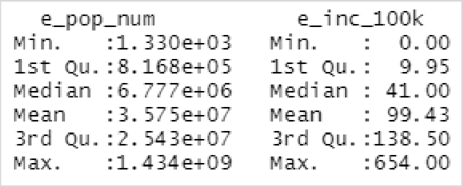
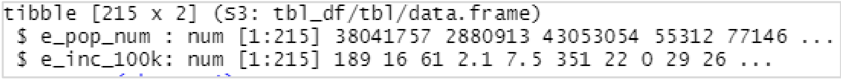
## Data

From pair assignment – 2, we analysed 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).

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. With the year filtered as "2019", we have taken 'Estimated Total Population Number' and '(Estimated incidence (all forms) per 100 000 population' as our correlation variables. From these two variables, we can find out the percentage of incidence of TB over the total population in a specific year.

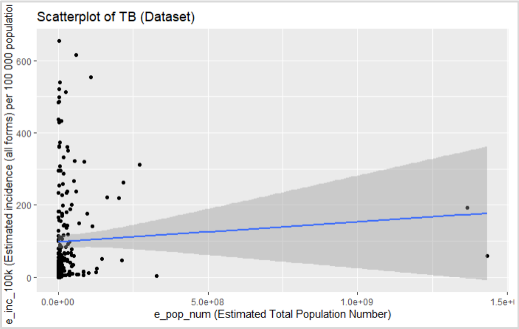


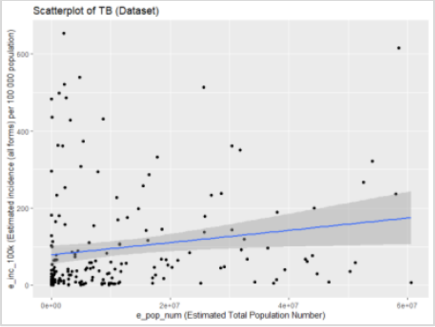
```
   e_pop_num             e_inc_100k    
 Min.   :1.330e+03   Min.   :  0.00  
 1st Qu.:8.168e+05   1st Qu.:  9.95  
 Median :6.777e+06   Median : 41.00  
 Mean   :3.575e+07   Mean   : 99.43  
 3rd Qu.:2.543e+07   3rd Qu.:138.50  
 Max.   :1.434e+09   Max.   :654.00  
```



```
tibble [215 x 2] (S3: tbl_df/tbl/data.frame)
 $ 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 ...
```

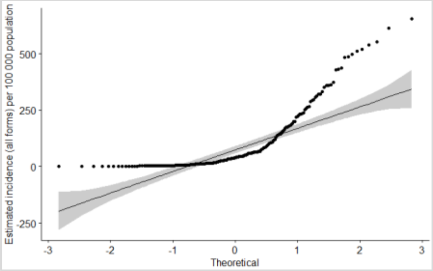
`e_pop_num` (Estimated total population number), e_inc_100k (Estimated incidence (all forms) per 100 000 population) are numerical values, currently stored as a double. For an integer value, the dataset is reloaded to explicitly parse those values as an integer. Moreover, we checked for "NA" values in the data set and it returned zero. As seen from the plots, there are 2 outliers. Since they distort the distribution of the data to a certain extent, they are being removed accordingly. Outlier plots are generated as follows:

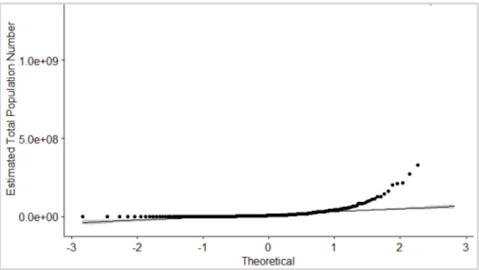




## Planning

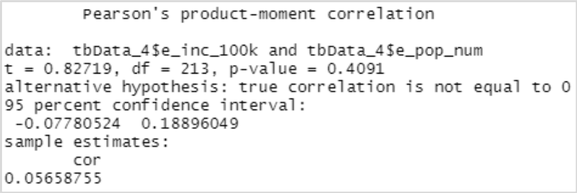
We plan to analyse the Pearson's R between **e_pop_num** and **e_inc_100k** because it is a measure of the linear relationship between two continuous random variables. It does not assume normality although it does assume finite variance and covariance. Prior to that, we check for the normality of 'e_inc_100k' and 'e_pop_num'. We visually looked at 2 different graphs and found that the data is not normally distributed.



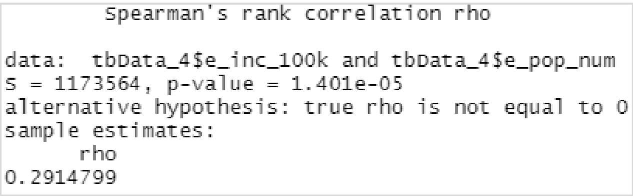


We have seen that the variables are not normally distributed or the relationship between the variables is not linear, it will be more appropriate to use the Spearman rank correlation method. However, we will analyse the result of the Pearson's r test as well since it measures the linear relationship between two continuous random variables. The distribution of either correlation coefficient will depend on the underlying distribution, although both are asymptotically normal because of the central limit theorem.

**Analysis**



```
        Pearson's product-moment correlation

data:  tbData_4$e_inc_100k and tbData_4$e_pop_num
t = 0.82719, df = 213, p-value = 0.4091
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07780524  0.18896049
sample estimates:
       cor
0.05658755
```

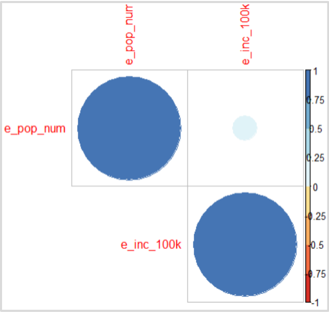


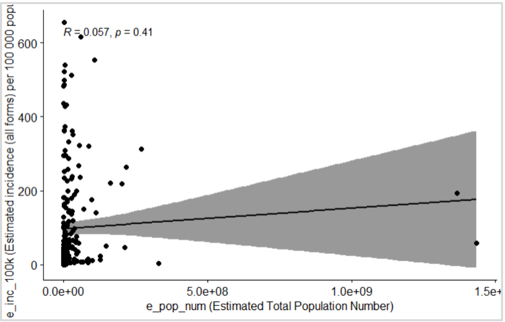
```
        Spearman's rank correlation rho

data:  tbData_4$e_inc_100k and tbData_4$e_pop_num
S = 1173564, p-value = 1.401e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2914799
```

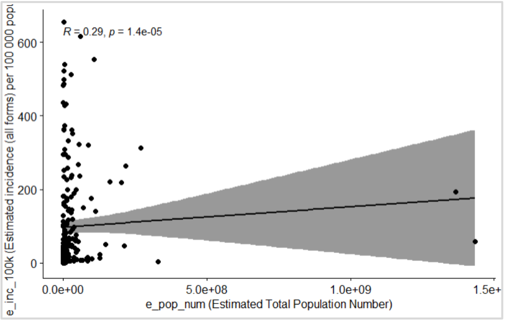
<u>In the (Pearson's R) result above:</u>
t is the t-test statistic value (t = 0.82719)
df is the degrees of freedom (df= 213) | p-value is the significance level of the t-test (p-value = 0.4091)
conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [-0.07780524, 0.1889604])
'sample estimates' is the correlation coefficient (Cor.coeff = 0.05658755) that is a moderate positive relationship.

<u>In the (Spearman's rho) result above:</u>
Correlation coefficient between 'e_inc_100k' & 'e_pop_num' are 0.2914799 that is closer to zero, the weaker the association between the ranks
p-value is 1.401e^{-05}

First of the below plots are showing the lower and the upper triangular part of a correlation matrix. In this plot, correlation coefficients are coloured according to the value. Correlation matrix can be also reordered according to the degree of association between variables. The second plot is the scatterplot of Pearson's R test and at the end, showing another scatterplot of Spearman's rho plot.







Bibliographic Reference:

Devinyak, O. (2013, April 13) Retrieved from https://www.researchgate.net/post/Which-correlation-coefficient-is-better-to-use-Spearman-or-Pearson

Frost, J. (2021, February 14) *Interpreting Correlation Coefficients* Retrieved from https://statisticsbyjim.com/basics/correlations/

(Anonymous, 2021) *Correlation Analyses in R* Retrieved from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

(Anonymous, 2021) *Correlation Analyses in R* Retrieved from http://www.sthda.com/english/wiki/correlation-analyses-in-r