A Project Report On

Oil and Natural Gas Production

Submitted by:

Sumit Gangwar	**21535032**
Ashish Kumar	**21911003**
Shubham Kumar Verma	**21535028**
Rajesh Singh Negi	**21535023**
Md Shahriar Tasjid	**21535013**
Mondira Ghosh	**21535017**

Under Guidance and Supervision of-
Dr. Durga Toshniwal

Department of Computer Science and Engineering
Indian Institute of Technology Roorkee
Roorkee, Haridwar
Uttarakhand- 247667

Declaration Certificate

We hereby certify that the work which has been presented in the Project entitled "**Oil and Natural Gas Production**", submitted in the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, is an authentic record of our work under the supervision of **Dr. Durga Toshniwal**.

Name:

Date: 26/03/2022

Sumit Gangwar
Ashish Kumar
Shubham Kumar Verma
Rajesh Singh Negi
Md Shahriar Tasjid
Mondira Ghosh

Acknowledgement

In the completion of our project on "**Oil and Natural Gas Production**" We would like to convey our special gratitude to **Dr. Durga Toshniwal** of the Computer Science and Engineering Department at Indian Institute of Technology Roorkee.

Your valuable guidance and suggestions helped us in various phases of the completion of this project. We will always be thankful to you in this regard.

We are ensuring that this project was finished by us and not copied.

Name:

Sumit Gangwar
Ashish Kumar
Rajesh Singh Negi
Shubham Kumar Verma
Md Shahriar Tasjid
Mondira Ghosh

Contribution

Name	Contribution	Page No
Mondira Ghosh	Data Collection	9
Rajesh Singh Negi	Data Extraction and Cleaning	10
Md Shahriar Tasjid	Data Transformation	11
Shubham Kumar Verma	Exploratory Data Analysis	12
Sumit Gangwar	Linear Regression Model	16
Ashish Kumar	Random Forest Model	17

Table of Contents

Table of Figures

Introduction

India is at present is the fastest growing major economy of the world. The robust growth in the economy has triggered the energy demand and India is positioned to drive the incremental demand growth in the global energy arena. GDP growth at constant (2011-12) prices has averaged 7.3 per cent for the period from 2015-16 to 2017-18, which is the highest among the major economies of the world. Real GDP, viz., GDP at constant prices for the year 2017-18 is estimated at Rs 130.11 lakh crore showing a growth rate of 6.7% over the year 2016-17 of Rs 121.96 lakh crore. The growth is against a backdrop of resurgence in exports coupled with global tensions on trade front and volatile global oil market with a distinct hardening crude oil price. On the domestic front, the strong growth is underpinned by robust private consumption, expectation of greater stability in GST and public investment as well as ongoing structural reforms. Demand for petroleum products having tapered off from double digits has stabilized in the region of 5.3% in 2017-18 and 2016-17 vis-à-vis corresponding period of last year.

Oil and Natural Gas Production

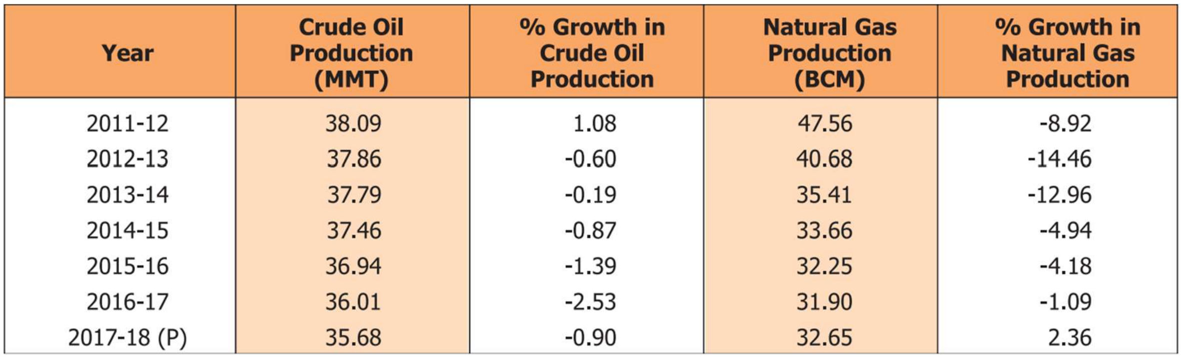
In India, the crude oil production during the year 2017-18 is at 35.68 million Metric Tons (MMT) as against production of 36,01 MMT in 2016-17, showing a decline of 0.9%. Production by OIL is mainly from matured fields where decline rate encountered was more than expected, less than planned contribution from work-over wells due to bandhs, blockades, miscreant activities which contributed to direct loss of production.

Natural Gas production during the year 2017-18 is at 32.65 billion Cubic Meters (BCM) which is 2.36% higher than production of 31.90 BCM in 2016-17. Shortfall in production by OIL was mainly attributed to less than planned production due to decline in associated gas production.

In the oil and gas industry, data and process mining would enable the following:

1. Discover previously unknown and possibly useful relationships in data, thus improving the understanding of the plant equipment, systems, operations, people, etc.
2. Forecast usage patterns and determine sustainable modes of operations at a granular level. This helps with accurate modeling, estimating and calculating plant throughput.
3. Track and predict oil and gas demand by using macro indicators, such as weather and production units/volume at a given time period.
4. Provide an objective, concrete, consistent and repeatable approach to analyzing the data continuously.

More than ever before, companies are dealing with significantly larger sets of data with more varied content, and therefore, need a big data strategy. Similarly, processes have grown complex with bottlenecks, deviations, long throughput times, and other roadblocks



Year	Crude Oil Production (MMT)	% Growth in Crude Oil Production	Natural Gas Production (BCM)	% Growth in Natural Gas Production
2011-12	38.09	1.08	47.56	-8.92
2012-13	37.86	-0.60	40.68	-14.46
2013-14	37.79	-0.19	35.41	-12.96
2014-15	37.46	-0.87	33.66	-4.94
2015-16	36.94	-1.39	32.25	-4.18
2016-17	36.01	-2.53	31.90	-1.09
2017-18 (P)	35.68	-0.90	32.65	2.36

Figure 1 Crude Oil and Natural Gas Production

Data Mining Process

1) Data Collection

The data for this study is a primary oil dataset from the U.S Energy Information administration. It lists the oil prices (in dollars) from 1987-2017 in a daily frequency fashion. The following graph shows the changes in oil prices with the passing of each year.

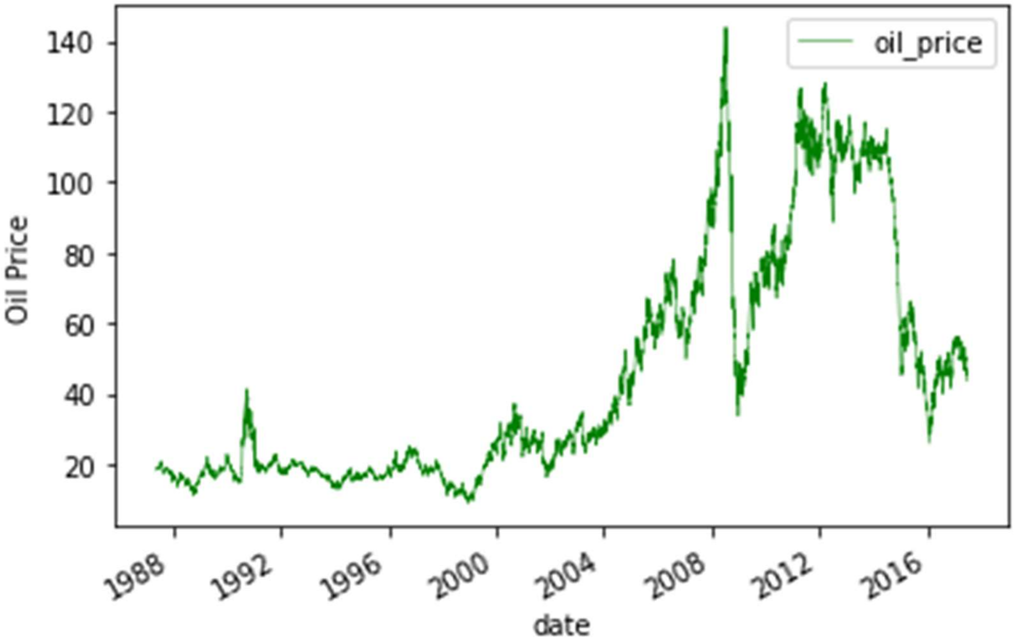


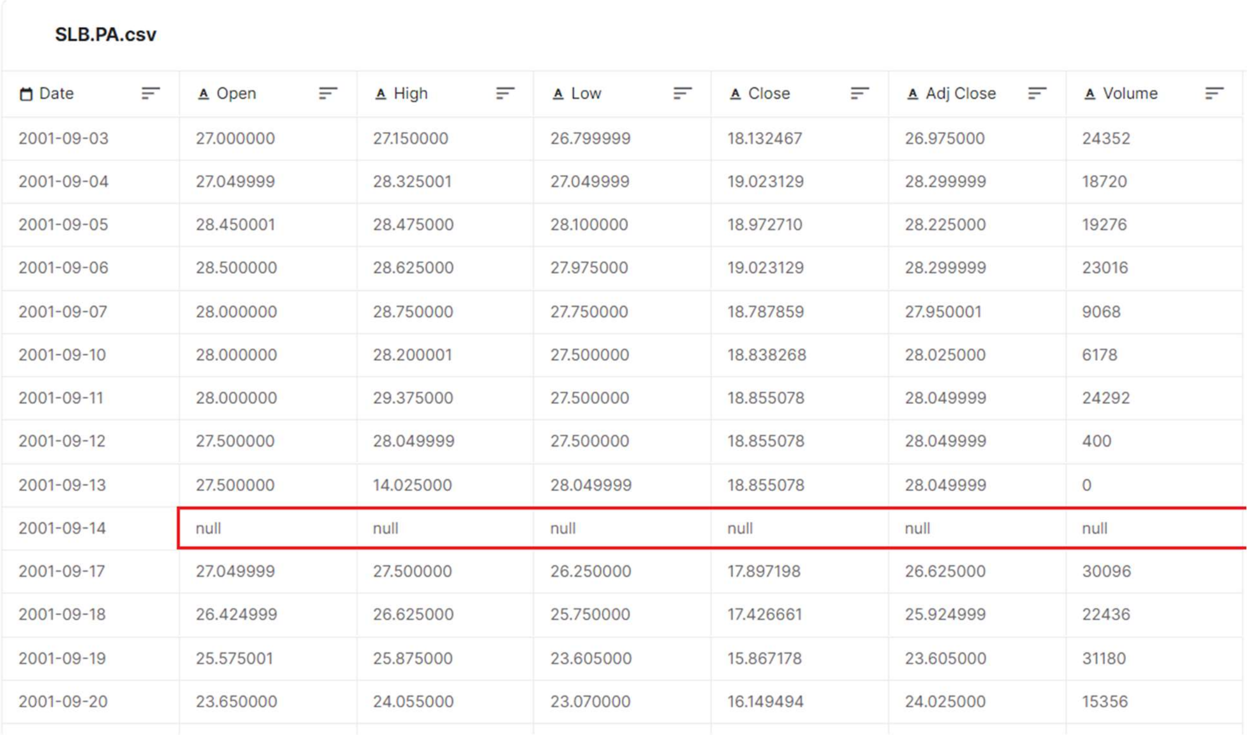
Figure 2 Oil Price vs Date

This dataset is additionally followed by the Share price dataset from Yahoo Finance in a daily frequency from the following companies.

1. Premier Oil
2. Cairn Energy
3. Total Oil
4. Stat Oil
5. Royal Dutch Shell
6. Engie

2) Data Extraction and Cleaning

All these datasets are publicly available in Kaggle. Thus, we have two different datasets: the first contains the oil prices in time series format and the second contains share prices information of top oil companies in a similar time series fashion. For a meaningful analysis, the data from these two sources will need to be merged and transformed into a new single dataset



SLB.PA.csv

Date	Open	High	Low	Close	Adj Close	Volume
2001-09-03	27.000000	27.150000	26.799999	18.132467	26.975000	24352
2001-09-04	27.049999	28.325001	27.049999	19.023129	28.299999	18720
2001-09-05	28.450001	28.475000	28.100000	18.972710	28.225000	19276
2001-09-06	28.500000	28.625000	27.975000	19.023129	28.299999	23016
2001-09-07	28.000000	28.750000	27.750000	18.787859	27.950001	9068
2001-09-10	28.000000	28.200001	27.500000	18.838268	28.025000	6178
2001-09-11	28.000000	29.375000	27.500000	18.855078	28.049999	24292
2001-09-12	27.500000	28.049999	27.500000	18.855078	28.049999	400
2001-09-13	27.500000	14.025000	28.049999	18.855078	28.049999	0
2001-09-14	null	null	null	null	null	null
2001-09-17	27.049999	27.500000	26.250000	17.897198	26.625000	30096
2001-09-18	26.424999	26.625000	25.750000	17.426661	25.924999	22436
2001-09-19	25.575001	25.875000	23.605000	15.867178	23.605000	31180
2001-09-20	23.650000	24.055000	23.070000	16.149494	24.025000	15356

Figure 3 Null Values in Data

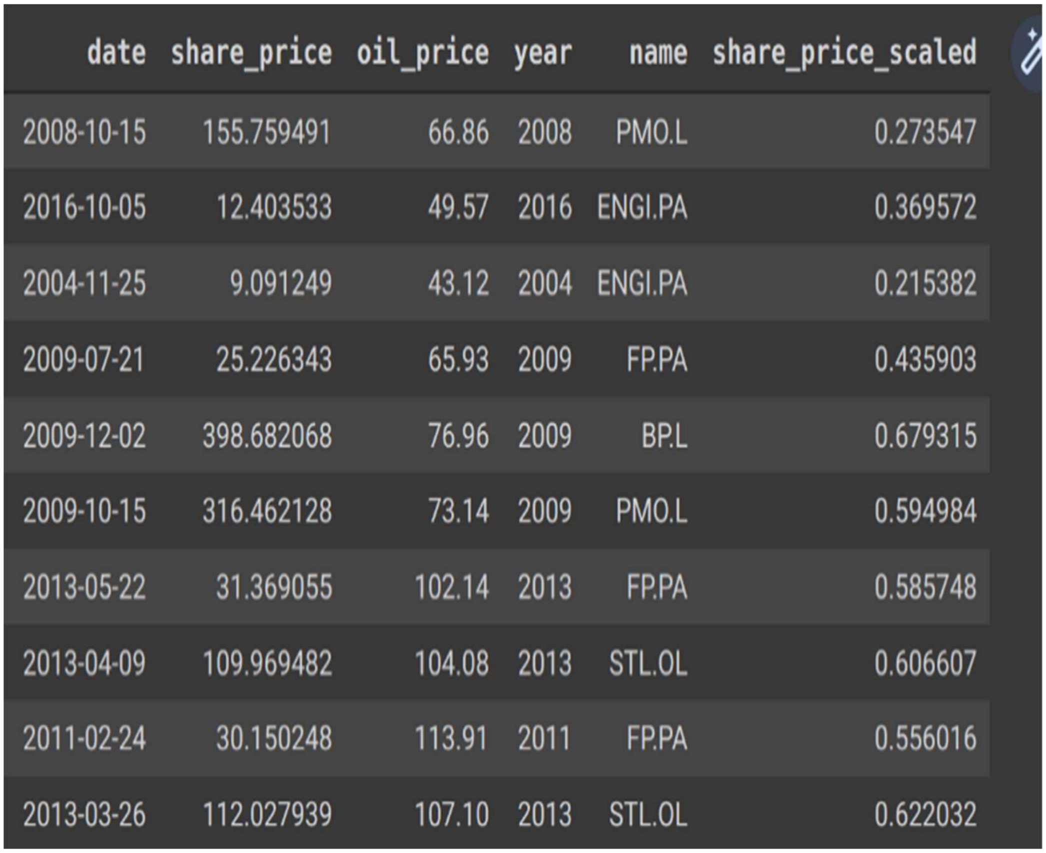
In Data Cleaning step ,we eliminate all the duplicates rows and rows that contain bad data. We replace all null data by zeros to prevent the problem when aggregationg the data.

We also filter out non-essential attributes.Nominal and numeric attribute types are identified along with class attribute.We combine tables for some tasks.

The next step Data Transformation will discuss these details.

3) Data Transformation

For our analysis purpose we will create a master data frame that will contain the information of oil price and share price of each company for the daily frequency series. We will merge the Oil Prices dataset with the Share price dataset for each company based on a joining condition which is ‘Date’ in our case. Additionally, we will drop the rows for which there is no merging possible i.e., the rows which yield *NAN* after merging. A bare minimum feature engineering by creating a new feature called *‘share_price_scaled’* which contains every share price scaled in the range between 0 to 1. The below figure shows a random sample of size 10 from the transformed dataset.



date	share_price	oil_price	year	name	share_price_scaled
2008-10-15	155.759491	66.86	2008	PMO.L	0.273547
2016-10-05	12.403533	49.57	2016	ENGI.PA	0.369572
2004-11-25	9.091249	43.12	2004	ENGI.PA	0.215382
2009-07-21	25.226343	65.93	2009	FP.PA	0.435903
2009-12-02	398.682068	76.96	2009	BP.L	0.679315
2009-10-15	316.462128	73.14	2009	PMO.L	0.594984
2013-05-22	31.369055	102.14	2013	FP.PA	0.585748
2013-04-09	109.969482	104.08	2013	STL.OL	0.606607
2011-02-24	30.150248	113.91	2011	FP.PA	0.556016
2013-03-26	112.027939	107.10	2013	STL.OL	0.622032

Figure 4 Data Transformation

4) Exploratory Data Analysis

To begin with analysis, we plot the scatter plot between *Oil Price* and *Share Price* for each of the companies for the last two years of data. Only last two years data was used to check for patterns in recent past and for lowering down the scale of analysis. From the scatter plot we check for correlations if any between the oil prices and share prices. It is a prior knowledge that share price of a company increase with the rise in Oil prices. Looking at the scatter plots we want to infer whether there exists any linear correlation if any between these two quantities. If linear correlation exists, then we can model the change in Share price as a linear function of Oil Price i.e., we can make use of Linear Regression to model this change. Following graphs shows the scatter plots for the various oil companies under consideration.

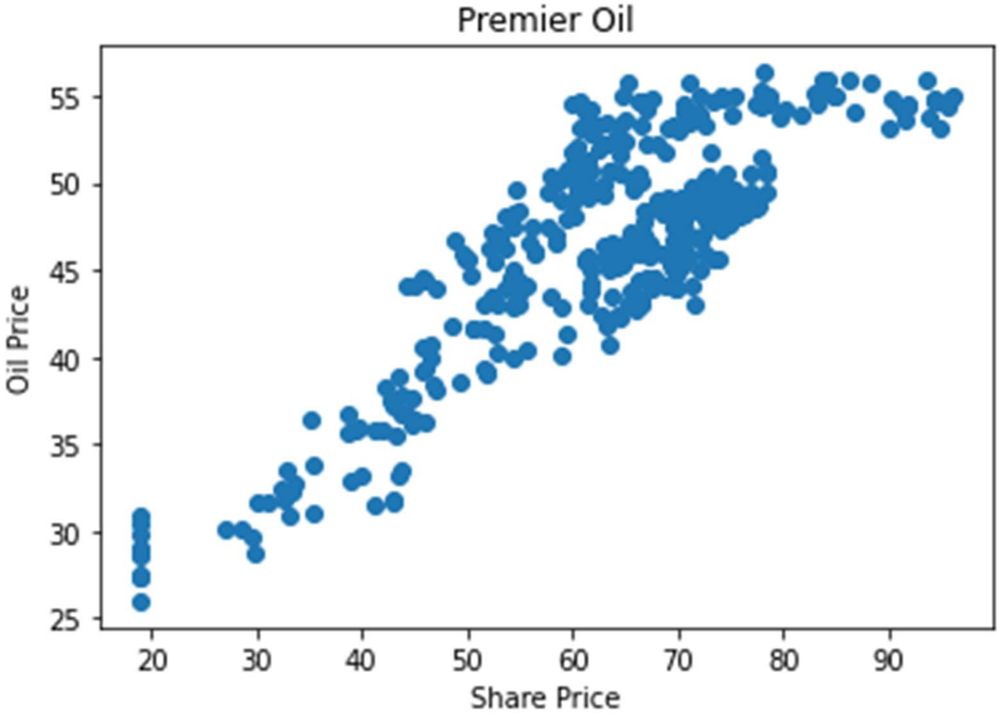


Figure 5 Premier Oil

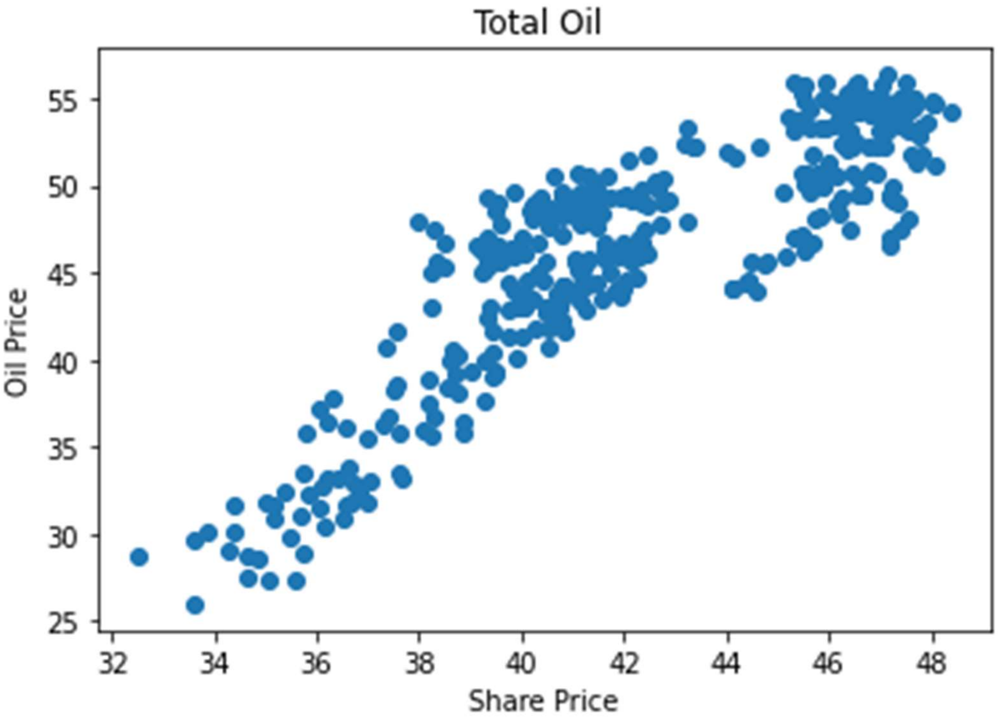


Figure 6 Total Oil

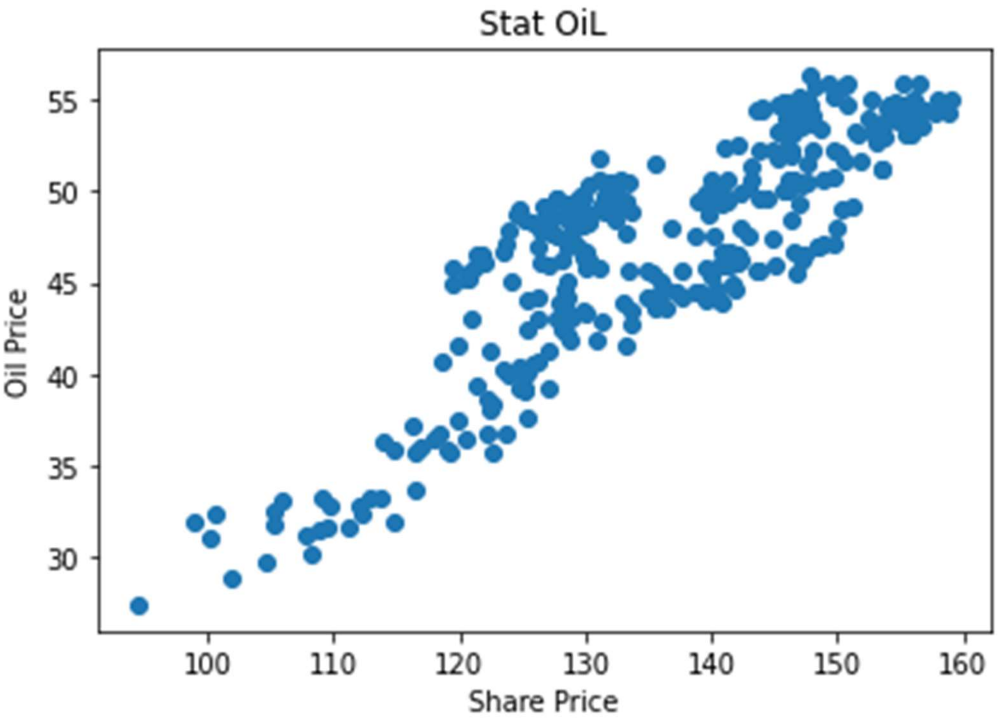


Figure 7 Stat Oil

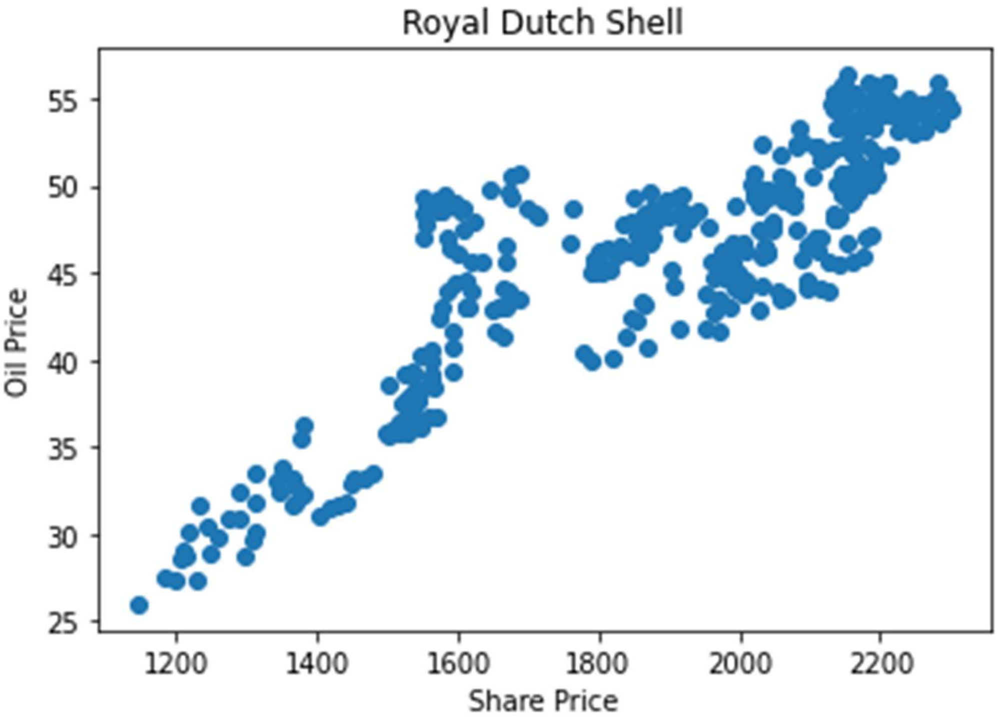


Figure 8 Royal Dutch Shell

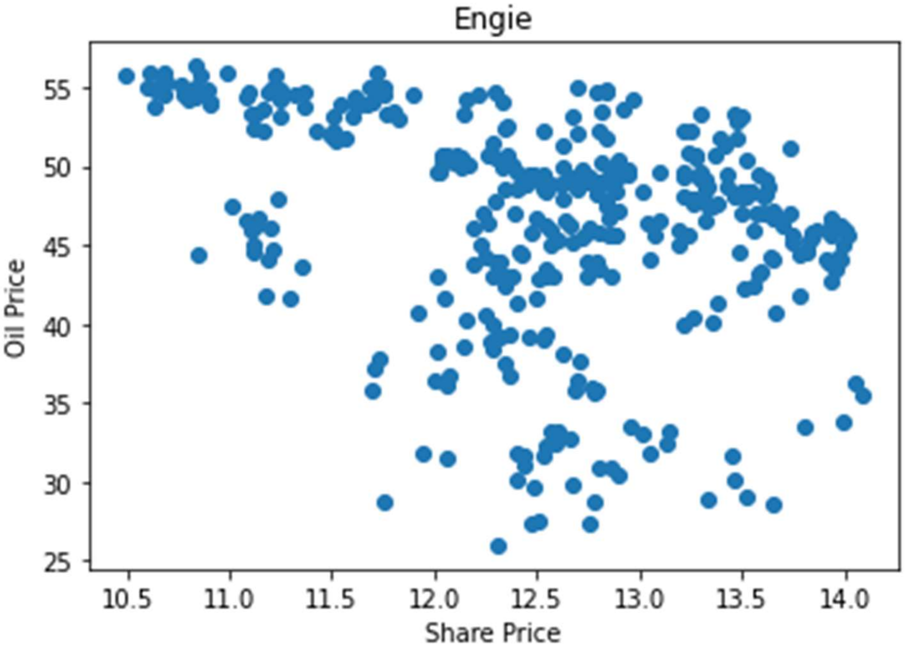
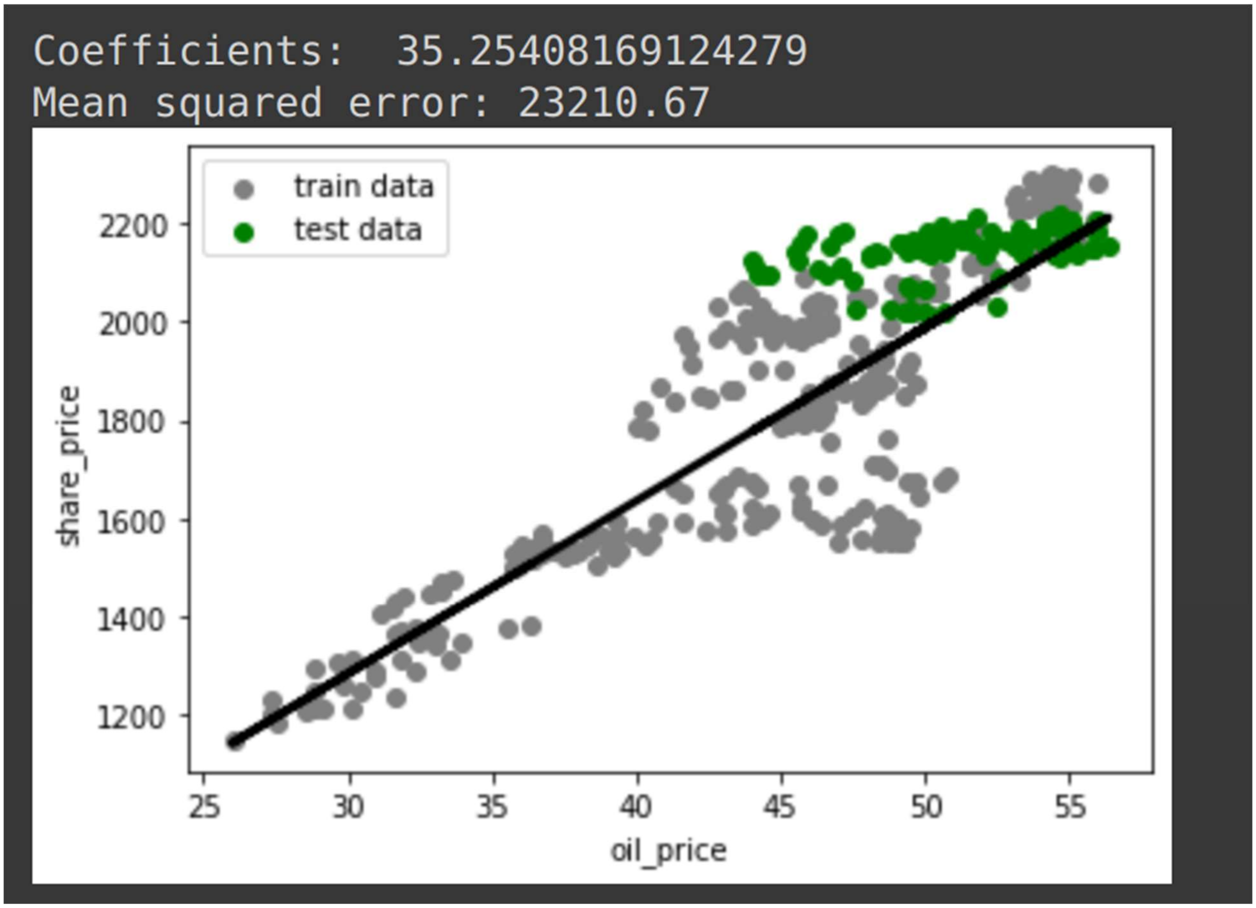


Figure 9 Engie Oil

There seems to exist some sort of correlations between share Price and Oil Price for all the companies except the Engie. However, this correlation does not seem to be a linear correlation. We will further verify this by fitting a linear regression model on this data and evaluating it fit on test data. For the data corresponding to the Engie company there does not exist any correlation so fitting a linear regression model on this data is not at all meaningful.

Linear Regression Model

We will construct a simple linear regression model using supervised learning. The objective is to evaluate the prediction of data from the last 100 days using data trained from years 2016/17 (excluding test data). Train data is the data used to construct the model and test data is the data we are trying to predict. Below figure shows the model trying to predict Share Price from the Oil Price for the Royal Dutch Shell Company.



```
Coefficients:  35.25408169124279
Mean squared error: 23210.67
```

Figure 10 Linear Regression Model

As the figure suggests the line is a poor fit for the data and there exists extremely high mean squared error (23210) in our fitted model. This brings us to a conclusion that no linear correlations exist in the data otherwise we would not have arrived at such a poor fit for our data. Similar regression model was tried for other Oil companies' data which showed promising correlations in the previous section. However linear regression is a poor fit in all these cases. Therefore, we have omitted their fitted results here.

Random Forest Model

Random forest is an ensemble tool which takes a subset of observations and a subset of variables to build a decision trees. It builds multiple such decision tree and amalgamate them together to get a more accurate and stable prediction. Random forest algorithm accepts more than one variable in the input data to predict the output. It runs very efficiently on large databases, it's very accurate, can handle many input variables, it has effective methods for estimating missing data and many more advantages. The main disadvantage is overfitting for some tasks or some sets of data.

In top of the oil price, we are going to use other variables to predict the share price of Shell. These are going to be the prices of Premier Oil, Cairn Energy, TOTAL, ENGIE and STATS Oil. The model looks good just predicting the training data. With quite a bit of overfitting.

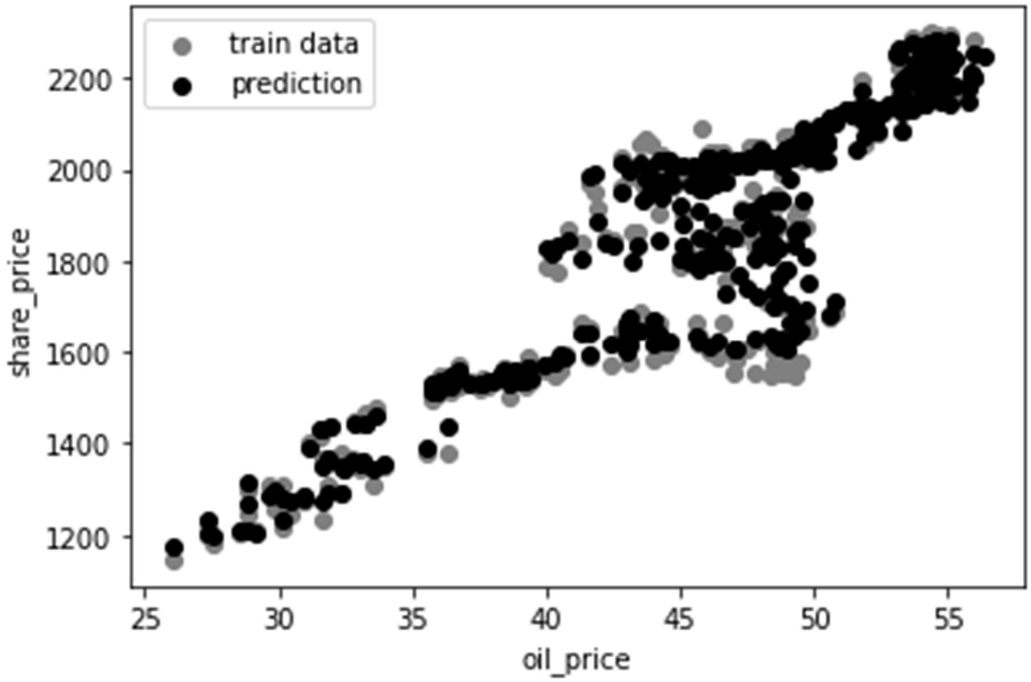
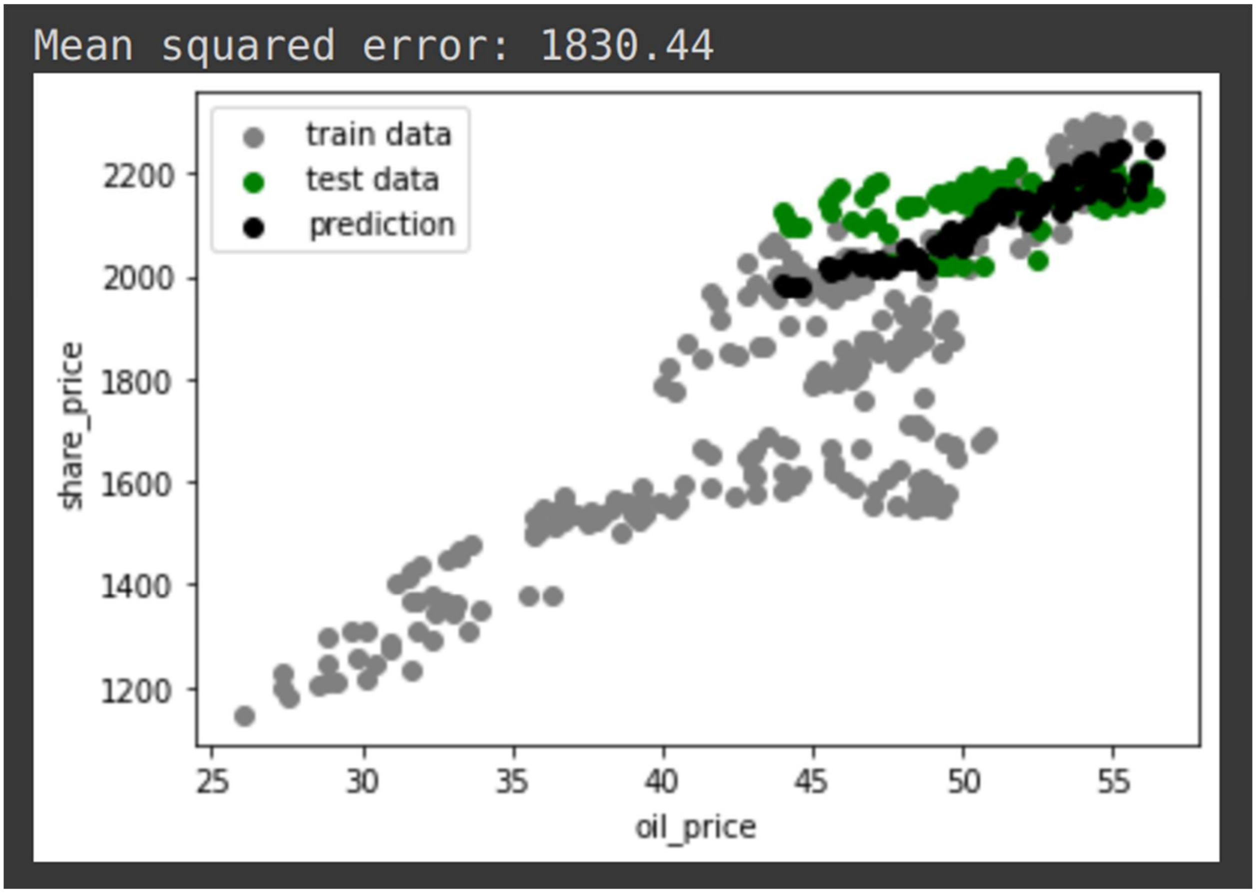


Figure 11 Random Forest on Training Data

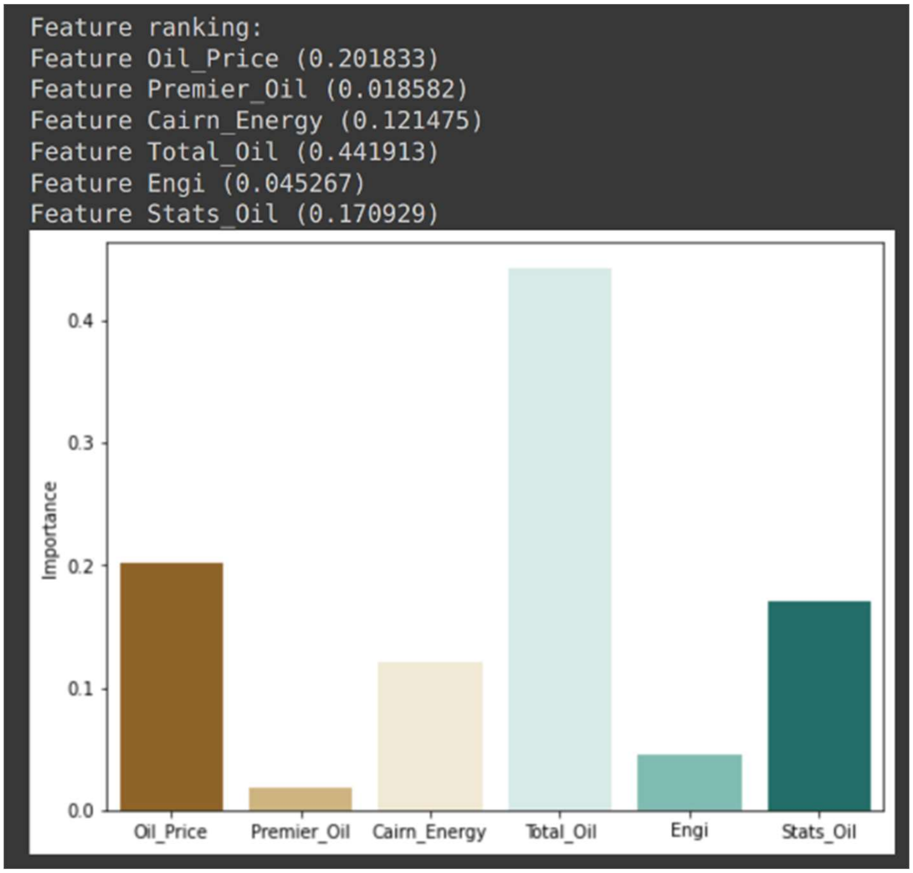
The prediction on the test data looks much better now, still somehow inaccurate for lower oil price environment. If you see the mean squared error, we manage to reduce the error from 23210 to 2709. That is 10 times lower than using linear regression.



```
Mean squared error: 1830.44
```

Figure 12 Random Forest on Test Data

Lastly, we can make use of Random Forest model to describe the relevancy of independent features involved in our predictions. The feature importance suggests that share price of Total Oil companies and Oil prices play significant role (almost 65 %) in the prediction of the share price of Royal Dutch Shell Company.



```
Feature ranking:
Feature Oil_Price (0.201833)
Feature Premier_Oil (0.018582)
Feature Cairn_Energy (0.121475)
Feature Total_Oil (0.441913)
Feature Engi (0.045267)
Feature Stats_Oil (0.170929)
```

Figure 13 Feature Ranking

Conclusion

We have performed Data Mining on Oil and Natural Gas Production Successfully

References

1) Oil price dataset from the U.S Energy Information administration.

2) Share price dataset from Yahoo Finance in a daily frequency from the following companies: