MSCI 718 Pair Assignment 2

Group 27
Anson Ma (#20916612) & Tahmid Bari (#20864394)

Data

In this assignment, we are analyzing the data set from the World Health Organization (WHO) which contains WHO-generated estimates of Tuberculosis (TB) mortality, incidence (including disaggregation by age, sex, risk factors, HIV status, rifampicin resistance), case fatality ratio, treatment coverage, proportion of TB cases that have rifampicin-resistant TB, and latent TB infection among children aged under 5 years.

The default data set comes with 4,272 observations and 50 variables. Each variable has a column, and each observation on those variables has one row. Variables in the data set are selected and filtered for our analysis. There are 215 observations and 4 variables.

```
country, the name of country or territory
year, the reported year of the data. This variable is filtered as "2019"
e_pop_num, the estimated total population number
e_inc_100k, the estimated incidence (all forms) per 100 000 population
```

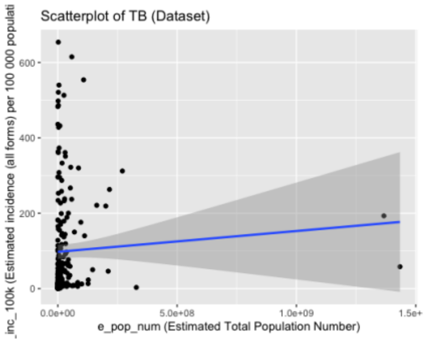
With the year filtered as "2019", the summary statistics of the variables are as follows.

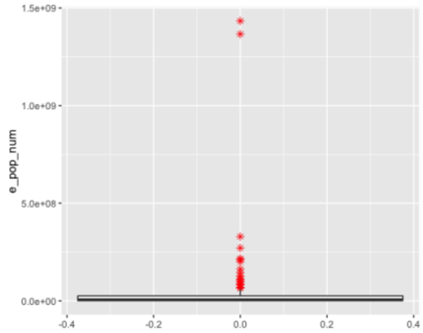
```
$ country   : chr [1:215] "Afghanistan" "Albania" "Algeria" "American Samoa" ...
$ year      : num [1:215] 2019 2019 2019 2019 2019 ...
$ e_pop_num : num [1:215] 38041757 2880913 43053054 55312 77146 ...
$ e_inc_100k: num [1:215] 189 16 61 2.1 7.5 351 22 0 29 26 ...
```

```
  country             e_pop_num           e_inc_100k
Length:215         Min.   :1.330e+03   Min.   :  0.00
Class :character   1st Qu.:8.168e+05   1st Qu.:  9.95
Mode  :character   Median :6.777e+06   Median : 41.00
                   Mean   :3.575e+07   Mean   : 99.43
                   3rd Qu.:2.543e+07   3rd Qu.:138.50
                   Max.   :1.434e+09   Max.   :654.00
```

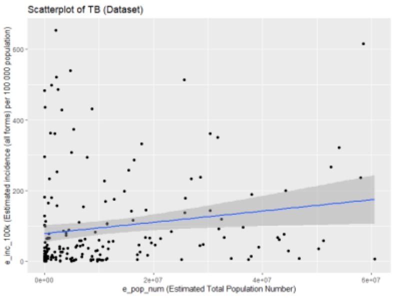
`e_pop_num` (Estimated total population number), e_inc_100k(Estimated incidence (all forms) per 100 000 population) are numerical values, currently stored as a double. We want this to be an integer value, so the dataset is reloaded to explicitly parse those values as an integer. Moreover, we checked for "NA" values in the data set and it returned zero.

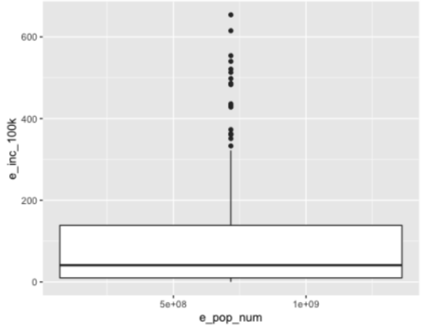
After tidying and cleaning the data set, the scatterplot is generated in this figure.





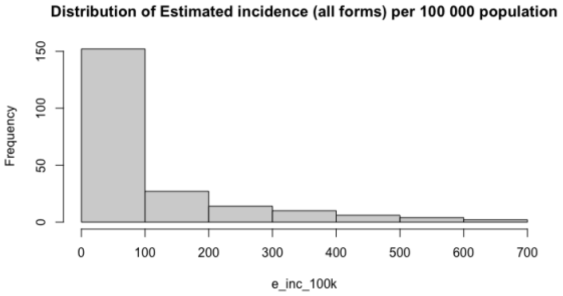
As seen from the plots, there are 2 outliers. Since they distort the distribution of the data to a certain extent, they are removed accordingly. The new plots are generated as follows.

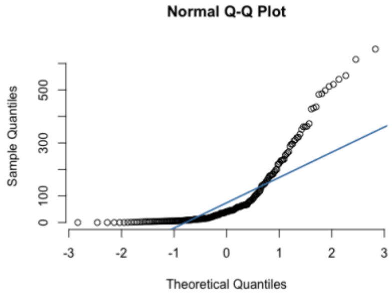




Planning

We plan to analyze the possible relationship between e_pop_num, e_inc_100k. Prior to that, we check for the normality of e_inc_100k. We visually looked at 2 different graphs to see if our data is normal: histograms and Q-Q Plots.

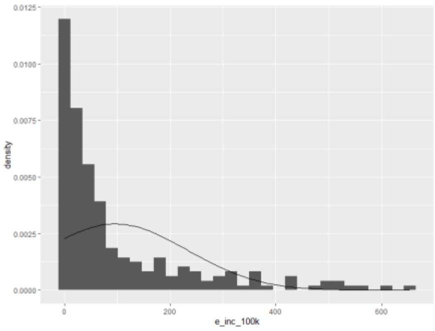




Another approach is to test for signifiers of non-normality through skewness and kurtosis.

```
skew.2SE:  5.735362
kurt.2SE:  4.753397
```

Because ["skew.2SE"] is not between -1 and 1 at 5.735362, we conclude that the skewness for `e_inc_100k` is different from 0 (at the 95\% level of confidence) with positively skewed distribution. Also, because ["kurt.2SE"] is greater than 1 at 4.753397, we can also conclude that the kurtosis for `e_inc_100k` is different from 0 (at the 95\% level of confidence) with positive kurtosis. Since the data is not normal, we further carried out Shapiro-Wilk normality test to check if data transformation is required. The results is shown as follows.



Since the data is not normal, and positively skewed, data transformation is needed. Upon transformation, the data is normally distributed. The results are included in the Appendix II. The data is fairly normally distributed based on the histograms after the data transformation.

Analysis

The hypothesis for this analysis is that e_pop_num, e_inc_100k have a positive correlation. From the aforementioned scatterplots before the data transformation, this is quite the case when the 2 outliers (China and India, both have an exceptionally high population) are taken out. In the next assignment, the assumptions will be tested, and correlation analysis will be conducted to further examine the possible positive relationship between the 2 variables.

Conclusion

In this assignment, it can be preliminary concluded that there is a possible relationship between the estimated total population number and the estimated incidence (all forms) per 100k population. Based on the above, the estimated incidence (all forms) per 100k population follows a positively skewed distribution with positive kurtosis. Data transformation is therefore required and performed to prepare for the next analysis.

Appendix I – Contributions of each group member

Anson is responsible for the research and checking of the analysis and writing up the report. Tahmid is responsible for the working and analysis on RStudio. Internet resources have been used as a guide to structure the analysis and the report.

Appendix II – Working file

A separate .Rmd file has been attached showing all workings.