**Iowa Liquor Sales Analysis**

**Arshita Dixit** (CIN: 305832363)

**Abhishek Shah** (CIN: 305084941)

**Kaustubh Padhya** (CIN: 305861379)

Department of Information Systems, California State University

Los Angeles

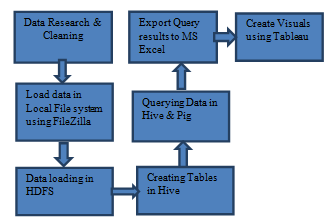
**Abstract**: This dataset consists of the complete liquor purchase information of Iowa Class “E” licensees. This data demonstrates the records by product and date of purchase from January 1, 2012 to 2015. The dataset can be used to analyze total spirit sales in Iowa for individual products at the store level. Using this dataset, liquor buying patterns can be predicted and can be categorized on the basis of maximum purchase of a particular brand on any day of the month/year. Plot trends can be identified with respect to overall liquor sales over the years. Moreover, it is intended to identify the top liquor consuming cities/areas in Iowa with this analysis. Also, a relational analysis between the retail sales and sales volume for particular categories of spirits can be generated.

**1. Introduction**

The goal of this project is to create a platform to visualize the complete liquor sales in the state Iowa to make a quick analysis on the amount and type of liquor sold every year. This analysis could be helpful for future investors in spirit business to come up with best suitable locations for liquor sales and to approach vendors perfect for better revenues.

2. Work Flow

To start with analysis, we first took data from data.iowa.gov in a .csv file and then went on with data cleaning. Then the data was uploaded into local file system using the tool FileZilla. Further, the data was loaded into Hadoop BigInsights cluster from IBM Bluemix. Next, we created external tables in Hive which helped us in the analysis of complete liquor sales throughout the state Iowa as we ran our queries using Hive and Pig. Next the query results were uploaded in Microsoft Excel. Finally, the graphs, maps and other visualizations were gathered using the tool Tableau. The analysis was made with respect to the results achieved.

****

**3. Background Work**

We have used the liquor sale dataset of the state of Iowa to carry out detailed analysis on liquor sales according to the categories, location, dates, brands and packsizes.

**4. Data Storage**

We intend to store our unstructured data in Hadoop Distributed File System (HDFS). HDFS provides reliable and scalable storage for storing and analyzing unstructured data. It runs on top of Linux operating system and this is designed to operate on commodity hardware. Hadoop stores data in blocks thereby providing strategic data processing and fault tolerance. Hadoop was designed to have a Master-Slave architecture in which slaves are data nodes and master is name node.

**5. Querying**

Data stored in HDFS can be queried using different programming languages, for our analysis we used Hive and Pig Query Language to extract the data as it provides an SQL-like structure and operates on tables just like an RDBMS, even though it does not strictly follow the full SQL standard. Hive queries are translated into Map reduce jobs which run on top of Hadoop cluster.

**6. Services**

We used IBM Bluemix cloud computing platform as it provides both PAAS (Platform as a service) and IAAS (Infrastructure as a service) in addition to supporting different programming languages. For using Hadoop on IBM Bluemix, Hadoop BigInsights must be created and deployed.

The cluster details are as follows:

* Number of Data Nodes – 1 node | vCPU = 4(24 GB RAM)
* Number of Management node-1 node | vCPU = 12 (48 GB RAM)
* Cluster Type – Hadoop IBM BigInsights
* Version – IOP 4.2 [IBM Open Version Platform]
* Operating System – CentOS 6.6 [Linux]
* Data Disk- 1 TB SATA | Data Storage- 244 GB
* CPU speed- 2.2 GHz

BigInsights provisions Hadoop clusters designed to store and analyze large data sets. For creating visualizations, we have used Tableau 10.0 software.

**7. Data Representation**

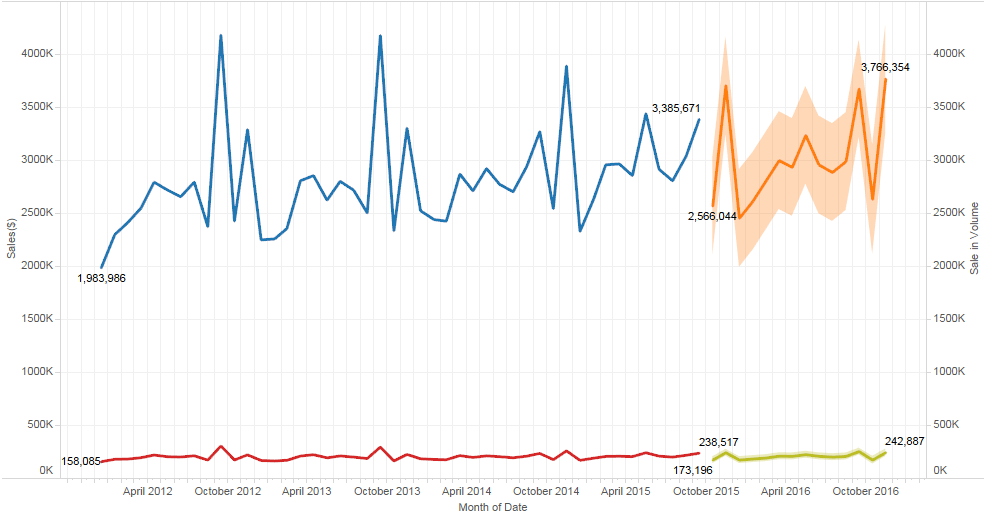
In this Project, we have considered Six main parameters to analyze the data.

1. Categories of liquor sold.
2. Cities with higher sales.
3. Sales on the basis of brand names in specific categories.
4. Yearly and Monthly sales analysis
5. Top Stores to be targetted for better revenues.
6. Suitable SKU sizes for Tequila.

**8. Analysis of the Dataset:**

Our Our data set consists of over 100 Million rows in comma separated values (csv) format, which provides information on types of liquor sold at any store in a particular city on any day of the year.The data was filtered and analyzed by writing suitable queries in Hive QL and Pig. Following Hive Queries were executed to obtain the above for mentioned results:

1. **Annual Liquor Sales in Iowa from 2012 to 2015**



This Time series graph shows the trend of present and future sales of liquor in Iowa. The highlighted graph represents the forecasted sale in dollars at top and volume in the bottom for next 2 years till 2017. Looking at it a clear seasonality occurs during the months of October- December, and so we suggest the companies to plan their Production and supplies accordingly to meet the market demand.

1. **Top five cities in Iowa according to Sales**



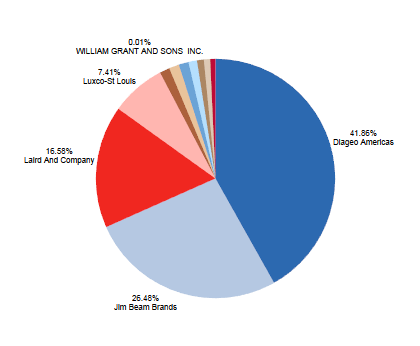
Here we are show casing the top five cities which have the highest sales in dollars and it comes out that Des Moines leads the race followed by Cedar Rapids, Davenport, Waterloo and West Des Moines. We suggest that these cities should be the first choice if you are planning to open a liquor store or want to sell your brand of liquor.

1. **Relationship between SKU size and Bottles sold for Tequila**



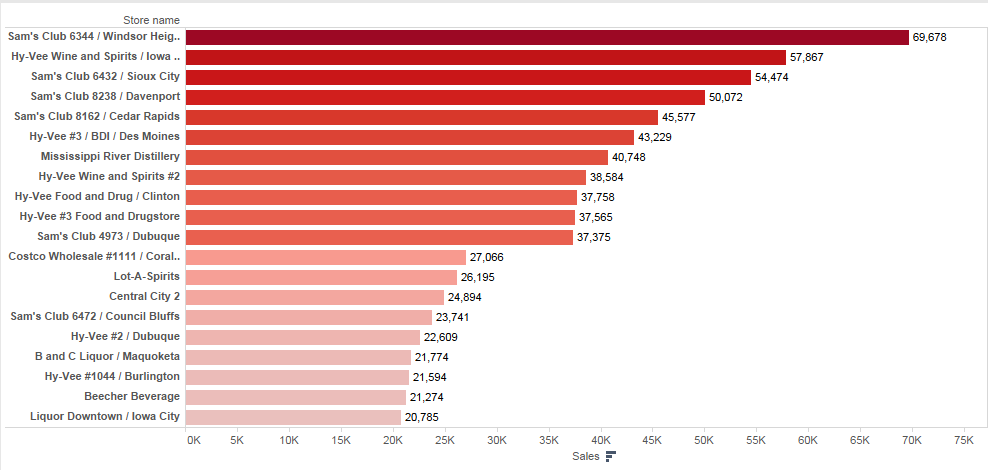
By the above Tree map, we are trying to draw a relation between different pack sizes and their sales. After this analysis we concluded that the tequila bottles for 750 ml are the most sold quantities followed by 1000 ml bottles. We also saw that 500ml, 300ml, 150 and 50ml are the least sold quantities for tequila, Hence; we suggest the manufacturers to produce the 750ml and 1000ml SKU sizes more in order to enhance sales and profitability and stop making the smaller SKU sizes for Tequila.

1. **Market share of companies selling Blended Whiskies**



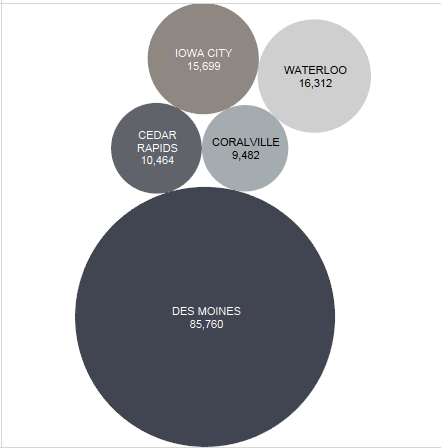
This visual represents the companies who have the highest sales for blended whiskies. From this, we can see that Diageo Americas has the highest market share which stands at 41.8% followed by Jim Beam brands having 26.4%. This shows that Diageo Americas is the favorite choice of blended whiskies of the people of Iowa.

1. **Top Dealers/Retailers of Blended Whiskies**



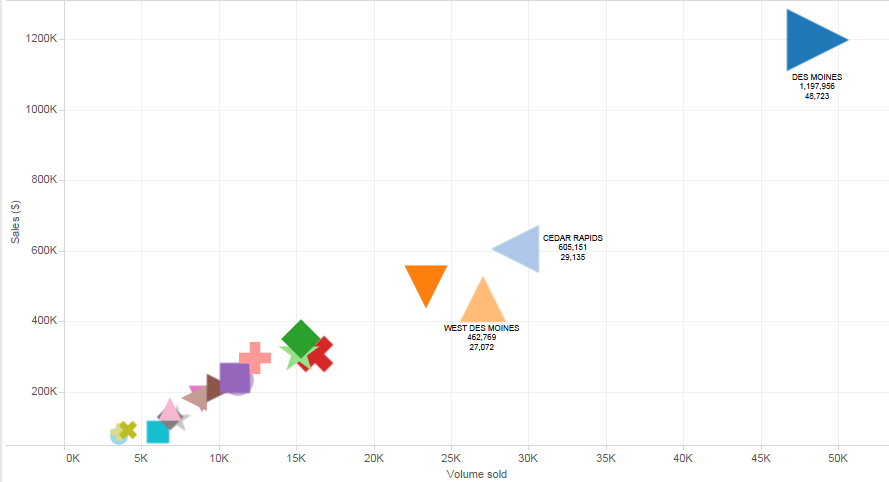
This visual shows the top dealer stores for blended whiskies and it appears that Sam’s Club from Windsor heights has the highest sales followed by Hy-Vee wine and spirits. We suggest that the company salespersons should target these stores at first for better business in the Blended whiskies category.

1. **Cities having Highest liquor sales on the Halloween**



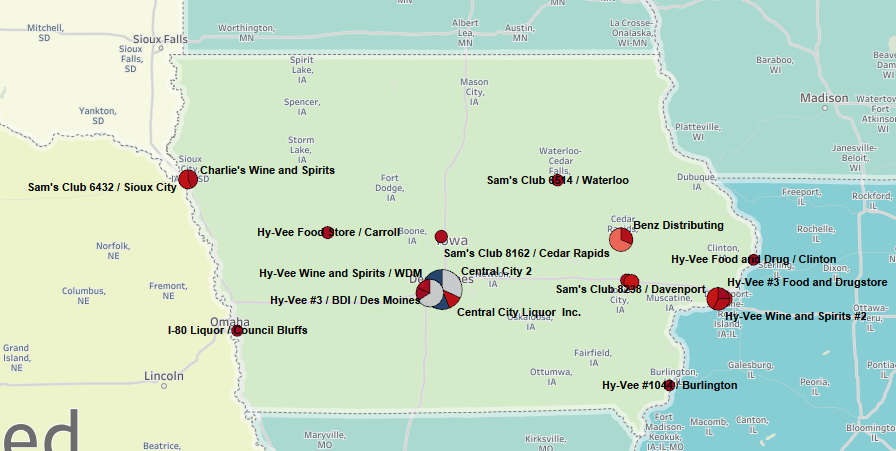
The cities with highest liquor sales on Halloween 2013 can be seen in this visual. We noticed the highest sales were recorded in Des Moines, Iowa city, Waterloo, Cedar Rapids and Coralville on the day of Halloween, which prompts us that the distributors and manufacturers should focus on their sales in these cities on the day of Halloween.

1. **Relation between Volume and Sales Turnover for Imported Vodka**



Here, a scatter plot for volume and sales turnover for imported vodka is represented, showing a linear relationship between Volume of Vodka sold and the Revenue in Dollars. Also, it shows the top 20 cities which can be targeted by liquor companies for selling Imported Vodka.

1. **Best Dealers/Retailers of Diageo Americas Blended Whiskies**



This Geospatial visual shows the top dealer stores for blended whiskies of Diageo Americas. Here we suggest that if Diageo Americas wants to award its dealers and run a loyalty program for them, it should start with these champion stores, starting with Hy-Vee Des Moines doing a business of 1.13 Billion for them in 4 years. By awarding these stores it can create a healthy relationship with them by boosting their morale to sell more and would in turn benefit the company.

**9. Conclusion**

In this project we have learnt the approach to leverage the Hadoop Distributed File System (HDFS) architecture. We have gained experience on querying the data stored in HDFS using the Hive Query Language. We also understand the mechanism used by Hive to convert queries to map reduce jobs and process the data. Also we understand the utility of using Hadoop in addition to its limitations. We also understand the difference between Hive and SQL.

In our detailed analysis we were able to project and visualize valuable data. The summary of the analysis is as follows:

* Highest liquor sales happens in the Months of October and December every year, and hence we suggest the companies to plan their supplies accordingly.
* Highest sales of liquor happens in Des Moines, Cedar Rapids, Waterloo, Davenport and West Des Moines.
* The 750ml and 1000ml SKU’s are sold the most for Tequila and the least sold are 50, 150 and 300ml bottles.
* There is a linear relationship between Volume sold and Sales in Dollars for the Imported Vodka. Also, it comes out that Des Moines is the city where imported vodka is sold the most.
* The Diageo Americas brand of Whisky has the highest market share (41.8%) for Blended whiskies. Also, we have shown the top sellers of Diageo Americas.

**Business Solution:**

On the basis of our analysis, we have found out that the highest liquor consuming city in Iowa is Des Moines across all the categories of liquors. Also, we have suggested the Manufacturers and distributors to plan their production and supplies of liquors across particular categories, SKU sizes and locations to serve the market demand and increase their profitability.

**Dataset URL**

<https://data.iowa.gov/Economy/Iowa-Liquor-Sales/m3tr-qhgy>

**GitHub URL**

<https://github.com/arshah137/CIS_5200>

**References**

* <https://data.iowa.gov/Economy/Iowa-Liquor-Sales/m3tr-qhgy>
* <https://filezilla-project.org/>
* <https://console.ng.bluemix.net/catalog/services/biginsights-for-apache-hadoop>
* [http://onlinehelp.tableau.com/current/pro/desktop/en-us/help.htm#maps\_geographicroles.html](http://onlinehelp.tableau.com/current/pro/desktop/en-us/help.htm)