Booze 'R' Us Final Report

Hailey Ernest, James Irwin, Elliot Kunz, Kaatje Matthews-vanKoetsveld

2025-10-16

Introduction

This document serves as a final analysis of the work that Team Tortilla has done for Booze 'R' Us. The goal of our collaboration was two-fold: to (1) accurately project sales in the coming year and (2) understand how to expand into other states and adjust sales tactics. Therefore, this report will elucidate the data, techniques, and methodologies that were used to generate predictions and provide relevant insight. It will also cover data preparation, model selection, and performance analysis. Finally, it will summarize our findings and make clear how Booze 'R' Us can best utilize our insights to improve its business.

To begin, we should first contextualize the data that was used for this analysis. The data that we used was provided by the state of Iowa and contains information about wholesale alcohol purchased by Iowa Class "E" liquor licensees. Class "E" licenses are for grocery stores, liquor stores, and convenience stores, among other establishments. The data source is linked here: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy/about_data. We felt that using this dataset would allow us to best understand how different factors affect store performance and make the most accurate predictions of future sales data. We have appreciated the opportunity to work with Booze 'R' Us and are excited to share our findings.

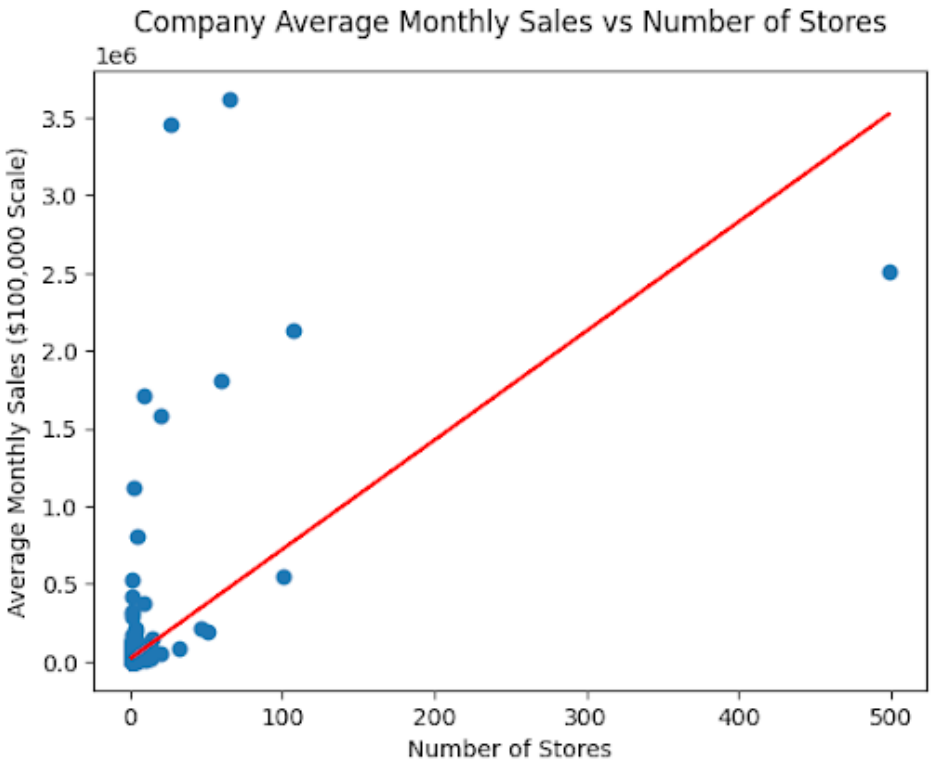
Data Preparation

To ensure the best possible results of our predictive modeling, we cleaned and transformed the data. Our goal during this stage was to create a clear, aggregated, and analysis-ready dataset that would allow us to both accurately predict future alcohol sales and identify factors of increased consumer activity. Furthermore, we only included data from January 1st, 2022 and onward to capture insights that are relevant to today's market. We felt that including any data before this date would hinder our ability to understand what factors are currently driving alcohol sales.

The initial inspection of the dataset revealed inconsistent company names due to variations in labeling, so we addressed this problem to make sure that each record was attributed to the correct parent company. We further processed the data by aggregating individual sales by year and month, which allowed us to conduct monthly trend analysis. We were also able to group the data by company and month to compute key performance indicators of store success, including total and average sales in dollars, total bottles sold, the number of transactions a store has in a month, and the number of stores a company owns. These metrics capture customer purchasing activity, thus allowing us form a complete basis for modeling sales performances by store.

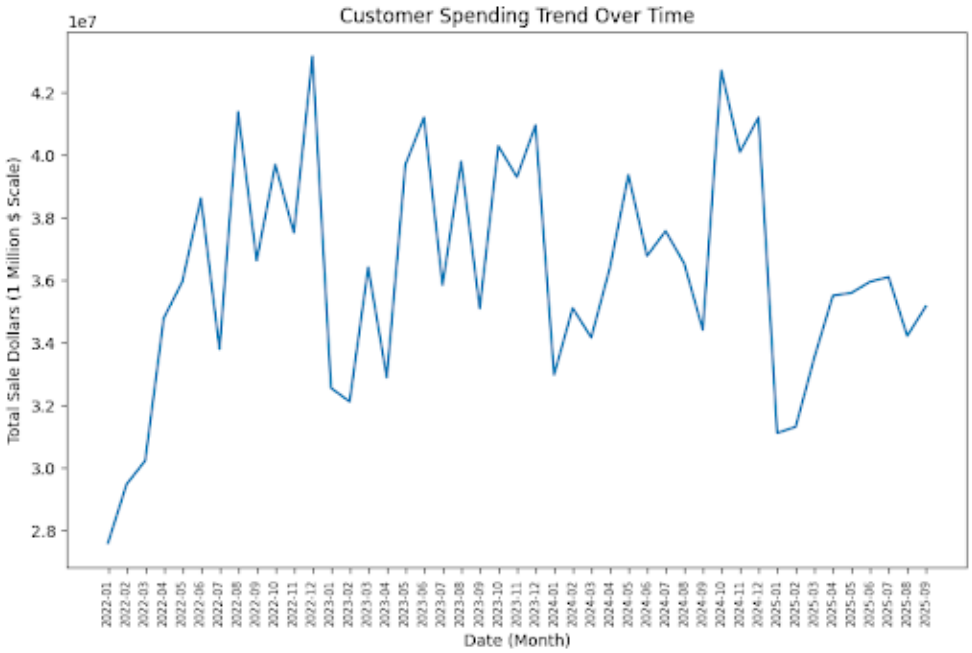
Model Selection Process

The goal of our process was to predict a company's liquor revenue as total sales in dollars per month. As such, we first had to look at the variables that can influence liquor sales. Below is a graph which shows how the average monthly sales of a store are correlated with the number of stores a company chain has, with each point being a liquor chain.

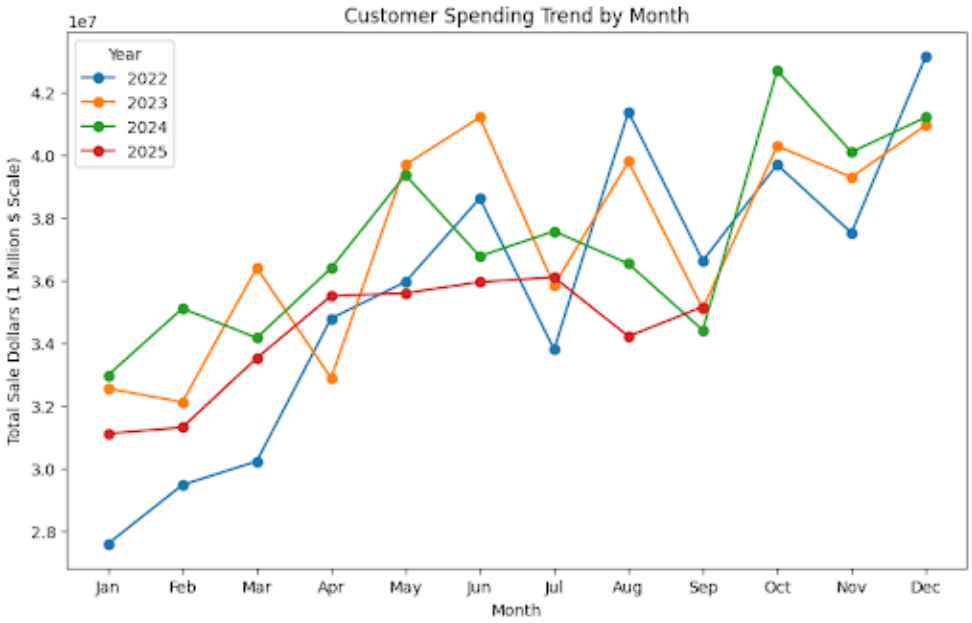


As you can see, there is a positive correlation between the number of stores and the average monthly sales for each store. There is a very uneven distribution in the number of stores; however, we decided that the potential positive correlation was worth adding to our model.

Next, we looked at spending over time from January 2022 to September 2025.



We saw an interesting distinction between the years, with December being a high point and January being a low-selling month for liquor stores. Seeing how each year did have some distinction, we decided it could be good for our model to incorporate the year as a potential variable. This also led us to want to overlay each year on top of each other to find out what other trends we could see on a monthly basis, which is shown in the graph below.



We noticed after overlaying the years separately that there were trends in the graph, with a general positive trend in all 4 years. We observed the most spending later in the year, around the holiday season from October to December. We also noticed a seasonal slump around July with lower sales. Seeing how there is a distinct positive linear fit between months and the total sales, we decided they would be good to fit into our model as they could have high predictive power.

These predictors were chosen because they capture some of the key factors influencing monthly revenue. Including the overall size of each chain (number of stores), long-term trends (year), and seasonal trends (month). Using matrix algebra to create custom linear regression functions, we fit the models. In the process of selecting the right model, we used cross-validation with 10 splits to test how well each model was performing. The metrics we were using to decide which model performed best were R^2 and MSE.

The first model we attempted was predicting revenue from the number of stores the brand owns in Iowa, the year, and the month. This model can account for 28.19% of the variation between company revenues and is, on average, about 211 thousand dollars off of the true revenue in testing. We additionally attempted to fit the same model, only using a subset of companies with at least 5 stores, to better reflect Booze 'R' Us as a large chain. However, that model performed much worse, with an average testing error of 981 thousand dollars. Due to this, we decided to stick with the model using all companies selling liquor in Iowa for the purposes of more accurate predictions. Then, this final model was fitted on the entire dataset with the selected predictors to find and interpret its coefficients and create sales predictions.

Model Summary

The final model achieved a Mean Squared Error (MSE) of $45,416,401,989.66 and a Root Mean Squared Error (RMSE) of $211,503.99, meaning that on average, the model's predictions are off by about $211,000. Since the typical values of sales in the dataset per month for store chains range from $100,000 to $1.5 million, the model's predictive correctness varies based on the store chain per month.

Approximately 28.19% of the differences we see in sales can be explained by the factors of Number of Stores, Year, and Month (R^2). The remaining percentage that isn't explained using these factors and should be assessed in further analysis includes county characteristics and demographics, competitor chains, and pricing by county. The combination of these findings suggests that it captures patterns and trends for predictive uses, but it tends to make errors in these predictions. How accurately we can predict may be inadequate for expansion predictions, although it does provide an understanding of the influence of store counts and seasonal changes on sales.

The coefficients for the final linear regression model are as follows:

Table 1: Model Coefficients

Feature	Coefficient
intercept	4,435,427.81
number of stores	7,191.87
year	-2,179.89
month	803.48

Conclusions

Our analysis showed that the number of stores a chain has and seasonal trends influence liquor sales across Iowa. The model found important relationships between store count and sales, showing that larger chains tend to make higher profits. Also, we found seasonal change trends across the years with more sales during holiday seasons and some summer months, and much less sales after the holidays. These findings give an understanding of how expansion and timing affect sales performances, while considering the large variation in the data does suggests that other factors also affect sales.

This model is best for identifying and predicting broad patterns and performance relative to individual store sales. We recommend using these to understand general relationships between chain sizes and sales and to plan according to the seasonal demand peaks and lows with inventory. The positive correlation between the number of stores and average monthly sales for each store is a useful finding when evaluating the expansion of chains.

Our model provides a strong start to planning and strategy on expansion and seasonal sales. To make detailed expansion decisions, we recommend combining this analysis with additional characteristics of the population of each county. This strategy will help Booze R Us maximize the success of future expansion and predictions.