Iowa Liquor Sales Analysis
Using HiveQL in Oracle Database Cloud Service (DBCS)

Authors: Katherine Belknap, Wethanie Law, Hsiu-Ping Lin, Shreyas Belur Manjunatha Swamy
Department of Information Systems, California State University Los Angeles
CIS5200-03 System Analysis and Design
kbelkna2@calstatela.edu, wlaw4@calstatela.edu, hlin54@calstatela.edu, sbelurm@calstatela.edu

Abstract: Tracking liquor sales is an essential part of regulating the sale and consumption of alcohol in Iowa. The Iowa Department of Commerce, Alcoholic Beverages Division (ABD) regulates the sale and distribution of alcoholic beverages in the state. They ensure that the state's alcohol industry operates safely and responsibly. Collecting proper tax amounts and preventing underage drinking and other alcohol-related problems lay within their purview. In this study, we delve into the insights gained from analyzing purchase information of Iowa Class 'E' liquor licensees. Although previous studies have explored this topic, they particularly emphasized best-performing sales locations and preferred categories. We aspire to present a novel and distinct perspective, which includes incorporating comprehensive sales data in terms of both monetary value and volume sold in liters, scrutinizing 11-year trends both in a broad context and for the top five volume-sold counties, providing store specifics for the first-place county, and presenting a categorization of overall sales by day. The dataset was analyzed using HiveQL to extract segments of interest. Furthermore, Excel and Tableau software applications were employed to produce visual representations that effectively demonstrate the discerned trends within the data.

1. Introduction

Iowa's liquor industry significantly contributes to the state's economy, generating millions of dollars in sales annually. Analyzing liquor sales trends in Iowa is vital for policymakers, regulators, and industry stakeholders to make well-informed decisions. This paper presents a comprehensive analysis of liquor sales in Iowa, focusing on trends and patterns observed over the past eleven years. The dataset utilized in this analysis comprises over fifty-one million individual sales transactions recorded by the Iowa Alcoholic Beverages Division (ABD). These records span from January 1, 2012, to March 31, 2023. It encompasses twenty-four columns of information, including Invoice/Item Number, Date, Store Number, Store Name, Address, City, Zip Code, Store Location, County Number, County Name, and specific details of the liquor sold. The size of the data file is approximately 6.35 GB. Through our analysis, we explored the evolution of sales trends over the years and examined the relationship between liquor sales and population changes. In addition, we identified the top-selling counties in Iowa and determined the most popular beverages during holidays. This comprehensive examination of liquor sales in Iowa, utilizing a robust dataset from the ABD, provides precise insights into the dynamics of the liquor industry. These findings enable informed decision-making and foster discussions on liquor-related policies and market strategies, benefiting policymakers, industry professionals, and businesses involved in the liquor industry.

2. Related Work

As we worked towards finishing our analytics project, we aimed to find diverse sources that could make us think in a new perspective and guide us on how to approach the data as well as what questions and queries could lead to unravelling insights. Additionally, we were cautious about avoiding plagiarism. Numerous publicly accessible works exist that document similar projects and analyses. Our project and analysis primarily center around a specific type of sales trend, explicitly regarding liquor sales in the state of Iowa. It was in our constant effort to approach the dataset in a unique way and to unearth and plot new insights from the data.

[1] In the analysis of Iowa Liquor Sales Data lead by Connor Toliver and team, reveals key insights such as, the top five cities for liquor sales are Des Moines, Cedar Rapids, Davenport, Waterloo, and Iowa City. December has the highest sales, followed by October, August, May, and June, indicating a seasonal trend. The top ten liquor brands include Black Velvet, Hawkeye Vodka, Five O'clock Vodka, and Captain Morgan Spiced Rum. Popular liquor types are Canadian Whiskies, Vodka 80 Proof, and Straight Bourbon Whiskies. Preference analysis shows whiskey is favored in Ames, while vodka is popular in Iowa City. This research adopted various Data cleaning procedures to address the missing values and inconsistencies.[2] In the second study by Evan Lutins explores the Ridge Regression Model to predict Total Sales for zip codes based on predictors such as the number of stores, volume sold, bottles sold, state profit, and sales per store. The model identified four promising zip codes (50314, 50320, 52807, and 50311) with high sales potential due to low competition and high sales indicators.[3] In the third reference material conducted by Richard Liang, he followed the data science pipeline to analyze liquor sales in midwestern Iowa. Through exploratory data analysis, he identified hotspots for liquor sales and examined the impact of location and time of year on sales. Hypothesis testing and machine learning revealed that the type of liquor and county location had some influence on spending but were inconclusive.

While there are shared objectives regarding sales trend analysis, our research stands out due to its distinct approach. We employed Hive on cloud computing to analyze the data, creating external tables and utilizing Hive queries to extract relevant information from the vast data set.

Additionally, our analysis incorporated tempo-spatial and geo-spatial representation by using Excel 3D maps to showcase the progression of liquor sales across Iowa's counties over time. The heat map categorizing sales by month and day in the past three years stood out amongst other analyses. The metrics and data points that we plotted in same chart representations differed from other studies, highlighting our unique implementation and approach to gaining insights.

3. Specifications

For this project we used the Hadoop Cluster on the Oracle Big Data Cloud Platform to extract and transform the Data, to create valuable insights.

Table 1. Hardware Specifications

Oracle BDCE	
Memory	390.7 GB
Cluster version	Hadoop 3.1.2
Hive version	Apache Hive version 3.1.2
Nodes	5 (3 Master, 2 worker)
CPU cores	8
CPU Speed	1995.312 GHz

Table 2. Data Set Specifications

Iowa Liquor Sales Dataset	
Data size	6.35GB
Total number of columns	24
Total number of rows	51,073,630

4. Working with the Data

The Iowa liquor sales dataset used in this analysis consists of over fifty-one million individual sales transactions recorded by the Iowa Alcoholic Beverages Division (ABD). These transactions represent the sales of liquor from the ABD to Iowa Class "E" liquor licensees, which include various commercial establishments such as grocery stores, liquor stores, and convenience stores. The dataset covers a time range from January 1, 2012, to March 31, 2023. The information available for each transaction are date, store, category, bottle size, and sales revenue, among other variables.

4.1 Project Workflow

This dataset was made available to us via Kaggle. We started by downloading the dataset from Kaggle, upload it Oracle DBCS, then put it Hadoop HDFS. We then used HiveQL to process and validate the data. Once the data was processed and validated, we continue to use HiveQL queries to analyze the data. The necessary tables were then extracted to our local computer. In order to create the 3D Map in Excel, we need to clean the latitude and longitude column for the PolkDetail.csv. Finally, we used Excel and Tableau to visualise the data and present our findings.

Figure 1 - Implementation Flowchart

4.2 Data Processing

During the data processing phase, we encountered challenges related to special characters present in the dataset, such as commas, backslashes, and other characters. To address this issue, we utilized the OpenCSVSerde function, which allowed us to parse the dataset accurately and ensure proper column alignment. By employing OpenCSVSerde, all columns were initially assigned the string data type by default. However, to overcome this limitation, we performed data type casting to assign the appropriate data types to each column, such as casting the "data" column to the "date" data type and numeric data to "bigint" or "decimal" types.

For geospatial analysis purposes, it was necessary to have separate columns for latitude and longitude information. The original dataset had the coordinates in one column labeled "store location", including both latitude and longitude. To separate the coordinates, we used the regexp_extract function to extract the latitude and longitude values and create two distinct columns. Consequently, the final table expanded from twenty-four to twenty-five columns, accommodating the additional latitude and longitude columns required for geospatial analysis.

The dataset initially contained information from the first quarter of 2023. However, due to the incomplete nature of this partial-year data, we decided to limit the transaction date range to span from 2012 to 2022, ensuring a more comprehensive analysis.

Additionally, for geospatial analysis and the creation of a 3D Map in Excel, it was essential to have consistent latitude and longitude information for each store number or store name. As we identified that the same store number or store name had different latitude and longitude values, we needed to resolve these inconsistencies. To do so, we replaced the inconsistent latitude and longitude values with consistent values in Excel for the PolkDetail.csv file. Then proceed with the creation of the 3D Map graph in Excel. This step was necessary to ensure accurate geospatial analysis and visualization.

Overall, these data processing steps helped us overcome challenges related to special characters, data type assignment, and geospatial analysis requirements, ensuring the dataset's accuracy and enhancing its usability for subsequent analyses and visualizations.

5. Analysis and Visualization

Looking at the difference between sales in terms of dollar amount and comparing that to the liters sold we decided to focus on liters for subsequent queries. There is a sizeable gap up post 2019 in this trend for the following speculative reasons: Increase in price per liter of alcohol due to inflation, shift in product mix and change in consumer preference to drink more premium alcohol, or changes in consumption patterns.

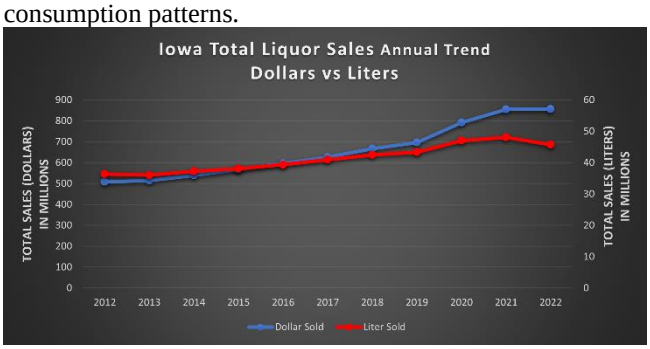

Figure 2: Iowa Total Liquor Sales Annual Trend Dollars versus Liters

The red trend line represents liquor sales annual percentage change, while the bar charts represent liquor sales in millions. The notable period in this representation is between 2019 and 2022, it can be speculated that quarantines and lockdowns resulted in a decline in the annual percentage change of liquor consumption. The data between 2021 and 2022 suggests that stricter lockdown measures and the impact of the second wave of the pandemic led to changes in consumer behavior, as liquor stores remained inaccessible.

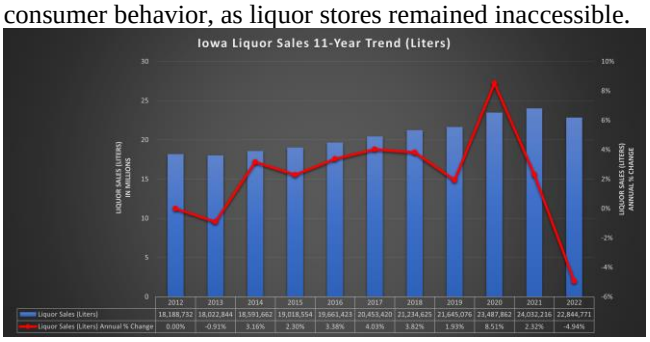

Figure 3: Iowa Liquor Sales 11-Year Trend in Liters

The red trend line represents population annual percentage change, while the bar chart represents liquor sales annual percentage change. The notable period in this representation is between 2019 and 2022. During this time, it appears that the increase in remote work led to a population influx in Iowa, resulting in a greater increase in liquor consumption. However, between 2020 and 2021, it can be speculated that pandemic-related factors caused a decrease in liquor consumption. The data between 2021 and 2022 suggests that the continued decline in population further contributed to a decrease in liquor consumption, pushing it into negative territory.

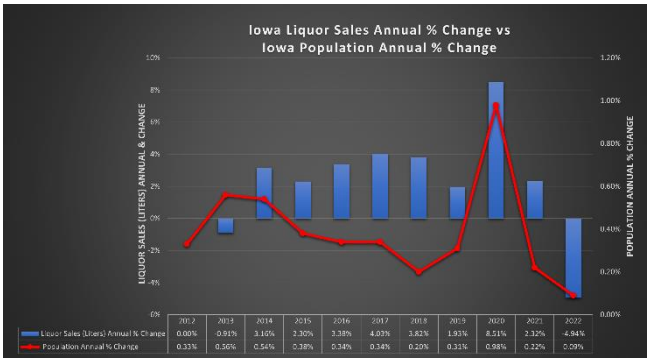

Figure 4: Iowa Liquor Sales Annual Percentage Change versus Iowa Population Annual Percentage Change

The most populous county in Iowa is Polk County, which is home to the state capital, Des Moines. According to the 2020 United States Census Polk County had a population of 498,823. Therefore, it is little wonder why sales are highest in this county, as evidenced by the following heatmap of Iowa.

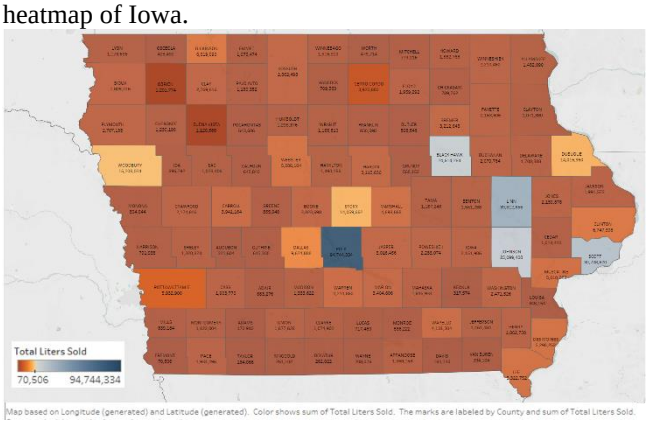

Figure 5: Total Liters Sold by County

We have compiled a list of the top five counties with the highest total liquor sales between 2012 and 2022. Upon examining the figure, it becomes evident that the trend of liquor sales for each year was predominantly increasing, albeit with a slight decrease observed in 2022. This outcome aligns consistently with our earlier findings illustrated in the figure.

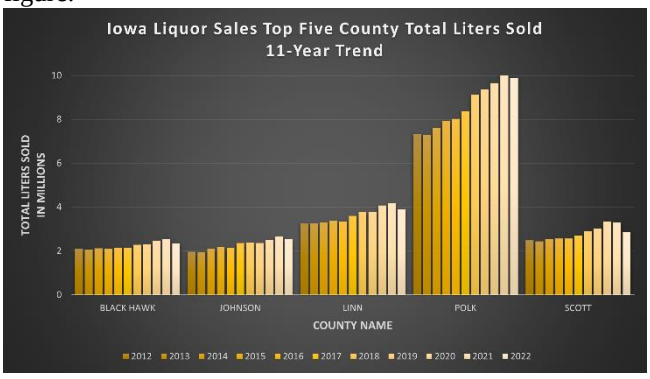

Figure 6: Iowa Liquor Sales Top County Total Liters Sold 11-Year Trend

Based on the preceding figure, it is evident that the county POLK exhibits the highest quantity of liquor sold. Furthermore, this figure surpasses the quantities recorded in the remaining four counties. Consequently, we have made the decision to place additional emphasis on analyzing the data pertaining to POLK. The subsequent figure illustrates the distribution of liquor stores within POLK, with the presence of a red color within a circle icon denoting a higher concentration of such establishments in a given area.

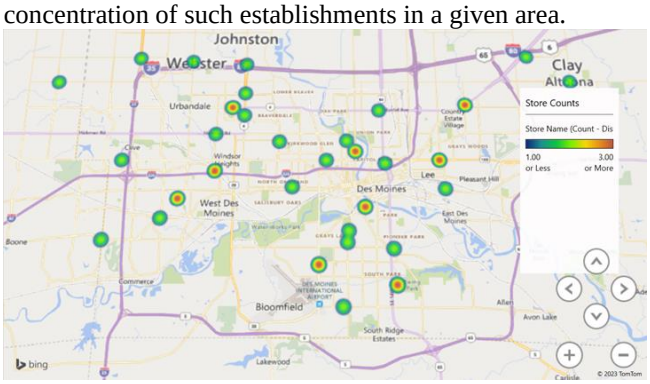

Figure 7: Polk County – Store Count

In continuation, the subsequent figure also pertains to the county of POLK, providing an overview of the quantity of liquor sold within its boundaries. By comparing this graph with the aforementioned figure, it becomes evident that areas with a greater number of liquor stores generally exhibit higher sales amounts. Additionally, our analysis reveals a correlation between higher sales amounts and proximity to freeways or main roads, as such locations tend to demonstrate greater sales volumes.

Figure 8: Polk County - Total Liters Sold by Store

[5] According to Alcohol.org, who polled over 1,000 Americans on drinking habits, the top drinking holidays are Mardi Gras, New Year's Eve, St. Patrick's Day, Fourth of July, Halloween, Cinco De Mayo, Memorial Day, Labor Day, Winter Holidays, and Thanksgiving. Given this data, we decided to create a heat map showing the sales by day and week. In general, however, we did not find much correlation save for perhaps winter holidays. The most consistent sales were between weeks 37 to 41 and week 49.

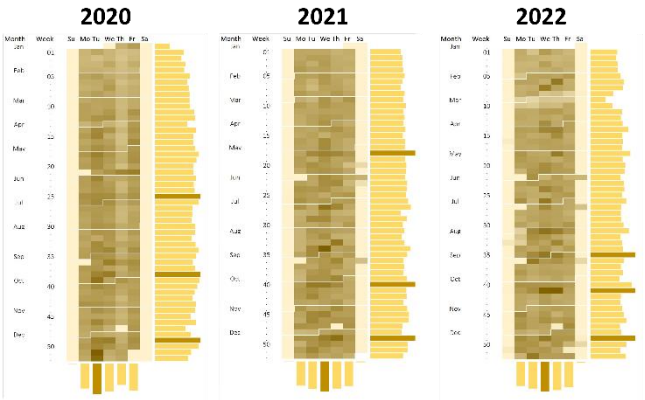

Figure 9: Heat Map – Total Liters Sold by Day / Week

6. Conclusion

In general, sales trends by liters, dollars were deemed too volatile, are consistent across counties with a mostly upward trend until the year 2020. Polk county consistently had the top sales, which made sense as the home of the state capital and the most populous county in Iowa. The maps also show a strong correlation between sales and main roads. We expected alcohol sales to corelate more with holidays, however, other than the 49th week we do not see much change overall.

References

[1] C. Toliver, "iowa-liquor-sales," 2019.
[2] E. Lutins, "Predicting-Iowa-Liquor-Sales," 2017.
[3] R. Liang, "Visualization and Analysis of Liquor Sales in Iowa," 2020.
[4] MacroTrends.net Editor, "Iowa Population 1900-2022," 2023.
[5] FindMeABrewery.com Editor, "A Boozy List of National & International Drinking Holidays in 2023," Jan 2023.

https://github.com/lovekangaroo/IowaLiquorSales[1]

[1] Github Link:
https://github.com/lovekangaroo/IowaLiquorSales