# World Happiness Data Analysis

Analysis of data from the World Happiness Report of the Year 2019

Prepared by:

Rashi Saki

May 16, 2023

# Introduction

A country's world happiness and the factors contributing to it can tell anyone examining it a lot about the past, current and future states of that country. While studying a single country can provide a lot of insight, those countries that surround it as well as looking at an entire continent's happiness can also provide additional, supporting evidence for high or low levels of happiness. Is the country an outlier or does the entire continent follow similar trends? Some countries are plagued by year-after-year government corruption, others by war. Some benefit from their country's rich resources and are beneficiaries of their country's location, however, not all within that country may benefit and rather benefit a select few. Others rely on their own altruistic behavior, that of their neighbors, and their society as a whole for parts of their happiness. Each of these factors by themselves are not sole determinants of happiness, but combined can paint a bigger picture. To what extent and to what degree do each of these factors play into the overall happiness of a country?
The dataset that was studied was from Kaggle and we choose to look at one individual year, 2019. While there were additional datasets after 2019, we chose to study the years prior to the COVID-19 outbreak to reduce the potential for outliers as the effects of this pandemic are continuing to be understood.

**World Happiness 2019 Dataset:** 2019 World Happiness report analyzes 156 countries and looks to understand how each individual score drives high and low ratings. To access each score, the Cantril ladder is used while asking and evaluating each respondent's answer. By having respondents think of their lives as a ladder, ranking each step from a 0 to 10 scale. At the top represents the best possible life, at the bottom represents the worst possible life. In this particular survey, there were 6 different categories analyzed to help conclude the happiness of a particular country and the overall happiness score of that particular country. Our team believed that among these factors, a country's overall happiness will be strongly correlated with their society's economic output and their average life in good health. Healthy life expectancy being the strongest driver for happiness of the 6 factors being studied.

| | | |
|---|---|---|
| 1. | Happiness Score | The calculation of the national average response to the 6 questions of life evaluations. |
| 2. | GDP per Capita | A core indicator of economic performance. A breakdown of a country's economic output per person. |
| 3. | Social Support | A network of that includes friends, family, neighbors, members of the community and their availability during times of need and reflect that someone. Can be reflected as emotional, physical, or financial support. |
| 4. | Healthy Life Expectancy | The indicator of one's average life in good health. |
| 5. | Freedom to Make Life Choices | The ability to choose how one decides to live one's life and to what extent. |
| 6. | Generosity | The number of citizens donating money to charity or giving back to their country. Not indicative of the quantity of the charity. |
| 7. | Perceptions of Corruption | The level of public sector corruption that are perceived by its society. |

# Data Extraction, Collection and Cleaning

Our group wanted to focus on 2019 for our study. However, while preparing our dataset we wanted to be able to review previous years in the case that data from 2019 was missing or had a zero value. As such, 2015 – 2018 were also imported into our notebook. The 2019 dataset was originally stored as a CSV file. To better read the file in our notebook we used, 'pd.read_csv', which placed the information into a chart. We added the 'str.replace' method to be able to change the spacing and allow for consistency and convenience. To be able to differentiate between the different years we were looking at and added a column for year.

```
>df_2019 = pd.read_csv('2019.csv')

>df_2019.columns = df_2019.columns.str.replace(' ', '_')

>df_2019.head()

>df_2019['Year'] = 2019

>df_2019.rename(columns = {'Country_or_region' : 'Country', 'Score':
'Happiness_score'}, inplace = True)

>df_2019 =
df_2019[['Country','Overall_rank','Year','Continent','Happiness_score','GDP_per_capita','
Social_support','Healthy_life_expectancy',
'Freedom_to_make_life_choices','Generosity','Perceptions_of_corruption']]
```

The exact process was followed in earlier years to ensure that all datasets were consistently named and that there would be no confusion. In previous datasets, names such as Score or Happiness Rank were used but ultimately, we chose titles that best and clearly reflected what our columns represented. Additionally, by restructuring our columns in a way that was consistent throughout our datasets, visualization and analysis of our code in our notebooks was simpler.

**Handling Missing Values:**

```
>missing_data_2019 = df_2019.isnull().sum()

>zero_val_2019 = (df_2019 == 0).sum()

>df_2019[df_2019.GDP_per_capita == 0]
```

```
zero_val_2019 = (df_2019 == 0).sum()
zero_val_2019
```

```
Country                         0
Overall_rank                    0
Year                            0
Continent                       0
Happiness_score                 0
GDP_per_capita                  1
Social_support                  1
Healthy_life_expectancy         1
Freedom_to_make_life_choices    1
Generosity                      1
Perceptions_of_corruption       1
dtype: int64
```

```
>df_2019.loc[df_2019['Country'] == 'Somalia', 'GDP_per_capita'] = 0.0243378507
```

- We found there to be no missing values from our 2019 dataset

- There were 6 values of zero:

- GDP per capita – Prior year GDP growth collected from macro trends was used to help retrieve 2019 data

- Social support – Prior years also had a social support of zero. Using the World Bank website, we determined that Central African Republic may indeed have a social support rating of zero

- Healthy life expectancy – 2015 was used to represent Swaziland's healthy life expectancy as a reference point and the WHO website was used to determine from 2015 to 2019 there was a 1.27% increase in healthy life expectancy.

- Freedom to make life choices – Based on Freedom house, from 2018 to 2019, there was a 237.5% increase. Bringing Afghanistan's score to 0.286875

- Generosity – No evidence could be found for Greece's Generosity score. Left within dataset as they have gone through years of economic hardship which could potentially represent a score of zero

- Perceptions of corruption – Transparency International calculated Maldova's score to have increased by 3.03% from 2017 to 2019. Previously their score was 0.10091(2017). New score is 0.010396(2019).

By looking at previous years and using additional data from websites such as The World Bank and Macrotrends, we were able to better understand whether these values were truly zero or needed to be replaced. Had there been missing values returned, they would have been handled in a similar manner to that of zero values. Overall and after further view of previous years, the dataset was complete and consistent.

**Removing Duplicates**

```
>duplicates = df_2019[df_2019.duplicated()]
```

- The World Happiness Report had values for each of 156 countries and none were repeated, making the handling of rows as well as the columns an efficient process.

**Standardizing Data:**

```
>df_2019.dtypes

>df_2019.shape
```

- By looking at all the data types of our dataset, we were able to see that all the necessary information that was needed for our analysis was either a float or an int. There were additional columns, Country and Continent, but these did not require any additional alterations.

# Data Transforming

Data transforming is the process of converting raw data into a format that is more suitable for analysis. It involves modifying, restructuring, or aggregating data to make it more useful and informative.

We've used the following Data transformation techniques:

**Joining/Merging:**

- Combining multiple datasets based on common variables or keys to create a single dataset. The continents column is added to the countries dataset to visualize data continent-wise.

```
#Adding continents dictionary. It has continent as the key and correspnding countries as the values.
continent_dict = {
    'Asia': ['Afghanistan', 'Bahrain', 'Iran', 'Iraq', 'Israel', 'Jordan', 'Kuwait', 'Lebanon', 'Oman', 'Pakistan', 'Qatar', 'Saudi Arabia', 'Syria', 'Turkey', 'United Arab Em
    'Africa': ['Algeria', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cabo Verde', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo (Brazzavill
    'North America': ['Antigua and Barbuda', 'Bahamas', 'Barbados', 'Belize', 'Canada', 'Costa Rica', 'Cuba', 'Dominica', 'Dominican Republic', 'El Salvador', 'Grenada', 'Guat
    'South America': ['Argentina', 'Bolivia', 'Brazil', 'Chile', 'Colombia', 'Ecuador', 'Guyana', 'Paraguay', 'Peru', 'Suriname', 'Trinidad & Tobago', 'Uruguay', 'Venezuela'],
    'Europe': ['Albania', 'Andorra', 'Austria','Azerbaijan', 'Belarus', 'Belgium', 'Bosnia and Herzegovina', 'Bulgaria', 'Croatia', 'Cyprus', 'Czech Republic', 'Denmark', 'Est
    'Oceania': ['Australia', 'Fiji', 'Kiribati', 'Marshall Islands', 'Micronesia', 'Nauru', 'New Zealand', 'Palau', 'Papua New Guinea', 'Samoa', 'Solomon Islands', 'Tonga', 'T
    }
```

```
#Adding continent column to the 2019 dataset using lambda function
df_2019['Continent'] = df_2019['Country'].apply(lambda x: next((k for k in continent_dict if x in continent_dict[k]), None))
df_2019.head()
```

| erall_rank | Country | Happiness_score | GDP_per_capita | Social_support | Healthy_life_expectancy | Freedom_to_make_life_choices | Generosity | Perceptions_of_corruption | Year | Continent |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Finland | 7.769 | 1.340 | 1.587 | 0.986 | 0.596 | 0.153 | 0.393 | 2019 | Europe |
| 2 | Denmark | 7.600 | 1.383 | 1.573 | 0.996 | 0.592 | 0.252 | 0.410 | 2019 | Europe |
| 3 | Norway | 7.554 | 1.488 | 1.582 | 1.028 | 0.603 | 0.271 | 0.341 | 2019 | Europe |
| 4 | Iceland | 7.494 | 1.380 | 1.624 | 1.026 | 0.591 | 0.354 | 0.118 | 2019 | Europe |
| 5 | Nether... | 7.488 | 1.396 | 1.522 | 0.999 | 0.557 | 0.322 | 0.298 | 2019 | Europe |

**Filtering:**

- Removing unwanted data from a dataset based on specified criteria. We dropped columns from the dataset that we'll not use.

```
df_2016.drop(['Lower_Confidence_Interval','Upper_Confidence_Interval','Dystopia_Residual','Region'], axis = 1, inplace = True)

df_2015.drop(['Region','Standard_Error','Dystopia_Residual'], axis = 1, inplace = True)
```

## Sorting:

- Tables are sorted in descending order to view countries having the highest happiness scores and countries having the lowest happiness score.

```
df_2019.sort_values(['Happiness_score'], ascending = False)

#Of all continents, Africa has a lower overall Happiness_score. Now we look into what factors impact this
```

| | Country | Overall_rank | Year | Continent | Happiness_score | GDP_per_capita | Social_support | Healthy_life_expectancy |
|---|---|---|---|---|---|---|---|---|
| **0** | Finland | 1 | 2019 | Europe | 7.769 | 1.340 | 1.587 | 0.986 |
| **1** | Denmark | 2 | 2019 | Europe | 7.600 | 1.383 | 1.573 | 0.996 |
| **2** | Norway | 3 | 2019 | Europe | 7.554 | 1.488 | 1.582 | 1.028 |
| **3** | Iceland | 4 | 2019 | Europe | 7.494 | 1.380 | 1.624 | 1.026 |
| **4** | Netherlands | 5 | 2019 | Europe | 7.488 | 1.396 | 1.522 | 0.999 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **151** | Rwanda | 152 | 2019 | Africa | 3.334 | 0.359 | 0.711 | 0.614 |
| **152** | Tanzania | 153 | 2019 | Africa | 3.231 | 0.476 | 0.885 | 0.499 |
| **153** | Afghanistan | 154 | 2019 | Asia | 3.203 | 0.350 | 0.517 | 0.361 |
| **154** | Central African Republic | 155 | 2019 | Africa | 3.083 | 0.026 | 0.000 | 0.105 |
| **155** | South Sudan | 156 | 2019 | Africa | 2.853 | 0.306 | 0.575 | 0.295 |

156 rows × 11 columns

## Aggregating:

- Combining data to create summary statistics, such as counts, averages, or percentages.

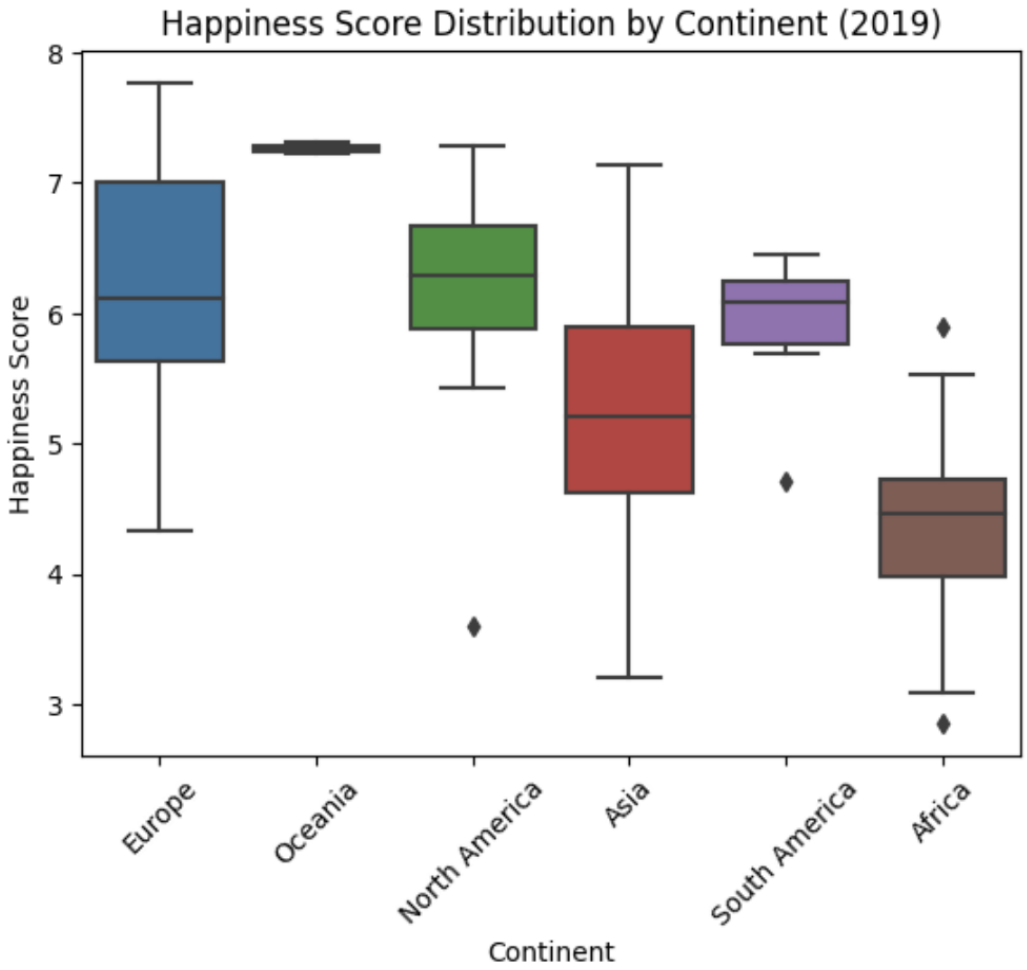
```
df_2019.groupby('Continent')['Happiness_score'].describe()
```

| Continent | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Africa** | 45.0 | 4.368289 | 0.643687 | 2.853 | 3.97500 | 4.4610 | 4.72200 | 5.888 |
| **Asia** | 43.0 | 5.224442 | 0.886827 | 3.203 | 4.62750 | 5.2080 | 5.89050 | 7.139 |
| **Europe** | 43.0 | 6.245000 | 0.861794 | 4.332 | 5.62550 | 6.1180 | 7.00300 | 7.769 |
| **North America** | 12.0 | 6.151583 | 0.971130 | 3.597 | 5.88250 | 6.2870 | 6.66925 | 7.278 |
| **Oceania** | 2.0 | 7.267500 | 0.055861 | 7.228 | 7.24775 | 7.2675 | 7.28725 | 7.307 |
| **South America** | 11.0 | 5.944909 | 0.477370 | 4.707 | 5.76100 | 6.0860 | 6.24250 | 6.444 |

# Data visualization

While conducting data visualization of the 2019 World Happiness Report, all graphs are displayed continent-wise to better understand which continent has the happiest countries and what factors affect their happiness score.

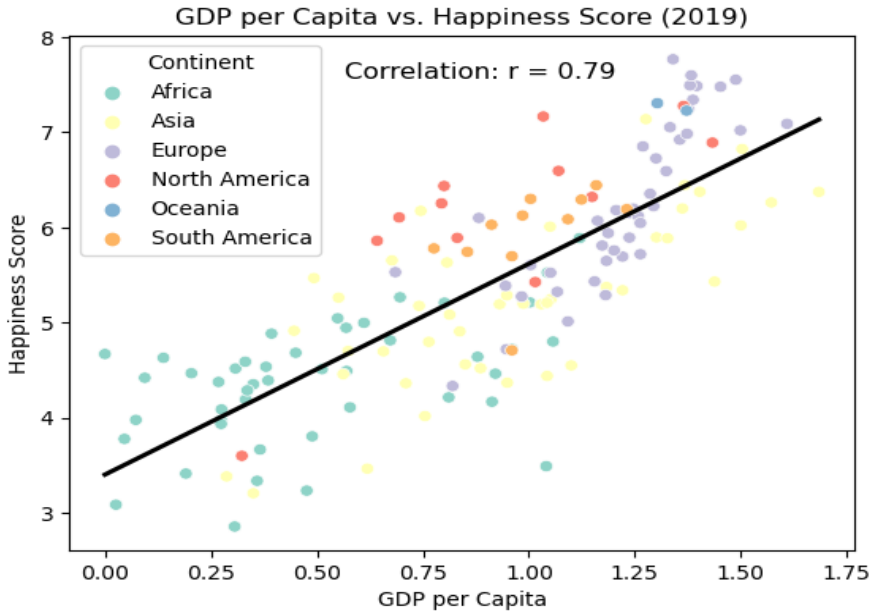
- To get an overall picture of the dataset, we used **Box Plot** using Seaborn library.
    - Europe has the highest happiness score followed by Oceania, North America, Asia, South America, and Africa.
    - Oceania has only 2 countries(New Zealand and Australia) hence it is displayed as a thin box plot. Europe has a maximum number of countries.
    - The Box plot shows that there are 2 outliers in Africa, they are the minimum & maximum happiness scores in Africa lying outside the distribution. Similarly in North America & South America their minimum lies outside box distribution. Both Europe and Oceania don't have outliers. Overall, there are only a total of 4 outliers as it is presented where the minimum and maximum values are outside of the distribution.



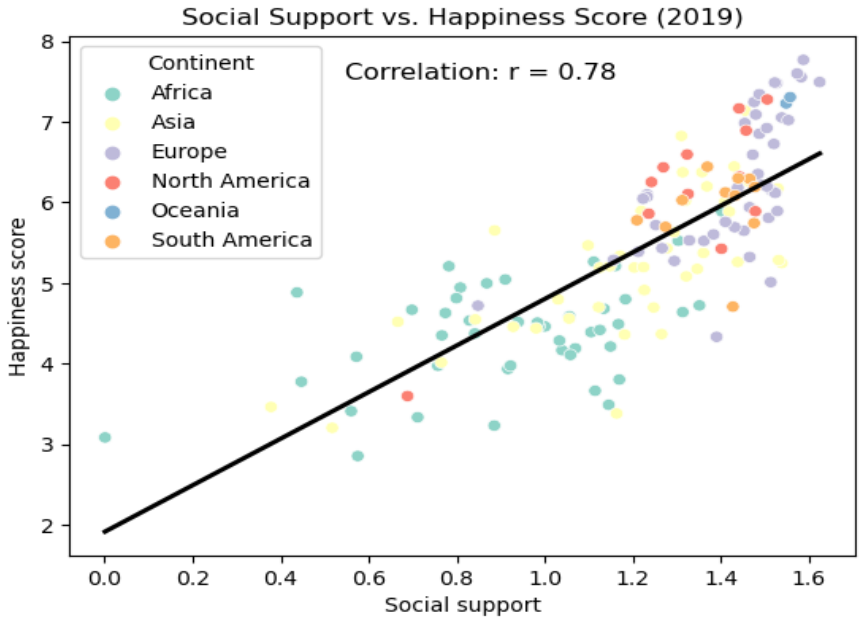

- To further look into how each factor is related to the happiness score, we used **Scatter Plots** to compare the happiness score with 6 factors- GDP per Capita, Social Support,

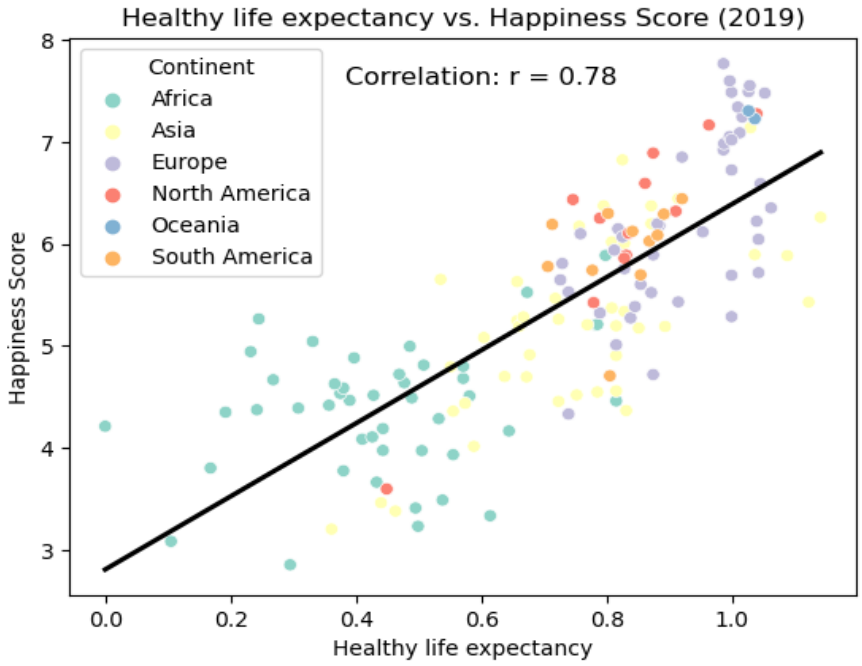
Healthy Life Expectancy, Freedom of making life choices, Generosity, and Perception of corruption. Scatter plots are used to help detect patterns or trends and compare groups (continents)

1. The scatter plot when comparing GDP per Capita vs. Happiness Score demonstrates a positive correlation with a correlation coefficient, r = 0.79.

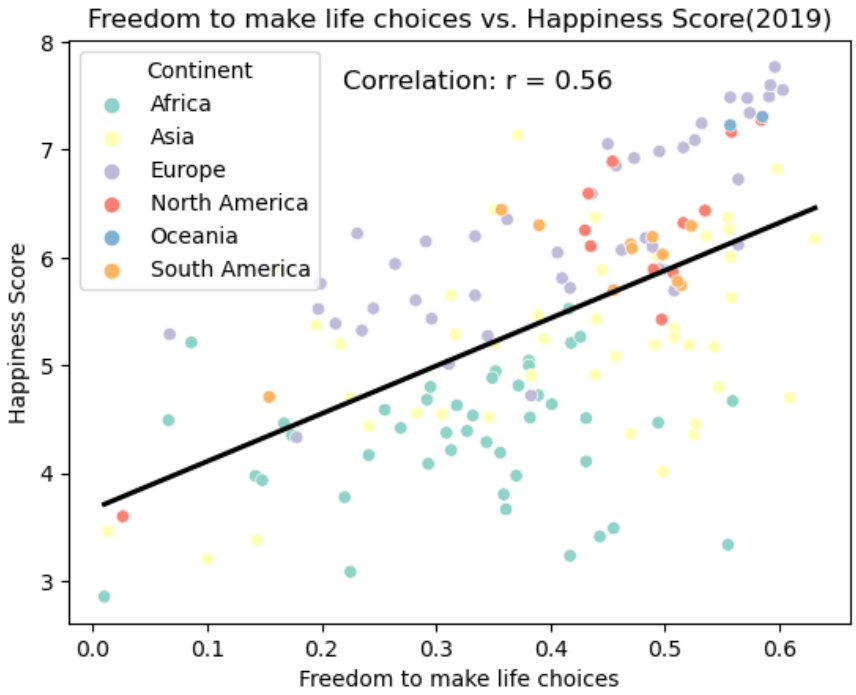



2. Both the scatter plots of Social Support and Healthy life expectancy are highly positive correlated to the happiness score, their correlation coefficient, r = 0.78.

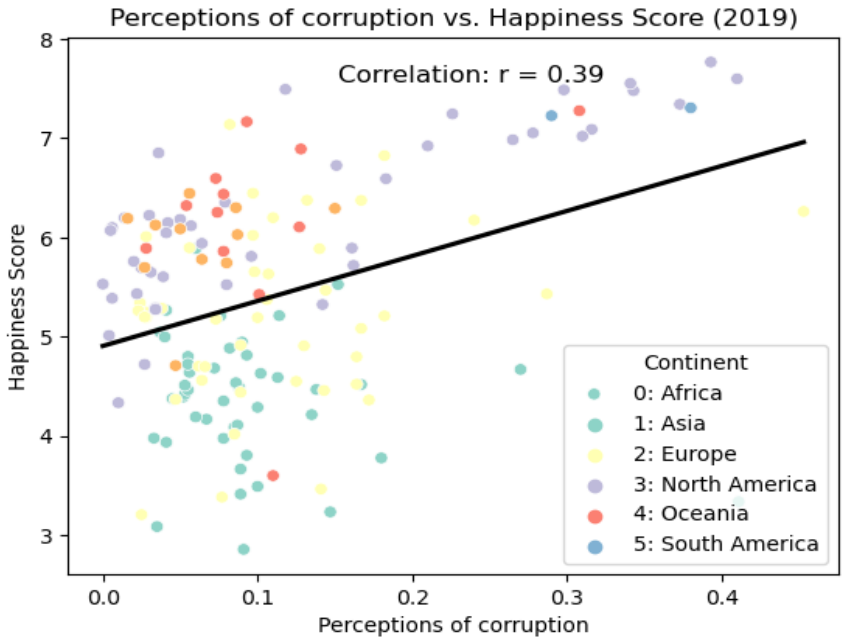




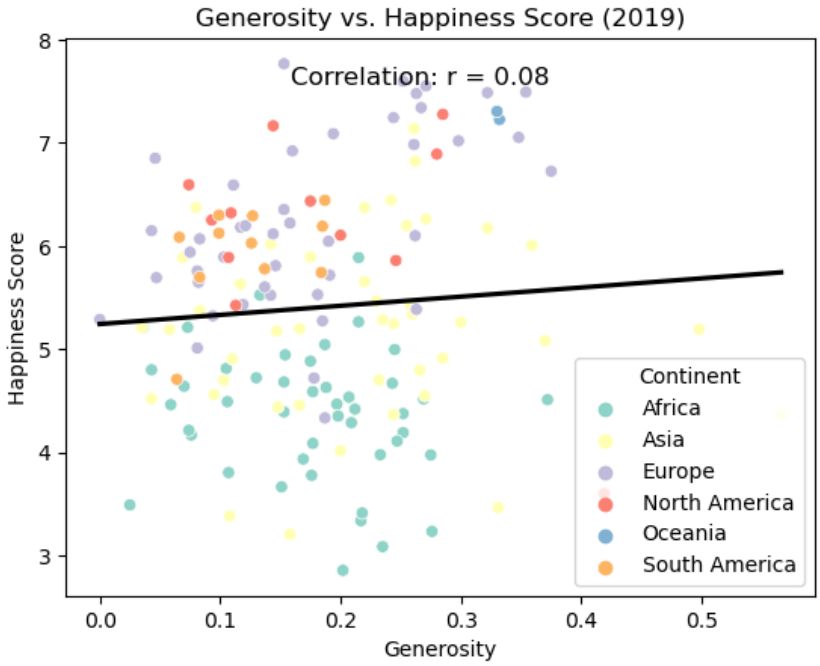

3. Comparing the Freedom to make life choices vs Happiness score shows that the scatter plot has as a low positive correlation with a value of correlation coefficient, r= 0.56.



4. This graph comparing the perceptions of corruption and the happiness score is showing that there is a moderate positive correlation while most of the data is moving to the left bound of the scatter plot.




5. This graph below is interpreted as having almost no correlation because the r value is 0.08 which is close to 0 when comparing Generosity with Happiness score.



# Conclusion

- Based on the analysis of the 2019 happiness score report from all countries continent-wise, we created scatter plots to examine the correlation between the happiness score and six factors: GDP per capita, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, and Perceptions of Corruption.

```
from tabulate import tabulate
A = 'Happiness score'
table = [[' ','Factors','Correlation Coefficient'],[A,'GDP per capita',0.79],[A,'Social Support',0.78],
print(tabulate(table,headers='firstrow'))
```

| | Factors | Correlation Coefficient |
|---|---|---|
| Happiness score | GDP per capita | 0.79 |
| Happiness score | Social Support | 0.78 |
| happiness score | Healthy Life Expectancy | 0.78 |
| Happiness score | Freedom to Make Life Choices | 0.56 |
| Happiness score | Perceptions of Corruption | 0.39 |
| Happiness score | Generosity | 0.08 |

- The highest correlation coefficient(r) is observed between the happiness score and **GDP per capita**, with a value of 0.79. This suggests a **strong positive relationship**, indicating that countries with higher GDP per capita tend to have higher happiness scores.

- Similarly, the r value between the happiness score and **Social Support**, as well as **Healthy Life Expectancy**, are both 0.78. These values indicate **strong positive correlations**, indicating that countries with greater social support and longer healthy life expectancy tend to have higher happiness scores.

- r value between the happiness score and **Freedom to Make Life Choices** is 0.56 and **Perceptions of Corruption** is 0.39, indicating a **moderate positive correlation**. This suggests that countries with greater freedom to make life choices and lower levels of perceived corruption tend to have higher happiness scores, although the relationship is not as strong as with other factors.

- In contrast, the correlation coefficient between the happiness score and **Generosity** is 0.08, indicating a **weak positive correlation**. This implies that there is little to no meaningful relationship between a country's generosity and its happiness score.

- Overall, the analysis reveals that factors such as GDP per capita, Social Support, and Healthy Life Expectancy have strong positive correlations with the happiness score. These findings emphasize the importance of economic prosperity, social support systems, and good health in fostering higher levels of happiness among countries. However, factors such as Freedom to Make Life Choices, Generosity, and Perceptions of Corruption have weaker correlations, suggesting that they may have a less direct influence on overall happiness scores.

# References


World Happiness Report dataset from Kaggle
https://www.kaggle.com/datasets/unsdsn/world-happiness

*Overview*. (n.d.). World Bank. https://www.worldbank.org/en/country/centralafricanrepublic/overview#1

Somalia GDP Per Capita 1960-2023. (n.d.). MacroTrends.
https://www.macrotrends.net/countries/SOM/somalia/gdp-per-capita

Countries. (n.d.-b). https://data.who.int/countries/748

Freedom House. (n.d.). Afghanistan. In *Freedom House*.
https://freedomhouse.org/country/afghanistan/freedom-world/2019

*2019 Corruption Perceptions Index - Explore Moldova's results*. (2020, January 24). Transparency.org.
https://www.transparency.org/en/cpi/2019/index/mda