Iowa Liquor Store Analysis
Executive Summary
Aakash Tandel
July 13th 2017

The goal of this analysis was to determine the best geographic location in Iowa in which to build a new liquor store. This analysis sought to determine location by maximizing the sales.

This data was a large dataset from the state of Iowa. It contained transaction level data for all stores holding a class E liquor license in 2015 and part of 2016. The full dataset contained upwards of 2.7 million transactions. Missing values and 2,973 duplicated columns were removed from the raw data. Because of our large number of observations, this had very little effect on the analysis. Based on location, there were 99 counties, 383 cities, and 676 zip codes represented. A large number of observations are found in Polk County, the city of Des Moines, and the zip code 50010 (Ames, Iowa). There were 72 different categories of alcohol, each of which was highly differentiated. There are 1400 unique stores in the data set. Lastly, the vast majority of sales were of quantities of less than 100 bottles and of transactions less than $1,000.

The correctness of the location and sales data was one major assumption. Data entry error, missing information, and other data problems could drastically affect our predictive model. The majority of the data refining aspect of this analysis was done with dummy variables. Dummy variables were created to indicate which county a transaction was in. This allowed the effect on sales to differ based on county.

Correlation matrices were used to perform feature selection for the model. High correlations between sales (the target variable) and various features suggested those features would be predictive in the model. This analysis resulted in the first model including location dummy variables, state bottle retail (or price per bottle), and the number of bottles sold as features.

The first model, hereby known as the Location Model, was a linear regression run through Python's Statsmodels. Both the number of bottles sold and state bottle retail were found to be statistically significant in effecting sales. Some of the location parameters were also statistically significant. The county that had the largest increase in sales was Dallas County and it was highly statistically significant. Based on this model, opening a liquor store in Dallas County would on average increase sales by about $56.25 above that of Polk County. Dallas County is located directly west of Des Moines. The analysis suggests opening a liquor store in Dallas County, Iowa would result in the highest sales.

There existed a risk of overfitting the model with the linear model including all of the counties in Iowa as dummy variables. Lasso Regression, or L1 regularization was used to reduce the error in the model. This ultimately did not change the conclusion that Dallas County was the best county in which to open a liquor store.

Lastly, a secondary model was run to determine which features most strongly affected sales in Dallas County specifically. This secondary model looked to determine how state bottle retail, the number of bottles sold, and the volume of liquor sold (in liters) affected the sales of a Dallas County liquor store. Using Lasso Regression, the variable which had the most affect on sales was number of bottles (followed by volume sold). Maximizing the number of bottles sold was the most effective way to increase sales, in

Dallas County, Iowa. Number of bottles sold was a better predictor of high sales than item type, average volume of bottle, or anything else.

In conclusion, the model run on Iowa's liquor sales data concluded that Dallas County, Iowa was the best locality in the state of Iowa in which to open a new liquor store and the best way to maximize sales in that county was to increase the number of bottles sold. Further analysis on the top performing stores in Dallas and the data associated with those stores would be necessary in order to develop a more robust business strategy.