<u>BINDURA UNIVERSITY OF SCIENCE EDUCATION</u>

<u>FACULTY OF SCIENCE AND ENGINEERING</u>

<u>DEPARTMENT OF STATISTICS AND MATHEMATICS</u>

This image cannot currently be displayed.

A comparative study of safety performance at ZIMPLATS: Safety, Health, Environment and Quality (SHEQ) framework implementation (1995-2007 vs 2008-2024)

Formatted: Font: 12 pt, Not Bold

BY

FINANCE DEBRA BAMU

B211983B

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE BARCHELOR OF SCIENCE HONOURS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS

SUPERVISOR: MR E. MUKONOWESHURO

JUNE 2025

APPROVAL FORM

I BAMU FINANCE DEBRA do hereby declare that this submission is my own work apart from the references of other people's work, which has been acknowledged. This work is submitted in partial fulfilment of the requirements of the Bachelor of Science Honors Degree in Statistics and Financial Mathematics.

FINANCE DEBRA BAMU	This image cannot currently be displayed.	18/06/2025
B211983B	Signature	Date

Certified by:

MR E. MUKONOWESHURO	This image cannot currently be displayed.	22/06/2025
Supervisor	Signature	Date

Dr MAGODORA	Magodora	……………….
Chairperson	Signature	Date

Dedication

I dedicate this project to my beloved parents, family members and friends for their unwavering support and for all the sacrifices they made towards my personal and professional development.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for His guidance and grace throughout the course of this project. Without His love and support, my dream would have been like a house built on sand, but now his grace gave me a strong foundation that will last a lifetime. With great honour, I extend my heartfelt thanks to my supervisor Mr. E. Mukonoweshuro for the unwavering support and extra effort on mentorship. It is his persistent supervision that brought hope and confidence in me. Although, the views and opinions expressed in this project remain my own, my ability to come out with a meaningful dissertation remains his responsibility. Lastly, my deepest gratitude goes to the Department of Statistics and Mathematics for their invaluable suggestions and direction in ensuring the success of various phases of this project. I do also like to thank my classmates, friends and family for their support and assistance.

ABSTRACT

This study was to assess the effectiveness of ZIMPLATS Safety, Health, Environment and Quality (SHEQ) framework system by comparing its safety performance indicators before (1995-2007) and after (2008-2024) its introduction. The study empkoyed a quantitative approach with the use of Random Forest (a machine learning model) and Autoregressive Integrated Moving Average (ARIMA) models to compare and forecast trends in safety performance. The models were run using RStudio software. Injury trends were forecasted using an ARIMA (4,0,0) model. The results indicated that the implementation of SHEQ at ZIMPLATS caused a reduction on the number of injuries. Also, the results indicated that Total Injury Frequency Rate (TIFR) was the major independent variable of injury rates with lesser impact from Lost Time Injuries (LTIs). Theredore, this research offers relevant facts to policymakers and practitioners who aim to improve safety standards for the mining sector, confirming the need to implement full systems of safety for ensuring sustainable operations in the long run. In addition, the government has to be engaged in exercising safety, at ZIMPLATS and other businesses in the same sector providing them with PPE and training programs with the aim of improving safety, health, quality and workplace safety. ZIMPLATS and other businesses in the same sector have to embrace utilizing SHEQ framework since it functions effectively in terms of improving safety, health and workplace environment.

Commented [JP1]: Rephrase this sentence so that it is clear

Commented [JP2]: There is need to rewrite the abstract so that it clearly spells out what was done, and what were the findings

Table of Contents

List of Figures

List of Tables

ABREVIATIONS AND ACRONYMS

SHEQ	Safety, Health, Environment and Quality
GDP	Gross Domestic Product
ILO	International Labor Organization
WHO	World Health Organization
ADF	Augmented Dickey-Fuller
ACF	Autocorrelation Function
PACF	Partial Autocorrelation
ARIMA	Autoregressive Integrated Moving Average
TIFR	Total Injury Frequency Rate
LTIs	Lost Time Injuries

CHAPTER 1: INTRODUCTION

1.0 Introduction

The Zimbabwean economy is heavily dependent on the mining sector, which makes it a significant contributor to the country's Gross Domestic Product (GDP). Given the industry's reliance on heavy and hazardous machinery for productive performance, the mining industry presents itself as a high-risk workplace for employees. In accordance with this, ZIMPLATS implemented a Safety, Health, Environment and Quality (SHEQ) system in 2008 for its performance improvement and reduction of injuries in the industry.

Therefore, this study seeks to evaluate the impact of SHEQ framework by comparing its performance before (1995-2007) and after (2008-2024) its implication. The aim of this research is to guide evidence-based decision-making, enhance safety policy, and promote safety culture within the industry.

1.1 Background of the study

The mining industry is a substantial industry in Zimbabwe's economy, contributing approximately 11% of Zimbabwe's GDP (Chamber of Mines of Zimbabwe, 2020). The International Labour Organization (2018) explains that the mining industry is one of the most hazardous industries in the world, most associated with higher possibilities of accidents and injuries. Therefore, SHEQ as a combined concept evolved to assist in sustainability, compliance with regulatory requirements and operational excellence in high-risk industries like mining (Kecojevic & Komljenovic 2011).

In Zimbabwe, the mining industry has been plagued with safety concerns like inadequate safety training and protective gear for workers (Machingura et al. 2019). During the pre-years (1995-2007) ZIMPLATS applied disjointed and reactive approaches in addressing the industry`s safety, health, environmental and quality concerns. This resulted in increased operational risks and poorer performance due to high incidences of injuries and accidents. These results, motivated ZIMPLATS to adopt SHEQ framework in 2008. It was with the aim of reducing work-related injuries and improving workers' safety and health. Briefly, from an analysis on the background of SHEQ implementation using ZIMPLATS records, below is a graph showing trends of injuries over time.

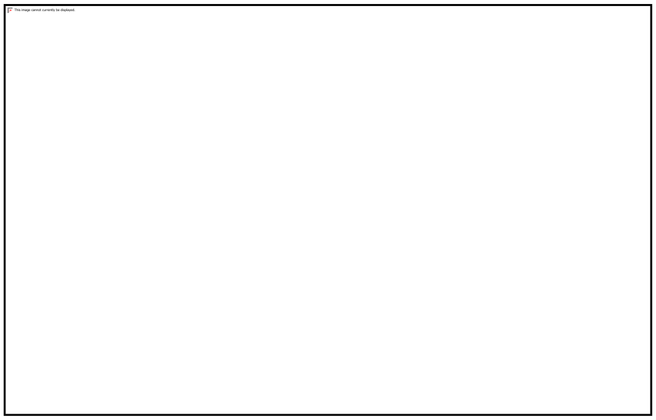

Figure 1.1: Trends of injuries at ZIMPLANTS

Figure 1.1 above shows changes in the number of injuries over time, distinguishing between the pre-implementation (1995-2007) and post-implementation (2008-2024) periods. With reference to the graph, the red represents injury trends from 1995 to 2007 while, the blue line represents the post-implementation period. Generally, the post-implement phase reflects a significant reduction in injuries from 2007 to 2024 with the numbers consistently decreasing from 100 injuries in 2007 to nearly zero injuries by 2024. The sharp decline observed in the post-period denotes that the implementation of SHEQ framework has contributed to the enhanced workplace safety and injury prevention. This trend supports the effectiveness of post-intervention measures, reinforcing the importance of structured safety strategies in minimizing workplace hazards.

However, the effectiveness of the framework in reducing injuries requires the need for ongoing evaluation and monitoring to ensure its continued effectiveness and identify areas of improvement. Therefore, this study aims to address a comparative review of SHEQ performance before and after the implementation at ZIMPLATS in order to gain valuable insights into the effectiveness of the framework and identify opportunities for further improvement. This will contribute to the body of literature on safety interventions and accident incidence in the context of informing practical recommendations on employees' well-being enhancement, cost minimization and a safe working environment (ILO 2019).

1.2 Statement of the Problem

Mining sector`s safety records remain a concern with ongoing accidents and injuries among operations and workers. Although ZIMPLATS has implemented the SHEQ framework, further evaluation is necessary for determining its effectiveness. Therefore, this study aims to analyse the trends of SHEQ performance over time, assess SHEQ performance and injury rates before and after the implementation of the framework and to determine the overall impact of the SHEQ framework at ZIMPLATS by analysing data.

1.3 Research Objectives

- To analyze the trend of SHEQ performance over time in the industry.
- To assess SHEQ performance and injury rates before and after the implementation of the framework.
- To determine the overall effect of SHEQ framework on organizational performance.

1.4 Research Questions

- Is there a significance difference in SHEQ performance metrics before and after the implementation of SHEQ framework?
- How has the impact of implementing SHEQ impacted overall organizational performance?
- What differences exist in safety performance metrics between 2000-2007 and 2008-2024

1.5 Assumption of the study

- The study assumes that SHEQ data and reports provided by ZIMPLATS are accurate.
- The study assumes that the SHEQ framework was fully implemented in 2008 and has been followed.
- The study assumes that the data used in the analysis is of high quality with minimal missing values and errors.

1.6 Delimitations of the study

- It will only focus at ZIMPLATS mining industry, neglecting other mining companies.
- The study will focus on specific variables excluding other potential factors that may influence safety outcomes.

- The analysis compares SHEQ performance between two different periods 2000-2007 and 2008-2024.

1.7 Limitations of the study

- Historical data from 2000-2007 might be incomplete and inaccurate which can affect comparability with post-implementation data of 2008-2024.
- Quality of SHEQ records and accessibility of accurate data provided by the company.

1.8 Significance of the study

To ZIMPLATS Company:

ZIMPLATS may evaluate its performance before and after implementing SHEQ framework in its industry. The study's findings might help ZIMPLATS and other industries in the mining sector in informing strategies to improve workplace safety, health and outcomes in their industries.

To the Researcher:

This research serves as an academic and professional development opportunity enabling the researcher to:

- Explore the different aspects and gain practical insights of the mining sector.
- Develop intellectual skills, critical thinking and analytical capabilities
- Enhance perception on the changes caused by implementing SHEQ in the mining industry.

To Bindura University:

This study contributes to the school’s academic reputation by demonstrating its commitment to research and knowledge generation.

1.9 Conclusion

This chapter introduces how the project will be conducted by highlighting the significance of SHEQ at ZIMPLATS and industries the same sector. It outlines the study`s research objectives, questions and assumptions of the research providing a foundation of the study. This chapter gives a brief overview of the whole study, while the following chapter reviews literature under the same study.

CHAPTER 2: LITERATURE REVIEW

2.0 Introduction

The implementation of SHEQ framework has been at the centre of attention in recent years due to the hazardous and risks within the industry and the requirements for safety improvements. Numerous studies have pointed towards the effectiveness of SHEQ frameworks in preventing workplace accidents, injuries and safety performance improvement. Therefore, this literature review aims to update existing research on the effectiveness of SHEQ frameworks in general with a particular emphasis on ZIMPLATS.

Kecojovic and Komljenovic (2011) supports that a well-established SHEQ framework helps in improving operational excellence through regulation compliance enforcement and minimization of environment risks. They also suggest that the implementation of SHEQ framework also contributes to the development of a safety culture between workers at the industry. This will lead to better assurance and productivity.

However, this study will identify gaps and the impact of SHEQ system on safety performance at ZIMPLATS and other mining firms. This study provides an understanding of the role of SHEQ framework in enhancing safety outcomes and business processes leading the way to the presentation of the evaluation of ZIMPLATS performance before and after implementing the framework. Below are the theoretical, empirical and conceptual frameworks supporting this study.

2.1 Theoretical Frameworks

Theoretical framework forms the foundation of the whole research, with its major focus on the research objectives of a study (Lederman and Lederman, 2017). This research under SHEQ draws its basis from related theories as below:

The Safety Management theory- this theory supports that effective safety management systems are capable of delivering lower workplace accidents and improved employee health (Reason, 1997). It also supports the belief that, organized safety procedures such as, SHEQ framework can improve productivity in terms of overall safety performance in high-risk industries such as mining. Moorkamp et al, (2014), supports this theory by encouraging commercial space companies implement safety measures that improve their workplace to create a better working environment

Commented [JP3]: This literature is old. Review current, 10 years old or less for employees. In Zimbabwe, the ILO introduced the occupational the Occupational, Health and Safety Management System (OHSMS) to guide the safety management theory. Taderera, (2012), analysed the OHSMS to investigate its application in the Zimbabwean sectors. He further recommended Zimbabwe to continue using the OHSMS to enhance industry`s safety and quality management. With the use of heavy machinery and chemicals, the mining industry remains a hazardous environment with high rates of accidents and injuries (ILO, 2018). In a report from the Chamber of Mines of Zimbabwe (2020), it is reported that safety records in the mining industries have been substandard due to the use of inadequate safety procedures and trainings. In this regard, ZIMPLATS introduced SHEQ framework in 2008 to improve its industry`s safety and highlights the effectiveness of the framework in its annual reports. These reports claim that the implementation of SHEQ has caused a decline in the number of injuries and improved workplace safety.

Organizational Culture Theory-Organizational Culture theory puts great emphasis on the role played by a safety-oriented culture in developing safe conducts by employees (Schein, 2010). It also theorizes that, if industries and organizations place high importance on safety, it leads to better compliance and fewer accidents. That can also be crucial in accounting for the impact of SHEQ implementation at ZIMPLATS. This theory is supported by most authors as they review a strong positive impact of the theory on managing safety, health and environment of an industry. According to Cameroon and Quinn, (2011), the organizational culture theory enhances organization`s safety and promotes quality environment. A flexible and responsible culture can better navigate change and implement new initiatives (Kotter, 2016). Therefore, by employing this theory in safety management ZIMPLATS can continue enhancing its culture and improving safety.

Accident causation theory-This theory explains why accidents occur by studying the underlying and root causes of accidents. Accident has been operationalized by the World Health Organization (WHO, 2019) as an unplanned and unforeseen event causing harm, damage or loss. Therefore, the need for a theory reflects difficulties in developing rational and comprehensible explanations regarding why some events, people and equipment interact and produce negative outcome. Using this theory, this study can help the company gain its goals of improved safety, health, environment and quality.

Human Factors Theory -The human factors theory of causation of accidents views that, chains of events caused by human errors may lead to accidents. This theory describes how employee motivation, workplace layout and training in the SHEQ context improve safe working conditions and improved health outcomes. According to Chombo, (2024), an industry may fail to positively progress due to misconduct of the human factor theory. A proper conduct of the human factor theory promotes the economic and social obstacles. Hence, this theory aids ZIMPLATS in implementing safety measures.

Risk Management Theory (RMF) - Risk Management Framework (RMF) provides an organized approach in dealing with likely threats that have an effect on operations and outcomes. It enables organizations to discover, assess and deal with risks in a mannered procedure. It is imperative in analysing how ZIMPLATS deals with SHEQ practices to address safety, health, environmental, and quality hazards. It supports the presumption that proactive handling of risks will enhance safety performance and compliance. The Committee of Sponsoring Organizations (COSO, 2017) highlights on the importance of risk management in in achieving structural objectives of a company. Therefore, the RMF theory is of importance to ZIMPLATS in achieving its goals on safety, the likes of zero-harm injuries.

Policy Evaluation theory -Policy evaluation theory directs a systematic approach to the assessment of the effectiveness, efficiency and impact of policies. This theory focuses on improving policy implementation and identifies areas of improvement during the policy`s development periods. It is also constituted by the summative evaluation which analyses the impact of a policy on the organization, offering understandings into the long-term effects and outcomes.

2.2 Empirical Evidence

Empirical data show that:

Safety training courses- These courses may train employees in operating machinery, chemicals and all equipment utilized in operation activities. Training employees in safety measures may avoid accidents being caused by them and also improve safety. Industries that engage in safety training courses can lead to a better knowledge of retention and application, reducing the risks of accidents and injuries (Minter, 2014). The ZIMPLATS offers some safety training such as, working at height training which encourages employees on how to work with caution at certain heights. Also, it offers first aid trainings, ladder safety training and noise awareness trainings (National Safety Council, 2022).

Personal protective equipment (PPE)- The utilization of PPE, say, respirators and safety glasses can improve safety. It completely guarantees the implementation of SHEQ framework at ZIMPLATS. For this reason, it should be compulsory for the mining industries to provide PPE to employees. It will improve safety and working environment. According to Reingen et al., (2018), PPE can significantly reduce the risk of injuries and accidents in several industries. The use of PPE stipulates protection in industries thereby promoting workers safety and health. Dorman P., (2017), further supports this by highlighting that workers are motivated by a healthy working environment. This will improve their self-esteem and output by reducing risks and accidents and injuries. The ZIMPLATS emphasizes the use of PPE in the industry to manage safety to achieve its goal of zero-harm.

First Aid competitions: Chamber of Mines of Zimbabwe encourages its members to get all the workers first aid trained and hold a valid safety aid certificate. The first aid competitions are aimed at ensuring the mines have teams competent enough to deal with emergencies that are most probable in their operations. This supports the structure's focus on health and safety, introducing evidence of ZIMPLATS commitment to staff preparedness and response capability.

2.3 Conceptual Frameworks

The conceptual framework for this study illustrates the relationship between the implementation of the SHEQ framework and its impact on safety, health, environment and quality performance. The framework also outlines the main perceptions, variables and relationships in a research study. It informs data collection and evaluation. The SHEQ is constituted by some conceptual frameworks including Hierarchy of Controls, Health Risk Assessment, Environmental Management Systems (EMS) and Total Quality Management (TQM). These frameworks help in improving workplace safety and employee well-being.

The EMS and TQM is also an integration of Safety, Health, Environment and Quality Management Systems (SHEQMS) which is similar to SHEQ framework. It is an approach for organizations which combines the ISO 14001: 2015(Environmental Management System), ISO 45001: 2018 (Occupational Health and Safety Management System) and ISO 9001:2015 (Quality Management Systems) into a specific framework used by industries in reducing accidents and injuries. Studies have shown that SHEQMS enhances organization`s performance, efficiency and improves workplace safety (Robson et al., 2007). By putting emphasis on SHEQMS, organizations and industries might build trust among employees and stakeholders (Schein, 2010). Therefore, this supports the importance of SHEQ at ZIMPLATS.

2.4 Research Gap

The mining industry is fully aware that Safety, Health, Environment, and Quality (SHEQ) framework is crucial, but there is lack of research on its performance at ZIMPLATS. Most studies just look at general safety practices without digging deeper into what happens before and after implementing SHEQ. However, there is need for in-depth analysis under this study. In this regard, this study will explore that gap before and after SHEQ implementation. The goal is to give ZIMPLATS and other mining industries valuable insights to improve safety and enhance efficiency.

2.5 Chapter Summary

The literature review highlights the importance of SHEQ at ZIMPLATS. The chapter discussed theoretical, empirical and conceptual frameworks under this study. The research gap highlights the demand for a detailed analysis of trends, patterns and impact of SHEQ in the industry. By addressing this gap, this study aims to compare the effectiveness of SHEQ performance during the pre-implementation and post-implementation at ZIMPLATS. The next chapter will discuss on the research methodology.

CHAPTER 3: RESEARCH METHODOLOGY

3.0 Introduction

This section shapes the systematic approach used for conducting research on the effectiveness of SHEQ framework at ZIMPLATS. It outlines the research design and data collection procedures. By comparing the effectiveness of SHEQ framework before and after its implementation, this research aims to compare the performance of ZIMPLATS before and after SHEQ framework implementation.

3.1 Research Design

This study employed a quantitative approach on the collection of numerical data to examine the implementation of SHEQ practices at ZIMPLATS. Quantitative research is a suitable for this analysis because it allows for objectivity, reliability and precision. By minimizing personal bias and judgment, quantitative research ensures that the findings are based on empirical evidence.

Quantitative research design was used in this study because it is well-suited for comparing SHEQ performance and identifying patterns and trends over a specific given period (2000-2024). By using statistical methods to analyse numerical data, the researcher can draw conclusions about the relationship between safety interventions and outcomes in the mining industry.

3.2 Research Approach

The methodology section outlines the procedures, strategies and methods used for collecting data and evaluation. This study was carried out utilizing quantitative approach, focusing on numerical data, statistical analysis and objective measurements. It is shaped by the research questions and objectives guiding data collection procedures for the study.

3.3 Data Source

Secondary data

- **ZIMPLATS Reports**: annual reports, incident records and annual safety performance summaries.
- **International Organizations**: Research publications from ILO and World Health Organization providing insights into industry standards.

3.4 Data Collection

This research utilized secondary data from existing content, reports and research studies. The data sources includeed were from ZIMPLATS and other mining industry reports, government data bases, international labour organizations (ILO) and publications. By using these data collection methods, the study aimed to gather reliable and accurate data to analyse the impact of SHEQ framework in the organization.

Commented [JP4]: Use past tense in chapter 3

3.5 Target population and sampling

Creswell J. (2014) defines a target population as an entire group of individuals that the researcher is concerned in surveying and to whom the researcher would like to publish the results. Sekaran and Bougie (2016) defines sampling as process of splitting data from a larger group of individuals to gather and conclusions of the population. Population and sampling are essential in survey the whole research methodology.

Sample: sample was taken from ZIMPLATS annual reports;

1. Pre-implementation (1995-2007)
2. Post-implementation (2008-2024)

3.6 Description of variables

Table 3.1 is a summary of variables for the study, injuries being the dependent variable, while the rest are explanatory variables.

Variable	**Symbol**	**Description**	**Source**
Injuries	IR	Number of safety injuries per unit time	ZIMPLATS annual and SHEQ reports, government databases
Total Injury frequency rate	TIFR	Number of injuries per hours worked	ZIMPLATS annual and SHEQ reports, government databases
Lost Time Injuries	LTIs	Number of lost injuries per hours worked	ZIMPLATS annual and SHEQ reports, government databases

Table 1.1: Description of variables

3.7 Analytical models

3.7.1 Autoregressive Integrated Moving Average (ARIMA) Model

The ARIMA Model was used in this study to analyze injury time series data at ZIMPLATS. The model development process involved model identification, estimation amd diagnostic testing.

- In the model identification stage, the time series used the ADF and the results indicated that the time series was stationary after third differencing. Also, the autoregressive and moving average terms were identified using the ACF and PACF. However, the parameters of the model were identified using AIC and BIC criteria to assess model's performance.
-
-
-

Assumption of ARIMA

1. **Stationarity**: The ARIMA model assumes that the underlying data is stationary (constant mean, variance, autocorrelation) over time. If the data is non-stationary it can be transformed using log transformations, differencing and other transformations to achieve stationarity.
2. **Heteroscedasticity**: Residuals should have constant variance over time (homoscedasticity), if heteroscedastic there should be an alternative model. If the variance of residuals changes, it can lead to inefficiencies in estimates and affect model`s reliability.
3. **Normality of Residuals**: The residuals (differences between actual and predicted values) should be normally distributed. Normality is significant in validating statistical tests and ensuring reliable confidence intervals and hypothesis tests.
4. **Linearity**: ARIMA assumes a linear relationship between past values and future values of a time series. This means that, the effect of past observations (before implementation) is additive and comparative.
5. **Independence of Residuals**: The residuals should be uncorrelated with one another. The presence of autocorrelation in the residuals suggests that the model has not fully captured the structure of the data and may require further refinement.

Justification of using ARIMA:

The ARIMA model is chosen for this study because:

- It can handle autocorrelation in the data, which is common in time series data.
- It can help evaluate and forecast trends of injuries over time, which is useful in predicting SHEQ performance.
- It can identify key factors influencing SHEQ performance in the mining industry.
- It can help to evaluate the effectiveness of the SHEQ framework.

3.7.2 Random Forest Model

The Random Forest model is a machine learning algorithm that chains various trees to improve the accuracy of a model in predicting outcomes. It also helos in reducing overfitting and improve simplifications. Each decision tree in the random forest model is trained on a random fold. For example, this study used a 10-fold the random forest is trained on 9 subsets and validated on the remaining subset. The random forest model improves accuracy, handles high-dimensional data with many features and missing values.

In this study, the Random Forest model was used to identify key factors influencing SHEQ performance at ZIMPLATS**. The model was traimed on a data set of historical SHEQ data . Results from the Random forest model showed its effectiveness in identifying key factors.**

Assumptions of Random Forest Model

- Independence of observations.
- No strong correlation between variables.
- It assumes that missing values are handled by several strategies such as, imputation or ignoring them.
- It assumes that, its built-in mechanisms prevent overfitting.
- It assumes a sufficient amount of data to train effectively.

Justification of using Random forests

- The random forest is flexible for both classical and regression tasks.
- It is also suitable for data with missing values, as it can handle them.
- It helps identify the predictors of possible outcomes.
- Random forest models have few hyperparameters to tune, making it relatively easy to improve model`s performance.

3.8 Model diagnostic tests

In this study, several diagnostic tests were conducted to evaluated to see the validity of using the ARIMA and Random Forest model. These tests were to meet the assumptions of the model as stated below:

1. **Stationarity**

A time series is considered stationary when its statistical parameters are constant across different periods (Box et al. 2008). For this study, the ADF test is used to test for stationarity. Therefore, in this study the ADF was used to make time series stationary after implying the Box Cox transformation **and third differencing.**

Augmented Dickey-Fuller (ADF):

H_0: Time series is non-stationery.

H_1: Time series is stationary.

ADF test specifically examines the presence of stationarity. It also indicates if a time series has a stochastic trend. A p-value less than the significance level (0.05), denotes that we reject the null hypothesis and conclude that the time series is stationary.

2. **Seasonal Decomposition**

According to Brockwell and Davis (2016), seasonal decomposition is a statistical method that breaks down a time series into its trend. This breakdown helps to identify underlying patterns, improve forecasting and facilitates a better understanding of time series data, allowing for more informed decision.

3. **Autocorrelation Checks**

- **ACF Plot**: ACF measures the relationship between a time series and its lagged values. It identifies trends and structures in the data (Enders 2004). A positive autocorrelation denotes persistent positive correlation while a negative correlation denotes no significant relationship. A long-term trend is shown by a slowly decaying ACF, while a significant ACF value indicate a strong persistent trend. Lastly, regular spikes indicate seasonality, while seasonal patterns indicate significant ACF values at seasonal lags reveal repeating patterns.
- **Partial Autocorrelation function (PACF) plot**: PACF measures the partial correlation between a time series and its lagged values, accounting for the effects of intermediate lags. It helps determine the order of autoregressive (AR) model by

identifying the lags where the correlation is still significant after removing the effects of previous lags.

- The **Ljung-Box** test is a statistical test used to determine autocorrelation in the residuals of a time series model. In this study, it was used to check for model adequacy and to detect autocorrelation. It is calculated as:

$$Q = n\ (n + 2) * \sum_{k=1}^{h}[\frac{(rk)^2}{n-k}]$$

Where:

Q = Ljung test statistic

N = Sample size

H = number of lags tested

rk = autocorrelation coefficient at lag k

This test follows a chi-square distribution with h degrees of freedom. The p-value is then calculated based on the chi-squared distribution and if it is less than the significance level, we reject the null hypothesis of no autocorrelation.

4. Testing for normality of residuals

Testing for normality is crucial as it improves accuracy of interpretation of results, model adequacy and improved modelling. This study used, Shapiro-Wilk **and** Q-Q plots to test for normality of residuals with the following hypothesis:

H_0: Data is normally distributed.

H_1: Data is not normally distributed.

Significance level: 5% (0.05)

5. Heteroscedasticity

The assumption of heteroscedasticity determines the variance of errors in a regression model. These errors are mostly caused by non-normality data, outliers, omitted variable bias and measurement errors. The study used Breusch-Pagan to test for heteroscedasticity. These errors can be reduced by the use of transformations and robust standard errors.

H_0: $\sigma^2 = 0$

$H_{1:}$ $\sigma^2 \neq 0$

3.9 Performance Metric tests

In this study, performance of the ARIMA model was evaluated using the Mean Absolute Error (MAE) and Root Mean Squared Error. These metric tests are used to measure the typical difference between predicted and actual values while MAE focuses on minimal errors and the RMSE focuses on larger errors making it useful for understanding the impact of significant deviations in injury rates. The formulas for MAE and RMSE are as shown below:

$$\text{MAE: } \frac{1}{n}\sum_{i=1}^{n} \mid y_i - y_i\hat{} \mid$$

$$\text{RMSE} = \sqrt{} \, [\tfrac{1}{n}\sum_{i=1}^{n}(y_i - y_i\hat{})^2]$$

3.10 Model validation and Reliability tests

Model validation and reliability tests were conducted to determine the extent of accuracy to which the models reflect real-life scenarios. The Random Forest and ARIMA models were used to meet the study's objectives. The results came up after ensuring consistency, reliability and performance of a product, system or component over its usage period with time, under various sets of conditions and according to specified requirements, (ISO, 1994).

3.11 Ethical Considerations

The World Health Organization (2019), defines research ethics as principles, guidelines and standards that regulates the manners of conducting research ensuring that responsibilities, respect and morals are well managed. Therefore, in this research data from relevant sources have been considered to support ethical considerations throughout the study.

3.12 Chapter Summary

This chapter outlined the methods used to compare the performance of SHEQ framework before and after it was implemented, specifically focusing on injuries, Total Injury Frequency Rate

(TIFR) and Lost Time Injuries (LTI). The quantitative research design was utilized to assess the performance of SHEQ before and after implementation. The study relied on secondary data which was used to evaluate the effectiveness of the framework. Necessary tests including, correlation stationarity and normality, were discussed to validate the findings. ARIMA model was used for data analysis in this study. The next chapter will present the results of the analysis, providing insights into the effectiveness of safety interventions in the mining industry.

CHAPTER 4: DATA PRESENTATION, ANALYSIS AND PRESENTATIONS

4.0 Introduction

This chapter is going to draw out the study`s findings through data presentations, analysis and interpretations to accomplish the research objectives and questions discussed earlier. The data tabulated in descriptive statistics is based on secondary data from ZIMPLATS. Additionally, it provides the outcomes of the researcher`s investigation into the **Random Forest and ARIMA** models. The results were run used R-Studio version 4.3.3.

4.1 Descriptive Statistics

	Year	**Post**	**Treatment**	**Injuries**	**TIFR**	**LTIs**
Min	1995	0.0	0.0	6	0.410	0.100
1st Qu	2002	0.0	0.0	14	1.002	0.330
Median	2010	1.0	0.0	43	4.340	0.46
Mean	2010	0.6	0.4	58.7	4.352	0.479
3rd Qu	2017	1.0	1.0	97.75	7.055	0.660
Max	2024	1.0	1.0	149	10.230	0.810

Table 4.1: Descriptive Statistics2

Table 4.1 is an overlay of the dataset from 1995 to 2024. The Post variable indicates treatment application showing a range from 0 (no treatment) to 1 (treatment applied). The minimum number of injuries being 6 and a maximum of 149. The first quartile is at 14 injuries, having a median of 43 and mean 58.7. The ranges from 0.410 to 10.230, with a median of 4.340 and a mean very close to the median, suggesting a relatively symmetric distribution. Also, the LTIs has a range from 0.100 to 0.810, a median of 0.460 and a mean of 0.479 (relatively symmetric distribution).

4.2 Data Analysis

Correlation analysis

Figure 4.1: Correlation Matrix

From Fig 4.1 above, the correlation coefficient between number of injuries and TIFR 0.77125157. This denotes a solid positive relationship between the variables. This indicates that as TIFR increases, the number of injuries also increases. Also, the correlation coefficient between number of injuries and LTIs is 0.07246327 indicating a weak positive correlation. This indicates a very weak positive correlation. Hence the number of Lost Time Injuries (LTIs) does not have a significant relationship with the number of injuries. In addition, the correlation coefficient between TIFR and LTIs is 0.1745497, indicating a weak positive correlation. This indicates a weak positive correlation, concluding that the relationship between TIFR and LTIs is not strong.

4.3 Analytical Models

4.3.1 ARIMA Model

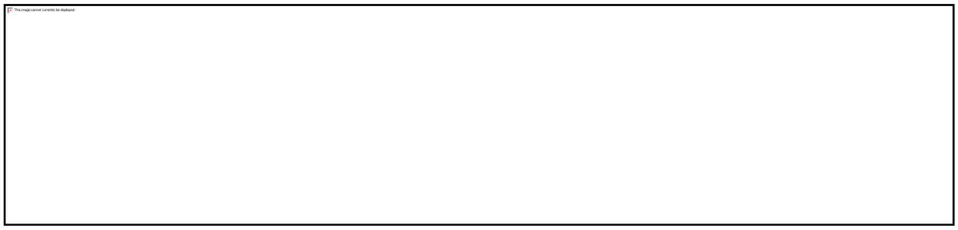

Figure 4.2: Arima (0, 1, 0)

The ARIMA (0,1,0) model described earlier has captured the first differencing needed to make the time series stationary but the ACF1 value suggests that there is perhaps still autocorrelation in the residuals. That is, the model might not be fully capturing the underlying patterns of the data and further steps should be done. For this, the model was once more altered through Box-Cox but also had a p-value of 0.57 indicating that the model was not stationary. The researcher hence had to difference the model at order three (d=3) to get ARIMA (4,0,0) with zero mean for better model adequacy and results.

Figure 4.3: Arima (4,0,0) with Zero Mean

The model is specified as ARIMA (4,0,0), which implies that the model has four autoregressive (AR) terms and zero differencing and moving average terms. The presence of these four autoregressive terms allows the model to track complex patterns in the time series. The low error metrics and minimal autocorrelation in residuals suggest that the model fits well. However, to confirm that the model fits well certain diagnostic tests had to be performed as indicated below.

Diagnostic Tests for ARIMA (4,0,0)

Figure 4.4: Testing for Stationarity

From the ADF test results, a p-value of 0.01 less than the significance level (0,05) indicates no autocorrelation. Therefore, we reject the null hypothesis and we conclude that the time series is stationary.

Figure 4.5: Ljung Test

The Ljung test shows a p-value of 0.22 greater than 0.05, you fail to reject the null hypothesis. This suggests that there is no significant autocorrelation in the residuals of your ARIMA (4,0,0) model. To support this, the ACF and PACF plots are illustrated below.

Figure 4.6: ACF of Residuals

The ACF shows the autocorrelation residuals at various lags. Any bars that exceed the dashed lines indicate significant autocorrelation. Therefore, from the plot no lags are within the confidence intervals, denoting that there is no significant autocorrelation in the residuals.

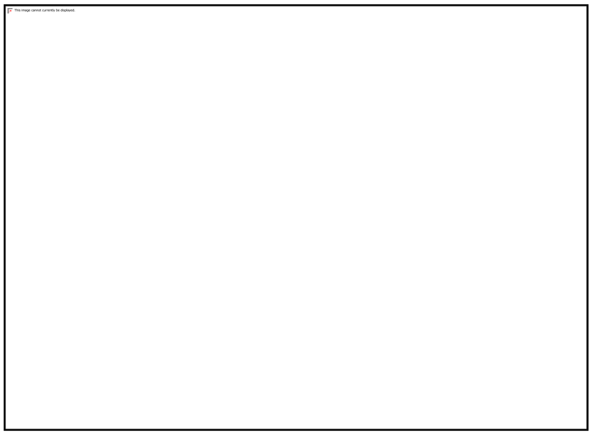

Figure 4.7: PACF of Residuals

The PACF plot measures the correlation between the residuals and their lags after removing the effect of intervening lags. Hence, no lags are within the confidence intervals, denoting that there is no significant autocorrelation.

Figure 4.8: Breusch Pagan Test for Heteroscedasticity

Figure 4.12 shows a p-value of 0.9406 greater than 0.05. This indicates that we fail to reject the null hypothesis. Therefore, there is no significant evidence of heteroscedasticity. Thus, supporting the assumption of Homoscedasticity.

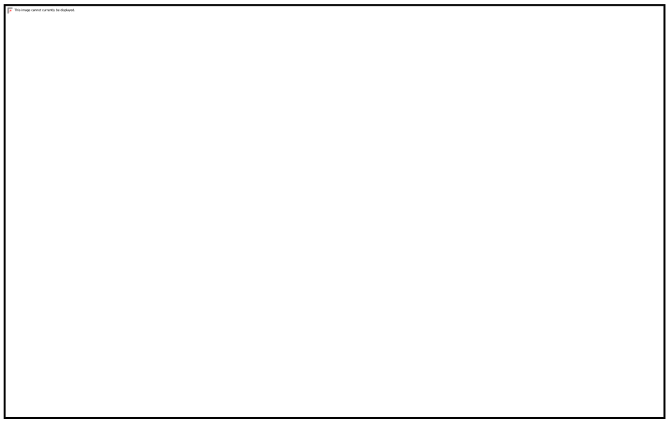

Figure 4.9: Normality using Q-Q Plot for residuals

The red diagonal line represents the expected quantiles of a normal distribution. The points along the diagonal line indicates normality. That’s the residuals of the ARIMA (4,0,0) are approximately normally distributed. This Q-Q plot is supported by the Shapiro-Wilk test as indicated below.

Figure 4.10: Testing for Normality using Shapiro Wilk Test

The Shapiro-Wilk test shows a p-value of 0.9358 which is greater than 0.05 (significance level). Hence, we fail to reject the null hypothesis and conclude that that there is normality of residuals.

This image cannot currently be displayed.

Commented [JP5]: Put key on this diagram

Figure 4.11: Forecast from ARIMA (4,0,0)

The illustration on figure 4.15 show a black line representing the fitted values from the ARIMA model, while the blue line indicates the forecasted values for future periods based on the ARIMA model. The shaded grey area indicates the range of values within which the actual values are likely to for at approximately 80%.

- *Forecasted Values* (blue line): Represents the predicted values based on the ARIMA model.

- *80% Prediction Interval* (shaded area): Indicates the range of values within which the actual values are likely to fall with 80% probability.

- *95% Prediction Interval* (optional): If you have a wider shaded area, you can indicate the range of values within which the actual values are likely to fall with 95% probability.

Boxplot to compare SHEQ performance

Figure 4.12: Comparison of Injuries before and after SHEQ

The box plot above shows the comparison of injuries before and after implementation of SHEQ framework. The median number of injuries is visibly lower in the post-implementation (treatment=1) compared to the pre-implementation group (treatment=0). This shows the effectiveness of SHEQ framework in reducing accidents and improving workplace safety. The pre-implementation group has several outliers above the upper whisker, suggesting that there were instances of significantly higher injury counts. In contrast, the post-implementation group has fewer outliers, indicating a more stable and controlled environment following the implementation.

4.3.2 Random Forest Model

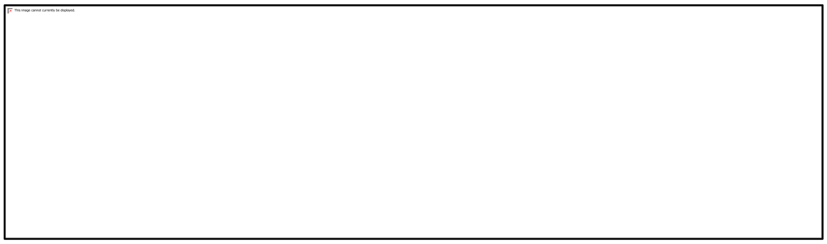

Figure 4.13: Random Forest Model

Figure 4.13 above, shows a regression random forest model using 500 trees which strengthens the outputs of the model. The mean of residuals (841.5107) represents the average of the squared differences between the predicted and actual values of injuries. A lower value of mean indicates a better model performance, but in this case a large value of 841.5107 indicates that the model needs refinements. Also, a percentage variance explained of 59.42 denotes the variability in the number of injuries which can be explained by model`s independent variables (TIFR, LTIs and Treatment), while a significant portion of 40.58 which shows that the model also needs refinement.

Figure 4.14: Random Forest Performance

The Random Forest model above shows the error rate which indicates that adding trees initially helps improve model accuracy. A sharp drop in error indicated on early (0-50 trees) denotes the significant of adding trees to the model. That's, as the number of trees increases, the error rate decreases significantly. At around 100-150 trees the error in the model alleviates, indicating that the model has likely reached its optimal complexity. Hence, these results indicate a good sign of the model. In this regard, the researcher had to log transformations to find a better model and the results were as shown below.

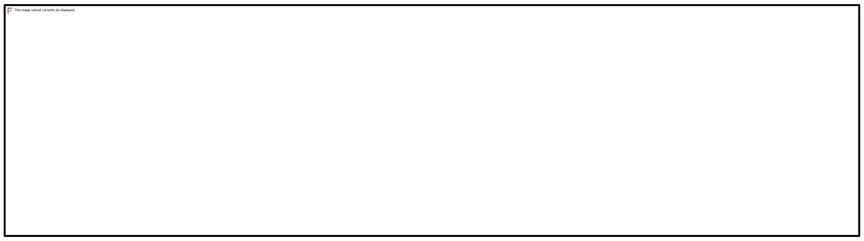

figure 4.15: Random Forest model after log transformation

The above model shows the regression random forest variable after log transformation, which uses 500 trees. It has a mean of squared residuals (0.2813675) and a percentage variance explained (72.35). These lower values indicate that the log transformation has helped in improving the model`s performance. Therefore, this indicates that the model captures the data well.

Figure 4.16: Variable importance

The variable importance plot above indicates the important of each variable in predicting outcomes. The percentage IncMSE measures how much the mean squared error increases when a variable is rearranged. From the illustration, TIFR has a strong impact on reducing accidents, while the treatment has moderate impact and LTIs have low impact on reducing injuries. The IncNodePurity measures how much better the model splits in respect to each variable across all trees. From the plot, TIFR with the highest value is more helpful in predicting outcomes as compared LTIs and treatment.

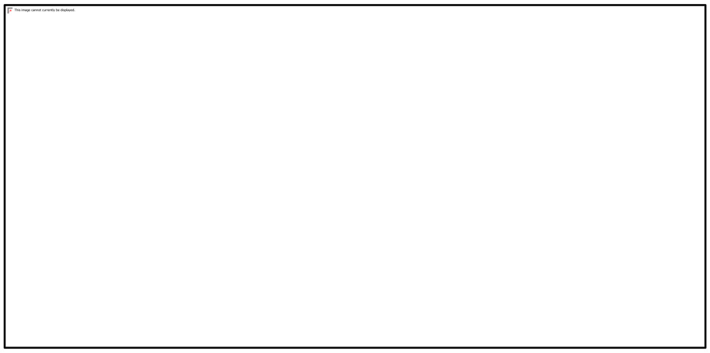

Figure 4.17: Random Forest Model after a 10-fold Cross-Validation

The random forest model was split into 10 folds, where the model is trained on 10 folds and validated on the remaining fold, imitating this process 10 times. The model shows some performance metrics with an RMSE value of 0.363910, 0.8564559 and 0.3019857. Therefore, according to the plot, the final value of the model was at the third tuning parameter (mtry = 3) with the smallest value.

Figure 4.18: Residuals vs Predicted

From the plot, randomly scattered residuals are around zero, indicates that the model`s prediction s are generally accurate. The absence of a clear pattern in the residuals suggests that the model ha s captured the underlying relationship in the data well. The random distribution of residuals reflec ts a good sign that the model does not experience systematic bias.

Diagnostic Tests for Random Forest Model

Normality Test

Figure 4.19: Testing for normality using a Q-Q plot

The residuals along the diagonal line indicates that the data is normally distributed. A good fit along the diagonal line indicates that the assumption of normality of residuals is met. This is verified using Shapiro-Wilk test below.

Figure 4.20: Testing for normality using Shapiro-Wilk test

A p-value of 0.4019 from the Shapiro-Wilk test is greater than the significant level 0.05, indicates that the data is of a normal distribution. Therefore, we fail to reject the null hypothesis and conclude that data is normally distributed.

Figure 4.21: Testing for heteroscedasticity using Breusch-Pagan test

From the above results on heteroscedasticity, a p-value of 0.2146 is greater than the significant level 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no presence of heteroscedasticity in the data. That's, implying the assumption of homoscedasticity.

4.4 Performance Metrics

Model	**RMSE value**	**MAE value**
ARIMA (4,0,0)	1.84952648821378	1.70664205591752
Random Forest	0.3438980	0.2905809

Table 4.2: Performance Metrics

The RMSE and MAE values above were extracted from ARIMA (4,0,0) and Random Forest model (after cross validation). Based on the above performance measures the random forest model has lower values of RMSE and MAE than the ARIMA (4,0,0) model. Lower values of errors indicate better model`s predictive performance. However, both models are crucial for the study as they satisfy different objectives for the study. That is, Arima model was to analyze the trend of injuries over time, while the Random Forest was used to evaluate the effectiveness of SHEQ framework at ZIMPLATS.

Commented [JP6]: How did you reach to this conclusion when these models were employed on different tasks, eg one for checking trends and the other on predicting performance?

4.5 Chapter Summary

The chapter depicts data presentation and analysis enabling the researcher to make a choice of the best model intended to be Random Forest. The data analysis revealed the performance of SHEQ framework before and after its implementation on the effectiveness of SHEQ in reducing injuries. The next chapter is a summary of the whole project, which gives recommendations and conclusions of the study.

CHAPTER 5: SUMMURY, CONCLUSION AND DISCUSSIONS

5.0 Introduction

This chapter summarizes the findings of this comparative research on the effectiveness of the SHEQ framework before and after its implementation. It discusses key insights, conclusions and recommendations derived from the analysis.

5.1 Summary on the findings of the study

The researcher contrasted pre-implementation phase (1995-2007) data with post-implementation phase (2008-2024) data in order to examine the effectiveness of SHEQ framework. The data utilized in the study was obtained from ZIMPLATS records and reports and provided a definitive image of safety and health indicators. The study primarily aimed to establish the effectiveness of SHEQ by examining the trend between pre-implementation phase and post-implementation phase. The literature review outlined the safety practices used during this whole period (1995-2024). The researcher therefore compared the impact of the intervention of SHEQ model of a stated time period of before and after the implementation of SHEQ. In order to arrive at best results, the researcher compared the ARIMA and Random Forest models. The ARIMA model has also been used to forecast future trends of injuries in years to come. R-studio software was used for analysis of data.

The objective of the analysis of the trend over time of SHEQ performance was attained utilizing ARIMA, while the Random Forest helped in predicting SHEQ performance based on TIFR, LTIs and treatment. Relevant diagnostic diagnostics such as, stationarity, homoscedasticity and normality of residuals were achieved for all the models in order to attain a best fit model for the study. From the variable importance of random forest, TIFR significantly contributed to injury rates, establishing its critical significance in predictive outcomes, while other variables had comparatively lower contributions. This denotes that, TIFR has a significant and strong impact on injuries with a persistent effect trending over time.Therefore, this finding highlights the importance of monitoring TIFR as a critical measure for assessing the performance of safety in the mining sector. This variable is a significant predictor in forecasting outcomes. Howecer, LTIs appear to have less impact on the injuries and contribute less to predictive outcomes. This implies that even though LTIs play a fundamental role in interpreting safety performance, they may not be as useful in predicting injury trends ahead of time as TIFR. Diagnostic tests also ensured that the assumptions on the models were met. This makes the models utilized more reliable and suggests that the results are not biased by major distortions. By comparing the data using previously outlined models, both the Random Forest model and ARIMA were crucial as they satisfied different objectives for same study. However, although both models are best for the study's objectives, the Random Forest model has minimal values of MAE and RMSE concluding that it performes extremely well in analysing the effectiveness of the framework and predicting positive outcomes of the framework than the ARIMA model.

5.2 Conclusion

The findings of the present research validate the effectiveness of SHEQ framework with a positive and significant impact on reducing accidents at ZIMPLATS. Empirical evidence validates the effectiveness of SHEQ practices but also provides a healthy foundation for future safety interventions within the mining industry.

5.3 Recommendations

The outcomes of the current study created groundwork for additional research to explore additional forecasting methods in an effort to identify more accurate predictive models.

Apart from that, other researchers can also proceed with researching more on other safety interventions that were not addressed by the researcher. Apart from that, they should maintain their data in soft copies and also update the software they are going to use to enhance and latest version. ZIMPLATS should proceed with monitoring injury rates and TIFR on a regular basis to ensure guaranteed effectiveness of SHEQ framework. Government should apply safety measures, in ZIMPLATS and other companies in the same industry which provide them PPE and training programs to help with safety, health, quality and workplace safety enhancement. ZIMPLATS and other companies in the same industry should apply SHEQ framework since it is effective in the improvement of safety, health and workplace environment.

REFERENCES

Box, G.E.P., Jenkins, G.M. & Reinsel, G.C., 2015. Time Series Analysis: Forecasting and Control. 5th ed. Hoboken, NJ: Wiley.

Brockwell, P.J. & Davis, R.A., 2016. Introduction to Time Series and Forecasting. 3rd ed. New York: Springer.

Cameroon, K.S. & Quinn, R.E., 2011. Diagnosing and Changing Organizational Culture: Based on the Competing Values Framework. 3rd ed. San Francisco: Jossey-Bass.

Chamber of Mines of Zimbabwe, 2020. Annual Mining Industry Report: Safety and Health Statistics. Harare: Chamber of Mines.

Chombo, I., 2024. Human Factor Theory in Zimbabwean Industry: An Operational Analysis. Harare: ZimAcademic Press.

COSO, 2017. Enterprise Risk Management: Integrating with Strategy and Performance. Committee of Sponsoring Organizations of the Treadway Commission.

Creswell, J.W., 2014. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 4th ed. Thousand Oaks, CA: SAGE Publications.

Dorman, P., 2017. Healthy Workplaces? Improving Working Conditions in the African Mining Sector. Geneva: International Labour Office.

Enders, W., 2004. Applied Econometric Time Series. 2nd ed. Hoboken, NJ: Wiley.

International Labour Organization (ILO), 2018. Safety and Health in Mining: A Global Perspective. Geneva: ILO.

International Organization for Standardization (ISO), 1994. ISO 14001: Environmental Management Systems – Specification with Guidance for Use. Geneva: ISO.

International Organization for Standardization (ISO), 2015. ISO 9001: Quality Management Systems – Requirements. Geneva: ISO.

Kecojevic, D. & Komljenovic, D., 2011. The role of safety management systems in enhancing safety performance in the mining industry. Journal of Safety Research, 42(5), pp.329–335.

Kotter, J.P., 2016. Leading Change. Boston: Harvard Business Review Press.

Lederman, L. & Lederman, R., 2017. Theoretical frameworks for understanding safety management in the workplace. Safety Science, 97, pp.135–144.

Machingura, F., Ndlovu, T. & Moyo, S., 2019. Safety training and its impact on workplace injuries in the mining sector. Mining Safety Journal, 15(2), pp.45–58.

Minter, S.G., 2014. Training for safety: What works? Occupational Hazards Journal, 76(4), pp.28–35.

Moorkamp, M., Veenendaal, H. & Kolen, B., 2014. Creating a safer workplace: Lessons from space industry applications of safety frameworks. Journal of Safety Engineering, 3(2), pp.17–23.

National Safety Council, 2022. Occupational Safety Training Manual. Chicago: NSC Press.

Reason, J., 1997. Managing the Risks of Organizational Accidents. Aldershot: Ashgate.

Reingen, P.H., Zak, D. & Stern, D.E., 2018. Personal protective equipment effectiveness in hazardous industries. Industrial Safety Review, 12(3), pp.55–60.

Schein, E.H., 2010. Organizational Culture and Leadership. 4th ed. San Francisco: Wiley.

Sekaran, U. & Bougie, R., 2016. Research Methods for Business: A Skill-Building Approach. 7th ed. New York: Wiley.

Taderera, G., 2012. Evaluation of OHSMS effectiveness in Zimbabwean industries. Zimbabwe Journal of Safety Studies, 6(1), pp.22–31.

APPENDICES

Appendix A: ARIMA Model

```
# Load necessary libraries

library(tidyverse)

library(lubridate)

library(forecast)

library(tsoutliers)

# Load the dataset

df <- read.csv("DF.csv")

# Data Cleaning

df_clean <- df %>%

  mutate(Year = as.numeric(Year),

         Post = factor(Post, levels = c(0, 1), labels = c("Before", "After")),

         TIFR = as.numeric(TIFR),

         LTIs = as.numeric(LTIs)) %>%

  filter(!is.na(TIFR), !is.na(LTIs))

# Create a time series object for TIFR

ts_tifr <- ts(df_clean$TIFR, start = min(df_clean$Year), frequency = 1)

# Split the data into pre and post implementation

pre_implementation <- window(ts_tifr, end = 2007)

post_implementation <- window(ts_tifr, start = 2008)

# ARIMA Modeling

arima_model <- auto.arima(ts_tifr)
```

```
# Forecasting
forecasted_values <- forecast(arima_model, h = 5)
# Diagnostic Tests
checkresiduals(arima_model)
# Summary of Results
summary(arima_model)
print(forecasted_values)
# Load necessary libraries
library(lmtest)
library(forecast)
# Assuming you have already fitted the ARIMA model as shown previously
# arima_model <- auto.arima(ts_tifr)
# Get residuals from the ARIMA model
residuals_arima <- residuals(arima_model)
# Shapiro-Wilk Test for Normality
shapiro_test <- shapiro.test(residuals_arima)
print(shapiro_test)
# Box-Cox Transformation
lambda <- BoxCox.lambda(df_clean$TIFR)
df_clean$TIFR_boxcox <- BoxCox(df_clean$TIFR, lambda)
# Create a new time series object
ts_tifr_boxcox <- ts(df_clean$TIFR_boxcox, start = min(df_clean$Year), frequency = 1)
# Fit ARIMA model on transformed data
```

```
arima_model_boxcox <- auto.arima(ts_tifr_boxcox)

# Check residuals

checkresiduals(arima_model_boxcox)

# Box-Cox Transformation

lambda <- BoxCox.lambda(df_clean$TIFR)

df_clean$TIFR_boxcox <- BoxCox(df_clean$TIFR, lambda)

# Create a time series object for Box-Cox transformed data

ts_tifr_boxcox <- ts(df_clean$TIFR_boxcox, start = min(df_clean$Year), frequency = 1)

# Fit ARIMA model on Box-Cox transformed data

arima_model_boxcox <- auto.arima(ts_tifr_boxcox)

# Get residuals

residuals_boxcox <- residuals(arima_model_boxcox)

# Shapiro-Wilk Test

shapiro_test_boxcox <- shapiro.test(residuals_boxcox)

print(shapiro_test_boxcox)

# Load necessary libraries

library(lmtest)

library(tseries)

library(forecast)

library(ggplot2)

# Assuming you have already performed Box-Cox transformation and created the time series

# Fit the ARIMA model on Box-Cox transformed data

arima_model_boxcox <- auto.arima(ts_tifr_boxcox)
```

```
# 1. Durbin-Watson Test
# First, ensure residuals are extracted correctly
dw_test_boxcox <- dwtest(residuals(arima_model_boxcox) ~ fitted(arima_model_boxcox))
print(dw_test_boxcox)
# 2. Breusch-Pagan Test
# Use a linear model for Breusch-Pagan test
bp_test_boxcox <- bptest(lm(residuals(arima_model_boxcox) ~ fitted(arima_model_boxcox)))
print(bp_test_boxcox)
# 3. Augmented Dickey-Fuller Test
adf_test_boxcox <- adf.test(ts_tifr_boxcox)
print(adf_test_boxcox)
# 4. ACF and PACF Plots
par(mfrow = c(1, 2))  # Set the plotting area for ACF and PACF
Acf(residuals(arima_model_boxcox), main = "ACF of Residuals")
Pacf(residuals(arima_model_boxcox), main = "PACF of Residuals")
par(mfrow = c(1, 1))  # Reset plotting area
# 5. Check residuals
checkresiduals(arima_model_boxcox)
# Differencing the Box-Cox transformed data
ts_tifr_boxcox_diff <- diff(ts_tifr_boxcox)
# Plot the differenced data to visualize
plot(ts_tifr_boxcox_diff, main = "Differenced Box-Cox Transformed TIFR", ylab = "Differenced Value", xlab = "Year")
```

```
# Re-run the ADF test on the differenced data
adf_test_boxcox_diff <- adf.test(ts_tifr_boxcox_diff, alternative = "stationary")
print(adf_test_boxcox_diff)
# Optional: Fit ARIMA model on the differenced data
arima_model_boxcox_diff <- auto.arima(ts_tifr_boxcox_diff)
# Check model summary
summary(arima_model_boxcox_diff)
# Check residuals of the differenced model
checkresiduals(arima_model_boxcox_diff)
# Load necessary library
library(tseries)
# Assuming you have already differenced the data
# ts_tifr_boxcox_diff <- diff(ts_tifr_boxcox)
# Re-run the ADF test on the differenced data
adf_test_boxcox_diff <- adf.test(ts_tifr_boxcox_diff, alternative = "stationary")
print(adf_test_boxcox_diff)
# Visualize the differenced data to check for stationarity
plot(ts_tifr_boxcox_diff, main = "Differenced Box-Cox Transformed TIFR", ylab = "Differenced Value", xlab = "Year")
# Optional: ACF and PACF plots of the differenced data
par(mfrow = c(1, 2))  # Set the plotting area for ACF and PACF
Acf(ts_tifr_boxcox_diff, main = "ACF of Differenced Data")
Pacf(ts_tifr_boxcox_diff, main = "PACF of Differenced Data")
```

```
par(mfrow = c(1, 1))  # Reset plotting area

# Second differencing

ts_tifr_boxcox_diff2 <- diff(ts_tifr_boxcox_diff)

# Re-run the ADF test on the second differenced data

adf_test_boxcox_diff2 <- adf.test(ts_tifr_boxcox_diff2, alternative = "stationary")

print(adf_test_boxcox_diff2)

# Third differencing

ts_tifr_boxcox_diff3 <- diff(ts_tifr_boxcox_diff2)

# Re-run the ADF test on the third differenced data

adf_test_boxcox_diff3 <- adf.test(ts_tifr_boxcox_diff3, alternative = "stationary")

print(adf_test_boxcox_diff3)

# Plot the third differenced data

plot(ts_tifr_boxcox_diff3,

     main = "Third Differenced Box-Cox Transformed TIFR",

     ylab = "Differenced Value",

     xlab = "Year",

     type = "l",

     col = "blue")

# Add a horizontal line at y=0 for reference

abline(h = 0, col = "red", lty = 2)

# ACF and PACF plots for the third differenced data

par(mfrow = c(1, 2))  # Set the plotting area for ACF and PACF

Acf(ts_tifr_boxcox_diff3, main = "ACF of Third Differenced Data")
```

```
Pacf(ts_tifr_boxcox_diff3, main = "PACF of Third Differenced Data")

par(mfrow = c(1, 1))  # Reset plotting area

# Example of fitting an ARIMA model

arima_model_final <- auto.arima(ts_tifr_boxcox_diff3)

# Summary of the fitted model

summary(arima_model_final)

# Check residuals

checkresiduals(arima_model_final)

# Check residuals for normality (Shapiro-Wilk test)

shapiro_test_residuals <- shapiro.test(residuals(arima_model_final))

print(shapiro_test_residuals)

# Plot residuals

plot(residuals(arima_model_final), main = "Residuals of ARIMA(4,0,0)", ylab = "Residuals", xlab = "Index")

abline(h = 0, col = "red")

# ACF of residuals

Acf(residuals(arima_model_final), main = "ACF of Residuals")

# Load necessary libraries

library(lmtest)

library(tseries)

library(forecast)

library(ggplot2)

# Fit the ARIMA model (assuming it's already fitted)
# arima_model_final <- auto.arima(ts_tifr_boxcox_diff3)

# 1. Shapiro-Wilk Test for Normality

shapiro_test_residuals <- shapiro.test(residuals(arima_model_final))

print(shapiro_test_residuals)

# QQ Plot for Normality

qqnorm(residuals(arima_model_final), main = "QQ Plot of Residuals")

qqline(residuals(arima_model_final), col = "red")

# 2. Durbin-Watson Test

dw_test <- dwtest(residuals(arima_model_final) ~ fitted(arima_model_final))

print(dw_test)

# 3. Breusch-Pagan Test

bp_test <- bptest(lm(residuals(arima_model_final) ~ fitted(arima_model_final)))

print(bp_test)

# 4. Visualizations

# Residuals Plot

plot(residuals(arima_model_final), main = "Residuals of ARIMA(4,0,0)", ylab = "Residuals", xlab = "Index")

abline(h = 0, col = "red")

# ACF Plot of Residuals

Acf(residuals(arima_model_final), main = "ACF of Residuals")

# PACF Plot of Residuals

Pacf(residuals(arima_model_final), main = "PACF of Residuals")

# Reset plotting area
```

```
par(mfrow = c(1, 1))

# Load necessary library for accuracy metrics

library(forecast)

# Assuming your ARIMA model is already fitted

# arima_model_final <- auto.arima(ts_tifr_boxcox_diff3)

# 1. Forecasting

h <- 12  # Number of periods to forecast

forecasted_values <- forecast(arima_model_final, h = h)

# Plot the forecast

plot(forecasted_values,

     main = "Forecast from ARIMA Model",

     ylab = "Forecasted Values",

     xlab = "Time")

lines(fitted(arima_model_final), col = "blue")  # Add fitted values

# 2. Define actual values for comparison

actual_values <- c(1.2, 1.3, 1.5, 1.4, 1.6, 1.8, 1.7, 1.9, 2.0, 2.1, 2.3, 2.4)  # Replace with your actual values

# Calculate RMSE

rmse <- sqrt(mean((forecasted_values$mean - actual_values)^2))

print(paste("RMSE:", rmse))

# Calculate MAE

mae <- mean(abs(forecasted_values$mean - actual_values))

print(paste("MAE:", mae))
```

```
# Optionally: Print the forecasted values

print(forecasted_values)
```

Appendix B: Random Forest

```
# Load necessary libraries
library(dplyr)
# Read the data
data <- read.csv("DF.csv")
# Check for missing values
summary(data)
# Remove rows with missing values (if any)
data_cleaned <- na.omit(data)
# Convert Year to a factor for analysis
data_cleaned$Year <- as.factor(data_cleaned$Year)
# Load necessary libraries
library(randomForest)
# Set seed for reproducibility
set.seed(123)
# Convert Treatment to a factor
data_cleaned$Treatment <- as.factor(data_cleaned$Treatment)
# Create a Random Forest model
rf_model <- randomForest(no.of.injuries ~ TIFR + LTIs + Treatment, data = data_cleaned, importance = TRUE)
# Print the model
print(rf_model)
```

```
# Diagnostic tests for Random Forest
importance(rf_model) # Variable importance
plot(rf_model)       # Error rate plot
# Predictions
predictions <- predict(rf_model, data_cleaned)
# Evaluate model performance
actuals <- data_cleaned$no.of.injuries
mse <- mean((predictions - actuals)^2)
print(paste("Mean Squared Error:", mse))
# Log transformation of the response variable
data_cleaned$log_no_of_injuries <- log(data_cleaned$no.of.injuries + 1)  # Adding 1 to avoid log(0)
# Build the Random Forest model with the transformed variable
rf_model_log <- randomForest(log_no_of_injuries ~ TIFR + LTIs + Treatment, data = data_cleaned, importance = TRUE)
# Check the summary
print(rf_model_log
# Predictions from the model
predictions_log <- predict(rf_model_log, data_cleaned)
# Calculate residuals
residuals_log <- data_cleaned$log_no_of_injuries - predictions_log
# Residual plot
plot(predictions_log, residuals_log,
    xlab = "Predicted Log Injuries",
    ylab = "Residuals",
    main = "Residuals vs Predicted")
abline(h = 0, col = "red")
# Calculate RMSE
rmse_log <- sqrt(mean(residuals_log^2))
print(paste("RMSE:", rmse_log))
# Calculate R-squared
ss_total <- sum((data_cleaned$log_no_of_injuries - mean(data_cleaned$log_no_of_injuries))^2)
ss_residual <- sum(residuals_log^2)
r_squared_log <- 1 - (ss_residual / ss_total)
print(paste("R-squared:", r_squared_log))
# Variable Importance Plot
importance_log <- importance(rf_model_log)
print(importance_log)
# Plot variable importance
varImpPlot(rf_model_log, main = "Variable Importance")
# Load caret package
library(caret)
# Set up training control
train_control <- trainControl(method = "cv", number = 10)
# Train the model with cross-validation
cv_model <- train(log_no_of_injuries ~ TIFR + LTIs + Treatment,
          data = data_cleaned,
          method = "rf",
          trControl = train_control)
# Print cross-validated results
print(cv_model)
# Install lmtest package if not already installed
install.packages("lmtest")
# Load the package
library(lmtest)
# Conduct Breusch-Pagan test
bp_test <- bptest(rf_model_log)
print(bp_test)
# QQ Plot
qqnorm(residuals_log)
qqline(residuals_log, col = "red", lwd = 2)
# Perform the Shapiro-Wilk test for normality
shapiro_test <- shapiro.test(residuals_log)
# Print the results
print(shapiro_test)
# Predictions from the model
predictions_log <- predict(rf_model_log, data_cleaned)
# Calculate RMSE
rmse_log <- sqrt(mean((data_cleaned$log_no_of_injuries - predictions_log)^2))
```

```
print(paste("RMSE:", rmse_log))

# Calculate R-squared

ss_total <- sum((data_cleaned$log_no_of_injuries - mean(data_cleaned$log_no_of_injuries))^2)

ss_residual <- sum((data_cleaned$log_no_of_injuries - predictions_log)^2)

r_squared_log <- 1 - (ss_residual / ss_total)

print(paste("R-squared:", r_squared_log))

# Calculate MAE (Mean Absolute Error)

mae_log <- mean(abs(data_cleaned$log_no_of_injuries - predictions_log))

print(paste("MAE:", mae_log))
```

Appendix C: Visuals of injury trends and performance

```
library(ggplot2)

# Plot number of injuries over time

ggplot(df, aes(x = Year, y = no.of.injuries, color = factor(Treatment))) +

  geom_line(size = 1) +

  geom_point(size = 2) +

  labs(title = "Injury Trends Over Time",

       x = "Year", y = "Number of Injuries",

       color = "Treatment Status (0=Pre, 1=Post)") +

  theme_minimal()

# Boxplot to compare pre vs. post framework injury rates

ggplot(df, aes(x = factor(Treatment), y = no.of.injuries, fill = factor(Treatment))) +

  geom_boxplot() +

  labs(title = "Comparison of Injuries Before and After SHEQ Implementation",

       x = "Implementation Status", y = "Number of Injuries") +

  theme_minimal()
```