Project Proposal:
Drinking Excess Alcohol is Dangerous (DEAD)

October 13, 2023

Submitted By:
Andrew Kerr - adkerr@calpoly.edu
Bella M - imccarty@calpoly.edu
Erik Luu - eeluu@calpoly.edu
Martin Hsu - mshsu@calpoly.edu
Matteo S - mshafe01@calpoly.edu

Drinking Excess Alcohol is Dangerous (DEAD) is interested in the driving factors behind small and large alcohol purchases in Iowa. We propose a machine learning model fitting a multiple linear regression model to a random sample of past sales data. The model will fit sales data and predict the quantity of alcohol sold. The features of the final model will represent the most influential factors driving alcohol purchasing behavior.

Data Collection

To analyze patterns in sales to determine what factors contribute to higher or lower alcohol purchases, we will collect a random sample of 100k alcohol purchases made in 2022 and later

Figure 1:
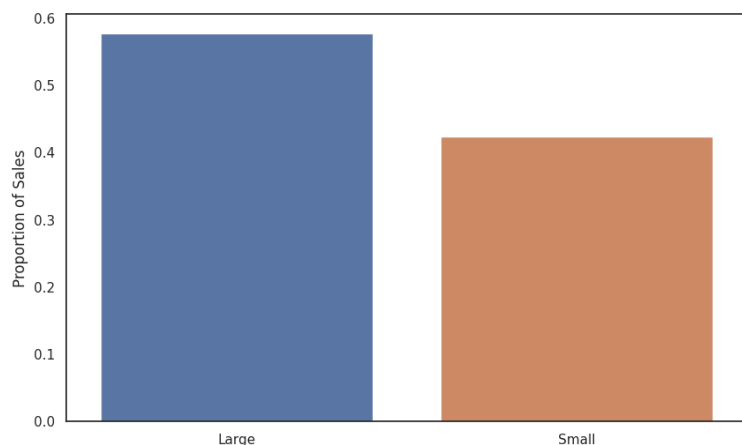
from the Iowa Liquor Sales dataset managed by the Iowa Department of Revenue, Alcoholic Beverages[1]. As shown in figure 1, initial investigation into this data shows that over half of the purchases made were considered large, greater than the median purchase of 4.5 sale liters, alcohol purchases. After thoroughly cleaning the collected data, we will create features for possible factors that drive alcohol purchases, such as the size of the college student population in each city, what type of alcohol

purchases, such as the size of the college student population in each city, what type of alcohol

---

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

was purchased, what day of the week the purchase was made, and whether the purchase was made during a holiday.

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:
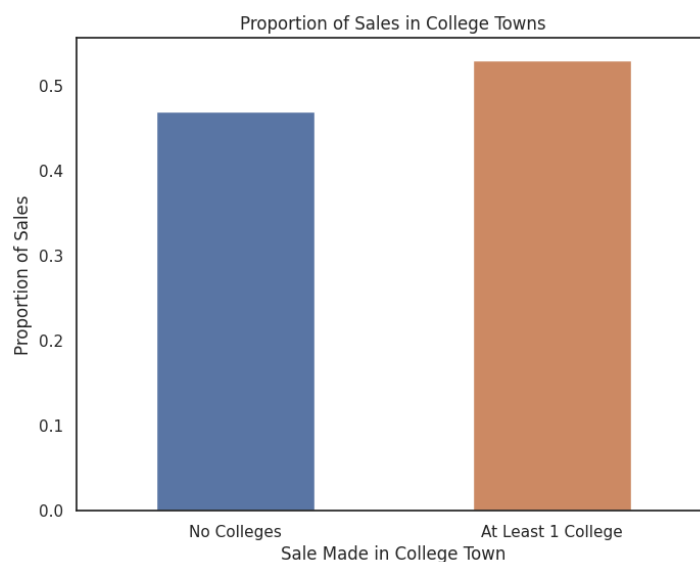
In Figure 2 shown to the left, it can be observed that over half of a random sample of purchases are made in a "College Town"; defined as having one or more collegiate institutions in the same town. It is imperative to investigate this factor when classifying high or low alcohol purchases because college 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

[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

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.

Modeling Process

The models we currently plan to fit are various Least Squares Regression models with optimized loss functions and necessary penalties composed of combinations of the following variables:

> Month, Day of the Week, City, County, Type of Alcohol, Sale Amount (in dollars), Near Holiday, Student Population (and relative size), College Count (in city), and Cost Per Liter

Utilizing K-Fold Cross Validation, we will train multiple models and assess their predictive capabilities on whether a purchase was a small or large amount of alcohol. To do so, we will compare various validation metrics, such as Mean Squared Error (MSE).

Project Outcomes

Our endeavor centers on the identification of optimal predictive features for distinguishing high and low alcohol purchases. Employing advanced data analysis methodologies, we will conduct an exhaustive examination of the dataset to pinpoint the most influential drivers of alcohol purchasing behavior. The presentation of our findings will be characterized by visually informative, professionally formatted outputs, including charts, tables, and succinct summaries, ensuring accessibility to a wide range of stakeholders, from executives to the analytical team. To provide a broader context for our results, we will incorporate relevant references, enabling the client's organization to make decisions rooted in data-backed insights.

Our system shall determine the most important features through sophisticated techniques such as feature importance ranking, correlation analysis, and predictive modeling. Once identified, our system shall employ these features to predict the nature of the alcohol purchase effectively.
The client-centric approach goes beyond technicalities. We shall employ advanced validation, ensuring the system we create aligns with the client's business objectives. The validation metrics we choose will serve as a measure of our models' accuracy in real-world applications. Our ultimate goal is to equip the client's organization with insights that transform decision-making about alcohol purchasing behaviors. By identifying influential factors affecting alcohol purchasing behavior, our system shall empower the client to optimize strategies and achieve client objectives. We eagerly anticipate the opportunity to partner with the client on this data-driven journey.