Mining of Large Datasets Report

# Study on the energy supply of countries

*Students:*
Saad Lahlali
Haileleul Haile

December 10, 2021

# Contents

# 1 Introduction

Since the first climate conference in 1995 in Berlin, the UN organizes a COP each year to organize a worldwide response to global warming. During each COP, new leaders take many commitments to change their country's energy supply politics.

Then, after the promises comes the time of action, however little has been done, change is lazy. On the one hand, major polluters such as the USA and China refuse to commit to reducing their CO2 emissions. On the other hand, emerging countries such as India or Brazil place their right to development before the climate issue and point the finger at the inaction of other powers.

In 2013, nearly $550 billion in public funds were spent on direct fossil fuel subsidies globally. This shows that the policy toward energy supply is very difficult to change in many countries.

The following study investigates the energy policy of countries for 25 years, between 1991 and 2014. This will show the usage of different sources of energy by countries for electricity production. The project will try to show changes in the energy policies of countries of the world.

# 2 Methodology

The project takes data showing the electricity production of countries from different sources of energy and produce a grouping/clustering based on similar trends of countries. To conduct such an investigation, a k-means clustering is best suited because it is used to find groups which have not been explicitly labeled in the data. First, the data will be prepared in a way columns/features reflect electricity production of countries for a given year from different sources of energy. Then, these set of features are fed into a k-means clustering algorithm. The algorithm will produce clusters of countries which has similar usage of different sources of energy for electricity production. Finally, the appearance of countries in different clusters will be displayed for the 25 years.

# 3 Data Exploration and Preparation

## 3.1 Dataset

The project uses a dataset supplied by the United Nations Statistics Division(UNSD). The data shows energy production, transformation and usage of countries from 1990 to 2014. It entails energy productions from different sources of energy per country making it a suitable dataset for the proposed analysis.

The dataset is presented in a CSV format which makes it easy to adopt to a python work-space and apply data mining algorithms.

Per the UNData terms of use: all data and metadata provided on UNdata's website are available free of charge and may be copied freely, duplicated and further distributed provided that UNdata[1] is cited as the reference.

## 3.2 Data Exploration

The dataset has a size of **124 MB**. It contains strings and numbers. The raw dataset has **1,189,482** rows and **7** columns. Each row presents **country name, commodity transaction, year, unit, quantity, quantity footnotes, and category**. The following figure shows the countries with the most data entries.

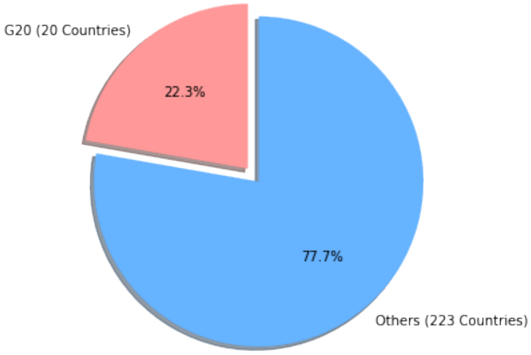


Figure 1: Ratio of data entries by country.

From the above figure, it can be seen that data from developed western nations constitute 22.25% of the entire dataset.

Analyzing the dataset with respect to the year column shows that there's a fair increment of the number of data entries each year. In the **1990s**, there's an average of **42,105** entries per year. In the **2000s**, the average number of data entries per year increases to **51,228**.

The units columns shows the measurement unit of the source of energy described on the entry. The dominant units of measurement are **metric tons, terajouls, and kilowatt-hours**; the most dominant being metric tons, **63.88%** of data entries having it as the unit of measurement.

To validate the correctness of the dataset, the project first conducts a preliminary data analysis. This analysis mines and plots the **conventional crude oil production** of countries and compares it to official reports by **OPEC**. The OPEC

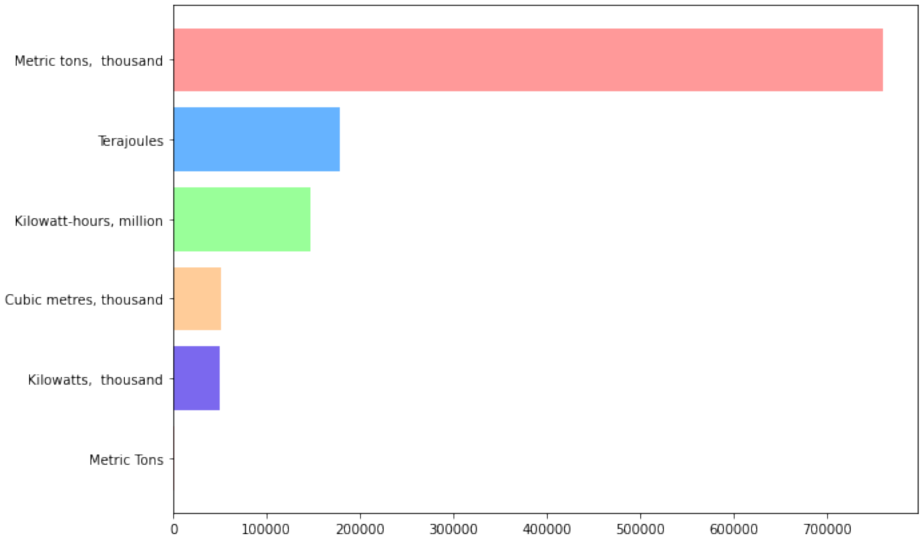


Figure 2: Ratio of data entries by country.

report, **"Annual Statistical Bulletin 2014"**, shows that in 2013 **Russia, Saudi Arabia and United States** were the leaders of crude oil production respectively. The following figure shows a plot of the yearly crude oil production using the dataset.

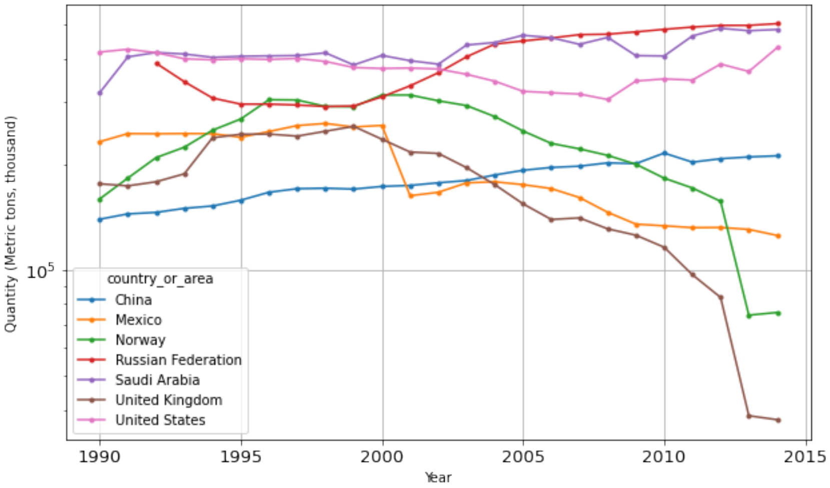
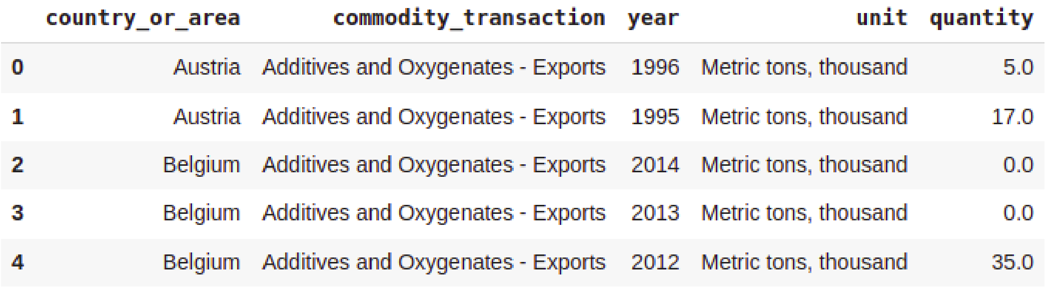


Figure 3: Crude oil production by country.

## 3.3 Data Preparation

The goal of this project is to group countries based on their sources of energy. To conduct such an investigation, the dataset should be arranged in way that each column shows the contribution of a specific energy source to the total energy production of the country. To do this the dataset should be rearranged because each source of energy is listed row-wise as shown below.

The first thing to do is to select the best features to cluster countries. The dataset presents a very wide range of energy production indicators. After reviewing



| | country_or_area | commodity_transaction | year | unit | quantity |
|---|---|---|---|---|---|
| **0** | Austria | Additives and Oxygenates - Exports | 1996 | Metric tons, thousand | 5.0 |
| **1** | Austria | Additives and Oxygenates - Exports | 1995 | Metric tons, thousand | 17.0 |
| **2** | Belgium | Additives and Oxygenates - Exports | 2014 | Metric tons, thousand | 0.0 |
| **3** | Belgium | Additives and Oxygenates - Exports | 2013 | Metric tons, thousand | 0.0 |
| **4** | Belgium | Additives and Oxygenates - Exports | 2012 | Metric tons, thousand | 35.0 |

Figure 4: Crude oil production by country.

these indicators, the most relevant ones were selected. The selection is a list of commodity production which describe electricity production by different sources of energy including **hydro, thermal, geothermal, wind, solar, nuclear** and **tide and wave**.After selecting the portion of data required for the project using the above procedure, the next step is flattening the row arranged values. This was done by one-hot encoding the **commodity transaction** column. This adds a column in the dataframe for all the unique values in the **commodity transaction** column, with its value set to **1** if the row contains the commodity's value and **0** if not. The one-hot encoding process returns a dataframe with **6** features. Next, the one-hot encoded columns are multiplied by the **quantity** column to place the actual quantities in places where the one-hot encode values are 1. Certain columns show **nan** values and they were dropped. It can be seen that there are many zero values but they are not considered as **nan** values. This is because the zero values indicate that a country did not produce a specific commodity for that year. This by itself is a good distinction feature for the clusters. Before performing the clustering task, normalization is done to get rid on any possible biases that may be caused due to differences in the magnitudes of the columns. The values in the rows are normalized in a way that they show their contribution to the total electricity installation. At this stage, the dataframe still shows a single energy source value per data entry. To display the contribution values of all the energy sources for each country-year, the dataframe is grouped by **country name** and **year** and aggregated by sum.

# 4 Energy Supply Analysis

This section will describe the data analysis conducted using K-means implementation provided by SciKit Learn.[2]

## 4.1 Finding the optimal number of clusters

The input of the KMeans algorithm will be a dataset containing **6** features which describe normalized percentage contribution of different energy sources to the total electricity production of a country for a given year.

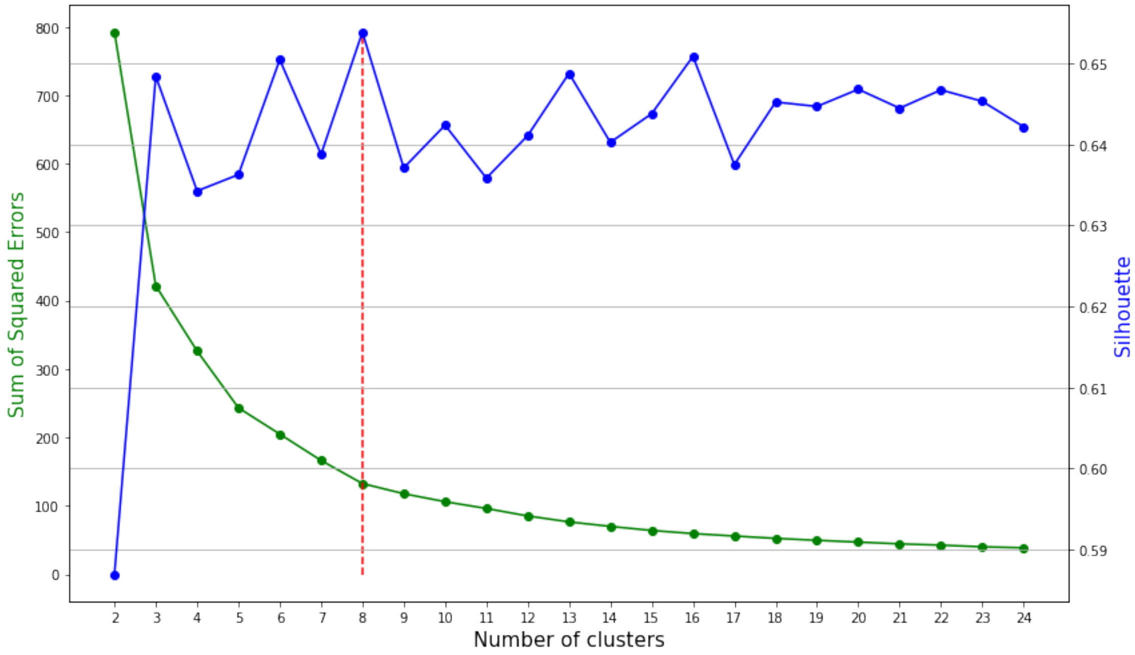


Figure 5: Applying the elbow method to find the optimal number of clusters.

From the previous graph, it is confirmed that the data is separable into groups since the sum of squared errors decreased sharply when increasing the number of clusters before converging. Applying the elbow method doesn't give one clear optimal number of cluster. There seems to be 3 possible values as optimal number of clusters which are **7**, **8** and **9**. Being unable to decide between these values, a silhouette curve is plotted on top of it. For **8** clusters the silhouette values is very high, making it the desirable choice. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually.[3]

## 4.2 Analysis of the clusters found

The k-means algorithm produces **8** cluster centers which are centroids for their respective clusters. The centroids show the average contribution of the specified feature energy sources to the total electricity production for countries in that cluster. The centroids are presented as follows.

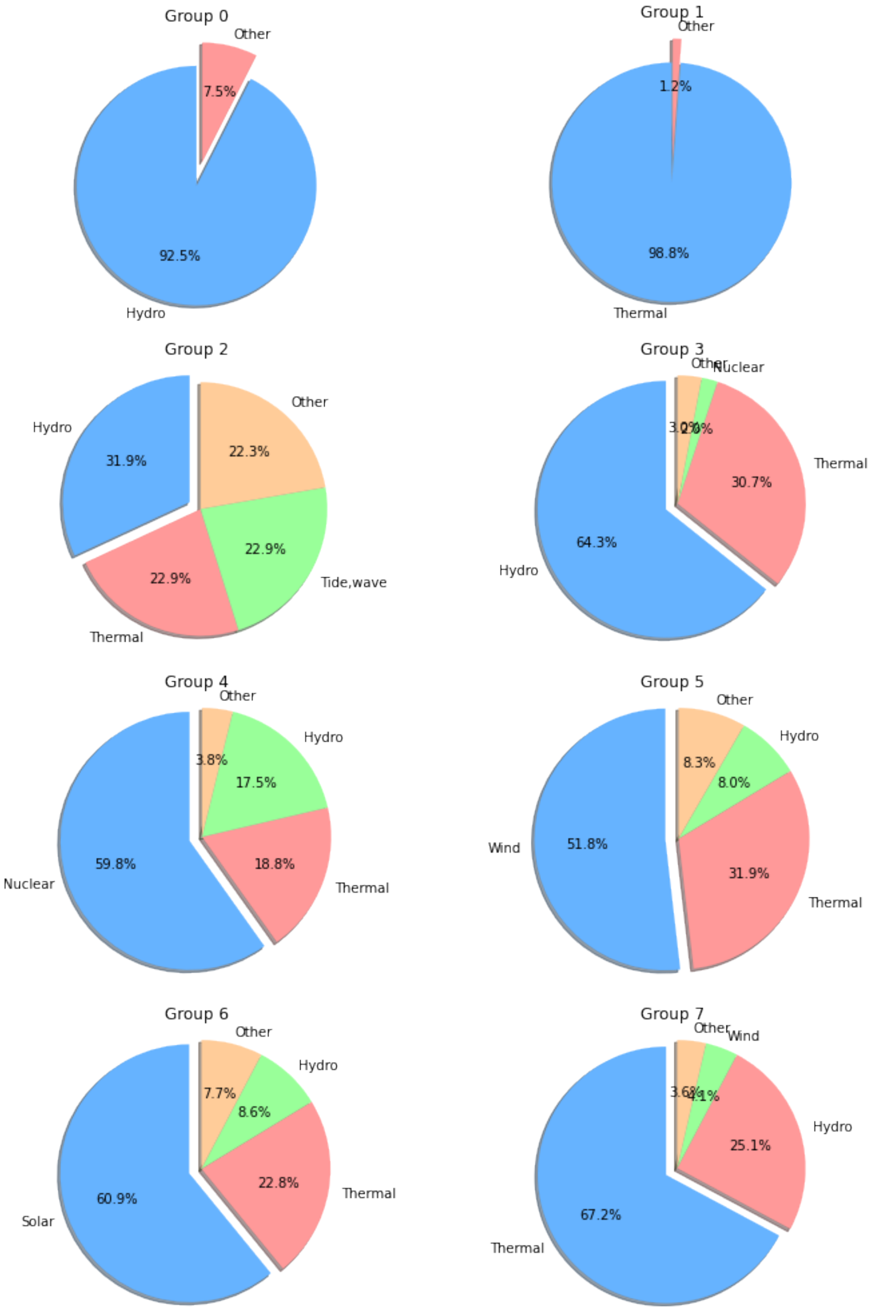


Figure 6: Applying the elbow method to find the optimal number of clusters.

# References

[1] UN Data: Dataset from the UNSD. http://data.un.org/Explorer.aspx. Accessed: 2021-12-10.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] SciKit Learn: Selecting the number of clusters with silhouette analysis on KMeans clustering. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html. Accessed: 2021-12-10.