CS/DS 732 Data Visualization - 2019-20

Assignment 4 - Europe Dataset

Aditya Vikram Srivastava

IMT2016004

December 21, 2019

Introduction

This report is in accordance to the course guidelines and deals with complete information on the work done with respect to Assignment 4. The task was to visualise and analyse Europe dataset obtained from Kaggle. The data set contained data for 32 European countries with values on 43 indicators. Some of the indicators were as follows - GDP, Population, Crime rate, Health, Political trust etc.
I have used Python as the preferred language for doing the assignment. The program requires other packages like Numpy, Pandas, Scipy and plotly.

Structure

The folder IMT2016004_A4.tar.gz contains one subfolder named images, two python files(*new.py* and *Plots.py*), original europe dataset folder and combined csv.

images

This subfolder contains all the images and plots generated using Plots python module.

How to run the program?

To run the modules, you need to have the following python packages installed:

- Numpy
- Pandas
- Seaborn
- Plotly
- Bubbly

The following command should be executed to run the modules:

```
$ python3 <module_name>.py
```

Analysis

The data provided can be divided into tracks of analysis based on the attributes present in the dataset. The two tracks are - General Demographics and Individual centeric information.

General Demographics	Individual centeric
Median Income	Job satisfaction
Population	Life satisfaction
GDP	Leisure satisfaction
Budget	Climate
Crime	Young population
Unemployment	Health
Political, Legal trust	Savings

General Demographics track will help analysts in government authorities, non profit organisations, international authorities (UN, World Bank etc) to identify, analyse and plan for their various schemes. The Individual centric helps regular humans who would want to migrate to Europe for work or some other reason. This is why we have divided the analysis task into the two tracks.

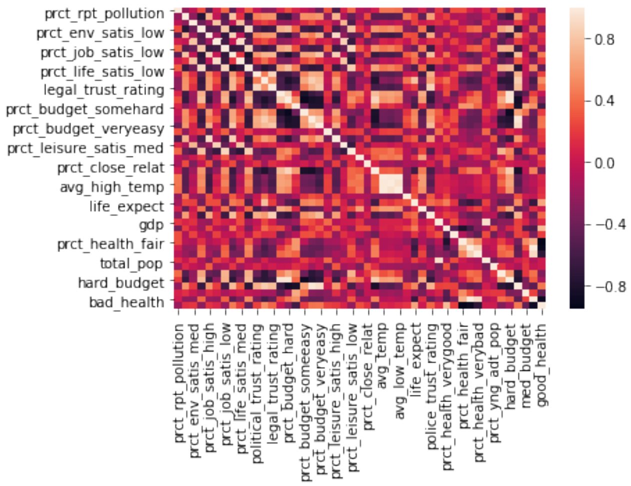

Figure 1: Correlation matrix

Problems faced

The data was provided in multiple individual CSVs, which had to be merged and cleaned for certain values. There are a total of 51 independent countries in europe but the dataset contained only 32. Baltic countries (Previouly Yugoslavia) were missing along with certain other small island countries. The reason for this can range from political unrest to unfeasability of data collection. All the data values were normalised to the range 0-100 for ease of data analysis.

Tasks

I have generated univariate, bivariate and multivariate plots to understand and analyse the data. Along with this geospatial visualisations on the map of Europe to identify region wise patterns have also been done.
Multivariate analysis has been done on attributes which had max correlation between them. This was identified using the correlation matrix plot(figure 1). The tasks were divided as per the two tracks described above

Task 1

The attributes used to analyse are(figure 2 - 21):

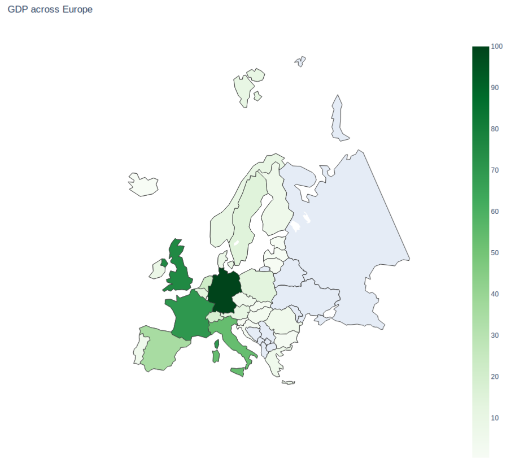

Figure 2: GDP map of Europe

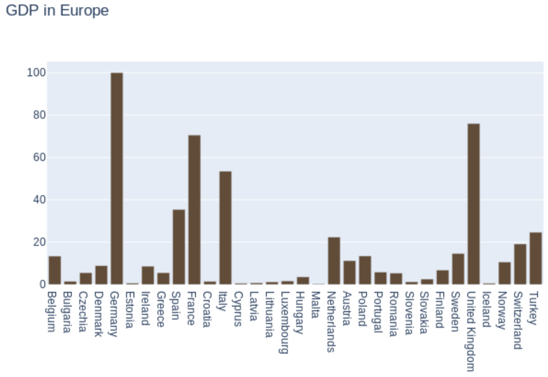

Figure 3: Univariate plot for GDP across Europe

1. GDP
2. Job satisfaction
3. Population
4. Median Income
5. Crime
6. Hours worked
7. Political trust
8. Police trust
9. Legal trust
10. Health
11. Unemployment
12. Budget
13. Pollution

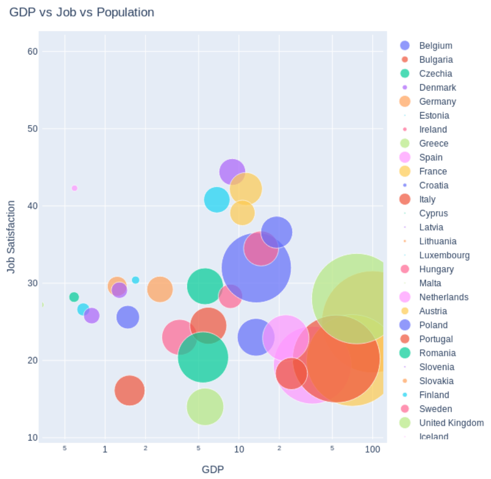

Figure 4: GDP vs Job Satisfaction vs Population multivariate bubble plot

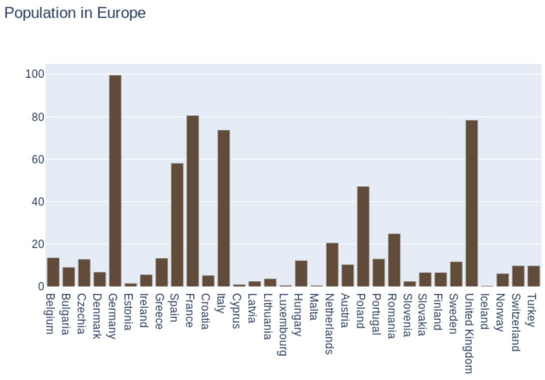

Figure 5: Univariate plot for Population across Europe

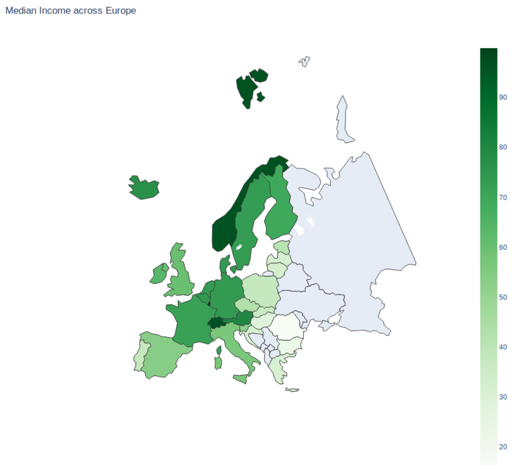

Figure 6: Median Income map of Europe

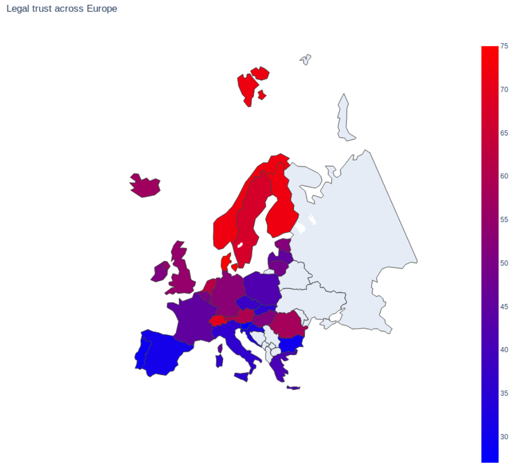

Figure 7: Legal map of Europe

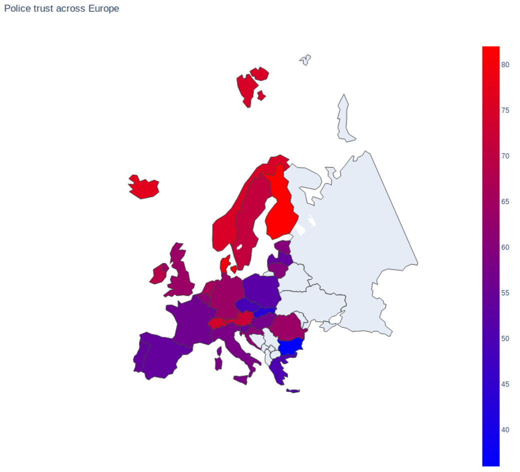

Figure 8: Police trust map of Europe

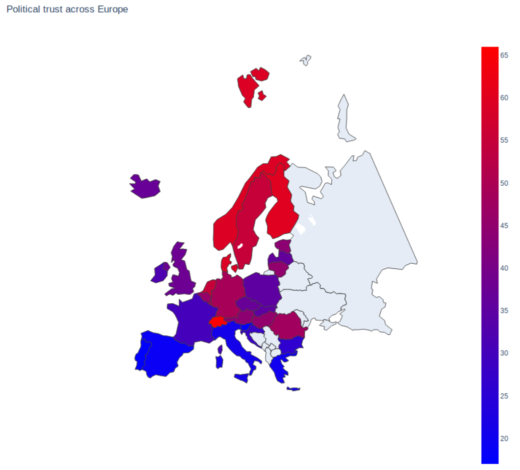

Figure 9: Political trust map of Europe

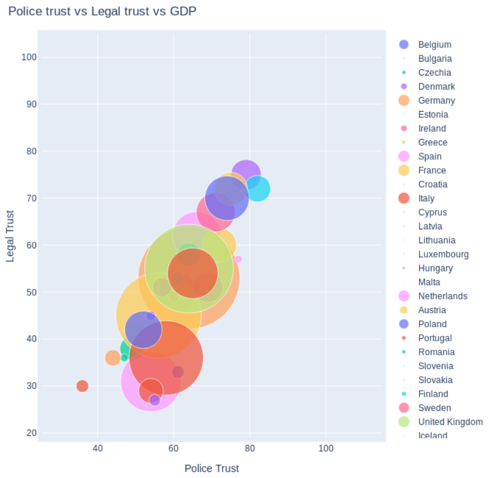

Figure 10: GDP vs Police trust vs Legal trust multivariate bubble plot

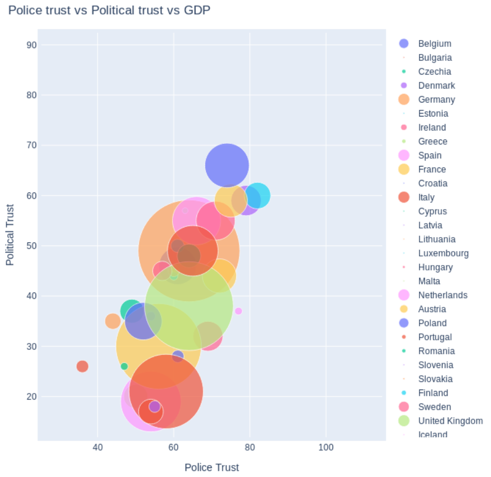

Figure 11: GDP vs Political trust vs Police trust multivariate bubble plot

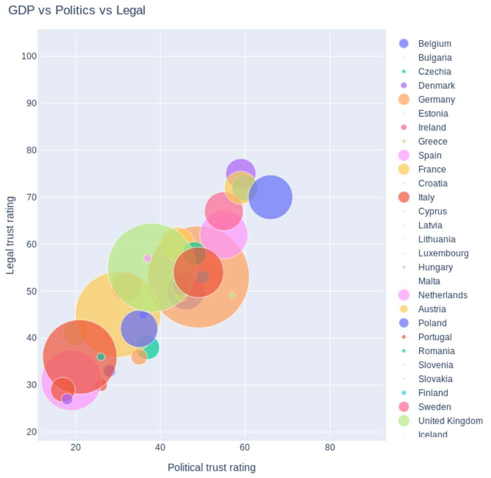

Figure 12: GDP vs Legal trust vs Political trust multivariate bubble plot

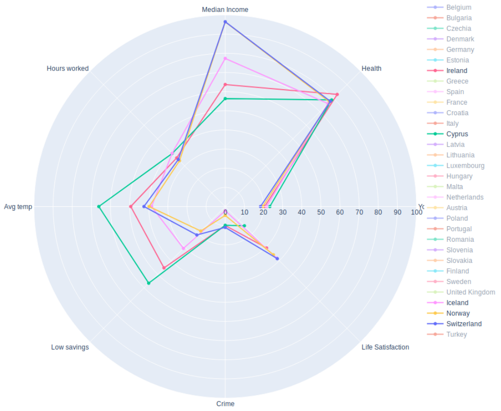

Figure 13: Radar plot for top health countries(Ireland, Cyprus, Iceland, Norway, Switzerland)

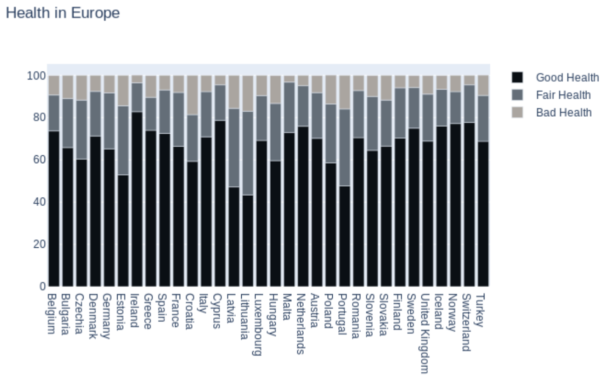

Figure 14: Health level bar across Europe

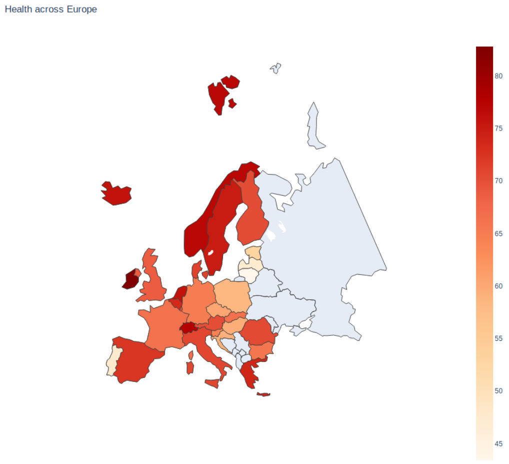

Figure 15: Health map of Europe

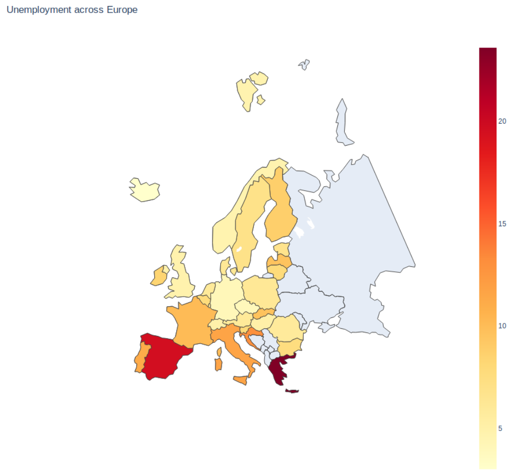

Figure 16: Unemployment map of Europe

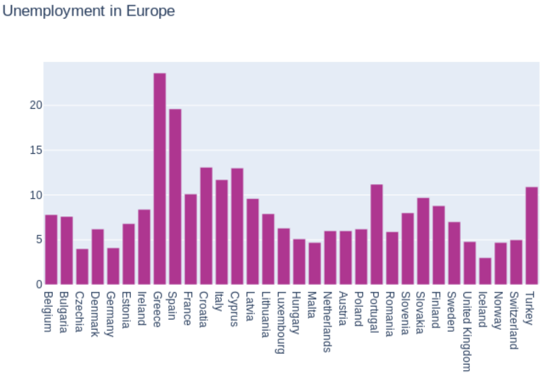

Figure 17: Unemployment bar across Europe

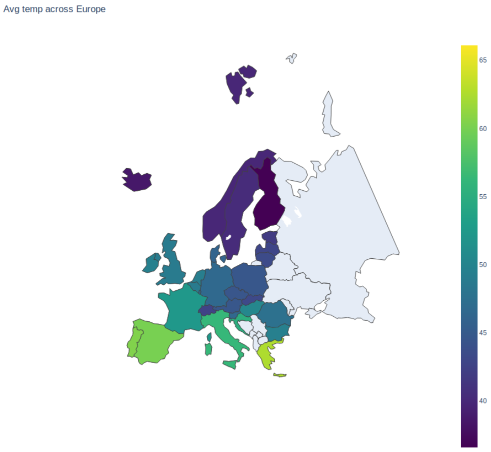

Figure 18: Temperature map of Europe

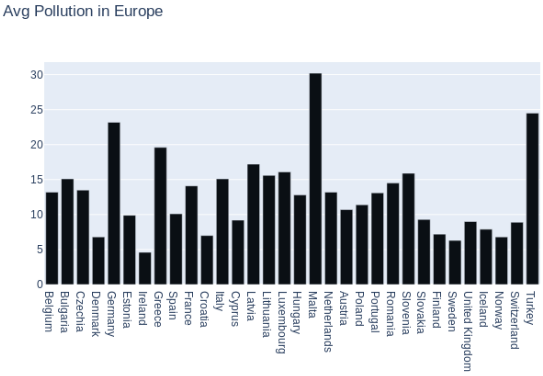

Figure 19: Pollution bar across Europe

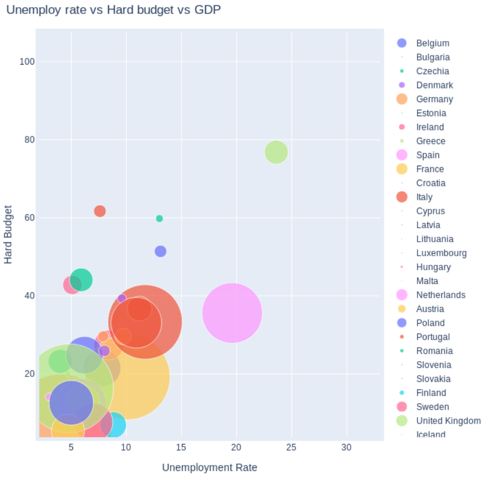

Figure 20: GDP vs Unemployment rate vs Hard budget multivariate bubble plot

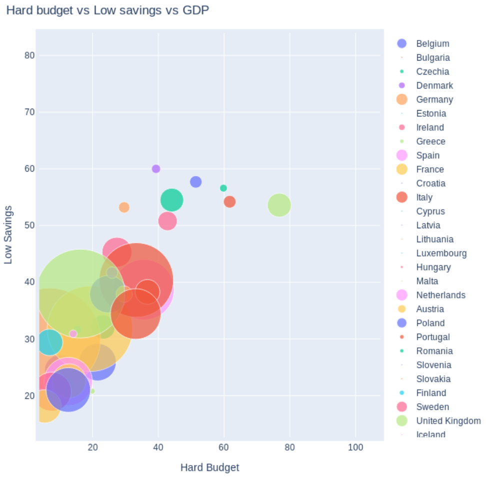

Figure 21: GDP vs Hard budget vs Low Savings multivariate bubble plot

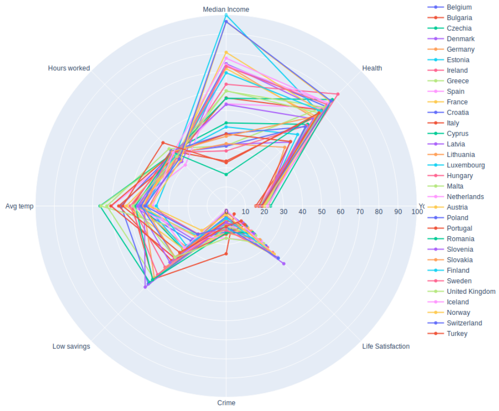

Figure 22: Radar plot of all given countries across Europe

Task 2

The attributes used to analyse Task 2 are(figure 22 - 36):

1. Life satisfaction
2. Environment satisfaction
3. Job satisfaction
4. Leisure satisfaction
5. Working hours per week
6. Low savings
7. Crime
8. Life Expectancy
9. Avg temperature
10. Health
11. Young population

Observations

Based on the analysis done on the data I found that Norway, Switzerland and Sweden were the best countries to migrate to for work and leading a healthier life. Similar observations could be determined from the plots under General Demographics task, which would be based on the analyst. For example, higher GDP is found in Germany, England, France while higher income is in the nordic countries of Norway, Sweden and Denmark.

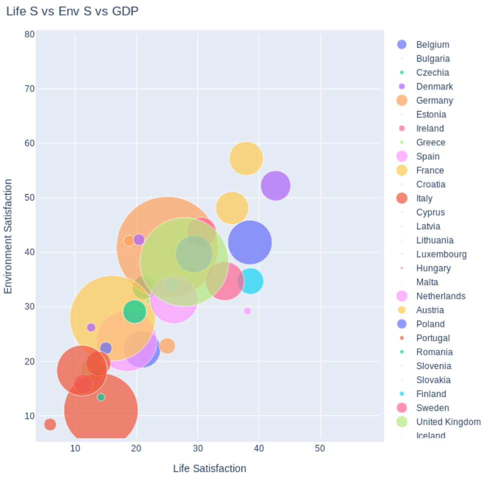

Figure 23: GDP vs Environment satisfaction vs Life satisfaction multivariate bubble plot

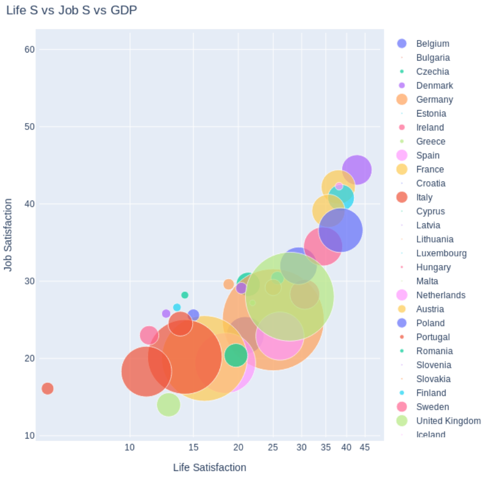

Figure 24: GDP vs Life satisfaction vs Job satisfaction multivariate bubble plot

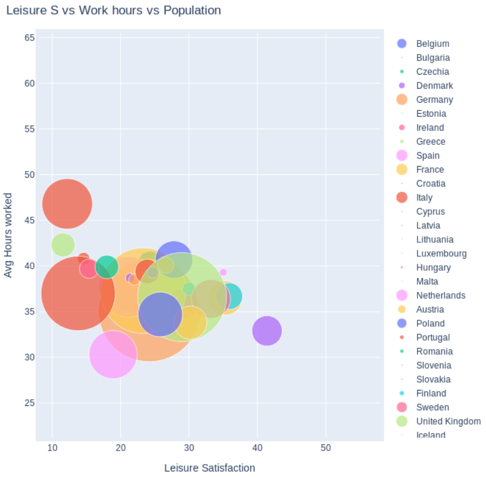

Figure 25: Population vs Work hours vs Leisure satisfaction multivariate bubble plot

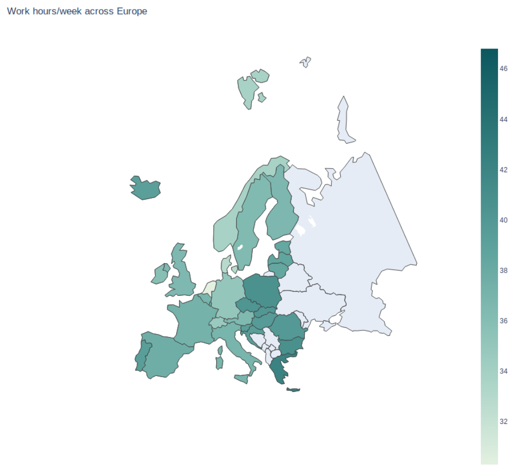

Figure 26: Work hours per week map of Europe

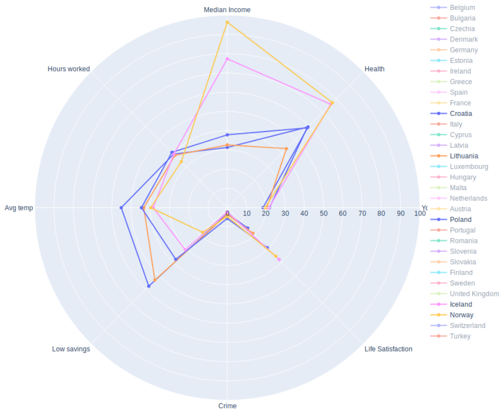

Figure 29: Radar plot for least reported crime countries(Croatia, Lithuania, Poland Iceland, Norway)

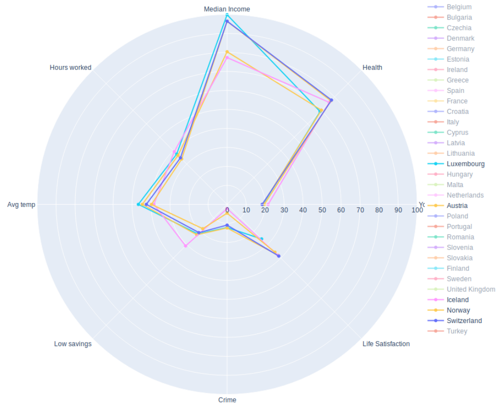

Figure 27: Radar plot for top median income countries(Luxembourg, Austria, Iceland, Switzerland, Norway)

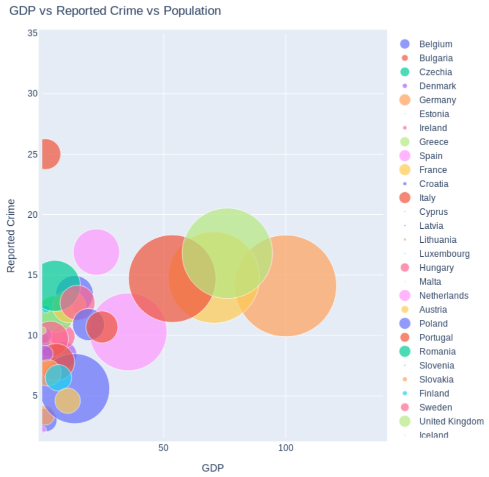

Figure 30: GDP vs Reported Crime vs Population multivariate bubble plot

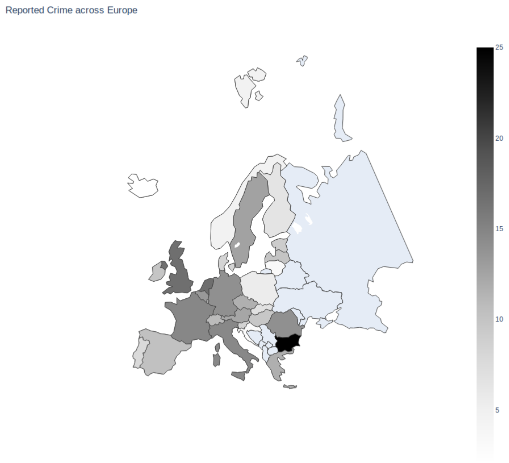

Figure 28: Reported Crime map of Europe

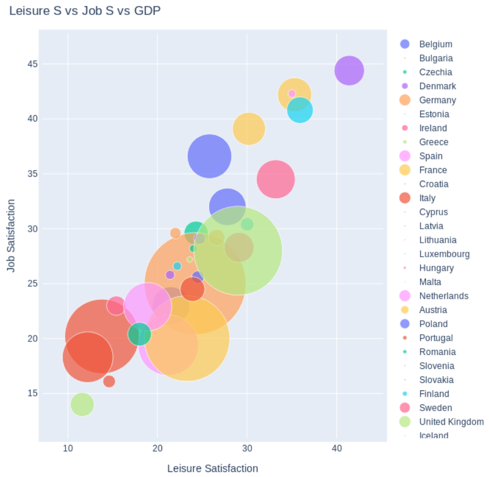

Figure 31: GDP vs Leisure satisfaction vs Job satisfaction multivariate bubble plot

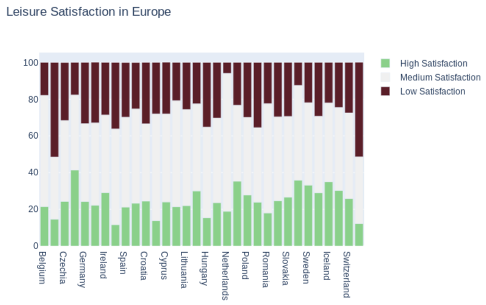

Figure 32: Leisure satisfaction levels bar across Europe

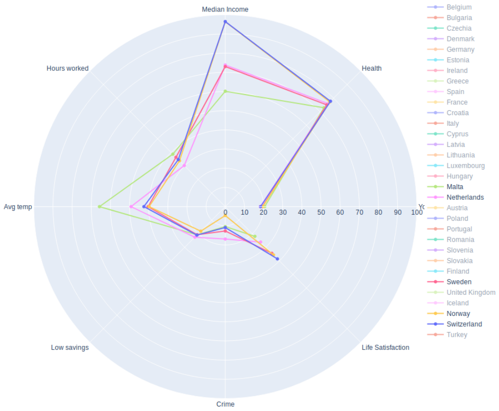

Figure 33: Radar plot of highest savings countries(Malta, Netherlands, Sweden, Norway, Switzerland)

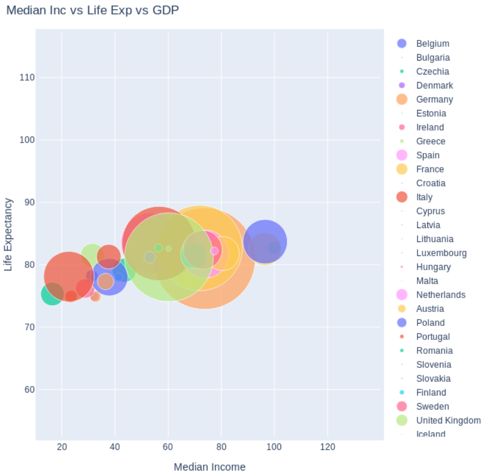

Figure 34: GDP vs Median Income vs Life expectancy multivariate bubble plot

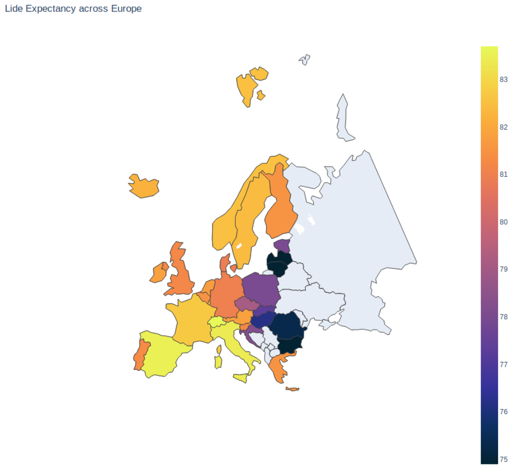

Figure 35: Median Income map of Europe

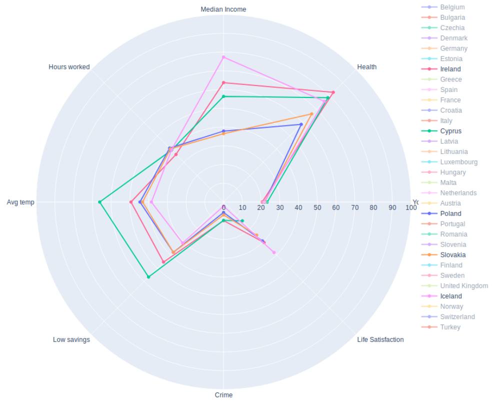

Figure 36: Radar plot for countries with highest young population (Ireland, Iceland, Cyprus, Poland, Slovenia)

References

[1] Numpy documentation

[2] Pandas documentation

[3] Matplotlib documentatuon

[4] Class slides