Global Power Plant Visualisation

Mariam Hersi
HER22555014
Submitted to The University of
Roehampton in partial fulfilment of
the requirements for the degree of
BACHELOR OF SCIENCE IN
COMPUTING

I. Introduction

Dataset Visualisation is a necessary and important tool for gathering information. Through the graphs and images, we create we can deduce what is a correlation and whether that is a causation. In this report, there will visualisation created to analyse a given dataset and come to a conclusion.

A. Purpose and Objectives

Power plants generate electricity that is needed in homes business and industries. They are essential to keep the everyday going. Therefore, understanding the performance of these power plants is also necessary. Creating visualisations that inform us on the power plants allows us to better optimise them and sustain them. The dataset that includes a variety of information which allows us to investigate the environmental impact, the economic impact and the human impact.

First, the initial data preparation is required and then some exploratory data analysis to get an initial insight. From three research questions are formed and that will allow a deeper dive. This report will explain each question and the context and what was found from the visualisations created and any challenges faced. Finally, there is an overall conclusion that summarises the findings and any challenges faced.

The objective is to be able to shed light on the overall landscape of the power plant industry but also discuss topics surrounding sustainability and equity in power generation. This kind of analysis becomes more important with climate change and its impact. We know that greenhouse gas emissions related to energy use account for around 70% of total global emissions, while power generation and heat supply specifically was responsible for 26% of total global greenhouse gas emissions in 2004.

II. Data Preparation and Exploratory Data Analysis (EDA)

There can be no data analysis without first data preparation and exploratory data analysis. It is the backbone of every project and essential to creating the research questions. This facilitates the understanding of the structure of our dataset and removes any inconsistencies. We can get an initial understanding of any patterns that form and chose where we want to explore further.

A. Dataset Overview

The dataset is an open-source open-access dataset of grid-scale (1 MW and greater) electricity generating facilities operating across the world. Using Jupyter notebook. The analysis begins by loading the dataset and then utilise Python to obtain general data info. It found that the dataset has 22 columns and 28664 entries. It includes the key variables:

- Plant Name: The name of the power plant.
- Location: Geographic details, including latitude, longitude, and region.
- Fuel Type: The energy source powering the plant (e.g., coal, gas, hydro, solar, wind, nuclear).
- Capacity (MW): The energy generation capacity of each plant, measured in megawatts.
- Generation: The estimated annual electricity generation in gigawatt-hours for the years 2013 - 2016
- Commissioning Year: The year the power plant began operation, indicating its age.

B. Dataset Cleaning and Preprocessing

Before any analysis, the data is pre-processed to ensure that it will be usable. To clean the data to make sure it was ready for analysis it is split into two categories: numerical and categorical.

1) Numerical Fields

First, it is made sure to deal with the missing values in the numerical columns filling any missing values with a median. This is because in this dataset it can be argued that missing values are missing not at random (MNAR). This means, it is important to deal with missing values as the presence of missing values leads to a smaller sample size than intended and eventually compromises the reliability of the study results (2). Next, there is focus on handling outliers that could occur in the numerical columns. For example, for the capacity column interquartile range (IQR)is used to ensure that extreme values are brought within a reasonable range without removing them. Then the is the conversion of the columns to integers which ensures that year data is treated as numeric rather than as strings. This will simplify any calculations such as age. The rows are filtered to ensure that latitude and longitude values where within specific ranges. This step removes invalid or incorrectly entered geolocation data, ensuring accurate geographic analysis.

2) Categorical Fields.

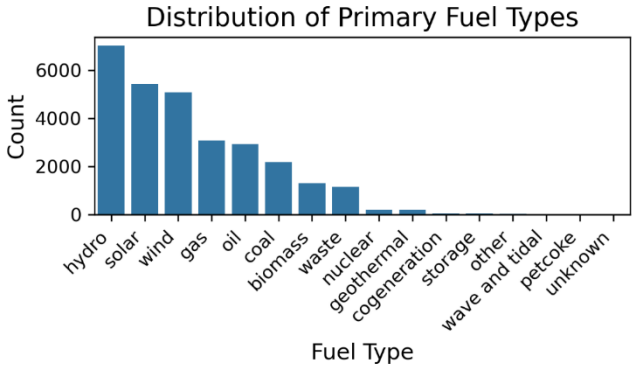
The missing values in the categorical columns were handled by filling them with ‘unknown’. Then for uniformity the categorical columns where standardized by converting text to lowercase and stripping any extra whitespace. This ensures text formatting is the same preventing any duplicates. The pre-processing requires the removal of any duplicates in the gppd_idnr which uniquely identifies power plants. This ensures that there were no redundant records in my dataset which could lead to inaccurate or biased graphs.

Finally, derived features were included to further enhance the analysis. A column called total generation data, a continent column and a plant age column are added. This is then all saved as a new csv file called cleaned dataset.

C. EDA

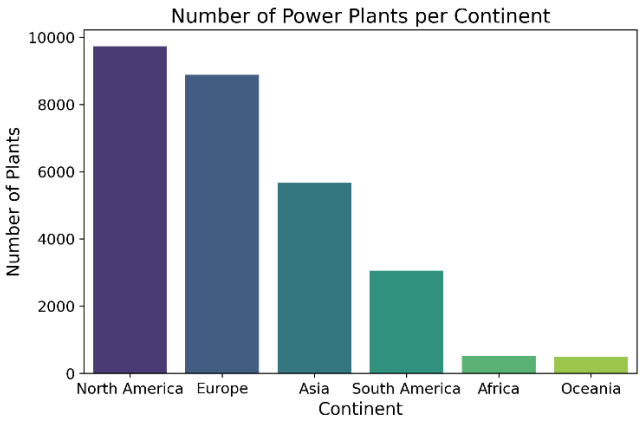
EDA was conducted to give a further insight into the connections between the features in the dataset so that research questions could be created. It allows a better understanding of trends and patterns relevant to the power plant industry and energy production. With EDA you can maximize insight into a dataset, detect outliers and anomalies, and test underlying assumptions (3).

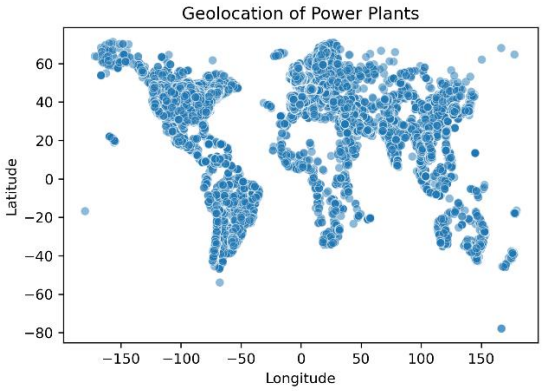
The first analyses to be looked at was the distribution of power plants by fuel types. A bar chart revealed that hydro solar and wind accounted for most power plants worldwide. This shows a global effort to transition to renewables. However, thermal fuel types such as gas oil coal are still used at a large amount. This reflects that there is a still a reliance on them despite the changes being made. It also reflected all the different kinds of fuel types in use around the world. This highlights the innovation and resources that are being put into the power plant industry.



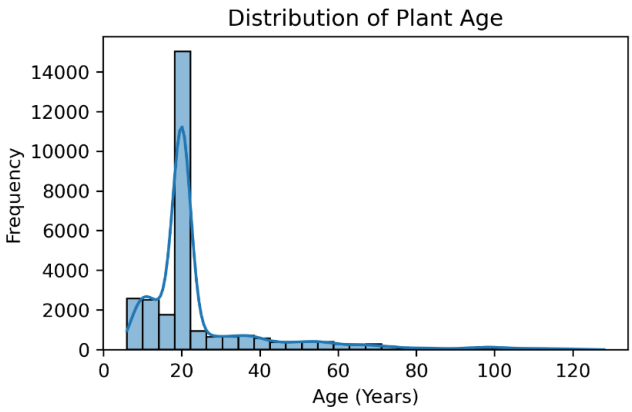
Next the number of power plants per continent was analysed to understand the main producers of these power plants. Industrialised continents, such as the North America and Europe have the highest concentration. This exhibits the energy demands and the needs as they require more power.

This is further shown in the geospatial analysis which highlights each continents differences in power plant distribution and energy strategies. From this we can see the gap and difference for example North America in comparison to Africa.





Lastly, the commissioning year data was analysed to understand the age profile of these power plants. Th bar chart indicates that the majority of the power plants are around 20 years old. This was also seen in the data overview from the summary statistics. The mean commissioning year was found to be 1994 which is exactly 20 years to this year. The are more power plants younger than 20 years than there are power plants older than 20 years. This may perhaps be because of the industrial innovations taking place such as renewable power plants. It could also be because of the increase in demand for energy as more countries begin to become industrialised.



III. Research Questions and Data Visualisation

As a result of the EDA, three research questions were formulated and will be analysed using visualisation. Those visualisations will help formulate conclusions. By leveraging Python libraries such as Pandas, Plotly, Matplotlib, and Seaborn, this analysis explores power plant efficiency across regions, evaluates renewable versus thermal energy trends, and highlights regional variations through visual storytelling.

There are many reasons as to why the gaps in performance can occur. It can be affected by many parameters including the power plant design, size, gas content, dissolved minerals content, parasitic load, ambient conditions and other parameters (5). Therefore, it is important to understand these geographical and technological limitations to understand what areas to improve on.

The has been a worldwide desire for the intensification of the use of different renewable energy sources is essential in order to reach the Paris Agreement or for achieving the goals of sustainable development (4). Therefore, the performance of the power plants is very important. With these research questions we will analyse the performance globally.

A. 1st Question

How does the generation output efficiency vary between renewable and thermal power plants across different geographies?

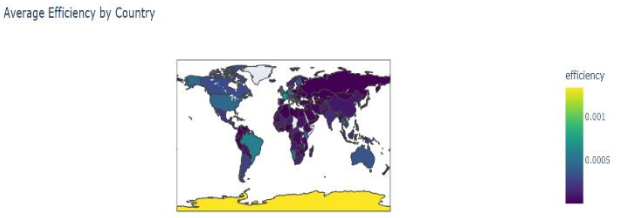
The first research question investigates how the efficiency of between renewable and thermal power plants differs depending on the continent. This report investigates the efficiency of the power plants which was calculated by generation against capacity.

```
df['efficiency'] =
(df['generation_gwh_2016'] /
(df['capacity_mw'] * 8760)).fillna(0)
```

By analysing efficiency of these plants, we can identify gaps in performance and work on improvements. To do this the Power Plants were categorised into two main categories:

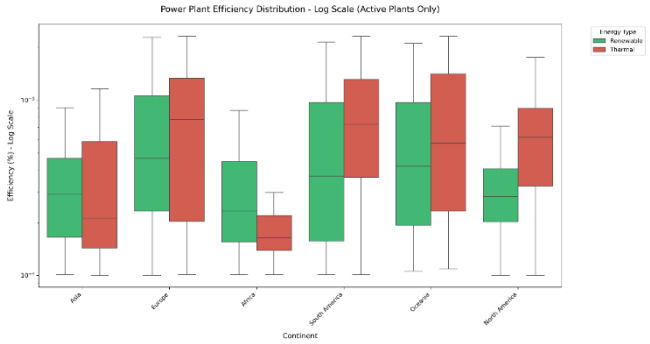
- Renewable: Wind, Solar, Hydro and Geothermal
- Thermal: Coal, Oil, Gas and Nuclear

To visualise the efficiency per country a choropleth was used to highlight the efficiency of power plants per country. We can see that Africa and Asia seem to have high efficiency rates in comparison to Europe, North and South America. In fact, we find that Asia enjoys the highest technical efficiency, and European countries suffer from the lowest technical efficiency among Europe, Asia, and America continents (6). The choropleth highlights the disparities in energy production efficiency, driven by factors such as technology, infrastructure, and energy policies.



In order to delve deeper into the comparison a box plot is put to use. A comparison is made by continent rather than country. When the efficiency is calculated, it is further converted to a log scale to visualise the box plots better and any negative efficiency were removed as those reflect power plant that and not generating energy so perhaps are not in use.

It reflects the difference in thermal and renewable energy. For each continent except Africa and South America we can see that thermal energy is more efficient than renewable energy. We can infer that this is because South America and Africa have better conditions for renewable energy. Thermal energy seems to have better efficiency as it is the most used energy source therefore has had more resources put into it. Its popularity in the modern society comes from its stability and controllability (7).



B. 2nd Question

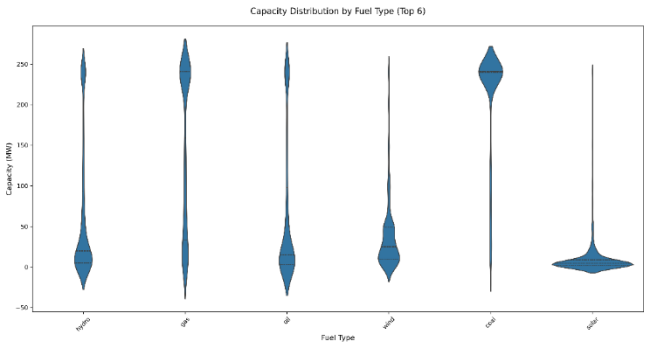
What is the relationship between the capacity (MW) of power plants and their commissioning year across different regions?

With this research question I can investigate whether older plants tend to have lower capacities and how newer plants compare in terms of scale. Understanding this relationship provides insights into how energy production infrastructure has evolved overtime especially regarding the kind of fuel types that exist today and the power plant capacities now. The analysis includes visualizations such as a violin plot and a line graph of capacity trends over time.

This dataset includes many fuel types which was found earlier in the data preparation. It contains around 15 different energy types. To focus on the analysis and improve

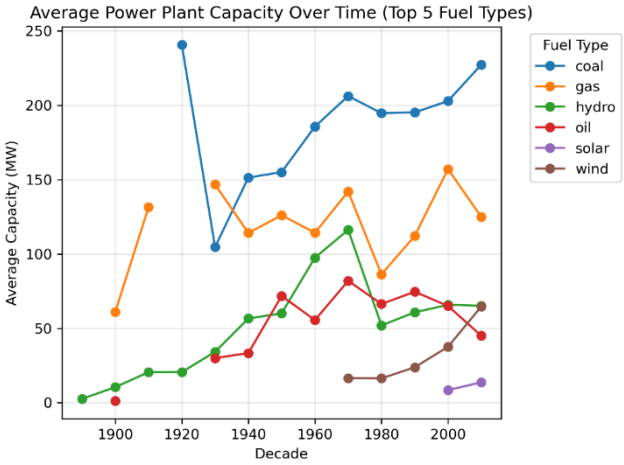
user understanding the top 6 most common fuel types were identified to ensure meaningful analysis of these trends.

The first visualisation graph is a violin graph. This reflects the density and spread of the power plants for each energy source. It suggests that coal and oil have the largest capacity distributions. This reflects the demand and dependency on the with coal and gas being the current primary source for energy generation. Hydro and Gas follow behind with the high capacities. This suggest that more efforts are being put into more renewable energy forms. However, wind and solar have the smallest capacities overall which reflects their reliance on nature and installations in particular locations.



The information that we have gathered can then be compared to time. To analyse how the capacity has evolved the commissioning year of each power plant was organised into descending order and then group by decade. This is important to consider as the percentage of plants currently in operation that are more than 30 years old is rising (7). Therefore, the average capacity for each fuel type will be investigated in and visualised over time.

- Early-20th Century – At this point there are few fuel types in use (coal, hydro) which highlights that this was the cause of the industrial revolution. Before the industrial revolution the energy source for homes and business was simply plant photosynthesis, but because of the revolution accumulated over a geological age in the form of coal (8). As a result, the capacity was also very low and things we just beginning to start out.
- Mid-20th Century – We now have different forms of energy sources being put to use in addition to the previous energy types. This includes oil and gas with coal being the main source of energy and having a steady increase in the capacity for all the types in use. This suggests that has country became more industrialised the demand for energy increased.
- Late -20th Century to 21st – The renewable energy is now being introduced and having an increase in capacity. This is driven the technological advancements and the push for climate initiatives that are better for the environment. We also begin to see a drop in some thermal energy types such as oil. This reflects the depleting stock or fossil fuels that are unrenewable (8).



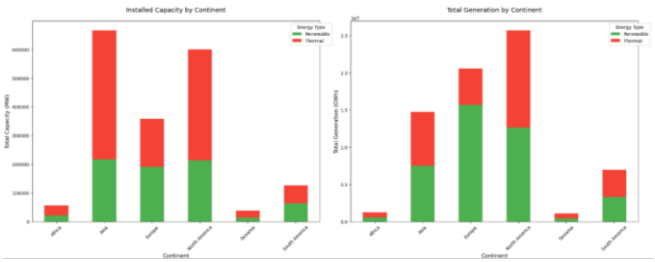
C. 3rd Question

Which energy source (fuel type) contributes the most to total energy generation in each continent?

This question will analyse the regional energy dependency on fossil fuels, renewables, or nuclear energy. Through we can see which kind of energy type continents are prioritising more and weather or the connection between industrialisation and power plants. We analyse this by creating to stacked box charts. One on Total capacity per continent and the other on total generation per continent. To begin with first we must calculate the sum of total energy generation per continent and fuel type.

The first graph shows that there is more capacity for the thermal energy types than there is renewable energy types in all continents except south America. The graph reveals that Asia has the largest capacity with the majority being thermal energy. This is not just unique to Asia, it sems to also be the case for North America and South America. North America and Asia having the largest capacity may perhaps be related to the sizes of the continents as they are the large. However, Africa is the largest continent and yet has one of the smallest capacity and generation. This is like because of demand and industrial requirements of each continent with Europe, Asia and North America requiring more energy and power.

Furthermore, this visualisation shows that is a more total energy generation of the renewable energy type than there is of the thermal energy type in comparison to its capacity. This indicates that there is more renewable energy being used which shows the gradual change towards the renewable energy source. The biggest driver in this direction being Europe although it does not have the largest capacity in comparison to Asia and North America.



IV. Findings

The analysis of power plants worldwide and their fuel types allowed for significant insights and understanding of the global energy industry. In this report, the analyses of efficiency, capacity, power generation, commissioning year and energy types were all analysed. Each showed the relevance of contextual changes and transitions but also the evolution the industries. This gives room for improvements to be made and predictions.

Several challenges arose during the investigation including:

Negative Data: During the creation of the box plot there were power plants that have negative efficiency. A negative efficiency indicates a power plant that is less documented or open but not generating energy. This therefore needed to be removed and focus the box plot only on active power plants.

Visualisation: The creation of the visualisations required careful consideration of the colour schemes and data normalisation to ensure clarity. The dataset was very large and contained a lot of information. This became difficult to visualise and led to graphs that were compressed and difficult to read. This required the grouping of numerical data and creation of the continent as there were too many countries included in the dataset. It also helped to find the average where necessary for the graphs.

V. Conclusion

With each analysis examined power plants across continents. The visualisations used included a log-scale boxplot, a choropleth, a violin graph, a line graph and a stacked bar graph. The box plot indicated that renewable energy sources thermal energy sources typically exhibit higher efficiency rates, and this can be linked to the higher capacity of thermal energy sources and longer-term use. It's an area that has had more resources and time put into it. It also revealed that more developed regions showed higher efficiency rates to. This is because of the higher GDP but also because of the demand for more energy increases and new machinery and technology is used. This evolution and increase in demand were reflected through the line graph. It showed the introduction of a new energy source and its increase in capacity overtime. Through the stacked bar chart, I was able to delve deeper into the production by each continent and compare them to one another. Asia, Europe and North America are leading in power plant capacity and generation. With Asia having the highest capacity and North America the highest generation.

Acknowledgment

I would like to express my gratitude to my professors for their dedication to teaching. I am especially grateful for the time and effort they have invested in teaching me data visualisations, which has contributed to my personal and professional *GROWTH*.

References

[1] T. K. Mideksa and S. Kallbekken, "The impact of climate change on the electricity market: A review," *Energy Policy*, vol. 38, no. 7, pp. 3579–3585, Jul. 2010, doi: https://doi.org/10.1016/j.enpol.2010.02.035

[2] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean Journal of Anesthesiology*, vol. 70, no. 4, pp. 407–411, 2017, doi: https://doi.org/10.4097/kjae.2017.70.4.407

[3] E. Camizuli and E. J. Carranza, "Exploratory Data Analysis (EDA)," *The Encyclopedia of Archaeological Sciences*, vol. 11, pp. 1–7, Nov. 2018, doi: https://doi.org/10.1002/9781119188230.saseas0271

[4] V. Sebestyén, "Renewable and Sustainable Energy Reviews: Environmental impact networks of renewable energy power plants," *Renewable and Sustainable Energy Reviews*, vol. 151, p. 111626, Nov. 2021, doi: https://doi.org/10.1016/j.rser.2021.111626. Available: https://www.sciencedirect.com/science/article/pii/S136403212100900X

[5] T.-Y. Chen, T.-L. Yeh, and Y.-T. Lee, "Comparison of Power Plants Efficiency among 73 Countries," *Journal of Energy*, vol. 2013, pp. 1–8, 2013, doi: https://doi.org/10.1155/2013/916413

[6] T. Zhang, "Methods of Improving the Efficiency of Thermal Power Plants," *Journal of Physics: Conference Series*, vol. 1449, p. 012001, Jan. 2020, doi: https://doi.org/10.1088/1742-6596/1449/1/012001

[7] C. Laire and M. Eyckmans, "Evaluating the condition and remaining life of older power plants," *VGB PowerTech*, vol. 81, Jul. 2001, Available: https://www.osti.gov/etdeweb/biblio/20244652. [Accessed: Jan. 08, 2025]

[8] E. A. Wrigley, "Energy and the English Industrial Revolution," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1986, p. 20110568, Mar. 2013, doi: https://doi.org/10.1098/rsta.2011.0568. Available: https://royalsocietypublishing.org/doi/full/10.1098/rsta.2011.0568

[9] *Types of Missing Values*. Available: https://synapse.koreamed.org/articles/1156715