World Happiness Report EDA

Jennifer Ruiz

DSC 530

Bellevue University

Introduction

To satisfy the term project requirements for this course, the author chose to perform exploratory

data analysis using the 2015 World Happiness Survey, available through Kaggle. It provides an

overall happiness score for each country represented in the survey, with 158 countries in total.

The study is conducted through a Gallup poll and is representative of the population in each

country. The happiness scores generated for each representative are calculated by aggregating all

the individual responses. The primary goal for the data analysis was to investigate the

relationship between the components and their collective influence on Happiness Score. The

following research questions were developed to support this analysis.

1. Are there any trends in Happiness rank based on Region?

2. What influence do Freedom and Trust have upon the Happiness Rank?

3. What impact does Health have on Happiness rank?

4. How does GDP impact the Happiness Score?

5. Which variables have the most significant impact upon the Happiness Score?

Outcome of EDA

Exploratory Data Analysis produced some interesting insights into the data. Countries in

Western Europe seem to have the highest happiness scores of those surveyed. Moreover, roughly

80% of the countries surveyed have a life expectancy of less than 80 years. Furthermore, the

region of the Middle East and North Africa seem to have the largest economy, as measured by

GDP. It is important to note that the region also has the lowest Happiness Rankings. Freedom

and Trust are both positively correlated to Happiness Score, but only at a mild to moderate level.

Multiple Regression analysis reveals that Economy, Family, and Health are the most significant

predictors of Happiness.

Missed Factors

After completing the analysis and reviewing the results, the author feels as though additional factors should have been included in the analysis to produce greater insight. Family was found to have a significant correlation (0.74) to the Happiness Score but was not further analyzed for insight. Furthermore, more analysis of the relationships between variables is needed to better understand the role they play in a country's Happiness Score.

Incorrect Assumptions

The author began the analysis with an assumption that the factors found to best predict Happiness Score would apply to the entire dataset. However, some paradoxical results were found during the analysis. These warrant further investigation and repeated testing for verification.

Challenges

The author faced a few significant challenges when performing the analysis. The largest was working with the probability mass function and properly interpreting the results. Additional moments of challenge were creating the Pareto distribution and understanding the information contained in the corresponding plot. Through additional research and reference back to the textbook, these challenges were easily overcome.