Getting to the Proof

Analyzing Liquor Store Franchise Transactions for Booze ‘R’ Us
Data-Driven Sales Predictions and Growth Insights

by

Andrew Kerr - adkerr@calpoly.edu
Bella McCarty - imccarty@calpoly.edu
Erik Luu - eeluu@calpoly.edu
Martin Hsu - mshsu@calpoly.edu
Matteo Shafer - mshafe01@calpoly.edu

Booze ‘R’ Us requested a method by which to predict future sales for growth purposes. We proposed a machine learning approach fitting a multiple linear regression model to historical monthly sales data from Booze ‘R’ Us’s storefronts and applied the modeling process to a case study similar in scope. The model was fit to historical data and accurately predicted future monthly sales for an average storefront. Furthermore, the features included reflected factors that we found tend to drive or have the greatest impact on monthly liquor sales. Through our analysis, we selected three main feature setups and evaluated each to finalize using month, size of bottles, price, and type of alcohol. In addition to developing a robust predictive model, we identified specific sizes, price ranges, and liquor types that serve as the primary drivers of sales.

I. Introduction

Accurate sales forecasting is critical for business growth planning. This project aimed to develop a recommendation for liquor sales growth at Booze ‘R’ Us storefront in Iowa based on statistical modeling. Historical yearly and monthly liquor sales data from franchise stores in Iowa was analyzed to uncover trends and relationships between sales and relevant factors like month, bottle size, bottle cost, and liquor type. A multiple regression model was developed that captures how these factors influence sales at a storefront level. This model can be fitted on Booze ‘R’ Us to strategically plan for expansion opportunities while quantifying uncertainty in predictions. Overall, statistical modeling provides real, customized insights to confidently grow their operations.

II. Data Preparation

The dataset we used as our main source for analysis and training was provided to us by the State of Iowa's public data platform, data.iowa.gov. We obtained the "Iowa Liquor Sales" dataset[1]. With attributes specifying details about every individual liquor purchase by Iowa Class "E" liquor licensees–grocery stores, liquor stores, convenience stores, and more, organized by product and date of purchase. With observations beginning on January 1, 2012, and running through the present, we had access to 27.5 million individual records to analyze.

To analyze patterns in sales to determine if Booze 'R' Us should expand its operations, we took subsets of the data, filtering for only a single franchise, "Casey's General Store." This franchise consists of multiple storefronts, closely matching the business model of Booze 'R' Us's multiple franchise locations, making it an ideal candidate for a case-study style application of the modeling process we hope to perform on Booze 'R' Us's sales data. Data from the years 2017-2020 were selected. After computing features, the data were aggregated into observations that represent a unique storefront-month-year combination, as seen in Table 2.1.

We aimed to use our model to estimate the average monthly sales per storefront for each new month. Sales were estimated from inventory, by multiplying the state bottle retail price by the number of bottles sold.

Table 2.1: Sample of Observations from Data

Store Number	Year	Month	Sale (Dollars)	Small	Large	…	Whiskey
4463	2017	1	10317.89	56	50	…	32
4463	2017	2	10824.47	64	56		38
…	…	…	…	…	…	…	…
6064	2020	12	9563.70	85	6	…	29

In order to break down sales volume, we categorized liquors bought in each into small and large sizes. Similarly, we categorized the sale price per bottle into three categories: cheap, average price, and expensive. These categories were created by examining the distribution of bottle size and cost and picking a reasonable value. Using other existing columns, we determined the year, month, number of full packs sold, and number of single bottles sold in excess of whole packs for each transaction. Additionally, we cleaned the liquor's "Category Name" data to create easily accessible and generalizable categories of alcohol such as Gin, Rum, Tequila, Vodka, and more.

[1] https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy

This allowed an analysis of alcohol transactions for each type of alcohol at the storefront in the corresponding year/month. These features provided us with more insights into what drives the most value in monthly sales to optimize a predictive model.

Table 2.2: Existing Feature Aggregates

Feature	**Description**
Store	A unique numeric identifier for the storefront.
Year	The year in which the sales were made
Month	The numeric month in which the sales were made, with each month's numeric label corresponding to the order in which they appear in the calendar year (beginning with 1 = January, and so on until 12 = December)

Table 2.3: Computed Features

Feature	**Description**
Full Packs Sold	The number of full packs sold during the storefront-month-year, estimated based on the number of total bottles sold floor divided by pack size on the transaction level data.
Full Bottles Sold	The number of full bottles sold during the storefront-month-year in excess of full packs, estimated based on the remainder of total bottles sold divided by pack size on the transaction level data.
Small	The number of transactions at the storefront-month-year where the bottles sold were less than 800mL
Large	The number of transactions at the storefront-month-year where the bottles sold were greater than 800mL
Cheap	The number of transactions at the storefront-month-year where the bottles sold were less than $25 per bottle
Mid-Priced	The number of transactions at the storefront-month-year where the bottles sold were between $25 and $50 per bottle
Expensive	The number of transactions at the storefront-month-year where the bottles sold were greater than $50 per bottle
Brandy	The number of transactions at the storefront-month-year where the bottles sold were brandy
Gin	The number of transactions at the storefront-month-year where the

	bottles sold were gins
Rum	The number of transactions at the storefront-month-year where the bottles sold were rums
Schnapps	The number of transactions at the storefront-month-year where the bottles sold were schnapps
Tequila	The number of transactions at the storefront-month-year where the bottles sold were tequilas
Vodka	The number of transactions at the storefront-month-year where the bottles sold were vodkas
Whiskey	The number of transactions at the storefront-month-year where the bottles sold were whiskeys
Other	The number of transactions at the storefront-month-year where the bottles sold were other types of alcohol not categorizable into the groups above

For the purposes of model fitting and validation, each variable was standardized by subtracting the average and dividing by the standard deviation.

III. Model Validation and Selection

We created three main candidate models using a subset of variables from our data preparation and engineering steps, as seen in Table 3.1.

Table 3.1: Models Tested

Model 0	Model 1	Model 2
Sales = Small + Large + Cheap + Mid Priced + Large + Whiskey + Vodka + Rum + Spirits + Brandy + Schnapps + Gin + Tequila + Other Alcohol	Sales = Year + Small + Large + Cheap + Mid Priced + Large + Whiskey + Vodka + Rum + Spirits + Brandy + Schnapps + Gin + Tequila + Other Alcohol	Sales = Month + Small + Large + Cheap + Mid Priced + Large + Whiskey + Vodka + Rum + Spirits + Brandy + Schnapps + Gin + Tequila + Other Alcohol

In each of our models, we ultimately elected to remove variables related to overall sales volume in favor of variables that broke down the sales volume into distinct categorized components related to volume of bottles sold by size, cost, and liquor type. In this way, we gave up using a

simpler and potentially closer fitting model in favor of creating a model that still provides robust predictive power, but also contains deep-level insights into what drives or inhibits sales. Each model represents a different approach to predicting sales. Model 0 predicts sales based on pure volume by type of inventory. Model 1 includes the Year variable to control for trend and provide a long-term outlook, while Model 2 includes the Month variable to control for seasonality and provide a short-term outlook.

Each model was evaluated and scored using a k-fold cross-validation process with 5 folds. In other words, the full dataset was separated into 5 parts, with each part interchangeably acting as a testing data set while the model was fit on the concatenation of the rest of the 4 parts. In this way, we could validate the predictive power multiple times of our model without needing external labeled data, and create an average score across all 5 parts that better reflects the model's true predictive power on future sales.

Furthermore, each model was evaluated using multiple different ridge regression penalties. The ridge penalty term is a model hyperparameter used to modify simple linear regression. This term is included to protect against overfitting the data.

The models were scored during the cross-validation process using three metrics: R-squared, mean squared error (MSE), and mean absolute error (MAE). In our model selection process, we sought to maximize R-squared and minimize MSE and MAE.

Table 3.2: Model Cross-Validation Metrics

Model Number	**Best Penalty Term**	**R-Squared**	**MSE**	**MAE**
0	10	0.675	140175	861
1	1	0.676	140094	861
2	5	0.676	139867	860

IV. Model Summary

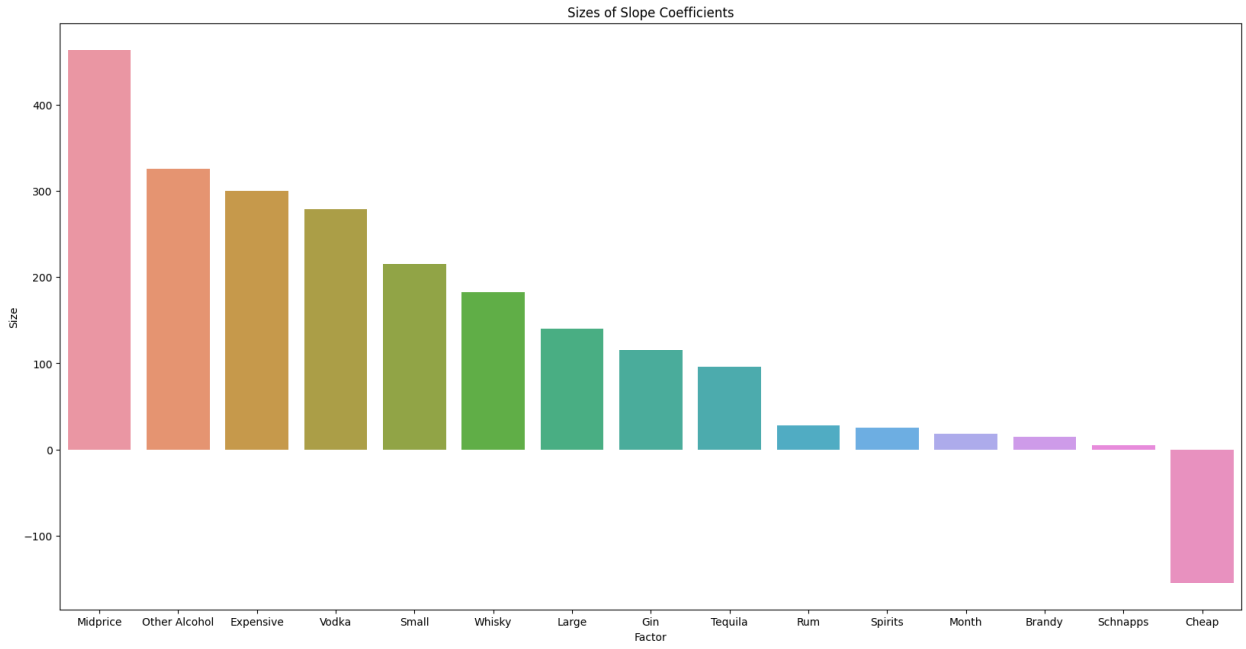
Following a comprehensive analysis, we chose our final model to be Model 2. This choice was driven by the model's notably high R-squared statistic which registered at an impressive 67.6%. This metric implies that 67.6% of the variation in sales can be attributed to the features we selected for our model. Furthermore, Model 2 exhibited superior performance as both Mean Squared Error and Mean Absolute Error were lower than most alternate models. These lower

metrics validate the model's accuracy in predicting sale outcomes. Additionally, this model contains meaningful predictors regarding sales volume of bottle size, cost, and liquor all while accounting for seasonality by including a month predictor. These features provide precise monthly predictions of the total sales for the average storefront.

The final model can be represented as follows, with all variables standardized:

Average Storefront Sales = 3210.22 + 463.40(Mid-Price) + 325.60(Other Alcohol) + 300.11(Expensive) + 278.45(Vodka) + 214.92(Small) + 182.81(Whiskey) - 154.43(Cheap) + 139.74(Large) + 115.83(Gin) + 95.63(Tequila) + 28.39(Rum) + 25.68(Spirits) + 17.94(Month) + 15.07(Brandy) + 4.91(Schnapps)

Figure 4.1: Final Model Coefficients



It is important to note that the interpretation is complex and does not necessarily reflect the real change in predicted sale price as the variables were standardized before modeling. The coefficients reflect the impact of a one standard deviation change in each predictor rather than a one unit change. Still, the relative magnitudes of coefficients reveal which alcohol price points and types drive monthly sales the most.

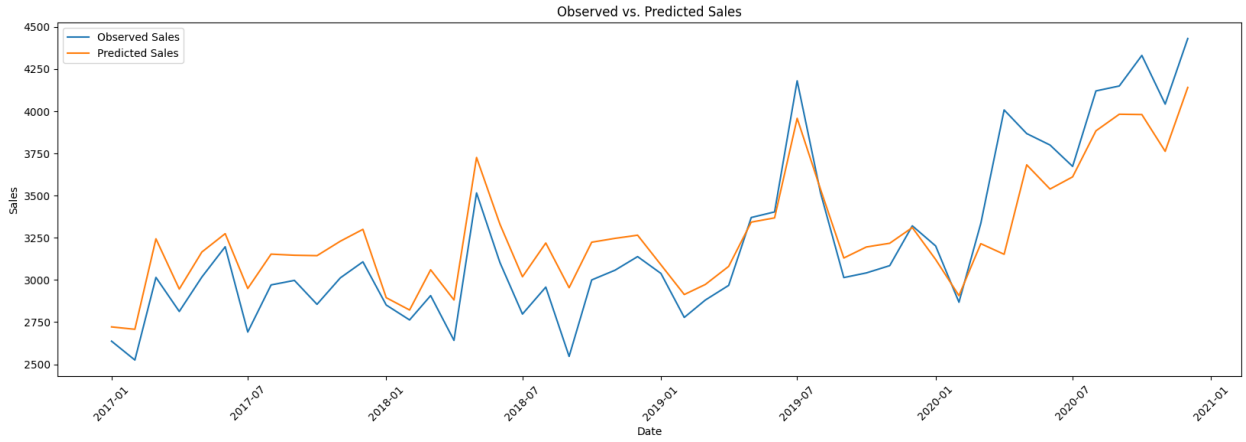
Nearly all of our predictors are categorical variables. This means, for example, the sale can only have one price category (*cheap*, *mid-price*, or *expensive*). Our most impactful attributes are the price category and type of alcohol. The positive coefficients for *Mid-Price* (463.40) and *Expensive* (300.11) compared to the negative coefficient for *Cheap* (-154.43) indicate that moving from cheap alcohol to mid-price or expensive alcohol is associated with a large increase

in average monthly sales. Specifically, selling *Mid-Price* rather than *Cheap* alcohol corresponds to an estimated $463.5 per standardized unit increase in sales on average. Selling *Expensive* rather than *Cheap* alcohol relates to a $300.11 standardized unit increase in average sales. This suggests that customers tend to purchase more alcohol, driving higher total monthly sales, when mid-price or expensive options are sold compared to cheap alcohol. The model captures how moving up the price range for alcohol generally boosts average sales.

V. Conclusions

In Figure 5.1, we can see an example of the accuracy of our predictions, with a root mean square error of $373. This metric serves as an indicator of the accuracy of our predictions, or in other words the size of the average error of our predictions.

Figure 5.1: Forecasted Compared to Observed Average Monthly Storefront Sales



Our model provides valuable insights into the factors that significantly influence sales performance and offers predictions for monthly sales trends. These insights were clearly demonstrated in our case study, where we identified specific sizes, price ranges, and liquor types that serve as the primary drivers of sales. Specifically, based on effect or coefficient size; mid-priced products ranging from $25 to $50, expensive products exceeding $50, and smaller bottle sizes (less than 750 mL) emerged as pivotal contributors to sales growth. Additionally, particular liquor categories, such as Vodka, Whiskey, and other exotic alcohol varieties, played an important role in shaping sales outcomes.

Conversely, our analysis also identified certain growth obstacles within the features used. Notably, cheaper bottles priced below $25 were associated with reduced overall sales performance. This insight highlights the importance of product selection and pricing strategy in optimizing sales and overall business success.

VI. Ethical Concerns

While the provided statistical model provides valuable insights for sales forecasting, there are important ethical considerations. As the providers, and for the benefactors at Booze ‘R’ Us, of this analysis and modeling, we must acknowledge possible limitations and biases in the data and methodology. The historical liquor sales data could contain inherent prejudices and while our analyses have led us to linear regression being the most appropriate solution, there is a chance it will not catch the true complexities of the system. The model predictions simply serve as a guide and should not be the sole basis for any business decision. Managers and analysts must also apply their experience and judgment. Over-reliance on any one source can prompt irresponsible practices surrounding alcohol sales and use. By incorporating ethical considerations into Booze ‘R’ Us’s business practices, predictive modeling can enable decision-making that meets obligations both to Booze ‘R’ Us and the greater public good.

VII. Recommendations

Based on the insights gained from our comprehensive case study conducted with “Casey’s General Store,” we recommend a more strategic focus toward expansion. Our analysis illustrates the significance of emphasizing specific product categories and characteristics that have proven to be primary drivers of sales for this particular franchise.

In particular, we advocate prioritizing the mid-priced ($25 - $50) and expensive ($50 and above) categories, as well as targeted liquor types such as Vodka, Whiskey, and other niche alcohol variants. These have consistently emerged as the key contributors to robust sales performance. Conversely, we advise a more cautious approach towards Cheap Bottles (priced below $25), as our findings indicate that these products tend to lead to lower sales.

Leveraging the predictive capabilities of our model we recommend applying this analytical framework to current Booze ‘R’ Us stores, using proprietary Booze ‘R’ Us monthly storefront sales data. Doing so will provide valuable insights into the feasibility of expansion for specific locations by offering predictions of future sales trends. This model also serves as a potent tool for discerning the unique purchase behaviors of “Booze ‘R’ Us” customers enabling a data-driven approach to inventory management and sales prediction.