
emergency care practitioner l4	1.3% (209)	2.7% (510)	2.1% (719)	29.1%	70.9%	173.4	171.5	172.0
emergency care officer grade 1	12.7% (2000)	13.4% (2556)	13.1% (4556)	43.9%	56.1%	161.9	161.7	161.8
mortuary attendant l4	0.5% (80)	1.9% (360)	1.3% (440)	18.2%	81.8%	147.9	157.5	155.8
pharmacist assistant (basic) grade 1	1.4% (221)	0.6% (106)	0.9% (327)	67.6%	32.4%	146.5	145.9	146.3
radiographer student l3	1.0% (154)	0.4% (67)	0.6% (221)	69.7%	30.3%	122.6	122.3	122.5
Total	**100.0% (15731)**	**100.0% (19092)**	**100.0% (34823)**	**45.2%**	**54.8%**	**297.5**	**236.6**	**264.1**

Earnings are in 1000 Rands

Random effects - health related

Looking at Table 16, positive estimates for the gender slope can be observed across various job title groups. This suggests that in this grouping more job titles are likely to have males dominating the earnings. The supplementary diagnostic radiographer senior l5 group stands out with the greatest variability in earnings between females and males. The estimates for this job title are statistically significant with true values likely to lie between 5.03 and 13.72 and have a standard error of 2.22. On the other hand, females are more likely to earn more than males in the pharmacist assistant (basic) grade 2 job title.

Table 16: Random effects-health related.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
supplementary diagnostic radiographer senior l5	-88.15	3.18	[-94.39;-81.92]	**9.37**	2.22	[5.03;13.72]
emergency care officer grade 5	-91.40	2.63	[-96.55;-86.25]	**5.64**	1.90	[1.9;9.37]
pharmacist grade 2	433.89	2.43	[429.12;438.65]	**5.14**	2.07	[1.1;9.19]
emergency care practitioner (intermediate) l5	-116.00	1.69	[-119.32;-112.69]	**4.89**	1.42	[2.1;7.68]
radiographer grade 1	36.77	1.47	[33.88;39.65]	**4.62**	1.95	[0.8;8.44]
forensic pathology officer grade 2	-76.52	4.71	[-85.75;-67.29]	**4.38**	3.02	[-1.55;10.3]
radiographer student l3	-122.15	2.06	[-126.19;-118.1]	**4.25**	2.33	[-0.33;8.82]
diagnostic radiographer (community service)	-38.43	1.66	[-41.68;-35.19]	**3.73**	2.18	[-0.54;8.01]
psychologist grade 1	377.13	3.76	[369.76;384.5]	**3.32**	4.16	[-4.82;11.47]
emergency care officer grade 4	-66.39	2.19	[-70.68;-62.1]	**2.36**	1.70	[-0.98;5.7]
emergency care officer grade 2	-110.40	1.00	[-112.36;-108.45]	**2.15**	1.08	[0.04;4.26]
emergency care practitioner l4	-99.16	1.91	[-102.91;-95.41]	**2.09**	1.43	[-0.71;4.9]
dental assistant grade 1	-127.09	3.51	[-133.97;-120.21]	**1.39**	4.08	[-6.6;9.39]
emergency care officer grade 3	-109.10	1.18	[-111.42;-106.79]	**1.35**	1.08	[-0.78;3.47]
forensic officer grade ii l6	-74.95	4.47	[-83.71;-66.19]	**0.98**	3.62	[-6.12;8.08]
pharmacist (community service)	177.25	1.39	[174.52;179.98]	**0.55**	1.76	[-2.89;3.99]
pharmacist (intern)	73.87	1.50	[70.93;76.81]	**0.44**	1.94	[-3.36;4.24]
emergency care officer grade 1	-131.09	1.07	[-133.18;-129]	**0.22**	1.26	[-2.25;2.69]
forensic officer grade i lev 5	-92.85	3.62	[-99.94;-85.76]	**-0.08**	3.16	[-6.28;6.12]
physiotherapist grade 1	37.33	1.66	[34.08;40.59]	**-0.11**	2.59	[-5.18;4.97]
mortuary attendant l4	-139.97	4.86	[-149.49;-130.44]	**-1.05**	4.12	[-9.13;7.03]
pharmacist assistant (post-basic) grade 3	-59.10	3.44	[-65.85;-52.35]	**-1.29**	3.95	[-9.03;6.45]
ems shift leader grade 3	-28.32	2.25	[-32.72;-23.91]	**-1.92**	1.38	[-4.62;0.77]
pharmacist grade 1	402.93	1.36	[400.25;405.6]	**-2.53**	1.67	[-5.81;0.75]
clinical associates	-21.76	2.30	[-26.26;-17.25]	**-2.95**	2.88	[-8.59;2.7]
physiotherapist (community service)	-34.34	2.06	[-38.38;-30.29]	**-4.59**	2.51	[-9.52;0.33]
pharmacist assistant (post-basic) grade 1	-72.74	1.52	[-75.71;-69.77]	**-4.72**	1.96	[-8.55;-0.88]
pharmacist assistant (post-basic) grade 2	-49.82	2.60	[-54.91;-44.73]	**-5.22**	2.62	[-10.34;-0.09]
forensic pathology officer grade 1	-94.15	4.01	[-102.01;-86.29]	**-5.74**	3.11	[-11.83;0.35]
pharmacist assistant (basic) grade 1	-141.58	1.86	[-145.23;-137.93]	**-5.77**	2.34	[-10.35;-1.18]
radiographer grade 2	67.34	3.08	[61.3;73.38]	**-6.41**	3.13	[-12.54;-0.28]
pharmacist grade 3	481.71	2.60	[476.61;486.81]	**-7.06**	2.98	[-12.9;-1.21]
pharmacist assistant (basic) grade 2	-102.77	1.73	[-106.16;-99.37]	**-7.46**	2.29	[-11.95;-2.97]

Visualization of random effects

Figure 20 visually depicts the variability in random effects for health related grouping across different levels of the overall model. It specifically illustrates the estimated intercepts and slopes for each job title in this grouping. It can be observed that the pharmacist and psychologists groups have a positive deviation for the intercept indicating that these groups are more likely to earn far above other job titles. In addition, it can be observed that gender wage gap is noticeable for both those earning higher and lower than average, indicating that all job title groups are likely to have gender wage gap despite whether their earnings are on the low or high side.

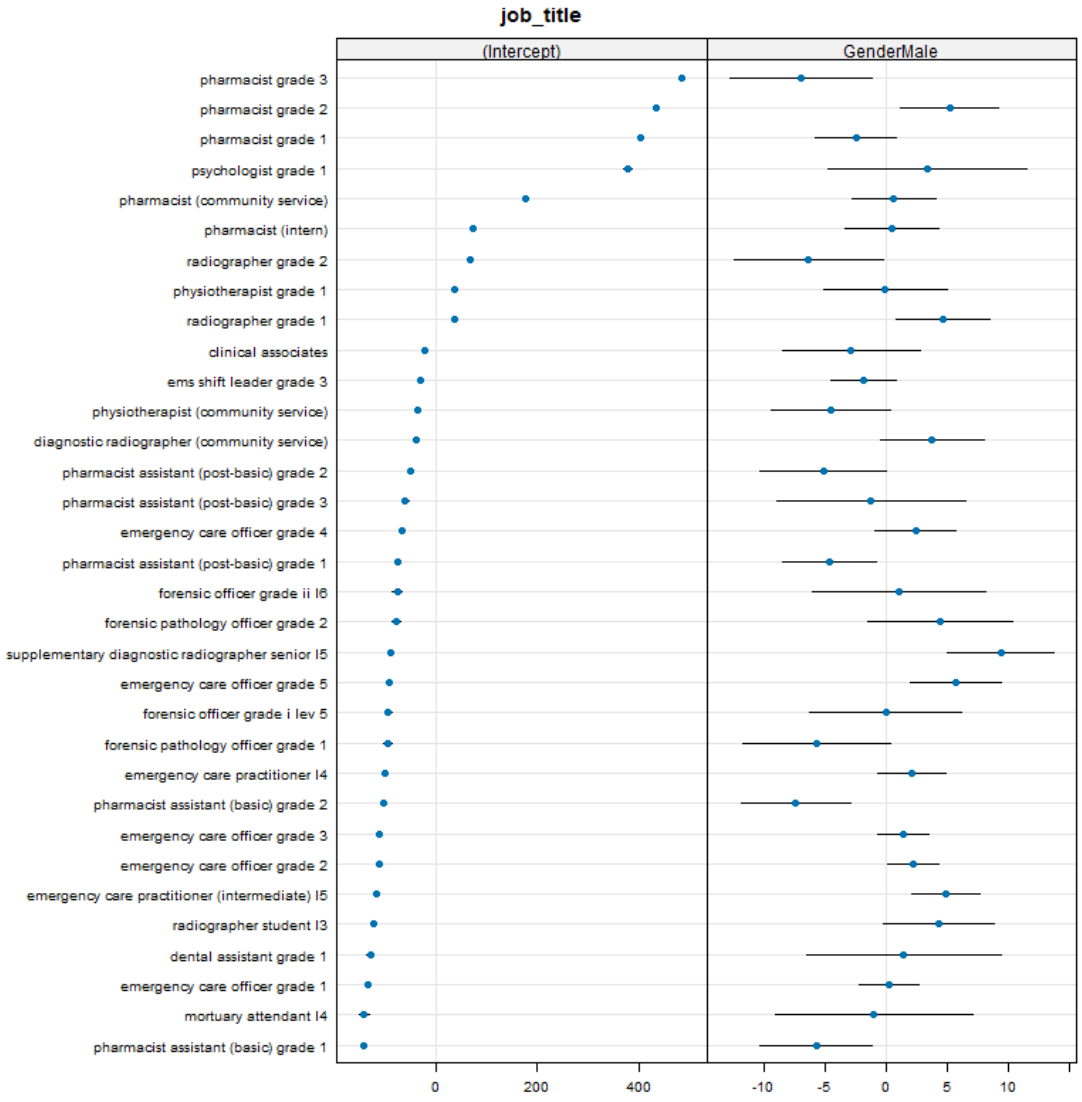

Figure 20: Departure from overall estimates-health related.

Normality of random effects

Looking at the job title intercept plot in Figure 21, there are outliers in the positives tails, indicating that pharmacists are more likely to be earning far above the rest of the groups as was noticed earlier on. On the other hand, the gender slope plot shows that all the groups lie within the confidence band. This suggests that though there are few job titles earning far above average, gender wage variability for the job titles in this grouping is within range and there are no outliers.

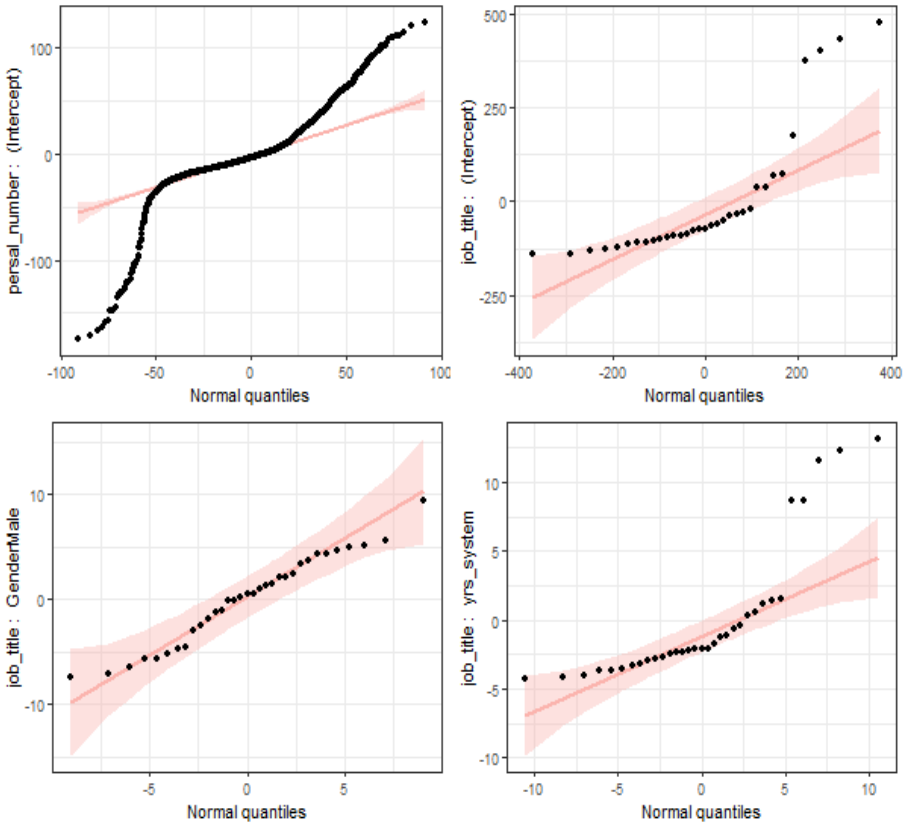

Figure 21: Normality of random effects-health related.

5.3.5 Model 4 managers/directors

This grouping comprise employees in managerial roles and has relatively high earnings averaging to R828,000. The average wages for males within this managerial group are even higher, reaching R982,800, compared to R727,300 for females. Notably, a significant proportion of females (60%) hold managerial positions, as indicated in Table 17. However, upon further analysis, it becomes evident from Table 17 that the top three highest earners in managerial positions are predominantly male. For instance, 77% of head clinical unit (medical) grade 2 are males.

Table 17: Descriptive statistics-managers/directors.

	Between job title percentages (frequencies)			Within percentages		Average earnings		
Job title	**Female**	**Male**	**Total**	**Female**	**Male**	**Female**	**Male**	**Overall**
head clinical unit (medical) grade 2	1.0% (53)	4.9% (175)	2.5% (228)	23.2%	76.8%	2,133.8	2,140.3	2,138.8
head clinical unit (medical) grade 1	3.9% (213)	15.9% (571)	8.6% (784)	27.2%	72.8%	1,917.5	1,878.3	1,889.0
clinical manager (medical) grade 1	3.4% (185)	13.3% (478)	7.3% (663)	27.9%	72.1%	1,297.2	1,300.5	1,299.6
manager pharmaceutical services assistant	6.5% (356)	5.7% (204)	6.2% (560)	63.6%	36.4%	998.8	978.5	991.4
chief executive officer l12	1.7% (93)	1.5% (55)	1.6% (148)	62.8%	37.2%	978.9	968.7	975.1
deputy director-senior:administration l12	1.6% (86)	1.9% (67)	1.7% (153)	56.2%	43.8%	968.6	966.0	967.5
pna8 deputy manager nursing (level 1 & 2 hospital)	16.1% (884)	3.4% (122)	11.1% (1006)	87.9%	12.1%	918.3	924.2	919.0
deputy director:finance l11	3.8% (211)	2.4% (86)	3.3% (297)	71.0%	29.0%	876.8	860.8	872.2
deputy director:administration l11	1.6% (89)	3.3% (117)	2.3% (206)	43.2%	56.8%	832.6	861.0	848.7
deputy director:hospitals l11	2.3% (128)	3.0% (107)	2.6% (235)	54.5%	45.5%	841.2	830.1	836.1
pnb4 assistant manager nursing (primary h care)	6.2% (339)	1.4% (51)	4.3% (390)	86.9%	13.1%	695.6	719.7	698.7
assistant director-senior:administration l10	9.5% (521)	9.6% (346)	9.5% (867)	60.1%	39.9%	541.2	536.9	539.5
assistant director-senior:finance l10	4.3% (238)	2.4% (88)	3.6% (326)	73.0%	27.0%	534.4	526.7	532.3
chief radiographer grade 1	3.8% (210)	2.0% (71)	3.1% (281)	74.7%	25.3%	523.8	523.7	523.8
assistant director-senior:hr management l10	1.5% (80)	2.3% (83)	1.8% (163)	49.1%	50.9%	524.8	516.7	520.7
assistant manager administration l9	0.9% (51)	1.6% (58)	1.2% (109)	46.8%	53.2%	536.5	420.3	474.7
assistant director:administration l9	5.6% (306)	7.5% (268)	6.3% (574)	53.3%	46.7%	460.7	464.2	462.4
assistant director:human resource management l9	2.7% (146)	3.0% (109)	2.8% (255)	57.3%	42.7%	429.0	441.7	434.5
assistant director:health l9	2.6% (145)	2.4% (85)	2.5% (230)	63.0%	37.0%	426.3	446.8	433.9
assistant director:logistic support l9	1.5% (82)	1.7% (60)	1.6% (142)	57.7%	42.3%	426.1	439.3	431.7
assistant director:finance l9	6.1% (335)	5.9% (211)	6.0% (546)	61.4%	38.6%	432.3	428.3	430.8
assistant director:information manager l9	2.2% (123)	1.9% (70)	2.1% (193)	63.7%	36.3%	418.9	444.5	428.2
assistant director:quality assurance l9	3.4% (188)	1.6% (58)	2.7% (246)	76.4%	23.6%	430.2	420.1	427.8
food service manager l7	7.8% (431)	1.6% (56)	5.4% (487)	88.5%	11.5%	286.2	279.6	285.5
Total	**100.0% (5493)**	**100.0% (3596)**	**100.0% (9089)**	**60.4%**	**39.6%**	**727.3**	**982.8**	**828.4**

Earnings are in 1000 Rands

Random effects-managers/directors

Table 18 reveals that the highest earners mentioned in the preceding paragraph that is the head clinical unit(medical) grade 1 and head clinical unit(medical) grade 2 exhibit substantial variability in earnings between females and males. Surprisingly, despite males constituting a larger proportion of employees within these groups, females earn more. This is evidenced by high negative gender slopes of -16.61 for head clinical unit(medical) grade 1 and -11.40 for clinical unit(medical) grade 2. However, these estimates come with wider confidence intervals and slightly higher standard errors, yet they remain statistically significant.

Table 18: Random effects-managers/directors.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
deputy director:administration l11	31.29	12.37	[7.05;55.54]	**7.87**	4.61	[-1.17;16.9]
deputy director:finance l11	-29.57	9.63	[-48.45;-10.69]	**7.78**	4.48	[-0.99;16.56]
assistant director:finance l9	-272.30	6.61	[-285.25;-259.35]	**7.65**	3.95	[-0.09;15.39]
assistant director:quality assurance l9	-291.03	10.64	[-311.87;-270.18]	**4.84**	4.62	[-4.21;13.89]
assistant director:administration l9	-270.78	7.61	[-285.69;-255.87]	**4.56**	4.11	[-3.51;12.62]
assistant director:human resource management l9	-298.75	10.25	[-318.85;-278.65]	**4.09**	4.52	[-4.77;12.95]
assistant director:logistic support l9	-277.13	14.39	[-305.34;-248.92]	**3.46**	4.80	[-5.94;12.87]
deputy director:hospitals l11	25.25	11.95	[1.82;48.68]	**3.09**	4.70	[-6.13;12.3]
assistant director:information manager l9	-241.77	12.68	[-266.62;-216.93]	**2.20**	4.72	[-7.04;11.45]
assistant director:health l9	-253.94	10.72	[-274.95;-232.92]	**1.92**	4.60	[-7.09;10.94]
assistant manager administration l9	-242.77	14.97	[-272.12;-213.42]	**0.92**	4.82	[-8.52;10.36]
assistant director-senior:finance l10	-175.90	9.97	[-195.45;-156.36]	**0.71**	4.25	[-7.62;9.04]
assistant director-senior:administration l10	-195.69	6.83	[-209.08;-182.29]	**-0.56**	3.82	[-8.04;6.92]
pnb4 assistant manager nursing (primary h care)	-61.52	12.11	[-85.26;-37.78]	**-0.98**	4.67	[-10.14;8.18]
manager pharmaceutical services assistant	188.94	8.50	[172.28;205.6]	**-1.38**	4.51	[-10.21;7.46]
clinical manager (medical) grade 1	461.50	8.17	[445.49;477.5]	**-1.40**	4.44	[-10.1;7.29]
chief radiographer grade 1	-170.60	15.26	[-200.51;-140.69]	**-1.58**	4.77	[-10.93;7.77]
assistant director-senior:hr management l10	-163.39	16.04	[-194.83;-131.95]	**-2.46**	4.80	[-11.87;6.95]
food service manager l7	-349.24	10.32	[-369.47;-329.02]	**-2.58**	4.67	[-11.74;6.57]
deputy director-senior:administration l12	114.26	14.84	[85.18;143.34]	**-2.90**	4.67	[-12.06;6.26]
pna8 deputy manager nursing (level 1 & 2 hospital)	138.58	6.94	[124.97;152.19]	**-3.54**	4.40	[-12.16;5.07]
chief executive officer l12	130.31	16.82	[97.36;163.27]	**-3.70**	4.73	[-12.96;5.57]
head clinical unit (medical) grade 2	1,152.85	13.75	[1125.9;1179.8]	**-11.40**	4.35	[-19.92;-2.89]
head clinical unit (medical) grade 1	1,051.40	7.59	[1036.52;1066.27]	**-16.61**	4.08	[-24.61;-8.61]

Visualization of random effects

Figure 22 shows that the confidence intervals for all the gender slope estimates are wider. This observation implies greater variability in earnings between females and males within these specific groups. Notably, since this particular grouping represents some of the highest earners, it is plausible that higher earnings contribute to increased variability. The wider confidence intervals indicate heightened uncertainty, leading to reduced confidence in the results. As a result, it appears that most of the estimates are not statistically significant.

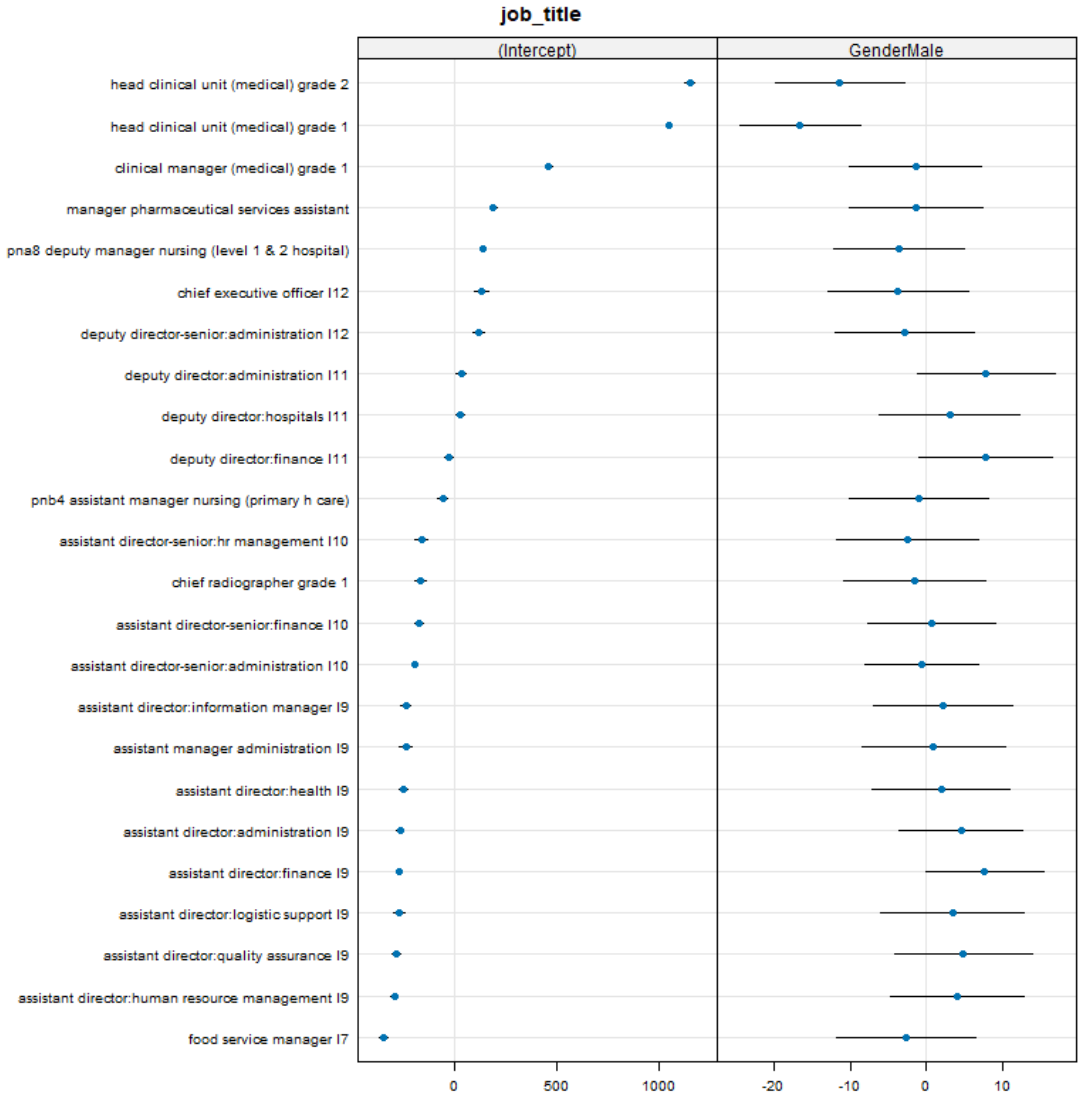

Figure 22: Departure from overall estimates-managers/directors.

Normality of random effects

The random effects normality plots in Figure 23 show that generally all the values lie on the confidence band with some slight deviation in the positive direction for job title intercept and negative deviations for the gender slope. The two groups showing some deviations are the same groups identified in the previous paragraph; head clinical unit (medical) grade 2 and head clinical unit (medical) grade 1 groups. As discussed before these two groups have higher earnings than average coupled by females earning higher than their male counterparts. In addition, the estimates for these two groups though they are wider just like the rest of the job titles, they are statistically significant.

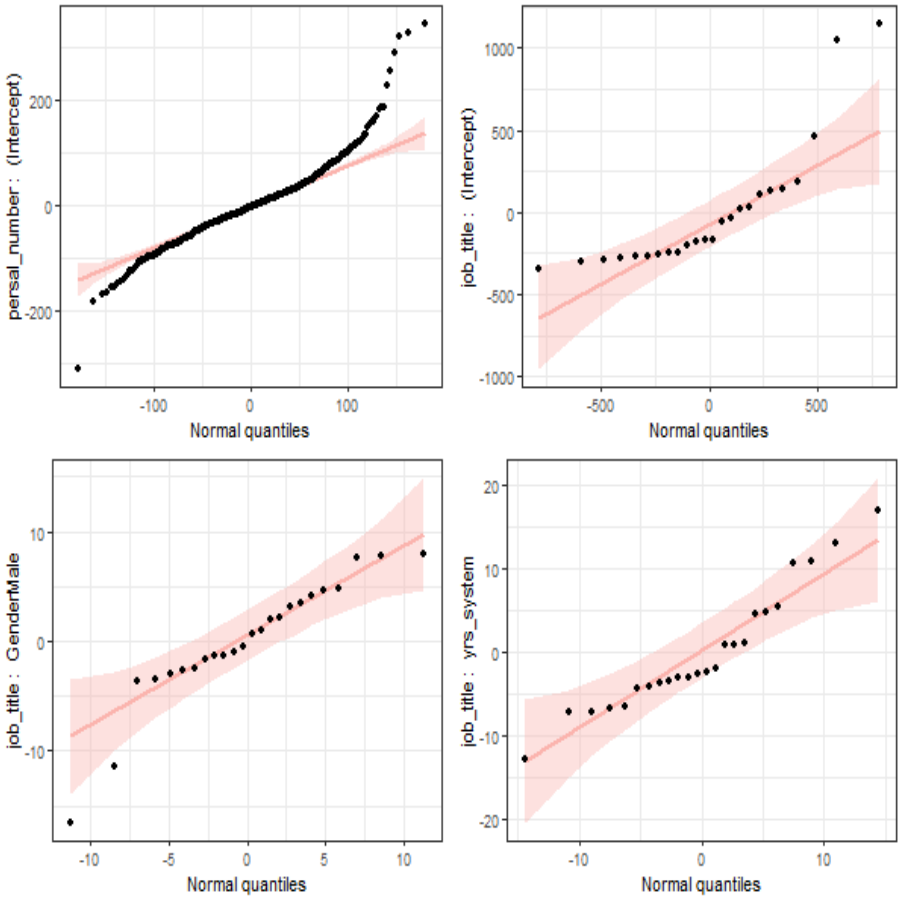

Figure 23: Normality of random effects-managers/directors.

5.3.6 Model 5 medical science

Within this specific grouping, individuals earn the highest average earnings, with an average of R894,000. Notably, males have higher average income, at R918,100, compared to females, whose average earnings are R869,200. Contrary to the assumption that females are primarily concentrated in low-skilled jobs, it is important to note that females are well represented in this grouping. There is a gender balance, with a nearly equal proportion of females (49%) and males (51%). This group comprise of medical specialists, medical officers and dentists. Among the various job titles, the most populated role is that of medical officer grade 1, constituting 29% of the population within this grouping, as indicated in Table 19.

Table 19: Descriptive statistics-medical science.

	Between job title percentages (frequencies)			Within percentages		Average earnings		
Job title	**Female**	**Male**	**Total**	**Female**	**Male**	**Female**	**Male**	**Overall**
medical specialist grade 2	0.8% (66)	1.0% (84)	0.9% (150)	44.0%	56.0%	1,432.9	1,410.0	1,420.1
medical specialist grade 1	2.9% (242)	3.3% (283)	3.1% (525)	46.1%	53.9%	1,224.4	1,236.5	1,230.9
medical officer grade 3	9.4% (790)	18.7% (1607)	14.1% (2397)	33.0%	67.0%	1,225.8	1,218.0	1,220.6
dentist grade 3	1.2% (99)	0.9% (79)	1.0% (178)	55.6%	44.4%	1,219.8	1,220.8	1,220.2
dentist grade 2	1.1% (91)	0.9% (77)	1.0% (168)	54.2%	45.8%	1,038.7	1,034.3	1,036.7
medical officer grade 2	9.6% (804)	10.7% (924)	10.2% (1728)	46.5%	53.5%	1,036.6	1,016.7	1,025.9
registrar (medical)	6.8% (573)	7.0% (600)	6.9% (1173)	48.8%	51.2%	917.0	914.7	915.8
medical officer grade 1	30.5% (2554)	27.5% (2370)	29.0% (4924)	51.9%	48.1%	907.5	909.2	908.3
dentist grade 1	4.9% (410)	4.2% (366)	4.6% (776)	52.8%	47.2%	861.6	863.1	862.3
medical officer (community service)	10.9% (910)	8.4% (723)	9.6% (1633)	55.7%	44.3%	720.1	717.1	718.8
dentist (community service)	1.3% (105)	0.8% (69)	1.0% (174)	60.3%	39.7%	714.5	701.3	709.3
medical officer (intern)	20.7% (1735)	16.6% (1433)	18.6% (3168)	54.8%	45.2%	546.7	545.6	546.2
Total	**100.0% (8379)**	**100.0% (8615)**	**100.0% (16994)**	**49.3%**	**50.7%**	**869.2**	**918.1**	**894.0**

Earnings are in 1000 Rands

Random effects medical science

It was observed from Table 19 that the highest earning groups have a higher proportion of males. Consequently these same groups, medical specialist grade 2 and medical specialist grade 1 demonstrate gender wage gap against females as seen in Table 20. On the other hand, the low earning groups such as medical intern and medical community service are predominantly female. It is interesting to note that within these lower-earning positions females are earning higher than males. For instance, dentist (community service) have a gender slope of -1.84, -1.31 for medical officer (community service) and -1.16 for medical officer (intern).

Table 20: Random effects-medical science.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
medical specialist grade 2	406.58	8.39	[390.13;423.03]	**3.27**	0.70	[1.89;4.64]
medical specialist grade 1	230.74	4.34	[222.24;239.24]	**1.84**	0.64	[0.59;3.09]
registrar (medical)	-39.16	3.00	[-45.04;-33.27]	**0.59**	0.62	[-0.62;1.8]
dentist grade 2	39.27	10.17	[19.33;59.22]	**0.42**	0.78	[-1.11;1.96]
medical officer grade 1	-53.27	1.32	[-55.85;-50.69]	**0.32**	0.56	[-0.78;1.42]
medical officer grade 3	185.30	2.22	[180.95;189.65]	**-0.22**	0.59	[-1.38;0.95]
medical officer grade 2	37.41	2.42	[32.66;42.15]	**-0.29**	0.59	[-1.45;0.88]
dentist grade 1	-92.59	3.17	[-98.81;-86.36]	**-0.55**	0.65	[-1.81;0.72]
dentist grade 3	141.34	9.50	[122.72;159.97]	**-1.08**	0.71	[-2.47;0.31]
medical officer (intern)	-392.56	1.54	[-395.58;-389.54]	**-1.16**	0.66	[-2.45;0.13]
medical officer (community service)	-227.91	1.90	[-231.62;-224.19]	**-1.31**	0.67	[-2.63;0.01]
dentist (community service)	-235.16	3.76	[-242.53;-227.8]	**-1.84**	0.99	[-3.78;0.11]

Visualization of random effects

Figure 24 shows the visual presentation of results discussed in Table 20. In summary, for the medical science grouping, it appears that higher-earning groups tend to be predominantly male, and within these groups, there exists a gender wage gap against females. Conversely, in lower-earning groups, which are more populated by females, there seems to be a gender wage gap in favor of females.

It is also important to note that the standard errors are remarkably low as observed from Table 20. In addition, Figure 24 shows that the confidence intervals are significantly narrower, spanning from -4 to 4. This indicates that the estimates are likely to be more precise and less variable. However, despite this increased confidence and certainty in the estimates, most of them overlap thereby not achieving statistical significance.

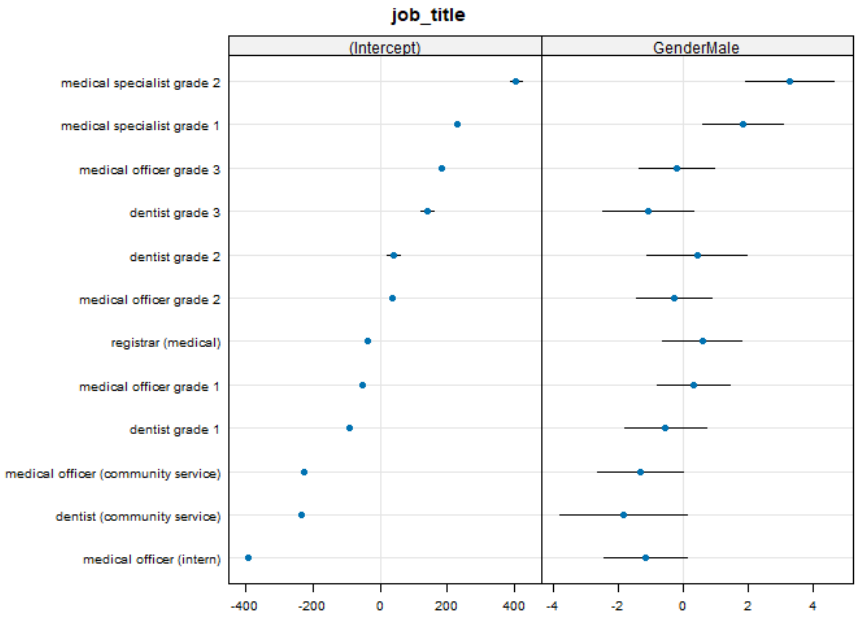

Figure 24: Departure from overall estimates-medical science.

Normality of random effects

The normality plots shown in Figure 25 have the confidence interval bands representing the range within which there is reasonable confidence that the true random effect lies. Upon examining the plots for job title intercept and gender slope it becomes evident that all the values fall within the band. This is in line with the small standard errors and narrow confidence intervals discussed in the previous paragraph. Therefore, the normality assumption for this model is satisfied. As a result, the estimates derived from this model are likely to be more robust.

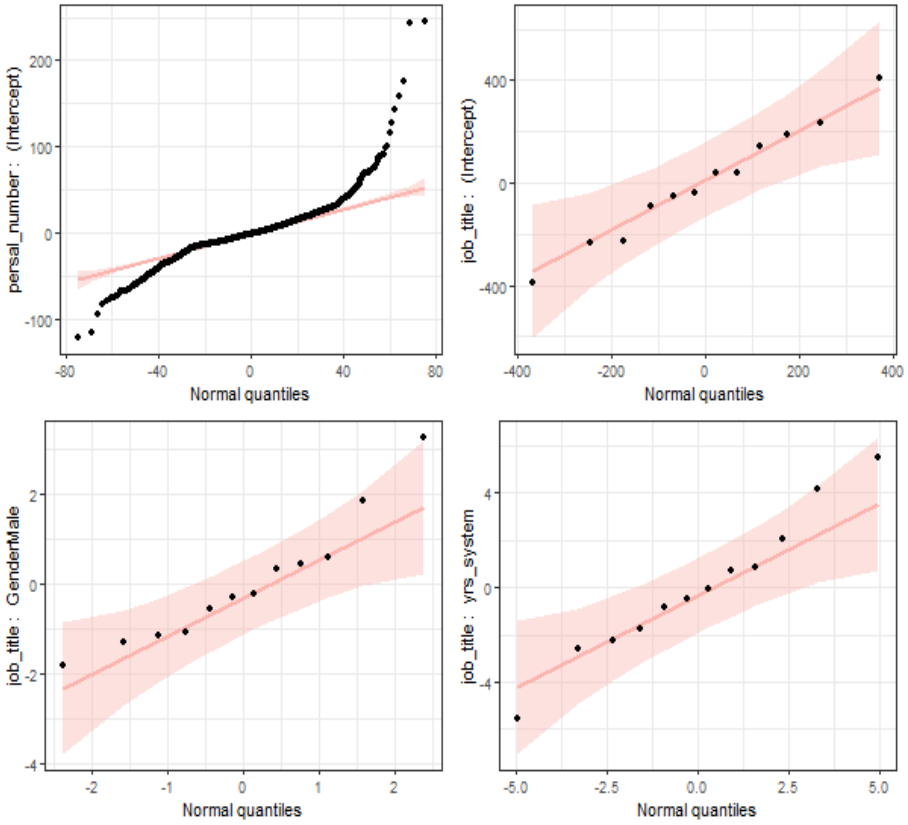

Figure 25: Normality of random effects-medical science.

5.3.7 Model 6 nurses

The nurses grouping is the most populated grouping, accounting for 42% of the total population. Table 21 reveals that this group is primarily female-dominated, with 87% of its employees being female. Within this nursing group, various types of nurses are represented, including professional nurses, nursing assistants, staff nurses and student nurses.

On average, earnings in this grouping amount to R212 500. Among the different job titles, the pnd1 lecturer nursing grade 1 earns the highest average income of R400 400, while the least paid position is held by the student nurse l3, earning R123 300.

Table 21: Descriptive statistics-nurses.

	Between job title percentages (frequencies)			Within percentages		Average earnings		
Job title	**Female**	**Male**	**Total**	**Female**	**Male**	**Female**	**Male**	**Overall**
pnd1 lecturer nursing grade 1	0.6% (699)	0.5% (105)	0.6% (804)	86.9%	13.1%	440.1	442.3	440.4
pnb1 clinical nurse practitioner gr 1 prim h care	1.6% (1819)	1.2% (249)	1.5% (2068)	88.0%	12.0%	434.0	437.7	434.5
pna2 professional nurse grade 1 (general nursing)	34.1% (38793)	32.4% (6874)	33.8% (45667)	84.9%	15.1%	285.7	284.7	285.6
pna1 professional nurse (community service)	4.1% (4698)	6.5% (1376)	4.5% (6074)	77.3%	22.7%	226.2	229.4	227.0
sn1 staff nurse grade 1	20.4% (23163)	20.6% (4383)	20.4% (27546)	84.1%	15.9%	189.3	189.2	189.2
na2 nursing assistant grade 2	4.3% (4871)	3.3% (694)	4.1% (5565)	87.5%	12.5%	177.2	178.7	177.4
na1 nursing assistant grade 1	31.5% (35873)	30.6% (6507)	31.4% (42380)	84.6%	15.4%	146.7	146.7	146.7
student nurse l3	3.4% (3838)	4.9% (1049)	3.6% (4887)	78.5%	21.5%	123.4	123.0	123.3
Total	**100.0% (113754)**	**100.0% (21237)**	**100.0% (134991)**	**84.3%**	**15.7%**	**213.0**	**210.2**	**212.5**

Earnings are in 1000 Rands

Random effects - Nurses

Despite the fact that females constitute the majority of the population in this grouping, examining the gender slope estimates from Table 22, reveals that males tend to earn more than females across most job titles. Specifically, within the pnd1 lecturer nursing grade 1 subgroup, male earnings surpass those of their female counterparts by a considerable margin compared to other groups. For instance, the gender slope estimate for pnd1 lecturer nursing grade 1 is 5.77 which is substantially high. On the other hand, females were earning higher than males within job titles such as na2 nursing assistant grade 2 with an estimate of -4.7.

Notably, the standard errors are generally small, and the confidence intervals are narrow across all groups, suggesting that these estimates are likely to be reliable. Furthermore, the fact that almost all estimates do not cross zero indicates statistical significance within these job titles.

Table 22: Random effects-nurses.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
pnd1 lecturer nursing grade 1	133.37	1.31	[130.81;135.94]	**5.77**	1.24	[3.34;8.2]
student nurse l3	-89.02	0.32	[-89.66;-88.39]	**2.11**	0.42	[1.27;2.94]
pna2 professional nurse grade 1 (general nursing)	25.98	0.18	[25.63;26.32]	**1.87**	0.31	[1.26;2.47]
pna1 professional nurse (community service)	-14.64	0.22	[-15.07;-14.21]	**0.82**	0.37	[0.09;1.54]
pnb1 clinical nurse practitioner gr 1 prim h care	110.55	0.84	[108.89;112.2]	**0.24**	1.14	[-2.01;2.48]
sn1 staff nurse grade 1	-45.92	0.19	[-46.3;-45.54]	**-2.18**	0.38	[-2.93;-1.44]
na1 nursing assistant grade 1	-76.96	0.16	[-77.28;-76.63]	**-3.91**	0.35	[-4.6;-3.21]
na2 nursing assistant grade 2	-43.37	0.48	[-44.3;-42.44]	**-4.70**	0.62	[-5.92;-3.48]

Visualization of random effects

As observed in the previous paragraphs, despite that females are more concentrated in this grouping, males are still dominating in higher paying job titles and earning more than females. For instance, Figure 26 shows that the job title groups with earnings higher than average are more likely to have gender wage gap against females while those with earnings below average are more likely to have gender wage gap in favour of females.

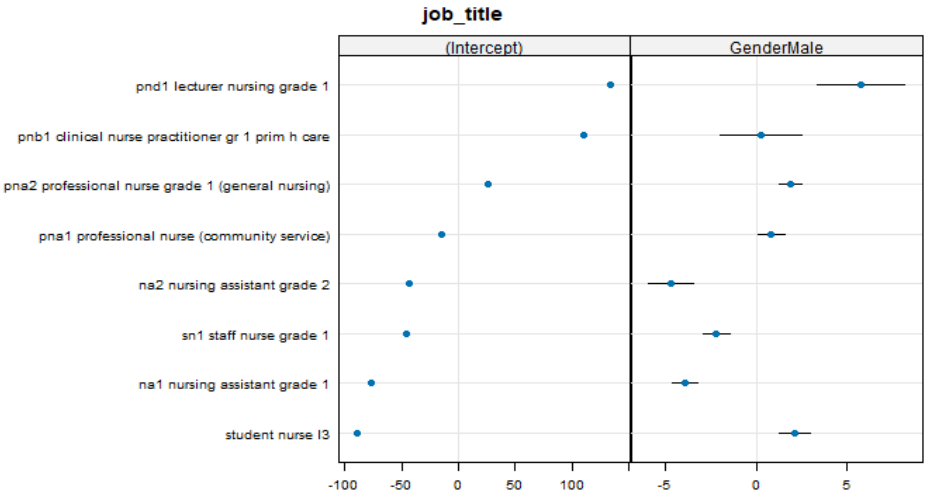

Figure 26: Departure from overall estimates-nurses.

Normality of random effects

Just like the random effect normality plots for medical science, both values for the job title intercept and gender slopes lie within the confidence band as shown in Figure 27. Thus the normality assumption for job title intercept and gender slope is holding. This suggests that the estimates are more likely to be robust.

This could be as a result that there are fewer levels in this group, only eight job titles. When a group has fewer distinct levels, the variability within each level is likely to be reduced because there will be fewer categories to account for differences. Therefore, with less variability, the data points within each level tend to cluster more closely together, resulting in a narrower spread. In such cases, the assumption of normality is more likely to hold more reliably.

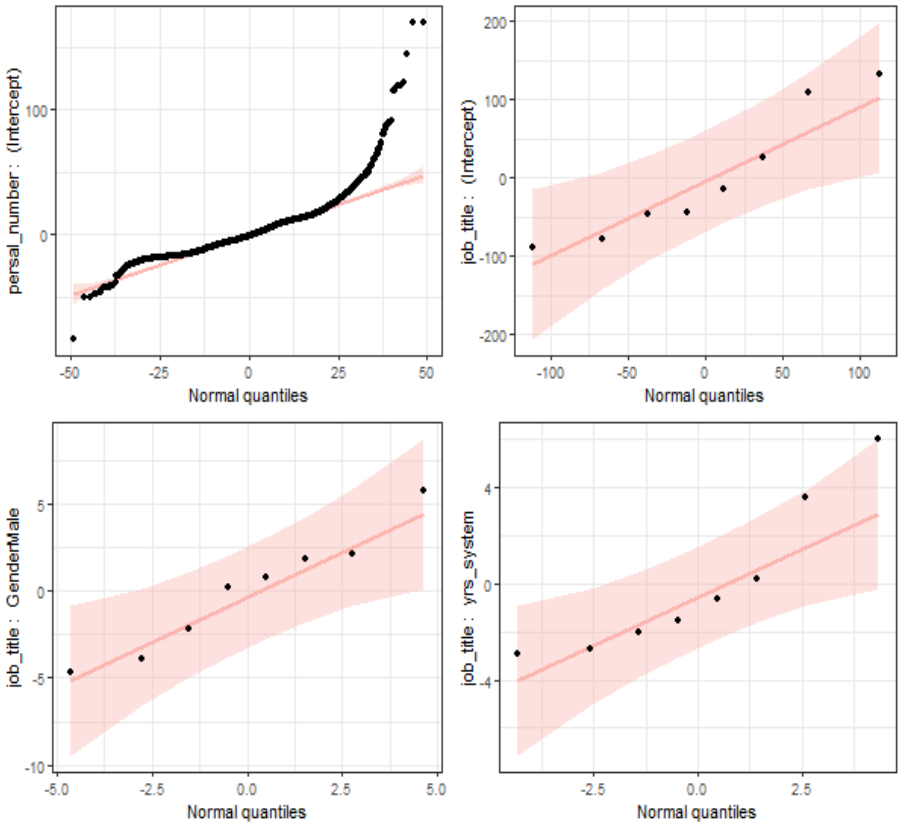

Figure 27: Normality of random effects-nurses.

5.3.8 Combined models

A separate model was fitted for all the six groupings constituting 142 different jobs titles. Table 23 represents the top ten job titles with a gender wage gap against females and the top ten job titles with a gender wage gap against males. The highest gender wage gap was observed for deputy director:administration l11 group (57.73), followed by assistant director:logistic support l9 (52.02) and then deputy director:hospitals l11 (38.24). On the other hand, gender wage gap was more

pronounced in favour of females for the following job titles; deputy director:finance l11 (-46.35), head clinical unit (medical) grade 1 (-25.98) and deputy director-senior:administration l12 (-23.93).

Looking at the job titles represented in Table 23. It is important to note that managers and directors held majority of the job titles, exhibiting gender wage gaps that affected both females negatively and favored them as well. Furthermore, the standard errors seem to be high though all the estimates represented are statistically significant.

Table 23: Random effects-combined models.

	Random intercepts-job title			Random slopes-genderMale		
Job title	**Estimate**	**Std error**	**95% CI**	**Estimate**	**Std error**	**95% CI**
deputy director:administration l11	243.82	5.40	[233.24;254.41]	**57.73**	5.68	[46.6;68.85]
assistant director:logistic support l9	11.39	6.87	[-2.08;24.86]	**52.02**	6.68	[38.92;65.11]
deputy director:hospitals l11	194.34	5.66	[183.24;205.44]	**38.24**	5.89	[26.7;49.78]
chief executive officer l12	410.79	7.28	[396.52;425.07]	**36.09**	6.04	[24.25;47.94]
pnb4 assistant manager nursing (primary h care)	181.48	4.61	[172.45;190.52]	**31.08**	8.05	[15.31;46.86]
assistant director:quality assurance l9	-66.98	3.90	[-74.62;-59.35]	**30.60**	4.64	[21.5;39.71]
pna8 deputy manager nursing (level 1 & 2 hospital)	349.90	2.88	[344.26;355.54]	**27.44**	5.24	[17.17;37.71]
administration clerk chief l7	-112.35	3.76	[-119.72;-104.97]	**21.94**	3.80	[14.49;29.4]
assistant director:administration l9	-27.90	2.72	[-33.22;-22.58]	**19.80**	2.62	[14.68;24.93]
supplementary diagnostic radiographer senior l5	-125.51	4.18	[-133.71;-117.31]	**16.34**	3.03	[10.39;22.29]
general worker gri	-201.15	1.59	[-204.27;-198.03]	**-13.04**	3.20	[-19.31;-6.78]
mortuary attendant l4	-175.62	2.99	[-181.49;-169.75]	**-14.33**	3.31	[-20.81;-7.85]
operator l3	-189.67	6.67	[-202.74;-176.6]	**-14.55**	5.74	[-25.8;-3.29]
food service manager l7	-91.56	2.96	[-97.36;-85.77]	**-15.10**	6.32	[-27.48;-2.73]
information officer l7	-83.27	3.79	[-90.69;-75.84]	**-15.96**	5.53	[-26.8;-5.11]
manager pharmaceutical services assistant	505.52	3.05	[499.54;511.5]	**-18.83**	3.86	[-26.4;-11.26]
financial clerk l6	-119.32	1.97	[-123.18;-115.46]	**-20.71**	2.50	[-25.62;-15.8]
deputy director-senior:administration l12	321.89	6.46	[309.23;334.54]	**-23.93**	6.95	[-37.55;-10.32]
head clinical unit (medical) grade 1	1,372.66	2.92	[1366.94;1378.39]	**-25.98**	2.84	[-31.54;-20.42]
deputy director:finance l11	205.34	4.64	[196.24;214.45]	**-46.35**	4.97	[-56.08;-36.62]

5.4 Model diagnostics

When working with linear mixed-effects models it is essential to assess model diagnostics to ensure the validity of the model assumptions. Earlier, the normality plots for random effects in each of the six fitted models were examined. These plots were not only assessing if the normality assumption was holding, they have been useful in identifying outliers and potential issues. It is important to note that normality of random effects has been covered and in this section the main focus will be on normality of residuals, linearity and homogeneous assumptions for each of the seven models.

5.4.1 Testing for linearity assumption

In the context of linear mixed effect regression models, the linearity assumption pertains to the relationship between independent variables and the response variable. There is an assumption that the effect of each independent variable on the response variable is linear. This can be assessed by creating a scatterplot where the differences between observed and predicted values are plotted against the predicted values from the model. By examining scatterplots for each model as shown in Figure 28, there is no evidence of a non-linear relationship. The scatter of points around the vertical line at zero appears random, indicating that the change in the response variable is proportional to the change in the predictor variables.

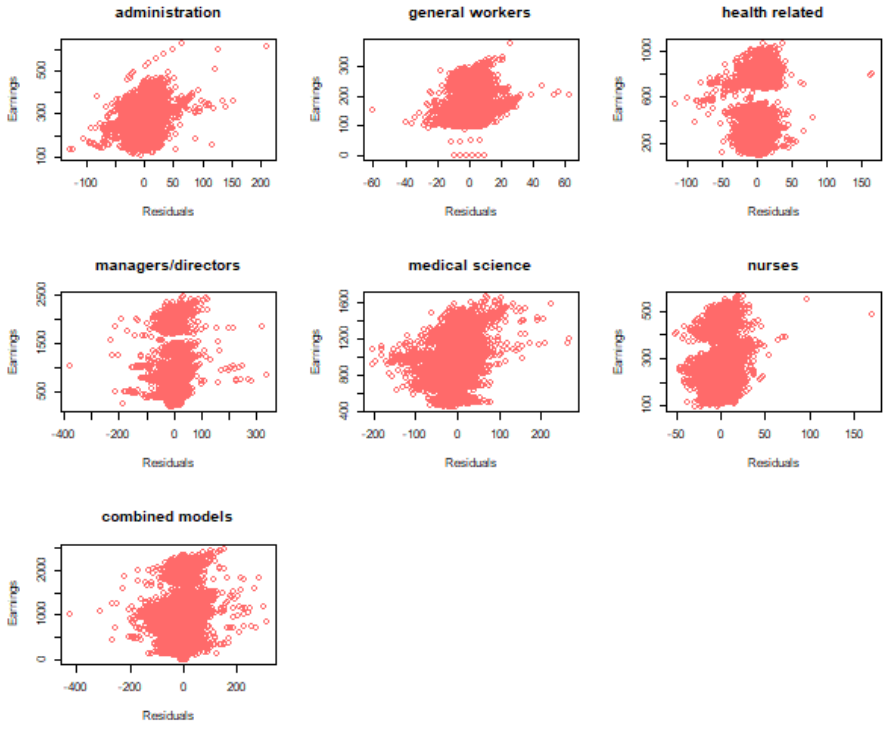

Figure 28: Linearity assumption.

5.4.2 Testing for homogeneity of variance

The assumption of homogeneity means that the variance of the residuals should be constant across all levels of the explanatory variables. By creating a fitted versus residual plot this assumption can be visualized. When examining the plots in Figure 29, an even spread of residuals around the centered line is observed. This suggests a relatively consistent variance across the entire range of fitted values. Consequently, it can be concluded that there is homogeneous dispersion of the residuals, indicating that the assumption of equal variance is upheld.

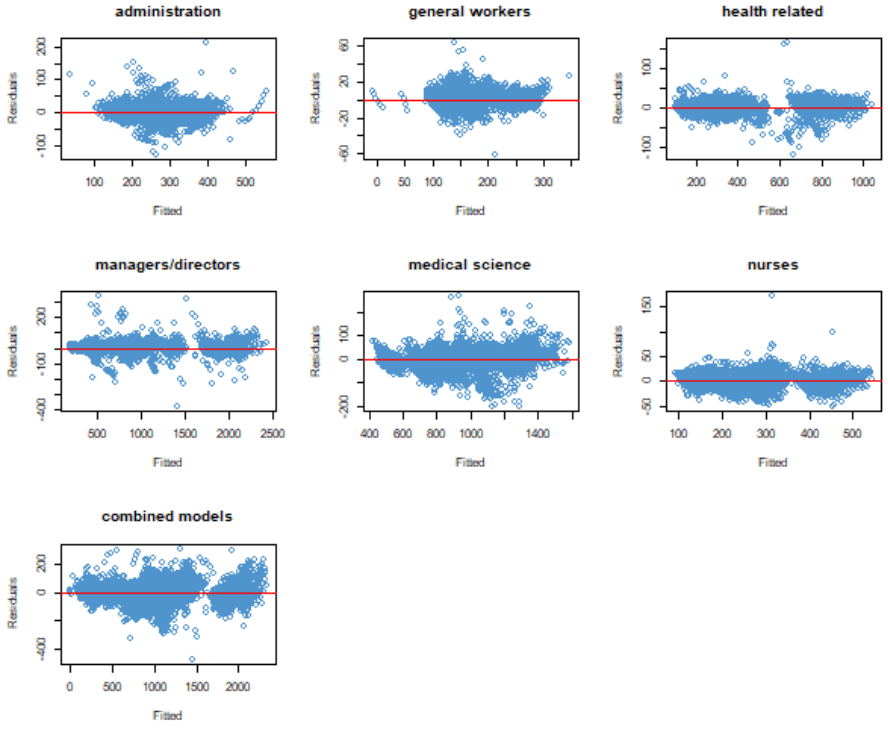

Figure 29: Homogeneity of variance assumption.

5.4.3 Testing for normality of residuals

Normality of residuals is typically tested using a quantile-quantile plot. Significant deviations from linearity of the observations or non-symmetric scales indicate a deviation from normality of residuals. The q-q plots in Figure 30 are deviating from normal especially at the tails but this is not much of a concern as the dataset constitute a large sample size which is less likely to be affected by deviations from normality. Additionally, a slight deviation at the tails is generally acceptable in practice.

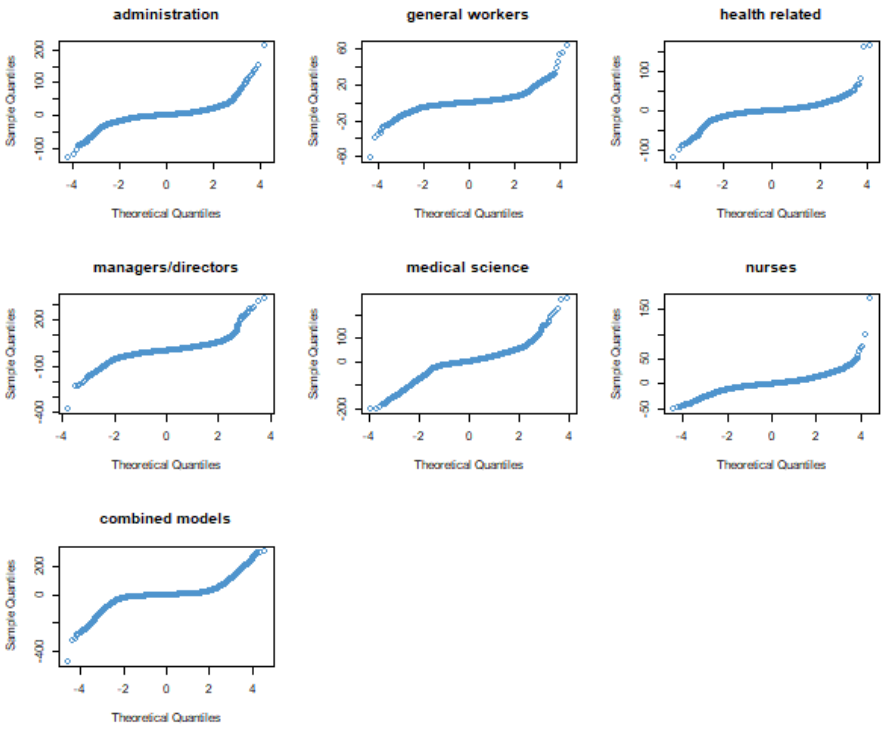

Figure 30: Normality of residuals assumption.

5.5 Regression summary output

Table 24 provides a summary of the regression output for seven models, which include models for the six groupings and the combined model. Focusing on the fixed coefficient of gender in the combined model, a positive and statistically significant gender coefficient of 3.88 is observed. This implies that, overall, males tend to earn higher than their female counterparts, while keeping other variables in the model constant. Furthermore, the gender coefficient for health related grouping is 5.38 and statistically significant. This suggests that there is significant gender wage gap against females in this grouping, holding all other variables constant. The other groupings that show positive coefficients are medical science with a value of 1.64 and nurses with a value of 2.08. This indicates that holding all other variables in the model constant, males are likely to earn higher than their female counterparts for medical science and nurses groupings. However, these estimates are not statistically significant.

On the other hand, the administration, general workers and managers/directors groupings have negative coefficients. This suggests that holding all other variables in the models constant, females are likely to earn higher than males in these groupings but the estimates are not statistically significant. Another variable of interest is years in the system. It can be observed that holding all other variables constant, as the number of years in the system increases the earnings also increase except for managers/directors.

The standard deviation of the random intercepts and slopes can be interpreted as the amount of variation in the intercepts and slopes that is not explained by the fixed effects. It can be observed from the models that much of the variability was explained by job title intercept (297.72), followed by persal number intercept (30.32). In addition, the random slopes for gender contributed more (13.21) compared to random slopes for the number of years in the system (7.85). It is interesting to note that the models have managed to capture most of the variability. This is evidence by very small residual standard deviation. In particular, the residual error for the combined models was 11.45, which accounted for only 3% of the total variance.

The regression summary output also shows the correlation between the job title intercept and the slopes. A closer look at the correlation between job title intercept and gender slope is negative for administration, general workers, health related and managers/ directors and positive for medical science, nurses and the combined models. A negative correlation implies that job titles with higher intercepts tend to have lower slopes and are less affected by the gender manipulation. On the other hand, a positive correlation imply that job titles with higher intercepts tend to be more affected by gender effects.

Looking at the R squared margin, about 0.9% of the variation in the combined models is explained by fixed effects in the model. On the other hand, the R squared conditional is showing a more comprehensive view of how well the models explain the data. For the combined models, 99.9% of the variance in the model is attributed to both fixed and random effects. In addition, the ICC for all the models is equal to 1, indicating that there is high similarity between values from the same subject.

It is important to note that the AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and RMSE (Root Mean Square Error) are used as the basis of model comparison, in this case the models under consideration come from different data sets therefore they are not comparable.

Table 24: Regression summary output.

	Administration	General workers	Health related	Managers/ directors	Medical science	Nurses	Combined Models
(Intercept)	-8502.05***	-3773.35***	-9297.88***	-43934.69***	-35751.99***	-8513.16***	-10073.64***
	(126.96)	(38.31)	(168.62)	(821.25)	(331.7)	(48.64)	(70.21)
GenderMale	-0.42	-0.44	5.38***	-2.72	1.64	2.08	3.88**
	(1.37)	(1.08)	(1.27)	(4.04)	(1.23)	(1.29)	(1.22)
yrs_system	2.88***	2.32***	3.78***	-0.63	2.49**	2.15+	5.27***
	(0.17)	(0.14)	(0.88)	(1.55)	(0.95)	(1.14)	(0.66)
SD (Intercept persal_number)	17.8	9.98	25.12	56.6	26.08	12.63	30.32
SD (Intercept job_title)	45.93	15.56	174.97	393.86	223.67	83.47	297.72
SD (GenderMale job_title)	6.9	5.66	4.96	7.23	1.59	3.53	13.21
SD (yrs_system job_title)	0.88	0.79	4.98	7.29	3.11	3.23	7.85
Cor (Intercept~GenderMale job_title)	-0.32	-0.04	-0.07	-0.61	0.71	0.61	0.1
Cor (Intercept~yrs_system job_title)	0.7	0.84	0.95	0.84	0.68	0.94	0.89
Cor (GenderMale~yrs_system job_title)	-0.3	0.25	0.03	-0.26	0.04	0.52	0.18
SD (Observations)	10.45	3.35	8.46	33	32	6.38	11.45
Num.Obs.	34502	64441	27670	7233	13652	108010	255834
R2 Marg.	0.151	0.241	0.016	0.024	0.062	0.030	0.009
R2 Cond.	0.969	0.986	0.998	0.995	0.983	0.997	0.999
AIC	274737.6	378547.5	215060.2	74746.4	137559.2	760740.8	2124855.2
BIC	275041.8	378883.2	215356.4	74994.3	137814.9	761066.9	2125241.9
ICC	1.0	1.0	1.0	1.0	1.0	1.0	1.0
RMSE	9.70	3.08	7.70	30.07	29.12	5.87	10.50

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

6 Conclusion

The objective of this study was to analyse gender wage gap within each of the finer occupational subcategories by incorporating both random and fixed effects. Though the Blinder-Oaxaca decomposition method has been widely used in estimating gender wage gap, a significant portion of the gender wage gap has been left unexplained and in some studies this unexplained portion has been found to be greater than the portion explained by the factors included in the model. Thus, by using linear mixed effects regression, the random effects account for both individual and group differences, while also addressing the inherent correlation among repeated observations. Consequently, the marginal R-squared for all the models was very low, specifically at 0.9% for the combined model. This suggests that the fixed effects alone do not explain a significant portion of the variability in the model. However, the conditional R-squared was remarkably high for all models, particularly reaching 99.99% for the combined models. This implies that incorporating both fixed and mixed effects is essential because it accounts for a significant portion of the variability in the model. In addition, findings from this study revealed that the majority of the variance was effectively accounted for by the factors in the model as the unexplained variance was only 3%. This underscores the robustness of mixed models in explaining the variation within the model.

Traditionally, the gender wage gap in South Africa has been examined at an aggregated level. While some studies have attempted to disaggregate the data, this has typically been limited to broad categories such as occupations, salary bands, and age groups. Unfortunately, this approach did not yield significantly better results because within those broad categories, women and men were often engaged in different job levels, with women typically taking lower positions which were paid less. This study took a different approach by analysing gender wage gap at finer-grained level, that is within each job title. This approach yields a more nuanced understanding of the gender wage gap, as it enables comparisons among employees with similar characteristics and at the same level, thereby providing a more accurate and comparable measure. Interestingly, the study's findings revealed that a significant 83% of the total variance in the data was attributed to differences in job titles. This emphasises the importance of disaggregating gender wage gap at more granulated and comparable levels. Moreover, it allows for the identification of specific groups that require targeted interventions, which is crucial for policymakers.

The findings of this study indicate the existence of a gender wage gap, disproportionately affecting females. In South Africa there have been conflicting reports regarding whether gender wage gap is widening or narrowing. According to a report by Pleace et al. (2023) gender wage gap was widening. However, other studies, including those by Mosomi (2019); Fisher et al. (2021) reported contracting results which showed a narrowing of gender wage gap. Consistent with observations from Pleace et al. (2023), the findings from this study indicate a widening of the gender wage gap over the years. However, upon closer analysis, it becomes evident that the gender wage gap narrowed from 2010, followed by a gradual widening starting from 2015. Mosomi (2019); Fisher et al. (2021) analysed the data covering 1995-2015 period while Pleace et al. (2023) analysed data from 2010 to 2021. Consequently, we can infer that the inequality gap may have narrowed between 1995 and 2015, only to begin widening again from 2015 to 2021.

Despite the widening of gender wage gap in South Africa over the recent years, interventions aimed at improving the productive characteristics of women seem to have yielded some positive results. This is evidenced by gender wage gap in favour of females in three of the six occupational groupings which are general workers, administration and managers or directors. Interestingly, when examining gender wage gap across all 142 job title groups analyzed, 42% of these groups showed inequality in favor of women. Despite the persistence of the overall gender wage gap,

a more detailed examination reveals that females actually hold an advantage in certain groups. Furthermore, it is noteworthy that a significant proportion of females hold managerial and highly skilled positions. For instance, 49% of medical professionals and 60% of managers/directors are female.

Studies on gender wage gap highlight that women, often considered primary caregivers, may prioritize occupations compatible with domestic responsibilities over promotions. Consistent with other studies, it was found in this study that female employees experience delayed promotions compared to males. Notably, most promotions were linked to facility changes. Thus, the reluctance of women to relocate due to family commitments may lead to missed promotional opportunities. Interestingly, it was evident from the study that female employees with longer tenure tend to earn more than their male counterparts. In addition, gender wage gap in favour of females was noticed for female permanent workers.

Several studies examining gender wage gap have consistently found that at the upper echelons, males tend to earn more than their female counterparts. Additionally, previous researches have highlighted a lack of women in high-ranking positions within institutional sectors. Findings from this study corroborates the same observations, indicating that gender wage gap is significantly higher particularly among high earners. For instance, it was observed that gender wage gap was especially significant for employees at the 90th percentile of the income level. Moreover, within the foreigners group who typically include highly skilled professionals such as medical specialists with substantial incomes, gender wage gap against females was also observed. Furthermore, findings from this study revealed that as employees advance to more senior positions, there is a notable decline in the proportion of females occupying those higher roles.

In South Africa, numerous studies have examined gender wage gap using household survey data. However, these datasets often lack specific information, such as detailed occupational levels for employees, and typically provide broader occupational categories. As a solution, Bhorat (2013) suggested that addressing the high unexplained gender wage gap requires the use of non-traditional survey instruments. Consequently, this study employed administrative data from human resource records for national department of health in the Eastern Cape, which offered insights into finely grained job titles.

However, while administrative data offers advantages such as accuracy, freedom from reporting bias and provides detailed information on occupational levels, it does come with limitations. In this study, the dataset lacks information on factors like education background, number of dependents, marital status, and whether the household is female or male-headed. Consequently, the impact of these factors on gender wage gap is restricted. Additionally, the data only covers formal employees, excluding informal workers who are predominantly female. The study's scope is limited to Eastern Cape province, making it non-nationally representative. Furthermore, unlike most gender wage gap studies that use hourly earnings, this study relied on annual earnings. This has a drawback of rendering data for part-time or casual appointments invalid especially where the number of hours worked is unspecified.

The Blinder-Oaxaca method has been a widely accepted approach for estimating gender wage gap. Therefore, it would be valuable to apply this method using the same dataset and assess the impact of incorporating random effects. While this current study focused on Eastern Cape province, extending the analysis to encompass all provinces within the country would provide a more comprehensive perspective on gender wage disparities. Furthermore, future research could adopt the positive deviance approach by studying groups that exhibit gender wage gap in favour of women, in order to identify unique characteristics that set them apart from those with gender

wage gap against females. Thus, understanding these distinguishing factors may guide efforts to reduce gender wage disparities effectively.

When analyzing the gender wage gap, relying on broad occupational groups, can inadvertently shift the focus from gender-based disparities to occupational gender segregation. This is because when grouping employees into broad occupational categories, we risk comparing individuals at different levels within those categories. For instance, considering dentists and dental assistants. Although they fall under the same occupational umbrella, their roles, education, and experience levels differ significantly. Therefore to accurately assess gender wage disparities, comparable groups are crucial for a comprehensive understanding of gender-based wage differences.

To sum up, incorporation of random effects techniques and more detailed job categories has considerably enriched this analysis, allowing for a deeper and more accurate understanding of the determinants of the gender wage gap. These approaches have opened up new possibilities for the study of gender inequality at the workplace that could guide future research and public policy more effectively.

Appendices

A Appendix

A.1 Additional regression output I

Table 25: Additional regression output I.

	Administration	General workers	Health related	Managers/ directors	Medical science	Nurses	Combined Models
FacilityClinic	7.32**	-11.94***	-1.2	14.93			4.44+
	(2.81)	(3.23)	(4.96)	(17.38)			(2.52)
FacilityCommunity HC	10.18***	-11.77***	3.86	-3.6	2.8	0.54**	5.22*
	(2.9)	(3.23)	(4.93)	(18.04)	(3.96)	(0.18)	(2.52)
FacilityDistrict hospital	7.11*	-10.43**	2.1	7.68	8.65*	-0.09	5.73*
	(2.8)	(3.22)	(4.9)	(17.16)	(3.63)	(0.14)	(2.52)
FacilityDistrict office	9.74***	-10.78***	-0.65	8.11	10.53**	-0.75***	4.51+
	(2.78)	(3.22)	(4.9)	(17.06)	(3.82)	(0.17)	(2.51)
FacilityMortuary	8.67*	-13.9**	-10.15	14.69	5.1		-1.7
	(4.1)	(4.34)	(8.31)	(18.92)	(6.9)		(3.17)
FacilityProvincial hospital	9.38***	-10.2**	1.55	29.28+	0.92	-0.09	5.38*
	(2.79)	(3.22)	(4.91)	(17.19)	(3.88)	(0.19)	(2.51)
FacilityRegion hospital	11.03***	-8.44**	4.87	21.74	15.4***	0.87***	7.53**
	(2.82)	(3.23)	(4.92)	(17.4)	(3.94)	(0.19)	(2.52)
FacilitySpecialised hospital	7.5**	-11.37***	2.27	14.26	11.71*	-0.05	4.91+
	(2.88)	(3.23)	(5.02)	(17.27)	(4.58)	(0.26)	(2.53)
districtBuffalo city	3.41***	0.49+	-0.66	3.66	-5.92**	0.05	-0.66*
	(0.77)	(0.29)	(0.85)	(5.1)	(2.25)	(0.21)	(0.29)
districtGqabi	3.14**	1.03*	1.54	-11.15+	-1.52	1.21***	1.52***
	(1)	(0.43)	(1)	(6.18)	(3.34)	(0.32)	(0.43)
districtHani	1.16	0.8**	1.69*	-3.99	4.69+	0.92***	0.52
	(0.81)	(0.3)	(0.85)	(5.46)	(2.5)	(0.22)	(0.32)
districtMandela	2.18*	3.15***	2.5**	-1.01	-8.96***	1.63***	-0.18
	(0.87)	(0.33)	(0.88)	(5.68)	(2.27)	(0.26)	(0.35)
districtNot_specified	5.92***	-0.19	30.9*	21.83***		-1.14**	4.33***
	(0.95)	(0.66)	(15.46)	(5.26)		(0.4)	(0.56)

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

A.2 Additional regression output II

Table 26: Additional regression output II.

	Administration	General workers	Health related	Managers/ directors	Medical science	Nurses	Combined Models
districtNzou	2.95**	1.74***	1.15	2.32	-0.44	0.3	0.3
	(0.95)	(0.38)	(0.96)	(5.46)	(2.87)	(0.26)	(0.39)
districtOR Tambo	3.8***	1.48***	0.74	6.12	0.14	-0.35	1.51***
	(0.79)	(0.31)	(0.87)	(5.15)	(2.22)	(0.22)	(0.31)
districtS Baartman	3.42***	1.77***	1.51	22.35***	-0.8	1.27***	1.66***
	(1.01)	(0.35)	(1.09)	(6.2)	(2.61)	(0.28)	(0.39)
raceAsian	42.4***	-2.53	8.98***	-18.81*	-2.43	7.71*	-3.87**
	(8.38)	(7.12)	(2.46)	(9.59)	(2.09)	(3.9)	(1.25)
raceColoured	2.53*	2.04***	6.58***	-0.01	-0.64	2.9***	3.64***
	(1.05)	(0.44)	(1.08)	(7.23)	(2.07)	(0.36)	(0.55)
raceWhite	8.26***	5.39***	5.42***	0.08	-0.02	7.07***	3.86***
	(1.44)	(0.95)	(1.24)	(5.71)	(1.36)	(0.77)	(0.7)
AppointmentPermanent	-9.9***	126.98***	6.92***	1.3	-8.28***	0.18	116.11***
	(1.83)	(7.14)	(1.53)	(4.95)	(1.7)	(0.28)	(30.7)
AppointmentProbation	-9.42***	128.21***	8.16***	-3.92	-10.5***	-0.73*	115.97***
	(1.84)	(7.14)	(1.52)	(5.05)	(1.59)	(0.28)	(30.7)
age	0.28***	0.14***	0.31***	1.24***	1.46***	-0.01	0.29***
	(0.04)	(0.01)	(0.05)	(0.21)	(0.1)	(0.01)	(0.02)
CitizenSouth Africans	19.26	-2.98**	1.14	-6.44	3.6*	-11.67***	-7.56***
	(18.24)	(0.95)	(3.44)	(7.09)	(1.73)	(2.29)	(0.94)
reporting_year	4.31***	1.87***	4.75***	22.15***	18.19***	4.34***	5.11***
	(0.06)	(0.02)	(0.08)	(0.41)	(0.16)	(0.02)	(0.03)
rural_urbanUrban	0.17	-0.17	-3.6***	2.03	0.83	-0.27*	-0.3
	(0.57)	(0.18)	(0.61)	(3.67)	(1.68)	(0.13)	(0.19)

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

References

Adelekan, A. M. and Bussin, M. H. (2018). Gender pay gap in salary bands among employees in the formal sector of south africa. *SA Journal of Human Resource Management*, 16(1):1–10.

Adeleken, A. and Bussin, M. H. (2022). Occupational segregation and gender pay gap dynamics in the formal sector of south africa. *SA Journal of Human Resource Management*, 20:1660.

Akinwande, M. O., Dikko, H. G., Samson, A., et al. (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. *Open journal of statistics*, 5(07):754.

Ansel, B. (2017). Do United States women choose low-paid occupations, or do low-paid occupations choose them?

Bargain, O., Etienne, A., and Melly, B. (2018). Public sector wage gaps over the long-run: evidence from panel administrative data. IZA Discussion Paper.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Bates, D. M. (2010). lme4: Mixed-effects modeling with r.

Bezuidenhout, C., Van Rensburg, C. J., Matthee, M., and Stolzenburg, V. (2019). Trading firms and the gender wage gap: evidence from south africa. *Agenda*, 33(4):79–90.

Bhorat, H. and Goga, S. (2013). The gender wage gap in post-apartheid south africa: A re-examination. *Journal of African Economies*, 22(5):827–848.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865.

Brauer, M. and Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, 23(3):389.

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920960351.

Burdín, G., De Rosa, M., Vigorito, A., and Vilá, J. (2022). Falling inequality and the growing capital income share: Reconciling divergent trends in survey and tax data. *world Development*, 152:105783.

Casale, D. and Posel, D. (2011). Unions and the gender wage gap in South Africa. *Journal of African Economies*, 20(1):27–59.

Casale, D., Posel, D., and Mosomi, J. (2021). Gender and work in South Africa.

Cho, I., Hu, B., and Berry, C. M. (2023). A matter of when, not whether: A meta-analysis of modesty bias in east asian self-ratings of job performance. *Journal of Applied Psychology*, 108(2):291.

Cho, S.-J., De Boeck, P., Naveiras, M., and Ervin, H. (2022). Level-specific residuals and diagnostic measures, plots, and tests for random effects selection in multilevel and mixed models. *Behavior Research Methods*, 54(5):2178–2220.

DPRU, D. P. R. U. (2024). Measuring the impact of the 2023 national minimum wage increas. *A report for the National Minimum Wage Commission.*

Fisher, B., Biyase, M., Kirsten, F., and Rooderick, S. (2021). Gender wage discrimination in south africa within the affirmative action framework. *The Journal of Developing Areas*, 55(2).

Fortin, N. M., Bell, B., and Böhm, M. (2017). Top earnings inequality and the gender pay gap: Canada, sweden, and the united kingdom. *Labour Economics*, 47:107–123.

Gradín, C. (2021). Occupational gender segregation in post-apartheid south africa. *Feminist Economics*, 27(3):102–133.

Grün, C. (2004). Direct and indirect gender discrimination in the south african labour market. *International Journal of Manpower*, 25(3/4):321–342.

Kerr, A. and Wittenberg, M. (2019). Earnings and employment microdata in south africa (wider working paper 2019/47).

Knief, U. and Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53(6):2576–2590.

Kollamparambil, U. and Razak, A. (2016). Trends in gender wage gap and discrimination in south africa: A comparative analysis across races. *Indian Journal of Human Development*, 10(1):49–63.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Loy, A., Hofmann, H., and Cook, D. (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*, 26(3):478–492.

Maas, C. J. and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3):86–92.

Mosomi, J. (2019). An empirical analysis of trends in female labour force participation and the gender wage gap in south africa. *Agenda*, 33(4):29–43.

Mosomi, J. N. (2018). Distributional changes in the gender wage gap in the post-apartheid south african labour market.

Muller, C. et al. (2009). Trends in the gender wage gap and gender discrimination among part-time and full-time workers in post-apartheid south africa. *Development Policy Research Unit, School of Economics, University of Cape Town.*

Musetsho, M., Isac, N., and Dobrin, C. (2021). Gender inequalities in the workplace: Case study of south africa. *Management and Economics Review*, 6(1):70–81.

Nadler, J. T. and Stockdale, M. S. (2012). Workplace gender bias: Not just between strangers. *North American Journal of Psychology*, 14(2).

Nagle, C. (2019). An introduction to fitting and evaluating mixed-effects models in r. *Pronunciation in Second Language Learning and Teaching Proceedings*, 10(1).

Oyenubi, A. and Mosomi, J. (2024). Utility of inequality sensitive measures of the gender wage gap: Evidence from south africa. *Economic Analysis and Policy*, 81:576–590.

Paterson, S. (2010). What's the problem with gender-based analysis? gender mainstreaming policy and practice in canada. *Canadian Public Administration*, 53(3):395–416.

Pinheiro, J. C. and Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56.

Pleace, M., Clance, M., and Nicholls, N. (2023). Sa-tied.

PricewaterhouseCoopers, P. (2019). Executive directors: Practices and remuneration trends report 2019.

Salinas Ruíz, J., Montesinos López, O. A., Hernández Ramírez, G., and Crossa Hiriart, J. (2023). Generalized linear mixed models with applications in agriculture and biology.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., and Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, 11(9):1141–1152.

Schielzeth, H. and Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral ecology*, 20(2):416–420.

Schweinberger, M. (2021). Fixed-and mixed-effects regression models in r.

Si, C., Nadolnyak, D., Hartarska, V., et al. (2021). The gender wage gap in developing countries. *Applied Economics and Finance*, 8(1):1–12.

Sinden, E. (2017). Exploring the gap between male and female employment in the south african workforce. *Mediterranean Journal of Social Sciences*, 8(6):37.

Singmann, H. and Kellen, D. (2019). An introduction to mixed models for experimental psychology. In *New methods in cognitive psychology*, pages 4–31. Routledge.

StatsSA (2022). Quarterly labourforcesurvey:quarter22022.

StatsSA (2024). Consumer price index. *STATISTICAL RELEASE P0141.*

Steyn, R. and Jackson, L. (2015). Gender-based discrimination in south africa: A quantitative analysis of fairness of remuneration. *South African journal of economic and management sciences*, 18(2):190–205.

Strittmatter, A. and Wunsch, C. (2021). The gender pay gap revisited with big data: Do methodological choices matter? *arXiv preprint arXiv:2102.09207*.

Toczek, L., Bosma, H., and Peter, R. (2021). The gender pay gap: income inequality over life course–a multilevel analysis. *Frontiers in sociology*, 6:815376.

Venables, W. N., Ripley, B. D., Venables, W., and Ripley, B. (2002). Random and mixed effects. *Modern applied statistics with S*, pages 271–300.

Wekwete, N. N. (2014). Gender and economic empowerment in africa: Evidence and policy. *Journal of African Economies*, 23(suppl_1):i87–i127.

West, B. T., Welch, K. B., and Galecki, A. T. (2022). *Linear mixed models: a practical guide using statistical software*. Crc Press.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Wodon, Q. and De La Briere, B. (2018). Unrealized potential: the high cost of gender inequality in earnings.

Xu, Y. (2015). Focusing on women in stem: A longitudinal examination of gender-based earning gap of college graduates. *The Journal of Higher Education*, 86(4):489–523.

Yu, D. and Fredericks, F. (2017). The effect of affirmative action on the reduction of employment discrimination, 1997-2015. Technical report, Economic Research Southern Africa.

Yu, H., Jiang, S., and Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social science research*, 53:118–136.

Zuur, A., Hilbe, J., and Ieno, E. (2013). A beginner's guide to glm and glmm with r, beginner's guide series. *Newburgh: Highland Statistics Limited*.