UNIVERSITY OF WOLVERHAMPTON
KNOWLEDGE ▪ INNOVATION ▪ ENTERPRISE

HERALD COLLEGE KATHMANDU | ing GROUP

5CS037 - Concepts of AI

Statistical Interpretation and Exploratory Data Analysis:

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

Name: Rohan Sitoula
Student Id: 240887
Module Leader: Siman Giri

Introduction:

The World Happiness Report is an annual report that ranks countries based on various factors influencing happiness: GDP, Freedom to make life choices, Social Support, Healthy life expectancy, and much more. By factoring all those things, it aims to create a Happiness score for each country. In total, there are 143 countries participating in this program.

About the analysis:

This report tries to analyze the dataset of **WHR** by proposing some data exploration and visualization techniques referring to ***South Asia*** and the ***Middle East***. We try to have an understanding, by looking through these two regions, over their ranking in happiness as well as disparities and its trends across various metrics by analyzing the dataset.

This report is divided into three main sections:

1. Data Exploration and Understanding
2. Advanced Data Exploration: *South Asia* Analysis
3. Comparative Analysis: *South Asia* vs. *Middle Eas*t

Each section is supported by detailed analysis, visualizations, and insights.

Tools and Techniques:

- Python as a programming language.
- Pandas for data Exploration.
- Seaborn and Matplotlib for different visualization methods.
- Google Collab as Code editor.

Problem 1: Data Exploration and Summary

Task 1: Basic Data Exploration

1. Load the dataset and display it.
2. Dataset Overview:
    - Number of Rows: 143
    - Number of Columns: 9
3. Data Types:
    - All columns are of type float, except Country Name, which is of type object.

Task 2: Descriptive Statistics

1. Happiness Score Analysis:
    1.1. Mean: 5.52
    1.2. Median: 5.78
    1.3. Standard Deviation: 1.17
2. Minimum and Maximum Happiness Score Country:
    2.1. Maximum: *Finland*
    2.2. Minimum: *Afghanistan*
3. Missing value
    3.1. There are 3 missing values in each column except the country name and score.

Task 3: Filtering and Sorting:

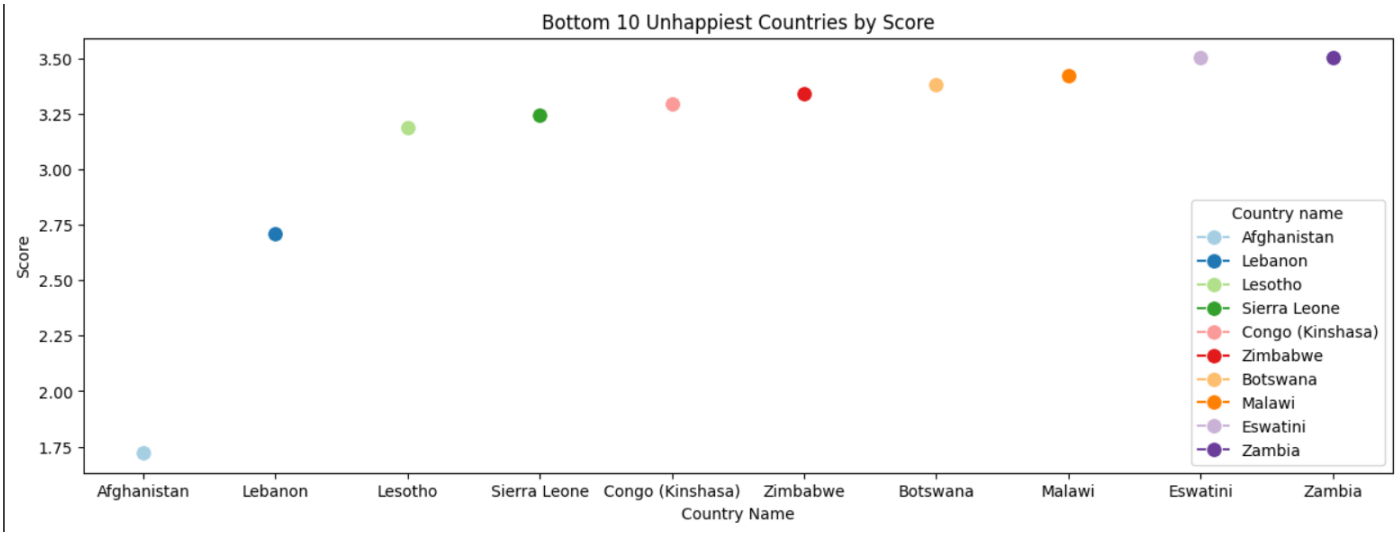
1. Countries with above 7.5 score are:
    a. *Finland*
    b. *Denmark*
    c. *Iceland*

Task 4: Creating new Column:

Here , we create a column name Happiness category to rate different happiness score range into 3 divisible parts High ,Medium and Low

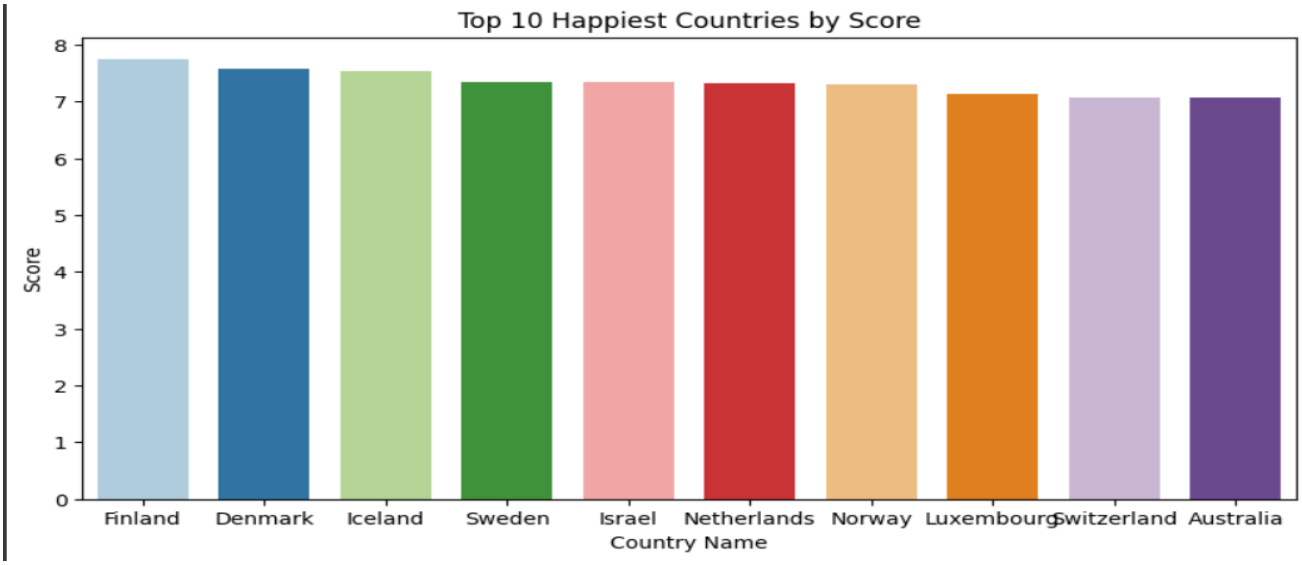
Task 5: Data Visualization:

1. Line Plot: Top 10 unhappiest Country:



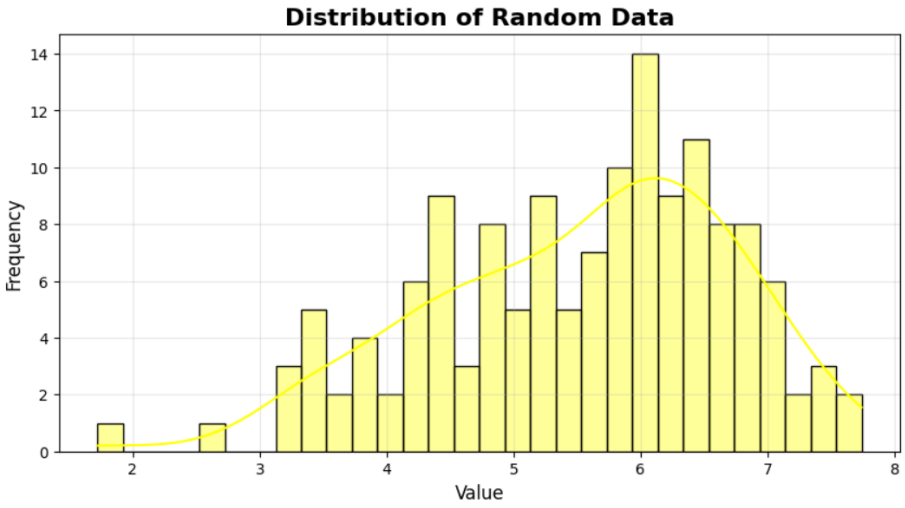
Top 10 unhappiest country across the globe according to WHR making *Afghanistan* the unhappiest country in the world.

2. Bar plot: Top 10 Happiest Country:



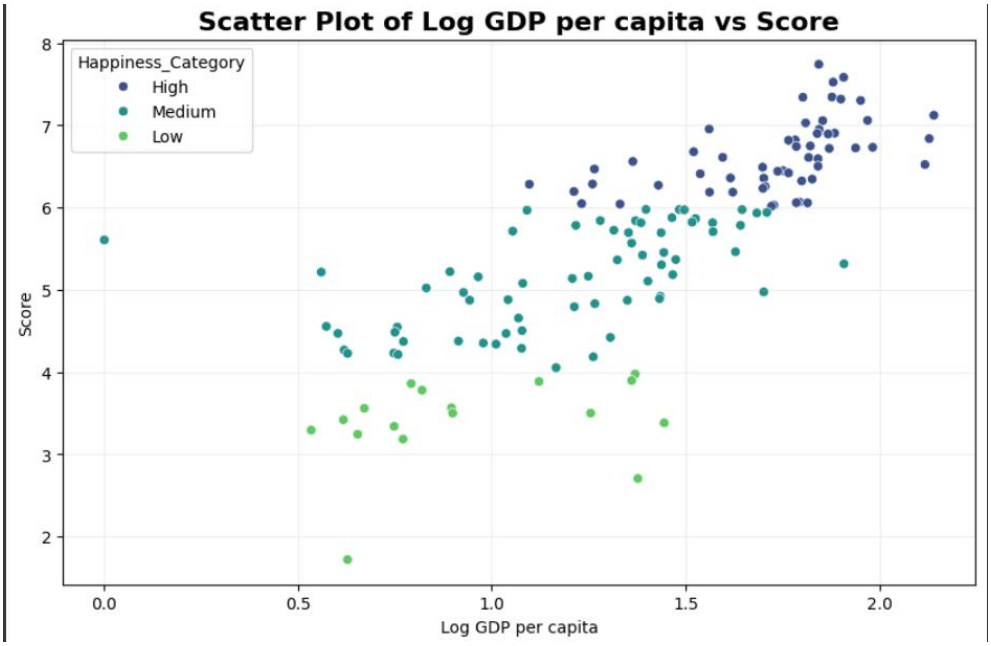
Top 10 Happiest countries in the world Finland Leading with above 7.5 , followed by *Denmark ,Iceland* and more.

3. Histogram:



Shows that most of the countries happiness lies between 5 to 7

4. Scatter Plot:

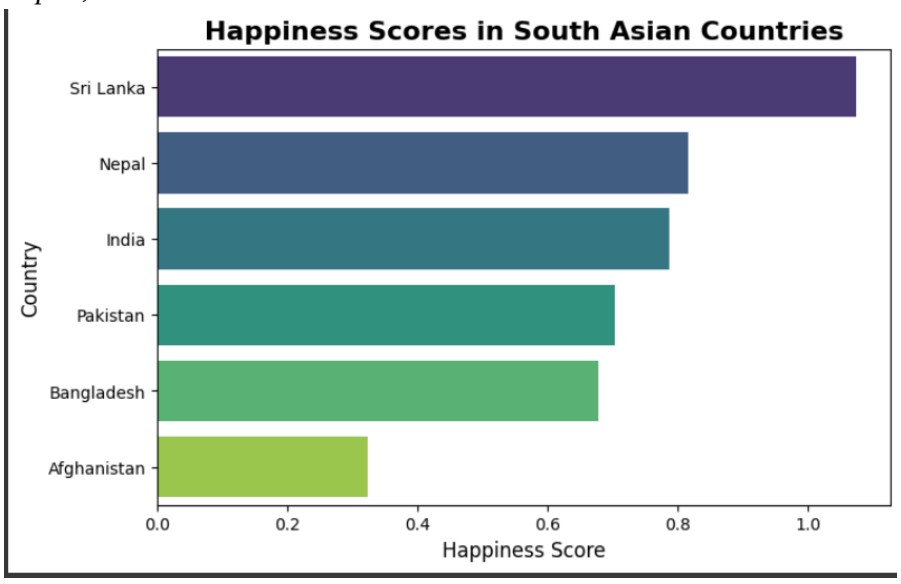


The Countries who have low GDP also have low Happiness scores. It suggests that GDP has a significant role in shaping the overall Happiness Rate of the country.

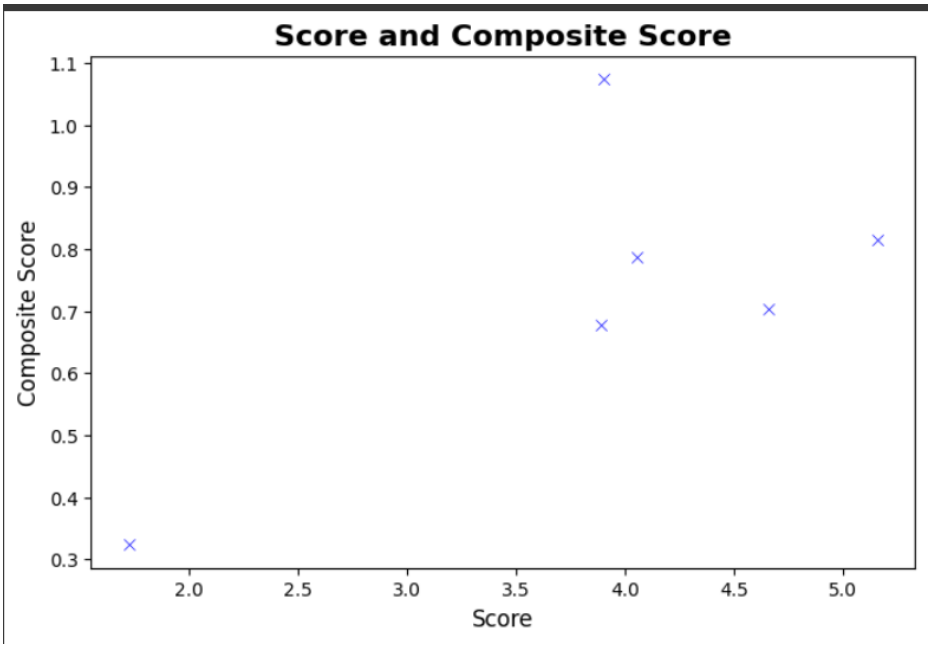
Problem 2: Data exploration.

Here, the major focus is kept on the data splitted from the real data frame which is of *South Asia*.

1. We have created a composite score combining GDP , Social Support and healthy life expectancy.
    a. Composite Score = 0.40 × GDP per Capita + 0.30 × Social Support + 0.30 × Healthy Life Expectancy
    b. Key findings of composite score:
        i. Out of the presented, *Sri Lanka* has the Highest composite score followed by *Nepal* , *India* and so on.



c. We have also compared the Composite score with the original score to see if the composite score matches with the original alignment or not.



Well, the composite score does somewhat match with the original score but not quite, because when the score is 5, the composite score is 0.8, while when the score is 3.8, the composite score reaches 1.1.

2. Outlier Detection

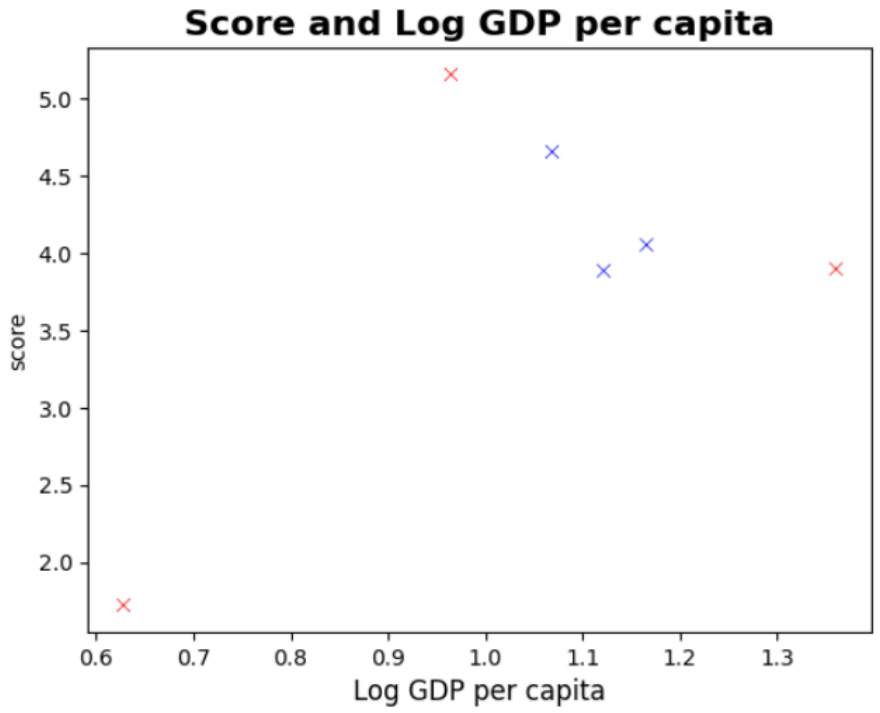
Here we have to find an outlier based on GDP and their score. In order to find the outlier we are using InterQuartile Range (IQR).

a. Based on out findings with score:
   i. Upper Limit is 5.05 , any score above 5.05 is considered as an outlier.
   ii. Lower Limit is 3.19 , any score below 3.19 is considered as an outlier.

   Based on Score *Nepal* has Happiness Score above 5.05 i.e 5.158 so Nepal is an Outlier. At the same time *Afghanistan* has a happiness Score below 3.19 i.e 1.172 so, *Afghanistan* is considered as an Outlier.

b. Based on out findings with GDP:
   i. Upper Limit is 1.3345 , any score above 1.33 is considered as an outlier.
   ii. Lower Limit is 0.82 , any score below 0.82 is considered as an outlier.

Based on Score *Sri Lanka* has Happiness Score above 1.361 i.e 1.361 so *Sri Lanka* is an Outlier. At the same time *Afghanistan* has a happiness Score below 0.628 i.e 0.628 so, *Afghanistan* is considered as an Outlier.



Plot marked with the red color is an outlier. As you can see *Nepal* at the top being outlier in Score and *Afghanistan* in the bottom as an outlier. *Sri Lanka* being an outlier at the last of x axis in case of GDP . *Sri Lanka* and *Nepal* are affecting the GDP and Happiness score respectively but *Afghanistan* is greatly affecting both GDP and Happiness.

a. The total mean including Outlier of GDP is 1.05.
b. The total mean removing the Outlier {*Afghanistan* and *Sri Lanka*} is 1.0805.
c. Total mean removing only the lower limit outlier is 1.1366.
d. Total mean removing only the upper limit outlier is 0.99.

It all boils down to the following conclusion:

Afghanistan, *Sri Lanka*, and *Nepal* are outliers that greatly affect the analysis of Happiness Score and GDP.
Afghanistan is an important outlier, with very low levels of Happiness and GDP, pulling the average down. It has underlined severe problems needing urgent attention.

Sri Lanka's high Happiness Score and high GDP skew the mean and might represent some unique progress or irregularities in data worth investigating.
This suggests influences are localized, since *Nepal* impacts the Happiness Score harder than it does GDP.

Removing outliers shows:

Excluding *Afghanistan*, the mean **GDP** improves to **1.1366**.
Without *Sri Lanka*, the **mean drops** to **0.99**, reflecting that *Sri Lanka* pulled the mean upward.

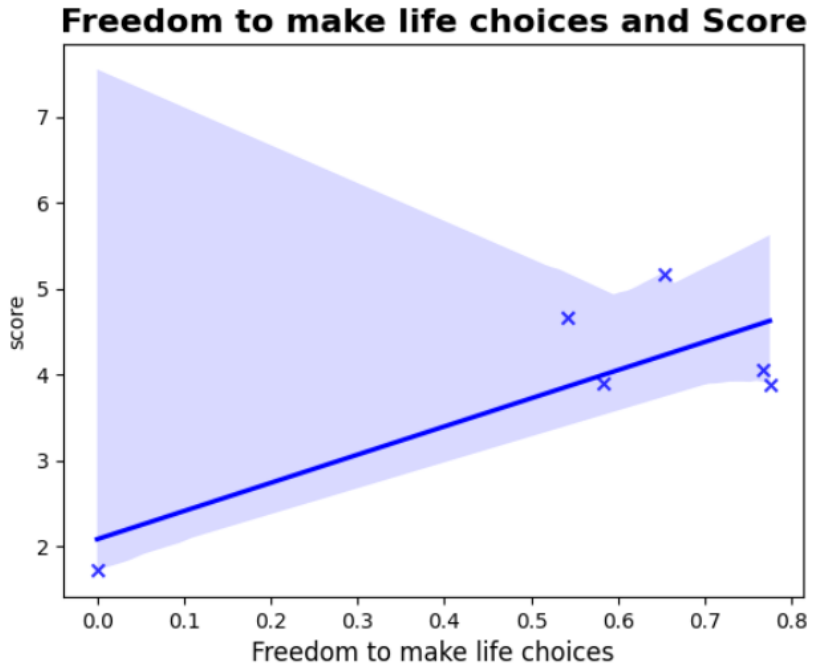
These outliers reveal shocking regional inequities. Scrutiny should be carried out to uplift *Afghanistan* while studying *Sri Lanka* and *Nepal* for best practices or anomalies.

3. Correlation Analysis:

We are going to compare different Features with score features.

a. Freedom to make life choices and score:

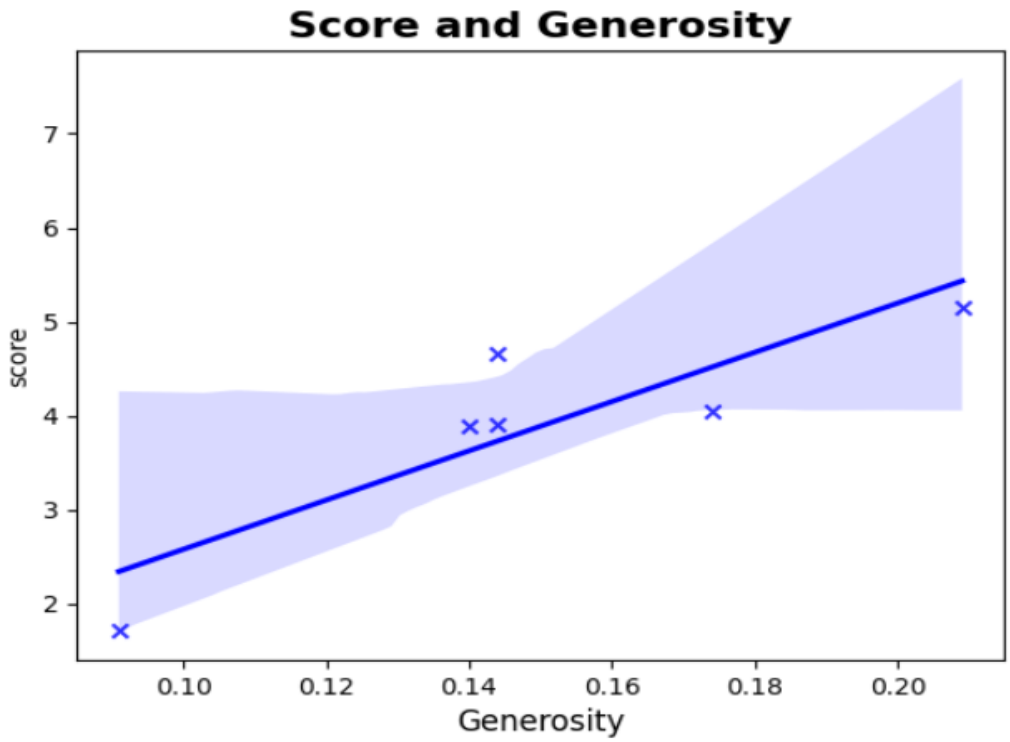
i. Freedom to make life choices and score have positive but not too strong relationships with each other.



As we can see , increment in freedom does increase score but not actively.

b. Generosity and Score

i. Generosity and Happiness Score should have a positive impact , because the more generous the people are, the more happy they will be. But let's see what statistics says:



Statistics do show the potential impact generosity has with a score but it's also not very strong , because there is some scatter in the middle where the score is getting lower even if the generosity is high. It's because generosity is not the only thing that helps to calculate scores.

4. Gap Analysis:

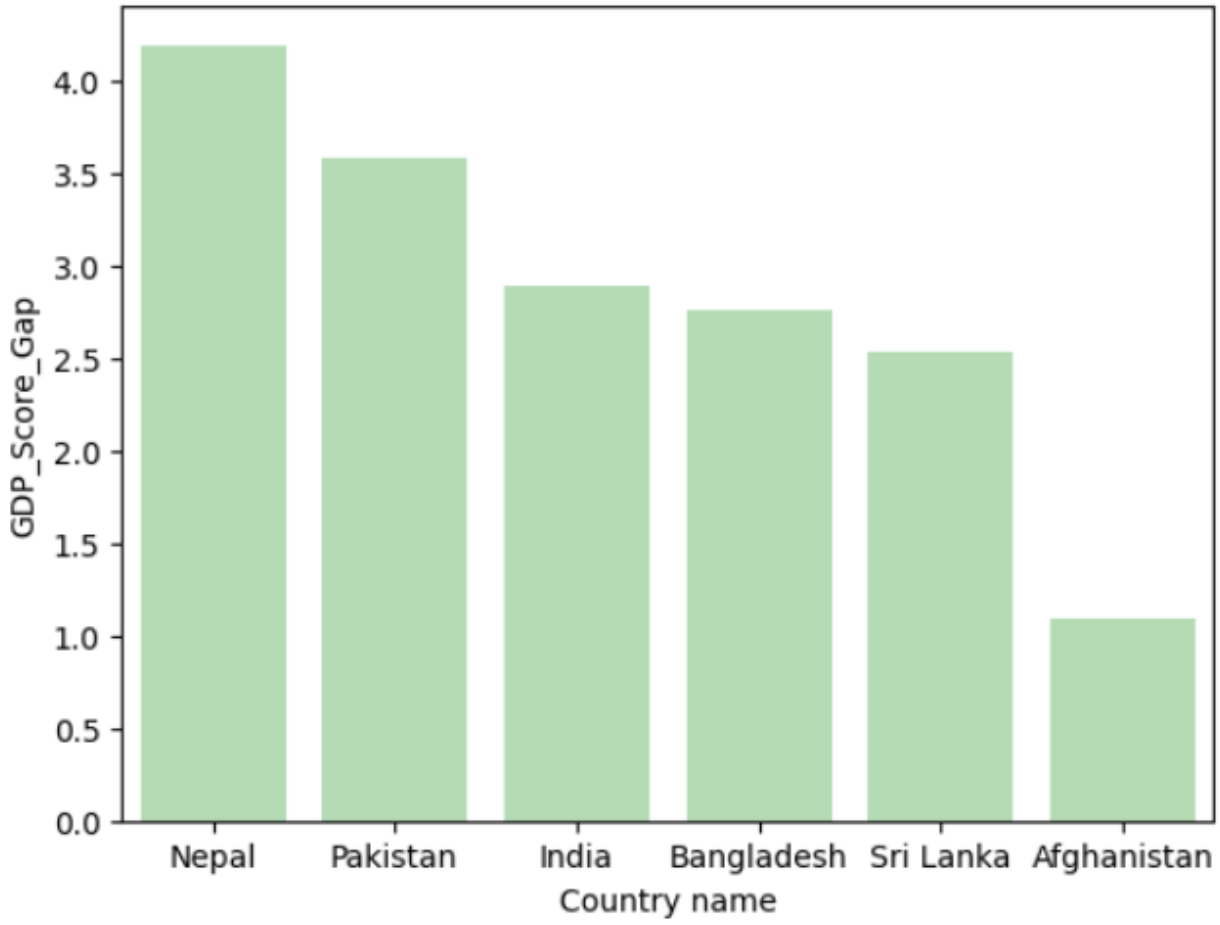
Here, we exclude the influence of GDP in the Happiness Score. This is because we are finding how happy people are in the country without directly considering the Economic outcome. The insights show that, for South Asian countries:

a. *Sri Lanka* is in the second last position in the gap analysis, having the highest composite score of 1.0739, which shows that happiness in the country corresponds well with its economic and social indicators. The happiness score of *Sri Lanka* is indicative of a balance among its various contributors like GDP, social support, healthy life expectancy, and freedom to make life choices.

b. *Nepal* is the opposite of *Sri Lanka*. This means that even though the GDP is low in *Nepal*, its Happiness Score is comparatively high. This suggests that happiness in *Nepal* does not depend on GDP but rather on other characteristics. For example, *Nepal* has a high degree of social support (0.990) and freedom to make life choices (0.653), which are significant determinants of well-being. This indicates that people in *Nepal* may feel a strong sense of community and autonomy in life, contributing greatly to their happiness.

c. *Afghanistan* is the lowest because everything in *Afghanistan* is low, and excluding GDP does not have much effect on its overall Happiness Score.

Here is a bar plot for overview of my analogy

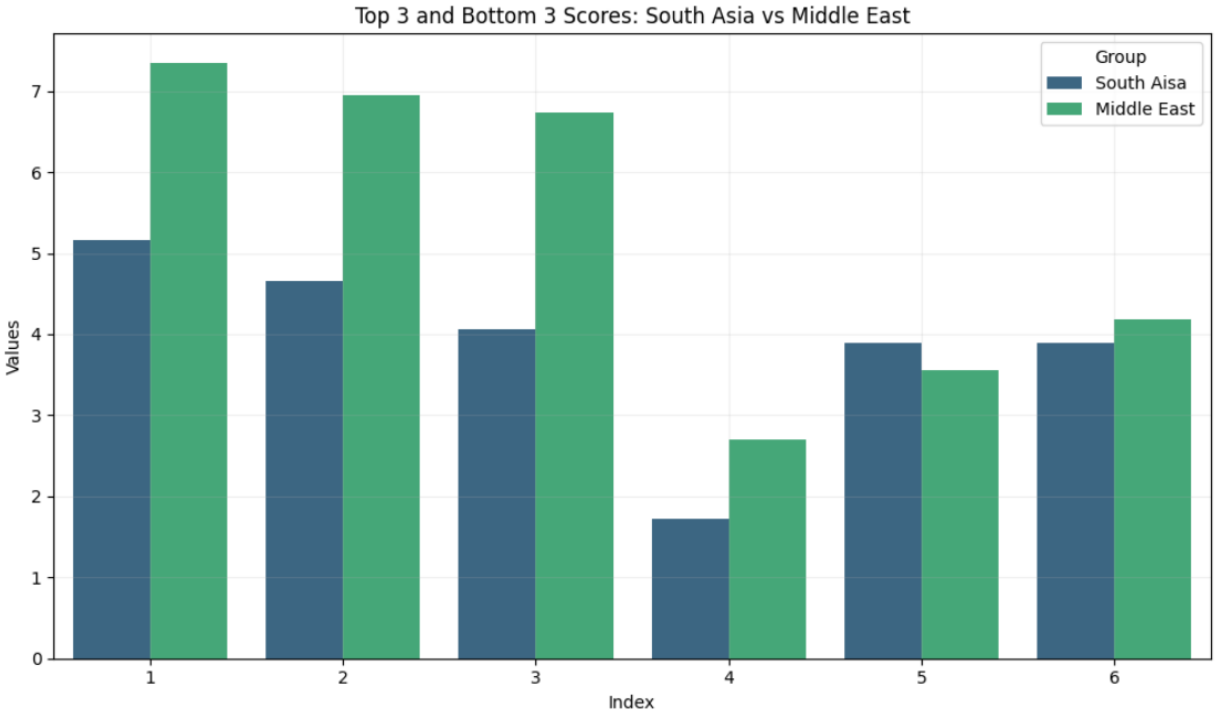


---

3. Comparative Analysis:

1. Here , we create one more dataframe or we can say a separate region which includes *Middle East* countries.(*two of them where null so we removed that countries*).
   Here are some info about the dataframe:
   1) Mean: 5.35
   2) Standard Deviation: 1.65
   3) Correlation between Freedom and Score : 0.86
   4) Correlation between Generosity and Score : 0.63

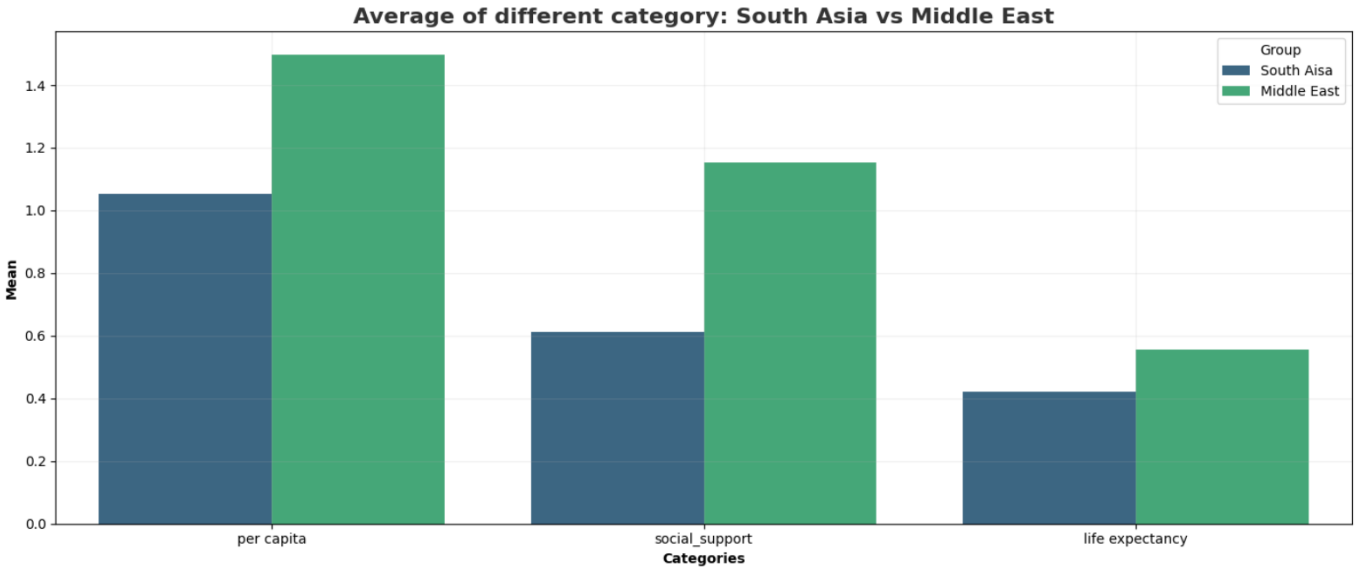
Some interesting comparative analysis of these data:



a. The *Middle East* region has a higher Happiness rate than *South Asian*.
b. The *Middle East* has 4 high happiness rate category countries and *South Asia* has none.
c. Highest happiness score of the *Middle East* is 7.341 which is of *Israel* whereas 5.158 goes to the top for *South Asia* which is of Nepal.
d. *South Asia* has more stability with their Countries happiness rate because their Standard deviation is 1.18 whereas 1.65 for *Middle East* countries.

2. Let's have come Metric Comparison between these two regions

(We are comparing per capita , social support , life expectancy of each region with each other.)



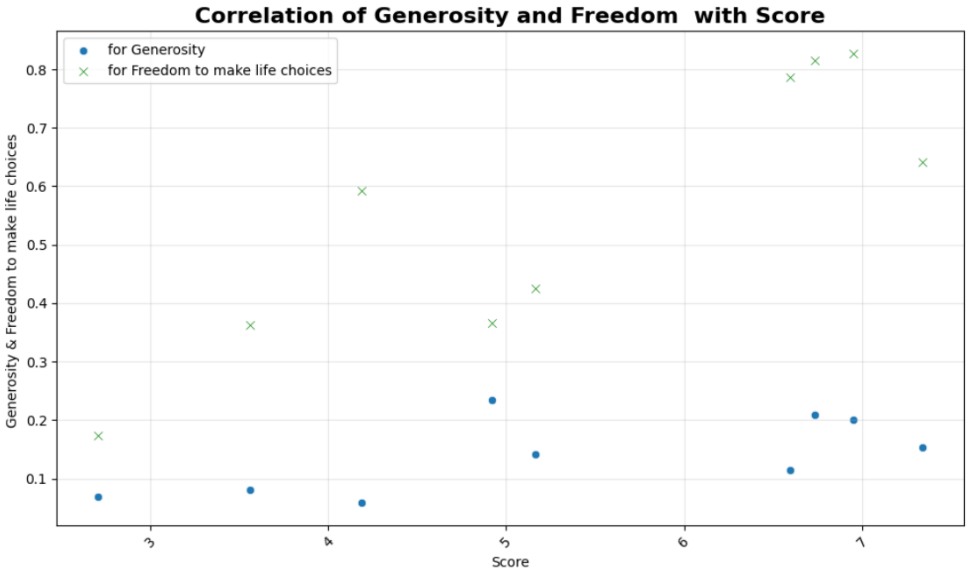
As seen in the bar plot , the *Middle East* is comparatively greater in every field , whether in per capita , social support or life expectancy.

3. Let's talk about Variations between these two regions:

a. **Coefficient of variation** of *South Asia* Happiness Score is **22.00**
b. **Coefficient of variation** of *Middle East* Happiness Score is **30.81**

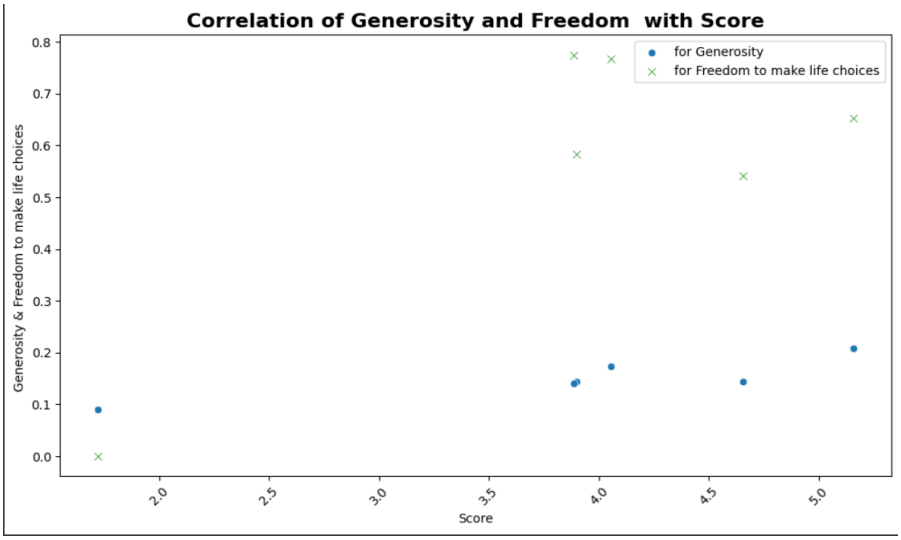
This clearly proves that the *Middle East* has greater variability which means the *Middle East* has a more scattered happiness rate than *South Asia.*

4. Correlation Metics:



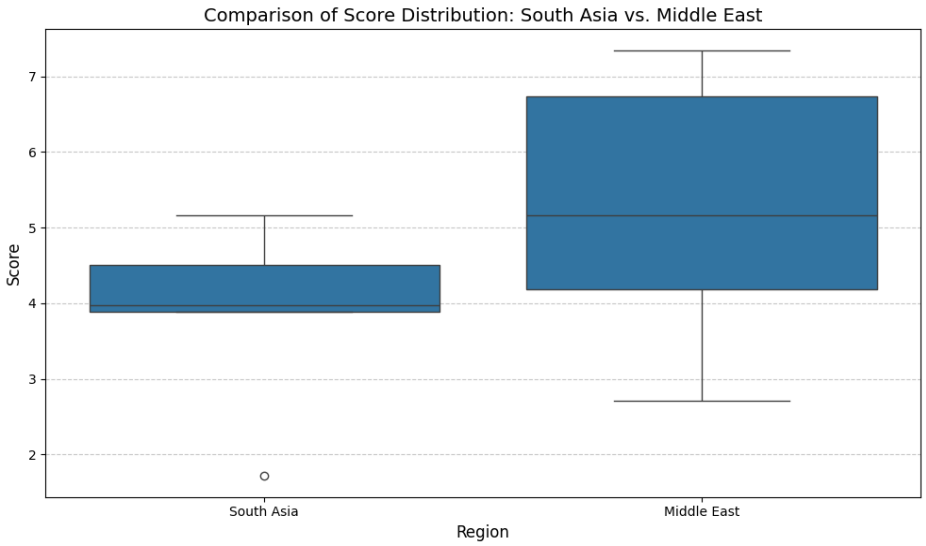
Here, the correlation metrics are remarkably different from *South Asian* correlations. Here, generosity has no role to play, or it can be said that it has less influence on Score, whereas Freedom plays an important role in the Increment of score. Higher the Freedom, the more happy the people will be.

Let's have a look on *South Asian* correlation:



Here we can see that data are not more scattered but more connected to each other , freedom and generosity seems to have a positive correlation but not very strong.

Let's have a final comparison with a boxplot:



South Asia is more compact with a smaller IQR, showing scores clustered around the median. In contrast, *the Middle East* with a larger IQR spans from 2.7 to over 7.2, indicating more variability. One outlier appears near 1.7 in *South Aisa*, while the *Middle East* scores show no outliers and mostly stay within the expected range despite their broader spread.