Tuberculosis (TB) is an infectious disease usually caused by the bacterium Mycobacterium tuberculosis (MTB). It generally affects the lungs, but can also affect other parts of the body. Now as part of this report, we will explore the data sets provided by WHO and conduct some analysis to check for the factors associated with TB.[1]

Data

World Health Organisation has provided a range of downloadable data sets which are categorized into two section: Data reported by countries and territories to WHO and also the estimation of burdens provided by WHO based on the previous data. For ease of use, we have also been provided by a data_dictionary. After overviewing the available data sets, we have decided to focus mainly on **TB Case notification data**. The reason is the availability of the variables in this data set which we believe are associated with TB.

TB notification Data set has **164 variables**, but we are interested to get an idea of the number of TB cases observed over the recent years globally. The data set has a variable, *g_whoregion* which groups the countries into different regions as decided by WHO. The data set contains **2144 observations** and we have created a subset with the variables used for further analysis. As we are interested to know more about the factors attributed to TB disease, we have observed the risk factor variable from data dictionary which lists the following: *Alcohol; Diabetes; HIV; Smoking and Undernourishment.* On further reviewing data sets provided by WHO, we observed that HIV has more information among all the other factors. Therefore, we create our subset data focusing on HIV and TB cases.

The data set for the analysis contains two **categorical variables**: country (with 216 levels) and WHO_region (with 6 levels as : Africa – AFR, America – AMR, Eastern Mediterranean Region – EMR, Europe – EUR, South East Asia – SEA, Western Pacific – WPR). Apart from that , we have Year as interval variables. Also we have the following **integer variables**:

- *TB_Cases*: Total of new and relapse cases and cases with unknown previous TB treatment history.
- *HIV_Cases*: Total number of people registered as HIV-positive regardless of year of diagnosis.
- *HIV_TB*: TB patients (new and re-treatment) recorded as HIV-positive. There are total 1651 missing values in the data set.

Cleaning the Data Set: As the background of the data collection is not known we can not replace NA'S with 0 because there are observations which has value 0. Hence we have removed the records with missing values in all three variables and the count of NA's has been reduced to **1489 observations**.

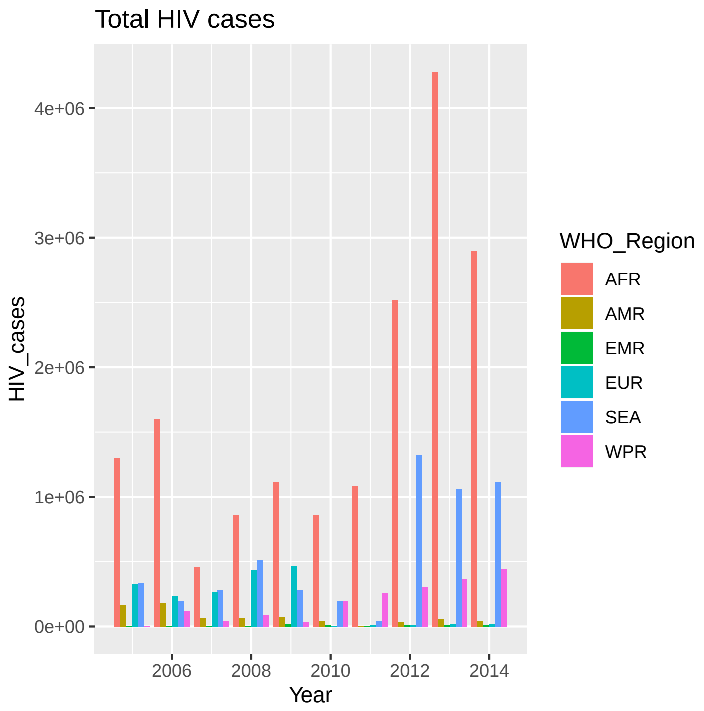
During our initial analysis we have observed that the total no of HIV_Cases are concentrated around year 2005 to 2014. Hence, we focus our analysis in the observed time range by filtering our data set.

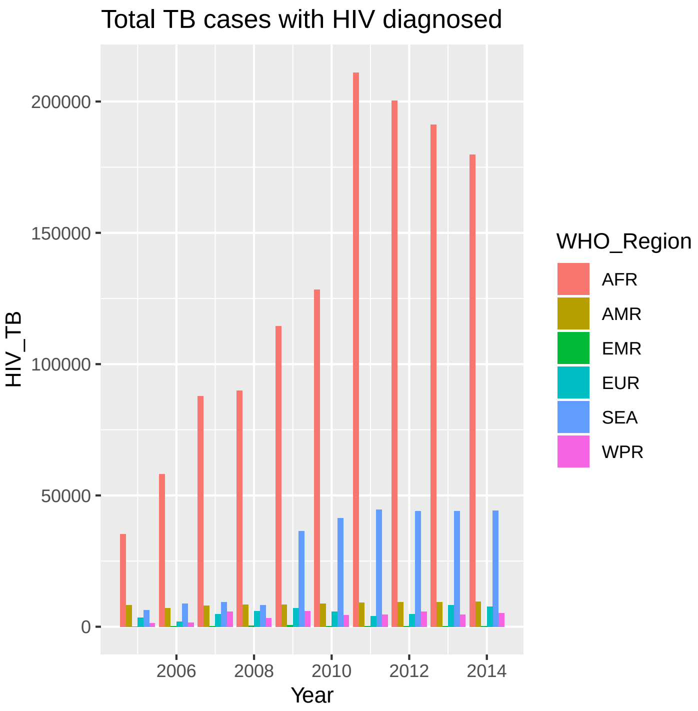
Planning

It will be good to have the idea of the number of total HIV_Cases across the year among the WHO decided regions. It is interesting to observe that AFR has significantly more number of HIV cases compared to any other region and AMR along with EUR have the lower number of total cases. Also we have noticed that after year 2012, HIV cases has suddenly began to rise in SEA.

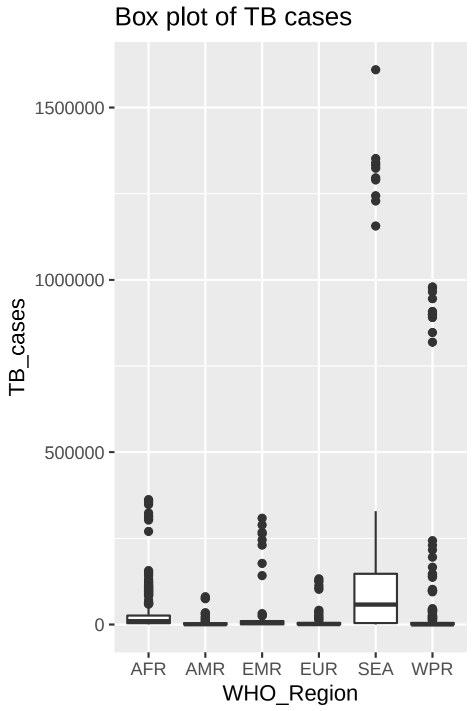
While preparing the report as we look through TB risk factors information, we noticed that *it is a general idea among people that people living with HIV are more likely than others to become sick with TB.* Considering this general idea as our prior hypothesis, we will check if the hypothesis has a statistical significance, by analysing our data set from the year 2005 to 2014.

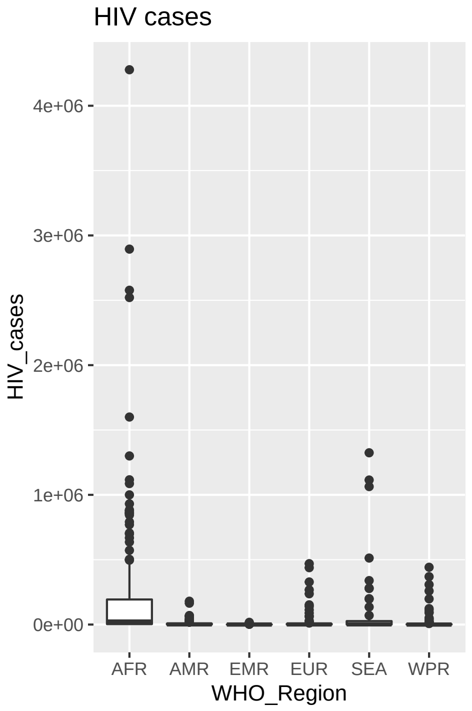
As the first step, we may want to explore the number of TB patients who also have diagnosed with HIV. From the below graph, it is not surprising to see that AFR has recorded the maximum number of TB patients who are also HIV patients. Also SEA recorded apparently higher number of these type of cases compared to EUR and AMR.

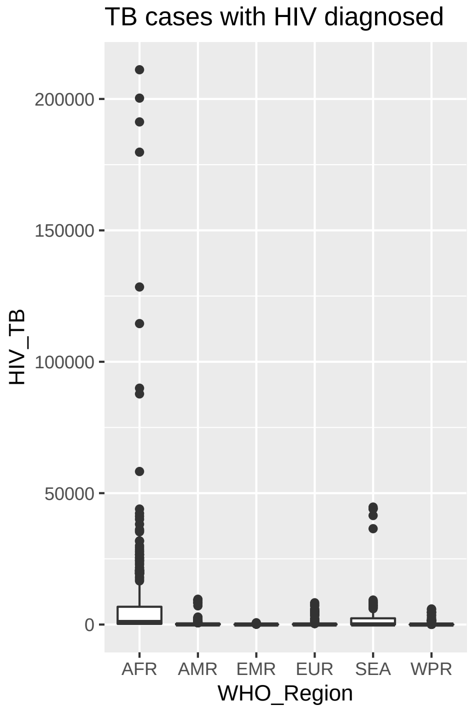




Now let's dig deeper into the data! As we observed the quantitative variables, we can notice that all the three variables: TB_Cases, HIV_Cases and HIV_TB cases have **outliers**. Now next question arise : *What are we going to do with the outliers?*





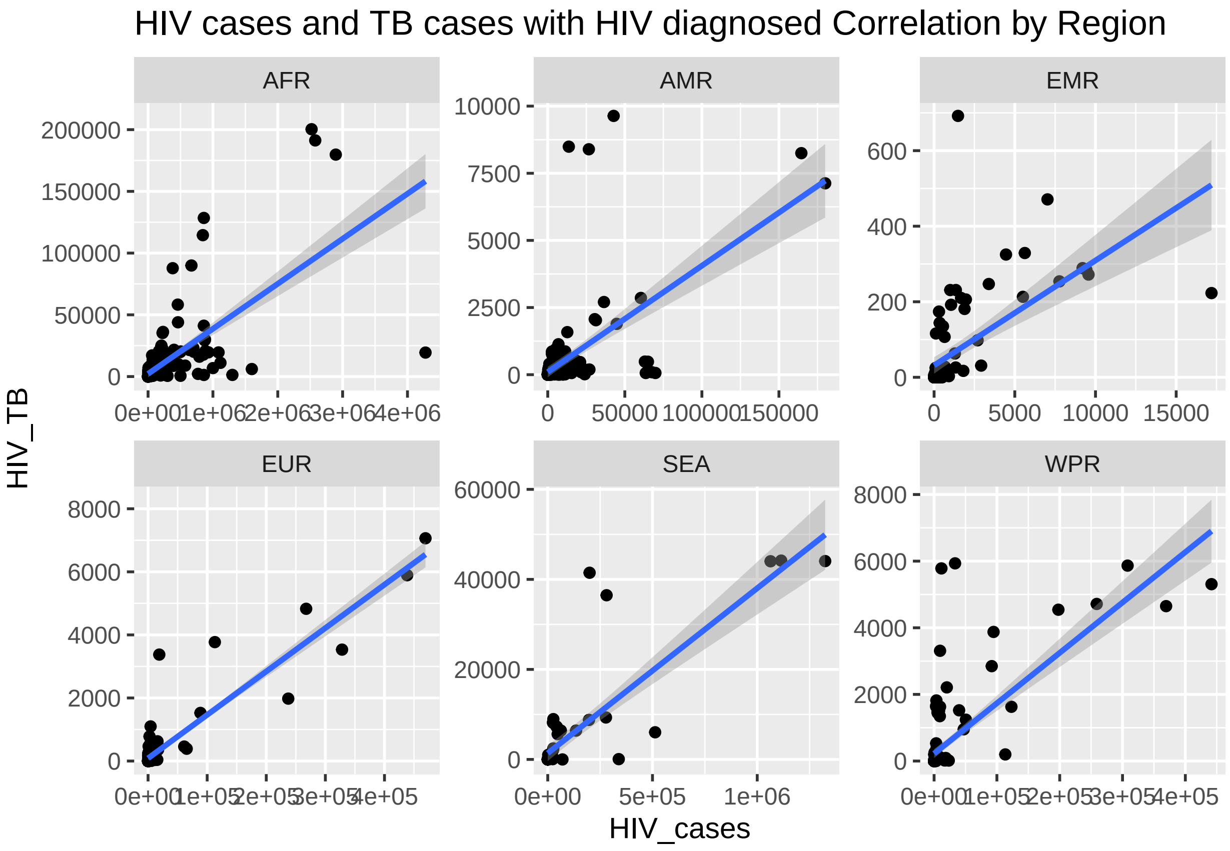


One possible reason for these outliers can be the population increase in some countries are higher than others or it can be possible to have data entry error. As we do not have prior information about how data has been collected and stored , our assumption will not be a good reason to exclude these outliers. Also if we exclude the outliers for TB_Cases using the Inter Quartile Range as thresholds, we will loose almost 17% of our data which will certainly impact our further analysis in the next step. Considering all these factors, here *we will not remove the outliers.*

Analysis

Now it's time to measure the extent to which total number of HIV cases and total number of cases of TB patients who are also diagnosed with HIV are related to verify our prior hypothesis. We will use the **correlation coefficient** to measure the pattern of responses across variables.

Scatter plot between two variables of our consideration is a good start. So, at first, we draw scatter plots for all the WHO regions and the regression line suggests that we do have some **positive correlation** between our two variables.



Next step, we want to conduct the test for coefficient value. For that we will have the following hypothesis as below:

- *Null hypothesis*: The correlation coefficient is not significantly different from 0, there is not a significant linear relationship (correlation) between TIB_cases and HIV_TB cases.
- *Alternative hypothesis*: The correlation coefficient is significantly different from 0, there is a significant linear relationship (correlation) between TIB_cases and HIV_TB cases.

Next we have checked whether the assumptions which are applicable for **Pearson Correlation Test** satisfies our data set. As our data set, each observation records are independent, the data is interval and also the sample sizes are larger than 20. Based on all these, we have decided to do the Pearson Test with **0.05 significance level**. As data is interval, there is no need to test the normality in this case. This test will remove data case wise (i.e., entire rows of data for any missing data).

```
##           HIV_cases    HIV_TB
## HIV_cases 1.0000000 0.7009763
## HIV_TB    0.7009763 1.0000000
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  data1$HIV_cases and data1$HIV_TB
## t = 28.991, df = 870, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6655603 0.7332425
## sample estimates:
##       cor 
## 0.7009763
```

As the values of correlation coefficient between the total number of HIV cases and total number of TB cases who are also diagnosed with HIV are **greater than 0.7** and also the Correlation test shows the **p value is < 0.001**, we can conclude that total HIV cases is **significantly correlated** with total TB cases who are also diagnosed with HIV with 95% confidence level.

Conclusion

As part of this report we have explored the data sets provided by WHO for TB from different countries and after careful consideration we have selected HIV as one of the risk factor associated with TB. We are inferring the definition of variables based on information provided in the data_dictionary data set.

Further we formulated a hypothesis based on our general knowledge to test if HIV cases are correlated with TB. We were able to conclude that total HIV cases are significantly correlated with total TB cases who are HIV diagnosed at 95% confidence level.

For future work we can use the given data set to analyse if there is any pattern of TB and HIV case incidences among different age groups. This analysis can be useful to spread the awareness among HIV patients about TB.

References:

[1] N. Suresh, K. Arulanandam. “A Study On Tuberculosis Analysis Using Data Mining Techniques”. IJARCCE Vol. 7, Issue 3 (2018).