



5CS037

## Concepts and Technologies of AI

Name: Swoyam Pokharel

Group: L5CG26

Canvas ID: 2431342

Tutor: Nabin Acharya

## Table Of Contents:

Introduction:	3
Problem - 1: World Health Report	4
1.1 Load the dataset and display the first 10 rows.	4
1.2 Identify the number of rows and columns in the dataset.	4
1.3 List all the columns and their data types.	4
2.1 Calculate the mean, median, and standard deviation for the Score column	5
2.2 Identify the country with the highest and lowest happiness scores	5
3.1 Check if there are any missing values in the dataset. If so, display the total count for each column	5
4.1 Filter the dataset to show only the countries with a Score greater than 7.5	6
4.2 For the filtered dataset - Sort the dataset by GDP per Capita in descending order and display the top 10 rows.	6
5.1 Create a new column called Happiness Category that categorizes countries into three categories based on their Score:	6
6.1 Bar Plot: Plot the top 10 happiest countries by Score using a bar chart.	7
6.2 Line Plot: Plot the top 10 unhappiest countries by Score using a Line chart	7
6.3 Plot a histogram for the Score column to show its distribution and also interpret	8
6.4 Scatter Plot: Plot a scatter plot between GDP per Capita and Score to visualize their relationship	8
Problem - 2 - Some Advance Data Exploration Task:	9
1.1 Define the countries in South Asia with a list for example	9
1.2 Use the list from step - 1 to filter the dataset (i.e. filtered out matching dataset from list.)	9
1.3 Save the filtered data frame as separate CSV files for future use	10
2.1 Using the SouthAsia DataFrame, create a new column called Composite Score that combines the following metrics	10
2.2 Rank the South Asian countries based on the Composite Score in descending order	10
2.3 Visualize the top 5 countries using a horizontal bar chart showing the Composite Score.	11
2.4 Discuss whether the rankings based on the Composite Score align with the original Score - support your discussion with some visualization plot	11
3.1 Identify outlier countries in South Asia based on their Score and GDP per Capita	12
3.2 Define outliers using the $1.5 \times \text{IQR}$ rule	13
3.3 Create a scatter plot with GDP per Capita on the x-axis and Score on the y-axis, highlighting outliers in a different color	14
3.4 Discuss the characteristics of these outliers and their potential impact on regional averages	14
4.1. Choose two metrics (e.g., Freedom to Make Life Choices and Generosity) and calculate their correlation {pearson correlation} with the Score for South Asian countries.	15
4.2. Create scatter plots with trendlines for these metrics against the Score	15
4.3. Identify and discuss the strongest and weakest relationships between these metrics and the Score for South Asian countries.	16
5.1. Add a new column, GDP-Score Gap, which is the difference between GDP per Capita and the Score for each South Asian country	17
5.2. Rank the South Asian countries by this gap in both ascending and descending order.	17
5.3 Highlight the top 3 countries with the largest positive and negative gaps using a bar chart.	18
5.4 Analyze the reasons behind these gaps and their implications for South Asian countries.	18
Problem - 3 - Comparative Analysis:	18
1.1 Similar in Task - 1 of Problem 2 create a dataframe from middle eastern countries. For hint use the following list:	18
2.1 Calculate the mean, Standard deviation of the score for both South Asia and Middle East.	19
2.2 Which region has higher happiness Scores on average?	19
3.1 Identify the top 3 and bottom 3 countries in each region based on the score.	20
3.2 Plot bar charts comparing these charts.	20
4.1 Compare key metrics like GDP per Capita, Social Support, and Healthy Life Expectancy between the regions using grouped bar charts.	21
4.2 Which metrics show the largest disparity between the two regions?	21
5.1 Compute the range (max - min) and coefficient of variation (CV) for Score in both regions.	21
5.2 Which region has greater variability in happiness?	21
6.1 Analyze the correlation of Score with other metrics Freedom to Make Life Choices, and Generosity within each region.	22
6.2 Create scatter plots to visualize and interpret the relationships	22
7.1 Identify outlier countries in both regions based on Score and GDP per Capita.	25
7.2 Plot these outliers and discuss their implications	26
8.1 Create boxplots comparing the distribution of Score between South Asia and the Middle East	26
8.2 Interpret the key differences in distribution shapes, medians, and outliers.	27
Conclusion:	28

## Introduction:

The World Happiness Report is an annual publication that ranks countries based on their levels of happiness and well-being. These rankings are determined by various factors, such as income, social support, life expectancy, freedom to life choices, generosity and perceptions of corruption.

In this analysis, we will explore the World Happiness Report dataset and perform a series of tasks to better understand happiness scores across South Asia and the Middle East. The task is divided into three sections, Data Exploration and Basic Analysis, Advanced Data Exploration and Comparative Analysis.

All the code related into making of this report is at this github repository: <https://github.com/PS-Wizard/School/tree/main/ai>

# Problem - 1: World Health Report

## 1.1 Load the dataset and display the first 10 rows.

	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
count	143.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
mean	5.527580	1.378807	1.134329	0.520886	0.620621	0.146271	0.154121	1.575914
std	1.170717	0.425098	0.333317	0.164923	0.162492	0.073441	0.126238	0.537459
min	1.721000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.073000
25%	4.726000	1.077750	0.921750	0.398000	0.527500	0.091000	0.068750	1.308250
50%	5.785000	1.431500	1.237500	0.549500	0.641000	0.136500	0.120500	1.644500
75%	6.416000	1.741500	1.383250	0.648500	0.736000	0.192500	0.193750	1.881750
max	7.741000	2.141000	1.617000	0.857000	0.863000	0.401000	0.575000	2.998000

Fig 1.1: Description of the Entire World Health Report

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
0	Finland	7.741	1.844	1.572	0.695	0.859	0.142	0.546	2.082
1	Denmark	7.583	1.908	1.520	0.699	0.823	0.204	0.548	1.881
2	Iceland	7.525	1.881	1.617	0.718	0.819	0.258	0.182	2.050
3	Sweden	7.344	1.878	1.501	0.724	0.838	0.221	0.524	1.658
4	Israel	7.341	1.803	1.513	0.740	0.641	0.153	0.193	2.298
5	Netherlands	7.319	1.901	1.462	0.706	0.725	0.247	0.372	1.906
6	Norway	7.302	1.952	1.517	0.704	0.835	0.224	0.484	1.586
7	Luxembourg	7.122	2.141	1.355	0.708	0.801	0.146	0.432	1.540
8	Switzerland	7.060	1.970	1.425	0.747	0.759	0.173	0.498	1.488
9	Australia	7.057	1.854	1.461	0.692	0.756	0.225	0.323	1.745

Fig 1.2: First 10 rows from the World Health Report

From the image, we can interpret that all of the columns are numerical with an exception for `Country name`.

## 1.2 Identify the number of rows and columns in the dataset.

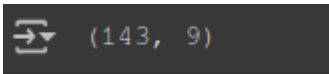
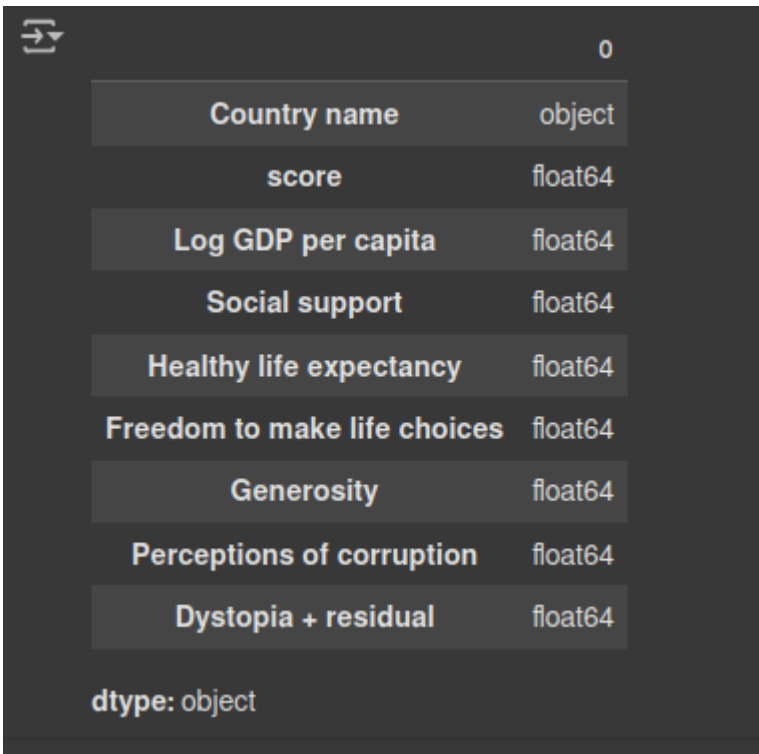


Fig 1.3: Number of rows and columns in the dataset

The results show that there are 143 country's data in the World Health Report.

## 1.3 List all the columns and their data types.



Country name	object
score	float64
Log GDP per capita	float64
Social support	float64
Healthy life expectancy	float64
Freedom to make life choices	float64
Generosity	float64
Perceptions of corruption	float64
Dystopia + residual	float64
dtype: object	

Fig 1.4: The datatypes of all the columns in the dataset

### 2.1 Calculate the mean, median, and standard deviation for the Score column

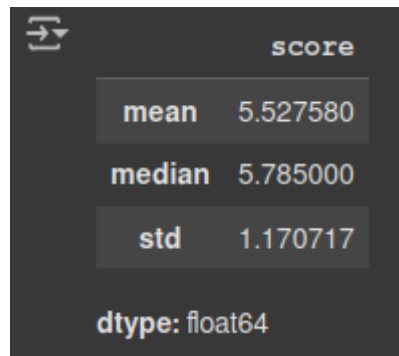


Fig 1.5: Mean, Median and standard deviation of the score column

The results show that the median is greater than mean, which indicates that the data is skewed in the lower end as countries with lower happiness scores are pulling the average mean down. The standard deviation of ~ 1.2 indicates some spread in the data.

### 2.2 Identify the country with the highest and lowest happiness scores

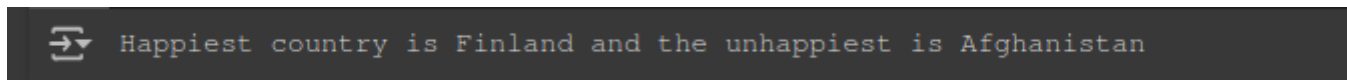


Fig 1.6: Identification of the happiest and unhappiest country in the dataset

### 3.1 Check if there are any missing values in the dataset. If so, display the total count for each column

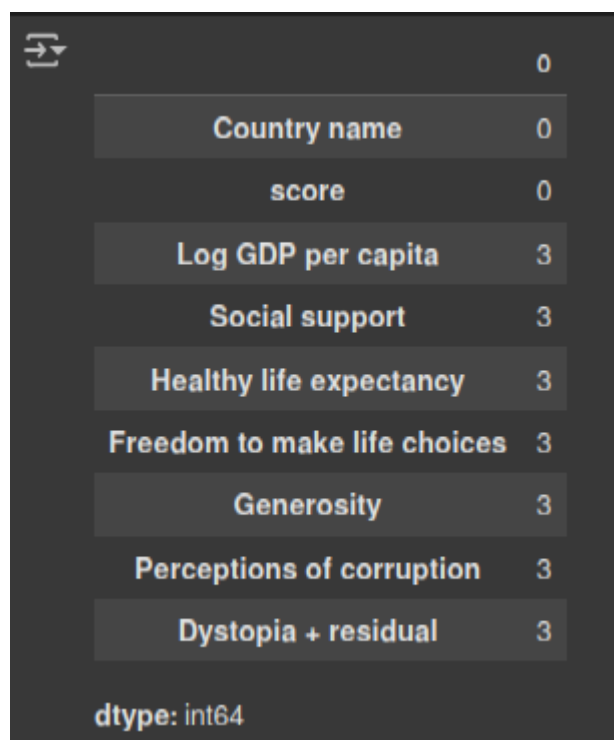


Fig 1.7: Image displaying the number of missing values in each column of the dataset

The results show a few missing values all in the numerical columns. Dropping these rows, would result in losing entire countries from the dataset, which was already containing only 143 countries. So the data were filled with the median. I chose median because we previously saw skewness in the data.

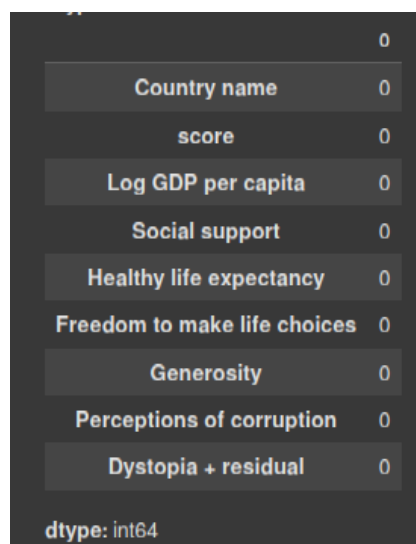


Fig 1.8: Image after treating the missing values.

4.1 Filter the dataset to show only the countries with a Score greater than 7.5

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
0	Finland	7.741	1.844	1.572	0.695	0.859	0.142	0.546	2.082
1	Denmark	7.583	1.908	1.520	0.699	0.823	0.204	0.548	1.881
2	Iceland	7.525	1.881	1.617	0.718	0.819	0.258	0.182	2.050

Fig 1.9: Countries That have a higher score than 7.5

4.2 For the filtered dataset - Sort the dataset by GDP per Capita in descending order and display the top 10 rows.

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
1	Denmark	7.583	1.908	1.520	0.699	0.823	0.204	0.548	1.881
2	Iceland	7.525	1.881	1.617	0.718	0.819	0.258	0.182	2.050
0	Finland	7.741	1.844	1.572	0.695	0.859	0.142	0.546	2.082

Fig 2.0: Image of the filtered countries sorted by GDP in descending order

5.1 Create a new column called Happiness Category that categorizes countries into three categories based on their Score:

Low – (Score < 4)

Medium – (4 ≤ Score ≤ 6)

High – (Score > 6)

	0	1	2	3	4	5	6	7	8	9	...	133	134	135	136	137	138	139	140	141
Country name	Finland	Denmark	Iceland	Sweden	Israel	Netherlands	Norway	Luxembourg	Switzerland	Australia	...	Zambia	Eswatini	Malawi	Botswana	Zimbabwe	Congo (Kinshasa)	Sierra Leone	Lesotho	Lebanon
score	7.741	7.583	7.525	7.344	7.341	7.319	7.302	7.122	7.06	7.057	...	3.502	3.502	3.421	3.383	3.341	3.295	3.245	3.186	2.707
Log GDP per capita	1.844	1.908	1.881	1.878	1.803	1.901	1.952	2.141	1.97	1.854	...	0.899	1.255	0.617	1.445	0.748	0.534	0.654	0.771	1.377
Social support	1.572	1.52	1.617	1.501	1.513	1.462	1.517	1.355	1.425	1.461	...	0.809	0.925	0.41	0.969	0.85	0.665	0.566	0.851	0.577
Healthy life expectancy	0.695	0.699	0.718	0.724	0.74	0.706	0.704	0.708	0.747	0.692	...	0.264	0.176	0.349	0.241	0.232	0.262	0.253	0.0	0.556
Freedom to make life choices	0.859	0.823	0.819	0.838	0.641	0.725	0.835	0.801	0.759	0.756	...	0.727	0.284	0.571	0.567	0.487	0.473	0.469	0.523	0.173
Generosity	0.142	0.204	0.258	0.221	0.153	0.247	0.224	0.146	0.173	0.225	...	0.168	0.059	0.135	0.014	0.096	0.189	0.181	0.082	0.068
Perceptions of corruption	0.546	0.548	0.182	0.524	0.193	0.372	0.484	0.432	0.498	0.323	...	0.109	0.116	0.136	0.082	0.131	0.072	0.053	0.085	0.029
Dystopia + residual	2.082	1.881	2.05	1.658	2.298	1.906	1.586	1.54	1.488	1.745	...	0.526	0.686	1.203	0.066	0.797	1.102	1.068	0.875	-0.073
Happiness_Category	High	High	High	High	High	High	High	High	High	High	...	Low	Low	Low	Low	Low	Low	Low	Low	Low

10 rows × 143 columns

Fig 2.1: Image after creating a new column Happiness\_Category

After counting, it is observed that 56 countries fell in the High category, 17 in the low, and a majority of 70 in the medium category. This suggests that most countries are positioned in the middle range.

## 6.1 Bar Plot: Plot the top 10 happiest countries by Score using a bar chart.

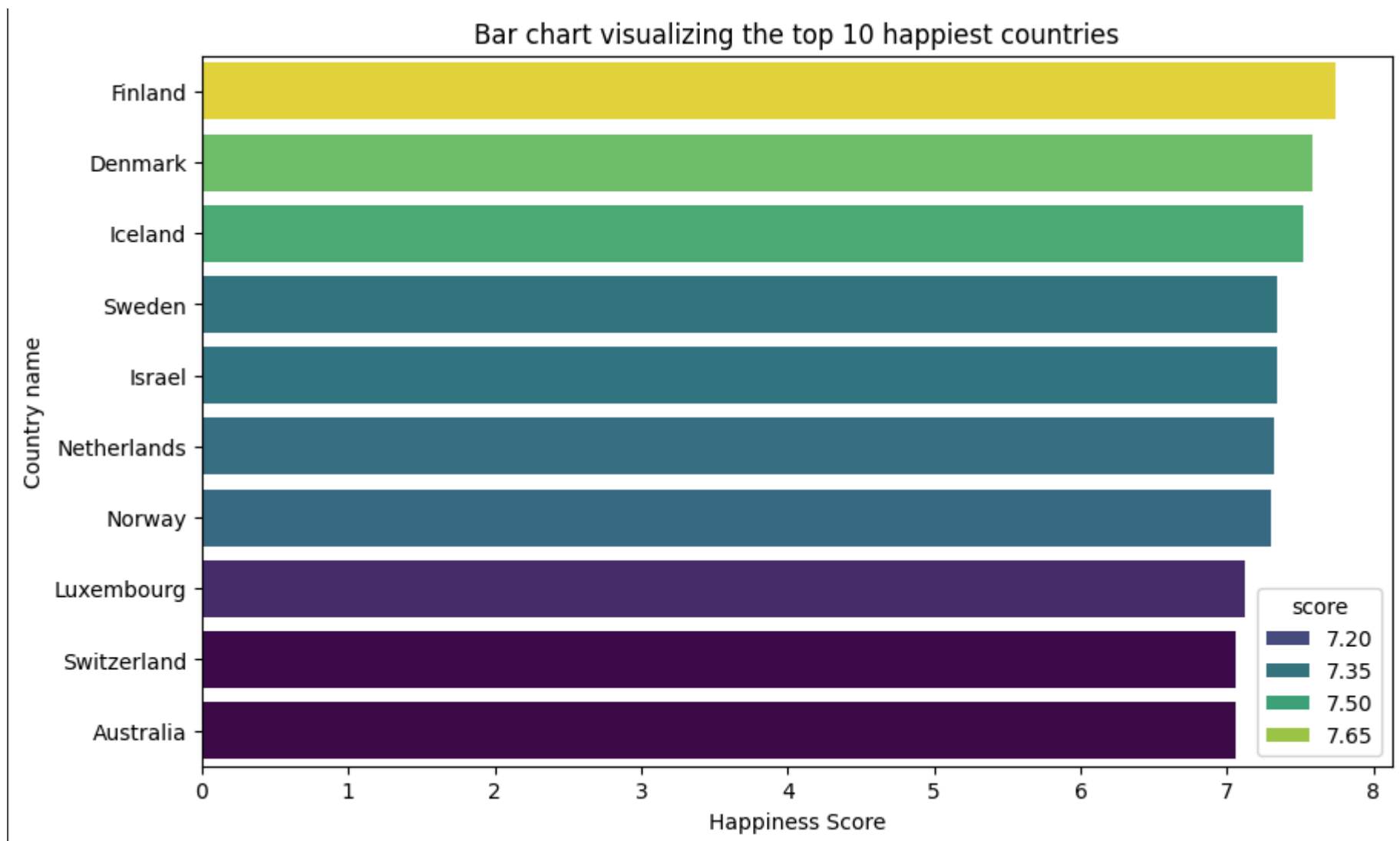


Fig 2.2: Result of plotting the top 10 happiest countries in a bar chart

## 6.2 Line Plot: Plot the top 10 unhappiest countries by Score using a Line chart

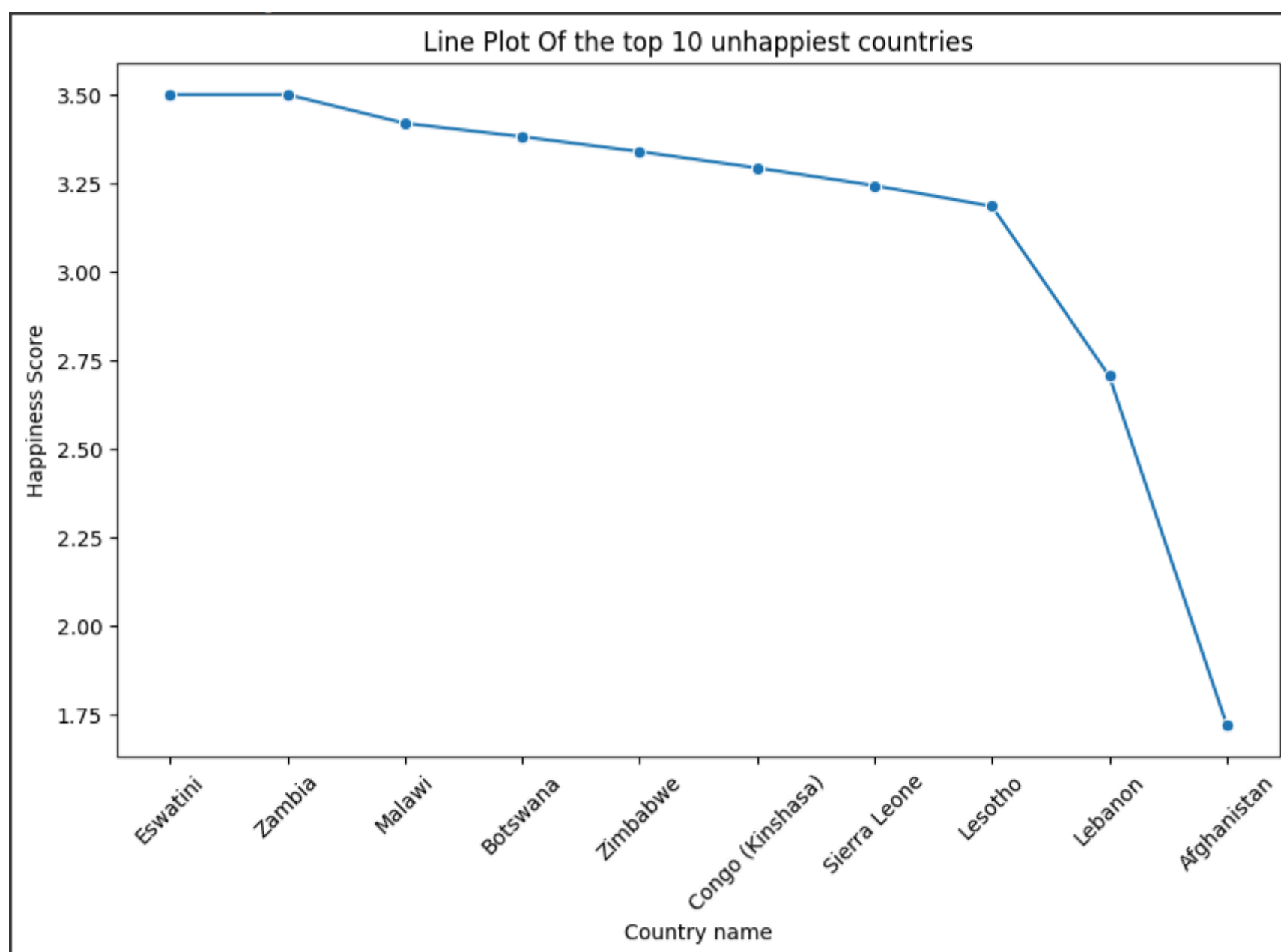


Fig 2.3: A Line Plot of the least 10 happiest countries.

The results indicate that the unhappiest countries almost all have a similar unhappiness score and a few have drastically low happiness scores.

### 6.3 Plot a histogram for the Score column to show its distribution and also interpret

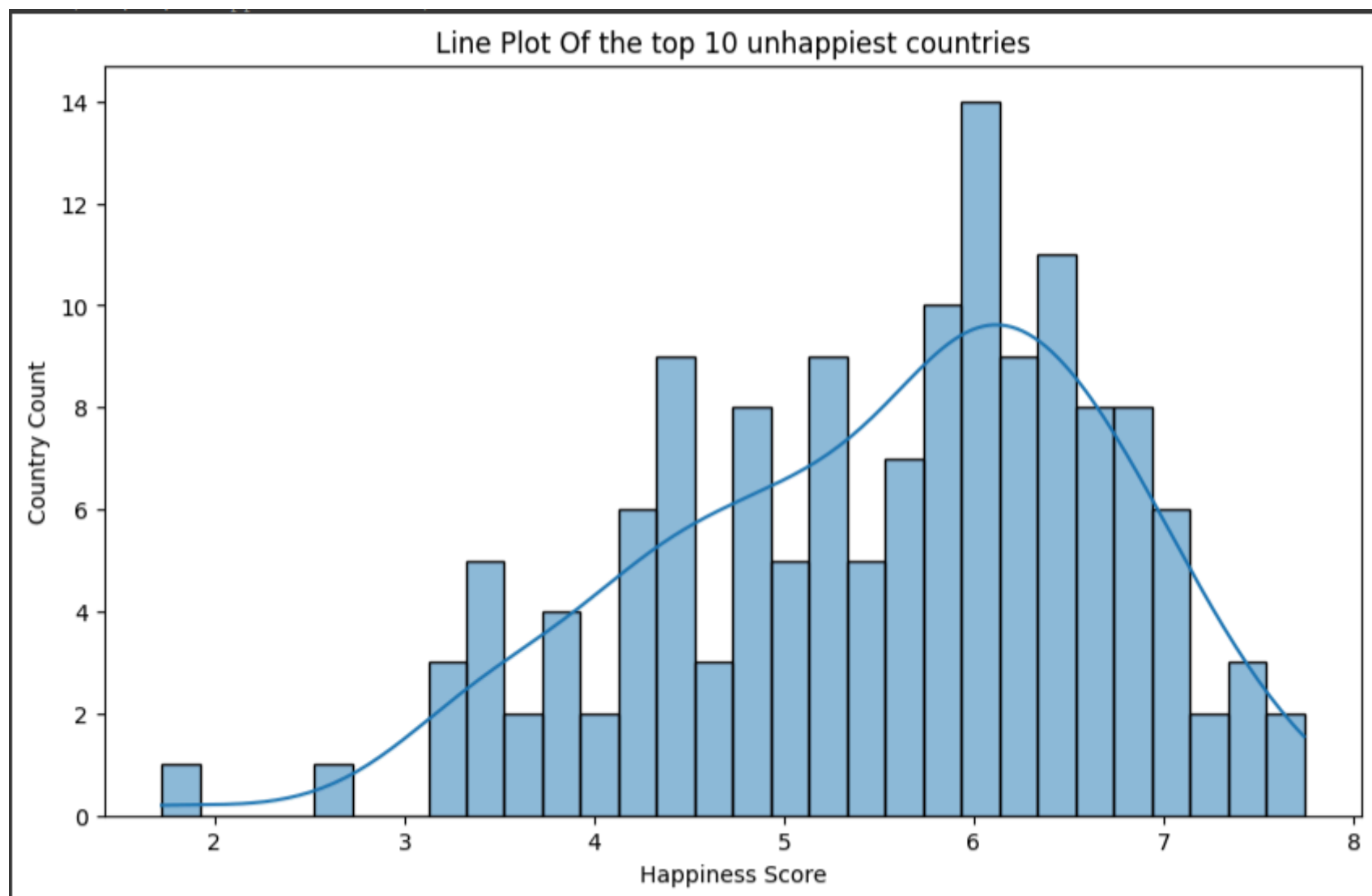


Fig 2.4: Histogram showing the distribution of the score column.

This visualization further solidifies our interpretation of the skewed distribution of the dataset. The data seems to be unimodal with a peak to the right, indicating a left skewed distribution.

### 6.4 Scatter Plot: Plot a scatter plot between GDP per Capita and Score to visualize their relationship

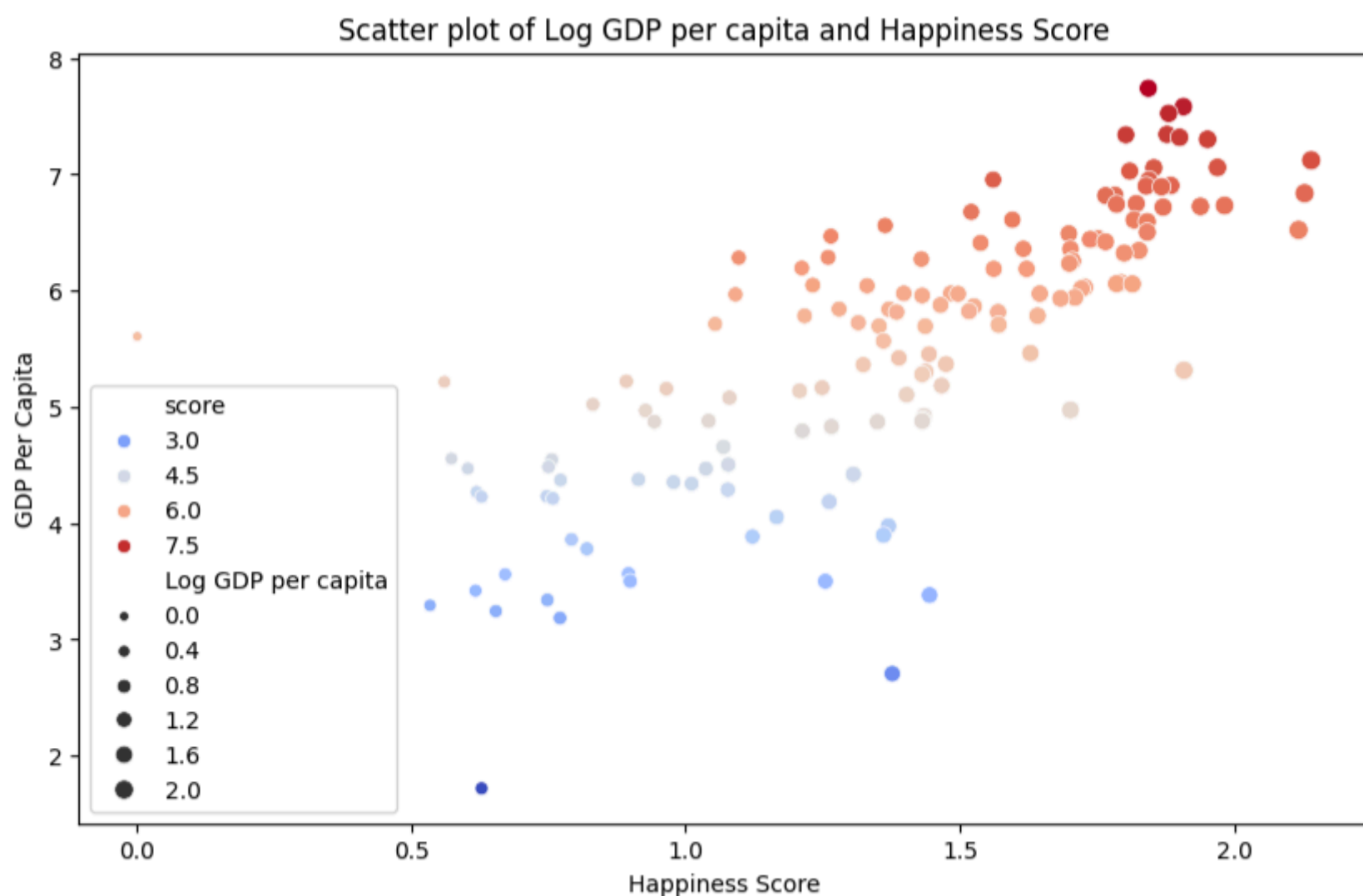


Fig 2.5: A scatter plot between GDP per capita and Score.



The data clearly indicates a positive correlation between these variables. It shows that the countries that are happier on average are also the ones with the highest gdp per capita. Majority of the data points are clustered toward the middle and upper spectrum of both axes, which suggests that most countries have a moderate to high happiness and GDP scores.

## Problem - 2 - Some Advance Data Exploration Task:

### 1.1 Define the countries in South Asia with a list for example

```
south_asian_countries = ["Afghanistan", "Bangladesh", "Bhutan", "India", "Maldives", "Nepal", "Pakistan", "Srilanka"]
```

```
south_asian_countries = ["Afghanistan", "Bangladesh", "Bhutan", "India", "Maldives", "Nepal", "Pakistan", "Srilanka"]
```

Fig 2.6: Declaring the countries in South Asia in a list.

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

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category
92	Nepal	5.158	0.965	0.990	0.443	0.653	0.209	0.115	1.783	Medium
107	Pakistan	4.657	1.069	0.600	0.321	0.542	0.144	0.074	1.907	Medium
125	India	4.054	1.166	0.653	0.417	0.767	0.174	0.122	0.756	Medium
128	Bangladesh	3.886	1.122	0.249	0.513	0.775	0.140	0.167	0.919	Low
142	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	0.088	0.672	Low

Fig 2.7: Image of the data frame containing South Asian Countries filtered from the original dataset

The results indicate that only 5 out of the 8 countries are present in the dataset. To address the missing countries, the values were calculated with the median of the dataset. While this doesn't accurately represent the actual situation in those countries, it merely serves as a strategy to counter the missing values.

	0	1	2	3	4	5	6	7
Country name	Nepal	Pakistan	India	Bangladesh	Afghanistan	Bhutan	Maldives	Srilanka
score	5.158	4.657	4.054	3.886	1.721	4.054	4.054	4.054
Log GDP per capita	0.965	1.069	1.166	1.122	0.628	1.069	1.069	1.069
Social support	0.99	0.6	0.653	0.249	0.0	0.6	0.6	0.6
Healthy life expectancy	0.443	0.321	0.417	0.513	0.242	0.417	0.417	0.417
Freedom to make life choices	0.653	0.542	0.767	0.775	0.0	0.653	0.653	0.653
Generosity	0.209	0.144	0.174	0.14	0.091	0.144	0.144	0.144
Perceptions of corruption	0.115	0.074	0.122	0.167	0.088	0.115	0.115	0.115
Dystopia + residual	1.783	1.907	0.756	0.919	0.672	0.919	0.919	0.919
Happiness_Category	Medium	Medium	Medium	Low	Low	Medium	Medium	Medium

Fig 2.8: Image after filling the values for the missing countries.

Description of the dataset:

	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
count	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000
mean	3.954750	1.019625	0.536500	0.398375	0.587000	0.148750	0.113875	1.099250
std	0.998194	0.168189	0.294194	0.081904	0.248355	0.033316	0.027132	0.470607
min	1.721000	0.628000	0.000000	0.242000	0.000000	0.091000	0.074000	0.672000
25%	4.012000	1.043000	0.512250	0.393000	0.625250	0.143000	0.108250	0.878250
50%	4.054000	1.069000	0.600000	0.417000	0.653000	0.144000	0.115000	0.919000
75%	4.204750	1.082250	0.613250	0.423500	0.681500	0.151500	0.116750	1.135000
max	5.158000	1.166000	0.990000	0.513000	0.775000	0.209000	0.167000	1.907000

### 1.3 Save the filtered data frame as separate CSV files for future use

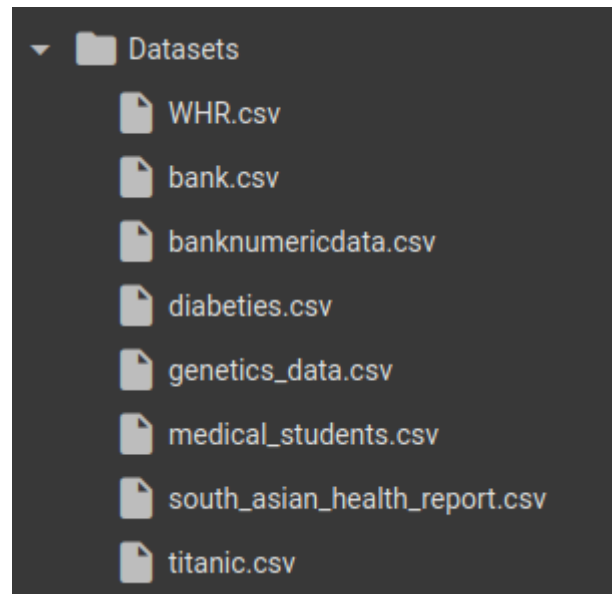


Fig 2.9: Image of the file tree where the data frame was stored as a CSV.

### 2.1 Using the SouthAsia DataFrame, create a new column called Composite Score that combines the following metrics

Composite Score =  $0.40 \times \text{GDP per Capita} + 0.30 \times \text{Social Support} + 0.30 \times \text{Healthy Life Expectancy}$

	Country name	Composite Score
0	Nepal	0.8159
1	Pakistan	0.7039
2	India	0.7874
3	Bangladesh	0.6774
4	Afghanistan	0.3238
5	Bhutan	0.7327
6	Maldives	0.7327
7	Srilanka	0.7327

Fig 3.0: Results of creating the Composite Score column

While the same composite scores for Bhutan, Maldives, and Sri Lanka may raise concerns, this is expected, as the missing values for these countries were missing in the original dataset and later estimated using the median.

### 2.2 Rank the South Asian countries based on the Composite Score in descending order

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category	Composite Score
0	Nepal	5.158	0.965	0.990	0.443	0.653	0.209	0.115	1.783	Medium	0.8159
2	India	4.054	1.166	0.653	0.417	0.767	0.174	0.122	0.756	Medium	0.7874
5	Bhutan	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327
6	Maldives	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327
7	Srilanka	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327
1	Pakistan	4.657	1.069	0.600	0.321	0.542	0.144	0.074	1.907	Medium	0.7039
3	Bangladesh	3.886	1.122	0.249	0.513	0.775	0.140	0.167	0.919	Low	0.6774
4	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	0.088	0.672	Low	0.3238

Fig 3.1: Result after ranking the South Asian Countries based off their Composite Score

## 2.3 Visualize the top 5 countries using a horizontal bar chart showing the Composite Score.

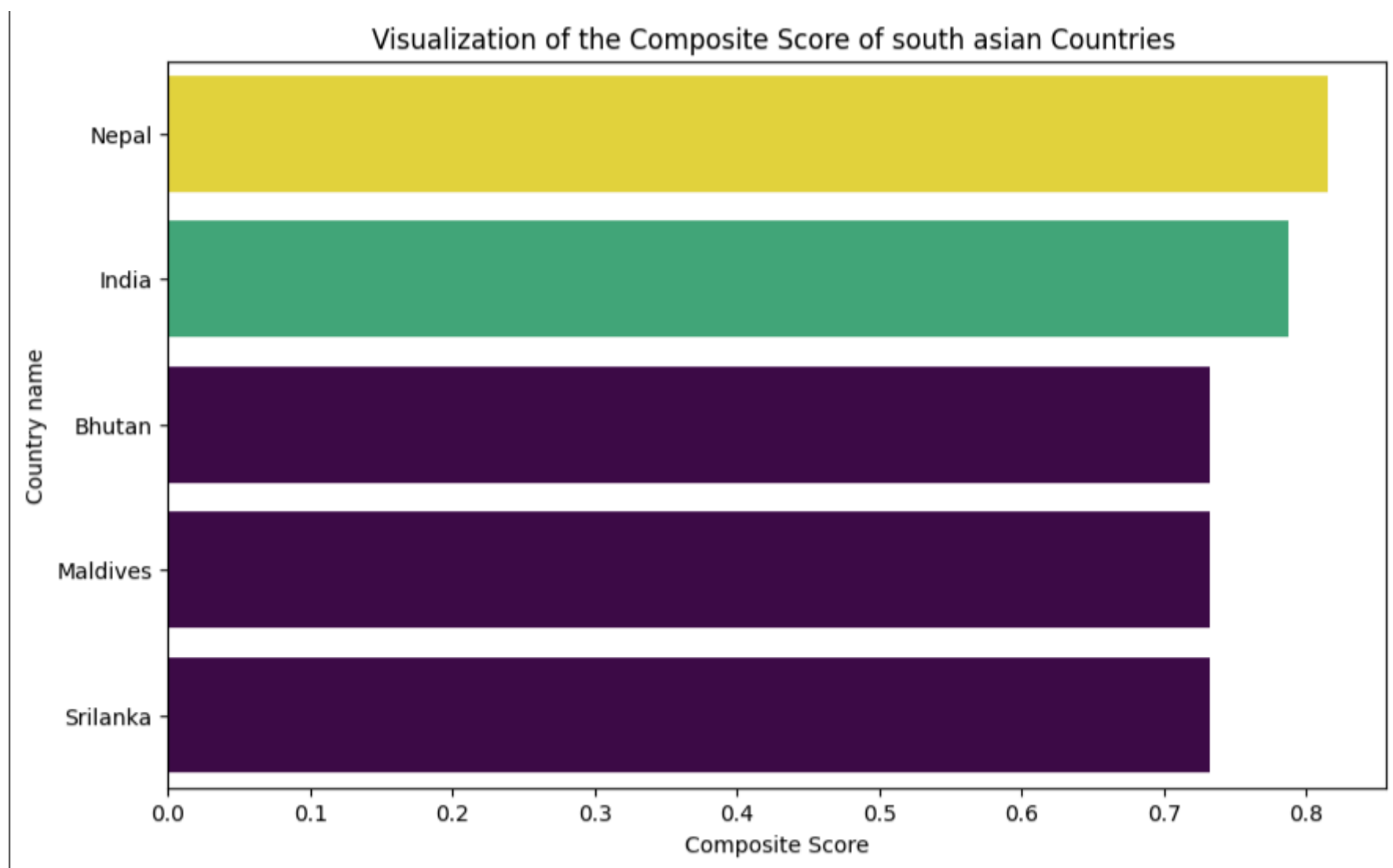
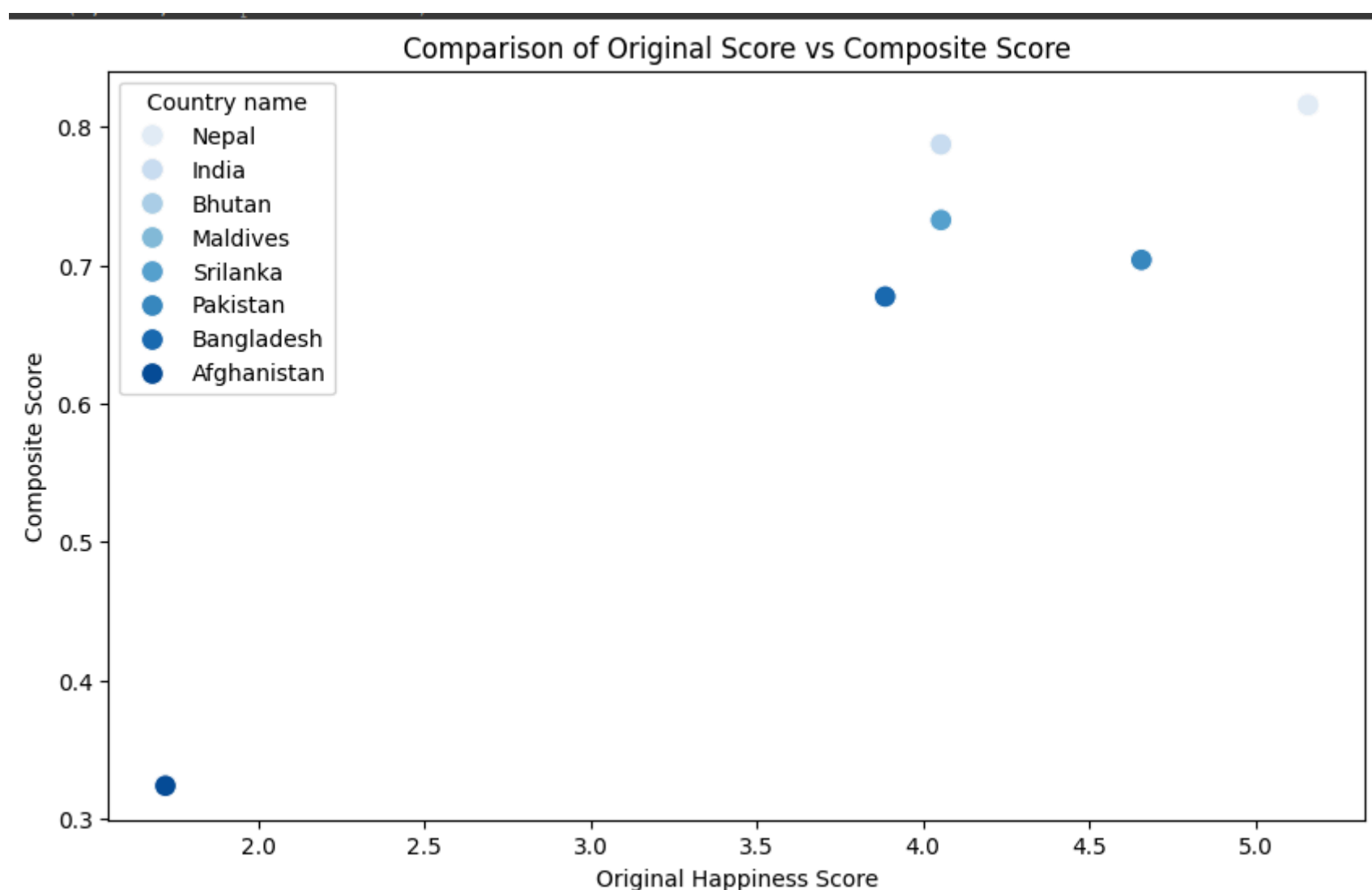


Fig 3.2: Visualization of the top 5 countries based on the composite score.

## 2.4 Discuss whether the rankings based on the Composite Score align with the original Score - support your discussion with some visualization plot



Both rankings have Nepal at the top, which shows that these two metrics align well at least for the highest ranked country. However, difference is seen from this point on, Pakistan which ranks 6th based off the Composite score ranks 2nd based on the original score. Bhutan, India, Maldives and Sri Lanka although have identical original scores, have different composite scores, which indicates that Composite Score weights the metrics differently.

The composite score considers factors such as GDP, Social support and Healthy Life expectancy which changes the rankings of some countries. This can imply that the Composite Score might give a nuanced picture compared to the original score.

### 3.1 Identify outlier countries in South Asia based on their Score and GDP per Capita

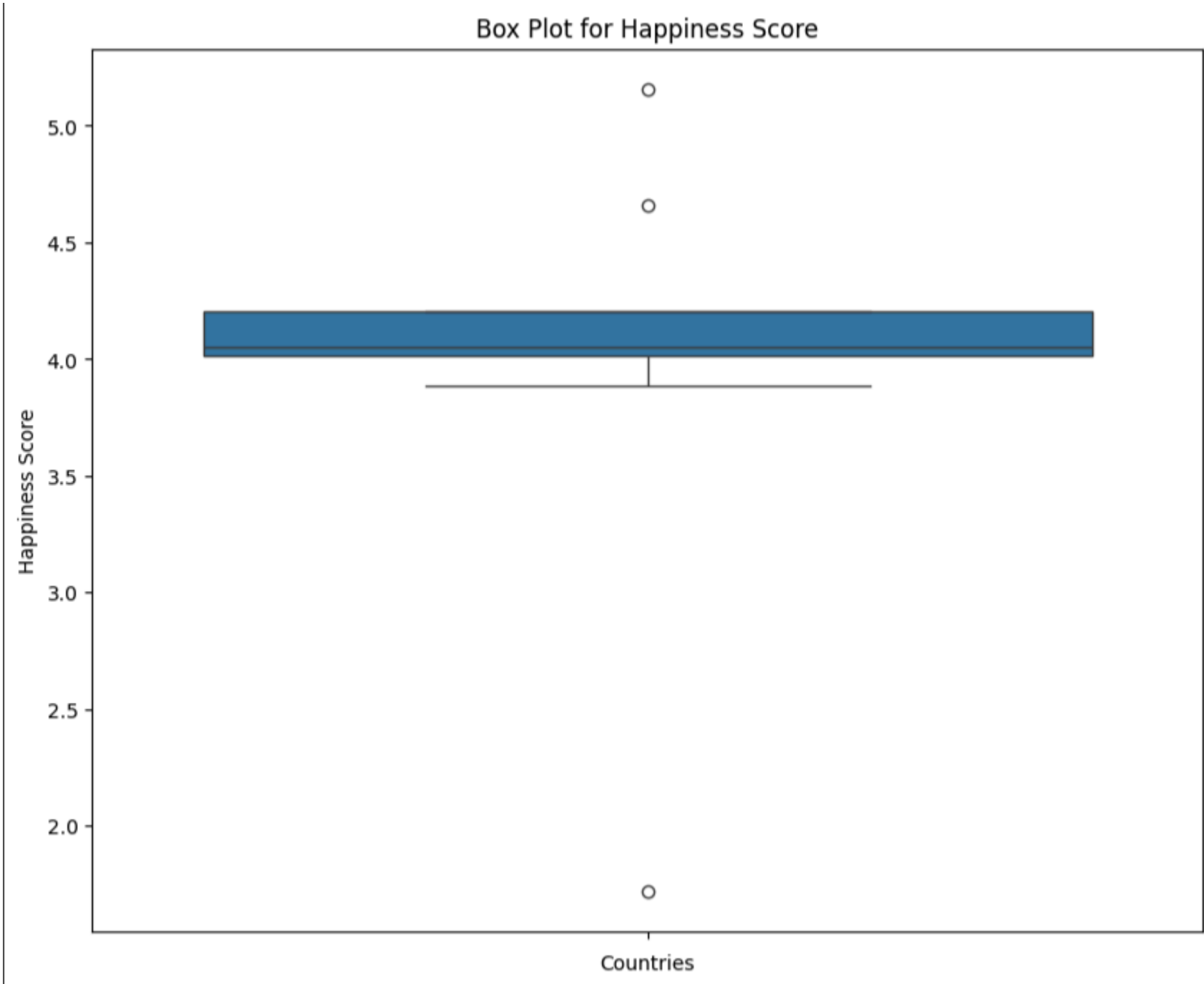


Fig 3.3: Box plot of the South Asian Countries based on their happiness scores

The result indicates that the median is much closer to the first quartile. The missing of the top whisker suggests that there are no values that extend the third quartile and the data is heavily skewed. The plot shows 3 outliers with 2 countries having exceptionally high happiness scores compared to the median, and a country having a very low happiness score. The majority of the countries have a happiness score hovering between ~4.0 - ~4.25 which suggests a low score for most countries.

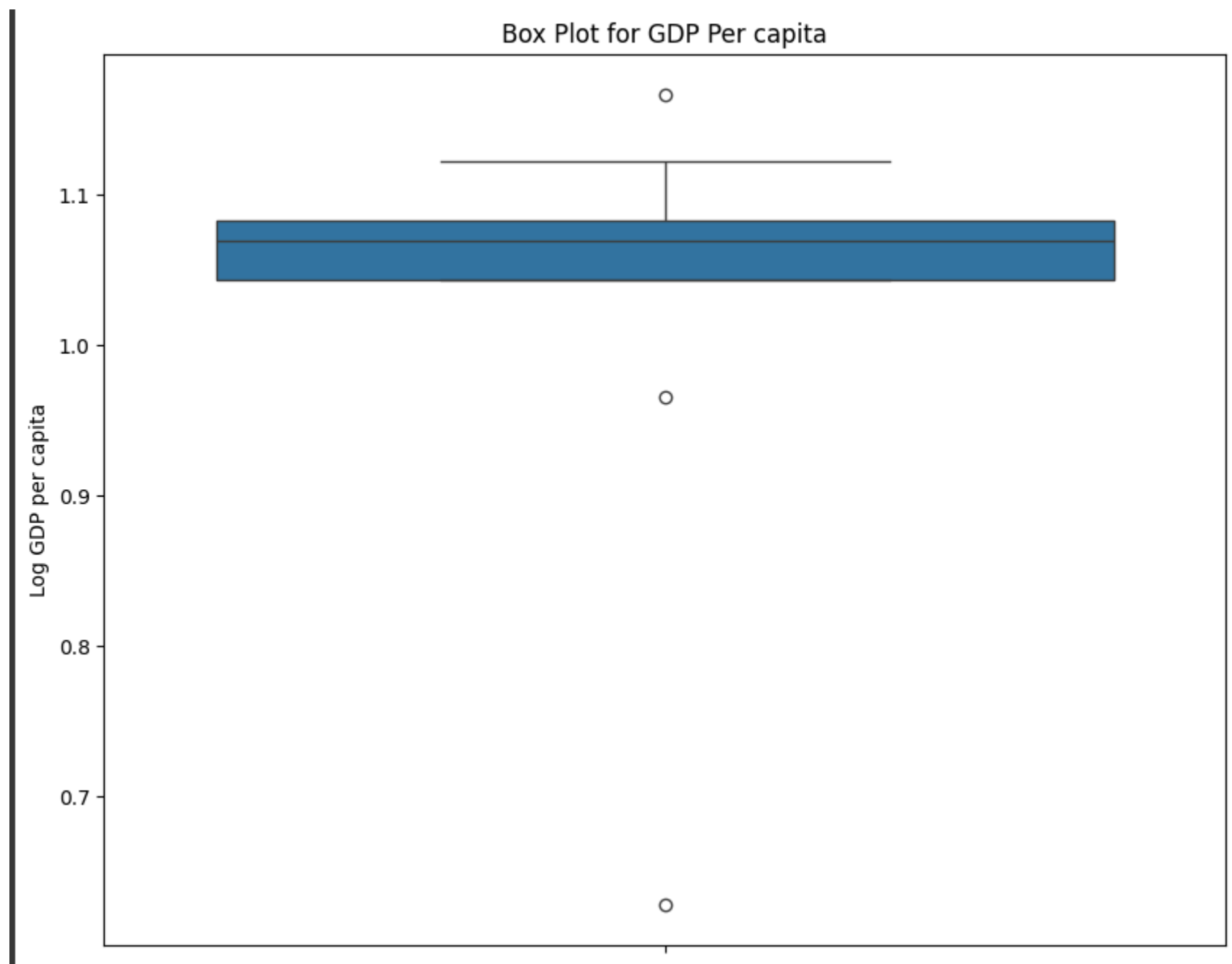


Fig 3.4: Box plot of the South Asian Countries based on their GDP per capita

Unlike the score, the median for `Log GDP per Capita` is closer to the 3rd quartile. The missing bottom whisker suggest that there are no values that extend the third quartile. The plot still shows 3 outliers, but majority fall under the  $1.5 \times \text{iqr}$  from  $q_1$ , indicating much lower gdp per capita from the average, which ranges from  $\sim 1.05$  -  $\sim 1.08$

### 3.2 Define outliers using the $1.5 \times \text{IQR}$ rule

Country name	
0	Nepal
2	India
4	Afghanistan
dtype: object	

Fig 3.5: Image of all the outliers in the dataframe.

### 3.3 Create a scatter plot with GDP per Capita on the x-axis and Score on the y-axis, highlighting outliers in a different color

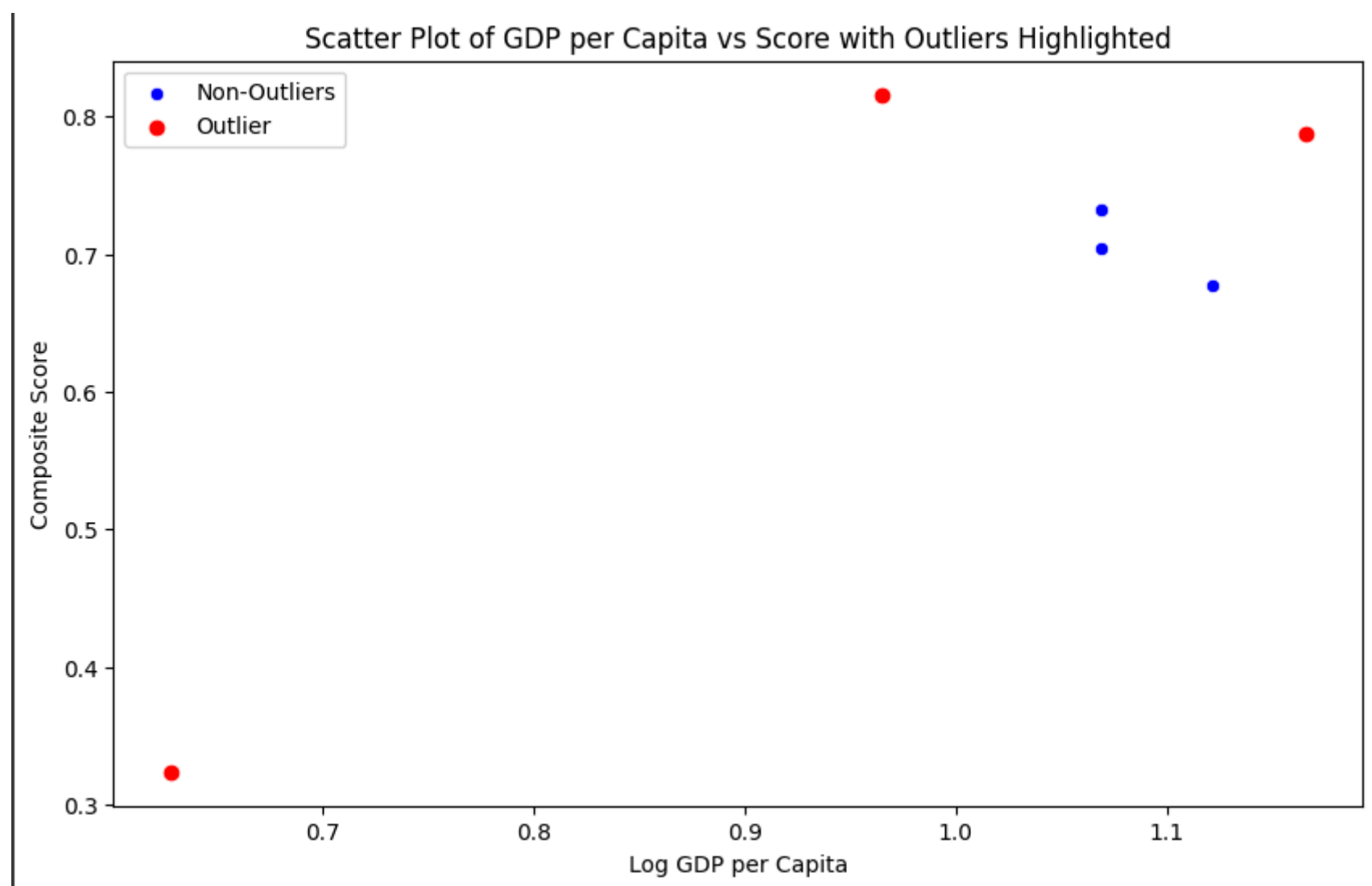


Fig 3.6: Scatter Plot of GDP and Happiness Scores with outliers highlighted.

The plot shows 3 outliers, Nepal, India and Afghanistan. Nepal and India are outliers since they have noticeably higher Composite score for their gdp per capita where as Afghanistan is exceptionally low in this scale

### 3.4 Discuss the characteristics of these outliers and their potential impact on regional averages

The outliers on the upper end of the spectrum are exceptionally high, whereas the one outlier on the lower end can pull down regional averages. These outliers skew the averages making the overall region seem happier or less happy than they truly are.

4.1. Choose two metrics (e.g., Freedom to Make Life Choices and Generosity) and calculate their correlation {pearson correlation} with the Score for South Asian countries.

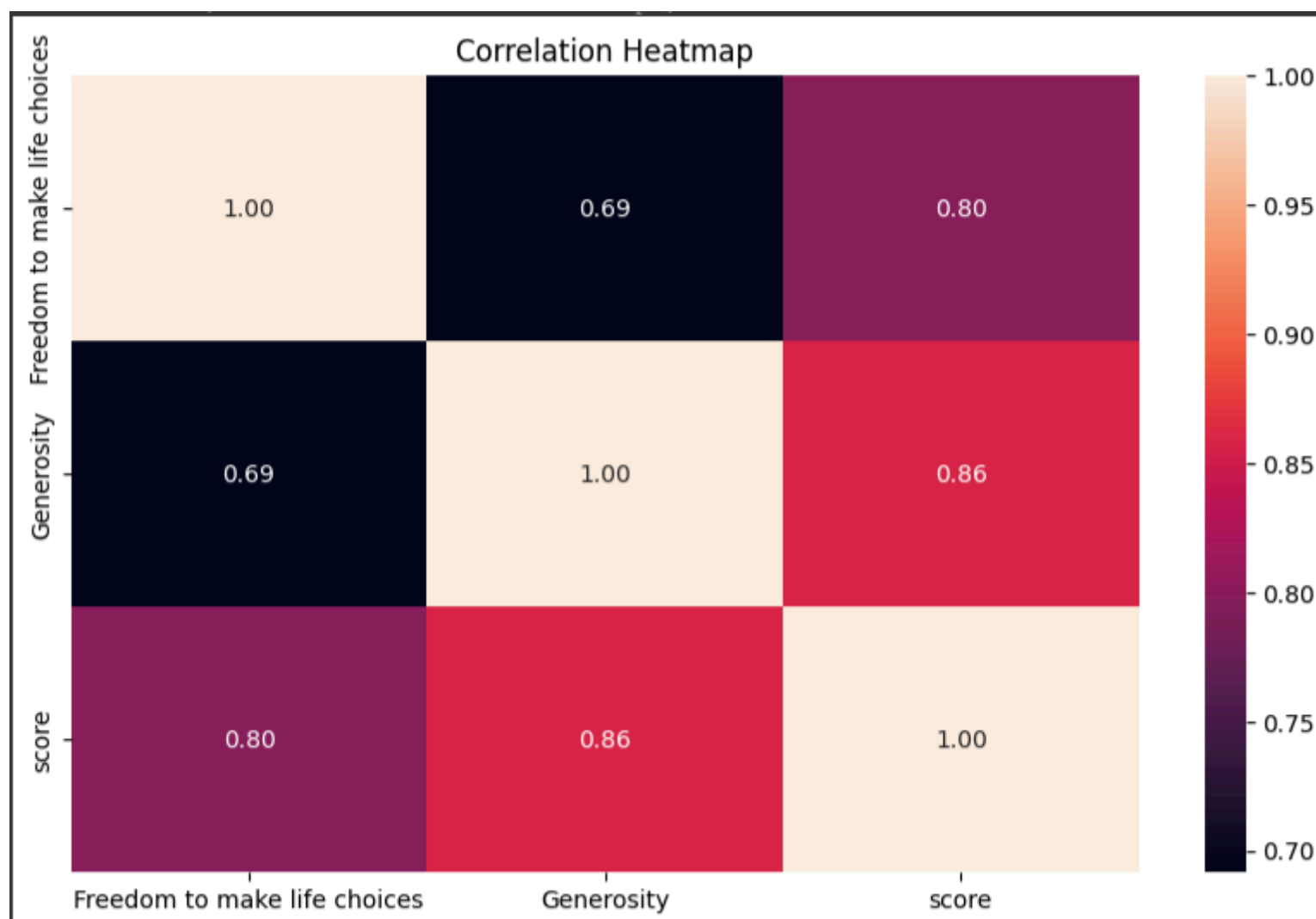


Fig 3.7: Correlation between the Happiness Score and the Freedom to Make Life Choices,Generosity.

4.2. Create scatter plots with trendlines for these metrics against the Score



Fig 3.8: Scatter Plot With Trendlines for Happiness Score and Freedom To Make Life Choices.

As Happiness Score increases, the freedom to make life choices also increases, so there is a positive correlation between these two metrics. The confidence interval is largest at the lower points which suggests a big uncertainty, partly due to the fact that there are not enough data points around that region. Most of the data seems to be clustered towards the ~4.0 happiness score.

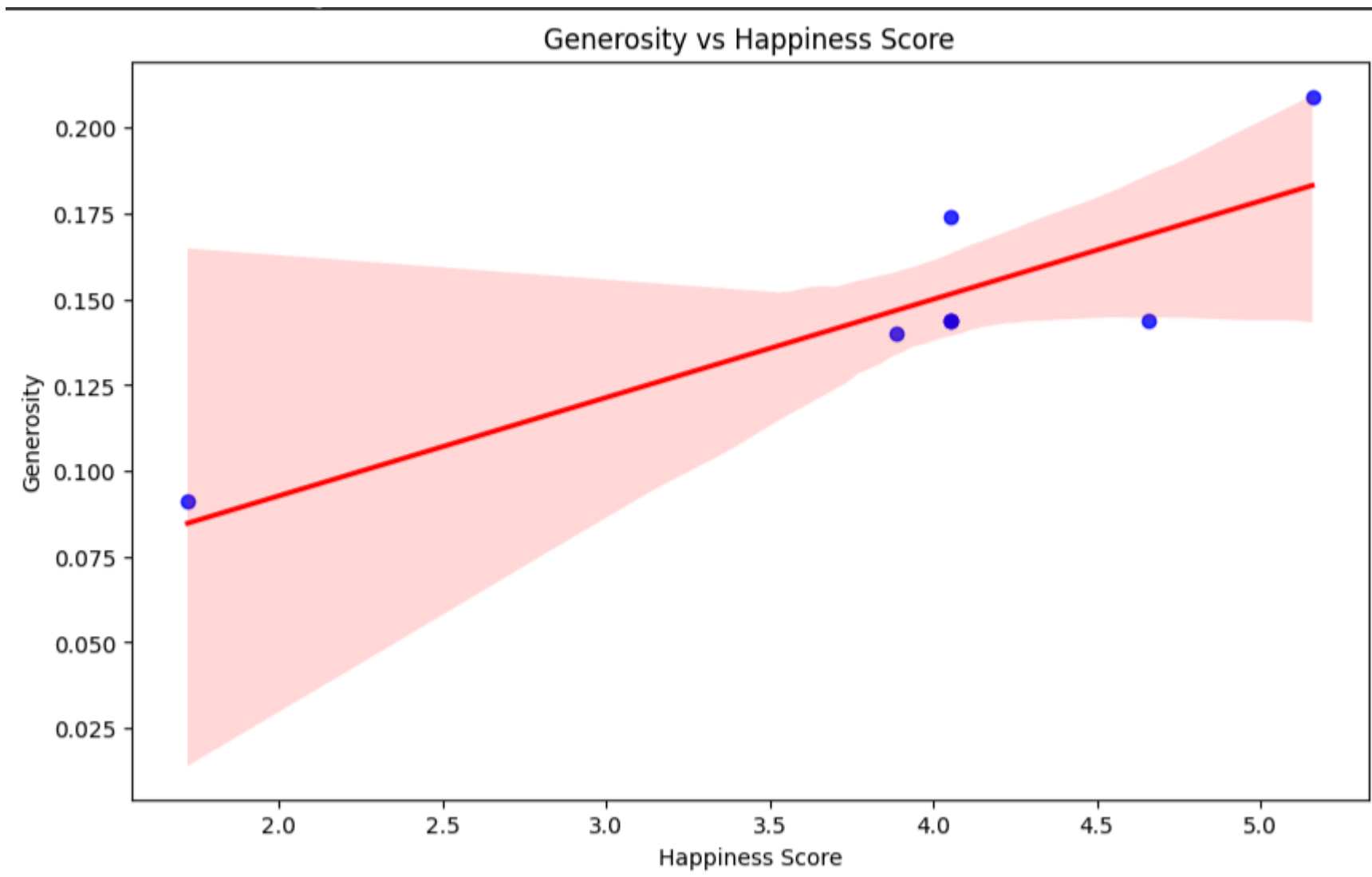


Fig 3.9: Scatter plot with trendlines for Happiness Score and Generosity

There also seems to be a positive correlation between the Happiness Score and Generosity. Similarly, the confidence interval is larger at the lower values which suggests variability, which also is partly due to the fact that there aren't enough points on that end of the spectrum. The data is also relatively clustered towards the ~4.0 happiness score.

#### 4.3. Identify and discuss the strongest and weakest relationships between these metrics and the Score for South Asian countries.

Amongst Generosity and the Freedom to make life choices, Generosity seems to be more related with the Happiness Score for south asian countries. Although both these metrics have a strong correlation, Generosity seems to be a closer metric for happiness, which could reflect a deeper social and cultural bond amongst the South Asian countries. Freedom to make life choices while also having a strong relationship with the happiness, it suggests that South Asian countries value generosity more than the freedom to make their life choices



5.1. Add a new column, GDP-Score Gap, which is the difference between GDP per Capita and the Score for each South Asian country

	0	2	5	6	7	1	3	4
Country name	Nepal	India	Bhutan	Maldives	Srilanka	Pakistan	Bangladesh	Afghanistan
score	5.158	4.054	4.054	4.054	4.054	4.657	3.886	1.721
Log GDP per capita	0.965	1.166	1.069	1.069	1.069	1.069	1.122	0.628
Social support	0.99	0.653	0.6	0.6	0.6	0.6	0.249	0.0
Healthy life expectancy	0.443	0.417	0.417	0.417	0.417	0.321	0.513	0.242
Freedom to make life choices	0.653	0.767	0.653	0.653	0.653	0.542	0.775	0.0
Generosity	0.209	0.174	0.144	0.144	0.144	0.144	0.14	0.091
Perceptions of corruption	0.115	0.122	0.115	0.115	0.115	0.074	0.167	0.088
Dystopia + residual	1.783	0.756	0.919	0.919	0.919	1.907	0.919	0.672
Happiness_Category	Medium	Medium	Medium	Medium	Medium	Medium	Low	Low
Composite Score	0.8159	0.7874	0.7327	0.7327	0.7327	0.7039	0.6774	0.3238
GDP Score Gap	-4.193	-2.888	-2.985	-2.985	-2.985	-3.588	-2.764	-1.093

Fig 4.0: Result of adding the column GDP-Score gap

5.2. Rank the South Asian countries by this gap in both ascending and descending order.

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category	Composite Score	GDP Score Gap
0	Nepal	5.158	0.965	0.990	0.443	0.653	0.209	0.115	1.783	Medium	0.8159	-4.193
1	Pakistan	4.657	1.069	0.600	0.321	0.542	0.144	0.074	1.907	Medium	0.7039	-3.588
5	Bhutan	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985
6	Maldives	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985
7	Srilanka	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985
2	India	4.054	1.166	0.653	0.417	0.767	0.174	0.122	0.756	Medium	0.7874	-2.888
3	Bangladesh	3.886	1.122	0.249	0.513	0.775	0.140	0.167	0.919	Low	0.6774	-2.764
4	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	0.088	0.672	Low	0.3238	-1.093
	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category	Composite Score	GDP Score Gap
4	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	0.088	0.672	Low	0.3238	-1.093
3	Bangladesh	3.886	1.122	0.249	0.513	0.775	0.140	0.167	0.919	Low	0.6774	-2.764
2	India	4.054	1.166	0.653	0.417	0.767	0.174	0.122	0.756	Medium	0.7874	-2.888
5	Bhutan	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985
6	Maldives	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985
7	Srilanka	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985
1	Pakistan	4.657	1.069	0.600	0.321	0.542	0.144	0.074	1.907	Medium	0.7039	-3.588
0	Nepal	5.158	0.965	0.990	0.443	0.653	0.209	0.115	1.783	Medium	0.8159	-4.193

Fig 4.1: Image after ranking the countries by both ascending and descending order based on the GDP-Score gap

5.3 Highlight the top 3 countries with the largest positive and negative gaps using a bar chart.

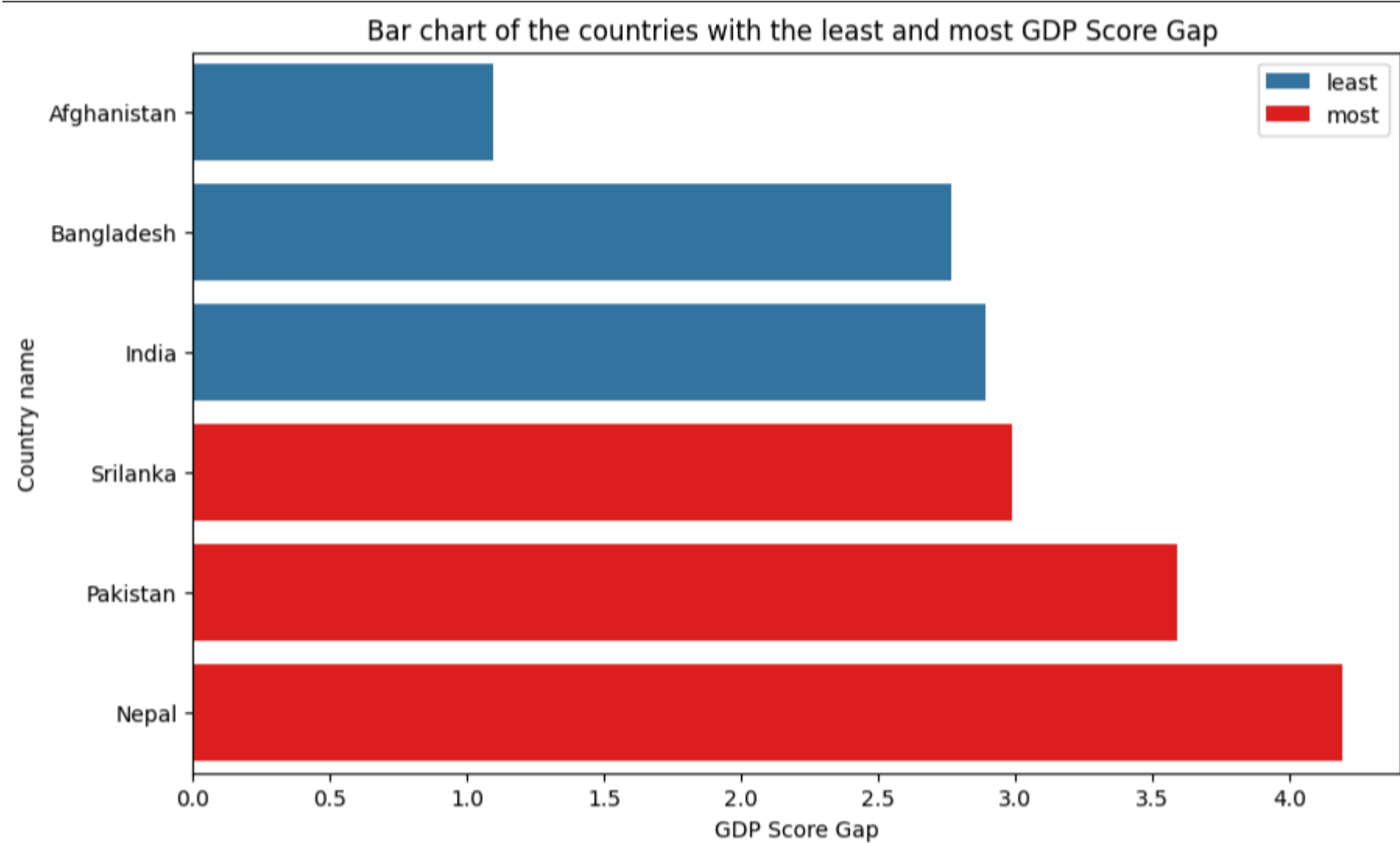


Fig 4.2: Countries with the largest and smallest GDP-Score gap.

5.4 Analyze the reasons behind these gaps and their implications for South Asian countries.

The GDP score gap helps us understand the difference between a country's GDP and its happiness score. Afghanistan having the lowest GDP per capita also has the lowest happiness score, which reflects the big challenges the country is facing. Nepal, although having a relatively low GDP, has a significantly higher happiness score, indicating that the country despite its financial troubles is still performing good in terms of happiness. Pakistan and Bangladesh have similar GDP scores, but noticeably different happiness scores. Bhutan, Maldives and Sri Lanka have the same scores which may raise concerns but its only due to the fact that their data were not present in the World Health Report but were rather imputed from the median. India’s gap suggests that factors like social support and healthy life expectancy could be suffering

Problem - 3 - Comparative Analysis:

1.1 Similar in Task - 1 of Problem 2 create a dataframe from middle eastern countries. For hint use the following list:

Description of the dataset:

	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	5.455071	1.473679	1.172536	0.555321	0.578429	0.139321	0.144536	1.441821
std	1.305064	0.328182	0.220213	0.096573	0.187109	0.050831	0.058561	0.641766
min	2.707000	0.671000	0.577000	0.293000	0.173000	0.059000	0.029000	-0.073000
25%	4.983750	1.390625	1.143000	0.551125	0.467000	0.119625	0.121125	1.203875
50%	5.562500	1.433250	1.200750	0.559500	0.617000	0.138750	0.147500	1.692750
75%	6.435250	1.711000	1.270125	0.569000	0.641000	0.150000	0.184000	1.777750
max	7.341000	1.983000	1.513000	0.740000	0.827000	0.235000	0.258000	2.298000

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

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category
4	Israel	7.341	1.8030	1.5130	0.7400	0.641	0.1530	0.1930	2.2980	High
12	Kuwait	6.951	1.8450	1.3640	0.6610	0.827	0.2000	0.1720	1.8840	High
21	United Arab Emirates	6.733	1.9830	1.1640	0.5630	0.815	0.2090	0.2580	1.7410	High
27	Saudi Arabia	6.594	1.8420	1.3610	0.5110	0.787	0.1140	0.1880	1.7900	High
61	Bahrain	5.959	1.4315	1.2375	0.5495	0.641	0.1365	0.1205	1.6445	Medium
91	Iraq	5.166	1.2490	0.9960	0.4980	0.425	0.1410	0.0480	1.8090	Medium
99	Iran	4.923	1.4350	1.1360	0.5710	0.366	0.2350	0.1230	1.0570	Medium
124	Jordan	4.186	1.2620	0.9830	0.5940	0.593	0.0590	0.1890	0.5040	Medium
132	Yemen	3.561	0.6710	1.2810	0.2930	0.362	0.0800	0.1130	0.7600	Low
141	Lebanon	2.707	1.3770	0.5770	0.5560	0.173	0.0680	0.0290	-0.0730	Low

Fig 4.3: Filtered Dataset of only the Middle Eastern Countries from the original dataset.

The results indicate that only 10 out of the 14 countries are present in the dataset. To address the missing countries, the values were calculated with the median of the dataset. While this doesn’t accurately represent the actual situation in those countries, it merely serves as a strategy to counter the missing values.

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category
0	Israel	7.3410	1.80300	1.51300	0.7400	0.641	0.15300	0.1930	2.29800	High
1	Kuwait	6.9510	1.84500	1.36400	0.6610	0.827	0.20000	0.1720	1.88400	High
2	United Arab Emirates	6.7330	1.98300	1.16400	0.5630	0.815	0.20900	0.2580	1.74100	High
3	Saudi Arabia	6.5940	1.84200	1.36100	0.5110	0.787	0.11400	0.1880	1.79000	High
4	Bahrain	5.9590	1.43150	1.23750	0.5495	0.641	0.13650	0.1205	1.64450	Medium
5	Iraq	5.1660	1.24900	0.99600	0.4980	0.425	0.14100	0.0480	1.80900	Medium
6	Iran	4.9230	1.43500	1.13600	0.5710	0.366	0.23500	0.1230	1.05700	Medium
7	Jordan	4.1860	1.26200	0.98300	0.5940	0.593	0.05900	0.1890	0.50400	Medium
8	Yemen	3.5610	0.67100	1.28100	0.2930	0.362	0.08000	0.1130	0.76000	Low
9	Lebanon	2.7070	1.37700	0.57700	0.5560	0.173	0.06800	0.0290	-0.07300	Low
10	Oman	5.5625	1.43325	1.20075	0.5595	0.617	0.13875	0.1475	1.69275	High
11	Palestine	5.5625	1.43325	1.20075	0.5595	0.617	0.13875	0.1475	1.69275	High
12	Qatar	5.5625	1.43325	1.20075	0.5595	0.617	0.13875	0.1475	1.69275	High
13	Syria	5.5625	1.43325	1.20075	0.5595	0.617	0.13875	0.1475	1.69275	High

Fig 4.4: Updated Dataset of the Middle Eastern Countries with missing countries imputed.

2.1 Calculate the mean, Standard deviation of the score for both South Asia and Middle East.

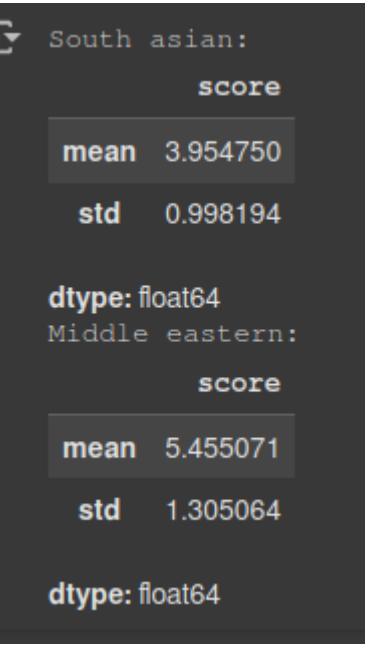


Fig 4.5: Result of Calculating the mean and standard deviation of both regions.

2.2 Which region has higher happiness Scores on average?

From the above results, we can conclude that Middle Eastern countries are happier on average than the South Asian countries.

3.1 Identify the top 3 and bottom 3 countries in each region based on the score.

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category	Composite Score	GDP Score Gap	Region
0	Nepal	5.158	0.965	0.990	0.443	0.653	0.209	0.115	1.783	Medium	0.8159	-4.193	South Asian
1	Pakistan	4.657	1.069	0.600	0.321	0.542	0.144	0.074	1.907	Medium	0.7039	-3.588	South Asian
2	India	4.054	1.166	0.653	0.417	0.767	0.174	0.122	0.756	Medium	0.7874	-2.888	South Asian
7	Srilanka	4.054	1.069	0.600	0.417	0.653	0.144	0.115	0.919	Medium	0.7327	-2.985	South Asian
3	Bangladesh	3.886	1.122	0.249	0.513	0.775	0.140	0.167	0.919	Low	0.6774	-2.764	South Asian
4	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	0.088	0.672	Low	0.3238	-1.093	South Asian
	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category			Region
0	Israel	7.341	1.803	1.513		0.740	0.641	0.153	0.193	2.298	High		Middle Eastern
1	Kuwait	6.951	1.845	1.364		0.661	0.827	0.200	0.172	1.884	High		Middle Eastern
2	United Arab Emirates	6.733	1.983	1.164		0.563	0.815	0.209	0.258	1.741	High		Middle Eastern
7	Jordan	4.186	1.262	0.983		0.594	0.593	0.059	0.189	0.504	Medium		Middle Eastern
8	Yemen	3.561	0.671	1.281		0.293	0.362	0.080	0.113	0.760	Low		Middle Eastern
9	Lebanon	2.707	1.377	0.577		0.556	0.173	0.068	0.029	-0.073	Low		Middle Eastern

Fig 4.6: Top and bottom 3 countries in each region based on the Happiness Score.

3.2 Plot bar charts comparing these charts.

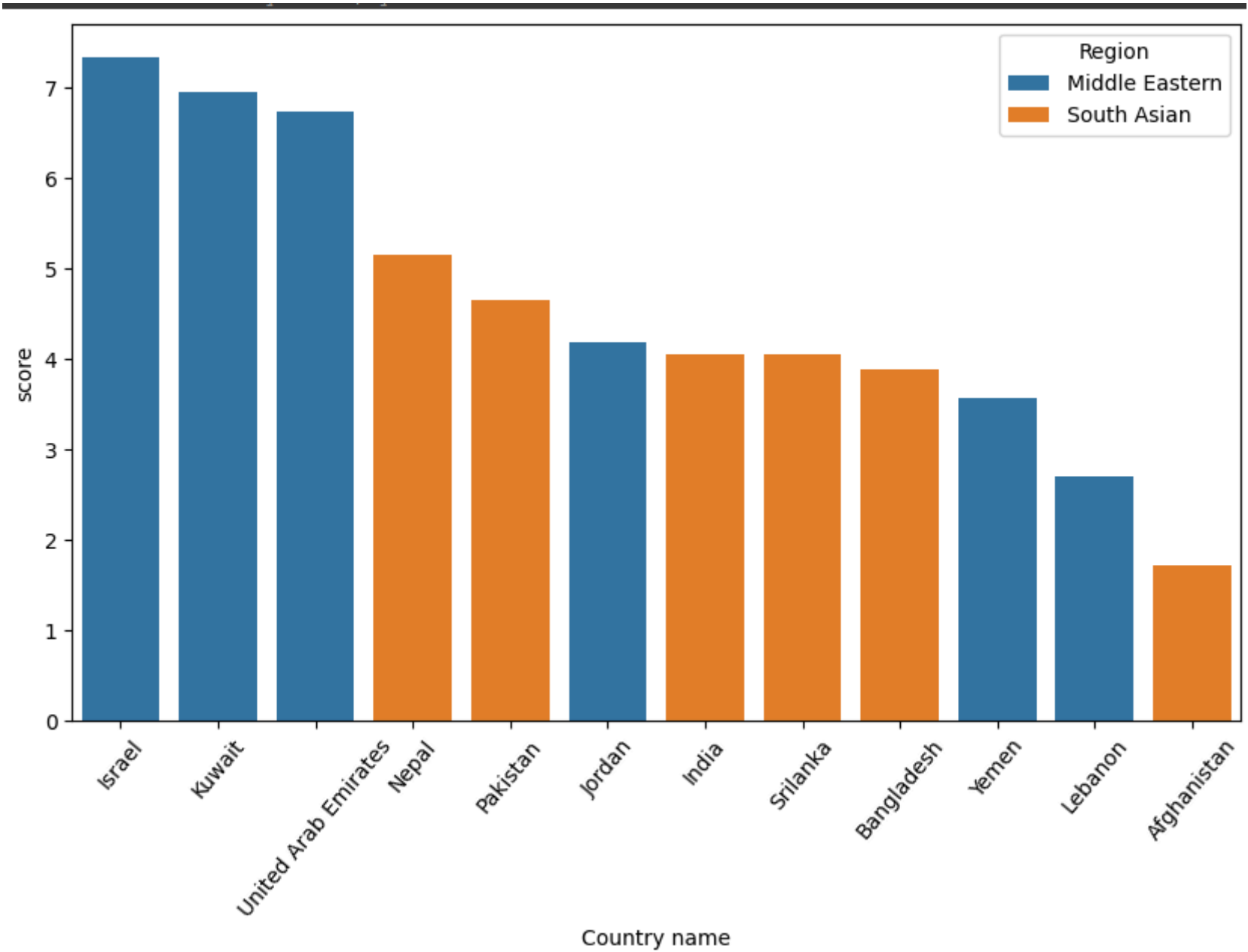


Fig 4.7: Bar plot comparing the three happiest and unhappiest countries in each region.

4.1 Compare key metrics like GDP per Capita, Social Support, and Healthy Life Expectancy between the regions using grouped bar charts.

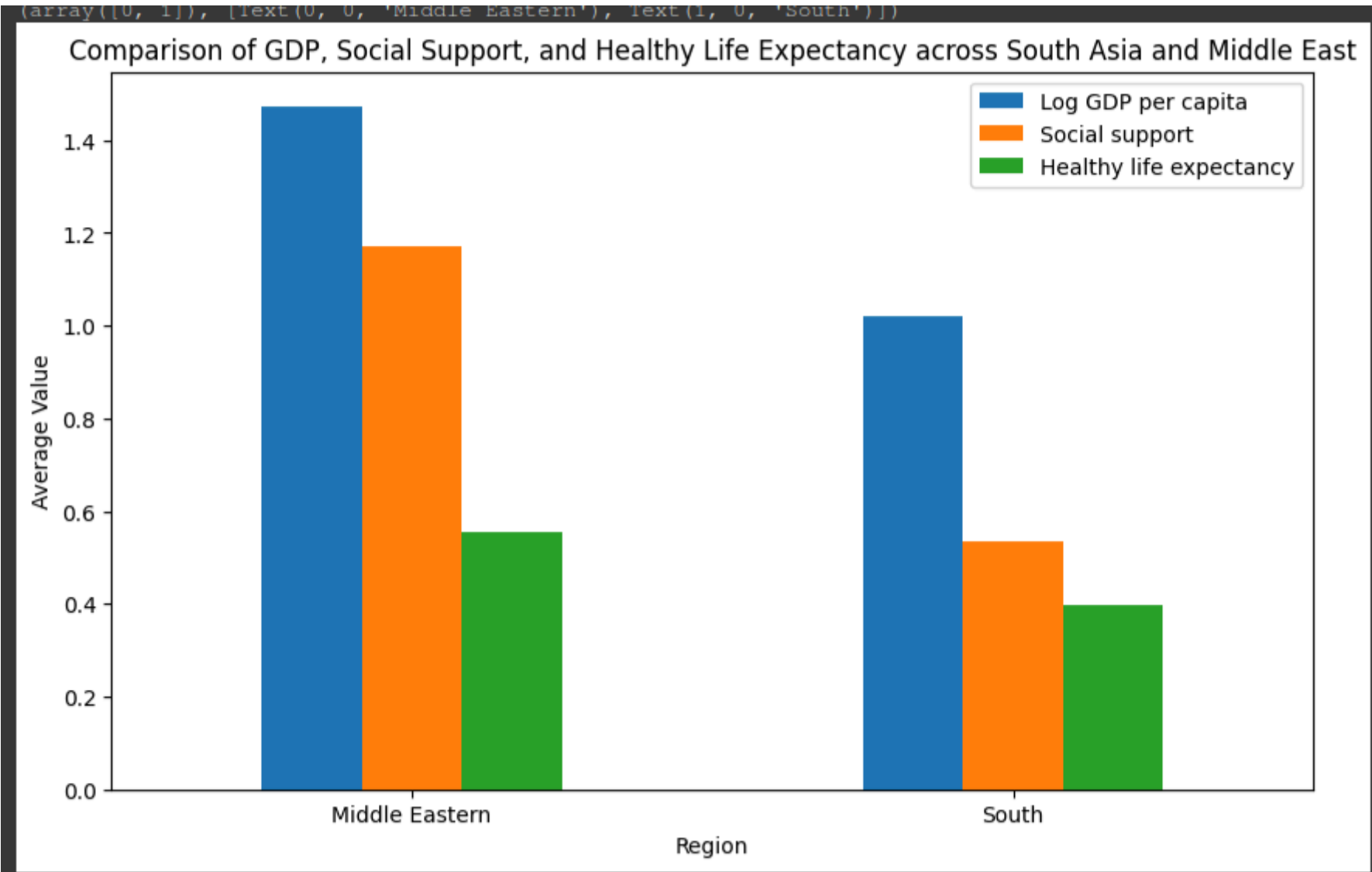


Fig 4.8: Bar chart grouped by region, comparing GDP per capita, Social Support and Healthy Life Expectancy between the regions.

Log GDP per capita	0.586688
score	2.026875
Healthy life expectancy	0.187562
Social support	0.682813

Fig 4.9: calculation of the gap between these metrics amongst these regions.

The chart shows that the Middle East is clearly better in every metric between these two regions. The difference in GDP is~ 0.58, the difference in Social support is ~0.68 and the difference between the Healthy life expectancy is only about ~0.18.

#### 4.2 Which metrics show the largest disparity between the two regions?

The largest disparity can be seen with the Social support, with a disparity of ~0.68

#### 5.1 Compute the range (max - min) and coefficient of variation (CV) for Score in both regions.

	Region	Range	CV (%)
0	South Asia	3.437	25.240372
1	Middle East	4.634	23.923870

Fig 5.0: Range and Coefficient of Variation for Happiness Score amongst the regions.

#### 5.2 Which region has greater variability in happiness?

South Asia seems to have a greater variability in happiness.

6.1 Analyze the correlation of Score with other metrics Freedom to Make Life Choices, and Generosity within each region.

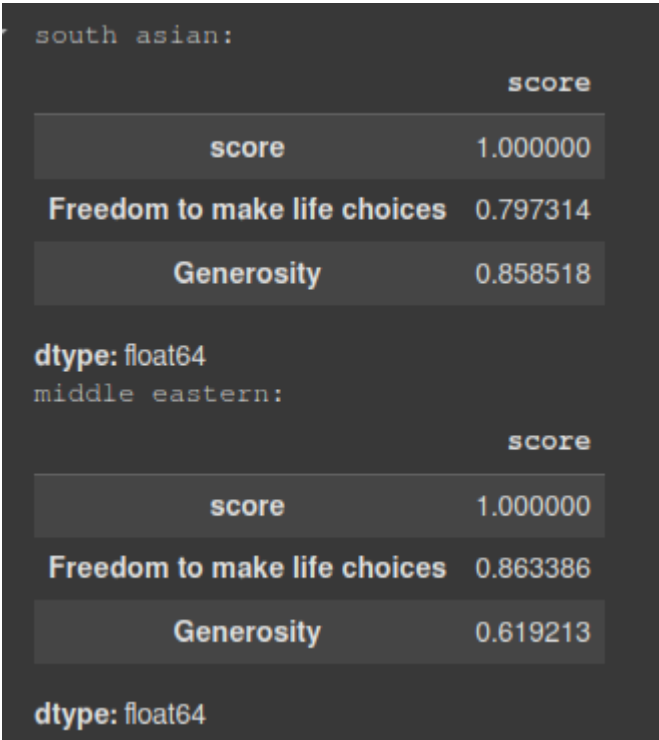


Fig 5.1: Correlation between the metrics

The results indicate that Generosity seems to be more correlated with the happiness score in the South Asian countries as compared to the Middle Eastern countries, while the opposite is true for the correlation between the freedom to make life choices and the happiness score between these two regions

6.2 Create scatter plots to visualize and interpret the relationships

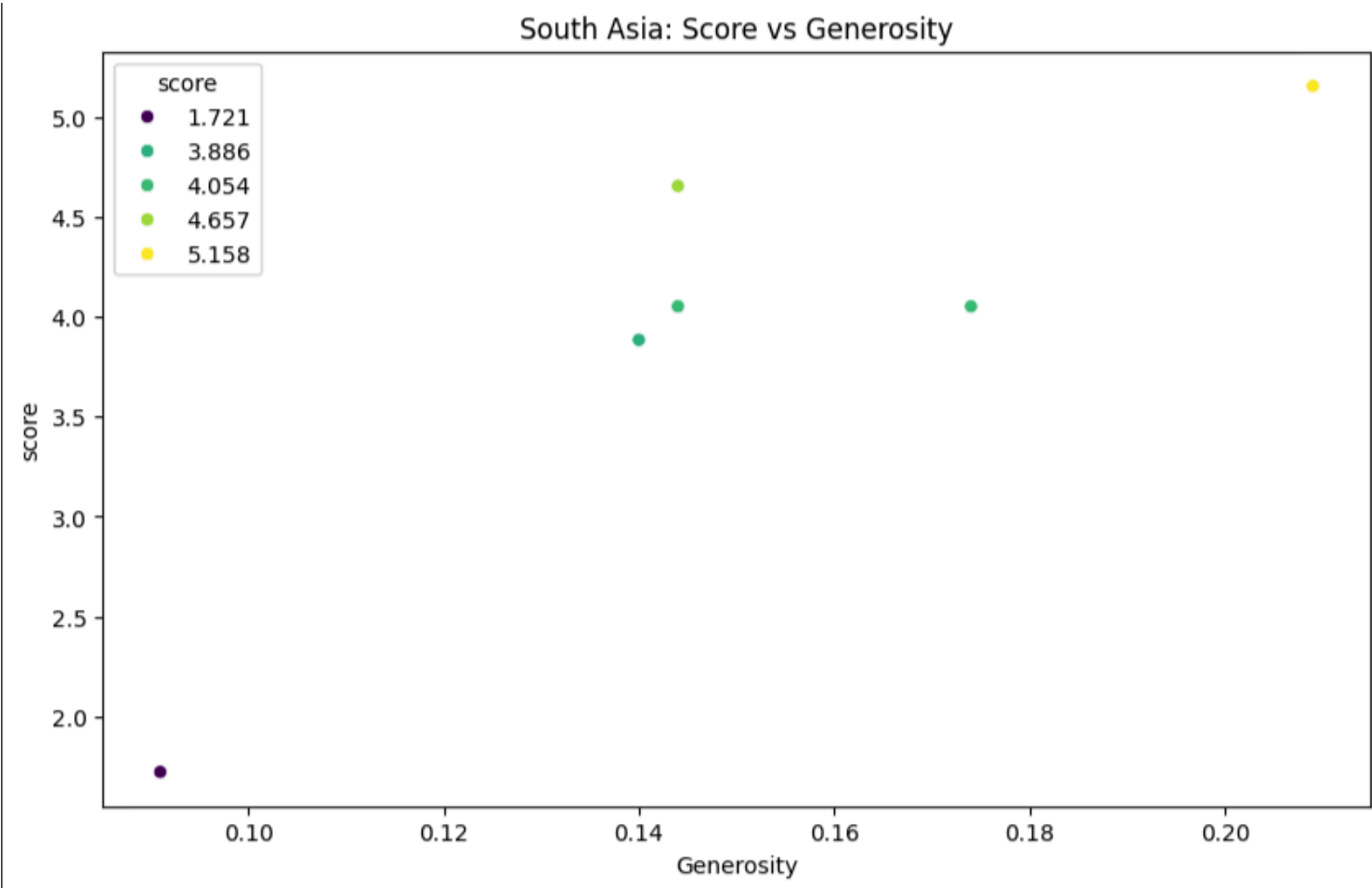


Fig 5.2: Scatter plot between Generosity and Happiness Scores For South Asia

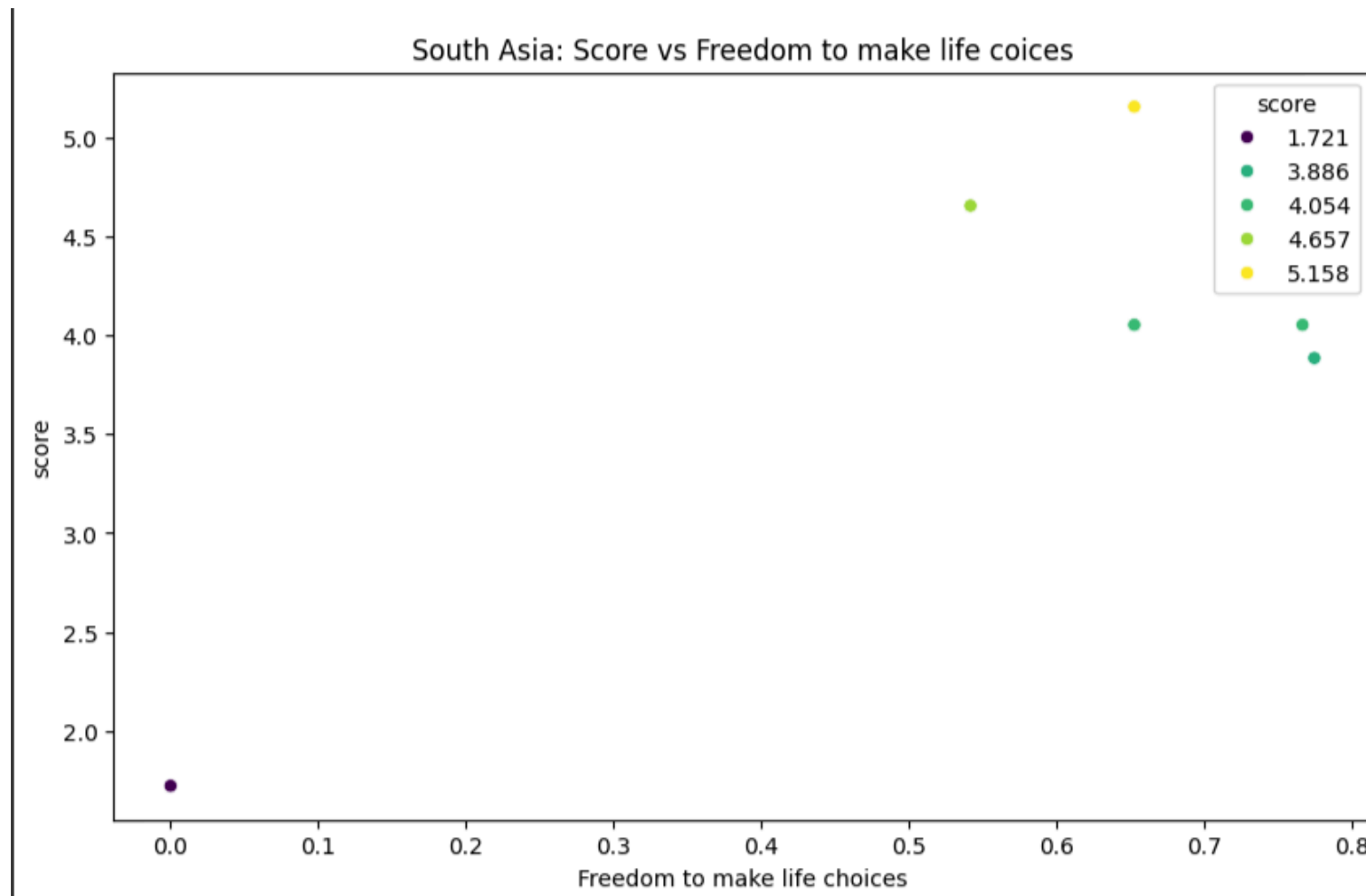


Fig 5.3: Scatter plot between Freedom to make life choices and Happiness Scores For South Asia

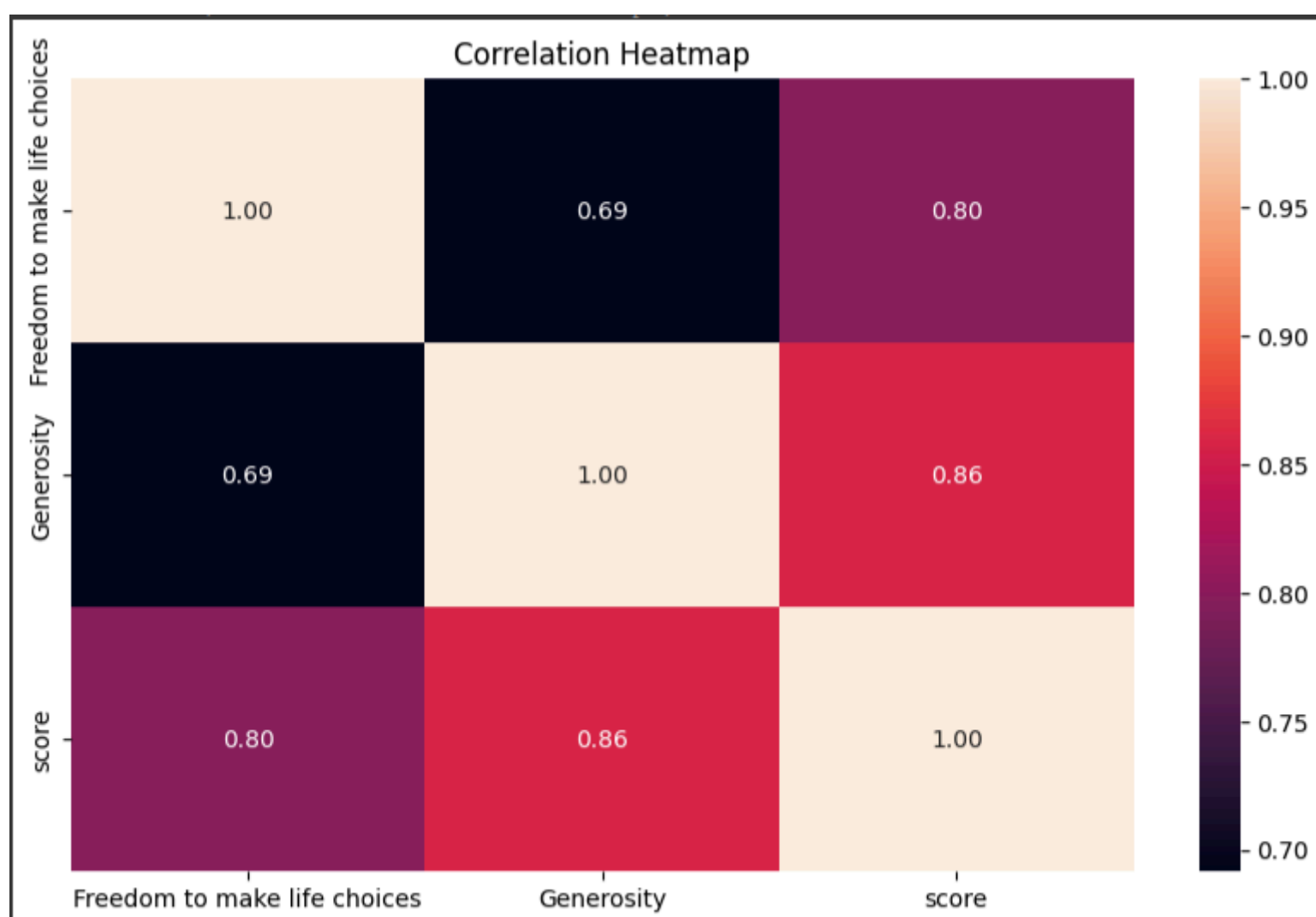


Fig 5.4: Heatmap of the correlations between these metrics for south asia



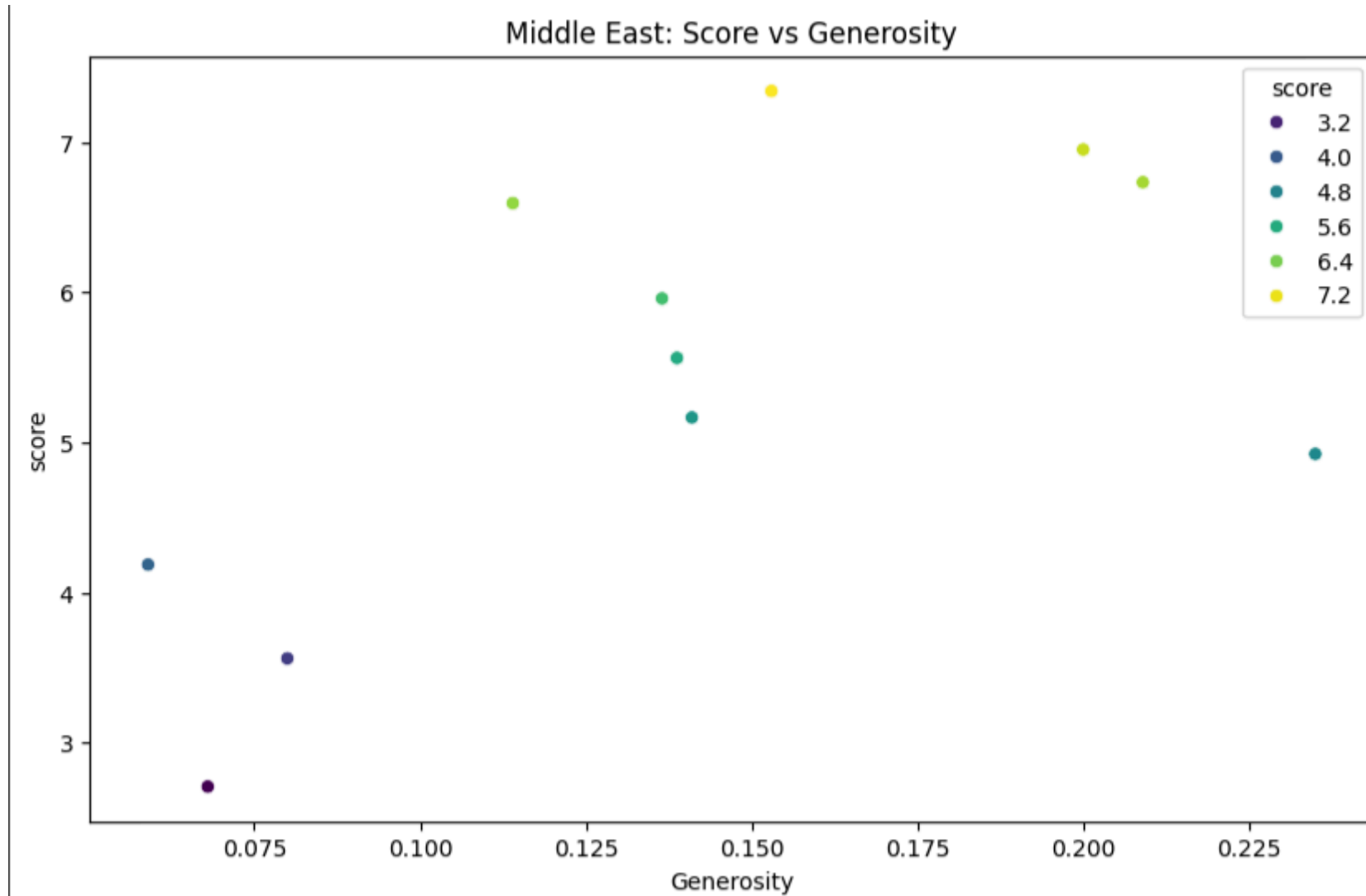


Fig 5.5: Scatter plot for Middle East between Happiness score and Generosity

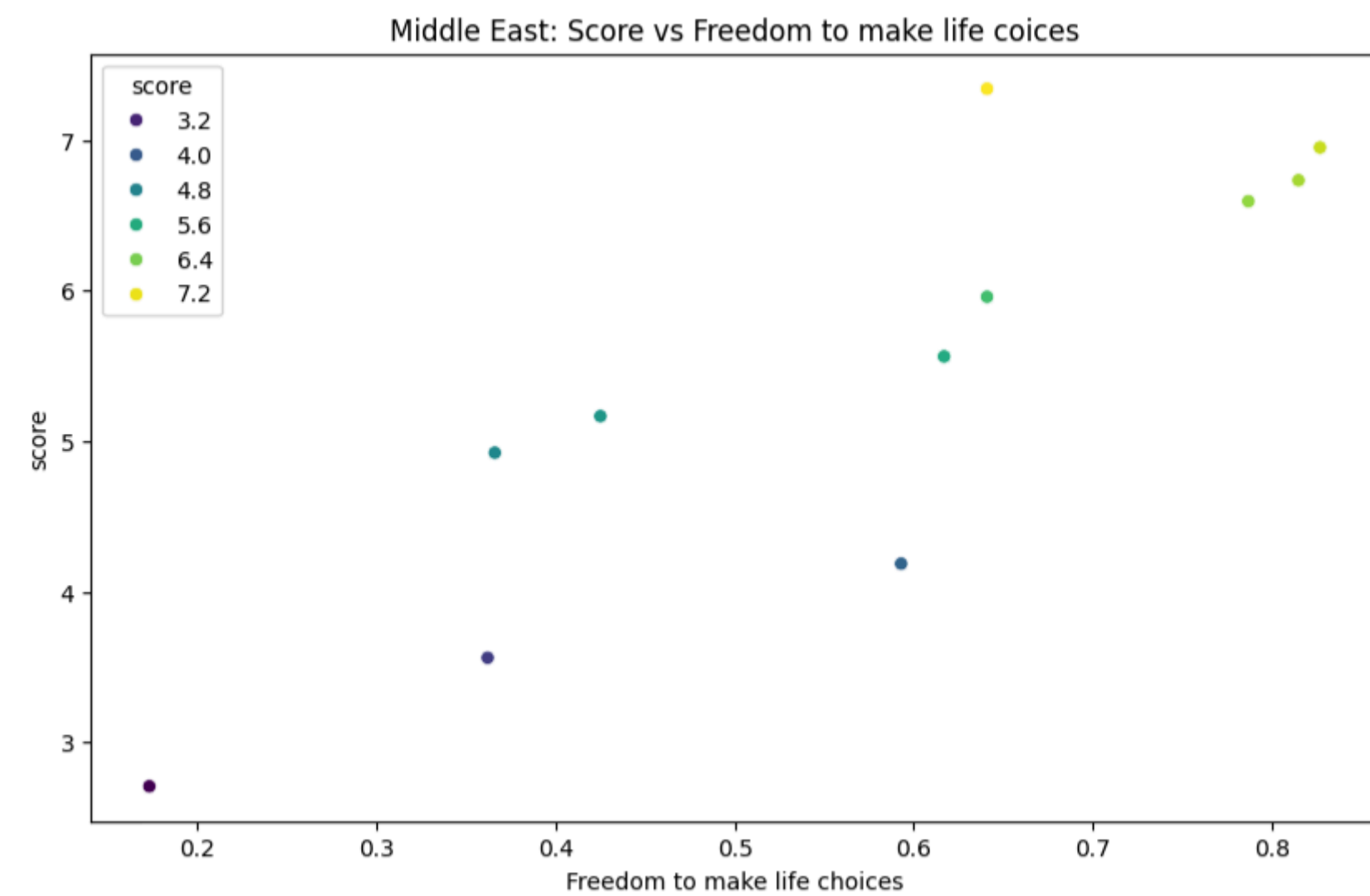


Fig 5.5: Scatter plot for Middle East between Happiness score and the Freedom to make life choices.



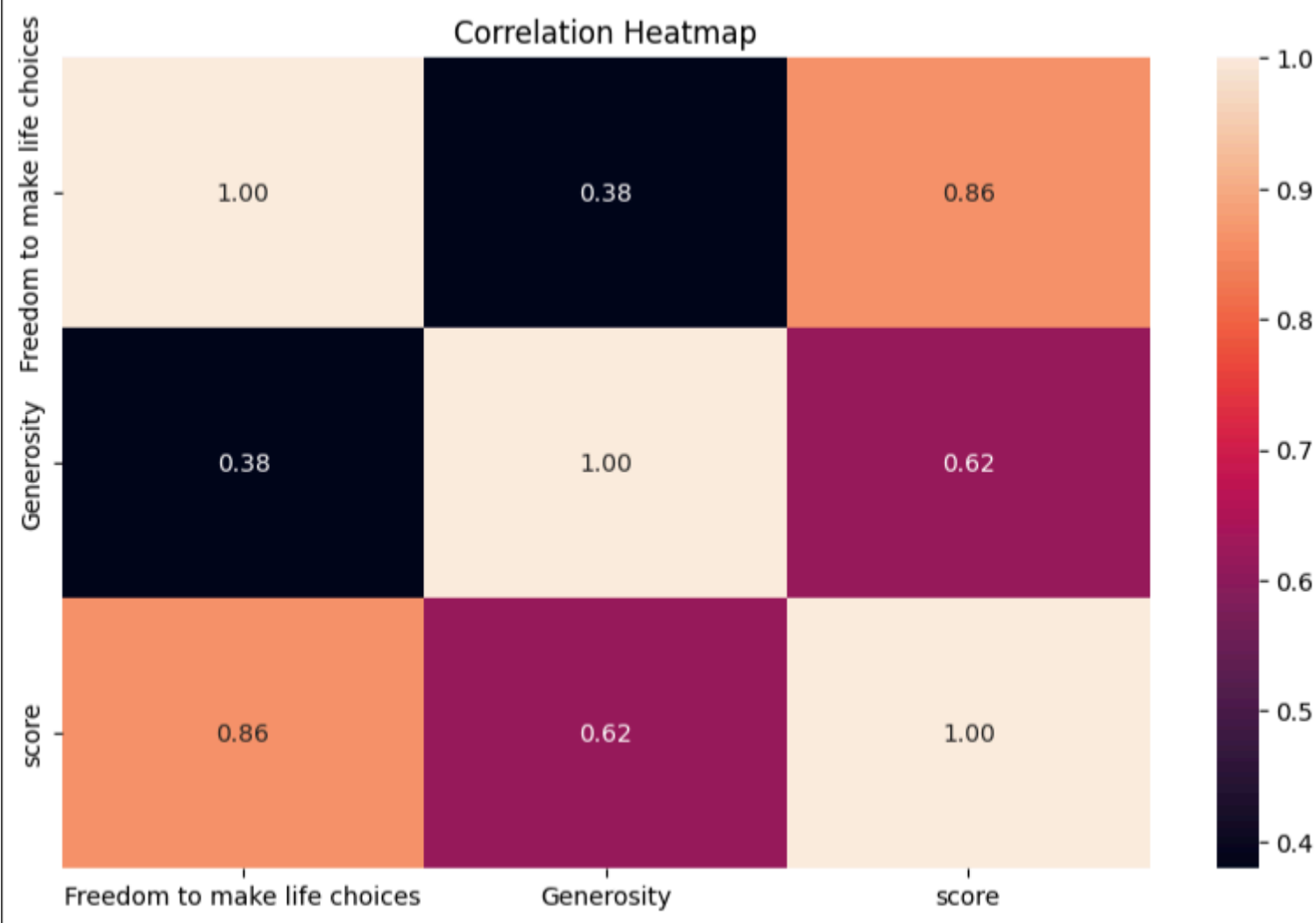


Fig 5.6: Heatmap of Middle East for the compared metrics.

7.1 Identify outlier countries in both regions based on Score and GDP per Capita.

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category	Composite Score	GDP Score Gap	Region
0	Nepal	5.158	0.965	0.990	0.443	0.653	0.209	0.115	1.783	Medium	0.8159	-4.193	South
1	Pakistan	4.657	1.069	0.600	0.321	0.542	0.144	0.074	1.907	Medium	0.7039	-3.588	South
2	India	4.054	1.166	0.653	0.417	0.767	0.174	0.122	0.756	Medium	0.7874	-2.888	South
4	Afghanistan	1.721	0.628	0.000	0.242	0.000	0.091	0.088	0.672	Low	0.3238	-1.093	South
	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Happiness_Category	Region		
8	Yemen	3.561	0.671	1.281	0.293	0.362	0.080	0.113	0.760	Low	Middle Eastern		
9	Lebanon	2.707	1.377	0.577	0.556	0.173	0.068	0.029	-0.073	Low	Middle Eastern		

Fig 5.7: Result of identifying outliers in both the regions.

## 7.2 Plot these outliers and discuss their implications

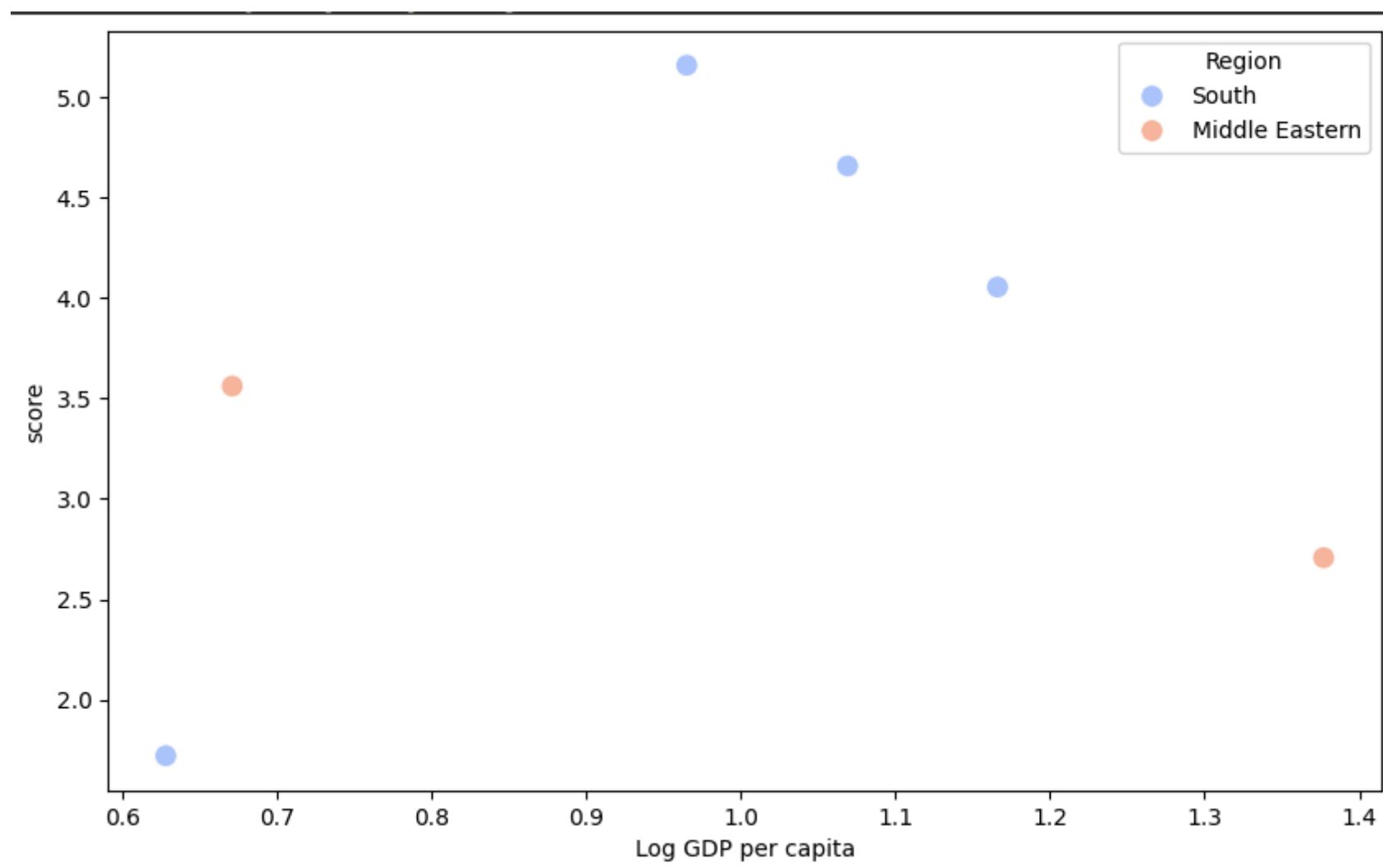


Fig 5.8: Scatter plot of combined outliers of both countries.

These outliers skew the data making these regions appear better / worse in different metrics which can mislead the interpretation of the actual situation in these regions. Especially in the case of South Asia, where 3 out of 8 countries are outliers, which is roughly ~37.5% of the data. These points make the data skewed which can exaggerate or minimize the true trends in the data, they can mess averages and any important decisions made from these data can lead to unwanted results. Thus, outliers, especially when they make up a significant portion of the available data, can significantly impact the overall result of the region's score.

## 8.1 Create boxplots comparing the distribution of Score between South Asia and the Middle East

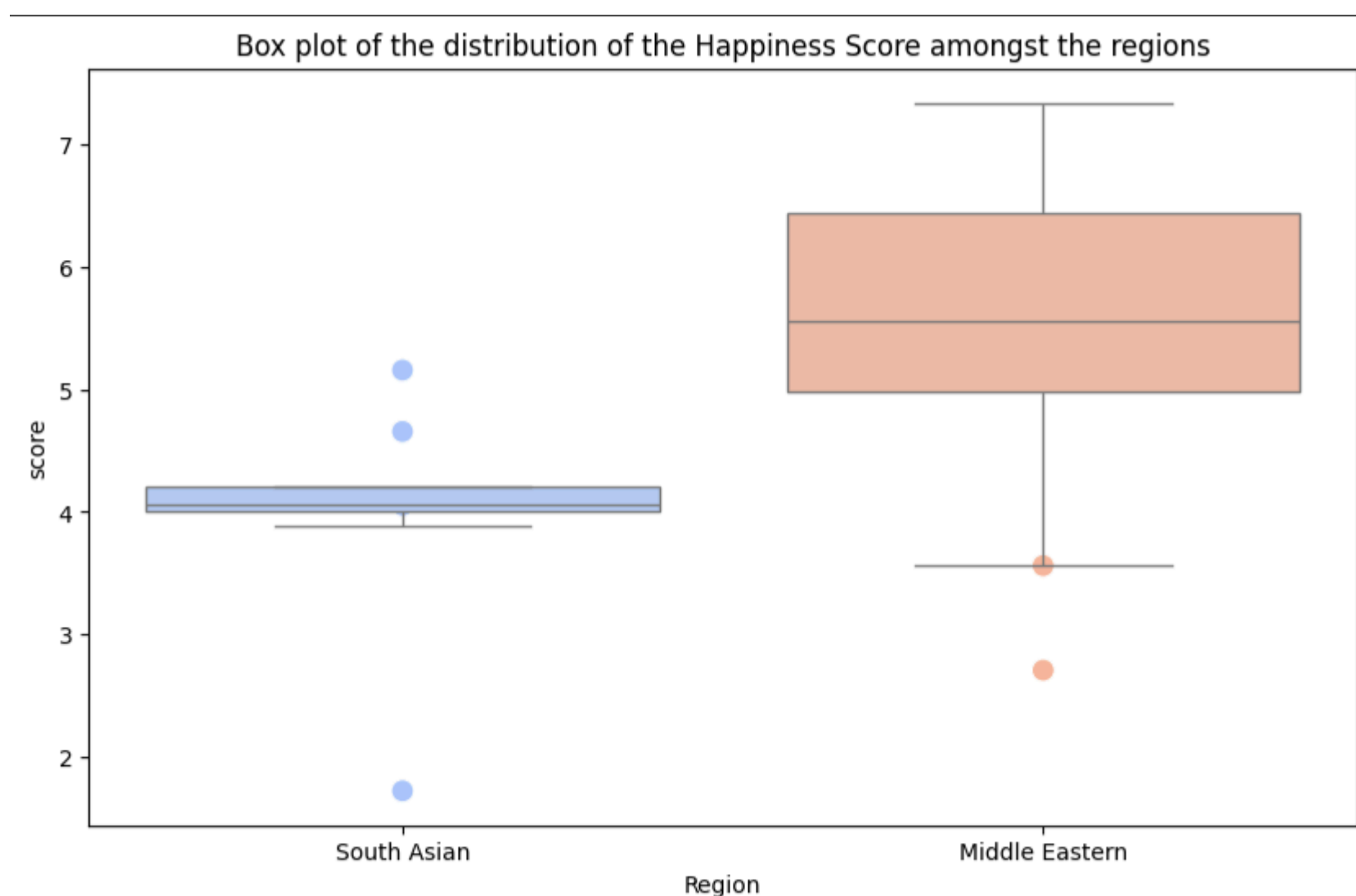


Fig 5.9: Box plot of the distribution of score amongst the regions with outliers.

## 8.2 Interpret the key differences in distribution shapes, medians, and outliers.

The Middle East has more countries, so naturally, their box is bigger. Even though there are more countries, it has fewer outliers. Outliers as we know mess with the overall average and can skew the distribution.

South Asia only has 8 countries, and 3 of them are outliers. Because of this, the size of the box for South Asia is smaller, and the median is closer to the first quartile. The whisker for the third quartile is missing, which means all the data is below Q3. The 3 outliers (two above Q3 and one below Q1) signal that a couple of countries are performing exceptionally well compared to the rest, while one is significantly lower. The countries doing well have a score of ~4.5 to 5.25, which is further than the mean of ~4. Whereas the outlier at the bottom, has a score of ~1.75 which is much lower than the average.

For the Middle East, the median is slightly not centered, but the skewness isn't as big as in South Asia. The whisker for Q1 is a bit longer than Q3, meaning the lower end of the data is more spread out. There are two outliers (one just past the  $1.5 * IQR$  range from Q1 and one further out). These outliers fall in the ~2.8–3.5 happiness score range, much lower than the average of ~5.5.

## Conclusion:

In conclusion, we have noticed that metrics such as GDP, healthy life expectancy, social support, generosity, freedom to make life choices and perception of corruption influence the happiness scores. We used the world health report to analyse these metrics and discover outliers and compare these metrics between these regions.

The findings from these exercises reveal a significant difference between South Asia and the Middle East. The Middle East has greater happiness and scores better than South Asia in all the metrics that were compared. Furthermore, South Asia shows a greater variability of happiness with countries like Nepal and India performing significantly better than expected, while other countries like Afghanistan are performing exceptionally low. Furthermore, the presence of outliers especially in South Asia has skewed the regional averages making the data possibly less representative of the actual situation in the region.

Overall, this analysis offers a view of the happiness scores along with other metrics, helping to highlight the differences between these regions and exploring complicated relationships amongst the metrics.