UNIVERSITY OF WOLVERHAMPTON

5CS037
Concepts and Technologies of AI

Analysis of the World Happiness Report:

A Data-Driven Exploration of Global and Regional Trends

Name: Sukriya Shrestha

Group: L5CG5

Canvas ID: 2407970

Lecturer: Prakriti Regmi

Tutor: Ronit Shrestha

Module Leader: Siman Giri

Submission Date : 20th Dec, 2024

Contents

An Overview of the World Happiness Report

World Happiness Report's main purpose is to analyze and interpret the statistical data of 144 countries by measuring, visually plotting, and considering different factors. It uses different tools to properly inspect the data by mainly emphasizing the South Asia and Middle East Region. Throughout the context, the report is objectified through different plots, diagrams, and other outputs calculated from different operation patterns like correlations, mean, standard deviation, and by identifying outliers with the help of different statistical methods.

Different Libraries and their roles in the Report Analysis

In data analysis, utilizing different libraries and tools is very essential to perform certain operations. The commonly used open source libraries in determining the report contains the following commands:

- import numpy as np(NumPy: used for numerical computing and array handling)
- import pandas as pd(Pandas: used for data manipulation)
- import matplotlib.pyplot as plt(Matplotlib: used to create static and interactive visualization)

Datasets Used in the Report

The datasets used in the World Happiness Report comprises 143 rows and 9 columns:

- Country Name
- Score(Happiness Score)
- Log GDP per Capita
- Social Support
- Healthy Life Expectancy
- Freedom to make life choices
- Generosity
- Perceptions of Corruption
- Dystopia + Residual

```
[141] df
```

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corr
0	Finland	7.741	1.844	1.572	0.695	0.859	0.142	
1	Denmark	7.583	1.908	1.520	0.699	0.823	0.204	
2	Iceland	7.525	1.881	1.617	0.718	0.819	0.258	
3	Sweden	7.344	1.878	1.501	0.724	0.838	0.221	
4	Israel	7.341	1.803	1.513	0.740	0.641	0.153	
...	...	...	...	...	...	...	...	
138	Congo (Kinshasa)	3.295	0.534	0.665	0.262	0.473	0.189	
139	Sierra Leone	3.245	0.654	0.566	0.253	0.469	0.181	
140	Lesotho	3.186	0.771	0.851	0.000	0.523	0.082	
141	Lebanon	2.707	1.377	0.577	0.556	0.173	0.068	
142	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	

143 rows × 9 columns

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

Data Exploration and Understanding

Dataset Overview

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
[140] df = pd.read_csv('/content/drive/MyDrive/Datasets/WHR-2024-5CS037.csv')
```

```
[141] df
```

Necessary libraries were imported in the first step for further calculations and the data from the data frame was converted to a csv file through the function ‘read_csv()’. Likewise, the data was printed out using the initialized variable name.

```
print("First ten rows of dataframe: \n")
df.head(10)
```

```
no_rows, no_column = df.shape
print("Number of rows: ", no_rows)
print("Number of column: ", no_column)
```

```
print("List of columns with their datatypes: \n")
df.dtypes
```

Similarly, the function of 'head()' is to read the top rows of the dataset therefore 10 rows were displayed in the output. Shape attribute is also used to show the number of rows and columns of the dataset and dtypes method for showing the data type of any object.

Basic Statistics

```
[ ] mean_of_score = df['score'].mean()
    print("Mean of score column: ", mean_of_score)

Mean of score column:  5.52758041958042

[ ] median_of_score = df['score'].median()
    print("Median of score column: ", median_of_score)

Median of score column:  5.785

[ ] standard_deviation = df['score'].std()
    print("Standard deviation of score column: ", standard_deviation)
```

To identify the mean, median and standard deviation of the data, panda's default methods: mean(), median() and std() were used.

```
[148] # country with highest happiness score
      highest_score_index = df['score'].idxmax()
      country_highest_score = df.loc[highest_score_index, 'Country name']

      print("the country with highest Happiness score is: \n",country_highest_score)

      the country with highest Happiness score is:
       Finland

[150] #country with lowest happiness score
      lowest_score_index = df['score'].idxmin()
      country_lowest_score = df.loc[lowest_score_index, 'Country name']

[151] print("The country with lowest Happiness score is: \n", country_lowest_score)
```

Moreover, to calculate the country with highest and lowest Happiness Score, idxmax() and idxmin() methods were used.

Missing Values

```
miss_value = df.isnull()
miss_value
```

```
total_count = df.isnull().sum()
total_count
```

To determine the missing values in the data frame as there was many non-given values, method like isnull() was used to find the missing values in the data set and sum() method to count the total number of elements in the column by excluding the NaNs automatically.

Filtering and Sorting

```
print("countries with greater score than 7.5")
fltr_df = df[(df['score'] > 7.5)]
fltr_df
```

```
sorted_data = fltr_df.sort_values(by = 'Log GDP per capita', ascending = False)
sorted_data.head(10)
```

The given condition was to sort and filter the value of score greater than 7.5 which we obtained through accessing the ‘score’ column and giving the term ‘less than 7.5’. Later, it was sorted using the sort_values() method with descending order. Only 10 rows were exhibited through the head() method.

Adding new Columns:

```
df['Happiness_Category'] = np.where(df['score'] < 4, 'Low', np.where((df['score'] >= 4) & (df['score'] <
```

A new column was added to the existing data set through the np.where() method from the NumPy library with certain condition to determine Low, Medium and High values.

2. Data Visualizations:

It implemented different methods to plot all the data to form diagrams and analyze it using matplotlib library.

Bar Plot:

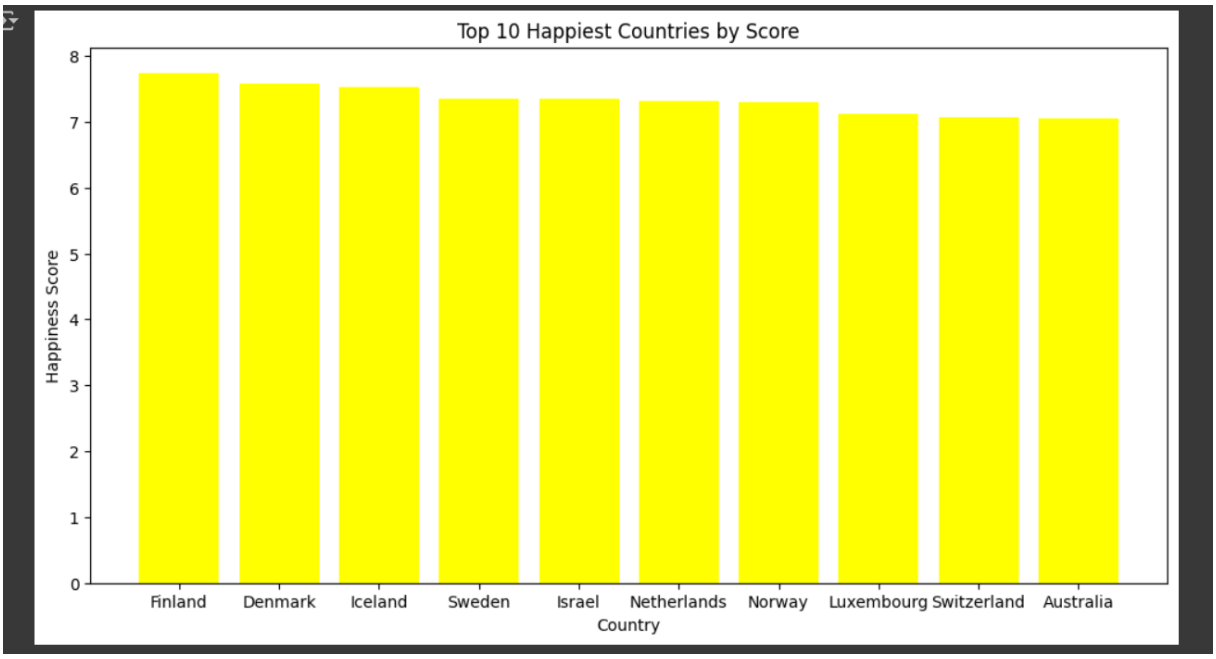


Fig1: Bar Chart to show top 10 Happiest Countries by Score

Line Plot:

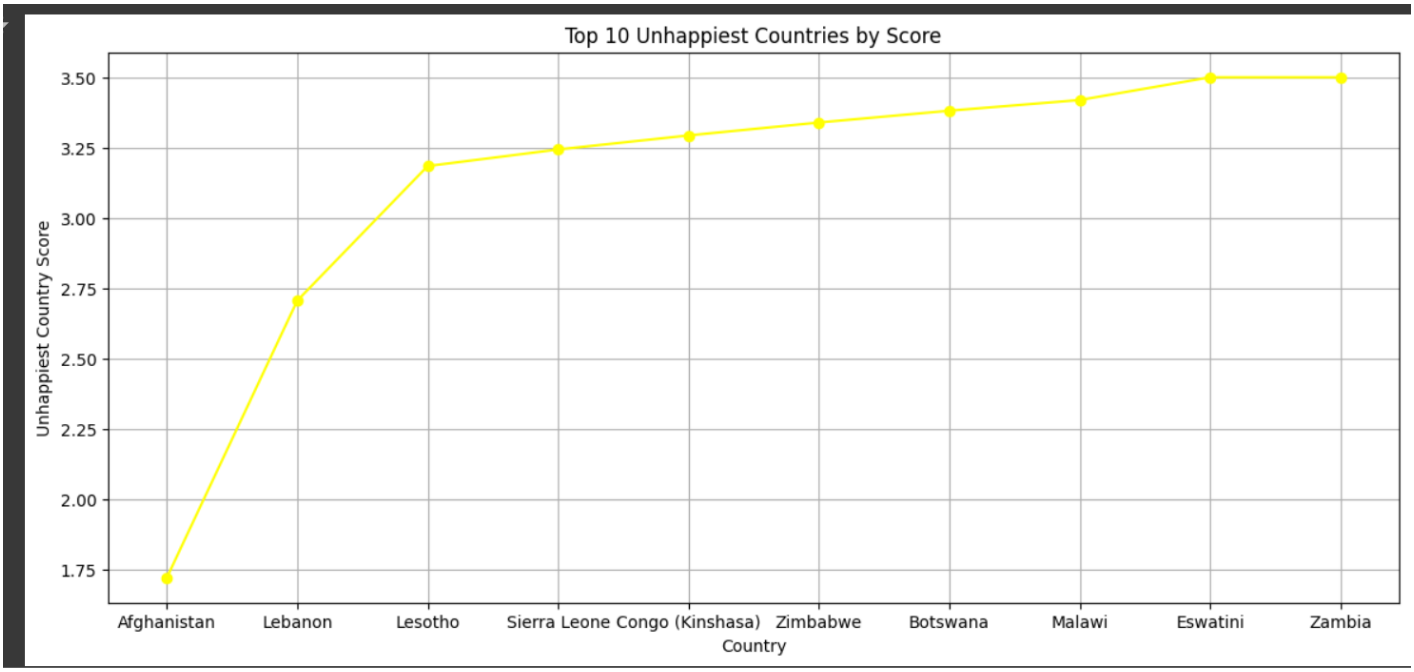


Fig2: Line Chart of top 10 unhappiest countries by Score

Histogram

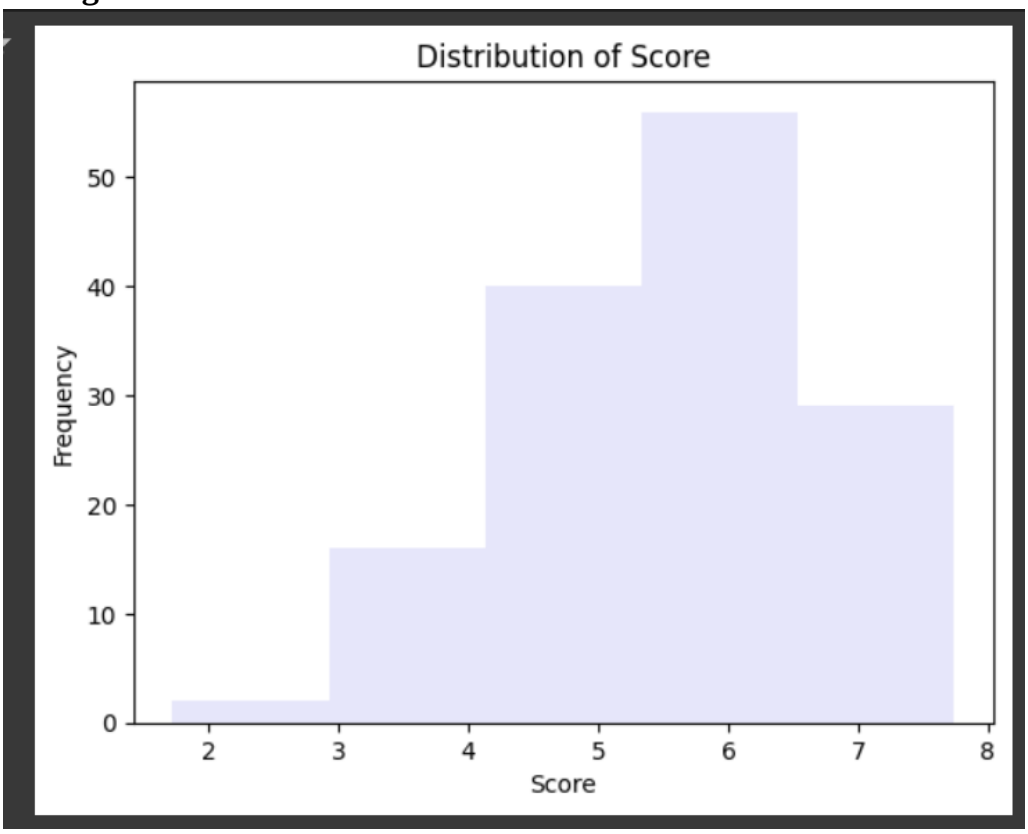


Fig3: Histogram diagram to show Distribution

The histogram plot shows that the majority of observation clustering ranges to 6 score which is the highest frequency above 60 for happiness score and the lowest scores observed in the lower range(2 -3).

Scatter Plot

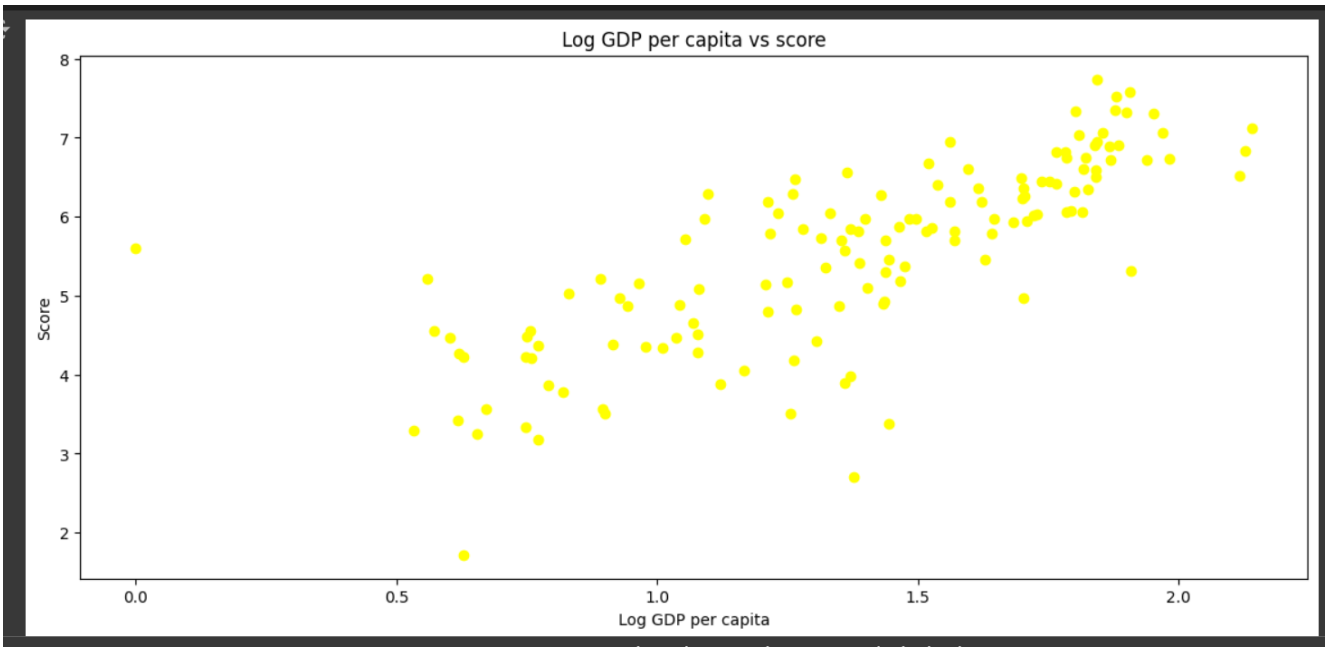


Fig4: Scatter plot to visualize between GDP per Capita and Score

We observe that both Happiness score and Log GDP per capita increases so there exists a positive relation between two variables. This indicates that higher economic condition gives higher happiness.

3.2 Problem- 2- Some Advance Data Exploration

Task1: Setup Task- Preparing the South-Asia Dataset:

```
[164] south_asian_countries = ["Afghanistan", "Bangladesh", "Bhutan", "India", "Maldives", "Nepal", "Pakistan", "Sri Lanka"]
```

2. Use the list from step- 1 to filtered the dataset {i.e. filtered out matching dataset from list.}

```
data = {
    'Country name': ['Afghanistan', 'Bangladesh', 'Bhutan', 'India', 'Maldives', 'Nepal', 'Pakistan', 'Sri Lanka']
}
df_south_asian = pd.DataFrame(data)
#filtered the dataframe
filtered_df = df[df['Country name'].isin(south_asian_countries)]
filtered_df
```

```
filtered_df.to_csv('South Asian countries.csv')
```

A list of South Asian Countries was already provided. The list values were added as key and values and stored in a new variable. Using pd.DataFrame() method the dictionary was converted to a data frame. Moreover, isin() function filtered the similar specified columns from the provided csv file and displayed it. Using to_csv() the data frame was converted to a csv file.

Task- 2- Composite Score Ranking

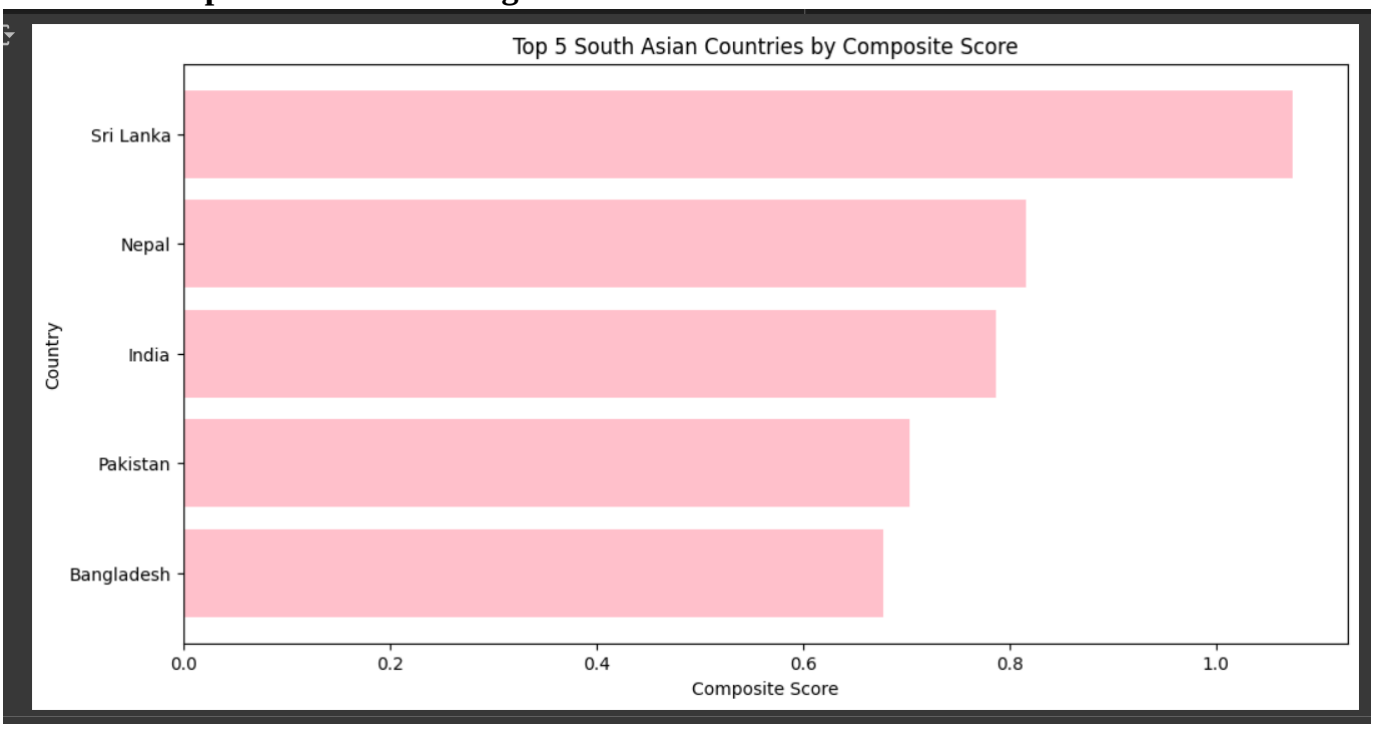


Fig: Composite Score of Top 5 countries with Bar Chart

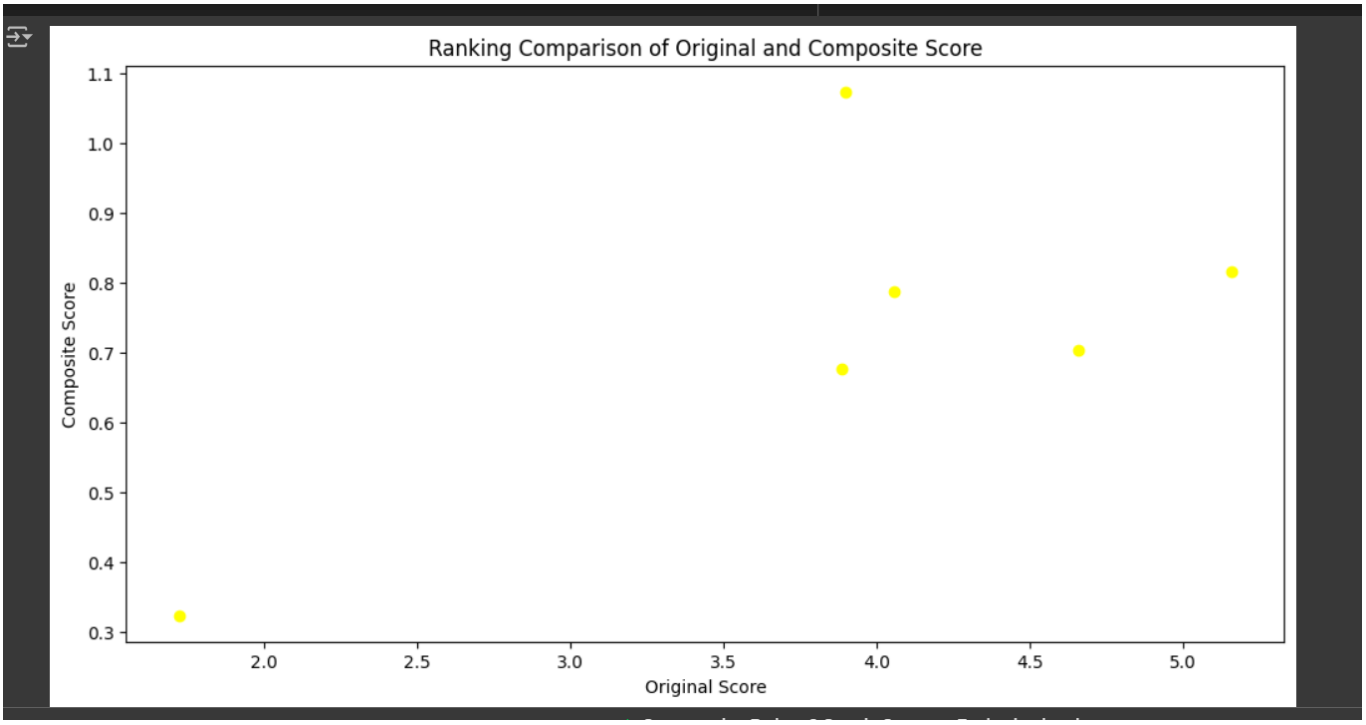


Fig: Ranking Comparison Original and Composite Score through scatter plot

A new Column ‘Composite Score’ was created by combining the given metrics:

Composite Score = 0.40×GDP per Capita+0.30×Social Support + 0.30 ×Healthy Life Expectancy

Similarly, the rank of south Asian countries was determined based on the Composite Score in descending order and again saved to the csv file.

The scatter plot proposes that there exists a general positive correlation between the Original and Composite Score but it doesn't perfectly align with the original scores.

Task- 3- Outlier Detection

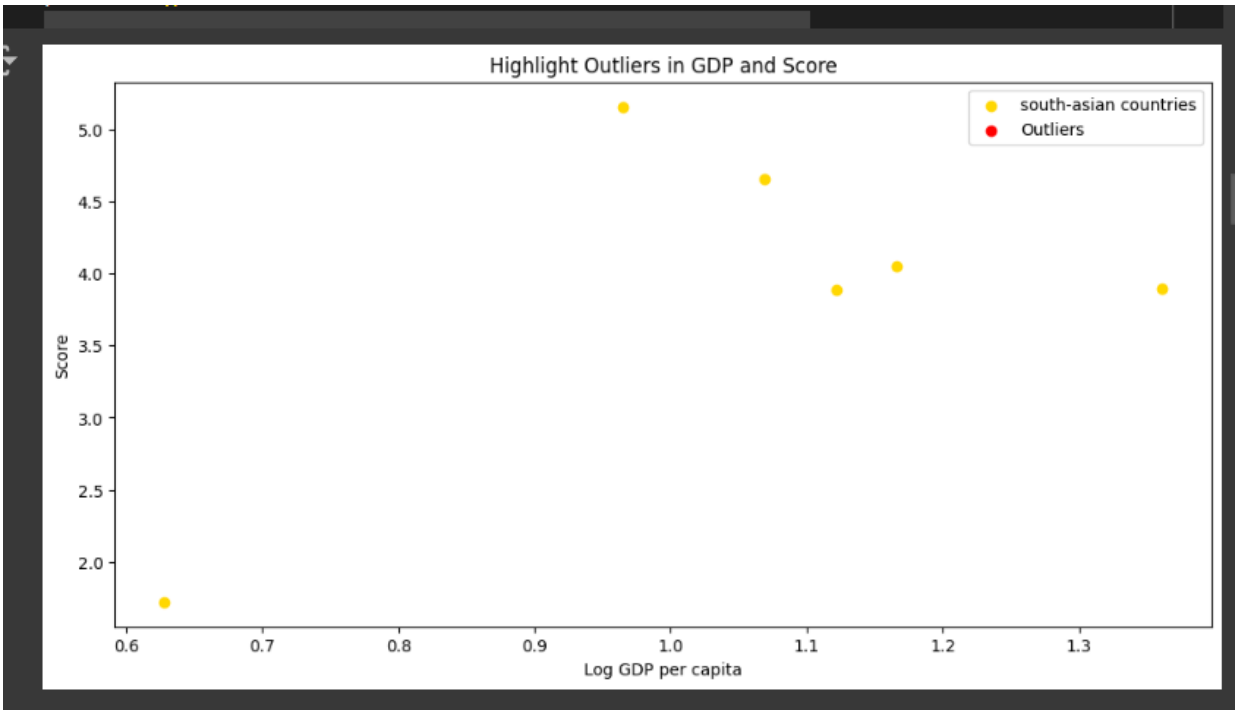


Fig: Scatter Plot of Outliers and Original data distribution

The Scatter plot illustrates the relation between Log GDP per capita on the x- axis and Happiness Score on the y-axis. In the context of Score and Log GDP per capita, outliers show the out of range value from the normal distribution. The plot indicated the absence of outliers as all the predictable data points falls within the range. In addition, the majority of the Yellow dots occurring together estimates that there is a positive correlation between GDP per capita and the Happiness Score. Hence, when GDP per capita increases, the happiness score also tend to incline.

Task- 4- Exploring Trends Across Metrics

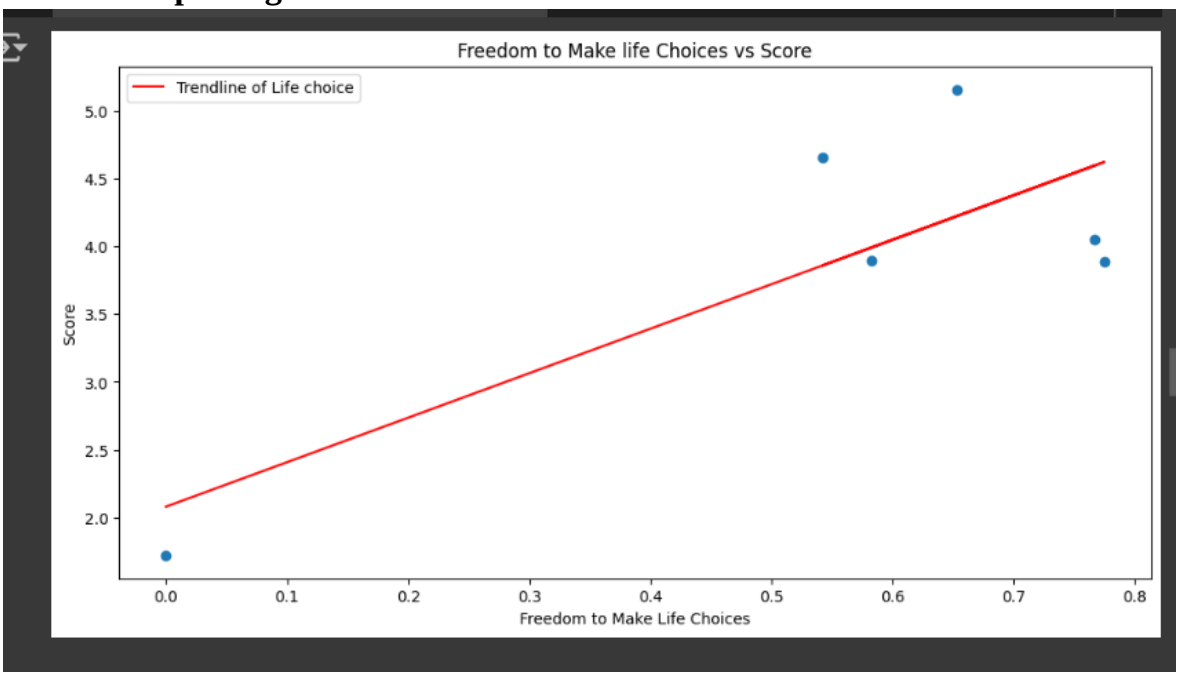


Fig: Scatter plots with trendlines with freedom vs Score

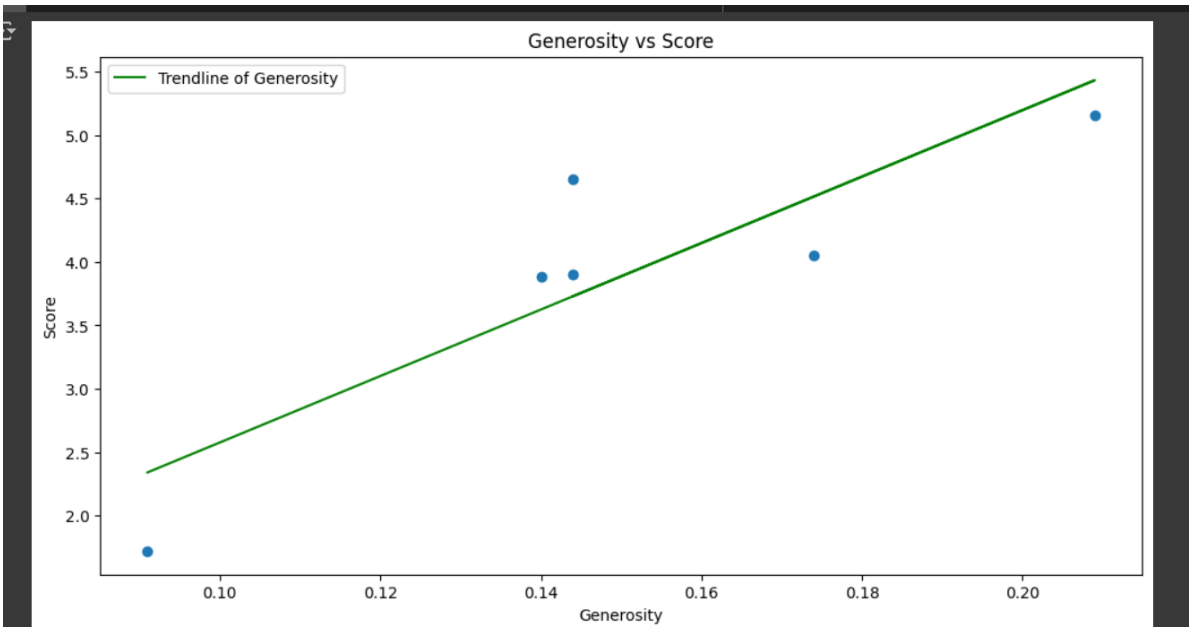
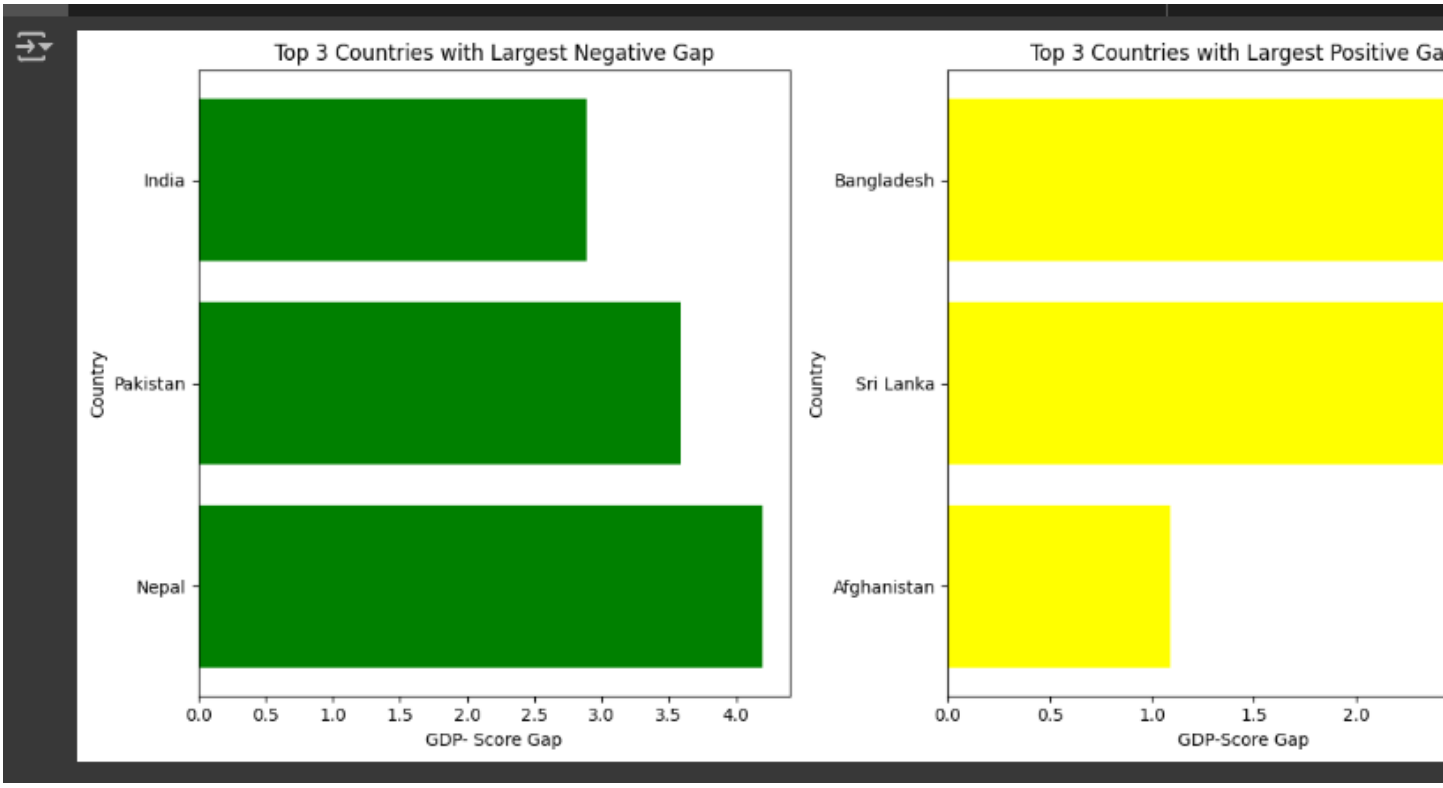


Fig: Scatter plot with trendlines with Generosity vs Score

The above two figures demonstrate the relationship between Generosity and Freedom to make life choices with respect to score. By calculating their correlation coefficient using corrcoef(), we can understand the correlation with Pearson's correlation coefficient which also teaches us about linear regression between two metrics.

Additionally, we can interpret that Freedom to make choice contains strong link with score meaning people of countries with freedom are tend to have higher happiness Score and vice versa.

Task- 5- Gap Analysis:



To calculate the gap, we need to find the difference between existing Log GDP per capita and score and display the data. The data are sorted in both ascending and descending order in a new column called GDP-score gap.

According to the first bar graph, India has the largest negative GDP – score gap, followed by Pakistan and Nepal. Even though they have lower GDP rate they seem happy. Whereas countries like Bangladesh, Sri Lanka and Afghanistan has happiness level according to their highest GDP rate.

3.3 Problem- 3- Comparative Analysis:

Task- 1- Setup Task- Preparing the Middle Eastern Dataset

```
52] middle_east_countries =  [ "Bahrain", "Iran", "Iraq", "Israel", "Jordan",
    "Kuwait", "Lebanon", "Oman", "Palestine", "Qatar", "Saudi Arabia", "Syria",
    "United Arab Emirates", "Yemen"]

53] data1 = {
        'Country name': ['Bahrain', 'Iran', 'Iraq', 'Israel', 'Jordan', 'Kuwait', 'Lebanon',
    }
    md_df = pd.DataFrame(data1)
    # filtering the data
    fr_df = df[df['Country name'].isin(middle_east_countries)]
    fr_df
```

From the given list of middle east countries I created a dictionary and converted it to data frame using pd.DataFrame() method and checked if the data and values are inside the data through isin() function as done above. I created a csv file of this data frame using to_csv() .

1. Descriptive Statistics

After calculating the mean and standard deviation of both regions(South Asia and Middle East). The average mean of Middle East Region was higher than that of South Asia. Therefore, Middle East has higher happiness score than South Asia indicating people are happier in Middle East countries.

2.Top and Bottom Performers:

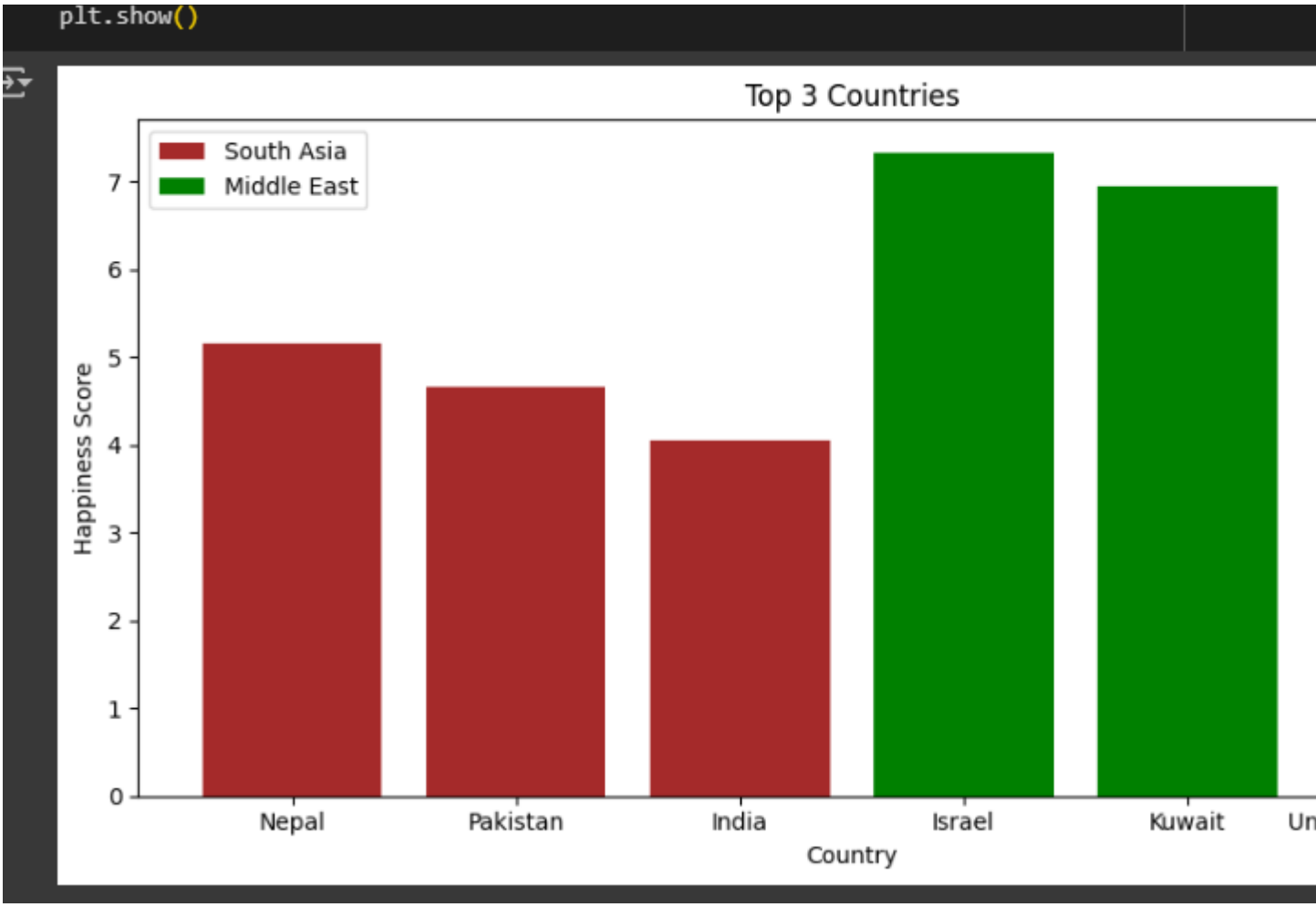


Fig: Top 3 Countries of both Regions

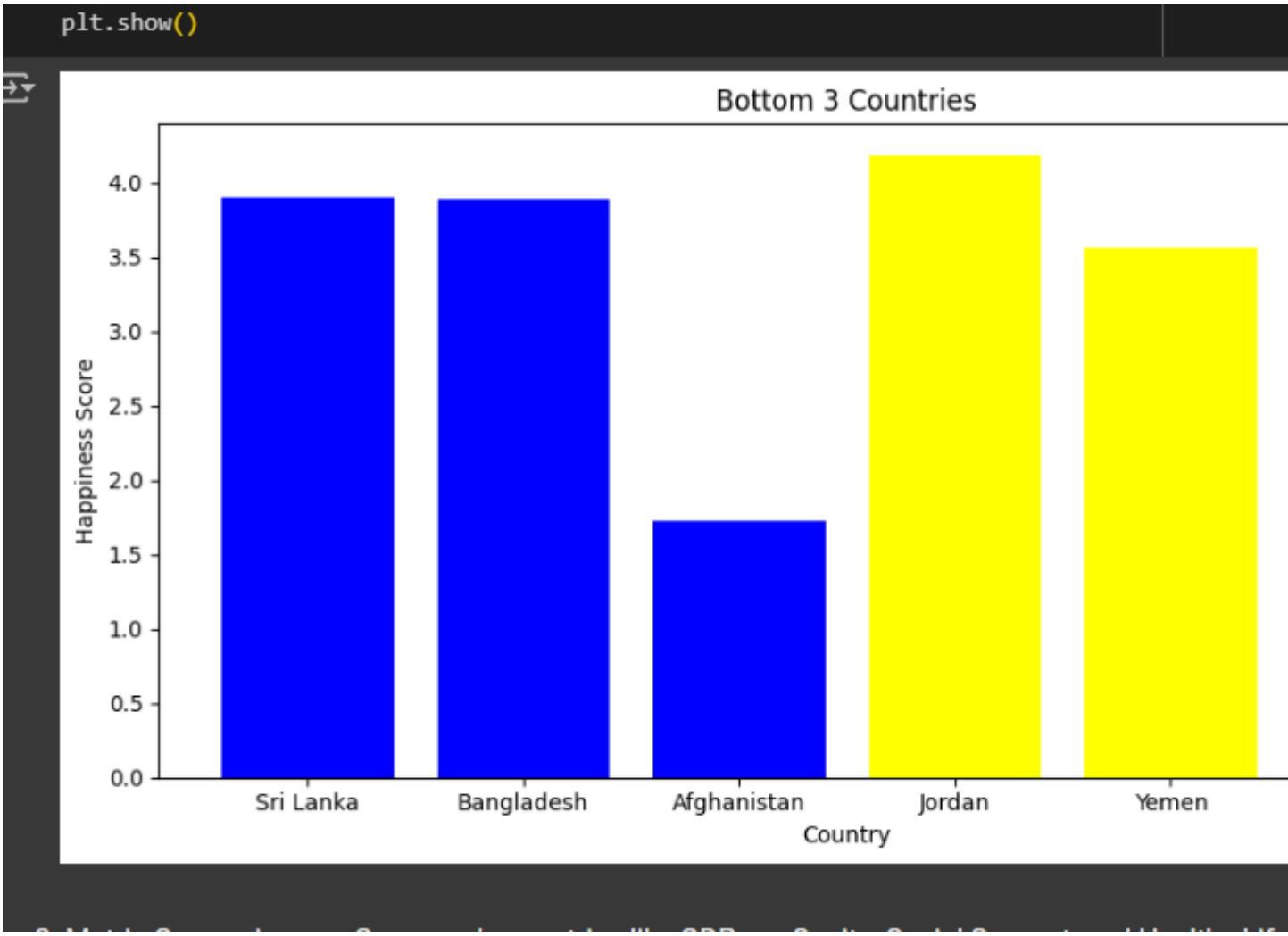
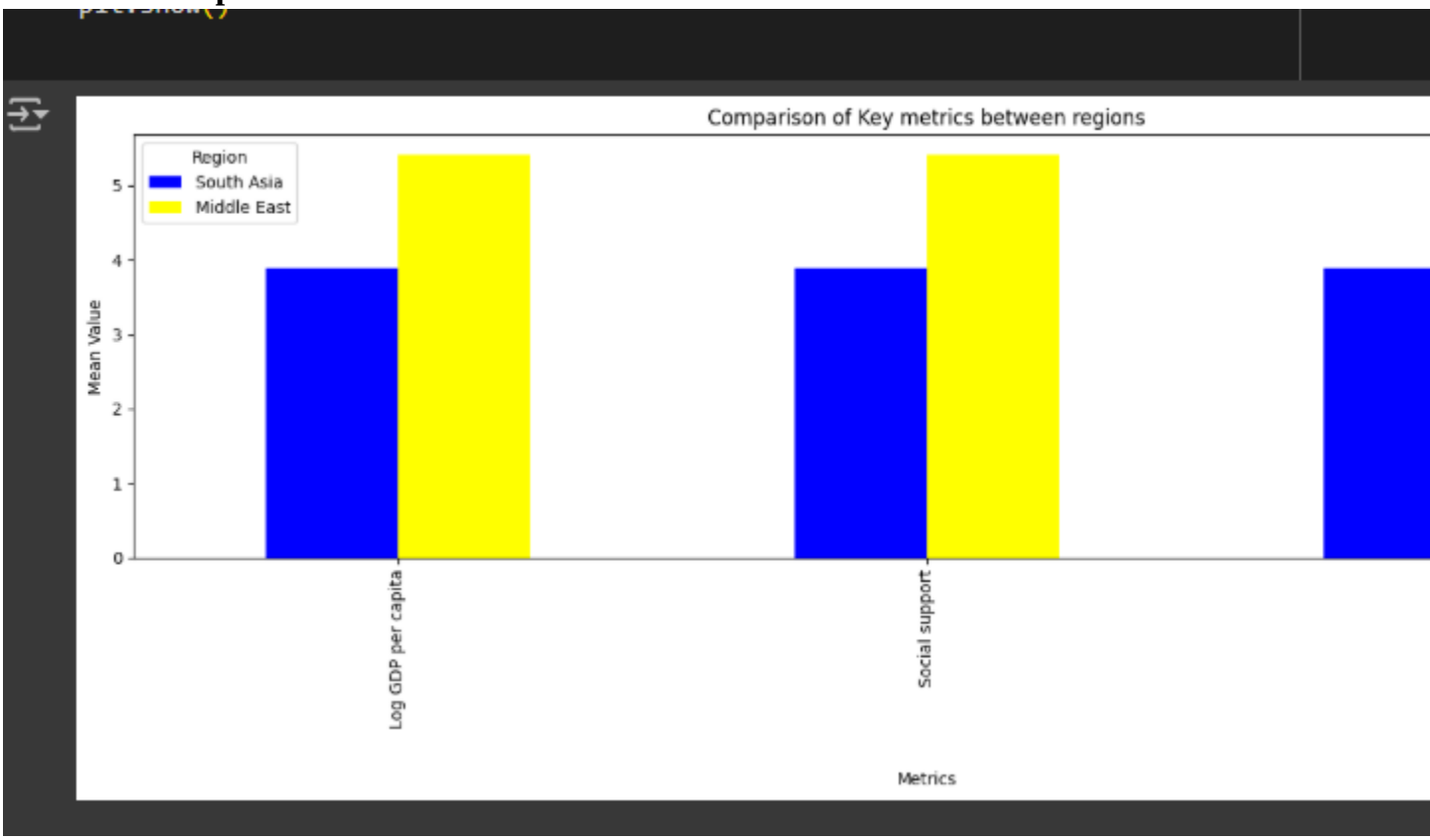


Fig: Bottom 3 Countries of both regions

After classifying the top 3 countries and bottom countries based on the score. The data was plotted in a bar graph for representation.

3. Metric Comparisons

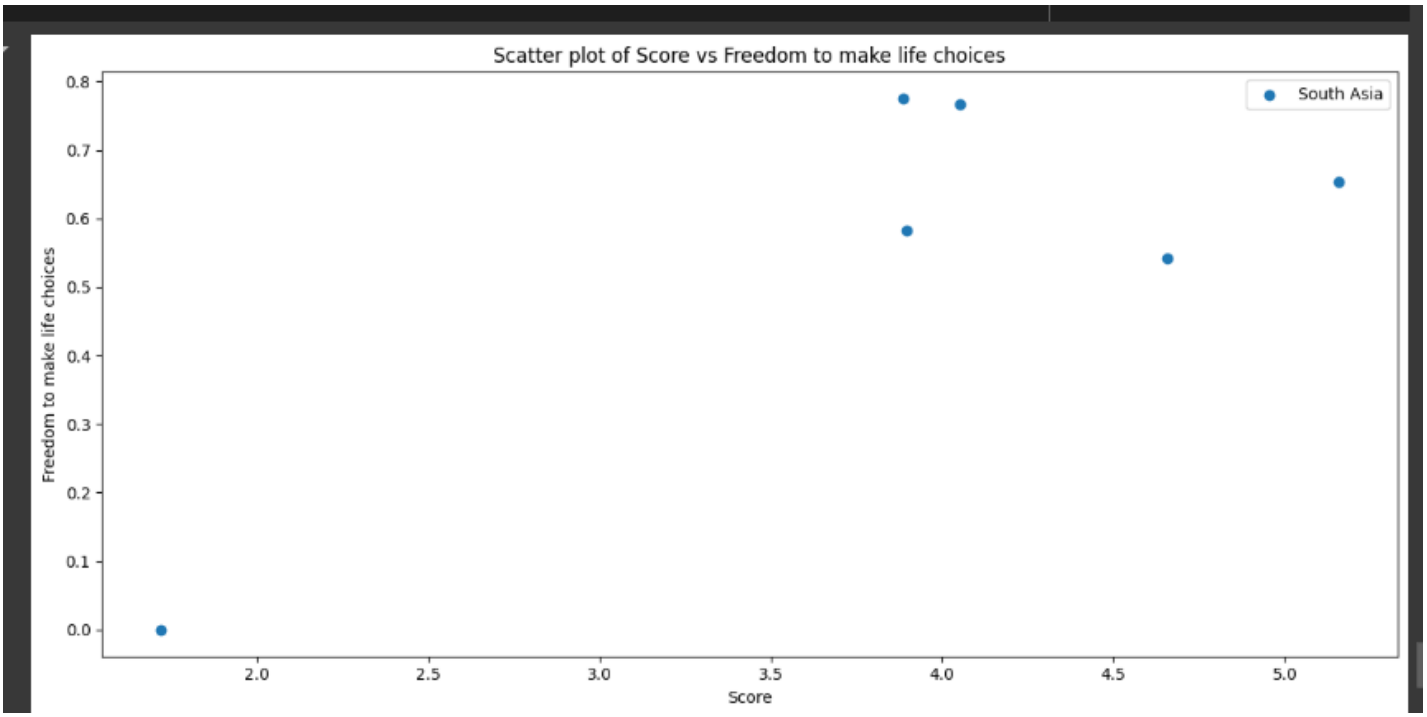


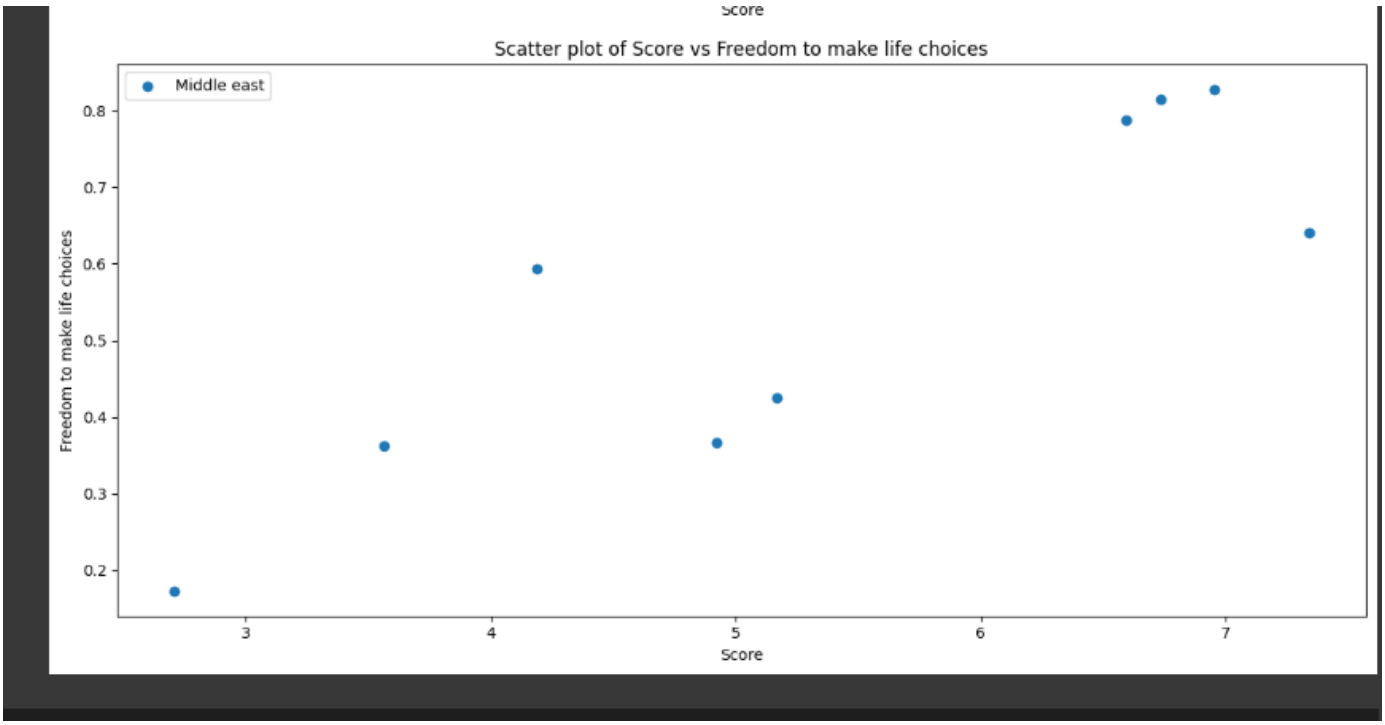
The bar chart compares three metrics: Log GDP per capita, Social Support and Healthy life expectancy between South Asia and Middle East. Blue representing South Asia and Yellow representing Middle East. Moreover, the crucial information that the bar plot is trying to show is that the Middle east has higher Mean Values in all three existing metrics compared to South Asia, pointing at major difference in economic and social support for a healthy life expectancy.

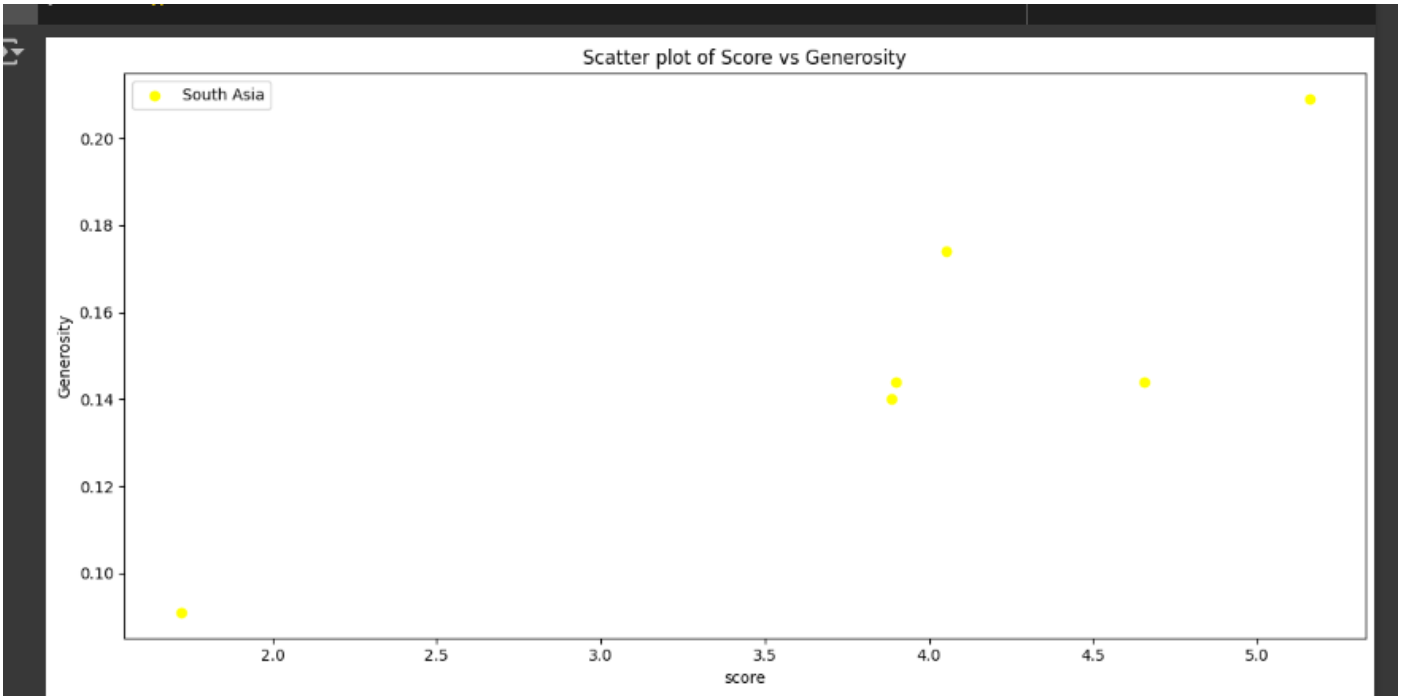
4. Happiness Disparity

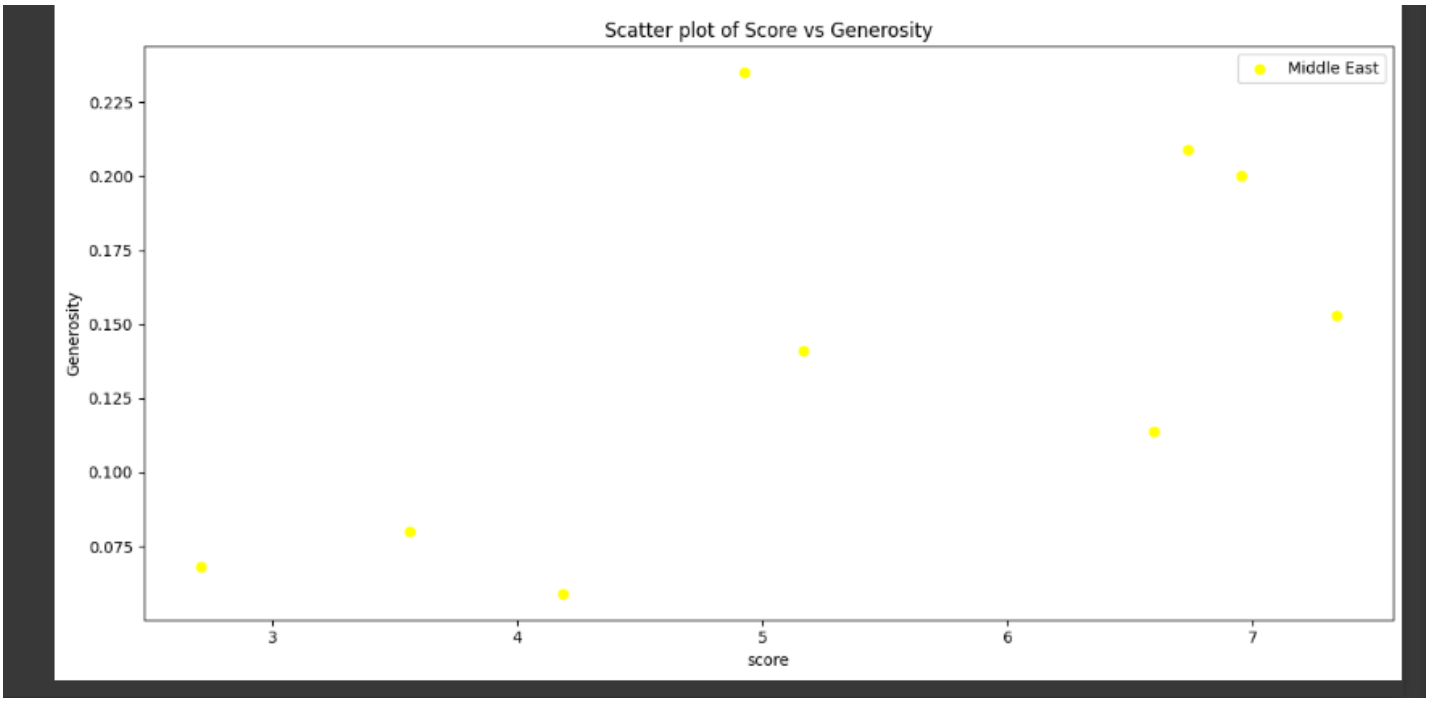
After comparing the two regions, we can see that the region with greater variability in happiness is South Asia as it's coefficient of variation (CV) is 30.21% compared to Middle East region, whose CV is 28.94%(approx.). Despite the Middle East having a larger range of 4.634 compared to 3.437 for South Asia, the result shows that South Asia has more variability of happiness.

5. Correlation Analysis


Scatter plot of Score vs Freedom to make life choices
South Asia
Freedom to make life choices
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Score
Scatter plot of Score vs Freedom to make life choices
Middle east
Freedom to make life choices
0.8
0.7
0.6
0.5
0.4
0.3
0.2
3
4
5
6
7
Score

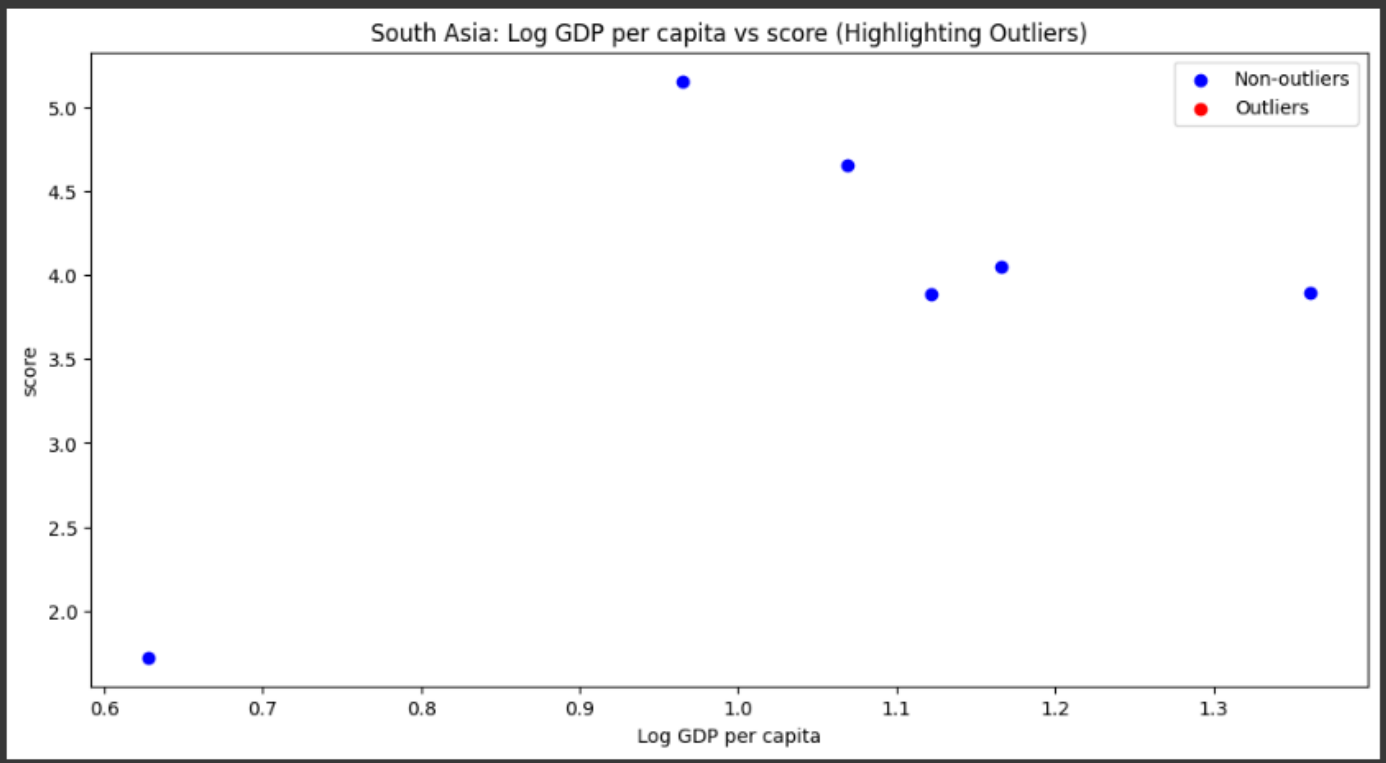


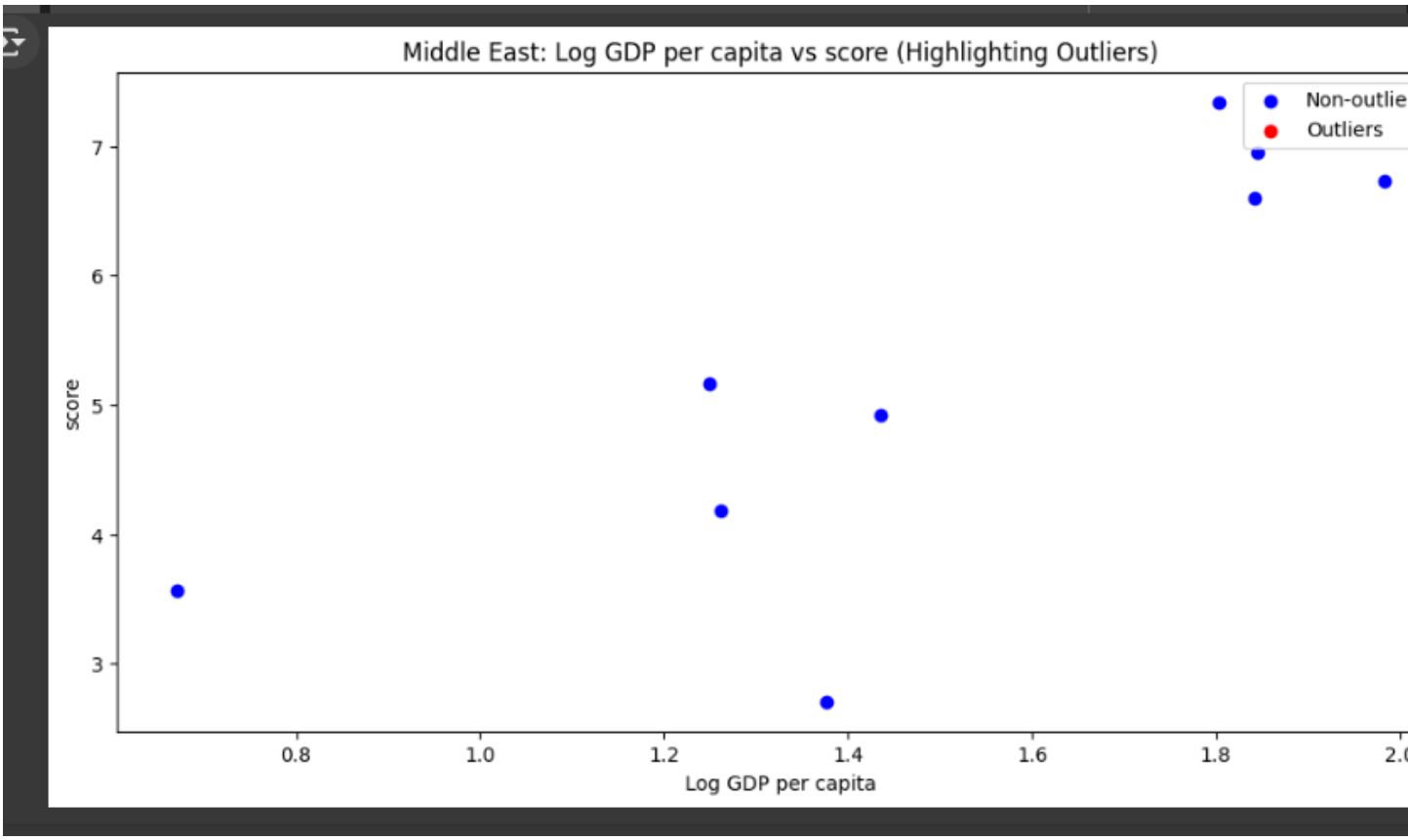


From the figures, we can interpret that , freedom to make life choices displays a direct correlation with Happiness score in both the regions whereas, Generosity shows a weak correlation with happiness score.

6. Outlier Detection

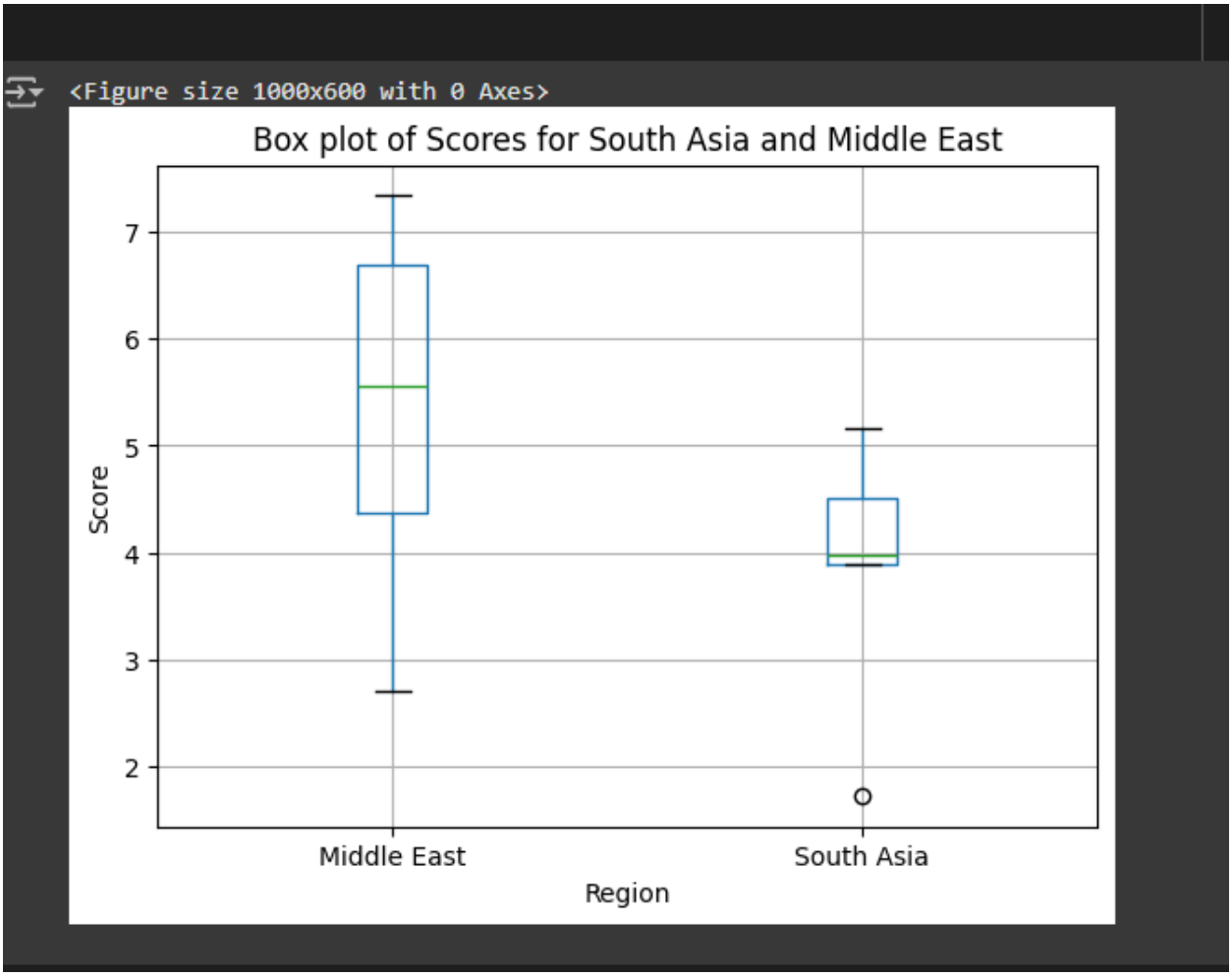
```
plt.show()
```


South Asia: Log GDP per capita vs score (Highlighting Outliers)
Non-outliers
Outliers
score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Log GDP per capita



In the above scatter plot, there contains no outliers for the following data as all the values lie in the normal distribution of numbers. Similarly, blue indicates non-outliers and red indicated outliers. Therefore, all the country has stable Happiness score with respect to GDP per capita.

7. Visualization



The box plot represents Happiness score between middle east and South Asia showing key differences in distribution to shapes, medians and outliers. The middle East has a higher median Happiness score (around 5.5) than South Asia (4.5) and Hence, the middle east usually has higher happiness level compared to South Asia.

Conclusion

To sum up, the interpretation of the World Happiness Report' mostly analyzes the Happiness scored focused mainly between 2 regions ('South Asia' and 'Middle East'). It can be clearly seen that south Asia is significantly lower in terms of Happiness and GDP per score. In terms of outliers, there was nothing as all the data were closely related to each other.