Project Report

On

A SATISTICAL STUDY ON CRUDE OIL PRICE IN INDIA

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

SANDHRA MARIAM JOHN

(Register No. SM21AS018)

(2022-2023)

Under the Supervision of

MRS. ANU MARY JOHN

DEPARTMENT OF STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

APRIL 2023

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **A SATISTICAL STUDY ON CRUDE OIL PRICE IN INDIA** is a bonafide record of the work done by Ms. **SANDHRA MARIAM JOHN** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Statistics** 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:
Place: Ernakulam

Mrs. Anu Mary John
Assistant Professor,
Department of Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

Mrs. Betty Joseph
HOD & Associate Professor,
Department of Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1:.............................

2: ....

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **MRS. ANU MARY JOHN**, Assistant Professor, Department of 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. **SANDHRA MARIAM JOHN**

Date: **SM21AS018**

ACKNOWLEDGEMENTS

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

I am very grateful to my project guide **MRS. ANU MARY JOHN** for the immense help during the period of work.

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

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

Ernakulam. SANDHRA MARIAM JOHN
Date: **SM21AS018**

Contents

Chapter 1

Introduction

Crude oil is a naturally occurring petroleum product composed of hydrocarbon deposits and other organic materials. Composition between 50% and 97% of oil is hydrocarbons. Between 6% and 10% of it is nitrogen, oxygen and sulphur. Less than 1% is metals. It is a nonrenewable resource and it's a limited resource.

Crude oil is a global commodity that trades in markets around the world, both as spot oil and via derivatives contracts. Many economists view crude oil as the single most important commodity in the world as it is currently the primary source of energy production. Crude oil is typically obtained through drilling. After its extraction, crude oil is refined and processed into a variety of forms, such as gasoline, kerosene and asphalt.

Crude oil is also called "black gold". The world 'Black' because of its appearance when it comes out of the ground and the word 'Gold' for its rarity. The top three oil producing countries are United States, Russia and Saudi Arabia. In world oil production India is the 23^{rd} production. The three largest producers of crude oil in India are Rajasthan(23.7%), Gujarat(12.5%) and Assam(12.1%).

1.1 Uses of Crude oil

Crude oil is the base for lots of products. These include transportation fuels such as gasoline, diesel and jet fuel. They also include fuel oils used for heating and electricity generation. It's used in carbon fiber in aircraft, PVC pipes, and cosmetics. Today, the world's economy is largely dependent on fossil fuels such as crude oil, and the demand for these resources often sparks political unrest.

The cost of a barrel of crude oil is determined by global supply and demand. If the worldwide demand for crude oil is high, and supply low, oil prices will be high. Oil prices are controlled by traders who bid on oil future contracts in the commodities market. That's why oil prices changes daily. It all depends on how trading went that day.

1.2 Three factors traders use to determine oil prices

There are three main factors that commodities traders look at when developing the bids that create oil prices. First is the current supply in terms of output. Since 1973, OPEC has a limited supply of 61 percent of the world's oil exports. But U.S. shale oil production doubled between 2011 and 2014. That created an oil glut. Traders bid the price down to $45 per barrel in 2014. Prices fell again in December 2015 to $36.87 a barrel. OPEC would normally cut supply to keep oil at its target of $70 a barrel. OPEC bet that the shale oil producers would go out of business. This would allow it to keep its dominant market share. The oil price forecast has shown such volatility in prices because of the changes in oil supply, dollar value, OPEC's actions and global demand.

Second is access to future supply. That depends on oil reserves. It includes what's available in U.S. refineries as well as in the Strategic Petroleum Reserves. These reserves can be accessed very easily to increase oil supply if prices get too high.

Third is oil demand, particularly from the United States. These estimates are provided monthly by the Energy Information Agency. Demand rises during the summer vacation driving season. To predict demand, forecasts for travel are used to determine potential gasoline use. During the winter, weather forecasts are used to determine potential home heating oil use.

Chapter 2

Literature Review

(1) A systematic review in crude oil markets: Embarking on the prices - Yuhang Zueng & Ziqing.

The role of crude oil in economic activities is significant, encompassing both commodity attributes and financial characteristics. After conducting a comprehensive literature review on crude oil prices, the following observations were made. Firstly, researchers continue to focus on forecasting and managing the risks associated with crude oil prices, but the uncertainty of economic activity has exacerbated fluctuations in crude oil prices. Secondly, the main drivers of crude oil price movements are attributed to factors from the supply and demand side, while investor sentiment has gradually become an important consideration in forecasting the expected level of crude oil prices. Thirdly, shocks to crude oil prices impact economic activities and financial stability, with many studies demonstrating asymmetric effects. However, due to changes in the external environment, more complex and non-linear time-varying features are emerging. Moreover, the advent of text mining and artificial intelligence technology provides new and effective methods to predict crude oil price trends and conduct risk measurement in the crude oil market.

(2) A study on impact of crude oil price in the price of gold - Monish.

P & Dhanabhakyam. M.

The word "oil" always generates attention in the news, but the average person has limited knowledge about this mysterious "mineral oil", even though most countries bear the cost of exploring or importing it. Despite the high associated risks, the current international market price of crude oil makes oil exploration economically feasible in India. Some argue that investing in gold is a response to concerns about rising inflation. Therefore, it is interesting to investigate the relationship between the prices of fuel and metals, particularly crude oil and gold.

This study aims to examine the impact of changes in oil prices on gold prices and analyze various factors that influence changes in oil prices. To analyze this relationship, monthly prices of both gold and crude oil from 2013 to 2017 are chosen for a period of five years. Regression analysis is utilised as a tool for analysis.

(3) An analytical view of crude oil prices and its impact on Indian economy - K. Soundarapandiyan & M. Ganesh.

Crude oil plays a crucial role in taming inflation and sustaining rapid economic growth in the volatile, uncertain, complex and ambiguous(VUCA) market. India is the fourth-largest consumer of crude oil globally, importing 100 million tons every year, accounting for 37 percent of the total imports. The price of crude oil is a determining factor among various commodities, as any rise or fall in price directly impacts the prices of other commodities and society as a whole.

(4) An analysis of crude oil prices in the last decade (2011-2020): with deep approach - Abhikasu Sen & Tapan Kumar Datta.

Crude oil is a highly significant commodity worldwide. In this study, we analyze the effects of crude oil inventories on crude oil prices from 2011 to 2020. We aim to determine how variations in crude oil prices respond to inventory announcements, as well as the relationship between crude oil variations and several other financial instruments.

(5) A study on impact of crude oil price fluctuation on Indian economy - Kali Charan Hadak & Pallabi Mukherjee.

The economic growth of any country is heavily dependent on crude oil prices. India, for example, imports over 70% of its crude oil requirements. In this study, we consider variables such as Gross Domestic Product(GDP), Crude oil Price(COP) and Wholesale Price Index(WPI). Our aim is to investigate the impact of fluctuations in oil prices on the growth of the Indian economy using time series data from 2000 to 2014. To analyse the data, we utilize multiple linear regression models.

(6) A study on the impact of crude oil prices on the Indian stock market - S. R. Raghunand & Smita Kavatekar.

The aim of this research is to add to the existing literature on the relationship between energy prices and stock markets by examining the impact of oil price changes on the Indian stock market. Various statistical tools such as regression analysis and ANOVA have been used to establish a correlation between Brent Crude Oil prices and the BSE Sensex and NSE Sensex stock market indices for the past five years (2013-17) in a six-monthly time series format.

Chapter 3

Objectives and Data Description

The main objectives of the study is as follows

- **To study the past behaviour of crude oil price in India using the time series analysis and to develop time series model.**
- **To forecast crude oil price in India from 2022-2024.**

Data consists of monthly data of crude oil(petroleum) prices in Indian rupee per barrel from March 1983 to March 2021.

3.1 Data Source

The data collected for the analysis of this project work is purely secondary data. The data crude oil price were collected from Kaggle.

Chapter 4

Methodology

Here we employed the statistical techniques to analyse the data are Time series analysis.

4.1 Time Series

A time series is a set of observations X_t, each one being recorded at a specific time t, at uniform time intervals. Time series models are designed to capture various characteristics of given data. Time series analysis has a wide range of applications. It is widely used in many disciplines in the science, humanities, engineering etc. The main objectives of investigating a time series data are

- Description of the data: To describe data using the summary statistics or by graphical methods. A time plot of the data is the particularly valuable.
- Modelling: To find suitable statistical model to describe the data generating process.
- Forecasting: To estimate the future values of the series.
- Control: Good forecast enables the analyst to take actions so as to control a given process.

4.2 Components of Time Series

The four components of time series are:

- **Trend(T_t)-These are the changes that occurred as a result of general tendency of the data to increase or decrease over a long period of time.**
- **Seasonal effects(S_t)-Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or during the same quarter every year.**
- **Cycle(C_t)-Any pattern showing an up and down movement around a given trend is identified as cycles.**
- **Irregular Variation(I_t)-This component is unpredictable. Every time series has some unpredictable component that makes it a random variable.**

4.3 Stationary

The joint probability of a series doesn't change over time that is mean and variance remain constant over time.

4.4 Weakly Stationary

A stochastic random variable(X_t) is said to be weakly stationary, if it satisfies the following properties:

- **The mean $E(X_t)$ is the same for all t.**
- **The variance of X_t is the same for all t.**
- **The covariance (and also correlation) between X_t and X_{t-h} is the same for all t, h.**

4.5 Strictly Stationary

A stochastic random variable is said to be strictly stationary, if the joint distribution of $X(t_1), X(t_2), \ldots, X(t_n)$ is identically distributed to the joint distribution of $X(t_1+h), X(t_2+h), \ldots, X(t_n+h)$ for all $t_1, t_2, \ldots, t_n$.

4.6 Auto Regressive Process(AR)

Autoregressive model of order p is defined by,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t \tag{4.1}$$

where ϕ_1, ϕ_2, ..., ϕ_p are weights called autoregressive parameters and a_t is white noise. In compact form, it is written as,

$$\Phi(B)X_t = a_t \tag{4.2}$$

where B is called backward shift operator.

4.7 Moving Average(MA) processes

In this series, the error of the past terms of the data set is written as infinite weighted linear sum. Its equation is,

$$X_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} \tag{4.3}$$

where a_t is white noise; θ_1, θ_2, . . . , θ_q is moving average parameters and q is the order of MA. Its compact form is,

$$X_t = \theta(B)a_t \tag{4.4}$$

where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$

4.8 Mixed autoregressive-moving average processes(ARMA)

Autoregressive moving average process of order (p, q) is the combination of autoregressive process of order (p) and moving average process of order (q).

The model can be written as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}$$

Or

$$\Phi(B)X_t = \Theta(B)a_t$$

is called the mixed autoregressive-moving average process of order (p, q), is abbreviated as ARMA(p, q).

4.9 Auto Regressive Integrated Moving Average (ARIMA) model

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series. They are applied in some cases where data shows evidence of non-stationarity, where an initial differing step (corresponding to the integrated part of the model) can be applied to remove the non stationarity. This model is generally referred to as ARIMA(p, d, q) model, where p, d, q are integers greater than or equal to zero.

4.10 Auto Correlation Function (ACF)

Auto correlation is the correlation between observations of a variable taken at different time point. Auto correlation plots are a commonly used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. ACF Identify appropriate time series model if data are not random. The $k-$th order ACF is defined as:

$$\rho(k) = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} = \frac{Cov(X_t, X_{t-k})}{Var(X_t)} = \frac{\gamma(k)}{\gamma(0)}.$$

4.11 Partial Auto Correlation Function (PACF)

This measure of correlation is used to identify the extent of relationship between current values of variable with earlier values of that same variable (values for various time lags) while holding the effects of all

other time lags constant. For a time series, the partial autocorrelation between x_t and x_{t-h} is defined as the conditional correlation between x_t and x_{t-h}, conditional on $x_{t-h+1}, \ldots, x_{t-1}$, the set of observations that come between the time points t and $t-h$. The first order partial autocorrelation will be defined to equal the 1^{st} order autocorrelation.

4.12 Augmented Dickey-Fuller Test

This test is a common statistical test used whether a given time series is stationary or not. It is an augmented version of the Dickey-Fuller test for a larger and more complicated set of time series models.

The hypothesis for testing is:

H0: Given series is not stationary, against

H1: Given series is stationary.

If the test statistic is greater than the critical value, we reject the null hypothesis. Otherwise we accept the null hypothesis.

4.13 Forecasting

Forecasting is defined as an activity to calculate or predict some future event or condition usually as result of rational study or analysis of relevant data. Once a time series model has been obtained, the minimum mean square error forecasts are easily calculated from the difference equation of the model. We also find the upper and lower confidence limit for the forecasts. It is to be emphasized that for practical computation of the forecasts, this approach via the difference equation is the simplest and most elegant.

4.14 Residual Analysis

When a model has been fitted to a time series it is advisable to check that the model really provide an adequate description of the data. This is usually done by looking at the residuals. For a good model, residuals

are stationary and uncorrelated and a model validation usually consists of plotting residuals in various ways. For many time series models, the residuals are equal to the difference between the observations and the corresponding fitted values:

Residual= Observation-Fitted value

A good forecasting method will yield residuals with the following properties:

- **The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.**
- **The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.**

4.15 Model Adequacy

Our selected model may appear to be best among those models considered, it is also necessary to do diagnostic checking to verify that the model is adequate. To check the model is adequate, the following procedure is adopted.

4.16 Akaike information criterion (AIC)

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

- **The number of independent variables used to build the model.**
- **The maximum likelihood estimate of the model (how well the model reproduces the data).**

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. The formula for AIC is:

$$AIC = 2K - 2\ln(L)$$

K is the number of independent variables used and L is the log-likelihood estimate.

4.17 Bayesian information criterion (BIC)

The Bayesian Information Criterion (BIC) or Schwartz criterion was developed by Schwartz (1978) and is a criterion for model selection among a finite set of model; the model with the lowest value BIC is preferred.

4.18 Ljung-Box Test

This test is to check whether the residuals from a time series resembles white noise. The Ljung Box Test can be defined as follows,

H0: **the data are independently distributed, i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process,**

H1: **the data are not independently distributed i.e. they exhibits serial correlation.**

The test statistic is:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

when n is the sample size , $\hat{\rho}_k^2$ is the autocorrelation at lag k, and h is the number of lags being tested. If the $p-$value associated with the Q statistics is small ($p-$value; 0.05), the model considered inadequate. If the model is inadequate, the model has to be modified and the analysis continued until the level of satisfactory for model achieved. Once an

adequate model has been selected, it was used to forecast for one period or several periods into the future.

4.19 Forecasting Method

First we prepare a time series plot for the crude oil price data. Then we find the ACF, PACF, trend and seasonal component of the data. To model a time series, the series has to be stationary. So we check stationarity using Augmented Dickey-Fuller test. If data is not stationary, to convert a non-stationary process to a stationary process, we apply the differencing method. Differencing a time series means finding the differences between consecutive values of a time series data. After making the data stationary, we plot ACF and PACF to ensure the stationarity.

To model the time series we find the value of the model coefficients which provide the best fit to the data. With the ACF and PACF we can determine the order of the model. Once we have determined the parameters we estimate the accuracy of the ARIMA model on a training dataset and then use the fitted model to forecast the values of the test dataset using a forecasting function. In the end, we cross check whether our forecasted values are in line with the actual values.

Now we obtain a point forecast as well as a lower bound and upper bound of forecasting for both 80% **and** 95% **confidence intervals by using ARIMA(1; 1; 2) model. Further we plot residuals for checking correlation and test the randomness by Ljung-Box test. If the values are uncorrelated the forecasts are good.**

Chapter 5

Data Analysis

5.1 Time Plot

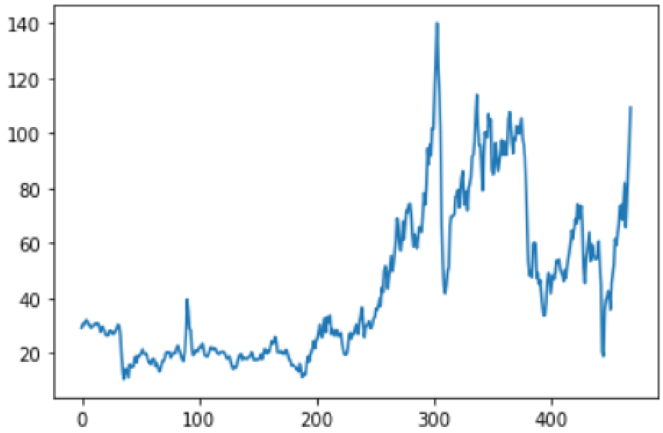

Figure 5.1: Time Series Plot of crude oil price

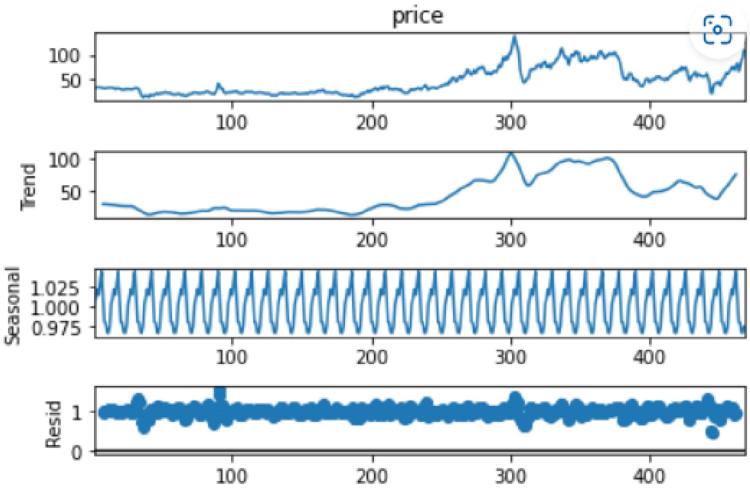

Figure 5.2: Decomposition of time series

5.2 ACF and PACF Plot

Figure 5.3 and 5.4 are ACF and PACF plot of the time series. It shows that the data is not stationary. Since the ACF plot does not approaches to zero monotonically.

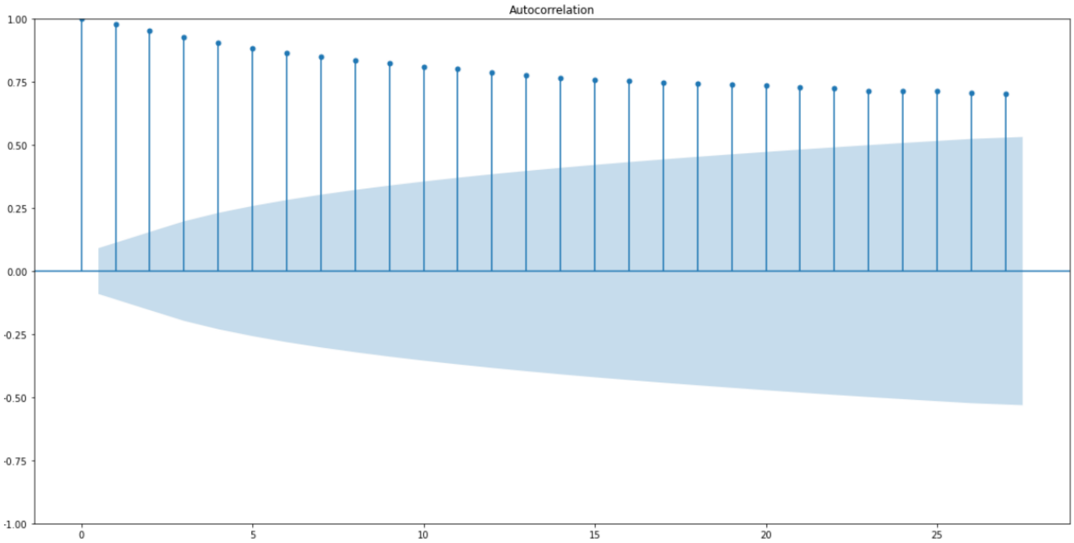

Figure 5.3: ACF Plot

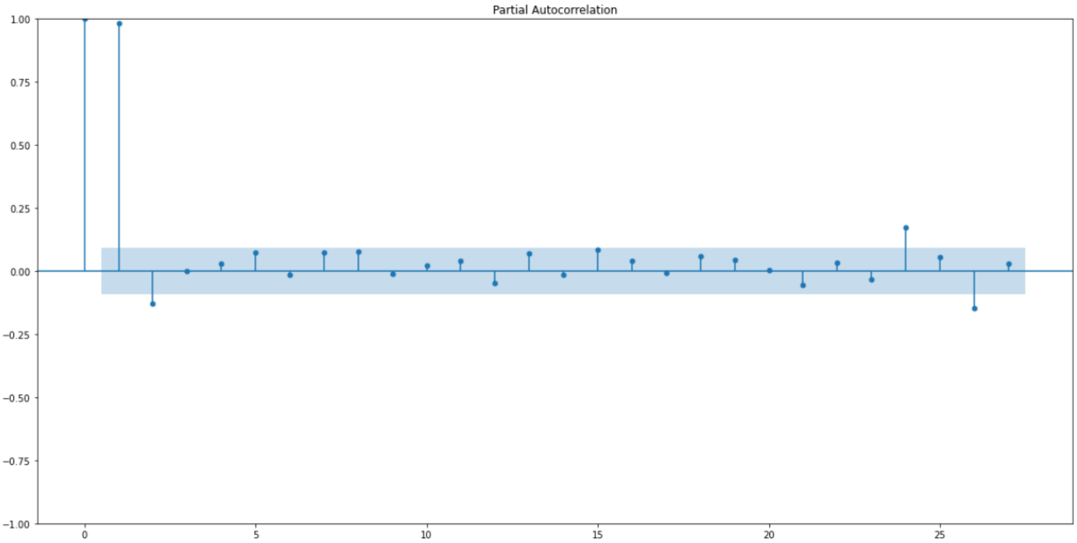

Figure 5.4: PACF Plot

5.3 Stationarity

Checking stationarity of the data using Augmented- Dickey Fuller test.

Here we test the hypothesis

H1: The crude oil price data is non-stationary. Against

H2: The crude oil price data is stationary.

Augmented Dickey-Fuller Test:

Dickey-Fuller	-1.782
Lag order	1
$p-$value	0.3890

Since $p-$value is greater than 0.05, we accept the null hypothesis. That is, the crude oil price data is non-stationary.

We can make it stationary by ordinary differencing. After differencing of order 1, the Augmented Dickey-Fuller Test of the differenced series is as follows:-

Dickey-Fuller	-3.681
Lag order	13
$p-$value	0.004

Since $p-$value is less than 0.05, we reject the null hypothesis. That is, the crude oil price data is stationary.

The time series plot of differenced series is,

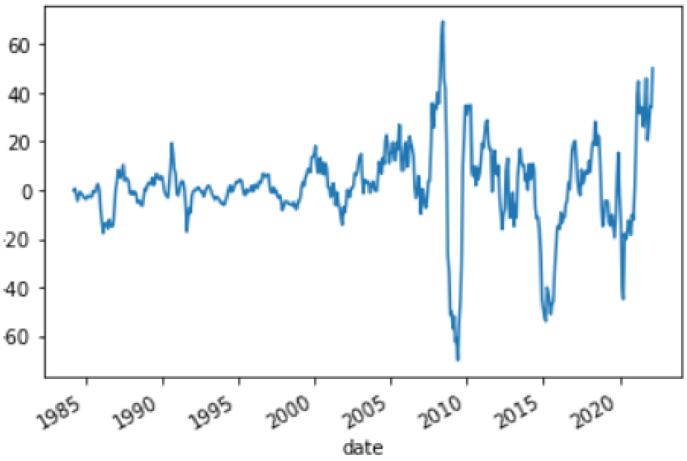

Figure 5.5: Time Series Plot of Differenced Series

ACF and PACF plots of differenced series are given by the figure 5.6 and 5.7.

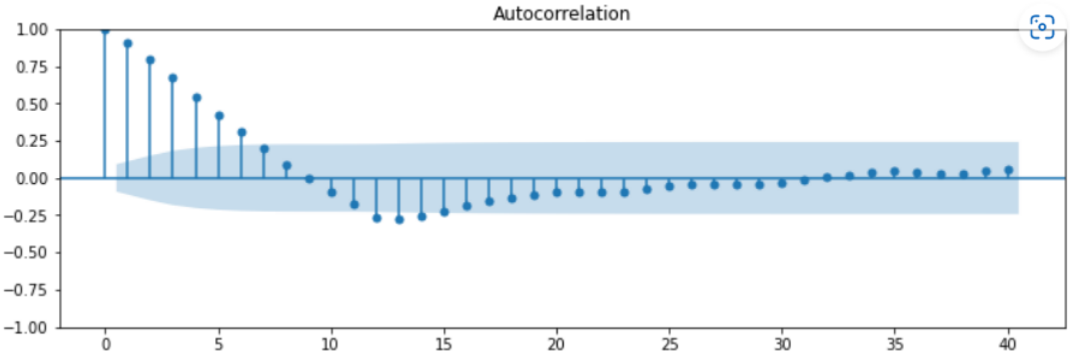

Figure 5.6: ACF Plot of Differenced Series

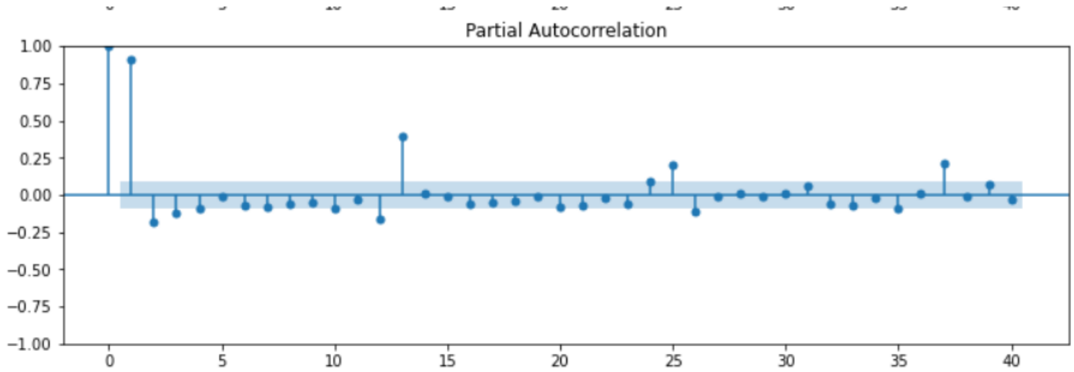

Figure 5.7: PACF Plot of Differenced Series

5.4 Modelling

Among a class of ARIMA models, the model with minimum AIC=2800.120 and BIC=2816.706 is ARIMA(2; 1; 1). The model coefficients are given in the following table.

	ar1	ar2	ma1
Estimates	1.1662	-0.2454	-0.9474
s.e.	0.044	0.032	0.043

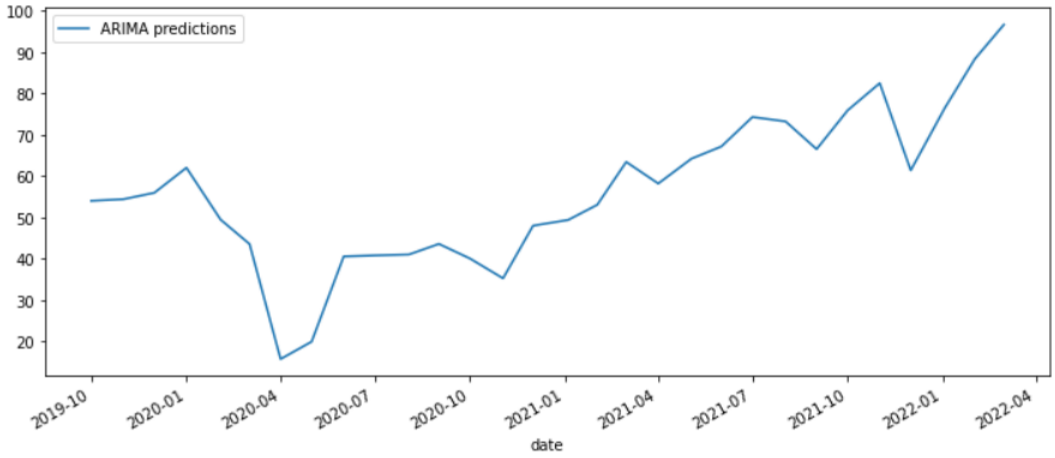

Figure 5.8: ARIMA Prediction

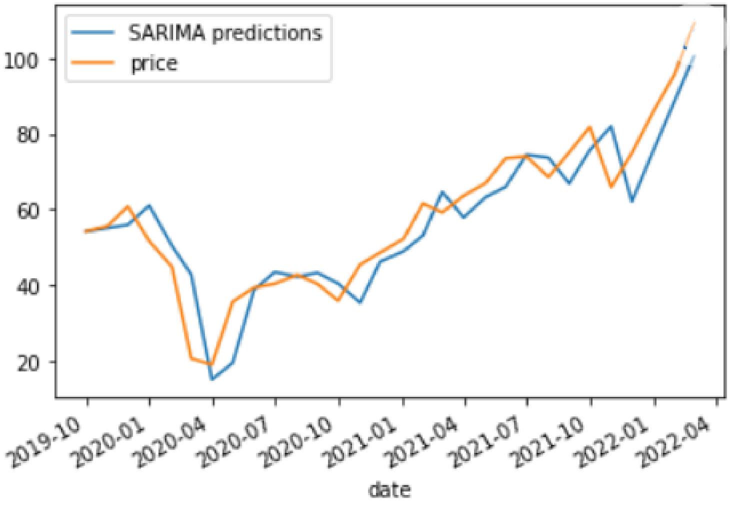

Figure 5.9: SARIMA Prediction

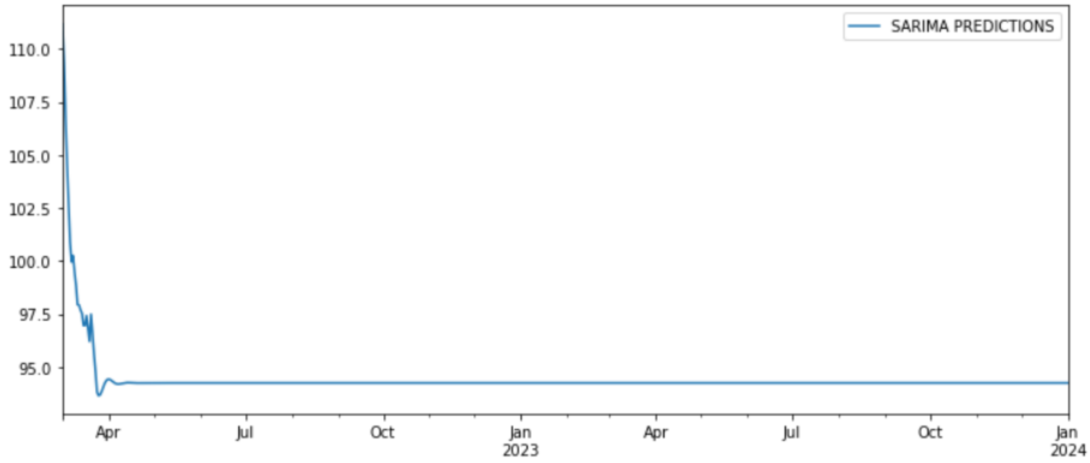

Figure 5.10: SARIMA Forecast

Chapter 6

Conclusion

In this study, we used the data of crude oil price from the year1983 to 2021. A time series analysis was conducted to forecast the crude oil price during the years 2022, 2023 and 2024. When it comes to the past behaviour of the data we can see that a trend pattern of the price but there is no seasonality in the data. After making the data stationary, we fitted the model using ARIMA for forecasting data with future intervals. We found that the ARIMA model for the monthly prices of the crude oil price is ARIMA(2; 1; 1).

Finally, we forecasted the monthly crude oil price for two years. We compared the actual data with forecasted data for first 6 months of the year 2021 and we got that the actual values lies with in the forecasted confidence intervals. The entire practical analysis was done through Python software.

REFERENCES

[1] Jonathan D. Cryer, Kung-Sik Chan-Time Series Analysis with Applications in R, Second Edition

[2] Spyros Makridakis, Steven C. Wheelwright, Rob J. Hyndman - Forecasting methods and applications, Third edition, John Wiley [2015]

[3] Vijay K Rohatgi, A. K. Md. Ehsanes Saleh - An introduction to probability and Statistics, Second Edition, John Wiley Inc.[2001]

[4] Gaynor, Patricia E and Kirkpatrick Rickey C.- Introduction to Time Series Modeling and Forecasting in Business and Economics, Tata Mc-Graw-Hill International, USA[1994]

[5] Gebhard Kirchg$\ddot{a}$ssner, J$\ddot{u}$rgen Wolters - Introduction to Modern Time Series Analysis, Springer

[6] https://www.indexmundi.com

[7] https://online.stat.psu.edu/stat510/book/export/html/665

[8] https://www.pluralsight.com

[9] https://www.statmethods.net