Concepts and Technologies of AI

5CS037

Report

Statistical Interpretation and Exploratory Data Analysis

Name: Sarthak Acharya

Group: L5CG6

Student ID: 2406785

Contents

Introduction

The World Happiness Report is a survey assessing global happiness based on various factors such as GDP per capita, social support, life expectancy, freedom, and generosity. This research aims to focus factors influencing happiness levels across nations.

This report explores the datasets in three parts: general data exploration, analysis of South Asia, and a comparative study between happiness level of South Asia and Middle East. Each section deals with specific tasks which uses simple statistics and easy to understand charts or visuals to identify patterns and data analysis.

Problem 1: Data Exploration and Understanding

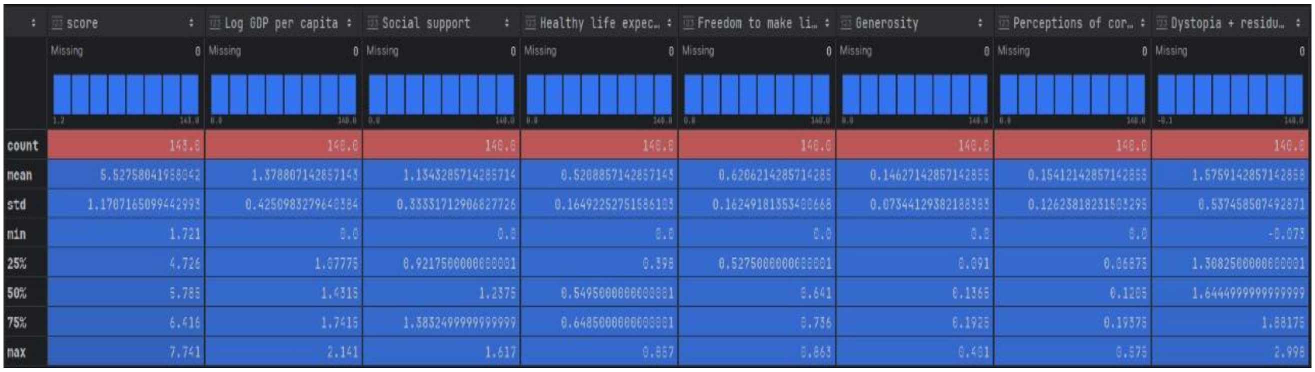
Overview

Initially, the exploration focused on the understanding of the data's structure, summarizing key metrices and visualizing happiness trends globally.

Tasks and Explanation

1. Dataset Overview

- Firstly, with importing necessary libraries and accessing the dataset csv file the overview of the data is shown using the data.head().
- After the data is checked, the rows and columns are calculated to know the size of data at the beginning.
- Then the most important part the description or main overview of data. The key statistics of the data is summarized using the data.describe() method as shown below:

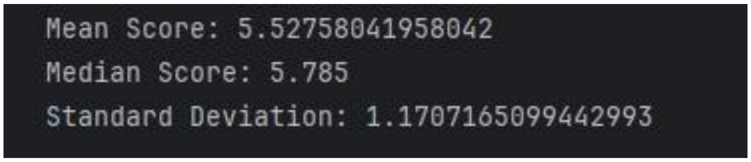


	score	Log GDP per capita	Social support	Healthy life expec…	Freedom to make li…	Generosity	Perceptions of cor…	Dystopia + residu…
Missing	0	0	0	0	0	0	0	0
count	143.0	140.0	140.0	140.0	140.0	140.0	140.0	140.0
mean	5.527580419580042	1.378807142857143	1.1343285714285714	0.5208857142857143	0.6206214285714285	0.14627142857142855	0.15412142857142855	1.5759142857142858
std	1.1707165099442993	0.4250983279640384	0.33331712906827726	0.16492252751586103	0.16249181353400668	0.07344129382188383	0.12623818231503295	0.537458507492871
min	1.721	0.0	0.0	0.0	0.0	0.0	0.0	-0.073
25%	4.726	1.07775	0.9217500000000001	0.398	0.5275000000000001	0.091	0.06875	1.3082500000000001
50%	5.785	1.4315	1.2375	0.5495000000000001	0.641	0.1365	0.1205	1.6444999999999999
75%	6.416	1.7415	1.3832499999999999	0.6485000000000001	0.736	0.1925	0.19375	1.88175
max	7.741	2.141	1.617	0.857	0.863	0.401	0.575	2.998

This dataset provides summary statistics for the factors influencing happiness levels across different countries, such as GDP, Social Support, Health, and freedom. These metrics help understand how different factors contributes to overall happiness levels globally.

2. Key Metrics

- After knowing about the data using the describe method we calculated the Mean, Median and Standard Deviation using their methods respectively which generated the output as below:

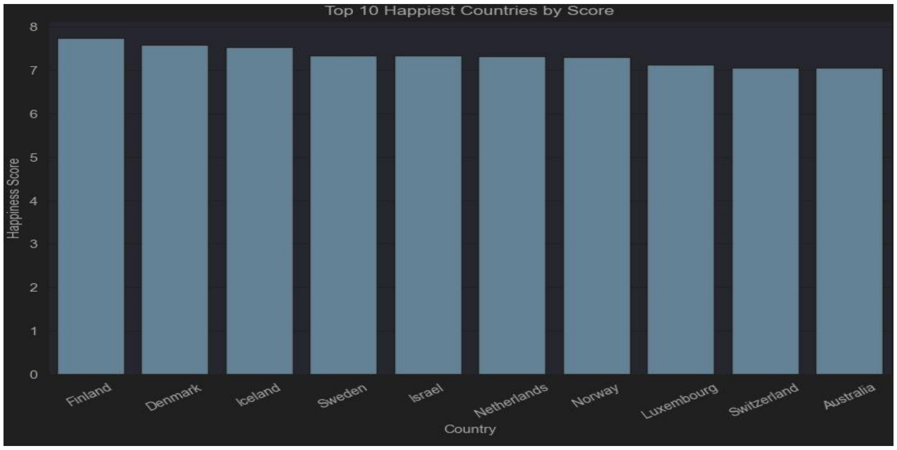


```
Mean Score: 5.52758041958042
Median Score: 5.785
Standard Deviation: 1.1707165099442993
```

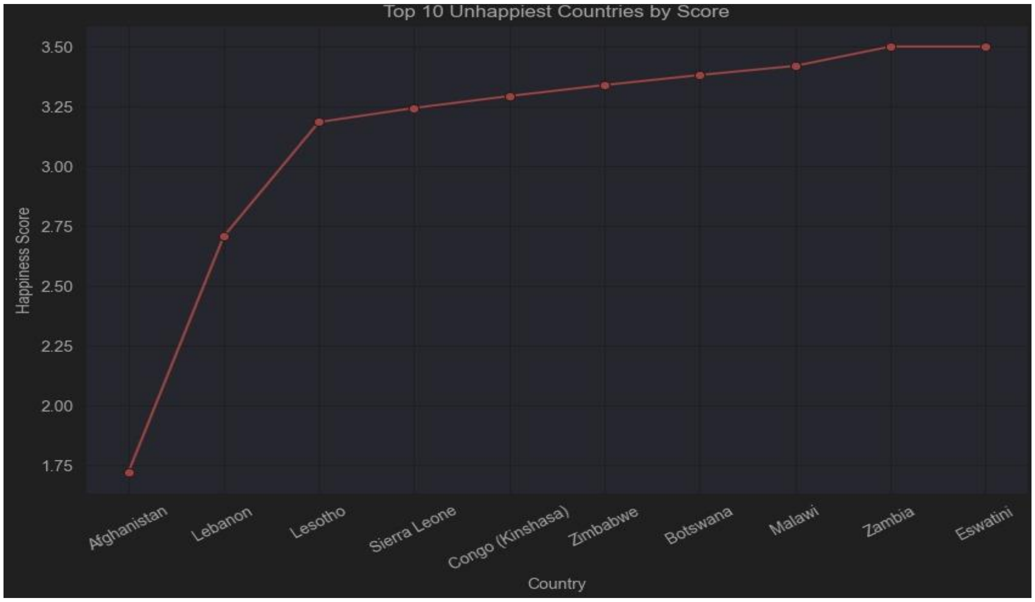
- This helps in the measurement of central tendency which basically provides a summary of the data using a single value that represents the entire distribution.
- The score-based analysis calculated showed that the happiest country on the dataset is Finland and the least-happiest is Afghanistan.
- Here then data is checked for missing values and data is filtered for countries whose score is more than 7.5.
- As per the requirement, the data is then sorted in descending order.
- At-last the data’s happiness score is categorized as low, medium and high.

3. Visualizations

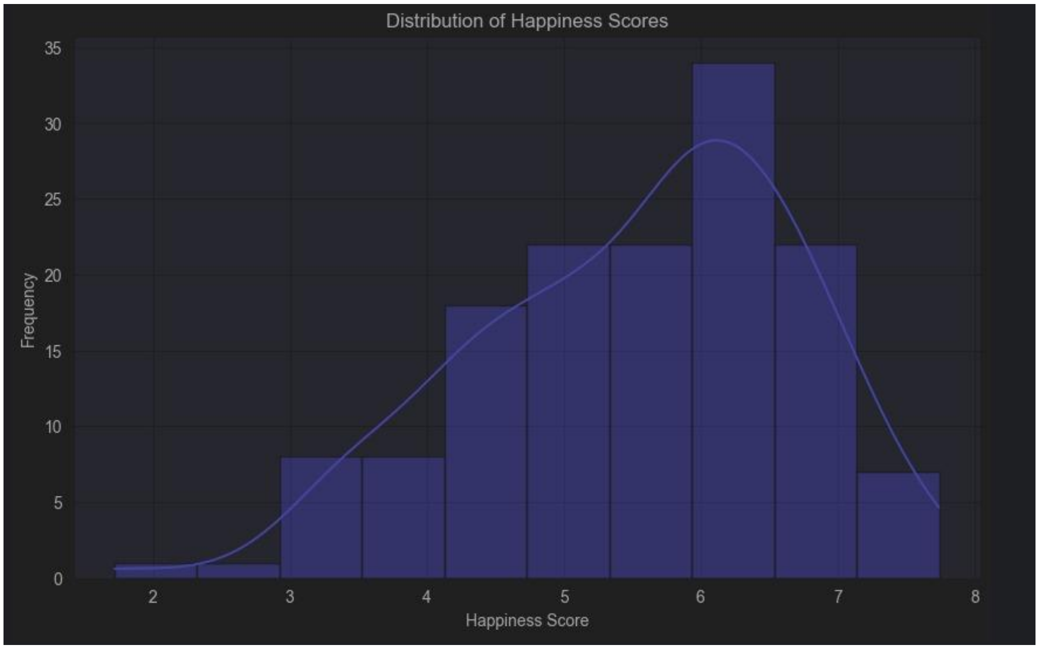
- The bar chart of the top 10 happiest countries is analyzed then.



- This bar graph shows the happiness level for the top 10 countries listed on the dataset.
- For the 10 unhappiest countries, line graph is used to analyze the data as shown below:

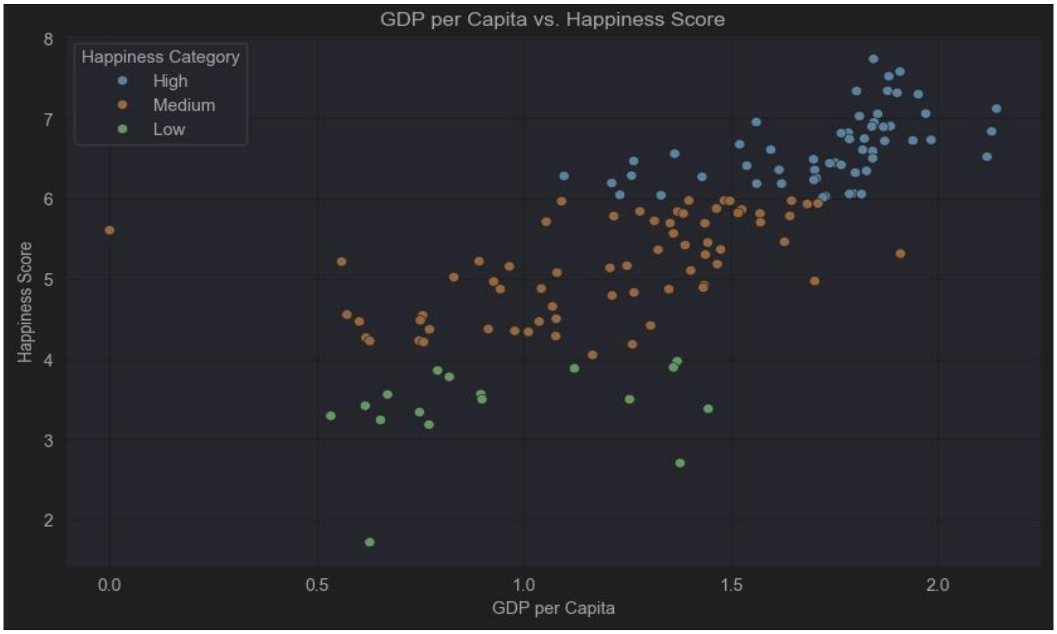


- The histogram is plotted which illustrates the distribution of happiness scores with additionally KDE added which is used to understand the underlying distribution of data points as shown below:

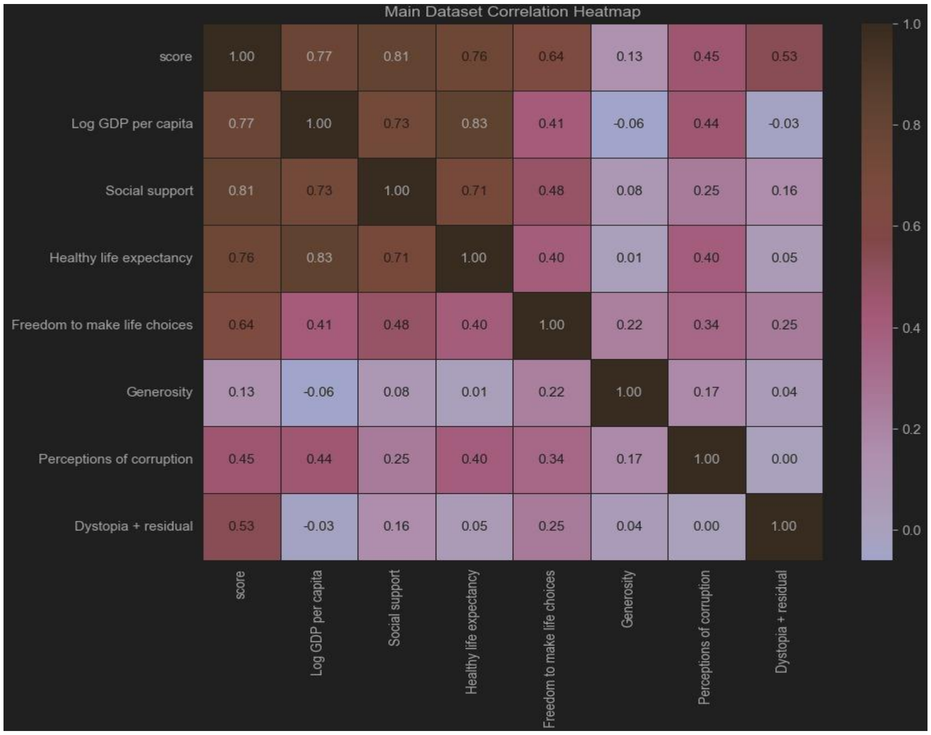


- Scatter plot:

It is a graph in which values of two variables are plotted along two axes, the patterns of the resulting points reveal any correlation present.



In this scatter plot, Happiness Score is shown with respect to the GDP per capita. We can see a trend in this plot, higher the GDP per capita score results in higher Happiness score.



	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
score	1.00	0.77	0.81	0.76	0.64	0.13	0.45	0.53
Log GDP per capita	0.77	1.00	0.73	0.83	0.41	-0.06	0.44	-0.03
Social support	0.81	0.73	1.00	0.71	0.48	0.08	0.25	0.16
Healthy life expectancy	0.76	0.83	0.71	1.00	0.40	0.01	0.40	0.05
Freedom to make life choices	0.64	0.41	0.48	0.40	1.00	0.22	0.34	0.25
Generosity	0.13	-0.06	0.08	0.01	0.22	1.00	0.17	0.04
Perceptions of corruption	0.45	0.44	0.25	0.40	0.34	0.17	1.00	0.00
Dystopia + residual	0.53	-0.03	0.16	0.05	0.25	0.04	0.00	1.00

The heatmap shows that "Log GDP per capita" and "Healthy life expectancy" have a strong positive relationship, highlighting their close association. On the other hand, "Generosity" exhibits weak or negative correlations, indicating its independence from other factors. Moderate relationships, such as between "Freedom to make life choices" and "Social support," suggest a balanced level of dependency. Overall, the visualization captures how some variables are highly connected while others stand apart.

4. Key Understanding

- Strong positive correlation was observed between GDP per capita and Happiness Score.
- The histogram showed a normal distribution of scores with slight peak.

Problem 2: South Asia Dataset and Advanced Tasks

Overview

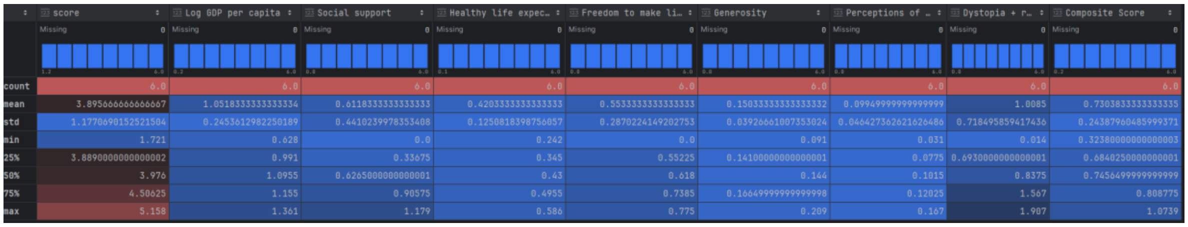
This section focuses on Suth Asian countries, calculating a composite score, detecting outliers, and analyzing the correlations.

Tasks and Explanation

1. Filtering South Asia

Firstly, a list of South Asian countries was extracted from the main dataset csv

file and saved the filtered dataset to a new csv file. The data is then tried to understand using the describe method.

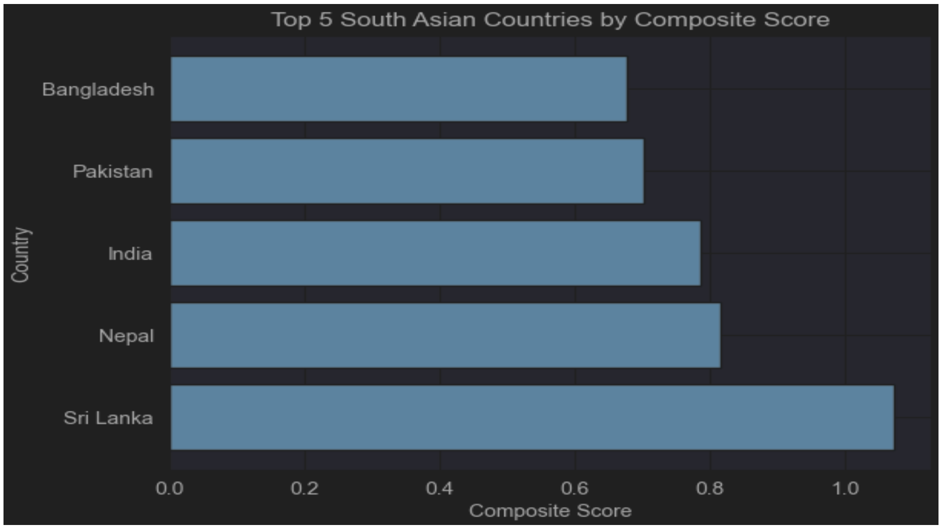


	score	Log GDP per capita	Social support	Healthy life expec…	Freedom to make li…	Generosity	Perceptions of …	Dystopia + r…	Composite Score
	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0
count	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0
mean	3.8956666666666667	1.0518333333333334	0.6118333333333333	0.4203333333333333	0.5533333333333333	0.15033333333333332	0.09949999999999999	1.0085	0.7303833333333335
std	1.1770690152521504	0.2453612982250189	0.4410239978353408	0.1250818398756057	0.2870224149202753	0.03926661007353024	0.046427362621626486	0.718495859417436	0.24387960485999371
min	1.721	0.628	0.0	0.242	0.0	0.091	0.031	0.014	0.32380000000000003
25%	3.8890000000000002	0.991	0.33675	0.345	0.55225	0.14100000000000001	0.0775	0.6930000000000001	0.6840250000000001
50%	3.976	1.0955	0.6265000000000001	0.43	0.618	0.144	0.1015	0.8375	0.7456499999999999
75%	4.50625	1.155	0.90575	0.4955	0.7385	0.16649999999999998	0.12025	1.567	0.808775
max	5.158	1.361	1.179	0.586	0.775	0.209	0.167	1.907	1.0739

2. Composite Score

Composite score was then created combining GDP per capita (40%), Social Support (30%), and healthy life expectancy (30%).

Then the top 5 countries based on their composite scored were viewed on a bar graph as shown below:

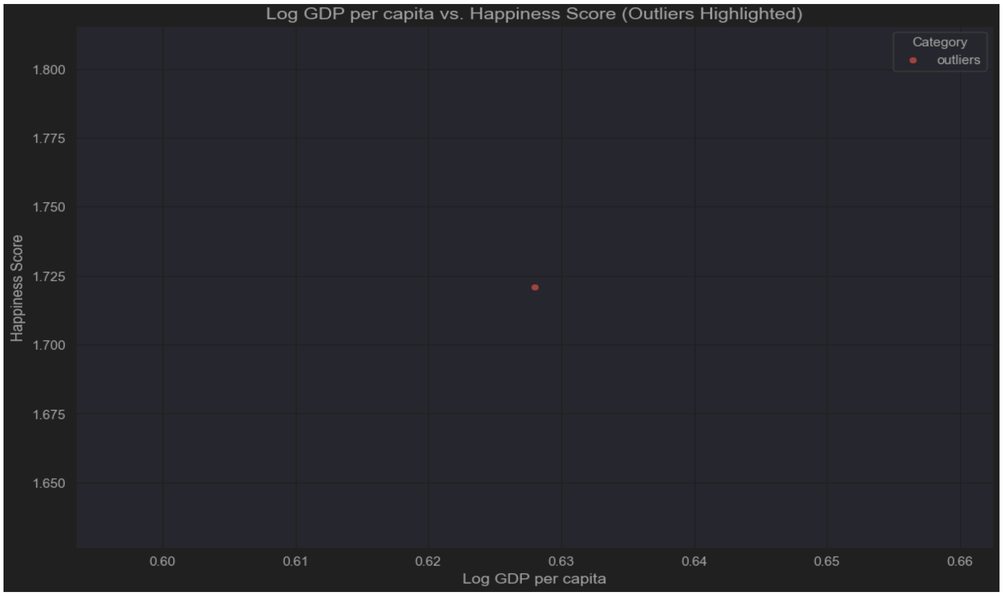


Then a simple comparison between the score and the composite score plotted in a bar graph.

3. Outlier Detection

Here, an important method IQR (Interquartile Range) is used to identify outliers in the Happiness scores. The IQR method is based on the spread of the middle 50% of the data.

Then after identifying the outliers, it is plotted in a scatter plot which compares the GDP per capita and Happiness Score.



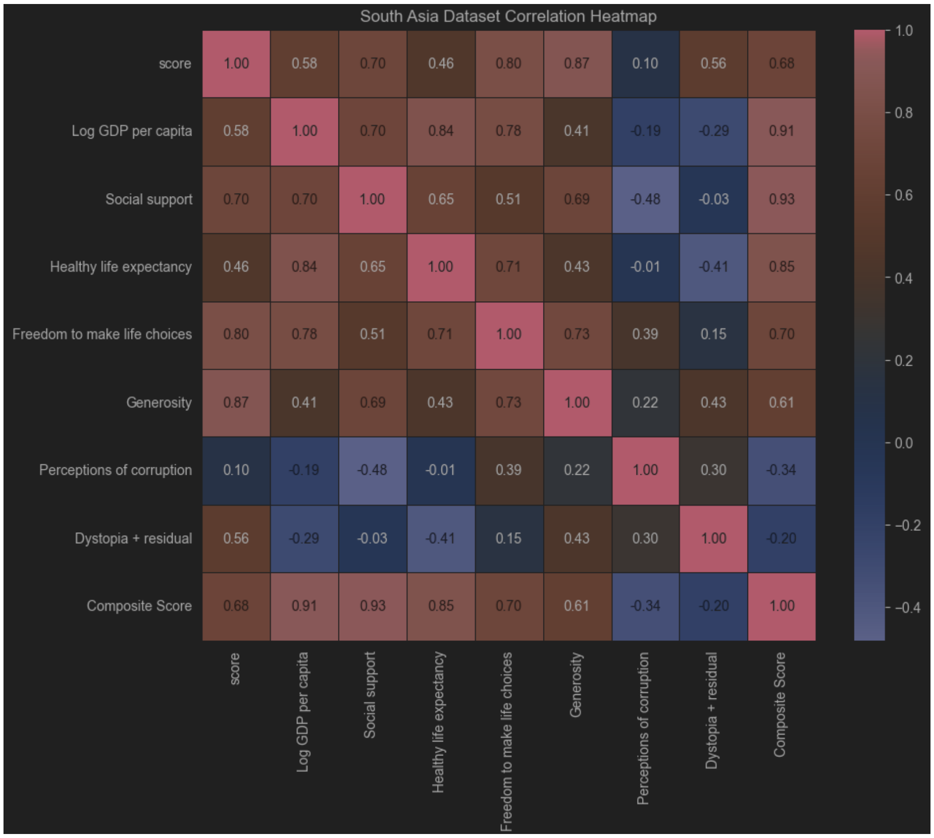
The outlier was basically calculated using the formula:

IQR=Q3−Q1

Where, Q3 = upper bound & Q1 = lower bound

4. Correlation Analysis

At the last a heatmap for South Asia is calculated which shows significant correlations between GDP per capita and other metrics.



The heatmap of the South Asia dataset highlights key relationships among variables. Strong positive correlations are seen between the composite score and factors like social support and GDP per capita, suggesting their significant influence on overall well-being. Conversely, perceptions of corruption show weak or negative correlations with most metrics, indicating a lesser impact on these outcomes. Generosity displays moderate positive associations with various factors, emphasizing its potential role in societal satisfaction. Overall, the heatmap underscores the interconnectedness of economic and social indicators in shaping well-being.

5. Key Understanding

On this section of South Asia data, Maldives ranked highest in Composite Score, comparing with its happiness score.

Problem - 3 - Comparative Analysis:

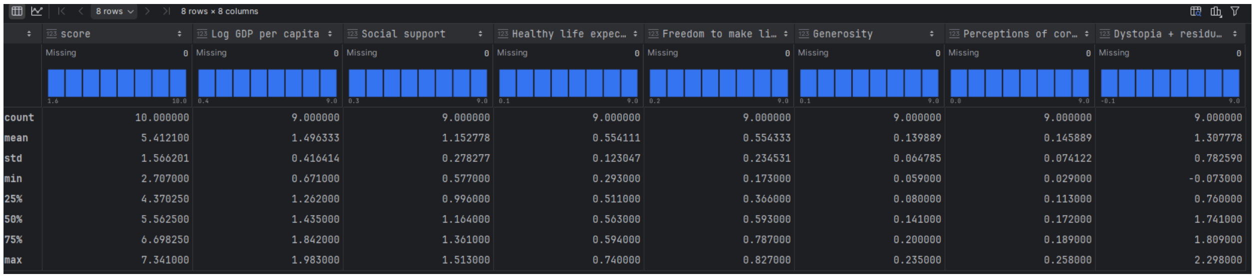
Overview

This section compares South Asia with the Middle East across key metrics, variability in scores, and outlier characteristics.

Tasks and Explanation

1. Filtering Middle East

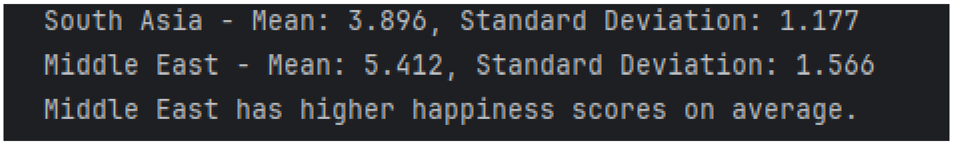
Firstly, a list of Middle East countries was extracted from the main dataset csv file and saved the filtered dataset to a new csv file. The data is then tried to understand using the describe method.



	score	Log GDP per capita	Social support	Healthy life expec…	Freedom to make li…	Generosity	Perceptions of cor…	Dystopia + residu…
	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0
count	10.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000
mean	5.412100	1.496333	1.152778	0.554111	0.554333	0.139889	0.145889	1.307778
std	1.566201	0.416414	0.278277	0.123047	0.234531	0.064785	0.074122	0.782590
min	2.707000	0.671000	0.577000	0.293000	0.173000	0.059000	0.029000	-0.073000
25%	4.370250	1.262000	0.996000	0.511000	0.366000	0.080000	0.113000	0.760000
50%	5.562500	1.435000	1.164000	0.563000	0.593000	0.141000	0.172000	1.741000
75%	6.698250	1.842000	1.361000	0.594000	0.787000	0.200000	0.189000	1.809000
max	7.341000	1.983000	1.513000	0.740000	0.827000	0.235000	0.258000	2.298000

2. Statistical Comparison

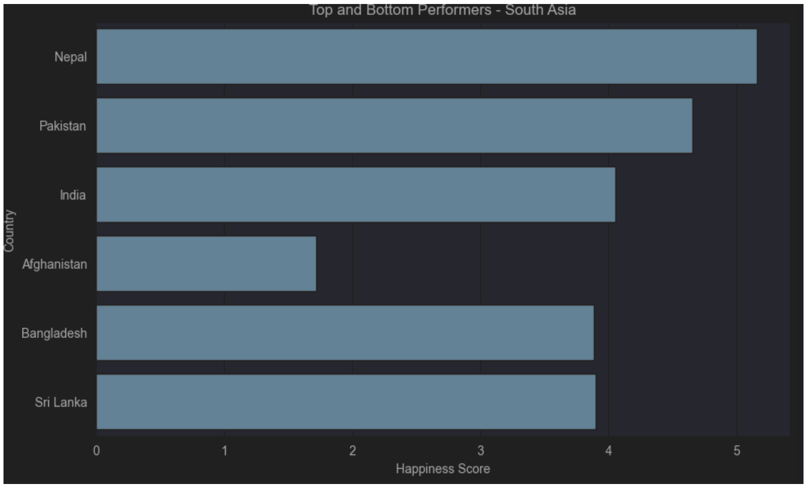
On this statistical comparison part, the means and standard deviation was calculated for the happiness score which resulted on having the highest score of Middle East.

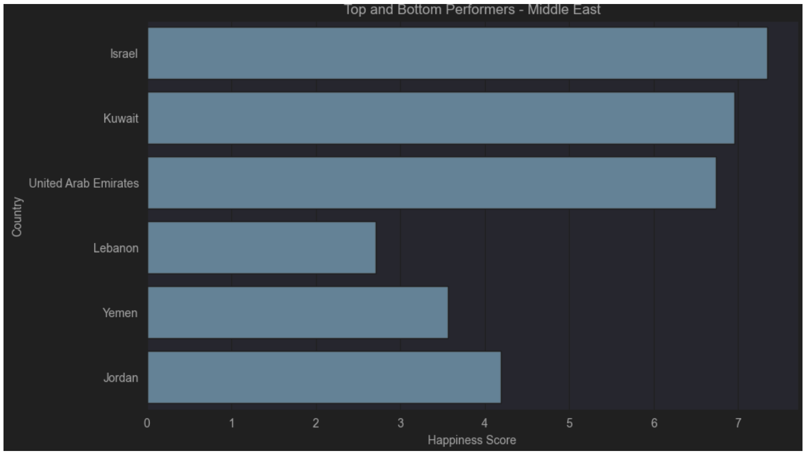


```
South Asia - Mean: 3.896, Standard Deviation: 1.177
Middle East - Mean: 5.412, Standard Deviation: 1.566
Middle East has higher happiness scores on average.
```

3. Top and Botton Performers

On this comparison, GDP per capita, social support, and healthy life expectancy is compared of the two regions, South Asia and Middle East using the bar graph.

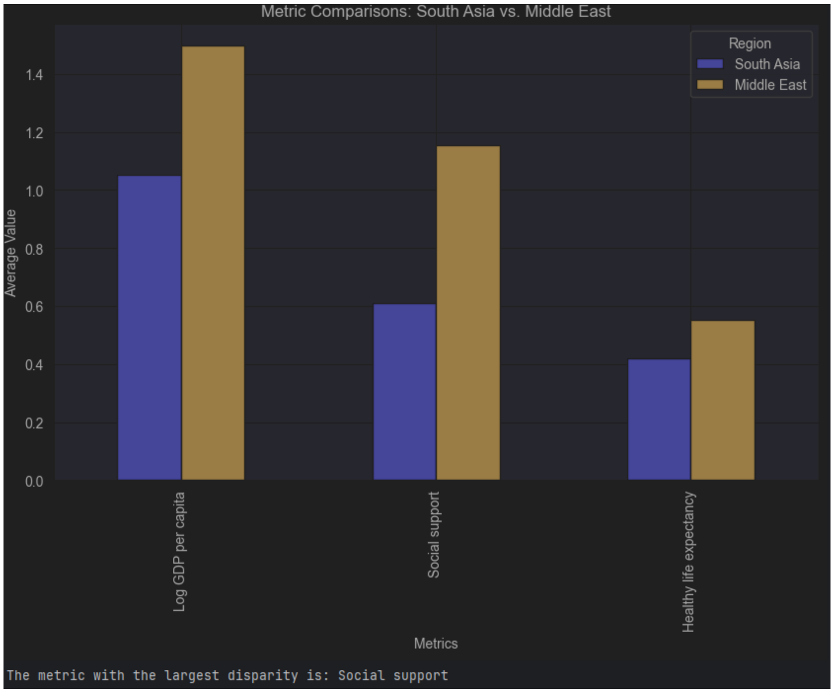




The bar charts show that Nepal, Pakistan, and India are the top performers in South Asia, while Afghanistan, Bangladesh, and Sri Lanka rank the lowest. In the Middle East, Israel, Kuwait, and the UAE lead with higher scores, while Lebanon, Yemen, and Jordan perform the worst. The Middle East shows greater disparities in happiness compared to South Asia.

4. Happiness Disparity

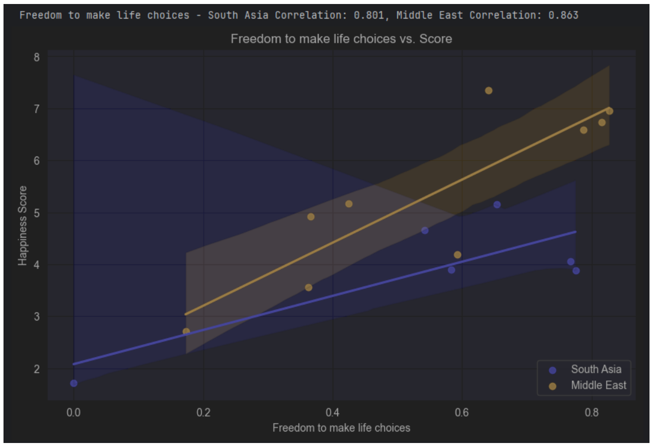
The bar chart below shows that the Middle East outperforms South Asia across all metrics: GDP per Capita, Social Support, and Healthy Life Expectancy. The metric with the largest disparity is Social Support, highlighting a significant gap in community or governmental assistance between the two regions.



```
The metric with the largest disparity is: Social support
```

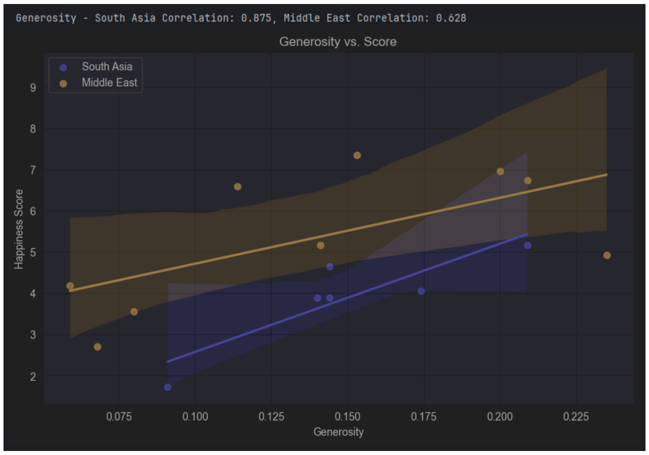
5. Correlation Analysis

The correlation analysis reveals that in both South Asia and the Middle East, "Freedom to Make Life Choices" has a strong positive correlation with Happiness Scores, with the Middle East showing a slightly higher correlation (0.863 vs. 0.801). Similarly, "Generosity" positively correlates with Happiness Scores in both regions, though the correlation is notably stronger in South Asia (0.875) compared to the Middle East (0.628). Scatter plots illustrate these relationships, highlighting stronger associations for both metrics in South Asia. These findings suggest regional variations in how these factors influence happiness.



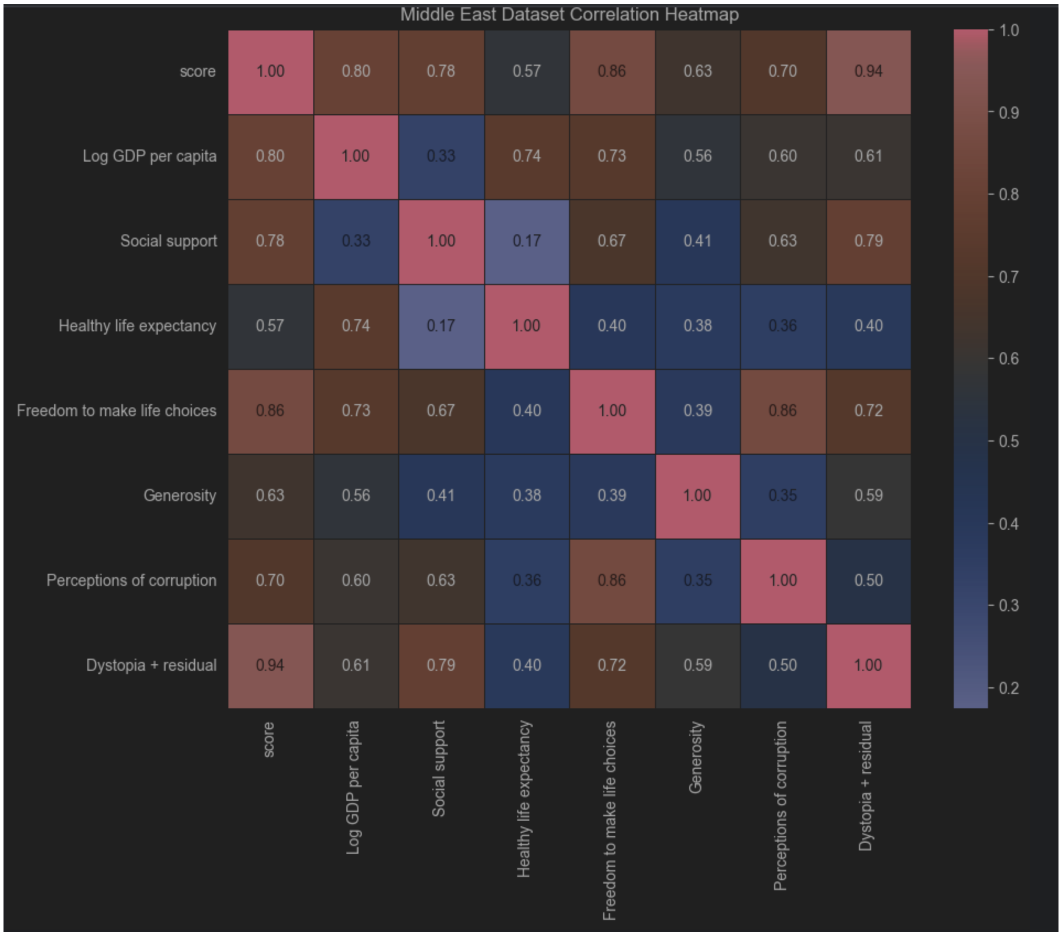
```
Freedom to make life choices - South Asia Correlation: 0.801, Middle East Correlation: 0.863
```

Freedom to make life choices vs. Score
Happiness Score
8
7
6
5
4
3
2
0.0
0.2
0.4
0.6
0.8
Freedom to make life choices
South Asia
Middle East



```
Generosity - South Asia Correlation: 0.875, Middle East Correlation: 0.628
```

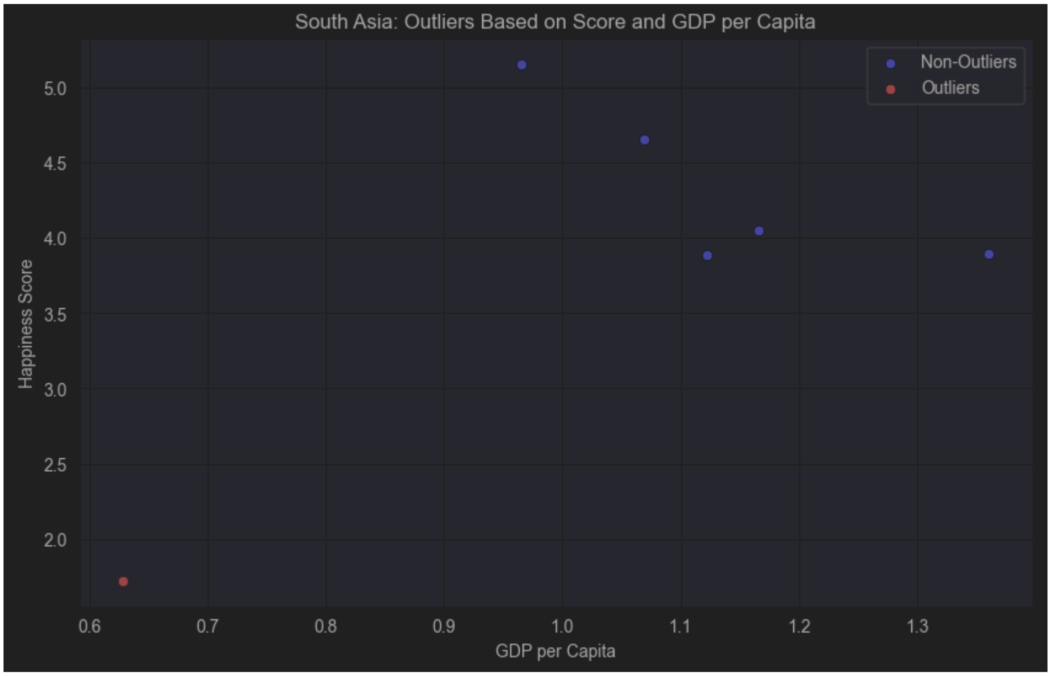
The heatmaps indicated stronger correlations in South Asia between happiness and GDP per capita compared to the Middle East.

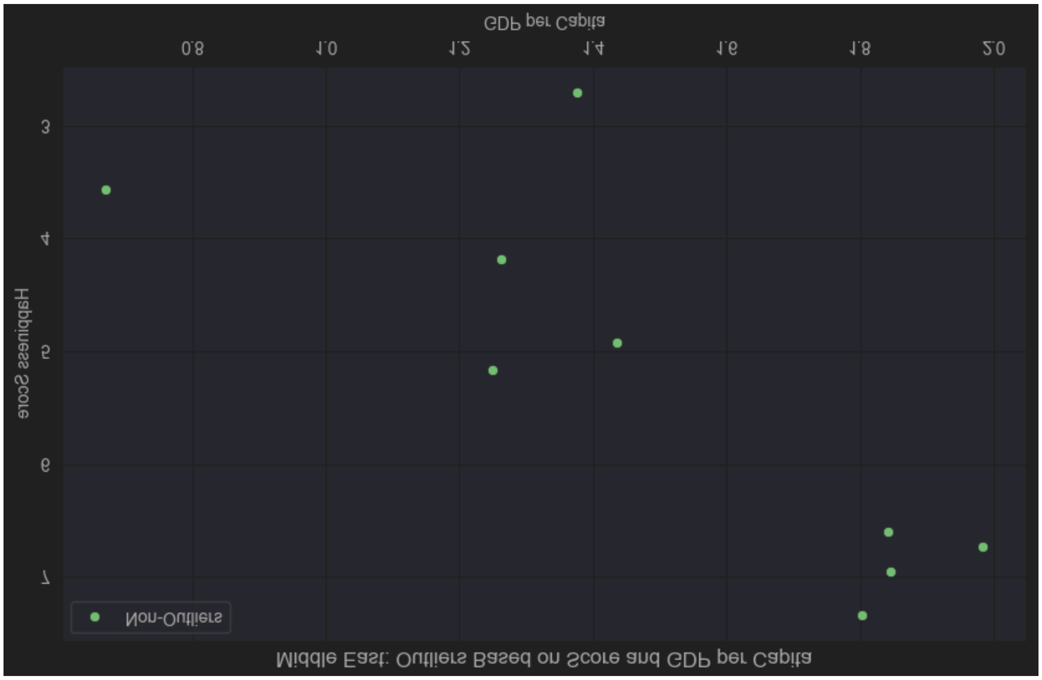


The heatmap shows that happiness in the Middle East is closely linked to "Freedom to Make Life Choices," "GDP per Capita," and "Social Support," as these factors have strong positive relationships with the happiness score. On the other hand, "Generosity" and "Healthy Life Expectancy" have weaker connections to happiness. Some factors, like "Freedom to Make Life Choices" and "Perceptions of Corruption," are also strongly connected to each other, showing how these aspects of life are related. Overall, the results highlight how important economic stability and personal freedoms are for happiness in the region.

6. Outliers

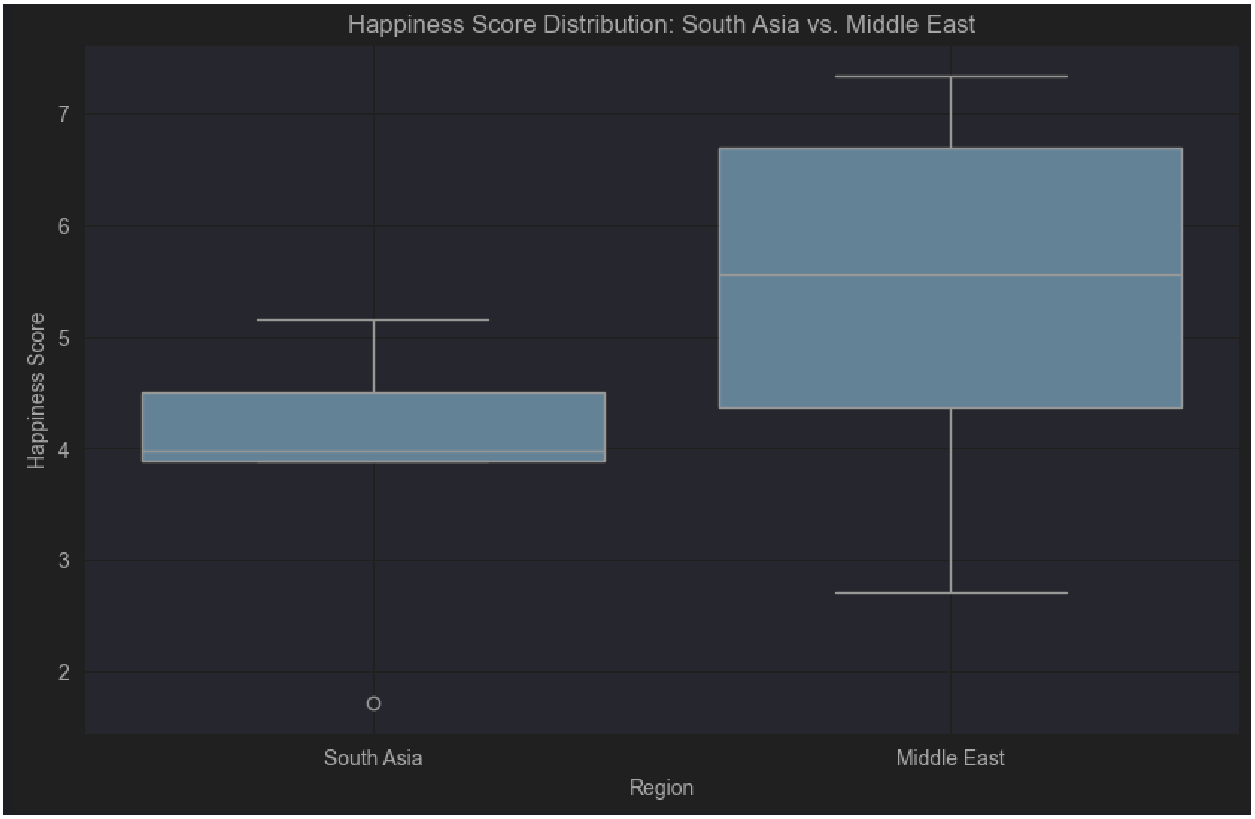
The analysis of outliers in the GDP per capita and Happiness Score relationship shows that South Asia has a noticeable outlier (marked in red), which significantly deviates from the main pattern of data. This suggests an unusual case where a country has a lower happiness score despite a moderate GDP per capita. On the other hand, the Middle East region does not display any outliers, with all data points closely aligning with the overall trend. This indicates a consistent relationship between GDP and happiness within the Middle East dataset.





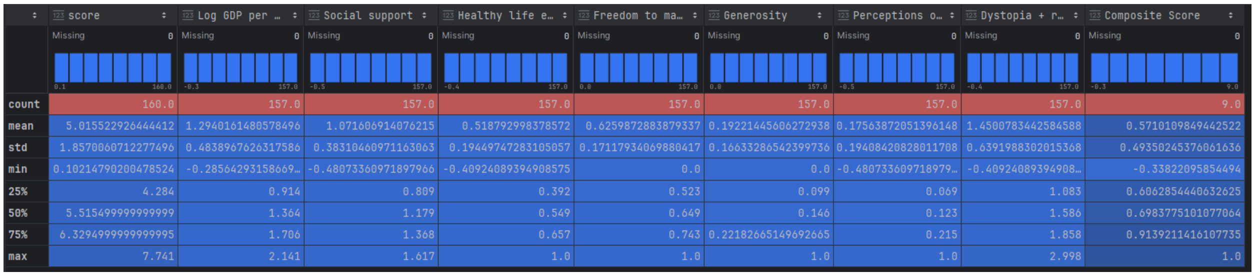
7. Visualization

The boxplot is created which is comparing the distribution of Score between South Asia and the Middle East.

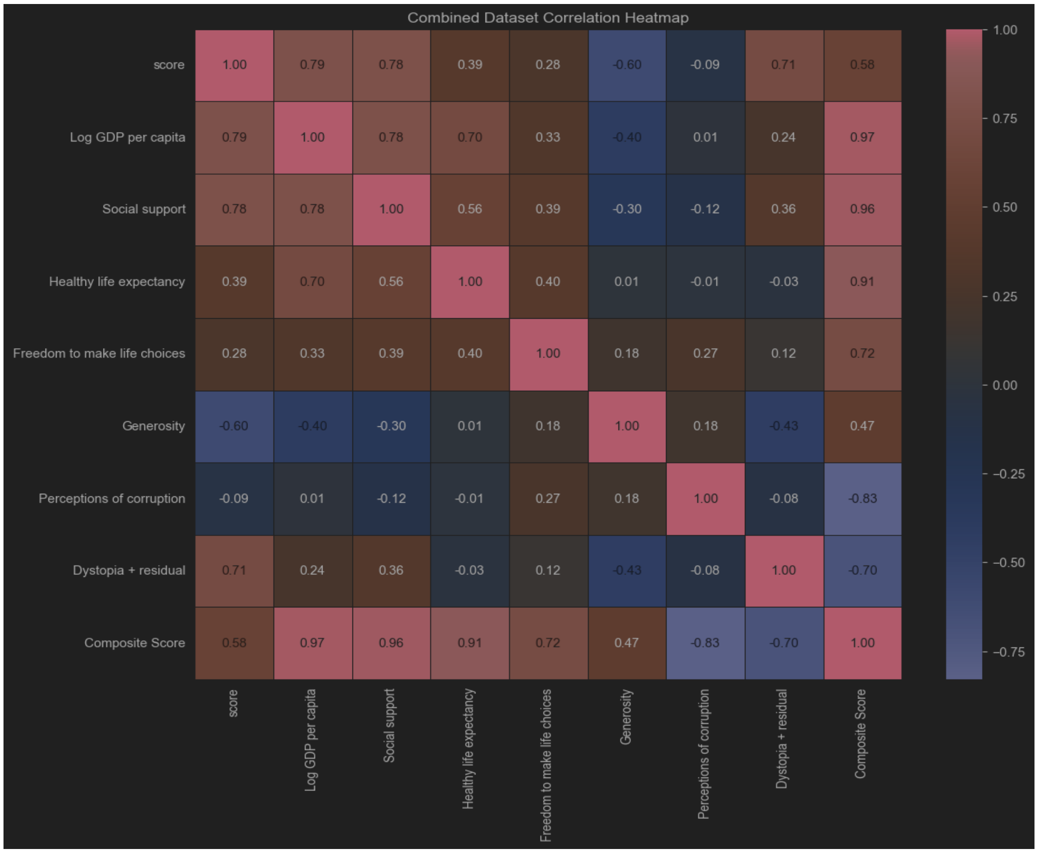


The boxplot provides a clear comparison of happiness score distributions between South Asia and the Middle East. South Asia's scores are tightly clustered, with a lower median and one visible outlier at the lower end. In contrast, the Middle East shows a broader range of happiness scores, a higher median, and no outliers. These differences highlight more variability and generally higher happiness levels in the Middle East.

Combined description and heatmaps:



	score	Log GDP per …	Social support	Healthy life e…	Freedom to ma…	Generosity	Perceptions o…	Dystopia + r…	Composite Score
Missing	0	0	0	0	0	0	0	0	0
	0.1 160.0	-0.3 157.0	-0.5 157.0	-0.4 157.0	0.0 157.0	0.0 157.0	-0.5 157.0	-0.4 157.0	-0.3 9.0
count	160.0	157.0	157.0	157.0	157.0	157.0	157.0	157.0	9.0
mean	5.015522926444412	1.2940161480578496	1.071606914076215	0.518792998378572	0.6259872883879337	0.19221445606272938	0.17563872051396148	1.4500783442584588	0.571010984944252
std	1.8570060712277496	0.4838967626317586	0.38310460971163063	0.19449747283105057	0.17117934069880417	0.16633286542399736	0.19408420828011708	0.6391988302015368	0.49350245376061636
min	0.10214790200478524	-0.28564293158669…	-0.48073360971897966	-0.40924089394908575	0.0	0.0	-0.480733609718979…	-0.40924089394908…	-0.33822095854494
25%	4.284	0.914	0.809	0.392	0.523	0.099	0.069	1.083	0.6062854440632625
50%	5.515499999999999	1.364	1.179	0.549	0.649	0.146	0.123	1.586	0.6983775101077064
75%	6.3294999999999995	1.706	1.368	0.657	0.743	0.22182665149692665	0.215	1.858	0.9139211416107735
max	7.741	2.141	1.617	1.0	1.0	1.0	1.0	2.998	1.0



	score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual	Composite Score
score	1.00	0.79	0.78	0.39	0.28	-0.60	-0.09	0.71	0.58
Log GDP per capita	0.79	1.00	0.78	0.70	0.33	-0.40	0.01	0.24	0.97
Social support	0.78	0.78	1.00	0.56	0.39	-0.30	-0.12	0.36	0.96
Healthy life expectancy	0.39	0.70	0.56	1.00	0.40	0.01	-0.01	-0.03	0.91
Freedom to make life choices	0.28	0.33	0.39	0.40	1.00	0.18	0.27	0.12	0.72
Generosity	-0.60	-0.40	-0.30	0.01	0.18	1.00	0.18	-0.43	0.47
Perceptions of corruption	-0.09	0.01	-0.12	-0.01	0.27	0.18	1.00	-0.08	-0.83
Dystopia + residual	0.71	0.24	0.36	-0.03	0.12	-0.43	-0.08	1.00	-0.70
Composite Score	0.58	0.97	0.96	0.91	0.72	0.47	-0.83	-0.70	1.00

The combined heatmap and descriptive table shows/reflects the key relationships between happiness factors. Log GDP per capita, social support, and healthy life expectancy are strongly correlated with the happiness score (correlations of 0.79, 0.78, and 0.39, respectively), indicating their importance in explaining happiness levels. Generosity shows a negative correlation with the happiness score (-0.60), suggesting it might play a less direct role. Overall, the composite score aggregates these relationships and aligns closely with the happiness score (correlation: 0.58). Descriptive statistics show that the different factors have different levels of impact on happiness.

Conclusion

The analysis of data revealed several important findings:

1. **Problem 1 (Global)**: GDP per capita is one of the strongest factors affecting happiness worldwide. Its high correlation with happiness scores suggests that economic stability plays a key role in overall well-being.
2. **Problem 2 (South Asia)**: In South Asia, economic and social indicators, such as GDP and social support, heavily influence happiness. The Maldives outperformed other countries in both the composite and happiness scores, emphasizing the importance of these factors in the region.
3. **Problem 3 (Middle East)**: The Middle East displayed greater variability in happiness levels, driven by wide disparities in GDP and social support among countries. This uneven distribution highlights the complexity of happiness factors in the region.

Combined-Analysis:

The correlation heatmap showed that GDP, social support, and health have a strong positive influence on happiness, while generosity has a weaker or even negative correlation. Descriptive statistics reinforced these findings, showing the varying strength of each factor.

In summary, the data confirms that economic and social factors are crucial for happiness but have different effects across regions. Future work can focus on tracking trends over time or exploring additional factors to provide a deeper understanding.