George Mason University

CS 504

Spring 2022

World Happiness and COVID-19

Final Paper

WFT

Mitch Breeden, Jacob Baisden, Delena Bell and Sudha Jain

Table of Contents

Abstract

We analyzed the World Happiness data collected from the Gallup World Survey from 2015-2021 in order to determine the factors that affect subjective happiness levels in each nation. In addition, we analyzed how the COVID-19 health crisis affected happiness across the world. The resulting analytics from our research can help governments and organizations that are interested in studying how the population's happiness and quality of life was affected by the COVID-19 pandemic. Companies and organizations could be interested in the difference of happiness levels in individual countries to help market their products and campaigns.

Our findings showed that the happiness scores during the years after the start of the pandemic (2020-2021) had strong resilience. However, the distribution of happiness scores were more right-skewed during COVID. When separating countries by wealth, we found that the poorer countries had stable happiness growth throughout all 7 years, whereas the richer countries remained stagnant.

GDP was found to be the most important variable in determining happiness levels. Generosity and Corruption were determined to have the least effect on happiness. During COVID, social support increased and had a stronger role in determining happiness whereas generosity decreased. Further work includes looking into other factors that could affect happiness, such as education level and exercise. We could also look further back and see the trends across the past couple of decades to identify trends and create predictive/descriptive models.

1 Introduction

1.1 Background and Rationale

The World Happiness Report examines the subjective happiness levels in different countries and attempts to determine the various economic, political, and cultural factors that affect the rankings. This data is collected through the Gallup World Poll. The report is published through the Sustainable Development Solutions Network, a nonprofit that works with the UN in order to "mobilize global scientific and technological expertise to promote practical problem solving for sustainable development and implement the Sustainable Development Goals". (https://www.un.org/en/academic-impact/page/sdsn)

1.2 Research

The data was collected by Gallup, a global analytics and advice firm with the mission to help organizations solve problems. They conducted their first survey which focused on the thoughts, feelings, and behaviors of people in the United Kingdom in 1938. The survey has since been expanded to over 160 countries. Gallup claims that the Gallup World Poll is the "most comprehensive and farthest-reaching survey of the world" (Gallup Global Research).

(https://www.gallup.com/analytics/318875/global-research.aspx)

1.3 Project Objectives

The overall objective of this project is to analyze the Word Happiness data collected from 2015 through 2021 to better understand happiness in different regions and the impact COVID-19 had on happiness. We want to analyze the data using various techniques including queries, visualizations, and tables to find interesting or notable factors. These factors can include relationships, outliers, and correlations. The reason and mission behind wanting to study these relationships through analytics is to better the understanding of happiness in certain regions to improve government policy and assist companies with marketing.

1.4 Problem Space

We will be analyzing World Happiness data collected from the Gallup World Survey from 2015-2021 in order to determine the factors that affect subjective happiness levels in each nation. We are also interested in how the COVID-19 health crisis affected happiness across the world, and if any regions were disproportionately affected.

1.5 Primary User Story (-ies):

User Story 1

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.

User Story 2

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.

1.6 Solution Space

Our system will provide valuable insights by showing the factors that contribute to happiness levels around the world. This analysis will help UN leaders pinpoint areas where potential action can be taken to improve happiness levels. Additionally, the responses to the COVID-19 pandemic will be evaluated to infer how or whether these actions affected happiness. We expect this report will be of great value to decision makers and users alike.

1.7 Product Vision - Sample scenarios (why would someone want to use this)

Scenario #1

The resulting analytics from our research can help governments and organizations that are interested in studying how the population's happiness and quality of life was affected by the COVID-19 pandemic. Results from the analytics acquired will give governments the ability to learn how to better prepare for the well-being of society during a pandemic or similar catastrophic, large-scale event. Using these analytics, the government can create organizations to focus on improving the mental and physical health of those struggling with their happiness levels during the trauma of a pandemic or the mental toll of a society-lockdown. There could be concerns with the validity of the analytics results caused by the survey. The way the survey was conducted could cause concerns of validity due to the subjectivity of the questions and answers.

Scenario #2

Companies and organizations could be interested in the difference of happiness levels in individual countries to help market their products and campaigns. The resulting analytics could give companies insights as to where and who to target certain products or services based on the happiness levels of where a person lives. There could be concerns with the validity of the analytics results caused by the survey. The way the survey was conducted could cause concerns of validity due to the subjectivity of the questions and answers.

1.8 Definition of Terms:

Gallup World Poll – Gallup World Poll tracks important issues worldwide, such as food access, employment, leadership performance, and well-being for various countries across the globe.

Life Ladder/Happiness Score/Score: **Main Happiness Index** – Ladder score or Overall Life Satisfaction Score of the country. It uses Cantril Ladder Scale to assess general life satisfaction.It's a national average response to the question of life evaluations.

Cantril Ladder – Its a simple visual scale to assess general life satisfaction. It evaluates a person's life as a whole using the image of a ladder, with the best possible life for him/her as a 10 and worst possible as a 0.

Logged GDP per Capita/Economy (GDP per Capita) – The total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period.

Social Support/Family – Social Support score for the country

Healthy Life Expectancy/Health (Life Expectancy) – Average Healthy Life Expectancy at Birth score for the country

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

Generosity – Generosity score of the country

Perceptions of Corruption/Trust (Government Corruption) – Overall Government/Business Corruption perceptions in the country

2 Data Acquisition

2.1 Overview:

For this project, we will be using World Happiness Report 2015 -2021 dataset from Kaggle.
https://www.kaggle.com/datasets/mathurinache/world-happiness-report-20152021

The dataset consists of 7 .csv files, one for each year from 2015-2021. The raw dataset consists of 9-18 variables (different across the years) affecting the Happiness Index/Ranking for about 160 countries across the globe, that include 99% of the world population. The main source of the data is **Gallup World Poll**, which asks respondents across various countries to evaluate their current lives. Each year, Gallup World Poll collaborates with the Sustainable Solutions Development Network, an initiative of the United Nations, to produce the World Happiness Report. The Gallup World Poll provides a uniform comparable basis across the world to measure how people value their own lives and how satisfied they are with their lives. The survey consists of around 100 global questions as well as regional specific questions. 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.

2.2 Field Descriptions:

- Country (Type: string) – The name of a country, there are approximately 160 unique values in this dataset. Regions can be derived from this field as well.
- Happiness Score (Type: decimal) – The numeric score given to each country using the Cantril Ladder Scale to assess general life satisfaction.
- Healthy Life Expectancy (Type: decimal) – The ratio of years individuals are expected to live at birth.
- Perceived Corruption (Type: decimal) – The average of binary answers to two questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?"
- Freedom (Type: decimal) – The average of binary responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
- Social Support (Type: decimal) – The average of the responses to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
- Generosity (Type: decimal) – The average of binary responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.
- Logged GDP per Capita (Type: decimal) – The total value of all finished goods and services produced.

2.3 Data Context:

The dataset we are using is from The World Happiness Report. This report attempts to determine the various economic, political, and cultural factors that affect happiness. The dataset, originally collected from the Gallup World Poll, consists of seven CSV files. There is a single CSV for each year from 2015 – 2021. The dataset consists of nine to twenty fields or columns, differing between years. Each year consists of results from approximately 160 countries, for a total of 1,085 records. The happiness index result uses the Cantril ladder scale to assess general life satisfaction. The fields social support, freedom, perceived corruption, and generosity are the average of scores taken during the Poll. The scores in the Poll are a choice of “No” or “Yes” which correspond to zero or one respectively. The field health is a ratio of healthy life expectancy to overall life expectancy.

2.4 Data Conditioning

There are 7 datasets: one for each year between 2015-2021. The fields that overlap between all years are Country, Happiness Score, Healthy Life Expectancy, Perceived Corruption, Freedom, Social Support, and Generosity. The statistical appendices made available by the Sustainable Development Solutions Network shows consistent collection and measurement methods for each dataset. The only metric that’s inconsistent across years is Log GDP per capita. This measure is only available for 2020-2021, it appears that for the years of 2015-2019 the aggregator of the data transformed the values through some type of normalization that we couldn’t reverse engineer. As a result, we will have to find a new source for log GDP per capita to use in conjunction with these datasets since it’s a major factor in the scope of our analysis. In addition, some of the datasets have extra redundant fields depending on the year. For example, there is sometimes a *region* field (Eastern Europe, South America, etc.) which can already be derived by its *country* field. There are also fields that contain statistical calculations done by the data collectors such as quartiles, rankings, and multiple regression coefficients. We’ll also be dropping these fields, since we want to conduct our own analyses. A more minor problem is the inconsistent names of countries, for example some years have *Hong Kong* while others have it as *Hong Kong S.A.R of China*, *Northern Cyprus* instead of *North Cyprus*, *Taiwan Province of China* instead of *Taiwan* and so on. Depending on our analyses we’ll most likely have to change the country field for certain records in order to be more consistent.

2.5 Data Quality Assessment:

- Completeness: There are no null values and each column is in the correct form as specified.
- Uniqueness: Every row is unique because it holds the values for an individual country and there are no duplicates apparent.
- Accuracy: Data is derived from the Gallup World Poll which is a renowned data source and therefore has undergone scrupulous accuracy tests.
- Atomicity: There is high atomicity as each column only has one value coinciding with its field.
- Conformity: Each column stays within the data type boundaries assigned to them.
- Overall Quality: This data is of extremely good quality since it is provided by Kaggle and sourced from a renowned data collection agency.

2.6 Other Data Sources

We had considered few other data sources for our analysis -

- Gallup Analytics - https://www.gallup.com/analytics/213617/gallup-analytics.aspx
  Gallup Analytics Data Access Portal is the original source of World Happiness Data. GMU students have subscription access through the campus library. But we selected the dataset from Kaggle because it is free and has instant access.
- Kaggle - https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021

We also considered this dataset from Kaggle. It consists of 2 .csv files, one for years 2005-2020 and one for year 2021. We didn't select this dataset because after studying the dataset, it looks manipulated/altered and not complete. Also, we didn't want to analyze data from 2005 to 2014 as Gallup World Poll has evolved over years and the variables collected in initial years are different than later years.

3 Analytics and Algorithms

3.1 Data preparation:

For the preliminary analytics that we wanted to conduct, we were interested in summary statistics to see the distribution of happiness levels across different countries, the effect of COVID on happiness levels using a paired t-test, and the strengths of the relationships between the happiness score and our different independent variables in our scope using a multiple regression.

For the paired t-test there had to be a dataset with a uniform number of countries for each year, so some records were removed if it was a country where data was only available for one year. Unfortunately, Qatar, Somalia, and South Sudan were all excluded from this test because they only had data for two or three years each, with nothing available for 2019 or 2020 (which is important since we're interested on COVID's effect on happiness. As a result, for each year there were 144 countries that were included for this test.

For the multiple regression that examined each variable's effect on happiness levels the most important aspect of the data was completeness; there were a few missing values for different variables depending on the year and we had to determine the best way to deal with this. If it was only one year's values that were missing for a country, we took the average of the two surrounding years (for example, if 2016's data for Switzerland was missing, we set that record equal to the average of 2015 and 2017). However, if the year was during COVID we generally kept those values null since those are anomalies and hard to predict reliably.

For all the tests in general we also had to prepare and fix other aspects of the data. FIrstly, the original data was separated into multiple csv files by year. Depending on the year, the datasets had a different name for the same variables (eg: "Ladder score" instead of "happiness score") and they all had to be named consistently. Also, each year's data had different numbers of variables and extra fields for statistical calculations made by the data collectors which had to be removed since we're conducting our own analyses. Problems like these were solved mostly with simple Python scripts that convert the csv files into dataframes which were altered and then appended on one another to create one master csv file.

3.2 Summary statistics:

For Summary Statistics, we wanted to analyze the distribution of Happiness Scores across different countries. For this we used 3 Analytics -

1. **Ranking Countries based on Happiness Score** - For this analytics, we are planning to use the centered data to rank countries based on their Happiness Score. For this, The countries will be ranked based on their deviation from the mean Happiness Score. The results will be plotted using Horizontal Bargraph using Python libraries Pandas and Matpotlib.

2. **Happiness Score across the Globe** - For this analytics, we are planning to plot Happiness Scores for various countries on the world map to analyze how the Happiness Score is distributed across the globe and if being in a particular region has affected a country's Happiness Score. The results will be plotted by mapping Happiness Scores for countries on world map using Python libraries GDAL and Geopandas.

3. **Density Analysis for Happiness Score** - For this analytics, we are planning to do Density Analysis for Happiness Score and analyze how it has changed over years 2015 to 2021. The Density Plot can also tell us how the distribution for Happiness Score is skewed - whether more countries have Happiness Score above or below the average score. The results will be plotted by mapping Happiness Scores for countries on world map using Python libraries Pandas and Matpotlib.

3.3 Paired t-test:

For the paired t-test, our first objective was to make sure the data had a normal distribution. We did this using three different methods. The first two were visualizations (matplotlib histogram and a pylab probability plot) which showed us that there was indeed a normal distribution, but we wanted to be certain. We used the Jarque-Bera test from scipy to get a p-value, which when looking at these results, we failed to reject the null hypothesis and concluded that the sample data follows a normal distribution.

Next, we performed the paired t-test from scipy using two subsets of the data: happiness scores for 2019 and 2020 respectively. This gave us a p-value of 0.0007 which is less than 0.05, so we rejected the null hypothesis (the mean pre-covid and post-covid scores are equal). We had sufficient evidence to say that the true mean is different for the happiness score before and after covid.

3.4 Multiple regression:

We used multiple regression to investigate the strength of the relationships between the dependent variable and the independent variables. The dependent variable is the Happiness Score and the independent variables are Social Support, Healthy Life Expectancy, Freedom, Corruption, Generosity, and GDP. The resulting regression coefficients, which statistically measure the significance of a relationship, are as follows respectively: 3.031, 0.0334, 1.731 -0.487, 0.335, 0.245. The multiple regression model was run using the Python libraries Pandas, Sklearn, and Numpy. The model led to an R-squared value of 0.815 which proves a relatively high level of correlation for the relationship of the independent variables to the Happiness Score. We received a Root Mean Square Error (RMSE) of 0.465, which shows the model relatively predicts the data somewhat accurately.

4 Visualization

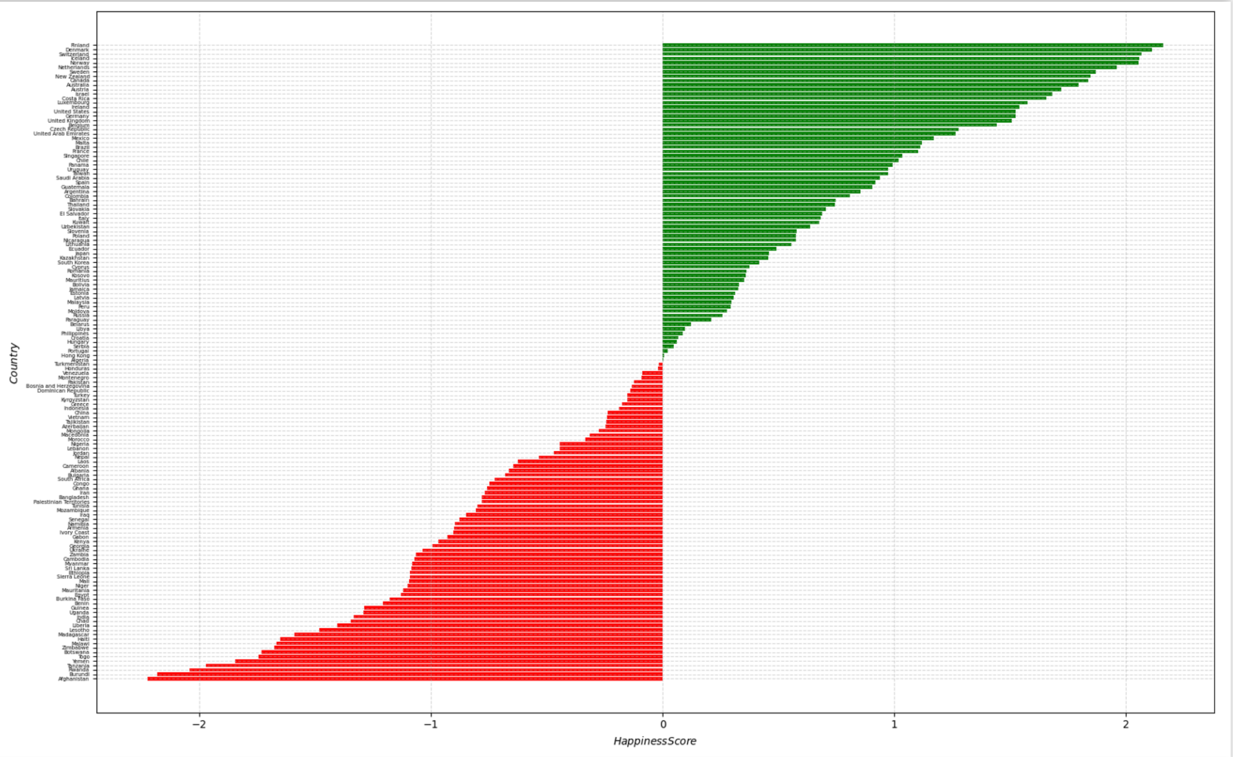
4.1 Visualization Goals:

For our visualizations sprint, we wanted to focus on illustrating how happiness levels have changed over time, happiness levels in the regions with the highest decrease in happiness over the seven years, the role of a country’s wealth in its happiness levels, and the relationships between variables and the shapes of their distributions.

4.2 Visualization Tools:

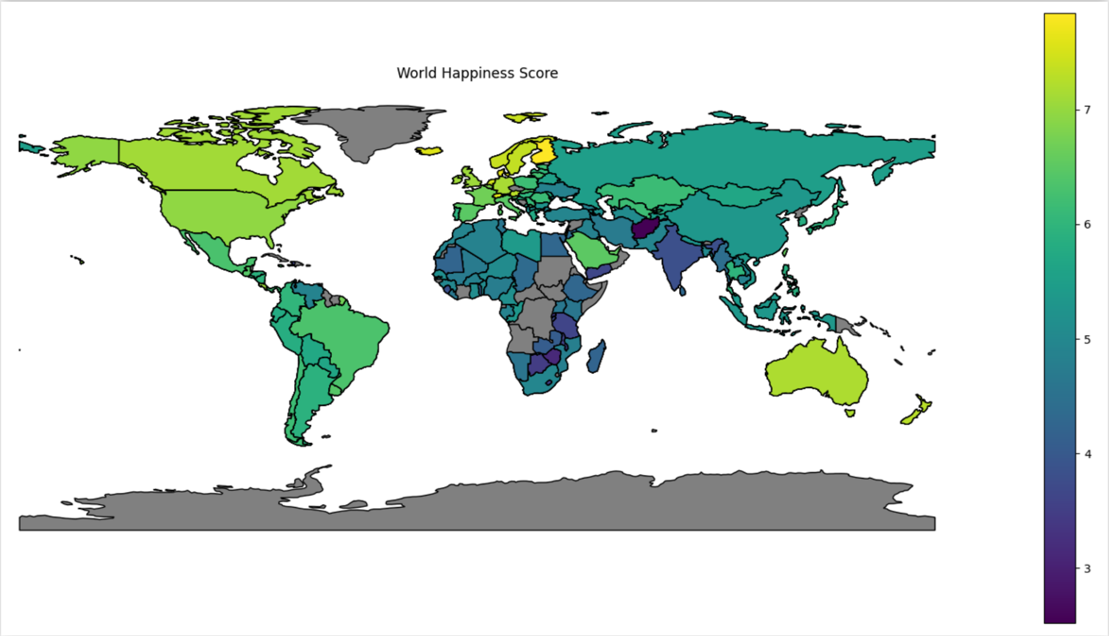
We used Python and several of its libraries to generate our visualizations. The Pandas library and Numpy were used to import and manipulate the datasets. Matplotlib, Seaborn, GDAL, and Geopandas were used to create graphs and visualizations. Scipy and Sklearn were also utilized to generate models and find the strengths of relationships between different variables.

4.3 Ranking countries based on happiness score:



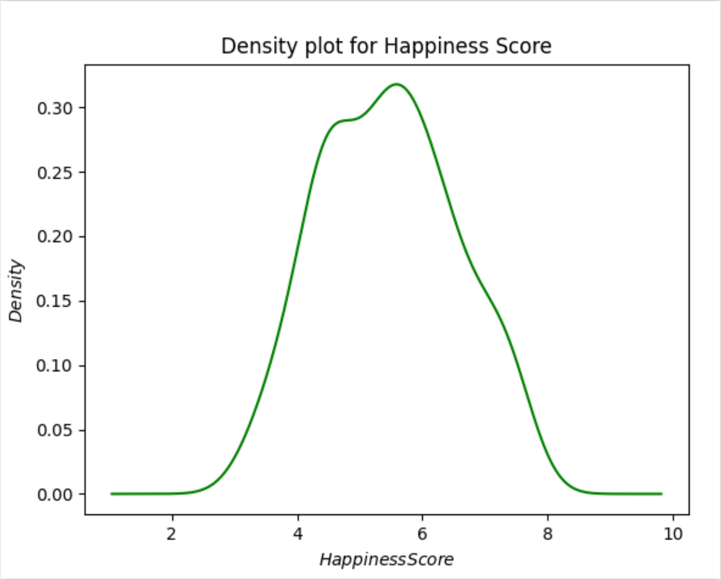
This simple visualization shows the happiness levels of each country in descending order, with the values normalized and the x-axis showing the standard deviation from the mean.

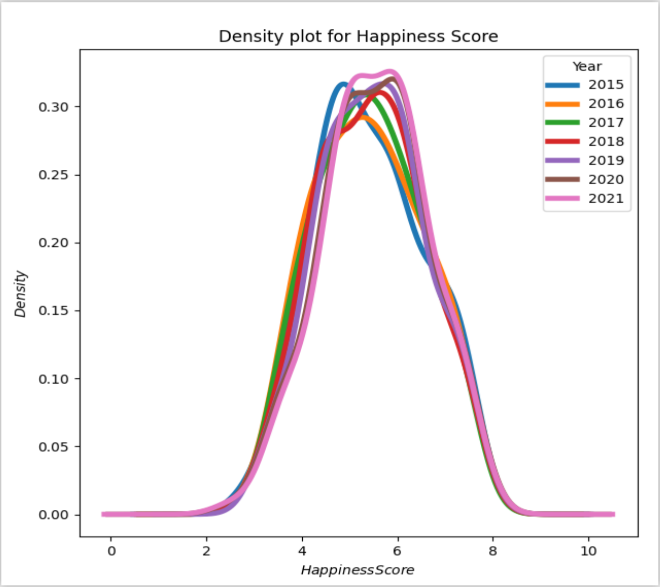
4.4 Happiness Score across the Globe:



This visualization shows a world map that assigns each country (for which there is available data) a color on the gradient that represents its happiness level. High happiness levels are more green/yellow, while low happiness levels are more blue/purple.

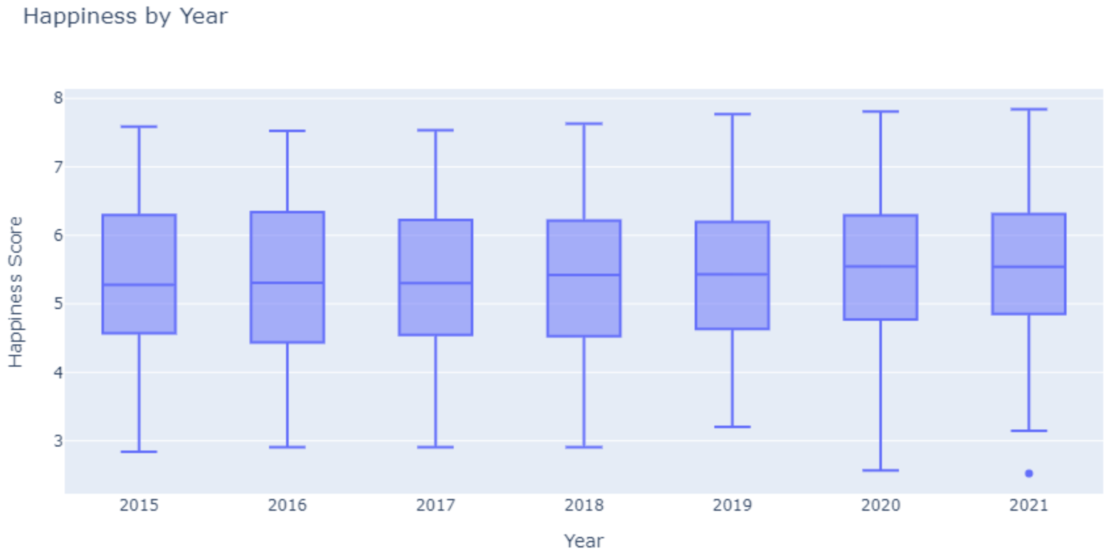
4.5 Density Analysis:





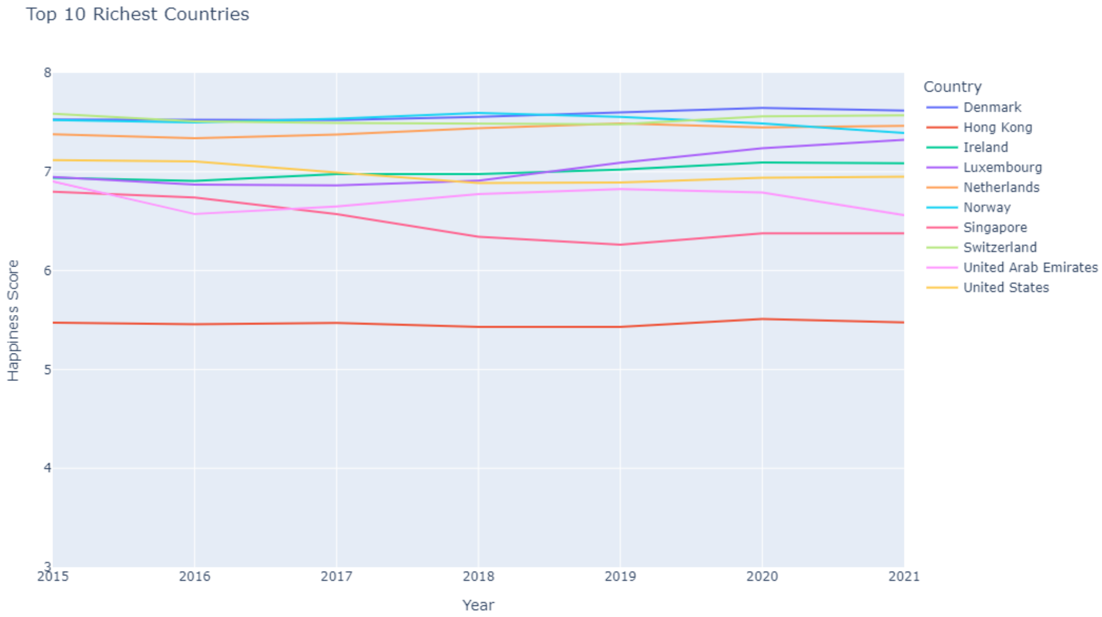
If we look at the density plots for the happiness score, we can see that it follows a relatively normal distribution with a slight skew to the right. The happiness distributions separated by year also seem to be mostly uniform, with the COVID years of 2020 having a more intense right-skew.

4.6 Happiness boxplots:



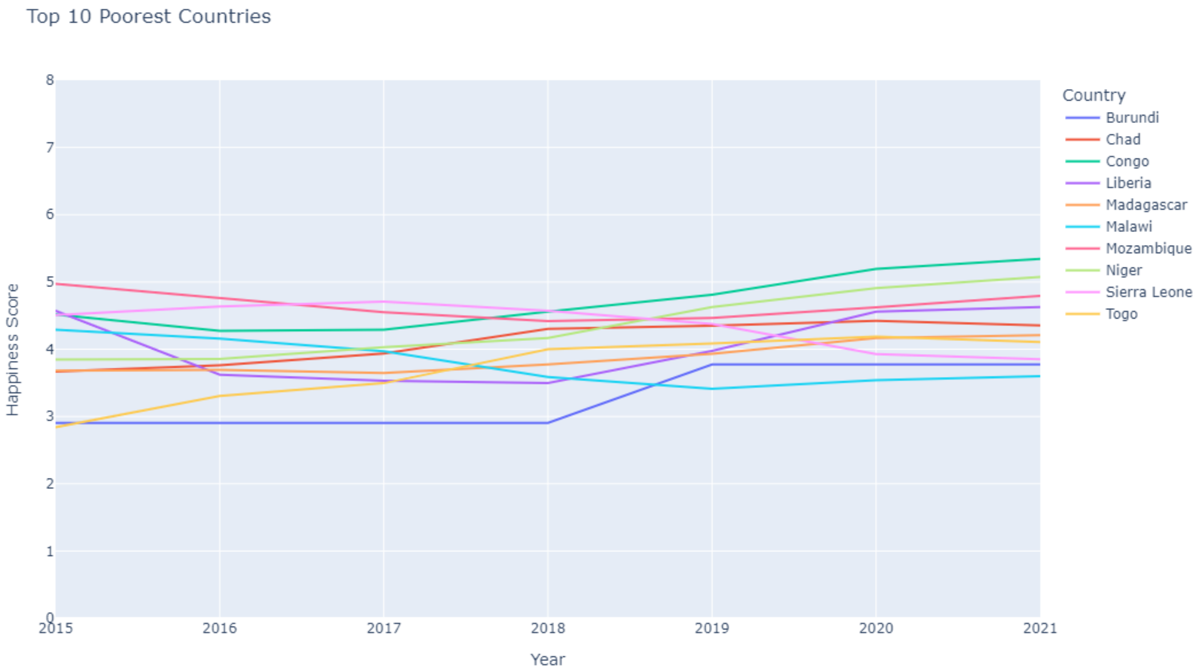
The most surprising finding from our analysis is the fact that happiness levels remained remarkably resilient during COVID, with the median happiness score rising steadily from 2015-2021. One interesting thing is that the distribution looks to be more right skewed during the COVID years in 2020 and 2021, which seems to indicate that even though the median is steadily rising, COVID still may have had an effect on the distribution of happiness scores.

4.7 Richest countries and their happiness levels:



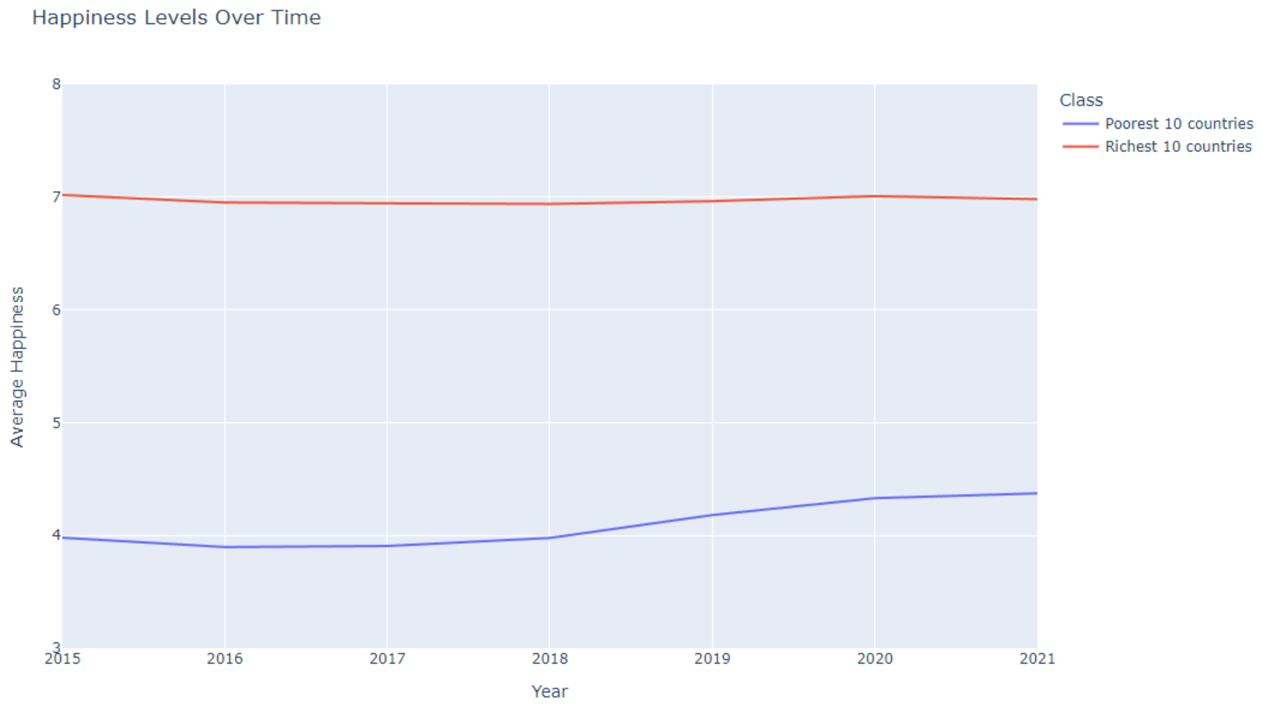
We also wanted to investigate how happiness levels have changed over time for the richer countries, so this multiple line graph shows the top 10 richest countries according to their average GDP per capita. As you can see, their happiness levels have remained relatively steady for all the consecutive years. And obviously Hong Kong stands out as the least happy of the top 10 rich countries. The top 10 richest countries by average GDP per capita are Denmark, Hong Kong, Ireland, Luxembourg, Netherlands, Norway, Singapore, Switzerland, United Arab Emirates, and the United States.

4.8 Poorest countries and their happiness levels:



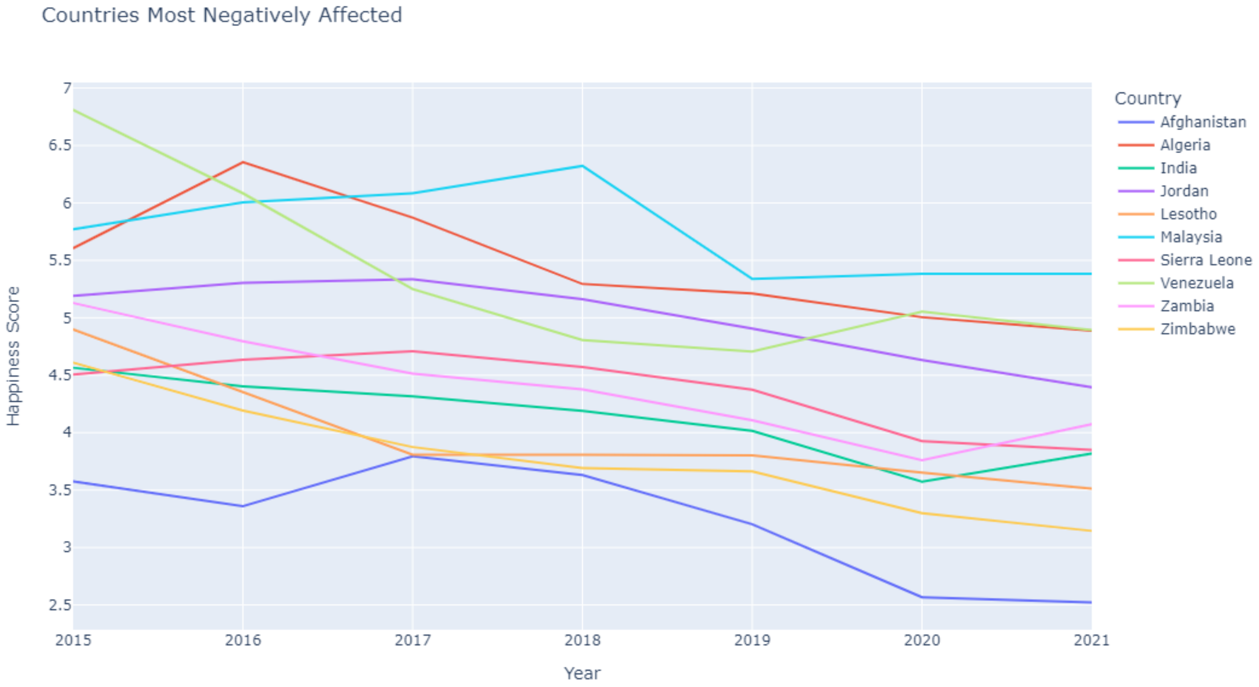
To compare to the richest countries, in this graph we identified the top 10 poorest countries by the same method. As you can see, the poorest countries have a lot more variation in their happiness levels, with the majority of them having increasing happiness levels even during COVID. The top 10 poorest countries by average GDP per capita are Burundi, Chad, Congo, Liberia, Madagascar, Malawi, Mozambique, Niger, Sierra Leone, and Togo.

4.9 Average happiness levels of the poorest and richest:



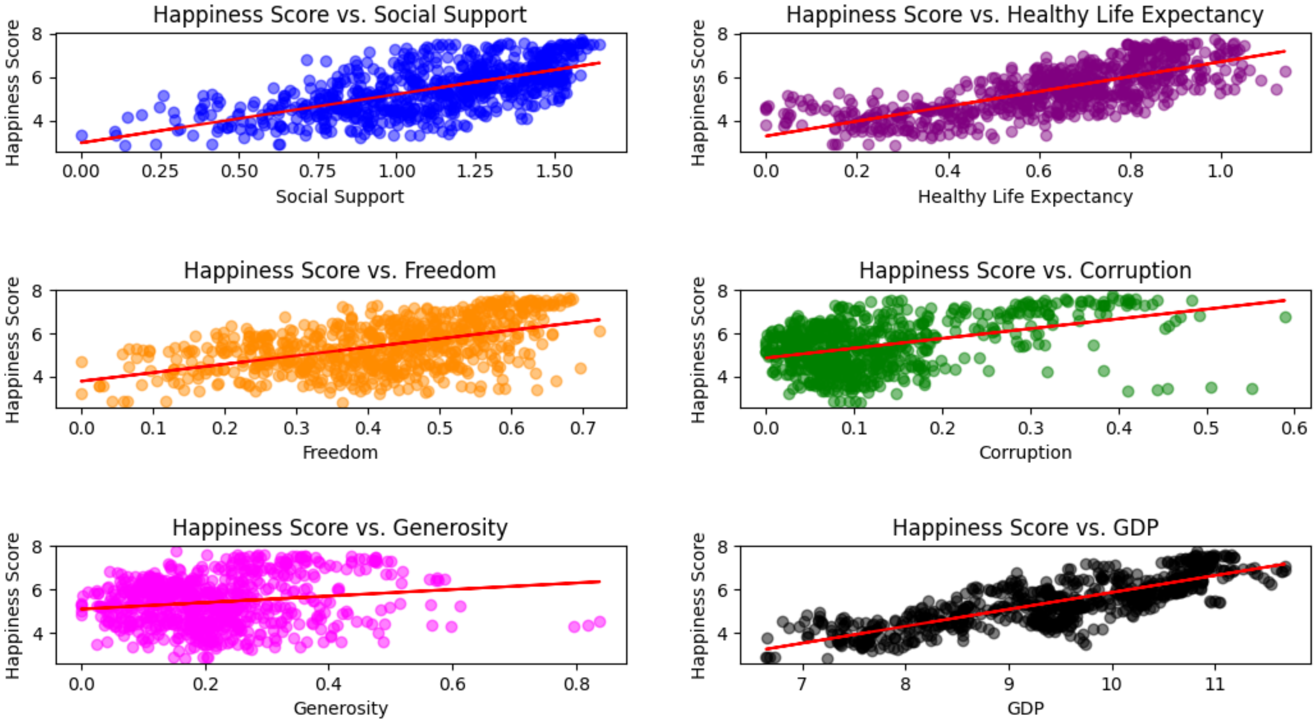
To better illustrate how the happiness levels have changed over time for the 10 poorest and richest countries, this line graph shows the average happiness level for each group. As you can see, the 10 poorest countries have had relatively stable happiness growth throughout the seven years that remained resilient throughout COVID. The average happiness levels for the top 10 richest countries remained remarkably stable.

4.10 Countries with the highest decreases in happiness:



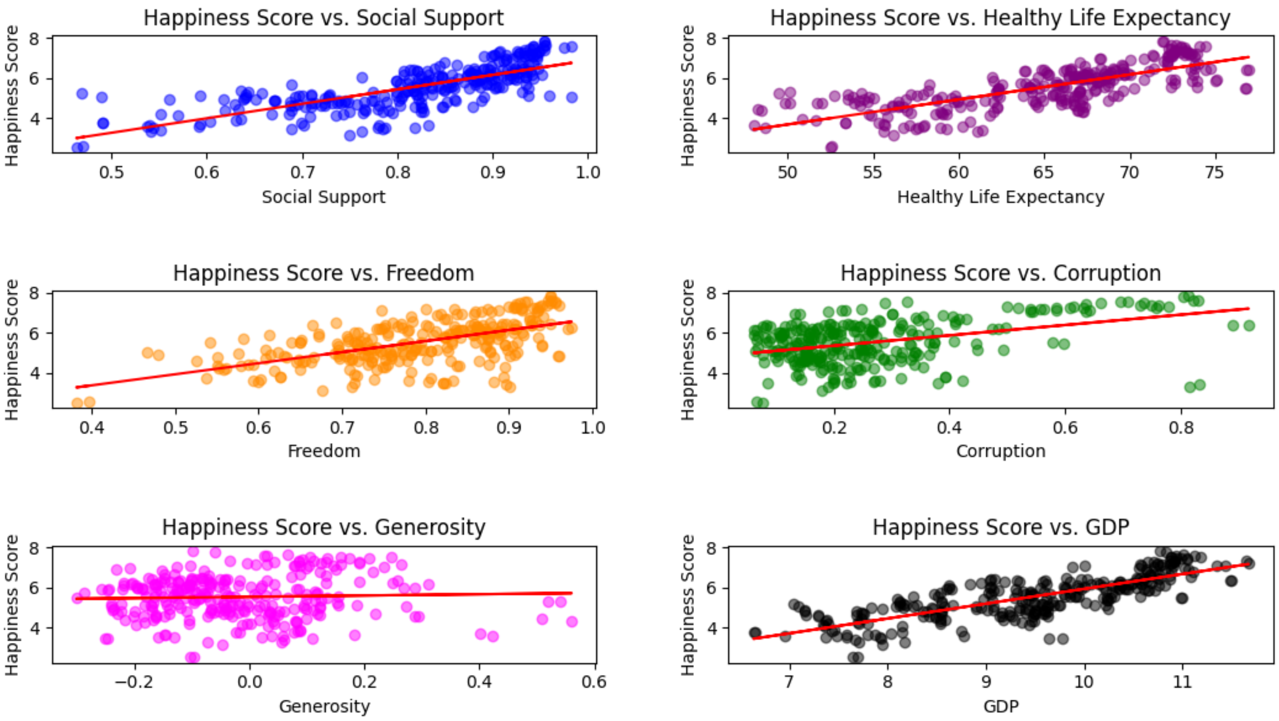
This final line graph shows the happiness levels over time for the 10 countries that had the highest decrease in happiness levels before and after COVID started (identified by comparing the average happiness levels for 2015-2019 with those of 2020-2021). You can see the general downward trends for each country. Afghanistan has had the lowest happiness levels for each of the measured years. It would be interesting to look more into Venezuela, since its average happiness levels started off much higher than the other countries in this list. Only the countries of Venezuela and Malaysia have increased happiness levels from 2019 to 2020 which would also be interesting to look into.

4.11 Linear relationships - Effect on happiness score 2015-2019:



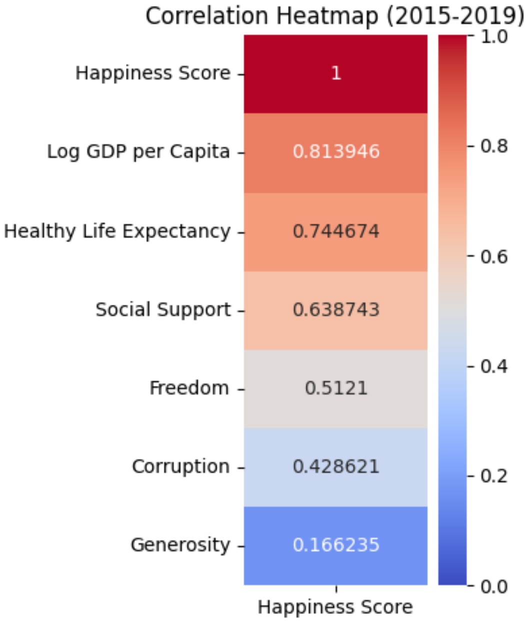
This visualization shows the scatterplots and regression lines for each variable’s relationship with happiness score for the years of 2015-2019. As you can see, the strongest relationships are with GDP per Capita and Healthy Life Expectancy.

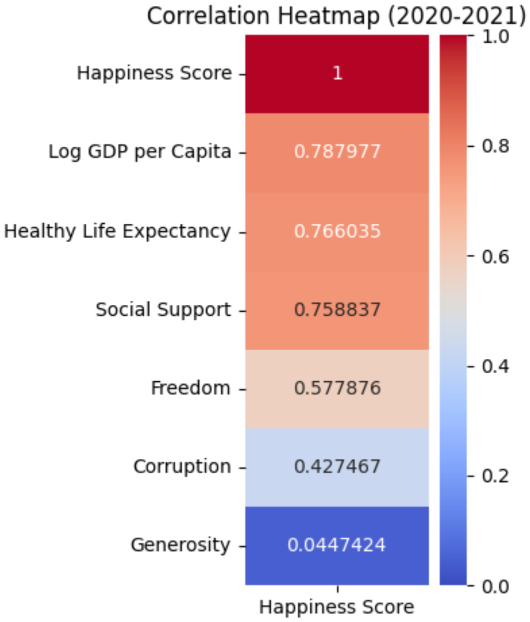
4.12 Linear relationships - Effect on happiness score 2020-2021:



This is the same visualization, but for the years 2020-2021. As you can see, the relationships all remain largely the same, but with a weaker relationship between happiness and generosity. It's possible that the negative economic effects of COVID made people less likely to be financially generous, which would have this effect on the scatterplot/regression line.

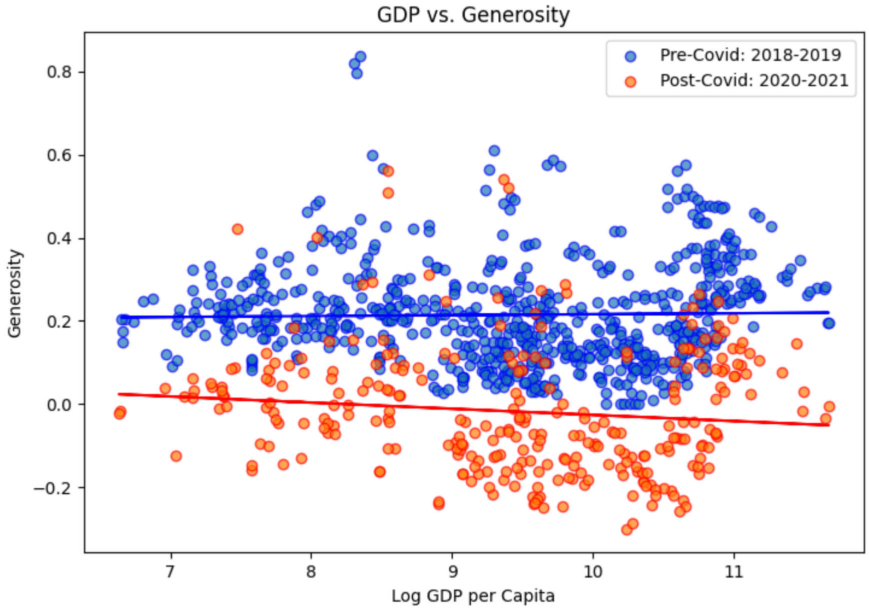
4.13 Correlation coefficients:

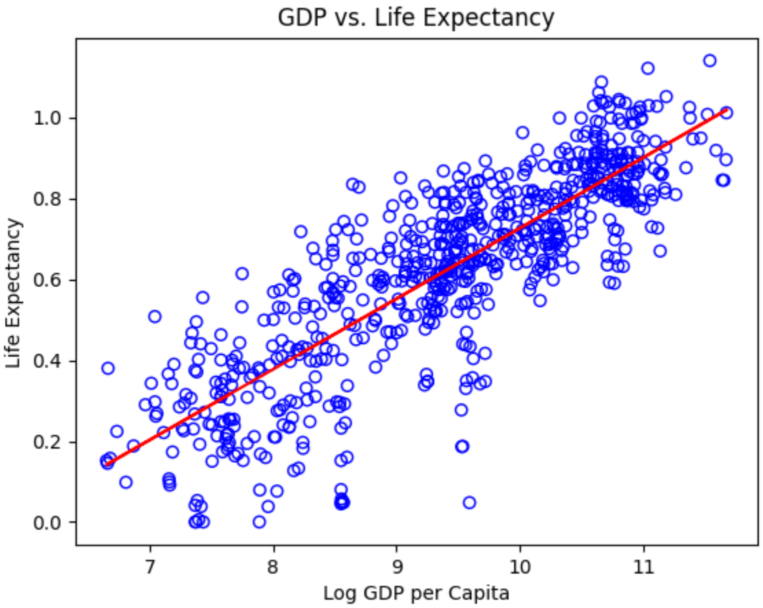




These are correlation heatmaps that show each variable's correlation coefficient with happiness score in descending order, separated by pre and post-COVID. The order has remained the same for both periods of time, but it's apparent that Social Support seems to have had a much more important role in happiness levels during the COVID years as indicated by its higher correlation coefficient.

4.14 Log GDP per Capita findings:

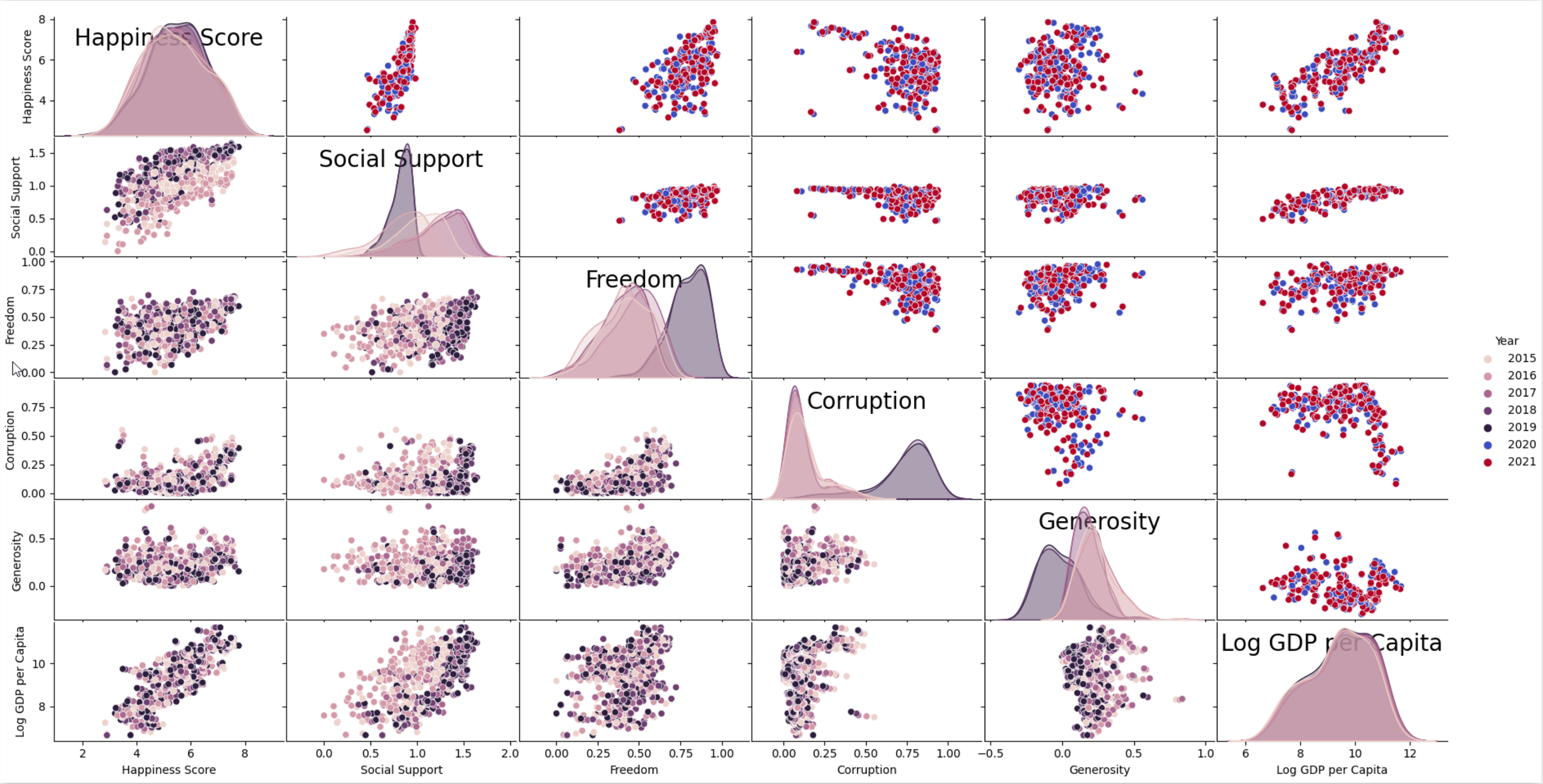




The first Log GDP per Capita visual, “GDP vs. Generosity”, shows the relationship between GDP and the Generosity score of a country. From the graph it can be seen that the two variables have a negative relationship where as GDP increases, Generosity decreases slightly. The blue data points are data from the years 2018-2019, or before COVID; and the red data points are the data from the first two years after the start of COVID, 2020-2021. This feature shows that after the pandemic, Generosity visibly decreased for countries. This can be attributed to many reasons, including that during uncertain times, people may be less likely to donate their money.

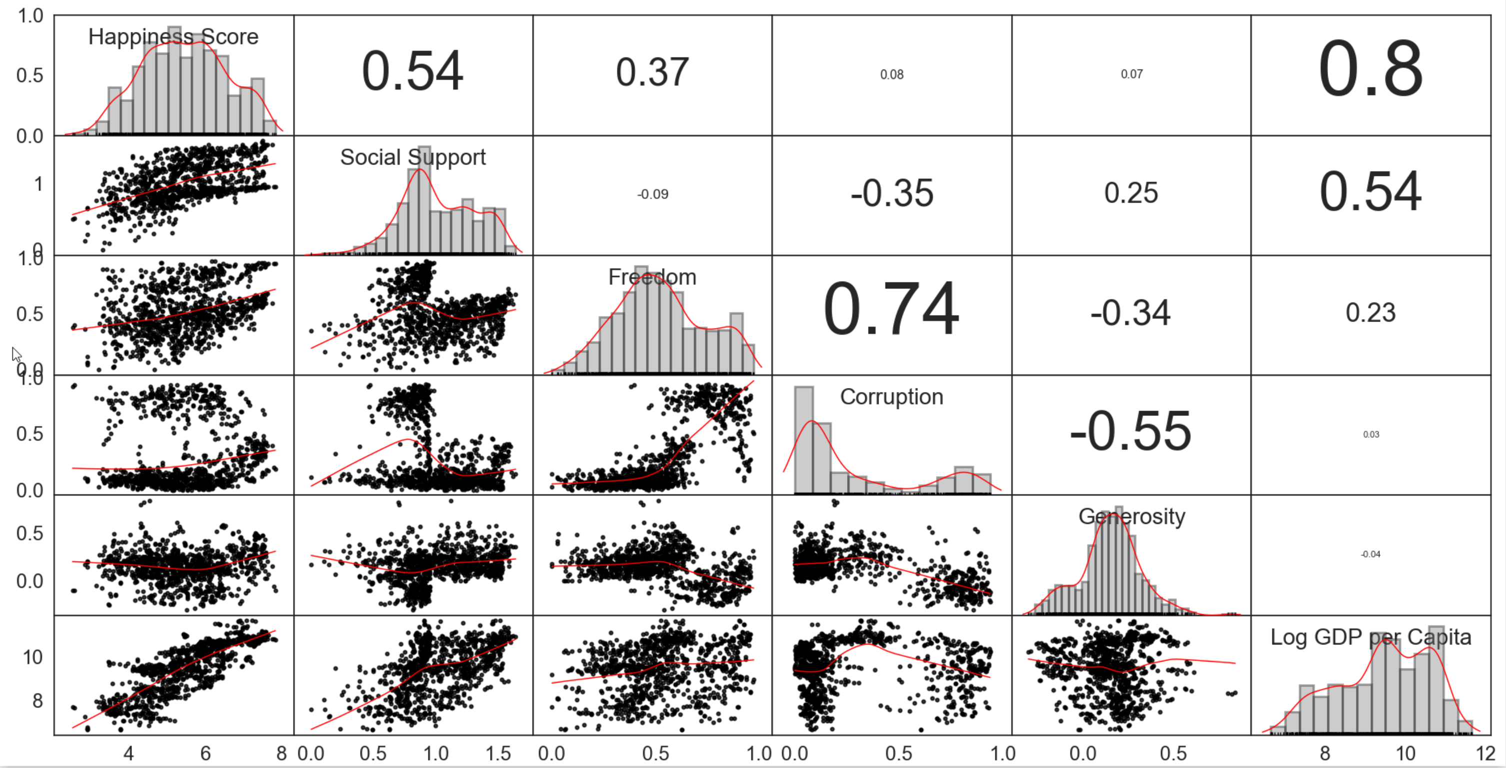
The second Log GDP per Capita visual, “GDP vs. Life Expectancy”, shows the relationship of GDP and Life Expectancy. The graph shows a strong linear relationship between GDP and Life Expectancy. This could be expected since a country with more money can have better access to good healthcare for its citizens.

4.15 Scatterplot matrix:



The visual above is a scatterplot matrix for the variables in the dataset. The pre-covid data is displayed in the lower left triangle and post-covid data is displayed in the upper right triangle of the matrix. The diagonal displays the density distribution for each variable. This visual shows an increase for Social Support and Perception of Freedom and a decrease for Corruption and Generosity during/after Covid.

4.16 Scatterplot matrix - Variable Distribution/Pairwise Relationships:



The visual above is a scatterplot matrix for the variables in the dataset. This matrix shows the variable distribution and pairwise relationships of the variables. The resulting coefficients can be seen in the upper right triangle. The diagonal displays the Histograms for individual variables.

5 Findings

5.1 Happiness levels over time and distribution

The most surprising finding from our analysis is the remarkable resilience of happiness scores during the COVID years of 2020 and 2021. The global mean and median for happiness score steadily increased from 2015-2021. One interesting thing to note, however, is that the distribution of happiness scores were more right-skewed during COVID, which implies that some countries may have been affected disproportionately.

When separating countries by wealth, we found that the poorer countries had stable happiness growth throughout all 7 years (despite COVID), whereas the richer countries had happiness levels that remained surprisingly stagnant. It’s possible that as overall economic conditions were still improving for the poorer countries overall throughout the years, there would be a more dramatic increase in quality of life compared to an already rich country raising its GDP by the same amount.

The top 10 countries that had the most dramatic decline in happiness scores for 2020-2021 were Afghanistan, Zimbabwe, Algeria, Sierra Leone, Zambia, Jordan, India, Venezuela, Lesotho, and Malaysia. Afghanistan has remained the country with the lowest happiness score for every year we studied, and its average happiness score for 2020-2021 was 27.5 percent lower than for 2015-2019.

5.2 Happiness variables

The factors that have the strongest relationships with happiness score are GDP per Capita, life expectancy, and social support. Corruption and generosity have the weakest effects on happiness in this dataset. During the COVID years 2020-2021, social support rose and its relationship with global happiness levels became stronger, which implies that during the pandemic people relied more on social support for their happiness.

For Log GDP per Capita, its strongest relationship was with life expectancy; the richer a country is, the higher its life expectancy. The ratio of generosity to GDP was also much lower during COVID, likely because people are less likely to donate to charity during an economic downturn.

6 Summary

Our findings showed that the happiness scores during the years after the start of the pandemic (2020-2021) had strong resilience. However, the distribution of happiness scores were more right-skewed during COVID. When separating countries by wealth, we found that the poorer countries had stable happiness growth throughout all 7 years, whereas the richer countries remained stagnant.

GDP was found to be the most important variable in determining happiness levels. Generosity and Corruption were determined to have the least effect on happiness. During COVID, social support increased and had a stronger relationship in determining happiness whereas generosity decreased. Further work to look at is other factors that could affect happiness, including education level and exercise. We could also look further back and see the trends across the past couple of decades to see which direction the world is trending.

7 Future Work

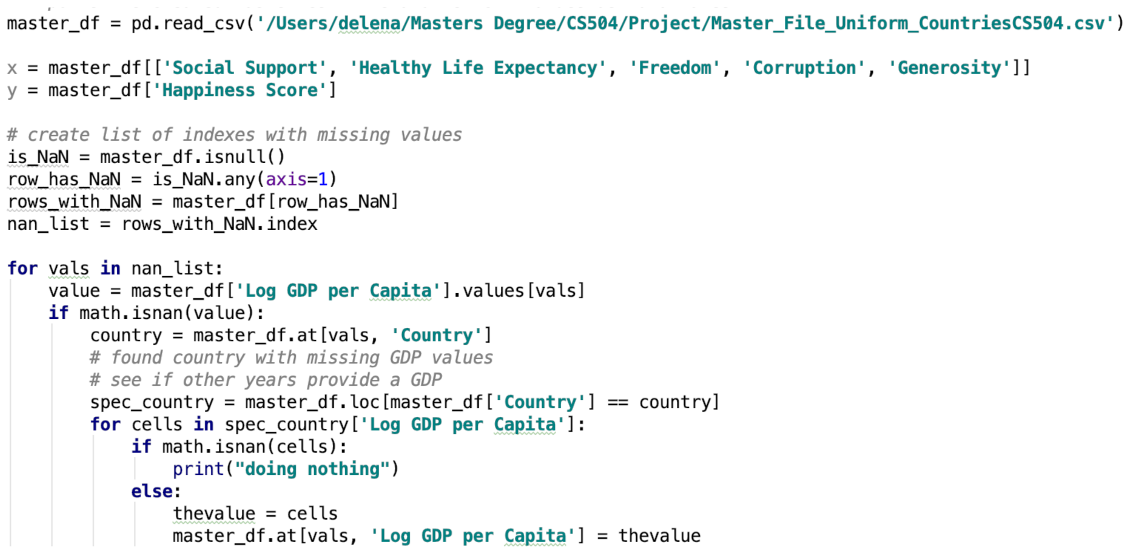
If we were to continue work on this project, we would like to look at other factors and how they affect happiness. Perhaps education level, exercise or a work-life balance could affect happiness levels and we could dig further into that. We could also look further back and see the trends across the past couple of decades to see which direction the world is trending.

8 Appendix

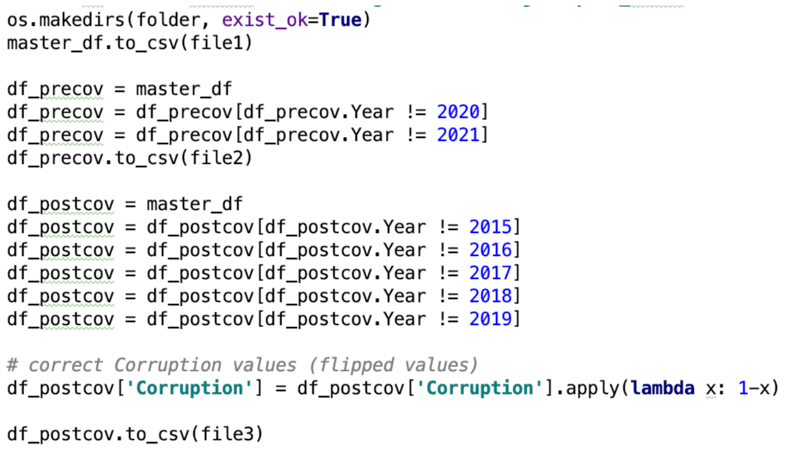
Appendix A

Code references – any code used for the analysis

Code used to remove “nan” data

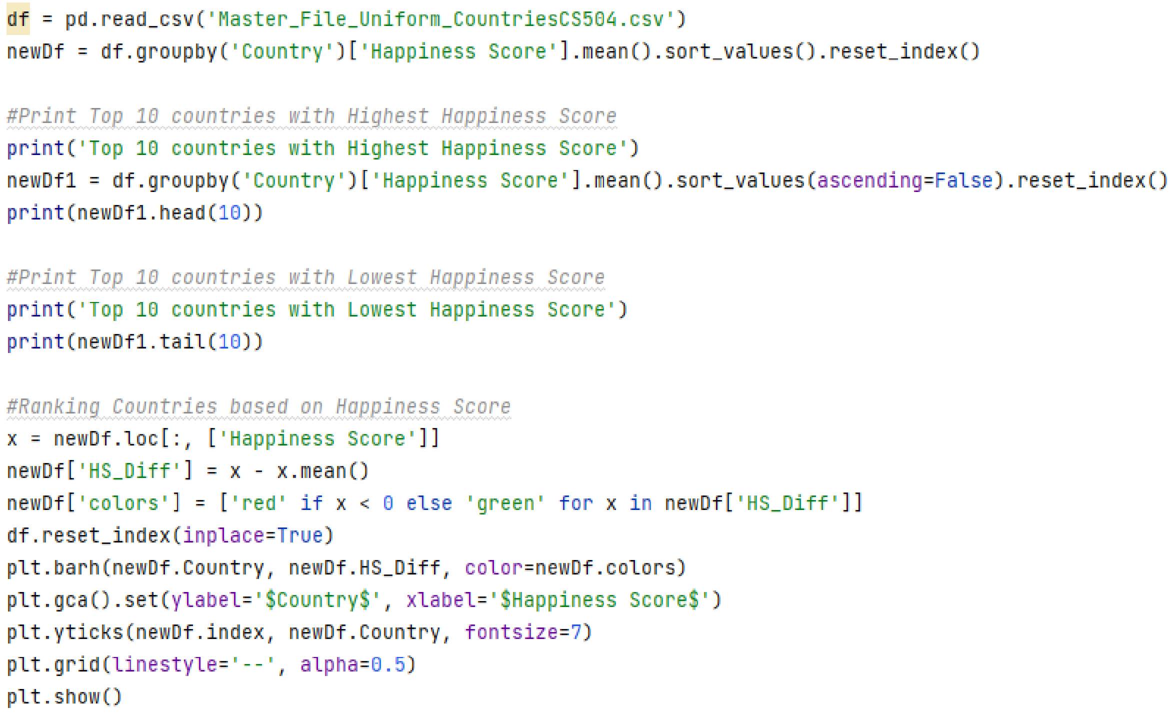


```
master_df = pd.read_csv('/Users/delena/Masters Degree/CS504/Project/Master_File_Uniform_CountriesCS504.csv')

x = master_df[['Social Support', 'Healthy Life Expectancy', 'Freedom', 'Corruption', 'Generosity']]
y = master_df['Happiness Score']

# create list of indexes with missing values
is_NaN = master_df.isnull()
row_has_NaN = is_NaN.any(axis=1)
rows_with_NaN = master_df[row_has_NaN]
nan_list = rows_with_NaN.index

for vals in nan_list:
    value = master_df['Log GDP per Capita'].values[vals]
    if math.isnan(value):
        country = master_df.at[vals, 'Country']
        # found country with missing GDP values
        # see if other years provide a GDP
        spec_country = master_df.loc[master_df['Country'] == country]
        for cells in spec_country['Log GDP per Capita']:
            if math.isnan(cells):
                print("doing nothing")
            else:
                thevalue = cells
                master_df.at[vals, 'Log GDP per Capita'] = thevalue
```



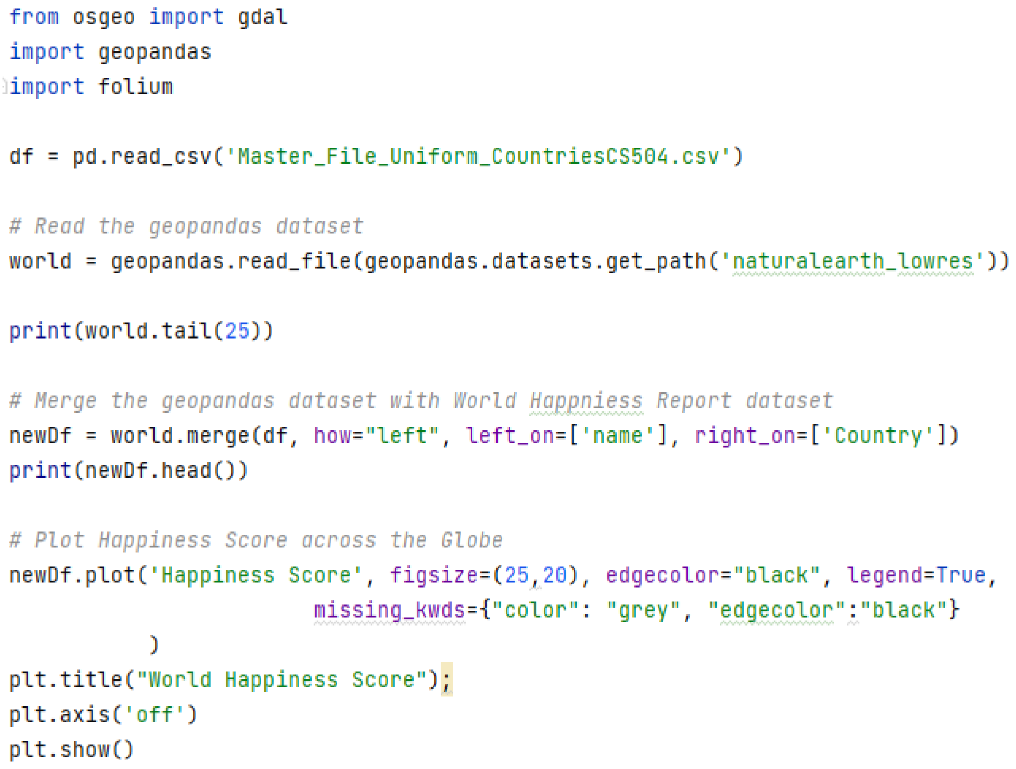
```
os.makedirs(folder, exist_ok=True)
master_df.to_csv(file1)

df_precov = master_df
df_precov = df_precov[df_precov.Year != 2020]
df_precov = df_precov[df_precov.Year != 2021]
df_precov.to_csv(file2)

df_postcov = master_df
df_postcov = df_postcov[df_postcov.Year != 2015]
df_postcov = df_postcov[df_postcov.Year != 2016]
df_postcov = df_postcov[df_postcov.Year != 2017]
df_postcov = df_postcov[df_postcov.Year != 2018]
df_postcov = df_postcov[df_postcov.Year != 2019]

# correct Corruption values (flipped values)
df_postcov['Corruption'] = df_postcov['Corruption'].apply(lambda x: 1-x)

df_postcov.to_csv(file3)
```

Code for Ranking Countries based on Happiness Score And Ranking Top 10 and Bottom 10 Countries



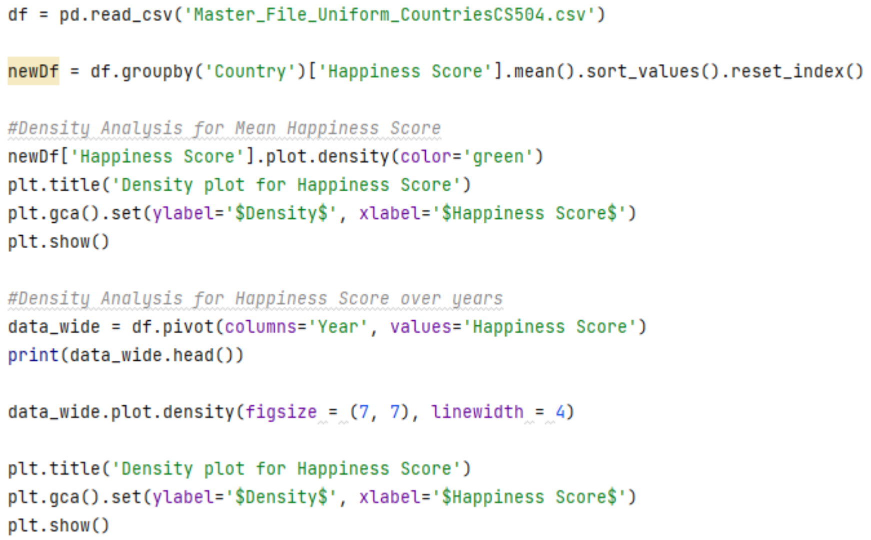
```
df = pd.read_csv('Master_File_Uniform_CountriesCS504.csv')
newDf = df.groupby('Country')['Happiness Score'].mean().sort_values().reset_index()

#Print Top 10 countries with Highest Happiness Score
print('Top 10 countries with Highest Happiness Score')
newDf1 = df.groupby('Country')['Happiness Score'].mean().sort_values(ascending=False).reset_index()
print(newDf1.head(10))

#Print Top 10 countries with Lowest Happiness Score
print('Top 10 countries with Lowest Happiness Score')
print(newDf1.tail(10))

#Ranking Countries based on Happiness Score
x = newDf.loc[:, ['Happiness Score']]
newDf['HS_Diff'] = x - x.mean()
newDf['colors'] = ['red' if x < 0 else 'green' for x in newDf['HS_Diff']]
df.reset_index(inplace=True)
plt.barh(newDf.Country, newDf.HS_Diff, color=newDf.colors)
plt.gca().set(ylabel='$Country$', xlabel='$Happiness Score$')
plt.yticks(newDf.index, newDf.Country, fontsize=7)
plt.grid(linestyle='--', alpha=0.5)
plt.show()
```

Code for Happiness Score across the Globe



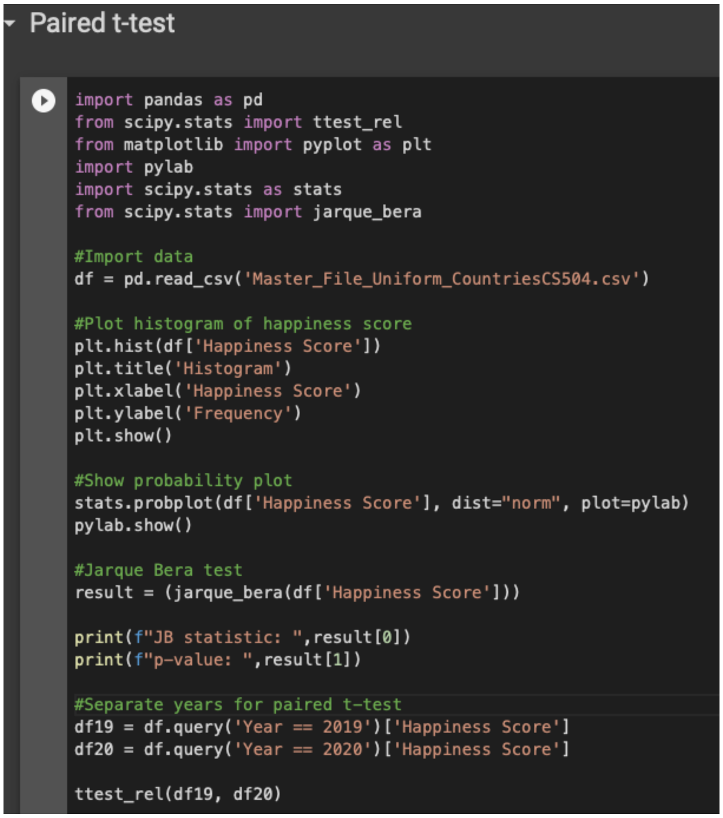
```
from osgeo import gdal
import geopandas
import folium

df = pd.read_csv('Master_File_Uniform_CountriesCS504.csv')

# Read the geopandas dataset
world = geopandas.read_file(geopandas.datasets.get_path('naturalearth_lowres'))

print(world.tail(25))

# Merge the geopandas dataset with World Happniess Report dataset
newDf = world.merge(df, how="left", left_on=['name'], right_on=['Country'])
print(newDf.head())

# Plot Happiness Score across the Globe
newDf.plot('Happiness Score', figsize=(25,20), edgecolor="black", legend=True,
                        missing_kwds={"color": "grey", "edgecolor":"black"}
           )
plt.title("World Happiness Score");
plt.axis('off')
plt.show()
```

Code for Density Analysis



```
df = pd.read_csv('Master_File_Uniform_CountriesCS504.csv')

newDf = df.groupby('Country')['Happiness Score'].mean().sort_values().reset_index()

#Density Analysis for Mean Happiness Score
newDf['Happiness Score'].plot.density(color='green')
plt.title('Density plot for Happiness Score')
plt.gca().set(ylabel='$Density$', xlabel='$Happiness Score$')
plt.show()

#Density Analysis for Happiness Score over years
data_wide = df.pivot(columns='Year', values='Happiness Score')
print(data_wide.head())

data_wide.plot.density(figsize = (7, 7), linewidth = 4)

plt.title('Density plot for Happiness Score')
plt.gca().set(ylabel='$Density$', xlabel='$Happiness Score$')
plt.show()
```

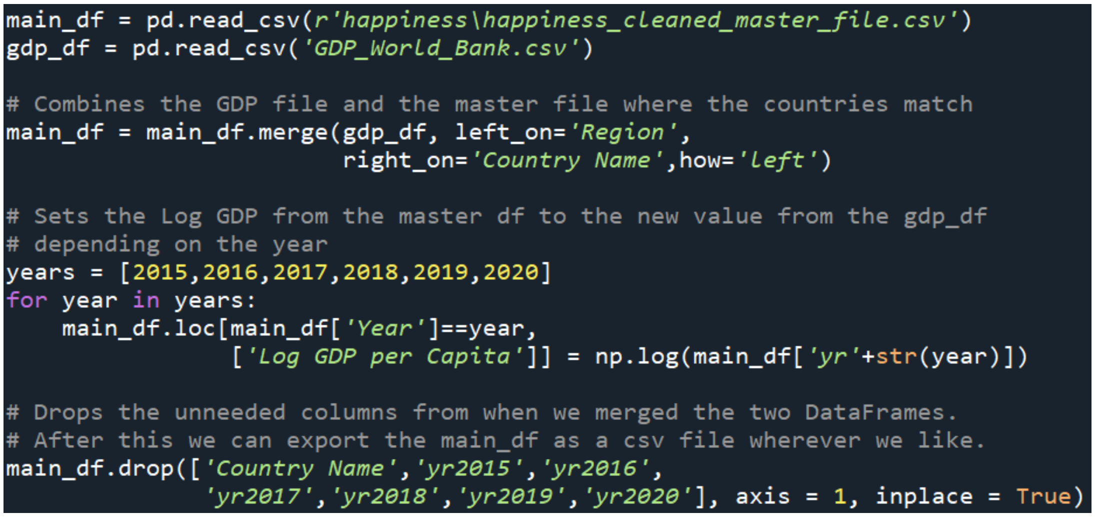
Code for normal distribution and paired t-test



Paired t-test

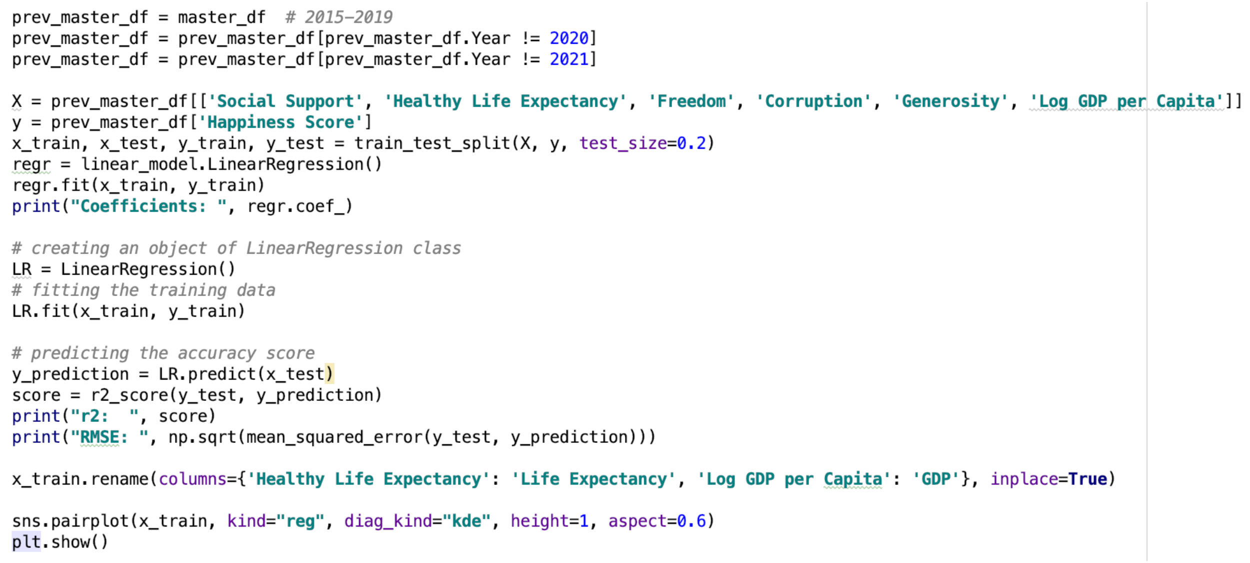
```
import pandas as pd
from scipy.stats import ttest_rel
from matplotlib import pyplot as plt
import pylab
import scipy.stats as stats
from scipy.stats import jarque_bera

#Import data
df = pd.read_csv('Master_File_Uniform_CountriesCS504.csv')

#Plot histogram of happiness score
plt.hist(df['Happiness Score'])
plt.title('Histogram')
plt.xlabel('Happiness Score')
plt.ylabel('Frequency')
plt.show()

#Show probability plot
stats.probplot(df['Happiness Score'], dist="norm", plot=pylab)
pylab.show()

#Jarque Bera test
result = (jarque_bera(df['Happiness Score']))

print(f"JB statistic: ",result[0])
print(f"p-value: ",result[1])

#Separate years for paired t-test
df19 = df.query('Year == 2019')['Happiness Score']
df20 = df.query('Year == 2020')['Happiness Score']

ttest_rel(df19, df20)
```

Code for replacing missing Log GDP per Capita



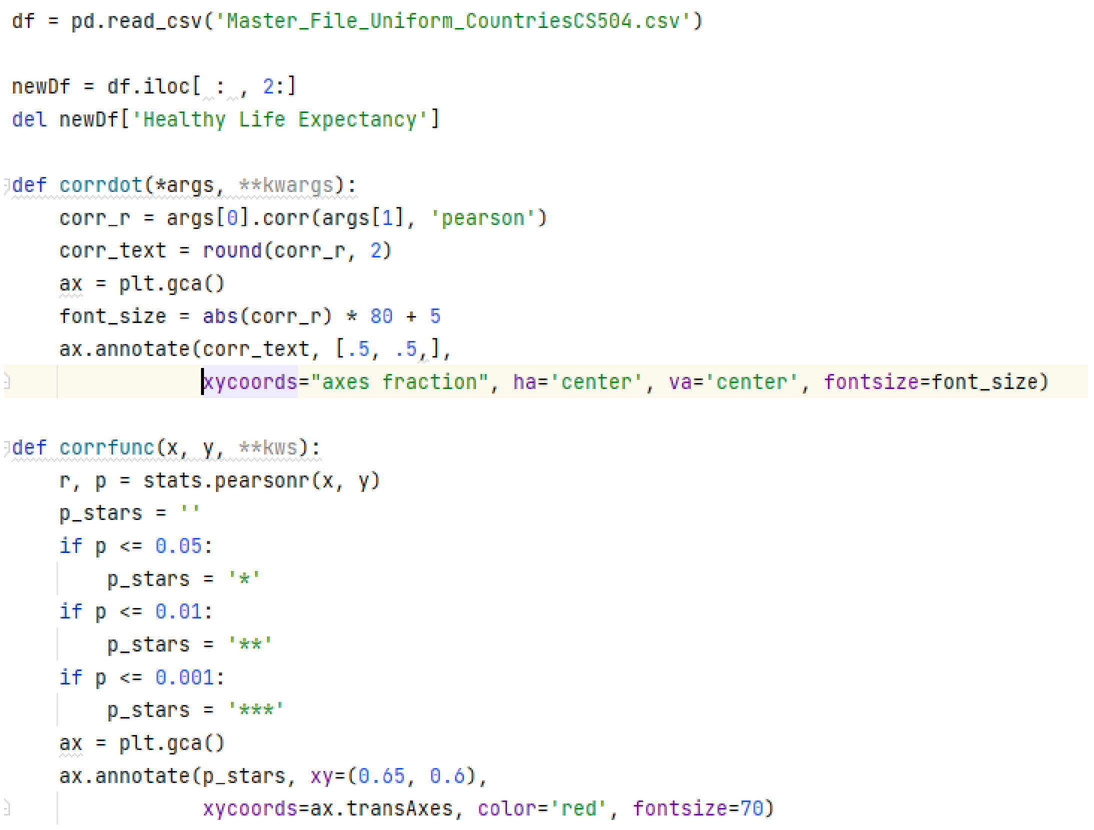
```
main_df = pd.read_csv(r'happiness\happiness_cleaned_master_file.csv')
gdp_df = pd.read_csv('GDP_World_Bank.csv')

# 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)
```

Multiple Regression model code

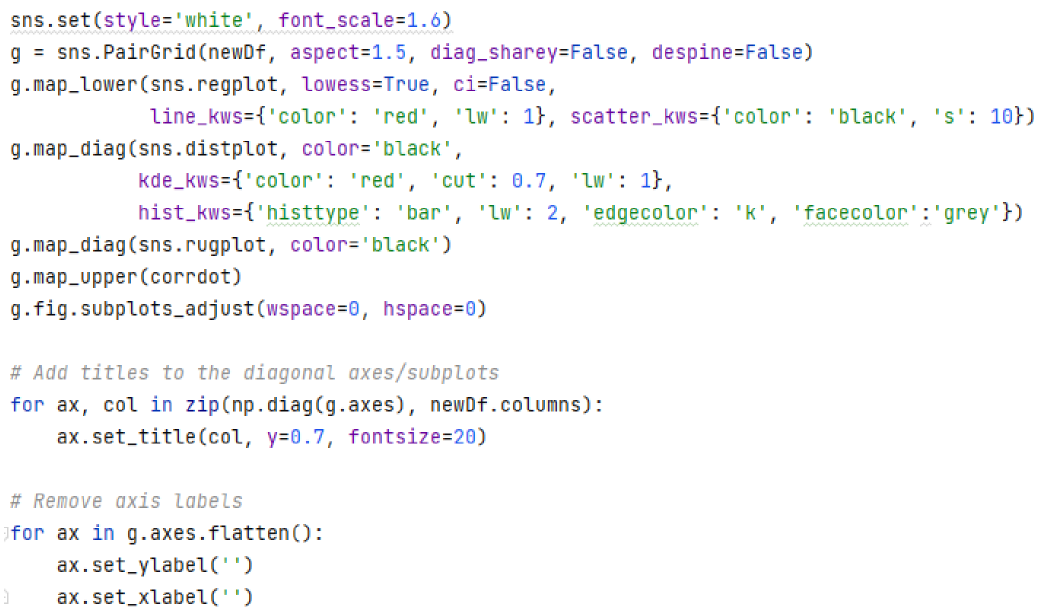


```
prev_master_df = master_df  # 2015-2019
prev_master_df = prev_master_df[prev_master_df.Year != 2020]
prev_master_df = prev_master_df[prev_master_df.Year != 2021]

X = prev_master_df[['Social Support', 'Healthy Life Expectancy', 'Freedom', 'Corruption', 'Generosity', 'Log GDP per Capita']]
y = prev_master_df['Happiness Score']
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
regr = linear_model.LinearRegression()
regr.fit(x_train, y_train)
print("Coefficients: ", regr.coef_)

# creating an object of LinearRegression class
LR = LinearRegression()
# fitting the training data
LR.fit(x_train, y_train)

# predicting the accuracy score
y_prediction = LR.predict(x_test)
score = r2_score(y_test, y_prediction)
print("r2:  ", score)
print("RMSE: ", np.sqrt(mean_squared_error(y_test, y_prediction)))

x_train.rename(columns={'Healthy Life Expectancy': 'Life Expectancy', 'Log GDP per Capita': 'GDP'}, inplace=True)

sns.pairplot(x_train, kind="reg", diag_kind="kde", height=1, aspect=0.6)
plt.show()
```

Code for ScatterPlot Matrix - Variable Distribution / Pairwise Relationships

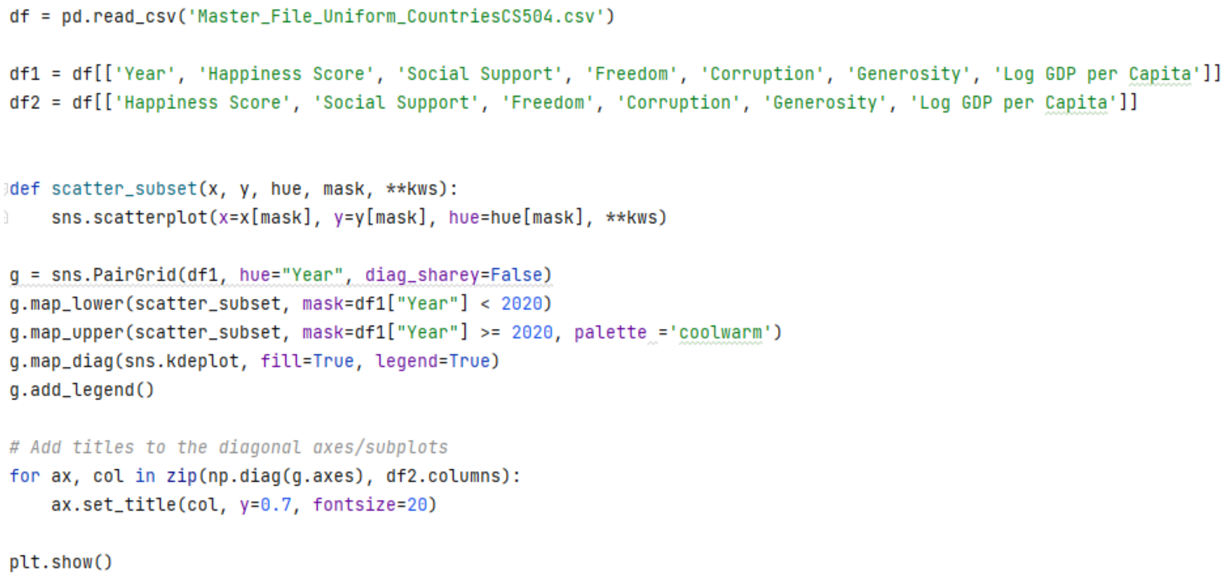


```
df = pd.read_csv('Master_File_Uniform_CountriesCS504.csv')

newDf = df.iloc[ : , 2:]
del newDf['Healthy Life Expectancy']

def corrdot(*args, **kwargs):
    corr_r = args[0].corr(args[1], 'pearson')
    corr_text = round(corr_r, 2)
    ax = plt.gca()
    font_size = abs(corr_r) * 80 + 5
    ax.annotate(corr_text, [.5, .5,],
                xycoords="axes fraction", ha='center', va='center', fontsize=font_size)

def corrfunc(x, y, **kws):
    r, p = stats.pearsonr(x, y)
    p_stars = ''
    if p <= 0.05:
        p_stars = '*'
    if p <= 0.01:
        p_stars = '**'
    if p <= 0.001:
        p_stars = '***'
    ax = plt.gca()
    ax.annotate(p_stars, xy=(0.65, 0.6),
                xycoords=ax.transAxes, color='red', fontsize=70)
```



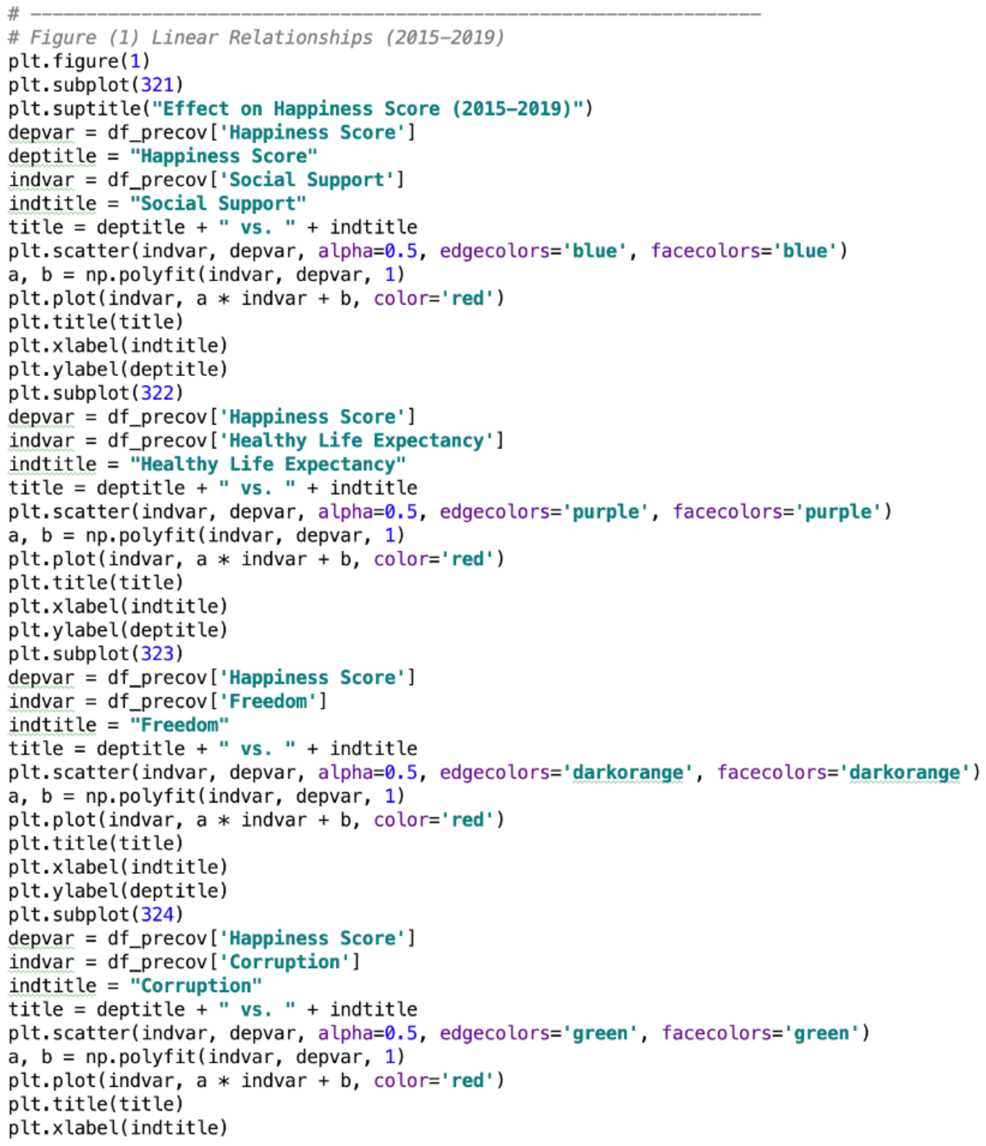
```
sns.set(style='white', font_scale=1.6)
g = sns.PairGrid(newDf, aspect=1.5, diag_sharey=False, despine=False)
g.map_lower(sns.regplot, lowess=True, ci=False,
            line_kws={'color': 'red', 'lw': 1}, scatter_kws={'color': 'black', 's': 10})
g.map_diag(sns.distplot, color='black',
           kde_kws={'color': 'red', 'cut': 0.7, 'lw': 1},
           hist_kws={'histtype': 'bar', 'lw': 2, 'edgecolor': 'k', 'facecolor':'grey'})
g.map_diag(sns.rugplot, color='black')
g.map_upper(corrdot)
g.fig.subplots_adjust(wspace=0, hspace=0)

# Add titles to the diagonal axes/subplots
for ax, col in zip(np.diag(g.axes), newDf.columns):
    ax.set_title(col, y=0.7, fontsize=20)

# Remove axis labels
for ax in g.axes.flatten():
    ax.set_ylabel('')
    ax.set_xlabel('')
```

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

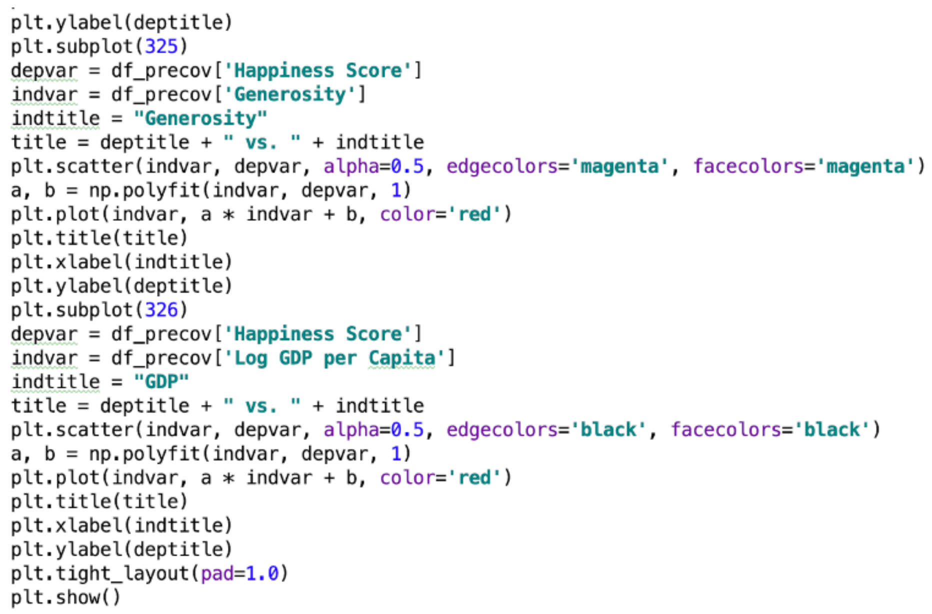


```
df = pd.read_csv('Master_File_Uniform_CountriesCS504.csv')

df1 = df[['Year', 'Happiness Score', 'Social Support', 'Freedom', 'Corruption', 'Generosity', 'Log GDP per Capita']]
df2 = df[['Happiness Score', 'Social Support', 'Freedom', 'Corruption', 'Generosity', 'Log GDP per Capita']]


def scatter_subset(x, y, hue, mask, **kws):
    sns.scatterplot(x=x[mask], y=y[mask], hue=hue[mask], **kws)

g = sns.PairGrid(df1, hue="Year", diag_sharey=False)
g.map_lower(scatter_subset, mask=df1["Year"] < 2020)
g.map_upper(scatter_subset, mask=df1["Year"] >= 2020, palette ='coolwarm')
g.map_diag(sns.kdeplot, fill=True, legend=True)
g.add_legend()

# Add titles to the diagonal axes/subplots
for ax, col in zip(np.diag(g.axes), df2.columns):
    ax.set_title(col, y=0.7, fontsize=20)

plt.show()
```

Code for Linear Relationships Plot for data from 2015 - 2019 (The same code was used for the data from 2020 - 2021)

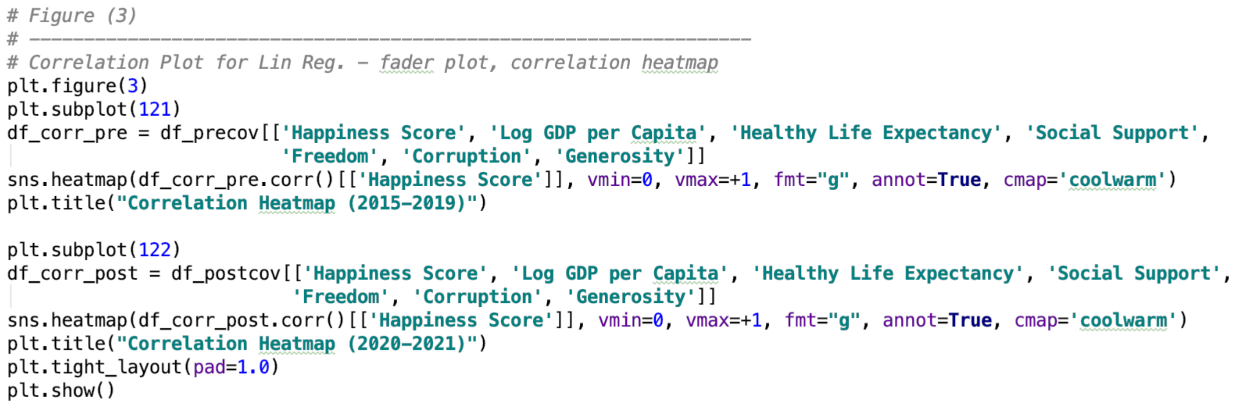


```
# ------------------------------------------------------------------
# Figure (1) Linear Relationships (2015–2019)
plt.figure(1)
plt.subplot(321)
plt.suptitle("Effect on Happiness Score (2015–2019)")
depvar = df_precov['Happiness Score']
deptitle = "Happiness Score"
indvar = df_precov['Social Support']
indtitle = "Social Support"
title = deptitle + " vs. " + indtitle
plt.scatter(indvar, depvar, alpha=0.5, edgecolors='blue', facecolors='blue')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title(title)
plt.xlabel(indtitle)
plt.ylabel(deptitle)
plt.subplot(322)
depvar = df_precov['Happiness Score']
indvar = df_precov['Healthy Life Expectancy']
indtitle = "Healthy Life Expectancy"
title = deptitle + " vs. " + indtitle
plt.scatter(indvar, depvar, alpha=0.5, edgecolors='purple', facecolors='purple')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title(title)
plt.xlabel(indtitle)
plt.ylabel(deptitle)
plt.subplot(323)
depvar = df_precov['Happiness Score']
indvar = df_precov['Freedom']
indtitle = "Freedom"
title = deptitle + " vs. " + indtitle
plt.scatter(indvar, depvar, alpha=0.5, edgecolors='darkorange', facecolors='darkorange')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title(title)
plt.xlabel(indtitle)
plt.ylabel(deptitle)
plt.subplot(324)
depvar = df_precov['Happiness Score']
indvar = df_precov['Corruption']
indtitle = "Corruption"
title = deptitle + " vs. " + indtitle
plt.scatter(indvar, depvar, alpha=0.5, edgecolors='green', facecolors='green')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title(title)
plt.xlabel(indtitle)
```



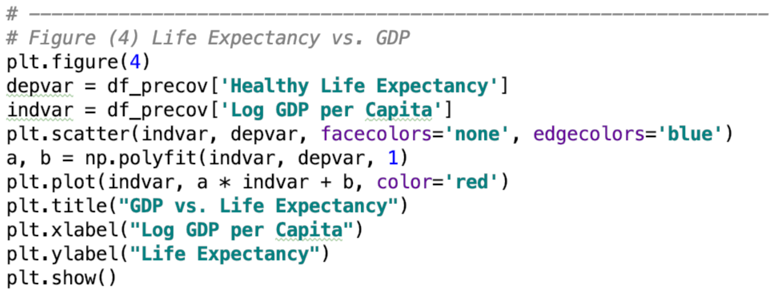
```
plt.ylabel(deptitle)
plt.subplot(325)
depvar = df_precov['Happiness Score']
indvar = df_precov['Generosity']
indtitle = "Generosity"
title = deptitle + " vs. " + indtitle
plt.scatter(indvar, depvar, alpha=0.5, edgecolors='magenta', facecolors='magenta')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title(title)
plt.xlabel(indtitle)
plt.ylabel(deptitle)
plt.subplot(326)
depvar = df_precov['Happiness Score']
indvar = df_precov['Log GDP per Capita']
indtitle = "GDP"
title = deptitle + " vs. " + indtitle
plt.scatter(indvar, depvar, alpha=0.5, edgecolors='black', facecolors='black')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title(title)
plt.xlabel(indtitle)
plt.ylabel(deptitle)
plt.tight_layout(pad=1.0)
plt.show()
```

Code for Correlation heatmap visual for 2015 - 2019 vs. 2020-2021



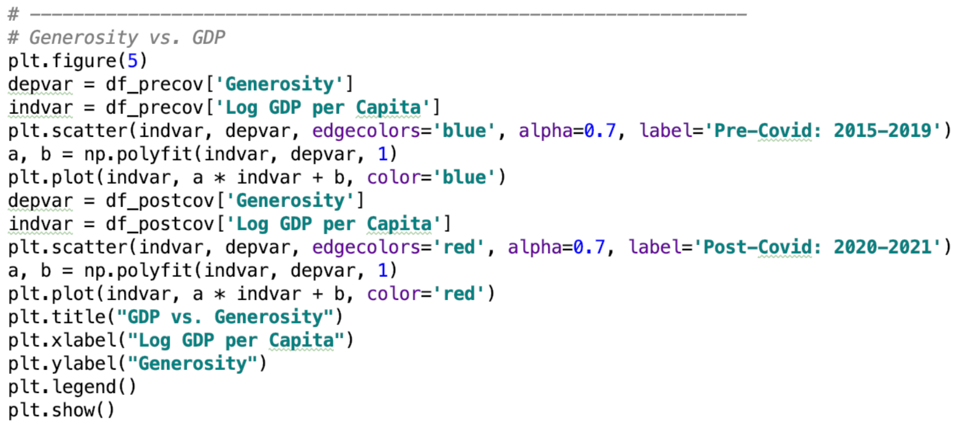
```
# Figure (3)
# ------------------------------------------------------------------
# Correlation Plot for Lin Reg. - fader plot, correlation heatmap
plt.figure(3)
plt.subplot(121)
df_corr_pre = df_precov[['Happiness Score', 'Log GDP per Capita', 'Healthy Life Expectancy', 'Social Support',
                         'Freedom', 'Corruption', 'Generosity']]
sns.heatmap(df_corr_pre.corr()[['Happiness Score']], vmin=0, vmax=+1, fmt="g", annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap (2015-2019)")

plt.subplot(122)
df_corr_post = df_postcov[['Happiness Score', 'Log GDP per Capita', 'Healthy Life Expectancy', 'Social Support',
                          'Freedom', 'Corruption', 'Generosity']]
sns.heatmap(df_corr_post.corr()[['Happiness Score']], vmin=0, vmax=+1, fmt="g", annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap (2020-2021)")
plt.tight_layout(pad=1.0)
plt.show()
```

Code for Life Expectancy and Log GDP per Capita relationship visual



```
# ------------------------------------------------------------------
# Figure (4) Life Expectancy vs. GDP
plt.figure(4)
depvar = df_precov['Healthy Life Expectancy']
indvar = df_precov['Log GDP per Capita']
plt.scatter(indvar, depvar, facecolors='none', edgecolors='blue')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title("GDP vs. Life Expectancy")
plt.xlabel("Log GDP per Capita")
plt.ylabel("Life Expectancy")
plt.show()
```

Generosity vs. Log GDP per Capita relationship visual



```
# -------------------------------------------------------------------
# Generosity vs. GDP
plt.figure(5)
depvar = df_precov['Generosity']
indvar = df_precov['Log GDP per Capita']
plt.scatter(indvar, depvar, edgecolors='blue', alpha=0.7, label='Pre-Covid: 2015-2019')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='blue')
depvar = df_postcov['Generosity']
indvar = df_postcov['Log GDP per Capita']
plt.scatter(indvar, depvar, edgecolors='red', alpha=0.7, label='Post-Covid: 2020-2021')
a, b = np.polyfit(indvar, depvar, 1)
plt.plot(indvar, a * indvar + b, color='red')
plt.title("GDP vs. Generosity")
plt.xlabel("Log GDP per Capita")
plt.ylabel("Generosity")
plt.legend()
plt.show()
```

Appendix B

Risk Section

With the data we chose the highest risk was inconsistent or incorrect data, especially since each year's values have their own separate dataset with different variables and measurements. Luckily, the statistical appendices from the nonprofit that aggregated the data for each year showed consistent measurement methods. Our only problem was that there was corrupted data for the Log GDP per Capita field in the years 2015-2019; it appears that the original values underwent some type of normalization that we couldn't reverse engineer. To replace the incorrect data, we imported GDP per Capita values from the World Bank and transformed them using the log function.

Another potential risk we identified early on was there being too much data to process locally. This was always a very unlikely scenario, since each of the seven csv files is less than 50 kilobytes. However, the datasets had a number of redundant fields: some years had a "Region" field that was nonexistent in others but could be inferred from the "Country" field. Some years also had statistics such as quartiles, rankings, and multiple regression coefficients that were calculated by the data collectors. All of these were removed (since we want to conduct our own analyses) and therefore the datasets were made much smaller and faster to process locally.

The final risk we were concerned with was the failure of our analytical methods. The solution we were ready to implement was to simply change the type of statistical tests/visualizations depending on what we learn about the data. For example, we conducted a t-test on the pre and post-COVID years to see if there was a statistically significant difference in their average happiness scores. To justify the t-test we first had to prove that the happiness scores followed a normal distribution. If that failed, we would have had to switch to a non-parametric test (such as the Wilcoxon Signed Rank test). Luckily, we were able to conduct the t-test as we originally planned, but if this problem did come up the impact would have been relatively low.

Appendix C

Agile development

In order to assign tasks, track progress, and organize our project we utilized YouTrack's SCRUM board, which is an online resource that each group member was able to view and edit as they completed their assigned tasks. SCRUM is part of the Agile framework, which is an efficient approach to work planning, management, and execution.

Appendix D

References

Dataset source -

Ache, Mathurin. "World Happiness Report 2015-2021." *Kaggle*, 19 Mar. 2021, https://www.kaggle.com/datasets/mathurinache/world-happiness-report-20152021

Alternate Data sources considered -

Ache, Mathurin. "World Happiness Report 2015-2021." *Kaggle*, 19 Mar. 2021, https://www.kaggle.com/datasets/mathurinache/world-happiness-report-20152021

Singh, Ajaypal. "World Happiness Report 2021." *Kaggle*, 22 Mar. 2021, https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021

Understanding Gallup Polls

Gallup. "How Does the Gallup World Poll Work?" *Gallup.com*, Gallup, 1 Feb. 2022, https://www.gallup.com/178667/gallup-world-poll-work.aspx

References for Python coding for various visualizations

Prabhakaran, Selva. "Top 50 Matplotlib Visualizations - the Master Plots (W/ Full Python Code): ML+." *Machine Learning Plus*, 6 May 2022, https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python

"Mapping and Plotting Tools." *Mapping and Plotting Tools - GeoPandas*, https://geopandas.org/en/stable/docs/user_guide/mapping.html?msclkid=eb063356d16511ec8413451ceba066a9

Panda, Aayush. "Installing Gdal with Python Binders." Python in Plain English, 21 July 2021, https://python.plainenglish.io/installing-gdal-with-python-binders-ce41c641808f

"Correlation Matrix Plot with Coefficients on One Side, Scatterplots on Another, and Distributions on Diagonal." *Stack Overflow*, https://stackoverflow.com/questions/48139899/correlation-matrix-plot-with-coefficients-on-one-side-scatterplots-on-another