Project Proposal

The dataset we'll be using is the World Happiness Report (https://www.kaggle.com/unsdsn/world-happiness), a survey on of the state of global happiness (available datasets are 2015-2017). This dataset ranks 155 countries by their "happiness" levels, which are based on six factors: economic production, social support, life expectancy, freedom, absence of corruption, and generosity. These are all variables in this dataset that contribute to another attribute "Happiness score". We are also given the region a country is categorized under and a "Dystopia residual" comparing the country to a non-existent, hypothetical "dystopia". Due to the subjectivity of the variables in the current dataset, we hope to use other datasets from other sources to see how other world factors affect a country's "happiness score".

In order to gain more predictors than the country data provided in the World Happiness Report Data, we plan to use supplementary data sets. The first supplementary dataset is Global Temperature Data (https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data/data).
This dataset (1750 - 2018) records Average Temperatures on around 140+ countries. More specifically for each country/region on each data there are Land Min, Max and Average Temperatures as well as Ocean Min, Max and Average Temperatures. Another dataset are the Global Peace Agreements (1990-2016). This dataset from the University of Edinburgh tracks ever major peace agreement undergone by a country. The variables for this dataset include the status of the agreement, the data of the agreement, and the other regions/countries involved.

In addition we plan to use financial data as further predictors for a countries "happiness score." The Huge Stock Market Dataset (https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs) contains the daily price and volume data for all US-based stocks and ETFs trading on the NYSE, NASDAQ, and NYSE MKT until November 2017. In addition, we will also utilize the Cryptocurrency Daily Market Price Dataset (2013 - 2018) (https://www.kaggle.com/jessevent/all-crypto-currencies/data). This dataset contains the daily max, min and spread in Market Price as well as the volume for each of the 13 major Cryptocoins.

Questions to be investigated: Our main question is "Can we develop an accurate model that predicts a countries happiness score utilizing predictors from demographic data, financial data, climate data and government/diplomatic data?" Once the data has been structured by country and then date, we can then answer subsequent questions between individual data segments, such as climate data or the cryptocurrency data, and their contribution to the prediction of a country's "happiness score."

Statistical learning tools to be used: Because we're interested specifically in how factors such as global finance, diplomatic agreements and local climate play roles in a country's "happiness" score, a continuous variable, we will likely be sticking to regression (linear regression, LDA, QDA) for the majority of this project - we could also use logistic regression or kNN to see if we can categorize a country as "happy" or "unhappy" based on how high the residuals are from "dystopia". Clustering countries together using PCA is also a possibility; however, our main goal is to see if there are correlations between certain major events or characteristics that impact whether the people in a country are "happy".

Data Sets

Introduction
Title Page
Shorten