It has been shown before that some areas have very low access to full grocery stores, but instead rely on convenience stores for food stuffs. I want to look at whether demographic data, including race, income, and education attainment, can be used to predict the density of these stores in a county. After creating these models, I want to look at which variables are the best predictors and how they differ between the two types of stores. The ratio of grocery to convenience stores will be the final model I want to make.

The main data source for this project is the United States Department of Agriculture's Food Access Research Atlas which contains county level data on the number of each type of store, population, racial makeup, and median income. To get a better idea of the distribution of income, I will use the Internal Revenue Service data on number of returns filed in each county in each of 8 income brackets. For the educational attainment, I will use the American Community Survey which gathers data in between the all inclusive 10-year census.

The first step to solving this problem is to join the data from the 3 sources into one data frame. After that, I will explore the data looking for variables that correlate to the densities of each store type and for any variables that correlate with each other. As an initial modelling step, I will fit the data with a linear regression. If this fits well enough, the coefficients can be used to compare how influential each variable is between the models. The second model I will use is (boosted?) random forest because it can typically fit the data better.

The final deliverables will be the code for data analysis and model creation as well as a slide deck discussing the problem and any interesting findings.