Monika Helak

Final Report:
Iowa Liquor Sales

Problem Statement

Predicting demand is a common problem for retailers in every domain across the country. It is not only useful to predict revenue and cash flow, but also in the interest of optimizing a supply chain. I was curious in seeing how to what extent demand for liquor tends to fluctuate with time and with local rates of unemployment.

By using data from the State of Iowa on the Liquor Sales over the last 10 years, I created a random forest model that can accurately predict the volume of liquor sold in a given week of the year across Hy-vee Supermarkets. The tool I created can be used to predict demand given the season and socio-economic condition of the State.

Data Wrangling

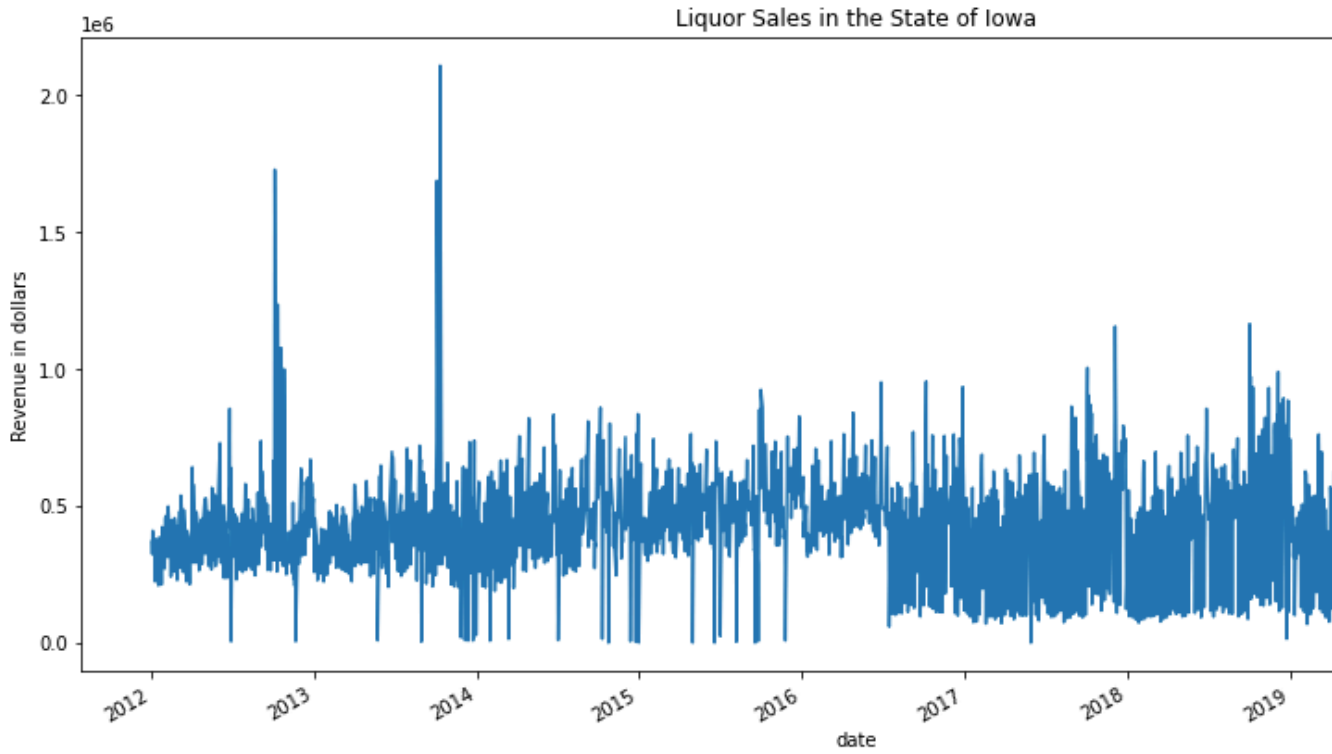
My raw dataset was downloaded in the form of a .csv file from the State of Iowa data.gov website, and its original dimensions were 19,445,831 rows and 24 columns. Since I was only interested in predicting the volume sold by a particular supermarket chain, I narrowed down the datasets to only observations from Hy-Vee Supermarkets. This reduced the file to (6458510, 24). I also downloaded another dataset from the same website that contained numbers of unemployment claims by county over the same time period. I merged these two datasets together on county, adding an additional column.

Exploratory Data Analysis

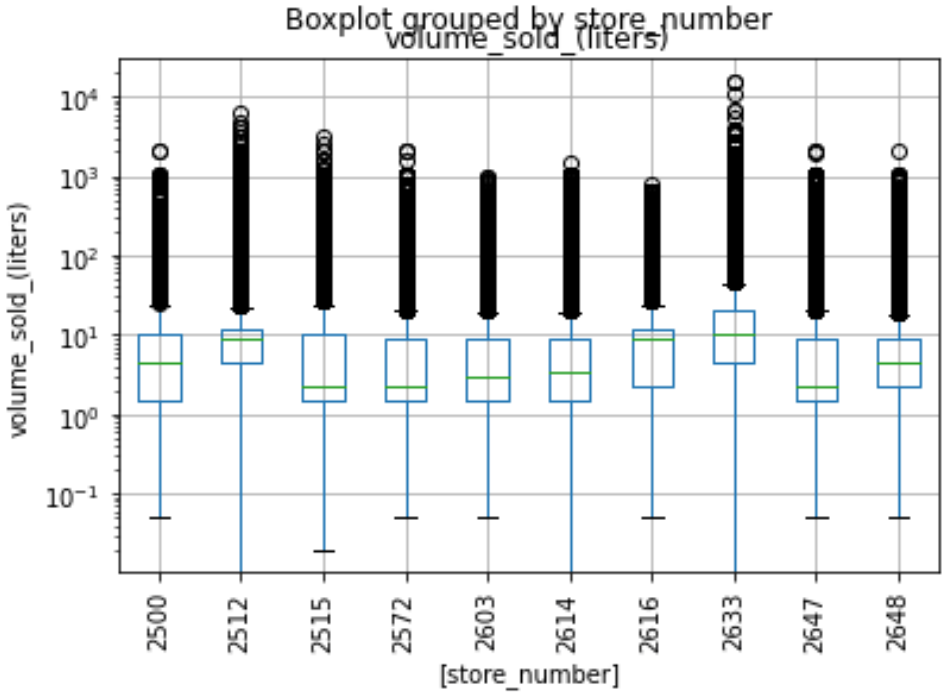
To draw more valuable insights and to counteract the curse of dimensionality, a reduction in the number of features was essential. With this in mind, exploratory data analysis was conducted where I looked at the relationship between independent variables and the target variable. Since the overwhelming amount of my features were categorical, a heatmap was not helpful in determining correlation due to the overwhelming amount of levels. The results of my correlation analysis are below:

- invoice number: unique for each transaction, do not need to keep
- store_name: correlated with store number
- address: also correlated with store number
- store location: also correlated with store number
- county number: correlated with county
- category: correlated with category name
- vendor name: correlated with vendor number
- item description: will keep item number instead
- bottle volume: not interested in bottle volume
- bottles sold: not interested in bottles sold
- volume sold (gallons): will keep volume sold (liters) instead

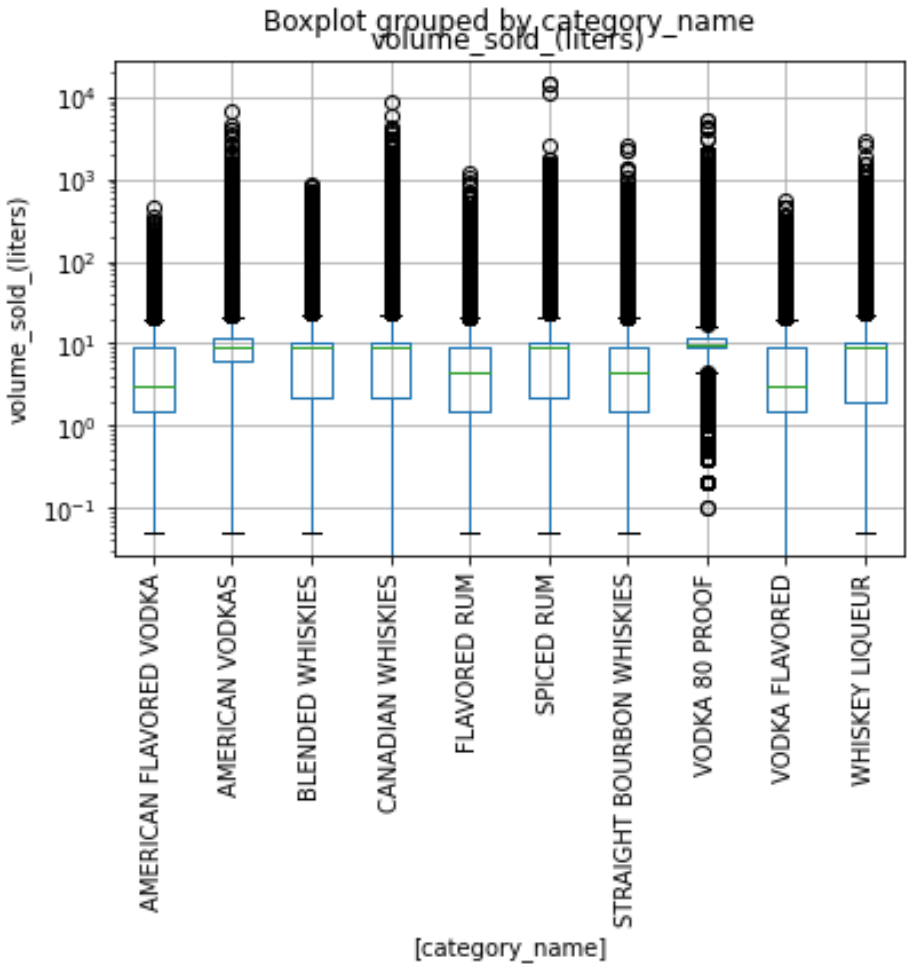
I examined what the volume of liquor sales looked like over time. As I had unemployment data available by week, I grouped the datetime data by week of the year.



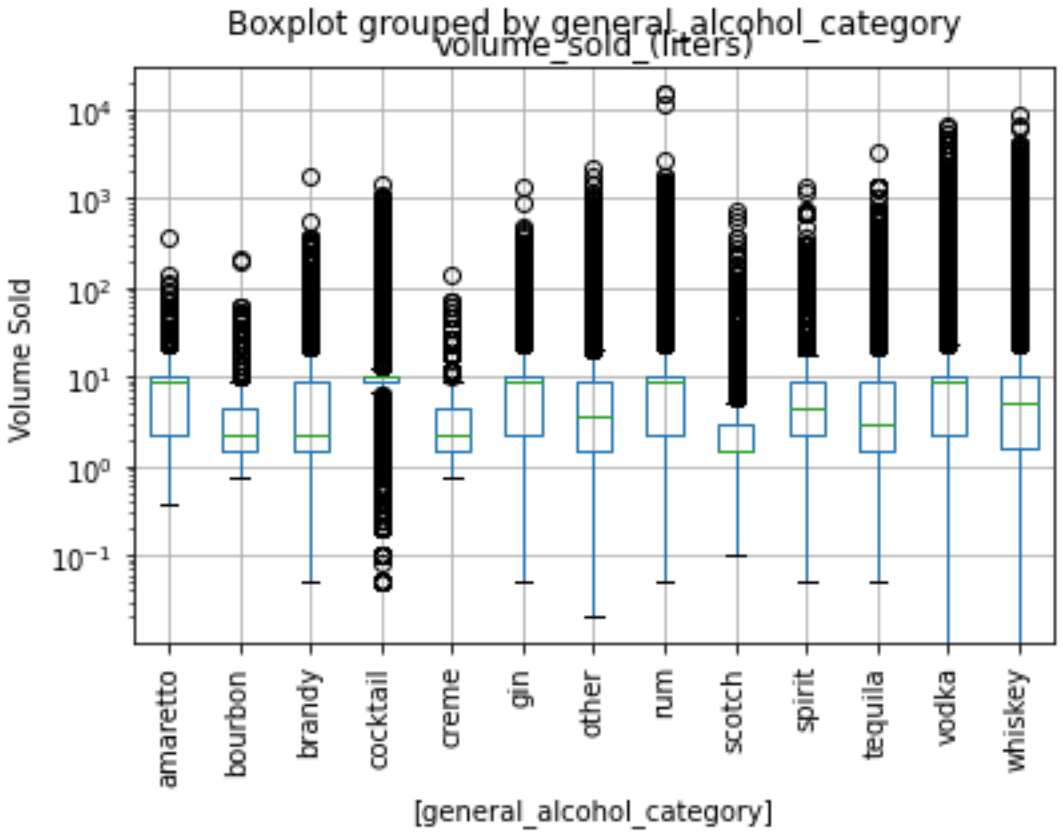
I made a series of bivariate charts to see which independent variables show significant variation among categories. As the number of stores was really large, I only looked at the 10 stores with the highest volume of sales.



The type of alcohol category also showed variation by the type of liquor sold. This makes sense, as different types of liquor are sold in different kinds of bottles and are consumed in varied quantities.



I wanted to dive further into this topic, so I feature engineered new categories that I grouped existing categories into. The logic was to narrow down labels like “Flavored Rum” and “Spiced Rum” into just “Rum”. This breakdown revealed some even clearer trends. I then ran an ANOVA to see whether there was a difference in the mean between all the categories. As expected, I was able to reject the null hypothesis that there was no difference.



Still, I was unsure whether crème and bourbon had the same mean. I wanted to investigate whether this was true, so I conducted a t-test to see if the volume sold in each of the two categories was the same. I calculated the mean, standard deviation and standard error on the difference between the samples. The p-value resulting from the t-test was 8.05 * 10^-8, yielding that I could reject the null hypothesis the means of the two samples were the same.

Upon narrowing the model down to just six predictors, I converted the categorical variables into dummy variables which greatly increased the dimensions of my dataset. I also split the dataset into a 75/25 train test split.

Model Selection

I approached the regression problem by trying out three different machine learning models: Linear Regression, Decision Tree and Random Forest.

Ultimately, after weighing the advantages and disadvantages, I decided that a random forest model is the best fit for solving the business problem. One of the advantages of random forest is that it can handle large datasets with high dimensionality, which was the case after converting all the categorical variables in my dataset to dummy variables.

Traditionally, linear regression is considered the model that is most interpretable. After fitting a linear regression, a stats model output shows all regression coefficients from which it is clear how each feature contributes to the slope of the line. While this is not nearly as clear in

the case of random forest, there is a method of looking at the feature importance. Nevertheless, the method of how the algorithm arrives at the feature importance is complex.

I chose mean absolute error as my evaluation metric. One of the advantages of using MAE is that outliers are not given as much weight as they are with MSE. It provides a rather generic measure of how our model is performing. Out of the three models I tested, random forest had an MAE of 69.48, which was lower than linear regression but higher than decision tree.

Another consideration in the selection was the time that it took the model to train. Out of the three models, Random Forest took a significantly longer time to train that the others. When optimizing the model, searching for the best parameters using Grid Search also took the longest for random forest (around 10 minutes).

Training time for each model:

Linear Regression	**Decision Tree**	**Random Forest**
6.09 s	18.72 s	121.49 s

However, given that this value was just around 2 minutes and time was not a key consideration in the selection process.

Takeaways

All things considered, I settled that Random Forest is still the best model for predicting volume of liquor sales across Hy-Vee Supermarkets in Iowa. The Random Forest model yielded the best performance. Results from 5-fold cross validation did not suggest overfitting, which could be a disadvantage of using Random Forest. Even though the training time for this model was the longest, I found that the superior performances outweighed the time cost, as the model was not intended for real-time analysis.

The model was also able to show feature importance. Unemployment rate was one of the most important features in the model, suggesting that the socioeconomic condition of the state should not be underestimated when forecasting liquor supply.

Future Considerations and Improvements

There are some limitations and shortcomings of this analysis that could lead way to opportunities to dive deeper into how the unemployment rate impacts volumes of liquor sales.

One of the areas for improvement could have been a closer analysis of the outliers. I found that the volumes of purchases varied greatly, likely because of a variety of customers ranging from individual buyers to bar and restaurant purchasers. More extensive outlier detection methods could be explored.

Other improvements to the model could be explored in the area of time optimization. One time-saving strategy could have been to train on a smaller dataset and plot the performance of each model as a function of the size of the training set. After choosing the size of the dataset at which performance hit an asymptote, model training of a diverse set of models could have taken place with a smaller dataset. This would eliminate the need to training on an extraordinarily large dataset, when similar results could have been achieved with a subset of the original.

In the future, other techniques for hyperparameter tuning could be explored. Grid search is not considered one of the best techniques for hyperparameter tuning because of the training time needed to evaluate all the different combinations of parameters. Random search would likely produce similar results and Bayesian methods could be tested.