Grocery stores and convenience stores are not equally distributed in counties in the United States. Grocery stores typically sell a wide variety of foods and household goods such as fresh fruits, vegetables, and meat as well as canned and prepared foods. Convenience stores normally only carry a limited variety of prepared foods at a higher cost than grocery stores. The unequal distribution of these means that some areas will not have as easy access to fresh and/or cheaper foods available at grocery stores. In this project I will create a model using socioeconomic variables to predict the ratio of convenience to grocery stores in a county. A high value for this ratio likely means that it is more difficult to access a grocery store.

The first step in this project was identifying which variables to use from the three datasets: the United States Department of Agriculture's Food Access Research Atlas, the Internal Revenue Service's Statistics of Income, and the American Community Survey's 5-year estimates of educational attainment. From the Food Access Research Atlas, I used variables from 3 sheets: 'Supplemental Data - County,' 'STORES,' and 'SOCIOECONOMIC.' I used the Supplemental Data as the master list of county names and population. From stores I used 'GROCPTH14,' 'SUPERCPTH14,' and 'CONVSPTH14' which are the number of each type of store per thousand residents. I also created a column that is the ratio of convenience stores to the sum of grocery stores and supercenters. I used the entirety of the socioeconomic sheet. From the Statistics of Income, I used only the agi_stub and N1 which gives the number of returns filed in each of eight income brackets for each county. I had to pivot this data to make only one row per county. From the educational attainment data I used the columns

'HC02_EST_VC17', and 'HC02_EST_VC18' which are the percent of the population who have graduated from high school and have a bachelor's degree respectively.

There are three rows with multiple variables missing which were dropped. To look for outliers, I used boxplots. All variables have values beyond the whiskers, but most continue smoothly to the minimum/maximum. There is one variable that does appear to have an outlier. The column 'income_pct4' shows 1 county with 100% of the population in a single income bracket. This county has a population of 115 so it is possible that this is real data.

After the data was merged and cleaned, I began looking for correlations in the data. The variables with the largest correlation are trivial such as the percentage of non-hispanic white people and the percentage of non-hispanic black people. The non-trivial relationships which I identified are: increasing median household income as the percentage of high school graduates or percentage of people with bachelor's degrees increases and an increase in poverty rate as the percentage of non-hispanic black population increases.

I then investigated which variables are most strongly correlated to my target variable, the ratio of convenience to grocery stores. Those with the largest relationship are: percentage of population who have graduated high school (-0.25); percentage of population with a bachelor's degree (-0.24); percentage of non-hispanic black population (0.24); poverty rate (0.21); and median household income (-0.20). Since income is correlated with the first four, there appears to be a complex web of interactions determining my target variable.

In order to test whether the categorical variables affect the ratio of convenience stores to grocery stores, I used 10,000 bootstrap replicates of the mean for each subgroup to calculate a 95% confidence interval for the difference of means. Out of the four categorical variables, I found that three of them: counties with persistent poverty, persistent child poverty, and population loss have a statistically significant difference of means from zero. Only the variable signifying counties in a metropolitan area did not show a difference in means. I then used a t-test to find the p-value and found the same results.

Since one of my significant continuous variables deals with poverty and two of my categorical variables do as well, I investigated how they are related. As would be expected, the mean poverty rates in counties with persistent poverty and counties with persistent child poverty is significantly different from the counties without. I also found that almost all counties that have persistent poverty also have persistent child poverty. However, only about half of counties that have persistent child poverty also have persistent poverty. These intercorrelations may mean removing at least one of these variables before modelling.