SSC442 Final Project Proposal

Project Ideas:

1. There's a bunch of public data surrounding schools so we could use that to see a correlation between wages and a whole lot of other data at a district or smaller level.

2. **Predicting unemployment in a few cities to capture what things have the largest effect on unemployment.**

3. Predicting demand and prices of renewable energy sources. Kaggle has a cool dataset for it form 2015-2018

4. Trying to determine how Right to Work laws (anti union laws) affect union strength. The data is easily obtainable from labor.gov

Our Plan

The final project will be an attempt to predict unemployment rates from different amenities provided by large cities in the USA. A linear model can help us to find some correlation between things like cost of living, indexes for rent, groceries, etc. as well as any environmental impacts. While we're still searching for a dataset that can help summarize community outreach programs we believe these variables will help to predict unemployment and we will have a large enough dataset to compare similar cities and find a degree of causality.

We plan to use the Bureau of Labor Statistics to find the data on unemployment for a select group of large cities and then gather data from other sources including Numbeo for cost of living indexing, a dataset from the EPA to gather information on the general health of the environment in each city and after stitching all this data together we will have a relatively strong set of data from which we will use to estimate unemployment. The analysis will be linear modeling formed from forward selection and hopefully will find some correlation between unemployment rates and the amenities a city offers. We expect primary contributors will be the community outreach programs and the cost of living indexes on the unemployment rate, but we include other variables as a way to hold cities 'constant' in another sense.

Five questions we are interested in:

1. What variables should be included to make the best linear model to predict unemployment rates?
2. What variables would heavily Bias our prediction?
3. What is the best visual tool to concisely and cleanly express what our data is telling us?
4. How well will our data predict unemployment, are we omitting important variables?
5. Chronologically how will a year variable capture the overall unemployment trend? Will it do better than more specific variables, or is it better to be precise?