Iowa Liquor Sales

Mayra Weidner Linnea Fritz-Watson Brittany Bilotti Margarita Espinoza

1 PROBLEM STATEMENT

This dataset contains information regarding the purchase of spirits in the state of Iowa by commercial establishments holding a [1]Class "E" liquor license. The original dataset contains purchasing data from January 1, 2012 to current. The following questions we seek to answer focus on trends and anomalies surrounding the sale of liquor to commercial vendors in the state of Iowa.

1. What times of the year have liquor sales been the highest?
2. Are there hotspots in the state where liquor sales have been higher than the average?
3. Can we see if there are preferred liquor types for regions or cities?
4. Is there a surprising/unexpected time of year when liquor sales have gone up?
5. Are there purchase trends during holidays and college football season?
6. Are there any alcohol types that are frequently bought together?

2 LITERATURE SURVEY

2.1 Iowa Liquor Sales Explorer: The State of Iowa used its liquor sales data to create an explorer that allows users to filter based on various attributes. The explorer's visualizations update based on user filtering [1].

2.2 Data on Liquor Sales Analysis in Iowa: Kang's data mining approach was based on total sales per store. Store location information was used to map store locations and determine active stores. Kang used K-means clustering to produce patterns in store locations and bottle sizes versus retail cost. Kang concluded that Des Moines, Polk would be an ideal location for a new store that would be most profitable if inventory bottle sizes ranged from 450-1750 ml. [2].

2.3 Iowa Retail Liquor Sales Data Analysis Using Linear Regression Model: Baskara used visualizations including a heatmap and geocoding to gain a preliminary understanding of the data. Based on the knowledge gleaned, Baskara built a machine learning linear regression model to predict sales in 2020. The linear regression model predicted 2020 liquor sales accurately with a variance score of 98%. Baskara comments that the high variance score is a direct result of the clean data and the fact that a sample population was used [3].

2.4 Visualization and Analysis of Liquor Sales in Iowa: Liang set out to determine what factors drive liquor sales ie., store location, time of the year, etc. Liang used a heat map to track counties with the highest sales known as "hotspots". Further exploration was performed with various visualizations. Liang also performed a two-sample hypothesis testing and machine learning multiple linear regression modeling [4].

3 PROPOSED WORK

3.1 Data Reduction: The original Liqour_Sales.csv dataset has over 19 million records from the years 2012 to 2020. We reduced the number of rows by randomly selecting a sample from the years 2012 to 2015. Our data mining will be based on this sample.

To further reduce our data we removed the following attributes that we considered unnecessary to answer the questions we proposed in our problem statement:

- Store Number
- Store Name
- Vendor Number
- Vendor Name
- Item Number
- State Bottle Cost
- State Bottle Retail

[1] *A Class "E" liquor license is a license required for grocery stores, liquor stores, convenience stores, and other commercial establishments to sell liquor for off-premises consumption in the original unopened containers.*

- Volume Sold (Gallons)

We removed Address, County Number, and Category as we deemed these attributes redundant because other attributes held similar information that we could use to answer our proposed questions.

Lastly, we split the Invoice/Item Number attribute into two separate fields to determine which invoices had only one line item. We removed rows with only one line item as these invoices would not be useful in determining purchase frequency.
Our final data sample has 1,004,380 rows.

3.2 Data Cleaning: We deleted rows that have null values as these rows make up a minimal percentage of total transactions (about 0.15%). The following is the breakdown of null values as a percentage of total rows:

- County - 0.10% (1,038 rows)
- Category Name - 0.05% (490 rows)
- Sale (Dollars - 0.0004% (4 rows)

We also deleted rows with $0 in sales.

3.3 Data Integration: Our full dataset was provided by the State of Iowa [1] so data integration from another source is unnecessary.

3.4 Data Transformation: We recategorized the Category Name attribute using a broader label as this attribute originally had many duplicate names with very specific descriptions of alcohol types. After re-grouping the liquor types we ended up with 15 unique labels.

3.5 Data Processing: We chose attributes from the dataset to find frequent patterns and correlations based on the problem statement questions we proposed. The problem statements are in the same vein as Liang's "hotspots" in that we set out to find what drives liquor sales by location and time of year. Unlike Liang's machine learning multiple linear regression modeling, we used frequent pattern analysis to find correlations [4]. For pattern evaluation we decided to use Apriori as Python has a library for this algorithm and it runs efficiently on our dataset. We decided against the use of Bayesian Classification and Decision Tree Classification to find correlations between the dataset attributes, as they are not suitable for our dataset. Furthermore, we evaluated the mean, minimum and maximum values of our dataset's attributes.

4 DATA SET

4.1 Dataset: The original Liqour_Sales.csv dataset was provided by the Iowa Department of Commerce, Alcoholic Beverages Division: it contains 19,666,763 unique values and 24 attributes, putting it at 4.77 GB file size [1]. Due to the enormity of the dataset, we created a sub-dataset to facilitate processing time. The sub-dataset was generated by randomly pulling a little over 1 million rows, resulting in nearly four consecutive years worth of sales data. We also removed several attribute columns that we deemed unnecessary for the purposes of this project. Of the original 24 attributes, we kept 14 for use. These remaining attributes provided us with information pertaining to the sell date, type of spirit, price, volume, and store location. This sub-dataset was downloaded onto each team members' local machine.

5 EVALUATION METHODS

During and after the processing of the dataset, we evaluated the results to determine whether the data could in fact provide answers to our proposed questions. To evaluate the results, we used the following techniques:

5.1 Pattern tracking allowed us to make intelligent inferences about liquor sale outcomes by identifying trends and patterns in our data.

5.2 Association was useful for examining buyer behavior. By identifying interesting relations between variables in our dataset, we used the results to help answer questions related to shopping basket data analysis.

5.3 Visualizations granted us the ability to visually highlight patterns in data, instead of simply using numerical outputs.

All of the above techniques provided insights that we used to answer our questions.

6 TOOLS

6.1 Project Management: We used Zoom and Discord for project planning and communication. GitHub was used as our project repository, along with the Google Drive (i.e. Google Docs, Google Sheets, Google Slides) to organize project artifacts.

6.2 Data Processing: To clean and process our data, we utilized a combination of Excel, Google Sheets, and Python within JupyterLab.

6.3 Data Visualization: For data visualization we are planning to use Tableau, Python, and/or PowerBi.

7 MILESTONES

We have outlined milestones for each remaining week until project parts 3 through 7 are due. The tasks have been split up according to skills and time availability of each group member while assuring that each member performs a relatively similar amount of work. Our general goal is to have all of the data processing finalized by April 24th. This will allow us two weeks to complete the final write-up, create any optional data visualizations, put together the voice-overed presentation, and complete any other last adjustments.

7.1 Milestones Completed

Week 3/6: Data Processing & Data Cleaning

- Decide which attributes to utilize and how to clean up NULL values in the data set - ***Completed***
- Assign roles for each task - ***Completed***

Week 3/13: Data Cleaning

- Data Reduction: Make tables based on attributes in excel - ***Completed***

Week 3/20: **Part 2 Due** & Data Cleaning

- Data Transformation with Excel **- *Completed***

Week 4/3: Data Processing

- Data processing discovery **- *Completed***

Week 4/10: Data Processing

- Finalize decisions regarding the appropriate algorithms to be used on apriori, fp growth, training model with Bayesian or decision tree. **- *Completed***
- Discover correlations and patterns with selected algorithm **- *Completed***

Week 4/17: Data Processing

- Finalize data processing **- *Completed***
- Review correlations and create graphs to visualize results **- *Completed***
- Update paper for Part 3 submission **- *Completed***

Week 4/24: **Part 3 Due**

- Compile completed milestones **- *Completed***

7.1 Milestones To-Do

Week 4/24:

- Update paper for final submission

Week 5/1: Finishing Touches

- Complete any remaining tasks before final submission.
- Create interactive data visualizations *(optional)*
- Record voice-overs and add to presentation
- Push all changes and artifacts to GitHub

Week 5/8: **Parts 4-7 Due**

- **Submit Project**
- **Project Presentation**
- **Peer Evaluation & Interview Questions**

8 RESULTS SO FAR

Each of the following subsections covers the results found so far regarding the questions stated in the section 1 Problem Statement.

8.1 *What times of the year have liquor sales been the highest?*

2012: In 2012, the sum of liquor sales for each month ranged from $1.9M to $4M. The month with the highest average sales is October at $4M which is more than $800k higher than December. December has the second highest sales with $3.2M. The 3rd highest month is May with $2.6M in total liquor sales.

1.	$4,022,641.40	OCTOBER 2012
2.	$3,217,669.67	DECEMBER 2012
3.	$2,659,377.44	MAY 2012

2013: Total liquor sales for the year 2013 are very similar to 2012, where again October comes in at the highest at $3.7M. The second and third highest months remain in the similar times of the year. December again has the second highest average at $3.3M and July has the third highest with $2.7M.

1.	$3,794,459.91	OCTOBER 2013
2.	$3,339,908.83	DECEMBER 2013
3.	$2,744,631.97	JULY 2013

2014: The months with the highest total sales in 2014 remain consistent with the previous years, however October has lost its place in first to December with $3.6M. October follows closely in second with approximately $400k less than December at $3.2M. June takes third place with $2.9M, which holds true that the summer months continue to be the 3rd highest sales per year.

1.	$3,689,630.75	DECEMBER 2014
2.	$3,257,069.45	OCTOBER 2014
3.	$2,936,948.46	JUNE 2014

2015: In 2015 we saw a shift in the month with the highest sales, June - one of our summer months - comes in at the top with $3.2M sales. October remains in the top 2 with $3M, and September has the third highest sales with $2.8M. This is the first time we see September in the top 3 and the first time a winter month is not the highest in total sales. Of note, December 2015 liquor sales were not included in our sub-dataset.

1.	$3,244,972.19	JUNE 2015
2.	$3,003,592.72	OCTOBER 2015
3.	$2,894,814.62	SEPTEMBER 2015

8.2 *Are there hotspots in the state where liquor sales have been higher than the average?*

Average Total Sales: We initiated this process by calculating the average total sales across all counties and cities in Iowa over the scope of all 4 years. The average total sales for all counties is $1,267,929.80 and the average total sales for all cities is $319,402.16. We then calculated the sum of total sales for each individual county and city and compared them to their averages.

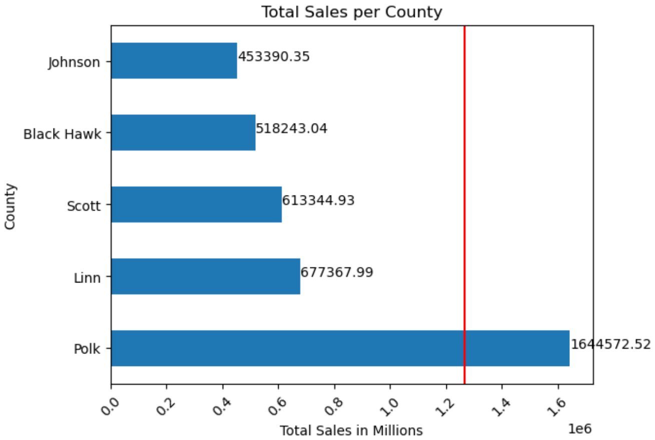

Figure 2.1 - Top 5 Total Sales per County

Total Sales by County: Our question asks if there are any ‘hotspots’ in the state where sales are higher than average. To find out which counties are considered ‘hotspots’ we took the top 5 counties with the highest total liquor sales and plotted them against the county average line.

According to our graph, out of the top 5 counties only Polk county has total sales over the $1.2M average with $1.6M. Linn county has the second highest total sales at $677k, this is a more than $900k difference between the top 2 counties.

1.	**$1,644,572.52**	**POLK**
2.	$677,367.99	LINN
3.	$613,344.93	SCOTT
4.	$518,243.04	BLACK HAWK
5.	$453,390.35	JOHNSON

This discovery leads us to believe that the average total sales across all counties is skewed by the tremendously large number of sales in Polk county. This leaves room for further investigation on how population or county size impacts these metrics.

Total Sales by City: We followed the same process for identifying if a city is considered a ‘hotspot’ as we did for counties, but we found that 4 of the top 5 performing cities were above the city average of $319,402.16.

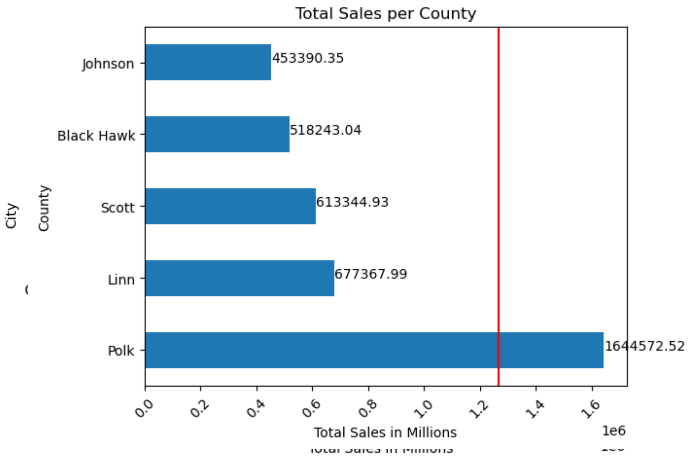

Figure 2.2 - Top 5 Total Sales per County

1.	**$1,048,027.83**	**DES MOINES**
2.	**$520,319.92**	**CEDAR RAPIDS**
3.	**$433,119.96**	**DAVENPORT**
4.	**$339,092.76**	**WATERLOO**
5.	$311,628.38	IOWA CITY

In conclusion, we can confidently say that Des Moines, Cedar Rapids, Davenport, and Waterloo can be considered ‘hotspots’ for liquor sales in the state of Iowa.

8.3 *Are there preferred liquor types in the state of Iowa?*

Over the four year time frame in our data, there are four liquors that were notably higher in sales; they are as follows:

Liquor Type	***Total Number of Sales***
Liqueur	101,368
Rum	125,322
Whiskey	192,711
Vodka	255,164

The other liquors shown in the data have significantly lower overall sale totals, all sitting under 60,000 units sold over the four years. For our final report we will attempt to narrow down the liquor preferences by sales amount per city. Since the sale reflects how much and what types of liquor vendors are purchasing, this information could help identify trends in liquor consumption by city.

8.4 *Is there a surprising/unexpected time of year when liquor sales have gone up?*

2012: In 2012, liquor sales began to increase after June with a noticeable drop off in July. It is likely that liquor stores, in anticipation of the Fourth of July holiday, purchased more inventory in June. Sales increased in August and continued to do so through November. In December sales decreased, but still remained higher than the first ten months of the year. It was surprising to see that fall and winter months have higher liquor sales than summer months as people tend to take vacations and spend more time outdoors during the summer. These activities tend to involve liquor consumption. The liquor sales average was highest in October which is 24% above the second highest month. This is true for 2013 through 2015 as well. We noted that October coincides with the college football season, baseball playoffs, Columbus Day, and Halloween so it’s likely that these holidays and major sporting events contributed to the continued increase in liquor sales.

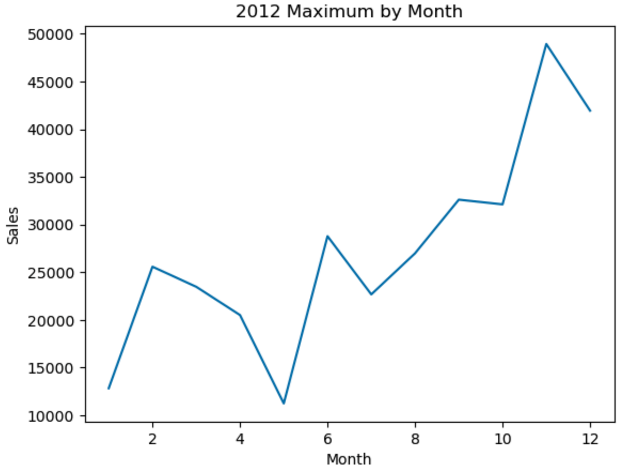

Figure 4.1 - 2012 maximum sale dollars by month

2013: There was an unexpected drop off in total liquor sales that occurred in 2013. We can only speculate as to the actual root cause of this reduction.

Unlike in 2012, Liquor sales increased drastically in March. March coincides with college spring break so this likely drove the increase in liquor sales. This trend was not present in 2014, but resurfaced in 2015. We expected to see a similar trend during March for all four years. We also noted that sales for September drastically decreased compared to 2012.

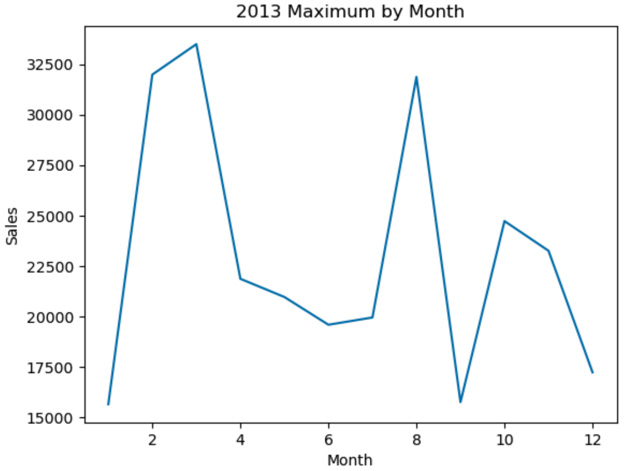

Figure 4.2 - 2013 maximum sale dollars by month

2014: In 2014, liquor sales in summer months were surprisingly lower than fall and winter months. We saw this as well in 2012. Unlike in prior years, liquor sales were the lowest in August compared to the rest of the year. This was unexpected as sales started to pick up in August in prior years and in 2015.

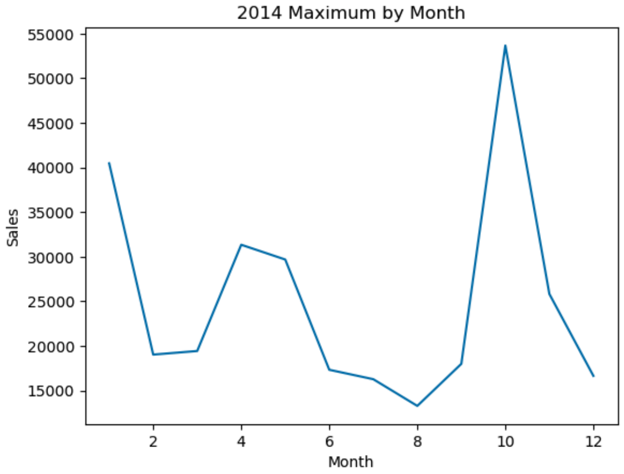

Figure 4.3 - 2014 maximum sale dollars by month

2015: Unlike in prior years, liquor sales in 2015 were the lowest in July compared to the rest of the year. In fact, July sales were 36% lower than the July sales in the prior years. We can only speculate as to what caused this drop in sales. There were no other unexpected trends that have not been previously discussed.

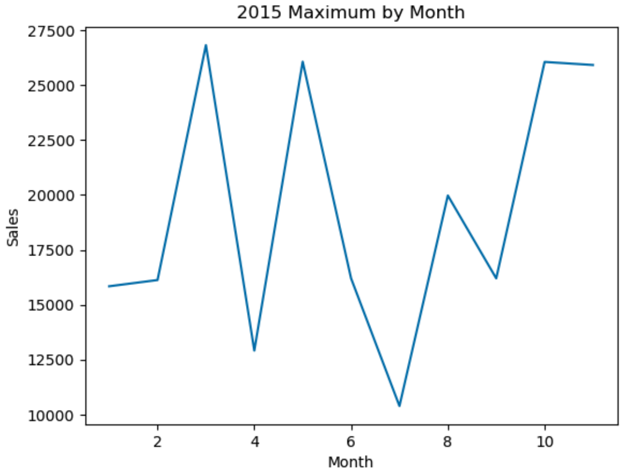

Figure 4.4 - 2015 maximum sale dollars by month

October average daily sales: As mentioned above, October has the highest average sales in the year for all years sampled so we drilled down into the daily average liquor sales for October.

2012: In 2012, we found that there were unexpected spikes that occurred on the 5th, 12th, 19th, and 26th days of the month.

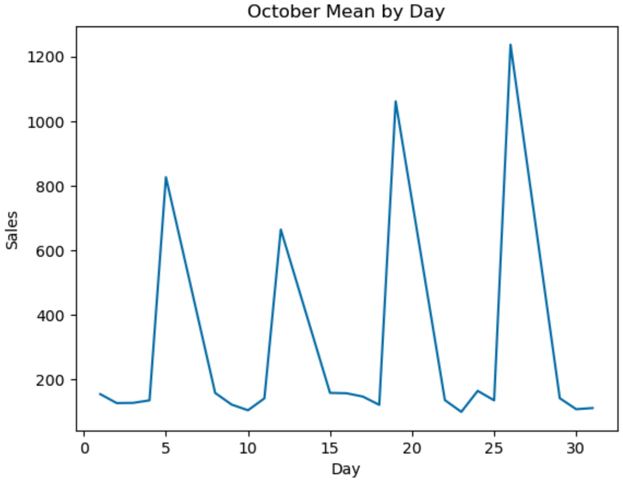

Figure 4.5 - 2012 mean sale dollars for October

2013: This thread continued in 2013 with the only exception being a liquor sales lag between the 16th until the next spike on the 25th. Of note, the 16th had the highest liquor sales average in October so it’s likely that liquor stores in general ordered more on this date and therefore needed less inventory for the following week. The results were unexpected as compared to last year simply because 2012's sales pattern was consistent.

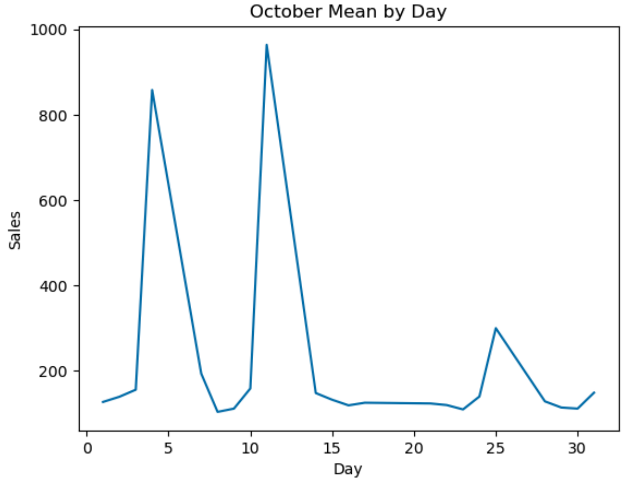

Figure 4.6 - 2013 mean sale dollars for October

2014: We found a similar trend in 2014 with spikes on the 2nd, 6th, 13th, 20th, and 27th days of the month which is more in line with what we saw in 2012. Interestingly though, these spikes were significantly smaller than the 2012 spikes, with a difference of approximately 450%.

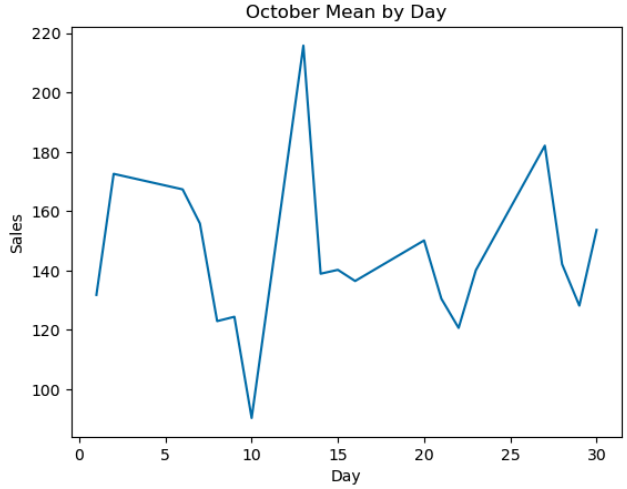

Figure 4.7 - 2014 mean sale dollars for October

2015: The 2015 trend was similar to 2014 with similar spikes on the 1st, 5th, 12th, 15th, 22nd, and 29th days of the month.

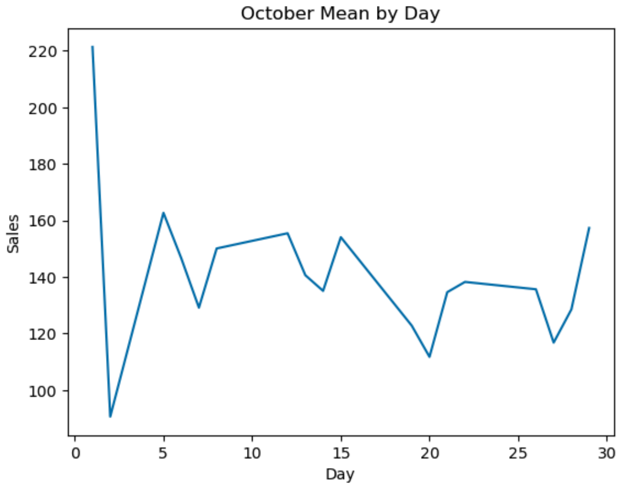

Figure 4.8 - 2015 mean sale dollars for October

We theorized that the spikes in liquor sales were due to liquor store reorder schedules. Liquor stores appear to be ordering in bulk about once a week in October.

8.5 *Are there purchase trends during holidays and college football season?*

It should be noted that we chose 'holidays' that we as a team felt were the most widely known/celebrated in the USA. These holidays are listed below.

The sales data in our dataset reflects the purchasing of liquor, which does not include beer or wine, by vendors in the state of Iowa. Therefore, sales spikes do not directly line up on the holiday's date, but rather spikes often appear prior to and/or shortly after a holiday, for what we believe are for preparatory and restocking purposes, respectfully. With this in mind when reading the results, discretion is used to make inferences regarding spikes and their correlation to a given holiday. We also took into account the difference in date(s) for holidays with fluctuating dates depending on the year. Discretion was also taken when speculating the sale of liquor during college football season.

St Patrick's Day: Sale spikes are relatively consistent throughout February and March of 2012 through 2014, with 2015 being the only year showing an outlying spike in sales right after St. Patrick's day. This could mean individual consumer sales were higher than usual around 2015's St Patrick's day and therefore restocking by vendors was required.

Easter: Sales dates surrounding Easter in 2012 remained relatively constant, however, there were small spikes in sales directly after Easter in 2013 and 2015, with a larger spike also shortly after Easter in 2014. This signifies that there was a trend of restocking by vendors after Easter, possibly because it's a more unpredictable holiday for liquor sales by vendors.

Cinco de Mayo: Sales spikes shortly before and after Cinco de Mayo in 2013 through 2015 suggest a trend in vendors either preparing or restocking liquor. Though it only appears to be one or the other considering that in any given year, there is either a pre-holiday or post-holiday spike, but not both in the same year. However, in 2012 there were no notable spikes in sales during the given time frame.

4th of July: July's average sales across all four years is surprisingly low. The main spike that stands out is in 2014. The lack of sales could be due to the fact that our dataset only contains liquor sales and does not account for beer and the like, which are potentially the more common beverage of choice in the warmer summer months.

Halloween: Average sales overall increase in October by over 200% in 2012 and 2013 and by approximately 30% in 2014. However in 2015, October average sales remained relatively the same as the previous month. This difference in 2015 could be due to vendors learning from previous years and preparing for an increase in individual consumer sales, but this is hard to verify without data for the following years.

Thanksgiving: In 2012, 2013, and 2015 there are sales spikes just prior to Thanksgiving, which shows a prestocking trend by vendors. In 2014 we don't see any spikes, though we do note that the average sales in the month of November are steadily higher than in the other years. Perhaps there is no spike in 2014 because vendors had plenty in stock from purchases earlier in the month.

Christmas & New Years (*Note that there is no data for December 2015.)***:** Average sales for all three years are higher in December, but there are no dramatic spikes neither prior to nor shortly after these holidays. This leads us to believe that the pre-holiday higher average sales trend is in preparation for these two big holidays.

College Football Season *(late August - early December)****:*** It should be noted that football season overlaps with several holidays, therefore the results may be slightly skewed. We think that the increase in average sales from August through December is most likely due to a combination of holiday and sports seasons over that time period.

From these findings we can conclude that there are yearly trends for certain holidays, namely St Patrick's Day, Easter, Cinco de Mayo, Halloween, Christmas/New Years. We were surprised to find that October held a higher sales average over all four years compared to November or December. This is especially interesting since December is said to be the busiest month for individual consumer sales. Because of this we expected to see spikes in sales in November and throughout December to account for increased consumer sales. However, perhaps the lack of spikes is due to the trend in average sales to vendors increasing throughout December, which would signify that said trend is working well for vendors.

8.6 *Are there any alcohol types that are frequently bought together?*

To find the frequent sets of alcohol that were purchased together we first used Excel to split the Invoice/Item Number attribute to determine if there were multiple items purchased per invoice. From this process we were able to determine that there are 148,431 invoice numbers that have more than 2 items within our dataset. With these invoice numbers we were able to create individual data frames that held the antecedents (item(s) purchased), consequents (item(s) likely to be purchased with the antecedent), support, confidence and lift for each county based on our chosen minimum support. We decided to choose a minimum support of 0.40 to filter out the antecedents and consequent pairings. We chose this because lower supports gave us many pairings without strong confidence and lifts and higher supports did not give us enough pairings for each county. With this we were able to generate a list of antecedent and consequent pairings and their frequencies that spanned across all counties. Below is a list of pairings with the counts of county frequent item set occurrences, average support, average confidence, and average lift for each frequent item set. Please note that the average support, confidence, and lift have been rounded to the nearest third decimal place:

Pairings	Count	Support	Conf	Lift
Whiskey→Vodka	44	0.521	0.743	1.017
Vodka→Whiskey	44	0.521	0.712	1.017
Rum→Vodka	25	0.465	0.807	1.047
Vodka→Rum	25	0.465	0.602	1.047
Rum→Whiskey	21	0.446	0.764	1.049
Whiskey→Rum	21	0.446	0.609	1.049
Liqueur→Vodka	8	0.448	0.845	1.055
Vodka→Liqueur	8	0.448	0.557	1.0547
Liqueur→Whiskey	5	0.470	0.843	1.068
Whiskey→Liqueur	5	0.469	0.591	1.068
(Rum, Whiskey) → Vodka	4	0.493	0.865	1.050
(Vodka, Rum) → Whiskey	4	0.493	0.847	1.050
(Vodka, Whiskey) →Rum	4	0.493	0.742	1.091
Rum → (Vodka, Whiskey)	4	0.493	0.719	1.091
Whiskey→ (Vodka, Rum)	4	0.493	0.608	1.050
Vodka → (Rum, Whiskey)	4	0.493	0.608	1.072

Given the frequent item sets above we can glean that the most frequently bought pairs of alcohol would be (Whiskey, Vodka), (Rum, Vodka), and (Rum, Whiskey) due to their occurring in many counties, their high average confidence and lift and their average support which exceeds our minimum support.

Our apriori algorithm also yielded results where the pairings seemed unusual, but only appeared in one or two counties. We felt that these pairings were interesting and show that specific counties can have strong preferences for certain types of liquor pairings and can be viewed as outliers in comparison to the rest of the state. Below you can see a table with the pairings and the counties they appeared in.

Pairings	Counties
(Schnapps, Rum) →Whiskey	Monona
(Liqueur, Vodka, Rum) →Whiskey	Monona
(Liqueur,Rum, Whiskey) → Vodka	Monona
(Liqueur,Vodka, Whiskey)→Rum	Monona
(Liqueur,Rum) →(Vodka, Whiskey)	Monona
(Liqueur,Vodka) →(Rum, Whiskey)	Monona
Whiskey→(Schnapps,Rum)	Monona

(Schnapps, Whiskey)→ Vodka	Jones, Monona
(Schnapps,Vodka) →Whiskey	Jones, Monona
Schnapps→(Vodka, Whiskey)	Jones, Monona

Given the results from the table, we can conclude that Jones and Monona county both have unusual purchase preferences in comparison to the rest of the state.

REFERENCES

[1] State of Iowa, 2012-present https://data.iowa.gov/stories/s/Iowa-Liquor-Sales-Explorer/mke2-7r5k

[2] Kang, J., 2022 https://nycdatascience.com/blog/student-works/data-on-liquor-sales-analysis-in-iowa/

[3] Baskara, J., 2021 https://jonando-baskara.medium.com/iowa-retail-liquor-sales-data-analysis-using-linear-regression-model-bbc22fc150f9

[4] Liang, R. https://richardcmsc320final.github.io/