UNIVERSITY OF CALIFORNIA

Los Angeles

Beer Production and Beer Judge in United States

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Applied Statistics

by

Yi Hu

2019

ABSTRACT OF THE THESIS

Beer Production and Beer Judge in United States

by

Yi Hu
Master of Applied Statistics
University of California, Los Angeles, 2019
Professor Yingnian Wu, Chair

The goal of this thesis is to analyze the current beer market in U.S. according to the manufacturing data. It will also give insights into beer judgement based on consumers' reviews. This thesis presents the market trends for producers and summarizes evaluations for beer lovers.

The thesis of Yi Hu is approved.

Rick Schoenberg

Nicolas Christou

Yingnian Wu, Committee Chair

University of California, Los Angeles

2019

Contents

1 Introduction

Beer is the third most popular drink over the world. While watching a sport game or hanging out with friends, many people like to grab a beer. However, it happens to all of us that a drink menu can be quite overwhelming. Even though most restaurants and bars have classified their beers by country or by styles, we still find it difficult to make a good choice. For example, many imported beers are named in Hochdeustch. You may have the question in mind: What does "Hochdeutsch" mean? Yes, this is exactly how we feel when we read an alcohol menu if we do not know any German and/or French. Most people love beer but do not have in-depth knowledge about beer. This paper will discuss several features of beers and the beer market in United States. Hopefully, readers will gain some insights into beers and find it easier to order a drink in future.

The following content can be separated into two major parts. The first part is from Chapter 2 to Chapter 4, introducing the beer manufacturing industry in U.S. from producers' perspective. This part applies both time series theory and regression analysis. It examines the relationship between manufacture outputs and consumed materials' volumes. The second part is Chapter 5 that analyzes beers by features from consumers' perspective. This part performs mostly hypothesis tests and classification tasks. It inspects both beer reviews and ABV values, as well as uses review scores and ABV values to predict a lager or an ale.

2 Beer Manufacture Output in U.S.

The beer manufacture data pictures the overall consumption and production condition in United States. Thinking from a management purpose and analyzing the output data will gain us many interesting findings about beer. For example, we find that the total output in June is about 36% higher than the total output in December. Also, based on the supply data, we can conjecture consumers' demand. Such as, consumption in summer must be significantly higher than that in winter.

Trend is the primary indicator that yields information. The trend plots in the following sections allow us to interpret the changes happening in the U.S. beer industry. Significantly, we find that the beer supply is decreasing by years; alternatively, beer demand is decreasing in U.S. in recent years.

2.1 Intro to Production and Exports Data

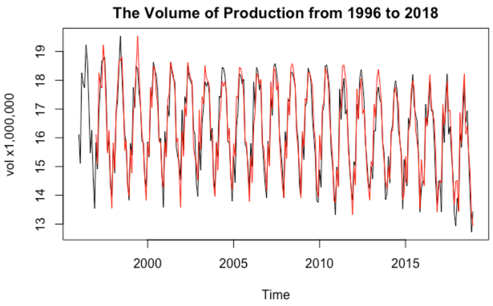
The whole manufacturing data are scraped from U.S. Department of Treasury, Alcohol and Tobacco Tax and Trade Bureau. The website contains monthly data from 1984 to 2018. Since data before 1996 are incomplete, the time horizon used in the following analysis is from 1996 January to 2018 December. The manufacture datasets are monthly data that contains the information of total volume of production, removals and stocks on hand end-of-month in barrels (1 Barrel is equivalent to 31 gallons). Within removals, there are taxable items including "in bottles and cans" and "in kegs" and tax-free items including "for exports", "for vessels and aircraft" and "consumed on brewery premises". After data cleansing and consolidation, three sub-categories from the cleaned data table will be included in the following analysis. That is, in bottles and cans, in kegs, and for exports.

2.2 Total Volume of Production

2.2.1 Time Series Analysis

The production's time series is as below in Figure 1. The black line is the real data; the red line represents a 12-month lag of the original data.

Figure 1



According to the almost overlapped two lines from the plot, we observe an evident seasonality factor. Therefore, we perform a box plot of production over each month, as in the left plot of Figure 2. We can observe that in general, beer production increases from the beginning of a year and reaches the peak in June and decreases until December. However, compared to January's volumes, decreased February's volumes seem to devaluate this argument. The cause could be the number of days in February is less than that of January. Thus, the adjusted production volume is the daily production volume in every month. After adjustment, the right plot below shows a clear trend of beer manufacturing within a year. $AdjustedProductionVol. = \frac{ProductionVol.EveryMonth}{TheNumberofdaysintheMonth}$

Figure 3 shows a decomposition of the time series. We observe a decreasing trend in total beer production volume in U.S. from 1996 to 2018. Next, the plot in the middle shows the lines by each year after seasonally adjustment. By using aggregate function, the right plot Figure 3 displays a trend on yearly basis. We observe a decreasing trend of total volume of beer production in U.S.

Figure 2

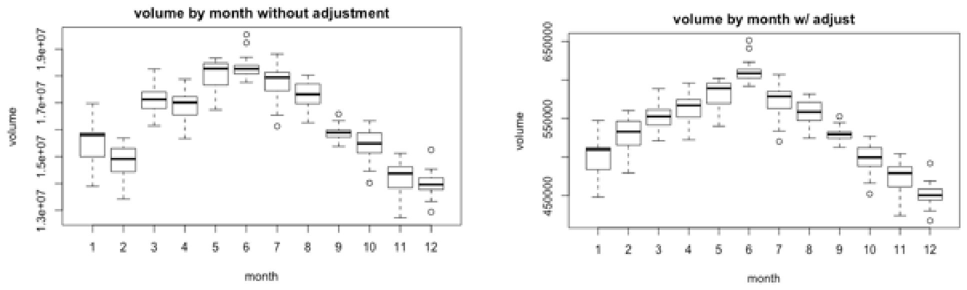
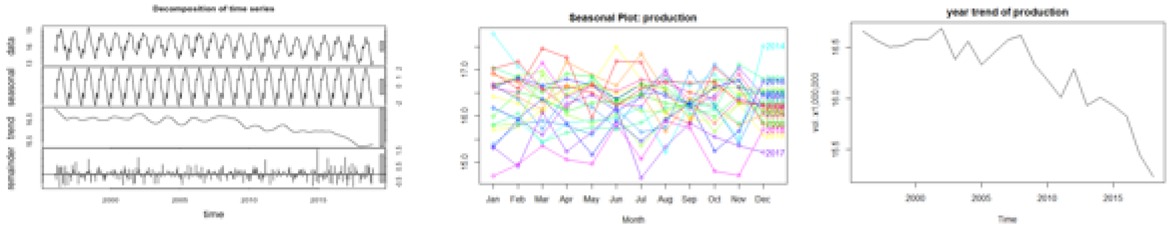
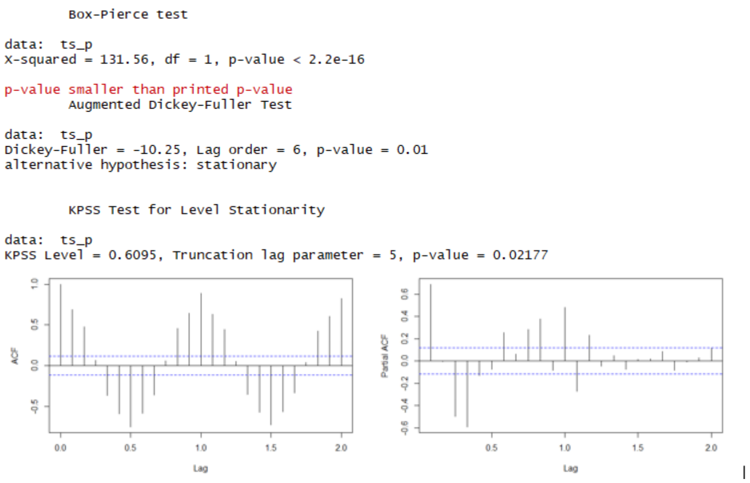


Figure 3



Next, I performed a few statistics tests on the time series of production volume. The Box-Ljung test rejects the null hypothesis that the series is independent. The Augmented Dickey-Fuller Test and KPSS test reject the null hypothesis that the serie is not stationary, since they both have p-values less than 0.05. Then through the ACF and PACF plots, ACF is a sine-wave shape pattern and PACF has a spike at lags 1. It indicates two autoregressive (p) parameters.

Figure 4

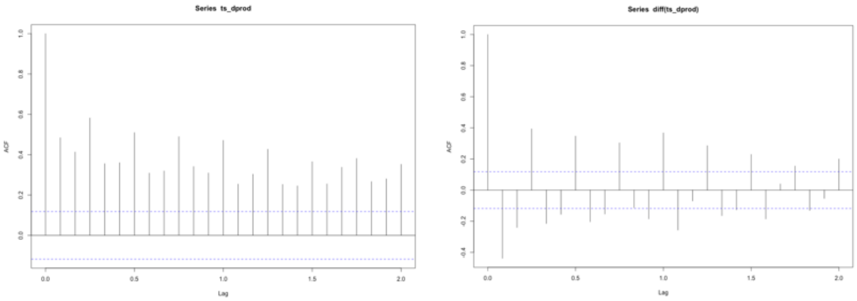


```
        Box-Pierce test

data:  ts_p
X-squared = 131.56, df = 1, p-value < 2.2e-16

p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  ts_p
Dickey-Fuller = -10.25, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary


        KPSS Test for Level Stationarity

data:  ts_p
KPSS Level = 0.6095, Truncation lag parameter = 5, p-value = 0.02177
```

2.2.2 Stationary Analysis

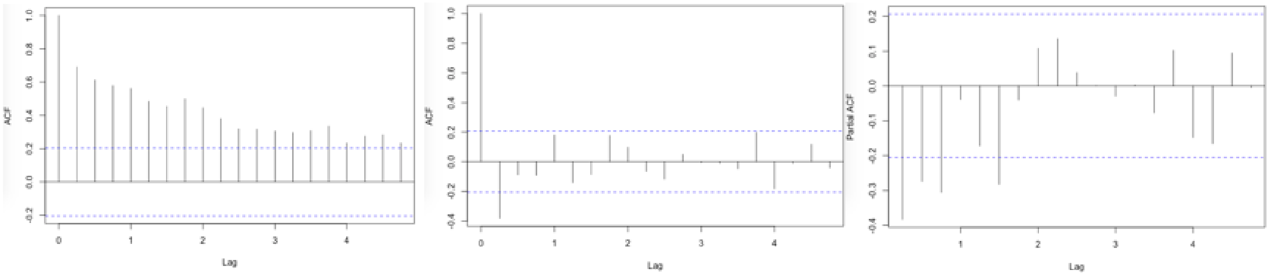
Although stationary tests in the previous unit reject the null hypothesis, we can observe a clear seasonal pattern in the ACF plot. Therefore, we will deseasonalize the series and then observe the result. By grouping by and then summarizing the averages of the production output by month, we obtain a deseasonized series. The new ACF plot is in the next figure. However, it shows no stationarity, so I take a difference. The ACF after a diff still exhibits a strong quarterly pattern. Therefore, I decide to aggregate the series by every three month.

Figure 5



Then the ADF test of the aggregated series do not reject the null hypothesis. Also, the ACF plot is in the left plot in below. It shows that the aggregated series has a long memory. Therefore, we take a difference, and finally, it renders a satisfying ACF and PACF plots. It also passes the ADF stationary test. Therefore, we can say that by aggregating by quarter and taking a one-order difference, we can obtain a stationary time series for production output.

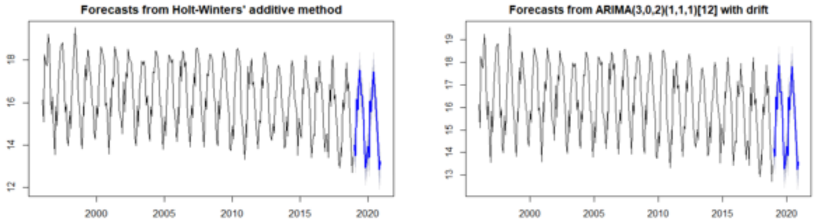
Figure 6

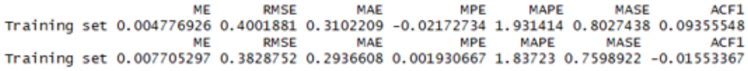


2.2.3 Forecasts

Forecasting and accuracy results from Holt-Winters and ARIMA are as below. Auto.Arima model gives lower scores for RMSE, MAE, absolute MPE, MAPE, MASE and absolute ACF1, so it performs better than Holt-Winters model in this case.

Figure 7



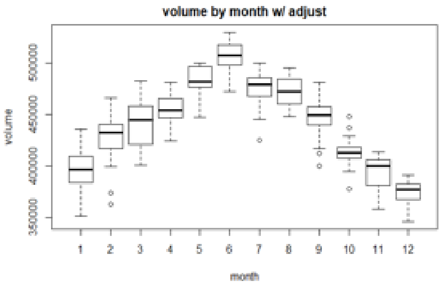


```
                    ME      RMSE       MAE         MPE     MAPE      MASE       ACF1
Training set 0.004776926 0.4001881 0.3102209 -0.02172734 1.931414 0.8027438 0.09355548
                    ME      RMSE       MAE         MPE    MAPE      MASE        ACF1
Training set 0.007705297 0.3828752 0.2936608 0.001930667 1.83723 0.7598922 -0.01553367
```

2.3 The Volume of Production in Bottles and Cans

2.3.1 Time Series Analysis

Similar to the process taken in 2.2, the results of bottles and cans are similar to those of the total production volume, according to plots below. However, it is noticeable that the trend of the production in bottles and cans has slight difference from that of in total production volume. Instead of a consistent decreasing trend, the line of bottles & cans displays an increasing trend before year 2007.



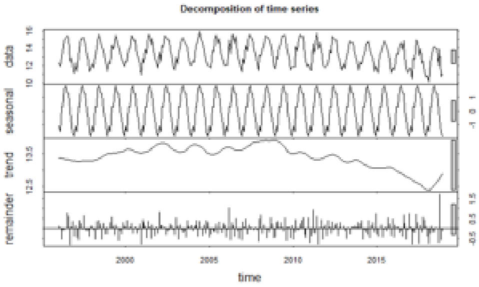
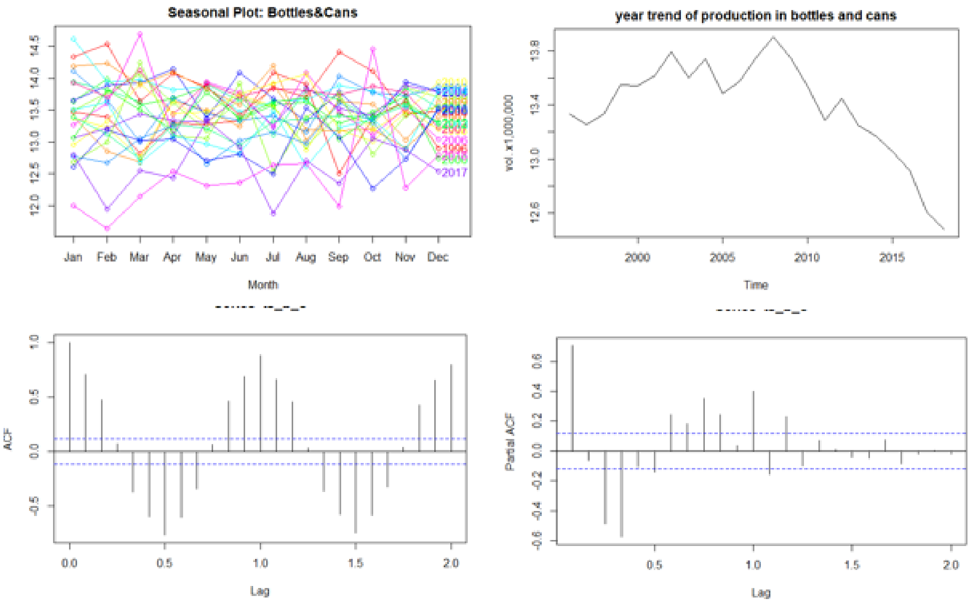
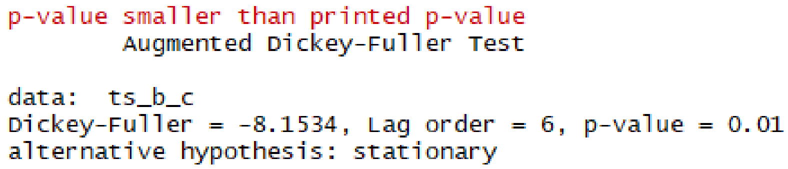


Figure 8

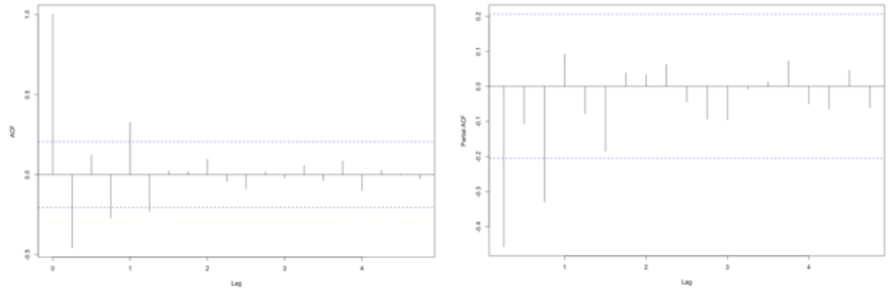



```
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  ts_b_c
Dickey-Fuller = -8.1534, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

2.3.2 Stationary Analysis

Similar to the production output data, although the time series of production volume in bottles and cans reject the null hypothesis of ADF test, its ACF shows a seasonal pattern that arouse our concerns. Therefore, we first deseasonalize the series, aggregate it by quarter and then take a difference, to see whether we can obtain a stationary time series. The outcome is confirmed in the Figure 9 below.

Figure 9

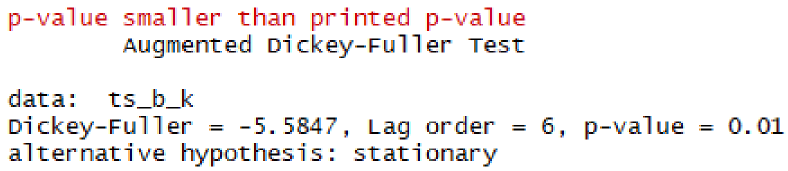


2.4 The Volume of Production in Kegs

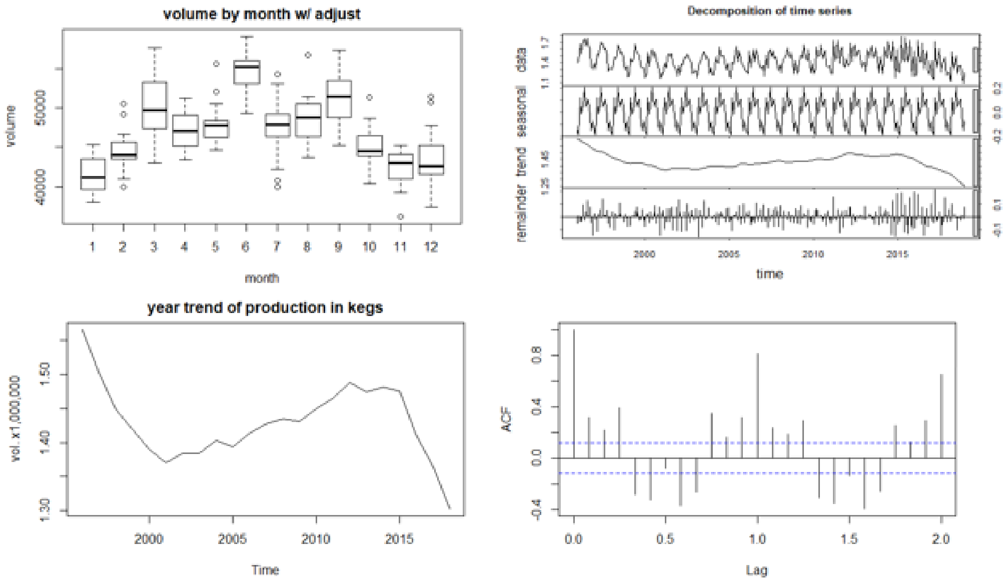
2.4.1 Time Series Analysis

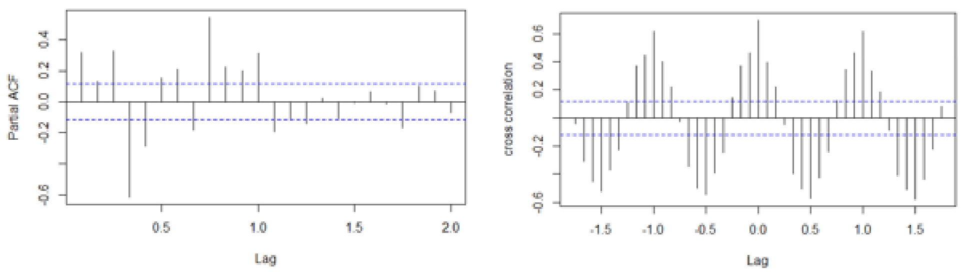
Similarly, the ADF test rejects the null hypothesis that the series is non-stationary. In the following plots, we can also observe that the pattern of monthly manufacturing volumes of kegs slightly differentiates from that of the total volume and that of the volume of bottles and cans. Instead of a smooth increase followed by a smooth decrease, the adjusted volumes of kegs show an increasing trend in each of the first three quarters and a decreasing trend in the last quarter of a year. A possible reasoning behind this finding is that there may exist a target output per quarter that drives up the production efficiency.

Not only the difference in the pattern of monthly output volume, the pattern in the year trend displays a decrease from 1996 to 2001, followed by an increase from 2001 to 2015 and next a decrease again from 2015 to 2018.



```
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  ts_b_k
Dickey-Fuller = -5.5847, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

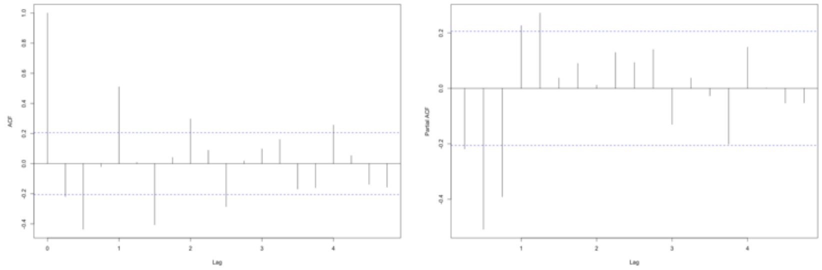




2.4.2 Stationary Analysis

To keep it in accord with the previous output analysis, the treated series has plots as below.

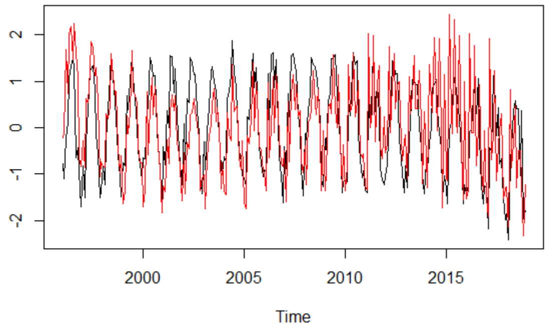
Figure 10



2.4.3 Correlation between the vol. of Bottles & Cans and the vol. of Kegs

Pearson correlation test yields a correlation result of 0.695, which is medium to strong. From Figure 11, we observe a high resemblance (Black line is bottles & cans, red line is kegs). Also, since year 2011, the volatility of kegs' output becomes high (more volatile fluctuations).

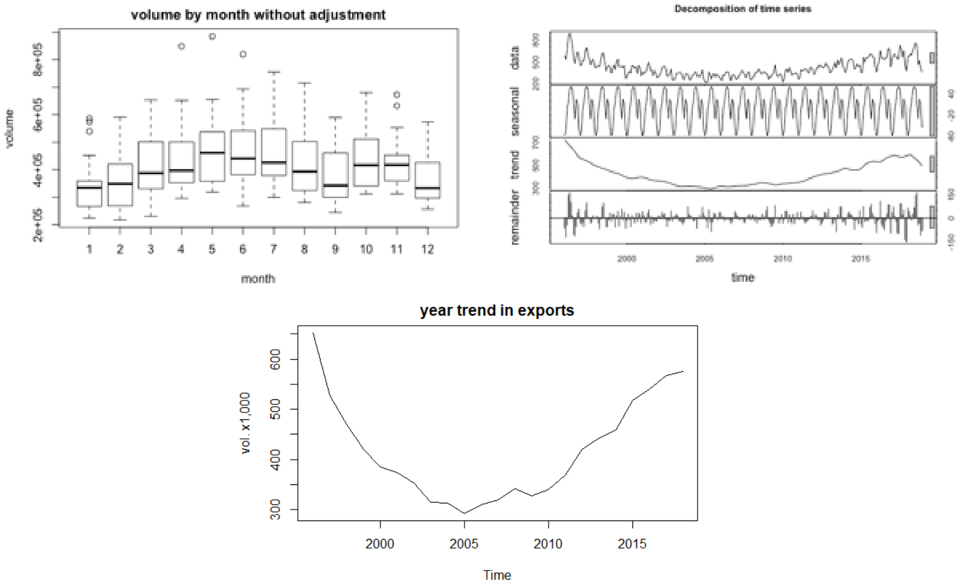
Figure 11



2.5 The Volume of Exports

In the volume of exports, the seasonality factor has smaller variance than that in previous sections. Next, from the trend plot, aggregated by mean, in Figure 12, we observe that the export amount reaches the minimum in 2005 and then gradually increases. The quantity of exports in late 1990s and recent year 2018 nearly double the quantity of export in 2005.

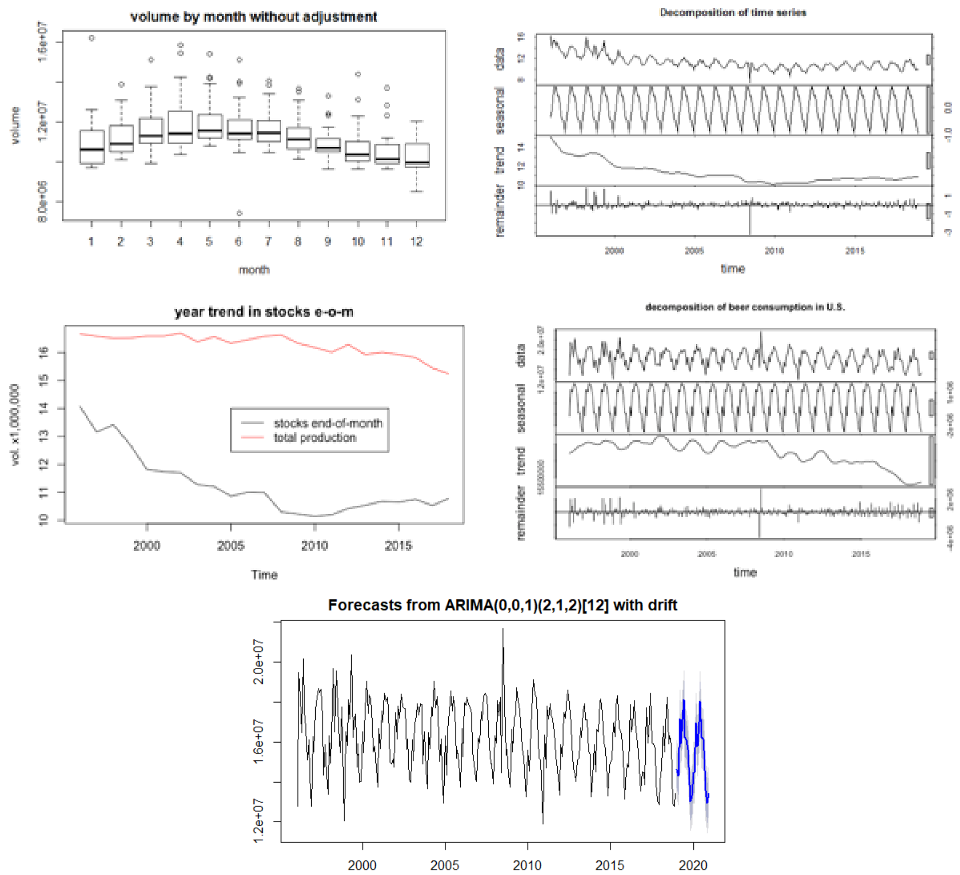
Figure 12



2.6 End-of-Month Stocks and the Consumption Volume in U.S.

The pattern of the seasonality of stocks end-of-month is similar to the pattern of the seasonality of total production volume, however, with a narrower variance. Next, from the trend plot in Figure 13, we observe a decreasing trend in stocks EOM (end-of-month) from 1996 to 2010 and a slowly increasing trend from 2010 to 2018. Suppose all beers are sold/consumed before expiration date, by adding the changes in the EOM stocks to the current month's production amount, we can roughly get the consumed quantity by month. From the second stl plot, the consumption maintains almost stable from 1996 to 2009, then mildly decreases from 2009 to 2016 and finally, significantly drops in 2016 - 2018.

Figure 13



From the analysis above, the numbers calculated for recent years (2016 to 2018 end) indicates a reduced popularity of beers in U.S. market. The forecast of the estimated consumption is in the figure above. The equation to calculate consumption is as below.

$$Consumption = TotalProduction_t - (StockEnd_t - StockEnd_{t-1})$$

3 Consumed Material in U.S.

In the process of brewing a beer, first is to malt grains. Next activating enzymes in the step of mashing causes grains to release sugar. Then we can add non-grain products such as hops during the boiling process. Last, we will add in yeasts to transform sugar into alcohol and to let beer carbonate. In this chapter, we will look into the consumption data of raw materials that include both grains and non-grain products.

It is important for us to understand changes and patterns in the consumption data of raw materials, since consumed materials significantly influence the tastes and the style of beers. For examples, American IPA consumes a lot of hops, in order to balance the sweetness with bitterness and to add fruity smells; Light Lager often uses recipes that adds corns and/or rice to get lighter-colored beers as well as to reduce the production cost. In the following units, you will see how much grains and non-grains are consumed every month in beer production.

3.1 Introduction to Consumed Material Data

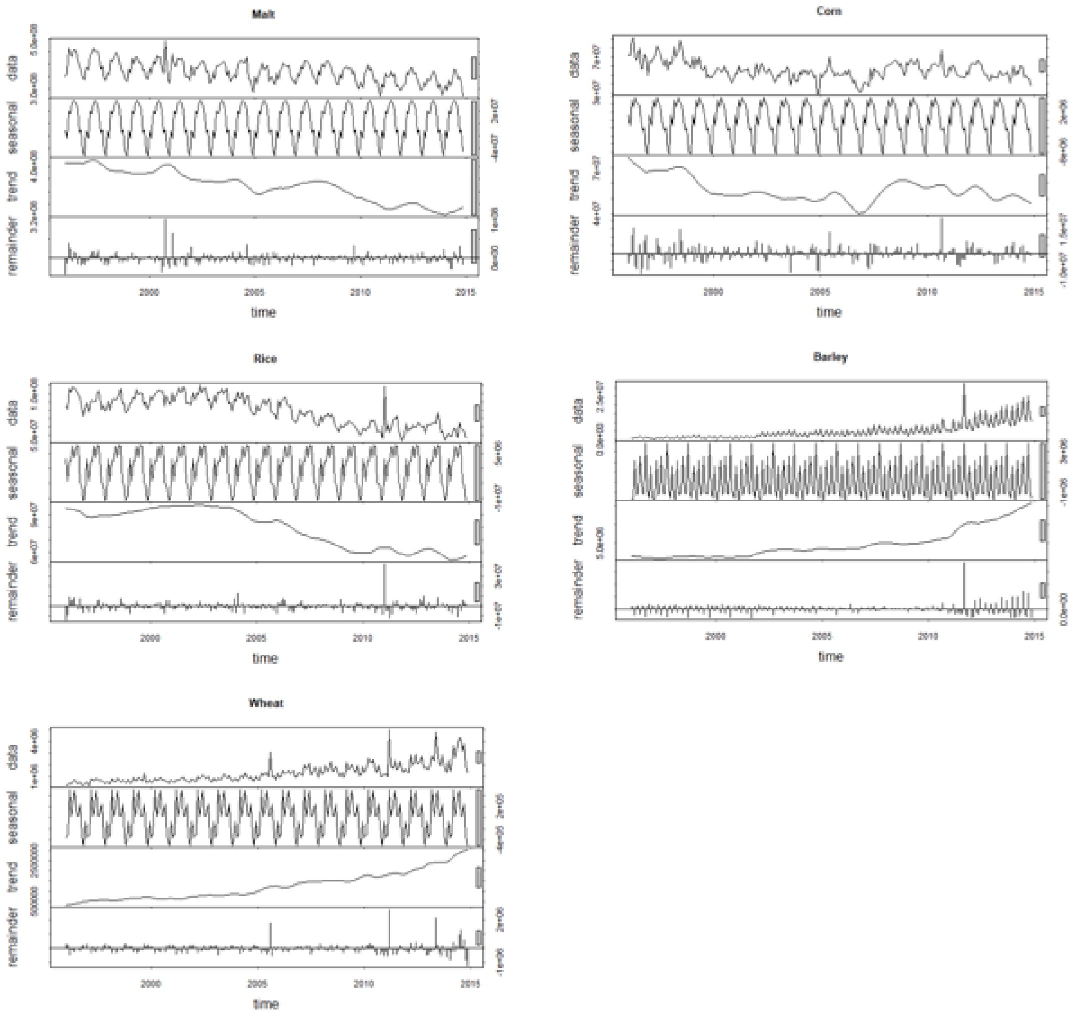
Same with the manufacture data, consumed materials' data is scraped from the U.S. Department of Treasury, Alcohol and Tobacco Tax and Trade Bureau website. Both data are stored in the same pdf. but in different tables. Both data are converted to csv. from pdf. and then passed to cleaning process. However, the government reformed the data format after 2014. That is, the materials' data after 2014 are either vacant or measured in a different scale without notes. Therefore, in order to keep the integrity of data, I apply a time horizon from 1996 to 2014 for the following analysis on the consumed raw materials. The material data includes both grain and non-grain raw materials. There are five grain products and four non-grain classifications – three non-grain products and other.

3.2 Grain Products

3.2.1 Decomposition of Five Grains

The five grains are malt, corn, rice, barley and wheat. Malt is prepared from cereal grain, chiefly barley, by allowing partial germination to modify the grain's natural food substances. However, the term, barley, usually refers to the seed of the barley plant in brewery. Corns and rice are often adjunct; malt and/or barley are usually the main materials, wheat can be used as both main material and adjunct. All beers use malts, except for few specialties.

Figure 14



From 1996 to 2014, there is a decreasing trend in malt consumption, corn consumption and rice consumption; there is an increasing trend in barley consumption and wheat consumption. The stl plots below indicate that from 1996 to 2014, there is a trend that beer manufactures intend to produce less lagers but more flavored beers that often use more wheat and barley in their recipes. For instance, wheat beers require a large proportion of wheat, while pale lager may only use malts and corns/rice.

Next, the seasonal component of barley looks distinctly more volatile than the other grains, because it has a much higher frequency of fluctuation. We will examine it further in the next paragraph.

3.2.2 Monthly Consumption of Five Grains with Adjustment

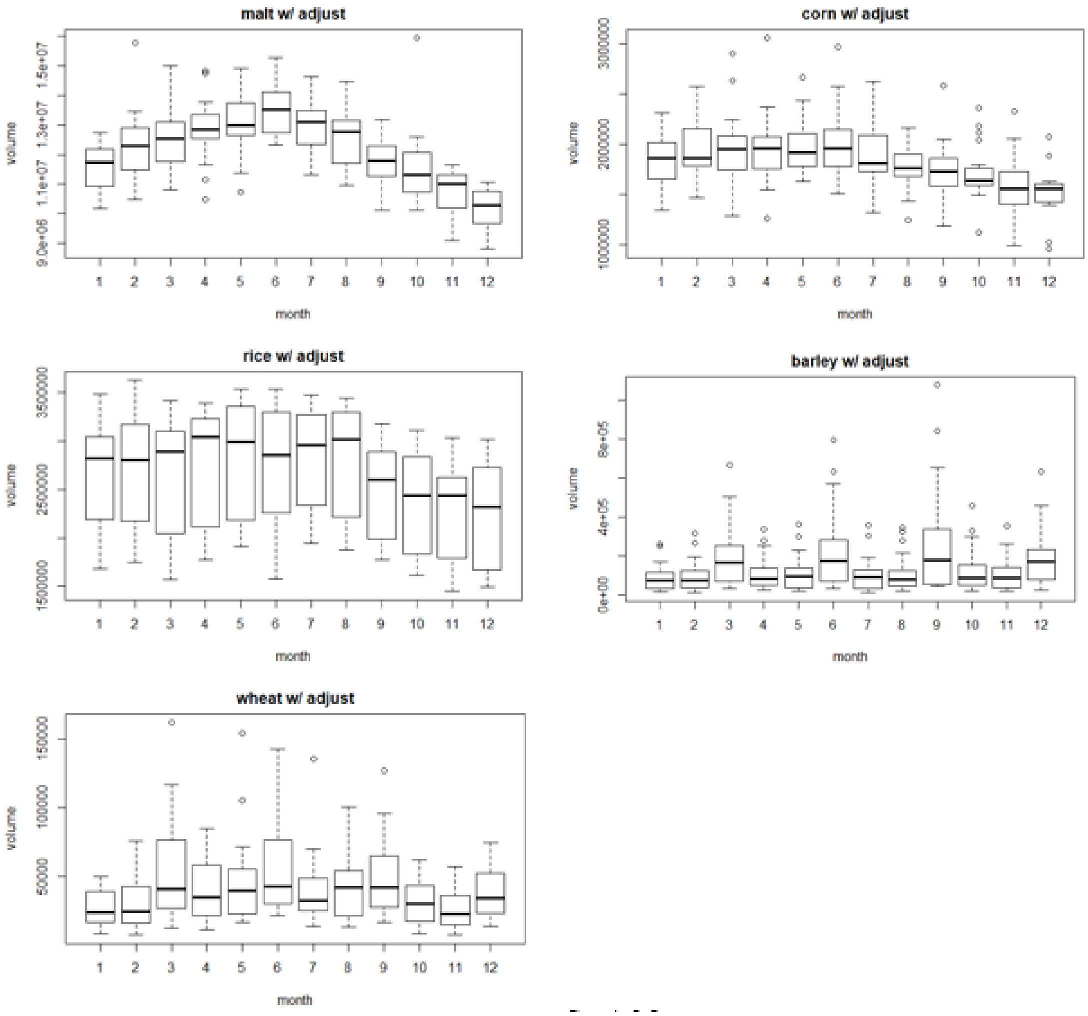
The Figure 15 shows the monthly consumption of grains after adjustment. The adjustment is based on the number of days per month. On average, malts' consumption reaches the peak in June and tumbles to the nadir in December. Corns and rice do not show a significant pattern from the plots. However, both barley and wheat seem to have the highest consumption in the last month of each quarter, although barley shows a more clear pattern. According to the results above, we can gauge that there are some special beers that use a lot of barley are quarterly produced, usually in March, June, September and December. The same fact may also be true for wheat, especially during the first three quarters.

3.3 Non-Grain Products

3.3.1 Decomposition of Three Non-Grains

Non-grain products include sugar, hops(dry), hops(as extract) and other. Here we only include analysis results of the three non-grain products.

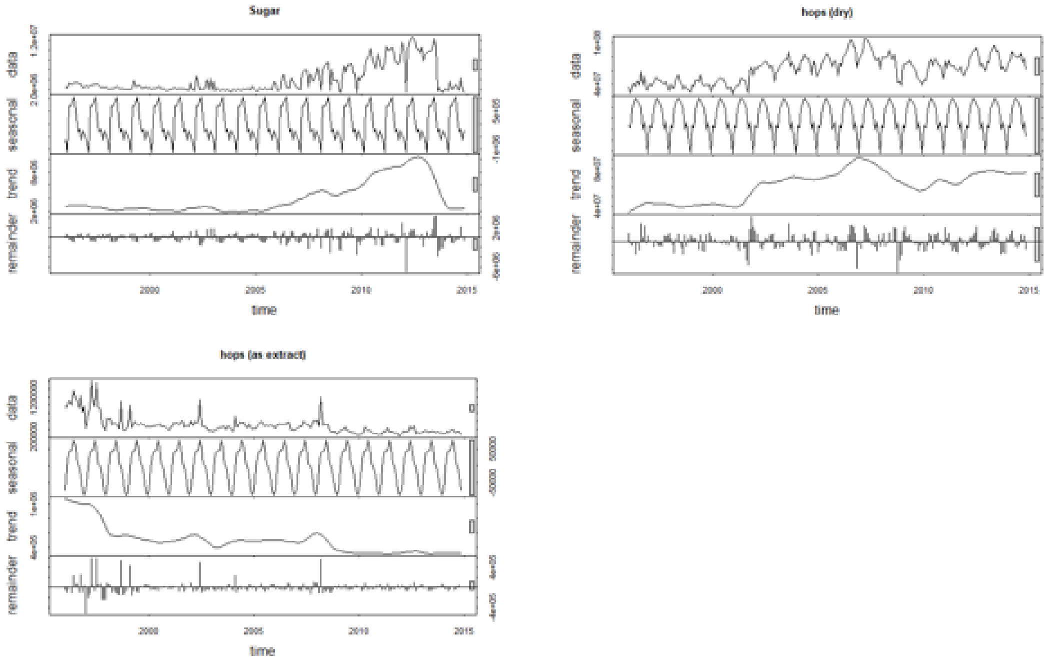
Figure 15



In the stl plot of sugar, Figure 16, we observe a turn from 2005. The consumption of sugar gradually increases till mid 2013, followed with a significant jump till end 2014. In beer production, sugar can be used to change flavors or add up the alcohol volume, because sugar will be transformed to alcohol by yeast. These turns in sugar consumption could indicate turns in producing preferences.

Next, there is an increasing trend in dry hops and a decreasing trend in hops as extract. It indicates that changes happen in the mainstream bittering process that uses hops.

Figure 16



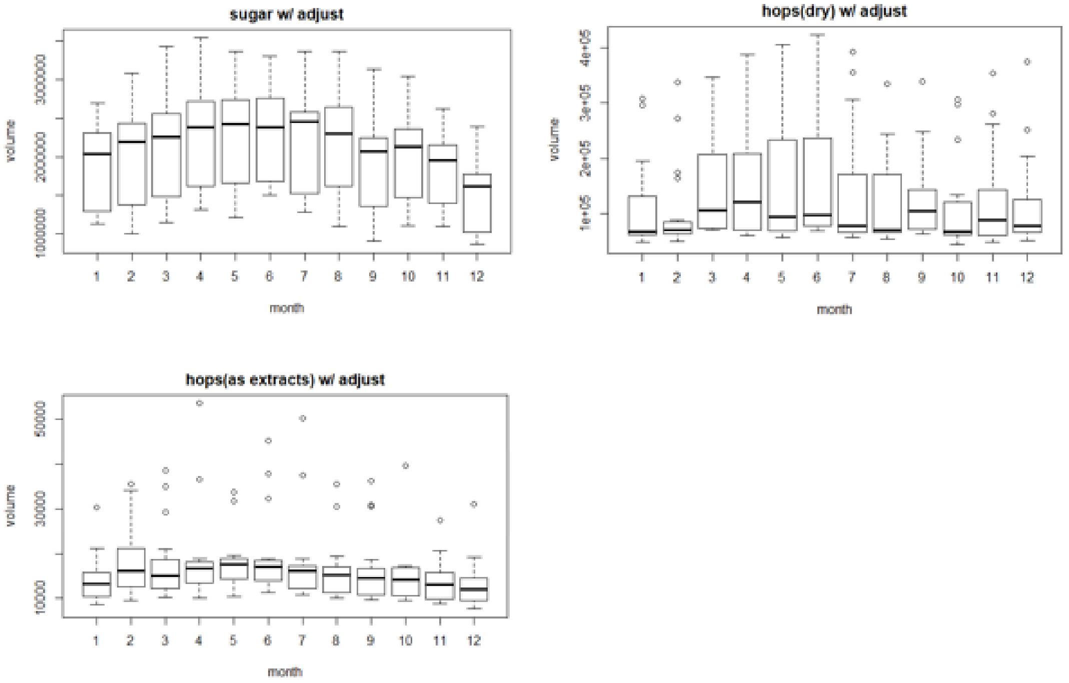
3.3.2 Monthly Consumption of Three Non-Grains with Adjustment

The plots in Figure 17 show that unlike the consumption of sugar, the consumption of dry hops and the consumption of hops as extract have many outliers. In addition, sugar's data are left skewed per month; dry hops' data are right skewed; hops' (as extract) data are mostly left-skewed. Also, all three materials roughly accord to the high consumption during Spring and Summer and low consumption during Fall and Winter.

4 Analysis on Manufacture Outputs and Materials

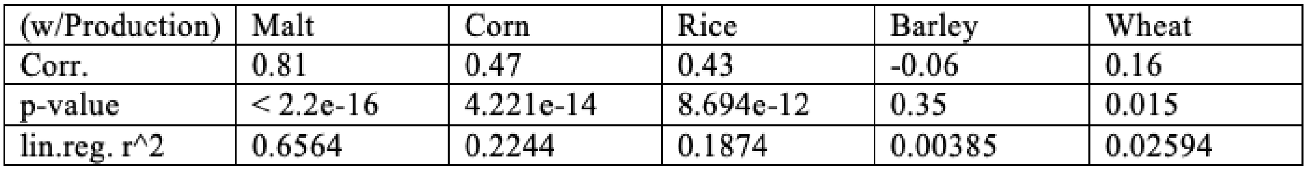
The goal of performing regression analysis in this section is to understand the relationships between variables in the two datasets, production data and material data. Part of results conform to our general knowledge about beers; the others will provide information and develop our new awareness about beers.

Figure 17



4.1 Correlations between Total Output and Eight Raw Materials

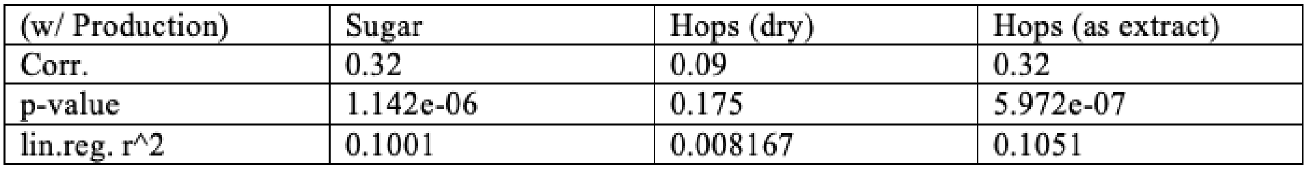
Within the correlation test between total production volume and 5 grains, malt, corn, rice and wheat are statistically significant. The consumption volume of malt has the highest also a strong level of correlation with the total production volume. Corn and rice have medium correlation with production. Wheat has a weak correlation with production.



(w/Production)	Malt	Corn	Rice	Barley	Wheat
Corr.	0.81	0.47	0.43	-0.06	0.16
p-value	< 2.2e-16	4.221e-14	8.694e-12	0.35	0.015
lin.reg. r^2	0.6564	0.2244	0.1874	0.00385	0.02594

Within non-grains, sugar and hops as extract have statistically significant correlation results with production. The monthly consumption volumes of both sugar and hops as extract have medium level of correlation with the total monthly production volume.

In addition, in univariate linear regressions, the R^2 of malt over production is 0.6564. That is, only the malt's consumption volume can explain 65.64% of the monthly production volume.



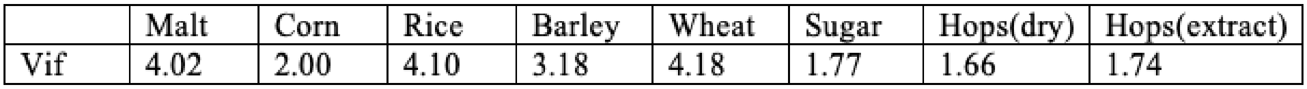
(w/ Production)	Sugar	Hops (dry)	Hops (as extract)
Corr.	0.32	0.09	0.32
p-value	1.142e-06	0.175	5.972e-07
lin.reg. r^2	0.1001	0.008167	0.1051

This result conforms to the fact that malt is the major and the most important material in brewery. The coefficient in the model of malt is 0.0291. Therefore, with one unit increase in malt consumption, there will be around 0.03 unit of increase in production. Next, corn and rice followed after with R^2 values of 0.2244 and 0.1874. The ranking, from high to low, of R^2 shows the significance of each material in influencing the total production volume. It implies that corns and rices are also widely used in beer production and therefore influenced the output. Corn has a coefficient of 0.0691, and rice has a coefficient of 0.0369. However, dry hops , barley and wheat barely influence the production output. It implies that these materials may not be widely used. Alternatively say, they are not mandatory in brewery.

4.2 Regression Analysis on Production Data and Material Data

In this section, I examined the relationships between each element in production data and materials, in order to understand how well the consumption of material can predict production outputs, export volume and stocks by the end of month. Full model on each dependent variables includes 8 independent variables in material data: malt, corn, rice, barley, wheat, sugar, hops (dry) and hops (as extract). There are six dependent variables: total production volume, production volume in bottles and cans, production volumes in barrels and kegs, export volume, stocks end-of-month, and monthly consumption. The variance inflation factor, VIF, will be referred to quantify the severity of multicolleariry. Also, stepwise models are utilized as the first step for variable elimination. Then both VIF and p-values of variables sponsor to select the ultimate independent variables for each model. The ultimate goal is to predict the outputs by achieving a high R^2 value with few independent variables.

A full model of 8 variables have a vif table below:

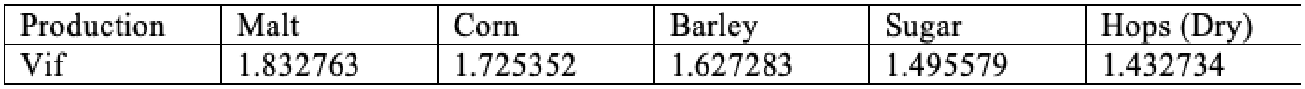


	Malt	Corn	Rice	Barley	Wheat	Sugar	Hops(dry)	Hops(extract)
Vif	4.02	2.00	4.10	3.18	4.18	1.77	1.66	1.74

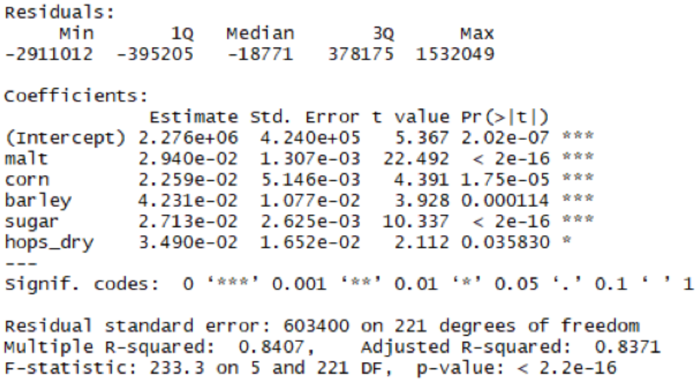
Obviously, malt, rice and wheat have relatively high vif values, because they are larger than 4. Barley is also worth paying attention as it has a value larger than 3. In the following regressions, we will consider eliminating a few of them to achieve lower vifs.

4.2.1 Total Production Volume

The full model of total production volume on eight variables gives R^2 of 0.8643. Backwise step model deletes barley from the full model. Then after further feature selection, five variables of malt, corn, barley, sugar and hops_dry can explain 84.07% of total production volume. VIF table for the new model is as below.



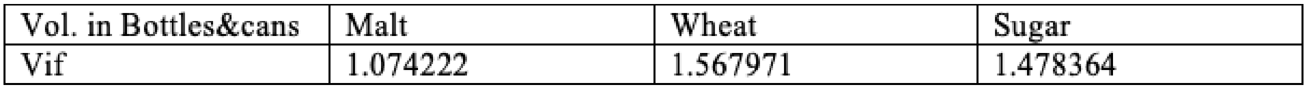
Production	Malt	Corn	Barley	Sugar	Hops (Dry)
Vif	1.832763	1.725352	1.627283	1.495579	1.432734



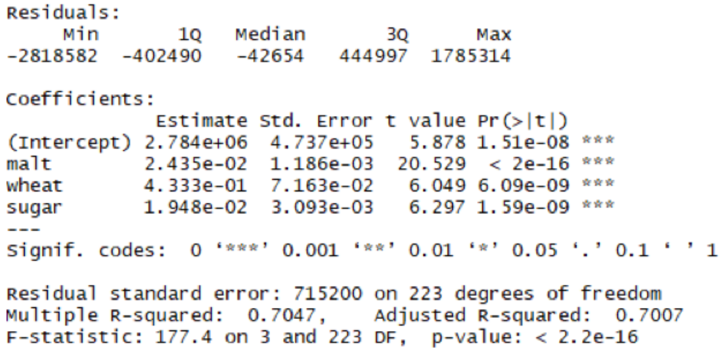
```
Residuals:
     Min       1Q   Median       3Q      Max
-2911012  -395205   -18771   378175  1532049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.276e+06  4.240e+05   5.367 2.02e-07 ***
malt        2.940e-02  1.307e-03  22.492  < 2e-16 ***
corn        2.259e-02  5.146e-03   4.391 1.75e-05 ***
barley      4.231e-02  1.077e-02   3.928 0.000114 ***
sugar       2.713e-02  2.625e-03  10.337  < 2e-16 ***
hops_dry    3.490e-02  1.652e-02   2.112 0.035830 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 603400 on 221 degrees of freedom
Multiple R-squared:  0.8407,    Adjusted R-squared:  0.8371
F-statistic: 233.3 on 5 and 221 DF,  p-value: < 2.2e-16
```

4.2.2 Production Volume in Bottles Cans

For production in bottles and cans, the full model gives an R^2 of 0.7186, while three variables of malt, wheat and sugar can explain 70.47% of the outcome data.



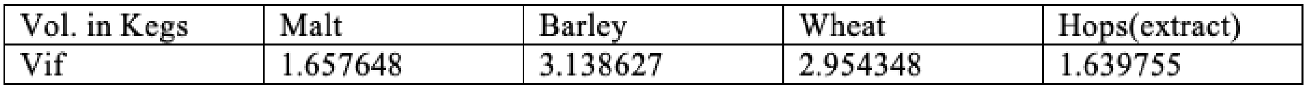
Vol. in Bottles&cans	Malt	Wheat	Sugar
Vif	1.074222	1.567971	1.478364



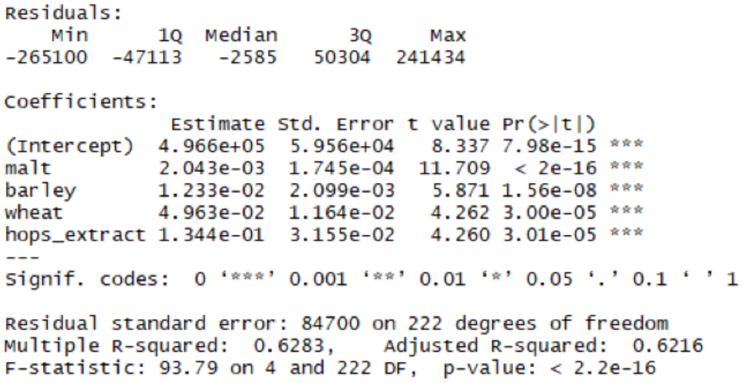
```
Residuals:
     Min       1Q   Median       3Q      Max
-2818582  -402490   -42654   444997  1785314

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.784e+06  4.737e+05   5.878 1.51e-08 ***
malt        2.435e-02  1.186e-03  20.529  < 2e-16 ***
wheat       4.333e-01  7.163e-02   6.049 6.09e-09 ***
sugar       1.948e-02  3.093e-03   6.297 1.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 715200 on 223 degrees of freedom
Multiple R-squared:  0.7047,    Adjusted R-squared:  0.7007 
F-statistic: 177.4 on 3 and 223 DF,  p-value: < 2.2e-16
```

4.2.3 Production Volume in Kegs

For production volumes in Kegs, the full model gives R^2 of 0.6456; now malt, barley, wheat and hops_extract can explain 62.83% of the production volume in kegs. It is noticeable that barley is significant in this model, although barley is not statistically significant in previous models. This outcome implies that compared to bottles cans, barley is more often involved in the production in kegs.



Vol. in Kegs	Malt	Barley	Wheat	Hops(extract)
Vif	1.657648	3.138627	2.954348	1.639755

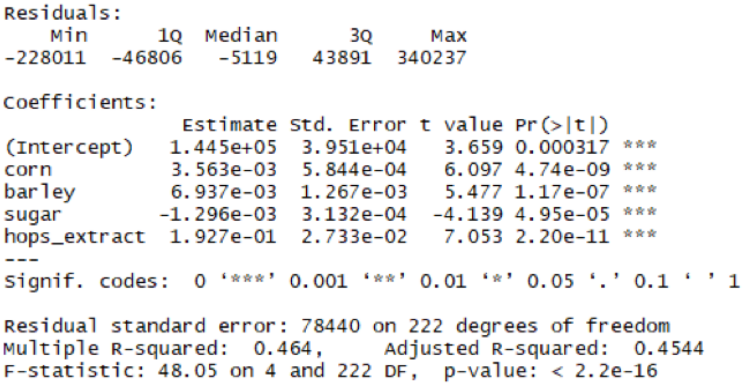


```
Residuals:
    Min      1Q  Median      3Q     Max
-265100  -47113   -2585   50304  241434

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.966e+05  5.956e+04   8.337 7.98e-15 ***
malt         2.043e-03  1.745e-04  11.709  < 2e-16 ***
barley       1.233e-02  2.099e-03   5.871 1.56e-08 ***
wheat        4.963e-02  1.164e-02   4.262 3.00e-05 ***
hops_extract 1.344e-01  3.155e-02   4.260 3.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84700 on 222 degrees of freedom
Multiple R-squared:  0.6283,    Adjusted R-squared:  0.6216 
F-statistic: 93.79 on 4 and 222 DF,  p-value: < 2.2e-16
```

4.2.4 Export Volume

The full model is with a R^2 of 0.4869, while corn, barley, sugar and hops_extract can explain 46.4%. Noticeably, malt is eliminated after feature selection since it is not strongly statistically significant in the full model. The final model without malt still renders a sound R^2. The stats can make sense, because export volume is a very small portion in the total production output. Therefore, it is possible that the consumption of malt does not count for the export volume.

Export Vol.	Corn	Barley	Sugar	Hops (extract)
Vif	1.316933	1.332128	1.259975	1.434066

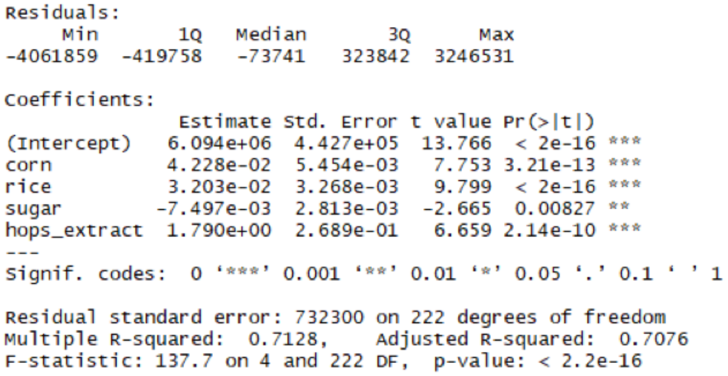


```
Residuals:
    Min      1Q  Median      3Q     Max 
-228011  -46806   -5119   43891  340237 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.445e+05  3.951e+04   3.659 0.000317 ***
corn         3.563e-03  5.844e-04   6.097 4.74e-09 ***
barley       6.937e-03  1.267e-03   5.477 1.17e-07 ***
sugar       -1.296e-03  3.132e-04  -4.139 4.95e-05 ***
hops_extract  1.927e-01  2.733e-02   7.053 2.20e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78440 on 222 degrees of freedom
Multiple R-squared:  0.464,	Adjusted R-squared:  0.4544 
F-statistic: 48.05 on 4 and 222 DF,  p-value: < 2.2e-16
```

4.2.5 Stocks by End of Month

The full model has a R^2 of 0.7201, corn, rice, sugar and hops_extract explain 71.28%. It is worth mentioning that malt is not significant again. In this model to predict end-of-month stocks, corns and rice caught my eyes. They are both statistically significant and contribute to explain end-of-month stocks. In beer production, corns and rice are often added into malts, mostly for cost reduction. Although they do add flavors to beers, however, using large amounts of corns and/or rice is considered as “detrimental” to beers by most beer lovers. Incorporated with the context, the positive coefficients of corn and rice indicate that the stocks are positively correlated to the consumption of corns and rices. As there are more rices used in production, end-of-month stocks’ quantity will be higher. The same conclusion is for hops (as extract). Contrarily, as there are more sugar consumed, less amounts will be stocked in that month.

Stocks e-o-m	Corn	Rice	Sugar	Hops (extract)
Vif	1.315949	1.387869	1.166602	1.592967

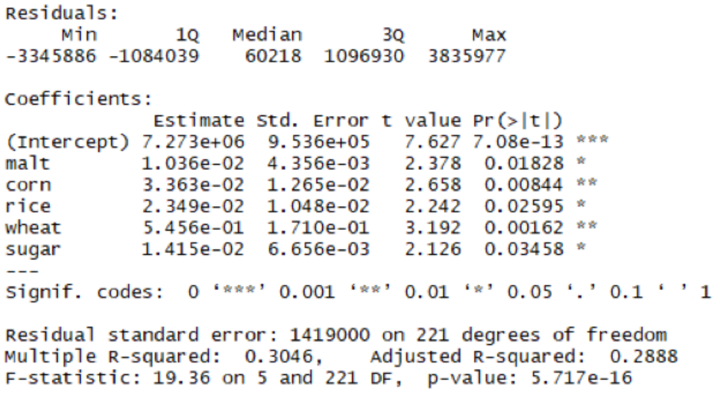


```
Residuals:
     Min       1Q   Median       3Q      Max 
-4061859  -419758   -73741   323842  3246531 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.094e+06  4.427e+05  13.766  < 2e-16 ***
corn          4.228e-02  5.454e-03   7.753 3.21e-13 ***
rice          3.203e-02  3.268e-03   9.799  < 2e-16 ***
sugar        -7.497e-03  2.813e-03  -2.665  0.00827 ** 
hops_extract  1.790e+00  2.689e-01   6.659 2.14e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 732300 on 222 degrees of freedom
Multiple R-squared:  0.7128,	Adjusted R-squared:  0.7076 
F-statistic: 137.7 on 4 and 222 DF,  p-value: < 2.2e-16
```

4.2.6 Consumption by Month

As calculated in 2.6, we obtain a new variable, Monthly Consumption. Since sales data are unavailable, we are unable to calculate the amounts of beer that are disposed for expiration. Therefore, we assume that there is no write down on inventory losses – all beer products are sold/consumed. According to the plots below, the full model is with R^2 of 0.3127, and now malt, corn, rice, wheat and sugar can explain 30.46% of consumption data.

Consumption	Malt	Corn	Rice	Wheat	Sugar
Vif	3.680033	1.885439	3.797245	2.268217	1.738312



```
Residuals:
     Min       1Q   Median       3Q      Max 
-3345886 -1084039    60218  1096930  3835977 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.273e+06  9.536e+05   7.627 7.08e-13 ***
malt        1.036e-02  4.356e-03   2.378  0.01828 *  
corn        3.363e-02  1.265e-02   2.658  0.00844 ** 
rice        2.349e-02  1.048e-02   2.242  0.02595 *  
wheat       5.456e-01  1.710e-01   3.192  0.00162 ** 
sugar       1.415e-02  6.656e-03   2.126  0.03458 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1419000 on 221 degrees of freedom
Multiple R-squared:  0.3046,	Adjusted R-squared:  0.2888 
F-statistic: 19.36 on 5 and 221 DF,  p-value: 5.717e-16
```

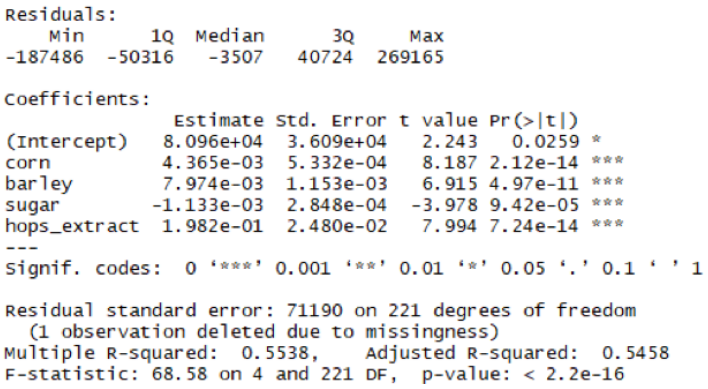
4.3 Further Regression Analysis with Lead and Lag Operators

There are many reasons that we should consider lead and lag operators in our models in order to improve prediction accuracy. For example, brewery takes time so that the relationship between production volumes and raw materials' consumption should be reexamined. Also, it takes time for producers to respond to overstock or understock. Therefore, further regressions will test on the lags and leads operators. One-month and two-month lags & leads are tested on all dependent variables.

Models with improved R^2 will be presented below. Although there are many reasons that lags and/or leads should be included in models, only two models eventually achieve higher R^2. Strong seasonality factor is one cause that most models are not improved with lead/lag, since one-month and two-month shifts will distort the seasonality pattern (same for 3 months and 6 months). Additionally, 12-month lags don't work either because of farness.

4.3.1 One-Month Leading in Export Volume

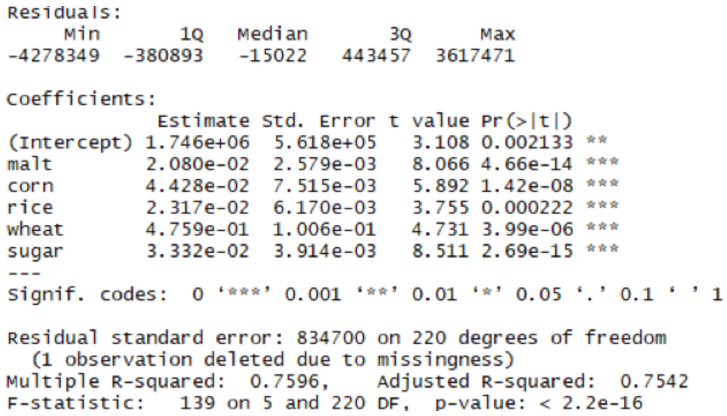
Export volume's model is improved with a lead operator on export. From the previous R^2 of 0.464, the same four independent variables give a new R^2 of 0.554, increased by 0.09. Therefore, we can say that the current consumption of corn, barley, sugar and hops as extract correlate better to the export volume in the next month than that in the current month.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-187486  -50316   -3507   40724  269165 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.096e+04  3.609e+04   2.243   0.0259 *  
corn          4.365e-03  5.332e-04   8.187 2.12e-14 ***
barley        7.974e-03  1.153e-03   6.915 4.97e-11 ***
sugar        -1.133e-03  2.848e-04  -3.978 9.42e-05 ***
hops_extract  1.982e-01  2.480e-02   7.994 7.24e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71190 on 221 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5538,	Adjusted R-squared:  0.5458 
F-statistic: 68.58 on 4 and 221 DF,  p-value: < 2.2e-16
```

4.3.2 One-Month Lagging in Consumption by Month

Taking a lag of monthly consumption variable significantly improves its model. Still, using stepwise model first and then referring to VIF and p-values to reduce the number of variables, the same five independent variables as the model in 4.2.6 give a R^2 of 0.7596! The lag operator improves the model from 0.3046. There is an increase over 0.45 in R^2. This result states that raw materials' consumption in the future month correlates to the current beer consumption. Conversely, the current month's beer consumption will influence producers' decision making in the material expenditure for next month.



```
Residuals:
     Min       1Q   Median       3Q      Max
-4278349  -380893   -15022   443457  3617471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.746e+06  5.618e+05   3.108 0.002133 **
malt        2.080e-02  2.579e-03   8.066 4.66e-14 ***
corn        4.428e-02  7.515e-03   5.892 1.42e-08 ***
rice        2.317e-02  6.170e-03   3.755 0.000222 ***
wheat       4.759e-01  1.006e-01   4.731 3.99e-06 ***
sugar       3.332e-02  3.914e-03   8.511 2.69e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 834700 on 220 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7596,    Adjusted R-squared:  0.7542
F-statistic:   139 on 5 and 220 DF,  p-value: < 2.2e-16
```

5 All Beers and Reviews

From this chapter, we will investigate beers from tasters' aspect. What does a beer look like? How does it smell? How does it taste? How does it feel? How do you like it overall? In Beer Judge program, tasters have to give five scores to these five questions. A great beer that give people rich and pleasant tasting experiences often achieves a score above 4 out of 5. The following content will reveal more about beer characteristics and how these characteristics influence drinking experiences.

We will first look into popular beer styles in United States. Next, we will focus on the ABV of beers. T-tests and anova tests will be used to interpret differences of means in groups of two and more than two. In addition, based on the scores given by beer lovers, we will examine the best and the worst beer families. In the end, I will use score data to predict the fermentation method of a beer. The goal of this chapter is to explore professional beer lovers' preferences, especially American drinkers.

5.1 Intro to All Beer Data

The dataset is downloaded from Kaggle, and the source of the data is from BeerAdvocate.com. This website is founded in Boston and attracts tons of beer lovers all over the world to add beers and reviews information to the website, although most contributors are English speakers living in North America. Therefore, the dataset is not inclusive: more than 70% beers in the data are from United States. As well, most reviews are given by American tasters. Therefore, from the data, we will primarily understand about American drinkers' beer taste rather than worldwide drinkers.

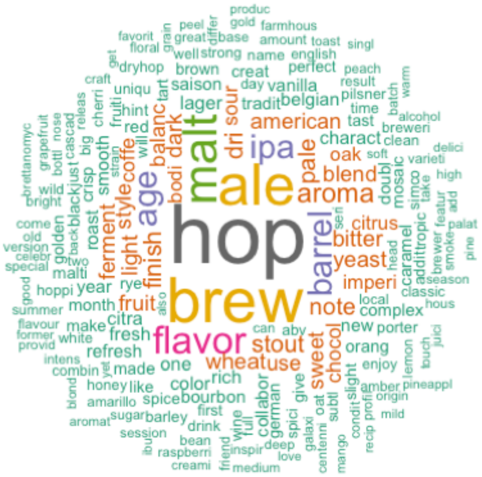
The Beer data include variables of beer' name, style, ABV value, availabilty, country, state, retired or not. According to the column of beer style and BJCP (Beer Judge Certification Program), I created a new column called "general style" (E.g. both American IPA and Belgian IPA go into the category of IPA). As well, since different styles of beers are assigned with different types of cups, another column of "Cup" has been created for analysis. In addition, new columns of "Country of Origin" and "Fermentation Method" have been generated according to style information of beers. I generate these new variables in order to understand how they relate to review scores and ABV of beers.

The Review data will be joined with Beer data. In review data, people give scores on each beer. The scores include 5 distinguished scores of look, smell, taste, feel and overall. There is also a final score that is calculated as the average of 5 scores.

5.2 Text Analysis Task on Consumers' Review Notes

In the Review data, in addition to scores, which range from 1.0 to 5.0, some tasters left comments. In the process of generating a word cloud, I obtained a frequency table for words appearing in comments. The Figure 18 shows the word cloud. The larger the word in this graph, the more times that the word appeared in comments. Therefore, we can see that beer tasters comment mostly on hop, malt, aged (stem-trimmed to age) beers, flavor, barrel, ipa, etc. This graph shows that American tasters care very much and discuss a lot about hops of beers.

Figure 18

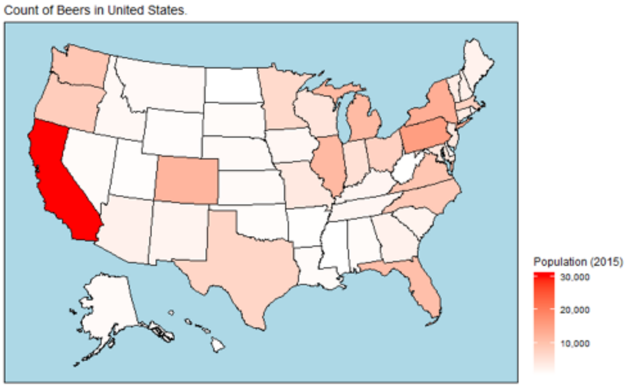


5.3 Beers by State in U.S.

The graph below reflects the number of beer in each state, the redder the color is, the more kinds of beers are produced in that state. According to the graph, Califonia is distinguishedly red: CA has more than 30,000 beers; the second place is PA, which has more 15,000+ beers.

To understand the most popular beer styles in each state, I grouped all beers by both style and state and sliced the top three beer styles in each state. There are in total 112 styles in the

Figure 19



Beer data. After groupby and selection, there are only 8 styles present to the top-three table. First of all, the most popular beer in 50 states and DC is American IPA, which appears 49 times; the other two do not have American IPA in their first place are DC and OK: Belgian Saison is the most common beer in DC and American Imperial Stout is the most common beer in OK. Next, I checked out three most common beers in each state. American IPA is in the list of every state and DC: American IPA has 51 times of appearances. The second popular beer is American Pale Ale, which makes 36 times of appearances. The following beers are American Imperial IPA (31 times), Belgian Saison (18 times), American Wild Ale (6 times), American Imperial Stout (6 times), American Amber / Red Ale (4 times), Berliner Weisse (1 time). Also, from this list, we can see that except for American beers, only Belgian Saison and Berliner Weisse are in. Berliner Weisse, only 1 time of appearance, is the second most popular beer in Florida. Belgian Saison are present 18 times. Belgian Saison is a pale ale beer, of high carbonation and with fruity taste.

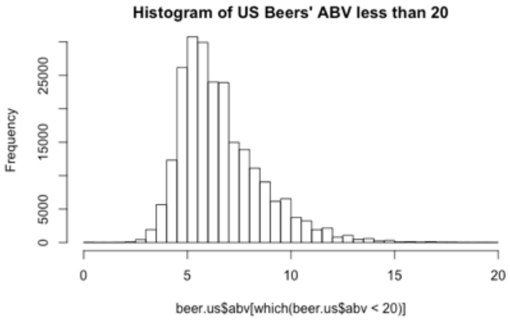
5.4 ABV of Beers

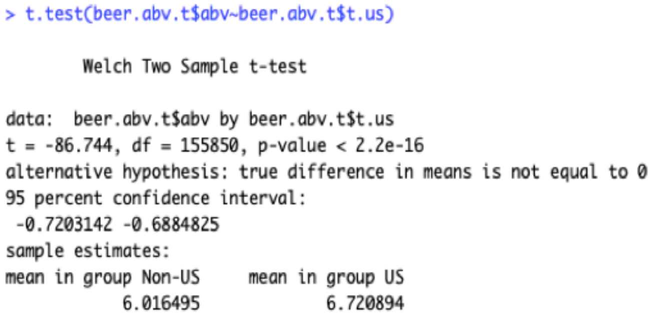
ABV represents Alcohol by Volume. It reflects the strength of ethyl alcohol. For ABV, BJCP gave four categories: session-strength (<4%), standard strength (4%-6%), high strength (6%-9%) and very-high-strength(>9%).

5.4.1 ABV of Beers in U.S. and outside U.S.

(ABV table)	Mean	Median	SD
Overall	6.53%	6%	2.055
U.S.	6.72%	6.3%	2.018
Non-U.S.	6.016%	5.5%	2.065

From the stat table above, we can see that beers in United States have a slightly higher average ABV than non-US areas. T-test result shows that the difference in means between these two groups are significant.




```
> t.test(beer.abv.t$abv~beer.abv.t$t.us)

	Welch Two Sample t-test

data:  beer.abv.t$abv by beer.abv.t$t.us
t = -86.744, df = 155850, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7203142 -0.6884825
sample estimates:
mean in group Non-US     mean in group US
            6.016495             6.720894
```

5.4.2 ABV by Cups

Beers are usually assigned to different shapes of cups based on styles and itself charateristics. There are in general 9 types of beer cups: Flute Glass, Goblet, Pilsner Glass, Pint, Snifter, Stange, Stein, Tulip and Weizen Glass. Grouped by the type of glass, the boxplot of abv values are as below.

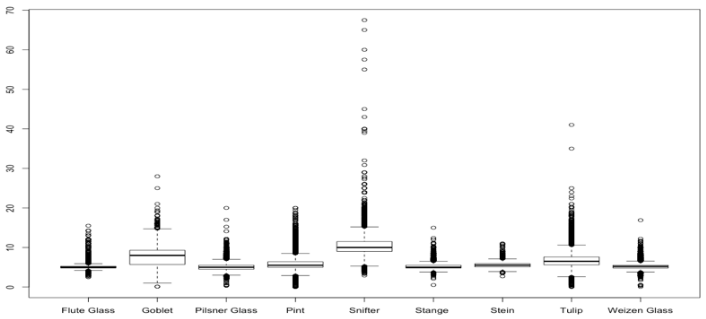
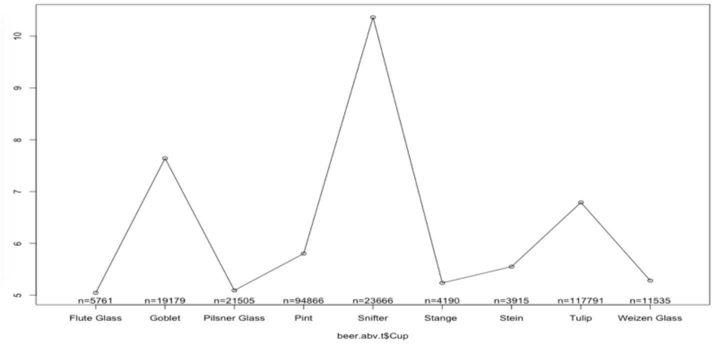
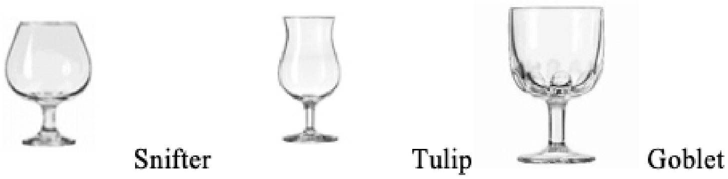


Figure 20



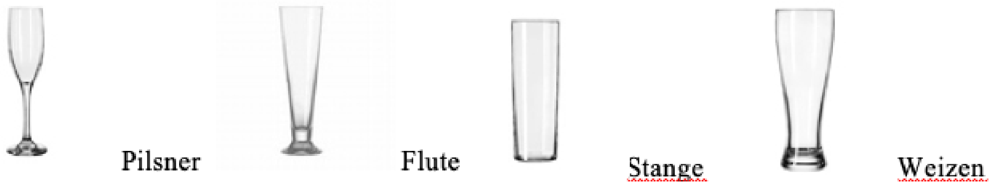
We see that for Goblet, Snifter and Tulip, beers' abv are significantly higher than the others. The shapes of these three cups are in the plots below. Surprisingly, they are all wide and stemmed. Considering that the warmth of hand will accelerate the metaphosis of beers, especially for those with high abv (alcohol evaporates), glass stems will avoid this problem.

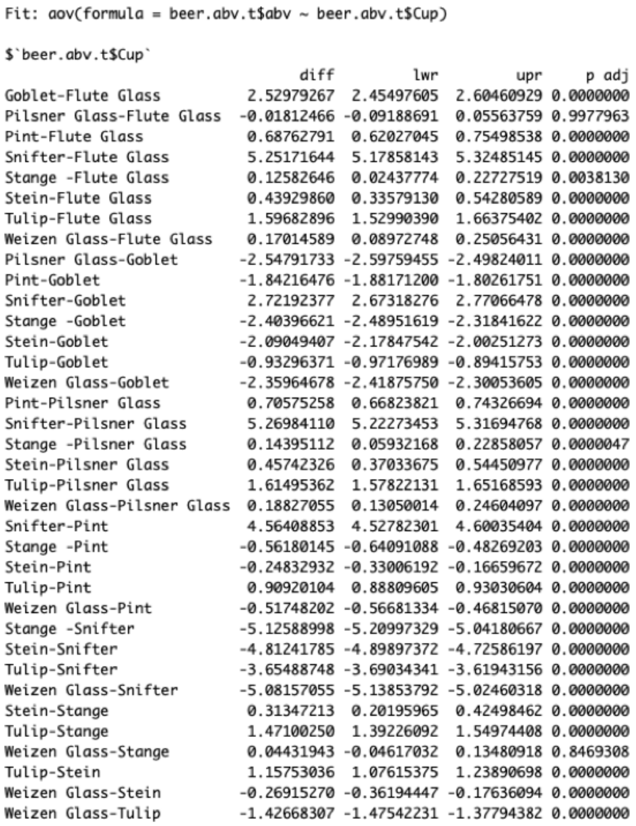
Figure 21



Next, let us look at the anova table by doing TukeyHSD. There are two comparisons that do not reject null hypothesis. They Pilsner Glass-Flute Glass and Weizen Glass-Stange. Again, from plots below, we find similarities between cups in these two groups. Both Pilsner and Flute Glasses can showcase carbonation and enhance volatiles in beers. In addition, Stange is special German beer glass for German altbier and German kolsch. Both of these beers have similar ABV with wheat beers that use Weizen Glass.

Figure 22

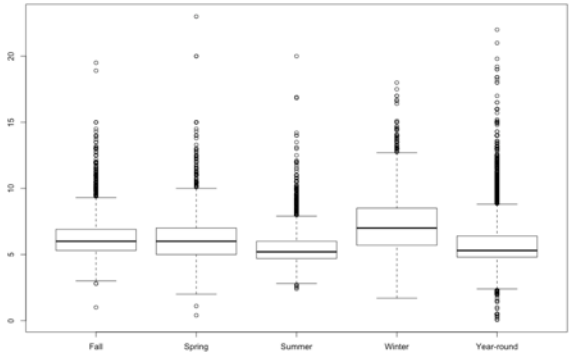


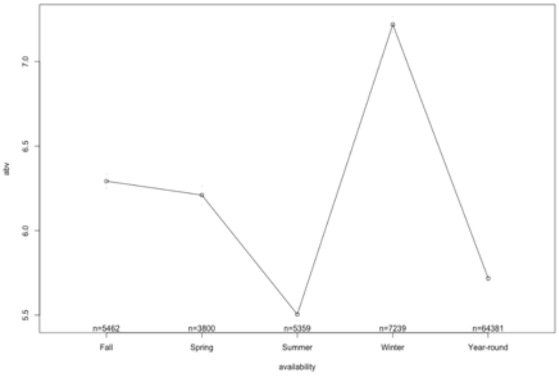


```
Fit: aov(formula = beer.abv.t$abv ~ beer.abv.t$Cup)

$`beer.abv.t$Cup`
                                  diff         lwr         upr     p adj
Goblet-Flute Glass          2.52979267  2.45497605  2.60460929 0.0000000
Pilsner Glass-Flute Glass  -0.01812466 -0.09188691  0.05563759 0.9977963
Pint-Flute Glass            0.68762791  0.62027045  0.75498538 0.0000000
Snifter-Flute Glass         5.25171644  5.17858143  5.32485145 0.0000000
Stange -Flute Glass         0.12582646  0.02437774  0.22727519 0.0038130
Stein-Flute Glass           0.43929860  0.33579130  0.54280589 0.0000000
Tulip-Flute Glass           1.59682896  1.52990390  1.66375402 0.0000000
Weizen Glass-Flute Glass    0.17014589  0.08972748  0.25056431 0.0000000
Pilsner Glass-Goblet       -2.54791733 -2.59759455 -2.49824011 0.0000000
Pint-Goblet                -1.84216476 -1.88171200 -1.80261751 0.0000000
Snifter-Goblet              2.72192377  2.67318276  2.77066478 0.0000000
Stange -Goblet             -2.40396621 -2.48951619 -2.31841622 0.0000000
Stein-Goblet               -2.09049407 -2.17847542 -2.00251273 0.0000000
Tulip-Goblet               -0.93296371 -0.97176989 -0.89415753 0.0000000
Weizen Glass-Goblet        -2.35964678 -2.41875750 -2.30053605 0.0000000
Pint-Pilsner Glass          0.70575258  0.66823821  0.74326694 0.0000000
Snifter-Pilsner Glass       5.26984110  5.22273453  5.31694768 0.0000000
Stange -Pilsner Glass       0.14395112  0.05932168  0.22858057 0.0000047
Stein-Pilsner Glass         0.45742326  0.37033675  0.54450977 0.0000000
Tulip-Pilsner Glass         1.61495362  1.57822131  1.65168593 0.0000000
Weizen Glass-Pilsner Glass  0.18827055  0.13050014  0.24604097 0.0000000
Snifter-Pint                4.56408853  4.52782301  4.60035404 0.0000000
Stange -Pint               -0.56180145 -0.64091088 -0.48269203 0.0000000
Stein-Pint                 -0.24832932 -0.33006192 -0.16659672 0.0000000
Tulip-Pint                  0.90920104  0.88809605  0.93030604 0.0000000
Weizen Glass-Pint          -0.51748202 -0.56681334 -0.46815070 0.0000000
Stange -Snifter            -5.12588998 -5.20997329 -5.04180667 0.0000000
Stein-Snifter              -4.81241785 -4.89897372 -4.72586197 0.0000000
Tulip-Snifter              -3.65488748 -3.69034341 -3.61943156 0.0000000
Weizen Glass-Snifter       -5.08157055 -5.13853792 -5.02460318 0.0000000
Stein-Stange                0.31347213  0.20195965  0.42498462 0.0000000
Tulip-Stange                1.47100250  1.39226092  1.54974408 0.0000000
Weizen Glass-Stange         0.04431943 -0.04617032  0.13480918 0.8469308
Tulip-Stein                 1.15753036  1.07615375  1.23890698 0.0000000
Weizen Glass-Stein         -0.26915270 -0.36194447 -0.17636094 0.0000000
Weizen Glass-Tulip         -1.42668307 -1.47542231 -1.37794382 0.0000000
```

5.4.3 ABV by Availability

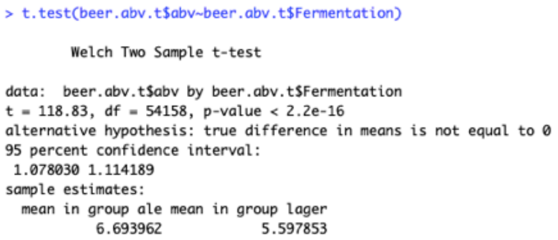
Availability refers to the seasonal, year-round, rotational and only-brewed-once production. In this chunk of analysis, we only include four seasons and year-round beers. Summer has on average lower ABV than Winter does, and there are also more kinds of beers in Winter than in Summer.





5.4.4 ABV by Fermentation Method

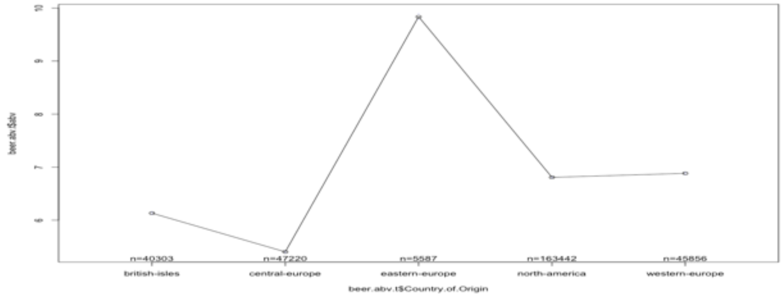
Beers can be classified into two groups: lager and ale. Lager beers are bottom-fermented and ales are top-fermented. Not considering beers involved in complecities in fermentation methods, I reclassify beers into lager and ale just by their styles. In graoh, I perform the t-test to compare the average of ABV between lagers and ales. The null hypothesis is rejected. On average, ales are 1 degree higher than lagers in ABV.



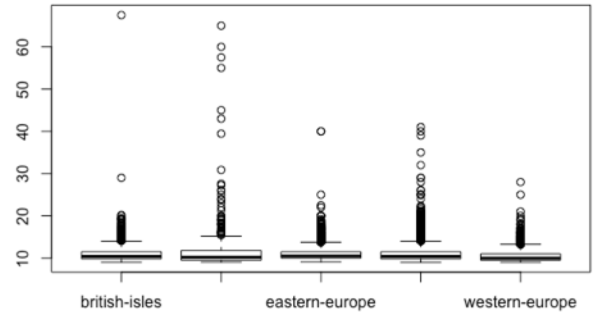
```
> t.test(beer.abv.t$abv~beer.abv.t$Fermentation)

	Welch Two Sample t-test

data:  beer.abv.t$abv by beer.abv.t$Fermentation
t = 118.83, df = 54158, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.078030 1.114189
sample estimates:
  mean in group ale mean in group lager
           6.693962            5.597853
```

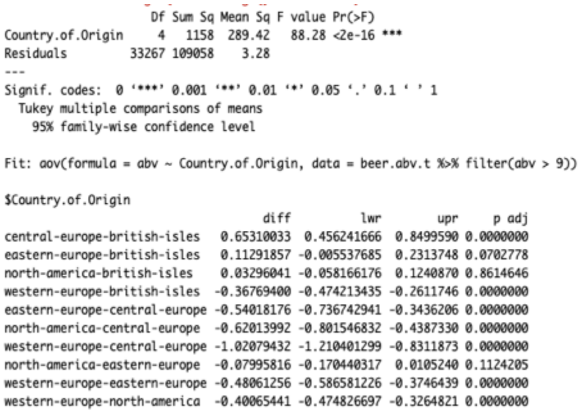
5.4.5 ABV by Country of Origin

By country of origin, most beers can be classified into five groups: North-America, Central-Europe, Eastern Europe, Western Europe and British-Isles. The ABV of beers by country of origin as in Graph 5.7, eastern europe has on average highest ABV and central europe has the lowest. Anova tests show that all differences in means are significant by rejecting all null hypothesis.



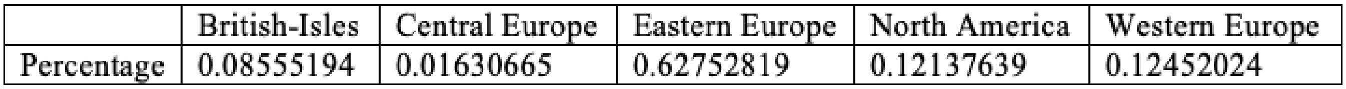
Next, let us filter out beers under ABV of 9%. In Beer Judge Certificate Program standard, beers above 9% are of very-high strength. Therefore, I want to look at the percentage of strong beers in each area. Still, eastern europe has the highest average ABV for beers above 9%, although there are a few insanely high ABV beers in central Europe. The Tukey Test reject three null hypotheses between north American and British Isles, between eastern Europe and British Isles, and between north American and eastern Europe at a confidence level of 95%.





```
                   Df Sum Sq Mean Sq F value Pr(>F)
Country.of.Origin     4   1158  289.42   88.28 <2e-16 ***
Residuals         33267 109058    3.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = abv ~ Country.of.Origin, data = beer.abv.t %>% filter(abv > 9))

$Country.of.Origin
                                   diff          lwr        upr     p adj
central-europe-british-isles   0.65310033  0.456241666  0.8499590 0.0000000
eastern-europe-british-isles   0.11291857 -0.005537685  0.2313748 0.0702778
north-america-british-isles    0.03296041 -0.058166176  0.1240870 0.8614646
western-europe-british-isles  -0.36769400 -0.474213435 -0.2611746 0.0000000
eastern-europe-central-europe -0.54018176 -0.736742941 -0.3436206 0.0000000
north-america-central-europe  -0.62013992 -0.801546832 -0.4387330 0.0000000
western-europe-central-europe -1.02079432 -1.210401299 -0.8311873 0.0000000
north-america-eastern-europe  -0.07995816 -0.170440317  0.0105240 0.1124205
western-europe-eastern-europe -0.48061256 -0.586581226 -0.3746439 0.0000000
western-europe-north-america  -0.40065441 -0.474826697 -0.3264821 0.0000000
```

To calculate the proportion of strong beers in each area, we obtain the following table. Eastern Europe has extremely high proportion of very-high-strength beers, most are Russian styles. Central Europe has the widest span in beer ABV but the lowest proportion in strong-strength beers. Therefore, we believe that most styles coming from eastern Europe are in strong strength, and most styles from central Europe have relatively low ABV.



	British-Isles	Central Europe	Eastern Europe	North America	Western Europe
Percentage	0.08555194	0.01630665	0.62752819	0.12137639	0.12452024

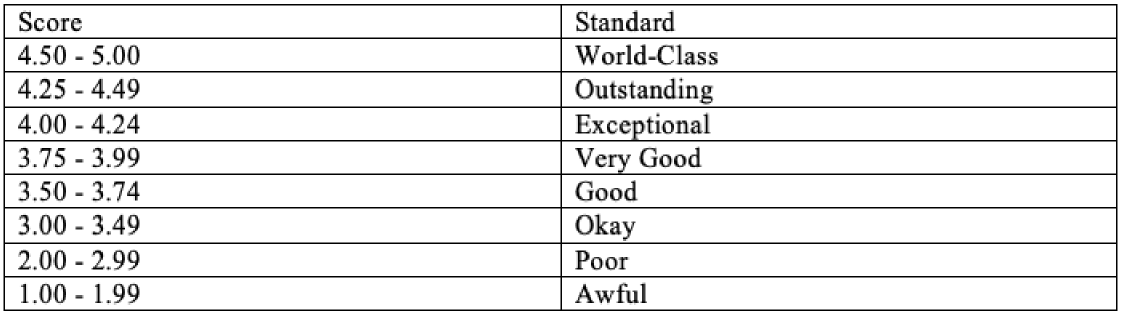
5.4.6 ABV by Styles and Families

Based on 112 specific styles, the strongest strength style is German Eisbock, which has an average of 13.92% ABV. Two styles have means lower than 4%, session-strength. They are Low Alcohol Beer and Russian Kvass.

According to 112 specific styles, I further generalize these 112 styles into 15 families based on style family in BJCP handbook. They are ipa-family, brown-ale-family, pale-ale-family, pale-lager-family, pilsner-family, amber-ale-family, amber-lager-family, dark-lager-family, porter-family, stout-family, bock-family, strong-ale-family, wheat-beer-family, specialty-beer. In this new category, wheat beer family has the lowest mean of 5.16% ABV; Strong ale family has the highest mean of 9.55% ABV and it's the only family whose average ABV falls in the very-strong-strength category.

5.5 Review Scores

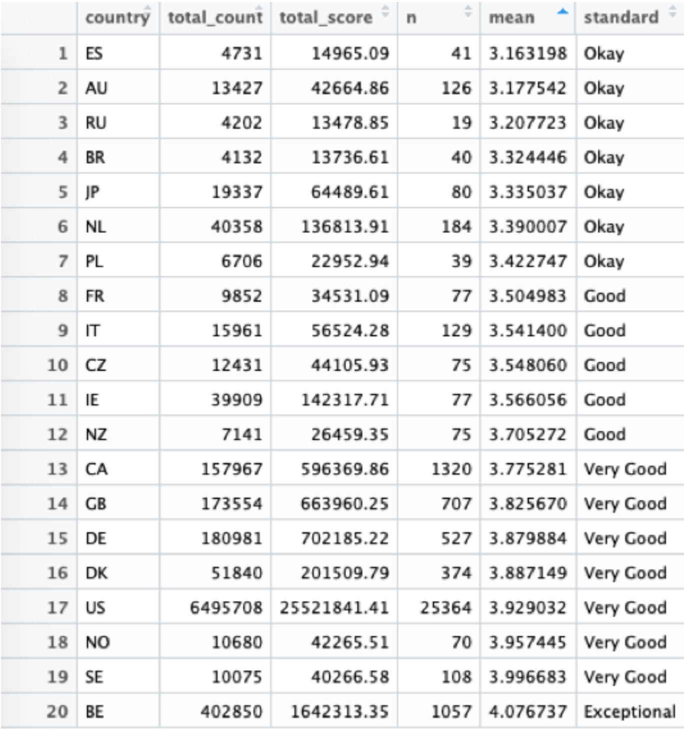
By joining the Beer Data and Review Data, we obtain a new table where beers have scores. However, not all beers have scores; it is similar to the fact that not all restaurants on yelp have scores. To maintain a fair value of score for each beer, we only include beers that have larger than or equal to 30 reviews. Eventually, 30,981 beers (out of 320,072) are qualified for later analysis. In addition, there is a standard table for review scores:



Score	Standard
4.50 - 5.00	World-Class
4.25 - 4.49	Outstanding
4.00 - 4.24	Exceptional
3.75 - 3.99	Very Good
3.50 - 3.74	Good
3.00 - 3.49	Okay
2.00 - 2.99	Poor
1.00 - 1.99	Awful

5.5.1 The Best Beer Producing Nation

Here we select 20 countries that have most kinds of beers from the Beer data. In the table below, "total_count" represents the the total number of reviews each country's beers have; "total_score" is the sum of all reviews' final scores; "n" is the number of beers considered (each beer has to have at least 30 reviews to be counted in); "mean" is the average scores of beers in that country; "standard" comments on beers by final score. From this table, we can see that only Belgian beers reach an "Exceptional" score on average, the highest standard in U.S. beer consumers. As well, both Northern European and Northern American beers are generally liked very much.



	country	total_count	total_score	n	mean	standard
1	ES	4731	14965.09	41	3.163198	Okay
2	AU	13427	42664.86	126	3.177542	Okay
3	RU	4202	13478.85	19	3.207723	Okay
4	BR	4132	13736.61	40	3.324446	Okay
5	JP	19337	64489.61	80	3.335037	Okay
6	NL	40358	136813.91	184	3.390007	Okay
7	PL	6706	22952.94	39	3.422747	Okay
8	FR	9852	34531.09	77	3.504983	Good
9	IT	15961	56524.28	129	3.541400	Good
10	CZ	12431	44105.93	75	3.548060	Good
11	IE	39909	142317.71	77	3.566056	Good
12	NZ	7141	26459.35	75	3.705272	Good
13	CA	157967	596369.86	1320	3.775281	Very Good
14	GB	173554	663960.25	707	3.825670	Very Good
15	DE	180981	702185.22	527	3.879884	Very Good
16	DK	51840	201509.79	374	3.887149	Very Good
17	US	6495708	25521841.41	25364	3.929032	Very Good
18	NO	10680	42265.51	70	3.957445	Very Good
19	SE	10075	40266.58	108	3.996683	Very Good
20	BE	402850	1642313.35	1057	4.076737	Exceptional

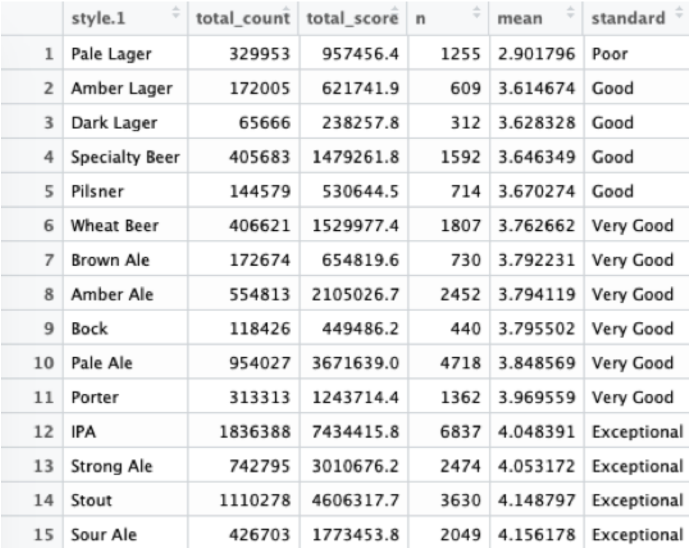
5.5.2 Horrible Lager?

When people mention bad beers, most often lagers are brought up. Sometimes, people consider “lager” as the representative of bad beers that can be boring, flavorless and watery. The table below demonstrates the average impression of beers by style families. The three groups with least average scores are all LAGER! Pale Lager even receives a score for “Poor”. Yes, lagers are generally considered bad; however, Bock is also a lager beer but has a score of “Very Good”, although many people do not know that Bock is a lager. However, in contrast to pale lager, Bock is dark-colored and has a strong alcohol strength. Remember the strongest strength beer from 5.4.6, the German Eisbock? It is one style in the Bock family.

Next, let us take a look at “Exceptional” beers. They are Sour Ale, Stout, Strong Ale and

IPA. They are all ale beers, top-fermented. From this table, apparently, ales receive higher scores than lagers. The common places among these exceptional beers are that they are all full of flavor and aroma. IPA is bitter and usually fruity; Strong Ale are strong in ABV; Sour ale tastes properly sour; Stout, usually with creamy hops, uses roasted malt to produce chocolate or coffee flavors.

Overall, Lager beers indeed have a bad reputation, especially Pale Lager. Ales are more favored by people. However, the theory of absolutely bad lager should be doubted. Instead, people appreciate stronger, more savory and aromatic beers. Complex taste experiences add values.



	style.1	total_count	total_score	n	mean	standard
1	Pale Lager	329953	957456.4	1255	2.901796	Poor
2	Amber Lager	172005	621741.9	609	3.614674	Good
3	Dark Lager	65666	238257.8	312	3.628328	Good
4	Specialty Beer	405683	1479261.8	1592	3.646349	Good
5	Pilsner	144579	530644.5	714	3.670274	Good
6	Wheat Beer	406621	1529977.4	1807	3.762662	Very Good
7	Brown Ale	172674	654819.6	730	3.792231	Very Good
8	Amber Ale	554813	2105026.7	2452	3.794119	Very Good
9	Bock	118426	449486.2	440	3.795502	Very Good
10	Pale Ale	954027	3671639.0	4718	3.848569	Very Good
11	Porter	313313	1243714.4	1362	3.969559	Very Good
12	IPA	1836388	7434415.8	6837	4.048391	Exceptional
13	Strong Ale	742795	3010676.2	2474	4.053172	Exceptional
14	Stout	1110278	4606317.7	3630	4.148797	Exceptional
15	Sour Ale	426703	1773453.8	2049	4.156178	Exceptional

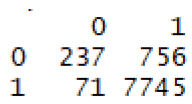
5.6 Predicting Fermentation Method by Scores

Can we predict a lager, bottom-fermented, or an ale, top-fermented, by inputting its scores and ABV value? Using three models, we will predict the fermentation method using the ABV from the Beer Data and the scores of look, smell, taste and feel from the Review Data. The 5-folds cross validation is applied in order to prevent overfitting. In the end, we want to pick the best model that can yield the highest prediction accuracy rate.

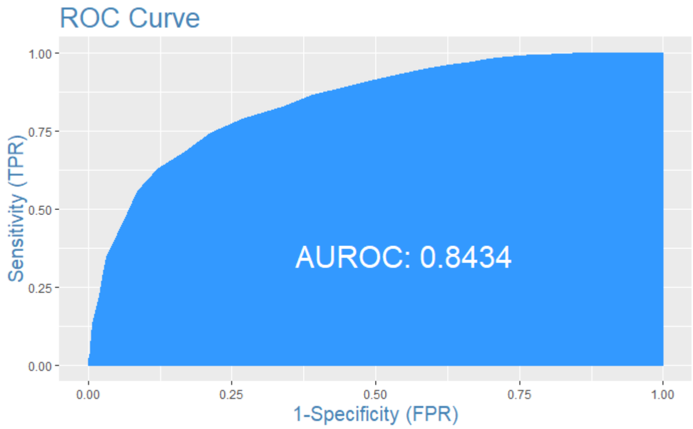
5.6.1 Bayesian Logistic Regression

In the R-package "caret", the function of "train" supports a cross-validated model on Bayesian Logistic Regression. The ROC plot is as below, with a value of 0.8434. Based on the confusion matrix below, the accuracy rate is 90.61%.

Figure 23

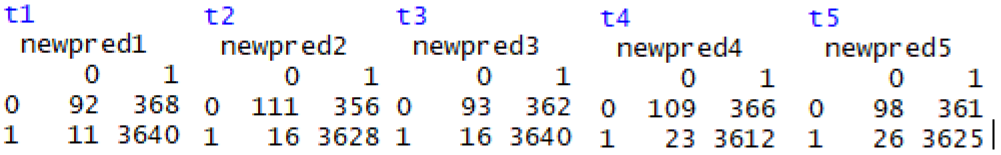


```
     0    1
0  237  756
1   71 7745
```



5.6.2 SVM with Radial Kernel in Classification

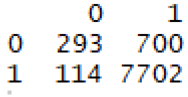
Another method is the Support Vector Machine. The kernel is radial. By patitioning the training set into five-folds and then running a SVM model on each fold, the confusion matrices render five close accuracy rates. They are 0.9078, 0.9095, 0.9081, 0.9054, 0.9058, in order. The confusion matrix of each model is in the figure below. The average accuracy rate is 0.9073. Therefore, we choose the first trained SVM model because its rate is the closest to the averaged rate. Eventually, applying the trained model on the test set gives an accuracy rate of 90.58%.



```
t1                t2                t3                t4                t5
 newpred1          newpred2          newpred3          newpred4          newpred5
      0    1            0    1            0    1            0    1            0    1
0    92  368   0  111  356   0    93  362   0  109  366   0    98  361
1    11 3640   1   16 3628   1    16 3640   1    23 3612   1    26 3625
```

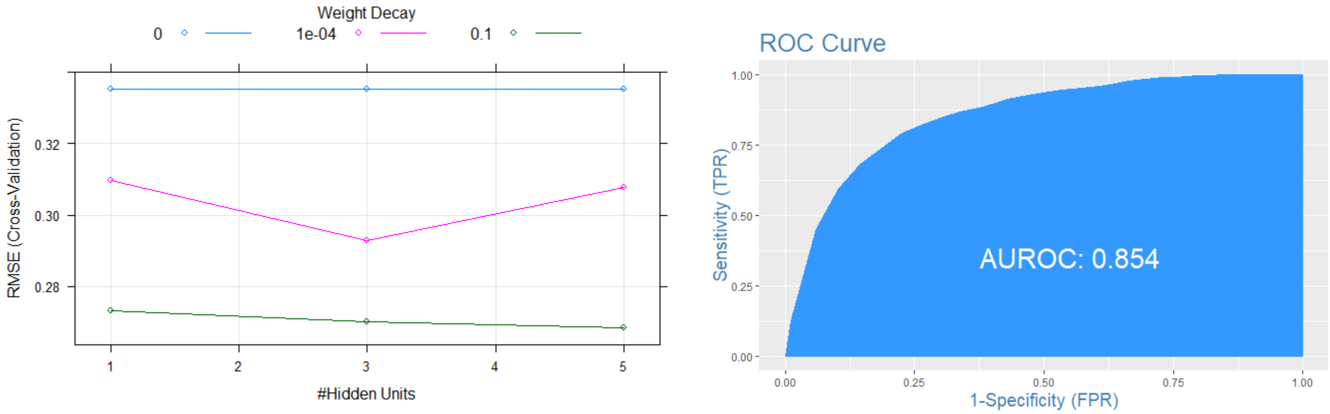
5.6.3 Neural Networks

The third method tested is Neural Networks with five hidden layers. Applying the trained NN onto the test set gives a confusion matrix as below. The ROC value is 0.854. The accuracy rate is 90.76%.



```
     0    1
0  293  700
1  114 7702
```

Figure 24



Although with slightly different results, Neural Networks wins over Bayesian Logistic Regression by 0.15% in accuracy rate. Its ROC value is also slightly higher than that of Bayesian Logistic Regression. Both indicators show that the Neural Networks model performs better in understanding the fermentation method with given variables.

6 Conclusion

Till now, we have learned how many gallons of beer America produce every year and which raw materials do producers use to brew beers. We've seen that manufactures produce more beers during summer than winter. More, the total output per year is decreasing. The same for yearly beer consumption. From Chapter 3, we have understood that the decreasing consumption of malts, corns and rice account for the decreasing total output. Meanwhile, we must have known that the consumption volumes of wheat and barley are increasing. This shift in raw materials' consumption indicate changes happening in consumers' tastes: the trend for less lagers but more wheat beers and strong ales.

Next, from the data from BeerAdvoate.com, we found that although pale lager accounts for over 50% of production amount in United States, it ranks the lowest place in beer lovers' group. Budweiser, Corona Light, Heineken, Bud Light, etc. are all pale lagers. These famous and massively produced beers explain a large proportion of yearly production output and consumption. However, beer lovers hate them. In conclusion, from both manufactures' data and beer lovers' data, we found that pale lagers are ill-reputed. Therefore, if you want to order a great beer, avoid the omnipresent pale lagers and pick a craft beer. Take American IPA or American Imperial (Double) IPA, if you prefer a bitter taste; Take a Stout or a Porter if you like flavors of coffee, chocolate or smoke and a thick taste. Also, it is always worth a try for prominent Belgian beers. If you still want a lager beer because you like a clear body color, you can try Koelsch, Oktoberfestbier/Maerzen, German Bock, etc...

Lastly, even though this paper has showed you a lot about beers such as comparisons in ABV, fermentation method and scores, there are many beer characteristics not discussed here. IBU (how bitter a beer is), color, hops' amount and yeasts used are all significant in understanding a beer. If with efficient datasets, a few meaningful regression and classification tasks can be implemented to interpret these characteristics. For instance, we can use characteristic variables to predict an ultimate score or to classify beers by styles/families. Meanwhile, this classifier project can eventually help waffling people make up their minds in ordering a good beer.

A Reference

1. Wolfe, Edward, et al. “BJCP BEER EXAM STUDY GUIDE .” BJCP Program.

2. Nrc. “Beer Statistics.” TTB, www.ttb.gov/beer/beer-stats.shtml.

3. “Homebrew Beer Recipes — Brewer’s Friend.” Brewer’s Friend Beer Brewing Software, www.brewersfriend.com/homebrew-recipes/.

4. “Glassware for Beer.” BeerAdvocate, www.beeradvocate.com/beer/101/glassware/.

5. “Beer Styles.” BeerAdvocate, www.beeradvocate.com/beer/styles/.

6. What Is the Difference between Barley and Malt? - Quora. www.quora.com/What-is-the-difference-between-barley-and-Malt.

7. “Beers, Breweries, and Beer Reviews.” Kaggle, www.kaggle.com/Ehallmar/Beers-Breweries-and-Beer-Reviews.

8. “Brewer’s Friend Home Brewing Software.” Brewer’s Friend Beer Brewing Software, www.brewersfriend.com/allgrain-ogfg/.