Proposal for Modeling World Happiness Data
Ross Brinkerhoff - Unit 7.2

For this project I will be analyzing data from the World Happiness Report.

Dataset:
https://www.kaggle.com/datasets/usamabuttar/world-happiness-report-2005-present

This report is based on a collection global survey data that measures the perceptions of the the citizens in different countries on a variety of data points. Some example include perceptions of corruption in government, freedom to control one life etc. All of these questions are combines with country GDP and life expectancy data to produce a broad picture of the situation in a particular country over time.

The key metric, and our dependent variable for this project, is the life ladder question. This question asks the respondent to rank their life as a whole from 1 to 10. 1 being the worst possible life and 10 being the best possible life.

The question I will be trying to answer here is how the responses on the various factors contribute to the final ladder question score. More specifically, I'll be trying to determine what factors, or collections of factors are most relevant in terms of explaining the variability in the ladder score. Building a model of these relationships will allow us to answer some interesting questions about the dynamics of citizen perception and how they relate to self reported well being.

For example, if perceptions of corruption were to increase in a particular country, how much change would we expect to see in terms of the ladder score. What if GDP drops, or perceptions of freedom become more positive? Answering these questions would enable us to better understand the general dynamics of reported happiness in a large population. If one or two factors explain most of the ladder score, additional questions might be introduced in these areas on future happiness surveys. The results, may also be of interest to people in leadership and governance, who would benefit from a better understanding of how perceptions of leadership actions might influence the happiness of constituents.

This project will be based on a single dataset from the world happiness report. This data set contains the citizen response data for multiple years and for most countries in the world. The data has also been processed already to some degree. It's not raw survey data, but consists of average response scores of the survey data. This means that the analysis will be high level, looking for large scale trends across countries and regions. Working with a single country over multiple years will be possible, but granularity beyond that will not be possible.

Other constraints are introduced by the fact that much of the data consists by design of self reported survey data, some of it gathered in countries that are unstable. Again, this means that any patterns found will be broad and general social patterns. It's also possible that some very important factors are simply not covered in the data we have.

After data cleaning and EDA, the overall approach to this project will be a regression model. We want to know how the various responses are related to the ladder score. To determine this, we will first run a PCA analysis on the data to see if any data features stand out in terms of explaining the ladder score. Based on those results, we will move forward with modeling using

various regression algorithms. The results of the algorithms will be compared and the best will be selected for the final model.

Because there are several years of data in the available data set, some portion of this dataset will be held back as test data. The selected model will be run on that data as a final test.