Project Report

On

A STATISTICAL STUDY ON ROAD ACCIDENTS IN ERNAKULAM

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

AISWARYA BINU

(Reg No. SM22AS001)

(2022-2024)

Under the Supervision of

VISMAYA VINCENT

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI – 682011

APRIL 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **A STATISTICAL STUDY ON ROAD ACCIDENTS IN ERNAKULAM** is a bonafide record of the work done by Ms. **AISWARYA BINU** under my guidance as partial fulfilment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 29/4/2024

Place : Ernakulam

Vismaya Vincent
Assistant Professor,
Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

Nisha Oommen
Assistant Professor & HOD
Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

External Examiners

1. ..... Chinu Joseph.

2. ..... LAKSHMI SURESH

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **VISMAYA VINCENT**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam

Date: 29/4/2024

AISWARYA BINU

SM22AS001

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout this study helped me to complete the work up to this stage.

I am very grateful to my project guide Vismaya Vincent for the immense help during the period of work.

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to the faculty, teaching, and non-teaching staff of the department and Colleagues.

I am also very thankful to HoD Mrs. Nisha Oommen for their valuable suggestions, critical examination of work during the progress.

Ernakulam

Date: 29/4/2024

AISWARYA BINU

SM22AS001

ABSTRACT

Road accidents are a major issue in Ernakulam, Kerala, causing loss of life and property. In this project, the focus was on forecasting road accident trends using statistical models like SARIMA (Seasonal Autoregressive Integrated Moving Average) and linear regression. The study included collecting historical data on road accidents in Ernakulam over a period of time. This data was analysed to identify patterns, trends and seasonal variations. The SARIMA model was used to capture the complex time series nature of road accident data, considering factors like seasonality and auto correlation. Linear regression was used to understand the relationship between the variables. By combining SARIMA and linear regression techniques, accurate forecasts were evolved for future road accidents. Over all these forecasts shows that it's important for government authorities, police, transportation authorities and the public to take action and make roads safer for everyone.

ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	AISWARYA BINU
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. Vismaya Vincent
Department	Post Graduate Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	A STATISTICAL STUDY ON ROAD ACCIDENTS IN ERNAKULAM
Similarity	0% AI 7%
Paper ID	1659038
Submission Date	2024-04-18 12:46:32

Signature of Student

Signature of Guide

Checked By
College Librarian

ST.TERESA'S COLLEGE LIBRARY ERNAKULAM

* This report has been generated by DrillBit Anti-Plagiarism Software

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

The study titled 'A Statistical Study on Road Accidents in Ernakulam' was an effort made to find out the number of accidents occurring in Ernakulam and to forecast the number of accidents in future. This project helps to understand the past and current situations in Ernakulam. Data used in this study was collected from the Police Department in Ernakulam.

Road accidents pose a significant challenge to public safety. Understanding the key patterns of road accidents was crucial for implementing effective prevent measures. This study aims to analyse the data on road accidents, identifying key trends and predict the future using advanced analytical techniques.

There are many factors contribute to road accidents. The main reasons for careless driving behaviour including over speeding, lane violations, type of roads etc. Also using mobile phones during driving causes major reasons for road accidents. Observing the types of vehicles that are involved in accidents is very important. There are two main vehicles involved in most of the accidents. They are two wheelers and cars. Lots of accidents happen with these vehicles. People who are travelling in two wheelers have more risk because they can get hurt easily. This leads to an increase in the number of death and injuries. Accidents with buses and trucks are also common. All of these together add to the overall number of accidents that happen. Other reasons of road accidents are:

1. Due to weather
2. Bad road
3. Malfunction of vehicle
4. Driving while drinking alcohol

Analysing the number of accidents shows major issues that need quick action. The accidents occur a lot for many different reasons and affect different kinds of people.

Understanding these patterns will help the government to create effective plans and reduce the number of road accidents and improve the lives of people.

OBJECTIVES:

The main objectives of the study is as follows:

1. To model and forecast road accidents in Ernakulam using Seasonal ARIMA model.
2. To model and forecast road accidents in Ernakulam using Linear Regression model.
3. To compare the forecast by Seasonal ARIMA model and Linear Regression model.

CHAPTER -2

LITERATURE REVIEW

This chapter shows the results from related research that analysed various road accident datasets and made predictions using a variety of statistical methods, different data mining techniques, machine learning algorithms etc.

Sikdar et al. (2017) focused on the importance of studying factors that affecting the road accidents to reduce their rate. They used the Chi Square method to analyse the data from different regions of India and found out the factors that contribute to accidents. They also investigate the relationship between different types of location, year-wise data, impact of time factors, weather etc. The study recommended the preventive measures to reduce the road accidents.

Sunny et al. (2018) investigated the road accidents in Kerala, point out the importance of data analysis in accident prevention. They used time series analysis to predict the road accidents in different places of Kerala. They studied the road accidents data from January 1999 to December 2016 to create a model that could predict the future patterns and they created predictive models like Holt- Winters exponential smoothing and Seasonal ARIMA, providing forecast values within the confidence interval.

Dutta et al. (2020) created an ARIMA model to analyse the road accidents in India from 1967 to 2015. They use secondary data collected from the National Crime Record Bureau. They found a considerable increase in the annual number of deaths during this period and they conclude that the ARIMA model was suitable for the dataset. The model's forecast for the next 10 years also indicated a continuing upward trend in the number of deaths caused by road accidents in India.

Yassin and Pooja (2020), focused on understanding the key factors affecting road accident severity. They employed a hybrid approach combining K means and random forest techniques to identify the most important variables. K-means helped uncover hidden information in the data and create a new feature in training set, while random forest predicted the severity of accident. Comparing the results with other techniques, the proposed method was very accurate, with a 99.86% success rate. Their findings indicated that driver experience, age, vehicle service year and day-light conditions were the important contributing factors for the injuries.

Rabbani et al. (2021) conducted a study to address the inadequacies of road safety guidelines and their failure to handle the increasing traffic volume and reduce accidents. Their aim was to provide valuable insights and predictions regarding accident rates in Pakistan. They utilized univariate time series analysis, specifically SARIMA and ES models, to create temporal patterns and forecast accident rates. The results showed that the ES model outperformed the SARIMA model in terms of various error metrics. The study's findings offer guiding principles for implementing forecasted accident rates in road design, prioritizing the safety of end users. These insights are beneficial for accident rate collection agencies, decision-makers, design consultants, and accident prevention departments.

Vipin & Rahul (2021) conducted a study on road traffic accidents (RTAs) and their mortality based on the time of occurrence in Kerala State, India. The study analyzed the data from 2005 to 2018 and found that RTA deaths showed a seasonal pattern and an overall upward trend during this period. The highest number of deaths occurred between 18:00 p.m. and 21:00 p.m., while the lowest number occurred between 00:00 a.m. and 03:00 a.m. Evening and night-time had relatively higher RTA death rates. Deaths before dawn (06:00 a.m. to 09:00 a.m.) and after dusk (15:00 p.m. to 18:00 p.m.) were also relatively high. The analysis revealed an increasing trend in RTA deaths during specific time zones since 2016, with the period between 21:00 p.m. and 24:00 p.m. showing a steady increase. Time series regression analysis indicated that for every 3-hour increase in time over the year, approximately four persons were killed due to RTA. The forecast for 2019-2021 suggested that most RTA

deaths would likely occur between 18:00 p.m. and 21:00 p.m., with the least number of deaths between 00:00 a.m. and 03:00 a.m. The forecast for 2020 showed a decrease in RTA deaths for every time zone compared to 2019, followed by an increase in 2021.

Babu and Sulaipher (2022) examined the main reasons for road accidents in Kerala. They observed the lifestyle and behaviours to understand the problems. By studying the previous research and the key factors such as driver behaviour, drunken driving they aimed to create a respondent profile and causes of accidents. They used percentage analysis and factor analysis with sample size 110. Their findings will help in identifying and deal with causes of road accidents in Kerala to lead a better traffic control and accident rate reduction.

Bhat et al. (2022), the researchers analysed accident-prone areas in Kerala, a state in southwest India. They employed data mining techniques to identify the key factors responsible for accidents. The researchers initially used cluster analysis to assess accident severity between 2007 and 2017, followed by principal component analysis to reduce the correlation data and identify major influencing factors. The cluster analysis revealed that South Kerala and Central Kerala were more affected than North Kerala in terms of accidents resulting in deaths and injuries. Principal component analysis was performed on the entire dataset due to sample adequacy violations in each cluster. The results highlighted that high vehicle speeds significantly influenced highway crashes. These multivariate techniques proved valuable in categorizing areas with high and low accident intensity, predicting contributing factors, and guiding efforts to reduce the impact of accidents on public health and improve administrative measures.

Deretic et al. (2022) studied traffic accidents to make traffic systems safer. Traffic accidents cause a lot of harm worldwide but thanks to efforts like the Decade of Action for Road Safety 2011–2020, many countries have seen a decrease in

accidents. This reduction is due to better vehicles and roads, driver education, and improved medical care. The study focused on analysing traffic accidents in Belgrade using data analysis and a prediction model. They found a seasonal pattern in accidents and their model predicted future accidents with about 5.22% error. Predicting accidents can help plan safety campaigns and strategies for safer roads.

Parthiban et al. (2022) conducted an analysis on road accidents in Kerala using a machine learning algorithm for prediction. Identifying the reasons behind accidents has been a major challenge, with existing methods lacking effectiveness and expertise. Therefore, the application of computer machine learning algorithms in road accident research is timely. By implementing algorithms and their technology, it becomes easier to predict and analyse the causes of frequent accidents in specific locations. In this study, Python programming and machine language patterns were utilized to create accurate prediction models and form clusters of accident data. These clusters were then used to derive cost-effective measures for preventing road accidents.

CHAPTER -3

MATERIALS AND METHODOLOGY

3.1 DATA DESCRIPTION

The dataset contains the number of road accidents in Ernakulam from the year 2018 to 2023. It is monthly data from the period of January 2018 to November 2023. Data used in the study was collected from the Police Department in Ernakulam.

3.2 METHODOLOGY

The initial and crucial step in this analysis was study the data in detail. The main purpose of the study was to forecast the future number of accidents in Ernakulam using time series model and linear regression model. First, the model was forecasted using time series technique. Finally, the model was forecasted using a linear regression model, and the MSE and RMSE values of both were calculated.

TOOLS FOR ANALYSIS & FORECASTING

1) Seasonal ARIMA (Autoregressive Integrated Moving Average) model
2) Linear Regression model

TOOL FOR COMPARISON

1) Mean Squared Error (MSE)
2) Root Mean Squared Error (RMSE)

3.3 PYTHON SOFTWARE

Python is a programming language that used to create websites and analyse the data. in this study python is used to forecast the number of road accidents using the time series model SARIMA and linear regression model. SARIMA is consistent for catching seasonal patterns and trends in a time series data. linear regression helps in understanding the relationship between variables.

CHAPTER –4

TIME SERIES

4.1 TIME SERIES ANALYSIS

Time series is the arrangement of statistical data according to the occurrence of time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future. If a random variable X is indexed to time, usually denoted by t, the observations $\{X_t, t \in T\}$ is called a time series, where T is a time index set. A time series is said to be continuous when the observations are made continuously on time.

Time series data is said to be discrete when the observations are taken only at specific times, usually equally spaced. Time series components refer to the different underlying structures that can be observed within a time series. These components help in understanding and modelling the various sources of variation present in the data. The main components of a time series are:

Trend: The long-term movement or direction in the data. It represents the overall growth or decline in the series over an extended period. Trends can be linear, nonlinear, or even absent.

Seasonality: The repeating, periodic patterns in the data that occur at fixed intervals. For instance, sales of winter clothing typically exhibit a seasonal pattern, with higher sales during colder months. Seasonality is usually influenced by factors like time of year, holidays, or weather.

Cyclic Variation: Similar to seasonality, but it doesn't have a fixed period. Cycles are longer term patterns that repeat at irregular intervals and are usually influenced by economic, political, or societal factors. Unlike seasonality, cycles are harder to predict accurately.

Irregular/Residual Component: This represents the random or erratic fluctuations that can't be explained by the trend, seasonality, or cyclic components. These irregular variations can result from unforeseen events, measurement errors, or other unpredictable factors.

The main objective of time series analysis is forecasting. In forecasting, we estimate the future values of the series. A given time series can be

1) Stationary
2) Non-stationary

4.2 STATIONARY TIME SERIES

A stationary time series exhibits consistent statistical properties throughout its entire duration, including constant mean, variance, and autocorrelation structure. This implies that the underlying processes generating the data remain stable over time, making it easier to model and predict future values. Stationary time series often enable the use of techniques like autoregressive integrated moving average (ARIMA) models, which assume a stable environment.

4.3 NON- STATIONARY TIME SERIES

Non-stationary time series display changing statistical properties over time, such as varying means, trends, and irregular fluctuations. These changes can arise from factors like seasonality, external influences, or inherent instability in the data-generating process. Non-stationary data can pose challenges for modelling and prediction because traditional methods may not accurately capture evolving patterns. To address non-stationarity, differencing is commonly used to transform the data into a stationary form. Differencing involves subtracting each data point from its lagged counterpart to remove trends or seasonality. Once differenced, the transformed data can often be modelled effectively using stationary techniques.

4.4 AUTOCORRELATION FUNCTION (ACF)

The Autocorrelation Function (ACF) is a graphical tool used to assess the correlation between a time series and its lagged versions. In other words, it quantifies the similarity between the values of a time series at different time points. By plotting the ACF, one can identify whether there is a significant correlation between the current observation and observations at various lags. A sharp drop in autocorrelation after a certain lag suggests the presence of a seasonality component with that lag. An ACF that gradually declines indicates a potential AR component in the time series. The ACF is pivotal for identifying the order of the moving average (MA) term in a model like ARIMA.

4.5 PARTIAL AUTOCORRELATION FUNCTION (PACF)

The Partial Autocorrelation Function (PACF) extends the concept of autocorrelation by capturing the direct correlation between two observations while accounting for the indirect correlations introduced by the intermediate lags. It essentially measures the correlation between the current observation and observations at specific lags, after removing the effects of the intervening lags. By plotting the PACF, analysts can identify the lag values where the direct correlation is significant. A sharp cutoff in the PACF after a certain lag indicates a potential autoregressive (AR) component in the time series. Just like the ACF, the PACF assists in determining the appropriate parameters for models like ARIMA.

4.6 ARIMA (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE)

ARIMA, which stands for Auto Regressive Integrated Moving Average, is a powerful time series forecasting model used to capture the patterns, trends, and dependencies present in time series data. ARIMA combines three main components: autoregression (AR), differencing (I for integrated), and moving average (MA). It is designed to

work with stationary and weakly stationary data, which means that the mean, variance, and autocorrelation structure remain relatively constant over time.

Auto Regressive (AR) Component:- The AR component captures the linear relationship between the current observation and a specified number of lagged observations. It models the dependency of the current value on its own past values. The order of the AR component is denoted as "p," which indicates how many lagged observations are included in the model.

Differencing (I) Component:- The differencing component is used to make the data stationary by subtracting each observation from its lagged observation. It accounts for trends and seasonality that might be present in the data. The "d" in ARIMA denotes the order of differencing required to achieve stationarity.

Moving Average (MA) Component:- The MA component models the relationship between the current observation and a specified number of past errors or residuals. It helps capture the impact of past shocks on the current value. The order of the MA component is denoted as "q." The ARIMA model is often denoted as ARIMA(p, d, q), where "p", "d", and "q" are the orders of the AR, differencing, and MA components, respectively.

Steps to be followed for ARIMA modelling:

1. Exploratory analysis
2. Fit the model
3. Diagnostic measure

The general form of an Autoregressive process of order *p* is denoted by AR(p) which is expressed as:

$$Xt = \phi 1Xt\text{-}1 + \phi 2Xt\text{-}2 + \ldots\ldots + \phi pXt\text{-}p + \varepsilon t$$

Where,

- Xt is the value at time t
- $\phi 1, \phi 2, \ldots\ldots, \phi p$ are the finite set of weight parameters.
- εt is the white noise term or errors at time t

The general form of a Moving Average process of order q denoted as MA(q), can be expressed as:

$$Xt = \Theta 1\varepsilon t\text{-}1 + \Theta 2\varepsilon t\text{-}2 + \ldots.. + \Theta q\varepsilon t\text{-}q + \varepsilon t$$

Where,

- Xt is the value at time t.
- $\Theta 1, \Theta 2, \ldots..,\Theta q$ are the moving average coefficients.
- q is the order of the MA process.
- εt is the white noise term.
-

4.7 SARIMA (SEASONAL ARIMA)

SARIMA, or Seasonal ARIMA, extends the basic ARIMA model to handle time series data with both non-seasonal and seasonal patterns. This is especially useful for data that exhibit recurring patterns at fixed intervals, such as sales data affected by yearly seasonality. SARIMA includes additional components to capture the seasonal variations:

Seasonal AutoRegressive (SAR) Component:- This component models the relationship between the current observation and past observations at the same lag in previous seasons. It captures the effect of seasonality on the data. The order of the seasonal AR component is denoted as "P."

Seasonal Differencing (S) Component:- Similar to the non-seasonal differencing, the seasonal differencing component is used to remove the seasonality from the data. The order of seasonal differencing is denoted as "D."

Seasonal Moving Average (SMA) Component:- This component models the relationship between the current observation and past errors or residuals at the same lag in previous seasons. It accounts for past shocks that persist across seasons. The order of the seasonal MA component is denoted as "Q." SARIMA is denoted as SARIMA (p, d, q) (P, D, Q, s), where "s" represents the season's length

4.8 AUGMENTED DICKEY-FULLER TEST

Augmented Dickey-Fuller test (ADF test) is the most common statistical test used to whether a given time series is stationary or not. Stationarity is an important factor in time series. In time series forecasting the first step is to determine the number of differences needed to make the data stationary because a model cannot forecast on non-stationary time series data.

4.9 SEASONAL DATA

A time series data is said to be seasonal if the data shows repeating patterns or fluctuations at regular interval over time

4.10 AKAIKE INFORMATION CRITERION (AIC)

The AIC is a method for examining how well a model fits the data. In statistics, AIC is used to compare different models and determine which one is best fit for the data. AIC does not test hypothesis but estimates information loss when a model represents data. Its formula,

AIC = 2k-2ln(L), consider the parameters (k) and likelihood function (L). Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

4.11 MODEL FITTING

Model fitting is the process of estimating the model's parameters. The main goal of the model fitting process to identify the model that best fits the data.

4.12 FORECASTING

Forecasting is the process of making predictions based on past and present data collected at regular intervals over time. Time series forecasting covers analysing historical patterns, identifying trends and seasonal fluctuations. Also use statistical models or machine learning algorithms to make predictions.

4.13 RESIDUAL ANALYSIS

Residual analysis is a component of model validation. Residuals are the differences between the observed values and the predicted values of the predicted model. Analyzing the residuals assess the model's performance.

4.14 MODEL EVALUATION

The accuracy of the model is evaluated using Mean Squared Error(MSE), Root Mean Squared Error (RMSE).

Mean Squared Error(MSE): MSE is the average of the squared difference between predicted values and actual values.

Root Mean Squared Error (RMSE): RMSE is calculated by taking the root of mean squared error.

CHAPTER-5
LINEAR REGRESSION

Linear Regression is a statistical method used for predicting the relationship between dependent and independent variables. It consider a linear relationship between the independent and dependent variables. The aim of linear regression is to find the best fit straight line which explains the relationship between the variables to make the predictions based on the new data points. Linear regression is widely used in healthcare, economics for forecasting, tend analysis and understanding the impact of one variable on another. Linear regression models are evaluated based on metrics like R-squared, MSE (Mean Squared Error), RMSE (Root Mean Squared Error) to measure the accuracy of predictions.

CHAPTER -6

RESULTS AND DISCUSSIONS

The analysis chapter of this study delves into two important techniques used for forecasting and understanding patterns in data: SARIMA modelling and Linear Regression.

SARIMA is a method that takes into account both seasonal and non-seasonal patterns in time series data. It helps us to forecast future values based on historical trends and seasonal variations making it useful for predicting patterns.

On the other hand, linear regression helps us to understand the relationship between two variables. In this context, linear regression is utilized to predict one variable based on another variable.

By using these techniques, the goal is to gain insights from the data, make predictions accurately and discover hidden patterns.

6.1 TIME SERIES PLOT

The primary step of time series analysis is to draw a time series plot. Time series plot of road accidents is given in fig 6.1.

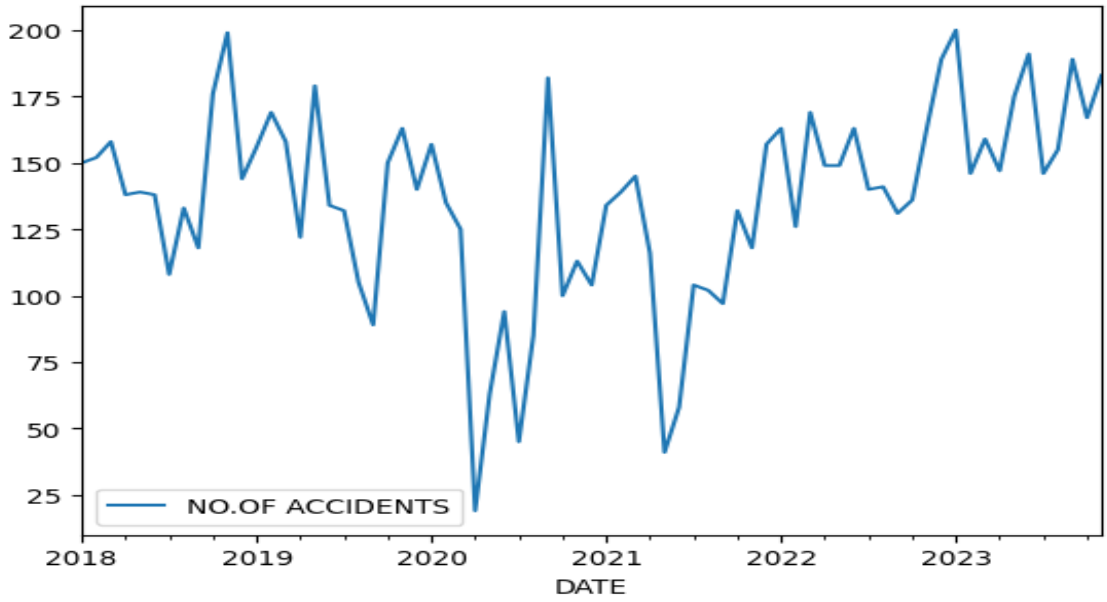

Fig.6.1

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
19	118	140	136.51	158.5	200

Table 6.1: Summary

6.2 DECOMPOSITION OF TIME

Now, perform seasonal decomposition for evaluating the trend, seasonal and random components of the time series and draw the seasonal plot. Seasonal plot is given in fig. 6.2.

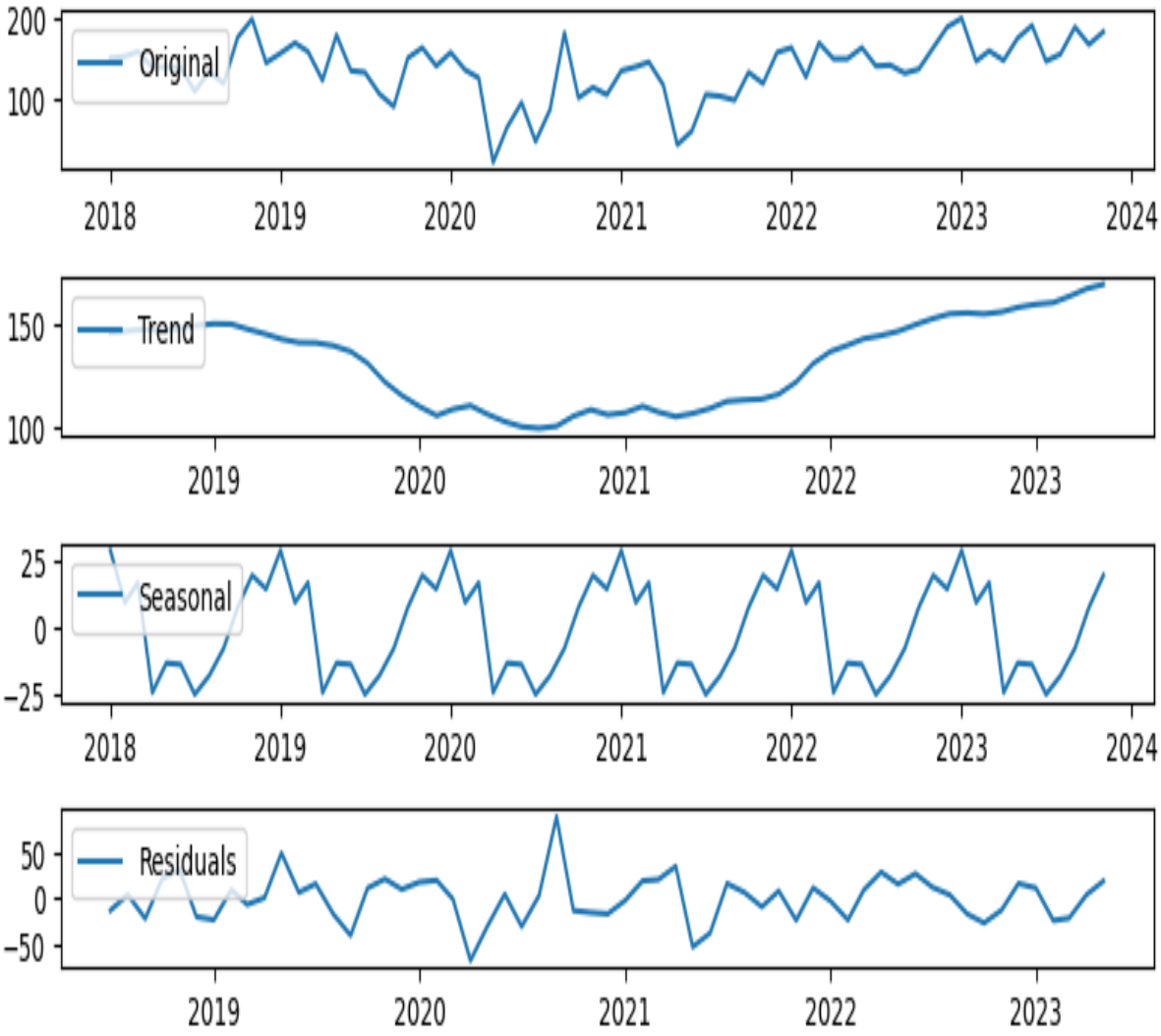

Fig. 6.2

From the fig.6.2, it can conclude that data has seasonality.

6.3 AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTION

Next step in analysing time series is to examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). To check for stationarity, plot the ACF and PACF values. ACF and PACF plot is given in fig 6.3.

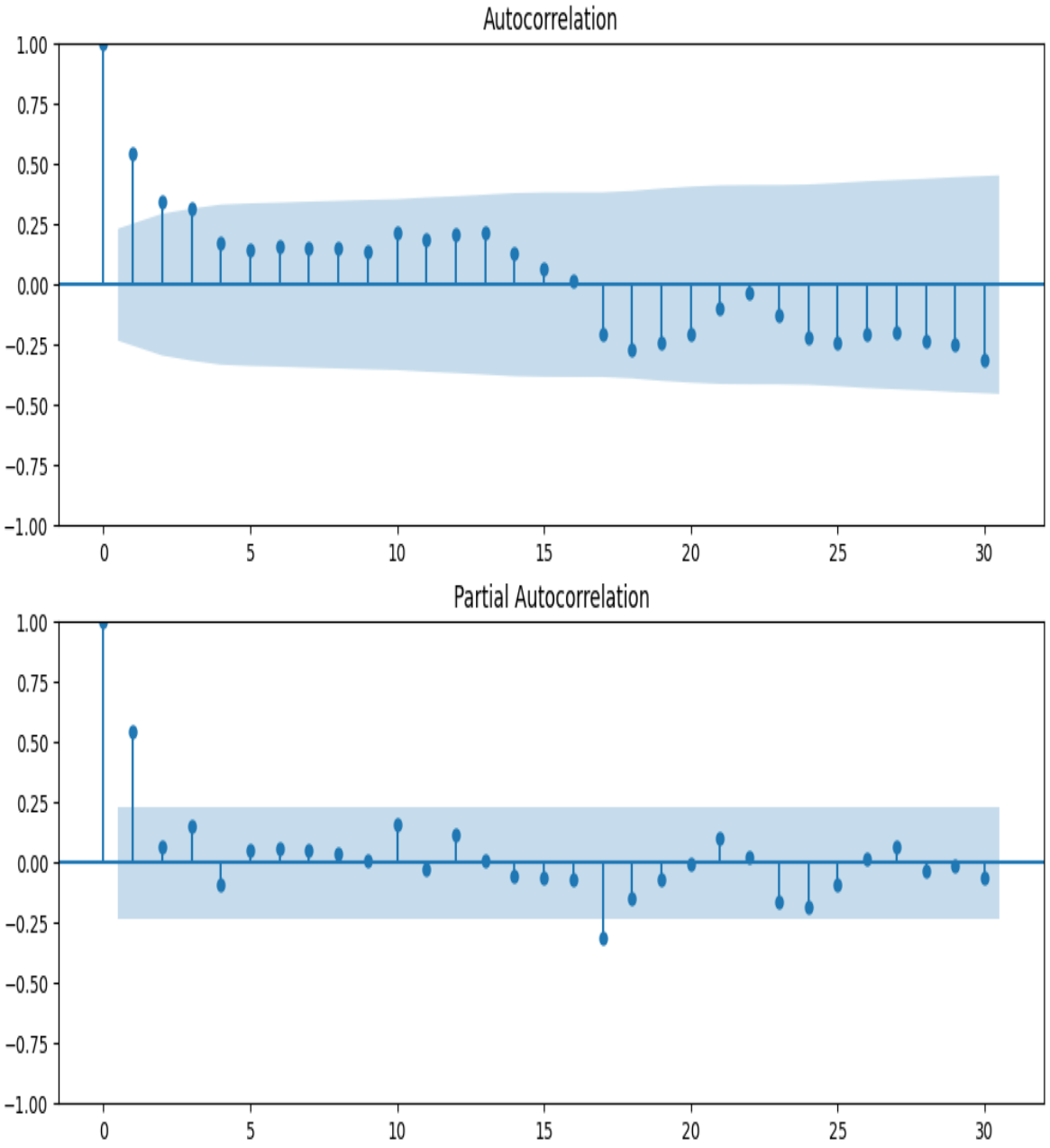

Fig. 6.3

6.4 STATIONARITY USING AUGMENTED DICKEY-FULLER TEST

To test the time series data for stationarity using ADF test, follows a hypothesis testing approach. The null hypothesis H_0 is given by,

H_0 : The data is non stationary.

The alternative hypothesis is given by,

H_1 : The data is stationary.

The outcome achieved,

ADF test statistic = -4.307426, Lag Order= 5, p-value=0.0004312

The ADF test gives the p-value 0.0004312, so fail to accept H_0 and hence can conclude that data is stationary.

6.5 SARIMA MODEL FOR ROAD ACCIDENTS

Next step is to choose the best model for forecasting. Choose the best model from all possible models according to Akaike Information Criterion (AIC). The model have lowest AIC value is the best model. Thus, the possible time series models along with their corresponding AIC statistic for the natural logarithm of road accident data are shown in table 6.2

SL NO.	MODEL ARIMA(p,d,q) x (P,D,Q)	AIC
1.	ARIMA(1, 1, 1)x(0, 0, 0)	667.28
2.	ARIMA(1, 1, 1)x(0, 0, 1)	556.97
3.	ARIMA(1, 1, 1)x(0, 1, 0)	584.75
4.	ARIMA(1, 1, 1)x(0, 1, 1)	451.13
5.	ARIMA(1, 1, 1)x(1, 0, 0)	565.29

6.	ARIMA(1, 1, 1)x(1, 0, 1)	556.41
7.	ARIMA(1, 1, 1)x(1, 1, 0)	470.53
8.	ARIMA(1, 1, 1)x(1, 1, 1)	453.12
9.	ARIMA(1, 1, 2)x(0, 0, 0)	659.36
10.	ARIMA(1, 1, 2)x(0, 0, 1)	548.00
11.	ARIMA(1, 1, 2)x(0, 1, 1)	434.64
12.	ARIMA(1, 1, 2)x(1, 0, 0)	565.79
13.	ARIMA(1, 2, 1)x(1, 1, 0)	474.85
14.	ARIMA(1, 2, 1)x(1, 1, 1,)	454.34
15.	**ARIMA(1, 2, 2)x(0, 1, 1)**	**432.01**

Table 6.2

Here the best model is ARIMA(1,2,2)$(0,1,1)_{[12]}$ with AIC value 432.02

Coefficients:

	ar1	**ma1**	**ma2**	**sma1**
	0.3672	-1.9511	1.0000	1.000
std err	0.175	1506.402	1544.123	1544.137

Table 6.2.1

6.6 DIAGNOSTIC CHECKING

Diagnostics checking is essential for ensuring the reliability, validity, and effectiveness of statistical models. It helps to choose the right model, estimate parameters correctly and improve prediction accuracy. Diagnostic plot is given below.

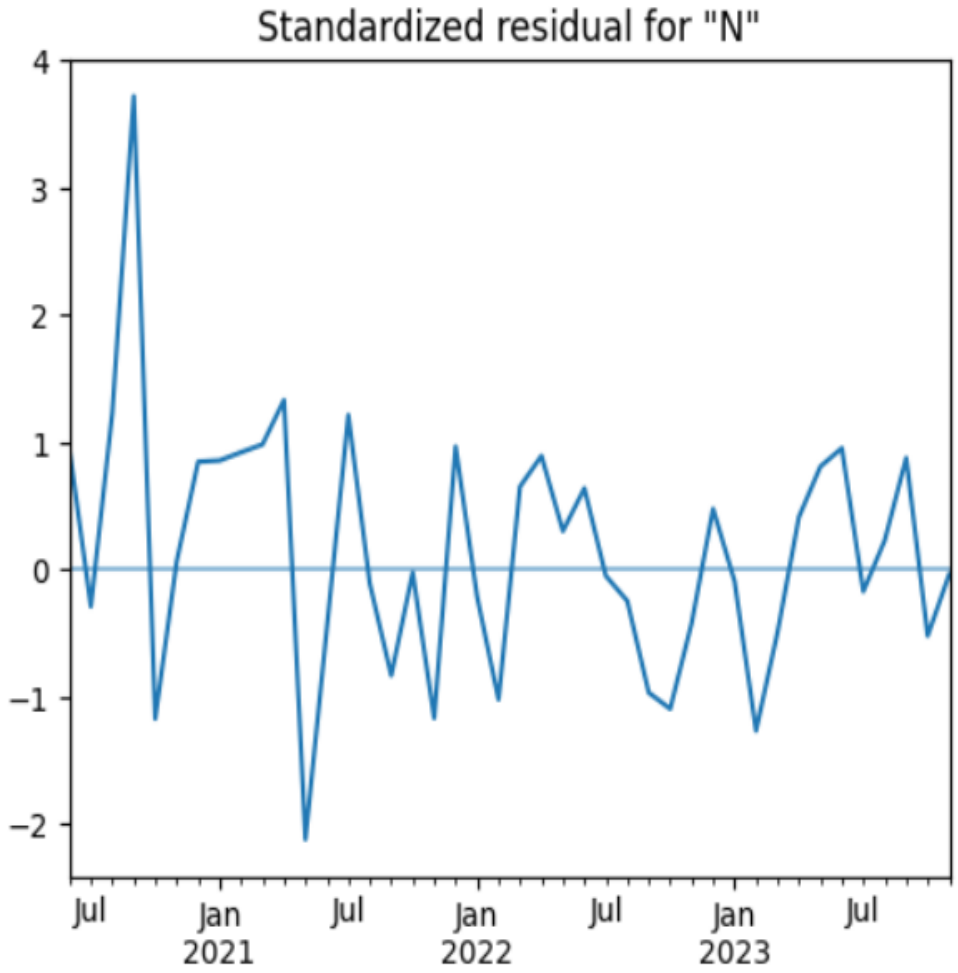

Fig 6.4

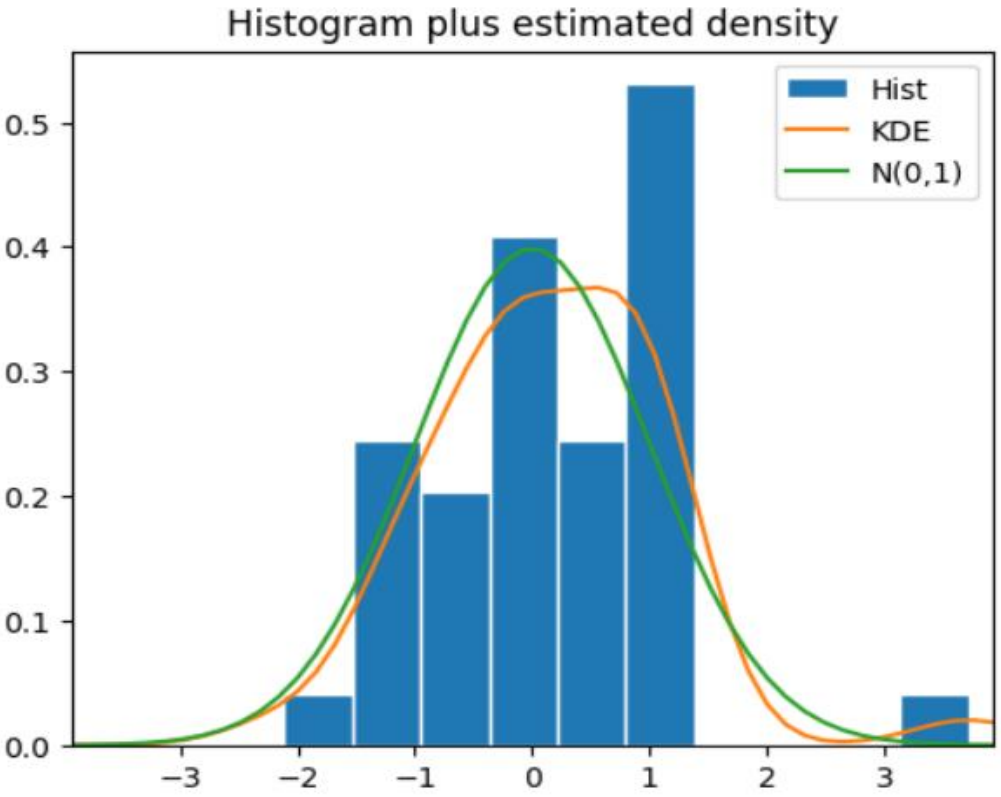
Histogram plus estimated density
Hist
KDE
N(0,1)
0.5
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3

Fig 6.5

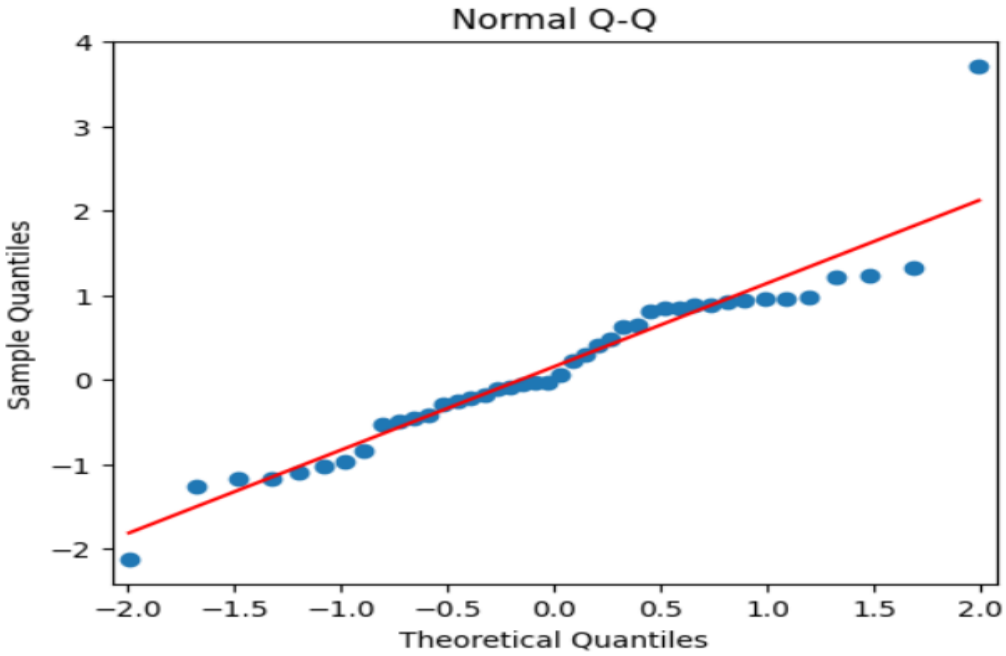
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Fig 6.6

From the Q-Q plot, most of the residuals are located on the straight line and so the standard residuals of fitted model seems to be normal.

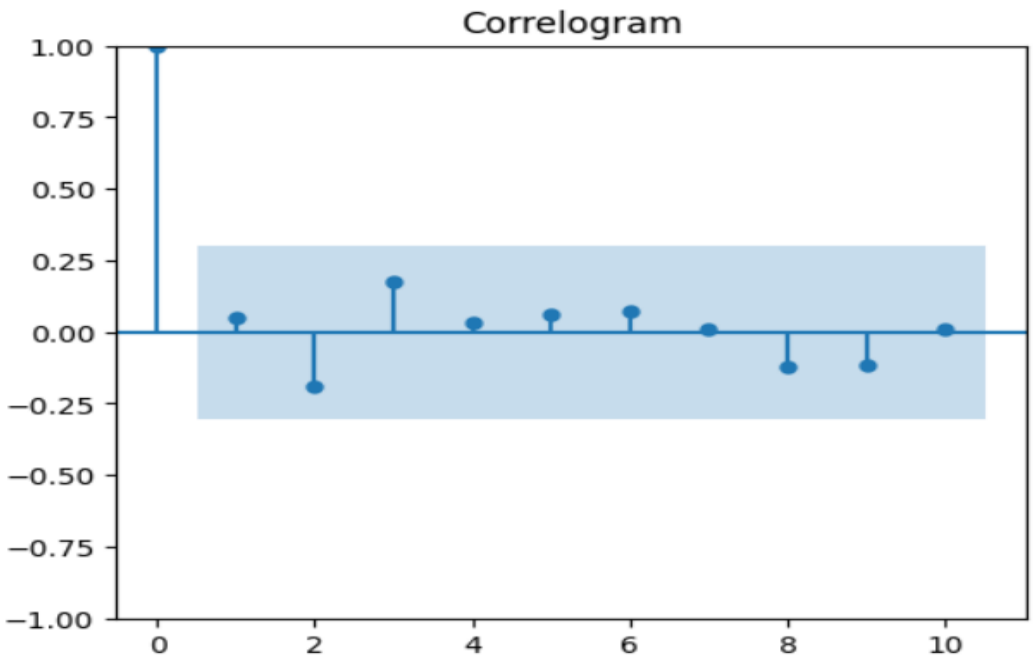

Fig. 6.7

The figure 6.7 is the correlogram which shows random fluctuations around zero in the autocorrelation function, which supports the assumption of independence of residuals.

6.7 IN SAMPLE FORECAST

In sample forecast refers to a prediction made by a model that uses data points within the range of data it was trained on. The table 6.3 is the actual and in sample forecasted values and fig. 6.8 is the plot of actual and predicted values.

MONTH	ACTUAL VALUE	PREDICTED VALUE
2021-01-01	134	100.12
2021-02-01	139	104.49

2021-03-01	145	108.20
2021-04-01	116	66.19
2021-05-01	41	120.24
2021-06-01	58	75.14
2021-07-01	104	58.81
2021-08-01	102	106.13
2021-09-01	97	127.97
2021-10-01	132	132.90
2021-11-01	118	161.39
2021-12-01	157	121.13
2022-01-01	163	170.88
2022-02-01	126	162.83
2022-03-01	169	145.56
2022-04-01	149	116.85
2022-05-01	149	138.13
2022-06-01	163	140.05
2022-07-01	140	141.82
2022-08-01	141	149.90

2022-09-01	131	165.80
2022-10-01	136	175.34
2022-11-01	163	178.12
2022-12-01	189	171.92
2022-09-01	131	165.80
2022-10-01	136	175.34
2022-11-01	163	178.12
2022-12-01	189	171.92
2023-03-01	159	176.33
2023-04-01	147	132.58
2023-05-01	175	146.76
2023-06-01	191	157.71
2023-07-01	146	152.01
2023-08-01	155	146.99
2023-09-01	189	158.39
2023-10-01	167	185.20
2023-11-01	183	184.39

Table 6.3

Plot of actual values vs predicted values

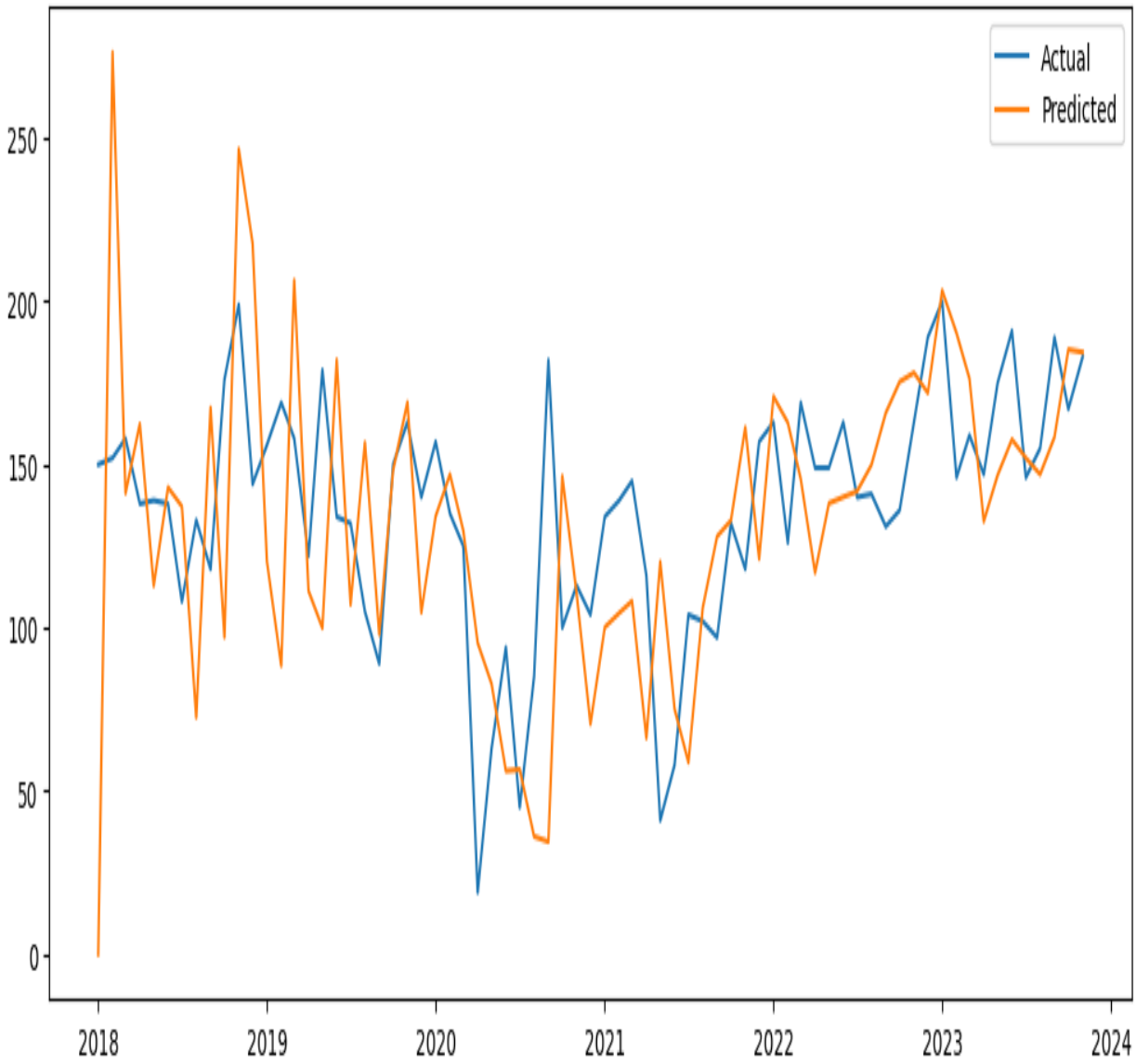

Fig. 6.8

6.8 FORECASTING

The forecast values of number of road accidents during December 2023 to November 2025 is given in the following tables.

Forecast values of road accidents from December 2023 to December 2024

MONTH	FORECASTED VALUES	LCL	UCL
DEC 2023	182.21	113.78	250.64
JAN 2024	199.68	124.94	274.42
FEB 2024	183.25	106.50	260.01
MAR 2024	191.00	111.93	270.06
APR 2024	154.10	71.99	236.22
MAY 2024	163.63	77.31	249.95
JUN 2024	169.37	77.47	261.27
JUL 2024	152.64	53.70	251.58
AUG 2024	160.74	53.33	268.15
SEPT 2024	175.34	58.11	292.58
OCT 2024	184.95	56.64	313.26
NOV 2024	198.39	57.85	338.92
DEC 2024	195.02	36.23	353.81

Forecast values of road accidents from January 2025 to November 2025

MONTH	FORECASTED VALUES	LCL	UCL
JAN 2025	211.05	36.41	385.70
FEB 2025	193.60	3.72	383.48
MAR 2025	200.47	0	406.91
APR 2025	162.76	0	386.76
MAY 2025	171.49	0	413.99
JUN 2025	176.45	0	438.31
JUL 2025	158.93	0	440.96
AUG 2025	166.25	0	469.20
SEPT 2025	180.06	0	504.66
OCT 2025	188.89	0	535.81
NOV 2025	201.54	0	571.43

Table 6.4

The graphical representation of the forecast values of number of road accidents is shown in the fig.6.9

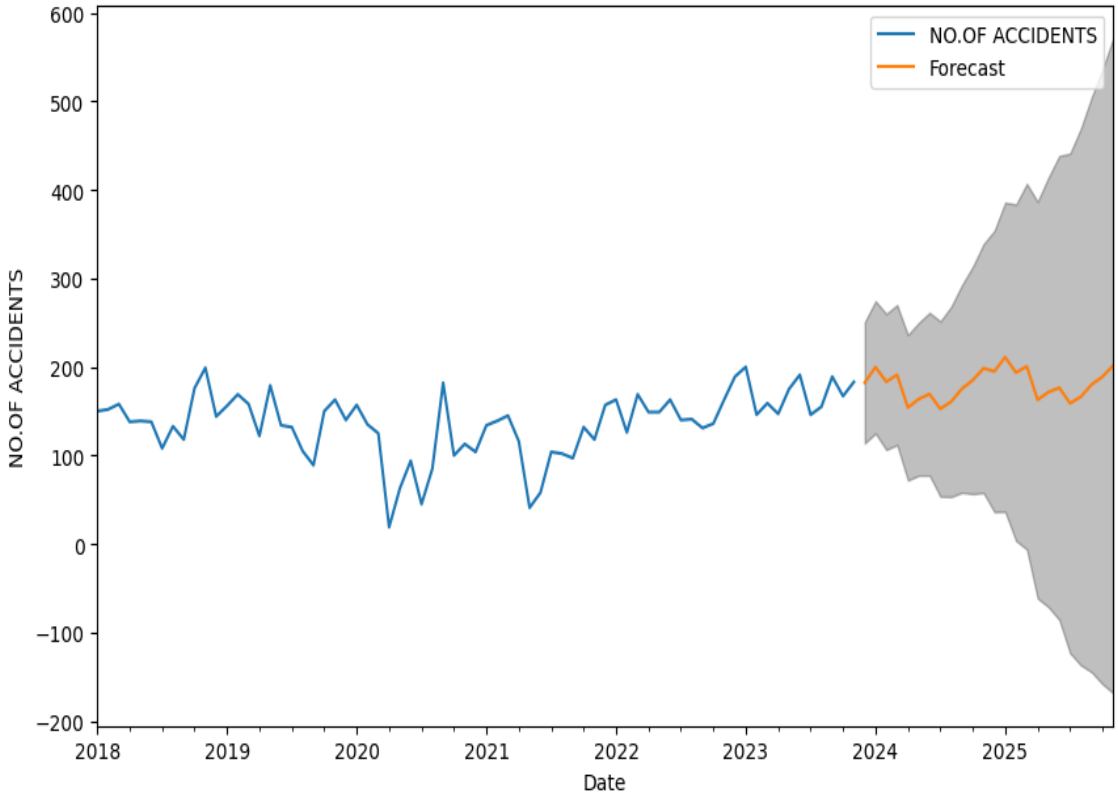

Fig.6.9

6.9 LINEAR REGRESSION ANALYSIS

The given forecasted values of road accidents for the period from December 2023 to November 2025 and the forecasting technique used to generate these forecasting values is linear regression. Linear regression used to predict the future trends based on historical data.

The forecast values of number of road accidents during December 2023 to November 2025 is given in the following table.

MONTH	FORECASTED VALUES
DEC 2023	155.73
JAN 2024	156.33
FEB 2024	156.90
MAR 2024	157.51
APR 2024	158.09
MAY 2024	158.70
JUN 2024	159.28
JUL 2024	159.89
AUG 2024	160.49
SEPT 2024	161.08
OCT 2024	161.68
NOV 2024	162.27
DEC 2024	162.87
JAN 2025	163.48
FEB 2025	164.03
MAR 2025	164.63
APR 2025	165.22

MAY 2025	165.82
JUN 2025	166.41
JUL 2025	167.01
AUG 2025	167.62
SEPT 2025	168.20
OCT 2025	168.81
NOV 2025	169.39

Table 6.5

The trend in the forecasted values shows a consistent increase over time. Starting from 155.733 in December 2023, the values slowly increase in each month, reaching 169.399 by November 2025.

Plotting predictions using Linear Regression

Fig 6.10 is the plot of the predicted vs actual accidents.

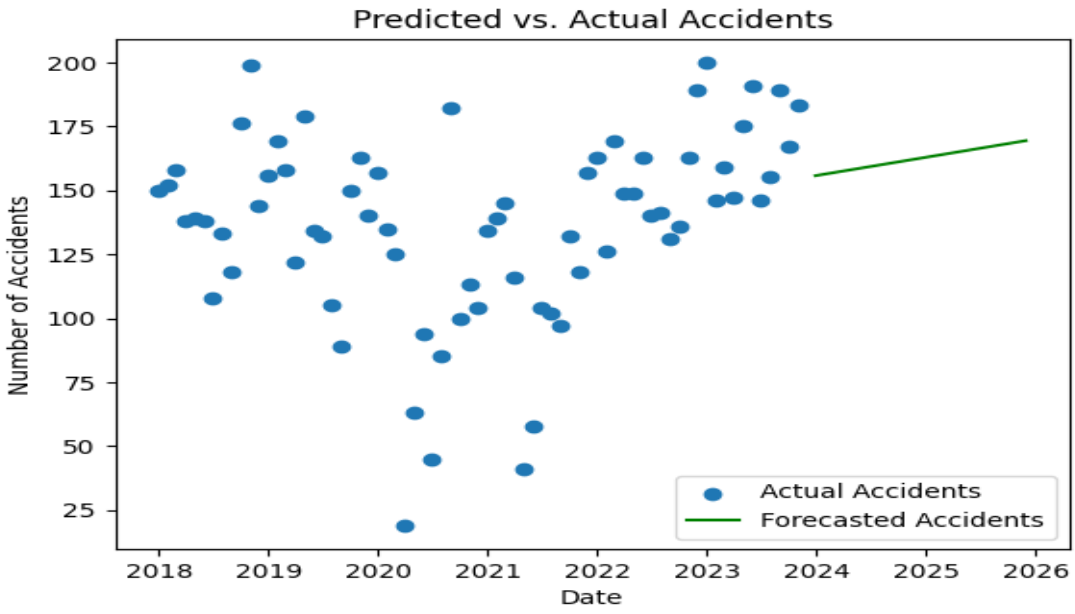

Fig 6.10

6.10 COMPARE MSE & RMSE VALUES

To determine which model is the best, the performance metrics should be compared and consider the context of the problem being addressed. In this case, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were provided for both a linear regression model and SARIMA model.

Comparing MSE and RMSE values of SARIMA model and Linear Regression model is given in the following table 6.6.

	SARIMA MODEL	**LINEAR REGRESSION MODEL**
MSE	2133.5	1282.09
RMSE	46.19	35.81

Table 6.6

The SARIMA model has a higher MSE and RMSE compared to the Linear regression model. Based on the provided metrics alone, the Linear regression model appears to perform better than the SARIMA model in terms of MSE and RMSE.

CHAPTER -7

CONCLUSION

Based on the results, after analysing the performance of SARIMA and linear regression models for forecasting road accidents based on the data from January 2018 to November 2023, the linear regression model appears to be more effective in predicting of road accidents. The accuracy of these forecasted values depends on the quality of data, regular updates are essential for better predictions.

In this study, a growing concern indicates that the forecasted road accident numbers from December 2023 to November 2025 shows a continuous increase. The rise in accidents due to the factors like population growth, increased vehicle use, changing traffic patterns, driver behaviours etc. To address this issues, proactive measures like public awareness campaign, better roads, promotion for safe driving are necessary. Over all these forecasts shows that its important for government authorities, police, transportation authorities and the public to take action and make roads safer for everyone.

REFERENCES

1. Aher, J. Analysis Of Road Accidents In Kerala, India, Using Data Mining Techniques.

2. Babu, A., & Sulaipher, M. (2022). A Study on Factors Influencing Road Accidents in Kerala. Journal of Algebraic Statistics, 13(2), 2520-2527.

3. Deretić, N., Stanimirović, D., Awadh, M. A., Vujanović, N., & Djukić, A. (2022). SARIMA modelling approach for forecasting of traffic accidents. Sustainability, 14(8), 4403.

4. Dutta, B., Barman, M. P., & Patowary, A. N. (2020). Application of Arima model for forecasting road accident deaths in India. International Journal of Agricultural and Statistical Sciences, 16(2), 607-615.

5. Parthiban, D., Vijayan, D. S., Shadhil, M. H., Reshma, U., & Krishnan, R. (2022, December). Analysis of road accident with prediction model using machine learning algorithm in the region of Kerala. In AIP Conference Proceedings (Vol. 2426, No. 1). AIP Publishing.

6. Rabbani, M. B. A., Musarat, M. A., Alaloul, W. S., Rabbani, M. S., Maqsoom, A., Ayub, S., ... & Altaf, M. (2021). a comparison between seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ES) based on time series model for forecasting road accidents. Arabian Journal for Science and Engineering, 46(11), 11113-11138.

7. Sikdar, P., Rabbani, A., & Dhapekar, N. K. (2017). Hypothesis of data of road accidents in India-review. Int. J. Civ. Eng. Technol, 8, 141-146.

8. Sunny, C. M., Nithya, S., Sinshi, K. S., Vinodini, V., KG, A. L., Anjana, S., & Manojkumar, T. K. (2018, August). Forecasting of road accident in Kerala: A case study. In 2018 International Conference on Data Science and Engineering (ICDSE) (pp. 1-5). IEEE.

9. Vipin, N., & Rahul, T. (2021). Road traffic accident mortality analysis based on time of occurrence: Evidence from Kerala, India. Clinical Epidemiology and Global Health, 11, 100745

10. Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Applied Sciences, 2, 1-13.