**Energy Trade Analysis Report – Nachiket Khambhati**

**Introduction**

This is an Energy Trade Analysis data science project, whose dataset is available here: <https://www.kaggle.com/datasets/unitednations/international-energy-statistics/>.

***Domain Understanding***

I started off in the first week with the following research on the Domain Understanding, Data identification, and Data Collection aspects of the analysis.

Primary/Secondary Energy

Primary and secondary energy delineate the initial and refined stages of energy conversion, respectively. Primary energy sources are directly extracted from nature, including fossil fuels like coal and oil, renewable sources like sunlight, wind, and water, and nuclear fuels. These raw materials are foundational to energy production and must undergo conversion into more practical forms for consumption.

In contrast, secondary energy emerges from the transformation of primary sources into usable formats. Examples of secondary energy include electricity derived from coal, solar power, or wind turbines, as well as refined fuels such as gasoline and diesel. Secondary energy plays a pivotal role in powering homes, industries, and transportation systems, offering a more accessible and versatile energy form.

Renewable Energy Sources:

Renewable energy sources are sustainable, environmentally friendly alternatives that harness natural processes to generate power. Examples include solar energy, derived from sunlight through photovoltaic cells, wind energy produced by turbines capturing wind motion, and hydropower generated from flowing water. Biomass energy utilizes organic materials like wood or agricultural residues, while geothermal energy taps into the Earth's internal heat.

Conventional and Non-Conventional Energy Sources:

Conventional energy sources encompass traditional, widely used methods like fossil fuels (coal, oil, and natural gas) and nuclear power. Non-conventional sources, on the other hand, include renewable options alongside innovative technologies. While renewable sources fall under the non-conventional category, unconventional oil and gas extraction methods, such as shale gas and oil sands, also fit this classification.

**Methodology**

***Data Identification***

The data for the energy trade analysis was provided by the United Nations statistics division on the UNData site. Available at: <http://data.un.org/Explorer.aspx>

***Data Collection***

The dataset was downloaded as a csv file and added to my Energy Trade Analysis repository as all\_energy\_statistics.csv. I used pandas to create a dataframe from the csv file as follows:

A screenshot of a computer

Description automatically generated

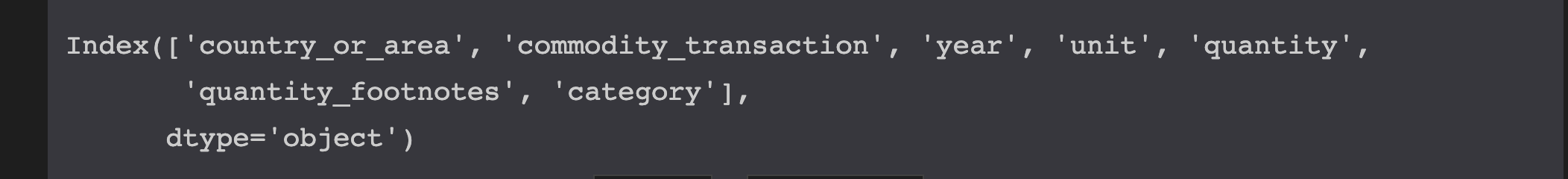
This would then facilitate in the next step which is to perform data cleaning and preprocessing.

***Data Exploration***

A screenshot of a graph

Description automatically generated

Based on this initial printout of the head of the dataframe obtained from the source csv, it appears that the relevant variables are:



The only ambiguous one here is the ‘quantity\_footnotes’ variable, which I don’t know what it means and has the value ‘nan’ for at least the elements in the dataframe head.

Then I looked at the data types for each column:

A screen shot of a computer

Description automatically generated

So we have types ‘object’, int64 and float64

I proceeded to find the unique values of each categorical variable i.e. country\_or\_area,

commodity\_transaction, unit, category and quantity\_footnotes. Upon doing this part of the exploration I also discovered that the quantity\_footnotes variable is always either nan or 1.0, seemed not relevant to the analysis.

***Data Cleaning***

I am removing the quantity\_footnotes column entirely as I don’t think it will be useful for the analysis. Additionally, I need to explore for further inconsistencies in the dataset. This will begin by checking for missing data. First we have to verify that all values are non-null. This was done as follows:

A screen shot of a computer screen

Description automatically generated

***Data Preprocessing and Feature Engineering***

Upon verifying this, we can continue preprocessing with categorical feature encoding, and feature engineering by numericizing quantity inputs through combining the unit and quantity variables. For now we are not differentiating between the export, import and other types of usage of energy, just going according to the energy category. I then created a dictionary that stores the unit associated with each energy type:

A screenshot of a computer program

Description automatically generated

**Results and Analysis**

***Part 1 – Total energy Production, Exchange, and Usage***

The first goal is to obtain the statistics according to each energy type. We have the dictionary storing the unit associated with each type, now we have to create a new dataframe that stores the total energy production, usage, and exchange of each energy type along with the unit. First created a dict called data, which stored the category of each energy type and the total PEU (production, exchange and usage) of it along with the unit as follows:

A screenshot of a computer program

Description automatically generated

Next, I created a dataframe with columns Category, Total PEU, and Unit using this dictionary as follows: (This will be called a PEU-style dataframe from here on)

A screenshot of a computer

Description automatically generated

This will have to be graphed and then analyzed, which will complete the first step of data analysis.

***Part 2 - Comparative Analysis:***

We will try to group the energy PEU dataframe according to the unique units, as that is the only way to compare across different energy types. This is done as follows:

A screenshot of a computer

Description automatically generated

Now that this unit-dataframe dictionary is obtained, we move on to grouping by countries. This will be done by creating a PEU dataframe for each country in the list as follows:

A screenshot of a graph

Description automatically generated

This gives us a dictionary with country key and PEU style dataframe value.

Same protocol was followed with countries as follows:

A screenshot of a computer

Description automatically generated

***Part 3 – Data visualization and analysis***

The first part is visualizing total energy production. Only meaningful way to do it on pie charts is grouping the ones with same unit together as follows:

A pie chart with different colored sections

Description automatically generated

This was done for all units.

Next part is different countries and different years. The dataframes are ready. They need to be plotted. Here is an example for additives and oxygenates over the years:

A graph with a line going up

Description automatically generated

The next part is visualizing the data for the top 10 energy producing nations for each energy type. I chose to use only the top 3 energy types as indicated from the pie charts of total energy consumption by energy category. These are coal, electricity and natural gas. The following is a sample of the code snippet for electricity, which was repeated for coal and natural gas. I wanted to obtain the top 10 electricity, coal and natural gas producing nations for the total consumption from 1990 to 2015:

A computer screen shot of a computer program

Description automatically generated

This was repeated for coal and natural gas as well, and the following plots were obtained:

A graph of a graph showing the amount of coal in the country

Description automatically generated with medium confidence

A graph of electricity usage

Description automatically generated with medium confidence

A graph of a number of gas users

Description automatically generated