DS 3606 Case Study

Brandon Sucrese

6/25/2021

Introduction: Welcome to the brewery and beer analysis. The following code will walk through important discovers about beer. We will be looking at quantity of breweries and the distribution of beers’ alcohol content. We will also explore the relationship between alcohol content and bitterness as well as how the style of beer effects these variables. Lastly we will look into which states have the highest different in alcohol among their beers. Understanding how beer works and it’s finer details, always us to better understand beers in the future.

Code Summary: Reads in necessary libraries.

library(naniar)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.1.1 ✓ dplyr 1.0.6  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1

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

library(mice)

##   
## Attaching package: 'mice'

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

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(ggplot2)  
library(htmlwidgets)  
library(caTools)  
library(class)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

library(stats)  
library(dplyr)  
library(agricolae)

Code Summary: Reads in the beer and breweries data sets. Sets the variables to the correct type and summaries the two data sets.

beer = read\_csv("https://raw.githubusercontent.com/BivinSadler/MSDS\_6306\_Doing-Data-Science/Master/Unit%208%20and%209%20Case%20Study%201/Beers.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## Name = col\_character(),  
## Beer\_ID = col\_double(),  
## ABV = col\_double(),  
## IBU = col\_double(),  
## Brewery\_id = col\_double(),  
## Style = col\_character(),  
## Ounces = col\_double()  
## )

breweries = read.csv("https://raw.githubusercontent.com/BivinSadler/MSDS\_6306\_Doing-Data-Science/Master/Unit%208%20and%209%20Case%20Study%201/Breweries.csv")  
breweries$State = as.factor(breweries$State) ## Sets the state variable to a factor  
summary(beer) ## Summarizes the beer data set

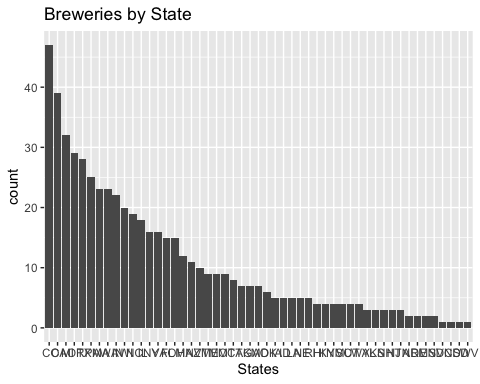
## Name Beer\_ID ABV IBU   
## Length:2410 Min. : 1.0 Min. :0.00100 Min. : 4.00   
## Class :character 1st Qu.: 808.2 1st Qu.:0.05000 1st Qu.: 21.00   
## Mode :character Median :1453.5 Median :0.05600 Median : 35.00   
## Mean :1431.1 Mean :0.05977 Mean : 42.71   
## 3rd Qu.:2075.8 3rd Qu.:0.06700 3rd Qu.: 64.00   
## Max. :2692.0 Max. :0.12800 Max. :138.00   
## NA's :62 NA's :1005   
## Brewery\_id Style Ounces   
## Min. : 1.0 Length:2410 Min. : 8.40   
## 1st Qu.: 94.0 Class :character 1st Qu.:12.00   
## Median :206.0 Mode :character Median :12.00   
## Mean :232.7 Mean :13.59   
## 3rd Qu.:367.0 3rd Qu.:16.00   
## Max. :558.0 Max. :32.00   
##

summary(breweries) ## Summarizes the breweries data set

## Brew\_ID Name City State   
## Min. : 1.0 Length:558 Length:558 CO : 47   
## 1st Qu.:140.2 Class :character Class :character CA : 39   
## Median :279.5 Mode :character Mode :character MI : 32   
## Mean :279.5 OR : 29   
## 3rd Qu.:418.8 TX : 28   
## Max. :558.0 PA : 25   
## (Other):358

Code Summary: Plots the amount of breweries in each State.

breweries %>% ggplot(aes(x=reorder(State, State, function(x) -length(x)))) + geom\_bar() + ggtitle("Breweries by State") + xlab("States")

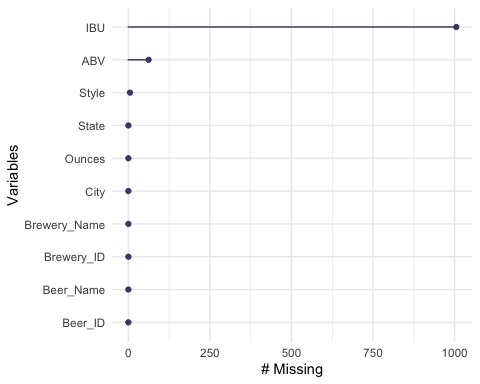


Code Summary: Merges the beer and breweries data set together and relabels the data set.

beer\_comp = merge(beer, breweries, by.x = "Brewery\_id", by.y = "Brew\_ID") ## Merges the beer and breweries data sets  
colnames(beer\_comp)[which(names(beer\_comp) == "Name.x")] <- "Beer\_Name" ## Renames column to Beer\_Name  
colnames(beer\_comp)[which(names(beer\_comp) == "Name.y")] <- "Brewery\_Name" ## Renames column to Brewery\_Name  
colnames(beer\_comp)[which(names(beer\_comp) == "Brewery\_id")] <- "Brewery\_ID" ## Renames column to Brewery\_ID

Code Summary: Graphs missing data and shows first and last observations in the full beer data set.

imp\_beer\_comp = beer\_comp ## Changes beer\_comp to imp\_beer\_comp  
gg\_miss\_var(imp\_beer\_comp) ## Graphs missing values by variable



head(imp\_beer\_comp, n=10) ## Displays first 10 observations

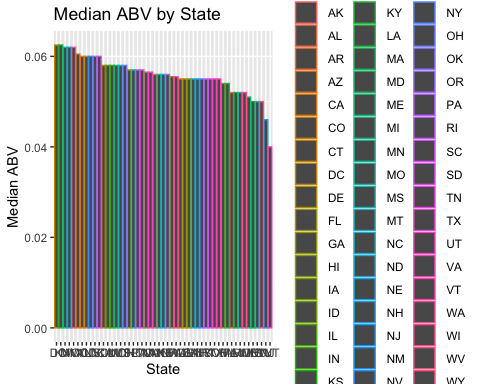
## Brewery\_ID Beer\_Name Beer\_ID ABV IBU  
## 1 1 Get Together 2692 0.045 50  
## 2 1 Maggie's Leap 2691 0.049 26  
## 3 1 Wall's End 2690 0.048 19  
## 4 1 Pumpion 2689 0.060 38  
## 5 1 Stronghold 2688 0.060 25  
## 6 1 Parapet ESB 2687 0.056 47  
## 7 2 Citra Ass Down 2686 0.080 68  
## 8 2 London Balling 2685 0.125 80  
## 9 2 35 K 2684 0.077 25  
## 10 2 A Beer 2683 0.042 42  
## Style Ounces Brewery\_Name  
## 1 American IPA 16 NorthGate Brewing   
## 2 Milk / Sweet Stout 16 NorthGate Brewing   
## 3 English Brown Ale 16 NorthGate Brewing   
## 4 Pumpkin Ale 16 NorthGate Brewing   
## 5 American Porter 16 NorthGate Brewing   
## 6 Extra Special / Strong Bitter (ESB) 16 NorthGate Brewing   
## 7 American Double / Imperial IPA 16 Against the Grain Brewery  
## 8 English Barleywine 16 Against the Grain Brewery  
## 9 Milk / Sweet Stout 16 Against the Grain Brewery  
## 10 American Pale Ale (APA) 16 Against the Grain Brewery  
## City State  
## 1 Minneapolis MN  
## 2 Minneapolis MN  
## 3 Minneapolis MN  
## 4 Minneapolis MN  
## 5 Minneapolis MN  
## 6 Minneapolis MN  
## 7 Louisville KY  
## 8 Louisville KY  
## 9 Louisville KY  
## 10 Louisville KY

tail(imp\_beer\_comp, n=10) ## Displays last 10 observations

## Brewery\_ID Beer\_Name Beer\_ID ABV IBU  
## 2401 552 Ice Pick Ale 160 0.068 NA  
## 2402 553 Mickey Finn's Amber Ale 174 0.056 NA  
## 2403 554 Heiner Brau Kölsch 129 0.050 NA  
## 2404 555 BrewFarm Select Golden Lager 110 0.055 NA  
## 2405 556 Pilsner Ukiah 98 0.055 NA  
## 2406 557 Heinnieweisse Weissebier 52 0.049 NA  
## 2407 557 Snapperhead IPA 51 0.068 NA  
## 2408 557 Moo Thunder Stout 50 0.049 NA  
## 2409 557 Porkslap Pale Ale 49 0.043 NA  
## 2410 558 Urban Wilderness Pale Ale 30 0.049 NA  
## Style Ounces Brewery\_Name  
## 2401 American IPA 12 Silverton Brewery  
## 2402 American Amber / Red Ale 12 Mickey Finn's Brewery  
## 2403 Kölsch 12 Covington Brewhouse  
## 2404 American Pale Lager 12 Dave's Brewfarm  
## 2405 German Pilsener 12 Ukiah Brewing Company  
## 2406 Hefeweizen 12 Butternuts Beer and Ale  
## 2407 American IPA 12 Butternuts Beer and Ale  
## 2408 Milk / Sweet Stout 12 Butternuts Beer and Ale  
## 2409 American Pale Ale (APA) 12 Butternuts Beer and Ale  
## 2410 English Pale Ale 12 Sleeping Lady Brewing Company  
## City State  
## 2401 Silverton CO  
## 2402 Libertyville IL  
## 2403 Covington LA  
## 2404 Wilson WI  
## 2405 Ukiah CA  
## 2406 Garrattsville NY  
## 2407 Garrattsville NY  
## 2408 Garrattsville NY  
## 2409 Garrattsville NY  
## 2410 Anchorage AK

Code Summary: Creates a new data set of Median ABV by state. Graphs and summarize this data set.

State\_ABV = imp\_beer\_comp %>% filter(!is.na(imp\_beer\_comp$ABV)) ## Filters out any rows with missing values for ABV  
  
State\_ABV = aggregate(x = State\_ABV$ABV, ## Creates data set of all the states and their corresponding median ABV values  
 by = list(State\_ABV$State),  
 FUN = median)  
  
colnames(State\_ABV)[which(names(State\_ABV) == "x")] <- "Median\_ABV" ## Renames column to Median\_ABV  
colnames(State\_ABV)[which(names(State\_ABV) == "Group.1")] <- "State" ## Renames column to State  
  
State\_ABV %>% ggplot(aes(x=reorder(State, -Median\_ABV), y=Median\_ABV, color=State)) + ## Graphs Median ABV by State in order  
 geom\_col() + ggtitle("Median ABV by State") + xlab("State") + ylab("Median ABV")



a = which.max(State\_ABV$Median\_ABV) ## Finds and displays the state with the greatest median ABV  
State\_ABV[a,]

## State Median\_ABV  
## 8 DC 0.0625

b = which.min(State\_ABV$Median\_ABV) ## Finds and displays the state with the lowest median ABV  
State\_ABV[b,]

## State Median\_ABV  
## 45 UT 0.04

Code Summary: Finds and displays the beer with highest ABV and IBU values. Also finds and displays the beer with the lowest ABV and IBU values.

c = which.max(imp\_beer\_comp$ABV) ## Finds and displays beer with the highest ABV value  
imp\_beer\_comp[c,]

## Brewery\_ID Beer\_Name Beer\_ID  
## 375 52 Lee Hill Series Vol. 5 - Belgian Style Quadrupel Ale 2565  
## ABV IBU Style Ounces Brewery\_Name City State  
## 375 0.128 NA Quadrupel (Quad) 19.2 Upslope Brewing Company Boulder CO

d = which.max(imp\_beer\_comp$IBU) ## Finds and displays beer with the highest IBU value  
imp\_beer\_comp[d,]

## Brewery\_ID Beer\_Name Beer\_ID ABV IBU  
## 1857 375 Bitter Bitch Imperial IPA 980 0.082 138  
## Style Ounces Brewery\_Name City  
## 1857 American Double / Imperial IPA 12 Astoria Brewing Company Astoria  
## State  
## 1857 OR

e = which.min(imp\_beer\_comp$ABV) ## Finds and displays beer with the lowest ABV value  
imp\_beer\_comp[c,]

## Brewery\_ID Beer\_Name Beer\_ID  
## 375 52 Lee Hill Series Vol. 5 - Belgian Style Quadrupel Ale 2565  
## ABV IBU Style Ounces Brewery\_Name City State  
## 375 0.128 NA Quadrupel (Quad) 19.2 Upslope Brewing Company Boulder CO

f = which.min(imp\_beer\_comp$IBU) ## Finds and displays beer with the lowest IBU value  
imp\_beer\_comp[d,]

## Brewery\_ID Beer\_Name Beer\_ID ABV IBU  
## 1857 375 Bitter Bitch Imperial IPA 980 0.082 138  
## Style Ounces Brewery\_Name City  
## 1857 American Double / Imperial IPA 12 Astoria Brewing Company Astoria  
## State  
## 1857 OR

Code Summary: Summarizes the ABV value and plots ABV distribution by state.

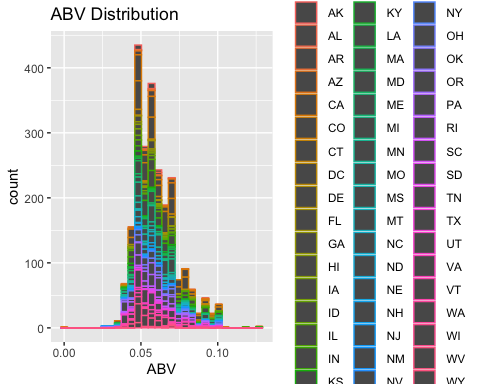
summary(imp\_beer\_comp$ABV) ## Summarizes ABV data

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

imp\_beer\_comp %>% ggplot(aes(x=ABV, color=State)) + geom\_histogram() + ggtitle("ABV Distribution") ## Histogram plot of ABV distribution

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

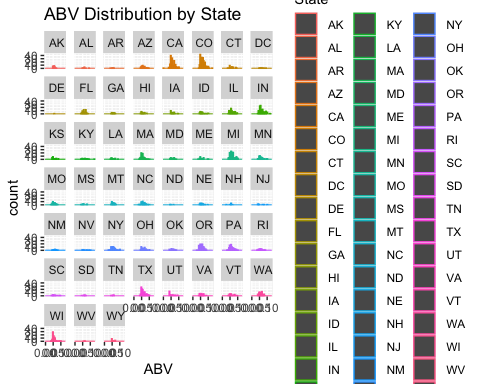
## Warning: Removed 62 rows containing non-finite values (stat\_bin).



imp\_beer\_comp %>% ggplot(aes(x=ABV, color=State)) + geom\_histogram() + ## Histogram plot of ABV distribution by state histogram  
 facet\_wrap(~State) + ggtitle("ABV Distribution by State")

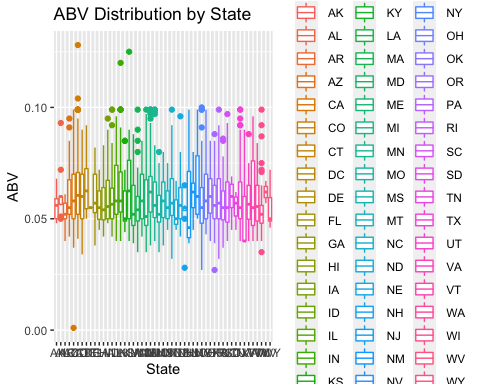
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 62 rows containing non-finite values (stat\_bin).



imp\_beer\_comp %>% ggplot(aes(x=State, y=ABV, color=State)) + geom\_boxplot() + ggtitle("ABV Distribution by State") ## Boxplot of ABV distribution by state

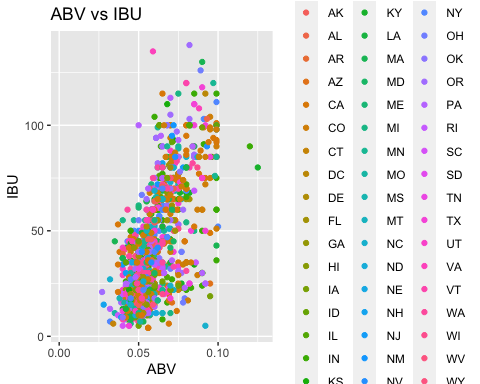
## Warning: Removed 62 rows containing non-finite values (stat\_boxplot).



Code Summary: Plots the relationship between ABV and IBU

imp\_beer\_comp %>% ggplot(aes(x=ABV, y=IBU, color=State)) + geom\_point() + ggtitle("ABV vs IBU") ## Scatter plot comparing ABV and IBU (Points colored by state)

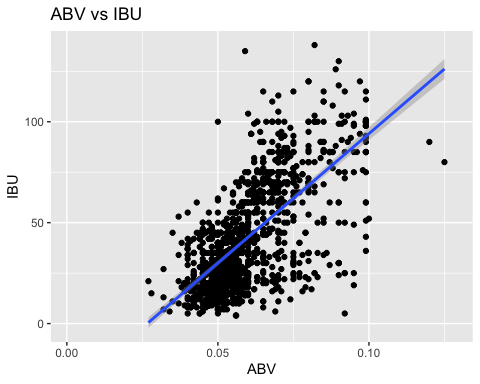
## Warning: Removed 1005 rows containing missing values (geom\_point).



imp\_beer\_comp %>% ggplot(aes(x=ABV, y=IBU)) + geom\_point() + ggtitle("ABV vs IBU") + geom\_smooth(method = "lm") ## Scatter plot comparing ABV and IBU with a linear regression line

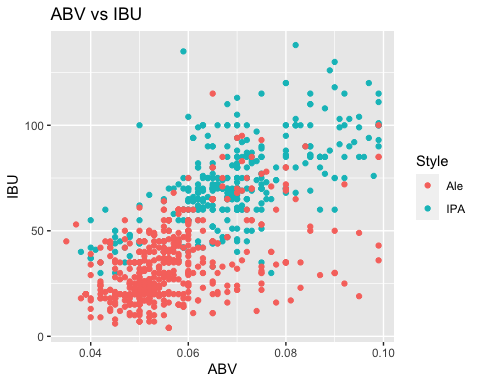
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 1005 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 1005 rows containing missing values (geom\_point).



Code Summary: Creates a data set of just IPA and Ale style of beers and plots this new data.

IPA\_Beer = imp\_beer\_comp %>% filter(str\_detect(imp\_beer\_comp$Style, "IPA")) ## Creates data set of just IPA beers  
IPA\_Beer$Style = "IPA"  
Ale\_Beer = imp\_beer\_comp %>% filter(str\_detect(imp\_beer\_comp$Style, "Ale")) ## Creates data set of just Ale beers  
Ale\_Beer$Style = "Ale"  
IPA\_Ale\_Beer = rbind(IPA\_Beer, Ale\_Beer) ## Combines the IPA and Ale beer data sets together  
IPA\_Ale\_Beer = na.omit(IPA\_Ale\_Beer) ## Removes any missing values from the data set  
  
IPA\_Ale\_Beer %>% ggplot(aes(x=ABV, y=IBU, color=Style)) + geom\_point() + ggtitle("ABV vs IBU") ## Scatter plot of IBU and ABV, style of Beers are colored



Code Summary: Splits the data into a training and test set. Runs a KNN classification model on the IPA and Ale beer data set.

set.seed(1) ## Sets the seed  
sample = sample.split(IPA\_Ale\_Beer[,1], SplitRatio = .70) ## Splits the data using 30/70 ratio  
beer\_train = subset(IPA\_Ale\_Beer, sample == TRUE) ## Creates training set  
beer\_test = subset(IPA\_Ale\_Beer, sample == FALSE) ## Creates test set  
  
v1 = beer\_train[,6] ## Creates classifiers references of IBA and IBU  
pred = knn(beer\_train[,c('ABV', 'IBU')], beer\_test[,c('ABV', 'IBU')], v1, k=2, prob = TRUE) ## Creates data set of predictions using the KNN model for the style of beers (IPA vs Ale)  
actual = beer\_test$Style ## Creates data set of the actual style of the beers  
confusionMatrix(pred,as.factor(actual)) ## Creates a confusion matrix of the predicted and actual style of the beers

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Ale IPA  
## Ale 148 20  
## IPA 26 80  
##   
## Accuracy : 0.8321   
## 95% CI : (0.7825, 0.8744)  
## No Information Rate : 0.635   
## P-Value [Acc > NIR] : 5.515e-13   
##   
## Kappa : 0.6424   
##   
## Mcnemar's Test P-Value : 0.461   
##   
## Sensitivity : 0.8506   
## Specificity : 0.8000   
## Pos Pred Value : 0.8810   
## Neg Pred Value : 0.7547   
## Prevalence : 0.6350   
## Detection Rate : 0.5401   
## Detection Prevalence : 0.6131   
## Balanced Accuracy : 0.8253   
##   
## 'Positive' Class : Ale   
##

Code Summary: Runs an ANOVA and Tukey test to determine which States have significant difference in ABV.

anova\_beer = imp\_beer\_comp %>% filter(!is.na(imp\_beer\_comp$ABV)) ## Filters out all NA values in ABV  
anova\_results = aov(ABV ~ State, data = anova\_beer) ## ANOVA test for ABV between all States  
summary(anova\_results) ## Summarizes the ANOVA results (results are significant)

## Df Sum Sq Mean Sq F value Pr(>F)   
## State 50 0.0218 0.0004359 2.451 9.07e-08 \*\*\*  
## Residuals 2297 0.4086 0.0001779   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

HSD\_results<-HSD.test(anova\_results,"State", group=FALSE) ## Runs Tukey pairwise comparison test  
state\_comparison = HSD\_results$comparison ## Makes data set related to comparisons including pvalues  
sign\_comp = filter(state\_comparison, pvalue <= .05) ## Makes data set of only comparison at are significant (pvalues <= 0.5)  
print(sign\_comp) ## Prints all significant comparisons

## difference pvalue signif. LCL UCL  
## CA - WI 0.007023594 0.0498 \* 2.080796e-06 0.0140451073  
## CO - MA 0.007689073 0.0063 \*\* 8.874533e-04 0.0144906931  
## CO - OR 0.006308000 0.0153 \* 4.532161e-04 0.0121627839  
## CO - UT 0.011487385 0.0264 \* 4.740684e-04 0.0225007008  
## CO - WI 0.009313176 0.0000 \*\*\* 2.602548e-03 0.0160238047  
## IN - MA 0.007711234 0.0306 \* 2.488927e-04 0.0151735748  
## IN - UT 0.011509545 0.0454 \* 7.636388e-05 0.0229427265  
## IN - WI 0.009335337 0.0005 \*\*\* 1.955836e-03 0.0167148379  
## MA - MI -0.007694557 0.0239 \* -1.502622e-02 -0.0003628883  
## MI - UT 0.011492868 0.0416 \* 1.445439e-04 0.0228411922  
## MI - WI 0.009318660 0.0003 \*\*\* 2.071325e-03 0.0165659945