Assignment 2

NK Ezeobele

2024-06-16

The dataset used is the “Global Power Plant Database” from the World Resources Institute. This dataset contains information about power plants around the world, including their location, type, capacity, and more. It is collected from various sources, including government reports, company disclosures, and other databases. I will be using this dataset to answer questions related to energy production, distribution, and the impact of power plants on the environment.

This dataset is saved in a CSV (Comma-Separated Values) file, which is a type of flat file. The data is delimited by commas, and CSV files can be opened using spreadsheet programs like Microsoft Excel or data analysis software such as R.

# Number of rows and columns  
num\_rows <- nrow(cleaned\_data)  
num\_columns <- ncol(cleaned\_data)  
  
# Inline R code to display in Markdown  
cat("This dataframe has", num\_rows, "rows and", num\_columns, "columns.")

## This dataframe has 34936 rows and 36 columns.

# Create a table with column names and descriptions  
library(knitr)  
  
# Specific descriptions for some columns  
specific\_descriptions <- c(  
 "The ISO 3166-1 alpha-3 country code",  
 "The full name of the country",  
 "The name of the power plant",  
 "The Global Power Plant Database identifier number",  
 "The installed capacity of the power plant in megawatts (MW)",  
 "The latitude coordinate of the power plant's location",  
 "The longitude coordinate of the power plant's location",  
 "The primary fuel used by the power plant",  
 "The secondary fuel used by the power plant (if any)",  
 "The tertiary fuel used by the power plant (if any)",  
 "The quaternary fuel used by the power plant (if any)",  
 "The year the power plant was commissioned",  
 "The owner of the power plant",  
 "The source of the power plant data",  
 "The URL to the source of the power plant data",  
 "The source of the geolocation data",  
 "The World Electric Power Plants database identifier",  
 "The year of the capacity data",  
 "The power generation in GWh for the year 2013",  
 "The power generation in GWh for the year 2014",  
 "The power generation in GWh for the year 2015",  
 "The power generation in GWh for the year 2016",  
 "The power generation in GWh for the year 2017",  
 "The power generation in GWh for the year 2018",  
 "The power generation in GWh for the year 2019",  
 "The source of the generation data",  
 "The estimated power generation in GWh for the year 2013",  
 "The estimated power generation in GWh for the year 2014",  
 "The estimated power generation in GWh for the year 2015",  
 "The estimated power generation in GWh for the year 2016",  
 "The estimated power generation in GWh for the year 2017",  
 "Notes on the estimated generation data for 2013",  
 "Notes on the estimated generation data for 2014",  
 "Notes on the estimated generation data for 2015",  
 "Notes on the estimated generation data for 2016",  
 "Notes on the estimated generation data for 2017"  
)  
  
# Generate descriptions for remaining columns  
if (length(specific\_descriptions) < num\_columns) {  
 descriptions <- c(  
 specific\_descriptions,  
 rep("Additional column with relevant information", num\_columns - length(specific\_descriptions))  
 )  
} else {  
 descriptions <- specific\_descriptions  
}  
  
# Create a dataframe for the column information  
column\_info <- data.frame(  
 Column\_Name = colnames(cleaned\_data),  
 Description = descriptions  
)  
  
kable(column\_info, col.names = c("Column Name", "Description"))

| Column Name | Description |
| --- | --- |
| country | The ISO 3166-1 alpha-3 country code |
| country\_long | The full name of the country |
| Plant\_Name | The name of the power plant |
| Plant\_ID | The Global Power Plant Database identifier number |
| Plant\_Capacity | The installed capacity of the power plant in megawatts (MW) |
| latitude | The latitude coordinate of the power plant’s location |
| longitude | The longitude coordinate of the power plant’s location |
| Primary\_Fuel | The primary fuel used by the power plant |
| other\_fuel1 | The secondary fuel used by the power plant (if any) |
| other\_fuel2 | The tertiary fuel used by the power plant (if any) |
| other\_fuel3 | The quaternary fuel used by the power plant (if any) |
| commissioning\_year | The year the power plant was commissioned |
| owner | The owner of the power plant |
| source | The source of the power plant data |
| url | The URL to the source of the power plant data |
| geolocation\_source | The source of the geolocation data |
| wepp\_id | The World Electric Power Plants database identifier |
| year\_of\_capacity\_data | The year of the capacity data |
| generation\_gwh\_2013 | The power generation in GWh for the year 2013 |
| generation\_gwh\_2014 | The power generation in GWh for the year 2014 |
| generation\_gwh\_2015 | The power generation in GWh for the year 2015 |
| generation\_gwh\_2016 | The power generation in GWh for the year 2016 |
| generation\_gwh\_2017 | The power generation in GWh for the year 2017 |
| generation\_gwh\_2018 | The power generation in GWh for the year 2018 |
| generation\_gwh\_2019 | The power generation in GWh for the year 2019 |
| generation\_data\_source | The source of the generation data |
| estimated\_generation\_gwh\_2013 | The estimated power generation in GWh for the year 2013 |
| estimated\_generation\_gwh\_2014 | The estimated power generation in GWh for the year 2014 |
| estimated\_generation\_gwh\_2015 | The estimated power generation in GWh for the year 2015 |
| estimated\_generation\_gwh\_2016 | The estimated power generation in GWh for the year 2016 |
| estimated\_generation\_gwh\_2017 | The estimated power generation in GWh for the year 2017 |
| estimated\_generation\_note\_2013 | Notes on the estimated generation data for 2013 |
| estimated\_generation\_note\_2014 | Notes on the estimated generation data for 2014 |
| estimated\_generation\_note\_2015 | Notes on the estimated generation data for 2015 |
| estimated\_generation\_note\_2016 | Notes on the estimated generation data for 2016 |
| estimated\_generation\_note\_2017 | Notes on the estimated generation data for 2017 |

knitr::opts\_chunk$set(echo = TRUE)  
# Select three columns for summary statistics  
selected\_columns <- cleaned\_data %>%  
 select(Plant\_Capacity, Primary\_Fuel, Plant\_ID) # Example columns  
  
# Summary statistics  
summary\_stats <- selected\_columns %>%  
 summarise(  
 Min\_Capacity = min(Plant\_Capacity, na.rm = TRUE),  
 Max\_Capacity = max(Plant\_Capacity, na.rm = TRUE),  
 Mean\_Capacity = mean(Plant\_Capacity, na.rm = TRUE),  
 Missing\_Capacity = sum(is.na(Plant\_Capacity)),  
 Unique\_Fuels = n\_distinct(Primary\_Fuel),  
 Missing\_Fuels = sum(is.na(Primary\_Fuel)),  
 Total\_Plants = n\_distinct(Plant\_ID)  
 )  
  
# Display the summary statistics  
summary\_stats

## # A tibble: 1 × 7  
## Min\_Capacity Max\_Capacity Mean\_Capacity Missing\_Capacity Unique\_Fuels  
## <dbl> <dbl> <dbl> <int> <int>  
## 1 1 22500 163. 0 15  
## # ℹ 2 more variables: Missing\_Fuels <int>, Total\_Plants <int>