Exploratory Data Analysis of the 2024 Happiness Index Report

Introduction to Happiness Index

The World Happiness Report is a landmark **survey of the state of global happiness**. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

Understanding the Dataset

Here's a brief explanation of each column in the dataset:

1. **Country name:** Name of the country.
2. **Regional indicator:** Region of the country.
3. **Ladder score:** Happiness score based on the Cantril Ladder question (scale 0–10).
4. **Upper whisker:** Upper bound of the happiness score.
5. **Lower whisker:** Lower bound of the happiness score.
6. **Log GDP per capita:** Logarithm of GDP per capita (PPP-adjusted).
7. **Social support:** Binary average for having someone to rely on in trouble.
8. **Healthy life expectancy:** Average years a newborn is expected to live in good health.
9. **Freedom to make life choices:** Average satisfaction with personal freedom.
10. **Generosity:** Residual of donating responses regressed on GDP.
11. **Perceptions of corruption:** Average responses on perceived corruption in government/business.
12. **Dystopia + residual:** Benchmark ensuring no score is below Dystopia.

Project Workflow

Data Cleaning

1. Handling missing values
2. Identifying Outliers
3. Normalisation

Exploratory Data Analysis

1. Visualization using Matplotlib and Seaborn
2. Correlation Analysis
3. Covariance Analysis

Statistical Analysis

1. Descriptive Statistics
2. Inferential Statistics
3. Hypothesis Testing

Data Cleaning

Handling Missing Values

- Missing values were identified using methods like *.isnull().sum()* to determine the extent of missing data in each column.
- For numerical columns, missing values were. imputed using the mean.

```
Country name                      0
Regional indicator                0
Happiness score                   0
upperwhisker                      0
lowerwhisker                      0
Log GDP per capita                3
Social support                    3
Healthy life expectancy           3
Freedom to make life choices      3
Generosity                        3
Perceptions of corruption         3
Dystopia + residual               3
dtype: int64
```

Finding Outliers

- Used box plots to visually detect outliers. Data points outside the whiskers of the box plot are considered potential outliers.

Normalisation

- Normalization transforms the data to a specific range, typically [0, 1]. This ensures all features contribute equally to model training by scaling them proportionally.
- Used *MinMaxScaler* from *sklearn.preprocessing*

Exploratory Data Analysis

Visualization using Matplotlib and Seaborn

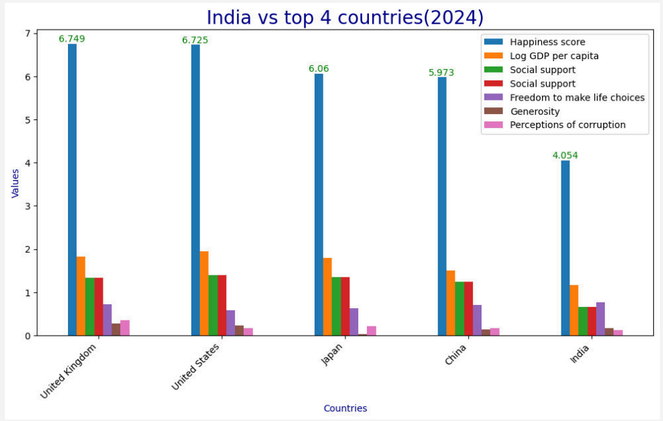
Histogram Plot:
• Analyzed frequency distribution.
• Explored relationships of key factors (e.g., GDP, social support, etc.) with the Happiness Score.

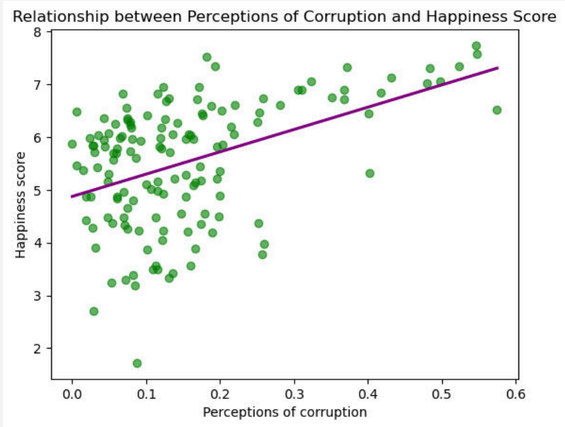
Boxplot:
• Compared the Happiness Score across different regions to identify variations and outliers.

Grouped Bar Graph:
• Compared key factors (e.g., GDP, social support, etc.) of India with other nations (United States, United Kingdom, China, and Japan).

Horizontal Bar Graph:
• Highlighted the happiest and unhappiest countries for a comparative view.

Scatter Plot:
• Visualized the relationship of the Happiness Score with other key factors





Exploratory Data Analysis

Correlation Analysis

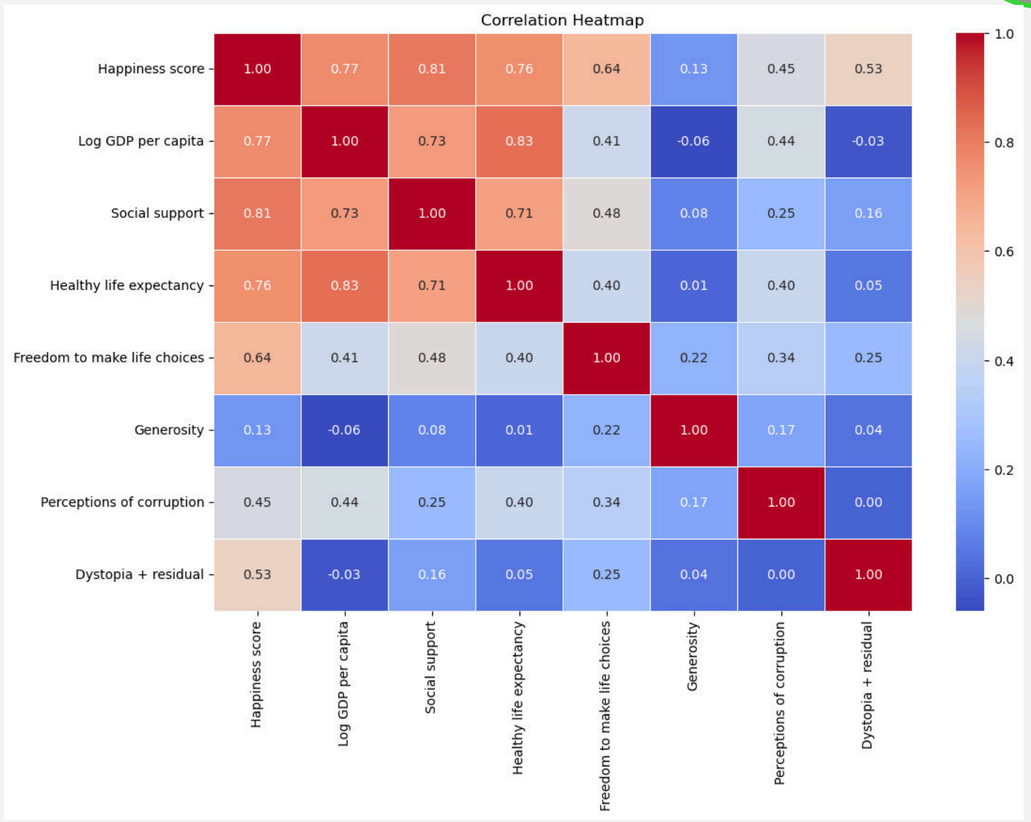
Why Correlation is Used for?

• Correlation helps identify the strength and direction of relationships between variables, enabling us to determine which factors most influence the Happiness score.

• Log GDP per capita, Social support, and Healthy life expectancy show a strong positive correlation with the Happiness score.

• Perceptions of corruption and Dystopia + residual exhibit a moderate correlation with the Happiness score.

• Generosity has a weak correlation with all other metrics, including the Happiness score.



Exploratory Data Analysis

Covariance Analysis

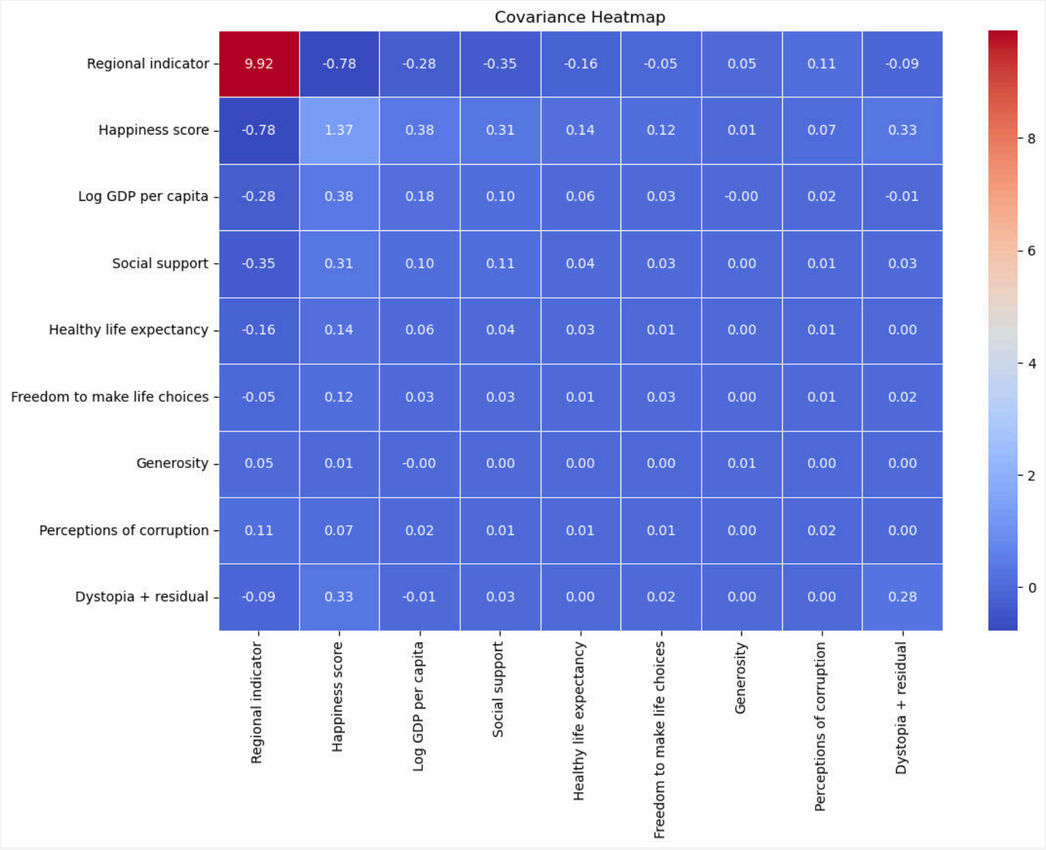
Why Covariance is Used for?

• Covariance measures the direction of the relationship between two variables, helping to identify how changes in one variable are associated with changes in another.

• Economic factors like Log GDP per capita and Social support show the strongest positive influence on Happiness score.

• Healthy life expectancy and Freedom to make life choices exhibit moderate positive covariance with Happiness score

• Psychological factors like Generosity and Perceptions of corruption show weaker covariance with the Happiness score.

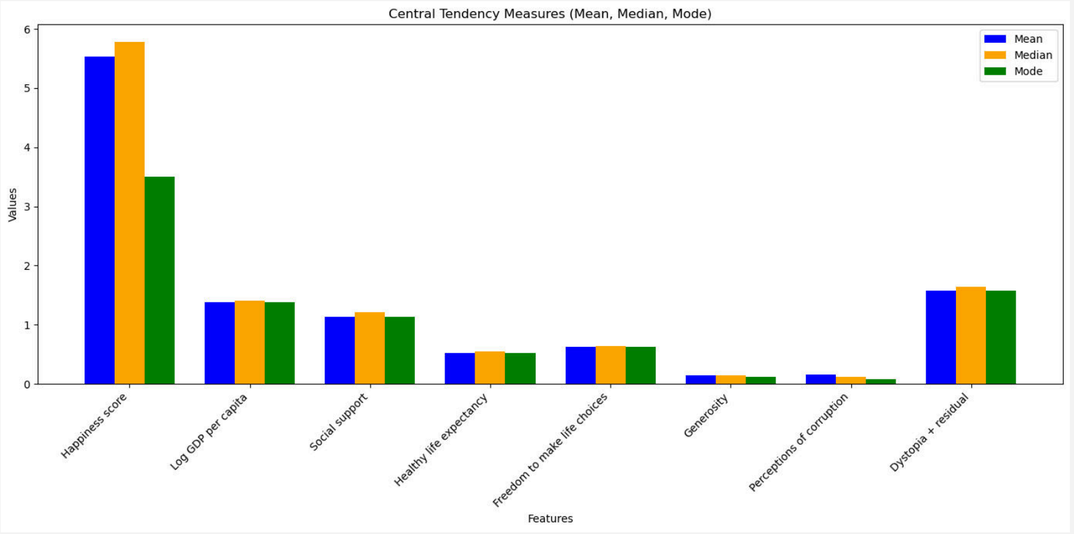


Statistical Analysis

Descriptive Statistical Analysis

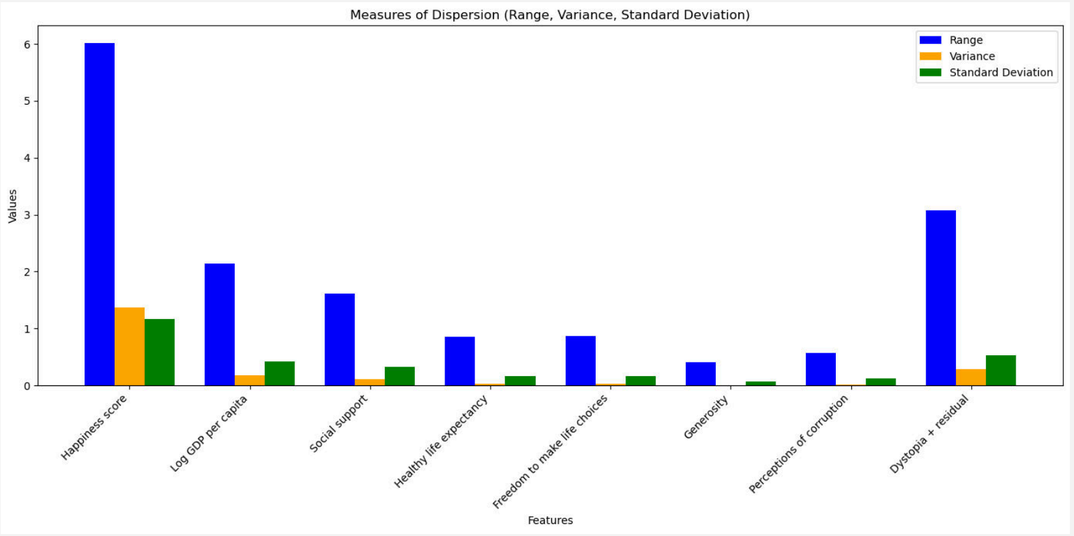
Measures of Central Tendency

• Happiness score has the highest central tendency values.
• Most variables have central tendencies between 0 and 2.
• The closeness of mean and median for most variables suggests symmetric distributions.



Measures of Dispersion

• Apart from Happiness score, the highest variance is observed in Log GDP per capita and Dystopia + residual
• These features show significant variability compared to others, reflecting their diverse distribution across countries.



Statistical Analysis

Inferential Statistical Analysis

Hypothesis Testing

Performing t-test
- Null Hypothesis: The sample mean is equal to the population mean.
- Alternating Hypothesis: The sample mean is different from the population mean.
- Test Result: No significant difference between sample and population mean.

Performing chi-squred test
- Null Hypothesis: The two variables are independent.
- Alternating Hypothesis:The two variables are not independent.
- Test Result: There is an association between GDP and Happiness categories.

Performing ANOVA test
- Null Hypothesis: The means of all groups are equal.
- Alternating Hypothesis: At least one group mean is different.
- Test Result: There is a significant difference in happiness scores across corruption categories.

Top Features Contributing to Happiness

1. Log GDP per Capita:
• Strong correlation and covariance with happiness scores.
• Higher GDP per capita indicates better economic stability, positively impacting happiness.

2. Social Support:
• Analysis highlights a strong link between social welfare and happiness.
• Countries with better social support systems tend to have happier citizens.

3. Healthy Life Expectancy:
• A prominent factor in both correlation and covariance analyses.
• Reflects the importance of health in determining overall happiness.

Key Insights

1. Economic factors, especially GDP, play a significant role in determining happiness.

2. Social metrics like support systems and life expectancy are equally crucial.

3. Psychological metrics like Generosity and Perception of Corruption have a weaker influence, suggesting secondary importance.

4. Happiness disparities exist regionally, with wealthier and socially supportive countries scoring higher.

CONCLUSION

The analysis highlights the multifaceted nature of happiness, driven by economic, social, and health factors. The insights derived can help policymakers and researchers design strategies to improve happiness levels, focusing on strengthening social welfare systems, improving health outcomes, and boosting economic stability.