Energy Dataset CLeaning

Anurag Thakur

Required packages

```
#install.packages("readr")
library(readr)
#install the tidyr package
#install.packages("tidyr")
#load the tidyr package
 library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(dplyr)
library(stringr)
```

Executive Summary

Data preprocessing is a must have tool in a data scientist's toolbox and I selected these data sets due to my interest in transcition from non-renewable to renewable energy for energy production.

Data

The datasets chosen were 2 data from kaggle. 1st Dataset :- This dataset contains energy statitics of production, trade, conversion and final consumption of different sources of energy. Source :- https://www.kaggle.com/unitednations/international-energy-statistics published by the United Nations Statistics Division

2nd Datasets :- This dataset contains observations renewable energy production from enviornment friendly sources. Source : https://www.kaggle.com/khadeejahalghadeer/renewable-energy-generation-world-1965-to-2018 published by http://www.bp.com/statisticalreview

```r
# This is the R chunk for the Data Section
Data1 <- read_csv("all_energy_statistics.csv")
```

```
## Rows: 1189482 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): country_or_area, commodity_transaction, unit, category
## dbl (3): year, quantity, quantity_footnotes

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
Data2  <- read_csv("modern-renewable-energy-consumption.csv")
```

```
## Rows: 5091 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Entity, Code
## dbl (5): Year, Hydropower (terawatt-hours), Solar (terawatt-hours), Wind (te...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
Data1.1 <- Data1 %>% select( - (2))
Data1.1 <- Data1.1 %>% select(- (5:6))
colnames(Data1.1)[1] = "Country"
colnames(Data1.1)[2] = "Year"
```

Explanation of relevant variables:- DATASET 1 country_or_area :- country, year = timeline , unit = unit of consuption, quantity = consumption

DATASET 2 Entity = country, year = timeline, Hydropower / Solar /Wind /Other renewables - profuction in terawatt ## Understand

Summarising the types of variables and data structures,

Checking the variable types of 1st DATASET

```r
print("Structure of 1st Dataset")
```

```
## [1] "Structure of 1st Dataset"
```

```r
str(Data1)
```

```
## spec_tbl_df [1,189,482 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ country_or_area      : chr [1:1189482] "Austria" "Austria" "Belgium" "Belgium" ...
##  $ commodity_transaction: chr [1:1189482] "Additives and Oxygenates - Exports" "Additives and Oxygena
##  $ year                 : num [1:1189482] 1996 1995 2014 2013 2012 ...
##  $ unit                 : chr [1:1189482] "Metric tons,  thousand" "Metric tons,  thousand" "Metric t
##  $ quantity             : num [1:1189482] 5 17 0 0 35 25 22 45 1 7 ...
##  $ quantity_footnotes   : num [1:1189482] NA NA NA NA NA NA NA NA NA NA ...
##  $ category             : chr [1:1189482] "additives_and_oxygenates" "additives_and_oxygenates" "addi
##  - attr(*, "spec")=
##   .. cols(
##   ..   country_or_area = col_character(),
##   ..   commodity_transaction = col_character(),
##   ..   year = col_double(),
##   ..   unit = col_character(),
##   ..   quantity = col_double(),
##   ..   quantity_footnotes = col_double(),
##   ..   category = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
class(Data1$country_or_area)
```

```
## [1] "character"
```

```
class(Data1$year)
```

```
## [1] "numeric"
```

```
class(Data1$quantity)
```

```
## [1] "numeric"
```

```
class(Data1$unit)
```

```
## [1] "character"
```

```
print("no change required")
```

```
## [1] "no change required"
```

Checking the variable types of 2st DATASET

```
print("Structure of 2nd Dataset")
```

```
## [1] "Structure of 2nd Dataset"
```

```
str(Data2)
```

```
## spec_tbl_df [5,091 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Entity                           : chr [1:5091] "Africa" "Africa" "Africa" "Africa" ...
##  $ Code                             : chr [1:5091] NA NA NA NA ...
##  $ Year                             : num [1:5091] 1965 1966 1967 1968 1969 ...
##  $ Hydropower (terawatt-hours)      : num [1:5091] 14.3 15.6 16.2 18.6 21.6 ...
##  $ Solar (terawatt-hours)           : num [1:5091] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Wind (terawatt-hours)            : num [1:5091] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Other renewables (terawatt-hours): num [1:5091] 0 0 0 0 0 0 0.164 0.165 0.17 0.175 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Entity = col_character(),
##   ..   Code = col_character(),
##   ..   Year = col_double(),
##   ..   `Hydropower (terawatt-hours)` = col_double(),
##   ..   `Solar (terawatt-hours)` = col_double(),
##   ..   `Wind (terawatt-hours)` = col_double(),
##   ..   `Other renewables (terawatt-hours)` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
print("no change required")
```

```
## [1] "no change required"
```

```
class(Data2$Entity)
```

```
## [1] "character"
```

```
class(Data2$Year)
```

```
## [1] "numeric"
```

```
class(Data2$`Hydropower (terawatt-hours)`)
```

```
## [1] "numeric"
```

```
class(Data2$`Solar (terawatt-hours)`)
```

```
## [1] "numeric"
```

```
class(Data2$`Wind (terawatt-hours)`)
```

```
## [1] "numeric"
```

```
class(Data2$`Other renewables (terawatt-hours)`)
```

```
## [1] "numeric"
```

```
print("no change required")
```

```
## [1] "no change required"
```

Tidy & Manipulate Data I

Removed the unnecessary columns comodity transaction(column 2) and quantity foot notes(column 6) and catagory(column 7) Renamed the columns

```
#removing unnecessary columns data1

Data1.1 <- Data1 %>% select( - (2))
Data1.1 <- Data1.1 %>% select(- (5:6))
colnames(Data1.1)[1] = "Country"
colnames(Data1.1)[2] = "Year"
head(Data1.1,10)
```

```
## # A tibble: 10 x 4
##    Country  Year unit                    quantity
##    <chr>   <dbl> <chr>                      <dbl>
##  1 Austria  1996 Metric tons,  thousand        5
##  2 Austria  1995 Metric tons,  thousand       17
##  3 Belgium  2014 Metric tons,  thousand        0
##  4 Belgium  2013 Metric tons,  thousand        0
##  5 Belgium  2012 Metric tons,  thousand       35
##  6 Belgium  2011 Metric tons,  thousand       25
##  7 Belgium  2010 Metric tons,  thousand       22
##  8 Belgium  2009 Metric tons,  thousand       45
##  9 Czechia  1998 Metric tons,  thousand        1
## 10 Czechia  1995 Metric tons,  thousand        7
```

removing unnecessary column - code (column 2)

```
#removing unnecessary columns data2
Data2.2 <-  Data2 %>% select( - (2))

colnames(Data2.2)[1] = "Country"

head(Data2.2,10)
```

```
## # A tibble: 10 x 6
##    Country  Year `Hydropower (terawatt-hours)` `Solar (terawat~ `Wind (terawatt~
##    <chr>   <dbl>                         <dbl>            <dbl>            <dbl>
##  1 Africa   1965                          14.3                0                0
##  2 Africa   1966                          15.6                0                0
##  3 Africa   1967                          16.2                0                0
##  4 Africa   1968                          18.6                0                0
##  5 Africa   1969                          21.6                0                0
##  6 Africa   1970                          27.1                0                0
##  7 Africa   1971                          25.8                0                0
##  8 Africa   1972                          29.8                0                0
```

```
##  9 Africa   1973                          29.8               0                0
## 10 Africa   1974                          35.1               0                0
## # ... with 1 more variable: Other renewables (terawatt-hours) <dbl>
```

##Merging the datasets Removing White spaces, Filtering data to year 2014, and removing time from 1st data set variable as we have already filtered the data, Triming the down to distinct country valuesin both datasets.

```
Data1.1 <- Data1.1 %>% mutate_if(is.character, str_trim)

DataDistict <-Data1.1 %>% filter(Year == 2014)
DataDistict <-DataDistict %>% distinct(Country, .keep_all = TRUE)
DataDistict <- DataDistict %>% select( -(2))
head(DataDistict,10)
```

```
## # A tibble: 10 x 3
##    Country        unit                   quantity
##    <chr>          <chr>                     <dbl>
##  1 Belgium        Metric tons,  thousand        0
##  2 France         Metric tons,  thousand      119
##  3 Greece         Metric tons,  thousand        2
##  4 Italy          Metric tons,  thousand        4
##  5 Netherlands    Metric tons,  thousand      390
##  6 Romania        Metric tons,  thousand        0
##  7 Serbia         Metric tons,  thousand        4
##  8 Ukraine        Metric tons,  thousand        0
##  9 United Kingdom Metric tons,  thousand       54
## 10 United States  Metric tons,  thousand     2484
```

```
Data2.2 <- Data2.2 %>%  mutate_if(is.character, str_trim)

DataDistict_2 <-Data2.2 %>% filter(Year == 2014)
DataDistict_2<-DataDistict_2 %>% distinct(Country, .keep_all = TRUE)
head(DataDistict_2,10)
```

```
## # A tibble: 10 x 6
##    Country       Year `Hydropower (terawa~ `Solar (terawatt-~ `Wind (terawatt-h~
##    <chr>        <dbl>                <dbl>              <dbl>              <dbl>
##  1 Africa        2014              124.                1.83               5.16
##  2 Algeria       2014                0.193             0.06               0.001
##  3 Argentina     2014               40.9               0.0159             0.619
##  4 Asia Pacific  2014             1511.               62.9              211.
##  5 Australia     2014               14.5               4.95               9.78
##  6 Austria       2014               41.0               0.785              3.85
##  7 Azerbaijan    2014                1.30              0.0029             0.0023
##  8 Bangladesh    2014                0.566             0.168              0.0051
##  9 Belarus       2014                0.121             0.002              0.009
## 10 Belgium       2014                0.292             2.88               4.62
## # ... with 1 more variable: Other renewables (terawatt-hours) <dbl>
```

Merging the data on the basis of column country using right join i.e. prioritizing the observations of dataframe DATA2.1 i.e., table containing observations of produciotn of energy using renewbale resources.

```
Final_data <- right_join(DataDistict,DataDistict_2, key = "Country")
```

```
## Joining, by = "Country"
```

```
head(Final_data,10)
```

```
## # A tibble: 10 x 8
##    Country        unit      quantity  Year `Hydropower (teraw~ `Solar (terawatt~
##    <chr>          <chr>        <dbl> <dbl>               <dbl>             <dbl>
##  1 Belgium        Metric t~        0  2014               0.292              2.88
##  2 France         Metric t~      119  2014              62.8                5.91
##  3 Greece         Metric t~        2  2014               4.48               3.79
##  4 Italy          Metric t~        4  2014              58.5               22.3
##  5 Netherlands    Metric t~      390  2014               0.112              0.785
##  6 Romania        Metric t~        0  2014              18.5                1.30
##  7 Ukraine        Metric t~        0  2014               8.48               0.429
##  8 United Kingdom Metric t~       54  2014               5.89               4.05
##  9 United States  Metric t~     2484  2014             256.                29.2
## 10 Bulgaria       Metric t~       36  2014               4.61               1.25
## # ... with 2 more variables: Wind (terawatt-hours) <dbl>,
## #   Other renewables (terawatt-hours) <dbl>
```

Tidy & Manipulate Data II

Creating 2 new variable from the existing variables (`Hydropower (terawatt-hours)` + `Solar (terawatt-hours)`+`Wind (terawatt-hours)`+`Other renewables (terawatt-hours)`).

Total_Renewable = sum of the energy produced from renewable resources Hydropower_Percentage = contribution of Hydropower to total energy produced by renewable resources

Renaming the column

Removing the variables not required anymore `Hydropower (terawatt-hours)` + `Solar (terawatt-hours)`+`Wind (terawatt-hours)`+`Other renewables (terawatt-hours)`

```
Final_data <- Final_data %>% mutate(
  Total_Renewable = `Hydropower (terawatt-hours)` + `Solar (terawatt-hours)`+`Wind (terawatt-hours)`+`O
  Hydropower_Percentage = `Hydropower (terawatt-hours)`/ Total_Renewable *100)
Final_data$Total_Renewable %>% as.numeric()
```

```
##  [1]   12.2145110   91.8100000   12.1770000  120.6784000   11.7072030
##  [6]   25.0474000   10.1640000   64.5224021  552.5271547    7.3887920
## [11]  416.3056872   29.5867562    9.9370000    3.1505400    1.5101000
## [16]   52.6765998    0.0012000   19.8450000   31.5620810    6.3390000
## [21]   85.7400000   40.6540505   52.6287140    6.1850083   43.8014603
## [26]   46.5411167   26.1964026  162.5251018  137.9256000   17.9909750
## [31]    6.3868358   19.8096540    0.7451534  110.2690836   23.9585000
## [36]    4.4161600   49.9775696   32.8073620   18.1197653    2.7642000
## [41]    0.3976300    0.8251000   14.2090781    0.3082000  132.1623967
## [46]  432.0498197 1277.1584790   11.9535920   25.4295600    8.2773351
## [51]   34.4071216    6.6120000    4.9045793    0.0040000   14.7546768
## [56]  202.0361925   14.5916800    3.5613010   34.4507000    0.2110000
```

```
## [61]    0.3180000    1.3738000    1.0236497    1.3919000    0.0033000
## [66]    0.2540000    2.8925020    0.0422000    0.1213000    0.0031175
## [71]  136.2447679 1928.7302828  222.1750816   31.2042109    9.1696730
## [76]   71.3918579 1191.1627281   32.7682319    0.1050020   14.7490000
## [81]    1.2916660   20.5551000   21.8914972 1021.5094418   53.6912022
## [86]   31.9018208    4.2040911    2.7501455           NA   71.4737434
## [91]    2.5611052  174.2049000  773.5307680   11.8673220    8.2147301
## [96]   78.7534778   61.3149698   18.1554233 5295.2445675
```

```
Final_data <- Final_data %>% select( -(5:8) )
head(Final_data,10)
```

```
## # A tibble: 10 x 6
##    Country        unit         quantity  Year Total_Renewable Hydropower_Percen~
##    <chr>          <chr>           <dbl> <dbl>           <dbl>              <dbl>
##  1 Belgium        Metric tons~        0  2014           12.2               2.39
##  2 France         Metric tons~      119  2014           91.8              68.4
##  3 Greece         Metric tons~        2  2014           12.2              36.8
##  4 Italy          Metric tons~        4  2014          121.               48.5
##  5 Netherlands    Metric tons~      390  2014           11.7               0.956
##  6 Romania        Metric tons~        0  2014           25.0              74.0
##  7 Ukraine        Metric tons~        0  2014           10.2              83.4
##  8 United Kingdom Metric tons~       54  2014           64.5               9.13
##  9 United States  Metric tons~     2484  2014          553.               46.3
## 10 Bulgaria       Metric tons~       36  2014            7.39             62.3
```

Diving the unit column into two

```
Final_data %>% separate(unit, into = c("Unit", " Multiplicant"), sep = ",")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 6 rows [31, 32,
## 33, 34, 62, 63].
```

```
## # A tibble: 99 x 7
##    Country Unit  ` Multiplicant` quantity  Year Total_Renewable Hydropower_Perc~
##    <chr>   <chr> <chr>              <dbl> <dbl>           <dbl>            <dbl>
##  1 Belgium Metr~ "  thousand"          0  2014           12.2              2.39
##  2 France  Metr~ "  thousand"        119  2014           91.8             68.4
##  3 Greece  Metr~ "  thousand"          2  2014           12.2             36.8
##  4 Italy   Metr~ "  thousand"          4  2014          121.              48.5
##  5 Nether~ Metr~ "  thousand"        390  2014           11.7              0.956
##  6 Romania Metr~ "  thousand"          0  2014           25.0             74.0
##  7 Ukraine Metr~ "  thousand"          0  2014           10.2             83.4
##  8 United~ Metr~ "  thousand"         54  2014           64.5              9.13
##  9 United~ Metr~ "  thousand"       2484  2014          553.              46.3
## 10 Bulgar~ Metr~ "  thousand"         36  2014            7.39            62.3
## # ... with 89 more rows
```

Scan I

Scanning the numeric attributes for missing values, special values and obvious errors (i.e. inconsistencies).

```
print("missing values in the entire dataset")
```

```
## [1] "missing values in the entire dataset"
```

```
sum(is.na(Final_data))
```

```
## [1] 60
```

```
print("missing values in the unit column")
```

```
## [1] "missing values in the unit column"
```

```
sum(is.na(Final_data$unit))
```

```
## [1] 29
```

```
print("missing values in the quantity column")
```

```
## [1] "missing values in the quantity column"
```

```
sum(is.na(Final_data$quantity))
```

```
## [1] 29
```

```
print("missing values in the Total_Renewable column")
```

```
## [1] "missing values in the Total_Renewable column"
```

```
sum(is.na(Final_data$Total_Renewable))
```

```
## [1] 1
```

```
print("missing values in the Hydropower_Percentage column")
```

```
## [1] "missing values in the Hydropower_Percentage column"
```

```
sum(is.na(Final_data$Hydropower_Percentage))
```

```
## [1] 1
```

Results for Inconsistency Operations for unit and multiplicant column are the same, so only one column is displayed in unit column NA values cannot be treated in quantity column NA values can be treated as the units are different, operations for conversion are irrelevant to this report treating NA values for Total_Renewable and Hydropower_Percentage by replacing them with mean

```
print("infinite values in the unit column")
```

```
## [1] "infinite values in the unit column"
```

```
sum(is.infinite(Final_data$unit))
```

```
## [1] 0
```

```
print("infinite values in the quantity column")
```

```
## [1] "infinite values in the quantity column"
```

```
sum(is.infinite(Final_data$quantity))
```

```
## [1] 0
```

```
print("infinite values in the Total_Renewable column")
```

```
## [1] "infinite values in the Total_Renewable column"
```

```
sum(is.infinite(Final_data$Total_Renewable))
```

```
## [1] 0
```

```
print("infinite values in the Hydropower_Percentage column")
```

```
## [1] "infinite values in the Hydropower_Percentage column"
```

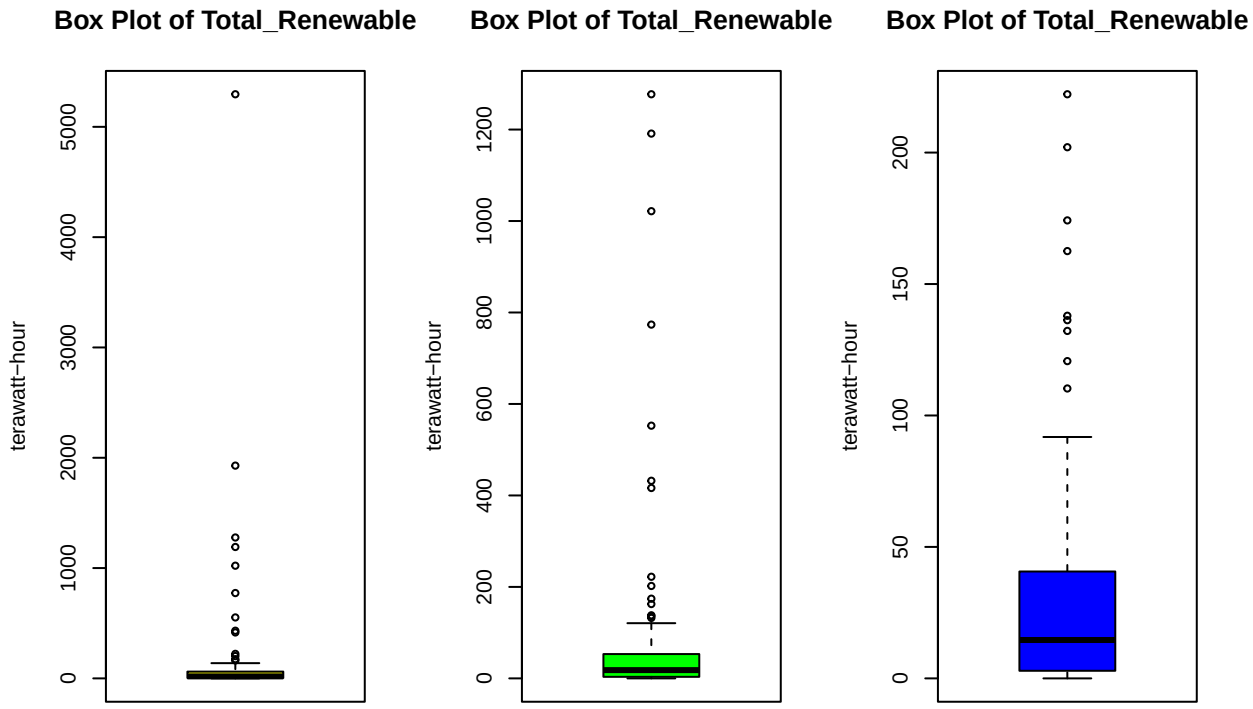
```
sum(is.infinite(Final_data$Hydropower_Percentage))
```

```
## [1] 0
```

Scan II

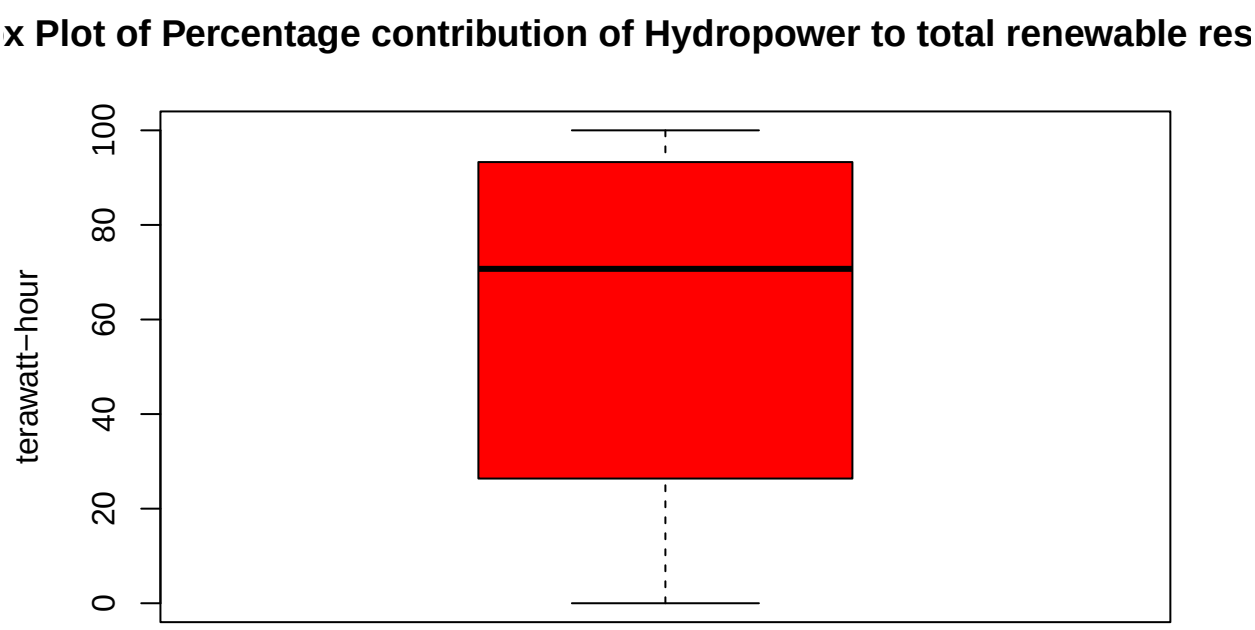
Scanning the numeric data(Total renewable and Hydropower_Percentage) for outliers.

```
par(mfrow=c(1,3))
# This is the R chunk for the Scan II
Final_data$Total_Renewable %>%  boxplot(main="Box Plot of Total_Renewable", ylab="terawatt-hour", col =

Final_data <-  Final_data %>%filter(Final_data$Total_Renewable < 1500)

Final_data$Total_Renewable %>%  boxplot(main="Box Plot of Total_Renewable", ylab="terawatt-hour", col =

Final_data <-  Final_data %>%filter(Final_data$Total_Renewable < 400)

Final_data$Total_Renewable %>%  boxplot(main="Box Plot of Total_Renewable", ylab="terawatt-hour", col =
```



No outliers were found for the total

```
# This is the R chunk for the Scan II
Final_data$Hydropower_Percentage %>%  boxplot(main="Box Plot of Percentage contribution of Hydropower t
```

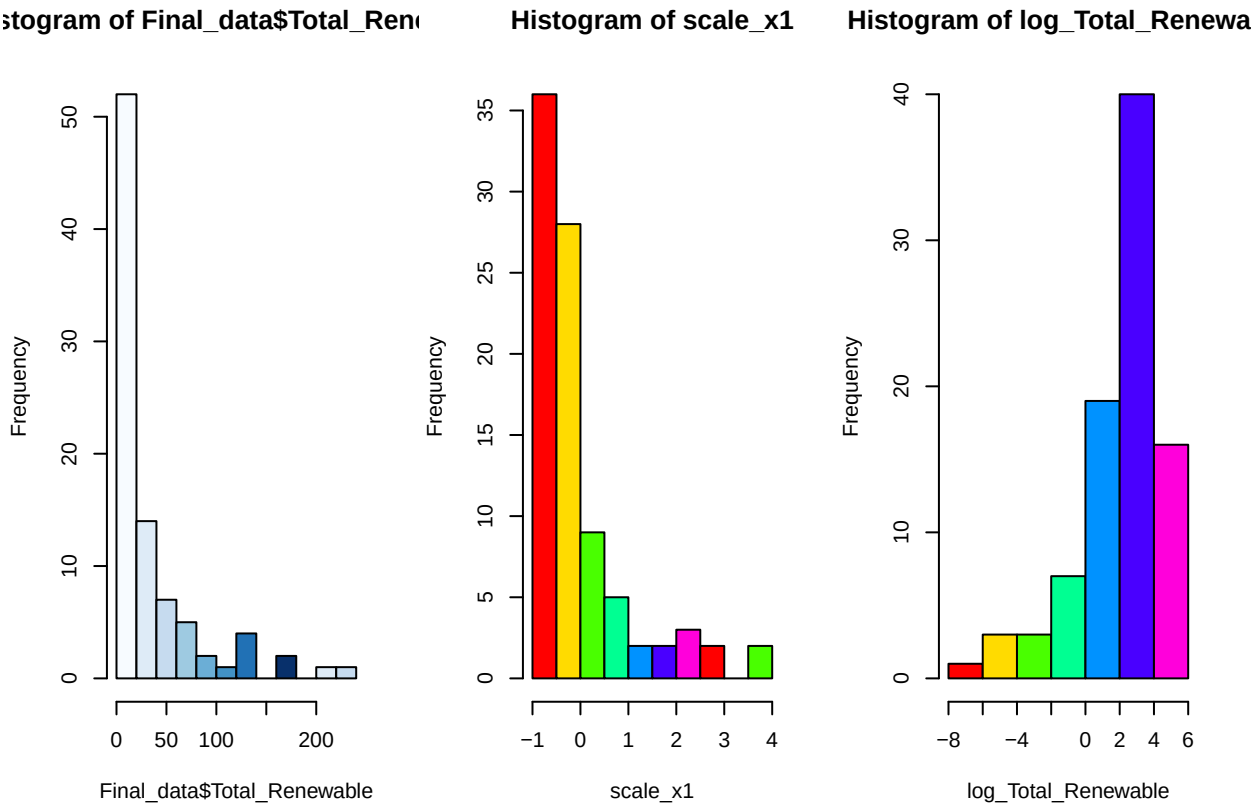


Transform

Transformation task was performed on 2 variables,

for Transformation the variable Total Renewable attribute, logarithmic transmission gave us better results

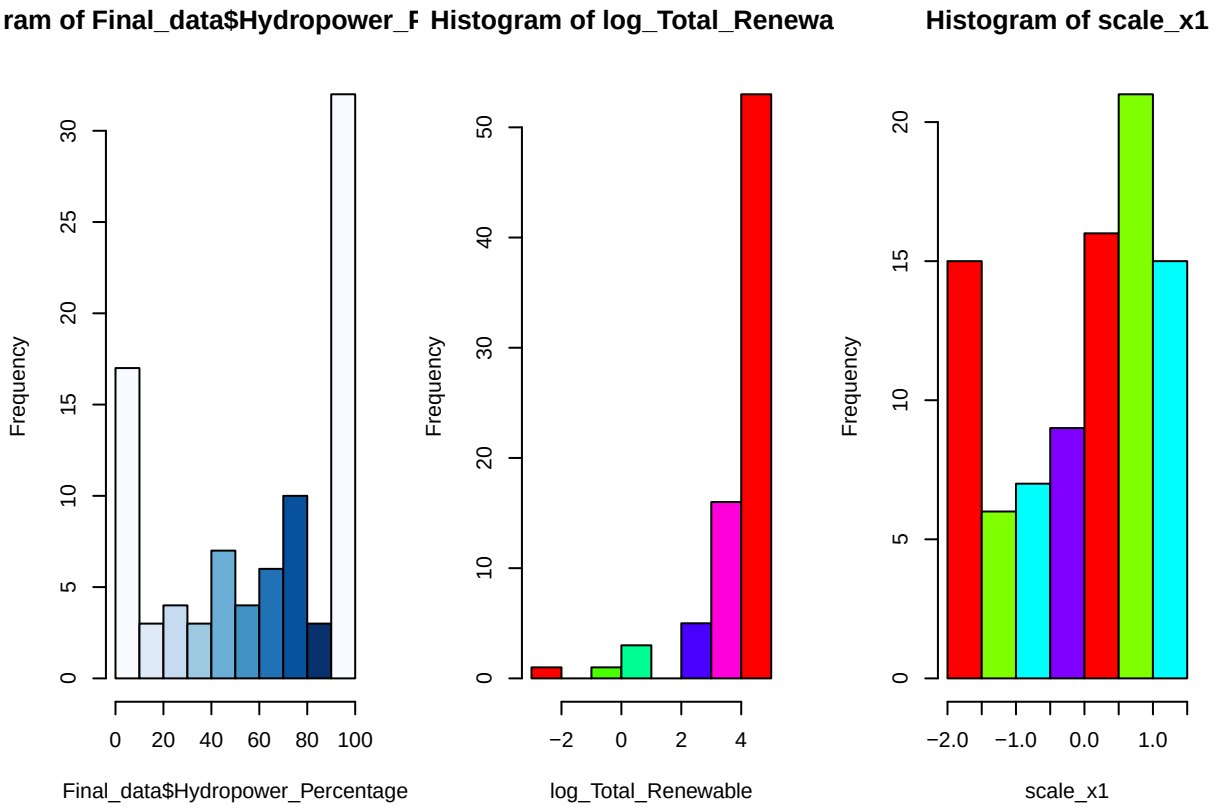
```
par(mfrow=c(1,3))
hist(Final_data$Total_Renewable,col = blues9)
scale_x1 <- scale(Final_data$Total_Renewable, center = TRUE, scale = TRUE)
hist(scale_x1,col = rainbow(7) )
log_Total_Renewable <- log(Final_data$Total_Renewable)
hist(log_Total_Renewable, ,col = rainbow(7))
```




```
#hist(Final_data$quantity)
```

Unlike with the Total Renewable attribute, transformation using logarithms wasn't as effective as scaling.

```
par(mfrow=c(1,3))
hist(Final_data$Hydropower_Percentage, col = blues9)
log_Total_Renewable <- log(Final_data$Hydropower_Percentage)
hist(log_Total_Renewable,col = rainbow(7))
scale_x1 <- scale(Final_data$Hydropower_Percentage, center = TRUE, scale = TRUE)
hist(scale_x1 , col = rainbow(4) )
```




```
#hist(Final_data$quantity)
```

Thank you!