Where to Campaign: Strategies for D.E.A.D.

James Rounthwaite: jrounthw@calpoly.edu

Sophia Chung: spchung@calpoly.edu

Lana Huynh: lmhuynh@calpolyledu

Jamie Luna: jluna28@calpoly.edu

Rachel Roggenkemper: rroggenk@calpoly.edu

Objective

The problem proposed to us was to find impactful trends and their variables in alcohol consumption given data from all of the liquor stores in the state of Iowa, so that DEAD can decide where and how to target their campaigns. With this in mind, our goal was to create a model that would present individual factors in alcohol consumption per year per alcohol type. Due to the importance of pattern recognition, we prioritized a model that would make precise trends. Additionally, a limitation of this model was finding a metric to represent liquor store sales, since that data is not available to us. In place, we are utilizing the total cost of a liquor order, which is the number of bottles multiplied by the state bottle retail.

Data Overview

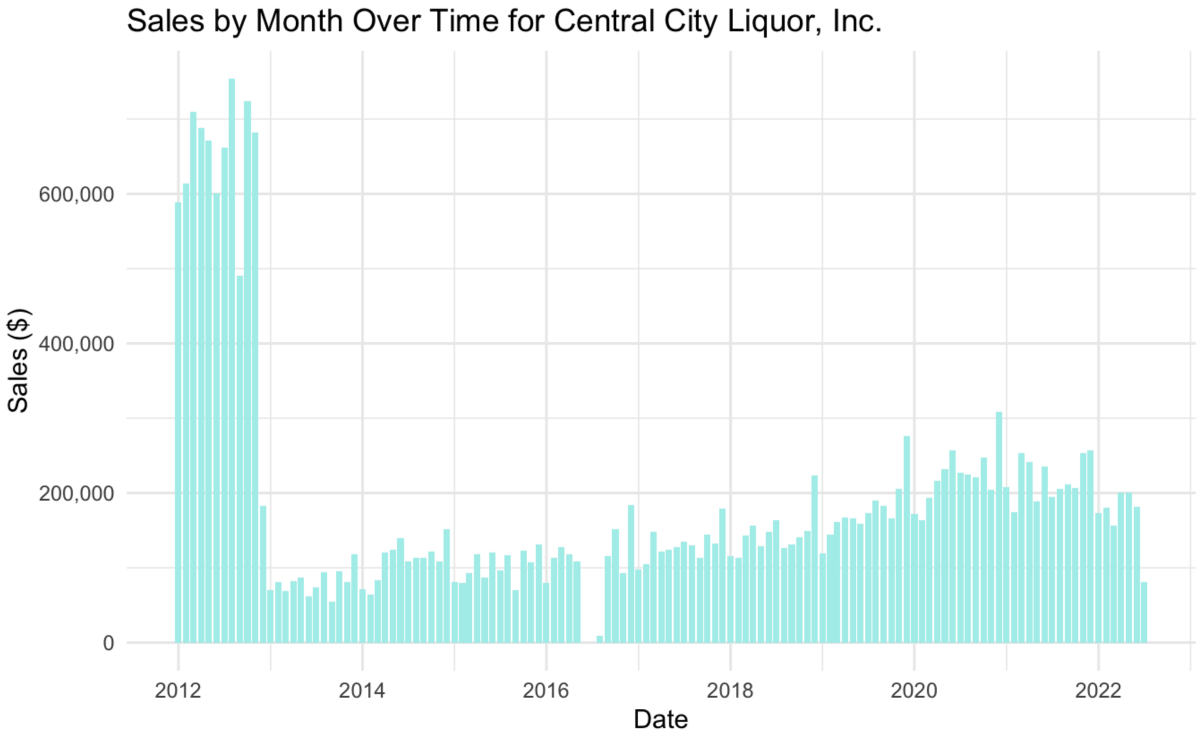
The dataset we used was data on alcohol sales in Iowa, which was collected by the state government. This dataset contains the spirits purchase information of Iowa Class "E" liquor licensees by product and date of purchase from January 1, 2012 to current. Class E liquor license, for grocery stores, liquor stores, convenience stores, etc., allows commercial establishments to sell liquor for off-premises consumption in original unopened containers. The dataset was acquired from the following website: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy.

The original dataset included twenty-four columns, such as the date of the order, the store number and name, information about the location of the store, the category of liquor ordered, the vendor of the brand of liquor ordered, and the cost and volume of the individual liquor products ordered. Since the original dataset contains over twenty-seven millions rows of individual product purchases, we only analyzed a subset of the original data.

The subset of data we utilized contained all the individual product purchases during a select year or two. We chose to only look at one or two year's worth of data per each model to get a better sense of changing trends throughout the decade. Lastly, we narrowed down each model to track a particular category of alcohol. After we had this subset of data, we first implemented feature engineering by using the date column to create three new columns: month, day, year. Feature engineering is the process of building new columns that provide more insight into data.

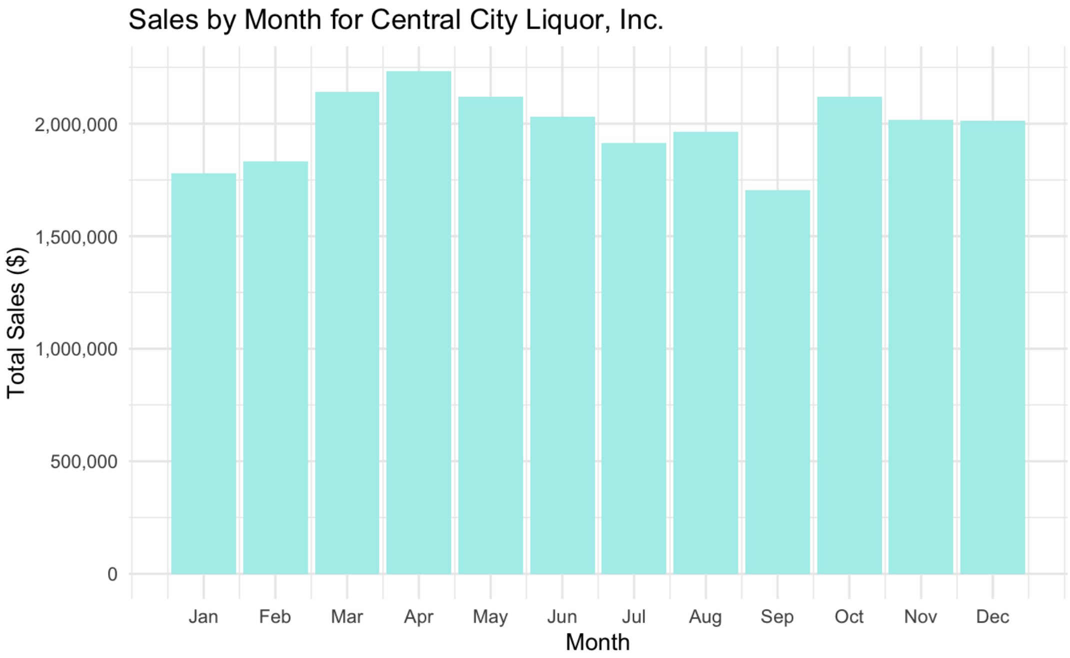
Therefore, we could use these as predictors. Not only did we want to look at where alcohol was impacting sales, we also wanted to look at when sales were happening for a better idea of when to campaign. To get an initial understanding of the total sales per month over time, in Figure 1 shown below, we have plotted the total sales to get an idea of the trend of sales over time. As we can see, the year of 2012 had a drastically greater number of sales per month compared to any other month and year. An additional point of interest from this figure is in the middle of 2016, there seem to be little to no sales.

Figure 1: Sales by Month Over Time for Central City Liquor, Inc.



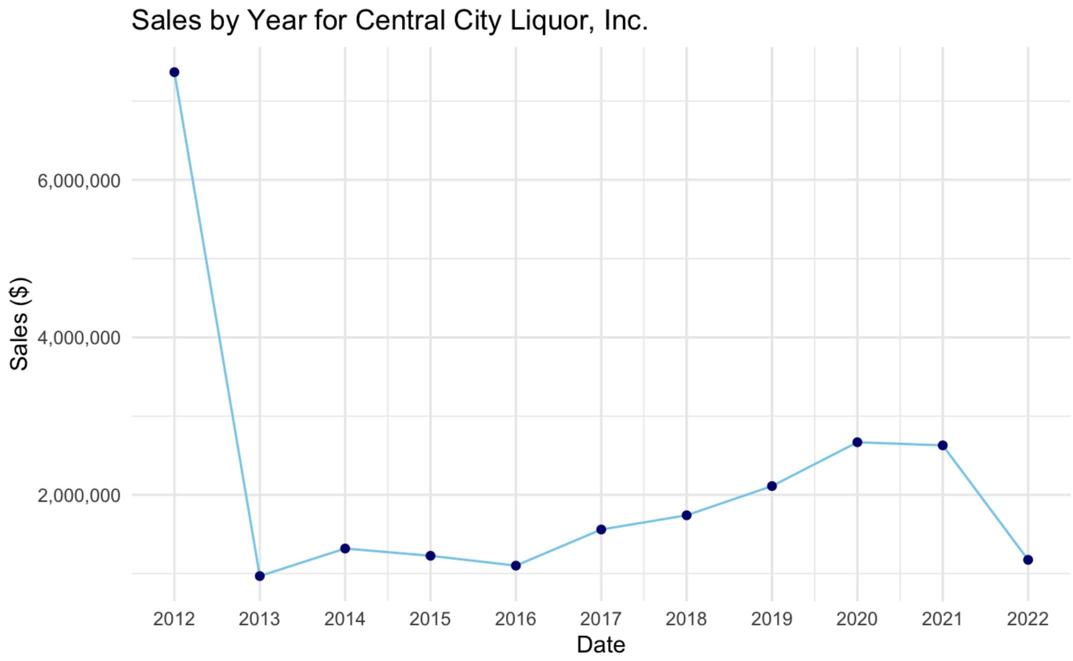
For a more fine grained look at a particular year we zoned in on the sales per month. This is especially helpful since one of the ways our models decide patterns of sales is by using the month. Figure 1, explores the trends of month and year separately as well. To do this, we have plotted the total sales per month in Figure 2 and total sales per year in Figure 3, both shown below. As we can see in Figure 2, we see that April is the month with the highest sales. Additionally, Figure 2 displays that September, January, and February have the lowest sales of the year. A potential reason why January and February have low sales could be because of New Years' Resolutions. Many people could be making resolutions to not drink or drink less, which decreases the amount of sales, which ultimately means the liquor stores are buying less alcohol.

Figure 2: Sales by Month for Central City Liquor, Inc.



To move on to Figure 3, we see a similar trend as witnessed in Figure 1. The year of 2012 had a drastically greater number of sales by year compared to any other year. Ultimately, now we have an initial idea of the patterns by month before we start trying to predict future sales if breaking down sales by month and by year is something we want to consider in our model.

Figure 3: Sales by Year for Central City Liquor, Inc.



Next, we filtered out columns that we deemed would not be relevant to our analysis, only keeping the cost of a liquor order, year, month, day, and counties of where the alcohol was purchased. We then aggregated the purchases by day, counting how many times each county had a purchase each day and the total cost of alcohol for that day. Thus, in this example, each row represented a day of alcohol Straight Bourbon Whiskeys. Lastly, we standardized the columns so we could achieve consistent model interpretations, especially since we used penalization such as Ridge when fitting our models.

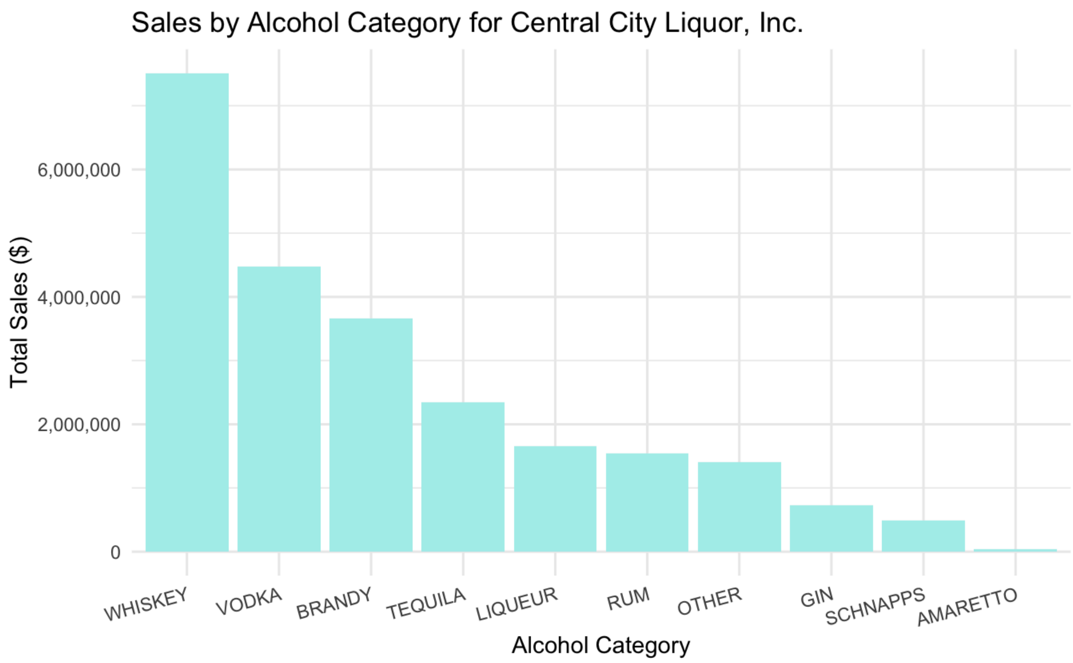
Table 2: Sample of our Final Dataset

Total cost of liquor order	**Month**	**Year**	**County 01**	**County 02**	**Country 03**	**…**	**Country 94**	**Country 95**	**Country 96**	**Country 97**	**Country 98**	**Country 99**
15922.40	-0.689512	2021	-0.4111	-0.4088	-0.4150	…	0.10956	0.1095	-0.4236	-0.46525	-0.7459	-0.4288
21295.84	-1.33321	2021	-0.4111	-0.4088	-0.41505	…	-0.4879	-0.4652	1.5558	-0.38217	-0.4288	-0.4288
74051.71	-0.36766	2021	-0.4111	-0.4088	-0.41505	…	-0.4871	-0.465	-0.5898	-0.38217	1.62052	0.207

5684.83	-0.04581	2021	1.2786	0.15128	0.950419	…	1.68415	-0.4236	-0.7459	-0.38217	-0.4288	0.718

To visualize what our final dataset looks like, Table 1 displays a sample of four rows of our data, including our column titles. Since we standardized the columns, the numbers don't make intuitive sense for interpretation. However, what the rows represent are on a particular day, for the county columns, it signifies how many (standardized) purchases were made in that county, and the total (standardized) cost of alcohol purchased for that day. Although each row represents a day of purchases, we decided to only use the standardized month and year in our analysis because we wanted to be able to predict monthly sales and did not think the day of purchase was relevant. During our model iteration we determined that the standardization was not entirely needed.

Figure 4: Sales by Alcohol Category for Central City Liquor, Inc.



One reason to justify first looking at whiskey is that it makes up most of the sales for Central City Liquor (a very large liquor store in Iowa) and for many other stores. This can be seen in the above figure 4 By first focussing on this largest demand we can see the most impactful trends that play into liquor purchases and use those insights to build better models for other alcohol categories.

Model Selection Process

There are several models we considered using, but in the end, we tested out three different models to predict sales: ordinary least squares (OLS), ridge, and log-cosh. Additionally, we utilized scikit-learn as a baseline model to compare our results.

OLS regression is an optimization strategy that helps you find a straight line as close as possible to your data points in a linear regression model. In other words, OLS regression tries to match the data as closely as possible.

Ridge regression is used in cases where there are many parameters or predictors that affect the outcome. However, the main drawback of ridge regression is it focuses on keeping the model simple while still trying to capture all the trends.

Log-cosh also tries to match our data, but it is much more resilient to outliers than the other methods described.

We also needed to consider how to validate our data. We utilized k-fold cross validation. In k-fold cross validation, the training data used in the model is split into k numbers of smaller sets, in which the model is then trained on k-1 folds of the training set. The remaining fold is then used as a validation set to evaluate the model.

In the end, we ended up choosing the OLS regression model, due to its simplicity in tandem with leveraging time-series cross validation to build the most optimal model.

Final Model Summary

After using OLS regression to build our model, we ended up with the model as the following:

$$\mathit{Sales} = \mathit{Year} + \mathit{Month} + \mathit{CountyCode\_01} + \mathit{CountyCode\_02} + \mathit{CountyCode\_03} + \dots + \mathit{CountyCode\_99} + \mathit{Other}$$

With this model, we ended up with an R^2 of 0.651 and a mean squared error (MSE) of 0.14. This means that our model explains about 65.1% of the variability in the data. Our MSE value of 0.14 is relatively low compared to other models that we fit.

Table 2. Coefficients of predictor variables

Year	**County 1**	**County 2**
2021	Adair	Polk
2020	Story	Polk

2019	Polk	Warren

As we can see from table 2, the county with the largest impact changes from year to year. However, we can see that Polk county has a strong relationship with increased alcohol sales of whiskey. Each model per year can give different results and more recent models can be used to gain insights on the coming year.

Ethical Concerns and Recommendations

There are several ethical considerations that should be taken into account in this analysis. Firstly, it is important to ensure that the data used does not contain any personally identifiable information about any individual. All personal information will be anonymized or redacted.

In our model, the predictions we make are based on historical trends and data. It is important to note that these are forecasts, not guarantees. This model can only make educated guesses based on the past information it has been given. It is built on the assumption that past trends will continue into the future. As a result, there will inherently be a certain degree of uncertainty in our model. There are many factors that can influence sales, some which are beyond the scope of the data used (i.e. economic changes or natural disasters). Another limitation of our data that is critical to acknowledge is that we used the number of bottles multiplied by the state bottle retail as a proxy for sales which may not be entirely accurate. Overall, we made numerous assumptions about our predictors and response, so it is of utmost importance to ensure the interpretations are correct.

This model is a decision support tool. It is a valuable resource that can inform decisions, but those decisions should also be informed by expert judgment and real world business insights.

Takeaways and Conclusions

Our model will be a vital asset to your non-profit organization, DEAD, as you look to target more high impact areas. With the structure behind our model, you will be able to plug in data for any type of alcohol, or any number of years, and see the most impactful counties. This will allow you to find your highest and lowest impactful counties and to determine where to be expending resources. Also, from the data you can see that some months have higher rates of selling alcohol than other months, and we wanted to ensure you would be able to see those trends so your campaigns can prepare accordingly. This gives a more granular view at a campaign's performance and can allow you to determine whether it is performing in a satisfactory manner on a monthly basis rather than only a yearly basis. Furthermore, each model is separated by alcohol type. By using our model, your organization will be able to see which counties impact each alcohol category to capitalize on the opportunity to maximize their campaigns.

You will be able to know how accurate our model is and how well it is performing based on the R^2 value for each calculation. This value will be between 0 and 1. R^2 tells us how much of the variation in total projected sales is explained by the model we provide. In other words, the closer the R^2 value is to 1, the more accurate it is at predicting the sales.