Datathon - 5

[1]Arjun Verma : IMT2017008
CS732/DS732: Data Visualization
Course Instructors : Prof. Jaya Sreevalsan Nair
Technical Report - 5

Abstract. *This technical report contains an overview of the methodology involved in generating the visualizations of the provided dataset consisting of an overview of countries. The report further attempts to generate a good narrative from the visualizations generated.*

1. Introduction

The provided dataset contains information about different statistical measures that can be used to track the development status of the countries for which these variables have been provided. These include a number of variables such as GDP per capita, life expectancy of men and women at various ages, economic activity, unemployment rates, etc. *In this assignment, we try to build a detailed case study of the economic development of the given countries.*

2. Tools and Methodology

2.1. Tools

The exhaustive set of libraries used for generating the final inference visualizations involve :

- os
- numpy
- pandas
- plotly
- seaborn
- Gephi

The usage of the aforementioned tools would be elaborated upon in the methodology section. Note, only those involved in the final generation of visualizations have been mentioned. The ones that were experimented with but discarded have been omitted.

2.2. Methodology

In this section, we provide the distinct characteristics involved in each technique and the reasonings behind them. However, before we dive into the individual techniques, let's have a clarity on the basic workflow of the implementation.

2.2.1. General Methodology

For getting to the final combined inference, the dataset has to go through three stages represented by three files of code to be run to get to the final output. The workflow proceeds in this order :

- First, the dataset is collected and passed along to the three files.
- Secondly, the datset is then pre-processed to capture only relevant information for the purposes of respective visualizations.
- Lastly, combined inferences are made from the outputs of each file.

A basic framework is represented as :

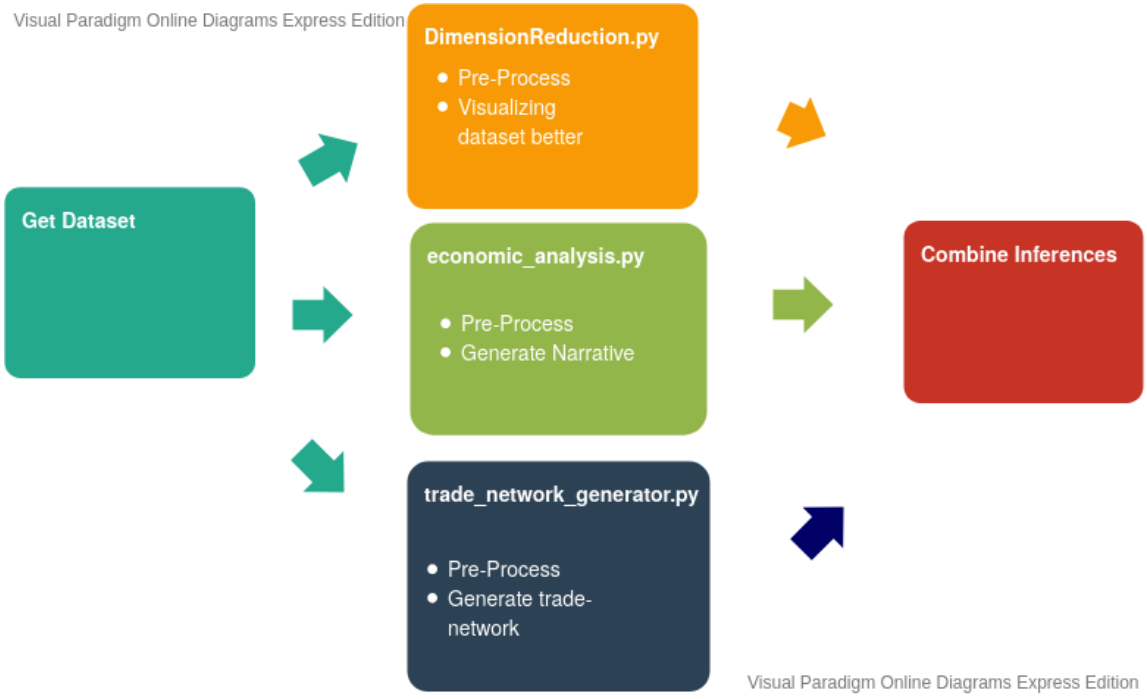

Figure 1. Framework

A general point to note with respect to the dataset is that the year 2016, has absolutely no values for the essential features across all countries and hence has been specifically truncated from the dataset. The latest trends have been analyzed upto 2015.

2.2.2. Methodology for Dimension Reduction

Distinct techniques (characteristics) for the dimension reduction :

- Computed principal components for the matrix.
- Projected the matrix onto components and reordered the matrix based on these projections.
- This was just an initial visualization technique to better grasp the dataset. It wasn't much helpful in making any inferences.

2.2.3. Methodology for Economic Analysis

Distinct techniques (characteristics) for these visualizations :

- Visualization Variety : Treemap, Sunburst, Scatterplot, Parallel Coordinates Plot, Correlation Heatmap, Bubbleplot, Lineplots, Barplots,Networks.
- Color Map for Nodes and Edges in Networks: Gradients of singular colors
- Edge Weighting Technique : Normalized Averages

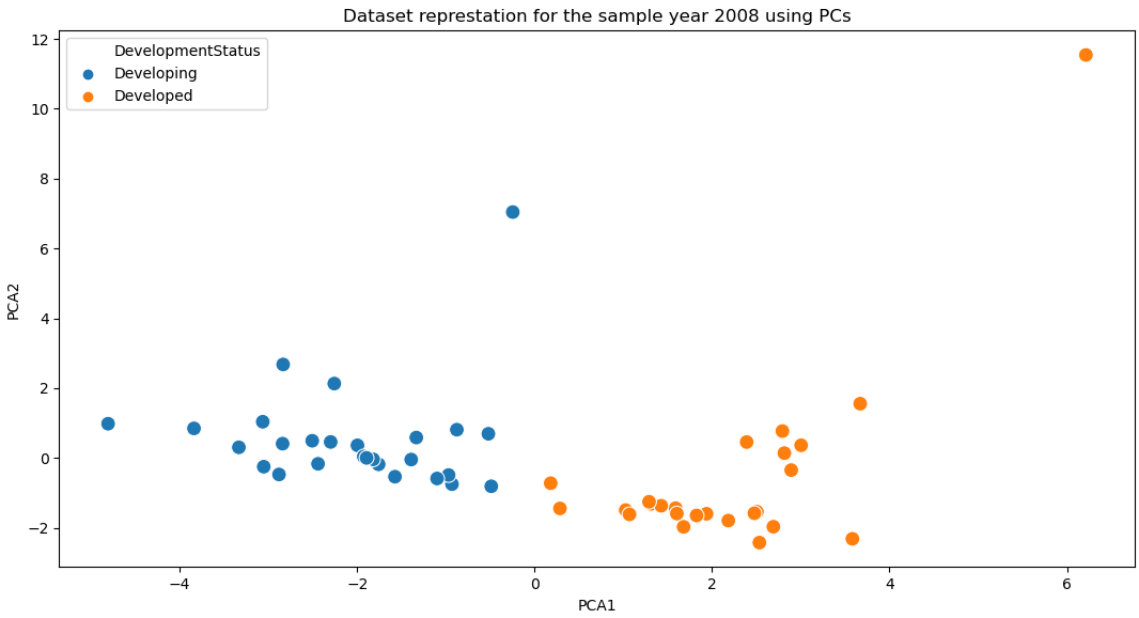

Figure 2. Sample Dataset visualization for the year 2008

3. Inferences

Since there are a lot of visualizations going around, it is better to have a look at the inferences from particular visualizations along with how they are being created (wherever necessary) in this section together.

1. ***Treemap Visualization for an over the top comparison across years :*** One of the primary factors on which the economies of various countries are compared is the GDP per capita. We thus begin our process of inferencing by visualizing the GDP per capitas (PPP based) of the varying countries over all year intervals using the treemap given below. From the map, we can make the following inferences :

    1. **Luxembourg** has held the position of having the **largest GDP** for a long period of time. **Norway** has always been a **second** from the list of majorly European countries provided and was only **overtaken** by **Ireland** in the year **2015**. This overtake would be highlighted again later, when we move on to the trade network. On the other end of the spectrum **Tajikstan** has always occupied the **last** position in terms of GDP per capita.

    2. From the color gradient across years, we can see that the **GDP per capita has been more or less increasing** across all the countries with the passing years with the countries being in a much better position than they were in at the beginning of 2000.

    3. Another thing to notice is the changing development status of some of the countries such as Cyprus which went from being a developing country in 2000 to a developed country by 2004. This was an added feature to the dataset which will again be touched upon in scatterplot visualizations.

Figure 3. Treemap Visualization of the GDP's of countries across years

2. ***Sunburst Visualization for choosing an optimal year of study :*** Sunburst visualizations for the purposes of the study chosen, give somewhat redundant inferences as those from treemap and thus the way I have utilized sunburst is to choose an optimal year of study when looking at a specific year study. We wish to study the countries at their latest stages and hence a subdataset was chosen which comprises of values for all countries for the last five years (**2015-2011**).The inferences made are as follows :

1. The factor that was being considered to choose an optimal year of study was to choose a year when a majority of the countries have fared their best until that period. With the exception of Norway, all countries have performed their best upto **2015** and hence this year was chosen. Also, it is a particularly interesting year to study so as to try to infer why did Norway see a sudden drop in it's GDP.

2. Another important thing to notice in contrast to **Norway** which saw a **significant drop** in it's GDP, **Ireland** saw a **significant increase** and we will try to address this as well in the trade network.

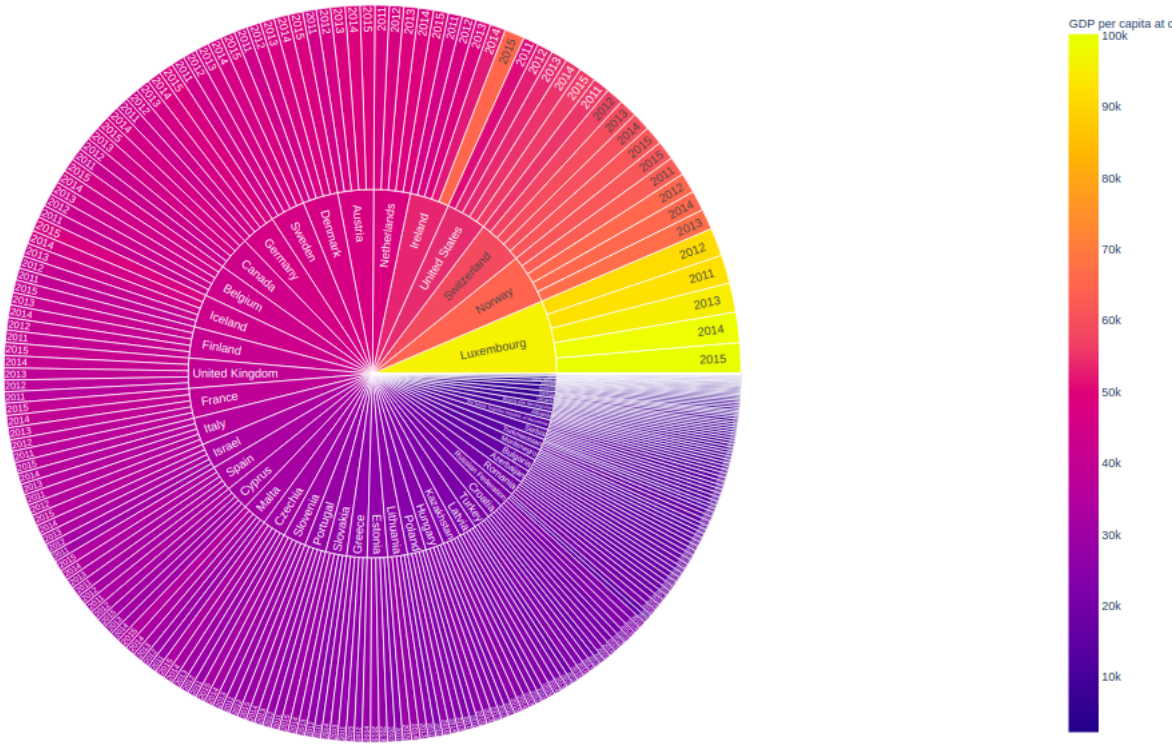

Figure 4. Sunburst visualizing a country wise analysis of GDP for the five latest years

3. ***Correlation Heatmap :*** Let us now start with our study of GDP per capita. We begin this we start with a **Correlation Heatmap** to get significant features in correlation with GDP. This will help us generate our **cause and effect** inferences in the latter sections. The sole information captures from this was :

1. GDP per capita, as expected, is in direct correlation with a number of columns such as Unemployment rate and Life Expectancies. We thus try to limit our study to one such category of columns (Life expectancies) as this is not our focus for datathon. Similar trends can be computed for remaining features. All in all, relations between 19 columns were visualized for this case study.

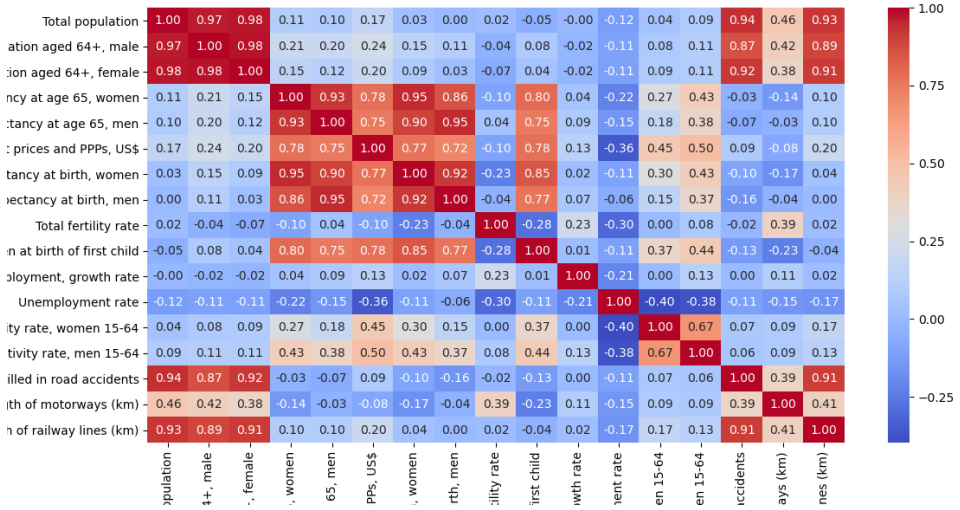

Figure 5. Correlation heatmap for nineteen features

4. ***Trend Analysis (All countries, fixed year = 2015) :*** In the next couple of points we will try to compare a few of the nineteen features' trends along with the GDP per capita. These have been included to make the study more complete and focus on establishing a **cause and effect** relation. We have divided the features across two cases, **fixed year study for across-countries analysis** and **fixed country study for in-country analysis**. All values have been normalized for these two case studies. Some of the inferences from the **fixed year** case are as follows :

1. Countries which have higher economic activity of men between ages 15-64, also tend to show a higher GDP per capita.

2. Countries which show a greater GDP per capita also tend to show a greater life expectancy at birth for men.

3. Countries which show a greater GDP per capita also tend to show a lower unemployment rate when compared to their counterparts.

4. The cause and effect relation that we hypothesize from the following visualizations is this : **Due to an increase in economic activity in a country, the GDP per capita of the country boosts, which also leads to better development and in turn increases life expectancy at birth.** This hypothesis will again be tested for the in-country analysis along with a few new features.

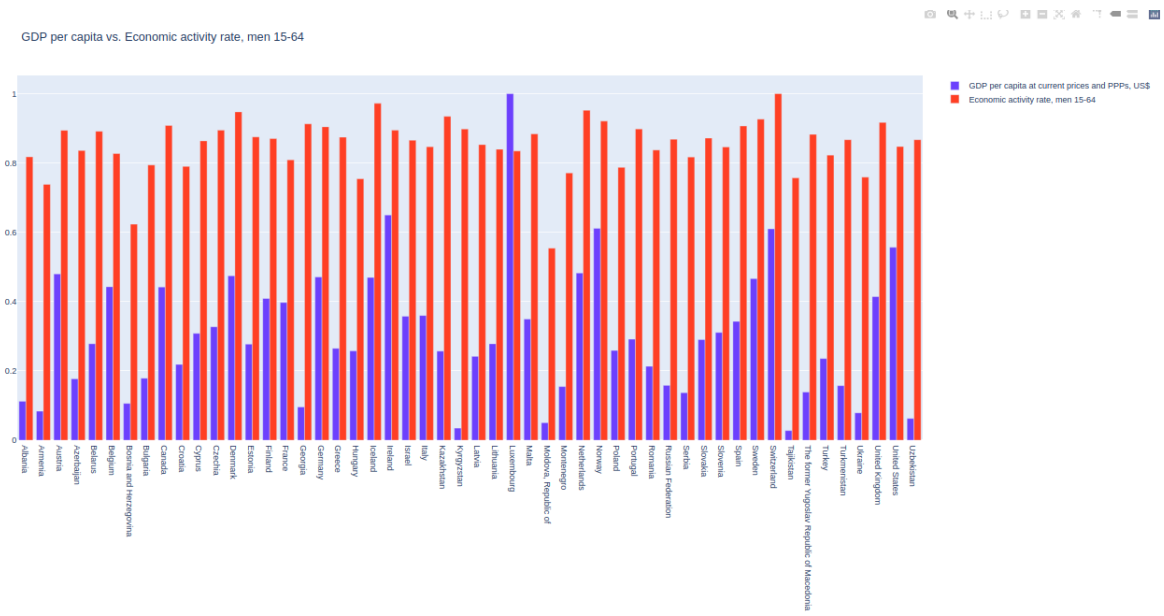

Figure 6. GDP per capita of countries vs. Economic activity of men in ages 15-64

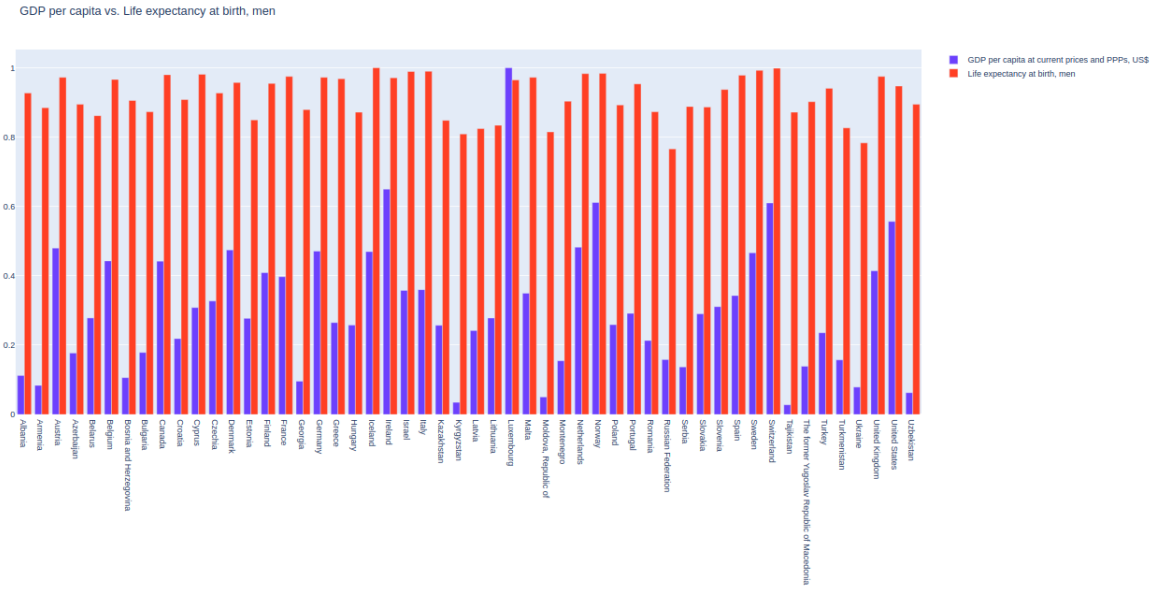

Figure 7. GDP per capita of countries vs. Life expectancy of men at birth

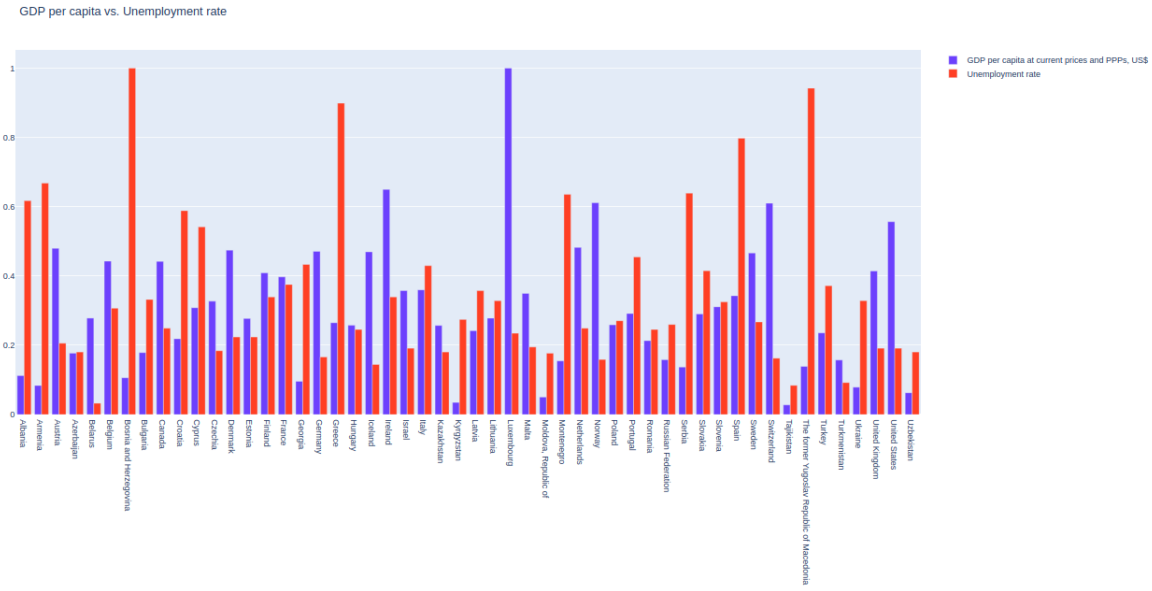

Figure 8. GDP per capita of countries vs. Unemployment rates

5. ***Trend Analysis (All years, fixed country = Norway) :*** Continuing with the same line of thought as the previous point, now let's look at some of the inferences from the **fixed country** case :

1. With an increase in economic activity of women between ages 15-64, the GDP per capita of the country increased.

2. With a greater GDP per capita there is also a greater life expectancy for women at birth as well as at age 65 and above.

3. With a greater GDP per capita there is lower unemployment rate over a period of time. As and when, there was a drop in Norway's GDP post 2014, there was also an increase in the unemployemnt rate.

4. With an increase in GDP, there is lesser number of deaths in accidents pointing to **better infrastructure**. Better has been highlighted as it is very evident that the quality of infrastructure is increasing as even with more kilometres of roadways being constructed, there are lesser number of deaths in accidents.

5. The cause and effect relation that we hypothesize from the following visualizations is this : **Due to an increase in economic activity in a country, the GDP per capita of the country boosts, which also leads to better development and in turn increases life expectancy at birth.** This hypothesis will again be tested for the in-country analysis along with a few new features.

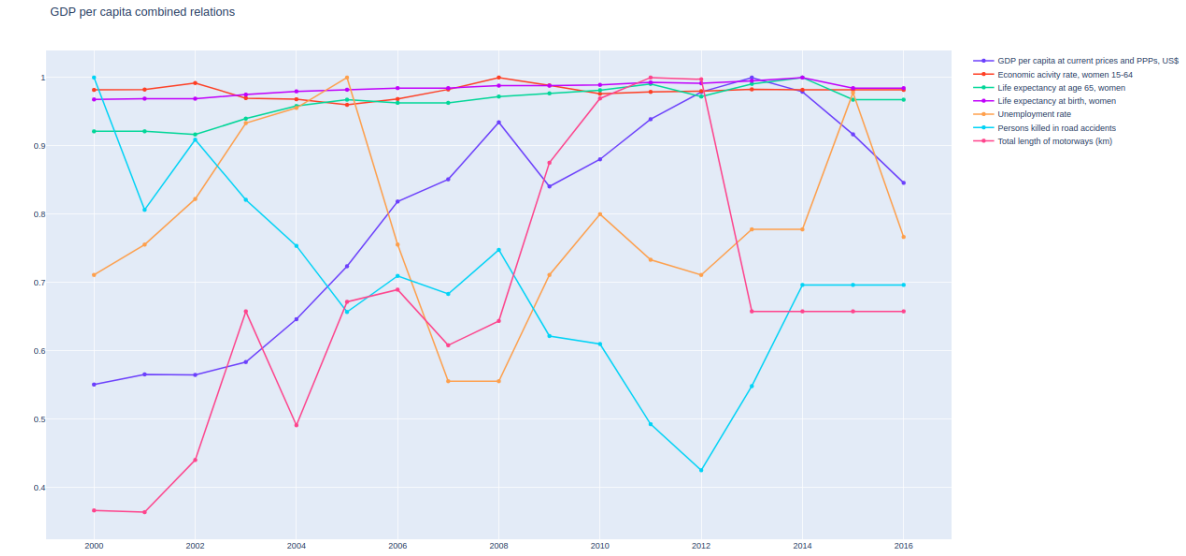

Figure 9. Combined Correlations with GDP for Norway across all years.

6. ***Scatter Matrix :*** The above comparisons of trends can also be found by plotting the scatter matrix. For this part of the visualization, I decided to create something interesting. Based on studies on GDP per capita, I have created a new feature of a country either being **developed** or **developing** and now, these features are compared across the two statuses. One can find many interesting relations apart from the one's mentioned in the previous points. For example, **if one were to compare unemployment rates in developed and developing countries, one can easily see that there is a larger rate of unemployment for a lower GDP per capita for developing countries than it is for developed countries.**

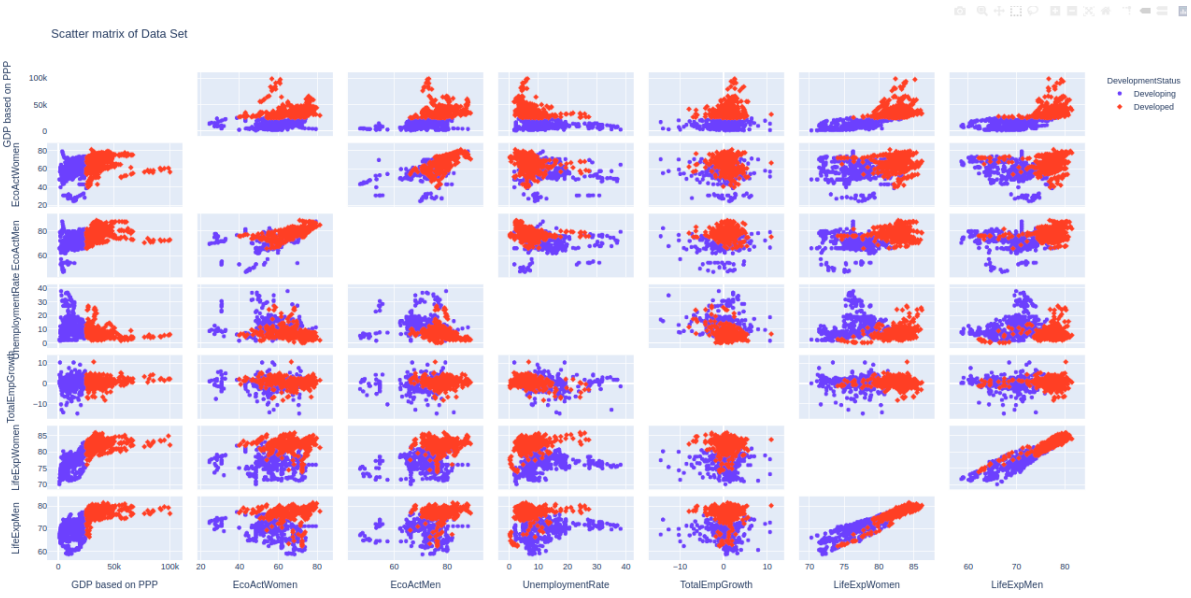

Figure 10. Scatterplot matrix for seven selected features

7. ***Parallel Coordinates Plot :*** The same correlations between features can also be visualized very easily using the parallel coordinates plot. The reason for including, both, multivariate analysis and pairwise analysis in the earlier section, was exactly to demonstrate the ability of the ease with which these plots show the same relations in a single figure. A very unusual inference from this plot is that **when countries are gradually developing towards a developed state, they follow the same trend between GDP per capita and economic activity for women and show exactly what one would expect. With an increase in economic activity of women, there is an increase in GDP per capita for the country. However, when we compare them to states who have long past reached the development status and have a significantly higher GDP per capita as compared to others (shown in yellow), one sees an unusual trend of higher GDP per capitas having lower ecomic activity in women aged 15-64.**

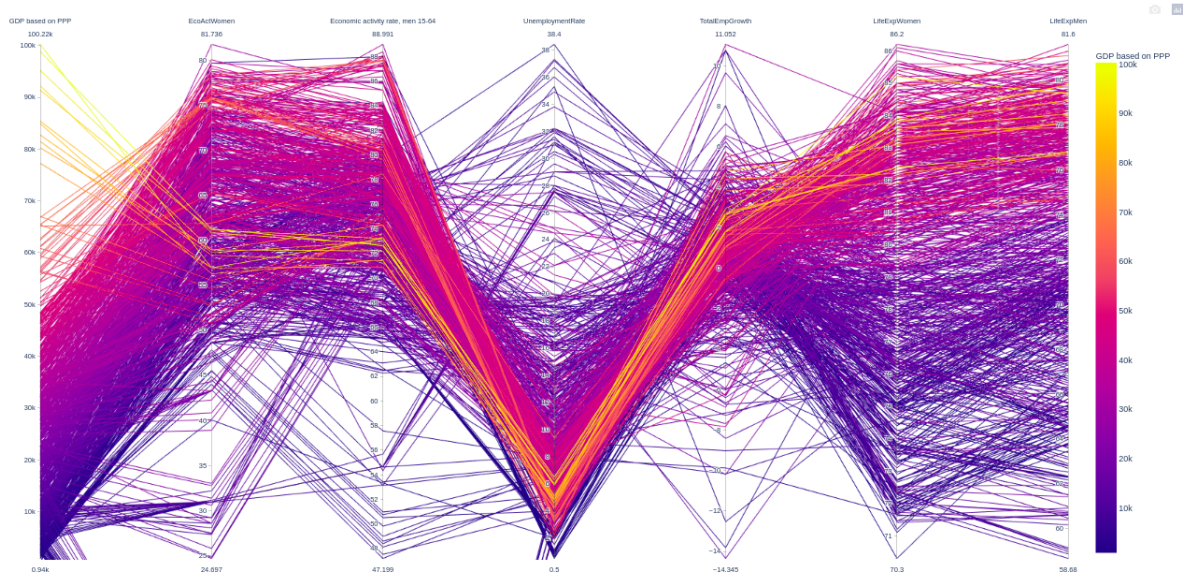

Figure 11. Parallel Coordinates Plot for seven selected features

8. ***Network Analysis : Trading amongst the European Union*** In a bid to understand the sudden increase in the GDP per capita for Ireland, I tried to create a network of trade relations between the countries. For the purposes of this analysis, I am limiting myself only to the European Union. A network assuming trading relations among all countries in the dataset was also created and the network has been included for reference, however it is not the highlight of our focus. The network has been created as follows : **All countries among the European Union (EU) have links between them. For the purposes of my study, I decided to check the dependency of the top three countries with the highest GDP's in the EU with the bottom ten in the EU. The node weights are thus the GDP per capitas and are colored with a single color of varying intensities. The edge weights are essentially the average of the external balance of the two connecting nodes (External Balance = Export - Import) as this captures the importance of trades between two connecting nodes. These graphs were plotted for four years : 2012, 2013, 2014 and 2015.** The inferences gathered from the networks are :

1. The top three GDP's among the EU have been **Luxembourg, Ireland and Netherlands**. These three also have the largest trade relations and control a good amount of the export market. Based on the analysis, the greater these countries export, the greater is their increase in GDP. In contrast, the more the countries import from these, the lower their GDP.

2. Now, we can clearly see why Ireland suddenly had a boost in it's GDP. We can see a significant increase in the weights of it's links in the year 2015, giving us a hint that their trade relations helped them significantly increase their GDP.

3. Another trend that is noted is, the stronger the links with these three countries, the lower the GDP of the countries, indicating of them only importing goods to sustain. Whenever, there has been an absence of trade with one or more with these countries, the bottom countries have shown a boost in GDP, indicating them exporting more for that particular year and less reliance on the imports from these three countries.

Figure 12. European Union Trade Market 2012

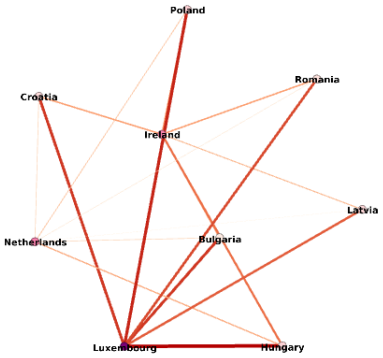

Figure 13. European Union Trade Market 2013

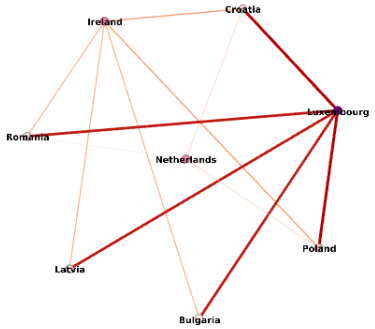

Figure 14. European Union Trade Market 2014

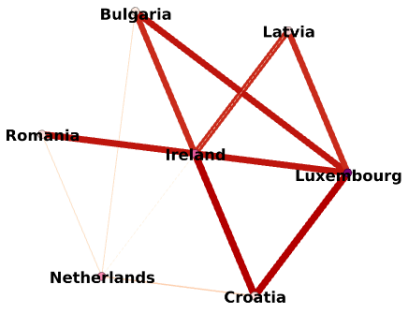

Figure 15. European Union Trade Market 2015

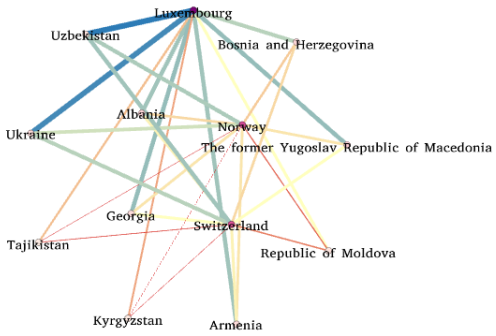

Figure 16. Hypothetical Trade Market assuming trades between all countries. This network compares the top three and the bottom ten in the entire dataset and uses a diverging colormap because of stark contrast.

9. ***Bubble Plots : An animated visualization representing Life Expectancy vs GDP per capita***
Lastly but not least, we have bubble plots depicting the entire growth rate of countries across different years using bubble plots. The increasing GDP of the countries across years can be very sweetly seen in the visualizations. The bubble size is in proportion to the total population of the country so as to showcase, the number of people who are being effected by increase or decrease in a country's GDP per capita.

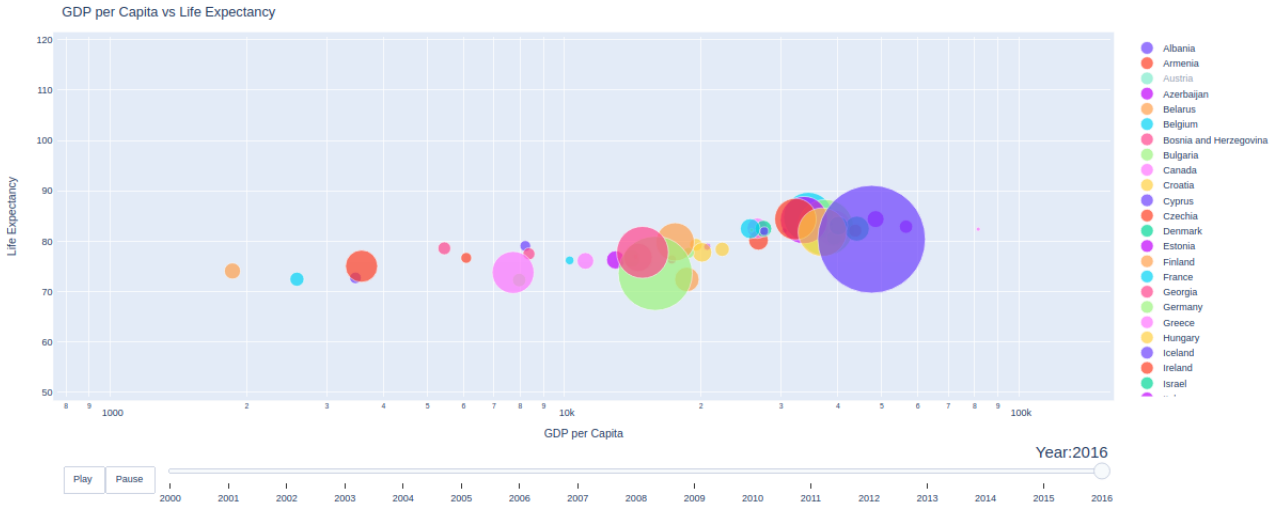

Figure 17. Bubbleplot showing the entire animated visualization along with an interface to stop at a particular year.

4. Conclusion

In conclusion, it is worth to mention that the report has successfully elaborated on all the deliverables. The tools and methodology section entails the questions :

1. Which parts of the dataset were you able to use, and how have you been able to use?

2. Which visualizations did you choose, why, what technologies (Python libraries, others) did you use for the visualizations ?

The indicative tasks of explaining inferences from the multivariate data visualizations and from remodeling the data using the hierarchical relationship has also been done. In addition to this, an entire narrative around economic analysis of countries has also been built.

5. References

1. Understanding Developing and Developed Countries.
2. Understanding effects of Exchange Rates.
3. Norway's GDP Statistics.
4. European Union Member States.