# Group ID - MSc in Data Analytics

Author: Afshan Mehmood

e-mail: 2025102@student.cct.ie

Student ID: 2025102

## Abstract

*My dataset comprises the Fuel Consumption (ktoe) data, sourced from the Energy Balance Statistics available on the CSO (Central Statistics Office) website. The data spans from the year 1990 to 2022. Using this dataset, I have analyzed the fuel consumption patterns across various sectors and fuel types in Ireland to identify which sectors and fuel types contribute the most to overall fuel consumption. Additionally, I have examined yearly consumption trends over nearly twenty three years and studied how different types of fuels (e.g. petroleum, natural gas) contribute to the total consumption. The analysis includes Exploratory Data Analysis (EDA), data preparation and visualization, as well as statistical evaluation. I have also applied machine learning models to support my findings. Apart from this, I have done sentimental analysis on the comments gathered from Reddit on the topic of Renewable Energy as well. By focusing on dominant fuel types and sectoral consumption trends, along with integrating public sentiment toward renewable energy, these findings can assist policy makers for designing data-driven strategies that support Ireland’s compliance with the 2009 Renewable Energy Directive and long-term sustainability goals.*

## Introduction

My report will majorly comprise four sections. These sections will correspond to the coding sections in my ipnyb code file.

1. Data Preparation and Visualization

2. Statistics for Data Analytics

3. Machine learning for Data Analytics

4. Programming for Data Analytics

First I will start with the Question Formulation part i.e. through my dataset analysis, what type of questions can arise:

## Question Formulation:

The dataset examines the monthly expenditure of overnight foreign visitors in Ireland, spanning from January 2003 to February 2025.

Based on the dataset, following questions can be formulated:

**Question 1:** How has the total consumption of fuel in Ireland changed over time?  
**Relevant Techniques:** Trend analysis (line plots).

**Question 2:** What is the distribution of fuel consumption by sector (e.g., Residential, Canada, Industrial, Transport)?  
**Relevant Techniques:** Grouping by sector, box-plots.

**Question 3:** Which fuel categories (e.g. petroleum, renewables) dominate the fuel consumption?  
**Relevant Techniques:** Grouping by fuel type, bar plots.

**Question 4:** Are there any yearly variations in the consumption of fuel overall in Ireland, or for a particular fuel type?  
**Relevant Techniques:** Year-over-year comparison. seasonality analysis.

**Question 5:** How does the consumption of fuel relate to the amount of consumption from each sector?

**Relevant Techniques:** Correlation analysis, scatter plots, regression.

**Question 6:** Based on historical data, can we predict the future expenditure of foreign visitors?  
**Relevant Techniques:** Regression models (e.g., linear regression, Random Forest regression).

### **Question 7:** Are there any extreme consumption values (outliers) for a fuel type/sector that significantly impact the overall consumption data? **Relevant Techniques:** Boxplots, IQR-based outlier detection.

**Question 8:** Can we predict the value of fuel consumption based on the sector and fuel type?  
**Relevant Techniques:** Regression models (e.g., Random Forest, Decision Trees)

**Question 9:** Can we predict the sector based on the fuel consumption value and fuel type?  
**Relevant Techniques:** Classification models (e.g., Random Forest, Decision Trees), KNN, etc.

**Question 10:** Can we compare Ireland’s energy demand/generation patterns with other countries?   
**Relevant Techniques:** Hypothesis testing (Wilcoxon test, Chi-square test, etc.).

## Section1: Data Preparation and Visualisation

Q1. Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data (This will require research outside of class material).

Q2. Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.

Q3. Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.

Q4. Modern energy distribution has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to utility companies and their customers, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development making reference to Tufts Principles. Note you may not use Powerbi, RapidMiner, tableau or other such tools to accomplish this (at this stage).

## 

**Question1: Data Acquisition Process**

For my project, I acquired the raw dataset from the Central Statistics Office (CSO) of Ireland (Central Statistics Office, 2025), specifically from the official webpage:<https://data.cso.ie/table/SEI06>. The dataset, titled *“SEI06: Fuel Consumption (ktoe)”*, provides comprehensive energy consumption statistics of Ireland from the year 1990 to 2022. The data gives fuel consumption across various sectors (e.g., transport, residential, agricultural) and fuel types (e.g., petroleum, natural gas, electricity, renewables) making it highly suitable for trend analysis, statistical modeling, and machine learning applications.

Rationale: The CSO is the authoritative provider of Irish energy statistics, ensuring accuracy and official recognition.

Access method: Direct download of csv file via web interface.

I got another dataset for my inferential statistics part from Ember (<https://ember-energy.org/data/yearly-electricity-data/>). Ember is an energy think tank that aims to accelerate the clean energy transition with data and policy. Ember is the trading name of Sandbag Climate Campaign CIC, a Community Interest Company registered in England & Wales #06714443. 'Ember' and 'Sandbag' are trademarks held at the United Kingdom and European Union Intellectual Property Offices. All content is released under a Creative Commons Attribution Licence (CC-BY-4.0).

My dataset titled ‘Yearly electricity data-long format’ provides contains yearly electricity generation, capacity, emissions, import and demand data for over 200 geographies, making it highly suitable to conduct research for finding similarities between some country(s) against Ireland and applying parametric and non-parametric inferential statistical techniques to compare them. Ember has collected this data from multi-country datasets (EIA, Eurostat, BP, UN) as well as national sources (e.g China data from the National Bureau of Statistics).

Rationale: Since, the Ember content is released under a Creative Commons Attribution Licence (CC-BY-4.0). This means it is free to share and adapt for our work – as long as we give credit.

Access method: Direct download of csv file via web interface.

### *Positive Aspects of the Acquisition Process:*

* Open Access and Transparency:   
  The CSO provides open access to this dataset under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This allows for reuse, distribution, and adaptation of the data for academic and research purposes, provided appropriate credit is given.

Similarly, Ember also provides open access to the dataset under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.This means it is free to share and adapt for our work – as long as we give credit.

This greatly enhances the usability of the data in an educational context without the need to seek additional permissions.

* Data Reliability and Credibility:  
  As an official governmental source, the data from CSO is highly reliable, consistent, and regularly updated. The Central Statistics Office (CSO) is Ireland's national statistical institute with the role to provide independent statistics about society, economy, and environment, which are freely available to everyone and support evidence-informed decision making. The CSO was established to provide independent, accurate, and verifiable facts about Irish society and economy.

Thus, CSO is Ireland’s official national statistics agency, and the data is collected from verified national surveys and energy accounts.

Similarly, Ember, being a global think tank, compiles its electricity data from official sources such as EIA, Eurostat, UN, BP, and national statistical offices (e.g., China’s National Bureau of Statistics).

This adds to the credibility of my analysis and ensures that any conclusions drawn are based on trusted information.

* Well-Structured Format and Content:  
  The data from both sources was in downloadable csv formats, making it easy to import into data analysis environments such as Python for further cleaning, transformation and analysis. Moreover, the content from both sources was valuable and elaborate enough to do my research.

Dataset from CSO provides detailed information about the contribution of sector and fuel types to the overall fuel consumption of Ireland from 1990 to 2022. Similarly, the dataset from Ember provides extensive content about energy demand, energy generation, etc. for countries around the world along with Irelnad, from 1990 to 2024. This helped me to do comparative analysis of energy demand and generation of various countries with Ireland through inferential statistics.

* Detailed Metadata:  
  Both datasets had accompanying metadata, including clear definitions of each variable and unit of measurement which helped in understanding the context and interpreting the values accurately. The categorical and numerical values could be easily distinguished which made it easy to understand the data.

### *Negative Aspects of the Acquisition Process:*

* Initial Complexity in Interpretation:

Although both datasets had clear and comprehensive data and terminologies, it was a challenge to explore, clean and do a detailed analysis on them due to their exhaustive nature.

In the CSO data, the structure and terminologies were initially complex to interpret, since it had hierarchical columns. For example, for every year, there were different rows with different fuel consumption values for a particular combination of sector and fuel type. This required additional research and critical analysis to fully understand and use the data correctly. The data had a lot of zero consumption values which made it difficult for analysis in the machine learning part since it was a challenge to choose the best features to train the machine learning models.

Similarly, in Ember data, the structure and terminology were initially complex to understand, since it also had hierarchical columns. Also, it had extensive information about various regions and countries of the world, with data divided into different categories, subcategories, values and units. This made it difficult to choose the right categories with relevant units to do my hypothesis testing for inferential statistics.

* Preprocessing Required:  
  Despite being well-structured, both the datasets required significant cleaning and preprocessing. For example, in the CSO data set, the aggregate categories showing the summation across the subcategories were also included in the sectors and fuel types. So, I had to filter out relevant columns/data, drop irrelevant columns, handle missing values, and apply scaling or transformation where necessary before the machine learning analysis.. Similarly, in the Ember data set, I had to filter out appropriate categories and energy units as well as restructure the dataset before hypothesis testing. While not a major limitation, it did require additional time and technical skills.

### *Licensing and Permissions:*

LICENSES for supplied datasets:

1.<https://data.gov.ie/organization/sustainable-energy-authority-of-ireland>

*Published by: Sustainable Energy Authority of Ireland*

*Licensed under: Creative Commons Attribution 4.0*

[*https://creativecommons.org/licenses/by/4.0/*](https://creativecommons.org/licenses/by/4.0/)

*Category: Government*

*2.*<https://ember-climate.org/data/>

*Published by:Ember (trading as Sandbag Climate Campaign CIC, reg. № 06714443)*

*Licensed under: Creative Commons Attribution Licence (CC‑BY‑4.0)*

[*https://creativecommons.org/licenses/by/4.0/*](https://creativecommons.org/licenses/by/4.0/)

*Category:Energy think tank / Non‑profit research*

Both the datasets are provided under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which is a very permissive and research-friendly license. The key implications of this license are:

* Proper credit must be given to the CSO and Ember when using or sharing the data. In my report, I have ensured to cite the sources clearly.
* The license allows the use of the data for educational, research, and even commercial purposes without the need for additional approvals, simplifying the legal and ethical aspects of my work.
* The data is free to use and the license permits adaptation and transformation of the dataset, which is essential for tasks like feature engineering and machine learning preprocessing.

*Implications of Licensing*

* Both datasets can be legally combined/integrated and redistributed, provided each source is credited appropriately.
* Any models, visualizations, or reports that we publish should include attribution statements and links to the original license texts.
* We may embed data in commercial software or dashboards, but we cannot use trademarks/logos without permission.

*Relevance:* The data used in this project was obtained from publicly accessible and reputable sources..The licensing terms, giving free access on the condition of giving due credit, were particularly beneficial in ensuring compliance with ethical and legal standards while maintaining the flexibility needed for my research and data analysis.

Licensing and metadata from both data providers were reviewed and cited to ensure ethical reuse and licensing compliance. This ensured the data is both ethically sourced and compliant with educational standards.

Overall, the availability of this type of open data from credible and licensed sources promotes transparency, reproducibility, and wider accessibility, making it highly relevant for academic research and analysis.

### *Conclusion:*

In summary, the process of acquiring data from the CSO website was largely positive, offering high-quality, transparent, and freely accessible data under a permissive license. While some effort was required to clean and preprocess the data, the overall experience was smooth and supportive of the objectives of my research project. The licensing terms made it easy to access the data for academic purposes.

**Question2: Data Preparation and Visualisation**

I will discuss the Exploratory Data Analysis and Data Preparation part. The visualizations according to Tufts principle will follow on throughout the report.

Exploratory Data Analysis:

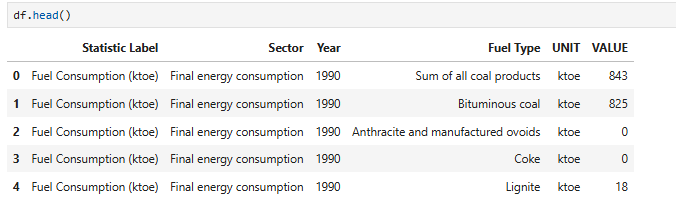
*Understanding my data structure:*

1.Initializing libraries

2.Reading the csv file: df=pd.read\_csv("Energy2.csv")

3.For displaying top 5 rows of my data set:

df.head()



I get the following insights:

● Categorical columns: Sector, Statistic Label, Fuel Type, UNIT

● Numeric column: VALUE

● Time column: Year

Each row gives a Value for the specific Statistic Label along with the Fuel Type and Sector, for a particular Year

The dataset reveals how Ireland's final energy consumption varies annually across different combinations of sectors and fuel types, highlighting trends in fuel usage and shifts in energy dependence over time.

Right now the data is in long format.

*Initial Checks:*

1. Checking the data types:

df.dtypes

Statistic Label object

Year int

Sector object

Fuel Type object

UNIT int

VALUE float64

This shows that except the VALUE and UNIT column, all other columns have non-numeric entries since the data type is ‘object’.

2. To get a summary of a DataFrame, including Class type, Index range, Column names, Non-null count and Data types:

df.info()

This shows there are 65340 entries in total i.e. 65340 rows and 6 columns.

3. To find dimensions of the dataset:

df.shape

I get (65340,6) confirming no of rows and columns.

4.To provide summary statistics for the numeric columns in a DataFrame by default.

df.describe()

It computes a set of descriptive statistics that are useful for understanding the distribution and spread of numerical data. However, in my case, the VALUE column has different values for different Sectors and Fuel Types, accordingly.

For VALUE column:

● Mean: 45.377

● Median (50%): 0

● Max: 13189.0

● Min: 0

● Std Dev: 372.367

Insights: Over 75 percent of the data is zero, indicating extreme sparsity. Minimum is zero indicating that some fuels/sectors have no consumption in certain years. However some entries are very large, inflating the mean and standard deviation. These high values may be due to national totals or large sectors like transport/electricity. The distribution is highly skewed indicating the need for a log-transformation for modeling in order to compress these large differences.

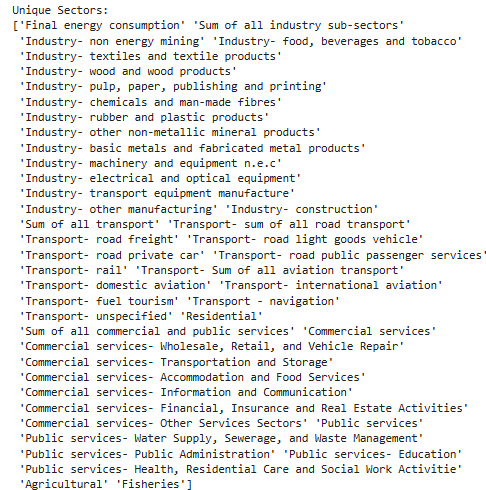
Rationale for log transformation: Since the data is highly skewed with substantial outliers, therefore, log transformation news to be applied to the fuel consumption values. This is because log transformation, when applied to numerical data, helps to stabilize variance, normalize skewed distributions, improve model performance and reduce the influence of large outliers.

5. There are unique values in the ‘Sector’ and ‘Fuel Type’ columns of the DataFrame. In order to see all the distinct values present in those specific columns, we use:

unique\_sectors = df['Sector'].unique()

print("Unique Sectors:")

print(unique\_sectors)



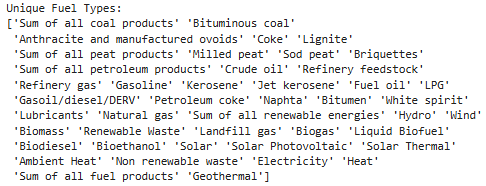
And

# Unique values in 'Fuel Type'

unique\_fuel\_types = df['Fuel Type'].unique()

print("\nUnique Fuel Types:")

print(unique\_fuel\_types)



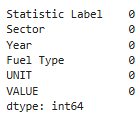
The Statistic Label, on the other hand, has just one unique value i.e. Fuel Consumption.

Rationale for formatting: My dataset has hierarchical columns. This suggests that my data is in long format and I need to arrange it in a proper format to do analysis in terms of Fuel Type, Sector, etc.

*6.* Check for missing values:

To check for missing values (null or NaN values) in my DataFrame:

df.isnull().sum()



This shows there are no missing values in my DataFrame.

Exploring data:

Further, in my report, I will discuss following methods of EDA:

*1.Univariate Non-Graphical EDA (Numerical Data):*

Descriptive Statistics:

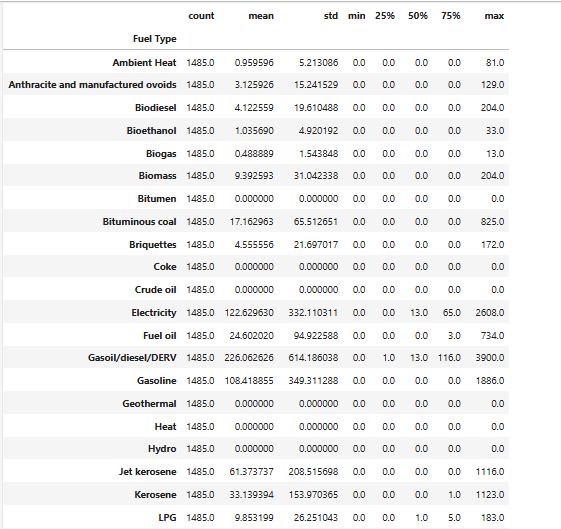
The reason for applying descriptive statistics (count, mean, std, min, quartiles, max), is that it will give me understanding about my data in terms of central tendencies and dispersion. It will also help me to identify possible outliers or unusual ranges and help me understand each metric's distribution before modeling.

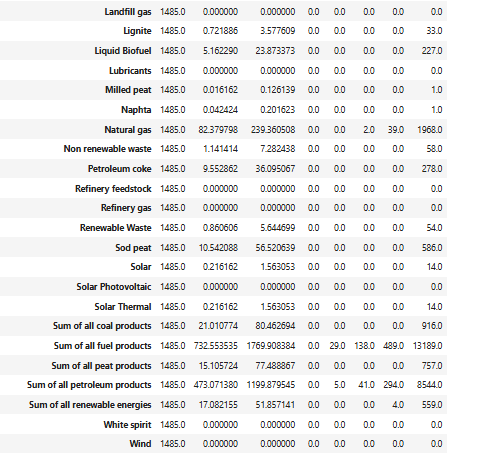
Checking the general distribution of each Fuel Type:

In order to check the general distribution of each Fuel Type:

df.groupby("Fuel Type")["VALUE"].describe()

This will group the dataset df by the Fuel Type column and then apply the describe() function to the VALUE column within each group. The describe() function will return key statistics for each group.





Key Insights:

* The above statistics show that the fuel type consumption data is highly skewed. Many fuel types have zero average consumption while Sum of all fuel products, Sum of all petroleum products, diesel and electricity have comparatively higher average fuel consumption. It is obvious that the large values are representing the aggregate or summed values of their subcategories e.g. sum of all peat products is the aggregate of sod peat, milled peat and briquettes.

Rationale for log transformation and for dropping sectors/categories: Due to skewness, we need to apply log transformation. This also highlights an important observation that in the column, we have the original subcategories as well their aggregates. This will result in duplication of values and thus, these categories or rows need to be dropped for further analysis, later.

* Electricity, Diesel, Gasoline, Jet Kerosene, Natural gas, Sum of all fuel products, Sum of petroleum products have large standard deviation suggesting the high consumption of those fuels across different sectors. These appear as outliers in the fuel consumption data and lead to a skewed distribution.
* Many fuel types have a median (50%) value of 0.0, indicating that they were used rarely or inconsistently across sectors or years. The fuels like Biodiesel, Gasoline, Refinery Waste, Non renewable waste, etc. show zero consumption across all sectors and years.

This may be due to the assumption that these fuels were not used in the particular sector in that particular year.

* Among renewable sources, Sum of all renewable energies has some modest mean consumption, however, wind, hydro, etc.consumption is still zero as compared to other non-renewable sources and fossil fuels.
* The aggregate fuels give greater insights since the sum of all fuel products has the highest mean and standard deviation, which can be set as a benchmark to compare other fuel types’ consumption. This is followed by the sum of petroleum products having the mean and standard deviation of approximately 473 and 1199.8, respectively. Overall, these aggregate groups help in tracking category-wise fuel trends and this information can be utilized for trend forecasting.However, if we are talking about comparison, then the aggregates of all categories will be used for comparison rather than their subcategories.

Checking the general distribution of each Sector::

In order to check the general distribution of each Sector:

df.groupby("Sector")["VALUE"].describe()

This will group the dataset df by the Sector column and then apply the describe() function to the VALUE column within each group. The describe() function will return key statistics for each sector.

Insights: The distribution of sector energy values is also highly right-skewed with many sectors having a median of zero, indicating a large number of zero or near-zero consumption values. A few sectors such as Residential, Final Energy Consumption, and Transport-related sub-sectors show extremely high maximum values and large standard deviations, highlighting the presence of significant outliers and variability. This also highlights an important observation that in the column, we have the original subcategories as well their aggregates. e.g. we have the sum of all transport along with its subcategories. Similarly, we have final energy consumption, showing the overall aggregate.

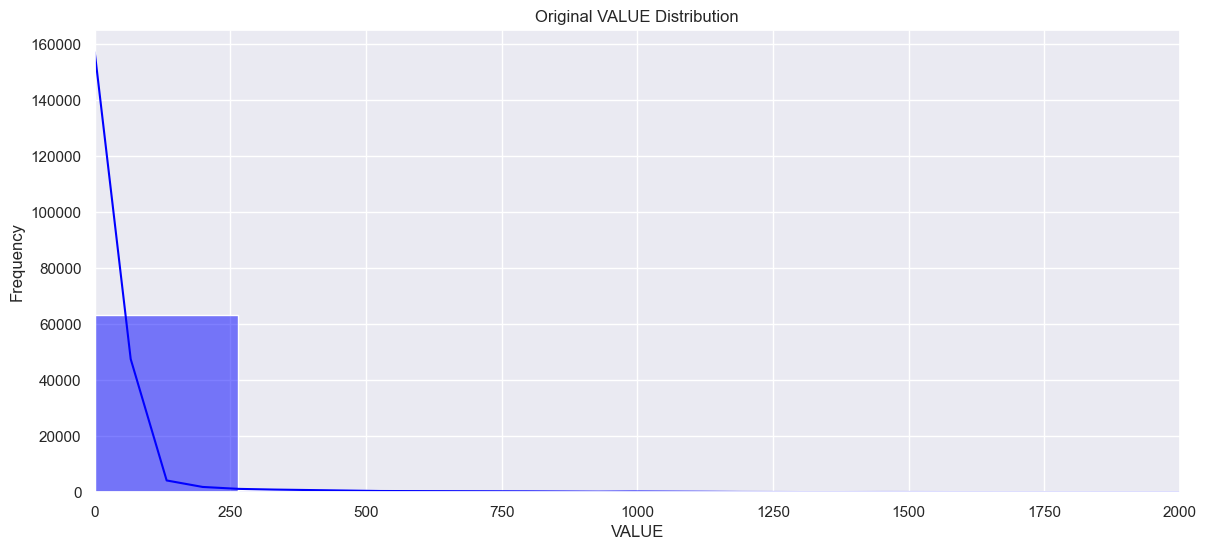
Rationale for log transformation and dropping categories: The above observation of skewed data, strongly justifies the use of log transformation to normalize the data and stabilize variance. Moreover, it also highlights that we have the original subcategories as well their aggregates which will result in duplication of values and thus, these categories or rows need to be dropped for further analysis, later.

*2. Univariate Graphical EDA (Numerical Data):*

In order to visualize the distribution of individual numerical variables to identify patterns, outliers, or skewness.

Histogram:

Plotting a histogram to visualize the distribution of the Fuel Consumption values. This helps assess whether the data is normally distributed or if there are any skewed patterns.



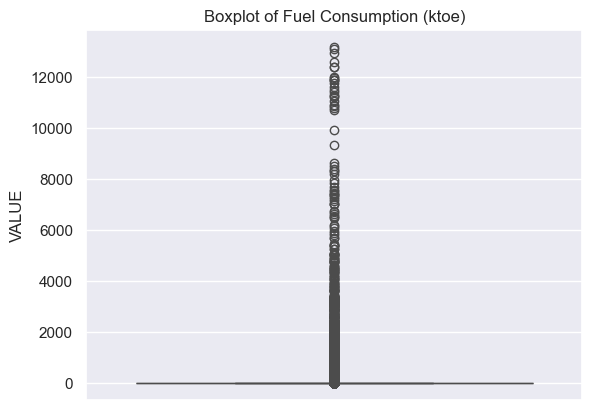
Insights: The histogram allows us to see the frequency of different fuel consumption values and whether they follow a specific distribution. In the above plot, it shows highly positively skewed distribution. In a right-skewed distribution, most of the data is clustered towards the lower end, but there is a long tail extending towards the higher values. This means, in our data, there are a small number of observations with very high fuel consumption values. However, the bulk of the values are zero showing zero consumption.

Impact on analysis: The mean (average) consumption will be higher than the median due to the influence of the higher values in the right tail. Further, the long right tail indicates the presence of outliers or a few very high-consumption values, which significantly increase the mean consumption. These indicate outliers possibly due to high fuel types like electricity, sum of all fuel products, etc. across various sectors like residential and transport, etc.. Understanding the high consumption behavior/trends might provide insights into targeted energy initiatives or policy recommendations.

Rationale: Right-skewed data can often be normalized or made more symmetric through a logarithmic transformation. When applying machine learning models or statistical techniques, skewness can influence predictions.

Boxplot:

The boxplot can be used to visualize the spread and identify any outliers in the expenditure data.



In the above boxplot for Fuel Consumption values, there are significant outliers, showing those values that are significantly different from the majority of the data. In this case, the 25th and 75th percentile are both zero resulting in interquartile range to be equal to zero as well. Moreover, a single line is seen in the graph, again because 25th, 50th, and 75th percentiles along with minimum value are all 0. The entire box has collapsed to a thin line at 0. Any non-zero value lies beyond the whiskers because the IQR is 0 and is considered an extreme outlier. Due to the max value being 13,189, the vertical scale is highly skewed. These high outliers may represent high consumption of a particular fuel type across a particular sector.

Impact on Analysis: Outliers can increase the consumption value significantly, making it higher than what most of the actual consumption values that are mostly zero. The median is still zero thus it may not provide a more reliable measure of central tendency in this case. Also, the presence of outliers leads to a right-skewed distribution (as already observed), where the majority of the data lies on the lower side, but a small number of very high values distort the overall shape of the distribution.

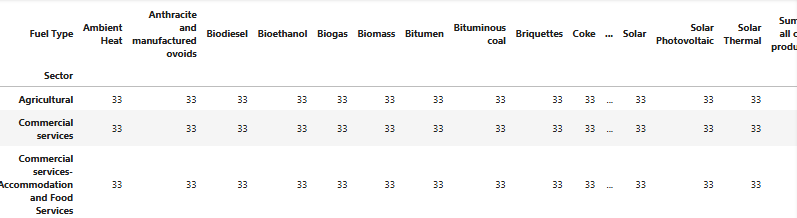
Rationale: Applying a log transformation to the data can reduce the effect of extreme values and bring the distribution closer to normal.

*3. Bivariate Non-Graphical EDA (Categorical & Quantitative Variables):*

In order to analyze the relationship between one categorical explanatory variable (e.g., Sector) and one quantitative outcome variable (e.g., Fuel Consumption value), we can use different methods like cross-tabulation or correlation.

Cross-tabulation:

It is a method for examining the relationship between two categorical variables. In my case, Sector and FuelType are categorical, while Fuel Consumption value is numerical.

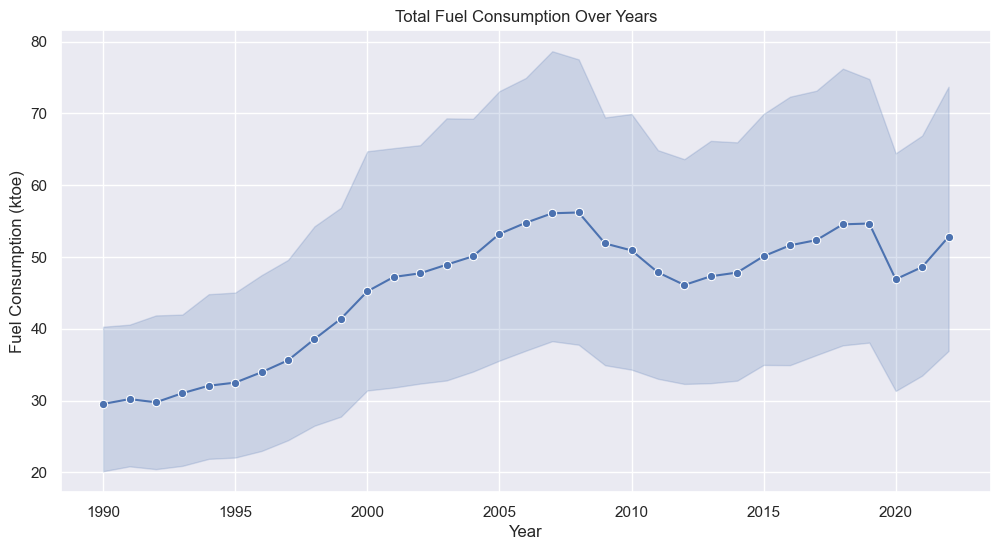


Rationale: As it can be noted, this helps to observe the frequency of each category combination and analyze the distribution of Fuel Type across different Sectors

*4. Bivariate Graphical EDA:*

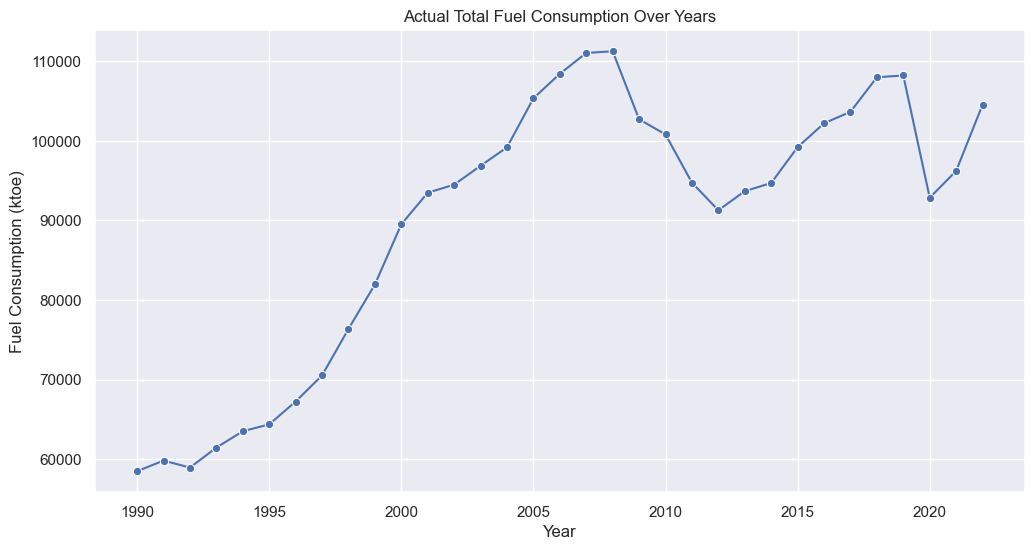
Line plot:

To observe the relationship of total fuel consumption with the number of years.



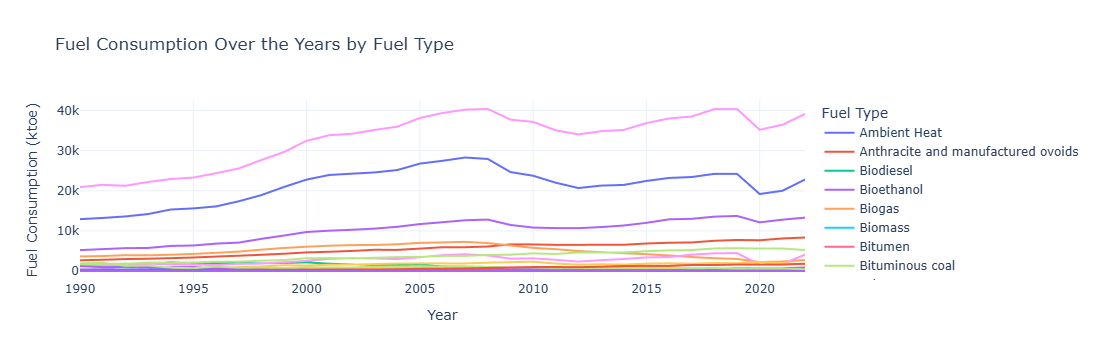
The above line plot of fuel consumption over years, without aggregation, simply connects individual data points corresponding to different fuel types and sectors for each year. This approach results in a cluttered visualization that misrepresents the overall yearly trend due to multiple values per year.

In order to accurately capture the total fuel consumption pattern over time, it is essential to aggregate the data i.e. by summing fuel consumption values for each year. This will provide a clear and meaningful trend of fuel consumption from 1990 to 2022.



Insights: This line plot shows the overall trend of fuel consumption from 1990 to 2022. The plot shows peaks in the years 2008, indicating highest fuel consumption of more than 110000 ktoe in that year, in Ireland. The fuel consumption decreased till 2012, after which it began increasing again reaching close to 108000 ktoe in 2019. Overall, this graph gives more reliable insights for analysis and policy making for energy consumption trends in Ireland.

Yearly fuel consumption by fuel type using a line plot:



Insights: The above plot shows the trend for yearly fuel consumption from 1990 to 2022 by fuel type. The plot shows highest fuel consumption across all fuel types in the years 2007-2008, which decreased after this period but went to the same peak of about 40k ktoe between 2018-2019.

Moreover, we can see that the highest consumption is due to the ‘sum of all fuel products’ followed by the ’sum of all petroleum products’ and ‘Gasoil’, over the years. However, since, ‘sum of all fuel products’ is just an aggregate, we can drop it for our analysis later. This makes Gasoil to be the highest consumed fuel.

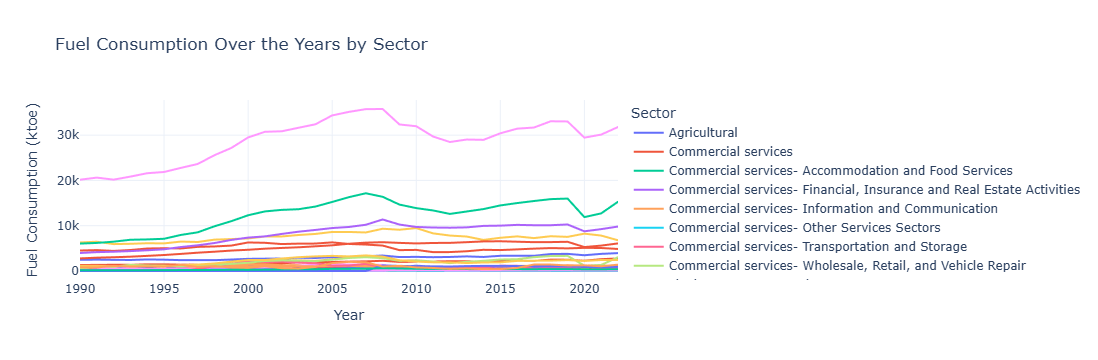
It can also be analysed that uptil 2009, there was not much fuel consumption from non-renewable energy sources, however, after 2009, an increase in the trend was observed. This may be due to the 2009 Renewable Energy Directive Ireland, under which Ireland had set a target of producing 16% of all its energy needs from renewable energy sources by 2020.

These statistics can be used by the policy makers to enhance consumption of renewable sources by investing in renewable rather than non-renewable fuel types.

Rationale for choosing selective categories: From the above figure, we can see that a lot of lines for different fuel types are close together and we cannot analyse their contribution clearly. Therefore, for further analysis, we can work on selective categories mainly representing aggregate values like ‘sum of all petroleum products’,etc.

Yearly fuel consumption by sector using a line plot:

We can also show the trend for yearly fuel consumption from 1990 to 2022 by sector.



Insights: The plot shows that the highest fuel consumption was in the year 2008 followed by a decline. But it started increasing again after 2012 rising to about 33k (ktoe) in 2019. The decrease after 2008 may be due to the 2008 financial crisis impacting the overall economy in Ireland.

Further, the plot specifically shows that the highest fuel consumption is by ‘final energy consumption’. Since it is just an aggregate, we can drop it for analysis, later. Apart from that, the highest fuel consumption trend is by ‘sum of all transport’ followed by road transport and residential sectors. Moreover, we can also see that the fuel consumption has been increasing in the residential sector over the years, with the highest consumption observed in 2010.

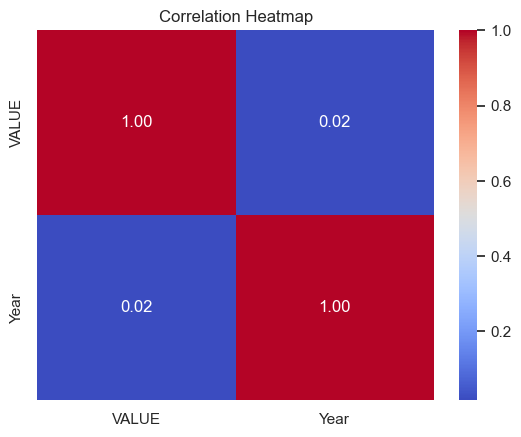
Moreover, the plot also shows that the highest fuel consumption across all sectors was in the year 2008 which was around 35k ktoe. However, this value is lower than that consumed by fuel type category in the same year, which was about 40k ktoe. So, we can deduce that in the year 2008, there was the highest fuel consumption across all the years, and it was predominantly by fuel type rather than sector.

*5.Multivariate Non-Graphical EDA:*

Used in order to analyze the relationships between multiple variables, including both categorical and numerical.

Correlation Matrix (for Numerical Variables):

We can create a correlation matrix to examine the linear relationships between all numerical variables in our dataset (e.g., Year and VALUE.).



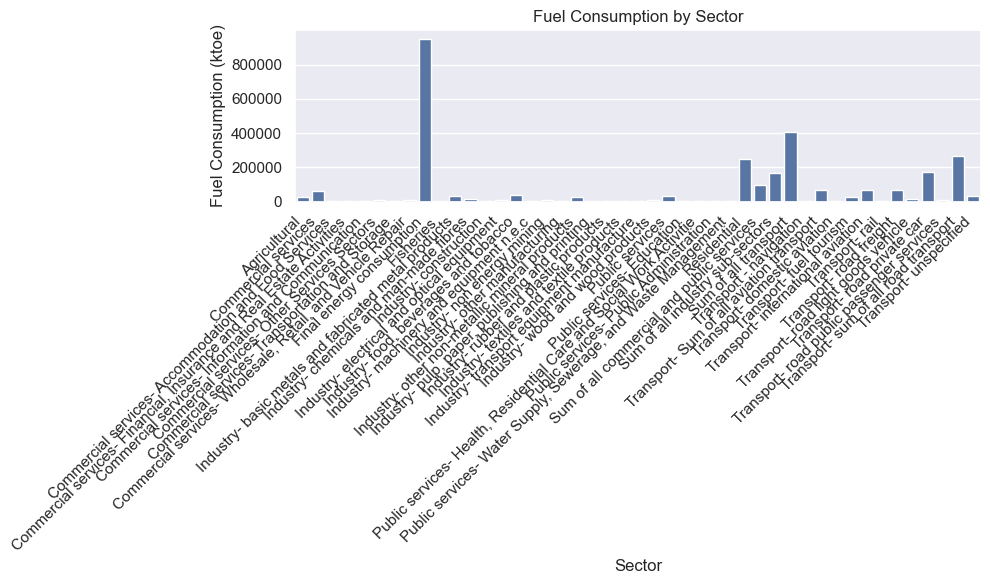
Correlation between Year and VALUE: 0.018

A correlation matrix helps to identify strong or weak relationships between numerical features, which can guide feature selection for modeling.The correlation value being almost 0.02 shows very weak (almost no) linear correlation between year and fuel consumption values. The value is close to zero, which suggests that there is no strong linear trend in overall fuel consumption across all years. This indicates that consumption may be fluctuating irregularly (ups and downs) and we need to explore other variables since it may be depending on variables like Sector or Fuel Type.

*6. Multivariate Graphical EDA:*

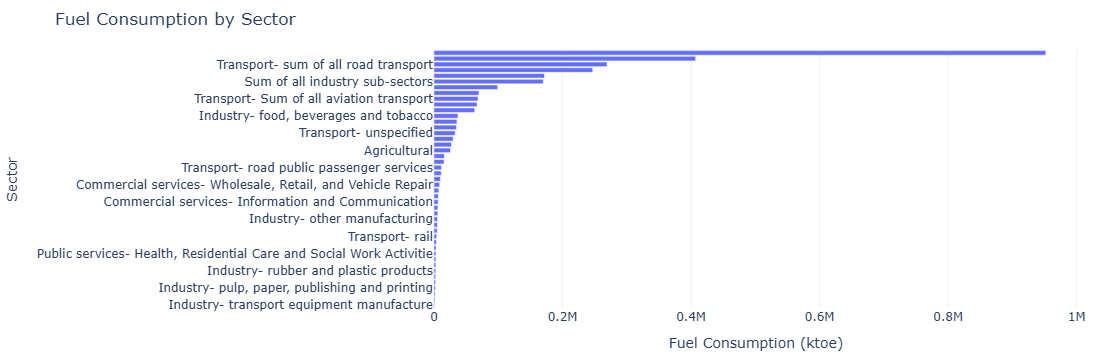
Used in order to visualize relationships between a categorical variable and quantitative variables to understand how consumption of fuel differs across categories.

We can use barplot to get meaningful insights into how fuel consumption varies by sector.This can help identify which sector has the highest or lowest consumption.



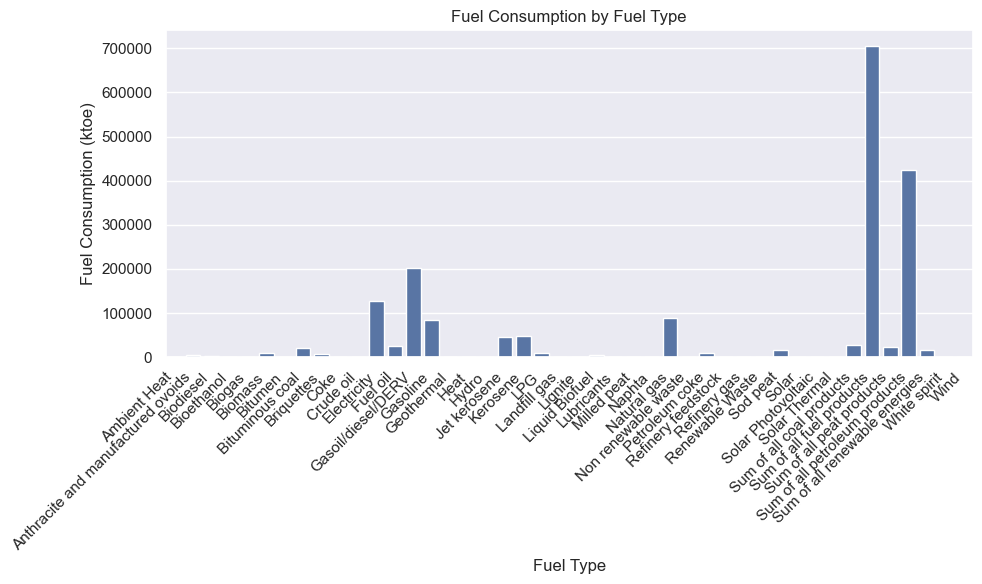
In the above plot, bars represent the total fuel consumption for each sector.   
The height of each bar tells us which sector consumes more fuel in total. For example, in the above plot, *Sum of all transport* has the tallest bar which suggests that transport sector consumes the most amount of fuel followed by Residential sector. Although the ‘Final energy consumption’ category has the highest bar, however, since it is just showing an aggregate over all the categories, therefore, we can drop it in further analysis later. The graph also shows that the fisheries sector consumes minimum fuel.

However, the graph looks cluttered as it has a lot of sectors. Therefore, in accordance with Tuft’s principles, which emphasize clarity, minimalism, and maximizing data-ink ratio without distorting the data and the need to have clear labels and more space for names, we can rotate the chart so sectors are on the y-axis. This gives more room for long names and is easier to read. We also use the Plotly bar chart with tooltips, so we don’t need to show full names on the axis.



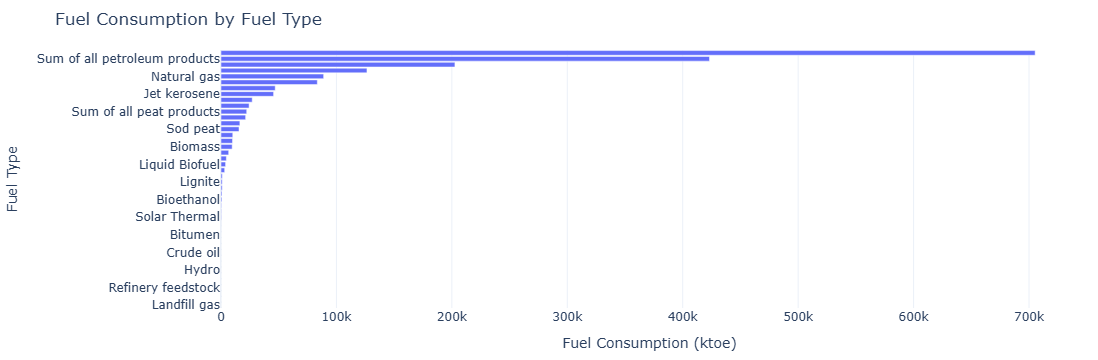
Policy Implications: If some sectors consume more consistently, government policies can be tailored to track them and provide subsidies e.g. high consumption in the residential sector. .

We can also use barplot to get meaningful insights into how fuel consumption varies by fuel types.This can help identify which fuel type has the highest or lowest consumption.

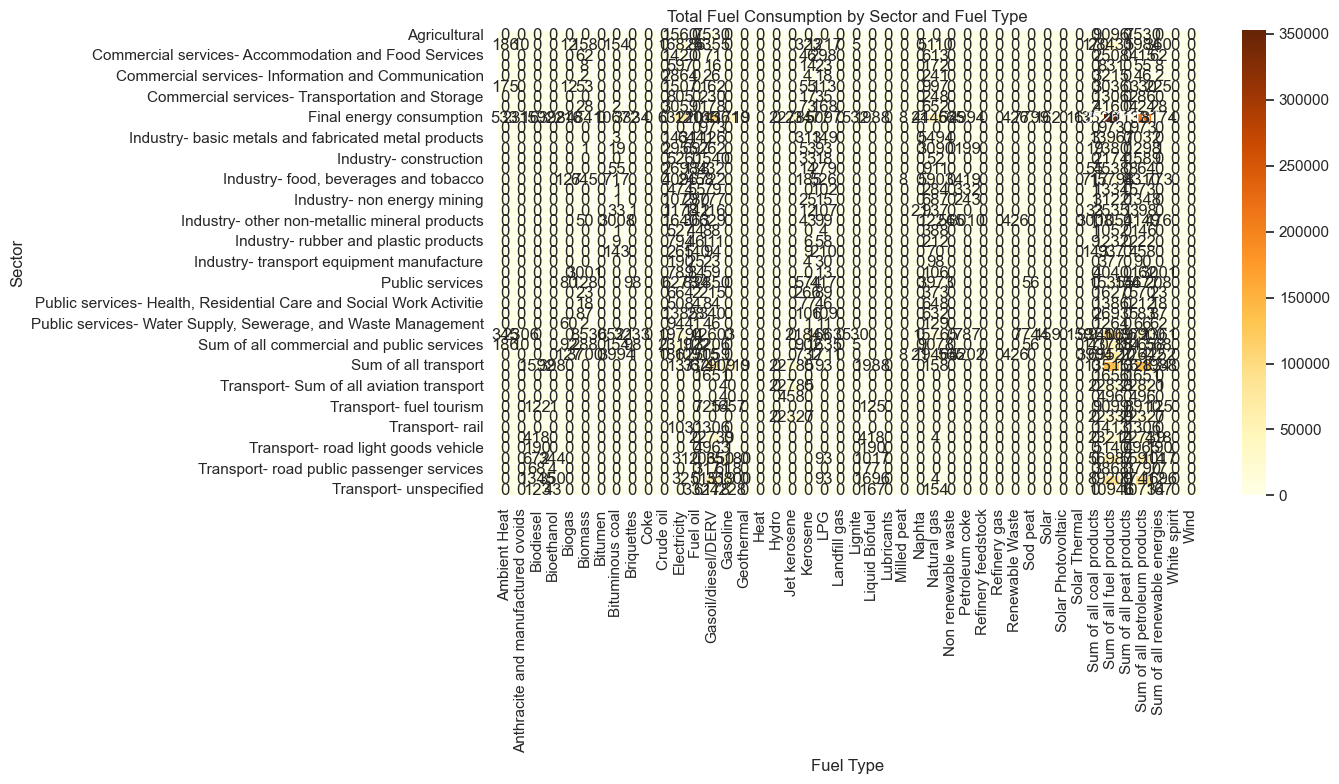


In the above plot, the bars represent the total fuel consumption for each fuel type.  
The length of each bar indicates which fuel type contributes most to overall fuel usage. For example, the ‘sum of all fuel products’ has the longest bar, followed by ‘sum of all petroleum products’, which suggests that they are the most consumed fuel type. However, since, ‘sum of all fuel products’ is just an aggregate, we can drop it for our analysis later.Apart from aggregates, highest fuel consumption is of Gasoil.

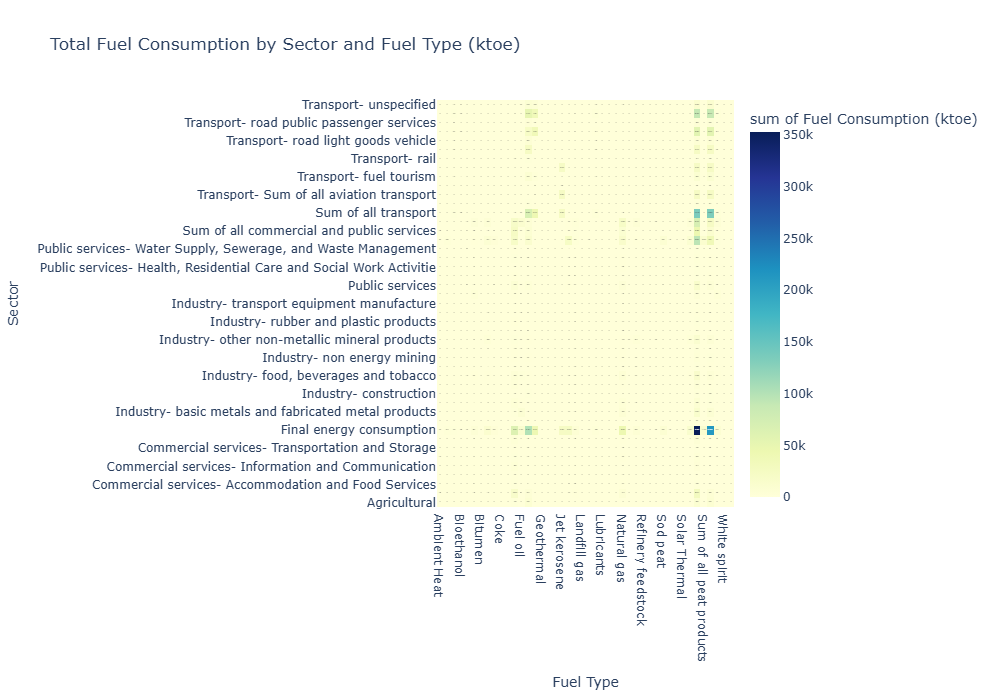
Moreover, as seen earlier as well, the chart looks cluttered due to the large number of fuel types and the length of their names. So, in line with Tufte’s principles for clarity, minimalism, and a high data-ink ratio, we can rotate the chart so that fuel types are placed along the y-axis. This layout allows for better readability of longer labels and avoids overlapping text. Additionally, by using an interactive Plotly bar chart with tooltips, we eliminate the need to display full names directly on the axis, further enhancing readability and interactivity. We also sort the values for better readability in order to assess the highest to lowest consuming fuel type.



Heatmap: It is also a way to see how different features are correlated. We can use the heatmap to see the correlation between sector and fuel types.



The above plot shows the heatmap of Total Fuel Consumption by Sector and Fuel Type and reveals several valuable insights regarding: Dominant fuel types per sector, fuel types with widespread use as well as underutilized or niche fuels. However, since it is very cluttered, therefore, for better and interactive visualisation, we can use plotly. The darker the color, the stronger the correlation. e.g. the darker blue shade between sum of all petroleum products and sum of all transport shows that both are highly correlated i.e. the transport sector consumes the most amount of petroleum products, after ‘final energy consumption’. However, since it just shows the aggregate, we can drop it later in our analysis. It further shows that commercial and public services consume the highest amount of electricity.



Question2: Now we will discuss the next part of data preparation.

**Data Cleaning and Preparation:**

The dataset was initially in a long format, with each row representing a combination of:

Year, Fuel Type, a Statistic Label, Sector, a corresponding VALUE and a UNIT. While this structure is ideal for flexible analysis and visualization, it is not directly suitable for machine learning models, which typically require one row per observation and one column per feature.

Therefore, the first step is to arrange the data in a proper format.

1. Drop unused or uninformative columns like "UNIT" since Statistic Label already has the unit in it. Also, filter the dataframe to only keep rows where 'VALUE' is valid (exists) and is zero or positive and remove any rows where 'VALUE' is missing or negative.

2. Pivot the table

df\_new = df.pivot\_table(

index=['Year', 'Sector', 'Fuel Type'],

columns='Statistic Label',

values='VALUE',

aggfunc='first' # if duplicates exist, take the first

).reset\_index()

This converts the data from long format to wide format. Now each unique value in Stat\_Unit becomes a new column (instead of rows). The new table has rows indexed by: Year, Sector and Fuel Type and columns as: Statistic Label called Fuel Consumption (ktoe).

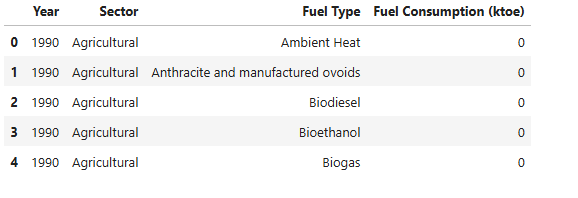
3. Flatten the column names (optional if you get a multi-index)

df\_new.columns.name = None # remove the name of the columns axis

This removes the column name label (which can appear in pivoted DataFrames) and prevents multi-level column headers. It also makes the DataFrame cleaner and easier to work with.

Overall, this transformation reshapes the dataset into a wide format, where each row corresponds to a unique combination of Year, Sector, Fuel Type and even the Statistic Label also becomes a separate column indicating values. This is essential for supervised ML models, which expect a consistent feature layout.

The data now looks like this:



Next we see the dimensions of the dataset after converting into wide format:

df\_new.shape

This gives (65340, 4) showing 65340 rows and 4 columns.

4. Check for missing values:

df\_new.isnull().sum()

The output is 0 for all columns. This indicates that the data does not have any missing values.

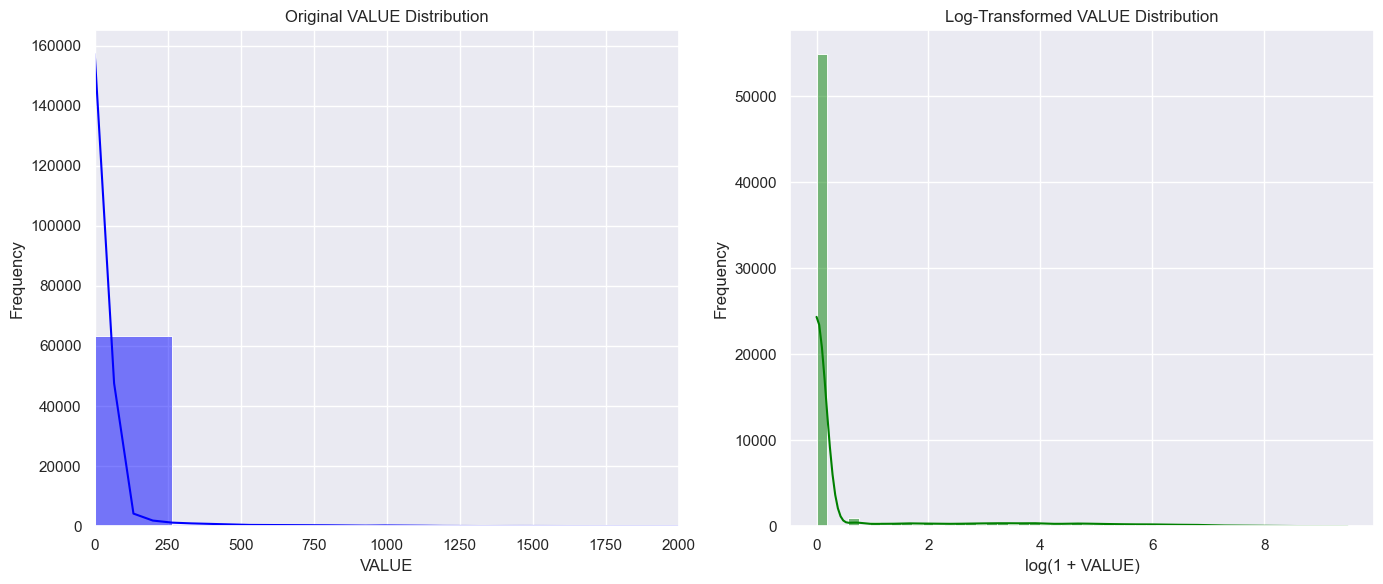
Next we reset the index to Year and drop any duplicates for cleaning our data.

5. Log Transformation:

From our EDA, we saw that the VALUE column has a lot of zeros and the distribution is heavily positively skewed therefore, we can apply log transformation.

So, in the next step, we apply np.log1p(x)which computes the natural logarithm of (x + 1), since it is numerically more stable and avoids issues with zero since log(0) is undefined.

Reason: The reason to apply log transformation is that the fuel consumption data is highly right-skewed so using log transformation compresses large values and spreads out smaller ones, making the data more normally distributed. It also helps in better visualization and stabilizes variance making it useful in statistical models and comparisons  
After log transformation, we compare our graphs by plotting them side by side.



The above plots show the distribution of fuel consumption before and after log transformation. The left figure shows that the original data is highly skewed and indicates extreme usage patterns i.e. either a lot of zeros or some extreme values. This ultimately results in outliers as indicated by boxplot plotted earlier. The figure on the right side shows log transformed data, which is again still skewed due to the large number of zeros in the original data. This shows that data has not been much normalized even after log transformation. This means that we may need to remove zeros values or scale our data further.

6. Encoding the data:

Next, we need to encode our categorical columns to be used in the Machine Learning models.

# One-hot encode categorical columns

df\_encoded = pd.get\_dummies(df\_new, columns=['Sector', 'Fuel Type'])

Here pd.get\_dummies() is a function in pandas (Pandas Development Team, 2025) that converts categorical columns into a set of binary columns (i.e., one-hot encoded columns). For each unique category in "Sector", a new column is created with a 1 or 0 to indicate whether that row belongs to that category.  
  
Reason: Here I am encoding the Sector and Fuel Type columns because most machine learning models require numerical inputs and they can not directly process categorical (text) data.

Further, I have chosen One-Hot Encoding here because these columns are nominal (no natural order), e.g., 'Sum of all transport’, ‘Residential’, etc. One-Hot Encoding creates a binary column for each category, preserving neutrality and it does not assume any ordinal relationship.

The reason for not choosing Label Encoding is that it assigns integers (e.g., 'Sum of all transport': 0, 'Residential':1), which implies order where none exists. This can mislead the algorithms into thinking one category is greater or less than another.

7. Scaling the data:

From earlier analysis, it was seen using histogram and box plot that ‘Fuel Consumption (ktoe) is positively skewed and has outliers, making it necessary to standardize it to a common scale.

Scaling the values is crucial because they have different magnitudes, which can affect the performance of machine learning algorithms.

from sklearn.preprocessing import StandardScaler

# Initialize the StandardScaler

scaler = StandardScaler()

# Apply the scaler to the selected columns

Reason: Here the StandardScaler is a tool from scikit-learn (Scikit-learn developers, 2025) used to standardize (or z-score normalize) our data. It transforms each feature (i.e., column) so that the mean becomes 0 and the standard deviation becomes 1.The result is a NumPy (DataCamp, 2025) array where all the numeric features are now on the same scale. Each column is now centered around 0 and spread equally. This is required for applying Machine Learning models that are sensitive to scale as they work better when features are on a similar scale.

For my dataset, scaling allows for better comparison and improved model performance, especially in algorithms sensitive to the scale of data, such as logistic regression and clustering.

Visualization:

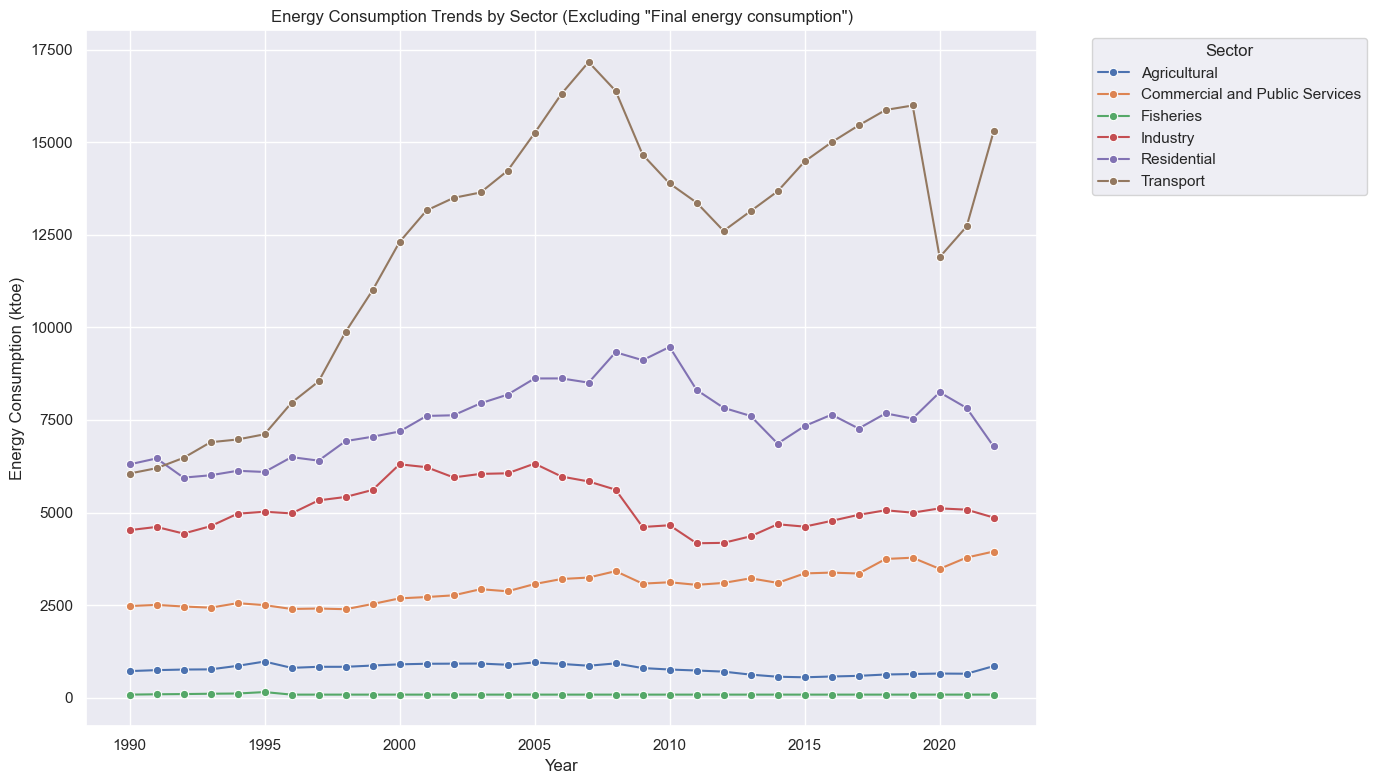
1.Extracting selective sector and fuel types:

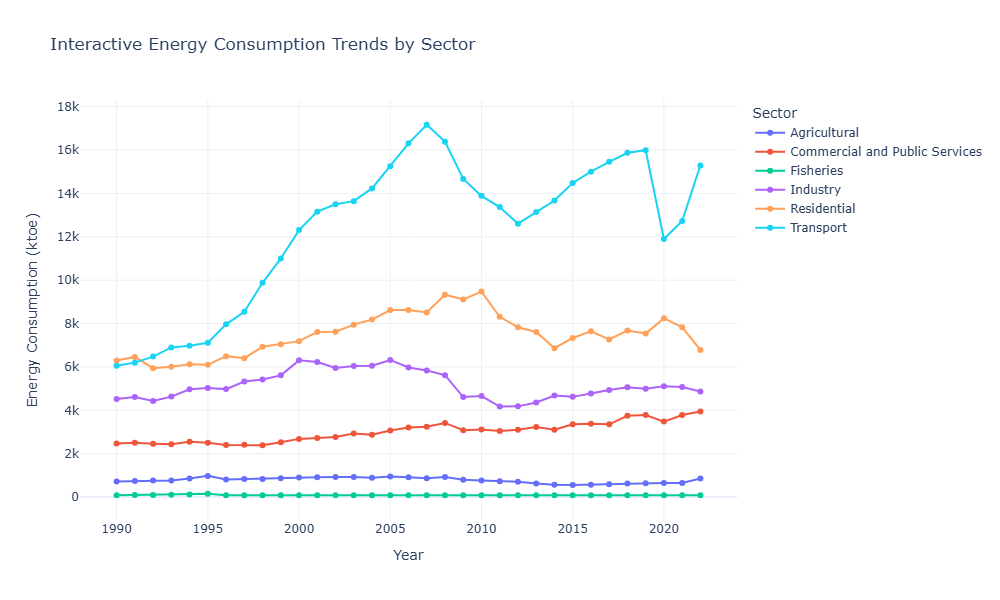
Reason: Since in my original dataset, there are 47 unique sectors and 17 unique fuel types, and after hot encoding, we have a large amount of dummy variables. Therefore, in order to reduce complexity and enhance interpretability, a domain-driven feature grouping strategy is applied. It involves manual feature selection using domain logic to group categories.

Specifically, we consolidate detailed sector categories into broader groups such as *Industry*, *Transport*, and *Commercial and Public Services*, while retaining key standalone sectors like *Residential* and *Agricultural*. Similarly, for fuel types, we filter to retain either aggregated categories (those starting with "Sum of all") or select a set of important fuels such as *Natural gas*, *Electricity*, *Heat*, and *Non-renewable waste*. However, while mapping, we drop ‘final energy consumption’ and ‘sum of all fuel products’’ because they represent the aggregate values only. This approach ensures meaningful representation of energy use patterns while minimizing the dimensionality introduced by one-hot encoding. It also simplifies the model, avoids overfitting, helps in faster training of our machine learning models and makes the data domain relevant as well.

Sector plots:

After choosing key sectors, we plot a line plot to see how total energy consumption in each high-level sector (e.g., Industry, Transport, Residential, etc.) has changed over time.





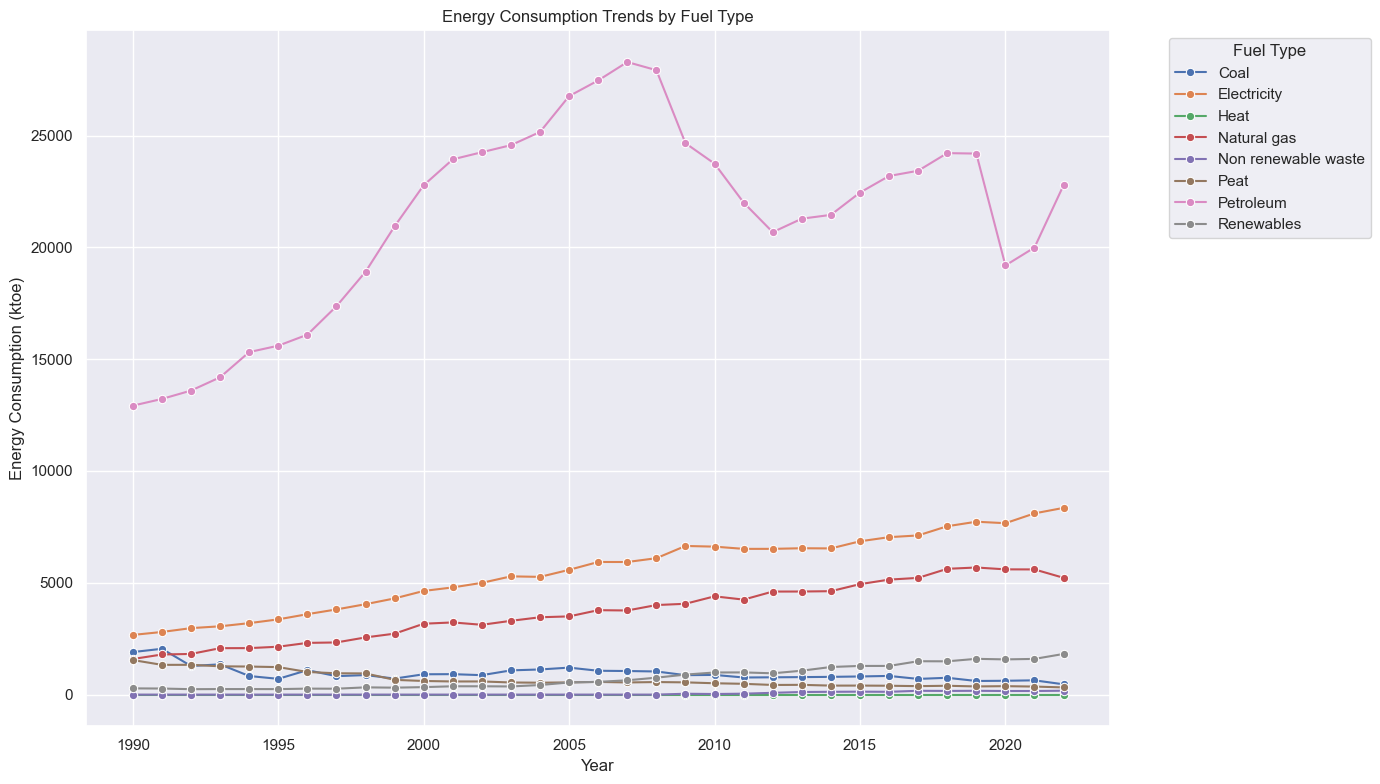
The above plot shows that the highest amount of fuel consumption has been in the transport sector across all the years from 1990 to 2022. This is followed by the residential sector and Industrial sector respectively. However, the lowest consumption is in the fisheries and agricultural sector.

Moreover, this plot also shows that overall, the highest amount of fuel consumption was in the year 2007 amounting to around 17k (ktoe) . Although after 2007, it started decreasing till 2012, however, after that it started rising again and reached a peak of around 16k (ktoe) in 2019. Post-2007 dip could suggest an economic slowdown reducing industrial and transport activities. So we can say that fuel consumption in the transport sector has been fluctuating over the years, which could be due to changing vehicle use, fuel prices, or mobility trends.On the other hand, the fisheries and agricultural sector’s consumption pattern remained somewhat stable throughout.

Therefore, using this information, we can assume higher CO2 emissions by the transport sector due to higher fuel consumption as well. This analysis can be used by policy makers to devise incentives to reduce fuel consumption thereby reducing CO2 emissions. Since Transport dominates consumption, therefore, policymakers can target decarbonization efforts there — such as promoting EV adoption, public transport, or fuel efficiency regulations. Moreover, Industrial and Residential sectors might benefit from renewable integration, insulation improvements, or energy-efficient appliances.

Fuel Type plots:

After choosing key fuel types, we plot a line plot to see how total energy consumption in each fuel type has changed over time.

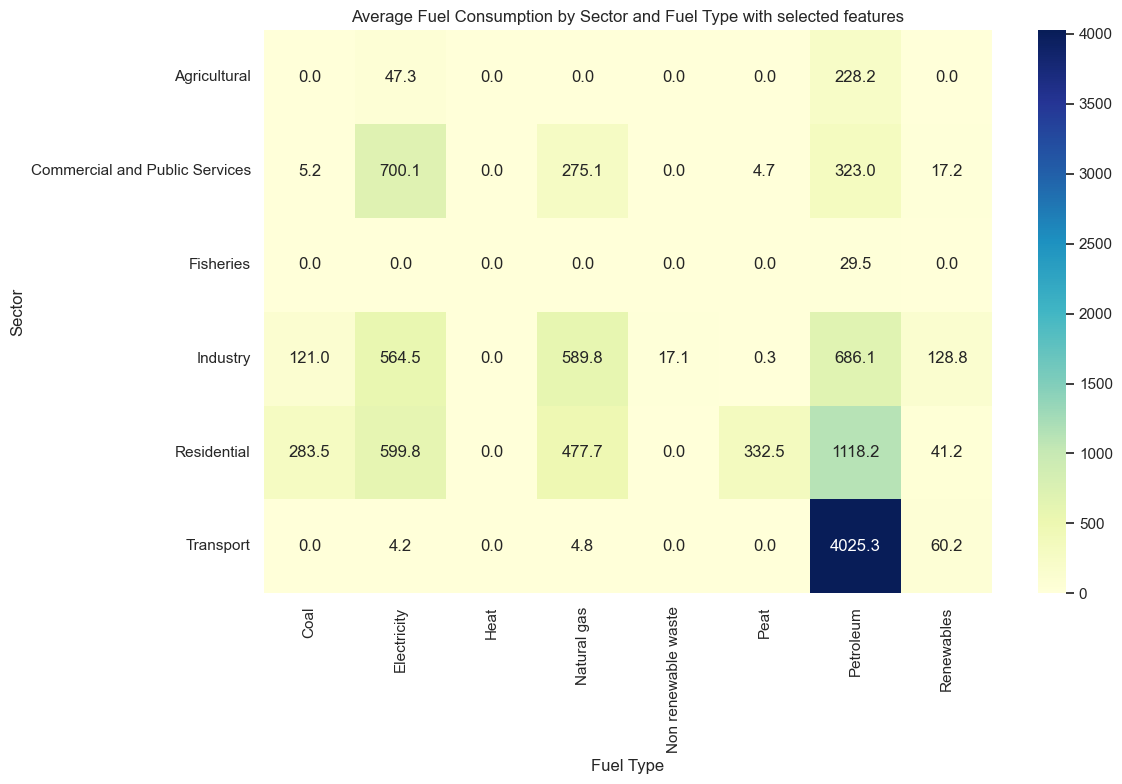


The above plot shows that the highest amount of fuel consumption has been in the petroleum category across all the years from 1990 to 2022. This is followed by electricity and Natural gas, respectively. However, the lowest consumption is in the heat and non renewable waste category. .

Moreover, this plot also shows that overall, the highest amount of fuel consumption was in the year 2007 amounting to around 28k (ktoe) . Although after 2007, it started decreasing till 2012, however, after that it started rising again and reached a peak of around 24k (ktoe) in 2019. Post-2007 dip could suggest an economic slowdown or 2008 financial crisis, reducing the consumption of fuel products , particularly petroleum.

Further, we can say that fuel consumption of coal has been fluctuating over the years, which could be due to changing prices, or mobility trends.On the other hand, electricity, natural gas and renewables consumption patterns have increased over the years. The particular rise in renewables after 2009 may be due to the 2009 Renewable Energy Directive that requires Ireland to meet 16% of our energy requirements from renewable sources by 2020.

2. For understanding the relationship between my categorical and numerical values, a heatmap is plotted. Below is a heatmap of average fuel consumption (in ktoe) across selected sectors and fuel types. In the plot, colors represent consumption values and darker color = higher average consumption.

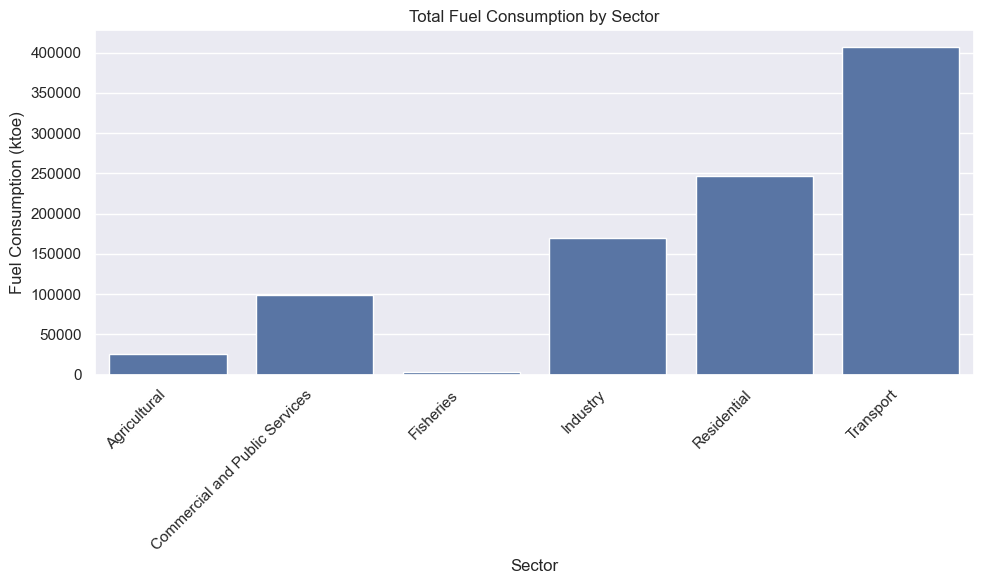


The heatmap shows the highest correlation between petroleum and transport which indicates that the transport sector relies heavily on fuel. Row patterns reveal which fuels dominate in each sector. We can also identify sectors that are more non-renewable fuel dependent v.s. those transitioning towards renewables e.g. the highest consumption of renewables is in the industrial sector as compared to other sectors.

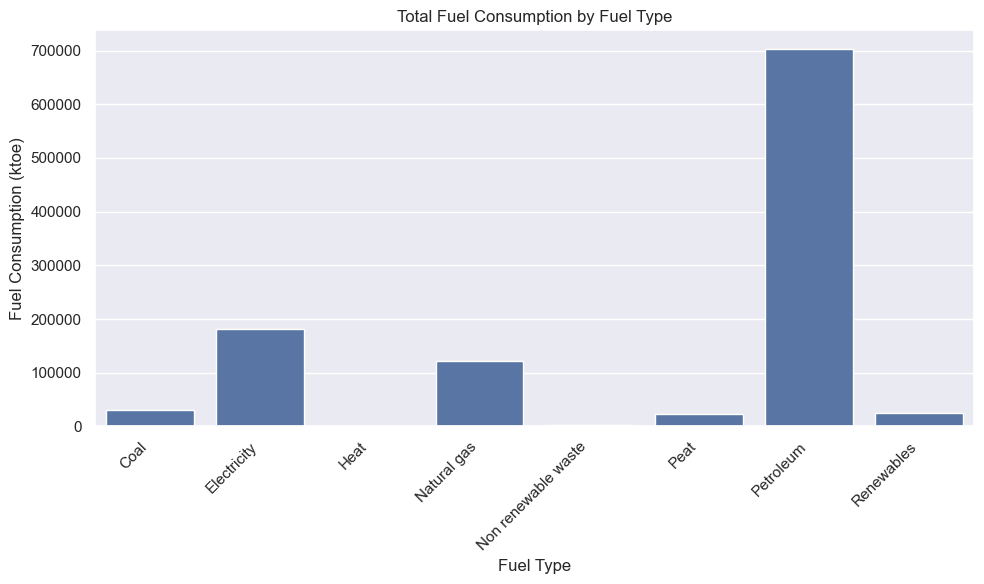
Another interesting insight is the high consumption of electricity in commercial and public services. This could indicate inefficiencies in heating, cooling, and lighting systems. This information can be used by policy makers to incentivize moving to green energy initiatives e.g. urging this sector to have on-site solar installations in commercial and government buildings and using green public procurement to ensure government facilities use clean energy.

Lastly, heat has no consumption in any sector i.e. it is redundant. Since it contributes zero to overall trends or sectoral analysis, i.e.it adds no analytical value and clutters visualizations so we can remove it for our analysis, later.

Fuel consumption by sector:



Fuel consumption by fuel type:



Question3: Interactive Dashboard Development Using Tufte’s Principles:

Modern energy systems are highly data-driven thus requiring the development of web-apps, mobile-apps and interactive dashboards. In compliance with developing a dashboard for my project on the analysis of energy statistics in Ireland, I have coded and developed an interactive dashboard ‘Energy ML Dashboard’ using Streamlit. Streamlit is an open-source Python library that makes it easy to create and share custom web apps for machine learning and data science. By using Streamlit, we can quickly build and deploy powerful data applications. The dashboard has been designed to depict my Machine Learning results.

Data Summary:

My dataset consisted of Fuel Consumption values across various sectors and fuel types in Ireland. I applied Regression models on my numerical data to predict fuel consumption trends in the future. I particularly used Decision Tree (DT) and Random Forest (RF) for my machine learning modelling. After training and testing my data, I produced three results:

1. A table for comparing R2 test and training score, RMSE, MAE values between DT (Iqbal,2025) and RF.
2. A plot for comparing Actual and Predicted values
3. A table for comparing mean R2 score in response to changing cv values i.e. cross validation scores for cv=5,10 and 15. These values were compared for both DT and RF.

The R2 score was found to be better for RF indicating its superior performance as compared to DT.

This data i.e. my results were then used to develop the interactive dashboard, in accordance with the Tufts principles.

In response to this, I developed a coded interactive dashboard using Streamlit (instead of Power BI, Tableau, or RapidMiner) to showcase the results of machine learning (ML) analysis applied to energy data. The dashboard was designed in alignment with Edward Tufte’s principles of data visualization, emphasizing clarity, precision, and efficient data communication. The Tufts principles are : Show data, maximize data-ink ratio, use precise and efficient graphics, show multivariate data, integrate text with visuals and encourage comparison. Keeping in mind these principles, I have divided my dashboard into four sections:

Dashboard Sectioning:

Section one : Model Evaluation Metrics

Section two : Actual vs Predicted (Log-Log Scale)

Section three : Cross Validation Results

Section four : Download Center

Development Approach:

The dashboard was developed using the Python programming (Weiss,2025) language, leveraging the following libraries. I exported my results from the jupyter file into the ml\_results.pkl file. Then I made an ‘[app.py](http://app.py)’ file in which I loaded my results and wrote my code for developing the dashboard. The following libraries were used to load the results, develop the dashboard and make it interactive.

* Streamlit: Streamlit is an open-source Python library that makes it easy to create and share custom web apps for machine learning and data science. By using Streamlit we can quickly build and deploy powerful data applications.
* Plotly: Plotly is a Python (PyPI, 2025) library which is used to design graphs, especially interactive graphs.
* Joblib: Joblib enables multiprocessing across several machines or cores on a single machine, which enables programmers to parallelize jobs across numerous machines. joblib. dump() and joblib. load() provides a replacement for pickle to work efficiently on arbitrary Python objects containing large data, in particular large numpy arrays. Joblib is a set of tools to provide lightweight pipelining in Python for loading pre-trained ML models and their results.
* Pandas: Pandas (Pandas Development Team, 2025) is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.
* Numpy: NumPy (DataCamp, 2025) is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform, and matrices.

Steps:

1. Machine Learning Preparation: As discussed above, I applied RF and DT models on my data to get R2 score, cross validation scores and actual vs predicted plot and these results were stored using joblib.
2. Streamlit Dashboard Construction:

I divide my dashboard into four logical sections (discussed above). To make the dashboard interactive, I used two tabs in the first sections for navigating between metrics and plots. I also provided user controls such as radio buttons and dropdowns for interactivity and comparison. Button could help the user choose DT , RF or both for visualising their results. In the second section of the plot, I used tool tips to see data points on the plot. Tooltips (also called hover labels) in Plotly are the little pop-up boxes that appear when you hover your mouse over a part of a chart. In the third section, I again provided radio buttons to see RF, DT or both of their CV results. In the last section, I enabled boxes to help the user download the statistics of results in a csv file

Visualization Choices and Tufte’s Principles

Each section of the dashboard was developed with a direct application of Tufte’s principles of graphical excellence.

Section 1: Model Evaluation Metrics Table and Bar Chart:

For developing this, I used the principles of minimal chartjunk and comparison. I did not include any unnecessary legends, borders, and grid lines and used bar plots to enable side-by-side comparison of metrics across models. I also followed the principle of data density by presenting my metrics in compact tables and their comparison in a clean bar chart.

Design Rationale:  
While designing tables, highlight\_max() function was used to identify best-performing models quickly. Moreover to avoid clutter, horizontal radio buttons and tabs were used to segment the content.

Section 2: Actual vs Predicted (Log-Log Scale) Visualization:

For developing this, I used the principles of clear comparison and small multiples.   
Scatter plots on log-log scale were used to allow clear visibility of deviations across magnitudes. Besides this, the side by side comparison of DT and RF allowed effective visual comparison. I also limited my use of colors, labels, etc. to focus attention on the core message of model accuracy thereby following the Data-Ink Ratio principle.

Design Rationale:  
I used a logarithmic scale to handle a wide range of prediction values and maintain visual clarity. The red dashed line represented the best predictions serving as a strong reference for deviations.

Section 3; Cross-Validation Summary

For developing this, I used the principles of integrity, simplicity and comparability. The raw cross-validation results were shown transparently and the users could use buttons to select and visualise a particular models’ performance across cv folds. A clean table layout was provided for ease of understanding.

Section 4: Download Center

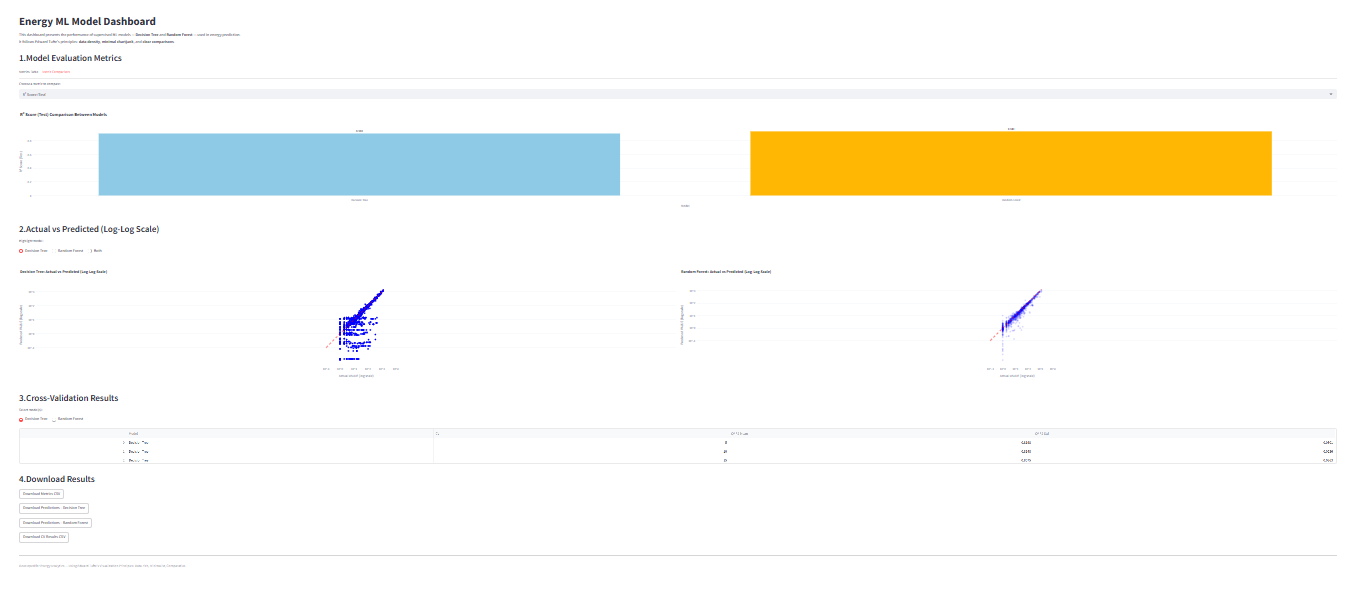
For developing this, I used the principles of function over form according to which the users were enabled to extract raw insights in CSV format for external validation or custom reporting.

User Interaction and Interface Design

* Wide Layout: Used layout="wide" to maximize screen real estate and avoid overwhelming vertical scrolling.
* Color Encoding: Limited palette (blue for Decision Tree (Iqbal,2025), amber for Random Forest) ensured consistent model tracking across visual elements.
* Opacity and Hover Controls: Used subtle visual cues (e.g., opacity, hover toggles) to emphasize selected models while de-emphasizing others without complete removal.

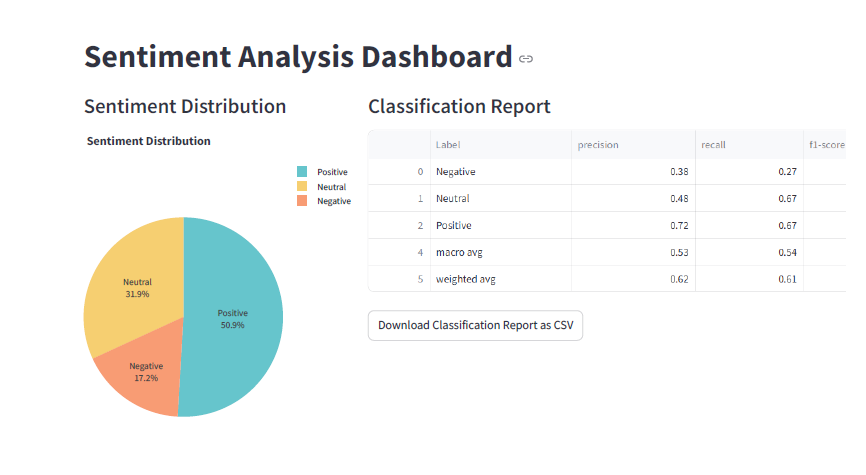
Conclusion

This dashboard demonstrates how Energy data related ML statistics can be represented in accordance with Tufte's principles. Following them not only enhances the interpretability of the results but also reduces clutter and maximizes clarity. In this way, users can navigate through the metrics and visualise the comparison among models. Thus by emphasizing on comparative and contextual insights, the dashboard serves as a technical tool as well as an effective communication medium, for decision-makers in the energy sector.



EXTRA WORK:

Apart from this I also developed dashboard for my Sentiment analysis

:

Now we will move to the Statistics section:

## Section2: Statistics for Data Analytics

Q1. Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models.

Q2. Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with international commerce, you could find a confidence interval for the population proportion of yearly gas energy imports out of all energy imports).

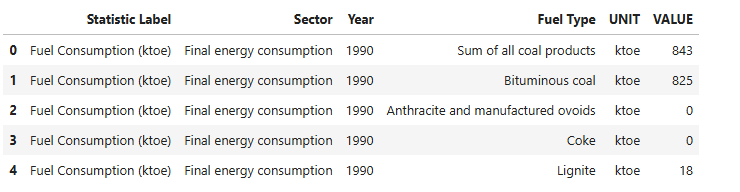
Q3. Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests.

Q4. Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process.

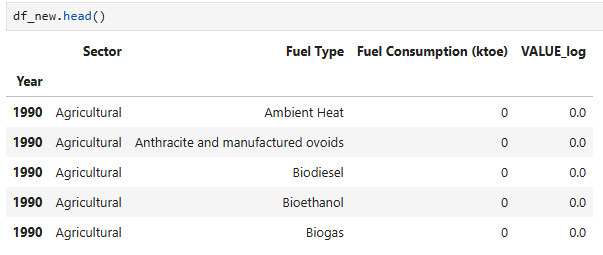
**Question 1: Statistics for Data Analytics**

Descriptive statistics and Plots:

The original dataset looks like this:



The dataset after formatting looks like this:



For descriptive statistic we need:

Central tendency: mean, median

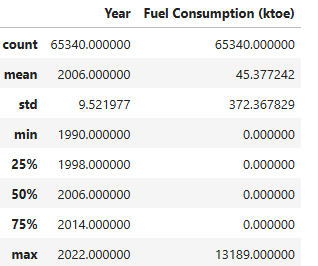
Spread: standard deviation, range, IQR

Shape: skewness, kurtosis

In our code, in order to get the summary of statistics for numerical columns, we use:

df\_new.describe()

From the data of Fuel Consumption (ktoe), we can deduce:



Fuel Consumption (ktoe)

● Mean: 45.377

● Median (50%): 0

● Max: 13189.0

● Min: 0

● Std Dev: 372.367

Insights: Over 75 percent of the data is zero, indicating extreme sparsity. Minimum is zero indicating that some fuels/sectors have no consumption in certain years. However some entries are very large, inflating the mean and standard deviation. These high values may be due to national totals or large sectors like transport/electricity. The distribution (Ahmed,2025) is highly skewed indicating the need for a log-transformation for modeling in order to compress these large differences.

Interquartile Range, Skewness and Kurtosis:

From code IQR:

iqr\_value = df\_new['VALUE(ktoe)'].quantile(0.75) - df\_new['VALUE(ktoe)'].quantile(0.25)

The values is coming out to be:

Interquartile Range (IQR): 0.0

Insights: Here Q1, Q2 and Q3 are all zero. Therefore IQR=Q3-Q1 is also zero. This shows that at least 75 percent of our values are exactly 0 confirming heavy sparsity in our dataset.This also suggests that the majority of fuels/sectors have no consumption in certain years.

Skewness:

skew\_value = skew(df\_new['VALUE(ktoe)'])

The values for skewness is: Skewness: 18.78295494200202

Insights: Skewness measures the asymmetry of the probability distribution of fuel consumption. Zero skewness indicates a symmetrical distribution while skewness greater than zero shows positive skewness indicating that the tail on the right side of the distribution is longer or fatter than the left side.

In our case, *Fuel Consumption value* shows a high positive skew of 18.78, indicating that while most values are relatively low and there are a few very high consumption records pulling the distribution to the right. This huge positive skew is due to many small (mostly 0) values and a few very large ones.These findings highlight the presence of outliers and the need for robust methods like log transformations when modeling.

Kurtosis:

kurt\_value = kurtosis(df\_new['VALUE(ktoe)'])

The kurtosis is: Kurtosis: 473.44782193280327

Insights: Kurtosis measures the "tailedness" of a probability distribution and indicates the presence of outliers in the dataset. A normal distribution has kurtosis = 3 while a kurtosis greater than 3 indicates that the dataset has heavy tails or outliers. This means that extreme values are more frequent than expected in a normal distribution.

The value of kurtosis in our case is extremely high, representing heavy tails and sharp peaks and suggests the presence of extreme outliers. This may be due to certain sectors contributing to the bulk of consumption and dominating the total.

Range, Variance, Coefficient of variation:

From code:

print(f"Range: {range\_value}")

print(f"Variance: {variance\_value}")

print(f"Coefficient of Variation: {cv\_value}%")

The values are as follows:

Range: 13189

Variance: 138657.79978816962

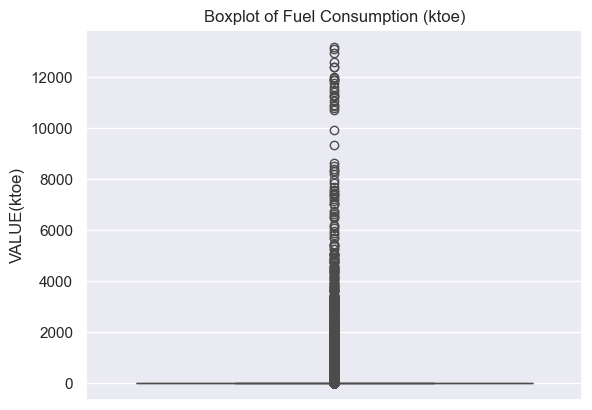
Coefficient of Variation: 820.6048037172725%

Insights: The range of 13,189 ktoe and variance of 138657.7 ktoe (leading to standard deviation of 372.37 ktoe calculated earlier) reflect the wide variability in consumption, driven by a small subset of high energy consuming fuels and sectors. The coefficient of variation (CV) is a measure of the dispersion of data points around the mean in a series. Here, the coefficient of variation (820.60%) further supports the fact that the dataset exhibits very high relative variability.

These patterns justify the use of log transformation to normalize the data and support the application of tree-based machine learning models that are robust to skewed and sparse distributions.

*Plotting:*

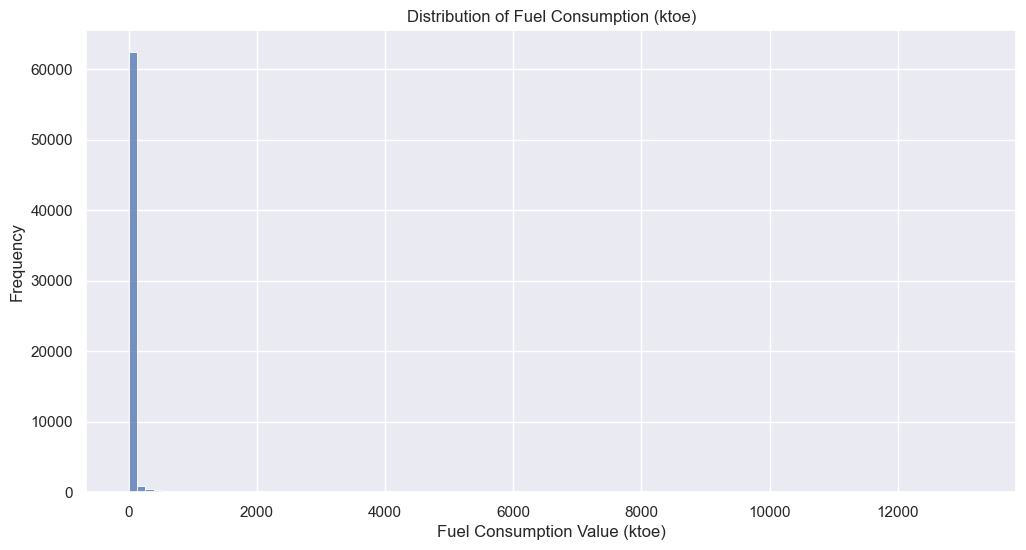
1. Boxplot: To examine outliers in Fuel Consumption



The box plot shows a compressed box at zero. Since 25th, 50th, and 75th percentiles are all 0, the entire box has collapsed to a thin line at 0. Any non-zero value lies beyond the whiskers because the IQR is 0. All non-zero values are considered extreme outliers, plotted individually. Due to the max value being 13,189, the vertical scale is highly skewed.

Overall, the boxplot confirms that distribution (Ahmed,2025) is heavily right-skewed, with the majority of entries being zero i.e. majority of fuels/sectors have no consumption in certain years with the outliers indicating a small number of large fuel consumption values in certain years.

1. Distribution plot to observe skewness:



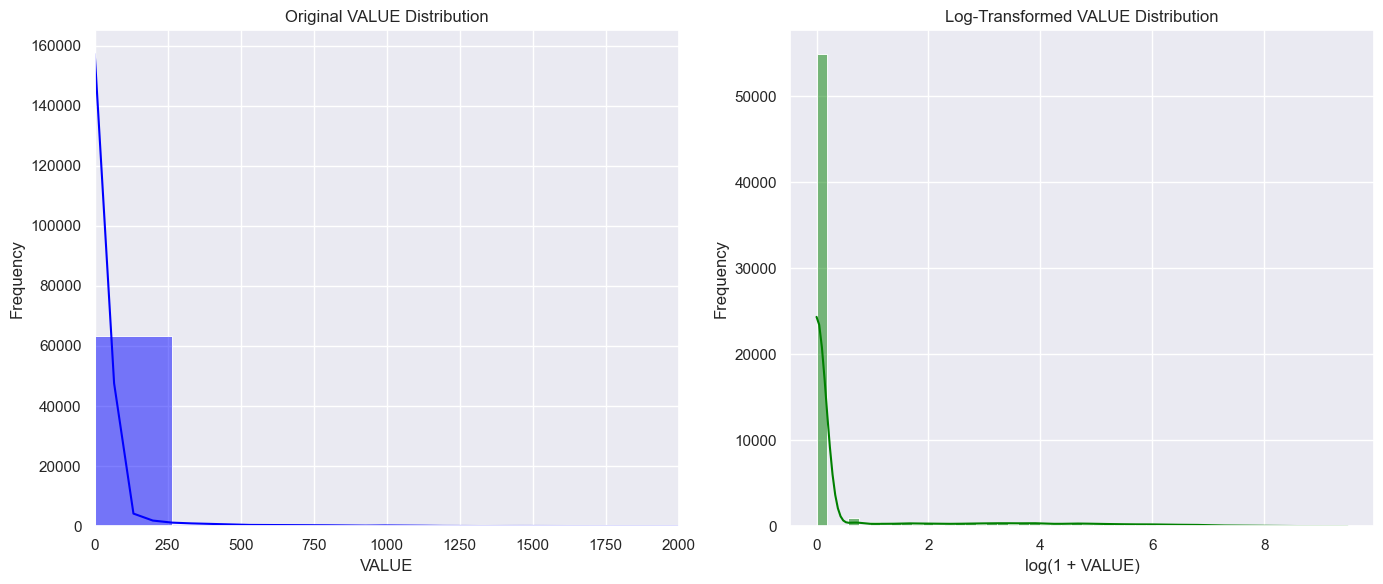
The histogram shows an extreme right-skewed (positive skew) distribution. A vast majority of data points are concentrated near zero, with very few instances of higher fuel consumption values. This is consistent with the earlier computed skewness of 18.78, indicating a heavy concentration of low values. As seen from the plot, zero values are dominant i.e. most records in the dataset have a VALUE of 0, as confirmed earlier by the median, Q1 and Q3, all being zero. This results in a single, very tall bar in the histogram near zero, overshadowing the distribution of the rest of the values.

While the overall standard deviation is high due to a few large consumption records, 50% of the data points lie exactly at zero.The highly imbalanced nature of the distribution justifies the need for data transformation e.g. log transformation.

Histogram before and after log transformation:

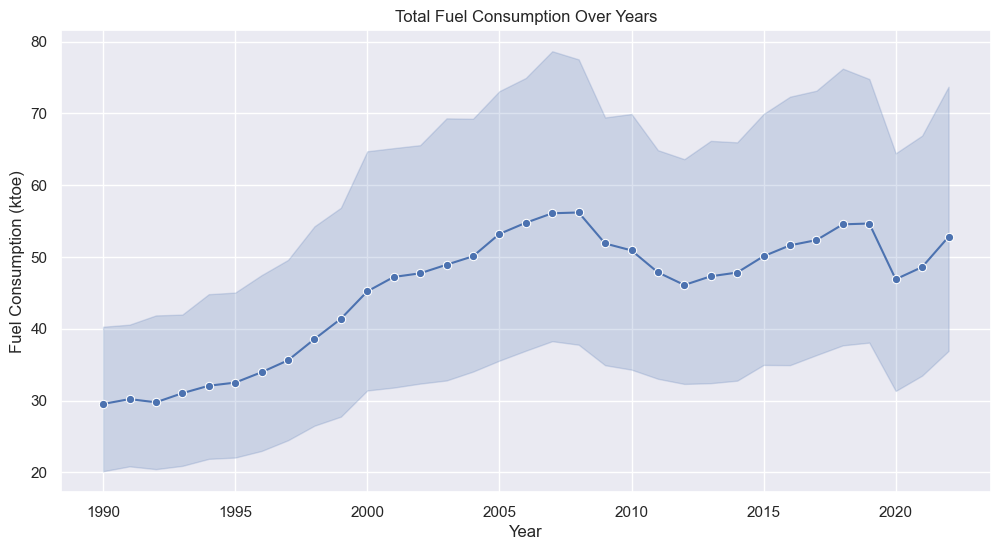
Since the original data is highly skewed, I have applied log transformation in order to compress the range of large values while spreading out the smaller ones, making the distribution more symmetric. I have applied log transformation using np.log1p(x) (which computes log(1 + x)) instead of just np.log(x) for mathematical stability and practical handling of zero values in my data. Since my VALUE column has a lot of zero values and logarithm function is undefined for zero therefore log(1+x) is a better choice.

Figure below shows the two histograms side by side, one for the original values and the other for the log-transformed values.



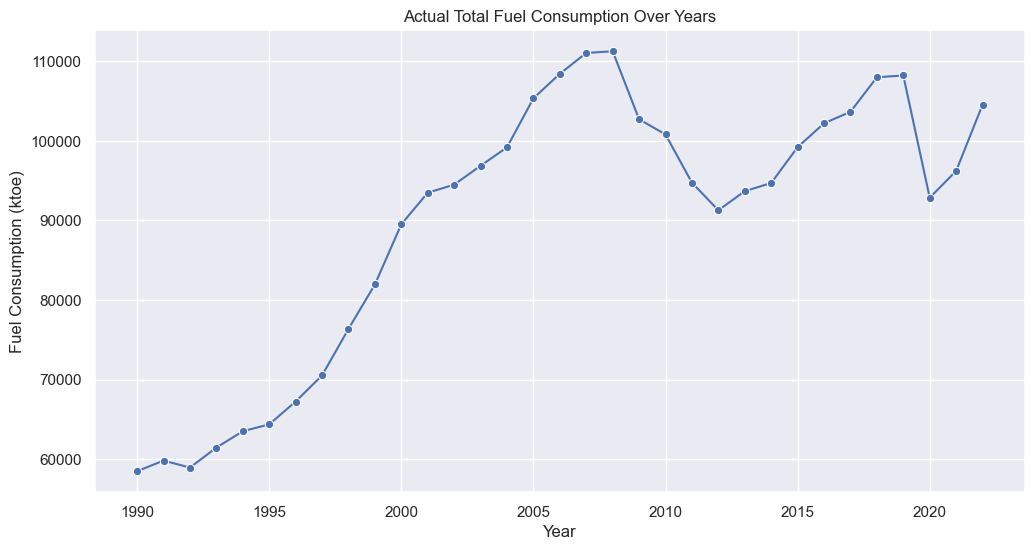
The above plots show the distribution of fuel consumption before and after log transformation. The left figure shows that the original data is highly skewed and indicates extreme usage patterns i.e. either a lot of zeros or some extreme values. This ultimately results in outliers as indicated by boxplot plotted earlier. The figure on the right side shows log transformed data, which is again still skewed due to the large number of zeros in the original data. This shows that data has not been much normalized even after log transformation. This means that we may need to remove zeros values or scale our data further.

1. Line plot: To observe the relationship of total fuel consumption with number of years



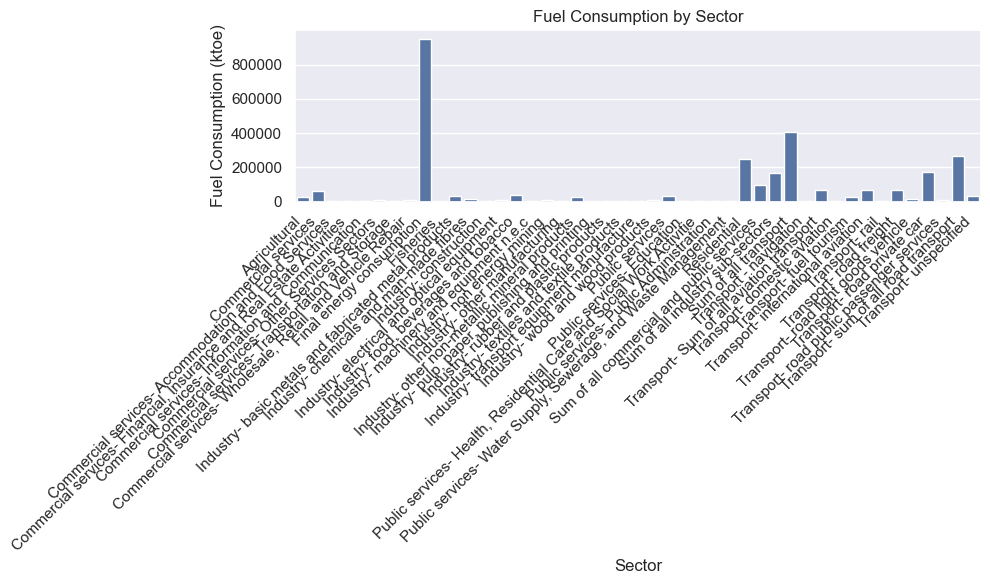
The above line plot of fuel consumption over years, without aggregation, simply connects individual data points corresponding to different fuel types and sectors for each year. This approach results in a cluttered visualization that misrepresents the overall yearly trend due to multiple values per year.

In order to accurately capture the total fuel consumption pattern over time, it is essential to aggregate the data i.e. by summing fuel consumption values for each year. This will provide a clear and meaningful trend of fuel consumption from 1990 to 2022.



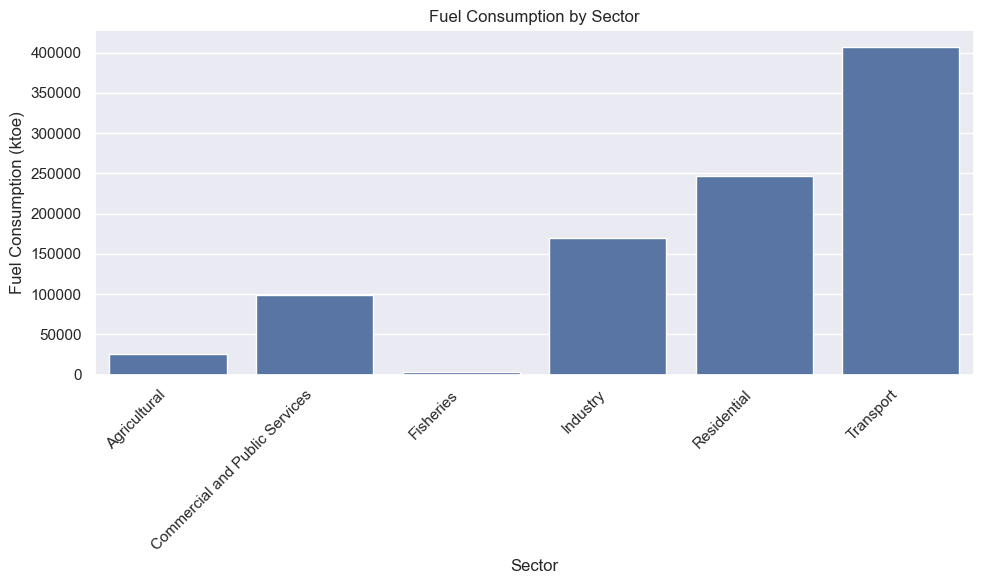
This line plot shows the overall trend of fuel consumption from 1990 to 2022. The plot shows peaks in the years 2007 and 2008, indicating highest fuel consumption of more than 110000 ktoe in those years, in Ireland. The fuel consumption decreased till 2012, after which it began increasing again reaching close to 110000 ktoe in 2018-2019. Overall, this graph gives more reliable insights for analysis and policy making for energy consumption trends in Ireland.

1. Bar plot: To visualise total fuel consumption per sector.



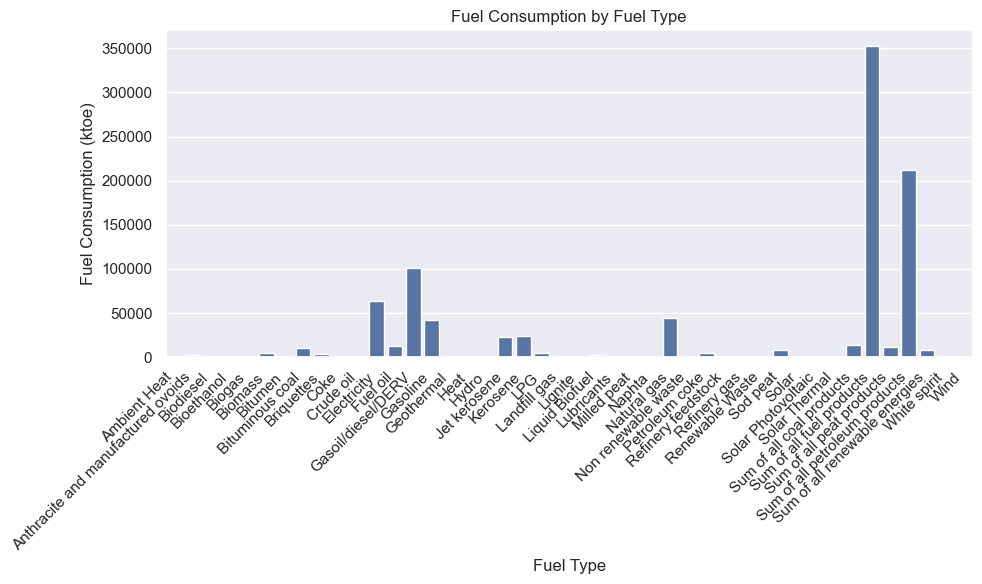
The graph is very cluttered. Therefore we can select a few relevant sectors which give us the aggregate across all subcategories.

The barplot below shows total fuel consumption across key sectors.



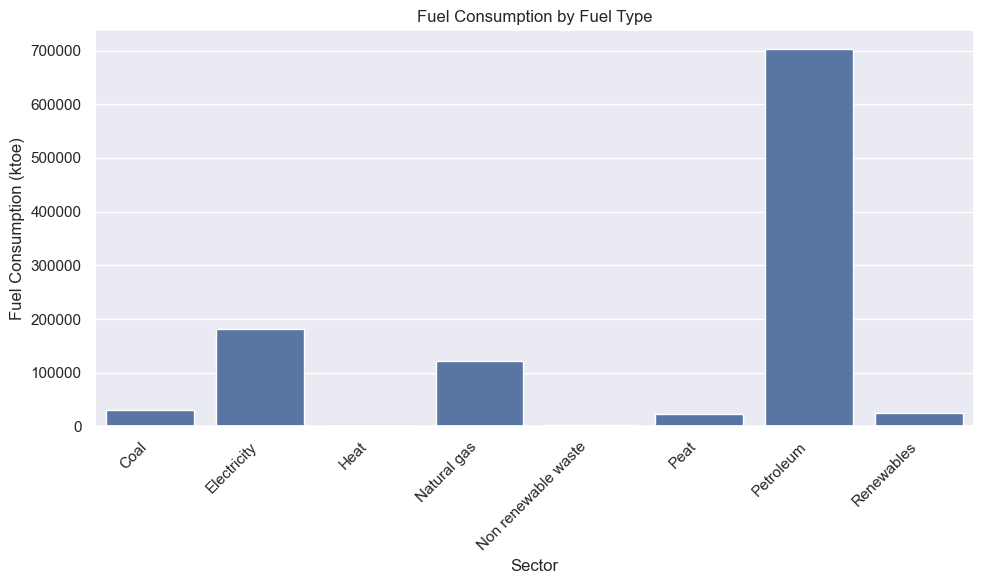
This bar plot illustrates the total fuel consumption across key sectors in Ireland from 1990 to 2022. The Transport sector has the highest fuel consumption, accounting for approximately 400,000 ktoe, followed by the Industry and Residential sectors. In contrast, the Fisheries sector shows the lowest fuel consumption, with the Agricultural sector coming next. Fuel consumption varies significantly across sectors, with Transport, Industry, and Residential being the top three consumers. Overall, this graph provides valuable insights into energy consumption trends in Ireland, which can be utilized for further analysis and policymaking across various sectors.

1. Bar plot:To visualise total fuel consumption per fuel type.



The graph is also very cluttered. Therefore we can select a few relevant fuel types which give us the aggregate across all subcategories along with some standalone ones.

The barplot below shows total fuel consumption across key fuel types.

.

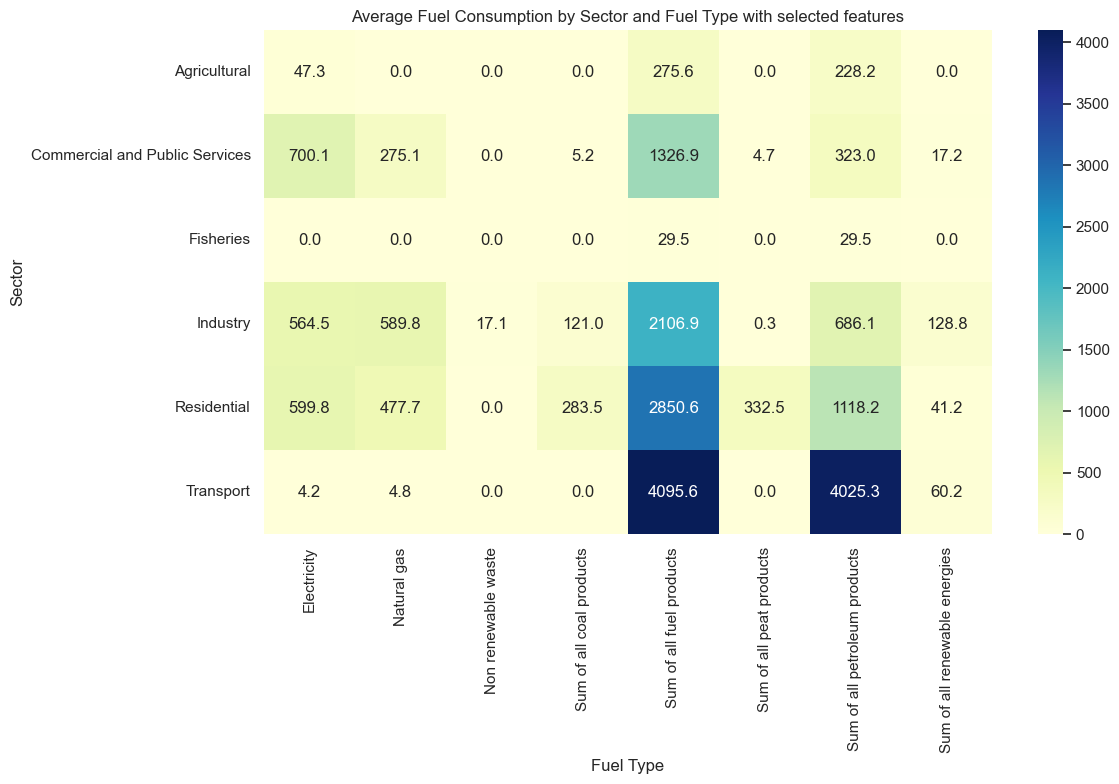
This graph illustrates the total fuel consumption across key fuel types in Ireland from 1990 to 2022. The data reveals that the highest fuel consumption is from the category "Sum of all fuel products," which exceeds 1,000,000 ktoe, followed by "Sum of all petroleum products" at more than 1,700,000 ktoe.

In contrast, the "Heat" category shows no fuel consumption, while the lowest consumption is observed in "Non-renewable waste," followed by "Sum of all peat products." Although fuel consumption varies across fuel types, the categories "Sum of all fuel products," "Sum of all petroleum products," and "Electricity" are the primary contributors to the highest fuel consumption.

Overall, this graph provides valuable insights into fuel consumption trends in Ireland, which can inform policymakers on areas to focus on, particularly when considering investments in renewable and non-renewable energy sources.

(For better visualisation, we can plot on a log scale as well. Plot shown in jupyter notebook)

1. Heatmap: To show the average fuel consumption for each Sector and Fuel Type.



This above plot shows a heatmap where the color intensity indicates the level of average fuel consumption and lighter or darker shades of blue/yellow represent lower or higher consumption, respectively.

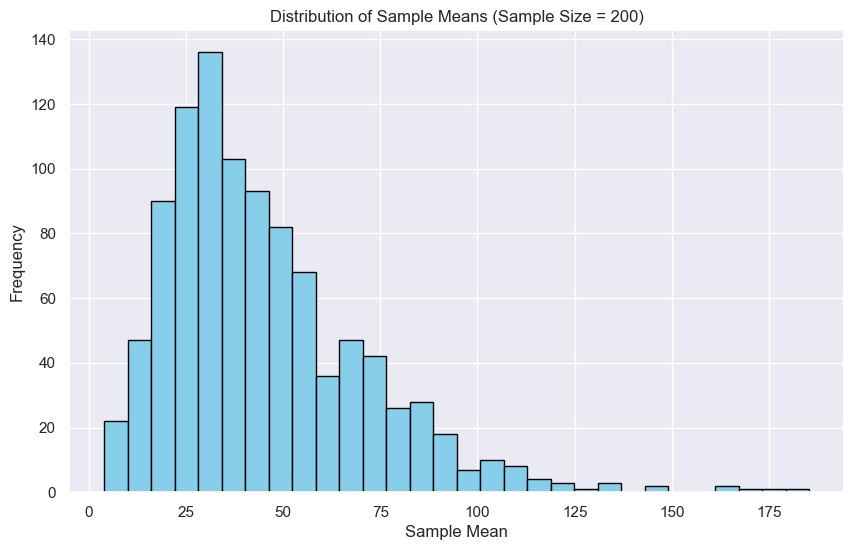
This heatmap visually compares the average fuel consumption of selected fuel types (e.g., electricity, natural gas, petroleum products) across various sectors in Ireland. It highlights which fuel types are dominant within each sector and which sectors depend most on particular [fuels. In](http://fuels.in) the map, the color intensity indicates the level of average fuel consumption and lighter or darker shades of blue/yellow represent lower or higher consumption, respectively.

From the figure, it is evident that the Transport sector shows a strong association with both *‘Sum of all fuel products’* and *‘Sum of all petroleum products’*, indicating that this sector consumes the highest amount of fuel, particularly petroleum-based. In contrast, there is minimal to no correlation between the Agricultural sector and Natural gas, suggesting limited use of this fuel in agriculture. Additionally, the Commercial and Public Services sector appears to consume the highest amount of electricity among all sectors. Overall, this visualization reveals clear energy consumption patterns across sectors and fuels, providing valuable insights for energy planning and targeted policy-making.

*Central Limit Theorem:*

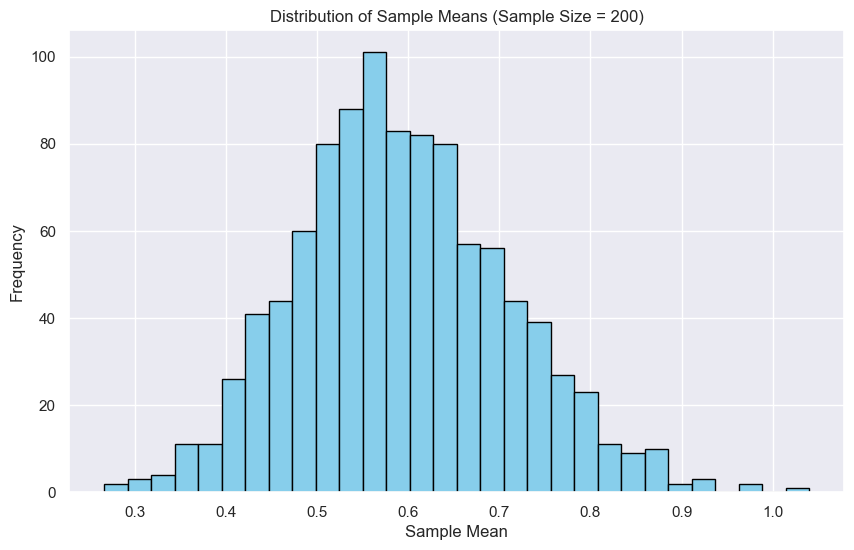
The Central Limit Theorem (CLT) states that the distribution of the sample mean of a random variable approaches a normal distribution as the sample size increases, regardless of the shape of the original distribution, under certain conditions. In our case, the fuel consumption data is not normally distributed. By repeatedly drawing random samples from the fuel consumption data and calculating their means, we observe that the distribution of these sample means approximates a normal distribution. Applying CLT in this context enables valid use of parametric statistical techniques, such as confidence intervals or hypothesis testing, on the sample means, thus strengthening the analytical foundation of the study despite the non-normality of the raw data.

The histogram below shows the distribution of sample means drawn from the fuel consumption values.



Despite the original values being highly skewed and non-normal, the histogram of the sample means appears close to normal (a bell-shaped curve) although it is still positively skewed. This confirms the theorem that even with skewed fuel consumption data, the sampling distribution of the mean becomes nearly normal.

For a better visualisation, we can take log of the values. This gives us the histogram shown below:



The histogram shows a bell-shaped, approximately normal distribution of sample means and confirms that the means of random samples from the fuel consumption data follow a normal distribution due to CLT. This provides a theoretical and visual basis for applying statistical inferential techniques to our dataset.

*Distributions:*

Discrete Distribution:

*Using Poisson Distribution:*

The Poisson distribution is used when:

● We are modeling the number of events in a fixed interval (time, space)

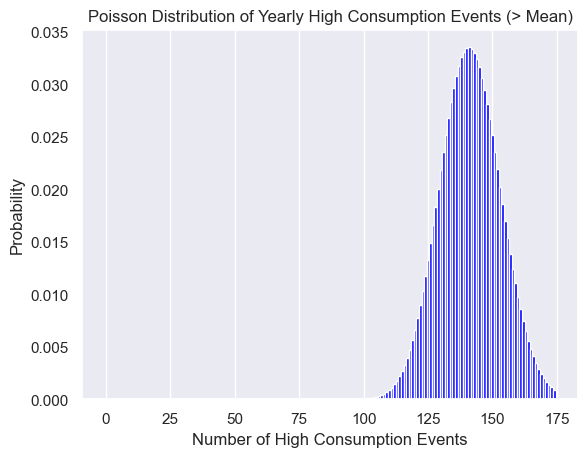
● The events occur independently, and the average rate (λ) is known.

Use case for my dataset:

Case1: Count the number of years with very high fuel consumption values (e.g., > threshold), assuming these are rare events. I will have to use a threshold since in the original dataset VALUE is a numerical column with continuous values and Poisson is a discrete distribution which cannot be applied on continuous values directly.

Steps: First we calculate the mean of all consumption values. Then count how many high consumption events occurred each year (i.e. when consumption is greater than the threshold ). After that we compute the average number (λ) of such events per year and generate poisson probabilities for a range of event counts (k = 0 to max).These are the theoretical probabilities of observing k high consumption events in any given year.

The figure below shows the poisson distribution for modelling and visualising the number of individual high fuel consumption records per year:

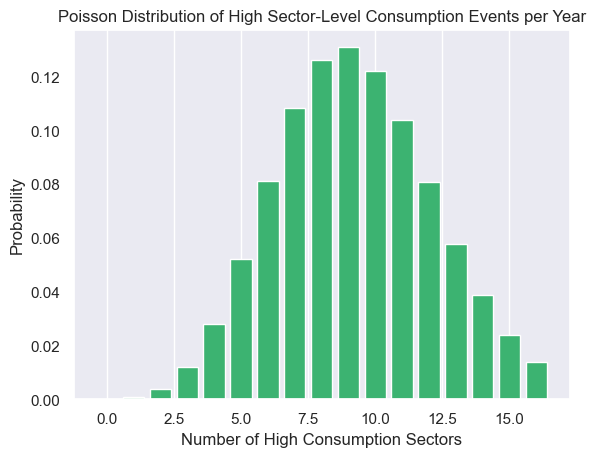


The figure shows that by setting the dataset’s mean fuel consumption as a threshold, we can identify and model “high consumption” events using a Poisson distribution. This model provides a theoretical expectation for how many high consumption cases occur each year, which helps in detecting the years with statistically unexpected energy surges.

Case 2: I can apply Poisson distribution to the Secor column as well since it has categorical values (discrete data), and I can count how many times each Sector appears per year, thus applying discrete distribution.

Steps: First we group by year and sector to get the total energy consumption for each sector within a given year. Then we find the average of sector consumption which is used as a threshold to determine whether a sector's consumption is high or not. After defining the high consumption sectors, we count high consumption sectors per year. Finally we calculate the average number of high consumption sectors per year i.e. λ for Poisson Distribution.

The figure below shows the poisson distribution to model the occurrence of high consumption sectors within each year.

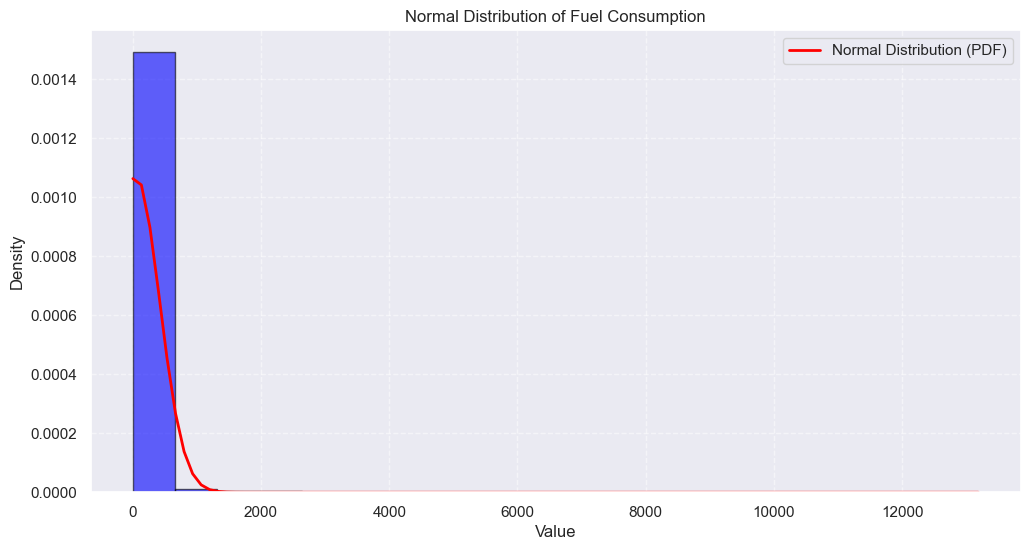


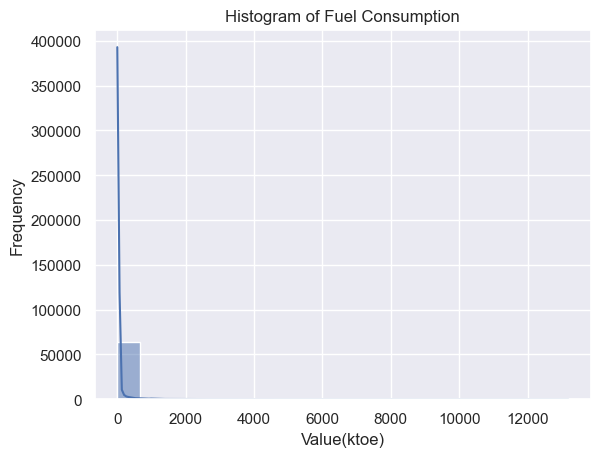
The plot reveals which counts of high consumption sectors are most probable in a typical year. For example, the peak is at 9, which means that observing 9 high consumption sectors per year is most common. Thus this distribution can help understand how likely it is to observe a certain number of high consumption sectors each year.

Rationale for using Poisson: The Poisson distribution is appropriate for modeling count data, especially when we're interested in the number of times an event occurs within a fixed interval (e.g., yearly fuel consumption from a specific sector).

*Normal distribution:*

Plotting histogram for VALUE column to see if it is following normal distribution





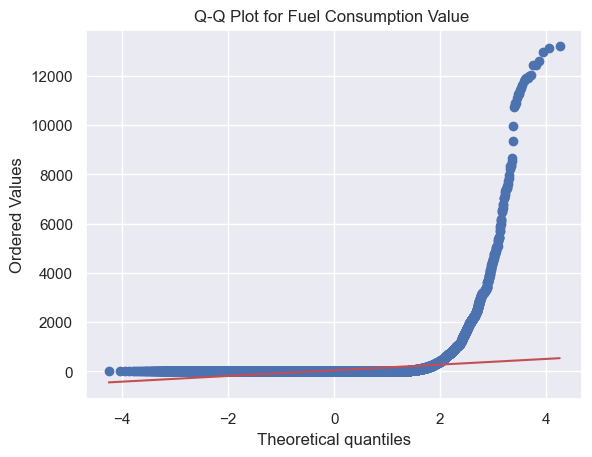
The histogram of fuel consumption values is showing a highly right-skewed distribution which indicates that most consumption values are clustered at the lower end, while a few observations have exceptionally high consumption.

This suggests that the data is not normally distributed, as a normal distribution would appear symmetric. Instead, it reflects majority zero values indicating no sector/fuel consumption in those years along with some occasional spikes in fuel consumption in certain years. This may be due to certain sectors contributing to the bulk of consumption and dominating the total. This skewness has important implications for analysis—such as the need for transformations or using non-parametric models that are less sensitive to non-normality.

QQ-Plot:

A QQ (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. The plot compares the quantiles of the observed data to the quantiles of the theoretical distribution. If the points on the QQ plot fall approximately along a straight line, this suggests that the data follows the assumed distribution. If the points deviate from the straight line, it indicates that the data does not follow the distribution.

The QQ-plot shown below also shows that data is not normally distributed.



Importance:

In this analysis, I used Poisson and Normal distribution (Ahmed,2025) to better understand the characteristics of the Fuel consumption in Ireland dataset. The choice of these distributions and the variables used were made based on the nature of the data, and their relevance to answering the research questions regarding fuel consumption across various sectors and fuel types and statistical modeling.

Importance of the Distributions Used:

Poisson Distribution: The Poisson distribution is used for modeling count data, particularly the number of events that happen in a fixed interval of time or space. It is ideal for modeling the count of sectors contributing to high yearly consumption. In real-life scenarios, this can be used to model the sectors demanding more fuel or the type of fuel contributing to higher consumption over the years.

Why Important: This distribution is helpful in modeling specific characteristics of the dataset if we want to understand how often certain events occur. It also provides insight into the probabilities of different categories or the frequency of occurrences, which can be critical for planning, policy-making, and forecasting the yearly fuel consumption.

Normal Distribution: The Normal distribution is the most widely used continuous probability distribution in statistics. It is important because many statistical models and machine learning algorithms assume that the data follows a normal distribution, particularly for the residuals (errors) in regression models, which are crucial for making inferences about the data.

Why Important:  
Normality Check: The Normal distribution helps to check whether the fuel consumption values are symmetrically distributed around the mean, which is a key assumption for many statistical techniques (such as Linear Regression).

Modeling Fuel Consumption: By checking if the fuel consumption data follows a normal distribution, we can better determine which modeling techniques to apply. For example, if the data is normally distributed, techniques like Linear Regression are suitable. On the other hand, if the data is skewed, transformations or non-parametric models may be necessary.

**Question 2: Inferential Statistics - Part 1**

Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with international commerce, you could find a confidence interval for the population proportion of yearly gas energy imports out of all energy imports).

To estimate the population mean fuel consumption and quantify uncertainty, a 95% confidence interval is computed using the t-distribution. Despite a large sample size, the use of t-distribution is justified due to the absence of the population standard deviation and the highly skewed nature of the data. The t-distribution offers a more conservative and appropriate interval estimate in such conditions.

Steps: After finding the average fuel consumption across our sample data, we use Standard Error of the Mean (SEM) to quantify the uncertainty around the sample mean as an estimate of the population mean.Lastly, we find the 95% confidence interval to get the range within which the true population mean fuel consumption is likely to lie with 95% confidence.

From code:

mean = np.mean(fuel\_values)

sem = stats.sem(fuel\_values) # standard error of mean

# 95% Confidence Interval

confidence\_interval = stats.t.interval(confidence=0.95, df=len(fuel\_values)-1, loc=mean, scale=sem)

This gives the following result:

Mean Fuel Consumption: 45.38 ktoe

95% Confidence Interval: 42.52 to 48.23 ktoe

Insights: Using inferential statistics, specifically a 95% confidence interval based on the sample mean and standard error, we estimate that the true average fuel consumption (in ktoe) across all sectors, fuel types, and years lies between 42.52 and 48.23 ktoe. This interval gives us a range within which the actual population mean is likely to fall, even though we are working with a sample.

This means we are 95% confident that the true population mean fuel consumption lies within this range. In this way, this method allows us to move beyond the specific sample values and draw a conclusion about the broader population. It is particularly useful for policymakers or utility companies in order to understand the expected average energy demand under uncertainty. The confidence interval reflects both the variability in the data and the reliability of the estimate.

In the next step, we find the confidence interval per sector as well. From code:

for sector, values in sector\_groups:

mean = np.mean(values)

sem = stats.sem(values)

ci = stats.t.interval(0.95, len(values)-1, loc=mean, scale=sem)

print(f"{sector} - Mean: {mean:.2f} ktoe, 95% CI: ({ci[0]:.2f}, {ci[1]:.2f}) ktoe")

This gives us the range in which the true mean fuel consumption per sector is likely to lie (with 95% confidence) giving us insight into the sectors consuming more or less fuel along with indicating how precise or variable the mean estimates are across sectors.

e.g. For agricultural sector, we get:

Agricultural - Mean: 17.71 ktoe, 95% CI: (14.47, 20.95) ktoe

From this, we are able to analyze the sector-wise fuel consumption pattern and estimate population-level behavior, with 95% confidence interval. In the above case, we estimate that the average fuel consumption for the agricultural sector is estimated to be 17.71 ktoe and we are 95% confident that the true mean fuel consumption for this sector lies between 14.47 and 20.95 ktoe.

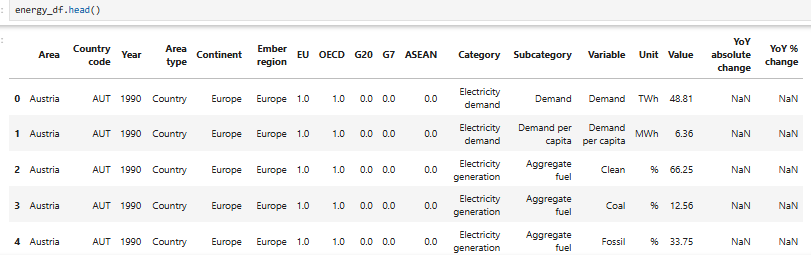
Thus, this inferential statistical method allows us to account for uncertainty in the sample data and make probabilistic inferences about the average consumption in each sector.

**Question 3:Inferential Statistics - Part 2**

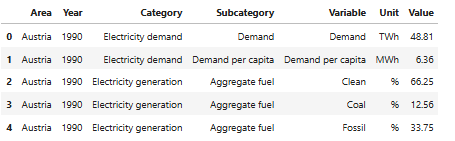
Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests.

Dataset description:

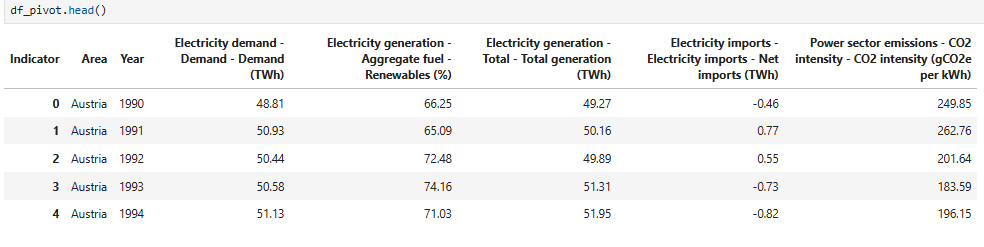
For this part, I have chosen a different dataset that includes statistics about different countries regarding energy demand, net imports, electricity generation, etc. from the year 1990 to 2024. This dataset named ‘europe\_yearly\_full\_release\_long\_format’ is taken from Ember (Ember,2024) and looks like this:



After dropping null values and selecting required columns, the dataset looks like this:



Finally, after selecting particular categories, sub-categories, variables and units and formatting, the dataset looks like this:



Objectives:

* The goal: To compare Ireland with other countries on energy metrics.
* The focus: Similarities/differences in variables like electricity demand, renewable energy generation, etc.
* The approach: Use parametric and non-parametric inferential statistical tests to justify findings.

Variable Selection & Assumption Testing:

* I have chosen two variables for my hypothesis testing:
  + Electricity demand (TWh)
  + Renewable electricity generation (%)
* For each variable, I am checking:  
  + Normality: Using Shapiro-Wilk test
  + Based on the results, parametric (e.g., t-test, ANOVA) or non-parametric (e.g., Wilcoxon, Kruskal-Wallis) tests are used.

*Variable 1: 'Electricity generation - Aggregate fuel - Renewables (%)'*

Test1: Shapiro-Wilk Test (Normality check)

Used to test if a single sample is normally distributed.

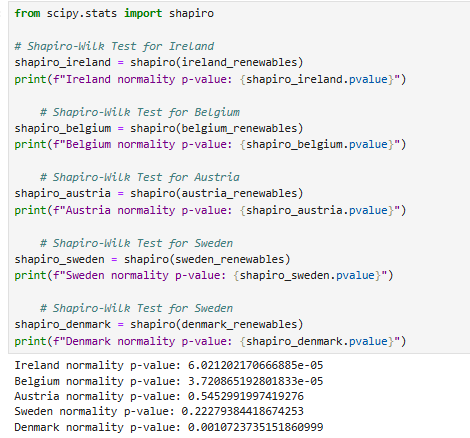
* Assumptions:
  + Data is continuous
  + Observations are independent

In my work, I have applied the shapiro test to the following countries: Ireland, Belgium, Denmark, Austria and Sweden.

Hypotheses: H0 is null hypotheses and H1 is alternate hypotheses

H0: Renewables Data is normally distributed.

H1: Renewables Data is not normally distributed.



The shapiro test gives p value. The p-value is a probability that helps you decide whether your observed data is significantly different from what you would expect under the null hypothesis. The p-value is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value of Ireland is less than 0.05

Therefore, Reject the null hypothesis. Ireland renewable electricity data is not normally distributed

p-value of Belgium is less than 0.05

Therefore, Reject the null hypothesis. Belgium renewable electricity data is not normally distributed

p-value of Austria is not less than 0.05

Therefore, accept the null hypothesis. Austria renewable electricity data is normally distributed

p-value of Sweden is not less than 0.05

Therefore, accept the null hypothesis. Sweden renewable electricity data is normally distributed

p-value of Denmark is less than 0.05

Therefore, Reject the null hypothesis. Denmark renewable electricity data is not normally distributed

Insights: From the above test, we can see that Ireland and Belgium renewable energy data is not normally distributed. Therefore, we cannot apply parametric tests and need to apply non-parametric tests on their data. Following this, I have applied Mann-Whitney U Test, Wilcoxon Signed-Rank Test and Chi-Square Test on Ireland and Belgium data for renewable energy generation %.

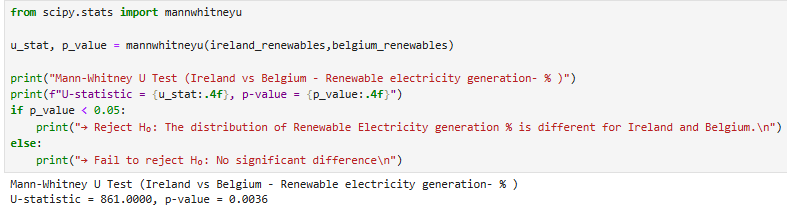
Test 2: Mann-Whitney U Test

Used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.

Hypotheses:

H0: The distribution of Renewable Electricity % is the same for Ireland and Belgium.

H1: The distribution is different.



The Mann Whitney test also gives p value which is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value is less than 0.05

→ Reject H₀: The distribution of Renewable Electricity generation % is different for Ireland and Belgium.

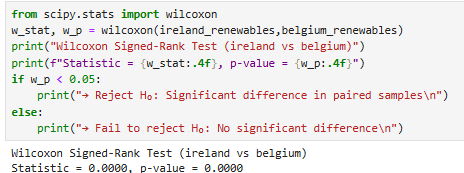
Test 2: Wilcoxon Signed-Rank Test

Used to compare two related samples, matched samples, or to conduct a paired difference test of repeated measurements on a single sample to assess whether their population mean ranks differ.

Hypotheses:

H0: Over the years 1990 to 2024, median difference between ireland and belgium electricity generation = 0 i.e. no significant difference

H1: Median difference ≠ 0 i.e. Significant difference in paired samples



The Wilcoxon test also gives p value which is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value is less than 0.05

→ Reject H₀: Significant difference in paired samples

Test 4: Chi-Square Test

Used to determine if there is a significant association between two categorical variables. It is commonly used when we have data in the form of counts or frequencies.

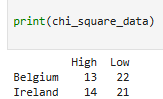
Steps for chi-square test:

1.Calculate Expected Frequencies

2. Calculate the Chi-Square Statistic

3.Compare with the Critical Value: After calculating the Chi-Square statistic, we compare it with the critical value from the Chi-Square distribution (Ahmed,2025) table for the desired significance level (e.g., 0.05) and degree of freedom. If the calculated Chi-Square statistic is greater than the critical value, we reject the null hypothesis.

My data is numerical so I need to make it categorical by setting a threshold of mean electricity production i.e. Renewable electricity above mean (High) vs below mean (Low). In my jupyter file, I have calculated the mean of each country, then classified its data into high and low based on country-specific means (as threshold). After that, I have created a contingency table.



Following this, I have applied chi-square test as follows:

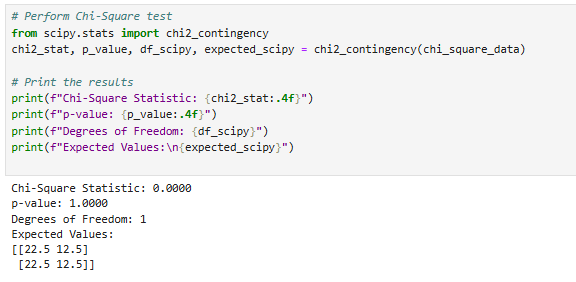
Hypotheses:

Null hypothesis (H₀):

There is no association between the country (Ireland or Belgium) and the classification of renewable generation ("High" or "Low").

Alternative hypothesis (H₁):

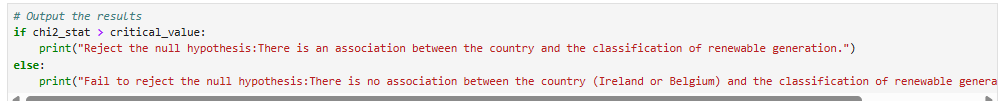
There is an association between the country and the classification of renewable generation.



Here the Chi-Square statistic, p-value, degree of freedom, and expected frequencies have been computed using scipy's chi2\_contingency. In the next step critical value is calculated using: critical\_value = chi2.ppf(1 - alpha, df\_scipy)

And comes out to be 3.841458820694124

In the last step, the calculated Chi-Square statistic is compared with the critical value and if it is found greater than the critical value, we reject the null hypothesis.



From the above results, we deduce that:

Chi-Square statistic < critical value

Fail to reject the null hypothesis:There is no association between the country (Ireland or Belgium) and the classification of renewable generation (High or Low).

→ i.e., the proportions of "High"/"Low" are similar in both countries.

Test 5: ANOVA (Analysis of variance) Test

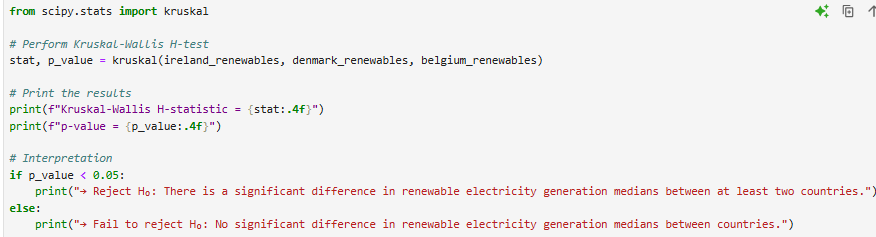
Used to compare the means of two or more groups by analyzing variance. ANOVA tells us whether there is a statistically significant difference between group means, but it does not specify which groups are significantly different from one another. However, it is applied to normally distributed data.

Since Ireland, Denmark and Belgium data is not normally distributed therefore, we can apply the Kruskal–Wallis test. It is a statistical test used to compare two or more groups for a continuous or discrete variable. It is a non-parametric test, meaning that it assumes no particular distribution of the data and is analogous to the one-way analysis of variance.

Hypotheses:

H0: The median renewable electricity generation is the same for Ireland, Denmark, and Belgium.

H1: At least one country differs in median electricity generation.



The Kruskal test also gives p value which is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value is less than 0.05

→ Reject H₀: There is a significant difference in renewable electricity generation medians between at least two countries.

*Variable 2: ''Electricity demand - Demand - Demand (TWh)'*

Test1: Shapiro-Wilk Test (Normality check)

Used to test if a single sample is normally distributed.

* Assumptions:
  + Data is continuous
  + Observations are independent

In my work, I have applied the shapiro test to the following countries: Ireland, Belgium, Austria and Sweden.

Hypotheses: H0 is null hypotheses and H1 is alternate hypotheses

H0: Electricity Demand Data is normally distributed.

H1: Electricity Demand Data is not normally distributed.



The shapiro test gives p value. The p-value is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value of Ireland is less than 0.05

Therefore, Reject the null hypothesis. Ireland electricity demand data is not normally distributed

p-value of Belgium is less than 0.05

Therefore, Reject the null hypothesis. Belgium electricity demand data is not normally distributed

p-value of Austria is not less than 0.05

Therefore, accept the null hypothesis. Austria electricity demand data is normally distributed

p-value of Sweden is not less than 0.05

Therefore, accept the null hypothesis. Sweden electricity demand data is normally distributed

Insights: From the above test, we can see that Ireland , Belgium and Austria energy demand data is not normally distributed. Therefore, we cannot apply parametric tests and need to apply non-parametric tests on their data. Following this, I have applied Mann-Whitney U Test, Wilcoxon Signed-Rank Test and Chi-Square Test on Ireland and Belgium data for electricity demand (TWh).

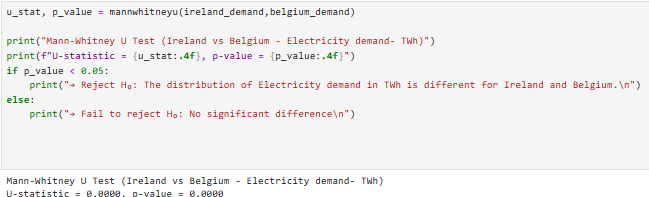
Test 2: Mann-Whitney U Test

Used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.

Hypotheses:

H0: The distribution of Electricity demand in TWh is the same for Ireland and Belgium.

H1: The distribution is different.



The Mann Whitney test also gives p value which is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value is less than 0.05

→ Reject H₀: The distribution of Electricity demand in TWh is different for Ireland and Belgium.

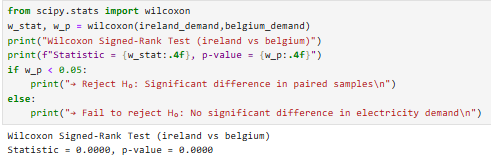
Test 2: Wilcoxon Signed-Rank Test

Used to compare two related samples, matched samples, or to conduct a paired difference test of repeated measurements on a single sample to assess whether their population mean ranks differ.

Hypotheses:

H0: Over the years 1990 to 2024, median difference between ireland and belgium electricity demand = 0

H1: Median difference ≠ 0



The Wilcoxon test also gives p value which is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value is less than 0.05

→ Reject H₀: Significant difference in paired samples

i.e. Over the years 1990 to 2024, there is significant difference between ireland and belgium electricity demand

Test 4: Chi-Square Test

Used to determine if there is a significant association between two categorical variables. It is commonly used when we have data in the form of counts or frequencies.

Steps for chi-square test:

1.Calculate Expected Frequencies

2. Calculate the Chi-Square Statistic

3.Compare with the Critical Value: After calculating the Chi-Square statistic, we compare it with the critical value from the Chi-Square distribution table for the desired significance level (e.g., 0.05) and degrees of freedom. If the calculated Chi-Square statistic is greater than the critical value, we reject the null hypothesis.

My data is numerical so I need to make it categorical by setting a threshold of mean electricity production i.e. Renewable electricity above mean (High) vs below mean (Low). In my jupyter file, I have calculated the mean of each country, then classified its data into high and low based on country-specific means (as threshold). After that, I have created a contingency table.



From the above table, we can see that it is clearly imbalanced. The two cells are zero so chi-square is invalid due to violated assumptions. Therefore, we can use Fisher’s Exact Test as it can handle this correctly, even with zeroes.

Fisher’s Exact Test is a statistical significance test used to determine if there is a non-random association between two categorical variables — especially when your data is in a 2x2 contingency table and some cells have low or zero counts.It’s called “exact” because it does not rely on approximations (like the chi-square test does), instead, it calculates the exact probability of the observed results happening by chance.

It is commonly used instead of the Chi-Square test when sample sizes are small or unbalanced.It calculates the exact probability (p-value) of getting a table this extreme (or more extreme) just by chance — under the null hypothesis that the row and column variables are independent.

In our context:

Null Hypothesis:

Country and demand level are not associated.

(The distribution of High/Low is the same in both countries.)

Alternative Hypothesis:

Country and demand level are associated.

(One country has more Highs or more Lows than the other.)



Here p-value < 0.05 therefore reject the null hypothesis.

→ Reject H₀: Country and demand level are associated i.e. One country has more Highs or more Lows than the other

Test 5: ANOVA (Analysis of variance) Test

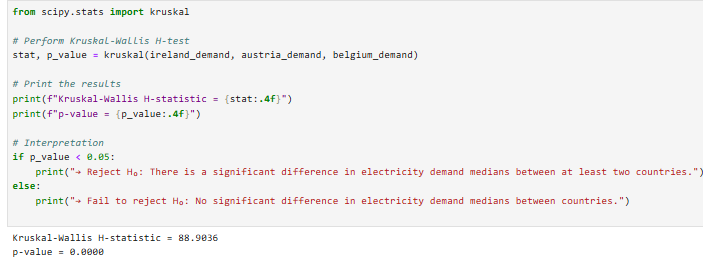
Used to compare the means of two or more groups by analyzing variance. ANOVA tells us whether there is a statistically significant difference between group means, but it doesn't specify which groups are significantly different from one another. However, it is applied to normally distributed data.

Since Ireland, Austria and Belgium data is not normally distributed therefore, we can apply the Kruskal–Wallis test. It is a statistical test used to compare two or more groups for a continuous or discrete variable. It is a non-parametric test, meaning that

Hypotheses:

HO: The median electricity demand is the same for Ireland, Austria, and Belgium.

HI: At least one country differs in median electricity demand



The Kruskal test also gives p value which is compared to a significance level (denoted by α, often set to 0.05).

If p < 0.05 then reject null hypotheses else, accept null hypotheses.

From the above results, we deduce that:

p-value is less than 0.05

→ Reject H₀: There is a significant difference in electricity demand medians between at least two countries.

**Question 4: Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process**

To deepen my research, I conducted two complementary streams of statistical analysis using energy-related datasets:

1. One focusing on fuel consumption patterns in Ireland from 1990 to 2022 across different sectors and fuel types.
2. Second on country-level renewable electricity generation and electricity demand, particularly comparing Ireland with other European countries, over the years 1990-2024.

In the first part, I applied descriptive statistics to a dataset covering multiple fuel types and sectors. The mean fuel consumption across all observations was approximately 45.377 ktoe, but the data revealed substantial differences between sectors i.e some sectors like transport and residential consume more of certain fuels. Through finding inter-quartile range and plotting box-plots, I found that the data has significant outliers. Moreover, after plotting the distribution, it was seen that data is not normally distributed but is heavily positively skewed. This indicated heavy sparsity in the dataset and suggested that the majority of fuels/sectors had no consumption in certain years. Moreover, the outliers suggested that certain sectors like transport contribute to the bulk of consumption and dominate the total. The bar chart indicated that highest consumption was seen to be of final energy consumption. Since this was an aggregate, it was deduced to drop it for further analysis. Moreover, after the findings highlighted the presence of outliers, the need for robust methods like log transformations was deemed necessary, when modeling.

Following the descriptive analysis, I applied inferential techniques—specifically 95% confidence intervals—to overall fuel consumption as well as to each sector individually. CI was applied using t- distribution as the data was not normally distributed. It was estimated that the true average fuel consumption (in ktoe) across all sectors, fuel types, and years lies between 42.52 and 48.23 ktoe. However, CI per sector revealed significant variation in fuel consumption patterns. The Industry - transport equipment manufacture sector had the lowest mean consumption at 0.58 ktoe with a narrow interval of (0.46, 0.69), indicating low and consistent energy use. In contrast, individual sectors, like the Residential sector stood out with a high mean of 170.09 ktoe with a wide interval of (145.44, 194.73)indicating high-consuming area and greater variability. The road private car transport also showed significant use at 118.39 ktoe, highlighting energy-intensive areas within the national profile. I also modeled the frequency of high-consumption sectors using the Poisson distribution, helping to assess the probability of extreme usage years and highlighting sectors or periods that may require special policy attention. However, the analysis highlighted the importance of taking caution when applying models like Poisson, which assume high-consumption events occur independently at a constant rate. Real-world data, on the contrary, is impacted by seasonality, policy shocks, and economic cycles, often violating these assumptions, especially for fuel types like gas used for heating, which spikes in winter. Ignoring such factors could lead to misestimating the probability of extreme usage and misrepresenting the effectiveness of policy measures.

In the second part, I shifted to a country-comparative perspective using the ‘europe\_yearly\_full\_release\_long\_format’ dataset from Ember (Ember,2020) . I focussed on two key variables like electricity demand (TWh) and renewable electricity generation (%), comparing Ireland with Belgium, Austria, and Sweden. The initial step involved testing for normality using the Shapiro-Wilk test. The results showed that renewable energy data for Ireland and Belgium was not normally distributed, guiding my decision to apply non-parametric methods (Mann-Whitney U, Wilcoxon Signed-Rank, Chi-Square test) rather than parametric ones (like the t-test). Mann-Whitney U Test and Wilcoxon Signed-Rank Test both indicated a statistically significant difference in the distribution and median values of renewable electricity generation between Ireland and Belgium. However, a Chi-Square test, after binning values into “High” and “Low,” found no significant association—suggesting that although the distributions differ, the long-term proportions of high vs. low renewable output years are similar i.e.there is no significant difference in proportions over time.  
This hinted that sector or fuel type may not be the only explanatory variable — time itself could be a major factor.The fact that both countries may experience a similar frequency of successful or unsuccessful renewable years, despite potential differences in their renewable strategies, prompted a shift in the research focus towards exploring the temporal dynamics (i.e. how things change over time means analyzing patterns, trends, or shifts in energy consumption or production) of energy trends, as well as the impact of policy changes and economic factors over segmented time periods, such as pre- and post-2010. Such segmentation can help assess whether significant shifts, like the EU Renewable Directive in 2009 or post-2008 financial crisis changes, accelerated renewable adoption or influenced fuel use patterns across particular sectors.

**Challenges Faced:**

Throughout the process, I encountered several challenges:

Challenges in the first dataset:

* The dataset used in the first part had hierarchical column headers that needed some restructuring before any statistical analysis could be done.
* On plotting the original fuel consumption VALUE (ktoe), the distribution was found to be highly positively skewed. Over 75% of fuel consumption values were zero, distorting the distribution and limiting the validity of parametric tests. On the other, there were extreme outliers as well. For this reason, log transformation had to be applied for better analysis.
* Another problem was the structure of the dataset, where multiple rows existed for the same year due to different sectors and fuel types being recorded separately. This granular level of detail made it difficult to directly visualize yearly trends using line plots, as each year had numerous entries. To overcome this, I had to aggregate the data by summing fuel consumption values across all categories for each year. This step was essential to transform the dataset into a suitable format for time series visualization (McQuaid,2025) and analyzing trends in total annual energy consumption.
* A key challenge in the analysis was dealing with numerous unique sectors and fuel types, including the "final energy consumption" sector, which is an aggregate category. This made it difficult to identify which specific sectors or fuel types to focus on. For example, the transport sector included many subcategories, such as transport-navigation, so I opted to choose a sector representing the aggregate/sum.
* Choosing the right model/distribution for the confidence interval was another challenge. Considering the non-normality of data, t-distribution was chosen.

Challenges in the second dataset:

* The dataset used in the second part also had hierarchical column headers that needed some restructuring before any statistical analysis could be done.
* Pre-processing was intensive as data cleaning and preparation was required that included removing null and duplicate values along with unnecessary columns and choosing only selective categories and subcategories with relevant variables and units.
* Not all countries passed the Shapiro-Wilk normality test, which made choosing between parametric and non-parametric approaches more nuanced.
* Since the Chi-Square test required categorical inputs, an appropriate threshold was required to create meaningful high/low categories based on each country's mean, which demanded justification and precision - this subjective threshold introduced interpretive complexity.
* Some tests (like Mann-Whitney and Wilcoxon) indicated significant differences among countries like Ireland and Belgium’s renewable electricity generation data, while others (like Chi-Square) did not. Interpreting and integrating these results meaningfully into a narrative required critical thinking.

Conclusion:

Overall, this analytical journey not only deepened my understanding of the energy consumption trends across various sectors and fuel types over the years in Ireland but also helped me gather insights about energy generation and demand in different countries in contrast to Ireland. It highlighted the importance of choosing the right test based on data distribution and clearly defining assumptions and thresholds and allowed me to analyze how different statistical techniques can highlight diverse aspects of the same energy system. Despite the challenges in data quality, test selection, and result interpretation, this process laid a strong foundation for developing more granular analysis, policy simulations, or predictive models in future energy research that can be used by governments in policy making and taking more green initiatives.

This section will focus on the ML part.

## Section3: Machine Learning for Data Analytics

The combination of these paradigms enabled a flexible, modular, and efficient approach to solving the complex challenges of analyzing and modeling my dataset.

Q1. Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

Q2. Collect and develop a dataset based on the energy topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate energy topic for producers and consumers point of view in Ireland.

Q3.You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

Q4. A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

**Question1:**

**Technique Selection:**

For my dataset on Fuel consumption across different sectors and fuel types, the appropriate machine learning technique depends on my Research Questions defined earlier. I have used Supervised learning approaches like Decision Trees and Random Forest (Iqbal,2025) in my project.

E.g. Question: Based on historical data, can we predict the future fuel consumption? Relevant Techniques: Regression models (e.g., Linear regression, Random Forest regression).

Below is an explanation of which type of learning technique (supervised, unsupervised, or semi-supervised) fits best for my dataset and why:

1.Supervised Learning:

Supervised learning involves training a model on labeled data, where the target variable is known. This is appropriate for regression or classification tasks where you have specific predictions to make

E.g. Question: Based on historical data, can we predict the future consumption of fuel?  
Relevant Techniques: Regression models (e.g., linear regression, Random Forest regression).

Why Supervised Learning is Appropriate for my dataset:

Prediction Task: Since my dataset includes target variables like fuel consumption value in ktoe, it is appropriate to use supervised learning to predict future consumption, identify trends, or forecast other related variables based on historical data.

Regression Model: For continuous values like fuel consumption values, regression models such as Linear Regression, Random Forest Regression, can be used to predict future consumption.

Classification Task: For grouping or classifying data into predefined categories (e.g. sector wise fuel consumption), classification algorithms like Logistic Regression, K-Nearest Neighbors (KNN), etc. can be applied.

Example Use Case: Predicting the fuel consumption value based on the sector and fuel type can be framed as a supervised learning problem where I can train a model to learn the relationships between input features (e.g., sector, fuel type) and the target variable (fuel consumption).

2.Unsupervised Learning:

Unsupervised learning is used when there is no labeled data and the goal is to uncover hidden patterns or groupings within the data. It is ideal for tasks like clustering and dimensionality reduction.

Why Unsupervised Learning Could Be Used:

Segmentation/Clustering: Unsupervised learning techniques like K-Means Clustering can be applied to segment the fuel consumption into groups based on the sector or fuel type, which would be useful for analyzing patterns, trends, market segmentation and policy making.

Example Use Case: Using K-Means Clustering to segment fuel consumption based on the sectors, helping to identify which sectors tend to consume the most fuel across all or in different categories of fuel types.

3.Semi-Supervised Learning:

Semi-supervised learning is a hybrid approach where a small amount of labeled data is combined with a large amount of unlabeled data. It is useful when labeling the entire dataset is costly or time-consuming.

Why Semi-Supervised Learning is Less Suitable Here:  
In my case, since I already have a dataset with labeled data (VALUE, sector, etc.), supervised learning is more suitable for most tasks. Semi-supervised learning is typically more beneficial when labeling data is difficult, which is not the case here.

*Machine Learning Models:*

Supervised Learning Approaches: Decision Trees and Random Forest

In my project, I have used Decision Trees and Random Forest for modelling. This is because I want to predict fuel consumption value (a continuous variable), making it a regression problem. Based on my dataset and goals, I used:

1. Decision Tree Regressor

Reason: Easy to interpret, handles categorical variables well, does not require feature scaling, captures non-linear relationships.  
Use Case for my dataset: Understanding how Sector, Fuel Type, and Year affect fuel consumption value.

2. Random Forest Regressor

Reason: An ensemble of Decision Trees, it improves accuracy by reducing overfitting and variance.  
Use Case for my dataset: More robust predictions; can also give the feature importance.

Justification: These are both relevant because my data is mixed (categorical + numerical). Since I am not working with an extremely huge dataset, both are efficient. Moreover, my objective is to gain insight as well as accuracy, and Decision Trees give interpretability while Random Forest gives performance.

*Steps for Modelling:*

1. Define X and y after cleaning and preparing phase

X = df[['Sector', 'Fuel Type', 'Year']] #features

y = df['VALUE\_log'] #log-transformed target

Since the fuel consumption value data is highly positively skewed therefore, to get better results, we are taking the log of the VALUE column. Moreover, I have removed some aggregate categories from fuel type and sector because they were duplicating the count.

2. Train/Test split the data

3. Model Building & Hyperparameter Tuning:

Define hyperparameter grids:

param\_grids = {

'Decision Tree': {

'max\_depth': [10, 20, 30 ],

'min\_samples\_split': [2, 5, 10, 20],

'min\_samples\_leaf': [1, 2, 4, 8],

'max\_features': ['sqrt', 'log2']

},

'Random Forest': {

'n\_estimators': [100, 300, 500],

'max\_depth': [5, 10, 20, None],

'max\_features': ['sqrt', 'log2']

}

}

Chosen parameters:

● max\_depth controls the depth of each tree. Using None means nodes are expanded until all leaves are pure.

● min\_samples\_split controls the minimum number of samples required to split an internal node. Higher values prevent overfitting. Min\_samples\_leaf controls the minimum number of samples that are required to be at a leaf node.

● n\_estimators is the number of trees in the forest (for Random Forest only).

● max\_features controls the number of features to consider when looking for the best split:

○ 'sqrt' is recommended for classification

○ 'log2' is another common choice

Next we initialize the models. After this, hyperparameter tuning is done using GridSearchCV. It is a model tuning tool from sklearn.model\_selection that automates the process of trying different combinations of hyperparameters to find the best-performing one based on a chosen scoring metric.

grid\_search = GridSearchCV(estimator=model, param\_grid=param\_grids[name], cv=5, n\_jobs=-1, verbose=1, scoring='r2')

Explanation:

1. estimator=model

● This is the machine learning model we are tuning e.g., a Decision Tree or Random Forest. The model here is selected from the models dictionary.

2. param\_grid=param\_grids[name]

● This is the dictionary of hyperparameters we are testing for that model. param\_grids contains possible values for model parameters like max\_depth, min\_samples\_split, etc. It will test all combinations of these values.

3. cv=5

● This means we are using 5-fold cross-validation. This implies that the training data is split into 5 parts. The model trains on 4 parts and validates on the 5th. This rotates 5 times so that every part is used for validation once.The reason for doing cross-validation is that it reduces overfitting and gives a better estimate of how the model performs on unseen data.

4. n\_jobs=-1

● Tells the computer to use all available CPU cores to speed up the grid search. -1 means “use everything you have got”. This helps to save time, especially when testing many hyperparameter combinations.

5. verbose=1

● Controls how much output is shown during the search. 1 means we will see progress updates printed to the console (like “Fitting 5 folds for each of 12 candidates…”).

6. scoring='r2'

● This tells GridSearchCV to evaluate each model based on the R² Score (coefficient of determination). It will choose the combination of parameters that gives the highest R² on average across the 5 folds.  
The reason behind using this score is that we are doing regression (predicting a numerical value like fuel consumption), and R² tells us how well the model explains the variance in the data.

After using Grid Search with Cross-Validation, we train the Best Model. Then use this model to predict fuel consumption value based on the testing features.

We evaluates the predictions using:

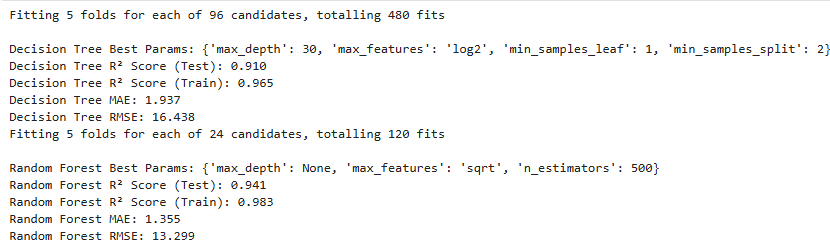
● R² Score: How well the model explains the variance.

● MAE (Mean Absolute Error): Average magnitude of errors.

● RMSE (Root Mean Squared Error): Penalizes large errors more.

One important point to highlight here is that after getting the best parameters, we invert the log transformation to get predictions in original VALUE scale

The results are as follows:



Insights: Higher R² on test data for Random Forest (0.941 vs 0.910) along with lower MAE and RMSE shows that it outperforms Decision Tree and gives better generalization and explanatory power.

**Question2:**

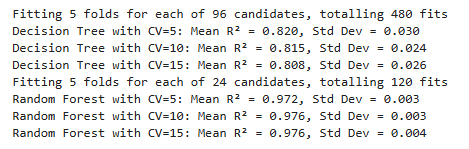
**Cross validation:**

Next we tune our model using GridSearchCV with 5-fold cross-validation to identify the best hyperparameters. This part *uses* cross-validation internally for hyperparameter tuning, but the performance scores we calculate are from single train/test splits, not CV averages.

The optimized model is then evaluated using three different cross-validation values on the training data to assess its generalizability. K-Fold is used to split the training data into five,ten and 15 shuffled folds for cross-validation (cv=5,10,15). Shuffling ensures varied data splits, while setting a random state guarantees reproducibility of results.

This two-step process ensures both model optimization and robust performance validation, with results reported as mean and standard deviation of R² scores across folds — a direct measure of the model’s generalizability and consistency.

Checking with cv=5,10 and 15:



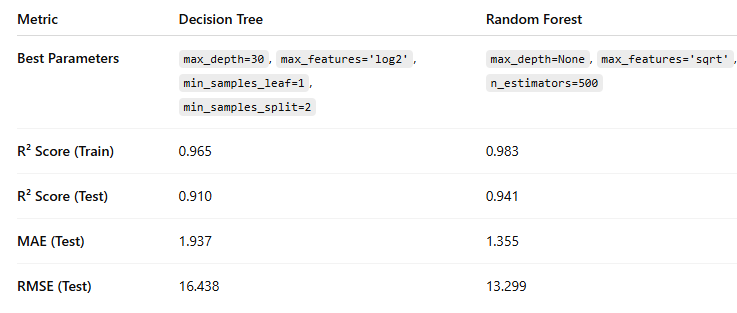
Insights: Across all CV splits, Random Forest shows remarkably high and stable R² scores (≥ 0.972), while Decision Tree performance is lower and less stable.

Next step will be the comparison of models.

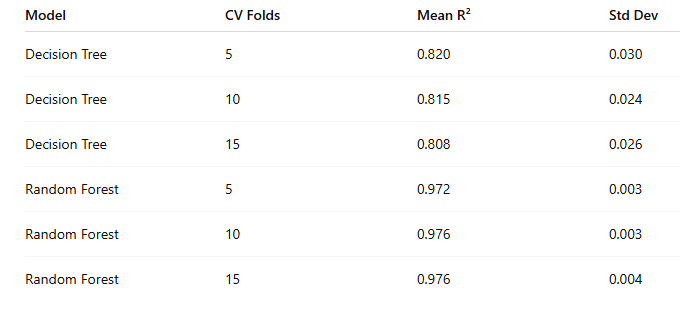
**Question4:**

**Comparison:**

The following table summarizes the results of both models after hyperparameter tuning:



The table below shows the results for R2 mean and standard deviation for different values of cv (5,10,15):



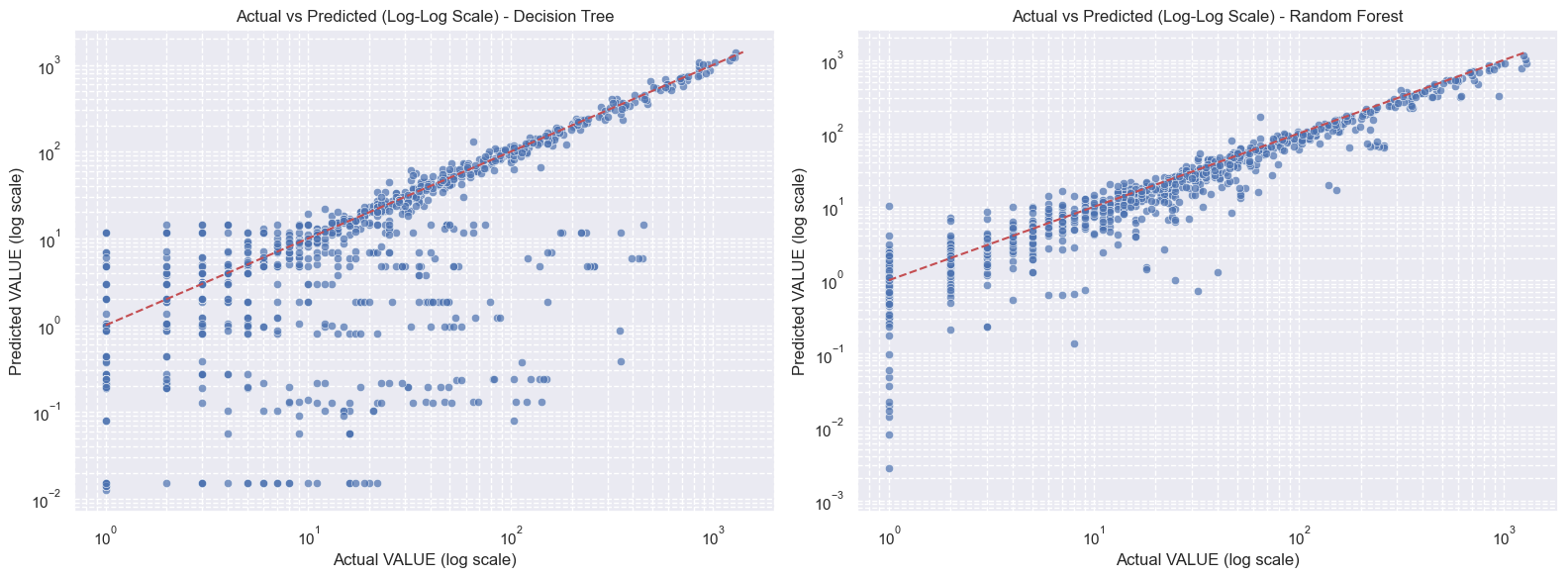
Key Insights:

● Random Forest outperforms Decision Tree overall in terms of better R² Scores as well as MAE and RMSE. This means the model explains a higher proportion of the variance in the target variable along with suggesting more accurate predictions.

● Random Forest also benefits from ensemble learning, which reduces overfitting and leads to more stable performance as seen in the minimal R² gap for Random Forest (0.983 vs 0.941). However, in the Decision Tree, the gap between train (0.965) and test (0.910) R² indicates mild overfitting, even after hyperparameter tuning. Since Decision Trees are prone to overfitting, especially with deep trees (max\_depth=30), this is expected.

* Across all CV splits, Random Forest shows better mean R2 score and standard deviation implying high reliability and low variance in performance. In contrast, Decision tree exhibits lower and more variable performance and is more affected by the increase in CV folds; its performance slightly drops showing that simpler models are more sensitive to changes in training set sizes and folds.
* Overall, minimal gain between CV=10 and CV=15 suggests CV=10 is sufficient for validation in future runs (saves compute time).

The following graphs also show the comparison for both the models based on actual and predicted plot:



Graphical analysis: Since dots are close to the red line so both models are performing well (predictions are close to actual values) however, as seen from the graph, dots are closer for Random Forest indicating its better performance.

Lastly, we will see the similarities and differences.

*Similarities and differences*:

Similarity Between the Models:

Both Decision Tree and Random Forest are tree-based models:

● They can handle categorical and numerical data effectively.

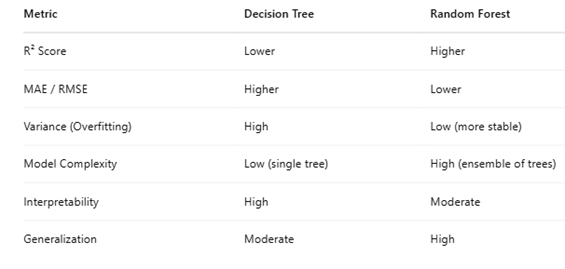
● Both are capable of capturing nonlinear patterns in the data.

● Neither requires feature scaling or normalization.

● Both models produce interpretable results such as feature importance.

During evaluation, both models achieved relatively high R² scores and low error metrics, indicating they are well-suited to the regression task.

Difference Between the Models:



Random Forest combines predictions from multiple decision trees, reducing the risk of overfitting and resulting in better generalization.

Decision Tree, while simpler and more interpretable, tends to overfit the data, leading to less accurate predictions on unseen data.

Overall, Random Forest achieved higher accuracy (e.g., R² score) and lower prediction error, indicating better reliability.

*Relevance:*

The Random Forest model demonstrates better predictive accuracy compared to the Decision Tree, making it more suitable for forecasting tasks where generalization and stability are essential. This is particularly relevant in real-world applications like in my project for predicting fuel consumption across various time periods or sectors, where the ability to make accurate predictions on new, unseen data is crucial.

On the other hand, the Decision Tree model excels in interpretability, which is beneficial when the goal is to understand how specific variables such as sector or fuel type influence the fuel consumption values. Its visual structure allows policymakers and stakeholders to easily grasp decision-making patterns. Additionally, Decision Trees are faster to train, making them practical when computational efficiency is prioritized.

Another reason is that my fuel consumption values data is positively skewed and I applied log transformation over it in order to get better results. While log transformation is commonly used to reduce skewness for linear models (like linear regression), Decision Trees and Random Forests are not sensitive to skewed distributions in the same way.

*Effectiveness:*

Random Forest Regressor proves more effective due to its ensemble nature, which leads to better predictive accuracy and robustness.  
For a task like predicting fuel consumption, where complex interactions may exist, Random Forest's ability to generalize well makes it the most suitable model for practical applications like forecasting, Renewable Energy initiatives and Policy decision support.  
Decision Tree Regressor, while less accurate, offers transparency and can help explain decisions e.g., identifying which year, sector or types of fuel are most influential and contribute more to Ireland’s fuel consumption. This makes it useful for exploratory analysis or stakeholder communication.

*Practical implications:*

The improved performance of Random Forest suggests that ensemble methods should be preferred for forecasting metrics like fuel consumption, where accuracy and generalization are crucial. However, Decision Trees offer interpretability, which might be useful when explaining individual decisions or identifying key consumption drivers, may it be a particular sector or fuel type.  
The findings can support data-driven energy policy, such as targeting non-renewable fuel types’ consumption and devising strategies towards higher use of renewable energy sources along with taking more green energy initiatives in order to have a good impact on fuel consumption costs as well as CO2 emissions.

In the next part of the report ,we will discuss Sentimental analysis.

**Question 3:**

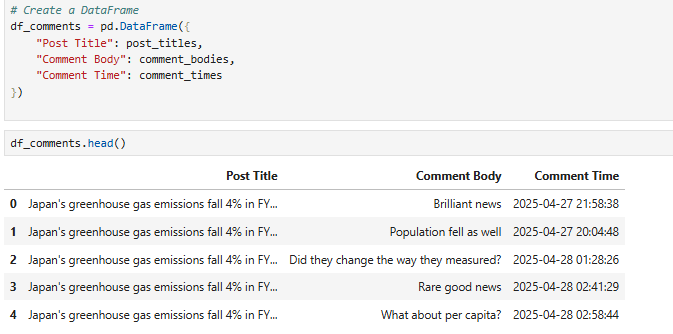
**Sentiment analysis:**

I have performed sentiment analysis on comments fetched from Reddit (Weiss,2025), particularly from subreddit ‘RenewableEnergy’. Below is the detailed description highlighting all the steps like data collection, cleaning, vader sentiment analysis and modeling.

Data Collection:

The data collection process involves creating a Reddit app to obtain API credentials (APP\_ID, APP\_NAME, APP\_SECRET, etc.), which are stored in a .env file using the dotenv package to keep the credentials safe. By using python (PyPI, 2025) dotenv, we can keep our settings in a separate file, making our code cleaner and easier to manage.

Then Python Reddit API Wrapper (PRAW), a Python module, is used to collect data from Reddit, by accessing the Reddit API. The subreddit ‘RenewableEnergy’ is accessed and the latest comments, post titles along with the comment times are fetched and stored in a CSV file (“Reddit\_Energy\_Comments.csv”) for further analysis.



Data preparation and cleaning:

In the next step, data is retrieved from the csv file. Different tools/libraries like NLTK , Porter Stemmer and RegEx, are used for cleaning the data.

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language. It supports classification, tokenization, stemming, parsing, and semantic reasoning functionalities. NLTK also provides Porter Stemmer that simplifies words by reducing them to their root forms, a process known as "stemming."

Similarly, a RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. Python has a built-in package called re, which can be used to work with Regular Expressions.

In my data cleaning work, first, all non-alphanumeric characters (such as symbols and punctuation) are removed using regular expressions. The text is then converted to lowercase to ensure uniformity. Each comment is split into individual words, and common English stopwords (e.g., "you", "also", "is") as well as punctuation are filtered out to retain only meaningful terms. Next, stemming is applied using the Porter Stemmer to reduce words to their root forms (e.g., "running" becomes "run"), which helps consolidate similar terms and reduce dimensionality. Finally, the cleaned and processed words are rejoined into a single string for each comment, resulting in a cleaned dataset suitable for further sentiment analysis.

VADER sentiment analysis:

In the next step, VADER sentiment analysis is performed on the cleaned comments. VADER stands for Valence Aware Dictionary and Sentiment Reasoner. It is a tool used for sentiment analysis, which is basically a way to figure out if a piece of text is expressing positive, negative, or neutral emotions. The polarity\_scores() function is applied to calculate sentiment scores, which extracts the four key components: compound (overall sentiment), pos (positive sentiment proportion), neu (neutral sentiment proportion), and neg (negative sentiment proportion). These values are collected into four separate columns and stored in a new csv file (“Reddit\_Energy\_Comments\_VADER\_analysis.csv”).

Next comments are categorized on the basis of their compound score:

if compound >= 0.05:

return 'Positive'

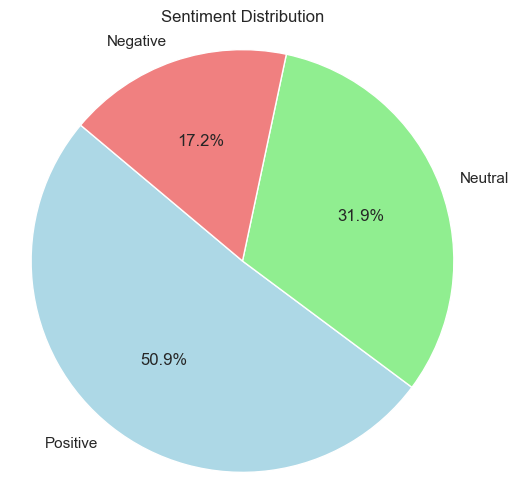
elif compound <= -0.05:

return 'Negative'

else:

return 'Neutral'

The pie below shows the sentiment distribution indicating a high percentage of positive sentiment.



Machine learning/modelling:

In the last part, text vectorization is done using CountVectorizer, which converts the cleaned textual comments into a bag-of-words matrix—a common format for machine learning models. CountVectorizer is a feature extraction technique in Natural Language Processing (NLP) that converts a collection of text documents into a matrix of token counts.

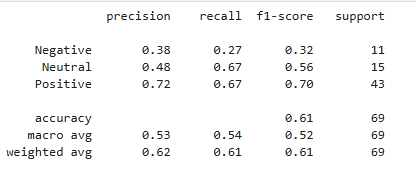
The ngram range=(1,1) specifies that the vectorizer is using unigrams only (i.e., single words). As a result, text\_counts i.e. a sparse matrix is generated in which each row represents a comment and each column a word (feature), with values indicating how often each word appears in the comment.

Next, the dataset is split into training and testing sets, with 75% used for training and 25% for testing. A Multinomial Naive Bayes classifier is trained on the training data to learn patterns associated with different sentiment classes (positive, negative, neutral). Multinomial Naive Bayes (MnB) is used for building a machine learning model to classify comments into different sentiment categories.

Finally, the model's performance is evaluated using accuracy score, providing a quantitative measure of how well the model predicted sentiments on unseen data.

The accuracy is found to be 0.608 by choosing alpha (smoothing parameter) as 1.

Besides this, classification report is generated as follows:



Insights:

Positive sentiment dominates and performs the best among all. Most samples are positive indicating class imbalance. Neutral sentiment has been fairly handled with moderate support (15) and decent recall (0.67), the model captures most Neutral cases. However, Negative sentiment is poorly predicted. Low recall (0.27) indicates most Negative samples are missed which may be due to class imbalance or small sample size.

**Section 4: Programming**

Q2. Data From Diverse Sources: In a dedicated section of your report, compare, contrast, and select relevant libraries/techniques to process data from diverse sources

Q3. Optimisation: In a dedicated section of your report, you are required to document and evaluate an optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured that the code is making good use of your system’s resources (eg CPU, RAM, time etc). Note any trade-offs that you've made in these areas.

**Question2:**

Data From Diverse Sources: In a dedicated section of your report, compare, contrast, and select relevant libraries/techniques to process data from diverse sources.

**Introduction:**

My project is based on analysing Ireland's National energy consumption and comparing the Irish Energy sector with other countries worldwide. This analysis also includes forecasting, sentiment analysis and evidence-based recommendations for the sector.

I have primarily used data from two sources:

1. Structured tabular data from CSO (Central Statistics Office, 2025)website of Ireland and Ember (https://ember-energy.org/)
2. Unstructured text data in the form of comments from Reddit (RenewableEnergy subreddit). (Reddit,2005)

Source 1: CSO and Ember Tabular data

*Nature of data:*

The tabular data from CSO (Central Statistics Office, 2025) is used for conducting Programming, Statistics, Visualization and Machine Learning tasks like observing current trends and predicting future trends of fuel consumption across different sectors and due to different fuel types in Ireland. The data from Ember (Ember,2020) is used for undertaking research in order to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them.

The dataset is available in CSV format with numeric and categorical variables. e.g. VALUE, Fuel Type,etc. in dataset-1 and VALUE, Area, Category, etc. in dataset-2.

The advantage of this data format is that it is clean, complete and ready for EDA and ML and Statistical analysis.

*Libraries used:*

The table below shows the relevant libraries/techniques to process data in different stages:

| Step | Purpose | Library/Technique |
| --- | --- | --- |
| Reading data | To read/import structured data | pandas (pd.read\_csv) |
| EDA | To analyze, explore, and manipulate data | seaborn, matplotlib, pandas |
| Preprocessing | For log transformation  scaling, encoding  plotting, visualising data | Numpy    Sklearn (e.g. from sklearn.preprocessing import StandardScaler, )  pandas, seaborn, matplotlib |
| Machine learning (applying models, hyperparameter tuning, grid search, cross validation) | To split the dataset,  apply ML models and perform exhaustive hyperparameter tuning to optimize those models | Sklearn (e.g. from sklearn.model\_selection  , import train\_test\_split,  GridSearchCV,  from sklearn.ensemble import RandomForestRegressor  from sklearn.tree import DecisionTreeRegressor),  numpy, pandas, seaborn, matplotlib |
| Evaluation | To evaluate the accuracy, r2 score, etc.  implement cross-validation to assess model performance | Sklearn (e.g. from sklearn.metrics import mean\_squared\_error, r2\_score, mean\_absolute\_error,  from sklearn.metrics import confusion\_matrix, accuracy\_score, precision\_score, recall\_score, f1\_score,  from sklearn.model\_selection import KFold, cross\_val\_score) |
| Interactive dashboard | To develop an interactive dashboard to showcase the information/evidence gathered from Machine Learning Analysis. Streamlit, a python library, used to create custom web app | Joblib (e.g. import joblib, import sys,  import subprocess,)  Streamlit (e.g. import streamlit as st),  Pandas,numpy, matplotlib, seaborn ,  Example usage:  import subprocess  subprocess.run(["streamlit", "run", "app.py"])  subprocess.check\_call([sys.executable, "-m", "pip", "install", package])  required\_packages = ['streamlit', 'pandas', 'numpy', 'scikit-learn', 'matplotlib', 'seaborn', 'joblib']  for pkg in required\_packages:  install(pkg) |
| Statistical analysis - Descriptive statistics | To compute and visualise descriptive statistics like kurtosis, skew, etc.  To visualize distribution plots, QQplots and poisson distribution. | pandas,  scipy.stats,  seaborn, matplotlib, numpy |
| Statistical analysis - Inferential statistics | To apply inferential statistical hypothesis testing through shapiro, mann-whitney, wilcoxon, chi-square, t-tests, etc. | pandas, numpy,  Scipy.stats (e.g. from scipy.stats import shapiro,  from scipy.stats import mannwhitneyu,  from scipy.stats import chi2\_contingency,  from scipy.stats import norm) |
| Optimisation | To check usage of system resources by calculating Timing,  CPU and  Memory usage | itime,  import io,  memory-profiler, multiprocessing, threading, sklearn |

Libraries like pandas, sklearn, and matplotlib are highly efficient for structured numerical data since they have built-in support for EDA, visualization (McQuaid,2025), statistical analysis and machine learning workflows.

*Justification of Library/Technique Choice:*

Pandas: It is a python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

In my work, I have used it for different purposes like reading csv files, encoding data during data pre-processing, etc. It is good for handling CSV files and performing data manipulation.

e.g. pd.get\_dummies(X, columns=['Sector', 'Fuel Type'])

NumPy: It is an open source python library that offers a collection of high-level mathematical functions including support for multi-dimensional arrays, masked arrays and [matrices. In](http://matrices.in) my work, I have used it for numerical operations like finding mean for inferential statistics or finding log of values during pre-processing, etc.

e.g. df\_new['VALUE\_log'] = np.log1p(df\_new['VALUE(ktoe)'])

Matplotlib and Seaborn: Matplotlib is a library in Python that enables users to generate visualizations like histograms, scatter plots, etc. Seaborn is a visualization library that is built on top of Matplotlib. It provides data visualizations that are typically more aesthetic and statistically sophisticated.

In my work, I have used them throughout for visualizing trends and patterns in the data including EDA, pre-processing, statistical analysis stages; for analysing data distribution using histogram, etc.

e.g sns.histplot(df\_new['VALUE(ktoe)'], bins=50, kde=True, color='blue')

plt.hist(sample\_means, bins=30, edgecolor='black', color='skyblue')

Sci-kit Learn (Sklearn): It is an open source python library that provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, etc. In my work, I have used it in the modelling phase for applying machine learning models like random forest and decision tree, etc.

e.g. from sklearn.ensemble import RandomForestRegressor

from sklearn.tree import DecisionTreeRegressor

Scipy Stats: It is the SciPy sub-package. It is mainly used for probabilistic distributions and statistical [operations. In](http://operations.in) my work, I have used it for descriptive statistics like finding kurtosis, skew, etc. of the distribution.

e.g. from scipy.stats import kurtosis

kurt\_value = kurtosis(df\_new['VALUE(ktoe)'])

Moreover, I have also used it in inferential statistics for hypothesis testing

e.g. for shapiro test: from scipy.stats import shapiro

shapiro\_ireland = shapiro(ireland\_renewables)

Lastly, it has been used in the evaluation as well to find metrics like r2 score, etc.

e.g. from sklearn.metrics import mean\_squared\_error, r2\_score, mean\_absolute\_error,

Memory-profier: It is a python module for monitoring memory consumption of a process as well as line-by-line analysis of memory consumption for python [programs. In](http://programs.in) my work, I have used it in the optimization phase for optimising memory usage e.g from memory\_profiler import memory\_usage

Apart from these, following libraries have been used:

multiprocessing : to run multiple processes in parallel, allowing tasks to be executed concurrently using multiple CPU cores.

psutil : for checking memory and cpu usage

threading : to create and manage multiple threads within a single process, enabling concurrent execution.

Time: for logging execution time.

Source 2: Reddit Unstructured Data

*Nature of data:*

The unstructured data in the form of comments, from Reddit, is used for sentiment analysis to get producers and consumers point of view in Ireland.

The raw text comments from Reddit (Reddit,2005) API are user-generated, unstructured and subjective. This data format needs to be organised and structured, and stored in a csv file. Then the data can be fetched and cleaned for sentiment analysis.

It also requires authentication, API handling, and text preprocessing

*Libraries used:*

The table below shows the relevant libraries/techniques to process data in different stages:

| Step | Purpose | Library/Technique |
| --- | --- | --- |
| API access | To provide a python interface to the Reddit API for comment fetching.To store and retrieve the environment variables and to convert reddit timestamps. | praw,  Dotenv (e.g. from dotenv import load\_dotenv),  Datetime (e.g. from os import getenv, from datetime import datetime as dt) |
|  |  |  |
| Data storage and retrieval | To store the comments in a structured format in a csv file and to fetch them for further analysis | pandas (df.to\_csv and then pd.read\_csv) |
| Data cleaning | To perform regular expression operations, remove stop words, punctuation and do stemming to clean data | pandas, nltk,  (e.g. from nltk.corpus import stopwords,  import string  from nltk.stem import PorterStemmer), import re |
| Sentiment analysis | To import the VADER tool to calculate sentiment polarity scores for analysis. | vaderSentiment,  (e.g.  from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer) |
| Machine learning | To convert text data into numerical feature vectors. To split data, apply machine learning models and do text classification. | Sklearn (e.g. from sklearn.feature\_extraction.text import CountVectorizer,from sklearn.model\_selection import train\_test\_split, from sklearn.naive\_bayes import MultinomialNB) |
| Evaluation | To evaluate and plot model performance using metrics like accuracy, etc. | Sklearn (from sklearn import metrics), matplotlib |

Python Reddit API Wrapper (PRAW) is a Python module that can be used to collect data from Reddit. It does so by accessing the Reddit API and is well-suited for quick prototyping. I have not used the Reddit API via Http explicitly in my project. It offers more control, but is more complex and wasn't directly used. PRAW abstracts away the complexity of making HTTP requests manually and also takes care of the OAuth authentication for Reddit API access. Therefore, we do not need to manually handle token generation as we would with the requests library.

Moreover, Reddit API responds with nested JSON data (i.e., JSON objects within JSON objects), especially for things like posts, comments, etc. PRAW simplifies this by converting the raw nested JSON into Python objects like, Comment, etc., making it easier to access attributes (e.g., post.title, comment.body) without parsing raw JSON ourselves.

*Justification of Library/Technique Choice:*

* PRAW (Python Reddit API Wrapper): It is a python module that can be used to collect data from Reddit. In my work, I used it to fetch recent posts and comments from the RenewableEnergy subreddit.e.g. Import praw

reddit = praw.Reddit(

client\_id=getenv("APP\_ID"), # Right below 'personal use script'

client\_secret=getenv("APP\_SECRET"), # secret

user\_agent=f"pda-2023 u/{getenv('USERNAME')}", # app-name u/username

)

* Dotenv: python-dotenv reads key-value pairs from a .env file and can set them as environment variables. By using python dotenv, we can keep these settings in a separate file, making our code cleaner and easier to manage. In my work, I have used it to load the environment variables like APP\_NAME, APP\_ID, etc.

e.g.

!pip install python-dotenv

from dotenv import load\_dotenv

* Datetime: It is a python module to work with dates and times. It provides a variety of classes for representing and manipulating dates and [times. In](http://times.in) my work, I have used it to see the time of creation of comments.

e.g. from datetime import datetime as dt

for comment in list(post.comments):

print(comment.body)

print(comment.created\_utc)

print(dt.fromtimestamp(comment.created\_utc))

* Nltk: The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming (Weiss,2025) language. It supports classification, tokenization, stemming, parsing, and semantic reasoning functionalities.

In my work, I have used it during data cleaning for stemming and removing stopwords.

e.g. from nltk.corpus import stopwords,

from nltk.stem import PorterStemmer

During cleaning, I have used ‘re’ and ‘string’ as well. These are built in python libraries.

import re: It imports the Regular Expression module, used for pattern matching and text manipulation (e.g., re.sub).  
import string: It imports the string constants and utility functions (e.g., string.punctuation)

* vaderSentiment: VADER stands for Valence Aware Dictionary and Sentiment Reasoner. It's a tool used for sentiment analysis, which is basically a way to figure out if a piece of text is expressing positive, negative, or neutral emotions. In my work, I have used it for sentiment scoring on Reddit comments. It is specifically designed for social media text, making it ideal for analyzing the informal and often ambiguous language used in Reddit comments.  
  e.g.

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

sentiment = SentimentIntensityAnalyzer()

scores = sentiment.polarity\_scores(str(comment))

* Sklearn: It is an open source python library that provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, etc.
* CountVectorizer is a feature extraction technique in Natural Language Processing (NLP) that converts a collection of text documents into a matrix of token counts.
* Multinomial Naive Bayes (MnB) is used for building a machine learning model to classify comments into different sentiment categories.

In my work, I have used sklearn during pre-processing for countvectorizer, during machine learning for splitting a dataset and applying Multinomial Naive Bayes (MnB) and lastly during evaluation for finding metrics like accuracy.

E.g.from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.model\_selection import train\_test\_split,

from sklearn.naive\_bayes import MultinomialNB, MNB = MultinomialNB(alpha=0.01)

from sklearn import metrics  
accuracy\_score = metrics.accuracy\_score(predicted, Y\_test)

* Apart from above, pandas, numpy, matplotlib, etc. have been used.

Data Collection and Processing Summary:

The process involved creating a Reddit app to obtain API credentials (APP\_ID, APP\_NAME, APP\_SECRET, etc.), which were stored in a .env file using the dotenv package to keep the credentials safe. Then Python Reddit API Wrapper (PRAW), a Python module, was used to collect data from Reddit, by accessing the Reddit API. The subreddit ‘RenewableEnergy’ was accessed and the latest comments and post titles were fetched and stored in a CSV file for further analysis.

Then data was retrieved from the csv file. The collected data was then preprocessed by performing text-cleaning operations like stemming, tokenization, removal of stop words, etc. After this step, VADER sentiment analysis was applied to assign sentiment scores. Finally, a machine learning model called the Multinomial Bayes Classifier was used to predict sentiment based on textual content. This sentiment analysis has helped in understanding the perspectives of renewable energy producers and consumers, specifically regarding pricing and policy discussions in Ireland.

Comparison:

*Energy Dataset (CSO and Ember Data):*

I have used two datasets for my project. One dataset (from CSO) includes various features like fuel consumption by sector and fuel types from 1990 to 2022. The other dataset (from Ember) includes features like CO₂ emissions, electricity generation and demand and geographical indicators like countries and regions. The analysis aims to compare Ireland’s energy statistics with other global regions such as the EU, OECD, G20, and ASEAN from 1990 to 2024.

The dataset from the CSO and Ember (Ember,2020) websites was already well-structured and formatted, requiring minimal preprocessing. On dataset-1, I had to do some pre-processing and log transformation using relevant libraries initially. Then, I primarily used descriptive statistics like mean, median, kurtosis, etc. and plots like histogram, boxplot, bar plot, etc. to understand data distribution. After that I analysed my data through inferential statistical techniques like confidence intervals and finally applied machine learning models like Decision Trees and Random Forests for analyzing relationships between variables.

Similarly, for the dataset-2 also, I had to do some pre-processing and cleaning and selection of relevant categories and subcategories for analysis. Then I applied inferential hypothesis testing on my data including tests like shapiro wilk test, chi-square test, etc. to understand about energy demand and generation in different countries including Ireland.

Libraries like Pandas (Pandas Development Team, 2025) and Matplotlib were ideal for working with tabular data and generating visualizations (McQuaid,2025) to summarize key insights. A description of relevant libraries at every stage has already been mentioned in the table above.

*Reddit Sentiment Analysis:*

Reddit (Reddit,2005) data, on the other hand, required a different approach because of its unstructured nature. I used PRAW for extracting comments from ‘Renewable Energy’ subreddit, but the main challenge was cleaning and processing text data. After that, I applied sentiment analysis using VADER to get sentiment scores for the comments. Finally I applied a machine learning model called Multinomial Naive Bayes classifier to classify the comment into segment categories. The steps also involved NLP (Natural Language Processing) techniques like tokenization, stopword removal, and text vectorization.

Libraries like praw and sklearn for using MultinomialNB were ideal for fetching comments and analysing the sentiments. A description of relevant libraries at every stage has already been mentioned in the table above.

| Aspect | Reddit data | CSO data |
| --- | --- | --- |
| Data format | Text, nested JSON from API | Categorical and numerical, csv file containing structured data |
| Complexity | Medium to high | low |
| Tools | praw, dotenv, nltk, etc. | pandas ,seaborn, sklearn, etc. |
| Data processing | Removing stop words, punctuation, spaces, etc. , stemming to clean data, calculate sentiment polarity scores | Encoding, scaling, log transformation, etc. Inferential statistical hypothesis testing |
| ML models | Multinomial NB classifier | Decision trees, random forest |
| Output | Sentiment scores | Regression predictions. Interactive dashboard |
| Metrics | Accuracy score, etc. | Accuracy, precision, mean square error, r2 score, etc. |

Conclusion:

The integration of diverse data sources i.e. reading structured data from csv file via official statistics (CSO and Ember websites) and fetching unstructured data in the form of comments via user-generated content (Reddit (Reddit,2005)), provided a comprehensive view of energy trends, both from a quantitative (statistical) and qualitative (consumer sentiment) perspective. By applying a variety of python libraries and machine learning techniques, I was able to preprocess, analyze, and model data from these sources, uncovering valuable insights related to energy usage and sentiment trends.

**Question3:**

Optimisation: In a dedicated section of your report, you are required to document and evaluate an optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured that the code is making good use of your system’s resources (eg CPU, RAM, time etc). Note any trade-offs that you've made in these areas.

**Planning:**

I will be applying optimization in different areas:

1. Optimization (Weiss,2025) of memory usage by loading data in chunks

2. Optimization by data types

3. Optimization of time, memory usage and CPU usage while applying machine learning models on my dataset taken from csv file (GridSearchCV and Cross-Validation part).

Apart from this, I will apply cross validation with cv=5,10,15 also to check the optimized results.

Lastly, I will check the optimized execution time with different CPU processors used.

Justification for chosen strategy : The reason for applying optimization in machine learning parts specifically, is that these sections of code require exhaustive search and computation, thus taking a lot of time and memory that need to be optimised.

**Documentation:**

1. Optimization of memory usage by loading data in chunks:

First we install memory-profiler, a Python package, which is used for monitoring memory usage of a Python program, especially useful for identifying memory bottlenecks during model training or optimization. Then we check the memory usage when we load the entire dataset at once and compare it to the scenario when we load the dataset in chunks. The chunksize is kept to be 10000.

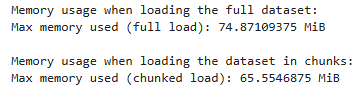
pip install memory-profiler

from memory\_profiler import memory\_usage

# Function to load the entire dataset

# Function to load the data in chunks and process it

Output:

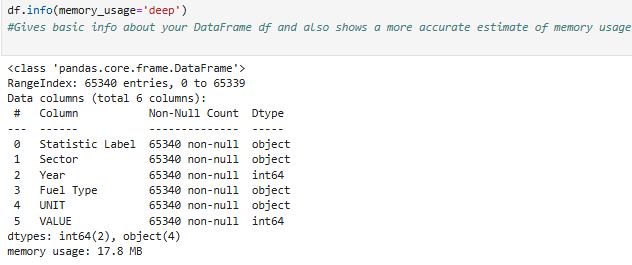


Insights: The above results show a clear decrease in memory usage when data is loaded in chunks as compared to loading the full dataset all together.

1. Optimization using data types:

Next, we look at memory usage by the data types and variables in our original dataset. Here [df.info](http://df.info)() with memory usage= deep gives information of memory usage as well along with information of data types.

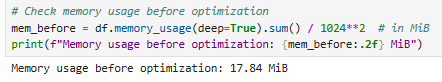
Original data types:



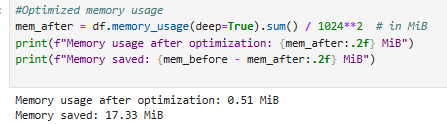
Here we can see that Statistic Label, Sector, Fuel Type and UNIT have object data type. Since object type takes more memory therefore we can down cast our data type to category, which requires less memory.

Similarly, Year and VALUE have data type int64. This again requires more memory so we can down cast our data type to int16, which requires less memory. We can do this because every year has four digits and it is an 11 bit number and int16 is sufficient to store an 11-bit number. Similarly, the VALUE column has a maximum value 13189, which comprises 5 digits and comes out to be a 14 bit number. Here, also, int16 is sufficient to store a 14-bit number.Thus we can achieve optimised results by down casting the data types.

The memory usage before optimizing the data types was:



The memory usage after optimizing the data types was:



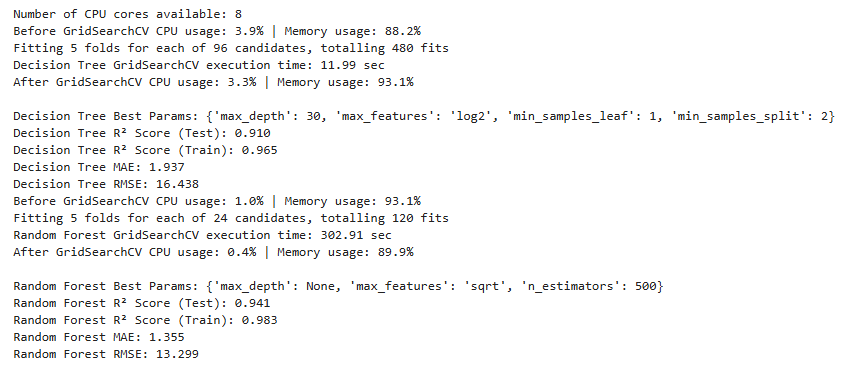
Insights: The above working shows clearly that as we downcast the data types, memory usage gets reduced significantly.

1. Optimization of time, memory usage and CPU usage while applying machine learning models:

Next, I will be applying optimization (Weiss,2025) to my dataset taken from the csv file (GridSearchCV and Cross-Validation part).

In this part, we perform optimization monitoring and model evaluation during the machine learning model training phase, specifically focusing on system resource usage (CPU, memory, and time). We use psutil , a cross-platform library for retrieving information on running processes and system utilization, in order to track CPU and memory usage before and after training. We also use time, the Python module that allows us to work with time in Python, in order to log execution time.

We perform GridSearchCV to tune hyperparameters for each model and see results.



The above results give following insights:

1. The optimization process using GridSearchCV is computationally efficient for the Decision Tree model, completing 480 fits in just 11.99 seconds with minimal CPU usage (~3.3–3.9%) and moderate memory increase (from 88.2% to 93.1%). In contrast, the Random Fores (Iqbal,2025)t model, despite performing only 120 fits, requires significantly more time (302.91 seconds) due to its computationally intensive nature, especially with 500 estimators, but interestingly shows lower CPU usage (0.4–1.0%) and memory fluctuation (from 93.1% to 89.9%). This highlights the trade-off between model complexity, resource efficiency, and execution time, where Random Forest offers superior performance metrics but at a higher computational cost.
2. Overall, Random Forest provides substantially higher R2 score and more stable results (lower mean square error), than Decision Tree, making it a better performing model.

Tradeoff: Overall, there is a trade off between performance and execution time, making Random Forest the better choice.

# *Cross validation with cv=5,10,15 to check the optimized results:*

In the next part, we monitor system resource usage (CPU and memory) and log execution time for different values of cv i.e. cv=5, 10 and 15. After running GridSearchCV, we check cross\_val\_scores with different values of cv. We use psutil to track CPU and memory usage.

In this process, we use following libraries:

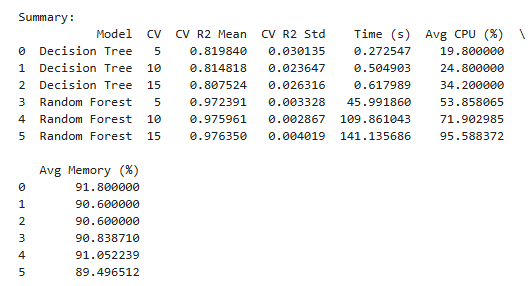
import time : for checking execution time

import multiprocessing : to run multiple processes in parallel, allowing tasks to be executed concurrently using multiple CPU cores.

import psutil : for checking memory and cpu usage

import threading : to create and manage multiple threads within a single process, enabling concurrent execution.

The results, including the R² mean and standard deviation from cross-validation, are shown below:



The above results give following insights:

1. As cross-validation (CV) folds increases from 5 to 15, both models require more time and CPU resources. For the Decision Tree, execution time grows from 0.27s to 0.62s, with CPU usage increasing from 19.8% to 34.2%, while memory stays stable (~90–92%), showing efficient scalability. In contrast, the Random Forest model shows significant increases in computational load i.e. time increases sharply from 46s (CV=5) to 141s (CV=15), and average CPU usage surges from 54% to 96%, reflecting its parallel complexity and larger model size.
2. Random Forest takes significantly longer time, due to its ensemble nature with many estimators. Decision Tree is much faster.
3. Despite these increases, R² values remained stable for Random Forest, with the Decision Tree slightly declining and the Random Forest consistently achieving high performance across CV folds, thus highlighting the trade-off between improved accuracy and resource demands at higher CV values.

.

Tradeoff: Therefore, we can say that the Decision Tree model is much faster and consumes fewer resources, making it ideal for quick benchmarking and exploratory testing. The Random Forest (Iqbal,2025) model, while significantly more resource-intensive, yields superior R² scores and low variance, making it preferable for final deployment or real-world forecasting.

*Time optimization with different CPU cores (n\_jobs):*

Next, we try to do time Optimization across cross validation value c=5 but with n\_jobs=4 and n\_jobs=-1:

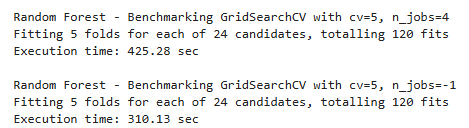
Case1:

When n\_jobs=4, it means we are utilising 4 of the available processors thus leading to higher execution time.

Case2:

When n\_jobs=-1, it means we are utilising only all available processors thus leading to lower execution time.

Tradeoff: Thus, there is a tradeoff between CPU usage (i.e. no of processors used) and execution time. Less no of processors lead to higher execution time.



**References**

[1]Central Statistics Office (2025) Central Statistics Office. Available at: <https://data.cso.ie/table/SEI06> (Accessed 17 May 2025).

Central Statistics Office (2025). SEI06 - Fuel Consumption (ktoe) [Dataset].

[2]Ember (2020) Ember Available at: https://ember-energy.org/data/yearly-electricity-data/ (Accessed 17 May 2025).

Ember (2025). Yearly electricity data - long format

[3]Pandas Development Team (2025) Pandas. Available at:<https://pandas.pydata.org/> (Accessed: 17 May 2025)

[4]Iqbal, M. (2025) “NLP and Topic Modelling” [Lecture], M4: Machine Learning for Data Analytics. CCT College Dublin. 16 April.

[5]Iqbal, M. (2025) “Classification using Decision Trees and RF” [Lecture], M4: Machine Learning for Data Analytics. CCT College Dublin. 26 February.

[6]Ali, M.M., et al. (2021). ‘Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison’, Computers in Biology and Medicine, 136, p.104672. <https://doi.org/10.1016/j.compbiomed.2021.104672>.

[7] Ahmed, M. (2025) “Uniform, Normal, Binomial, Poison, and Geometric distributions” [Lecture], Statistics for Data Analytics. CCT College Dublin. 13 March.

[8]McQuaid, D. (2025) “What is Exploratory Data Analysis?” [Lecture], M3: Data Preparation and Visualisation. CCT College Dublin. 26 February.

[9] McQuaid, D. (2025) “Principles of Data Visualization” [Lecture], M3: Data Preparation and Visualisation. CCT College Dublin. 12 March.

[10]McQuaid, D. (2025) “Encoding Categorical Data ” [Lecture], M3: Data Preparation and Visualisation. CCT College Dublin. 26 March.

[11]DataCamp, Inc. (2024) Datacamp. Available at: https://www.datacamp.com/blog/introduction-to-programming-paradigms (Accessed 6 April 2025).

[12]Scikit-learn developers (2025) Scikit-learn. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html> (Accessed: 16 May 2025)

[13] Snowflake Inc (2025) Streamlit. Available at: <https://streamlit.io/> (Accessed 17 May 2025)

[14] Reddit (2005) Reddit. Available at: <https://www.reddit.com/r/RenewableEnergy/>

(Accessed 7 May 2025).

[15]Weiss, S. (2025) “Databases, MySQL (and Mongo DB)”[Lecture], M1: Programming for Data Analytics. CCT College Dublin. 3 May.

[16] PyPI (2005) PyPI. Available at: <https://pypi.org/project/memory-profiler/>

(Accessed 9 May 2025).