**Herald College, Kathmandu**

****

**Concepts and Technologies of AI**

**5CS037**

Assignment-1 - Statistical Interpretation and Exploratory Data Analysis

Analysis of the World Happiness Report: A Data-Driven

Exploration of Global and Regional Trends.

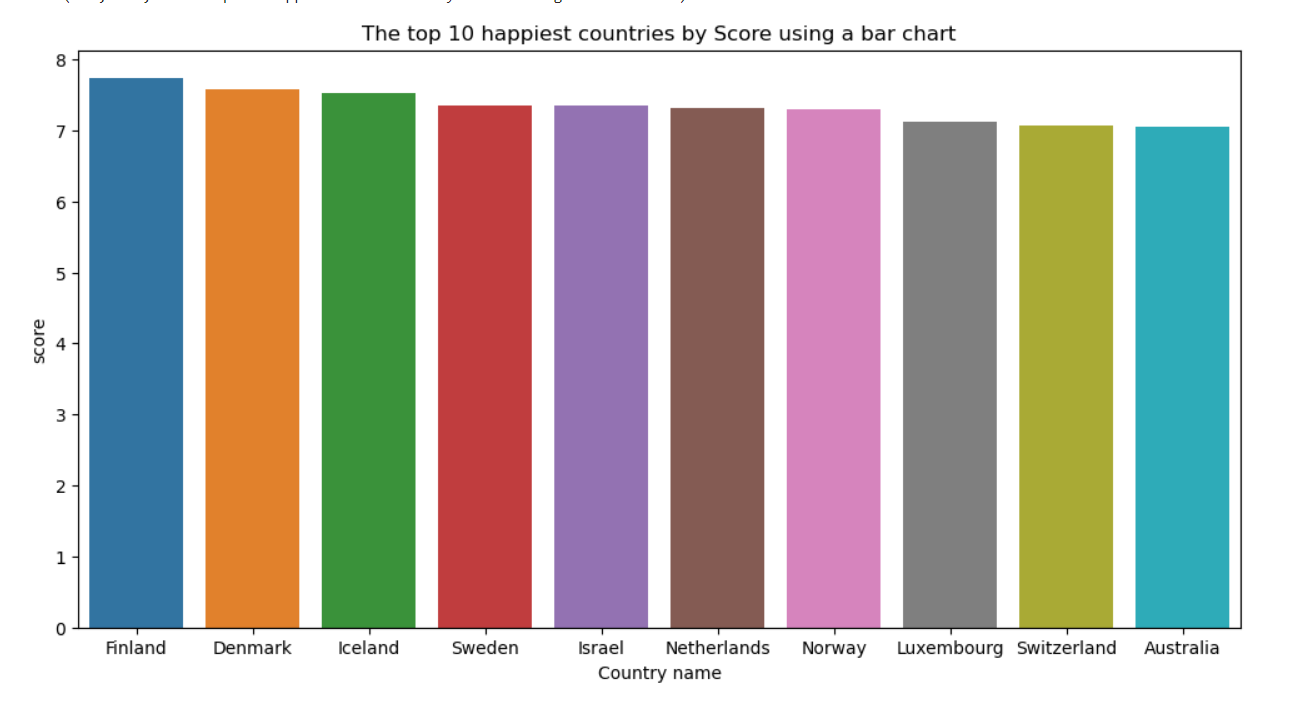
December 02, 2024

Interpretation for the Numerical data of original dataset (WHR-2024-5CS037.csv)

* The given data set is about the details of the country and their important indexes i.e. Score, GDP per capita. There are many countries. When the shape of the dataset is calculated the shape was found to be (143 rows & 9 columns).
* The repeated data types of the given original dataset were found to be ‘float64’ and the unique data type was found to be Object that is ‘Country name’.
* While getting the basic statistics data for the dataset, the following the dates were found:
  + Count: 143
  + Mean: 5.52
  + Standard Deviation: 1.17
  + Minimum value: 1.721
  + First Quartile: 4.72
  + Second Quartile: 5.785
  + Third Quartile: 6.416
  + Maximum Value: 7.741
* The highest score on the original dataset is 7.741 which happens to be of Finland.
* The lowest score on the original dataset is 1.721 which happens to be of Afghanistan.
* While checking for any missing values in the dataset , it appears that all the column has 3 missing values except for the column ‘Country name’ & ‘score’.
* While filtering for the score whose value were greater than 7.5 was found to be only of three countries i.e. Finland, Denmark, Iceland.
* We are now adding the new column ‘Happiness\_category’ to the dataset which categorizes on the basis of score. We made a function which categorizes the county happy or not according to the score it has.

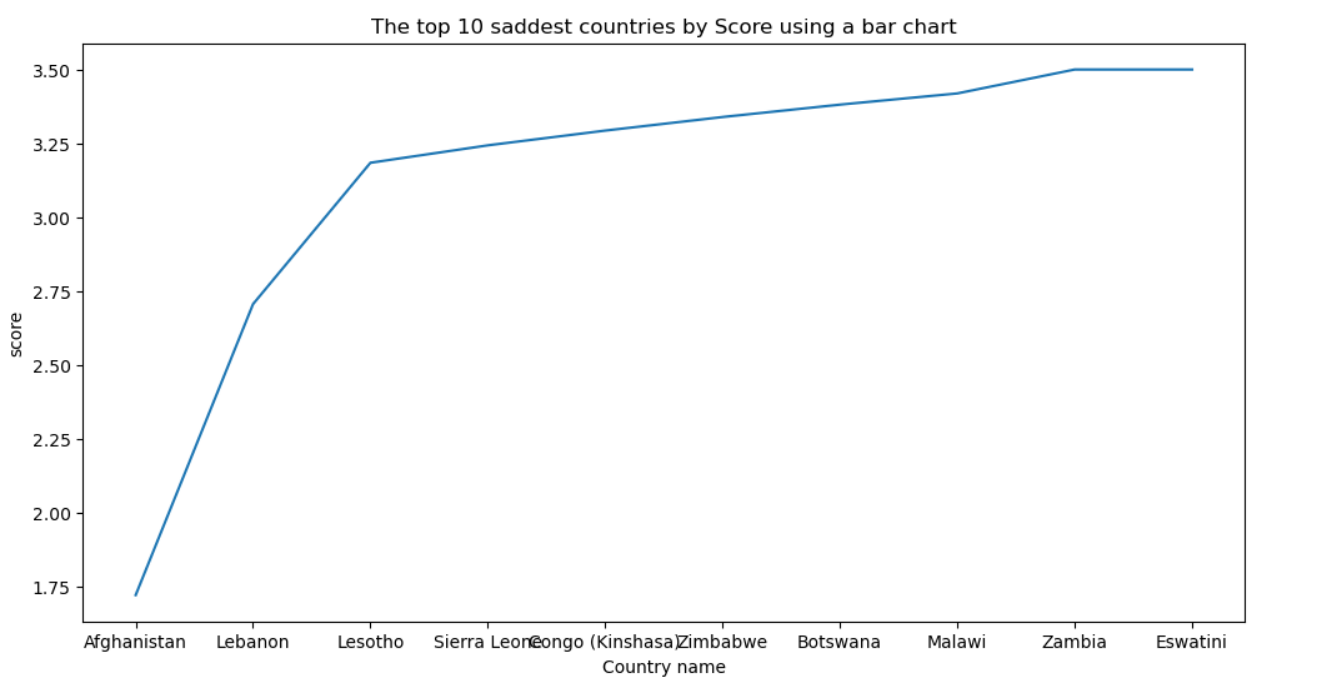
1. Interpreting the visualization of original dataset (WHR-2024-5CS037.csv)

While plotting the most happiest country from the dataset , the followings were my founding from analyzing the data.

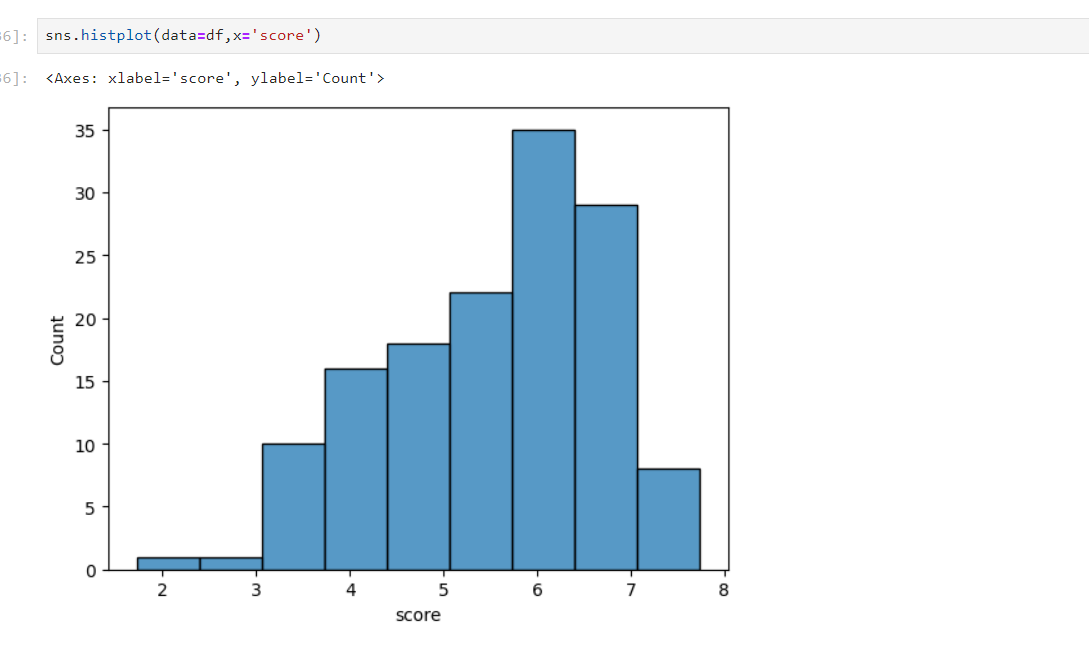


According to the above bar chart, Finland is the most happy country with the score of ‘7 – 7.8’.

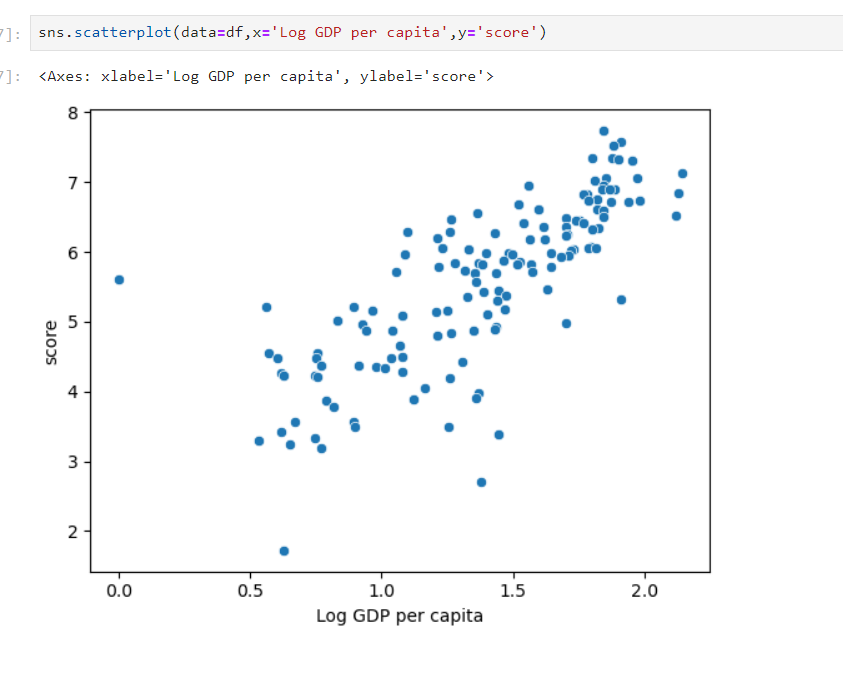
While plotting the line plot for the saddest countries in the dataset, the following seems to be the generalization,



According to the line plot, the saddest the country from the dataset, Afghanistan is the saddest country with score 1.6.



According to the above hist plot, The most common score in the data frame is somewhere between 6 – 7.



From the above scatterplot, it can be seen that the relationship between ‘Log GDP per capita’ and ‘score’ seems to be somewhat linear.

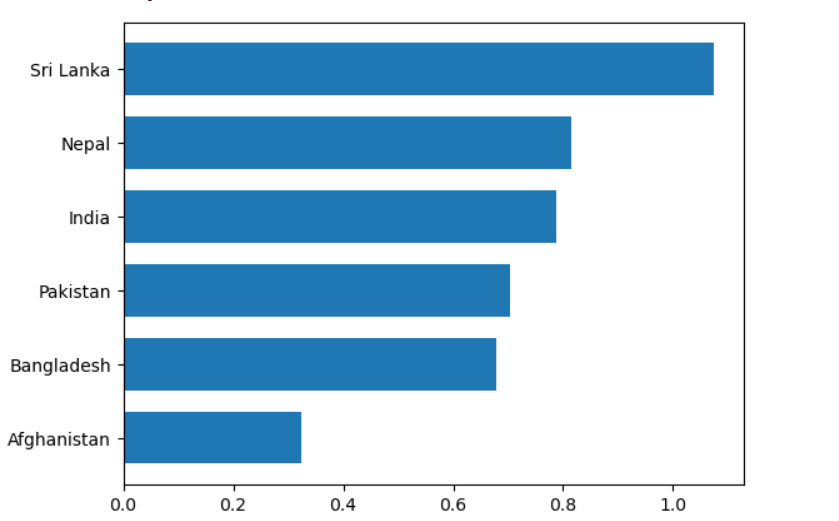
1. Analysis of the filtered dataset which contains south Asian countries.

* At first, I created the function to checks whether the given countries are in the Data frame or not and push into the new csv.
* A new column called Composite score was created by using the formula,

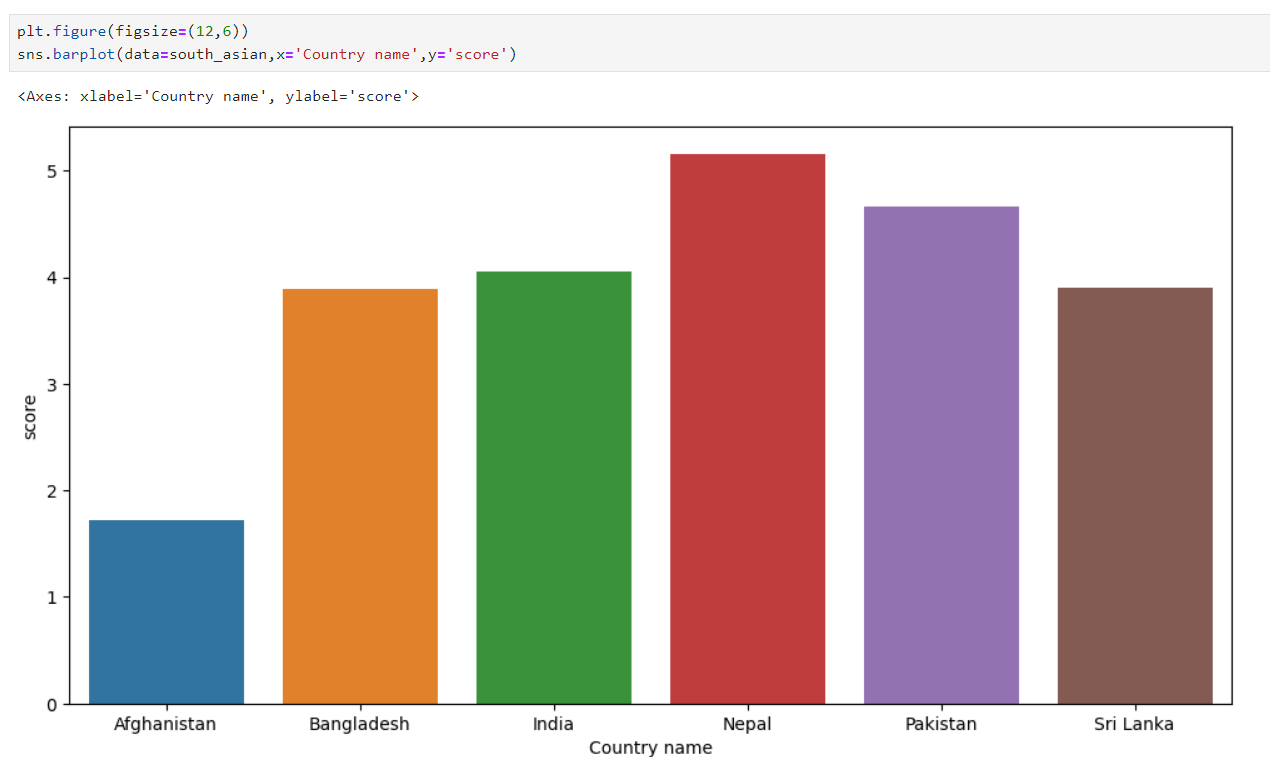
Composite Score = 0.40 × GDP per Capita + 0.30 × Social Support

+ 0.30 × Healthy Life Expectancy.

* The new dataset was sorted in a descending order and was plotted in a horizontal bar graph

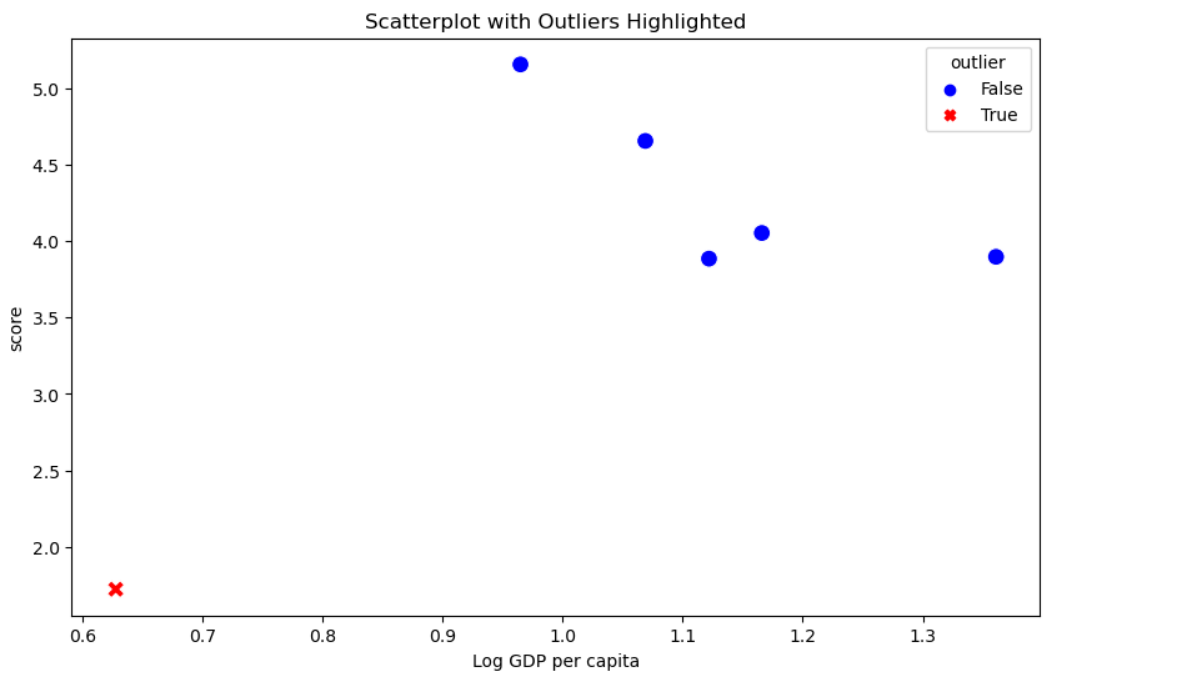


According to the above bar plot , Sri Lanka is at the top as it’s Composite Score is high in dataset of the given South Asian Countries and Afghanistan seems to at the bottom of the list due to its lower Composite Score which is affected by various factors.

But however, the ranking based on the Composite score doesn’t align with the original score as

The following graph denotes that our country Nepal is at the top but in the figure where compared with respect to Composite score Sri Lanka was at the top of the line.

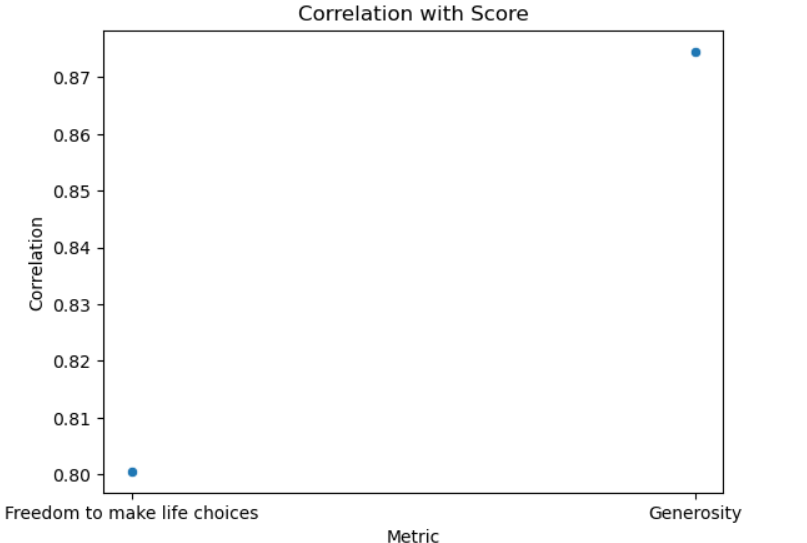
* Outlier Detection



Outlier is the type of data which behaves abnormally from the rest of the data. The upper scatterplot shows the outlier based on the score and Log GDP per capita. In this scatterplot the outlier country was found to be ‘Afghanistan’.

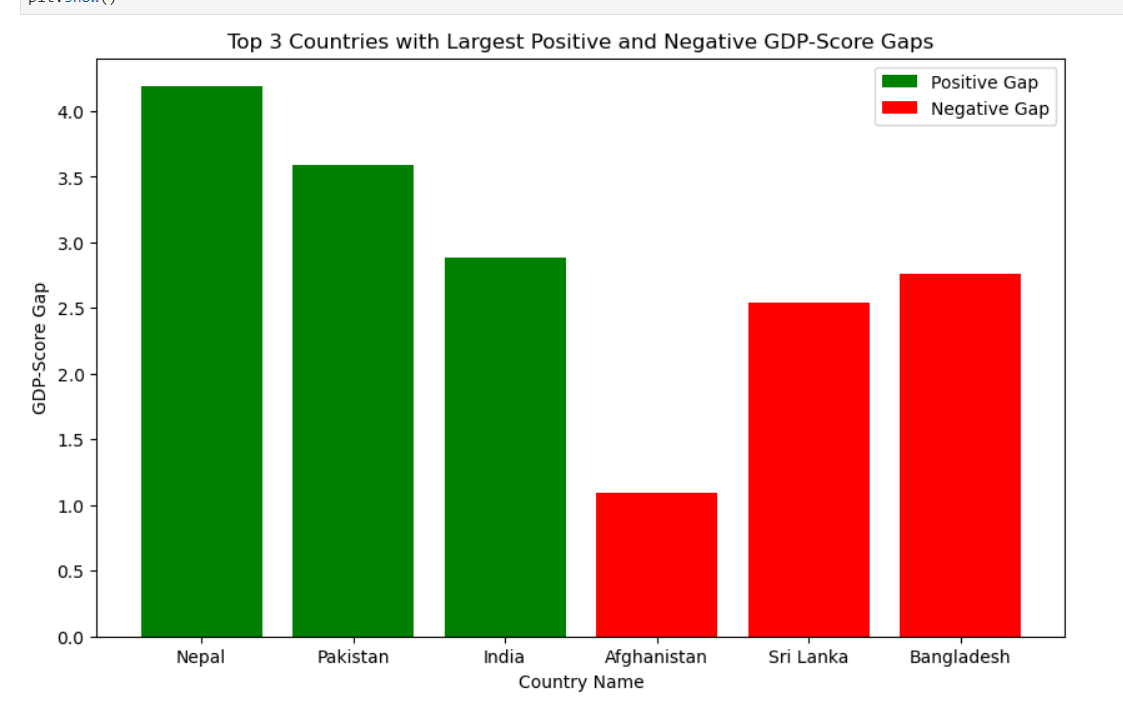
The main characteristic of outlier is that it behaves as a noise in the dataset and can be a potentially dangerous as outlier is generally an exception which may lead our machine learning model to perform badly.

* Exploring Trends Across Metrics



The above scatterplot shows the correlation between Freedom to make like choices and Generosity from the dataset of South Asian Countries. This scatterplot shows that they are correlated with each other. But, Scatterplot is not the efficient way to show the correlation between any variables, heatmap is more efficient and convenient.

* Gap Analysis

We added a new column called “GDP-Score Gap” in the south Asian country’s dataset. So that, we can see the positive and negative performing country’s in the visualization.

The above bar chart, shows the Largest Positive and Negative GDP-Score Gaps of the south Asian countries. According to the bar chart, Nepal is in the top hierarchy in the chart which shows that Nepal has the less gap then that of other south Asian countries.

Problem - 3 - Comparative Analysis

**Descriptive Statistics**

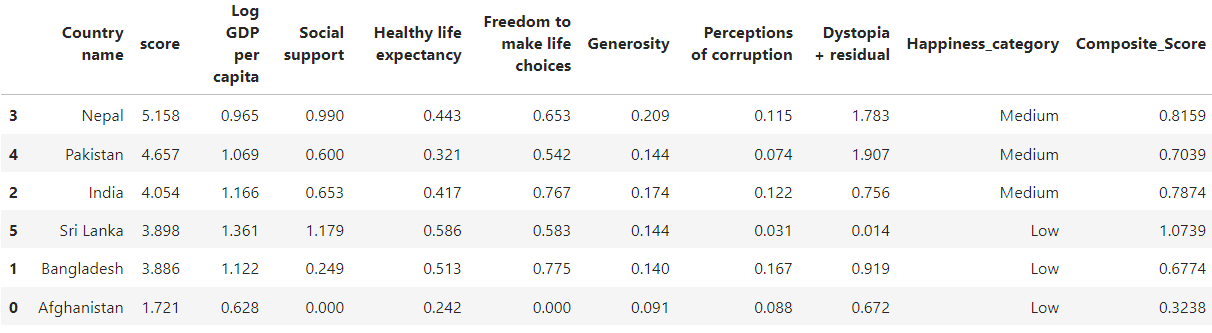
While calculating the basic statistics, the mean and Standard Deviation of the score of the both Middle east and South Asian regions were found to be:

Mean & Standard Deviation (Middle east): 5.351333333333333, 1.648656346847335

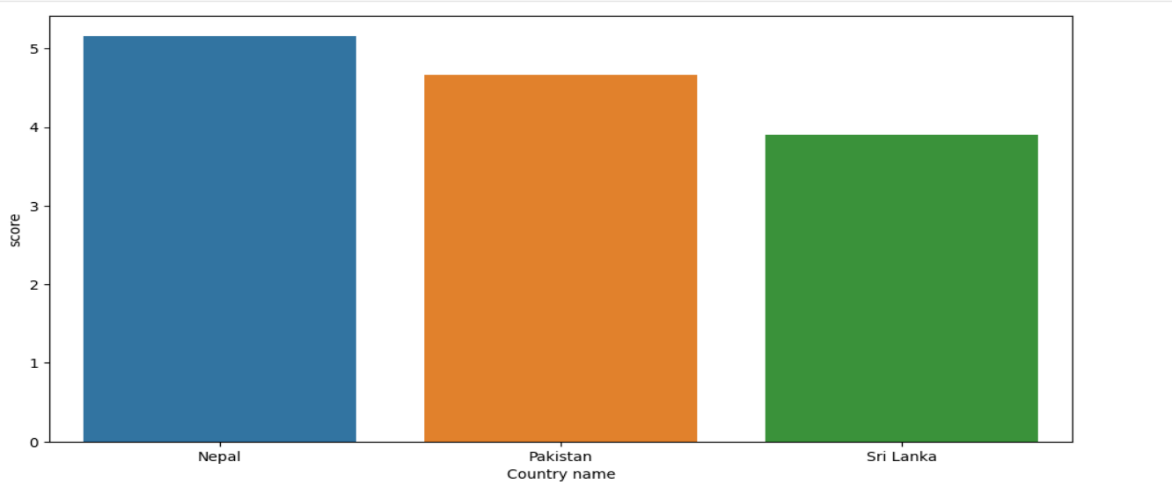
Mean & Standard Deviation (South Asia): 3.895666666666667, 1.1770690152521501

**Top and Bottom Performers**

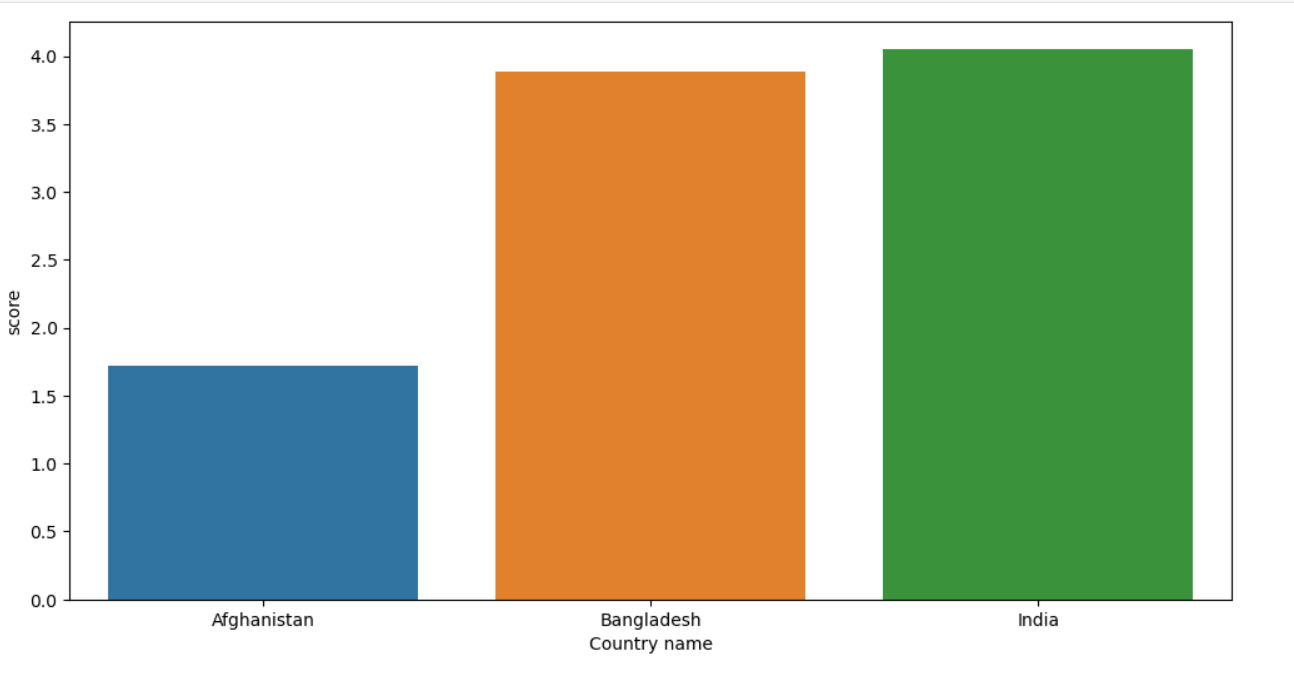
The top and bottom performer countries on the basis of the score are:



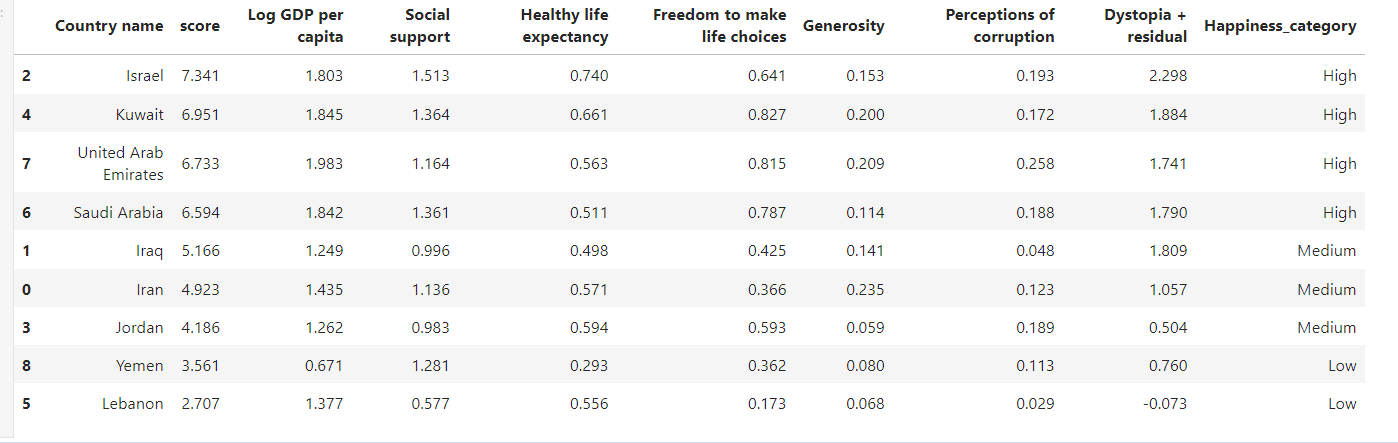
According to the table, Nepal is the top performer on the basis of score, while Afghanistan is the bottom performer with the score of 0.628.



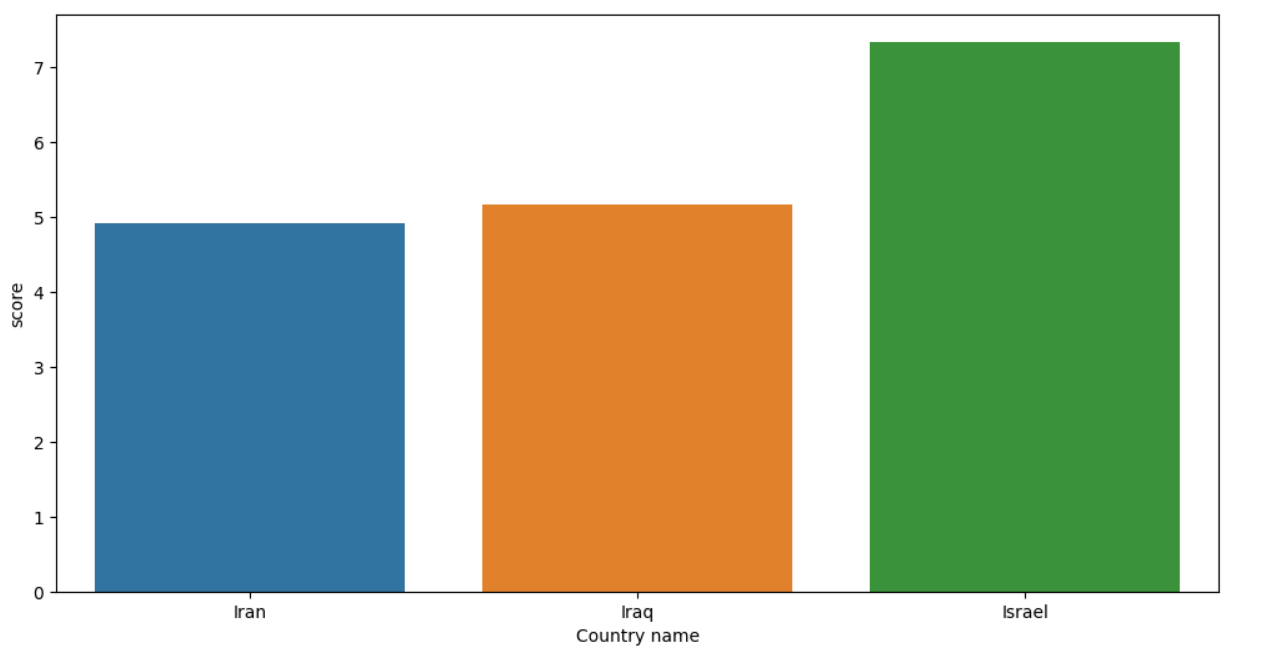
The above given bar chart is of the top performer countries, which includes Nepal, Pakistan & Sri Lanka with Nepal being at top.



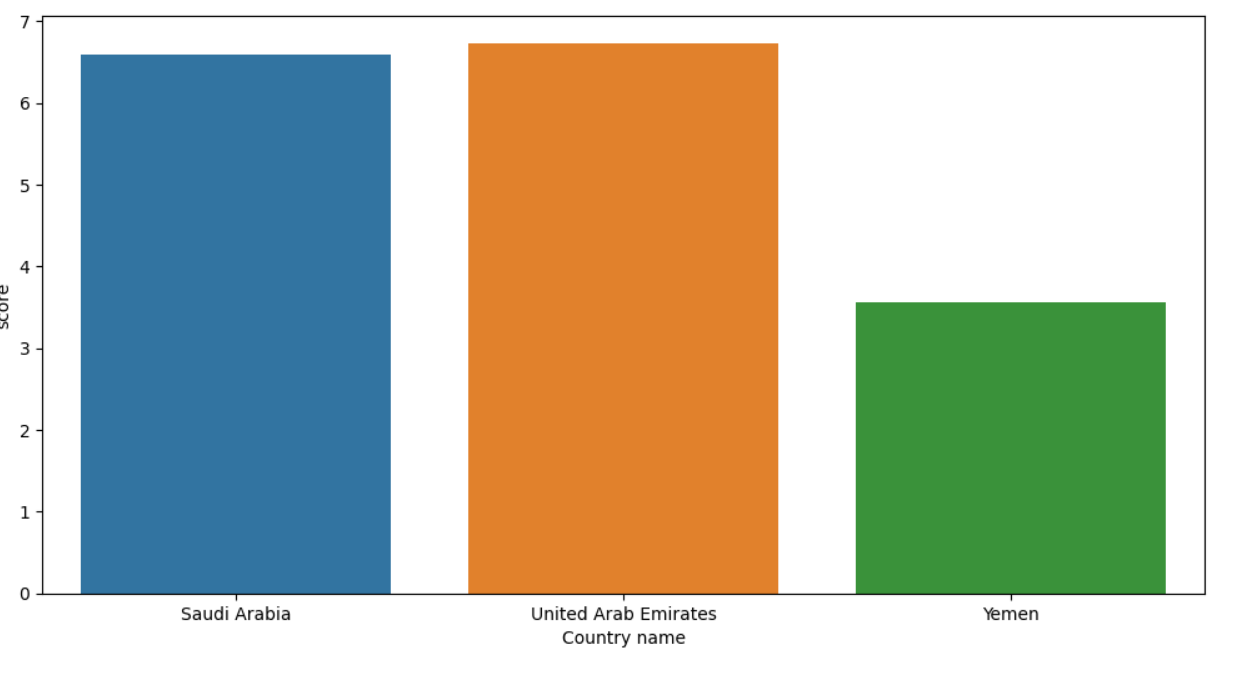
The above given bar chart is of the bottom performing countries, which includes Afghanistan , Bangladesh, India with Afghanistan being at bottom.



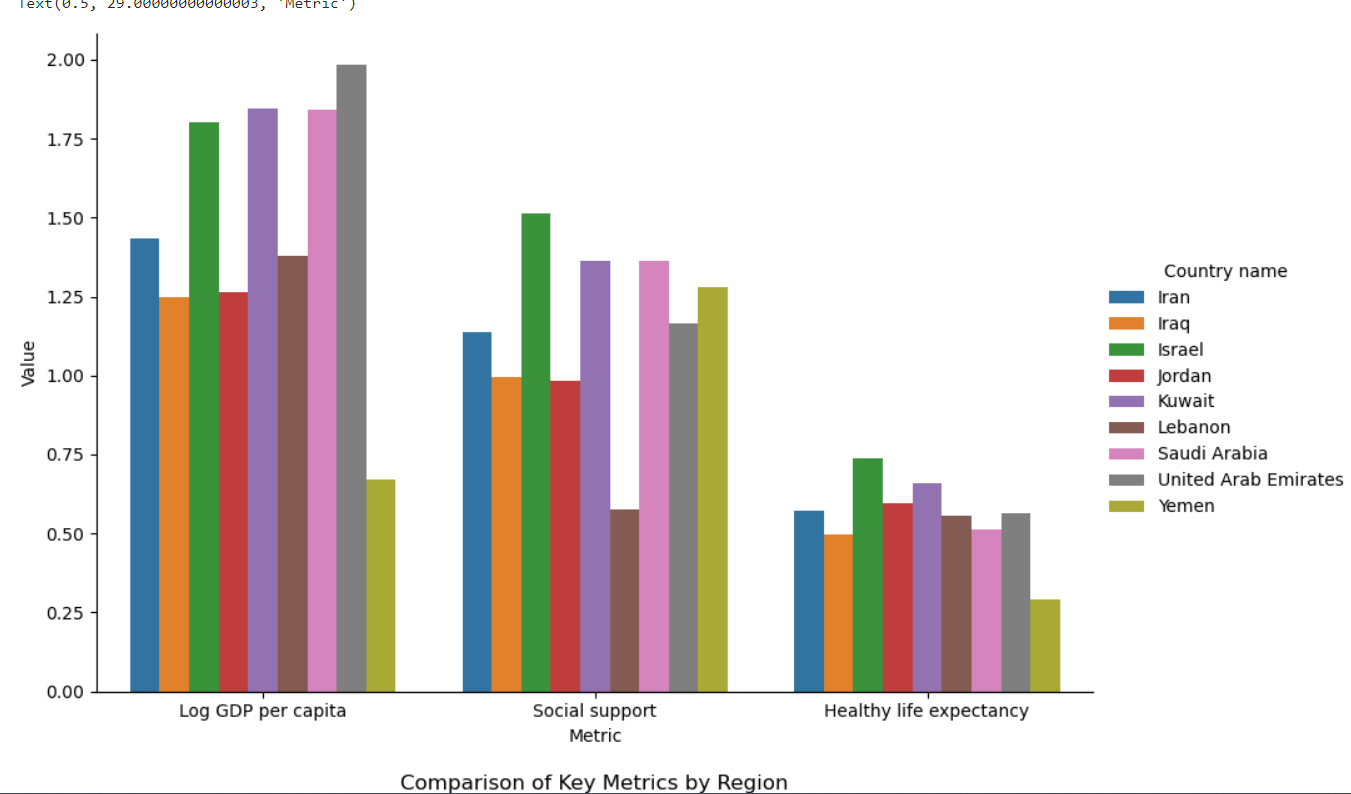
According to the table, Isreal is the top performer on the basis of score, while Lebanon is the bottom performer with the score of 2.708.

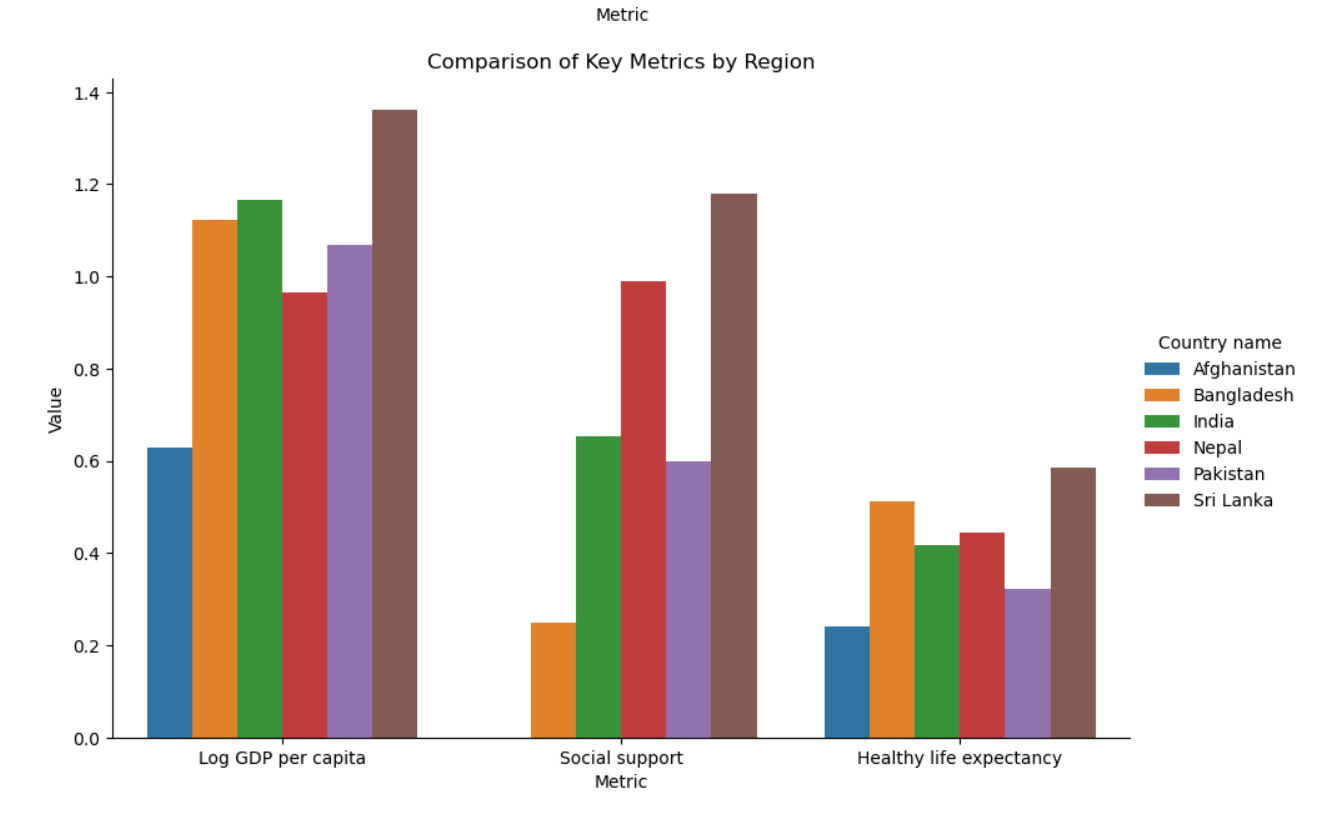


According to the above bar plot, Iseral, Iran and Iraq were the top formers.



According to the above bar plot, it represents the bottom performers from middle east region. Saudi Arabia, UAE and Yemen were the bottom performers.





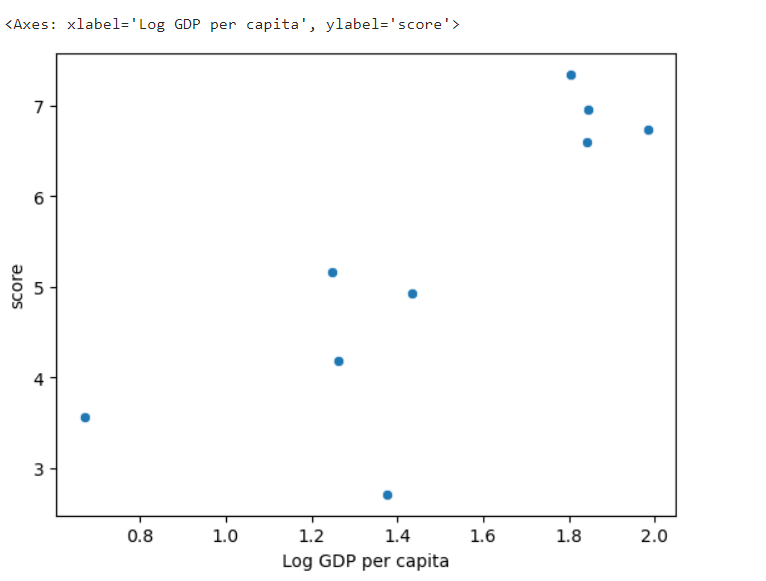
The above categorical plot , shows that the middle east is more developed and has more average score then that of South Asia.

**Happiness Disparity**

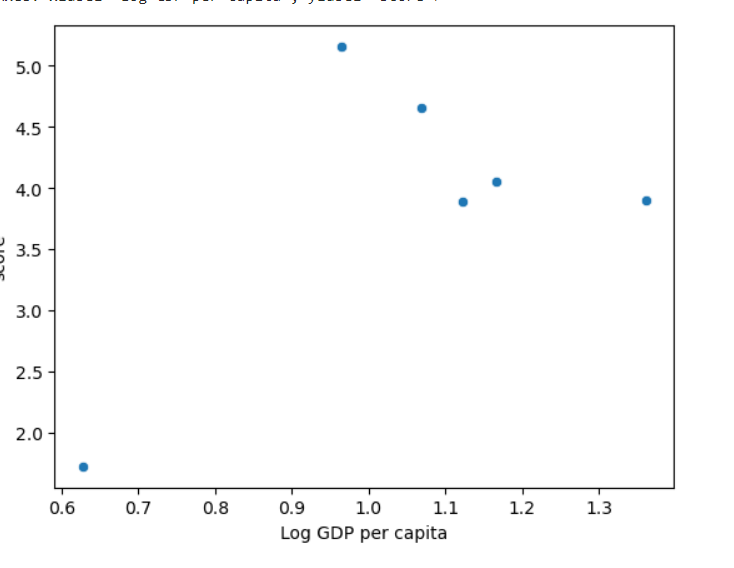
While calculating range for Middle east and South Asia, it was found to be 4.634

And 3.4370000000000003 respectfully. The covariance for both regions were found to be 30.214828833374263 for South Asia and 30.808328395054225 middle east.

**Outlier Detection**

****

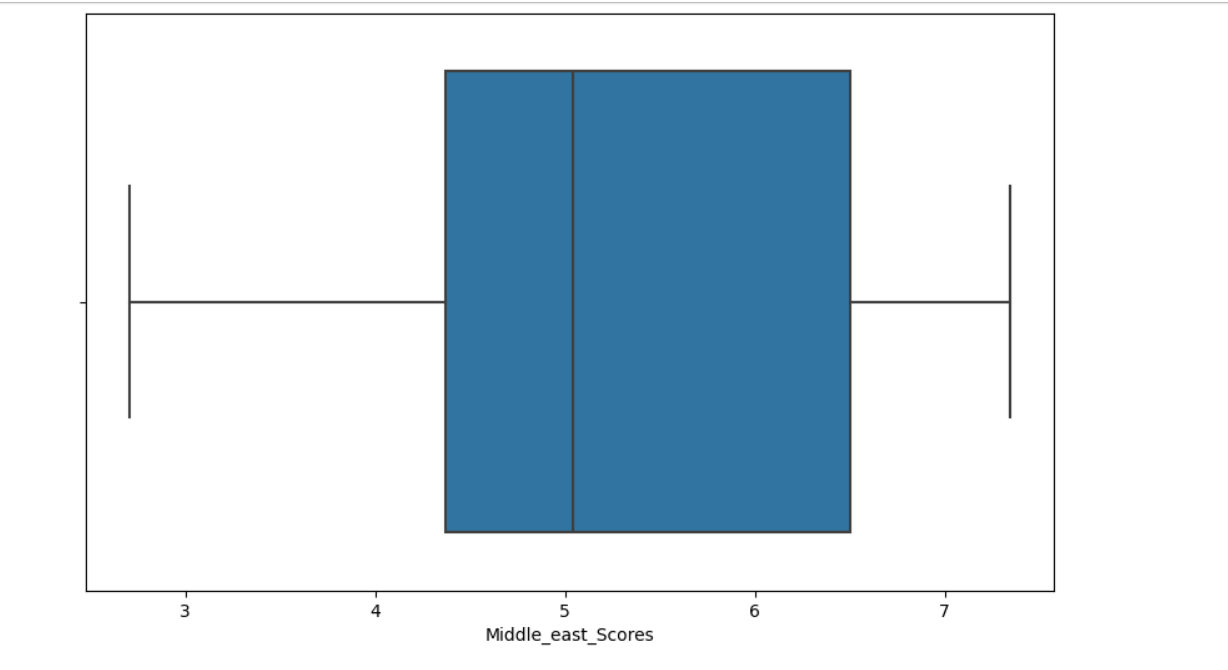
While outliers are the data which behave abnormally with respect to the other data, Here according to me is only one outlier that being the country Afghanistan, the other may be above of 1.4 , but it’s not because it can just be the data which didn’t perform well on the score .



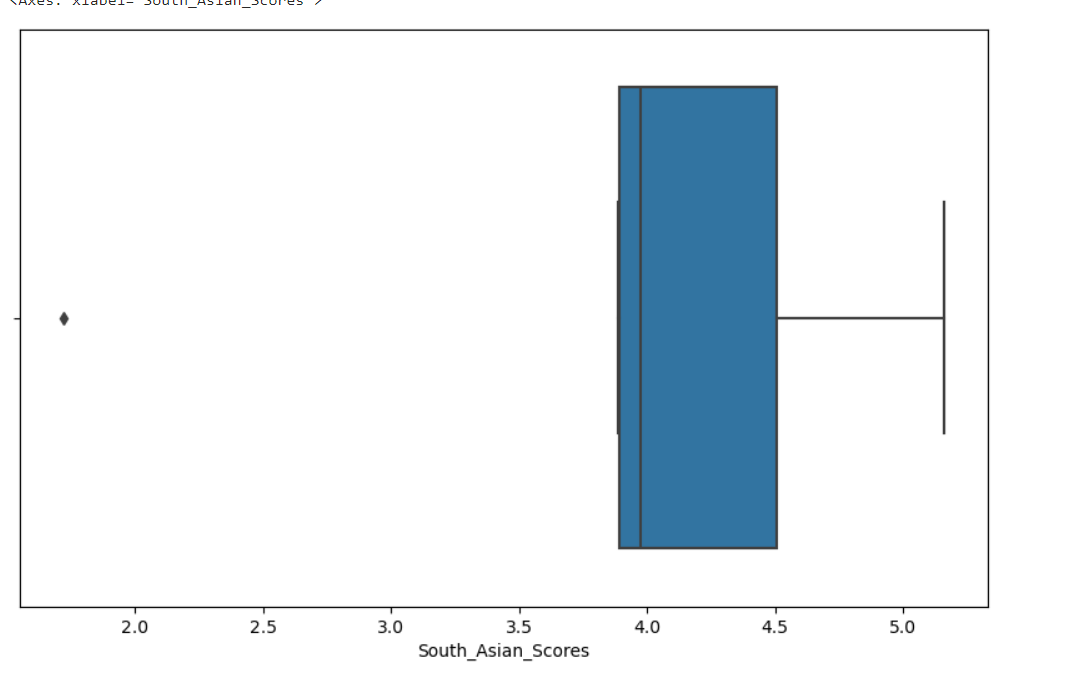
While in this scatterplot, the outliers are the one on 0.65.

The effect of the outlier can be dangerous for our machine learning model because it doesn’t know the difference between the exception and generalization. So due to it, our machine learning model’s accuracy could drop down drastically.

**Box Plots**



The above boxplots is of Middle East countries with respect to their scores, the above boxplot shows that the minimum value is about 2.5 and the maximum value is 7.8.  
  
The frequency (No. of times the number was repeated) was shown between 4.3 to 6.7.



The above boxplots are of South Asian countries with respect to their scores, the above boxplot shows that the maximum value is 5.1.  
  
The frequency (No. of times the number was repeated) was shown between 3.7 to 4.5.