Analyzing Liquor Sales for Consumption Trends

Drinking Excess Alcohol is Dangerous (D.E.A.D)

OCTOBER 2023

ISSUED BY
CALIFORNIA POLYTECHNIC STATE UNIVERSITY SAN LUIS OBISPO

CONSULTANTS
BRENDAN CALLENDER
JADYN ELLIS
KYLE LEW
SOREN PATEAU
ANAGHA SIKHA

Introduction

Your company (Drinking Excess Alcohol is Dangerous) is looking to change drinking culture in Iowa for the better. The problem your company faces is identifying how, where, and when to focus your campaigning efforts in order to have the biggest effect. In an effort to address this problem, our consultation group was hired by your company to help you understand which factors drive higher and lower purchases of alcohol. Throughout the analysis process, our team went through rounds of data preparation and exploration, eventually convening our findings into one final statistical model.

Background

Throughout our process, the main dataset (Iowa Liquor Sales) we used for data exploration was pulled from the Iowa Department of Revenue, Alcoholic Beverages section. This dataset provided information describing each purchase of alcohol for liquor stores, including various information on the type and amount of alcohol purchased.

Throughout our exploration, we used 'Volume Sold (Gallons)' as our response variable. This variable tracked the total volume of alcohol per liquor store order, rather than customer purchases. With the assumption that the volume of alcohol purchased by liquor stores is heavily associated with the volume of alcohol being sold at the given liquor store, we used this response variable as a marker for drinking rates.

Data Preparation

Before our model tuning process, we prepped our data by creating multiple potential explanatory variables that we thought may have an effect on alcohol consumption and therefore may be included in our final model.

Holidays

Firstly, we explored how holidays affect alcohol purchases thinking people may drink more during holidays. To do this we used a dataset with government-designated holidays[1], looking at store alcohol purchases 10 days before a holiday. We used this '10 Days Before Holiday' variable as one of our explanatory variables.

Population

We also wanted to explore city populations, so we included the Iowa City[2] dataset which provided information about the populations for different cities in Iowa. With this variable we grouped cities into population brackets and used this population bracket variable in our model.

Alcohol Category

Due to the multiple unique alcohol types, we cleaned the data by creating larger groups of each alcohol category. These categories included but were not limited to: Whiskey, Vodka, Tequila, Gin, and Rum.

Day of Week

We created a 'Day of Week' variable to explore how each day of the week affected alcohol purchases. We thought this would be an interesting variable to study, thinking people may tend to buy more alcohol on specific days of the week.

Weekend vs Weekday

We created a 'Weekend vs Weekday variable' in order to explore whether people may tend to buy more alcohol on weekends compared to weekdays.

Major College Town

The two largest public universities in Iowa are Iowa State University and University of Iowa which are in Ames and Iowa City respectively. We created a binary variable to indicate whether the liquor store is in one of these college cities/towns, thinking the large populations of college students may lead to a trend of higher alcohol consumption.

---

[1] Python Package Index - PyPI. (n.d.). Python Software Foundation. Retrieved from https://pypi.org/

[2] Retreived from https://www.iowa-demographics.com/cities_by_population

Season

We created a season variable to determine the season in which a liquor store purchased alcohol and whether there are seasonal increases and decreases in alcohol consumption.

Model Selection

While creating this model, our focus was to find variables that were the most significant in predicting volume of alcohol purchased by stores in gallons. Since we have multiple features we want to focus on, we used ridge regression with a sum of squared error plus a lambda ridge penalty to prevent overfitting in our model. This prevents the model from putting more weight on certain features, and allows the model to focus on all the variables included. We used a 5 fold cross-validation tuning process to find the lambda value that provided the highest R^2.

The first model decision we had to make was how to group values. When we first ran our model without grouping, we found a low R^2. To improve the amount of variation the model explained, we grouped our values by city and date. This meant each of our observations was the total alcohol purchased within a city on a given date, rather than for each liquor store. This grouping strategy resulted in a better R^2, leading that to be our final grouping in our model.

The next decisions we made for our model was which explanatory variables to include. Using R^2 as our metric, we were able to decide which variables to add to our final model. Our population bracket variable turned out to result in a higher R^2 compared to using the pure population counts. Using 'Day of Week' instead of 'Weekend vs Weekday' was another choice we made to optimize our R^2. And through the same process we found that including '10 Days Before Holidays' turned out to be a better account of variance compared to using 'Holiday'.

Final Model

Our final selected model is a Ridge Regression model with $\lambda = 100$ and includes the following explanatory variables on Volume Sold (Gallons): '10 Days Before Holiday', 'Major College Town', 'Day of Week', 'Season', and 'Population Bracket'. This model was selected due to its high yield of an R^2 in comparison to our other potential models. We also considered that it would be best to include population brackets in our model, since we would be able to interpret our other coefficients while holding population constant, which made this model even more useful for D.E.A.D's interest. In other words, it is intuitive that we would see increases in alcohol volume purchased by stores in larger cities since they have higher numbers of stores and larger numbers of people to sell to. Utilizing $\lambda = 100$ and these explanatory variables, we found a final **R^2 = 43%**, meaning our model was able to explain 43% of the variation in volume of alcohol in gallons purchased by stores.

Our model coefficients in table and plot form and interpretations can be found below:

Table 1. Intercept

Intercept
25.88

The expected volume of alcohol in a non-college town with a population of less than 1,000 people purchased on a Friday that is not 10 days within a holiday in the fall is 25.88 gallons.

This resulting intercept is quite specific in its interpretation, so we do not have any recommendations to provide for your company utilizing results of this intercept.

Table 2. 10 Days Before Holiday

10 Days Before Holiday
20.02

The mean predicted volumes of alcohol purchased 10 days before a holiday is 20.02 gallons higher compared to the mean predicted volumes of alcohol purchased not 10 days before a holiday, holding all other predictors constant.

From our results above, we see that stores purchase more volume of alcohol intake before holidays, so it would be recommended to hold more campaign efforts close to major holidays, such as Christmas, 4th of July, Thanksgiving, etc. However, the overall effect of an increase in 20.02 gallons of alcohol purchased, holding other predictors constant, 10 days before a holiday versus days that are not 10 days before a holiday is not too large.

Table 3. Major College Town

Major College Town
-186.59

The mean predicted volumes of alcohol purchased for stores in a major college town is 186.59 gallons lower compared to the mean predicted volumes of alcohol purchased for stores not in a major college town, after adjusting for other variables in the model.

Overall we found that stores in college towns are surprisingly associated with decreases in expected total alcohol volume purchased compared to non-college towns, holding other variables in the model constant. We should mention that this relationship may be due to other confounding variables in the model, as we cannot conclude causation in our particular model. Intuitively, it would not be smart to stop campaigns in college towns due to common college drinking cultures. That being said, this shows that efforts should be put in place to campaign in cities that do not contain college towns, as there is a predicted increase in the gallons of alcohol purchased.

Table 4. Population Bracket: ***Baseline Population Bracket 1 (0-1000)***

Pop. Bracket 2 1,000 - 10,000	Pop. Bracket 3 10,000-20,000	Pop. Bracket 4 20,000 - 50,000	Pop. Bracket 5 50,000- 100,000	Pop. Bracket 6 100,000 +
91.14	300.22	468.03	1054.06	1785.16

Note: The below interpretation is for Population Bracket 6, but can be tailored for each coefficient above

The expected volume of alcohol purchased by a store on a given day in a city with a population greater than 100,000 people is 1,785.16 gallons higher than the expected volume of alcohol purchased in a city with a population less than 1,000.

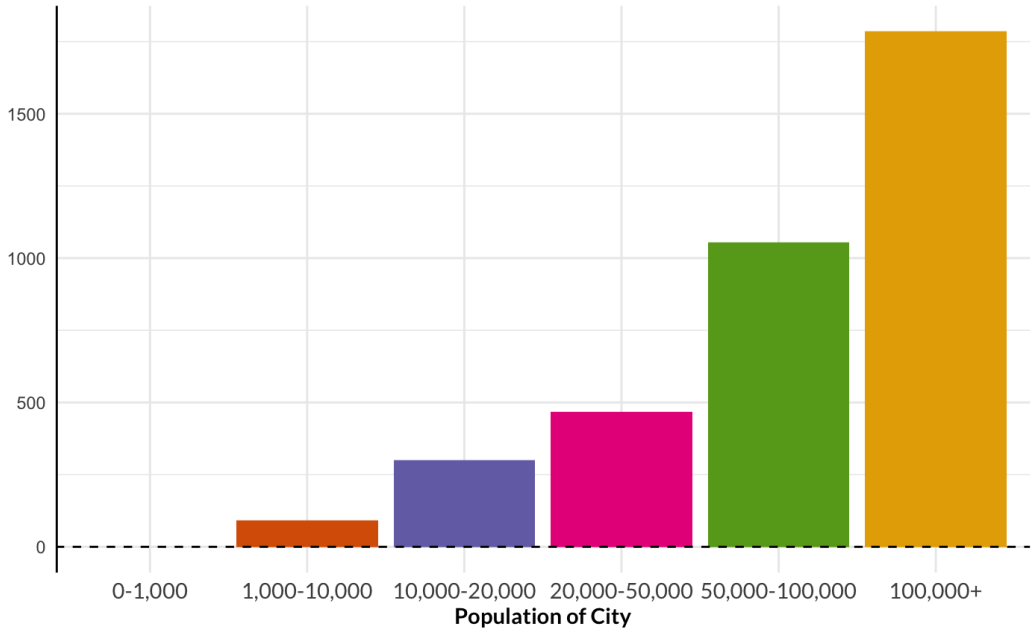

Figure 1. Magnitude of population bracket coefficients

The results seen above are intuitive. Cities with larger populations will tend to have more stores, and in turn, more volume of alcohol purchased by stores. If not already in place, we would suggest that D.E.A.D hold more campaigns in larger cities in order to achieve more outreach in places that have more alcohol at their disposal. However, we will not suggest that outreach efforts are ignored in smaller towns.

Table 5: Season - *Baseline Fall*

Spring	Summer	Winter
4.18	3.03	-12.77

Note: The below interpretation is for Winter, but can be tailored for each coefficient above

The expected volume of alcohol purchased by a store on a given day in winter is 12.77 gallons lower than the expected volume of alcohol purchased on a given day in fall.

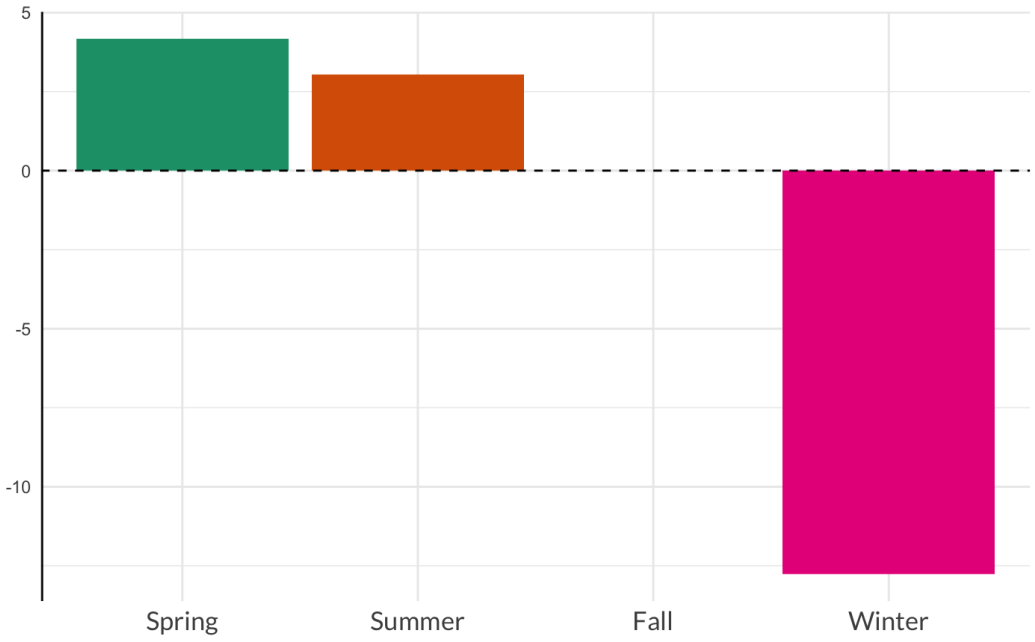

Figure 2. Magnitude of season coefficients

As seen above, winter purchases are associated with decreases in total volume purchased relative to other seasons. There could be a decrease in populations over winter seasons due to holiday travel along with other possibilities, resulting in stores decreasing their intake of alcohol inventory. We would suggest focusing more on anti-excessive-drinking outreach during other seasons, particularly spring, which is associated with larger increases in volume of alcohol purchased compared to other seasons.

Table 6: Day of Week - *Baseline Friday*

Monday	Tuesday	Wednesday	Thursday	Saturday	Sunday
23.39	23.38	82.37	167.85	-0.74	-494.49

Note: The below interpretation is for Thursday, but can be tailored for each coefficient above

The mean predicted volumes of alcohol purchased for stores on a Thursday is 167.85 gallons higher compared to the mean predicted volumes of alcohol purchased on a Friday, after adjusting for other variables in the model.

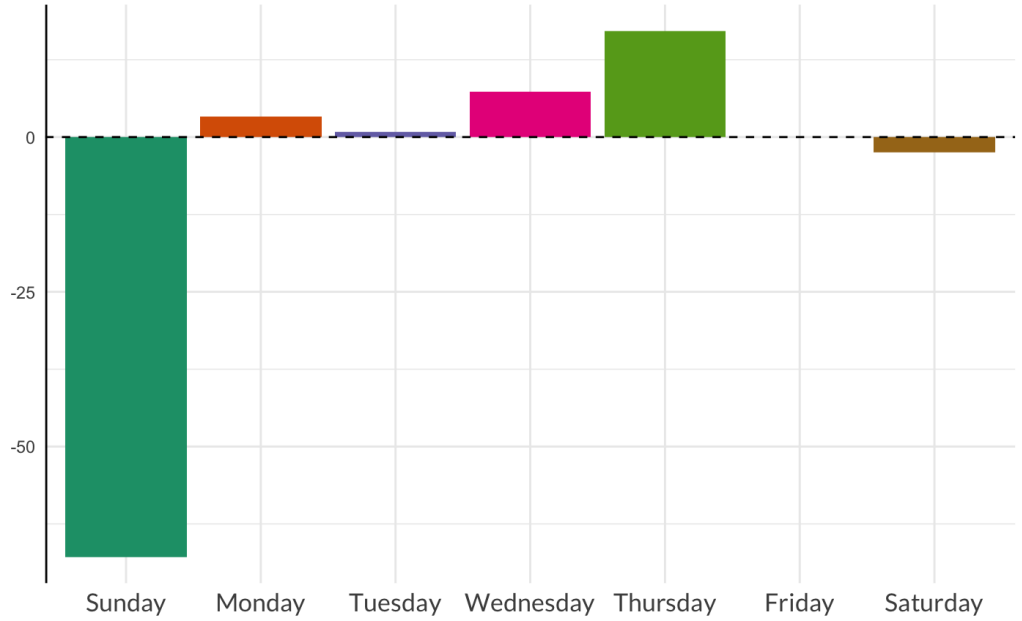

Figure 3. Magnitude of day of week coefficients

We discovered that Thursday orders are associated with increases in predicted total volume of alcohol purchased while Sundays are associated with decreases in predicted total volume of alcohol purchased. This suggests potentially to move campaigns for preventing excessive drinking to Thursdays, when alcohol stores are gearing up for the weekend rush, in order to be timely in efforts. It would not be as effective in the interests of prohibiting excessive drinking to hold campaigns on Sundays, as stores are not purchasing as much alcohol that will be seen in store in the coming days.

Additional Exploration

With our final model already built, we decided to do some preliminary additional analysis to show the potential uses beyond our initial analysis. Using the same model, we grouped by the 3 most popular alcohol types (*Vodka, Whiskey, and Tequila*). From this slight alteration, we found some interesting interaction effects between type of alcohol and other variables. During spring and winter, whiskey has a positive change in expected total volume of alcohol ordered while vodka and tequila have a negative change. During the 10 Days Before a Holiday, whiskey and vodka are associated with higher expected total volumes of alcohol ordered compared to tequila. While we didn't do a deep dive into this model, these initial findings point to the potential future uses of our model. Beyond looking at the variable effects on alcohol consumption, you could also use our model and group by certain variables to analyze interaction effects, opening up the possibility of many future analyses.

Project Takeaways

How to use our model:

We suggest you focus your attention towards the model coefficients and our recommendations based on the results (See Final Model). These coefficients offer quantifiable values explaining the effects for each variable on alcohol consumption. Using these coefficients, you will be able to get a better sense of which explanatory variables most contribute to drinking culture. It is important to keep in mind that this model should not be used to shame liquor stores or consumers, and instead be used to increase awareness around dangerous drinking. With this given insight on the overall trends of alcohol consumption across the state of Iowa, we hope you can better decide where, when, and how to focus D.E.A.D's efforts to change drinking culture.

Cautions with our model:

Our final model was created with the primary goal of analyzing drinking patterns and was not created to predict future values. This means that if our model was used to predict the

alcohol volume sold for specific liquor stores, we cannot promise the validity of these predictions and advice against using our model this way. Similarly, the significant effects we found on 'Volume Sold (Gallons)' should be used solely as exploration values and should not be analyzed as cause and effect relationships.