Introduction

The objective of this project is to advise The Iowa State legislature if they should consider changes in the liquor tax rates using the current liquor sales by county and projecting the total sales for the rest of the year.

Data Overview

The data used in this project was obtained through Iowa's open data portal https://data.iowa.gov/. The Iowa Liquor Sales dataset contains all the transactions for all stores that have a class E liquor license[1].

In this project, we will only be looking at 10% of the data to train and test a model.

Methods

Data Cleaning and Engineering

I started with looking at the raw dataset and immediately noticed redundant columns such as having both "Volume Sold (Gallons)" and "Volume Sold (Liters)". I dropped "Volume Sold (Gallons)" because the bottle volume in the dataset is in milliliters. I also noticed that the cost of the bottle both actual and retail as well as the sales are not in a numerical format. I removed the dollar signs for these columns and converted them to numeric so that we can use these when we build our model.

In this project, I am using data from 2015 to make a linear model to predict the yearly sales of each store. As such, I had to split the data into 2015 and 2016 subsets and since 2016 only has sales from January to March per store, I also broke down the subsets by month. I did this by converting the Date into datetime format where I then extracted the month and year out of the dates. I also checked for missing values and found that the highest number of missing values is only 4% of our data, I decided to drop the rows with these null values.

I started with visually analyzing the 2015 sales dataset to gain insights on sales trend per month. I compared this to the first quarter sales of 2016 and saw that they both generally have an increasing trend. I also looked at the correlation between the variables and found that there's a high correlation between sales features and volume sold. There is no correlation between bottles sold and volume sold as well as the sales features and bottles sold.

I aggregated the sum of sales per store per month and transposed my data frames so that my features become my months and my indexes are the store numbers. I combined the sales for the first quarter and dropped the other quarters since our target only has the first three months. I proceeded to do this for the volume sold, total sales and bottles sold. I then merged these data frames into one.

Feature Selection

For this project, I wanted to test different combinations of the features that would give me the best projections for the rest of the year. Initially, I wanted to have store as a feature, but there are instances when a store shuts down or a brand new store opens. As such, it does not make sense to have the stores as each of our features but we can make use of the counties to be part of our feature set.

The first combination was to consider all columns. From there, I ran linear regression, ridge regression and lasso regression. I looked at the coefficients of the Lasso Regression and with its

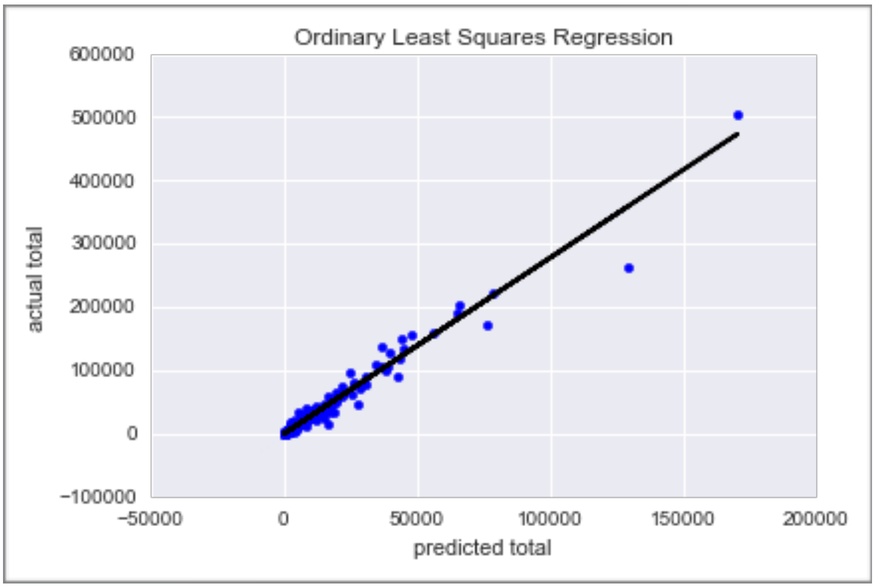
[1] For grocery, liquor and convenience stores, etc. Allows for the sale of alcoholic liquor for off-premises consumption in original unopened containers. No sales by the drink. Sunday sales are included. Also allows wholesale sales to on-premises Class A, B, C and D liquor licensees but must have a TTB Federal Wholesale Basic Permit.

two highest features (volume sold and first quarter sales), I reran the models using only this feature. From the correlation matrix that I did earlier, recall that there was a high correlation between sales features and volume sold. As such, I wanted to test out removing volume sold and rerunning the models using only the counties and the first quarter sales. This gave a very high R score but running the LassoCV, I saw that the coefficients of the counties were insignificant. I wanted to try seeing how volume sold would do when we take out the first quarter sales and leave the counties so I reran the models using those features and surprisingly, we ended up with a bad R score. Finally, I reran the models using only the first quarter sales and as expected, this gave us our best score.

Optimization
After finding the features (in our case feature) that gave the best R score, I ran a train test split so that we can fit the model on randomized data, after which I did a grid search to get the best parameters and score for our linear regression.

Results

The results show that the first quarter sales is the best predictor of our total sales. The model produced an R Squared of .985 and a root mean squared error (RMSE) of 8345. What does this all mean?



As can be seen in the figure above, the optimal regression model did not miss any of the data points very much, which means that the model was able to predict the total sales very well. This resulted in having a high R squared score. The RMSE is the square root of the mean of the square of all of the error. In this case, the RMSE represents the average amount that the model was off by.

We can see from the regression model that there is an upward trend for the total sales and it would be good for The Iowa State legislature to increase the rate of liquor tax.