Project Proposal BRSM

2021114009 – Aditya Raghuvanshi
2021101057 - Pranit Khanna
2021114012 – Yash Bhaskar

Topic Selection & Datasets

World Happiness Report: (2012 – 2023)

The World Happiness Report provides a rich dataset on various factors related to happiness and well-being across countries. While the report analyzes key variables, there is an opportunity for further investigation into the relationships between happiness and other potential factors not directly included in the report. The aims of this study are to thoroughly analyze the report data, evaluate the robustness of the methodology, perform comparative analysis across years, and explore additional variables that may correlate with happiness index scores.

Going beyond this dataset, we explore two further datasets-

♦ Alcohol use (https://www.kaggle.com/datasets/marcospessotto/happiness-and-alcohol-consumption)
♦ Population (https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset)

Aim of the study

We want answers to questions such as the following, the methodologies of which are covered in a later section-

- Can we trust this data?
- How does each variable relate to the happiness index?
- Are the factors in the analysis robust which corresponds to verifying whether the dataset results in understandable curves and reasonable number and quality of outliers.
- Compare happiness data and key variables between pre-COVID and COVID pandemic years. Statistical tests will determine significant differences across years.
- To find a new variable which affects happiness, an attempt to go beyond the dataset or read in between the lines by finding some new variable or a combination of variables which affects the happiness outside the dataset, such as the following.
  - Alcohol use
  - Median age
  - population

Proposed Hypotheses

The World Happiness Report provides data on various factors related to well-being and happiness globally. The following hypotheses aim to investigate potential relationships and differences between happiness and other variables using statistical analysis of the report data across countries and years.

We have 1 major hypothesis with a couple of sub-hypothesis validating the major one, and 2 minor hypothesis.

- There is a significant relationship between a country's socio-economic factors or major events, and the happiness of its citizens:
    - Happiness has no relationship with Economy (GDP per capita).
    - **Sri Lankan** Recession had no significant effect on the happiness in Sri Lanka.
    - **United States** had no significant difference in average happiness scores during Biden's presidency as compared to Obama.
    - There is no significant difference in average life evaluation scores between countries before (2017-2019) and during (2020-21) the COVID-19 pandemic.
- Do the countries with higher alcohol consumption tend to be happier than others?
- There is no significant difference between the average 'Freedom to make life choices' rankings as measured at the 1st quartile versus that as measured at the 3rd quartile.

Logistic Regression

Using logistic regression, predicting next year's happiness scores for each country and from the weights of the features identifying which factor is not contributing to the countries happiness significantly and try to verify this using other statistical methods.

Identifying importance of various factors based on the weight obtained by the trained logistic regression model and then compare it with statistical models and correlational matrix.

Methodologies for analysis

1. First, we will do EDA (Exploratory Data Analysis) and answer questions like:
    a. Which country has the highest score? Why?
    b. Are there any null values? How does this dataset need to be cleaned?
    c. Is there any correlation between the features?
    d. Are the minimum and maximum happiness scores reasonable? Are there any outliers?
    e. What is the mean happiness score?
2. Assess assumptions and use various statistical methods to measure aspects such as:
    - Similarity of each variable to the normal distribution
    - Internal consistency

- Correlation
- Outlier detection
3. Use Statistical Methods such as:
    - Shapiro-Wilk test to check normality of prediction data.
    - Box plots to detect outliers in predictions.
    - Cronbach alpha
    - Pearson correlation coefficient
4. Visual inspection through graph plotting:
    - Histograms, world map charts, violin plot, line charts, pie charts, box plots, and other useful graphs also including interactive charts and maps (if time permits).
5. Finding new variables which affects happiness.

Expected Results, Predictions, Outcomes & Learnings

The overall expected outcome of this project is to derive comprehensive, data-driven insights into national-level factors correlated with or predictive of happiness globally across countries and years. Specifically, the key goals of the analysis include:

- Assess the validity and reliability of the happiness data through statistical tests of internal consistency
- Identify the socio-economic factors most correlated with happiness scores through regression and correlation analysis
- Compare average happiness levels and associated variables between pre-COVID and pandemic periods using hypothesis testing
- Uncover additional variables positively or negatively related to happiness, beyond those already included
- Profile groups of nations with highest versus lowest happiness using clustering analysis
- Model temporal trends and trajectories in happiness rankings using time series analysis
- Forecast future year happiness scores and rankings using ARIMA modeling.

Achieving these outcomes will provide a multidimensional, rigorous understanding into predictors of happiness at a national level. It will help distinguish the most relevant determinants of well-being and evaluate the data reliability. Predictive modeling can further validate the robustness.

The insights can inform government policies aimed at improving societal happiness through interventions on influential factors. Trend analysis will reveal effects of economic and social events. In sum, this study aims to explain, compare, predict, and ultimately provide recommendations for raising national happiness levels globally.