|  |
| --- |
| Craft Beer Analysis Project |
| Technical Paper: Trends in the Craft Brewing Industry - Standard-Knapp |
| September 23rd, 2021  Syracuse University  IST 687  Noah Laraway |

Table of Contents

* Report Overview (page 1)
* Analysis Questions (page 1)
* Data Acquisition, Data Cleansing, Transforming, and Munging (page 1)
* Descriptive Statistics (page 10)
* Modeling Techniques (page 11)
* Overall Conclusions and Interpretation of Results (page 14)
* Appendix (page 15)

Report Overview

As an avid craft beer enthusiast, I have always been interested in different styles of beer and where they’re produced. I’ve also been interested in the difference in variety and availability throughout the US. This project is an analysis of craft breweries, beer styles and their specifications in the US. The first part of this project was importing raw data from CSV files into Rstudio that consisted of 557 breweries and 2,410 different beers. The data was then cleaned, transformed and further prepared for analysis. The analysis of the data utilized aggregate functions and modeling to produce statistics and visualizations to make conclusions about the data.

Analysis Questions

Question #1: What styles of craft beer are produced most frequently in the US?

Question #2: What cities and states produce the most beers?

Question #3: Which cities and states are producing the most bitter beer (highest IBU)?

Question #4: Which cities and states are producing beers with the highest alcohol content (highest ABV)?

Question #5: Are the ABV and IBU of the beers related and can we create a model that can predict ABV from IBU?

Data Acquisition, Cleansing, Transforming and Munging

The data sets used for this project consisted of 2 CSV files from Kaggle that can be found here: [Craft Beer Dataset](https://www.kaggle.com/nickhould/craft-cans/version/1?select=beers.csv)

* The first data set is for US craft breweries. It contains 559 rows and 4 columns. Each row is a different brewery and the columns contain the brewery name and location details. Data dictionary for brewery data set:



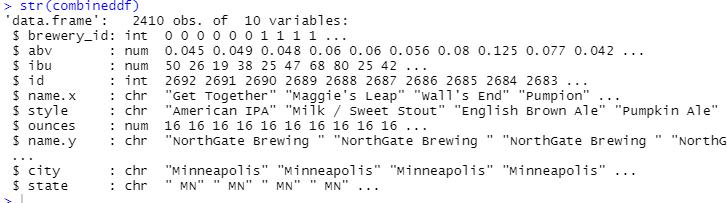
* The second data set is for US craft beers and contains 2,411 rows and 8 columns. Each row is a different beer and the columns contain the beer name, brewery and specifications about the beer.

Data dictionary for beer data set:



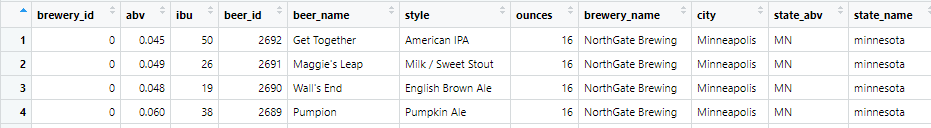
The data was downloaded from Kaggle and brought into Rstudio using read.csv. The two data sets were merged on the brewery id which was in both data frames.

The resulting data frame was then inspected:



From the inspection it was found the data had good quality and all the data types were in correct format, but it was discovered that some cleansing of the data frame needed to be performed. The name.X column was removed since it was redundant. The column names were changed for easier reading. The state\_abv column was altered to remove any white spaces. Finally, a state\_name column was added that would be used later for map visualizations.

Cleansed data frame head:

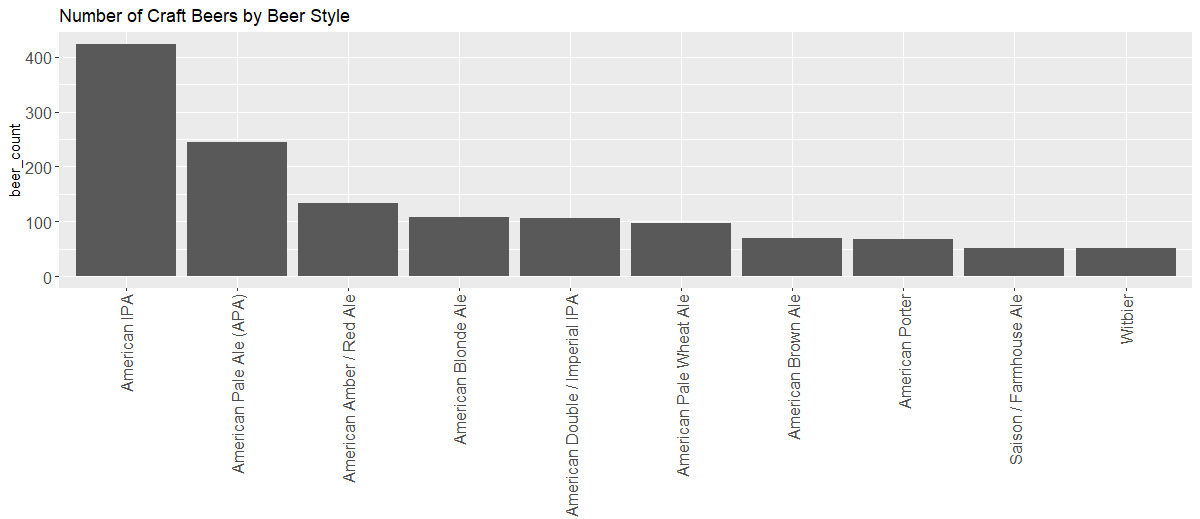


At that point the data was ready to be analyzed to answer the business questions. The first question was what styles of craft beer are produced most frequently in the US? To do this, the data frame was sorted by beer styles and by count of beers and then narrowed down to the top ten by count. A bar plot was also created for the top then beer style.

Data frame for top ten craft beer styles:



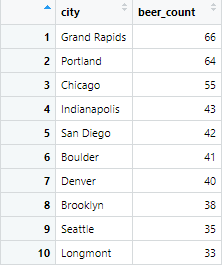
Bar plot of top ten styles of craft beer in the US:



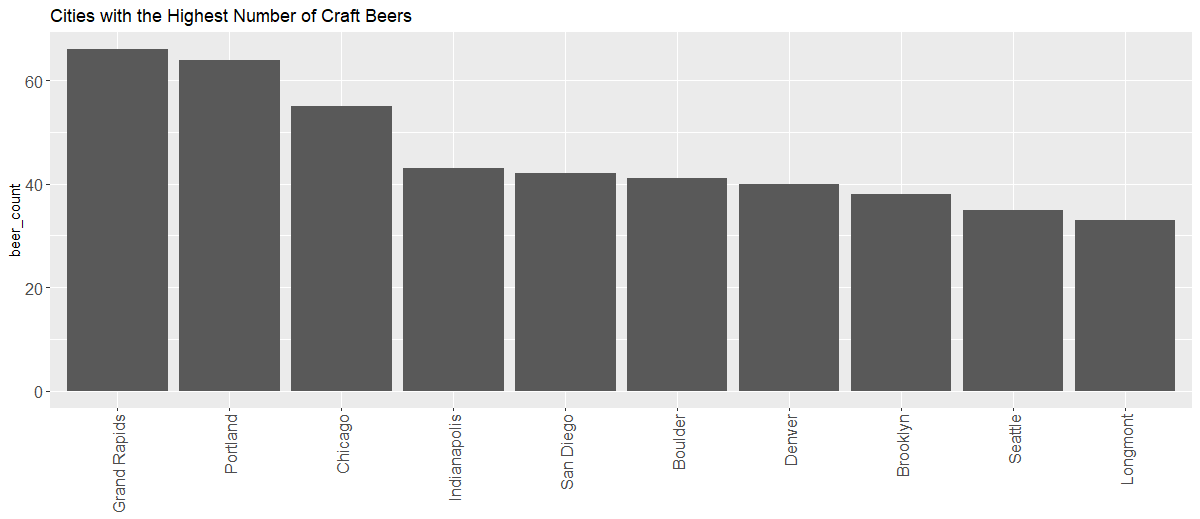
The data frame and bar plot show that the most frequent beer style produced is American IPA.

The second question was what cities and states produce the most beers? I first looked at the top cities for beer. To answer this question the data frame was sorted by city and by count of the number of beers. I then narrowed it down to the top ten. A bar plot was also created for the top ten beer producing cities.

Data frame for the top ten cities by craft beer count:



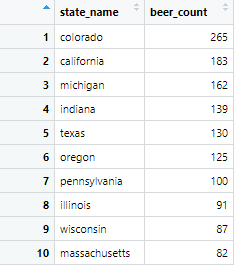
Bar plot of top ten cities for craft beer in the US:



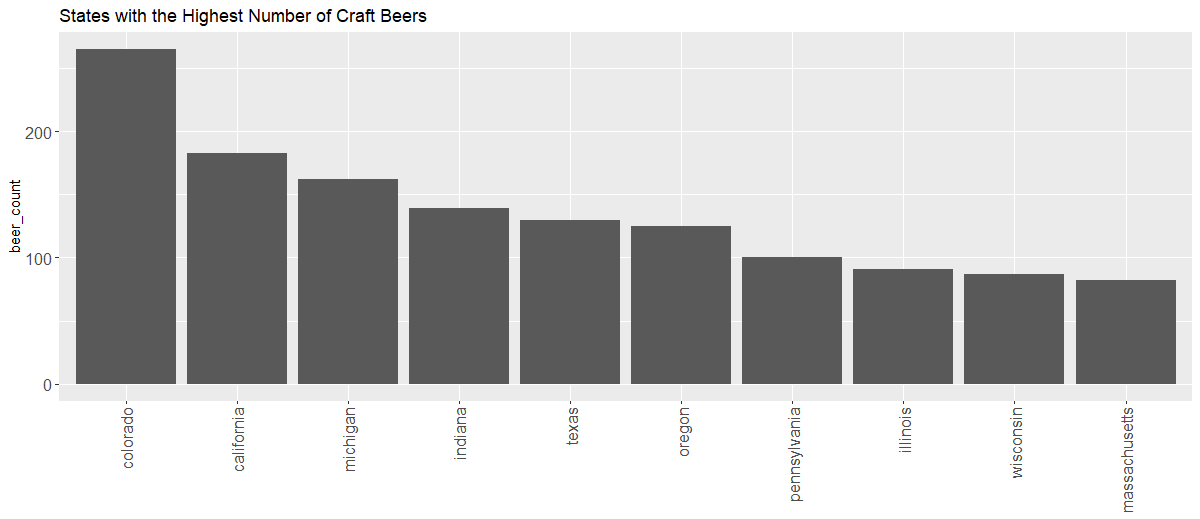
The data frame and bar plot show that the city with the highest number of craft beers produced is Grand Rapids.

I also looked at the top states for craft beer production. This was performed in a similar way to the cities except sorting by state and narrowing it down to the top ten. For visualizations a bar plot of the top ten beer states was created and also a map of the US which highlights each state by color to show the number of beers produced in that state.

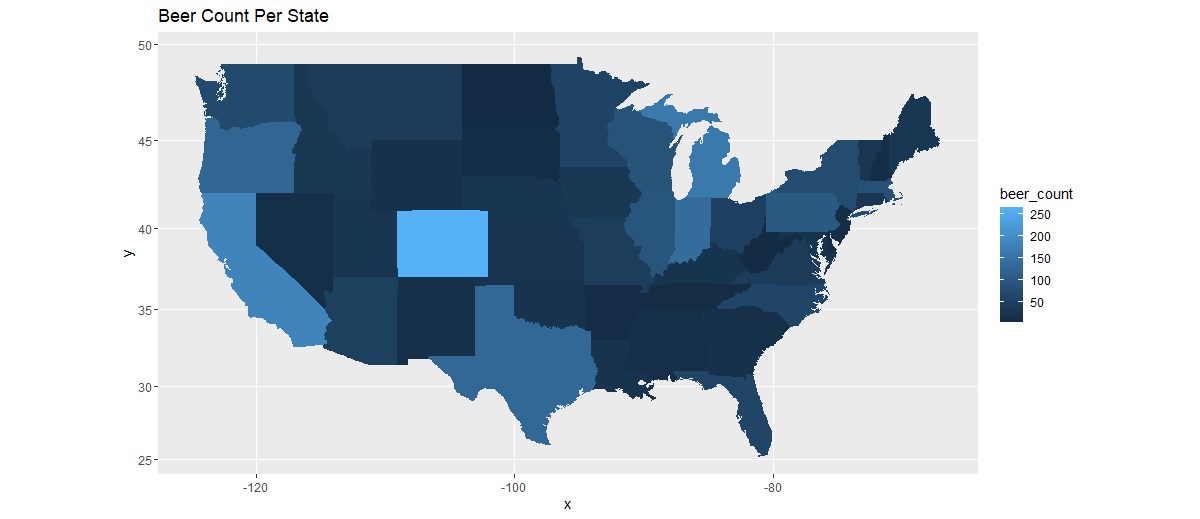
Data frame for the top ten states by craft beer count:



Bar plot of top ten states for craft beer in the US:



Map of US showing craft beer count for each state by color:



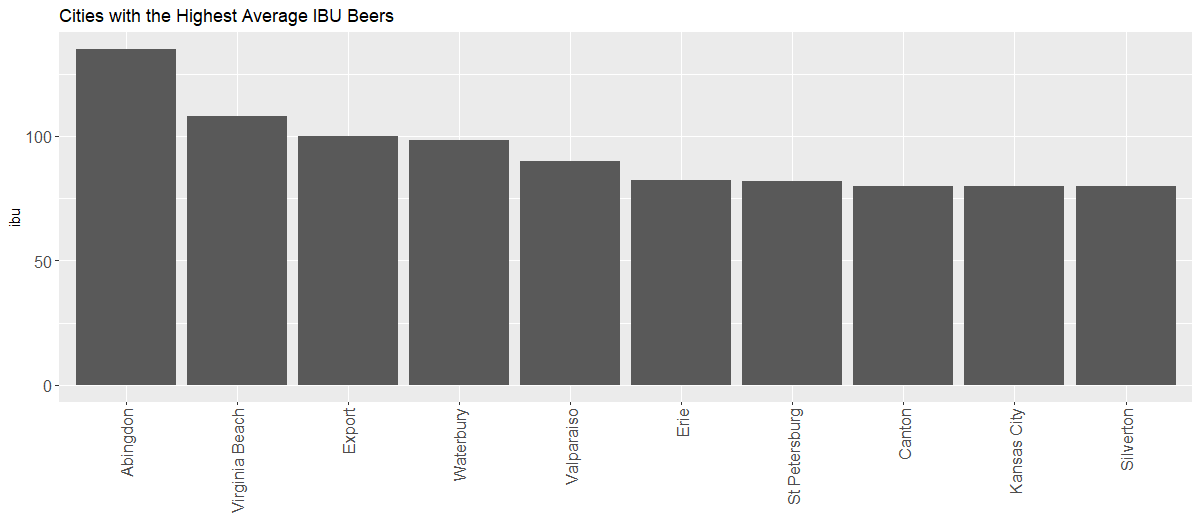
The data frame, bar plot and map show that Colorado is the state with the highest number of craft beers produced.

The next question was which cities and states are producing the most bitter beer (highest IBU)? I had to first remove the rows where the IBU column contained NAs. I first looked at cities and aggregated by the mean of IBU for that city. The data was then narrowed down to the top 10 cities and a bar plot was also created.

Data frame for the top ten cities by average IBU:



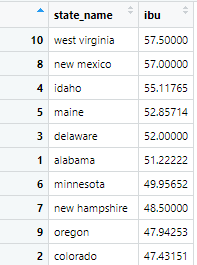
Bar plot of top ten cities by average IBU:



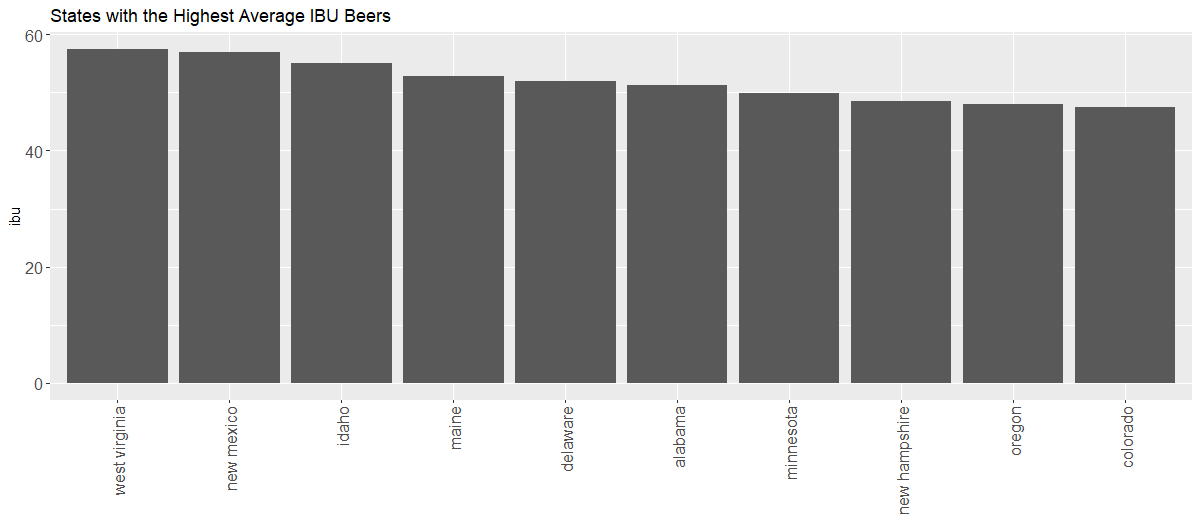
From the data frame and bar plot Abingdon is the city with the highest average IBU.

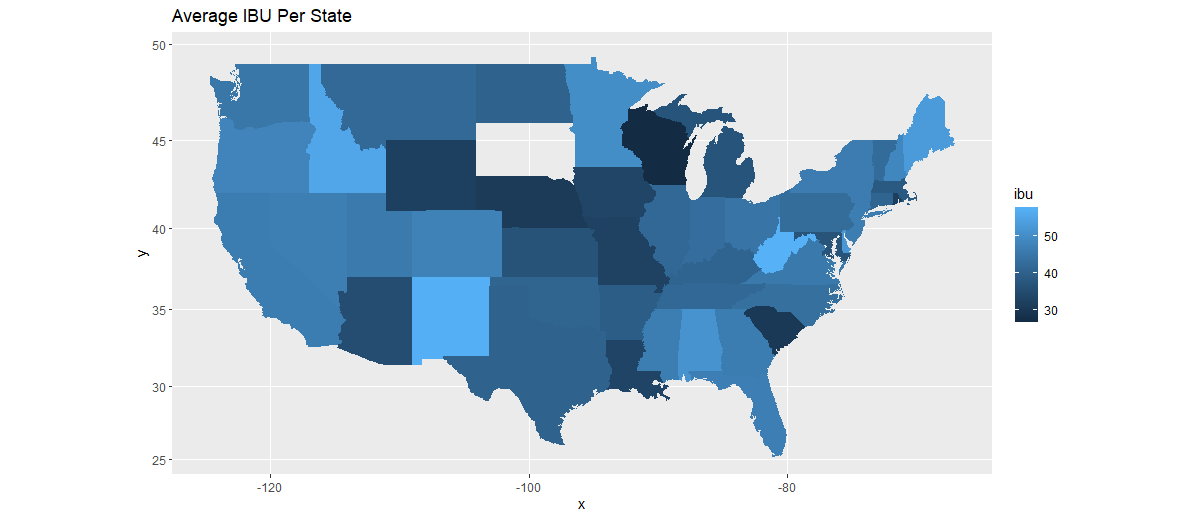
I then looked at the top states by average IBU. This was performed similar to the cities by aggregating by mean IBU and then narrowing down to the top ten. A bar plot and map were also created.

Data frame for the top ten states by average IBU:



Bar plot for the top ten states by average IBU:

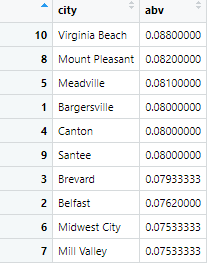
Map of US showing average beer IBU for each state by color:



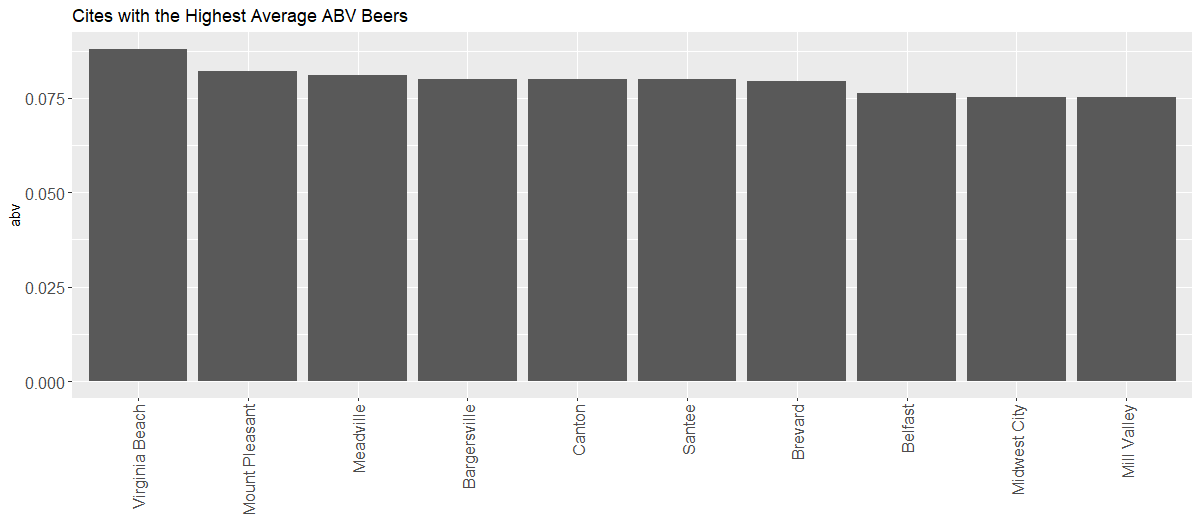
From the data frame, bar plot and map West Virginia is the state with the highest average IBU.

The next question is which cities and states are producing beers with the highest alcohol content (highest ABV)? I had to first remove the rows where the ABV column contained NAs. I first looked at cities and aggregated by the mean of ABV for that city. The data was then narrowed down to the top 10 cities and a bar plot was also created.

Data frame for the top ten cities by average ABV:



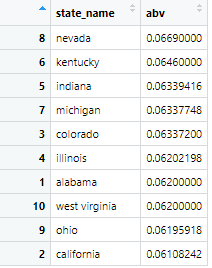
Bar plot of top ten cities by average ABV:



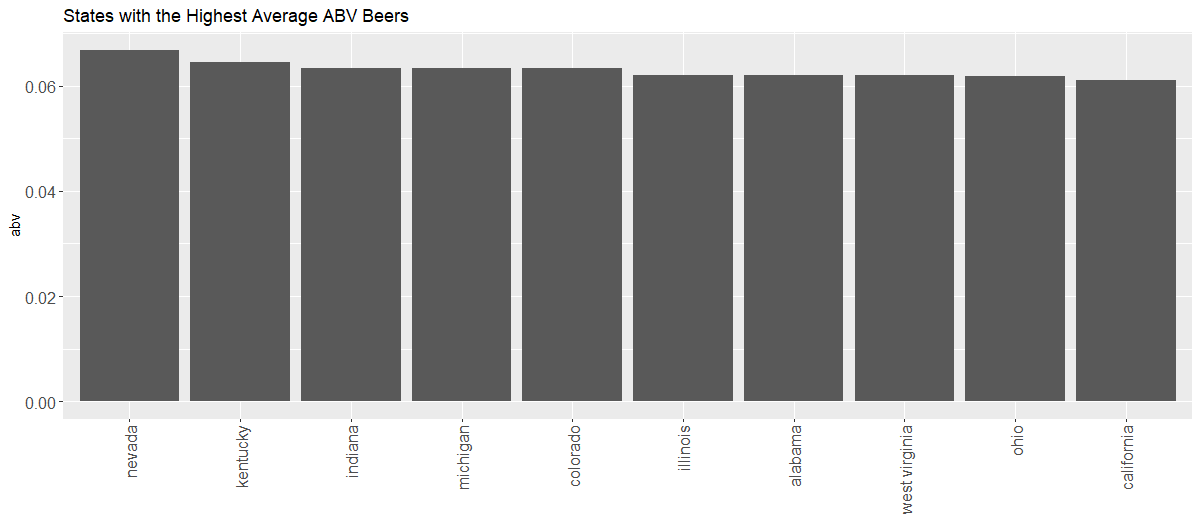
From the data frame and bar plot Virginia Beach is the city with the highest average IBU.

I then looked at the top states by average ABV. This was performed similar to the cities by aggregating by mean ABV and then narrowing down to the top ten. A bar plot and map were also created.

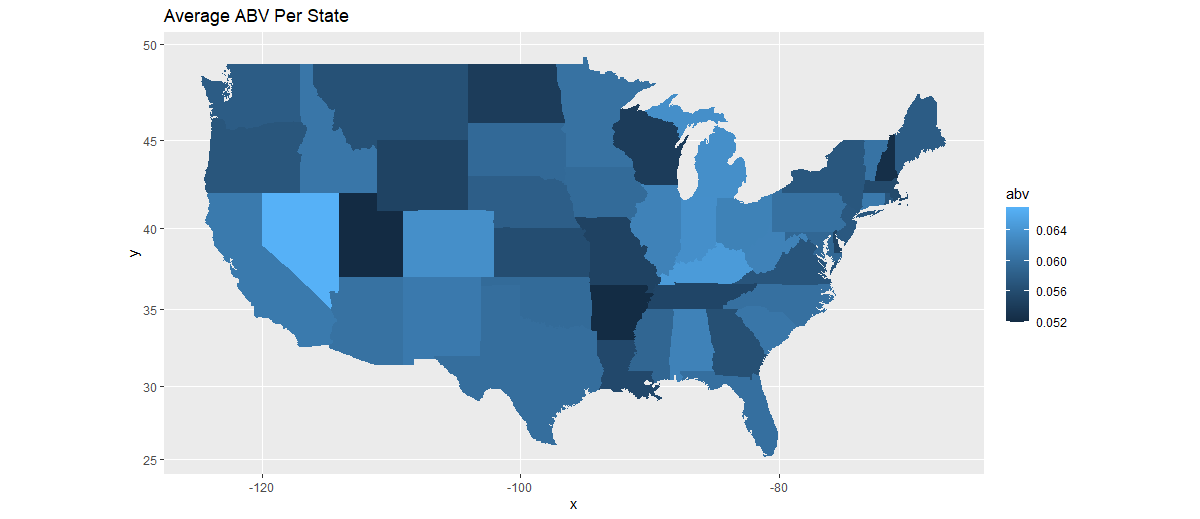
Data frame for the top ten states by average ABV:



Bar plot for the top ten states by average ABV:



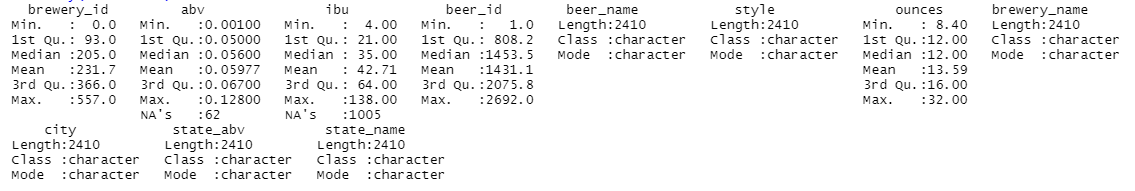
Map of US showing average beer ABV for each state by color:



From the data frame, bar plot and map Nevada is the state with the highest average ABV.

Descriptive Statistics

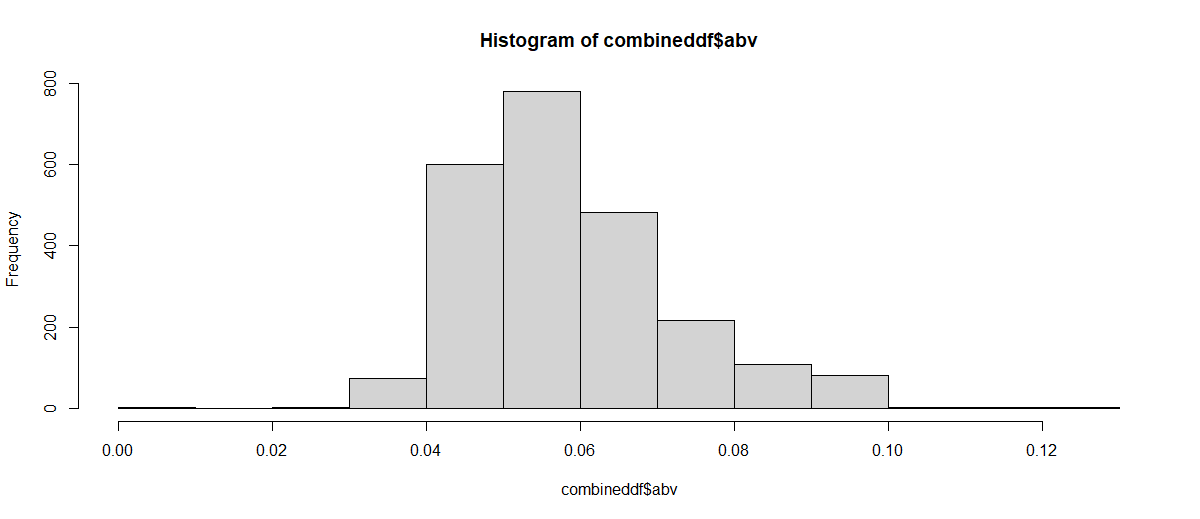
Descriptive statistics on the data were also performed and the summary function in R was used. Summary of data frame:



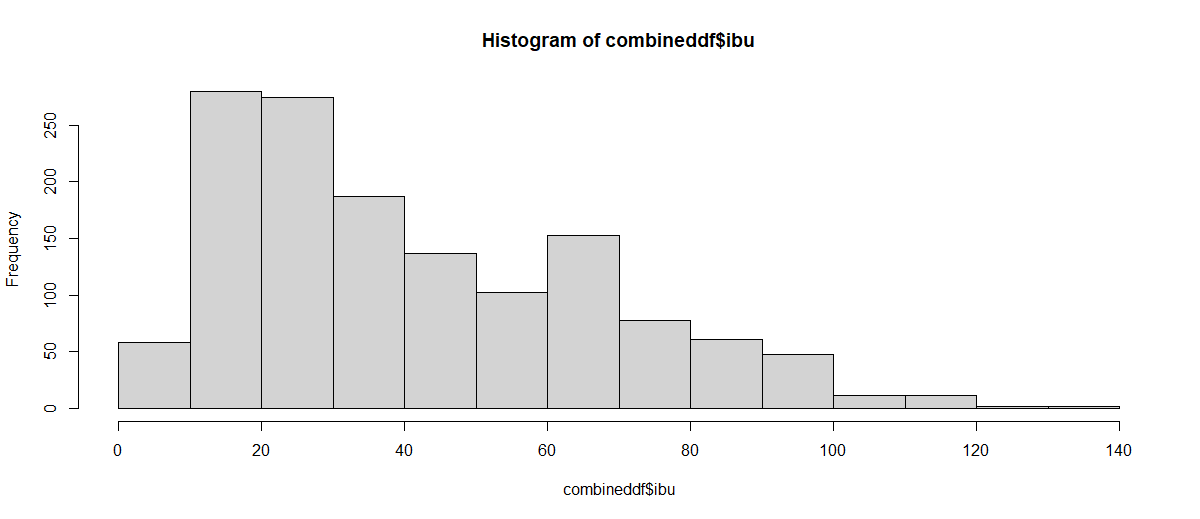
From the statistics summary the median alcohol by volume is 5.6% and the median IBU is 35. These are fairly typical for craft beers and there isn’t any data that stands out from the summary.

Histograms were also created to look at frequencies of ABV and IBU.

Histogram of ABV:



Histogram of IBU:

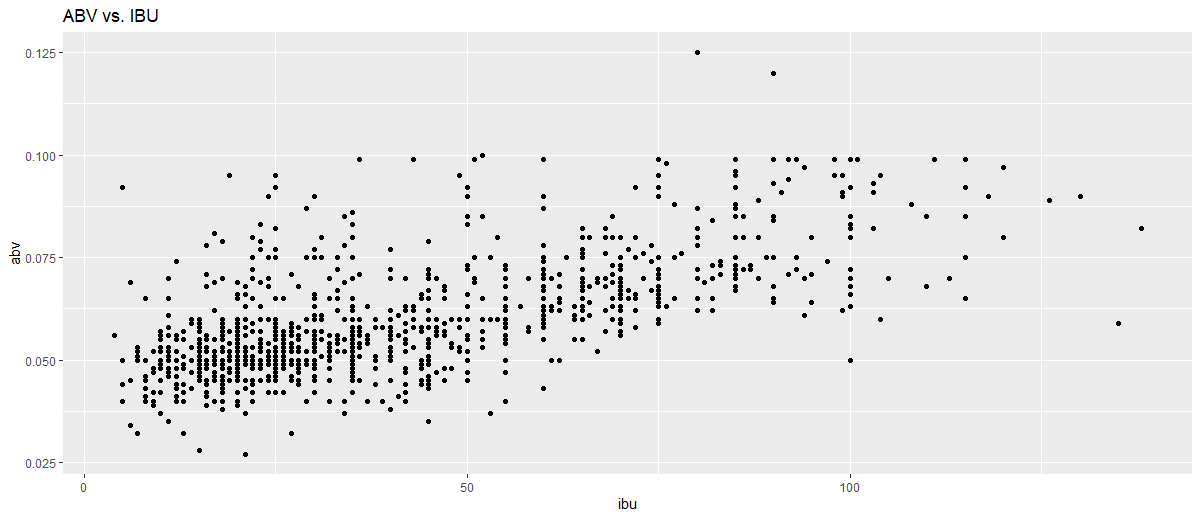


The histogram for ABV shows that the data is mostly centered around the mean of 5.97%. For IBU the highest frequency is between 10 and 30 IBUs, but the histogram is skewed towards the right.

Modeling Techniques

The last question is how are ABV and IBU of the beers related and can we create a model that can predict ABV from IBU? To answer this question linear modeling and support vector modeling were used. A new data frame was created that only contained the ABV and IBU columns of each beer. Then all the NAs were omitted from the data frame so that each row had an ABV and IBU which left 1,405 rows (beers). A scatter plot was created showing ABV vs. IBU.

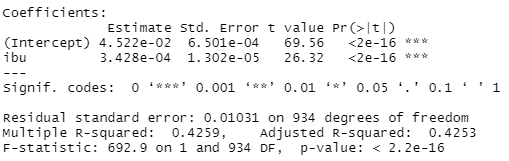
Scatter plot of ABV vs. IBU



This plot shows a trend with ABV increasing with increasing IBU.

The next step was taking random samples of the data to use for modeling. The train data set included 2/3 of the rows (936) and the test data set contained 1/3 (469). Next the lm model in R was used for linear regression.

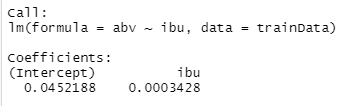
Summary of the linear regression:



The summary shows an adjusted R-squared of 0.4253, an F-statistic of 692.9 with a p-value of 2.2e-16. This R-squared means that 42.53% of the variation in ABV can be explained by variation in IBU. Also, the p-value is very low (less than 0.05) so this equation is significant.

Next, I looked at the y-intercept and coefficient for IBU which is needed for the prediction value equation.

Linear model results from train data set:

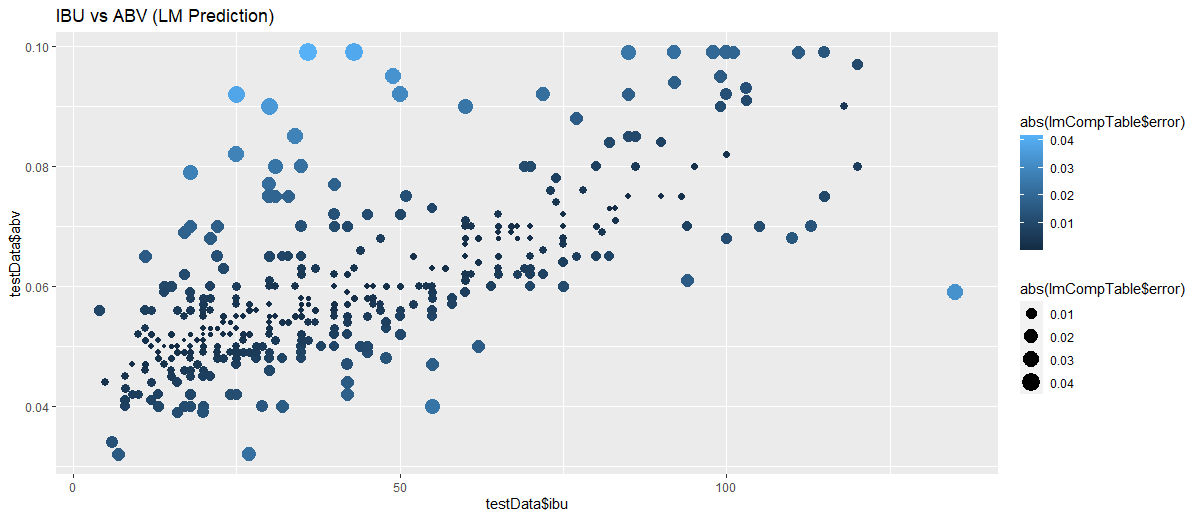


The intercept is 0.0452188 and coefficient for IBU is 0.0003428. These values were put into an equation and then used to predict the ABV for the test data. The predicted ABV values were put into a table with the actual test ABV values to calculate the root mean squared value.



The RMSE was calculated to be 0.009603749. An error column between predicted and actual values was then added to the table. A scatter plot of IBU vs. ABV was created showing the difference in error by color and size.

Scatter plot of IBU vs. ABV with color and size representing error:



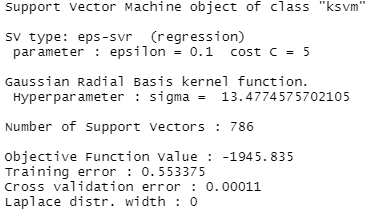
Some of the larger errors are at low IBUs, but there are larger errors at higher IBUs as well.

The KSVM model was also ran on the data in R. The same train data set was used with the following model parameters:



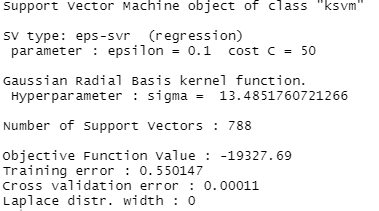
I then looked at the output of the KSVM model.

The resulting output from the model:



The training error is 0.553375 and cross validation error is 0.00011. I then adjusted the cost from 5 to 50 to check to see if that would reduce the training error.

KSVM Model output with c = 50



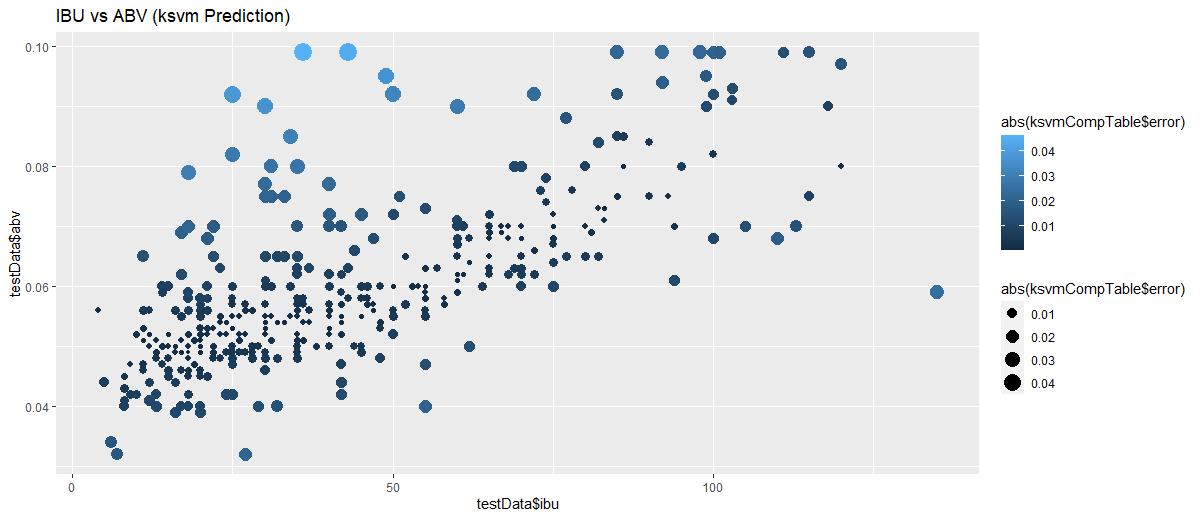
Since this model shows only a slight decrease in the training error. It was decided to use the or first KSVM model with c=5 for modeling. The model was used to run predictions on the test data next. The predicted ABV values were put into a table with the actual test ABV values to calculate the root mean squared value.



The RMSE for the KSVM model was 0.009779093 which is slightly higher than the linear model.

An error column between predicted and actual values was then added to the table similar to the linear model scatter plot.

Scatter plot of IBU vs. ABV with color and size representing error:



This shows similar results to the linear model plot with larger error values at both low and high IBUs.

Overall Conclusion and Interpretation of Results

With the drastic increase in popularity of craft beer in recent years this was an interesting topic to pursue. When the data was initially read into R it became apparent that cleansing and transforming the data early on would be critical to later analysis. While analyzing the data to find information on popular craft beer styles and where they’re produced it showed some results that I expected, but others that were surprising. Also, using modeling techniques was a valuable experience for me in understanding how to make predictions using R.

Overall, the data sets were in good condition, but with over 557 breweries and 2,410 different beers between the two data sets it was somewhat of a challenge to get the data cleaned and in a useful format. An understanding of what the results were going to look like needed to occur before merging, renaming columns and removing unnecessary data. If the data was not cleansed and transformed properly it would create issues later on in the project.

Based on my experience with craft beer there were some initial assumptions that were confirmed during the analysis. The top produced craft beer style was American IPA with 424 different beers. This beer style is very common in breweries and it was not surprising to see this at the top of the list. The top two cities for beers produced were Grand Rapids (66) and Portland (64) and these cities are well known for their craft beer production. Likewise, the top state for beer production was Colorado which again is a top location for craft beer.

Other results from the analysis were unexpected. When looking at highest IBU beers the top two cities were Abingdon and Virginia Beach both in the state of Virginia and the top two states were West Virginia and New Mexico. The west coast of the US is well known for producing “West Coast IPAs” which are typically very bitter with high IBUs. It was surprising to not see states like California and Oregon at the top of the IBU lists. For highest ABV the top cities and states were spread out throughout the country. Again, I expected the top ABV locations to be on the west coast since they’re known for IPAs with higher ABV content.

For the modeling portion of the project, it was interesting to see the connection between ABV and IBU of the beers. The initial scatter plot showed a trend with ABV increasing with increasing IBU. The linear model showed a very low p-value meaning the equation calculated for ABV and IBU was significant. For the KSVM model I was only able to get the training error slightly reduced by increasing the cost from 5 to 50. The resulting RSME from the KSVM model was slightly higher than for the linear model. Based on the RSME, the linear model performed the best for this analysis.

Overall, this project was an interesting and valuable experience in working with data, performing an analysis and making interpretations using R. The data sets from Kaggle were in good shape, but had to be cleansed and transformed which required extra steps before the analysis could be started. During the analysis, additional data frames were created along with visualizations to answer the questions. The modeling part of the project was beneficial in understanding how to use different techniques to make predictions in R. Going forward, it would be interesting to continue to add data to the analysis and see how beers styles, specifications and their production locations change over time.

Appendix (Code)

#Read in CSV files and assign to data frames

beersdf <- read.csv("C:/Users/Noah/Desktop/Syracuse/IST 687/Project/beers.csv")

breweriesdf <- read.csv("C:/Users/Noah/Desktop/Syracuse/IST 687/Project/breweries.csv")

breweriesdf

#Merge beersdf and breweriesdf on breweryid and X

combineddf <- merge(beersdf, breweriesdf, by.x = "brewery\_id", by.y = "X" )

str(combineddf)

#Begin processing data

#Remove column for X(not needed)

combineddf$X <- NULL

combineddf

#Change column names for id=beer\_id, name.x=beer\_name, name.y = brewery\_name and state=state\_abb

colnames(combineddf)[which(names(combineddf) == "id")] <- "beer\_id"

colnames(combineddf)[which(names(combineddf) == "name.x")] <- "beer\_name"

colnames(combineddf)[which(names(combineddf) == "name.y")] <- "brewery\_name"

colnames(combineddf)[which(names(combineddf) == "state")] <- "state\_abv"

#Remove blank spaces from state\_abv column

combineddf$state\_abv <- gsub(" ","", combineddf$state\_abv)

#Add column for full state name

combineddf$state\_name <- tolower(state.name[match(combineddf$state, state.abb)])

#Check data types are correct

str(combineddf)

#Confirmed that data types are correct

#Check Summary of data frame

summary(combineddf)

#Create histograms for IBU and ABV

hist(combineddf$abv)

hist(combineddf$ibu)

#Begin analyzing data and answering questions

#What styles of beer are produced the most?

library(dplyr)

#Find the top ten beer styles

beerstyles <- combineddf %>% count(style, sort=TRUE) %>% top\_n(10)

colnames(beerstyles)[which(names(beerstyles) == "n")] <- "beer\_count"

beerstyles

#Create bar plot

library(ggplot2)

gstyle <-ggplot(beerstyles, aes(x = reorder(style, -beer\_count), beer\_count)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gstyle <- gstyle + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("Number of Craft Beers by Beer Style")

gstyle

#What cities and states have the most beers?

#Find the top ten cities

beercities <- combineddf %>% count(city, sort=TRUE) %>% top\_n(10)

colnames(beercities)[which(names(beercities) == "n")] <- "beer\_count"

beercities

#Create bar plot

gcities <-ggplot(beercities, aes(x = reorder(city, -beer\_count), beer\_count)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gcities <- gcities + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("Cities with the Highest Number of Craft Beers")

gcities

#Order states by beer count for all states

beerstatesall <- combineddf %>% count(state\_name, sort=TRUE)

colnames(beerstatesall)[which(names(beerstatesall) == "n")] <- "beer\_count"

beerstatesall

#Find the top ten states by number of beers

beerstates <- combineddf %>% count(state\_name, sort=TRUE) %>% top\_n(10)

colnames(beerstates)[which(names(beerstates) == "n")] <- "beer\_count"

beerstates

#Create bar plot

gstates <-ggplot(beerstates, aes(x = reorder(state\_name, -beer\_count), beer\_count)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gstates <- gstates + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("States with the Highest Number of Craft Beers")

gstates

#Create map for each state with color representing beer count

library(ggmap)

us <- map\_data("state")

map.beer <- ggplot(beerstatesall, aes(map\_id = state\_name))

map.beer <- map.beer + geom\_map(map = us, aes(fill=beer\_count))

map.beer <- map.beer + expand\_limits(x = us$long, y = us$lat)

map.beer <- map.beer + coord\_map() + ggtitle("Beer Count Per State")

map.beer

#Which states and cities are producing beers with the highest IBU(bitterness)?

#Remove blanks for IBU column

ibudf <- combineddf[!is.na(combineddf$ibu),]

ibudf

#Find the top ten cities and aggregate by mean ibu

ibucities <- aggregate(x = ibudf$ibu,

by = list(ibudf$city),

FUN = mean) %>% top\_n(10)

colnames(ibucities)[which(names(ibucities) == "Group.1")] <- "city"

colnames(ibucities)[which(names(ibucities) == "x")] <- "ibu"

ibucities <- ibucities[order(-ibucities$ibu),]

ibucities

#Create bar plot

gibucities <-ggplot(ibucities, aes(x = reorder(city, -ibu), ibu)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gibucities <- gibucities + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("Cities with the Highest Average IBU Beers")

gibucities

#Find the top ten states

ibustates <- aggregate(x = ibudf$ibu,

by = list(ibudf$state\_name),

FUN = mean)

colnames(ibustates)[which(names(ibustates) == "Group.1")] <- "state\_name"

colnames(ibustates)[which(names(ibustates) == "x")] <- "ibu"

ibustates

#Narrow down to top 10 states

ibustatestop10 <- ibustates %>% top\_n(10)

ibustatestop10 <- ibustatestop10[order(-ibustatestop10$ibu),]

ibustatestop10

#Create bar plot

gibustates <-ggplot(ibustatestop10, aes(x = reorder(state\_name, -ibu), ibu)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gibustates <- gibustates + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("States with the Highest Average IBU Beers")

gibustates

#Create map for each state with color representing mean ibu

us <- map\_data("state")

map.ibu <- ggplot(ibustates, aes(map\_id = state\_name))

map.ibu <- map.ibu + geom\_map(map = us, aes(fill=ibu))

map.ibu <- map.ibu + expand\_limits(x = us$long, y = us$lat)

map.ibu <- map.ibu + coord\_map() + ggtitle("Average IBU Per State")

map.ibu

#Which states and cities are producing beers with the highest alcohol content?

#Remove blanks for abv column

abvdf <- combineddf[!is.na(combineddf$abv),]

abvdf

#Find the top ten cities and aggregate cites by mean abv

abvcities <- aggregate(x = abvdf$abv,

by = list(abvdf$city),

FUN = mean) %>% top\_n(10)

colnames(abvcities)[which(names(abvcities) == "Group.1")] <- "city"

colnames(abvcities)[which(names(abvcities) == "x")] <- "abv"

abvcities <- abvcities[order(-abvcities$abv),]

abvcities

#Create bar plot

gabvcities <-ggplot(abvcities, aes(x = reorder(city, -abv), abv)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gabvcities <- gabvcities + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("Cites with the Highest Average ABV Beers")

gabvcities

#Aggregate by mean abv for all states

abvstates <- aggregate(x = abvdf$abv,

by = list(abvdf$state\_name),

FUN = mean)

colnames(abvstates)[which(names(abvstates) == "Group.1")] <- "state\_name"

colnames(abvstates)[which(names(abvstates) == "x")] <- "abv"

abvstates

#Find the top ten states

abvstatestop10 <- abvstates %>% top\_n(10)

abvstatestop10 <- abvstatestop10[order(-abvstatestop10$abv),]

abvstatestop10

#Create bar plot for top 10

gabvstatestop10 <-ggplot(abvstatestop10, aes(x = reorder(state\_name, -abv), abv)) + geom\_bar(stat="identity")+ theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))

gabvstatestop10 <- gabvstatestop10 + theme(axis.text=element\_text(size=12), axis.title.x = element\_blank()) + ggtitle("States with the Highest Average ABV Beers")

gabvstatestop10

#Create map for each state with color representing mean abv

us <- map\_data("state")

map.abv <- ggplot(abvstates, aes(map\_id = state\_name))

map.abv <- map.abv + geom\_map(map = us, aes(fill=abv))

map.abv <- map.abv + expand\_limits(x = us$long, y = us$lat)

map.abv <- map.abv + coord\_map() + ggtitle("Average ABV Per State")

map.abv

#Predict ABV value as a function of IBU value using linear modeling

#Import Libraries

library(kernlab)

library(e1071)

#Create new data frame for modeling with only the IBU and ABV columns from beersdf

modeldf <- data.frame(beersdf$abv, beersdf$ibu)

#Change column names

colnames(modeldf)[which(names(modeldf) == "beersdf.abv")] <- "abv"

colnames(modeldf)[which(names(modeldf) == "beersdf.ibu")] <- "ibu"

#Omit NAs

modeldf <- na.omit(modeldf)

str(modeldf)

#Plot abv vs. ibu

gIbuAbv <-ggplot(modeldf, aes(ibu, abv)) + geom\_point()

gIbuAbv <- gIbuAbv + ggtitle("ABV vs. IBU")

gIbuAbv

#Create train and test data sets

#Create randindex

randIndex <- sample(1:dim(modeldf)[1])

#Create cut point at 2/3rds of data

cutpoint2\_3 <- floor(2 \* dim(modeldf)[1]/3)

#Create trainData set

trainData <- modeldf[randIndex[1:cutpoint2\_3],]

#Create testData set

testData <- modeldf[randIndex[(cutpoint2\_3 + 1):dim(modeldf[1])],]

#Look at structure of testdata and traindata

str(testData)

str(trainData)

#Compute model for lm with trainData set

regressionlm <- lm(formula = abv~ibu, data=trainData)

summary(regressionlm)

regressionlm

#Test the lm model on the testData set

lmPredicted <- 0.0452188 + 0.0003428\*testData$ibu

lmPredicted

#Create comparison table for actual and predicted

lmActual = testData[,1]

lmCompTable <- data.frame(lmActual,lmPredicted)

lmCompTable

#Compute the Root Mean Squared Error for the model

lmRMSE = sqrt(mean((lmCompTable$lmActual - lmCompTable$lmPredicted)^2))

lmRMSE

# Calculate error and add to table

lmCompTable$error <- lmActual - lmPredicted

#Create Plot

lmScat <- ggplot(testData, aes(x=testData$ibu, y=testData$abv, color=abs(lmCompTable$error), size=abs(lmCompTable$error))) + geom\_point()

lmScat <- lmScat + ggtitle("IBU vs ABV (LM Prediction)")

lmScat

#Compute model ksvm with trainData set

ksvmOutput <- ksvm(abv ~ ibu,data=trainData, kernel="rbfdot",kpar="automatic",

C=5,cross=3,prob.model=TRUE)

ksvmOutput

#run new model with c=50

ksvmOutput2 <- ksvm(abv ~ ibu,data=trainData, kernel="rbfdot",kpar="automatic",

C=50,cross=3,prob.model=TRUE)

ksvmOutput2

#Test the model on the testData set

ksvmPred <- predict(ksvmOutput, testData, type="votes")

#Create comparison table for actual and predicted

ksvmActual = testData[,1]

ksvmPredicted = ksvmPred[,1]

ksvmCompTable <- data.frame(ksvmActual,ksvmPredicted)

#Compute the Root Mean Squared Error for the model

ksvmRMSE = sqrt(mean((ksvmCompTable$ksvmActual - ksvmCompTable$ksvmPredicted)^2))

ksvmRMSE

# Calculate error and add to table

ksvmCompTable$error <- ksvmActual - ksvmPredicted

#Create Plot

ksvmScat <- ggplot(testData, aes(x=testData$ibu, y=testData$abv, color=abs(ksvmCompTable$error), size=abs(ksvmCompTable$error))) + geom\_point()

ksvmScat <- ksvmScat + ggtitle("IBU vs ABV (ksvm Prediction)")

ksvmScat