Capstone Exploratory Data Analysis

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

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

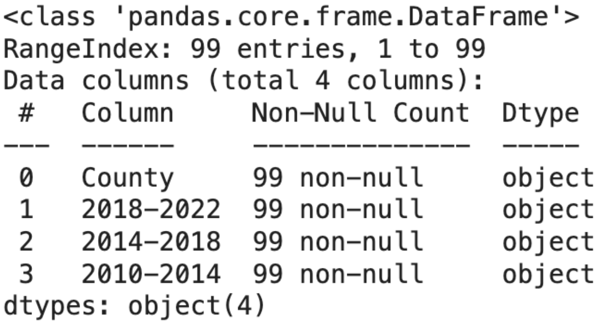
September 22, 2024

1 Data Sources, Preparation, and Cleaning

We used three different datasets for this exploratory data analysis of study the various liquor trends in Iowa to see if there are any notable changes over the years. As a quick side note for our geospacial analyses, we created a geospacial borders of Iowa's counties and standardized the counties' names since they were not exactly the same as the rest our data frames case-wise.

1.1 Income Dataset

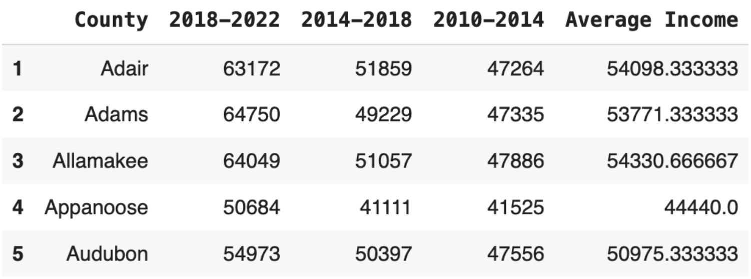
The first dataset [3] we received off of the Iowa State Data Center website, which uses the United States Census Bureau and Iowa state agencies to provide statistics about Iowa's income. It contains 99 entries, for each county in Iowa, and 4 columns that were named Unnamed: 0, 2018-2022, 2014-2018, 2010-2014. The first row of this dataset prints out as County Name, Estimate, Estimate, and Estimate, so we started off by dropping that first row and renaming the Unnamed: 0 to County. After changing that we looked at some information about the dataset and found that all the data types were objects so our next task was focusing on making the estimates numerical values, see Figure 1.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 1 to 99
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   County     99 non-null     object
 1   2018-2022  99 non-null     object
 2   2014-2018  99 non-null     object
 3   2010-2014  99 non-null     object
dtypes: object(4)
```

Figure 1: Some data info and dtype of each column in our income dataframe

We converted the data types for 2018-2022, 2014-2018, and 2010-2014 to numerical values and although the df_income.info() already returned all of the instances as non-null values we still checked to see if there were any NA values, there were not. We then calculated the total average income of each county and added it into the dataset as a new column as our final step of cleaning of preparing this dataset, see Figure 2.

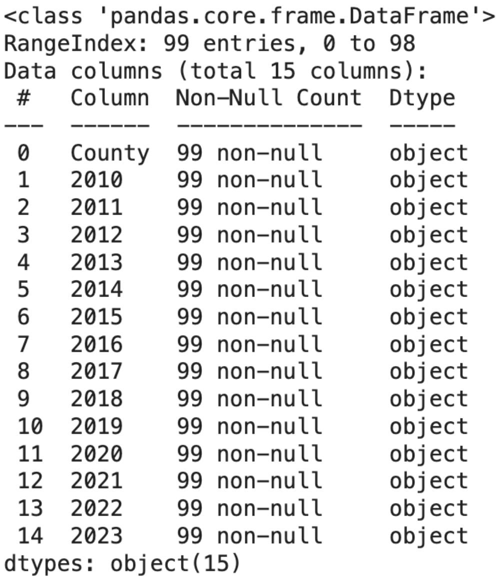


	County	2018-2022	2014-2018	2010-2014	Average Income
1	Adair	63172	51859	47264	54098.333333
2	Adams	64750	49229	47335	53771.333333
3	Allamakee	64049	51057	47886	54330.666667
4	Appanoose	50684	41111	41525	44440.0
5	Audubon	54973	50397	47556	50975.333333

Figure 2: Snapshot of our cleaned up income dataframe

1.2 Population Dataset

The second dataset [1] we received off of the United States Census Bureau website on each Iowa county's population count. It also has 99 entries for each county, but 14 columns that include County Name, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, and 2023. We looked at some information about this dataset and found that all the data types were objects too as well as a Non-Null Count of 99.



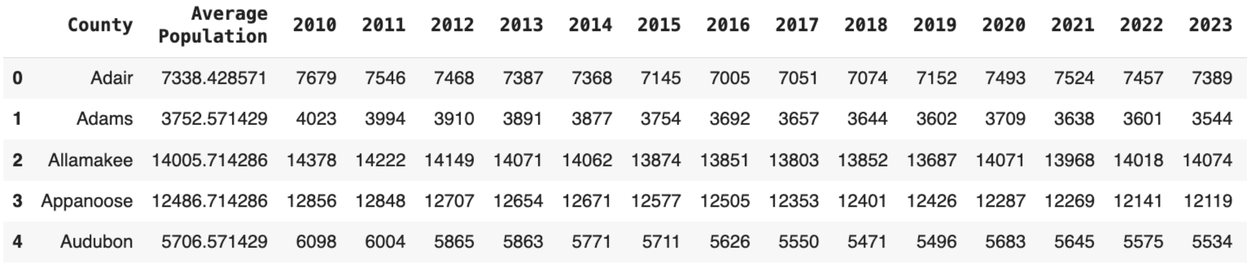
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 15 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   County  99 non-null     object
 1   2010    99 non-null     object
 2   2011    99 non-null     object
 3   2012    99 non-null     object
 4   2013    99 non-null     object
 5   2014    99 non-null     object
 6   2015    99 non-null     object
 7   2016    99 non-null     object
 8   2017    99 non-null     object
 9   2018    99 non-null     object
 10  2019    99 non-null     object
 11  2020    99 non-null     object
 12  2021    99 non-null     object
 13  2022    99 non-null     object
 14  2023    99 non-null     object
dtypes: object(15)
```

Figure 3: Some data info and dtype of each column in our population dataframe

After looking at this our next task was focusing on making the years' instances into numerical values. We converted the specified columns into numeric values and then double checked to see if there were any NA values, which there were not, and the shape of the dataframe to wrap up this dataframe. We also calculated the Average Population for each county and added it as a new column, see Figure 4.

1.3 Alcohol Sales Dataset

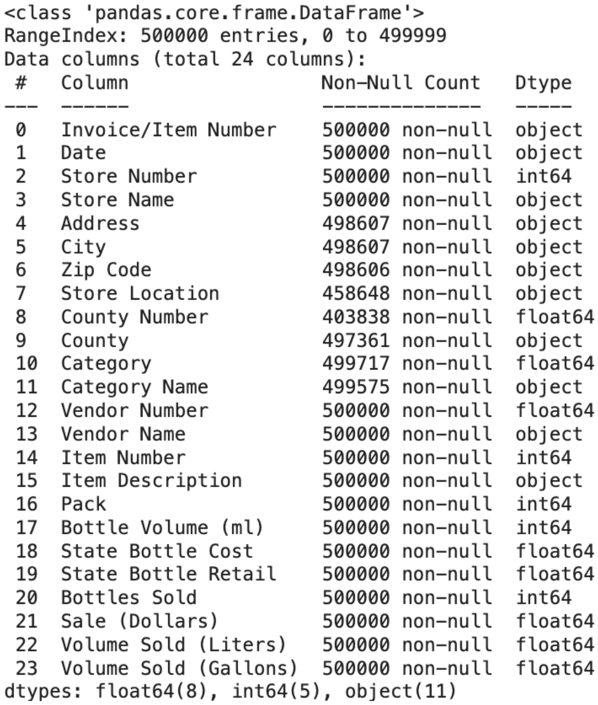
The final dataset [2] is published and maintained through the Iowa Department of Revenue, Alcoholic Beverages and contains data starting from January 1, 2012 to currently and is typically updated the first of every month. As of September 1, 2024, there are 29.9M instances, individual product purchases, in this dataset with 24 different variables. The 24 variables are Invoice/Item Number, Date, Store Number, Store



	County	Average Population	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
0	Adair	7338.428571	7679	7546	7468	7387	7368	7145	7005	7051	7074	7152	7493	7524	7457	7389
1	Adams	3752.571429	4023	3994	3910	3891	3877	3754	3692	3657	3644	3602	3709	3638	3601	3544
2	Allamakee	14005.714286	14378	14222	14149	14071	14062	13874	13851	13803	13852	13687	14071	13968	14018	14074
3	Appanoose	12486.714286	12856	12848	12707	12654	12671	12577	12505	12353	12401	12426	12287	12269	12141	12119
4	Audubon	5706.571429	6098	6004	5865	5863	5771	5711	5626	5550	5471	5496	5683	5645	5575	5534

Figure 4: Snapshot of our cleaned up population dataframe

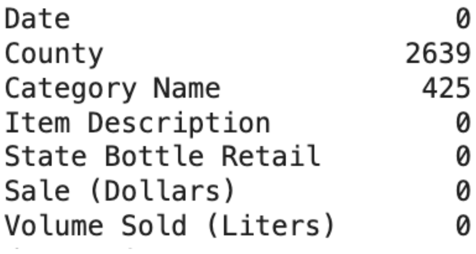
Name, Address, City, Zip Code, Store Location, County Number, County, Category, Category Name, Vendor Number, Vendor Name, Item Number, Item Description, Pack, Bottle Volume (ml), State Bottle Cost, State Bottle Retail, Bottles Sold, Sale (Dollars), and Volume Sold (Gallons). We did have to scale down this dataset and we ended up using 500,000 instances of it at random using df.sample() with the 'frac' argument set to the value that would yield 500,000 records. When we ran df_alcohol.info() we could tell that we were missing some values and had a few different data types, see Figure 5.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500000 entries, 0 to 499999
Data columns (total 24 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Invoice/Item Number    500000 non-null  object
 1   Date                   500000 non-null  object
 2   Store Number           500000 non-null  int64
 3   Store Name             500000 non-null  object
 4   Address                498607 non-null  object
 5   City                   498607 non-null  object
 6   Zip Code               498606 non-null  object
 7   Store Location         458648 non-null  object
 8   County Number          403838 non-null  float64
 9   County                 497361 non-null  object
 10  Category               499717 non-null  float64
 11  Category Name          499575 non-null  object
 12  Vendor Number          500000 non-null  float64
 13  Vendor Name            500000 non-null  object
 14  Item Number            500000 non-null  int64
 15  Item Description       500000 non-null  object
 16  Pack                   500000 non-null  int64
 17  Bottle Volume (ml)     500000 non-null  int64
 18  State Bottle Cost      500000 non-null  float64
 19  State Bottle Retail    500000 non-null  float64
 20  Bottles Sold           500000 non-null  int64
 21  Sale (Dollars)         500000 non-null  float64
 22  Volume Sold (Liters)   500000 non-null  float64
 23  Volume Sold (Gallons)  500000 non-null  float64
dtypes: float64(8), int64(5), object(11)
```

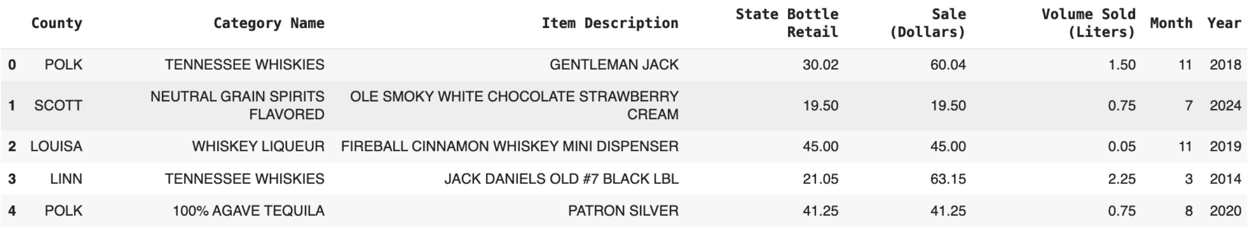
Figure 5: Some data info and dtype of each column in our alcohol sales dataframe

Before dealing with the missing values we dropped the columns we were not planning to use for this analysis from our dataframe. After dropping those we were left with 7 columns. We then printed to see if these columns had any NA values and they did, see Figure 6.



```
Date                     0
County                2639
Category Name          425
Item Description         0
State Bottle Retail      0
Sale (Dollars)           0
Volume Sold (Liters)     0
```

Figure 6: Missing value count and column names in our alcohol sales dataframe



	County	Category Name	Item Description	State Bottle Retail	Sale (Dollars)	Volume Sold (Liters)	Month	Year
0	POLK	TENNESSEE WHISKIES	GENTLEMAN JACK	30.02	60.04	1.50	11	2018
1	SCOTT	NEUTRAL GRAIN SPIRITS FLAVORED	OLE SMOKY WHITE CHOCOLATE STRAWBERRY CREAM	19.50	19.50	0.75	7	2024
2	LOUISA	WHISKEY LIQUEUR	FIREBALL CINNAMON WHISKEY MINI DISPENSER	45.00	45.00	0.05	11	2019
3	LINN	TENNESSEE WHISKIES	JACK DANIELS OLD #7 BLACK LBL	21.05	63.15	2.25	3	2014
4	POLK	100% AGAVE TEQUILA	PATRON SILVER	41.25	41.25	0.75	8	2020

Figure 7: Snapshot of our cleaned up alcohol sales dataframe

We dropped those values and were left with 496,944 instances. After that we separated the date into a month and year format and added two new columns with that. We dropped the original date column as well, see Figure 7, to bring us to the end of our data preparation and cleaning.

2 Exploratory Data Analysis

2.1 Overview of Census Data

2.1.1 Income

Through the analysis of the Iowa income dataset, we calculated the average income across counties for three periods: 2010-2014, 2014-2018, and 2018-2022. The dataset revealed that average incomes have steadily increased across most counties, with the highest average income found in *Adams County*, reaching $64,750 in the 2018-2022 period. The income distribution histograms also show a consistent rise, which can be linked to economic growth in the state. This growth correlates with increased purchasing power, which may influence liquor sales in counties with higher income levels.

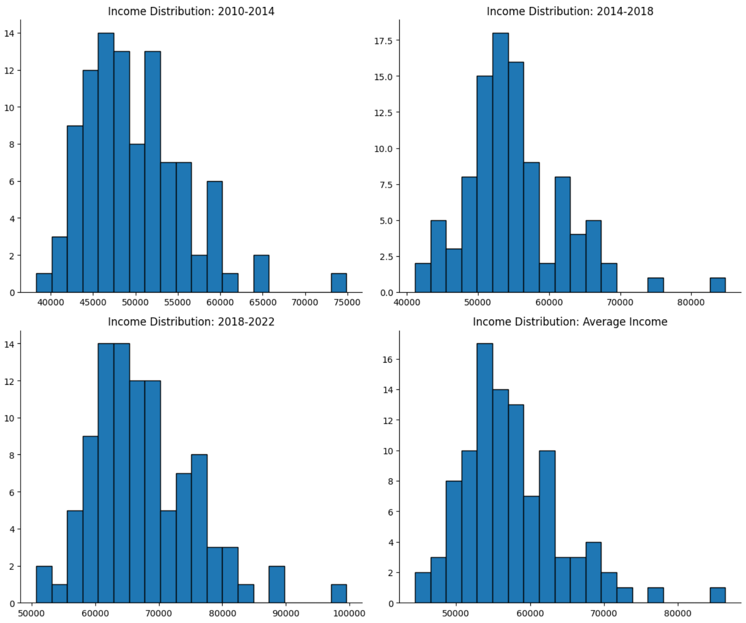


Figure 8: Distribution of Average Income Over Time

2.1.2 Population

The population data analysis highlighted growth trends from 2010 to 2023. Counties such as *Polk* and *Linn* showed substantial population increases, reflecting urban expansion in Iowa. On the other hand, smaller counties experienced stagnation or slight decline. The average population growth across the state provides insight into demographic changes, which, in turn, can impact consumer behavior, including liquor sales. A correlation between larger populations and increased alcohol sales was evident in more populous counties. This relationship was examined through the scatterplot between population size and liquor sales, which showed a strong positive correlation.

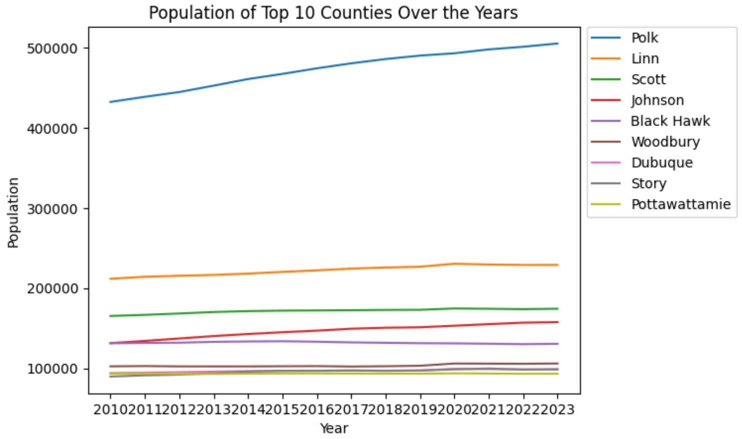


Figure 9: Top 10 Counties by Population

2.2 Overview of Alcohol Sales Data

2.2.1 Sales by County

The analysis of the Iowa liquor sales dataset revealed key consumption trends across counties. The distribution of alcohol sales by county was right-skewed, indicating that a few counties, such as *Polk* and *Linn*, dominated the total sales. These counties have higher population densities and higher incomes, which were found to be positively correlated with alcohol sales. The scatterplot analysis revealed a strong positive correlation between population size and alcohol sales, as well as a moderate correlation with average income. This suggests that both population size and income levels significantly impact liquor consumption in Iowa.

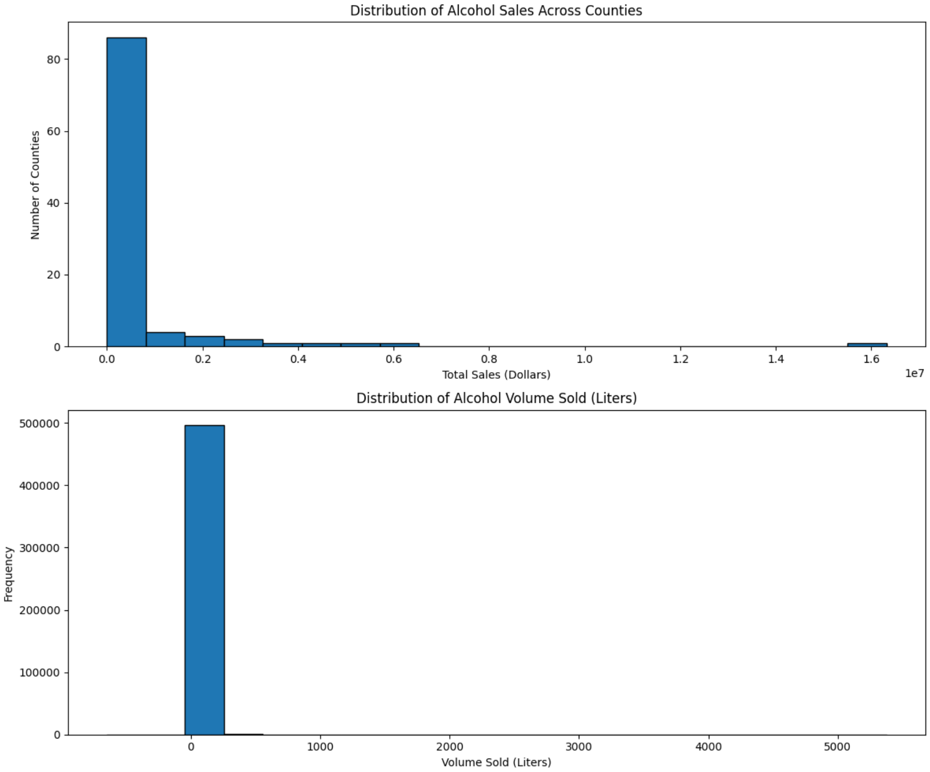


Figure 10: Sales by County

2.2.2 Alcohol Sales Over Time

Additionally, sales trends over time from 2012 to 2024 showed a consistent increase in both total revenue and volume of alcohol sold. Peaks in sales were observed around the holiday seasons, indicating seasonal spikes in consumption. This trend was evident from the bar plots showing year-over-year increases in sales, with 2023 marking the highest recorded sales in both revenue and volume.

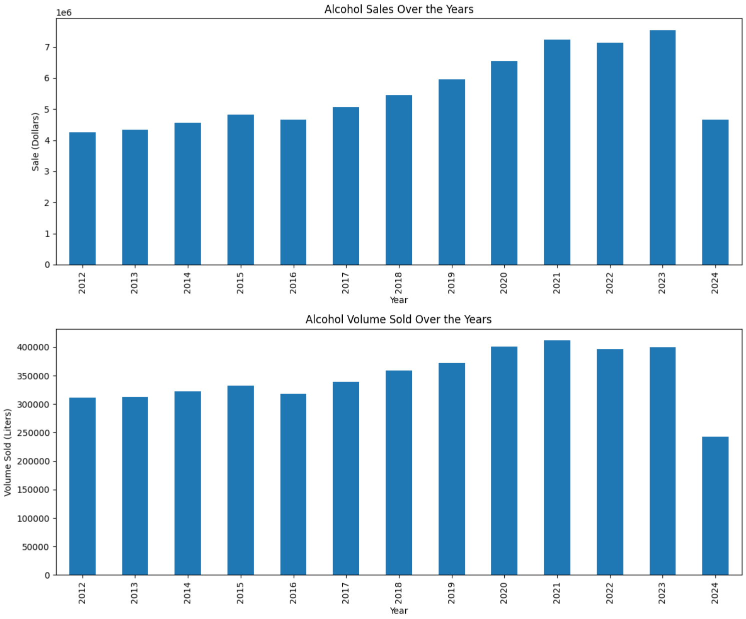
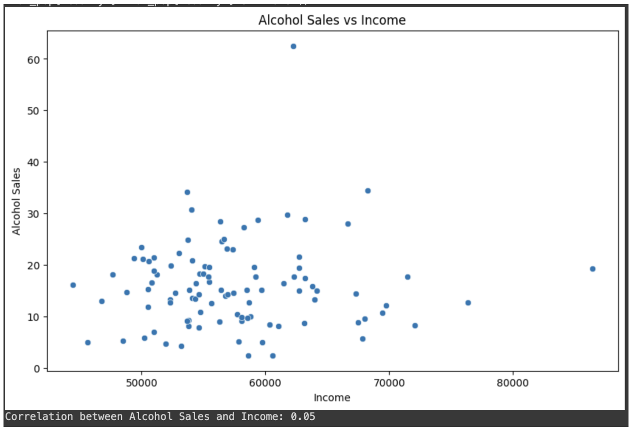


Figure 11: Alcohol Sales Over Time

2.2.3 County Income vs Alcohol Sales

A comparison of average income per county against alcohol sales per capita revealed a more nuanced insight. While wealthier counties such as *Dallas* and *Polk* have higher sales per capita, the correlation between income and sales was weaker than the correlation between population size and sales. This suggests that while income influences alcohol consumption, population density and overall size are stronger predictors of liquor sales trends.



```
Correlation between Alcohol Sales and Income: 0.05
```

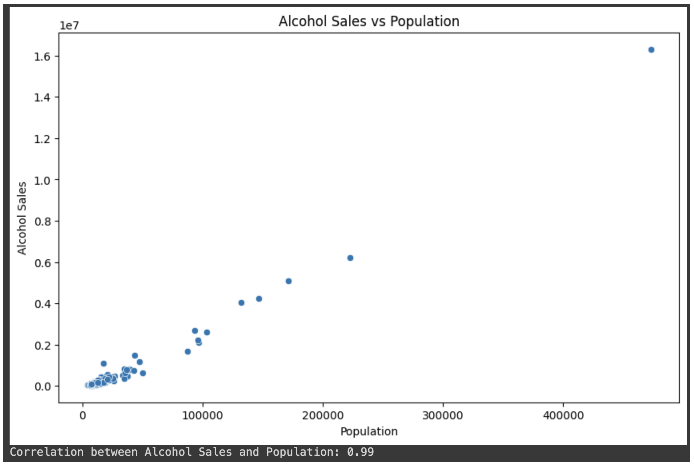
Figure 12: Income vs Alcohol Sales

2.3 Overview of their Relationships

Here we looked at how a handful of variables relate to each other in order to look for patterns in the data. We wanted to see how strong the correlation was between sales and population both in raw totals as well as per capita. We also wanted to see how the average income of a specific county related to the total sales in that county.

2.3.1 Alcohol Sales vs Population

First we wanted to look at how the population of a county translated to the total alcohol sales in that county. We had already assumed that the larger the population was would translate into more sales. However we just wanted confirmation of that as well as to see if there were any outliers. The results were exactly as expected and the correlation coefficient shows an almost perfect relation.

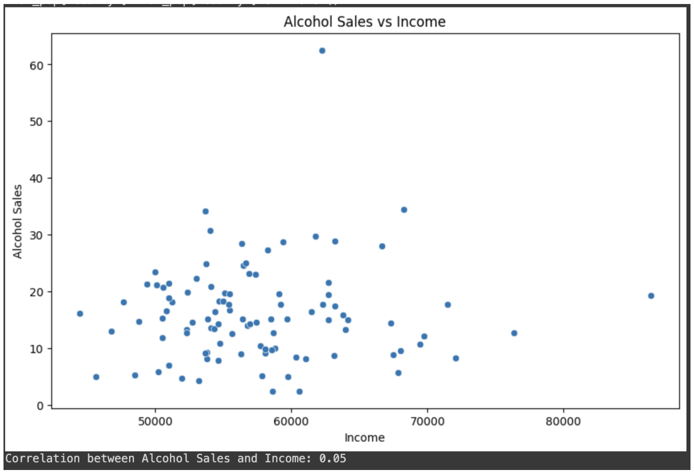


```
Correlation between Alcohol Sales and Population: 0.99
```

Figure 13: Scatter plot of population with total sales

2.3.2 Alcohol Sales vs Income

Next we wanted to look at how the average income of a county factored into the total sales. In order to eliminate the bias of a county's population we added a new feature to our data known as the sales per capita, which – . Our goal here was to see what kind of patterns would arise, as a study [5] identified within our literature review suggested that certain income groups have higher rates of binge drinking. However when we plotted these variables our results differed and we found a weak correlation of 0.05.



```
Correlation between Alcohol Sales and Income: 0.05
```

Figure 14: Scatter plot of population with total sales

2.3.3 County vs Income

Next we wanted to get a good idea of the overall distribution of income across all Iowa counties. This is more to get a picture of where most counties sit in relation to the entire state. What is interesting is we found that the income distribution has a skew to it. While there is a decent range from the mid 40k to over 80k, the bulk of the counties are in the lower half with the median appearing to be in-between 55k and 57k.

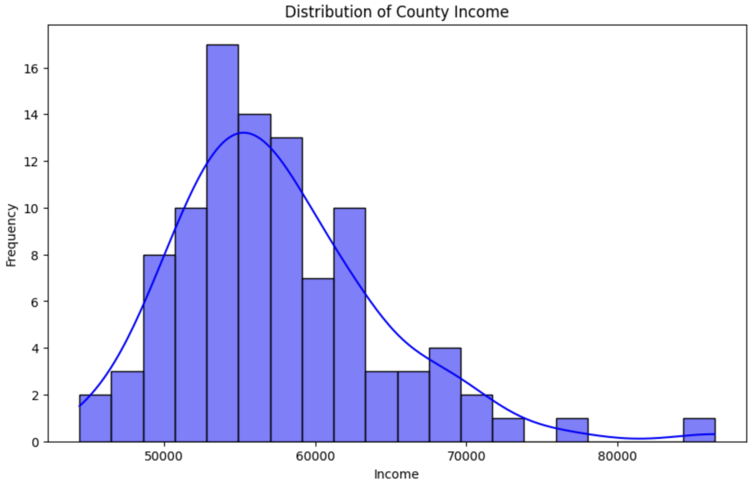


Figure 15: Income distribution by county

2.3.4 County vs Alcohol Sales

Finally, we wanted to check the sales per capita for each county as a histogram. What is interesting is the distribution is very similar to the county vs income distribution. This can be shown as the kernel density estimates have the same basic shape. Not only that but both have an outlier that is pretty far from the rest. We decided to check if they were the same county and they were not. Dickinson County has a per capita sales that is almost double that of the second highest. This is something we will be looking into but a quick look on its wikipedia page has it as a popular tourist hotspot which could influence its alcohol sales.

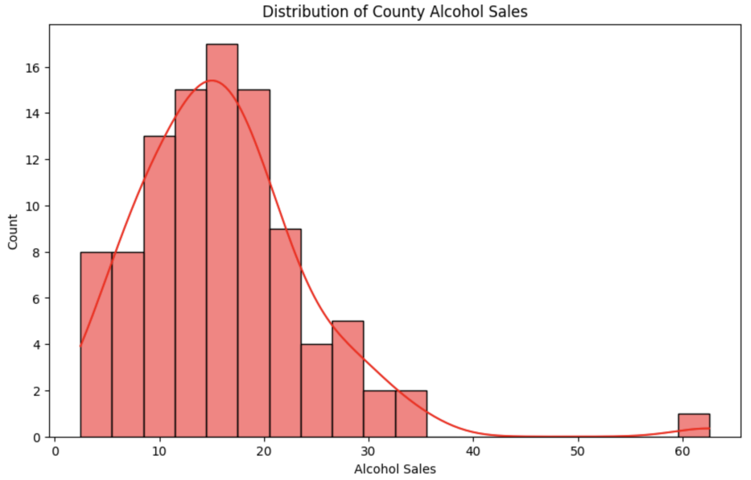


Figure 16: Sales distribution by county

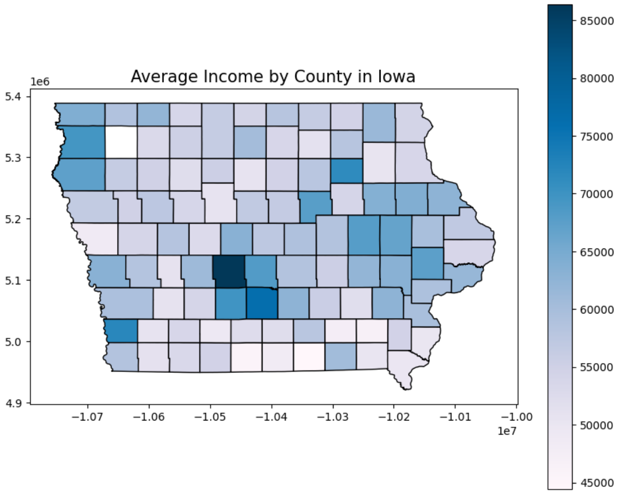


Figure 17: Geographic representation of income distribution by county

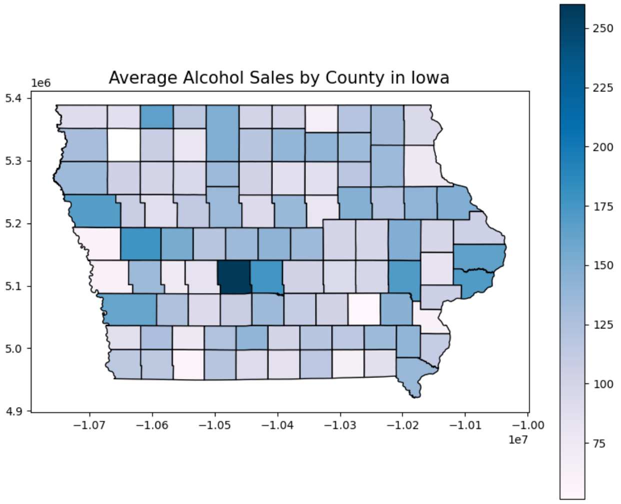


Figure 18: Geographic representation of sales distribution by county

3 Conclusion

For our initial EDA, the focus was on understanding population changes within the state of Iowa. Overall, state population reflects a slight increase. This increase in population is reflected in the State's most populous counties, such as Polk and Linn counties, whereas less populous counties have shown stagnation or decline in their population. This reveals specific counties of interest for further study, since our EDA has also revealed that larger populations result in higher alcohol sales. Another focus of the EDA was investigating income data within the state of Iowa to understand the influence of income on choice of alcohol and amount of alcohol purchased. Though the state of Iowa has seen steady increases in average income between 2010-2022, our initial study shows that income is not a strong predictor within liquor sales. The relationship between income and liquor sales will continue to be investigated within our project and will be further detailed in the

Methodology section. Finally, this initial EDA has shown steadily increasing alcohol sales with a peak in 2023. Alcohol volume purchased peaked in 2021, which was also when Iowa passed a new law that allowed to-go alcohol sales [4] to help businesses impacted by COVID-19. Additional analysis into any changes within the types of alcohol will be conducted as part of the Methodology.

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] S. L. of Iowa. State data center of iowa. Accessed: 09-16-2024.

[4] I. D. of Revenue. To-go carryout cocktails.

[5] L. Ta. Binge drinking rates are higher in iowa than other states, 2024.