Quantitative Data Analysis Report:
Observing Liquor Brand Sales and Rating Trends Across Iowa Counties

Sofie Coopersmith & Sofia Horenstein

I. Introduction

In this data analysis, we chose to work with two csv format files from kaggle.com and data.world. The first dataset is of liquor sales data including city, ZIP code, county, brand, bottles sold, and sale amounts in dollars across Iowa counties across the years 2012-2017. The second dataset is reviews of different alcohol including brand, name, date of review, review, text and review, url.

In our analysis we tried to answer questions such as :

- What is the average rating for each liquor brand?
- What is the total sales amount in dollars for each liquor brand?
- How many purchases were made for each liquor brand?
- What is the total sale amount for each brand by county?
- How many purchases were made for each brand by county?
- Which counties had the most purchases?

II. Methods

Before performing any analysis, we thoroughly cleaned both datasets with Python by dropping null values, removing columns, renaming values, and filtering for specific years. The Iowa liquor sales dataset needed to be reduced considerably from about 12 million values to close to 3 million once several years were removed. The final Iowa liquor sales data set has values only for the years 2015 through 2017. The Iowa liquor sales dataset also had repeat values for each county with different spelling e.g. ADAIR vs. Adair, so we had to replace all those values for only the latter version of the spelling for more accurate data analysis. Once the datasets were cleaned, we merged them by brand. In the merged dataset we removed the "$" from all the sale amounts and converted that column data type to a float for more streamlined analysis.

In order to connect our data frames to sql we needed to drop 2016 also. We first cleaned the data to drop columns we no longer needed. We then renamed the columns in order to create tables in a sql database. We then created a database in sql and ran a query to create 3 tables: iowa liquor sales, iowa liquor reviews and the merged table. We then created a connection in python in order to import the csv files. We were able to successfully import the reviews and checked by running a query in our jupyter notebook file.

III. Analysis

After replacing all the county name values with a uniform spelling and creating a new csv file from this data, we were finally ready to run an analysis on brand ratings and sales.

The first analysis we ran was to find the **average review ratings for each brand**. This was done by reducing the merged data frame to only the brand and reviews.rating columns, grouping it by brand, and finding the mean values.

We then ran an analysis on **total sales by brand** by reducing the merged data frame to only the brand and Sales (Dollars) columns, grouping it by brand, and finding the sum of the values.

We then ran an analysis on **total purchase count by brand** by reducing the merged data frame to only the brand and Sales (Dollars) columns, grouping it by brand, and finding the count of the values.

We then ran an analysis on **total purchase count by brand and county** by reducing the merged data frame to only the brand, Sales (Dollars), and County columns, grouping it by brand and County, and finding the count of the values.

We then ran an analysis on **total sales by brand and county** by reducing the merged data frame to only the brand, Sales (Dollars), and County columns, grouping it by brand and County, and finding the sum of the values.

We then ran an analysis of **highest and lowest amounts of liquor purchases by county** by reducing the merged data frame to only the Sales (Dollars) and County columns, grouping it by County, and finding the count of the values. We then sorted these values by Sale (Dollars) in descending order, allowing us to see which counties had the most purchases. When sorted in ascending order, we could see the counties that had the least purchases made.

Lastly, we ran an analysis of **highest and lowest liquor sales by county** by reducing the merged data frame to only the Sales (Dollars) and County columns, grouping it by County, and finding the sum of the values. We then sorted these values by Sale (Dollars) in descending order, allowing us to see which counties had the highest liquor sales. When sorted in ascending order, we could see the counties that had the lowest liquor sales amounts.

IV. Results

Fig1: Mean rating values across liquor brands in the data set.

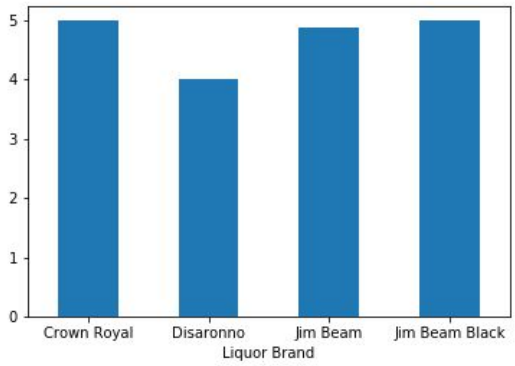


Fig2: Total sales values across all liquor brands in the dataset.

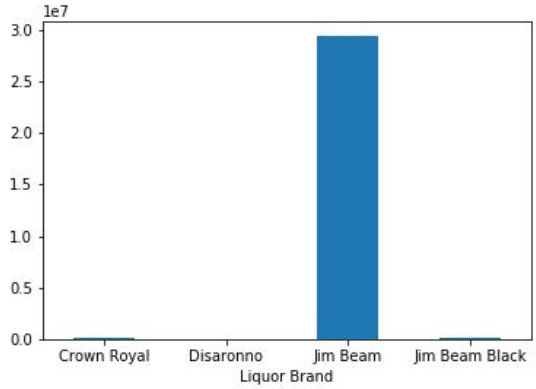


Fig 3: Total purchases made across all brands in the dataset.

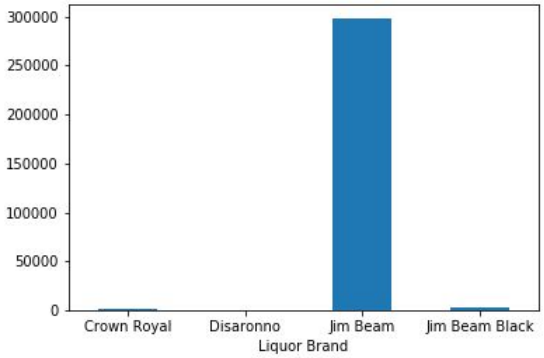


Fig 4: Top 5 Highest Liquor Sales in Iowa Counties

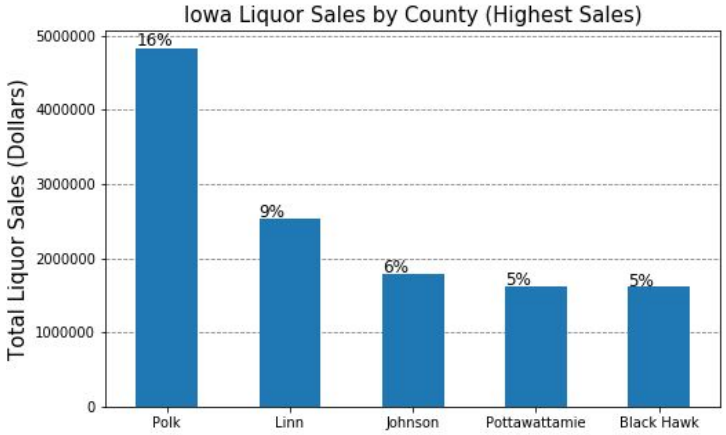
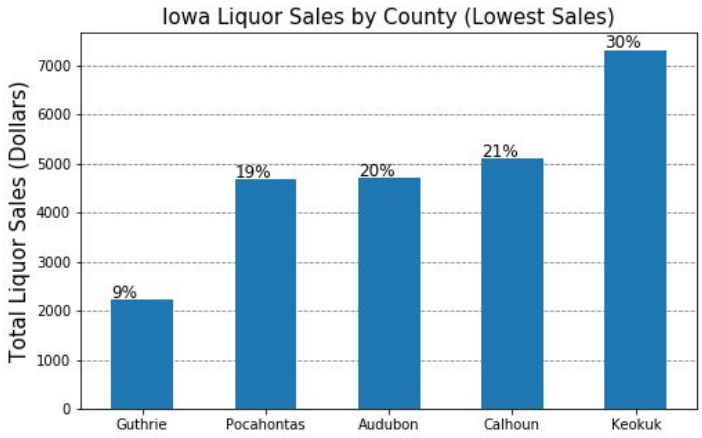


Fig 5: Lowest 5 Liquor Sales in Iowa Counties



V. Conclusion:

From our analysis, some key insights we gathered from the data were:

1. The liquor brand with the highest average review rating is tied between Crown Royal Whiskey and Jim Beam Black Whiskey.
2. The liquor brand with the highest sales is Jim Beam Whiskey with $29,440,760 in total sales.
3. The liquor brand that sold the most items was Jim Beam Whiskey with a count of 298,080 purchases made.
4. The Iowa county with the highest liquor sales is Polk county with $4,824,606 in total sales. Polk county also sold the most items with a count of 53220 purchases made.
5. The Iowa county with the lowest liquor sales is Guthrie county with $2236.41 in total sales. Fremont county sold the least items with a count of 135 purchases made.

Sources

Iowa Liquor Sales CSV Source:
https://www.kaggle.com/residentmario/iowa-liquor-sales#Iowa_Liquor_Sales.csv

Liquor Reviews CSV Source:
https://data.world/datafiniti/wine-beer-and-liquor-reviews