Unveiling Insights

Analyzing Factors Influencing Alcohol Purchases in Iowa for DEAD

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

Drinking Excess Alcohol is Dangerous (DEAD) is interested in the driving factors behind small and large alcohol purchases in Iowa. We proposed a machine learning model fitting a multiple linear regression model to a random sample of past sales data. The model fits sales data and predicts the quantity of alcohol purchased. The features of the final model represent the most influential factors driving alcohol purchasing behavior. This investigation has revealed valuable insights that can guide DEAD in shaping responsible alcohol retailing initiatives, with a focus on seasonal, category-specific, and college town-targeted campaigns, while upholding ethical principles of transparency and data privacy.

I. Introduction

Your organization, Drinking Excess Alcohol is Dangerous (DEAD), has undertaken a mission to uncover the driving factors behind both small and large alcohol purchases in this state. To achieve this, we proposed a powerful solution: the implementation of a learning model that employs a multiple linear regression approach. By analyzing a random sample of historical sales data, our model not only fits the sales patterns but also predicts the quantities of alcohol purchased. The most important features extracted from this model provide a profound understanding of the key influences behind alcohol purchasing behavior in Iowa.

II. Data Preparation

To analyze patterns in sales to determined what factors contribute to higher or lower alcohol purchases, we collected a random sample of 100k alcohol purchases made in 2022 and later from the Iowa Liquor Sales dataset managed by the Iowa Department of Revenue, Alcoholic Beverages[1]. This dataset contains information on purchases of alcohol made by stores in Iowa, as well as location data on the stores. As shown in Figure 2.1, initial investigation into this data showed that over half of the purchases made were considered large, greater than the median purchase of 4.5 sale liters, alcohol purchases. After thoroughly cleaning the collected data, we created features for possible factors that drive alcohol purchases, such as the size of the college student population in each city, what type of alcohol was purchased, what day of the week the purchase was made, and whether the purchase was made during a holiday, as seen below in Table 2.1.

Figure 2.1: Proportion of Alcohol Order Size

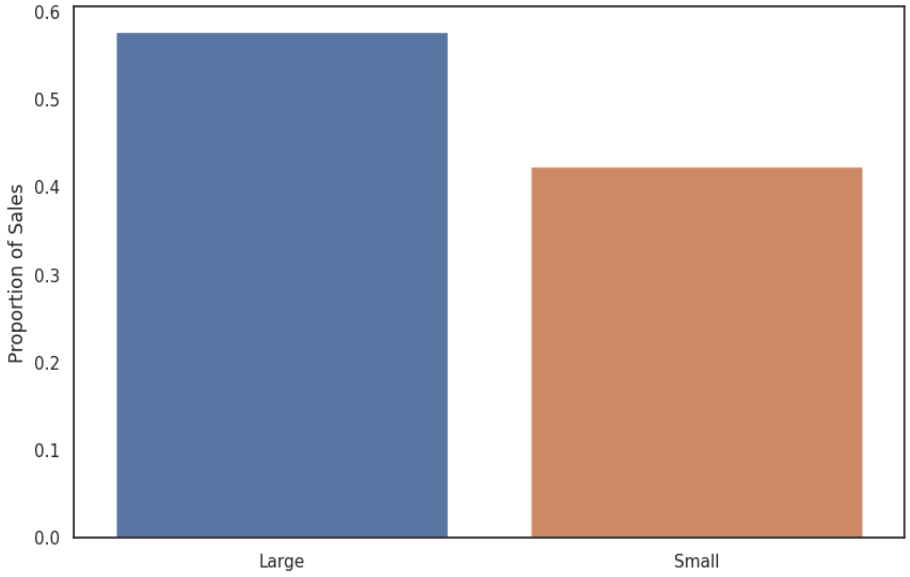


Table 2.1: Sample of Observations from Data

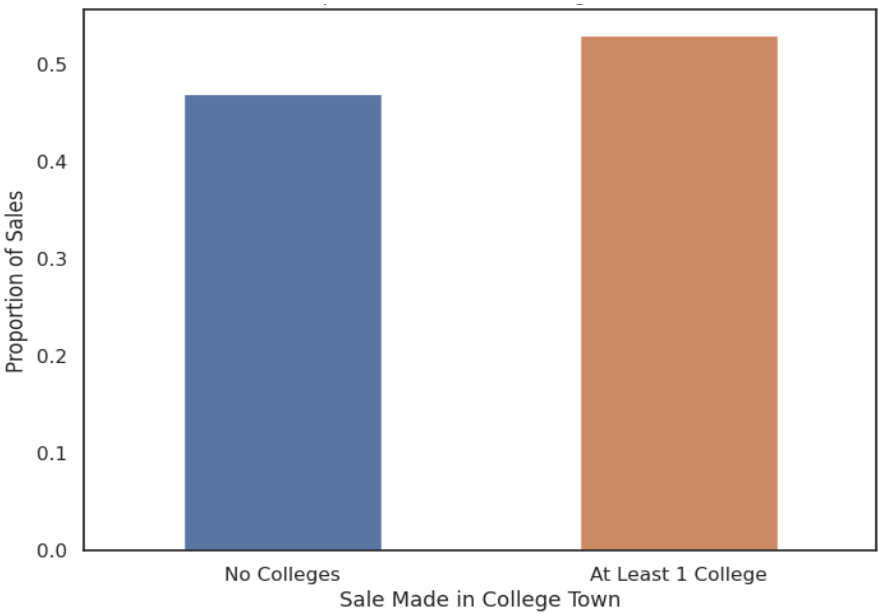
Date	Month	Day of Week	NearHoliday	Category	…	Sale Amount
2023-04-13	4	4	0	GIN	…	Small
2023-04-13	4	4	0	VODKA	…	Large
2022-08-16	8	2	0	VODKA	…	Large

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

We are motivated to include features related to holiday alcohol purchases due to the significant impact of holidays on alcohol consumption, which can often lead to excessive drinking. As mentioned by the American Addiction Centers Organization, holidays are occasions for fun and celebration, and alcohol plays a prominent role in the festivities. However, this combination of good cheer and abundant alcohol can lead to binge drinking during holidays, becoming a major law enforcement and public health concern. Statistics and data reveal that holiday-related drinking can result in negative consequences. For instance, the holiday season, including Thanksgiving, Christmas, and the Fourth of July, can increase high-risk drinking, contributing to a dangerous and often deadly series of risk factors, such as more people driving late at night and in adverse weather conditions.

Another important motivation for creating features related to the size of the college student population in each city is the well-documented issue of harmful and underage college drinking, as highlighted by the National Institute of Health[2]. College drinking has become a common practice among students and is often seen as an integral part of the higher education experience. The college environment can influence established drinking habits and lead to problematic drinking behaviors. According to the 2021 National Survey on Drug Use and Health (NSDUH), a significant percentage of full-time college students engage in alcohol consumption and binge drinking[3]. This data emphasizes the importance of exploring how the presence of a significant student population in a city may influence alcohol sales patterns.

Figure 2.2: Proportion of Sales in College Town



[2] https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/college-drinking
[3] https://www.samhsa.gov/data/release/2021-national-survey-drug-use-and-health-nsduh-releases

In Figure 2.2, shown above, it can be observed that over half of a random sample of purchases are made in a “College Town”; defined as having one or more collegiate institutions in the same town. It is imperative to investigate this factor when classifying high or low alcohol purchases because college towns represent a specific subset of locations where alcohol sales might be significantly influenced by the presence of a collegiate population. College students, known for higher alcohol consumption rates, and their unique events and preferences can contribute to increased alcohol sales in these areas. By considering whether a purchase occurs in a college town, the analysis can better account for the distinct dynamics and behaviors at play in such locations, thereby enhancing the precision of high and low alcohol purchase classification and the overall understanding of alcohol consumption patterns within the dataset.

III. Model Selection and Validation

The models we fit are various Ridge Regression models, which are enriched with a penalization term for enhanced accuracy when identifying the most influential features driving a purchase, composed of combinations of the following variables:

> Month, Day of the Week, City, County, Category (type of alcohol), Near Holiday, Student Population (and relative size), Institution (college count in city), and Cost Per Liter

Table 3.1: Main Models Tested

Model Number	Features in Model
1	Category, NearHoliday, Cost Per Liter, Student Population
2	Category, NearHoliday, Cost Per Liter, Institution
3	Category, NearHoliday, Cost Per Liter, Student Population, Day of Week
4	Category, NearHoliday, Cost Per Liter, Institution, Student Population, County, Month
5	Category, NearHoliday, Cost Per Liter, Student Population, Month

We standardized the quantitative variables (Student Population and Cost Per Liter) to be able to compare the coefficients of the factors as a way of measuring their importance in the model. Utilizing K-Fold Cross Validation with 5 folds, we trained multiple models and assessed their predictive capabilities on the quantity of alcohol purchased. 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. The scores we compared were R-Squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The top 5 models ranked by R-Squared, as seen in Table 3.1 and Table 3.2, were all composed of Category, NearHoliday, and Cost Per Liter in addition to other combinations of factors.

Table 3.2: Test Metrics of Main Models

Model Number	Ridge Penalty Term	R-Squared	MSE	MAE
1	100	0.007324	1440.438900	7.888863
2	100	0.007701	1440.009767	7.893249
3	100	0.006961	1440.952302	7.904119
4	100	0.006595	1441.478604	8.145308
5	1000	0.005087	1443.624394	8.044913

IV. Model Summary

We chose our final model based on the largest R-Squared value and what factors it is composed of. Models 1 and 2 achieved similar R-Squared values, and differed only in a single factor; while Model 1 included Student Population, Model 2 included Institution. Model 2 was selected since "Institution" is likely to be a better predictor for the store's sales due to its direct relevance to the store's location, potential stability over time, local impact, niche market appeal, and reduced risk of collinearity compared to "Student Population"; which might potentially be correlated with some of the other predictors such as "NearHoliday" as it relates to the academic calendar, which could indirectly affect the local student population. These characteristics make it a more robust and relevant factor for modeling the store's sales, providing greater confidence in the model's ability to explain and predict store purchases.

Additionally, the stability over time of the "Institution" factor is a key advantage. When using "Student Population," which is based on a single year's data (2012), there's a risk of missing the broader trends or seasonality that might influence store sales. If student populations fluctuate significantly from year to year, it could lead to inconsistencies in the model's predictions. "Institution," on the other hand, is more likely to remain relatively constant, making it a more reliable predictor over time.

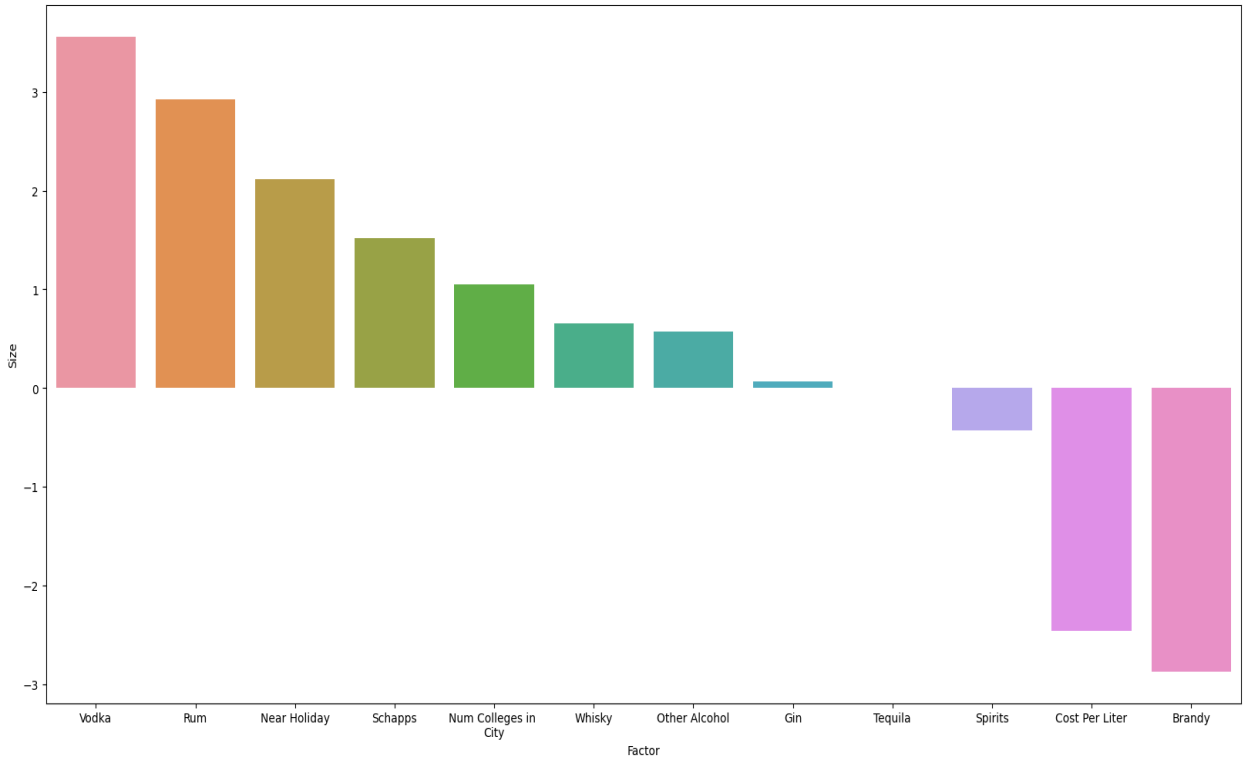
V. Conclusions

Our model provides valuable insights into the factors that significantly influence the amount of alcohol purchased. Below in Figure 5.1, we can see that the factors Vodka, Rum, and Near Holiday all increase the predicted amount of alcohol being purchased by larger amounts than other predictors. The interpretation of these factors is as follows:

- When the type of alcohol being purchased by a store is Vodka, the expected amount of liters being purchased increases by 3.56 liters.
- When the type of alcohol being purchased by a store is Rum, the expected amount of liters being purchased increases by 2.93 liters.
- When the store orders alcohol within two weeks prior to a holiday, the expected amount of liters being purchased increases by 2.11 liters.

Conversely, our analysis also identified certain factors that decreased the amount of alcohol being purchased. Notably, purchases of Brandy tend to signify smaller quantities. Additionally, as the cost per liter of a purchase rises, the predicted amount of alcohol being purchased decreases.

Figure 5.1: Final Model Coefficients



VI. Ethical Concerns

In the context of analyzing store purchases rather than individual customer data, ethical concerns still arise, and a duty-based ethical approach is highly relevant. One primary ethical consideration is the duty to use the insights gained from this analysis responsibly and in the best interests of the community. While the primary duty of "Drinking Excess Alcohol is Dangerous (DEAD)" is to understand the factors behind store alcohol purchases, there's a responsibility to ensure that these insights are not exploited to encourage excessive alcohol consumption or harm to the community. The organization must consider its duty to the well-being of society and prioritize actions that promote responsible alcohol retailing.

Furthermore, there is a duty to safeguard the privacy of individual stores and businesses that contributed to the historical sales data. Sharing or analyzing this data without their consent or without respecting their privacy preferences can raise ethical concerns. DEAD should ensure that our data usage is in compliance with legal and ethical standards, and that the analysis respects the confidentiality and proprietary nature of the data provided by stores when or if publishing findings. In the application of a learning model, there is also a duty to be transparent about the model's methodologies, assumptions, and limitations, such as our limited ability to capture variation in sale liters based on our predictors. This transparency is vital to ensure that the analysis can be independently assessed and that any potential biases or limitations are openly acknowledged.

Ultimately, the ethical duty in this scenario is not only to understand the driving factors behind store alcohol purchases but also to apply this knowledge in a responsible and ethical manner that prioritizes the well-being of society and respects the rights and interests of all stakeholders involved, including stores and the broader community.

VII. Recommendations

Based on the analysis, DEAD should strategically target campaigns to promote responsible and safe alcohol consumption in Iowa. First and foremost, the organization should launch seasonal campaigns that focus on responsible drinking during holidays with high-risk drinking patterns, such as Thanksgiving, Christmas, and the Fourth of July. These initiatives should emphasize moderation and safety, collaborating with local authorities, law enforcement, and public health agencies to raise awareness about the potential risks associated with excessive alcohol consumption during these times.

Additionally, DEAD should tailor campaigns specifically for college towns, where students are known for higher alcohol consumption rates. Collaborating with educational institutions and student organizations, these campaigns should focus on alcohol education, harm reduction, and

the consequences of excessive drinking to create a culture of moderation and safety. Furthermore, considering the impact of specific alcohol categories on purchase quantities, such as Vodka and Rum, DEAD can design category-specific campaigns that encourage informed and responsible choices. These campaigns should provide educational materials, tips for responsible drinking, and information on this specific alcohol content. Additionally, DEAD should run initiatives to raise awareness about the financial implications of alcohol consumption, given that a higher cost per liter tends to decrease predicted purchase quantities, encouraging individuals to make more conscious decisions about their drinking habits. Lastly, the organization should continuously assess the impact of its efforts, adapting strategies as needed, and being open to feedback from the community and stakeholders to better serve its mission of promoting responsible alcohol retailing and safe consumption.