l— output: word\_document: default html\_document: default —

# Questions - this chunk just holds all the questions for my reading  
  
# 1. How many breweries are present in each state?  
  
# 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.)  
  
# 3. Address the missing values in each column.  
  
# 4. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.  
  
# 5. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?  
  
# 6. Comment on the summary statistics and distribution of the ABV variable.  
  
# 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.  
  
# 8. 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.   
  
# 9. 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  
  
  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.4  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## Warning: package 'dplyr' was built under R version 4.0.4

## Warning: package 'forcats' was built under R version 4.0.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(maps)

##   
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':  
##   
## map

library(mapproj)  
library(usmap)  
library(curl)

##   
## Attaching package: 'curl'

## The following object is masked from 'package:readr':  
##   
## parse\_date

library(class)  
library(e1071)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

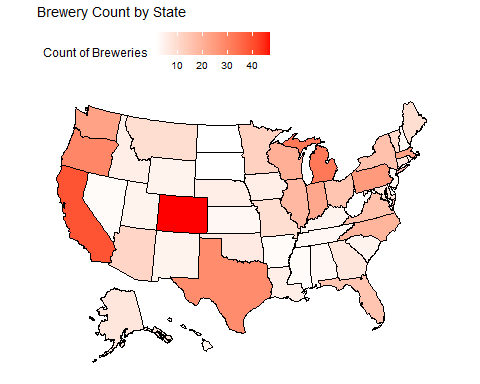
library(fuzzyjoin)  
library(RCurl)

##   
## Attaching package: 'RCurl'

## The following object is masked from 'package:tidyr':  
##   
## complete

Answering question 1 by creating a heat map: How many breweries are there for each state? Answer: top 5 are Colorado, California, Michigan, Oregon, Texas

# write up of the code  
  
# Bringing in the brewery data  
brew\_data <- read.csv('https://raw.githubusercontent.com/scottdyl/MSDS6306\_CastStudy1/main/data/Breweries.csv')  
  
# Counting each brewery by each state and renaming the count column to count  
brewery\_count = count(brew\_data,State)  
  
#trimming the extra spaces off of state  
brewery\_count$State = str\_trim(brewery\_count$State)  
  
#getting the fips for each state for my map  
brewery\_count$fips = fips(brewery\_count$State)  
  
#renaming the 'n' column to count   
colnames(brewery\_count)[2] = 'Count'  
  
# creating the heat map  
plot\_usmap(data=brewery\_count, values="Count", color = "black") + scale\_fill\_gradient(name = "Count of Breweries", low = "white", high = "red") + labs(title="Brewery Count by State") + theme(legend.position = "top")



# getting the top 3 states by brewery for presentation  
brewery\_count$State[order(brewery\_count$Count,decreasing = T)[1:5]]

## [1] "CO" "CA" "MI" "OR" "TX"

Merging the two datasets from github and printing the first and last 6.

brew\_data <- read.csv('https://raw.githubusercontent.com/scottdyl/MSDS6306\_CastStudy1/main/data/Breweries.csv')  
beer\_data <- read.csv('https://raw.githubusercontent.com/scottdyl/MSDS6306\_CastStudy1/main/data/Beers.csv')  
  
# changed the ID column name to match in each dataset and did a full join  
colnames(beer\_data)[5] = 'Brew\_ID'  
bud\_data <- full\_join(brew\_data,beer\_data,'Brew\_ID')  
  
head(bud\_data,6)

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

tail(bud\_data,6)

## Brew\_ID Name.x City State  
## 2405 556 Ukiah Brewing Company Ukiah CA  
## 2406 557 Butternuts Beer and Ale Garrattsville NY  
## 2407 557 Butternuts Beer and Ale Garrattsville NY  
## 2408 557 Butternuts Beer and Ale Garrattsville NY  
## 2409 557 Butternuts Beer and Ale Garrattsville NY  
## 2410 558 Sleeping Lady Brewing Company Anchorage AK  
## Name.y Beer\_ID ABV IBU Style Ounces  
## 2405 Pilsner Ukiah 98 0.055 NA German Pilsener 12  
## 2406 Heinnieweisse Weissebier 52 0.049 NA Hefeweizen 12  
## 2407 Snapperhead IPA 51 0.068 NA American IPA 12  
## 2408 Moo Thunder Stout 50 0.049 NA Milk / Sweet Stout 12  
## 2409 Porkslap Pale Ale 49 0.043 NA American Pale Ale (APA) 12  
## 2410 Urban Wilderness Pale Ale 30 0.049 NA English Pale Ale 12

Address the missing values in each column we removed 1005 beers due to missing values

# is it best to just filter out the missing data?  
# how else would I fill it in?   
#unclean data count  
dim(bud\_data)

## [1] 2410 10

#summing possible missing values to see how much data we will lose  
sum(is.na(bud\_data$ABV))

## [1] 62

sum(is.na(bud\_data$IBU))

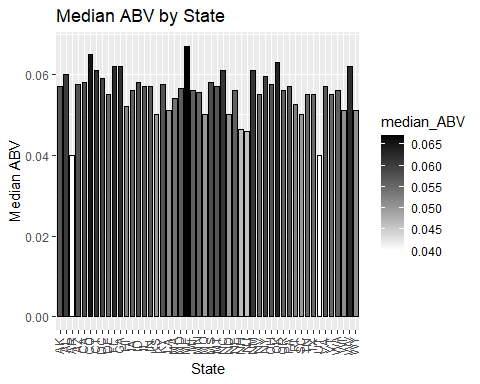
## [1] 1005

#filtering out the missing values and leaving the full data in new dataframe  
bud\_data\_clean<- filter(bud\_data,!is.na(bud\_data$ABV)&!is.na(bud\_data$IBU)&!is.na(bud\_data$Style))  
colnames(bud\_data\_clean)[2] = "Brew Name"  
colnames(bud\_data\_clean)[5] = "Beer Name"  
#clean data count  
dim(bud\_data\_clean)

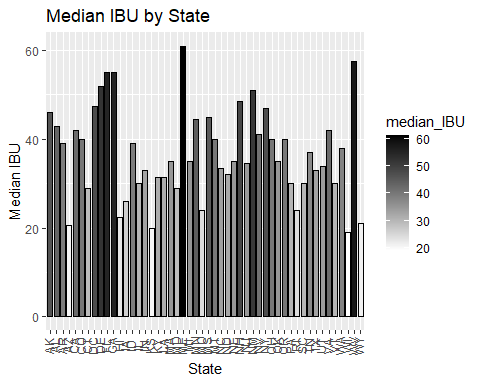
## [1] 1405 10

we would like to find the median alcohol content and IBU for each state this will be presented in a bar chart ggplot used to show the median ABV and IBU for each state using the clean data

# I pipe in the data and group by state then summarize by median abv giving me median abv by state  
bud\_data\_clean %>%   
 group\_by(State) %>%  
 summarise(median\_ABV = median(ABV)) %>%  
 #I use ggplot bar to plot the data  
 ggplot(aes(State,median\_ABV,fill = median\_ABV)) +  
 geom\_bar(stat = 'identity',width = .75,color = 'black')+  
 scale\_fill\_gradient(low = 'white', high = 'black') +  
 labs(title="Median ABV by State", x="State", y="Median ABV")+  
 theme(axis.text.x = element\_text(angle = 90, vjust = .3))



# This code does the same as the ABV but by IBU  
bud\_data\_clean %>%   
 group\_by(State) %>%  
 summarise(median\_IBU = median(IBU)) %>%  
 #I use ggplot bar to plot the data  
 ggplot(aes(State,median\_IBU,fill = median\_IBU)) +  
 geom\_bar(stat = 'identity',width = .75,color = 'black')+  
 scale\_fill\_gradient(low = 'white', high = 'black') +  
 labs(title="Median IBU by State", x="State", y="Median IBU")+  
 theme(axis.text.x = element\_text(angle = 90, vjust = .3))



we will calculate which state has highest ABV and IBU beer finding the max and selecting the row

# use grep to find the max value then select the columns from the data to show which city, state, name of the beer and name of the brewery  
bud\_data\_clean[grep(max(bud\_data\_clean$ABV),bud\_data\_clean$ABV),] %>% select(City, State, `Beer Name`, `Brew Name`, ABV, Style)

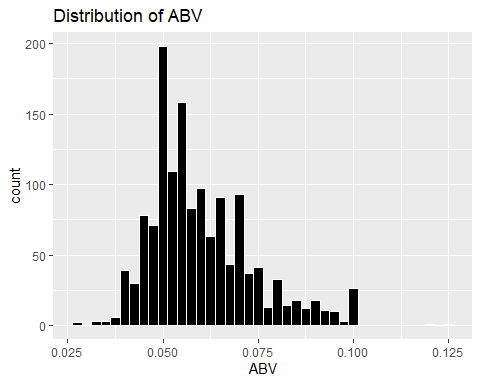
## City State Beer Name Brew Name ABV  
## 8 Louisville KY London Balling Against the Grain Brewery 0.125  
## Style  
## 8 English Barleywine

bud\_data\_clean[grep(max(bud\_data\_clean$IBU),bud\_data\_clean$IBU),] %>% select(City, State, `Beer Name`, `Brew Name`, IBU, Style)

## City State Beer Name Brew Name IBU  
## 1134 Astoria OR Bitter Bitch Imperial IPA Astoria Brewing Company 138  
## Style  
## 1134 American Double / Imperial IPA

Comment on the summary statistics and distribution of the ABV variable. for this I will be using a histogram to show the distribution we showed how the distribution of ABV was by using a histogram

# here I pipe the budwiser data into a ggplot using the variable ABV and create a histogram  
bud\_data\_clean %>%  
 ggplot(aes(ABV)) +  
 geom\_histogram(fill ='black', binwidth = .0025, color = 'white') +  
 labs(title="Distribution of ABV")



summary(bud\_data\_clean$ABV)

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

#summarize the data by min, max, mean, median, and standard deviation  
summarise(bud\_data\_clean,  
 max=max(ABV),  
 min=min(ABV),  
 mean = mean(ABV),  
 median = median(ABV),  
 sd=sd(ABV))

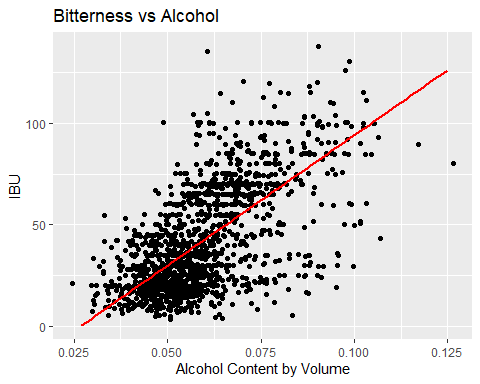
## max min mean median sd  
## 1 0.125 0.027 0.05991388 0.057 0.01357633

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.

Answer: their appears to be a direct correlation between ABV and IBU. this is presented in the graph below with a strong positive relationship.

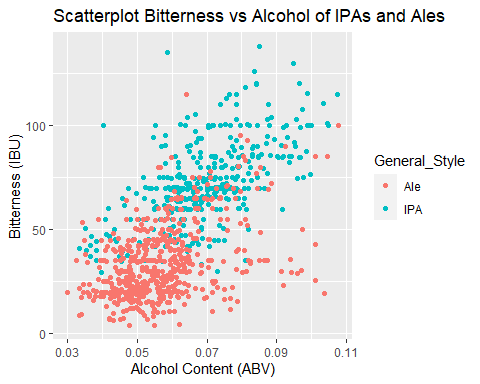
bud\_data\_clean %>%  
 ggplot(aes(ABV, IBU)) +  
 geom\_jitter(width = .01) +  
 geom\_smooth(method = 'lm', se = F, color = 'red')+  
 labs(title="Bitterness vs Alcohol", x="Alcohol Content by Volume", y="IBU")

## `geom\_smooth()` using formula 'y ~ x'

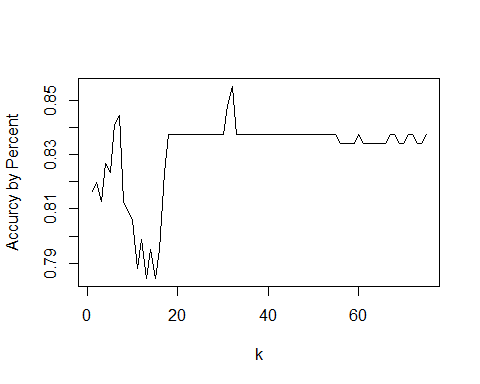
 # 8. Budweiser would also like to investigate the difference with respect to IBU and ABV between IPAs (India Pale Ales) and other types of Ale (any beer with “Ale” in its name other than IPA). You decide to use KNN classification to investigate this relationship. Provide statistical evidence one way or the other. You can of course assume your audience is comfortable with percentages … KNN is very easy to understand conceptually.

Answer: IBU is a better variable to tell IPA from an Ale over ABV. Overall this model was 86% accurate. IBU was a better descriptor than ABV shown by the clustering of the two beers

# include IPA into one category and ales into another but exclude "india pale ale" from the ale section  
# test example is "english india pale ale (IPA)"  
# how to do this? filter out all the IPAs first to ensure you don't get them put into the ale section - filter out the ales and merge with IPA data  
  
IPA\_Brew <- bud\_data\_clean %>% filter(grepl("IPA",Style) | grepl("India Pale Ale", Style))  
IPA\_Brew$General\_Style<-"IPA"  
  
Ale\_Brew <- bud\_data\_clean[grepl("Ale",bud\_data\_clean$Style) & !grepl("India Pale Ale",bud\_data\_clean$Style) & !grepl("IPA", bud\_data\_clean$Style),]  
Ale\_Brew$General\_Style<-"Ale"  
  
Beer\_train<- rbind(IPA\_Brew,Ale\_Brew)  
  
Beer\_train %>% ggplot(aes(ABV, IBU, color = General\_Style)) + geom\_jitter(width=.01) + labs(title="Scatterplot Bitterness vs Alcohol of IPAs and Ales", x="Alcohol Content (ABV)", y="Bitterness (IBU)")



#KNN train where we set seed to 99 and do a 70/30 split of the data  
set.seed(99)  
split\_data = .70  
train\_split = sample(1:dim(Beer\_train)[1],round(split\_data \* dim(Beer\_train)[1]))  
train = Beer\_train[train\_split,]  
test = Beer\_train[-train\_split,]  
  
  
#setting up the data frame to find the most accurate K value  
accs = data.frame(accuracy = numeric(75), k = numeric(75))  
  
# for loop that runs 75 times over upping the k value by 1 each time k=1:75  
for(i in 1:75)  
{  
 classifications = knn(train[,c(7,8)],test[,c(7,8)],train$General\_Style, prob = TRUE, k = i)  
 table(test$General\_Style,classifications)  
 CM = confusionMatrix(table(test$General\_Style,classifications))  
 accs$accuracy[i] = CM$overall[1]  
 accs$k[i] = i  
}  
# plotting k and accuracy to see visually how k performs over time  
plot(accs$k,accs$accuracy, type = "l", xlab = "k", ylab = "Accurcy by Percent")



#filtering by the best accuracy value  
filter(accs, accuracy == max(accs$accuracy))

## accuracy k  
## 1 0.8551237 32

#we will use K = 32 here because that gave us the best results given seed 99  
classifications = knn(train[,c(7,8)],test[,c(7,8)],train$General\_Style, prob = TRUE, k = 32)  
table(test$General\_Style,classifications)

## classifications  
## Ale IPA  
## Ale 137 20  
## IPA 21 105

confusionMatrix(table(test$General\_Style,classifications))

## Confusion Matrix and Statistics  
##   
## classifications  
## Ale IPA  
## Ale 137 20  
## IPA 21 105  
##   
## Accuracy : 0.8551   
## 95% CI : (0.8086, 0.894)  
## No Information Rate : 0.5583   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.7065   
##   
## Mcnemar's Test P-Value : 1   
##   
## Sensitivity : 0.8671   
## Specificity : 0.8400   
## Pos Pred Value : 0.8726   
## Neg Pred Value : 0.8333   
## Prevalence : 0.5583   
## Detection Rate : 0.4841   
## Detection Prevalence : 0.5548   
## Balanced Accuracy : 0.8535   
##   
## 'Positive' Class : Ale   
##

# This is the code Dylan used to answer qustion 9

# 9. 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

# is one state more likly to purchase local? Where is there a gap in that states beer?

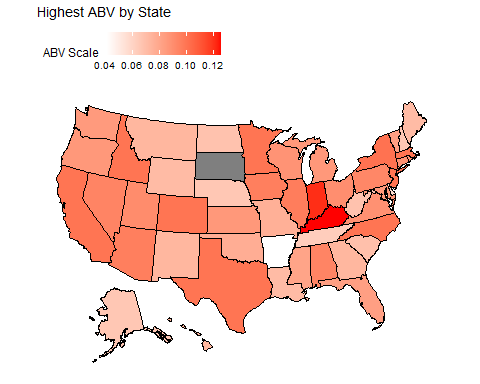
# This is using data from another study which can be found on the github and is cited in the slides

Overall Maine had the best chance for a successful booze stout by having a high ABV average but a middle of the road ABV max

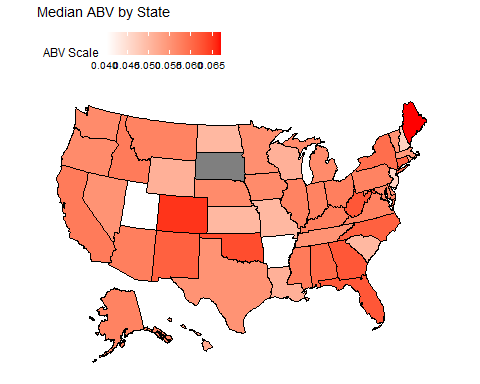
# we want to plot the max and median ABV by state and eliminate a state if their regulations are too high  
# next we will find a gap in the market by showing number of stouts/porters by states and give reccomondations  
  
#filter the data to stout and get a summary to see what the max abv is  
stout\_porter <- bud\_data\_clean %>% filter(grepl("Stout",Style) | grepl("Porter", Style))  
summary(stout\_porter$ABV)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.03700 0.05500 0.06000 0.06572 0.07100 0.12000

#getting the max and median abv by state by using aggregate  
state\_abv = aggregate(ABV~State,data = bud\_data\_clean,max)  
avg\_abv = aggregate(ABV~State,data = bud\_data\_clean,median)  
  
#trimming off those empty spaces of state from the data \*cleaning\*  
state\_abv$State = str\_trim(state\_abv$State)  
avg\_abv$State = str\_trim(avg\_abv$State)  
  
#getting the fips for each state for my map  
state\_abv$fips = fips(state\_abv$State)  
avg\_abv$fips = fips(avg\_abv$State)  
  
#renaming the 'n' column to count   
colnames(state\_abv)[2] = 'abv'  
colnames(avg\_abv)[2] = 'abv'  
  
# creating the heat map  
plot\_usmap(data=state\_abv, values="abv", color = "black") +   
 scale\_fill\_gradient(name = "ABV Scale", low = "white", high = "red") +   
 labs(title="Highest ABV by State") + theme(legend.position = "top")



plot\_usmap(data=avg\_abv, values="abv", color = "black") +   
 scale\_fill\_gradient(name = "ABV Scale", low = "white", high = "red") +   
 labs(title="Median ABV by State") + theme(legend.position = "top")



#counting stouts and porters by state  
stout\_count = count(stout\_porter,State)  
#changing the column name from n to count  
colnames(stout\_count)[2] = 'Count'  
#creating a bar graph by state to show number of stouts  
stout\_count %>%  
 ggplot(aes(State,Count,fill = Count)) +  
 geom\_bar(stat = 'identity',width = .75,color = 'black')+  
 scale\_fill\_gradient(low = 'white', high = 'black') +  
 labs(title="Number of Stouts or Porters by State", x="State", y="Count of Stouts or Porters")+  
 theme(axis.text.x = element\_text(angle = 90, vjust = .3))

