STATS 415 Project Proposal

Names:
- Nicholas Szerman(Li's Lab)
- Nihar Bhingaradia(Yujia's Lab)
- Mahir Taqbeem(Eli's Lab)
- Parth Aggarwal(Eli's Lab)

Selected Lab: Eli Schultz 8:30-10:00 am

Designated Speaker: Parth Aggarwal

Dataset:

We will be using the Quality of Government(QoG) Standard Data obtained from the Quality of Government Institute. The dataset is a collection of data from various sources. It contains nearly 2100 variables from 100 sources. The variables are grouped in categories such as education, infrastructure, energy, employment, health, religion, welfare.

https://qog.pol.gu.se/data/datadownloads/qogstandarddata
https://www.gapminder.org/data/

Questions:
- Can economic performance (i.e. GDP) be predicted from a small number of key conditions of the country (e.g. available agricultural land, economic liberalism index)?
- How is happiness related to sociological variables (e.g. fertility rate, political inclination)?
- What is a better predictor of happiness: a strong economy or a strong social structure?
- How the quality of government (e.g. corruption index) affects different indicators?

Methods we might use:
- Since we have a lot of variables, we can use Principal Component Analysis for dimensionality reduction. (Not quite sure how well that would work because we haven't learned PCA yet.)
- We can also use subset selection and shrinkage to simplify our models.
- We can do significance tests on linear models to find out which are the most significant predictors of economic performance.
- Use KNN to predict happiness using sociological and economic predictors.
- Build an LDA/QDA classifier to separate happy and unhappy countries.

Potential Issues:
- There are tons of variables, so we can find false correlations. Before running any analysis we must narrow down the predictors to use.
- There are only ~200 countries, so the size of the dataset is not very large.

- There might be a lot of variance in the data because of the diversity among countries.
- Since the data is collected from various sources, we might have to read through the documentation to make sure that the methodologies are consistent across all countries.