5CS037

Concepts and Technologies of AI

Analysis of the World Happiness Report:

A Data-Driven Exploration of Global and Regional Trends

Name: Prashanna Karki

Group: L5CG16

Canvas ID: 2418331

Lecturer: Siman Giri

Tutor: Ronit Shrestha

Module Leader: Siman Giri

Submission Date : 12/20/2024

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

Objective of the report:

The primary goal of this written report is to examine the data analysis of the World Happiness Report dataset and perform a statistical explanation. It includes both global and regional patterns, relationships, and outliers. Additionally, it does a qualified study among the nations of South Asia. Additionally, it looks at proportion and disproportion among the regions.

Library Used on the Analysis:

The data analysis was done in a Python programming language which used the following libraries mentioned below:

- Numpy Library as (np)
- Pandas Library as (pd)
- Seaborn Library as (sns)
- Matplotlib Library as (plt)

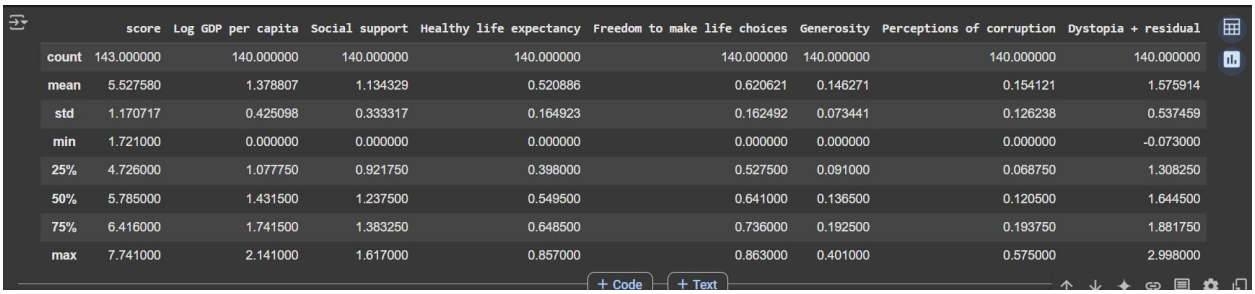
About the Dataset:

This dataset is the World Happiness Report of 143 different countries and there are a total of nine columns which are written below:

- Country Name
- Score (happiness score)
- Log GDP per capita (Economic Production of countries per person)
- Social Support

- Health Life Expectancy
- Freedom to make life choices
- Generosity
- Perceptions of corruption
- Dystopia + residual (irrational value of a country's happiness score)

With the help of the dataset shown below, the provided questions were solved which are explained in other pages:



	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

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

1. Data Exploration and Understanding

Dataset Overview

The CSV file was loaded in the IDE using the "read_csv()" method, and the required libraries were imported to provide an overview of the provided dataset. The "head ()" technique is used to display the first ten rows.

```
[2]  import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
```

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

1. Data Exploration and Understanding

Dataset Overview

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

df = pd.read_csv("/content/drive/MyDrive/dataset/WHR-2024-5CS037.csv")
df.head(10)
```

The dataset's row and column counts are displayed using the Shape attribute. Moreover, the ‘dtypes’ technique is used to list the columns and specific data types.

Identify the number of rows and columns in the dataset:

The (shape) attribute of the dataset was utilized to identify its dimensions. It revealed the total number of rows and columns present, providing an overview of the dataset's size.

```
# 2. Identify the number of rows and columns in the dataset.
rows = df.shape[0]
columns = df.shape[1]
print(f"Number of rows: {rows}\nNumber of columns: {columns}")
```

```
Number of rows: 143
Number of columns: 9
```

List all the columns and their data types:

The (dtypes) method was used to list all the columns along with their respective data types. This step is crucial for understanding the nature of the data in each column and planning further preprocessing steps accordingly.

```
# 3. List all the columns and their data types.
print(f"columns = {df.columns}\n")

print("All data types: ")
print(df.dtypes)
```

```
columns = Index(['Country name', 'score', 'Log GDP per capita', 'Social support',
       'Healthy life expectancy', 'Freedom to make life choices', 'Generosity',
       'Perceptions of corruption', 'Dystopia + residual'],
      dtype='object')

All 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
```

Basic Statistics:

The default methods mean (), median (), and std () {for standard deviation} of the pandas library were utilized to calculate the mean, median, and standard deviation of the score columns. Similarly, we utilized the max() and min() methods to determine which countries had the highest and lowest happiness scores.

Basic Statistics

```
[ ] # 1. Calculate the mean, median, and standard deviation for the Score column.
    print(f"The mean is {df['score'].mean()}")
    print(f"The median is {df['score'].median()}")
    print(f"The standard deviation is{df['score'].std()}")
```

```
The mean is 5.52758041958042
The median is 5.785
The standard deviation is1.1707165099442995
```

```
[ ] # 2. Identify the country with the highest and lowest happiness scores.
    print("The country with highest happiness score: ")
    df[['Country name','score']][df.score == df['score'].max()]
```

```
The country with highest happiness score:
```

	Country name	score
0	Finland	7.741

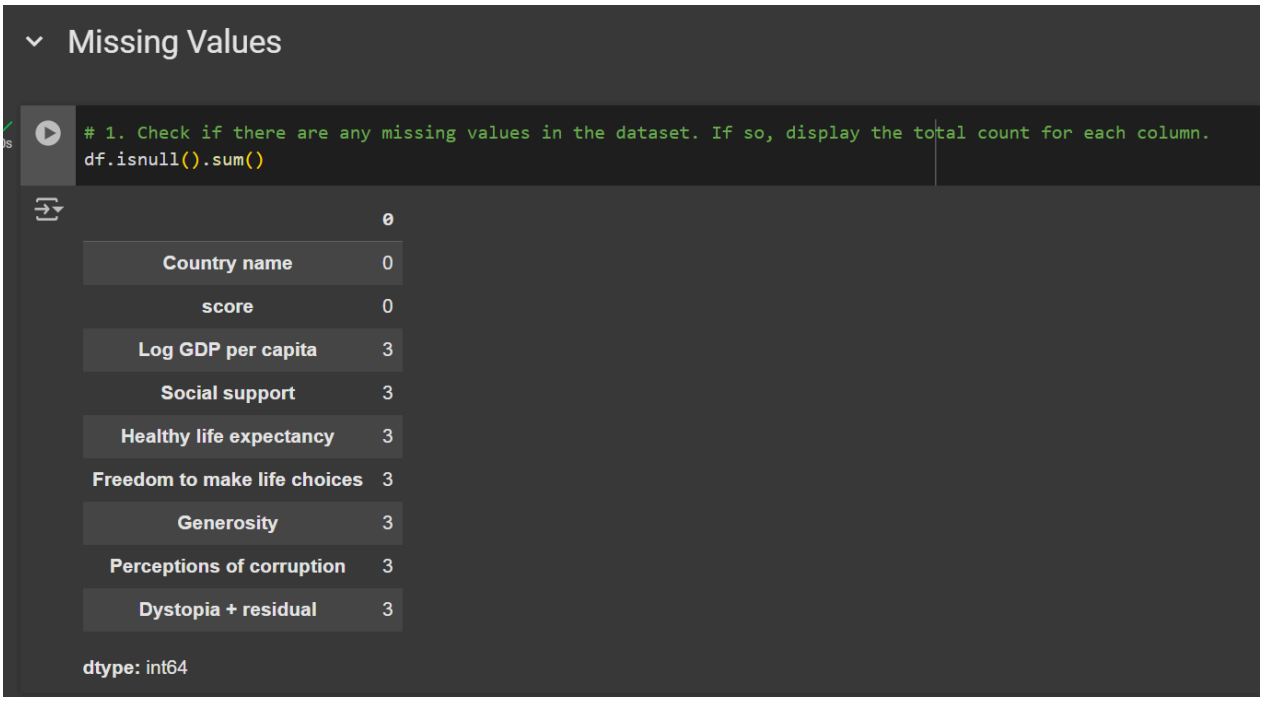
```
[ ]
    print("The country with lowest happiness score: ")
    df[['Country name','score']][df.score == df['score'].min()]
```

```
The country with lowest happiness score:
```

	Country name	score
142	Afghanistan	1.721

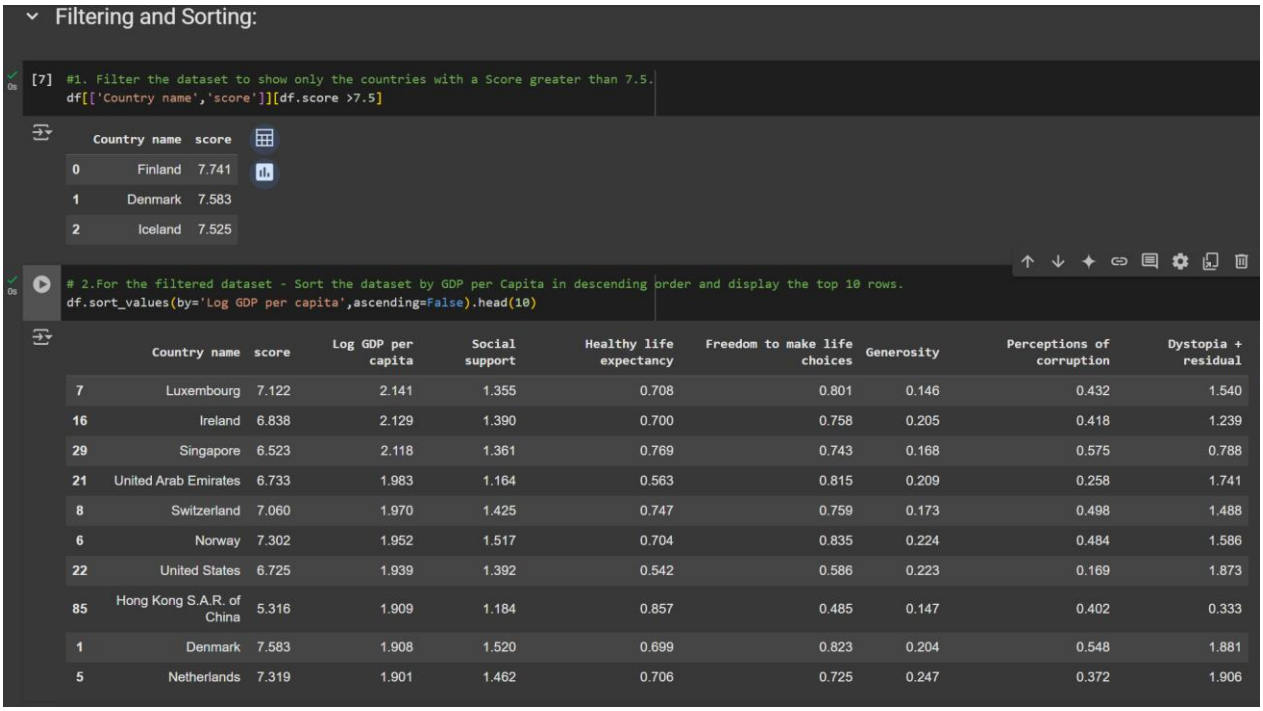
Missing Values:

There weren't many missing values in the dataset that we were given—that is, data that wasn't collected throughout the collection procedure. We used the sum() function to gather them column-by-column and the isnull() method to calculate them, which helped us find the missing values in our dataset.



Filtering and Sorting:

Given criteria ‘dataframe.score > 7.5’ was used to filter the provided dataset to reveal the countries with the score larger that that of 7.5. The dataset was sorted by Log GDP per capita using the sort_values() method, with ascending set to false. Ten rows were then shown using the head() method.



Filtering and Sorting:

```
[7] #1. Filter the dataset to show only the countries with a Score greater than 7.5.
df[['Country name','score']][df.score >7.5]
```

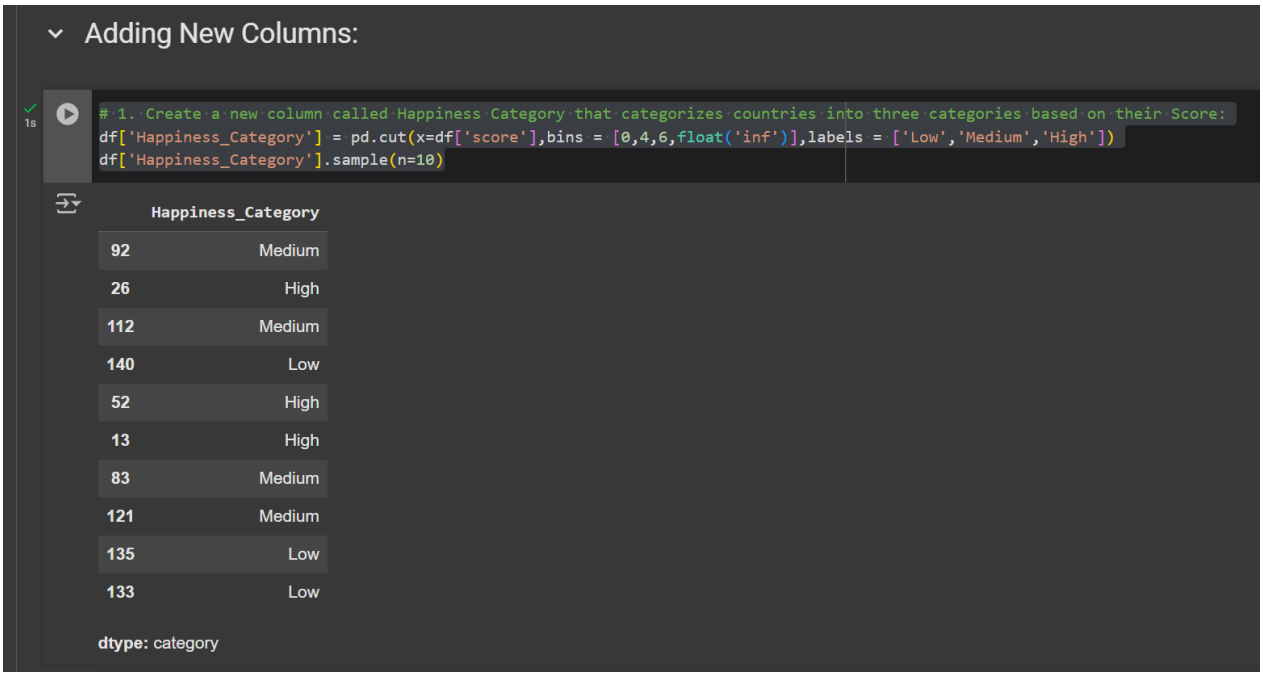
	Country name	score
0	Finland	7.741
1	Denmark	7.583
2	Iceland	7.525

```
# 2.For the filtered dataset - Sort the dataset by GDP per Capita in descending order and display the top 10 rows.
df.sort_values(by='Log GDP per capita',ascending=False).head(10)
```

	Country name	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
7	Luxembourg	7.122	2.141	1.355	0.708	0.801	0.146	0.432	1.540
16	Ireland	6.838	2.129	1.390	0.700	0.758	0.205	0.418	1.239
29	Singapore	6.523	2.118	1.361	0.769	0.743	0.168	0.575	0.788
21	United Arab Emirates	6.733	1.983	1.164	0.563	0.815	0.209	0.258	1.741
8	Switzerland	7.060	1.970	1.425	0.747	0.759	0.173	0.498	1.488
6	Norway	7.302	1.952	1.517	0.704	0.835	0.224	0.484	1.586
22	United States	6.725	1.939	1.392	0.542	0.586	0.223	0.169	1.873
85	Hong Kong S.A.R. of China	5.316	1.909	1.184	0.857	0.485	0.147	0.402	0.333
1	Denmark	7.583	1.908	1.520	0.699	0.823	0.204	0.548	1.881
5	Netherlands	7.319	1.901	1.462	0.706	0.725	0.247	0.372	1.906

Adding new columns:

To add new columns the method pd.cut () was used to classify countries into 3 categories which are low, medium and high based on their score. Following that new column Happiness_Category was made.



2. Data Visualization:

As seen in the provided graphic, a bar chart representing the ten happiest nations was used in this section. The bar graph clearly shows the happiness score for each nation, showing that the top-ranked countries differ very little from one another. Even at a glance, the data may be easily interpreted because to the clear labels and titles.
To illustrate the data in the dataset, all types of diagrams were plotted using the Seaborn and Matplotlib packages.

Bar Plot:

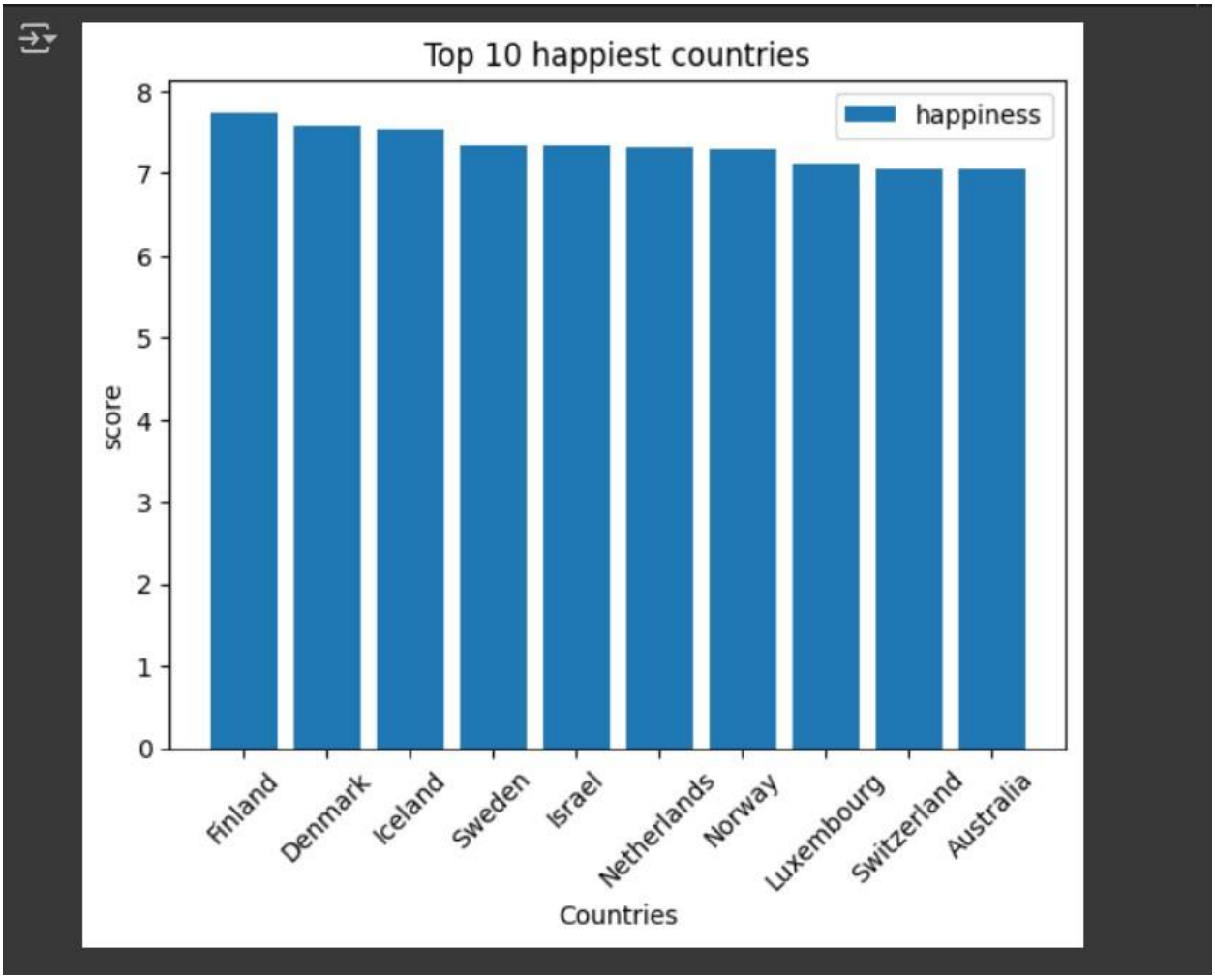


Fig1: bar chart to show top 10 happiest countries by score from the data set

Line Plot:

To illustrate the ten most unhappy nations according to their score, a line plot was made in this section. The graphic illustrates the plot, which shows the countries' dissatisfaction scores in ascending order. By showing how the scores progressively rise, the graph sheds light on the nations with the lowest levels of happiness.

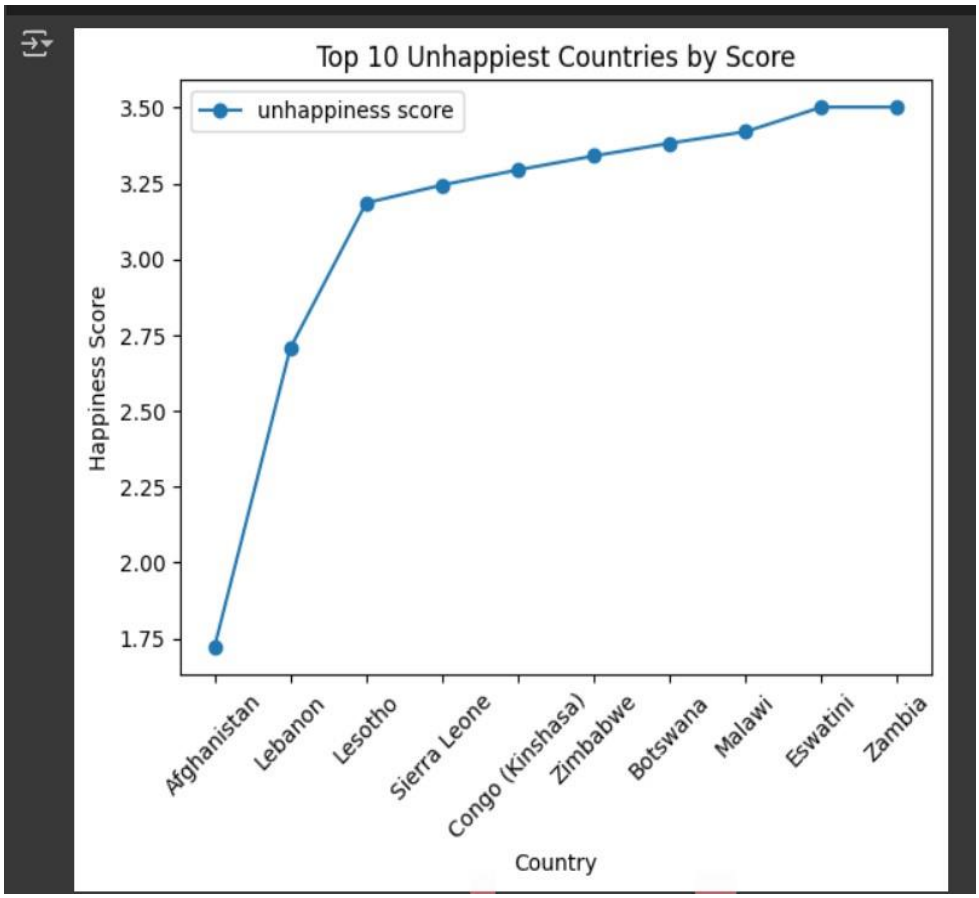


Fig 2: line plot to show top 10 unhappiest countries by score.

Histogram:

By observing the histogram below we can tell that Most people have their happiness score in the mid to high range (5-7) whereas fewer people have their happiness score in the extreme low or extreme high.

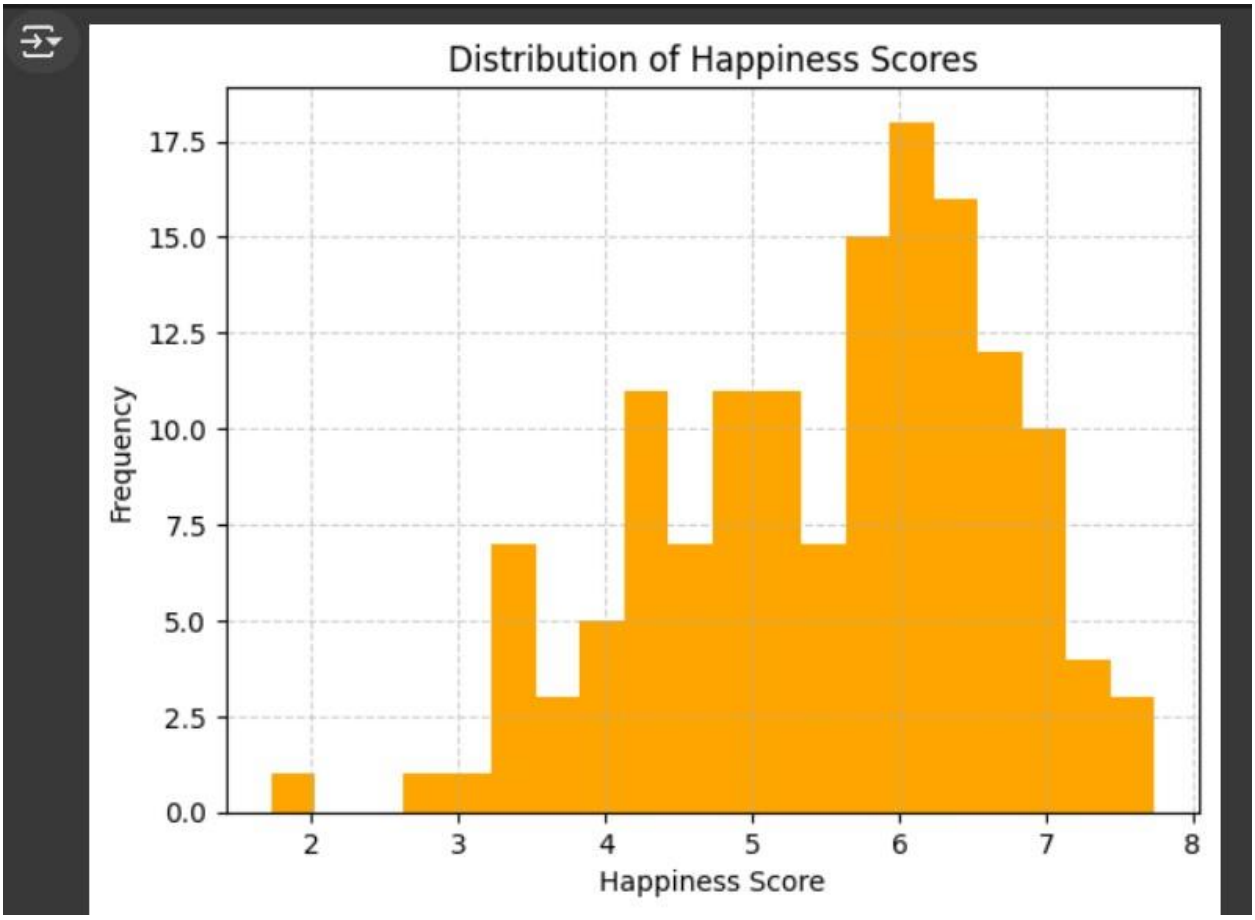


Fig 3: Histogram showing the happiness score.

Scatter Plot:

By analyzing the scatter plot, it is evident that as the **Log GDP per capita** increases, the happiness **score** also shows a significant upward trend. This reveals a **positive correlation** between these two variables, indicating that countries with higher GDP per capita tend to have higher happiness scores.

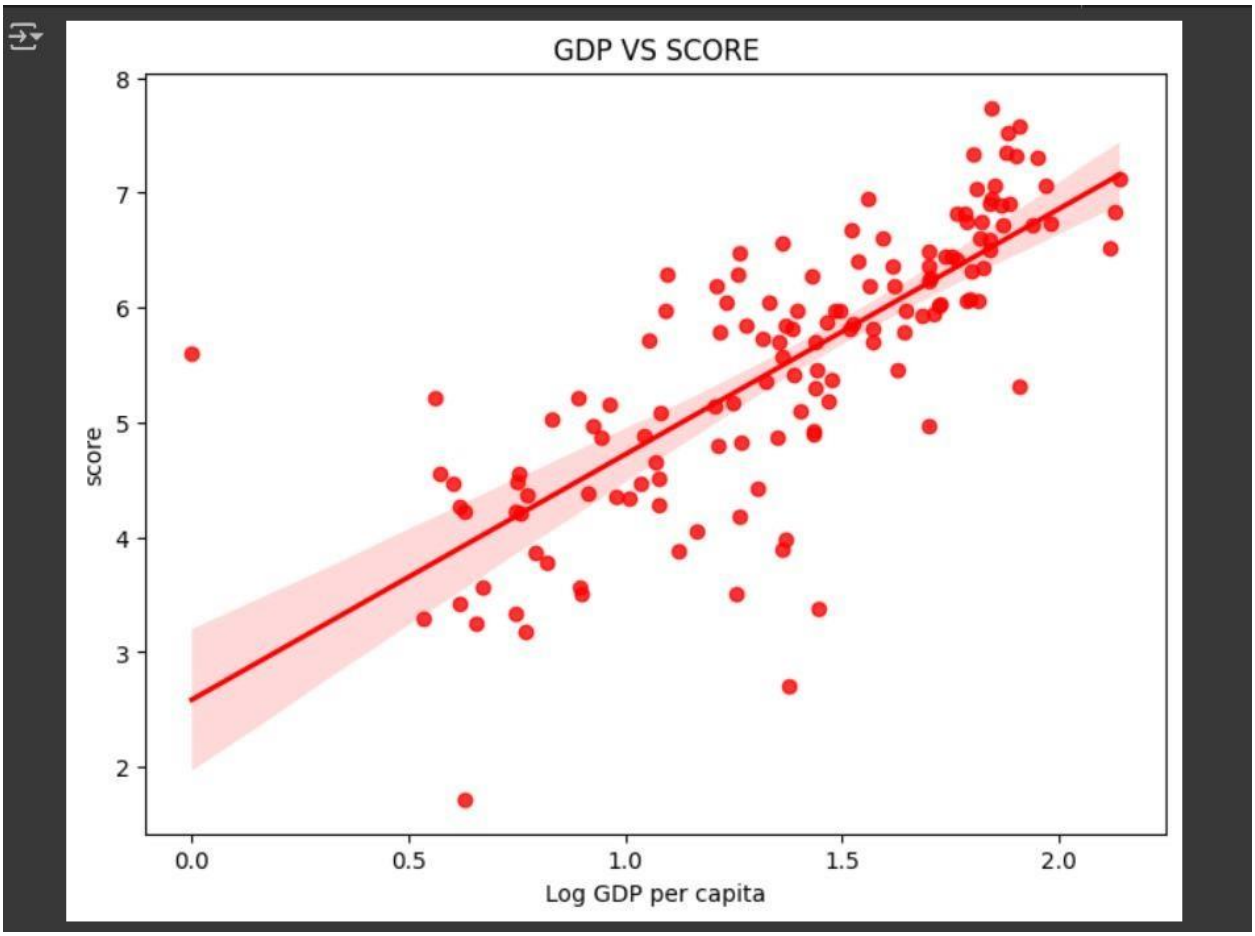


Fig 4: Scatter Plot to show the relation between the Log GDP per capita and Score.

Problem 2 – Some Advance Data Exploration Task:

Task 1: Setup Task – Preparing the South Asia Dataset.

The list called south_asian_countires, which includes nations like Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka, was created in order to prepare the South Asia dataset. The isin() method is used to filter the dataset's similarities from the list. This will allow us to extract full rows by Country Name if the nations are present in the dataset above. The resulting dataset will then be stored as a new CSV file using the to_csv method.
().

```
# 1. Defining countries in the South Asia with a list:

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

# 2. Using list from step - 1 to filter the dataset.
south_asian_country = df[df['Country name'].isin(south_asian_countries)]

# 3. Saving the filtered dataframe as  a separate CSV files to use in future.
south_asian_country.to_csv("drive/MyDrive/ML WORKSHOPS/DataSet/south_asian.csv",index = False)
south_asian_country
```

```
asian_df = pd.read_csv("drive/MyDrive/ML WORKSHOPS/DataSet/south_asian.csv")
asian_df
```

```
[ ]  asian_df.isnull().sum()
```

Task 2: Composite Score Ranking:

The new dataset is imported as asian_df. Then a new column is created as Composite Score in the dataset for calculating the composite score.

To calculate it this formula was used:

COMPOSITE SCORE: 0.40 * GDP per capita + 0.30 * Social Support + 0.30 * Healthy Life Expectancy

Following that data frame ranked itself in descending order on the basis of Composite Score.

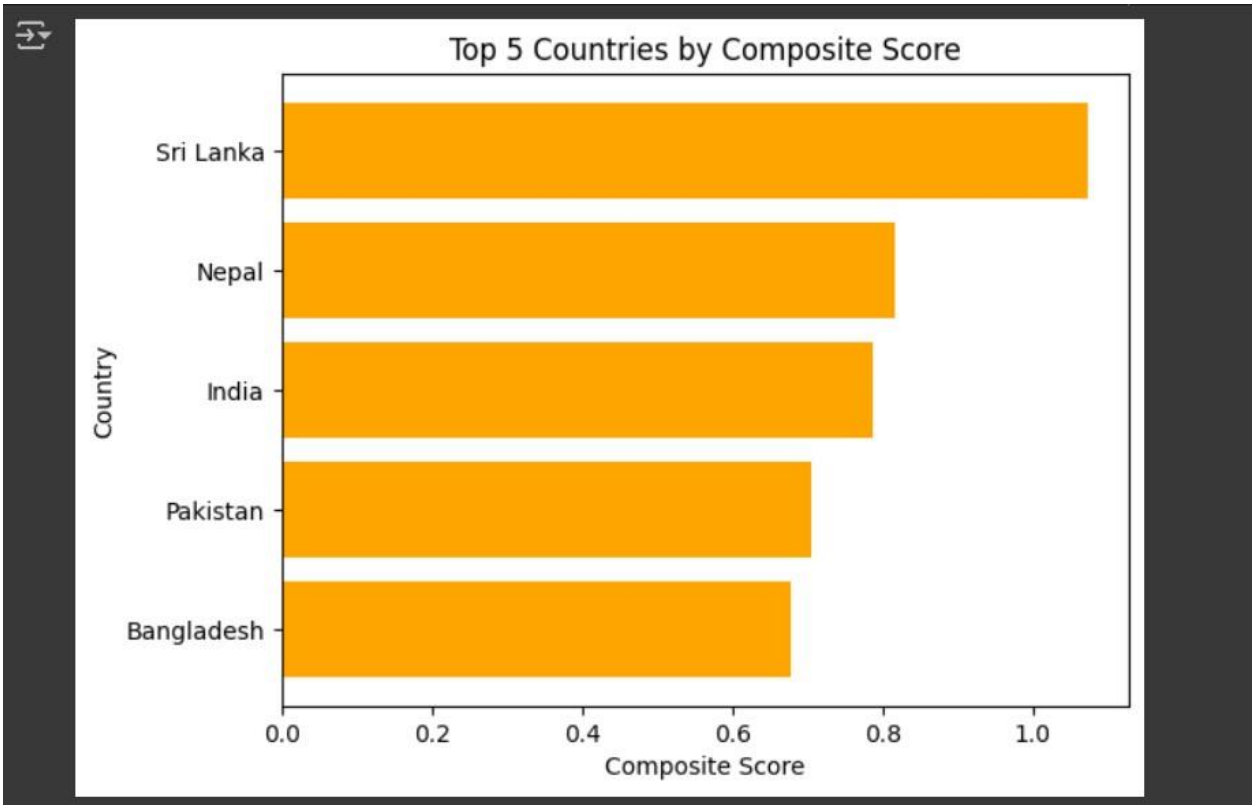


Fig 5: Bar Chart of top 5 countries by Composite Score.

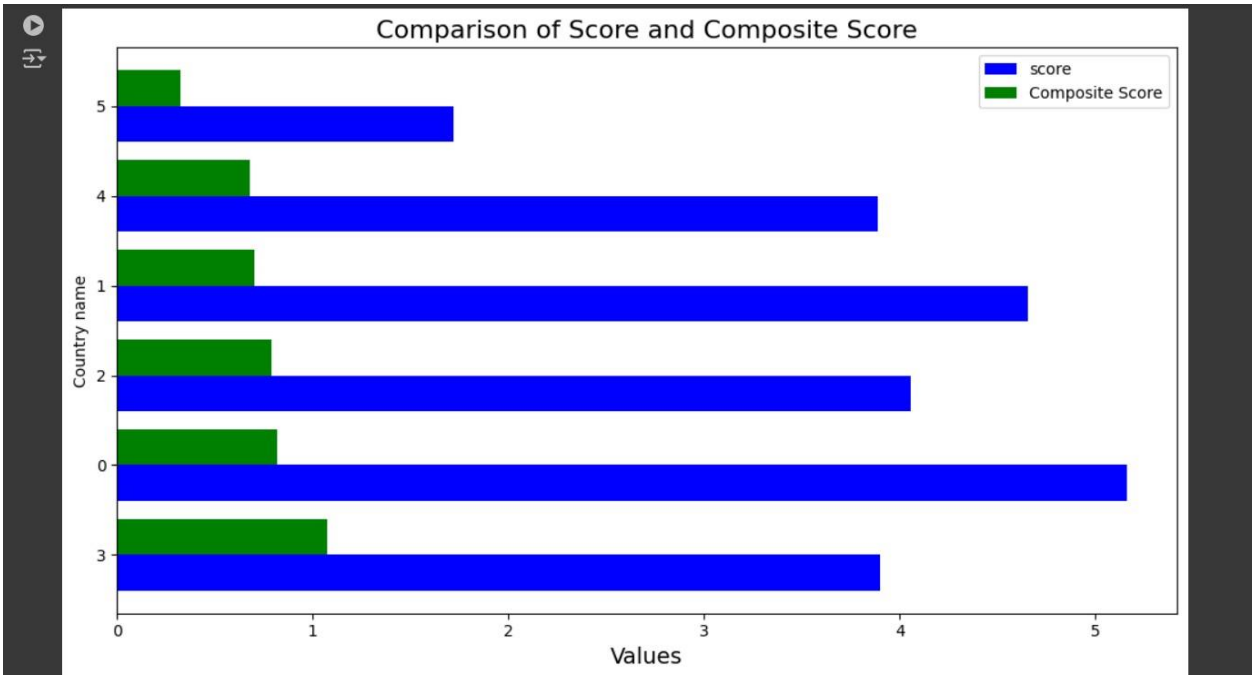


Fig 6: Comparison of Composite Score and Score using Bar Chart.

The ranking of Composite Score doesn't line up with Score as seen from the above Figure.

For Example:

Afghanistan and Bangladesh both have poor scores, and their combined scores remain low.
For Nepal's case it has the highest Score although Sri Lanka has the highest Composite Score so does India and Pakistan.

Task 3: Outlier Detection:

Those countries who are absolutely distinct from other countries because to their incredibly high or low GDP per capita or due to really high or low Score these countries are known as Outliers.
The following formula can be used to determine whether countries are outliers: 1.5 * IQR. We may observe that only one nation emerges as an outlier after using this formula.

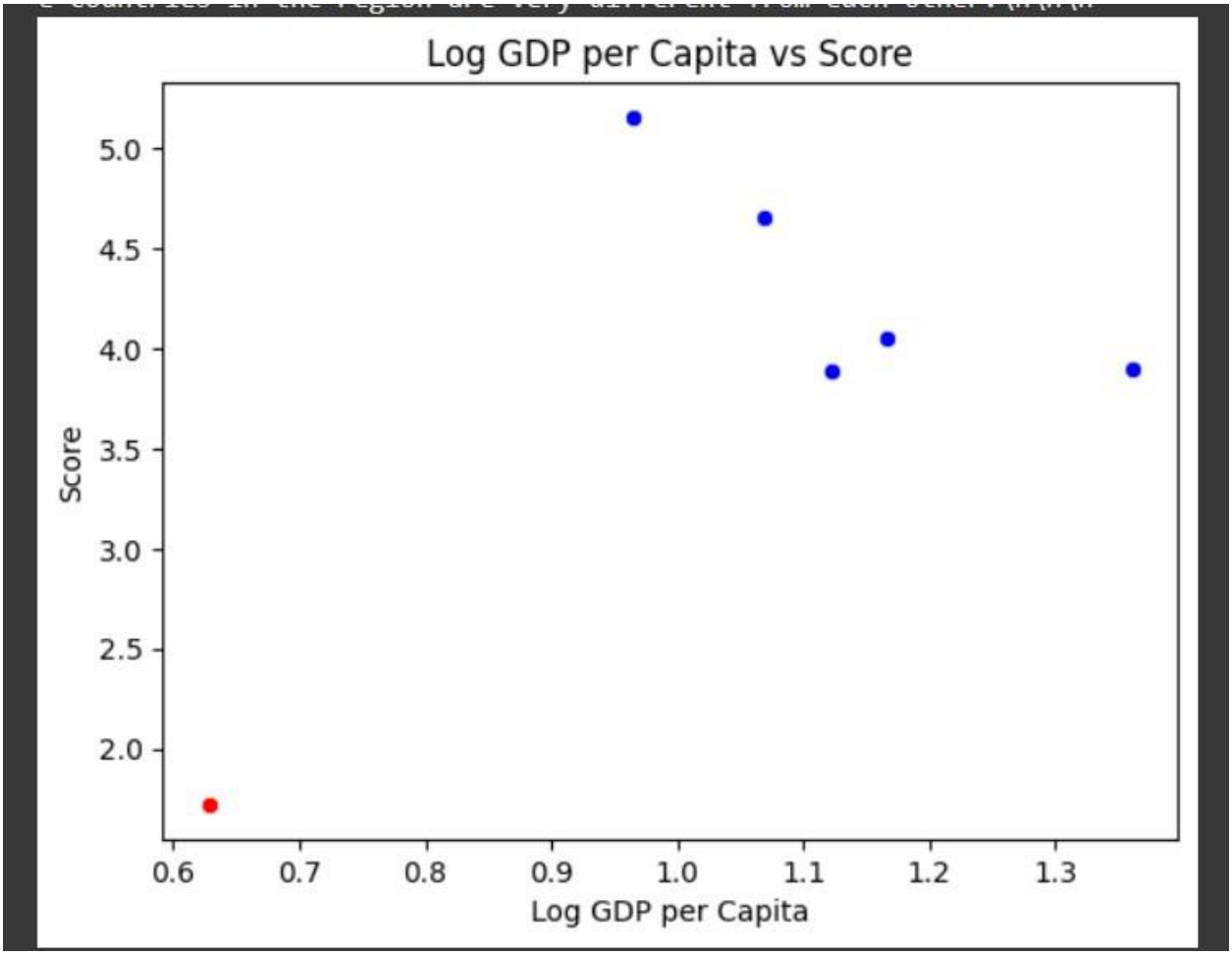


Fig 7: Scatter Plot of Outliers and non-Outliers.

As seen the outlier country is Argentina which is lower than that of other countries.

Some Potential Impacts of Outlier are as follows:

- Weather a country is richer than other countries it will make the whole region look wealthy that it actually is.
- So goes for the Poor conditions it can make the whole region look Poor that it really is.

- Outliers can make the data look huge and messy.

Task 4: Exporting Trends Across Metrics:

Choosing 2 metrics Generosity and Freedom to make life choices from our data frame. Calculating the correlation between them using panda's library's method corr (). By using this method it helps us to calculate the correlation using person's correlation.

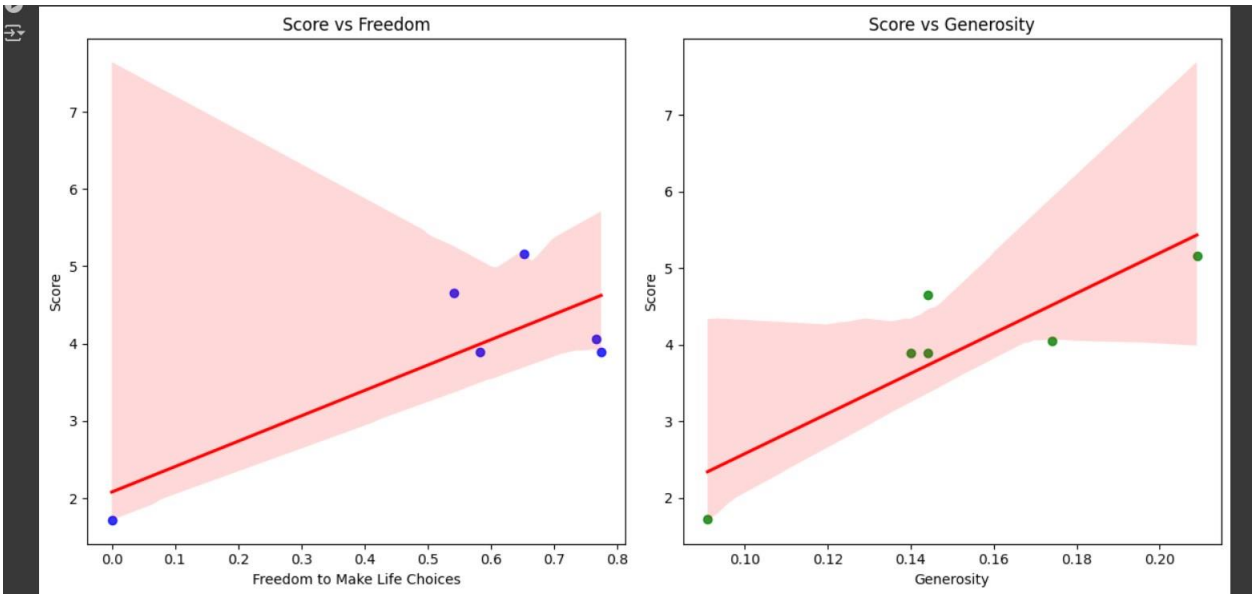


Fig 8: Scatter plots of trendlines for metrics against the Score.

From the above figure we can see both Strong and Weak relation:

Strong Relation: Freedom to make life choices contains powerful link with Score. Which means country where people feel freedom have higher happiness score.

Weak Relation: Generosity contains delicate link with Score. Which means generosity won't impact the happiness score that much.

Task 5: Gap Analysis:

Calculating the gaps between scores and Log GDP per capita we need to subtract the Log GDP per capita by score. Following the calculations storing it on a new column named 'GDP-Score Gap' of south_asain_df.

Dispalying the South Asian Countries rank in both ascending and descending order in a Bar Chart.

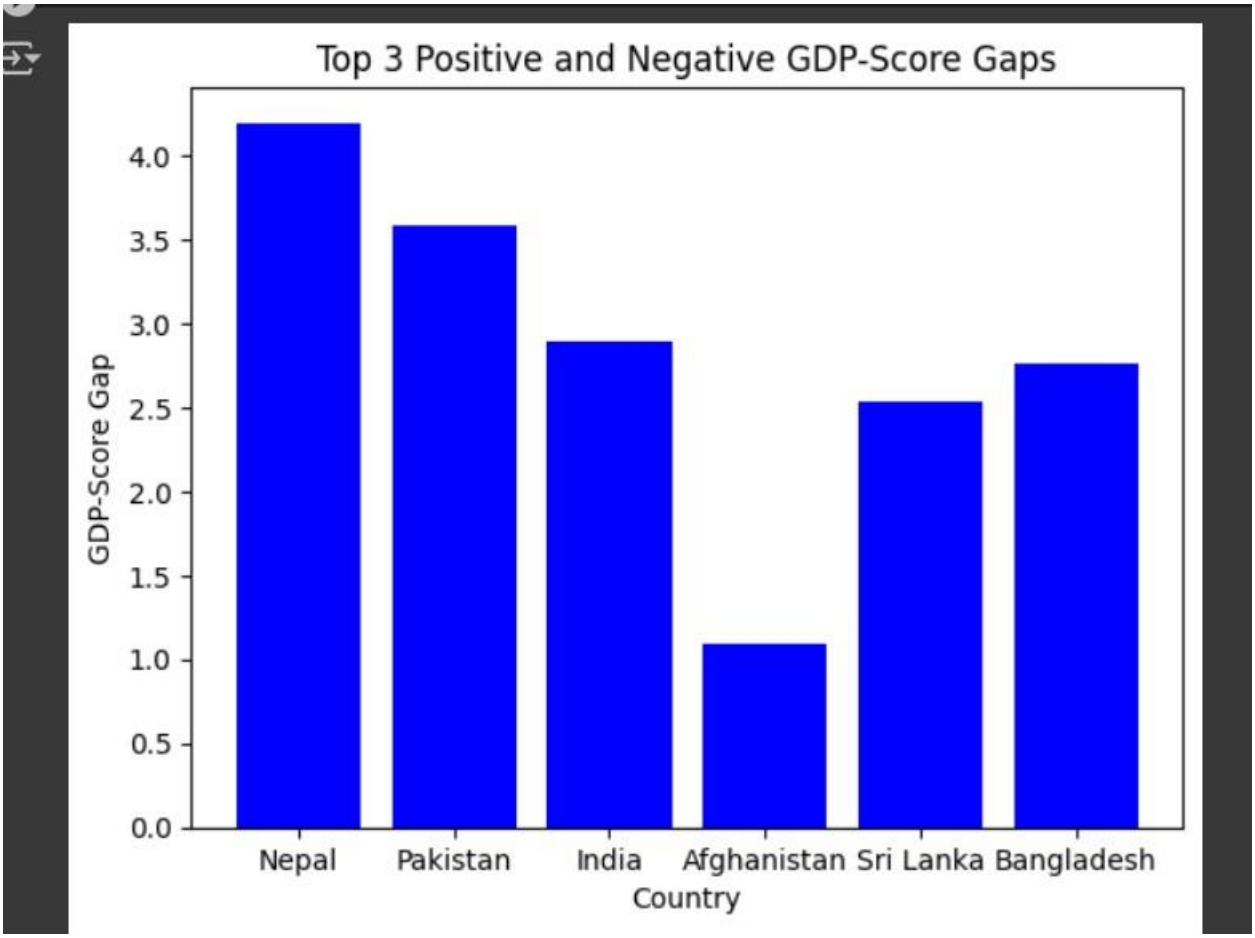


Fig 9: Bar chart of positive and negative gaps

Forget negative scores, happiness and wealth seem to go hand-in-hand in some places. Countries like Afghanistan, Sri Lanka, and Bangladesh show this - their happiness levels are closer to their wealth (measured by GDP).

On the other hand, Nepal, Pakistan, and India tell a different story. While people there seem pretty happy, their wealth is much lower. This could be because of poverty, economic struggles, or other social issues they face.

Problem 3: Comparative Analysis:

Task 1: Setup Task – Preparing the Middle Eastern Dataset:

Making a list of Middle Eastern countries. Which, I used to find and select the corresponding rows in my main dataset. This created a new dataset specifically for Middle Eastern countries, which I saved as a CSV file.

When I looked at this new dataset, I noticed that the information for Bahrain and the State of Palestine was completely missing. To avoid introducing inaccurate or 'fake' data, I decided to remove these countries from the dataset entirely. This is because filling in the missing information with guesses could lead to incorrect results in my analysis."

So overall here new data frame named as middle_esat_df was created and was exported to a CSV file using method to_csv () and for any null values isnull () method was used. After finding the null rows which were 2 countries Palestine and Bahrain I dropped them using loc () method.

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

     middle_east_countries = [ "Bahrain", "Iran", "Iraq", "Israel", "Jordan", "Kuwait", "Lebanon", "Oman", "State of Palestine", "Qatar",
                               "Saudi Arabia", "Syria","United Arab Emirates", "Yemen"]
     middle_east_df = df[df['Country name'].isin(middle_east_countries)]
     middle_east_df.to_csv("/content/drive/MyDrive/dataset/middle_east.csv",index=False)
```

```
[ ]  east_df = pd.read_csv("/content/drive/MyDrive/dataset/middle_east.csv")
     east_df
```

```
[ ]  east_df.isnull().sum()
```

```
     east_df = east_df.loc[(east_df['Country name'] != 'Bahrain') & (east_df['Country name'] != 'State of Palestine')]
     east_df
```

```
     east_df
```

1. Descriptive Statistics:

I determined the mean score and the standard deviation of scores for both Middle Eastern and South Asian nations. People in the Middle East reportedly reported higher levels of happiness on average, as evidenced by the fact that their average score was higher than that of South Asian nations.

```
[ ]  # Calculate the mean, Standard deviation of the score for both South Asia and Middle East.

     print("The mean of asian dataframe", asian_df['score'].mean())
     print("The Standard deviation of asian dataframe",asian_df['score'].std())
     print("The mean of middle east dataframe",east_df['score'].mean())
     print("The Standard deviation of east dataframe",east_df['score'].std())

     # Which region has higher happiness Scores on average?
     """
     Since, the mean (score) of middle east is greater than south asian so, middle east region has higher happiness score.
     """

     The mean of asian dataframe 3.895666666666667
     The Standard deviation of asian dataframe 1.1770690152521504
     The mean of middle east dataframe 5.351333333333333
     The Standard deviation of east dataframe 1.648656346847335
     '\nSince, the mean (score) of middle east is greater than south asian so, middle east region has higher happiness score.\n'
```

2. Top and Bottom Performers:

I identified the top 3 and bottom 3 happiest countries in both regions. Then, I combined the results from both regions into a single view for easy comparison.

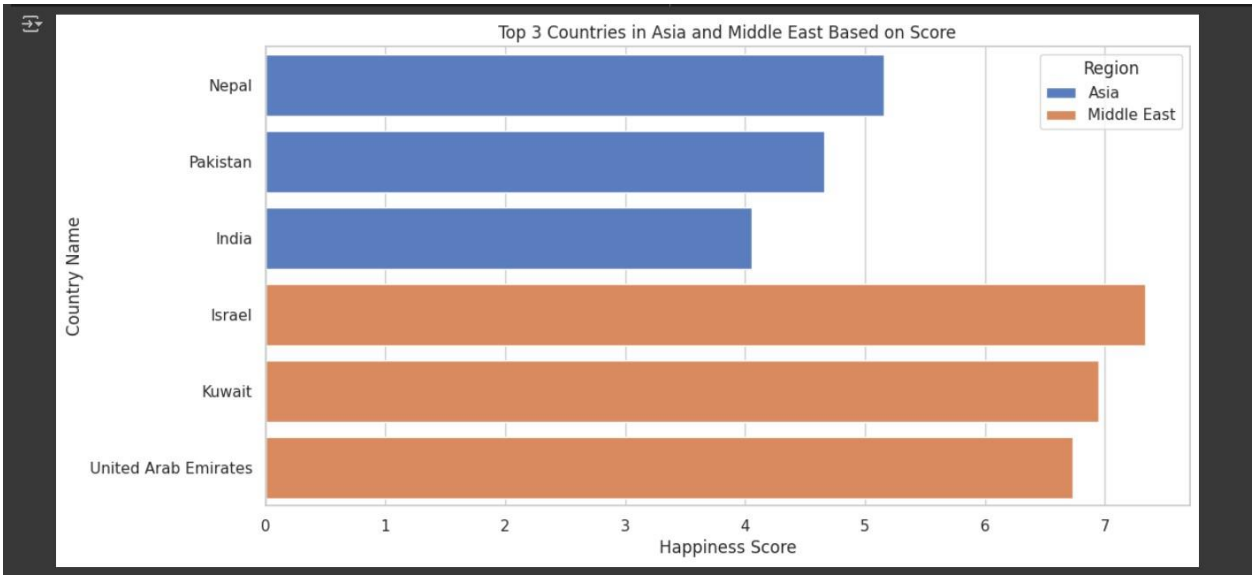


Fig 10: Bar Chart of top 3 happiest countries from both data frames.

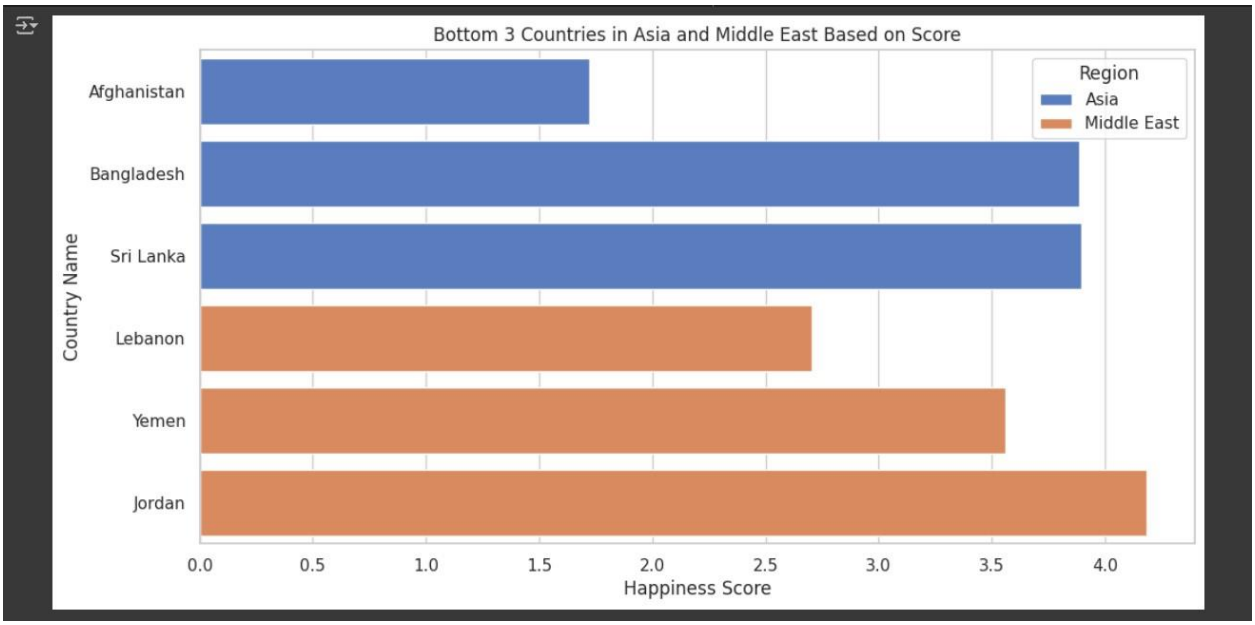


Fig 11: Bar Chart of bottom 3 happiest countries from both data frames.

3. Metric Comparisons:

The first step involves creating columns on both data frames called "Region," then filling in the new columns for "South Asia" and "Middle East" in the regions of "South Asia" and "Middle East," respectively.
Concat was used to merge the two data frames, and the resulting data frame was named combined_df. Extracting metrics Country name, Region, Log GDP per capita, Social support, Healthy life expectancy form this new integrated data frame.
reshaping the data according to region using the melt() function. displaying the information as a bar chart with groups.

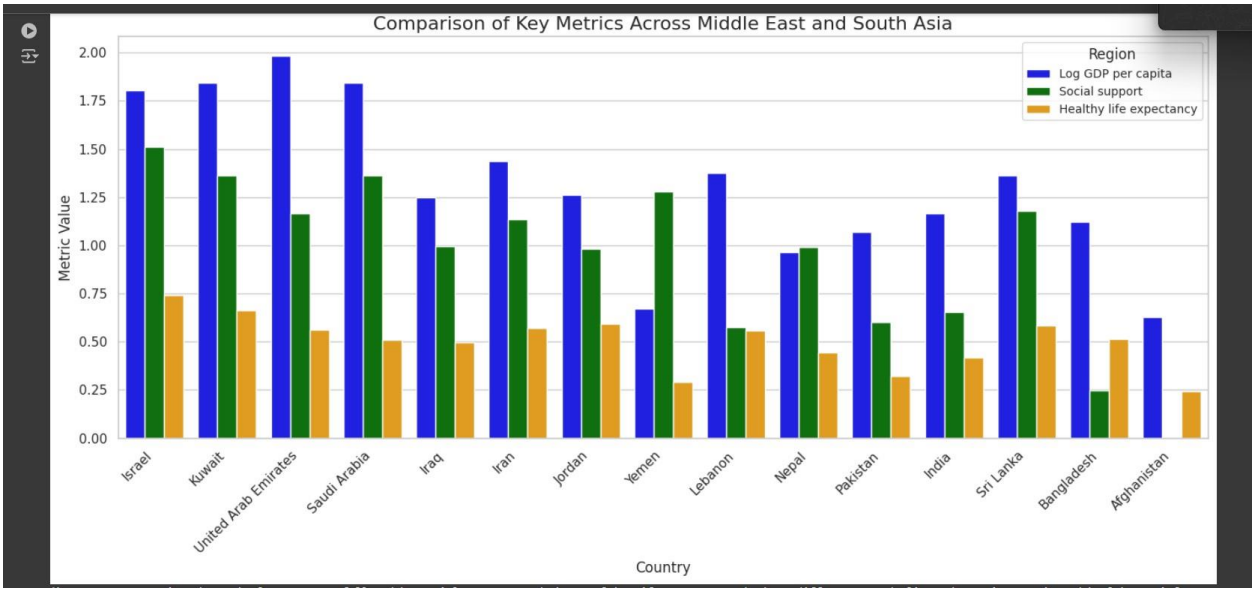


Fig 12: Metric Comparison between Data frames.

From the above figure we can see that Log GDP per capita showing the largest gap followed by Social Support and Healthy Life Expectancy.

4. Happiness Disparity:

I measured the range (difference between the highest and lowest scores) and the coefficient of variation (a measure of relative variability) in happiness scores for both regions.

Results:

- **South Asian Region:** Range: 3.4370000000000003, Coefficient of Variation:
  - 30.21482883337427.
- **Middle Eastern Region:** Range: 4.634, Coefficient of Variation: 30.808328395054225.

The Middle Eastern region shows greater variability in happiness scores, as indicated by a larger range and a slightly higher coefficient of variation. This suggests that happiness levels in the Middle East are more widespread compared to South Asia.

5. Correlation Analysis:

Using the groupby() method on the basis of the Region column, determine the link between Score and other metrics such as Generosity and Freedom to make life choices within each region. The outcome is shown below:

Region		score	Freedom to make life choices	Generosity
Middle East	score	1.000000	0.863220	0.627524
	Freedom to make life choices	0.863220	1.000000	0.388854
	Generosity	0.627524	0.388854	1.000000
South Asia	score	1.000000	0.800519	0.874512
	Freedom to make life choices	0.800519	1.000000	0.733396
	Generosity	0.874512	0.733396	1.000000

Fig 13: Corelation of Score with other Metrics.

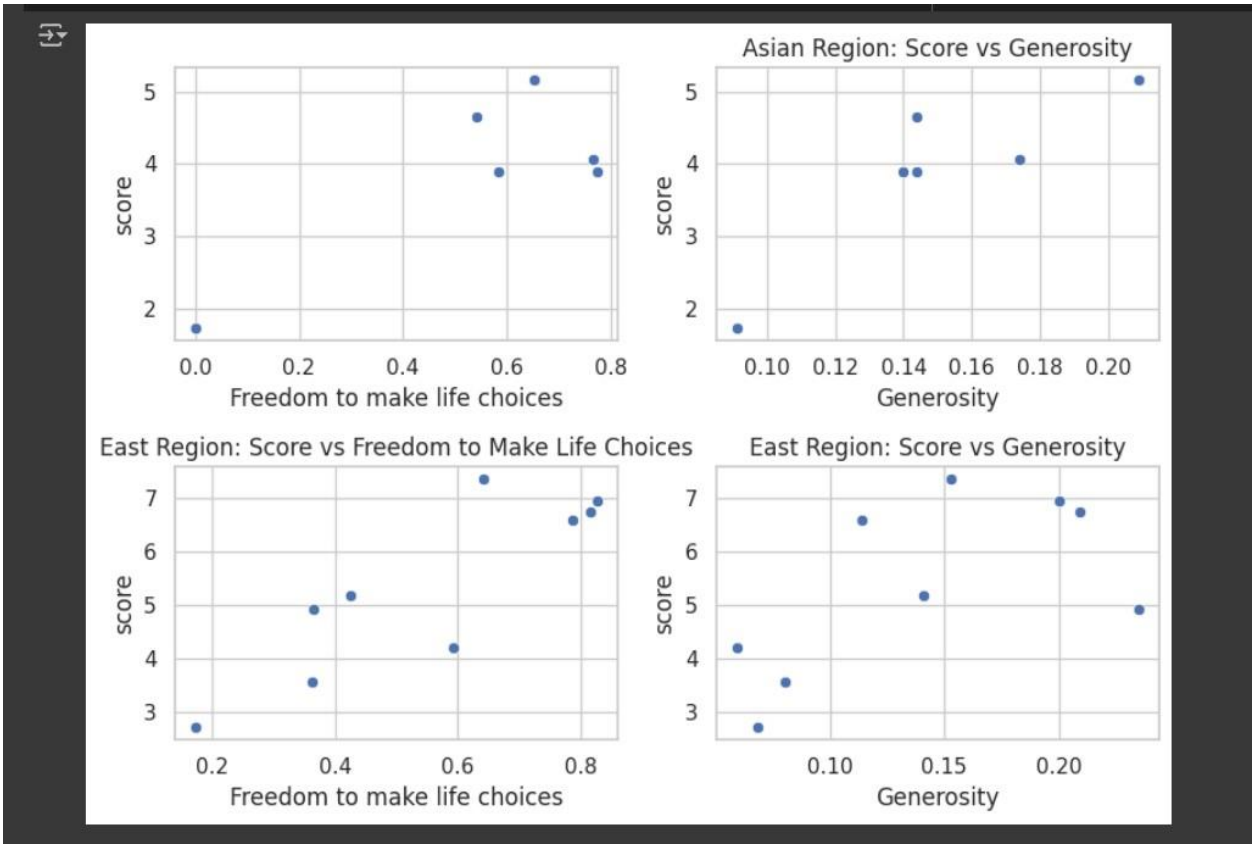


Fig 14: scatter plot comparing score with other metrics.

As can be seen from the preceding numbers, there is a direct positive association between happiness scores in both regions and the freedom to make life decisions.

Generosity, on the other hand, does not significantly correlate with happiness. For this reason, eastern nations score higher than those in the south.

6. Outlier Detection:

As mentioned in above the same formula is used for the detection of outlier in both regions 1.5 * IQR. From out detection we found that middle_east_df didn't have any outliers whereas asian_df had one.

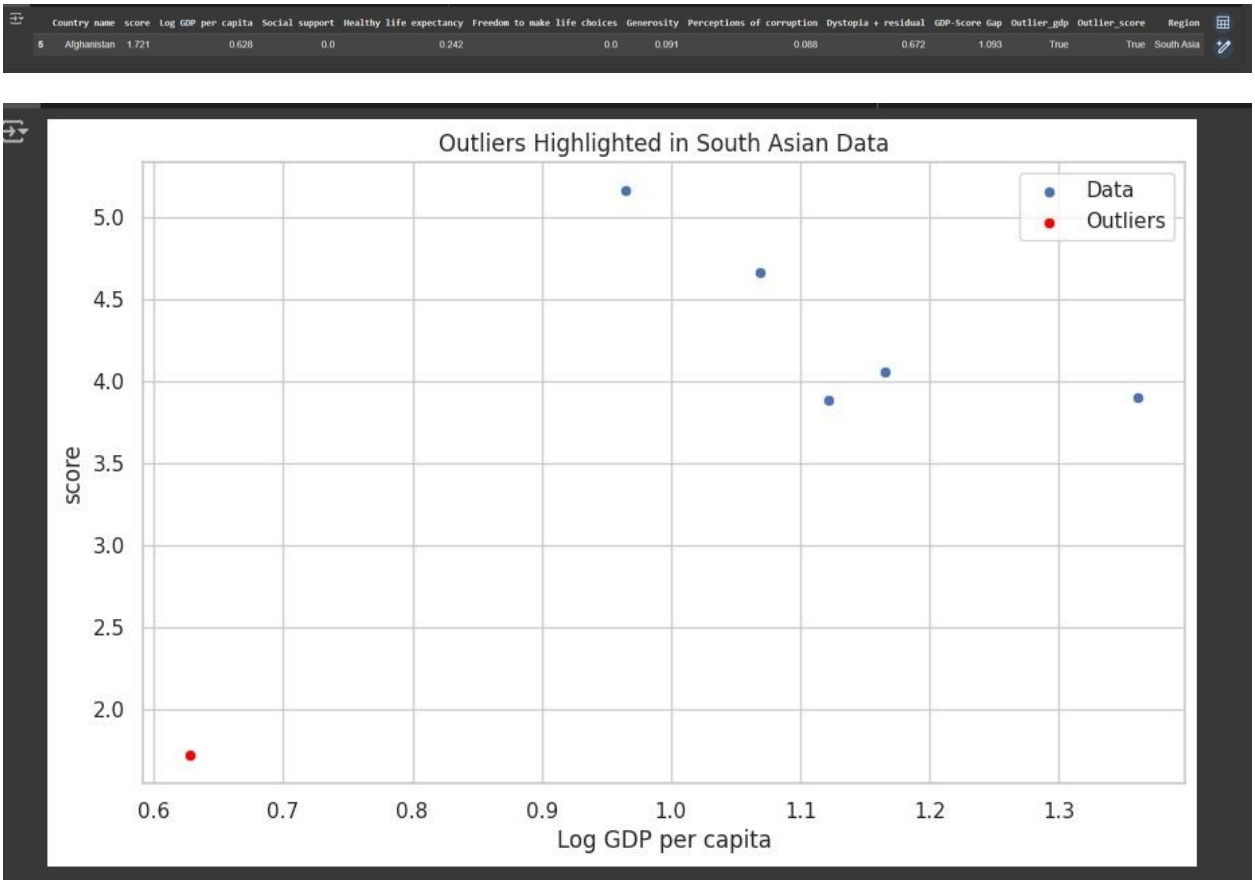


Fig 17: Outliers in Shout_asain_df.

The only country that stands out significantly in the South Asian data is Afghanistan. It shows unusual values for both its economic output (GDP) and happiness score. This outlier suggests that the people of Afghanistan are experiencing significantly lower levels of happiness compared to the average in South Asia.

FOR MIDDLE EASTERN:

Using the same formula for outlier detection, 1.5×IQR1.5 \times \text{IQR}1.5×IQR, it was determined that the **Middle Eastern dataset (middle_east_df) did not have any outliers**, as shown in the figure above. The distribution of data points in this region is consistent, with no values significantly deviating from the expected range.

In contrast, for the South Asian dataset (asian_df), one prominent outlier was identified—Afghanistan. This country exhibits significantly lower values for both economic output (measured by GDP) and happiness score. The unusual positioning of Afghanistan in the scatter plot highlights its distinctive characteristics compared to other South Asian nations, emphasizing the relatively low happiness levels of its citizens.

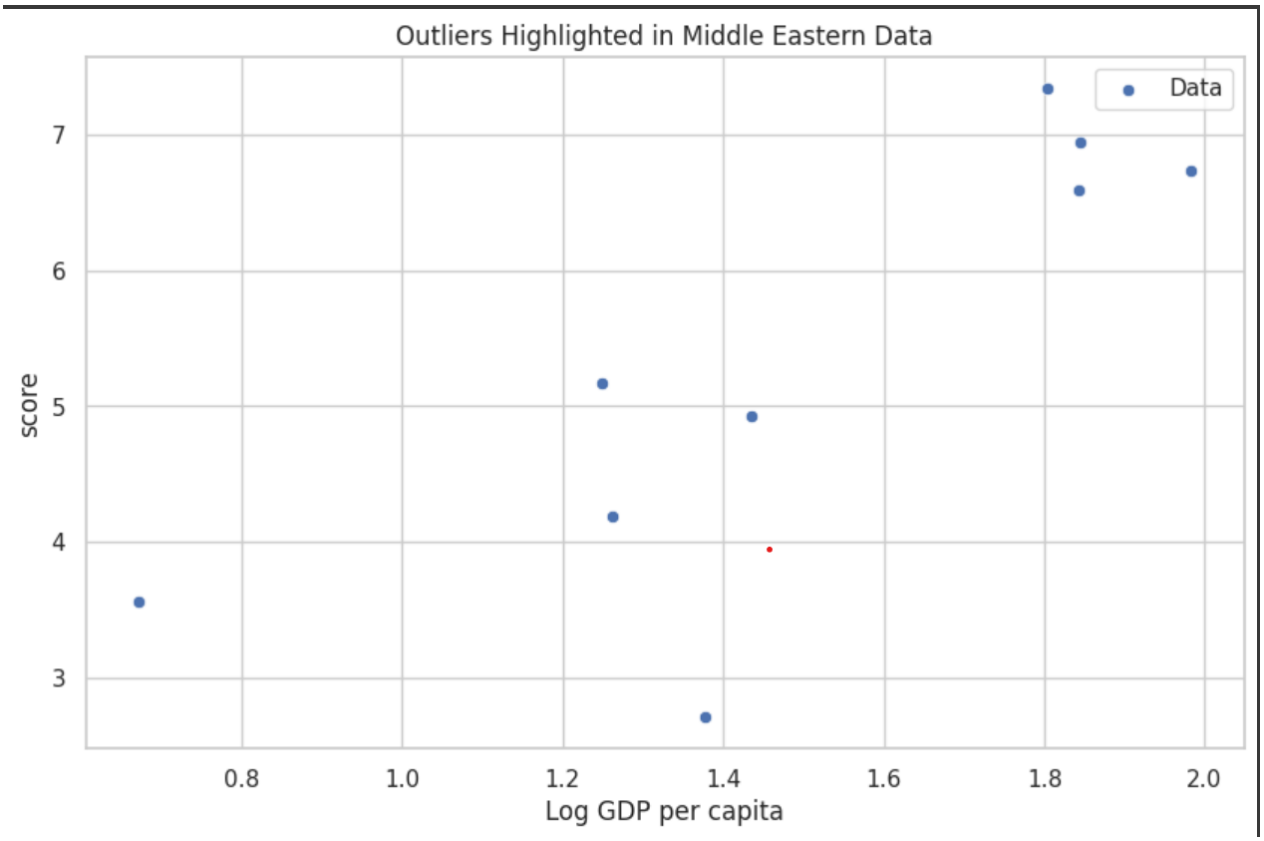


***Fig 18**: Scatter plot for the **Middle Eastern Data** with no outliers detected.*

7. Visualization:

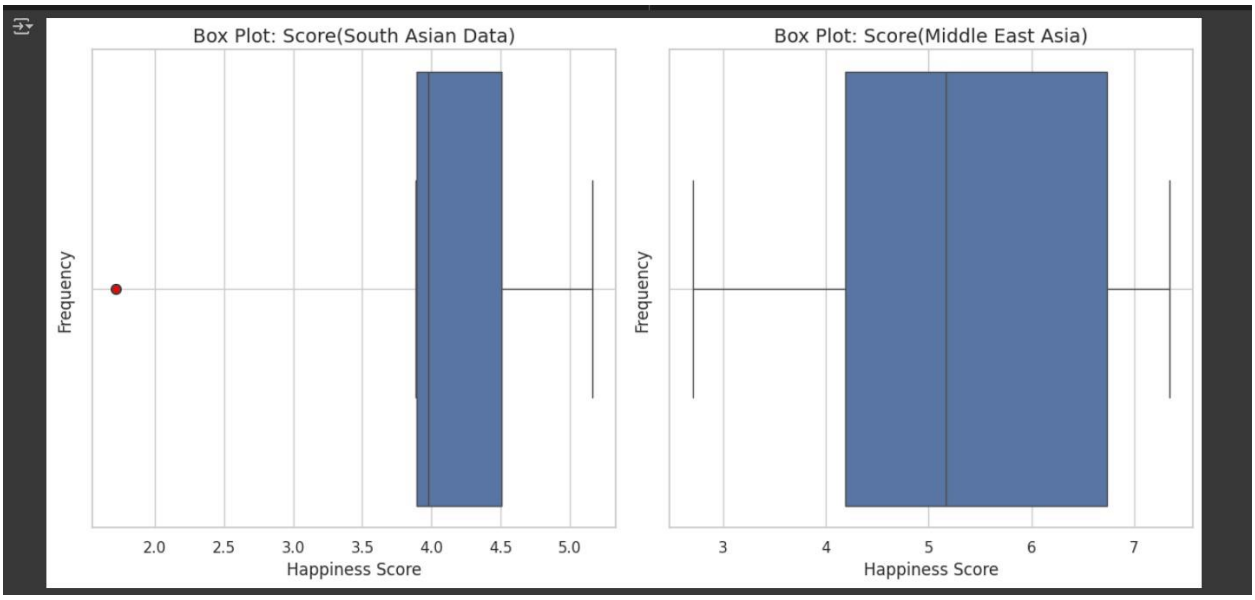


Fig 19: Boxplot comparing distribution of Score.

From the displayed figure it shows that:

In South Asia, mostly Scores of happiness is closer to 4.0 to 4.5. It shows that almost all the countries have the same Score but on the lower side. There is one Score less that 2 which is much lower than other countries so that country stands out as an outlier.

Whereas in Middle East , the Scores are higher and more spread out from 4 to 7.

Overall South Asia has a lower Score of happiness than compared to Middle East.

Conclusion:

In conclusion, the analysis of the World Happiness Report was successful. The initial data exploration phase provided a solid understanding of the dataset through statistical and visual analysis. The focus on happiness scores from South Asia and the Middle East revealed distinct regional trends. South Asian countries generally had lower happiness scores, around 4, with Afghanistan standing out as an outlier due to its significantly lower score. In contrast, Middle Eastern countries exhibited higher happiness scores, sometimes reaching 7, with no detected outliers. This comparative analysis highlights clear regional differences in happiness levels and provided meaningful insights.