Craft Beer - Data Analysis

Mor Baruch, Omer Avraham & Itamar Avieli

June 2021

Introduction

At first, it is more than recommend to start reading this project when you have a good beer next to you.

In this research we will analyze the world of craft beers!

The market we chose to focus on and analyze is the USA Craft Beer industry. Our dataset contains more than 1000 beers from different breweries.

In this project you will expand your knowledge of the different types of craft beers, discover the meaning of flavors and ingredients, all combined with statistical analysis.

Table of contents

1. Data Import & Tidying.
2. Beers Tutorial & Visualization.
3. Hypothesis test - What you taste is what you get!
4. Linear regression statistic test - Correlation between IBU&ABV.

Let’s start !

Part 1 - Data Import & Tidying

The main package we used for this project is tidyverse

```
library(tidyverse)
```

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

```
## <U+221A> ggplot2 3.3.3     <U+221A> purrr   0.3.4
## <U+221A> tibble  3.0.6     <U+221A> dplyr   1.0.4
## <U+221A> tidyr   1.1.2     <U+221A> stringr 1.4.0
## <U+221A> readr   1.4.0     <U+221A> forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Explaining The Dataset

The “Craft Beer Dataset” were taken from Kaggle. The dataset was collected in January 2017 from CraftCans.com. The data comes in a CSV file.

Content :

beers.csv - Contains data on 2000+ craft canned beers.

breweries.csv - Contains data for 500+ breweries in the United States

```
beers <- readr::read_csv('\\Users\\mbaru\\Desktop\\High_Tech\\semester b\\data analyze\\Proje
ct\\beers.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   abv = col_double(),
##   ibu = col_double(),
##   id = col_double(),
##   name = col_character(),
##   style = col_character(),
##   brewery_id = col_double(),
##   ounces = col_double()
## )
```

```
breweries <- readr::read_csv('\\Users\\mbaru\\Desktop\\High_Tech\\semester b\\data analyze\\P
roject\\breweries.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   name = col_character(),
##   city = col_character(),
##   state = col_character()
## )
```

About the contents :

Beers.csv

X1 - Row number.

ABV - The alcoholic content by volume with 0 being no alcohol and 1 being pure alcohol.

IBU - International bittering units, which describe how bitter a drink is.

ID - Unique ID for each beer.

Name - Name of the beer.

Style - Beer style (lager, ale, IPA, etc.).

Brewery_id - Unique identifier for brewery that produces this beer; can use to join with brewery info.

Ounces - Size of beer in ounces.

Breweries.csv

Brewery_id - Unique ID for each brewery.

Name - Name of the brewery.

City - City that the brewery is located in.

State - State that the brewery is located. in.

Tidying our data :

As you can see we have two separated datasets. Both are linked together by the “brewery_id”.

The problems we encountered to tidy our data :

1. There were many beers without IBU index and some without ABV index as well.
2. The names of the columns in each dataset where different or too general.
3. The data included two unimportant columns for our analyzing.

So we make it tidy :

```
colnames(breweries) <- c('brewery_id', 'brew_name', 'city', 'state')
colnames(beers)[4:5] <- c("beer_id","beer_name")
main_data <- merge(beers, breweries, by="brewery_id") %>%
  select(c(-"X1", -"ounces")) %>%
  filter(!is.na(ibu) , !is.na(abv))
```

Part 2 - Beers Tutorial & Visualization

So before we dive into some statistical tests, it is important to know what are the four ingredients a beer of any type must have:

1. Grain

Usually barley & wheat. The grains have to go through a malting process which metabolizes the natural grain sugars (more grains you use more sweetness you get), which is what the yeast feeds on during fermentation. To do so, the seed is soaked in water until the plant starts to grow. Just before it emerges from the seed it is put in a kiln and dried. The method of drying can make different colors and flavors of malt, which will affect the beer color.

2. Hops

Hops provide beer with piquant aroma, a variety of flavors, and a delicate-to-intense bitterness that balances the sweetness of the malt. The less time the hops are boiled, the less bitterness in the beer. There are many different hop varieties, just as there are different kinds of tomatoes. Each variety has a flavor and aroma of its own.

3. Yeast

During fermentation, yeast consumes the sugars derived from the malted grain and excretes ethyl alcohol and carbon dioxide in return. (Carbon dioxide is responsible for the gases in beer). There are hundreds of different yeast strains within these categories. Certain strains are suited to making specific beer styles. Some brewers believe the yeast used is the most important element in determining their beer’s character.

4. Water

Considering that beer consists of up to 95% water, the quality of the water is of great importance.

So as you understood every ingredient has is own effect. The yeasts responsible for the type (lager / ale), more hops you use the more bitterness and fruity beer you get, and if you want a beer with high ABV you should use large amount of grains (but don't be surprise about the sweetness).

So now it's the time to explore some of the main types!

Just a reminder:
ABV - Alcohol By Volume
IBU - International Bitterness Units.

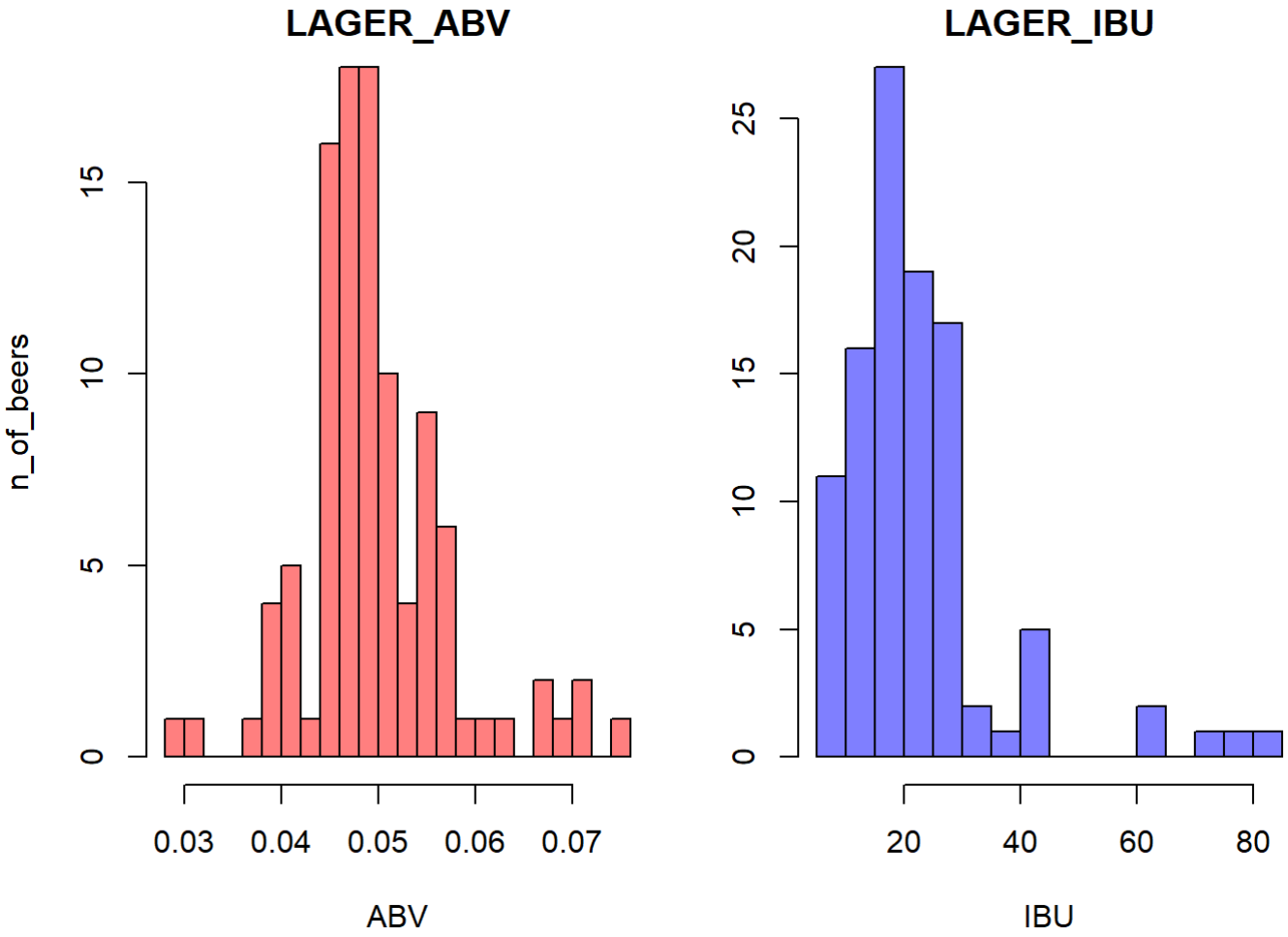
LAGER

Lagers are group of beers and aren't a specific type. Lagers can be pale (named also pilsner), amber, or dark. Usually taste light and a little bitter. Lagers can be a good place to start as you work your way up the flavor ladder.

```
# Filtering the whole beers that contains 'Lager|Pilsner' in style
LAGER_data <- main_data %>%
  filter(grepl("Pilsner|Lager", style))

# Geom_histograms showing the distribution of ABV and IBU in Lager types
par(mfrow=c(1,2), mar=c(4,4,1,0))

hist(LAGER_data$abv, breaks=30, col=rgb(1,0,0,0.5),
     xlab="ABV", ylab="n_of_beers" , main="LAGER_ABV")
hist(LAGER_data$ibu, breaks=20, col=rgb(0,0,1,0.5),
      xlab="IBU" , ylab="" , main="LAGER_IBU")
```



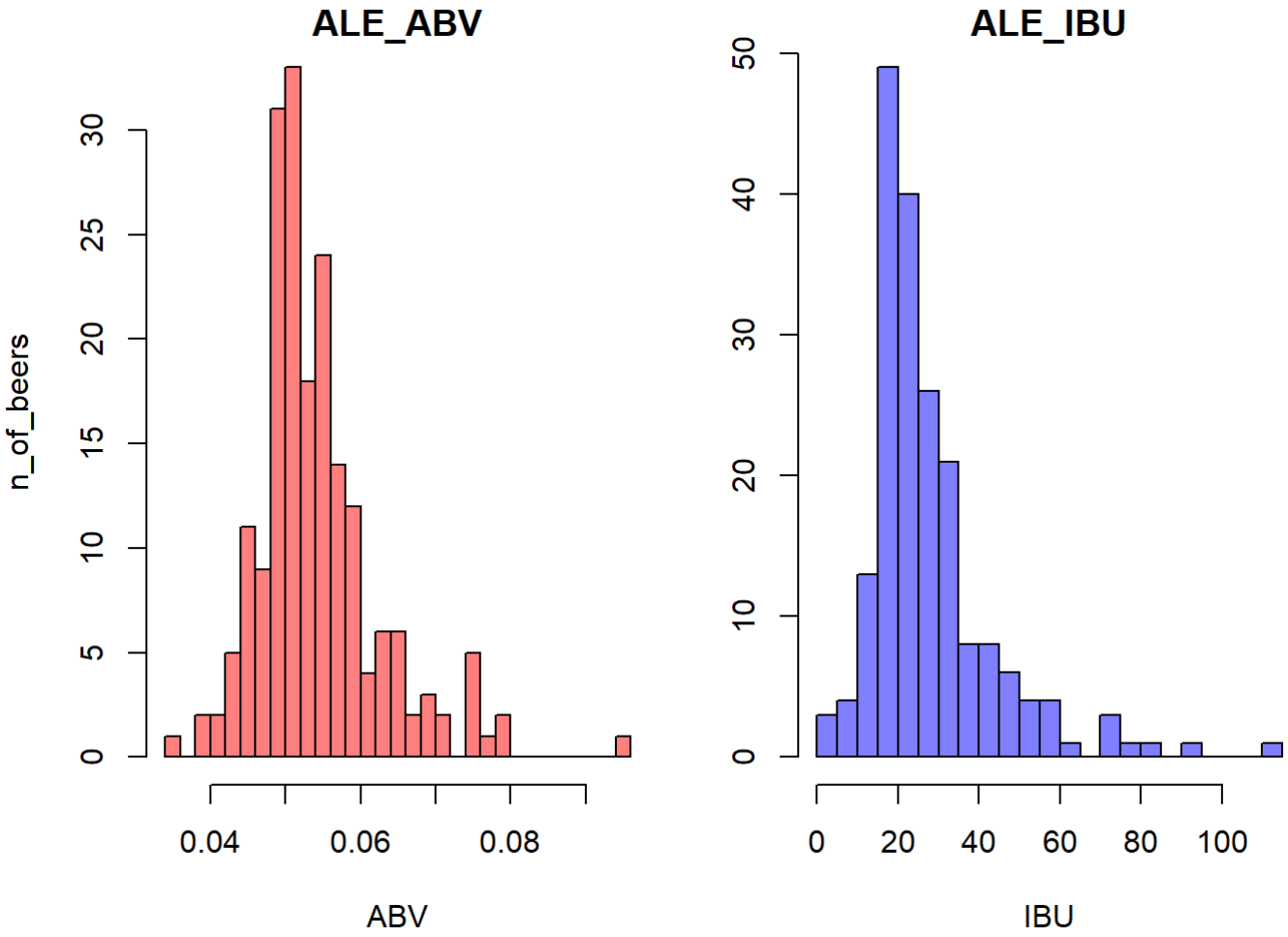
LIGHT ALES

The same as Lagers, Ales refer to a family of beers. We decided to sample some of the most famous types of “Light Ales”. Those beers are characterized by a sweet, full-bodied and fruity taste. As with most beers, ale typically has a bittering agent.

```
# Those are some of the main types we chosed to sample as light ales
light_ale_types = c("American Amber / Red Ale", "American Blonde Ale","American Brown Ale",
"Cream Ale")
# Filtering the whole beers that contains 'light_ale_types' in style
ALE_data <- main_data %>%
  filter(style %in% light_ale_types)

# Geom_histograms showing the distribution of ABV and IBU in Light Ales types
par(mfrow=c(1,2), mar=c(4,4,1,0))

hist(ALE_data$abv, breaks=25, col=rgb(1,0,0,0.5),
     xlab="ABV", ylab="n_of_beers" , main="ALE_ABV")
hist(ALE_data$ibu, breaks=25, col=rgb(0,0,1,0.5),
      xlab="IBU" , ylab="" , main="ALE_IBU")
```



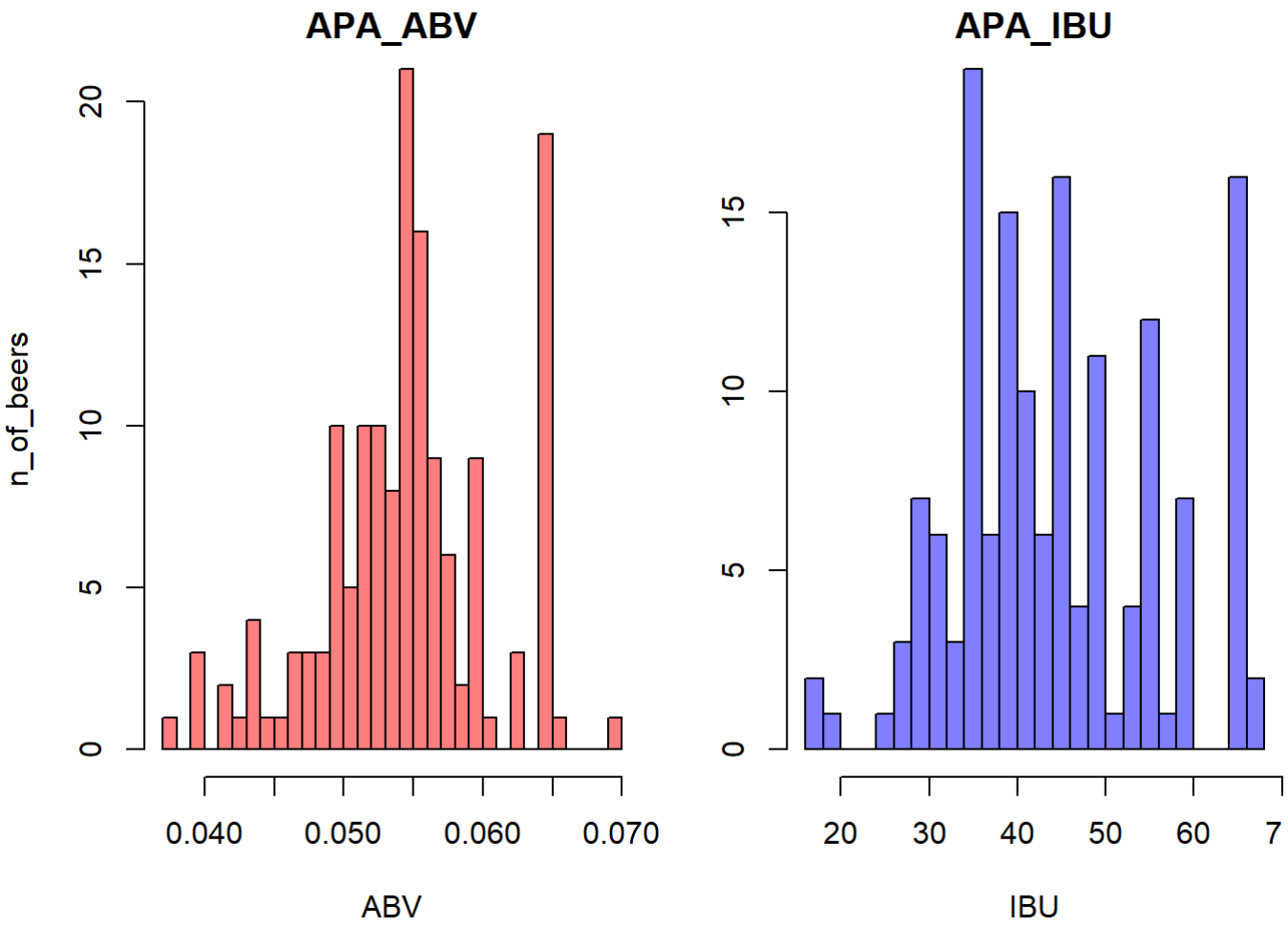
APA - American Pale Ale

As we go with the flavor ladder up we arrive to the American Pale Ale, a specific type of ale. APA is characterized by floral, fruity and hoppy (from hops). We'll find that beer style, a little more bitter than typical light Pale Ale, but yet not bitter as IPA.

```
# Filtering by APA
APA = "American Pale Ale (APA)"
APA_data <- main_data %>%
  filter(style == APA)

# Geom_histograms showing the distribution of ABV and IBU in APA type
par(mfrow=c(1,2), mar=c(4,4,1,0))

hist(APA_data$abv, breaks=30, col=rgb(1,0,0,0.5),
     xlab="ABV", ylab="n_of_beers" , main="APA_ABV")
hist(APA_data$ibu, breaks=30, col=rgb(0,0,1,0.5),
      xlab="IBU" , ylab="" , main="APA_IBU")
```

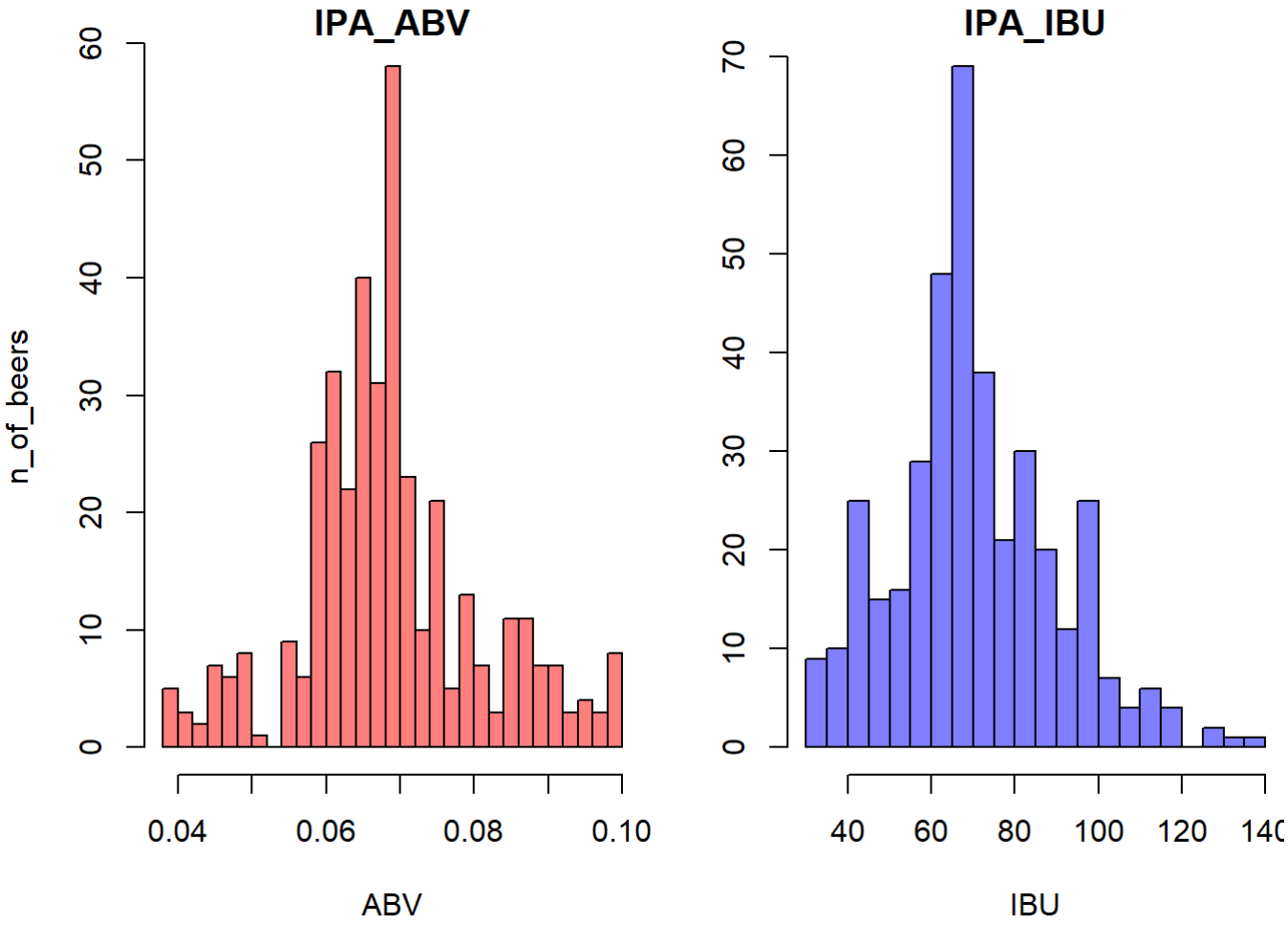


IPA - Indian Pale Ale

IPA originated in England during the 19th century and were made extra strong and hoppy to survive the ocean journey to India (the British soldiers couldn't fight without beers of course). The beer is characterized by powerful and often fruity flavors. Brewed with many hops from different types. A beer for those who understand.

```
# Filtering the whole beers that contains 'IPA' in style
IPA_data <- main_data %>%
  filter(grepl('IPA', style))
# Geom_histograms showing the distribution of ABV and IBU in IPA type
par(mfrow=c(1,2), mar=c(4,4,1,0))

hist(IPA_data$abv, breaks=30, col=rgb(1,0,0,0.5),
     xlab="ABV", ylab="n_of_beers" , main="IPA_ABV")
hist(IPA_data$ibu, breaks=30, col=rgb(0,0,1,0.5),
      xlab="IBU" , ylab="" , main="IPA_IBU")
```

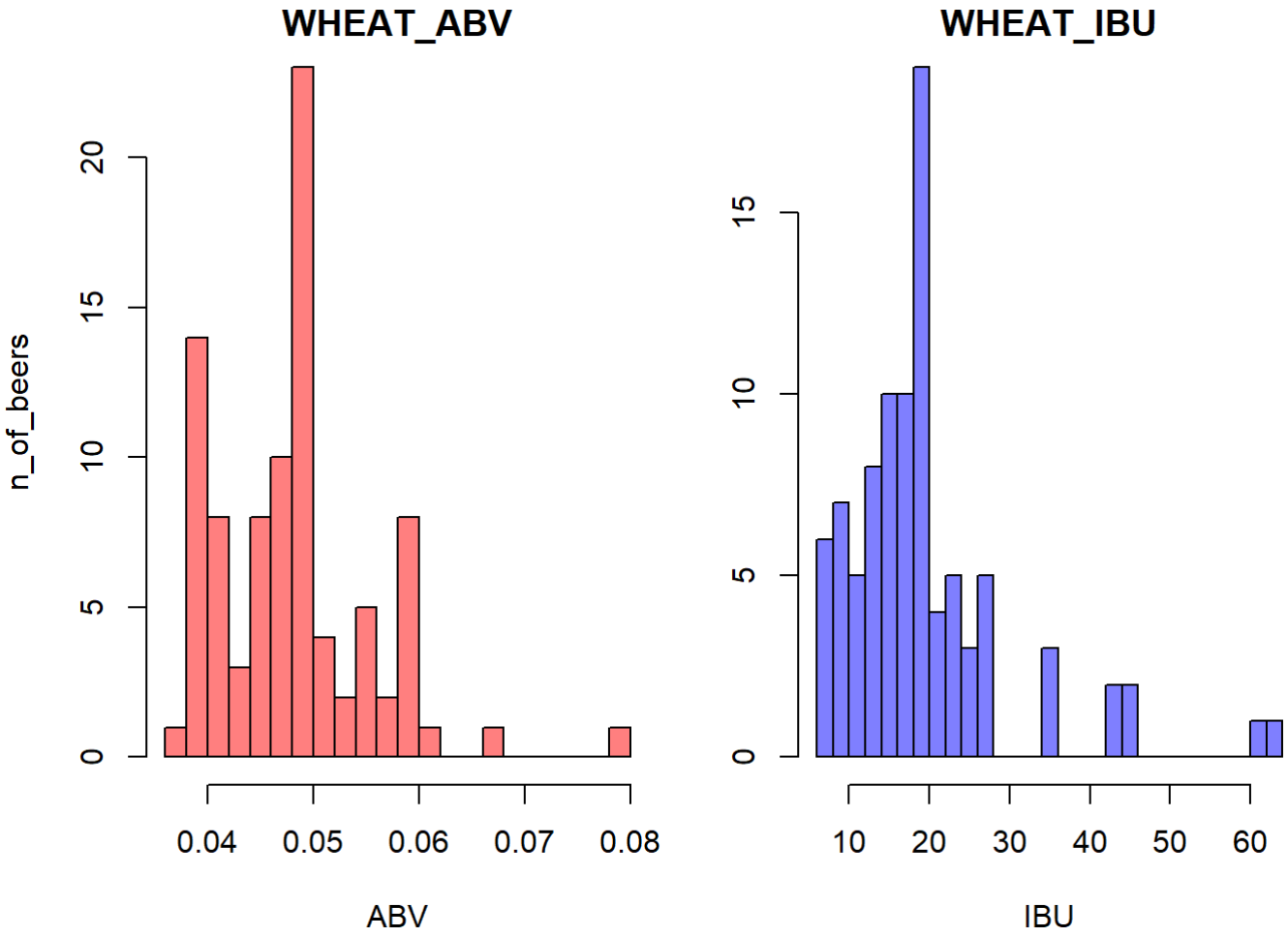


WHEAT

Wheat beer originating in Bavaria. They are brewed with a generous amount of grains, which adds body and flavor. Wheat beers are typically light in color, low to medium in alcohol content, and can be cloudy or clear in appearance, tend to lack bitterness, making them easy drinkers.

```
# Filtering the whole beers that contains 'Wheat|Witbier' in style
WHEAT_data <- main_data %>%
  filter(grepl("Wheat|Witbier", style))
# Geom_histograms showing the distribution of ABV and IBU in Wheat type
par(mfrow=c(1,2), mar=c(4,4,1,0))

hist(WHEAT_data$abv, breaks=20, col=rgb(1,0,0,0.5),
     xlab="ABV", ylab="n_of_beers" , main="WHEAT_ABV")
hist(WHEAT_data$ibu, breaks=25, col=rgb(0,0,1,0.5),
      xlab="IBU" , ylab="" , main="WHEAT_IBU")
```

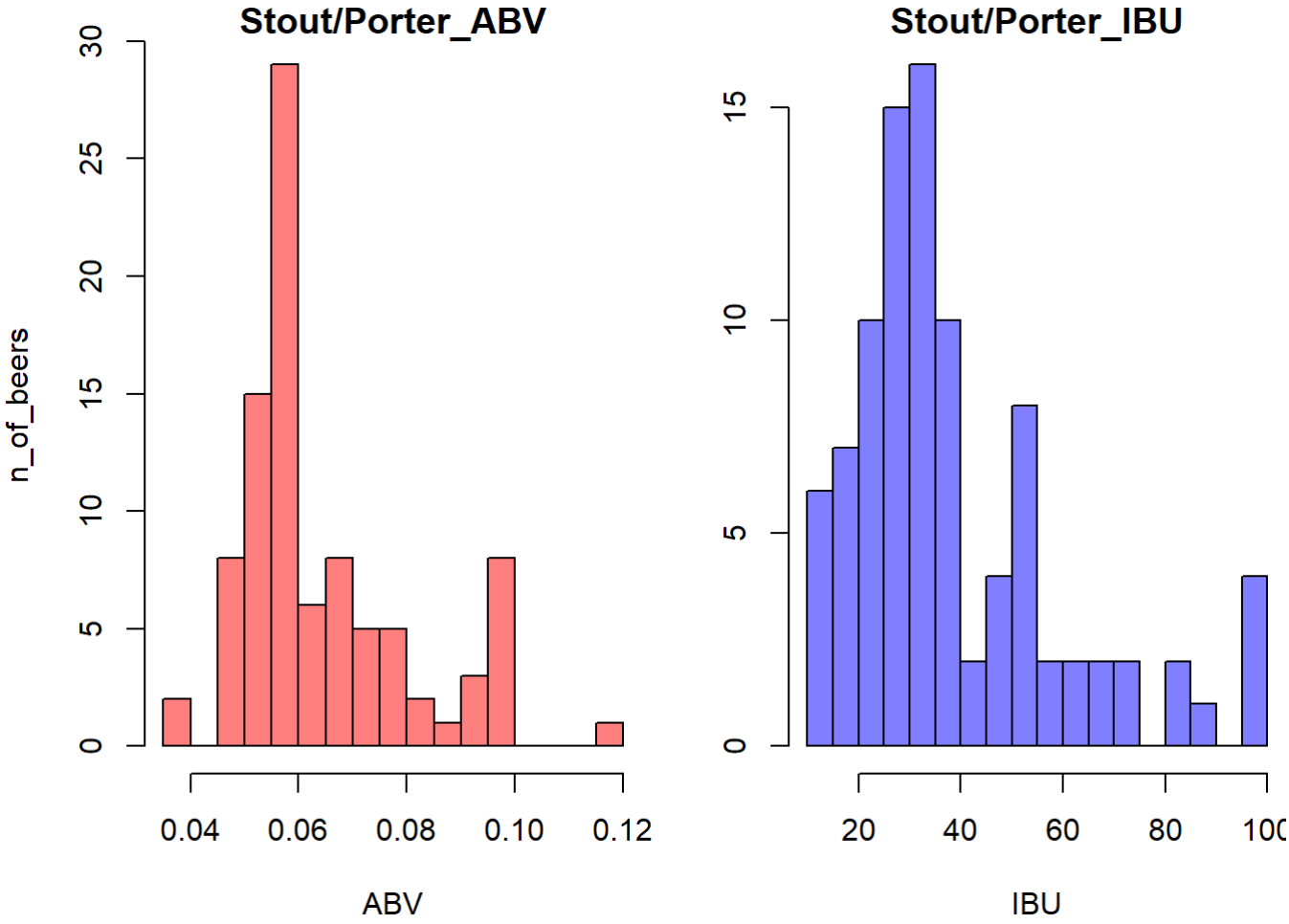


STOUT & PORTER

Stouts are dark-colored ales made with roasted barley that impart chocolate or coffee flavors. The Dark color comes from the long roasting process of the seeds until they burn. The Porter type is a slightly gentle version of the stout.

```
## Filtering the whole beers that contains 'Stout|Porter' in style
STOUT_PORTER_data <- main_data %>%
  filter(grepl("Stout|Porter", style))
# Geom_histograms showing the distribution of ABV and IBU in Stout/Porter types
par(mfrow=c(1,2), mar=c(4,4,1,0))

hist(STOUT_PORTER_data$abv, breaks=20, col=rgb(1,0,0,0.5),
     xlab="ABV", ylab="n_of_beers" , main="Stout/Porter_ABV")
hist(STOUT_PORTER_data$ibu, breaks=30, col=rgb(0,0,1,0.5),
      xlab="IBU" , ylab="" , main="Stout/Porter_IBU")
```



Part 3 - Hypothesis test : What you taste is what you get!

So now it is the time to add some statistical tests to the story.

As we explored in the last chapter there are many differences between the beers styles. In this Hypothesis test we will examine the differences between ALE, APA & IPA.

As we have seen, the bitterness index (which represented by the IBU) plays an important role. Therefore we will perform a statistical test to compare the means of the bitterness index in each type./ To know if we can reject the null hypothesis we will use alpha = 5%.

Main Hypothesis Test :

H_0: The three means are equal.

H_1: IPA is the bitterest, after it APA, then LIGHT_ALES.

Stages for every test:

1. Decide if they are Unpaired or not.
2. Check if the variances are known, and if needed check comparability by F-test.
3. Hypothesis Test by T-test.

First Test :

H_0: \(\mu\)$_{IPA\_IBU}$ \(=\) \(\mu\)$_{APA\_IBU}$

H_1: \(\mu\)$_{IPA\_IBU}$ \(>\) \(\mu\)$_{APA\_IBU}$

1. The variables are unpaired.
2. The variances are unknown, so we have to check if they are equal or not, by F test.

```
# checking if the variance are equal
var.test(IPA_data$ibu, APA_data$ibu, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  IPA_data$ibu and APA_data$ibu
## F = 2.8673, num df = 391, denom df = 152, p-value = 1.224e-12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.180947 3.710707
## sample estimates:
## ratio of variances
##           2.867257
```

Result : The variances are not equal

3. T-test :

```
t.test(IPA_data$ibu, APA_data$ibu, alternative = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  IPA_data$ibu and APA_data$ibu
## t = 19.881, df = 459.07, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  24.76881      Inf
## sample estimates:
## mean of x mean of y
##  71.94898  44.94118
```

OK it seems that we were right (Reject **H_0**)!/ KEEP GOING !

Second Test :

H_0: \(\mu\)$_{APA\_IBU}$ \(=\) \(\mu\)$_{LightAle\_IBU}$

H_1: \(\mu\)$_{APA\_IBU}$ \(>\) \(\mu\)$_{LightAle\_IBU}$

1. The variables are unpaired.
2. The variances are unknown, so we have to check if they are equal or not, by F test.

```
var.test(APA_data$ibu, ALE_data$ibu, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  APA_data$ibu and ALE_data$ibu
## F = 0.52322, num df = 152, denom df = 193, p-value = 3.72e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3880450 0.7093426
## sample estimates:
## ratio of variances
##          0.5232152
```

Result : The variances are not equal

3. T-test :

```
t.test(APA_data$ibu, ALE_data$ibu, alternative = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  APA_data$ibu and ALE_data$ibu
## t = 10.959, df = 342.58, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  13.75661      Inf
## sample estimates:
## mean of x mean of y
##  44.94118  28.74742
```

Reject $\mathbf{H_0}$ again, So… we were right ! :)

Conclusion :

We reject the null hypothesis in both tests, as a result we conclude that we can reject the main null hypothesis. Hence, the IPA style is the bitterest, after it the APA & then LIGHT_ALES.

Part 4 - Linear regression statistic test : Correlation test between IBU&ABV

As we have explained so far about the world of beers, it will be very interesting to see and predict the relationship between the level of alcohol in the beer and the level of bitterness of the beer. We decided to build a simple regression model to get a perspective on the relationship between them. We chose that the independent variable will be the ABV and the dependent variable will be the IBU. That means, by given the level of alcohol we can try to predict the level of bitterness in a given beer.

Our model: we used alpha = 5%. \(Y\) \(=\) $\mathbf{b_0}$ \(+\) $\mathbf{b_1}$\(X\) + $\mathbf{e_i}$

```
ibu_abv <- main_data %>%
select(ibu, abv)

cor(ibu_abv)
```

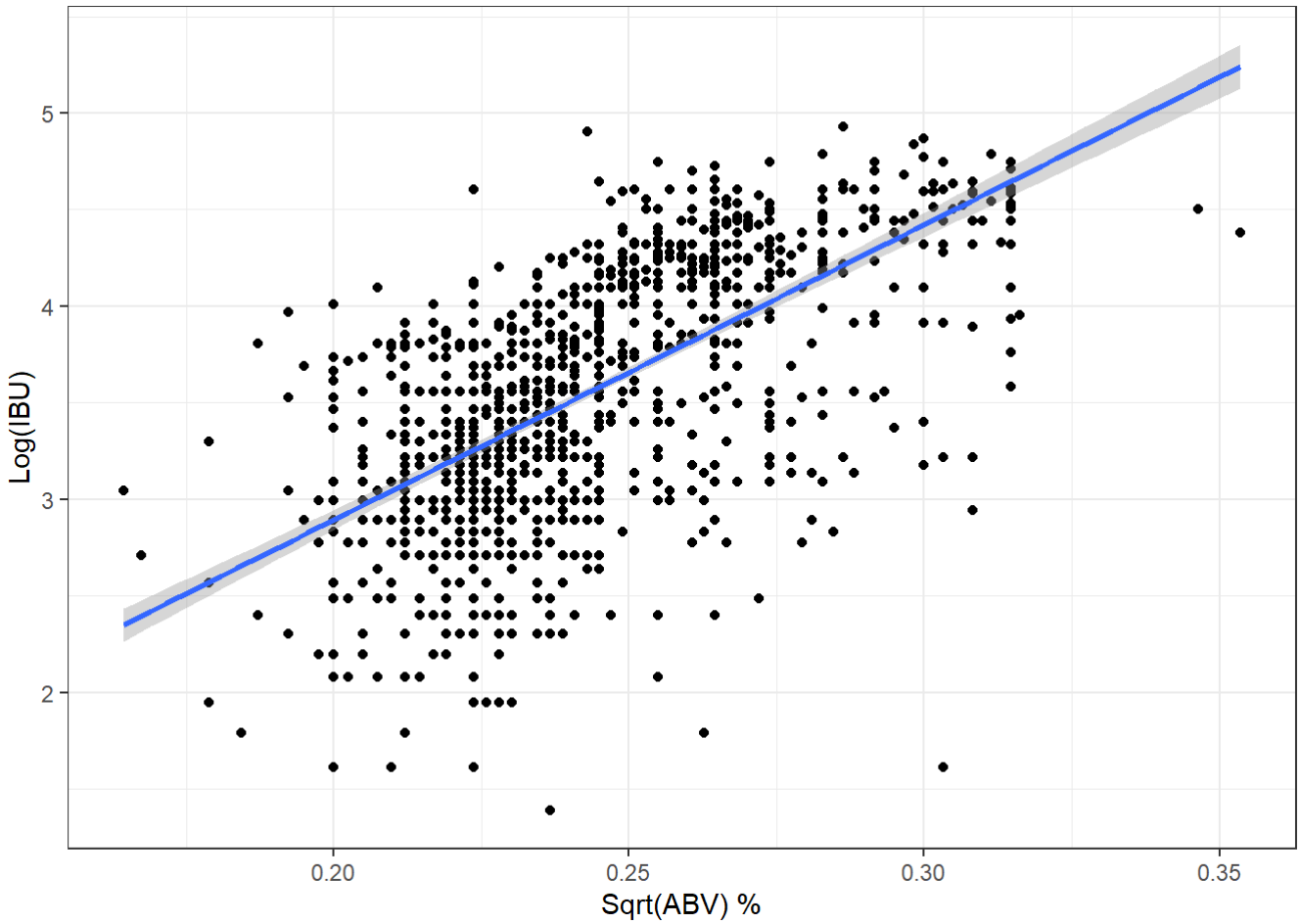
```
##           ibu       abv
## ibu 1.0000000 0.6706215
## abv 0.6706215 1.0000000
```

We can see that our test gives the impression that there is a high correlation between the variables.

Let's see how it looks :

```
# the y column is in log(10) *** ask ADI how to explain
IBU_ABV_DOTS <- ggplot(main_data, aes(x = sqrt(abv), y = log(ibu))) +
  geom_point() + theme_bw() +
  stat_smooth(method = "lm") + xlab("Sqrt(ABV) %") + ylab("Log(IBU)")

IBU_ABV_DOTS
```



So as you can see from the graph it seems that there is indeed a correlation between the variables, and as the level of alcohol increases so does the level of bitterness.

Let's investigate more thoroughly.

```
ibu_abv_lm <- lm(formula = log(ibu) ~ sqrt(abv), data = main_data)
summary(ibu_abv_lm)
```

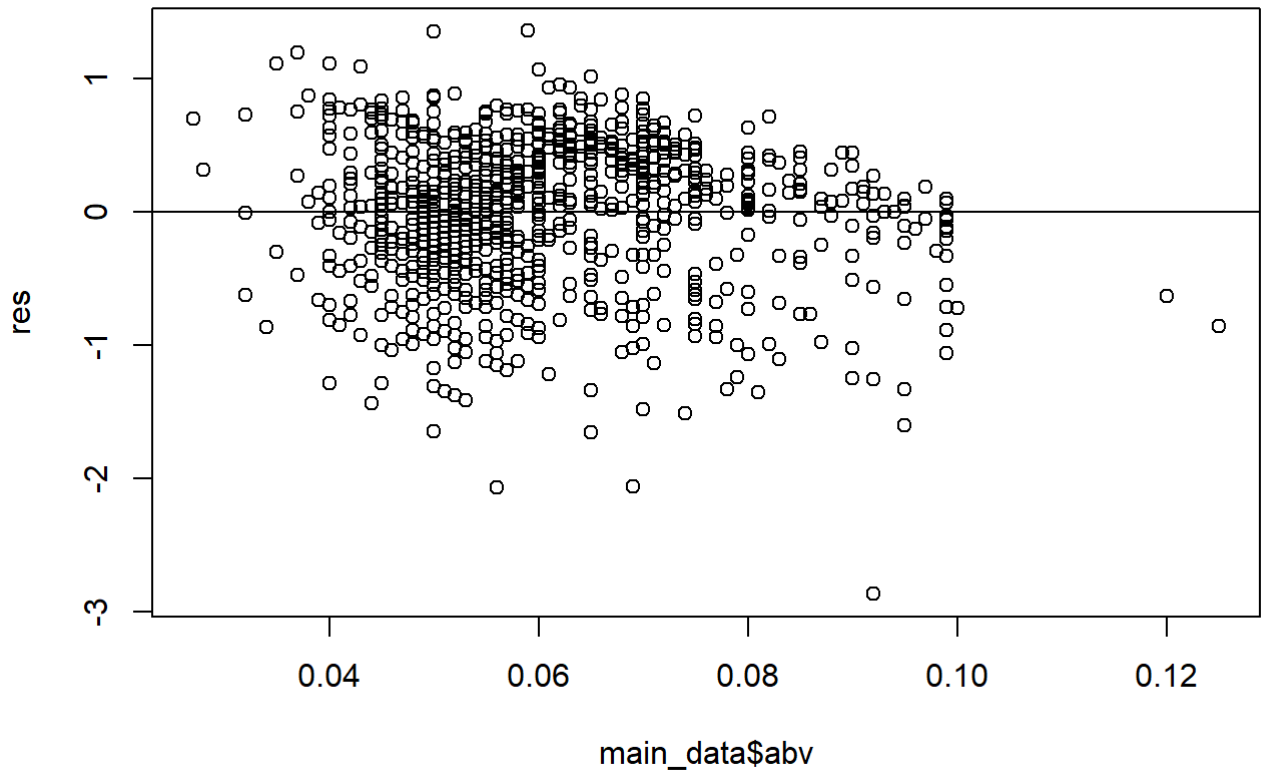
```
##
## Call:
## lm(formula = log(ibu) ~ sqrt(abv), data = main_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86318 -0.29880  0.07778  0.39173  1.35600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1630     0.1258  -1.295    0.195
## sqrt(abv)    15.2832     0.5141  29.729   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5166 on 1403 degrees of freedom
## Multiple R-squared:  0.3865, Adjusted R-squared:  0.386
## F-statistic: 883.8 on 1 and 1403 DF,  p-value: < 2.2e-16
```

$LOG(IBU) = -0.163 + sqrt(ABV) \times 15.2832$

It can be seen that we received a very low P-value (<2.2e-16). Which indicates that the choice of the alcohol index to estimate the bitterness index is a good choice.

How do we know if our model is good? We'll check R-squared, it seems that it is not so high (we would like to get R-squared that is close to one), Therefore we will try to understand if there is Homoscedasticity by looking at the residuals.

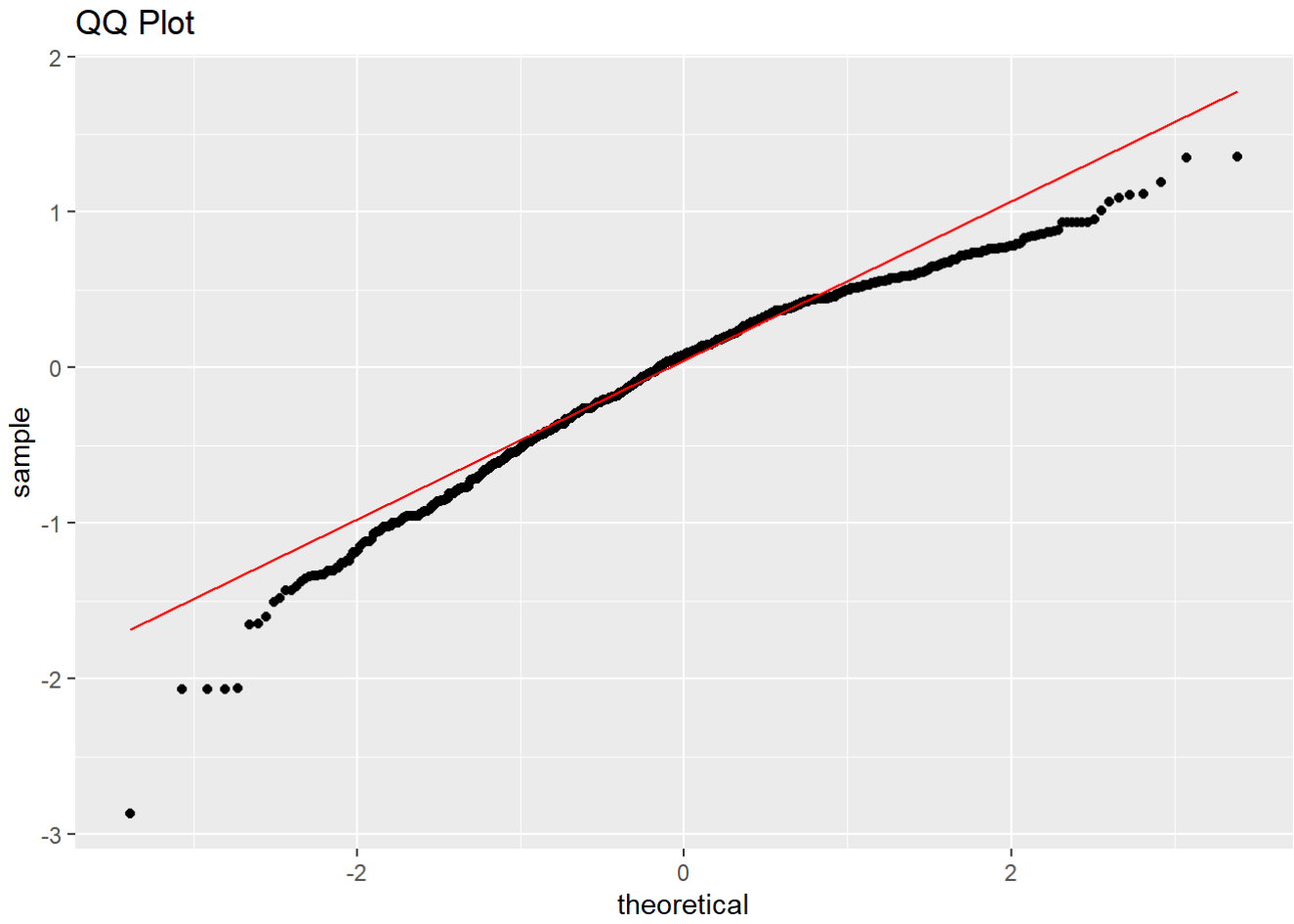
```
res <- resid(ibu_abv_lm)
plot(main_data$abv, res)
abline(0,0)
```



You can see that we got a result that indicates Homoscedasticity, which means that the scattering of our point does not depend on X.

In addition we will check how does the qqplot reflects our model, qqplot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution, in our case a Normal distribution.

```
ibu_abv_lm %>% ggplot(aes(sample=.resid)) +
 geom_qq() + geom_qq_line(col="red") +
 labs(title="QQ Plot")
```



The qq_plot shows us that most of our data is distributed normally, However in the edges there is little irregularity.

Conclusion

In this project we checked two main questions:

How does the level of bitterness affects the beer style?

Is there a correlation between the amount of alcohol in a beer to the amount of bitterness in it?

Both of our tests had interesting results. In the first one we showed that the differences between the beer styles are statistically significant. The second test showed a correlation between the amount of alcohol in a beer to the amount of bitterness, now we have a statistical equation for predicting the level of alcohol in the beer according to the amount of bitterness at a significance level of 95 percent.

Thank you for reading!