CD 732: Data Visualization
Report for Datathon-5

Swasti Shreya Mishra (IMT2017043)
International Institute of
Information Technology Bangalore
Email: SwastiShreya.Mishra@iiitb.org

1. Introduction

We are given a tabular dataset published by United Nations Economic Commission for Europe (UNECE) which has the *Country Overview* data. It consists data for 52 distinct countries from 2000 to 2016. It has 79 columns that ranges over different characteristics of measuring performance of the nations. I will be using a subset of this dataset to test some of my hypotheses and will make inferences from the same.

2. Methods

The data is mainly visualized using *plotly.express* and *pandas* library has been used for data manipulation. In most of the visualizations, data from the years 2015 and 2016 has not be included as it consisted of a lot of null values. For every other numeric data column, the null values have been replaced with the column mean.

2.1. Sunburst plots

Sunburst plots are used to visualize hierarchical data spanning outwards radially from root to leaves.

- The mean life expectancy of humans at birth for countries over the years 2000 to 2014 has been visualized. The mean life expectancy of humans is calculated by summing up (life expectancy at birth men * total population male) and (life expectancy at birth women * total population female) and then dividing it by the total population. This is shown in figure 1. The reason behind choosing a diverging colormap is that, we can see magnitude of the increase in life expectancy for all the countries over the years 2000-2014. Figures 10 and 11 represents the same information for Switzerland and United States respectively. The fans are arranged in the order of increasing values of mean life expectancy and this is nearly same as the order of the years from 2000-2014.
- The total population of Germany, France, United Kingdom and Italy over the years 2000 to 2014 has been visualized in the figure 2. The fans are arranged in the order of increasing values of population during a particular year. For France, United Kingdom and Italy there is steady rise in population with increase in years. Whereas the opposite is true for Germany. There has been a steady decline in the population with increase in years.
- The population in million per square kilometer is visualized is in figure 3. This plot shows us the population density of European nations.

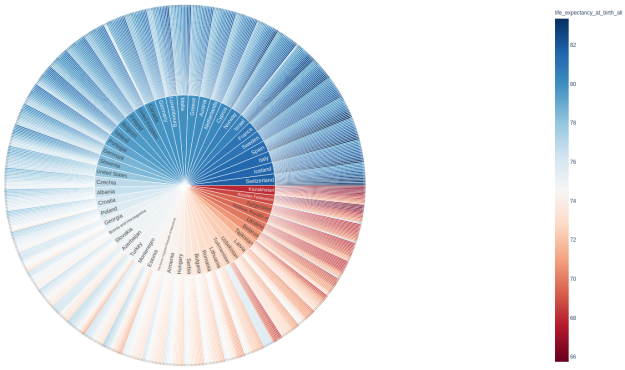

Figure 1. Sunburst plot showing mean life expectancy of countries from 2000-2014

2.2. Treemaps

A Treemap displays hierarchical data as a set of nested rectangles. Each group is represented by a rectangle, which area is proportional to its value.

The gender pay gap has been visualized using a treemap in figure 4. I think using a treemap to visualize gender pay gap is better as we can visualize which country has the largest gaps along with how the gap has changed over the years.

2.3. Parallel Coordinates Plots

In a parallel coordinates plot, each row of the *pandas DataFrame* is represented by a polyline mark which traverses a set of parallel axes, one for each of the dimensions.

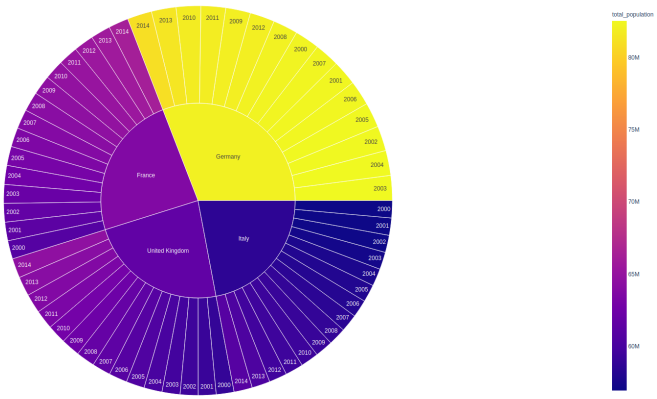

Figure 2. Sunburst plot showing total population of countries from 2000-2014

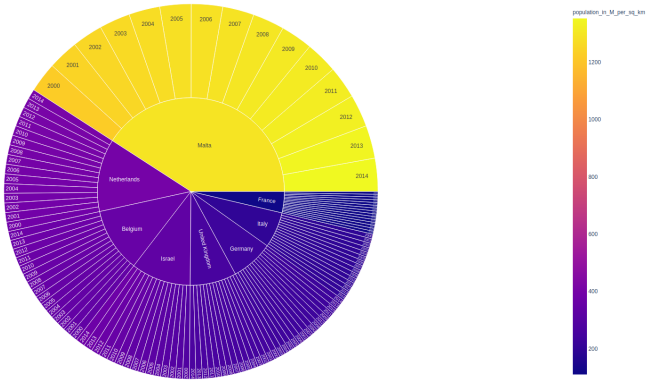

Figure 3. Sunburst plot showing total population in millions per square kilometer from 2000-2014

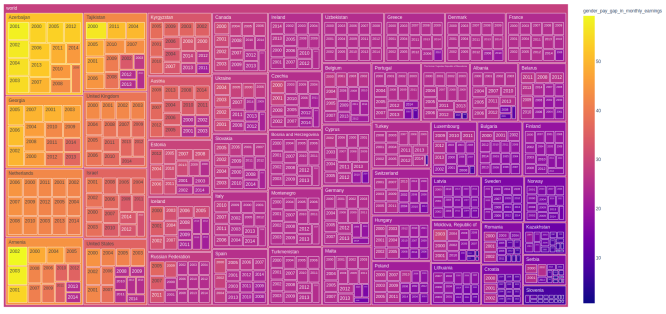

Figure 4. Treemap depicting gender pay gap in countries over the years 2000-2014

This can be used to get a high level idea of the correlation of multiple variables at the same time.

- Adolescent fertility rate is defined as the adolescent birth rate measures the annual number of births to women 15 to 19 years of age per 1,000 women in that age group. Figure 5 displays the correlation of adolescent fertility rate with the life expectancy of women after 65. The upper half of the parallel axes

Figure 5. Parallel Coordinates Plot for adolescent fertility and life expectancy of women after 65

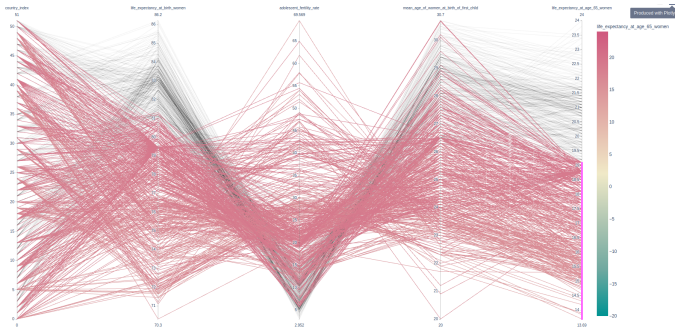

Figure 6. Parallel Coordinates Plot for adolescent fertility and life expectancy of women after 65

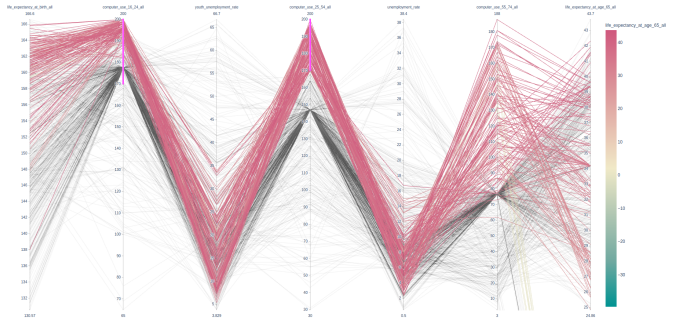

Figure 7. Parallel Coordinates Plot for computer usage and unemployment rate

of *life expectancy of women after 65* is selected to show that, as the adolescent fertility rate decreases, the life expectancy of women after 65 increases. The lower half is selected in figure 6 for comparison.

- The figure 7 depicts the correlation of computer usage and unemployment rate. A trend can be observed from the plot that, unemployment almost decreases with increase in computer usage.

2.4. Scatterplot Matrices

A scatterplot matrix is a matrix associated to n numerical arrays (data variables), $X_1, X_2, ..., X_n$, of the same length.

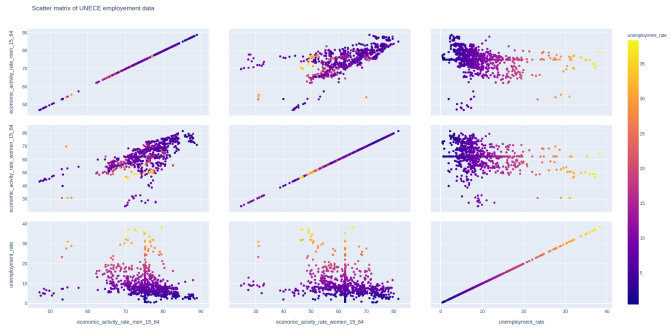

Figure 8. Scatterplot matrix for unemployment rate, economic activity rate for men and women of the age group 15-64

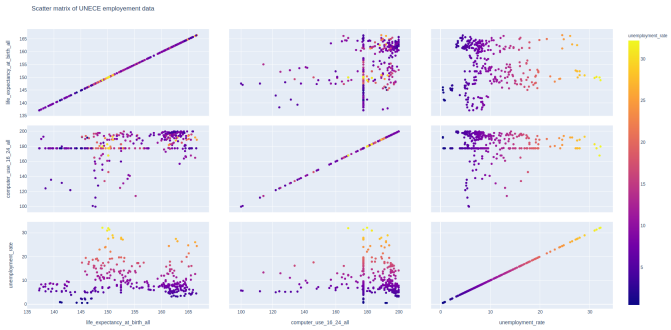

Figure 9. Scatterplot matrix for unemployment rate, computer usage of the age group 16-24 and life expectancy at birth

The cell (i, j) of such a matrix displays the scatter plot of the variable X_i versus X_j.

- The figure 8 shows the scatterplot matrix for unemployment rate, economic activity rate for men and women of the age group 15-64. Here it is evident that the economic activity rate for men goes up as it goes up for women and vice-versa.
- The figure 9 shows the scatterplot matrix for unemployment rate, computer usage of the age group 16-24 and life expectancy at birth. This was plotted to get more insight for the inferences made from plot 7.

3. Inferences

- From figure 1, we can infer that mean life expectancy at birth is very high for countries like Switzerland, France, Netherlands, Germany, Canada and United Kingdom. These also have world's best health case facilities and therefore, this trend is expected.
- Figures 10 and 11 represent that there has almost been a steady rise in the life expectancy at birth from the years 2000-2014. This is in general true for all the countries in the dataset and can be verified from the figure 1.

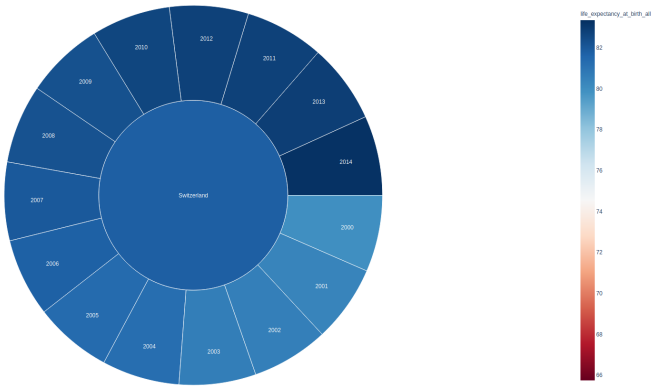

Figure 10. Sunburst plot showing mean life expectancy of Switzerland from 2000-2014

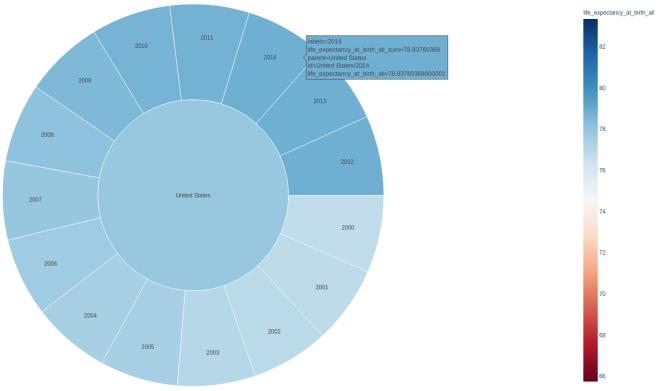

Figure 11. Sunburst plot showing mean life expectancy of United States from 2000-2014

- In the figure 1, from the divergence of the colors in the fans of the countries, we can infer the countries for which the net increase in mean life expectancy at birth from 2000-2014 has been the highest. It is evident from the figure that Turkmenistan and Kazakhstan have the highest increase with 11.7% and 8.3% respectively. Next are countries like Switzerland and France with 4.1% and 4.5% increase respectively. United States can be seen to have little change in the mean life expectancy at birth and therefore has a 2.7% increase.
- Figure 2 is consistent with the fact that there was a steady decline of the population of Germany over the years 2000-2014. This issue also has a positive correlation with total fertility rate of Germany and is depicted in figure 12
- Figure 3 depicts this fact that Malta has the highest population density among European countries. Along with that, Netherlands, Belgium, Israel and United Kingdom also have high population densities among the European nations.
- As shown in figure 4, the gender pay gap is very high

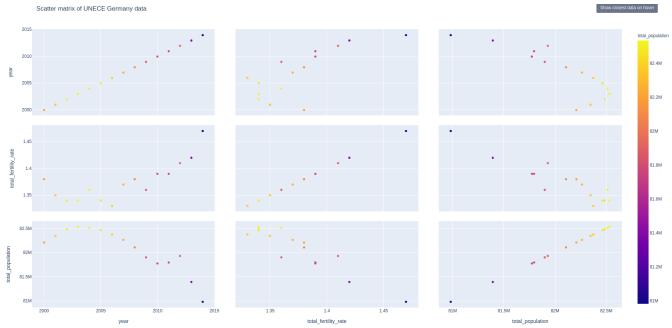

Figure 12. Scatterplot matrix for Germany data: Total fertility rate and Total population

for advanced countries like United Kingdom, United States and Canada. Even though United States has managed to reduce the gender pay gap over the years 2000-2014, United Kingdom has failed to achieve the same. The gender pay gap is also very high in Netherlands and from the plot we can infer that the gap almost remained the same throughout 2000-2014.

- The comparison of figure 5 and figure 6 shows that the increase in adolescent fertility negatively affects the life expectancy of women after 65.
- The figure 7 shows the relation between computer usage and unemployment rate. Even though decrease in unemployment rate with increase in computer usage isn't highly correlated, I feel that it is significant. As during the years 2000-2014, there was a bloom in technology and it can be assumed that with the advent of computers a lot of new opportunities were created as well as the literacy went up due to the easier access of information.
- The figure 8 shows that the economic activity rate for men goes up as it goes up for women and vice-versa. Almost a regression like model can be fit to find the value of one variable, given the other.

4. Conclusion

The possibilities of interpretations of the data we have visualized is very large. I have only been able to look at a few of them but they provide us with a lot of information. I approached this datathon by first making a few hypotheses and then testing them using the data and inferring results from the visualizations. It was a fun exercise as I got to verify my intuition using the data visualization tools.