Liquor Sales in Iowa

John Montes
Mario Monterrosa
Danny Olivas
California State University, Los Angeles

Abstract: Iowans of all counties love their liquor. Due to laws in the state, alcohol vendors must log all liquor sales in order to maintain their license/ permit. The log must include the name of the bottle, bottle volume, how many bottles were sold, total sales amount, total volume of sale, and the alcohol category. We used this dataset to find that you are more likely to find Canadian whiskey in the typical Iowan home. Iowans' favorite bottle of alcohol is Black Velvet Whiskey and that not only does it top the charts in terms of dollar but also in volume of liquid sold. Iowans in Polk County always have their flasks filled and ready to go as they spend the most money on alcohol compared to other counties.

Cleaning, sorting, and analyzing the dataset of this size proved to be impossible using simple data tools like Excel. We tried opening the files on our personal devices and they would crash immediately. The processing power and the advantage of having 3 nodes working in parallel was immediately apparent.

1. Introduction

Alcohol has been used since ancient times for a variety of purposes. There are different forms of alcohol, but the one that is the most popular is ethanol. Ethanol's popularity stems from it producing psychoactive effects to the brain when consumed. People consume it for different reasons so there are many types to choose from. In the state of Iowa, all alcohol sales must be recorded by vendors in order to maintain their alcohol license. We uploaded the Iowan alcohol sales dataset into the Hadoop File system and used Hive to create the database. With Beeline, we defined our tables and transfered the data from the original dataset into our Hive tables. With the tables defined we used HiveQL to create queries and views that we exported into data visualization sofware Tableau and Microsoft Excel.

2. Hardware and Software Specifications

We are using an Amazon Web Services server to run Hadoop within a Linux instance. The server has 3 Hadoop clusters running the Intel Xeon E5-2670 v2 Ivy Bridge processor with 15 GB of memory and 80 GB (2 40GB SSDs) of storage space. The software includes Hive 2.3.5, Pig 0.17.0, Hue 4.4.0, Ganglia 3.7.2, Tez 0.9.2, Sqoop 1.4.7, HCatalog 2.3.5. The dataset used is 3.48 GB in size and was downloaded from Kaggle.

3. Iowa's Alcohol Sales

In order to maintain an alcohol license/permit in Iowa, alcohol vendors must report information to the state on every sale. The dataset contains the sales information going back to 2012. The information reported by vendors includes the following: Invoice/Item, Number, Date, Store Number, Store Name, Address, City, Zip Code, Store Location, County Number, County Name, Category Number, 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).

3.1 Top 10 Selling Bottles

To find the best selling bottles in the dataset we have to include the Item Description column, which gives us the name of the bottle, and the SUM(Sales in Dollars), which gives us the total sales generated by each bottle. Of the top 10 selling bottles, 5 of them are whiskeys, 3 are vodkas, 1 rum and 1 jägermeister. It is interesting that there is no tequila in the top ten.

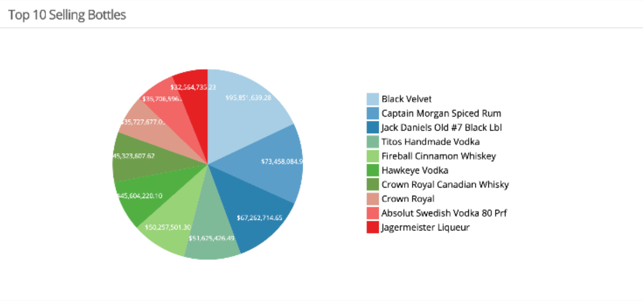

Figure 3.1 Pie chart of top selling liquor bottles

3.2 Best-Selling Liquor Category

To find the best selling liquor category in the dataset, we include the Category Name column and the SUM(Sales in Dollars), which gives us the total sales generated by each category. You can see that Canadian Whiskeys category is listed twice with a different spelling, if both categories were added together you would find that Canadian whiskey is the best-selling category of alcohol in Iowa and not American vodkas like the chart suggests. If you add the two Canadian whiskey bars you get $285,149,124.70 which nearly doubles the sales of the American vodka category. Tequila lands 7th in the list but not a single tequila bottle shows up in the top 10 selling bottles.

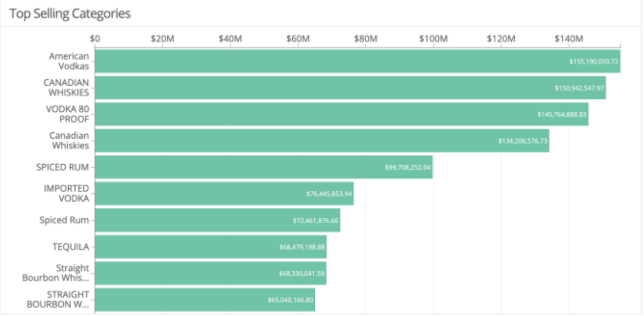

Figure 3.2 Bar chart of top selling liquor categories

3.3 County with the Most Amount of Alcohol Sold

The dataset contains the addresses of the stores that sold the alcohol so we used Microsoft Excel's 3D mapping feature to visualize which county had the most alcohol sales. With the information overlaid on a map you can start to make out where in the state the bigger cities lie. You can quickly find Des Moines, which is in Polk County as it has the largest amount of alcohol sales and that is no surprise as it is the most populated city in Iowa at 217,521 residents as of 2017. You see medium-sized spikes in the Linn County, which is home to Cedar Rapids, and Scott County, which is home to Davenport and Bettendorf. Polk, Linn, and Scott are counties with the most alcohol sales in that order.

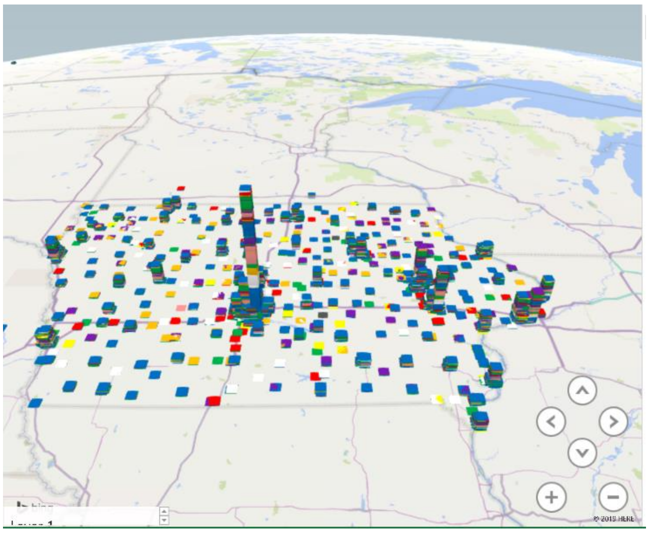

Figure 3.3 3D Map of Iowa illustrating sales per county (height) and category (color)

3.4 Top Sales in Dollars and Volume (Correlation)

We wanted to see if the best-selling bottle was also had the most liquid volume sold. We took the Item Description column, the SUM(sales in dollars), and the SUM(volume sold in gallons) to compare. We find that the Black Velvet places first in both tables so there is positive correlation between sales in dollars and sales in volume. Other bottles that make an appearance in both lists include: Captain Morgan, Jack Daniels, Tito's, Fireball, Hawkeye, Crown Royal, Absolute, and Bacardi. Jägermeister makes an appearance in the top selling in dollars table but not the volume table. Again, we see an absence of tequila despite its place in the top 10 selling categories.

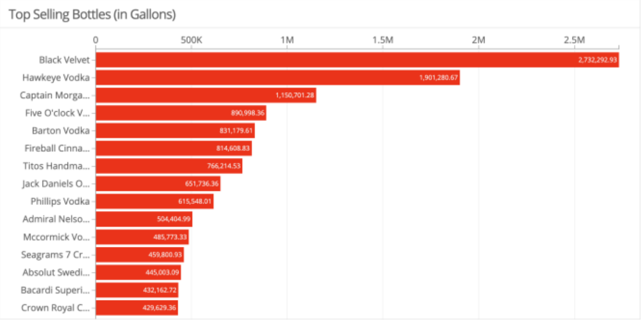

Figure 3.4.1 Bar chart of the top selling bottle in terms of gallons.

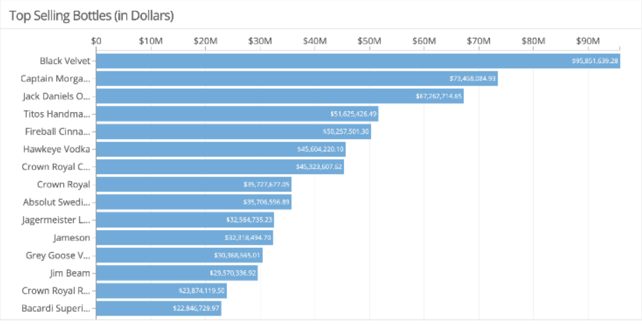

Figure 3.4.2 Bar chart of the top selling bottle in terms of dollars.

3.5 Bottle Unit Sales Over Time

The dataset contains a Date Sold column so we thought it would be interesting to see when people are buying the most alcohol. The dataset contains data starting from the beginning of 2012. We took the Dates column and the SUM(Bottles Sold) to look for seasonality. We see that from 2012 to 2015 there is always a spike in October, maybe people in Iowa really enjoy Halloween parties or need alcohol soon after the school year begins in August. In more recent years the spikes are happening in June and December, you have the beginning of summer and the end of the year, both of which are times of celebration for many people.

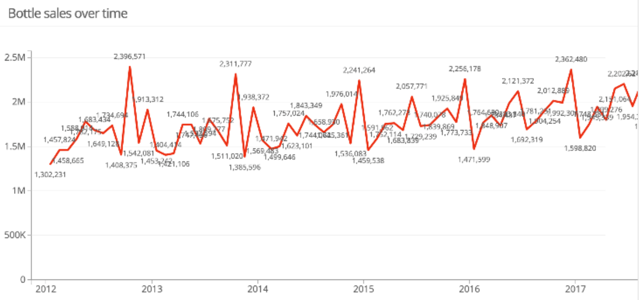

Figure 3.5 Timeline chart of bottle unit sales over time

4. Conclusion

We uploaded the dataset to Hadoop by connecting to the AWS server via ssh then using the scp command to upload the .csv file into HDFS. Once the file is in HDFS we connect to Hive to create a database for our group and then use Beeline to create tables using HiveQL commands. Once the tables are built, we can run queries using HiveQL to pull information from the database. The data is downloaded and used in Microsoft Excel to create the 3D Map to visualize the sales per county and Tableau to create other visualizations. We find interesting results, especially using the 3D mapping feature of Excel, the sales totals being to shape the state of Iowa and you can draw other information based on the chart, like city population. It was interesting to see that despite being in the top 10 selling liquor categories, Tequila does not have a representative in the top selling bottles, both by dollar and volume. It would be interesting to see the data for the state of California due to the diversity of not only people, but climate and land compared to Iowa.

References

[1] Iowa Liquor Sales
Aleksey Bilogur - https://www.kaggle.com/residentmario/iowa-liquor-sales#Iowa_Liquor_Sales.csv

[2] Group Five Github
John Montes, Mario Monterrosa, Danny Olivas – https://github.com/John2324/Group-5-Big-Data