Datathon 5
Sushranth Hebbar

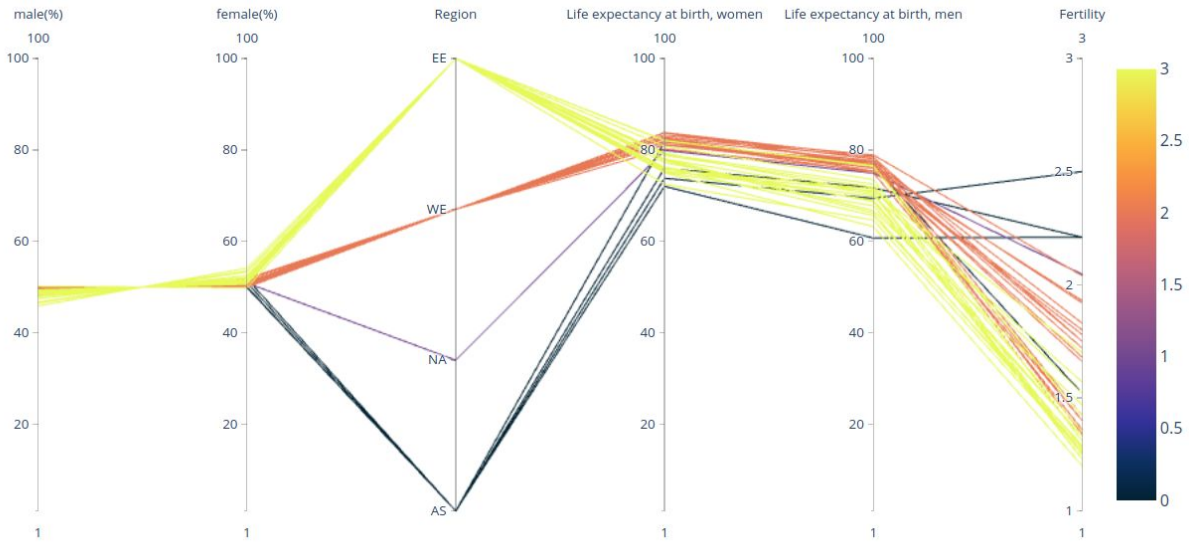
Introduction:- Given the tabular uneces-country dataset, use the data as is and visualise using parallel coordinates and scatterplot matrices. Then build hierarchical relationships based on time and space and use the treemap and sunburst visualisation.

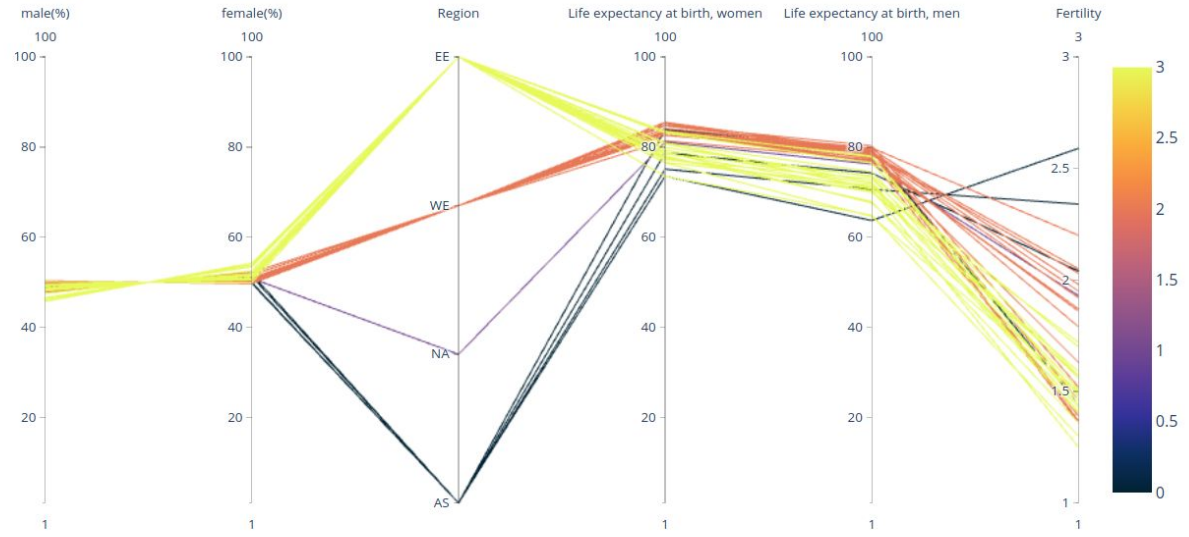
Methods:- The dataset has about 80 columns. The first step would be to select appropriate columns to use as variables for parallel coordinates and scatterplot matrices. So, the dataset had to be explored and variables which were similar to each other were picked (Details have been described below). The second step would be to identify variables which can be part of a hierarchical relationship.

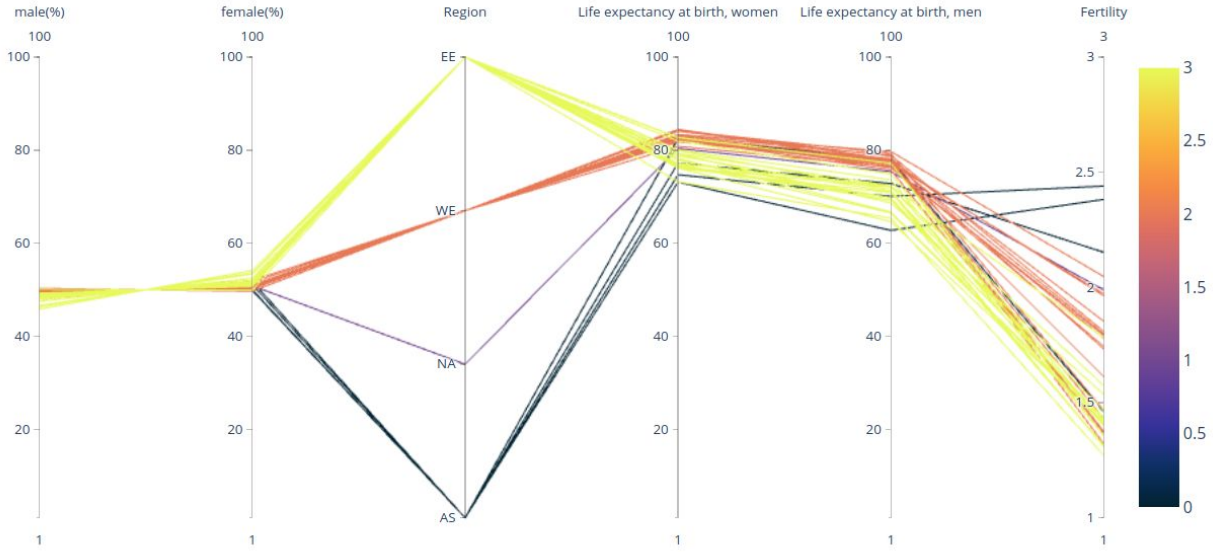
Exploration and Preprocessing:- I started by understanding the distribution of countries across continents. Majority of the countries were European followed by Asian countries and lastly two countries which belong to North America. I decided to further group the European countries into Western and Eastern Europe. Then a dictionary was created where the key values were the four regions mentioned above. The values were the list of countries that belong to that region. The next step was to handle the missing values in the dataset. If the number of missing values for a country of a particular variable/column is higher than a threshold, then that country is removed from the dictionary. If a lot of countries have been removed from the dictionary, then I will not include that variable as it has too many missing values for a majority of countries. Otherwise, I select another variable and repeat the process again. This way I select upto 7 variables. Once this dictionary is finalized, I handle the missing values across the selected variables by replacing them with the mean value of that variable for a specific country. Finally, I create a new dataframe containing these variables.

The process is almost identical for generating hierarchical relationships. The only difference here is that instead of selecting the variables based on empirical analysis, they are chosen only if they have a hierarchical relationship with each other. As an example a region can be further divided into countries. Then a color can be assigned to each of these countries. Then the color of the parents will be the weighted average of the color of their children. The color is usually a column that belongs to the dataframe. Ex: If the value associated with a node in the treemap is area, then the color of the node can be the population density of that node. The value of the parent node will be the sum of it's child values. The color of the parent will be the average of color values of the children weighted with respect to the values of the child nodes. In this case, it's the area.

EE: Eastern Europe, WE: Western Europe, NA: North America, AS: Asia. This is the terminology used for the region section in the below diagrams. Each line corresponds to a row in the data frame where the columns in the dataframe are the variables selected below. The color of the line represents the region to which it belongs. Each diagram is with respect to a particular time stamp. The images below are for the years 2004, 2010 and 2016 respectively.

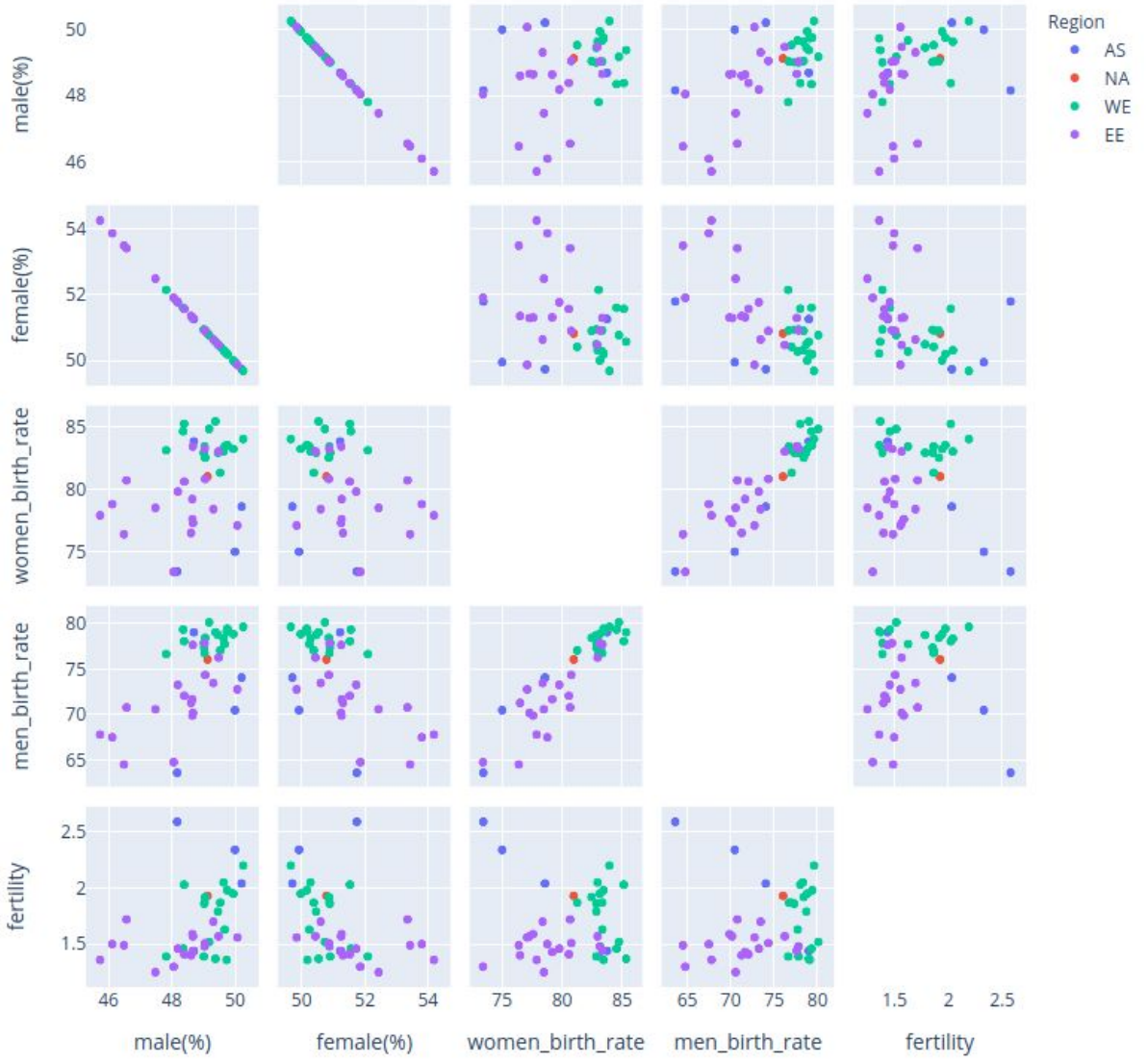




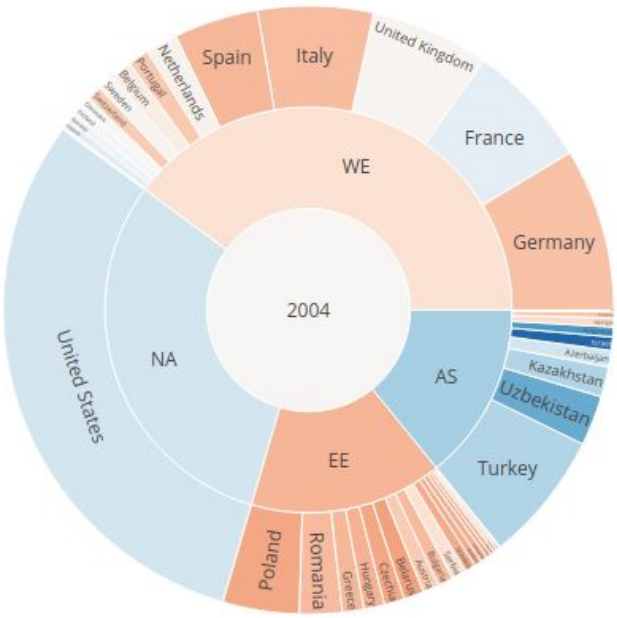


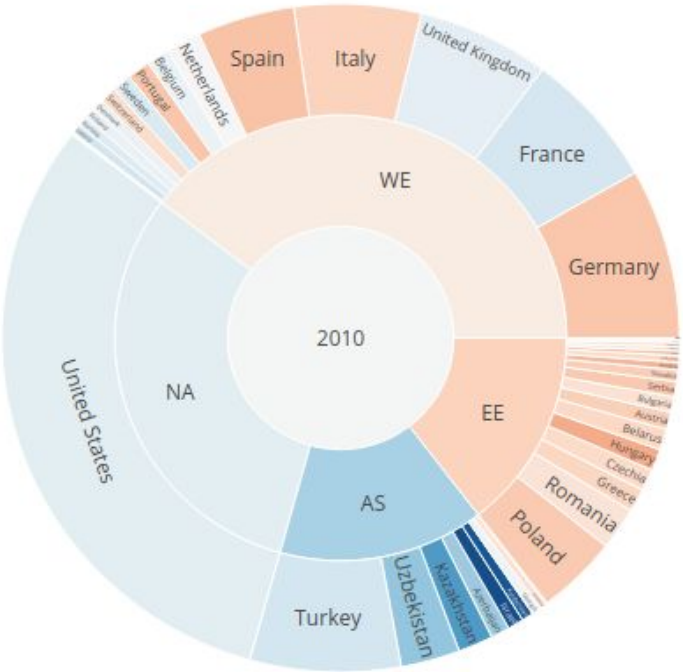
The obvious observation is that the life expectancy of women is higher than men across all regions. Let us focus on the red and yellow lines as they are higher in number. The red lines are closer to each other across all variables except for fertility. This means that the rows in the data frame are similar for these variables across countries in Western Europe. The yellow lines on the other hand are not very close. This could mean that the quality of healthcare is not uniform across Eastern Europe. It could be more advanced in some regions and still under development for other regions. But it is more uniform for Western European countries. The red lines are farther apart than the yellow lines when it comes to fertility. It is possible that there is more cultural diversity in the west than the east and this might influence the number of children in the household of these countries. Another point worth mentioning is that the life expectancy of men and women has slightly increased from 2004 to 2016 signifying the advancement of health care across the world.

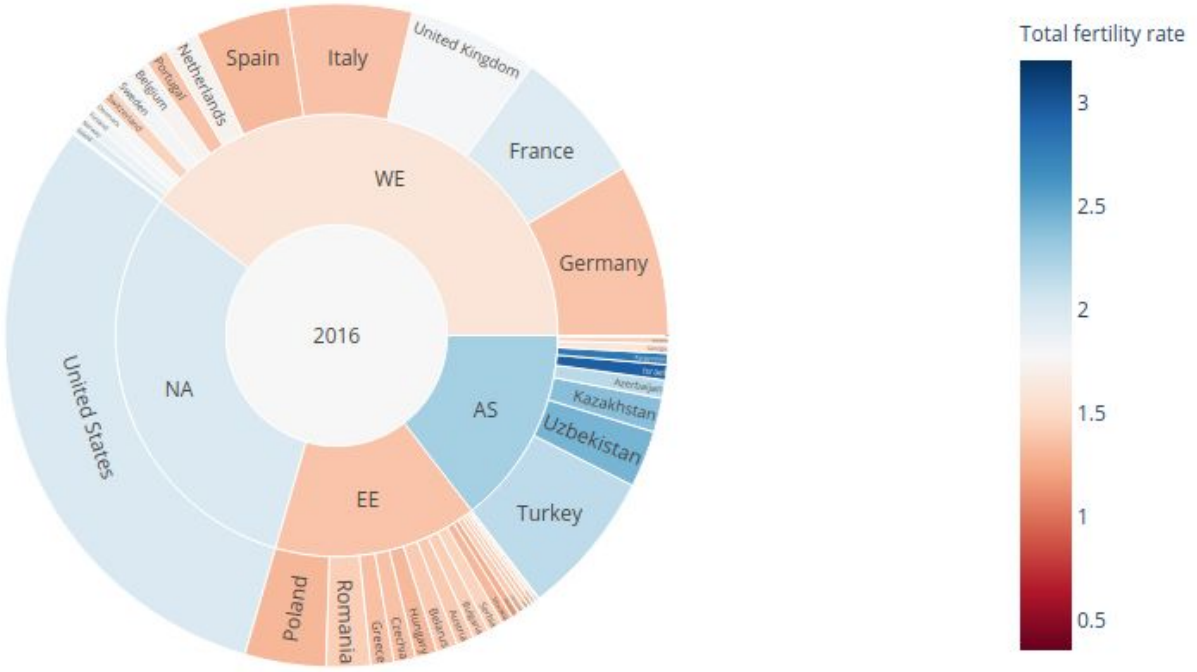
Consider the ScatterPlot Matrix for the year 2010. There does not appear to be any obvious pattern for the fertility row. This means that fertility is independent of those variables. It depends more on other factors like culture. Another observation is that the fertility rate is higher in Asian Countries compared to European and North American countries.



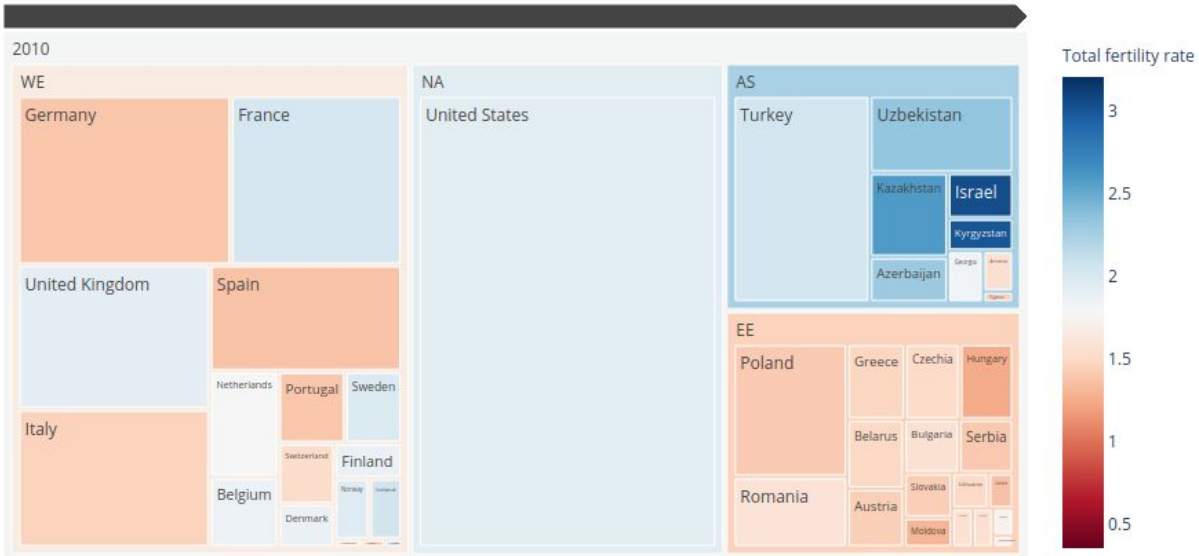
The hierarchical relationship is Region followed by country. The size of each node is determined by the total female population of that node. The color of each node is determined by the fertility rate of that node.The color of the parent will be the average of color values of children nodes where the weights are the values associated with the children. In this case, the total female population of that node.


2004
WE
NA
AS
EE
Spain
Italy
United Kingdom
France
Germany
Netherlands
Portugal
Belgium
Sweden
United States
Kazakhstan
Uzbekistan
Turkey
Poland
Romania
Greece
Hungary
Czechia
Belarus
Austria
Bulgaria
Serbia
Azerbaijan
Total fertility rate
3
2.5
2
1.5
1
0.5


2010
WE
NA
AS
EE
Spain
Italy
United Kingdom
France
Germany
Netherlands
Belgium
Portugal
Sweden
United States
Uzbekistan
Kazakhstan
Turkey
Poland
Romania
Greece
Czechia
Hungary
Belarus
Austria
Bulgaria
Serbia
Total fertility rate
3
2.5
2
1.5
1
0.5



The structure of the graph is almost the same from 2004 to 2016. This means that the relative female population across regions has stayed constant across this time period. The fertility rate increased from 2004 to 2010 and then reduced in 2016 for Western Europe. This is indicated by the change in color. The below diagram is the tree map for the year 2010.



References:-

https://plotly.com/python/parallel-coordinates-plot/
https://plotly.com/python/splom/
https://plotly.com/python/sunburst-charts/
https://plotly.com/python/treemaps/
https://en.wikipedia.org/wiki/Western_Europe
https://en.wikipedia.org/wiki/Eastern_Europe