Case Study 01

Thad Schwebke

2/18/2020

## Instructions

You can assume that your audience is the CEO and CFO of Budweiser (your client) and that they only have had one class in statistics and have indicated that you cannot take more than 7 minutes of their time. 20% of your grade will be based on the presentation.

They have hired you to address the 9 questions / items described in the next section. Beyond those general questions you may and should speculate / anticipate what may be of interest to them.

## Deliverables

1. A GitHub repository (Due Saturday March 7th 11:59pm CST

The GitHub repo should contain the following items and a link to the GitHub repo should be placed on a Word Doc (or PDF) and submitted to 2DS for Unit 8.

The final repo which will be checked after 11:59pm CST March 7th should contain:

1. An RMarkdown file containing the following:
2. The introduction needs to be written as if you are presenting the work to the CEO and CFO of Budweiser (your client) and that they only have had one class in statistics. If it sounds like a student presentation, that is not acceptable. You may assume that the CEO and CFO gave you the data and gave you the directive to report any interesting finding that you may uncover through your analysis.
3. Briefly explain the purpose of the code. The explanations should appear as a sentence or two before or after the code chunk. Even though you will not be hiding the code chunks (so that I can see the code), you need to assume that the client can’t see them.
4. Use R to address the 9 questions / items below:
5. How many breweries are present in each state?
6. Merge beer data with the breweries data. Print the first 6 observations and the last six observations to check the merged file. (RMD only, this does not need to be included in the presentation or the deck.)
7. Address the missing values in each column.
8. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.
9. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?
10. Comment on the summary statistics and distribution of the ABV variable.
11. Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot. Make your best judgment of a relationship and EXPLAIN your answer.
12. Budweiser would also like to investigate the difference with respect to IBU and ABV between IPAs (India Pale Ales) and other types of Ale (any beer with “Ale” in its name other than IPA). You decide to use KNN classification to investigate this relationship. Provide statistical evidence one way or the other. You can of course assume your audience is comfortable with percentages … KNN is very easy to understand conceptually.

In addition, while you have decided to use KNN to investigate this relationship (KNN is required) you may also feel free to supplement your response to this question with any other methods or techniques you have learned. Creativity and alternative solutions are always encouraged.

1. Knock their socks off! Find one other useful inference from the data that you feel Budweiser may be able to find value in. You must convince them why it is important and back up your conviction with appropriate statistical evidence.

Directives on RMD File:

1. Give clear, explicit answers to the questions. Just the code to answer the questions is not nearly enough, even if the code is correct and gives the correct answer. You must state the answer in a complete sentence outside the code chunk.
2. Conclusion to the project. Summarize your findings from this exercise. The file must be readable in GitHub. In other words, don’t forget to keep the md file!!
3. Knit HTML file.  
   In fact, you will also upload the knit html file to GitHub as well. This will allow for plots and tables to supplement your answers (part e) to the 7 questions below. You are already making an Rmd file, we should take advantage of it and knit a nice presentation of the project!
4. Codebook, Both CSV files and a ReadMe.md The Readme file describes the purpose of the project and codebook. The repo can be structured however you like, but it should make sense and be easily navigated.
5. PPT Slides Described more below and should have the link to your You Tube presentation … described further below as well.)

B. UNIT 8 Live Session: EDA

Each team will need to meet with the professor and present their EDA in Unit 8. It is up to the teams and the professor when and how to set up these meetings. They may be during the schedule live session time, but given time constraints, some teams will need to schedule times outside of their scheduled live session time. With that said, these are one on one meetings between each team and the professor. Your only scheduled time commitment in Unit 8 is to attend this 10ish minute meeting with your professor. The rest of the time is reserved to work on your project with your partner. Your goal is to present your EDA (Answers to Questions 1,3,4,5,6,7). At this point, teams should have presentation quality slides and presentation prepared. Responses to each of the questions listed above should be prepared and addressed in this meeting.

The grade for this portion is based on the slide deck and the presentation. If the team is prepared and delivers a well-practiced presentation it should be easy to earn close to a 100% here.

1. Unit 9 Live Session: Q & A.

During Live Session for Unit 9, the professor will be available for a live Q & A about the project / presentation. Attendance at this live session is not required (attendance is optional). NOTE ABOUT POWERPOINT … You may use the same powerpoint or develop them separately. I would imagine that even if you develop the powerpoint together that each student’s final powerpoint will be a little different just based on individual presentation style. Everyone has their own unique style and delivery.

C. Final YouTube Video

Each team member will need to record and upload to YouTube a 7-minute or less presentation of your findings. At this point you should know your presentation backwards and forwards. If you trip up too much in your recording, you should start over until you have a very polished presentation that does not exceed 7 minutes.

To record you can download Camtasia (free trial) which is a video software available at <https://www.techsmith.com/video-editor.html> or use your preferred screen capture software (like QuickTime if you have a Mac.) The presentation slides that include a link to your video should be in the Case Study Github repo as well as on the Google Doc provided by your professor. The goal is to communicate the findings of the project in a clear, concise and scientific manner. Also, uploading to YouTube is not difficult. Here is a YouTube video to help: <https://www.youtube.com/watch?v=VtF2AgFSLAw> Your professor will make the Google Doc link available to everyone in the class so that your peers can benefit from your work and so that you can benefit from theirs. Student’s presentation links will be available for a week at which time you may take your video off of YouTube if you wish.

## Collaboration and Peer Review

This will be a team project. We expect that all team members will do equal work and give their best to advance the knowledge of both themselves and their teammate. All members will need to push, add, commit, and pull to GitHub. This is a collaborative project, be sure and communicate early and often; mutual respect is key.

You will be providing two peer reviews that will be submitted to 2DS in Unit 8 and Unit 9 under: Project 1: EDA and Peer Review (by Saturday Feb 29 11:59pm / Unit 8) and Project 1: Final Documentation, Presentation and Peer Review Assignment (by Saturday March 7 11:59pm / Unit 9) . See the Rubric for detailed information on the peer review.

## EDA:

Questions 1,3,4,5,6,7

## Load Libraries

#install.packages("vctrs")  
library(vctrs)

## Warning: package 'vctrs' was built under R version 3.6.2

library(qwraps2)

## Warning: package 'qwraps2' was built under R version 3.6.2

library(tidyverse) # super library that loads dplyr, ggplot2, lubridate, tidyr, stringr, etc.

## Warning: package 'tidyverse' was built under R version 3.6.2

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

library(DataExplorer)

## Warning: package 'DataExplorer' was built under R version 3.6.2

library(maps) # used for heat map

## Warning: package 'maps' was built under R version 3.6.2

library(openintro) # used to convert state abbreviation to full state name  
library(gdata)

## Warning: package 'gdata' was built under R version 3.6.2

library(mapdata)

## Warning: package 'mapdata' was built under R version 3.6.2

library(mime)

## Warning: package 'mime' was built under R version 3.6.2

library(treemap)

## Warning: package 'treemap' was built under R version 3.6.2

library(VIM)

## Warning: package 'VIM' was built under R version 3.6.2

## Warning: package 'data.table' was built under R version 3.6.2

library(inspectdf)

## Warning: package 'inspectdf' was built under R version 3.6.2

library(GGally)

## Warning: package 'GGally' was built under R version 3.6.2

library(fiftystater)  
library(ggthemes)

## Warning: package 'ggthemes' was built under R version 3.6.2

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.6.2

#theme\_set(theme\_classic())

## Load Data

setwd("C:/Users/thads/OneDrive/SMU Data Science/DS6306/DDS6306-Group-Project-1")  
  
beers\_df <- read.csv("./data/Beers.csv", header = T, sep = ",")  
breweries\_df <-  
 read.csv("./data/Breweries.csv", header = T, sep = ",")

## Inspect the data

*TODO add notes about what inspection found*

#View(beers\_df) # view the complete dataframe  
names(beers\_df) # view the variable names

## [1] "Name" "Beer\_ID" "ABV" "IBU" "Brewery\_id"  
## [6] "Style" "Ounces"

dim(beers\_df) # view the dimensions of the data

## [1] 2410 7

str(beers\_df) # view the structure of the data

## 'data.frame': 2410 obs. of 7 variables:  
## $ Name : Factor w/ 2305 levels "#001 Golden Amber Lager",..: 1638 577 1705 1842 1819 268 1160 758 1093 486 ...  
## $ Beer\_ID : int 1436 2265 2264 2263 2262 2261 2260 2259 2258 2131 ...  
## $ ABV : num 0.05 0.066 0.071 0.09 0.075 0.077 0.045 0.065 0.055 0.086 ...  
## $ IBU : int NA NA NA NA NA NA NA NA NA NA ...  
## $ Brewery\_id: int 409 178 178 178 178 178 178 178 178 178 ...  
## $ Style : Factor w/ 100 levels "","Abbey Single Ale",..: 19 18 16 12 16 80 18 22 18 12 ...  
## $ Ounces : num 12 12 12 12 12 12 12 12 12 12 ...

head(beers\_df, n = 5) # view the first 5 records

## Name Beer\_ID ABV IBU Brewery\_id  
## 1 Pub Beer 1436 0.050 NA 409  
## 2 Devil's Cup 2265 0.066 NA 178  
## 3 Rise of the Phoenix 2264 0.071 NA 178  
## 4 Sinister 2263 0.090 NA 178  
## 5 Sex and Candy 2262 0.075 NA 178  
## Style Ounces  
## 1 American Pale Lager 12  
## 2 American Pale Ale (APA) 12  
## 3 American IPA 12  
## 4 American Double / Imperial IPA 12  
## 5 American IPA 12

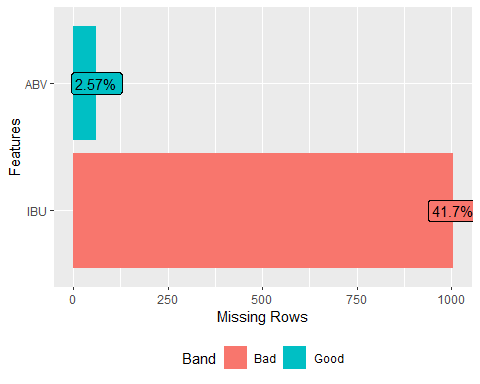
tail(beers\_df, n = 5) # view the last 5 records

## Name Beer\_ID ABV IBU Brewery\_id Style  
## 2406 Belgorado 928 0.067 45 425 Belgian IPA  
## 2407 Rail Yard Ale 807 0.052 NA 425 American Amber / Red Ale  
## 2408 B3K Black Lager 620 0.055 NA 425 Schwarzbier  
## 2409 Silverback Pale Ale 145 0.055 40 425 American Pale Ale (APA)  
## 2410 Rail Yard Ale (2009) 84 0.052 NA 425 American Amber / Red Ale  
## Ounces  
## 2406 12  
## 2407 12  
## 2408 12  
## 2409 12  
## 2410 12

summary(beers\_df) # view basic model data i.e. min, max, mdedium, mean, etc.

## Name Beer\_ID ABV   
## Nonstop Hef Hop : 12 Min. : 1.0 Min. :0.00100   
## Dale's Pale Ale : 6 1st Qu.: 808.2 1st Qu.:0.05000   
## Oktoberfest : 6 Median :1453.5 Median :0.05600   
## Longboard Island Lager: 4 Mean :1431.1 Mean :0.05977   
## 1327 Pod's ESB : 3 3rd Qu.:2075.8 3rd Qu.:0.06700   
## Boston Lager : 3 Max. :2692.0 Max. :0.12800   
## (Other) :2376 NA's :62   
## IBU Brewery\_id Style   
## Min. : 4.00 Min. : 1.0 American IPA : 424   
## 1st Qu.: 21.00 1st Qu.: 94.0 American Pale Ale (APA) : 245   
## Median : 35.00 Median :206.0 American Amber / Red Ale : 133   
## Mean : 42.71 Mean :232.7 American Blonde Ale : 108   
## 3rd Qu.: 64.00 3rd Qu.:367.0 American Double / Imperial IPA: 105   
## Max. :138.00 Max. :558.0 American Pale Wheat Ale : 97   
## NA's :1005 (Other) :1298   
## Ounces   
## Min. : 8.40   
## 1st Qu.:12.00   
## Median :12.00   
## Mean :13.59   
## 3rd Qu.:16.00   
## Max. :32.00   
##

plot\_missing(beers\_df, missing\_only = TRUE)



#plot\_histogram(beers\_df)  
#plot\_density(beers\_df)  
# length(unique(beers\_df$Name)) # Check for unique value  
# length(unique(beers\_df$Beer\_ID)) # Check for unique value  
  
#View(breweries\_df) # view the complete dataframe  
names(breweries\_df) # view the variable names

## [1] "Brew\_ID" "Name" "City" "State"

dim(breweries\_df) # view the dimensions of the data

## [1] 558 4

str(breweries\_df) # view the structure of the data

## 'data.frame': 558 obs. of 4 variables:  
## $ Brew\_ID: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Name : Factor w/ 551 levels "10 Barrel Brewing Company",..: 355 12 266 319 201 136 227 477 59 491 ...  
## $ City : Factor w/ 384 levels "Abingdon","Abita Springs",..: 228 200 122 299 300 62 91 48 152 136 ...  
## $ State : Factor w/ 51 levels " AK"," AL"," AR",..: 24 18 20 5 5 41 6 23 23 23 ...

head(breweries\_df, n = 5) # view the first 5 records

## Brew\_ID Name City State  
## 1 1 NorthGate Brewing Minneapolis MN  
## 2 2 Against the Grain Brewery Louisville KY  
## 3 3 Jack's Abby Craft Lagers Framingham MA  
## 4 4 Mike Hess Brewing Company San Diego CA  
## 5 5 Fort Point Beer Company San Francisco CA

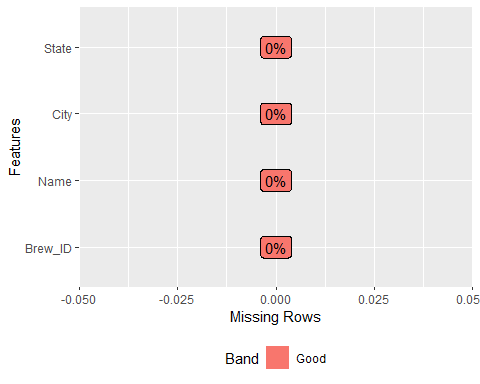
tail(breweries\_df, n = 5) # view the last 5 records

## Brew\_ID Name City State  
## 554 554 Covington Brewhouse Covington LA  
## 555 555 Dave's Brewfarm Wilson WI  
## 556 556 Ukiah Brewing Company Ukiah CA  
## 557 557 Butternuts Beer and Ale Garrattsville NY  
## 558 558 Sleeping Lady Brewing Company Anchorage AK

summary(breweries\_df) # view basic model data i.e. min, max, mdedium, mean, etc.

## Brew\_ID Name City State   
## Min. : 1.0 Blackrocks Brewery : 2 Portland: 17 CO : 47   
## 1st Qu.:140.2 Blue Mountain Brewery : 2 Boulder : 9 CA : 39   
## Median :279.5 Lucette Brewing Company: 2 Chicago : 9 MI : 32   
## Mean :279.5 Oskar Blues Brewery : 2 Seattle : 9 OR : 29   
## 3rd Qu.:418.8 Otter Creek Brewing : 2 Austin : 8 TX : 28   
## Max. :558.0 Sly Fox Brewing Company: 2 Denver : 8 PA : 25   
## (Other) :546 (Other) :498 (Other):358

plot\_missing(breweries\_df)



#plot\_histogram(breweries\_df)  
#plot\_density(breweries\_df)  
# length(unique(breweries\_df$Name)) # Check for unique value  
# length(unique(breweries\_df$Brew\_ID)) # Check for unique value  
  
brewery\_dupes <-  
 as.character(breweries\_df[which(duplicated(as.character(breweries\_df$Name))), "Name"]) # Check for duplicates  
breweries\_df %>% filter(Name == "Summit Brewing Company") # Brew\_ID, City (St. Paul, St Paul)

## Brew\_ID Name City State  
## 1 59 Summit Brewing Company St. Paul MN  
## 2 139 Summit Brewing Company St Paul MN

breweries\_df %>% filter(Name == "Lucette Brewing Company") # Brew\_ID, City (Menominee, Menominie)

## Brew\_ID Name City State  
## 1 378 Lucette Brewing Company Menominee WI  
## 2 457 Lucette Brewing Company Menominie WI

breweries\_df$City <-  
 str\_replace\_all(breweries\_df$City, "St. Paul", "St Paul") ## sync spelling  
breweries\_df$City <-  
 str\_replace\_all(breweries\_df$City, "Menominee", "Menominie") ## sync spelling

## Sanitize Variable Names In Data Sets

*TODO add note about why this is needed*

options(show.error.messages = FALSE)  
try(breweries\_df <- rename(breweries\_df, Brewery\_id = Brew\_ID))  
try(breweries\_df <- rename(breweries\_df, Brewery\_Name = Name))  
try(beers\_df <- rename(beers\_df, Beer\_Name = Name))  
options(show.error.messages = TRUE)

## Get mapping information

states\_df <- map\_data("state") # get US states and their coordinates  
  
#Add Hawaii and Alaska - https://stackoverflow.com/questions/13757771/relocating-alaska-and-hawaii-on-thematic-map-of-the-usa-with-ggplot2

## Question 1:

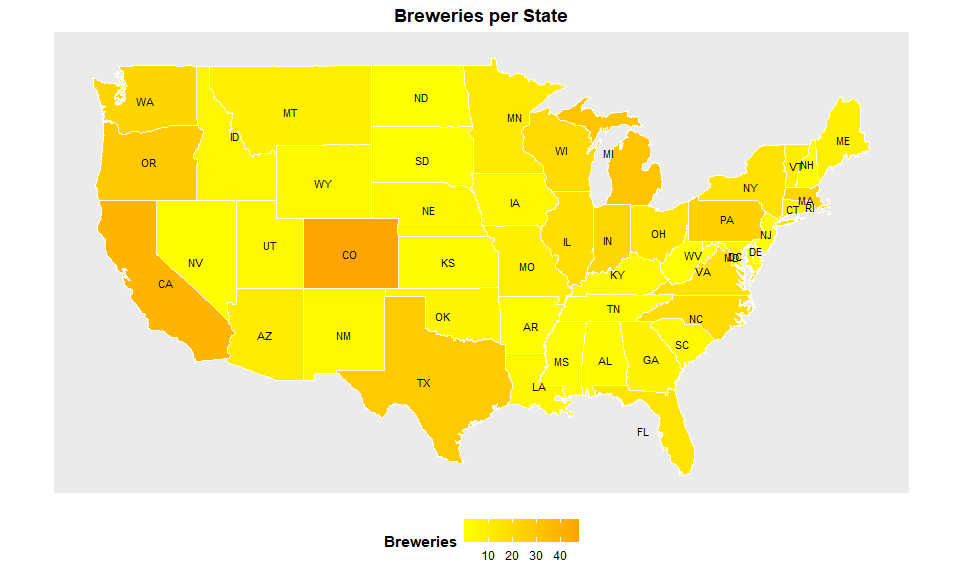
How many breweries are present in each state?

*TODO Finish write up Number of breweries in 51 states has been calculated and provided as the below table, together with a plot with numbers in order.*

brewery\_states\_df <-  
 breweries\_df %>% group\_by(State) %>% distinct(Brewery\_id)  
brewery\_count <- aggregate(Brewery\_id ~ State, breweries\_df, length)  
names(brewery\_count) <- c("State", "Breweries")  
  
brewery\_count$State <-  
 trim(brewery\_count$State, recode.factor = TRUE)  
brewery\_count$region <- tolower(abbr2state(brewery\_count$State))  
  
brewery\_map\_df <- left\_join(states\_df, brewery\_count, by = "region")  
state\_center <-  
 aggregate(  
 cbind(long, lat) ~ State,  
 data = brewery\_map\_df,  
 FUN = function(x)  
 mean(range(x))  
 )  
state\_center\_region <- brewery\_map\_df %>% group\_by(State) %>%  
 summarize\_at(vars(long, lat), ~ mean(range(.)))  
all.equal(state\_center, as.data.frame(state\_center\_region))

## [1] TRUE

ggplot(data = brewery\_map\_df) +  
 geom\_path(aes(x = long, y = lat, group = group), color = "white") +  
 geom\_polygon(aes(  
 x = long,  
 y = lat,  
 group = group,  
 fill = Breweries  
 ), color = "white") +  
 coord\_map() +  
 scale\_fill\_gradient(low = "yellow" , high = "orange") +  
 geom\_text(aes(x = long, y = lat, label = State),  
 data = state\_center,  
 size = 3) +  
 ggtitle("Breweries per State") +  
 theme(  
 axis.text = element\_blank(),  
 axis.line = element\_blank(),  
 axis.ticks = element\_blank(),  
 panel.border = element\_blank(),  
 panel.grid = element\_blank(),  
 axis.title = element\_blank(),  
 legend.position = "bottom",  
 legend.title = element\_text(face = "bold"),  
 legend.direction = "horizontal",  
 plot.title = element\_text(face = "bold",  
 hjust = 0.5)  
 )



## Question 2:

Merge beer data with the breweries data. Print the first 6 observations and the last six observations to check the merged file. (RMD only, this does not need to be included in the presentation or the deck.)

*We merged the two data set into one and named it as ‘mergeDF’, the primary key being used is ‘Brewery\_id’ from Beer data set, and ‘Brew\_ID’ from Breweries data set. We also changed the two columns’ name for clear understanding. The first and last 6 observations were showed there with head/tail command.*

beer\_brewery\_merged\_df <-  
 full\_join(beers\_df, breweries\_df, by = "Brewery\_id")  
head(beer\_brewery\_merged\_df, n = 5) # view the first 5 records

## Beer\_Name Beer\_ID ABV IBU Brewery\_id  
## 1 Pub Beer 1436 0.050 NA 409  
## 2 Devil's Cup 2265 0.066 NA 178  
## 3 Rise of the Phoenix 2264 0.071 NA 178  
## 4 Sinister 2263 0.090 NA 178  
## 5 Sex and Candy 2262 0.075 NA 178  
## Style Ounces Brewery\_Name City State  
## 1 American Pale Lager 12 10 Barrel Brewing Company Bend OR  
## 2 American Pale Ale (APA) 12 18th Street Brewery Gary IN  
## 3 American IPA 12 18th Street Brewery Gary IN  
## 4 American Double / Imperial IPA 12 18th Street Brewery Gary IN  
## 5 American IPA 12 18th Street Brewery Gary IN

tail(beer\_brewery\_merged\_df, n = 5) # view the last 5 records

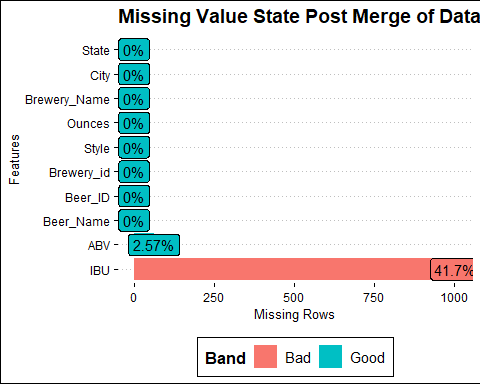
## Beer\_Name Beer\_ID ABV IBU Brewery\_id Style  
## 2406 Belgorado 928 0.067 45 425 Belgian IPA  
## 2407 Rail Yard Ale 807 0.052 NA 425 American Amber / Red Ale  
## 2408 B3K Black Lager 620 0.055 NA 425 Schwarzbier  
## 2409 Silverback Pale Ale 145 0.055 40 425 American Pale Ale (APA)  
## 2410 Rail Yard Ale (2009) 84 0.052 NA 425 American Amber / Red Ale  
## Ounces Brewery\_Name City State  
## 2406 12 Wynkoop Brewing Company Denver CO  
## 2407 12 Wynkoop Brewing Company Denver CO  
## 2408 12 Wynkoop Brewing Company Denver CO  
## 2409 12 Wynkoop Brewing Company Denver CO  
## 2410 12 Wynkoop Brewing Company Denver CO

## Question 3:

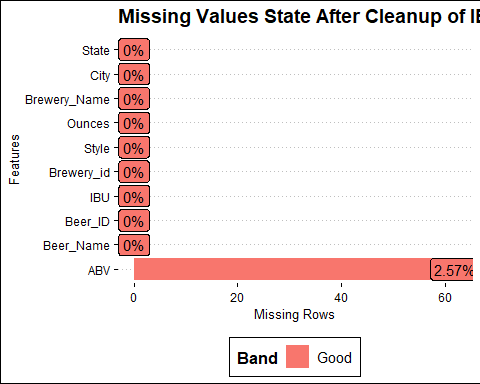
Address the missing values in each column.

*There are 62 observations where both ABV and IBU are NA’s, 943 observations where the IBU only are NA’s. We replaced 62 NA’s in ABV with state level median. Replacing 1005 NA’s in IBU with state level median led to an 18% reduction in the accuracy of the correlation model. Predicted values from simple linear regression model are used to replace missing values in IBU to improve the accuracy of the model.*

plot\_missing(beer\_brewery\_merged\_df,  
 title = "Missing Value State Post Merge of Data Sets",  
 ggtheme = theme\_clean())



#aggr(beer\_brewery\_merged\_df, prop = FALSE, numbers = TRUE)  
beer\_brewery\_merged\_df <-  
 beer\_brewery\_merged\_df %>% mutate(IBU = ifelse(is.na(IBU) |  
 is.na(ABV), 0, IBU))  
#aggr(beer\_brewery\_merged\_df, prop = FALSE, numbers = TRUE)  
  
# ibu\_na <- beer\_brewery\_merged\_df %>% filter(is.na(IBU))  
# abv\_na <- beer\_brewery\_merged\_df %>% filter(is.na(ABV))  
# ibu\_abv\_na <-  
# beer\_brewery\_merged\_df %>% filter(is.na(ABV) & is.na(IBU))  
plot\_missing(beer\_brewery\_merged\_df,  
 title = "Missing Values State After Cleanup of IBU",  
 ggtheme = theme\_clean())



## Question 4:

Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.

*Median ABV and IBU data have been sorted and plotted.*  
*The median ABV per state appears fairly consistent with an overall ABV median of 0.056.*  
*Kentucky has the highest median ABV at 0.062 and Utah has the lowest at 0.04.*  
*The median IBU per state appears to change considerably between states with an overall IBU median of 37.*  
*West Virginia has the highest median IBU at 57.5 and Kansas has the lowest at 22.*

Arkansas and Utah are tied for lowest ABV at 4.0% Maine has highest ABV at 6.7%

Wisconsin has the lowest IBU @ 19 Maine has the highest IBU @ 61

## Calculate IBU and ABV median by state  
  
ibu\_median <-  
 aggregate(  
 beer\_brewery\_merged\_df$IBU,  
 by = list(beer\_brewery\_merged\_df$State),  
 FUN = mean  
 )  
colnames(ibu\_median) <- c("State", "median")  
ibu\_median <- filter(ibu\_median, ibu\_median$median > 0)  
ibu\_median <- ibu\_median[order(ibu\_median$median),]  
ibu\_median$State <-  
 factor(ibu\_median$State, levels = ibu\_median$State)  
head(ibu\_median, 5)

## State median  
## 3 AR 7.800000  
## 23 MI 8.617284  
## 7 CT 9.074074  
## 41 SC 10.785714  
## 30 NE 11.040000

tail(ibu\_median, 5)

## State median  
## 24 MN 41.78182  
## 2 AL 46.10000  
## 32 NJ 46.37500  
## 26 MS 46.45455  
## 49 WV 57.50000

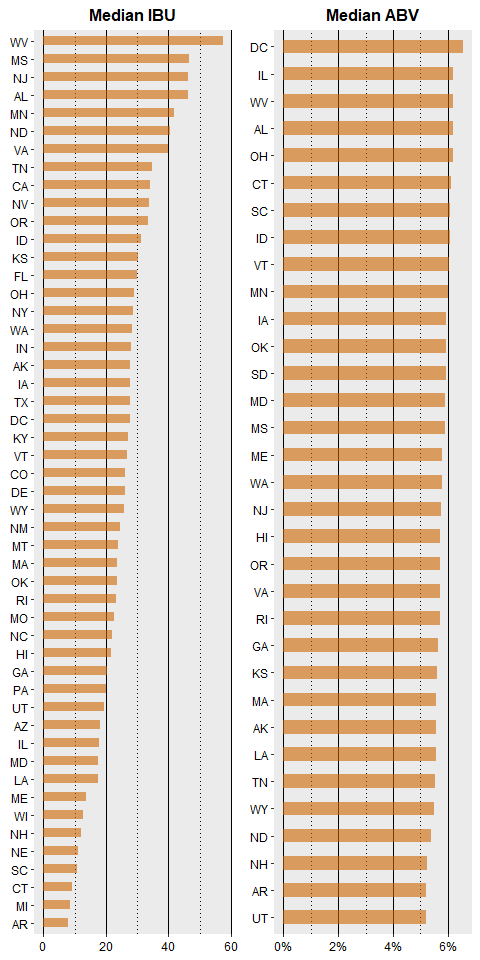
abv\_median <-  
 aggregate(  
 beer\_brewery\_merged\_df$ABV,  
 by = list(beer\_brewery\_merged\_df$State),  
 FUN = mean  
 )  
colnames(abv\_median) <- c("State", "median")  
abv\_median <- filter(abv\_median, abv\_median$median > 0)  
abv\_median <- abv\_median[order(abv\_median$median),]  
abv\_median$State <-  
 factor(abv\_median$State, levels = abv\_median$State)  
head(abv\_median, 5)

## State median  
## 28 UT 0.05188462  
## 3 AR 0.05200000  
## 19 NH 0.05237500  
## 18 ND 0.05400000  
## 33 WY 0.05486667

tail(abv\_median, 5)

## State median  
## 21 OH 0.06195918  
## 2 AL 0.06200000  
## 32 WV 0.06200000  
## 10 IL 0.06202198  
## 5 DC 0.06562500

p1 <-  
 ggplot(ibu\_median, aes(  
 x = State,  
 y = median,  
 width = 0.5  
 )) +  
 geom\_bar(  
 stat = "identity",  
 fill = "darkorange3",  
 alpha = .6,  
 width = .4  
 ) +  
 coord\_flip() +  
 ggtitle("IBU Median") +  
 theme(  
 panel.grid.major.y = element\_blank(),  
 plot.title = element\_text(face = "bold", size = 12),  
 axis.title = element\_blank()  
 ) + theme(  
 panel.grid.major = element\_line(colour = "black"),  
 panel.grid.minor = element\_line(colour = "black",  
 linetype = "dotted"),  
 axis.text = element\_text(colour = "black"),  
 plot.title = element\_text(hjust = 0.5),  
 legend.position = "none"  
 ) + labs(title = "Median IBU")  
  
p2 <-  
 ggplot(abv\_median, aes(  
 x = State,  
 y = median,  
 width = 0.5  
 )) +  
 geom\_bar(  
 stat = "identity",  
 fill = "darkorange3",  
 alpha = .6,  
 width = .4  
 ) +  
 scale\_y\_continuous(labels = scales::percent\_format(accuracy = 1)) +  
 coord\_flip() +  
 ggtitle("ABV Median") +  
 theme(  
 panel.grid.major.y = element\_blank(),  
 plot.title = element\_text(face = "bold", size = 12),  
 axis.title = element\_blank()  
 ) + theme(  
 panel.grid.major = element\_line(colour = "black"),  
 panel.grid.minor = element\_line(colour = "black",  
 linetype = "dotted"),  
 axis.text = element\_text(colour = "black"),  
 plot.title = element\_text(hjust = 0.5),  
 legend.position = "none"  
 ) + labs(title = "Median ABV")  
  
grid.arrange(p1, p2, ncol = 2)



# plot IBU and ABV median by state in a map  
# ibu\_median$State <-  
# trim(ibu\_median$State, recode.factor = TRUE)  
# ibu\_median$region <- tolower(abbr2state(ibu\_median$State))  
# ibu\_map\_df <- left\_join(states\_df, ibu\_median, by = "region")  
# ggplot(ibu\_map\_df, aes(x = long, y = lat), color = white) +  
# ggtitle("IBU median by State") +  
# geom\_polygon(aes(fill = median, group = group), stat = "identity") +  
# geom\_path() +  
# scale\_fill\_gradientn(colours = rainbow(5)) +  
# coord\_map() +  
# theme(axis.text = element\_blank(),  
# axis.line = element\_blank(),  
# axis.ticks = element\_blank(),  
# panel.border = element\_blank(),  
# panel.grid = element\_blank(),  
# axis.title = element\_blank())  
#  
# abv\_median$State <-  
# trim(abv\_median$State, recode.factor = TRUE)  
# abv\_median$region <- tolower(abbr2state(abv\_median$State))  
# abv\_map\_df <- left\_join(states\_df, ibu\_median, by = "region")  
# ggplot(abv\_map\_df, aes(x = long, y = lat), color = white) +  
# ggtitle("ABV median by State") +  
# geom\_polygon(aes(fill = median, group = group), stat = "identity") +  
# geom\_path() +  
# scale\_fill\_gradientn(colours = rainbow(5)) +  
# coord\_map() +  
# theme(axis.text = element\_blank(),  
# axis.line = element\_blank(),  
# axis.ticks = element\_blank(),  
# panel.border = element\_blank(),  
# panel.grid = element\_blank(),  
# axis.title = element\_blank())

## Question 5:

Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?

*Maximum ABV and IBU data have been sorted and plotted. The maximum ABV per state appears to vary between states with an overall median of all max values at 0.09. Colorado has the highest Max ABV at 0.128 and Delaware has the lowest at 0.055. The maximum IBU per state appears to vary between states with an overall median of all max values at 92.82. Oregon has the highest Max IBU at 138 and Arkansas has the lowest at 44.11*

ibu\_max <-  
 aggregate(  
 beer\_brewery\_merged\_df$IBU,  
 by = list(beer\_brewery\_merged\_df$State),  
 FUN = max  
 )  
colnames(ibu\_max) <- c("State","max")  
ibu\_top <- ibu\_max %>% top\_n(10, max)  
ibu\_top <- ibu\_top[order(ibu\_top$max),]  
ibu\_top$State <-  
 factor(ibu\_top$State, levels = ibu\_top$State)  
head(ibu\_top, 5)

## State max  
## 1 CA 115  
## 2 DC 115  
## 3 IN 115  
## 5 MI 115  
## 9 TX 118

tail(ibu\_top, 5)

## State max  
## 11 VT 120  
## 7 OH 126  
## 4 MA 130  
## 10 VA 135  
## 8 OR 138

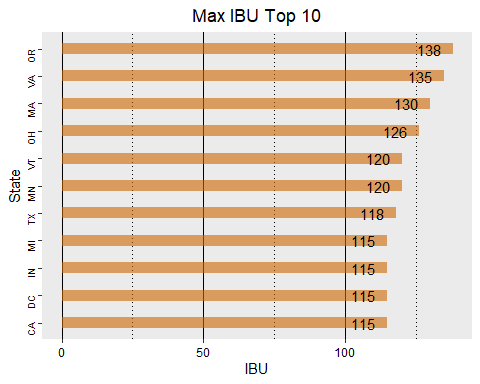
abv\_max <-  
 aggregate(  
 beer\_brewery\_merged\_df$ABV,  
 by = list(beer\_brewery\_merged\_df$State),  
 FUN = max  
 )  
colnames(abv\_max) <- c("State", "max")  
abv\_top <- abv\_max %>% na.omit() %>% top\_n(10, max)  
abv\_top <- abv\_top[order(abv\_top$max),]  
abv\_top$State <-  
 factor(abv\_top$State, levels = abv\_top$State)  
head(abv\_top, 5)

## State max  
## 1 IA 0.095  
## 3 IL 0.096  
## 10 VT 0.096  
## 9 SC 0.097  
## 2 ID 0.099

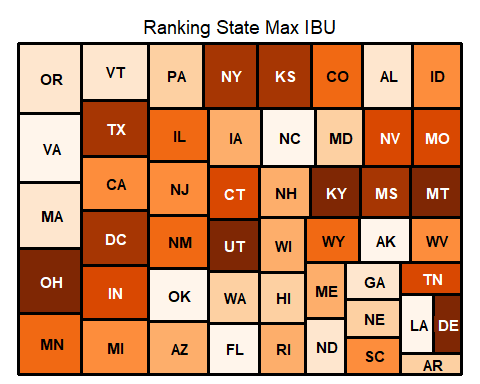
tail(abv\_top, 5)

## State max  
## 4 MA 0.099  
## 5 ME 0.099  
## 6 MN 0.099  
## 7 NJ 0.099  
## 8 OH 0.099

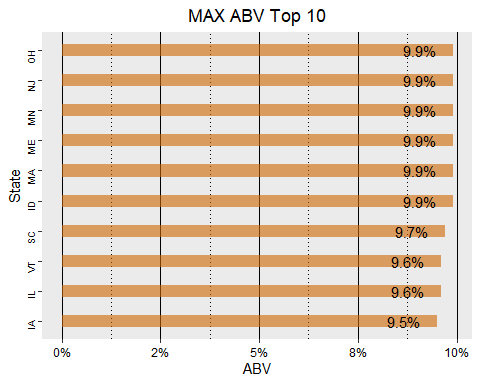
# plot IBU top by state and beer  
ggplot(ibu\_top, aes(x = State, y = max, fill = max)) + geom\_bar(  
 stat = "identity",  
 fill = "darkorange3",  
 alpha = .6,  
 width = .4  
) + ggtitle("Max IBU Top 10") +  
 geom\_text(aes(label = max), hjust = 1.5) +  
 theme(  
 axis.text.y = element\_text(size = rel(0.8), angle = 90),  
 panel.grid.major.y = element\_blank(),  
 panel.grid.major = element\_line(colour = "black"),  
 panel.grid.minor = element\_line(colour = "black",  
 linetype = "dotted"),  
 axis.text = element\_text(colour = "black"),  
 plot.title = element\_text(hjust = 0.5)  
 ) +  
 labs(x = "State", y = "IBU") +  
 coord\_flip()



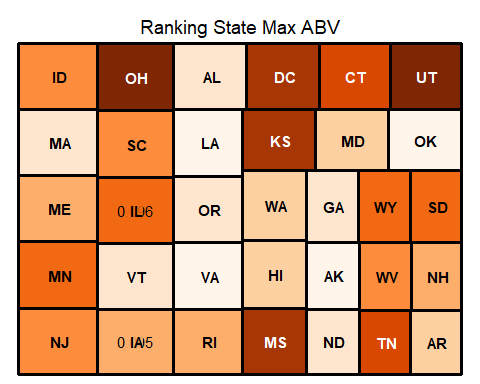
treemap(  
 ibu\_max,  
 index = c("State", "max"),  
 vSize = "max",  
 type = "index",  
 palette = "Oranges",  
 title = "Ranking State Max IBU",  
 fontsize.title = 14  
)



# ggplot(ibu\_top, aes(x=Beer\_Name, y=IBU, fill=State)) + geom\_bar(stat = "identity") + ggtitle("Top 5 IBU by Beer")+  
# geom\_text(aes(label=IBU), hjust=1.5) +  
# theme(axis.text.y = element\_text(size=rel(0.8))) +  
# labs(x="Beer",y="IBU") +  
# coord\_flip()  
  
# plot ABV top by state and beer  
ggplot(abv\_top, aes(x = State, y = max, fill = max)) + geom\_bar(  
 stat = "identity",  
 fill = "darkorange3",  
 alpha = .6,  
 width = .4  
) + ggtitle("MAX ABV Top 10") +  
 geom\_text(aes(label = scales::percent(max, accuracy = .1)), hjust = 1.5) +  
 scale\_y\_continuous(labels = scales::percent\_format(accuracy = 1)) +  
 theme(  
 axis.text.y = element\_text(size = rel(0.8), angle = 90),  
 panel.grid.major.y = element\_blank(),  
 panel.grid.major = element\_line(colour = "black"),  
 panel.grid.minor = element\_line(colour = "black",  
 linetype = "dotted"),  
 axis.text = element\_text(colour = "black"),  
 plot.title = element\_text(hjust = 0.5)  
 ) +  
 labs(x = "State", y = "ABV") +  
 coord\_flip()



treemap(  
 abv\_max,  
 index = c("State", "max"),  
 vSize = "max",  
 type = "index",  
 palette = "Oranges",  
 title = "Ranking State Max ABV",  
 fontsize.title = 14  
)



# ggplot(abv\_top, aes(x=Beer\_Name, y=ABV, fill=State)) + geom\_bar(stat = "identity") + ggtitle("Top 5 ABV by Beer")+  
# geom\_text(aes(label=ABV), hjust=1.5) +  
# theme(axis.text.y = element\_text(size=rel(0.8))) +  
# labs(x="Beer",y="ABV") +  
# coord\_flip()

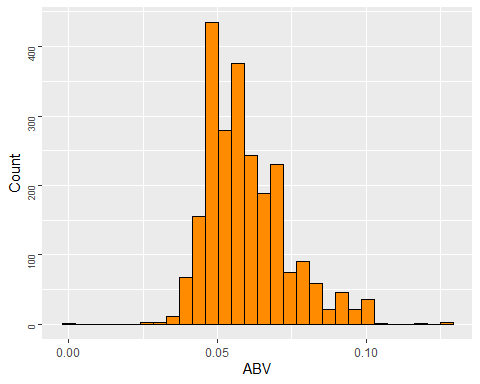
## Question 6:

Comment on the summary statistics and distribution of the ABV variable.

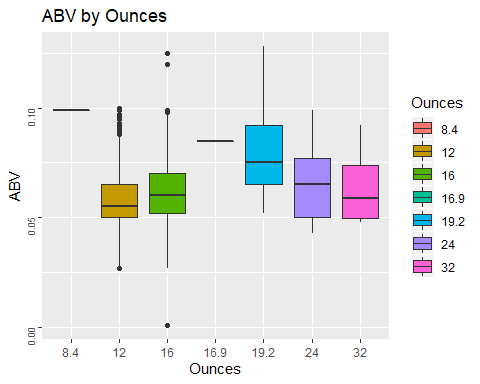
*The ABV clearly illustrates that the IPA variety has more Alcohol By Volume than other varieties. This is across all can sizes with one exception (can size=19.2) in the “other” type. The ABV per size of can (oz) clearly illustrates that the IPA variety has more Alcohol By Volume than other varieties. This across all can sizes with one exception (can size=19.2).*

*IPA style beer is the predominant variety when ABV exceeds 0.06. This again indicates that IPA beer tends to have higher alcohol content than other variety.*

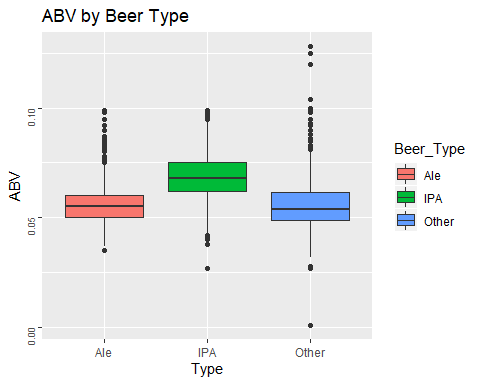
# Create new column for IPA/Ale breakdown  
beer\_brewery\_merged\_df$Beer\_Type <-  
 ifelse(str\_detect(beer\_brewery\_merged\_df$Style, regex("IPA", ignore\_case = TRUE)),  
 "IPA",  
 ifelse(str\_detect(  
 beer\_brewery\_merged\_df$Style, regex("Ale", ignore\_case = TRUE)  
 ), "Ale", "Other"))  
  
## Histogram  
beer\_brewery\_merged\_df %>% ggplot(aes(x = ABV)) + geom\_histogram(  
 fill = "darkorange",  
 color =  
 "black",  
 bins = 30,  
 na.rm = TRUE  
) + theme(axis.text.y = element\_text(size = rel(0.8), angle = 90)) +  
 labs(x = "ABV", y = "Count" )



## Bloxplot  
beer\_brewery\_merged\_df$Ounces = as.factor(beer\_brewery\_merged\_df$Ounces)  
beer\_brewery\_merged\_df %>% ggplot(aes(  
 y = ABV,  
 x = Ounces, fill = Ounces  
)) + geom\_boxplot(na.rm = TRUE) +  
 theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
 labs(x="Ounces",y="ABV") +  
 ggtitle("ABV by Ounces")



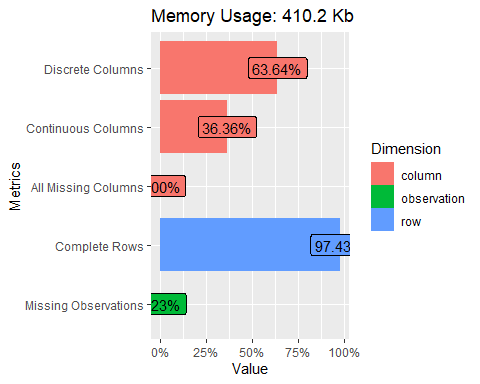
beer\_brewery\_merged\_df %>% ggplot(aes(  
 y = ABV,  
 x = Beer\_Type, fill = Beer\_Type  
)) + geom\_boxplot(na.rm = TRUE) +  
 theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
 labs(x="Type",y="ABV") +  
 ggtitle("ABV by Beer Type")



summary(beer\_brewery\_merged\_df)

## Beer\_Name Beer\_ID ABV   
## Nonstop Hef Hop : 12 Min. : 1.0 Min. :0.00100   
## Dale's Pale Ale : 6 1st Qu.: 808.2 1st Qu.:0.05000   
## Oktoberfest : 6 Median :1453.5 Median :0.05600   
## Longboard Island Lager: 4 Mean :1431.1 Mean :0.05977   
## 1327 Pod's ESB : 3 3rd Qu.:2075.8 3rd Qu.:0.06700   
## Boston Lager : 3 Max. :2692.0 Max. :0.12800   
## (Other) :2376 NA's :62   
## IBU Brewery\_id Style   
## Min. : 0.0 Min. : 1.0 American IPA : 424   
## 1st Qu.: 0.0 1st Qu.: 94.0 American Pale Ale (APA) : 245   
## Median : 18.0 Median :206.0 American Amber / Red Ale : 133   
## Mean : 24.9 Mean :232.7 American Blonde Ale : 108   
## 3rd Qu.: 41.0 3rd Qu.:367.0 American Double / Imperial IPA: 105   
## Max. :138.0 Max. :558.0 American Pale Wheat Ale : 97   
## (Other) :1298   
## Ounces Brewery\_Name City   
## 8.4 : 1 Brewery Vivant : 62 Length:2410   
## 12 :1525 Oskar Blues Brewery : 46 Class :character   
## 16 : 841 Sun King Brewing Company : 38 Mode :character   
## 16.9: 1 Cigar City Brewing Company: 25   
## 19.2: 15 Sixpoint Craft Ales : 24   
## 24 : 22 Hopworks Urban Brewery : 23   
## 32 : 5 (Other) :2192   
## State Beer\_Type   
## CO : 265 Length:2410   
## CA : 183 Class :character   
## MI : 162 Mode :character   
## IN : 139   
## TX : 130   
## OR : 125   
## (Other):1406

plot\_intro(beer\_brewery\_merged\_df)



introduce(beer\_brewery\_merged\_df)

## rows columns discrete\_columns continuous\_columns all\_missing\_columns  
## 1 2410 11 7 4 0  
## total\_missing\_values complete\_rows total\_observations memory\_usage  
## 1 62 2348 26510 420056

## Stacked histogram  
# beer\_brewery\_merged\_df %>% ggplot(aes(x = ABV, fill = Beer\_Type)) + geom\_bar(color = "black") + xlim(c(0.025, 0.100)) +  
# theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="",y="ABV") + ggtitle("Beer Type vs ABV")  
#   
# beer\_brewery\_merged\_df %>% ggplot(aes(x = ABV, fill = as.factor(Ounces))) + geom\_bar(color =  
# "black") + xlim(c(0.025, 0.100)) +  
# theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="",y="ABV") + ggtitle("Ounce vs ABV")  
  
# Scatter plot ABV vs. IBU  
# beer\_brewery\_merged\_df %>% na.omit() %>% ggplot(aes(x = IBU, y = ABV)) + geom\_point(shape=0) + ggtitle("ABV vs IBU") + geom\_smooth(method = lm) + theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="IBU",y="ABV")  
# beer\_brewery\_merged\_df %>% ggplot(aes(  
# y = ABV,  
# x = as.factor(Ounces),  
# fill = Beer\_Type  
# )) + geom\_boxplot() +  
# theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="Ounces",y="ABV") + ggtitle("Boxplot of ABV by Ounces")

## Question 7:

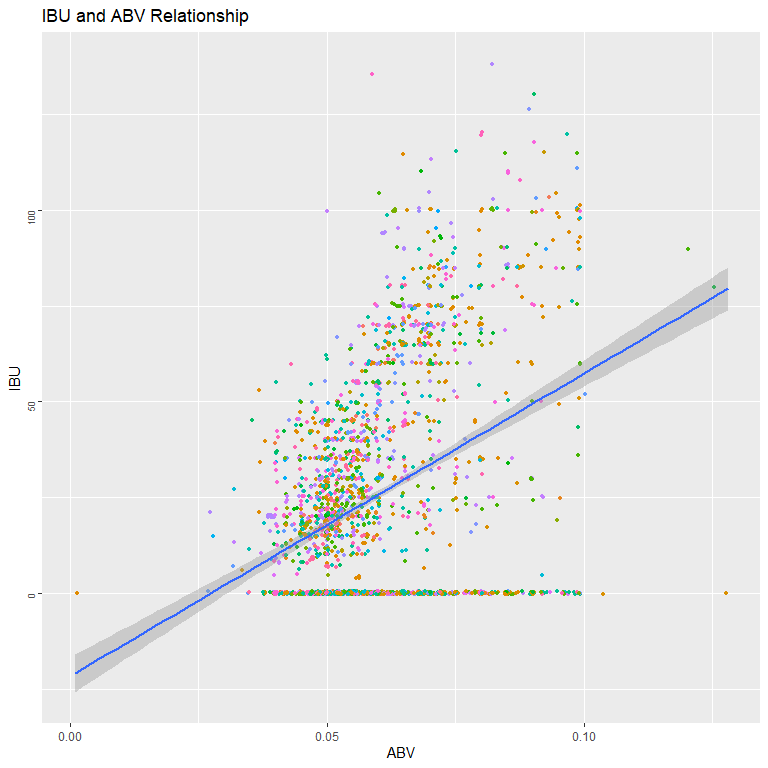
Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot. Make your best judgment of a relationship and EXPLAIN your answer.

*The point plot below illustrates what appears to a positive linear relationship between the ABB and IBU. The addition of line trend to the point plot below confirms the presence of a positive linear relationship between the ABV and IBU.*

*The correlation between IBU and ABV is 0.757878, with a p-value < 2e-16, Multiple R-squared is 0.5744, meaning that 57% changes in exploratory variable ABV can explain the changes in response variable IBU.*

*GGPairs Plot below shows a strong correlation between ABV and IBU across all styles of beers. The strongest correlation is in other beer style category at 0.787 followed by IPA at 0.689.*

# beer\_brewery\_merged\_df %>% na.omit() %>% ggplot(aes(x=IBU, y=ABV, color=Beer\_Type)) + geom\_point() + theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="State",y="IBU") +ggtitle("ABV vs IBU") + facet\_wrap(~State)  
# beer\_brewery\_merged\_df %>% na.omit() %>% ggplot(aes(x=IBU, y=ABV, color=Beer\_Type)) + geom\_point() + theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="State",y="IBU") +ggtitle("ABV vs IBU") + facet\_wrap(~as.factor(Ounces))  
  
# Scatter plot and Pearson's correlation with original data  
beer\_brewery\_merged\_df %>% na.omit() %>%  
 ggplot(aes(x=ABV, y=IBU)) +   
 geom\_point(aes(color=State), size=1, position="jitter") +   
 geom\_smooth(method = lm) +   
 theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
 labs(x="ABV",y="IBU") +  
 ggtitle("IBU and ABV Relationship") + theme(plot.subtitle = element\_text(vjust = 1),   
 plot.caption = element\_text(vjust = 1),   
 legend.position = "none")



cor.test(beer\_brewery\_merged\_df$IBU, beer\_brewery\_merged\_df$ABV)

##   
## Pearson's product-moment correlation  
##   
## data: beer\_brewery\_merged\_df$IBU and beer\_brewery\_merged\_df$ABV  
## t = 19.228, df = 2346, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3334942 0.4033996  
## sample estimates:  
## cor   
## 0.3689686

mod <- lm(beer\_brewery\_merged\_df$IBU ~beer\_brewery\_merged\_df$ABV)  
summary(mod)

##   
## Call:  
## lm(formula = beer\_brewery\_merged\_df$IBU ~ beer\_brewery\_merged\_df$ABV)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.49 -21.00 -1.81 18.84 110.05   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -21.688 2.519 -8.608 <2e-16 \*\*\*  
## beer\_brewery\_merged\_df$ABV 790.436 41.109 19.228 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.97 on 2346 degrees of freedom  
## (62 observations deleted due to missingness)  
## Multiple R-squared: 0.1361, Adjusted R-squared: 0.1358   
## F-statistic: 369.7 on 1 and 2346 DF, p-value: < 2.2e-16

#Correlations  
# brew\_corr <- inspect\_cor(beer\_brewery\_merged\_df)  
# show\_plot (brew\_corr) + theme(axis.text.y = element\_text(size=rel(0.8), angle=90)) +  
# labs(x="State",y="IBU") +ggtitle("Correlation Between All Attributes")  
  
# Plot GGpairs  
beer\_brewery\_merged\_df %>% na.omit() %>%  
select(ABV, IBU, Beer\_Type, Ounces) %>%  
ggpairs(mapping = aes(color = "ABV"))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

