Iowa Liquor Sales

Khatereh Mohajery

Scenario 1: State tax board

Dec 16, 2016

Summary

- To be presented to Iowa state tax board.
- Goal:

  Summarizes the current class E liquor sales in the Iowa state and the projections of the sale for the rest of the year of 2016 .

- Includes:
  - The source of the data
  - Steps and assumptions in exploring, processing and mining the dataset
  - The methods and models used for the projections of the sale in 2016
  - The results of the models

The Data

- The original data from Iowa.gov
- This data in csv format contains only 10% of the available data
- It is assumed that this 10% was collected randomly and therefore can represent the whole dataset.
- Each entry:
    - Information on a single transaction between a liquor store and the state vendor
    - Store information (location,zip code, ..)
    - Amount, type and value of the liquor
    - Vendor information and dates

Initial Cleaning Steps

- Handling missing values in Category_Name, County (County_Number)
- Removing $ sign, change to numerical values for calculation purposes
- Parsing date column

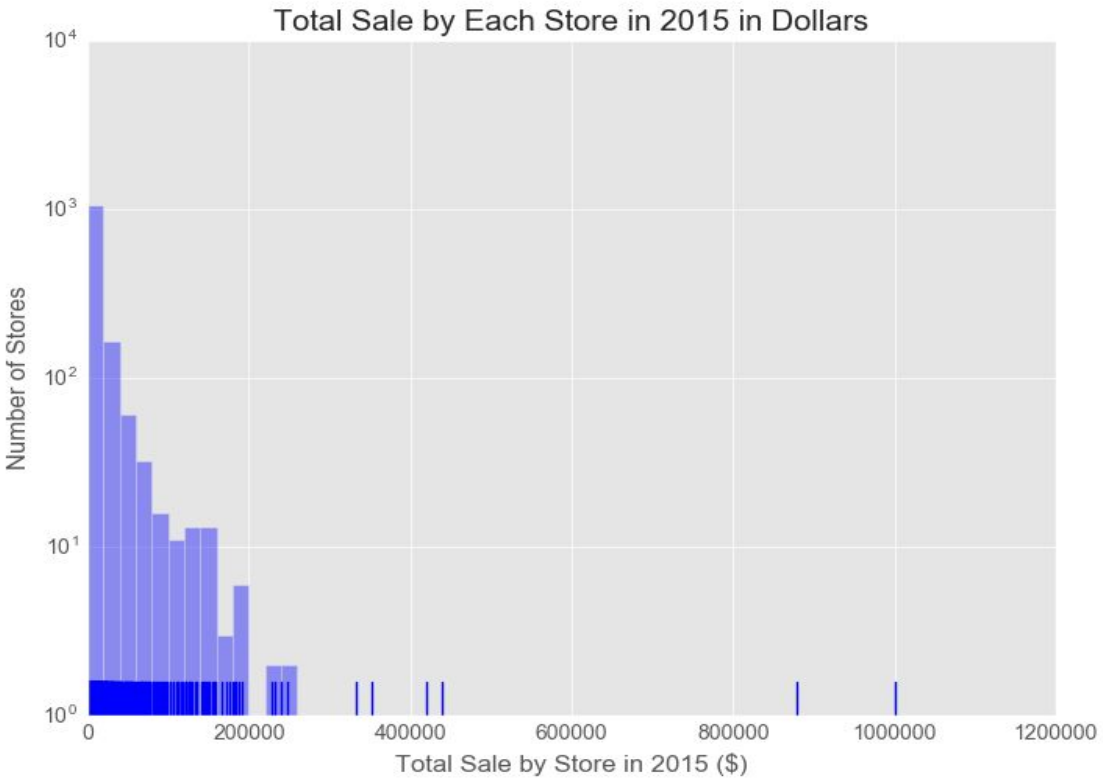
- Liquors were assigned to 9 general categories

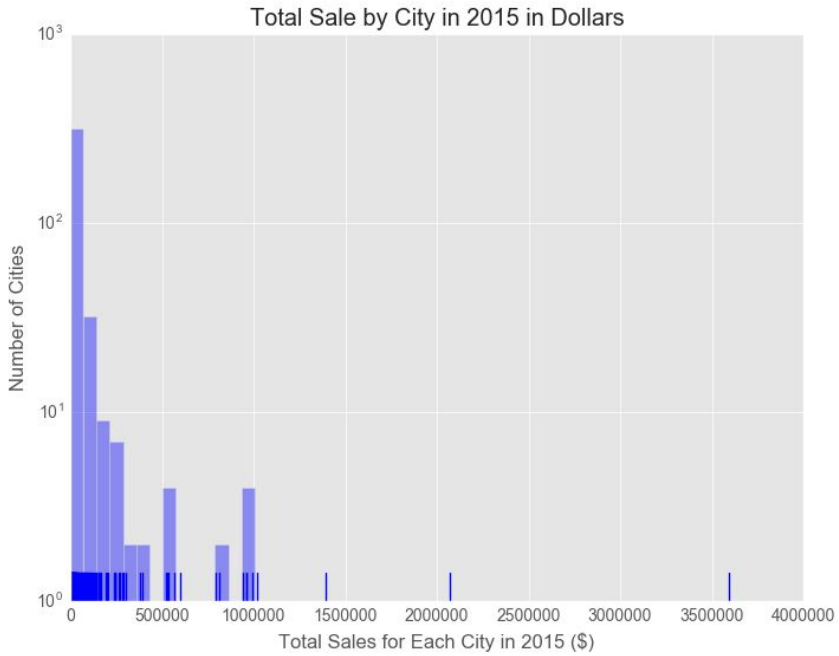
- A new column for projected profit based on state bottle cost and the retail price of the same bottle.

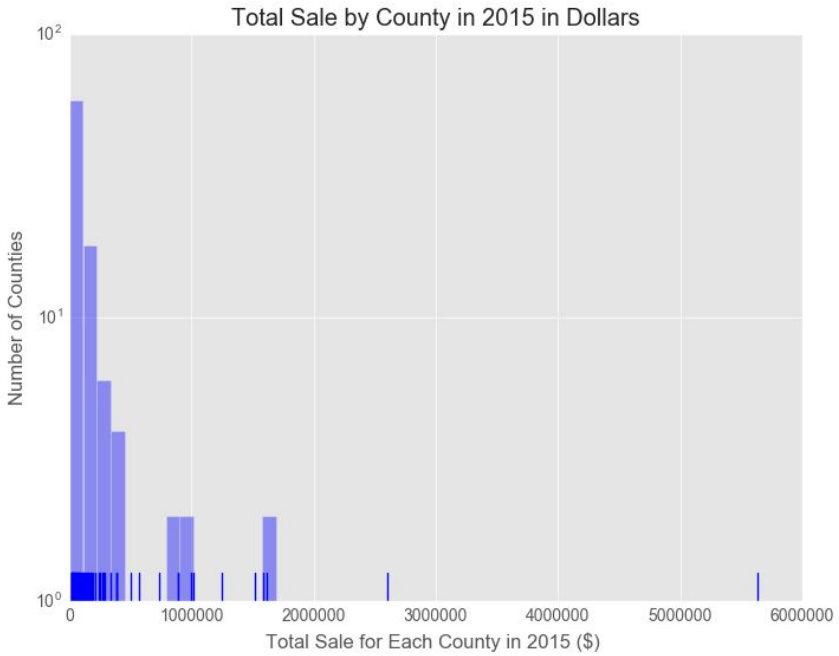
- Invalid Zip_Code: 712-2

Total Sales

- 2,174,546 bottles
- 1,985,754.2 liters
- retail sale value of $28,516,695.5
- Average price of $13.1 per bottle.
- The average sales of stores in 2015 was $20,784.77
- High sales : Central City Liquors and Hy-Vee number #3 in Des Moines city



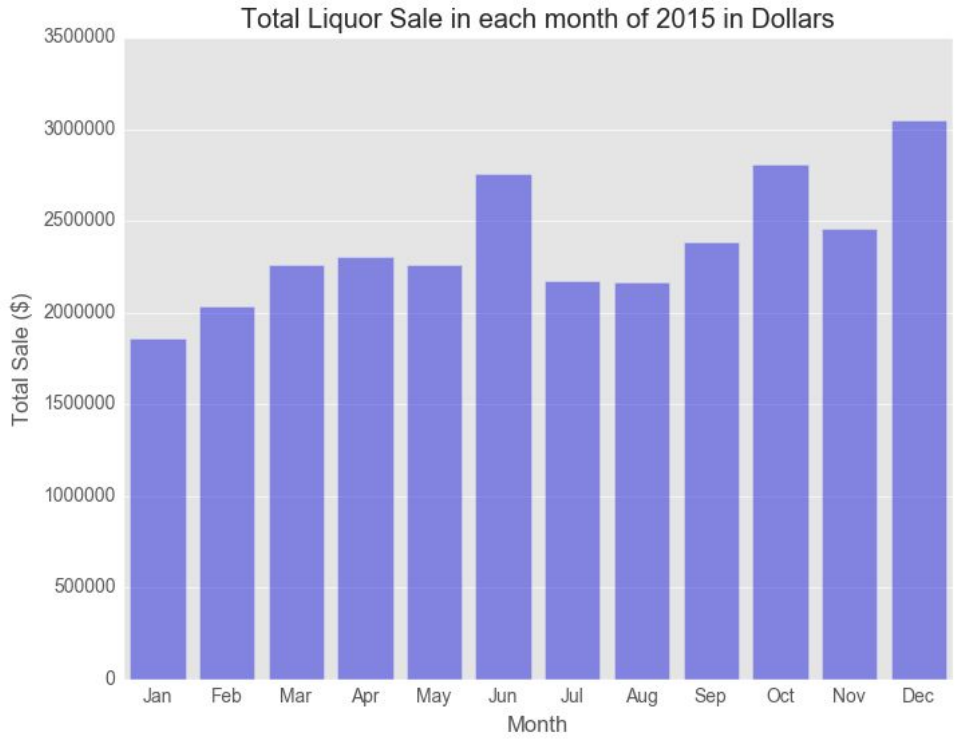




Highest sales in 2015: City of Des Moines and Polk County

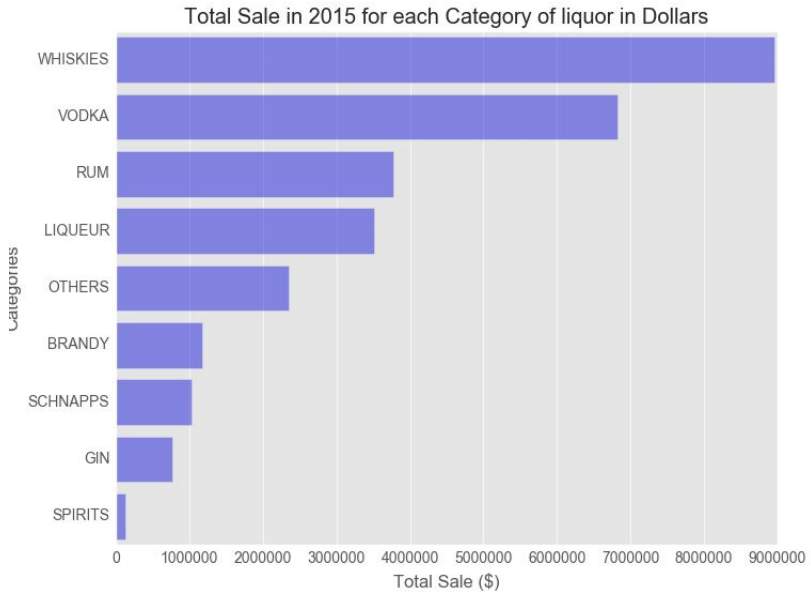
Cities average sales:$75,044

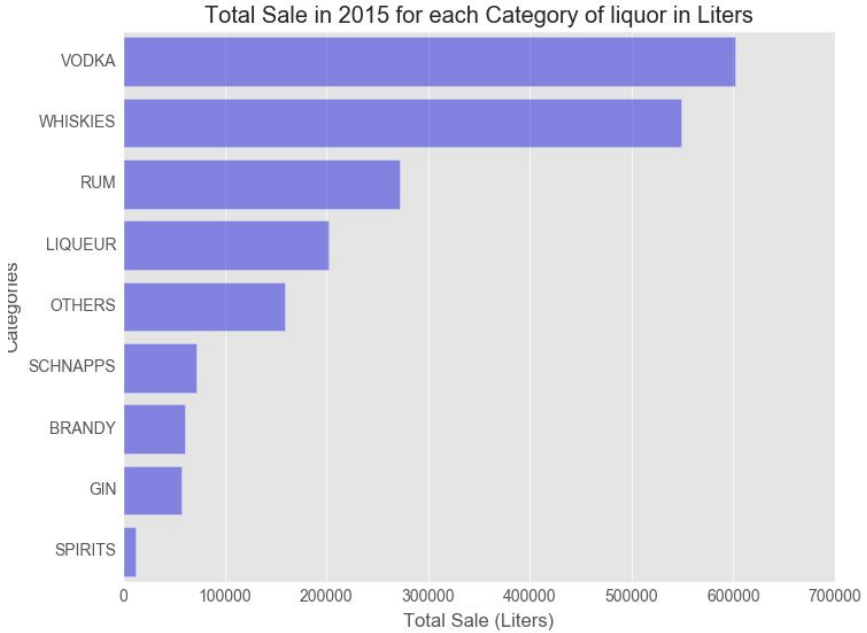
Counties average sales:$288047.4



It is interesting that the first three months of the year have below average sales i.e. $2,376,391

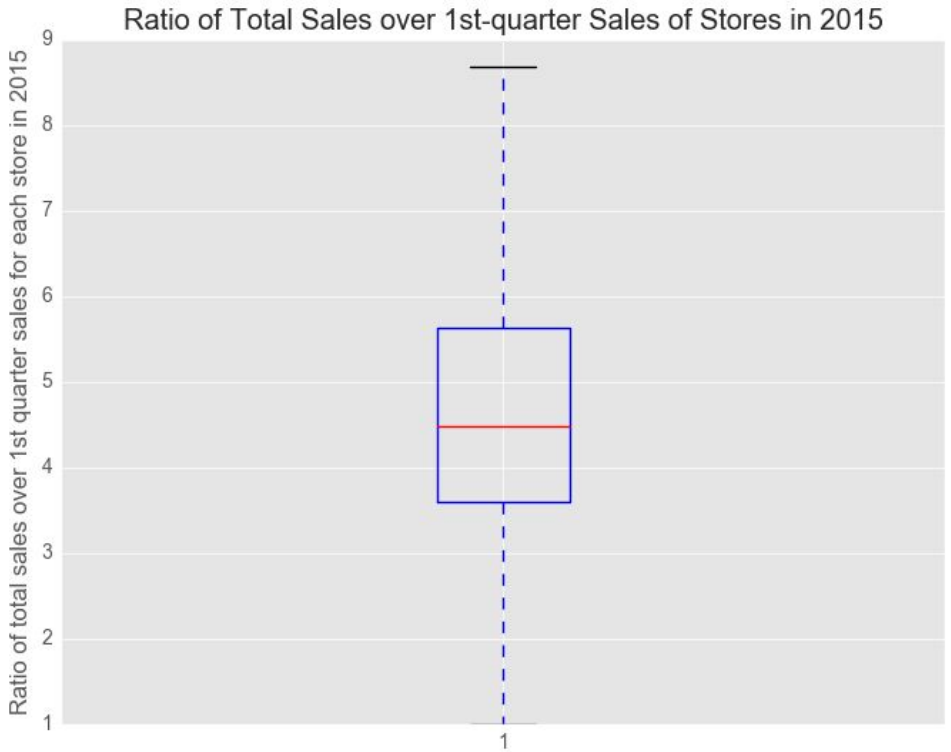
While Vodka category has the highest volume of sales, Whiskies have higher sale values.





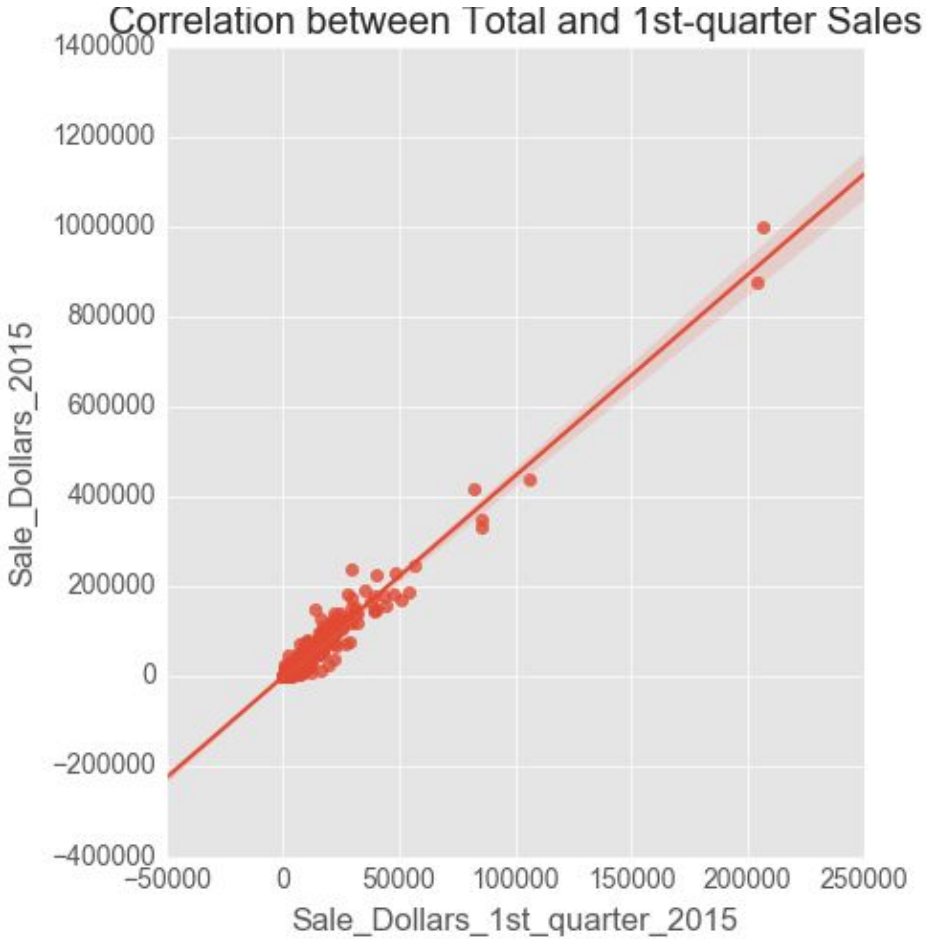
Mine and Refine the Data

- Creating a column for 1st-Quarter sales
- Separating 2015 and 2016
- Group by stores
- Sales vs 1st-quarter sales for each store
- Few stores with very high ratio New stores maybe!!



- There is very strong correlation between sales in the 1st-quarter and the annual sales

Correlation matrix	1st quarter	annual
1st quarter	1.0	0.98144
annual	0.98144	1.0



First 3 Models

Model 1: Using linear regression with train and test split gave R^2 score of 0.945.

Model 2: Using linear regression with train and test split and 5 fold cross validation gave the same R^2 score of 0.945.

Model 3: Using linear regression with no train and test split gave R^2 score of 0.963.

formula of : Total sale =282.065 + 4.4688 * 1st_quarter sale

Dep. Variable:	Sale_Dollars_2015	**R-squared:**	0.963
Model:	OLS	**Adj. R-squared:**	0.963
Method:	Least Squares	**F-statistic:**	3.300e+04
No. Observations:	1262	**Prob (F-statistic):**	0.00
Df Residuals:	1260	**Log-Likelihood:**	-13428.
Df Model:	1	**AIC:**	2.686e+04
Covariance Type:	nonrobust	**BIC:**	2.687e+04

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	282.06	308.993	0.913	0.361	-324.132 888.262
1st quarter	4.4688	0.025	181.662	0.000	4.421 4.517

Models -Outliers Removed

Model 4: Using linear regression with train and test split gave R^2 score of 0.952.

Model 5: Using linear regression with train and test split and 5 fold cross validation gave the same score of 0.952.

Model 6: Using linear regression with no train and test split gave R^2 score of 0.933

Total sale = 698.5949 + 4.4688*1st_quarter sale

Dep. Variable:	Sale_Dollars_2015	**R-squared:**	0.933
Model:	OLS	**Adj. R-squared:**	0.933
Method:	Least Squares	**F-statistic:**	1.752e+04
No. Observations:	1260	**Prob (F-statistic):**	0.00
Df Residuals:	1258	**Log-Likelihood:**	-13369.
Df Model:	1	**AIC:**	2.674e+04
Covariance Type:	nonrobust	**BIC:**	2.675e+04

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	698.59	314.99	2.218	0.027	80.627 1316.56
1st quarter	4.3699	0.033	132.3	0.000	4.305 4.435

Final Results

All models predicted to have around 4% increase in sales in 2016 compare to 2015 based on the sales in 1st quarter of the 2016.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Ratio of predicted total sale in 2016 to total sale in 2015	1.037	1.037	1.040	1.040	1.039	1.059

Steps to Improve the Model

- Using Lasso regularization and counties as dummy variables for the linear regression

- Replacing the value of 1st-quarter sales for the stores with high ratio of total sales to 1st-quarter with the median value