Project Report

On

FUTURE OF FUEL PRICE: PREDICTING FUEL PRICE OF DELHI – THE CAPITAL OF INDIA

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

RINU BABU

(Register No. SM22AS016)

(2022-2024)

Under the Supervision of

Ms. RAHNA BABU

THE FEAR OF THE LORD IS THE BEGINNING OF WISDOM

ST. TERESA'S COLLEGE
(AUTONOMOUS)
ESTD 1925

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI -682011

MAY 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

THE FEAR OF THE LORD IS THE BEGINNING OF WISDOM

ST. TERESA'S COLLEGE
(AUTONOMOUS)
ESTD 1925

CERTIFICATE

This is to certify that the dissertation entitled, **FUTURE OF FUEL PRICE: PREDICTING FUEL PRICE OF DELHI – THE CAPITAL OF INDIA** is a Bonafide record of the work done by **RINU BABU** 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: 29/04/2024

Place: Ernakulam

Ms. Rahna Babu
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College (Autonomous),
Ernakulam.

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

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

External Examiners

1: ....

CHINU JOSEPH
29/4/2024

2: ....

LAKSHMI SURESH
29/04/2024

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **Ms. RAHNA BABU**, 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

Rinu Babu

Date: 29/04/2024

SM22AS016

ACKNOWLEDGEMENTS

I take this opportunity to thank everyone who has encouraged and supported me to carry out this project.

I am very grateful to my project guide Ms. Rahna Babu for her immense help during the period of work.

In addition, I acknowledge with thanks to the Department for all the valuable support and guidance that has significantly contributed to the successful completion of this project.

I would also like to thank the HOD for her valuable suggestions and critical examinations of the project.

Ernakulam

Date: 29/04/2024

Rinu Babu

SM22AS016

ABSTRACT

The fluctuation of fuel prices has a significant impact on the economy and daily life of citizens. This study focuses on predicting petrol and diesel prices in New Delhi, India, using statistical analysis techniques such as Time Series Analysis and Regression Analysis. The primary objective is to determine the most accurate forecasting model between Time Series (SARIMA) and Linear Regression. By visualizing and comparing the results, the study concludes that the Linear Regression model provides the best forecast for the selected fuel price data. Additionally, the study demonstrates that even in a stable market environment, it is feasible to predict fuel prices with a high degree of accuracy. The findings of this research offer valuable insights for policymakers, businesses, and consumers, aiding in informed decision-making and strategic planning related to fuel consumption and pricing in India's capital region.

ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	RINU BABU
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. RAHNA BABU
Department	Post Gradate Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	FUTURE OF FUEL PRICE: PREDICTING FUEL PRICE OF DELHI – THE CAPITAL OF INDIA
Similarity	0% AI 3%
Paper ID	1663397
Submission Date	2024-04-19 10:31:15

Signature of Student

Signature of Guide

Checked By
College Librarian

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

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Fuel prices are constant concerns for everyone, from individuals to businesses. As a result, predicting the fuel prices is important for making effective decision in various sectors such as transportation, agriculture, etc. Also increase in fuel prices have a negative impact on entire economic growth. It will lower consumer spending and make a way for smuggling also. Fuel price have increased dramatically in recent years, which have greatly affected the country's mobility.

The prediction of fuel prices, particularly petrol and diesel, plays a vital role in various sectors. This includes transportation, manufacturing, and also in household budgeting. In the history of Delhi. Where these fuels are essential to everyday life? Forecasting their prices accurately can help policymakers, businesses, and individuals in making reasonable decisions? In the crowded streets of Delhi, vehicles of all shapes and sizes are steering through the daily chaos. Here petrol and diesel are like the vital principle that keeps the city move. Whether it's shuttle to work, transporting goods, or simply running errands, these fuels are essential. But have you ever wondered how the prices of petrol and diesel are determined? More importantly, is it possible to predict how much they will cost in the future?

Since 2022, the prices of petrol and diesel in Delhi have remained constant. This is influenced by different factors like international crude oil prices, government policies, taxes etc. Understanding the historical trends and patterns of fuel prices is very important to get effective predictive models. Previous studies used time series analysis to predict fuel prices with different expand of accuracy. SARIMA is a powerful tool for holding seasonality and trends in time series data. Linear regression can strengthen the accuracy and prediction.

Petrol and diesel prices will affect our daily lives. Predicting their prices are essential for budgeting and policy-making. Fuel prices are stable since 2022. But global events and regulations represent uncertainties. Accurate predictions provide benefits like logistics planning. Ensuring affordability and sustainability is important here. Public awareness play an important role in notifying energy-related problems.

Imagine that it can predict the future of fuel prices. It is like having a crystal ball that helps in budget planning, changes in transportation costs and arrange the fuel purchases to save money. That is where we use predictive modeling. It can build a model that predict future prices with accuracy, by analyzing historical data and finding the patterns. As a result, this project will introduce the topic of fuel price prediction of Delhi.

This project doesn't depend on guesswork or intuition. It gets straight to business by analyzing the data deeply. This project gathers historical data of fuel price from government websites. Here is the use of time series analysis and machine learning comes in. This project makes use of powerful tools like SARIMA modelling and Linear Regression. This project try to create a model that predict fuel price more accurately.

The goal of project is simple but ambitious. That is to develop models that can predict petrol and diesel prices in Delhi. But the thing to be noted that the prices have been stable since 2022. While stability is good news for consumers, but it makes a challenge for predictive modeling. Is it still possible to accurately forecast prices in a stable market? That's what this project aims to find out.

This project work deals with the prediction of fuel price of India's capital using SARIMA (Seasonal Autoregressive Moving Average) modelling and Linear Regression model. This project also performs the comparative study of accuracy of the SARIMA modelling and Linear Regression algorithms for fuel price prediction of the original data.

ABOUT THE DATA

For the purpose of the present study, data was collected from **(https://ppac.gov.in/).** The data consist of petrol and diesel from 2003 to 2023 of India's capital. The government changed over time. The new fuel charge update rule came into effect.

OBJECTIVES OF THE STUDY

The main objectives of the study is as follows:

- To forecast future fuel (petrol and diesel) prices of Delhi using SARIMA modelling.
- To predict fuel prices of Delhi using Linear Regression algorithms.
- To compare the accuracy of the SARIMA modelling and Linear Regression algorithms for fuel price prediction.

Dept of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam

CHAPTER 2

LITERATURE REVIEW

- Rao and Parikh (1996) conducted study forecasting and analyzing the demand for petroleum products in India. They used a combination of econometric and time-series models to evaluate the order for different petroleum products. Their methodology contains a set of regression analysis, error correction models and ARIMA models. This is to assess the order for each product. Additionally, they used artificial variables to catch the influence of factors like price change, income growth, etc. The analysis by Rao and Parikh shows that the income and price as the renter of petroleum product order in India. A positive correlation between infrastructure development and demand levels are also noted. Their project suggest increase in demand for petroleum products in India. And also cause of increasing income levels and a population growth.

- Sajal Ghosh (2006) explores the future demand for petroleum products in India. This study investigates the long-term relationship between overall petroleum products consumption and economic growth in India through cointegration analysis in time series and error-correction modeling in scenario analysis. It predicts the demand for total petroleum products and middle-distillates and offers insights into the necessary investments in the Indian refinery sector. The research concludes that India's demand for petroleum products will continue to grow rapidly, driven by factors like population growth, urbanization, and industrialization. Furthermore, it highlighted the imperative for increased investment in infrastructure and alternative energy sources to accommodate this increasing demand.

- Sonia Yeh’s (2007) research aimed to assess the factors influencing the adoption of alternative fuel vehicles (AFVs) in the United States. Employing a mixed-methods approach, the study encompassed a survey involving 1,000 households and interviews with key stakeholders in the AFV industry. Data analysis utilized statistical software tools, including SPSS and STATA, with econometric modeling employed to estimate the impact of these variables on the probability of AFV adoption. The findings of this study are potentially applicable to other types of AFVs as well. The research finds that both fuel prices and vehicle costs significantly affect the likelihood of AFV adoption, with higher fuel prices and lower vehicle costs increasing the probability of households adopting AFVs. Additionally, the study finds the significance of consumer attitudes towards AFVs, particularly concerning environmental and energy security, in shaping important adoption decisions.

- Gawande and Kaware's (2013) research on the Fuel Adulteration Consequences in India aimed to identify various types of adulterants in fuel and assess their impacts on the environment and human health. Utilizing a combination of experimental and analytical techniques, the study investigated the presence of adulterants and their effects. Gas Chromatography (GC), Infrared Spectroscopy (IR), Ultraviolet-Visible Spectroscopy (UV-Vis) were used as analytical tools to detect fuel adulteration with kerosene and subsequent tailpipe emissions that cause environmental risks. The study revealed that GC proved to be the most efficient method for detecting fuel adulteration, boasting a success rate of 98%. Although IR and UV-Vis were also effective, but their success rates were comparatively lower at 85% and 75%, respectively.

- Vashist and Ahmad (2014) paper delves into the statistical analysis of diesel engine performance when using blends of castor and jatropha biodiesel. This contribution stands as a noteworthy addition to biodiesel research. It explores the application of statistical tools in analyzing engine performance with biodiesel blends, offering valuable insights into their potential for diesel engines. Employing ANOVA, regression analysis, and correlation analysis the authors examines engine performance, enabling them to identify significant differences among biodiesel blends, and to elucidate the relationships between different engine parameters and fuel characteristics. The research revealed that the use of blended fuel resulted in reduced emissions of carbon monoxide and hydrocarbons, but there is an increase in nitrogen oxide emissions. Moreover, it identified the optimal blend ratio of castor and jatropha biodiesel was 20:80.

- Sasikumar and Abdullah (2017) explore the correlation between the stock price of Oil India Limited and fuel prices in India by using a vector autoregressive (VAR) approach and Granger Causality Test. This study holds significance because the influence of oil prices on the Indian economy, and understanding the relationship between oil prices and fuel prices are important for policymakers and investors. The study findings revealed that there is a significant impact of oil India stock prices on fuel prices in India. The VAR model able to depict the dynamic relationship between these two variables, while the Granger Causality Test confirmed the causal bond between them.

- Vempatapu and Kanaujia (2017) focuses on the use of spectroscopy techniques to identify and measure adulteration in petroleum fuels. The authors argue that traditional methods of detecting fuel adulteration are both time-consuming and ineffective, supporting for spectroscopy as a more accurate and dependable approach. Authors used a combination of gas chromatography-mass spectrometry (GC-MS) and Fourier transform infrared (FTIR) spectroscopy to find and measure the presence of adulterants in petroleum fuel samples. The research found the frequent occurrence of petroleum fuel adulteration in India, with a significant proportion of samples containing illegal additives like kerosene, diesel, and naphtha.

- Lahari et al. (2018) examines the use of Recurrent Neural Networks (RNN) in forecasting fuel prices. The authors argue that accurate prediction of fuel prices is essential for industries like transportation and logistics, facilitating informed decision making based on pricing and supply chain management. Here, RNNs were used for fuel price prediction, using python programming language for the model implementation, and Keras, a high-level neural networks API, were employed for the model development. The findings revealed that the LSTM model is better than the RNN model in terms of accuracy. Furthermore, the study also reveals that using historical fuel price data and economic indicators as input features yielded in better predictions compared to utilizing only historical fuel prices as input variables.

- Bhuvandas and Gundimeda (2020) analyzed the welfare impacts of transport fuel price changes on Indian households, with a specific focus on assessing the distributional effects across different income groups. The study employs household survey data and employed a computable general equilibrium model to simulate the impacts of fuel price variations on different economic indicators. The study also used a computable general equilibrium (CGE) and Microsimulation models to simulate the effects of fuel price changes on households in India. The findings indicated that increase in fuel price have a negative impact on the welfare of Indian households, particularly those with lower incomes. Furthermore, the results suggests that policy interventions aimed at reducing fuel consumption, like fuel taxes or subsidies for alternative fuels, could have a significant influence on household welfare.

- Anusree and Sarika (2022) explore the impact of increasing petrol and diesel prices on both the economy and households across fourteen cities in India. The authors used a case study methodology to analyze the impact of fuel price hikes on different sectors, including transportation, agriculture, and manufacturing. The purpose of this study is to utilize the regression and correlation analysis to assess the increasing fuel prices across fourteen different Indian cities. The statistical methods like correlation, regression and ANOVA were used to analyze the collected data. Authors revealed the significant impact on the transportation sector where increased fuel price resulted in increased financial burden for commuters and businesses. Also, increasing fuel prices were observed to have a streaming impact on other sectors, including agriculture and manufacturing, which depends on transportation for their operational needs.

CHAPTER 3

MATERIALS AND METHODOLOGY

3.1 DATA COLLECTION

For the study purpose data set over Delhi were downloaded from the official website of Petroleum Planning and Analysis Cell

https://ppac.gov.in/retail-selling-price-rsp-of-petrol-diesel-and-domestic-lpg/rsp-of-petrol-and-diesel-at-delhi-up-to-15-6-2017.

3.2 DATA DESCRIPTION

The dataset contains petrol and diesel price from 2003 to 2023 of India’s capital, i.e., New Delhi. The government changed over time. The data consist of the date, petrol price and diesel price. The fuel price is Indian rupees.

3.3 METHODOLOGY

The crucial step in this analysis was to study the data in detail. The main purpose of the study was to predict the future fuel prices in New Delhi using time series model and predict the actual price using linear regression model. Initially, the model was forecasted using time series technique. Then, linear regression was employed for prediction, and the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values of both approaches were calculated.

➢ TOOLS FOR ANALYSIS AND FORECASTING

- SARIMA (Seasonal Autoregressive integrated moving average) modelling
- Linear Regression algorithm

➢ TOOLS FOR COMPARISON

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

3.3.1 Time series analysis

Time series analysis means examining data points that is collected at regular intervals to find underlying patterns and trends over time. It is a statistical method that remove meaningful statistics and features from data change with time. Data consist of a sequence of data points arranged in chronological order. It is commonly used in different fields like statistics, finance, etc. This analysis helps to understand how variables change over time. The main components of time series data include trends and seasonality. They define the patterns and behavior of the data. Also, this analysis contain visually and statistical analysis of data. That is to understand its characteristics like trends and seasonality.

Time Series Analysis is a statistical method used to estimate data collected over successive interval of time. Its aim is to identify trends and make future forecasts according to historical data. Different models like moving averages, ARIMA, SARIMA, and machine learning are used here. This is to represent the fundamental structure of the time series data. Model is trained and confirmed using historical data. It can be used to predict future values of data. And it will provide valuable information for decision-making.

Analyzing the performance of the is crucial. It is fulfilled by using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE). This will help to calculate the accuracy of the prediction.

In summary, Time Series Analysis is a powerful tool used to understanding historical trends, forecasting future outcomes, and making decisions based on the collected data.

3.3.2 Trend

The trend shows the overall tendency of the data to increase or decrease over a long period. It represents a general, long-term, average tendency. It's worth nothing that the direction of increase or decrease may not remain constant throughout the given period of time.

A trend refers to the general direction in which data points are progressing over time. That indicates if there is any increase, decrease, or stability in the overall pattern. Identifying trends is important for understanding patterns and predicting future values. Techniques like moving averages, regression analysis, and decomposition methods are used to identify and model trends in the data.

3.3.3 Seasonality

Seasonality in time series analysis refers to repetitive patterns within a dataset. It will occur at regular intervals. It is often assigned with seasonal factors such as months or years. It is essential for accurately analyze and forecast time series data. It allows analysts to recognize cyclic changes influenced by external factors like weather, holidays. Businesses can enhance resource allocation by combining seasonality into models

3.3.4 Cyclic variation

Cyclic variation in time series analysis describes fluctuations in data. This variation will not follow a consistent cycle like seasonal variations. But it will display repetitive patterns over extended periods. Cyclic variations often mirror economic or business cycles. But they can also arise from factors like technological innovation or demographic changes. Seasonality occur at regular intervals. But cyclic variations happen at irregular intervals. It can be more challenging to identifying and modeling. Understanding cyclic variation is important. Because recognizing long-term trends and making informed decisions based on historical data indifferent fields is required.

3.3.5 Irregular or Residual Component

The irregular component in time series analysis captures unpredictable variations in data that cannot be explained by trends, seasonality, or cyclic patterns. It includes random fluctuations and noise inherent in the dat. It makes challenges for systematic modeling. Understanding this component is essential. It is because the accurate forecasting and detecting anomalies is important.

3.3.6 Stationary time series

A stationary time series maintains consistent statistical properties like mean, variance, and autocorrelation. This implies that the average value of the data points remains constant over time. The dispersion of data points around the mean remains constant. Also, the relationship between observations at different time periods remains unchanged. Attaining stationarity is important for reliable time series analysis and prediction, as many modeling techniques and statistical tests depends on this assumption. Different methods like differencing and transformations, can be utilized to achieve stationarity, and statistical tests like the Augmented Dickey-Fuller test are employed to confirm stationarity.

3.3.7 Non stationary time series

A non-stationary time series is defined by fluctuations in statistical properties like mean, variance, and autocorrelation structure over time, which present difficulties in modeling and prediction. Unlike stationary time series, non-stationary time series often display trends, seasonality, or other time-dependent patterns that make accurate forecasting challenging without appropriate adjustments. Dealing with non-stationarity usually involves detrending, or de seasonalizing to stabilize the data before modeling. Common techniques for addressing non-stationarity involve differencing and transformation methods like logarithmic or Box-Cox transformations. Like stationarity, statistical tests like the ADF test are used to identify non-stationarity and help the selection of suitable modeling approaches.

3.3.8 Augmented Dickey-Fuller (ADF) test

The ADF test is a statistical hypothesis test used to determine whether a given time series data is stationary or not. Stationarity, characterized by a constant mean, variance, and autocovariance over time, is critical for accurate modeling and forecasting in time series analysis. The test evaluates a null hypothesis (H_0) that assumes the existence of a unit root, indicating non-stationarity, in contrast to an alternative hypothesis (H_1) proposing stationarity. Comparing computed test statistic and critical values from a predefined distribution is required. It will helps ADF test to determine whether to reject the null hypothesis. Rejection of (H0) indicates that the time series is stationary. Otherwise it is non-stationarity. The outcomes of the ADF test provide insights into the stationarity of the time series. Also directing subsequent modeling and prediction strategies for increased accuracy will be provided.

3.3.9 Auto Regressive Integrated Moving Average or ARIMA

ARIMA is a commonly used time series forecasting model. It is known for its combination of autoregressive (AR), differencing (I), and moving average (MA) components. It is functioned by capturing the historical relationships between data points, trends and seasonality. The autoregressive component is considered as the influence of past observations on current values. The differencing component adjusts for trends or seasonality by differencing the data. Finally, the moving average component will model the influence of previous forecast errors on future values.

The model parameters, ARIMA (p, d, q) are identified by analyzing the autocorrelation and partial autocorrelation functions of the data. The model is fitted to the historical data to estimate coefficients and optimize performance. ARIMA is then used to produce forecasts for future time periods. At last, the accuracy of the model is evaluated using MSE or RMSE.

3.3.10 Autocorrelation function (ACF)

The autocorrelation function is a statistical tool. It used in time series analysis to measure the correlation between a time series and its lagged versions. It computes correlation coefficients at different time lags. This is done by indicating the strength of the relationship between observations at various points. A positive coefficient indicates a positive correlation. A negative coefficient shows negative correlation. A coefficient near to zero implies no correlation. Understanding the autocorrelation structure support in interpreting the reliability of forecasts. It also helps in identifying potential issues like seasonality or trend effects in the data. It enables relevant decision-making in time series analysis.

3.3.11 Partial Auto Correlation Function (PACF)

The Partial Auto Correlation Function is a statistical tool used in time series analysis to evaluate the correlation between data points at different lags while considering the influence of intermediate data points. Unlike ACF, which measures direct correlations at different lags, the PACF separates the relationship between observations at specific lags from the effects of shorter lags. This makes analysts to identify significant lagged relationships more directly, supporting in model selection and forecasting. Peaks in the PACF plot shows strong correlations at specific lags, helping to identify vital time intervals with predictive power in the time series data.

3.3.12 Auto Regressive (AR) Components

The AR components in a time series model is the relationship between a data point and its lagged values. These components assume that current value of the time series data can be forecasted based on its past values, following a linear relationship. In an autoregressive model, each observation is represented as a weighted sum of its own past observations, with the weights identified by the parameters of the model. The order of autoregression (p) in an ARIMA (p, d, q) model, indicates the number of lagged observations considered in the model. Parameters influencing this relationship are estimated from past data, indicating the strength and direction of influence of each lagged observation on the current value. AR models are mainly used for forecasting future values of a time series, utilizing past behavior to generate predictions and discover potential trends and patterns in the data.

3.3.13 Moving Average (MA) component

The MA component is vital aspect of the ARIMA model in time series analysis. It mainly focuses on predicting future values by considering the average of past forecast errors. Unlike the autoregressive component, which directly models the relationship between current and past observations, the MA component shows the influence of previous errors on future values. The order of the moving average component(q), in the ARIMA (p, d, q) model, determines the number of lagged forecast errors integrated into the model. By smoothing out short-term fluctuations and filtering random variations, the MA component helps in capturing underlying trends in the data. When combined with the AR component, the ARIMA can effectively model a wide range of temporal dependencies present in the time series data. Parameters of the MA component is estimated from past data to determine the strength and influence of past forecast errors. It shows that the MA component plays a vital role in achieving accurate forecasting within the ARIMA model by incorporating past errors to predict future values.

3.3.14 Seasonal ARIMA or SARIMA

SARIMA is the ARIMA model by integrating seasonal components to improve forecasting accuracy for time series data with appearing seasonal patterns. It includes seasonal counterparts for the autoregressive, differencing, and moving average components, indicated by seasonal order parameters P, D, and Q respectively. SARIMA combines both seasonal and non-seasonal components to capture the combined effects of trends, seasonality, and other temporal patterns in the data. By jointly modeling the seasonal and non-seasonal variations, SARIMA offers a comprehensive approach for predicting that effectively combine the complex dynamics of seasonal time series data. Once trained on historical data, SARIMA can produce forecasts for future time periods, including both seasonal and non-seasonal patterns determined in the data.

3.3.15 Seasonal Auto Regressive (SAR) Components

The Seasonal Auto Regressive component in a SARIMA model indicates the correlation between a data point and its seasonal lagged values. It is designed for time series data with recurring seasonal patterns. The SAR component focuses on modeling the relationship between the current observation and its past values. Seasonal autoregressive order parameter is P. It determines the number of seasonal lags considered in the model. Then it will accounts for the order of SAR dependencies. SARIMA models effectively account for seasonal fluctuations by integrating the SAR component. This will improve their ability to forecast future observations in seasonal time series data.

3.3.16 Seasonal differencing

Seasonal differencing is a data adjustment method. It is used in time series analysis for eliminating seasonal patterns by subtracting values from previous seasons. Its goal is to make the data more stationary. This is done by neutralizing cyclic fluctuations. This will simplify the modeling process and enhancing the accuracy of forecast. The procedure includes identifying seasonal patterns and analyzing the resulting differences. This is done to uncover underlying trends or irregular components in the time series data. Seasonal differencing facilitates more reliable modeling by attaining stationarity. Also, forecasting of time series data makes them valuable in time series analysis.

3.3.17 Seasonal Differencing Components

The Seasonal Differencing Components aims to eliminate seasonal patterns from the data. This is done by implementing adjustments to achieve stationarity. These adjustments include computing the difference between successive observations at fixed seasonal intervals. The order of seasonal differencing (D), shows the number of seasonal differences needed for making the data stationary. Combine seasonal differencing with non-seasonal differencing components. Then SARIMA models effectively address both seasonal fluctuations and other non-stationarities in the data. Parameter is estimated from historical data.

3.3.18 Akaike Information Criterion (AIC)

AIC is a statistical measure used to compare the goodness of fit of different statistical models. The computation of AIC includes considering the number of parameters in the model. Also, the maximum likelihood estimation of the model's fit to the data. Lower AIC values is the best fitted model. It helps researchers to choose the most suitable model from a set of candidates. AIC fulfills valuable tool for model selection by balancing model complexity with goodness of fit.

3.3.19 Forecasting

Forecasting indicates predicting future values through the analysis of historical data. It includes identifying trends and patterns in the data. Also, selecting suitable forecasting techniques and estimating model parameters. In general, forecasting fulfills a crucial role for organizations. It is to anticipate future developments and plan effectively.

3.3.20 Linear regression

Linear regression is a machine learning method. It is used to establish the relationship between a dependent variable and more than one independent variables. It assumes a linear connection between the variables. The change in the independent variables results in proportional changes in the dependent variable. Linear regression facilitates predictions of the dependent variable based on the values of the independent variables.

3.3.21 Mean Squared Error (MSE)

MSE is a statistical measure used to evaluate the accuracy of predictions. It is the average squared difference between the predicted values and the actual values. Smaller MSE values gives better model performance. Because they show only smaller errors between predictions and actual observations. Lower MSE values considered to have best predictive accuracy.

3.3.22 Root Mean Squared Error (RMSE)

RMSE is a statistical measure used to evaluate the accuracy of a predictive model. It is done by evaluating the average magnitude of errors between predicted and actual values. It is the square root of the Mean Squared Error (MSE). A smaller RMSE value gives better model performance. It implies smaller errors and higher prediction accuracy.

CHAPTER 4

STATISTICAL ANALYSIS AND RESULT

This section of the study includes data analysis for the fuel price in New Delhi from January 2003 to December 2023. The analysis contains two technique parts for predicting and understanding data patterns. They are SARIMA and Linear Regression.

SARIMA is a method that incorporates both seasonal and non-seasonal patterns in time series data. It helps us to forecast future values based on historical trends and seasonal variations. This makes it a valuable tool for predicting patterns and trends in time series data.

Conversely, linear regression helps us to understand the relationship between two variables by establishing a linear equation that predicts one variable based on another. In this scenario, linear regression is used to anticipate the values of one variable by analyzing its correlation with another variable.

These two techniques are used with the goal of extracting insights from data, ensuring accurate predictions, and discover hidden patterns.

4.1 TIME SERIES PLOT OF PETROL PRICE

The primary step of time series analysis is to draw a time series plot of the given dataset. The visualization of petrol price in New Delhi is given below.

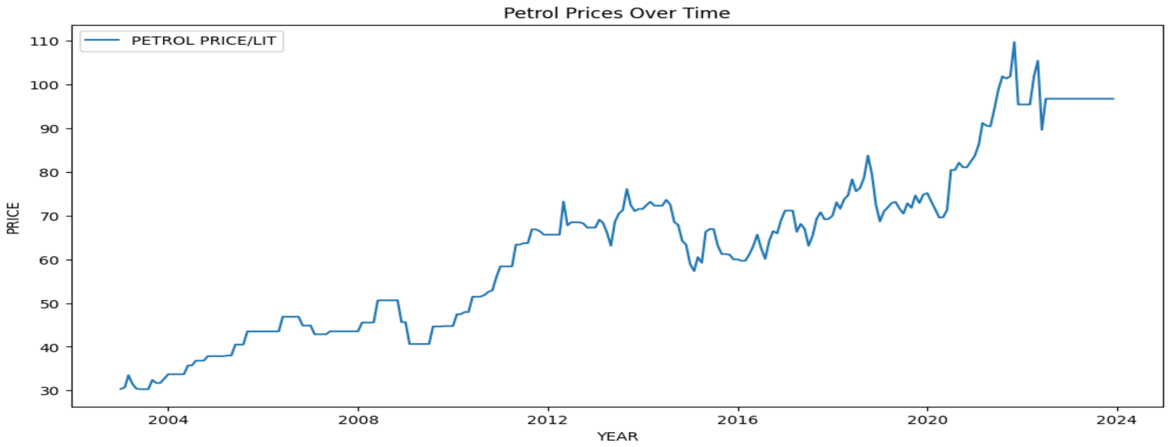

Fig 4.1

4.2 DECOMPOSITION OF TIME

Perform seasonal decomposition for evaluating the trend, seasonal and residual components of the given data. Visualization of seasonal plot is given below.

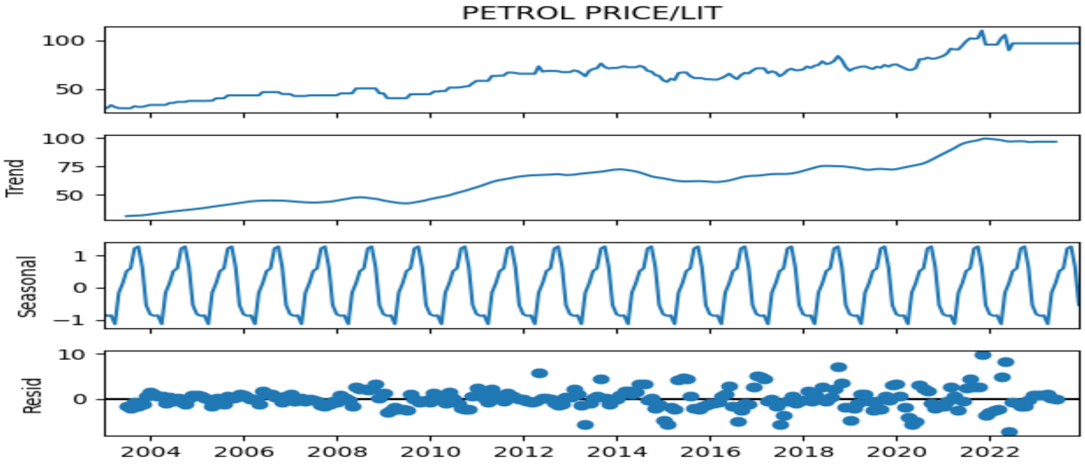

Fig 4.2

From this visualization, it is clear that the data has seasonality. But it is not stationary. Hence conduct a seasonal difference in the given data. The visualization of seasonal differenced data is given below.

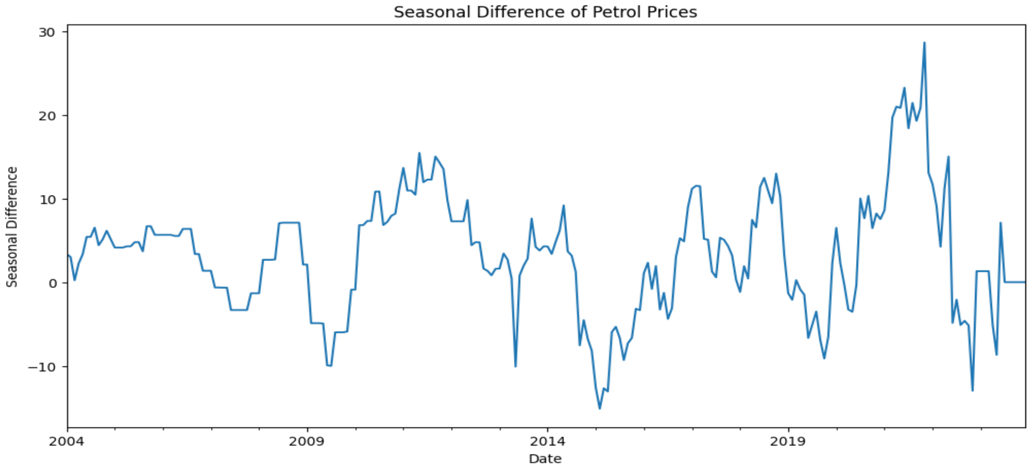

Fig 4.2.1

4.3 STATIONARITY USING AUGMENTED DICKEY-FULLER TEST

To test the data for stationarity using ADF test again with seasonal differenced data, follows a testing of hypothesis approach. The null hypothesis H_0 is given by,

H_0: The data is non stationary.

The alternative hypothesis H_1 is given by,

H_1: The data is stationary.

The result achieved is given by,

ADF test statistic = -2.994916, Lag Order= 5, p-value=0.035371

The ADF test gives the p-value 0.035371, therefore fail to accept H_0 and hence it can be concluded that the given data is stationary.

4.4 AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTION

Next step in analysis is to examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the seasonal differenced data. To check for stationarity, plot the ACF and PACF values. ACF and PACF plot is given in fig 4.4.

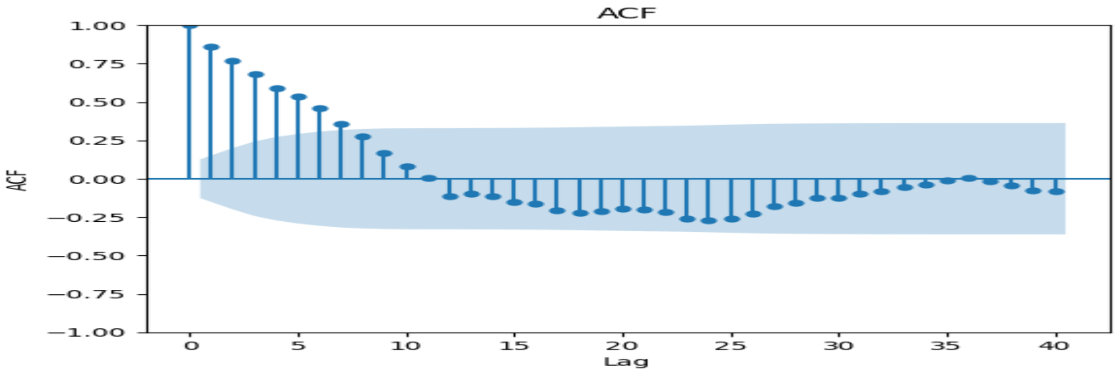

Fig 4.4(a)

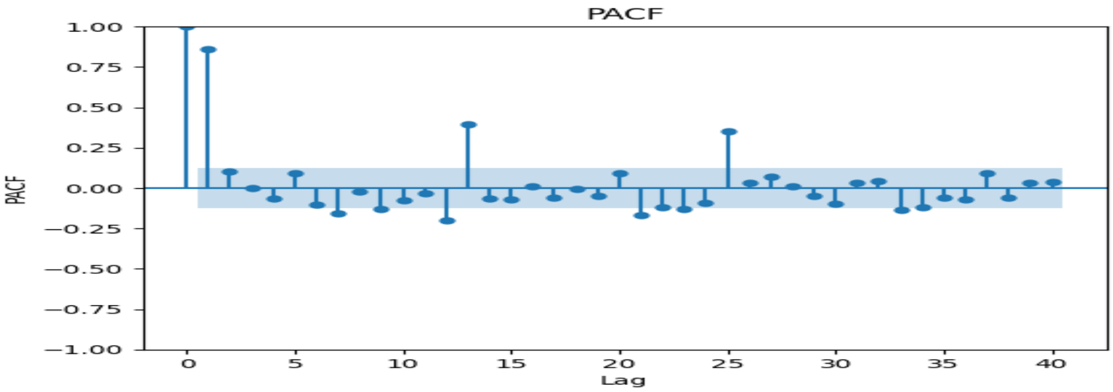

Fig 4.4(b)

4.5 SARIMA MODEL FOR PETROL PRICE

Next step is to find the best model for forecasting. The better model can choose from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value gives the best model. Thus, the possible time series models along with their corresponding AIC statistic for the natural logarithm of petrol prices are shown in table 4.5.

SL NO.	MODEL ARIMA (p, d, q) x (P, D, Q)	AIC
1.	ARIMA (0, 1, 2) x (2, 2, 1, 12)	753.1096642111524
2.	ARIMA (0, 1, 0) x (2, 2, 1, 12)	749.2541002102748
3.	ARIMA (0, 1, 2) x (2, 1, 1, 12)	746.755028118705
4.	ARIMA (0, 1, 0) x (2, 2, 2,12)	745.1123270746023
5.	ARIMA (0, 1, 1) x (2, 1, 1,12)	744.7612171846199
6.	ARIMA (0, 1, 0) x (2, 1, 1, 12)	743.0853914145239
7.	ARIMA (0, 1, 1) x (2, 2, 2,12)	742.3704464999616
8.	ARIMA (1, 1, 0) x (2, 1, 1, 12)	740.8407634876314
9.	ARIMA (0, 1, 2) x (2, 2, 2,12)	739.5618233672767
10.	ARIMA (0, 1, 0) x (2, 1, 2, 12)	737.7549077279516
11.	ARIMA (0, 1, 0) x (1, 1, 2, 12)	736.0308139395543
12.	ARIMA (0, 1, 0) x (0, 1, 2, 12)	735.8172583576918

13.	ARIMA (0, 1, 1) x (0, 1, 2, 12)	734.3572089719203
14.	ARIMA (0, 1, 0) x (0, 2, 2, 12)	729.4597975509296
15.	**ARIMA (0, 1, 2) x (0, 2, 2, 12)**	**724.6451386282131**

Table 4.5

The best model obtained here is ARIMA (0, 1, 2) x (0, 2, 2, 12) with AIC value 724. 6451.

4.6 DIAGNOSTIC CHECKING

Diagnostics checking is necessary for ensuring the reliability, validity, and effectiveness of statistical models. It enables to select the right model, estimate parameters correctly and also improve prediction accuracy.

Diagnostic plot is given below.

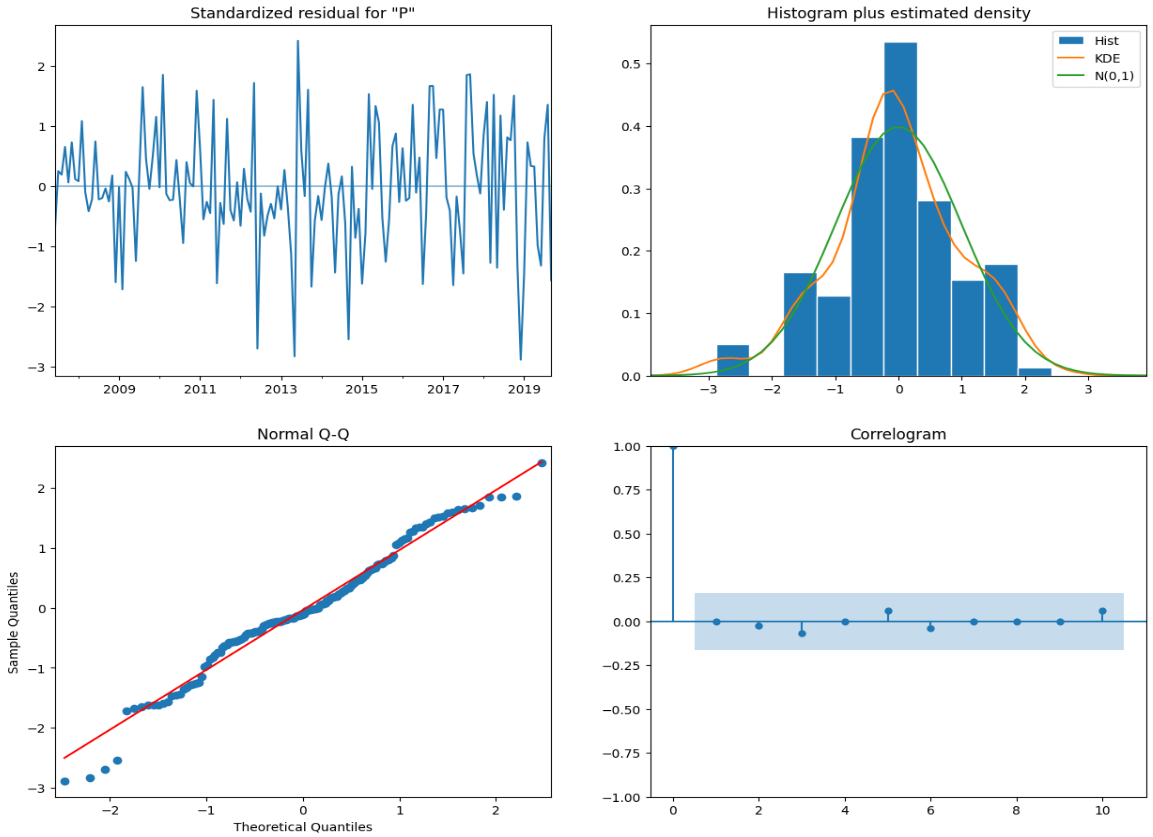

Fig 4.6(a)

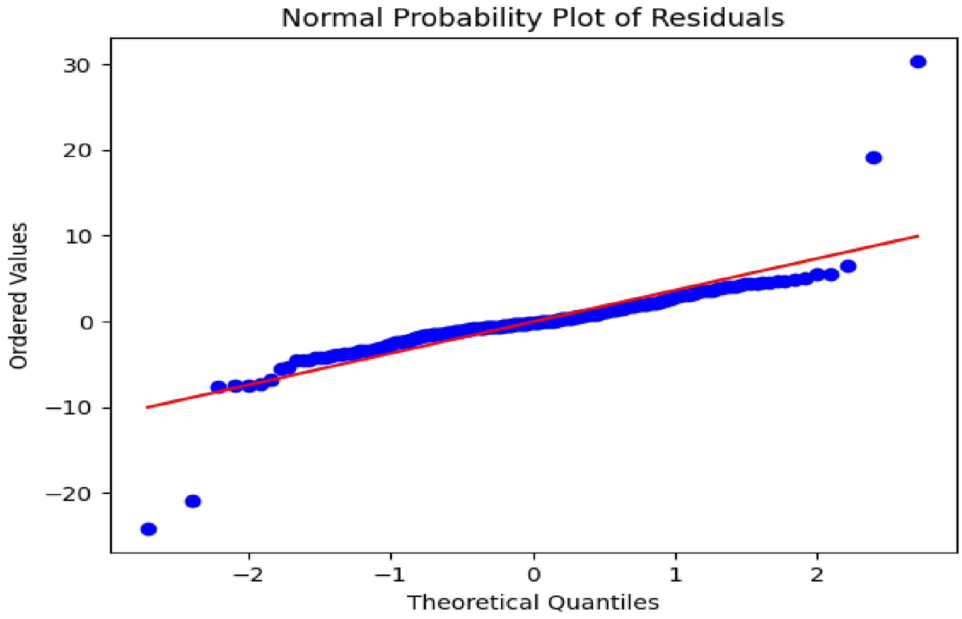

Fig 4.6(b)

From the Q-Q plot and P-P plot, most of the residuals are located on the straight line. Therefore, the standard residual of best fitted model is said to be normal.

Also, here given a plot of residuals vs fitted values:

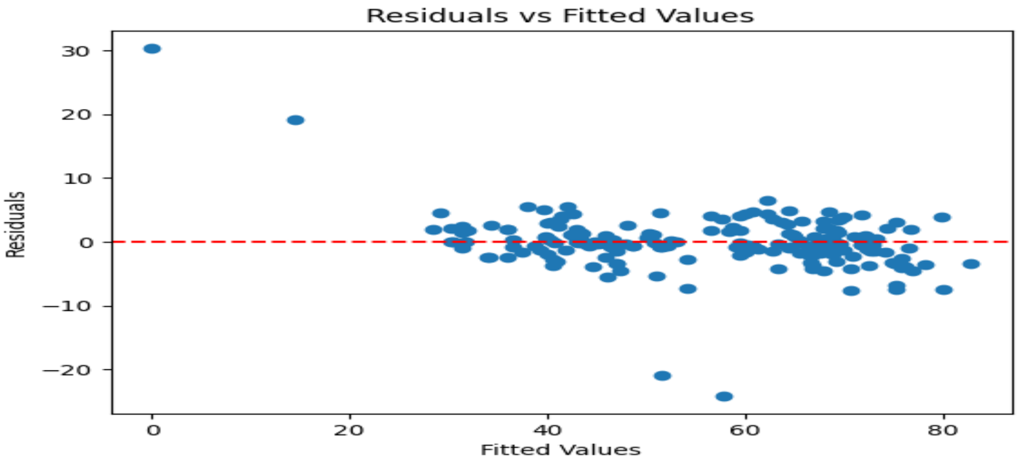

Fig 4.6(c)

4.7 IN SAMPLE FORECAST

In sample forecast is the prediction generated by the best fitted model using data points within the range it was trained on. The table 4.7 given below illustrates the plot of actual and in sample forecasted values of petrol price.

DATE	***PETROL PRICE/LIT***	**PREDICTED PRICE**
2022-01-01	95.41	90.3017
2022-02-01	95.41	90.6808
2022-03-01	95.41	90.8678
2022-04-01	101.81	98.5987
2022-05-01	105.41	100.0308
2022-06-01	89.62	84.2500
2022-07-01	96.72	95.6465
2022-08-01	96.72	95.8967
2022-09-01	96.72	95.8759
2022-10-01	96.72	95.3621
2022-11-01	96.72	95.3067
2022-12-01	96.72	96.7052
2023-01-01	96.72	96.4455
2023-02-01	96.72	96.8252
2023-03-01	96.72	96.0126
2023-04-01	96.72	96.7440
2023-05-01	96.72	98.1756
2023-06-01	96.72	96.3937
2023-07-01	96.72	96.7899
2023-08-01	96.72	98.0411
2023-09-01	96.72	98.0194
2023-10-01	96.72	98.5055
2023-11-01	96.72	98.4501
2023-12-01	96.72	98.8488

Table 4.7

4.7.1 Plot of actual values vs predicted petrol price

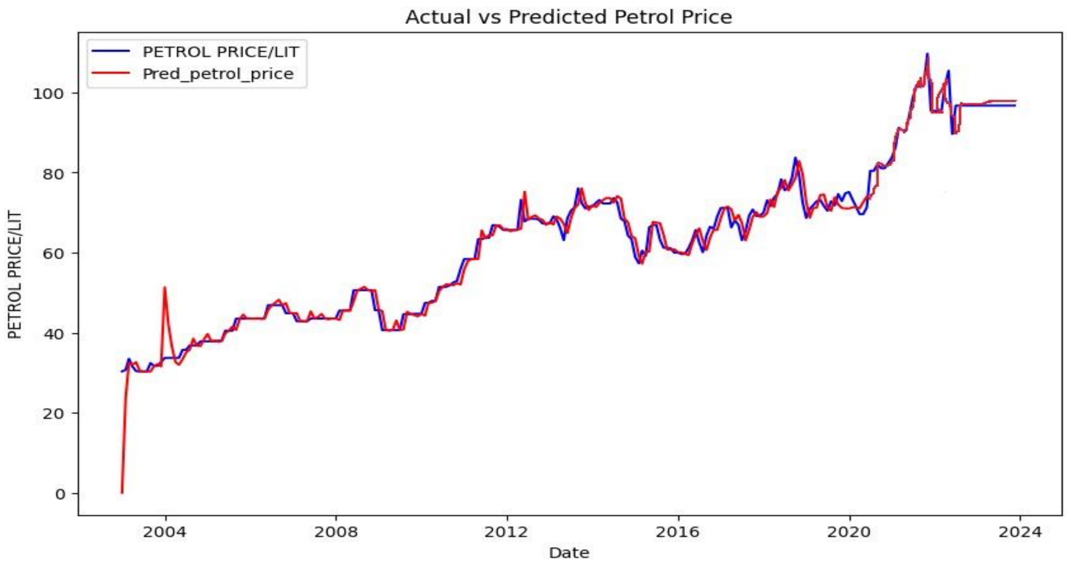

Fig 4.7.1

4.8 FORECASTING

The forecast values of petrol prices during January 2024 to December 2025 is given in the following table.

DATE	*FORCASTED* *PETROL PRICE/LIT*
2024-01-01	98.5892
2024-02-01	98.9690
2024-03-01	96.1563
2024-04-01	98.8876
2024-05-01	97.3193
2024-06-01	97.5374
2024-07-01	98.9337
2024-08-01	98.1848
2024-09-01	98.1631

2024-10-01	100.6493
2024-11-01	100.5939
2024-12-01	98.9926
2025-01-01	120.7330
2025-02-01	98.1127
2025-03-01	110.3
2025-04-01	125.0314
2025-05-01	126.4630
2025-06-01	126.6812
2025-07-01	126.0775
2025-08-01	130.3285
2025-09-01	130.3069
2025-10-01	128.7930
2025-11-01	130.1363
2025-12-01	130.7376

Table 4.8

The graphical representation of the forecast values of petrol price is shown in the below fig.4.8.

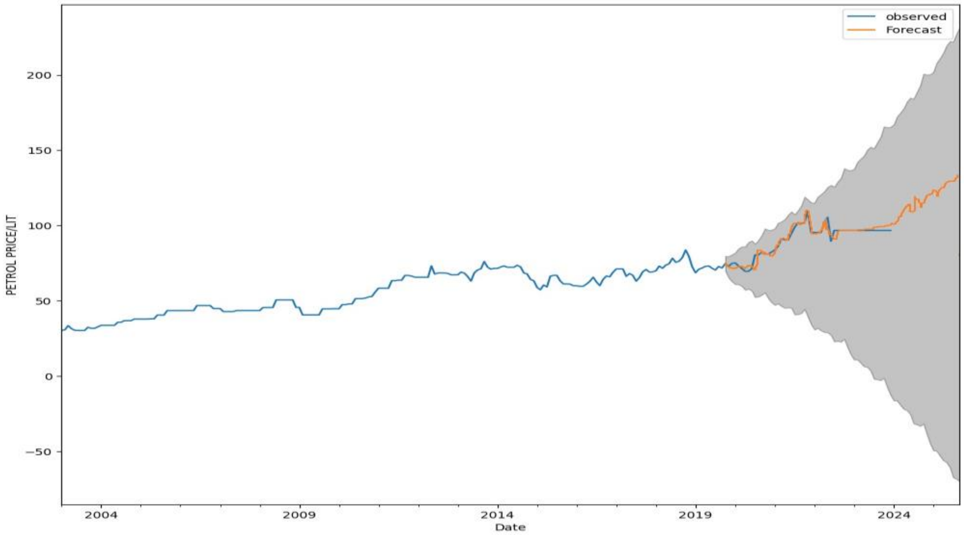

Fig 4.8

This forecasted plot of petrol prices in New Delhi shows an upward trend.

4.9 TIME SERIES PLOT OF DIESEL PRICE

The primary step of time series analysis is to draw a time series plot of the given dataset. The visualization of diesel price in New Delhi is given below.

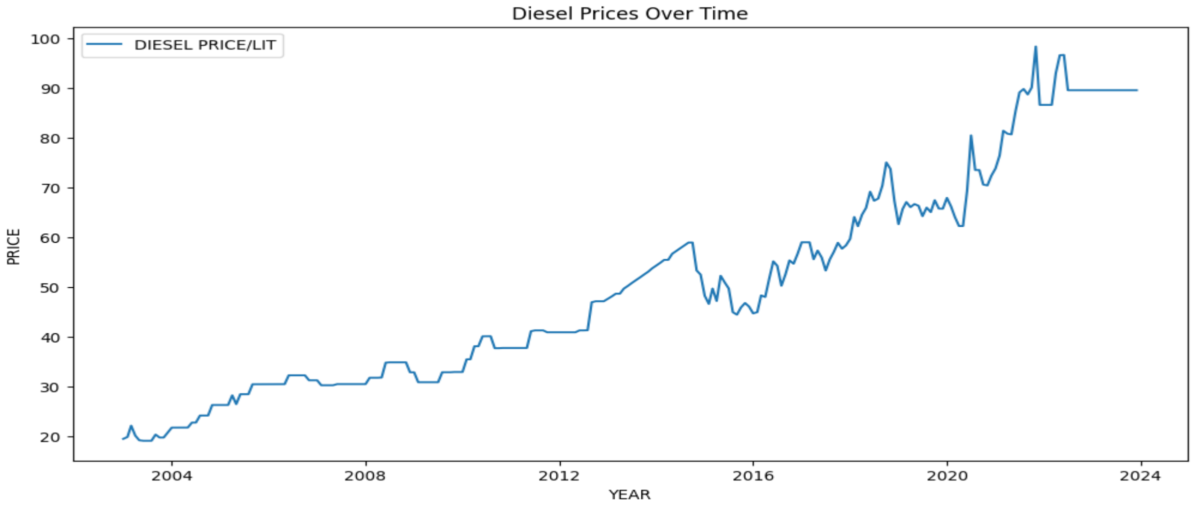

Fig 4.9

4.10DECOMPOSITION OF TIME

Perform seasonal decomposition for evaluating the trend, seasonal and residual components of the given data. Visualization of seasonal plot is given below.

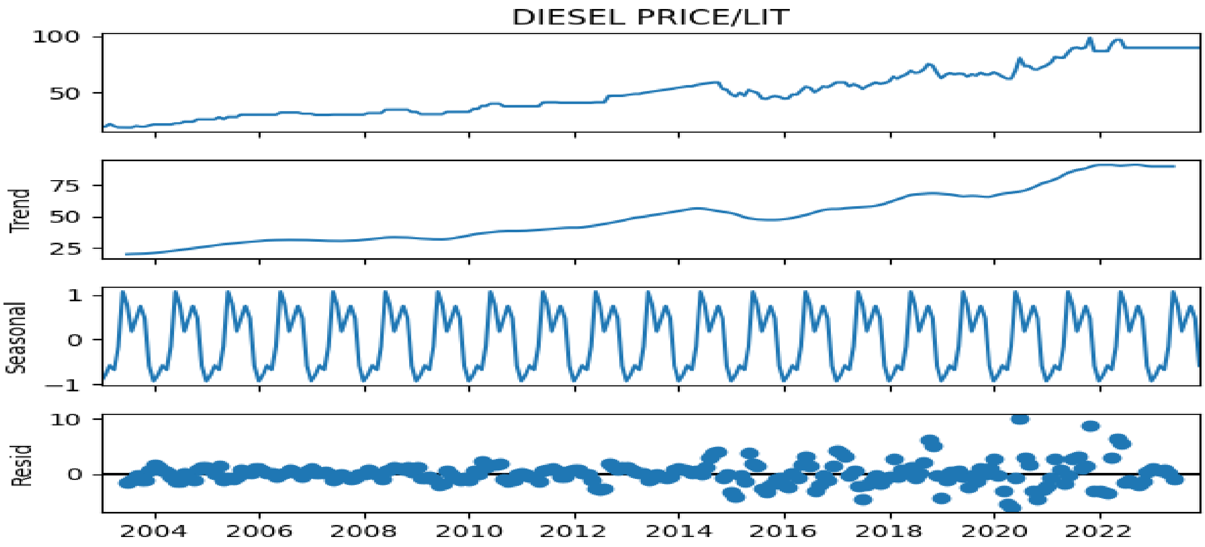

Fig 4.10

From this visualization, it is clear that the data has seasonality. But it is not stationary. Hence conduct a seasonal difference in the given data. The visualization of seasonal differenced data is given below.

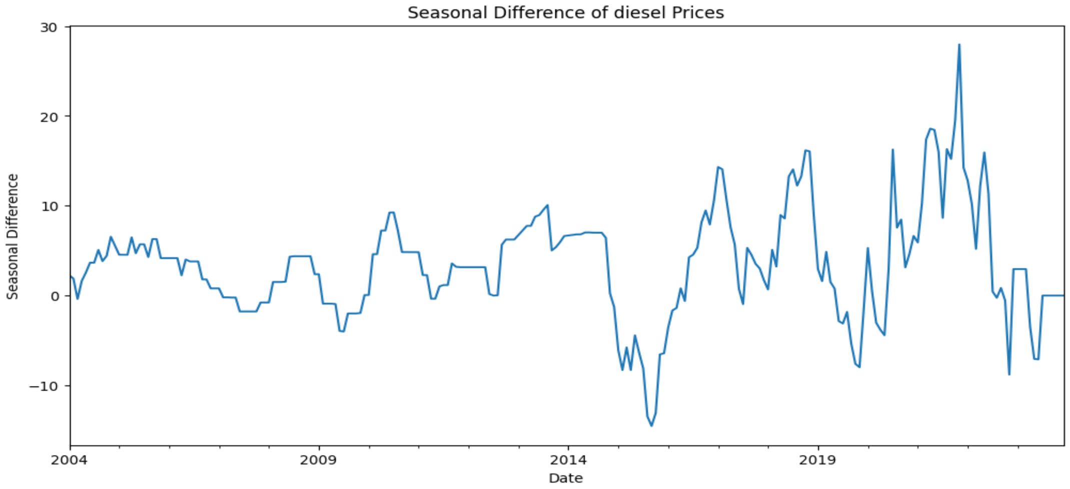

Fig 4.10.1

4.11 STATIONARITY USING AUGMENTED DICKEY-FULLER TEST

To test the data for stationarity using ADF test again with seasonal differenced data, follows a testing of hypothesis approach. The null hypothesis H_0 is given by,

H_0: The data is non stationary.

The alternative hypothesis H_1 is given by,

H_1: The data is stationary.

The result achieved is given by,

ADF test statistic = -3.53692378, Lag Order= 5, p-value= 0.00708793147

The ADF test gives the p-value 0.00708793147, therefore fail to accept H_0 and hence it can be concluded that the given data is stationary.

4.12 AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTION

Next step in analysis is to examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the seasonal differenced data. To check for stationarity, plot the ACF and PACF values. ACF and PACF plot is given in fig 4.12.

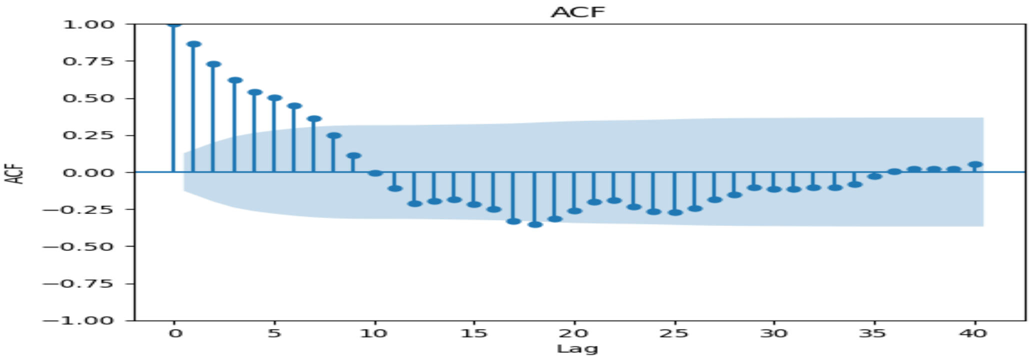

Fig 4.12(a)

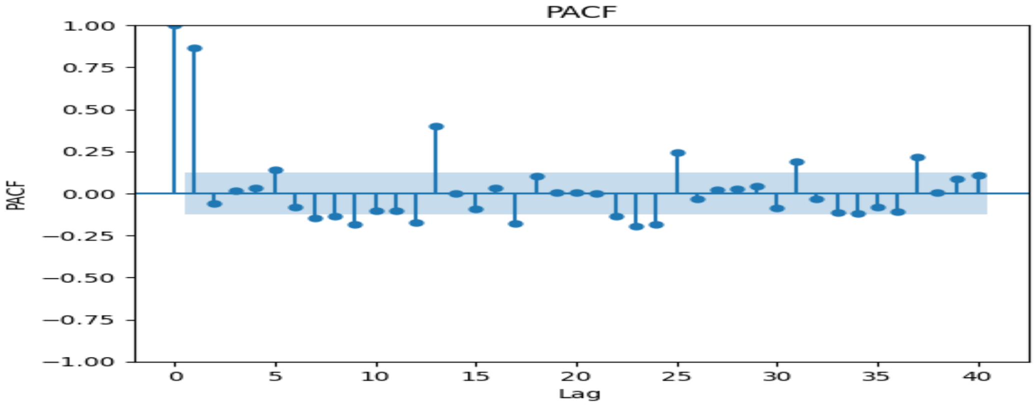

Fig 4.12(b)

4.13 SARIMA MODEL FOR PETROL PRICE

Next step is to find the best model for forecasting. The better model can choose from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value gives the best model. Thus, the possible time series models along with their corresponding AIC statistic for the natural logarithm of petrol prices are shown in table 4.13.

SL NO.	**MODEL**	**AIC**
	ARIMA (p, d, q) x (P, D, Q)	
1.	ARIMA (0, 1, 1) x (2, 2, 1,12)	668.8943914325005
2.	ARIMA (0, 1, 0) x (2, 2, 1,12)	667.5216629510834
3.	ARIMA (0, 1, 0) x (2, 2, 2, 12)	664.5563514415239
4.	ARIMA (0, 1, 0) x (2, 1, 1,12)	662.7406761199416
5.	ARIMA (0, 2, 2) x (2, 1, 2,12)	659.0975152357692
6.	ARIMA (0, 1, 1) x (2, 1, 2,12)	658.3881525345207
7.	ARIMA (0, 1, 0) x (1, 2, 2, 12)	655.3539052525427
8.	ARIMA (0, 1, 0) x (0, 1, 2, 12)	654.3949932939066
9.	ARIMA (0, 1, 1) x (0, 1, 2,12)	653.2305732505388
10.	ARIMA (0, 1, 2) x (0, 1, 2,12)	651.5130314655823
11.	ARIMA (0, 1, 2) x (1, 2, 2, 12)	649.688651756021
12.	ARIMA (0, 1, 0) x (0, 2, 2, 12)	646.2926472392098
13.	ARIMA (0, 1, 1) x (0, 2, 2,12)	644.5707274903523
14.	ARIMA (0, 1, 1) x (0, 2, 2,12)	639.9583178340788
15.	**ARIMA (0, 1, 2) x (0, 2, 2, 12)**	**637.9293345701016**

Table 4.13

The best model obtained here is ARIMA (1, 1, 2) x (0, 2, 2, 12) with AIC value 637.929334.

4.14 DIAGNOSTIC CHECKING

Diagnostics checking is necessary for ensuring the reliability, validity, and effectiveness of statistical models. It enables to select the right model, estimate parameters correctly and also improve prediction accuracy.

Diagnostic plot is given below.

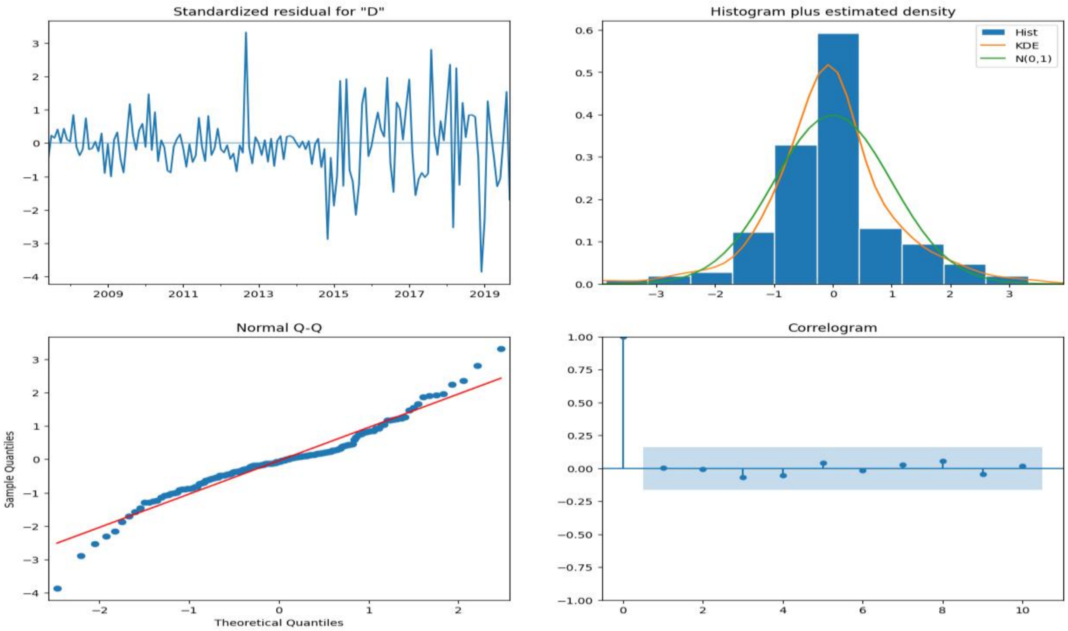

Fig 4.14(a)

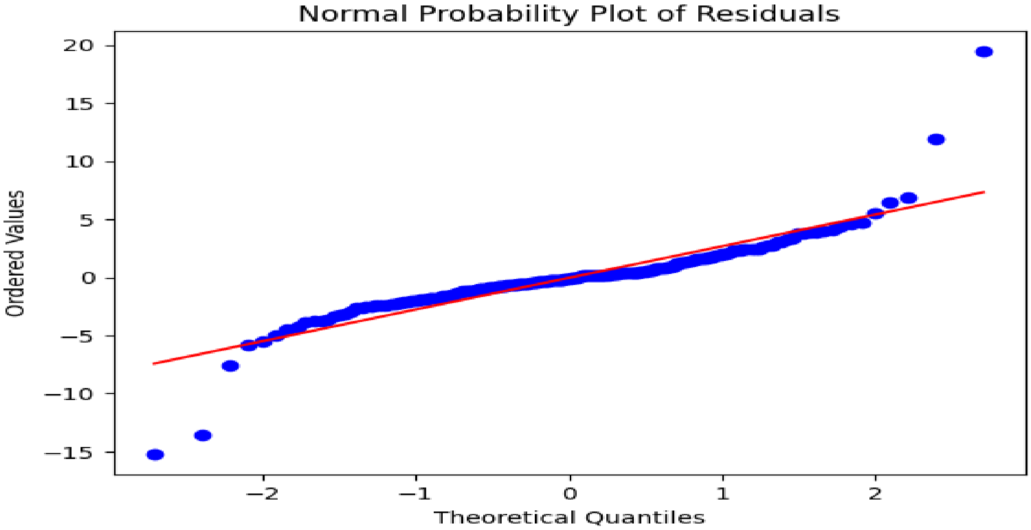

Fig 4.14(b)

From the Q-Q plot and P-P plot, most of the residuals are located on the straight line. Therefore, the standard residual of best fitted model is said to be normal.

Also, here given a plot of residuals vs fitted values:

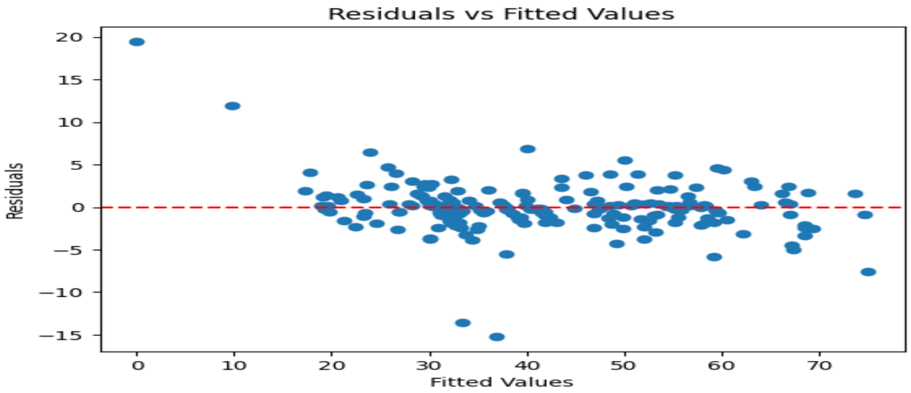

Fig 4.14(c)

4.15 IN SAMPLE FORECAST

In sample forecast is the prediction generated by the best fitted model using data points within the range it was trained on. The table 4.15 given below illustrates the plot of actual and in sample forecasted values of diesel price.

DATE	***DIESEL PRICE/LIT***	**PREDICTED PRICE**
2022-01-01	86.67	80.04313
2022-02-01	86.67	80.4566
2022-03-01	86.67	82.1036
2022-04-01	93.07	90.6684
2022-05-01	96.67	95.2049
2022-06-01	96.72	95.1063
2022-07-01	89.62	90.854
2022-08-01	89.62	90.9009
2022-09-01	89.62	90.837
2022-10-01	89.62	92.5058
2022-11-01	89.62	92.40
2022-12-01	89.62	92.2879
2023-01-01	89.62	90.8096
2023-02-01	89.62	92.42
2023-03-01	89.62	92.4708

2023-04-01	89.62	90.4355
2023-05-01	89.62	90.9717
2023-06-01	89.62	92.4727
2023-07-01	89.62	92.4208
2023-08-01	89.62	92.467
2023-09-01	89.62	92.42
2023-10-01	89.62	92.4719
2023-11-01	89.62	92.4664
2023-12-01	89.62	92.4541

Table 4.15

4.15.1 Plot of actual values vs predicted diesel price

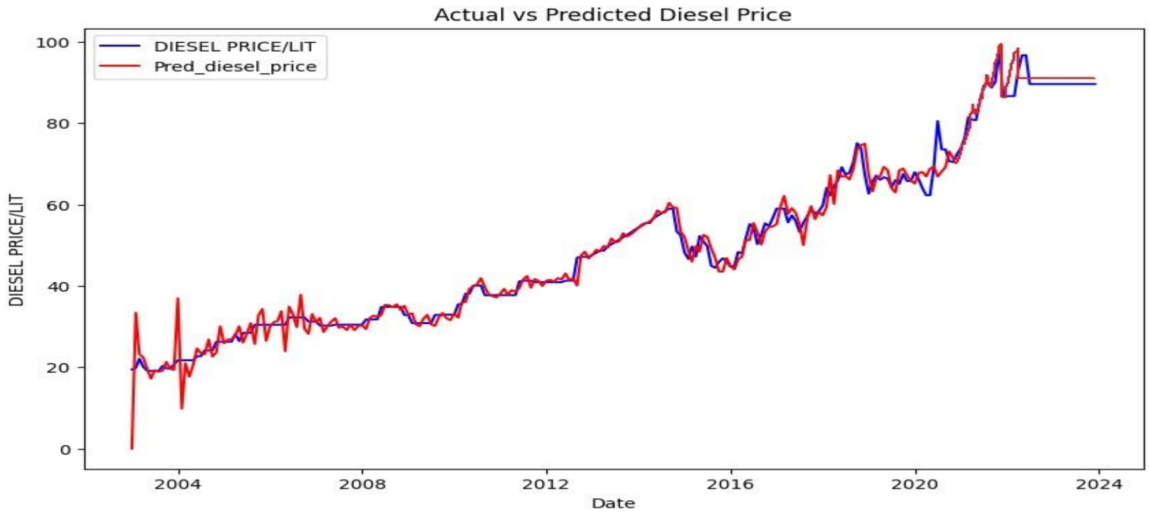

Fig 4.15.1

4.16 FORECASTING

The forecast values of diesel prices during January 2024 to December 2025 is given in the following table.

DATE	***FORCASTED DIESEL PRICE/LIT***
2024-01-01	92.575
2024-02-01	92.9899
2024-03-01	95.637
2024-04-01	98.201
2024-05-01	95.7379
2024-06-01	98.6390
2024-07-01	100.3871
2024-08-01	96.43
2024-09-01	96.369
2024-10-01	98.238
2024-11-01	98.332
2024-12-01	98.2204
2025-01-01	103.3422
2025-02-01	106.75
2025-03-01	105.4033
2025-04-01	108.968
2025-05-01	108.5042
2025-06-01	113.405
2025-07-01	110.1534
2025-08-01	116.2002
2025-09-01	114.135
2025-10-01	118.0044
2025-11-01	120.0990
2025-12-01	123.986

Table 4.16

The graphical representation of the forecast values of petrol price is shown in the below fig.4.16.

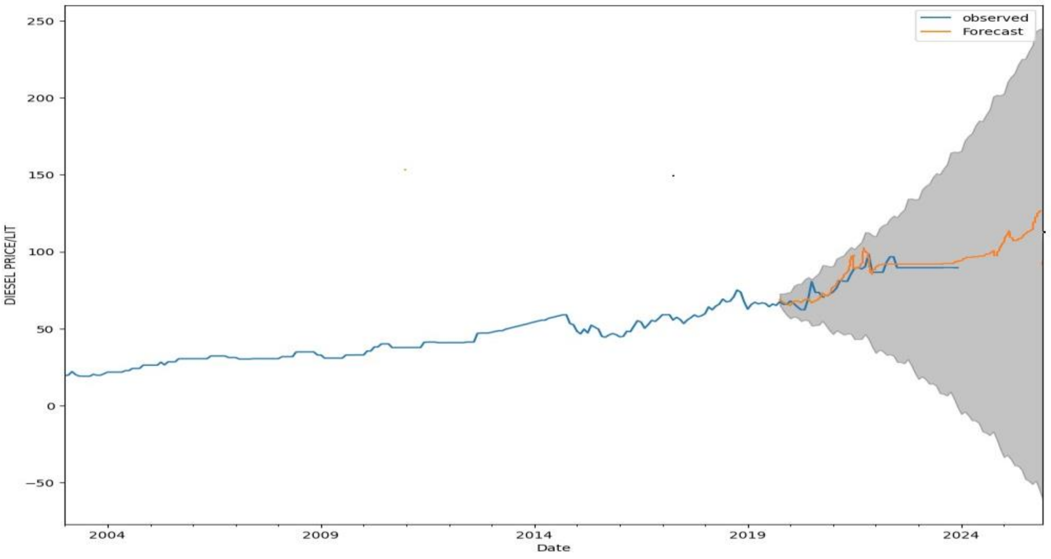

Fig 4.16

This forecasted plot of diesel prices in New Delhi shows an upward trend.

4.17 LINEAR REGRESSION ANALYSIS OF PETROL PRICE

The primary step of Linear Regression analysis is to visualize the given dataset. The visualization of petrol price in New Delhi is given in the figure below.

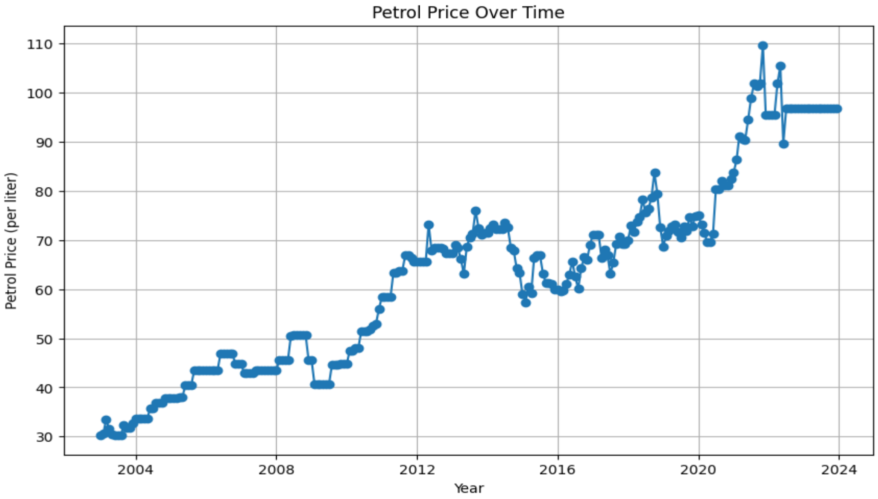

Fig 4.17

4.18 In sample forecast of petrol price using Linear Regression

Here exploratory data analysis has been conducted to check the null values, and to analyze the relationship between different variable.

The predicted values of petrol prices from 2016 to 2023 is given in the following table.

DATE	PETROL PRICE/LIT	PREDICTED PRICE
2016-01-01	59.99	57.84891
2016-04-01	61.13	60.78203
2016-07-01	62.51	66.35497
2016-10-01	66.45	67.33268
2017-05-01	68.09	69.08367
2017-08-01	65.4	67.51044
2017-11-01	69.14	69.42142
2018-03-01	71.57	73.43891
2018-04-01	73.73	75.50988
2018-07-01	75.55	77.99859
2019-05-01	73.13	77.40308
2019-08-01	72.8	76.77201
2020-02-01	73.19	76.96755
2020-10-01	81.06	80.88727
2021-03-01	91.17	90.52215
2021-04-01	90.56	89.98885
2021-11-01	109.69	105.5877
2022-08-01	96.72	97.76608
2023-03-01	96.72	97.76608

Table 4.18

14.18.1 Plot of actual values vs predicted petrol price

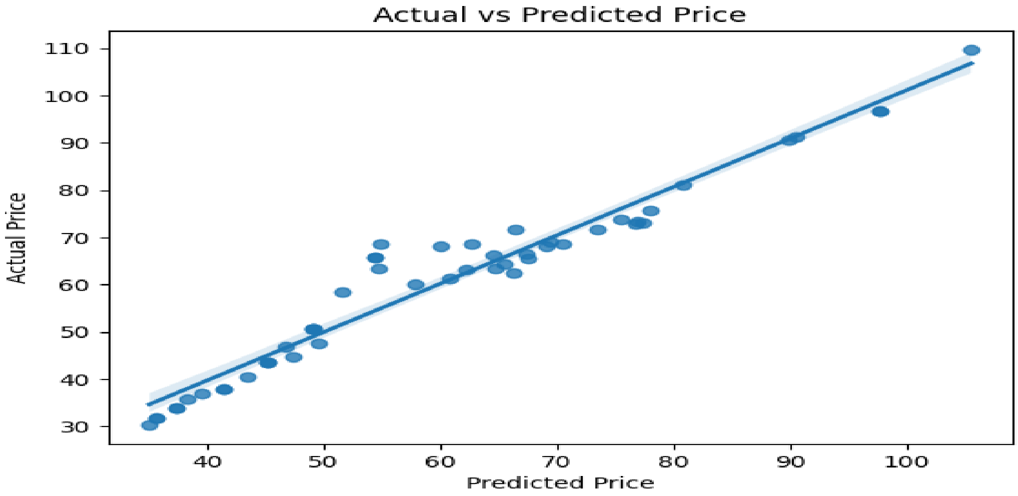

Fig 4.18.1

4.19 LINEAR REGRESSION ANALYSIS OF DIESEL PRICE

The primary step of Linear Regression analysis is to visualize the given dataset. The visualization of diesel price in New Delhi is given in the figure below.

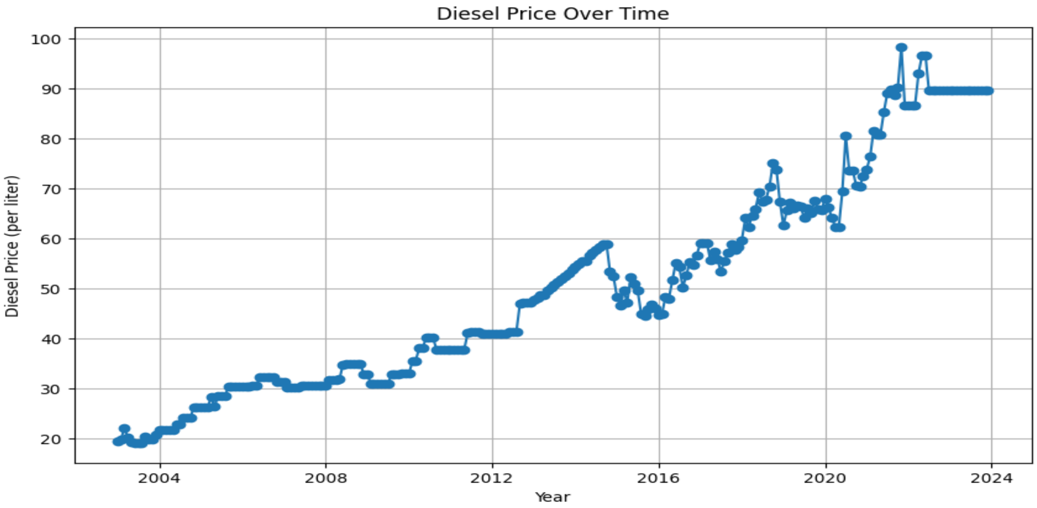

Fig 4.19

4.20 In sample forecast of diesel price using Linear Regression

Here exploratory data analysis has been conducted to check the null values, and to analyze the relationship between different variable.

The predicted values of diesel prices from 2016 to 2023 is given in the following table.

DATE	DIESEL PRICE/LIT	PREDICTED PRICE
2016-01-01	44.71	47.3182
2016-04-01	48.01	48.52751
2016-07-01	54.28	49.99142
2016-10-01	55.38	54.17098
2017-05-01	57.35	55.91069
2017-08-01	55.58	53.05714
2017-11-01	57.73	57.02454
2018-03-01	62.25	59.60228
2018-04-01	64.58	61.89362
2018-07-01	67.38	63.82428
2019-05-01	66.71	61.25714
2019-08-01	66	60.90707
2020-02-01	66.22	61.32078
2020-10-01	70.63	69.6693
2021-03-01	81.47	80.39401
2021-04-01	80.87	79.74692
2021-11-01	98.42	100.0401
2022-08-01	89.62	86.28146
2023-03-01	89.62	86.28146

Table 4.20

14.20.1 Plot of actual values vs predicted petrol price

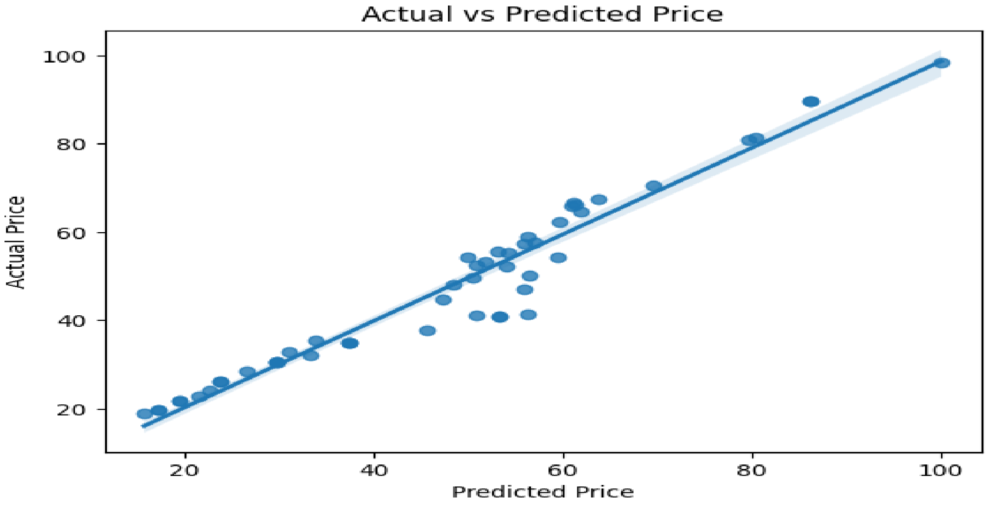

Fig 14.20.1

4.21 COMPARING MSE & RMSE VALUES

To determine the best model, it is important to compare the performance metrics and consider the context of the problem being addressed. For the particular case, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated for both SARIMA model and Linear Regression, providing insights into their predictive accuracy and effectiveness.

MSE and RMSE values of SARIMA model and Linear Regression model is given in the following table 4.21.

TOOL USED	**SEASONAL ARIMA**		**LINEAR REGRESSION**	
	PETROL	DIESEL	PETROL	DIESEL
MSE	65.2	41.38	21.27048	25.20367
RMSE	8.07459	6.43297	4.61199	5.02033

Table 4.21

The SARIMA model has higher MSE and RMSE values compared to Linear regression model. Based on the provided metrics alone, the Linear regression gives the best model than SARIMA in terms of MSE and RMSE values.

CHAPTER 5

CONCLUSION

The study was undertaken to analyze whether Time series model or Linear regression model is best forecast. The monthly data of fuel price in Delhi from 2003 to 2023 are used for the study. The SARIMA and Linear Regression models forecast is then compared. The MSE and RMSE were used to compare the forecast. It is done by both models. Also, by this comparison Linear Regression gives the best forecast for fuel price data chosen.

In this project, the forecasted fuel prices for Delhi shows the importance of proactive measures. It is to manage potential fluctuations and ensure stability in fuel costs. In New Delhi, the fuel prices are stable since 2022, July. However, the future prediction indicates an upward trend in fuel cost for both models. Policymakers, industry stakeholders, and consumers should collaborate to address challenges related to fuel pricing, ensuring a sustainable and affordable energy future for the city. And finally it concludes that it is still possible to accurately forecast prices in a stable market.

REFERENCES

1. Anusree, M., & Sarika, S. G. (2022). A Case Study on Rising Petrol and Diesel Prices in Fourteen Cities of India. *Journal of Pharmaceutical Negative Results*, 417-420.

2. Bhuvandas, D., & Gundimeda, H. (2020). Welfare impacts of transport fuel price changes on Indian households: An application of LA-AIDS model. *Energy Policy*, *144*, 111583.

3. Gawande, A. P., & Kaware, J. P. (2013). Fuel adulteration consequences in India: a review. *Sci. Rev. Chem. Commun*, *3*(3), 161-171.

4. Ghosh, S. (2006). Future demand of petroleum products in India. *Energy Policy*, *34*(15), 2032-2037.

5. Lahari, M. C., Ravi, D. H., & Bharathi, R. (2018, September). Fuel price prediction using RNN. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1510-1514). IEEE.

6. Rao, R. D., & Parikh, J. K. (1996). Forecast and analysis of demand for petroleum products in India. *Energy policy*, *24*(6), 583-592.

7. Sasikumar, R., & Abdullah, A. S. (2017). Vector autoregressive approach for impact of oil India stock price on fuel price in India. *Communications in Statistics: Case Studies, Data Analysis and Applications*, *3*(1-2), 41-47.

8. Vashist, D., & Ahmad, M. (2014). STATISTICAL ANALYSIS OF DIESEL ENGINE PERFORMANCE FOR CASTOR AND JATROPHA BIODIESEL-BLENDED FUEL. *International Journal of Automotive & Mechanical Engineering*, *10*.

9. Vempatapu, B. P., & Kanaujia, P. K. (2017). Monitoring petroleum fuel adulteration: A review of analytical methods. *TrAC Trends in Analytical Chemistry*, *92*, 1-11.

10. Yeh, S. (2007). An empirical analysis on the adoption of alternative fuel vehicles: The case of natural gas vehicles. *Energy policy*, *35*(11), 5865-5875.