MATH2349 Semester 2, 2019

Assignment 3

Sarthak Sirari (S3766477) | Zhuoming Li (S3815870) | Lei Wang (S3412072)

Required packages

Before we begin with our project, we need to import all the necessary libraries in this part.

```
library(readr)
library(dplyr)
library(tidyr)
library(knitr)
library(forecast)
library(editrules)
library(kableExtra)
```

Executive Summary

- In this pre-processing, we discussed the relations between environmental and socioeconomic sustainability w.r.t. Human Development Index (HDI) ranking of the 189 countries.
- We found two datasets and used inner-join operator to merge them together by the common variable “Country” so that a new data frame is created.
- To be more convenient, we convert the data in “Country” column into factor and all the other columns with values are converted from character to numeric.
- Then we found that the dataframe is untidy, there exists a duplicated column “HDI.rank.x” same as “HDI.rank.y”, only one is needed.
- Then we create 2 factor variables “Red.List.Index.Category” & “Forest.Cover.Change.Category” defined from existing “Red.List.Index” (from Very Low to Very High) and “Forest.Cover.Change” (to see if the forest cover increased, decreased or stayed the same) variables.
- Then in the dataframe, we scan all variables and try to find the columns which contain missing values, then we replace the missing values by the calculated mean values in those columns.
- For some special values, we created new function and apply for the dataframe to find special values (infinity or nan), also apply. Then applied couple of rules over the variables to keep the data in check.
- Then generate a new function called “outlier” that can be applied to the dataframe for the numeric variables.
- Also, generate a function called “cap” and apply it for replacing the exist outlier(s) with the nearest neighbours that is non-outlier.
- After that, plot a histogram based on a numeric column of the dataframe, we now apply Box-Cox transformation with “lamda=auto” so that the transform the skewed data to normal distribution, since the normality assumption is crucial when doing statistical hypothesis test or analysis.

Data

- The first data set “Environmental sustainability” was downloaded from United Nations Development Programme (UNDP) - Human Development Reports (HDR) under the following URL:- http://hdr.undp.org/en/composite/Dashboard4 (http://hdr.undp.org/en/composite/Dashboard4)
- The data set “Environmental sustainability” contains a total of 12 variables that cover environmental sustainability and environmental threats.

- The first two are the common variables - HDI Rank and Country Name.
- The next seven level and change indicators variables on environmental sustainability are energy consumption, carbon dioxide emissions, change in forest area and fresh water withdrawals.
- The next three environmental threats indicators are mortality rate, which is attributed to household and ambient air pollution and another is attributed to unsafe water, sanitation and hygiene service, and the Red List Index deviced by the International Union for Conservation of Nature and Natural resources which is a measure to aggregate extinction risk across different species.
- The data is available for 189 countries in total.

- The second data set “Socioeconomic sustainability” was downloaded from United Nations Development Programme (UNDP) - Human Development Reports (HDR) under the following URL:- http://hdr.undp.org/en/composite/Dashboard5 (http://hdr.undp.org/en/composite/Dashboard5)
- The data set “Socioeconomic sustainability” contains a total of 13 variables for the economic and social sustainability.
- The first two are the common variables - HDI Rank and Country Name.
- The next six economic sustainability indicators are adjusted net savings, total debt service, gross capital formation, skilled labour force, diversity of exports and expenditure on research and development.
- The next four social sustainability indicators are the ratio of education and health expenditure to military expenditure, change in overall loss in HDI value due to inequality, and changes in gender and income inequality.
- The data is available for 189 countries in total.

- First, we set up a working directory to conveniently read the files
- Read the csv file and store the dataset in “environmental_sustainability” and convert factor to character then view the dataset
- Read the csv file and store the dataset in “socioeconomic_sustainability” and convert factor to character then view the dataset
- Use inner join to combine two dataset with a common column, the variable “country” and print out the new dataset (view the table)

```
# Set the working directory
setwd("C:\\Users\\abhis\\OneDrive\\Desktop\\Master of Data Science\\Sem 2\\Data Preprocessing
MATH2349\\Assignments\\Assignment 3")

# Check the working directory
getwd()
```

```
## [1] "C:/Users/abhis/OneDrive/Desktop/Master of Data Science/Sem 2/Data Preprocessing MATH2
349/Assignments/Assignment 3"
```

```
# Read the csv file and store the dataset in "environmental_sustainability" and set stringsAs
Factors as FALSE
environmental_sustainability <- read.csv("Environmental_Sustainability.csv", stringsAsFactors
= FALSE)

# Check the data set
environmental_sustainability
```

HDI.rank	Country	Fossil.fuel.energy.consumption
<int>	<chr>	<chr>
1	Norway	58.5

HDI.rank	Country	Fossil.fuel.energy.consumption
<int>	<chr>	<chr>
2	Switzerland	50.1
3	Australia	93.4
4	Ireland	85.4
5	Germany	79.8
6	Iceland	11.5
7	Hong Kong, China (SAR)	93.2
7	Sweden	26.8
9	Singapore	97.5
10	Netherlands	91.4

1-10 of 189 rows | 1-3 of 12 columns Previous **1** 2 3 4 5 6 … 19 Next

```
#read the csv file and store the dataset in "socioeconomic_sustainability" and set stringsAsF
actors as FALSE
socioeconomic_sustainability <- read.csv("Socioeconomic_Sustainability.csv", stringsAsFactors
= FALSE)

# Check the data set
socioeconomic_sustainability
```

HDI.rank	Country	Adjusted.net.savings	Total.de
<int>	<chr>	<chr>	<chr>
1	Norway	15.8	..
2	Switzerland	16.4	..
3	Australia	5.3	..
4	Ireland	20.6	..
5	Germany	13.6	..
6	Iceland	21.8	..
7	Hong Kong, China (SAR)	..	..
7	Sweden	20.2	..
9	Singapore	32.7	..
10	Netherlands	16.5	..

1-10 of 189 rows | 1-4 of 13 columns Previous **1** 2 3 4 5 6 … 19 Next

```
# Use inner-join to combine 2 dataset with a common column, the varibale "country"
environmental_and_socioeconomic_sustainability <- inner_join(environmental_sustainability, so
cioeconomic_sustainability, by="Country")

# Check the data set
environmental_and_socioeconomic_sustainability
```

HDI.rank.x <int>	Country <chr>	Fossil.fuel.energy.consumption <chr>
1	Norway	58.5
2	Switzerland	50.1
3	Australia	93.4
4	Ireland	85.4
5	Germany	79.8
6	Iceland	11.5
7	Hong Kong, China (SAR)	93.2
7	Sweden	26.8
9	Singapore	97.5
10	Netherlands	91.4

1-10 of 189 rows | 1-3 of 24 columns Previous **1** 2 3 4 5 6 … 19 Next

Understand

- Use str() to check structure of the data set to view each column names, values and variable type in the new dataframe after using inner join
- Convert the data type in column “Country” from character to factor
- Convert the data type in numeric columns from character to numeric
- Repeat this procedure for other columns contain numbers, convert them from character to numeric
- Finally verify whether all the conversion are successful

```
# Check structure of the data set
str(environmental_and_socioeconomic_sustainability)
```

```
## 'data.frame':    189 obs. of  24 variables:
##  $ HDI.rank.x                                                                : int  1 2 3
4 5 6 7 7 9 10 ...
##  $ Country                                                                   : chr  "Norwa
y" "Switzerland" "Australia" "Ireland" ...
##  $ Fossil.fuel.energy.consumption                                            : chr  "58.5"
"50.1" "93.4" "85.4" ...
##  $ Renewable.energy.consumption                                              : num  57.8 2
5.3 9.2 9.1 14.2 77 0.9 53.2 0.7 5.9 ...
##  $ Carbon.dioxide.emissions.Per.Capita                                       : num  9.3 4.
3 15.4 7.3 8.9 6.1 6.4 4.5 10.3 9.9 ...
##  $ Carbon.dioxide.emissions.KG.per.PPP.GDP                                   : chr  "0.15"
"0.08" "0.35" "0.15" ...
##  $ Forest.Cover                                                              : chr  "33.2"
"31.7" "16.2" "10.9" ...
##  $ Forest.Cover.Change                                                       : chr  "-0.2"
"9" "-2.9" "62.2" ...
##  $ Fresh.water.withdrawals                                                   : chr  "0.8"
"3.7" "3.4" "1.5" ...
##  $ Mortality.Rate.attributed.to.Household.and.ambient.air.pollution          : chr  "8.6"
"10.1" "8.4" "11.9" ...
##  $ Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services: chr  "0.2"
"0.1" "0.1" "0.1" ...
##  $ Red.List.Index                                                            : num  0.943
0.982 0.828 0.917 0.983 0.872 0.823 0.992 0.862 0.943 ...
##  $ HDI.rank.y                                                                : int  1 2 3
4 5 6 7 7 9 10 ...
##  $ Adjusted.net.savings                                                      : chr  "15.8"
"16.4" "5.3" "20.6" ...
##  $ Total.debt.service                                                        : chr  ".."
".." ".." ".." ...
##  $ Gross.capital.formation                                                   : chr  "28.8"
"23.3" "24.2" "24.3" ...
##  $ Skilled.labour.force                                                      : chr  "82.4"
"85.7" "78.3" "82.8" ...
##  $ Concentration.index.exports                                               : chr  "0.31
5" "0.288" "0.244" "0.242" ...
##  $ Research.and.development.expenditure                                      : chr  "1.9"
"3.0" "2.2" "1.5" ...
##  $ Education.and.health.expenditure.versus.military.expenditure              : chr  "1.7"
"0.7" "2.0" "0.3" ...
##  $ Ratio.of.education.and.health.expenditure.to.military.expenditure         : chr  "" "2
5.6" "8.0" "32.0" ...
##  $ Overall.loss.in.HDI.value.due.to.inequality.Change                        : chr  "3.0"
"1.5" "0.6" "-0.3" ...
##  $ Gender.Inequality.Index.Change                                            : chr  "-3.6"
"-3.9" "-1.8" "-3.6" ...
##  $ Income.quintile.ratio.Change                                              : chr  "-0.7"
".." "0.4" "-0.6" ...
```

```
# Covert the character variable to factor variable
environmental_and_socioeconomic_sustainability$Country <- factor(environmental_and_socioeconomic_sustainability$Country, ordered = FALSE)

# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Fossil.fuel.energy.consumption <- as.numeric(environmental_and_socioeconomic_sustainability$Fossil.fuel.energy.consumption)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Carbon.dioxide.emissions.KG.per.PPP.GDP <- as.numeric(environmental_and_socioeconomic_sustainability$Carbon.dioxide.emissions.KG.per.PPP.GDP)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Forest.Cover <- as.numeric(environmental_and_socioeconomic_sustainability$Forest.Cover)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Forest.Cover.Change <- as.numeric(environmental_and_socioeconomic_sustainability$Forest.Cover.Change)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Fresh.water.withdrawals <- as.numeric(environmental_and_socioeconomic_sustainability$Fresh.water.withdrawals)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Mortality.Rate.attributed.to.Household.and.ambient.air.pollution <- as.numeric(environmental_and_socioeconomic_sustainability$Mortality.Rate.attributed.to.Household.and.ambient.air.pollution)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services <- as.numeric(environmental_and_socioeconomic_sustainability$Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Adjusted.net.savings <- as.numeric(environment
al_and_socioeconomic_sustainability$Adjusted.net.savings)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Total.debt.service <- as.numeric(environmental
_and_socioeconomic_sustainability$Total.debt.service)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Gross.capital.formation <- as.numeric(environm
ental_and_socioeconomic_sustainability$Gross.capital.formation)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Skilled.labour.force <- as.numeric(environment
al_and_socioeconomic_sustainability$Skilled.labour.force)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Concentration.index.exports <- as.numeric(envi
ronmental_and_socioeconomic_sustainability$Concentration.index.exports)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Research.and.development.expenditure <- as.num
eric(environmental_and_socioeconomic_sustainability$Research.and.development.expenditure)
```

```
## Warning: NAs introduced by coercion
```

```
environmental_and_socioeconomic_sustainability$Education.and.health.expenditure.versus.milita
ry.expenditure <- as.numeric(environmental_and_socioeconomic_sustainability$Education.and.hea
lth.expenditure.versus.military.expenditure)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Ratio.of.education.and.health.expenditure.to.military.expenditure <- as.numeric(environmental_and_socioeconomic_sustainability$Ratio.of.education.and.health.expenditure.to.military.expenditure)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Overall.loss.in.HDI.value.due.to.inequality.Change <- as.numeric(environmental_and_socioeconomic_sustainability$Overall.loss.in.HDI.value.due.to.inequality.Change)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Gender.Inequality.Index.Change <- as.numeric(environmental_and_socioeconomic_sustainability$Gender.Inequality.Index.Change)
```

```
## Warning: NAs introduced by coercion
```

```
# Convert the type from character to numeric
environmental_and_socioeconomic_sustainability$Income.quintile.ratio.Change <- as.numeric(environmental_and_socioeconomic_sustainability$Income.quintile.ratio.Change)#All above are convert to numeric variable
```

```
## Warning: NAs introduced by coercion
```

```
# Verify all the data type conversion
str(environmental_and_socioeconomic_sustainability)
```

```
## 'data.frame':    189 obs. of  24 variables:
##  $ HDI.rank.x                                                          : int  1 2 3
4 5 6 7 7 9 10 ...
##  $ Country                                                             : Factor w/ 1
89 levels "Afghanistan",..: 127 164 9 82 65 77 75 163 153 122 ...
##  $ Fossil.fuel.energy.consumption                                      : num  58.5 5
0.1 93.4 85.4 79.8 11.5 93.2 26.8 97.5 91.4 ...
##  $ Renewable.energy.consumption                                        : num  57.8 2
5.3 9.2 9.1 14.2 77 0.9 53.2 0.7 5.9 ...
##  $ Carbon.dioxide.emissions.Per.Capita                                 : num  9.3 4.
3 15.4 7.3 8.9 6.1 6.4 4.5 10.3 9.9 ...
##  $ Carbon.dioxide.emissions.KG.per.PPP.GDP                             : num  0.15
0.08 0.35 0.15 0.2 0.15 0.12 0.1 0.13 0.22 ...
##  $ Forest.Cover                                                        : num  33.2 3
1.7 16.2 10.9 32.7 0.5 NA 68.9 23.1 11.2 ...
##  $ Forest.Cover.Change                                                 : num  -0.2 9
-2.9 62.2 1.1 ...
##  $ Fresh.water.withdrawals                                             : num  0.8 3.
7 3.4 1.5 21.4 0.2 NA 1.5 NA 11.8 ...
##  $ Mortality.Rate.attributed.to.Household.and.ambient.air.pollution    : num  8.6 1
0.1 8.4 11.9 16 8.7 NA 7.2 25.9 13.7 ...
##  $ Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services: num  0.2 0.
1 0.1 0.1 0.6 0.1 NA 0.2 0.1 0.2 ...
##  $ Red.List.Index                                                      : num  0.943
0.982 0.828 0.917 0.983 0.872 0.823 0.992 0.862 0.943 ...
##  $ HDI.rank.y                                                          : int  1 2 3
4 5 6 7 7 9 10 ...
##  $ Adjusted.net.savings                                                : num  15.8 1
6.4 5.3 20.6 13.6 21.8 NA 20.2 32.7 16.5 ...
##  $ Total.debt.service                                                  : num  NA NA
NA NA NA NA NA NA NA NA ...
##  $ Gross.capital.formation                                             : num  28.8 2
3.3 24.2 24.3 19.8 22.2 22.3 25.7 27.6 20.2 ...
##  $ Skilled.labour.force                                                : num  82.4 8
5.7 78.3 82.8 86.5 73.4 76.9 84.9 81.7 77.4 ...
##  $ Concentration.index.exports                                         : num  0.315
0.288 0.244 0.242 0.106 0.441 0.268 0.091 0.24 0.072 ...
##  $ Research.and.development.expenditure                                : num  1.9 3
2.2 1.5 2.9 2.2 0.8 3.3 2.2 2 ...
##  $ Education.and.health.expenditure.versus.military.expenditure        : num  1.7 0.
7 2 0.3 1.2 0 NA 1 3.2 1.2 ...
##  $ Ratio.of.education.and.health.expenditure.to.military.expenditure   : num  NA 25.
6 8 32 13.5 NA NA 16.5 2.2 14 ...
##  $ Overall.loss.in.HDI.value.due.to.inequality.Change                  : num  3 1.5
0.6 -0.3 0.1 -1.3 NA 1.2 NA -0.3 ...
##  $ Gender.Inequality.Index.Change                                      : num  -3.6 -
3.9 -1.8 -3.6 -3.2 -4.2 NA -1.4 -5 -3.9 ...
##  $ Income.quintile.ratio.Change                                        : num  -0.7 N
A 0.4 -0.6 1 -1.2 NA 1.1 NA 0 ...
```

Tidy & Manipulate Data I

- In this step, we check three interrelated rules which make a dataset tidy (introduced by Hadley Wickham and Grolemund in 2016) over our dataset as following: -
    - Each variable must have its own column
    - Each observation must have its own row

    - Each value must have its own cell
- When we checked the rules over our dataset, we found that our data is already in tidy form.
- Since our data set follows all the tidy rules, so there is no need to apply any transformation to make the data tidy.
- However, we found column “HDI.rank” is reapted twice and stored as “HDI.rank.x” and “HDI.rank.y” in the dataframe, so remove one “HDI.rank.y” and renamed the column “HDI.rank.x” as “HDI.rank”
- View the updated dataframe

```
# Remove the "HDI.rank.y" this duplicate column
environmental_and_socioeconomic_sustainability <- select(environmental_and_socioeconomic_sust
ainability, -HDI.rank.y)

# Rename the "HDI.rank.x" as "HDI.rank"
colnames(environmental_and_socioeconomic_sustainability)[colnames(environmental_and_socioecon
omic_sustainability)=="HDI.rank.x"] <- "HDI.rank"

# Verify the updated data set
environmental_and_socioeconomic_sustainability
```

HDI.rank	Country	Fossil.fuel.energy.consump
<int>	<fctr>	<
1	Norway	
2	Switzerland	
3	Australia	
4	Ireland	
5	Germany	
6	Iceland	
7	Hong Kong, China (SAR)	
7	Sweden	
9	Singapore	
10	Netherlands	

1-10 of 189 rows | 1-3 of 23 columns Previous **1** 2 3 4 5 6 … 19 Next

Tidy & Manipulate Data II

- In this step, we added a new column “Red.List.Index.Category” with categorical variables (very low, low, medium, high very high) to define the index values in column “Red.List.Index”
- We used ifelse() function to determine the range of index for each level, for example, [0.8,1.0] is “very high” and between [0.5, 0.599] is “low”, all other less than 0.5 is “very low”
- Check the class of the column “Red.List.Index.Category” is character in data frame
- Rearrange the information in column “Red.List.Index.Category” to follow the order level “Very Low”, “Low”, “Medium”, “High”, “Very High” and also convert to factor variables
- Verify the variables in this column are factors
- Check and show the levels in order is “Very Low”, “Low”, “Medium”, “High”, “Very High”

- Similarly, we added a new column “Forest.Cover.Change.Category” with categorical variables (Incr, Decr, Same) to define the values of change in column “Forest.Cover.Change”
- Use ifelse() function to determine that if the values of changes are “>0”, call “Incr”, if the changes are “<0”, call “Decr”; otherwise (“=0”, No difference), call “Same”.
- Check the class of the column “Forest.Cover.Change.Category” is character in data frame.
- Convert the character variables in column “Forest.Cover.Change.Category” to factor variables, and we don not need to order them, because no ordered for “increase, decrease, same”.
- Verify the variables in column “Forest.Cover.Change.Category” are factors
- Check and show the levels as “Decr, Same, Incr”
- Print and view the updated dataframe

Red.List.Index	Red.List.Index.Category
1.000 - 0.800	Very High
0.799 - 0.700	High
0.699 - 0.600	Medium
0.599 - 0.500	Low
0.499 - 0.400	Very Low

```
# Create new variable using mutate() function to have categories for Red List Index
environmental_and_socioeconomic_sustainability <- mutate(environmental_and_socioeconomic_sust
ainability, Red.List.Index.Category = ifelse(Red.List.Index<=1.000 & Red.List.Index>=0.800,"V
ery High", ifelse(Red.List.Index<=0.799 & Red.List.Index>=0.700,"High", ifelse(Red.List.Index
<=0.699 & Red.List.Index>=0.600,"Medium", ifelse(Red.List.Index<=0.599 & Red.List.Index>=0.50
0,"Low","Very Low")))))

# Check the class of the new variable
environmental_and_socioeconomic_sustainability$Red.List.Index.Category %>% class()
```

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

```
# Convert the variable to ordered factor
environmental_and_socioeconomic_sustainability$Red.List.Index.Category  <- factor(environment
al_and_socioeconomic_sustainability$Red.List.Index.Category , labels=c("Very Low", "Low", "Me
dium", "High", "Very High"), ordered=TRUE)

# Verify the column in convereted to factor
environmental_and_socioeconomic_sustainability$Red.List.Index.Category %>% is.factor()
```

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

```
# Show the levels in order, very low - very high
environmental_and_socioeconomic_sustainability$Red.List.Index.Category %>% levels()
```

```
## [1] "Very Low"  "Low"       "Medium"    "High"      "Very High"
```

```
# Create new variable using mutate() function to have categories for Forest Cover Change
environmental_and_socioeconomic_sustainability <- mutate(environmental_and_socioeconomic_sust
ainability, Forest.Cover.Change.Category = ifelse(Forest.Cover.Change > 0, "Incr", ifelse(For
est.Cover.Change < 0,"Decr","Same")))

# Check the class of the new variable
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category %>% class()
```

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

```
# Convert the variable to factor
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category  <- factor(enviro
nmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category , labels=c("Decr", "Sam
e", "Incr"), ordered = FALSE)

# Verify the column in convereted to factor
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category %>% is.factor()
```

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

```
# Show the levels of the column
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category %>% levels()
```

```
## [1] "Decr" "Same" "Incr"
```

```
# Verify the data set
environmental_and_socioeconomic_sustainability
```

HDI.rank	Country	Fossil.fuel.energy.consump
<int>	<fctr>	<
1	Norway	
2	Switzerland	
3	Australia	
4	Ireland	
5	Germany	
6	Iceland	
7	Hong Kong, China (SAR)	
7	Sweden	
9	Singapore	
10	Netherlands	

1-10 of 189 rows | 1-3 of 25 columns Previous **1** 2 3 4 5 6 … 19 Next

Scan I

- In this step, first we check the missing values count in each variable
- Now we have found all the missing in columns, then from the dataframe,
    - Use “round” function and “mean(…na.rm=TRUE)” function to calculate the mean(average) of the “weight” exclude all NA(missing value); then round the numbers up to the decimal places in which each column values currently are.
    - Use “ifelse()” function to make conditions for the “weight” column: if the value is missing, replace the “NA” values by the mean of the “weight”, otherwise keep the original values in “weight” column.
- For “Forest.Cover.Change.Category” column, since it is dependent on “Forest.Cover.Change”, we fill in the values again in the “Forest.Cover.Change.Category” column as per the “Forest.Cover.Change” as we have filled in the missing values in “Forest.Cover.Change” column by repeating the steps performed in the previous task

- Now we created a function isSpecial() to check for special values, i.e. infinite (Inf and -Inf) and NaN.
- Apply the function to each column to check for special values
- No specials values were found, so we proceed.

- Then we created two rules as following:-
    - Check if the HDI rank from 1 to 189, as only 189 countries are available in the data set
    - Check if the Red List Index value is in between 0 to 1
- Apply both the rules over the dataset

```
# Check the missing values in each column
environmental_and_socioeconomic_sustainability %>% is.na() %>% colSums()
```

```
##                                                                    HDI.rank
##                                                                           0
##                                                                     Country
##                                                                           0
##                                                  Fossil.fuel.energy.consumption
##                                                                          52
##                                                   Renewable.energy.consumption
##                                                                           0
##                                               Carbon.dioxide.emissions.Per.Capita
##                                                                           0
##                                            Carbon.dioxide.emissions.KG.per.PPP.GDP
##                                                                           4
##                                                                  Forest.Cover
##                                                                           2
##                                                        Forest.Cover.Change
##                                                                           8
##                                                      Fresh.water.withdrawals
##                                                                          83
##           Mortality.Rate.attributed.to.Household.and.ambient.air.pollution
##                                                                           8
## Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services
##                                                                           8
##                                                               Red.List.Index
##                                                                           0
##                                                          Adjusted.net.savings
##                                                                          36
##                                                            Total.debt.service
##                                                                          72
##                                                       Gross.capital.formation
##                                                                          15
##                                                            Skilled.labour.force
##                                                                          48
##                                                     Concentration.index.exports
##                                                                           2
##                                               Research.and.development.expenditure
##                                                                          64
##               Education.and.health.expenditure.versus.military.expenditure
##                                                                          28
##          Ratio.of.education.and.health.expenditure.to.military.expenditure
##                                                                          73
##                           Overall.loss.in.HDI.value.due.to.inequality.Change
##                                                                          57
##                                                 Gender.Inequality.Index.Change
##                                                                          45
##                                                   Income.quintile.ratio.Change
##                                                                          65
##                                                        Red.List.Index.Category
##                                                                           0
##                                                   Forest.Cover.Change.Category
##                                                                           8
```

```
# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Fossil.fuel.energy.consumption = ifelse(is.na(Fossil.fuel.energy.consumption),
round(mean(Fossil.fuel.energy.consumption, na.rm = TRUE), 1), Fossil.fuel.energy.consumptio
n))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Carbon.dioxide.emissions.KG.per.PPP.GDP = ifelse(is.na(Carbon.dioxide.emission
s.KG.per.PPP.GDP), round(mean(Carbon.dioxide.emissions.KG.per.PPP.GDP, na.rm = TRUE), 2), Car
bon.dioxide.emissions.KG.per.PPP.GDP))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Forest.Cover = ifelse(is.na(Forest.Cover), round(mean(Forest.Cover, na.rm = TR
UE), 1), Forest.Cover))#.....
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Forest.Cover.Change = ifelse(is.na(Forest.Cover.Change), round(mean(Forest.Cov
er.Change, na.rm = TRUE), 1), Forest.Cover.Change))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Fresh.water.withdrawals = ifelse(is.na(Fresh.water.withdrawals), round(mean(Fr
esh.water.withdrawals, na.rm = TRUE), 1), Fresh.water.withdrawals))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Mortality.Rate.attributed.to.Household.and.ambient.air.pollution = ifelse(is.n
a(Mortality.Rate.attributed.to.Household.and.ambient.air.pollution), round(mean(Mortality.Rat
e.attributed.to.Household.and.ambient.air.pollution, na.rm = TRUE), 1), Mortality.Rate.attrib
uted.to.Household.and.ambient.air.pollution))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services = i
felse(is.na(Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services), roun
d(mean(Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services, na.rm = TR
UE), 1), Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Adjusted.net.savings = ifelse(is.na(Adjusted.net.savings), round(mean(Adjuste
d.net.savings, na.rm = TRUE), 1), Adjusted.net.savings))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Total.debt.service = ifelse(is.na(Total.debt.service), round(mean(Total.debt.s
ervice, na.rm = TRUE), 1), Total.debt.service))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Gross.capital.formation = ifelse(is.na(Gross.capital.formation), round(mean(Gr
oss.capital.formation, na.rm = TRUE), 1), Gross.capital.formation))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Skilled.labour.force = ifelse(is.na(Skilled.labour.force), round(mean(Skilled.
```

```
labour.force, na.rm = TRUE), 1), Skilled.labour.force))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Concentration.index.exports = ifelse(is.na(Concentration.index.exports), round
(mean(Concentration.index.exports, na.rm = TRUE), 3), Concentration.index.exports))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Research.and.development.expenditure = ifelse(is.na(Research.and.development.e
xpenditure), round(mean(Research.and.development.expenditure, na.rm = TRUE), 1), Research.an
d.development.expenditure))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Education.and.health.expenditure.versus.military.expenditure = ifelse(is.na(Ed
ucation.and.health.expenditure.versus.military.expenditure), round(mean(Education.and.health.
expenditure.versus.military.expenditure, na.rm = TRUE), 1), Education.and.health.expenditure.
versus.military.expenditure))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Ratio.of.education.and.health.expenditure.to.military.expenditure = ifelse(is.
na(Ratio.of.education.and.health.expenditure.to.military.expenditure), round(mean(Ratio.of.ed
ucation.and.health.expenditure.to.military.expenditure, na.rm = TRUE), 1), Ratio.of.educatio
n.and.health.expenditure.to.military.expenditure))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Overall.loss.in.HDI.value.due.to.inequality.Change = ifelse(is.na(Overall.los
s.in.HDI.value.due.to.inequality.Change), round(mean(Overall.loss.in.HDI.value.due.to.inequal
ity.Change, na.rm = TRUE), 1), Overall.loss.in.HDI.value.due.to.inequality.Change))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Gender.Inequality.Index.Change = ifelse(is.na(Gender.Inequality.Index.Change),
round(mean(Gender.Inequality.Index.Change, na.rm = TRUE), 1), Gender.Inequality.Index.Chang
e))

# Fill missing values with the mean value
environmental_and_socioeconomic_sustainability <- environmental_and_socioeconomic_sustainabil
ity %>% mutate(Income.quintile.ratio.Change = ifelse(is.na(Income.quintile.ratio.Change), rou
nd(mean(Income.quintile.ratio.Change, na.rm = TRUE), 1), Income.quintile.ratio.Change))

# Fill the values in new values in Forest.Cover.Change.Category column
environmental_and_socioeconomic_sustainability <- mutate(environmental_and_socioeconomic_sust
ainability, Forest.Cover.Change.Category = ifelse(Forest.Cover.Change > 0, "Incr", ifelse(For
est.Cover.Change < 0,"Decr","Same")))

# Check the class of Forest.Cover.Change.Category variable
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category %>% class()
```

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

```
# Convert the variable to factor
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category  <- factor(enviro
nmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category , labels=c("Decr", "Sam
e", "Incr"), ordered = FALSE)

# Verify the Forest.Cover.Change.Category column in convereted to factor
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category %>% is.factor()
```

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

```
# Verify the levels of the column
environmental_and_socioeconomic_sustainability$Forest.Cover.Change.Category %>% levels()
```

```
## [1] "Decr" "Same" "Incr"
```

```
# Verify the missing values is removed
environmental_and_socioeconomic_sustainability %>% is.na() %>% colSums()
```

```
##                                                                   HDI.rank
##                                                                          0
##                                                                    Country
##                                                                          0
##                                                 Fossil.fuel.energy.consumption
##                                                                          0
##                                                   Renewable.energy.consumption
##                                                                          0
##                                               Carbon.dioxide.emissions.Per.Capita
##                                                                          0
##                                            Carbon.dioxide.emissions.KG.per.PPP.GDP
##                                                                          0
##                                                                 Forest.Cover
##                                                                          0
##                                                           Forest.Cover.Change
##                                                                          0
##                                                        Fresh.water.withdrawals
##                                                                          0
##           Mortality.Rate.attributed.to.Household.and.ambient.air.pollution
##                                                                          0
## Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services
##                                                                          0
##                                                              Red.List.Index
##                                                                          0
##                                                            Adjusted.net.savings
##                                                                          0
##                                                              Total.debt.service
##                                                                          0
##                                                         Gross.capital.formation
##                                                                          0
##                                                            Skilled.labour.force
##                                                                          0
##                                                     Concentration.index.exports
##                                                                          0
##                                               Research.and.development.expenditure
##                                                                          0
##               Education.and.health.expenditure.versus.military.expenditure
##                                                                          0
##          Ratio.of.education.and.health.expenditure.to.military.expenditure
##                                                                          0
##                           Overall.loss.in.HDI.value.due.to.inequality.Change
##                                                                          0
##                                                 Gender.Inequality.Index.Change
##                                                                          0
##                                                   Income.quintile.ratio.Change
##                                                                          0
##                                                        Red.List.Index.Category
##                                                                          0
##                                                   Forest.Cover.Change.Category
##                                                                          0
```

```
# Function made to check for special values, i.e. infinite (Inf and -Inf) and NaN
is.special <- function(x){
  if (is.numeric(x)) (is.infinite(x) | is.nan(x))
}

# Apply the function on each varible
sapply(environmental_and_socioeconomic_sustainability, function(x) sum(is.special(x)))
```

```
##                                                                HDI.rank
##                                                                       0
##                                                                 Country
##                                                                       0
##                                              Fossil.fuel.energy.consumption
##                                                                       0
##                                                Renewable.energy.consumption
##                                                                       0
##                                           Carbon.dioxide.emissions.Per.Capita
##                                                                       0
##                                        Carbon.dioxide.emissions.KG.per.PPP.GDP
##                                                                       0
##                                                              Forest.Cover
##                                                                       0
##                                                       Forest.Cover.Change
##                                                                       0
##                                                    Fresh.water.withdrawals
##                                                                       0
##           Mortality.Rate.attributed.to.Household.and.ambient.air.pollution
##                                                                       0
## Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services
##                                                                       0
##                                                            Red.List.Index
##                                                                       0
##                                                          Adjusted.net.savings
##                                                                       0
##                                                            Total.debt.service
##                                                                       0
##                                                        Gross.capital.formation
##                                                                       0
##                                                           Skilled.labour.force
##                                                                       0
##                                                  Concentration.index.exports
##                                                                       0
##                                          Research.and.development.expenditure
##                                                                       0
##               Education.and.health.expenditure.versus.military.expenditure
##                                                                       0
##          Ratio.of.education.and.health.expenditure.to.military.expenditure
##                                                                       0
##                           Overall.loss.in.HDI.value.due.to.inequality.Change
##                                                                       0
##                                               Gender.Inequality.Index.Change
##                                                                       0
##                                                 Income.quintile.ratio.Change
##                                                                       0
##                                                      Red.List.Index.Category
##                                                                       0
##                                                 Forest.Cover.Change.Category
##                                                                       0
```

```
# Make a new rule Rule1 that HDI Rank is between 1 to 189, as total of 189 countries are pres
ent in the data sets
(Rule1 <- editset(c("HDI.rank > 0", "HDI.rank < 190")))
```

```
##
## Edit set:
## num1 : 0 < HDI.rank
## num2 : HDI.rank < 190
```

```
# Make a new rule Rule2 to check the Red List Index is 0 to 1
(Rule2 <- editset(c("Red.List.Index > 0", "Red.List.Index < 1")))
```

```
##
## Edit set:
## num1 : 0 < Red.List.Index
## num2 : Red.List.Index < 1
```

```
# Check the Rule1 on the data set
sum(violatedEdits(Rule1, environmental_and_socioeconomic_sustainability))
```

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

```
# Check the Rule2 on the data set
sum(violatedEdits(Rule2, environmental_and_socioeconomic_sustainability))
```

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

Scan II

- In this step, we created a function called “outliers()” to seek the outliers in each variable of the data set
- Apply this function by sapply() function to call outliers() function over the dataframe on only column 1 and columns 3 to 23, since those are numeric
- Create a function called “cap()” which define as “replace the outliers with its nearest neighbour which is not an outlier”
- Apply cap() function by using sapply() function to the dataframe which only column 1 and columns 3 to 23, since those are numeric

```
# Function to check the outliers in each variable of the data set
outliers <- function(x) {
  boxplot(x, plot= FALSE)$out
}

# Apply the function outlier() on each numeric column
sapply(environmental_and_socioeconomic_sustainability[,c(1, 3:23)], FUN = outliers)
```

```
## $HDI.rank
## numeric(0)
##
## $Fossil.fuel.energy.consumption
##  [1] 11.5 26.8 12.3 24.2 33.7 17.2 10.6 30.7 15.8 29.9 14.4 29.1 19.0 17.2
## [15] 31.8 22.0 26.5  6.1  5.4 22.3 12.6 24.1
##
## $Renewable.energy.consumption
## numeric(0)
##
## $Carbon.dioxide.emissions.Per.Capita
##  [1] 15.4 15.1 16.5 17.4 14.8 23.3 45.4 22.1 19.5 23.4 15.4 25.2 14.4 34.2
##
## $Carbon.dioxide.emissions.KG.per.PPP.GDP
##  [1] 0.55 0.53 0.61 0.96 1.10 0.59 0.59 0.64 0.63 0.64 0.53 0.56 0.87 0.72
## [15] 0.52
##
## $Forest.Cover
## numeric(0)
##
## $Forest.Cover.Change
##  [1]  62.2 205.6  33.2 144.1  32.1 131.3  81.2  60.5  34.6  32.6  79.5
## [12]  61.9  65.9  65.6  55.7 -43.6  34.7 -41.7  32.1 -36.6 -59.4  50.9
## [23] -45.9 -56.4 -72.6 -37.5 -41.3
##
## $Fresh.water.withdrawals
## [1] 943.3 117.8 822.9 126.6  74.4
##
## $Mortality.Rate.attributed.to.Household.and.ambient.air.pollution
## [1] 307.4 324.1
##
## $Mortality.rate.attributed.to.Unsafe.water..sanitation.and.hygiene.services
##  [1]  51.2  48.8  45.2  68.6  44.4  59.7  50.7  41.6  47.2  43.7  44.6
## [12]  59.8  45.6  41.5  70.7  49.6  81.3  65.4 101.0  63.3  82.1  70.8
##
## $Red.List.Index
## [1] 0.401 0.569
##
## $Adjusted.net.savings
##  [1]  32.7  34.1  30.3  29.9  41.5 -22.6 -14.8 -38.4  32.5 -20.7 -31.0
## [12] -28.5 -19.1 -16.9 -18.7 -15.8 -39.3 -29.5 -28.9
##
## $Total.debt.service
##  [1] 95.5 34.9 44.3 39.3 30.0 37.6 41.4 59.5 51.2 34.1 26.8 29.3 28.9 28.5
## [15] 40.4 39.6 28.1 26.5 49.1
##
## $Gross.capital.formation
##  [1] 45.2 43.8 47.8 43.6 56.5 47.2 47.2  7.8 42.5 56.8 43.4 50.4 43.1  1.7
##
## $Skilled.labour.force
## numeric(0)
##
## $Concentration.index.exports
## [1] 0.876 0.937 0.892 0.934 0.876
##
## $Research.and.development.expenditure
##  [1] 1.9 3.0 2.2 2.9 2.2 3.3 2.2 2.0 3.0 2.8 1.7 2.9 2.5 3.3 3.1 4.3 4.2
```

```
## [18] 2.2 2.2 1.9 2.1
##
## $Education.and.health.expenditure.versus.military.expenditure
##  [1]  4.7  5.6 10.2  4.0 12.0  4.2  5.7  3.8  4.7  3.9  5.9  4.8  9.1  3.8
## [15]  5.6  4.1  4.0  4.6
##
## $Ratio.of.education.and.health.expenditure.to.military.expenditure
##  [1] 25.6 32.0 16.5 15.6 17.4 14.9 21.0 24.4 33.2  1.8  1.7 15.0  0.5 57.9
## [15] 16.7  1.8 51.4 18.2 21.9 31.1 17.2  1.5 23.4 23.0  1.4  0.4
##
## $Overall.loss.in.HDI.value.due.to.inequality.Change
##  [1]  3.0  2.4  3.0  2.1  8.2  4.6  2.0 -5.9 -5.9 -4.9  5.9 -5.3 -5.3  7.1
## [15] -5.0  2.0 -4.7
##
## $Gender.Inequality.Index.Change
##  [1] -3.6 -3.9 -3.6 -4.2 -5.0 -3.9 -3.6 -4.5 -3.3 -4.8 -3.6 -3.4 -3.9 -5.1
## [15] -4.2 -3.6 -3.3 -5.1 -5.4 -4.3 -3.8 -3.6  1.1 -3.5  0.7
##
## $Income.quintile.ratio.Change
##  [1]  1.3  1.4  2.6  1.6  1.9 -2.2 -3.1 -4.8 -2.1 -3.1 -2.6 -5.1 -2.3 -3.7
## [15] -3.0 -2.1 -3.8 -5.4 -4.0 -2.2 -2.1 -2.7 -2.9 -4.4 -2.7  1.6  1.9  2.3
## [29] -3.3  4.1 -3.5 -2.4  2.0 13.3 -2.5  3.3  2.9  6.3 -4.2 -3.3 10.0  3.9
## [43] -2.7 -3.4 -2.4  3.2  8.6 -3.5
```

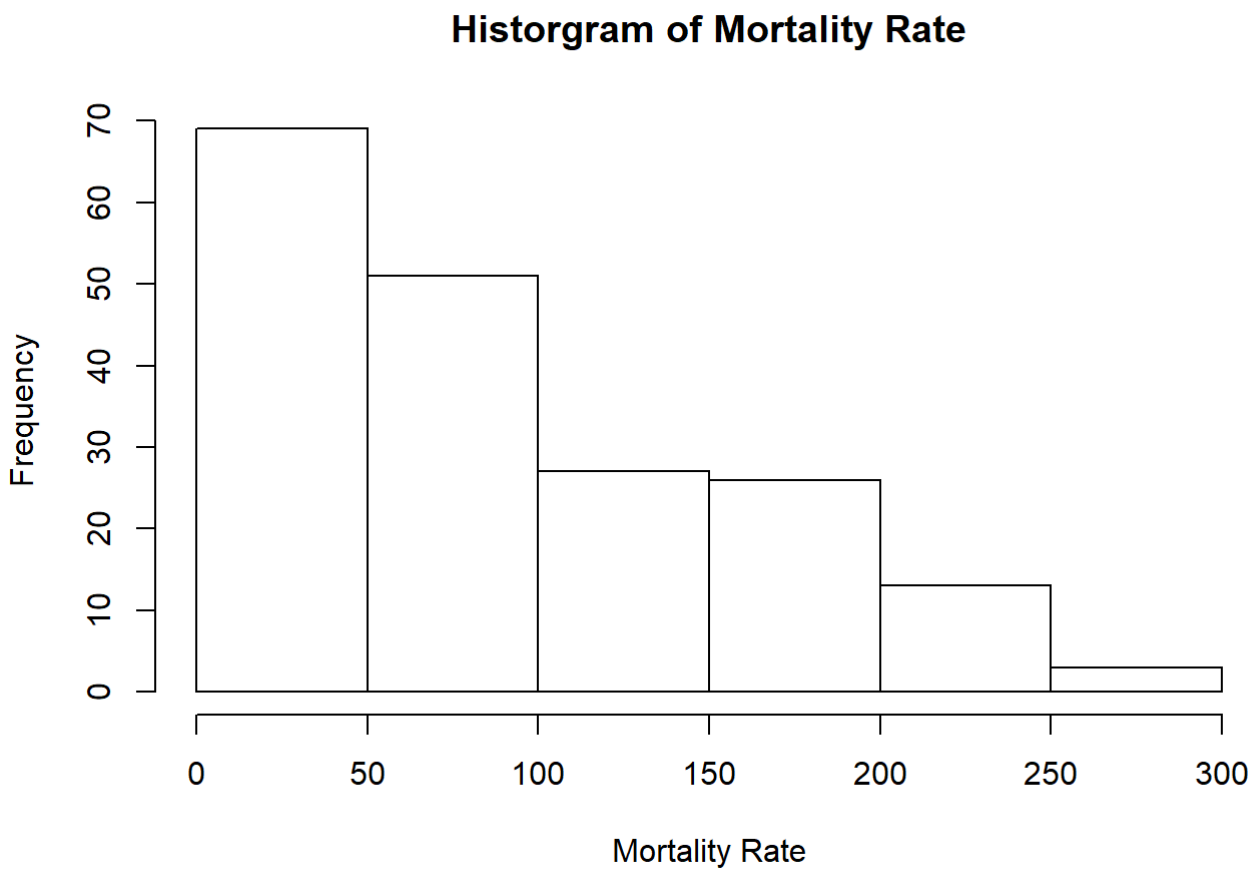
```
# Function to replace the outliers with its nearest neighbour which is not an outlier
cap <- function(x){
  quantiles <- quantile( x, c(.05, 0.25, 0.75, .95 ) )
  x[ x < quantiles[2] - 1.5*IQR(x) ] <- quantiles[1]
  x[ x > quantiles[3] + 1.5*IQR(x) ] <- quantiles[4]
  x
}

# Apply the function outlier() on each numeric column
environmental_and_socioeconomic_sustainability[,c(1, 3:23)] <- sapply(environmental_and_socio
economic_sustainability[,c(1, 3:23)], FUN = cap)
```

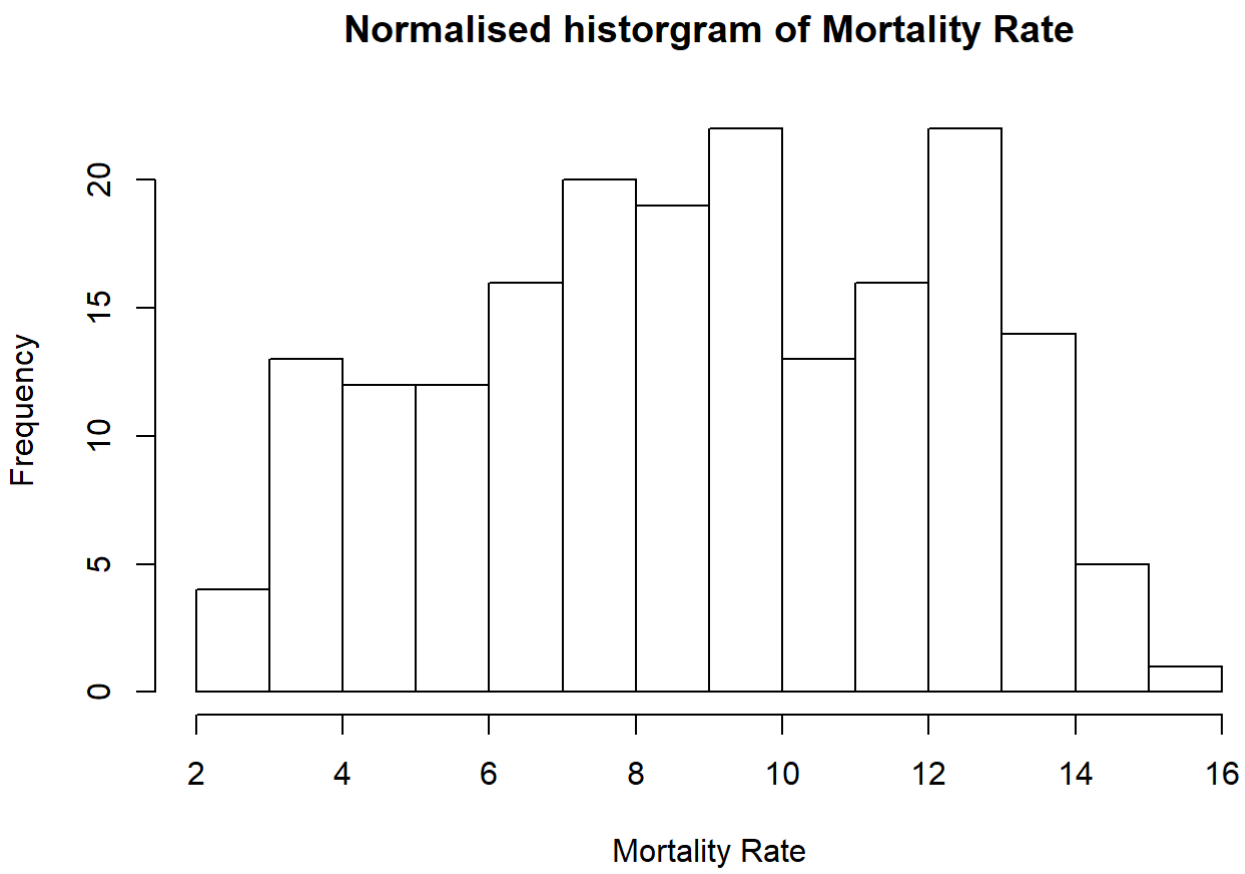
Transform

- In this step, we applied data transformation on the column “Mortality.Rate.attributed.to.Household.and.ambient.air.pollution”
- We checked the histogram of the column “Mortality.Rate.attributed.to.Household.and.ambient.air.pollution” and found that it is “Skewed to the Right”
- To make the data normalised, we applied Box-Cox transformation with “lamda=‘auto’”
- This function can easily transform the skewed data to normal distribution, since normal distribution is important for statistical hypothesis testing
- We the plot the histogram of the transformed data to verify that the data is transformed to a normal distribution

```
# Plot the histogram for the column "Mortality.Rate.attributed.to.Household.and.ambient.air.p
ollution"
hist(environmental_and_socioeconomic_sustainability$Mortality.Rate.attributed.to.Household.an
d.ambient.air.pollution, main="Historgram of Mortality Rate", xlab="Mortality Rate")
```




```
# Apply Box-Cox transformation on the column and plot its histogram
hist(BoxCox(environmental_and_socioeconomic_sustainability$Mortality.Rate.attributed.to.House
hold.and.ambient.air.pollution, lambda = "auto"), main="Normalised historgram of Mortality Ra
te", xlab="Mortality Rate")
```


Normalised historgram of Mortality Rate
Frequency
0
5
10
15
20
2
4
6
8
10
12
14
16
Mortality Rate