```
---
title: "Global Power Plant Dataset Report"
date: "2023-04-12"
output:
  word_document:
    toc: yes # toc = table of contents
    toc_depth: 2 # refers to the number of "levels" of headings shown
  pdf_document:
    toc: yes
    toc_depth: 2
    fig_width: 7
    fig_height: 6
    fig_caption: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Intorduction

The power station also known as a power plant first invented in the late 18's by the Belgian engineer Zénobe Gramme. He made the invention of the world's first electric generator which produces electricity to support commercial purposes. After his invention, many engineers from different parts of the world started to work on building machines like a generator and soon in 1882 world is successful to create its first coal-fired power station in London. As time passes away evolution happened in the industry of power plants and nowadays we have more than 9 different types of energy power plants that help us to fulfill our energy requirements.
For more details on the history of power plants [Click Here]
(https://en.wikipedia.org/wiki/Power_station#:~:text=9%20External%20links-,History,estate%20to%20power%20Siemens%20dynamos).

As we know nowadays power plants play an important role in the completion of energy requirements of the world. In this data set of **Global Power Plant** We are going to analyze the data of 9 different types of power plants around the globe. In this data we will discuss the types of power plants**(primary_fuel)**, their capacity to store energy** (capacity_mw)**,their location**(longitude,latitude)**,the region where they are belongs to**(country_long)**, and many more. This data is provided by the International institute resources which gives you the record till 2021. For more details you can visit <https://datasets.wri.org/dataset/globalpowerplantdatabase>.

In this report we will discuss the major steps include in data analysis and figure out the answers of our research questions and prove our answers with the help of visualization of data in the form of charts or presentation. Here is the summary of the data.

```{r}
global_power_plant_database <- read.csv("F:/Data Work/global_power_plant_database_v_1_3/global_power_plant_database.csv")
data(global_power_plant_database)
View(global_power_plant_database)
summary.data.frame(global_power_plant_database)
```

Research Questions

**Q1. ** An appropriate location for the plantation of power plants is very important as it helps to minimize transmission cost and losses. So it is too important to know What is the most preferred geographical location for the plantation of power plants?

**Q2. ** Power plants consume lot of energy in order to deliver their purpose. The energy consumption is the major factor when it comes to cost evaluation and yield efficiency. The Question is which primary fuel for power plants is preferred with respect to the continent.

**Q3. **

Variables of Interest##

Following are the variables we are going to use in our analysis in order to find the answer of the above questions.

1. country:

This variable is used to represent the data of the countries involve in the Global Power Plant database. It contains String form of data which shows the short form of every country name like it contains **USA** instead of **United States of America**. Here is the code.
```{r}
barplot(table(global_power_plant_database$country))
```


2. country_long:

This variable also works like the previous country variable. But in case of country_long it will return the complete name of the country present in the database. For example: it will return **Pakistan** instead of **PAK**. Here is the code.
```{r}
barplot(table(global_power_plant_database$country_long))
```

3. capacity_mw:

This variable will give you detail info regarding the capacity of power plants to store energy in the unit of megawatt. The
```

minimum value of an entity in the data set is 1.00mw while the maximum value is 22500.00mw.
Here is the code.
```{r}
barplot(table(global_power_plant_database$capacity_mw))
```

4. latitude:

Will give information regarding the latitude of every power plant available in the data set. the minimum latitude regarding during analysis was -77.85 while the maximum latitude was 71.29. These insights will help us to conclude the perfect environment for the plantation of power station.Here is the code.
```{r}
barplot(table(global_power_plant_database$latitude))
```

5. longitude:

This variable will help us to find the longitudinal position of a power house. The minimum value of a power house recorded was -179.978 and the maximum value was 179.389. This info will help us to give the answers of our research questions.Here is the code.
```{r}
barplot(table(global_power_plant_database$longitude))
```

6. primary_fuel:

will give us information regarding the types of fuels us to run power stations all over the world there are multiple types of fuels use in the world but the most prominent are gas, coal, hydro etc. This information is useful to find the most abundant source of fuel for power plant around the globe. Here is the code.
```{r}
barplot(table(global_power_plant_database$primary_fuel))
```

Data cleaning/wrangling

By using the function of **summary()** we will get to know the actual amount of observations in the data which is **34936**. Analyzing this amount of data is to hectic and so we will divide this data into small chunks to make it more readable and error free. Following are the methods we use in data cleaning/wrangling.

1. Filter Data:
One of the most used function when it comes to data cleaning is **filter()** this will help you to filter the you need in you analysis. As our data is consist on the global data collection and in the world we have 6 continents, So make it more sense I divided this data on the basis of continents and filter each continent countries data into a separate Data Frame. The code for the filter() is given below.

*Asia Code**

```{r}
library(dplyr)
Asia <- filter(global_power_plant_database, country_long == 'Pakistan' | country_long == 'China' |
                 country_long == 'Indonesia' | country_long == 'Iran' | country_long == 'Israel'
               | country_long == 'Japan' | country_long == 'Japan' | country_long == 'Singapore'
               | country_long == 'Philippines' | country_long == 'Bangladesh' | country_long == 'Uzbekistan'
               | country_long == 'Kazakhstan'  | country_long == 'Bahrain' | country_long == 'Myanmar'
               | country_long == 'Jordan' | country_long == 'Afganistan' | country_long == 'Oman'
               | country_long == 'Qatar' | country_long == 'Syria' | country_long == 'Tajikistan'
               | country_long == 'Thailand' | country_long == 'Yemen' | country_long == 'Cambodia'
               | country_long == 'Vietnam' | country_long == 'Laos' | country_long == 'Turmenistan'
               | country_long == 'Saudia Arabia' | country_long == 'Kuwait' | country_long == 'Lebanon'
               | country_long == 'Iraq' | country_long == 'United Arab Emirates' | country_long == 'Nepal'
               | country_long == 'Sri Lanka' | country_long == 'Malaysia' | country_long == 'South Korea'
               | country_long == 'Maldives' | country_long == 'Mongolia' | country_long == 'Bhutan'
               | country_long == 'Palestine' | country_long == 'North Korea' | country_long == 'Armenia'
               | country_long == 'Azerbaijan' | country_long == 'Turkiye' | country_long == 'Georgia'
               | country_long == 'Taiwan' | country_long == 'Russia' | country_long == 'Macao' | country_long == 'Hong
Kong')
```

Africa

```{r}
Africa <- filter(global_power_plant_database, country_long == 'Nigeria' | country_long == 'Ethiopia' |
                 country_long == 'Egypt' | country_long == 'Tanzania' | country_long == 'South Africa'
               | country_long == 'Kenya' | country_long == 'Uganda' | country_long == 'Algeria'
               | country_long == 'Angola' | country_long == 'Morocco' | country_long == 'Sudan'
               | country_long == 'Mozambique'  | country_long == 'Ghana' | country_long == 'Madagascar'
               | country_long == 'Burkina Faso' | country_long == 'Niger' | country_long == 'Cameroon'
               | country_long == 'Mali' | country_long == 'Malawi' | country_long == 'Zambia'
               | country_long == 'Somalia' | country_long == 'Chad' | country_long == 'Senegal'
               | country_long == 'Zimbabwe' | country_long == 'Guinea' | country_long == 'Rwanda'
               | country_long == 'Tunisia' | country_long == 'Burundi' | country_long == 'Benin'
               | country_long == 'South Sudan' | country_long == 'Togo' | country_long == '     Sierra Leone'
               | country_long == 'Liberia' | country_long == 'Congo' | country_long == 'Libya'
               | country_long == 'Central African Republic' | country_long == 'Mauritania' | country_long == 'Eritrea'
               | country_long == 'Botswana' | country_long == 'Gambia' | country_long == 'Namibia'
               | country_long == 'Gabon' | country_long == 'Lesotho' | country_long == 'Equatorial Guinea'
               | country_long == 'Seychelles' | country_long == 'Comoros' | country_long == 'Cabo Verde' | country_long ==
'Mauritius')
```

```

**Europe**

```{r}
Europe <- filter(global_power_plant_database, country_long == 'Germany' | country_long == 'United Kingdom' |
                 country_long == 'France' | country_long == 'Itlay' | country_long == 'Spain'
               | country_long == 'Ukraine' | country_long == 'Poland' | country_long == 'Romania'
               | country_long == 'Netherlands' | country_long == 'Belgium' | country_long == 'Czech Republic'
               | country_long == 'Sweden'  | country_long == 'Portugal' | country_long == 'Greece'
               | country_long == 'Hungry' | country_long == 'Belarus' | country_long == 'Austria'
               | country_long == 'Serbia' | country_long == 'Switzerland' | country_long == 'Bulgaria'
               | country_long == 'Denmark' | country_long == 'Finland' | country_long == 'Slovakia'
               | country_long == 'Norway' | country_long == 'Ireland' | country_long == 'Croatia'
               | country_long == 'Albania' | country_long == 'Bosnia and Herzegovina' | country_long == 'Moldova'
               | country_long == 'Lithuania' | country_long == 'Russia' | country_long == 'North Macedonia'
               | country_long == 'Estonia' | country_long == 'Latvia' | country_long == '       Slovenia'
               | country_long == 'Montenegro' | country_long == 'Luxembourg' | country_long == 'Malta'
               | country_long == 'Monaco' | country_long == 'Andorra' | country_long == 'Iceland'
               | country_long == 'Liechtenstein' | country_long == 'San Marino' | country_long == 'Holy See')

```


South America

```{r}
South_America <- filter(global_power_plant_database, country_long == 'Brazil' | country_long == 'Colombia' |
                   country_long == 'Venezuela' | country_long == 'Peru' | country_long == 'Argentina'
                 | country_long == 'Chile' | country_long == 'Ecuador' | country_long == 'Bolivia'
                 | country_long == 'Guyana' | country_long == 'Uruguay' | country_long == 'Paraguay'
                 | country_long == 'Suriname'  | country_long == 'Portugal' | country_long == 'French Guiana'
                 | country_long == 'Falkland Islands')
```

North America

```{r}
North_America <- filter(global_power_plant_database, country_long == 'United States of America' | country_long == 'Mexico'
|
country_long == 'Haiti' | country_long == 'Guatemala' | country_long == 'Canada'
| country_long == 'Cuba' | country_long == 'Dominican Republic' | country_long == 'Honduras'
| country_long == 'Costa Rica' | country_long == 'El Salvador' | country_long == 'Nicaragua'
| country_long == 'Panama'  | country_long == 'Jamaica' | country_long == 'Puerto Rico'
| country_long == 'Belize' | country_long == '  Guadeloupe' | country_long == 'Trinidad and Tobago'
| country_long == 'Bahamas' | country_long == 'Martinique' | country_long == 'Barbados'
| country_long == 'Saint Vincent and the Grenadines' | country_long == 'Grenada' | country_long == 'Saint Lucia'
| country_long == 'Aruba' | country_long == 'United States Virgin Islands' | country_long == 'Antigua and Barbuda'
| country_long == 'Bermuda' | country_long == 'Cayman Islands' | country_long == 'Dominica'
| country_long == 'Greenland' | country_long == 'Saint Kitts and Nevis' | country_long == 'Sint Maarten'
| country_long == 'British Virgin Islands' | country_long == 'Saint Martin' | country_long == 'Turks and Caicos Islands'
| country_long == 'Caribbean Netherlands' | country_long == 'Anguilla' | country_long == 'Saint Pierre and Miquelon'
| country_long == 'Monaco')
```

We didn't include Antarctica as there is no power plant so it will not effect our analysis.As now we make 5 data frames from the main data set it will be esay to make visualization of data and makes our analysis a bit more easy.

2. Sorting

After dividing data into small data frames now it is time to sort data make it ready for our analysis. In sorting we have 2 ways to sort one sort data in ascending order and one in descending order the choose in yours. in this data we sort data in descending order. To arrange data in descending order we use arrange function with name of data set by placing **-** sign at the start of the data frame.

Here is the code to sort the data by using **arrange()** function.

```{r}
##arrange(Asia,desc(-capacity_mw))
## I put this code in a comment as it generate and place all of the data of database in a word file

```

That's how you can arrange data in the remaining data frames too. Sometimes it's important to rearrange data but sometimes it doesn't matter. In our case to answer question 1 sorting of data doesn't matter. So it depends on you. If you want to sort it then above code will help you out in it.

Choice of data visualisations and rationale

Now it's time to visualize our data by using some functions and packages in this assignment I use **ggplot2** for data visualization the procedure to use it to first install the package by using command **install.packages("ggplot2")**. Once it installed then you need to import it by using **library(ggplot2)**. When it comes to visualization you have so many ways to visualize you data by using different kinds of graphs and Visualizations, So it's hard to decide which one is best for your visualization. So there are some tips to choose the best way to represent your data.

1. The visualization should be clear.
2. Should be easy to understand.
3. It should be fine and engaging.
```

4. The representation of data should be prominent.

As our data frames has data of thousands of rows so it's really hard to present it. So I chosen point graph to give the answer of my first question by making a visualization of every data frame I created. Here is the code of the visualization.

Visualizations For Question 1

Asia

```{r}
library(ggplot2)
ggplot(data = Asia) + geom_point(mapping = aes(x = latitude, y = longitude,
color = primary_fuel))+labs(title = "Power Plants in Asia",
subtitle = "Geographical Location and Abundance",
caption = "World Resources Institue")
```

Africa

```{r}
ggplot(data = Africa) + geom_point(mapping = aes(x = latitude, y = longitude,
color = primary_fuel))+labs(title = "Power Plants in Africa",
subtitle = "Geographical Location and Abundance",
caption = "World Resources Institue")
```

North America

```{r}
ggplot(data = North_America) + geom_point(mapping = aes(x = latitude, y = longitude,
color = primary_fuel))+labs(title = "Power Plants in North America",
subtitle = "Geographical Location and Abundance",
caption = "World Resources Institue")
```

South America

```{r}
ggplot(data = South_America) + geom_point(mapping = aes(x = latitude, y = longitude,
color = primary_fuel))+labs(title = "Power Plants in South America",
subtitle = "Geographical Location and Abundance",
caption = "World Resources Institue")
```

Europe

```{r}
ggplot(data = Europe) + geom_point(mapping = aes(x = latitude, y = longitude,
color = primary_fuel))+labs(title = "Power Plants in Europe",
subtitle = "Geographical Location and Abundance",
caption = "World Resources Institue")
```

Visualization For Question 2

As for question 2 we need to find out the Percentage of abundance of power plants in every continent so that we will figure out how power plants vary from continent to continent and which form of fuel used as the most preffered fule in the region. So let's satrt creating the visualizations. Below are the codes and visualization for every data frame we filter from the original data set.

Asia

```{r}
ggplot(data = Asia, aes(x = primary_fuel)) +
  geom_density(fill = "lightblue", color = "darkblue", alpha = 0.5) +
  labs(x = "Primary Fuel", y = "Abundance %", title = "Asia Power Plants")
```

Europe

```{r}
ggplot(data = Europe, aes(x = primary_fuel)) +
  geom_density(fill = "yellow", color = "red", alpha = 0.5) +
  labs(x = "Primary Fuel", y = "Abundance %", title = "Europe Power Plants")
```

Africa

```{r}
ggplot(data = Africa, aes(x = primary_fuel)) +
  geom_density(fill = "green", color = "black", alpha = 0.5) +
  labs(x = "Primary Fuel", y = "Abundance %", title = "Africa Power Plants")
```

South America

```{r}
ggplot(data = South_America, aes(x = primary_fuel)) +
```

```
  geom_density(fill = "pink", color = "lightblue", alpha = 0.5) +
  labs(x = "Primary Fuel", y = "Abundance %", title = "South America Power Plants")
```

North America

```{r}
ggplot(data = North_America, aes(x = primary_fuel)) +
  geom_density(fill = "yellow", color = "blue", alpha = 0.5) +
  labs(x = "Primary Fuel", y = "Abundance %", title = "North America Power Plants")
```

These are the visualizations that will help us to answer all the questions in out research.

Conclusions

for the final conclusion the data insights we get from our analysis is when it comes to **answer** the **Question 1** then in case of Asia we saw that most of power plants are satiated between **latitude 25 to 50 and longitude 50 to 100** which shows the shows diverse and complex nature of an environment, which may be influenced from the nature and human factors. Studies shows that in Asia there is diversity of every natural phenomena due to which most of the resources are around the areas where we saw diverse behavior of weather, wildlife, climate and many more. In the case of Africa there is no specific environment restrictions as in Africa the climate is consistent at every place that's why the power plants in Africa are scattered. When it comes to North America then there most of power plants are located in between the **latitude 30 to 50 and longitude -75 to -125** which shows that there the area is consist on rocks and mountains which face high level of every weather and have rational human activities. In South America the weather is extremely hot and tropical as it is almost at equator. Most of the power plant in that region is located at **latitude -25 to 0 and longitude -60 to -40**. Last but not the least In Europe you will see that the abundance of power plants is greater at **latitude 30 to 60 and longitude 0 to 50**. These area usually have cold weather through out the year. In concluding question 1's answer when we talk about the world most of power plants are planted at the **latitude 30 to 50 and longitude -50 to 50**. In short where we find diverse nature and environment that will be the more appropriate place for the plant neither too near the equator nor too far.

Now in order to give the an answer to the **Question 2** we will take a reference from the answer of Question 1 and continue to answer. As for Asia most of the power plants are using coal as their primary fuel in the power plants, it is because Asia is full of coal natural resources and makes it a cheap fuel and high yield efficiency.In Asia almost 65% of power plants are run on Coal. When it comes to Europe, North and South America, and Africa then these continents use biomass as their primary fuel as their is no abundance of natural resources as the Asia has so they use biomass to run their plants and produce the energy to fulfill their requirements. When you consider all over the world then Biomass plants has far more higher percentage then any other category of power plants at second their is Gas power plant and so on.