Analysis of Mortality & Population in the World

Ammarah Naveed

Setup

Loading packages that are needed in the report

```
library(magrittr)
library(dplyr)
library(outliers)
library(knitr)
library(tidyr)
library(openxlsx)
library(writexl)
library(deducorrect)
library(Hmisc)
library(igraph)
```

Health authorities make extensive use of mortality data, for example in monitoring the health of their local populations over time; drawing comparisons with other health authorities; and exploring variations in health within their local populations.

Two datasets that are presented here in which one is explaining population by year 2020 for different countries throughout the world, it's area covered by population, yearly change in population and fertility rate etc. . . .. The other one is explaining death causes from different diseases/accidents in different countries throughout the world. Here is a complete list of the variables that are available in the 2 datasets.

- Death Causes by countries:
  - Country_names
  - Cardiovascular_diseases
  - Kidney_diseases
  - Interpersonal_violence
  - Tuberculosis
  - Lower_respiratory_infections
  - Diarrheal_diseases
  - Alzheimer_disease
  - Fire_heat

  - Drug_use_disorders

  - Diabetes
  - Covid19_Deaths
  - Respiratory_diseases
  - Malaria

  - HIV_AIDS
  - Maternal_disorders
  - Alcohol_use_disorders
  - Poisoning
  - Parkinson_disease
  - Drowning

  - Road_injuries

- Countries population by year 2020:

  - Country.names
  - Population.2020
  - Yearly.change
  - Net.Change
  - Density.per.kilometer
  - Land.Area.km
  - Migrants.net
  - Fertility.rate
  - Median.age
  - Urban.pop
  - World.share

First dataset:

I have picked worldwide dataset of death cause's by country. It is important to understand what is meant by the cause of death:

In the epidemiological framework of the Global Burden of Disease study each death has one specific cause. In their own words: 'each death is attributed to a single underlying cause — the cause that initiated the series of events leading to death'.

THis date of death cause by country helps for an estimation of the reduction of the number of deaths that would be achieved if the risk factors to which a population is exposed would be eliminated (in the case of tobacco smoking, for example) or reduced to an optimal, healthy level. (Causes of Death)

The first one explain the death reasons as different countries have different major reasons of deaths.Across low- and middle-income countries, mortality from infectious disease, malnutrition, nutritional deficiencies, neonatal and maternal deaths are common – and in some cases, dominant. In Kenya, for example, diarrhea infections are still the primary cause of death. HIV/AIDS is the major cause of death in South Africa and Botswana. However, in high-income countries, the proportion of deaths due by these causes is quite low.

Second dataset: Countires population by year 2020

World Population World Population and top 20 Countries Live Clock. Population in the past, present, and future. Milestones. Global Growth Rate. World population by Region and by Religion. Population Density, Fertility Rate, Median Age, Migrants. All-time population total.

Import dataset 1:

I have downloaded my first dataset from Kaggle (Feb 2022,Death Cause by Country) as CSV format and then imported by using the read library.This dataset contains 34 columns and 191 observations which is too

large and a little messy to get meaningful insights.I have reduced by sub-setting the columns by removing certain columns like column1 & colum2 that are not useful in this report. After reducing, now my dataset contains 22 columns. This dataset has each variable name with one or two dots in between which is not much readable and also not best practice so I renamed every variable name by replacing dots with "_" sign which makes it more readable.

```
data <- read.csv("/Users/ammarahnaveed/downloads/Death Cause Reason by Country.csv")
df1 <- data.frame(data)
df = subset(df1, select = -c(6:7,17,20:24,29:31,33))
df <-  rename(df, "Country_names" = "Country.Name","Fire_heat"= "Fire..heat","Drug_use_disorders" = "Dru

colnames(df)
```

```
##  [1] "Country_names"                "Covid19_Deaths"
##  [3] "Cardiovascular_diseases"      "Respiratory_diseases"
##  [5] "Kidney_diseases"              "Malaria"
##  [7] "Interpersonal_violence"       "HIV_AIDS"
##  [9] "Tuberculosis"                 "Maternal_disorders"
## [11] "Lower_respiratory_infections" "Alcohol_use_disorders"
## [13] "Diarrheal_diseases"           "Poisoning"
## [15] "Alzheime_disease"             "Parkinson_disease"
## [17] "Fire_heat"                    "Drowning"
## [19] "Drug_use_disorders"           "Road_injuries"
## [21] "Diabetes"
```

Import dataset 2:

Here is my second imported dataset from kaggle (2020, Population by Country) as csv format.This dataset is quite clean as compared to the first one, it contains 11 columns with 235 observations.It also has dots in variable names, i have replaced it with "_" to make it more clear. For detailed work on my report i have created the data frames for both oif the datasets.

```
data1 <- read.csv("/Users/ammarahnaveed/downloads/population_by_country_2020.csv")
df2 <- data.frame(data1)
df2 <-
  rename(df2, "Fertility_rate" = "Fert..Rate",
         "Median_age" = "Med..Age","Population_2020" = "Population..2020.","Yearly_Change"= "Yearly.Cha
colnames(df2)
```

```
##  [1] "Country_names"          "Population_2020"       "Yearly_Change"
##  [4] "Net_change"             "Density_per_kilometer" "Land_Area(km)"
##  [7] "Migrants_net"           "Fertility_rate"        "Median_age"
## [10] "Urban_pop"              "World_share"
```

Merged dataset:

I have merged both data frames and created a new data frame using common variable of country name. Both datasets have country name variable in them which gives information of death causes and population in 2020 and land area etc. . . , for these countries.

```
merge_data <- merge(df2,df, by = "Country_names")
head(merge_data)
```

```
##         Country_names Population_2020 Yearly_Change Net_change
## 1         Afghanistan        39074280        2.33 %     886592
## 2             Albania         2877239       -0.11 %      -3120
## 3             Algeria        43984569        1.85 %     797990
## 4             Andorra           77287        0.16 %        123
## 5              Angola        33032075        3.27 %    1040977
## 6 Antigua and Barbuda           98069        0.84 %        811
##   Density_per_kilometer Land_Area(km) Migrants_net Fertility_rate Median_age
## 1                    60        652860       -62920            4.6         18
## 2                   105         27400       -14000            1.6         36
## 3                    18       2381740       -10000            3.1         29
## 4                   164           470           NA           N.A.       N.A.
## 5                    26       1246700         6413            5.6         17
## 6                   223           440            0            2.0         34
##   Urban_pop World_share Covid19_Deaths Cardiovascular_diseases
## 1      25 %      0.50 %           2201                   61995
## 2      63 %      0.04 %           1181                   12904
## 3      73 %      0.56 %           2762                   97931
## 4      88 %      0.00 %             84                     169
## 5      67 %      0.42 %             33                   25724
## 6      26 %      0.00 %              5                     200
##   Respiratory_diseases Kidney_diseases Malaria Interpersonal_violence HIV_AIDS
## 1                 7082            5637     530                  5015      318
## 2                  815             329       0                    57        2
## 3                 7528            8201       0                   459      264
## 4                   39              16       0                     0        3
## 5                 3934            2464   10784                   974    16802
## 6                   11              36       0                     5        7
##   Tuberculosis Maternal_disorders Lower_respiratory_infections
## 1         3627               4038                       18697
## 2           11                  3                         457
## 3          445                638                        5786
## 4            0                  0                          20
## 5        11752               2069                       12783
## 6            0                  0                          30
##   Alcohol_use_disorders Diarrheal_diseases Poisoning Alzheime_disease
## 1                   147               4320       525             1775
## 2                    18                  7        11              917
## 3                   111                527       351             5209
## 4                     1                  1         0               36
## 5                   211              12936       433             1143
## 6                     7                  2         0               16
##   Parkinson_disease Fire_heat Drowning Drug_use_disorders Road_injuries
## 1               560       485     1687                406          8254
## 2               248        18       36                 29           243
## 3              1283       782      526                526         11051
## 4                 7         0        0                  0             8
## 5               267       513      793                 80          9253
## 6                 5         2        4                  0             7
##   Diabetes
## 1     4817
## 2      175
## 3     5328
## 4        9
```

```
## 5     4033
## 6       57
```

Understanding:

This data is useful for certain organizations like WHO, UNICEF etc that are working in health sector to identify which counties are suffering from diseases and certain outbreaks. By merging this health data with the population one, we can correlate the effect of population with certain death causes.

This merged dataset is explaining number of people, net change and by what percentage population changed in 2020 for every country, land area per kilometer and density of population per kilometer,how many migrants arrives in 2020 for every country and fertility rate which directly effects on increase in population. Data also gives information of median age per country and by what percent this all information share world. As fertility rate describes increase in population, death rate results decrease in population. This data gives information about the death causes in every country which helps world health organisations to work on that causes to reduce death rate.

Here is the code chunks giving the information about the attributes and classes of data set. This dataset have integers and characters.

```
attr(x = "merge_data", which = "colnames")
```

```
## NULL
```

```
class(merge_data)
```

```
## [1] "data.frame"
```

Tidy and Manipulate data 1:

This dataset still have 33 variables which can be reduced by explaining similar kind of information in a single variable like lower respiratory and respiratory infection can be explained under one variable, alcohol_use_disorders and drug_use_disorders can comes under drug_use_ orders and deaths from fire, heat, drowning and road injuries can all be explained altogether under one variable of accidental deaths.

```
merge_data<- mutate(merge_data,respiratory_infections = Lower_respiratory_infections + Respiratory_disea
       drug_use_disorders = Alcohol_use_disorders + Drug_use_disorders, accidental_deaths = Fire_heat +

merge_data <- subset(merge_data, select = -c(11,14,21,22,24,27:28,29,29))

head(merge_data)
```

```
##         Country_names Population_2020 Yearly_Change Net_change
## 1         Afghanistan        39074280        2.33 %     886592
## 2             Albania         2877239       -0.11 %      -3120
## 3             Algeria        43984569        1.85 %     797990
## 4             Andorra           77287        0.16 %        123
## 5              Angola        33032075        3.27 %    1040977
## 6 Antigua and Barbuda           98069        0.84 %        811
##   Density_per_kilometer Land_Area(km) Migrants_net Fertility_rate Median_age
## 1                    60        652860       -62920            4.6         18
## 2                   105         27400       -14000            1.6         36
## 3                    18       2381740       -10000            3.1         29
```

```
## 4                  164           470           NA           N.A.       N.A.
## 5                   26       1246700         6413           5.6         17
## 6                  223           440            0           2.0         34
##   Urban_pop Covid19_Deaths Cardiovascular_diseases Kidney_diseases Malaria
## 1      25 %           2201                  61995            5637     530
## 2      63 %           1181                  12904             329       0
## 3      73 %           2762                  97931            8201       0
## 4      88 %             84                    169              16       0
## 5      67 %             33                  25724            2464   10784
## 6      26 %              5                    200              36       0
##   Interpersonal_violence HIV_AIDS Tuberculosis Maternal_disorders
## 1                   5015      318         3627               4038
## 2                     57        2           11                  3
## 3                    459      264          445                638
## 4                      0        3            0                  0
## 5                    974    16802        11752               2069
## 6                      5        7            0                  0
##   Diarrheal_diseases Alzheime_disease Parkinson_disease Road_injuries Diabetes
## 1               4320             1775               560          8254     4817
## 2                  7              917               248           243      175
## 3                527             5209              1283         11051     5328
## 4                  1               36                 7             8        9
## 5              12936             1143               267          9253     4033
## 6                  2               16                 5             7       57
##   respiratory_infections drug_use_disorders accidental_deaths
## 1                  25779               553              2172
## 2                   1272                47                54
## 3                  13314               637              1308
## 4                     59                 1                 0
## 5                  16717               291              1306
## 6                     41                 7                 6
```

Tidy & Manipulate Data 2:

In second dataset, there is urban population variable and we can extract the information about rural population out of this existing one which gives more useful and detailed information. In this variable values are given in percentage format and i can"t use extraction method with these percentage signs. To solve this problem i have used gsub fuction first to remove % sign and subract the given observation from 100 and then placing the % sign again by paste function to the extracted values.

```
merge_data$Rural_pop <- paste(100 - as.numeric(gsub(" %$","",merge_data$Urban_pop)),"%")
```

```
## Warning in paste(100 - as.numeric(gsub(" %$", "", merge_data$Urban_pop)), : NAs
## introduced by coercion
```

```
head(merge_data)
```

```
##         Country_names Population_2020 Yearly_Change Net_change
## 1         Afghanistan        39074280        2.33 %     886592
## 2             Albania         2877239       -0.11 %      -3120
## 3             Algeria        43984569        1.85 %     797990
## 4             Andorra           77287        0.16 %        123
## 5              Angola        33032075        3.27 %    1040977
```

```
## 6 Antigua and Barbuda           98069        0.84 %        811
##   Density_per_kilometer Land_Area(km) Migrants_net Fertility_rate Median_age
## 1                    60        652860       -62920            4.6         18
## 2                   105         27400       -14000            1.6         36
## 3                    18       2381740       -10000            3.1         29
## 4                   164           470           NA           N.A.        N.A.
## 5                    26       1246700         6413            5.6         17
## 6                   223           440            0            2.0         34
##   Urban_pop Covid19_Deaths Cardiovascular_diseases Kidney_diseases Malaria
## 1      25 %           2201                   61995            5637     530
## 2      63 %           1181                   12904             329       0
## 3      73 %           2762                   97931            8201       0
## 4      88 %             84                     169              16       0
## 5      67 %             33                   25724            2464   10784
## 6      26 %              5                     200              36       0
##   Interpersonal_violence HIV_AIDS Tuberculosis Maternal_disorders
## 1                   5015      318         3627               4038
## 2                     57        2           11                  3
## 3                    459      264          445                638
## 4                      0        3            0                  0
## 5                    974    16802        11752               2069
## 6                      5        7            0                  0
##   Diarrheal_diseases Alzheime_disease Parkinson_disease Road_injuries Diabetes
## 1               4320             1775               560          8254     4817
## 2                  7              917               248           243      175
## 3                527             5209              1283         11051     5328
## 4                  1               36                 7             8        9
## 5              12936             1143               267          9253     4033
## 6                  2               16                 5             7       57
##   respiratory_infections drug_use_disorders accidental_deaths Rural_pop
## 1                  25779               553              2172      75 %
## 2                   1272                47                54      37 %
## 3                  13314               637              1308      27 %
## 4                     59                 1                 0      12 %
## 5                  16717               291              1306      33 %
## 6                     41                 7                 6      74 %
```

Scanning of dataset: There are some NA values I can see in observations of dataset,to check the number of NA values in dataset I have used Colsums(is.na()) and it results in 14 NA values in this whole dataset. There are 11 NA's in migrant_net column and 3 NA's in Covid19_Deaths. To fix these NA's values in migrant_net i have used mean function for migrant_net which is the average of all values in the set,I can see missing values in Fertility_rate column, Urban_pop and median_age column as well but these are the character vectors and is.na function can't detect these NA's from it. I need to convert the character vector into numeric vector first before detection of NA values. Urban_pop variable have percentage values and it can"t be converted into numeric vector without removing percentage signs, so i have removed the percentage sign first from urban_pop variable and then converted it into numeric vector.

```
merge_data$Urban_pop <- gsub(" %$","",merge_data$Urban_pop)

merge_data$Fertility_rate <- as.numeric(merge_data$Fertility_rate)
```

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

```
merge_data$Median_age <- as.numeric(merge_data$Median_age)
```

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

```
merge_data$Urban_pop <- as.numeric(merge_data$Urban_pop)
```

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

```
class(merge_data$Fertility_rate)
```

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

```
class(merge_data$Median_age)
```

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

```
class(merge_data$Median_age)
```

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

Missing values

Now checking the positions of these missing values I have used colSums() function which explains the number of NA's in each column separately.Fertility rate,urban_pop,Median_age,Migrant_net, Covid19_-Deaths have 11,8,11, 11 and 3 respectively. To know the exact place I have use which() function that shows the exact place of NA from specific mentioned column.

```
colSums(is.na(merge_data))
```

```
##           Country_names         Population_2020           Yearly_Change
##                       0                       0                       0
##              Net_change   Density_per_kilometer           Land_Area(km)
##                       0                       0                       0
##           Migrants_net          Fertility_rate              Median_age
##                      11                      11                      11
##              Urban_pop          Covid19_Deaths Cardiovascular_diseases
##                       8                       3                       0
##         Kidney_diseases                 Malaria  Interpersonal_violence
##                       0                       0                       0
##                HIV_AIDS            Tuberculosis      Maternal_disorders
##                       0                       0                       0
##      Diarrheal_diseases        Alzheime_disease       Parkinson_disease
##                       0                       0                       0
##           Road_injuries                Diabetes  respiratory_infections
##                       0                       0                       0
##      drug_use_disorders       accidental_deaths               Rural_pop
##                       0                       0                       0
```

```
which(is.na(merge_data$Median_age))
```

```
##  [1]   4  20  39  46  62 102 109 116 122 137 166
```

```
which(is.na(merge_data$Fertility_rate))
```

```
##  [1]   4  20  39  46  62 102 109 116 122 137 166
```

```
which(is.na(merge_data$Migrants_net))
```

```
##  [1]   4  20  39  46  62 102 109 116 122 137 166
```

```
which(is.na(merge_data$Covid19_Deaths))
```

```
## [1]  62 117 165
```

```
which(is.na(merge_data$Urban_pop))
```

```
## [1]  68  83 102 109 122 130 141 174
```

To check whether we can use mean, median or mode function for missing values, First we can do boxplot or distrbution plot, These boxplots also shows the outliers in variable's which needs to be addressed before working on the dataset.

We can use mean value to replace the missing values in case the data distribution is symmetric. Consider using median or mode with skewed data distribution. Migrant_net has large number of Negative and positive values in data, positive values shows the number of people coming inside the country after leaving their home country and negative values shows the number of people who are leaving their home country as migrant.I have used mean function for missing value in this variable which gives average values for all of the neagtive and positive values as it has normal skewed data. Covid19_Deaths are not normally distributed data with small number of deaths even 0 to very large number of deaths. I have checked it's histogram, it is right skewness data so I used mean here to fill th NA's. For Urban_pop, the data is also normally distributed values ranging from 0 to 100. Mean is the best to find out the values filling for NA's. Fertility rate have smaller values ranging from 1.1 to 5 and after checking with histogram the graph is rightly skewed. I have used median here to fill the NA's.

```
merge_data$Migrants_net[is.na(merge_data$Migrants_net)] <- mean(merge_data$Migrants_net,na.rm = TRUE)
sum(is.na(merge_data$Migrants_net))
```

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

```
merge_data$Covid19_Deaths[is.na(merge_data$Covid19_Deaths)] <- median(merge_data$Covid19_Deaths, na.rm =
sum(is.na(merge_data$Covid19_Deaths))
```

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

```
merge_data$Urban_pop[is.na(merge_data$Urban_pop)] <- mean(merge_data$Urban_pop, na.rm = TRUE)
sum(is.na(merge_data$Urban_pop))
```

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

```
merge_data$Median_age[is.na(merge_data$Median_age)] <- mean(merge_data$Median_age, na.rm = TRUE)
sum(is.na(merge_data$Median_age))
```

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

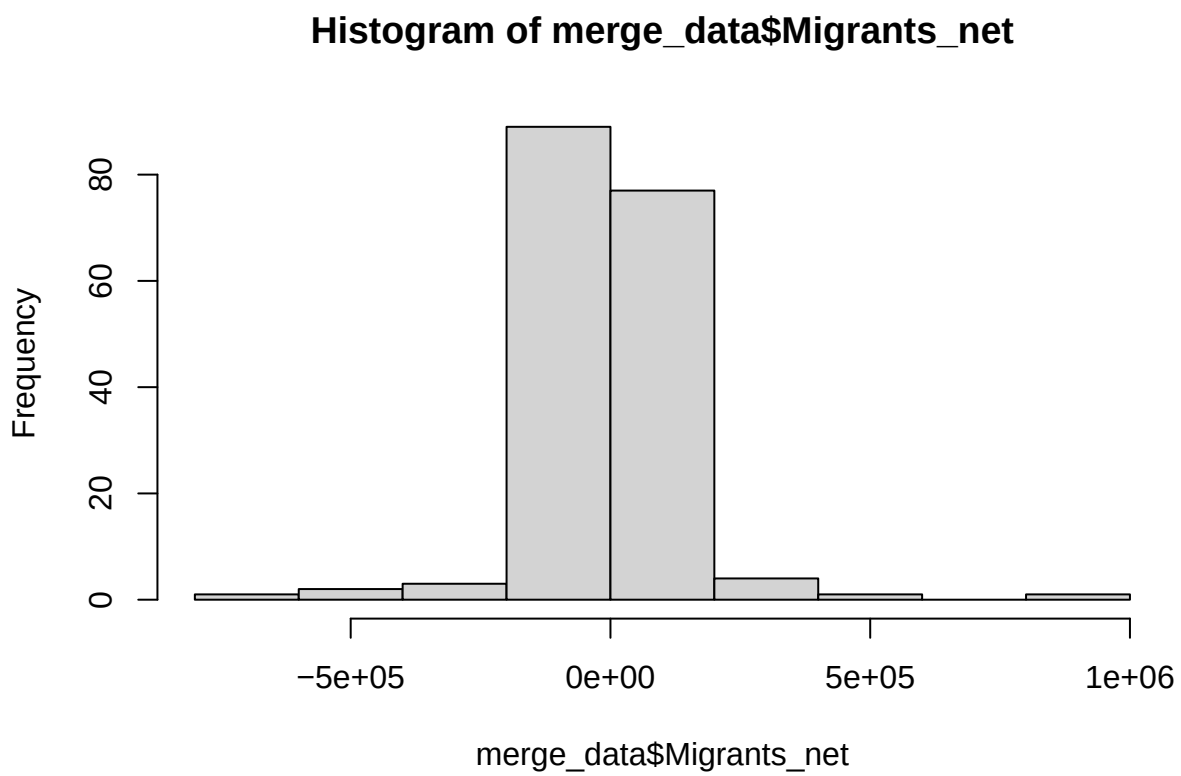
```
merge_data$Fertility_rate[is.na(merge_data$Fertility_rate)] <- median(merge_data$Fertility_rate, na.rm =
sum(is.na(merge_data$Fertility_rate))
```

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

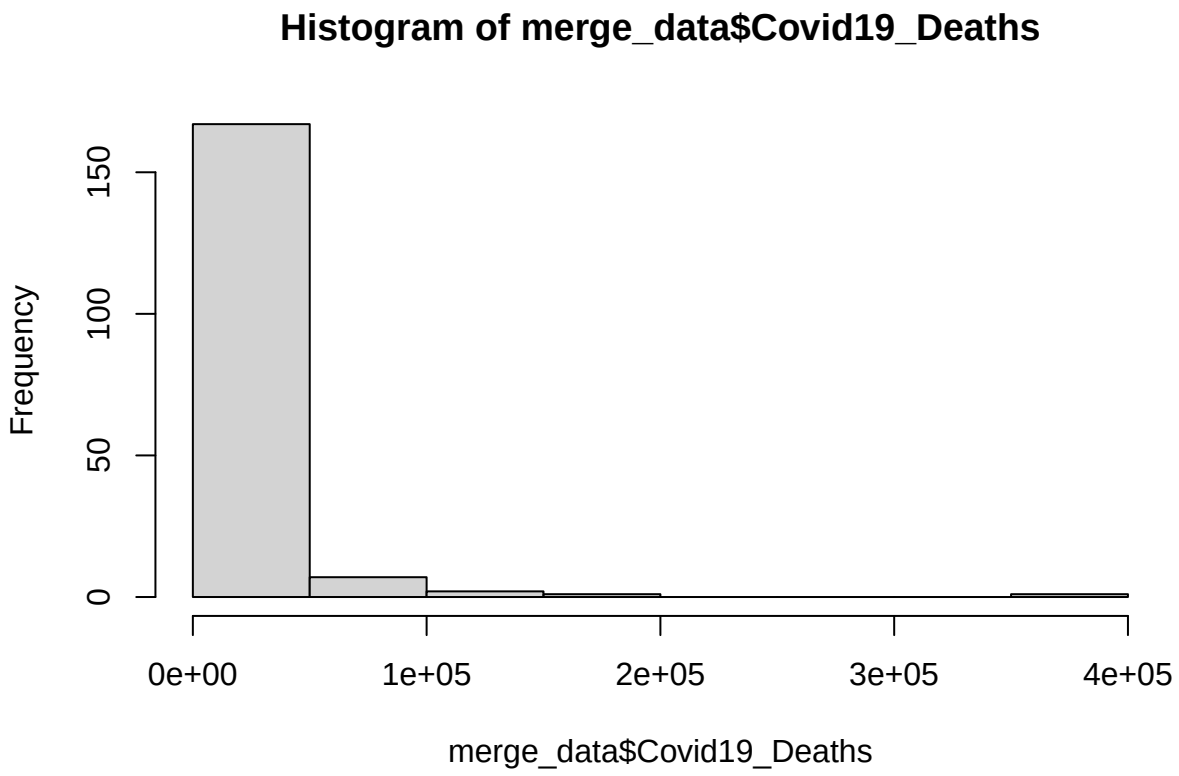
Data Skewness

By checking the histogram graphs or distribution plots, we are checking for data skewness. Following are the functions for plotting the graphs for 5 of the merged data variables. Migrant_net, Urban population and median age have normally distributed data with normal skewness but Covid19 Deaths and Fertility rate have right skewness which needs to be fixed.

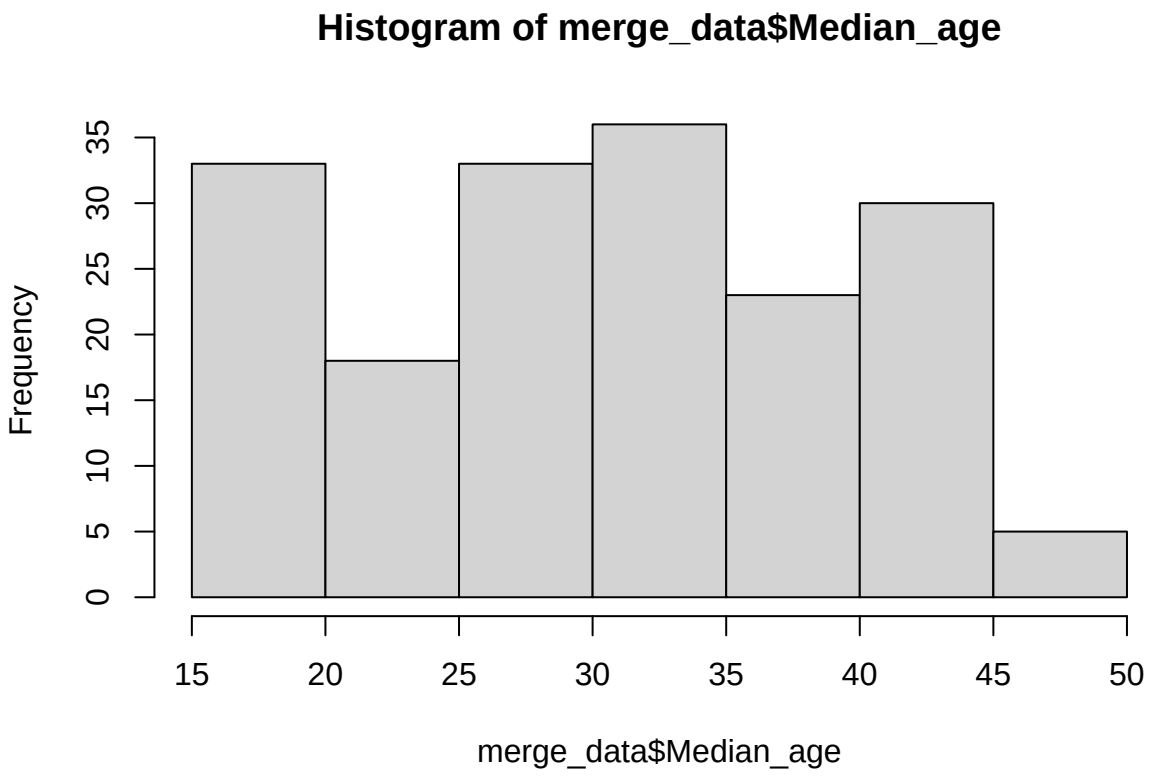
```
hist(merge_data$Migrants_net)
```



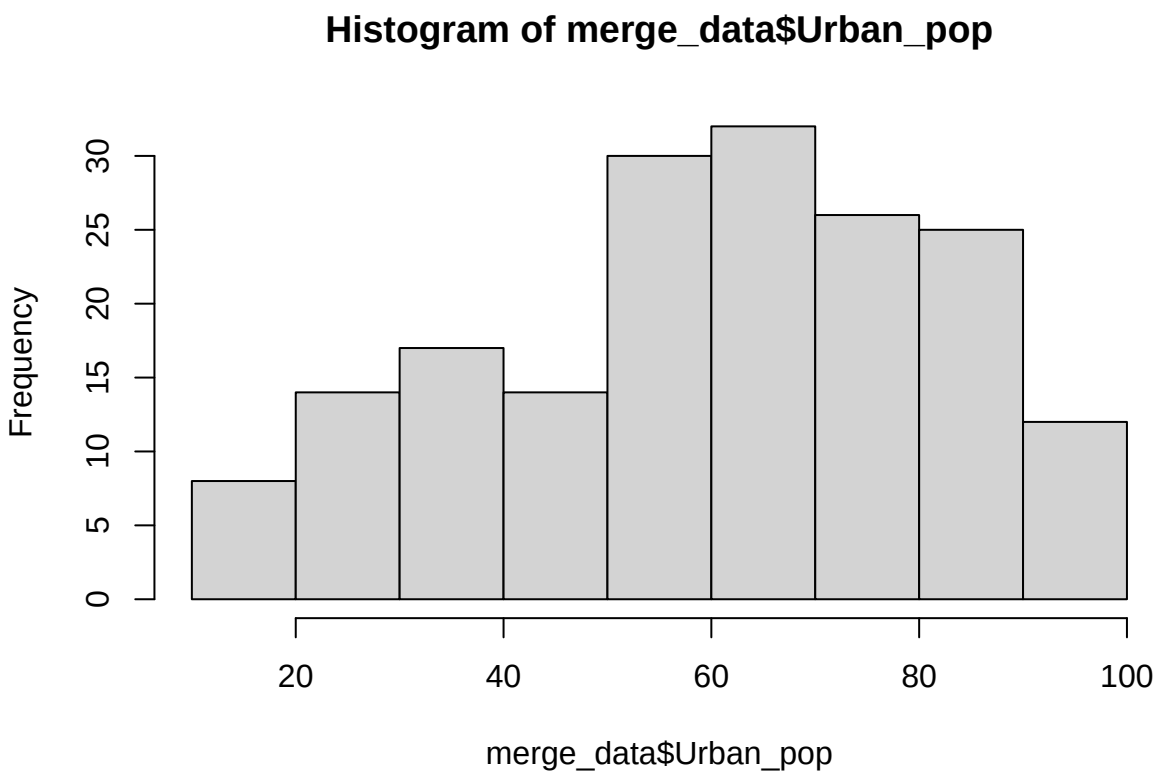
```
hist(merge_data$Covid19_Deaths)
```



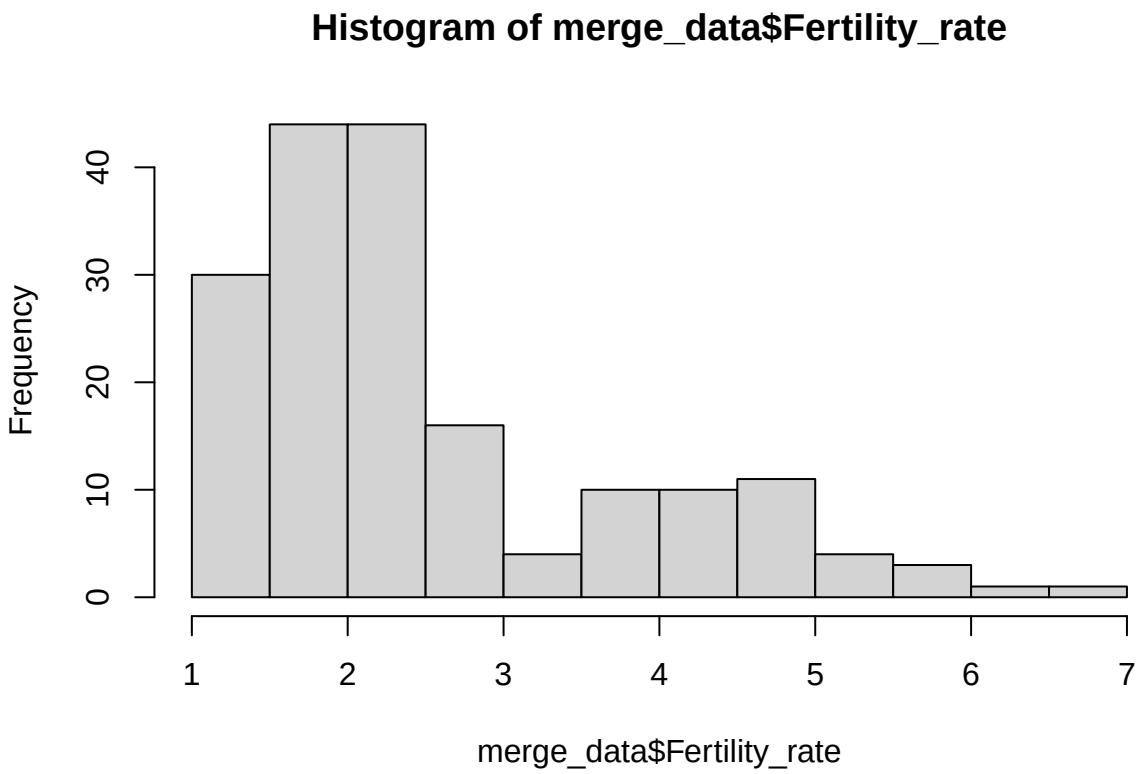

```
hist(merge_data$Median_age)
```




```
hist(merge_data$Urban_pop)
```

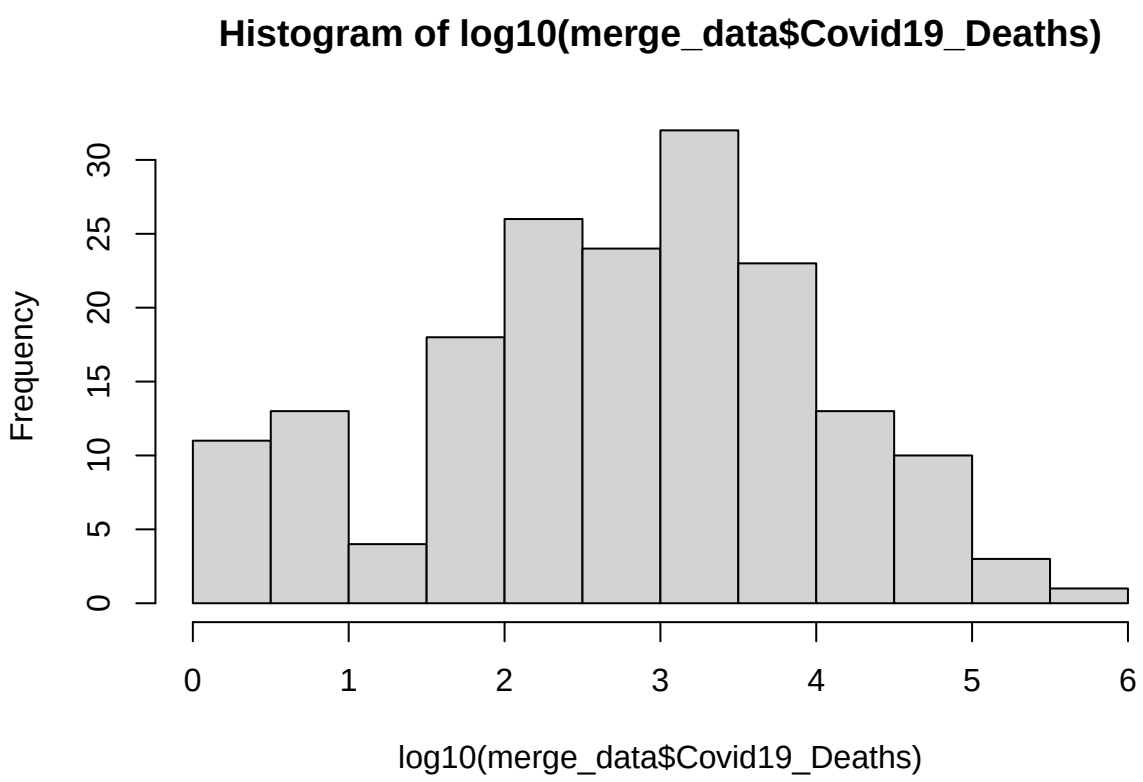



```
hist(merge_data$Fertility_rate)
```

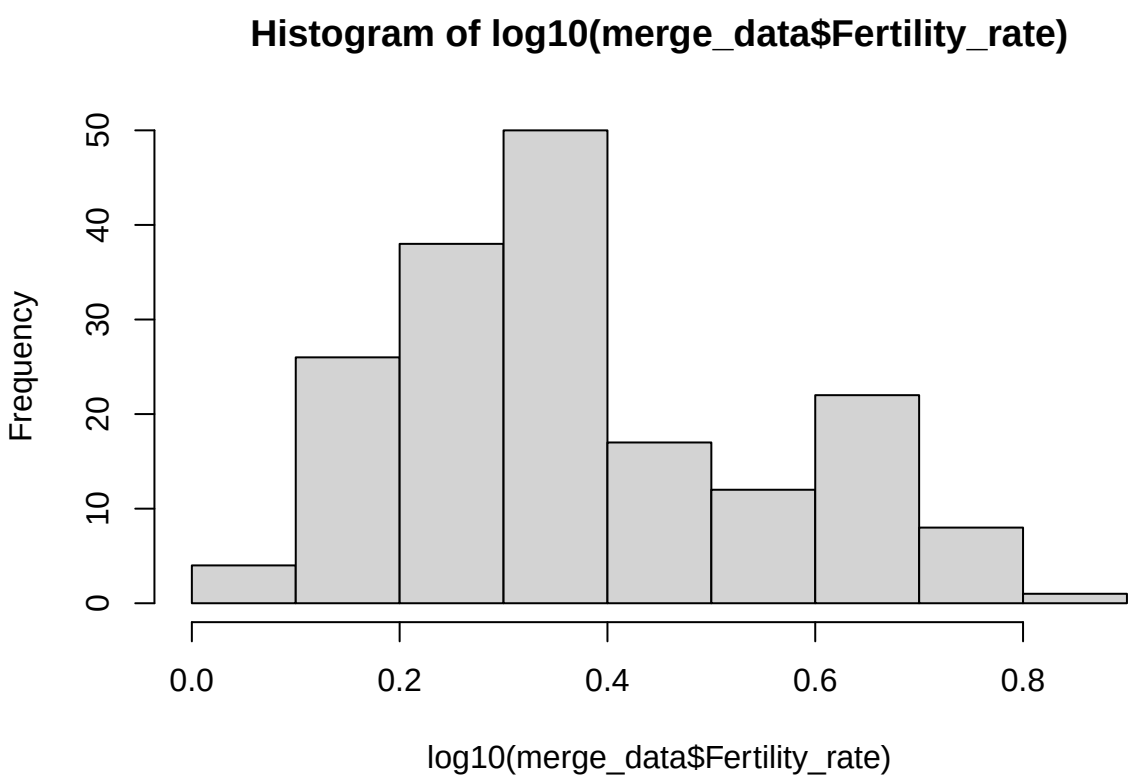


Transformation: A log transformation is a process of applying a logarithm to data to reduce its skew. This is usually done when the numbers are highly skewed to reduce the skew so the data can be understood easier. Log transformation in R is accomplished by applying the log() function to vector, data-frame or other data set.While log functions themselves have numerous uses, they can be used to format the presentation of data into an understandable pattern. They are handy for reducing the skew in data so that more detail can be seen. In R, they can be applied to all sorts of data from simple numbers, vectors, and even data frames. The usefulness of the log function in R is another reason why R is an excellent tool for data science. To fix the skeweness of right hand side data we need to raise the values of variable by using log of 10 function.

```
hist(log10(merge_data$Covid19_Deaths))
```




```
hist(log10(merge_data$Fertility_rate))
```



outlier:

An outlier is defined as an observation that stands far away from the pattern of the other observations. An outlier deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. As we can see in the following boxplot for road injuries shows some unusual numbers from other observations, this variable has 16 outliers as shown in the box plot. For fixing these outliers, I have used mean function as mean gives the average of all the values. After checking for road injuries variable, we found 2 high outliers and few medium outliers. The 2 high outliers were causing skewness in the data and got replaced by the mean. After plotting again, the outliers are no longer shown.

```
merge_data$Road_injuries %>%
  boxplot(main = "distribution Plot of Road Injuries", ylab = "Road Injuries", col = "grey")
```

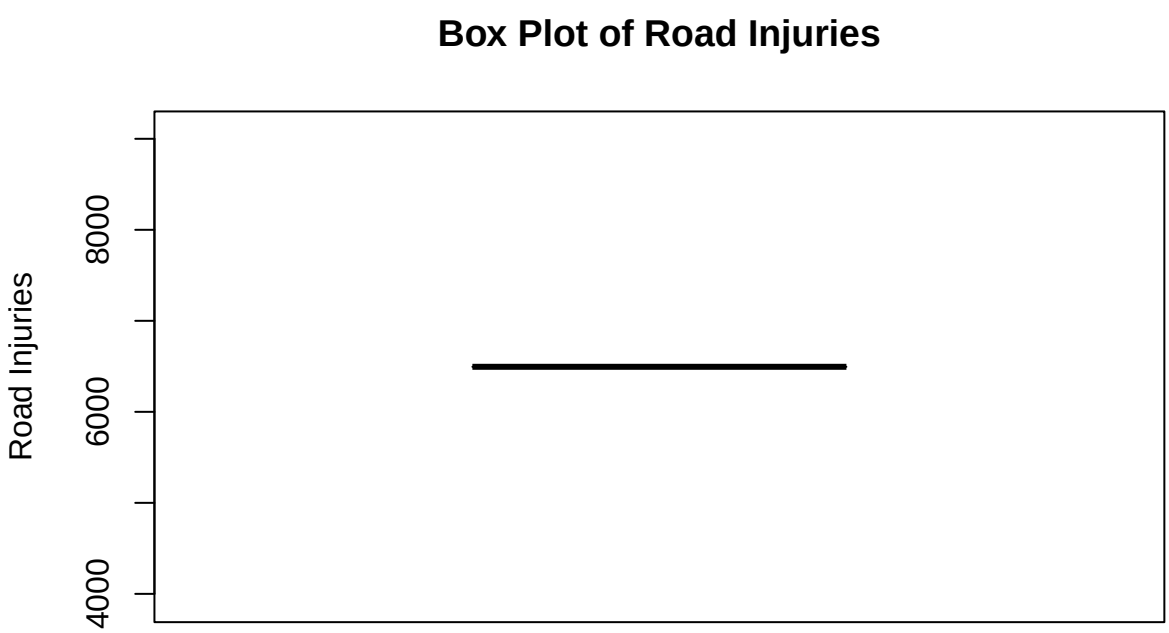

```
outliers_Road_injuries <- boxplot(merge_data$Road_injuries, plot=FALSE)$out

outliers_Road_injuries
```

```
##  [1]  44529 250025  29490 211975  37004  21122  21103  18508  19348  20867
## [11]  21316  19239  19541  41362  24153  11717
```

```
merge_data$Road_injuries <- mean(merge_data$Road_injuries)

merge_data$Road_injuries %>%
boxplot(main = "Box Plot of Road Injuries", ylab = "Road Injuries", col = "grey")
```



Reflective journal:

The most tricky part of this assignment was picking datasets. I looked up for 4 to 5 topics for my report like weather, crime rate, covid19, etc.... but as per requirement I can't find the 2nd dataset which I can correlate and merged by common variable with first dataset. Finally I found the dataset from Kaggle website about countries population of year 2020 and the other one is Death cause's by countries. There are so many columns in my first data set which explains death cause's but this information looks too much for my report. What I end up doing is I eliminating some of the unnecessarily columns from dataset 1. When Tidying up and manipulating dataset 2, I wanted to create a new variable rural population out of urban population to explains little more about the population density in every country but the problem I have faced is the urban pop were in percentage format and I was not able to extract it since it was a character variable with a % sign in it. So I googled it and find out the gsub function to remove the percentage sign. After using it, I have extracted the Rural values and I added the % back to both the columns, rural and urban population. Next road block that I encountered was that I can't find NA's for character vector. I then realized that I have to convert the character vector into numeric vector which I have learned during my course. After little digging from course module, I found the vocareum link for this task at hand. Another lesson I learned during this assignment was when I was detecting outliers for my dataset.It is a little tricky topic to understand so I went through all the course material and googled certain websites for more detail about it.

After this third assignment, concepts are more clear in my mind now and I can explain more about my report in comprehensive way and can use code in an efficient & meaningful manner. I am now confident enough to use R markdown for making reports from different datasets.

References:

MAJYHAIN, Feb 2022,Death Cause by Country,15th April 2022, kaggle, [id]:https://www.kaggle.com/datasets/majyhain/death-cause-by-country

Mohamed Fadl, 2020, Population by Country - 2020, kaggle,15th April 2022,[id]:https://www.kaggle.com/datasets/eng0mohamed0nabil/population-by-country-2020

Hannah Ritchie & Max Roser, Feb 2018, Causes of Death, [id]:https://ourworldindata.org/causes-of-death

Presentation Link:

https://www.loom.com/share/02494a1abff44c22a4b56afa961e65f4