5CS027

Concepts and Technologies of AI

Exploratory Data Analysis of World Happiness dataset

Name: Rohit Khadka

Group: 15

Student ID: 2407942

Lecturer: Siman Giri

Tutor: Siman Giri

Tutor’s Signature

____

Analysis of the World Happiness Report: Exploring South Asia and Middle East Perspectives.

Introduction

World Happiness Report is a comprehensive report that numerically tries to describe happiness levels of people within a country based on various measurable life factors. The scores and rankings in the report are based on data from the Gallup World Poll. The dataset consists of columns as follows: country name, log GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and alongside a calculated baseline value called 'dystopia + residual'

Understanding the Attributes in the Dataset

1) **Country name:** Lists the name of the countries in the report

2) **Log GDP per capita**: Logarithmic value of GDP per capita of a country. Taking logarithmic helps to counter extremeness in the data.

3) **Social Support**: It measures the social support given by the country to its people.

4) **Healthy life expectancy**: It represents the number of years people can expect to live in good health.

5) **Freedom to make life choices**: This captures the implicit perception of freedom while making decisions regarding life without any outer intervention.

6) **Generosity**: It numerically describes the inclination of people towards charity, donation, and volunteerism.

7) **Perceptions of corruption**: This describes the perceptions of people regarding the existence of corruption, bribery, and other undutiful activities in the government.

8) **Dystopia + residual**: Dystopia represents a hypothetical country with lowest national averages that serves as a benchmark for comparison. The "residual" part accounts for any happiness factors not captured by the other metrics.

9) **Score**: It is a composite attribute that shows the overall happiness rate among people in the country.

Problem - 1: Getting Started with Data Exploration - Some Warm up Exercises:

In this section, I performed some basic data explorations operations with the help of some pre-built methods of pandas library.

1. Data Exploration and Understanding

Dataset Overview

1. I loaded the world happiness dataset with the help of read_csv() method**.**
2. Then, I used shape attribute of the pandas dataframe to check the number of rows and columns
3. Further, I used dtypes attribute to list all the columns and their data types

Basic Statistics

1. For the score column, I calculated mean, median, standard deviation using mean(), median(), and std() methods respectively.
2. Using max() and min(), I identified the countries with lowest and highest happiness scores.

Missing Values

1. Used isnull() method to check the null values and chained it with sum() method to calculate total number of null values in each column.

Filtering and Sorting

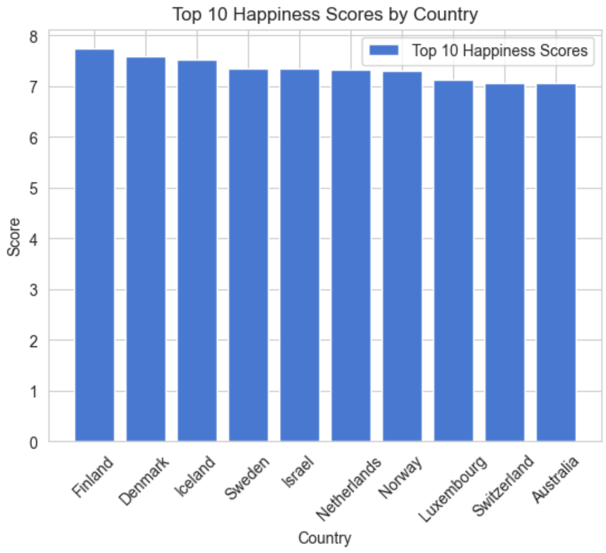
1. Used loc() method to display the countries with happiness greater than 6.5 only.
2. Used sort_values() method with arguments “ascending” set to false and “by” set to GDP column.

Adding New Columns

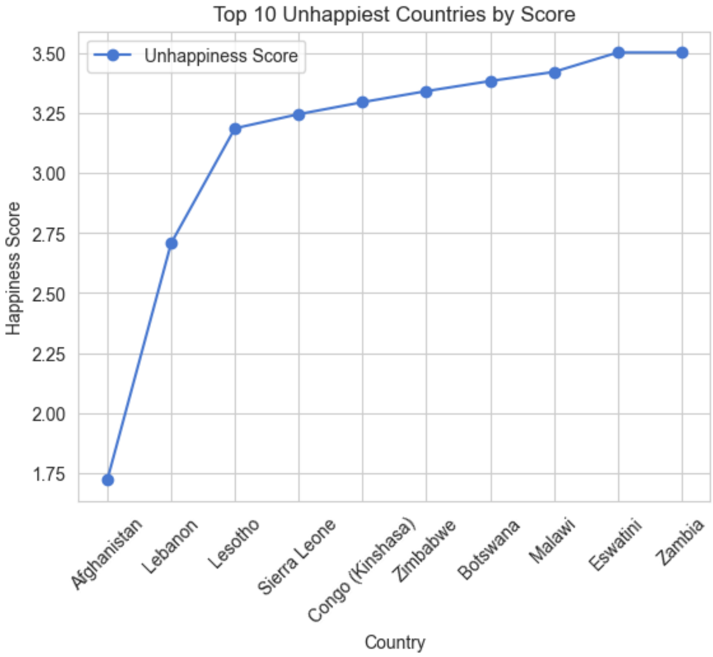
1. Used pd.cut() method to divide the scores into three labels low, medium, and high**.**

2. Data Visualizations:

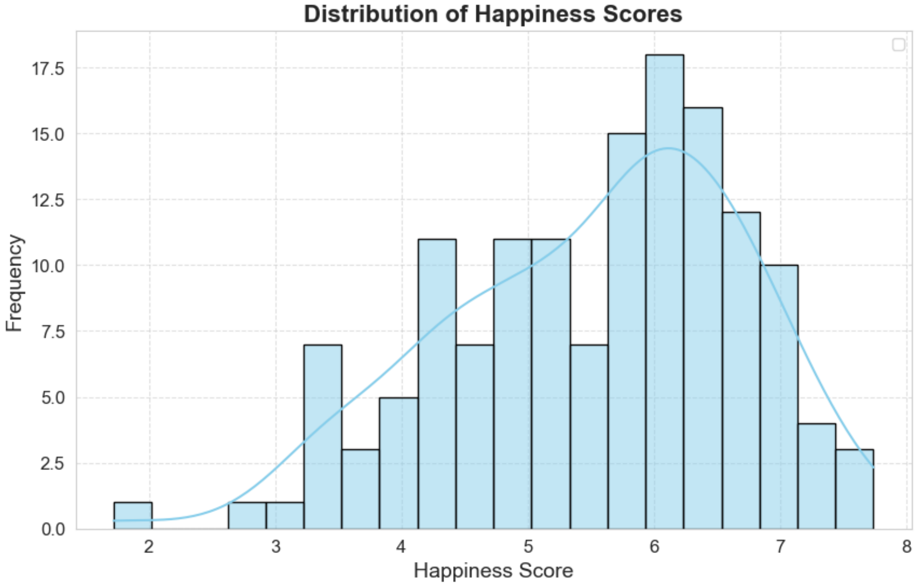
1. Plot of top 10 countries by Happiness score



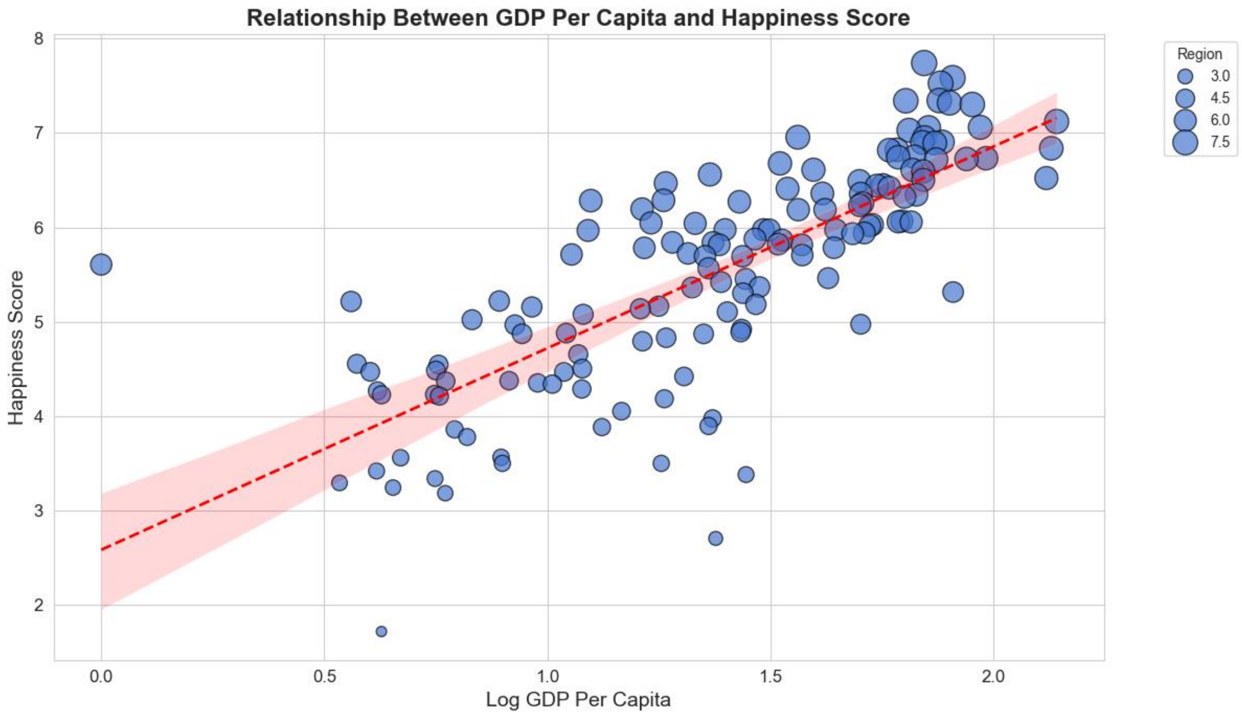
2) Line chart of below 10 countries by happiness score



3) Histogram for the Happiness score column



4) Scatterplot between GDP per Capita and Score



3.2 Problem - 2 - Some Advance Data Exploration Task: Task - 1 - Setup Task - Preparing the South-Asia Dataset:

Task - 1 - Setup Task - Preparing the South-Asia Dataset:

1. Defined a list with South Asian country names, and make a filtered dataset with south Asian countries using isin() method.
2. Saved the filtered dataset as csv file using to_csv() method.

Task - 2 - Composite Score Ranking:

1. Created a new column “Composite Score” using the following formula: Composite Score = 0.40 × GDP per Capita + 0.30 × Social Support + 0.30 × Healthy Life Expectancy
2. Ranked the south Asian countries using sort_values method in descending order based on Composite Score.

3.

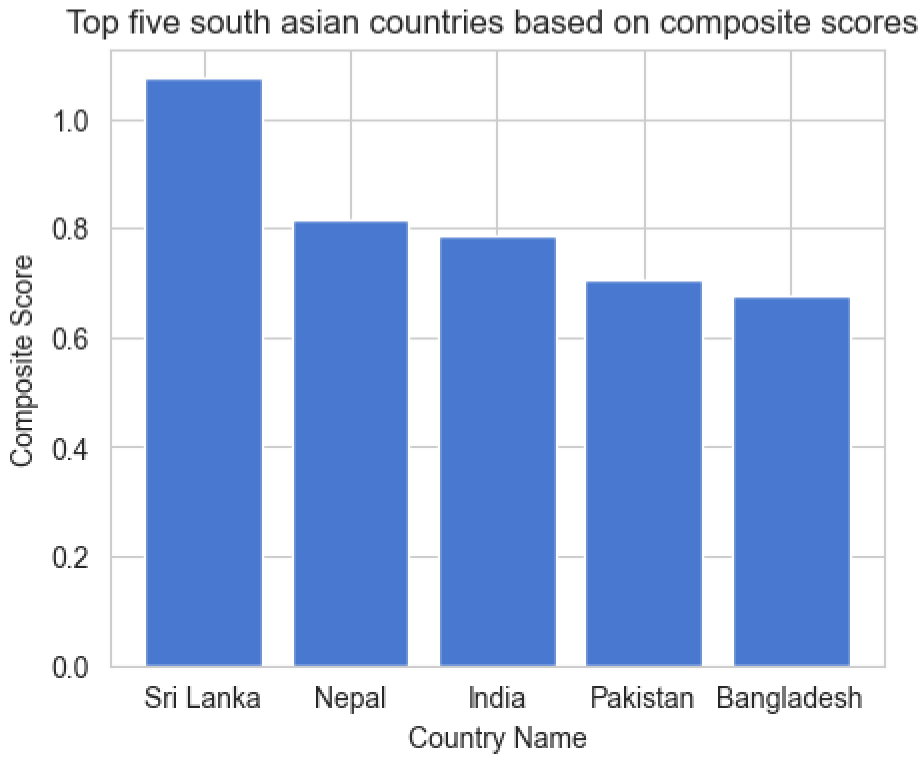
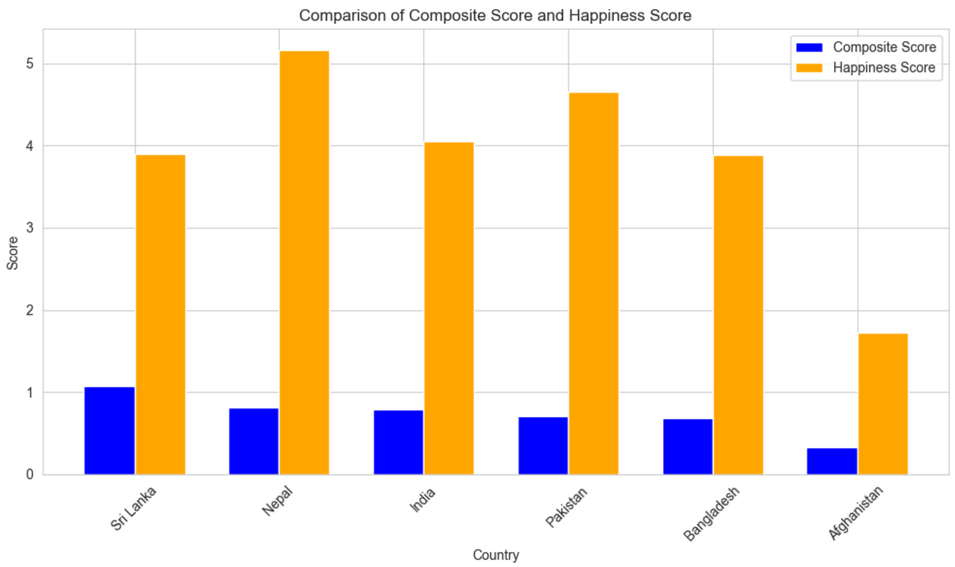


Figure: Visualization of Top 5 south Asian countries based on composite score.

4.



Based on the bar plots, we can infer that composite scores do not align with the original happiness scores.

Task - 3 - Outlier Detection:

1. Defined outliers in the score and Log GDP per capita column using the 1.5 * IQR rule.
2.

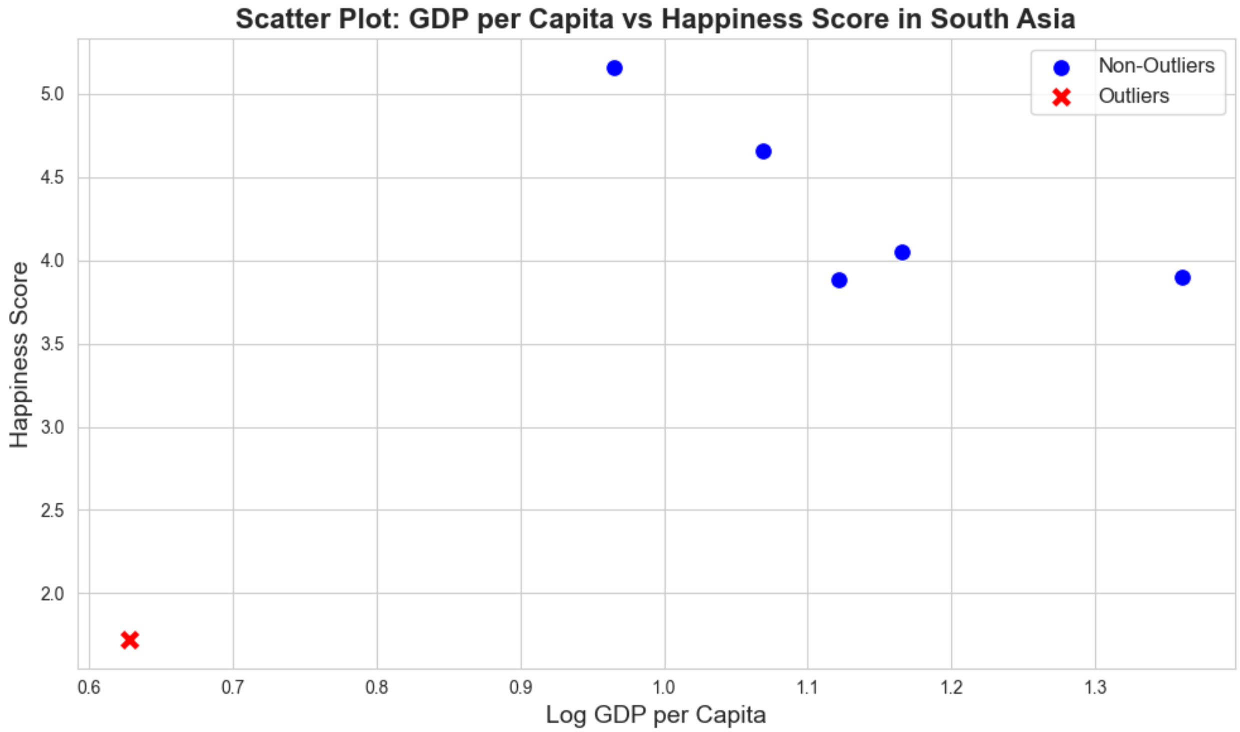


Figure: Scatter plot with Log GDP per capita on the x-axis and Happiness score on the y-axis with outliers represented by cross symbol.

Task - 4 - Exploring Trends Across Metrics:

1. Chose “Social support” and “Healthy life expectancy”, and calculated their pearson correlation coefficient using numpy’s corrcoef method.
2. Plotted scatterplots of these metrics against scores.

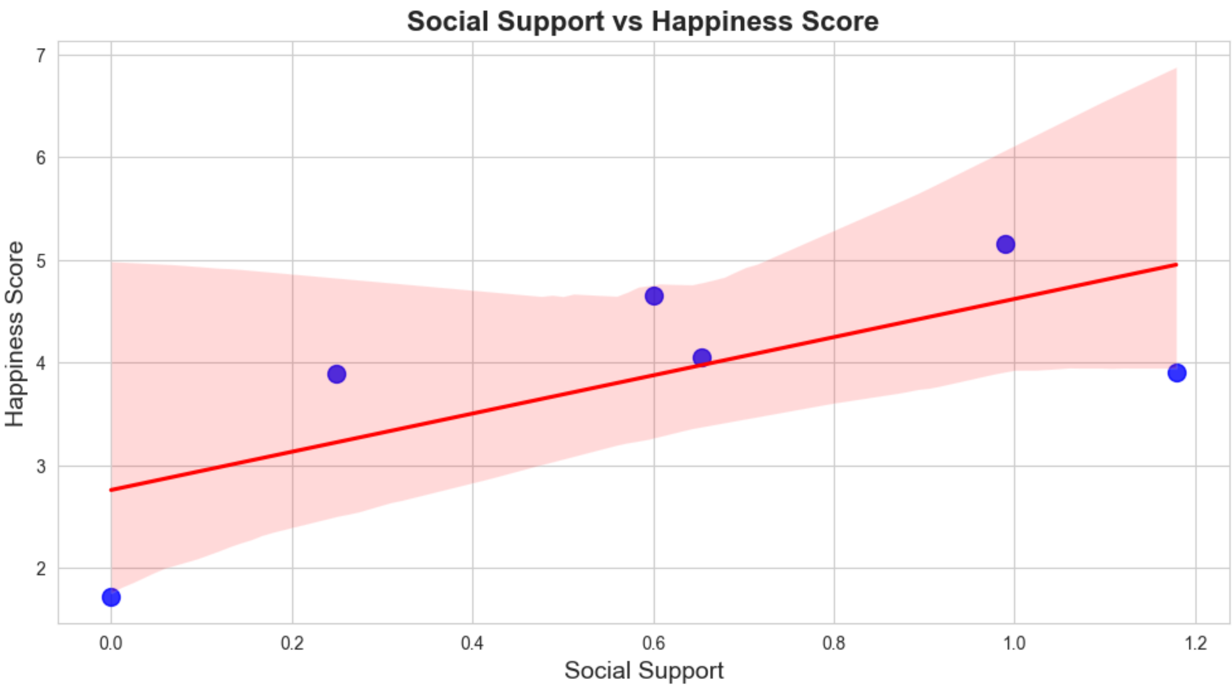


Figure: Scatterplot of Social Support against Happiness score

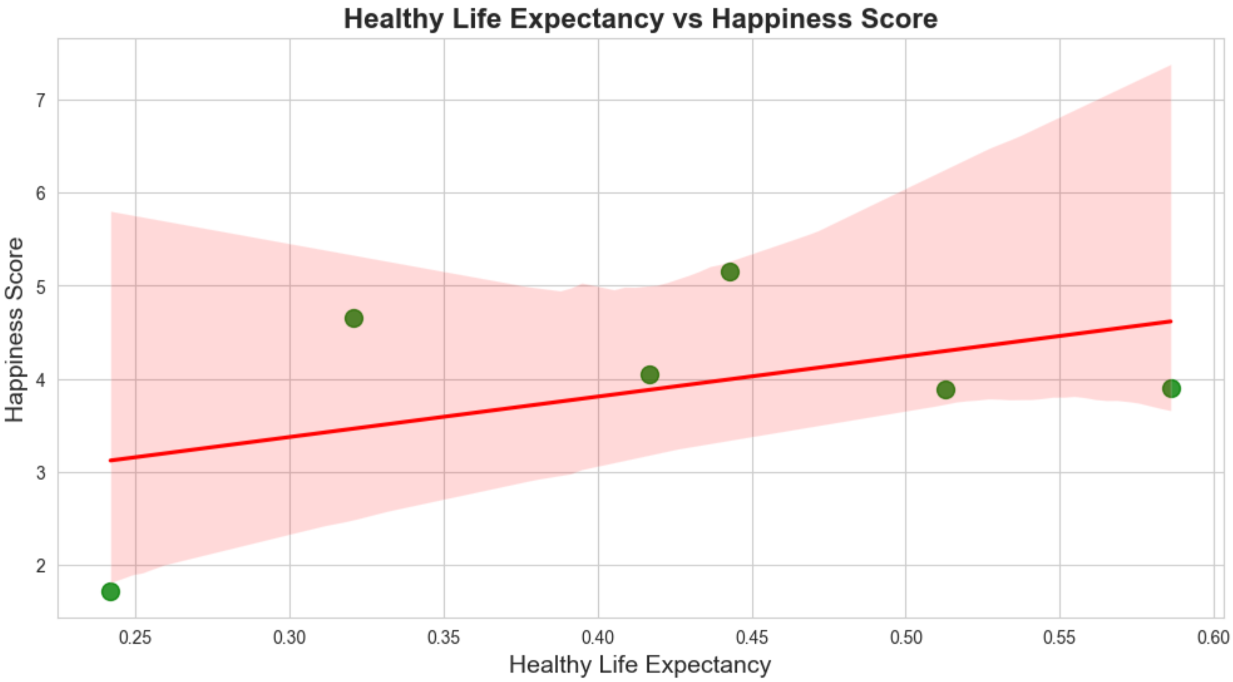


Figure: Scatterplot of Healthy Life Expectancy against Happiness score

Task - 5 - Gap Analysis:

1. Added a new column, “GDP-Score Gap” by computing the difference between GDP per Capita and the score.
2. Arranged south Asian countries as per the gap in both ascending and descending order.
3. Visualized positive gaps using a bar chart

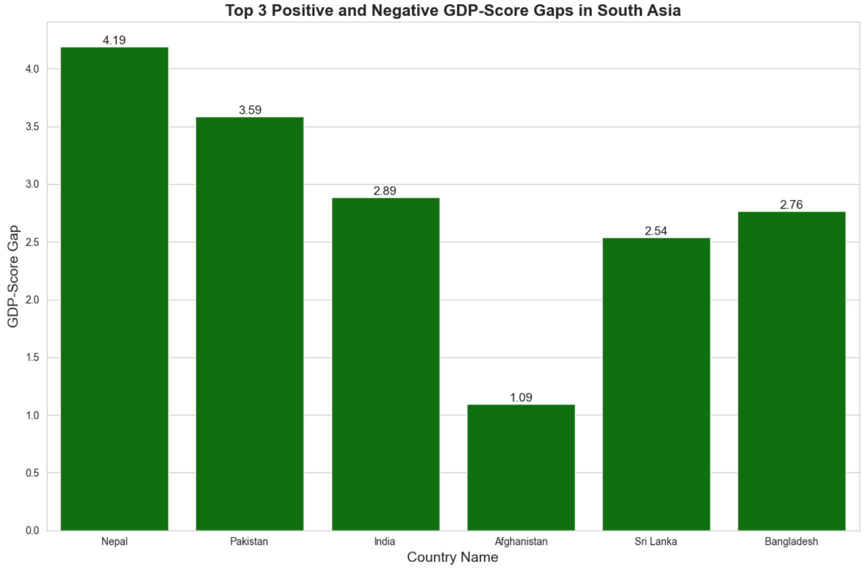


Figure: Gaps between GDP per capita and the happiness score in the south Asian countries

3.3 Problem - 3 - Comparative Analysis:

Created a list named middle_east_countries including names of middle east Asian countries ("Bahrain", "Iran", "Iraq", "Israel", "Jordan", "Kuwait", "Lebanon", "Oman", "Palestine", "Qatar", "Saudi Arabia", "Syria", "United Arab Emirates", "Yemen”)

Then, created a separate dataframe for the Middle East Asian countries in the same way I did for South Asian countries.

Tasks:
1) Descriptive Statistics
Compared the descriptive statistics measures such as mean, standard deviation for both south Asia and Middle East

Based on the analysis, middle east countries have generally higher happiness level with mean happiness score of 5.41 whereas south Asian countries have mean happiness score of 3.89

2) Top and Bottom Performers:

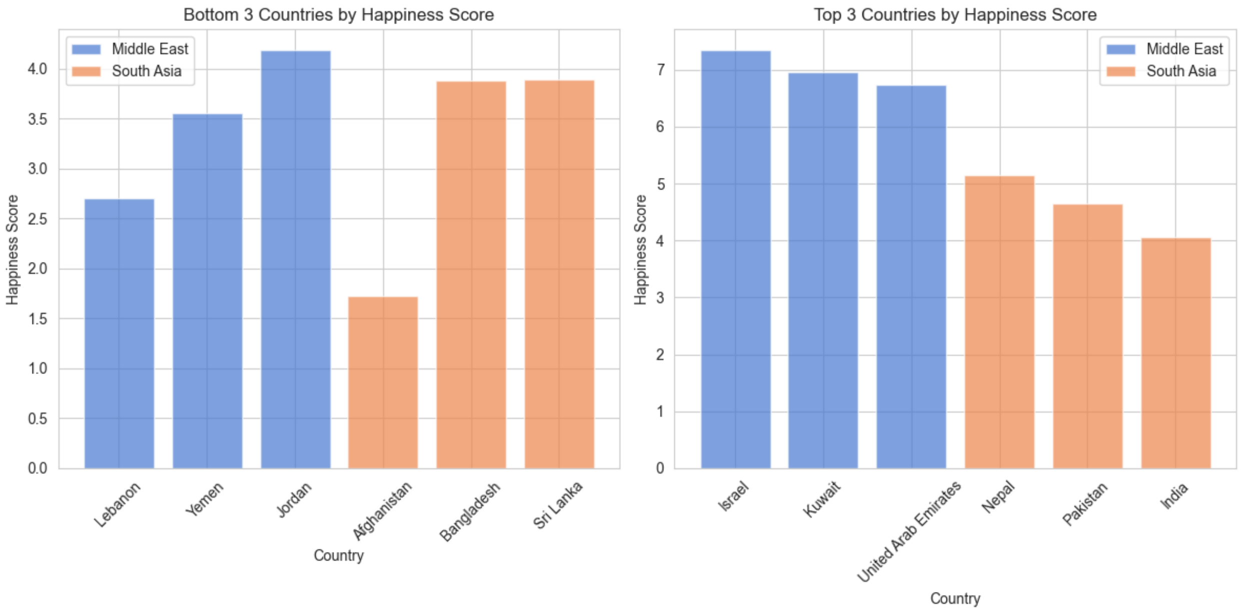


Figure: Bar plots showing top 3 and bottom 3 countries in each region based on happiness score.

3) Metric Comparisons

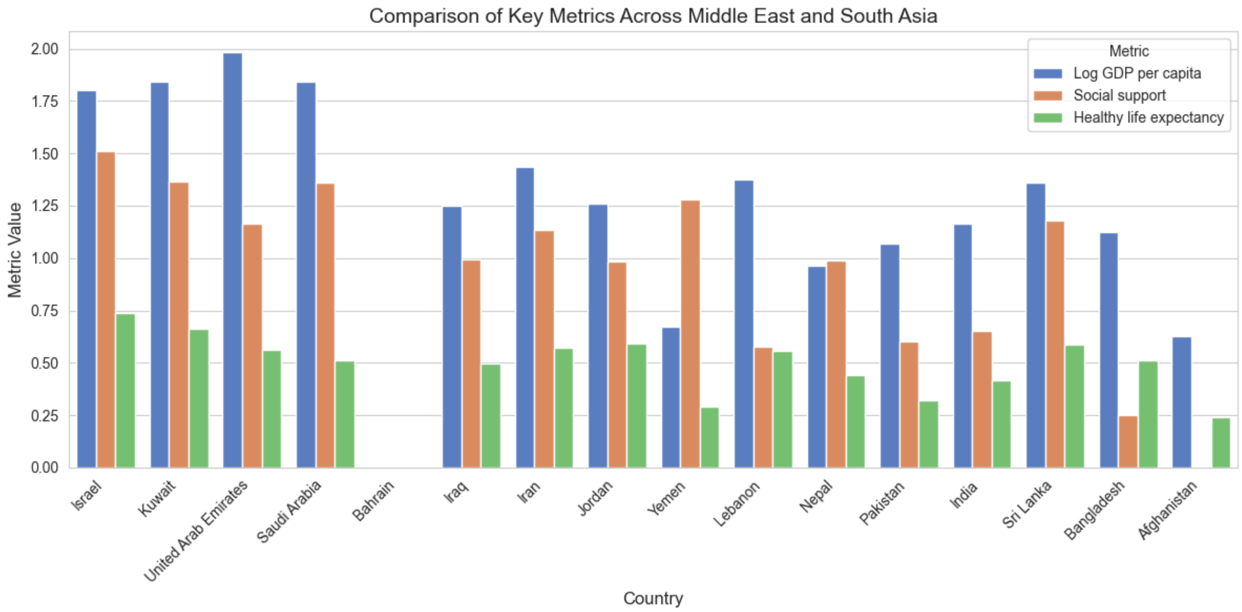


Figure: Grouped Bar charts showcasing the comparison between two region based on GDP, Social Support, and Healthy Life Expectancy

Among the three key metrics, Log GDP per capita shows the largest disparity between two regions.

4) Happiness Disparity

Computed range and coefficient of variation for score in both regions to measure the Spreadness among the data in both regions.

From the computation we can infer that middle east region has greater variability in Happiness score with both coefficient of variation and range being slightly greater than that of south Asian region.

5) Correlation Analysis

Firstly, calculated the correlation of score with other metrics: Freedom to make life choices and generosity within both each region. Then plotted a heatmap and scatterplots to visualize the relationships among metrics.

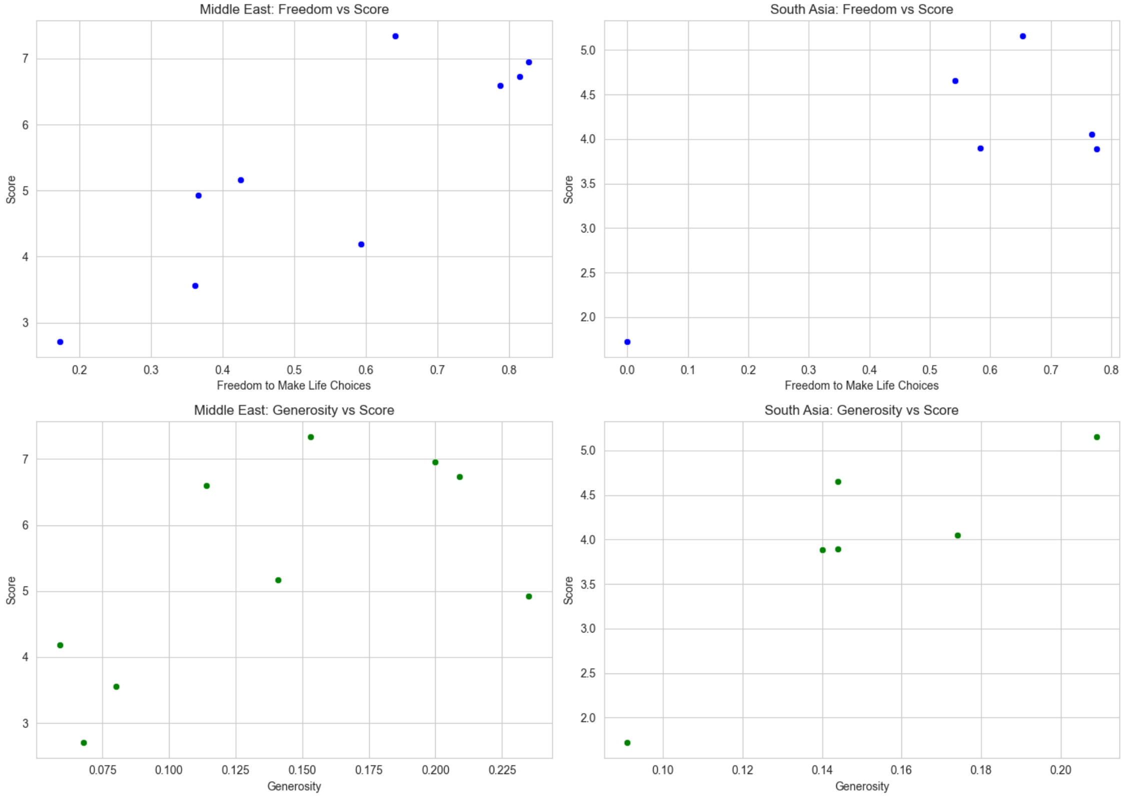


Figure: Scatterplots of score against freedom to make life choices and generosity.

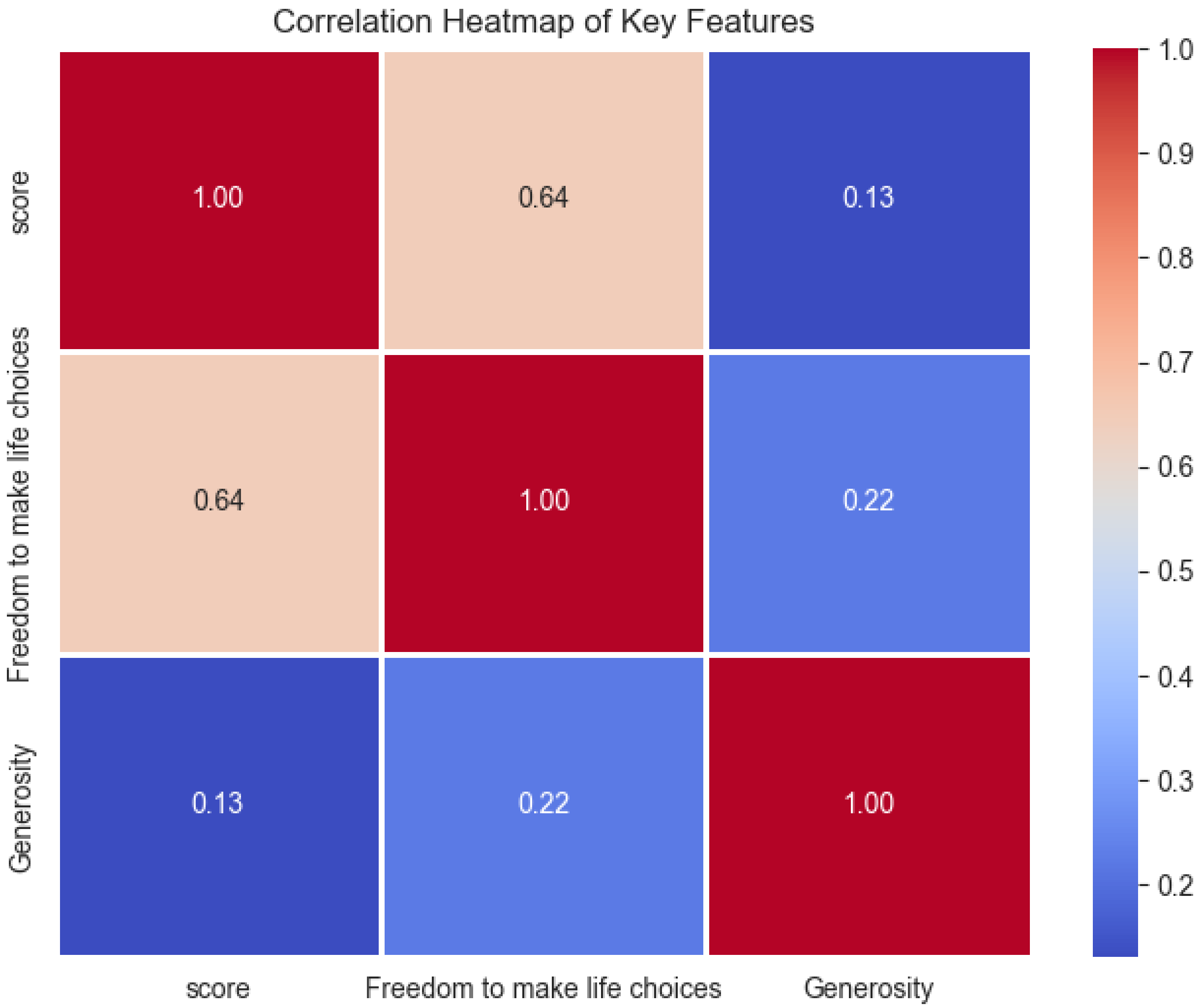


Figure: Heatmap annotating the correlation and variance score of happiness scores against freedom to make life choices and generosity.

6) Outlier Detection

Used 1.5 *IQR rule to detect outliers in the score and GDP per capita section in both regions.

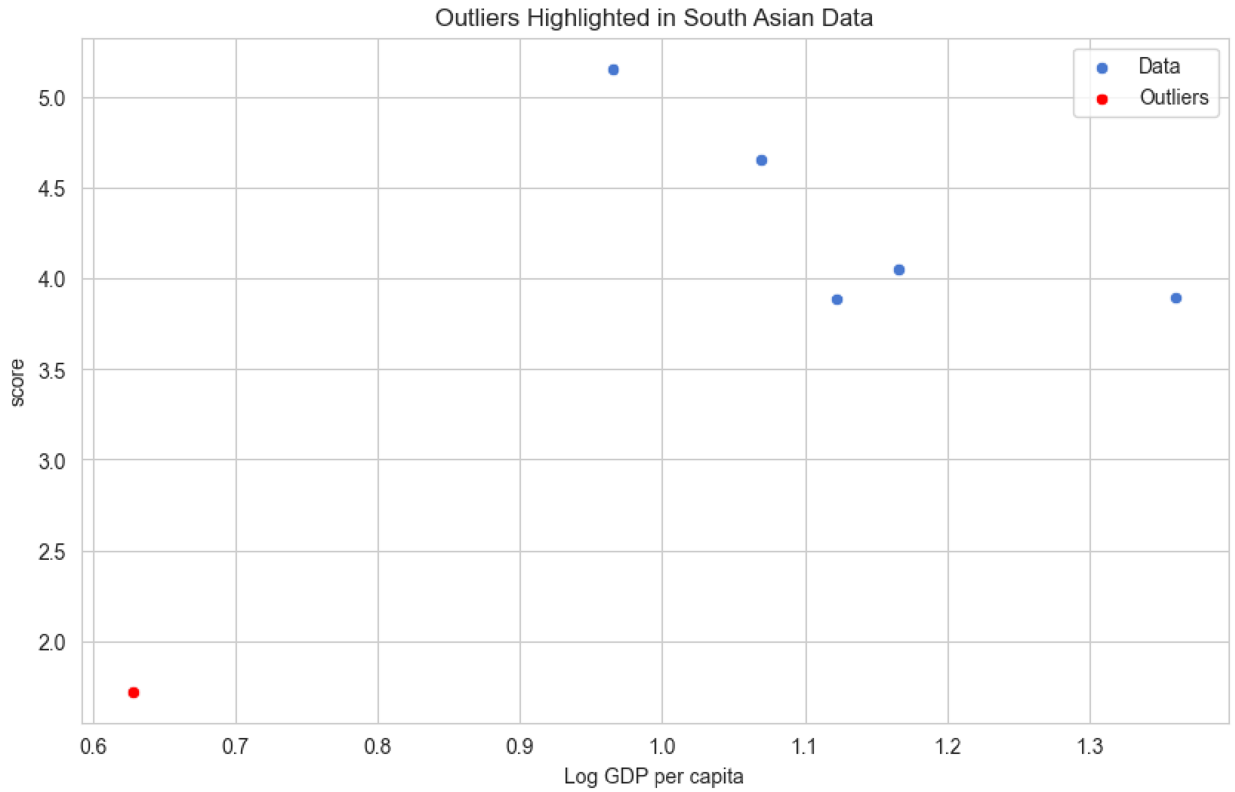
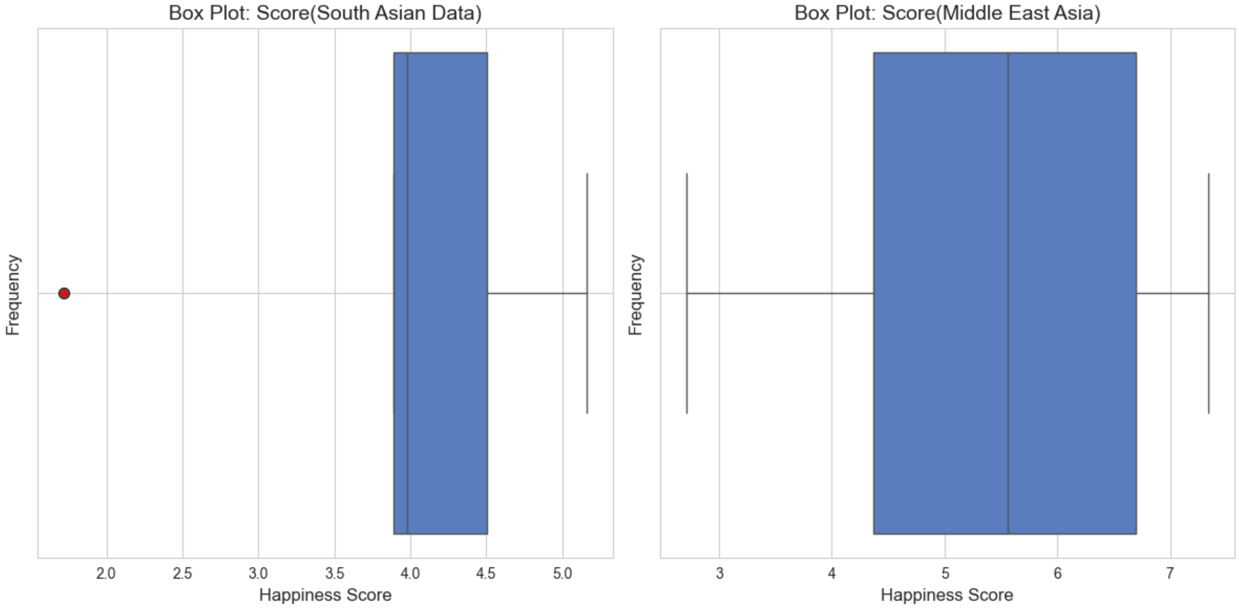


Figure: Plot representing outlier(red) and non-outlier data (blue)

7) Visualization

Created boxplots comparing the distribution of score between South Asia and Middle East.



Based on the preliminary analysis, we can infer that distribution of south Asian data has more symmetricity, but the statistical metrics such as mean and median of middle east region are greater than that of south Asian. South Asian data has one outlier whereas middle east region does not have any.

Conclusion

I used various built in functions of pandas library in order to load the dataset and explore the trends and patterns within it. I used statistical measures such as mean and median to infer description of the dataset.

Talking about the first problem, I generally performed basic data exploration operations. I checked for null values, outliers, visualized top and bottom 10 countries as per the happiness score.

Secondly, I generally did bivariate analysis. I created a new column “Composite Score” and saw the trends that newly created feature and score followed.

Thirdly, I performed comparative analysis based on two main regions: south Asian and middle east Asian. After performing all the exploratory analysis, it can be inferred that people in Middle East Asian countries are generally happier than in the South Asian countries.