J

Team WFT

Final presentation

CS504 Team Project

Spring 2022

GEORGE MASON UNIVERSITY

J

Team Formation

Team Members

Jacob Baisden
SCRUM Master

Mitch Breeden
Product Owner

Delena Bell
Developer

Sudha Jain
Developer

J

Project Schedule

- The goals for this project were accomplished in 4 sprints, using YouTrack to monitor progress
- Sprint 1 (3/28 - 4/3): Problem definition
  - Define the potential problem, identify potential data sources and analytics
- Sprint 2 (4/4 - 4/10): Data Sets
  - Finalization and initial processing
- Sprint 3 (4/11 - 4/24): Algorithms & Analytics
  - Definition and coding of algorithms
- Sprint 4 (4/25 - 5/8): Visualization
  - Define and implement visualization concepts

M

Problem Definition

Problem -

We are analyzing World Happiness data collected from the Gallu World Survey from 2015-2021. Our goal is to use this data to determine the factors that affect happiness levels in each country. We are also interested in how the COVID-19 health crisis affected happiness across the world and if any regions were disproportionately affected.

Life Evaluations from the Gallup World Poll provide the basis for the annual happiness rankings and they are based on answers to the main life evaluation questions asked in the poll.

Typically, around 1,000 responses are gathered annually for each country. Weighted averages are used to construct population-representative national averages for each year in each country

Primary User Stories

- As a decision maker in the UN, I want to know the specific factors that have the largest effect on happiness in different countries in order to prioritize policies and apportion resources that would have the most dramatic impact on the well-being of citizens across the world.
- As a decision maker in the UN, I also want to know how COVID-19 affected the well-being of different regions across the world. We must ensure that our response to the health crisis and the distribution of our resources are based on the extent to which each country was affected.

M

Data source

- World Happiness Report 2015-2021 dataset
    - Sourced from Kaggle
- A publication of the Sustainable Development Solutions Network
    - Global Initiative for the United Nations
- Powered by Gallup World Poll data, which has been collected every year since 2005
- License Concerns: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

World Happiness Report

S

Dataset Description

- Description
  - 9-20 fields/columns (different across years)
  - About 160 countries
  - Total records = 1,085 spread across 7 csv files
- Variables
  - Main - Happiness Index - Happiness Score/Life Ladder
    - uses Cantril Ladder Scale to assess general life satisfaction
  - Social support, freedom, perceived corruption, generosity
    - Binary data - 0 or 1 (no or yes)
    - Data value is the averages
  - Healthy life expectancy
    - The average amount of healthy years over a person's life
  - GDP per Capita - measure of economic prosperity of country
- Data to be altered for consistency:
  - Some years have missing log GDP per capita values
  - Fields that are outside our scope or only included in part of the data will be excluded

S

Lexicon

Gallup World Poll: Tracks important issues worldwide, such as food access, employment, leadership performance, and well-being

Happiness Score/ Ladder Score: Measure of Overall Life Satisfaction Score

Uses Cantril Ladder Scale

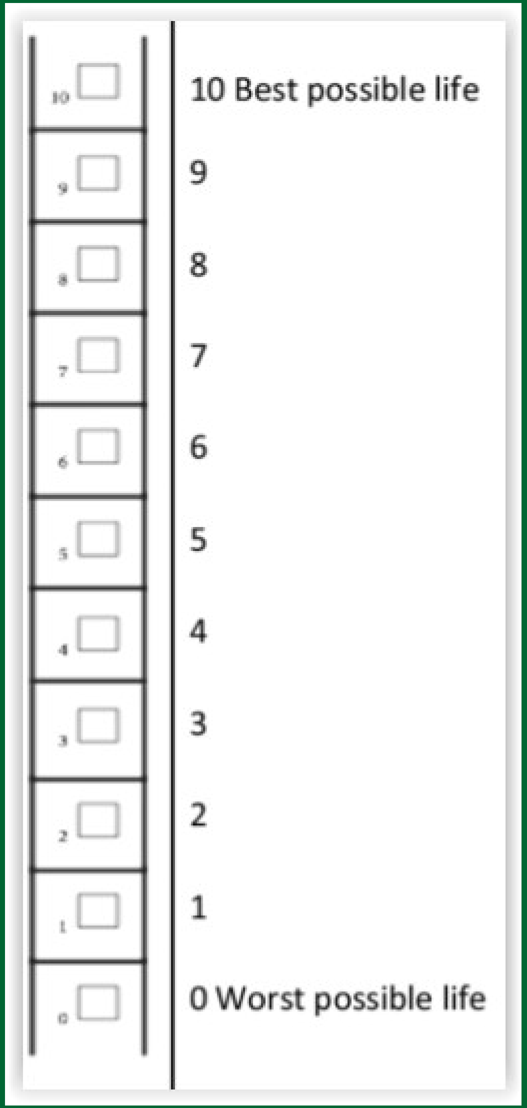
Cantril Ladder: Simple visual scale to assess general life satisfaction

Log GDP per Capita: Total value of all finished goods & services produced as a proportion to a country’s population

Social Support: Social support score for the country

Healthy Life Expectancy: Avg Healthy Life Expectancy at Birth score

Freedom to make life choices: Freedom score of people in the country to choose what they do with their life



Cantril Ladder

s

Other Data Sources

- Gallup Analytics - https://www.gallup.com/analytics/213617/gallup-analytics.aspx
  - Needs subscription access

- Kaggle - other datasets
  h
  ttps://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021
  - Looks Altered/Manipulated
  - Not complete
  - Data from 2005-2021 - not interested in data from 2004-2014

Risks and Mitigation

Possible Risks	Mitigation
● Inconsistent or incorrect data for metrics such as GDP per capita ○ Medium probability ○ Low impact	● Using multiple sources where appropriate in order to verify accuracy ● Transformation of the preexisting data for consistency between datasets
● Too much data to process locally ○ Very low probability ○ Medium impact	● Use cloud service such as AWS RDS and a database management system such as Oracle MySQL ● Drop unneeded fields from each dataset that are beyond our scope
● Analytical methods fail ○ Low probability ○ Medium impact	● Change type of statistical tests and visualizations implemented ○ Parametric vs. non-parametric tests depending on distributions ○ Tests that are appropriate for ordinal data such as the Wilcoxon Signed Rank test

J

Project Risk - Inconsistent data across years

- Each dataset for every year from 2015-2021 have consistent availability for these fields:
    - Country, Happiness Score, Healthy Life Expectancy, Perceived Corruption, Freedom, Social Support, and Generosity
    - Statistical appendices from nonprofit show consistent measurement methods
- Log GDP per Capita is missing from years 2015-2019
    - Needed additional source, low impact
    - Missing values were sourced from the World Bank

J

Project Risk - Too much data to process locally

- Redundant fields were removed from datasets
  - Region - only a field in certain years and can be derived from the country
  - Statistics such as quartiles, rankings, and multiple regression coefficients that were calculated by the data collectors
- Sizes of the datasets are therefore smaller and easier to process locally

D

Preliminary Analytics - Initial Goals

- Overall: Analyze World Happiness data collected from 2015-2021
  - factors, correlations, relationships, outliers
  - through visualizations, tables, queries
- Determine importance of factors that affect happiness levels
  - how the factors vary between countries
- Analyze how the COVID-19 health crisis affected happiness across different regions - Years before 2020 vs. Years during/after 2020

D

Analysis - Tools Used

- Python
  - Pandas
  - numpy
  - matplotlib
  - seaborn
  - scipy
  - sklearn
  - plotly
  - GDAL
  - GeoPandas
- Blackboard Collaborate for Teams
- GroupMe
- YouTrack

D

Analytics Overview

- Summary statistics - Happiness distribution across different countries
- Paired t-test - Determine if the happiness scores decreasing due to COVID are statistically significant
- Multiple regression - Examine the strength of the relationship between the happiness score and our independent variables (social support, generosity, corruption, GDP)

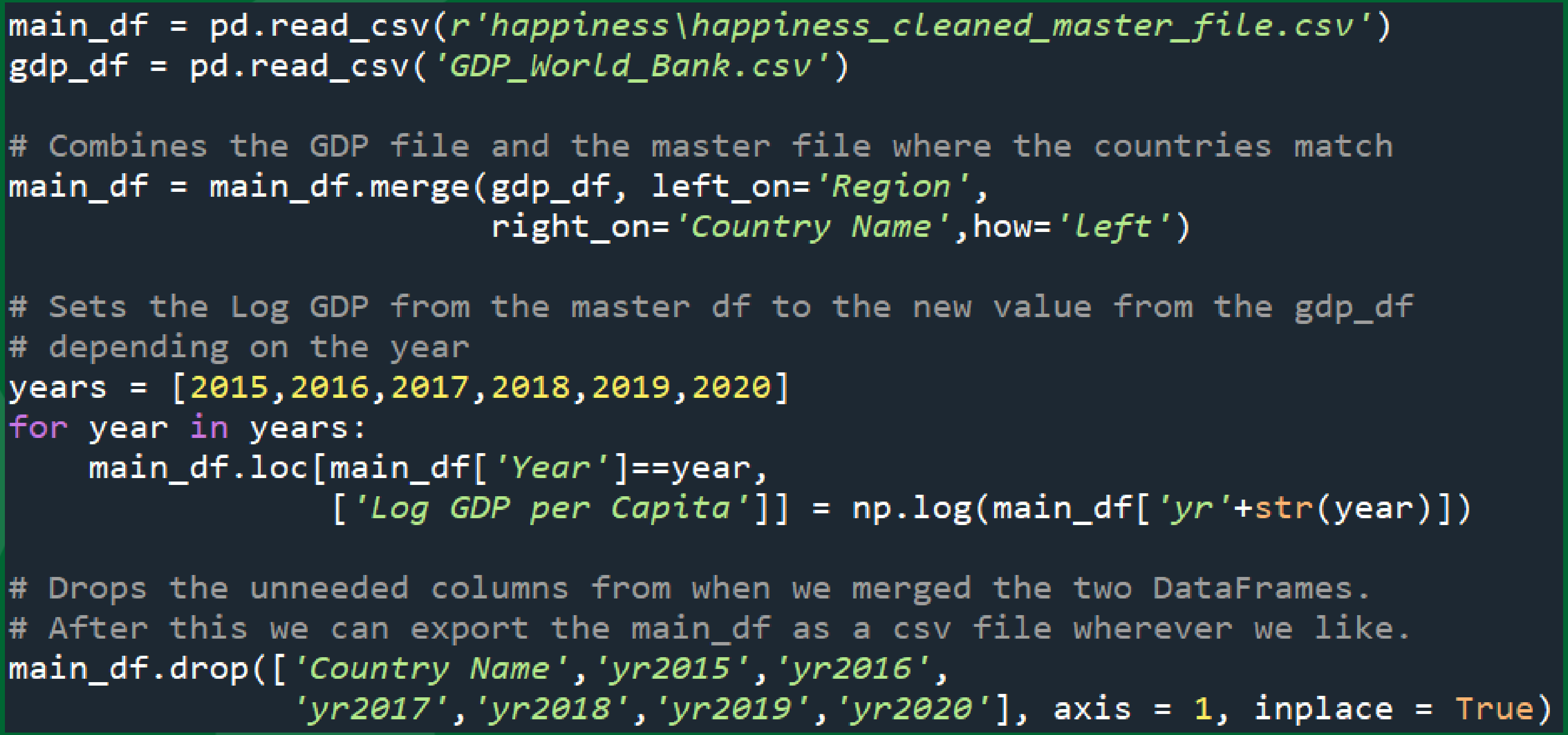
J

Data preparation

- Paired t-test
  - Countries between years had to be consistent
  - 144 countries
- Multiple regression
  - Missing values for various variables
- Other
  - Inconsistent variable names across years
  - Redundant columns
  - Missing Log GDP

Data preparation

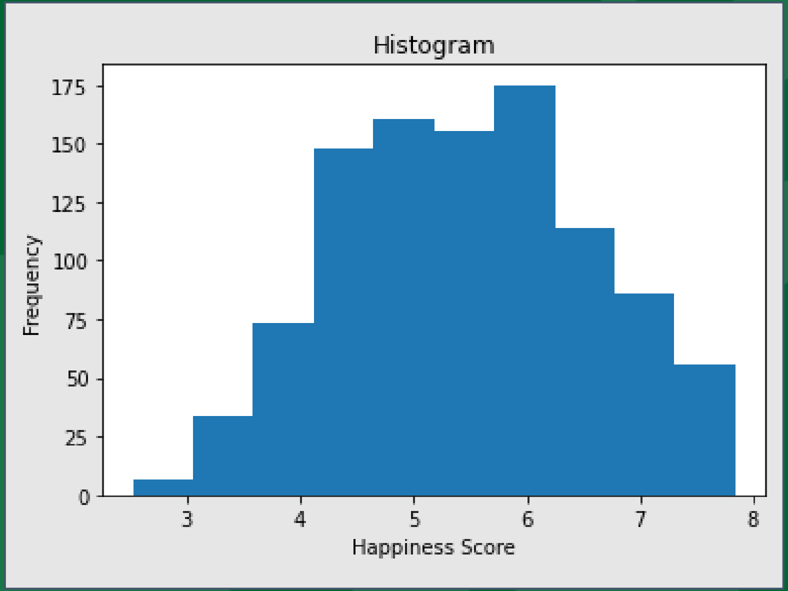
- Log GDP was missing/incorrect for the years 2015-2019
  - Replacement data is collected from the World Bank

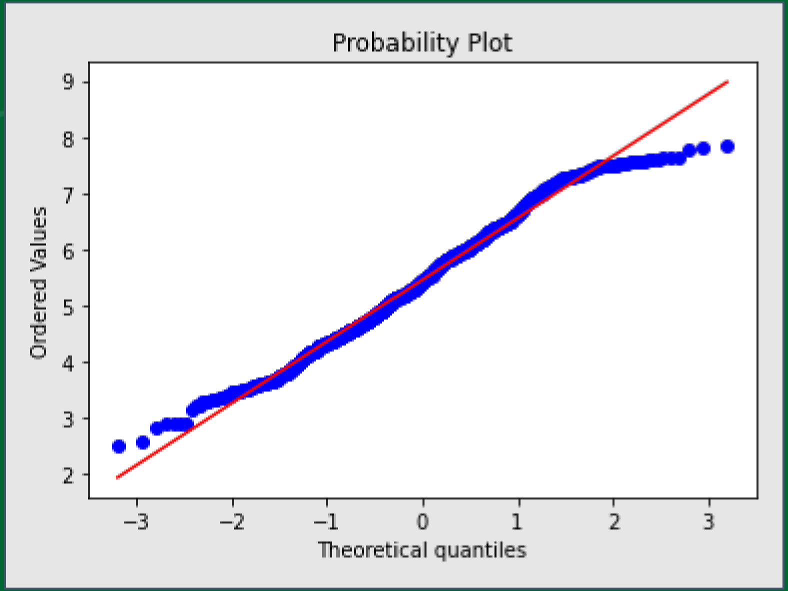


```
main_df = pd.read_csv(r'happiness\happiness_cleaned_master_file.csv')
gdp_df = pd.read_csv('GDP_World_Bank.csv')

# Combines the GDP file and the master file where the countries match
main_df = main_df.merge(gdp_df, left_on='Region',
                        right_on='Country Name',how='left')

# Sets the Log GDP from the master df to the new value from the gdp_df
# depending on the year
years = [2015,2016,2017,2018,2019,2020]
for year in years:
    main_df.loc[main_df['Year']==year,
                ['Log GDP per Capita']] = np.log(main_df['yr'+str(year)])

# Drops the unneeded columns from when we merged the two DataFrames.
# After this we can export the main_df as a csv file wherever we like.
main_df.drop(['Country Name','yr2015','yr2016',
              'yr2017','yr2018','yr2019','yr2020'], axis = 1, inplace = True)
```

M

Normality





Jarque-Bera statistic: 20.282173131318714
p-value: 3.942594067030125e-05

Looking at these results, we fail to reject the null hypothesis and conclude that the sample data follows normal distribution.

M

Paired t-test

```
Ttest_relResult(statistic=-3.4481190392741348, pvalue=0.0007418106233208259)
```

Ho: The mean pre-covid and post-covid scores are equal

Ha: The mean pre-covid and post-covid scores are not equal

Since the p-value (0.0007) is less than 0.05, we reject the null hypothesis. We have sufficient evidence to say that the true mean is different for the happiness score before and after covid.

D

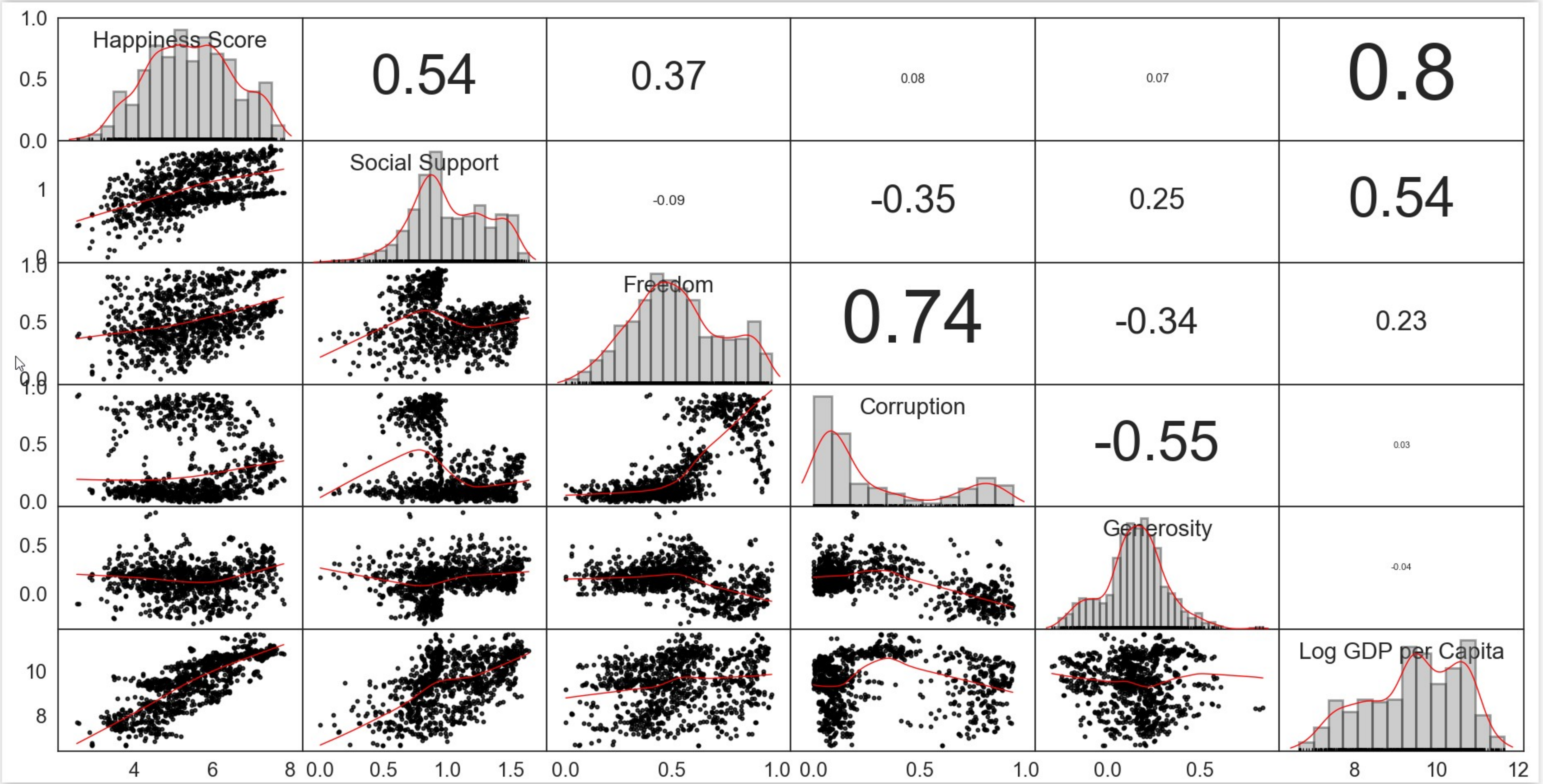
Multiple Regression - Results (2015 - 2019)

- Pandas
- Numpy
- Sklearn

- Strength of relationship between the happiness score and independent variables
  - Social Support, Healthy Life Expectancy, Freedom, Corruption, Generosity, GDP
- Regression Coefficients: relationship significance
  - [0.6399 0.8473 0.9906 1.0538 0.8037 0.4359]
- R-Squared: level of correlation
  - 0.823
- Root Mean Square Error (RMSE): prediction accuracy
  - 0.468

S

ScatterPlot Matrix - Variable Distribution / Pairwise Relationships



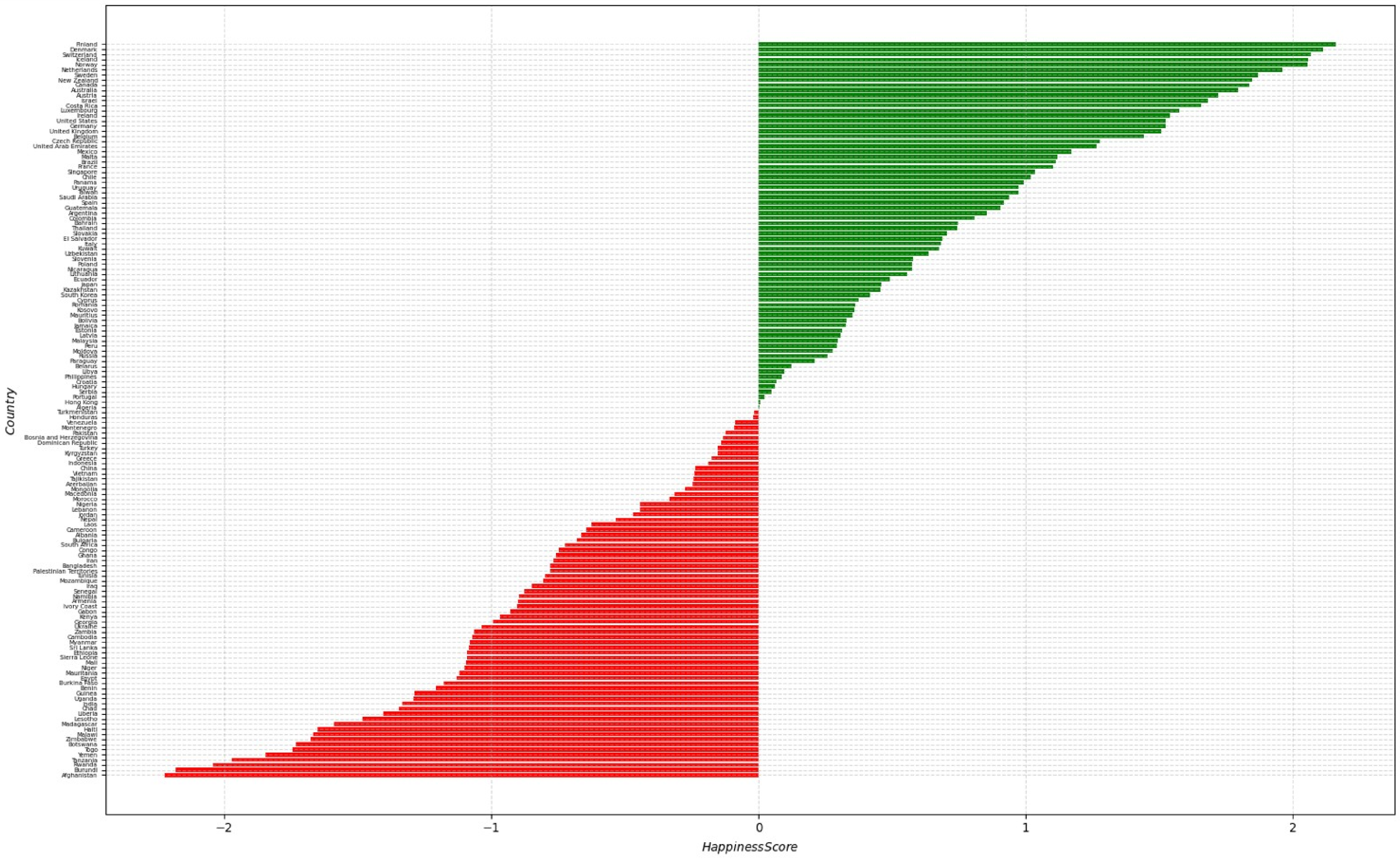
S

Visualization Goals

- Show how happiness levels have changed over time
  - COVID-19’s effects (if any)
- Examine happiness levels in the most affected regions
- Investigate the role of a country's wealth in its happiness levels over time
- Find relationships between variables and their distributions

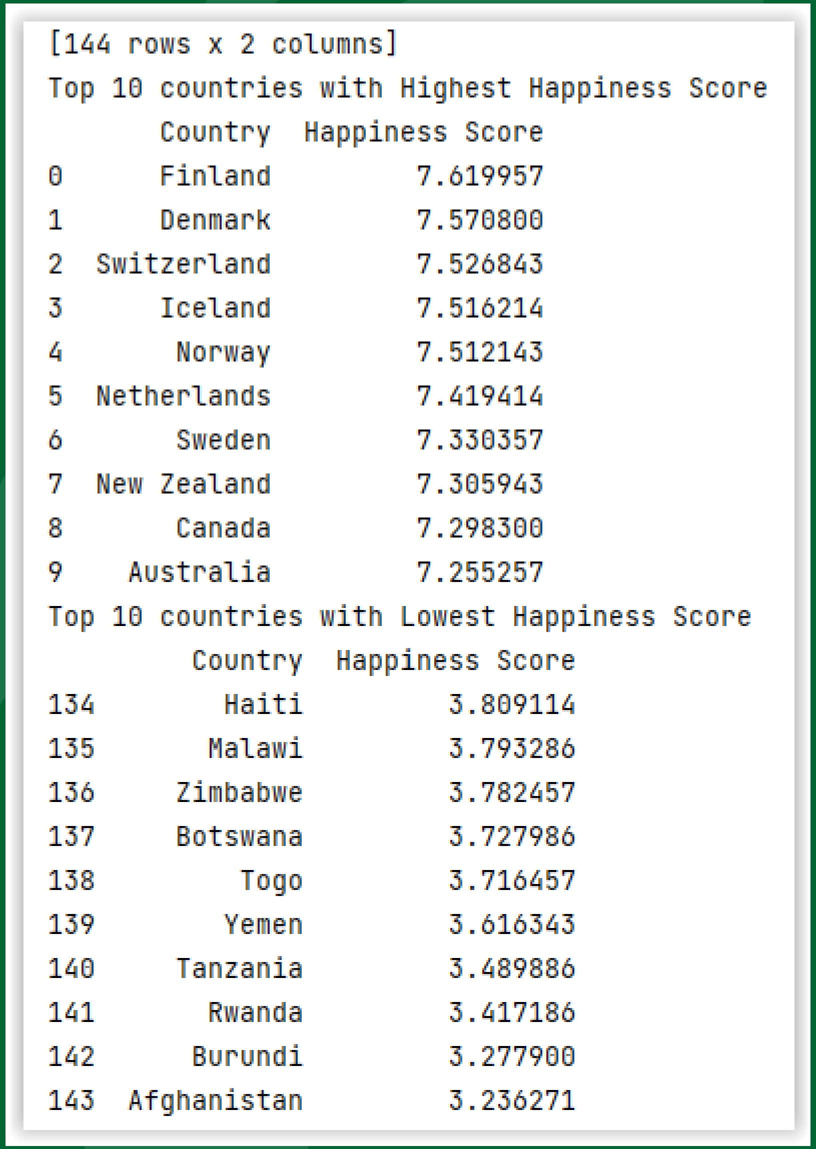
S

Ranking Countries based on Happiness Score



S

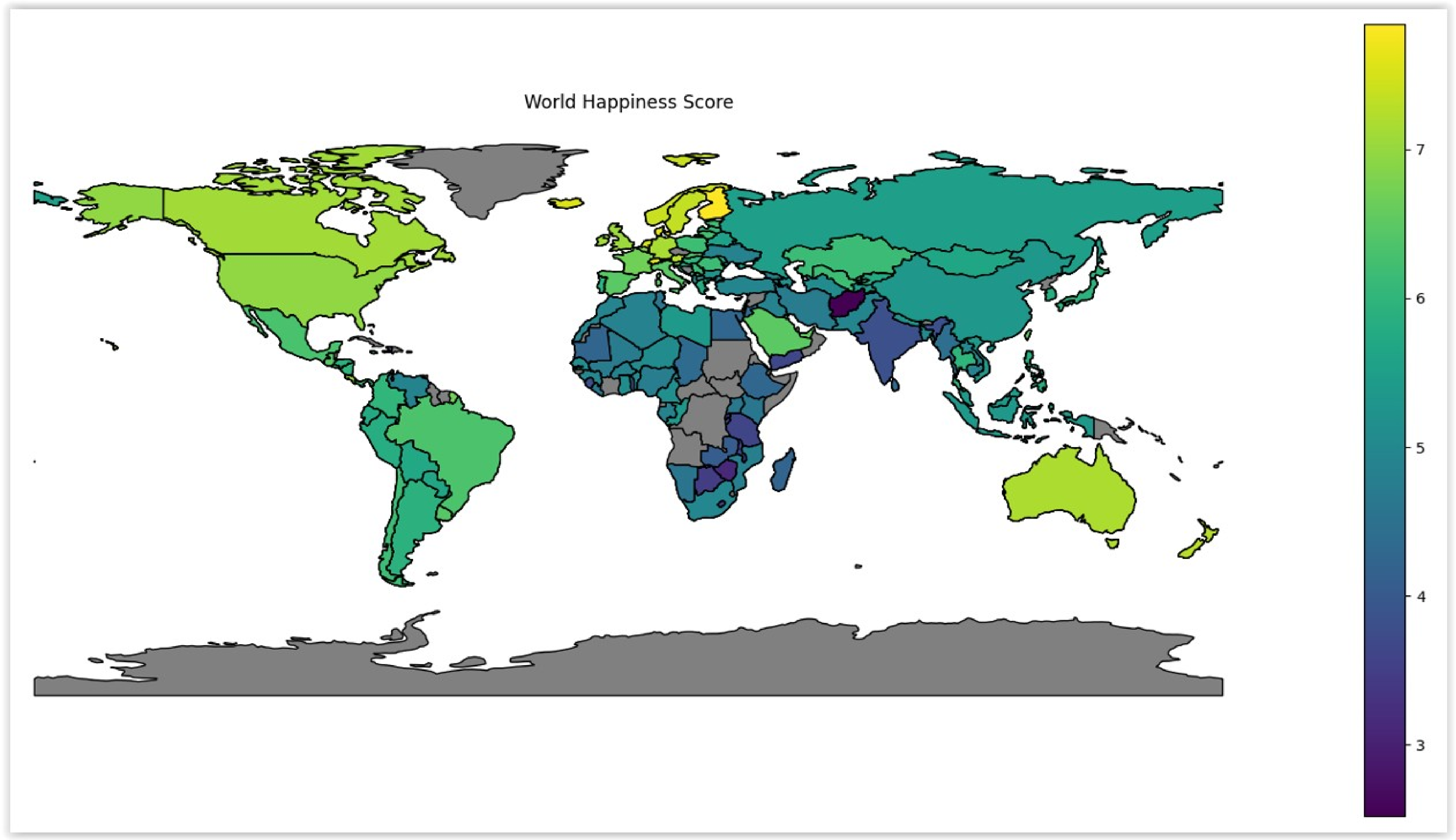
Ranking Countries based on Happiness Score



```
[144 rows x 2 columns]
Top 10 countries with Highest Happiness Score
       Country  Happiness Score
0      Finland         7.619957
1      Denmark         7.570800
2  Switzerland         7.526843
3      Iceland         7.516214
4       Norway         7.512143
5  Netherlands         7.419414
6       Sweden         7.330357
7  New Zealand         7.305943
8       Canada         7.298300
9    Australia         7.255257
Top 10 countries with Lowest Happiness Score
         Country  Happiness Score
134        Haiti         3.809114
135       Malawi         3.793286
136     Zimbabwe         3.782457
137     Botswana         3.727986
138         Togo         3.716457
139        Yemen         3.616343
140     Tanzania         3.489886
141       Rwanda         3.417186
142      Burundi         3.277900
143  Afghanistan         3.236271
```

S

Happiness Score across the Globe



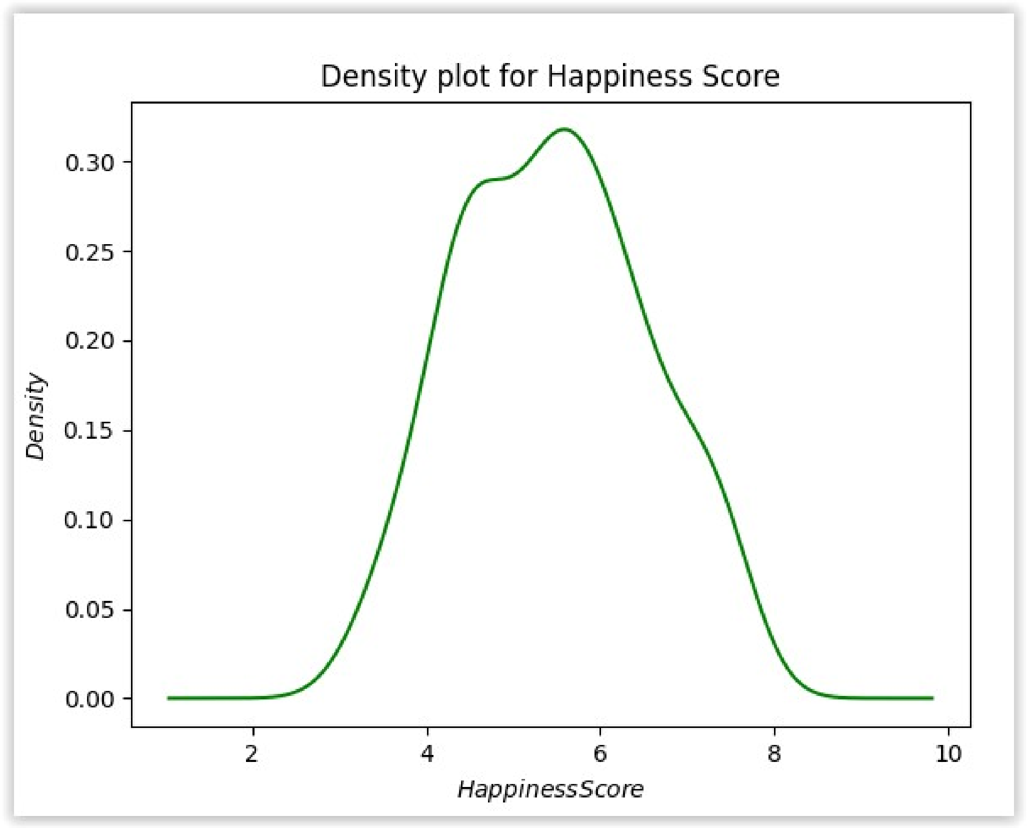
Python Libraries Used

GDAL

Geopandas

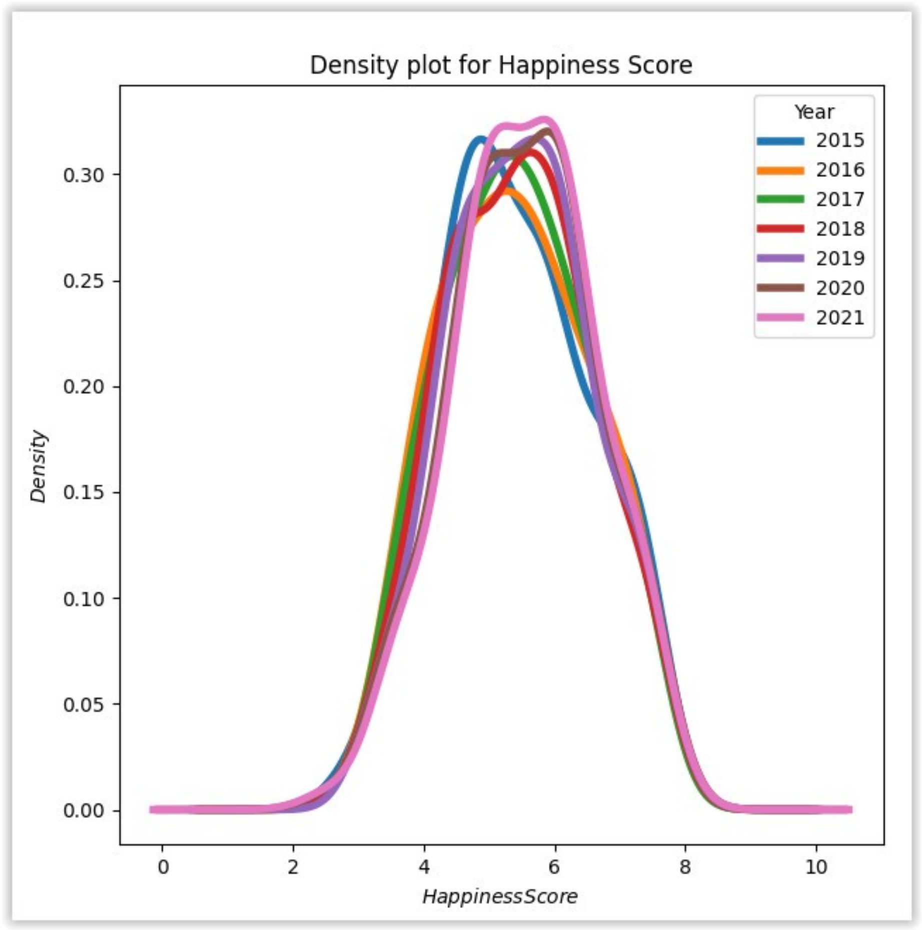
S

Density Analysis



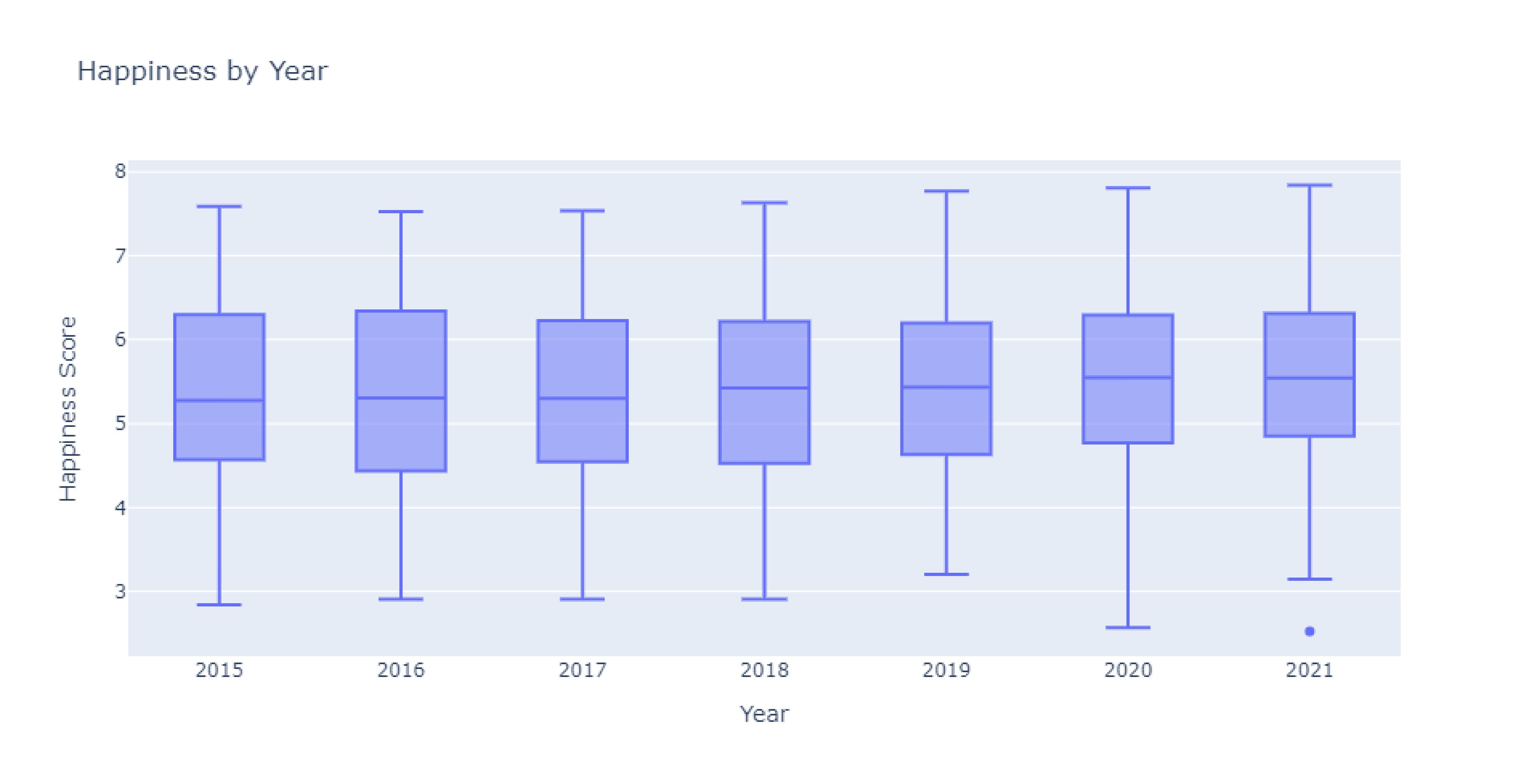
Density Analysis for Mean Happiness Score

Density Analysis for Happiness Score over years



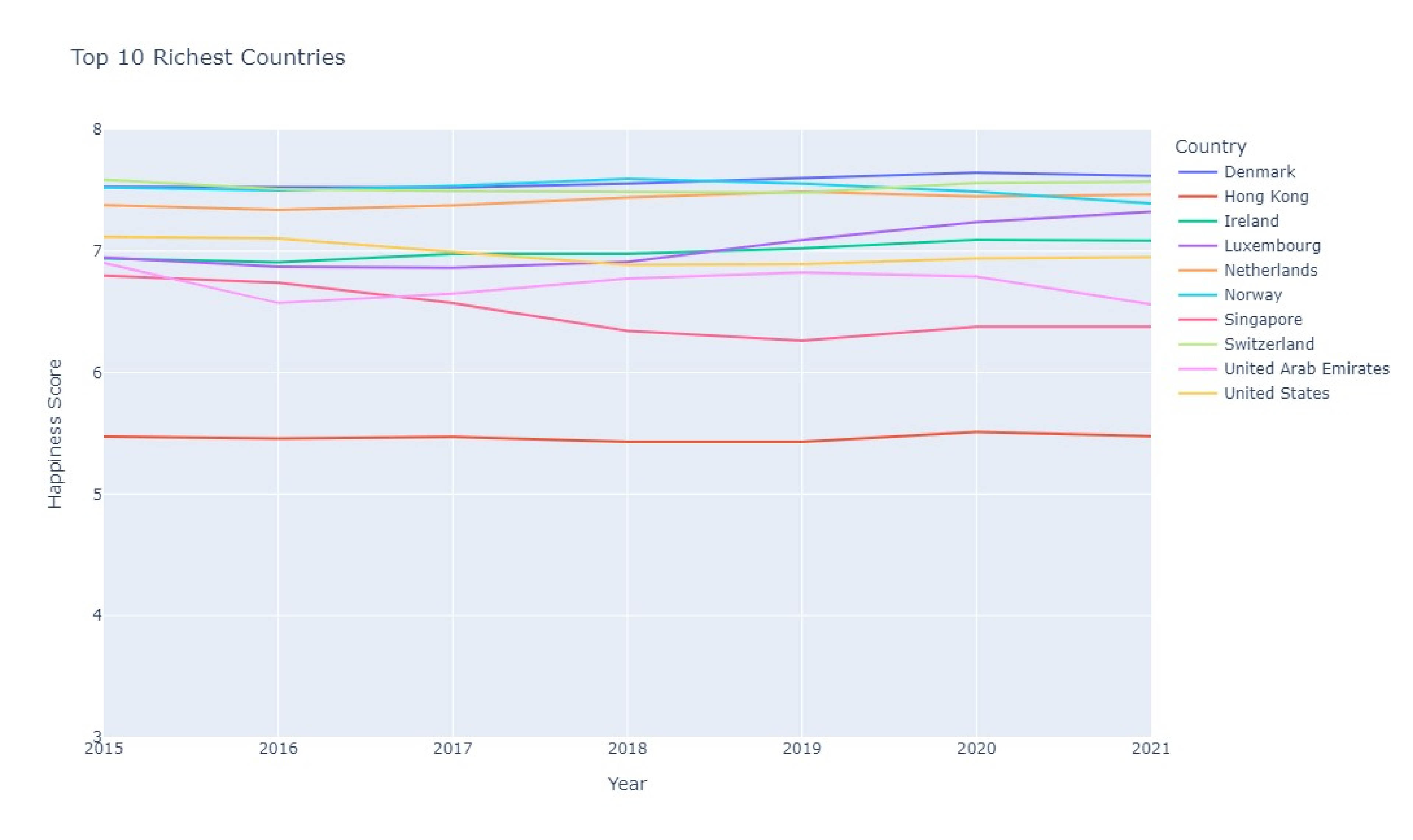
J

Happiness boxplots 2015-2021



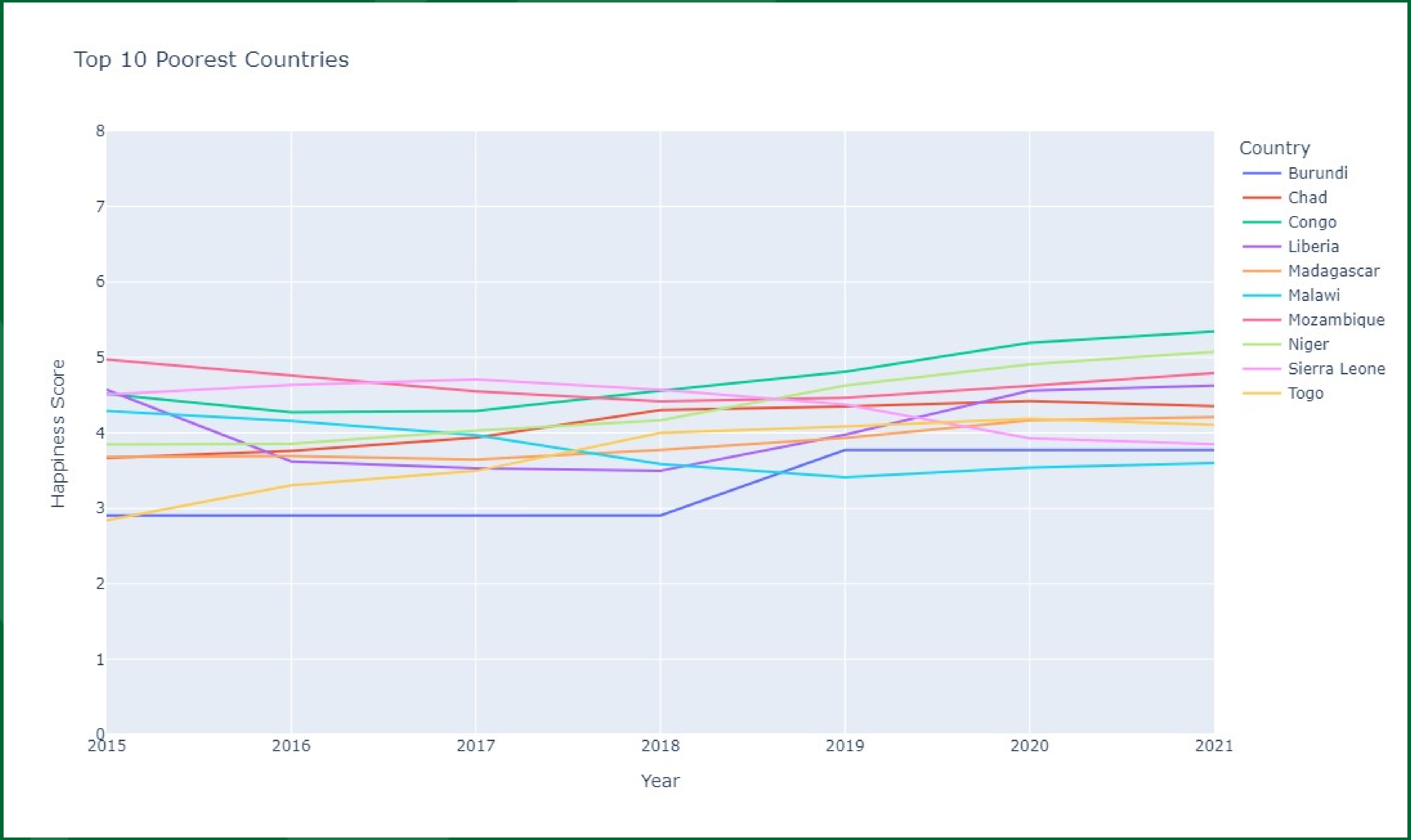
J

Richest countries and their happiness levels



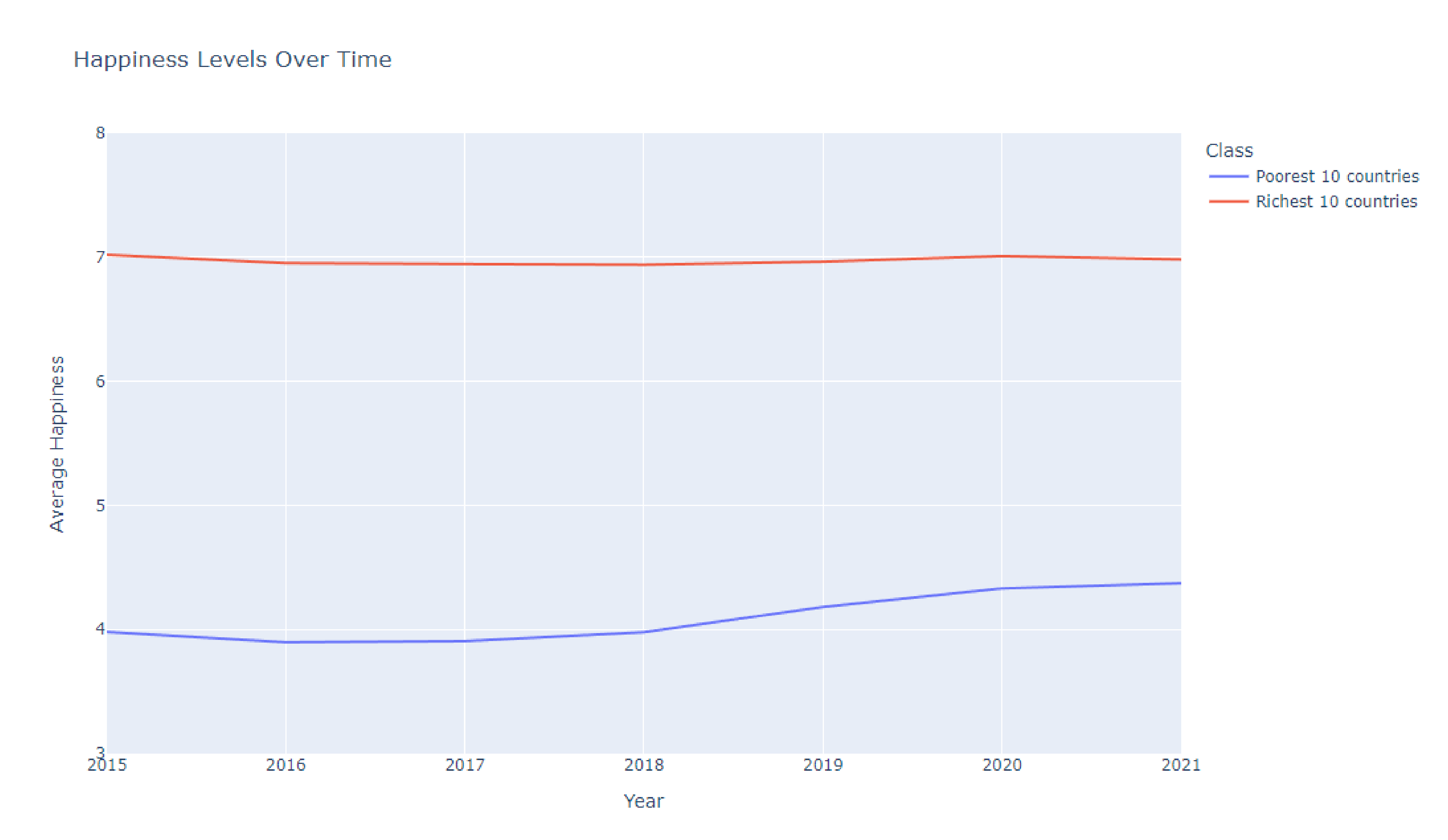
J

Poorest countries and their happiness levels



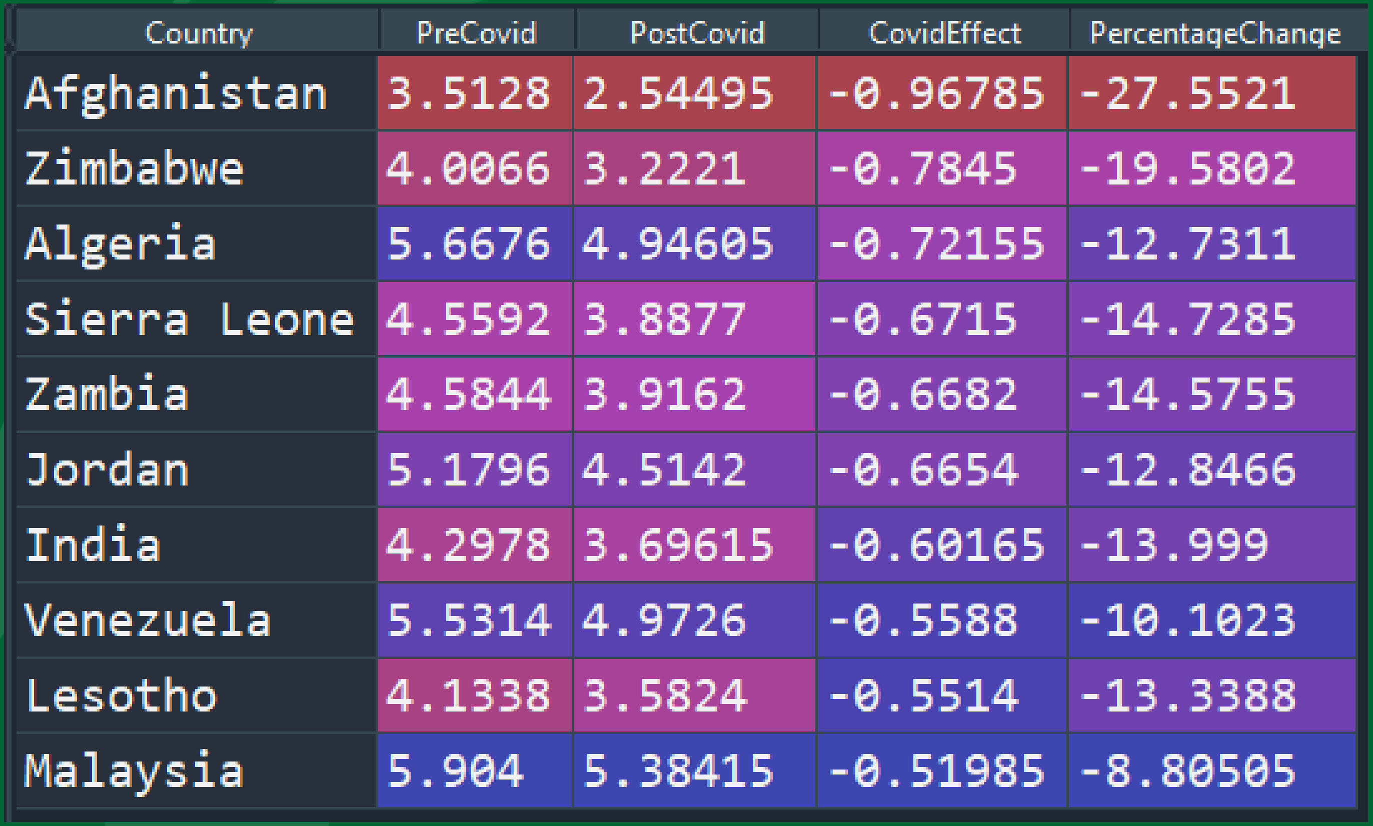
J

Average happiness levels separated by wealth



J

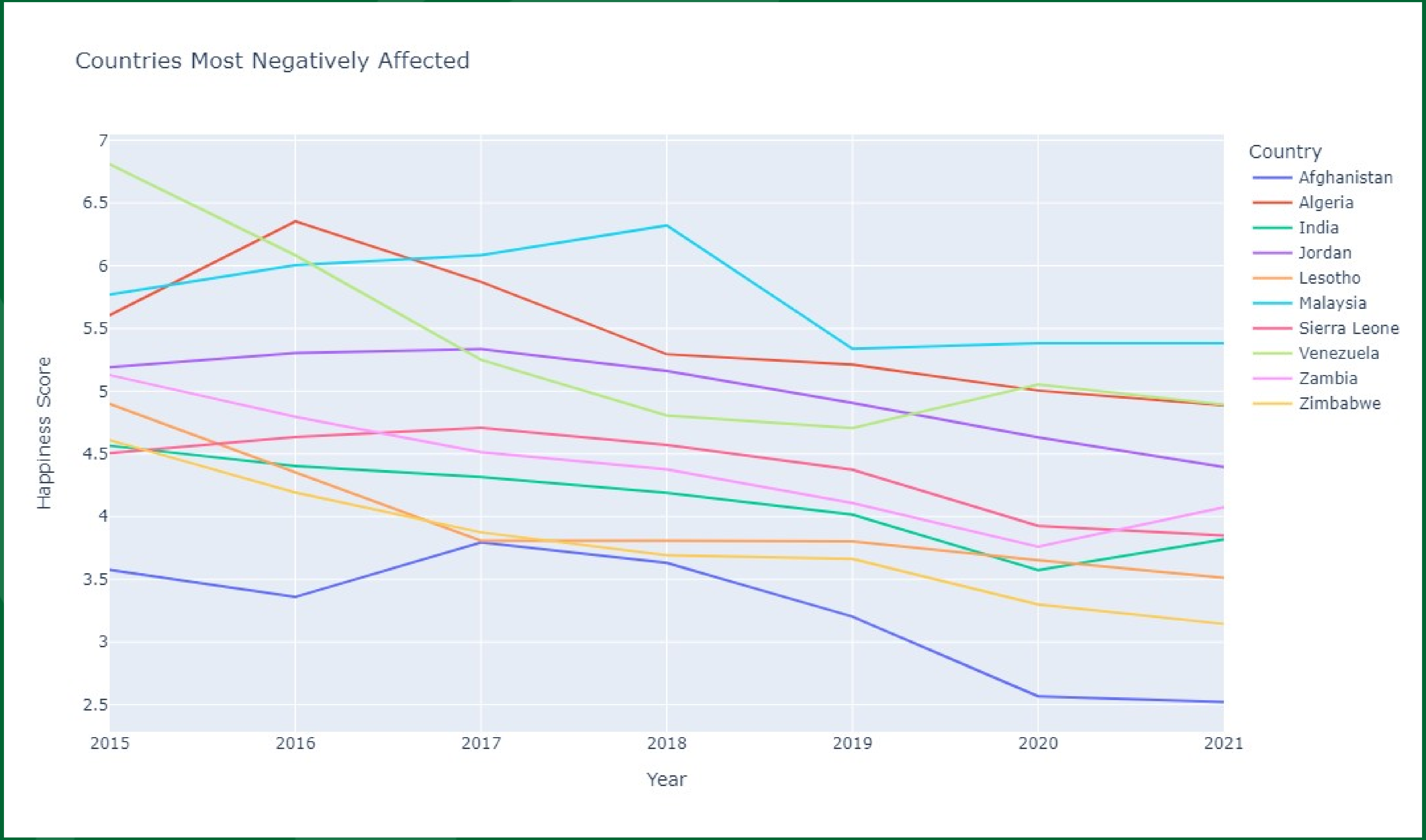
Most negatively affected countries



Country	PreCovid	PostCovid	CovidEffect	PercentageChange
Afghanistan	3.5128	2.54495	-0.96785	-27.5521
Zimbabwe	4.0066	3.2221	-0.7845	-19.5802
Algeria	5.6676	4.94605	-0.72155	-12.7311
Sierra Leone	4.5592	3.8877	-0.6715	-14.7285
Zambia	4.5844	3.9162	-0.6682	-14.5755
Jordan	5.1796	4.5142	-0.6654	-12.8466
India	4.2978	3.69615	-0.60165	-13.999
Venezuela	5.5314	4.9726	-0.5588	-10.1023
Lesotho	4.1338	3.5824	-0.5514	-13.3388
Malaysia	5.904	5.38415	-0.51985	-8.80505

J

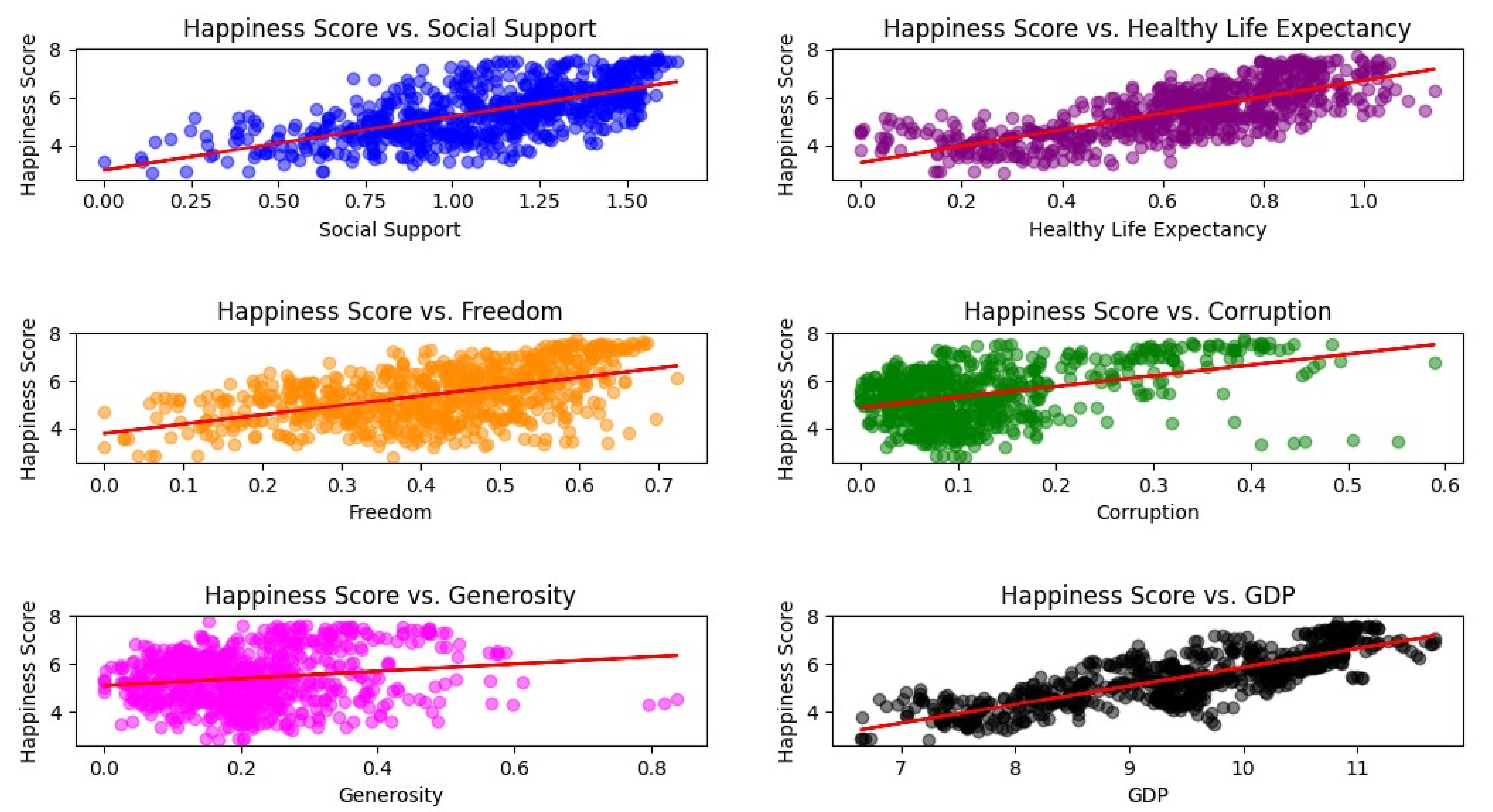
Most negatively affected countries



D

Linear Relationships - Effect on Happiness Score (2015 - 2019)

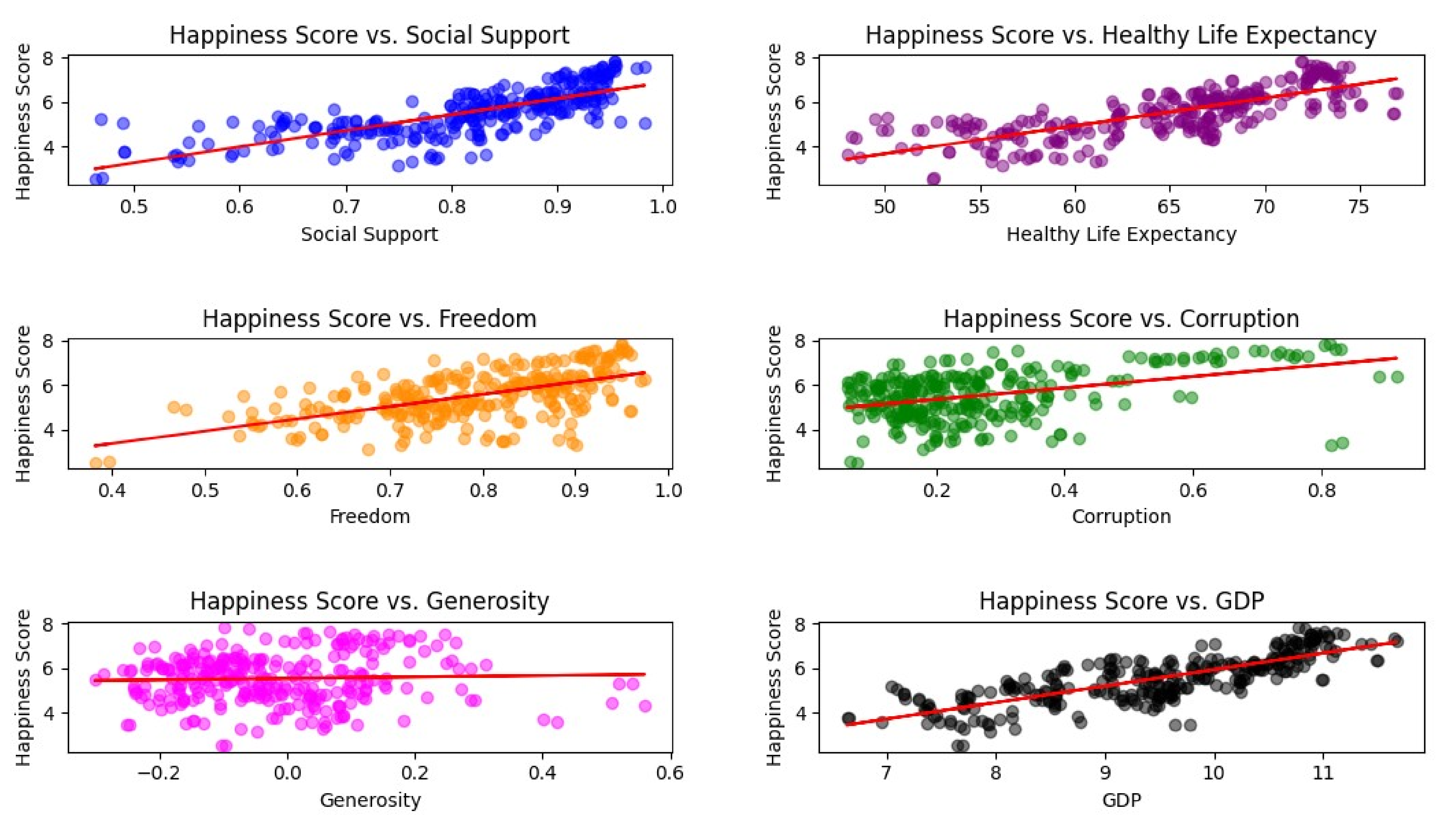
- Pandas
- Numpy
- Matplotlib



D

Linear Relationships - Effect on Happiness Score (2020 - 2021)

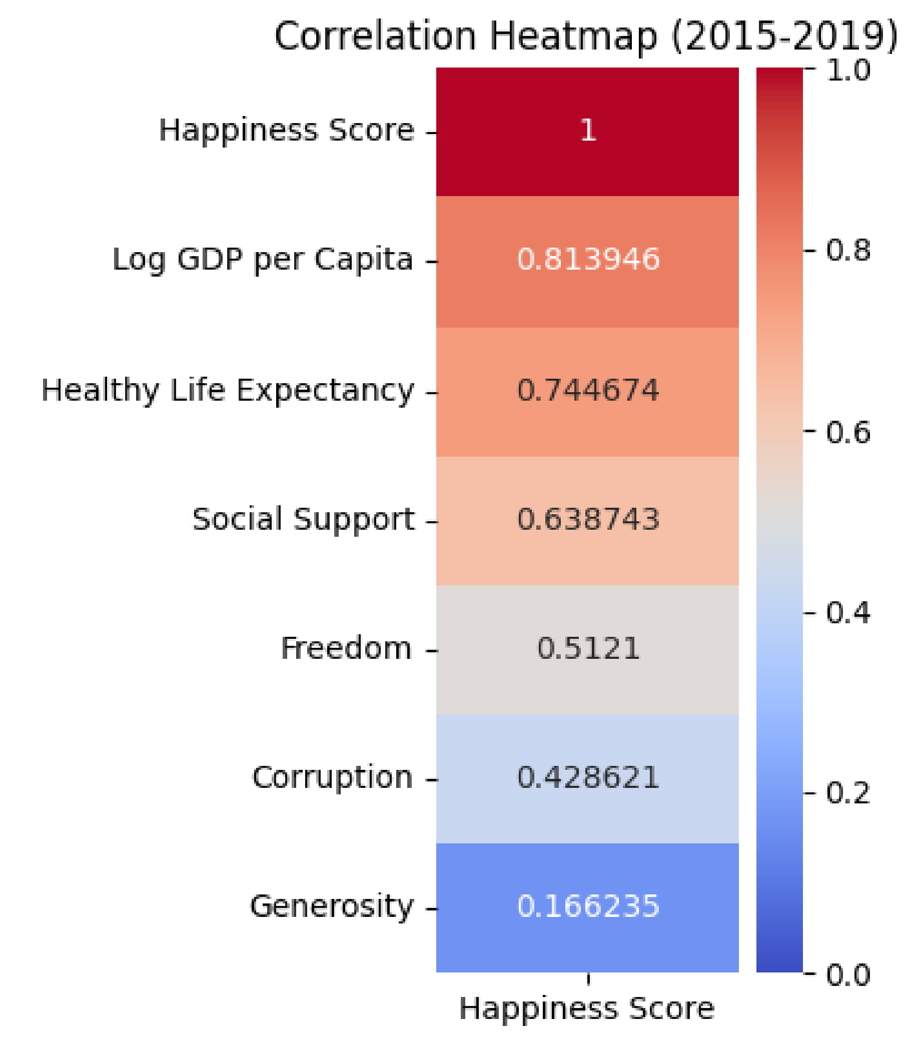
- Pandas
- Numpy
- Matplotlib

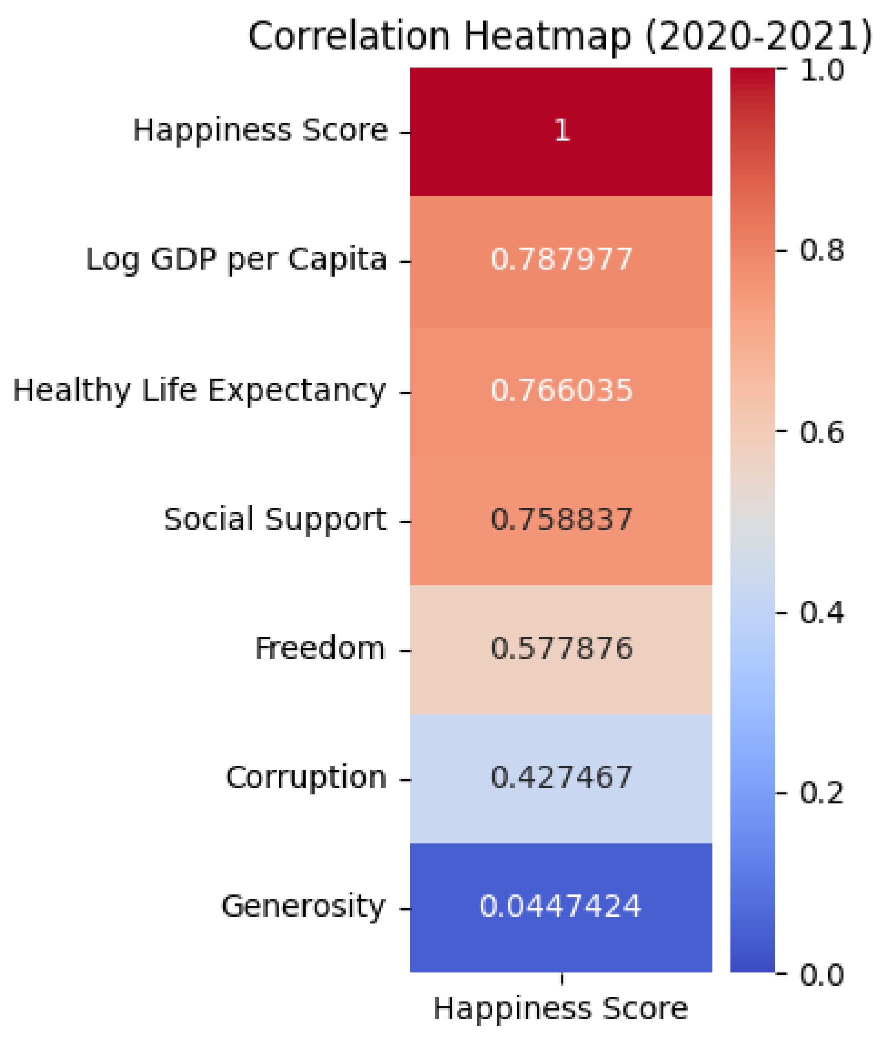


D

Correlation Coefficients

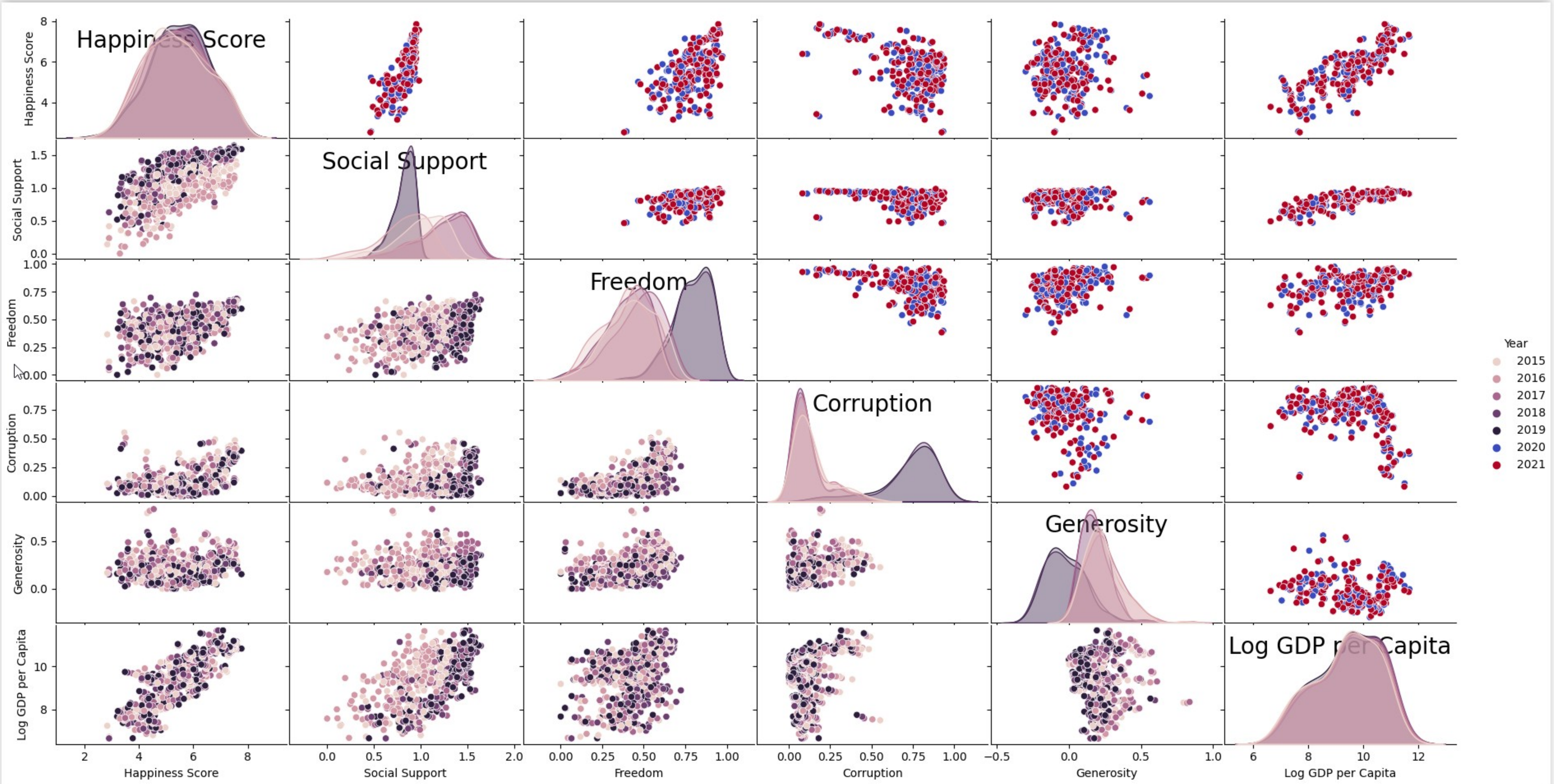
- Pandas
- Numpy
- Matplotlib
- Seaborn





M

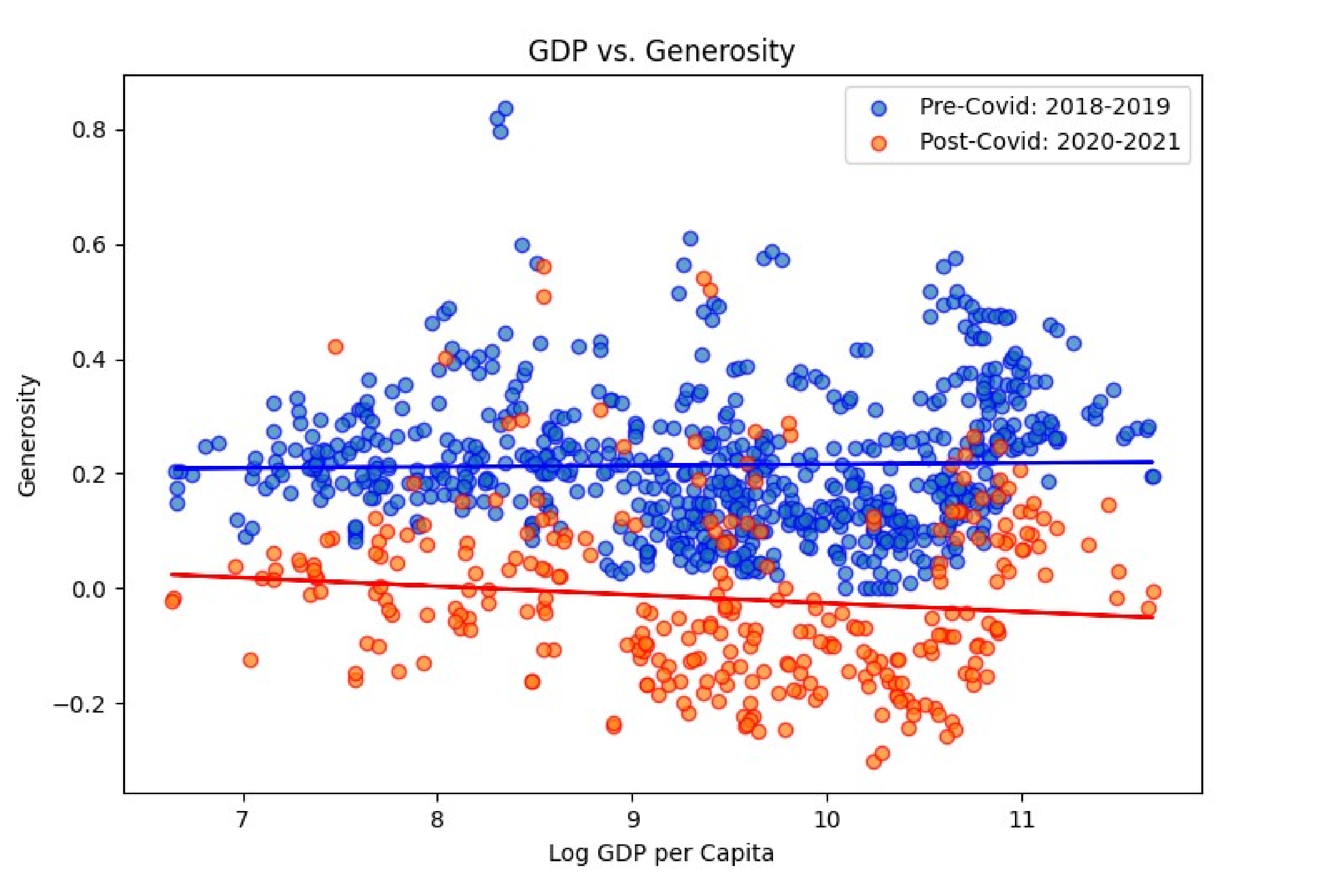
Scatterplot Matrix (Lower Triangle - PreCovid, Upper Triangle - Post Covid)

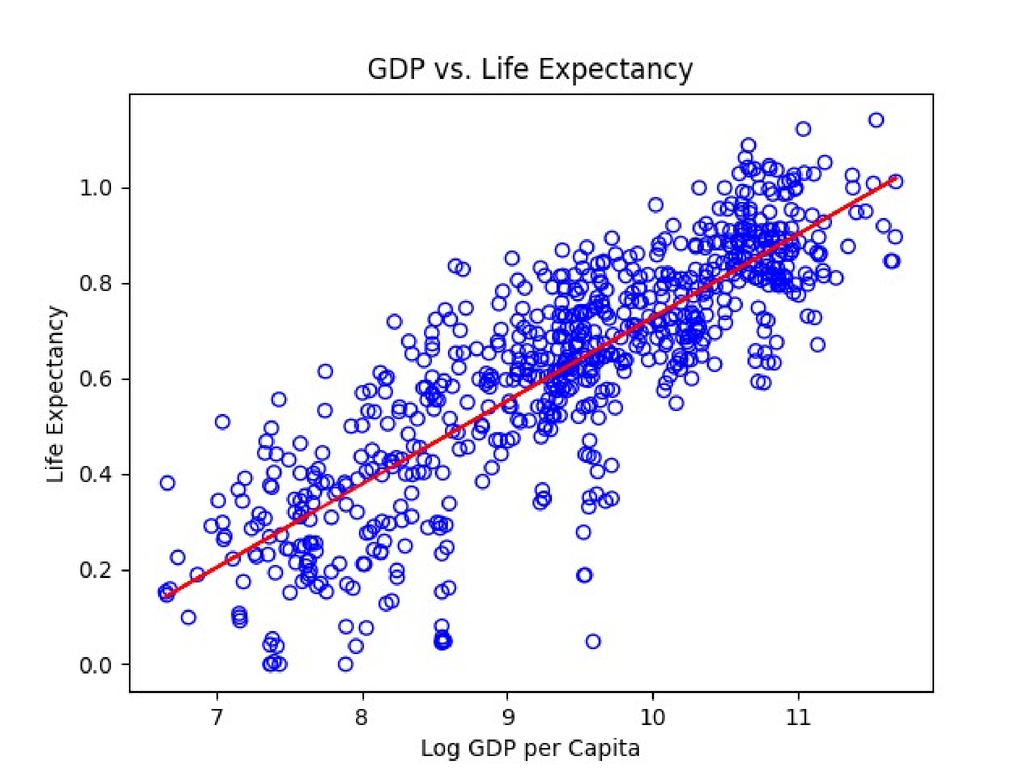


M

Log GDP per Capita Findings

- Pandas
- Numpy
- Matplotlib





M

Conclusion - Findings

- Happiness levels remained resilient during COVID
- The most important variable to determine happiness levels is GDP
- During COVID social support increased and had a stronger relationship with happiness
- Generosity and Corruption have the least effect on happiness
- Happiness score has a normal distribution, but more right-skewed during COVID
- Poor countries had stable happiness growth throughout all 7 years, while the richer countries remained largely the same
- People were less generous during COVID (less likely to donate money)

M

Future Work

- We would like to look at other factors and how they affect happiness
    - Education level
    - Exercise
    - Work-life balance
    - Most negatively affected countries and their commonalities
- Look at past decades
    - Search for trends

Questions?

GEORGE MASON UNIVERSITY