Report

Dhruv jain & Venkata Dhanush Kikkisetti

2022-12-16

ACKNOWLEDGEMENTS

"I am not what happened to me, I am what I choose to become" by Christopher Gardner, The Pursuit of Happiness.

It is always a pleasure to remember the fine people who guided me in the Data Science program. I received to uphold my practical and theoretical skills during the respective session. Firstly, I would like to thank Pro. Rabya Ghafoor and secondly, I want to thank my family & friends for their love, motivation, and support during this semester in American university. Thanks for all the ideas, opinions, knowledge, and suggestions given to me to help me to complete this report. We are very thankful to American University for giving us the opportunity to pursue this project.

1. Title Page with Executive Summary

Title: ESTIMATING POWER PLANT GENERATION USING GLOBAL POWER PLANT DATA

Type of analysis: Application analysis

Table 1:

Name	course
Dhruv Jain	Data -612
Venkata Dhanush Kikkisetti	Data -612

Summary: The type of analysis is **application based**. This dataset is collected from different open-source databases manually and later combined. Only 29.4 percent of capacity in database is geolocated using national data. All the other data is taken from other secondary sources or manually identified via satellite imagery and most of the nationals did not reveal their power plant details.

The database collects the following characteristics and indicators: All types of fuel

- Technical characteristics (fuel, technology, ownership)
- Operational characteristics (generation)
- Plants' geolocation
- Plants over 1 megawatt (MW)
- Plants in operation only (in first iteration)

Data source: **website name**: WORLDS RESOURCES INSTITUE.**Citation:** Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, Enipada, World Resources Institute. 2018. Global Power Plant Database. Published on Resource Watch. Website link

Overall approach is Cleaning, analyzing, again cleaning, answering the five questions, playing with data to deliver more better output and finally graphical representation. **Defining** the issues and trying to resolve that by presenting the power plant data. **Measure** overall what can be done. **Analyze** the data to use for future capability. **Improving** You will use information gathered in the previous phases to design and implement improvements in processing with consistency. Overall Approach to this question is using various tools and coding sets with providing the statistical data with convincing evidence.

Clean –> Design –> Plot —> compare –> hypothesis –> statistical findings –>conclusion

Primary findings:
- Choosing a particular country and analyzing what type of powerplant has more electricity generation.

- Analyzing which type of powerplant has more electricity generation and doing a little research on it.

- Analyzing if there is a difference in generated electricity and estimated electricity with respect to the country.
- Effect of power plant type in generating electricity.
- Comparing electricity generation in different years with respect to country

Recommendations: Using R-studio one can achieve high graphical results with better quality and one can also conclude that big data can be easily handled on this platform. Using this platform, we will continue to evaluate the power plant generation data using this tool. The issue is that what can be done with the data can we provided some explanation to justify the results. We would like to address those five questions with different visual ideas.

IMPORTANT KEYWORDS:

- Power generation growth
- estimated power generation
- Country
- type of fields
- P-value

2. Introduction: provide context for why and how you are doing the analysis.

The main **motivation** of analysis is to figure out whether the data provided is accurate or not with statistical findings and comparing them after log transformation. The **overarching research** of interest is choosing global data and explaining the visual content in a manner that one can differentiate whether it is right or wrong.

Data source is from the Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, Enipada, World Resources Institute. 2018. Global Power Plant Database. Published on Resource Watch. The **data time period** is from 2013 to 2019. The number of rows is more than **35,000** and it contains **33 different observations**.

Brief Literature review:

This dataset consists of electricity generated by different countries around the globe using different power plants (Hydro, Gas, Oil, solar, wind, nuclear, biomass, waste) from 2013 to 2019. The data is more than 35,000 or rows and contains 33 different variables. The data is reliable, and the results kind is supporting the graphs and all the assumptions we have claimed in this report making.

The methodology blends statistical regression with machine learning techniques. Explanatory variables include plant-level characteristics such as plant size and fuel type, and country-level characteristics, such as country- and fuel-specific average generation per megawatt of installed capacity [1]. We show that fuel-specific models can provide more accurate results for wind, solar, and hydro plants. Estimations for natural gas plants also improve, but the error remains high, especially for smaller plants [1].

3. Initial Hypotheses:

The Dataset currently contains more than 30,000 power plants in 164 countries, representing about 82% of the world's capacity between 2014 to 2018. We will find if total current generation increase with increase in the years and check the pair wise comparison between the years.

Based on the questions answered in the following report one can say that the variable name generation is associated with the type of fuel used in a particular year to generate electricity for our country.

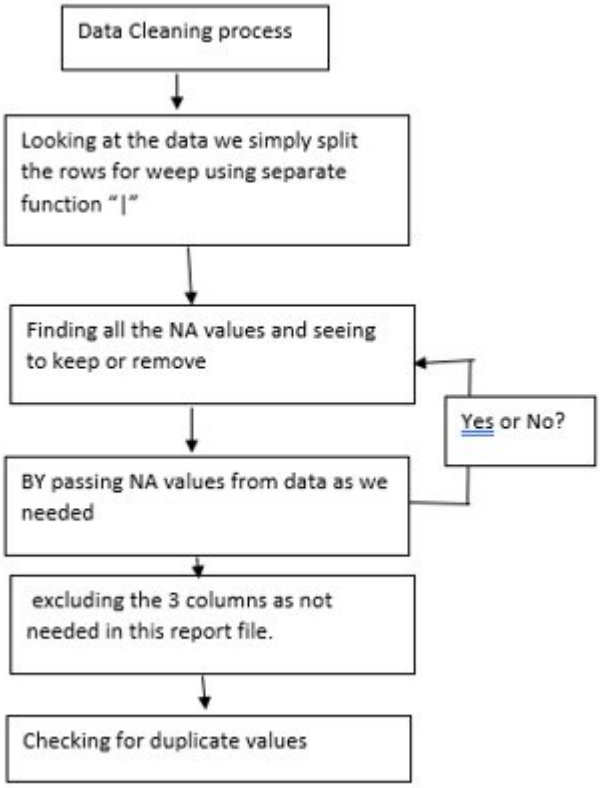
- To statistically perform a test for comparing electricity generation for different years, we should have to consider some assumptions that are true for the sample.
- Data should have a normally distribution
- Homogeneity of variance

4. Data Preparation:

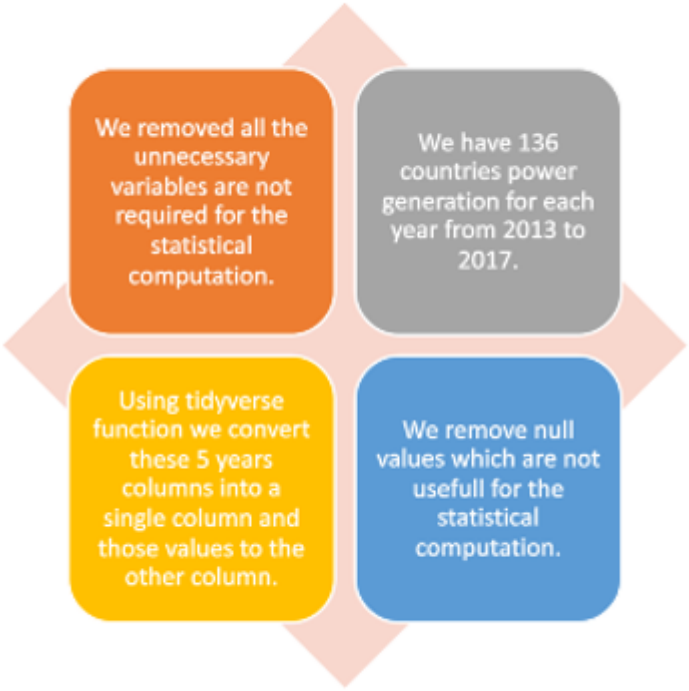
initially figuring out the **data collection** was a little bit challenging for us because the site contains 30% of data and the rest data was collected from satellites and different types of sources other than that to you get some values, we did some research and from the other source we combined the data into a single excel sheet. Dealing with the large data set was difficult after more than 35,000 rows.

The dataset consists of many variables which are not required for making statistical conclusions for our analysis. We have current generation for 164 countries and each country has power generation for all these 5 years. So, we considered taking all the 5-year columns into a new table and started our analysis. Using Tidy verse function to convert years columns into single columns and their values to the other column. The dataset consists of many NA values as most countries do not publicly report their power sector data.

Data cleaning process is explained below:



For the analysis part again the tiding you can't was taken into place, and we remove all the unnecessary columns then we did the log transformation for the data after that we were able to see much more better results and we concluded the hypothesis using some test.



5. Statistical Analysis with Interpretations

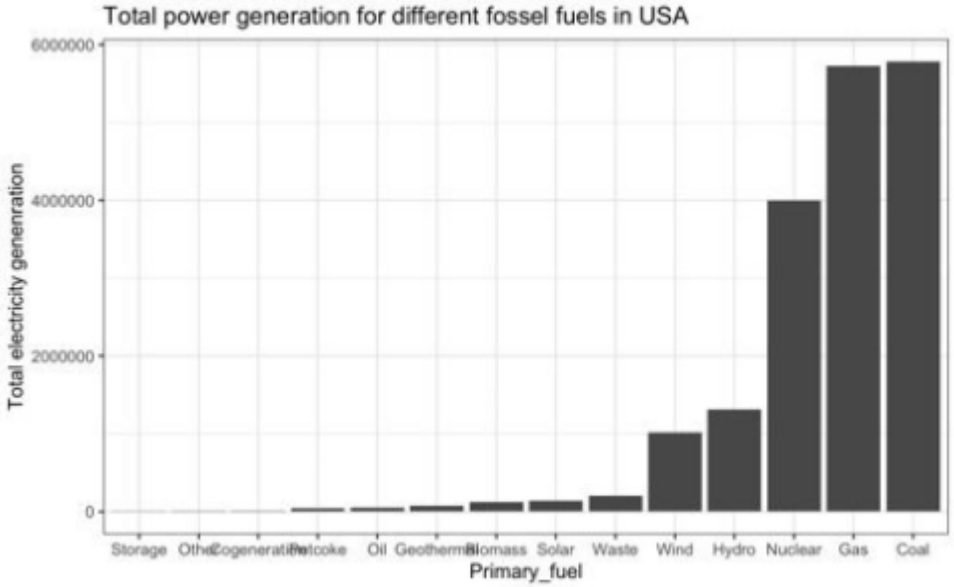
A. Top 10 countries with high power generation

After cleaning and tidying process, we Are showing the top ten countries which has highest power generation among all the other countries. One can observe from this table below that the US stands the first rank in terms of total power generation across the years from 2013 to 2018. India stands 2nd and Australia stands at the third rank across the globe.

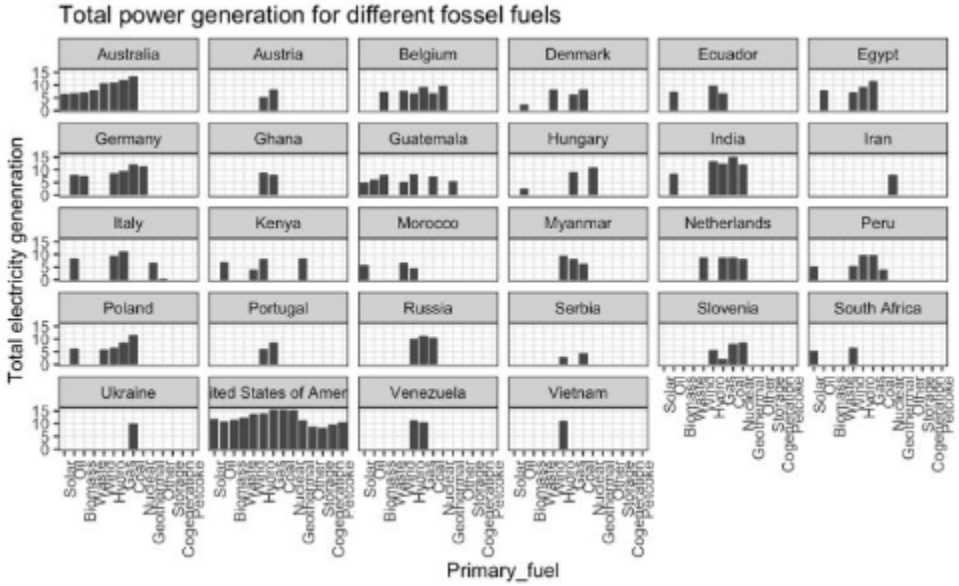
country <chr>	Total_power_generation <dbl>
USA	18498152
IND	5060381
AUS	961431
DEU	275922
RUS	143644
EGY	124140
VEN	120004
POL	114285
ITA	91183
HUN	69864

B. Country Wise analyzing type of powerplant has more electricity generation.

one can observe from the below graph that it states all types of fuels and from that how much a country can generate this chart is particularly for US. Below this graph one can observe the global flower plants across the world and can state which country has which power generation more.



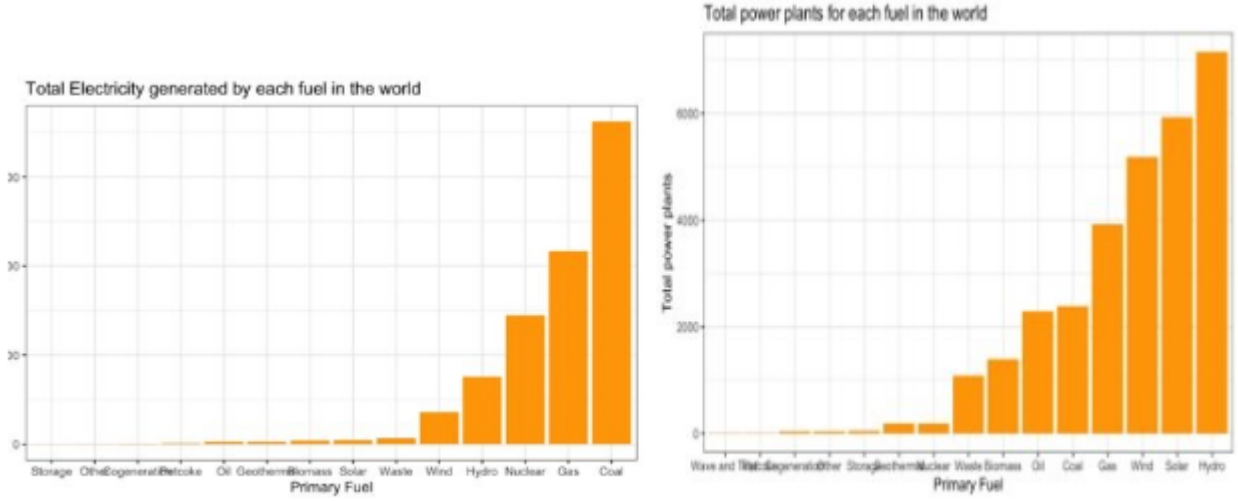
It compares all the countries power plant generation.



C. No. Of power plant vs their generation growth

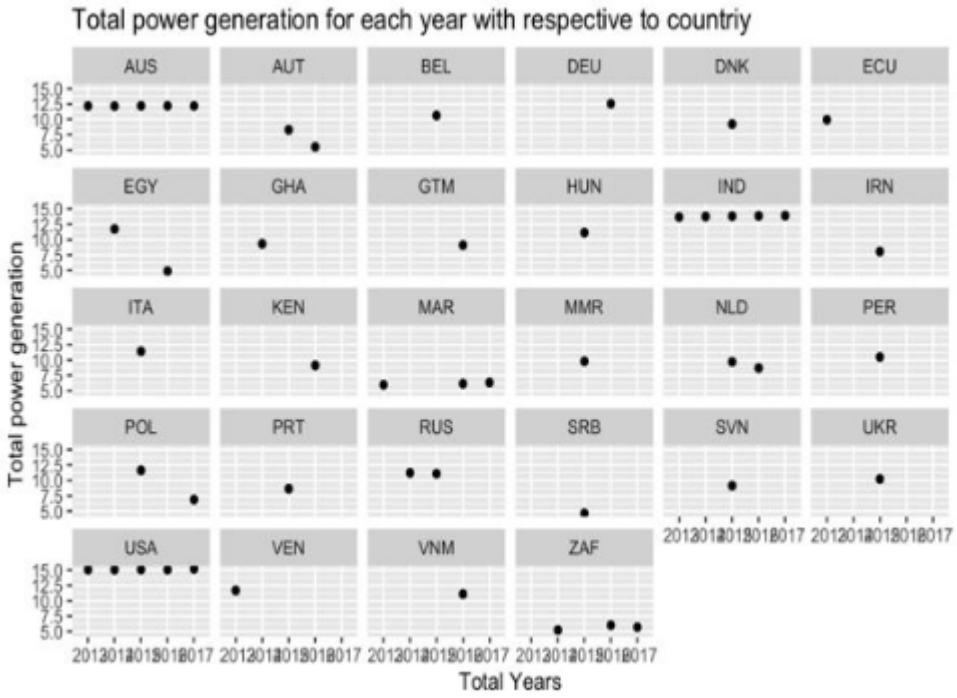
The first chart shows that coal has the highest generation of electricity, and the gas is the 2nd highest generating electricity across the globe. We also see that in the second chart the number of plants is hydro, solar and wind but still we generate more electricity from coal which is a nonrenewable resource. renewable

resources are more and can be used more often but the generation rate is way too low even though we have a high number of plants in that sector.



D. Country wise power generation After log transformation

One can clearly observe from this chart that the US stands number one in terms of total power generation across the globe. We can also see different types of countries showing their production unit across the years from 2013 to 2017. from this data we can see that year wise generation has been increased for few countries and the year wise generation was stable for some countries. A little bit strange that the country Netherlands' production has been decreased from 2015 to 2016. and for some countries it has increased like India.



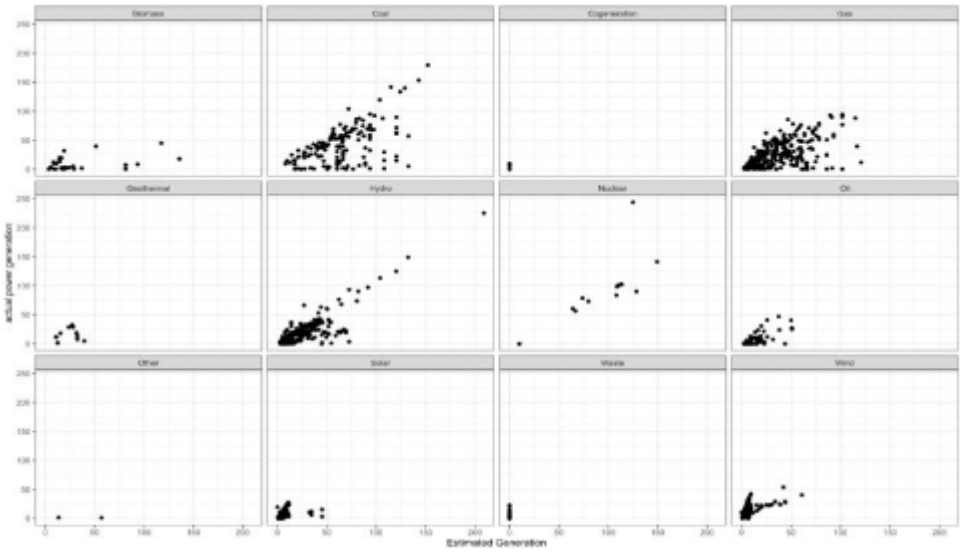
E. Analyzing if there is a difference in generated electricity and estimated electricity with respect to the country.

In the below table one can clearly observe the estimated power generation versus total power generation across the globe.

country_long	Total_power_generation	Estimated_power_generation
Australia	9604	23006.4
Austria	4202	9299.7
Belgium	40549	62774.3
Denmark	9978	20081.1
Ecuador	20610	21713.7
Egypt	133	14591.1
Germany	275922	353268.0
Guatemala	9053	9752.4
Hungary	69864	18510.5
India	121441	462896.0
Iran	3198	4472.0
Italy	91183	153997.1
Kenya	9320	9085.9
Morocco	1428	565.7
Myanmar	17814	12225.1
Netherlands	22471	43405.2
Peru	36496	43676.0
Poland	114285	114630.7
Portugal	5689	10579.6
Russia	66720	60366.6
Serbia	103	26578.8
Slovenia	9284	12899.7
South Africa	723	102.0
Ukraine	27518	24019.4
United States of America	299571	469795.7
Venezuela	120004	90006.5
Vietnam	66775	58228.7

F. Actual vs estimated current generation

This graph shows the actual versus predicted power generation across the planet we can see the generation estimation for particular coal fuel was much more accurate rather than the renewable energy generations the major difference where the country was not able to deliver this because of time duration or the money factor.



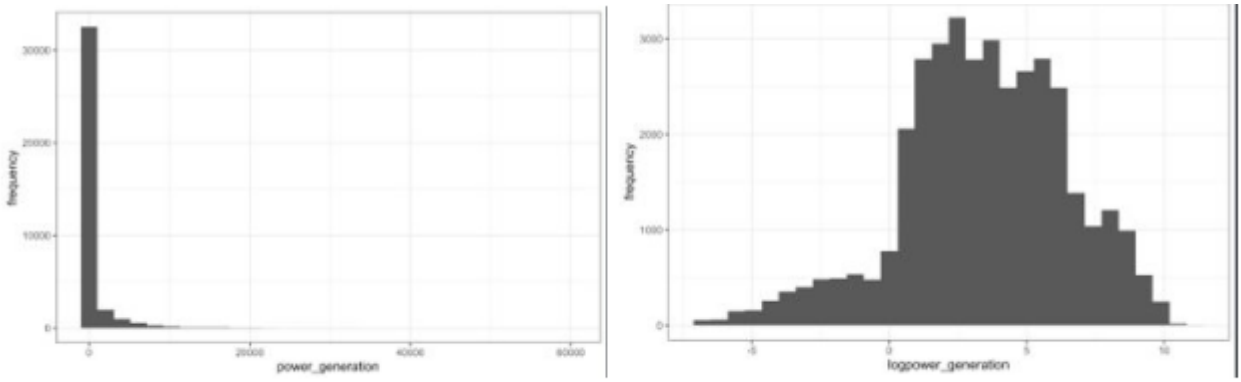
G. EDA:We get a simple statistic on power generation for each year

The below table represents the average, median, standard deviation and interquartile range of total power plants generation of electricity from 2013 to 2017.

Years	Total_power_plants	Average	Median	SD	IQR
2013	6258	771.8	42.56	2451	305.3
2014	6640	745.9	34.79	2378	278.5
2015	7257	746.3	35.22	2428	295.0
2016	7952	659.3	30.35	2217	246.4
2017	8716	613.4	26.07	2116	216.4

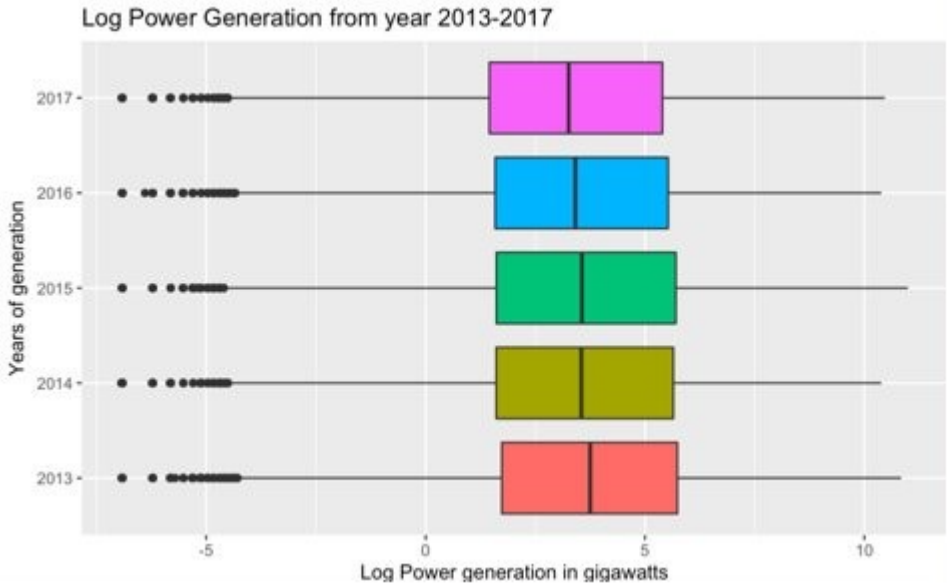
H. Skewness

And can clearly observe and the first chart the histogram graph was skewed towards right after the log transformation it was uniformly distributed.

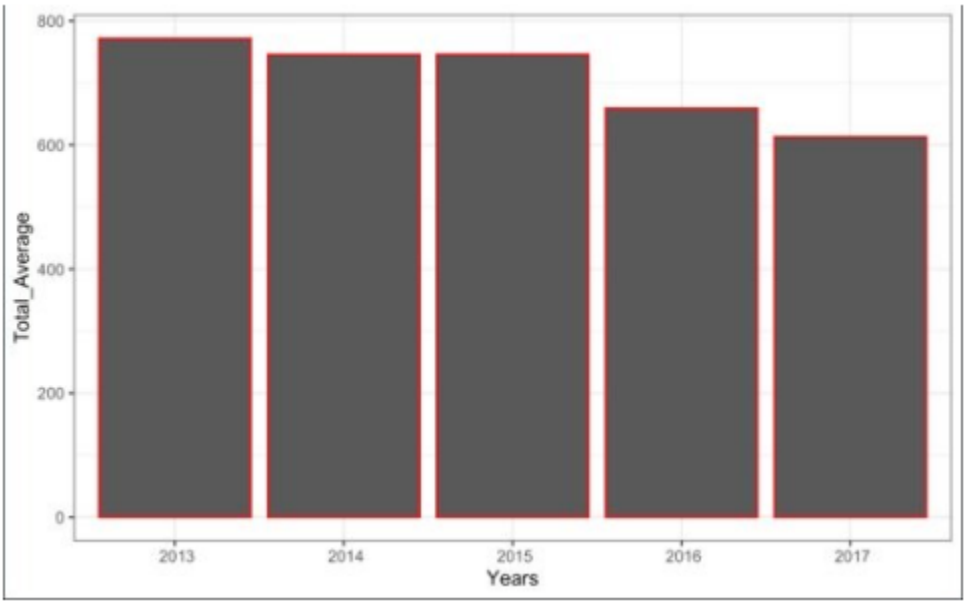


I. Box plot

just a normal visualization representation of how the log of power generation in gigawatts versus the years of generation looks in the data. We can clearly observe the decrease in and the increase using the below box plot chart.



Histogram is also plot and from the year 2013 to 2017 one can clearly see that there is a decrease in the total number of averages of our generation.



J. Statistical analysis

When we did the analysis of variance tests for years, we're going to clearly see that the P value is less than 4 decimal places which means the test which we ran shows that this is significant. all the hypothesis assumed, and the data provided was correct according to the statistics that the energy generation was decreased

```
anova_model<-aov(log_generation~Years,data=powerplant_new)
anova(anova_model)
```

```
Analysis of Variance Table

Response: log_generation
            Df Sum Sq Mean Sq F value    Pr(>F)
Years        4    293    73.3    8.01 0.0000019 ***
Residuals 36818 337309     9.2
```

K. Confidence interval level

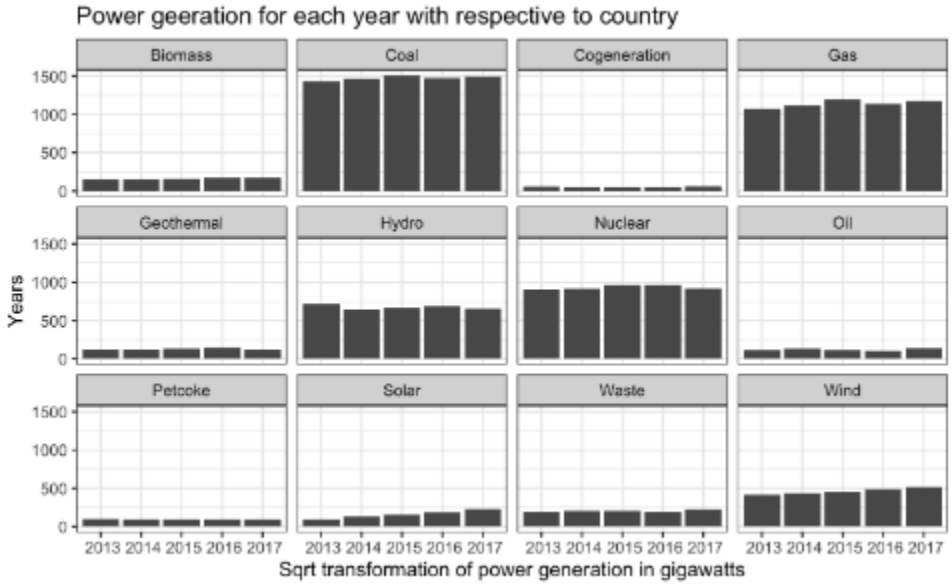
Year wise comparison of generated electricity as an experiment we can observe that from 2017 to 2013 the P value is less than 0.0001 this means that there is strong evidence of decrease in the power generation across the years from 2013 to 2017.

```
$Years
               diff   lwr.ci   upr.ci      pval
2014- 2013 -0.07070 -0.17522  0.03382   0.18491
2015- 2013 -0.01903 -0.12137  0.08331   0.71555
2016- 2013 -0.12555 -0.22580 -0.02530   0.01411 *
2017- 2013 -0.24061 -0.33890 -0.14231 0.0000016 ***
2015- 2014  0.05167 -0.04908  0.15242   0.31478
2016- 2014 -0.05485 -0.15347  0.04377   0.27570
2017- 2014 -0.16990 -0.26654 -0.07327   0.00057 ***
2016- 2015 -0.10652 -0.20283 -0.01021   0.03018 *
2017- 2015 -0.22158 -0.31585 -0.12730 0.0000041 ***
2017- 2016 -0.11506 -0.20706 -0.02306   0.01424 *

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L. year-wise all types of fuel generation from year 2013-2017

One can clearly see in the below a visual work that coal generation has increased over the years for all the countries the second highest production for electricity is gas and the third highest is nuclear. So, the primary source of generation is still nonrenewable resources which are hard to reproduce again over a short period of time. On the other hand, we can see that renewable resources like wind solar biomass are really taking place nowadays and increasing the generation year on year. However, we see a stable observation in the petcock generation.



6. Summary:

From the statistical test we have strong evidence that the current generation is not same for all the years with p values less than 0.001. The overall visualization, we can say that there is a decline trend as year passes by. From a little research we found that, with shortages of natural gas and coal leading to volatile prices, demand destruction and negative effects on power generators, retailers and end users, notably in China, Europe and India. In the year 2014 global growth in electricity demand took place in China, where demand grew by an estimated 10%. China and India suffered from power cuts at certain points in the second half of the year because of coal shortages.

The easiest and the best possible solution when there is a crisis is solar energy. in upcoming days solar energy will create a good impact, harmless and best way to regenerate electricity with simplicity. solar generation is something that the governments are giving subsidies to the owners to install and generate more electricity. some countries have lands which is useful where they can install the solar plants and generate good electricity if they set the solar panel at a right angle.

Future implementation:

- One could take the excel sheet and fill in all the missing data to get much more accurate and precise results than what we are getting right now.
- Another possible could be taking the data and plotting several types of scatter plots instead of distinct types of histograms or bar charts to get a better visualization.
- one can also try to do a deep dive research in the solar industry that will it be the next future.

7. Appendix

Appendix A:

References

[1] Cawse-Nicholson, K., Townsend, P. A., Schimel, D., Assiri, A. M., Blake, P. L., Buongiorno, M. F., et al. (2021). NASA's surface biology and geology designated observable: A perspective on surface imaging algorithms. Remote Sensing of Environment, 257, 112349.link

[2] Cusworth, D. H., Duren, R. M., Thorpe, A. K., Tseng, E., Thompson, D., Guha, A., et al. (2020). Using remote sensing to detect, validate, and quantify methane emissions from California solid waste operations. Environmental Research Letters, 15(5), 054012.link

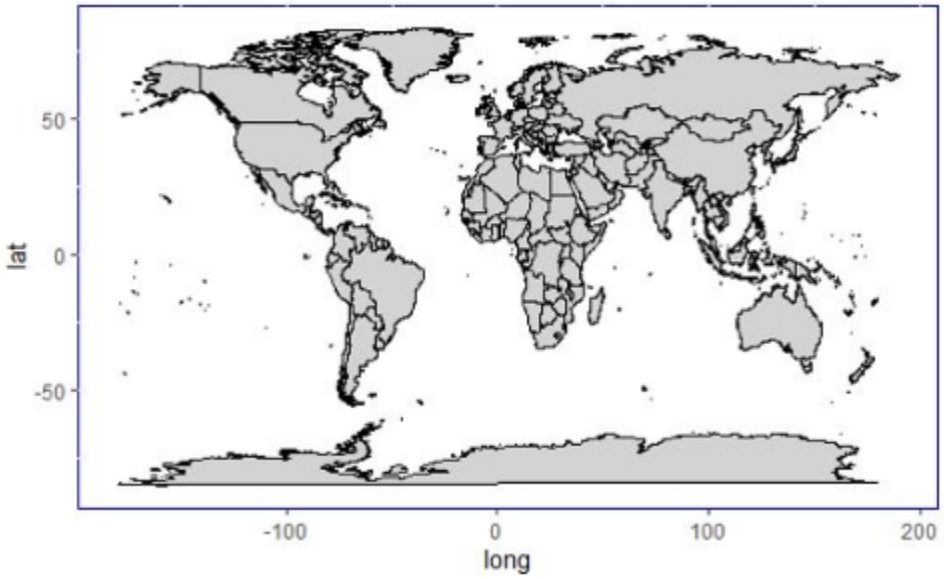
[3] Green, R. O., Mahowald, N., Ung, C., Thompson, D. R., Bator, L., Bennet, M., et al. (2020). The Earth surface mineral dust source investigation: An Earth science imaging spectroscopy mission. In 2020 IEEE Aerospace Conference (pp. 1– 15). IEEE.

[4] Hockstad, L., & Hanel, L. (2018). Inventory of U.S. greenhouse gas emissions and sinks. United States Environmental Protection Agency. link

Appendix B:

figure 1:

This figure is world map done in r-code using inbulid library.


50
0
-50
lat
-100
0
100
200
long