Methodology

Anthony Rosamond[1], Madison Brown[1], Jade Towgood[1], and Jasmine Gomez[1]

[1]Arizona State University, Tempe, AZ 85281, USA

September 23, 2024

1 Extended EDA

With our preliminary findings from our exploratory data analysis, EDA, after cleaning up our 3 datasets [1, 2]—income, population, and alcohol sales in Iowa—we have decided that there are a few things we would like to investigate within the methodology portion of this project. During our EDA we have discovered that, in regards to the income trends, there has been a steady increase over the years, which tracks with inflation and the rise in prices, and could potentially influence liquor sales. When it comes to population trends, the more urban counties experienced an increase while the smaller ones were either stagnant or decreased. We discovered a positive correlation between liquor sales and population size. The alcohol sales have shown an increase over the years, spiking during 2021 most likely due to the new to-go alcohol sales law passed during COVID-19 [3] and during the holidays in 2023. When looking at counties with higher average incomes there seems to be a correlation between that and higher sales. We do want to look into the spread of data per capita as well after our preliminary results since, as of right now, it seems that population has a stronger correlation between alcohol sales than income, which is not too surprising that the larger counties have a higher consumption rate.

2 Description of Methodology

In our analysis of liquor sales trends in Iowa, we aim to explore how shifts in income, population, and alcohol sales reflect broader lifestyle changes in the state's population. By conducting a detailed exploratory data analysis (EDA) using three key datasets—income, population, and alcohol sales—we have already uncovered notable patterns that suggest strong relationships between these variables. Our methodology involves the use of inferential statistics to dig deeper into these relationships, allowing us to make informed generalizations about the population's changing behavior and consumption habits.

This analysis focuses on several key areas:

- The impact of population size and urbanization on liquor sales.
- The relationship between income growth and alcohol consumption.
- The influence of external factors, such as the COVID-19 to-go alcohol sales law, on consumption trends.

By understanding these patterns, we can infer lifestyle changes and socioeconomic shifts that may affect recreational spending and social behaviors.

2.1 Initial Results

In our preliminary findings from the EDA, we identified the following trends:

2.1.1 Income Growth

Income across Iowa has steadily increased from 2010 to 2022, with the highest average incomes found in more urban counties, such as Polk and Linn. The growth in income correlates with increased purchasing power, which could influence spending on discretionary items like alcohol. However, the correlation between income and liquor sales is weaker than expected. We found only a moderate relationship between income levels and alcohol sales, suggesting that while income may contribute to higher liquor consumption, it is not the primary driver.

2.1.2 Population Trends

Iowa's urban counties, like Polk and Linn, have shown significant population increases, while many smaller, rural counties have stagnated or declined. This urbanization trend is accompanied by higher liquor sales in more populous areas, reinforcing the idea that population size is a stronger predictor of alcohol sales than income. Scatterplots between population size and alcohol sales reveal a near-perfect correlation, with larger counties consuming more alcohol, both in total and per capita. This suggests that urban areas may be driving lifestyle changes, potentially due to factors like higher density, greater socialization opportunities, and more diverse recreational activities.

2.1.3 Alcohol Sales Trends

Liquor sales have risen steadily over the years, with noticeable spikes during 2021, likely due to the passage of the to-go alcohol sales law during COVID-19, and in the 2023 holiday season. Seasonal patterns also emerged, with alcohol consumption peaking during the holidays. This aligns with social behaviors tied to special occasions and celebrations, offering insights into how certain events influence spending on alcohol.

3 Inferential Statistics

To further explore these trends, we will apply various inferential statistical techniques to quantify the relationships between income, population, and liquor sales. The goal is to move beyond descriptive analysis and test hypotheses about how socioeconomic factors influence alcohol consumption.

3.1 Correlation and Regression Analysis

- **Population vs. Sales**: Using regression models, we will quantify the impact of population size on liquor sales. The scatterplots from our EDA show a strong positive correlation, which we will explore further to test the significance of this relationship. We will also examine whether this relationship holds on a per capita basis, as our initial findings suggest.
- **Income vs. Sales**: Although income appears to play a role in liquor sales, the relationship is weaker than expected. We will use correlation analysis to determine how much income influences consumption, controlling for population size, and test whether this relationship varies across different income levels.

3.2 Hypothesis Testing

We will use hypothesis testing to determine whether the observed trends, such as the spike in liquor sales in 2021, are statistically significant. Specifically, we will assess whether the to-go alcohol sales law had a measurable impact on overall consumption by comparing sales before and after the law's passage. Additionally, we will test whether counties with higher incomes have significantly different consumption patterns compared to counties with lower incomes.

3.3 Per Capita Consumption Analysis

Since population appears to be the strongest driver of liquor sales, we will normalize the data by calculating sales per capita for each county. This will allow us to better understand consumption rates relative to county size and identify whether certain counties are outliers in terms of per capita alcohol consumption. For example, our EDA found that Dickinson County had notably high per capita liquor sales, likely due to its status as a tourist destination. We will investigate whether other counties display similarly unique consumption patterns.

3.4 Seasonality and Time Series Analysis

We will perform time series analysis to examine trends in liquor sales over time, particularly focusing on seasonal spikes in consumption. This analysis will help us understand whether certain times of the year consistently see higher sales and whether those patterns have shifted over time, possibly due to external factors like economic conditions or changes in legislation.

4 Supervised Learning

A portion of our analysis utilizes supervised learning methods to predict future alcohol sales based on features identified during the EDA. Two such features, population and income, are noted above and will be used in regression models for predicting sales.

4.1 Regression Models: *Linear Regression and Random Forest Regression*

The two different types of regression models that will be used for the liquor sales prediction are a linear regression model and a random forest regression model. Since the linear regression model was chosen to establish the relationship between the predictors of income and population size within a county with its liquor sales. However, as linear regression models can be prone to over fitting datasets, we will also analyze the data using a random forest regression model. This type of model is typically more flexible than a linear model, since it uses the average outcome of many decision trees, thus also making it less prone to over fitting. Furthermore, the random forest model has a feature importance ranking that will identify which feature is more influential in predicting sales. To process the data before implementing each model, numeric data will be scaled using the StandardScaler, whereas categorical data will be encoded using the OneHotEncoder strategy. The dataset will then be split into train/test datasets using a 80/20 split.

5 Unsupervised Learning Methods

For our unsupervised learning methods we will be employing a couple of different techniques where the goal is to cluster the data in a way that makes sense. We will be using both the k-means clustering algorithm as well as an auto-encoder.

5.1 Clustering: *K-Means Clustering*

We will be using the K-Means Clustering algorithm to try and find patterns within our data. Specifically we are looking for sales trends within certain time frames as well as how these trends relate to factors such as income and population. In order to get clusters that make sense we need to pick an appropriate amount of clusters. In order to do so we will be using the elbow method. This involves looking at a graph of the within-cluster-sum-of-squares compared to the total cluster amount and finding the point that forms a sort of "elbow". From there we will use that number of clusters and perform the cluster analysis. After grouping the data into clusters we will then append the cluster values to the actual dataset in order to analyze them. For the analysis part we will look at statistical data of each cluster, observing how income and population is

distributed amongst these clusters as well as sales per capita. This will hopefully give us more insight into how the alcohol sales are distributed amongst various variables.

5.2 Deep Learning: *Auto-encoder*

In conjuction with the k-means algorithm we will also utilize an auto-encoder as a more "advanced" form of clustering. An auto-encoder is a neural network that takes the data and encodes it into fewer dimensions and then converts the encoded data back to its original form. Due to it being a deep learning method and having all kinds of relationships within the architecture of the neural network, the goal here is to see if there are any complex/non-linear relationships in the data.

In order to create the auto-encoder we will build a neural network architecture that is reasonable and pick a number of layers that isn't too much but will still do what it needs to do. This may require some tuning in order to see what architecture performs best. After the model is trained its performance will be tested based on the rate at which the encoded data is reconstructed correctly. After we are satisfied with the results we will employ the algorithm and extract the encoded data. Using the encoded data we will perform a k-means clustering on that data and then run the different clusters through the decoder in order to get the original data back. Using the reconstructed data we will then go through the methods described above for analyzing these clusters.

References

[1] U. S. C. Bureau. State iowa, 2022. Accessed: 09-01-2024.

[2] A. B. D. Commerce. Iowa liquor sales, 2024. Accessed: 09-01-2024.

[3] I. D. of Revenue. To-go & carryout cocktails, 2020. Accessed: 09-22-2024.