Data621 Final Project

Jim Lung

May 21, 2018

Iowa Liquor Sales

Abstract:

The goal of this project is to determine a geographic location (county) in Iowa that will yield the highest amount of liquor sales. Once we have the location specified according to highest amount of liquor sales, we can look and see if there is anything specifically that helps total volume sold. ((coupled with number of current liquor stores per area, can offer recommendations on where to open a new liquor store)). High sales and low numbers of existing stores will most likely lead to higher profits for the new liquor store, on average.

The criteria for success would be determining the inventory forecast by Bottles sold predication in which the company could plan and determine which factors effect total sales and how. Using the correlation and forward stepwise regression with linear method are applied to perform predication the bottles sold. The dataset used consisted of data regarding sales of liquor from different stores in different counties within the state of Iowa.

In order to obtain the specific trend and predication, the dataset was subset into the highest volume sold location at DES MOINES City in 2017. Before outputting the models, the influential points were all removed. For each target variable, two models were rendered. All the three models, the variables that showed significance were Whiskies and distillery Whiskies. Although the country and city highest volume sold are also separately Polk country and Des Moines, the volume sold in Gallons are definitely decreased comparing between 2016 and 2017.

Key words:

- Liquor Sales,
- Naive Forecast,
- Linear Regression
- Inventory Forecast
- Prediction

Introduction:

The objective of this report is to create a statistical model for the number of bottles sold of whiskey which is within the state of Iowa. This can help us make informed decisions on inventory prediction, sales, and assist wholesale distributors to plan for the predicted volume of distribution. To perform exploratory analysis with visualizations and statistical analysis, this is a large dataset and that is a great thing. When loading the full dataset, there will be upwards of 2.7 million observations. We removed the 2,973 duplicated columns from the dataframe, as well as the all of the null values. Because of our large number of observations, this should have very little effect on our analysis. For our location data, we can see that there are 100 county numbers, 99 counties, 383 cities, and 676 zip codes. It would be wise to cross-reference this data with the state's municipality records to make sure the location variables are properly matched across city, county, and zip code.

We see that a large number of observations are found in Polk County, the city of Des Moines, and the zip code 50010 (Ames, Iowa). Ames is the home of Iowa State. This makes logical sense because these are the main urban centers in the state of Iowa and a larger number of people should correlate positively with a higher number of liquor sales. We have 72 different categories of alcohol. These are highly differentiated. If we were to analyze the categories further, it may be wise to group in broader categories. For example, all whiskeys and bourbons could be in one category, all vodkas in another, etc. There are 1400 unique stores in the data set. The vast majority of sales are of quantities of less than 100 bottles and of transactions less than $1,000. This establishes that spirit sales in the Unites States is a valuable market worth exploring for a more detailed and statistical understanding of sales and volume.

We hope to more thoroughly understand what impact specific store sights may have accounting for the seasonal impact in effect liquor sales. We set up the range of our analysis to the City of Des Moines in 2017. The part of the year has an decreasing tread in sales in total capacity of alcohol, so the time of interest for this analysis will be in 2017.

Background

The main goal that has to be achieved in inventory prediction is increasing the efficiency without decreasing the service value offered to the customers. When managing the levels of inventory, it is important to maintain moderate level(s) - not too high and not too low. If the inventory level is excessive, business funds can get wasted. These funds would not be able to be used for any other purpose, thus involving an opportunity cost. The costs of shortage, handling insurance, recording and inspection would proportionately increase along with inventory volume, thus impairing profitability.

Literature review:

Book: An Overview of Forecasting Methodology

Author: David S. Walonick (1993)

Trend extrapolation - These methods examine trends and cycles in historical data, and then use mathematical techniques to extrapolate to the future. The assumption of all these techniques is that the forces responsible for creating the past, will continue to operate in the future. This is often a valid assumption when forecasting short term horizons, but it falls short when creating medium and long term forecasts. The further out we attempt to forecast, the less certain we become of the forecast.

The most common mathematical models involve various forms of weighted smoothing methods. Another type of model is known as decomposition. This technique mathematically separates the historical data into trend, seasonal and random components. A process known as a “turning point analysis” is used to produce forecasts. ARIMA models such as adaptive filtering and Box-Jenkins analysis constitute a third class of mathematical model, while simple linear regression and curve fitting is a fourth. Makridakis (one of the gurus of quantitative forecasting) correctly points out that judgmental forecasting is

superior to mathematical models, however, there are many forecasting applications where computer generated forecasts are more feasible. For example, large manufacturing companies often forecast inventory levels for thousands of items each month. It would simply not be feasible to use judgmental forecasting in this kind of application.

Consider Timing - When planning to effectively forecast your inventory levels it is important to consider both the life cycle of your products as well as how far in the future your forecast must reach. When it comes to the life cycle of your products, understanding how much stock to keep of certain items to avoid waste is very important. For example, assuming that your most expensive items are also your mist profitable may not always be the case and could actually prevent cost flow when lower costing items have a higher turnover rate. Understanding how to balance inventory levels when forecasting can be beneficial to your budget. Knowing just how your products and sales effects your business is key for any businesses success.

Experimentation and Results

Data Exploration

The data set contains the spirits purchase information of Iowa Class "E" liquor licensees by product and date of purchase from January 2013 to December 2017. The data set is provided by the Iowa Department of Commerce, Alcoholic Beverages Division, click here to view the data set at Data.Iowa.Gov. As previously discussed, the data set is 3.3 GB in total size and much to large to use in a meaningful model.

Achieve the dataset from:

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

Month	Year	City	Category Name	County	Bottles Sold	Sale (Dollars)	Bottle Volume (ml)	State Bottle Cost	State Bottle Retail	Volume Sold (Gallons)
1	2013	ACKLEY	BLENDED WHISKIES	HARDIN	12	117.48	750.000	6.53	9.79	2.38
1	2013	ACKLEY	BLENDED WHISKIES	WEBSTER	9	104.37	1750.000	15.33	22.99	4.16
1	2013	ACKLEY	CANADIAN WHISKIES	HARDIN	38	487.50	1030.000	37.69	56.52	13.71

1	2013	ACKLEY	CANADIAN WHISKIES	WEBSTER	15	193.53	1416.667	25.99	39.00	5.35
1	2013	ACKLEY	STRAIGHT BOURBON WHISKIES	HARDIN	12	170.52	750.000	9.22	14.21	2.38
1	2013	ACKLEY	STRAIGHT BOURBON WHISKIES	WEBSTER	6	80.28	1750.000	8.92	13.38	2.77

Data Preparation

Ensuring Correct Calculations:

- Ensuring bottle size (e.g., 750 ml) x bottles sold = volume liters sold
- Ensuring bottle retail value x bottles sold = sale dollars
- I found no problems with the math, but it was good to check all the same

Type of dataset preparation

VAR	TYPE
Month	Integer
Year	Integer
City	Char
CategoryName	Char
County	Char
Bottles Sold	double
Sale (Dollars)	double
Bottle Volume (ml)	Integer
State Bottle Cost	Integer
State Bottle Retail	Integer
Volume Sold (Gallons)	Integer

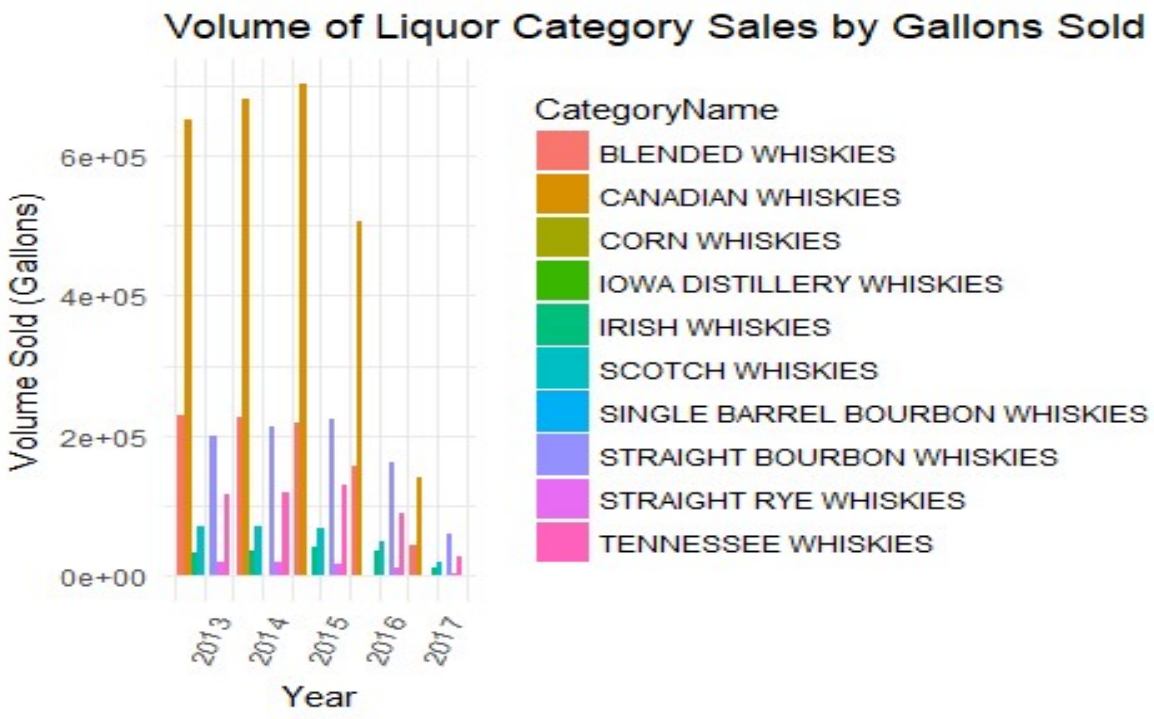
```
# comparing between volume sold and category name
##    Year CategoryName                  `Volume Sold (Gallons)`
##   <dbl> <fct>                                           <dbl>
## 1  2013 BLENDED WHISKIES                               228340
## 2  2013 CANADIAN WHISKIES                              651162
## 3  2013 IRISH WHISKIES                                  32933
## 4  2013 SCOTCH WHISKIES                                 71387
## 5  2013 SINGLE BARREL BOURBON WHISKIES                   1040
## 6  2013 STRAIGHT BOURBON WHISKIES                      199607


# comparing between volume sold and category name at 2017
##    Year CategoryName             `Volume Sold (Gallons)`
##   <dbl> <fct>                                      <dbl>
## 1  2017 BLENDED WHISKIES                           43696
## 2  2017 CANADIAN WHISKIES                         141743
## 3  2017 CORN WHISKIES                                808
## 4  2017 IOWA DISTILLERY WHISKIES                      97
## 5  2017 IRISH WHISKIES                             12381
## 6  2017 SCOTCH WHISKIES                            20176


# comparing between volume sold and county at 2017
##    Year County      `Volume Sold (Gallons)`
##   <dbl> <chr>                        <dbl>
## 1  2013 BLACK HAWK                   67295
## 2  2013 CERRO GORDO                  31422
## 3  2013 DUBUQUE                      41406
## 4  2013 JOHNSON                      62294
## 5  2013 LINN                        101739
## 6  2013 POLK                        218528


# comparing between volume sold and county at 2017
##    Year CategoryName                   `Sale (Dollars)`
##   <dbl> <fct>                                    <dbl>
## 1  2013 BLENDED WHISKIES                     8082719
## 2  2013 CANADIAN WHISKIES                   30783180
## 3  2013 IRISH WHISKIES                       3457948
## 4  2013 SCOTCH WHISKIES                      5046236
## 5  2013 SINGLE BARREL BOURBON WHISKIES        140029
## 6  2013 STRAIGHT BOURBON WHISKIES           12748832
```

Below is a plot of the distribution of counts for the Volume Sold (Gallons) variable.

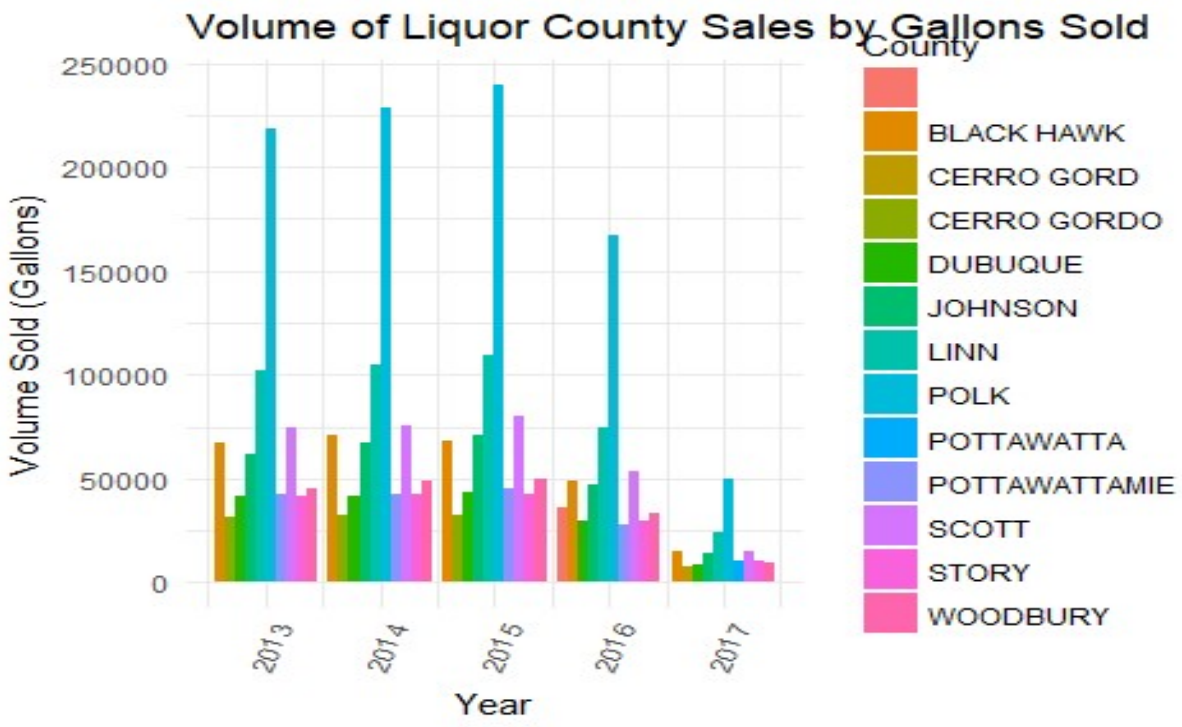
We reviewed the liquor sales by gallons sold per year by Liquor Category. Initially, we viewed the top 5 Liquor Categories by volume sold but there were large disparities

between years, suggesting that the top 5 change often and is likely due to changing consumer tastes. We do see a more stable set of liquor categories for the top 10 category which suggests that while tastes may change we don't see large movements in liquor categories at this level.

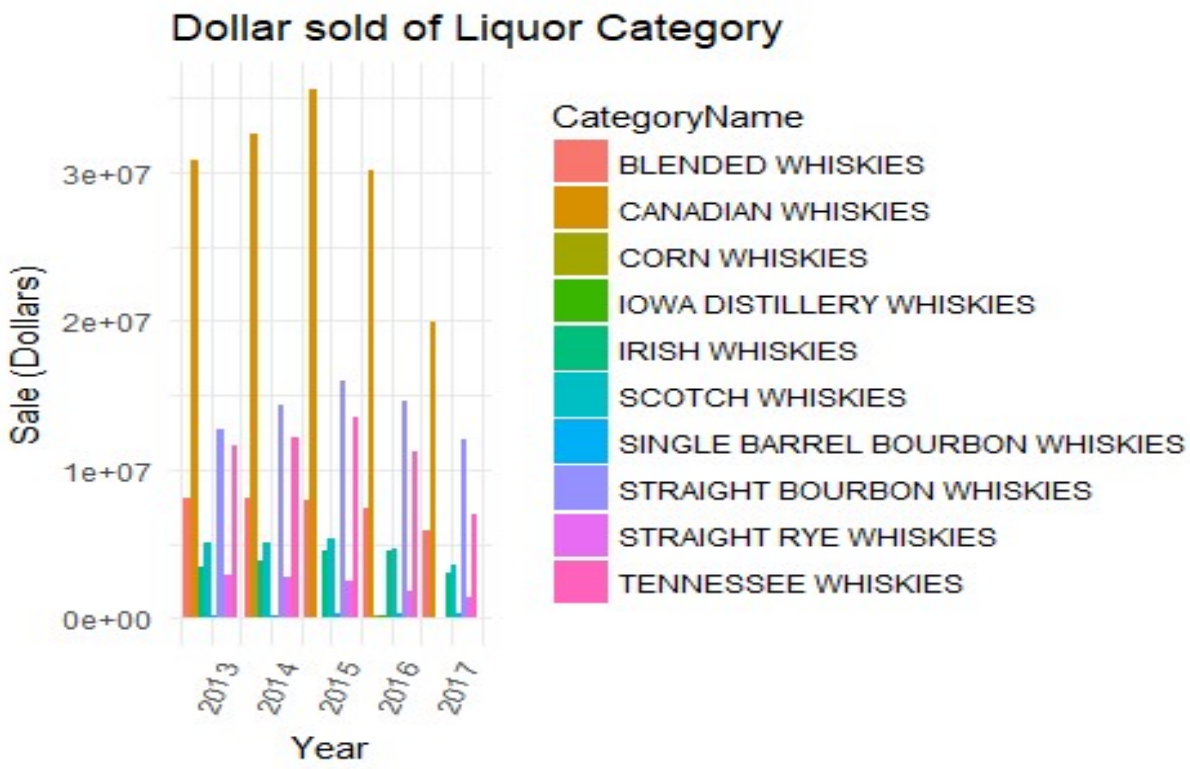
We can further see that the canadian whiskies are the highest volume sold from 2012 to 2016, but the number of straight bourbon whiskies is the highest in 2017.



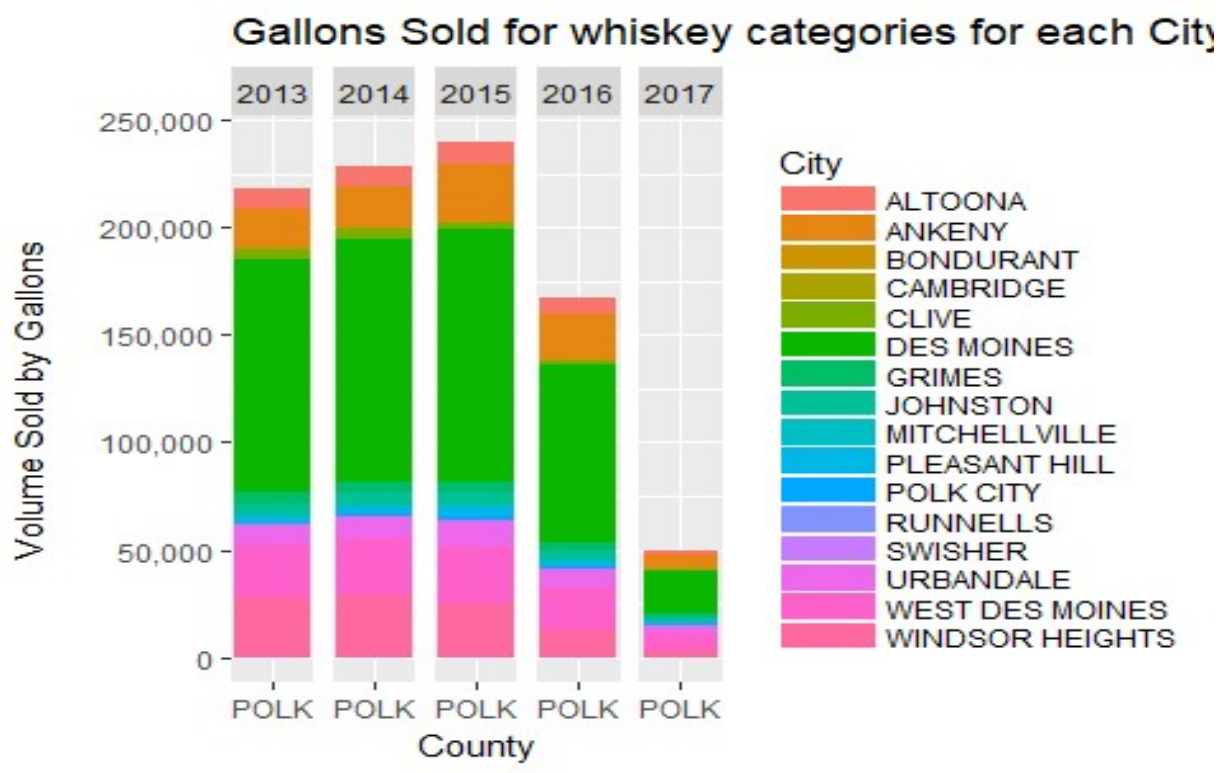
The highest sold for volume of Liquor is located at Polk County.



The dollar sold of liquor category is same proportation with volume sold of liquor gallon, the canadian whiskies are the highest volume sold from 2012 to 2016, but the number of straight bourbon whiskies is the highest in 2017.



We can see that Des Moines accounts for a significant portion of the liquor sales in Polk County. We will focus our analysis on Polk County.



Build Models

Bottles Sold Model

We used forward selection method for our initial model for the Bottles Sold. However, we expect some high degrees of multicollinearity as some of our variables can be easily explained by other variables in the data set. We see a very high degree of multicollinearity in our independent variables for Bottles Sold and with good reason. If more bottles sold then certainly the volume sold by gallons would increase as would the sale dollars, we therefore removed volume sold by gallons. Below is the table that highlights the high levels of multicollinearity for Volume Sold by Gallons and Sale Dollars.

```
# Build model by selection the highest volume location at DES MOINES City
```

Month	Year	City	CategoryName	Bottles Sold	Sale (Dollars)
1	2017	DES MOINES	BLENDED WHISKIES	990	35100.55
1	2017	DES MOINES	CANADIAN WHISKIES	2933	112301.42
1	2017	DES MOINES	CORN WHISKIES	10	948.75
1	2017	DES MOINES	IOWA DISTILLERY	8	716.73

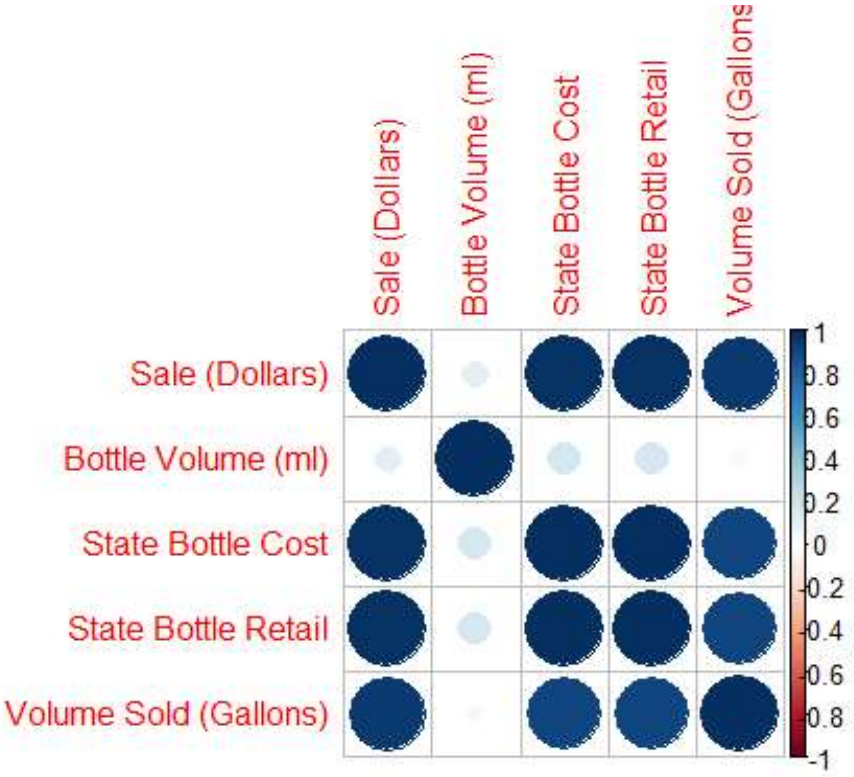
			WHISKIES		
1	2017	DES MOINES	IRISH WHISKIES	574	31019.63
1	2017	DES MOINES	SCOTCH WHISKIES	420	35540.77

Bottle Volume (ml)	State Bottle Cost	State Bottle Retail	Volume Sold (Gallons)
912.5701	3333.05	5000.55	196.99
802.6042	11247.43	16925.69	552.23
750.0000	157.50	236.25	1.91
750.0000	136.50	204.78	1.54
786.3014	3562.69	5344.42	116.86
981.5647	4906.90	7364.12	99.03

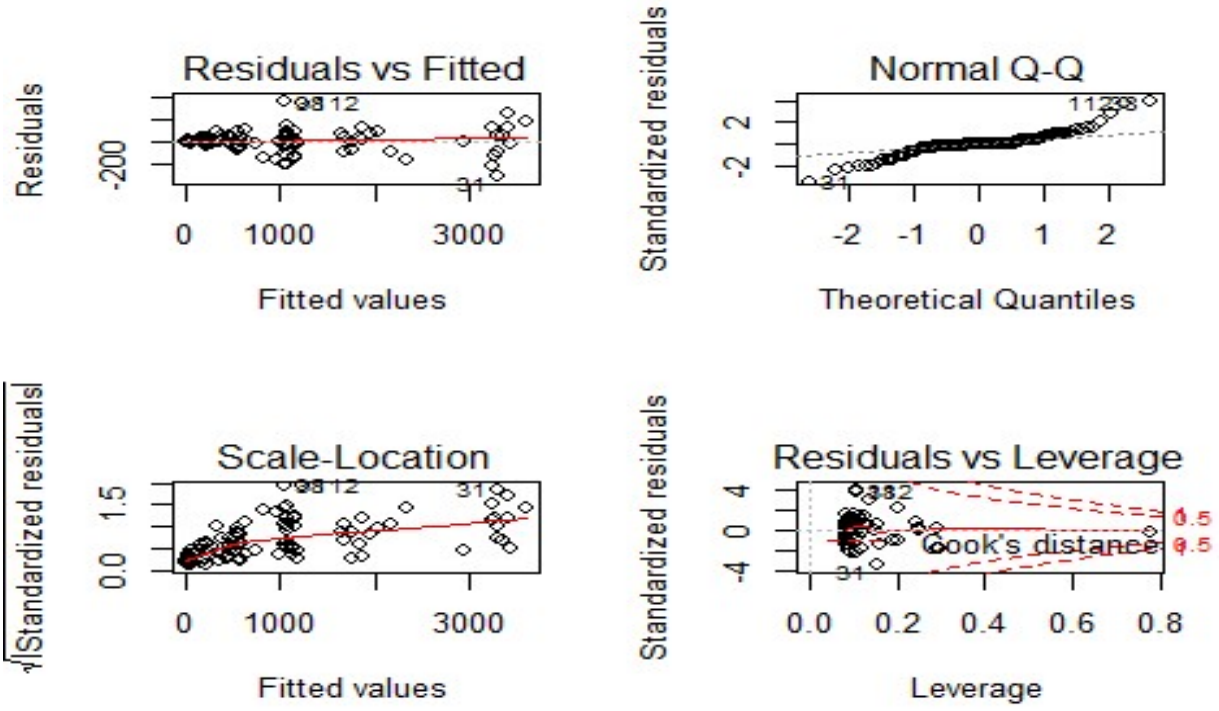
Correlation (Model 1)

** Check multicollinearity See correlation between all variables and keep only one of all highly correlated pairs **

```
model1 <- lm(`Bottles Sold` ~ CategoryName + `Sale (Dollars)`+ `Bottle Volume
(ml)`+ `State Bottle Cost`+ `State Bottle Retail`+ `Volume Sold (Gallons)`,
data=iowa_data_reduced2)
```



```
# VIF for an X variable should be less than 4 in order to be accepted as not
causing multi-collinearity. The cutoff is kept as low as 2
model1 <- lm(`Bottles Sold` ~ CategoryName + `Sale (Dollars)`+ `State Bottle
Cost`+ `State Bottle Retail`, data=iowa_data_reduced2)
```



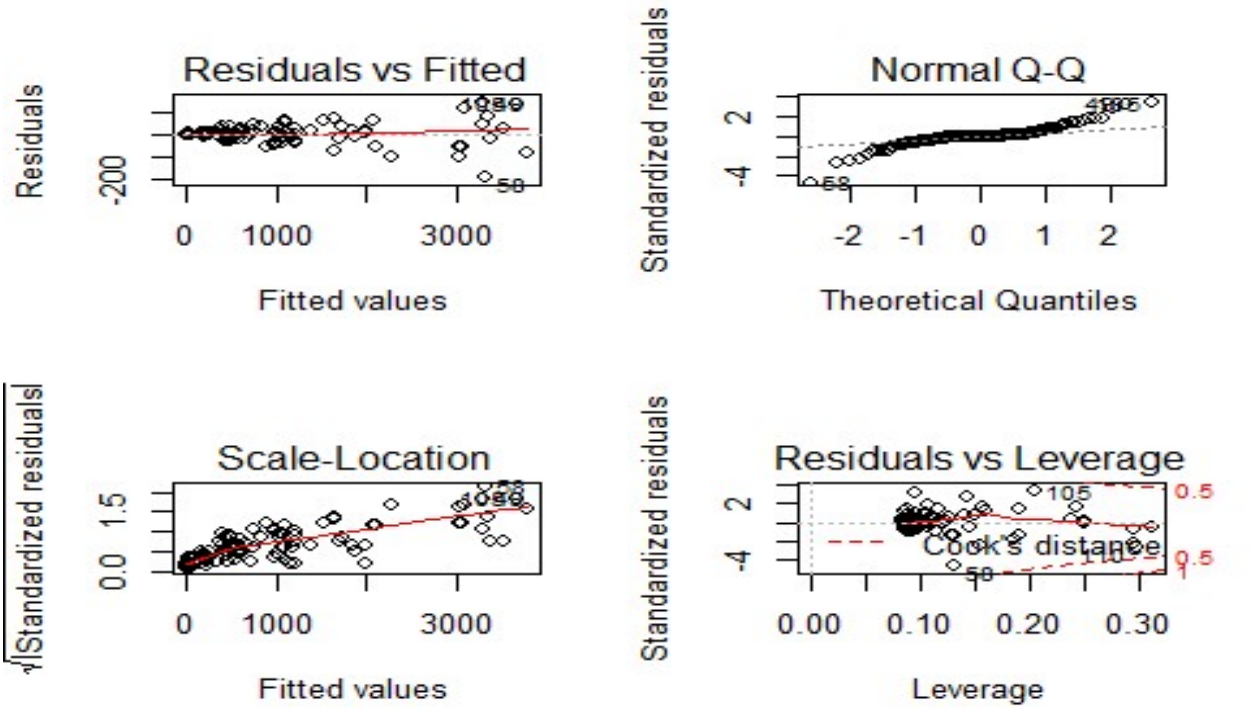
rn	GVIF	Df	GVIF^(1/(2*Df))	Adjusted_GVIF
CategoryName	255.8244	9	1.360738	1.851608e+00
Sale (Dollars)	277.4708	1	16.657454	2.774708e+02
State Bottle Cost	1823143.9656	1	1350.238485	1.823144e+06
State Bottle Retail	1820815.5428	1	1349.375983	1.820816e+06

Forward regression (model 2)

It is using forward selection method for the initial model for the Bottles Sold. However, we expect some high degrees of multicollinearity as some of our variables can be easily explained by other variables in the data set. We see a very high degree of multicollinearity in our independent variables for Bottles Sold and with good reason.

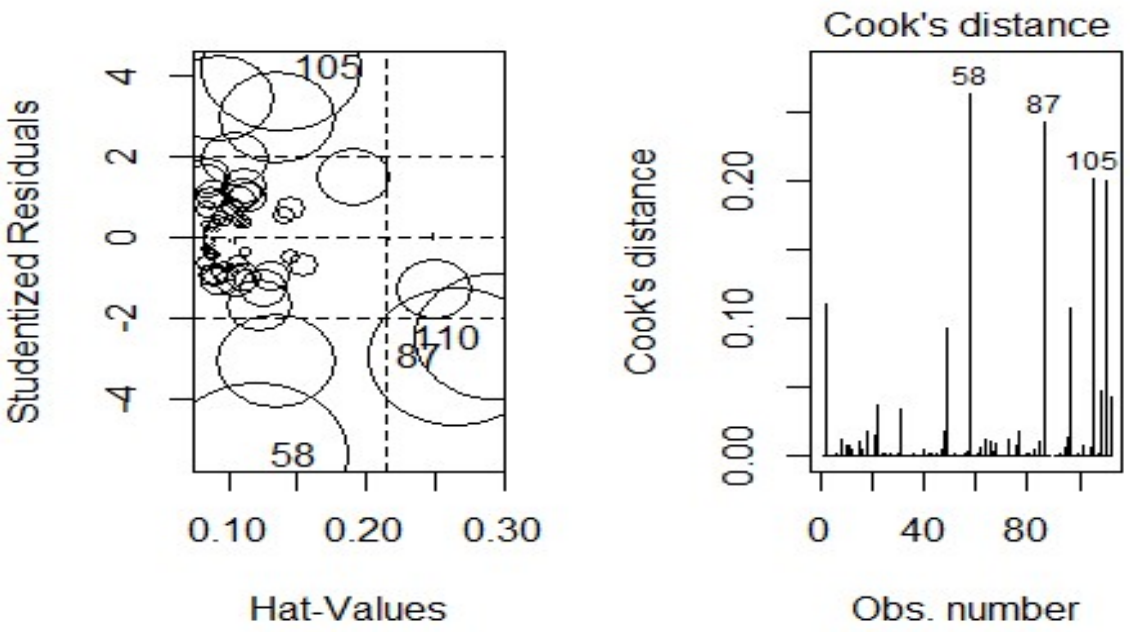
```
#Forward step
forward <- step(lm(`Bottles Sold`~1,data=iowa_data_reduced2),direction =
"forward",
                scope=~CategoryName +`Sale (Dollars)`+ `Bottle Volume (ml)`
```

```
              +`State Bottle Cost`+`State Bottle Retail`+`Volume Sold
(Gallons)`,trace = FALSE)
```



rn	GVIF	Df	GVIF^(1/(2*Df))	Adjusted_GVIF
Volume Sold (Gallons)	57.24044	1	7.565741	57.240438
CategoryName	546.51180	9	1.419348	2.014549
State Bottle Retail	242.41628	1	15.569723	242.416284
Sale (Dollars)	270.73867	1	16.454138	270.738670

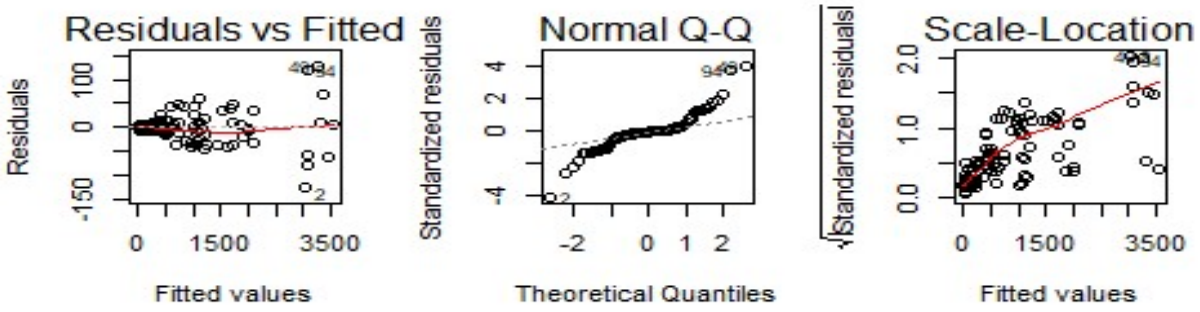
If more bottles sold then certainly the volume sold by gallons and sale dollars would increase, we therefore removed Volume Sold by Gallons and Sale Dollars. Below is the table that highlights the high levels of multicollinearity for Volume Sold by Gallons and Sale Dollars.

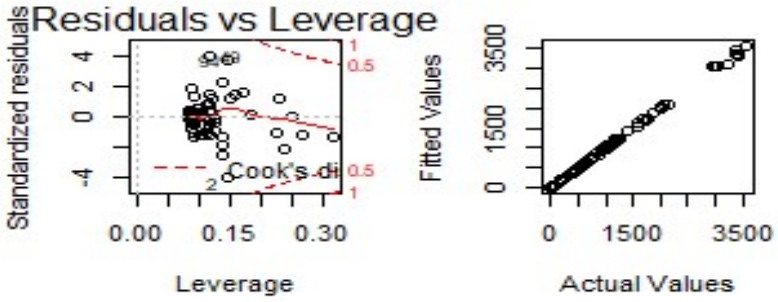


Model 2 remove influencePlot

Several values may have undue influence on the final form of our model. Using the influencePlot function from the car package and Cooks Distance plot, we can see which values that have the greatest impact on our model and we removed the observations indicated in the Cook's distance plot for 58, 87, 105 and 227.

Below are the diagnostic plots for our Model 1, without influential points. Unfortunately, we see a non-normal distribution in residuals of the qq plot and we see a linear relationship for the fitted and actual values plot.





Our model has an extremely good Adjusted R^2 at 0.99 but we see that the distribution of the residuals is not normally distributed and the fitted values plotted to the actual values do show a clearly linear relationship. We will need to further transform the variables in order to have a more normal distribution of our residuals.

Bottles Sold Model with Log Transformation (Model 3)

The adjusted model uses the same selection method of forward and keeps Bottles Sold as our dependent variable. And we use the BoxCox transformation method to transform our dependent variable. The resulting λ is 0 for transformation as log. By using this selection method and dependent variable transformation, the final model excludes the Sales Dollars variable. Additionally, we have high multicollinearity between the Bottle Cost and the Retail variables, because Bottle Cost has a high impact on Retail price. We remove the Bottle Cost variable as impacts Retail Price may have on our dependent variable.

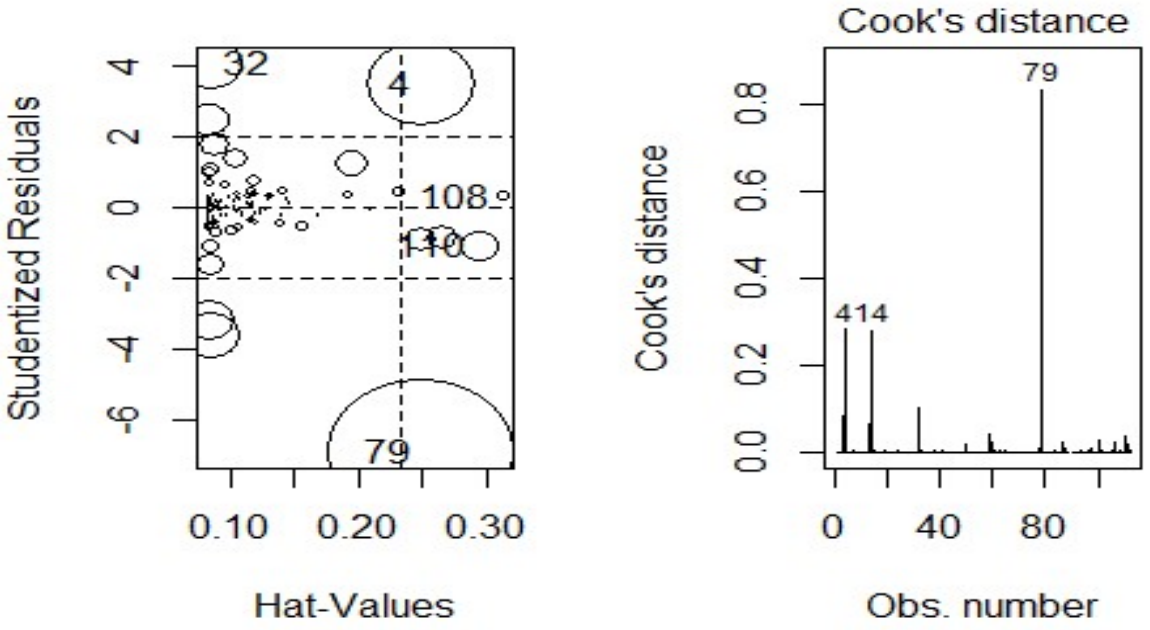
```
# forward regression with log transformation
model3 <- step(lm(log(`Bottles Sold`)~1,data=iowa_data_reduced2),direction =
"forward",
                 scope=~CategoryName +`Sale (Dollars)`+ `Bottle Volume (ml)`
               +`State Bottle Cost`+`State Bottle Retail`+`Volume Sold
(Gallons)`,trace = FALSE)
```

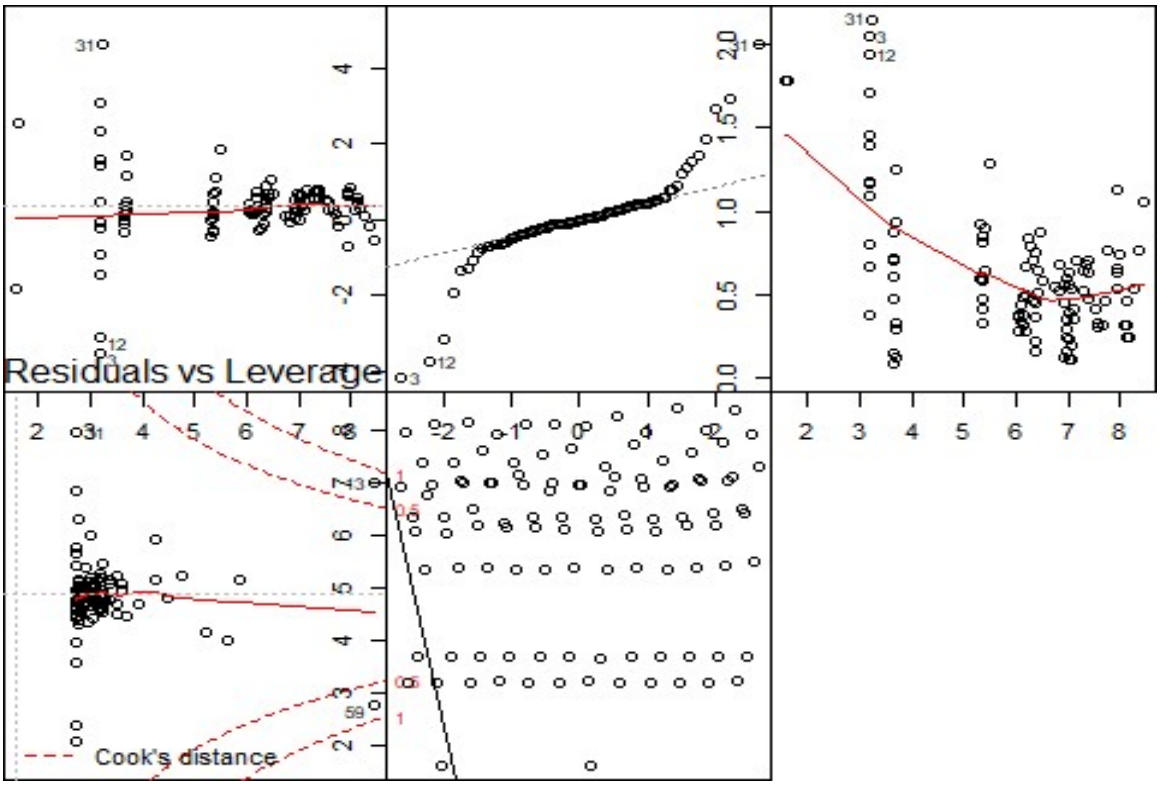
rn	GVIF	Df	GVIF^(1/(2*Df))	Adjusted_GVIF
State Bottle Cost	2.016193e+06	1	1419.926956	2.016193e+06
CategoryName	8.984032e+03	9	1.658202	2.749632e+00

State Bottle Retail	2.015624e+06	1	1419.726705	2.015624e+06
Volume Sold (Gallons)	6.517572e+01	1	8.073148	6.517572e+01
Bottle Volume (ml)	2.323615e+01	1	4.820388	2.323615e+01

We select the values that have the greatest influence on our model and remove them to improve the model performance. The observations removed in this model is 4, 32, 79. By excluding these values from our evaluation data set we are able to fit a more appropriate model.

```
## ----------------------------------------
##       StudRes     Hat      CookD
## --------- --------- --------- ----------
##   **4**     3.498     0.25      0.2818
##
##  **32**     4.05     0.08449    0.1008
##
##  **79**    -6.915    0.2501     0.833
##
##  **108**   0.3069     0.312    0.003316
##
##  **110**   -1.087     0.294    0.03778
## ----------------------------------------
##
## Table: Influential points in Log of Bottles Sold Model from influencePlot
```





we see a much more normal distribution of the residuals.

The residual distribution are more normal, it is almost remain unchange for log bottle sold when state bottle retail change.

We would expect a negative correlation with bottles price and bottles sold. WHISKIES, DISTILLERY WHISKIES, and BOURBON WHISKIES were shown to be the most significant.

It would suggest that larger bottles are correlated with better sales. The Adjusted R^2 was improved from 0.979 to 0.984 by removing the influence points.

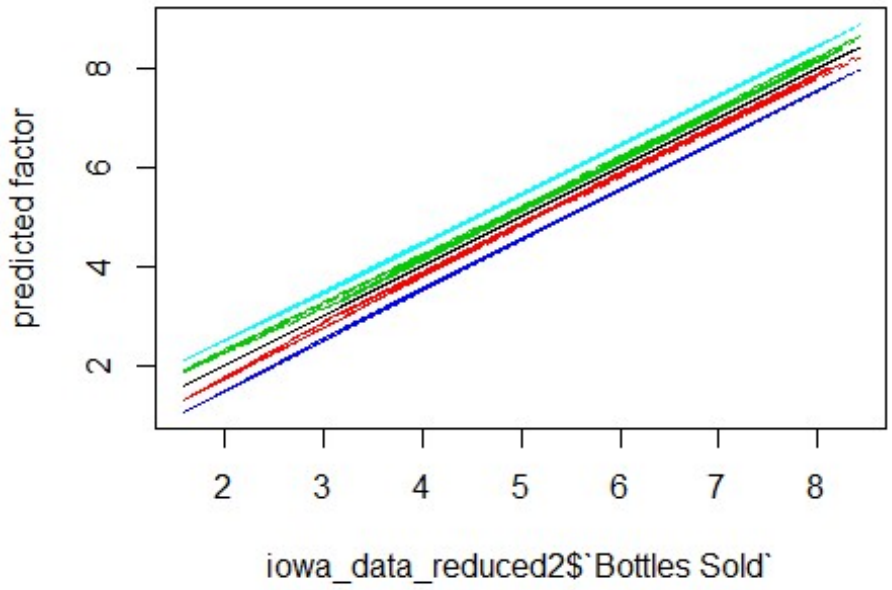
Selection Model

Also, the AIC of this model is r round(AIC(model3),3) which is a much better AIC than r round(AIC(model1),3) and (AIC(modelTo compare all our regular models first, we build a dataframe which contains all the performance parameters of the models. Out of the four regular models, it is clear that the Log transformation with forward regression is the best model. It has the lowest AIC and BIC.

Models	AIC	BIC	Deviance	df.residual
Correlation	1363.90365	1401.96263	9.928481e+05	99
Forward regression	1079.26792	1114.13563	1.087422e+05	96
Log transformation	-11.08298	26.72375	4.514541e+00	97

Prediction

Try To generate 30 data for next period for the Bottle sold, and predicted values obtained by evaluating the regression function in the frame new data by using model3 of log transformation with forward linear regression.



Discussion and Conclusions

The resulting models allow us to model in Des Moines for both Bottles Sold. We can utilize a naive forecast, assuming that the prior year of 2017 is predictive of the year 2018. However, there are further analysis types that may result in more robust predictions.

The performance of the proposed method was evaluated using a real data set provided by Iowa Department of Commerce, Alcoholic Beverages Division. The results of the evaluation indicated that the proposed method can cope with the low number of past records while accurately forecasting sales.

An evaluation of the liquor data set using these techniques may provide greater insight as the vast number of records could produce a more accurate model.

After exploratory analysis was done on the data, it was concluded that most liquor sold had different and specific characteristics and sales behavior, it was impractical to make a single

prediction model for all medicines, and most sales records had nonlinear relationships per years.

Appendices

AIC Value Comparison

```
## -------------------------------------
##          Model Name             AIC
## ------------------------- ----------
##  Correlation Bottles Sold   1363.904
##
##         Bottles Sold         1079.268
##
##      Log Bottles Sold        -11.083
## -------------------------------------
##
## Table: AIC Values
```

Supplemental tables and figures

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Month	1	112	6.437500	3.479104	6.500000	6.433333	4.447800	1.000000	1.200000e+01	1.100000e+01	0.0021616	-1.2519544	0.3287445
Year	2	112	2017.000000	0.000000	2017.000000	2017.000000	0.000000	2017.000000	2.017000e+03	0.000000e+00	NaN	NaN	0.000000
City*	3	112	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
CategoryName*	4	112	5.607143	2.957224	6.000000	5.633333	4.447800	1.000000	1.000000e+01	9.000000e+00	-0.0984481	-1.3050306	0.2794314
Bottles Sold	5	112	5.875522	1.773434	6.341651	6.001152	1.475627	1.581315	8.419349e+00	6.838035e+00	-0.7242590	-0.5534474	0.1675737

Sale (Dollars)	6	112	4333 2.31 5000	3992 4.57 3536	3598 6.99 0000	3867 5.26 3333	3235 6.97 4048	157. 500 000	1.39 5039 e+05	1.39 3464 e+05	0.8 935 026	-0.3 890 212	3772 .517 5992
Bottle Volume (ml)	7	112	799. 9501 04	88.6 0493 9	782. 0746 13	795. 9030 89	47.5 5382 2	619. 069 069	1.01 8885 e+03	3.99 8158 e+02	0.5 077 436	0.1 330 022	8.37 2379 8
State Bottle Cost	8	112	4749 .230 893	4120 .845 114	3836 .410 000	4335 .610 333	3346 .280 091	17.5 000 00	1.51 4666 e+04	1.51 2916 e+04	0.7 910 598	-0.5 637 597	389. 3832 629
State Bottle Retail	9	112	7126 .403 482	6184 .555 936	5755 .845 000	6505 .656 333	5020 .016 883	26.2 500 00	2.27 2218 e+04	2.26 9593 e+04	0.7 912 515	-0.5 642 293	584. 3856 063
Volume Sold (Gallons)	10	112	180. 4049 11	191. 6324 43	117. 8150 00	147. 4265 56	156. 7997 76	0.19 000 0	7.47 8500 e+02	7.47 6600 e+02	1.2 626 115	0.6 915 185	18.1 0756 38

Session Information

```
toLatex(sessionInfo())
```

```
## \begin{itemize}\raggedright
##   \item R version 3.4.4 (2018-03-15), \verb|x86_64-w64-mingw32|
##   \item Locale: \verb|LC_COLLATE=English_United States.1252|,
\verb|LC_CTYPE=English_United States.1252|, \verb|LC_MONETARY=English_United
States.1252|, \verb|LC_NUMERIC=C|, \verb|LC_TIME=English_United States.1252|
##   \item Running under: \verb|Windows 10 x64 (build 17134)|
##   \item Matrix products: default
##   \item Base packages: base, datasets, graphics, grDevices,
##     methods, stats, utils
##   \item Other packages: bindrcpp~0.2.2, car~3.0-0, carData~3.0-1,
##     corrplot~0.84, data.table~1.11.2, dplyr~0.7.4, forcats~0.3.0,
##     ggplot2~2.2.1, knitr~1.20, lubridate~1.7.4, magrittr~1.5,
##     pander~0.6.1, psych~1.8.4, purrr~0.2.4, randomForest~4.6-14,
##     readr~1.1.1, stargazer~5.2.1, stringr~1.3.1, tibble~1.4.2,
##     tidyr~0.8.0, tidyverse~1.2.1
##   \item Loaded via a namespace (and not attached): abind~1.4-5,
##     assertthat~0.2.0, backports~1.1.2, bindr~0.1.1, broom~0.4.4,
##     cellranger~1.1.0, cli~1.0.0, colorspace~1.3-2, compiler~3.4.4,
##     crayon~1.3.4, curl~3.2, digest~0.6.15, evaluate~0.10.1,
```

```
##     foreign~0.8-69, glue~1.2.0, grid~3.4.4, gtable~0.2.0,
##     haven~1.1.1, highr~0.6, hms~0.4.2, htmltools~0.3.6,
##     httr~1.3.1, jsonlite~1.5, labeling~0.3, lattice~0.20-35,
##     lazyeval~0.2.1, mnormt~1.5-5, modelr~0.1.2, munsell~0.4.3,
##     nlme~3.1-137, openxlsx~4.0.17, parallel~3.4.4, pillar~1.2.2,
##     pkgconfig~2.0.1, plyr~1.8.4, R6~2.2.2, raster~2.6-7,
##     Rcpp~0.12.16, readxl~1.1.0, reshape2~1.4.3, rio~0.5.10,
##     rlang~0.2.0, rmarkdown~1.9, rprojroot~1.3-2, rstudioapi~0.7,
##     rvest~0.3.2, scales~0.5.0, sp~1.2-7, stringi~1.1.7,
##     tools~3.4.4, utf8~1.1.3, xml2~1.2.0, yaml~2.1.19
## \end{itemize}
```

R programming code

See Final Project.rmd on GitHub for source code.

https://github.com/fung1091/data621/blob/master/finalproject/data621final.Rmd