Summary:

I began this project by finding and transforming this data into a usable format. I then explored the data using various types of graphs and statistical techniques. I found multiple features that are correlated with the ratio of convenience stores to grocery stores in a county including:

- **percent of the population who identifies as non-hispanic black**
- **percent of the population who has graduated high school**
- **percent of the population who has a bachelor's degree**
- **poverty rate**
- **median household income**
- **The first four are also correlated to the last.**

Finally, I trained a predictive model using this dataset. Gradient boosting preformed the best, but even after tuning the hyper-parameters, it still had a fairly high RMSE and low R^2

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.

After completing the statistical analysis of these data, I began working on creating the model to predict my target variable which is the ratio of convenience stores to grocery stores. The first step was to get the data ready. A convenient way to segment the counties geographically is by state so I moved it from the index of the DataFrame to a column and created dummy variables for all the categorical features. I then split the data into three parts: 70% for training the model, 15% for evaluating the model as I tried various regression techniques and feature engineering, and 15% for a final evaluation after all parameters are set.

For scoring all of the models I decided to use the root mean squared error (RMSE). Smaller RMSE are better, but there is no absolute scale to determine what a good RMSE is. I have found two ways people have used to assess RMSE scores. The first is to compare it to the standard deviation of the test set. The second is to compare it to the RMSE of using the mean of the train set as the predicted value. For this test set both are ~1.88. After looking at the math of these methods, I realized that these are the same equation except the first uses the mean of the test set and the second, the mean of the train set so the values should be similar.

For a baseline measurement of performance, I used a scaled version of the training set to fit a few linear models including a basic linear, lasso, ridge, and elastic net regression. None of these scored very well (1.77 - 1.79), and somewhat counter-intuitively, the regularized versions scored worse than the basic version. I then moved on to the tree methods. For the basic random forest, I changed the number of estimators to 100, which is the soon to be default, and the max number of features to be considered in each split to square root, which is the value I have heard is often better. I fitted the random forest to the training data multiple times to see randomness affected the RMSE score and found that while it was sometimes slightly better than the linear regression, it also sometimes scored roughly the same. For the final simple model, I used a gradient boosting regressor (GBR) with the max features set to square root and the max depth of a single tree set to one. This consistently had a lower RMSE score than the other models, but not by much.

The food atlas dataset contained a feature showing the median household income in each county. In order to have more information about the distribution of income, I added features containing the percentage of income tax returns filed in each of eight brackets in each county. I decided to test if the model performance changed if only one of these two were used. After training a similar GBR on the reduced datasets, I found that neither variant scored differently from the full dataset so I continued on with all the features.

The next step I took was to use the PolynomialFeatures function from scikit-learn which creates interactions terms and squares the features. I created interactions for

only the non-dummy features because the dummy features are only 0 or 1 which would likely not add much information, but would have meant there were more features than training set points. The RMSE for this expanded dataset was worse than the original.

After trying some feature selection and engineering and not finding any improvement, I began using grid search cross validation to tune the hyperparameters of the GBR including: the number of estimators, the max features used in a split, and the max depth of each estimator. Using the square root of the number of features for max features clearly was the best, but the others had no clear winners. I then ran another search on the two remaining features and found that a max depth of 3 was best and using a smaller number of estimators appeared better. To see if I should use fewer than 50 estimators, I added 20 to the list of parameters. This however was too small.

With the features and hyperparameters set, it was time to evaluate the model. I began by plotting the residuals vs the predicted value and the distribution of the residuals. The distribution is fairly normal with a long tail in the negative (under-predictions) with the mode in the positive. The sparsity of counties with large ratios of convenience to grocery stores leads the model to rarely predict a value larger than 5. These large underpredictions appear to be one of the drivers of the relatively poor RMSE scores. As a second evaluation method, I used the final holdout data to calculate a new RMSE score. At first it appeared that the model performed much worse on this dataset; however, the standard deviation of the true values is also much larger than for the test data (2.48 vs 1.88). As a percent of the standard deviation, the RMSE

of the final dataset is approximately the same as for the test dataset. The residuals graph is also similar to the test set.

Someone interested in everyone having access to the cheaper and fresh food available at grocery stores would likely have two main concerns: improving access and keeping access. Counties where the first is most needed can be identified by looking for the largest values of the conv_to_groc column. My model provides a way of finding counties which may be most at risk of access decreasing by identifying the counties where the predicted values is much larger than the current value. I identified 15 counties that have this property. These counties may also be able to provide insight into how to increase the number of grocery stores in a county.

For possible future work, I have found data on the percent of a county's population living within a certain distance of a grocery store. Since this directly measures access, it would be interesting to perform this analysis again with this variable as the target and comparing it with this indirect measurement of access.