Academic Year	Module	Assessment Number	Assessment Type
2024	Concepts and Technologies of AI	1	Report

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

Student ID: 2414183

Student Name: Swapnil Bhattarai

Section: L5CG07

Module Leader: Mr. Siman Giri

Tutor: Mr. Siman Giri

Submitted on: 20th December 2024

Table of Contents

Introduction:

The report of a study performed annually to determine the state of happiness within the people of different countries and analyze the study comparing it to the result of different other countries data is World Health Report. The World Happiness Report is a very important way of computing the worldwide happiness through different aspects. The World Health Report is based on the different key factors of life which is used to measure the happiness report such as: Economic production, Healthy life expectancy, Generosity, Freedom to make life choices and more. The data from the Gallup World Poll is accessed to rank the countries with the specific aspects to identify as well as classify the happiness scoring.

The importance of

Objectives:

The objective of this report is to understand how the data can be analyzed and manipulated as per required. In problem 1, it is required to understand the data to explore that data by loading it to the data frame as well as filter and manipulate the data. In problem 2, it is required to filter the data frame based on the list provided and perform various operations such as detecting outliers, calculating Pearson correlation coefficient and performing gap analysis. In problem 3, it is required to compare the different metrics that affect the happiness score of two regions: South Asia and the Middle East.

Report:

The World Happiness Dataset is provided to work on the analysis which consists of the different based factors of calculating the happiness scores and the actual scores of different countries. The provided dataset consists of a total of 143 countries and their scores as well as the data of different factors which is used to calculate the happiness score.

First the dataset is accessed from the drive through google colab by:

```
from google.colab import drive
drive.mount('/content/drive')
```

And the further problems are computed.

Problem-1:

Data Exploration and Understanding:

In the first problem, the task to be performed is to load the dataset and display the first 10 rows. The pandas library as pd is imported and then created a data frame with a variable named df where the dataset is read including a syntax to print the first 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
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

After reading and observing the first ten rows the second question requires to identify the total number of rows and columns. Using the function shape the total number of rows and columns is found from the dataset assigning it to variables total_rows and total_columns respectively.

After the count of total number of rows and columns, the columns are listed with their datatypes using the function dtype.

After the overview, the mean, median and standard deviation is calculated in the problem using the data from the ‘score’ column where the average of column is computed using the function (.mean()), the median of the score column using (.median()) and standard deviation of the column by (.std()) stored in a variable mean_of_score, median_of_score and standard_dev_of_score respectively.

Then the identification of the country with highest score and lowest happiness score is found which results in Finland with highest and Afghanistan with the lowest. The function df.loc is used to locate the country as the function specifically observes each of the value of given column when the .idxmax and .idxmin function classifies the country with highest and lowest score respectively. The data is stored in the variable country_with_highest_score and country_with_lowesr_score respectively.

The total count of missing values from the each column is calculated and stored in the variable missing_values using the function isnull() and .sum() for the total count and displayed with their respective column.

Then the dataset is filtered to display the countries with score which is higher than 7.5 and stored in filtered_ds variable. Then the filtered dataset is sorted using the column Log GDP per capita in a descending way. After the sorting of the dataset the top 10 rows of the dataset is printed where the dataset only provides three countries as those countries only consisted of score higher than 7.5.

A new column is added in the dataset by first declaring an empty list and then with the conditions given as:

Low − (Score < 4)

Medium − (4 ≤ Score ≤ 6)

High − (Score > 6)

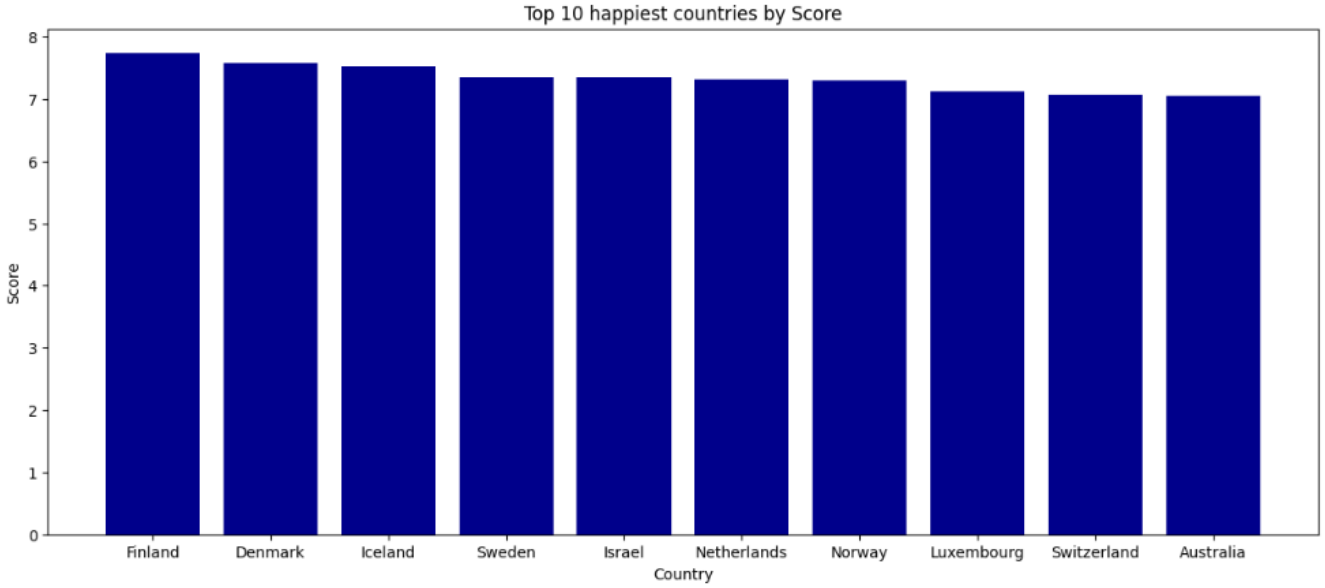
There is If else statement implemented where score of a country is less than 4, then it is categorized as low and the “Low” value is inserted in the empty list using append function. Similarly, if the value is <=4 and <=6, “medium” value is appended and if the score doesn’t meet the two

conditions the else statement is executed resulting to append of “High” value to the empty list.

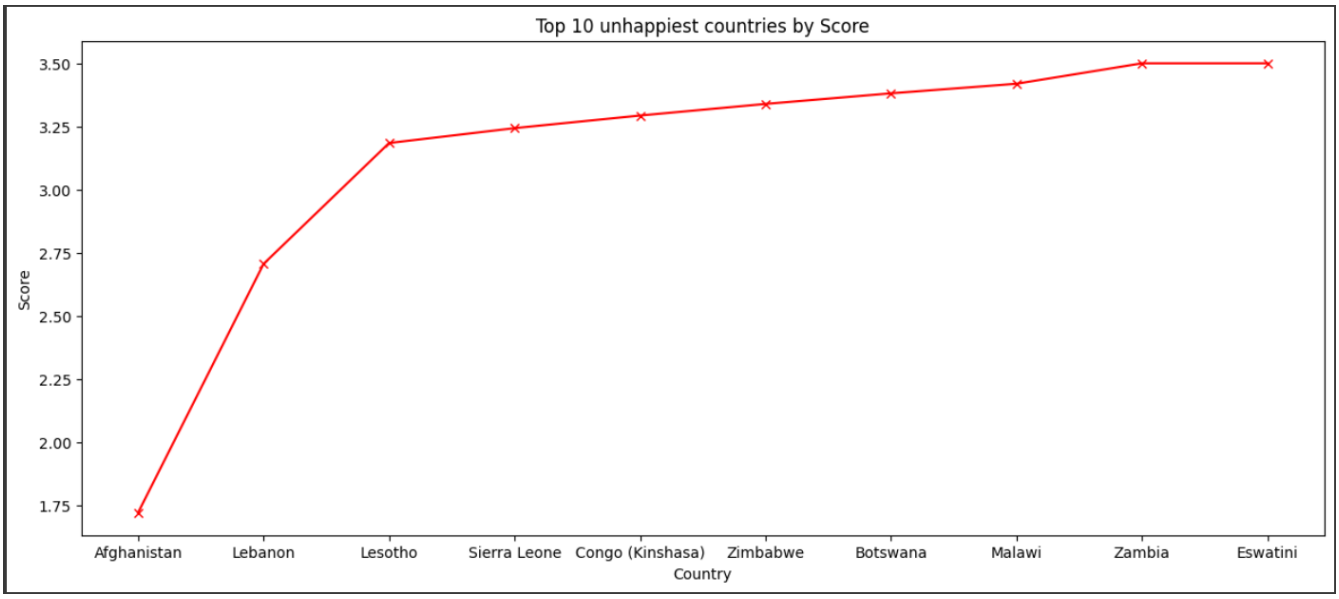
Data Visualization:

First we have imported matplotlib.pyplot to plot the different visualization of the given data in different forms.

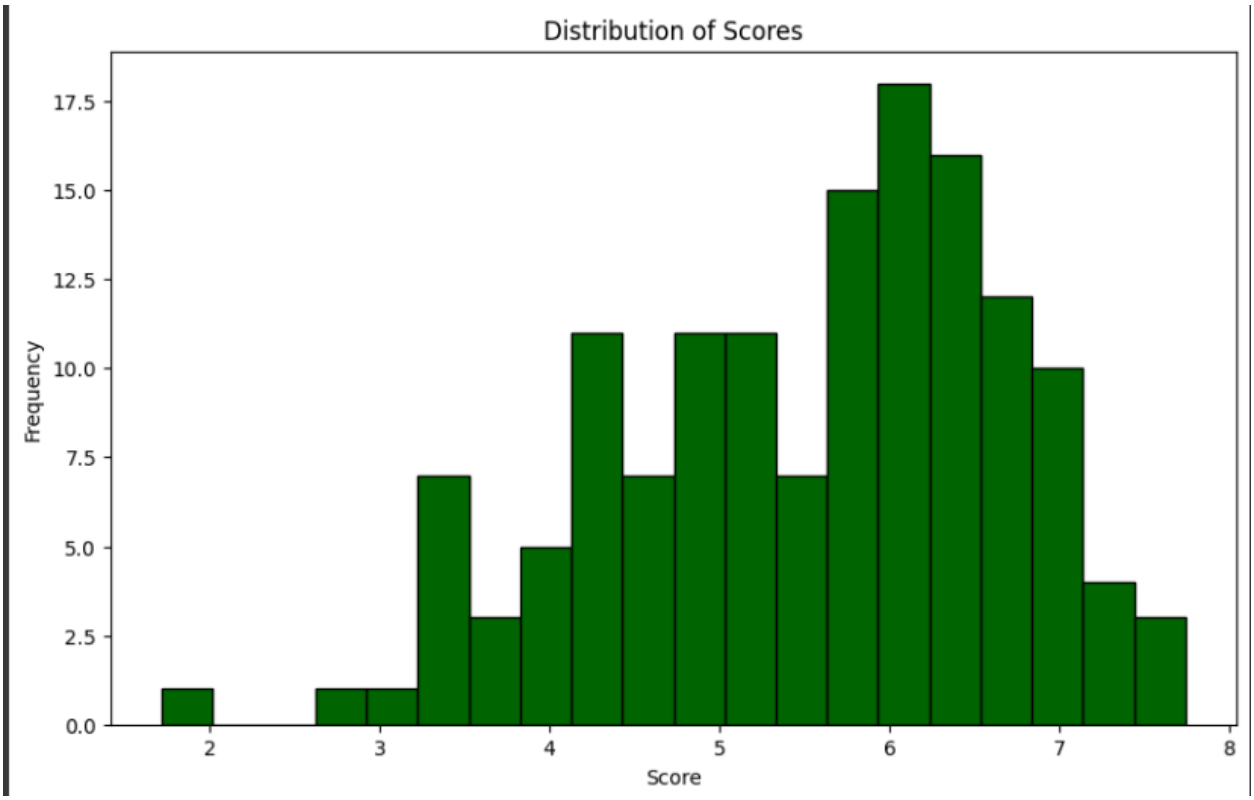
In the data visualization, firstly the bar plot for the top 10 happiest country is plotted with the names based on their score using .nlargest function. Observing the bar plot, Finland is listed as the happiest country followed by Denmark, Iceland, Sweden, Israel, Netherlands, Norway, Luxembourg, Switzerland and Australia.



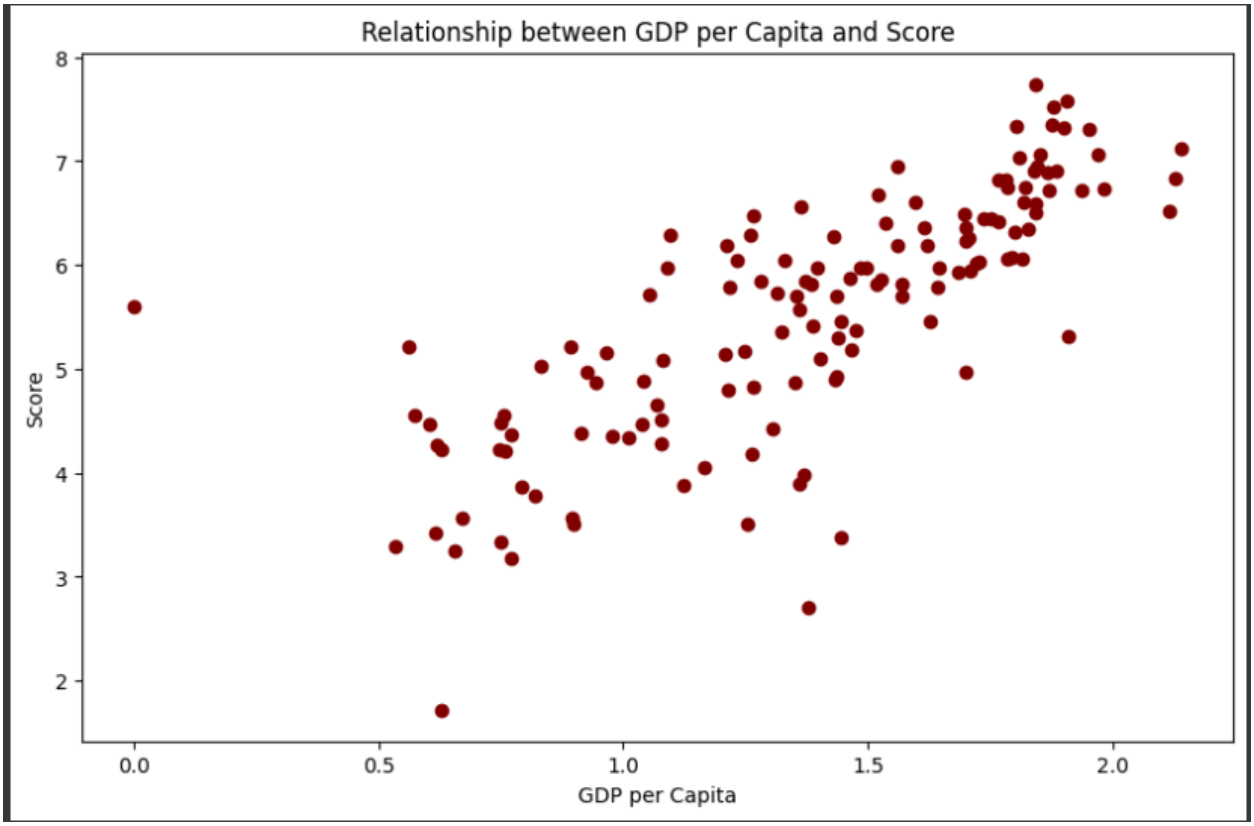
And the second visualization required to plot a line plotting of top 10 unhappiest countries by score where .nsmallest function is used to identify the countries with lowest happiness score. And we can observe that afghanisthan has the lowest happiness score followed by Lebanon, Lesotho, Sierra Leone and more.



Then the histogram is plotted for the score column to represent the distribution of the score column data.



Then the scatter plot is plotted in the task based on the relationship between the Log GDP per capita and score of each countries where we can observe the relationship between the two columns is similar.



Problem 2:

Task-1: Preparing the South Asia Dataset:

A variable is declared with the name south_asian_countries with the given list of countries as the value. Then the list is filtered out if the data set is present in the new data containing the south Asian countries using .isin function and stored in variable filtered_dataset. Then the filtered data set is separated and stored in a newly created csv file named filtered_dataset.csv with index=false using .tocsv function.

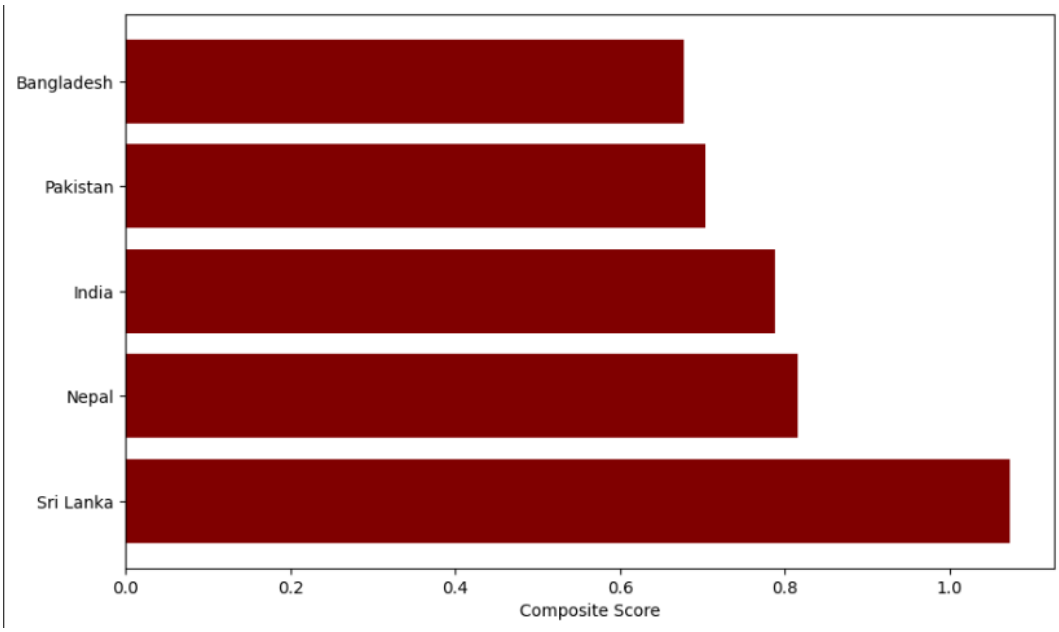
Task-2: Composite Score Ranking:

Using the filtered dataset, a new column is added to the dataset with column name ‘Composite Score’ using the formula given as:

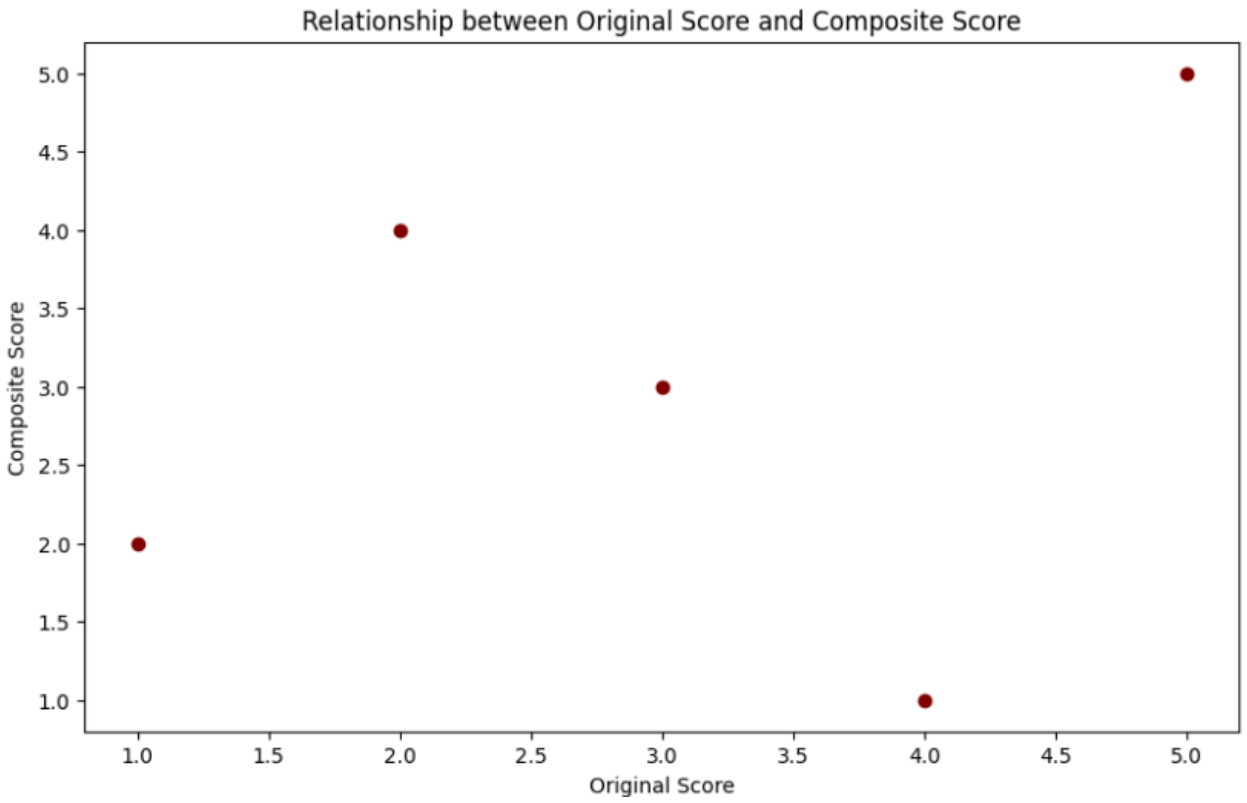
$$\texttt{Composite Score} = 0.40 \times \texttt{GDP per Capita} + 0.30 \times \texttt{Social Support} + 0.30 \times \texttt{Healthy Life Expectancy}$$

Then after creating a new column, the countries are ranked in a descending order based on the new composite score computed in the previous step using .sort_values and stored in variable south_asian_countries_rank.

Then the top 5 countries from the ranked data is visualized using a horixontal bar graph based on their respective composite score with their respective country name.



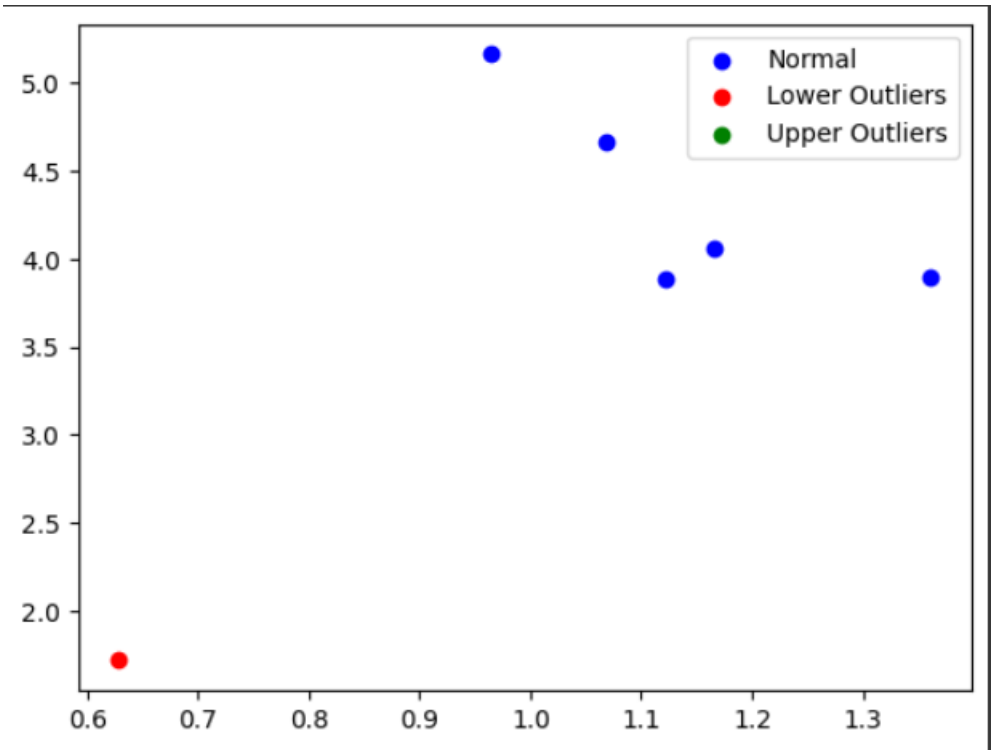
Then the relation between the composite and original score is visualized in the task with the help of scatter plot. First the filtered_dataset is ranked in a descending way based on composite score and similarly ranked in a descending way based on original score. Then the scatter plot visualization is observed and discussed showing the relationship between the original and composite score of the countries.



Task-3: Outlier Detection

While identifying the outlier, we calculate the q1 and q3 using quantile function and iqr is calculated based on the score then the outliers is defined using the 1.5 x iqr rule for the score-based calculation. Then for the iqr of Log GDP per capita-based calculation is performed to obtain the outliers.

Then a scatter plot is made using the Log GDP per capita on the x axis and score on the y axis showing the outliers in a different color.

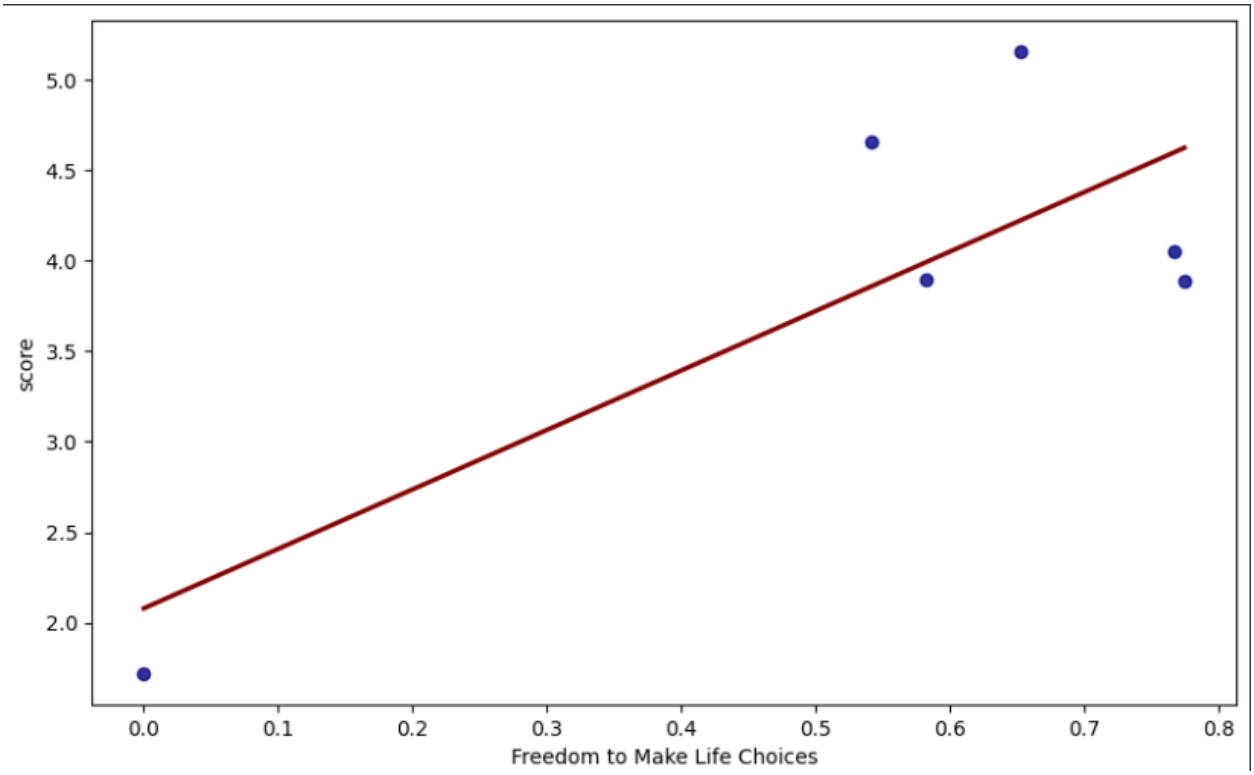


Observing all of the calculations and the scatter plot, Afghanistan has the lowest score and Log GDP per capita, indicating economic underdevelopment when compared to the other four South Asian countries. It also has low values in certain linked metrics. Afghanistan's lower score and Log GDP per capita may cause regional averages to fall, misinterpreting South Asia's overall social progress and economic development.

Task-4: Exploring Trends Across Metrics

In this task, two metrics Freedom to make life choices and generosity is chosen to calculate correlation with the score for the filtered south Asian countries.

A scatter plot is created with trednlines against the score of the countries.



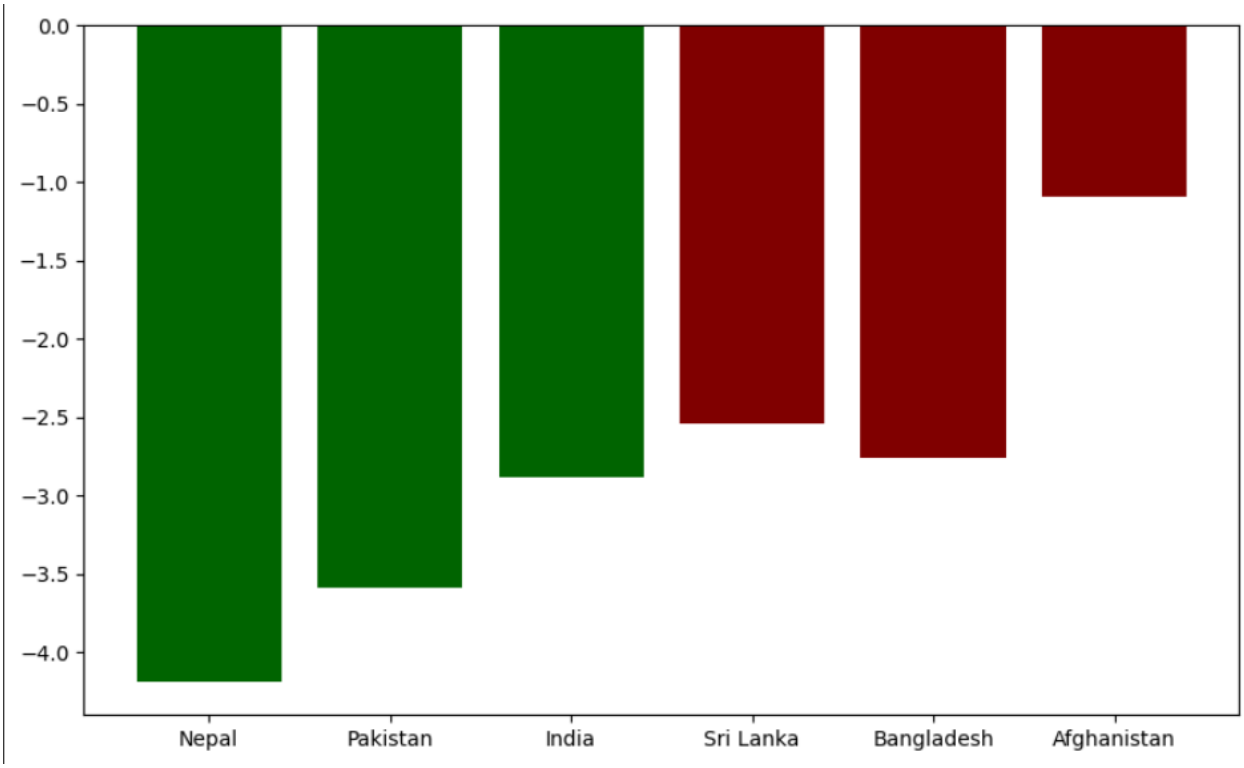
After calculating the Pearson Correlation Coefficients of both freedom to make life choices and Generosity with score, we can observe that the values of both the coefficients are close to 1 which indicates a strong positive relationship of freedom to make life choices with score and generosity with score. This results that the score increases when the freedom to make life choices increases, and when the generoisty increases, score increases as well.

We can see that the trendlines for freedom to make life choices and generosity with score is found to be upward slope which shows that it has a positive relation that is higher generosity or higher freedom to make life choices is associated with higher score.

Task 5: Gap Analysis

A new column is added with the variable name GDP_Score_Gap which computes and stores the difference between the log gpd per capita and score. Then the countries are ranked based on the GDP score gap in a descending way as well as ascending way.

The top three countries with largest positive and negative gaps are visualized using a bar chart and analyzed the reasons behind the gaps.



Here, we know that the GDP-Score Gap is the difference between the values in the Log GDP per capita and score columns. The gap has the following implications for South Asian countries: the GDP-Score Gap values are all negative, indicating that the score is substantially larger than Log GDP per capita. So, the bar graph we created for the top three countries with positive and negative gaps extends downward to reflect negative values.

Problem 3:

Task :

Firstly, similar as problem 2, the list of countries are provided in the question to create a data frame out of which consists of middle eastern countries. Then the countries are filtered whether it is present in the data frame using the (.isin) function. After filtering, the null value containing rows are dropped using dropna function.

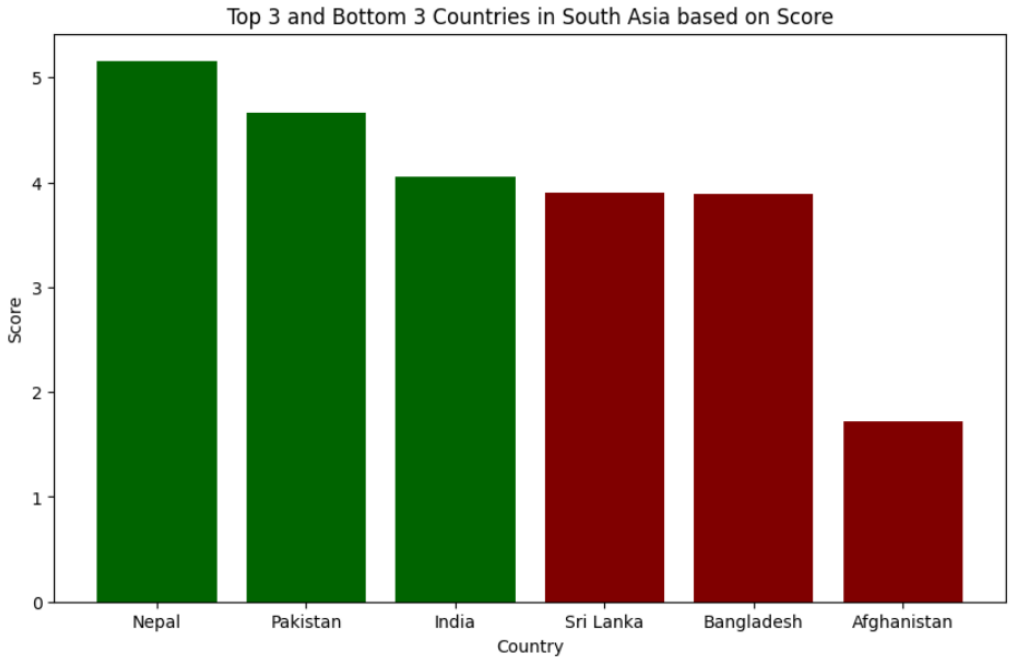
1. Descriptive Statistics:

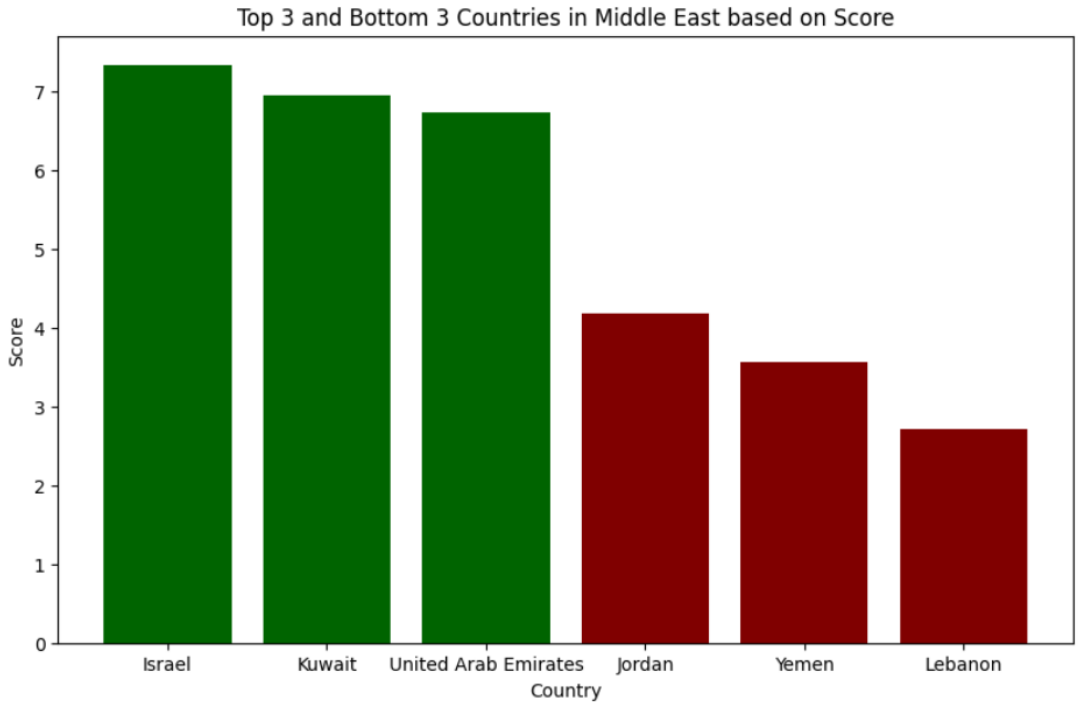
The mean and standard deviation of the south Asian data is calculated as well as the mean and standard deviation of the middle eastern data suing the mean and std function. And then the region with higher happiness score on average is observed using if else function with applying a condition of which has the higher score averages results in printing the name of the region.

2.Top and Bottom Performers:

In this task the dataset is classified in top3 based on score from South Asia region as well as Middle east and bottom 3 based on score from South Asia as well as Middle East suing head and tail function.

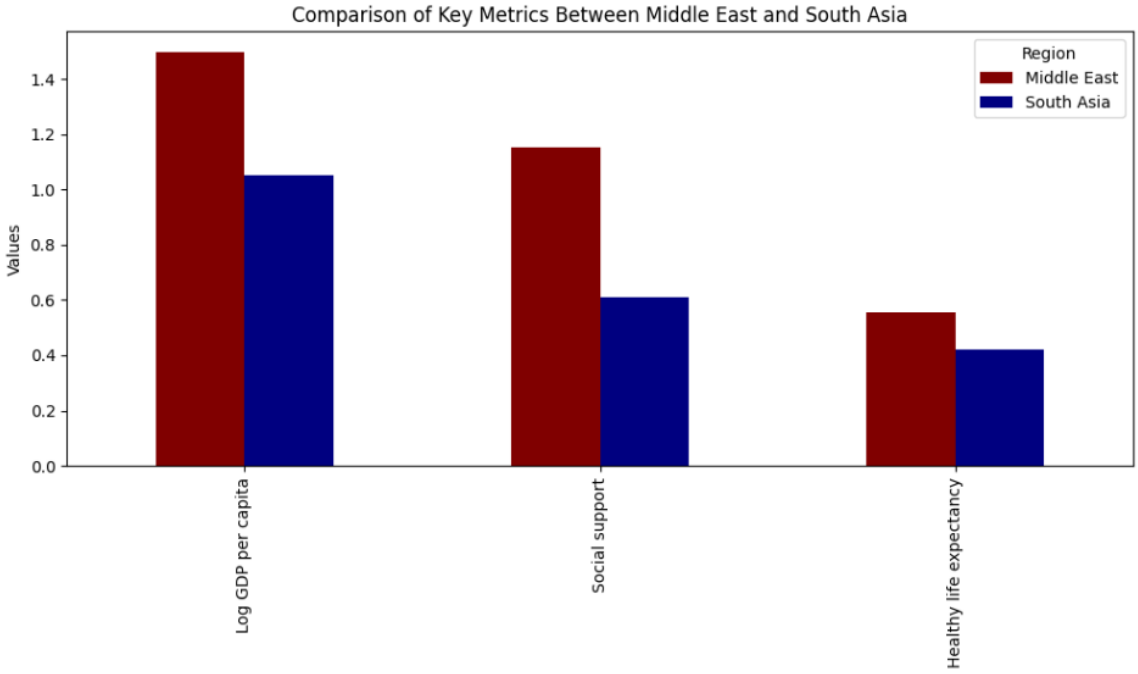
Based on the top 3 and bottom 3, two bar graphs are plotted showing the visualization of both South Asia and Middle east.





Metric Comparisons:

The metrics of the south asia and middle east is calculated based on the key metrics such as country name, log gdp, social support and healthy life expectancy and compared using a grouped bar chart.



The comparison of important metrics between the Middle East and South Asia is shown in the bar graph above. We may observe that the two regions' values for the Social Support metric differ more. As a result, the biggest difference between the two regions is seen in social support.

4.Happiness Disparity:

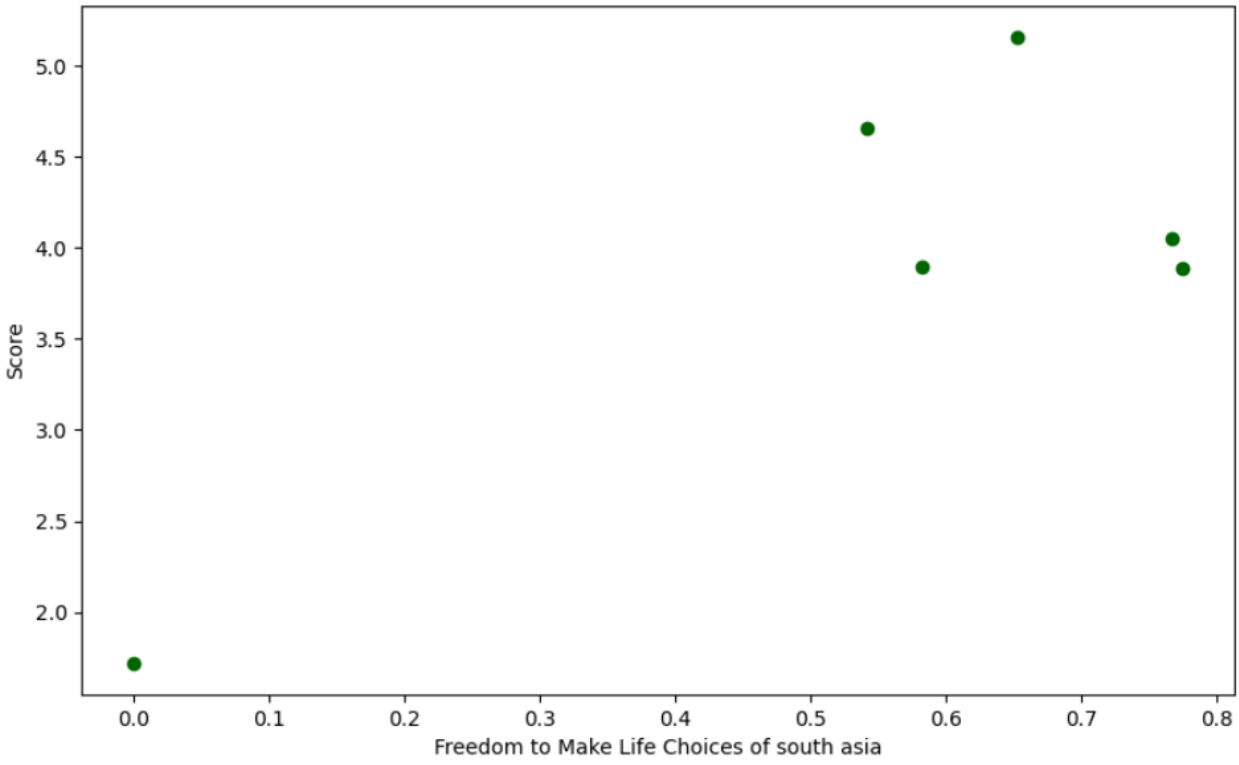
The difference of maximum and minimum based on score of happiness is computed and stored in variable range_of_score_south_asia and range_of_score_middle_east. The coefficient of variation based on score is calculated for both south Asian and middle east region using the data of standard deviation and mean computed in the previous tasks.

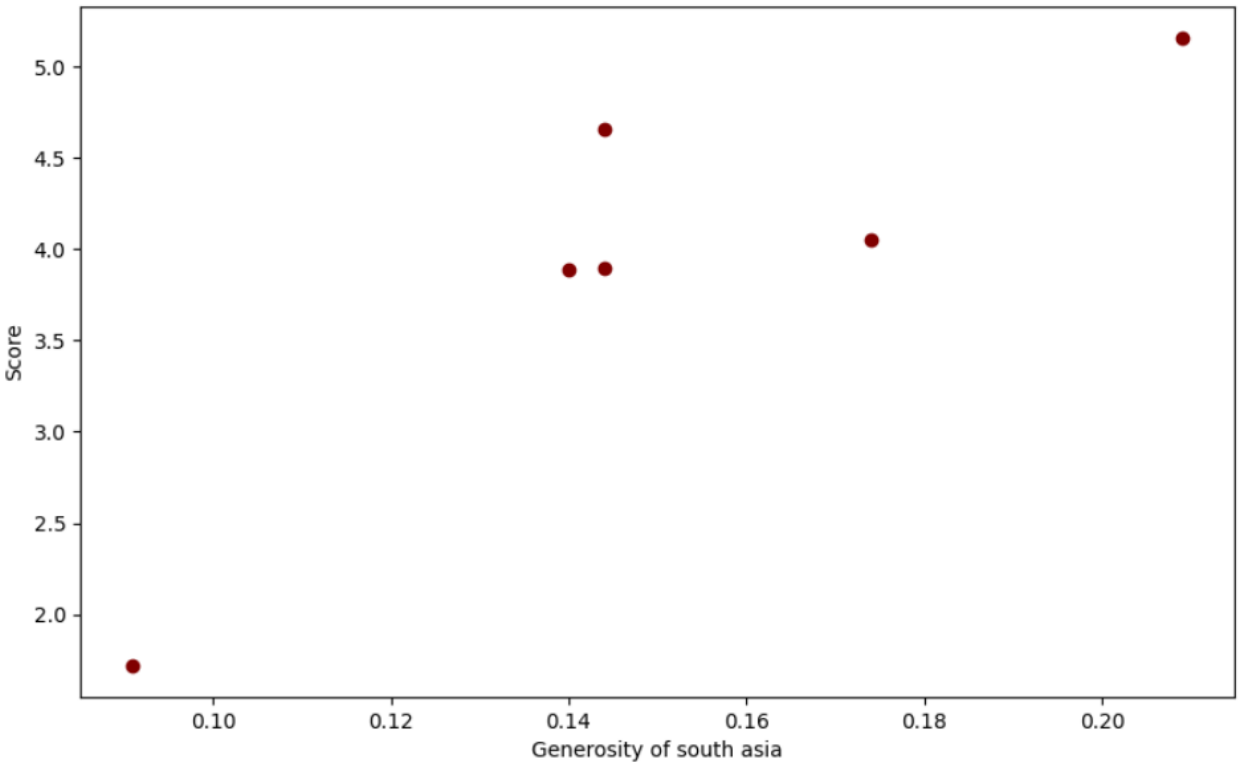
The middle east region was is observed to have greater variability in happiness and it is obtained by using if else conditional statement.

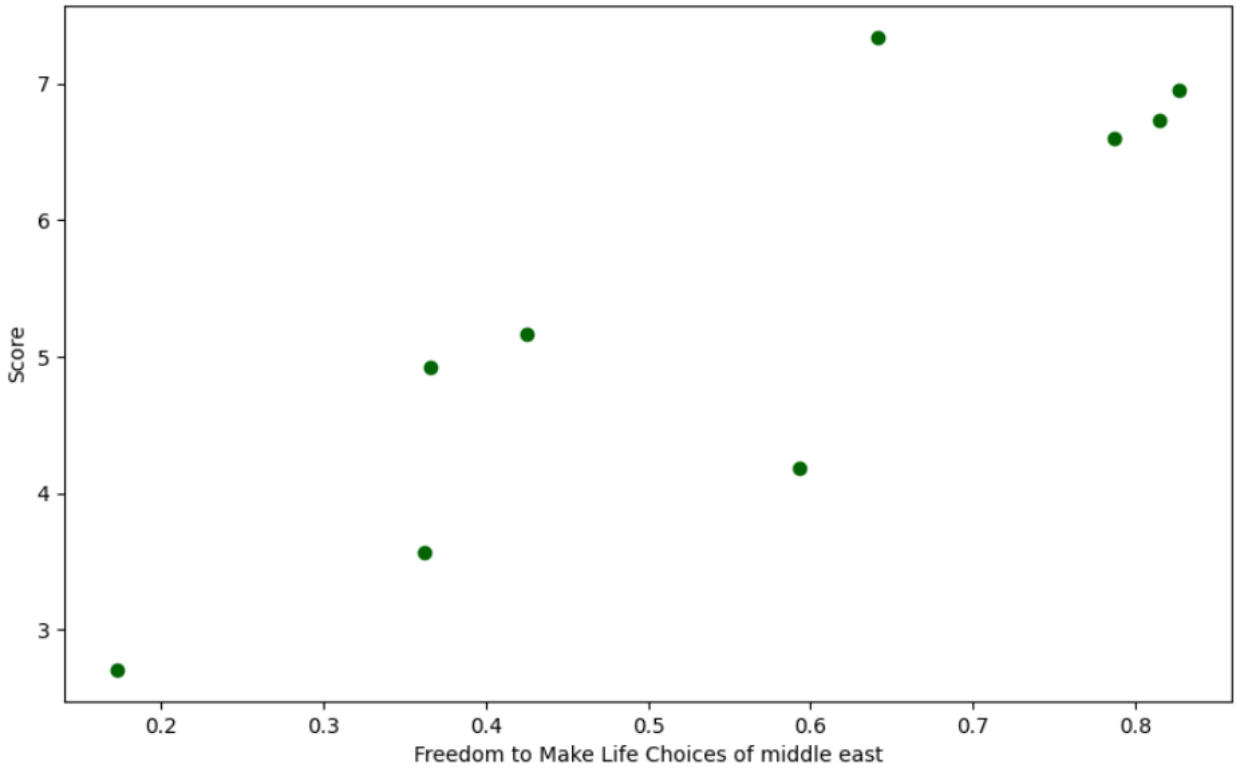
5. Correlation Analysis:

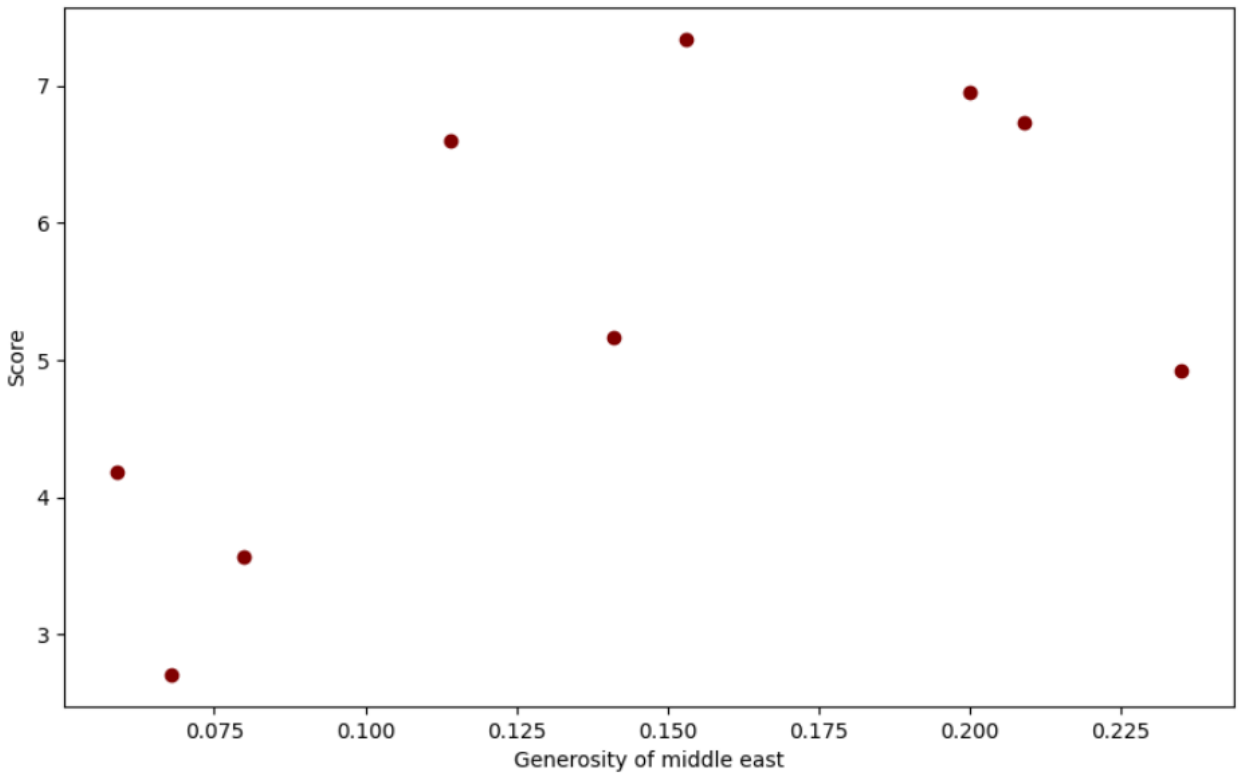
In this task, the correlation based on score with freedom to make life choices and generosity is calculated using the corr function within both the south Asian and middle east region each.

After this, scatter plot of the freedom to make life choices and generosity is computed of both the regions south asia and middle east.




Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
0.10
0.12
0.14
0.16
0.18
0.20
Generosity of south asia

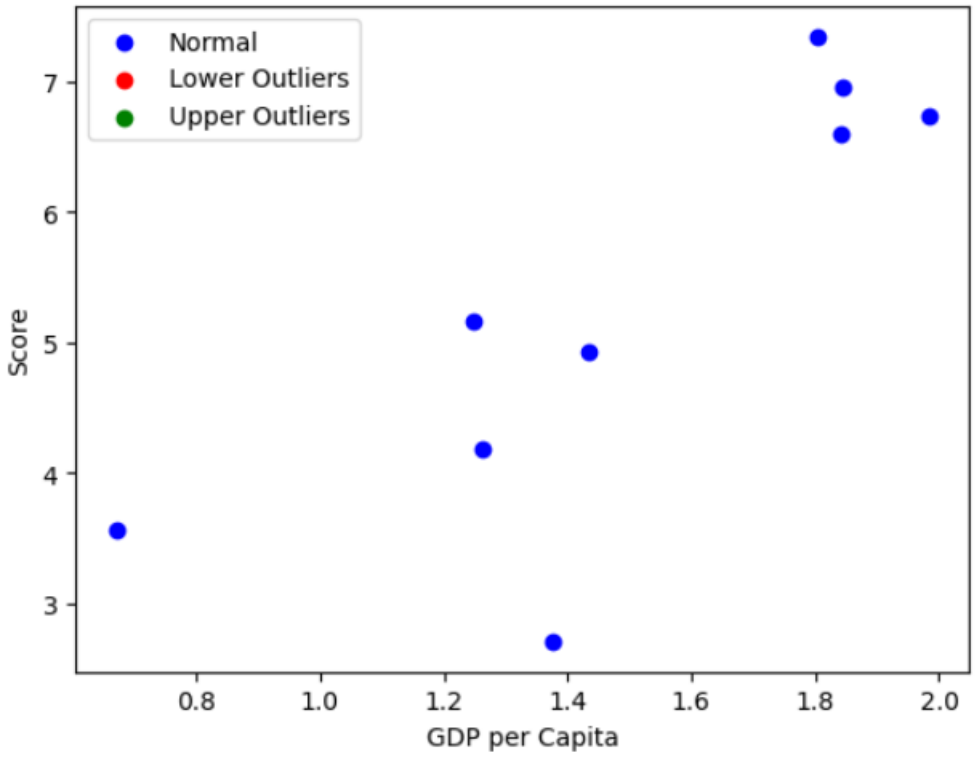

Score
7
6
5
4
3
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Freedom to Make Life Choices of middle east



6.Outlier Detection:

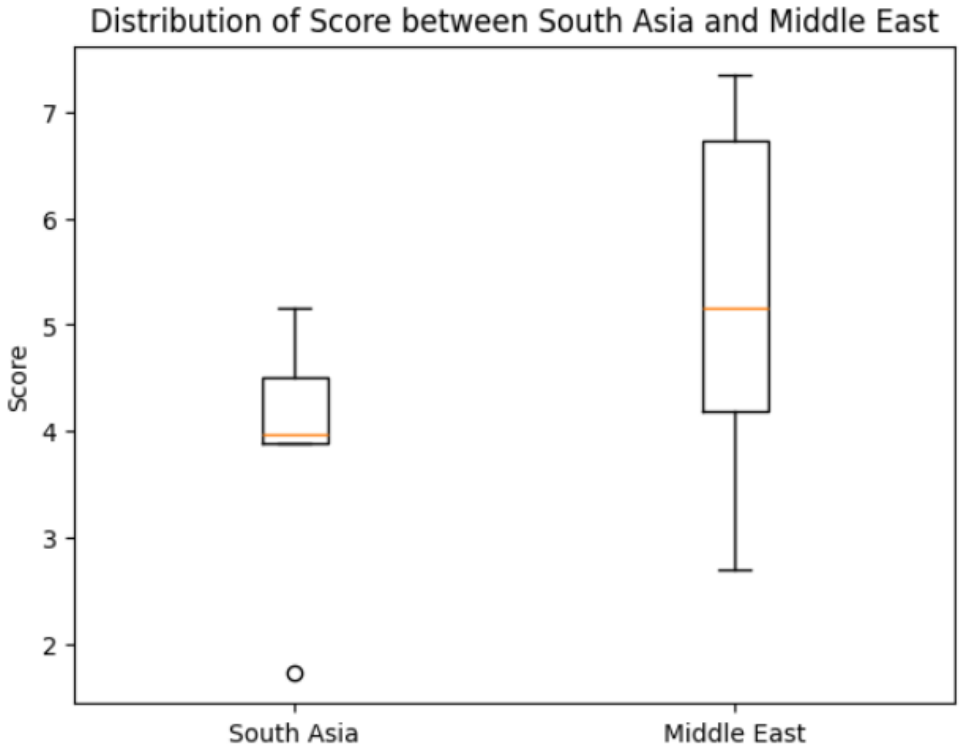
The outlier is calculated of the middle eastern region using the quantile function and iqr, computing lower as well as upper bound based on score and GDP per capita.

The outliers are plotted based on the column of log GDP per capita and score using scatter plot.



7.Visualization:

Boxplot is plotted which compared the distribution of score between the south Asian countries as well as middle eastern countries.



Shape: The box plot's length indicates the interquartile range (IQR), and the Middle East region has a higher IQR and a wider range of scores than South Asia. The range of non-outlier data that falls between the lower and higher bounds is indicated by the margins outside the box. Compared to the Middle East, South Asia has less variety because its whiskers are often shorter.

Median: We can see from the boxplot above that the middle line inside the Middle East box is higher. This indicates that, in comparison to South Asia, the Middle East region has a higher average score.

Outliers: The dots outside the whiskers are known as outliers. We can see that the Middle East region has no lower outliers, whereas South Asia has one below the bottom whisker, indicating the occurrence of lower outliers. This suggests that there is a nation in the South Asian region with an exceptionally low score.

Conclusion:

In conclusion, the findings that can be observed in Problem 1 is that GDp per capita is strongly related to happiness score. The findings that can be observed in Problem 2 is that Afghanistan is an outlier country in South Asia, both metrics GDP and Freedom to Make Life choices have a positive relationship with happiness score and there is a negative gap between gdp and score indicating the values of score is greater than that of gdp. The findings that can be observed in Problem 3 is that the Middle East region has greater variability and disparity than the South Asia region and no outliers are present in the Middle East region. Therefore, the main objective of this report is to perform a data exploratory task in the provided dataset and provide visual interpretations for the relationships between regions or metrics included in the dataset.