World Happiness Report 2024: A Comprehensive Analysis

Author: ORBIN SUNNY
Date: January 12, 2025

1. Executive Summary

1.1 Objective

To explore and analyse the factors contributing to happiness across countries globally, identify significant predictors of happiness, and derive actionable insights.

1.2 Dataset Overview:

- **Source:** Kaggle
- **Total Features:** 12
- **Target Variable:** Happiness Score
- **Key Features:** Log GDP per Capita, Social Support, Healthy Life Expectancy, Generosity, Perception of Corruption.

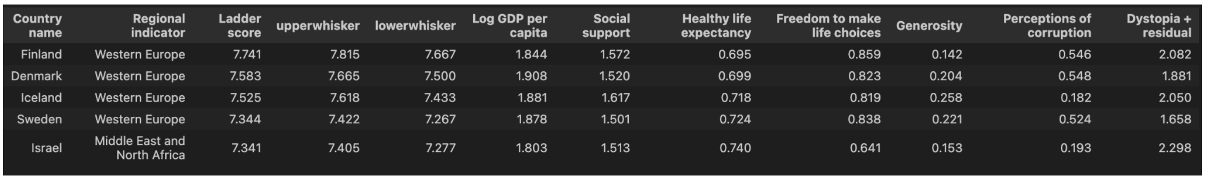
1.3 Scope of Analysis:

The project focuses on identifying patterns, relationships, and key drivers of happiness using exploratory data analysis (EDA), visualisation, and statistical testing.

Table of Contents

2. Data Preprocessing

Data preprocessing is a critical step to ensure the dataset is clean, consistent, and ready for analysis.



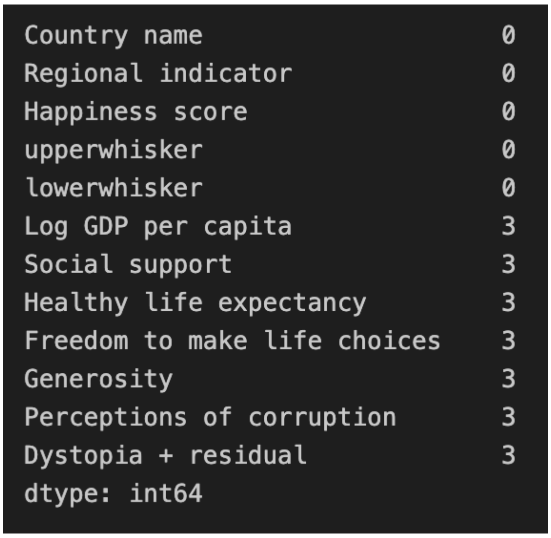
Country name	Regional indicator	Ladder score	upperwhisker	lowerwhisker	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
Finland	Western Europe	7.741	7.815	7.667	1.844	1.572	0.695	0.859	0.142	0.546	2.082
Denmark	Western Europe	7.583	7.665	7.500	1.908	1.520	0.699	0.823	0.204	0.548	1.881
Iceland	Western Europe	7.525	7.618	7.433	1.881	1.617	0.718	0.819	0.258	0.182	2.050
Sweden	Western Europe	7.344	7.422	7.267	1.878	1.501	0.724	0.838	0.221	0.524	1.658
Israel	Middle East and North Africa	7.341	7.405	7.277	1.803	1.513	0.740	0.641	0.153	0.193	2.298

2.1 Handling Missing Values

Steps Taken:

- Identifying the count of missing values in each column

```
# Checking for missing values
happy_data.isnull().sum()
```

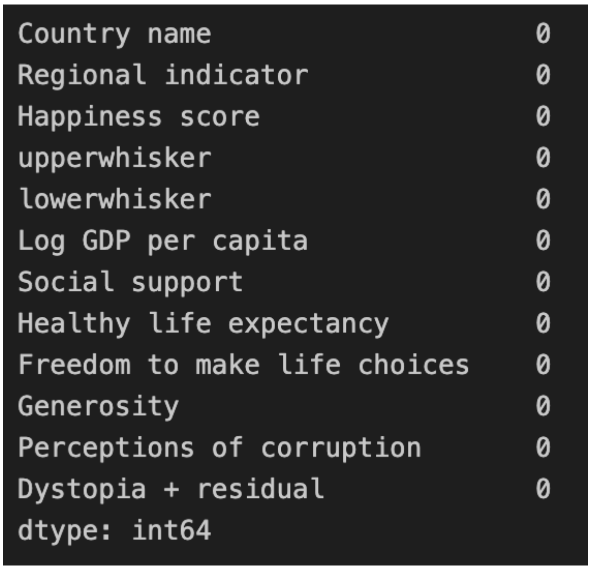


```
Country name                    0
Regional indicator              0
Happiness score                 0
upperwhisker                    0
lowerwhisker                    0
Log GDP per capita              3
Social support                  3
Healthy life expectancy         3
Freedom to make life choices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

- Imputed missing numerical values using the mean.

```
# Exclude non-numeric columns to fill missing values
numeric_cols = happy_data.select_dtypes(include=np.number).columns
happy_data[numeric_cols] =
happy_data[numeric_cols].fillna(happy_data[numeric_cols].mean())
```

Result: Missing data resolved without significant information loss.



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

2.2 Outlier Detection

Importance of Finding Outliers

Outliers are data points that deviate significantly from the rest of the dataset. Identifying outliers is crucial because they can:

- Skew the statistical analysis, affecting measures like mean and variance.
- Introduce bias into machine learning models, leading to inaccurate predictions.
- Represent significant insights, such as unique trends or anomalies in the data.

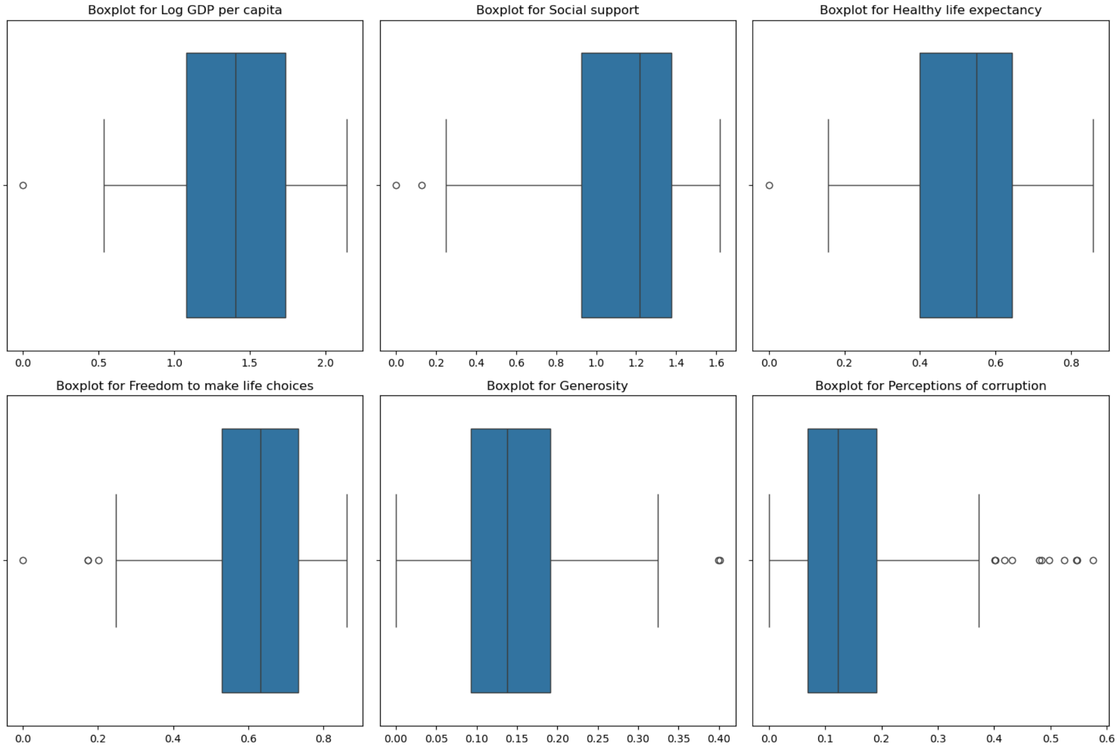
Technique Used for Detecting Outliers

To detect outliers, we used boxplots for visual analysis of each numerical feature:

- Subplots of Boxplots: Created a grid of boxplots to examine all numerical variables side by side.
- Each boxplot represents the distribution of a variable, highlighting the median, interquartile range (IQR), and potential outliers (points outside 1.5 times the IQR).

Observations on Outliers

- Perceptions of Corruption: This feature displayed several outliers, indicating countries with exceptionally high or low corruption perception scores.
- Freedom to Make Life Choices: A few countries stood out as outliers due to their extreme scores on this metric.
- Other variables like Healthy Life Expectancy and Generosity had relatively fewer outliers.



2.3 Normalization

Importance of Normalization

Normalization is a crucial step in preprocessing, especially for machine learning and statistical analysis, because:

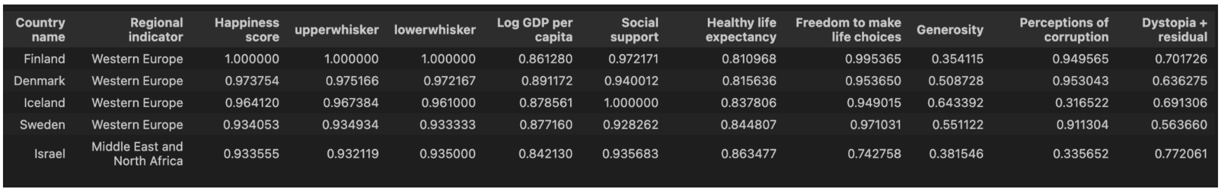
- It scales the data to a uniform range, making all features comparable.
- It reduces the impact of varying scales across features, preventing larger-scale variables from dominating smaller-scale ones.
- It helps improve the performance and stability of machine learning algorithms, particularly those relying on distance metrics (e.g., k-Nearest Neighbors, SVM).

Implementation

The *MinMaxScaler* from the *sklearn.preprocessing* module was used for normalization:

```
scaler = MinMaxScaler()
happy_data2[numeric_cols2] = scaler.fit_transform(happy_data2[numeric_cols2])
```

Result:



Country name	Regional indicator	Happiness score	upperwhisker	lowerwhisker	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
Finland	Western Europe	1.000000	1.000000	1.000000	0.861280	0.972171	0.810968	0.995365	0.354115	0.949565	0.701726
Denmark	Western Europe	0.973754	0.975166	0.972167	0.891172	0.940012	0.815636	0.953650	0.508728	0.953043	0.636275
Iceland	Western Europe	0.964120	0.967384	0.961000	0.878561	1.000000	0.837806	0.949015	0.643392	0.316522	0.691306
Sweden	Western Europe	0.934053	0.934934	0.933333	0.877160	0.928262	0.844807	0.971031	0.551122	0.911304	0.563660
Israel	Middle East and North Africa	0.933555	0.932119	0.935000	0.842130	0.935683	0.863477	0.742758	0.381546	0.335652	0.772061

3. Exploratory Data Analysis (EDA)

3.1 Visualisations

Effective visualizations were utilized to analyze and understand the distribution of numerical features, relationships between variables, and regional patterns in happiness. Below are the key visualizations and insights:

1. Histograms: Analyzing Distribution

- Objective: To evaluate the skewness and kurtosis of each numerical feature.
- Approach:
  - Plotted histograms for all numerical variables, including Log GDP per capita, Social support, Healthy life expectancy, and Generosity.
  - Assessed the shape of the distributions (e.g., normal, right-skewed, left-skewed).

- Insights:
  - Features like Log GDP per capita and Healthy life expectancy show a slightly right-skewed distribution, indicating that most countries have moderate values with a few having significantly high values.
  - Generosity and Perceptions of corruption display highly skewed distributions, with most values concentrated near zero.

2. Boxplot: Regional Happiness Comparison

- Objective: To visualise region-wise disparities in happiness (Ladder score).
- Approach:

    - Created a boxplot of Ladder score grouped by Regional indicator.
    - Compared the median, interquartile range, and outliers for each region.

- Insights:
    - Happiest Regions: Western Europe and North America have the highest median happiness scores.
    - Least Happy Regions: Sub-Saharan Africa and South Asia exhibit lower happiness scores with wider variability.
    - Significant outliers exist in each region, such as countries with unexpectedly high or low happiness scores.

3. Grouped Bar Plot: India vs. Other Nations

- Objective: To compare key metrics of India with four other nations (United States, United Kingdom, Japan, and China).
- Approach:
    - Created a grouped bar plot for Log GDP per capita, Social support, Healthy life expectancy, and Freedom to make life choices.
    - Each country was represented with separate bars for the selected metrics.

- Insights:
    - India lags behind the other four countries in most metrics but shows comparable levels of Freedom to make life choices.

    - The United States and the United Kingdom consistently outperform in economic and social metrics, correlating with higher happiness scores.

4. Scatter Plot: Relationships with Happiness Score

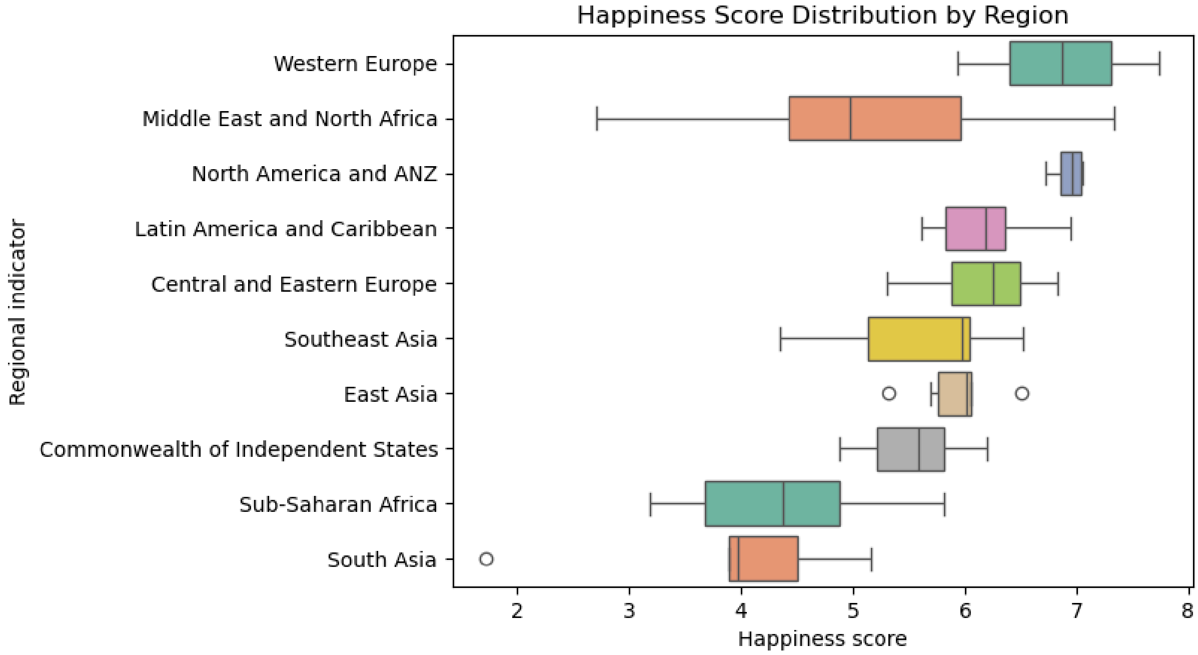
- Objective: To assess the relationships between selected features and the target variable (Happiness score).
- Approach:
    - Plotted scatter plots for Log GDP per capita, Social support, Healthy life expectancy, and Freedom to make life choices against Happiness score.
    - Fitted trend lines to observe the nature of relationships.

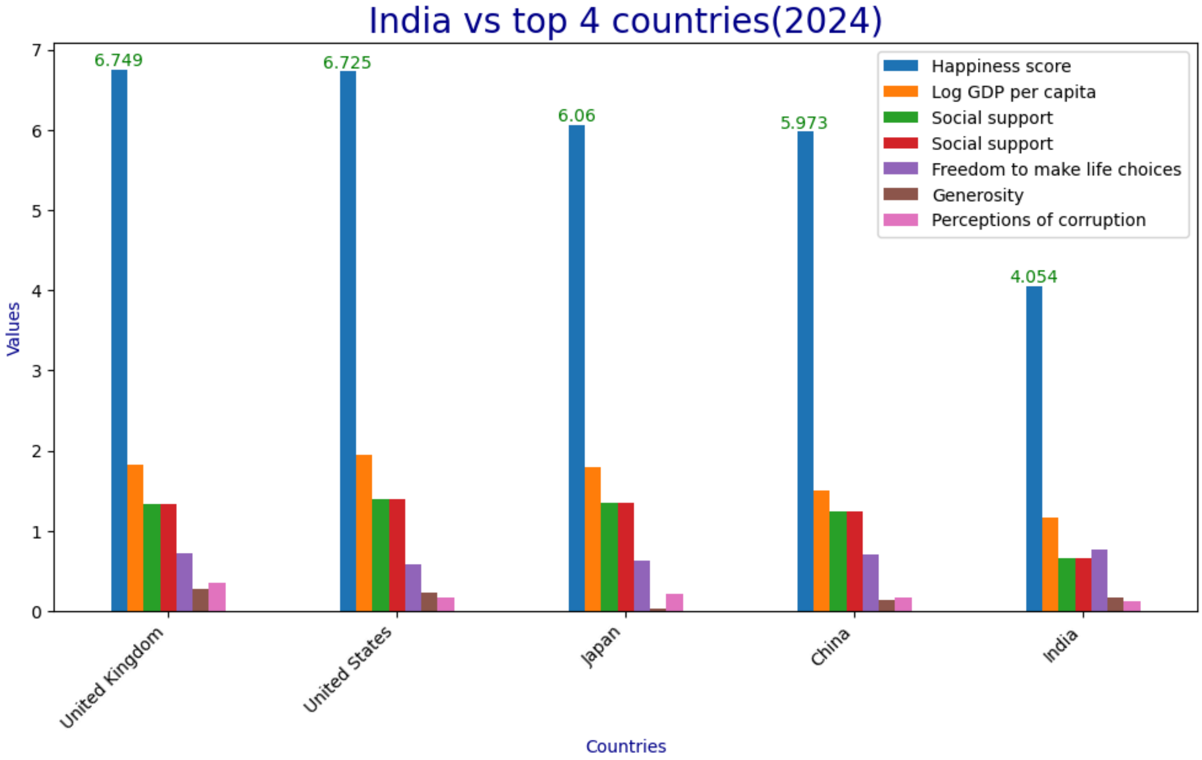
- Insights:
    - Strong Positive Correlations:
        - Log GDP per capita: Countries with higher economic output tend to have higher happiness scores.
        - Healthy life expectancy: A strong relationship between longevity and happiness is evident.

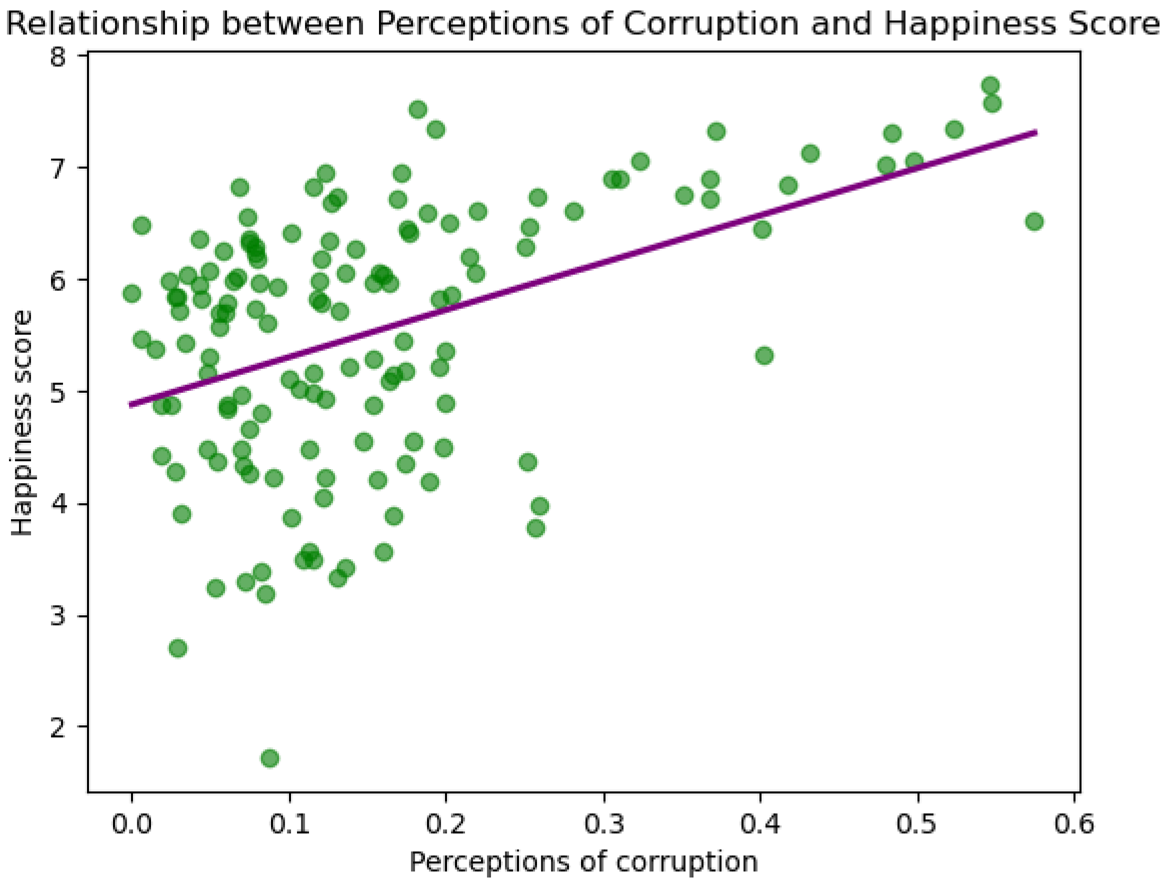
    - Moderate Relationships:
        - Freedom to make life choices and Social support show moderate positive relationships, indicating their influence on happiness but with some variability.

    - Weak Relationships:
        - Features like Generosity and Perceptions of corruption exhibit weaker or no significant correlation with happiness scores.

Example:


Happiness Score Distribution by Region
Regional indicator
Western Europe
Middle East and North Africa
North America and ANZ
Latin America and Caribbean
Central and Eastern Europe
Southeast Asia
East Asia
Commonwealth of Independent States
Sub-Saharan Africa
South Asia
2
3
4
5
6
7
8
Happiness score


India vs top 4 countries(2024)
Values
0
1
2
3
4
5
6
7
6.749
6.725
6.06
5.973
4.054
Happiness score
Log GDP per capita
Social support
Social support
Freedom to make life choices
Generosity
Perceptions of corruption
United Kingdom
United States
Japan
China
India
Countries



3.2 Correlation Analysis

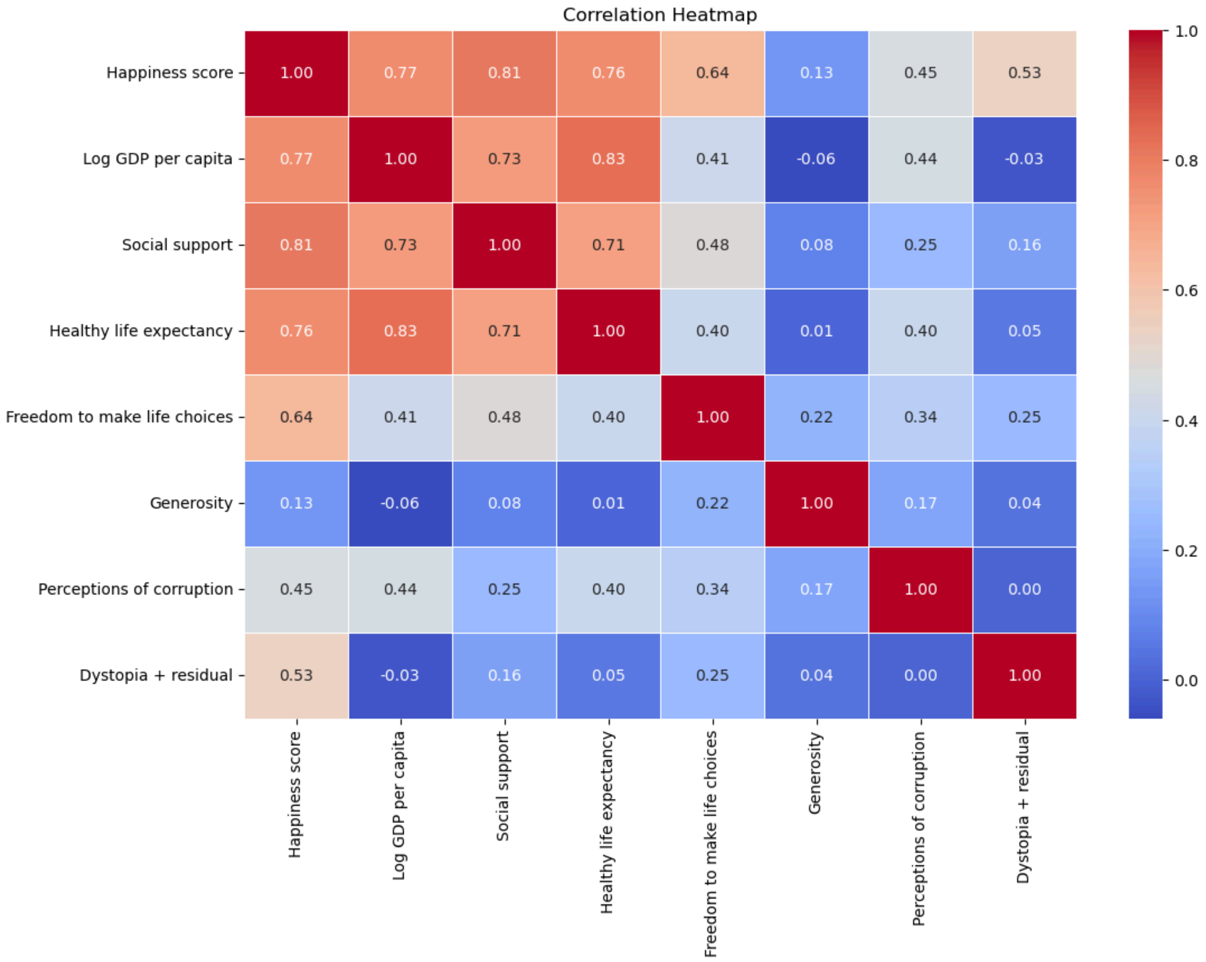
Correlation analysis measures the linear relationship between numerical variables. It helps identify how changes in one feature are associated with changes in another.

Implementation:
Step 1: Dropped all the unwanted data for analysis

```
# Drop Country name, upperwhisker, lowerwhisker and Regional indicators for correlation analysis
data_encoded = happy_data.drop(columns=['Country name', "upperwhisker", "lowerwhisker", 'Regional indicator'])
```

Step 2: Computed the correlation matrix and plotted a heatmap for the result.

```
# Compute correlation matrix
correlation_matrix = data_encoded.corr()
```



	Happiness score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
Happiness 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

Key features to note:

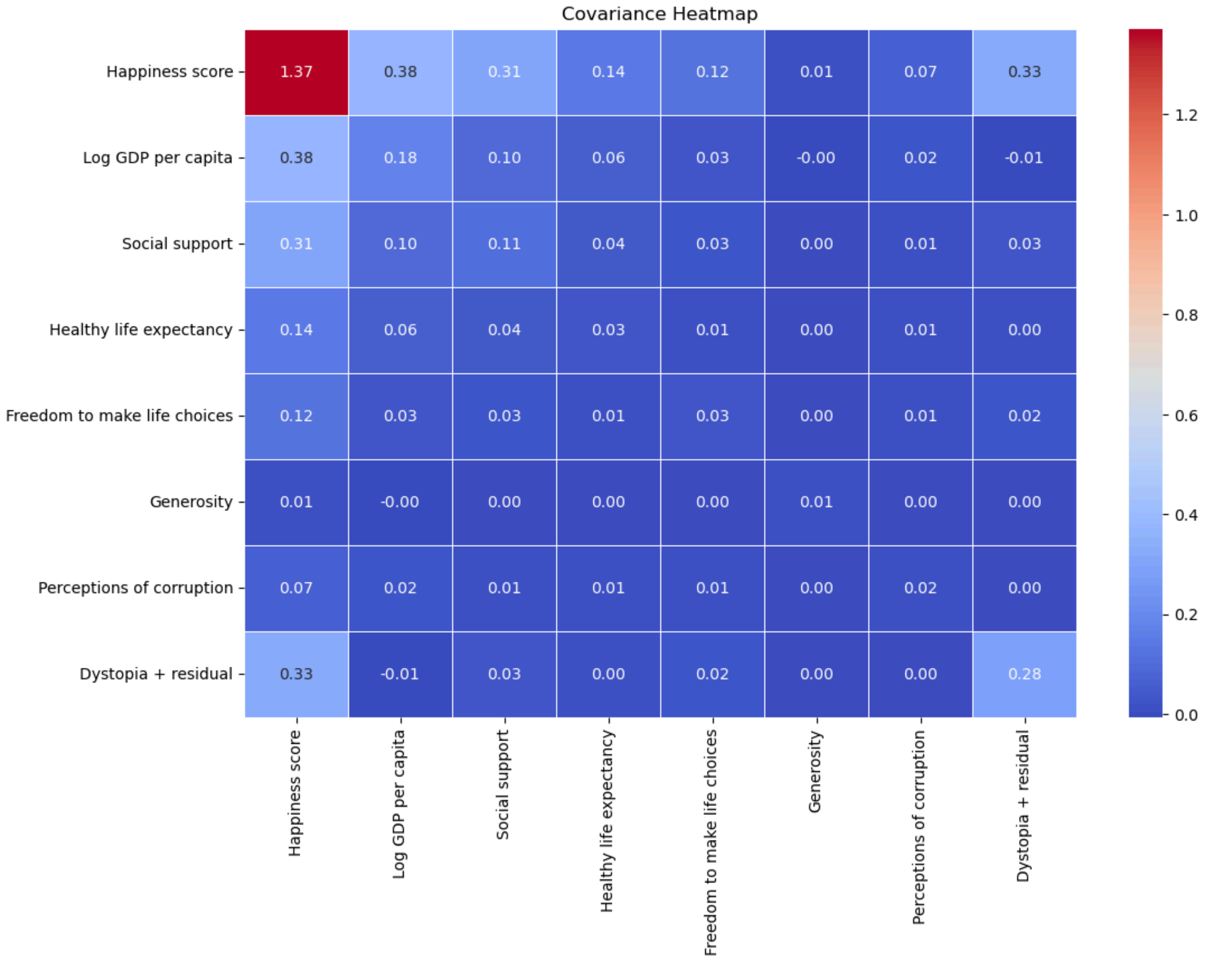
- Log GDP per capita, Social support, and Healthy life expectancy are the factors that have high correlation with Happiness scores.
- Perceptions of corruption and Dystopia + residual have only moderate correlation with the Happiness score.
- Generosity has a weak correlation with every other metrics.

3.3 Covariance Analysis

Covariance measures the directional relationship between two variables, showing how they vary together. Unlike correlation, it does not normalize values between -1 and 1.

Implementation:
Computed the covariance matrix and plotted a heatmap for the result.

```
# Compute the covariance matrix
covariance_matrix = data_encoded.cov()
```



	Happiness score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
Happiness score	1.37	0.38	0.31	0.14	0.12	0.01	0.07	0.33
Log GDP per capita	0.38	0.18	0.10	0.06	0.03	-0.00	0.02	-0.01
Social support	0.31	0.10	0.11	0.04	0.03	0.00	0.01	0.03
Healthy life expectancy	0.14	0.06	0.04	0.03	0.01	0.00	0.01	0.00
Freedom to make life choices	0.12	0.03	0.03	0.01	0.03	0.00	0.01	0.02
Generosity	0.01	-0.00	0.00	0.00	0.00	0.01	0.00	0.00
Perceptions of corruption	0.07	0.02	0.01	0.01	0.01	0.00	0.02	0.00
Dystopia + residual	0.33	-0.01	0.03	0.00	0.02	0.00	0.00	0.28

Based on the covariance analysis of the World Happiness Report data, we can see that economic factors (GDP) and social support have the strongest positive influence on happiness scores. The psychological factors like generosity and corruption perception show weaker correlations.

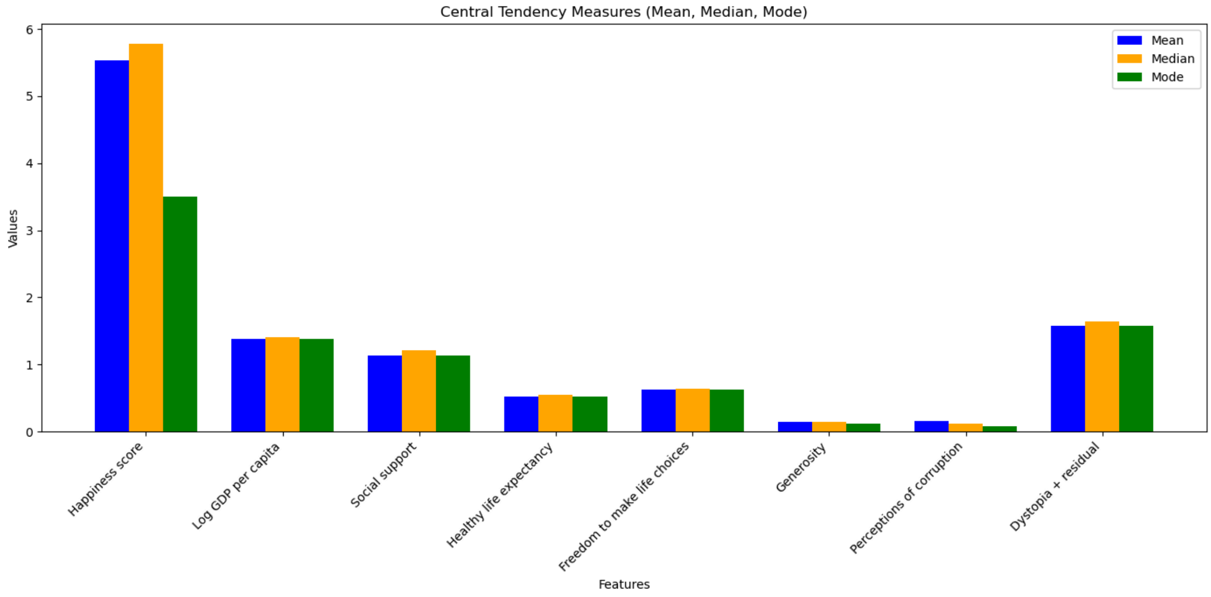
4. Statistical Analysis

4.1 Descriptive Statistics

Descriptive statistics provide a summary of the central tendency and variability of the dataset.

4.1.1 Central Tendency measures

The analysis focused on three key measures: mean, median, and mode, which were visualized using grouped bar graphs for better interpretation.



Objective

- To understand the central tendency (mean, median, mode) of numerical variables.
- To identify the spread and symmetry of the data.

Key Observations:

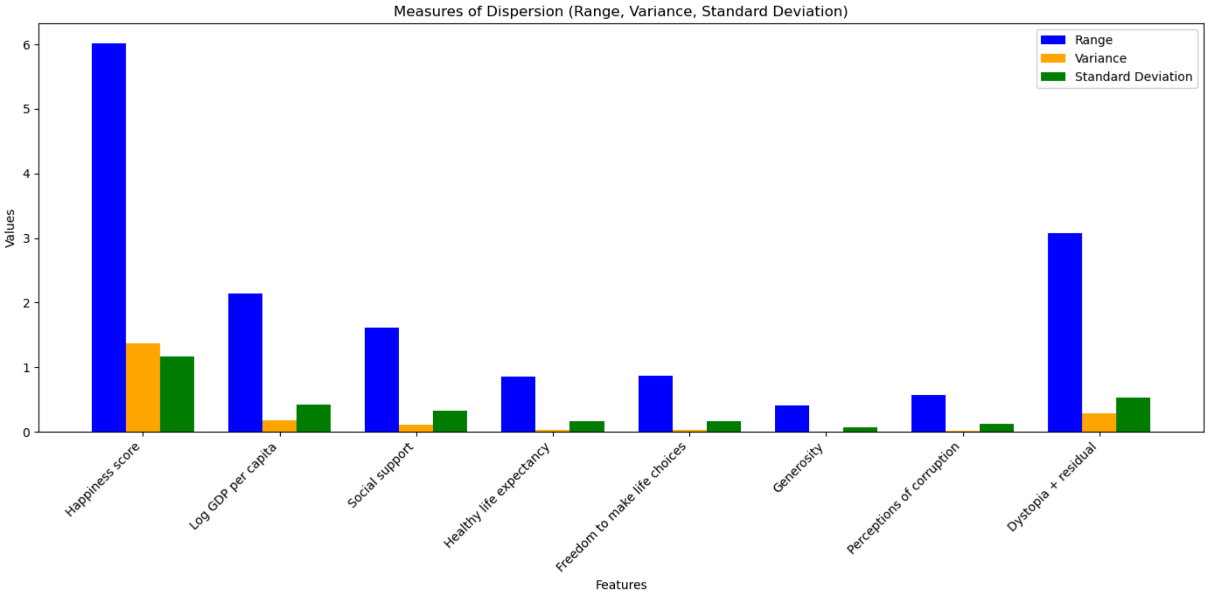
- Happiness Score exhibited the highest central tendency values.
- Generosity and Perception of Corruption had the lowest values.
- Most features displayed symmetric distributions.

4.1.2 Measures of Dispersion

Measures of dispersion provide insights into the spread or variability of the data. By analyzing metrics such as range, variance, and standard deviation, we can better understand the distribution and consistency of numerical features in the dataset.

Objective

- To measure how much the data deviates from the central tendencies.
- To identify variables with high variability, which may impact the overall analysis.



Key Insights

- High variance was observed in Log GDP per Capita and Dystopia + Residual.
- Generosity had low variability, reflecting consistency.

4.2 Inferential Statistics

Importance of Inferential Statistics in Data Analysis

Inferential statistics plays a critical role in data analysis as it allows us to draw conclusions and make predictions about a population based on a sample of data. Unlike descriptive statistics, which focuses on summarizing and visualizing data, inferential statistics helps us understand relationships, test hypotheses, and generalize findings to larger groups. It enables data analysts to assess the significance of observed patterns, evaluate the reliability of results, and make data-driven decisions. By employing techniques like hypothesis testing, confidence intervals, and regression analysis, inferential statistics provides a foundation for making sound inferences, identifying trends, and deriving actionable insights in real-world scenarios. This is particularly important in datasets like the World Happiness Report, where understanding the impact of key factors on happiness scores across different regions and populations is essential for meaningful conclusions.

4.3 Hypothesis Testing

Importance of Hypothesis Testing in Data Analysis
Hypothesis testing is a critical statistical method used to make data-driven decisions and draw reliable conclusions about a population based on sample data. It helps validate assumptions, determine the significance of observed patterns, and assess relationships between variables. By providing a structured framework to test claims, hypothesis testing minimizes the risk of errors and ensures objectivity in analysis. In projects like the Happiness Index 2024, hypothesis testing allows us to evaluate the influence of key factors like GDP, social support, and health on happiness scores, enabling a deeper understanding of global well-being trends.

1. T-Test:

Problem statement: The mean happiness score of a sample significantly differs from the population mean.

Null Hypothesis: The sample mean is equal to the population mean.

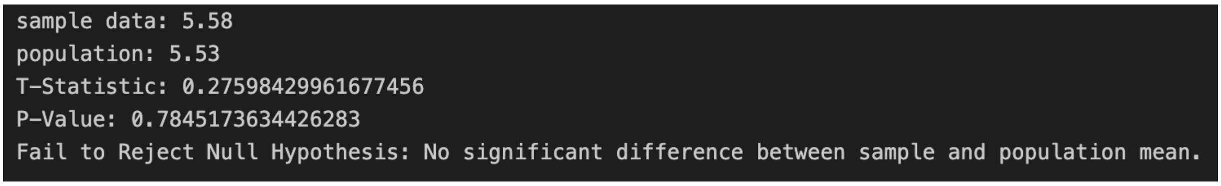
Alternating Hypothesis: The sample mean is different from the population mean.

Choose the Significant Level(α) to be: 5%

Implementation:

```python
from scipy.stats import ttest_1samp

# Random sampling
sample_data = data_encoded['Happiness score'].sample(30, random_state=42)
population_mean = data_encoded['Happiness score'].mean()

print(f"sample data: {sample_data.mean():.2f}")
print(f"population: {population_mean:.2f}")


t_stat, p_value = ttest_1samp(sample_data, population_mean)

# Results
print("T-Statistic:", t_stat)
print("P-Value:", p_value)


if p_value < 0.05:
    print("Reject Null Hypothesis: The sample mean is significantly different from the population mean.")
else:
    print("Fail to Reject Null Hypothesis: No significant difference between sample and population mean.")
```

Output:



```
sample data: 5.58
population: 5.53
T-Statistic: 0.27598429961677456
P-Value: 0.7845173634426283
Fail to Reject Null Hypothesis: No significant difference between sample and population mean.
```

2. Performing chi-squared test

Problem statement: Testing Association Between Log GDP per capita and Happiness Score Categories

NOTE: Discretize Log GDP per capita and Happiness score into categories using quantiles (e.g., high, medium, low).

Null Hypothesis: The two variables are independent.

Alternating Hypothesis: The two variables are not independent.

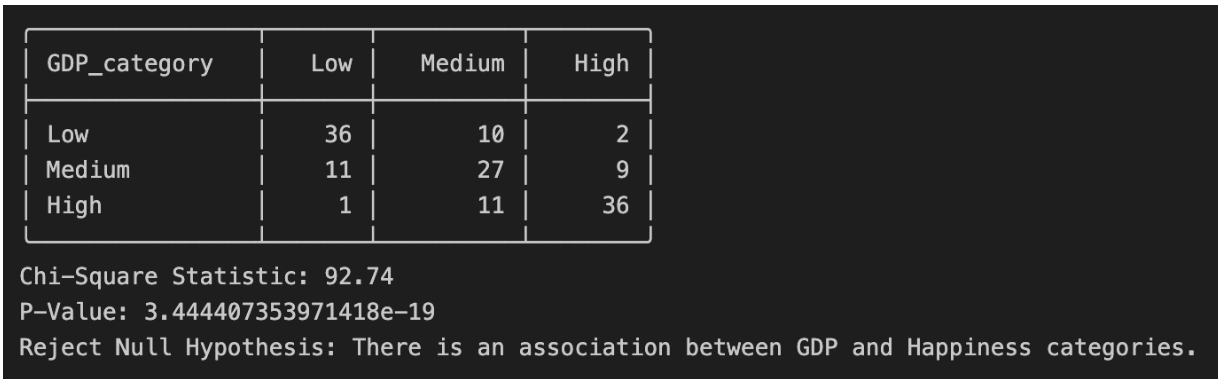
Choose the Significant Level(α) to be: 5%

Implementation:

```
from scipy.stats import chi2_contingency
import pandas as pd

# Discretize variables into categories
data_encoded['GDP_category'] = pd.qcut(data_encoded['Log GDP per capita'], 3, labels=['Low', 'Medium', 'High'])
data_encoded['Happiness_category'] = pd.qcut(data_encoded['Happiness score'], 3, labels=['Low', 'Medium', 'High'])

# Create a contingency table
contingency_table = pd.crosstab(data_encoded['GDP_category'], data_encoded['Happiness_category'])

# Display table
print(tabulate(contingency_table, headers='keys', tablefmt='rounded_outline'))



# Perform chi-square test
chi2, p, dof, expected = chi2_contingency(contingency_table)
```

```
# Results
print("Chi-Square Statistic:", round(chi2, 2))
print("P-Value:", p)
if p < 0.05:
    print("Reject Null Hypothesis: There is an association between GDP and Happiness categories.")
else:
    print("Fail to Reject Null Hypothesis: No association between GDP and Happiness categories.")
```

Output:



```
╒════════════════╤═══════╤══════════╤════════╕
│ GDP_category   │   Low │   Medium │   High │
╞════════════════╪═══════╪══════════╪════════╡
│ Low            │    36 │       10 │      2 │
│ Medium         │    11 │       27 │      9 │
│ High           │     1 │       11 │     36 │
╘════════════════╧═══════╧══════════╧════════╛
Chi-Square Statistic: 92.74
P-Value: 3.444407353971418e-19
Reject Null Hypothesis: There is an association between GDP and Happiness categories.
```

3. Performing ANOVA test

Problem statement: Testing the Effect of Perceptions of corruption on Happiness Score

NOTE: Discretize Perceptions of corruption and Happiness score into categories using quantiles (e.g., high, medium, low).

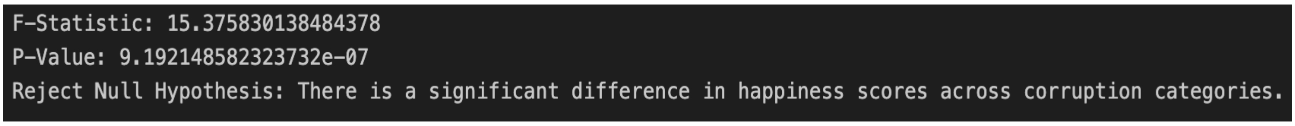
Null Hypothesis: The means of all groups are equal. No significant difference in happiness scores across corruption categories.

Alternating Hypothesis: At least one group mean is different. There is a significant difference in happiness scores across corruption categories.

Choose the Significant Level(α) to be: 5%

```
from scipy.stats import f_oneway

# Step 1: Discretize Perceptions of Corruption into categories
data_encoded['Corruption_category'] = pd.qcut(data_encoded['Perceptions of corruption'],
                                   3,
                                   labels=['Low Corruption', 'Medium Corruption', 'High Corruption'])

# Step 2: Create groups based on Corruption categories
low_corr = data_encoded[data_encoded['Corruption_category'] == 'Low Corruption']['Happiness score']
medium_corr = data_encoded[data_encoded['Corruption_category'] == 'Medium Corruption']['Happiness score']
high_corr = data_encoded[data_encoded['Corruption_category'] == 'High Corruption']['Happiness score']

# Step 3: Perform ANOVA test
f_stat, p_value = f_oneway(low_corr, medium_corr, high_corr)

# Step 4: Display results
print("F-Statistic:", f_stat)
print("P-Value:", p_value)
if p_value < 0.05:
    print("Reject Null Hypothesis: There is a significant difference in happiness scores across corruption categories.")
else:
    print("Fail to Reject Null Hypothesis: No significant difference in happiness scores across corruption categories.")
```

Output:



```
F-Statistic: 15.375830138484378
P-Value: 9.192148582323732e-07
Reject Null Hypothesis: There is a significant difference in happiness scores across corruption categories.
```

Key Findings from Statistical Tests

1. T-Test

- The mean happiness score of the sample (5.58) is not significantly different from the population mean (5.53).
- Conclusion: Fail to reject the null hypothesis. No significant difference between the sample and population mean.

2. Chi-Square Test

- There is a significant association between GDP per capita categories and Happiness Score categories.
- Conclusion: Reject the null hypothesis. GDP per capita and happiness are related.

3. ANOVA Test

- A significant difference in happiness scores is observed across different levels of perceptions of corruption.
- Conclusion: Reject the null hypothesis. Perceptions of corruption significantly affect happiness scores.

5. Feature Selection

5.1 Top Predictors of Happiness

1. Log GDP per Capita:

- Strongest correlation and covariance with Happiness Score; indicates economic stability as a key factor.
- Countries with higher GDP consistently rank higher on the happiness index.

2. Social Support:

- Countries with robust social systems report higher happiness.
- Better social support systems lead to happier citizens.

3. Healthy Life Expectancy:

- A direct link between health and happiness was evident.
- Life expectancy has a profound impact on happiness levels.

5.2 Key Insights

- **Economic Factors:** Countries with higher GDP consistently rank higher on the happiness index, highlighting the critical role of financial stability in happiness.
- **Social Factors:** Nations with stronger social support systems exhibit higher happiness levels, indicating the importance of community and interpersonal relationships.
- **Health Factors:** Higher life expectancy correlates with greater happiness, emphasizing the value of health and well-being in quality of life.
- **Psychological Factors:** While generosity and perceptions of corruption influence happiness, their impact is secondary compared to economic, social, and health factors.

6. Limitations

- **Limited Feature Assessment:** The analysis could be expanded to include more individual-level features, such as cultural or geographical variables, for a deeper understanding.
- **Interpretability Challenges:** Certain features, like Dystopia + Residual, lack clear interpretability, limiting their utility in driving actionable insights.
- **Redundant Factors:** Factors such as upper and lower whiskers in visualizations may be unnecessary and could be removed to improve focus and clarity.
- **Lack of Historical Data:** The absence of data from previous years prevents a comprehensive, longitudinal analysis and limits the ability to identify trends over time.

7. Conclusion

This project provided a comprehensive analysis of the factors influencing happiness in the World Happiness Index 2024, using data preprocessing, exploratory data analysis, and statistical techniques. Economic stability, measured by GDP per capita, emerged as the strongest predictor of happiness, followed by social support and healthy life expectancy, highlighting the importance of economic prosperity, community bonds, and healthcare. Psychological factors like perceptions of corruption and generosity also showed notable, though secondary, impacts. The analysis revealed regional disparities and significant relationships among variables, offering actionable insights for policymakers to prioritize economic equity, social cohesion, and healthcare improvements. While the findings were robust, limitations such as the lack of historical data and interpretability challenges for certain features indicate opportunities for further exploration. Overall, the project demonstrates the value of a holistic, data-driven approach to understanding and fostering happiness across nations.

8. Reference and Future Work

References:

- Dataset: Kaggle - World Happiness Report
  https://www.kaggle.com/datasets/jainaru/world-happiness-report-2024-yearly-updated/data

- Python Libraries: Pandas, NumPy, Matplotlib, Seaborn, SciKit-Learn.

Future Work:

- Time-Series Analysis: Expand the analysis by incorporating historical happiness data to identify trends and shifts over time.
- Individual-Level Analysis: Investigate individual-level data to uncover granular insights into the personal factors influencing happiness.
- Predictive Modeling: Leverage machine learning algorithms to build predictive models for happiness scores, enabling better identification of key contributing factors.
- Feature Engineering: Investigate additional variables or latent factors to improve the depth of insights and model performance.