Case Study 1

By Roslyn Smith, David George and Varun Gopal

This presentation investigates a few trends found in the Beers and Breweries dataset. We reviewed trends across states in the U.S as well as a deep dive on both ABV and IBU values. We also explored how ABV and IBU can predict whether or not a beer is an IPA or an Ale.

Importing packages

```
#Libraries
library(tidyverse)
library(caret)
library(class)
library(e1071)
library(maps)
library(mapproj)
library(plotly)
library(data.table)
library(formattable)
library(tidyr)
library(dplyr)
```

Importing datasets

```
beers=read.csv("Beers.csv")
breweries=read.csv("Breweries.csv")
```

Question 3: Filter out NAs in beers

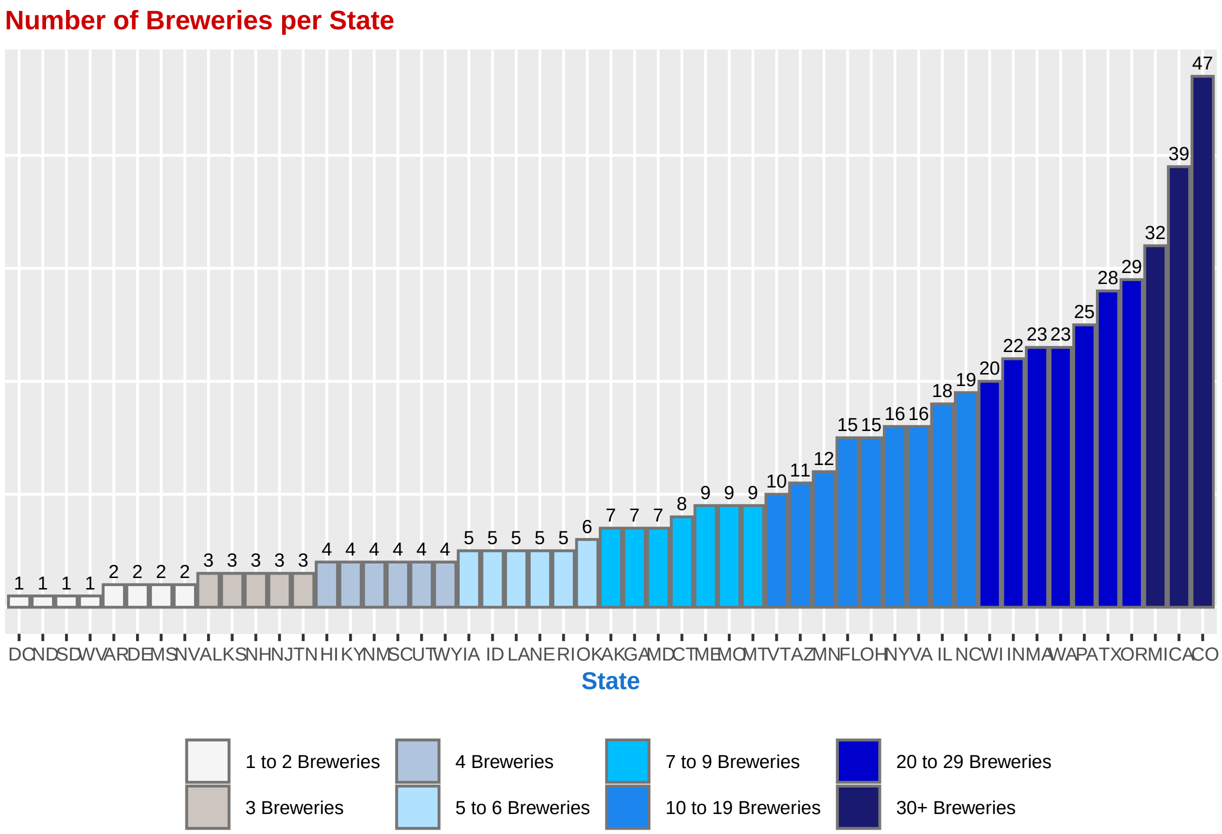
```
#Filter out NAs in beers
beersclean=beers%>%filter(!(is.na(IBU)|is.na(ABV)))
ABVclean=beers%>%filter(!is.na(ABV))
IBUclean=beers%>%filter(!is.na(IBU))
```

In general, we decided to delete the entire row if an IBU or an ABV was missing. However, there are some questions in which we only filtered out missing ABV values when investigating only ABV values so that we could keep more data. The same concept applied when investigating only IBU values.

Question 1: Find how many breweries there are per state

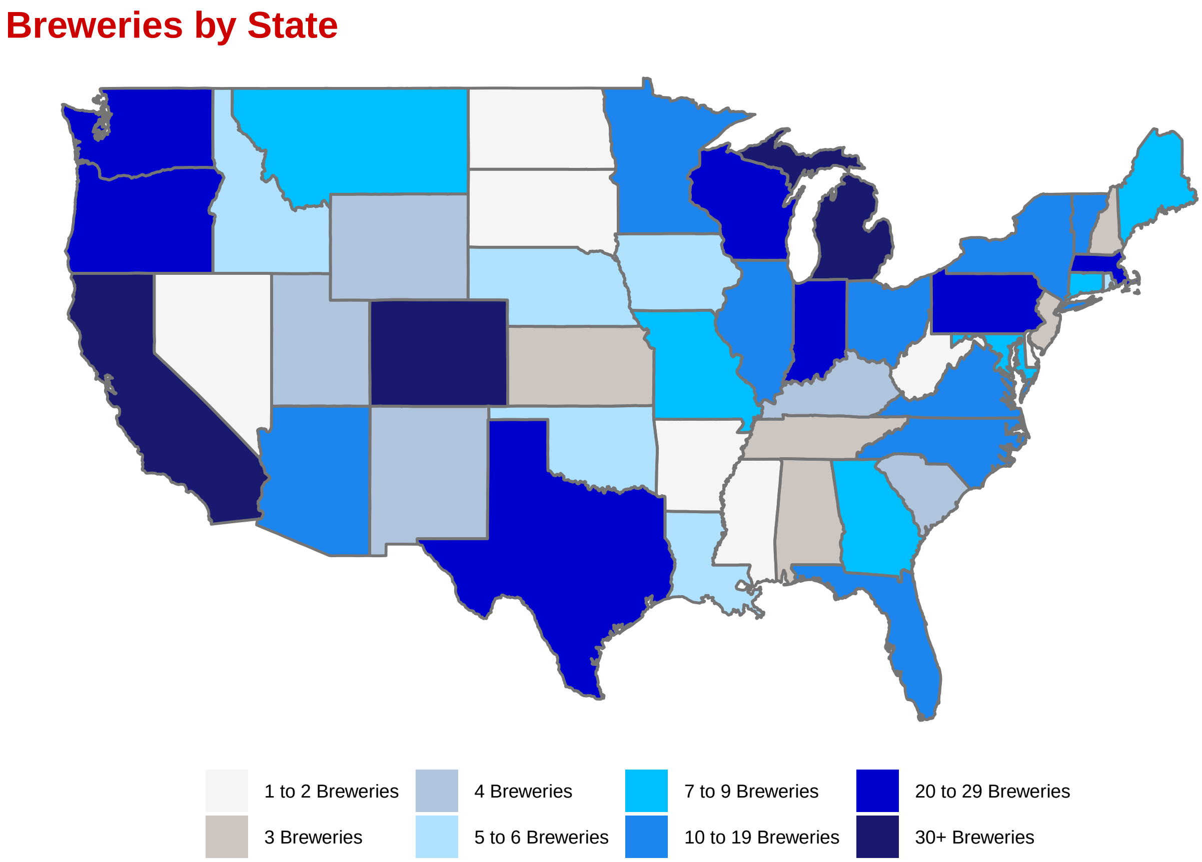
```
#Find how many breweries there are per state
# Grouped Breweries by volume
breweriesclean = breweries%>%group_by(State)%>%
  summarize(count=n())

breweriessummary = breweriesclean%>%
  mutate(Group = case_when(
```

```r
    between (count,1,2)~"1 to 2 Breweries",
    count ==3 ~"3 Breweries",
    count ==4 ~"4 Breweries",
    between (count,5,6)~"5 to 6 Breweries",
    between (count,7,9)~"7 to 9 Breweries",
    between (count,10,19)~"10 to 19 Breweries",
    between (count,20,29)~"20 to 29 Breweries",
    between (count,30,50)~"30+ Breweries"
  ))


#Barplot of Breweries in Each State.  Note: x=reorder, "-" before count orders decending, no "-" orders

breweriessummary%>%
  ggplot(aes(x=reorder(State, count),y=count, fill=Group))+
  geom_bar(stat='identity', color = "grey46")+
  geom_text(aes(label  = count), vjust = -0.5, size = 2.5, color = "black")+
  ggtitle('Number of Breweries per State')+
  xlab('State')+
  ylab('Number of Breweries')+
  scale_fill_manual("Group",values = c("1 to 2 Breweries" = "gray96"
                                       ,"3 Breweries" = "seashell3"
                                       ,"4 Breweries" = "lightsteelblue"
                                       ,"5 to 6 Breweries"  = "lightskyblue1"
                                       ,"7 to 9 Breweries" = "deepskyblue"
                                       ,"10 to 19 Breweries" = "dodgerblue2"
                                       ,"20 to 29 Breweries" = "blue3"
                                       ,"30+ Breweries" = "midnightblue") )+
  theme(legend.position = "bottom",
        legend.text = element_text(size=7),
        legend.title = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks.y = element_blank(),
        title = element_text(face="bold", color = "red3", size = 8),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
        axis.text.x = element_text(size = 7))+
  scale_y_continuous(minor_breaks = seq(0,50,10),breaks=seq(0,50,10))
```




```r
#Create Heat Map
FiftyStates=data.frame(State=state.abb,Name=state.name) #Create and Name columns of DF
brewstates=breweries%>%group_by(State)%>%
  summarize(count=n())
brewstates=data.frame(breweriessummary)
for (i in 1:dim(brewstates)[1]){#Make sure the string values of FiftyStates and brewstates match for th
  brewstates$State[i]=str_extract(brewstates$State[i],"\\b[A-Za-z]+\\b")
}
brewstates=merge(brewstates,FiftyStates,'State',all.x=TRUE)
brewmapdata=data.frame(region=tolower(brewstates$Name),Breweries=brewstates$count, State=brewstates$Sta

States=map_data('state')
map.df=merge(States,brewmapdata,"region",all.x=T)#Finalize Map Data
map.df=map.df[order(map.df$order),]

map.df%>%ggplot(aes(x=long,y=lat,group=group))+
  geom_polygon(aes(fill=Group))+
  geom_path(color = "grey46")+
  theme_void()+
  scale_fill_manual("Group",values = c("1 to 2 Breweries" = "gray96"
                                       ,"3 Breweries" = "seashell3"
                                       ,"4 Breweries" = "lightsteelblue"
                                       ,"5 to 6 Breweries"  = "lightskyblue1"
                                       ,"7 to 9 Breweries" = "deepskyblue"
                                       ,"10 to 19 Breweries" = "dodgerblue2"
                                       ,"20 to 29 Breweries" = "blue3"
```

```r
                                        ,"30+ Breweries" = "midnightblue"))+
  theme(legend.position = "bottom",
        legend.text = element_text(size=7),
        legend.title = element_blank(),
        title = element_text(face="bold", color = "red3", size = 12),
        axis.text = element_blank(),
        axis.title = element_blank(),
        axis.ticks = element_blank())+
  ggtitle('Breweries by State')
```



The barplot above shows the number of breweries for each state. There are three states that have 30 or more breweries and those are Michigan, California and Colorado. The heat map of the states also demonstrates the same trend.

Question 2: Merge beer and breweries datasets

```r
#Merge beer and breweries
beersclean$Brew_ID=beersclean$Brewery_id
beerbreweries=merge(beersclean,breweries,by="Brew_ID",all.x = TRUE)
beerbreweries=beerbreweries%>%select(!Brewery_id)

#Merge cleaned ABV data with breweries
ABVclean$Brew_ID=ABVclean$Brewery_id
ABVbreweries=merge(ABVclean,breweries,by="Brew_ID",all.x = TRUE)
ABVbreweries=ABVbreweries%>%select(!Brewery_id)
```

```
#Merge cleaned IBU data with breweries
IBUclean$Brew_ID=IBUclean$Brewery_id
IBUbreweries=merge(IBUclean,breweries,by="Brew_ID",all.x = TRUE)
IBUbreweries=IBUbreweries%>%select(!Brewery_id)

#Print first 6 and last 6 observations
head(beerbreweries)
```

```
##   Brew_ID        Name.x Beer_ID   ABV IBU                               Style Ounces             Name
## 1       1  Get Together    2692 0.045  50                        American IPA     16 NorthGate Brewi
## 2       1 Maggie's Leap    2691 0.049  26                  Milk / Sweet Stout     16 NorthGate Brewi
## 3       1    Wall's End    2690 0.048  19                   English Brown Ale     16 NorthGate Brewi
## 4       1       Pumpion    2689 0.060  38                         Pumpkin Ale     16 NorthGate Brewi
## 5       1    Stronghold    2688 0.060  25                     American Porter     16 NorthGate Brewi
## 6       1   Parapet ESB    2687 0.056  47 Extra Special / Strong Bitter (ESB)     16 NorthGate Brewi
```

```
tail(beerbreweries)
```

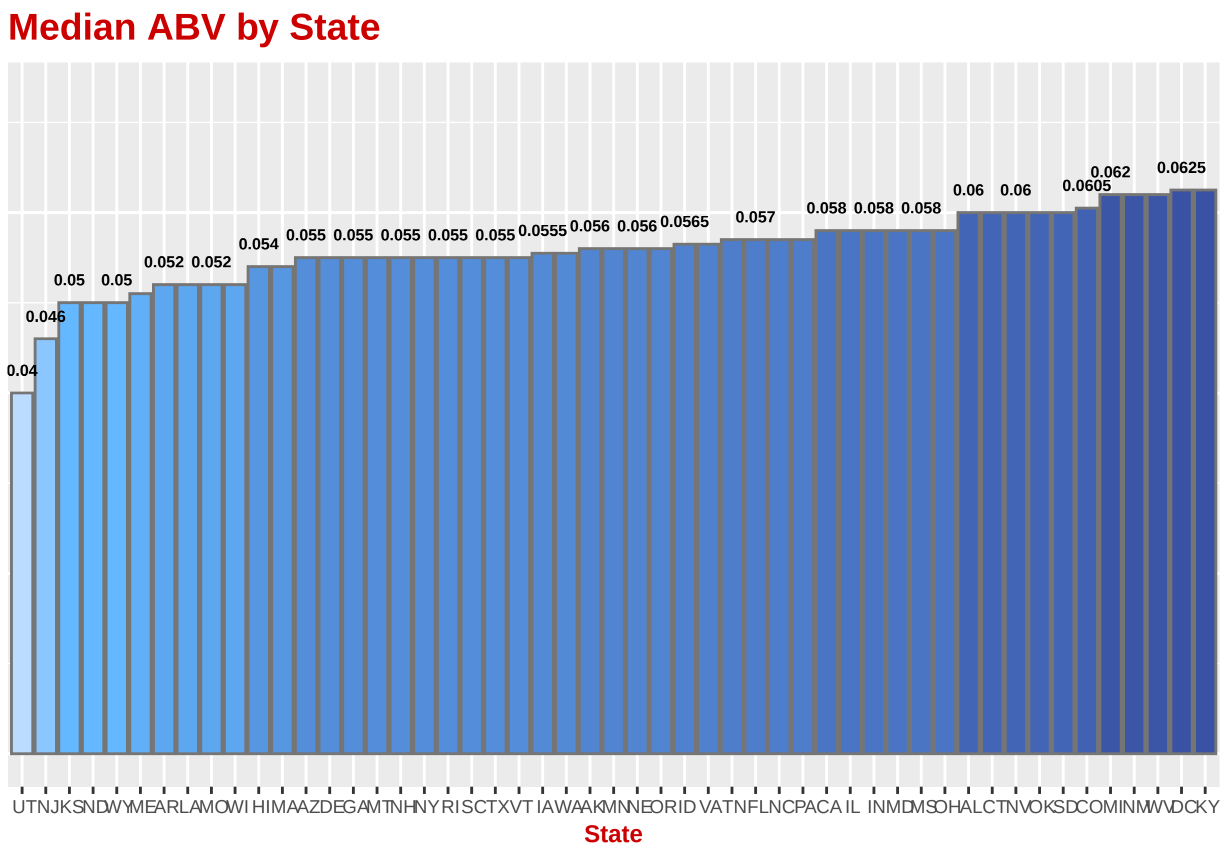
```
##      Brew_ID                           Name.x Beer_ID   ABV IBU                Style Ounces
## 1400     545        Pyramid Hefeweizen (2011)     399 0.052  18           Hefeweizen     12
## 1401     545        Haywire Hefeweizen (2010)      82 0.052  18           Hefeweizen     16
## 1402     546           Rumspringa Golden Bock     392 0.066  30 Maibock / Helles Bock     12 Lancaste
## 1403     546   Lancaster German Style KÃ¶lsch     195 0.048  28               KÃ¶lsch     12 Lancaste
## 1404     547 Common Sense Kentucky Common Ale     382 0.053  22    American Brown Ale     16   Upsta
## 1405     547                   Upstate I.P.W.     381 0.065  70         American IPA     12   Upsta
```

We merged the beers and breweries datasets by brewery id and by using a left join. We also merged the dataset containing non-NA ABV's and the breweries dataset for investigating ABV values only. The same concept applied when investigating only IBU values.

Question 4: Find median ABV and IBU per state

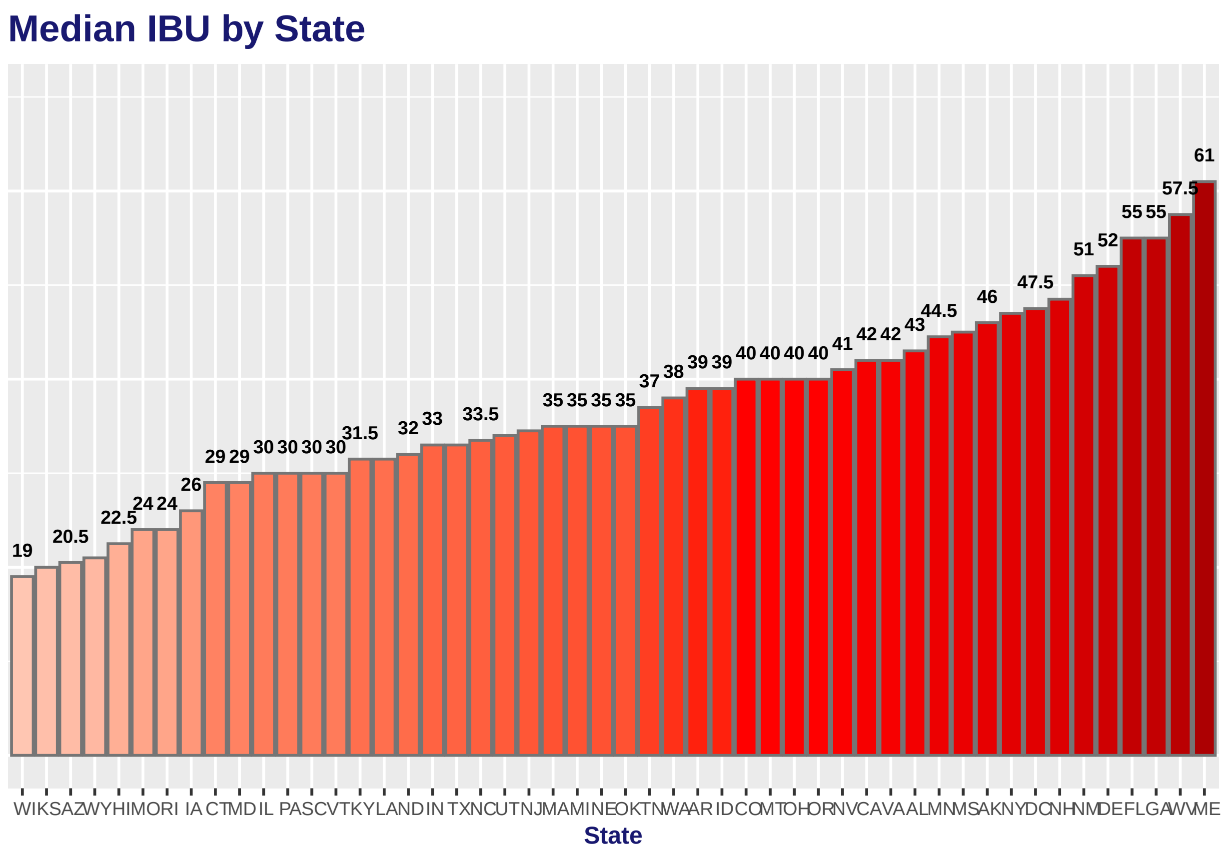
```
#Find median ABV and IBU per state
colnames(beerbreweries)[8]='Brewery_Name'
colnames(beerbreweries)[2]='Beer_Name'

colnames(ABVbreweries)[8]='Brewery_Name'
colnames(ABVbreweries)[2]='Beer_Name'

colnames(IBUbreweries)[8]='Brewery_Name'
colnames(IBUbreweries)[2]='Beer_Name'


#Gather median abv and ibu per each state
ABVIBUData=beerbreweries%>%group_by(State)%>%summarize(medianABV=median(ABV),medianIBU=median(IBU),coun
ABVData=ABVbreweries%>%group_by(State)%>%summarize(medianABV=median(ABV),count=n())
IBUData=IBUbreweries%>%group_by(State)%>%summarize(medianIBU=median(IBU),count=n())


#barplot ABV Updated

ABVData%>%
  ggplot(aes(x=reorder(State,medianABV),y=medianABV,fill=medianABV))+
```

```
  geom_bar(stat='identity', color = "grey46")+
  geom_text(aes(label  = medianABV), vjust = -1.5, size = 2.2, color = "black",fontface =  "bold",check_
  ylim(0,.073)+
  ggtitle('Median ABV by State')+
  xlab('State')+
  ylab('Median ABV')+
  scale_fill_gradient2(name='Median ABV',low = "white", mid = "steelblue1", high = "midnightblue",midpoi
                       breaks=c(0.03,0.038,0.046,0.054,0.062,0.07), na.value = "grey50")+
  theme(legend.position = "none",
        title = element_text(face="bold", color = "red3", size = 12),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.x = element_text(size = 7),
        axis.title.x = element_text(face="bold", color = "red3", size = 9))
```




```
#barplot IBU Updated
IBUData%>%ggplot(aes(x=reorder(State,medianIBU),y=medianIBU,fill=medianIBU))+
  geom_bar(stat='identity', color = "grey46")+
  geom_text(aes(label  = medianIBU), vjust = -1.5, size = 2.3, color = "black",fontface =  "bold",check_
  ylim(0,70)+
  ggtitle('Median IBU by State')+
  xlab('State')+
  ylab('Median IBU')+
  scale_fill_gradient2(name='Median IBU',low = "white", mid = "red", high = "red4",
```

```
                       midpoint = 40, limits = c(10,70),
                       breaks=c(10,22,34,46,58,70), na.value = "grey50")+
theme(legend.position = "none",
      title = element_text(face="bold", color = "midnightblue", size = 12),
      axis.text.y = element_blank(),
      axis.title.y = element_blank(),
      axis.ticks.y = element_blank(),
      axis.title.x = element_text(face="bold", color = "midnightblue", size = 9),
      axis.text.x = element_text(size = 7))
```



The bar charts of median ABV per state and the median IBU per state are shown above. The two states with the highest median ABV were Kentucky and Washington DC with 0.625. The state with the lowest median ABV was Utah (0.04). The state with the highest median IBU was Maine (score of 61) whereas the state with the lowest median IBU was Wisconsin (19).

```
#Find State w/ Max ABV and IBU when all NAs are deleted
ABVData%>%filter(medianABV==max(medianABV))
```

```
## # A tibble: 2 x 3
##   State medianABV count
##   <chr>     <dbl> <int>
## 1 " DC"    0.0625     8
## 2 " KY"    0.0625    20
```

```
IBUData%>%filter(medianIBU==max(medianIBU))
```

```
## # A tibble: 1 x 3
##   State medianIBU count
##   <chr>     <dbl> <int>
## 1 " ME"        61     7
```

```
ABVIBUData%>%filter(medianIBU==max(medianIBU))
```

```
## # A tibble: 1 x 4
##   State medianABV medianIBU count
##   <chr>     <dbl>     <dbl> <int>
## 1 " ME"     0.067        61     7
```

```
HighABVIBU=ABVIBUData%>%filter(medianABV==max(medianABV))



#Find State w/ Max ABV and IBU based on updated cleanup (separate for ABV and IBU)
HighABV=ABVData%>%filter(medianABV==max(medianABV))



HighIBU=IBUData%>%filter(medianIBU==max(medianIBU))
```

The first table shows that when accounting for all present ABV values, KY and DC have the highest median ABV. In the second table, when accounting for all present IBU values, Maine has the highest median ABV. Both of these findings match the barplots above. The third table above shows that Maine has the highest ABV and IBU when all NAs are deleted.

Question 5: Find States with Max ABV and IBU

```
#Find State w/ Max ABV and IBU Version 2

maxABVData=ABVbreweries%>%filter(ABV==max(ABV))
maxIBUData=IBUbreweries%>%filter(IBU==max(IBU))
maxABVData=data.frame(maxABVData$Beer_Name, maxABVData$Style, maxABVData$Brewery_Name, maxABVData$City,
names(maxABVData) <- c("Beer_Name","Style","Brewery_Name","City","State","Value")
maxABVData = maxABVData%>% mutate(Type = "Maximum ABV")
maxIBUData=data.frame(maxIBUData$Beer_Name, maxIBUData$Style, maxIBUData$Brewery_Name, maxIBUData$City,
names(maxIBUData) <- c("Beer_Name","Style","Brewery_Name","City","State","Value")
maxIBUData = maxIBUData%>% mutate(Type = "Maximum IBU")
maxABVIBUData = union(maxABVData,maxIBUData)
maxABVIBUData = maxABVIBUData [, c("Type","Beer_Name","Style","Brewery_Name","City","State","Value")]

formattable(maxABVIBUData)
```

Type

Beer_Name

Style

Brewery_Name

City

State

Value

Maximum ABV

Lee Hill Series Vol. 5 - Belgian Style Quadrupel Ale

Quadrupel (Quad)

Upslope Brewing Company

Boulder

CO

0.128

Maximum IBU

Bitter Bitch Imperial IPA

American Double / Imperial IPA

Astoria Brewing Company

Astoria

OR

138.000

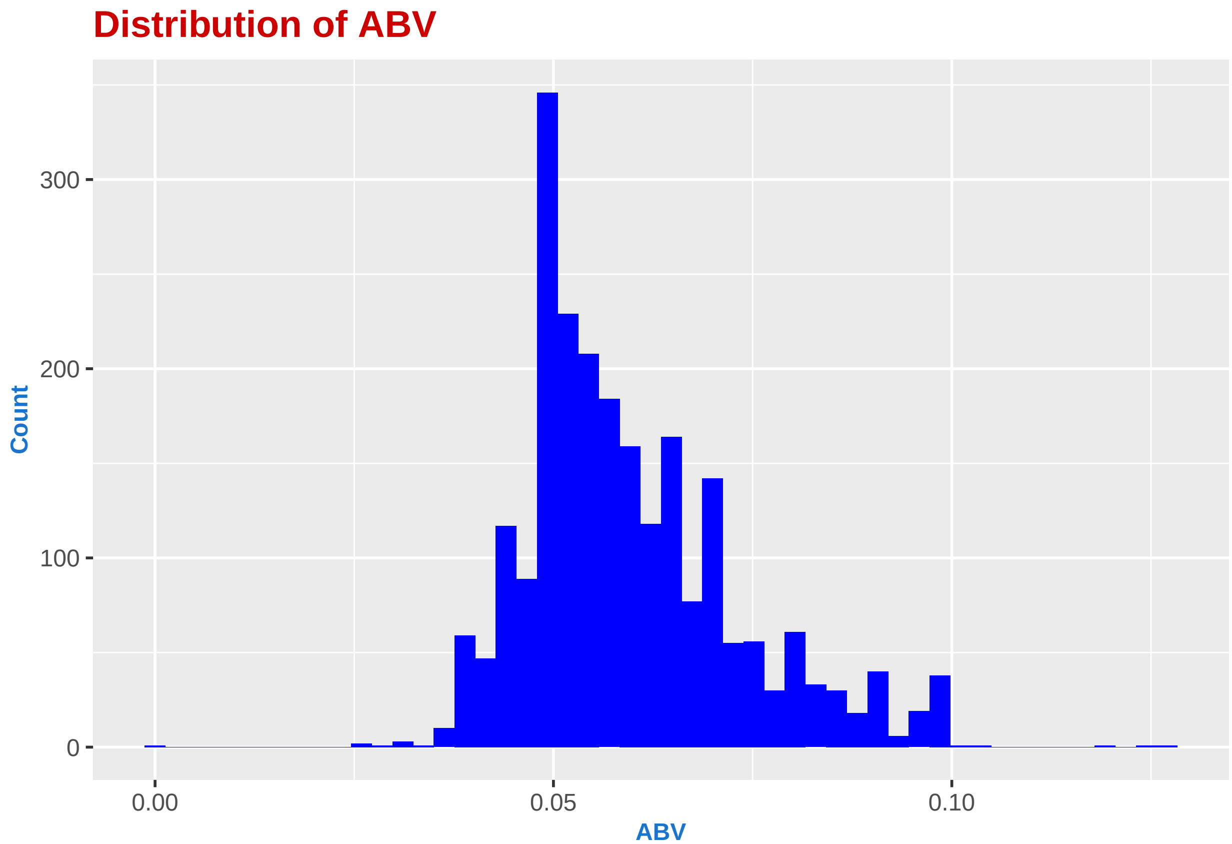
The table above shows that Colorado has the highest ABV beer and Oregon has the Highest IBU beer.

Question 6: Summary statistics and Distribution of ABV

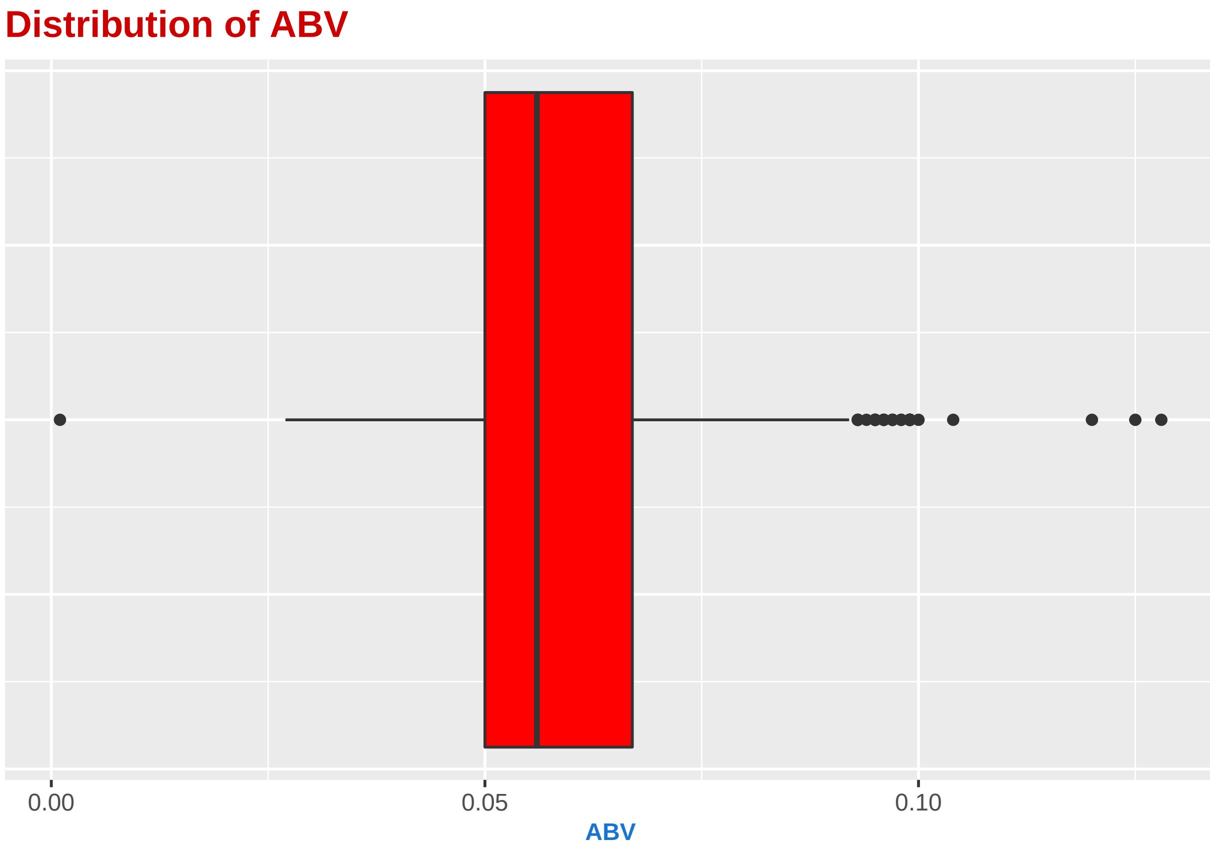
```
summary(beerbreweries$ABV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.02700 0.05000 0.05700 0.05991 0.06800 0.12500
```

```
#Histogram
ABVclean%>%ggplot(aes(x=ABV))+
  geom_histogram(bins=50,aes(),fill='blue')+
  xlab('ABV')+
  ylab('Count')+
  ggtitle('Distribution of ABV')+
  theme(legend.position = "bottom",
        legend.text = element_text(size=7),
        legend.title = element_text(face="bold", color = "black", size = 8),
        title = element_text(face="bold", color = "red3", size = 12),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
        axis.title.y = element_text(face="bold", color = "dodgerblue3", size = 9))
```




```r
#Boxplot
ABVclean%>%ggplot(aes(x=ABV))+
  geom_boxplot(aes(),fill='red')+
  ylab('ABV')+
  ggtitle('Distribution of ABV')+
  theme(legend.position = "bottom",
        legend.text = element_text(size=7),
        legend.title = element_text(face="bold", color = "black", size = 8),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks.y = element_blank(),
        title = element_text(face="bold", color = "red3", size = 12),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),)
```




```
SummaryABV = summary(ABVclean$ABV)

SummaryABV
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00100 0.05000 0.05600 0.05977 0.06700 0.12800
```

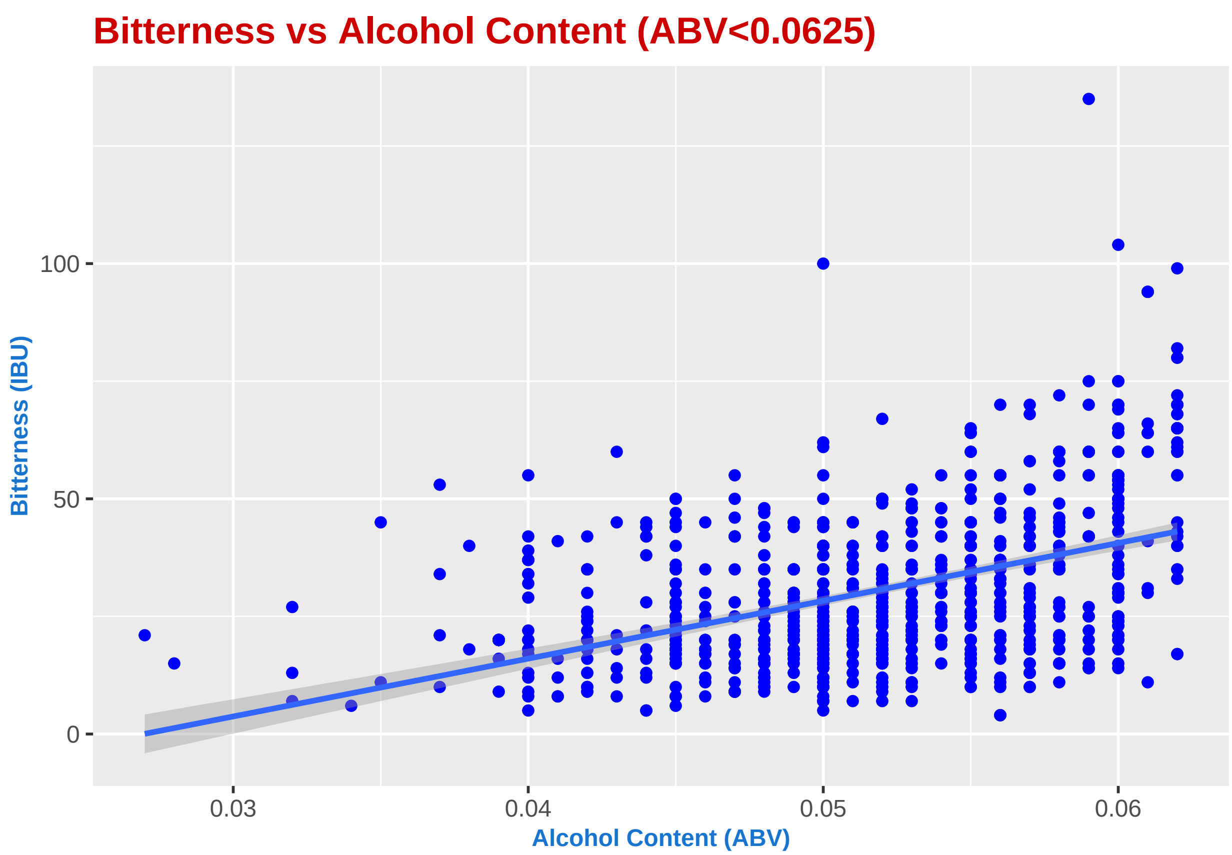
The first data table shows the summary statistics of ABV values if we excluded all NA values for ABV and IBU. The histogram, boxplot and the second data table account for all present ABV values. From histogram and boxplot, we can see that the distribution of ABV's is right skewed. This is confirmed by the second table in that the median is less than the mean. The average ABV is 0.05977 with the lowest ABV at 0.001 and the highest at 0.128. The interquartile range of the beers is between 0.05 and 0.067.

Question 7: Relationship between Bitterness of beer and its Alcohol Content

```
#How does the correlation between ABV and IBU change as either increase?
#Scatter Plot ABV and IBU

#Check correlation at lower ABV values
beerbreweries%>%filter(ABV<0.0625)%>%
  ggplot(aes(x=ABV,y=IBU))+
  geom_point(aes(),color='blue')+
  geom_smooth(method="lm")+
  ggtitle('Bitterness vs Alcohol Content (ABV<0.0625)')+
  xlab('Alcohol Content (ABV)')+
```

```r
  ylab('Bitterness (IBU)')+
  theme(title = element_text(face="bold", color = "red3", size = 12),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
        axis.title.y = element_text(face="bold", color = "dodgerblue3", size = 9))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



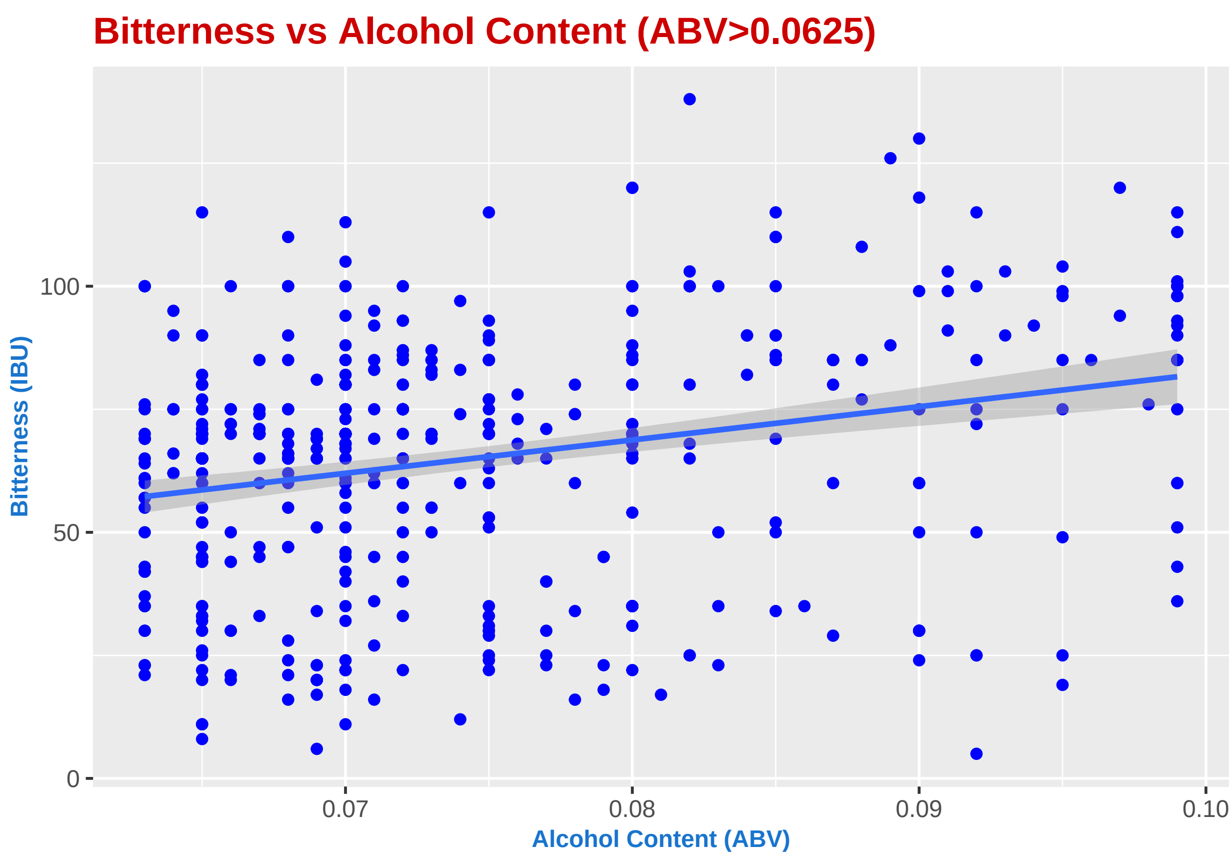

```r
#Find R^2 at lower values
LowABV=as.data.frame(beerbreweries%>%filter(ABV<0.0625))
LmodLow=lm(IBU~ABV,LowABV)
summary(LmodLow)#r^2=0.1971
```

```
##
## Call:
## lm(formula = IBU ~ ABV, data = LowABV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.684 -10.226  -2.684   8.720  95.629
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -33.138      4.291  -7.723    3e-14 ***
## ABV          1228.960     82.261  14.940   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.93 on 909 degrees of freedom
## Multiple R-squared:  0.1971, Adjusted R-squared:  0.1963
## F-statistic: 223.2 on 1 and 909 DF,  p-value: < 2.2e-16
```

```
#Check correlation at higher ABV values
beerbreweries%>%filter(ABV>0.0625&ABV<0.1)%>%
  ggplot(aes(x=ABV,y=IBU))+
  geom_point(aes(),color='blue')+
  geom_smooth(method="lm")+
  ggtitle('Bitterness vs Alcohol Content (ABV>0.0625)')+
  xlab('Alcohol Content (ABV)')+
  ylab('Bitterness (IBU)')+
  theme(title = element_text(face="bold", color = "red3", size = 12),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
        axis.title.y = element_text(face="bold", color = "dodgerblue3", size = 9))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Find R^2 at higher values
HighABV=as.data.frame(beerbreweries%>%filter(ABV>0.0625))
LmodHigh=lm(IBU~ABV,HighABV)
summary(LmodHigh)#r^2=0.07411
```

```
##
## Call:
## lm(formula = IBU ~ ABV, data = HighABV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.154 -14.866   6.087  14.507  68.255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.193      7.721   2.227   0.0264 *
## ABV          640.882    102.129   6.275 7.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.13 on 492 degrees of freedom
## Multiple R-squared:  0.07411,    Adjusted R-squared:  0.07222
## F-statistic: 39.38 on 1 and 492 DF,  p-value: 7.664e-10
```

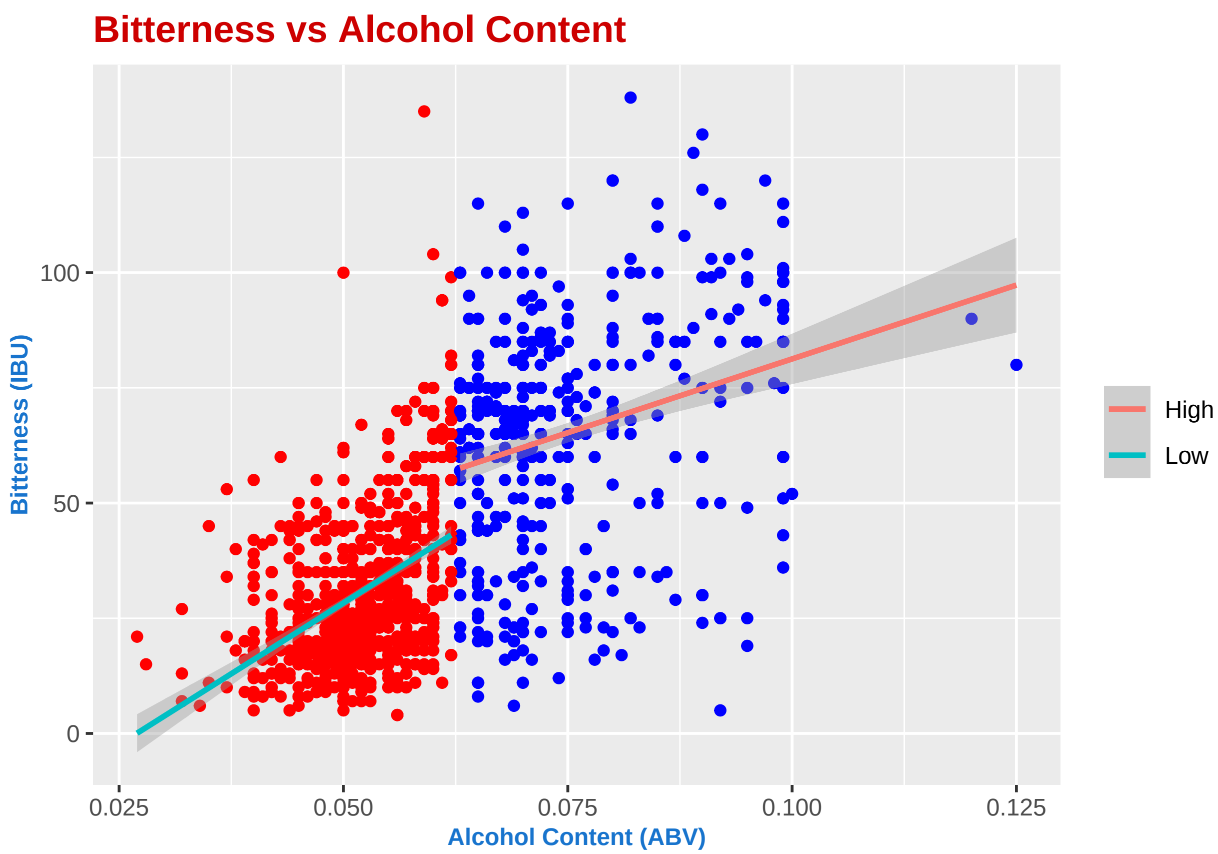
```
#Find overall r^2
LmodTotal=lm(IBU~ABV,beerbreweries)
summary(LmodTotal)#r^2=0.4497
```

```
##
## Call:
## lm(formula = IBU ~ ABV, data = beerbreweries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.849 -11.977  -0.721  13.997  93.458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.099      2.326  -14.66   <2e-16 ***
## ABV         1282.037     37.860   33.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.26 on 1403 degrees of freedom
## Multiple R-squared:  0.4497, Adjusted R-squared:  0.4493
## F-statistic:  1147 on 1 and 1403 DF,  p-value: < 2.2e-16
```

```
#Both high and low lines on the same graph
HighLowAbv=ifelse(beerbreweries$ABV>0.0625,'High','Low')
beerbreweries%>%mutate(HighLowAbv)%>%
  ggplot(aes(x=ABV,y=IBU))+
  geom_point(aes(),col=ifelse(beerbreweries$ABV>0.0625,'blue','red'))+
  geom_smooth(method='lm',aes(col=HighLowAbv))+
  ggtitle('Bitterness vs Alcohol Content')+
  xlab('Alcohol Content (ABV)')+
  ylab('Bitterness (IBU)')+
  theme(title = element_text(face="bold", color = "red3", size = 12),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
```

```
        axis.title.y = element_text(face="bold", color = "dodgerblue3", size = 9),
        legend.title=element_blank())
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the third graph, there appears to be a positive correlation between alcohol content and bitterness. It is of note, that the correlation appears weaker at the higher ABV and IBU. The percentage of estimated variance in IBU is explained by changes in ABV is 44.97%. Thus, we divided the graph where the correlation between IBU and ABV increased which was approximately at an ABV of 0.0625. The first scatterplot and data table show the distribution for ABV and IBU for ABV values less than 0.0625 where the least squares regression line has an R^2 = 0.1971. The second scatterplot and data table show the distribution for ABV and IBU for ABV values greater than 0.0625 where the least squares regression line has an R^2 = 0.07411. This supports our original claim that the correlation appears weaker at the higher ABV and IBU.

Question 8: Predicting ALE or IPA based on ABV and IBU Content

```
#Filter out beers that are either an Ale or an IPA
AleData=beers%>%filter((grepl("(IPA)",beers$Style))|(grepl("(Ale)",beers$Name)))


#Classify the drinks as either Ale or IPA
AleData$IPAorALE=ifelse(grepl("IPA",AleData$Style),"IPA","Ale")
AleClean=AleData%>%filter(!(is.na(IBU)|is.na(ABV)))

#Set up Matrix to store Accuracy, Specificity,and Sensitivity values for the upcoming confusion matrix
iterations=500
```

```r
numks=30
masterAcc=matrix(nrow=iterations,ncol=numks)
masterSen=matrix(nrow=iterations,ncol=numks)
masterSpec=matrix(nrow=iterations,ncol=numks)

for (j in 1:iterations){
  #70-30 Training-Test Split
  set.seed(sample(1:100000,1))
  trainInd=sample(1:dim(AleClean)[1],round(0.7*dim(AleClean)[1]))
  trainAle=AleClean[trainInd,]
  testAle=AleClean[-trainInd,]

  for(i in 1:numks){
    #k-NN to predict whether the drink is an IPA or an Ale
    AlePredictions=knn(trainAle[,c('ABV','IBU')],testAle[,c('ABV','IBU')],trainAle$IPAorALE,k=i,prob=TRU
    AleTable=table(AlePredictions,testAle$IPAorALE)
    AleCM=confusionMatrix(AleTable)
    masterAcc[j,i]=AleCM$overall[1]
    masterSen[j,i]=AleCM$byClass[1]
    masterSpec[j,i]=AleCM$byClass[2]
  }
}

#Collect the mean stats for each k-val
meanAcc=colMeans(masterAcc)
meanSen=colMeans(masterSen)
meanSpec=colMeans(masterSpec)

#Create dataframe with all stats
AleStats=data.frame(k=1:30,Mean_Accuracy=meanAcc,Mean_Sensitivity=meanSen,Mean_Specificity=meanSpec,Sum_

#Tune k-val based on all three stats
HighAleStats=AleStats%>%filter(Sum_Stat==max(Sum_Stat))
formattable(HighAleStats)
```

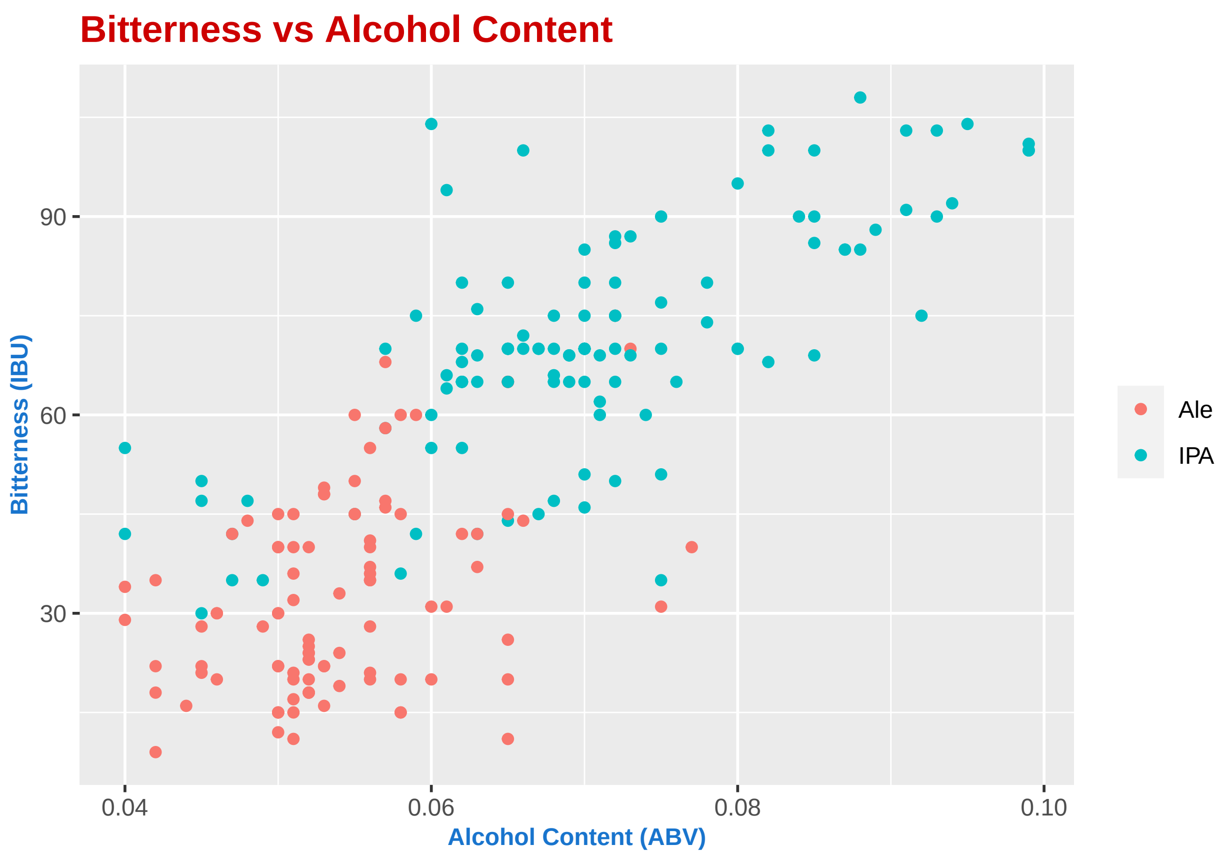
k	Mean_Accuracy	Mean_Sensitivity	Mean_Specificity	Sum_Stat
4	0.8612837	0.8634406	0.860476	2.5852

```r
#k=4 seemed to give the best balance between accuracy, sensitivity, and specificity
#I would prefer to use an odd number, but k=3,4, or 5 should provide good results regardless
```

```r
#70-30 Training-Test Split
set.seed(sample(1:100000,1))
trainInd=sample(1:dim(AleClean)[1],round(0.7*dim(AleClean)[1]))
trainAle=AleClean[trainInd,]
testAle=AleClean[-trainInd,]

#3-NN to predict whether the drink is an IPA or an Ale
AlePredictions=knn(trainAle[,c('ABV','IBU')],testAle[,c('ABV','IBU')],trainAle$IPAorALE,k=3,prob=TRUE)
AleTable=table(AlePredictions,testAle$IPAorALE)
confusionMatrix(AleTable)
```

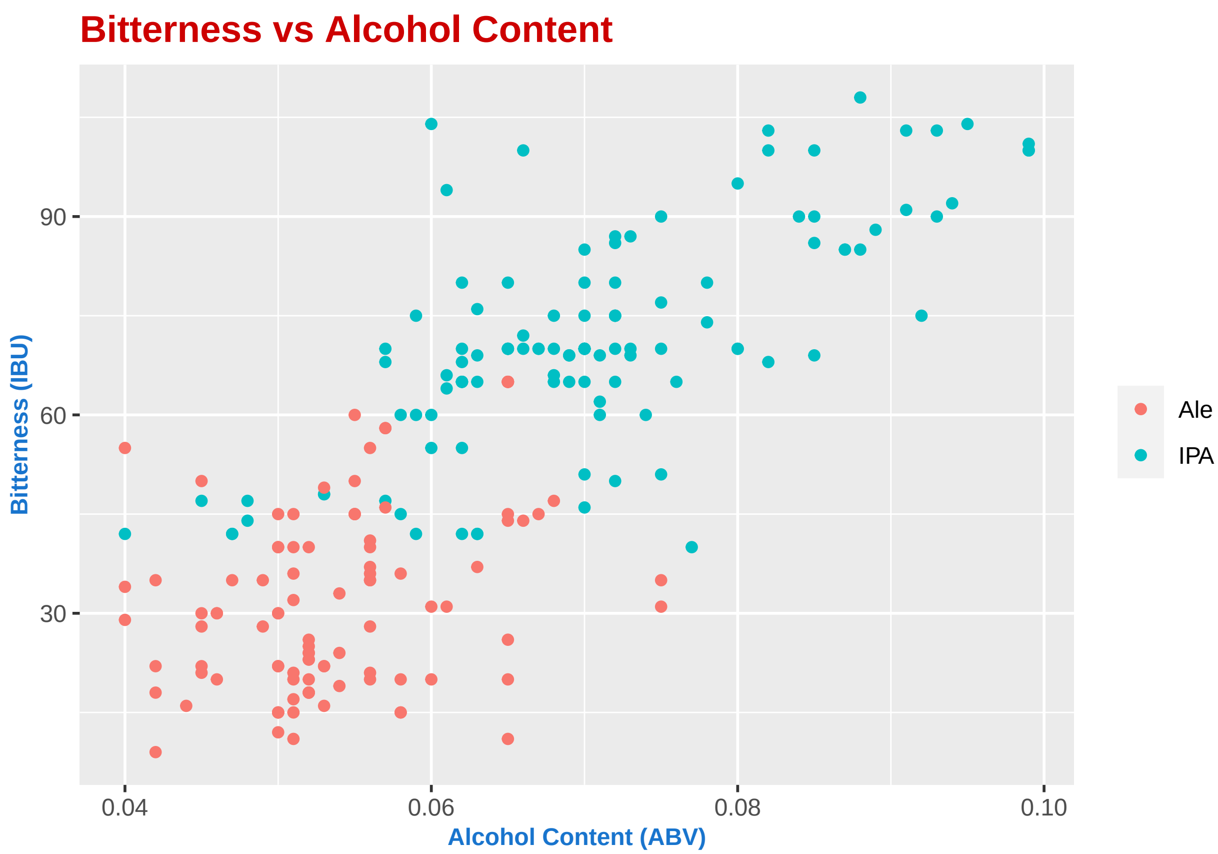
```
## Confusion Matrix and Statistics
##
##
## AlePredictions Ale IPA
##            Ale  85  15
##            IPA  14 101
##
##                Accuracy : 0.8651
##                  95% CI : (0.8121, 0.9078)
##     No Information Rate : 0.5395
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7287
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8586
##             Specificity : 0.8707
##          Pos Pred Value : 0.8500
##          Neg Pred Value : 0.8783
##              Prevalence : 0.4605
##          Detection Rate : 0.3953
##    Detection Prevalence : 0.4651
##       Balanced Accuracy : 0.8646
##
##        'Positive' Class : Ale
##
```

```r
#Scatterplots of prediction and actual classifications
#Scatterplot of actual classifications
testAle%>%ggplot(aes(ABV,IBU,color=IPAorALE))+
  geom_point()+
  ggtitle('Bitterness vs Alcohol Content')+
  xlab('Alcohol Content (ABV)')+
  ylab('Bitterness (IBU)')+
  theme(title = element_text(face="bold", color = "red3", size = 12),
        legend.title = element_blank(),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
        axis.title.y = element_text(face="bold", color = "dodgerblue3", size = 9))
```



We tested a variety of k from 1 through 30 for our knn model over 500 iterations to find the optimal average accuracy, specificity and sensitivity. The optimal k was found to be 4 but we decided to use k=3 since an odd k was more reasonable. We split the merged datasets into a 70-30 training and test set and found that the accuracy of this specific model was 0.8837, the sensitivity was 0.8835, and the specificity was 0.8839.

```
#Scatterplot of predicted classifications
testAle%>%mutate(AlePredictions)%>%
  ggplot(aes(ABV,IBU,color=AlePredictions))+
  geom_point()+
  ggtitle('Bitterness vs Alcohol Content')+
  xlab('Alcohol Content (ABV)')+
  ylab('Bitterness (IBU)')+
  theme(title = element_text(face="bold", color = "red3", size = 12),
        legend.title=element_blank(),
        axis.title.x = element_text(face="bold", color = "dodgerblue3", size = 9),
        axis.title.y = element_text(face="bold", color = "dodgerblue3", size = 9))
```



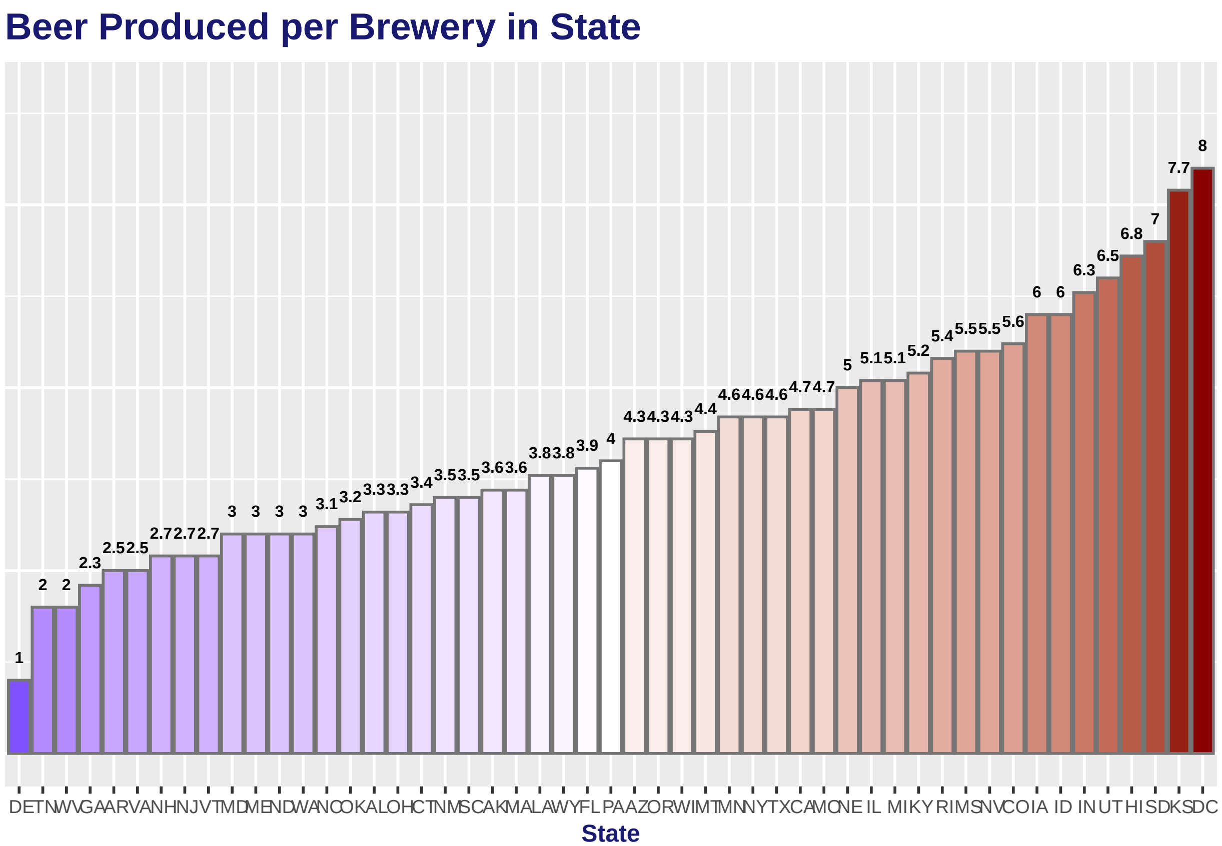
Above shows the scatterplot of the predicted classifications of either Ale or IPA. It appears IBU and ABV levels is very accurate at explaining whether or not a beverage is an IPA or an ale.

Question 9: Number of Beers per Brewery per State

We also found the number of beers produced per brewery for each state. A bar graph is shown below. Washington DC was found to have the most beers per brewery of all states.

```
beerbreweriesall = merge(beers,breweries, by.x="Brewery_id",by.y="Brew_ID")
colnames(beerbreweriesall)[1]='Brewery_ID'
colnames(beerbreweriesall)[2]='Beer_Name'
colnames(beerbreweriesall)[8]='Brewery_Name'

beerbreweriesallsummarybystate = beerbreweriesall%>%group_by(State)%>%
  summarize(Brewery_Count=n_distinct(Brewery_ID),Beer_Count=n_distinct(Beer_ID))
beerbreweriesallsummarybystate$Beer_Per_Brewery <- round(beerbreweriesallsummarybystate$Beer_Count/beerb

beerbreweriesallsummarybystate%>%
  ggplot(aes(x=reorder(State,Beer_Per_Brewery),y=Beer_Per_Brewery,fill=Beer_Per_Brewery))+
  geom_bar(stat='identity', color = "grey46")+
  geom_text(aes(label  = Beer_Per_Brewery), vjust = -1.5, size = 2.2, color = "black",fontface =  "bold
  ylim(0,9)+
  ggtitle('Beer Produced per Brewery in State')+
  xlab('State')+
  ylab('Avg. Beer per Brewery')+
  scale_fill_gradient2(low = "blue", mid = "white", high = "red4",
                       midpoint = 4, limits = c(1,8),
```

```
                        breaks=c(1,2,3,4,5,6,7,8), na.value = "grey50")+
  theme(legend.position = "none",
        title = element_text(face="bold", color = "midnightblue", size = 12),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.x = element_text(size = 7),
        axis.title.x = element_text(face="bold", color = "midnightblue", size = 9))
```




```
beerbreweriesallsummarybybrewery = beerbreweriesall%>%group_by(State,Brewery_Name)%>%
  summarize(Brewery_Count=n_distinct(Brewery_ID),Beer_Count=n_distinct(Beer_ID))
```

```
## `summarise()` has grouped output by 'State'. You can override using the `.groups` argument.
```

```
beerbreweriesallsummarybybrewery$Beer_Per_Brewery <- round(beerbreweriesallsummarybybrewery$Beer_Count/

Top_Brewery=beerbreweriesallsummarybybrewery%>%filter(State==" DC")
Top_Brewery_print=data.frame(Top_Brewery$State,Top_Brewery$Brewery_Name,Top_Brewery$Beer_Count)
names(Top_Brewery_print)=(c('State','Brewery Name','Beer Count'))
formattable(Top_Brewery_print)
```

State

Brewery Name

Beer Count

DC

DC Brau Brewing Company

8

From investigating the Beers and Breweries datasets, we found at least one brewery in each state with CO, MI, and CA having at least 30. Kentucky and Washington DC had the highest median ABV and Maine had the highest median IBU. Colorado had the beer with the highest ABV (Quadrupel (Quad)) whereas Oregon had the beer with the highest IBU (IPA). The state with the highest number of beers per brewery was Washington DC.

We found that ABV and IBU seemed to be positively correlated although correlation seemed to weaken for higher values of ABV and IBU. Both variables were approximately 88% accurate in predicting whether a beer was an Ale or an IPA. We appreciate the opportunity to work on this analysis. If you have an questions, feel free to contact us at davidg@mail.smu.edu, varung@mail.smu.edu, or roslyns@mail.smu.edu.