Dissertation presented for the degree of
Master of Science
in the
Department of Statistics
University of Cape Town

Generating New Data Points Using Singular Value Decomposition

Author:
Tlhologello Biyana

Supervisor:
Dr Juwa Nyirenda

August 21, 2024

Abstract

This study presents an innovative solution to the challenge of generating new data points for small data sets. It introduces a Single Value Decomposition (SVD)-based model that draws inspiration from the ability of SVD to estimate a lower rank matrix. This approach seeks to overcome the limitations imposed by sample size constraints by expanding available data. Motivated by challenges faced during algorithm development due to small data sets, the study proposes the SVD-based model, evaluates its efficacy in replicating original data attributes and compares model performance with new and original data. The method involves utilising SVD to generate new data, mimicking a predictive modelling formula by combining systematic and error components. The generated data set retains the distribution of the original data but introduces distinct error values, facilitating efficient data generation. Through graphical and quantitative assessments, including histograms, box plots, correlation analysis and reconstruction error evaluations, the effectiveness of the method is demonstrated. The study focuses on comparing SVD-generated data sets with original data across three data sets: Abalone, Life Expectancy and NBA. Findings indicate close approximation of distribution, correlation and model performance attributes between SVD-generated and original data sets. Improved similarity with increasing observation count enhances comparability and model performance of SVD-generated data. While minor deviations are noted in specific scenarios, the study underscores potential of SVD in generating new data points from the original data sets, making it a valuable tool for data augmentation and analysis across diverse data sets.

Dedication

This research paper is dedicated with profound love and deep appreciation to those who have played an immeasurable role in shaping my journey and inspiring me to reach greater heights:

To my family, who have been my unwavering pillars of strength, offering boundless sacrifices and unwavering support throughout this arduous yet rewarding academic pursuit. Your boundless love and encouragement have fueled my determination and I am endlessly grateful for the enduring bond we share.

To the memory of my beloved grandmother, whose nurturing presence and profound wisdom continue to guide me, even in her absence. Though she didn't witness this momentous milestone, her unyielding love and guidance during my formative years have been the cornerstone of my character and aspirations.

To my incredible parents, Thembe Biyana and Johannes Nkwe, who have bestowed upon me more than I could have ever asked for in parents. Your unwavering belief in my potential and the unending support you have provided throughout my life have been the wind beneath my wings, propelling me forward with confidence and determination.

To my cherished friends, who have walked this journey by my side, offering encouragement, laughter and understanding. Your presence has illuminated even the darkest moments and your camaraderie has made every step of this journey worthwhile.

And a special dedication to Keneilwe Mmako, whose unwavering support and consistent encouragement during the most challenging times have been a beacon of light in my life. Your belief in my abilities and your steadfast presence have given me the strength to persevere and never lose sight of my goals.

This work is a testament to the love, support and belief bestowed upon me by these exceptional individuals. Each one of you has left an indelible mark on my heart and soul, shaping not only this academic achievement but also the person I am today. I carry your love and inspiration with me and I dedicate this paper to you as a token of my gratitude and love.

Acknowledgements

I would like to express my sincere gratitude to the following individuals and organi sations for their invaluable contributions to this research:

- Dr. Juwa Nyirenda: I am deeply thankful to Dr. Nyirenda for serving as my academic supervisor throughout this research project. His guidance, expertise and unwavering support have been instrumental in shaping the direction and quality of this work.
- Dr. R.K Thobejane: I extend my heartfelt appreciation to Dr. Thobejane for providing financial support, which made this research possible. His belief in the significance of this study encouraged and motivated me to pursue my academic aspirations.
- Isaac Mophatlane: I am grateful to Mr. Mophatlane for his generous financial assistance, which significantly contributed to the successful completion of this research. His support eased the financial burdens and allowed me to focus on the study's objectives.
- Kaggle: I would like to acknowledge Kaggle for providing access to valuable datasets. The data contributions from Kaggle played a pivotal role in the empirical analysis, enriching the outcomes of this research.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Context

The true power of a supervised learning algorithm resides in its intrinsic ability to detect and discover elaborate patterns that lie within a data set. The effectiveness of the algorithm in detecting these patterns relies on their inherent strength, as stronger and more prominent patterns are readily identifiable, ultimately enhancing the capacity and overall performance of the algorithm. The difficulty of supervised learning relates to two crucial questions:

1. What is the minimum amount of data required to approximate the unknown relationship between the inputs and output? In various real-world scenarios, inputs and their corresponding outputs are often present without an explicitly known mathematical relationship. The objective is usually to estimate this unknown relationship using the available data. The minimum data amount needed represents the smallest data set size enabling the development of a reasonably accurate model that captures the core relationship between the inputs and the output.

2. How much data is required to estimate the performance of an approximation of the mapping function? The size of a data set has a significant impact on the reliability of performance metrics in machine learning. With a larger data set, performance results tend to be more statistically significant, reducing sampling variability and leading to more stable and reliable metrics.

Interestingly, not all data falls into the category of Big Data, despite the growing popularity of Big Data tools and expertise. Smaller sample sizes are quite prevalent in fields such as medicine, sociology, psychology, geology and so on (EduPristine, 2016). Small data sets may occur because of the cost involved in obtaining a sample. For example, conducting in-person interviews is expensive so that the majority of studies that involve primary research with individuals result in small data. Furthermore, small data sets may occur because the population from which the sample is extracted is small to begin with. For example, the population of nations in the world; the number of schools in a region or a group of elite athletes who compete at the highest level in a particular sport. In statistical analysis, the significance of small data sets should not be underestimated, as it has the potential to amplify specific challenges that may impact the integrity of the analysis. These challenges relate to:

- **Outliers** - refers to a data point that significantly deviates from the typical patterns or distribution of the rest of the data. Managing outliers is crucial for several models, but it can be manageable if the number of outliers is insignif-

icant. This is not the case for small data sets since even a small number of outliers can make up a significant portion and significantly impact the model (EduPristine, 2016).

- **Train and test data** - in model building, it is typically preferable to split the data into two parts: the "training set" on which the model is trained and the "test set" on which the general performance of the model is evaluated. If the test set is also used for parameter tuning, it is known as the cross-validation set, which requires dividing the data into three sets. However, when the amount of available data is small, excluding too many samples may result in insufficient observations in the training, test or cross-validation sets, making it difficult to train models that are able to generalise well, generate meaningful performance estimates or adequately optimise parameters (EduPristine, 2016).
- **Overfitting** - occurs when a model learns to perform exceptionally well on the training data but fails to generalise effectively to new, unseen data. Instead of capturing the underlying patterns and relationships in the data, an overfitted model memorises noise and specific details of the training set, leading to poor performance on real-world data. When the training data set is small, there is an increased probability of encountering the problem of overfitting. Furthermore, using cross-validation as a preventive strategy against overfitting presents a similar concern. This concern involves the possibility of excluding an excessive number of samples, leading to a shortage of observations within the training, testing, or cross-validation datasets. This scarcity of data can create challenges in training models capable of effective generalisation, as emphasised in the previous point (EduPristine, 2016).
- **Missing values** - Small data sets pose challenges for imputing missing values due to limited representativeness, increased impact of outliers, higher risk of overfitting, limited feature relationships and reduced robustness of imputation methods (EduPristine, 2016).
- **Sampling Bias** - the difficulties posed by small data sets can be compounded if the data is biased and not selected at random from the population. This is frequently a problem in sociology studies because the test subjects are usually individuals in the same social circle or environment as the researcher, such as undergraduates at the university of the research (EduPristine, 2016).

Given the prevalence of problems with small data sets, various strategies have been proposed to address them. These approaches include, but are not limited to:

- **Data review** – in the case of small data sets, it is essential to spend time evaluating and managing the data to account for irregularities that can significantly

affect the accuracy of predictions. This includes identifying outliers, dealing with missing values and being aware of the implications of measurement errors (EduPristine, 2016).

- **Simpler models** - the lower the degree of freedom in proportion to the total number of training observations, the more robust the parameter estimates become. It is advisable to prioritise simpler models with fewer parameters to estimate, wherever possible (EduPristine, 2016).
- **No cross-validation data set** - this builds on the idea of using simpler models. Be cautious when using cross-validation data sets for tuning hyper-parameters. If the number of observations is small, it may be best to avoid using a cross-validation data set for training the model (EduPristine, 2016).
- **Regularisation** - is a technique used in machine learning to prevent overfitting of a model on the training data set. It adds a penalty term to the loss function of the model, which discourages the model from assigning too much importance to any one feature or having excessively large parameter values. The goal of regularisation is to encourage the model to find a simpler solution that can generalise well to new data. (EduPristine, 2016).

While these are some of the ways to address the challenges of working with small data sets, there is also an emerging solution inspired by human cognition that aims to give machine learning algorithms the ability to create new objects: generative modelling. This technique, which seeks to replicate the human ability to imagine and create, could provide a valuable tool for working with small data sets (Lamb, 2021).

Generative models not only offer ambitious possibilities, but they also have practical applications. According to some experts, generative models can be used to perform supervised and reinforcement learning tasks with less labelled data. For instance, when teaching a new language to children, explicit feedback is sporadic and reward signals are scarce, so cognitive scientists are intrigued by how humans can learn without overfitting. People have access to vast amounts of unlabelled data, such as sensory information, that can be used to train generative models without overfitting (Lamb, 2021). This is known as the "lack of stimuli" problem. Unsupervised generative models are used to construct robust representations of reality, which can then be used to engage in supervised and reinforcement learning with minimal labelled data.

Generative models have become increasingly popular in deep learning due to their

diverse applications. Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) are two of the most used techniques for generative modelling. Unfortunately, despite numerous publications on the use cases and modifications of these algorithms, the methods are unsuitable for small data sets precisely because they are "data hungry". That is, they require an excessive amount of data to train to accurately capture underlying patterns.

This paper proposes an alternative approach for generating new data points using a model based on Singular Value Decomposition (SVD). The suggested methodology is motivated by the capability of SVD to produce a lower rank estimation of a data matrix. By harnessing this capability, an alternative approach is proposed to generate new data points even for small data sets. This approach aims to expand the available data and ultimately mitigate the limitations posed by small sample sizes.

1.2 Motivation

The inspiration for this paper stems from a predicament encountered during the course of a separate research project whose objective was to propose an algorithm aimed at enhancing predictive performance when training on cytokine data. At the outset, the problem appeared insignificant. However, it soon became apparent that the number of data points was not only smaller but also lower than the number of variables being analysed. As a result, the data set was unable to train a model that could uniquely represent the relationship between the predictor and response variables. Additionally, due to the limited number of data points, splitting the data set into test and training sets was not ideal. Moreover, the classes of the response variable were not equally represented in the data set.

To address the data shortage problem, Synthetic Minority Oversampling Technique (SMOTE) was utilised to generate additional data points (Chawla et al., 2002). SMOTE is an advanced form of oversampling or a specialised data augmentation method that creates synthetic data points based on existing ones. Unlike traditional oversampling methods, SMOTE generates new data points that are slightly different from the original data points. Although SMOTE effectively balanced the classes of the response variable, it is sensitive to noise in the minority class. During the synthetic sample generation process, SMOTE amplified noisy or outlier instances, leading to a less effective oversampling strategy and adversely affecting the performance of the model (Chawla et al., 2002).

This thesis presents a novel method for creating new data points from a small data set using a model based on Singular Value Decomposition (SVD). The method is inspired by the ability of SVD to approximate a data matrix with a lower rank matrix. By exploiting this ability, we propose a new way to create new data points even when the data set is small. Our method aims to increase the amount of data and overcome the challenges caused by small sample sizes.

1.3 Research Objectives and Hypotheses

The purpose of this study is threefold. The first is to develop an algorithm based on SVD for generating new data from an original data set such that the new data exhibit the same distributional properties as the original data set. The second objective is to evaluate the effectiveness of the algorithm at generating new data sets from small data sets and the third and final objective is to compare the performance of regression and classification models on the new and original data.

The study aims to test the following hypotheses:

- H1: The data points generated from singular value decomposition (SVD) have the same distribution as the original data set.
- H2: The model trained on SVD-generated data set performs as well or marginally less than the model trained only on the original data set.

1.4 Performance Metrics

The study will use different ways to measure the performance of the model in light of the three hypotheses. These will include:

- looking at histograms that show the same variables in the original and the new data sets and seeing how similar they are;
- using boxplots to compare the distributions of the variables in the original and the new data sets;
- finding a correlation matrix for each of the original and the new data sets and then checking how much the correlation matrices differ in terms of the Frobinius norm;
- using the Kolmogorov-Smirnov test on each variable in the two data sets to see how alike or different they are; and
- comparing how well a regression or classification model trained on a mix of synthetic data points created by the SVD-based method and a smaller version

of the original data set performs against a model trained on the original data set. This comparison will show how good the SVD-based method is at making synthetic data that has the same important features as the original data set. It will also explore how adding these synthetic data affects the overall performance of regression or classification models.

1.5 Significance of Study

This study addresses a commonly encountered challenge by both researchers and practitioners in many fields - that of small samples which make it impossible to conduct rigorous statistical analyses. By examining how SVD can generate new data points, this study puts forth a potential solution that could simplify the process of augmenting the size of a data set; while still retaining the relationship between variables. As a result, this research paper has implications for various disciplines, as they can utilise the additional data points to facilitate their analyses.

The paper is divided into several distinct sections. In Section two, a review of the existing literature pertaining to the topic is provided. Section three offers an overview of the three data sets used in the research and includes an exploratory analysis of these data sets. The fourth section outlines and explains the experimental process in detail. Finally, Section five focuses on the practical application, evaluation and impact of SVD-generated data points on the prediction accuracy of the model.

2 Literature Review

In fields such as medicine, sociology, psychology and geology, amongst others, it is normal practise to work with relatively small data sets. This is because in these fields data collection is time-consuming, cost-prohibitive, or logistically challenging. The use of small data sets can lead to biased estimates, decreased statistical power and limited inferential scope, posing challenges in drawing reliable conclusions from the data. Consequently, researchers need to exercise caution when working with small data set and explore alternative approaches such as statistical models, data imputation, or aggregating data from multiple sources to enhance the validity and dependability of their findings.

2.1 How Researchers have Tackled the Small Data Problem

In their study, Tai Le Quy et al. (2020) explored the use of data augmentation methods to overcome the challenge of low-frequency signal sampling in Non-Intrusive Load Monitoring (NILM) due to insufficient sampling rates and incomplete records in real-world data sets. Non-Intrusive Load Monitoring (NILM) is a state-of-the-art technology to disaggregate and estimate the power consumption of individual appliances from the aggregated signal in households or companies.

To tackle the problem of low frequency signal sampling in NILM, Tai Le Quy et al. (2020) proposed various data augmentation techniques, such as Stepwise, Cubic Spline, Denton-Cholette and Device interpolation approaches, to increase the sampling rate of the data.

It was found that the Stepwise method was the best performing method and the Cubic Spline method was the least performing method, it generated overly smooth time series, which hindered incident recognition and affected the inferred power consumption of the appliances.

Stepwise interpolation generates augmented data between two time-stamps, t_1 and t_2, by dividing the time gap into n equal parts. For each part, a value is estimated using the formula:

$$\text{Interpolated value} = \text{Value at } t_1 + \frac{\text{Value at } t_2 - \text{Value at } t_1}{n} \times i$$

where i is the index of the segment. This means that for each segment between t_1 and t_2, it calculates the difference in data values at those time-stamps, divides that

difference by n and adds it to the value at t_1 to obtain values for the intermediate time points. This method allows for the creation of higher granularity data from lower sampling rate signals.

Furthermore, it was found that their data augmentation methodology is not reliant on the specific NILM technique employed, making it applicable to various other NILM techniques. This versatility stems from the fact that the technique was implemented at the data level, enabling its integration with different approaches.

Shaikhina et al. (2017) investigated the development of subject-specific models for predicting hip fractures in osteoarthritis (OA) and a generic method for applying regression Neural Networks (NNs) to small data sets. The researchers were motivated by the challenges of collecting patient data, which can be complex and costly and the small sample sizes typically seen in single-centre medical studies. The study investigated various Neural Network (NN) variants, including an ensemble NN and found that when dealing with small data sets, NN ensemble approach was inferior, in predicting hip fractures in osteoarthritis, to NNs constructed within a multiple runs framework. The study also discovered that a significant reduction in the size of the data set (by 18-fold) had only a minor impact (2.12% decrease) on the accuracy of the model. This trade-off should be considered in single-centre studies where data sets are typically small and Neural Networks are used.

In recent years, deep learning has emerged as a highly effective technique for several classification tasks. However, a major hurdle in utilising deep learning algorithms is the requirement for large amounts of training data. This issue is particularly evident in deep convolutional Neural Networks, which exhibit remarkable performance on large data sets such as ImageNet but tend to suffer from overfitting when trained on small data sets.

To address this challenge, a modified Deep Neural Network was introduced by Mengying Shu et al. (2019), which incorporated transfer learning to effectively train on small data sets while mitigating the risk of overfitting.

To evaluate the effectiveness of their approach, the researchers compared the performance of their modified model to a baseline model consisting of a 5-layer Convolutional Neural Network optimised by a binary classifier with no dropout or data augmentation. They discovered that their modified model achieved higher accuracy and was computationally more efficient than the baseline model.

In summary, the study underscores the potential of transfer learning and modified Deep Neural Networks in addressing the issue of image classification on small data sets.

The paper titled "Extreme Data Mining: Inference from Small Data sets" by Razvan Andonie in 2010 offers a comprehensive examination of computational solutions aimed at addressing challenges posed by small data sets. The paper explored the use of Artificial Neural Networks and fuzzy logic algorithms to extract information from small data sets, with the goal of addressing the limitations of traditional statistical approaches. The findings indicated that optimised Fuzzy Adaptive Resonance Theory Mapping (ARTMAP) (FAMR) models showed significantly better adaptability to the size of the training data compared to standard FAMR models. In this context, Standard FAMR refers to a type of Neural Network architecture used for classification and function approximation, which operates without any optimisations applied to the training data or relevance factors. Additionally, among the three tested FAMR models, the Genetic Algorithm (GA)-FAMR model demonstrated the highest level of fit to the training data. Overall, the paper highlights the potential of these algorithms to extract meaningful insights from small data sets. The paper noted that while these models can be effective, they can also incur significant computational overhead, making them less scalable for larger data sets. The paper concluded by suggesting that further research is needed to explore the full potential of these techniques, particularly in the context of real-world applications where small data sets are common.

Kamath et al. (2018) proposed a novel method to address regression problems using small data sets. According to Kamath, high-quality surrogates can be constructed for small data sets by carefully selecting input samples and employing suitable regression techniques. The study focused on additive manufacturing (AM), also known as 3-D printing, which involves building objects layer by layer. In AM, understanding the impact of the control variables on the properties and quality of the final product is challenging and expensive and thus surrogate models are utilised. Surrogate modelling is a technique that constructs statistical models to replicate simulation outcomes, allowing researchers to bypass the issue of small data sets. Kamath et al. (2018) considered five different surrogate models for this task: Locally Weighted Kernel Regression (LWKR), Regression trees, Multivariate Adaptive Regression Splines (MARS), Support Vector Regression (SVR) and Gaussian processes (GP). The study found that weighted distances in the input space significantly improved nearest-neighbor methods and generated effective surrogate models.

The results indicated that the three remaining techniques - LWKR with optimised weights, MARS and GP - exhibited satisfactory performance. However, SVR did not perform well on smaller data sets with considerable volatility and limited sample points to capture the variation. GP was found to outperform the other methods in terms of predictive accuracy and functionality. Additionally, the study emphasised that surrogates yield optimal results when applied to continuous and smoothly varying functions. However, they may struggle to accurately identify discontinuities and steep gradients in the function when working with small samples.

This section offered an overview of the strategies employed by researchers across various fields to tackle the challenge of small data. While some researchers have concentrated on improving existing techniques, others have ventured into exploring novel models capable of effectively handling small data sets. Furthermore, the generation of new data points has emerged as an area of significant interest, as it has the potential to supplement existing data and enhance the development of more accurate models.

2.2 Recent Developments and the Application of Variational Autoencoders and Generative Adversarial Network

Generative modelling is a powerful branch of machine learning that focuses on the creation and understanding of data. Unlike discriminative models that aim to classify or predict existing data, generative models learn the underlying distribution of the data and can generate new samples that closely resemble the original data distribution. Generative models find applications in various fields, including computer vision, natural language processing and data synthesis. They enable tasks such as image synthesis, text generation, anomaly detection and data augmentation. Generative modelling techniques have witnessed significant advancements in recent years. The two most prevalent generative deep learning architectures are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

Wei et al. (2020) conducted a comprehensive analysis of the performance of many variants of VAE at the data generation task. The results showed that no single VAE design may be deemed as the best and that performance depends on VAE type, capacity of network, network size and batch size, suggesting that a well-designed architecture is required for maximum VAE performance.

GANs have been successfully applied to the task of generating realistic images of

human faces, including those that do not correspond to any specific individual. In 2017, Karras et al. introduced the "Progressive Growing" method as a novel approach for training GANs. This technique allows for the generation of high-quality images with enhanced stability and increased variation. Through experimentation and comparison with other state-of-the-art methods, the authors demonstrated the effectiveness of their approach. In 2018, Brock et al. proposed a new training method that utilised large-scale parallelisation to improve the efficiency of GAN training for generating photorealistic and high-quality images. Their method surpassed previous state-of-the-art approaches in terms of visual precision and diversity.

Academic studies have investigated various techniques to address the challenges associated with small data sets, including data augmentation, transfer learning and modified neural networks. These methods have demonstrated promising outcomes, yet there remains a need for innovative approaches to effectively solve this issue. Although generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have been extensively applied in domains like medicine, natural language processing and computer vision, their usage has predominantly targeted visual data such as the MNIST and molecular data sets. The application of VAEs and GANs to small data sets is still a relatively underexplored area within machine learning. This gap exists because generative models typically demand a large amount of data to accurately learn underlying patterns. As a result, small data sets present a significant obstacle in achieving optimal results with these models. To address the constraints of applying VAEs and GANs to small data sets, it is crucial to explore alternative generative techniques that can operate effectively on small data set.

3 Overview of Data

This chapter presents a detailed and thorough examination of the data sets employed in the study. This includes information about the sources of the data, their specific characteristics, as well as the preprocessing steps undertaken.

3.1 Data Sets used for Analysis

In this paper, three data sets will be employed namely: Abalone Age, Life Expectancy (WHO) and NBA Rookie data set. All three data sets were sourced from Kaggle.

3.1.1 Abalone Age

Abalone, scientifically recognised as a sea snail, holds significant commercial importance due to its substantial export value. In the Afrikaans language, it is referred to as 'perlemoen'. This marine organism is exclusively endemic to the rocky habitats found in the shallow coastal waters of South Africa. The term 'perlemoen' originates from the Dutch expression 'Paarlemoer,' signifying 'mother of pearl,' which alludes to the glowing hues and texture exhibited on the internal surface of the abalone shell. The determination of the age of an abalone involves a laborious and time-consuming process of piercing the shell through the cone, staining it and counting the number of rings under a microscope. As an alternative, simpler measures may be utilised to predict the age, but this requires supplementary information, such as weather conditions and geographical location, to accurately address the issue. In this study, the Abalone data set will consist of the following variables:

Table 1: Description of the Abalone Data Set

Name	**Data Type**	**Measurement Unit**	**Description**
Sex	Nominal	–	M, F and I (Infant)
Length	Ratio	mm	Longest shell measurement
Diameter	Ratio	mm	Perpendicular to length
Height	Ratio	mm	With meat in shell
Whole Weight	Ratio	grams	Whole abalone
Shucked Weight	Ratio	grams	Weight of meat
Viscera Weight	Ratio	grams	Gut weight (after bleeding)
Shell Weight	Ratio	grams	After being dried
Rings	Ratio	–	Plus 1.5 gives the age in years

The data set contains 4177 observations, each having 9 variables. Among these

variables, the variable "Rings" or "Age" serves as the response variable, while the other eight variables act as predictors. The "Rings" or "Age" of the abalone will be predicted based on physical characteristics of the snail.

3.1.2 Life Expectancy (WHO)

The Life Expectancy (WHO) data set is a collection of health-related data for 193 countries, with a focus on life expectancy and other health-related factors. The data set was compiled by merging multiple data sets from 2000 to 2015. The selection of relevant key variables from all categories of health-related factors was a crucial step in the data collection process. In addition to health-related factors, economic information was also incorporated in the data set. The objective of the data set was to identify patterns in health-related factors and economic indicators that contribute to changes in life expectancy rates over time. However, upon visual inspection of the data set, missing values were identified, particularly in population, Hepatitis B and GDP indicators. Several countries, such as Vanuatu, Tonga, Togo and Cape Verde, among others, had incomplete data and it was impossible to obtain complete data for these nations. As a result, these countries were excluded from the final model data set. Nonetheless, the remaining observations provide valuable insights into the relationship between various health-, economic-related factors and life expectancy rates across different countries over the 15-year period. Table 2 offers a comprehensive description of the data set, outlining the selected health-related and economic indicators for each country included in the study.

Table 2: Description of the Life Expectancy Data Set

Name	**Data Type**	**Measurement Unit**	**Description**
Country	Nominal	–	Names of the country
Year	Interval	–	Year of recording
Status	Nominal	–	Developed or developing
Life Expectancy	Ratio	years	Age
Adult Mortality	Ratio	–	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant Deaths	Ratio	–	Number of Infant Deaths per 1000 population
Alcohol	Ratio	–	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Percentage Expenditure	Ratio	%	Expenditure on health as a percentage of Gross Domestic Product per capita

Hepatitis B	Ratio	%	Hepatitis B (HepB) immunisation coverage among 1-year-olds
Measles	Ratio	–	Measles - number of reported cases per 1000 population
BMI	Ratio	kg/m^2	Average Body Mass Index of entire population
Under-five deaths	Ratio	–	Number of under-five deaths per 1000 population
Polio	Ratio	%	Polio (Pol3) immunisation coverage among 1-year-olds
Total Expenditure	Float	%	General government expenditure on health as a percentage of total government expenditure
Diphtheria	Ratio	%	Diphtheria tetanus toxoid and pertussis (DTP3) immunisation coverage among 1-year-olds
HIV/AIDS	Ratio	–	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Ratio	usd	Gross Domestic Product per capita
Population	Ratio	–	Population of the country
Thinness 10-19 years	Ratio	%	Prevalence of thinness among children and adolescents for Age 10 to 19
Thinness 5-9 years	Ratio	%	Prevalence of thinness among children for 5 to 9
Income Composition of Resources	Ratio	–	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Ratio	years	Number of years of Schooling

The data set resulting from the study comprises 2938 observations and 22 variables, with 20 predictor variables and a single dependent variable, namely "Life expectancy". The "Country" variable was excluded from the analysis due to its high cardinality. Transforming country names into a numerical format typically involves one-hot encoding, which significantly increases the dimensionality of the feature space. This heightened complexity can lead to overfitting, thereby reducing the generalisation performance and computational efficiency of the model. Consequently, the exclusion aimed to enhance model robustness.

3.1.3 NBA Rookie

The National Basketball Association (NBA) is widely regarded as the leading professional basketball league in the world. As a result, there is a great deal of competitiveness for entry into the NBA. Only around one percent of college basketball players in the National Collegiate Athletics Association (NCAA) get selected for the NBA Draft. Newly drafted players in the league are expected to consistently showcase their skills and performance on the basketball court in order to secure their spot and remain on the team and consequently the league. The data is a composite of two data sets: one containing rookies (recruits) and the other containing active players. The rookies data set contains information on all of the players that were drafted as rookies between the years 1980 and 2015. The active players data set contains a list of active players for each season from 1980 to 2017, starting with the first season.

Table 3: Description of the NBA Rookie Data Set

Name	**Data Type**	**Measurement Unit**	**Description**
Name	Nominal	–	Name of the player
GP	Ratio	–	Games played
MIN	Ratio	–	Minutes played
PTS	Ratio	–	Points per game
FGM	Ratio	–	Field goals made
FGA	Ratio	–	Field goal attempts
FG	Ratio	%	Field goal percent
3P Made	Ratio	–	3-point made
3PA	Ratio	–	3-point attempts
3P	Ratio	%	3-point percent
FTM	Ratio	–	Free throw made
FTA	Ratio	–	Free throw attempts
FT	Ratio	%	Free throw percent
OREB	Ratio	–	Offensive rebounds
DREB	Ratio	–	Defensive rebounds
REB	Ratio	–	Rebounds
AST	Ratio	–	Assists
STL	Ratio	–	Steals
BLK	Ratio	–	Blocks
TOV	Ratio	–	Turnovers
Target_5Yrs	Nominal	–	Outcome: 1 if career length >= 5 yrs, 0 otherwise

The NBA Rookie data set contains 21 columns and 1340 rows. The "TARGET 5Yrs"

column is the dependent variable, while the remaining columns serve as predictor variables. This data set will be utilised to predict whether a player will remain in the league for five or more years, based on their performance statistics.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a fundamental step in the data analysis process that aims to uncover patterns, trends and insights from the available data. It involves examining and visualising the data to gain a deeper understanding of its characteristics, relationships and potential outliers. EDA typically involves a combination of statistical techniques, data visualisation and data manipulation. Descriptive statistics, such as measures of central tendency and dispersion, help summarise the overall characteristics of the numerical variables in the data sets. Data visualisation techniques, including histograms, scatter plots and box plots, provide visual representations of the data distribution and relationships between numerical variables, as these are not applicable to all data types. The main objectives of exploratory data analysis are to summarise the main variables of the data set, detect any data quality issues or missing values, explore the distribution and variability of variables, identify potential patterns or correlations and generate initial hypotheses for further analysis (Wongsuphasawat et al., 2019).

3.2.1 Abalone Age

As mentioned earlier, the Abalone Age data set comprises of 4177 observations and 9 variables. Prior to establishing relationships between the variables, data cleaning is crucial. This process involves identifying and addressing abnormal data, missing values and outliers. In the Abalone age data set, there were no missing values and outliers that were detected were removed from the data set. Apart from outliers, the other anomaly observed in the data set was the presence of numerical variables measured in different units. To ensure consistency among the variables, the data set was scaled, bringing all the variables within the same range of between 0 and 1. By scaling the variables, the focus is shifted towards their relationship with the response variable rather than their individual magnitudes.

To detect and remove outliers, the Interquartile Range (IQR) Method was employed. The IQR itself does not directly detect outliers; rather, it is used to establish fences, with values falling beyond these fences being flagged as outliers. It is important to note its drawbacks, which include, but are not limited to, insensitivity to distribution shape, reliance on fixed thresholds, limited scope and simplicity that may not always capture complex outlier patterns. Understanding these limitations is crucial for using

the method correctly. It is important to note that removing data points is generally discouraged because it can result in the loss of valid observations and unfairly reduce their influence on the analysis, possibly leading to biased or inaccurate outcomes. To ensure robust and reliable insights, it is important to retain as much data as possible. However, for the purposes of this analysis, this requirement is relaxed. Furthermore, the following are some notable characteristics found in the data set:

- Table 4 shows a summary of the basic statistics for each of the variables in the Abalone age data derived from IQR method:

Table 4: Eight Number Summary for Abalone Data Set

	Count	Mean	Std	Min	25%	50%	75%	Max
Length	3781	0.521	0.112	0.205	0.450	0.535	0.610	0.760
Diameter	3781	0.405	0.092	0.155	0.345	0.420	0.475	0.600
Height	3781	0.137	0.035	0.040	0.110	0.140	0.165	0.240
Whole Weight	3781	0.792	0.445	0.043	0.433	0.767	1.118	2.128
Shucked Weight	3781	0.347	0.204	0.017	0.181	0.327	0.492	0.960
Viscera Weight	3781	0.174	0.101	0.001	0.091	0.164	0.244	0.492
Shell Weight	3781	0.226	0.123	0.013	0.125	0.220	0.315	0.625
Age	3781	10.931	2.330	5.500	9.500	10.500	12.500	16.500

  Based on Table 4 the data set exhibits notable characteristics with regard to the distribution of its variables. Specifically, the median of each variable lies within one standard deviation of its mean, indicating that the distribution of the variables approximates a symmetrical frequency curve. However, if the mean is greater than the median, the variable is positively skewed and if the mean is less than the median, the variable is negatively skewed. It is important to highlight that the mean is not a robust statistic, meaning that the presence of an outlier can have a substantial influence on its value. In contrast, the median is a robust statistic and the presence of an outliers do not have a significant impact on its value.

- Figure 1, shows distributions of the variables in the Abalone age data set in terms of histograms. The histograms of the variables show that height and age approximately follow a normal distribution, although they are slightly skewed

to the right or left. These distributions have a bell-shaped curve, with most data points concentrated around the center, featuring a single peak and extending tails. In contrast, the other variables do not closely resemble a normal distribution.

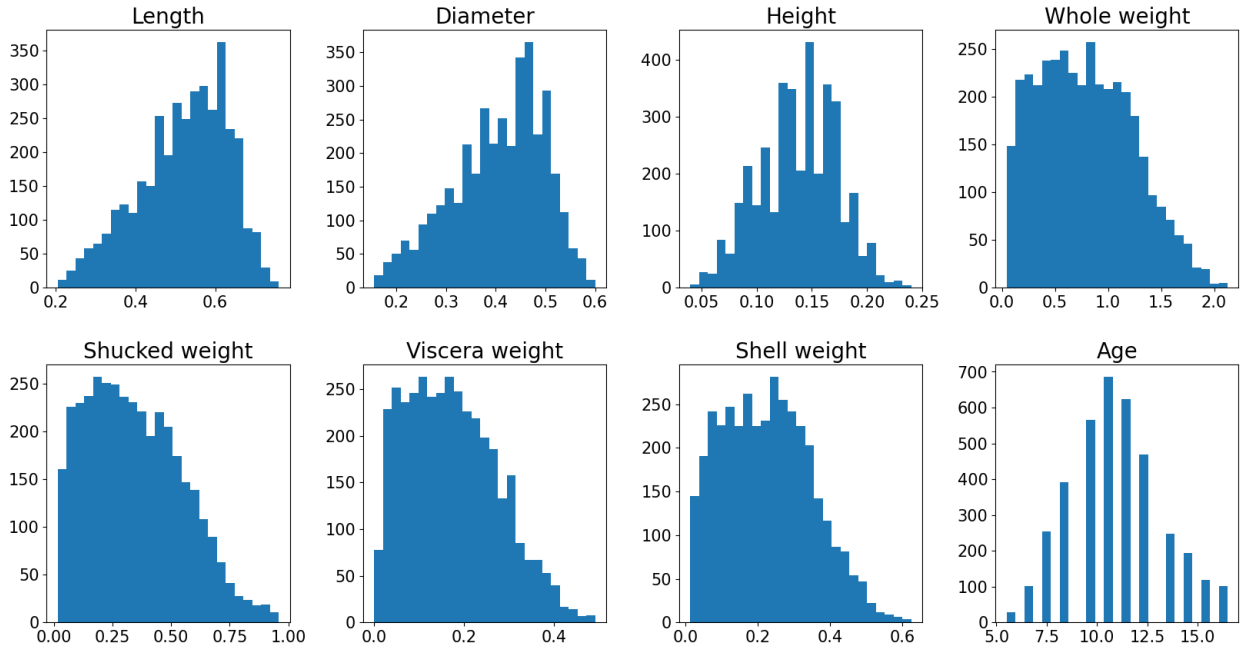


Figure 1: Histogram of the Variables of the Abalone Data Set

However, after removing outliers, shucked weight, viscera weight, whole weight and shell weight exhibited the highest degree of skewness. The skewness coefficient provides a measure of the asymmetry in a distribution. The coefficient quantifies how much a distribution deviates from symmetry. Positive values indicates that the distribution has a longer or fatter tail on the right side, whereas negative values indicates a longer or fatter tail on the left side. While it does not directly measure the "heaviness" of the tails, it is related to the distribution's asymmetry, which can be influenced by heavy tails. Compared to other variables, the age and height variables are relatively closer to normality, as indicated in the Table 5.

Table 5: Skewness of the Variables of the Abalone Data Set

Variables	**Skewness Degree**
Shucked weight	0.455
Viscera weight	0.455
Whole weight	0.345
Shell weight	0.342
Age	0.269
Height	-0.153
Diameter	-0.482
Length	-0.495

- It was observed that the number of male abalone entries was higher than the number of infant and female abalone entries. This suggests that the data set is skewed towards males, but the difference in the number of observations across sexes does not indicate an imbalanced data set (as exhibited in Figure 2).

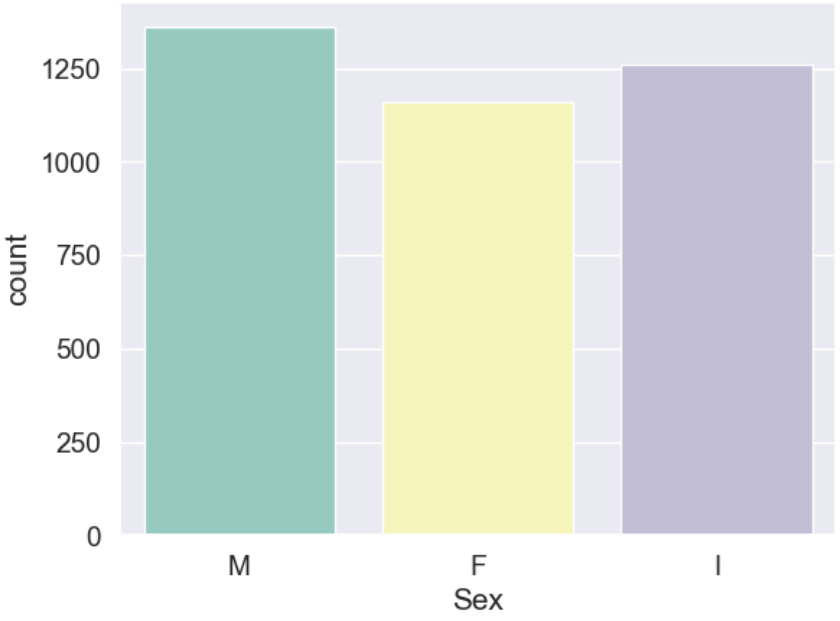


Figure 2: Distribution of the Sex in the Abalone Data Set

The data set shows that a significant portion of male, female and infant abalone entries are concentrated within specific age ranges. More specifically, the majority of male abalone fall within the age range of 10.5 to 12.5 years, whereas female abalone are predominantly found in the age range of 10.5 to 14 years. Infant abalone, on the other hand, are mostly in the age range of 7 to less than

10 years. The observed pattern indicates that, on average, female abalone are older than male abalone and male abalone are older than infant abalone (average age of 11.9, 11.6 and 9.4 respectively). This finding is reinforced by calculating the average ages for each category of sex. It should be noted that while infant Abalone can be male or female, the "infant" category in the sex variable pertains to age rather than gender. These are Abalones whose age cannot be determined simply by counting the rings and adding 1.5; other characteristics are needed to determine their age.

Figure 3: Distribution of the Sex vs Age in the Abalone Data Set

Additionally, female abalone exhibit larger values of length, diameter, height and overall weight than those of male or infants abalone.

- Upon analysing the data using bivariate analysis (see Figure 4), it can be detected that all the variables exhibit a positive correlation with one another, which is further supported by the correlation matrix in Figure 5. However, not all the relationships between the variables are linear in nature. Hence, utilising a linear algorithm to model the data would not lead to an optimal solution for predictive modelling purposes. It is necessary to consider other non-linear techniques to construct a model that captures the complex relationships between the variables and provides accurate predictions.
- Within the data set, several variables, including Whole Weight, Diameter, Length, Sex, Shucked Weight, Shell Weight and Viscera Weight exhibit high values of Variance Inflation Factor (VIF), indicating the existence of multi-collinearity (see Table 18 in Appendix A.1.1). Multi-collinearity occurs when

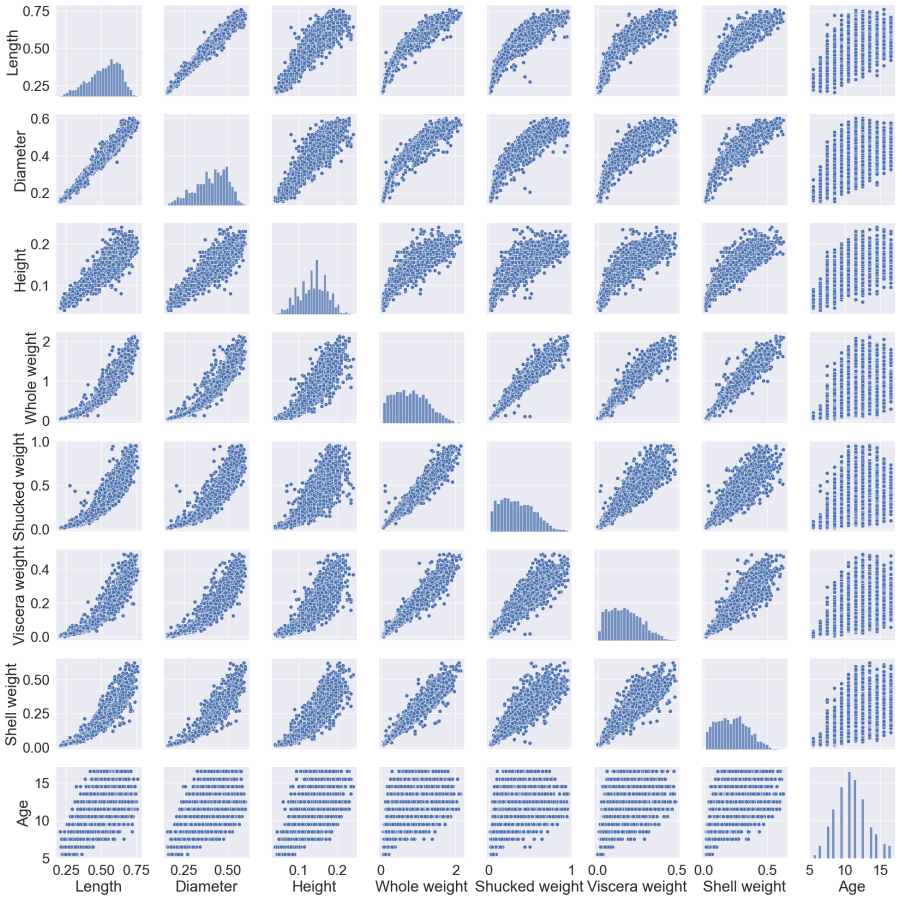


Figure 4: Bivariate Analysis of the Abalone Data Set

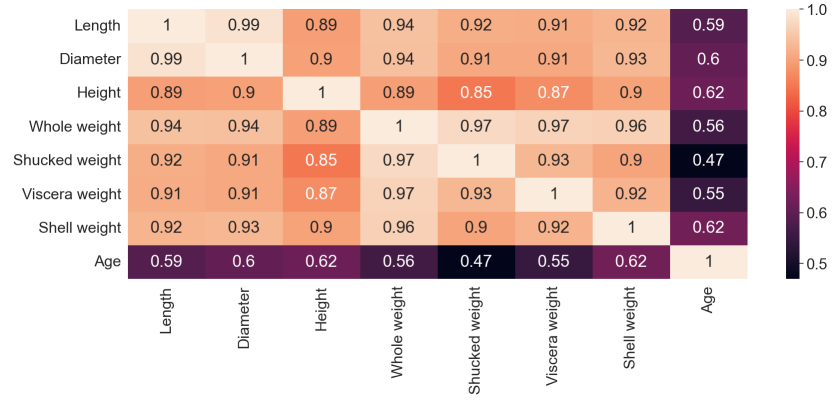


Figure 5: Correlation Heatmap of the Abalone Data Set

two or more predictor variables are highly correlated, leading to redundancy in the model. In such cases, it becomes difficult to isolate the effect of individual predictor variables on the response variable. Therefore, it is necessary to remove a subset of such variables from the model or consider alternative methods, such as regularisation, to reduce the impact of multi-collinearity on the performance of the model.

3.2.2 Life Expectancy (WHO)

This data set initially included 2938 observations and 22 variables. Prior to analysing the relationships between variables, the data was cleaned. The combination of the "Country" and "Year" variables provides a unique identifier for each entry, but for the purpose of this study, which aims to determine life expectancy regardless of origin or time, these columns were removed from analysis. Several variables, including life expectancy, adult mortality, alcohol, hepatitis B, BMI, polio, total expenditure, diphtheria, GDP, population, thinness 1-19 years, thinness 5-9 years, income composition of resources and schooling contain missing values that are not random, but rather due to non-response from participants. Imputation of these missing values may introduce biases to the study and therefore, records with missing values have been omitted. Outliers were detected and eliminated using the interquartile range (IQR) method. The decision to remove outliers was justified based on the same rationale as explained in the preceding subsection. The only irregularity in the data set was that the values were measured using different units and were therefore scaled in the range between 0 and 1. Points worth noting about the data set are:

- The table in Appendix Section A.2.1 shows a summary of the basic statistics for each of the variables in the Life Expectancy data derived from IQR method.

  In terms of the variables, the median of all variables is within one standard deviation of the mean, implying that the distribution of the variables is nearly symmetrical. However, a few variables, such as Alcohol, Adult Mortality, Total expenditure and Life expectancy, show significant differences between the minimum and 25th percentile value or 75th percentile and the maximum value, indicating that these variables are highly skewed and do not adhere to the properties of a normal distribution. Conversely, other variables display only slight differences between these metrics, suggesting a moderate degree of skewness.

- The distributions of the variables can be visualised using histograms, which show the frequency distribution of the data (see Figure 6). The histogram show that only one variable (specifically Schooling) exhibits a distribution that is close to a normal distribution, with a fairly symmetrical bell-shaped

curve. However, the rest of the variables have skewed distributions with Measles, Under-five deaths, Infant deaths, HIV/AIDS, Population, Thinness 10-19 years, Thinness 5-9 years, Percentage expenditure and GDP in that order exhibiting the highest degree of skewness (Refer to the Table 20 in Appendix A.2.2).

This asymmetry in the distribution can be problematic when analysing data, as it can affect the accuracy of statistical models and make it difficult to draw meaningful conclusions from the data.

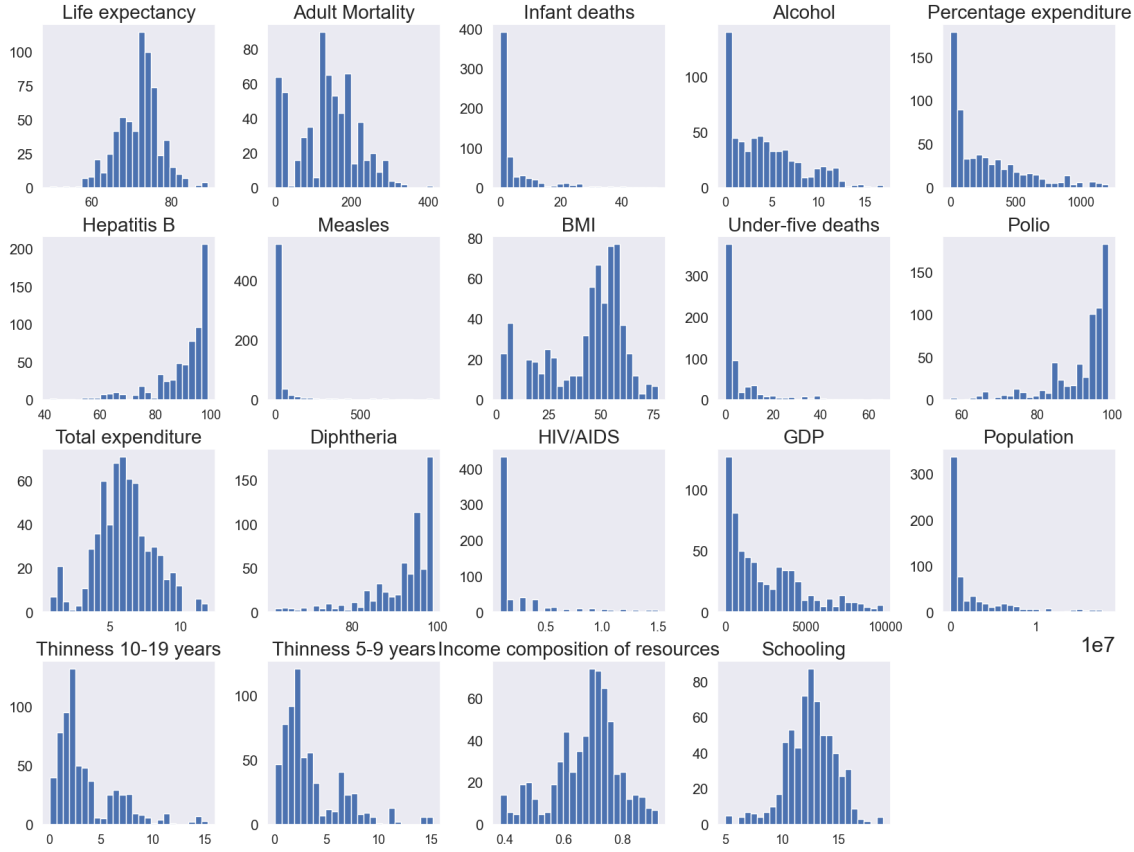


Figure 6: Histogram of the Variables of the Life Expectancy Data Set

- Referring to Figure 7, which presents the count of observations (representing countries) categorised into developed and developing statuses. The dataset contains more information from developing countries than from developed ones. This difference shows that the variable is not evenly distributed. However, this difference accurately represents the distribution of the world's population, as most people live in developing countries. Nonetheless, when examining the data, it is important to remember this unbalance, as it could affect the results and conclusions of the study.

Figure 7: Distribution of the Status Variable

- Figure 8 shows, that Developing countries have consistently reported a higher adult mortality rate than Developed countries throughout the recorded time. This trend is in line with the common perception that developing countries face greater health challenges and have weaker healthcare systems than their developed counterparts. The data indicates that the gap in adult mortality rates between the two types of countries varies over time.

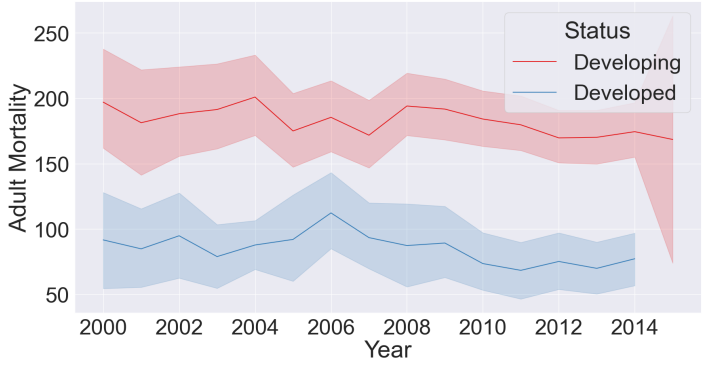


Figure 8: Adult Mortality Rate vs Status over Time

As anticipated (see Figure 9), individuals in Developed countries have a greater Life expectancy than those in Developing countries. Specifically, a large proportion of the Life expectancy in Developed countries falls within the range of 75 to 83 years, while for Developing countries, it falls within 70 to 76 years.

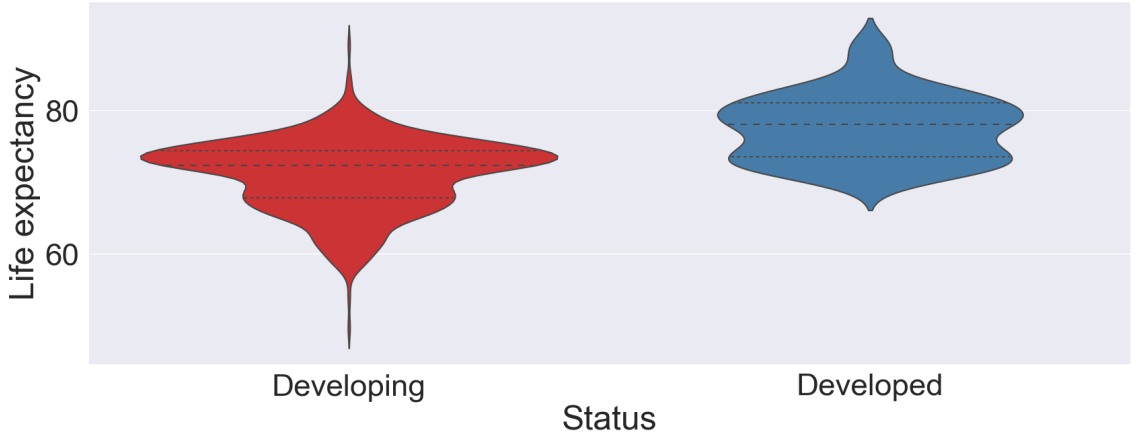


Figure 9: Life Expectancy vs Status

Similarly, developed countries exhibit lower rates of infant deaths, under-five deaths, prevalence of measles per 1000 population, thinness among children and adolescents and adult mortality rate for both sexes than Developing countries. Furthermore, Developed countries demonstrate higher average mass index for the entire population, expenditure on health as a percentage of Gross Domestic Product (GDP) per capita, alcohol consumption per capita, General government expenditure on health as a percentage of total government expenditure, years of schooling completed by the population, Hepatitis B (HepB) immunisation coverage among 1-year-olds, Polio (Pol3) immunisation coverage among 1-year-olds and Diphtheria tetanus toxoid and pertussis (DTP3) immunisation coverage among 1-year-olds than Developing countries.

- Upon conducting bivariate analysis on the data set, it is evident from the correlation matrix in Figure 10 that some variables exhibit positive, negative or no relationship with each other. Specifically, the variables: Alcohol & Life expectancy; percentage expenditure & Life expectancy; BMI & Life expectancy; Income composition of resources & Life expectancy; Schooling & Life expectancy exhibit a positive correlation, ranging from moderate to strong. whilst the variables: thinness 5-9 years & Life expectancy; thinness 1-19 years Life expectancy; HIV/AIDS & Life expectancy; Adult Mortality & Life expectancy exhibit a negative correlation, ranging from moderate to strong. The finding that appears counterintuitive is the positive relationship

between alcohol consumption and life expectancy, despite the well-established evidence linking excessive alcohol intake to various adverse health outcomes and a potential reduction in life expectancy. Moreover, some of the relationships between life expectancy and the other variables do not follow a linear pattern, as indicated by the first row or column of the plot present in the Appendix section A.2.3. In other words, the strength and direction of the relationship between life expectancy and other variables are not constant throughout the range of values of the variables.

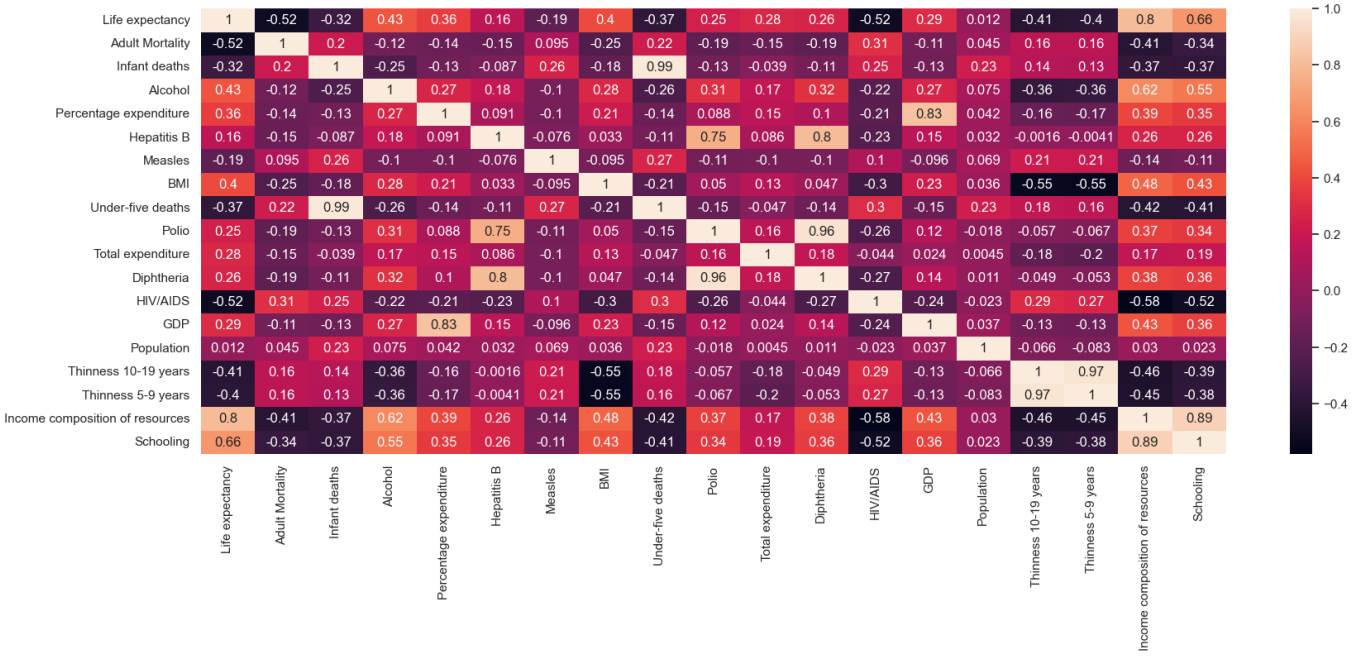


Figure 10: Correlation of the Life Expectancy Data set

- Variables such as Status, Under-five deaths, Infant deaths, Thinness 10-19 years, Thinness 5-9 years, Diphtheria, Polio, Income composition of resources and Schooling exhibit notably high values of variance inflation factor (VIF), suggesting a high degree of correlation among them (see Table 21 in Appendix A.2.4). As a result, including these variables in a model would introduce significant multicollinearity.

3.2.3 NBA Rookie

The initial data set comprised 1340 observations and 21 variables. Prior to analysing any relationships, data cleaning and manipulation were performed, following a similar approach as used on the other data sets. One of the variables, namely "name,"

exhibited high cardinality as it predominantly consisted of unique values (96% of values in the column were unique). Dealing with high cardinality variables, such as "name," posed challenges with one-hot encoding, primarily in terms of space usage and the curse of dimensionality. The curse of dimensionality refers to the exponential growth in data requirements to accurately distinguish between variables and generalise the model as the number of variables increases (Jain, 2020). Consequently, the "name" variable was excluded from the analysis. The variable "3-Point %" was the only one in the data set with missing values. Consistent with previous data sets, observations with missing values were removed. Outliers were eliminated using the Interquartile Range (IQR) method. As for the other two data sets, the values of the variables were measured using different units. Hence, the data set was scaled or normalised. The following characteristics of the data were noted:

- Based on summary statistics of the NBA data set in Appendix Section A.3.1,

  Each variables in the data set exhibits a median that is within one standard deviation of its mean, suggesting that the distributions of the variables are approximately symmetrical. However, there exist significant differences between the minimum and 25th percentile, or 75th percentile and maximum values for most variables, excluding Minutes Played, Field Goal %, 3 Points %, Free Throw %, target 5yrs and Games Played. These variables with considerable gaps between the minimum and 25th percentile or 75th percentile and maximum values exhibit a high degree of skewness and do not conform to the typical characteristics of a normal distribution.

- The histogram analysis of the variables, in Figure 11, reveals that only a few variables, such as Free Throws % and Field Goals %, exhibit a relatively close approximation to a normal distribution.These distributions have a bell-shaped curve, with most data points concentrated around the center and featuring extending tails. However, the majority of the remaining variables display slight skewness except for the variables: 3-Points Made (3pm), Blocks (blk), 3-Points Attempts (3pa), Assists (ast), Field Goals Attempts (fgm), Field Goals Made (fgm) and Free Throw Made which exhibit the most significant skewness as shown in the Appendix section A.3.2.

- Based on the data set, it can be observed from Figure 12 that a higher proportion of basketball players have had a career lasting five or more years compared to those who had a career of less than five years. This finding suggests that a significant number of players have accumulated substantial experience and longevity in their professional basketball careers.

- Consistent with expectations, Figure 13, illustrates that athletes who had a career lasting longer than five years tend to have a higher average points per

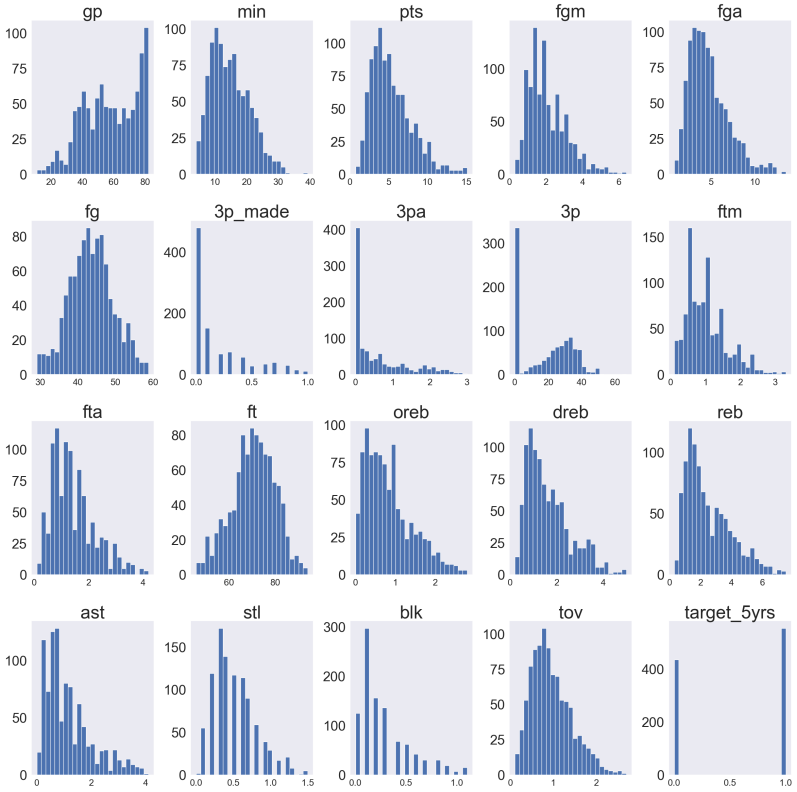


Figure 11: Histogram of the Variables of the NBA Data Set

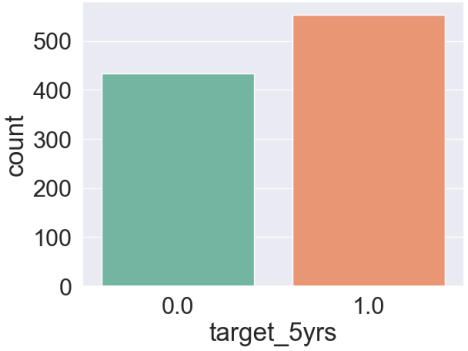


Figure 12: Distribution of the Target Variable

game compared to those with a career duration of less than five years.

The same can be said about number of minutes played. Figure 14 shows that those players whose careers were longer than 5 years played more minutes on average in games than those players with careers which lasted less than 5 years.

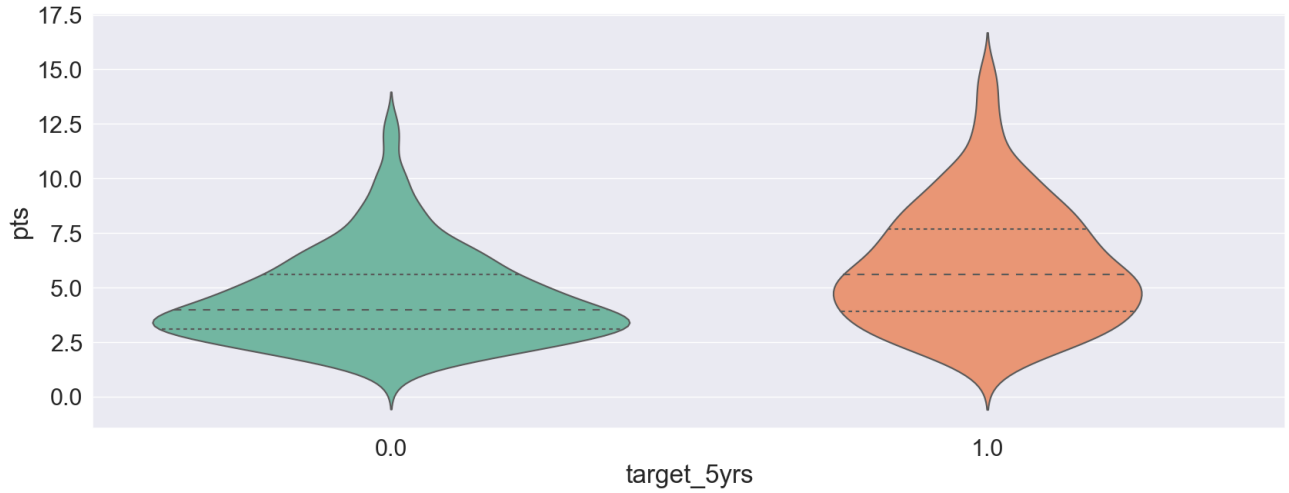


Figure 13: Target Variable vs Points

Figure 14: Target Variable vs Minutes Played

The analysis indicates that, in general, players who had a career lasting more than five years have better performance in most statistics compared to those with a shorter career, indicating a positive relationship between career length and overall performance. However, there is one exception to this trend: in the case of 3-point shots made, 3-point attempts and 3-point percentage, players with shorter careers have higher values than those with longer careers.

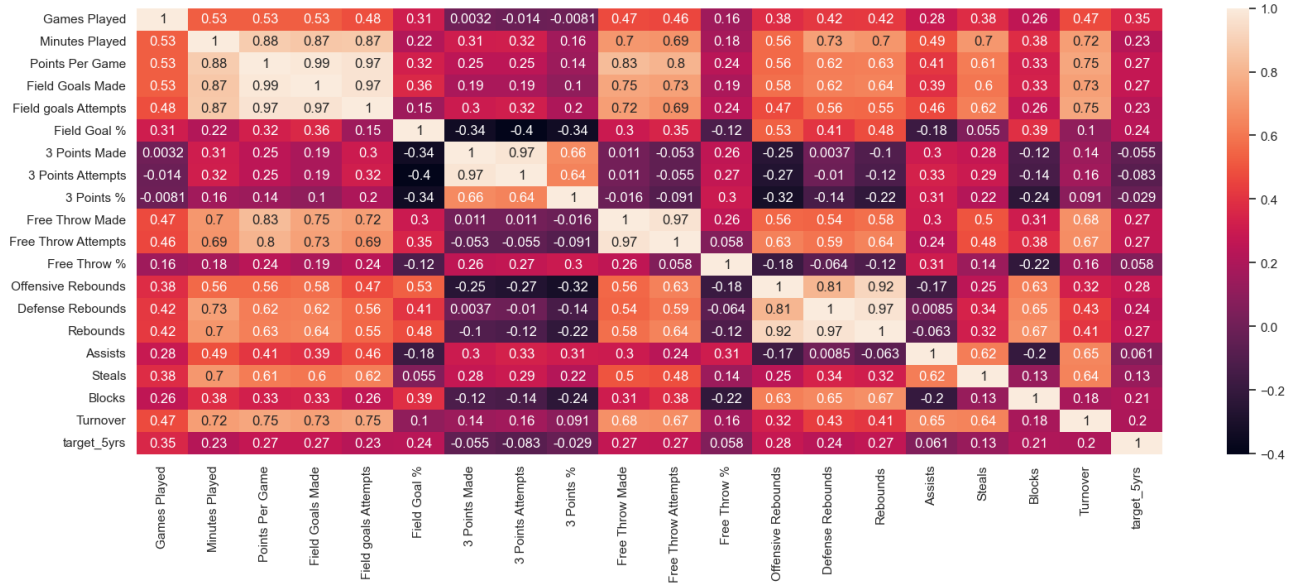


Figure 15: Correlation Heatmap of the NBA Data Set

The correlation matrix in Figure 15 shows that , there is a positive correlation between the number of minutes played by a player and their performance in various statistics such as scores, field goal attempts, free throw attempts, rebounds, steals, blocks, turnovers and assists. This finding is consistent with prior expectations. In general, there are positive correlations between most variables, with a few exceptions. Furthermore, the last row or column of the plot in Appendix Section A.3.3 shows that the relationship between the response variable and the other variables is not linear. This means that there may be other factors at play that affect the response variable in a non-linear manner and need to be taken into consideration when interpreting the data.

The exploratory data analysis (EDA) provided a robust foundation for understanding the various structures within the three datasets. By meticulously cleaning, transforming and interpreting the data, we were able to uncover critical insights into the variables' distributions, correlations and associations, which ranged from linear to non-linear and were measured on different scales. The EDA highlighted the diverse dimensions and characteristic of each data set, establishing a solid guideline that will be instrumental in addressing the research objectives and hypotheses. The resulting dimensions of the three data sets were as follows: The Abalone Age data set contained 3781 observations and 11 variables; the Life Expectancy data set comprised 646 observations and 21 variables; and the NBA data set encompassed

987 observations and 20 variables.The next section will delve into the derivation of method to be used to generate new data points from these small samples, leveraging the foundational insights gained from the EDA.

4 Singular Value Decomposition

Singular Value Decomposition (SVD) is a matrix factorisation method that is widely utilised in diverse disciplines, including linear algebra, signal processing, data analysis and computer vision. This method involves breaking down a rectangular matrix into three matrices, which helps reveal important information about the structure and characteristics of the original matrix. The conventional way to represent singular value decomposition (SVD) is as follows: any matrix $\mathbf{M} \in R^{m \times n}$ can be decomposed into three matrices,

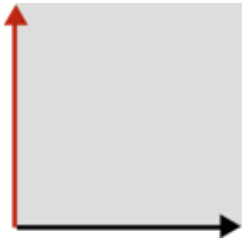
$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}, \tag{1}$$

where $\mathbf{U}$ is an $m \times m$ unitary matrix, $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix and $\mathbf{V}$ is an $n \times n$ unitary matrix. $\mathbf{V}^T$ is a transpose of $\mathbf{V}$.

In order to provide a comprehensive understanding of the subject matter, Gregory Gundersen (2018) takes a step-by-step approach to present Equation 1, starting from fundamental principles instead of directly providing its final form. Gregory begins by explaining Singular Value Decomposition (SVD) in a clear manner, avoiding complex terminologies or jargon. This gradual progression then leads to a more formal expression of the concept, ultimately arriving at the precise definition embodied in Equation 1 (Gundersen, 2018).

4.1 SVD without jargon

Suppose there is a square. Its orientation can be represented graphically by arrows, similar to a hand forming an "⌊" (as shown in Figure 16).

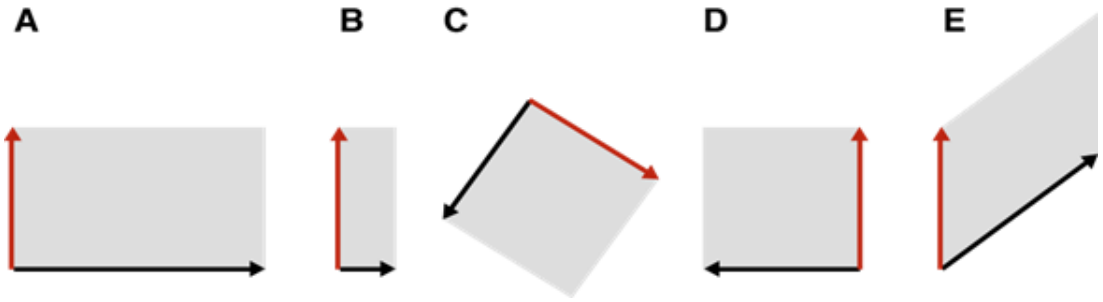


Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 16: Square with Orienting Arrows

There are various ways to manipulate the square. For instance, one can apply a force to one of its edges by pushing or pulling, causing it to expand or contract (see Figure 17A and 17B). The square can also be rotated (as demonstrated in Figure

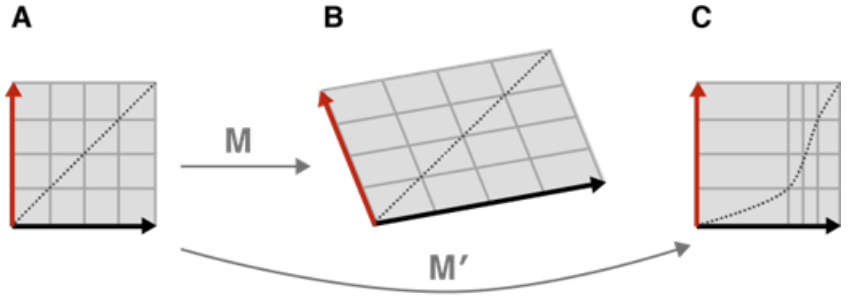
17C) or flipped to change its orientation, similar to the transformation of an "⌊" shape to a "⌋" shape by rotating a hand (as shown in Figure 17D). Moreover, the square can be sheared, meaning its shape can be altered by applying a force in an upward, downward, leftward, or rightward direction at one of its corners (as depicted in Figure 17E) (Gundersen, 2018).



Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 17: Original Square under different Types of Transformations: (A) Stretched, (B) Compressed, (C) Rotated, (D) Reflected or Flipped and (E) Sheared.

The transformation under consideration is subject to a unique constraint: it must be linear. In other words, a linear transformation preserves the straightness of lines before and after the transformation. To illustrate this concept, consider a square lattice consisting of intersecting vertical and horizontal lines. When a diagonal line is superimposed on the square and a linear transformation is applied, the straightness of the diagonal line is preserved throughout the transformation process (see Figure 18B). On the contrary, a nonlinear transformation (shown in Figure 18C) can be visualised as the deformation of an engineering paper sheet when its midsection is pressed and curved (Gundersen, 2018).



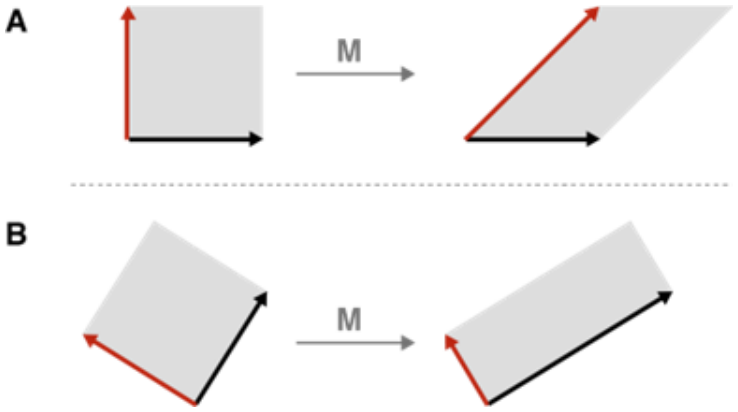
Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 18: (A) Original Square under a Linear Transformation M (B) and a Non-linear Transformation M(C)

Having established the characteristics of the square and its possible transformations, an important mathematical observation emerges. Let **M** represent a linear transfor-

mation to be applied to the square. If the square undergoes a prior rotation, there exists a specific rotation that can transform it into a rectangle after the application of **M**. This implies that by first rotating the square and then applying **M**, the resulting transformation will solely involve stretching, compressing, or flipping the square, without any shearing (Gundersen, 2018).

To illustrate this concept, consider an instance where the square is subjected to horizontal force, resulting in a sheared square (as illustrated in Figure 19A). However, if the square is first rotated before applying horizontal force, the resulting shear will only lead to stretching and compression of the square in a different orientation (as depicted in Figure 19B) (Gundersen, 2018).



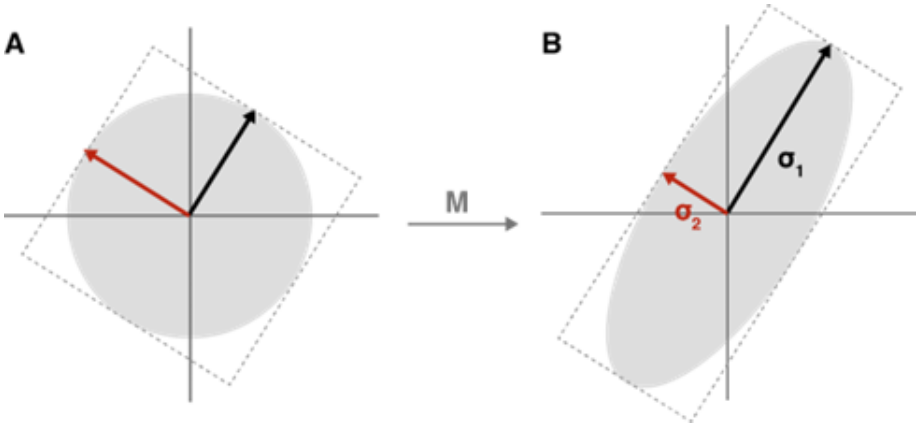
Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 19: The geometric essence of SVD: any linear transformation **M** of square (A) can be thought of as simply stretching, compressing, or reflecting that square, provided the square is rotated before and after (B)

The geometric essence of SVD can be understood as the fact that any linear transformation can be considered as merely stretching, compressing or flipping a square, provided that it is rotated prior to the application of the transformation. The resulting transformed square or rectangle may exhibit a new orientation (Gundersen, 2018).

This geometric concept is highly practical. The singular values that are the basis of the "singular value decomposition" are merely the length and width of the transformed square. These values can provide a wealth of information; for example, if one of the singular values is 0, this implies that the transformation has flattened the square. Furthermore, the larger of the two singular values can describe the maximum "action" of the transformation (Gundersen, 2018).

To clarify the previous statement, it may be helpful to visualise the transformation



Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 20: (A) Oriented Circle; imagine that circle inscribed in the original square. (B) the Circle Transformed into an Ellipse. The length of the major and minor axes of the ellipse have values σ_1 and σ_2 respectively, called the singular values.

without the additional rotation, which has no impact on the size of the resulting rectangle. In this scenario, one can perceive the rectangle in the bottom-right subplot of Figure 19 as having the same orientation as the rotated square in the bottom-left subplot. Additionally, instead of imagining the stretching (or flattening) of a square into a rectangle, it can be beneficial to consider the stretching of a circle into an ellipse. This is illustrated in Figure 20.

The significance of Figure 20 lies in its depiction of the singular values within the framework of an ellipse. More precisely, the major axis of the ellipse corresponds to the larger of the two singular values. When a perfect circle is transformed, all radii along its circumference are stretched to the edge of the resulting ellipse. However, the amount of stretching is not uniform across all radii. Rather, the radius pulled along the major axis experiences the greatest stretching. Consequently, the magnitude of the largest singular value precisely matches the amount by which the most stretched radius is extended (Gundersen, 2018)

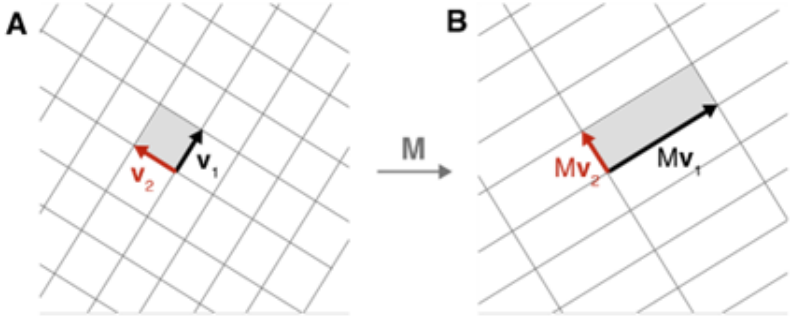
4.2 From Intuition to Definition

Having gained a geometrical intuition for SVD, it is now necessary to formalise these ideas. Firstly, it is essential to give names to the elements involved. It is a known fact that any two orthogonal vectors in two-dimensional space serve as a basis for that space. In the current context, the orthogonal vectors in the input space will be referred to as $\mathbf{v_1}$ and $\mathbf{v_2}$ (as illustrated in Figure 21A). Once a matrix transformation $\mathbf{M}$ is applied to these vectors, $\mathbf{Mv_1}$ and $\mathbf{Mv_2}$ are obtained (as shown

in Figure 21B) (Gundersen, 2018). Furthermore, it is important to decompose these two transformed vectors into unit vectors, namely $\mathbf{u}_1$ and $\mathbf{u}_2$, multiplied by their respective magnitudes, σ_1 and σ_2:

$$\begin{aligned} \mathbf{M}\mathbf{v_1} &= \mathbf{u_1}\sigma_1 \\ \mathbf{M}\mathbf{v_2} &= \mathbf{u_2}\sigma_2. \end{aligned} \tag{2}$$

Thus far, no new concepts have been introduced, as the only step taken has been assigning labels.



Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 21: Formalising the Geometric Essence of SVD: by properly rotating the domain defined by the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, then any linear transformation $\mathbf{M}$ is just a transformation by a diagonal matrix (dilating, reflecting) in a potentially rotated range defined by $\mathbf{u}_1$ and $\mathbf{u}_2$.

With the notation established, it is now possible to perform algebraic manipulations. In particular, it is worth noting that any vector $\mathbf{x}$ can be expressed in terms of the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, as follows:

$$\mathbf{x} = (\mathbf{x} \cdot \mathbf{v}_1)\mathbf{v}_1 + (\mathbf{x} \cdot \mathbf{v}_2)\mathbf{v}_2. \tag{3}$$

The notation "$\mathbf{a} \cdot \mathbf{b}$" denotes the dot product, or scalar product, of two vectors $\mathbf{a}$ and $\mathbf{b}$. Equation 3 involves the projection of $\mathbf{x}$ onto $\mathbf{v}_1$ and $\mathbf{v}_2$, accomplished through the utilisation of the dot product. This is followed by the decomposition of the resultant terms with respect to the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ (Gundersen, 2018).

The following step involves the left-multiplication of both sides of Equation 3 with the matrix transformation $\mathbf{M}$. Due to the scalar nature of the dot product, it is possible to distribute and commute $\mathbf{M}$ in the manner demonstrated below:

$$\mathbf{M}x = (\mathbf{x} \cdot \mathbf{v}_1)\mathbf{M}\mathbf{v}_1 + (\mathbf{x} \cdot \mathbf{v}_2)\mathbf{M}\mathbf{v}_2.$$

The subsequent step involves the conversion of $\mathbf{M}\mathbf{v}_i$ to $\mathbf{u}_i\sigma_i$:

$$\mathbf{M}x = (\mathbf{x} \cdot \mathbf{v}_1)\mathbf{u}_1\sigma_1 + (\mathbf{x} \cdot \mathbf{v}_2)\mathbf{u}_2\sigma_2.$$

The final steps involve observing the commutativity of the dot product, which allows for switching the order of the terms. For instance, $\mathbf{x} \cdot \mathbf{v}_1 = \mathbf{x}^T\mathbf{v}_1 = \mathbf{v}_1{}^T\mathbf{x}$. As the dot product produces scalar values, it is possible to move each term to the end of its respective expressions:

$$\mathbf{M}\mathbf{x} = \mathbf{u}_1\sigma_1\mathbf{v}_1{}^T\mathbf{x} + \mathbf{u}_2\sigma_2\mathbf{v}_2{}^T\mathbf{x}.$$

The elimination of the variable $\mathbf{x}$ from both sides of the equation is feasible since the equation $\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}$ generally implies that $\mathbf{A}$ is equal to $mathbfB$. For the sake of clarity and understanding, it may be beneficial to rephrase the statement as follows $(\mathbf{A} - \mathbf{B})\mathbf{x} = 0$, which results in $(\mathbf{A} - \mathbf{B}) = \mathbf{0}$. By removing the variable $\mathbf{x}$, the resulting equation is obtained:

$$\mathbf{M} = \mathbf{u}_1\sigma_1\mathbf{v}_1{}^T x + \mathbf{u}_2\sigma_2\mathbf{v}_2{}^T. \tag{4}$$

If matrices are defined appropriately, Equation 4 can be transformed into the canonical form of SVD (Equation 1) for 2x2 matrices:

$$\mathbf{M} = \underbrace{\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_1\end{bmatrix}}_{U} \underbrace{\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix}\mathbf{v}_1{}^T \\ \mathbf{v}_1{}^T\end{bmatrix}}_{V^T}. \tag{5}$$

Moreover, it is important to have a sense of understanding about what this implies. Every matrix transformation can be expressed as a diagonal transformation (scaling or reflection) that is determined by $\mathbf{\Sigma}$, given that the domain and range have been appropriately rotated beforehand (Gundersen, 2018).

The left singular vectors are denoted as $\mathbf{u}_i$, while the right singular vectors are represented by $\mathbf{v}_i$. However, this terminology can be misleading since the terms "left" and "right" are derived from the equation mentioned earlier, whereas in diagrams of rectangles and ellipses, the $\mathbf{v_i}$ vectors are illustrated on the left-hand side.

4.3 The Standard Formulation

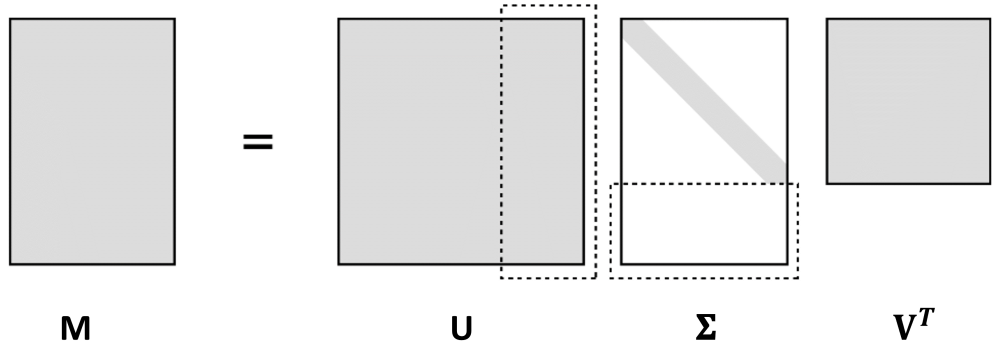
With a clear geometric understanding of the singular value decomposition (SVD) and a formal definition for all 2×2 matrices, the subsequent step is to restate the problem in a general and standard format. Given the definitions of $\mathbf{v}_i$, $\mathbf{u}_i$ and σ_i, it is possible to rephrase Equation 2 for any $m \times n$ matrix $\mathbf{M}$ as:

$$\begin{bmatrix} & & \\ & \mathbf{M} & \\ & & \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 | \mathbf{v}_2 | \cdots | \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 | \mathbf{u}_2 | \cdots | \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}.$$

This leads to the equation 5 again, but it is applicable for matrices with dimensions of $m \times n$:

$$\mathbf{MV} = \mathbf{U\Sigma},$$
$$\mathbf{MVV^T} = \mathbf{U\Sigma V^T},$$
$$\mathbf{M} = \mathbf{U\Sigma V^T}.$$

In this context, $\mathbf{U}$ represents an $m \times m$ unitary matrix, $\mathbf{V}$ represents an $n \times n$ unitary matrix and $\mathbf{\Sigma}$ is a diagonal matrix of size $m \times n$. The matrix $\mathbf{V}$ is orthonormal and therefore, $\mathbf{VV}^T = \mathbf{I}$. A diagrammatic representation of this can be seen in Figure 22.



Source: Singular Value Decomposition as Simply as Possible (2018)

Figure 22: The Canonical Diagram of the SVD Decomposition of a Matrix $\mathbf{M}$. The columns of $\mathbf{U}$ are the orthonormal left singular vectors; $\mathbf{\Sigma}$ is a diagonal matrix of singular values; and the rows of V^T are the orthonormal right singular vectors. The dashed areas are padding.

The singular values of $\mathbf{\Sigma}$ are found in the diagonal elements. It is possible to assume that these values are ordered. The geometric interpretation of singular values holds true in higher dimensions. For example, if the SVD is performed on an $m \times n$ matrix and the bottom k singular values are smaller than some epsilon, this can be visualised as flattening a hyperellipse along those k dimensions. Exponential decay in the singular values indicates that the matrix transformation primarily occurs in a few dimensions. This is essentially what the SVD is. To obtain a mathematically rigorous derivation of the Singular Value Decomposition (SVD), interested readers may refer to Chapter 1 of the book "Data Driven Science and Engineering: Machine Learning, Dynamical Systems and Control" authored by Steven L Brunton and Nathan J Kutz in 2019.

4.4 Leveraging SVD for the Generation of New Data Points

An important application of singular value decomposition (SVD) is to calculate the optimal rank-k approximation of a given matrix $\mathbf{X}$. The resulting low-rank matrices have a wide range of applications, including but not limited to dimensionality reduction, image compression, recommender systems, natural language processing and information retrieval (Phillips, 2016). To derive the best rank-k approximation of $\mathbf{X}$, the SVD of $\mathbf{X}$, which can be expressed as $\mathbf{U\Sigma V^T}$, is typically represented as a summation of rank-1 matrices. That is:

$$\mathbf{X} = \mathbf{U\Sigma V}^T = \sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u_i v_i}^T, \tag{6}$$

where, σ_i,$\mathbf{u}_i$ and $\mathbf{v}_i{}^T$ are the singular value, the left singular vector and right singular vector of the i^{th} rank-1 matrix respectively. Moreover, it is typically assumed that the singular values have been arranged in descending order, such that $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0)$. Consequently, the first rank-1 matrix corresponds to the largest singular value, the second rank-1 matrix corresponds to the second largest singular value and so forth. As each rank-1 matrix captures the underlying correlations or patterns present in the data, the first rank-1 matrix represents the most dominant correlation or pattern observed in the data. Subsequent rank-1 matrices encode progressively fewer of these correlations or patterns as one moves towards the $\min(m,n)^{th}$ rank-1 matrix (Phillips, 2016).

The phenomenon described can be observed by analysing the shapes of the histograms presented in Figure 25. Specifically, the histograms illustrate the distribution of values for a variable (sepal length) that has been extracted from the four

rank-1 matrices obtained through SVD of the iris data set (Fisher, 1988). The histogram of sepal length associated with the rank-1 matrix with the largest singular value depicts a non-normal distribution with a mean approximately equal to that of the sepal length variable in the original data set. Subsequent histograms of sepal length associated with rank-1 matrices corresponding to the second largest through to the smallest singular value encode deviations of sepal lengths from the average sepal length of the sample. The deviations have a mean which changes from a non-zero value to approximately zero for rank 1 matrices ranging from the rank 1 matrix associated with the second largest singular value to the rank 1 matrix associated with the smallest singular value. Furthermore, deviations associated with larger singular values exhibit an asymmetric distribution whilst deviations associated with smaller values of the singular values exhibit symmetric distributions. For example, the histogram of deviations associated with the second largest singular value is non-normal with three modes, possibly indicating the three types of flowers in the sample. The histograms of the sepal length variable for rank-1 matrices with smaller singular values gradually approach a normal distribution with a mean of zero and a positive variance. Additionally, the variance of these distributions decreases as the singular values become smaller (Phillips, 2016).

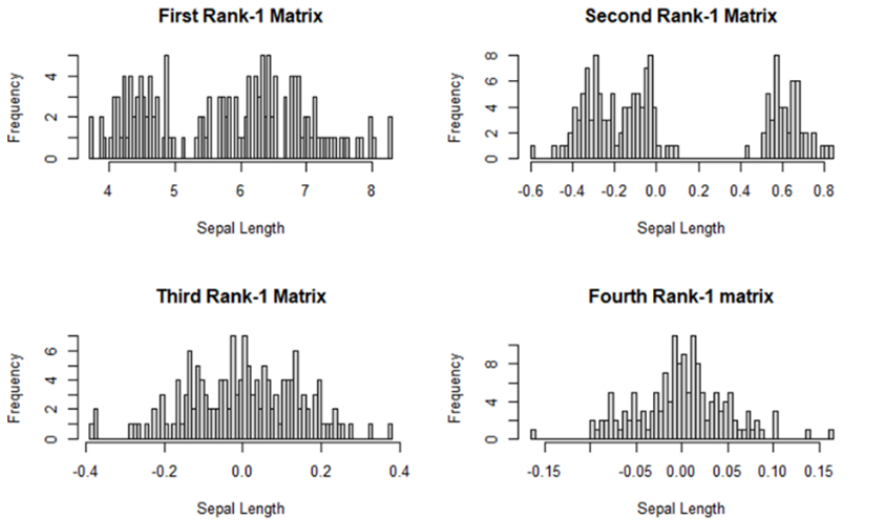


Figure 23: Values of a Variable, Sepal Length, Selected from Each of the Four Rank-1 Matrices Derived from the SVD of the Iris Data Set.

The objective of low rank approximation is to retain only the first k terms, where k is much smaller than $\min(m, n)$, associated with the first k eigenvalues on the right-hand side of (6). These k terms are deemed to represent the most significant correlations or patterns observed in the data $\mathbf{X}$, as depicted in Figure 25. The

remaining $\min(m,n)-k$ terms are regarded as noise and are therefore discarded. The optimal low rank approximation is demonstrated to be optimal in both Frobenius and L_2 norms. An immense benefit of the best rank k approximation is that it only takes $O(k(m+n))$ space to store the k rank-1 matrices, as opposed to the $O(mn)$ space required to store the original matrix $\mathbf{X}$. This represents a notable advantage, especially when the value of k is relatively small, while the dimensions m and n are significantly large, as commonly encountered in various applications (Phillips, 2016).

4.5 Proposed SVD based algorithm for Data Generation

In the preceding sections, it was discussed that SVD can be applied to find the optimal rank-k approximation of a matrix $\mathbf{X}$ by retaining only the first k terms of the expression $\sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^T$, where k is much smaller than $\min(m,n)$. These k terms are considered to explain most of the variation in the data, while the remaining $\min(m,n)-k$ terms are considered as noise and are ignored. In this chapter, an algorithm is proposed for generating new samples from an existing data set by combining the two components, $\sum_{i=1}^{k} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^T$ and $\sum_{i=k+1}^{\min(m,n)} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^T$, in a natural manner. The idea is rooted in the remarkable resemblance between the roles played by the components in the Singular Value Decomposition (SVD) equation and those in the general formula for predictive modeling, as depicted in (7),

$$y = f(H) + \epsilon, \tag{7}$$

in which, $f(H)$, denotes the true or underlying or systematic signal and ϵ represents the random variation or error or residual signal. In 7, the distribution of ϵ is typically assumed to be normal or Gaussian with a mean of zero and a constant variance. Furthermore, ϵ is assumed to be independent of H, while errors themselves are assumed to be independent of each other.

Let $\mathbf{X} \in R^{m \times n}$, representing a data matrix. In analogy to (7), the matrix $\mathbf{X}$ can be interpreted as representing the observed response, y, while the components $\sum_{i=1}^{k} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^T$ and $\sum_{i=k+1}^{\min(m,n)} \sigma_i \mathbf{u}_i {\mathbf{v}_i}^T$ can be viewed as representing the true signal, $f(H)$ and the error term ϵ, respectively. In this study, we will assume that $k = [\min(m,n) - 1]$, that is, restrict the error term, ϵ, to be made up of the rank 1 matrix associated with the smallest singular value only. This is because theoretically rank 1 matrices associated with the smallest singular values represent noise in the data. Furthermore, although the demonstration in Section 4.4 appeared to suggest that the histograms of deviations associated with the rank-1 matrix having

the smallest singular value follow a normal distribution with a mean of zero and positive variance, this fact is not always true, as will be seen in Chapter 5. As such, in this study, we will assume that the deviations for such matrices follow a symmetric distribution (not necessarily normal) with a mean of zero. The symmetric but non-normal distributions seem to occur in cases where the variables in the data do not exhibit significant correlations between themselves. Thus, setting $\epsilon = \sigma_{\min(m,n)}\mathbf{u}_{\min(m,n)}\mathbf{v}^T_{\min(m,n)}$, and noting that the eigenvalue and eigenvector pair, $(\sigma^2_{\min(m,n)}, \mathbf{v}_{\min(m,n)})$, is fixed, it is clear that the randomness in ϵ comes from the vector, $\mathbf{u}_{\min(m,n)}$. Assuming that the samples in the original data set are independent, it is reasonable to assume that the values in the vector, $\mathbf{u}_{\min(m,n)}$, are also independent of each other. Consequently, it is possible to draw a sample of errors of size, m, by randomly redistributing the values of m in the vector, $\mathbf{u}_{\min(m,n)}$, followed by multiplying the resulting vector, ${\mathbf{u}^*}_{\min(m,n)}$, by $\sigma_{\min(m,n)}\mathbf{v}^T_{\min(m,n)}$. By adding errors to the systematic component $\sum_{i=1}^{\min(m,n)-1}\sigma_i\mathbf{u}_i{\mathbf{v}_i}^T$ new samples of data would be obtained. By repeatedly employing this approach, a new data set is generated that possesses the same underlying distribution as the original data.

In practice, to make sampling more general, instead of sampling through the repeated random redistribution of the same values in the vector, $\mathbf{u}_{\min(m,n)}$, an empirical distribution of the values in the vector, $\mathbf{u}_{\min(m,n)}$, is formed. The use of the empirical distribution in sampling ensures a proportional representation in the sample. Samples of size m, are drawn randomly from the empirical distribution to form a vector ${\mathbf{u}^{**}}_{\min(m,n)}$. The corresponding rank-1 matrix of errors is given by the product, $\sigma_{min(m,n)}{\mathbf{u}^{**}}_{\min(m,n)}\mathbf{v}^T_{\min(m,n)}$. New data is then generated by adding the rank-1 matrix of errors to $\sum_{i=1}^{\min(m,n)-1}\sigma_i\mathbf{u}_i{\mathbf{v}_i}^T$. This approach enables the simultaneous modification of all the variable values associated with the relevant rank-1 matrix, thereby facilitating the prompt generation of any desired quantity of data. To ensure that representative samples are drawn from the empirical distribution of the values of $\mathbf{u}_{\min(m,n)},$, it is essential to conduct the procedure separately within each class (i.e., for instance where the response variable has various classes), rather than between different classes. This is because there may be differences in the mean, variance, skewness, kurtosis or general distribution characteristics of the deviations between classes. To accomplish this, it is necessary to group the classes and obtain a sample that matches the size of the original sample for each class. This sample should be drawn from the empirical distribution of $\mathbf{u}_{\min(m,n)},$, utilising the indices associated with each class.

The proposed method offers a solution for generating new data points from small data sets, tackling a significant challenge across multiple disciplines. The following

section will explore the methodology in detail, which will provide a roadmap to assess the effectiveness of the method in generating data points that preserve key characteristics seen in the observed data set.

5 Methodology

The main objectives of the research were outlined in Chapter 1 and are repeated here for the sake of completeness. The first is to develop an algorithm based on SVD for generating new data from an original data set such that the new data exhibit the same distributional properties as the original data set. The second objective is to evaluate the effectiveness of the algorithm in generating new data sets from small data sets and the third and final objective is to compare the performance of regression and classification models on the new and original data. There are also two hypotheses that need to be tested. The two hypotheses are the following.

- The data points generated from singular value decomposition (SVD) have the same distribution as the original data set.
- The model trained on the SVD-generated data set performs as well or marginally less than the model trained only on the original data set.

5.1 Performance measures

5.1.1 Performance Evaluation of SVD algorithm at generated new data points

To assess the similarity or dissimilarity between the observed and generated data sets, it is crucial to establish a set of criteria to use. These criteria should consider various factors, including the size, structure, content and statistical properties of the data sets. Evaluation criteria commonly employed to compare data sets include measures of central tendency (e.g., mean, median and mode), measures of dispersion (e.g., standard deviation, variance and IQR), correlation coefficients, hypothesis testing and visualisation techniques such as scatter plots and histograms, which apply to numerical data sets only. The selection of appropriate evaluation criteria depends on the characteristics of the data sets and the specific research question under investigation. For this study, the following criteria have been chosen.

Histogram
This evaluation approach involves the selection of variables from both the observed and generated data. Two histograms are created, one representing a variable from the observed dataset and the other representing a corresponding variable from the generated data set. The histograms are examined to assess the level of similarity between the selected variables. Histograms are commonly used to visualise the shape of a distribution - whether it is symmetric, skewed or bimodal. Histograms can also help identify outliers or gaps in the data. If the histograms have a similar shape and location, then the variables may be considered to be of the same distribution.

However, if the histograms have different shapes or locations, then the variables may not have come from the same distribution.

Boxplot
A different set of variables is selected from both the observed and generated data and boxplots are then constructed and analysed to determine the similarity between them. If the box plots are in a similar location and have similar spreads, then the variables may have come from the same distribution. If the boxes are in different locations or have different spreads, then the variables may not have come from the same distribution. By examining the boxplot, the distribution of the data sets can be visualised and any outliers or extreme values can be identified.

Correlation
This approach involves first, computing the dissimilarity matrices of the correlation matrices of the observed data and the generated data and second, computing the Frobenius norm of their differences. The Frobenius norm between two matrices A and B is calculated as

$$d_2(A, B) = \sqrt{\sum_{i}^{n} \sum_{j}^{n} (a_{ij} - b_{ij})^2},$$

where a_{ij} given by $a_{ij} = 1 - |\rho_{ij}|$ is the dissimilarity value of the element in row i and column j of matrix A corresponding to the correlation coefficient ρ_{ij} in row i and column j of he correlation matrix. b_{ij} is similarly defined (Solomon, 2023). If the Frobenius norm value is small, it suggests that the correlation matrices of the two data sets are similar, indicating that the relationship between the variables is preserved in the newly generated data set. Conversely, a large Frobenius norm value would indicate that the correlation matrices of the two data sets are dissimilar, suggesting that the relationship between the variables is not preserved in the newly generated data set.

Reconstruction Error
This reconstruction error is the difference between the observed data and the data generated from the SVD-base method. A lower reconstruction error indicates that the method provides a better approximation of the observed data. Conversely, a higher reconstruction error indicates that the method performs poorly in approximating the observed data.

Kolmogorov–Smirnov (KS) Test
This is a statistical test used to compare whether the same variables in two different data sets follow the same distribution. The KS test checks how likely it is that two sets of data come from the same distribution. It does this by comparing the pattern of values of the same variables in the two data sets. The test calculates a test statistic, which represents the largest vertical distance between the cumulative distribution functions (CDF) of the same variables in the two data sets. If this test statistic is small for every variable, it suggests that the two data sets have similar distributions. On the other hand, a large test statistic indicates that the data sets likely come from different distributions. If the p-value is below a predetermined level of significance, typically 0.05, the null hypothesis that the two data sets come from the same distribution is rejected in favour of the alternative hypothesis that they come from different distributions (Lopes, 2011).

5.1.2 Performance measures for modelling data with regression and classification models

The following section discusses the evaluation metrices that were used to compare the performance of the regression and classification models on the original train data set versus the corresponding new train data generated from six samples of different sizes drawn from original train data.

Metrics for Classification tasks
The following are the most widely used metrics for assessing classification algorithms.
Accuracy
Accuracy measures the percentage of correctly predicted instances out of the total number of instances in the data set. In other words, it is the ratio of the number of correct predictions to the total number of predictions made by the model (Goyal,2021).

Area Under the ROC Curve (AUC)
AUC is a common metric used to evaluate the performance of binary classification models, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds. AUC measures the degree of separability between the positive and negative classes and ranges from 0 to 1, with higher values indicating better model performance (Goyal,2021).

Recall
Recall is a metric used to evaluate the performance of a machine learning model

in binary classification tasks. It measures the proportion of true positive instances that are correctly classified as positive out of the total number of positive instances in the data set. In other words, it is the ratio of the number of true positives to the sum of true positives and false negatives (Goyal,2021).

Precision
Precision measures the proportion of true positive instances that are correctly classified as positive out of the total number of instances that are classified as positive. In other words, it is the ratio of the number of true positives to the sum of true positives and false positives (Goyal,2021).

Matthews Correlation Coefficient (MCC)
MCC is a more reliable statistical rate that yields a high score only if the prediction achieved good results in all four confusion matrix categories (i.e., true positives, false negatives, true negatives and false positives), proportional to the size of both positive and negative elements in the data set (Li,2020).

Metrics for Regression tasks
The most widely used metrics for assessing regression tasks are briefly discussed below.

Mean Squared Error (MSE)
MSE measures the average squared difference between the predicted and actual values of the target variable. In other words, it is the average of the squared residuals (the difference between the predicted and actual values) of the model (mathematical formulation can be found in appendix A.4.1) (Rowe, 2018).

Root Mean Square Error (RMSE)
RMSE calculated as the square root of the average of the squared differences between the predicted and actual values of the target variable. In other words, it is the square root of the mean squared error (MSE) (mathematical formulation can be found in appendix A.4.2) (Moody, 2019).

R-Squared Score (R^2)
Contrary to other metrics, such as MSE and RMSE, R^2 is not a measure of the accuracy of the predictions, but rather a measure of how well the predictions fit the data. R-Squared measures how much of the variation in the dependent variable is

explained by the independent variables of the model (mathematical formulation can be found in appendix A.4.3) (Allwright, 2022).

Mean Absolute Percentage Error (MAPE)
MAPE measures the average percentage difference between the predicted and actual values of a data set. MAPE is expressed as a percentage and it indicates how far off the predictions are from the actual values. A smaller MAPE value indicates that the model is better at predicting the actual value (mathematical formulation can be found in appendix A.4.4) (Ahmed,2023).

5.2 Addressing the objectives and hypotheses

This section provides details of how the three main objectives of the study were addressed.

5.2.1 Objective 1

The first objective was to develop an algorithm based on SVD for generating new data from an original data set such that the new data exhibit the same distributional properties as the original data set. The first part of the objective, that is, the development of the SVD-based algorithm for generation of new data was addressed in the latter part of Chapter 4. To evaluate the ability of the method in generating new data points with similar characteristics as the original data, the algorithm was applied to the three observed data sets: Abalone Age, Life Expectancy and NBA. This was performed on the data sets that resulted after EDA. New data sets of the same size as the original data sets were generated. The pseudocode used by the algorithm to generate the new data is as follows:

Algorithm 1 An SVD based algorithm for generating new data

Data $\leftarrow$ [Abalone, NBA, LifeExp]
Error $\leftarrow$ [ ]
while $j \leq 3$ **do**
  $k \leftarrow$ min(nrow(Data[j]),ncol(Data[j]))
  Apply SVD to dataset and generate the matrices U, Σ and V
  Using values in column k of U, denoted U_k, form an empirical
  cumulative distribution function, $ECDF_k$
  while $i \leq m$ **do** ▷ m is the number of rows in Data[j]
    Randomly draw a value, u, from the $ECDF_k$
    Error.append($\Sigma_{kk} * u * V_k^T$)
  end while
  NewData $\leftarrow \sum_i^{k-1} \Sigma_{ii} U_i V_i^T + Error$
  Save New Data as NewData+'j'
end while

The null hypothesis related to the first objective is that the data points generated from singular value decomposition (SVD) based algorithm will have the same distribution as the original data set. To test this hypothesis and assess the effectiveness of the proposed model to generate new data points the criteria discussed in section 5.1.1 were used.

5.2.2 Objective 2

The second objective was to evaluate the effectiveness of the algorithm at generating new data sets from small data sets. To simulate small data sets, six samples from each of the original data sets were drawn of sizes equal to 25, 50 and 100 observations and 25%, 50%, 75% of the original data respectively. Considering each of the six samples as the "original data sets", the SVD-based algorithm was applied to each sample in turn as discussed above and from each sample, new data sets of the same size as the original data set was generated. The original data set for Abalone Age or Life Expectancy or NBA is divided into train and test sets in a ratio of approximately 85% to 15%, however, for this objective the test sets were not used. For example, because the total number of observations in the Abalone data set after EDA is 3781 and training set used to evaluate the effectiveness of the generated data points is 3210 it implies that from each of the six samples of the Abalone age data set 3210 new observations were generated. Evaluating the effectiveness of the algorithm at generating new data sets from small data sets involved comparing distributional properties of the six new data sets with the properties of the original dataset from which the small data set used to generate the new data sets were sampled using

several criteria as discussed in section 5.1.1. The figure below can be applied to other data sets using the parameters specified for each of those data sets as outlined in Table 6.

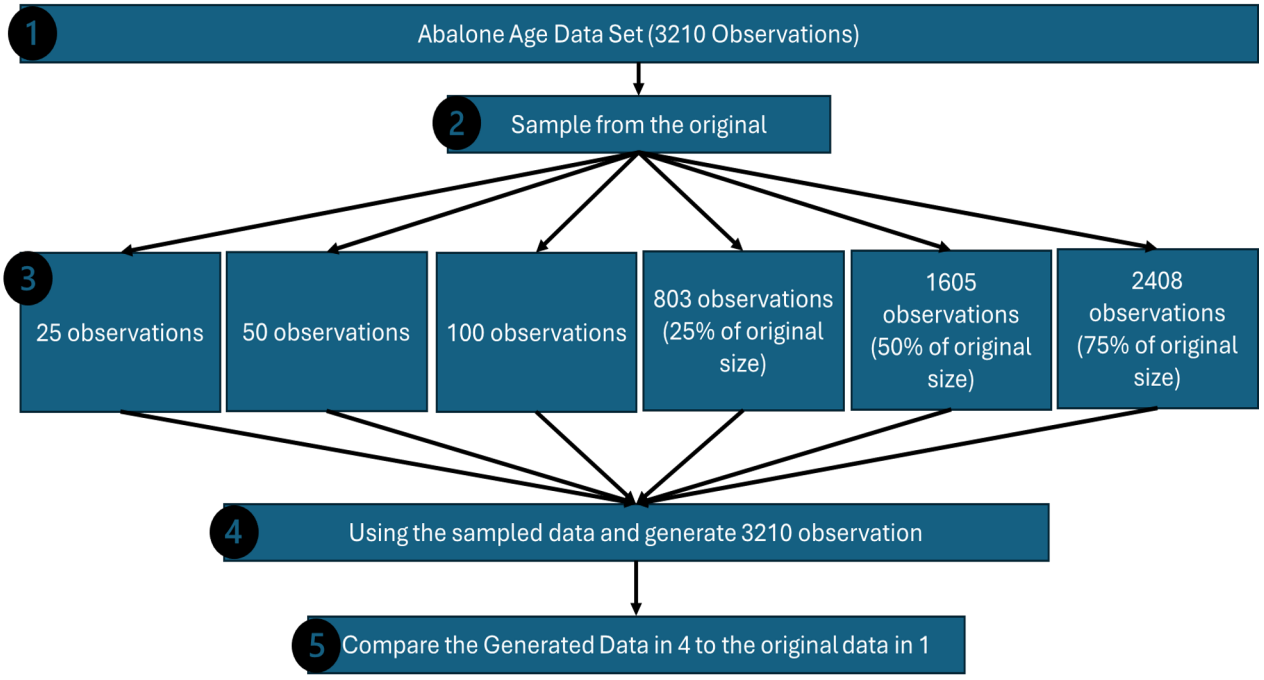


Figure 24: This is flow chart depicting how objective 2 will achieved when working with Abalone Age data set.

5.2.3 Objective 3

The third objective was to compare the performance of regression and classification models on the new and original data. The null hypothesis in this case is that the model trained on the data set generated using the SVD based algorithm performs as well as the model trained on the original data set. Similar to what was done under objective 2, to address the stated objective as well as test the associated hypothesis, first the original data set for Abalone age or Life expectancy or NBA is divided into train and test sets in a ratio of approximately 85% to 15%. To mitigate the risks associated with data snooping, where data is repeatedly examined without accounting for the heightened chance of false positive outcomes, the splitting of the data into train and test sets were therefore done before commencing with any analysis. The training sets are employed for training the models whilst the testing sets are utilised to assess the accuracy of the models. Given that there are three different data sets to work with, implies that after the split there will be a total of three train sets and

three test sets.
Six samples from each of the three train sets were drawn of sizes equal to 25, 50 and 100 observations and 25%, 50%, 75% of train data respectively. As before, considering each of the six train samples as the "original train set", the SVD-based algorithm was applied to each train sample in turn as discussed above and from each train sample, new train data sets of the same size as the original train sample were generated. The six new train sets plus the original train set are each used to train a particular type of model (e.g. Logistic Regression) whilst the corresponding test sample is used to evaluate the performance of all the resulting seven trained models. In this study, the performance of three types of models on each of the three original data sets and their corresponding six new data sets were examined. Similarly, the figure below can also be applied to other data sets using the parameters specified for each of those data sets outline in the Table 6.

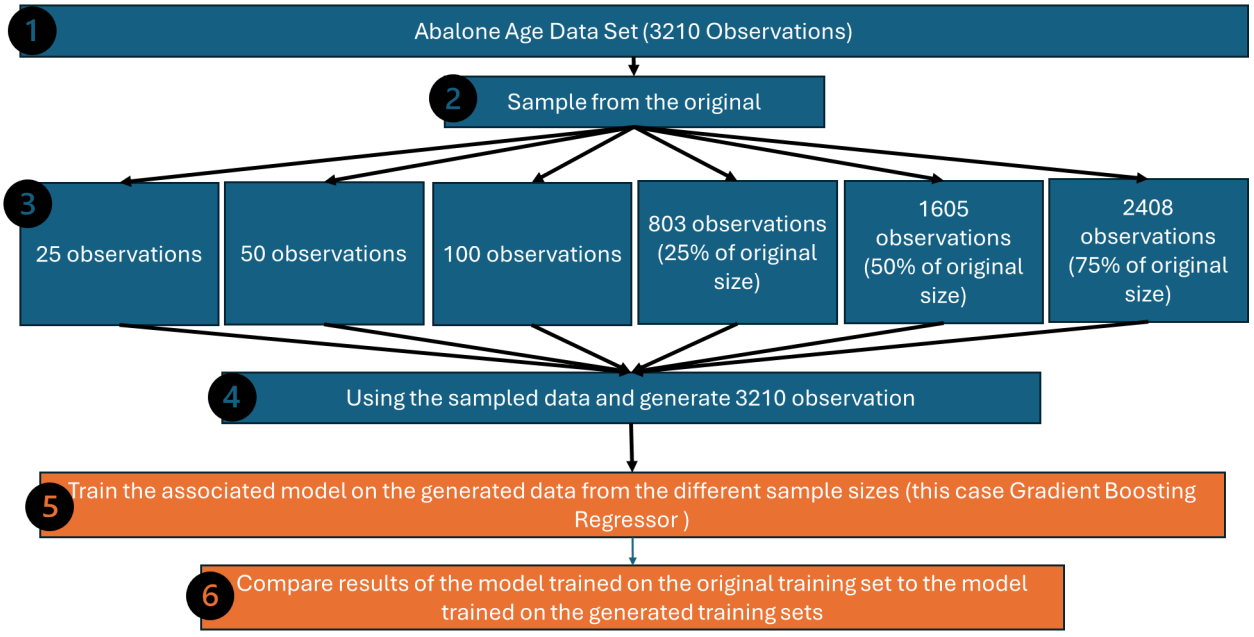


Figure 25: This is flow chart depicting how objective 3 will achieved when working with Abalone Age data set.

The three types of models considered were Gradient Boosting Regressor, Extra Trees Regressor and Logistic Regression. A brief description of these models is given below.

Gradient Boosting Regressor
The Gradient Boosting Regressor is an ensemble technique where models are incre-

mentally added in sequence. It starts with an initial model, often something simple like the mean target value. In each iteration, a new model is trained to predict the errors of the current ensemble. This process is optimised through gradient descent, which minimises the loss function, such as mean squared error. The new model is then added to the ensemble, updating the overall predictions. The final prediction is a weighted sum of the outputs from all the models in the ensemble (Saini,2021).

Extra Trees Regressor
This a machine learning algorithm used for regression tasks. It is an ensemble method that creates multiple decision trees and combines their predictions to make a final prediction. Extra Trees employs the complete data set to train decision trees. To provide significant variations between distinct decision trees, the algorithm randomly determines the values at which to split a feature and construct child nodes, which leads to greater randomness and a potential reduction in variance (Thankachan,2022).

Logistic Regression
Logistic regression is a statistical approach for performing binary classification. It predicts the probability that a given instance falls into one of two categories. Essentially, logistic regression is a statistical model that employs a logistic function to represent the relationship between a dependent variable and one or more independent variables. This model operates using odds and log-odds, with coefficients determined through Maximum Likelihood Estimation (MLE). Once trained, it provides the probability of an outcome for new data points and classifies these data points based on a set threshold, typically 0.5 (Al-Serw, 2021).

The performance of the three models on each of the three sets of data where each set is made up the original data plus the six new data sets derived from the original data as discussed were assessed using the metrics described in 5.1.2. The hypothesis related to objective 3 was assessed using a comparative approach to the original, rather than employing a formal statistical test, as exercised in various papers by Kamath et al. (2018), Brownlee (2019) and Zohair (2019).

5.2.4 Experiment Parameters and Environment for Implementation of the Regression and Classification Models

To implement the three algorithm the Pycaret Package was used. PyCaret is an opensource Python library that aims to simplify the end-to-end machine learning workflow. It provides a high-level interface and automates various tasks involved in

building and deploying machine learning models. PyCaret combines several popular machine learning libraries, such as scikit-learn, XGBoost, LightGBM and CatBoost, to offer a unified framework for streamlined model development. The package offers various additional features and functionalities, including data preprocessing, automatic setup, model training and selection, model comparison, interpretability and model deployment. However, for this study, as the exploratory data analysis was primarily conducted in chapter three, the package was utilised specifically for tasks such as model training, optimising and deployment.

Table 6: Parameters

Data sets	**Model used**	**Target variable**	**Various numbers of observations to be explored**					
			25-Obs	**50-Obs**	**100-Obs**	**25%**	**50%**	**75%**
Abalone Age	Gradient Boosting Regressor	Age	25	50	100	803	1605	2408
Life Expectancy	Extra Trees Regressor	Life Expectancy	25	50	100	138	275	413
NBA	Logistic Regression	Target_5yrs	25	50	100	210	420	630

The outlined methodology offers a structured approach to evaluating the effectiveness of the proposed method through thorough analysis and systematic assessment. The following section will present the findings obtained from this methodology, highlighting the impact and effectiveness of the method.

6 Results and Discussion

This chapter presents and discusses the results of the research study in relation to the objectives and hypotheses set out in Chapter 1.

6.1 Comparing the Distribution of Original and the SVD-Generated Data Sets

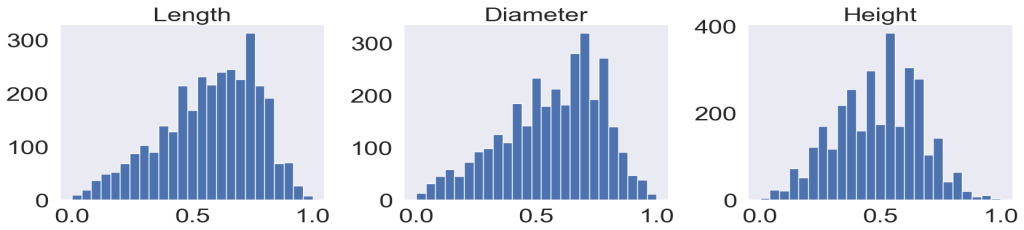
The second part of the first research objective was to assess whether the original and the generated data sets share the same distribution. Several evaluation metrics were employed, including histograms, box plots, correlation analysis with the Frobenius norm, reconstruction error and the KS test. Because the results comparing the distributional properties of the original and the SVD generated data involves a lot of graphical output and tables, to avoid clutter, in the following subsection, only the results of the Abalone age data set will be presented in the text while corresponding results for the other two data sets will be presented in the Appendix. However all the results relating to the three data sets will be discussed below.

6.1.1 Abalone Age Data Set

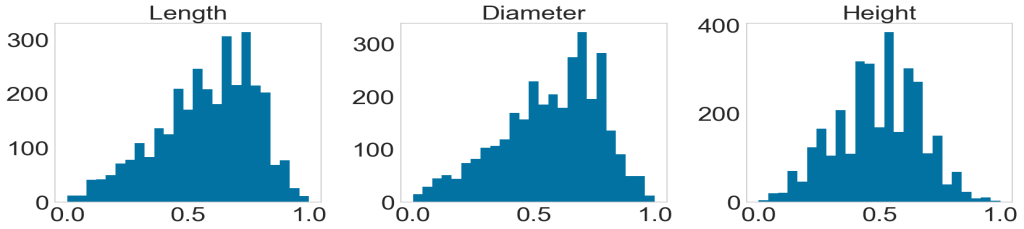
Histogram
The histograms in Figure 26 compare the distribution of the first three variables in the original and generated Abalone Age data sets.

The histograms provide visual evidence that the newly generated Abalone data set approximates the distribution of the original Abalone data set. The shape, center and spread of the distribution in the newly generated Abalone data set are comparable to those of the original Abalone data set, as seen in Figure 26b. Whilst there are some minor differences in the frequency of certain bins, the overall layout of the distributions remains similar. These findings, though empirical, suggest a high level of comparability between the newly generated Abalone data set and the original Abalone data set, particularly for the first three variables.

The same trend was observed in the histograms of both the Life Expectancy and NBA data as can be seen from Figure 36 in Appendix Section A.5.2) and Figure 40 in Appendix Section A.5.3 respectively.



(a) Original Data Set



(b) Data Set Generated from 3210 Observations using the Proposed Model

Figure 26: Histogram of the First Three Variables from the Abalone Age Data Set and the New Generated Data Set Derived from a Subset of the Abalone Age Data Set

Boxplot

The boxplots in Figure 27 compare the distribution of the fourth, fifth and sixth variables in the original and newly generated Abalone Age data sets.The boxplots reveal that the distributions of the variables in the original and generated data sets are similar. The boxplots demonstrate near identical shapes, central values, spreads and 5-number summaries, highlighting a high degree of similarity. Furthermore, the dispersion of the boxes and whiskers is comparable across all variables. Similar observations were observed among boxplots comparing the distributions of the fourth, fifth and sixth variables in the original and generated data for Life Expectancy and NBA, as depicted in Figure 37 in Appendix Section A.5.2 and Figure 41 in Appendix Section A.5.3 respectively. Although, some minor deviations in distribution are discernible from the boxplots, these deviations are of marginal significance in nature.

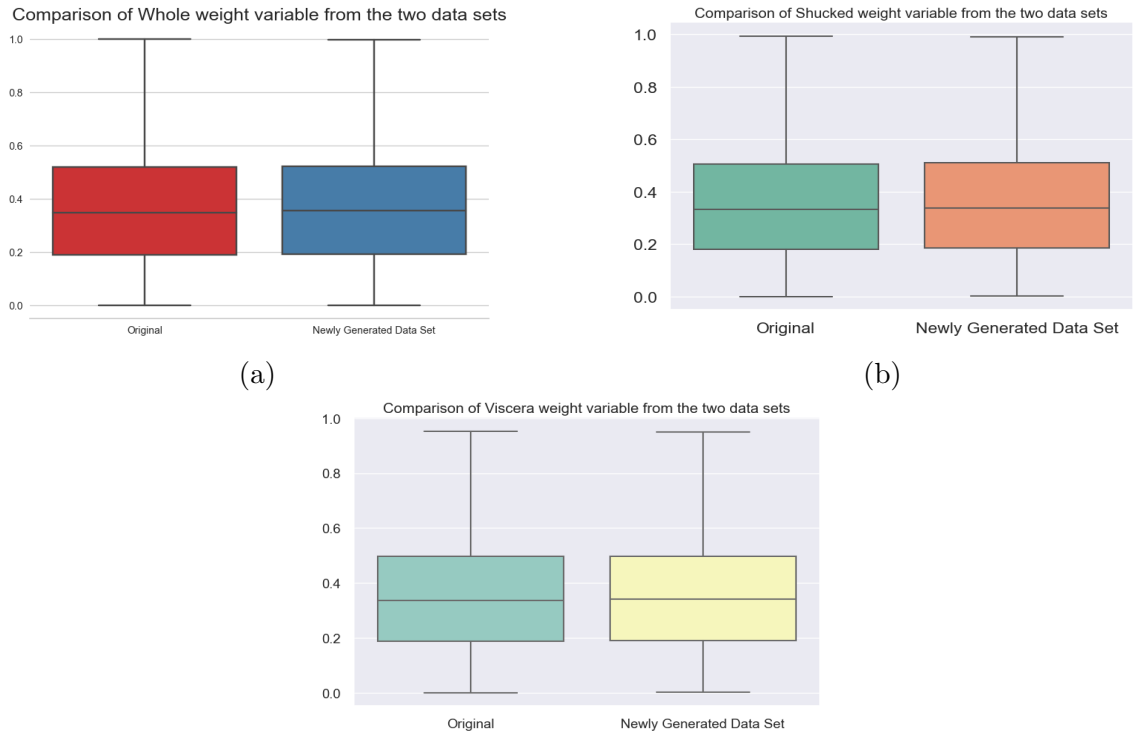


Figure 27: Boxplots Comparing the Variables found in the Original Data Set and the Newly Generated Data Set For the Abalone Age Data Set

Correlation

Using the correlation matrices of both the newly generated Abalone age data set and the original Abalone age data sets, corresponding dissimilarity matrices were calculated and from these, the Frobenius norm was computed. A value of 0.119 was obtained from the Abalone data set generated from 3210 observations of the original Abalone data set. The value indicates a relatively small difference between the matrices. The proximity of most of these values to zero suggests that the matrices are fairly similar, indicating that the newly generated Abalone data set closely approximates the original Abalone data set in terms of correlation.

The corresponding values of the Frobenius norm obtained for the Life Expectancy and NBA data sets were 0.682 and 0.272 respectively. In general, these values may be regarded as small, signifying similarity in correlation matrices between the original and newly generated data sets. The preservation of variable relationships in the new data set is evident, most notably in the Abalone Age data, followed by the

NBA data and finally the Life Expectancy data.

Reconstruction Error
Table 7 shows the reconstruction error between newly generated Abalone age data and the original Abalone age data. The majority of variables in the data demonstrate a reconstruction error of less than 0.68 with the exception of the Age variable, as it is due to its higher measurement unit.
Similarly, the reconstruction errors between newly generated data and the original data for Life Expectancy and NBA variables are generally low as can be seen from Table 34 and 46 respectively.

Table 7: Reconstruction Error between the Original and SVD-generated Data Set using 3210 Observations for Abalone Age Data Set

Columns	**Recon Error**
Length	0.290
Diameter	0.300
Height	0.255
Whole weight	0.312
Shucked weight	0.313
Viscera weight	0.301
Shell weight	0.293
Age	3.279
Sex_F	0.667
Sex_I	0.655
Sex_M	0.675

Kolmogorov–Smirnov (KS) Test
Table 8 shows the calculated p-values from the Kolmogorov-Smirnov test, which were obtained for all the comparisons between the generated Abalone age data set and the original Abalone age data set, exceed the significance level of 0.05. This indicates that the null hypothesis, assuming no noteworthy difference between the compared samples (i.e., the two distinct samples are drawn from the same distribution), cannot be rejected.

Similarly, calculated p-values for the Kolmogorov-Smirnov (KS) Test obtained for all comparisons between the newly generated and original data of Life Expectancy and NBA data as shown in Table 35 and 47) in Appendix Section A.5.2 and A.5.3, respectively, provide no substantial evidence suggesting divergence between the distributions of newly generated and original data sets.

Table 8: Kolmogorov–Smirnov (KS) Test between the Original and SVD-generated Data Set using 3210 Observations for Abalone Age Data Set

Column Names	**KS Stat**	**P value**
Length	0.012	0.983
Diameter	0.012	0.983
Height	0.012	0.983
Whole weight	0.012	0.983
Shucked weight	0.012	0.983
Viscera weight	0.012	0.983
Shell weight	0.012	0.983
Age	0.012	0.983
Sex_F	0.012	0.983
Sex_I	0.012	0.983
Sex_M	0.012	0.983

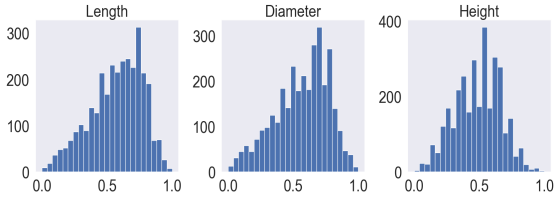
6.2 Evaluate the effectiveness of the algorithm at generating new data sets from small data sets

To assess whether the original and the generated data sets, derived from small data set, share the same distribution, the same evaluation metrics that were used in the previous subsection were employed. Again as before, to avoid clutter, only results for the Abalone age data set will be included in the text. The results of the other two data sets will be placed in the Appendix.

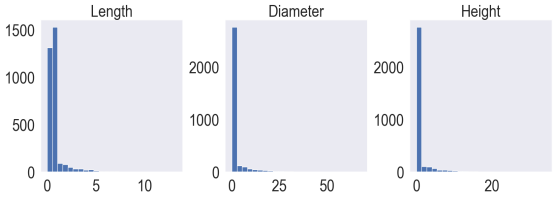
6.2.1 Abalone Age Data Set

Histogram
The histograms in Figure 28 compare the distributions of the first three variables in the original and generated Abalone Age data sets derived from samples of 25, 50, 100 and 803.

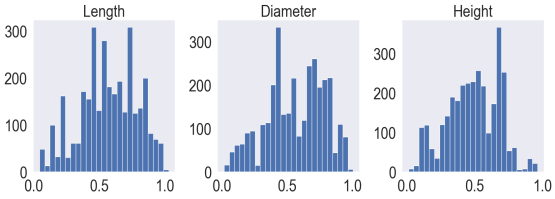
The histograms show that the newly generated Abalone age data set approximates the distribution of the original data set except when the input data set to the algorithm is a sample of 25 observation of the original Abalone age data set. In that case, the data set generated does not reflect the characteristics seen in the original data set. Additionally, as more observations of the original data are used as input to the algorithm to generate new data points, the shape, center and spread of the distribution in the newly generated data sets become more comparable to those of the original data set as can be seen from Figure 28c and the following figures thereafter. Whilst there are some minor differences in the frequency of certain bins, the



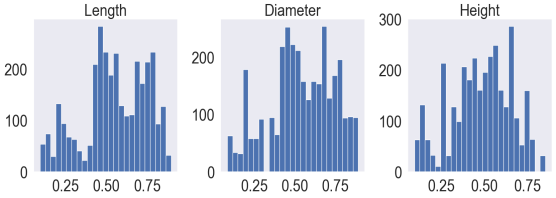
(a) Original Data Set



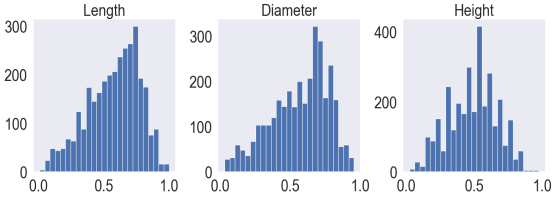
(b) Data Set Generated from 25 Observations using the Proposed Model



(c) Data Set Generated from 50 Observations using the Proposed Model



(d) Data Set Generated from 100 Observations using the Proposed Model



(e) Data Set Generated from 803 Observations using the Proposed Model

Figure 28: Distributions of the first Three Variables in the Original Abalone Age Data set and Generated Abalone Age Data Sets Derived from Samples of 25, 50, 100 and 803

overall layout of the distributions remains similar. These findings suggest a fairly high level of comparability between the newly generated Abalone data sets and the original Abalone data set. Additional histograms, derived from using samples of size 1605 and 2408 observations of the original Abalone age data set, can be found in Figure 32 in Appendix A.5.1, which further supports the findings above.

Figures 38 and 42 in Appendix Section A.5.2 and A.5.3, show histograms comparing the distributions of the first three variables of the original and generated data sets for Life Expectancy and NBA, respectively. Similar to the Abalone findings, these histograms provide evidence that the newly generated data sets approximates the distribution of the original data sets. However, as in the case of the Abalone data set, the data sets generated from only 25 observations of Life Expectancy or

NBA data set do not fully capture the characteristics present in the original data sets. And as expected, increasing the number of data points or observations used as input to the algorithm for generating the new data sets results in greater similarity between the histograms of newly generated data and that of the original data for both Life Expectancy and the NBA data sets.

Boxplot

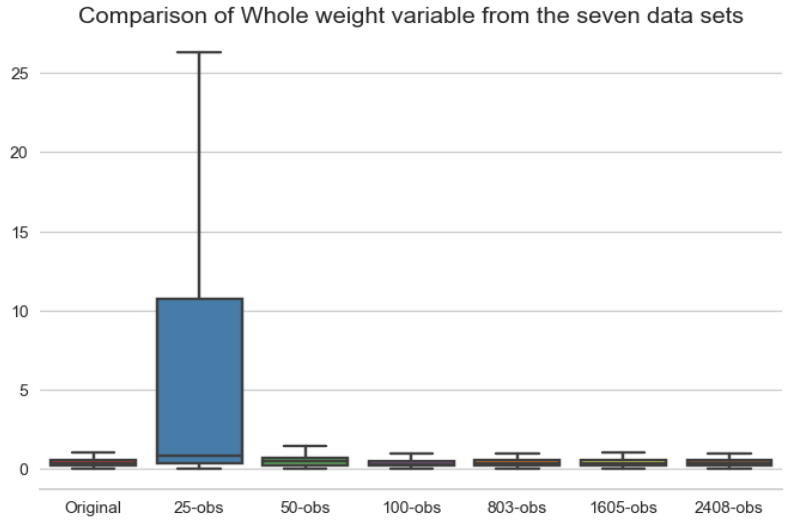
The boxplots in Figure 29 compare the distribution of the fourth, fifth and sixth variables in the original and newly generated Abalone Age data sets.

The boxplots reveal that all of the distributions are highly similar, with the exception of the boxplot derived from the data set generated using 25 observations. The boxplots demonstrate almost identical shapes, central values, spreads and 5-number summaries, highlighting a fair degree of similarity between variables despite minor differences. Furthermore, the dispersion of the boxes and whiskers is comparable across all variables. As for the histograms, as the sample size of the input observations to the algorithm for generating data grows, the boxplots of the variables in the newly generated data become increasingly similar to the boxplots of the variables in the original data set. The boxplot for the sixth variable of the Abalone age data set can be found in Figure 34 in Appendix A.5.1.

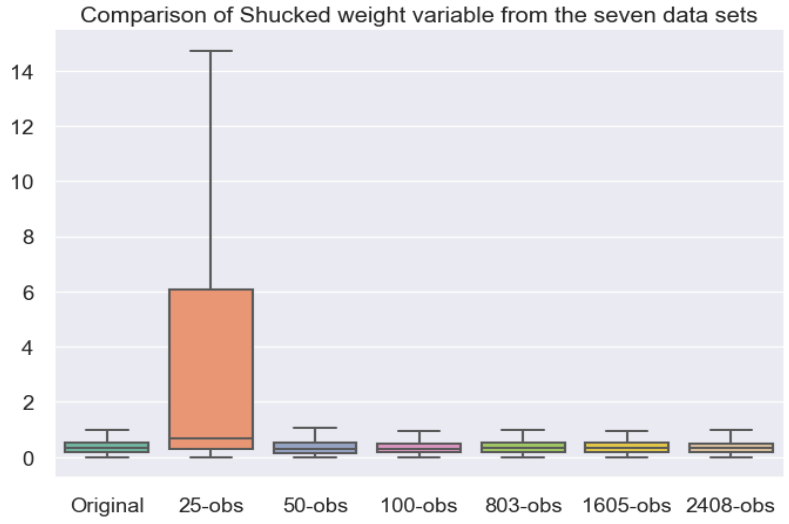
The Boxplots in Figure 39 and 43 of Appendix Section A.5.2 and A.5.3 respectively, compare the distributions of the fourth, fifth and sixth variables in the original and newly generated Life Expectancy and NBA data sets. These comparisons demonstrate a substantial resemblance between the variables of Life Expectancy and NBA data sets in the newly generated and original data sets. In the case of Life Expectancy and NBA data sets, even the boxplot derived from the data set generated using 25 observations exhibits comparable characteristics with the original data sets. This resemblance encompasses the shapes, central values, spreads and 5-number summaries, underscoring a substantial similarity despite minor distinctions.

Correlation

Using the correlation matrices of each newly generated Abalone data set and the original Abalone data set, corresponding dissimilarity matrices were calculated and from these, the Frobenius norm was computed. A value of 2.08, 2.33, 0.46, 0.19, 0.16 and 0.07 was obtained from Abalone age data sets generated from 25, 50, 100, 803, 1605 and 2408 observations of the original Abalone data set, respectively. These values indicate a small difference between the matrices except for the Abalone age



(a)



(b)

Figure 29: Boxplots Comparing the Variables found in the Original Data Set and the Newly Generated Data Set from samples of 25, 50, 100, 803, 1605 and 2408 for the Abalone Age Data Set

data sets generated from 25 and 50 observations, where the Frobenius norm of the differences is relatively larger. The proximity of most of these values to zero suggests that the matrices are fairly similar, indicating that the newly generated data sets closely approximates the original data set in terms of correlation.

The corresponding Frobenius norm values obtained for input samples of size 25, 50, 100, 803, 1605 and 2408 observations for Life Expectancy and NBA data sets were 3.799, 2.400, 1.693, 1.445, 0.769 and 0.635 and 2.968, 1.478, 1.294, 0.866, 0.683 and 0.421. These values underscore the resemblance in correlation matrices between the original and generated data sets. The retention of variable relationships in the new data set is clearly observable for both data sets.

Reconstruction Error

Tables 11 shows the reconstruction error between the original Abalone age data set and newly generated Abalone Age data sets derived from samples of 25 and 50 observations.

Table 9: Reconstruction Error for Data Set Generated from 25 Observations of the Abalone Age Data Set

Columns	**Recon Error**
Length	0.967
Diameter	5.306
Height	2.631
Whole weight	100.013
Shucked weight	55.587
Viscera weight	3.798
Shell weight	53.280
Age	5.814
Sex_F	0.670
Sex_I	0.659
Sex_M	0.679

Table 10: Reconstruction Error for Data Set Generated from 50 Observations of the Abalone Age Data Set

Columns	**Recon Error**
Length	0.294
Diameter	0.303
Height	0.258
Whole weight	0.496
Shucked weight	0.321
Viscera weight	0.322
Shell weight	0.301
Age	3.137
Sex_F	0.644
Sex_I	0.676
Sex_M	0.677

Table 11: Reconstruction Error between the Original and SVD-generated Data Sets For Abalone Age Data Set

The majority of variables in the data sets demonstrate a reconstruction error (i.e., between newly generated Abalone age data sets and the original Abalone age data set) of less than 0.7 except for the Age variable when input sample to the SVD algorithm is 50 and some of the variables from the data set generated using an input of 25 observations. The reconstruction error decreases, across all the variables, as more data points are used as input to the algorithm to generate new data sets (the rest of the tables can be found in Table 28 in Appendix A.5.1). A lower reconstruction error indicates higher quality of reconstruction of the original data set.

For the "Life Expectancy" data set, in Table 40 in Appendix Section A.5.2, most variables exhibited reconstruction errors below 0.6 except for the "Life Expectancy" variable which had a comparatively higher error due to its larger unit of measurement. Similarly, in the NBA data set, in Table 52 in Appendix Section A.5.3,most variables displayed errors below 0.4, reinforcing similarity between new and original data sets. The "target_5yrs" variable ranged from 0.6 to 0.73, confirming replication quality. As for the other previous cases, as the input sample increases for generating data sets, reconstruction errors decreases across variables. This trend, is consistent across both data sets, highlights how reduced error corresponds to heightened precision in capturing original data characteristics.

Kolmogorov–Smirnov (KS) Test

Tables 13 and 12 show calculated p-values from the Kolmogorov-Smirnov test, which were obtained for all the comparisons of variables between the generated Abalone data sets and the original Abalone data set, when the generated data was derived from a sample of 25 and 50 original observations respectively. In all cases the p-values exceeded the significance level of 0.05. This indicates that the null hypothesis, which states that the variables are drawn from the same distribution cannot be rejected.

Table 12: KS Test for Data set Generated from 25 Observations of the Abalone Age Data Set

Column Names	**KS Stat**	**P value**
Length	0.013	0.936
Diameter	0.013	0.936
Height	0.013	0.936
Whole weight	0.013	0.936
Shucked weight	0.013	0.936
Viscera weight	0.013	0.936
Shell weight	0.013	0.936
Age	0.013	0.936
Sex_F	0.013	0.936
Sex_I	0.013	0.936
Sex_M	0.013	0.936

Table 13: KS Test for Data Set Generated from 50 Observations of the Abalone Age Data Set

Column Names	**KS Stat**	**P value**
Length	0.001	1
Diameter	0.001	1
Height	0.001	1
Whole weight	0.001	1
Shucked weight	0.001	1
Viscera weight	0.001	1
Shell weight	0.001	1
Age	0.001	1
Sex_F	0.001	1
Sex_I	0.001	1
Sex_M	0.001	1

Table 14: Kolmogorov–Smirnov (KS) Test between the Original and SVD-generated Data Sets For Abalone Data Set

As a rule of thumb it seems employing 50 or more observations for generating new data sets brings about a notable improvement in the various evaluation measures.

This is illustrated by the fact that the newly generated data sets progressively adopt the distinctive features and variations present in the original data, as highlighted in Figures 33 and 35 in Appendix A.5.1.

In both the Life Expectancy and NBA data sets, the p-values resulting from the KS test, in Table 45 and 57 in Appendix Section A.5.2 and A.5.2 respectively, indicate that comparisons between the generated and original data sets consistently yield values greater than the significance level of 0.05. This implies that the null hypothesis, which suggests no significant difference between the two sets of samples and assumes they are drawn from the same distribution, stands and cannot be rejected.

6.3 Performance of Model Trained on SVD-Generated Data Sets to Original: Case Study

The aim of this section is to assess the similarity in performance between the model trained on the newly generated data sets and on the original data set. To measure this, several evaluation metrics will be employed, as highlighted in Section 5.5. The experiment will follow the methodology outlined in Section 5.6.

6.3.1 Abalone Age Data Set

Table 15 displays the evaluation results of the model trained on the numerous data sets, including MSE, RMSE, R^2 and MAPE. These data sets correspond to both the original one and newly generated ones. The initial row in the table serves as a reference point, showcasing the performance of the model trained on the original data set. Subsequent rows present the performances of models trained on the newly generated data sets. This data will be trained using a Gradient Boosting Regressor, while the Life Expectancy data will be trained using an Extra Trees Regressor and the NBA data will be trained using the Logistic Regression method, as shown in Table 6.

For the model trained on the original data set, following performance outcomes were realised:

- **MSE of 2.600** - Rowe (2018) explains that the MSE score of 2.600 represents the magnitude of squared difference between the predicted and actual Age. A low MSE value, which is near zero, indicates a model that fits the data set exceptionally well. However, in the current instance, the model is deemed to have a moderate fit with the data set, as indicated by the MSE value.
- **RMSE of 1.607** - this is the square root of the mean squared error, the difference between the actual and predicted Age. A RMSE score between 1.0

and 2.0 signifies that the model can reliably predict the data to a satisfactory degree, according to a rule of thumb (Davtalab, 2023). As a result, the model makes reasonable predictions.

- **R^2 of 0.567** - R^2 evaluates how much of the variation in the Age variable is explained by the independent variables of the model. If the realised R^2 value is between 0.5 and 0.7, it is said to explain a good amount of the variation (Allwright, 2022). This is the case for the model, since the independent variables of the model explain 56.7 percent of the variation in the Age variable.
- **MAPE of 11.407%** - A MAPE greater than 10% but less than 25% indicates low, but acceptable accuracy (Swanson, 2015).

Table 15: Experiment Results from the Abalone Age Data Set

Number of Observations Used to Generate a New Data Set	**MSE**	**RMSE**	**R^2**	**MAPE**
Original	2.600	1.607	0.567	11.407
25 Observations	4.986	2.233	0.164	14.576
35 Observations	3.931	1.983	0.341	13.707
50 Observations	3.727	1.931	0.375	13.146
100 Observations	3.695	1.922	0.381	13.573
803 Observations	3.123	1.767	0.477	12.515
1605 Observations	3.188	1.785	0.466	12.214
2408 Observations	2.698	1.643	0.548	11.554

Table 15 illustrates that employing data points generated through Singular Value Decomposition (SVD) can produce outcomes similar to those achieved using the original data set. In the case where the model is trained on data set generated from 25 observations, there is a noticeable decline in performance metrics. Specifically, the MSE, RMSE, R^2 and MAP show degradation of 91.769%, 38.524%, 70.922% and 27.168%, respectively, when compared to the performance of the model trained on the original data set. Moreover, the results obtained across several metrics shows an improvement as more observations are utilised to generate data points using the suggested method. This pattern is consistent across all the evaluation metrics.

6.3.2 Life Expectancy Data Set

Table 16 displays the evaluation results of displays the evaluation results of the model trained on the numerous data sets, including MSE, RMSE, R^2 and MAPE. These data sets correspond to both the original one and newly generated ones. The

initial row in the table serves as a reference point, showcasing the performance of the model trained on the original data set. Subsequent rows present the performances of models trained on the newly generated data sets.

Table 16: Experiment Results from the Life Expectancy Data Set

Number of Observations Used to Generate a New Data Set	**MSE**	**RMSE**	**R^2**	**MAPE**
Original	3.826	1.956	0.858	1.791
25 Observations	12.844	3.584	0.522	3.512
50 Observations	12.622	3.553	0.531	3.689
100 Observations	8.040	2.835	0.701	2.800
138 Observations	6.060	2.462	0.775	2.332
275 Observations	5.543	2.354	0.794	2.144
413 Observations	4.025	2.006	0.850	1.885

For the model trained on the original data set, following performance outcomes were realised:

- **MSE of 3.826** - this is the squared distance between the actual and projected Life Expectancy. If a model has a low MSE value, that is, a number close to 0, it is an ideal fit for the data set (Rowe, 2018). In this instance, the model is deemed to fit the data set moderately so.
- **RMSE of 1.956** - this is the square root of the mean squared error, the difference between the actual and predicted Life Expectancy. A RMSE score between 1.0 and 2.0 signifies that the model can reliably predict the data to a satisfactory degree, according to a rule of thumb (Davtalab, 2023). As a result, the model makes reasonable predictions.
- **R^2 of 0.858** - R^2 evaluates how much of the variation in the Life Expectancy variable is explained by the independent variables of the model. If the realised R^2 value is between 0.75 and 1, it is said to explain a significant proportion of variation (Allwright, 2022). This is the case for the model, since the independent variables of the model explains 85.8 percent of the variation in the Life Expectancy variable.
- **MAPE of 1.791%** - A MAPE less than 5% indicates that the predictions are acceptably accurate (Swanson, 2015).

Table 16 demonstrates that utilising SVD-generated data points can lead to results comparable to those obtained from the original data. The data set derived from

25 observations demonstrates a noticeable decrease in performance across multiple evaluation metrics compared to the model trained on the original data set. Specifically, the MSE shows a degradation of 235.703%, the RMSE of 83.231%, the R^2 of 39.161% and the MAP of 96.092%. Moreover, the results obtained across several metrics shows an improvement as more observations are utilised to generate data points using the suggested method. This pattern is consistent across all the evaluation metrics.

6.3.3 NBA Data Set

Table 17 displays the evaluation results of the model trained on the numerous data sets, including Accuracy, AUC, Recall, Precision and MCC. These data sets correspond to both the original and newly generated data sets. The initial row in the table serves as a reference point, showcasing the performance of the model trained on the original data set. Subsequent rows present the performances of models trained on the newly generated data sets.

Table 17: Experiment Results from the NBA Data Set

Number of Observations Used to Generate New Data Set	**Accuracy**	**AUC**	**Recall**	**Precision**	**MCC**
Original	0.694	0.677	0.761	0.736	0.358
25 Observations	0.605	0.581	0.705	0.660	0.166
50 Observations	0.633	0.615	0.705	0.689	0.231
100 Observations	0.639	0.607	0.773	0.673	0.226
210 Observations	0.667	0.657	0.705	0.729	0.312
420 Observations	0.660	0.643	0.727	0.711	0.288
630 Observations	0.701	0.675	0.807	0.724	0.363

For the model trained on the original data set, following performance outcomes were realised:

- **Accuracy of 0.694**: a metric frequently used to assess the effectiveness of a classification model is known as accuracy. This metric determines the percentage of instances that were classified correctly out of the total instances. Essentially, it emphasises how frequently the model makes accurate predictions about the class. Specifically, in this scenario, the model makes correct predictions approximately 69.4% of the time (Hastie et al., 2008).
- **AUC of 0.677**: is common way to evaluate the effectiveness of a binary classification model is through a metric called the AUC score. This score evaluates the model's ability to distinguish between positive and negative classes. The

AUC score typically ranges from 0 to 1, where a score of 1 indicates perfect performance, meaning that the model correctly predicts all positive and negative instances. On the other hand, a score of 0.5 is equivalent to random guessing, where the model performs no better than chance. In this situation, the performance of the model is superior to that of a model that randomly guesses the classification of instances (Hastie et al., 2008).

- **Recall of 0.761**: is the evaluation of the ability of a classification model to identify all relevant instances in a data set is commonly done using a performance metric known as recall. This metric measures the proportion of positive instances that are correctly identified by the model, expressed as a percentage. When the model has a high recall score, it signifies that it is accurately identifying a significant proportion of positive instances in the data set. In this particular case, the model has a high recall score, indicating that it is correctly identifying a large percentage of positive instances (Hastie et al., 2008).
- **Precision of 0.736**: is a metric used to assess how effectively a model can identify positive instances among all the instances it labels as positive. The calculation involves dividing the number of true positive instances by the sum of true positive and false positive instances. When a model has a high precision score, it means that it can accurately identify true positive instances while minimising the number of false positives. In this specific case, the model exhibits a high precision score, indicating that it is proficient at identifying true positive instances while minimising false positives (Hastie et al., 2008).
- **MCC of 0.358**: is a metric that evaluates the effectiveness of binary (twoclass) classifications. It considers true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) and generates a value ranging from -1 to 1, where 1 signifies a perfect prediction, 0 indicates a random prediction and -1 represents a completely inaccurate prediction. In this case, since the value is 0.358, it implies that the classifier has a moderate positive correlation between predicted and actual classes (Hastie et al., 2008).

Table 17 demonstrates that utilising SVD-generated data points can lead to results comparable to those obtained from the original data. Moreover, the results across different metrics show improvement as more observations are utilised to generate data sets using the suggested method.

The proposed method for generating new data points in situations involving small data sets shows significant potential to enhance the robustness and reliability of research. The results highlight its potential to mitigate the issues associated with

small data sets, facilitating more accurate and comprehensive analyses. Proceeding to the next section, the overall conclusions and implications of the study will be discussed.

7 Conclusions and Recommendations

7.1 Conclusions

This research was concerned with developing an SVD based algorithm for generating new data sets from an already existing data set. The need was motivated by the fact that many times researchers are faced with small data sets or imbalanced data and such an algorithm could be used to generate more data. The performance of the algorithm at generating new data was assessed on a number of criteria using simple statistical tools both quantitative and graphical. The algorithm was assessed on (1) its effectiveness to generate new data from an existing data set such that the new data exhibits the same distributional properties as the original data set, (2) its effectiveness at generating new data sets from small existing data sets and (3) how the performance of regression and classification models compare on the new and original data.

7.1.1 Distribution Similarity

On assessing the effectiveness of the algorithm to generate new data from an existing data set such that the new data exhibits the same distributional properties as the original data the results of the comprehensive evaluation, which includes studying histograms, box plots, analysing reconstruction errors and using the Kolmogorov-Smirnov test, consistently supported a strong similarity between the distribution properties of the data set generated using SVD and those of the original data set. This alignment in distribution highlights the ability of the SVD method to faithfully replicate the inherent characteristics of the original data set. This conclusion remains consistent across the Abalone Age, Life Expectancy and NBA data sets.

7.1.2 Effectiveness of the Algorithm at Generating New Data Sets From Small Data Sets

The comprehensive assessment, involving the analysis of histograms, boxplots, reconstruction error and the Kolmogorov-Smirnov test, consistently demonstrates a notable similarity between the data sets generated through SVD and the original data set. This alignment in distribution underscores the ability of the SVD-based method in accurately replicating the intrinsic characteristics of the original data set, thereby substantiating its capability to generate data points with similar distributions. This outcome remains consistent across the three data sets, namely: Abalone Age, Life Expectancy and NBA data sets. Moreover, the effectiveness increased as the size of sample used to generate new data points grew. It is also important to note that the outcomes were not favorable when 25 observations were used to generate new data points across all the data sets examined.

7.1.3 Performance Similarity

In conclusion, the evaluation of a model trained on data sets generated through Singular Value Decomposition (SVD) demonstrated performance similar to that of the model trained solely on the original data set. This aligns with the research hypothesis that the model trained on SVD-generated data would perform as well as or better than the model trained on observed data. For different data sets—such as Abalone, Life Expectancy and NBA—the models' predictive abilities were assessed using various metrics. Notably, when utilising SVD-generated data, performance outcomes closely resembled those achieved with the original data. Even in cases where data was generated from a lower number of observations, a pattern of improvement emerged as the size of the observation pool increased. In essence, the study suggests that SVD-based data generation can maintain comparable model performance, potentially offering a means to expand data availability for analysis purposes.

7.2 Recommendations

7.2.1 Usuage

Based on the findings from the analysis of the Abalone Age data set, it is evident that achieving a comparable data set in terms of characteristics and distribution across all variables requires a minimum of 50 observations. It was not until this point that the newly generated data closely resembled the original data. Therefore, it is recommended to have a data set size of at least 50 observations when utilising this method.

7.2.2 Limitations of the Proposed Method

This section examines two primary limitations associated with the proposed approach. While the method showcases significant strengths, acknowledging and understanding these limitations is crucial for a comprehensive assessment of its applicability and potential implications.

- Firstly, due to the alteration being confined to the final column of the U matrix, the resultant effect translates to a marginal deviation from the original data. As a consequence, the newly generated data tend to cluster around the patterns present in the original data set. Thus, the proposed method in its current configuration is fairly limited in how novel the newly generated points are.
- Secondly, the efficacy of the proposed method appears to diminish when tasked with generating binary variables. In the study, all variables exhibited notably

low reconstruction errors; however, the binary variable demonstrated a slightly elevated reconstruction error compared to other variables, except for the Life Expectancy and Age variables. While the method still demonstrates proficiency in generating new data points, this observation highlights an area that could benefit from refinement.

7.2.3 Further research

This segment explores potential paths for future research, with the goal of expanding upon the conclusions drawn from the current study. Through the examination of unexplored facets and innovative applications, it aspires to make meaningful contributions to the advancement of the field.

- Subtle Refinements to the Proposed Method
  To address the primary constraints of the study, an alternative approach could be considered. Instead of restricting modifications solely to the last column of the U matrix, a strategy involving the redistribution of the last three or more columns within the U matrix could be pursued. Following this redistribution, the application of the Gram-Schmidt method would serve to establish orthogonality among these columns. Subsequently, the data reconstruction process could be executed using this adjusted U matrix. This adaptation aims to prevent the formation of clusters exclusively around the original dataset, thereby mitigating the aforementioned limitations.
- Application of the Method to Real-World Challenges
  In light of the successful application of this method on Abalone age, Life Expectancy and NBA data sets, it is recommended that further research explore the potential of applying this methodology to a diverse range of data sets from various domains. Investigating the effectiveness and adaptability of the method on different types of data, such as financial, environmental, or social data sets, could yield valuable insights into its generalisability and utility. Additionally, assessing the method's performance across data sets with varying levels of complexity and characteristics could provide a deeper understanding of its strengths and limitations in diverse contexts. This would not only contribute to the method's robustness but also offer valuable implications for its practical application across a wider spectrum of data-driven research and analysis.
- How the proposed method behaves on small and sparse data sets
  Research should be undertaken to fully explore the data generation process, especially in instances involving small and sparse data sets. These data sets are particularly intriguing for several reasons: they often exhibit high dimen-

sionality, which challenges traditional methods and models; they necessitate strategies for improved storage efficiencies and computational speed; and they underscore the importance of careful feature selection to improve model performance. This will provide a thorough comprehension of proposed method, thus deriving essential insights into possible biases, inconsistencies and structural deficiencies within the data.

References

Abu Zohair, Lubna Mahmoud (Aug. 2019). "Prediction of Student's performance by modelling small dataset size". In: *International Journal of Educational Technology in Higher Education* 16. DOI: 10.1186/s41239-019-0160-3.

Ahmed, M. Waqar (Aug. 2023). *Understanding Mean Absolute Error (MAE) in Regression: A Practical Guide*. Medium. URL: https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df.

Allwright, Stephen (Aug. 2022). *What is a good R-Squared value? (simply explained)*. Stephen Allwright. URL: https://stephenallwright.com/good-r-squared-value/.

Andonie, Rzvan (2010). "Computer Sciences Commons, and the Data Science Commons Int". In: *Communications Control* V, pp. 280–291. URL: https://digitalcommons.cwu.edu/cgi/viewcontent.cgi?article=1200&context=cotsfac (visited on 08/31/2023).

Brock, Andrew, Jeff Donahue, and Karen Simonyan (2018). *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. arXiv.org. URL: https://arxiv.org/abs/1809.11096.

Brownlee, Jason (Aug. 2019). *Impact of Dataset Size on Deep Learning Model Skill And Performance Estimates*. Machine Learning Mastery. URL: https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/.

Brunton, Steven and J Kutz (Apr. 2019). "Chapter 1: Singular Value Decomposition". In: Data Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge, p. 3.

Chawla, N. V. et al. (June 2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. DOI: 10.1613/jair.953.

Davtalab, Rahman (Jan. 2023). *What's the acceptable value of Root Mean Square Error (RMSE), Sum of Squares due to error (SSE) and Adjusted R-square?* ResearchGate. URL: https://www.researchgate.net/post/Whats-the-acceptable-value-of-Root-Mean-Square-Error-RMSE-Sum-of-Squares-due-to-error-SSE-and-Adjusted-R-square.

EduPristine (Feb. 2016). *Problems of Small Data and How to Handle Them*. EduPristine. URL: https://www.edupristine.com/blog/managing-small-data.

Gareth, James et al. (2021). *An Introduction to Statistical Learning*. Springer US. DOI: 10.1007/978-1-0716-1418-1.

Goyal, Shweta (July 2021). *Evaluation Metrics for Classification Models*. Analytics Vidhya. URL: https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69.

Gundersen, Gregory (Dec. 2018). *Singular Value Decomposition as Simply as Possible*. gregorygundersen.com. URL: https://gregorygundersen.com/blog/2018/12/10/svd/.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2008). *Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition*. URL: https://hastie.su.domains/Papers/ESLII.pdf.

Hofmann, Thomas (1999). "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*. DOI: 10.1145/312624.312649. URL: https://dl.acm.org/citation.cfm?id=312649.

Jain, Jitesh (June 2020). *The Curse of Dimensionality - Vision and Language Group - Medium*. Medium. URL: https://medium.com/vlgiitr/the-curse-of-dimensionality-15f950e519d2 (visited on 08/11/2024).

Ji, Guoli and Chao Wang (Jan. 2022). "A Denoising Method for Seismic Data Based on SVD and Deep Learning". In: *Applied Sciences* 12, p. 12840. DOI: 10.3390/app122412840. URL: https://www.mdpi.com/2076-3417/12/24/12840 (visited on 04/20/2023).

Jin, Yanghua et al. (Aug. 2017). "Towards the Automatic Anime Characters Creation with Generative Adversarial Networks". In: *arXiv:1708.05509 [cs]*. URL: https://arxiv.org/abs/1708.05509.

Kamath, Chandrika and Ya Ju Fan (Mar. 2018). "Regression with small data sets: a case study using code surrogates in additive manufacturing". In: *Knowledge and Information Systems* 57, pp. 475–493. DOI: 10.1007/s10115-018-1174-1. (Visited on 08/31/2023).

Karras, Tero et al. (2017). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. arXiv.org. URL: https://arxiv.org/abs/1710.10196.

Kumar, Sandeep (Apr. 2021). *How Does Variational Autoencoder Work? Explained!* AITUDE. URL: https://www.aitude.com/how-does-variational-autoencoder-work-explained/.

Lamb, Alex (Feb. 2021). "A Brief Introduction to Generative Models". In: *arXiv:2103.00265 [cs]*. URL: https://arxiv.org/abs/2103.00265.

Lopes, Raul H. C. (2011). "Kolmogorov-Smirnov Test". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 718–720. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_326. URL: https://doi.org/10.1007/978-3-642-04898-2_326.

Maklin, Cory (May 2022). *Synthetic Minority Over-sampling TEchnique (SMOTE)*. Medium. URL: https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c.

"MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)" (2020). In: *Encyclopedia of Production and Manufacturing Management*, pp. 462–462. DOI: 10.1007/1-4020-0612-8_580. URL: https://link.springer.com/referenceworkentry/10.1007%2F1-4020-0612-8_580.

Moody, James (Sept. 2019). *What does RMSE really mean?* Medium. URL: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e.

Natsu (Oct. 2018). *Variational Autoencoder Explained*. Mohit Jain. URL: https://mohitjain.me/2018/10/26/variational-autoencoder/.

Phillips, Bei Wang (2016). *Lecture 18: The SVD: Examples, Norms, Fundamental Subspaces, Compression*. The University of Utah. URL: https://www.sci.utah.edu/~beiwang/teaching/cs6210-fall-2016/lecture18.pdf.

Quy, Tai Le et al. (Mar. 2021). *Data augmentation for dealing with low sampling rates in NILM*. arXiv.org. DOI: 10.48550/arXiv.2104.02055. URL: https://arxiv.org/abs/2104.02055 (visited on 08/31/2023).

Roughgarden, Tim and Gregory Valiant (2021). *CS168: The Modern Algorithmic Toolbox Lecture 9: The Singular Value Decomposition (SVD) and Low-Rank Matrix Approximations*. URL: https://web.stanford.edu/class/cs168/l/l9.pdf.

Rowe, Walker (2018). *Mean Squared Error, R2, and Variance in Regression Analysis*. BMC Blogs. URL: https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/.

Sadek, Rowayda A. (Nov. 2012). "SVD Based Image Processing Applications: State of The Art, Contributions and Research Challenges". In: *arXiv:1211.7102 [cs]*. URL: https://arxiv.org/abs/1211.7102.

Saini, Anshul (Sept. 2021). *Gradient Boosting Algorithm: A Complete Guide for Beginners*. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/.

Sathe, Saket et al. (Nov. 2017). *Kernel-Based Feature Extraction for Collaborative Filtering*. IEEE Xplore. DOI: 10.1109/ICDM.2017.138. URL: https://ieeexplore.ieee.org/document/8215601 (visited on 04/20/2023).

Schmidt, Marius et al. (Mar. 2003). "Application of Singular Value Decomposition to the Analysis of Time-Resolved Macromolecular X-Ray Data". In: *Biophysical Journal* 84, pp. 2112–2129. DOI: 10.1016/S0006-3495(03)75018-8. URL: https://www.sciencedirect.com/science/article/pii/S0006349503750188 (visited on 04/20/2023).

Al-Serw, Nour Al-Rahman (Apr. 2021). *Logistic Regression: the Maths behind it, How It works, and an Example*. Medium. URL: https://medium.com/analytics-vidhya/logistic-regression-the-maths-behind-it-how-it-works-

and-an-example-7d50b778183#:~:text=Logistic%20regression%20uses%20something%20called%20a%20Cross-Entropy%2C%20or.

Shafkat, Irhum (Feb. 2018). *Intuitively Understanding Variational Autoencoders*. Medium. URL: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf.

Shaikhina, Torgyn and Natalia A. Khovanova (Jan. 2017). "Handling limited datasets with neural networks in medical applications: A small-data approach". In: *Artificial Intelligence in Medicine* 75, pp. 51–63. DOI: 10.1016/j.artmed.2016.12.003. (Visited on 11/24/2019).

Shu, Mengying (2019). *Deep Learning for Image Classification on Very Small Datasets Using Transfer Learning*. URL: https://dr.lib.iastate.edu/server/api/core/bitstreams/e72a8f4f-adb3-48d3-9831-0bc95e7e1edb/content (visited on 08/31/2023).

Sisters, Li (May 2020). *Matthews Correlation Coefficient: when to use it and when to avoid it*. Medium. URL: https://towardsdatascience.com/matthews-correlation-coefficient-when-to-use-it-and-when-to-avoid-it-310b3c923f7e.

Solomon, Martin (Sept. 2023). *Matrix Norms: Concepts Applications*. Martin Solomon. URL: https://martinsolomon.io/concepts/matrix-norms/.

Sunitha Reddy, M. and T. Adilakshmi (Jan. 2014). *Music recommendation system based on matrix factorization technique -SVD*. IEEE Xplore. DOI: 10.1109/ICCCI.2014.6921744. URL: https://ieeexplore.ieee.org/document/6921744 (visited on 04/20/2023).

Swanson, David (2015). *UC Riverside UC Riverside Previously Published Works Title On the Relationship among Values of the same Summary Measure of Error when used across Multiple Characteristics at the same point in time: An Examination of MALPE and MAPE Journal Author Publication Date*. URL: https://escholarship.org/content/qt1f71t3x9/qt1f71t3x9.pdf?t=o5wul1.

Thankachan, Karun (Aug. 2022). *What? When? How?: ExtraTrees Classifier*. Medium. URL: https://towardsdatascience.com/what-when-how-extratrees-classifier-c939f905851c.

Wei, Ruoqi et al. (2020). "Variations in Variational Autoencoders - A Comparative Evaluation". In: *IEEE Access* 8, pp. 153651–153670. DOI: 10.1109/access.2020.3018151.

Wisz, M. S. et al. (Sept. 2008). "Effects of sample size on the performance of species distribution models". In: *Diversity and Distributions* 14, pp. 763–773. DOI: 10.1111/j.1472-4642.2008.00482.x.

Wongsuphasawat, Kanit, Yang Liu, and Jeffrey Heer (Nov. 2019). "Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study". In: *arXiv:1911.00568 [cs]*. URL: https://arxiv.org/abs/1911.00568.

Zhang, Han et al. (2016). *StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks*. arXiv.org. URL: https://arxiv.org/abs/1612.03242.

Zhang, He, Vishwanath Sindagi, and Vishal M. Patel (June 2019). "Image De-raining Using a Conditional Generative Adversarial Network". In: *arXiv:1701.05957 [cs]*. URL: https://arxiv.org/abs/1701.05957 (visited on 04/08/2023).

A Appendix

A.1 Abalone Age Data Set EDA

A.1.1 VIF of the variables

Table 18: VIF for the Variables in the Abalone Age Data Set.

Feature	**VIF**
Whole weight	109.48719
Diameter	40.95762
Sex_M	40.09568
Length	39.94141
Sex_F	35.11226
Sex_I	33.70086
Shucked weight	31.95909
Shell weight	23.16848
Viscera weight	17.31054
Height	6.61809
Age	2.03543

A.2 Life Expectancy Data Set EDA

A.2.1 8-Number Summary

Table 19: Eight Number Summary for Life Expectancy Data Set

	Count	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Life expectancy	646.000	71.868	5.392	49.700	68.300	72.700	75.000	89.000
Adult mortality	646.000	136.416	78.120	1.000	81.000	138.000	186.000	411.000
Infant deaths	646.000	4.833	8.195	0.000	0.000	1.000	6.000	52.000
Alcohol	646.000	4.313	3.740	0.010	1.083	3.640	6.640	16.990
Percentage expenditure	646.000	271.752	287.061	0.108	40.681	170.810	430.345	1212.666
Hepatitis B	646.000	90.146	10.392	42.000	87.000	94.000	97.000	99.000
Measles	646.000	52.533	148.478	0.000	0.000	0.000	20.000	926.000
BMI	646.000	43.252	18.171	2.000	29.400	48.500	56.300	77.100
Under-five deaths	646.000	6.149	10.785	0.000	0.000	1.000	6.000	66.000
Polio	646.000	91.768	8.381	57.000	88.000	95.000	98.000	99.000
Total expenditure	646.000	6.011	2.021	0.740	4.700	5.920	7.230	11.970
Diphtheria	646.000	91.851	8.052	62.000	88.000	95.000	98.000	99.000
HIV/AIDS	646.000	0.262	0.319	0.100	0.100	0.100	0.300	1.500
GDP	646.000	2559.968	2389.427	8.376	548.761	1766.948	3956.692	9834.473
Population	646.000	2529741.263	3914647.244	34.000	96210.000	674932.500	3325630.750	18472228.000
Thinness 10-19 years	646.000	3.569	2.951	0.100	1.700	2.400	4.775	15.300
Thinness 5-9 years	646.000	3.578	2.968	0.100	1.600	2.500	4.875	15.200
Income composition of resources	646.000	0.677	0.111	0.385	0.613	0.700	0.745	0.920
Schooling	646.000	12.547	2.164	5.000	11.200	12.700	14.000	19.000

A.2.2 Skewness Table

Table 20: Skewness of the Variable of the Life Expectancy Data Set

Features	**Skewness degree**
Measles	3.890
Under-five deaths	2.703
Infant deaths	2.651
HIV/AIDS	2.277
Population	2.084
Thinness 10-19 years	1.562
Thinness 5-9 years	1.559
Percentage expenditure	1.238
GDP	1.026
Alcohol	0.777
Adult mortality	0.054
Total expenditure	-0.001
Life expectancy	-0.266
Year	-0.285
Schooling	-0.400
Income composition of resources	-0.549
BMI	-0.808
Polio	-1.656
Diphtheria	-1.657
Hepatitis B	-1.858

A.2.3 Bivariate Analysis of Life Expectancy Data Set

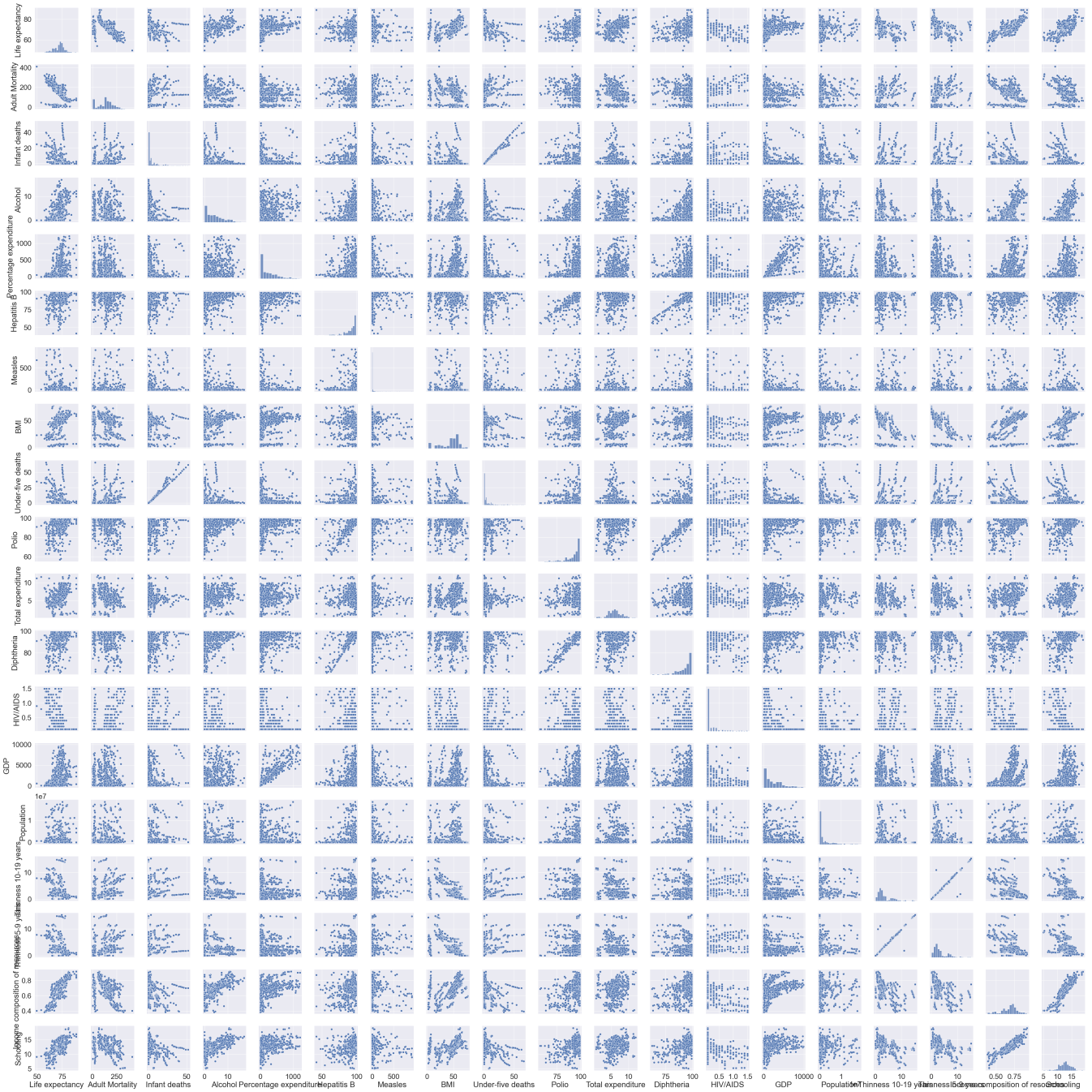


Figure 30: Bivariate Analysis of Life Expectancy Data Sets

A.2.4 VIF of the variables

Table 21: VIF for the Variables in the Life Expectancy Data Set

Feature	**VIF**
Status_Developing	599.033
Status_Developed	95.324
Under-five deaths	55.565
Infant deaths	52.871
Thinness 10-19 years	17.888
Thinness 5-9 years	17.728
Diphtheria	16.515
Polio	13.321
Income composition of resources	11.110
Schooling	5.289
Life expectancy	3.941
GDP	3.841
Percentage expenditure	3.643
Hepatitis B	2.941
Alcohol	2.321
HIV/AIDS	1.696
BMI	1.659
Adult Mortality	1.463
Total expenditure	1.230
Measles	1.155
Population	1.113

A.3 NBA Data Set EDA

A.3.1 8-Number Summary

Table 22: Eight Number Summary for NBA Data Set

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Games played	987.000	57.382	17.139	11.000	44.000	57.000	74.000	82.000
Minutes played	987.000	14.797	5.956	4.100	10.100	14.100	19.000	39.200
Points per game	987.000	5.356	2.631	0.700	3.300	4.800	6.850	15.100
Field goals made	987.000	2.080	1.056	0.300	1.300	1.900	2.700	6.400
Field goals attempts	987.000	4.727	2.244	0.800	3.100	4.300	5.950	13.500
Field goals %	987.000	43.558	5.832	29.100	39.600	43.500	47.400	58.700
3 Points made	987.000	0.187	0.260	0.000	0.000	0.100	0.300	1.000
3 Points attempts	987.000	0.615	0.745	0.000	0.000	0.300	1.000	3.000
3 Points %	987.000	18.684	15.442	0.000	0.000	22.000	32.000	66.700
Free throw made	987.000	1.012	0.579	0.100	0.600	0.900	1.300	3.300
Free throw attempts	987.000	1.433	0.793	0.100	0.800	1.300	1.900	4.200
Free throw %	987.000	70.352	9.317	45.900	64.700	70.700	77.200	93.800
Offensive rebounds	987.000	0.847	0.585	0.000	0.400	0.700	1.200	2.800
Defensive rebounds	987.000	1.636	0.950	0.200	0.900	1.400	2.200	5.000
Rebounds	987.000	2.481	1.462	0.300	1.300	2.100	3.400	7.500
Assists	987.000	1.167	0.865	0.000	0.500	0.900	1.600	4.100
Steals	987.000	0.508	0.283	0.000	0.300	0.500	0.700	1.500
Blocks	987.000	0.274	0.253	0.000	0.100	0.200	0.400	1.100
Turnover	987.000	0.964	0.452	0.100	0.600	0.900	1.200	2.700
Target_5yrs	987.000	0.560	0.497	0.000	0.000	1.000	1.000	1.000

A.3.2 Skewness Table

Table 23: Skewness of the Variable of the NBA Data Set

Features	**Skewness Degree**
3 Points made	1.405
Blocks	1.286
3 Points attempts	1.249
Assists	1.147
Field goals attempts	1.034
Field goals made	1.024
Free throw made	1.021
Points per game	0.976
Offensive rebounds	0.937
Rebounds	0.918
Defense rebounds	0.914
Free throw attempts	0.908
Steals	0.847
Turnover	0.745
Minutes played	0.562
Field Goal %	0.042
3 Points %	0.002
Free throw %	-0.238
target_5yrs	-0.243
Games Played	-0.253

A.3.3 Bivariate Analysis

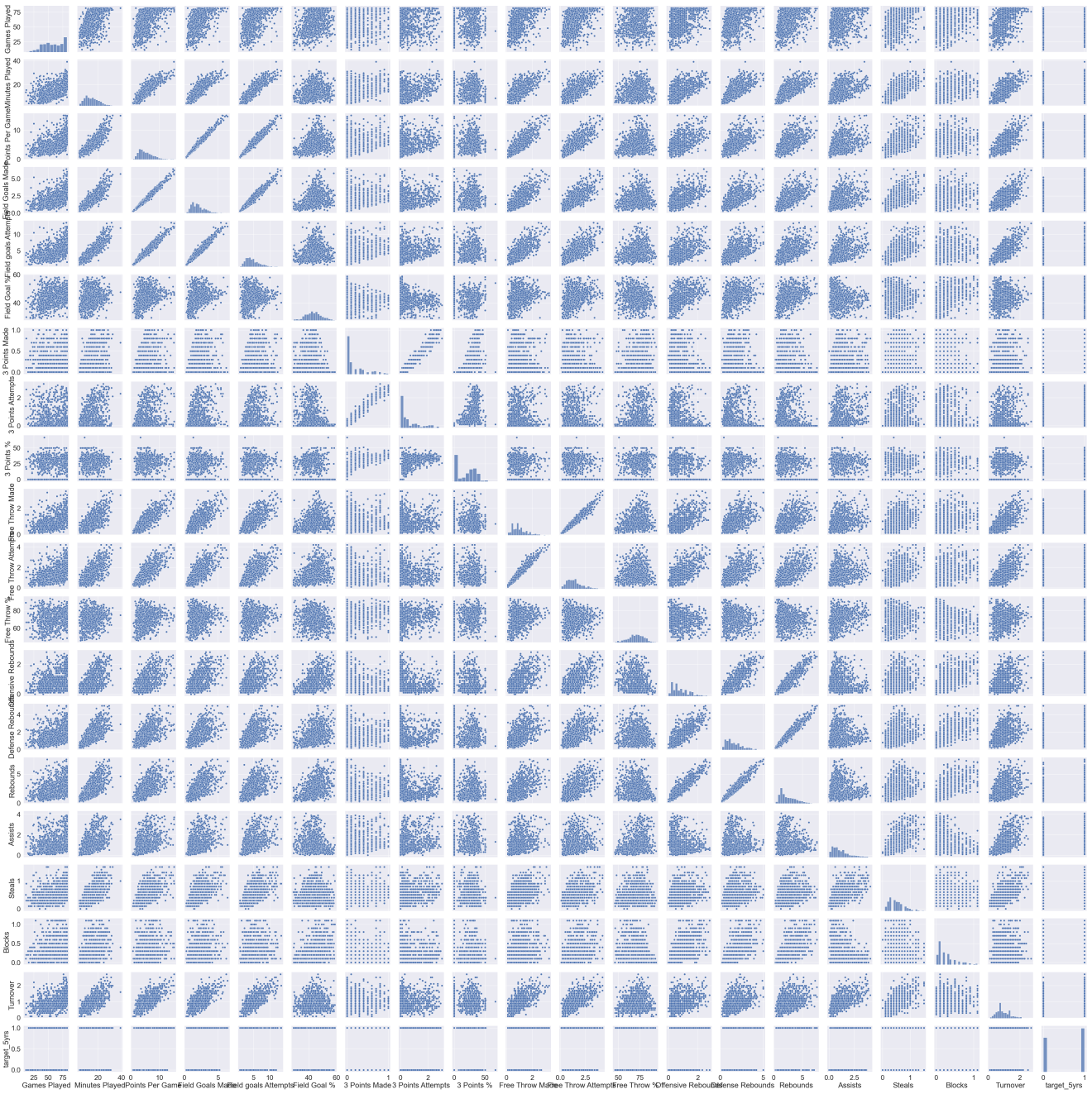


Figure 31: Bivariate Analysis of NBA Data sets

A.4 Mathematical Formulations of Evaluation Metrics

A.4.1 MSE:

$$MSE = \frac{1}{n}\sum(y - \hat{y})^2$$

A.4.2 RMSE:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

A.4.3 R2:

$R^2 = 1 - \frac{RSS}{TSS}$
R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares
$RSS = \sum(y_i - \hat{y}_i)^2$, where: y_i is actual value and, $\hat{y}_i$ is the predicted value.
$TSS = \sum(y_i - \bar{y})^2$
where: y_i is the actual value and $\bar{y}$ is the mean value of the variable/feature

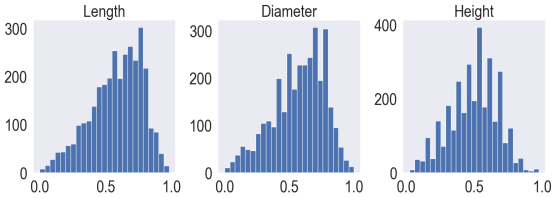
A.4.4 MAPE:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} |\frac{Y_i - \hat{Y_i}}{Y_i}|$$

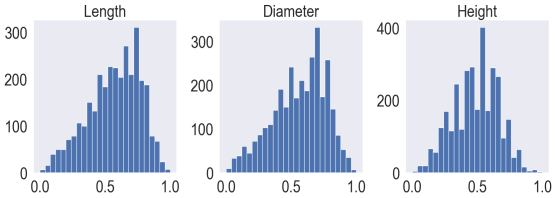
A.5 Results

A.5.1 Abalone Age Data Set

Histogram

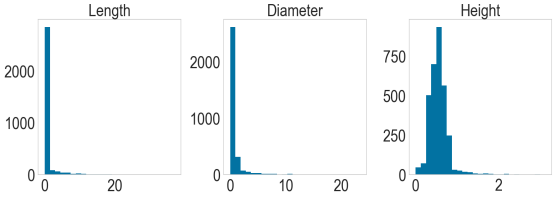


(a) Data Set Generated from 1605 Observations using the Proposed Model

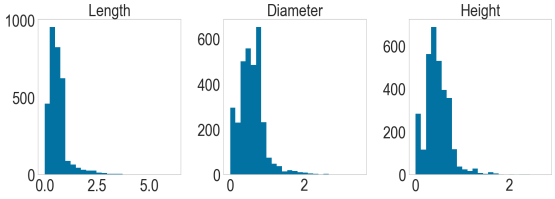


(b) Data Set Generated from 2408 Observations using the Proposed Model

Figure 32: Histogram of the Remaining Data Sets



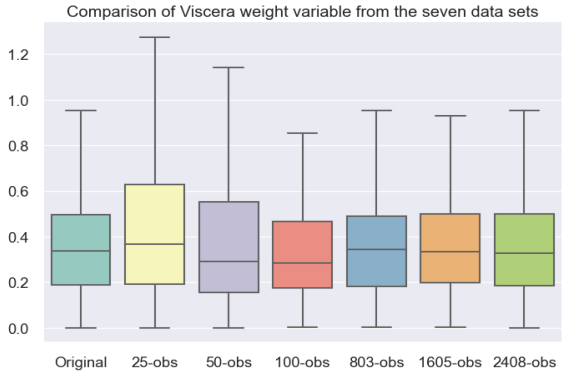
(a) Data Set Generated from 30 Observations using the Proposed Model



(b) Data Set Generated from 35 Observations using the Proposed Model

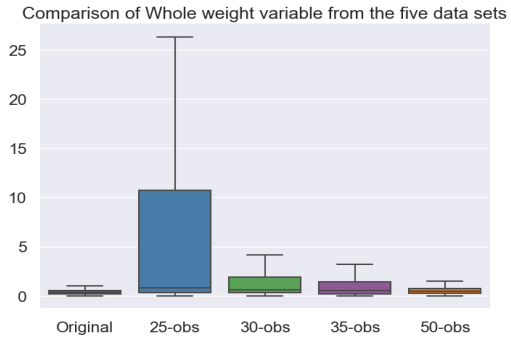
Figure 33: Histogram of the 30- and 35-obs generated data sets

Boxplots

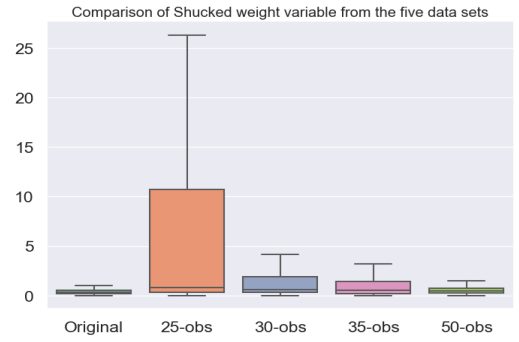


(a)

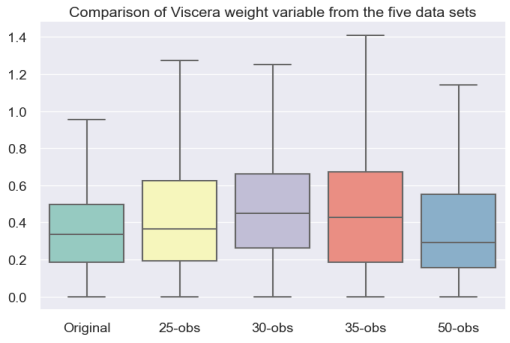
Figure 34: The Remaining Boxplot



(a)



(b)



(c)

Figure 35: Threshold boxplots

Reconstruction Error

Table 24: Reconstruction Error for Data Set Generated from 100 Observations of the Abalone Age Data Set

Columns	**Recon Error**
Length	0.277
Diameter	0.283
Height	0.244
Whole weight	0.295
Shucked weight	0.299
Viscera weight	0.282
Shell weight	0.273
Age	3.274
Sex_F	0.639
Sex_I	0.679
Sex_M	0.662

Table 25: Reconstruction Error for Data Set Generated from 803 Observations of the Abalone Age Data Set

Columns	**Recon Error**
Length	0.288
Diameter	0.295
Height	0.253
Whole weight	0.306
Shucked weight	0.310
Viscera weight	0.296
Shell weight	0.286
Age	3.324
Sex_F	0.639
Sex_I	0.671
Sex_M	0.685

Table 26: Reconstruction Error for Data Set Generated from 1605 Observations of the Abalone Age Data Set

Columns	**Recon Error**
Length	0.285
Diameter	0.294
Height	0.252
Whole weight	0.302
Shucked weight	0.303
Viscera weight	0.290
Shell weight	0.285
Age	3.226
Sex_F	0.653
Sex_I	0.657
Sex_M	0.682

Table 27: Reconstruction Error for Data Set Generated from 2408 Observations of the Abalone Age Data Set

Columns	**Recon Error**
Length	0.284
Diameter	0.293
Height	0.251
Whole weight	0.302
Shucked weight	0.301
Viscera weight	0.294
Shell weight	0.284
Age	3.280
Sex_F	0.658
Sex_I	0.661
Sex_M	0.669

Table 28: The Remaining Reconstruction Error Tables For Abalone Age Data Set

Kolmogorov-smirnov (KS) test

Table 29: KS Test for Data set Generated from 100 Observations of the Abalone Age Data Set

Column Names	KS Stat	P value
Length	0.033	0.064
Diameter	0.033	0.064
Height	0.033	0.064
Whole weight	0.033	0.064
Shucked weight	0.033	0.064
Viscera weight	0.033	0.064
Shell weight	0.033	0.064
Age	0.033	0.064
Sex_F	0.033	0.064
Sex_I	0.033	0.064
Sex_M	0.033	0.064

Table 30: KS Test for Data set Generated from 803 Observations of the Abalone Age Data Set

Column Names	KS Stat	P value
Length	0.005	1
Diameter	0.005	1
Height	0.005	1
Whole weight	0.005	1
Shucked weight	0.005	1
Viscera weight	0.005	1
Shell weight	0.005	1
Age	0.005	1
Sex_F	0.005	1
Sex_I	0.005	1
Sex_M	0.005	1

Table 31: KS Test for Data set Generated from 1605 Observations of the Abalone Age Data Set

Column Names	KS Stat	P value
Length	0.006	1
Diameter	0.006	1
Height	0.006	1
Whole weight	0.006	1
Shucked weight	0.006	1
Viscera weight	0.006	1
Shell weight	0.006	1
Age	0.006	1
Sex_F	0.006	1
Sex_I	0.006	1
Sex_M	0.006	1.

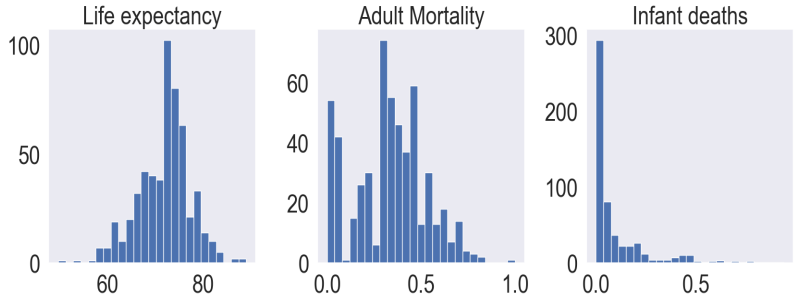
Table 32: KS Test for Data set Generated from 2408 Observations of the Abalone Age Data Set

Column Names	KS Stat	P value
Length	0.004	1
Diameter	0.004	1
Height	0.004	1
Whole weight	0.004	1
Shucked weight	0.004	1
Viscera weight	0.004	1
Shell weight	0.004	1
Age	0.004	1
Sex_F	0.004	1
Sex_I	0.004	1
Sex_M	0.004	1

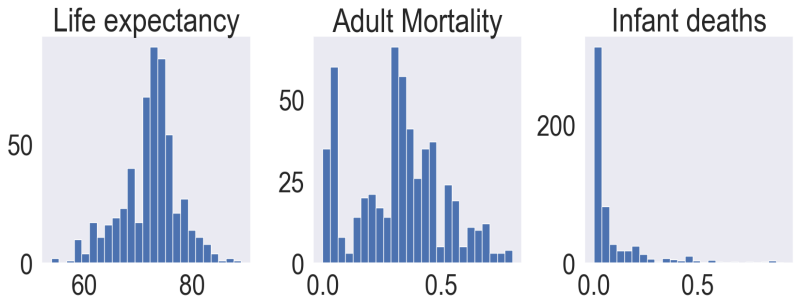
Table 33: The Remaining KS-test Tables For Abalone Age Data Set

A.5.2 Life Expectancy Data Set

Objective 1

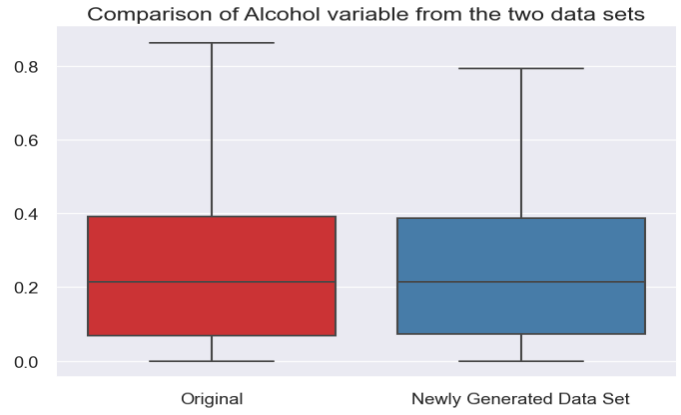


(a) Original Data Set

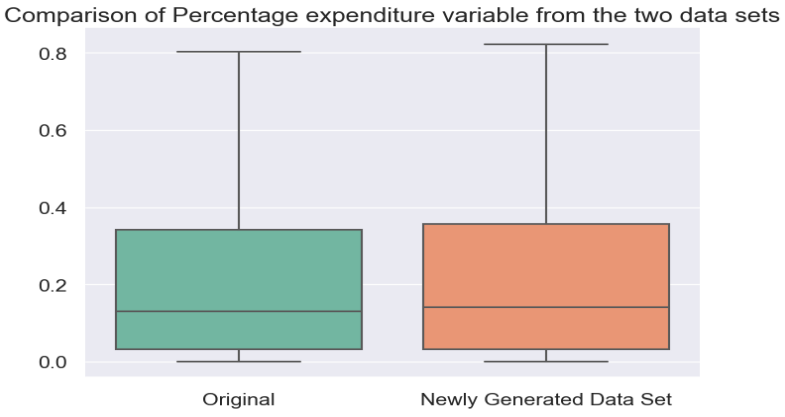


(b) Data set Generated from 550 Observations using the Proposed Model

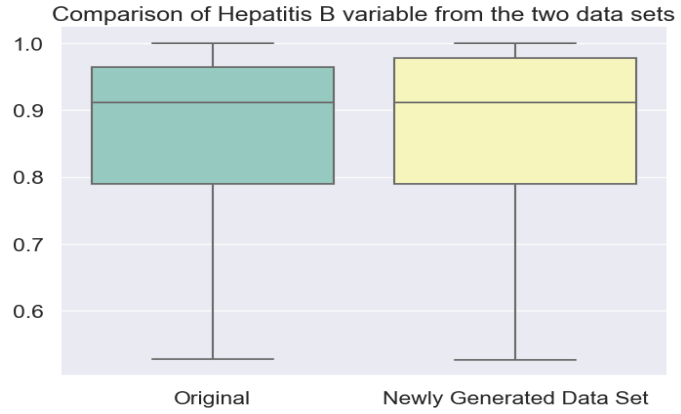
Figure 36: Histogram of the First Three Variables from the Life Expectancy Data Set and the New Generated Data Set Derived from a Subset of the Life Expectancy Data Set



(a)



(b)



(c)

Figure 37: Boxplots Comparing the Variables found in the Original Data Set and the Newly Generated Data Set For the Life Expectancy Data Set

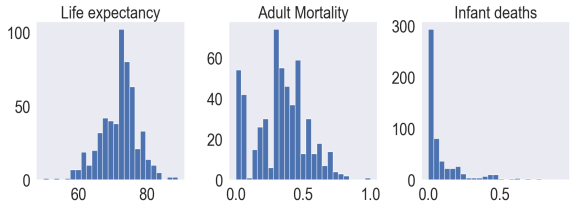
Table 34: Reconstruction Error between the Original and SVD-generated Data Set using 550 Observations For Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	7.325
Adult Mortality	0.269
Infant deaths	0.223
Alcohol	0.298
Percentage expenditure	0.324
Hepatitis B	0.243
Measles	0.256
BMI	0.365
Under-five deaths	0.236
Polio	0.274
Total expenditure	0.253
Diphtheria	0.300
HIV/AIDS	0.300
GDP	0.330
Population	0.301
Thinness 10-19 years	0.296
Thinness 5-9 years	0.297
Income composition of resources	0.290
Schooling	0.215
Status_Developed	0.474
Status_Developing	0.474

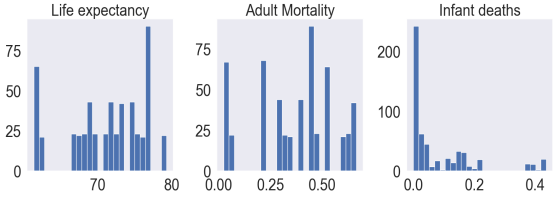
Table 35: Kolmogorov–Smirnov (KS) Test between the Original and SVD-generated Data Sets using 550 Observations For Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.002	1.0
Adult Mortality	0.002	1.0
Infant deaths	0.002	1.0
Alcohol	0.002	1.0
Percentage expenditure	0.002	1.0
Hepatitis B	0.002	1.0
Measles	0.002	1.0
BMI	0.002	1.0
Under-five deaths	0.002	1.0
Polio	0.002	1.0
Total expenditure	0.002	1.0
Diphtheria	0.002	1.0
HIV/AIDS	0.002	1.0
GDP	0.002	1.0
Population	0.002	1.0
Thinness 10-19 years	0.002	1.0
Thinness 5-9 years	0.002	1.0
Income composition of resources	0.002	1.0
Schooling	0.002	1.0
Status_Developed	0.002	1.0
Status_Developing	0.002	1.0

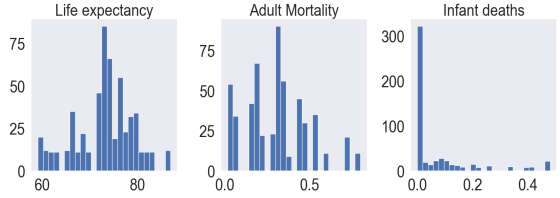
Objective 2 Histogram



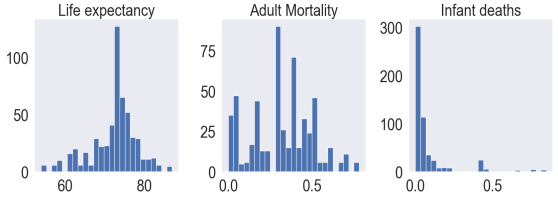
(a) Original Data Set



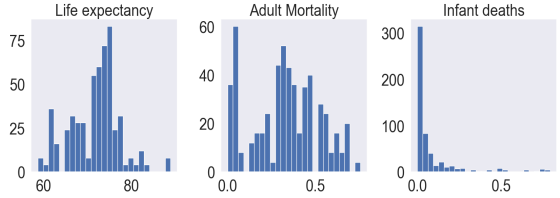
(b) Data set Generated from 25 Observations using the Proposed Model



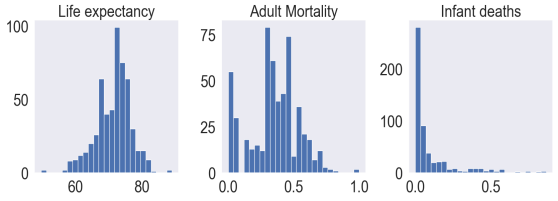
(c) Data set Generated from 50 Observations using the Proposed Model



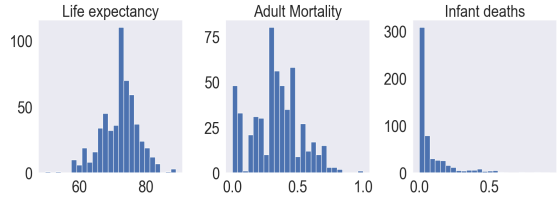
(d) Data set Generated from 100 Observations using the Proposed Model



(e) Data set generated from 138 Observations using the Proposed Model



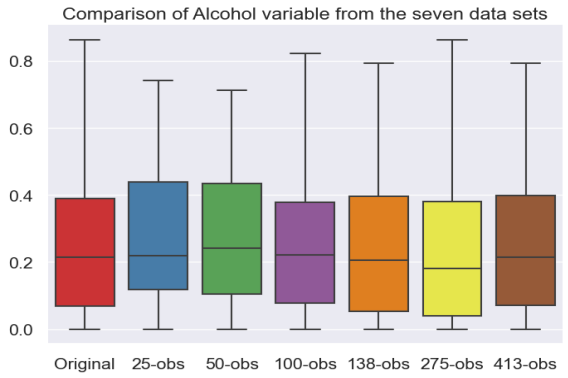
(f) Data Set Generated from 275 Observations using the Proposed Model



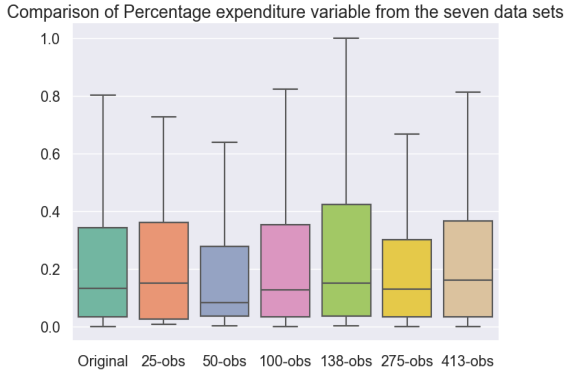
(g) Data Set Generated from 413 Observations using the Proposed Model

Figure 38: Distributions of the first Three Variables in the Original Life Expectancy Data set and Generated Life Expectancy Data Sets Derived from samples of 25, 50, 100, 138, 275 and 413

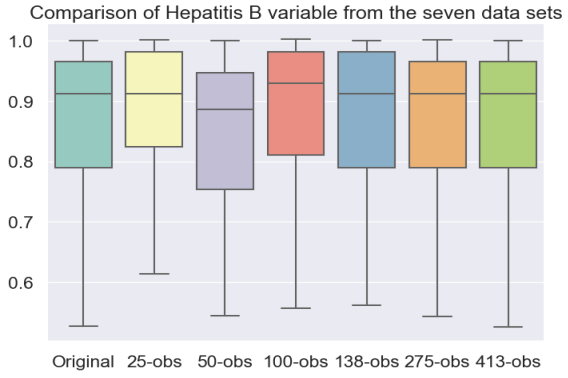
Boxplots



(a)



(b)



(c)

Figure 39: Boxplots Comparing the Variables found in the Original Data Set and the Newly Generated Data Set from Samples of 25, 50, 100, 138, 275 and 413 for the Life Expectancy Data

Reconstruction Error

Table 36: Reconstruction Error for Data Set Generated from 25 Observations of the Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	7.758
Adult Mortality	0.284
Infant deaths	0.201
Alcohol	0.307
Percentage expenditure	0.317
Hepatitis B	0.239
Measles	0.285
BMI	0.319
Under-five deaths	0.218
Polio	0.257
Total expenditure	0.281
Diphtheria	0.284
HIV/AIDS	0.343
GDP	0.360
Population	0.319
Thinness 10-19 years	0.294
Thinness 5-9 years	0.283
Income composition of resources	0.316
Schooling	0.241
Status_Developed	0.458
Status_Developing	0.458

Table 37: Reconstruction Error for Data Set Generated from 50 Observations of the Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	8.551
Adult Mortality	0.272
Infant deaths	0.200
Alcohol	0.307
Percentage expenditure	0.348
Hepatitis B	0.232
Measles	0.214
BMI	0.331
Under-five deaths	0.209
Polio	0.255
Total expenditure	0.263
Diphtheria	0.274
HIV/AIDS	0.352
GDP	0.354
Population	0.257
Thinness 10-19 years	0.295
Thinness 5-9 years	0.299
Income composition of resources	0.332
Schooling	0.230
Status_Developed	0.522
Status_Developing	0.522

Table 37: Reconstruction Error for Data Set Generated from 100 Observations of the Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	8.074
Adult Mortality	0.268
Infant deaths	0.224
Alcohol	0.310
Percentage expenditure	0.340
Hepatitis B	0.228
Measles	0.198
BMI	0.355
Under-five deaths	0.241
Polio	0.250
Total expenditure	0.256
Diphtheria	0.281
HIV/AIDS	0.313
GDP	0.348
Population	0.277
Thinness 10-19 years	0.300
Thinness 5-9 years	0.297
Income composition of resources	0.305
Schooling	0.215
Status_Developed	0.523
Status_Developing	0.523

Table 38: Reconstruction Error for Data Set Generated from 138 Observations of the Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	8.038
Adult Mortality	0.263
Infant deaths	0.226
Alcohol	0.331
Percentage expenditure	0.346
Hepatitis B	0.260
Measles	0.233
BMI	0.335
Under-five deaths	0.240
Polio	0.286
Total expenditure	0.261
Diphtheria	0.317
HIV/AIDS	0.335
GDP	0.346
Population	0.305
Thinness 10-19 years	0.278
Thinness 5-9 years	0.282
Income composition of resources	0.316
Schooling	0.243
Status_Developed	0.504
Status_Developing	0.504

Table 38: Reconstruction Error for Data Set Generated from 275 Observations of the Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	7.521
Adult Mortality	0.254
Infant deaths	0.228
Alcohol	0.298
Percentage expenditure	0.327
Hepatitis B	0.246
Measles	0.249
BMI	0.339
Under-five deaths	0.237
Polio	0.276
Total expenditure	0.248
Diphtheria	0.300
HIV/AIDS	0.314
GDP	0.327
Population	0.270
Thinness 10-19 years	0.272
Thinness 5-9 years	0.272
Income composition of resources	0.289
Schooling	0.204
Status_Developed	0.450
Status_Developing	0.450

Table 39: Reconstruction Error for Data Set Generated from 413 Observations of the Life Expectancy Data Set

Columns	**Recon Error**
Life expectancy	7.745
Adult Mortality	0.273
Infant deaths	0.220
Alcohol	0.316
Percentage expenditure	0.344
Hepatitis B	0.245
Measles	0.246
BMI	0.344
Under-five deaths	0.228
Polio	0.279
Total expenditure	0.274
Diphtheria	0.298
HIV/AIDS	0.308
GDP	0.337
Population	0.316
Thinness 10-19 years	0.274
Thinness 5-9 years	0.277
Income composition of resources	0.304
Schooling	0.221
Status_Developed	0.504
Status_Developing	0.504

Table 40: Reconstruction Error between the Original and SVD-generated Data Sets for Life Expectancy Data Set

Kolmogorov-smirnov (KS) test

Table 41: KS Test for Data set Generated from 25 Observations of the Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.062	0.244
Adult Mortality	0.062	0.244
Infant deaths	0.062	0.244
Alcohol	0.062	0.244
Percentage expenditure	0.062	0.244
Hepatitis B	0.062	0.244
Measles	0.062	0.244
BMI	0.062	0.244
Under-five deaths	0.062	0.244
Polio	0.062	0.244
Total expenditure	0.062	0.244
Diphtheria	0.062	0.244
HIV/AIDS	0.062	0.244
GDP	0.062	0.244
Population	0.062	0.244
Thinness 10-19 years	0.062	0.244
Thinness 5-9 years	0.062	0.244
Income composition of resources	0.062	0.244
Schooling	0.062	0.244
Status_Developed	0.062	0.244
Status_Developing	0.062	0.244

Table 42: KS Test for Data set Generated from 50 Observations of the Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.042	0.723
Adult Mortality	0.042	0.723
Infant deaths	0.042	0.723
Alcohol	0.042	0.723
Percentage expenditure	0.042	0.723
Hepatitis B	0.042	0.723
Measles	0.042	0.723
BMI	0.042	0.723
Under-five deaths	0.042	0.723
Polio	0.042	0.723
Total expenditure	0.042	0.723
Diphtheria	0.042	0.723
HIV/AIDS	0.042	0.723
GDP	0.042	0.723
Population	0.042	0.723
Thinness 10-19 years	0.042	0.723
Thinness 5-9 years	0.042	0.723
Income composition of resources	0.042	0.723
Schooling	0.042	0.723
Status_Developed	0.042	0.723
Status_Developing	0.042	0.723

Table 42: KS Test for Data set Generated from 100 Observations of the Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.009	1.0
Adult Mortality	0.009	1.0
Infant deaths	0.009	1.0
Alcohol	0.009	1.0
Percentage expenditure	0.009	1.0
Hepatitis B	0.009	1.0
Measles	0.009	1.0
BMI	0.009	1.0
Under-five deaths	0.009	1.0
Polio	0.009	1.0
Total expenditure	0.009	1.0
Diphtheria	0.009	1.0
HIV/AIDS	0.009	1.0
GDP	0.009	1.0
Population	0.009	1.0
Thinness 10-19 years	0.009	1.0
Thinness 5-9 years	0.009	1.0
Income composition of resources	0.009	1.0
Schooling	0.009	1.0
Status_Developed	0.009	1.0
Status_Developing	0.009	1.0

Table 43: KS Test for Data set Generated from 138 Observations of the Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.013	1.0
Adult Mortality	0.013	1.0
Infant deaths	0.013	1.0
Alcohol	0.013	1.0
Percentage expenditure	0.013	1.0
Hepatitis B	0.013	1.0
Measles	0.013	1.0
BMI	0.013	1.0
Under-five deaths	0.013	1.0
Polio	0.013	1.0
Total expenditure	0.013	1.0
Diphtheria	0.013	1.0
HIV/AIDS	0.013	1.0
GDP	0.013	1.0
Population	0.013	1.0
Thinness 10-19 years	0.013	1.0
Thinness 5-9 years	0.013	1.0
Income composition of resources	0.013	1.0
Schooling	0.013	1.0
Status_Developed	0.013	1.0
Status_Developing	0.013	1.0

Table 43: KS Test for Data set Generated from 275 Observations of the Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.027	0.987
Adult Mortality	0.027	0.987
Infant deaths	0.027	0.987
Alcohol	0.027	0.987
Percentage expenditure	0.027	0.987
Hepatitis B	0.027	0.987
Measles	0.027	0.987
BMI	0.027	0.987
Under-five deaths	0.027	0.987
Polio	0.027	0.987
Total expenditure	0.027	0.987
Diphtheria	0.027	0.987
HIV/AIDS	0.027	0.987
GDP	0.027	0.987
Population	0.027	0.987
Thinness 10-19 years	0.027	0.987
Thinness 5-9 years	0.027	0.987
Income composition of resources	0.027	0.987
Schooling	0.027	0.987
Status_Developed	0.027	0.987
Status_Developing	0.027	0.987

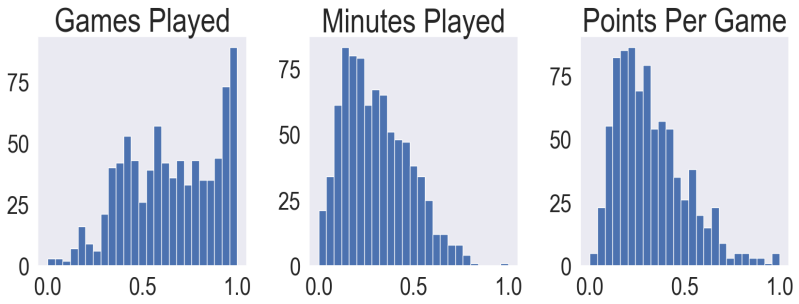
Table 44: KS Test for Data set Generated from 413 Observations of the Life Expectancy Data Set

Column Names	**KS Stat**	**P value**
Life expectancy	0.013	1.0
Adult Mortality	0.013	1.0
Infant deaths	0.013	1.0
Alcohol	0.013	1.0
Percentage expenditure	0.013	1.0
Hepatitis B	0.013	1.0
Measles	0.013	1.0
BMI	0.013	1.0
Under-five deaths	0.013	1.0
Polio	0.013	1.0
Total expenditure	0.013	1.0
Diphtheria	0.013	1.0
HIV/AIDS	0.013	1.0
GDP	0.013	1.0
Population	0.013	1.0
Thinness 10-19 years	0.013	1.0
Thinness 5-9 years	0.013	1.0
Income composition of resources	0.013	1.0
Schooling	0.013	1.0
Status_Developed	0.013	1.0
Status_Developing	0.013	1.0

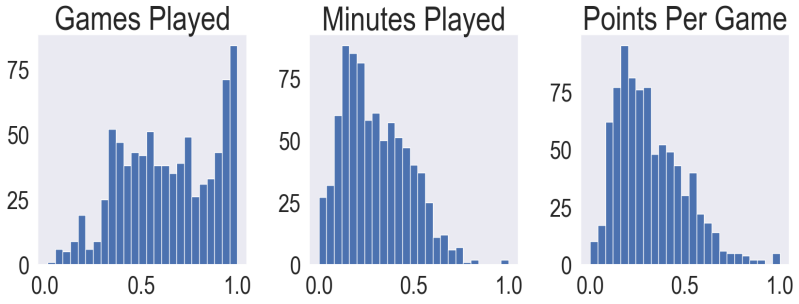
Table 45: Kolmogorov–Smirnov (KS) Test between the Original and SVD-generated Data Sets for Life Expectancy Data Set

A.5.3 NBA Data Set

Objective 1

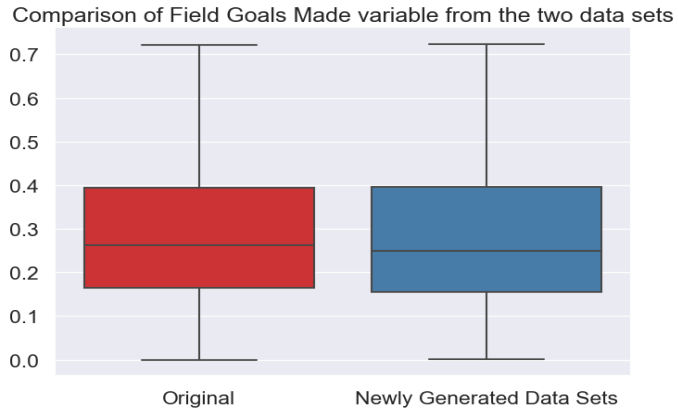


(a) Original Data Set

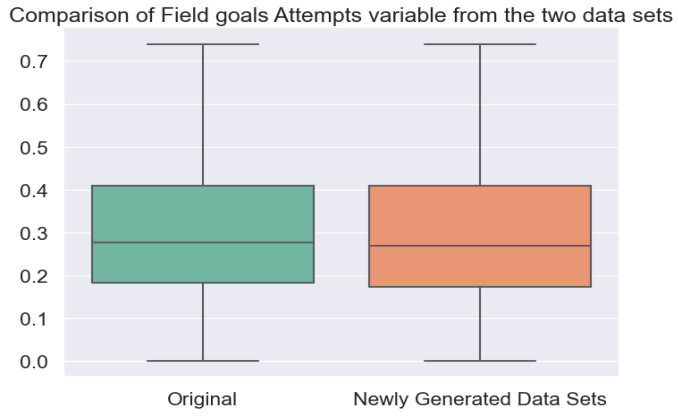


(b) Data Set Generated from 840 Observations using the Proposed Model

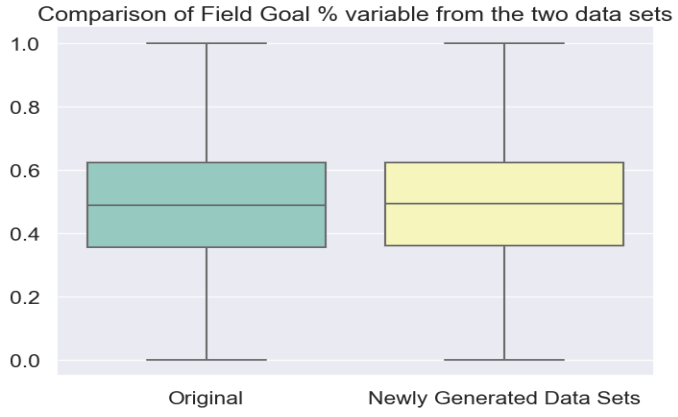
Figure 40: Histogram of the First Three Variables from the NBA Data Set and the New Generated Data Set Derived from a Subset of the NBA Data Set



(a)



(b)



(c)

Figure 41: Boxplots Comparing the Variables Found in the Original Data Set and the Newly Generated Data Set For NBA Data Set

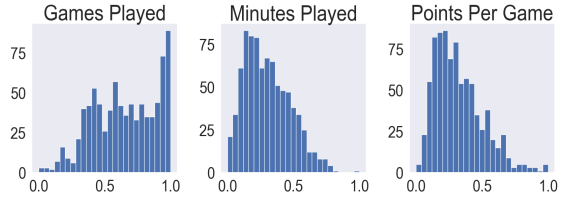
Table 46: Reconstruction Error between the Original and SVD-generated Data Set using 840 Observations For NBA Data Set

Columns	**Recon Error**
Games Played	0.358
Minutes Played	0.241
Points Per Game	0.255
Field Goals Made	0.243
Field goals Attempts	0.247
Field Goal %	0.277
3 Points Made	0.355
3 Points Attempts	0.339
3 Points %	0.341
Free Throw Made	0.255
Free Throw Attempts	0.274
Free Throw %	0.268
Offensive Rebounds	0.301
Defense Rebounds	0.291
Rebounds	0.298
Assists	0.295
Steals	0.265
Blocks	0.324
Turnover	0.238
target_5yrs	0.719

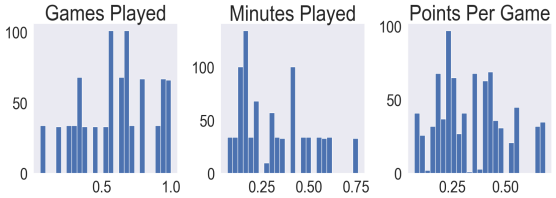
Table 47: Kolmogorov–Smirnov (KS) Test between the Original and SVD-generated Data Set using 840 Observations For NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.029	0.883
Minutes Played	0.029	0.883
Points Per Game	0.029	0.883
Field Goals Made	0.029	0.883
Field goals Attempts	0.029	0.883
Field Goal %	0.029	0.883
3 Points Made	0.029	0.883
3 Points Attempts	0.029	0.883
3 Points %	0.029	0.883
Free Throw Made	0.029	0.883
Free Throw Attempts	0.029	0.883
Free Throw %	0.029	0.883
Offensive Rebounds	0.029	0.883
Defense Rebounds	0.029	0.883
Rebounds	0.029	0.883
Assists	0.029	0.883
Steals	0.029	0.883
Blocks	0.029	0.883
Turnover	0.029	0.883
target_5yrs	0.029	0.883

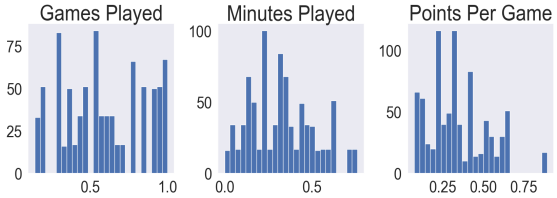
Objective 2 Histogram



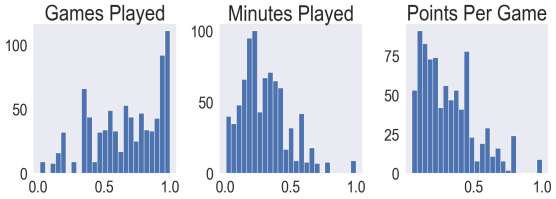
(a) Original Data Set



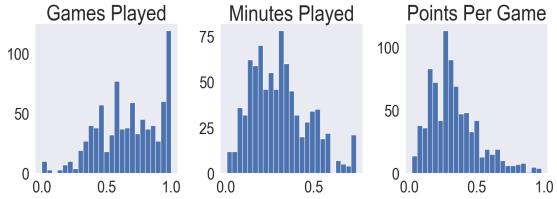
(b) Data Set Generated from 25 Observations using the Proposed Model



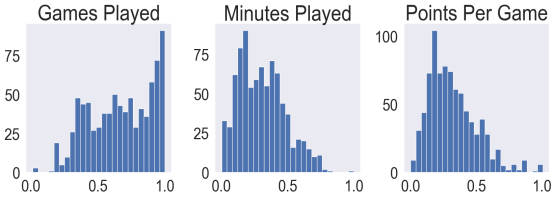
(c) Data Set Generated from 50 Observations using the Proposed Model



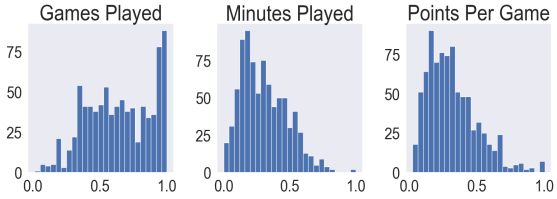
(d) Data Set Generated from 100 Observations using the Proposed Model



(e) Data Set Generated from 210 Observations using the Proposed Model



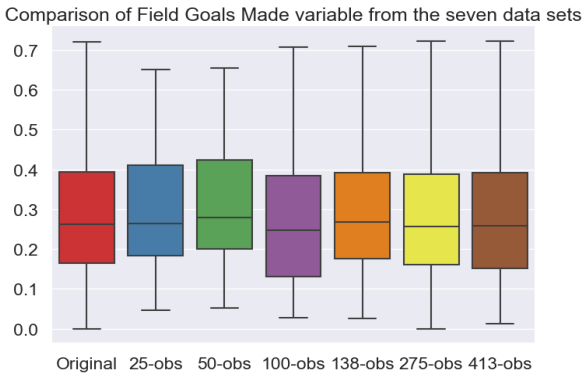
(f) Data Set Generated from 420 Observations using the Proposed Model



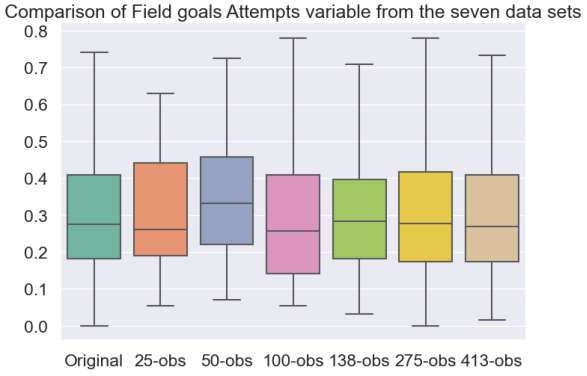
(g) Data Set Generated from 630 Observations using the Proposed Model

Figure 42: Distributions of the first Three Variables in the Original NBA Data set and Generated NBA Data Sets Derived from samples of 25, 50, 100, 210, 420 and 630

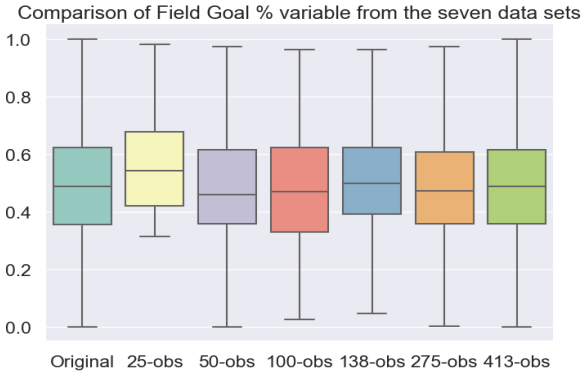
Boxplots



(a)



(b)



(c)

Figure 43: Boxplots Comparing the Variables found in the Original Data Set and the Newly Generated Data Set from samples of 25, 50, 100, 138, 275 and 413 for the NBA Data Set

Reconstruction Error

Table 48: Reconstruction Error for Data Set Generated from 25 Observations of the NBA Data Set

Columns	**Recon Error**
Games Played	0.364
Minutes Played	0.255
Points Per Game	0.249
Field Goals Made	0.238
Field goals Attempts	0.241
Field Goal %	0.285
3 Points Made	0.355
3 Points Attempts	0.316
3 Points %	0.328
Free Throw Made	0.251
Free Throw Attempts	0.259
Free Throw %	0.256
Offensive Rebounds	0.256
Defense Rebounds	0.292
Rebounds	0.275
Assists	0.249
Steals	0.289
Blocks	0.328
Turnover	0.222
target_5yrs	0.705

Table 49: Reconstruction Error for Data Set Generated from 50 Observations of the NBA Data Set

Columns	**Recon Error**
Games Played	0.373
Minutes Played	0.253
Points Per Game	0.256
Field Goals Made	0.243
Field goals Attempts	0.239
Field Goal %	0.281
3 Points Made	0.385
3 Points Attempts	0.368
3 Points %	0.322
Free Throw Made	0.262
Free Throw Attempts	0.278
Free Throw %	0.268
Offensive Rebounds	0.298
Defense Rebounds	0.285
Rebounds	0.294
Assists	0.335
Steals	0.289
Blocks	0.326
Turnover	0.271
target_5yrs	0.718

Table 49: Reconstruction Error for Data Set Generated from 100 Observations of the NBA Data Set

Columns	**Recon Error**
Games Played	0.361
Minutes Played	0.246
Points Per Game	0.268
Field Goals Made	0.256
Field goals Attempts	0.256
Field Goal %	0.290
3 Points Made	0.354
3 Points Attempts	0.341
3 Points %	0.329
Free Throw Made	0.252
Free Throw Attempts	0.260
Free Throw %	0.276
Offensive Rebounds	0.283
Defense Rebounds	0.288
Rebounds	0.288
Assists	0.305
Steals	0.246
Blocks	0.314
Turnover	0.233
target_5yrs	0.710

Table 50: Reconstruction Error for Data Set Generated from 210 Observations of the NBA Data Set

Columns	**Recon Error**
Games Played	0.337
Minutes Played	0.238
Points Per Game	0.256
Field Goals Made	0.245
Field goals Attempts	0.245
Field Goal %	0.278
3 Points Made	0.382
3 Points Attempts	0.365
3 Points %	0.338
Free Throw Made	0.244
Free Throw Attempts	0.262
Free Throw %	0.271
Offensive Rebounds	0.291
Defense Rebounds	0.285
Rebounds	0.287
Assists	0.295
Steals	0.254
Blocks	0.326
Turnover	0.235
target_5yrs	0.703

Table 50: Reconstruction Error for Data Set Generated from 420 Observations of the NBA Data Set

Columns	**Recon Error**
Games Played	0.329
Minutes Played	0.227
Points Per Game	0.248
Field Goals Made	0.236
Field goals Attempts	0.237
Field Goal %	0.271
3 Points Made	0.375
3 Points Attempts	0.350
3 Points %	0.321
Free Throw Made	0.249
Free Throw Attempts	0.267
Free Throw %	0.277
Offensive Rebounds	0.278
Defense Rebounds	0.263
Rebounds	0.268
Assists	0.291
Steals	0.263
Blocks	0.307
Turnover	0.239
target_5yrs	0.708

Table 51: Reconstruction Error for Data Set Generated from 630 Observations of the NBA Data Set

Columns	**Recon Error**
Games Played	0.342
Minutes Played	0.249
Points Per Game	0.265
Field Goals Made	0.251
Field goals Attempts	0.254
Field Goal %	0.267
3 Points Made	0.369
3 Points Attempts	0.352
3 Points %	0.330
Free Throw Made	0.262
Free Throw Attempts	0.278
Free Throw %	0.279
Offensive Rebounds	0.301
Defense Rebounds	0.289
Rebounds	0.293
Assists	0.287
Steals	0.281
Blocks	0.330
Turnover	0.244
target_5yrs	0.716

Table 52: Reconstruction Error between the Original and SVD-generated Data Sets for NBA Data Set

Kolmogorov-smirnov (KS) test

Table 53: KS Test for Data set Generated from 25 Observations of the NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.006	1.0
Minutes Played	0.006	1.0
Points Per Game	0.006	1.0
Field Goals Made	0.006	1.0
Field goals Attempts	0.006	1.0
Field Goal %	0.006	1.0
3 Points Made	0.006	1.0
3 Points Attempts	0.006	1.0
3 Points %	0.006	1.0
Free Throw Made	0.006	1.0
Free Throw Attempts	0.006	1.0
Free Throw %	0.006	1.0
Offensive Rebounds	0.006	1.0
Defense Rebounds	0.006	1.0
Rebounds	0.006	1.0
Assists	0.006	1.0
Steals	0.006	1.0
Blocks	0.006	1.0
Turnover	0.006	1.0
target_5yrs	0.006	1.0

Table 54: KS Test for Data set Generated from 50 Observations of the NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.013	1.0
Minutes Played	0.013	1.0
Points Per Game	0.013	1.0
Field Goals Made	0.013	1.0
Field goals Attempts	0.013	1.0
Field Goal %	0.013	1.0
3 Points Made	0.013	1.0
3 Points Attempts	0.013	1.0
3 Points %	0.013	1.0
Free Throw Made	0.013	1.0
Free Throw Attempts	0.013	1.0
Free Throw %	0.013	1.0
Offensive Rebounds	0.013	1.0
Defense Rebounds	0.013	1.0
Rebounds	0.013	1.0
Assists	0.013	1.0
Steals	0.013	1.0
Blocks	0.013	1.0
Turnover	0.013	1.0
target_5yrs	0.013	1.0

Table 54: KS Test for Data set Generated from 100 Observations of the NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.038	0.576
Minutes Played	0.038	0.576
Points Per Game	0.038	0.576
Field Goals Made	0.038	0.576
Field goals Attempts	0.038	0.576
Field Goal %	0.038	0.576
3 Points Made	0.038	0.576
3 Points Attempts	0.038	0.576
3 Points %	0.038	0.576
Free Throw Made	0.038	0.576
Free Throw Attempts	0.038	0.576
Free Throw %	0.038	0.576
Offensive Rebounds	0.038	0.576
Defense Rebounds	0.038	0.576
Rebounds	0.038	0.576
Assists	0.038	0.576
Steals	0.038	0.576
Blocks	0.038	0.576
Turnover	0.038	0.576
target_5yrs	0.038	0.576

Table 55: KS Test for Data set Generated from 210 Observations of the NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.005	1.0
Minutes Played	0.005	1.0
Points Per Game	0.005	1.0
Field Goals Made	0.005	1.0
Field goals Attempts	0.005	1.0
Field Goal %	0.005	1.0
3 Points Made	0.005	1.0
3 Points Attempts	0.005	1.0
3 Points %	0.005	1.0
Free Throw Made	0.005	1.0
Free Throw Attempts	0.005	1.0
Free Throw %	0.005	1.0
Offensive Rebounds	0.005	1.0
Defense Rebounds	0.005	1.0
Rebounds	0.005	1.0
Assists	0.005	1.0
Steals	0.005	1.0
Blocks	0.005	1.0
Turnover	0.005	1.0
target_5yrs	0.005	1.0

Table 55: KS Test for Data set Generated from 420 Observations of the NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.01	1.0
Minutes Played	0.01	1.0
Points Per Game	0.01	1.0
Field Goals Made	0.01	1.0
Field goals Attempts	0.01	1.0
Field Goal %	0.01	1.0
3 Points Made	0.01	1.0
3 Points Attempts	0.01	1.0
3 Points %	0.01	1.0
Free Throw Made	0.01	1.0
Free Throw Attempts	0.01	1.0
Free Throw %	0.01	1.0
Offensive Rebounds	0.01	1.0
Defense Rebounds	0.01	1.0
Rebounds	0.01	1.0
Assists	0.01	1.0
Steals	0.01	1.0
Blocks	0.01	1.0
Turnover	0.01	1.0
target_5yrs	0.01	1.0

Table 56: KS Test for Data set Generated from 630 Observations of the NBA Data Set

Column Names	**KS Stat**	**P value**
Games Played	0.005	1.0
Minutes Played	0.005	1.0
Points Per Game	0.005	1.0
Field Goals Made	0.005	1.0
Field goals Attempts	0.005	1.0
Field Goal %	0.005	1.0
3 Points Made	0.005	1.0
3 Points Attempts	0.005	1.0
3 Points %	0.005	1.0
Free Throw Made	0.005	1.0
Free Throw Attempts	0.005	1.0
Free Throw %	0.005	1.0
Offensive Rebounds	0.005	1.0
Defense Rebounds	0.005	1.0
Rebounds	0.005	1.0
Assists	0.005	1.0
Steals	0.005	1.0
Blocks	0.005	1.0
Turnover	0.005	1.0
target_5yrs	0.005	1.0

Table 57: Kolmogorov–Smirnov (KS) Test between the Original and SVD-generated Data Sets for NBA Data Set