Results, Implications, Challenges

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

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

October 1, 2024

1 Analysis and Results

1.1 Exploratory Data Analysis

1.1.1 Income Dataset

When looking at the income dataset for Iowa we decided to graph the average total income per county in 2010-2014, 2014-2018, and 2018-2022 because that is how our dataset had the information grouped just to see if there were any major changes. As displayed in Figure 1 below, you can see these first three graphs their average income across the board is around $50,000, $55,000, and $67,000 respectively.

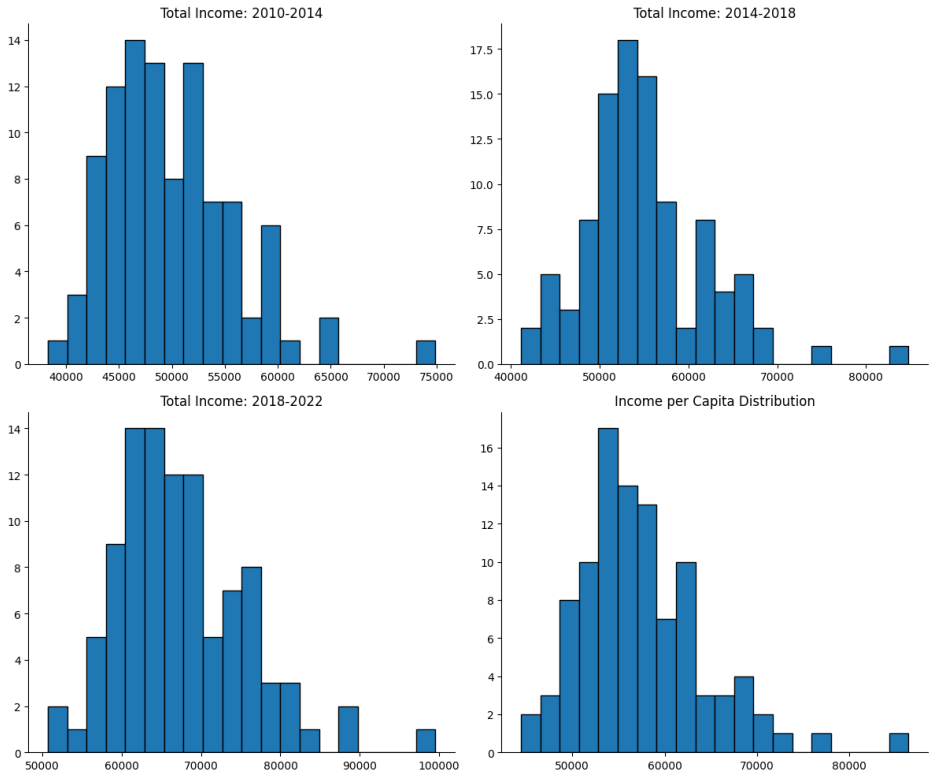

Figure 1: Four graphs displaying the distribution of total income in Iowa

The fourth graph in Figure 1 shows the total income per capita across all the counties for all the years recorded, which we did a second graph of, but added a distribution curve to help visualize the distribution more easily. When looking at Figure 2 you can see that the average income is around $58,000 across all counties in Iowa in all the years recorded. Looking at the distribution of income most of it seems to be spread around the same amount and mean other than a couple outliers on the higher end.

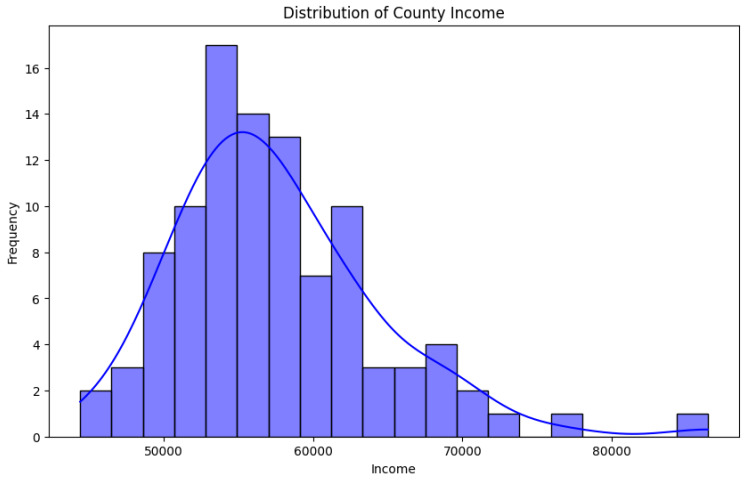

Figure 2: A second graph with a distribution curve to display the total counties' income per capita

The final figure looking at just the income data, is Figure 3, which is a geospacial map of the income per capita across all years. Looking at this map, it appears to be spread out with the highest concentration being in and near Dallas county and the rest being relatively similar.

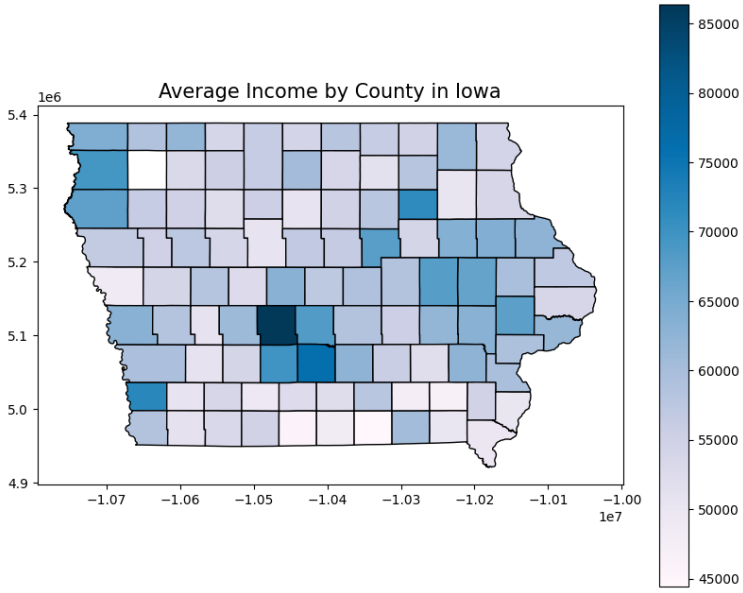

Figure 3: A geospacial map to show the spread of income across counties in Iowa per capita/average

1.1.2 Population Dataset

When beginning to look at the population dataset based on Iowa's counties we decided to take a little glance at the population size of Iowa's counties over the years. We decided to look at the counties with the top ten highest amount of people, which is pictured in Figure 4 below.

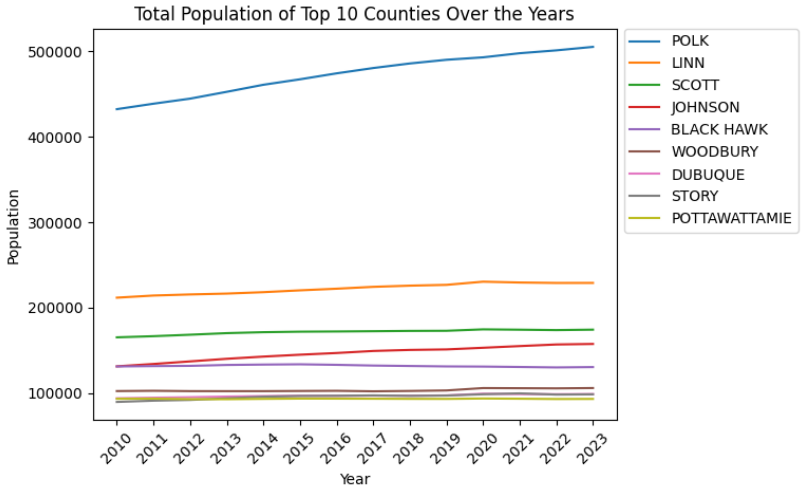

Figure 4: Total population count for the top 10 counties

After seeing the results in Figure 4, we decided to create a distribution graph to show the population growth rate across all the counties from 2010-2023. This graph shows that, for the most part, the population sizes stayed the same, they did not really grow exponentially and decrease drastically. Just small changes here and there.

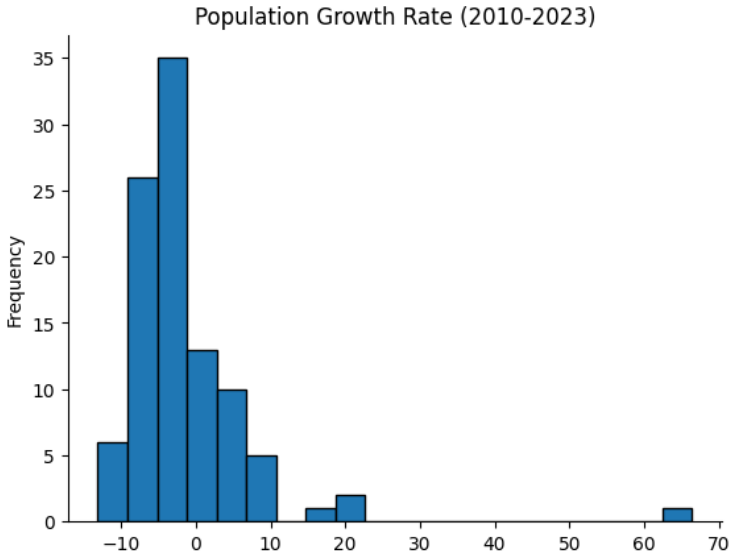

Figure 5: Distribution of growth rate in Iowa based on Counties

1.1.3 Alcohol Sales Dataset

The alcohol dataset had many different components, but for this portion we focused on the alcohol sales and volume of alcohol sold. Below, in Figure 6, we graphed the distribution of both based on county per capita. Looking at these two graphs they seem have a high frequency on the lower end, which makes sense since they are dependent on each other.

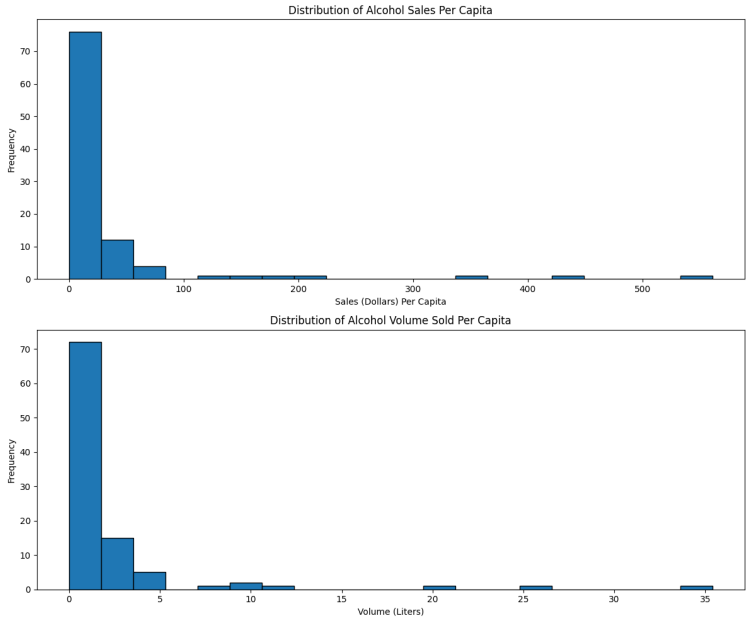

Figure 6: Alcohol sales and volume of alcohol sold distribution per capita

After confirming that if the alcohol sales go up so does the volume of alcohol sold and vice versa, we created another distribution graph and curve on alcohol sales per capita based on counties. But we condensed the scale to see the spread closer. This shows the average being $15 in alcohol sales, whereas the first graph, it was difficult to see that.

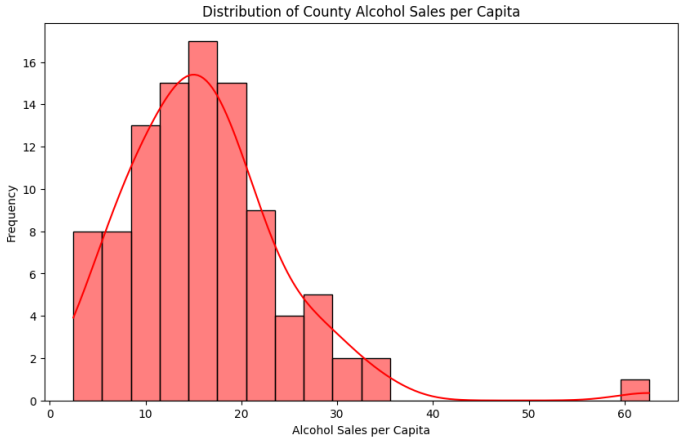

Figure 7: Distribution of alcohol sales per capita

We also created two other visualizations, geospacial ones, to display the similarities between the alcohol sales per capita and volume of alcohol sold per capita. Figure 8 and Figure 9 provide further evidence of their correlation.

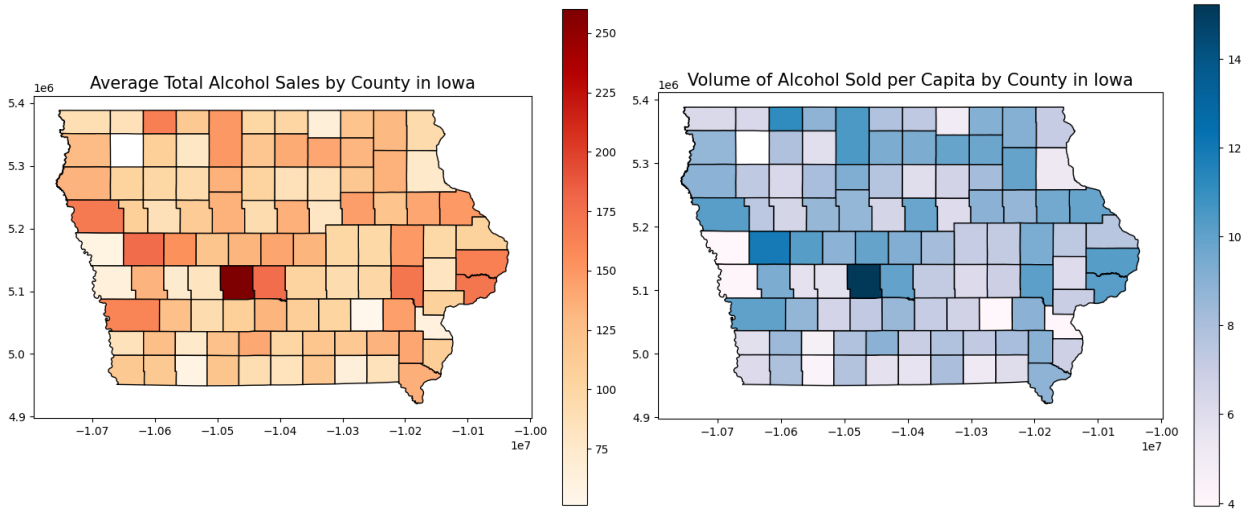

Figure 8: A geospacial map to show the spread of alcohol sales across counties in Iowa per capita/average

Figure 9: A geospacial map to show the spread of volume of alcohol sold across counties in Iowa per capita/average

1.1.4 Comparisons

In addition to analyzing the datasets based on the counties and capita, we analyzed the correlation of alcohol sales with population and income. In Figure 10, we created a scatter plot to see the correlation between the alcohol sales and population. The correlation came out to 0.46, which displays a moderately positive correlation between the two. Figure 11 we looked at the correlation between alcohol sales and income. It ended up being 0.05, which is a very weak positive correlation.

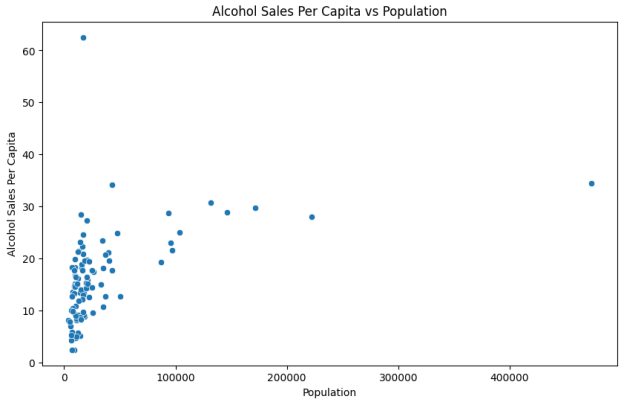

Figure 10: The correlation between Alcohol Sales Per Capita and Population is 0.46

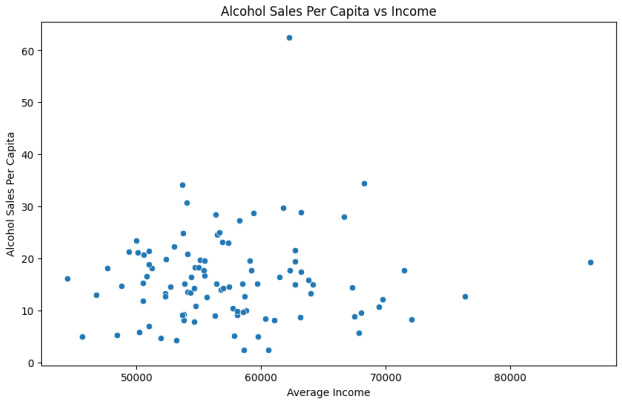

Figure 11: The correlation between Alcohol Sales Per Capita and Income is 0.05

1.2 Supervised Learning

1.2.1 Time Series Analysis

For the time series analysis and future sales forecasting, the first step was to visually inspect the sales data. Plotting the monthly average sales data revealed that the data potentially has both trend and seasonality.

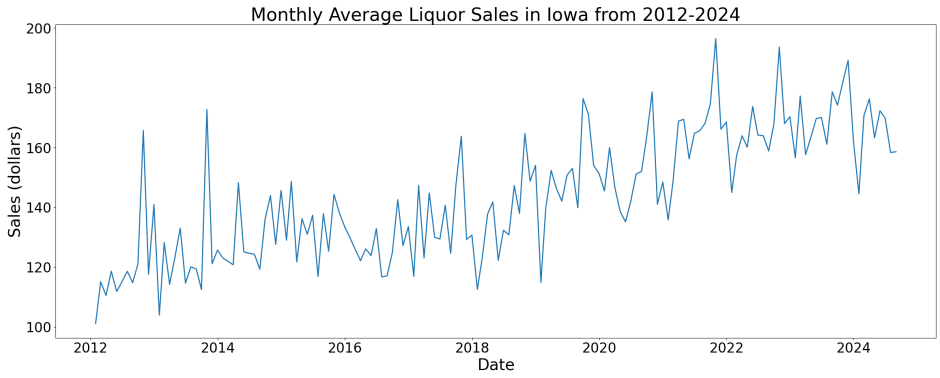

Figure 12: The monthly average liquor sales in Iowa.

To better visualize these components within the time series, the seasonal_decompose function from the statsmodels library is utilized which breaks down the seasonal components. Figure 13 confirms that this datasets contains both trend and seasonality components.

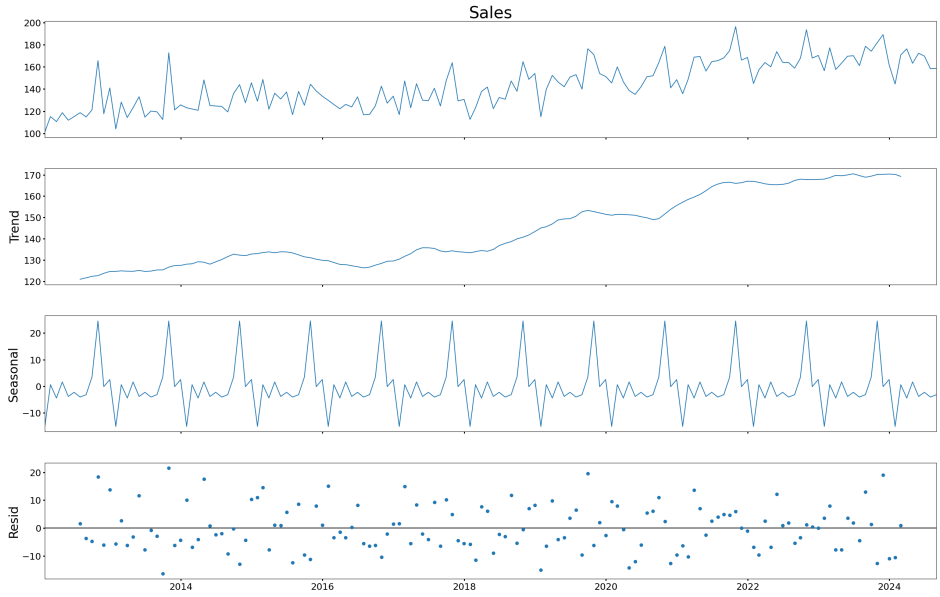

Figure 13: The trend, seasonality, and residuals of the time series against the full liquor sales.

To check for stationarity, an Augmented Dickey-Fuller (ADF) Test was performed on the data, revealing that is non-stationary. In order to make the data stationary so that we can use a SARIMA model to predict future sales, the data needs to be differenced. This entails removing the previously identified trend component from the data. For the differenced data, the autocorrelation (ACF) and partial autocorrelation (PACF) were plotted.

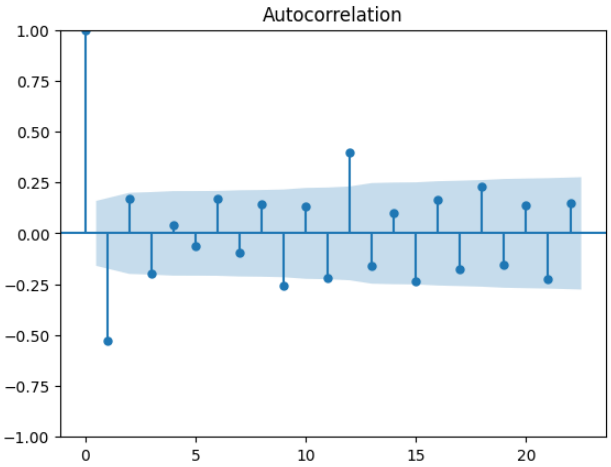

Figure 14: Autocorrelation plot.

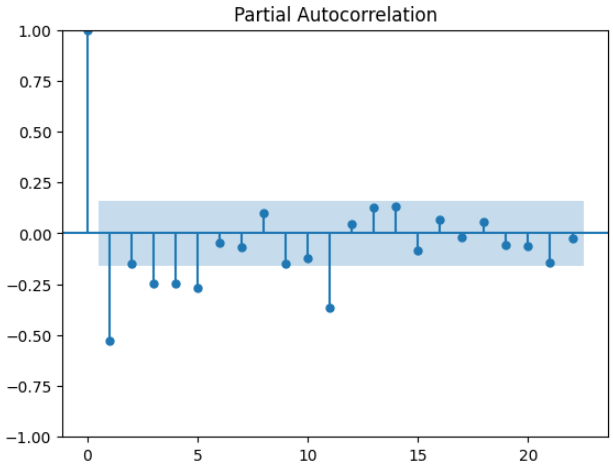

Figure 15: Partial autocorrelation plot.

Though the ACF and PACF plots can be used to determine the parameters for the SARIMA model, we used the auto_arima function from the pmdarima library to determine the best SARIMA model. This function determined that the best model had the parameters (1,0,1)(1,0,1)[12], with the first set of three terms describing the non-seasonal order and the last set of four terms describing the seasonal order.

Once the model was created and trained on the sales data, we plotted the forecasted data against the testing data [Figure 16], which were the most recent 12 months of sales data. Then, Figure 17 shows the forecasted historical data against all the sales data. These two visualizations demonstrate how the model is forecasting sales data in comparison to the actual sales data. The mean absolute error of the models was 9.359 and the root mean squared error of the model was 10.953.

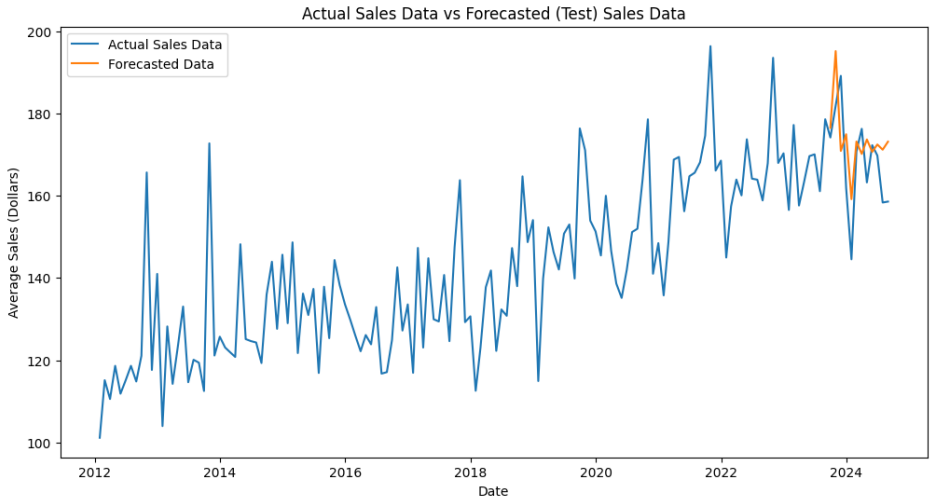

Figure 16: SARIMA model's forecasted sales over testing sales data.

The final part of the analysis was predicting future liquor sales using the model. As shown in Figure 18, the model predicts a decrease in future sales. Within our exploratory data analysis, we had previously identified the peak in liquor sales to have been in 2023. Therefore, this predicted decrease in sales in interesting to note.

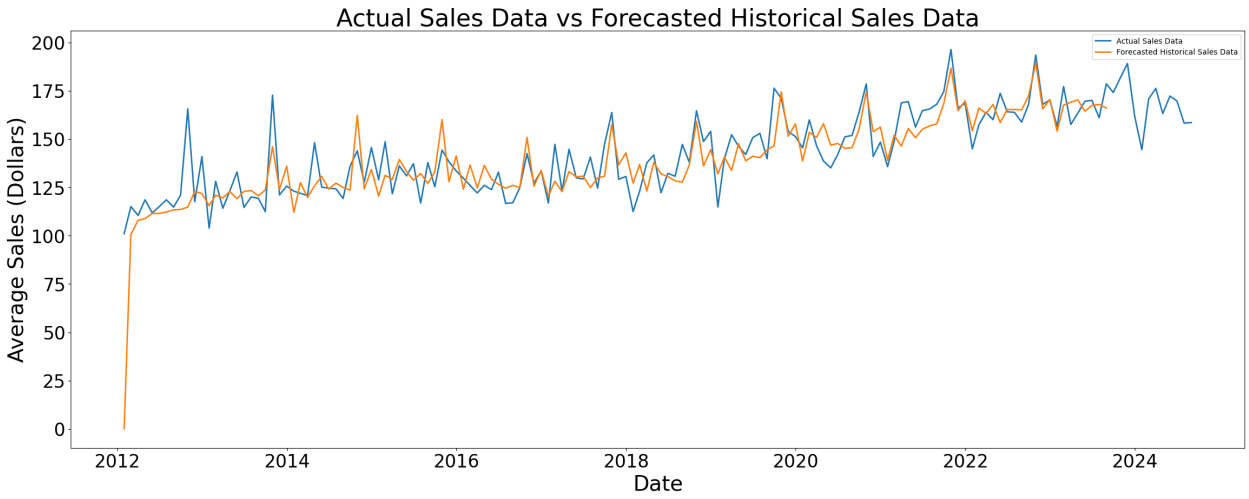

Figure 17: SARIMA model's forecasted historical sales over entire sales data.

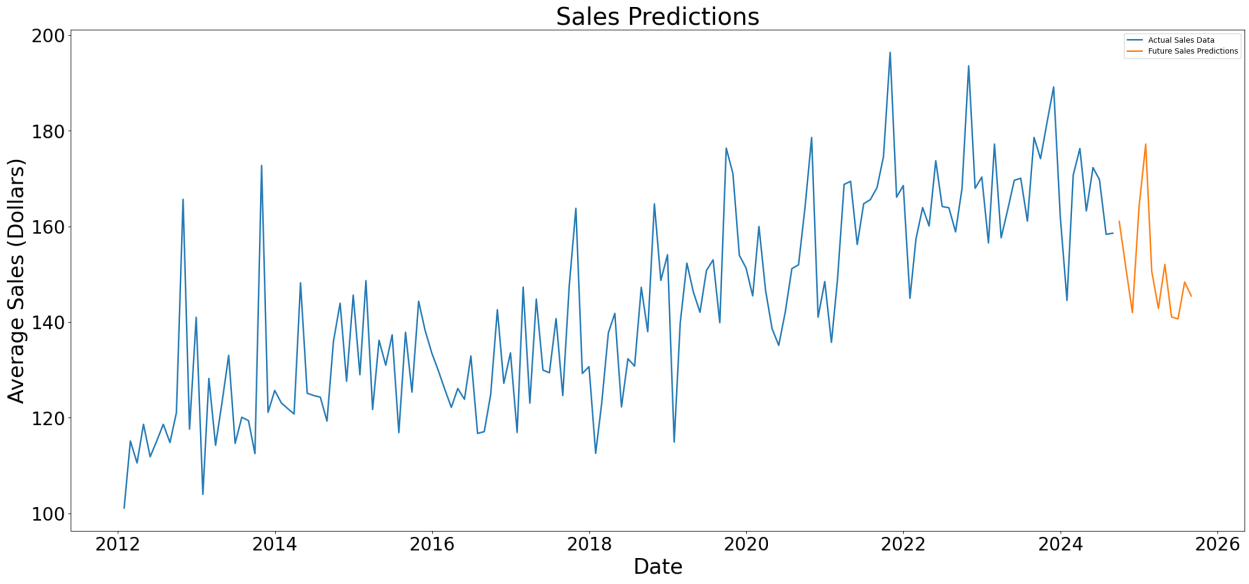

Figure 18: SARIMA model's future sales prediction for next 12 months.

1.2.2 Linear Regression and Random Forest Regression Models

For the supervised regression models, we sought more data relevant to our project question of demographics affecting liquor sales within Iowa. Relevant features we identified from our existing dataset and additional census data for each county were:

- average income
- average population
- average volume (liters) of alcohol sold
- total number of alcohol vendors
- average cost of one bottle of alcohol
- total number of different alcohol categories sold
- total number of religion congregations
- percent of population that adhered to a religious congregation,
- median age
- total number of universities
- majority urban versus rural classification
- percentage of population in urban areas
- percentage of population in rural areas

The goal of each of these regression models was to predict the Average Liquor Sales (dollars). Numeric columns were scaled using a Standard Scaler, and the one categorical feature was One Hot Encoded.

The linear regression model had a root mean squared error of 11.6503 and an R^2 score of 0.8190. The random forest model performed slightly worse, and had a root mean squared error of 12.2815 and an R^2 score of 0.7988. Using the random forest feature_importance method, it is clear that the most influential feature in this model for predicting a county's sales was the average volume (liters) of alcohol sold, as shown in Figure 19.

Since this single feature was the main feature behind the model's predictions, the models were recreated dropping the average volume (liters) of alcohol sold column to see which other features impacted liquor sales. This significantly impacted model performance. The second linear regression model had a root mean squared error of 25.3986 and an R^2 score of 0.1395. The second random forest regression model had a root mean squared error of 27.2497 and an R^2 score of 0.009537.

Once again utilizing the feature_importance function, the second random forest regression model ranked the total number of alcohol vendors, the total number of alcohol categories sold, and the average cost of a bottle of alcohol to be the top three most important features in a county's liquor sales. The new ranking of features is shown in Figure 20.

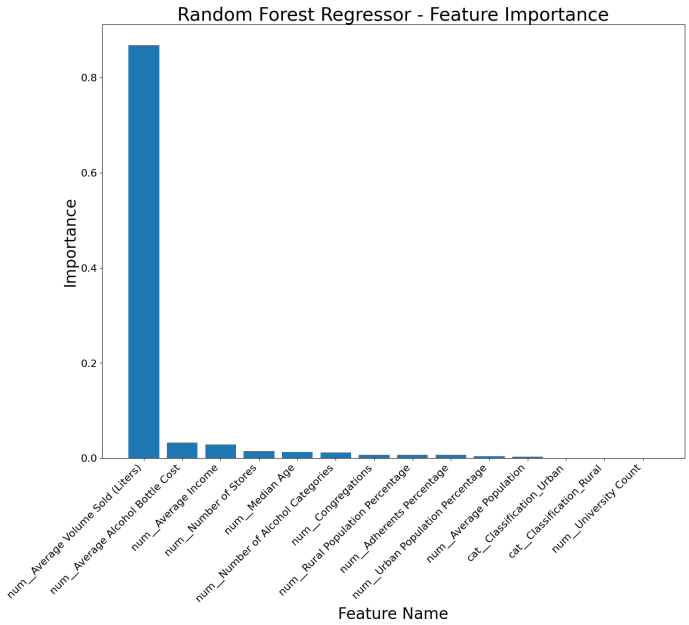

Figure 19: Random Forest Regression feature importance, when including Average Volume (liters) of Alcohol Sold as a feature.

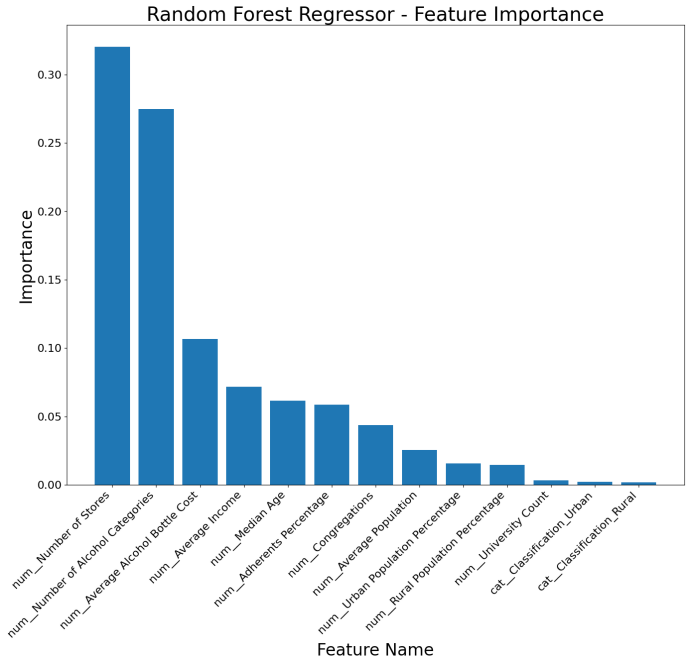

Figure 20: Random Forest Regression feature importance, when removing Average Volume (liters) of Alcohol Sold as a feature.

1.3 Unsupervised Learning

1.3.1 K-Means Clustering

In order to try and learn some more behaviors we decided to employ some clustering techniques in order to see how an unsupervised learning algorithm would group the data and whether any patterns previously unseen would show up. At first our original plan was to perform both K-Means clustering on its own and then utilize an auto-encoder as well in order to try and draw out more complicated relations. Unfortunately the auto-encoder ended up being a total wash, the details of which will be discussed later on. As for the more traditional K-Means, we first wanted to figure out both on what variables the cluster analysis would be performed on, and also how many clusters would be ideal for the analysis. For the variables we wanted to look at total sales versus revenue. We also decided to look at sales per capita when compared to both average income and average population. Now in order to pick an appropriate amount of clusters we used the elbow method. This method creates a visual graph that represents the "explained variation" within the clusters. In other words the variation is low enough within each cluster that the amount of outliers in each is reduced as much as possible without creating too many clusters.

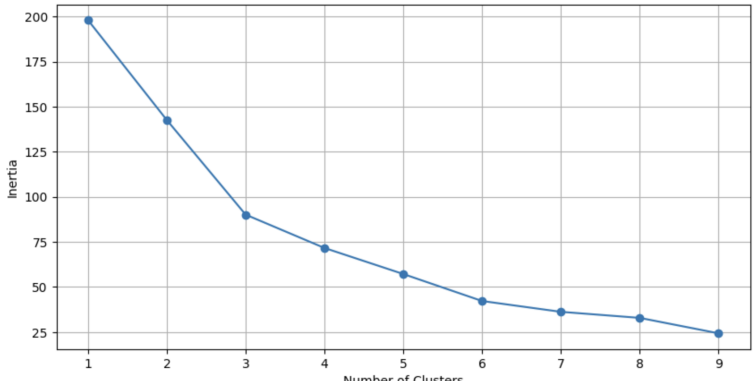

Figure 21: Example of elbow method from our analysis.

The first clustering we did was revenue versus sales. While the two variables are clearly closely related, we wanted to see the behavior of how each county was distributed within each. Immediately we see some

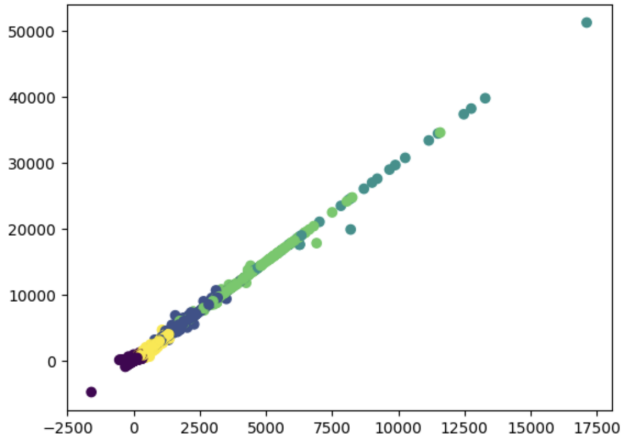

Figure 22: Sales vs Revenue

interesting clustering behavior as the clusters actually ended up overlapping. We weren't sure what was

causing this but we decided to move forward on this clustering. The main difference between the clusters themselves was the average value of the sales field. Some clusters contained sales that were extremely large while others were more mainstream. We decided to see what the top 5 most prevalent counties were in each cluster. Not surprisingly it was generally the top 5 most populous counties. However the 2 clusters that contained the largest transactions contained Dallas County. Dallas ranks 10 for average population within our data so we were a bit surprised to see it here. Another thing to note is Dallas ranks number 1 in average income and is adjacent to Polk County which is the most populous. Our conclusion for this discrepancy is that the large transactions are more of an extension of Polk county and is amplified by a higher income earning population.

Next we wanted to look at average income and sales per capita. From our EDA earlier we had learned that there is essentially no correlation between these two variables but we thought that clustering on them would help bring out some form of pattern within.

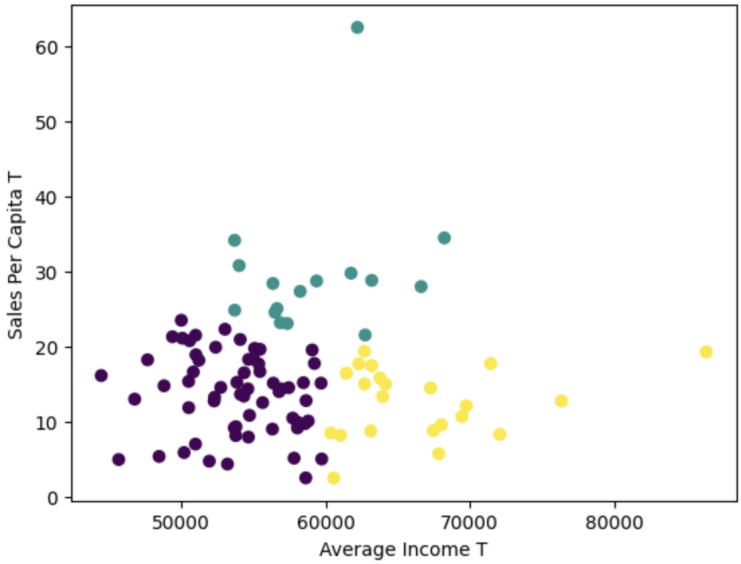

Figure 23: Sales per capita vs Income

While there is no statistical correlation we see something interesting in the clusters. For the 3 clusters we have 2 that separate the lower income brackets from the higher income brackets. We then have a third cluster who's income range falls within the other 2 clusters but has a much higher sales per capita. We can conclude here that while average income can't be used to predict sales, it seems that counties with a mainstream middle class income are far more likely to have higher sales in alcohol. For our last cluster we decided to look at sales per capita vs average population. Here we have something a little more exciting. We have 3 clusters, although one cluster is Polk county far away from everything else due to it having a much higher population so we will focus on the other two. Those two clusters break up the lower population and the higher population counties and there is an obvious difference in sales per capita. Counties that have a larger population all have a more moderate sales per capita. While the sales per capita for the low population cluster do contain more moderate numbers, there is a huge chunk that have much lower sales per capita. While there isn't a strong correlation between the two variables we do see that having a higher population makes it more likely to have higher sales per capita.

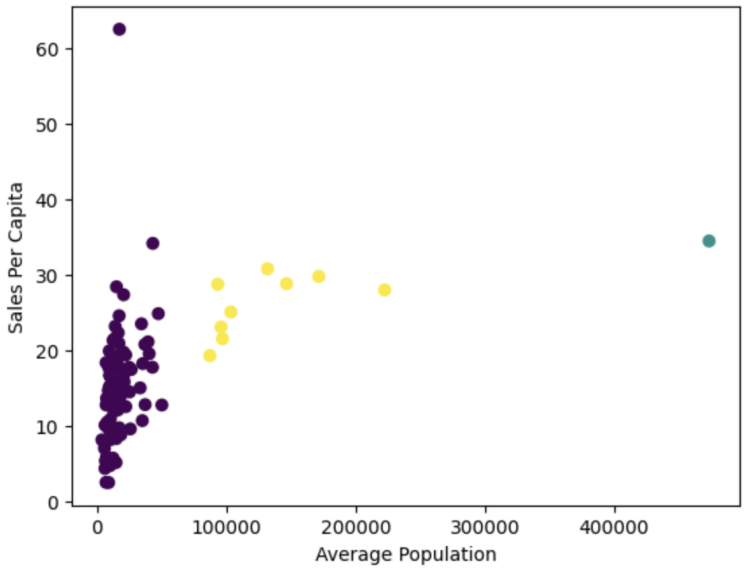

Figure 24: Sales per capita vs Population

2 Implications

2.1 Scientific Implications

1. **Correlation Between Population and Alcohol Sales**: The positive correlation between population size and alcohol sales (0.46) aligns with existing research that suggests higher population densities drive greater alcohol consumption due to increased access to vendors and more social opportunities. Larger urban populations, such as in *Polk* and *Linn* counties, exhibited significantly higher overall sales volumes [7]. This finding implies that public health efforts aimed at reducing alcohol consumption in urban areas should consider both the social and environmental factors influencing consumption behavior.

2. **Weak Correlation Between Income and Alcohol Sales**: Contrary to the assumption that higher incomes result in greater alcohol consumption, our analysis found a weak correlation (0.05) between income and sales. This result challenges traditional economic models, suggesting that factors beyond income, such as social norms and education, may play a more critical role in alcohol consumption [5]. This finding mirrors broader literature on the inelasticity of alcohol consumption, meaning people do not necessarily adjust their drinking habits in response to income fluctuations.

3. **Seasonal and Long-Term Trends**: The SARIMA model identified periodic spikes in alcohol sales during holidays but projects a potential decline in sales post-2023. This supports the hypothesis that public health campaigns and changes in consumer behavior, potentially accelerated by the COVID-19 pandemic, could lead to long-term reductions in alcohol consumption [4]. This decline highlights the need for further investigation into the effectiveness of public health interventions and shifts in societal drinking habits.

2.2 Social and Ethical Implications

1. **Urban vs. Rural Alcohol Consumption Patterns**: The K-Means clustering analysis revealed distinct consumption patterns between rural and urban counties. Urban counties displayed higher volumes of alcohol sales but lower per capita consumption, suggesting that accessibility to vendors plays a critical role in consumption levels. In contrast, rural counties exhibited greater variability in consumption patterns, raising ethical concerns about access to healthcare and public awareness campaigns in those areas [3]. The discrepancy between urban and rural alcohol consumption patterns suggests that tailored public health

interventions are needed to address the specific needs of each community.

2. **Alcohol Vendor Density and Consumption**: The analysis highlighted a strong relationship between the number of alcohol vendors and total sales, particularly in urban areas. This suggests that easy access to alcohol increases consumption, underscoring the ethical need for regulating vendor density. Public policy measures such as limiting the number of alcohol outlets in high-consumption areas could effectively reduce alcohol-related harm while maintaining consumer freedom [8]. This finding supports the growing body of literature advocating for harm-reduction strategies based on regulating alcohol availability.

3 Challenges

As with any research project done we had our fair share of complications along the way. While creating the regression models we quickly realized that the "average volume sold" feature completely overshadowed all of the other variables. Thinking about it this makes sense considering that the more you purchase of something the more it will cost. In order to mitigate this we dropped the feature and recreated the model to get a better idea of how the other features influenced sales.

Now moving on to the unsupervised learning we had a very interesting hiccup. We wanted to try out deep learning methods for clustering, specifically an auto-encoder, but we had a multitude of issues. The first thing we noticed was the time to train the model. Typically you will run several epochs however each epoch took anywhere from 40 seconds to over a minute and we were hoping to get at least 100. As each epoch was run we had the loss print and it wouldn't go under 0.75 which is not a useful model at all. We tried messing around with all of the hyper-parameters but any change was marginal at best so we decided to scrap the model.

These challenges were the two main ones that impacted what we had planned to do. Other challenges that were presented acted more like nuisances than anything else. An example is we had trouble formatting several images into our document and trying to get them to show in the correct spot.

4 Conclusion

The research conducted on Iowa's liquor sales trends has revealed significant insights into the broader lifestyle changes and population shifts within the state. The core question driving this study was: *How can analyzing liquor sales trends provide insights into broader lifestyle changes and shifts in Iowa's population?*

Through a detailed analysis of the correlation between liquor sales, population size, and income distribution, our findings indicate that population size is a far more significant predictor of liquor consumption than income [2]. This supports the notion that in urbanized, high-density areas, social and cultural behaviors, rather than purely economic factors, drive alcohol consumption [7]. Counties such as *Polk* and *Linn*, with larger urban populations, consistently displayed higher overall alcohol sales compared to smaller, rural counties. This suggests that population growth and urbanization in Iowa are key contributors to rising alcohol consumption.

However, despite the overall increase in sales from 2012 to 2023, our SARIMA forecast indicates a possible decline in future sales [4]. This projected decrease may point to broader shifts in consumer behavior, potentially driven by increased health consciousness, policy interventions, or societal changes post-COVID-19. The seasonality seen in the sales data—where spikes occurred during holiday periods—highlights the strong role of cultural and social traditions in shaping consumption patterns [6].

Income, while traditionally considered a driver of consumer spending, showed only a weak correlation with liquor sales [5]. This finding challenges the assumption that wealthier counties would inherently consume more alcohol. Instead, the cultural and social context of a region, along with access to alcohol vendors, appears to be a more significant factor in influencing consumption.

These findings provide critical insights into how lifestyle changes, such as urbanization and evolving social norms, are influencing Iowa's population dynamics. The forecasted decrease in sales further suggests that

lifestyle shifts, including changes in health priorities and alcohol-related policies, may be driving long-term changes in consumption behavior.

4.1 Next Steps

Moving forward, several key areas require further exploration:

- **Investigating Public Health Interventions:** Given the projected decline in alcohol sales, future research should explore the effectiveness of public health campaigns, taxation policies, and regulations on alcohol availability [8]. Tracking how these interventions shape consumer behavior over time will provide valuable insights into their long-term impact.
- **Demographic and Social Analysis:** To further refine our understanding of consumption drivers, future studies should incorporate additional demographic factors such as education, employment status, and ethnic composition [6]. This will help identify how social determinants of health interact with alcohol consumption trends.
- **Vendor Density Regulation:** Policymakers should consider regulating the density of alcohol vendors, particularly in high-consumption urban areas [8]. Examining the relationship between vendor accessibility and sales will provide a clearer picture of how consumption can be controlled through policy.

4.2 Final Conclusions

In conclusion, the analysis of Iowa's liquor sales data provides a clear indication that population growth and urbanization are the primary drivers of alcohol consumption in the state, more so than income [1]. As population density increases, particularly in urban counties, the demand for alcohol increases. However, the projected decline in future sales suggests that Iowa may be entering a phase where broader societal changes, such as health-conscious behavior or alcohol-related legislation, begin to shape consumer habits [3].

The key takeaway is that by analyzing liquor sales trends, we gain valuable insights into how Iowa's population is evolving in terms of lifestyle and consumption habits. As the state's population continues to shift, particularly with growing urban centers, alcohol consumption patterns are likely to be influenced by not just economic factors, but also social, cultural, and public health interventions [7]. This study provides a foundation for policymakers, public health officials, and researchers to monitor and address alcohol consumption trends as Iowa's population dynamics continue to evolve.

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] M. Davis and R. Peterson. Rural vs. urban alcohol consumption patterns: A comparative analysis. *Rural Health Journal*, 28(2):78–85, 2019.

[4] L. Garcia. Time series analysis of alcohol sales using sarima models. *Time Series Analysis Journal*, 11(3):201–210, 2022.

[5] E. Johnson. The impact of income on alcohol consumption: A review. *Economic Behavioral Studies*, 42:123–130, 2021.

[6] M. Lee. Social and cultural factors driving alcohol consumption. *Sociology of Health and Illness*, 39:299–312, 2021.

[7] J. Smith and J. Doe. Urban alcohol consumption trends in the u.s. *Journal of Urban Studies*, 34(4):456–467, 2023.

[8] A. Thompson. Alcohol vendor density and its impact on public health policy. *Public Health Policy Review*, 15(1):67–80, 2020.