Primary data set – EDA:

* We have 296 rows which means 296 countries information, which means there is information for both countries and sub regions in the countries. So we are filtering the rows which only have country code(which means the rows which have country codes are countries and rest all are sub-regions)
* Removed the rows where there are missing values in the target variable. As algorithms like KNN cannot handle(to my knowledge).
* After doing the above steps we are now ended up having 196 rows (=countries).
* Now, refactored the country and country\_code columns.
* There are some missing values, as the boxplots between those columns showed there are some outliers. Replacing the missing values with median of their respective columns.
* Have looked into the correlation matrix and correlation plot. It seems there is almost no correlation between the predictor values('poverty\_gap','public\_health\_exp','malnutrition\_death\_rates','Infant\_mortality\_rate','GDP\_per\_capita','annual\_health\_care\_per\_capita') to the target -'homicide\_Rate'.

Merging the data to have the first and second secondary datasets:

After downloading the data from our source, we had to merge them to have tidy datasets to work with. I loaded the data into my workspace. After checking that everything was imported properly, I made some preprocessing like changing the columns names and filter them to have only the data from 2012. Then I made successive out join to have our two sample datasets. I made out join because our predicators had different sets of primary keys and we did not want to lose data at this step of the project. Then I exported it to excel format file.

Final First Secondary Dataset:

To fill the missing values of our first secondary data we assumed that the data from prevois years (before 2012) has not changed since the past registration. For the female\_manager for instance if we don’t have the value of 2012, we take the previous value registered. Afterwards, we have assumed that is fine to take the data from the direct next year (2013 or 2014).

I checked the amount of common countries between the predicators and the primary data. I figured that "Female\_senior\_and\_middle\_management\_position" predicator has very low amount of countries in common with the primary data, so I decided to get rid of it because it obviously would not help since it’s impossible to assign a mean or the median of values to the missing values because the predicator can change drastically from one country to another. We ended up with 116 countries in common with the primary data without missing values for "youth\_literacy\_female", "Women\_in\_parlements", "Firms\_with\_female\_top\_manager" that are our final first secondary predicators.

* EDA:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

As you can see we have some outliers around 58, Need to investigate for the presentation.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

As you can see we have some outliers around 58, Need to investigate for the presentation.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

As you can see we have some outliers around 58, Need to investigate for the presentation.

# Secondary EDA:

* Filtered out the rows which are not having any country code.
* Removed the rows in which there is NA’s in the target variable.
* After doing this, we have replaced the NA’s in government\_expenditure by its mean, since there are no outliers.
* Gross\_enroll\_primary indicator missing values are replaced by its median.
* Gross\_enrol\_secondary indicator missing values are replaced my mean under the assumption that the corresponding primary enrollment has pretty good ratio, and the min enrollment ratio is 80% and the mean of gross\_Enroll\_secondary(excluding NA) is also around 80. So under this assumption we have replaced by mean.
* Lower secondary completion rate is also replaced by its mean(excluding NA), because of the above assumption. Since enrollment is replaced by min 80% , completion rate would be less than or equal to that. Since mean is satisfying that condition considered this.