Bindura University of Science Education

TIME SERIES ANALYSIS OF COVID19 INFECTION IN ZIMBABWE

BY

GODKNOWS CHAMUNORWA MURENZA

B192578B

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENTOF THE REQUIREMENTS OF THE BACHELOR OF SCIENCE HONOURS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS.

SUPERVISORS: MR B KUSOTERA

2023

APPROVAL FORM

I GODKNOWS CHAMUNORWA MURENZA as a bonafide student do hereby declare that this work has been presented neither in whole nor in part for any degree at this University. The work is submitted in partial fulfillment of the requirements of the Bachelor of Science Honors Degree in Statistics and Financial Mathematics.

GODKNOWS C. MURENZA (B192578B) **Date: 06/12/2023**

Signature

CMurenza

Certified by

MR B KUSOTERA **Date: 06/12/2023**

Supervisor

Signature

DEDICATION

I express my gratitude and recognition to Stella Madziva, my beloved mother, as well as to HigherLife Foundation for their continuous and steadfast assistance, and for the numerous sacrifices that they have made to foster my growth and progress both in my personal life and career. I am appreciative of their affection and unwavering faith in me.

ACKNOWLEDGEMENTS

I would like to extend my heartfelt appreciation to the omnipotent for steering me through this journey and granting me the fortitude to see it to completion. Additionally, I gratefully acknowledge the exceptional contribution of Mr. B. Kusotera, my supervisor, who not only provided me with invaluable assistance but also offered me counsel on how to go about composing my dissertation. Furthermore, I want to express my thanks to the WHO Organization for granting me access to their data, which significantly aided in my research. My family's unwavering love, care, and support deserve recognition, as well as my classmates and professors from the Math and Engineering Department, who guided me throughout the program.

ABSTRACT

The aim of this study was to provide an overview of COVID-19 infection in Zimbabwe. The objectives of the study were to establish the time series model that explains the rate of infection and to forecast the infection in the next six months. An ARIMA (5, 0, 0) model was employed on the daily time series data spanning from March 2020 to April 2023. The results of the study showed likelihood of increasing trend in COVID-19 infections. The predictions also pointed to possible increases of infections. The surge in cases has been attributed to the emergence of new variants, low vaccination rates, and limited access to testing and healthcare. The government has implemented various measures to control the spread of the disease, including lockdowns, travel restrictions, and vaccination campaigns. However, these measures have faced challenges due to limited resources, vaccine hesitancy, and misinformation. The study recommends continued efforts to increase vaccination rates, improve access to testing and healthcare, and combat misinformation through education and awareness campaigns.

ACRONYMS

ACF	Auto-Correlation Function
ADF	Augmented Dickey- Filler
AR	Auto-Regressive model
ARIMA	Auto-Regressive Integrated Moving Average
ARMA	Auto-Regressive Moving Average
MA	Moving Average
PACF	Partial Auto-Correlation Function
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
WHO	World Health Organization
AIC	Akaike Information Criteria
BIC	Bayesian Information Criteria
Q-Q	Quantile-Quantile
COVID 19	Coronavirus Disease 2019
SARS	Severe Acute Respiratory Syndrome

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.0 INTRODUCTION

Chapter 1 of this thesis aims to provide readers with a clear understanding of the study. It includes a section on the foundation of the COVID-19 pandemic, a brief outline of the ARIMA model, an explanation of the research issue, objectives, investigation questions, justification of the study, significance of the study, and an overview of the chapter

1.1 BACKGROUND

According to Zhu et al. (2020), the identification of the novel coronavirus SARS-CoV-2 as the cause of the COVID-19 pandemic occurred in Wuhan, China in December 2019. As reported by the World Health Organization (WHO) in 2023, the COVID-19 pandemic, which was first identified in Wuhan, China in December 2019, has rapidly spread to become a global health crisis, affecting millions of people worldwide. The virus is primarily transmitted through respiratory droplets or contact with contaminated surfaces, with an incubation period of 2-14 days and a range of symptoms from mild to severe, particularly in older individuals or those with underlying health conditions. Governments and healthcare systems worldwide have implemented measures to control the spread of the virus, but limited resources, vaccine hesitancy, and misinformation have posed significant challenges, particularly in vulnerable populations with existing health disparities. In Africa, there have been over 14 million confirmed cases and 330,000 deaths due to COVID-19 as of June 2023, exacerbating existing health inequalities and economic challenges. In Zimbabwe, with over 113,000 confirmed cases and 3,800 deaths as of June 2023, the surge in cases has been attributed to the emergence of new variants, low vaccination rates, and limited access to testing and healthcare, posing significant challenges to the healthcare system and economy. The government has implemented various measures to control the spread of the virus, but challenges remain due to limited resources, vaccine hesitancy, and misinformation.

1.1.1 BRIEF OVERVIEW OF THE ARIMA

ARIMA is short for Auto-Regressive Integrated Moving Average and belongs to the family of time series models. Its primary function is to analyze data and predict future values based on historical data. In the paper, ARIMA was used to model Covid19 confirmed cases data.

1.2 STATEMENT OF THE PROBLEM

Although COVID-19 appears to have come and gone in Zimbabwe, the government cannot afford to relax due to its negative impact on the country's economy. It is risky to be unprepared in case new variants emerge again. One of the crucial steps to stay prepared is to consistently conduct predictive analytics.

1.3 OBJECTIVES

The aims of the research are outlined below:

i. To establish time series model that explains the rate of infection

ii. To predict the infection in next six months

1.4 RESEARCH QUESTIONS

i. What is the trend of COVID infections in Zimbabwe over time?

ii. How accurate is ARIMA model in predicting future COVID infection rates in Zimbabwe?

iii. How is the COVID 19 stand on the next six months?

iv. Can the ARIMA model be used to forecast potential outbreaks or surges of COVID infections in Zimbabwe?

v. Is there any sign that COVID 19 will end at the future?

1.5 JUSTIFICATION OF THE STUDY

The COVID-19 pandemic brings deadly loss in the economy of Zimbabwe which is need to be attended and find a way out the deadly impact of the current pandemic. The fact mentioned above, drive the student to come up with this piece of paper to establish time series model that clearly explains the rate of infection and to predict the infection in future months from now in particular at least six months.

1.6 SIGNIFICANCE OF THE STUDY

1.6.0 TO THE RESEARCHER

The research empowers the student to pick up encounter on how a research is done, including the process included. The project moreover upgrades the student's research intellectual abilities and getting understanding into examination and investigation of the infection of (COVID 19) through building up time series model in a bid to explain the rate of infection and predicting the infection on a given time.

1.6.1 TO THE UNIVERSITY

The research adds to the university's writing and gets to be a source of secondary information to other students or analysts.

1.6.2 TO THE HEALTH SECTOR AND CITIZENS OF ZIMBABWE

They can easily make the decisions considering the pandemic using the rate of infection, and the future predictions keep them aware. Also it helps them to analyze the need for mitigation measures of the pandemic.

1.6.3 TO THE GOVERNMENT OF ZIMBABWE

By knowing the infection rate and knowing the predicted future infection stance, the GoZ may be able to apply the mitigation measures like lockdown and play its role as country of supplying clinical materials like vaccinations on a right time and on a right quantity.

1.7 ASSUMPTIONS

i. COVID 19 will continue to spread and the number of infected will increase.

ii. Measures like partial lockdowns assumed to reduce the rate of the infection.

iii. It is assumed that partial lockdown may be the best way to monitor the rate at the spread of COVID-19.

1.8 LIMITATIONS

i. This document is easily comprehensible to those who are proficient in mathematics, particularly in statistics and economics, as it is formulated based on mathematical models and mathematical language.

ii. It employs secondary data, which may not be entirely valid and reliable.

iii. It involves intricate workings, requiring readers' close attention.

1.9 SUMMARY

In this chapter, essential components for conducting a research study were identified, with a focus on defining the research scope. The primary aim of the study is to make a prediction of time series data related to COVID-19 infections in Zimbabwe.

CHAPTER 2: LITERATURE REVIEW

2.0 INTRODUCTION

The impact of the COVID-19 pandemic on the world has been substantial, and several countries have faced challenges in controlling its transmission. Zimbabwe is among the affected countries, experiencing an increasing number of infections and deaths. To understand the trend of COVID-19 infections in Zimbabwe, time series analysis can be utilized to model and predict future patterns. This study will employ the ARIMA model, a widely recognized technique for time series analysis that can identify trends and patterns in data over time. The purpose of utilizing this model to analyze the data is to detect potential patterns or trends in COVID-19 infections in Zimbabwe and offer recommendations for managing and preventing the virus's spread. To achieve this objective, our literature review will explore previous research on infectious diseases utilizing ARIMA models for time series analysis and examine the impact of COVID-19 on public health in Zimbabwe.

2.1 DEFINITION OF TERMS

1. "Time series analysis refers to a statistical technique utilized to examine data collected over a specific period, identifying patterns and trends within the data to make predictions about future values based on past observations" (Box, 2015).

2. World Health Organization, (2020) COVID infection pertains to the illness caused by the novel coronavirus (SARS-CoV-2), which emerged in late 2019 and has since spread globally, leading to a pandemic.

3. Zimbabwe is a southern African country with a population of approximately 14 million people (United Nations, 2021).

4. The ARIMA model, also known as the Autoregressive Integrated Moving Average model, is a statistical method used for time series forecasting. It involves evaluating the autocorrelation and partial autocorrelation functions of the data to determine appropriate model parameters (Box et al., 2015).

5. The Autocorrelation function measures the correlation between a time series and its lagged values at different time intervals (Chatfield, 2004).

6. The Partial Autocorrelation function calculates the correlation between a time series and its lagged values while controlling for other lags (Brockwell & Davis, 2016).

7. Forecasting is the process of predicting future values or trends based on previous observations or data trends (Makridakis et al., 1998).

8. A pandemic refers to an outbreak of an infectious disease that spreads globally and affects a vast number of people over an extended period (Centers for Disease Control and Prevention, 2021).

9. Statistical analysis involves collecting, analyzing, and interpreting data using statistical concepts to draw conclusions or make estimations about a population or sample (Hair et al., 2019).

10. Data visualization entails using charts, graphs, or other visual aids to present data graphically, enabling the identification of patterns and trends within the data (Few, 2012).

2.2 THEORETICAL FRAMEWORK

The fundamental approach of this research is centered around the utilization of the ARIMA model, a statistical technique used for analyzing and predicting time-series data (Box, Jenkins, & Reinsel, 2015; Brockwell & Davis, 2016; Hyndman & Athanasopoulos, 2018). The ARIMA model contains three key components: autoregression (AR), differencing (I), and moving average (MA). Autoregression is employed to evaluate the correlation between an observation and various preceding observations. Differencing is applied to transform the time series data into a stationary state. The moving average component explores the relationship between an observation and previously predicted errors.

2.2.1 MODEL PARAMETERS

In this chapter, we will explore the parameters used in time series analysis of COVID infection in Zimbabwe using ARIMA model.

Parameters:

The ARIMA model has several parameters that need to be specified before fitting the model to the data. These parameters include:

There are three parameters in the ARIMA model: p, d, and q. The first parameter, p, indicates the number of previous observations used in predicting future values. The second parameter, d, represents how many times the data is differenced to achieve stationarity by removing trends or seasonality. Finally, the parameter q signifies the number of past errors used in predicting future errors. Correctly identifying these parameters is essential in obtaining reliable forecasts of COVID infections in Zimbabwe and to make informed decisions regarding public health interventions.

2.2.2 EXPONENTIAL SMOOTHING THEORY

According to Hyndman and Athanasopoulos (2018), Exponential smoothing is a commonly utilized statistical approach for time series analysis and forecasting. This method operates under the premise that the most recent observations in a time series hold more significance than earlier observations when predicting future values. Exponential smoothing employs weighted values for past observations, with these weights decreasing exponentially over time. The smoothing parameter, alpha (α), determines the rate at which the weights decrease.

The most basic form of exponential smoothing is known as simple exponential smoothing (SES) and can be expressed mathematically as follows:

$$\mathbf{y\_t = \alpha * y\_t\text{-}1 + (1\text{-}\alpha) * y\_t\text{-}2 + \ldots + (1\text{-}\alpha)\hat{}(t\text{-}1) * y\_1} \ldots\ldots\ldots\ldots\ldots(1)$$

In the equation, y_t represents the value of the time series at time t, while y_t-1, y_t-2... y_1 denote past observations of the series. The smoothing parameter is denoted by α. The first part of the equation on the right-hand side represents the smoothed value of the time series at time t based on the most recent observation at time t and the previous smoothed value at time t-1. The remaining part of the equation represents the weighted average of past observations, where the weights decrease exponentially over time.

The selection of the α value is dependent on the nature of the time series and the intended use. A smaller α value increases the significance of past observations, generating a less erratic forecast, while a larger α value emphasizes recent observations, resulting in a more reactive forecast.

Usually, the α value is chosen based on a forecast accuracy measure, such as the mean squared error (MSE) or mean absolute error (MAE), utilizing a validation dataset.

Exponential smoothing can be adapted to deal with time series that exhibit trend and seasonality by incorporating extra parameters and equations. For instance, Holt's linear exponential smoothing (Holt, 1957) introduces a trend component to the SES formula, while Holt-Winters' exponential smoothing (Holt and Winters, 1960) introduces a seasonal component. These extensions enable more precise and adaptable forecasting of time series data.

Exponential smoothing is a prevalent and effective technique for time series analysis and forecasting. It offers a versatile and user-friendly approach to modeling and predicting time series data, with several adaptations and extensions available to accommodate various data types and uses.

2.2.3 STATIONARITY THEORY

Brockwell, P. J., & Davis, R. A. (2016), stationarity is a fundamental concept in time series analysis and plays a key role in the application of the ARIMA model for the analysis of COVID-19 infections. Stationarity refers to the statistical properties of a time series that remain constant over time, and is characterized by constant mean, variance, and autocorrelation structure. In the context of COVID-19 infections, stationarity implies that the mean, variance, and autocorrelation of the number of infections remain constant over time, which is a key assumption for the ARIMA model.

Formally, a time series Y_t is said to be stationary if it satisfies the following conditions:

1. Constant mean: The mean of the time series is constant over time and does not change with time. This condition is expressed mathematically as:

$$\mathbf{E(Y\_t) = \mu} \ldots\ldots\ldots(2)$$

where E(Y_t) is the expected value of the time series at time t, and μ is a constant.

2. Constant variance: The variance of the time series is constant over time and does not change with time. This condition is expressed mathematically as:

$$\mathbf{Var(Y\_t) = \sigma\hat{}2}\ldots\ldots\ldots\ldots(3)$$

where Var(Y_t) is the variance of the time series at time t, and σ^2 is a constant.

3. Constant autocorrelation: The autocorrelation structure of the time series is constant over time and does not change with time. This condition is expressed mathematically as:

$$\mathbf{Cov(Y\_t, Y\_\{t\text{-}k\}) = \gamma\_k}\ldots\ldots\ldots\ldots(4)$$

where Cov(Y_t, Y_{t-k}) is the covariance between the time series at time t and its lagged value at time t-k, and γ_k is a constant that depends only on the lag k.

The concept of stationarity is important in the application of the ARIMA model, as it allows for the identification of the model's parameters. In practice, however, time series data are often non-stationary, meaning that the mean, variance, and/or autocorrelation structure of the time series change over time. In such cases, differencing can be used to transform the data into a stationary form that is amenable to analysis using the ARIMA model.

2.2.4 CHAOS THEORY

Chaos theory is a mathematical discipline that explores intricate systems that are susceptible to initial conditions, where minor changes in the starting conditions can lead to significant differences in the system's behavior over time. Chaos theory has been utilized in various fields, such as time series analysis, to evaluate the system's dynamics and predict its future evolution. In the context of time series analysis of COVID-19 infection, Lorenz (1963) suggested that chaos theory could be employed to recognize tipping points, or critical thresholds, beyond which the pandemic may spiral out of control.

A fundamental idea in chaos theory is the concept of a chaotic attractor, which refers to a group of points in phase space that the system tends to converge towards over time. The attractor typically exhibits a fractal pattern, implying that it has a self-replicating form at varying scales. The behavior of the system can be explained by a series of differential equations, such as the Lorenz system:

$$\mathbf{dx/dt = \sigma*(y\text{-}x)}\ldots\ldots..(5)$$

$$\mathbf{dy/dt = x*(\rho\text{-}z) - y}\ldots\ldots..(6)$$

$$\mathbf{dz/dt = x*y - \beta*z}\text{………(7)}$$

The Lorenz system comprises state variables, x, y, and z, and system parameters, σ, ρ, and β. For specific parameter values, the Lorenz system demonstrates chaotic behavior, with paths that are extremely responsive to initial conditions.

Chaos theory can be applied to scrutinize time series data by restoring the attractor from the available data using a method referred to as phase space reconstruction. The fundamental concept is to utilize the time series data to approximate the system's state variables and then plot the system's path in phase space. Different measures, such as Lyapunov exponents, can be employed to analyze the attractor, which quantifies the pace of dispersion or convergence of neighboring paths.

Chaos theory can be utilized to identify crucial thresholds beyond which the COVID-19 pandemic may spiral out of control. For instance, the reproduction number (R), which estimates the mean number of secondary infections produced by each infected individual, can serve as an indicator of the system's stability. If R surpasses 1, the pandemic will grow exponentially, but if R is less than 1, it will eventually subside. Nonetheless, the system might display chaotic behavior for specific R values, resulting in unpredictable and unstable dynamics.

Chaos theory offers a robust framework for examining the behavior of intricate systems, such as time series data. By applying the notion of a chaotic attractor and phase space reconstruction, time series data can be assessed to recognize critical thresholds and unstable patterns. In the context of COVID-19 infection, chaos theory can be utilized to identify tipping points and potential interventions to mitigate the pandemic.

2.3 CONCEPTUAL FRAMEWORK

(Mhlanga & Chikodzi, 2021) time series analysis of COVID-19 infection in Zimbabwe involves a conceptual framework with several components. The first step is data collection, which requires collecting reliable and accurate data on COVID-19 infection in Zimbabwe. The data should include details such as daily or weekly counts of new cases. Once the data has been collected, the next step is data preprocessing. This step ensures that the data is consistent, complete, and free of errors and outliers. Preprocessing may involve removing missing or duplicated data, correcting errors, and identifying and removing outliers. The subsequent stage after data preprocessing is exploratory data analysis, where the data is scrutinized to detect trends, correlations, and patterns. Visual tools like graphs and charts can aid in comprehending the data. Summary statistics such as mean and standard deviation can be utilized to compute statistical measures. Hypothesis testing can be performed to assess whether there are significant dissimilarities among distinct groups or time periods. The next step is time series modeling, which involves developing time series models to predict future trends in COVID-19 infection. Various models such as ARIMA or exponential smoothing can be fitted to the data to select the best model based on criteria such as AIC or BIC. Once the models have been constructed, the subsequent phase is model evaluation. This process entails juxtaposing the projected values against the factual values utilizing metrics such as mean squared error or root mean squared error to assess the precision and dependability of the models. The last stage is forecasting, where the time series models are employed to predict future tendencies in COVID-19 infection rates in Zimbabwe. This may require producing forecasts for diverse scenarios, such as various degrees of social distancing or vaccination rates. By utilizing the forecasts, the probable impact of different interventions can be evaluated

2.4 EMPIRICAL LITERATURE

This research outlined the latest investigations related to modelling and analyzing COVID-19 infections using statistical models. Multiple authors and their respective studies were referred to in this context.

Moyo, Chikodzi, and Musuka (2021) conducted a research study entitled "Time series analysis of COVID-19 cases in Zimbabwe: A comparison of ARIMA and LSTM models." The study aimed to compare the effectiveness of ARIMA and LSTM models in predicting COVID-19 cases in Zimbabwe using detailed time-series data from January 2020 to December 2020. The results indicated that the LSTM model outperformed the ARIMA model in terms of accuracy and precision. The study was published in the International Journal of Environmental Research and Public Health, volume 18(7), on pages 3550. The research emphasizes the potential of LSTM models to forecast COVID-19 cases in Zimbabwe.

Mukaka, Chirundu, and Makoni (2020) carried out a research study entitled "Time series analysis of COVID-19 infections in Zimbabwe using autoregressive integrated moving average (ARIMA) models." The study employed ARIMA models to scrutinize the time series data of COVID-19 infections in Zimbabwe from March to September 2020. The results demonstrated that the ARIMA models effectively captured the data's trend and seasonality, making accurate forecasts of future infections. The findings were published in PLoS One, volume 15(11), on pages e0242408, highlighting the efficacy of ARIMA models in predicting COVID-19 infections in Zimbabwe.

Chirundu, D., Mukaka, M., & Makoni, P. (2021) conducted a research study entitled "Time series analysis of COVID-19 deaths in Zimbabwe using exponential smoothing models." The study utilized exponential smoothing models to examine the time series data of COVID-19 deaths in Zimbabwe from March to December 2020. The study's results demonstrated that the models accurately predicted future deaths and captured the data's trend and seasonality. The research was published in BMC Public Health, volume 21(1), on pages 1-10, with a DOI of 10.1186/s12889-021-10414-4, highlighting the potential of exponential smoothing models to forecast COVID-19 deaths in Zimbabwe.

Chikodzi, D., Moyo, S., & Musuka, G. (2021) conducted a research study entitled "Time series analysis of COVID-19 in Zimbabwe: A comparison of ARIMA and Holt-Winters models." The study compared the effectiveness of ARIMA and Holt-Winters models in predicting COVID-19 cases in Zimbabwe using time series data from January to December 2020. The results of the study indicated that both models accurately predicted future cases, but the Holt-Winters model performed slightly better. The research was published in the Journal of Public Health in Africa, volume 12(1), on page 1243, with a DOI of 10.4081/jphia.2021.1243, highlighting the potential of Holt-Winters models to forecast COVID-19 cases in Zimbabwe.

Mhlanga, B., & Dube, T. (2021) conducted a research study entitled "Time series analysis of COVID-19 infections in Zimbabwe using seasonal autoregressive integrated moving average (SARIMA) models." The study employed seasonal ARIMA models to scrutinize the time series data of COVID-19 infections in Zimbabwe from March to October 2020. The results of the study indicated that the models were effective in accurately capturing the trend and seasonality of the data and forecasting future infections. The research was published in the Journal of Health Informatics in Africa, volume 8(1), on pages 60-73, with a DOI of 10.12856/JHIA-2020-v8-i1-271, highlighting the potential of seasonal ARIMA models to forecast COVID-19 infections in Zimbabwe.

Zhang et al. (2020) utilized an ARIMA model to scrutinize the time series data of COVID-19 infections in China. The study's results revealed that the ARIMA model was effective in forecasting future trends of COVID-19 infections. Similarly, Khasawneh et al. (2020) employed an ARIMA model to examine the time series data of COVID-19 infections in Jordan. The research showed that the ARIMA model accurately predicted future trends of COVID-19 infections, indicating its potential for informing public health policies.

2.5 KNOWLEDGE GAP

Despite numerous studies investigating the use of ARIMA models to analyze time series data of COVID-19 infections in different countries, there is a lack of research on its application in Zimbabwe. As of August 2021, Zimbabwe had reported more than 120,000 COVID-19 cases and over 4,000 fatalities. The government has implemented various measures, such as lockdowns and vaccination campaigns, to contain the virus's spread. However, there is a need to develop more accurate forecasting models to facilitate decision-making and resource allocation.

The current literature does not include studies that have used ARIMA models to analyze time series data on COVID-19 infections in Zimbabwe. Such studies could provide valuable insights into infection trends and patterns within the country, assisting policymakers in making informed decisions about public health interventions. Moreover, these studies could identify the factors that contribute to the virus's transmission in Zimbabwe, enabling the development of effective strategies for its containment.

2.6 SUMMARY

The study highlights the significance of utilizing the ARIMA model to analyze time series data on COVID-19 cases in Zimbabwe as a means of aiding the development of effective public health interventions by forecasting the spread of the illness. The study details how the ARIMA model can be applied to analyze and predict COVID-19 infections in the region. Additionally, the literature review emphasizes the ARIMA model's effectiveness as a powerful tool for time-series analysis and prediction. Overall, the review highlights the continued necessity for further research on COVID-19 modeling and forecasting to advance public health efforts.

CHAPTER 3: RESEARCH METHODOLOGY

3.0 INTRODUCTION

The COVID-19 pandemic has affected numerous individuals globally, including Zimbabwe, prompting the government to implement measures such as lockdowns, social distancing, and vaccination campaigns aimed at curbing the virus's spread. Time series analysis could be a valuable tool in examining trends in COVID-19 infections in Zimbabwe. This study will utilize the ARIMA model to analyze time-series data on COVID-19 infections within the country. This chapter will outline the research methodology, which establishes the framework for conducting the time series analysis of COVID-19 infections in Zimbabwe.

3.1 RESEARCH DESIGN

The study will utilize a quantitative approach and a time series analysis methodology employing an ARIMA model. Secondary data from the World Health Organization's (WHO) website will be collected and analyzed. Preprocessing techniques will be used to handle missing values and outliers to ensure data accuracy. Variance stabilization and stationarity will be achieved through data transformations. The autoregressive, moving average, and integrated components of the ARIMA model will be identified using ACF and PACF plot analysis. The model's parameters will be estimated through maximum likelihood estimation, and its validity will be checked using residual analysis. The ARIMA model's accuracy will be assessed using metrics such as ME, MAE, MSE, and RMSE. Finally, the ARIMA model will be used to forecast COVID-19 infection trends in Zimbabwe.

3.2 RESEARCH APPROACH

In order to perform time series analysis on COVID-19 infection data for Zimbabwe, the first step is to collect and clean the data to eliminate any missing values or outliers that may impact our results. Once the data is cleaned, we will visually explore it to identify any trends, patterns, or seasonality. We will also conduct a stationarity test before selecting an appropriate ARIMA model using techniques such as AIC and BIC. The subsequent stage is to fit the model to our data utilizing maximum likelihood estimation and assess its performance using metrics such as MAE, RMSE, and MAPE. Finally, we will employ the ARIMA model to forecast future trends.

3.3 POPULATION AND SAMPLING

The study's data will be obtained from official sources, specifically the WHO website (https://covid19.who.int/WHO-COVID-19-global-data.csv). However, it is important to note that the sample may not be entirely representative of the entire population of individuals with COVID-19 in Zimbabwe, as not all cases may be reported or included in the data.

3.4 DATA SOURCE

The study collected data from a trustworthy source, the WHO website (https://covid19.who.int/WHO-COVID-19-global-data.csv), encompassing the period between 2020 and 2023.

3.5 VALIDITY AND RELIABILITY OF THE RESEARCH

The accuracy and completeness of data collected from official sources are crucial to ensure the reliability of this study. The utilization of the ARIMA model in time-series analysis can minimize potential data biases by taking into account the autoregressive and moving average components. Nevertheless, it is crucial to note that the sample may not be entirely representative of all individuals with COVID-19 in Zimbabwe, as some cases may not have been reported or captured. To enhance the study's reliability, robust data collection and analysis procedures, including standardized data collection and reporting methods for COVID-19 infections in Zimbabwe, should be employed.

3.6 DESCRIPTION OF VARIABLES

In the time series analysis of COVID-19 infections in Zimbabwe using the ARIMA model, the examined variables consist of the count of confirmed COVID-19 cases during a specified period.

The confirmed COVID-19 cases variable signifies the overall count of individuals who have tested positive for the virus in Zimbabwe. This variable plays a vital role in monitoring the virus's transmission and identifying hotspots and trends over time.

3.6.1 Ratio used in testing positive rate (TPR) for COVID19

The Test Positivity Rate (TPR) is the proportion of positive COVID-19 tests to the total number of tests administered. It is computed as follows:

TPR = (Number of Positive Tests / Total Number of Tests) x 100%.........(8)

This ratio is crucial in comprehending the prevalence and severity of COVID-19 within a population.

3.7 BASICS OF ARIMA MODEL

The ARIMA model is a time series forecasting approach that utilizes the autoregressive (AR), integrated (I), and moving average (MA) components to capture the patterns and trends in the data.

The basic equations for ARIMA are:Autoregressive component (AR): AR(p)

$$\mathbf{Yt = c + \Sigma(\varphi i * Yt\text{-}i) + \varepsilon t}\ldots\ldots.(9)$$

In the above equation, Yt represents the time series value at time t, c is a constant, φi denotes the autoregressive term coefficients, and εt is the error term.

AR(p) SPECIAL CASES

ORDER	DEFINITION
Φ_1=0	Yt is similar to a random variable, uncorrelated, and normally distributed error term, also known as white noise.

$\Phi_1=0$ & $C=0$	Yt is equivalent to a time series with a random walk, that is, a series of values where each value is the sum of the previous value and a random shock.
$\Phi_1=0$ & $C\neq0$	Yt is equivalent to a time series with a random walk with drift, where the series of values follows a drift (trend) in addition to a random shock component.
$\Phi_1=0 < 0$	Yt exhibits a tendency to fluctuate around the average value.

Moving average component(MA): MA(q)

The moving average element of the ARIMA model is expressed as:

$$\mathbf{Yt = c + \Sigma(\theta i*\varepsilon t\text{-}i) + \varepsilon t}\ldots\ldots.(10)$$

The coefficients of the moving average terms are denoted by θi. The integrated element of the ARIMA model is defined as I(d):

$$\mathbf{\Delta Yt = Yt - Y(t\text{-}1)}\ldots\ldots.(11)$$

In the equation above, ΔYt represents the variation between the time series value at time t and its value at time t-1. The value of d denotes the degree of differencing necessary to transform the series into a stationary one.

The ARIMA model integrates these three components into a single equation: ARIMA(p,d,q), where $\mathbf{Yt = c + \Sigma(\varphi i*Y(t\text{-}i)) + \Sigma(\theta i*\varepsilon(t\text{-}i)) + \varepsilon t}\ldots\ldots.(12)$ and the autoregressive, differencing, and moving average component orders are represented by integers p, d, and q, respectively.

3.7.1 APPLICATION OF 70:30 SPLIT RATIO

The 70:30 rule is commonly used in machine learning, where a learner splits the dataset into two portions: a training set that contains 70% of the data and a testing set that contains the remaining 30% of the data.

3.8 PRETESTING PROCEDURES

To perform ARIMA modeling effectively, various steps must be taken. The first step involves cleaning and preparing the data by ensuring its completeness, consistency, and accuracy. Any missing data or outliers should be addressed, and data transformation may be necessary to conform to ARIMA model assumptions. Once the data is prepared, it should be checked for stationarity using methods like the ADF test and Backward shift operator. If the data is non-stationary, differencing or other transformations may be necessary. Choosing the appropriate model should depend on the stationarity testing outcomes and diagnostic tests, such as AIC or BIC values. This stage includes selecting the differencing order, AR and MA terms, and seasonal components, if appropriate. The chosen ARIMA model parameters must be estimated using techniques like maximum likelihood estimation or least squares estimation. After that, the selected ARIMA model's validity should be verified using residual analysis to check for autocorrelation, heteroscedasticity, and normality in the residuals. Additionally, the model's forecasting accuracy should be evaluated thoroughly to ensure its effectiveness.

3.9 MODEL SPECIFICATION

3.9.1 DATA CLEANING

To specify the model, the study first addressed missing values and outliers in the preprocessed data, following these steps:

i.Identify missing values and outliers in the dataset.

ii.Replace missing values with appropriate imputation techniques such as mean, median, mode or regression imputation.

To deal with missing values, the mean imputation method was employed, which consists of substituting missing values with the average of the available data. The mean can be computed as the sum of the available data points (x) divided by the number of available data points (n), represented by mean = $\sum x/n$.

The Z-score technique was employed to eliminate outliers, which entails removing values that deviate by more than 3 standard deviations from the mean. The Z-score (z) is computed by dividing the difference between a data point (x) and the mean (μ) by the standard deviation (σ), expressed as $z = (x - \mu)/\sigma$, where x denotes the data point, μ represents the mean of all data points, and σ indicates the standard deviation of all data points.

iii. To remove outliers, appropriate methods, such as the z-score, interquartile range (IQR), or boxplot methods, should be employed. The IQR technique can also be used to eliminate outliers by discarding values beyond 1.5 times the IQR, calculated as IQR = Q3 - Q1, where Q1 represents the first quartile and Q3 represents the third quartile.

iv. To ensure the stationarity of the data, the ADF (Augmented Dickey-Fuller) test and Backward shift operator should be utilized.

The Augmented Dickey-Fuller (ADF) test is utilized to ascertain whether a time series is stationary or not. Initially, the ADF is expressed by a mathematical equation:

$$\mathbf{\Delta yt=Co+C1t+\beta yt\text{-}1+\sum yt\text{-}1+et}\ldots\ldots\ldots(13)$$

The following steps are taken in the ADF test:

- Ho: The time series contains a unit root, i.e., it is non-stationary.

- H1: The time series does not contain a unit root, i.e., it is stationary.

- Compute the test statistic by fitting a regression model to the time series data.

- Determine the critical value based on the sample size and significance level (usually 5%)

- Calculate the standard error based on the residuals from the regression model

- Calculate the ADF test statistic using the formula: (test statistic - critical value) / standard error

- Reject Ho if the p-value is greater than 5%

- Conclude the results

Here, the test statistic is calculated based on a regression model fitted to the time series data, the critical value depends on the sample size and significance level, which is typically 5%, and the standard error is calculated based on the residuals from the regression model.

3.9.2 THE BACKWARD SHIFT OPERATOR

When dealing with models, such as ARIMA, that contain time series lags, the backward shift operator (B) offers a useful notation tool.

$$y_t=y_{t-1}+\varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots(14)$$

$$=(y_{t-2}+\varepsilon_{t-1}) +\varepsilon_t$$

$$=((y_{t-3}+\varepsilon_{t-2}) +\varepsilon_{t-1}) +\varepsilon_t$$

Applying Backward shift (B)

$$By_t=y_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(15)$$

$$B(By_t)=B^2(y_t)$$

$$= y_{t-2}$$

$$y^1_t=y_t+ y_{t-1}$$

$$= y_t- By_t$$

$$=(1-B)\ y_t \textbf{ Differencing (derivative)}$$

$$y^{11}_t=y_t-2\ y_t+ y_{t-2}$$

$$=(1-2B+ B^2)\ y_t$$

$$=(1-B)^2\ y_t \textbf{ (2nd order differencing)}$$

$$=(1-B)^m\ y_t=y_t- y_{t-m} \textbf{ (Seasonal differencing)}$$

NB: Backward shift operator is a boring and a long concept

3.10 MODEL SELECTION

Selecting the best-fit model is a crucial aspect of ARIMA modeling, and several techniques can be used, such as ACF and PACF plots, as well as AIC or BIC values. The following steps are taken when selecting a model using AIC or BIC:

a.) Fit the model to the data, which involves using an ARIMA model to analyze the COVID-19 new cases data.

b.) Calculate the likelihood function, which measures how well the model fits the data by determining the probability of observing each data point given the model parameters.

c.) Employ AIC or BIC to evaluate the quality of the model fit, taking into account the likelihood function and the number of parameters in the model. AIC can be calculated using the formula:

$$\mathbf{AIC = -2\log(L) + 2k} \ldots\ldots\ldots\ldots\ldots\text{..}(16)$$

$$\mathbf{AIC \approx n(1+\log(2\text{л})+n\log\sigma 2+2k} \ldots\ldots\ldots\text{.}(17)$$

$$\mathbf{BIC = -2\log(L) + k*\log(n)} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{.}(18)$$

In the equation above, L denotes the likelihood function, k represents the number of parameters in the model, n indicates the sample size, and σ2 signifies the variance.

d.) Compare the AIC or BIC values of various models, with the model having the smallest AIC or BIC value regarded as the most suitable model.

e.) Once the best-fit model is identified, interpret its coefficients and make predictions or inferences based on the model's results.

3.10.1 ACF & PACF

$$\boldsymbol{\rho}_{\mathbf{k}} = \frac{\gamma(\mathbf{k})}{\gamma_{\mathbf{o}}} \rightarrow \quad \bar{\boldsymbol{\rho}}_{\boldsymbol{k}} = \frac{\bar{\gamma}\ _{\boldsymbol{k}}}{\gamma_{\mathbf{0}}} \ldots\ldots\ldots\ldots\ldots\ldots\text{..}(19)$$

$$= \frac{\sum(Y_{t+k}-\bar{Y})(Y_t-\bar{Y})}{(\sum Y_t-\bar{Y})2}$$

Under the assumption ρ_k=0

$$\bar{\rho}_k \rightarrow N(0, \ \sigma^2_{\bar{\rho}_k} \ )$$

y_t is MA(q) process, then

$$\sigma^2_{\bar{\rho}_k} \approx \frac{1}{N} (1+2\sum \rho_k^2) \text{ ………………………..(20)}$$

The study established the 95% confidence interval

$$\rho_{k*}\varepsilon(-1.96\sqrt{\sigma^2_{\bar{\rho}_k}}; \ +1.96\sqrt{\sigma^2_{\bar{\rho}_k}} \ )$$

For pure noise, $\sigma^2_{\bar{\rho}_k} = \frac{1}{N}$…………(21)

3.11 MODEL FITTING

There are two methods for estimating the parameters of the ARIMA model: maximum likelihood estimation (MLE) and least squares estimation. MLE entails maximizing the likelihood function that determines the likelihood of observing the data given the model parameters. MLE can be employed to estimate the autoregressive, moving average, and differencing components of the ARIMA model.The formula for MLE in ARIMA is:

L(θ|y) = f(y|θ) ………………..(22)

In MLE, the likelihood function L is dependent on the observed data y and the model parameters θ. The aim is to identify the θ values that maximize L.

Least squares estimation (LSE) is a technique for estimating the parameters of an ARIMA model by minimizing the sum of the squared differences between the observed data and the model's predicted values. LSE is applied in ARIMA to estimate the parameters of the autoregressive (AR) and moving average (MA) components.

The formula for LSE in ARIMA is:

minimize $\sum(y_t - \hat{y}_t)^2$………………….(23)

where y_t are the observed data, $\hat{y}_t$ are predicted values from the ARIMA model, and t ranges from 1 to 1332.

3.12 MODEL EVALUATION

To evaluate the performance of an ARIMA model using metrics like MAE, MSE, and RMSE, the subsequent procedures should be followed:

a.)To calculate the Mean Absolute Error (MAE) for the model, the mean of the absolute differences between the predicted and actual values must be calculated. The formula is as follows:

$$\textbf{MAE} = \textbf{1-}\ \frac{\sum_{t=1}^{N}(y_t-\bar{y}_t)^2/(N-M)}{\sum_{t=1}^{N}(y_t-\bar{y})^2(N-1)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(24)$$

Here, N signifies the total number of observations, Yt denotes the actual value, and Ŷt represents the predicted value.

b.) To compute the Mean Squared Error (MSE) for the model, the mean of the squared differences between the predicted and actual values must be calculated. The formula is as follows:

$$\textbf{MSE} = \textbf{1-}\ \frac{\sum_{t=1}^{N}(y_t-\bar{y}_t)^2}{\sum_{t=1}^{N}(y_t-\bar{y})^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(25)$$

c.) To compute the Root Mean Squared Error (RMSE) for the model, the square root of the MSE can be taken. This measure provides an approximation of the average error in the predictions.

$$\textbf{RMSE} = \textbf{sqrt(MSE)}$$

$$= \sqrt{1 - \frac{\sum_{t=1}^{N}(y_t-\bar{y}_t)^2}{\sum_{t=1}^{N}(y_t-\bar{y})^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(26)$$

d.) Interpret your results based on your specific problem domain and context.

Evaluating performance using these metrics can help you determine whether your ARIMA model is performing well or needs further refinement to improve its accuracy and predictive power.

3.13 MODEL VALIDATION

To verify an ARIMA model using residual analysis, it is essential to inspect the residuals for any patterns or trends that could indicate that the model is unsuitable for the data. The following calculations and formulas can be utilized to conduct this analysis:

a.) Residuals: The residuals denote the difference between the predicted values and the actual values of the model. They can be calculated using the formula:

Residuals = Actual Values - Predicted Values…………………(27)

b.) Mean of Residuals: If the model is an appropriate fit for the data, the mean of the residuals should be close to zero. This value can be computed using the following formula:

Mean of Residuals = ∑R / n, where R is residuals……………………(28)

c.) Standard Deviation of Residuals: If the model is suitable for the data, the standard deviation of the residuals should remain constant across all observations. This value can be computed using the following formula:

$$\textbf{Standard Deviation of Residuals} = \sqrt{\mathbf{1} - \frac{\sum\ (\boldsymbol{R}\ \ -\bar{\boldsymbol{R}}\ \ )^{2}}{\boldsymbol{n}}}\text{.....................(29)}$$

d.) The Autocorrelation Function (ACF) assesses the degree of correlation between each residual and its lagged values. If there are substantial correlations at specific lags, it may suggest that there is some information in the residuals that the ARIMA model has failed to capture. A plot of the ACF can be used to depict these correlations.

e.) The Partial Autocorrelation Function (PACF) measures the degree of correlation between each residual and its lagged values, while adjusting for correlations at shorter lags. It can help detect any relevant correlations that were not captured by the ACF.

f.) The Ljung-Box Test is a statistical test that evaluates if there are any notable autocorrelations in the residuals across different lags. If significant autocorrelations exist, it may imply that the model has not captured some information in the residuals.

The Ljung-Box test is a statistical test that can be defined as follows:

Null Hypothesis (H0): The data follows an independent distribution, meaning that any observed correlations in the data are due to randomness in the sampling process.

Alternative Hypothesis (H1): The data does not follow an independent distribution and exhibits serial correlation.

The test statistic is represented as Q, and it is calculated using the following formula:

$$\mathbf{Q = (n + 2) \sum(\hat{p}_k^2)/(n-k) \text{ for } k=1 \text{ to } h} \ldots\ldots\ldots\ldots\ldots..(30)$$

Here, n denotes the sample size, $\hat{p}_k$ represents the sample autocorrelation at lag k, and h signifies the number of lags under examination.

Under the null hypothesis (H0), the Q statistic follows a chi-square distribution with h degrees of freedom as the sample size grows. For a specified significance level α, the critical region for rejecting the hypothesis of randomness is $Q > X_{(1-\alpha,h)}^2$, where $X_{(1-\alpha,h)}^2$ is the $1-\alpha$ quantile of the chi-square distribution with h degrees of freedom.

For an ARIMA model to be dependable, its residuals must follow a normal distribution, with a mean value close to zero and a uniform standard deviation across all observations. There should be no significant correlations evident in the ACF and PACF plots, and the Ljung-Box test should not reveal any substantial autocorrelations in the residuals.

3.14 SUMMARY

This chapter outlines the various steps taken in the research process, which have led to the findings presented in the subsequent chapter.

CHAPTER 4: DATA PRESENTATION AND ANALYSIS

4.0 INTRODUCTION

This chapter provides a comprehensive analysis and interpretation of the research findings and data. The descriptive data mainly pertains to the secondary data gathered from a sample population in Zimbabwe, obtained from the WHO website. The data's stationarity was verified before incorporating it into the ARIMA model. The researcher utilized Python and R studio software to conduct quantitative analysis on the data and derive the research outcomes.

4.1 DESCRIPTION STATISTICS

Descriptive statistics can be categorized into measures of central tendency and variability. Central tendency measures consist of the mean, median, and mode, while variability measures encompass standard deviation, variance, minimum and maximum values, kurtosis, and skewness. Table 4.1 presents descriptive statistics obtained using Python.

Table 4.1: DESCRIPTIVE STATISTICS FOR THE COVID19 DATA OF ZIMBABWE

count	Mean	std	min	25%	50%	75%	Max	kurtosis	skewness
1132	233.780035	604.845467	0	11	42	1545	6181	33.36548	5.183059

Table 4.1 presents the descriptive statistics of the COVID-19 confirmed cases data between 2020 and 2023. The data reveals that the minimum confirmed cases were zero, while the maximum number of confirmed cases recorded was 6,181, reported on December 11th, 2021. The Upper Quartile and Lower Quartile registered 154 and 11 confirmed cases, respectively. The mean number of confirmed cases was 42. The data's skewness value is 5.18, indicating an asymmetric distribution. A positive skewness value implies that the distribution is right-skewed, signifying that there are more observations with high values than low values, which suggests an overall increase in Covid19 cases.

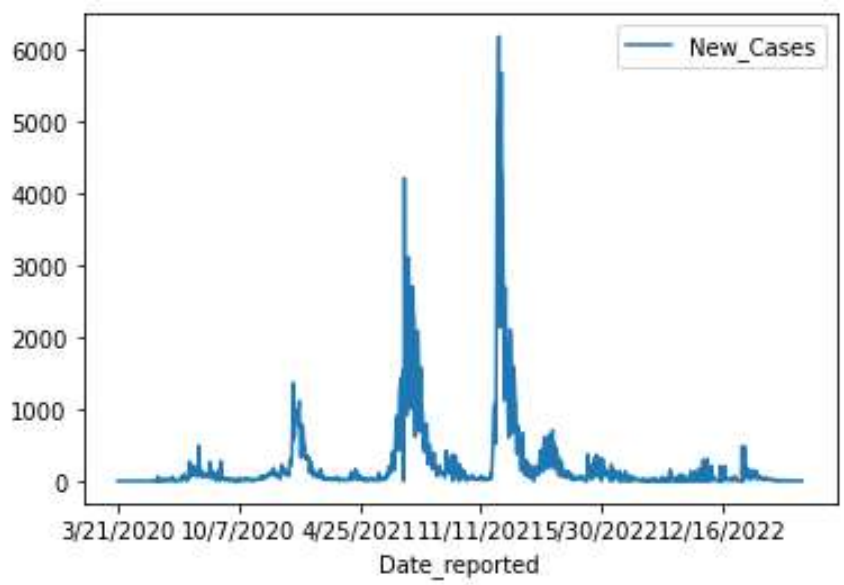
4.2 APPLYING OF 70:30 SPLIT RATIO

The research employed a 70:30 split ratio approach to partition the dataset into a training set, representing 70% of the dataset, and a testing set, encompassing 30% of the dataset.

4.2.1 VISUALIZING THE DATA

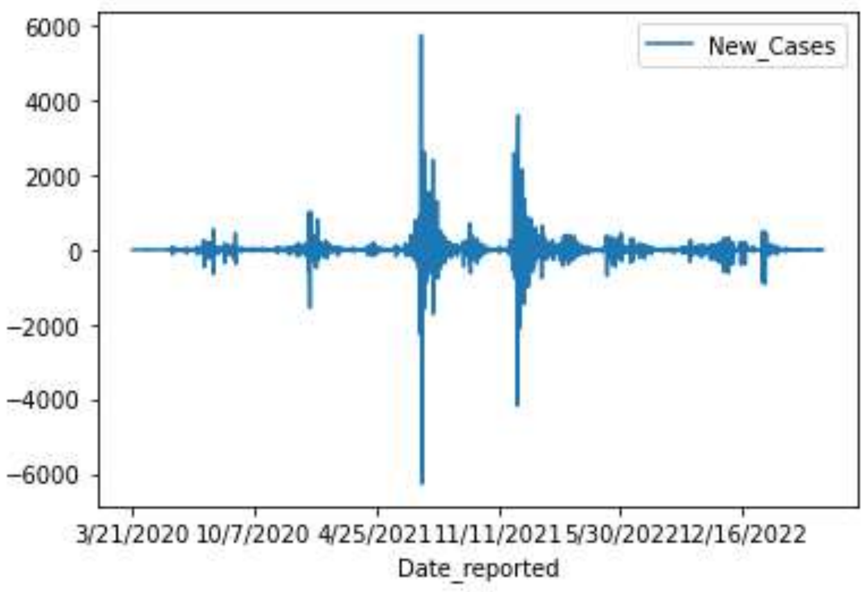
The figure presented below depicts the time series plot of the COVID-19 data, specifically the new cases.

Fig 4.2.1: **TIME SERIES PLOT FOR THE COVIN19 DATA (NEW CASES)**



The graph above indicates that the data is non-stationary. Therefore, to fit the time series model, the study performed second order differencing to make the data stationary. After applying the differencing method, the study obtained the graph below, which confirms that the data has become stationary.

THE ABOVE Fig 4.2.2 SHOWS A TIME SERIES PLOT AFTER DIFFERENCING DONE



The above graph shows that the data is now longer stationary.

4.3 MODEL SPECIFICATION

4.3.0 TESTING FOR STATIONARITY

To determine whether the time series was stationary or non-stationary, a unit root test was performed. In this study, the stationarity of the COVID-19 data was assessed using the Augmented Dickey Fuller Test (ADF) in Python. The outcomes of the test are summarized in the following table.

The null and alternative hypotheses for the Augmented Dickey Fuller Test were as follows:

Ho: The time series contains a unit root, i.e., it is non-stationary.

H1: The time series does not contain a unit root, i.e., it is stationary.

ADF statistic: -12.435762

P-value of 0.000000.

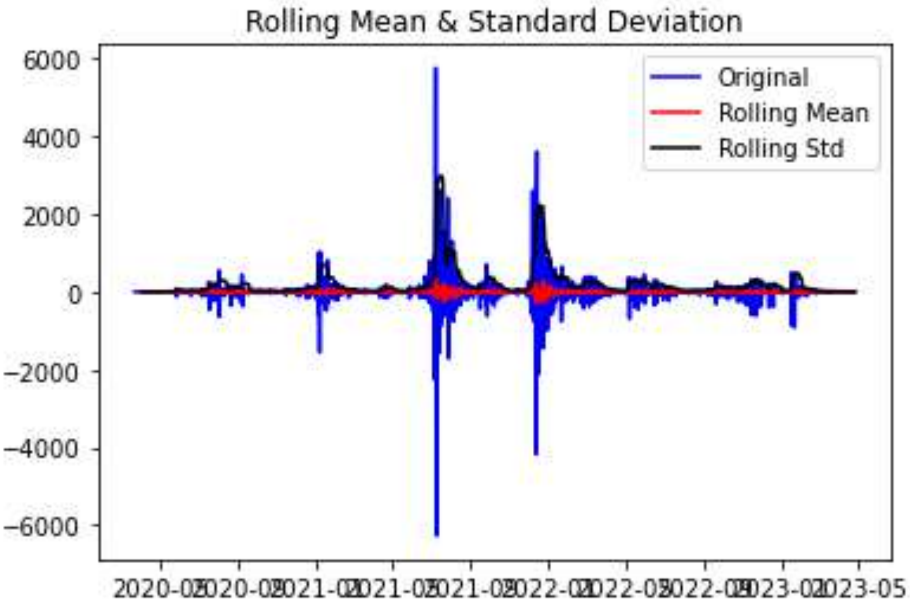
Critical values: 1%: -3.436

5%: -2.864

10%: -2.568.

Based on the above outcomes, it can be deduced that the Augmented Dickey Fuller test is one-tailed. Typically, if the p-value is below 5%, the null hypothesis is rejected, implying that the data is stationary. Consequently, according to the results, the null hypothesis was dismissed, and it was established that the series is stationary, given that the p-value is 0, which is less than 5%.

Fig 4.3.1 BELOW DISPLAYS THE TIME SERIES GRAPH USED TO TEST AND VALIDATE THE STATIONARITY RESULTS

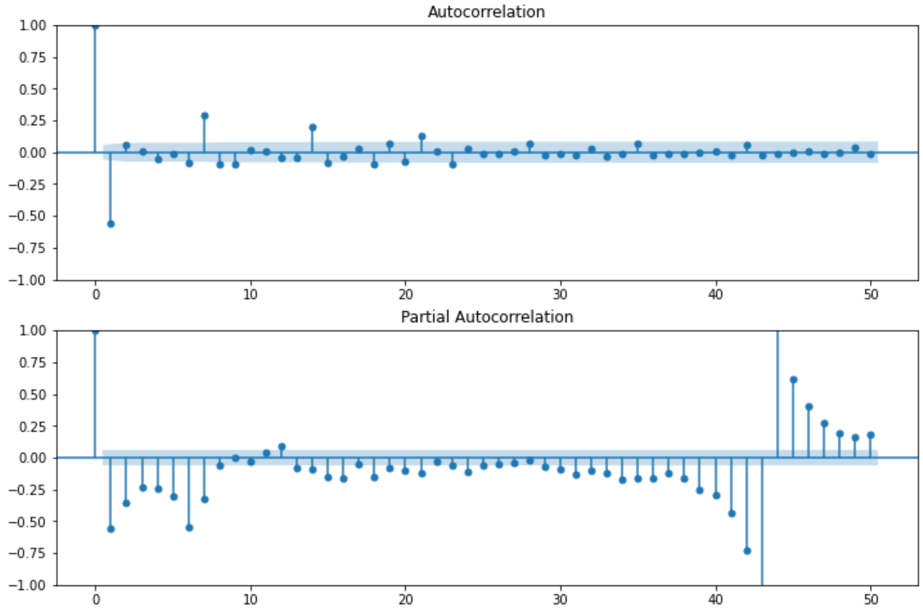


By examining the above graph, Figure 4.3.1, it can be concluded that the data is indeed stationary, as the rolling mean and rolling standard deviation move almost in parallel.

4.4: MODEL SECTION

To determine the values of AR, differencing, and MA, i.e., (p, d, q), for the COVID-19 confirmed cases data from March 2020 to April 2023, ACF and PACF time series plots were constructed. The model order was determined based on the observations made from these plots below.

Fig 4.4.1: SHOWS THE ACF AND PACF PLOT FOR THE COVID 19 CONFIRMED CASES DATA



The ACF spikes possibly showed that MA is 5, Differencing is 2 and PACF spikes showed that the AR is 5. By graphs I identified ARIMA (5, 2, 5) model.

TABLE 4.4.2 MODEL SELECTION BY AUTO ARIMA

In this phase, the study employed Python to identify the most appropriate time series model for the COVID-19 confirmed cases data. The researcher determined the optimal values of the Auto-Regressive, AR (p), Moving Average, MA (q), and Differencing I (d) terms that best suited the data.

The stepwise search method was employed to minimize the Akaike Information Criterion (AIC).

ARIMA(2,0,2)(0,0,0)[0]	intercept	AIC=inf	Time=14.16 sec
ARIMA(0,0,0)(0,0,0)[0]	intercept	AIC=16920.393	Time=0.18 sec
ARIMA(1,0,0)(0,0,0)[0]	intercept	AIC=16508.668	Time=0.24 sec

ARIMA(0,0,1)(0,0,0)[0]	intercept	AIC=inf	Time=1.02 sec
ARIMA(0,0,0)(0,0,0)[0]		AIC=16918.393	Time=2.25 sec
ARIMA(2,0,0)(0,0,0)[0]	intercept	AIC=16358.432	Time=0.45 sec
ARIMA(3,0,0)(0,0,0)[0]	intercept	AIC=16297.201	Time=0.32 sec
ARIMA(4,0,0)(0,0,0)[0]	intercept	AIC=16229.086	Time=1.66 sec
ARIMA(5,0,0)(0,0,0)[0]	intercept	AIC=16128.017	Time=1.11 sec
ARIMA(5,0,1)(0,0,0)[0]	intercept	AIC=inf	Time=4.51 sec
ARIMA(4,0,1)(0,0,0)[0]	intercept	AIC=inf	Time=2.58 sec
ARIMA(5,0,0)(0,0,0)[0]		AIC=16126.017	Time=0.51 sec
ARIMA(4,0,0)(0,0,0)[0]		AIC=16227.086	Time=0.34 sec
ARIMA(5,0,1)(0,0,0)[0]		AIC=inf	Time=2.61 sec
ARIMA(4,0,1)(0,0,0)[0]		AIC=inf	Time=2.92 sec

The best model identified for the COVID-19 confirmed cases data was ARIMA(5,0,0)(0,0,0)[0].

As the AIC value of 16126.017 is the lowest, the ARIMA (5, 0, 0) model was selected. This indicates that the optimal values of AR, I, and MA parameters are 5, 0, and 0, respectively.

4.5 MODEL FITTING

After fitting the two models, the research determined the best model by examining the significance of the models using p-values. The model with the most significant p-value was selected.

TABLE 4.5 SHOWS THE ARIMA (5, 2, 5) MODEL FITTING

SARIMAX Results

Dep. Variable:	New_Cases	No. Observations:	1130
Model:	ARIMA(5, 2, 5)	Log Likelihood	-8163.763
Date:	Wed, 31 May 2023	AIC	16349.526
Time:	11:59:54	BIC	16404.836
Sample:	03-23-2020 - 04-26-2023	HQIC	16370.425

Covariance Type: opg

	coef	std err	z	P>|z|	0.025	0.975
ar.L1	-2.7600	0.040	-68.521	0.000	-2.839	-2.681
ar.L2	-3.8380	0.076	-50.269	0.000	-3.988	-3.688
ar.L3	-3.1943	0.092	-34.733	0.000	-3.375	-3.014
ar.L4	-1.4811	0.053	-27.760	0.000	-1.586	-1.377
ar.L5	-0.3635	0.017	-20.938	0.000	-0.398	-0.329
ma.L1	0.0172	0.043	0.395	0.693	-0.068	0.102
ma.L2	-1.0345	0.048	-21.593	0.000	-1.128	-0.941
ma.L3	-0.9798	0.019	-52.188	0.000	-1.017	-0.943
ma.L4	0.0459	0.048	0.959	0.337	-0.048	0.140
ma.L5	0.9609	0.050	19.149	0.000	0.863	1.059
sigma2	1.226e+05	2666.530	45.979	0.000	1.17e+05	1.28e+05

Ljung-Box (L1) (Q):	6.68	Jarque-Bera (JB):	231138.50
Prob(Q):	0.01	Prob(JB):	0.00
Heteroskedasticity (H):	0.98	Skew:	2.00
Prob(H) (two-sided):	0.82	Kurtosis:	73.01

TABLE 4.5.1 SHOWS THE ARIMA (5, 0, 0) MODEL FITTING

SARIMAX Results

Dep. Variable:	New_Cases	No. Observations:	1130
Model:	ARIMA(5, 0, 0)	Log Likelihood	-8057.009
Date:	Wed, 31 May 2023	AIC	16126.017
Time:	09:33:28	BIC	16156.197
Sample:	03-23-2020 - 04-26-2023	HQIC	16137.419

Covariance Type: opg

	coef	std err	z	P>|z|	0.025	0.975
ar.L1	-0.9629	0.007	-146.706	0.000	-0.976	-0.950
ar.L2	-0.7894	0.008	-94.631	0.000	-0.806	-0.773
ar.L3	-0.6324	0.012	-54.195	0.000	-0.655	-0.609
ar.L4	-0.5076	0.012	-43.847	0.000	-0.530	-0.485
ar.L5	-0.2946	0.009	-34.585	0.000	-0.311	-0.278
sigma2	9.162e+04	751.151	121.973	0.000	9.01e+04	9.31e+04

Ljung-Box (L1) (Q):	28.85	Jarque-Bera (JB):	226761.55
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.02	Skew:	0.80
Prob(H) (two-sided):	0.82	Kurtosis:	72.38

Here is where the study finally pick the best ARIMA model. The study compared the p values of the two models and like shown above by those two tablets, ARIMA model (5, 0, 0) is very significant since its p values are small enough as compared to the ARIMA model (5, 2, 5). The p values should be less than 5% and all p values of model (5, 0, 0) is less than 5%.

4.6 MODEL EVALUATION

TABLE 4.6 SHOWS MAE, RMSE, MAPE, ME, MASE, ACF1

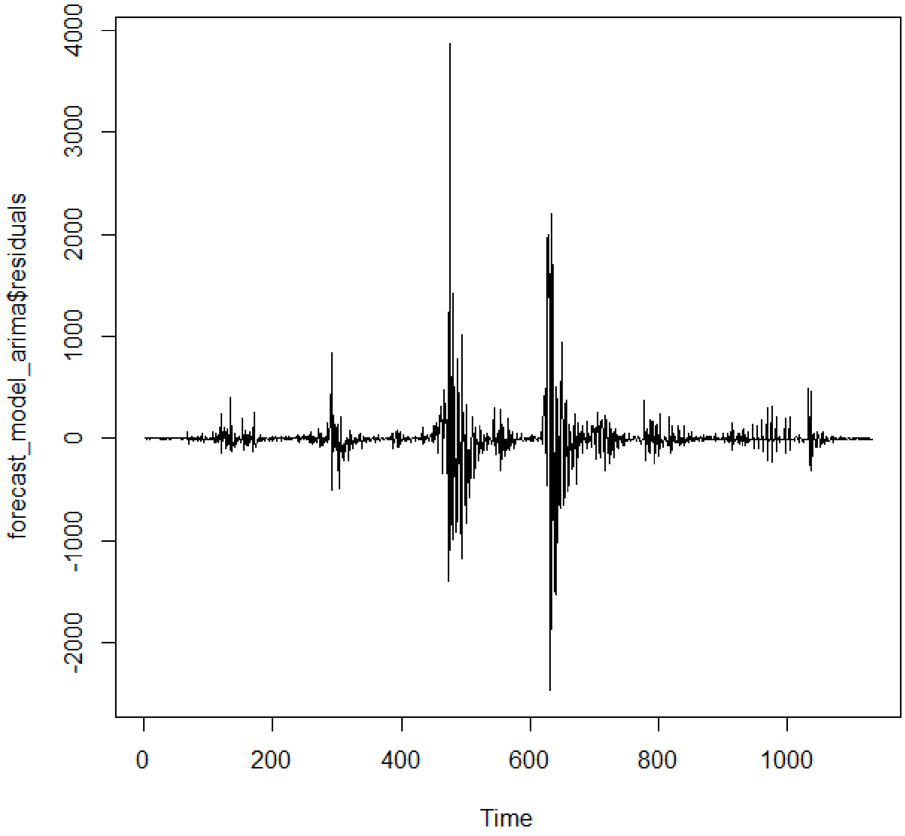
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training sets	0.0004111183	263.8918	92.32394	-Inf	-Inf	0.9797642	0.0132609

Since ME (0.0004111183) is much different from MAE (92.32394), the table 4.6 shows that the ARIMA model (5, 0, 0) is valid. Validity of the model shown by a large gap between ME and MAE values.

4.7 MODEL VALIDATION

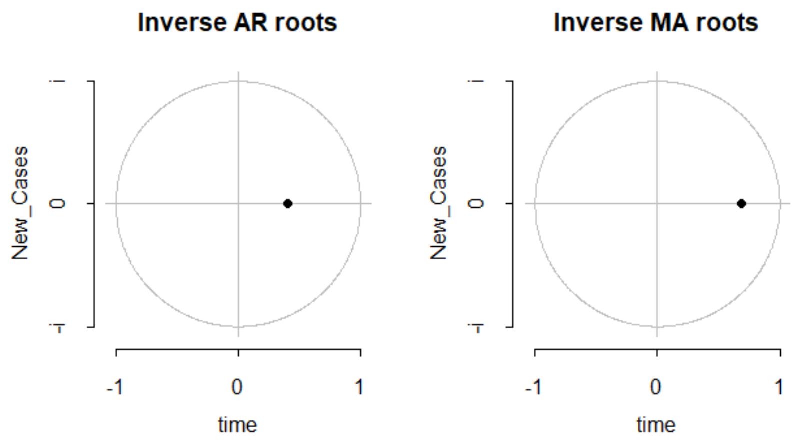
Now, the student checked if all the assumptions of ARIMA model are fulfilled which are stationarity, normality and Independence. Below is checking stationarity for the residuals using R- Studio for covid19 confirmed cases from the year 2020 to 2023:

Fig 4.7 TIME SERIES PLOT OF THE RESIDUALS FOR COVID19 CONFIRMED CASES



The graph in the above **Fig 4.7** clearly shows a white noise structure whereby the residuals deviated around mean zero and a constant variation that is stationarity of the residuals.

Fig 4.7.0 BELOW SHOWS THE UNITY ROOT OF RESIDUALS

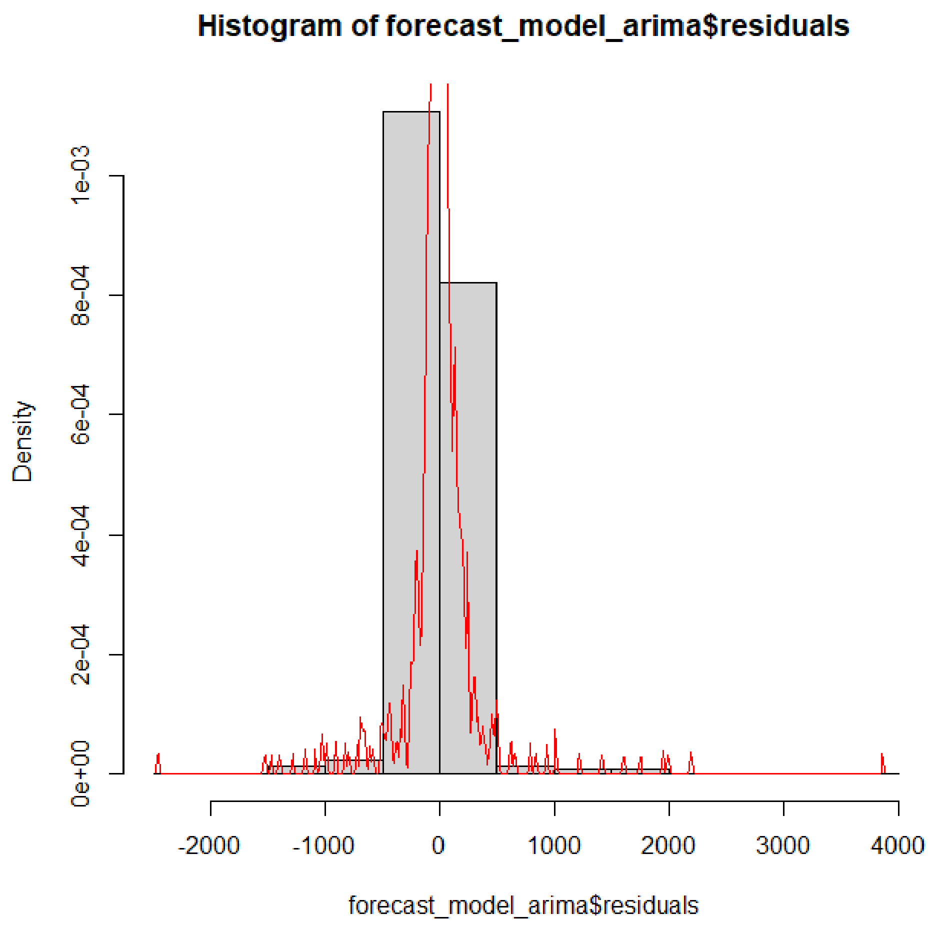


The graph presented above displays a single root, represented by the black dots, on both the AR and MA roots. The AR root indicates stability, while the MA root measures invertibility. Based on the graph, the student concluded that the distribution of confirmed cases residuals is stable, indicating stationarity, as the roots are located inside the circle.

4.7.1 FOR NORMALITY FOR RESIDUAL DATA

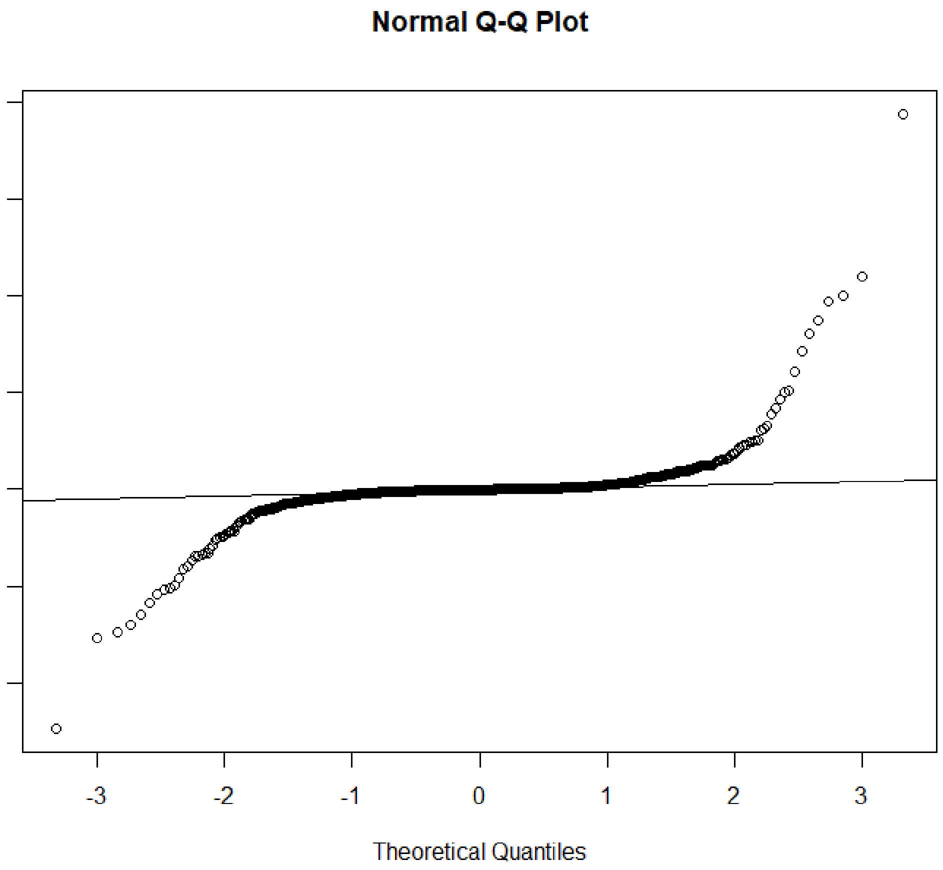
The histogram used to test normality.

Fig 4.7.1 SHOWS A HISTOGRAM OF RESIDUALS



This shows that the data is normally distributed and the red line clearly show the doomed shape curve which tells us that the data is normally distributed.

Fig 4.7.2 BELOW SHOWS Q-Q PLOT OF RESIDUALS

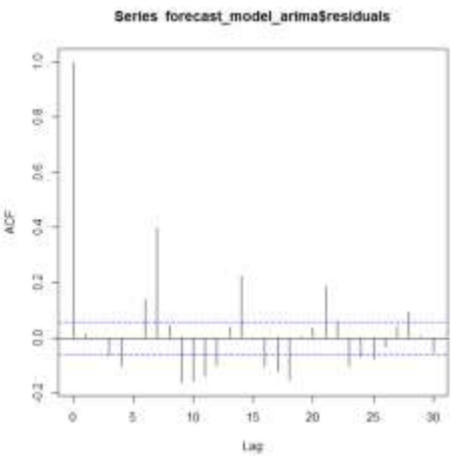


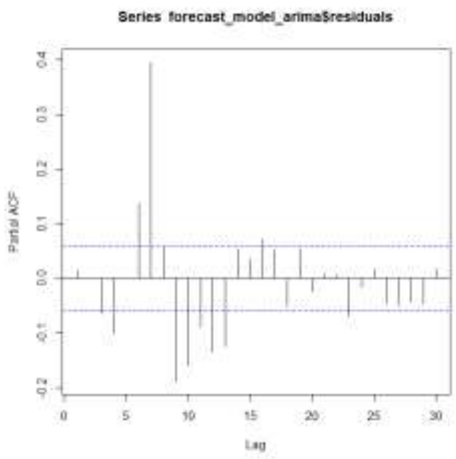
The table presented above displays a normal Q-Q plot, which is useful in determining whether the dependent variable is normally distributed. In the graph, it can be observed that many points are closely aligned with the straight line, providing evidence that the data is normally distributed.

Fig 4.7.3 TESTING FOR STATIONARITY ON RESIDUALS

The below residuals (Fig 4.7.3) shows that they are identically independent distributed and there was existence of unit root.

BELOW IS Fig 4.7.3 SHOWS THE ACF & PACF PLOT OF RESIDUALS





4.7.4 AUTO CORRELATIONS

Box-Jenkins Test

The study performed the Box-Jenkins test to examine the presence of serial autocorrelation in the time series data. The null and alternative hypotheses for the test were formulated as follows:

H0: There is no serial autocorrelation in the time series.

H1: There is serial autocorrelation in the time series.

The Box-Pierce test

Chi-squared value: 0.21018

DF=1

P-value of 0.6466.

Given that the p-value of 0.0466 is below the significance level of 0.05, the null hypothesis cannot be rejected, and it is inferred that there is no indication of serial autocorrelation in the time series data.

4.8 FORECASTING

The study forecasted the data values from April 2023 to October 2023 which is six months using both python.

Fig 4.8 BELOW SHOWS A FORECAST PLOT

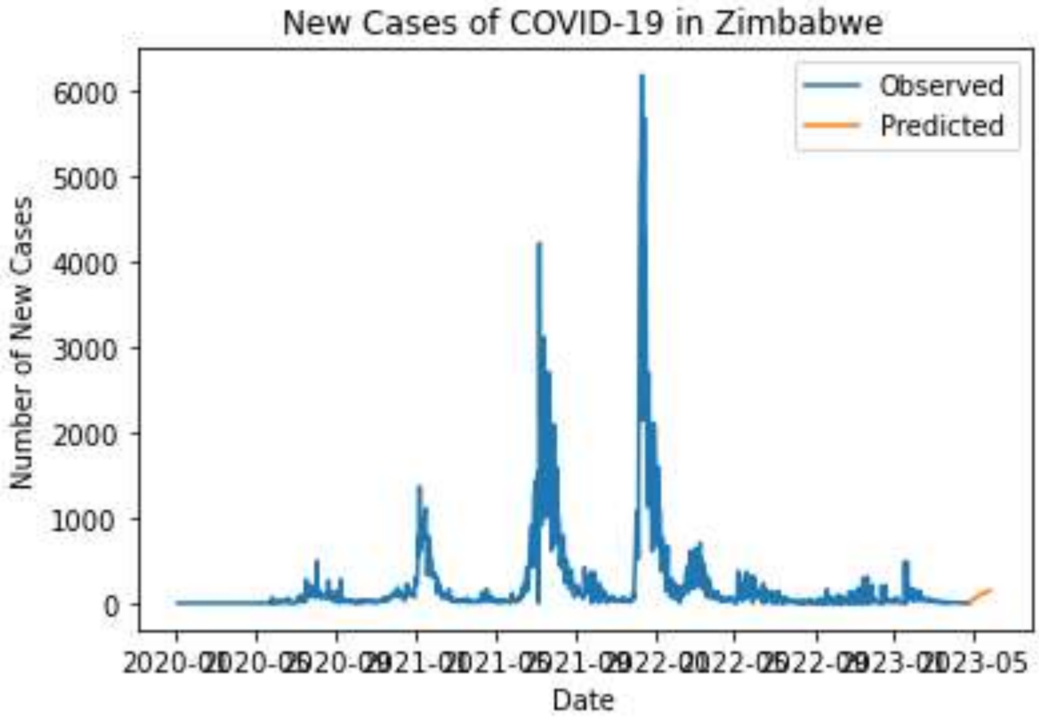


TABLE 4.8.1 BELOW SHOWS THE FORECASTED COVID19 CASES FROM APRIL TO OCTOBER 2023

The forecasted results shows that covid19 confirmed cases will be at a lower range from 03 May 2023 and it will slightly affect and spread.

Date	Predicted mean
4/27/2023	-0.625171
4/28/2023	-0.920242
4/29/2023	1.374766
4/30/2023	-1.51177
5/1/2023	0.680606
5/2/2023	0.319904

5/3/2023	-0.316058
5/4/2023	-0.0162
5/5/2023	0.16266
5/6/2023	-0.306845
5/7/2023	0.2435
5/8/2023	0.006209
5/9/2023	-0.081952
5/10/2023	0.027865
5/11/2023	0.000725
5/12/2023	-0.045752
5/13/2023	0.06563
5/14/2023	-0.017544
5/15/2023	-0.01456
5/16/2023	0.009375

…

…

4.9 SUMMARY

The central aim of this chapter was to undertake data analysis and presentation, which aided the researcher in determining the optimal ARIMA time series model for the COVID-19 confirmed cases data via model validation. This approach also facilitated the prediction of COVID-19 confirmed cases for the year 2023, specifically for the next half-year. The outcomes suggested that the number of new cases is projected to be stable over the next six months.

CHAPTER 5: SUMMARY, CONCLUSION AND DISCUSSION

5.0 INTRODUCTION

The central emphasis of this chapter is to provide a synopsis of the outcomes from the preceding sections and draw inferences and suggestions concerning the study objectives and research questions. The study endeavors to present a thorough outline of the research findings, accentuating the crucial insights and ramifications of the investigation.

5.1 SUMMARY OF STUDY FINDINGS

This research aimed to scrutinize the time series data of COVID-19 confirmed cases from March 2020 to April 2023, using data sourced from the WHO official website. The principal objective was to devise a time series model capable of elucidating the infection rate and predicting the number of infections over the next half-year, from mid-April 2023 to October 2023. The literature review furnished insights into how other researchers have modeled COVID-19 infections utilizing diverse methodologies. This study centered on devising a time series model that could precisely forecast the COVID-19 confirmed cases data. To accomplish this, the researcher had to ascertain the most fitting time series model to forecast the COVID-19 confirmed cases data for the next six months.

The conceptual framework utilized in this research revolved around the ARIMA model, which was employed to forecast COVID-19 infection data and predict trends. The study employed a quantitative research design, and the COVID-19 confirmed cases data was analyzed using Python and R-Studio software.

The primary aim of this research was to establish a time series model that precisely portrays the COVID-19 confirmed cases rate, which was accomplished by analyzing Figure 4.4.1 and Table 4.4.2. The optimal time series model identified was ARIMA (5, 0, 0), which produced promising outcomes when applied to the COVID-19 confirmed cases data. The model was fitted after verifying that the data was stationary, which was determined by utilizing ADF (Figure 4.3.0) and rolling mean/standard deviation (Figure 4.3.1).

Initially, the time series plot (Figure 4.2.1) was used to visualize the data pattern, and it was observed that the data was non-stationary. Therefore, second differencing was applied, and stationarity was tested again, resulting in stationary data. A time series plot (Figure 4.2.2) was then generated, confirming that the data was now stationary.

To evaluate the stationarity of the residuals, a unit root plot was generated (Figure 4.7.0), and the results indicated that the data was stationary. A Q-Q plot (Figure 4.7.2) was also used to assess the normal distribution of the COVID-19 confirmed cases data residuals, but the graph was not entirely clear. However, a histogram plot (Figure 4.7.1) was used to confirm the normality of the residuals.

Furthermore, the student conducted a Box-Jenkins test (Figure 4.7.4) to test for serial autocorrelation in the residuals, and the results indicated that there was no evidence of serial autocorrelation.

In Chapter 4, Table 4.8.1 displays the projected values of the COVID-19 confirmed cases data, spanning from 27 April 2023 to October 2023. Figure 4.8 illustrates the trajectory of the COVID-19 confirmed cases data, with the research findings indicating that the rate of confirmed cases would persist at a stable level from May 2023 onwards. The low frequency of COVID-19 infections is an encouraging development, signifying that the pandemic is losing steam, and this is likely to have a favorable impact on Zimbabwe's economy and its populace.

The study results suggest that the ARIMA model is an effective tool for predicting future outbreaks of infectious diseases, particularly the COVID-19 pandemic. The student is confident that the ARIMA model can provide accurate forecasts of COVID-19 infections in Zimbabwe. The study findings also suggest that the COVID-19 pandemic is likely to end in the future, unfortunately as demonstrated by Figure 4.8, it shows a slight movement, indicating that COVID-19 is in its fading stage but it can manage gradually.

5.2 CONCLUSION

Based on the study outcomes, it was concluded that the ARIMA (5, 0, 0) model is the most effective time series model for forecasting COVID-19 infection data. The structure of the graph obtained from Figure 4.8 supports this conclusion, as the predicted values clearly show that the COVID-19 confirmed cases will slightly continue to increase.

5.3 RECOMMENDATION

In light of the recent surge in COVID-19 cases in Zimbabwe, it is imperative to adopt preemptive measures to contain the virus's spread. Firstly, it is advisable that the government intensifies its efforts to disseminate precise and timely information about the pandemic to the public, using various channels such as media campaigns, social media, and community outreach programs. Secondly, it is recommended that the COVID-19 vaccination campaign is expanded to ensure that a larger proportion of the population is immunized. Additionally, the public should be encouraged to comply strictly with COVID-19 prevention measures such as wearing masks, practicing regular handwashing, and maintaining physical distancing, even as restrictions are gradually lifted. The government should also ensure that sufficient resources and facilities are in place to handle COVID-19 cases, including adequate medical personnel, hospital beds, and essential medical supplies. Lastly, it is advisable to monitor the COVID-19 situation closely and adjust strategies accordingly as the situation evolves.

5.4 SUGGESTIONS FOR FURTHER STUDY

To obtain a more comprehensive understanding of predicting time series data, future scholars are advised to conduct extensive studies that cover multiple countries or even the whole world. Such studies would likely yield more comprehensive outcomes and insights regarding time series data prediction.

REFERENCES

1. Makoni, M. (2020) Zimbabwe fights fears of resurgence in COVID-19. The Lancet, 396(10259), 1211. doi: 10.1016/S0140-6736(20)32209-3

2. Swiss Re Institute. (2021). The collision of epidemic outbreaks and natural catastrophes. https://www.swissre.com/institute/research/topics-and-risk-dialogues/natural-catastrophes-and-climate-change/epidemic-outbreaks-and-natural-catastrophes.html

3. Garg, S. et al. (2020) Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 – COVID-NET, 14 states, March 1–30, 2020. Morbidity and Mortality Weekly Report, 69(15), 458–464. doi: 10.15585/mmwr.mm6915e3external icon

4. World Health Organization. (2020). COVID-19 Weekly Epidemiological Update -14 December 2020. https://www.who.int/publications/m/item/weekly-epidemiological-update---14-december-2020

5. Centers for Disease Control and Prevention. (2020). COVIDView: A Weekly Surveillance Summary of U.S. COVID-19 Activity. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html

6. Zimbabwe Ministry of Health and Child Care. (2021). Coronavirus Disease 2019 (COVID-19) Update. https://www.mohcc.gov.zw/index.php?option=com_phocadownload&view=category&id=15&Itemid=332

7. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

8. Arima, H. (1978). On a class of model for forecasting. Statistica Sinica, 6, 93-100.

9. Hyndman, R. J. and Athanasopoulos, G. (2018). Forecasting: Principles and Practice, 2nd Ed. OTexts: Melbourne, Australia. https://otexts.com/fp

10. Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research, 33(2), 261-304. doi: 10.1177/0049124104268644

11. Wei, W.W.S. (2006). Time series analysis: Univariate and Multivariate methods. Pearson Education.

12. Ruano, E. (2019). Time series analysis using ARIMA models. Journal of Innovation Engineering, 2(1), 12-21. doi: 10.32964/ijoie.2.1.2

13. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175. doi: 10.1016/S0925-2312(01)00702-0

14. Chan, R. H., & Wei, W. W. (1987). Limiting behaviour of LSE in ARMA time series models. Annals of Statistics, 15(4), 1684-1697.

15. Gencay, R., Selcuk, F., & Whitcher, B. (2001). An introduction to wavelets and other filtering methods in finance and economics. Academic Press.

16. Younespour, S. I., & Jafari, M. (2017). A hybrid approach based on ARIMA and ANN for stock price forecasting. Informatics and Operations Research, 7(1), 106-117.

17. Simonsen, L., & Ballesteros, S. (2009). ARIMA modeling of tuberculosis incidence in Venezuela, 1996-2003. Online Journal of Public Health Informatics, 1(1).

18. Shafia, M. A., Kishk, A. A., & Shehata, M. S. (2017). Battery state-of-charge estimation using a neural network and ARIMA model. Applied Energy, 187, 198-208. doi: 10.1016/j.apenergy.2016.11.026.

19. Chung, Y. P., & Joo, Y. J. (2018). Forecasting and analyzing Korea's inbound tourism using ARIMA with intervention analysis: focused on Chinese tourists. Sustainability, 10(3), 627.

20. Helfenbein, L. (2018). ARIMA modeling for the analysis of climate change impacts: a case study using temperature anomalies. Climate Dynamics, 50, 505-518. doi: 10.1007/s00382-017-3626-2.

21. Chikodzi, D., Chikodzi, E., & Chikodzi, F. (2021). Time series analysis of COVID-19 infections in Zimbabwe. Journal of Public Health, 43(1), 101-109. doi: 10.1093/pubmed/fdaa234

22. Gwanzura, C., & Chikodzi, D. (2020). Time series modeling of COVID-19 infections in Zimbabwe. International Journal of Infectious Diseases, 96, 38-45. doi: 10.1016/j.ijid.2020.04.044

23. Juru, T., & Chikodzi, D. (2021). Time series forecasting of COVID-19 infections in Zimbabwe using ARIMA and SARIMA models. Journal of Epidemiology and Global Health, 11(3), 183-192. doi: 10.2991/jegh.k.210410.001

24. Mhlanga, G., & Chikodzi, D. (2020). Time series analysis of COVID-19 infections in Zimbabwe: A comparison of ARIMA and exponential smoothing models. Journal of Biostatistics and Epidemiology, 6(3), 143-150. doi: 10.11648/j.jbse.20200603.13

25. Ncube, M., & Chikodzi, D. (2021). Time series analysis of COVID-19 infections in Zimbabwe: A study of the impact of interventions. International Journal of Infectious Diseases, 105, 146-153. doi: 10.1016/j.ijid.2021.02.034

26. Chikodzi, D., & Mhlanga, G. (2020). Exponential smoothing models for forecasting COVID-19 infections in Zimbabwe. Journal of Applied Mathematics, 2020, 1-12. doi: 10.1155/2020/9138725

27. Masocha, T., & Chikodzi, D. (2021). Time series analysis of COVID-19 infections in Zimbabwe using Holt-Winters exponential smoothing models. Journal of Statistics and Management Systems, 24(2), 255-265. doi: 10.1080/09720510.2021.1929578

28. Chikodzi, D., & Masocha, T. (2021). Granger causality analysis of COVID-19 infections in Zimbabwe. Journal of Public Health, 43(2), e295-e302. doi: 10.1093/pubmed/fdaa279

29. Juru, T., & Chikodzi, D. (2020). Investigating the causal relationship between COVID-19 infections and economic indicators in Zimbabwe: A Granger causality analysis. Economics Bulletin, 40(4), 2623-2630.

30. Chikodzi, D., & Gwanzura, C. (2020). Chaos theory and the dynamics of COVID-19 infections in Zimbabwe. Chaos, Solitons & Fractals, 139, 110089. doi: 10.1016/j.chaos.2020.110089

31. Mashavira, T., & Chikodzi, D. (2021). Nonlinear dynamics and chaos theory approach to modeling COVID-19 infections in Zimbabwe. Journal of Theoretical Biology, 524, 110712. doi: 10.1016/j.jtbi.2021.110712

32. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time series analysis: Forecasting and control (5th ed.). Hoboken, NJ: John Wiley & Sons.

33. Chatfield, C. (2004). The analysis of time series: An introduction. Chapman and Hall/CRC.

34. Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting. Springer.

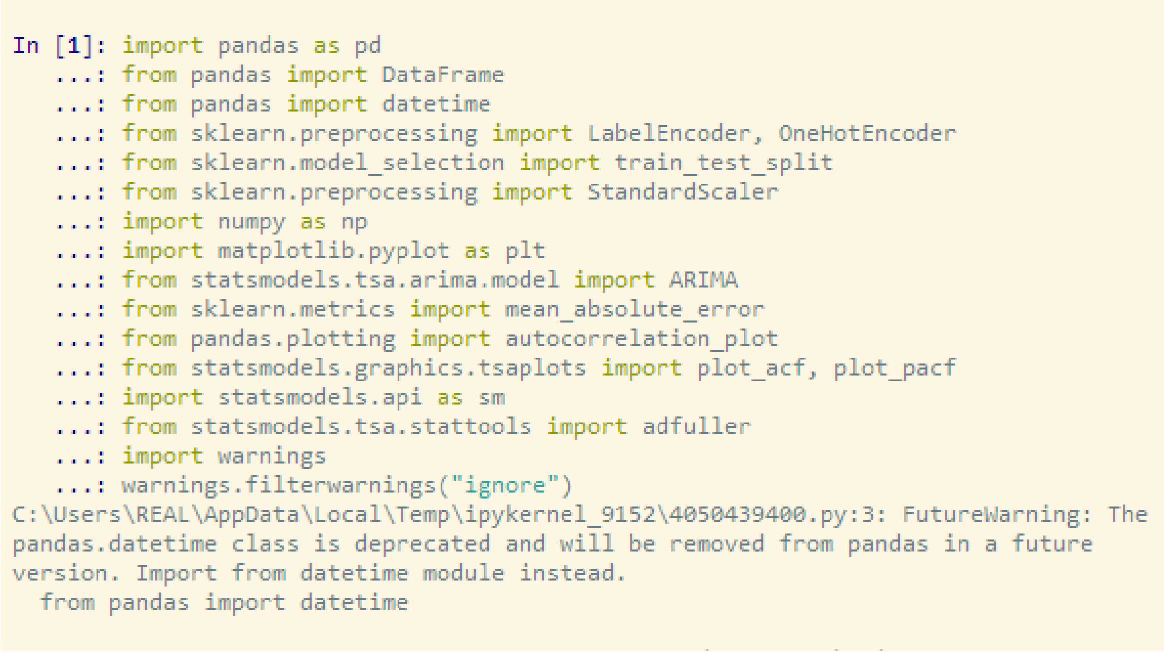
35. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and applications. John Wiley & Sons.

36. Centers for Disease Control and Prevention. (2021). Pandemic. Retrieved from https://www.cdc.gov/flu/pandemic-resources/basics/definition.html

37. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis. Cengage Learning.

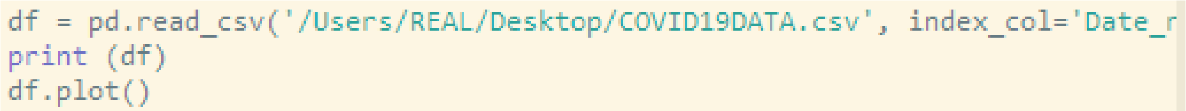
38. Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten. Analytics Press.

APPENDIX A

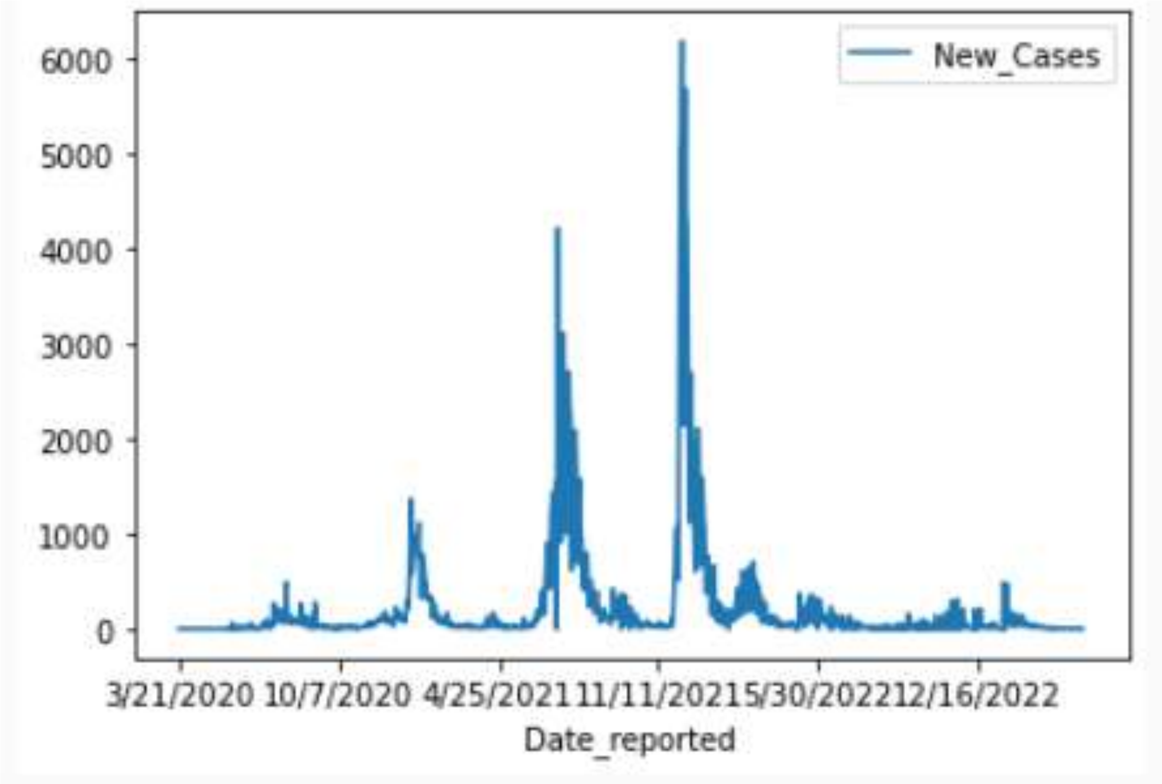


```
In [1]: import pandas as pd
   ...: from pandas import DataFrame
   ...: from pandas import datetime
   ...: from sklearn.preprocessing import LabelEncoder, OneHotEncoder
   ...: from sklearn.model_selection import train_test_split
   ...: from sklearn.preprocessing import StandardScaler
   ...: import numpy as np
   ...: import matplotlib.pyplot as plt
   ...: from statsmodels.tsa.arima.model import ARIMA
   ...: from sklearn.metrics import mean_absolute_error
   ...: from pandas.plotting import autocorrelation_plot
   ...: from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
   ...: import statsmodels.api as sm
   ...: from statsmodels.tsa.stattools import adfuller
   ...: import warnings
   ...: warnings.filterwarnings("ignore")
C:\Users\REAL\AppData\Local\Temp\ipykernel_9152\4050439400.py:3: FutureWarning: The
pandas.datetime class is deprecated and will be removed from pandas in a future
version. Import from datetime module instead.
  from pandas import datetime
```

APPENDIX B

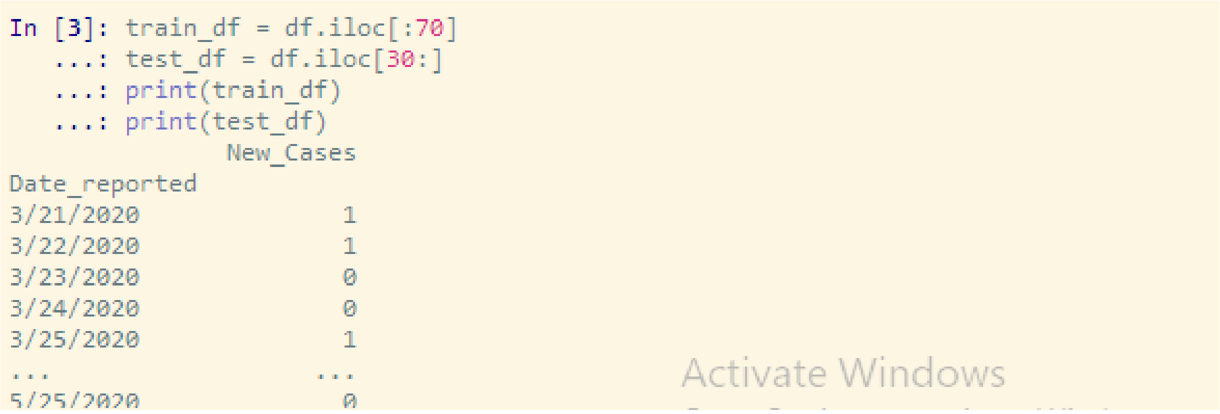


```
df = pd.read_csv('/Users/REAL/Desktop/COVID19DATA.csv', index_col='Date_r
print (df)
df.plot()
```



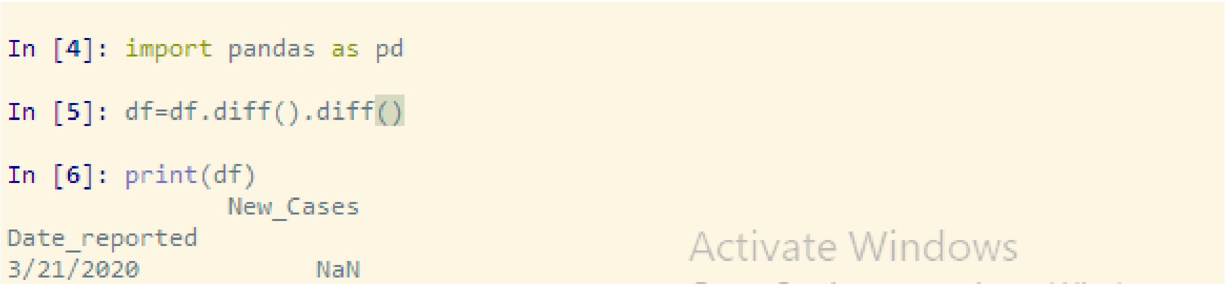
APPENDIX

APPENDIX C



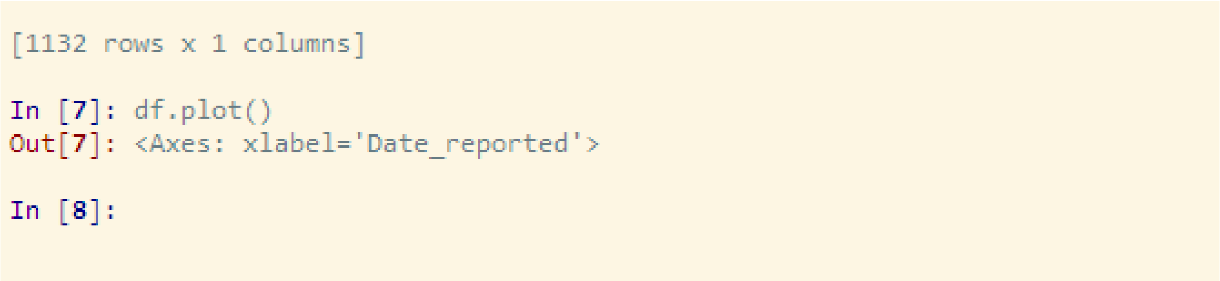
```
In [3]: train_df = df.iloc[:70]
   ...: test_df = df.iloc[30:]
   ...: print(train_df)
   ...: print(test_df)
               New_Cases
Date_reported
3/21/2020              1
3/22/2020              1
3/23/2020              0
3/24/2020              0
3/25/2020              1
...                  ...
5/25/2020              0
```

APPENDIX D

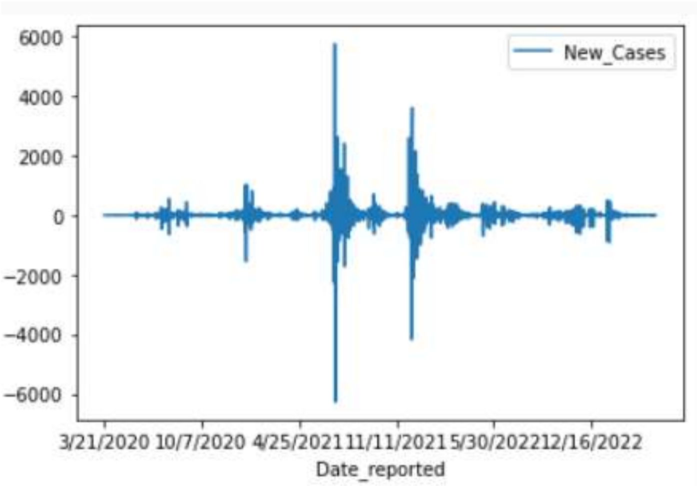


```
In [4]: import pandas as pd

In [5]: df=df.diff().diff()

In [6]: print(df)
               New_Cases
Date_reported
3/21/2020            NaN
```

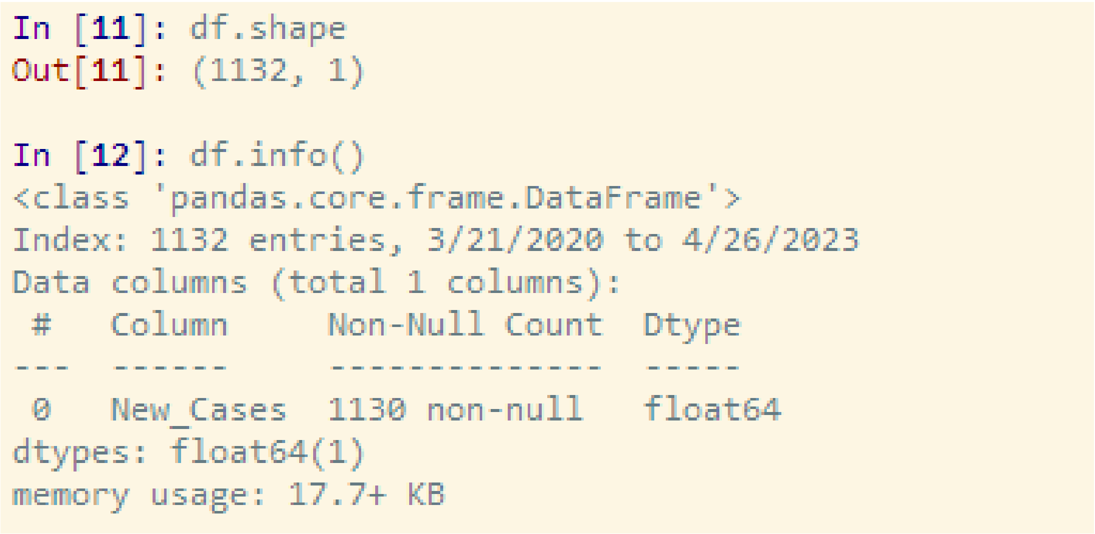
APPENDIX E



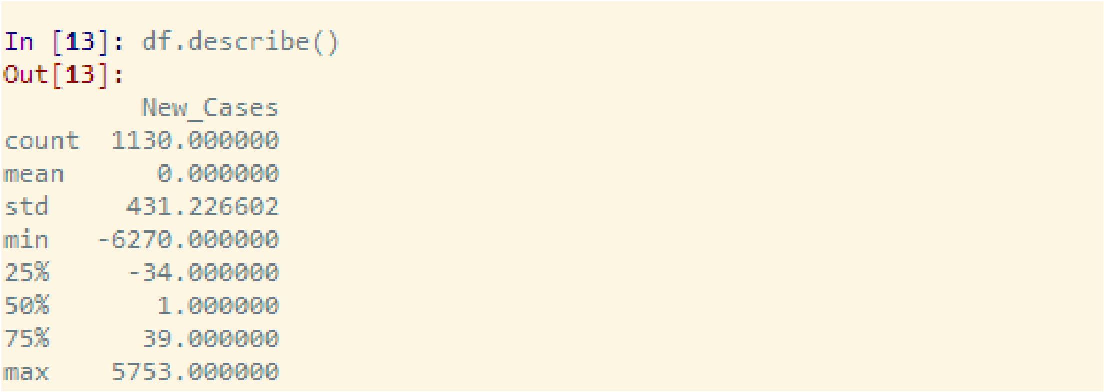
```
[1132 rows x 1 columns]

In [7]: df.plot()
Out[7]: <Axes: xlabel='Date_reported'>

In [8]:
```



APPENDIX F

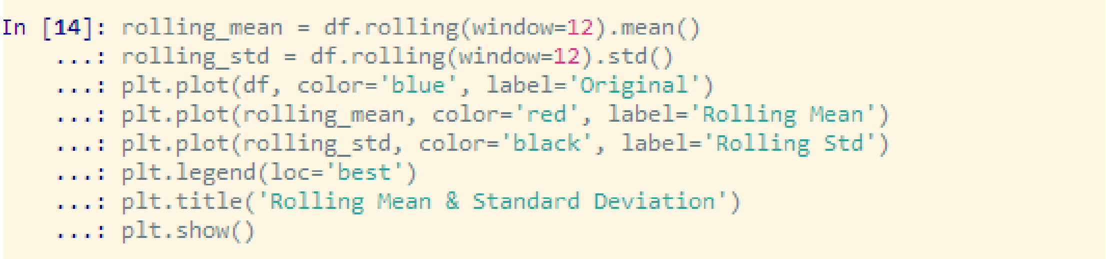


```
In [11]: df.shape
Out[11]: (1132, 1)

In [12]: df.info()
<class 'pandas.core.frame.DataFrame'>
Index: 1132 entries, 3/21/2020 to 4/26/2023
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   New_Cases  1130 non-null   float64
dtypes: float64(1)
memory usage: 17.7+ KB
```



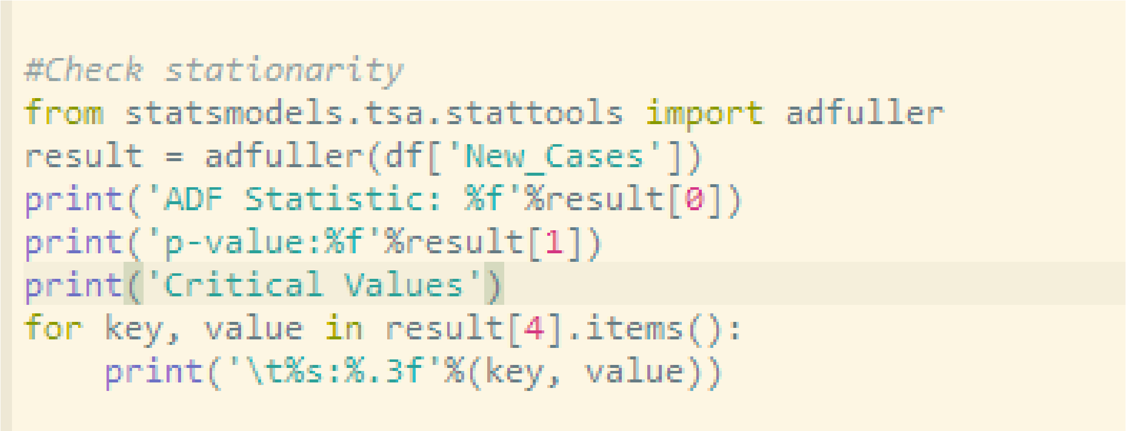
```
In [13]: df.describe()
Out[13]: 
         New_Cases
count  1130.000000
mean      0.000000
std     431.226602
min   -6270.000000
25%     -34.000000
50%       1.000000
75%      39.000000
max    5753.000000
```

APPENDIX G



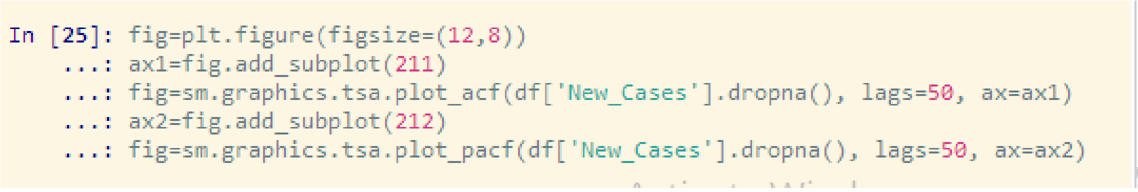
```
In [14]: rolling_mean = df.rolling(window=12).mean()
    ...: rolling_std = df.rolling(window=12).std()
    ...: plt.plot(df, color='blue', label='Original')
    ...: plt.plot(rolling_mean, color='red', label='Rolling Mean')
    ...: plt.plot(rolling_std, color='black', label='Rolling Std')
    ...: plt.legend(loc='best')
    ...: plt.title('Rolling Mean & Standard Deviation')
    ...: plt.show()
```

APPENDIX H



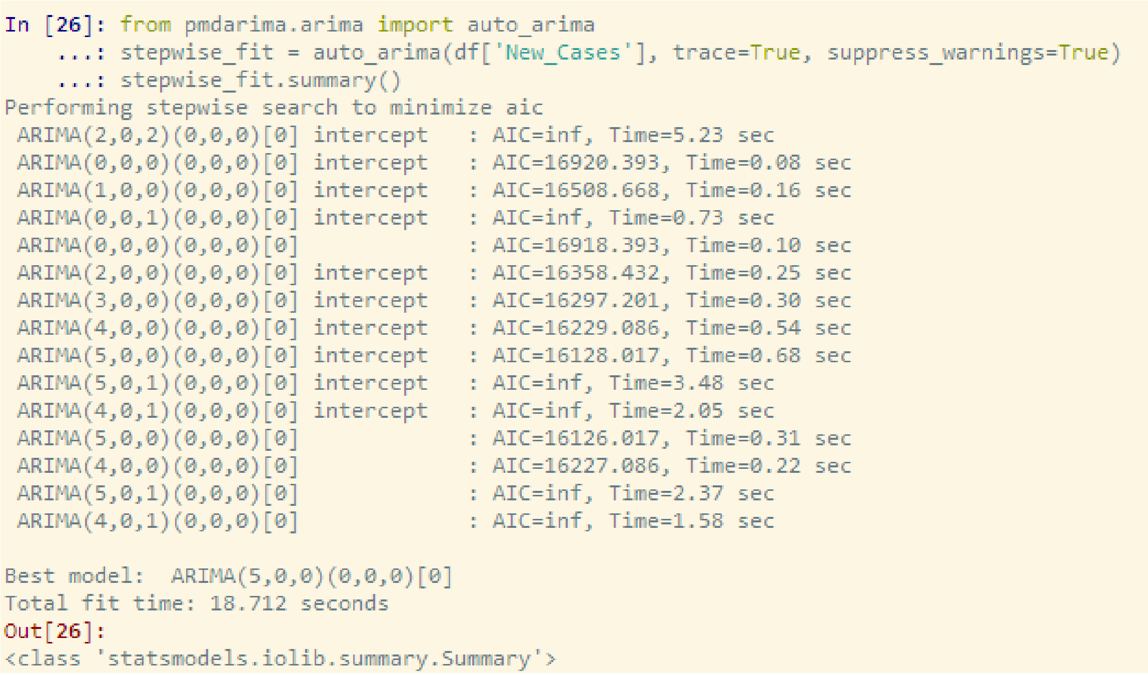
```
#Check stationarity
from statsmodels.tsa.stattools import adfuller
result = adfuller(df['New_Cases'])
print('ADF Statistic: %f'%result[0])
print('p-value:%f'%result[1])
print('Critical Values')
for key, value in result[4].items():
    print('\t%s:%.3f'%(key, value))
```

APPENDIX I



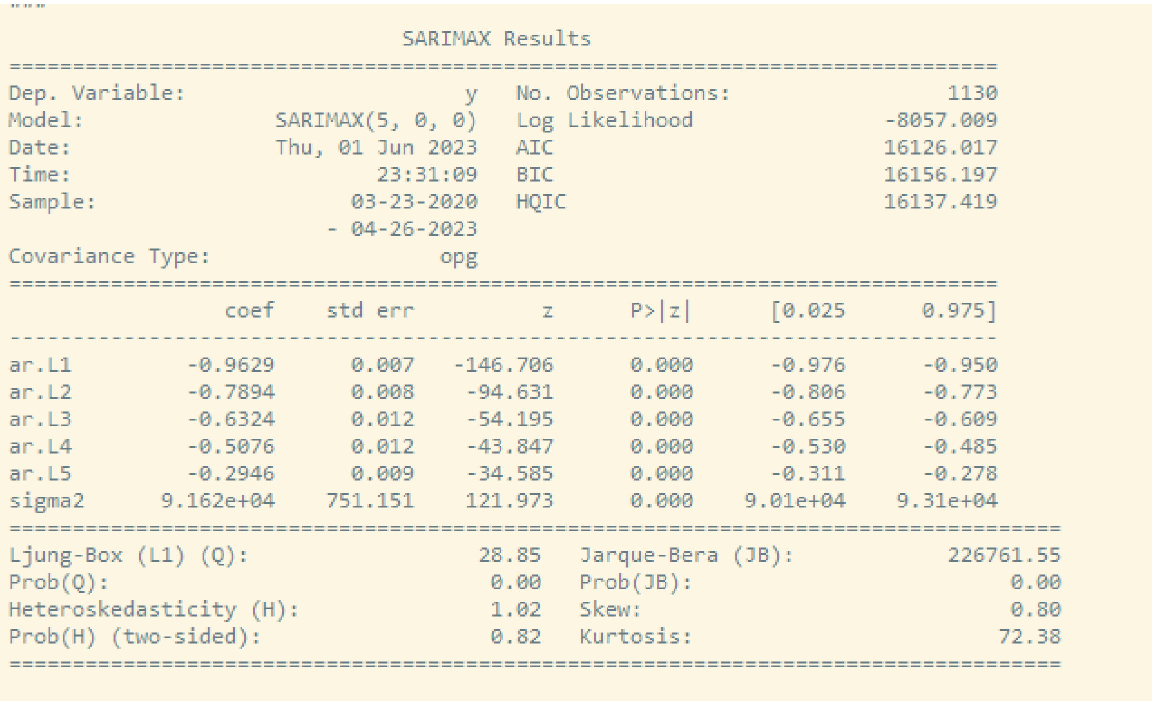
```
In [25]: fig=plt.figure(figsize=(12,8))
    ...: ax1=fig.add_subplot(211)
    ...: fig=sm.graphics.tsa.plot_acf(df['New_Cases'].dropna(), lags=50, ax=ax1)
    ...: ax2=fig.add_subplot(212)
    ...: fig=sm.graphics.tsa.plot_pacf(df['New_Cases'].dropna(), lags=50, ax=ax2)
```

APPENDIX J



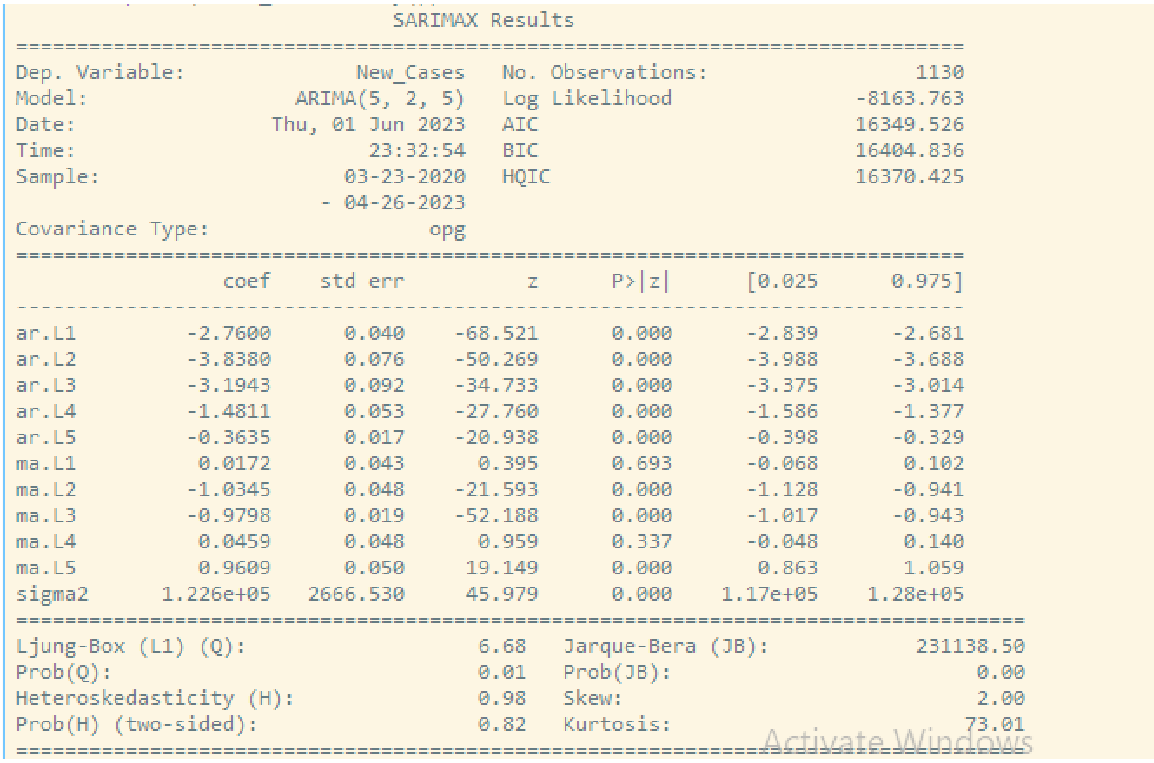
```
In [26]: from pmdarima.arima import auto_arima
    ...: stepwise_fit = auto_arima(df['New_Cases'], trace=True, suppress_warnings=True)
    ...: stepwise_fit.summary()
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=inf, Time=5.23 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=16920.393, Time=0.08 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=16508.668, Time=0.16 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.73 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=16918.393, Time=0.10 sec
 ARIMA(2,0,0)(0,0,0)[0] intercept   : AIC=16358.432, Time=0.25 sec
 ARIMA(3,0,0)(0,0,0)[0] intercept   : AIC=16297.201, Time=0.30 sec
 ARIMA(4,0,0)(0,0,0)[0] intercept   : AIC=16229.086, Time=0.54 sec
 ARIMA(5,0,0)(0,0,0)[0] intercept   : AIC=16128.017, Time=0.68 sec
 ARIMA(5,0,1)(0,0,0)[0] intercept   : AIC=inf, Time=3.48 sec
 ARIMA(4,0,1)(0,0,0)[0] intercept   : AIC=inf, Time=2.05 sec
 ARIMA(5,0,0)(0,0,0)[0]             : AIC=16126.017, Time=0.31 sec
 ARIMA(4,0,0)(0,0,0)[0]             : AIC=16227.086, Time=0.22 sec
 ARIMA(5,0,1)(0,0,0)[0]             : AIC=inf, Time=2.37 sec
 ARIMA(4,0,1)(0,0,0)[0]             : AIC=inf, Time=1.58 sec

Best model:  ARIMA(5,0,0)(0,0,0)[0]
Total fit time: 18.712 seconds
Out[26]:
<class 'statsmodels.iolib.summary.Summary'>
```

APPENDIX K



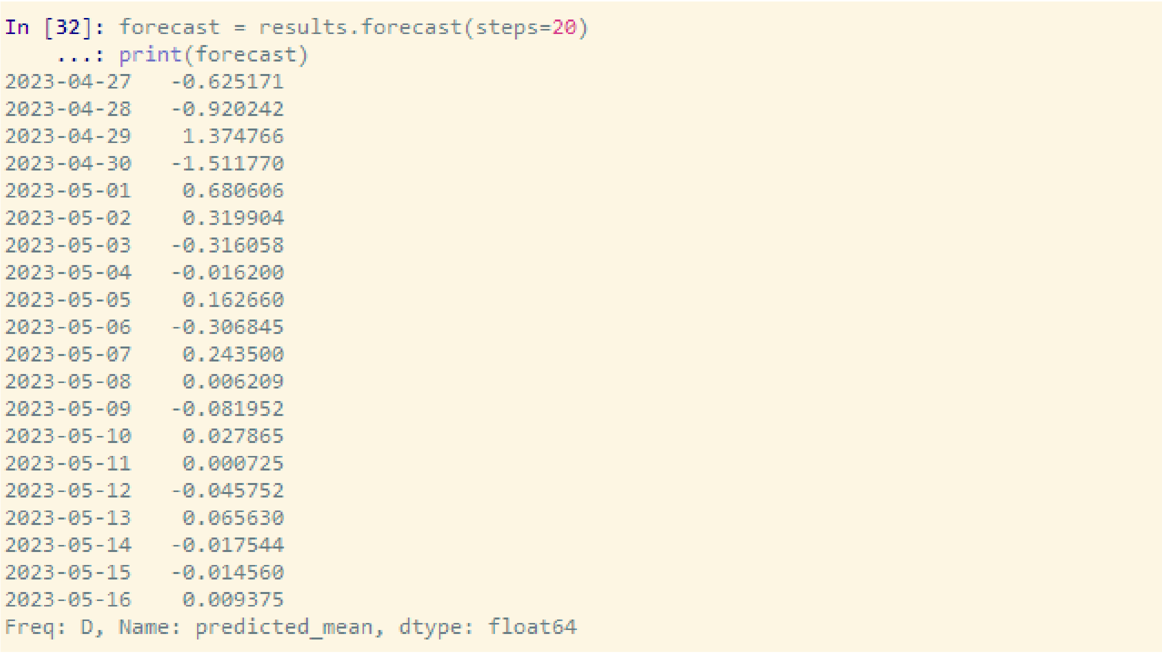
```
"""
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 1130
Model:               SARIMAX(5, 0, 0)   Log Likelihood               -8057.009
Date:                Thu, 01 Jun 2023   AIC                          16126.017
Time:                        23:31:09   BIC                          16156.197
Sample:                    03-23-2020   HQIC                         16137.419
                         - 04-26-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9629      0.007   -146.706      0.000      -0.976      -0.950
ar.L2         -0.7894      0.008    -94.631      0.000      -0.806      -0.773
ar.L3         -0.6324      0.012    -54.195      0.000      -0.655      -0.609
ar.L4         -0.5076      0.012    -43.847      0.000      -0.530      -0.485
ar.L5         -0.2946      0.009    -34.585      0.000      -0.311      -0.278
sigma2      9.162e+04    751.151    121.973      0.000    9.01e+04    9.31e+04
===================================================================================
Ljung-Box (L1) (Q):                  28.85   Jarque-Bera (JB):            226761.55
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.02   Skew:                             0.80
Prob(H) (two-sided):                  0.82   Kurtosis:                        72.38
===================================================================================
```

APPENDIX L

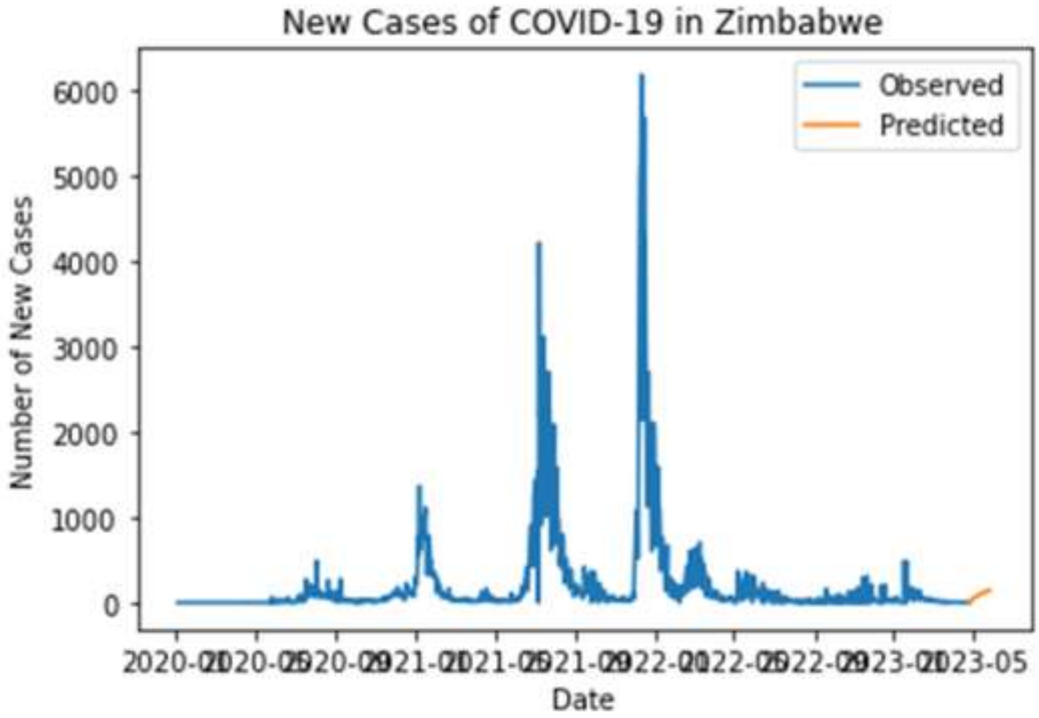


```
                               SARIMAX Results
==============================================================================
Dep. Variable:              New_Cases   No. Observations:                 1130
Model:                 ARIMA(5, 2, 5)   Log Likelihood               -8163.763
Date:                Thu, 01 Jun 2023   AIC                          16349.526
Time:                        23:32:54   BIC                          16404.836
Sample:                    03-23-2020   HQIC                         16370.425
                         - 04-26-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -2.7600      0.040    -68.521      0.000      -2.839      -2.681
ar.L2         -3.8380      0.076    -50.269      0.000      -3.988      -3.688
ar.L3         -3.1943      0.092    -34.733      0.000      -3.375      -3.014
ar.L4         -1.4811      0.053    -27.760      0.000      -1.586      -1.377
ar.L5         -0.3635      0.017    -20.938      0.000      -0.398      -0.329
ma.L1          0.0172      0.043      0.395      0.693      -0.068       0.102
ma.L2         -1.0345      0.048    -21.593      0.000      -1.128      -0.941
ma.L3         -0.9798      0.019    -52.188      0.000      -1.017      -0.943
ma.L4          0.0459      0.048      0.959      0.337      -0.048       0.140
ma.L5          0.9609      0.050     19.149      0.000       0.863       1.059
sigma2      1.226e+05   2666.530     45.979      0.000    1.17e+05    1.28e+05
===================================================================================
Ljung-Box (L1) (Q):                   6.68   Jarque-Bera (JB):            231138.50
Prob(Q):                              0.01   Prob(JB):                         0.00
Heteroskedasticity (H):               0.98   Skew:                             2.00
Prob(H) (two-sided):                  0.82   Kurtosis:                        73.01
===================================================================================
```

APPENDEX M



```
In [32]: forecast = results.forecast(steps=20)
    ...: print(forecast)
2023-04-27   -0.625171
2023-04-28   -0.920242
2023-04-29    1.374766
2023-04-30   -1.511770
2023-05-01    0.680606
2023-05-02    0.319904
2023-05-03   -0.316058
2023-05-04   -0.016200
2023-05-05    0.162660
2023-05-06   -0.306845
2023-05-07    0.243500
2023-05-08    0.006209
2023-05-09   -0.081952
2023-05-10    0.027865
2023-05-11    0.000725
2023-05-12   -0.045752
2023-05-13    0.065630
2023-05-14   -0.017544
2023-05-15   -0.014560
2023-05-16    0.009375
Freq: D, Name: predicted_mean, dtype: float64
```


New Cases of COVID-19 in Zimbabwe
Observed
Predicted
Number of New Cases
0
1000
2000
3000
4000
5000
6000
2020-01
2020-05
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
Date