Project Proposal - World Happiness Report

Nimrah Ashfaq   AUIC-24FL-MSDS-9570

0.1   Section 1 - Introduction

Happiness is increasingly recognized as a vital measure of human well-being, often offering a more holistic perspective on quality of life than traditional economic metrics alone. The World Happiness Report spanning 2019 to 2024 compiles data from countries around the globe to assess how various factors, ranging from income and social support to health, governance, and freedom, contribute to overall national happiness levels. But beyond the commonly measured economic indicators, how do social and governance factors impact a nation's happiness? What other variables might influence how happiness is experienced and reported across different regions? One of the key columns happiness score serves as our primary measure of national well-being. Using this score as an indication of a country's overall happiness, general research question is:

What are the most significant economic, social, and governance predictors of national happiness levels, and how can statistical modeling be used to analyze regional variations and forecast future trends using World Happiness Report data from 2019 to 2024?

To address this question, statistical techniques to examine trends, model relationships between predictors and happiness scores will be applied. My analysis is driven by the following hypotheses:

H1: Higher GDP per capita, social support, and life expectancy are significantly associated with higher happiness scores across countries.

H2: Perceived freedom and generosity positively influence happiness, but their effect size differs by region - suggesting regional interaction effects.

H3: Countries with low perceived corruption tend to report higher happiness, and this relationship strengthens when combined with strong governance indicators (e.g., social support and life expectancy).

The dataset is derived from the **World Happiness Report**, an annual study based on data collected by the **Gallup World Poll**, supplemented with official statistics from sources such as the **World Bank, World Health Organization (WHO), and the United Nations**. The data is collected through large-scale surveys where individuals rate their overall life satisfaction on a scale of **0 to 10**. This dataset is derived from the **World Happiness Report**.

0.2   Section 2 - Data

The dataset has been placed in the `/data` folder for organization and reproducibility. A **README.md** file within this folder documents the dataset's dimensions and a codebook explaining each variable.

0.2.0.1   Dataset Overview   Below is an overview of the dataset using `glimpse()` to provide an initial summary.

```
## Rows: 875
## Columns: 13
## $ Year                      <int> 2024, 2023, 2022, 2021, 202~
## $ Rank                      <int> 1, 143, 137, 146, 150, 153,~
```

```
## $ Country.name                             <chr> "Finland", "Afghanistan", "~
## $ Ladder.score                             <dbl> 7.736, 1.721, 1.859, 2.404,~
## $ upperwhisker                             <dbl> 7.810, 1.775, 1.923, 2.469,~
## $ lowerwhisker                             <dbl> 7.662, 1.667, 1.795, 2.339,~
## $ Explained.by..Log.GDP.per.capita         <dbl> 1.749, 0.628, 0.645, 0.758,~
## $ Explained.by..Social.support             <dbl> 1.783, 0.000, 0.000, 0.000,~
## $ Explained.by..Healthy.life.expectancy    <dbl> 0.824, 0.242, 0.087, 0.289,~
## $ Explained.by..Freedom.to.make.life.choices <dbl> 0.986, 0.000, 0.000, 0.000,~
## $ Explained.by..Generosity                 <dbl> 0.110, 0.091, 0.093, 0.089,~
## $ Explained.by..Perceptions.of.corruption  <dbl> 0.502, 0.088, 0.059, 0.005,~
## $ Dystopia...residual                      <dbl> 1.782, 0.672, 0.976, 1.263,~
```

0.3 Section 3 - Data Analysis Plan

Response Variable (Y): - Happiness Score (Ladder Score) – Measures overall well-being on a scale of 0 to 10.

Explanatory (Predictor) Variables (X): - GDP per Capita - Social Support - Healthy Life Expectancy - Freedom to Choose - Perception of Corruption - Generosity

0.3.0.1 Comparison Groups

- **Regional Comparisons**: Comparing happiness scores across different continents
- **Economic-Based Comparisons**: Do wealthier countries report significantly higher happiness?
- **Governance & Corruption Comparisons**: Do people in countries with lower perceived corruption report higher happiness?
- **Health & Well-being Comparisons**: Do longer lifespans correlate with higher happiness?
- **Freedom & Rights-Based Comparisons**: Does personal freedom significantly impact happiness?
- **Social & Cultural Factors**: How crucial is social bonding for happiness?

## 0.4 Prelim EDA

Data Cleaning and Preprocessing

1. **Check for missing values**: All missing values were removed from the dataset to ensure the accuracy and reliability of the subsequent statistical analyses.
2. **Standardization**: To enhance readability and streamline the analysis process, column names are standardized using the `dplyr::rename()` function.
3. A new variable, Continent, is created using the countrycode package to classify each country into its corresponding continent. This allows for more effective aggregate and comparative analysis across global regions.

0.4.1 1. Summary Statistics

```
##   Summary_Statistics_Happiness              Value
## 1                      Mean    5.53524769585253
## 2                    Median                5.68
## 3        Standard Deviation    1.12436986004367
## 4                     Range       1.364 to 7.842
## 5                  Skewness   -0.351673785183297
```

0.4.1.0.1 Interpretation: *The mean and median of happiness scores indicates that most scores are close to the middle of the scale. The standard deviation suggests moderate variability in the data. The range of happiness scores reflects a broad distribution of values. The negative skewness indicates that, while there are some countries with low happiness scores, the majority have scores closer to the higher end of the scale.*

0.4.2 2. Happiness Score Distribution
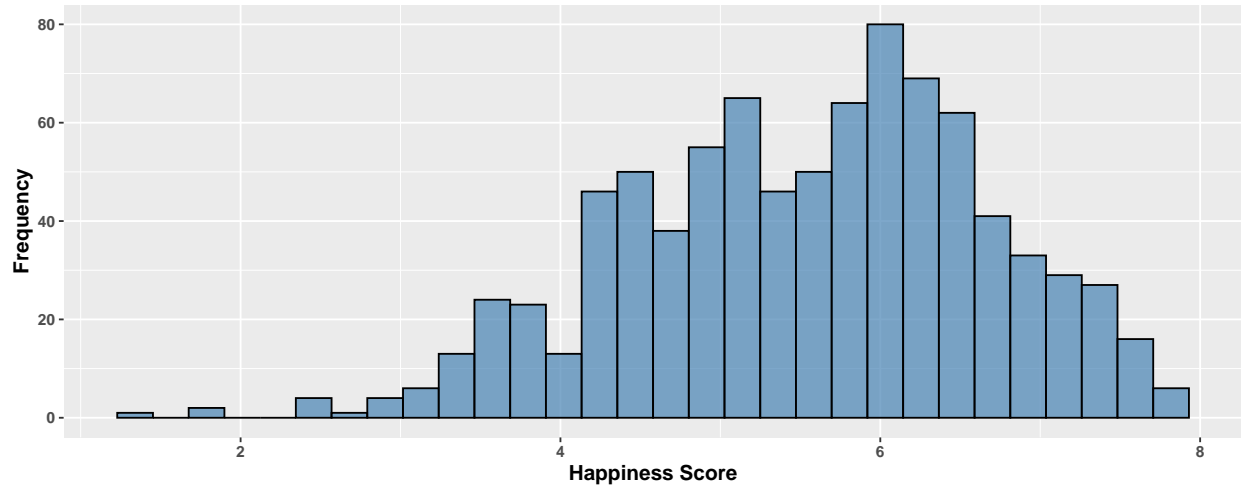
0.4.2.0.1 Interpretation: *The histogram displays the distribution of happiness scores across countries. The data shows a left-skewed distribution with most countries having happiness scores between 5 and 7, which aligns with the fact that the mean and median values are relatively high. The distribution also reveals a long tail towards the lower end, indicating a smaller number of countries with significantly lower happiness scores.*

0.4.3 3. GDP per Capita Distribution
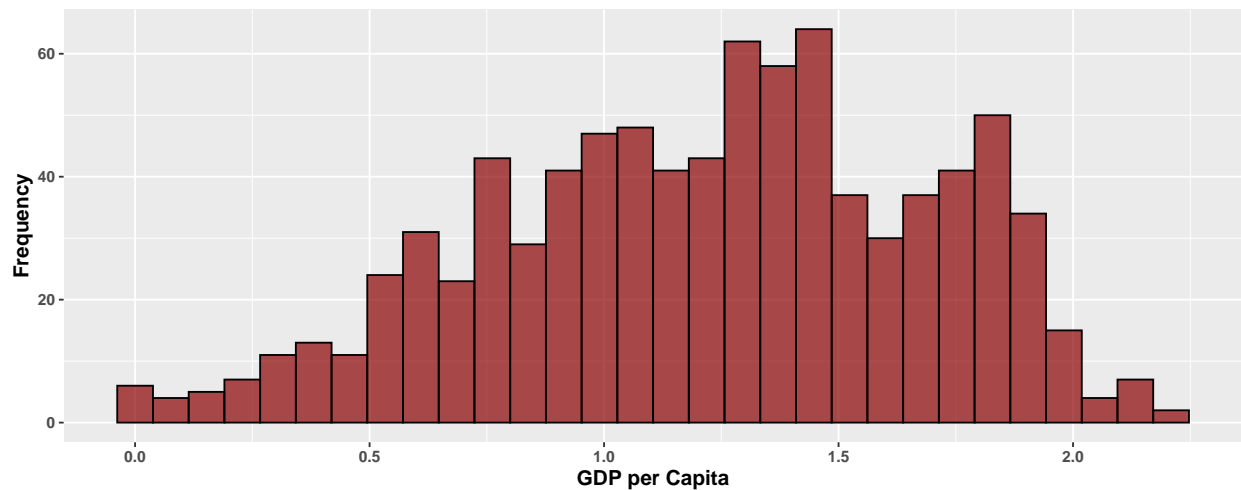
0.4.3.0.1 Interpretation: *The histogram of GDP per Capita shows a right-skewed distribution, with most countries clustered around a value of 1.5. This indicates that while many countries have moderate economic prosperity, a few wealthier nations contribute to the long tail on the right, highlighting global*

economic inequality. This distribution underscores the disparity between wealthier and less affluent nations, essential for understanding economic factors that may influence national well-being and happiness.

0.4.4   4. Bivariate Analysis of Factors Influencing National Happiness
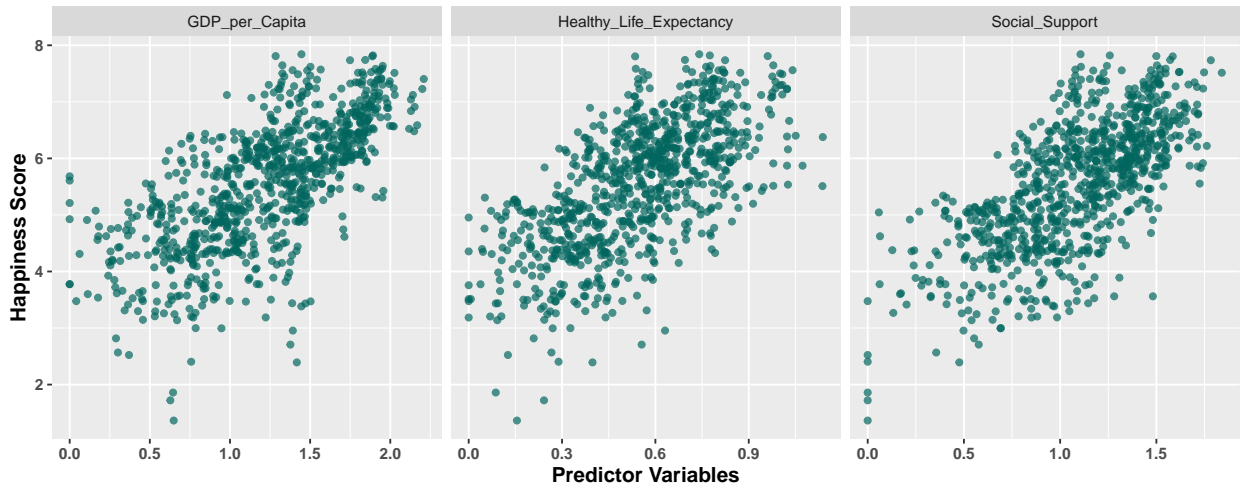
0.4.4.0.1   Interpretation: *Each of these factors - GDP per capita, social support, and healthy life expectancy - shows a positive correlation with happiness. This indicates that economic prosperity, strong social networks, and better health contribute significantly to higher levels of national happiness.*

0.4.5   5. Median Happiness Score by Region and Year
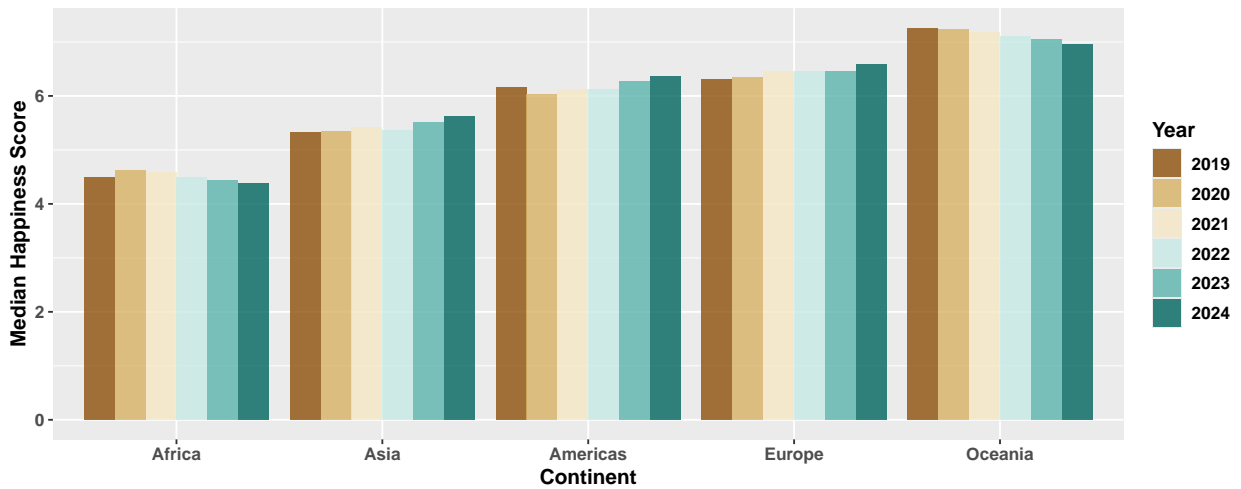
0.4.5.0.1   Interpretation: *The bar plot illustrates the median happiness scores by region from 2019 to 2024, showing a general upward trend across all regions. Europe and Oceania consistently report higher happiness scores, while Africa tends to have the lowest scores. Over time, all regions show improvement, with Americas and Asia falling in the middle range.*

0.5 Analysis Approach

Exploratory data analysis will be conducted to examine distributions, detect outliers, and summarize key variables. For H1, the association between happiness and GDP per Capita, Social Support, and Life Expectancy will be assessed using scatter plots and linear regression.

H2 will be explored by incorporating perceived Freedom and Generosity into the model, with interaction terms included to examine regional differences in effect sizes.

For H3, the relationship between Corruption Perception and Happiness will be analyzed, both independently and in interaction with governance-related variables such as Social Support and Life Expectancy, to assess compound effects.

0.6 Data Dictionary

The data dictionary can be found **here**.