Drinking Excess Alcohol is Dangerous 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 Drinking Excess Alcohol is Dangerous (DEAD). The goal of our collaboration is clear: to analyze patterns in alcohol sales, understand what factors influence alcohol purchases, and make drinking culture safer. 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 DEAD can best utilize our insights to make drinking safer in Iowa.

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. Using this dataset would allow us to best understand how different factors affect alcohol purchases and make specific, impactful recommendations to increase alcohol safety.

Additionally, we collected Iowa state population data from 2010 to 2025. To use this dataset for our analysis, we aggregated the population by county and joined this dataset with the alcohol wholesale dataset that is described above. The population data source is linked here: https://catalog.data.gov/dataset/city-population-in-iowa-by-county-and-year

We have appreciated the opportunity to work with DEAD and are excited to share our findings.

Data Preparation

Before making models or conducting statistical analysis, we prepared the data so that we could identify factors associated with excessive and dangerous drinking. Both the wholesale and population datasets described above required targeted cleaning and preprocessing to yield the best possible results.

For the population data, we kept only the relevant columns: county, city, year, and population estimate. Keeping only these columns allowed us to understand population data without losing focus on the wholesale data, which is what drives this analysis. We also converted the year to a standardized date format so that we could conduct time-based analysis. Then, we aggregated the data by county and year to obtain the annual population estimates for each county.

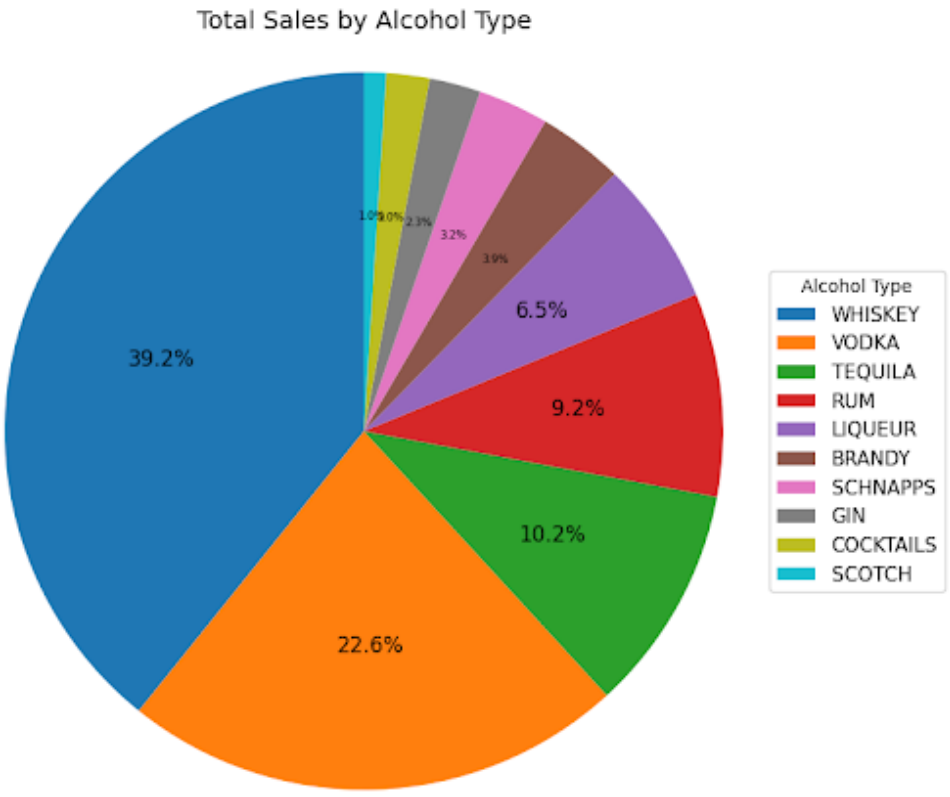
For the wholesale data, we used feature engineering to make the data more interpretable and reduce complexity. Sales were grouped into broader alcohol types (whiskey, vodka, rum, tequila, etc) using string detection, which allowed for comparisons across beverage types. We felt that this was an important step because understanding what kinds of alcohols are selling (and which are not) can help DEAD create targeted campaigns to reduce overconsumption.

Additionally, we filtered both datasets to only include information from January 1st, 2022, and onward. We felt that including any data from before this year would hinder our ability to understand current drinking habits and prevent us from gaining the most up-to-date insight.

Finally, both datasets were merged together by county names and years. This merged dataset was grouped by county, year, and alcohol type to calculate different key measurements for drinking. These measurements include the total sale dollars, average sale dollars, and the population for a county (not factoring in alcohol). Therefore, with our final dataset, we are able to predict these measurements for each alcohol type, organized by both county and year.

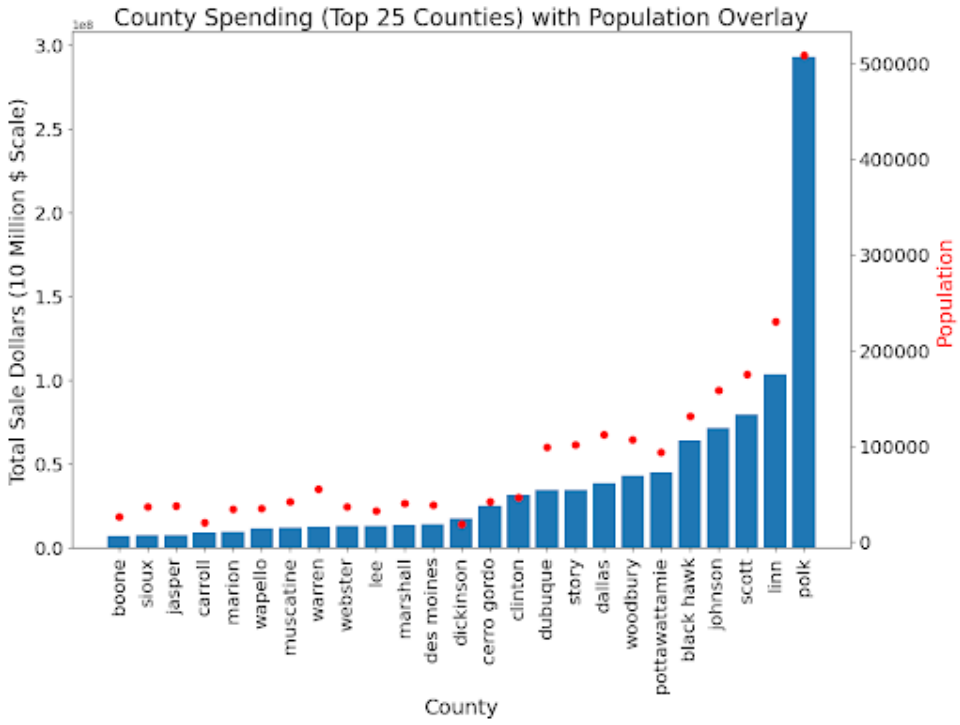
Model Selection Process

The goal of our process was to find out what factors led to increased alcohol consumption. As such, we first looked at liquor sales to see what the current market for liquor is like to better understand what is being consumed and where it is being consumed in excessive amounts. Below is a graph that shows the distribution of liquor sales in the state of Iowa.



As shown in the pie chart, the majority of sales come from Whiskey and Vodka. Whiskey is a very high-proof alcohol, which can make excess drinking very dangerous, as very little volume can have a tremendous effect. Vodka is also associated with high alcohol cocktails, which can lead to excessive drinking without even realizing how much alcohol a person really consumes.

Then we looked at the top 25 counties with the highest liquor spending in the state of Iowa. We also overlaid the population data from the Iowa census to get an image of the trend for the population of a county and its total sales.



As shown above, there is a strong correlation between the population of a county and the total sales dollars that a county spends on alcohol. However, there are some outliers shown in this graph, which are Dickinson, Clinton, Cerro Gordo, and Polk counties, which have very high alcohol sales compared to their population, unlike other counties with similar populations.

We believe the above features capture some of the key factors influencing alcohol purchases and where excessive drinking might be prevalent. Using matrices to create custom linear regression functions, we then 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.

To start, we decided on using a linear regression to model sales by county because the model is highly interpretable and will make understanding the factors that influence alcohol sales easiest. The first model tested used just county population and year, which was able to capture 33.98% of the variation in sales by county on average in testing. Then, another model was tested using county population and alcohol type to predict sales. This model captured 45.03% of the variation on average in testing and lowered the mean squared error. Finally, we attempted to add the variable year back into the model containing population and alcohol type, but the additional predictor did not improve the model in testing. Thus, we decided that using just county population and alcohol type was the best model for predicting total sales of an alcohol category in dollars by year by county.

Model Summary

The final model we created for DEAD achieved a Mean Squared Error (MSE) of $1,483,680,726,214.02 and a Root Mean Squared Error (RMSE) of $1,139,957.55, meaning that on average, the model's predictions are off by about $1.2 million. The model uniquely analyzes patterns for each county, year, and alcohol type combination, so predictive correctness varies.

The standardized population coefficient of $1,091,078 means that a one standard deviation increase in county population corresponds to over $1 million more in sales for an alcohol category. Since population varies a lot across counties, the model's pattern analysis captures effects of population differences that could be worth millions of dollars in sales. Therefore, the RMSE of $1.2 million is not unreasonably big relative to the transactions of alcohol, as the model handles a variety of sales in dollars across both smaller and larger counties.

Approximately 45.03% of the differences we see in sales can be explained by the factors of population and alcohol type (R^2). The combination of these findings shows that population and alcohol type are key drivers of sales. Other factors also play important roles in shaping alcohol driver patterns across Iowa Liquor stores in counties not captured in this analysis, but will be explored further in the next analysis for your nonprofit.

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

Table 1: Model Coefficients by Feature

Feature	Coefficient
intercept	162,567.80
population normalized	1,091,077.94
COCKTAILS	-80,473.45
GIN	-65,840.63
LIQUEUR	109,262.51
RUM	221,723.67
SCHNAPPS	-28,896.96
SCOTCH	-123,716.32
TEQUILA	266,783.56
VODKA	784,907.24
WHISKEY	1,481,377.67

Conclusions

As shown in the table below, which shows the top 10 counties with the highest sales per person, there are clear differences between the top 5 and the bottom 5, with the bottom 5 being on the

higher end of the typical alcohol sales per person. Especially Dickinson and Howard counties, which spend almost $300 more per person in their county than the third-highest ranking.

Table 2: County Sales and Population Summary

County	Total Sales ($)	Average Yearly Population	Sales per Person ($)
Dickinson	$17,635,729	18099	$974.40
Howard	$3,040,670	3136	$969.60
Clinton	$31,707,054	46139	$687.21
Cerro Gordo	$25,084,898	42457	$590.83
Polk	$292,961,467	508311	$576.34
Black Hawk	$64,030,118	131203	$488.02
Pottawattamie	$45,234,974	93354	$484.55
Scott	$79,552,458	174725	$455.30
Carroll	$9,329,494	20495	$455.22
Johnson	$71,479,688	158366	$451.36

From these, we would like to recommend targeted advertising and resource management for these five counties of Dickinson, Howard, Clinton, Cerro Gordo, and Polk. They have a far larger alcohol culture than other counties, which can lead to unsafe drinking habits, thus requiring attention, especially in Howard, as it has such a small population compared to other counties.

In addition to this, we would like to target the alcohol types with the most prevalence, especially Whiskey and Vodka. We believe that targeting the most prolific alcohol types will lead to safer drinking habits around these spirits. That can just be in the form of explaining the factors that make these spirits dangerous and the potential for excess drinking, even without having much volume.