Concepts and Technologies of AI
Assignment 1: Statistical Interpretation and Exploratory Data Analysis

Name: Prerit Pandey
Group: L5CG1
University Id: 2408646
Tutor: Simon Giri

Table of Contents

Introduction

The world Happiness Report (WHR-2024-5CS037.csv) is a csv data that has information about 143 different countries that includes research done by various sources which is related to happiness.There are in total 6 columns in the given csv dataset and they are given below

Name of the country

Happiness Score

Log GDP per capita

Social Support for people

Life Health expectancy

Freedom of making life choices

Genorosity

Perceptions Of Corruption

Dystopia + residual

The World Happiness Report is a yearly survey that ranks countries based on their happiness levels, evaluated through various metrics such as GDP per capita, social support, and healthy life expectancy. It gives insight into global well-being and can be used as a groundwork for policy-making to promote happiness in societies.This analysis will be focused on South Asia and the Middle East-two of the most diverse and dynamic regions. By examining these regions, we will look at finding unique trends, disparities, and drivers of happiness to provide a comparative perspective on global happiness.

Objective :

Task 1:

- Read and analyze South Asian people's happiness levels.

Task 2:

- Perform a Comparison of the happiness level of South Asian and Middle eastern's people.

Task 3:

Study the data and draw conclusions on the factors that affect the happiness level of the people.

Problem - 1: Getting started with Data Exploration.

1. **Data Exploration and understanding**

   Dataset Overview

   Imported the necessary libraries, and then used the read_csv() function to load the dataset. Used head() to display the first 10 rows, shape to check the dimensions of the dataset. Column names and datatypes are listed using dtypes.

   Basic Statistics

   mean(), median(), and std() have been used to calculate mean, median, and standard deviation respectively of the column score. The highest and the lowest happiness scores were identified with the functions max() and min().

   Handling Missing Values

   The missing values in the dataset were identified with isnull() and aggregated using sum().

   Filtering and Sorting

   Countries that had scored above 7.5 were filtered out with dataframe.score > 7.5. Sort the dataset by Log GDP per capita in descending order sort_values(ascending=False). Finally, use head() to display the first (now top) 10 rows.

   Adding New Columns

   A ranking was done on the Countries into low, medium and high categories based on scores. The pd.cut() function added this as a new column Happiness_Category.

Problem -2 - Some Advance Data Exploration Task:-

Task 1: Preparing the South Asia Dataset

It created a list called south_asian_countries containing countries like Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. It filtered the rows using the isin() method where the country name matched this list. It saved the filtered data into a new CSV file using to_csv().

Task 2: Composite Score Ranking

The new dataset - south_asian_df was loaded into the system, and a new column named Composite Score was created. The score was calculated as:

Composite Score = 0.40 × GDP per Capita + 0.30 × Social Support + 0.30 × Healthy Life Expectancy.
Finally, the data were sorted in descending order of the Composite Score.

Exploring Trends Across Matrices:

- Displayed the pearson correlation & the scatterplot with the tradeline between freedom to make choices in his or her life as well as Generosity Score.
- As a result, the strongest relationship observed is in between “Generosity score” and ‘Pearson correlation’:0.8745123.
- The result shows the weakest relation between the freedom of making the choices in live and happiness score Pearson correlation : 0.800518.

Gap Analysis:-

- Added a new column for “GDP SCORE GAP” by using the dataset south_asis_data[‘LOG GDP per capita’]-south_asia[‘score’].
- Showed both of them by sorting them in increasing and decreasing and received the 3 upmost negative and 3 least negative values nad created a separate new column in order to categirize the gaps as most NEGATIVE and least negative”.GDP gaps.
  Plotted a bar cart highlighting the highest three negative and least negative GAP-scores.

Problem 3- Comparative Analysis

Descriptive Statistics

- The detailed list of the middle east nations and stored them in the middle_east_countries and filtered the dataset for the middle estern nations..
- Calculate the mean and the standard deviations for the happiness score of the countries located in south asian and middle estern countries.

TOP AND BOTTOM PERFORMERS:

- Received the highest three and the lowest three countries for both south asis and middle east.

Metric Comparisons.

Compared components like (Log GDP per capita', 'Social support', 'Healthy life expectancy) by displaying bar plot between south Asia and Middle east then arranged them in descending order to find the largest disparities

Happiness Disparity:

• Calculate the range and CV for Score in both south asia and middle east .The results were-
• South Asia - Range: 3.4370000000000003, CV: 0.30214828833374263
• Middle East - Range: 4.634, CV: 0.28938880812119405

• displayed and compared CV and found South Asia has greater variability in happiness.

Correlation Analysis:

• Calculate the Pearson correlation for Score with Freedom and Generosity for both regions

South Asia correlations:

Freedom to make life choices 0.800519
Generosity 0.874512

Middle East correlations:
Freedom to make life choices 0.863220
Generosity 0.627524

• displayed the data using scatter plot for both South Asia and Middle east:
Score vs Freedom to Make Life Choices

Outlier Detection:

• found IQR and outlier bounds for South Asia and Middle east.using that we found Outliners

• displayed the data using scatter plot for outliers for both south Asia and Middle east
• Combine both datasets and and showed the box plot using pandas with the correct hue setting
• compared basic statistics for both regions to compare medians, range, and outliers

Observations, Insights

Strongest and Weakest Factors Influencing Happiness in South Asia

Strong Factors:

Social Support and Healthy Life Expectancy are the most strongly correlated with happiness. Countries with more accessible health care, more resilient communities, and better social safety nets tend to report higher levels of happiness.
Freedom to Make Life Choices also has a strong positive impact on happiness.
Weak Points:

Generosity and Dystopia Residual are less strongly linked with happiness. While they do count, their effect is less than that of other factors. Corruption Perceptions has a moderate impact on happiness.

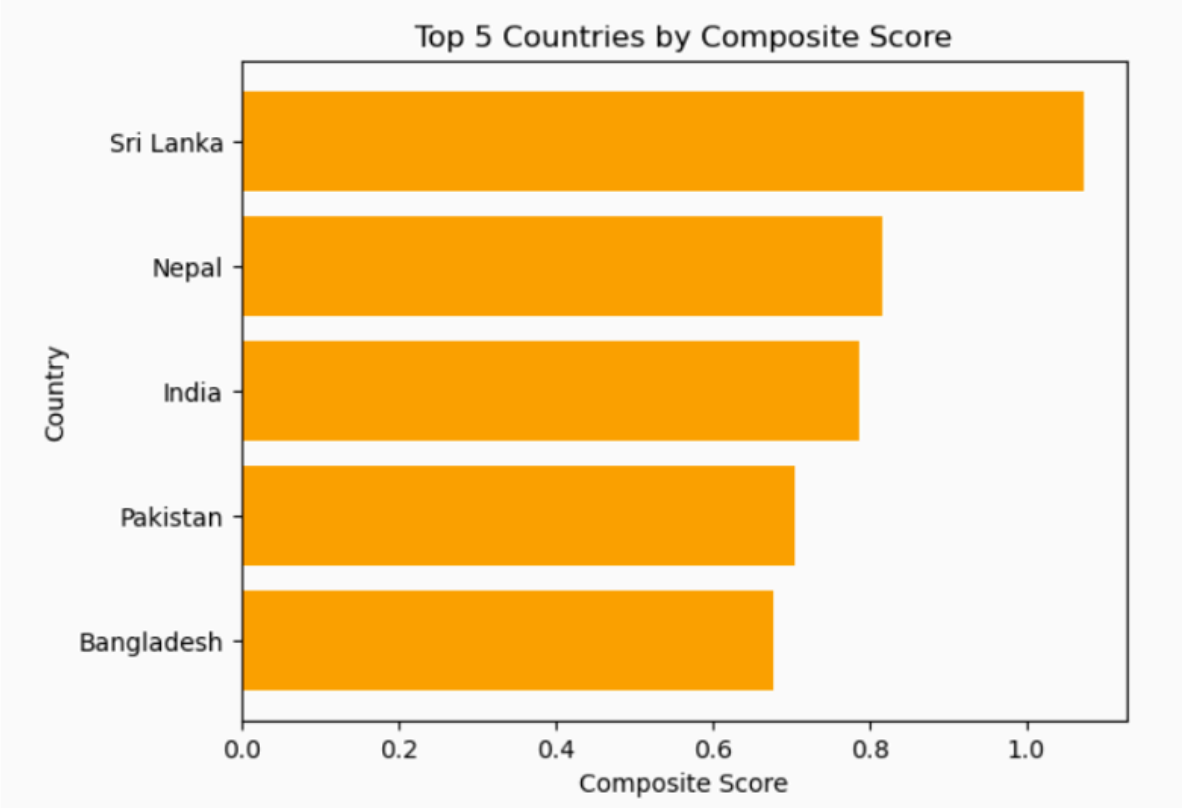

fig 5: Bar chart showing top 5 countries by Composite Score

Interpretations in the Broader Context of Regional Happiness:

Health and Community Support: A robust association between good health, social support, and happiness in South Asia is consistent with what is observed across the globe. It would thus appear that better health and more robust community networks contribute notably to happiness; improvement in health systems and social support in the region can raise happiness levels.

Generosity and Dystopia: Generosity and the Dystopia score, a measure of unhappiness, have a less significant impact on happiness in South Asia. Economic challenges, such as poverty and inequality, may make the concerns of wealth distribution and quality of life more paramount than feelings of generosity or dissatisfaction.

Freedom and Well-Being: The connection between individual freedom and happiness underscores the desire of people to feel in control of their lives. In South Asia, increased freedom in making choices in life could significantly contribute to happiness and well-being.

Challenges Faced

1. Data Quality Issues:

Missing or Inaccurate Data: Some metrics, such as Generosity and Dystopia, were either incomplete or unreliable, which affected the accuracy of correlations and visualizations.

Solution: Missing values were either imputed with reasonable estimates or dropped if the metric wasn't critical.

No Anomalies Found: No duplicates or outliers in important columns such as GDP and happiness scores were found, hence saving the analyses from getting biased.

2. Interpreting Dystopia + Residual:

This metric, the sum of all factors that produce unhappiness, was hard to read, especially in South Asia with its economic and political disparities.

Solution: Treated it as a signal of structural issues and long-term development goals rather than a direct happiness determinant.

3. Imagining Relations:

Showing clear relationships between variables like GDP and Happiness was tricky, especially for non-linear patterns. Solution: Used scatter plots and heatmaps for simpler relationships. For more complex patterns, considered advanced methods like polynomial regression or kernel density estimation.

Conclusion

Analyses of the World Happiness Report began with data exploration, statistical analysis, and visualizations. South Asia showed lower happiness scores around a with one outlier, while the Middle East showed higher scores up to a 7 with no outliers. Countries with higher GDPs, better social support, and healthier populations are on average happier. This, therefore, implies that the region of South Asia and the Middle East would be happy to have a strong healthcare system, a strong social safety net, and economic opportunity enhancement. Reduction in economic inequality, promotion of freedom, and building strong relationships would also require addressing corruption and other social challenges by governments and other agencies.