560M Final Project

Group 72

Yvie Lin 491764, Huina Cao 500775, Lanlin Su 487930, Rainy Chen 500856, Tian Liu 502454

1. Executive Summary

To provide a region-specific recommendation and marketing strategy for a new liquor entrepreneur in order to maximize its profit, we used several data process tools such as Hive to process the dataset 'Iowa Liquor Sales structured data'.

First, we consulted with our client to make the objectives, then we set up 8 research problems accordingly with the dataset. To process these problems, we used Mapreduce, Hive, and tableau as methods. We further generated 8 results in terms of location, city, brands, volumes, profits, etc., which can be selected to make better decisions for the entrepreneur to maximize profit.

2. Introduction

a. Data Introduction:

We used Iowa Liquor Sales structured data. It contains purchase information of Iowa Class "E" liquor. Specifically, information on the name, kind, price, quantity, and location of sale, sales of individual containers, or packages of containers of alcoholic beverages from January 1, 2012 to current (from Kaggle dataset's description).

b. Data Source:

Original link: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy

We downloaded the following files Iowa_Liquor_Sales.csv from the cloud server, representing the sales activity data from Iowa state in the U.S.

c. Data Collection:

This dataset was collected by The Iowa Department of Commerce includes purchase information since 2014 and keeps updating.

d. Data Size: 3.47 GB.

e. Data Dimension and Dictionary: 12591077 rows and 25 columns.

Column Name	**Data Type**	**Description**
Invoice/Item Number	string	Invoice of each product
City	string	City of store
Zip Code	bigint	Zipcode of store
Store Name	string	Name of store
State Bottle Cost	double	The amount that Alcoholic Beverages Division paid for the bottle
Store Location	string	Location of store
Item Description	string	Item Description
Category Name	string	Category for the liquor type
Pack	bigint	The number of bottles in a case
County Number	double	Iowa county number where store is located\n
Store Number	bigint	Store Number
Volume Sold (Liters)	double	Volume of bottle sold in liters
Volume Sold (Gallons)	double	Volume of bottle sold in gallons
County	string	County where store is located

Vendor Number	double	The vendor number of the company for this brand of liquor
Date	string	Date of order
Address	string	Address of store
Category	double	Category code for liquor type
Vendor Name	string	The vendor name of the company for this brand of liquor
Bottle Volume (ml)	bigint	The metric size of a bottle
State Bottle Retail	double	The amount the store paid for the bottle
Item Number	bigint	Item Number
Sale (Dollar)	double	The price the store sell for the bottle
Profit	double	The total amount of the sale (number of bottles multiplied by the bottle price)

3. Problem Statement

We aim to provide data-driven region-specific recommendations and market strategy for liquor entrepreneurs in order to optimize the storage and distribution plan, elevate the

R&D production strategy and earn higher return. Therefore, we set up 8 research questions as follows:

1. What are the top 5 cities sold the most each year between 2012 and 2017?
2. Which area has the highest concentrated amount of sales?
3. Who are the best sellers (vendors) for the top 10 most popular liquor?
4. What are the monthly top 10 consumed liquor categories on average?
5. What are the monthly least 10 consumed liquor categories on average?
6. How many liquor stores are built in each city?
7. What are the top 10 stores that sell the most gallons of liquor each month?
8. What are the top 10 consumed liquors for each city in terms of profit?

4. Why is this big data?

- Volume: The data is 3.47 GB large and has 12591077 unique values, the huge volume of data is larger than the common datasets, which can provide us with more data available to analyze.

- Stability: According to the data, every row would be created if one alcohol sale order takes place in Iowa. During the months from January 2012 to 30 October 2017, the sales order recorded in each month kept the similar volume and there were no big fluctuations between it, which means this dataset grows in a high stability.

- Variety: What the dataset recorded is not only numerical data like price, but also categorical data, like the store number, the date, which help us analyze in time dimension.

- Value: The sales order data is valuable as it can help the entrepreneur to analyze critical factors that affect the alcohol sales in Iowa, as well as to make better decisions regarding market segmentation, location of stores and price of products to improve sales and maximize profits.

- Veracity: This raw dataset is collected by the The Iowa Department of Commerce and logged in the Commerce department system, which was in turn published as open data by the state of Iowa. There is every reason to believe it has a high degree of veracity.

- Tools used to process: since the dataset is large and structured, it cannot be opened by traditional data analysis tools such as Excel which would exceed the maximization limit. In this case, we use bash server, Hive/Impala/Spark, and tableau to visualize the data.

5. Methods

(1) Data preprocessing

- ❖ Download the raw data from the cloud.
- ❖ Processed the data with Python: we converted the data in the 'date' column into time and date format, and converted the strings of 'State Bottle Retail', 'State Bottle Cost' and 'Sale (Dollars)' into floats Format, and by traversing the data, we found that among the 12591077 pieces of data, 95103 pieces of data contained missing values, accounting for only 0.755% of the total data volume, so we chose to delete these data directly. In the end, we got a total of 12,495,974 pieces of valid data.
- ❖ Upload the data to the server and copy them to HDFS.
- ❖ Created a table called “iowa_liquor_sales_group_72” in Hive and loaded the CSV file into the table.

(2) Using mapreduce and Hive to analyze the data

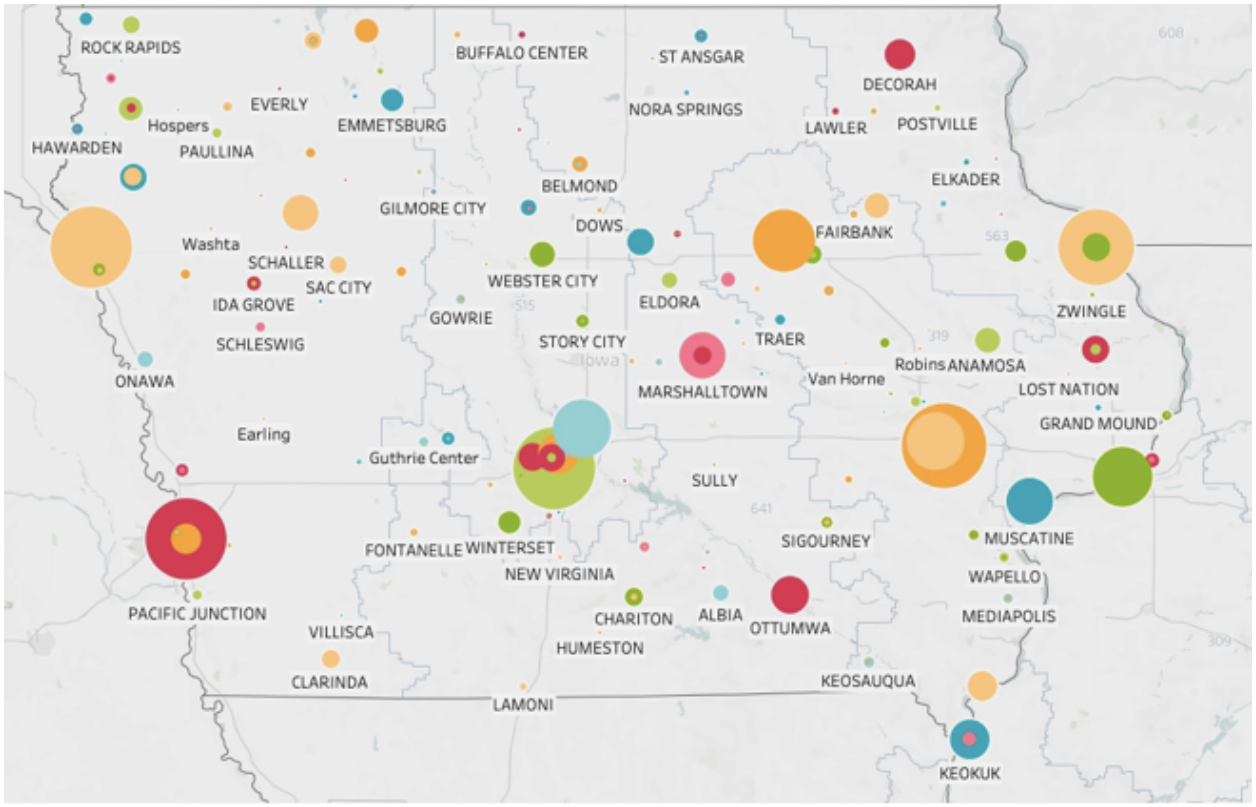
Given the goals of providing data-driven region-specific recommendations and market strategy, we went through the data and focused on some specific key variables that might help us draw up a business strategy, and did some related queries to gain structured data for further analysis.

We used SUM() function to calculate the total gallons of Liquor and used GROUP BY function to sort them by date, category and so on. Then, we used ORDER BY function to arrange the data from largest to smallest, in order to get the top 10 categories of Liquor sold the most in terms of

volume for each month. We change the descending order to the ascending order to get the least 10 categories of Liquor sold the most in terms of volume for each month. We used COUNT() function to get the total number of stores in each city.

(3) Visualization in Tableau

We drew a geographical map of the Iowa State to see the distribution of cities in terms of volume sold. The dot represents the volume sold of each city, each city represented by different colors and the volumes are proportional to the size of dots. According to the visualization map, we found a tendency that the right area of Iowa state tends to have higher demand for liquor. This geographic feature is helpful for us to make future optimization of storage and distribution strategy.



6. Result

(1) The steady Top 5 cities sold the most each year

From 2012 to 2017, the top 5 cities sold the largest amount of liquor in general. They are Des moines, Cedar rapids, Davenport, Iowa city, and Waterloo. These cities should be the main target city in regards to advertising activity. Volume sold of Des moines is way more larger than other cities. Regardless of renting cost or site conditions, selecting the location of the warehouse in Des moines can minimize transportation/distribution cost.

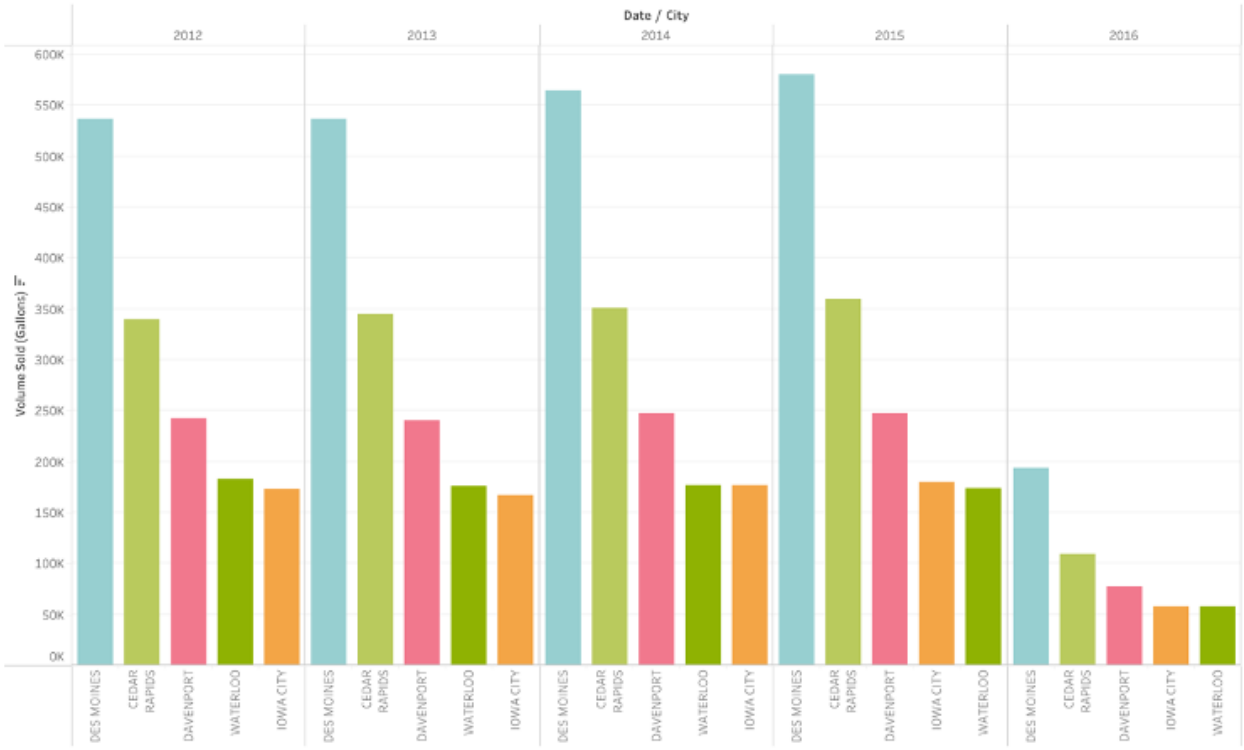
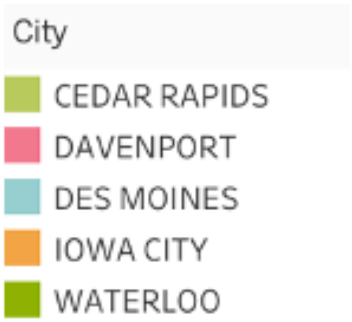


Figure 1



(2) Regional concentration of sales

From the tableau visualization map, we study the geographical information on regional concentration tendency of sales in Iowa (Figure 2). We found that the right area of Iowa tends to have higher demand for liquor. We first pinned the top 5 cities from result (1) on the map, then found most of them are concentrated on the right area (Figure 3). In terms of the site selection question, we prefer to locate our warehouse in the right part of the state, oriented to the best sellers cities.

We also estimated the Iowa Fair Market Rent for each city. We learned that Oskaloosa has a relatively low rent and has a nice geographical location, which is near to Des moines and other four cities (Figure 4).

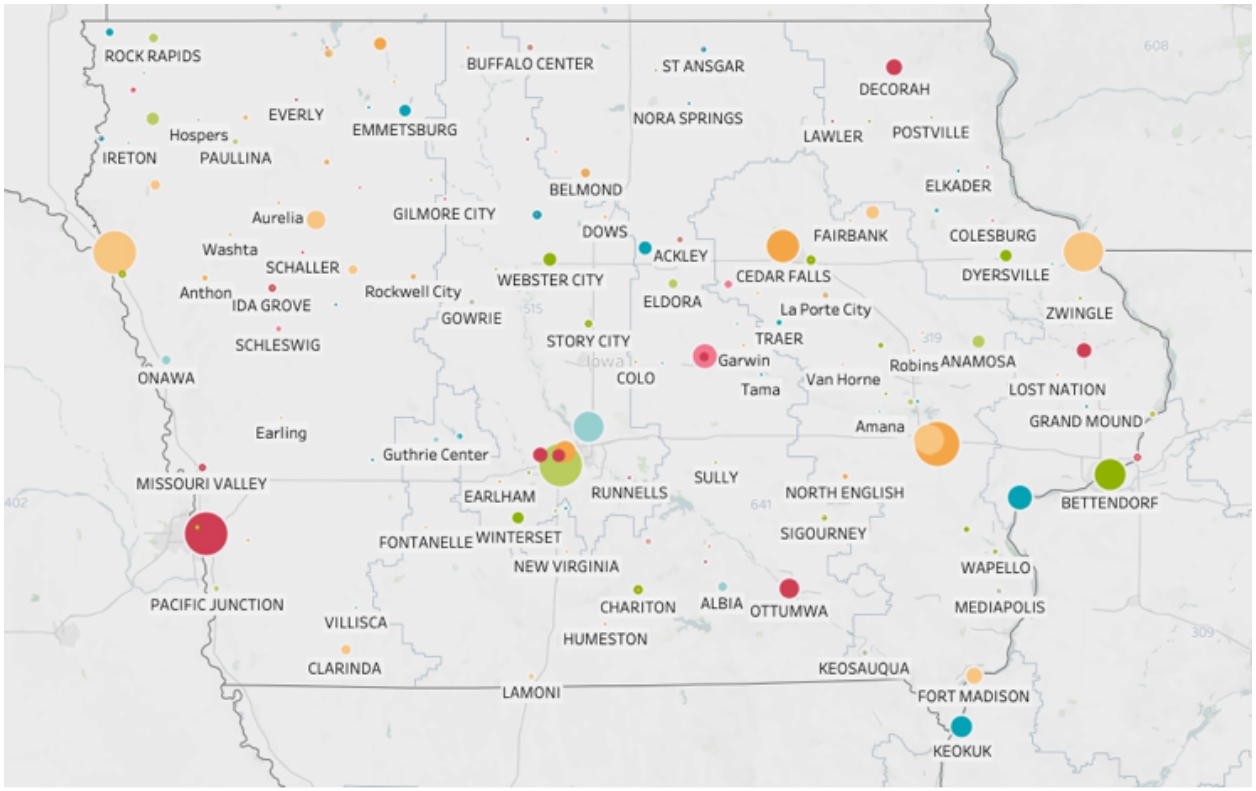


Figure 2

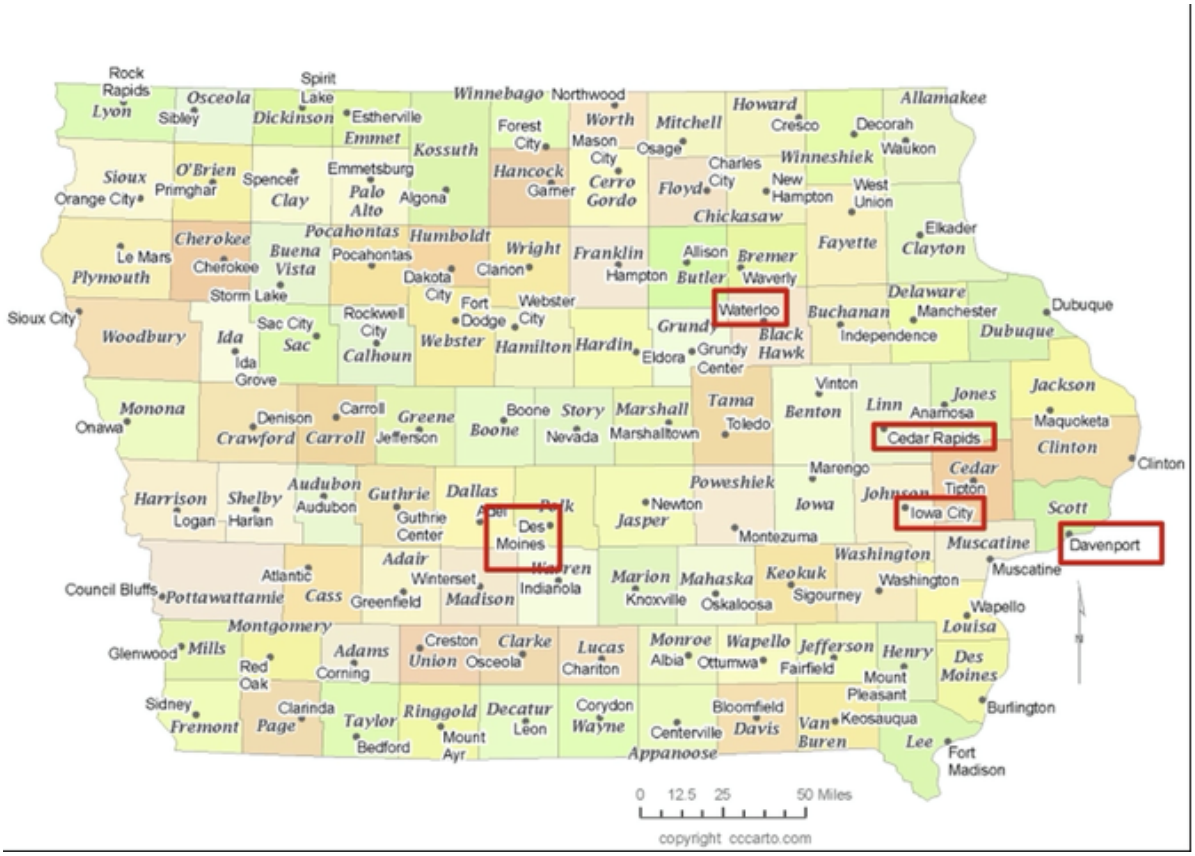

Rock Rapids
Spirit Lake
Lyon
Osceola
Sibley
Dickinson
Estherville
Emmet
Winnebago
Northwood
Kossuth
Forest City
Worth
Mitchell
Howard
Cresco
Decorah
Allamakee
Waukon
Mason City
Osage
Winneshiek
Charles City
New Hampton
West Union
Sioux
O'Brien
Primghar
Orange City
Spencer
Clay
Emmetsburg
Palo Alto
Algona
Hancock
Garner
Cerro Gordo
Floyd
Chickasaw
Elkader
Clayton
Fayette
Cherokee
Le Mars
Plymouth
Buena Vista
Pocahontas
Humboldt
Wright
Franklin
Hampton
Allison
Butler
Bremer
Waverly
Dakota City
Clarion
Storm Lake
Rockwell City
Webster City
Fort Dodge
Sioux City
Sac City
Woodbury
Ida
Sac
Calhoun
Webster
Hamilton
Hardin
Eldora
Grundy Center
Grundy
Black Hawk
Waterloo
Buchanan
Delaware
Manchester
Dubuque
Independence
Ida Grove
Monona
Onawa
Crawford
Carroll
Denison
Greene
Jefferson
Boone
Story
Nevada
Marshall
Marshalltown
Tama
Toledo
Benton
Vinton
Linn
Anamosa
Jones
Jackson
Maquoketa
Cedar Rapids
Clinton
Cedar
Tipton
Harrison
Logan
Shelby
Harlan
Audubon
Guthrie
Guthrie Center
Dallas
Polk
Des Moines
Jasper
Newton
Poweshiek
Marengo
Iowa
Johnson
Iowa City
Scott
Davenport
Muscatine
Washington
Montezuma
Adair
Atlantic
Winterset
Madison
Warren
Indianola
Marion
Knoxville
Mahaska
Oskaloosa
Keokuk
Sigourney
Washington
Muscatine
Council Bluffs
Pottawattamie
Cass
Greenfield
Wapello
Louisa
Montgomery
Glenwood
Mills
Red Oak
Adams
Corning
Creston
Union
Osceola
Clarke
Lucas
Chariton
Monroe
Albia
Wapello
Ottumwa
Jefferson
Fairfield
Henry
Mount Pleasant
Des Moines
Burlington
Sidney
Fremont
Clarinda
Page
Taylor
Bedford
Ringgold
Mount Ayr
Decatur
Leon
Corydon
Wayne
Centerville
Bloomfield
Davis
Van Buren
Keosauqua
Lee
Fort Madison
Appanoose
0 12.5 25 50 Miles
copyright cccarto.com

Figure 3

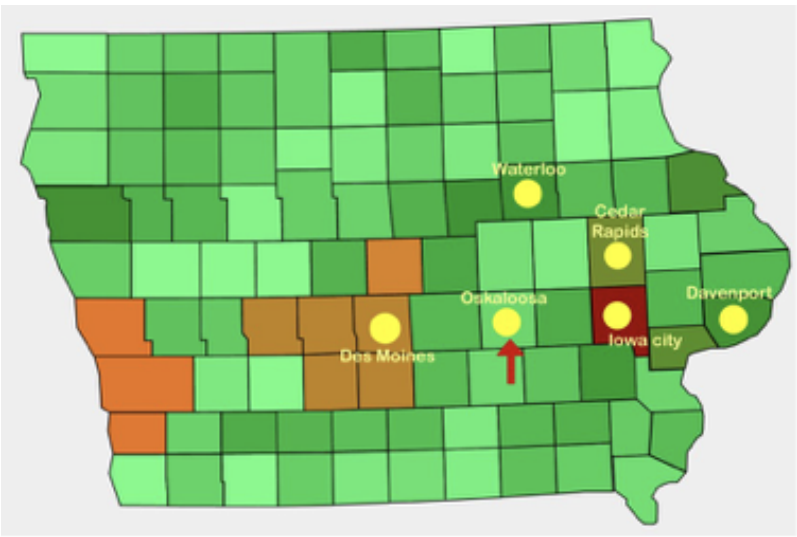
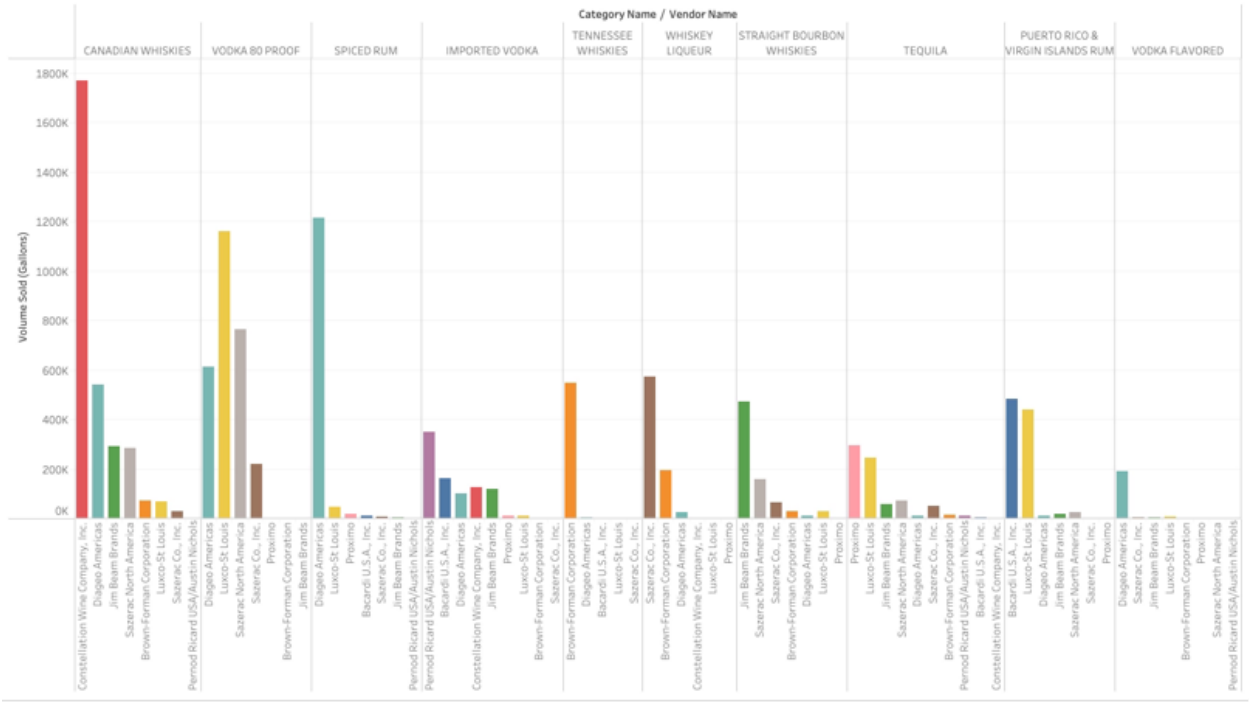
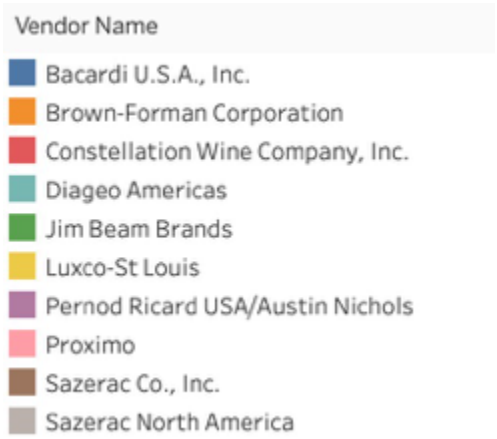


Figure 4 Iowa Fair Market Rent Map

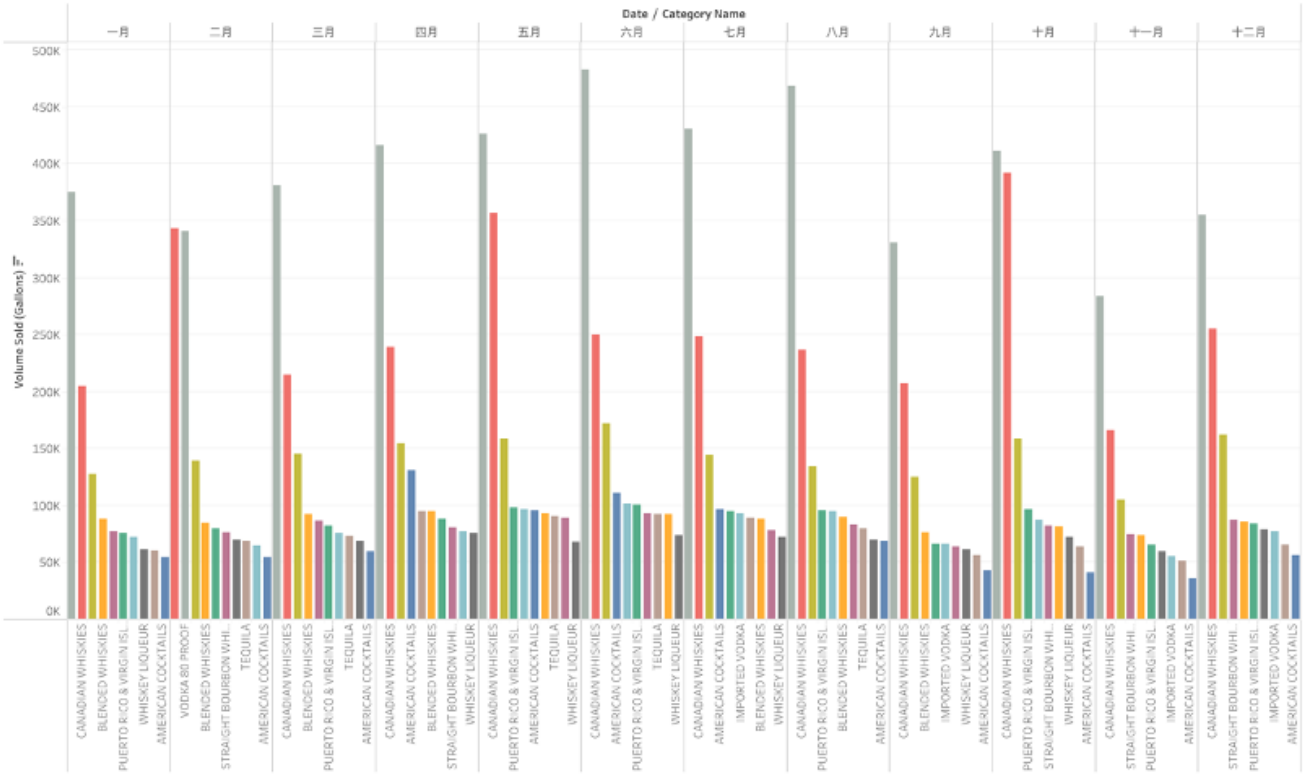
(3) Vendor





In particular categories, there are specific leading vendors with the most volumes sold, like the Constellation Wine Company, Inc., Diageo Americas and Luxco-St Louis. We can analyze the reason why they are popular, and set less stores selling such categories to avoid loss from competition with them. In contrast, we can set more stores selling such categories with the least vendors as the competition there is smaller.

(4) Top 10 Popular Consumed Liquors for Each Month



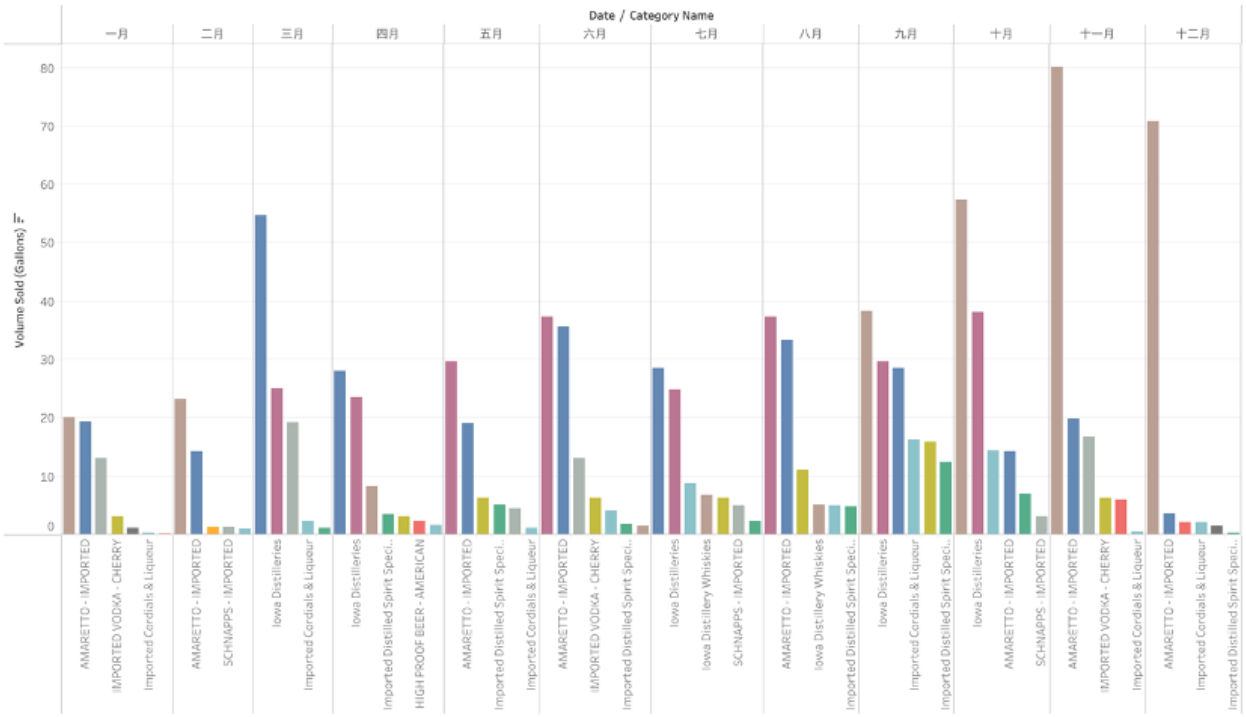
Among the top 10 popular consumed liquors in a year, there is a stable trend that VODKA 80 PROOF, CANADIAN WHISKIES and SPICED RUM have always been the top 3

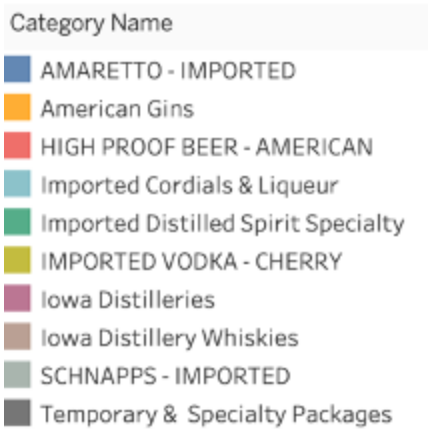
in each month. Notably, the monthly top 1 liquor has larger amounts of sales than the second rank, while there is no difference among the rest of the liquor in each month.

Another trend shows that whole volumes of the top 10 liquors in the spring and summer are apparently larger than the volumes in the winter, especially the volumes reached to a peak in June.

Therefore, we suggest that entrepreneurs can make more production on VODKA 80 PROOF, CANADIAN WHISKIES and SPICED RUM. We recommend increasing the VODKA 80 PROOF in stock in June. In winter seasons, entrepreneurs should consider decrease the production number to avoid loss.

(5) Least 10 brands sold the most in term of volume sold

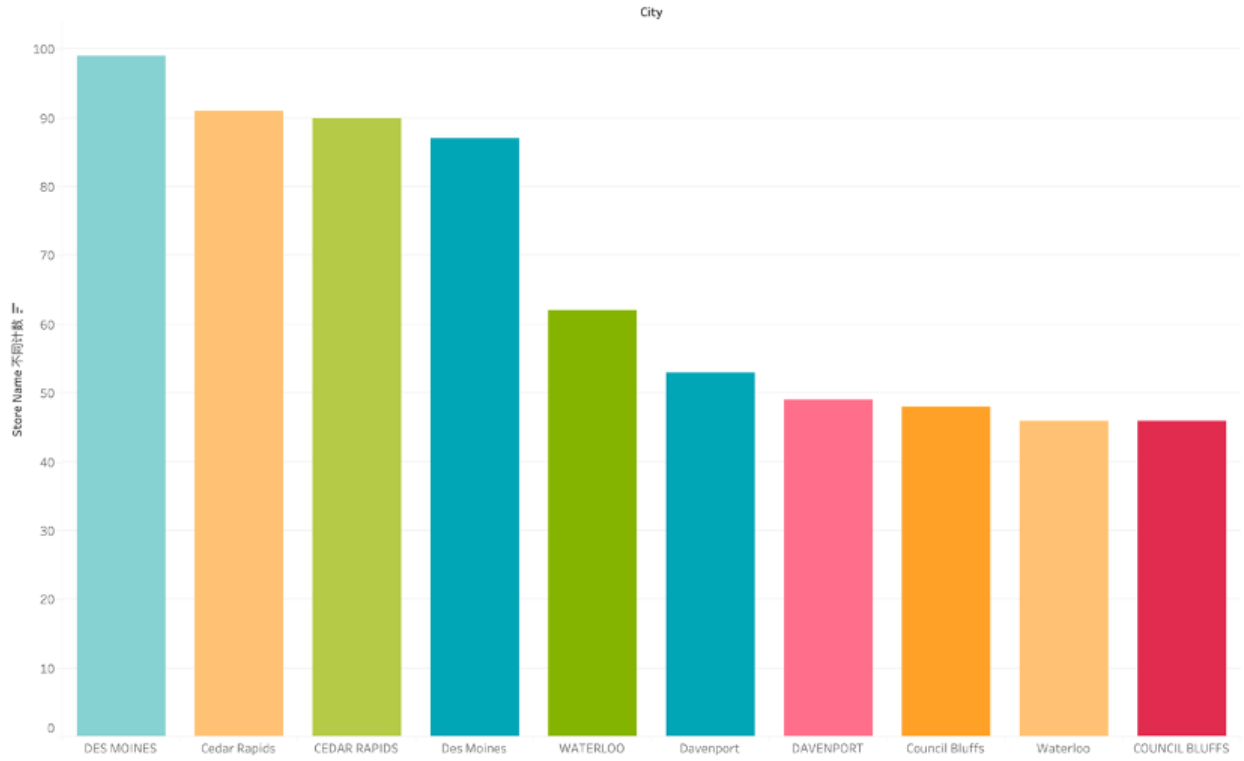




There is a fluctuation between the least 10 brands sold the most in terms of volume sold, Iowa Distillery Whiskies has been the least popular brand from Sep to Feb, while AMARETTO – IMPORTED and Iowa Distilleries has been the least 2 popular brands during the remaining months.

Therefore, we suggest that the new company could create marketing search to the above 3 companies, analyzing the reason why they are unpopular in such specific seasons and make reversions to its own products, to produce more popular products in such seasons.

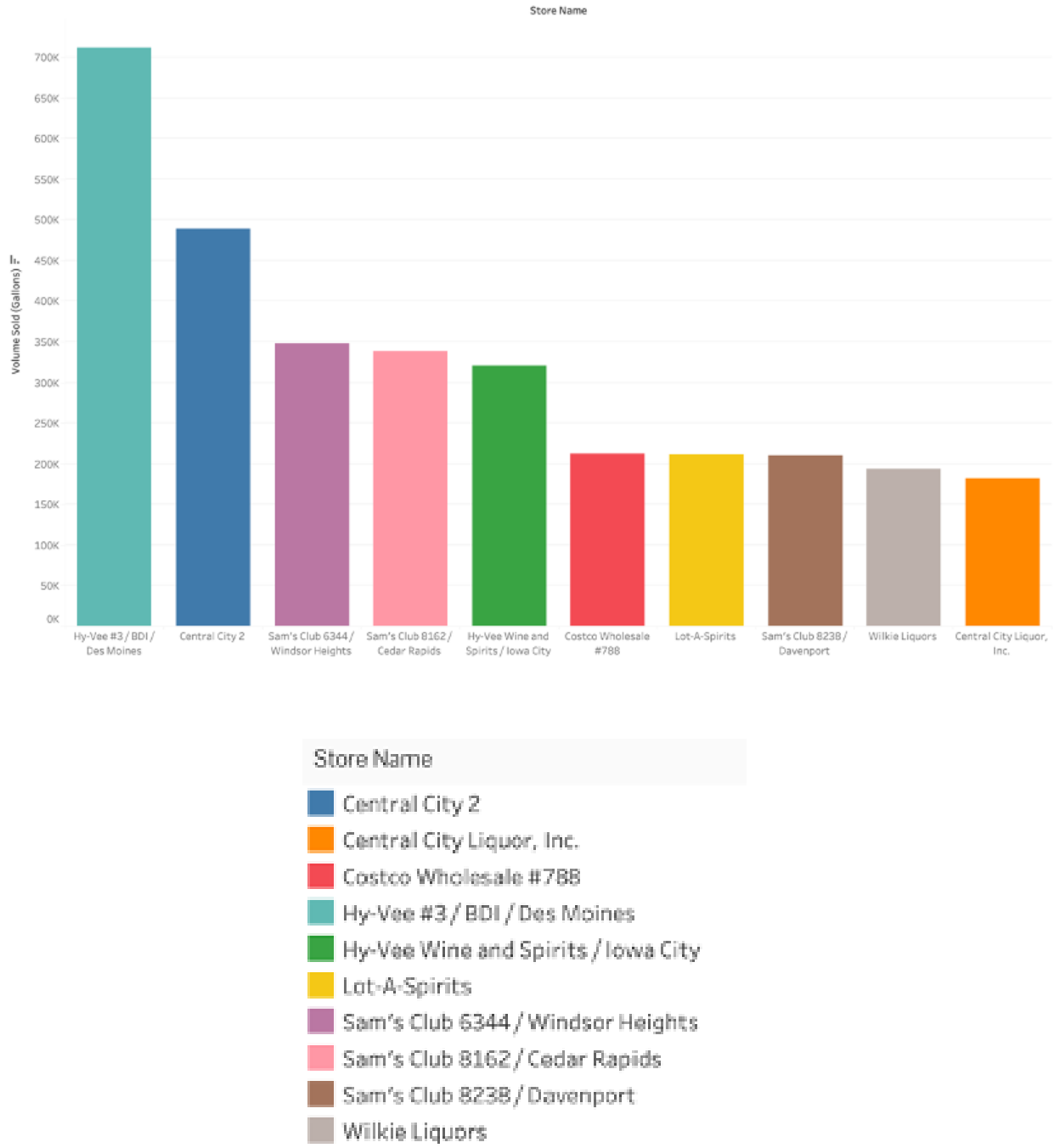
(6) Number of Stores for Each City



According to the above data, we find that there is an apparent difference between the 4th city and the 5th city. We need to consider the competitor factor when deciding the numbers of stores of our brand put in each city. For the top 3 cities, there may be more competitors, while for the least 5 cities, there may be location restrictions limiting the sales.

Therefore, it is better to consider the 4th and 5th city. There is a gap between them, which is a great opportunity. We could analyze the reason for the gap: from the result 1, we found that the gap is possibly due to geography or transportation factors. So we can put stores in the two cities with more stores in the Des Monies and less in WATERLOO. In a word, we need regard these 2 cities as our target markets.

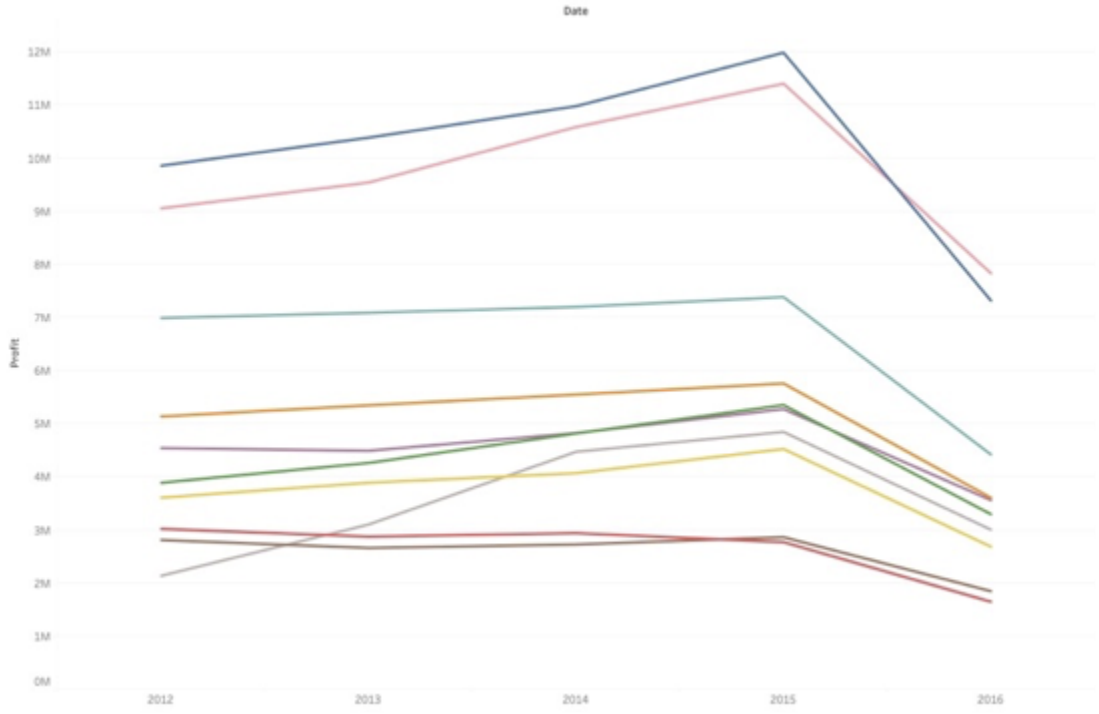
(7) Top 10 Stores that Sell the Most Gallons of Liquor

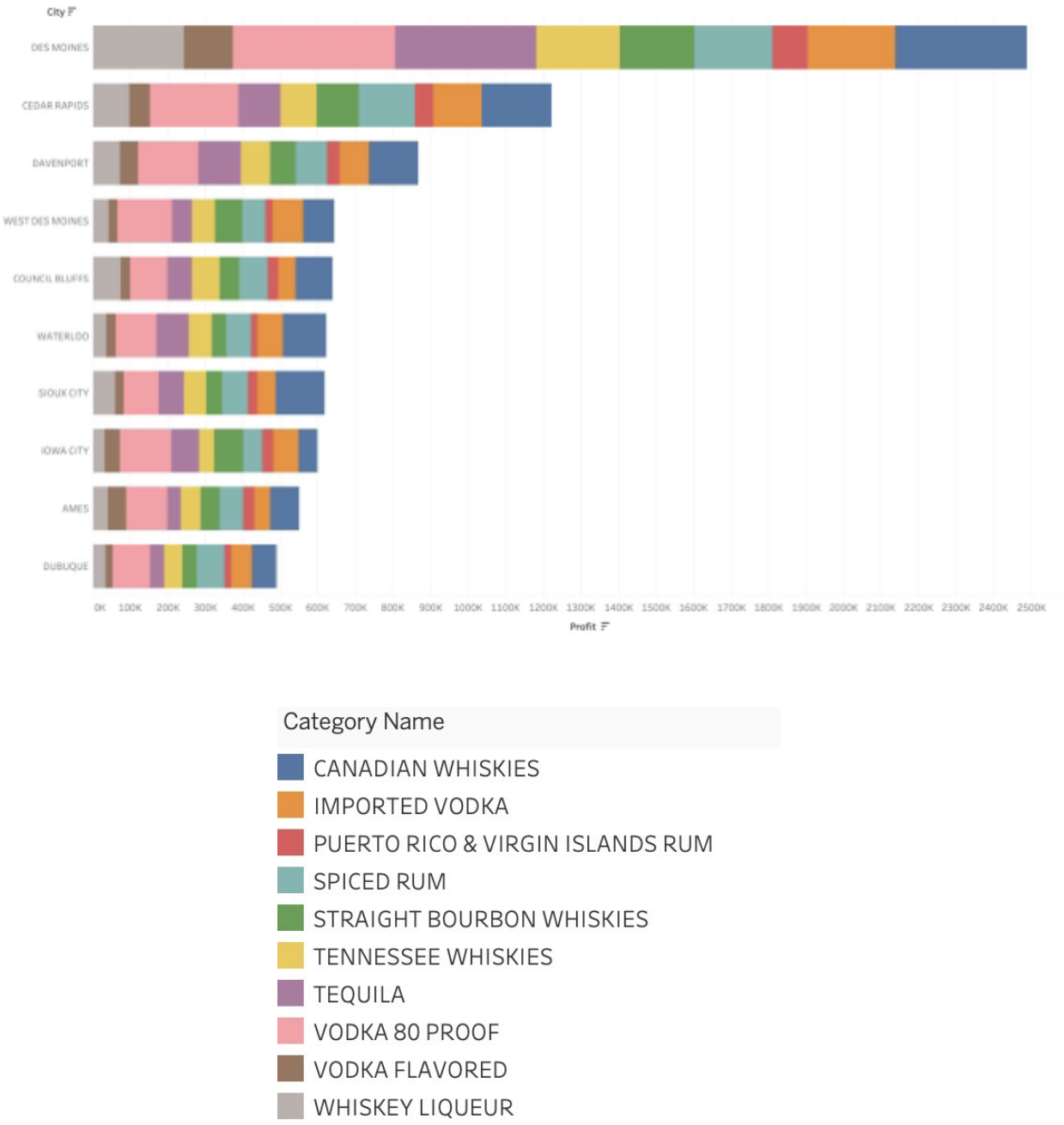


From the above data we can find that the first store has a large difference from the other stores, while there is little difference between the 3~5 and the 6~10 stores. Through analyzing

the top 2 cities & least 5 stores' locations, we could identify the existing numbers of competitors in the market.

(8) Top 10 Popular Consumed Liquors for each city in terms of profit





Besides consideration of regional concentration sales, we need to think about a good location to start the business. The most apparent trend about this data is that top 10 liquors have much larger volumes in the Des Monies than that in the other cities, which indicates that it might have a better market environment to sell the products. Therefore, Des Monies is our

recommended first choice. CEDAR RAPIDS and DAVENPOT are also potential target markets because they have higher volume than the rest.

7. Conclusion

(1) To make better decisions for the new company to maximize profit, there are various factors to be considered, such as: location, popularity of brands, seasons, competitors and so on. By combining these factors, we make a overall suggestion with specific details to the entrepreneur as follows:

(2) Considering the location and the volume sold conditions, selecting the location of the warehouse in Des moines can minimize transportation costs.

(3) In terms of the site selection question, it is better for us to locate our warehouse in the right part of the state, oriented to the best sellers cities. Also, Des Monies and WATERLOO could be the target markets.

(4) Entrepreneurs can make more production on VODKA 80 PROOF, CANADIAN WHISKIES and SPICED RUM. We recommend increasing the VODKA 80 PROOF in stock in June. In winter, entrepreneurs should consider decreasing the production number to avoid loss.

(5) The new company could make marketing searches to the top and least popular competitors, analyze the reason why they are popular/unpopular in such specific seasons/categories/cities and make reversions to its own products, to produce more popular products in such seasons/categories/cities.

Appendix

Code 1: Data preprocessing

```
import pandas as pd
df = pd.read_csv('Iowa_Liquor_Sales.csv')

df['Date'] = pd.to_datetime(df['Date'])
df['State Bottle Retail'] = df['State Bottle Retail'].str.replace('$', '').astype('float')
df['State Bottle Cost'] = df['State Bottle Cost'].str.replace('$', '').astype('float')
df['Sale (Dollars)'] = df['Sale (Dollars)'].str.replace('$', '').astype('float')
df.dropna(inplace = True)
df.to_csv('iowa_liquor_sales.csv')
```

Code 2: Find the top 5 cities sold the most each year between 2012 and 2017

Map code:

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
        line = line.strip()
        city = line.split(",")[6]
        year = line.split(",")[2]
        sale = line.split(",")[22]
        try:
                year = year.split('-')[0]
                print '%s\t%s\t%s'%(city,year,sale)
        except:
                continue
```

Reduce code:

```
#!/usr/bin/env python
import sys
sale_2012 = {}
sale_2013 = {}
sale_2014 = {}
sale_2015 = {}
sale_2016 = {}
sale_2017 = {}
for line in sys.stdin:
        city, year,sale = line.strip().split('\t')
        try:
                year = int(year)
                sale = float(sale)
                count_year = str(year)+"  "
        except:
                continue
        if year == 2012:
                try:
                sale_2012[count_year+city]+=sale
                except:
                sale_2012[count_year+city] = sale
        elif year == 2013:
        try:
                sale_2013[count_year+city]+=sale
                except:
                sale_2013[count_year+city] = sale
        if year == 2014:
                try:
              sale_2014[count_year+city]+=sale
                except:
```

```
            sale_2014[count_year+city] = sale
        elif year == 2015:
        try:
            sale_2015[count_year+city]+=sale
          except:
            sale_2015[count_year+city] = sale
        if year == 2016:
            try:
            sale_2016[count_year+city]+=sale
            except:
            sale_2016[count_year+city] = sale
        elif year == 2017:
        try:
            sale_2017[count_year+city]+=sale
            except:
            sale_2017[count_year+city] = sale
print(sorted(sale_2012.items(),key = lambda x:x[1],reverse = True)[0:5])
print(sorted(sale_2013.items(),key = lambda x:x[1],reverse = True)[0:5])
print(sorted(sale_2014.items(),key = lambda x:x[1],reverse = True)[0:5])
print(sorted(sale_2015.items(),key = lambda x:x[1],reverse = True)[0:5])
print(sorted(sale_2016.items(),key = lambda x:x[1],reverse = True)[0:5])
print(sorted(sale_2017.items(),key = lambda x:x[1],reverse = True)[0:5])
```

Bash code:

```
#!/bin/bash
hadoop jar
/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hadoop-streaming-3.1.1.7.1.7.0-5
51.jar \
        -Dmapred.reduce.tasks=1 \
        -input /user/lanlin.s/Iowa_Liquor_Sales_group72.csv \
```

```
    -output /user/lanlin.s/final_output \
    -file iowa_liquor_sales_map.py\
  -file iowa_liquor_sales_red.py \
    -mapper "python iowa_liquor_sales_map.py" \
    -reducer "python iowa_liquor_sales_red.py"
```

Code 3: Find which city has the highest concentrated amount of sales

```
SELECT city , sum( sale  ) as TotalSale
FROM `default`.`iowa_liquor_sales_group_72`
GROUP BY city;
```

Code 4: Find who are the best sellers (vendors) for the top 10 most popular liquor

```
SELECT categor_ name , vendor_name, sum( volume_sold_gallons  ) as TotalGallons
FROM `default`.`iowa_liquor_sales_group_72`
GROUP BY categor_ name and vendor_name
ORDER BY TotalGallons DESC
LIMIT 10;
```

Code 5: Find the top 10 categories of Liquor sold the most in terms of volume for each month

```
SELECT category, sale_date , sum( volume_sold_gallons  ) as TotalGallons
FROM `default`.`iowa_liquor_sales_group_72`
GROUP BY category and sale_date
ORDER BY TotalGallons DESC
LIMIT 10;
```

Code 6: Find the top 10 categories of Liquor sold the least in terms of volume for each month

```
SELECT category, sale_date , sum( volume_sold_gallons  ) as TotalGallons
FROM `default`.`iowa_liquor_sales_group_72`
```

```
GROUP BY category and sale_date
ORDER BY TotalGallons ASC
LIMIT 10;
```

Code 7: Find how many stores for each city

```
SELECT city,  count( store_number ) as Number
FROM `default`.`iowa_liquor_sales_group_72`
GROUP BY city
ORDER BY Number DESC
LIMIT 10;
```

Code 8: Find top 10 Stores that Sell the Most Gallons of Liquor

```
SELECT store_name, sum( volume_sold_gallons  ) as TotalGallons
FROM `default`.`iowa_liquor_sales_group_72`
GROUP BY store_name
ORDER BY TotalGallons DESC
LIMIT 10;
```

Code 9: Find top 10 Popular Consumed Liquors for each city in terms of profit

```
SELECT category, city , sum((state_bottle_retail -  state_bottle_cost )*bottles_sold) as profit
FROM `default`.`iowa_liquor_sales_group_72`
GROUP BY category and city
ORDER BY profit DESC
LIMIT 10;
```