UCDPA - Certificate in Introductory Data Analytics

PROJECT REPORT

WORLD HAPPINESS INDEX REPORT ANALYSIS

- Sarthak Kapoor

GitHub URL:

https://github.com/SarthakKapoor1/UCDPA_SathakKapoor

Abstract:

To begin our data analysis project, we start by choosing a dataset. In this project we will be analysing the World Happiness Index[1] data from the year 2021 and then will also analyse the World Happiness data from 2017 to 2021 for a select few regions or countries. The reason why I have chosen this particular topic is because there are multiple factors which affect the happiness core of a country, and thus will give us a healthy number of insights. After going through the data, I also realised, that the possibilities to analyse and visualize this particular dataset are very vast.

Next, we can move on to importing the data and libraries we will use to analyse the data. After the data is imported into the form of Pandas dataframes, we can prepare, analyse and visualise the data. Based on which, we can generate valuable insights.

I have included all the milestones for the project and have inculcated all the leanings from both DataCamp and the weekly online lectures. Getting a hands-on experience in data analysis has truly been an enriching process.

Introduction:

The data used in this visualization come from the Gallup World Poll (GWP)[2], which ranks countries based on answers to the main life evaluation question asked in the poll. This question is called the Cantril ladder, and asks respondents to think of a ladder, where the best possible life would be a 10, and the worst a 0. They then rate their own lives on that 0 to 10 scale. This is also called Ladder Score. There are six main factors on which the happiness scale is dependent upon. These are:

1. **GDP per capita:** GDP per capita is a measure of a country's economic output that accounts for its number of people. It divides the country's gross domestic product by its total population
2. **Healthy Life Expectancy:** The estimated extent to which Health (Life Expectancy) contributes to the calculation of the Happiness Score. The time series of healthy life expectancy at birth are constructed based on data from the World Health Organization and World Development Indicators.

3. **Social Support:** Equal to the national average of the responses to the GWP 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?"
4. **Generosity:** Equal to the residual of regressing the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.
5. **Perceptions of Corruption:** Perceptions of corruption (Trust) are the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?"
6. **Freedom to make Life Choices:** Equal to the national average of binary responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

Each country is also compared against a hypothetical nation called Dystopia. Dystopia as a benchmark against which to compare contributions from each of the six factors. Dystopia is an imaginary country that has the world's least-happy people. Since life would be very unpleasant in a country with the world's lowest incomes, lowest life expectancy, lowest generosity, most corruption, least freedom, and least social support, it is referred to as "Dystopia," in contrast to Utopia.

Dataset:

The reason I selected the following datasets, from Kaggle.com[3] is because I was curious to analyse the impact of various factors on the world happiness report.

https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report-2021.csv

https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report.csv

This data is collected by Gallup World Poll. The Gallup World Poll—surveys are conducted on sample sizes of approximately 1000, depending on the size of the country and are done through telephone or face-to-face for developing countries. I used Kaggle, to download and import my datasets, because it has a wide collection and selection of open datasets. Kaggle is a website where one can find open source, and free to use datasets. This tool/ website is used by data analysts, data engineers and data scientists all over the world. It also lets users upload their own datasets and their own reports on any topic. There are datasets available for almost any topic, that one can think of. These datasets are also updated either annually, monthly or over a period of time, which makes the data relevant in the current world scenario.

Implementation Process:

I started by selecting a topic for my data analysis project. This took quite some time, but once I chose to go forward with data from the World Happiness Index, it was quite streamlined. I started by searching for sources for this data on the internet and came across Kaggle.com, from where I downloaded the csv files for this topic.

I started by importing a few libraries to the Jupyter Notebook that I was working on. I imported the following libraries, each having a different purpose for importing, analysing and visualising the datasets.

- NumPy[4] is one of the most commonly used packages for scientific computing in Python
- Pandas[5] package is a high-level data manipulation tool built on top of the NumPy package in Python
- Matplotlib[6] is a comprehensive library for creating static, animated, and interactive visualizations in Python
- Seaborn[7] is used for data visualization and exploratory data analysis and is built on top of Matplotlib in Python

After importing these libraries, I created dataframes of the csv files. Next, I viewed the data to identify which columns to drop from the dataset. We create a list with all the columns that we want to include and use the .copy() method to create a final dataframe. Post this, I changed the names of a few columns using the .rename() method. I did the same for both the datasets. I also changed the starting Index from zero to one. The following methods were used to sort and clean the data: .info(), .describe(), .shape and .dropna(). The null values were removed from our dataset to ensure the data was optimized.

The following step was to group the data by various factors to analyse the data more easily. I grouped the data by region and noticed there were 10 regions which did not make much sense. So, to fix this, I merged a few of these regions to form 5 new regions namely: Asia, Africa, Europe, South America and North America. Australia and New Zealand were grouped together with the United States of America and Canada to form the 'North America' region. I also used the .iloc[] method to create a dataframe of the top five and bottom five countries in the world.

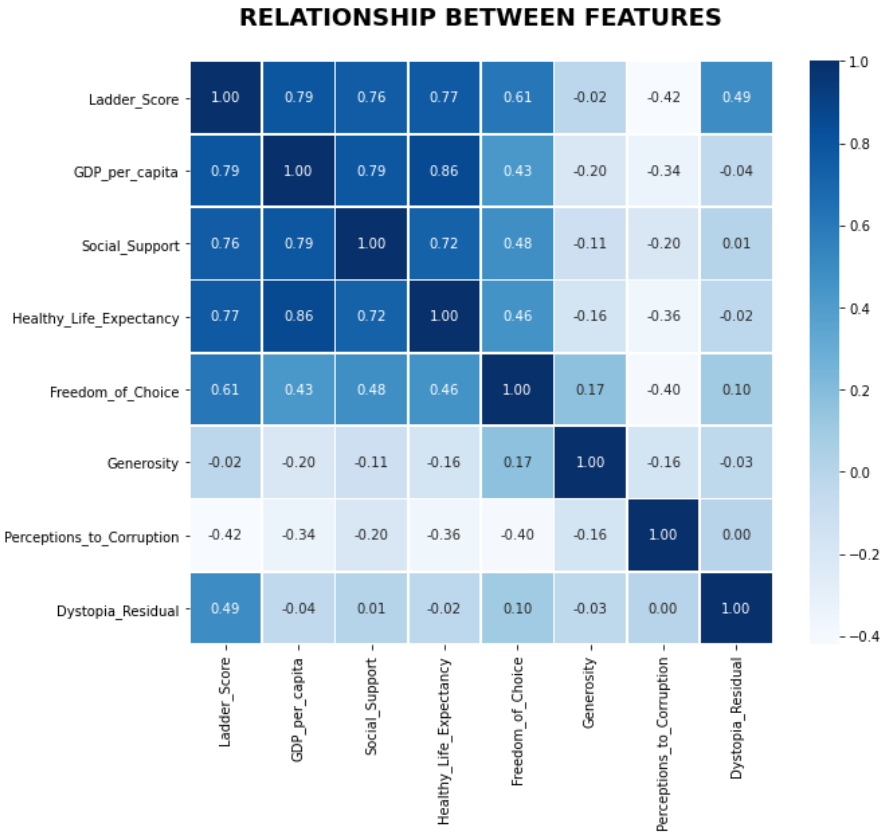
Next, it was time to visualize data. I visualized the data using the matplotlib.pyplot library and the Seaborn library, and was slightly inclined towards using the seaborn library more. The visualizations include the relationship between the features, GDP per capita by regions, view the country rankings based on regions, ladder score distribution and mean by countries, visualize the top and bottom 5 countries and analyse the happiness score trends by countries over a span of years. I used a lineplot, barplot, regplot, kdeplot, pie chart, bar graph, boxplot and a heatmap to show these visualizations.

I ended the notebook by listing a few conclusions and insights.

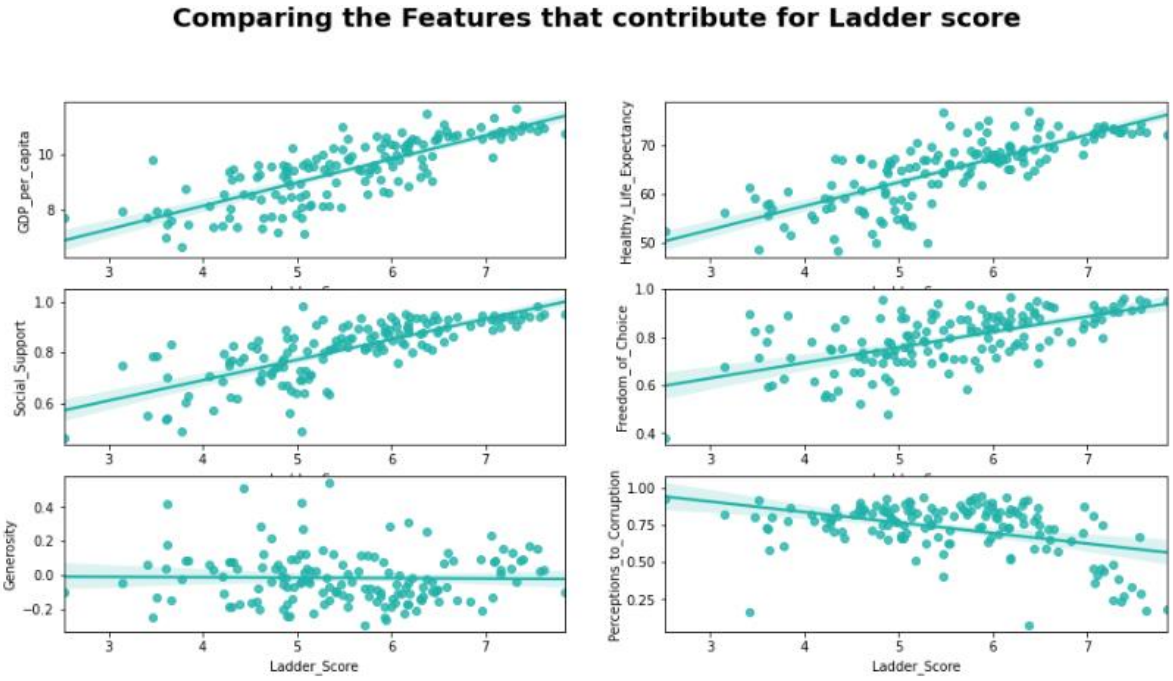
Results:

The following are a few of the visualizations from the plots.

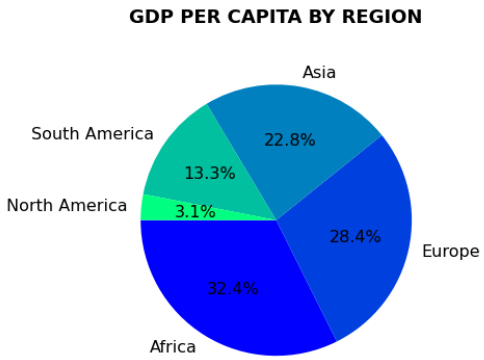
- Relationship between Features using Seaborn Heatmap:
  This is a heatmap showing how each of the features are dependent on each other. The heatmap can interpret the relationship between the features like Ladder Score, GDP per capita, Social Support, Healthy Life Expectancy, Freedom to make Life Decisions, Generosity, Perceptions to Corruption and Dystopia Residual



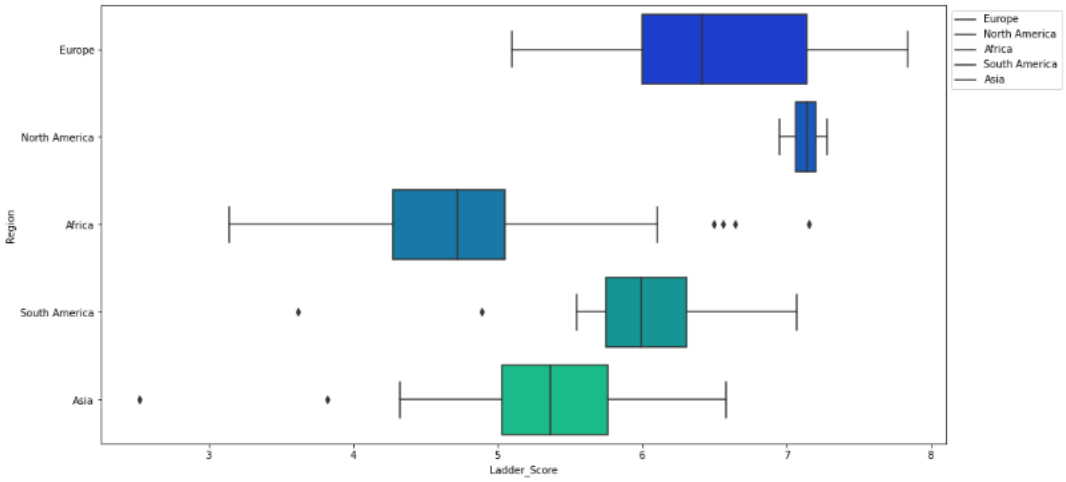
- Comparing the features that contribute to the Ladder Score using a Seaborn regplot: This shows that the higher the ladder score of a country is, the higher is the GDP per capita, healthy life expectancy, social support and freedom to make life decisions. A country with a higher ladder score, also tends to have lower perceptions to corruption within their government. Surprisingly, the "happier" countries are not higher on the generosity scale.



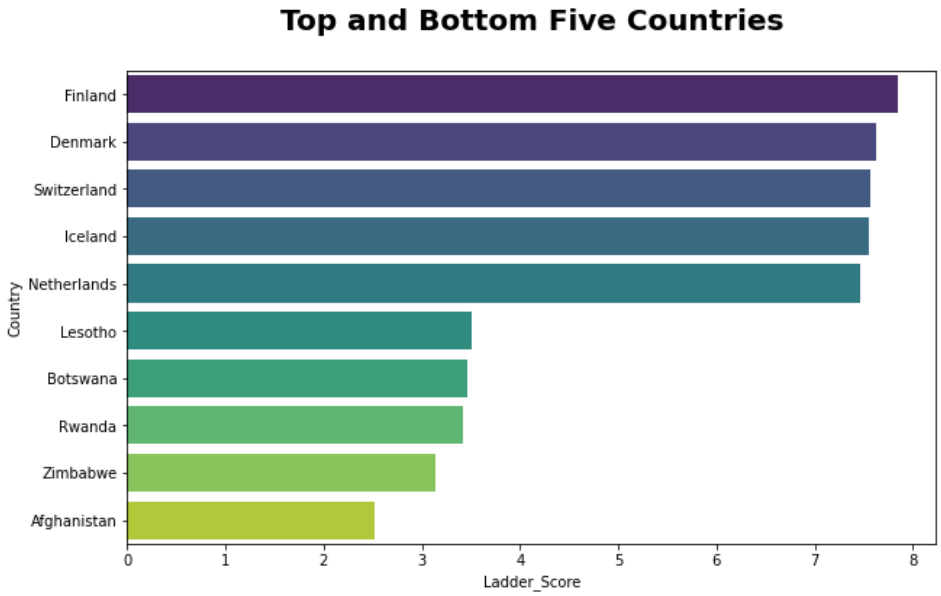
- GDP per Capita by Region using Pie Chart: We can see the GDP per capita by region



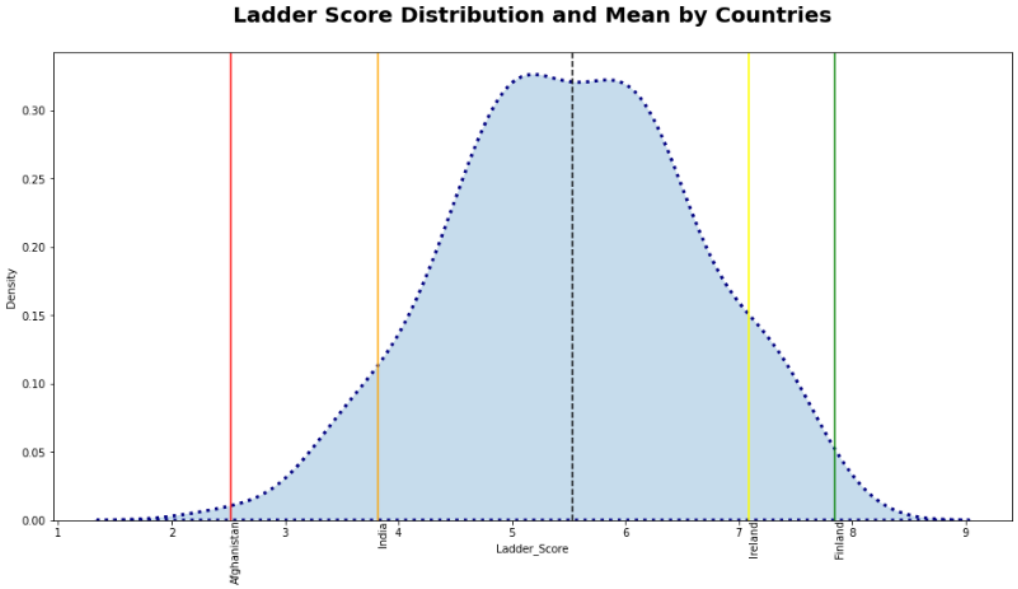
- Ladder Score vs Region using Seaborn boxplot:
  In this plot, we can see the lower limit, upper limit and mean of the ladder score for countries in a particular region. There are a few outliers in the plot as well.



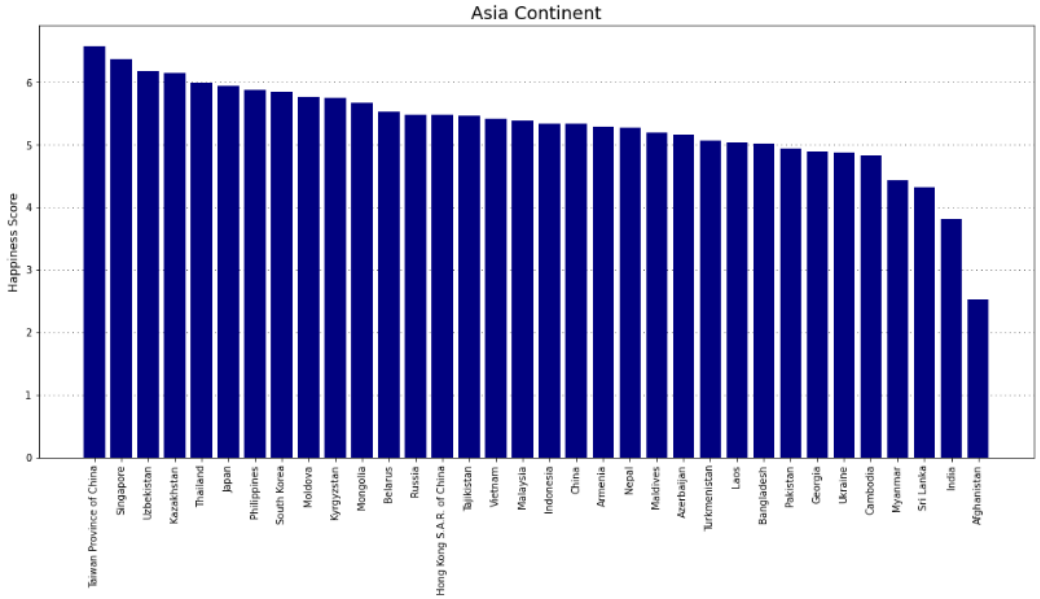
- Top Five and Bottom Five Countries using Seaborn barplot:
  Viewing the top and bottom five countries based on their happiness scores



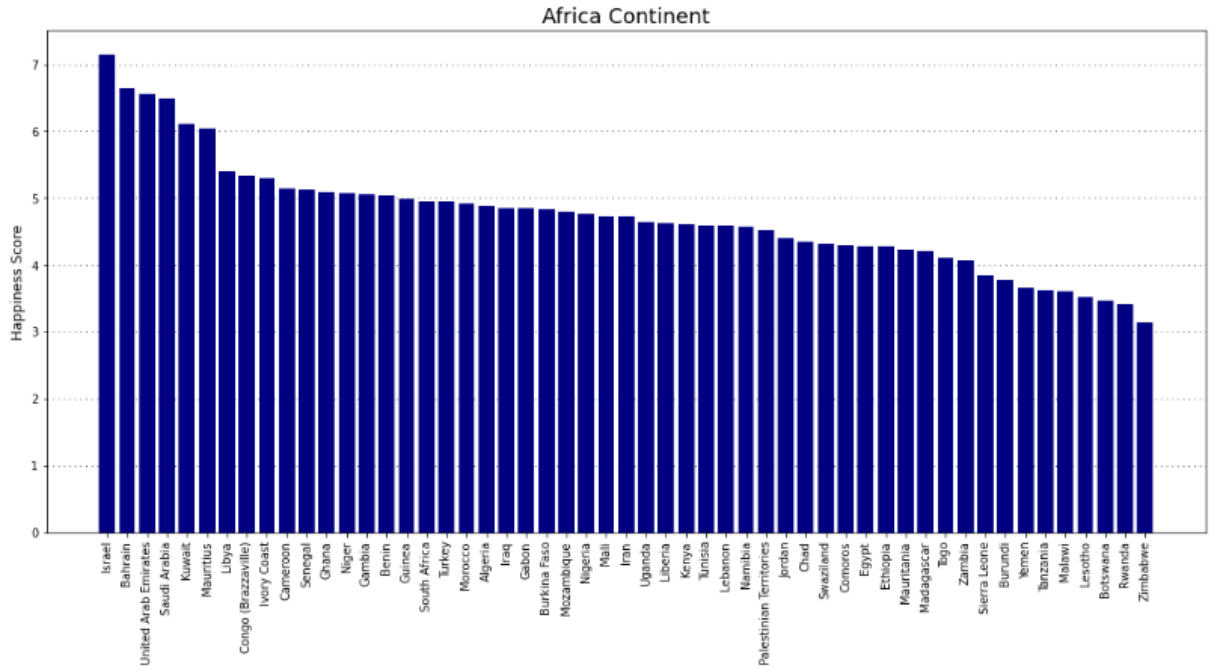
- Ladder Score Distribution and Mean by Countries using Seaborn kdeplot:
  Here, the red line Indicates the "unhappiest" country Afghanistan, followed by the orange line that indicates the Ladder score for India. The black dotted line is the mean of the ladder score for all countries. The next line, denotes Ireland, which is represented by the colour yellow. Lastly, the "happiest" country Finland is represented as a green vertical line.



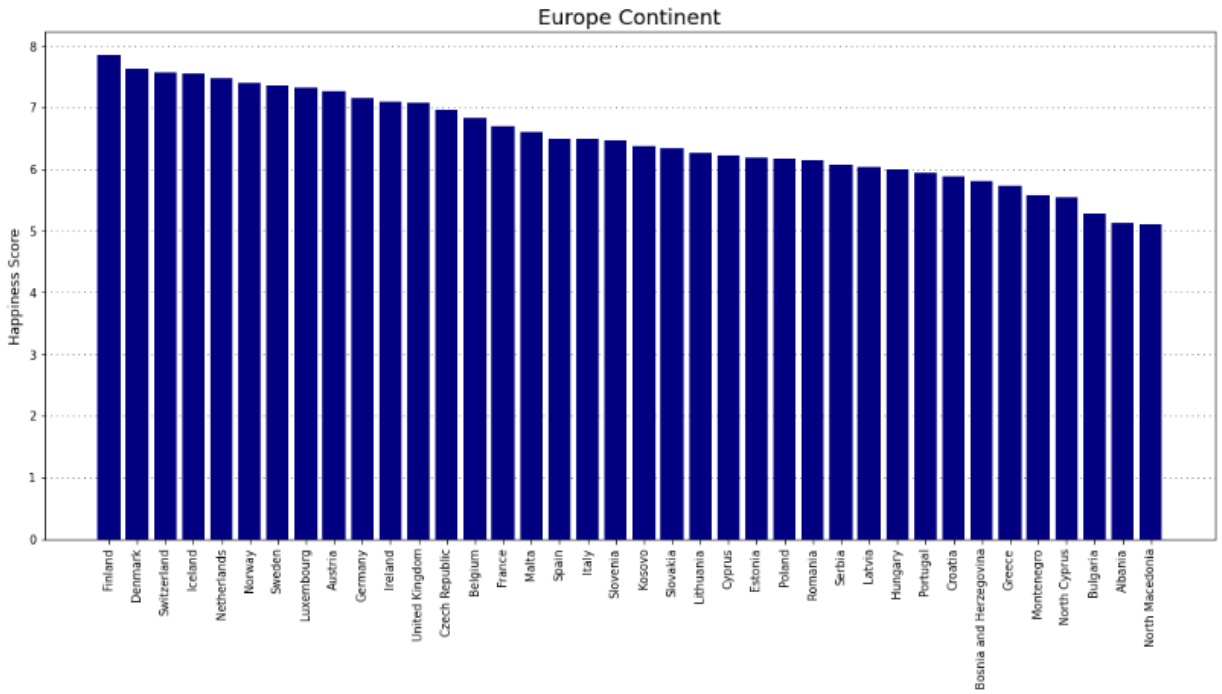
- World's Happiest Countries by Region using matplotlib.pyplot bar graph:
  - Asia:
    The 'happiest' country in Asia is Taiwan, and the 'unhappiest' country is Afghanistan



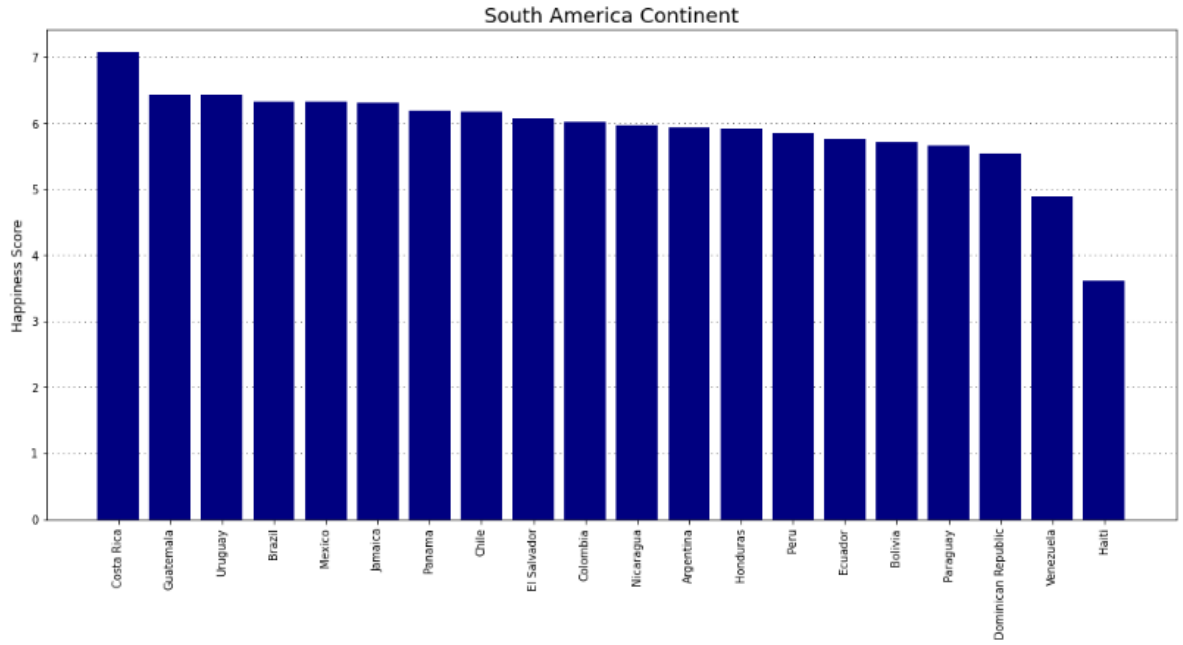
- Africa:
  The 'happiest' country in Africa is Israel, and the 'unhappiest' country is Zimbabwe



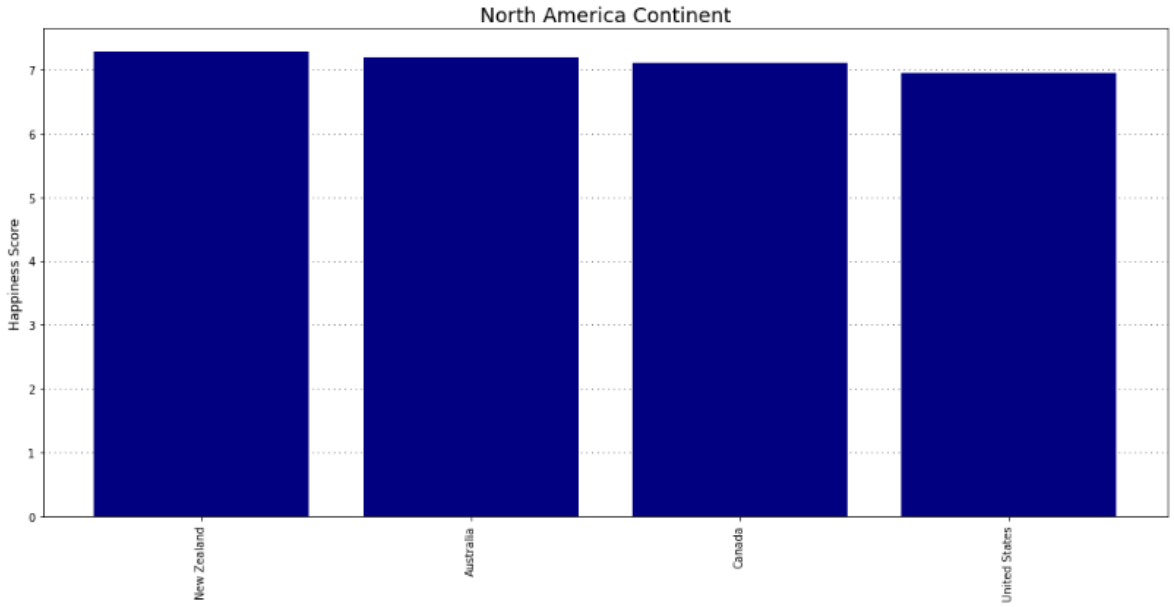
- Europe:
  The 'happiest' country in Europe is Finland, and the 'unhappiest' country is North Macedonia



- South America:
  The 'happiest' country in South America is Costa Rica, and the 'unhappiest' country is Haiti



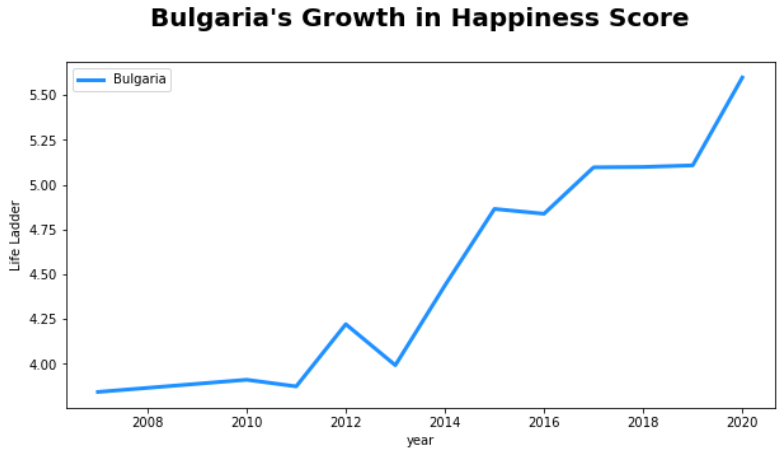
- North America:
  The 'happiest' country in North America is New Zealand, and the 'unhappiest' country is USA



Now, let's see the growth or decline of a select few countries over a period of 15 years, right from 2006 to 2020. I have visualized the trend for Ireland, Finland, Bulgaria, Afghanistan, the top 5 countries and the bottom 5 countries.

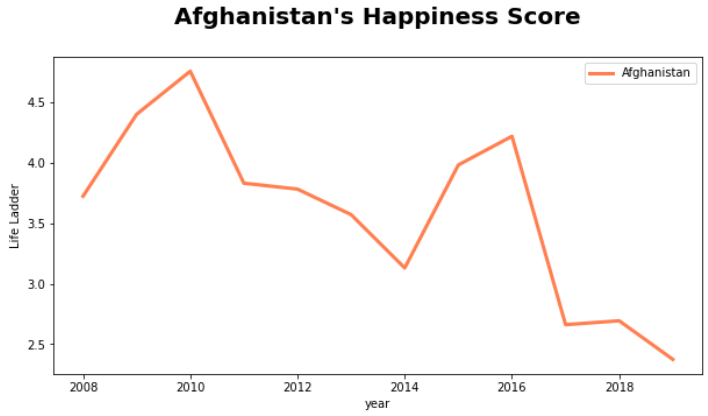
- Ireland's Happiness Score over the years, using Seaborn lineplot:
  As one can notice, Ireland's happiness scores have remained in a similar range of plus/minus 0.4 over the last 15 years. Ireland ranked 15th in the 2021 rankings.

- Bulgaria's Happiness Score over the years, using Seaborn lineplot:
  Bulgaria has improved its happiness score by over 1.5 points, which is the most by a country during the 2007 to 2020 period. Bulgaria ranked 88th in the 2021 rankings.

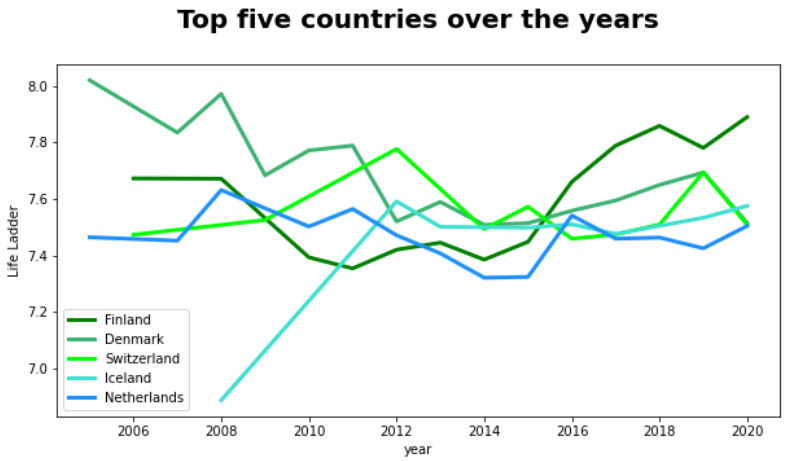


- Finland's Happiness Score over the years, using Seaborn lineplot:
  Finland is the happiest country in the world for 5 years now, getting a score of 7.842

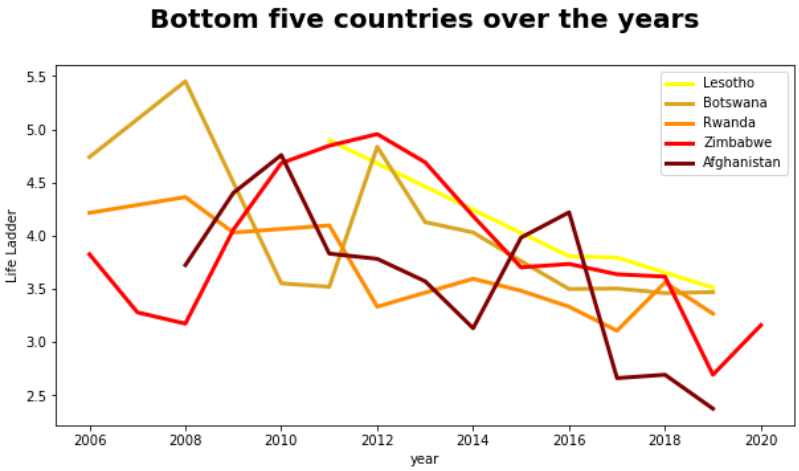
- Afghanistan's Happiness Score over the years, using Seaborn lineplot:
  Afghanistan has had a low ladder score for a few years now, as evident from the plot. This could be due to their unsteady political situation.



- Top 5 Country's Happiness Score over the years, using Seaborn lineplot:
  We analyse the top 5 countries based on the 2021 report over 15 years, we can see that apart from Iceland and Denmark, all the countries have maintained their ladder score values.



- Bottom 5 Country's Happiness Score over the years, using Seaborn lineplot:
  We analyse the top 5 countries based on the 2021 report over 15 years, we can see that Lesotho, Afghanistan and Botswana have had a decline in their ladder score values.



Insights:

The following are a few insights that I have gained by analysing and visualizing the data for the world happiness index scores from 2021 and over the years from 2006 to 2020.

- We see that there are many clear distinctions between happy and unhappy countries. Generally, happier countries tend to be wealthier, be less corrupt, be freer, have healthier life expectancy and have a lot of social support
- Something I noticed from the heatmap, was that generosity and GDP per capita are negatively correlated. This isn't surprising, but I was disappointed to see that wealthier nations tend to be less generous regardless.
- Bulgaria improved by over 1.5 points between 2007 and 2020, the most by any country. So, there is hope that unhappy countries can break their own historic trends.
- Finland was the happiest country in 2021 while Afghanistan was the unhappiest country during that same year.
- The top 5 happiest countries in 2021 are Finland, Denmark, Switzerland, Iceland and Netherlands in that order, and the bottom 5 countries in 2021 are Lesotho, Botswana, Rwanda, Zimbabwe and Afghanistan in that order. From this, we can see that, all the top 5 countries are from the European region, and on the other hand, 4 out of the bottom 5 countries are from the African region.
- Africa as a region also has the lowest region wise happiness score, social support, GDP per capita and healthy life expectancy.
- The features can be ranked based on importance in the following manner:
  GDP per capita – Life Expectancy – Social Support – Freedom of Choice – Perceptions of Corruption – Generosity

To wrap things up, this was an exploratory data analysis into why some countries are considered "happier" than others. We found that GDP per capita is the most important factor, which makes sense because money allows countries to afford luxuries along with basic resources. Despite all this, however, there are many more things that affect a country's happiness and there aren't models sophisticated enough to visualize them. To change this however, we can add machine learning in the form of neural networks to improve the data collection, grouping, analysis and visualization.

A good way to analyse the success of a machine learning network will be to check the answer to the question: "Is (Country Name), a Happy Country?" The answer would be a simple YES or NO.

The deep learning model can predict the Happiness Score of a country based on its historic trends and other factors such as GDP, Life expectancy, Social Support, etc. The output can be classification based to begin with, for example, the model could classify Ireland as a Happy country and Haiti as an Unhappy country. Once we have a basic machine learning model implemented, we can add regression/ back propagation to find specific values for happiness score or other factors.

Using the plotly.express library, one can also make a visualization on the map of the world, assigning colour scale to the happiest and unhappiest countries.

References:

1. https://worldhappiness.report/
2. https://www.gallup.com/home.aspx
3. https://www.kaggle.com/
4. https://numpy.org/doc/
5. https://pandas.pydata.org/docs/index.html
6. https://matplotlib.org/stable/plot_types/index.html
7. https://seaborn.pydata.org/examples/index.html
8. https://app.datacamp.com/learn/custom-tracks/custom-certificate-in-introductory-data-analytics-3b1e9b6a-346a-4086-9c5a-95daf66d4b6f