Project Report – Dillon Constantine

# GitHub URL

<https://github.com/dconstantine22/UCDPA_DillonConstantine>

# Abstract

(Short overview of the entire project and features)

1. Purpose of Study and Research problems you investigated.
2. Design/ Layout of the study.
3. Trends found as a result.
4. Summary of interpretations / conclusion

This report is an exploratory data analysis into petrol prices worldwide. With the petrol prices crisis that is currently unfolding right here in Ireland there was an opportunity to further understand the problem at hand worldwide. The report looks at trends in the data, looking into the top oil consumers, most affordable petrol prices worldwide and the countries that have the most gross domestic product per capita. The study took shape using an open-source petrol prices dataset that was readily available for use. The analysing was carried out using various python libraries such as Pandas and NumPy and visualisation libraries; matplotlib, seaborn and plotly.

# Introduction

(Explain why you chose this project use case)

Current inflation in Ireland and cost of living rising particularly hot topic the rise in Petrol and Diesel prices.

Problems Statements

* Top / Lowest 10 Oil Consumers
* Prices of Petrol (top end and low end)
* Top GDP per capita
* Highest Yearly Gallon consumption per capita
* Possibility of Modeling?

# Dataset

The dataset that was used in this project contained information on the petrol/oil prices worldwide (Usmani, 2022). The open-source dataset was collected by Zeeshan Ul Hassan Usmani and is available on Kaggle.com, an online collaborative sharing platform (Uslu, 2022). According to the metadata, (Usmani, 2022), of this dataset it was collected using Google API from public sources such as International Monetary Fund (IMF), the United Nations and World Bank. Note throughout this report oil and petrol are the key words referenced when talking about the data.

The original data set included 11 fields in total and 181 rows of data.

The fields were;

* Serial Number (Used as the Index Column)
* Country
* Daily Oil Consumption in Barrels
* World Share of Oil Consumption
* Yearly Gallon Per Capita
* Price Per Gallon in USD
* Price Per Liter in USD
* Price Per Liter in PKR (Pakistani Rupee)
* Gross Domestic Product (GDP) Per Capita
* Gallons GDP Per Capita Can Buy
* Number Times the GDP Per Capita can be Purchased of Petrol

The reasons behind the selection of this dataset were simple. The price of petrol/oil is a current topic, so it is on the mind of a lot of people. With the prices continuously rising right here in Ireland (O'Sullivan, 2022), looking into price trends worldwide was of interest. The dataset contained a good sample size with 181 countries included, clearly defined fields and the data was up to date as of June 2022.

# Implementation Process

This section of the report with step through the process in which the exploratory data analysis of Worldwide Petrol Prices took place from beginning to end.

## Step 1: Importing the Libraries

After selecting the dataset, the python packages that were going to be used throughout this study needed to be installed and imported. There were five libraries used in this study NumPy, Pandas to work with the data and then visualisation libraries such as Matplotlib, Seaborn and Plotly.

The following is some background on the selection of the libraries used and their main purpose in this exploratory analysis of petrol prices worldwide.

* The NumPy library is used within Python to work with numerical data (Developers, 2008-2022).
* The Pandas library is used for data analysis to generate data structures that make it easy and intuitive when working with different data types (Development, 2008-2022).
* The Matplotlib and Seaborn packages are used in this study to generate visual graphs to depict the data while the use of Plotly allows for the creation of interactive graphs.

## Step 2: Importing the Dataset

The dataset selected in this study was a csv file so using the Pandas library was ideal. Pandas has a function called read\_csv () which imported the data into a pandas data structure called df\_ppww (Data Frame Petrol Prices Worldwide).

Initially a ‘*UnicodeDecodeError’* error occurred when importing the dataset without any parameters within the pd.read\_csv("Petrol Dataset June 23 2022 -- Version 2.csv").

This error was due to the data requiring a larger width encoding parameter to be used such as ‘latin-1’ or ‘ISO-8859-1’. The ‘latin-1’ encoding parameter was used in this study. In the import statement the ‘Serial Number’ (S#) field was also defined as the index column as it contained unique values for each row of data (i.e., Each country).

Table

Description automatically generated

## Step 3: Data Understanding & Cleaning

It was important to look behind the data being used in this study. As it was a real-world data set there was potential to be flaws in the data.

### Shape and Data Types

To start the shape and data types of each field were established. A conversion was required on some object fields to convert them to ints and floats. The reason of the data type being object was due to special characters (‘%’, ‘,’) being included in the data. Included in this stage was a quick fix on spelling mistakes in the field names. Once the conversion and column name fixes were carried out a check was done to confirm the change.

### Create / Drop Fields

The next stage of this process was to create any new fields for analysis or drop unnecessary fields. To investigate the ‘daily liter price’ per country a new column was needed. This column was generated by taking the Daily Oil Consumption which was in barrels and multiplying by the quantity of litters in a barrel, 159L according to (Termeer, 2013), and then multiply by the price per liter.

*(Daily Oil Consumption (Barrels) \* 159 Liters) \* Price Per Liter (USD)*

Also, in this stage an unnecessary field to this study was dropped from the dataset, Price Per Liter (PKR). The reason behind this was this study uses the United States Dollar (USD) throughout making the price per liter in Pakistan Rupees an unnecessary field.

### Null & Unique Value Check

A null value check was carried out next to check for any missing data. This check came back clear, and all data was accounted for as seen below in figure.

A picture containing table

Description automatically generated

Another check was carried out to check for duplicate entries in the data and again this came back clear there was no duplications of the represented countries. Evidence of this can be seen in the figure below.

Table

Description automatically generated

### Descriptive Statistics

The final part of this step was to return the basic descriptive statistics of the petrol prices data. The output of this gives a clearer picture of the data as the min, max and mean of each field is returned as well as other useful information.

For instance, some of the basic stats of the Petrol Prices Worldwide date are;

* The Average Price for a Liter of Petrol: $1.50
* The Max Price for a Liter of Petrol: $14.50
* The Min Oil Consumed Daily: 51 Barrels
* The Average Oil Consumed Daily: 500,000 Barrels

## Step 4: Visualisation

Once the Data was cleaned, prepped and basic stats understood it was time to visualise the data. In this study there was three python libraries used as explained prior in this report, Matplotlib, Seaborn and Plotly. Through visualisations the trends within the data could be easily displayed using bar charts, histograms, and scatter plots. The trends and results of the EDA will be discussed in the results section of this report.

# Results

### Price Per Litre – Data Dispersion

The figure below shows the count of the Price Per Liter of the Petrol Prices Worldwide dataset. The graph shows us where most countries have a liter petrol price that is let then $3. There are two countries where a liter of petrol is greater the $4.

Chart, histogram

Description automatically generated

## Correlation Graph

### Heat Map of all Field Correlations

The Heatmap below gives a good depiction of the various relationships between fields in the Petrol Prices Worldwide data. Corelations of 0.95 and above can be suspicious and with more data collection these relationships could be investigated further. Corelations of between 0.5 – 0.9 show a strong relationship between fields and this can be seen in the heatmap below.

A picture containing text, monitor, screen, several

Description automatically generated

Based off the above heat map there is a strong corelation of 0.622 between Yearly Gallons Per Capita and GDP Per Capita (USD). In the below figure it can be seen that many of the data points are along a similar area of the Scatter plot showing a strong relationship between the two fields.

Chart, scatter chart

Description automatically generated

## Petrol/Oil Prices

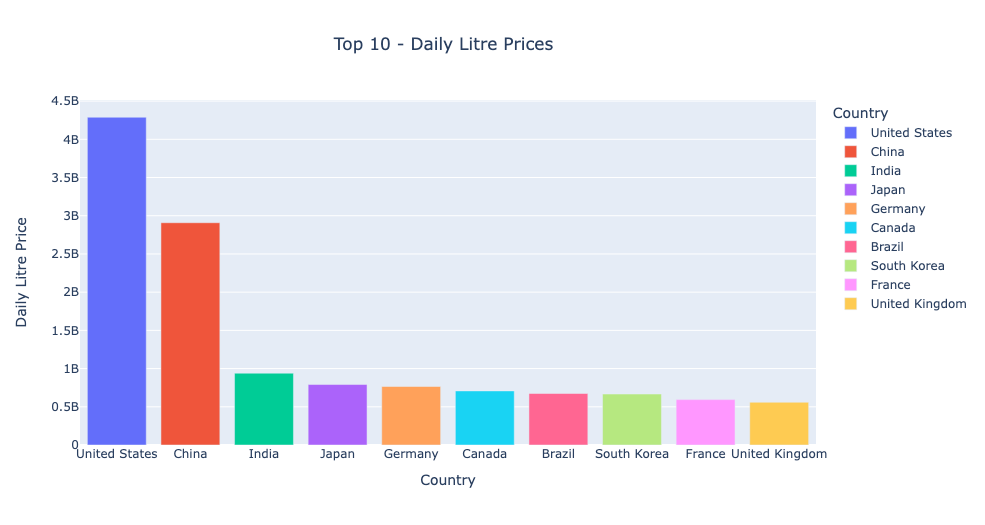
### Most expensive countries to buy Petrol/Oil

The graph below shows us the top 10

Chart

Description automatically generated with medium confidence

### Daily Prices Petrol/Oil



## Petrol/Oil Consumption

### Top Oil Consumers

A picture containing bar chart

Description automatically generated

### Lowest Oil Consumers

Chart, bar chart

Description automatically generated

## GDP Per Capita

### Yearly Gallons Consumption Per Capita

Chart, bubble chart

Description automatically generated

### Countries with the Highest Gross Domestic Product (GDP) Per Capita

Chart, bubble chart

Description automatically generated

### Linear Regression Model

Chart

Description automatically generated with medium confidence

# Insights

(Point out at least 5 insights in bullet points)

* In USD majority of petrol prices are below $3 per liter.
* Most Expensive place to buy petrol North Korea, Tonga, Nieu (lowest oil consumers)
* Top oil consumers US & China (Top manufacturing countries) compared to lowest Nieu & Saint Helena.
* Leading the way with the highest individual consumption of Oil is Singapore (3679.5 Gallons) and closest countries behind are less then 2000 gallons per year.
* The Highest Gross Domestic Product is Luxemburg followed closely by 4 other countries Switzerland, Macao, Ireland, and Cayman Islands.
* Linear Regression Model – Needs more data collection to be able to improve the predictability accuracy.

# Bibliography

Developers, N., 2008-2022. *NumPy: the absolute basics for beginners.* [Online]   
Available at: https://numpy.org/doc/stable/user/absolute\_beginners.html  
[Accessed July 2022].

Development, P., 2008-2022. *Pandas Package overview.* [Online]   
Available at: https://pandas.pydata.org/docs/getting\_started/overview.html#package-overview  
[Accessed July 2022].

O'Sullivan, J., 2022. *Cost of diesel and petrol up more than 40% on last year.* [Online]   
Available at: https://www.rte.ie/news/2022/0614/1304877-fuel-prices/  
[Accessed 17 July 2022].

Termeer, C., 2013. *Fundamentals of Investing in Oil and Gas.* illustrated ed. s.l.:Chris Termeer.

Uslu, C., 2022. *What is Kaggle?.* [Online]   
Available at: https://www.datacamp.com/blog/what-is-kaggle  
[Accessed July 2022].

Usmani, Z., 2022. *Petrol/Gas Prices Worldwide.* [Online]   
Available at: https://www.kaggle.com/datasets/zusmani/petrolgas-prices-worldwide  
[Accessed 11 July 2022].