Mining Iowa Liquor Sales

CSPB 4502 Semester Project

Part 4: Final Report

Team #4: Nicholas Stafford, Alexander Sueppel, Alex Thomas, Kevin Vick

ABSTRACT

Interesting Questions

The questions listed below are categorized into the respective data mining technique used to answer, or at least explore, the question. Our group sought to answer these interesting questions based on a dataset containing store liquor sales from the state of Iowa.

Trend Analysis

- What dates were alcohol sales at their highest?
- What is the most common alcohol volume bought?
- Which countries made up the majority of sales?

Classification

- Can the dataset be used to accurately predict whether an alcohol sale takes place within the vicinity of a university?

Clustering

- How can the interaction between liquor sales and volume sold be grouped?
- What categories of liquor have similar gross margins?

Outlier Analysis

- How is data distributed?
- Are there outliers and how many?

Summary of Results

Overall, we discovered alcohol sales are highest in the month of December and lower in the summer months than expected. The most popular volume of alcohol purchased was 750ml followed by 1.75ml. Polk, Linn, Johnson, Scott, Blackhawk (ranked from most to least) are the top five counties for alcohol sales. These five counties made up the significant majority of alcohol sales.

A classification analysis based on sales information and university location revealed the dataset is not ideal for this type of prediction. The highly skewed nature of the label support (i.e., many more sales not made near a university campus than made near a university campus) made it difficult to create an accurate classifier.

Among the clustering methods, sales and volume were grouped distinctly about the former using K-Means when optimized with four clusters. Additionally, liquors from related geographies and subsequently those with similar purities had the most similar sales margins after hierarchical investigation.

Many confirmed outliers exist in the numerical attributes. Without removing outliers, most of the numerical attributes are heavily skewed left on a distribution (low values) with significant high-value outliers on the right. Removing outliers using IQR (≈7% of the data) resulted in a more normal and less dense distribution.

INTRODUCTION

Question Descriptions and Importance

Alcohol sales throughout the state of Iowa were examined within the timeframe from January 1, 2012 to present. Long-term trends, such as preferences for certain brands, volumes, and varieties of liquors (e.g., vodka, gin, etc.), as well as shorter, seasonal trends (e.g., an increase in rum sales in the summer months) were of particular interest.

Further, geographic differences in preferences in Iowa, such as what liquor varieties and brands are popular in college towns throughout the state, are other interesting questions explored in this study.

The information derived from our data analysis can be applied to a number of commercial purposes. Liquor brands could use this information to decide where in Iowa to introduce new varieties of liquor. Additionally, companies could identify liquor varieties that are not popular in certain regions of the state and cut supply to save money.

RELATED WORK

We looked at a variety of different studies that had been done previously, relating to the sale of liquor, particularly in the state of Iowa. Through this, we hoped to gather ideas regarding what sort of connections could be made about the data. It also gave us a better understanding of what questions hadn’t been asked, and what new insights our study could bring to the plethora of knowledge surrounding this subject.

Weekdays/Months with largest liquor sales

As would be expected, the study shows a significant increase in liquor sales on the weekends as well as around specific holidays. The study also goes into the specific locations around the United States where liquor sales were highest, not surprisingly the East Coast dominated lesser populated Midwest states. When conducting our study, we found similar results in that alcohol sales rose around the holidays. However, this seems to only be true for more winter-based holidays rather than for summertime holidays [1].

Alcohol purchases vs other beverage purchases

Alcohol came in 4th place in terms of total purchases compared to other beverages in the U.S. falling just behind coffee, water, and soft drinks. Of the harder spirits that were purchased, vodka and whiskey together made up over half of the sales, however beer was by far the predominant seller in terms of total sales by one specific type of liquor [2].

Alcoholic Sales in Iowa Counties

This study looked at the total sales of each Iowa county over multiple years. Unsurprisingly, Polk County, Iowa's most populated county, had the largest sales of any of the counties featured. When looking at our dataset, there are really only 6 or 7 counties that make significant alcoholic sales in the state of Iowa. This is not surprising, as Iowa's main population resides in the more urban towns, whereas more rural counties have smaller populations, thus leading to less sales [3].

Preferred alcoholic beverages in Iowa

This study shows that Iowans follow the U.S.’s lead in preferring whiskey and vodka over other spirits. Interestingly, Iowa ranks in the top 5 in the U.S for excessive drinking. While the study shows that 40% of Iowans rarely drink at all, around 23% of Iowans report indulging in excessive drinking habits. Unfortunately, our dataset only parses out the information based off of brand rather than off of specific liquor type. Thus, it would be difficult to replicate this study [4].

Alcohol and college football

The findings in this study are staggering. Close to 50% of the individuals in this study reported excessive drinking when attending college football games. Only

12% of the individuals surveyed in this study abstained from alcohol completely when at a college football game. On average, participants guzzled down 5 drinks during the course of a single day of college football festivities [5].

DATA SET

The Iowa Liquor Sales data set contains information on liquor sales throughout the state of Iowa from January 1, 2012 to present and is updated on a monthly basis. The data set is maintained by the Alcoholic Beverages Division (Commerce) of the State of Iowa, so we can be confident in its correctness and integrity.

The liquor data set is extremely large, containing over 22 million data tuples. The set covers 24 attributes (columns), with information primarily on date of sale, type of product, transaction amount and location. Specifically, the attributes are titled as follows: Invoice/Item Number; Date; Store Number; Store 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); Volume Sold (Liters); and Volume Sold (Gallons).

Please note that given the size of the data set, some analysis described below was performed on less than the full data set. In certain cases, the algorithms being used could not efficiently run on the entire data set.

The data set can be found at the following location: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy.

MAIN TECHNIQUES APPLIED

Collection

The data and the majority of the code was run on the Kaggle platform. Kaggle allowed us to see each other’s progress and keep everything in one, centralized location.

Data Clean/Preprocessing

The dataset in its raw, untouched state did not require much preprocessing before mining. There is no need for any data integration (combining with other relevant/similar datasets) as the one data set already includes 24 attributes across 22+ million rows. Refer to the *Data Set* section for more details. The data set is rich in potential information and did not need anything more to ask and answer interesting data mining questions.

Additionally, data transformation was unnecessary. There are no apparent redundancies or obvious errors in the data.

The two preprocessing steps performed on this data set were data reduction and data cleaning. Data reduction focused on eliminating two attributes: “Store Location” and “Invoice/Item Number”. “Store Location”, containing latitude/longitude coordinates, was ignored because 1) there are missing values making it an incomplete attribute and 2) the necessary geographic information is already present in both the “city” and “zip code” attributes. There is also no need to be as precise as a coordinate for this mining project and with our objectives.

The other attribute that was ignored and eliminated was “Invoice/Item Number” because it is specific to one transaction and contains no meaningful or interesting information by itself. There are no relationships with other attributes that can be made with this number. Eliminating the two attributes “Store Location” and “Invoice/Item Number” reduce our data set to something slightly more manageable and, more importantly, make navigating and analyzing the set more efficient.

In a data cleaning preprocessing step, outliers were analyzed and identified. The outlier analysis is focused on the numerical attributes such as “Bottles Sold” and “Sale (Dollars)” to capture errors and anything to skew the analysis.

Outlier Detection/Analysis

The outlier analysis exercise for this project consisted of three parts: 1) determining if outliers were present by Grubb’s test 2) testing various outlier detection methods and 3) viewing the results of removing outliers.

First, Grubb’s test was performed on six numerical attributes to determine which one may contain at least one outlier. If the calculated value is greater than the critical value, an outlier may be present in the data.

Next, an exploration of different outlier methods was conducted using the attribute with the highest calculated Grubb’s test value and the most interesting attribute which is “Sale”. We ran the z-score method, interquartile range (IQR) method, and Winsorization method which is a tighter IQR method.

The final step in the outlier analysis exercise was removing the outliers detected by each of the methods and viewing the results through box plots, distributed plots, and pair plots.

Trend Correlation Analysis

Through analyzing different data points within the dataset, we were able to create charts that show the relationships between different attributes and use general statistics to determine trends. While memory limitations prohibited access to all attributes at the same time, we were able to create visual representations, particularly related to the total sales, dates of purchase, bottle volume, and counties that show the most activity.

Classification Analysis

Classification analysis was performed on only a select few features. The naïve Bayes and the k-nearest neighbor methods were applied to our data set. There is substantial conditional dependence between attributes in the data set. With inventory specific to a store and vendors specific to a region as common examples, having knowledge of one attribute can provide information about another. As such, the simplifying assumptions of the naïve Bayes approach failed. As the k-nearest neighbor method does not make assumptions about the relationship of the data, we found this method to be more appropriate for our data set.

Clustering Analysis

By treating the data as if it was unlabeled for clustering, intrinsic structure was analyzed beyond the explicit category-based rules of the supervised methods. K-Means was run on the numerical volume and price attributes to jointly partition the data over four clusters, an optimization based on relative contribution to fit suggested by the “Elbow method” heuristic [6].

Separately, by incorporating the categorical attribute of liquor type in conjunction with the quantitative abstraction of gross margin (the simple difference between price and cost) reduced to its product-specific mean, a dendrogram based on visualizing margin similarity was produced. This agglomerative hierarchical approach was undergirded by Ward’s method (a more complex distance metric that minimizes sum of squared error) and allowed us to cascade a measure of relative comparison over increasingly abstract tiers.

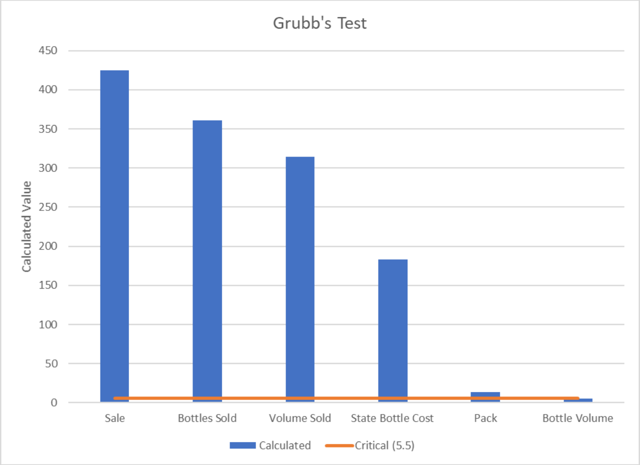
KEY RESULTS

Preprocessing

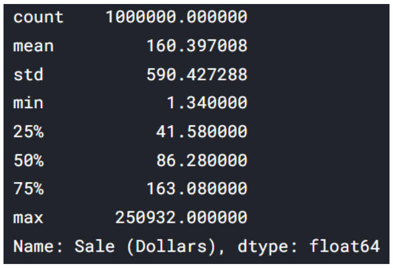
As anticipated, the ‘Store Location’ attribute contains a significant percentage of missing (empty/null) values. Of the million sample size rows, nearly 10% of values were missing. ‘Store Location’ was therefore excluded from our analysis.

Outlier Detection/Analysis

Using Grubb’s test, five of the six numerical attributes (Sale, Pack, Bottles Sold, State Bottle Cost, and Bottle Volume) contain at least one outlier. The numerical attribute that does not have an outlier (by Grubb’s hypothesis) is Bottle Volume.



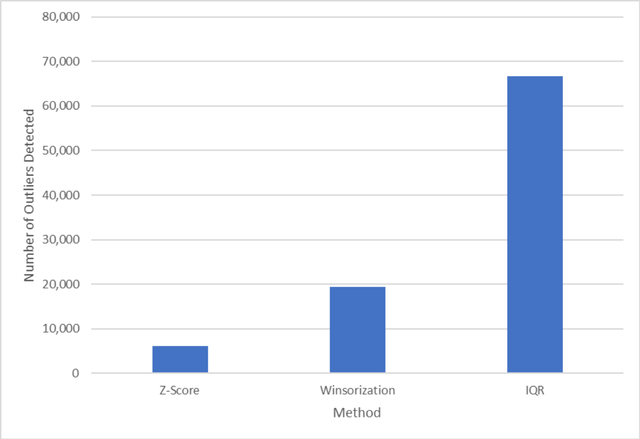
Three outlier detection methods were implemented and were tested on the ‘Sale’ attribute. By Grubb’s test, the ‘Sale’ attribute had the greatest difference in the Grubb’s calculated value versus the Grubb’s critical value. Additionally, simply using the pandas describe method shows the maximum value in ‘Sale’ is greater than 400 standard deviations from the mean.



```
count    1000000.000000
mean         160.397008
std          590.427288
min            1.340000
25%           41.580000
50%           86.280000
75%          163.080000
max       250932.000000
Name: Sale (Dollars), dtype: float64
```

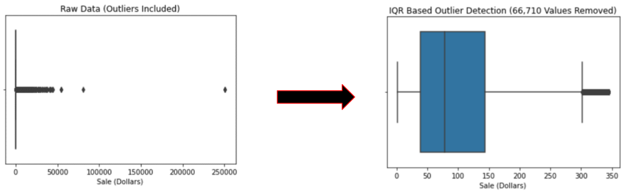
The three outlier methods analyzed in this project are Z-Score based, interquartile range (IQR) based, and Winsorization based. The table and bar graph below shows the number of outliers detected using each of the methods using a sample size of 1 million:

Method	**# of Outliers Detected (out of 1 million)**
Z-Score	6,129
Winsorization	19,480
IQR	66,710

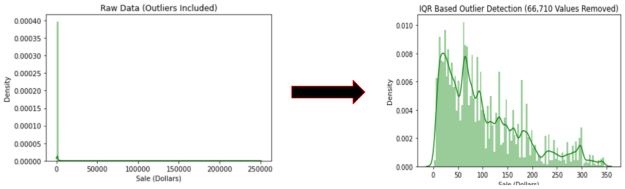


The IQR method detects significantly more than the other methods. If the outliers IQR detects are removed from the dataset, the distribution of the data is significantly altered. Below are boxplot and distribution plots of the original, unaltered data and data with outliers removed (using IQR).

Box plot transformation from raw data to removing outliers using IQR:



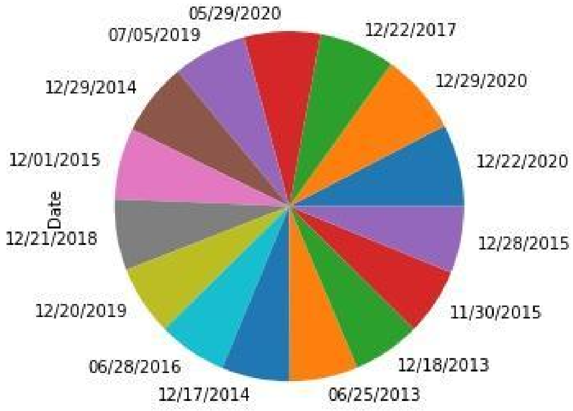
Density plot transformation from raw data to removing outliers using IQR:



Removing these outliers significantly affects the spread of the data and the distribution of the data. The box in the box plot is actually presented as well as the distribution of the data in the density plot.

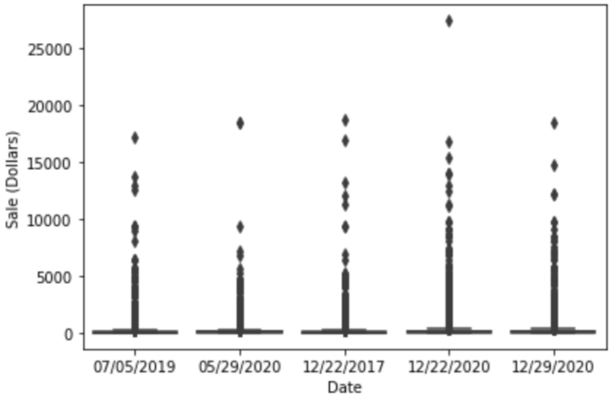
Trend Correlation Analysis

A count function was run on all of the calendar dates in which alcoholic sales were made. This was used to create a pie chart showing the 15 most common dates in which liquor was purchased:

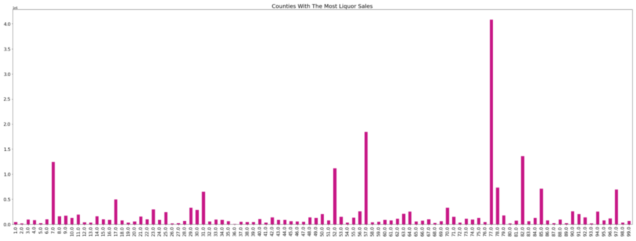


December clearly dominates this category. Also, surprisingly, there are no dates that are relatively close to July 4th, as you might expect there would be during the holiday.

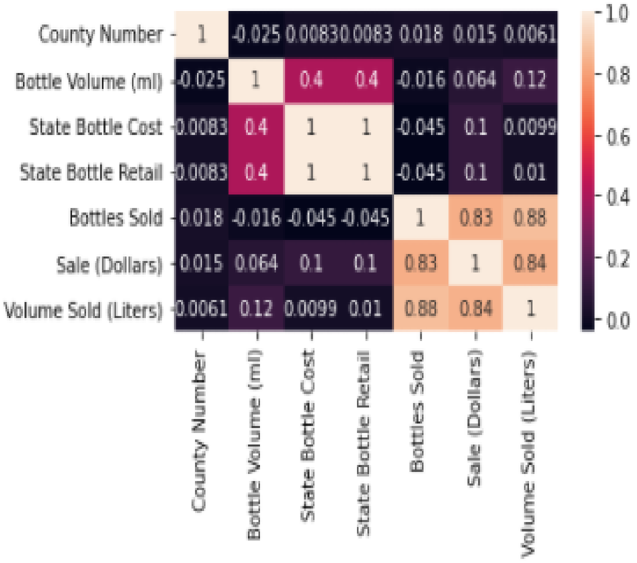
In the next visual we can see the average sales on the five most popular liquor sale dates within the dataset. The graph confirms that the average sale is higher for the December dates vs that of the summer dates.



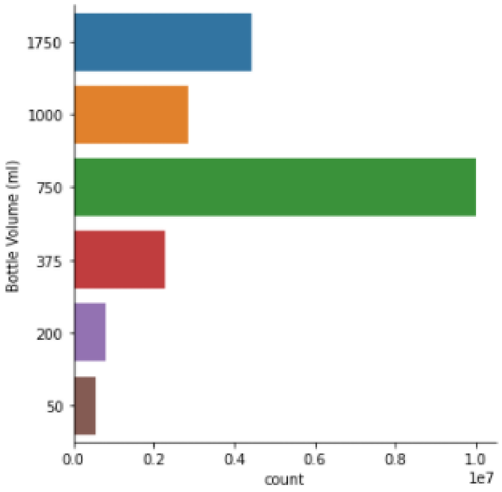
Additionally, a simple bar graph shows the counties that received the largest amount of sales. As the plot shows, there are really only 6 or 7 counties that make up >90% of the total sales. In the following visual we can see the differences of bottles sold between one major state county vs another. Polk County had the largest amount of bottles sold by far which is unsurprising as Polk County is home to the state capital of Des Moines, and makes up roughly 15% of Iowa's total population.



The next visual is a correlation table showing some of the correlation between specific attributes within the database. As you would expect the table verifies the connection between number of bottles sold, and sale dollars.



Finally, we have a simple horizontal bar graph showing the trend of Iowa consumers buying larger volume containers of hard liquor rather than smaller volumes. The graph shows that the most purchased volume, by quite a large margin I may add, is that of 750 ml. We found this surprising as we expected the 1750 ml quantity would be just as high if not higher than that of the 750 ml quantity. The reason why we expected this, is that at a higher quantity, cost per unit almost always goes down. However, it appears as though the people of Iowa predominantly prefer the 750 ml quantity instead by a factor of around 2.5x.

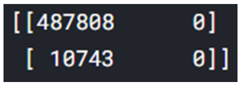


Classification Analysis

We examined two methods of classification during our data analysis. Namely, we utilized the naïve Bayes algorithm and the k-nearest neighbor algorithm using the sci-kit learn library in Python. We used these classification tools to predict whether a purchase was made in a zip code with a university or not. To label the training data, we accessed a list of universities in the state of Iowa and located their zip code. Based on this information, we labeled the tuples with a binary target class. A “True” value represented that a university was located in the zip code, while a “False” value represented that no university was located in the zip code. The inputs for the classification tools were the county the sale took place in and the retail price per bottle. Please note, the classification analysis was performed on 2.5 million tuples of data, with 80% (2 million tuples) being used as the training data and 20% (500,000 tuples) being used as the test data.

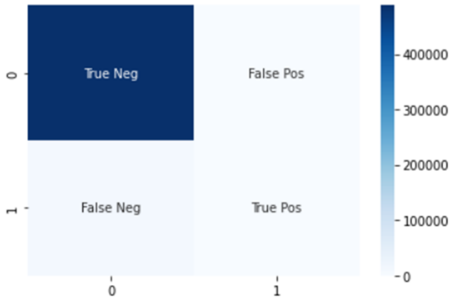
As can be seen in the analysis below, the naïve Bayes method was unsuccessful in its analysis. While at first glance the naïve Bayes approach seemed accurate, with an accuracy score of 0.978, closer examination revealed that the classifier predicted that no sales were made in the vicinity of a university. We hypothesized that this was due to the dependence and distribution of the underlying data. Naïve Bayes classifiers assume that the features are independent, which may not be true. There may indeed be dependence between various features in our dataset. Additionally, we specifically used a Guassian naïve Bayes classifier, which assumes that there is a normal distribution of values. The highly skewed distribution of the target support made a naïve Bayes classifier a poor choice for our dataset.

Naïve Bayes Confusion Matrix

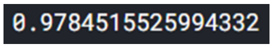


```
[[487808      0]
 [ 10743      0]]
```

Naïve Bayes Confusion Matrix Heatmap

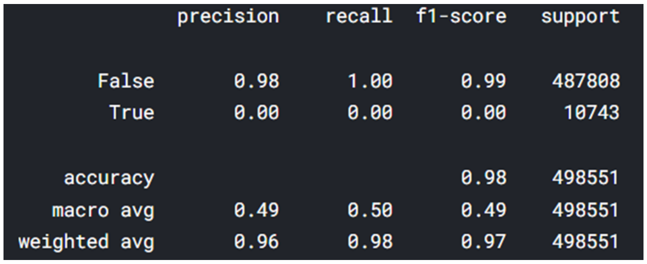


Naïve Bayes Accuracy Score



```
0.9784515525994332
```

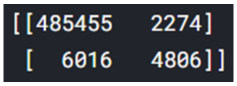
Naïve Bayes Classification Report



```
              precision    recall  f1-score   support

       False       0.98      1.00      0.99    487808
        True       0.00      0.00      0.00     10743

    accuracy                           0.98    498551
   macro avg       0.49      0.50      0.49    498551
weighted avg       0.96      0.98      0.97    498551
```

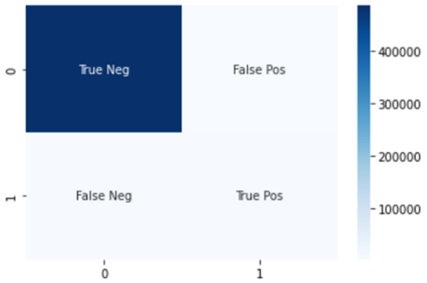
Given the issues laid out above with the independence, we hypothesized that the k-nearest neighbor algorithm would be better suited to our data. The k-nearest neighbor method is non-parametric, meaning that it makes no assumptions about the data, unlike a Gaussian naïve classifier. Our hypothesis turned out to be true, as the k-nearest neighbor analysis was more accurate, as can be seen below. The k-nearest neighbor classifier dramatically improved in the prediction of true positives, meaning sales were correctly predicted as taking place near a university. However, it still only predicted about half of the university transactions correctly, as demonstrated by the F1-score of 0.54.

K-Nearest Neighbor Confusion Matrix

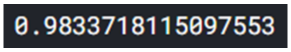


```
[[485455   2274]
 [  6016   4806]]
```

K-Nearest Neighbor Confusion Matrix Heatmap

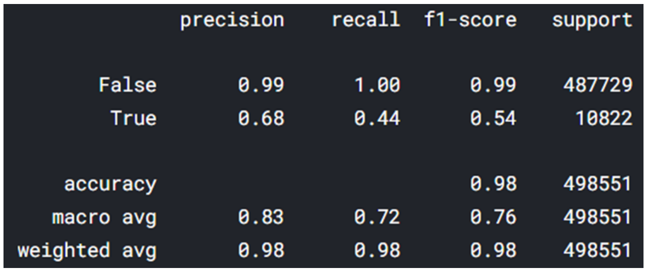


K-Nearest Neighbor Accuracy Score



```
0.9833718115097553
```

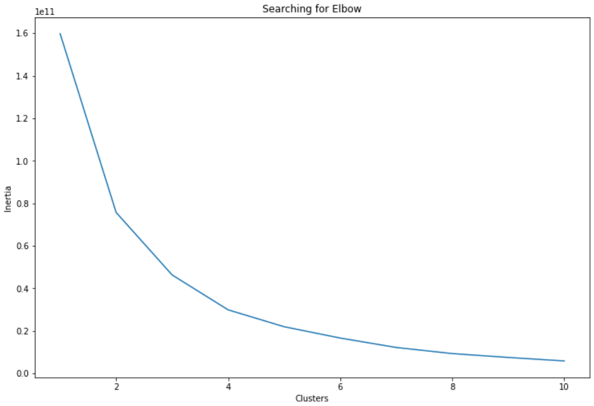
K-Nearest Neighbor Classification Report



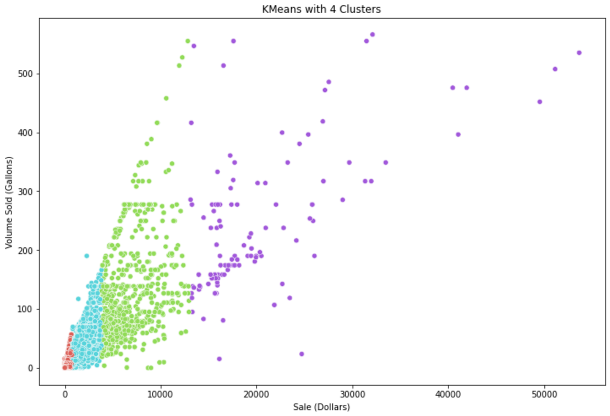
```
              precision    recall  f1-score   support

       False       0.99      1.00      0.99    487729
        True       0.68      0.44      0.54     10822

    accuracy                           0.98    498551
   macro avg       0.83      0.72      0.76    498551
weighted avg       0.98      0.98      0.98    498551
```

Clustering Analysis

To exogenously define the number of clusters required for K-Means, the following elbow diagram was analyzed to determine that the addition of clusters meaningfully contributed to the fit between the Sale (Dollars) and Volume (Gallons) attributes up to a total of four:

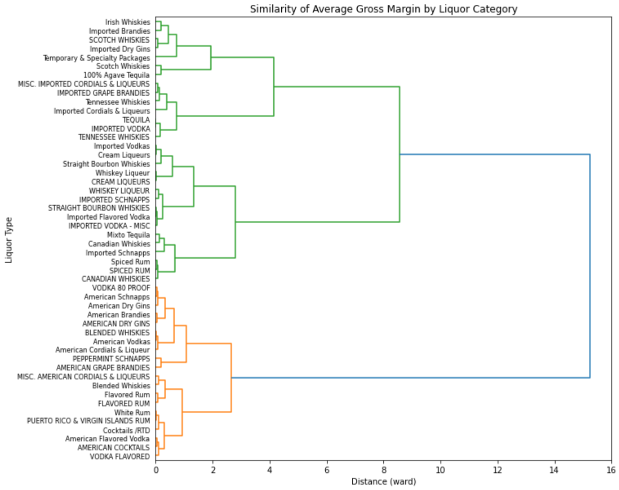


Using this cluster count for K-Means over the first million rows, groupings were developed most cleanly around basket size, with the size of each increasing in relative breadth as both attributes increased. The spatial relationship, however, may mask the extreme concentration and potential significance of the smallest clusters with respect to the largest.



An additional analysis of gross margin based on hierarchical clustering over Ward's method indicated substantial discrimintation in product markup. For legibility, liquor categories were restricted to the fifty most common, and margin aggregated to the average by category to summarize intra-group variation.

As visualized in the subsequent dendrogram, imported liquors (most of the green bracket) and domestic liquors (most of the orange bracket) had the most dissimilar margins at higher levels of agglomeration, each with their own subset of regional variation. At lower levels within each of these groupings, natural and artificial/blended liquors also showed a degree of discretization.



APPLICATIONS

The insights gained from our analysis have many applications, particularly as it relates to business and public health. Liquor distributors could use the data to help match supply with demand. For example, 750 ml is by far the most bottle size purchased, so it would be prudent to keep this size in stock. Additionally, December has the highest liquor sales of any month, so liquor stores could anticipate this uptick in demand by being well stocked with liquor going into the month.

Cluster analysis revealed interesting patterns that could be exploited commercially. Referring to the K-Means chart in the previous section, each cluster could be used to represent a potential customer profile. There is then demonstrated demand for dedicated account representatives at high volumes (purple), live shared support for those that do not quite meet this level of concierge (green), an asynchronous invoice inbox for smaller queries (blue), and a self-directed system for the most retail-oriented transactions (red).

Building on the dendrogram, liquor sales representatives could similarly be given a tool for recommendations based on the tier hierarchy to increase a revenue base. Depending on the customer situation, it may be more appropriate to suggest similar-margin products to promote volume in a transaction, or complementary ones priced at a premium given openness to exploring new tastes.

The data insights could also be used by both public and private groups to combat alcohol abuse. Resources would be best directed towards the counties with the highest consumption of liquor. As stated above, 7 counties in the state of Iowa account for approximately 90% of total liquor sales, so resources should be focused on those locations. Ad campaigns and resources, such as alcohol abuse rehabilitation centers, would be most effective in these high consumption counties.

REFERENCES

[1] *Analysis: Liquor store sales data reveals the biggest drinking days of the Year: Womply*. Womply helps small businesses thrive in a digital world. (2020, October 22). Retrieved October 25,2021,from https://www.womply.com/blog/analysis-liquor-store-sales-data-reveals-the-biggest-drinking-days-of-the-year/.

[2] Admin. (n.d.). *Alcoholic beverage market overview in the United States*. Park Street Imports. Retrieved October 25, 2021, from https://www.parkstreet.com/alcoholic-beverage-market-overview/.

[3] Iowa Department of Commerce, A. B. D. (2021, October 1). Iowa liquor sales by year and County. Retrieved October 25, 2021, from https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales-by-Year-and-County/ahiv-u4uz.

[4] Schroeder, S. (2020, December 18). *Increased alcohol sales come with costs for Iowans*. Des Moines Register. Retrieved October 25,2021,from https://www.desmoinesregister.com/story/opinion/columnists/iowa-view/2020/12/18/alcohol-sales-increase-costs-for-iowans/3943579001/.

[5] Merlo, L. J., Ahmedani, B. K., Barondess, D. A., Bohnert, K. M., & Gold, M. S. (2011, July 1). *Alcohol consumption associated with collegiate American football pre-game festivities*. Drug and alcohol dependence. Retrieved October 25, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101303/.

[6] Dangeti, P. (n.d.). *Statistics for Machine Learning*. O'Reilly Online Learning. Retrieved November 30, 2021, from https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml.