Brewer's Associate of American Independent Study

James Vasquez, Daniel Serna, Kumar Ramasundaram, Lance Dacy

June 17, 2018

Introduction

According to a recent article in Forbes, MillerCoors **(which owns Blue Moon, Pilsener Urquell and numerous brands besides Miller and Coors products)**, ranks No. 2 in volume of beer produced for sale. Constellation Brands is No. 3 with its various brands, including Corona and Modelo.

The traditional giants have bought some craft breweries in recent years, blurring the lines between craft and non-craft brewing companies. Anheuser-Busch bought 10 Barrel Brewing of Bend, Oregon, and Constellation Brands purchased Sand Diego-based Ballast Point for $1 billion.

This signifies that Craft beers have grown in popularity since the early 2000's and have skyrocketed in 2010 (see Figure 1 below).

In the US alone there are well over 150 styles of beer with as many breweries in each state. Knowing the palate of potential customers can gain a brewery an edge in profit margins by staying close to what the customer wants. Even more important is that the various regions in the USA might actually have variying taste / alcohol content preferences from other regions.

Our client, Brewer's Association of America, wanted to conduct a study for would-be investors to help understand the best chances of product success in today's hyper-competative market.

```r
#Load necessary libraries for the project.
library(ggplot2)
library(DataExplorer)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

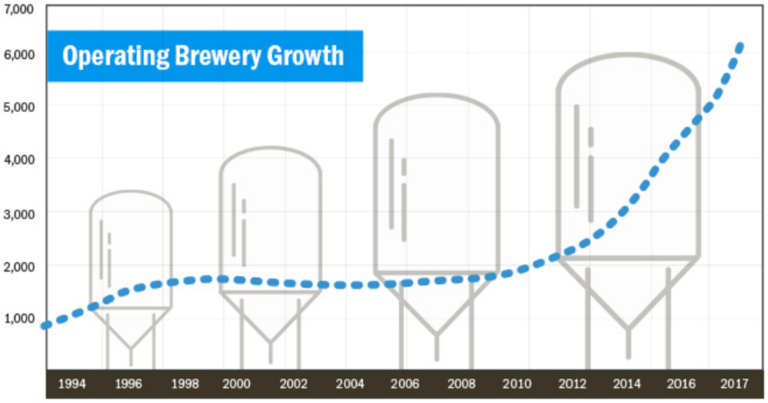


Figure 1:

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Question 1

Where and how many breweries are in the US?

The results indicate which states may be over populated with breweries and states that have not seen a wild growth in breweries.

```
###############################################
#####                          ###############
#####         Load Data        ###############
#####        Basic details     ###############
#####                          ###############
###############################################

#Read CSV file into R
Beers <- read.csv("Beers.csv", header=TRUE, sep=",", strip.white = TRUE)
Breweries <- read.csv("Breweries.csv", header=TRUE, sep=",", strip.white = TRUE)

###############################################
#####                          ###############
#####                          ###############
#####    Make State DB         ###############
#####                          ###############
###############################################

#Create State DB data frame
StateDB <- data.frame(state.name, state.abb, state.region)
#Rename columns for readability
colnames(StateDB)[colnames(StateDB)=='state.name'] <- 'StateName'
colnames(StateDB)[colnames(StateDB)=='state.abb'] <- 'State'
colnames(StateDB)[colnames(StateDB)=='state.region'] <- 'StateRegion'

#Add district of Columbia to StateDB Data Frame
DistrictColumbia <- data.frame("District of Columbia","DC", "South")
names(DistrictColumbia) <- c("StateName","State", "StateRegion")
StateDB <- rbind(StateDB, DistrictColumbia)

#head(StateDB,2) #Assuming this was for debugging purposes.

###############################################
#####                          ###############
#####         Question 1       ###############
#####    Breweries per state   ###############
#####                          ###############
###############################################
#count of breweries by state
BreweryCounts <- data.frame(table(Breweries$State))

#rename column names
```

```r
colnames(BreweryCounts)[colnames(BreweryCounts)=='Var1'] <- 'State'
colnames(BreweryCounts)[colnames(BreweryCounts)=='Freq'] <- 'NumberOfBreweriesByState'

#Merge the StateDB and sort by count of breweries by state
BreweryCounts <- merge(BreweryCounts, StateDB, by.x=("State"), by.y=("State"))
BreweryCounts <- BreweryCounts[order(BreweryCounts$NumberOfBreweriesByState, decreasing=TRUE),c(3,2)]
BreweryCounts
```

```
##               StateName NumberOfBreweriesByState
## 6              Colorado                       47
## 5            California                       39
## 23             Michigan                       32
## 38               Oregon                       29
## 44                Texas                       28
## 39         Pennsylvania                       25
## 20        Massachusetts                       23
## 48           Washington                       23
## 16              Indiana                       22
## 49            Wisconsin                       20
## 28       North Carolina                       19
## 15             Illinois                       18
## 35             New York                       16
## 46             Virginia                       16
## 10              Florida                       15
## 36                 Ohio                       15
## 24            Minnesota                       12
## 4               Arizona                       11
## 47              Vermont                       10
## 22                Maine                        9
## 25             Missouri                        9
## 27              Montana                        9
## 7           Connecticut                        8
## 1                Alaska                        7
## 11              Georgia                        7
## 21             Maryland                        7
## 37             Oklahoma                        6
## 13                 Iowa                        5
## 14                Idaho                        5
## 19            Louisiana                        5
## 30             Nebraska                        5
## 40         Rhode Island                        5
## 12               Hawaii                        4
## 18             Kentucky                        4
## 33           New Mexico                        4
## 41       South Carolina                        4
## 45                 Utah                        4
## 51              Wyoming                        4
## 2               Alabama                        3
## 17               Kansas                        3
## 31        New Hampshire                        3
## 32           New Jersey                        3
## 43            Tennessee                        3
## 3              Arkansas                        2
## 9              Delaware                        2
## 26          Mississippi                        2
## 34               Nevada                        2
## 8  District of Columbia                        1
## 29         North Dakota                        1
## 42         South Dakota                        1
## 50        West Virginia                        1
```

Question 2

Merging Data

The team merged both the beer and breweries data sets in order to get a better wholelistic view of the data and determine how they relate to each other. Below will show the first and last six rows of the data sets.

```
##############################################
#####                          ###############
#####         Question 2       ###############
#####       Merge Data Sets    ###############
#####                          ###############
##############################################

#join data on Brewery_id and Brew_ID
BeersAndBreweries <- merge(Beers, Breweries, by.x=("Brewery_id"), by.y=("Brew_ID"))

#list column names on the joined data frame
#colnames(BeersAndBreweries) #assuming this was for debugging purposes

#rename the name.x(Beer) and name.y(Brewery) after the merger
colnames(BeersAndBreweries)[colnames(BeersAndBreweries)=='Name.x'] <- 'BeerName'
colnames(BeersAndBreweries)[colnames(BeersAndBreweries)=='Name.y'] <- 'BreweryName'

# I don't think we don't need to create this again.
# #Create State DB data frame
# StateDB <- data.frame(state.name, state.abb, state.region)
# colnames(StateDB)[colnames(StateDB)=='state.name'] <- 'StateName'
# colnames(StateDB)[colnames(StateDB)=='state.abb'] <- 'State'
# colnames(StateDB)[colnames(StateDB)=='state.region'] <- 'StateRegion'
#
# #Add district of Columbia to StateDB Data Frame
# DistrictColumbia <- data.frame("District of Columbia","DC", "South")
# names(DistrictColumbia) <- c("StateName","State", "StateRegion")
# StateDB <- rbind(StateDB, DistrictColumbia)


#Merge data with State DB
BeersAndBreweries <- merge(BeersAndBreweries, StateDB, by="State", all = TRUE)

# dont think we need to show this.
# #find dimensions of data frames
# dim(BeersAndBreweries)
# dim(BeersAndBreweries)
```

```r
#Show first and last 6 entries of merged files
head(BeersAndBreweries)
```

```
##   State Brewery_id                   BeerName Beer_ID   ABV IBU
## 1    AK        103            King Street IPA    1667 0.060  70
## 2    AK        103                  Amber Ale    2436 0.051  NA
## 3    AK        494             Polar Pale Ale     920 0.052  17
## 4    AK        459          Sunken Island IPA     349 0.068  NA
## 5    AK        103        King Street Pilsner    1706 0.055  NA
## 6    AK        459 Skilak Scottish Ale (2011)     348 0.058  NA
##                      Style Ounces                 BreweryName      City
## 1             American IPA     12  King Street Brewing Company Anchorage
## 2 American Amber / Red Ale     12  King Street Brewing Company Anchorage
## 3  American Pale Ale (APA)     12 Broken Tooth Brewing Company Anchorage
## 4             American IPA     12  Kenai River Brewing Company  Soldotna
## 5           Czech Pilsener     12  King Street Brewing Company Anchorage
## 6             Scottish Ale     12  Kenai River Brewing Company  Soldotna
##   StateName StateRegion
## 1    Alaska        West
## 2    Alaska        West
## 3    Alaska        West
## 4    Alaska        West
## 5    Alaska        West
## 6    Alaska        West
```

```r
tail(BeersAndBreweries)
```

```
##      State Brewery_id                         BeerName Beer_ID   ABV IBU
## 2405    WY        458    Saddle Bronc Brown Ale (2013)    1198 0.048  16
## 2406    WY        192                     Pako’s EyePA     393 0.068  60
## 2407    WY        192               Snow King Pale Ale    1606 0.060  55
## 2408    WY        458             Wagon Box Wheat Beer    1197 0.059  15
## 2409    WY        458            Indian Paintbrush IPA    1199 0.070  75
## 2410    WY        458 Bomber Mountain Amber Ale (2013)    1200 0.046  20
##                         Style Ounces                     BreweryName
## 2405        English Brown Ale     12 The Black Tooth Brewing Company
## 2406             American IPA     12     Snake River Brewing Company
## 2407  American Pale Ale (APA)     12     Snake River Brewing Company
## 2408  American Pale Wheat Ale     12 The Black Tooth Brewing Company
## 2409             American IPA     12 The Black Tooth Brewing Company
## 2410 American Amber / Red Ale     12 The Black Tooth Brewing Company
##          City StateName StateRegion
## 2405 Sheridan   Wyoming        West
## 2406  Jackson   Wyoming        West
## 2407  Jackson   Wyoming        West
## 2408 Sheridan   Wyoming        West
## 2409 Sheridan   Wyoming        West
## 2410 Sheridan   Wyoming        West
```

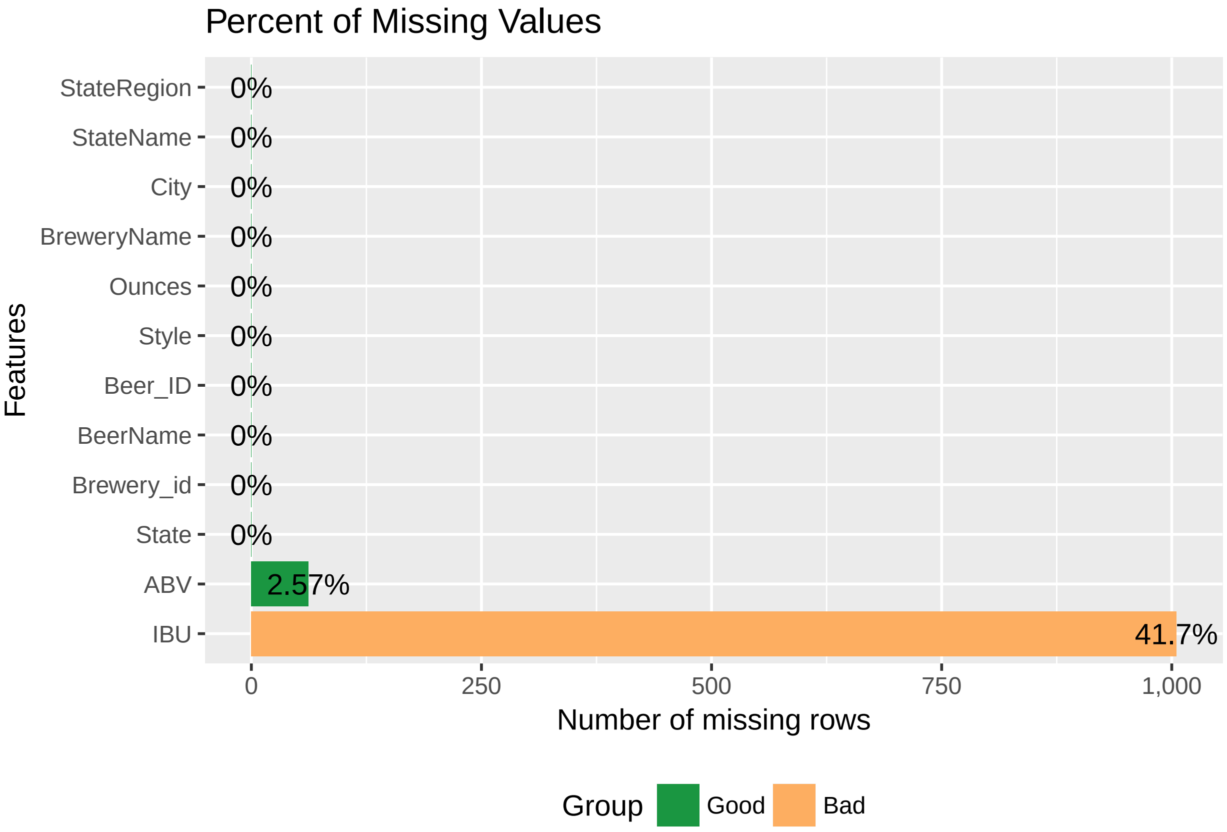
Question 3

Report Missing Values

To do a complete analysis the team needed to asses if the merged data set has any missing values. The team created a report for both a graphical and tabular representation of the results. The team utilized code from http://www.gettinggeneticsdone.com/2011/02/summarize-missing-data-for-all.html to count NA's in columns.

The data shows the variables with missing data are the IBU and ABV values.

```
#Graphical representation of missing vaules using 'DataExporer' library
plot_missing(BeersAndBreweries, title = "Percent of Missing Values")
```



```
#Function to count all NA's in columns (sourced from the internet)
#http://www.gettinggeneticsdone.com/2011/02/summarize-missing-data-for-all.html
propmiss <- function(dataframe) {
  m <- sapply(dataframe, function(x) {
    data.frame(
      na_count=sum(is.na(x)),
      Obs=length(x),
      perc_missing=sum(is.na(x))/length(x)*100
    )
  })
  d <- data.frame(t(m))
  d <- sapply(d, unlist)
  d <- as.data.frame(d)
  d$variable <- row.names(d)
  row.names(d) <- NULL
  d <- cbind(d[ncol(d)],d[-ncol(d)])
  return(d[order(d$na_count, decreasing=TRUE), ])
}

#show results of NA's counted
BeerColumnInventory_nacount <- propmiss(BeersAndBreweries)
BeerColumnInventory_nacount
```

```
##       variable na_count  Obs perc_missing
## 6          IBU     1005 2410    41.701245
## 5          ABV       62 2410     2.572614
## 1        State        0 2410     0.000000
## 2   Brewery_id        0 2410     0.000000
## 3     BeerName        0 2410     0.000000
## 4      Beer_ID        0 2410     0.000000
## 7        Style        0 2410     0.000000
## 8       Ounces        0 2410     0.000000
## 9  BreweryName        0 2410     0.000000
## 10        City        0 2410     0.000000
## 11   StateName        0 2410     0.000000
## 12 StateRegion        0 2410     0.000000
```

Question 4

Plotting Data

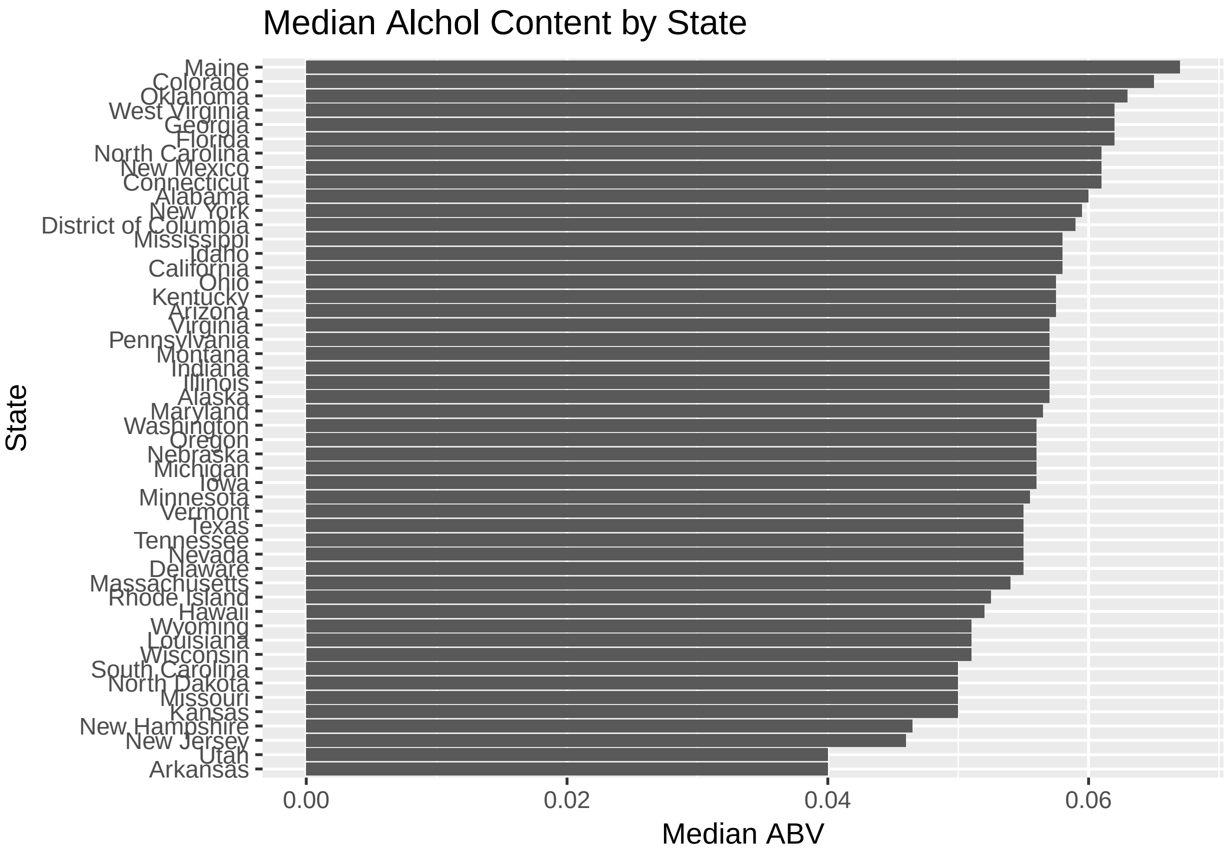
The team in order to look for trends plotted the ABV and IBU against the states to determine which states had the highest median value of each of tthe states by value.

- Process to analyze
  - Calculate the median values of ABV & IBU by state
  - Plot the data against states and sort by highest value

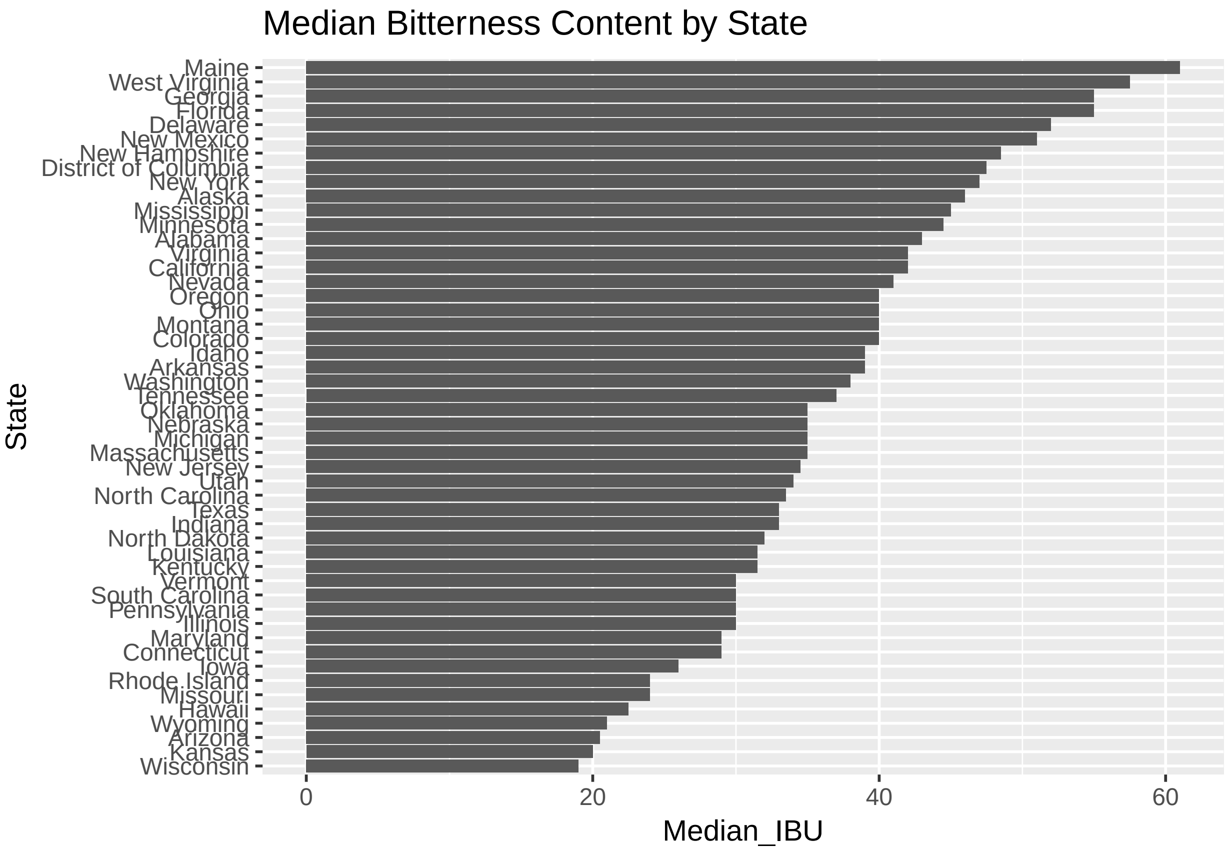
As requested by the client all NA's have been removed

```
#Make data frame with only State, ABV, IBU
DF_ABV_IBU <- BeersAndBreweries[,c("StateName","ABV","IBU")]
#head(DF_ABV_IBU)  #I don't think we don't need to show this.

#remove any rows with a NA value using 'complete.cases'
DF_ABV_IBU_noNA <- DF_ABV_IBU[complete.cases(DF_ABV_IBU),]
#head(DF_ABV_IBU_noNA) #I don't think we don't need to show this.

#Calculate MEDIAN values for ABV&IBU by State
MEDIAN_ABV_IBU_by_State <- aggregate(DF_ABV_IBU_noNA[, 2:3],list(DF_ABV_IBU_noNA$StateName), median)
#head(MEDIAN_ABV_IBU_by_State) #I don't think we don't need to show this.

#Rename column names
colnames(MEDIAN_ABV_IBU_by_State)[colnames(MEDIAN_ABV_IBU_by_State)=='Group.1'] <- 'State'
colnames(MEDIAN_ABV_IBU_by_State)[colnames(MEDIAN_ABV_IBU_by_State)=='ABV'] <- 'Median_ABV'
colnames(MEDIAN_ABV_IBU_by_State)[colnames(MEDIAN_ABV_IBU_by_State)=='IBU'] <- 'Median_IBU'
```

```
#Check data
#head(MEDIAN_ABV_IBU_by_State,10) #I don't think we don't need to show this.

#######  Plot MEDIAN ABV By State   #########
BarPlot_ABV_byState <- ggplot(data=MEDIAN_ABV_IBU_by_State,
                              aes(x=reorder(State, Median_ABV),
                                  y=Median_ABV)) +
                                  geom_bar(stat="identity")+
                                  coord_flip() +
                                  labs(x="State",
                                       y="Median ABV",
                                       title = "Median Alchol Content by State")
BarPlot_ABV_byState
```



```
#######  Plot MEDIAN IBU By State   #########
BarPlot_IBU_byState <- ggplot(data=MEDIAN_ABV_IBU_by_State,
                              aes(x=reorder(State, Median_IBU),
                                  y=Median_IBU)) +
                              geom_bar(stat="identity")+
                              coord_flip()+
                              labs(x="State",
                                   y="Median_IBU",
                                   title = "Median Bitterness Content by State")
BarPlot_IBU_byState
```



Question 5

States with highest ABV and IBU

For a quick reference the team identified the states with the highest ABY and IBU recorded within the data set.

```
#Find MAX ABV with State
MAX_ABV_byState <- head(BeersAndBreweries[order(BeersAndBreweries$ABV, na.last = TRUE, decreasing=TRUE)
MAX_ABV_byState
```

```
##     StateName   ABV
## 533  Colorado 0.128
```

```
#Find MAX IBU with State, column has missing values
MAX_IBU_byState <- head(BeersAndBreweries[order(BeersAndBreweries$IBU, na.last = TRUE, decreasing=TRUE)
MAX_IBU_byState
```

```
##      StateName IBU
## 1824    Oregon 138
```

Question 6

Summary of ABV

As part of the analysis the team has provided the summary results for the ABV variable.

```
##############################################
#####                          ###############
#####     Question 6           ###############
##### Summary of ABV Variable  ###############
#####                          ###############
##############################################

#Summary Stats of the ABV variable
SUMMARY_ABV <- summary(BeersAndBreweries$ABV)

#Show ABV SUmmary
SUMMARY_ABV
```

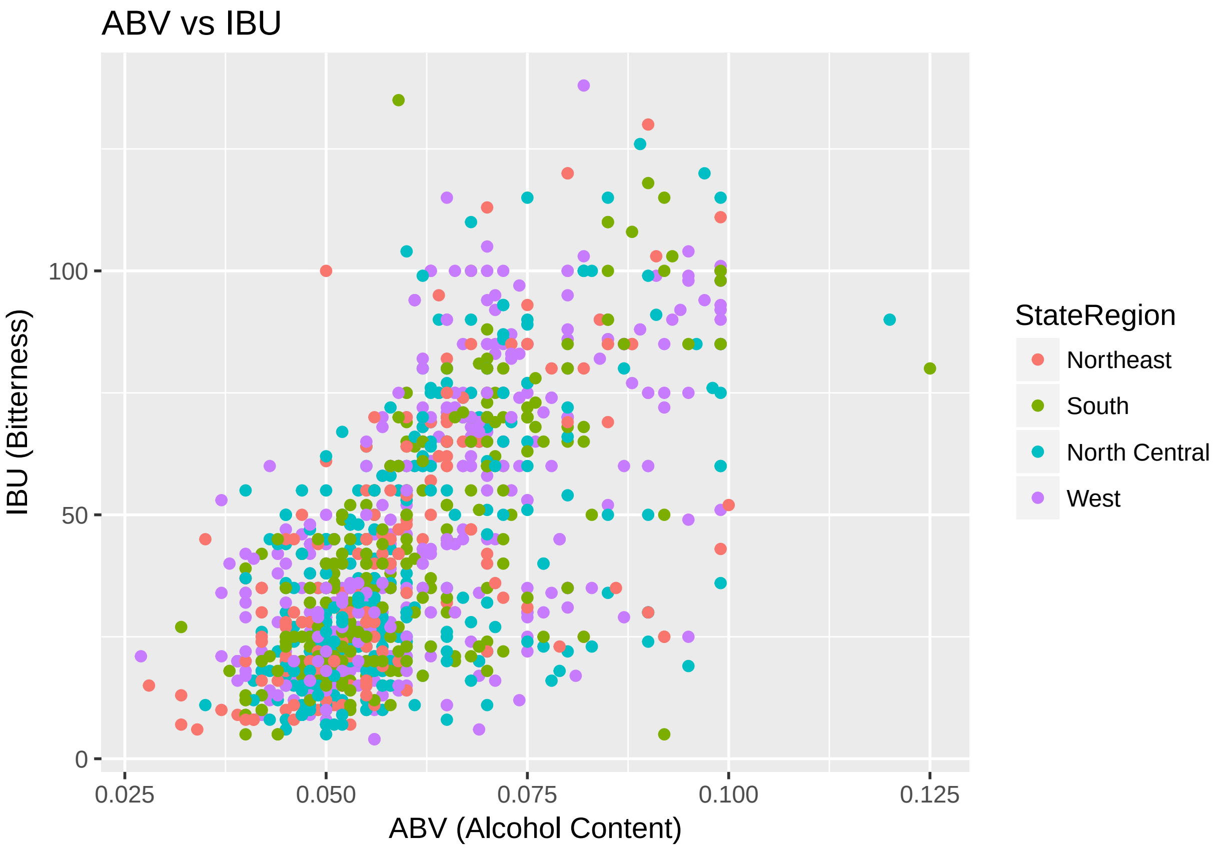
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.00100 0.05000 0.05600 0.05977 0.06700 0.12800      62 
```

Question 7

Relationship between ABV & IBU

To determine any realtionships between ABV and IBU a scatterplot was created. In addition the data was color coded by the region of the brewery.

```
#Merge data with State DB
DF_ABV_IBU_noNA <- merge(DF_ABV_IBU_noNA, StateDB, by="StateName", all = TRUE)

#Remove rows with NA's
DF_ABV_IBU_noNA <- DF_ABV_IBU_noNA[complete.cases(DF_ABV_IBU_noNA),]

# I don't think we need to show this.
#Check merge and top of the file
# propmiss(DF_ABV_IBU_noNA)
# head(DF_ABV_IBU_noNA)

#Scatter plot ABV vs IBU and color by StateRegion
ABUvsIBU <- qplot(ABV, IBU,
                  xlab = "ABV (Alcohol Content)",
                  ylab = "IBU (Bitterness)",
                  main= "ABV vs IBU",
                  colour=StateRegion,
                  data=DF_ABV_IBU_noNA)

#Show Scatter Plot
ABUvsIBU
```


ABV vs IBU
IBU (Bitterness)
100
50
0
0.025
0.050
0.075
0.100
0.125
ABV (Alcohol Content)
StateRegion
● Northeast
● South
● North Central
● West