STATISTICAL ANALYSIS OF CRUDE OIL PRICE

www.dataintelligenceandmagic.com

15-Dec-2022

CONTENTS

- OBJECTIVE
- MOTIVATION
- METHODOLOGY
- DATA
- TRAINING A MODEL
- ESTIMATION

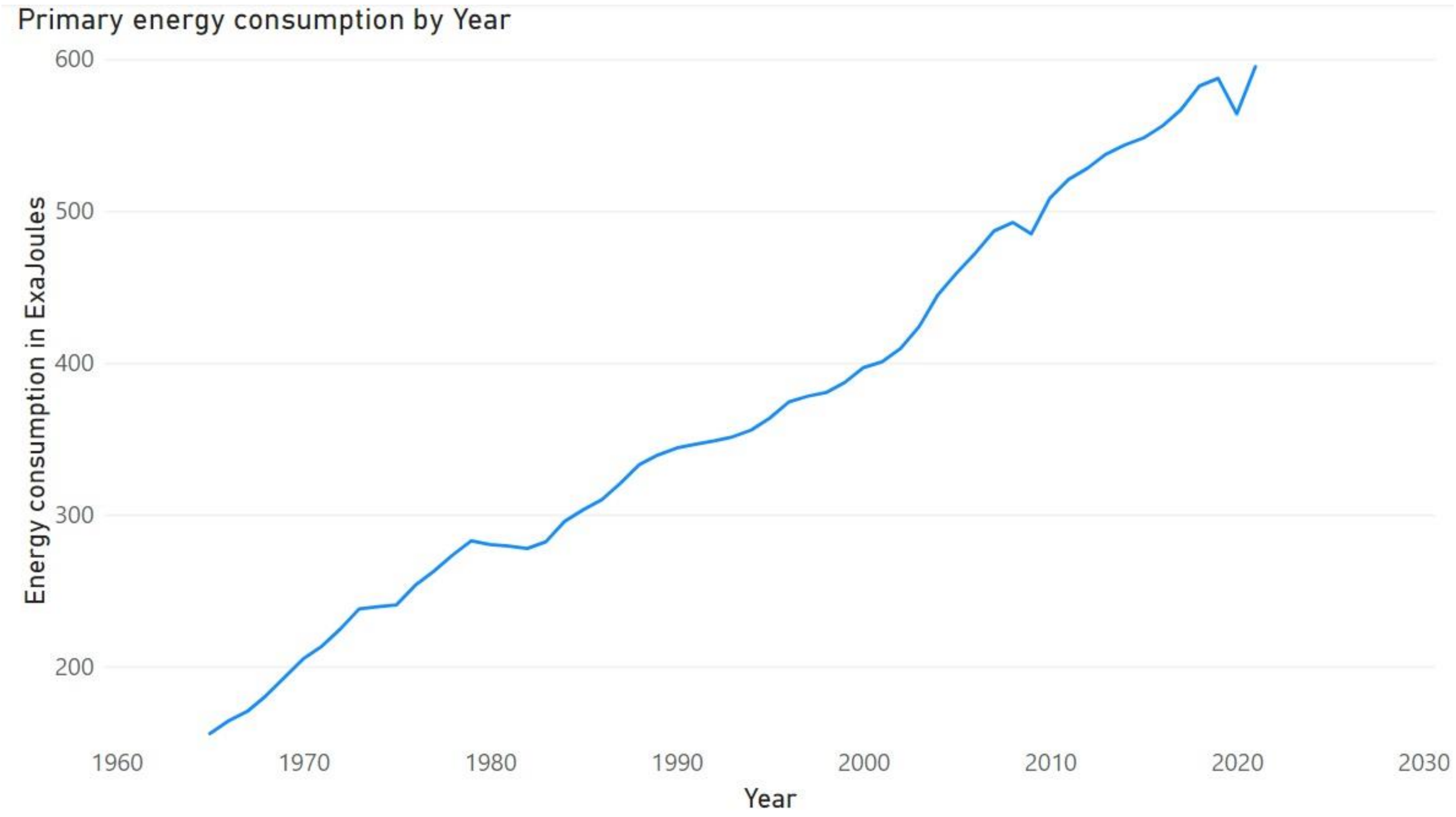
OBJECTIVE

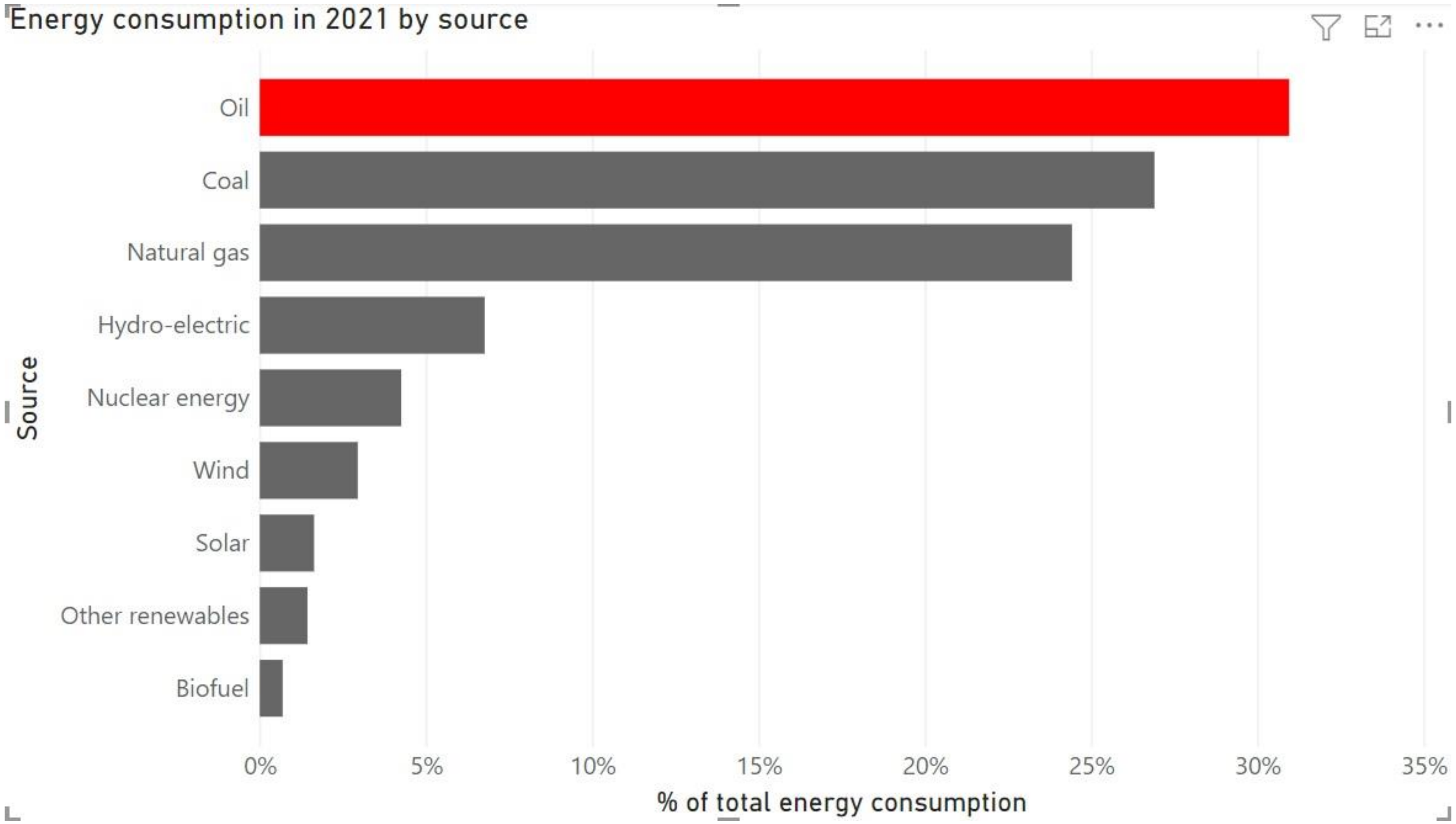
To perform a statistical analysis of historical crude oil prices and determine whether crude oil was accurately valued during the year 2021

MOTIVATION

Crude oil has been a major source of primary energy for human activities for many decades. The demand for energy has been increasing forever and the total energy consumption for the year 2021 stood at 595.15 Exajoules (1 EJ = 277.778 Terawatt hours). Crude oil accounted for almost 31% of the primary energy consumed during the year 2021.

ENERGY CONSUMPTION OF THE WORLD





For electricity generators using non-fossil fuel sources, primary energy consumption is calculated on an input equivalent basis - i.e. based on the equivalent amount of fossil fuel input required to generate that amount of electricity in a standard thermal power plant. For example, in the case of a Solar electricity generator with an output of 100 Terawatt hours (TWh), the primary energy consumption is 100/0.4 = 250 TWh (assuming the efficiency of a thermal power plant is 40%)

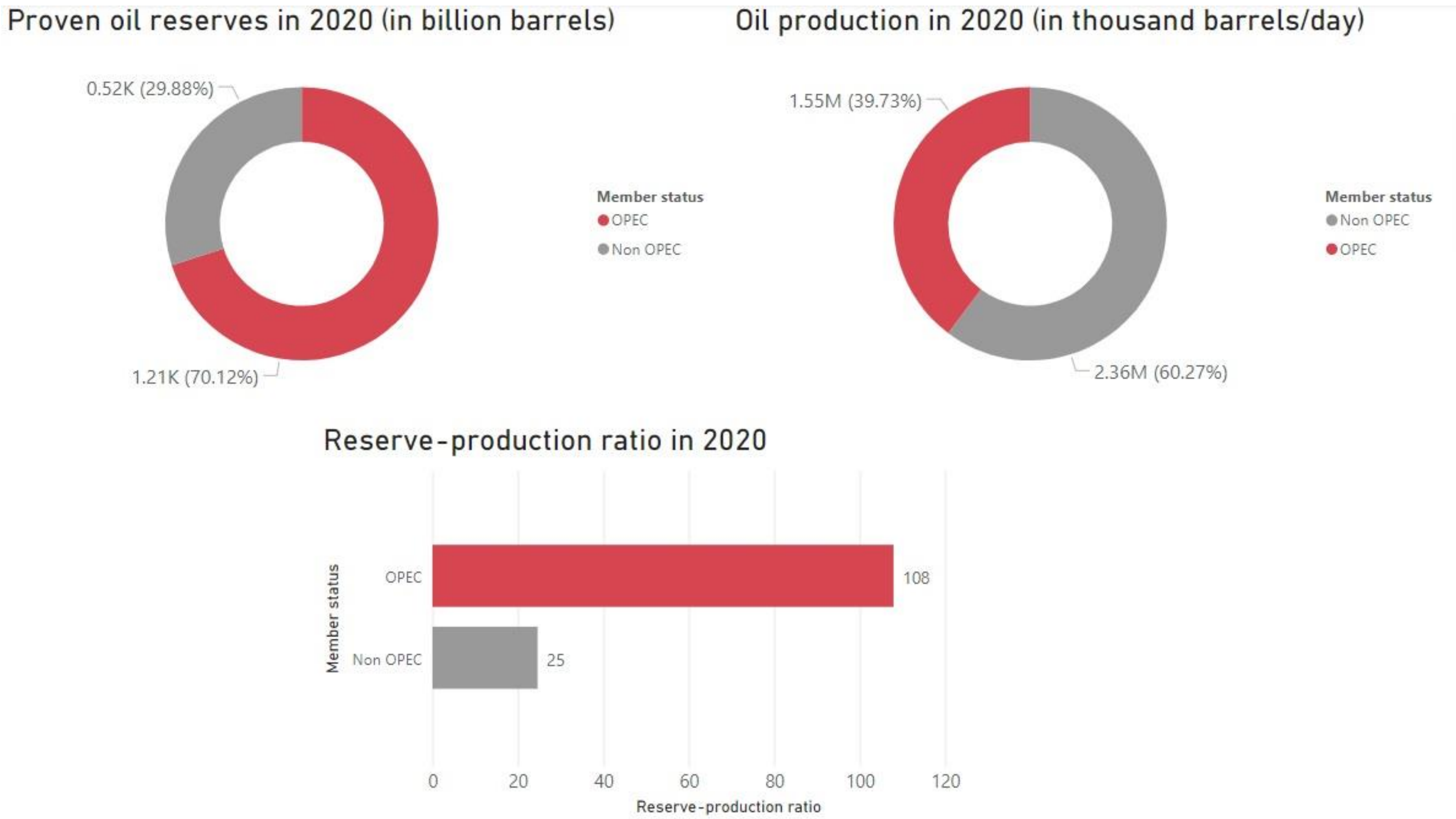
Even though there is strong urge to transition from fossil fuels to renewable sources of energy, oil is expected to remain as the most consumed primary energy source for many more decades in the future. So, it is necessary to analyze whether global markets have priced oil correctly considering our ever increasing demand for energy.

METHODOLOGY

LET'S DIVE IN

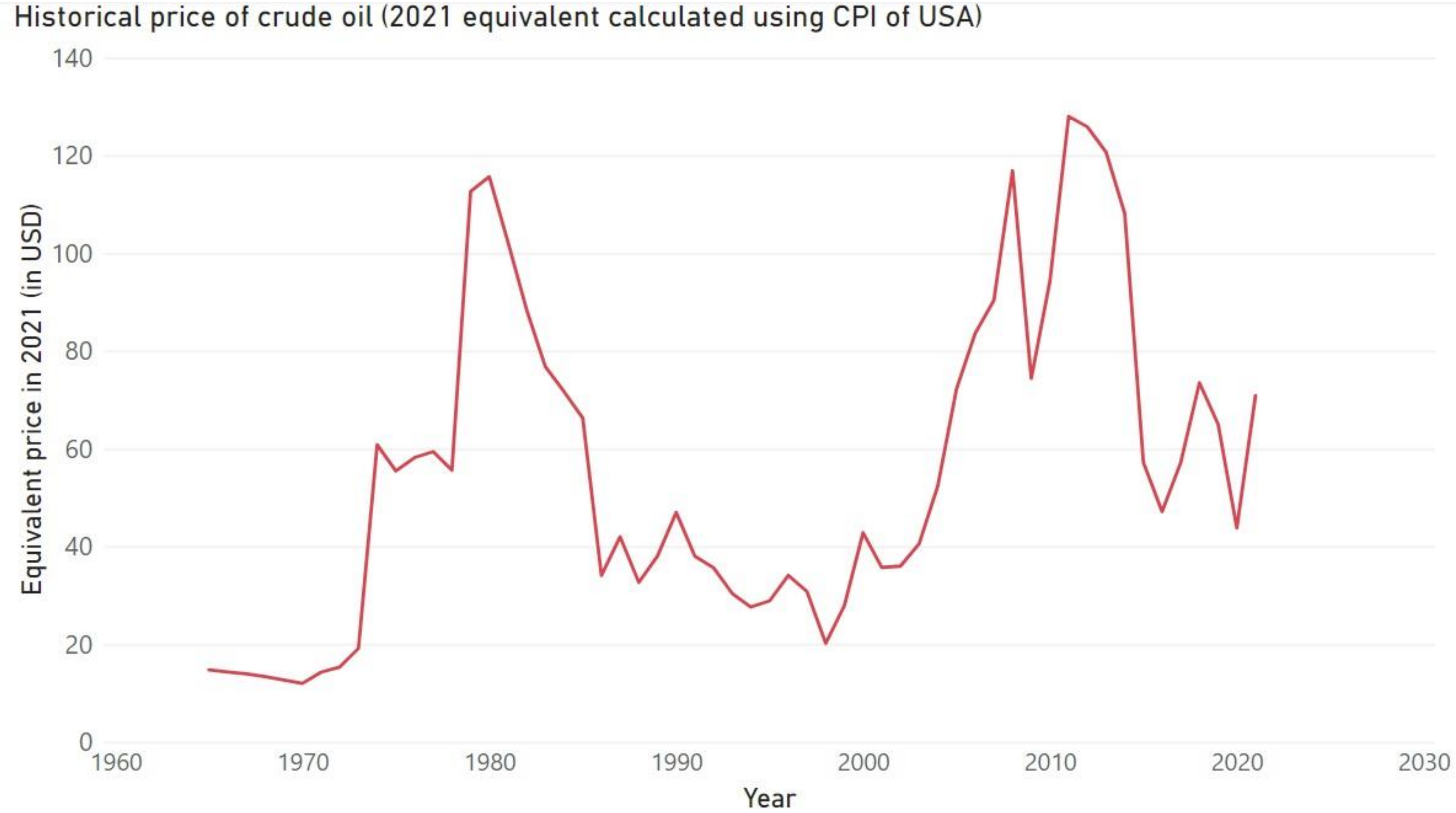
METHODOLOGY

This is not a Time-series analysis of the price of crude oil and the reason for not using a time-series method is the peculiar behavior of crude oil prices. The members of the OPEC are the major producers of oil and they exercise significant influence on its price in the global markets. Instead of letting the market to decide the oil price based on the demand-supply metrics, the price of oil in most cases is a value which the OPEC members are willing to trade for depending on several factors including geo-political issues. Due to the significant portion of oil reserves held by OPEC members, it is easy for them to influence the supply of oil in the global markets; they are rightly called the oil cartel.



The reserve-production ratio is an indicator of how many years the proven reserves will last at the given rate of oil production.

Also, the price of crude oil as a commodity has not followed the behavior of the prices of other traded commodities. Since 1980 the cost of many essential goods, commodities, and indexes have increased while crude spot price has remained relatively flat by comparison. Between 1980 and 2020, average house price has grown by 79.8%, a loaf of bread has risen by 77.2%, the S&P 500 surged 98.6%, and gold has increased 64.6%. Over the same period oil has had a 1.6% upward movement.



Due to the afore-mentioned reasons, we are not using a time-series analysis methodology for this problem. Rather this analysis tries to find the relationship between a bunch of predictor variables and price of crude oil, the response variable. This bunch of predictors includes oil consumption (in EJ), proven oil reserves (in billion barrels), reserve-to-production ratio (an indicator of how long the reserves will last), consumption of renewables - solar, wind and others(including geothermal and biomass) and their relative proportions to the total energy consumption.

To be specific, the analysis tries to estimate the relationship between the 5-year rolling average of oil price and the 5-year rolling averages of afore-mentioned predictor variables. The rolling averages are used instead of yearly values to reduce the impact of short-term fluctuations caused due to geo-political issues like what happened in 1973 (Arab oil embargo) and 1979 (Iranian revolution).

In this methodology, we try to estimate the nature of relationship between the different predictors and oil price for the past several years. The parameters of the relationship are estimated using a machine learning model which is then used to predict what should have been the price of oil in 2021.

For training the ML model, we use the values of predictors and oil prices from 1965 to 2020. For the oil price, we use the inflation adjusted price of crude oil from 1965 to 2020 using the CPI inflation rate in the USA. Doing so will effectively result in a model that tries to estimate the oil price for different values of predictors, as if time was frozen. Additionally, new engineered features that represent the trend information of the different predictors are also used to train the model. For example, one of the newly engineered features represents whether the percentage of solar energy consumption in a particular year has increased or decreased in comparison to the previous year.

```
ion:absolute;z-index:99
x 5px #ccc}.gbrtl .gbm
;display:block;positio
pacity:1;*top:-2px;*le
/;top:-4px\0/;left:-6
ne-box;display:inline
display:block;list-s
ne-block;line-height
pointer;display:bl
tive;z-index:1000}
padding-right:9px
```

DATA

DATA

The main source of data used for this analysis is the publication 'Statistical Review of World Energy' by British Petroleum, the latest edition of which was released in June-2022. The detailed report and data-sets can be accessed at the below weblink.

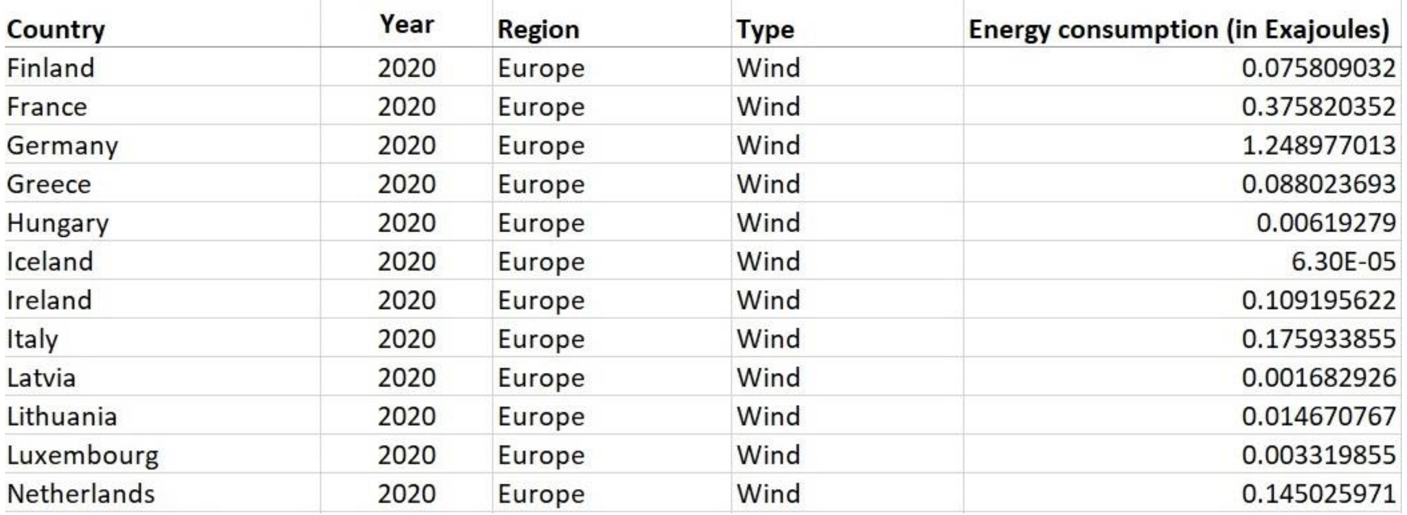
https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html

This data was supplemented with the data on proved oil reserves (1965-1979) published by OPEC, which can be accessed at the below weblink.

https://asb.opec.org/data/ASB_Data.php

Both the data-sets didn't require cleaning as these are the official publications from two major organizations. But it required substantial pre-processing to arrive at a form suitable for training the ML model. A subset of this processed data which represents the primary energy consumption of the World for every year from 1965 to 2021 segregated by country and fuel source can be accessed at the below Kaggle repository.

https://www.kaggle.com/datasets/nirmalprasad/world-energy-consumption



Country	Year	Region	Type	Energy consumption (in Exajoules)
Finland	2020	Europe	Wind	0.075809032
France	2020	Europe	Wind	0.375820352
Germany	2020	Europe	Wind	1.248977013
Greece	2020	Europe	Wind	0.088023693
Hungary	2020	Europe	Wind	0.00619279
Iceland	2020	Europe	Wind	6.30E-05
Ireland	2020	Europe	Wind	0.109195622
Italy	2020	Europe	Wind	0.175933855
Latvia	2020	Europe	Wind	0.001682926
Lithuania	2020	Europe	Wind	0.014670767
Luxembourg	2020	Europe	Wind	0.003319855
Netherlands	2020	Europe	Wind	0.145025971

The data pre-processing was performed using Python and the Jupyter notebook can be found at the below GitHub repository.

https://github.com/code-nirmalprasad/Statistical-Analysis-of-Crude-Oil-price.git

All the required data of predictors and oil price was extracted from the two afore-mentioned data-sets. The oil price considered for this analysis is as given below:

1965-1983 → Arabian Light posted at Ras Tanura
1984-2021 → Brent

```
ion:absolute;z-index:99
x 5px #ccc}.gbrtl .gbm
;display:block;positio
pacity:1;*top:-2px;*le
/;top:-4px\0/;left:-6
ne-box;display:inline
display:block;list-s
ne-block;line-height
pointer;display:bl
tive;z-index:1000}
padding-right:9px
```

TRAINING A MODEL

TRAINING A MODEL

The objective of training a model is to estimate the price of crude oil for 2021 based on historical data. This estimated price is compared to the actual price in 2021 to determine whether crude oil was accurately priced. So, the target (or response) variable in this problem is crude oil price.

For the oil price, we use the inflation adjusted price of crude oil from 1965 to 2020 using the CPI inflation rate in the USA. Doing so will effectively result in a model that tries to estimate the oil price for different values of predictors, as if time was frozen.

To perform a statistical analysis and estimate the price for 2021, a set of predictor variables is required. The following data is compiled for each year between 1965 and 2020.

THE BASE PREDICTORS

- Energy consumption (in EJ) in the form of Oil
- Energy consumption (in EJ) - Solar
- Energy consumption (in EJ) - Wind
- Energy consumption (in EJ) - Other renewables (Geothermal, Biomass)
- Proven oil reserves (in billion barrels)
- Oil production (in thousand barrels/day)

Using the above features, following new features are engineered.

- Percentage of energy consumed in the form of oil (against total energy consumption for a year)
- Percentage of energy consumed - Solar
- Percentage of energy consumed - Wind
- Percentage of energy consumed - Other renewables
- Reserve-production ratio

Reserve-production ratio is the ratio of proven oil reserves to the oil production in a year. It is an indicator of how many years the oil reserves will last at the given rate of production.

The afore-mentioned variables are the predictors used to estimate the price of crude oil. A machine learning model is trained to find the relationship between the set of predictors and the crude oil price. These predictor variables are not used as they are. Instead, their 5-year rolling averages is used to train the machine learning model. This approach is chosen to reduce the impact of short term fluctuations and cut the noise in the data-set. So the oil consumption for year 2010 is represented as the 5 year rolling average of the consumption between 2006 and 2010.

To enrich the data-set, the trend information of each predictor variable is also extracted which is the percentage change in the predictor value in comparison to the previous year. This information represents whether a predictor variable is having an upward or downward trend.

The data processing results in the following engineered features (predictors) for years from 1965 to 2020.

THE ACTUAL PREDICTORS

- 5-year average of energy consumption in the form of Oil
- 5-year average of energy consumption - Solar
- 5-year average of energy consumption - Wind
- 5-year average of energy consumption - Other renewables
- 5-year average of proven oil reserves (in billion barrels)
- 5-year average of reserve-production ratio
- 5-year average of percentage of energy consumption in the form of oil
- 5-year average of percentage of energy consumption - Solar
- 5-year average of percentage of energy consumption - Wind
- 5-year average of percentage of energy consumption - Other renewables

In addition, we also use the trend information for each of the afore-mentioned engineered features.

The data preprocessing was performed using Python and the Jupyter notebook can be found at the below GitHub repository.
https://github.com/code-nirmalprasad/Statistical-Analysis-of-Crude-Oil-price.git

THE MODEL

Because we are using rolling average values for the predictors, we need to use the rolling average of oil price as well. Effectively, the trained model tries to estimate the relationship between rolling averages of predictor and response variables.

Before training the model, a standardization of the predictor variables is performed. An XGBoost regressor model is trained using the enriched data-set (data-points for years from 1965 to 2020). The trained model is later used to estimate the 5-year average price of crude oil for 2021. XGBoost regression is a method of ensemble learning using decision trees and the objective function of the model is chosen as 'squared error' with number of estimators as 50.

x 5px #ccc}.gbrtl .gbm
display:block;positio
pacity:1;*top:-2px;*l
;top:-4px\0/;left:-6
e-box;display:inline
display:block;list-s
e-block;line-height
pointer;display:bl
tive;z-index:1000;
padding-right:9px

ESTIMATION

ESTIMATION

After the model was trained using data-points for years from 1965 to 2020, the set of predictors for the year 2021 was supplied to the model which estimated oil to be priced between USD 70 and USD 75. Please note that the price estimated by the model is the 5-year average and not the actual price for 2021.

The average spot price of Brent in 2021 was in fact USD 70.91 and the 5-year average (2017-2021) was USD 62. This shows that the model estimates the spot price of Brent (from 2017 to 2021) to be more than what it was traded at.

The low price of crude oil in 2020 can be justified because of the Covid-19 pandemic and the dull economic outlook. But oil was undervalued during the years from 2017 to 2019.

WHAT’S NEXT

LOOKING AHEAD

WHAT'S NEXT

The model under discussion was not trained using an indicator that represents the economic outlook. The economic outlook is a major factor deciding the energy demand of the World. The next step in this statistical analysis is to include the outlook for the world economy as an additional predictor in the model. This should provide us a more accurate estimation of crude oil price.

THANK YOU

☎ +1 (438) 525-9554

✉ connect.nirmalprasad@gmail.com

www.dataintelligenceandmagic.com