The Sales Situation of Liquors in Different Regions in Iowa 2017
Final Report

Lei Teng
lete3485@colorado.edu

Mingxuan Zhang
mizh1382@colorado.edu

Yuxiang Wang
yuwa4103@colorado.edu

ABSTRACT

Nowadays, alcoholic beverage becomes one of the most important things in our life. People use liquors in many different ways. Such as cooking, medicine, etc. But for most of us, alcoholic beverage is used for drinking and then release our pressure. And what we did in this project is to analyze the sales situation of liquors in different regions in Iowa. What's more, we built up some models to predict missing values and mining some other information.

KEYWORDS

Liquors, alcoholic beverage

1 INTRODUCTION

More specifically, our project works for analyzing the name, date, kind, price, quantity, and location of sales of individual containers or packages of containers of alcoholic beverages in 2017 to get the sales situation and drinking behaviors of people in different regions.

1.1 Problem Statement

1.1.1 Sales Situation

1.1.1.1 Description

For the sales situation, we can analyze the data to gain some correlation results. Such as the total amount of alcoholic beverage sold and consumed in months, years and regions. The alcoholic beverage sells best in different regions. Which region has the most liquors' store. By analyzing the correlation of price and locations, we can get the region which has the highest price. By analyzing the correlation of date (months, years), and sales volume, we can get the tendency of months and years in different regions. Analyzing the correlation of name (or kind) and price to get the alcoholic beverage which gains the maximum profit. Analyzing the correlation of date and price to get the distribution in different regions, etc.

2.1.2 Specific Questions

- As for each vendor, which regions gains the maximum revenue?
- Which category of alcoholic beverages is the most popular in different regions?
- Which category of alcoholic beverages gains the maximum revenue?
- Which brand of alcoholic beverages gains the maximum revenue?

These questions may help the beverage producers know more about the sales situation to help them realize if these kinds of alcoholic beverages is popular or not.

1.1.2 Drinking Behaviors

1.1.2.1 Description

For the drinking behavior, according to the results of sales situation, such as the total sales amount of alcoholic beverage, can tell us people in which region drink the most alcoholic beverage or in which season people drink more. Also, from the dataset, we can get different sales rate of different alcoholic beverage in a specific region, then we will know which alcoholic beverage is the most popular in that region.

2.2.2 Specific Questions

- Which category of alcoholic beverages is the most popular?
- People in which regions consumed the most alcoholic beverages?
- How the alcoholic beverage sale situation distributes in the first half year and what it means?

These questions may help us and the beverage producers know more about the drinking behaviors of people and thus, they can make a better market planning.

2 RELATED WORK

2.1 Related work (1)

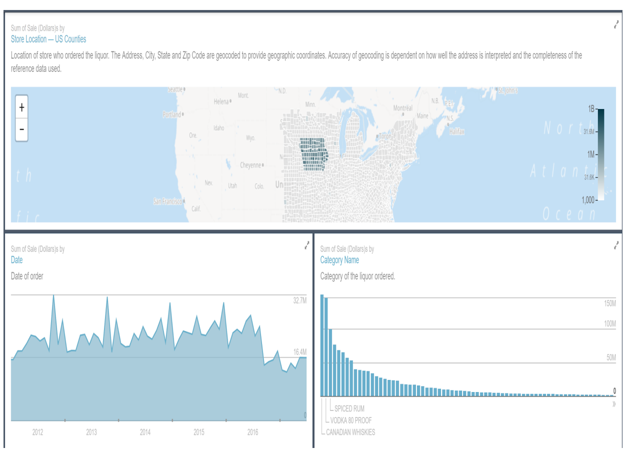


Figure 1 [1]

Figure 1: This work is about Iowa Liquor Sales in Dollars. It contains a map and two plots, which give us the sales in dollars respect to locations and times and categories.

2.2 Related work (2)

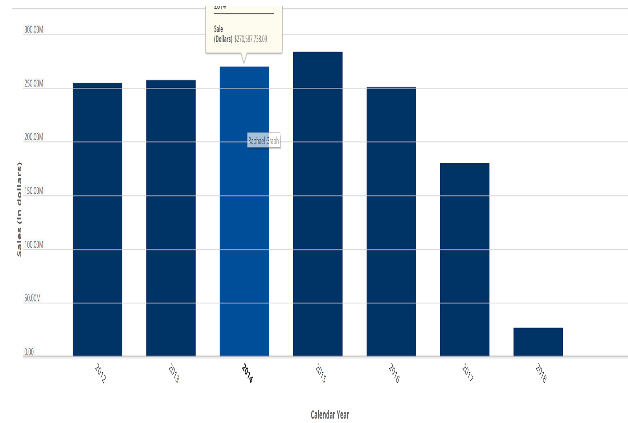


Figure 2 [2]

Figure 2: This work gives us the histogram about Iowa Liquor Sales in Dollars by Year.

2.3 Related work (3)

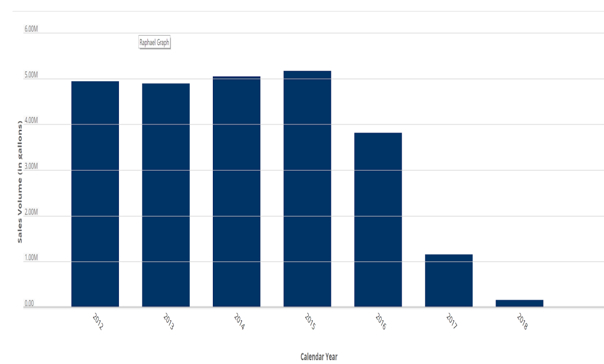


Figure 3 [3]

Figure 3: This work gives us the histogram about Iowa Liquor Sales in Gallons by Year.

2.4 Related work (4)

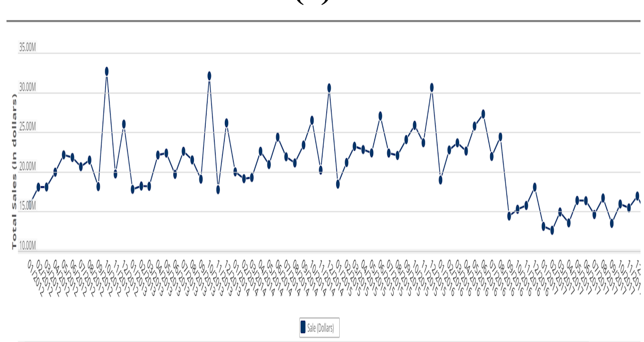


Figure 4 [4]

Figure 4: This work gives us the histogram about Iowa Liquor Sales in Gallons by Year.

2.5 Related work (5)

	Date	County	Sale (Dollars)	Volume Sold (Gallons)
1	2018	Adair	$22,292.57	203.18
2	2018	ADAIR	$25,368.84	156.89
3	2018	ADAMS	$4,709.95	86.47
4	2018	ALLAMAKEE	$76,742.45	575.00
5	2018	APPANOOSE	$84,414.18	469.24
6	2018	AUDUBON	$14,764.40	349.15
7	2018	BENTON	$95,766.11	735.35
8	2018	Black Hawk	$43,930.93	466.45
9	2018	BLACK HAWK	$1,599,401.10	9,119.71
10	2018	Boone	$18,989.58	226.71
11	2018	BOONE	$213,670.17	1,228.03
12	2018	BREMER	$183,483.74	1,389.14
13	2018	Buchanan	$26,826.39	304.41
14	2018	BUCHANAN	$112,582.86	953.62
15	2018	BUENA VIST	$176,043.82	1,425.38
16	2018	Butler	$2,946.21	40.47
17	2018	BUTLER	$36,219.57	262.02
18	2018	CALHOUN	$44,428.77	293.93
19	2018	CARROLL	$179,503.17	1,165.40

Figure 5 [5]

Figure 5: This work gives us the table about Iowa Liquor Sales by Year and County. In this table, we can get the sales in gallons and dollars in different time and county.

3 Data Set

URL: https://www.kaggle.com/residentmario/iowa-liquor-sales/data[6]

This dataset contains information on the name, kind, price, quantity, and location of sale of sales of individual containers or packages of containers of alcoholic beverages.

Our datasets have around 12 million objects and 24 different attributes [7]:

Invoice/Item Number: Concatenated invoice and line number associated with the liquor order.
Date: Date of order.
Store Number: Unique number assigned to the store who ordered the liquor.
Store Name: Name of store who ordered the liquor.
Address: Address of store who ordered the liquor.
City: City where the store who ordered the liquor is located.
Zip Code: Zip code where the store who ordered the liquor is located.
Store Location: Location of store who ordered the liquor.
County Number: Iowa county number for the county where store who ordered the liquor is located.
County: County where the store who ordered the liquor is located.
Category: Category code associated with the liquor ordered.
Category Name: Category of the liquor ordered.
Vendor Number: The vendor number of the company for the brand of liquor ordered.
Vendor Name: The vendor name of the company for the brand of liquor ordered.
Item Number: Item number for the individual liquor product ordered.
Item Description: Description of the individual liquor product ordered.
Pack: The number of bottles in a case for the liquor ordered.
Bottle Volume (ml): Volume of each liquor bottle ordered in milliliters.
State Bottle Cost: The amount that Alcoholic Beverages Division paid for each bottle of liquor ordered.
State Bottle Retail: The amount the store paid for each bottle of liquor ordered
Bottles Sold: The number of bottles of liquor ordered by the store.
Sale (Dollars): Total cost of liquor order (number of bottles multiplied by the state bottle retail).
Volume Sold (Liters): Total volume of liquor ordered in liters.
Volume Sold (Gallons): Total volume of liquor ordered in gallons.

4 MAIN TECHNIQUES

4.1 Data Cleaning

We delete the rows which have missing values, in order to build an accurate Bayesian Classification Model, the missing values may make some negative effects to the veracity for our model.

4.2 Data Reduction

The original dataset contains from 2012 to current, its over 12 million while our project is working for 2017. Thus, we did the reduction only left datasets about 2017.

Since our data set has 24 attributes, some of them are not useful for our data analyzing, we did the dimensionality reduction. We deleted those attributes and removed those redundant attributes, thus, we can improve our efficiency and insure that our results are correct.

For example, we deleted: ‘Store location’, ‘County’, ‘Invoice/Item Number’, etc. Because they are the redundant attributes.

What’s more, since Volume Sold (Liters) and Volume Sold (Gallons) are almost the same. So, we can use Volume Sold(Liters) instead of Volume Sold (Gallons).

4.3 Data Transformation

“Sales numbers” contains “$” which causes panda to fail to identify them as numbers. So, we deleted the “$” and transform it from type string to type float.

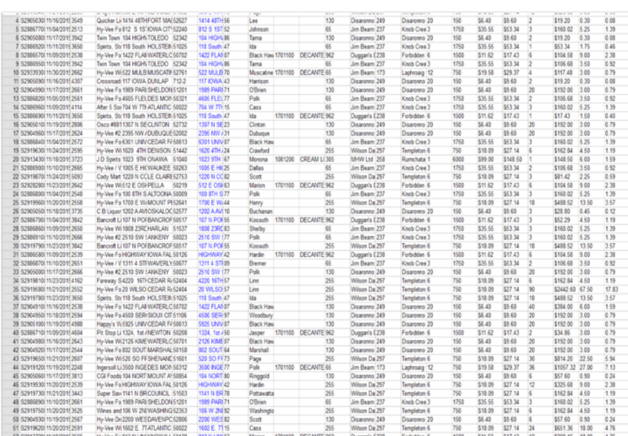


Figure 6

Figure 6: Before the data preparation work, our data set looks intricate and complex.

		Number	Number			Name	Number	Name	Number	Description		(ml)	Cost	Retail	Sol
10708015	01/03/2017	4312	78.0	POTTAWATTA	1012200.0	Scotch Whiskies	55.0	SAZERAC NORTH AMERICA	8208	House Of Stuart	6	1750	$10.52	15.78	
10708016	01/03/2017	4312	78.0	POTTAWATTA	1042100.0	Imported Dry Gins	35.0	BACARDI USA INC	28206	Bombay Dry Gin	12	750	$10.50	15.75	
10708017	01/03/2017	4312	78.0	POTTAWATTA	1082000.0	Imported Cordials & Liqueurs	259.0	Heaven Hill Brands	65195	Hpnotiq	6	750	$9.83	14.75	
10708018	01/03/2017	4312	78.0	POTTAWATTA	1081200.0	Cream Liqueurs	260.0	DIAGEO AMERICAS	68037	Bailey's Original Irish Cream	12	1000	$16.50	24.75	
10708019	01/03/2017	4312	78.0	POTTAWATTA	1012100.0	Canadian Whiskies	65.0	Jim Beam Brands	10627	Canadian Club Whisky	12	1000	$9.71	14.57	
10708020	01/03/2017	4312	78.0	POTTAWATTA	1062200.0	White Rum	55.0	SAZERAC NORTH AMERICA	44217	Barton Rum Light	12	1000	$4.00	6.00	
10708021	01/03/2017	4312	78.0	POTTAWATTA	1062400.0	Spiced Rum	260.0	DIAGEO AMERICAS	43338	Captain Morgan Spiced Rum	6	1750	$18.00	27.00	
10708022	01/03/2017	4312	78.0	POTTAWATTA	1031000.0	American Vodka	55.0	SAZERAC NORTH AMERICA	35318	Barton Vodka	6	1750	$6.92	10.38	
10708023	01/03/2017	4312	78.0	POTTAWATTA	1032200.0	Imported Flavored Vodka	370.0	PERNOD RICARD USA	34051	Absolut Raspberri	12	1000	$14.99	22.49	
10708024	01/03/2017	4312	78.0	POTTAWATTA	1032000.0	Imported Vodka	370.0	PERNOD RICARD USA	34007	Absolut Swedish Vodka 80 Prf	12	1000	$14.99	22.49	
10708025	01/03/2017	4312	78.0	POTTAWATTA	1031000.0	American	461.0	Skyy Spirits	37987	Skyy Vodka	12	1000	$12.35	18.53	

Figure 7

Figure 7: After do the preparation work, our data got simplified and looks good.

4.4 Classification - Bayesian Classification

We built a model, by using this conditional probability:

P (category | store number, vendor number)

If we have the store number and vender number, we can get the posterior probability distribution for the categories. We used this to predict the values for category attribute.

We can pick the one with the highest probability or we can use this posterior distribution to draw samples for category attribute.

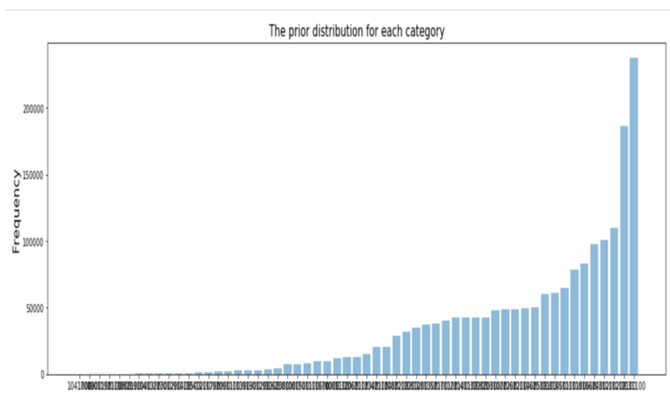


Figure 8

Figure 8: This is the prior distribution for category by scanning the entire data set for 2017, which is not useful, since we don't have other information, but if we use my model, we get some really useful results, like figure 9.

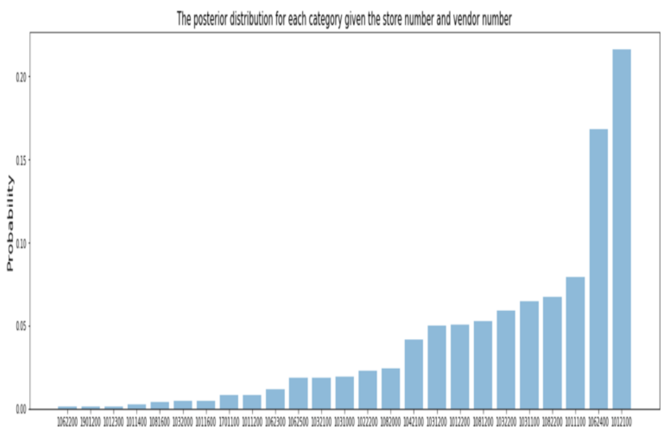


Figure 9

Figure 9: This is figure 9 is the example of using the Bayesian classification model, the values on y axis are the corresponding probabilities for each category, now we can use this posterior distribution to classify the category, we can either pick the category with highest probability or we can use this distribution to draw random samples.

4.5 Data Analysis

4.5.1 Build Function (1)

We build a function to analyze the data to get the top 8 counties which brought the best revenue for each vender. (Vender number as an input.)

```
In [98]: def best8_counties(vendor_number):
             sorted_v = sorted(vendor_county[vendor_number].items(), key=operator.itemgetter(1))
             x, y = map(list, zip(*sorted_v))
             c8 = y[len(y)-8:]
             numbers = x[len(x)-8:]
             objects = []
             for number in numbers:
                 objects.append(list(df.loc[df['County Number'] == number, 'County'])[0])
             y_pos = np.arange(len(objects))
             plt.figure(figsize=(24, 8))
             plt.bar(y_pos, c8, align='center', alpha=0.5)
             plt.xticks(y_pos, objects)
             plt.ylabel('Revenue', fontsize=14)
             title = 'The 8 counties that like vendor ' + list(df.loc[df['Vendor Number'] == vendor_number, 'Vendor Name'])[0]
             plt.title(title)
             plt.show()

In [99]: best8_counties(260)
```

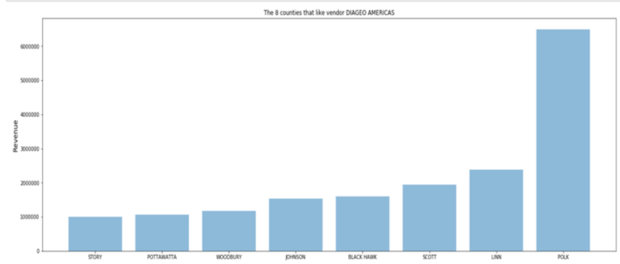


Figure 10

Figure 10: This is the function we built to analyze the sale situation at each county for a specific vendor. This function takes the vendor number as the input, the output are top 8 counties which will bring the vendor the highest revenue. For example, in this graph the vendor is DIAGEO AMERICAS, we can see that for this vendor, people in POLK seems really likes their products. Also since this is a function, we can use this function to do analysis for multiple companies, which is really useful.

```
best8_counties(461)
```

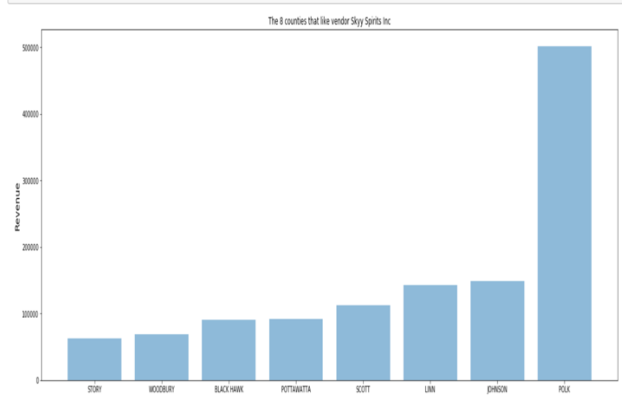


Figure 11

Figure 11: This is another outcome from this function, this is the top 8 counties for vendor Skyy Spirits Inc, we can see that both vendors are popular POLK, I think this because the population at POLK is higher than other counties, but there are differences too, like

the county JOHNSON prefer Skyy Spirits Inc than DIAGEO AMERICAS.

4.5.2 Build Function (2)

We build a function to analyze the data to get the top 8 categories consumed in different regions.
(County number as an input.)

```
In [101]: def best_cate_county(CN):

              sorted_c = sorted(county_cates[CN].items(), key=operator.itemgetter(1))
              x, y = map(list, zip(*sorted_c))
              c6 = y[len(y)-8:]
              numbers = x[len(x)-8:]
              objects = []
              for number in numbers:
                  objects.append(list(df.loc[df['Category'] == number, 'Category Name'])[0])
              y_pos = np.arange(len(objects))
              plt.figure(figsize=(24, 8))
              plt.bar(y_pos, c6, align='center', alpha=0.5)
              plt.xticks(y_pos, objects)
              plt.ylabel('Total volume consumed at 2017', fontsize=14)
              title = 'The 8 most popular types of liquors at county ' + list(df.loc[df['County Number'] == CN, 'County'])[0]
              plt.title(title)
              plt.show()
```

```
In [102]: best_cate_county(78)
```

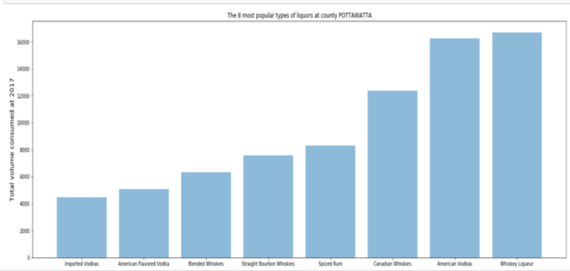


Figure 12

Figure 12: We also built another function which takes the county number as the input. The output is the 8 most popular types of liquors in that county. For example, if the county number is 78 which corresponding to POTTAWATTA, we can see the people in this county really like to consume Whiskey. Thus, vendors can use this function to analyze the drinking behavior in different regions, then they can promotion that specific type of liquor, for example, vendor can promotion their Whiskey productions in POTTAWATTA.

```
best_cate_county(7)
```

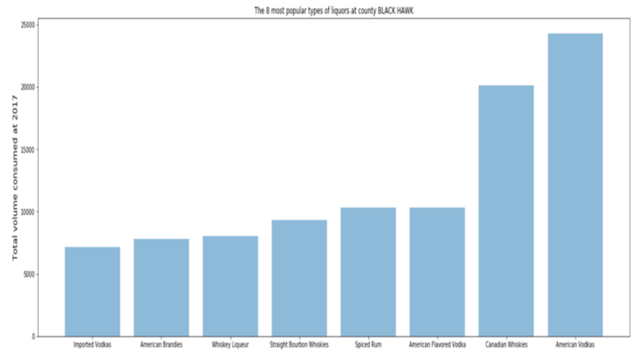


Figure 12

Figure 12: This is another graph for this function for county BLACK HAWK, we can see there are difference for the drinking behavior in these two counties, the people in POTTAWATTA likes Whiskey liqueur most, but people in BLACK HAWK likes America Vodkas best, and the Whiskey liqueur is not popular at BLACK HAWK compare to POTTAWATTA.

4.5.3 Other Analysis

4.5.3.1 Revenue and Category Analyzing

We generate a bar chart diagram of the attribute ‘Category Number’ to get the revenue of each kind of alcoholic beverage.

Which means we can know which kind of alcoholic beverage gains the most revenue in Iowa in 2017.

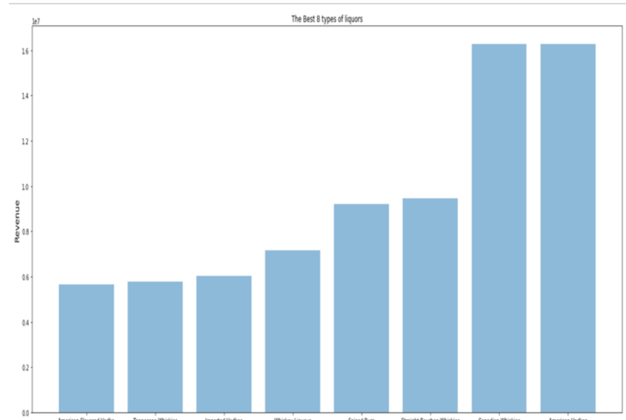


Figure 13

Figure 13: For example, in 2017 the American Vodkas gains the highest revenue.

4.5.3.2 Revenue and Brand Analyzing

We generate a bar chart diagram of the attribute 'Item Number' to get the revenue of each kind of alcoholic beverage.

Which means we can know which kind of alcoholic beverage gains the most revenues in Iowa in 2017.

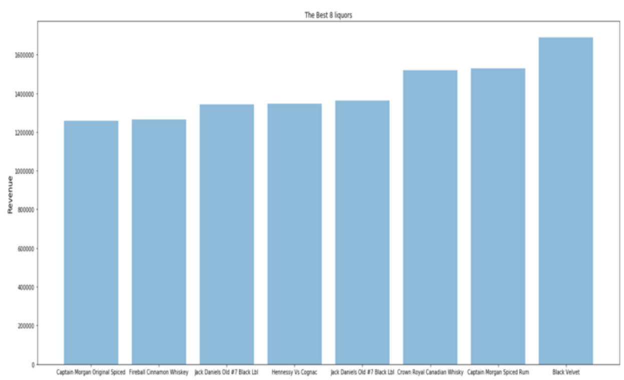


Figure14

Figure 14: For example, in 2017, the Black Velvet gains the highest revenue.

4.5.3.3 Total volume sold and Category analyzing

We built this graph by scanning entire data set, during the scanning, I use the dictionary to manage the data, the keys for the dictionary are the category number, the values are the total volume consumption in 2017.

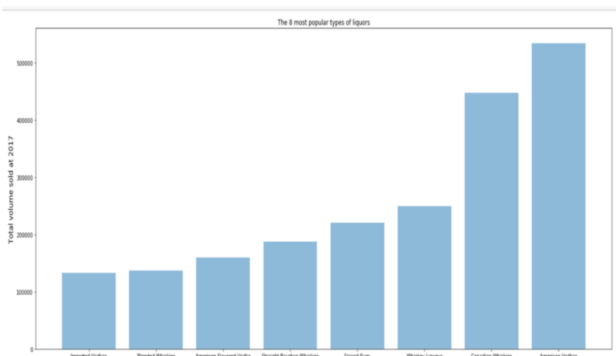


Figure 15

Figure 15: In this graph, we can see the most popular type of liquor is American Vodkas, the second best is Canadian Whiskies.

4.5.3.4 The consume distribution for each county

We built the graph by scanning the entire data set, before the scanning, I built a dictionary, the keys are the county numbers, and the values are the total volume consumption for each county at year 2017, during the scanning of the data set.

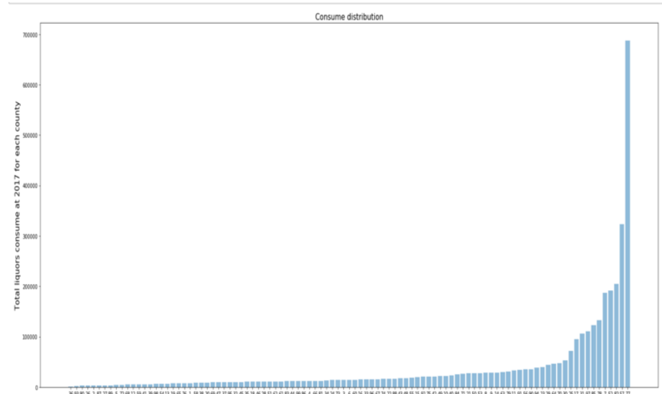


Figure 16

Figure 16: This graph shows the total liquor volume consumption at year 2017, the values on y-axis is the corresponding total volume consumption to each category. We can see that the county 77 has really high liquor consumption compare to other counties, one possible reason could be the large population compare to other counties, the other reason could be the people in this county really like to drink liquors.

4.5.3.5 Sales distribution of the first half year

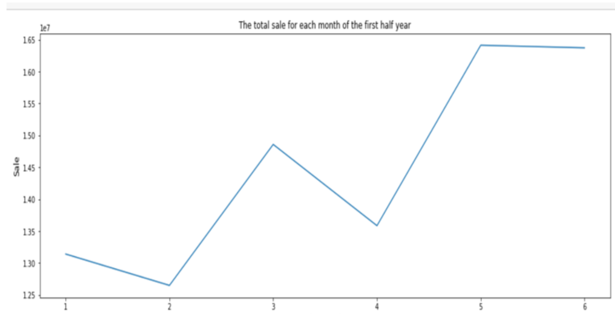


Figure 17

Figure 17: This is the broken line graph of the total sale for each month of the first half year

From Figure 17, we can easily see that the sales increased by month. One of our guess is that from January to June, the temperature went up and people may drink more alcoholic beverage when the weather become warmer and warmer.

4.6 Tools

4.6.1 Pandas

4.6.2 Numpy

4.6.3 Python

4.6.4 Matplotlib

4.6.5 JupyterNotebook

4.6.6 Overleef

5 KEY RESULTS

5.1 Information gained

5.1.1

Q: As for each vendor, which regions gains the maximum revenue in 2017?

A: As for Diageo America Company, the county POLK brought the maximum revenue in 2017.

5.1.2

Q: Which category of alcoholic beverages is the most popular in Iowa in 2017?

Q: Which category of alcoholic beverages gains the maximum revenue?

A: The American Vodkas is the most popular category of alcoholic beverages in 2017, it gains the maximum volume sold and maximum revenue.

5.1.3

Q: Which brand of alcoholic beverages gains the maximum revenue in 2017?

A: The Brand which gains the maximum revenue in 2017 is the Black Velvet. Which is a Canadian Whiskies.

5.1.4

Q: People in which regions consumed the most alcoholic beverages?

A: The Whisker Liqueur is the most popular category of alcoholic beverages in county POTTAWATTA in 2017.

5.1.5

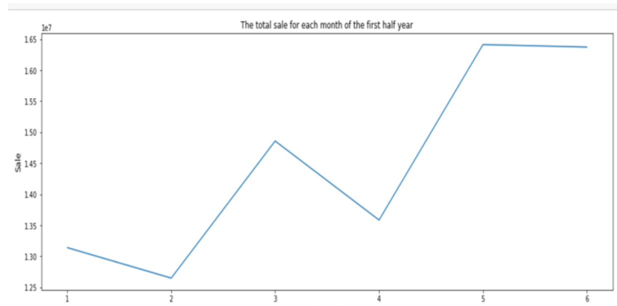
Q: People in which regions consumed the most alcoholic beverages?

A: People in POLK consumed the most alcoholic beverages in 2017.

5.1.6

Q: How the alcoholic beverage sale situation distributes in the first half year and what it means?

A:



What we know from the sales distribution is that people drink more alcoholic beverage when the weather become warmer and warmer.

5.1.7

The American Vodkas and Canadian Whiskies have the pretty same volume sold in 2017, but they have huge difference on revenue.

Which means **high Volume sold doesn't mean high revenue.**

5.2 Results evaluation

5.2.1 Compare to previous work

Since we use bar chat diagram to analysis Iowa Liquor Sales in Dollars by Year, which is similar to previous work (2), and we also use bar chart diagram to analysis Iowa Liquor Sales in Gallons by Year, which is similar to previous work (3). We can compare the graphic to previous work (2) and (3) to check if we are right.

5.2.2 Compare to intra-dataset mining result

For the question 'People in which region drink more alcoholic beverage?', we use two different ways to get the final result. We can compare these two results to check if they are matched. This can proof our results are correct if they are mostly matched.

6 APPLICATIONS

6.1 On venders' side

- Let vendor know people in which regions consumed more alcoholic beverages then vender can increase the supply for those regions.
- Let vendor know which regions brought less revenue and then the vender can increase the promotion and publicity to attract more users in those regions.
- Let vendor know which kinds of alcoholic beverages can brought more revenue, then the vender can extend the production for those kinds of alcoholic beverages.
- Let vendor know which alcoholic beverages gains more volume sold and then the vender can make a good market planning to get more profits.
- Let vendor know people drink more in summer days, thus they can prepare for the huge consumption in advance.

6.2 On consumers' side

We can let customer know which alcoholic beverage is the most popular to help them to make a right choose.

What's more we may predict the reason why people like or do not like a specific kind of alcoholic beverage by doing surveys which are combined with the knowledge we gained from this project.

For example, we can make online surveys in the county which gains less profit on a specific kind of alcoholic beverage. And ask people for what reasons they are not willing to consume this kind of alcoholic beverage, such as category, flavor, alcoholicity, taste flavor, liquid color, yeast, manufacture methods, etc.

After this, we can gain some other information by mining the online survey results, and find out why people do not like that kind of alcoholic beverage, and then we can send it to the beverage producers. Thus, they can make some improvement.

6.3 Some other interesting applications

- Use the Bayesian classification to predict the missing value (such as those alcoholic beverages which has no category label)
- Use the Bayesian classification to classify the category attribute after 2017(like 2018)

REFERENCES

[1] Iowa Liquor Sales in Dollars
https://data.iowa.gov/Economy/Iowa-Liquor-Sales-in-Dollars/8epw-u33y

[2] Iowa Liquor Sales in Dollars by Year
https://data.iowa.gov/Economy/Iowa-Liquor-Sales-in-Dollars-by-Year/wwyw-7at4

[3] Iowa Liquor Sales in Gallons by Year
https://data.iowa.gov/Economy/Iowa-Liquor-Sales-in-Gallons-by-Year/7uuv-irpi

[4] Total Liquor Sales in Iowa by Month
https://data.iowa.gov/Economy/Total-Liquor-Sales-in-Iowa-by-Month/xiyh-fbvw

[5] Iowa Liquor Sales by Year and County
https://data.iowa.gov/Economy/Iowa-Liquor-Sales-by-Year-and-County/ahiv-u4uz

[6] Kaggle data set of Iowa Liquor sales

https://www.kaggle.com/residentmario/iowa-liquor-sales/data

[7] Iowa Liquor Sales
https://data.iowa.gov/Economy/Iowa-Liquor-Sales/m3tr-qhgy