

# Unit 8

[Code ▾](#)

Kyle Kuberski, Pejal Rath

2023-03-03

[Hide](#)

##Introduction:

#Hello Team Budweiser, I am excited to share with you an in-depth view of gathered data relating to both beers and breweries from around the United States. Today we will be walking through a multitude of cleaned, insightful, and easily interpreted data sets. From these datasets we will begin to understand differences in preference of beer, states that include the highest amounts of brewed beers, and overarching similarities/differences between ABV and IBU values. By the end of this in depth review, you find this EDA both thought provoking and impactful on the current industry of beer. This may assist in guiding future business decisions and shape the way you perceive the current platform.

#Here we will load our libraries and display the amount of breweries present in each state. ##  
Breweries Per State

[Hide](#)

```

library(e1071)
library(tm) #text mining library provides the stopwords() function
library(tidyr)
library(plyr)
library(jsonlite)
library(dplyr)
library(tidyverse)
library(mvtnorm)
library(caret)
library(class)
library(ggplot2)
library(plotly)
library(ggthemes)

#1. How many breweries are present in each state?
beer <- read.csv("https://raw.githubusercontent.com/KyleKuberski/MSDS_6306_Doing-Dat
a-Science/Master/Unit%208%20and%209%20Case%20Study%201/Beers.csv")
brewer <- read.csv("https://raw.githubusercontent.com/BivinSadler/MSDS_6306_Doing-Da
ta-Science/Master/Unit%208%20and%209%20Case%20Study%201/Breweries.csv")

beer=read.csv(file.choose(),header = TRUE)
brewer=read.csv(file.choose(),header = TRUE)

brewer %>% ggplot(aes(x=State,fill=State))+geom_histogram(stat="count")+theme(legen
d.position = "none")
brewerByState=brewer %>% count(State)
colnames(brewerByState)[2]="Count"

```

#Now, we will merge the two datasets by "Brew\_ID" to make one large dataset containing all information. ##Merge Datasets

[Hide](#)

```

#2.
newbeer <- beer
colnames(newbeer)[5]<- "Brew_ID"
mergeboth <- merge(newbeer, brewer, by =c("Brew_ID"))
printmerge <- head(mergeboth, 6)
printmerge

```

#Due to NA values, errors can be thrown when trying to interpret the data. Here we clean these missing values. ## Missing Values

[Hide](#)

```
#3. Address the missing values in each column.
#Fill IBU missing value
set.seed(5)
IBU_na=which(is.na(beer$IBU))
IBU_fill=sample(4:138, 1005,replace=TRUE)
for(i in 1:1005)
{
  beer$IBU[IBU_na[i]]=IBU_fill[i]
}
#Fill ABV missing value
set.seed(5)
ABV_na=which(is.na(beer$ABV))
ABV_fill=sample(0.001:0.128, 62,replace=TRUE)
for(i in 1:62)
{
  beer$ABV[ABV_na[i]]=ABV_fill[i]
}
# Change Ounces to factor type
beer$Ounces=factor(beer$Ounces)
```

#Lets find the median values of both ABV and IBU within all beers. ## Median ABV/IBU

[Hide](#)

```
#4. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.
beer_data <- merge(beer,brewer, by.x="Brewery_id",by.y="Brew_ID")
colnames(beer_data)[2]="BeerName"
colnames(beer_data)[8]="BreweryName"
#beer_data$ABV=as.numeric(beer_data$ABV)
summary_ABVIBU<- beer_data %>% group_by(State) %>% summarize(median_ABV = median(ABV),median_IBU = median(IBU))

summary_ABVIBU %>% ggplot(aes(x=State))+geom_bar(aes(y=median_IBU),stat="identity",fill="Red")+labs(title = "Median IBU by State", x = "State",y = "Median IBU") +theme_bw()

summary_ABVIBU %>% ggplot(aes(x=State))+geom_bar(aes(y=median_ABV),stat="identity",fill="blue")+labs(title = "Median Alcohol Content by State", x = "State",y = "Median Alcohol Content") +theme_bw()
```

#We will now find the minimum and maximum values for ABV and IBU with all beers. ## Min/Max ABV & IBU

[Hide](#)

```

#5.
#Max for ABV
beer_data[which.max(beer_data$ABV),]$State
beer_data[which.max(beer_data$ABV),]$BreweryName
beer_data[which.max(beer_data$ABV),]$BeerName
beer_data[which.max(beer_data$ABV),]$ABV
#Min for ABV
beer_data[which.min(beer_data$ABV),]$State
beer_data[which.min(beer_data$ABV),]$BreweryName
beer_data[which.min(beer_data$ABV),]$BeerName
beer_data[which.min(beer_data$ABV),]$ABV
#Max for IBU
beer_data[which.max(beer_data$IBU),]$State
beer_data[which.max(beer_data$IBU),]$BreweryName
beer_data[which.max(beer_data$IBU),]$BeerName
beer_data[which.max(beer_data$IBU),]$IBU
#Min for IBU
beer_data[which.min(beer_data$IBU),]$State
beer_data[which.min(beer_data$IBU),]$BreweryName
beer_data[which.min(beer_data$IBU),]$BeerName
beer_data[which.min(beer_data$IBU),]$IBU

max_ABV_IBU<- beer_data %>% group_by(State) %>% summarize(max_ABV = max(ABV),max_IBU
= max(IBU))

maxsummary_ABVIBU<- beer_data %>% group_by(State) %>% summarize(max_ABV = max(ABV),m
ax_IBU = max(IBU))

maxsummary_ABVIBU %>% ggplot(aes(x=State))+geom_bar(aes(y=max_ABV),stat="identity",f
ill="blue")+labs(title = "ABV Content by State", x = "State",y = "Max Alcohol Conten
t") +theme_bw()

maxsummary_ABVIBU %>% ggplot(aes(x=State))+geom_bar(aes(y=max_IBU),stat="identity",f
ill="red")+labs(title = "IBU Content by State", x = "State",y = "Max Alcohol Conten
t") +theme_bw()

```

#Lets make a visual graph showing the overall Distribution of ABV. ## Distribution

Hide

```
#6 summary ABV and distribution of ABV

beer_data[which.min(beer_data$ABV),]$State

summabv<-summary(beer_data$ABV)

beer_data %>% ggplot(aes(x=ABV)) + geom_histogram(binwidth =0.01, fill = "blue", col
= "black")+
labs(title = "Summary of ABV Across all States", x = "Alcohol by Volume (ABV)", y=
"Count")

ggplot(beer_data, aes(ABV), col=State) +
  geom_boxplot()
```

#We will use GGPlot to view the overall relationships between ABV and IBU values in beer. ## GGPlot & Relationship

[Hide](#)

```
#7
ggplot(beer_data, aes(x = ABV, y = IBU)) +
  geom_point() +
  labs(x = "Alcohol Content", y = "Bitterness")+geom_smooth()

# Yes, there is an apparent relationship between the bitterness of the beer and its
alcoholic content. Base on visualization, there are evidences to show the higher alc
ohol content then the beer also has higher IBU.
```

#Using a KNN Classifier as well as a confusion matrix, we will take a look at different levels of IBU/ABV within Ales and IPAs. ## KNN/CFM

[Hide](#)

#8

#create IPA/Ale dataframe

```
ipa_ale <- beer %>% filter(grepl('IPA|Ale', Style))
ipa_ale$Style <- factor(ifelse(grepl("IPA", ipa_ale$Style), "IPA", "Ale"), levels =
c("Ale", "IPA"))
```

#KNN classifier training and test sets

set.seed(123) # for reproducibility

trainIndex &lt;- createDataPartition(ipa\_ale\$Style, p = .8, list = FALSE)

train &lt;- ipa\_ale[trainIndex, ]

test &lt;- ipa\_ale[-trainIndex, ]

#set 'k' and create model

k &lt;- 3

model &lt;- knn(train[, c("ABV", "IBU")], test = test[, c("ABV", "IBU")], cl = train\$Style, k = k)

#levels will not work unless they are the same level and both factors

# Using grepl to check variables for IPA, and assigning it to IPA as a level (same for Ale)

```
ipa_ale$Style <- factor(ifelse(grepl("IPA", ipa_ale$Style), "IPA", "Ale"), levels =
c("Ale", "IPA"))
```

confusionMatrix(model, reference = test\$Style)

cmf&lt;- confusionMatrix(model, reference = test\$Style)

fourfoldplot(as.table(cfm),color=c("green","red"),main = "Confusion Matrix")

```
ggplot(ipa_ale, aes(x = ABV, y = IBU, color = Style)) +
  geom_point() +
  labs(title = "Relationship between ABV and IBU for IPAs and Ales",
       x = "ABV", y = "IBU")
```

#Here we find interesting characteristics in the overarching dataset and visually graph them for clean interpretation. ##Creative EDA

[Hide](#)

```
#9
data_ca <- beer_data %>% filter(State==" CO")
group_style<- data_ca%>% group_by(Style) %>%
  summarize(count = n())

top_5_styles <- group_style %>%
  top_n(5,count)

ggplot(top_5_styles, aes(x = Style, y = count)) +
  geom_bar(stat = "identity", fill = "orange")+geom_text(aes(label = count), vjust =
-0.5) +
  xlab("Beer Style") +
  ylab("Count") +
  ggtitle("Number of Beers by Style")+ theme_economist()+theme(legend.position = "no
ne",axis.title = element_text(size = 25),plot.title = element_text(size = 30, face =
"bold"))
# Popular size

group_size<- data_ca%>% group_by(Ounces) %>%
  summarize(count = n())

top_5_size <- group_size %>%
  top_n(5, count)

ggplot(top_5_size, aes(x = Ounces, y = count)) +
  geom_bar(stat = "identity", fill = "orange") +geom_text(aes(label = count), vjust
= -0.5)+
  xlab("Beer Size") +
  ylab("Count") +
  ggtitle("Number of Beers by Size")+ theme_economist()+theme(legend.position = "non
e",axis.title = element_text(size = 25),plot.title = element_text(size = 30, face =
"bold"))
```