Mining Iowa Liquor Sales

CSPB 4502 Semester Project

Part 3: Project Progress Report

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

Updated Proposal

Problem Statement

We will examine alcohol sales throughout the state of Iowa from January 1, 2012 to present. We intend to examine long-term trends, such as preferences for certain brands 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).

We seek to identify geographic differences in preferences in Iowa, such as what liquor varieties and brands are particularly popular in college towns throughout the state.

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.

Literature Survey

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 about 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.

<u>Weekdays/Months with largest liquor sales</u>

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 seem to only be true for more winter based holidays rather than for summertime holidays. [1]

<u>Alcohol purchases vs other beverage purchases</u>

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]

<u>Alcoholic Sales in Iowa Counties</u>

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 about 6 or 7 counties that make significant alcoholic sales in the state of iowa. This is not surprising, as Iowa main population resides in the more urban towns, where more rural counties have smaller populations, thus leading to less sales [3]

Prefered 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]

Proposed Work

Collection

Initially, we proposed to locally download the data set for the preprocessing phase, and then re-upload to a cloud service such as Google Drive to work on individually. However, through more research on viable tools, we decided to store and code our data on the Kaggle platform. Kaggle allows us to see each other's progress and keep everything in one, centralized location.

Preprocessing

The dataset in its raw, untouched state does 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 does not need anything more to ask and answer interesting data mining questions. The only external source that could supplement this data would be a table matching zip code or city with a state if we wanted to conduct any geographical analysis by state.

Additionally, data transformation may also be an unnecessary exercise. There are no apparent redundancies or obvious errors in the data, at least visually. A more thorough outlier analysis and data distribution analysis will confirm this judgement.

The two preprocessing steps that will be performed on this data set are data reduction and data cleaning. Data reduction will focus on eliminating two attributes: "Store Location" and "Invoice/Item Number". "Store Location", containing lat/long coordinates, can be 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 lat/long coordinate for this mining project and with our objectives.

The other attribute that can be ignored and eliminated is "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 can be identified and inconsistencies resolved. 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. Resolving inconsistencies includes unifying the syntax on the stringed attributes. An example is the varying capitalization like "POLK" and "Polk" in the "County" attribute.

Data Mining

There will be 3-4 independent data mining techniques and studies applied to the data set:

1. Attribute association/correlation
2. Trend analysis
3. Clustering
4. Classification
5. Outlier analysis

Depending on difficulty, each technique has a different milestone in our proposed schedule. Refer to the *Milestone* section for more details.

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).

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

Evaluation Methods

Market Basket Analysis

Given the transactional nature of the data, association rules are of consequence. By using Apriori logic to prune candidate sets, exploratory analysis can be prioritized by interestingness. Calculations for support, confidence, and lift will provide an overview of our options, but we anticipate the need to identify a null-invariant metric to judge pattern recognition given relatively sparse binary vectors.

Bayesian Classification

With the sets of input conditions reduced to interesting, we can then compute the probability of specific outputs for attributes such as liquor type or volume using a Bayesian approach and compare against the frequency of actual occurrences. Statistical analysis would then assess the ability of the mined relationship to significantly predict the population.

There is substantial conditional independence between attributes in the data. 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 a naïve Bayesian approach fail and a Bayesian network should be used.

Information Visualization

The human optical system is valuable in pattern recognition because it processes information in parallel [6]. While some overhead is required to transform data into forms people can preconsciously interpret, visual depictions implicitly summarize dense results as efficient estimates of reasonableness.

Tools

A series of common Python libraries will be used to both conduct the Data Mining and apply the Evaluation Methods. With built-in optimizations to support more complex calculations and no restrictions to query-based constraints, there are advantages to using these over comprehensive relational models.

NumPy

For matrix operations; also required as a base for subsequent libraries.

pandas

Using DataFrames as the core idiom for the project simplifies filtering and aggregation.

SciPy

Contains many functions useful for summarizing the Data Set and for conducting descriptive statistical inquiries.

scikit-learn

A core package with methods for many data mining methods, including classification and clustering.

seaborn

Plotting library that standardizes design decisions to emphasize interpretation over visualization.

apyori

A simple implementation to apply the Apriori algorithm to data given minimum support and confidence.

PyStan

Allows Python-based access to Stan, an open-source program that supports Bayesian networks.

Milestones - Original

The schedule below shows the planned progress of the project with what is currently known about the data set and the data mining requirements.

10/25
- data set downloaded, ready for preprocessing
- proposal due
- assign mining tasks to members

11/1
- data set preprocessed
- data mining technique #1 complete

11/8
- data mining technique #2 complete
- begin result evaluation

11/15
- data mining technique #3 complete
- draft progress report

11/22
- data mining technique #4 complete
- progress report due

11/29
- data mining complete

- result evaluation complete
- work on report and presentation

12/6
- final report and final presentation complete

12/10
- final report, presentation, interview due

Completed Milestones

- Downloaded dataset locally
- Uploaded dataset to Kaggle
    - import attributes into pandas dataframe
    - limit # of rows for efficiency
- Clustering analysis
- Trend/correlation analysis
- Outlier detection/analysis

Outstanding Milestones

- Apply methods independent of outliers
- Regression analysis
- Further classification analysis
- Final project deliverables
    - Presentation
    - Report

Results

Preprocessing

As anticipated, the 'Store Location' attribute contains a significant percentage of missing (empty) values. Of the million sample size rows, nearly 10% of values were missing. 'Store Location' is not an attribute to be used unless cautiously.

Outlier Detection/Analysis

Using Grubb's test, four of the five 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 have been implemented so far and are 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.

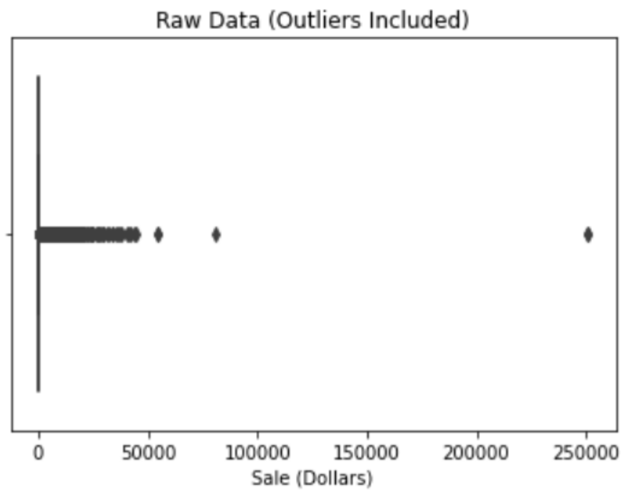
```
count    1000000.000000
mean         160.397008
std          590.427288
min            1.340000
25%           41.580000
50%           86.280000
75%          163.080000
max       250932.000000
Name: Sale (Dollars), dtype: float64
```

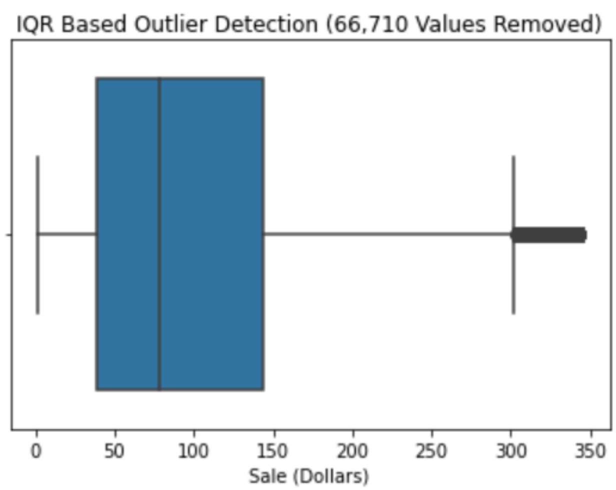
The three outlier methods are Z-Score based, interquartile range (IQR) based, and Winsorization based. The table below shows the number of outliers detected using each of the methods using a sample size of 1 million:

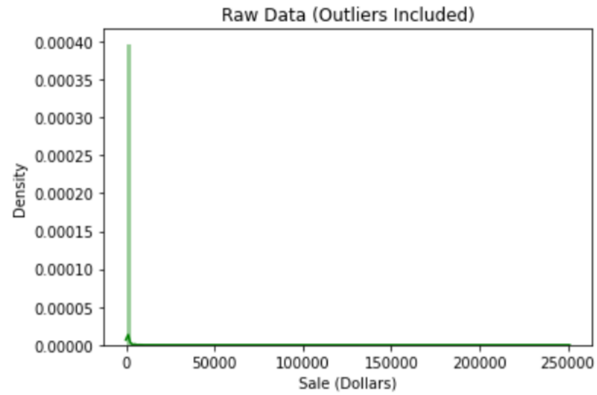
Method	# of Outliers Detected
Z-Score	6,129
Winsorization	19,480

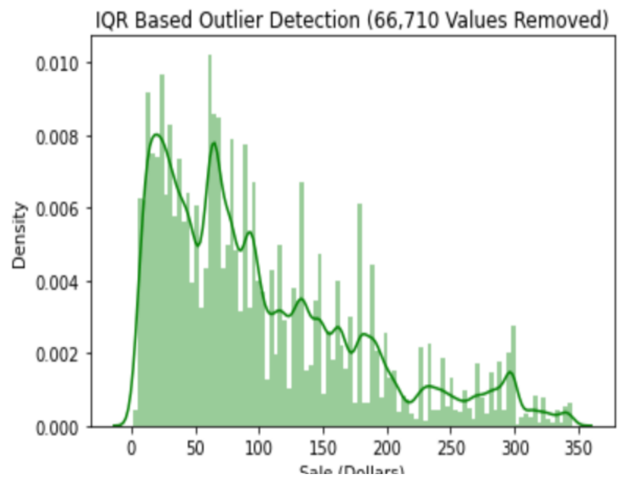
IQR	66,710

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).







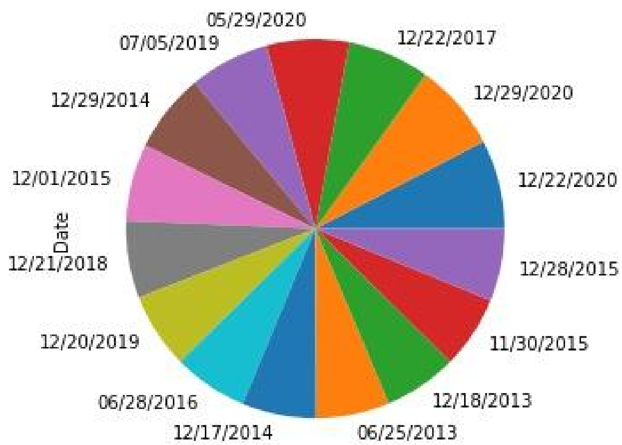


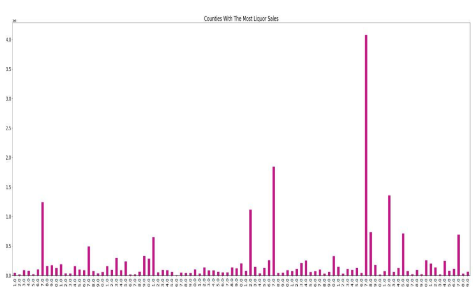
<u>Trend Correlation Analysis</u>

Through analyzing different attributes in the dataset with one another, we were able to create charts that start to show the relationship between different attributes. While memory issues have not allowed us to access all attributes at the same time, we were able to create visual representations particularly related to the total sales, dates of purchase, and counties that show the most activities. Matplotlib has a savefig function that allows for easy exportation of visuals created within python.

Through running a count function on all of the dates in which alcoholic sales were made, we were able to visualize a pie chart showing the 15 most common dates in which liquor was purchased. As you can see, December dominates that category. Surprisingly, there are no dates that are super close to July 4th, as you might expect there would be.

Additionally, a simple bar graph shows the counties that received the most amount of sales. As the visual shows, there are really only 6 or 7 counties that make up >95% of the total sales. Moving forward, it would be interesting to maybe eliminate the smaller sub counties and compare the big ones with one another.





<u>Classification Analysis</u>

We examined a few 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 classify whether a purchase was made in a zip code with a university/college 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. The inputs for the classification tools were the city the sale took place in and the retail price per bottle.

As can be seen in the analysis below, the naïve Bayes method was unsuccessful in its analysis. The algorithm predicted that no tuple belonged to the "College" class. We hypothesized that this was due to the independence 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 of features. Again, this may not be true; the data may be heavily skewed. These factors combined with the fact that the vast majority of the dataset are labeled as not belonging to the "College" class makes a naïve Bayes classifier a poor choice for our dataset.

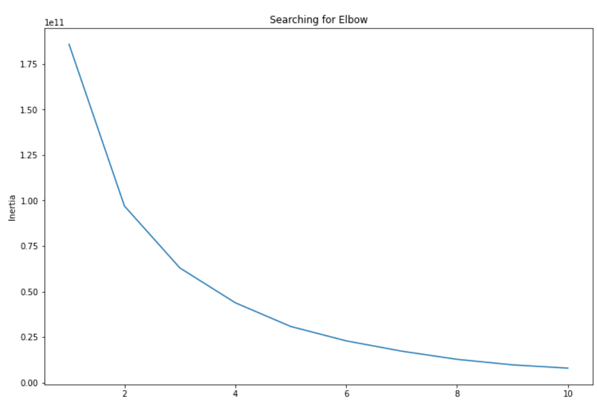
```
[[1633    0]
 [ 367    0]]
0.8165
              precision    recall  f1-score   support

           0       0.82      1.00      0.90      1633
           1       0.00      0.00      0.00       367

    accuracy                           0.82      2000
   macro avg       0.41      0.50      0.45      2000
weighted avg       0.67      0.82      0.73      2000
```

Given the issues laid out above with the independence and distribution of the data, we hypothesized that the k-nearest neighbor algorithm would be better suited to our data. The k-nearest neighbor approach 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 much more accurate, as can be seen below.

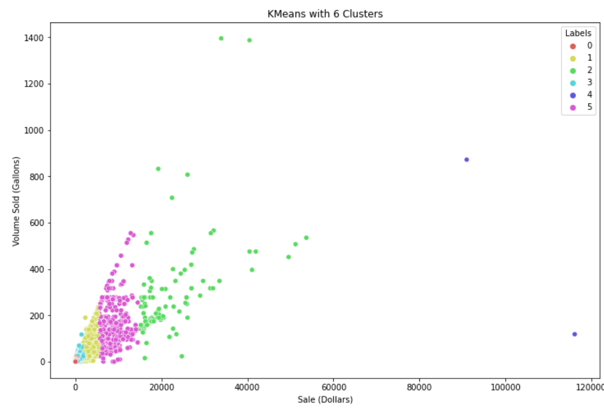
```
[[1226   78]
 [ 127  169]]
0.871875
              precision    recall  f1-score   support

          No       0.91      0.94      0.92      1304
         Yes       0.68      0.57      0.62       296

    accuracy                           0.87      1600
   macro avg       0.80      0.76      0.77      1600
weighted avg       0.87      0.87      0.87      1600
```

Clustering Analysis

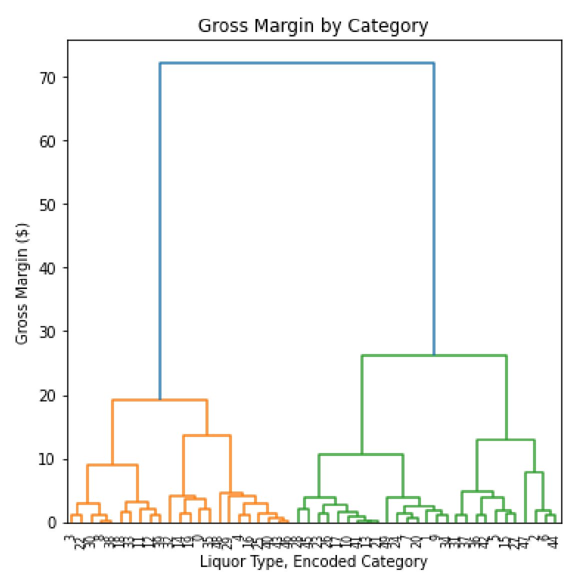
By treating the data as if it was unlabeled for clustering, intrinsic structure can be analyzed beyond the explicit category-based rules of the supervised methods. To apply K-Means on the Sale (Dollars) and Volume (Gallons) attributes, it was necessary to identify an ideal number of clusters, an approach simplified by using the “Elbow method” heuristic [7] which indicated an optimal count of six:



Using this for K-Means on the Sale and Volume over the first million rows, six clearly defined clusters were developed, increasing in relative breadth as both attributes increased. One cluster was populated with only two distant data points, indicating the influence of outliers; further analysis should be conducted on the full data set reduced to central tendency by the methods described above.



In a further trial involving a hierarchical clustering method, a dendrogram was created for the intersection of Gross Margin (the difference between retail cost and price) and Liquor Type (encoded as an arbitrary numeric category) to investigate potential markups. For legibility in proof-of-concept, the plot was restricted to 50 arbitrary rows of data. Future iterations should focus on the most frequent liquor types, and restore the categorical string label from the encoded version for clarity in interpretation.



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 October25,2021,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 October25,2021,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] Munzner, T. (2015). 1.5: Why Depend on Vision? In *Visualization Analysis & Design* (pp. 6–7). CRC Press, Taylor & Francis Group.

[7] 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.