Iowa Sales Analysis

Biostat 625 Final Project Report

Group 15: Hua Jiang, Haisheng Xu, Longfei Zhang, Jessica Zhao

12/14/2021

Background

The pandemic of COVID-19 hit the world by surprise in early 2020 and had since left profound impacts to social, economical, and physiological aspects. As a worldwide pandemic that's not going away anytime soon, it has certainly caused significant changes to public behaviors. Many industries such as airlines, leisure facilities, auto parts, restaurants, and etc. were negatively impacted, so was the alcoholic beverages consumption. But according to the surveillance report from the National Institute on Alcohol Abuse and Alcoholism, alcoholic beverages sales increased while the most sales of food and other drinks decreased during the pandemic of COVID-19. Hence, our group was particularly interested in the economical impacts due to the pandemic. Through investigation on alcohol sales in Iowa, our group tried to figure out economic influences of the pandemic. And we believe that the analysis of the Iowa alcohol sales data provided a good reference to the overall trend in the retail sector.

Dataset

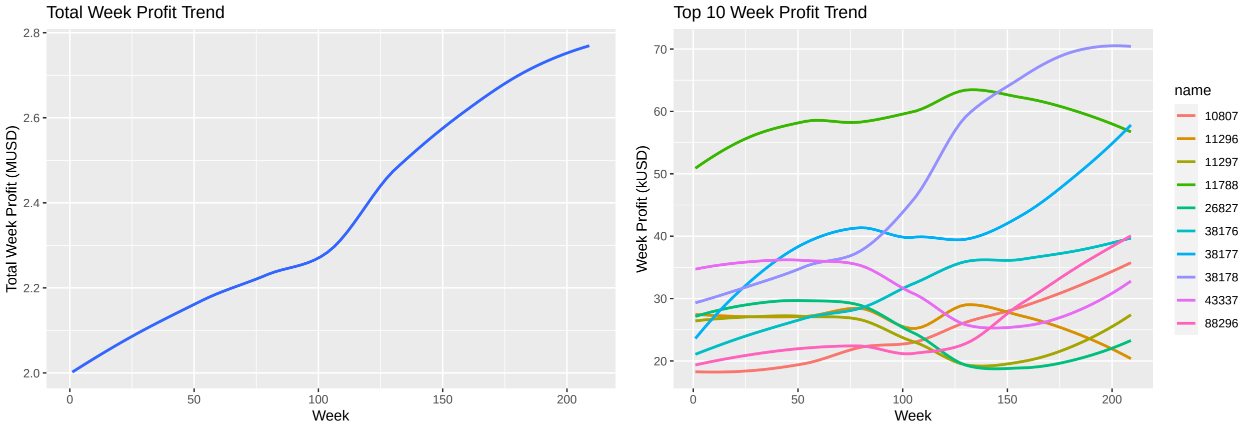
The dataset chosen for this analysis is the "Iowa Liquor Sales" data publicly available from the state of Iowa. More specifically, this dataset contains the alcohol purchase records of Iowa Class "E" liquor licensees from January 1, 2012 to current. Class E liquor license, for grocery, liquor, convenience stores, etc., allows the above to sell liquor for off-premises consumption in unopened original containers in a commercial setting. For our analysis, sales data from Oct. 2017 to Oct. 2021 was selected, such as store, brand, sales volume, and sales price, considering data duplication, completion, etc. Overall, more than 9,900,000 observations and 24 variables were evaluated. The dataset is extremely large and intricate. It would be hard to do the analysis on the dataset directly. We need to clean such a large data (more than 2.4G) in different ways for each method, and we need to make sure each function we use works perfectly on our large dataset. In the stage of data processing, we removed data entries with missing values due to the limitation in analytical methods and tried to use vectorization for better performance.

Method

To figure out the trends of alcohol sales and have a better understanding on the characteristics of popular alcohol drinks, we decided to use the following five methods: Exploratory Data Analysis, Sales Trend Analysis, Correlation Analysis, Classifications, and Visualizations. In the Exploratory Data Analysis part, we summarized the statistics of our data using overall sales trends and tried to find the best methods for further analysis. In the Sales Analysis section, we mainly used the Time Series method to see if there was an obvious trend and seasonality in overall sales. At the same time, we predicted the future trends of sales with and without seasonality in this part. Next, we simply conducted correlation analysis to see the associations for the quantitative variables between bottle sold and other variables. After that, we tried to use classification for two different response variables to see the characteristics of popular alcohol drinks. Finally, we summarized and reported our results by visualizations.

Results (Trend)

To evaluate the influence of COVID-19 on liquor sales in Iowa, we first drew the overall sales trend during the four years. As the Loess smoothed weekly profit trend of all liquors shows, during the four years, liquor profit grew steadily and eventually increased by 20 percent compared to 2017. At the start of the pandemic, around week 105, the profit was influenced a little bit, but quickly

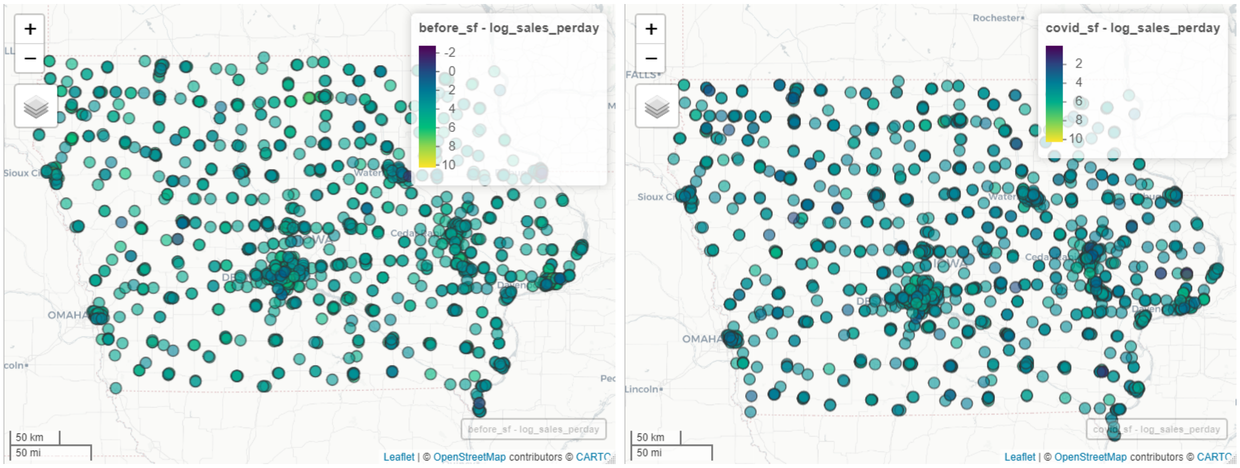


Weekly Profit Trend Plot

returned to a strong growth trend and grew even faster, which was consistent with previous research.

Based on this result, we further checked the trend for top 10 liquor sales. The trends of those liquors generally follow the total trend. Most maintained strong growth trends during the pandemic, while only 2 of 10 shows a significant decreasing trend of profits.

Results (Map visualization)

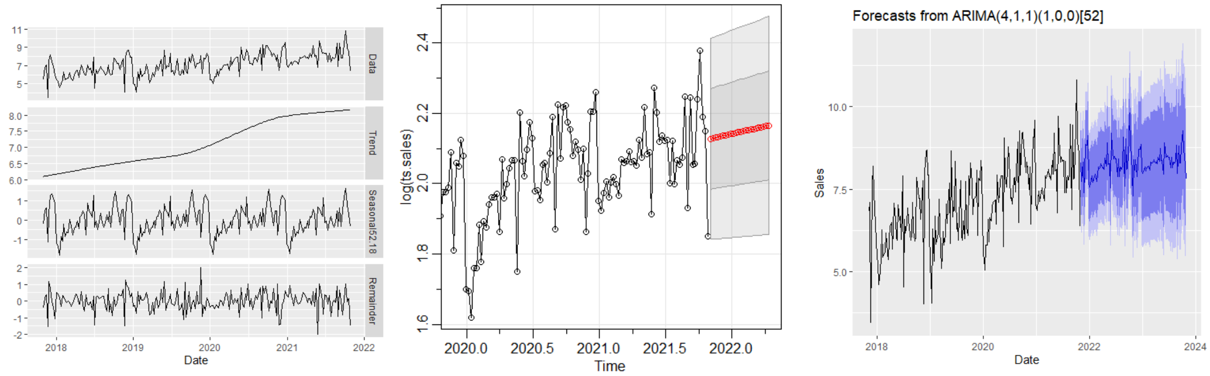

before_sf - log_sales_perday
-2
0
2
4
6
8
10
Sioux C
OMAHA
coln
IOWA
Cedar
Water
Pec
50 km
50 mi
before_sf - log_sales_perday
Leaflet | © OpenStreetMap contributors © CARTO
covid_sf - log_sales_perday
2
4
6
8
10
Rochester
FALLS
Sioux C
OMAHA
Lincoln
IOWA
Water
Ceda
50 km
50 mi
covid_sf - log_sales_perday
Leaflet | © OpenStreetMap contributors © CARTO

Map Before and During Covid

To further study the distribution of alcohol sales and the magnitude of those sales, we displayed the sales per day for each location using map visualization. For the faster data importing process, we used the fread function to first get the column names that needed to be analyzed, then we used fread again to get the columns that we wanted. Doing so was faster than importing the entire data all at once. We first separated the data into two groups using the date of purchase. The first group contained the sales record between 2017-10-31 and 2019-10-31 which was before COVID-19, and the second group was the sales record between 2019-11-01 and 2021-10-31 which was during the COVID-19. Since the sales data were highly skewed to the left, we decided to log transform the sales per day for better illustration.

The first map was showing the location of each store and the log-transformed sales per day before COVID-19. The second map was showing the location of each store and the log-transformed sales per day during the COVID-19 period. We can see that more stores were located in bigger cities, but only one or two stores in that city had skyrocketed sales compared to all the others. And the others had relatively the same amount of sales as stores in suburban areas. Moreover, stores located in suburban areas tend to open at the major intersections. Compare the two maps, we can see that there was an increase in sales during COVID-19 for stores located in suburban areas. The increase in sales could be caused by leisure drinking while spending more time at home or working from home due to COVID, or could be using alcohol as a destresser for quarantine or unemployment.

Results (Time series)



Time Series Overview and Prediction

For time series analysis, we sum up the weekly sales(million dollars) for all kinds of alcoholic beverages as one data point from Oct. 31, 2017 to Oct. 31, 2021. Thus, we have 208 data points in total to conduct the time series analysis.

The time series is not stationary. Overall, The sales increase through the time, and this time series does not show obvious heteroskedasticity by looking at the plot. At the same time, the trend shows a pattern of seasonality: At the beginning of a year, the sales are usually at the lowest point of the year. Then the sales continue to increase until the end of the year close to the Christmas Day, which reaches the peak of the sales in that year. Then the sales decrease suddenly after the peak and reach the lowest point in the next year.

For the seasonality version, we made a prediction for the future 104 weeks which is shown in the plot above. Because we only have four years data, so the confidence interval might be slightly large. According to the forecasts from ARIMA(4,1,1)(1,0,0)[52], in the future, the sales might continue increasing, and the peak of the sales in one year would be close to the end of the year.

We also did a version of time series analysis without considering the seasonality. In that version, we used log() and diff() to remove the trend and seasonality. In that case, we use acf and pacf plots to determine the ARIMA model. We finally used the MA(1) model which has lower AIC and BIC compared to AR(7) to predict the next 12 weeks of Oot. 31, 2021.

Results (Correlation)

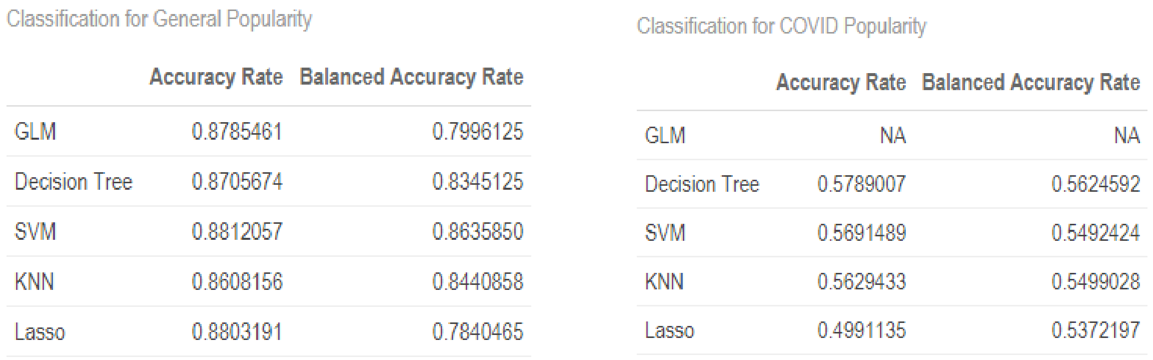
We conducted a correlation matrix to present the correlations visually between the variables we were interested in. In a brief word, we can see that the bottles sold might be positively associated with how many stores(0.45), cities (0.39), counties (0.29) are selling that alcoholic beverage. The more counties, cities, stores selling a specific alcoholic beverage, the more popular that alcohol product is. And the number of packs might be also slightly positively associated with bottles sold. When there are more bottles in the pack, the bottles sold might be higher.

There might be a multicollinearity issue between Store.Number, County, and City at the same time because of the high correlations (greater than 0.80 for each pair of these three variables).

Results (Classification)

For the classification section, we tried to use Generalized Linear Model (GLM), Decision Tree, Random Forest, Neural Network Regression, Support Vector Machines (SVMs), K-Nearest Neighbors algorithm (KNN), and Least Absolute Shrinkage and Selection Operator (LASSO) respectively. However, because there are too many levels (more than 100 levels) in some categorical variables such as "Vendor" and "Category", Random Forest and Neural Network Regression might be applied to our dataset.

We implemented classification for two response variables: One is for the general popularity of alcoholic beverages; The other is for the COVID popularity of alcoholic beverages. We defined the overall top 10% bottle sold alcoholic beverages as the general popular products, and we defined the alcoholic beverages that have more sales during the pandemic are the COVID popular products. The explanatory variables include Vendor, Category, Pack, Bottle Volume (ml), and so on. We randomly split the data into a training set (60%) and a test set (40%) to measure the accuracy of prediction which can reflect the reliability of the model. The table below shows the accuracy.



Classification for General Popularity

	Accuracy Rate	Balanced Accuracy Rate
GLM	0.8785461	0.7996125
Decision Tree	0.8705674	0.8345125
SVM	0.8812057	0.8635850
KNN	0.8608156	0.8440858
Lasso	0.8803191	0.7840465

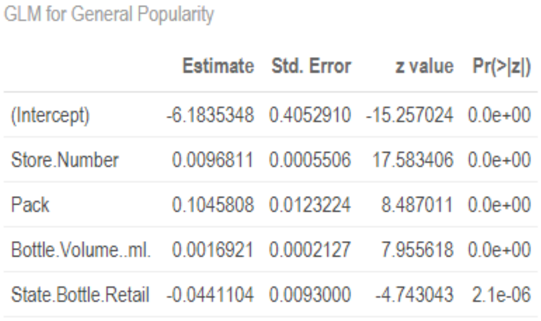
Classification for COVID Popularity

	Accuracy Rate	Balanced Accuracy Rate
GLM	NA	NA
Decision Tree	0.5789007	0.5624592
SVM	0.5691489	0.5492424
KNN	0.5629433	0.5499028
Lasso	0.4991135	0.5372197

Accuracy Tables for Classifications

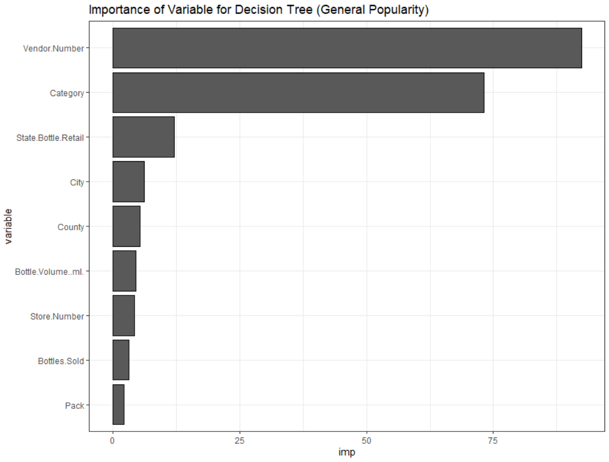
We can see that the models for general popularity all have good accuracy, but for COVID popularity, the models look not accurate enough. The GLM method is not applied for COVID popularity because there is only one significant predictor after stepwise selection. We cannot find a good threshold for that prediction.

We can find that for the general popularity, the accuracies are relatively high and all results almost keep consistent (except the decision tree: it might not avoid the multicollinearity issue). The most important predictors are Store.Number, (City, County,) Category, Vendor, Retailed price, Bottle Volume, and Pack. When an alcoholic beverage has a larger pack, larger bottle volume, lower retailed price, and more places people can buy, it tends to be a popular product(higher bottle sold).



GLM for General Popularity

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.1835348	0.4052910	-15.257024	0.0e+00
Store.Number	0.0096811	0.0005506	17.583406	0.0e+00
Pack	0.1045808	0.0123224	8.487011	0.0e+00
Bottle.Volume..ml.	0.0016921	0.0002127	7.955618	0.0e+00
State.Bottle.Retail	-0.0441104	0.0093000	-4.743043	2.1e-06



Significant Variables from GLM and Importance of Variables from Decision Tree

For the COVID popularity, some methods keep consistent while some are not. For example, GLM after stepwise selection shows that the only significant predictor is retail prices, while the LASSO method indicates that the significant variables converged include Category, Vendor, Bottle Volume, and so on. It might be because the predictors in the dataset are not significant enough. There might be none or very few significant variables. The final accuracy for each model all proves it. We think we cannot find the characteristics of increasing sales of alcohol in the pandemic due to the very low accuracy of our classifications. The variables in the dataset may not answer the question of which kinds of alcohol products are more popular during the pandemic.

Conclusion

Based on Iowa liquor sales data from Oct. 31, 2017 to Oct. 31, 2021, we firstly verified that the alcohol industry maintained a strong growth trend during the covid pandemic. Then, through GeoMap visualization, we found that stores located in suburban areas tend to gain more increase in sales. Time series analysis shows that sales of liquor in Iowa will still grow but at a slower pace with the similar seasonality compared to the trend before, maybe because of public adaptation to life under the pandemic. And the classification results show that the alcoholic beverage with a larger pack, larger bottle volume, lower retailed price, and more places people can buy would be more popular. We did not find the characteristics of the alcohol products which became more popular during the pandemic of COVID-19 based on our data.

However, we still have some noticeable limitations on our study. For map visualization, around 100 stores had a missing value for the store location which could not be plotted on the map and led to flaws in our analysis. For time series analysis, we only retrieved four years of data which is not enough for more detailed time series analysis. Some variables have too many levels which are not compatible for some classifications. The variables in the dataset for classification might be insignificant. For the classifications, we cannot identify whether those characteristics, like bottle volume, pack, retailed price, and so on, make the alcohol products become popular. To investigate more for the popular alcohol products in the future study, more survey data is needed.