Deccan Education Society's
Fergusson College (Autonomous), Pune
Department of Statistics

STS3609: Statistics Practical 111

CERTIFICATE

This is to certify that Mr. Mahesh Madan Roll No. 214527, has satisfactorily completed the project work entitled "Rudis Oleum" towards the partial fulfillment of B.Sc. (Statistics), Semester VI during the academic year

2021-2022.

Place: Pune

Date:21/ 05 / 2022

Dr. Subhash S. Shende

Mr. Shital K. Gadekar

1. (Internal Guide)

2. (External Guide, if applicable)

External Examiner

Internal Examiner

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the various individuals for supporting us throughout our project. Firstly, we would like to thank our project guide, Professor Shital K. Gadekar, for his enthusiasm, patience, helpful comments and unceasing ideas that have helped us at all times during our research and writing of this project. Without his support and guidance, this project would not have been possible.

We also wish to express our thanks to the Department of Statistics, Fergusson College, and the Head of the Department, Dr Subhash S. Shende, for providing us with an opportunity to work on this project.

We would also like to take this moment to extend our gratitude to Fergusson College and the principal, Dr Ravindrasinh Pardeshi, for providing us with the infrastructure and facilities utilised to complete this project.

We are also very grateful to the people who have helped us, directly or indirectly. We want to thank our parents, friends and family members for their constant support and encouragement throughout the project.

INDEX

	Title	Page. No
i.	*Abstract*	*1*
ii.	*Objectives*	*1*
iii.	*Data Collection*	*1*
iv.	*Softwares Used*	*2*
v.	*Statistical Techniques*	*3*
vi.	*Introduction*	*13*
vii.	*Price of Crude Oil*	*15*
viii.	*Import of Crude Oil*	*35*
ix.	*Refineries of India*	*45*
x.	*Oil Consumption*	*57*
xi.	*Natural Gas Consumption*	*67*
xii.	*Conclusion*	*77*
xiii.	*Bibliography*	*79*
xiv.	*Limitations and Further Scope*	*81*
xv.	*Appendix*	*83*

ABSTRACT

This project aims to analyse various factors in the Indian crude oil sector like the price of a barrel of oil, import, consumption and refinery capacity of crude oil, natural gas consumption and prices of petrol and diesel. It involves fitting time series models and regression models and checking their accuracy. Using these models, we have tried to forecast values to see their future behaviour based on the past several years of data available to us.

OBJECTIVES

- *To analyse crude oil prices of both Brent and WTI market indices which decide the prices for the barrels of crude oil*
- *To explore and give inferences on other factors related to crude oil like import, consumption, refinery capacity, etc.*
- *To fit time series and regression models based on the previous years' data*
- *To forecast future values of the above-analysed components from their past behaviour.*

DATA COLLECTION

To collect the data, we scoured through tons of sites to try and find the most accurate data for our project. The data used for this project has information of countries all around the world. Though the data we have collected dates back to the 1980s and in some cases even more, its accuracy has been verified. Therefore, our analysis is based on reliable numbers due to which the inferences drawn are very relatable to the real world.

SOFTWARES USED

R and R Studio:

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques and is highly extensible. It was developed by Ross Ihaka and Robert Gentleman of New Zealand and has been improving ever since!

R studio is an extension environment of R software. It uses the R language to develop statistical and data science programs. It makes it easier to develop programs, view the charts / outputs and check environment variables easily. It can even run Python codes!!! However, R studio must be used alongside R to function properly. For the entirety of this project, we have used R Studio to create, develop and execute the calculations and computations used in our project.

The RStudio Logo (Source: rstudio.com)

Microsoft Excel:

Microsoft Excel is a spreadsheet program which offers very helpful tools for data manipulation, analysis and visualization. For this project, Excel helped us to sort and filter the data according to our requirements and use it later to analyse and give inferences.

Microsoft Word:

Microsoft Word is a graphical word processing software that allows users to type and save documents. It has many tools to make the documents look attractive and professional. For our project, we have used Microsoft Word for the Project Compilation.

STATISTICAL TECHNIQUES

1. Time Series:

A time series is a collection of data points in successive order over a fixed period, say days, months, years, etc. It has four components, trend, seasonality, cycles and irregularity. Ex, Price of SBI stock, rainfall recorded monthly, etc.

Applications of Time Series (Source: medium.com)

2. Exponential Smoothing:

Exponential smoothing is a time series forecasting method for univariate data. Exponential smoothing forecasting methods are a weighted sum of past observations, but the model explicitly uses an exponentially decreasing weight for past observations. Specifically, past observations are weighted with a *geometrically decreasing* ratio.

Types of Exponential Smoothing: -

There are three main types of exponential smoothing time series forecasting methods.

a. Single Exponential Smoothing:

It requires a single parameter, called alpha (α), also called the smoothing constant or smoothing coefficient. The smoothing constant, α, takes values between 0 and 1. The formula is given by:

$$\text{Current forecast} = \alpha * \text{previous value} + (1-\alpha) * \text{previous forecast value}$$

b. Double Exponential Smoothing:

In addition to the alpha parameter for controlling the smoothing factor for the level, an additional smoothing factor is added to control the decay of the influence of the change in the trend called beta (β). The formula is given by:

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + hb_t \\ \text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}, \end{aligned}$$

c. Triple Exponential Smoothing:

In addition to the alpha and beta smoothing factors, a new parameter is added called gamma (ϒ) that controls the influence on the seasonal component. Triple exponential smoothing is the most advanced variation of exponential smoothing and through configuration, it can also develop double and single exponential smoothing models. Additionally, to ensure that the seasonality is modelled correctly, the number of time steps in a seasonal period (Period) must be specified.

$$\begin{aligned} \hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\ \ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\ s_t &= \gamma \frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m} \end{aligned}$$

3. Stationarity in time series:

When we make a model for forecasting purposes in time series analysis, we require a stationary time series for better prediction. So, the first step to work on modelling is to make a time series stationary. Testing for stationarity is a frequently used activity in autoregressive modelling.

i. ***Augmented Dickey-Fuller (ADF)*** test is a statistical significance test which means the test will give results in hypothesis tests with null and alternative hypotheses.

Before going into the ADF test, we must know about the unit root test because the ADF test belongs to the unit root tests. **Unit Root Test**: A Unit Root Test tests whether a time series is not stationary and consists of a unit root in time series analysis. The presence of a unit root in the

time series defines the null hypothesis, and the alternative hypothesis defines the time series as stationary. Mathematically the unit root test can be represented as:

$$y_t = D_t + z_t + \varepsilon_t$$

where,

- Dt is the deterministic component.
- z_t is the stochastic component.
- ε_t is the stationary error process.

There are various tests which include unit root tests:

- Augmented Dickey-Fuller test.
- Phillips-Perron test.
- KPSS test.
- ADF-GLS test
- Breusch-Godfrey test.
- Ljung-Box test.

The Augmented Dickey-Fuller Test (ADF) is a unit root test for stationarity. The *Augmented* Dickey-Fuller test can be used with serial correlation. But this test should be used with caution because, like most unit root tests, it has a relatively high Type I error rate.

The hypotheses for the test:

- The null hypothesis for this test is that there is a unit root.
- There is no unit root in the time series

In general, a p-value of less than 5% means you can reject the null hypothesis that there is a unit root.

ii. KPSS Test

The **Kwiatkowski–Phillips–Schmidt–Shin** (KPSS) test figures out if a time series is stationary around a mean or linear trend, or is non-stationary due to a unit root. The hypothesis for the test is as follows:

- The null hypothesis for the test is that the data is stationary.
- The alternate hypothesis for the test is that the data is *not* stationary.

A major disadvantage of the KPSS test is that it has a high rate of Type I errors. One way to deal with the potential for high Type I errors is to combine the KPSS with the ADF test. If the result from both tests suggests that the time series is stationary, then it probably is.

4. ARIMA:

ARIMA, short for 'Auto-Regressive Integrated Moving Average' is a class of models that 'explains' a given time series based on its past values, that is, its own and the lagged forecast errors, so that equation can be used to forecast future value. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modelled with ARIMA models. An ARIMA model is characterized by 3 terms: p, d, q where,

p is the order of the AR term

d is the number of differencing required to make the time series stationary

q is the order of the MA term

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'.

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. The equation, in general, becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

ARIMA model in words:

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags).

5. Regression:

Regression analysis is used to predict the value of a variable based on the value of another variable. The value of the variable is called dependent variable which the value using which it is predicted is called independent variable.

a. Simple Linear Regression

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B_0** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B_1** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**) and should be independent of each other.

- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient, which should be normally distributed.

b. Second Degree Curve Fitting:

Let a parabola $y = a + bx + cx^2$

which is fitted to a given data (x_1, y_1), (x_2, y_2), (x_3, y_3) ,....., (x_n, y_n).

Let y_λ be the theoretical value for x_1 then $e_1 = y_1 - y_\lambda$

$$\Rightarrow \qquad e_1 = y_1 - (a + bx_1 + cx_1^2)$$

$$\Rightarrow \qquad e_1^2 = (y_1 - a - bx_1 - cx_1^2)^2$$

Now we have

$$S = \sum_{i=1}^{n} e_i^2$$

$$S = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

By the principle of least squares, the value of S is minimum, therefore

$$\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0 \quad \text{and} \quad \frac{\partial S}{\partial c} = 0$$

Solving equation (2) and dropping suffix, we have

$$\sum y = na + b\sum x + c\sum x^2$$

This model has the same assumptions as of the simple linear regression model.

c. Robust Regression

The linear regression models are based on the normal distribution of errors. If the distribution of errors is prone to outliers, the result of linear regression becomes unreliable. By assigning a weight to each observation, the robust fitting method is less sensitive than ordinary least squares to the outliers. Weighting is done automatically and iteratively using a process called iteratively reweighted least squares. Robust regression is a technique that can reduce the impact of outliers, violation of the distribution assumption and heterogeneity in variance. <u>It should be noted that the linearity assumption is still needed for proper inference using robust regression.</u>

i. Least Median Squares (LMS)

One of the earliest types of robust regression is called median regression, which has the advantage of diminishing the influence of the residuals. This algorithm minimizes the median of ordered squares of residuals to obtain the regression coefficient, b:

$$Least\ median\ of\ squares\,(LMS) = \min_{b} \underset{i}{median} |\, y_i - x_i b \,|^2$$

ii. Least Trimmed Squares (LTS)

One way to eliminate possible outliers is to run the analysis on trimmed or winsorized distributions. The trimmed distribution is estimated by minimizing the sum of squared absolute residuals. By trimming the alpha rejection region, the distorting effects of influential outliers could be pruned from the variables before processing.

$$Least\ trimmed\ squares\,(LTS) = \min_{b} \sum_{i=1}^{q} |\, y_i - x_i b \,|_{(i)}{}^2$$

iii. Huber Loss

The Huber loss function is a combination of the mean squared error function and the absolute value function. The intention behind this is to make the best of both worlds. The Huber Loss offers the best of both worlds by balancing the MSE and MAE together. We can define it using the following piecewise function:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \le \delta, \\ \delta\,|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

This equation essentially says: for loss values less than delta, use the MSE; for loss values greater than delta, use the MAE. This effectively combines the best of both worlds from the two loss functions!

iv. Least Absolute Deviation (LAD)

The Least Absolute Deviation model minimizes the absolute value of the residuals:

$$MAE = \min \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

This provides a more robust solution when outliers are present, but it does have some undesirable properties, most notably that there are some situations where there is no unique solution, and an infinite number of different regression lines are possible. When a unique solution does exist, then the LAD model has the desirable property that if there is one independent variable, then the regression line will pass through at least two of the data points; if there are *k* independent variables, then the residual of at least *k* + 1 of the data elements will be zero.

v. S-estimator and MM-estimators

These estimators are extensions of the Maximum Likelihood Estimator and Robust Estimator techniques, which have been developed based on the M-estimators. In this method, it is possible to eliminate some of the data, which in some cases is not always appropriate to do especially if it is eliminating important data. The M-estimator is defined as follows:

$$\hat{\beta}_M = \min_{\beta} \rho(y_i - \sum_{j=0}^{k} x_{ij}\beta_j).$$

We have to solve

$$\min_{\beta} \sum_{i=1}^{n} \rho(u_i) = \min_{\beta} \sum_{i=1}^{n} \rho(\frac{e_i}{\sigma}) = \min_{\beta} \sum_{i=1}^{n} \rho(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\sigma})$$

to obtain (2), and we set estimator for σ:

$$\hat{\sigma} = \frac{MAD}{0.6745} = \frac{median|e_i - median(e_i)|}{0.6745}.$$

For ρ function we use the Tukey's bisquare objective function:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4}, & |u_i| \leq c \\ \frac{c^2}{6}, & |u_i| > c. \end{cases}$$

Using the Tukey's Biweight function,

The function

$$\psi(x) = \begin{cases} x\left(1-\frac{x^2}{c^2}\right)^2 & \text{for } |x| < c \\ 0 & \text{for } |x| > c \end{cases} \tag{1}$$

sometimes used in robust estimation. It has a minimum at $x = -c/\sqrt{5}$ and a maximum at $x = c/\sqrt{5}$, where

$$\psi'(x) = \frac{(c-x)(c+x)(c^2-5x^2)}{c^4} = 0, \tag{2}$$

and inflection points at $x = 0$ and $x = \pm c/\sqrt{5}$, where

$$\psi''(x) = -\frac{4x(3c^2-5x^2)}{c^4} = 0. \tag{3}$$

,

the x in this above function is u_i in the M-estimator. Then,

the S-estimator is defined by $\hat{\beta}_s = \min_\beta \hat{\sigma}_s(e_1, e_2, \ldots, e_n)$ with determining minimum robust scale estimator $\hat{\sigma}_s$ and satisfying

$$\min \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{i=1}^{n} x_{ij}\beta}{\hat{\sigma}_s}\right)$$

where

$$\hat{\sigma}_s = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i e_i^2}$$

$K = 0.199, w_i = w_\sigma(u_i) = \frac{\rho(u_i)}{u_i^2}$, and the initial estimate is

$$\hat{\sigma}_s = \frac{median|e_i - median(e_i)|}{0.6745}.$$

The solution is obtained by differentiating to β so that

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}_s}\right) = 0, j = 0, 1, \ldots, k \tag{8}$$

ψ is a function as derivative of ρ:

$$\psi(u_i) = \rho'(u_i) = \begin{cases} u_i\left[1-\left(\frac{u_i}{c}\right)^2\right]^2, & |u_i| \le c \\ 0, & |u_i| > c. \end{cases}$$

The MM-estimator is an extension of the M-estimator and the S-estimator. It is as follows:

MM estimation procedure is to estimate the regression parameter using S estimation which minimize the scale of the residual from M estimation and then proceed with M estimation. MM estimation aims to obtain estimates that have a high breakdown value and more efficient. Breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model [3]. MM-estimator is the solution of

$$\sum_{i=1}^{n} \rho_1'(ui)X_{ij} = 0 \text{ or } \sum_{i=1}^{n} \rho_1'\Big(\frac{Y_i - \sum_{j=0}^{k} X_{ij}\hat{\beta}_j}{s_{MM}}\Big)X_{ij} = 0$$

where s_{MM} is the standard deviation obtained from the residual of S estimation and ρ is a Tukey's biweight function:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^2}, & -c \leq u_i \leq c; \\ \frac{c^2}{6}, & u_i < -c \text{ or } u_i > c. \end{cases}$$

6. Root Mean Square Error:

The **root-mean-square deviation** (**RMSD**) or **root-mean-square error** (**RMSE**) is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. The RMSD serves to aggregate the magnitudes of the errors in predictions for various data points into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}.$$

7. Ljung-Box test:

The Ljung-Box test, named after statisticians Greta M. Ljung and George E.P. Box, is a statistical test that checks if autocorrelation exists in a time series. The test is applied to the residuals of a time series after fitting an ARIMA (p, d, q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit a significant lack of fit.

The Ljung-Box test uses the following hypotheses:

H_0: The residuals are independently distributed.

H_A: The residuals are not independently distributed; they exhibit serial correlation.

The test statistic for the Ljung-Box test is as follows:

$$\mathbf{Q = n\,(n+2)\,\Sigma p_k^2 \,/\, (n-k)}$$

Where:

n = sample size

p_k = sample autocorrelation at lag *k*

The test statistic *Q* follows a chi-square distribution with *h* degrees of freedom; that is, $Q \sim X^2(h)$. We reject the null hypothesis and say that the residuals of the model are not independently distributed if $Q > X^2_{1-\alpha,\, h}$

INTRODUCTION

India is the third-largest energy consumer in the world after China and USA. It is also the fastest-growing energy consumer. With a share of 5.7% of the World's primary energy consumption, India's energy requirement is fulfilled primarily by Coal, Crude Oil, Natural Gas and Renewable Energy. Oil and gas within the energy mix play an important role as over one-third of the energy required is met by hydrocarbons. The growing economy and population growth are the main drivers for oil & gas demand increasing every year.

"*A country with over 1.4 billion people, India imports around 84% of its crude oil needs. Also, India is the third-largest oil consumer and importer in the world. With a marked increase in oil prices, increase in imports, and decline in domestic production, India's crude oil import bills are rising.*

LNG, which is about one-fourth of the total gas demand in India, is also imported to a large extent. India is the fourth largest importer of LNG too. There is a growing demand for oil and natural gas in India"

-Groww

India is not an oil-rich country; hence we need to import most of our oil from OPEC member countries. This is important as its prices are dependent on the *Brent Crude Oil Index* which has been marked to rise and has been rising significantly over the past few months. Hence, we need to analyse what the price will be in the future months so that the Indian Government can decide the best time to purchase crude oil based on past trends.

This is what crude oil looks like (Source: eworldtrade.com)

Once we get crude oil, we need to refine it. So, we study the refinery capacity to know if the refineries can keep up with the demand and deliver crude oil products faster. We also analyse the consumption pattern to see how much the demand will rise year on year and see the demand for the coming years. And since India's gas consumption is also rising due to population growth, we need to decide on plans well in advance.

Since we are analysing crude oil prices and other factors, we should know a little bit about the composition of crude oil too.

Crude oil is a complex liquid mixture made up of a vast number of hydrocarbon compounds that consist mainly of carbon and hydrogen in differing proportions. In addition, small amounts of organic compounds containing sulphur, oxygen, nitrogen and metals such as vanadium, nickel, iron and copper are also present. The main constituents of crude oil are:

1- Paraffins
Example: methane, ethane, propane, butane, isobutane, pentane, hexane
2- Olefins (also known as alkenes)
Example: ethylene, butene, isobutene
3- Naphthenes (cycloalkanes)
Example: cyclohexane, methyl cyclopentane
4- Aromatics
Example: toluene and xylene.
5- Sulphur Compounds
The Sulphur content of crude oils varies from less than 0.05 to more than 10 wt.% but generally falls in the range of 1–4 wt.%. Crude oil with less than 1 wt. % sulphur is referred to as low sulphur or sweet, and that with more than 1 wt.% sulphur is referred to as high sulphur or sour.
6-Nitrogen Compounds
Crude oils contain very low amounts of nitrogen compounds, less than 1%, The nitrogen compounds in crude oils may be classified as basic or non-basic.
7- Oxygen Compounds
Less than 1% (found in organic compounds such as carbon dioxide, phenols, ketones and carboxylic acids) occur in crude oils in varying amounts.
8- Metals Compounds
Metallic compounds exist in all crude oil types in very small amounts.

Petroleum Products Obtained from a Barrel of Crude Oil (Source: Howden)

PRICE of CRUDE OIL

We now move on to the analysis of crude oil prices. There are two main indices which measure and mark the price of a barrel of oil ($) after the crude oil is dug out and pumped into barrels. They are: the WTI and the Brent indices. Like any industry, supply and demand heavily affect the prices and profitability of crude oil. The United States, Saudi Arabia, and Russia are the leading producers of oil in the world. The price of the spot contract reflects the current market price for oil, whereas the futures price reflects the price buyers are willing to pay for oil on a delivery date set at some point in the future.

The futures price is no guarantee that oil will actually hit that price in the current market when that date comes. It is just the price that, at the time of the contract, purchasers of oil are anticipating. The actual price of oil on that date depends on many factors. These contracts are basically what these two indices are about. Both these indices are publicly traded, Brent on the Intercontinental Exchange (ICE) and WTI on NYMEX exchange of the CME group.

Here's a quick comparison between BRENT and WTI:

BRENT:

Brent crude refers to the price of the ICE Brent Crude Oil futures contract. Brent blend is a blend of crude oil extracted from the oilfields in the North Sea between the United Kingdom and Norway. It is an industry-standard because it is "light", meaning not overly dense and "sweet", meaning it is low in sulphur content. It is one of the two main benchmarks for the prices of crude oil. This index is used to price two-thirds of the crude oil traded internationally. The contract settlement of the Brent index happens in London, UK. India buys oil from countries based on the price of barrels on this index. The prices of BRENT are mostly controlled/ influenced by the decisions made by OPEC+.

The BRENT Council Civic Centre in London, UK. (Source: Wikipedia)

WTI:

West Texas Intermediate (WTI) crude oil is a specific grade of oil and the other main benchmark in oil pricing, along with Brent. WTI is known as a light and sweet oil because it contains a very low amount of sulphur and a low density. WTI is the main benchmark for North America as it is sourced from the United States, mainly from the Permian Basin. It is the second main benchmark for prices of crude oil. The oil comes mainly from Texas. It then travels through pipelines to Cushing, Oklahoma after getting refined. The main point for physical exchange and price settlement is Cushing, which is also known as the "Pipelines Capital of the World".

Pipelines Crossroads of the World;

(Above)

Oil Storages in Cushing, Oklahoma

(Below)

(Source: Wikipedia and Reuters)

Price Comparision between Brent and WTI up to Apr 2022:

Historically, price differences between Brent and other index crudes have been based on physical differences in crude oil specifications and short-term variations in supply and demand. Many reasons have been given for this divergence ranging from regional demand variations to the depletion of the North Sea oil fields.

What is OPEC?

The Organization of the Petroleum Exporting Countries (OPEC) consists of 13 major players in the global oil industry. The primary objectives of OPEC were:

- Coordinating and unifying petroleum prices across member countries
- Ensuring the stabilization of oil markets
- Securing an efficient, regular, and economical supply to consumers
- Ensuring that producers and companies investing in the industry get a fair return on their investments

The countries in OPEC are:

- Algeria
- Angola
- Equatorial Guinea
- Gabon
- Iran
- Iraq
- Kuwait
- Libya
- Nigeria
- The Republic of Congo
- Saudi Arabia
- The United Arab Emirates
- Venezuela

The OPEC headquarters in Vienna, Austria (Source: Britannica)

In 2016, a larger group was formed called OPEC+ to increase the organisation's control over the global crude oil market. These non-OPEC countries are:

- Azerbaijan
- Bahrain
- Brunei
- Kazakhstan
- Malaysia
- Oman
- Russia
- Sudan
- South Sudan

These countries help major importers like India diversify their energy import basket and get oil at reasonable prices. While India is still primarily importing from OPEC countries, oil imports from the United States are now the second-largest in India.

Now, we analyse the Brent index prices first and WTI later. We will then compare the predicted values for the next year. Following is the BRENT crude oil price chart up to April 2022. As we can see, the variation and increase have not been constant. Taking that into consideration, we will try and fit models to the data.

We divide the data into training and test set. The training set has values from July 1988 to July 2015. The test data has values from August 2015 to April 2022. Since the variance is not constant, we decompose the training set into various components viz. trend, seasonality and randomness w.r.t. a multiplicative model.

We fit Holt's model on the training data:

```
## Holt-Winters exponential
smoothing with trend and
multiplicative seasonal component.

## Call:

## HoltWinters(x = train, seasonal
= c("multiplicative"))

## Smoothing parameters:

##  alpha: 0.886391

##  beta : 0

##  gamma: 1

## Coefficients:

##           [,1]

## a   51.4336195

## b    0.2384659

## s1   1.0056421

## s2   0.9509448

## s3   0.9204578

## s4   0.8920006

## s5   0.8837637

## s6   0.9125547

## s7   0.9839786

## s8   0.9764985

## s9   1.0303304

## s10  1.0184624

## s11  1.0394512

## s12  1.0150948
```

The following chart shows the Holt-Winters fitted values and training data:

We now forecast values for the length of the test set and plot it with the actual test data to get a general idea of the model.

Now, we perform the KPSS test and ADF test to check for stationarity of the training set:

```
## # KPSS Unit Root Test #
## Test is of type: mu with 5 lags.
## Value of test-statistic is:
4.374
## Critical value for a
significance level of:
##                 10pct  5pct
2.5pct  1pct
## critical values 0.347 0.463
0.574 0.739
```

We check if any differencing is required for the data to make it stationary. In this case, we need to difference the data once.

```
## [1] 1    #Differencing required
of lag 1
##  Augmented Dickey-Fuller Test
## data:  train
## Dickey-Fuller = -2.2367, Lag
order = 6, p-value = 0.4767
## alternative hypothesis:
stationary
```

From both the KPSS test and the ADF test, we see that the data is not stationary. We now fit an ARIMA model to try and make the model stationary. Following are different combinations which were tried and the best model was found using AIC as a criterion.

```
##  ARIMA(0,1,0)                              : 1946.488
##  ARIMA(0,1,0)            with drift        : 1948.339
##  ARIMA(0,1,0)(1,0,0)[12] with drift        : 1950.339 …
##  ARIMA(0,1,0)(1,0,1)[12]                   : 1950.31
##  Best model: ARIMA(1,1,0)
```

The following graph shows training data vs the ARIMA model:

We now plot the test values and the forecasted values on the same chart to understand the model accuracy:

We have stored the actual values and the predicted values by the two models to calculate RMSE values to determine which model is better:

```
## [1] "The predicted values by the Holt-Winters and ARIMA methods and the
test data is shown in the table below:"

##          Year   Test Holts_Method    ARIMA

## 1  2015-08-01  54.15     51.96363 48.73196

## 2  2015-09-01  48.37     49.36407 47.66898

## 3  2015-10-01  49.56     48.00097 47.34410 …

## 81 2022-04-01 107.14     72.89522 47.20111
```

THE RMSE values for the two models:

```
# [1] "The RMSE of the models are:"

##                Holt's Smoothing    ARIMA

## RMSE of models         14.87895 20.79111
```

We can see that the RMSE values are really small for both models, indicating that both models are very accurate. Below, we find the residuals, Ljung-Box test and runs test for the residuals of both the models to find the goodness of fit of the model:

```
##  Box-Ljung test

## data:  resid(expo2)

## X-squared = 35.737, df = 1, p-
value = 2.259e-09

##  Approximate runs rest

## data:  resid(expo2)

## Runs = 142, p-value = 0.07928

## alternative hypothesis:
two.sided
```

Runs test p-value is indicating that the sequence of residuals is random, which implies that the model might be a good fit but the Ljung-Box test is rejected.

Analysis of the ARIMA model:

```
##  Ljung-Box test

## data:  Residuals from
ARIMA(1,1,0)

## Q* = 29.304, df = 23, p-value =
0.1704

## Model df: 1.   Total lags used:
24

##  Approximate runs rest

## data:  resid(acons)

## Runs = 177, p-value = 0.1336
```

We can see that the ARIMA model is a good fit. We now predict the price of a barrel of oil for the next 1 year using both, Holt-Winters and ARIMA.

```
## [1] "The predicted values for 2021 predicted by the different models are:"

##         Years Holts_method
ARIMA

## 1  2022-05-01     105.7580
106.9286

## 2  2022-06-01     105.3532
106.8706

## 3  2022-07-01     104.2504
106.8546

## 4  2022-08-01     104.9076
106.9286

## 5  2022-09-01     105.9865
106.8706

## 6  2022-10-01     106.8195
106.8546

## 7  2022-11-01     107.0568
106.9286

## 8  2022-12-01     109.2044
106.8706

## 9  2023-01-01     107.2491
106.8546

## 10 2023-02-01     107.8334
106.9286

## 11 2023-03-01     110.8652
106.8706

## 12 2023-04-01     110.0016
106.8546
```

We can see that the prices are marked to increase over the year. So, the only way to stop this is to pump out more oil from the reserves and 'oversupply' so that the prices will come down. But this is not always possible.

We now analyse the prices of the WTI index. Below is the original data plotted along with the values upto April 2022.

We divide the data into training and test set. The training set has values from July 1988 to July 2015. The test data has values from August 2015 to April 2022. Since the variance is not constant, we decompose the training set into various components viz. trend, seasonality and randomness using a multiplicative model.

We fit Holt's model on the training data:

```
## Holt-Winters exponential
smoothing with trend and
multiplicative seasonal component.

## Call:

## HoltWinters(x = train, seasonal
= c("multiplicative"))

## Smoothing parameters:

##  alpha: 1

##  beta : 0

##  gamma: 0

## Coefficients:

##           [,1]

## a   52.0579731

## b    0.2756425

## s1   0.9134807

## s2   0.9695813

## s3   0.9639662

## s4   0.9702911

## s5   1.0771869

## s6   0.9622831

## s7   1.0118239

## s8   1.0984257

## s9   1.0781921

## s10  1.0255519

## s11  1.0240725

## s12  0.9051447
```

The following chart shows the Holt's Model values and the training data:

We now forecast values for the length of the test set and plot it with the actual test data to get a general idea of the model

Now, we perform the KPSS test and ADF test to check for stationarity of the training set:

```
## # KPSS Unit Root Test #                    ## Value of test-statistic is:
                                               4.3947
## Test is of type: mu with 5 lags.
```

```
## Critical value for a
significance level of:
##                  10pct  5pct
2.5pct  1pct
## critical values 0.347 0.463
0.574 0.739
## [1] 1
```

```
##  Augmented Dickey-Fuller Test
## data:  train
## Dickey-Fuller = -2.2619, Lag
order = 6, p-value = 0.4661
## alternative hypothesis:
stationary
```

From both the KPSS test and the ADF test, we see that the data is not stationary. We now fit an ARIMA model to try and make the model stationary. Following are different combinations which were tried and the best model was found using AIC as a criterion.

```
##  ARIMA(0,1,0)                                : 1959.505
##  ARIMA(0,1,0)            with drift          : 1961.411
##  ARIMA(0,1,0)(0,0,1)[12]                     : 1961.134
##  ARIMA(0,1,0)(0,0,2)[12]                     : 1962.374 …
##  Best model: ARIMA(5,1,0)
```

The following chart shows the ARIMA model and the training data to get a general idea of the fit of the model. We also plot the test values and the forecasted values on the same chart to understand the model accuracy:

Below are the actual test values and the predicted values from the different methods using which we will find the RMSE of the models and fit the best model for the whole data.

```
## [1] "The predicted values by the Holt-Winters and ARIMA methods and the
test data is shown in the table below:"

##          Year   Test Holts_Method    ARIMA

## 1  2015-08-01  49.20     47.80575 44.54068

## 2  2015-09-01  45.09     51.00895 41.16333

## 3  2015-10-01  46.59     50.97926 39.99491 …

## 81 2022-04-01 103.03     80.20133 43.47315
```

We now calculate the RMSE values of both models:

```
## [1] "The RMSE of the models are:"

##                Holt's Smoothing    ARIMA

## RMSE of models         15.96224 19.59952
```

Since RMSE is low for both models, we fill check the goodness of fit for both these models.

Analysis of Holt-Winters:

```
##  Box-Ljung test

## data:  resid(expo2)

## X-squared = 8.1404, df = 1, p-
value = 0.004329

##  Approximate runs rest

## data:  resid(expo2)

## Runs = 158, p-value = 0.9547

## alternative hypothesis:
two.sided
```

We can see that the p value is 0.004329, which means that the residuals are autocorrelated, signifying that the model is not a good fit. But the runs test p-value is indicating that the sequence of residuals is random, which implies that the model is a good fit. We perform the same tests for the other model.

Analysis of the ARIMA model:

```
##  Ljung-Box test

## data:  Residuals from
ARIMA(5,1,0)

## Q* = 28.427, df = 19, p-value =
0.07556

## Model df: 5.   Total lags used:
24
```

```
##  Approximate runs rest

## data:  resid(acons)

## Runs = 173, p-value = 0.2911

## alternative hypothesis:
two.sided
```

From the above values, we can see that the ARIMA model is a good fit. We now fit the models for the whole data and predict future 1-year values. Predicting the WTI prices for the future 1-year using the models:

```
## [1] "The predicted values for 2021 predicted by the different models are:"

##         Years Holts_method
ARIMA

## 1  2022-05-01     102.3302
103.2390

## 2  2022-06-01     102.5946
102.6561

## 3  2022-07-01     100.4840
102.6127

## 4  2022-08-01     100.9030
103.2390

## 5  2022-09-01     102.3336
102.6561

## 6  2022-10-01     102.4939
102.6127

## 7  2022-11-01     102.9020
103.2390

## 8  2022-12-01     105.3451
102.6561

## 9  2023-01-01     103.3366
102.6127

## 10 2023-02-01     104.4968
103.2390

## 11 2023-03-01     106.3787
102.6561

## 12 2023-04-01     106.3377
102.6127
```

We can see that the price is predicted to increase over time just like Brent's predicted prices. Thus, we should take measures that it does not be so like this and we do not pay too much money for crude oil. All measures should be taken so that geopolitical tensions, country elections, etc. do not affect the oil prices much.

Comparison of WTI and Brent future prices:

We now compare the predicted prices of WTI and Brent by both Holt-Winters and ARIMA methods using correlation. Here BHolts means Brent forecasts by Holt-Winters method and BARIMA means Brent forecasts by ARIMA. The other two columns are for the WTI prices respectively.

We are finding the correlation between the prices as a comparison. Why correlation? It's because we want to see that does the change in one's price affects the other one or are they uncorrelated.

```
##    Holts_method    ARIMA
BHolts   BARIMA

## 1      102.3302 103.2390
105.7580 106.9286

## 2      102.5946 102.6561
105.3532 106.8706

## 3      100.4840 102.6127
104.2504 106.8546

## 4      100.9030 103.2390
104.9076 106.9286

## 5      102.3336 102.6561
105.9865 106.8706

## 6      102.4939 102.6127
106.8195 106.8546

## 7      102.9020 103.2390
107.0568 106.9286

## 8      105.3451 102.6561
109.2044 106.8706

## 9      103.3366 102.6127
107.2491 106.8546

## 10     104.4968 103.2390
107.8334 106.9286

## 11     106.3787 102.6561
110.8652 106.8706

## 12     106.3377 102.6127
110.0016 106.8546
```

From the above chart, we see that some prices are positively correlated while some are negatively correlated. Therefore, changes in one's prices (for example ARIMA predicted WTI price) do affect the other one (say ARIMA predicted Brent price). Thus, <u>we can say that if the US decides to increase its oil prices, OPEC will have to increase their oil prices too.</u>

PETROL and DIESEL PRICE ANALYSIS

We have also analysed the petrol and diesel prices in India as a consequence of crude oil prices since the prices of these two fuels affect us the most in day-to-day life. Below is the chart which compares the yearly prices of petrol and diesel prices from 2003.

From the above line graph, we can easily visualize that the trend is increasing over the period from 2003 to 2022. So, considering this data, we can forecast the price of Petrol as well as Diesel for the near future.

A. PETROL

The price has been taken as an average of the prices of the months. We will fit only a 2nd degree linear regression model and an ARIMA model to the data. The following is the price of petrol year wise:

1. RMSE for 2nd Order Regression:	2. RMSE for ARIMA
RMSE = 6.277019	RMSE = 6.209720

The RMSE for the ARIMA model is lowest. So, we will use it for forecasting. The forecasted average yearly values for the future 3 years are as follows:

Year	Point Forecast
2023	109.1953
2024	112.9805
2025	116.7658

B. DIESEL

This is what the price of diesel over the years looks like. As it can be seen, the price has been constantly rising.

To forecast the diesel price from the previous years' data, we will again fit 2nd degree Regression and ARIMA. The following are the RMSE values:

1. RMSE for 2nd Degree Regression	2. RMSE for ARIMA
RMSE = 3.787507	RMSE = 4.925929

We will predict the prices of diesel for the next three years using the ARIMA model since it is a better model and can be adjusted according to the observation. The forecasted value for the next three years on an average is:

Year	Point Forecast
2023	100.5937
2024	104.5174
2025	108.4411

We can see how much the petrol and diesel costs are expected to rise in the next 3 years. Nowadays, the sky-touching cost of petrol is directly affecting the day-to-day life of the common man. The price is recording the new heights every day. So, we must have to find out other alternative resources over the use of such crude oil. There are other few potential alternative resources are available that we need to adopt such as Electric Vehicles, CNG Vehicles, or the more use of public transport. The graph given below is showing that what will be the price of petrol in the year 2042 if the current scenario continues to happen...

The estimated value is 192 rupees per litre with an upper limit of 251 rupees per litre. These prices may increase exponentially in the coming future because of the consistently increasing demand and drying resources. So, we should try and go for alternative sources of energy or use more public transports to try and save these fuels so that we can keep the prices from rising exponentially!!

IMPORT of CRUDE OIL

India is the world's third-largest consumer of oil, after the United States and China. India imports nearly 84% of its crude oil to meet its energy needs. The country has few oil reserves, and production is limited, which is why it has to import most of its oil. India is looking to reduce its dependence on oil imports, but this will be a challenge because demand for oil is expected to grow in the coming years. The crude oil that is brought in from other countries is turned into more valuable products like petrol and diesel at oil refineries. These products are then sold to automotive vehicles and used in other ways. India has plenty of capacity to refine and export petroleum products, but it doesn't have enough role in making cooking gas, LPG, bought from countries like Saudi Arabia.

An Oil Transport Pipeline (Source: energyeducation.ca)

The main suppliers of oil to India are Middle East countries. Their share of the import basket of oil is 52.7%. Then it's Africa's —15% and then the United States'—14%. India is working on diversifying the country's energy basket with crude oil supplies from non-organisation of the Petroleum Exporting Countries (OPEC) nations. Most of the crude oil imported to India comes in oil tankers (ships).

An Oil Tanker used to transport oil (Source: International Chamber of Shipping)

Import Data from the Top 5 Nations:

1. Iraq—867,500 BPD (barrels per day)

2. United States—-545,300 BPD—14% of India's overall imports in FEB 2021

3. Nigeria —-472,300 BPD

4. Saudi Arabia—-445,200 BPD.

5. UAE

Crude Oil Reserves Around the World (Source: howmuch.net)

We will now analyse India's import data and make some important inferences from the analysis. Following is the plot of the import data over the years. As we can see that the trend has been increasing and will only continue to increase further still.

Now, we divide the data into training and test set to better analyse the data. The training data is from 1980 to 2011 and the test data is from 2012 to 2020. We fit Holt's model on the training data:

```
## Holt-Winters exponential
smoothing with trend and without
seasonal component.

## Call:

## Holt Winters (x = train, gamma =
F)

## Smoothing parameters:

## alpha: 1

## beta: 1

## gamma: FALSE

## Coefficients:

## [,1]

## a 3365.5976

## b 606.9488
```

The following chart shows the Holt's Model values and the training data:

We now forecast values for the length of the test set and plot it with the actual test data to get a general idea of the model.

Now, we perform KPSS test and ADF test to check for stationarity of the training set

```
## # KPSS Unit Root Test #

## Test is of type: mu with 3 lags.

## Value of test-statistic is:
0.8296

## Critical value for a
significance level of:

##                  10pct 5pct
2.5pct 1pct

## critical values 0.347 0.463
0.574 0.739
```

We check if any differencing is required for the data to make it stationary. In this case we need to difference the data twice.

```
## [1] 2

## Augmented Dickey-Fuller Test

## data:  train

## Dickey-Fuller = 0.0027572, Lag
order = 3, p-value = 0.99

## alternative hypothesis:
stationary
```

<u>From both the KPSS test and the ADF test, we see that the data is not stationary.</u> We now fit an ARIMA model to try and make the model stationary. Following are different combinations which were tried and the best model was found using AIC as a criterion.

```
## ARIMA (0,2,0)              : 380.0123

## ARIMA (0,2,1)              : 382.2442 …

## ARIMA (5,2,0)              : 386.1154

## Best model: ARIMA (0,2,2)
```

The following chart shows the ARIMA model and the training data to get a general idea of the fit of the model.

We now plot the test values and the forecasted values on the same chart to understand the model accuracy:

Below is the chart which compares the different regression model vs the training data. From the regression line, we can guess the best regression model for the training data:

Below are the actual test values and the predicted values from the different methods using which we will find the RMSE of the models and fit the best model to the whole data.

```
## [1] "The predicted values by the models and the test data is shown in the
table below:"

##   Year     Test Holts_Method    ARIMA LS_Predict PLS_Predict H_Predict

## 1 2012 3682.236     3972.546 3842.515   2623.603    3402.784 2623.603

## 2 2013 3792.576     4579.495 4116.187   2713.947    3634.798 2713.947 …

## 9 2020 4033.050     8828.137 6031.889   3346.354    5492.227 3346.354

##   LMS_Predict LTS_Predict LAD_Predict S_Predict MM_Predict

## 1    1031.135    1048.406    2570.864 967.7911   2575.980

## 2    1056.203    1074.114    2659.239 990.5867   2664.744 …

## 9    1231.677    1254.065    3277.866 1150.1561   3286.096
```

We now calculate the RMSE values of the different models for comparison:

```
##                Holt's Smoothing    ARIMA Huber Loss      LAD      LMS
LS

## RMSE of models         2662.594 978.5078   1136.778 1196.502 2984.083
1136.778
```

```
##                      LTS MM-estimator      PLS S-estimator
## RMSE of models 2964.237      1189.92 577.1354    3056.572
## [1] "The minimum RMSE values among all the models is:"
## [1] 577.1354
#Project compiled by Sombit Ghosh.
```

We can see that the minimum RMSE is for the PLS model. But we analyse the residuals of the best 3 models by their ACF, noise plot, Ljung-Box test and runs test to check the fit of the models.

```
## Box- Ljung test                          ## X-squared = 0.0003394, df = 1,
                                            p-value = 0.9853
## data:  resid (expo2)
```

<u>We can see that the p value is 0.9853, which means that the residuals are not autocorrelated, signifying that the model is a good fit.</u>

```
## Approximate runs rest                    ## alternative hypothesis:
                                            two.sided
## data:  resid (expo2)
## Runs = 19, p-value = 0.2649
```

<u>Runs test p-value is also indicating that the sequence of residuals is random, which implies that the model is a good fit</u>. We perform the same tests for the other two models.

Analysis of the ARIMA model:

```
## Ljung-Box test

## data:  Residuals from ARIMA
(0,2,2)

## Q* = 3.3093, df = 4, p-value =
0.5075

## Model df: 2.   Total lags used:
6
```

```
## Approximate runs rest

## data:  resid (acons)

## Runs = 22, p-value = 0.07234

## alternative hypothesis:
two.sided
```

The above tests imply that the ARIMA model is also a good fit.

Analysis of the 2nd Degree Linear Regression:

```
## Box-Ljung test

## data:  fitPLS$residuals

## X-squared = 8.4489, df = 1, p-
value = 0.003653

## Approximate runs rest
```

```
## data:  fitPLS$residuals

## Runs = 12, p-value = 0.07234

## alternative hypothesis:
two.sided
```

The above tests indicate that the second-degree model is not a good fit for the data. We now fit the above 3 selected models on the whole data and predict the future 3 years values. We select the PLS model for prediction too. Predicting the import for the future 3 years using the models selected:

```
## [1] "The predicted values for 2021,2022 and 2023 predicted by the
different models are:"

##   Years      PLS Holts_method    ARIMA

## 1 2021 5211.305     3952.313 3873.835

## 2 2022 5463.927     3871.576 3959.037

## 3 2023 5722.790     3790.840 4044.240
```

From the forecasts, <u>we can see that the import is only going to increase over the years</u> unless we totally shift towards other sources of energy like electric energy. Thus, to tackle this problem, we should build good relations with Oil Exporting nations so that we can get uninterrupted supply of crude oil and cause no problems to the citizens of our country.

REFINERIES of INDIA

India is emerging as a refinery hub and refining capacity exceeds the demand. The country's refining capacity has increased from a modest 62.00 million Metric Tonnes. The Indian Refining Industry has established itself as a major player globally.

Refinery Capacity of crude oil in India financial year 2020-2021 is expected to be one of the largest contributors to non-OECD petroleum consumption. The refinery capacity of crude oil in India was approximately 250 million metric tons per annum in financial year 2021. The economic growth in the country is closely related to energy demand, and with increasing industrialization and automation, the demand for petroleum products has been on a rise over the years. The volume of the crude oil import in the country was approximately around 226 million metric tons during financial year 2019, most likely due to the increased energy consumption over the region. With a workforce of over 30 thousand, Indian Oil Corporation is a major contributor in the refining and development sector.

Indian Oil Corporation Limited's Refinery at Paradip, Odisha. (Source: iocl.com)

The Indian petroleum refining sector has come a long way since crude oil was discovered and the first refinery was set up at Digboi in 1901. The present Mumbai Refinery of HPCL was the first modern refinery to be setup after independence by ESSO in 1954. India has witnessed a spectacular growth in the refining sector over the years. From deficit scenario in 2021, the

country achieved self-sufficiency in refining and today is the major exporter of quality petroleum products. There are 23 refineries in the country, 18 in the public sector, 2 in the joint venture and 3 in the private sector.

The Jamnagar Refinery in Gujarat, India (Source: Business Standard)

The majority of world's 10 largest refineries are situated in the Asia pacific region, with India hosting the world's largest refinery complex. Reliance Jamnagar refinery is the world's largest oil refinery with an aggregate capacity 1.24 million barrels per day. The refinery complex is located at Jamnagar in Gujarat, India. The following table shows the refineries and their capacities in India:

Sl. No	Refineries	Name of the company	Name Plate Capacity (MMTPA)
PSU Refineries			
1	Digboi-1901#	Indian Oil Corporation Limited	0.65
2	Guwahati-1962		1.00
3	Barauni-1964		6.00
4	Koyali-1965		13.70
5	Bongaigaon-1974		2.35
6	Haldia-1975		8.00
7	Mathura-1982		8.00
8	Panipat-1998		15.00
9	Paradip-2016		15.00
10	Mumbai-1954	Hindustan Petroleum Corporation Limited	7.50
11	Visakhapatnam-1957		8.30
12	Mumbai-1955	Bharat Petroleum Corporation Limited	12.00
13	Kochi-1963		15.50
14	Manali-1965	Chennai Petroleum Corporation Limited	10.50
15	Nagapattinam-1993		0 @
16	Numaligarh-2000	Numaligarh Refinery Limited	3.0
17	Mangalore-1996	Mangalore Refinery and Petrochemicals Limited	15.0
18	Tatipaka, AP-2001	Oil and Natural Gas Commission	0.066
Total			**141.566**

Sl. No	Refineries	Name of the company	Name Plate Capacity (MMTPA)
JV Refineries			
19	Bina-2011	Bharat Oman Refinery Ltd.	7.80
20	Bathinda-2012	HPCL Mittal Energy Ltd.	11.30
Total			**19.10**
Private Sector Refineries			
21	DTA-Jamnagar-1999	Reliance Industries Limited	33.00
22	SEZ, Jamnagar-2008		35.20
23	Vadinar-2006	Essar Oil Limited	20.00
Total			88.20
		Grand Total	**248.866**

Just to get an idea of where India stands on the world's standard, we are comparing the refinery capacities of the top ten GDP producing countries:

We can see from the chart that from 1980 to 2020, the refining capacity of almost every country has been in a fixed range except for China where the capacity has increased exponentially, marking the industrialisation and advancement of China in every field like technology, manufacturing, etc.

Below we analyse the refining capacity of India and try to predict the refining capacity for the future 3 years.

The following chart shows the actual values of the refinery capacity in 1000 bpd year wise for India. It is clearly visible that it has been increasing over the years as India has been moving towards advancement in the field of Oil and Gas:

Now, we divide the data into training and test set to better analyse the data. The training data set is from 1980 to 2011. The test set is from 2012 to 2020.

We fit Holt's model on the training data:

```
## Holt-Winters exponential
smoothing with trend and without
seasonal component.

## Call:

## HoltWinters(x = train, gamma =
F)

## Smoothing parameters:

##  alpha: 0.8746089

##  beta : 0.127363

##  gamma: FALSE

## Coefficients:

##        [,1]

## a 3803.6019

## b  158.0212
```

The following chart shows the Holt's Model values and the training data:

We now forecast values for the length of the test set and plot it with the actual test data to get a general idea of the model.

Now, we perform KPSS test and ADF test to check for stationarity of the training set

```
## # KPSS Unit Root Test #

## Test is of type: mu with 3 lags.

## Value of test-statistic is:
0.8438

## Critical value for a
significance level of:

##                 10pct  5pct
2.5pct  1pct
```

```
## critical values 0.347 0.463  0.574 0.739
```

We check if any differencing is required for the data to make it stationary. In this case we need to difference the data once.

```
## [1] 1     #Differencing of lag 1
required

##  Augmented Dickey-Fuller Test

## data:  train

## Dickey-Fuller = -0.98573, Lag
order = 3, p-value = 0.9248

## alternative hypothesis:
stationary
```

From both the KPSS test and the ADF test, we see that the data is not stationary. We now fit an ARIMA model to try and make the model stationary. Following are different combinations which were tried and the best model was found using AIC as a criterion.

```
##  ARIMA(0,1,0)                    : 421.7188

##  ARIMA(0,1,0) with drift         : 415.2296

##  ARIMA(0,1,1)                    : 423.011

##  ARIMA(0,1,1) with drift         : 417.5821  …

##  Best model: ARIMA(0,1,0) with drift
```

The following chart shows the ARIMA model and the training data to get a general idea of the fit of the model.

We now plot the test values and the forecasted values on the same chart to understand the model accuracy:

Below is the chart which compares the different regression models vs the training data.

Below are the actual test values and the predicted values from the different methods using which we will find the RMSE of the models and fit the best model to the whole data.

```
## [1] "The predicted values by the models and the test data is shown in the
table below:"

##   Year     Test Holts_Method    ARIMA LS_Predict PLS_Predict H_Predict

## 1 2012 4278.828     3961.623 3899.017   3353.334    4026.972  3353.334

## 2 2013 4319.025     4119.644 4003.455   3452.282    4248.400  3452.282 …

## 9 2020 4796.265     5225.792 4734.521   4144.918    6000.126  4144.918

##   LMS_Predict LTS_Predict LAD_Predict S_Predict MM_Predict

## 1        1646    1840.786    3299.769  3264.009   3330.231

## 2        1673    1877.571    3392.439  3352.210   3426.267 …

## 9        1862    2135.071    4041.132  3969.613   4098.524
```

We now calculate the RMSE values for the different models to find the best models.

```
##                Holt's Smoothing    ARIMA Huber Loss      LAD      LMS
LS

## RMSE of models         222.1981 247.9901   799.9832 877.4143 2794.234
799.9832

##                      LTS MM-estimator      PLS S-estimator

## RMSE of models 2559.487      834.096 618.6418    930.4747

## [1] "The minimum RMSE values among all the models is:"

## [1] 222.1981
```

We can see that the minimum RMSE is for the Holt's smoothing model. But we analyse the residuals of the best 3 models by their ACF, noise plot, Ljung-Box test and runs test to check the fit of the models.

```
##  Box-Ljung test                          ## X-squared = 0.079224, df = 1, p-
                                            value = 0.7784
## data:  resid(expo2)
```

We can see that the p value is 0.7784, which means that the residuals are not autocorrelated, signifying that the model is a good fit.

```
##  Approximate runs rest                   ## alternative hypothesis:
                                            two.sided
## data:  resid(expo2)

## Runs = 12, p-value = 0.1372
```

Runs test p-value is also indicating that the sequence of residuals is random, which implies that the model is a good fit. We perform the same tests for the other two models.

Analysis of the ARIMA model:

```
##  Ljung-Box test                          ##  Approximate runs rest

## data:  Residuals from                    ## data:  resid(acons)
ARIMA(0,1,0) with drift
                                            ## Runs = 17, p-value = 1
## Q* = 3.2481, df = 5, p-value =
0.6618                                      ## alternative hypothesis:
                                            two.sided
## Model df: 1.   Total lags used:
6
```

Analysis of the 2nd Degree Linear Regression:

```
##  Box-Ljung test

##

## data:  fitPLS$residuals

## X-squared = 9.6244, df = 1, p-value = 0.00192

##
```

```
##  Approximate runs rest

##

## data:  fitPLS$residuals

## Runs = 13, p-value = 0.1506

## alternative hypothesis: two.sided
```

From the above tests it can be seen that 2nd Degree Regression model is not a good fit but the ARIMA model is a good fit. We now fit the above 3 selected models on the whole data and predict the future 3 years data. We are fitting 2nd degree regression model even though it is not a good fit on the data. Predicting the refinery output for the future 3 years using the models selected:

```
## [1] "The predicted values for 2021,2022 and 2023 predicted by the different models are:"

##   Years      PLS Holts_method    ARIMA

## 1  2021 5540.173     4924.616 4902.246

## 2  2022 5759.914     5036.940 5008.228

## 3  2023 5984.354     5149.264 5114.209
```

Thus, we can assume that the refinery capacity will only increase for the years to come and expect that India will become a central refinery hub for the South Asian countries. India's actual refinery capacity is a lot higher than is being currently utilised. So, we can expect that even if the demand of the refineries increases in the future, the Indian Oil Refinery industry will be able to keep up with the demands and exceed the refinery capacities of other countries.

OIL CONSUMPTION

Asia is the world's largest and fastest-growing consumer of energy as well as the largest emitter of CO2. The continent has become the biggest threat due to a single factor - the complete domination of the fossil fuel industry, with coal as the primary source of energy.

- For oil, the international energy outlook 2019 report projects that refineries throughout Asia will increase by 60% between 2018 and 2050, as seen above. Oil demand in the region to go up to 9 million barrels per day by 2040 from 6.5 million barrels per day today. The monthly consumption of petrol and diesel by the country hit an all-time high during March 2022, even as retail prices of the two key commodities shot up to their highest levels for that month.
- India's consumption growth was aided by economic activity inching back to normalcy as well as common man and industries stocking on the key commodities in anticipation of the government resuming the fuel price mechanism post Assembly elections in five States.
- India's demand for high-speed diesel (HSD) stood at 7.71 million tonnes (MT) during March this year, which is the second highest consumption by the country, barring May 2019 (7.79 MT). Petrol consumption stood at 2.91 MT — the highest ever for the key fuel.

Oil being refined at a refinery in the United States (Source: Bureau of Labour Statistics)

China is the largest consumer of oil in the Asia-Pacific region with the country consuming 14,225 thousand barrels of oil per day in 2020. In the same year, India consumed around 4,669 barrels of oil per day making it the second largest consumer of oil in the region.

India's oil demand is projected to jump 8.2 per cent to 5.15 million barrels per day in 2022 as the economy continues to rebound from the devastation caused by the pandemic.

We will now analyse the oil consumption of India and draw some inferences from it.

Now, we divide the data into training and testing set. The training set has data from 1965 to 2008 and test set has data from 2009 to 2020.We fit Holt's model on the training data to fit the model.

```
## Holt-Winters exponential
smoothing with trend and without
seasonal component.

## Call:

## HoltWinters(x = train, gamma =
F)

## Smoothing parameters:

##  alpha: 1

##  beta : 0.3909699

##  gamma: FALSE

## Coefficients:

##        [,1]

## a 3073.5431

## b  137.1437
```

The following is the plot of Holt's smoothing vs the training data:

We plot the forecasts of the Holt model and the actual test values to get an idea of our model.

Most of the test data falls in the 80% confidence interval, indirectly implying that the model is a good fit. We now perform the KPSS test and ADF test to check for stationarity. We also fit an ARIMA model to the training data and check the model fit against the training data.

```
## # KPSS Unit Root Test #

## Test is of type: mu with 3 lags.

## Value of test-statistic is:
1.1396

## Critical value for a
significance level of:

##                 10pct  5pct
2.5pct  1pct

## critical values 0.347 0.463
0.574 0.739
```

```
## [1] 2   #differencing of lag 2
required

##  Augmented Dickey-Fuller Test

## data:  train

## Dickey-Fuller = -0.34014, Lag
order = 3, p-value = 0.9845

## alternative hypothesis:
stationary
```

<u>The training data is not stationary.</u> We will now fit an ARIMA model on the training data and find the best model based on the AIC values.

```
##  ARIMA(0,2,0)                    : 443.4713

##  ARIMA(0,2,1)                    : 438.7659

##  ARIMA(0,2,2)                    : 438.0476 …

##  ARIMA(5,2,0)                    : 447.6212

##  Best model: ARIMA(0,2,2)
```

The ARIMA model vs the training data and the forecasts vs test data are shown on the graphs below:

Following the above two methods, we fit regression models to the training data and try to find the best regression model.

We can see that the 2nd Order Regression fits the model well. Below is the table which stores all the test values and the predicted values from the different models:

```
## [1] "The predicted values by the models and the test data is shown in the
table below:"

##    Year     Test Holts_Method    ARIMA LS_Predict PLS_Predict H_Predict

## 1  2009 3232.727     3210.687 3194.010   2654.510    3187.990  2654.510

## 2  2010 3308.084     3347.830 3314.160   2717.730    3322.341  2717.730 …

## 12 2020 4669.078     4719.267 4515.663   3349.928    4835.940  3349.928

##    LMS_Predict LTS_Predict LAD_Predict S_Predict MM_Predict

## 1     1635.537    1632.325    2596.459  1903.627   2601.203

## 2     1667.994    1664.783    2659.901  1943.348   2663.071 …

## 12    1992.571    1989.359    3294.314  2340.559   3281.751
```

We now calculate the RMSE values to find the best models:

```
##                Holt's Smoothing    ARIMA Huber Loss     LAD      LMS
LS

## RMSE of models          279.689 380.2165   1200.581 1253.44 2367.621
1200.581

##                     LTS MM-estimator      PLS S-estimator
```

```
## RMSE of models 2370.75     1258.825 251.4869    2062.786
## [1] "The minimum RMSE values among all the models is:"
## [1] 251.4869
```

We now perform Ljung-Box Test, runs test and residual plots to find the goodness of fit for the selected models, PLS, Holt's and ARIMA.

```
##  Box-Ljung test
## data:  fitPLS$residuals
## X-squared = 27.881, df = 1, p-value = 1.29e-07
#Project Compiled by SOMBIT GHOSH
```

```
##  Approximate runs rest
## data:  fitPLS$residuals
## Runs = 8, p-value = 4.735e-06
## alternative hypothesis: two.sided
```

From the above test values, we can see that <u>the 2nd degree regression model is not a good fit for the data.</u> We do the same tests for the Holt's and ARIMA model.

```
##  Box-Ljung test

## data:  resid(expo2)

## X-squared = 0.70218, df = 1, p-
value = 0.4021

##  Approximate runs rest

## data:  resid(expo2)

## Runs = 19, p-value = 0.3486

## alternative hypothesis:
two.sided
```

We can see that Holt's model is a good fit for the training data.

Now we analyse the ARIMA model for the goodness of fit.

```
##  Ljung-Box test

## data:  Residuals from
ARIMA(0,2,2)

## Q* = 9.2228, df = 7, p-value =
0.2371

## Model df: 2.   Total lags used:
9

##  Approximate runs rest

## data:  resid(acons)

## Runs = 21, p-value = 0.5418

## alternative hypothesis:
two.sided
```

The above tests show that second-order linear regression is not a good fit. But we will still use it to predict for comparison's sake. Also, the ARIMA model is a good fit for the training data.

```
## [1] "The predicted values for 2021,2022 and 2023 predicted by the
different models are:"

##   Years      PLS Holts_method    ARIMA

## 1  2021 5212.406     4286.504 4190.581

## 2  2022 5397.202     3903.929 3712.084

## 3  2023 5585.457     3521.354 3233.587
```

As India is the world's third largest oil consumer, it is expected that the oil consumption will increase in the upcoming years. The projected demand is set to rise up 10 million barrels per day by 2030 in Asia alone. So, it's high time we move to alternative sources of energy or use our fuels prudently.

NATURAL GAS CONSUMPTION

Over the past decade, gas demand growth has been driven largely by the power and industrial sectors, as rapidly growing economies and populations require increasing volumes of energy and living standards improve. Environmental and economic efficiency gains from natural gas have led to an ongoing displacement of more carbon-intensive fossil fuel sources of energy like coal and oil. Burning of natural gas for energy results in fewer emissions of nearly all types of air pollutants and carbon dioxide (CO2) than burning coal or petroleum products to produce an equal amount of energy. It is expected that natural gas consumption growth will continue over the next decades due to economic competitiveness, national climate pledges, and new natural gas capacity infrastructure coming online while aging coal and oil-fired power plants enter retirement phase.

Components of Natural Gas (Source: Dialog Group)

Asia lacks a pre-existing gas-user culture of the kind found in North America and Europe, where an infrastructure for distributing gas was developed long ago to provide light, and later heat. Asia, on the other hand, developed much of its centralized energy infrastructure later, when electricity distribution systems were used to meet most energy needs, so it has lacked a gas-distribution system upon which modern gas use could build. For most Asian countries, natural gas is locally produced and consumed. Unlike oil, there is no international gas market to which Asian countries can link their domestic natural gas prices. In some countries, local gas prices are loosely linked with the prices of fuel oil. In others, including China and India, natural gas prices are determined and regulated by the governments, often set at low levels to benefit industrial sectors or to subsidize the residential sector in areas adjacent to natural gas fields. Excessive government intervention in natural gas pricing has discouraged exploration, development, and production of natural gas in many Asian countries, leading to less natural gas consumption.

A Natural Gas Processing Plant in Austria (Source: Wikipedia)

Below is a chart which compares the natural gas consumption of different Asian countries:

Natural gas consumption of Asian countries has significantly increased over the years. China is leading consumer in Asia followed by South Korea and India. But consumption of gas decreased in Japan.

Below we analyse the natural gas consumption of India and try to predict the demand of gas for the future 3 years. The following chart shows the natural gas demand in India year wise. It is clearly visible that it has been increasing over the years because of rise in population.

Now, we divide the data into training and test set to better analyse the data. The training data contains values from 1965 to 2008 and test set contains values from 2009 to 2020.

We fit Holt's model on the training data:

```
## Holt-Winters exponential
smoothing with trend and without
seasonal component.

## Call:

## HoltWinters(x = train, gamma =
F)

## Smoothing parameters:

##  alpha: 1

##  beta : 0.3107889

##  gamma: FALSE

## Coefficients:

##        [,1]

## a 39.987666

## b  1.949935
```

The following chart shows the Holt's Model values and the training data:

We now forecast values for the length of the test set and plot it with the actual test data to get a general idea of the model.

```
## # KPSS Unit Root Test #

## Test is of type: mu with 3 lags.

## Value of test-statistic is:
1.1069

## Critical value for a
significance level of:
##                 10pct  5pct
2.5pct  1pct

## critical values 0.347 0.463
0.574 0.739
```

We check if any differencing is required for the data to make it stationary. In this case we need to difference the data twice.

```
## [1] 2

## Augmented Dickey-Fuller Test

## data:  train

## Dickey-Fuller = -0.59504, Lag
order = 3, p-value = 0.9718

## alternative hypothesis:
stationary
```

From both the KPSS test and the ADF test, we see that the data is not stationary. We now fit an ARIMA model to try and make the model stationary. Following are different combinations which were tried and the best model was found using AIC as a criterion.

We now fit the best ARIMA model based on the least AIC values.

```
##  ARIMA(0,2,0)                    : 104.2112

##  ARIMA(0,2,1)                    : 91.7081

##  ARIMA(1,2,1)                    : 93.63158….

##  Best model: ARIMA(1,2,0)
```

The following chart shows the ARIMA model and the training data to get a general idea of the fit of the model.

We now plot the test values and the forecasted values on the same chart to understand the model accuracy:

Below is the chart which compares the different regression models vs the training data. From the regression line, we can guess the best regression model for the training data:

Below are the actual test values and the predicted values from the different methods using which we will find the RMSE of the models and fit the best model to the whole data.

```
##    Year     Test Holts_Method    ARIMA LS_Predict PLS_Predict H_Predict
## 1  2009 49.10808     41.93760 42.17146   32.40658    42.42313  32.40658
## 2  2010 58.96162     43.88754 43.77837   33.31976    44.67184  33.31976 …
## 12 2020 59.60083     63.38688 62.02572   42.45157    70.35272  42.45157
##    LMS_Predict LTS_Predict LAD_Predict S_Predict MM_Predict
## 1     35.99999    36.72130    33.15137  36.94708   31.70930
## 2     37.27616    38.03064    34.11682  38.27133   32.60903 …
## 12    50.03783    51.12402    43.77135  51.51391   41.60636
```

We now calculate the RMSE values for the different models to find the best models.

```
##              Holt's Smoothing ARIMA  Huber Loss  LAD      LMS       LS
## RMSE of models 7.377753    7.203841 17.53623  16.57192 12.54494 17.53623
##                   LTS       MM-estimator   PLS       S-estimator
## RMSE of models  11.77571   18.27183      8.938505    11.52209
## [1] "The minimum RMSE values among all the models is:"
## [1] 7.203841
```

We can see that the minimum RMSE is for the ARIMA model. But we analyse the residuals of the best 3 models by their ACF, noise plot, Ljung-Box test and runs test to check the fit of the models.

```
##  Ljung-Box test                          ## Q* = 11.996, df = 8, p-value =
                                            0.1514
## data:  Residuals from
ARIMA(1,2,0)                                ## Model df: 1.   Total lags used:
                                            9
```

<u>We can see that the p value is 0.1514, which means that the residuals are not autocorrelated, signifying that the model is a good fit.</u>

```
##  Approximate runs rest                   ## alternative hypothesis:
                                            two.sided
## data:  resid(acons)

## Runs = 23, p-value = 1
```

Runs test p-value is also indicating that the sequence of residuals is random, which implies that the model is a good fit. We perform the same tests for the other two models.

We will now analyse the Holt's model:

```
##  Box-Ljung test

## data:  resid(expo2)

## X-squared = 0.14003, df = 1, p-value = 0.7082

##  Approximate runs rest

## data:  resid(expo2)

## Runs = 19, p-value = 0.3486

## alternative hypothesis: two.sided
```

The tests show that the Holt's model is also a good fit for the training data. We now analyse the 2nd order regression.

```
##  Box-Ljung test

## data:  fitPLS$residuals

## X-squared = 29.411, df = 1, p-value = 5.856e-08

##  Approximate runs rest

## data:  fitPLS$residuals

## Runs = 9, p-value = 1.945e-05

## alternative hypothesis: two.sided
```

The above results indicate that the 2nd Order linear Regression is not a good model since the residuals are not unautocorrelated. But for the sake of comparison since its RMSE value is the smallest, we will forecast values using this model for the future 3 years viz. 2021,2022 and 2023. But Holt's Method is a better model as it provides more weight to the recent data than the past data. Thus, we will use it to forecast too since it will be beneficial in the long run

We now fit the above 3 selected models on the whole data and predict the future 3 years data.

Predicting the consumption output for the future 3 years using the models selected:

```
## [1] "The predicted values for 2021,2022 and 2023 predicted by the different methods are:"

##   Years   PLS     Holts_method   ARIMA

## 1  2021 68.18076     59.95134   59.57797

## 2  2022 70.73159     60.30186   60.46427

## 3  2023 73.32862     60.65237   61.50506
```

Thus, we can assume that the consumption will increase in coming years. India will require more imports of natural gas or find alternative sources of gas like biogas, etc. We should also be judicious about the gas usage and not waste LPG (cooking gas) or CNG by using personal means of transport irresponsibly. We should also be aware of the fact that using petroleum products increase the amount of greenhouse gases in the atmosphere.

CONCLUSION

The main idea behind this project was to see how this important yet difficult to obtain commodity called crude oil / black gold (Rudis Oleum in Latin) affects the daily life of people through various of its factors like price, import, consumption and others.

From the above analyses, all the variables are pointing towards increasing magnitude of values, whether it be the price, oil consumption, import quantities, etc. This is *not* a good sign. An increase in prices of crude oil means that the common man has to pay more money for the same quantity of oil than he used to, which is not possible since the mean increase in salary in India is not adjusted to the rise in inflation rates. This also means that transportation and manufacturing of goods become expensive since diesel prices will increase too, and mostly diesel is used for industrial purposes to generate energy along with electrical energy. This further implies that basic necessities will become costlier as cost of transportation increases, which is basically what inflation is, rise in prices of goods and services in an economy.

Increase in oil consumption means burning more crude oil, which increases the level of greenhouse gases, since burning crude oil products gives out carbon dioxide (CO_2), carbon monoxide (CO), etc. Hence, we should take care to not overuse crude oil products. Increase in oil consumption also implies digging up more oil to keep up with the demand. This is very bad for the environment in the sense that faster pumping of crude oil will lead to oil drilling plans being disrupted and oil being extracted in unplanned and harmful ways.

The increase in consumption of natural gases can be linked to the population boom of the late 1900s. Use of CNG in cars, LPG for transportation as well as cooking has increased the gas consumption to unthinkable levels. Indian government's plans to build pipelines for natural gas might bring the demand and prices under control but as of now, India still needs to import most of the required natural gas.

Hence, to keep up with the demands of crude oil and natural gas, India needs to import the required quantities from OPEC+ countries. Thus, people should take care and judiciously use petroleum products so that we can conserve petroleum for us and the future generations as well.

BIBLIOGRAPHY

- https://www.bp.com
- https://asb.opec.org/data/ASB_Data.php
- https://groww.in/blog/where-does-india-get-its-oil-supply
- https://en.wikipedia.org
- https://www.researchgate.net/publication/266599562
- Alternative Methods of Regression by David Birkes, Yadolah Dodge
- Introduction to Time Series Analysis and Forecasting by Douglas Montgomery
- The Analysis of Time Series: An Introduction by Chris Chatfield
- https://www.investopedia.com
- https://www.google.com
- www.machinelearningmastery.com and many more sites and books have helped us gain knowledge to analyse the data better.

LIMITATIONS and FURTHER SCOPE

In this project/ report, we could only analyse the basic and most important factors of crude oil like the price of crude oil, oil consumption, oil import, oil refineries and natural gas consumption. But there are many more factors like Gas Production, Oil Energy and Coal Energy as a comparative study, and Oil Reserves and Oil Production can be factors under analysis.

Therefore, in a future project, all these variables can be analysed along with the ones analysed in the project to give better analysis and prediction as a whole. Also, the prices can be analysed with the effect of other variables like taxation, import duties policies, changes in OPEC policies, diplomatic policy changes between countries, USD-INR exchange rates, etc.

There is ample scope for future research in this field because India's major oil companies like Indian Oil Corporation Ltd. (IOCL), Hindustan Petroleum Corporation Ltd. (HPCL), Oil and Natural Gas Corporation (ONGC), etc. have a whole Research and Development (R&D) section dedicated for these kinds of research.

So, yeah there is a future ahead of us in the field of Petroleum and Natural Gas until the time that we go for alternative and better sources of energy.