AN EXPLORATORY ANALYSIS AND FORECAST ON ENERGY GENERATION

TEAM 6

ABHISHIEK KURRA

SATYAM SINGH

THANVI MALYALA

THANMAYEE AKKINENI

BHARATH KUMAR DINDIGALA

George MASON uNIVERSITY

AIT 614-DL2

Big Data Essentials

Dr. Duoduo Liao

MARCH 14, 2023

Index:

1. Introduction:

The universe is a collection of energy in numerous forms. Humans, as minuscule as they are, have created various ways to convert one form of energy into another so that it becomes easier to transport and use. The most used form of energy to date is Electrical energy. Most of the methods and techniques convert different forms of energy into electrical energy. This process is called power generation. Although it is called “generation”, it doesn’t mean that the energy is generated from nothing.

There are many ‘**plants**’ located across the globe that generate electricity. There are numerous types of such plants, each described by the original form of energy, also called **fuel** they are converted from. Some of them are:

- Thermal Energy Plant
- Nuclear Energy Plant
- Solar Energy Plant
- Hydro Energy Plant
- Wind Energy Plant

These can be classified into 2 types depending on their accessibility and resource availability. They are Renewable and Non-Renewable Energy Plants. Renewable for plants that have potentially unlimited access to its fuel within the lifetime of the plant, while non-renewable for plants that have limited access to fuels that are either perishable within the lifetime of the plant or are extinguishable.

Energy is at the epicenter of human evolution (Smil, 2018). As such, it is essential to understand and learn how much energy is being generated and used by the world. This project is an attempt to analyze the energy generated by the world and attempt to forecast future energy generation at high accuracy with the available information.

2. Related Work:

2.1. Global Electricity Generation

This paper gives us details of electricity generation across the world. It involves the energy sources used for generation, i.e. renewable or non-renewable. It contains research done on data from 1980 to 2021 in terawatt-hours. The paper also shares the distribution of energy sources and how each energy source contributes to the world’s energy consumption for each particular year. It explores all countries and their consumption and capacity for different years. Lastly, it also compares the electricity prices in US dollars for each country.

Our paper would not only explore the consumption and capacity of energy generation for each country, but will also forecast future generation and consumption for specific countries. We can use machine learning models which will train on our huge dataset and, based on that, it will help us predict future requirements or trends.

2.2. Energy Production in the US

This paper presents graphs and figures for energy generation in the US. It focuses on energy production by different energy sources, like fossil fuels and renewable sources. The paper explains how each state in the US has different power consumption and shows an overall increasing trend in energy generation from 1980 to 2021. The paper digs deeper to explore the production of energy across renewable and non-renewable energy across all the years. Nonrenewable energy sources like mining of coal, natural gas, and oil are explained via bar graphs. Then we also see the same for renewable energy sources like solar, wind, hydro, and others. Finally, we see the electricity generation in the US by each source of energy, either renewable or non-renewable. Also, the paper gives us a forecast of how energy consumption and generation would look through 2050 for the US.

We plan to find the distribution of energy consumption for the whole world. The paper does not forecast energy consumption across all countries. Using Machine learning models can help predict how energy generation and consumption can look over the years. This would help governments throughout the world to plan better for future requirements.

3. **Objectives:**

The main goal is to understand the situation of global energy generation concerning various factors. These include:

- Does the geolocation of a plant affect the type of power plant?
- Are there any countries inclined to a specific type of plant/plant?
- Is there any growth or decline in the energy generation rates?
- What proportions of the world's energy are each generated using each fuel type?
- Which type of plants tend to have higher plant capacities.
- Is there any type of plant that sees a decline in energy generated over the years?
- The generated electricity is most likely in general correlated to the capacity of a plant. If so, then by how much.

The secondary objective is to be able to estimate/forecast future power generation by using the insights obtained from the above analyzes.

4. **Proposed Selected Dataset:**

The Global Power Plant dataset is an open-source, open-access free-to-use database consisting of details of about 35,000 energy generation plants across the world. The data consists of 36 features/fields of data which include:

S.no	Field	Field Details	Field Type
1	country	ISO 3166-1 alpha-3 standard of country codes	Text
2	country_long	The longer form name of each country	Text
3	name	Name of the plant	Text
4	gppd_idnr	10 or 12-character identifier for the power plant	Numeric
5	capacity_mw	The total capacity of the plant in megawatts	Numeric
6	latitude	geolocation in decimal degrees; WGS84 (EPSG:4326)	Numeric
7	longitude	geolocation in decimal degrees; WGS84 (EPSG:4326)	Numeric
8	primary_fuel	The primary fuel used for electricity generation	Text
9	other_fuel1	Optional or secondary energy source	Text

10	other_fuel2	Optional or secondary energy source	Text
11	other_fuel3	Optional or secondary energy source	Text
12	commissioning_year	year of plant operation	Numeric
13	owner	The majority shareholder of the power plant	Text
14	source	The data source for the plant	Text
15	geolocation_source	attribution for geolocation information	Text
16	url	Web document corresponding to 'source'	Text
17	wepp_id	a reference to a unique plant identifier in the widely-used PLATTS-WEPP database	Text
18	year_of_capacity_data	year the capacity information was reported	Numeric
19	generation_gwh_2013	Electricity generated in the year 2013 (gigawatt-hours)	Numeric
20	generation_gwh_2014	Electricity generated in the year 2014 (gigawatt-hours)	Numeric
21	generation_gwh_2015	Electricity generated in the year 2015 (gigawatt-hours)	Numeric
22	generation_gwh_2016	Electricity generated in the year 2016 (gigawatt-hours)	Numeric
23	generation_gwh_2017	Electricity generated in the year 2017 (gigawatt-hours)	Numeric
24	generation_gwh_2018	Electricity generated in the year 2018 (gigawatt-hours)	Numeric
25	generation_gwh_2019	Electricity generated in the year 2019 (gigawatt-hours)	Numeric
26	generation_data_source	attribution for the reported generation information	Text
27	estimated_generation_gwh_2013	Estimated energy generation for the year 2013	Numeric
28	estimated_generation_gwh_2014	Estimated energy generation for the year 2014	Numeric
29	estimated_generation_gwh_2015	Estimated energy generation for the year 2015	Numeric
30	estimated_generation_gwh_2016	Estimated energy generation for the year 2016	Numeric
31	estimated_generation_gwh_2017	Estimated energy generation for the year 2017	Numeric
32	estimated_generation_note_2013	Label for the model used to estimate energy generation in the year 2013	Text
33	estimated_generation_note_2014	Label for the model used to estimate energy generation in the year 2014	Text
34	estimated_generation_note_2015	Label for the model used to estimate energy generation in the year 2015	Text
35	estimated_generation_note_2016	Label for the model used to estimate energy generation in the year 2016	Text
36	estimated_generation_note_2017	Label for the model used to estimate energy generation in the year 2017	Text

Although the information provided by all the fields is important, some of the data is plain redundant for analysis and will be memory hogs due to the size of the data. Hence, it is important to remove such fields and then continue to further filter and clean data.

The potentially useful fields of data are '**country**', '**name**', '**gppd_idnt**','**capacity_mw**', 'latitude', 'longitude', '**primary_fuel**', all the 3 **other_fuel_***, '**commissioning_year**', all the **genereation_gwh_*year***'s for the year 2013 to 2017. However, the estimated values are skipped and filtered out as we create a new model to forecast energy generation.

5. Methodology:

Any data-driven analytics problem requires to be solved using a standardized methodology. Looking at the data we have, the potential methods that might solve the analytics problem here could be:

a. Data filtering by removing unnecessary data.
b. Eliminating missing values by either dropping the records with missing values or by using intuitive means of replacing missing values with aggregated or summarized values to prevent loss of data.
c. Using summary statistics and visualizations to obtain clear insights.
d. Making proper use of machine learning techniques to forecast worldwide electricity generation.

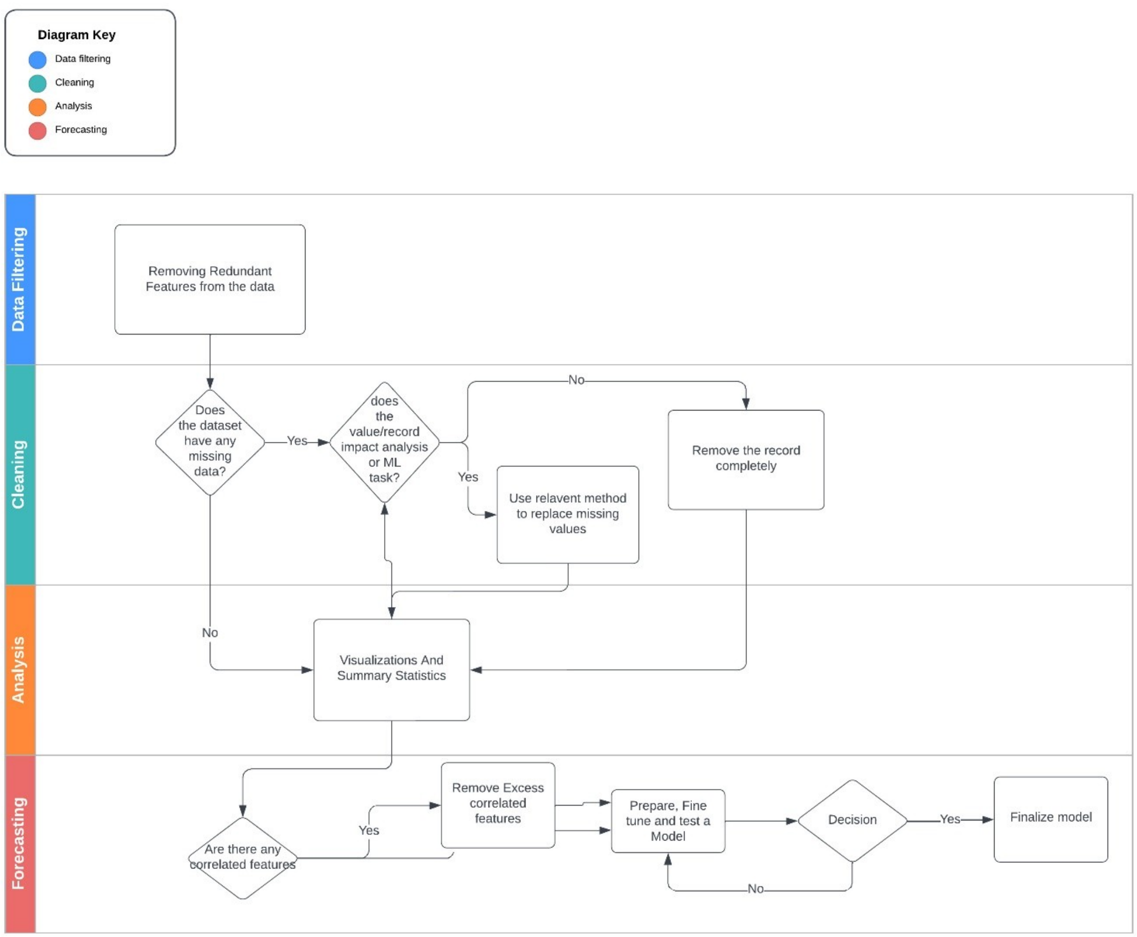

Image Generated Using Lucidchart

6. List of Tasks And Timeline:

Task	Assignment	Start Date	End Date
Select the data to be used	Abhishiek Kurra	3/19/23	3/21/23
Skim through documentation	Satyam Singh	3/21/23	3/24/23
Import into Databricks without data loss	Thanvi Malyala	3/22/23	3/24/23
Identify Redundant Columns	Thanmayee Akkineni	3/25/23	3/28/23
Filter Columns	Bharath Kumar Dindigalla	3/28/23	3/30/23
Detect rows with missing data	Abhishiek Kurra	3/29/23	4/2/23
Identify non normal quantitative data	Thanvi Malyala	4/4/23	4/6/23
Remove recordswith missing data that have least significance	Satyam Singh	4/6/23	4/9/23
Summarize for missing data that belong to significant records	Bharath Kumar Dindigalla	4/6/23	4/8/23
Normalize data	Thanmayee Akkineni	4/6/23	4/9/23
Basic individual feature analysis	Thanmayee Akkineni	3/30/23	4/2/23
Examining interrelations prior to data cleaning	Bharath Kumar Dindigalla	3/31/23	4/4/23
Examinimng Relations post data cleaning	Abhishiek Kurra	4/9/23	4/14/23
Framing Analysis questions	Thanvi Malyala	3/26/23	4/4/23
Answering analysis questions	Satyam Singh	4/9/23	4/13/23
Standardize all data for Machine Learning process	Satyam Singh	3/30/23	4/9/23
Prepare a basic forecasting model	Thanmayee Akkineni	4/6/23	4/9/23
Test Initaial Model	Thanvi Malyala	4/9/23	4/9/23
Use analysis results for simplifying model and improving Confidence Intervals	Bharath Kumar Dindigalla	4/14/23	4/21/23
Test and apply final model	Abhishiek Kurra	4/19/23	4/21/23

Table generated using a template for Gantt chart from **Vortex42.com**

7. Proposed Development Platforms

7.1. Databricks DBFS

Databricks Distributed File System is a file system mounted into AZURE data bricks workspace and available in the Databricks cluster. It allows you to interact and mount cloud object storage into the Databricks directory and workspace instead of using cloud-specific API commands. It simplifies object storage and allows virtual machines and volume storage to be safely deleted after cluster termination. It also provides a convenient location to store scripts, libraries, and configurations for each cluster. FileStore is a special folder within DBFS that helps you save files and access them locally with your web browser.

7.2. PySpark

Apache Spark is an open-source, distributed processing system for big data workloads written in Scala. PySpark is a python API for spark, which enables collaboration between python and spark. This helps in data-related tasks on a single node or multiple clusters. As python is an open-source language, it has benefits, such as multiple libraries, which boost data manipulation tasks. It also acts as an interface for Resilient Distributed Datasets by using the Py4j library.

7.3. Spark MLlib

MLlib is a library over PySpark and Spark, which enables us to apply machine learning models on huge datasets across all clusters. It uses data parallelism techniques to store and work with large amounts of data. It is scalable and consists of machine learning algorithms, such as classification, regression, clustering, filtering, dimensionality reduction, and many more. Spark MLlib can integrate with other spark libraries like SparkSQL, and Spark Streaming as per the user's requirements. It can help in preprocessing, training, and making predictions on the scale.

7.4. Databricks

The Databricks platform provides a combination of tools to build, train, share, deploy, and maintain big data solutions on a scale. Databricks platform can be used for many applications or to convert BI solutions to Machine Learning Solutions for different enterprises. Data bricks users can take full advantage of all the available resources provided by the platform, including interactive notebooks, SQL editor, pipelining, discovery, compute management, workflow scheduler, etc.

8. References:

Smil, V. (2018). *Energy and civilization: A history*. The MIT Press.

Jocelyn, V., & Biagi, L. (n.d.). *Energy production in the US* Statista. Retrieved March 16, 2023, from https://www.statista.com/study/48975/energy-production-in-the-united-states/?locale=en

Jocelyn, V., & Biagi, L. (n.d.). *Global Electricity*. Statista. Retrieved March 16, 2023, from https://www.statista.com/study/74593/electricity-worldwide/

What is the Databricks File System (DBFS)? What is the Databricks File System (DBFS)? | Databricks on AWS. (n.d.). Retrieved March 16, 2023, from https://docs.databricks.com/dbfs/index.html

What is pyspark?: Domino data science dictionary. What is PySpark? | Domino Data Science Dictionary. (n.d.). Retrieved March 16, 2023, from https://www.dominodatalab.com/data-science-dictionary/pyspark

What is Databricks? What is Databricks? | Databricks on AWS. (n.d.). Retrieved March 16, 2023, from https://docs.databricks.com/introduction/index.html

Intelligent diagramming. Lucidchart. (n.d.). Retrieved March 16, 2023, from https://www.lucidchart.com/pages

Simple gantt chart. Vertex42.com. (n.d.). Retrieved March 19, 2023, from https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html?utm_source=ms&utm_medium=file&utm_campaign=office&utm_content=url

9. Appendix:

Reference Data Truncated to the first 25 records.

country	country_	name	gppd_idr	capacity_	latitude	longitude	primary_	other_fue	other_fue	other_fue	commiss	owner	source	url	geolocati	wepp_id	year_of_
AFG	Afghanis	Kajaki Hy	GEODB0	33	32.322	65.119	Hydro						GEODB	http://glot	GEODB	1009793	2017
AFG	Afghanis	Kandaha	WKS007	10	31.67	65.795	Solar						Wiki-Sol	https://ww	Wiki-Solar		
AFG	Afghanis	Kandaha	WKS007	10	31.623	65.792	Solar						Wiki-Sol	https://ww	Wiki-Solar		
AFG	Afghanis	Mahipar	GEODB0	66	34.556	69.4787	Hydro						GEODB	http://glot	GEODB	1009795	2017
AFG	Afghanis	Naghlu D	GEODB0	100	34.641	69.717	Hydro						GEODB	http://glot	GEODB	1009797	2017
AFG	Afghanis	Nangarh	GEODB0	11.55	34.4847	70.3633	Hydro						GEODB	http://glot	GEODB	1009787	2017
AFG	Afghanis	Northwes	GEODB0	42	34.5638	69.1134	Gas						GEODB	http://glot	GEODB		2017
AFG	Afghanis	Pul-e-Kh	GEODB0	6	35.9416	68.71	Hydro						GEODB	http://glot	GEODB		2017
AFG	Afghanis	Sarobi D	GEODB0	22	34.5865	69.7757	Hydro						GEODB	http://glot	GEODB	1009799	2017
ALB	Albania	Bistrica 1	WRI1002	27	39.9116	20.1047	Hydro				1965		Energy C	http://ww	GEODB	1021225	
ALB	Albania	Fierza	WRI1002	500	42.2514	20.0431	Hydro				1978		Energy C	http://ww	GEODB	1021231	
ALB	Albania	Koman	WRI1002	600	42.1033	19.8224	Hydro				1985		Energy C	http://ww	GEODB	1021233	
ALB	Albania	Lanabreg	WRI1002	5	41.3428	19.8964	Hydro				1951		Energy C	http://ww	GEODB	1021236	
ALB	Albania	Shkopet	WRI1002	24	41.6796	19.8305	Hydro				1963		Energy C	http://ww	GEODB	1021238	
ALB	Albania	Ulez	WRI1002	25	41.6796	19.8936	Hydro				1958		Energy C	http://ww	GEODB	1021241	
ALB	Albania	Vau i Dij	WRI1002	250	42.0137	19.6359	Hydro				1971		Energy C	http://ww	GEODB	1021242	
ALB	Albania	Vlora	WRI1002	98	40.4874	19.434	Other						Energy C	http://ww	GEODB	1021244	
DZA	Algeria	Adrar	WKS006	20	27.908	-0.317	Solar						Wiki-Sol	https://ww	Wiki-Solar		
DZA	Algeria	Ain Azel	WKS006	20	35.88	5.475	Solar						Wiki-Sol	https://ww	Wiki-Solar		
DZA	Algeria	Ain Djas	WRI1023	520	35.8665	6.0262	Gas	Oil				SociÃƒÅ	Arab Uni	http://ww	KTH	1069670	
DZA	Algeria	Ain Sekh	WKS006	20	34.532	0.804	Solar						Wiki-Sol	https://ww	Wiki-Solar		
DZA	Algeria	Ain el Ibe	WKS006	20	34.346	3.164	Solar						Wiki-Sol	https://ww	Wiki-Solar		
DZA	Algeria	Ain el Ibe	WKS007	53	34.342	3.169	Solar						Wiki-Sol	https://ww	Wiki-Solar		
DZA	Algeria	Ain el Me	WKS006	20	34.861	4.204	Solar						Wiki-Sol	https://ww	Wiki-Solar		
DZA	Algeria	Algerie S	WKS006	43.5	27.908	-0.317	Solar						Wiki-Sol	https://ww	Wiki-Solar		

name	generatic	generatic	generatic	generatic	generatic	generatic	generatic	generatic	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimatec	estimated_generation_note_
Kajaki Hy									123.77	162.9	97.39	137.76	119.5	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Kandaha									18.43	17.48	18.25	17.7	18.29	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Kandaha									18.64	17.58	19.1	17.62	18.72	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Mahipar									225.06	203.55	146.9	230.18	174.91	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Naghlu D									406.16	357.22	270.99	395.38	350.8	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Nangarh									58.77	54.42	42.71	59.72	46.12	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Northwes														NO-EST	NO-EST	NO-EST	NO-EST	NO-ESTIMATION
Pul-e-Kh									21.99	21.19	18.4	25.34	19.74	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Sarobi D									123.23	82.87	69.15	93.83	80	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Bistrica 1									105.17	75.26	79.5	105.45	88.45	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Fierza									1976.01	1276.61	1503.72	1795.15	1648.24	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Koman									2072.13	1618.73	1805.63	2434.84	1982.72	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Lanabreg									20.37	12.89	14.64	20.04	15.23	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Shkopet									93.52	69.86	77.51	96.2	83.57	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Ulez									97.42	72.77	80.74	100.21	87.06	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Vau i Dij									895.02	561.94	614.47	897.47	703.64	HYDRO-	HYDRO-	HYDRO-	HYDRO-	HYDRO-V1
Vlora														NO-EST	NO-EST	NO-EST	NO-EST	NO-ESTIMATION
Adrar										35.22	34.22	35.33	35.17	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Ain Azel										38.68	37.56	38.37	38.75	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Ain Djas													2171.28	NO-EST	NO-EST	NO-EST	NO-EST	CAPACITY-FACTOR-V1
Ain Sekh										34.85	33.67	34.54	35.46	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Ain el Ibe										33.42	33.58	34.75	34.81	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Ain el Ibe										80.98	81.84	85.66	85.55	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Ain el Me										33.64	33.68	33.63	33.75	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE
Algerie S										73.79	72.11	74.36	74.02	NO-EST	SOLAR-'	SOLAR-'	SOLAR-'	SOLAR-V1-NO-AGE