جامعة مصر للمعلوماتية
EGYPT UNIVERSITY
OF INFORMATICS

Egypt University of Informatics
Computer and Information Systems

The World Happiness Score Data Analysis

Submitted by:
Abdelkareem Ahmed Abdelkareem Mohamed Soliman [22-101251]
Mohamed Mohamed Hussien Ahmed [22-101158]
Mohanad Abd El Fattah Samir Eladl [22-101235]
Hazem Refaat Mohamed
[21-101105]

Supervised by:
Dr. Mohamed Taher El Refaei PhD

Eng. Nadine El Saeed

8/16/2024

Introduction

This project aims to analyze the relationship between various factors and countries' overall happiness and well-being globally. Specifically, the study focuses on how economic performance (GDP per capita), social support, life expectancy, perceptions of freedom, trust in government, and generosity contribute to a country's happiness. Additionally, the research will explore whether these relationships vary across different regions of the world.

Research Question

1. **GDP and Life Expectancy:**
    - What is the nature of the relationship between a country's GDP and life expectancy?
    - Can a higher GDP predict a longer life expectancy in a country?
2. **GDP and Happiness Score:**
    - How does a country's GDP relate to the happiness scores of its citizens?
    - Can a higher GDP predict greater happiness among citizens?
3. **Happiness Score and Freedom**:
    - How do happiness scores vary with perceptions of freedom in different countries?
    - Are higher happiness scores associated with greater perceptions of freedom?

Hypothesis

Hypothesis 1:

Correlation between economy (GDP) and health life expectancy.

- **Question**: Are countries with higher GDP have a longer life expectancy?

Null Hypothesis (H_0):

There is no correlation between a country's Gross Domestic Product (GDP) and its health life expectancy. In other words, the GDP of a country does not have a significant impact on the life expectancy of its population.

Alternate Hypothesis (H_1):

There is a positive correlation between a country's Gross Domestic Product (GDP) and its health life expectancy. In other words, countries with higher GDP tend to have a longer life expectancy.

Hypothesis 2:

Correlation between a country's GDP and the happiness scores of its citizens.

- **Question**: Is higher GDP in a country is associated with higher happiness scores among its citizens?

Null Hypothesis (H_0):

There is no correlation between a country's Gross Domestic Product (GDP) and the happiness scores of its citizens. In other words, the GDP of a country does not have a significant impact on the happiness scores.

Alternate Hypothesis (H_1):

There is a positive correlation between a country's Gross Domestic Product (GDP) and the happiness scores of its citizens. In other words, countries with higher GDP tend to have higher happiness scores among their citizens.

Hypothesis 3:

Correlation between happiness scores and perceptions of freedom in a country.

- **Question**: Is higher happiness scores are positively correlated with greater perceptions of freedom in a country?

Null Hypothesis (H_0):

There is no correlation between happiness scores and perceptions of freedom in a country. In other words, the perceptions of freedom in a country do not have a significant impact on the happiness scores of its citizens.

Alternate Hypothesis (H_1):

There is a positive correlation between happiness scores and perceptions of freedom in a country. In other words, higher happiness scores are associated with greater perceptions of freedom in a country.

Population of Interest

Countries worldwide.

Sampling Method

This study uses a secondary dataset from the 2019 World Happiness Report, which includes data from 158 countries around the world. The original data collection involved various international surveys and official sources to measure factors such as GDP, life expectancy, and happiness scores across these countries.

Since the dataset is already compiled and published, the sampling method for this study can be considered as **convenience sampling** of the available data. It relies on pre-existing, publicly available data rather than a new, randomized sampling method.

Bias Identification

1. Selection Bias

- **Description**: Selection bias occurs when the sample is not representative of the population. In our case, the dataset includes 158 countries, but it may not cover all countries or regions equally. Some countries with less reliable data or those not surveyed may be missing, which could skew the results.
- **Impact**: This bias could lead to over- or underrepresentation of certain regions, potentially affecting the generalizability of your findings.

2. Reporting Bias

- **Description**: Reporting bias occurs when the data collected may not accurately reflect the true values. For example, self-reported measures of happiness or perceptions of freedom could be influenced by cultural factors, political environments, or social desirability.

- **Impact**: If people in certain countries are more likely to report higher or lower levels of happiness or freedom, it could distort the relationship between these variables and other factors like GDP.

3. Survivorship Bias

- **Description**: This bias occurs when only certain observations (such as countries or populations) are included because they have "survived" or remained relevant by the time of data collection. In this context, countries that have stable governance and reliable data collection practices are more likely to be included.
- **Impact**: Countries with less stable conditions might be excluded, leading to an overrepresentation of more developed or politically stable nations, which could skew the analysis of global trends.

4. Temporal Bias

- **Description**: Since the data is from 2019, there is a potential temporal bias, meaning that the relationships observed in the data may have changed over time. Economic conditions, public health, and social factors could have evolved, affecting the current relevance of the findings.
- **Impact**: Conclusions drawn from this dataset may not fully apply to the present day, especially if significant global events or trends have occurred since 2019.

5. Cultural Bias

- **Description**: Cultural bias arises when the data collection methods or interpretations are influenced by the cultural context of the researchers or respondents. Different cultures might have different perceptions of concepts like happiness or freedom.
- **Impact**: This could affect the validity of cross-country comparisons, as the same score might not mean the same thing in different cultural contexts.

Collected Data/Dataset

The dataset used in this analysis is derived from the **201 World Happiness Report**, which is publicly available on Kaggle. This dataset includes data from 158 countries, capturing various economic, social, and health-related indicators that contribute to the overall happiness and well-being of populations.

Source of the Data: The data was originally collected and published by the Sustainable Development Solutions Network (SDSN) under the auspices of the United Nations. The data compilation involved several international surveys and official statistics from reliable sources such as the World Bank, World Health Organization (WHO), and the Gallup World Poll.

Key Variables:

- **Country**: The name of each country included in the dataset.
- **Region**: The geographical region where the country is located.
- **Happiness Rank**: The ranking of the country based on its happiness score.
- **Happiness Score**: A composite score representing the overall happiness of the population, calculated from various factors.
- **Economy (GDP per Capita)**: A measure of the country's economic performance, calculated as GDP per capita.
- **Family**: A variable reflecting the strength of social support and family connections.
- **Health (Life Expectancy)**: The average life expectancy in each country, representing the overall health of the population.
- **Freedom**: A score representing the freedom to make life choices.
- **Trust (Government Corruption)**: A measure of the perceived trust in government and the level of corruption.
- **Generosity**: A measure of generosity and the willingness to help others.

Data Collection Method: The data was collected through a combination of surveys, statistical databases, and economic reports. The happiness score was derived from survey responses related to personal satisfaction and well-being, while economic and health indicators were sourced from global databases maintained by international organizations.

Data Structure: The dataset is organized into rows representing individual countries, with columns for each of the variables listed above. This structured format allows for straightforward analysis and comparison of the different factors influencing happiness across countries.

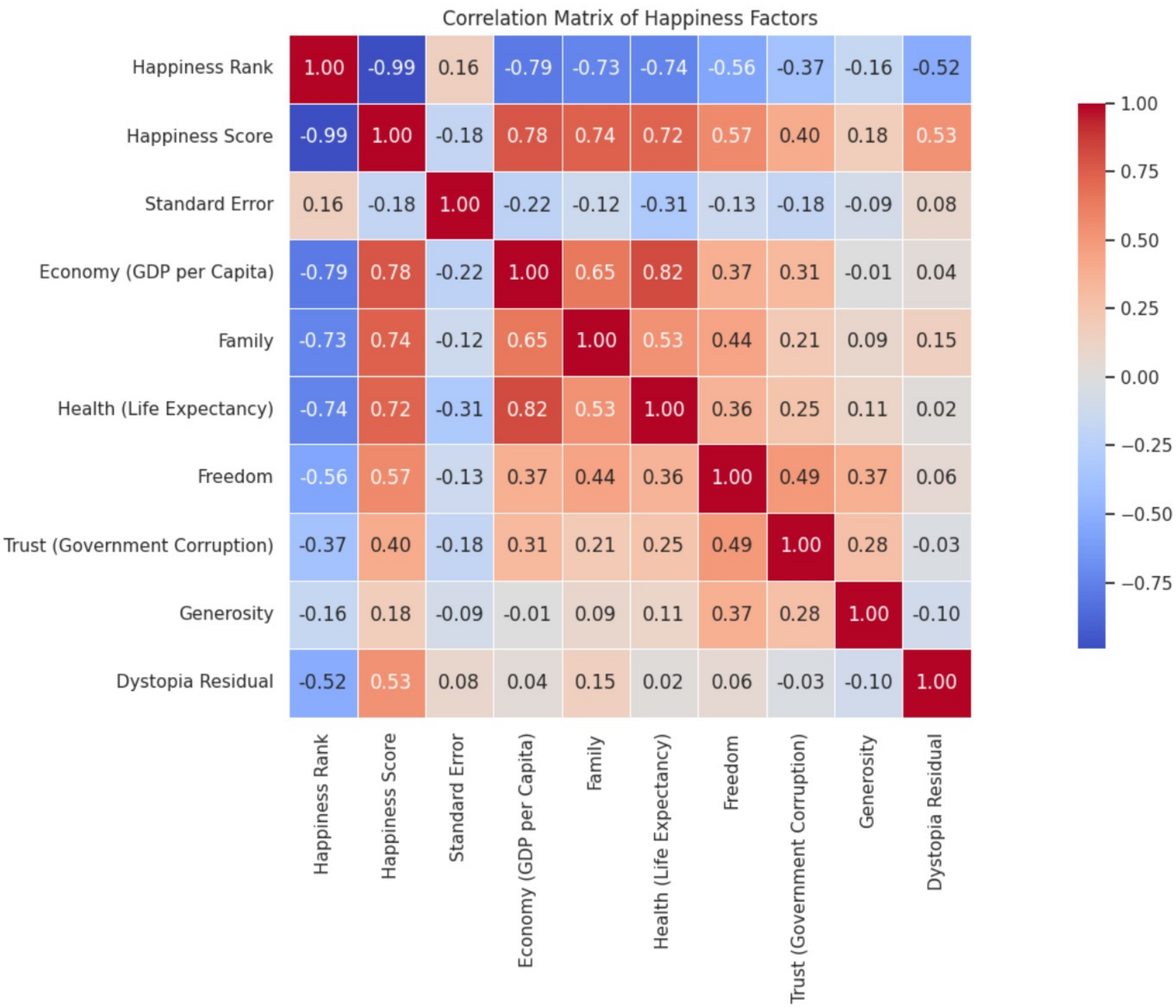
Improvements:

Null Data: We used a dataset that did not have any null data, ensuring that all records were complete and usable without the need for imputation or exclusion of samples.

Data Cleaning: We cleaned up the dataset by removing columns that were deemed unnecessary for our analysis. For example, we excluded the 'Dystopia Residual' column, which represented the follow-up period, as it was not directly relevant to our research questions and hypotheses.

Analysis:

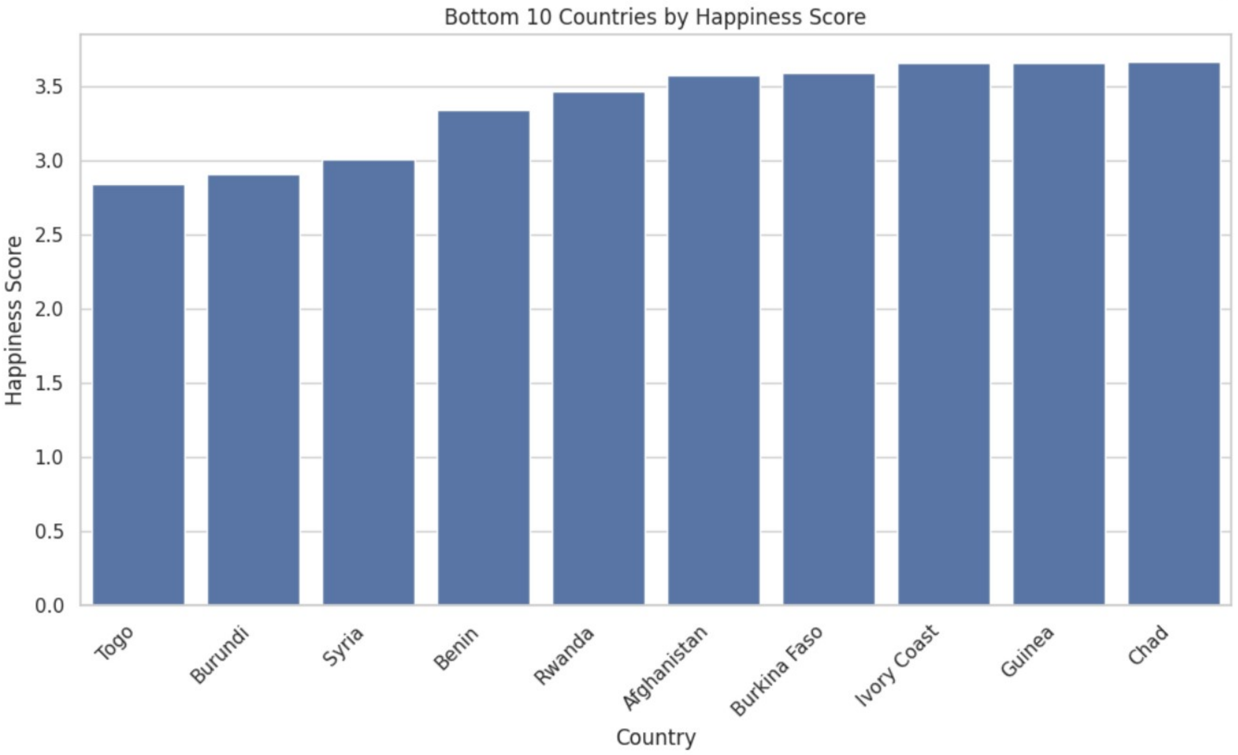
We wanted to study more about the sectors we have so we made some graphs to know visualize more the data.

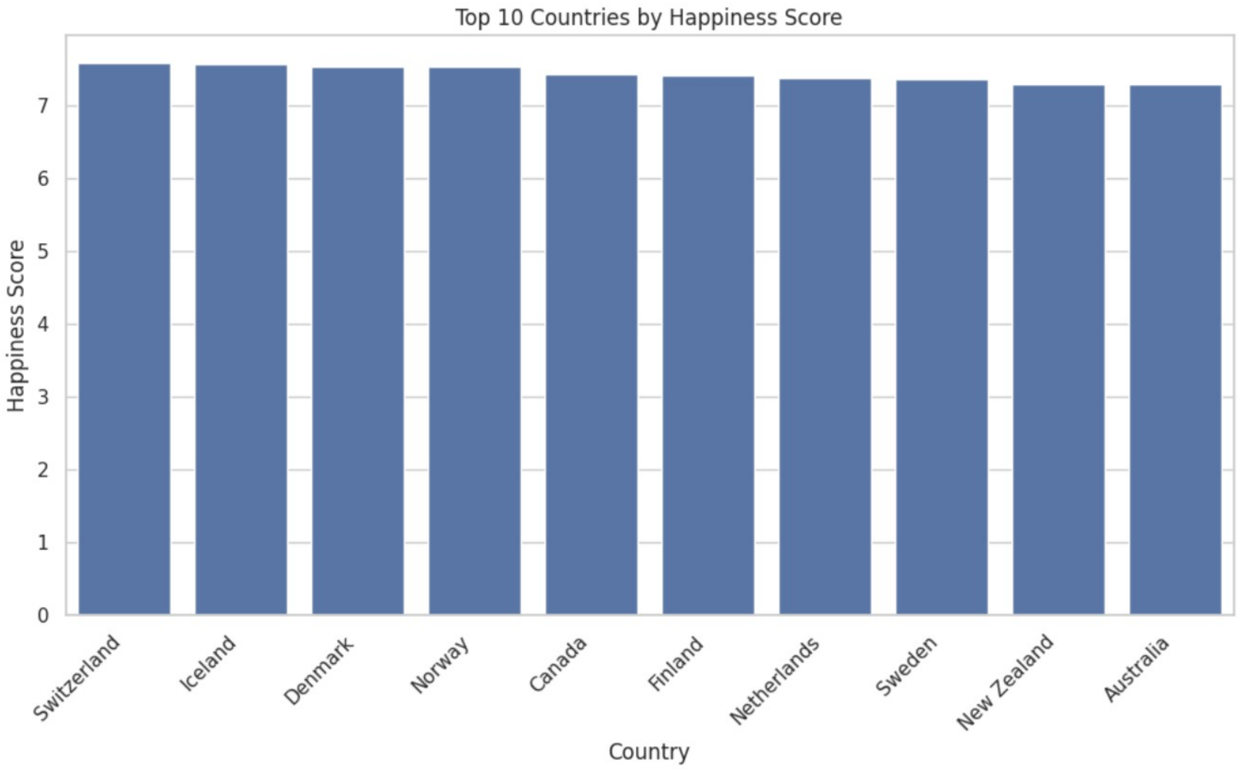


Correlation Matrix of Happiness Factors

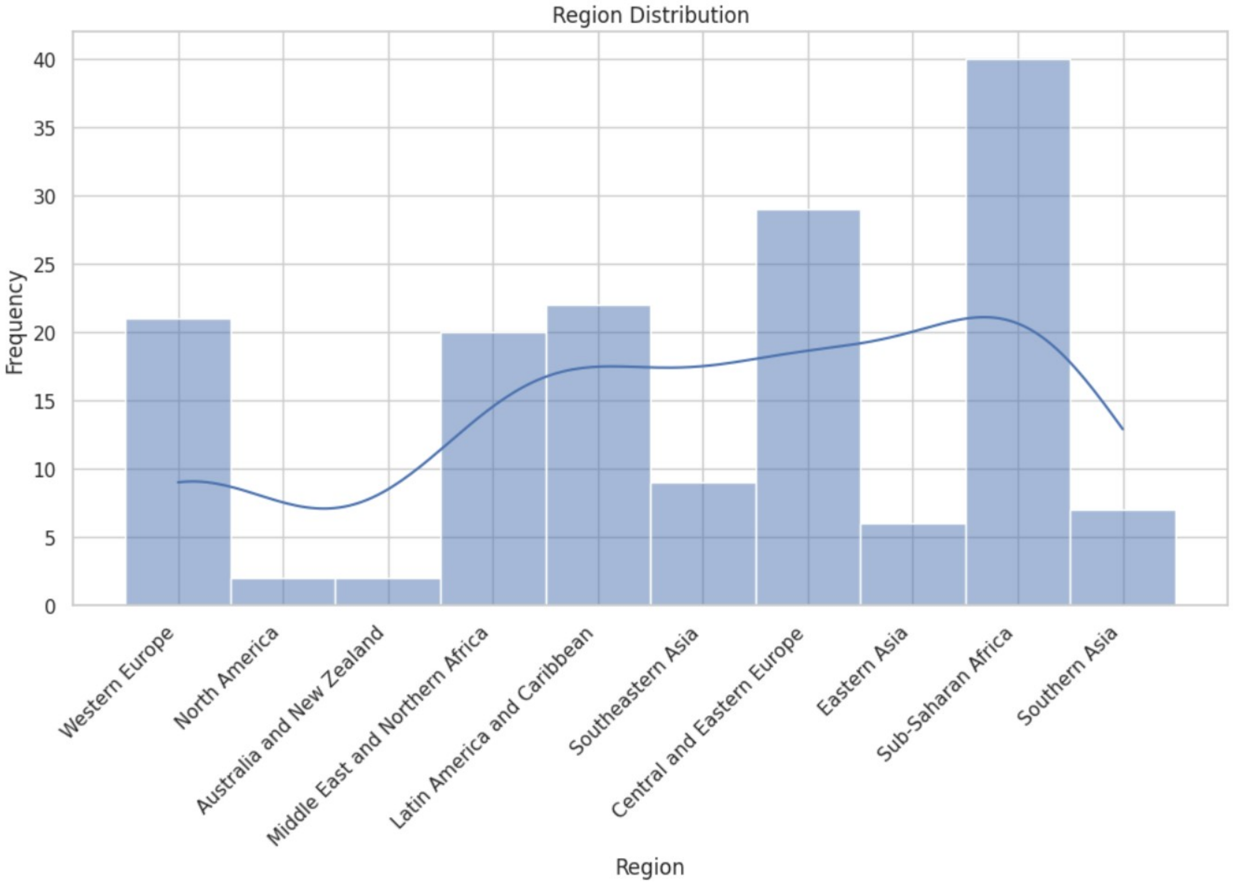
	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
Happiness Rank	1.00	-0.99	0.16	-0.79	-0.73	-0.74	-0.56	-0.37	-0.16	-0.52
Happiness Score	-0.99	1.00	-0.18	0.78	0.74	0.72	0.57	0.40	0.18	0.53
Standard Error	0.16	-0.18	1.00	-0.22	-0.12	-0.31	-0.13	-0.18	-0.09	0.08
Economy (GDP per Capita)	-0.79	0.78	-0.22	1.00	0.65	0.82	0.37	0.31	-0.01	0.04
Family	-0.73	0.74	-0.12	0.65	1.00	0.53	0.44	0.21	0.09	0.15
Health (Life Expectancy)	-0.74	0.72	-0.31	0.82	0.53	1.00	0.36	0.25	0.11	0.02
Freedom	-0.56	0.57	-0.13	0.37	0.44	0.36	1.00	0.49	0.37	0.06
Trust (Government Corruption)	-0.37	0.40	-0.18	0.31	0.21	0.25	0.49	1.00	0.28	-0.03
Generosity	-0.16	0.18	-0.09	-0.01	0.09	0.11	0.37	0.28	1.00	-0.10
Dystopia Residual	-0.52	0.53	0.08	0.04	0.15	0.02	0.06	-0.03	-0.10	1.00

The correlation matrix shows that GDP per capita is strongly positively correlated with both Happiness Score (0.78) and Life Expectancy (0.82). This suggests that higher GDP is associated with greater happiness and longer life expectancy etc.

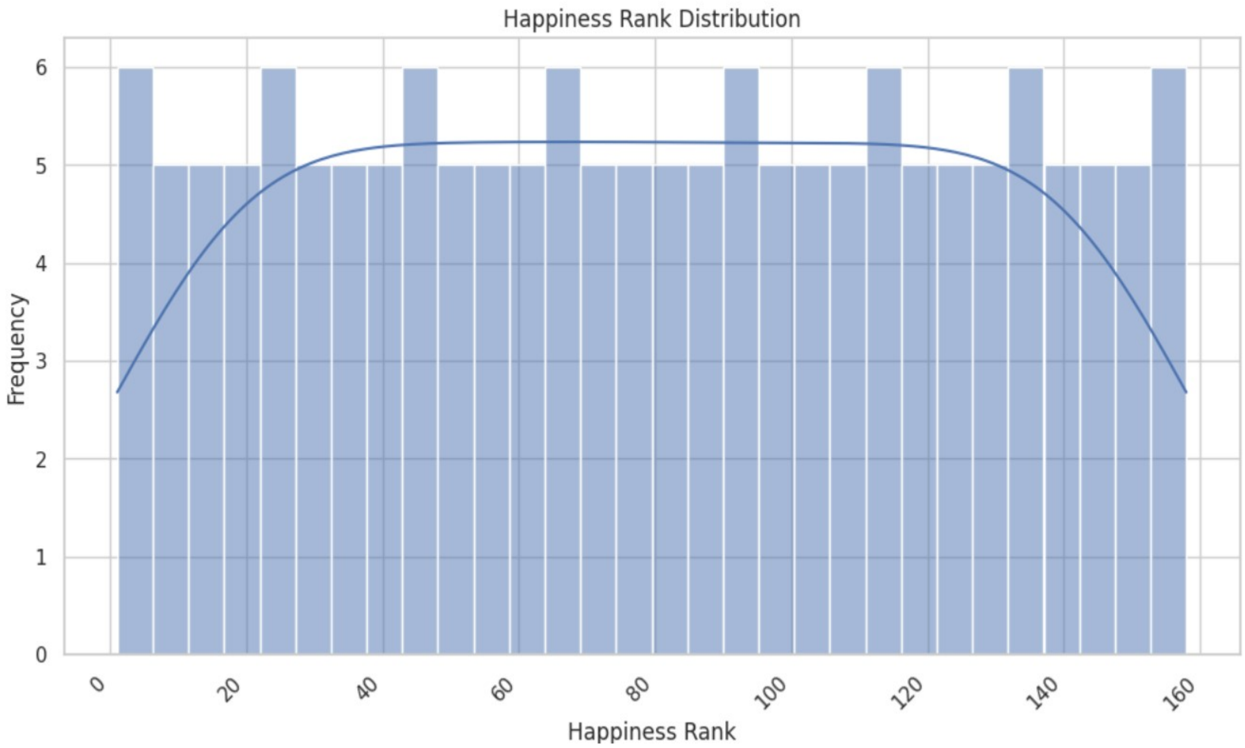




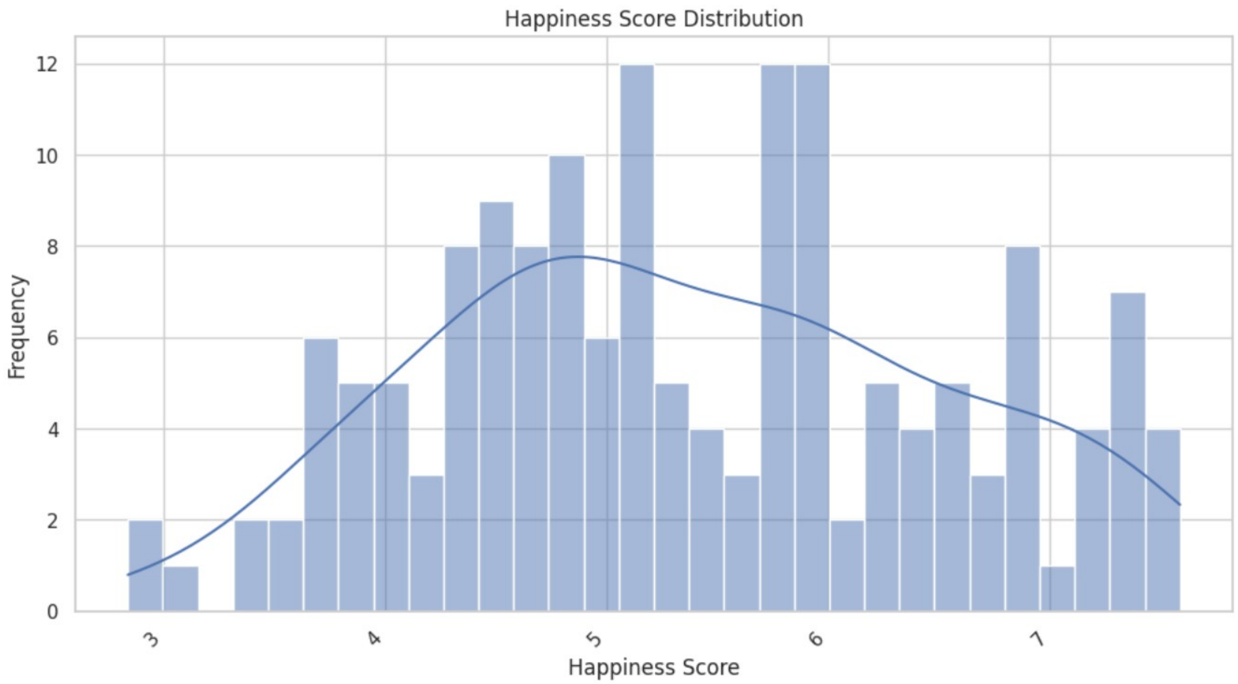
The two bar graphs display the top 10 and bottom 10 countries by happiness score. These graphs were created separately to maintain clarity, as including all 150+ countries in a single graph would have made it difficult to read. By focusing on the highest and lowest rankings, the graphs effectively highlight the contrast between the happiest and least happy countries.



The bar chart illustrates the distribution of frequency across different regions. The x-axis represents the regions while the y-axis indicates frequency. A trend line is overlaid on the bars, suggesting a potential pattern in the data.

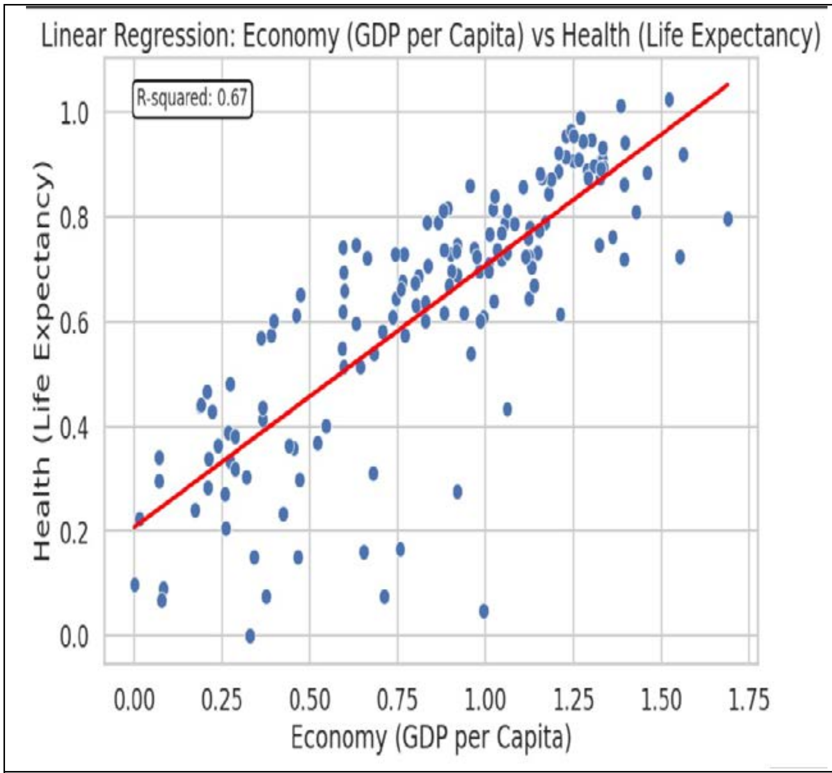
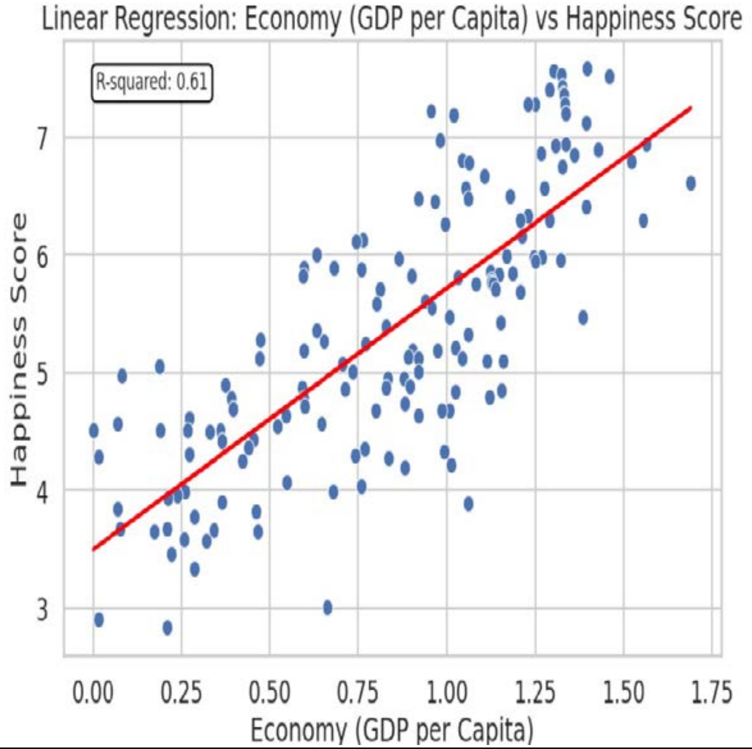


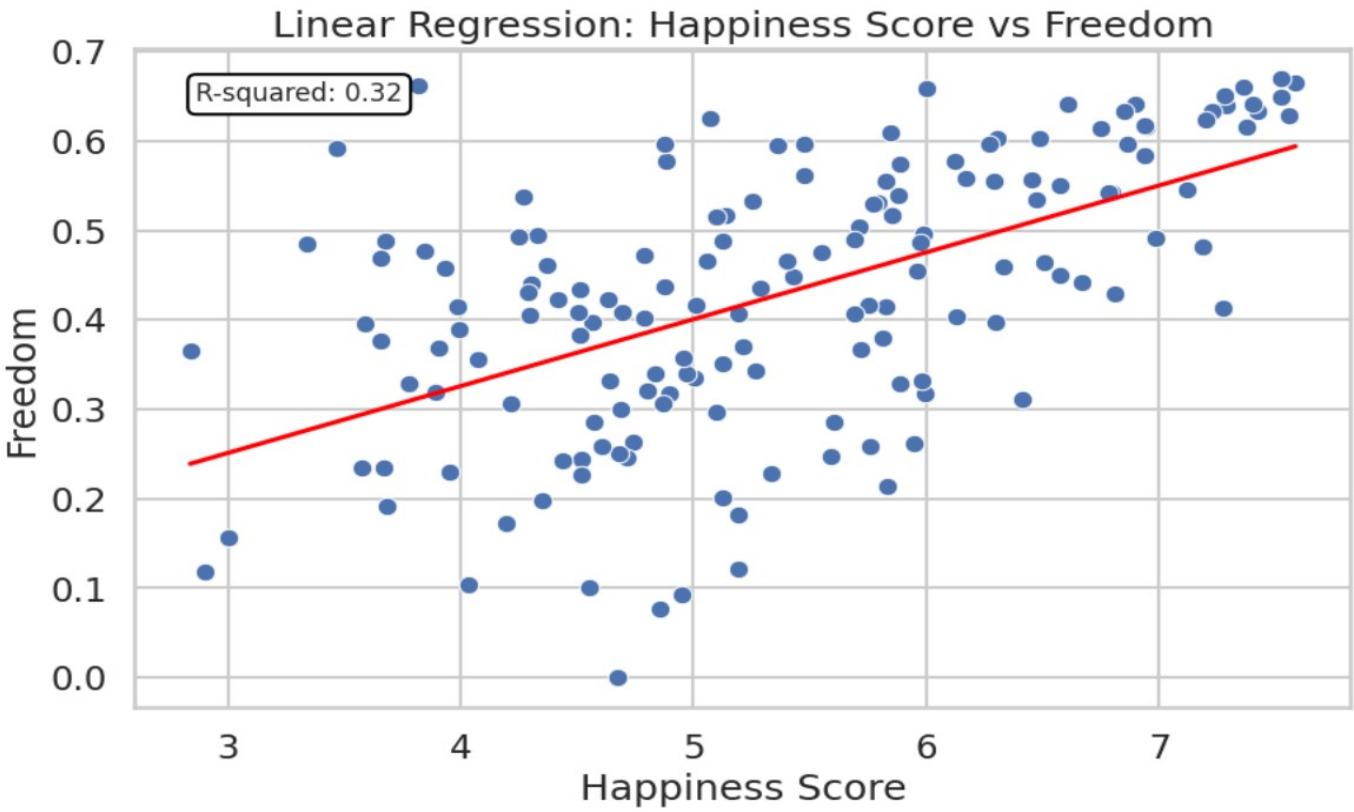
The image presents a histogram depicting the distribution of happiness ranks. The x-axis represents the happiness rank, ranging from 0 to 160, while the y-axis indicates the frequency of countries within each rank. The histogram displays a generally uniform distribution of happiness ranks, with some slight variations. A smooth curve is overlaid on the bars, potentially representing a probability density function.



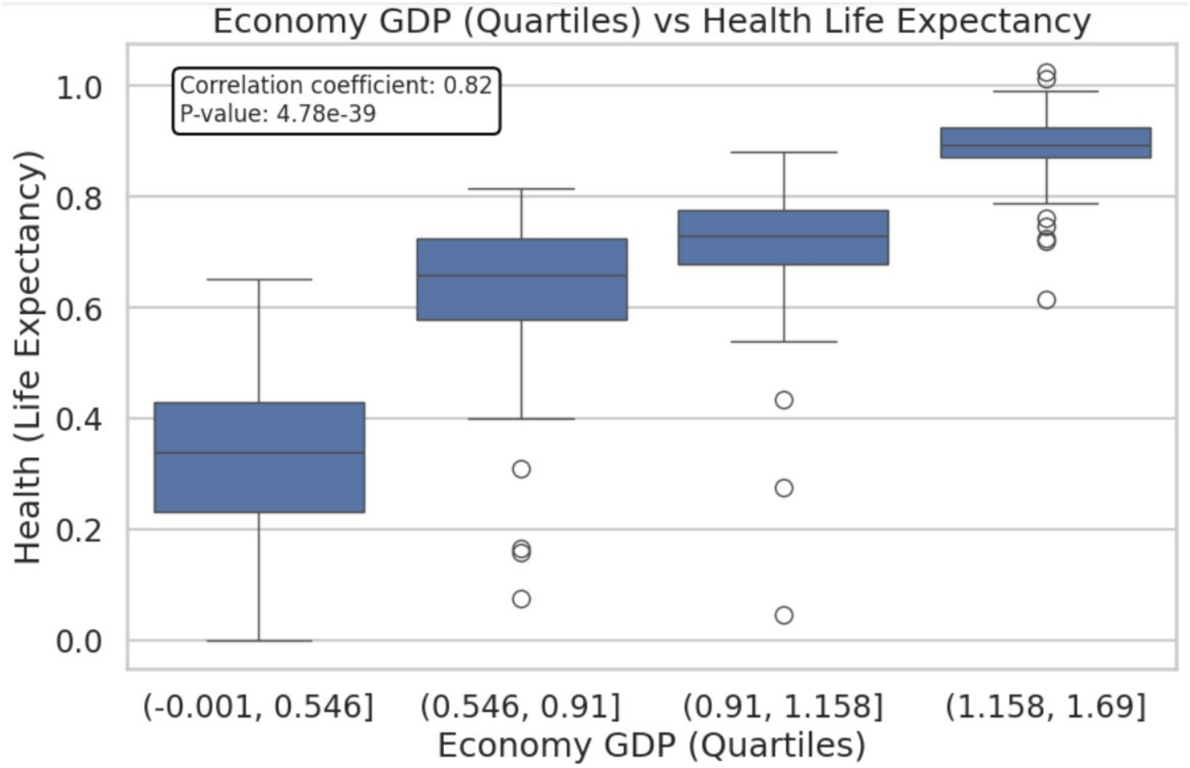
The image presents a histogram showing the distribution of happiness scores. The x-axis represents happiness scores, ranging from approximately 3 to 7.5, while the y-axis indicates the frequency of countries within each score range. The histogram exhibits a bell-shaped curve, suggesting a normal distribution of happiness scores. A superimposed smooth curve reinforces this observation.

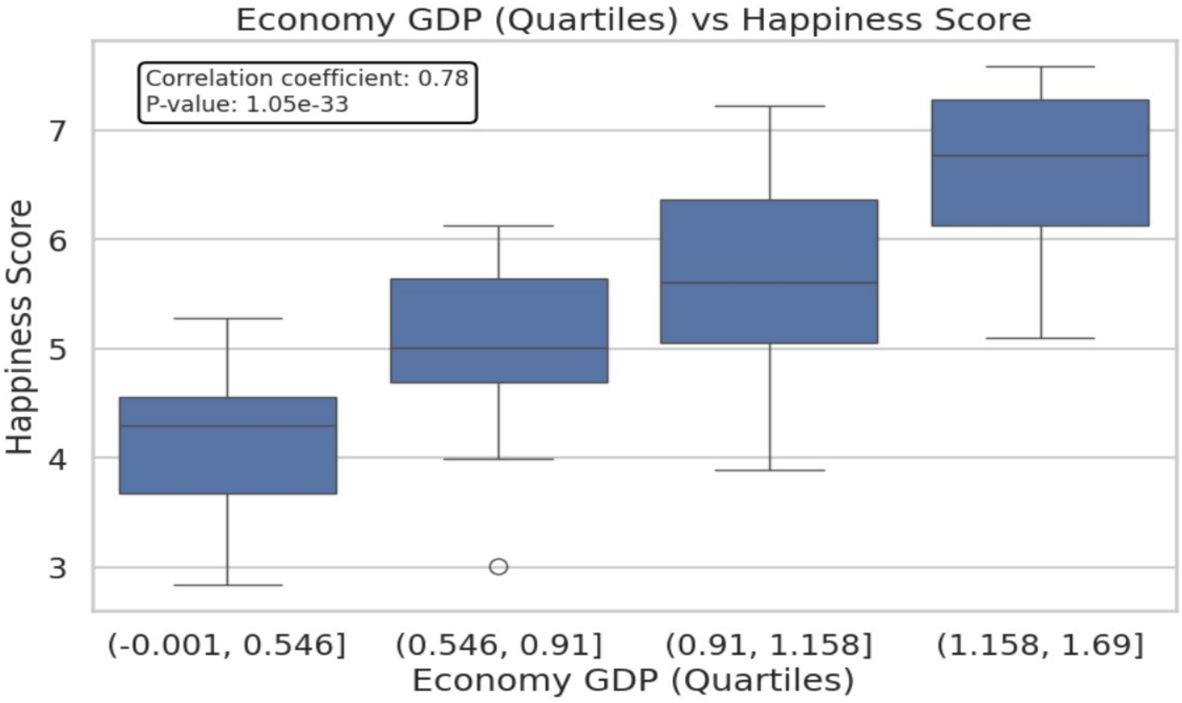
There are more graphs like this but to not make it long they won’t be included

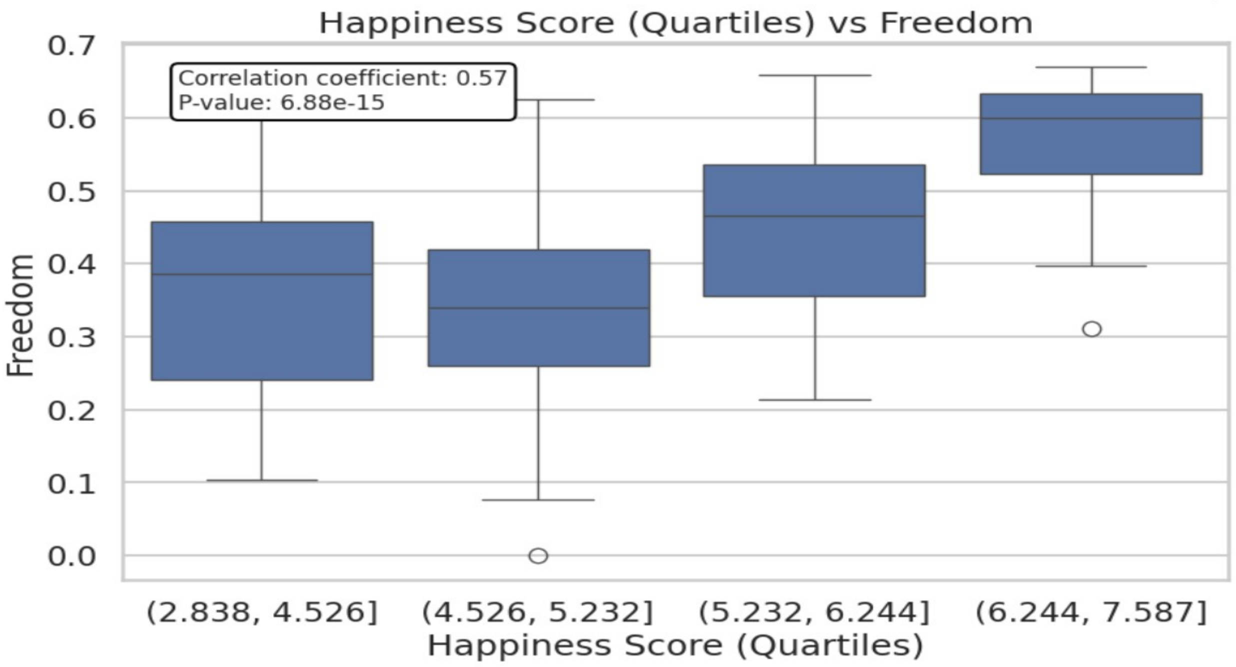
	
The scatter plot reveals a strong positive correlation between a country's economy (GDP per capita) and its citizens' life expectancy, suggesting that increased economic prosperity is linked to improved health outcomes.	The scatter plot illustrates a strong positive correlation between a country's economy (GDP per capita) and its happiness score. As GDP per capita increases, there's a general upward trend in happiness scores. The regression line further emphasizes this positive relationship.



The scatter plot illustrates a moderate positive correlation between Happiness Score and Freedom. While there's a general trend suggesting that higher happiness scores are associated with greater freedom, the relationship is not as strong as in some other correlations. This indicates that while freedom contributes to happiness, other factors also play significant roles.


Economy GDP (Quartiles) vs Health Life Expectancy
Correlation coefficient: 0.82
P-value: 4.78e-39
Health (Life Expectancy)
1.0
0.8
0.6
0.4
0.2
0.0
(-0.001, 0.546]
(0.546, 0.91]
(0.91, 1.158]
(1.158, 1.69]
Economy GDP (Quartiles)


Economy GDP (Quartiles) vs Happiness Score
Correlation coefficient: 0.78
P-value: 1.05e-33
Happiness Score
7
6
5
4
3
(-0.001, 0.546]
(0.546, 0.91]
(0.91, 1.158]
(1.158, 1.69]
Economy GDP (Quartiles)



Based on these last three graphs they contain the correlation coefficient and P-value for each hypothesis.

Interpretation of Correlation Scores and P-Values

Hypothesis 1: There is a positive correlation between a country's GDP per capita and its citizens' life expectancy.

Correlation Coefficient (0.82)

The correlation coefficient measures the strength and direction of the linear relationship between two variables. In this case, it's between Economy GDP (Quartiles) and Health (Life Expectancy).

- A value of 0.82 indicates a **strong positive correlation**. This means that as the Economy GDP increases (moving from lower to higher quartiles), the Health Life Expectancy also tends to increase.

P-value (4.78e-39)

The p-value is a statistical measure used to assess the likelihood of observing a result as extreme as the one obtained, assuming the null hypothesis is true.

- A p-value of 4.78e-39 is extremely small, which is much less than the typical significance level of 0.05. This means we can **reject the null hypothesis**.

- **In simpler terms:**
- There is a strong positive association between a country's economic development (as measured by GDP) and its population's life expectancy.
- The statistical analysis provides very strong evidence to support this relationship.
-

Hypothesis 2: There is a positive correlation between a country's GDP per capita and its citizens' happiness scores.

Correlation Coefficient: 0.78

This indicates a strong positive correlation between economy GDP (quartiles) and happiness score. A higher value implies that as the economic status increases (moving from lower to higher quartiles), the happiness score also tends to increase.

P-value: 1.05e-33 The extremely low p-value is significantly less than the standard alpha level of 0.05. This suggests that the observed correlation is highly unlikely to occur by chance. We can confidently **reject the null hypothesis** and conclude that there is a statistically significant positive relationship between economy GDP and happiness score.

In simpler terms:

- There is a strong positive association between a country's economic development (as measured by GDP quartiles) and its citizens' happiness.
- The statistical analysis provides overwhelming evidence to support this relationship.

Hypothesis 3: There is a positive correlation between perceptions of freedom and happiness scores within a country.

Correlation Coefficient: 0.57 This indicates a moderate to strong positive correlation between happiness score quartiles and freedom levels. A higher value implies that as happiness scores increase, freedom levels also tend to increase.

P-value: 6.88e-15 The extremely low p-value is significantly less than the standard alpha level of 0.05. This suggests that the observed correlation is highly unlikely to occur by chance. We can confidently **reject the null hypothesis** and conclude that there is a statistically significant positive relationship between happiness scores and freedom levels.

In simpler terms:

People who feel happier also tend to feel freer. This suggests that there's a connection between happiness and freedom.

Summary:

From our analysis, the correlation scores and p-values indicate the following:

- **Strong positive correlations** were found between:
  - GDP per capita and life expectancy o
    GDP per capita and happiness scores o
    Perceptions of freedom and happiness scores
- The **p-values** for all these correlations were **highly significant**, suggesting that these relationships are unlikely to be due to chance.
- These findings support the hypothesis that economic factors and perceived well-being are interconnected.

Conclusions

Hypothesis 1: There is a correlation between a country's GDP and life expectancy.

The analysis supports this hypothesis, showing a positive correlation between GDP and life expectancy. This indicates that countries with higher economic prosperity tend to have better health outcomes, as reflected in longer life expectancies.

Hypothesis 2: There is a correlation between a country's GDP and the happiness scores of its citizens.

The findings also confirm this hypothesis, demonstrating that GDP is strongly associated with higher happiness scores. This suggests that wealthier countries generally report higher levels of happiness among their citizens.

Hypothesis 3: There is a correlation between happiness scores and perceptions of freedom in different countries.

The analysis reveals a positive correlation between happiness and perceptions of freedom, though the strength of this relationship varies. This suggests that while freedom contributes to happiness, it is not the sole determinant.

Collectively, these results highlight that while GDP is a crucial driver of happiness and life expectancy, other factors like family support, health, freedom, trust in government, and generosity significantly influence the overall happiness of a country. Therefore, improving national happiness requires a comprehensive approach that goes beyond economic growth to include social, health, and governance factors.

Potential Issues

While the analysis provides valuable insights into the relationships between economic factors, health, and happiness, it is essential to acknowledge potential limitations. Firstly, correlation does not imply causation; although we observe strong correlations between variables, it cannot be definitively concluded that one variable directly causes changes in the other. There may be confounding factors or underlying mechanisms that influence the observed relationships. Secondly, the data used in this analysis may not be entirely representative of the global population, potentially limiting the generalizability of the findings. Additionally, the focus on GDP as a measure of economic development may overlook other important economic indicators that could influence the results.

Furthermore, the analysis relies on aggregated data, which masks variations within countries. A more granular analysis at the individual level could provide additional insights into the factors driving the observed relationships.