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 from 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 sales of liquor 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?
    a. Based on the above three questions, can we tailor advertisements to increase sales of a certain liquor?
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. What have the average liquor prices in cities been during holidays?
    a. With this information we can predict what the highest price a customer is willing to pay before sales start to drop off
7. Are there any category names of 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 have reduced the number of rows by randomly selecting a sample from the years 2012 to 2015.

To further reduce our data we will be removing the following attributes that we consider unnecessary to answer the questions we proposed in our problem statement:

- Store Number
- Store Name
- Vendor Number

[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.*

- Vendor Name
- Item Number
- State Bottle Cost
- State Bottle Retail
- Volume Sold (Gallons)

3.2 Data Cleaning: We plan to delete 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)

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

3.4 Data Transformation: The category name attribute has many duplicate names with very specific descriptions of alcohol types that we plan on broadening. This attribute will be recategorized using excel.

3.5 Data Processing: Based on the problem statement questions we intend to answer, we will choose attributes from the dataset to find frequent patterns and correlations. The problem statements are in the same vein as Liang's "hotspots" in that we are setting out to find what drives liquor sales by location and time of year. Unlike Liang's machine learning multiple linear regression modeling, we intend to use frequent pattern analysis to find correlations [4]. For pattern evaluation we will decide between using Apriori and frequent pattern growth based on the efficiency and speed of the algorithm on our dataset. We will also decide between using Bayesian Classification and Decision Tree Classification to build a training model that can find correlations between the dataset attributes. Furthermore, we will evaluate the standard deviation, chi-squared, 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 have 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 will also remove several attribute columns that we feel are unnecessary for the purposes of this project. Of the original 24 attributes, we plan to keep 13 for use. These remaining attributes will provide us information pertaining to the sell date, type of spirit, price, volume, and store location. This sub-dataset will be downloaded onto each team members' local machine.

5 EVALUATION METHODS

During and after the processing of data, we will evaluate the results to see if the data is in fact able to provide answers to any of our questions. To evaluate the results, we plan on using techniques including, but not limited to, the following:

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

5.2 Association is useful for examining customer behavior. By identifying interesting relations between variables in our dataset, we can use the results to help answer questions related to shopping basket data analysis and product clustering.

5.3 Clustering analysis can be used to create consumer profiling by discovering groups and clusters in our data such that the degree of association between objects is highest if they belong together, and lowest otherwise. We can then use these results to understand questions pertaining to purchasing behaviors.

5.4 Regression can be used for prediction and forecasting of sales by identifying and analyzing the relationship among variables. Though none of our questions are concerned with prediction or forecasting, we are interested in what these results might show us.

5.5 Sequential patterns can reveal what other items customers are more likely to buy after an initial purchase. In theory, understanding sequential patterns could help liquor stores recommend additional items to customers to spur sales. For our purposes, we may see patterns within sales of specific liquors.

5.6 Visualizations grant us the ability to visually highlight patterns in data, instead of simply using numerical outputs of statistical models.

All of the above techniques have the potential to provide insight that could be used to answer our questions. We can then further assess the quality of any correlations, patterns and anomalies we find by potentially using the following evaluation methods:

5.7 Cross-validation is a pattern evaluation method that we can use to assess the accuracy of a model. If we use regression analysis or similar techniques to try to answer any questions, then splitting our data into training and test sets will give us the ability to evaluate the accuracy of such techniques. Bootstrapping might be similarly useful for establishing accuracy.

5.8 Clustering accuracy measures how accurately the patterns discovered by our algorithm can be used to cluster new data. We would likely measure our clustering algorithm by stability.

6 TOOLS

6.1 Project Management: As a group, we are using Zoom and Discord for project planning and communication. GitHub is 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 are utilizing a combination of Excel, Google Sheets, and Python within JupyterLab.

6.3 Data Visualization: For data visualization we are planning to use Tableau 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 minute adjustments.

7.1 Outline

Week 3/6: Data Processing & Data Cleaning

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

Week 3/13: Data Cleaning

- Data Reduction: Make tables based on attributes in excel

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

- Data Transformation with Excel

Week 4/3: Data Processing

- Data processing discovery

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.
- Discover correlations and patterns with selected algorithm

Week 4/17: Data Processing

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

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

- Compile completed milestones
- 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**

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/