Project Proposal - Analysis, Visualization and Prediction on World Happiness Report Dataset

Lingyi Cai
lc3352@columbia.edu

Mengyu Ji
mj2857@columbia.edu

1 Goal and problem clarification

In this project, we use "The World Happiness Report [1]" as our dataset, which is a landmark survey of the state of global well-being. The goal of our project is to find out the most important factors for citizens to live happier lives and predict their happiness scores. This problem will be divided into two parts: visualization and prediction. The first part is to visualize relationship between happiness scores and various factors. Factors include [2]:

Economics (GDP per capita) indicates Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank in 2017.

Social support indicates the national average of the binary responses to "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

Life expectancy indicates the time series of healthy life expectancy at birth are constructed based on data from the World Health Organization (WHO) and WDI.

Perception of corruption indicates the average of binary answers to "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?"

Freedom to make life choices indicates the national average of binary responses to "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

Generosity indicates the residual of regressing the national average of responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.

The second part is a comparative study of the predictive model of the happiness score using this dataset.

2 Proposed approach

In the first part, our hypothesis is that there are relationships among happiness, economics, family, and etc [3]. We will use NumPy and Pandas libraries, which are two of the most widely used python libraries in data science, to analyze and visualize datasets.

In the second part, our experiment will base on independent variables which are family, economy, life expectancy, trust, freedom, and generosity to predict the happiness score. We will implement several machine learning models by using R programming language and find the model with the best prediction performance. Algorithms include multiple linear regression, decision tree regression, support vector regression, random forest regression, neural net and etc.

3 Evaluation plan

In data analysis and visualization, we need to consider several important factors: whether the visualized data can clearly reflect the trend of happiness score when a particular factor changes and how factors affect each other. For the prediction process, we will evaluate it by comparing the accuracy of different models and find the best model.

4 Timeline

Generally, we will propose new ideas and solve all the problems together. Our timeline is as below:

Date	Event
Mar 13 - Mar 20	Literature review
Mar 21 - Mar 24	Data pre-process
Mar 25 - Apr 8	Data analysis and visualization
Apr 9 - Apr 15	Happiness score prediction
Apr 15 - Apr 30	Report and presentation slides

References

[1] World happiness report dataset. https://www.kaggle.com/unsdsn/world-happiness.

[2] Jeffrey D. Richard Layard and John F. World happiness report 2018. Technical report, 2018.

[3] Peggy Schyns. Crossnational differences in happiness: Economic and cultural factors explored. *Social Indicators Research*, 43(1-2):3–26, 1998.