Introduction

The project looks into the a data set over viewing certain beers and rating them by a criteria.

```
#install packages
library(dplyr)
library(tidyverse)
library(ggplot2)
library(readr)
library(data.table)
library(corrplot)
library(corrgram)
library(forcats)
```

Overview of the Data

The data set goes over beer that were reviewed by people based on characteristic.

The data set has 13 variables/columns as follows:

beer_ABV: alcohol by volume, is the standard measurement, used worldwide, to assess the strength of a particular beer

beer_beerId: Id number that was given for the beer_name

beer_brewerId: Id number that given for the beer_style

beer_name: Name of the beer

beer_style: Style of the beer

review_appearance: How the look of the beer looked to reviewer overall. Scored by 1 out of 5.

review_palette: The range of taste of the beer.Scored by 1 out of 5.

review_overall: The overall review of the beer. Taken the mean of all the other reviews. Scored by 1 out of 5.

review_taste : How the beer tasted. Scored by 1 out of 5.

review_profileName: The profile name of the reviewer.

review_aroma: The scent of the beer. Scored by 1 out of 5.

review_text: The comment the reviewer left.

review_time: When the reviwer made the review.

```
# puts the data set into variable BP
BP <- read.csv("BeerProject.csv")
# shows the first 5 rows of the data set
head(BP)
```

```
##   beer_ABV beer_beerId beer_brewerId              beer_name
## 1      5.0       47986         10325           Sausa Weizen
## 2      6.2       48213         10325               Red Moon
## 3      6.5       48215         10325 Black Horse Black Beer
## 4      5.0       47969         10325             Sausa Pils
## 5      7.7       64883          1075          Cauldron DIPA
## 6      4.7       52159          1075    Caldera Ginger Beer
##                        beer_style review_appearance review_palette
## 1                     Hefeweizen              2.5            2.0
## 2             English Strong Ale              3.0            2.5
## 3         Foreign / Export Stout              3.0            2.5
## 4                German Pilsener              3.5            3.0
## 5 American Double / Imperial IPA              4.0            4.5
## 6           Herbed / Spiced Beer              3.5            3.5
##   review_overall review_taste review_profileName review_aroma
## 1            1.5          1.5            stcules          1.5
## 2            3.0          3.0            stcules          3.0
## 3            3.0          3.0            stcules          3.0
## 4            3.0          2.5            stcules          3.0
## 5            4.0          4.0     johnmichaelsen          4.5
## 6            3.0          3.0            oline73          3.5
##
## 1
## 2
## 3
## 4
## 5 According to the website, the style for the Caldera Cauldron changes every year. The current releas
## 6
##   review_time
## 1  1234817823
## 2  1235915097
## 3  1235916604
## 4  1234725145
## 5  1293735206
## 6  1325524659
```

```
# Shows the variable and how many rows the data set has
str(BP)
```

```
## ’data.frame’:    528870 obs. of  13 variables:
##  $ beer_ABV          : num  5 6.2 6.5 5 7.7 4.7 4.7 4.7 4.7 4.7 ...
##  $ beer_beerId       : int  47986 48213 48215 47969 64883 52159 52159 52159 52159 52159 ...
##  $ beer_brewerId     : int  10325 10325 10325 10325 1075 1075 1075 1075 1075 1075 ...
##  $ beer_name         : chr  "Sausa Weizen" "Red Moon" "Black Horse Black Beer" "Sausa Pils" ...
##  $ beer_style        : chr  "Hefeweizen" "English Strong Ale" "Foreign / Export Stout" "German Pilse
##  $ review_appearance : num  2.5 3 3 3.5 4 3.5 3.5 3.5 3.5 5 ...
##  $ review_palette    : num  2 2.5 2.5 3 4.5 3.5 3.5 2.5 3 3.5 ...
##  $ review_overall    : num  1.5 3 3 3 4 3 3.5 3 4 4.5 ...
##  $ review_taste      : num  1.5 3 3 2.5 4 3 4 2 3.5 4 ...
##  $ review_profileName: chr  "stcules" "stcules" "stcules" "stcules" ...
##  $ review_aroma      : num  1.5 3 3 3 4.5 3.5 4 3.5 4 4 ...
##  $ review_text       : chr  "A lot of foam. But a lot. In the smell some banana, and then lactic and
##  $ review_time       : int  1234817823 1235915097 1235916604 1234725145 1293735206 1325524659 131899
```

```
# Shows the overall summary of the data set
summary(BP)
```

```
##     beer_ABV       beer_beerId    beer_brewerId    beer_name
##  Min.   : 0.010   Min.   :    3   Min.   :    1   Length:528870
##  1st Qu.: 5.300   1st Qu.: 1745   1st Qu.:  132   Class :character
##  Median : 6.500   Median :14368   Median :  394   Mode  :character
##  Mean   : 7.017   Mean   :22098   Mean   : 2598
##  3rd Qu.: 8.500   3rd Qu.:40528   3rd Qu.: 1475
##  Max.   :57.700   Max.   :77310   Max.   :27980
##  NA's   :20280
##   beer_style        review_appearance review_palette  review_overall
##  Length:528870      Min.   :0.000     Min.   :1.000   Min.   :0.000
##  Class :character   1st Qu.:3.500     1st Qu.:3.500   1st Qu.:3.500
##  Mode  :character   Median :4.000     Median :4.000   Median :4.000
##                     Mean   :3.865     Mean   :3.759   Mean   :3.833
##                     3rd Qu.:4.000     3rd Qu.:4.000   3rd Qu.:4.500
##                     Max.   :5.000     Max.   :5.000   Max.   :5.000
##
##   review_taste   review_profileName  review_aroma   review_text
##  Min.   :1.000   Length:528870      Min.   :1.000   Length:528870
##  1st Qu.:3.500   Class :character   1st Qu.:3.500   Class :character
##  Median :4.000   Mode  :character   Median :4.000   Mode  :character
##  Mean   :3.766                      Mean   :3.817
##  3rd Qu.:4.000                      3rd Qu.:4.500
##  Max.   :5.000                      Max.   :5.000
##
##   review_time
##  Min.   :8.844e+08
##  1st Qu.:1.175e+09
##  Median :1.240e+09
##  Mean   :1.225e+09
##  3rd Qu.:1.289e+09
##  Max.   :1.326e+09
##
```

Cleaning the data set

The cleaning process involved checking to see if there were any NA's or blanks in the column. Removing the following columns;review_text, review_time and review_profileName, as I felt those columns would not affect the results of the findings of the this project. Making a function to get the mean review score of the beers as beers were reviewed more then once.

```r
colSums(is.na(BP))  # Shows any columns that contains NA's or Blanks
```

```
##           beer_ABV        beer_beerId      beer_brewerId          beer_name
##              20280                  0                  0                  0
##         beer_style  review_appearance     review_palette     review_overall
##                  0                  0                  0                  0
##       review_taste review_profileName       review_aroma        review_text
##                  0                  0                  0                  0
##        review_time
##                  0
```

```r
# functions gets the overall average review score of the beers as beers were reviewed more then once
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
```

```
beer_ABV_mode<-getmode(BP$beer_ABV)
BP$beer_ABV[which(is.na(BP$beer_ABV))] <- beer_ABV_mode

colSums(is.na(BP))
```

```
##           beer_ABV        beer_beerId      beer_brewerId          beer_name
##                  0                  0                  0                  0
##         beer_style  review_appearance     review_palette     review_overall
##                  0                  0                  0                  0
##       review_taste review_profileName      review_aroma        review_text
##                  0                  0                  0                  0
##        review_time
##                  0
```

```
# removes columns
BP2 <- BP %>% subset(select=-c(review_text,review_time,review_profileName))
```

```
# sees the first 5 rows of the edited data set
head(BP2)
```

```
##   beer_ABV beer_beerId beer_brewerId              beer_name
## 1      5.0       47986         10325           Sausa Weizen
## 2      6.2       48213         10325               Red Moon
## 3      6.5       48215         10325 Black Horse Black Beer
## 4      5.0       47969         10325             Sausa Pils
## 5      7.7       64883          1075          Cauldron DIPA
## 6      4.7       52159          1075    Caldera Ginger Beer
##                       beer_style review_appearance review_palette
## 1                     Hefeweizen               2.5            2.0
## 2             English Strong Ale               3.0            2.5
## 3         Foreign / Export Stout               3.0            2.5
## 4                German Pilsener               3.5            3.0
## 5 American Double / Imperial IPA               4.0            4.5
## 6           Herbed / Spiced Beer               3.5            3.5
##   review_overall review_taste review_aroma
## 1            1.5          1.5          1.5
## 2            3.0          3.0          3.0
## 3            3.0          3.0          3.0
## 4            3.0          2.5          3.0
## 5            4.0          4.0          4.5
## 6            3.0          3.0          3.5
```

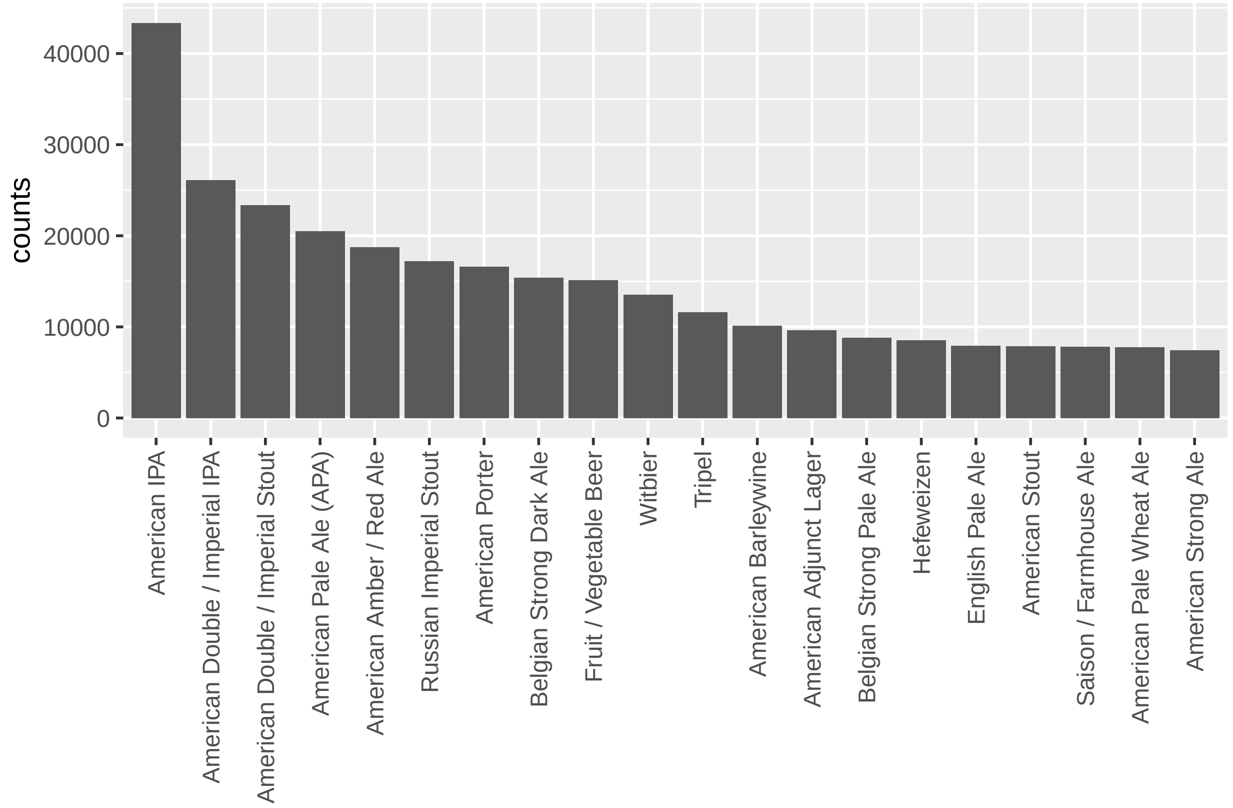
Visualizations of the characteristics that were reviewed for the beers.

The following code chucks shows aggregation and visualizations of characteristics of the beers that were scored. The following visualization were are table that shows the count or the average score of the beers. A point chart or a bar chart to visualize the results from the table. A histogram to shows the frequency of the score,

```
# counts the different types of beers.Outputs only the first 5 rows
type_df <- BP2 %>% group_by(beer_style) %>% summarize(counts=n())
head(type_df)
```

```
## # A tibble: 6 x 2
##   beer_style                  counts
##   <chr>                        <int>
## 1 Altbier                       3708
## 2 American Adjunct Lager        9613
## 3 American Amber / Red Ale     18731
## 4 American Amber / Red Lager    2935
## 5 American Barleywine          10108
## 6 American Black Ale            3055
```

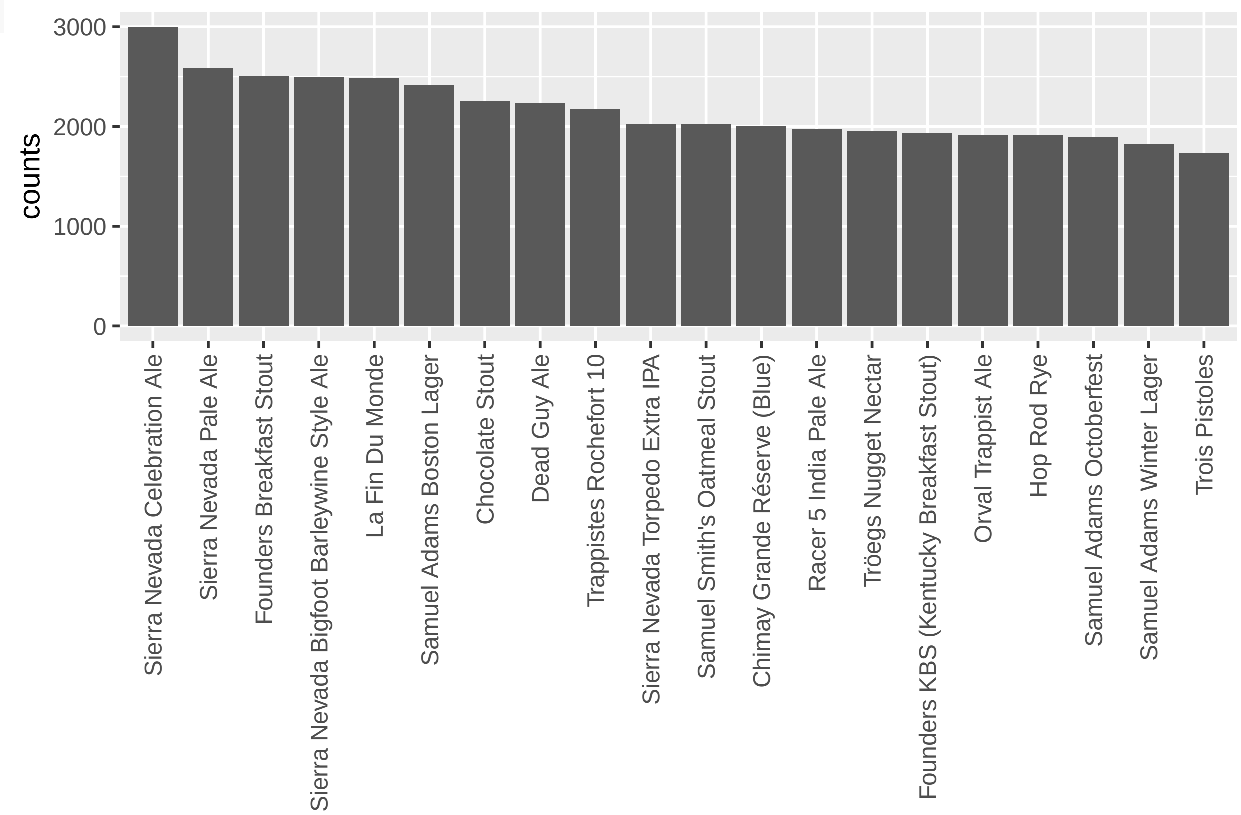
```
# Charts the counts of the type of beers in a bar chart
type_df %>% top_n(n=20) %>% mutate(beer_style = fct_reorder(beer_style,desc(counts))) %>%
ggplot(aes(x=beer_style,y=counts)) +
```



```
# counts the names of the beers that were reviewed
name_df <- BP2 %>% group_by(beer_name) %>% summarize(counts=n())
name_df
```

```
## # A tibble: 18,339 x 2
##    beer_name                                          counts
##    <chr>                                               <int>
##  1 '99 Wee Heavy Scotch Ale                                2
##  2 'Bout Time Barley Wine                                  5
##  3 'Pooya Porter                                           2
##  4 'Sconnie Pale Ale                                       1
##  5 'Sconnie Rustic Trail Amber                             2
##  6 'Sconnie Tall Blonde Ale                                1
##  7 't Gaverhopke / Tired Hands Bitter Sweet Symphony       2
##  8 't Gaverhopke De Kriek (Red Cap)                        1
##  9 't Gaverhopke Den Blond 8° (White Cap)                6
## 10 't Gaverhopke Den Bruin 8° (Blue Cap)                 3
## # ... with 18,329 more rows
```

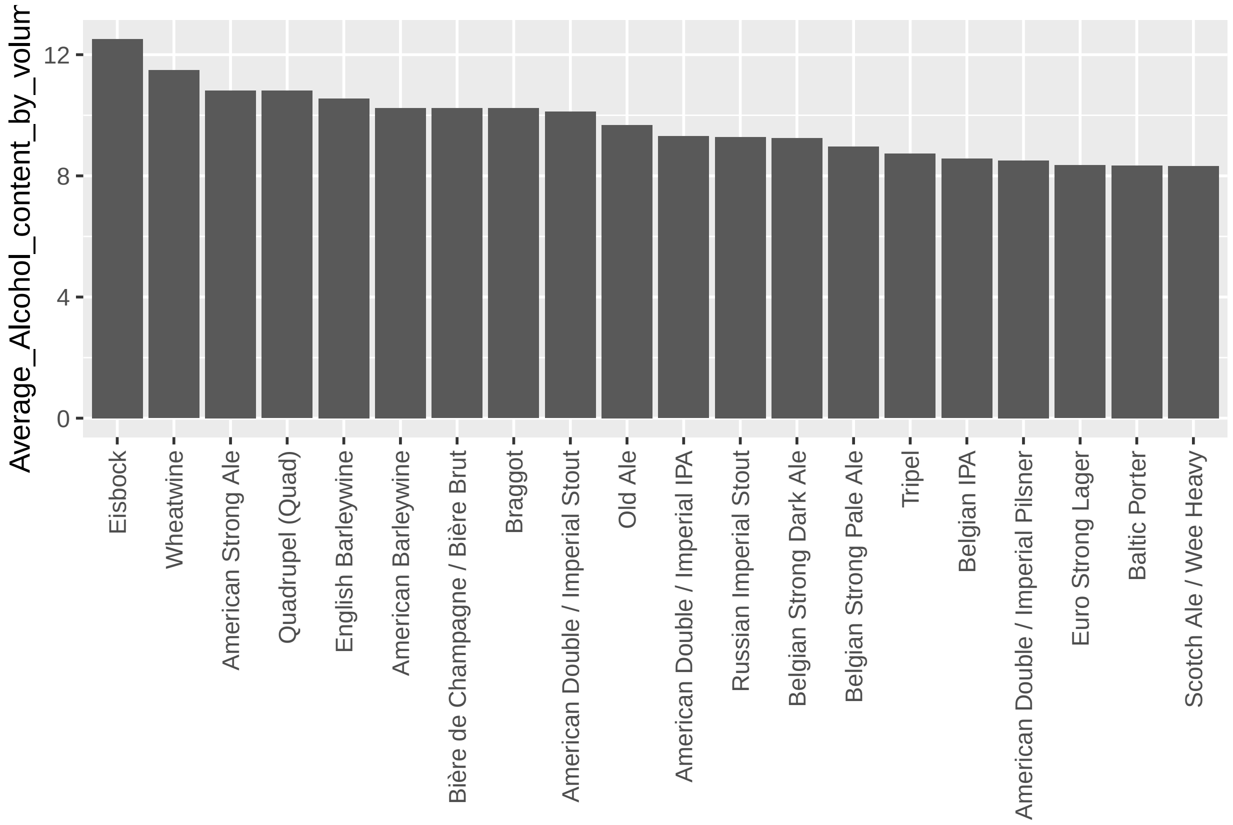
```
# Charts the counts of the name of the beers in a bar chart
name_df %>% top_n(n=20) %>% mutate(beer_name = fct_reorder(beer_name,desc(counts))) %>%
ggplot(aes(x=beer_name,y=counts)) +
```



```
# shows the average alcohol content of the beers. Output only shows the first top 5 of the data set
beer_ABV_df <- BP2 %>% group_by(beer_style) %>%  summarize(mean(beer_ABV))
names(beer_ABV_df)[2] <- "Average_Alcohol_content_by_volume"
beer_ABV_df <- beer_ABV_df %>% arrange(desc(Average_Alcohol_content_by_volume))
head(beer_ABV_df)
```

```
## # A tibble: 6 x 2
##   beer_style          Average_Alcohol_content_by_volume
##   <chr>                                           <dbl>
## 1 Eisbock                                          12.5
## 2 Wheatwine                                        11.5
## 3 American Strong Ale                              10.8
## 4 Quadrupel (Quad)                                 10.8
## 5 English Barleywine                               10.6
## 6 American Barleywine                              10.2
```

```
# charts the results of the table in a bar chart. Outputs only the top 20
beer_ABV_df %>% top_n(n=20) %>%
mutate(beer_style = fct_reorder(beer_style, Average_Alcohol_content_by_volume, .desc = TRUE)) %>% ggplo
geom_bar(stat = "identity")+theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



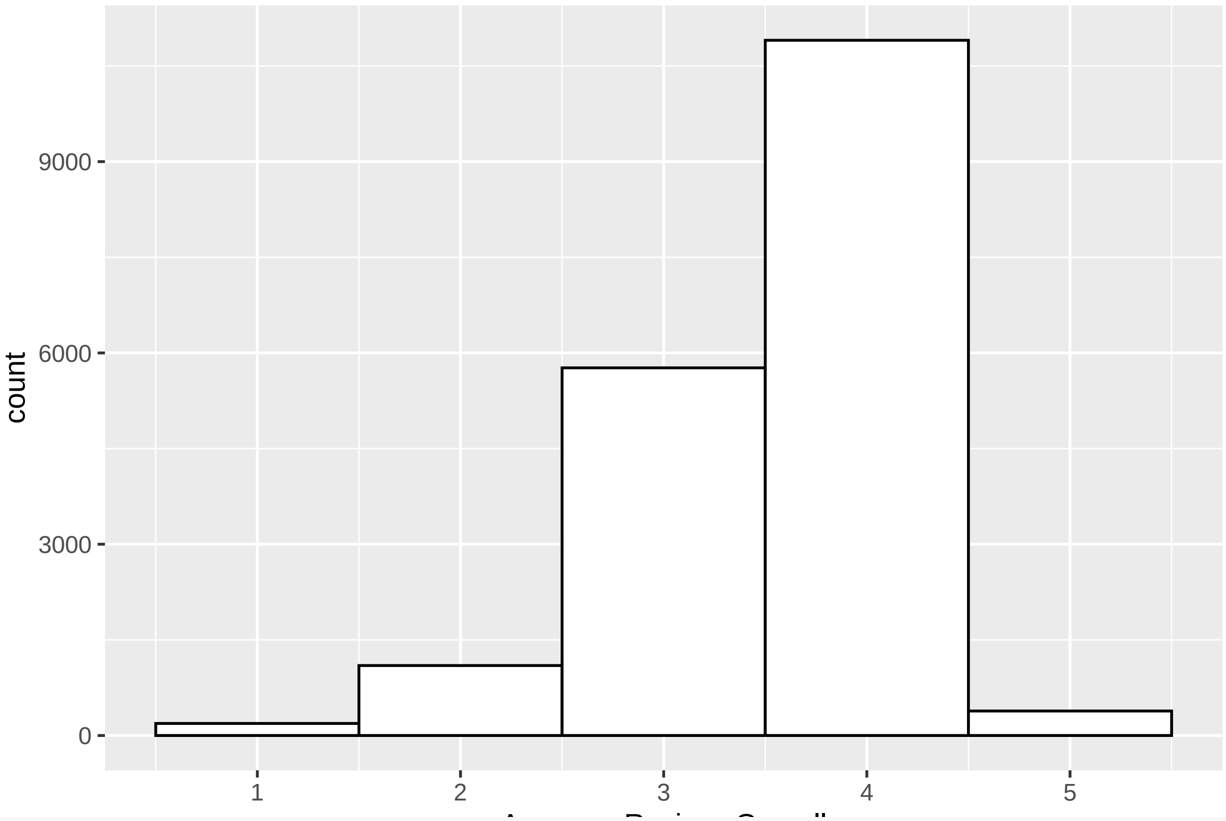
```
# Shows the overall rating of the beers. Outputs only the first 5 rows of the data set
beer_name_rating <- BP2 %>% group_by(beer_name) %>% summarize(mean(review_overall))

names(beer_name_rating)[2] <- "Average_Review_Overall"

beer_name_rating <- beer_name_rating %>% arrange(desc(Average_Review_Overall))
beer_name_rating
```

```
## # A tibble: 18,339 x 2
##    beer_name                                                Average_Review_Overa~
##    <chr>                                                                    <dbl>
##  1 '99 Wee Heavy Scotch Ale                                                     5
##  2 10th Anniversary Strong Belgian                                              5
##  3 2005 Grand Cru                                                               5
##  4 3X IPA                                                                       5
##  5 Ackerman's Imperial Double Stout (Winterfest Replicale~                      5
##  6 AleSmith Speedway Stout - Oak Aged                                           5
##  7 All About Amber                                                              5
##  8 Amarillo Single Hop Pale Ale                                                 5
##  9 Amber Classic                                                                5
## 10 Auger Falls Dark Amber Ale                                                   5
## # ... with 18,329 more rows
```

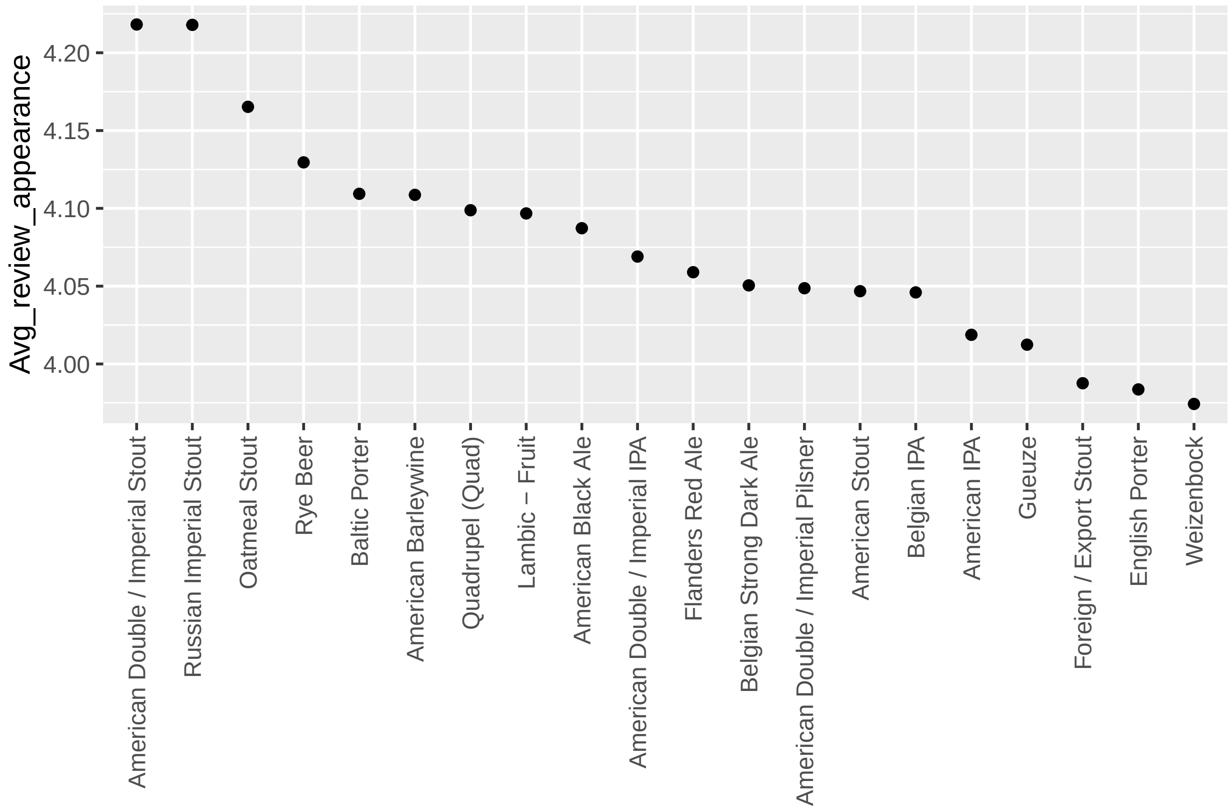
```
# shows the frequency of each review score
beer_name_rating %>% ggplot(aes(x=Average_Review_Overall)) +
geom_histogram(color="black", fill="white",binwidth=1)
```



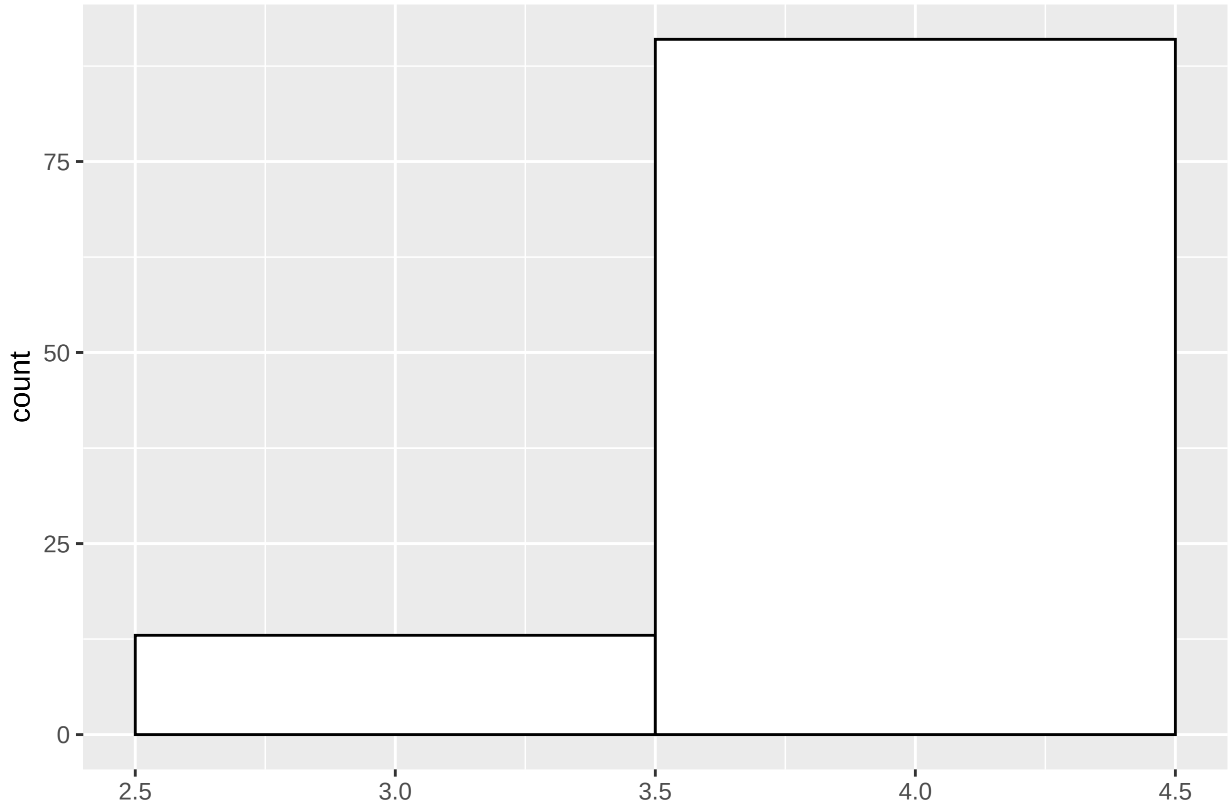
```
# Shows the average appearance score by beer style. Outputs only the first 5 rows
review_appearance_df <- BP2 %>% group_by(beer_style
) %>%  summarize(mean(review_appearance))
names(review_appearance_df)[2] <- "Avg_review_appearance"
review_appearance_df<- review_appearance_df %>% arrange(desc(Avg_review_appearance))
head(review_appearance_df)
```

```
## # A tibble: 6 x 2
##   beer_style                           Avg_review_appearance
##   <chr>                                                <dbl>
## 1 American Double / Imperial Stout                      4.22
## 2 Russian Imperial Stout                                4.22
## 3 Oatmeal Stout                                         4.17
## 4 Rye Beer                                              4.13
## 5 Baltic Porter                                         4.11
## 6 American Barleywine                                   4.11
```

```
# Charts the results from the table in a point chart. Only output the top 20.
review_appearance_df %>% top_n(n=20)%>%
mutate(beer_style = fct_reorder(beer_style, Avg_review_appearance, .desc = TRUE)) %>%
ggplot(aes(beer_style,Avg_review_appearance))+geom_point(stat = "identity")+
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



```
# shows the frequency of each review score
review_appearance_df %>% ggplot(aes(x=Avg_review_appearance)) +
```

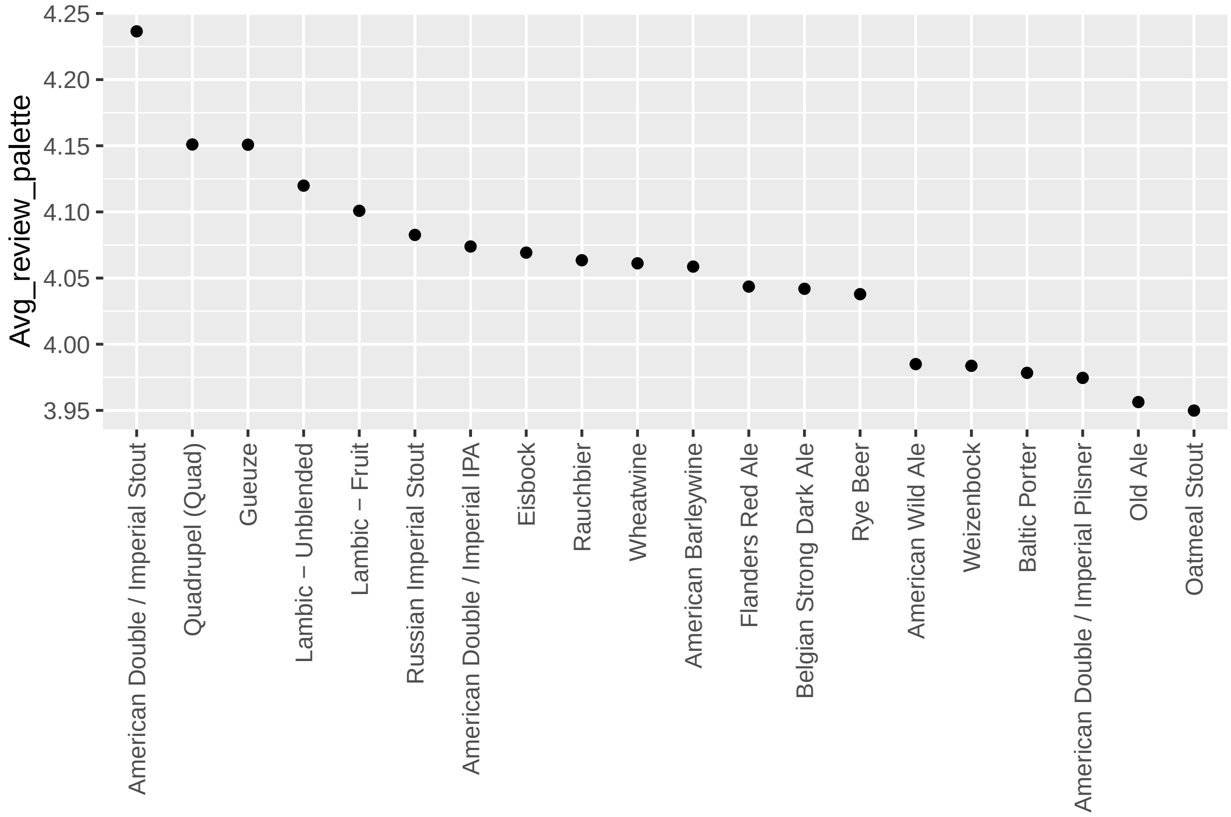


```
# Shows the average review palette by beer style. Outputs only the first 5 rows
review_palette_df <- BP2 %>% group_by(beer_style) %>% summarize(mean(review_palette))
names(review_palette_df)[2] <- "Avg_review_palette"
```

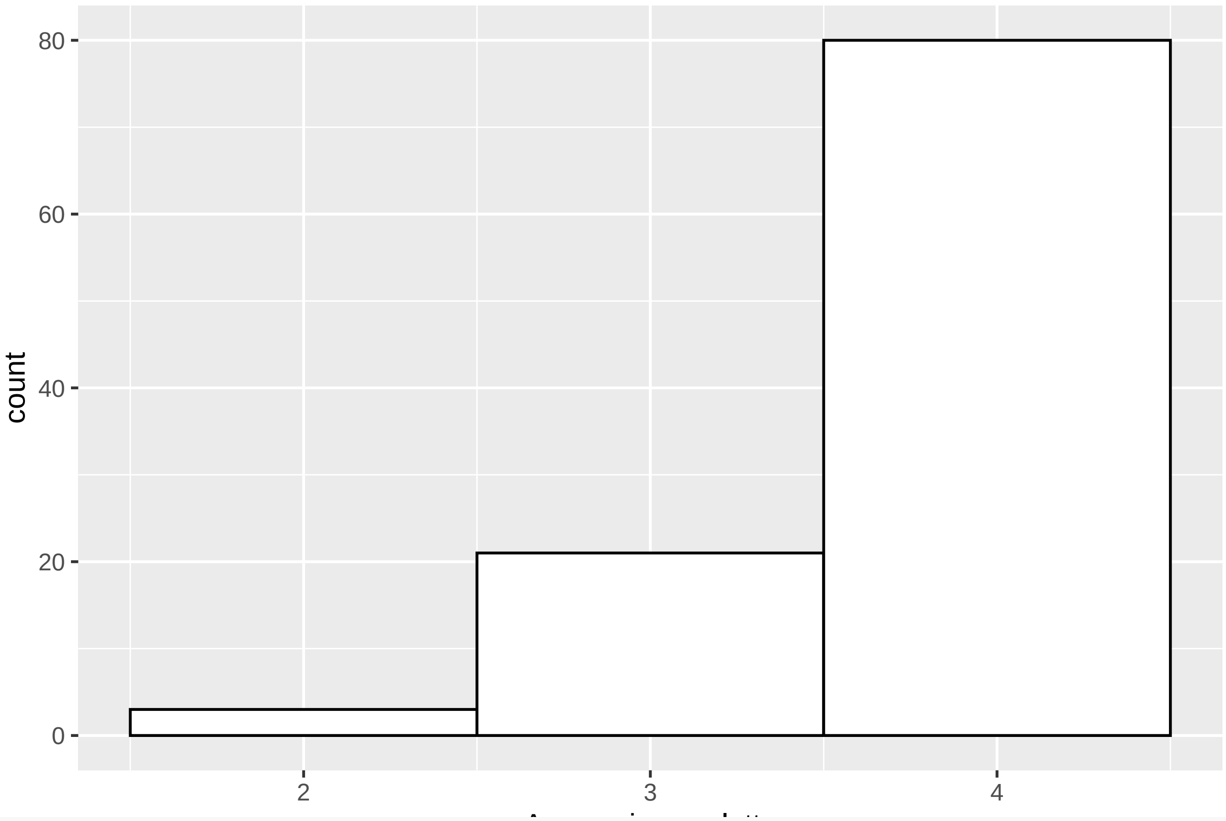
```
review_palette_df <- review_palette_df %>% arrange(desc(Avg_review_palette))
head(review_palette_df)
```

```
## # A tibble: 6 x 2
##   beer_style                           Avg_review_palette
##   <chr>                                             <dbl>
## 1 American Double / Imperial Stout                   4.24
## 2 Quadrupel (Quad)                                   4.15
## 3 Gueuze                                             4.15
## 4 Lambic - Unblended                                 4.12
## 5 Lambic - Fruit                                     4.10
## 6 Russian Imperial Stout                             4.08
```

```
# Plots the results from the table in a graph. Only outputs the top 20
review_palette_df %>% top_n(n=20) %>%
mutate(beer_style =  fct_reorder(beer_style, Avg_review_palette, .desc = TRUE)) %>%
ggplot(aes(beer_style,Avg_review_palette)) + geom_point(stat ="identity" )+
```



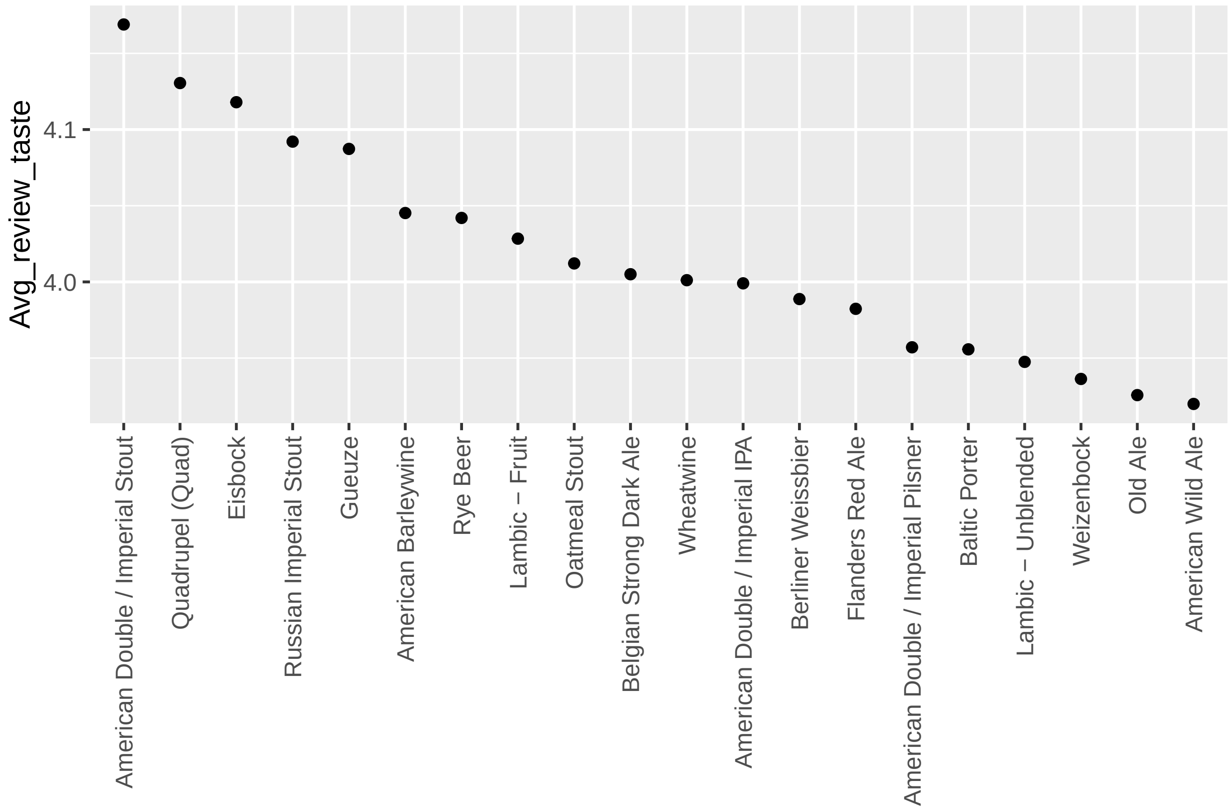
```
# shows the frequency of each review score
review_palette_df %>% ggplot(aes(x=Avg_review_palette)) +
geom_histogram(color="black", fill="white",binwidth=1)
```



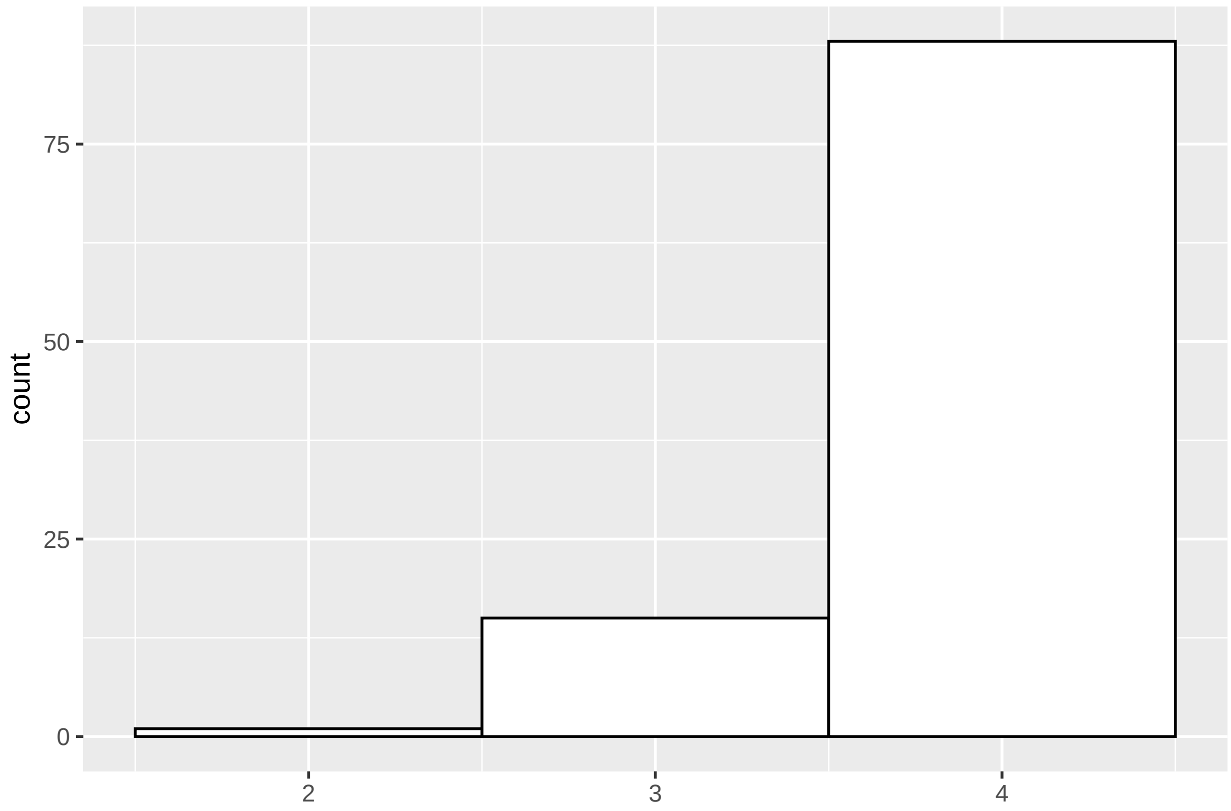
```
# Shows the average review taste by beer style. Outputs only the first 5 rows
review_taste_df <- BP2 %>% group_by(beer_style) %>% summarize(mean(review_taste))
names(review_taste_df)[2] <- "Avg_review_taste"
review_taste_df <- review_taste_df %>% arrange(desc(Avg_review_taste))
head(review_taste_df)
```

```
## # A tibble: 6 x 2
##   beer_style                      Avg_review_taste
##   <chr>                                      <dbl>
## 1 American Double / Imperial Stout            4.17
## 2 Quadrupel (Quad)                            4.13
## 3 Eisbock                                     4.12
## 4 Russian Imperial Stout                      4.09
## 5 Gueuze                                      4.09
## 6 American Barleywine                         4.05
```

```
# Plots the results from the table in a graph. Only outputs the top 20
review_taste_df %>% top_n(n=20) %>%
mutate(beer_style =  fct_reorder(beer_style, Avg_review_taste, .desc = TRUE)) %>%
ggplot(aes(beer_style,Avg_review_taste)) + geom_point(stat ="identity" )+
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



```
# shows the frequency of each review score
review_taste_df %>% ggplot(aes(x=Avg_review_taste)) +
```

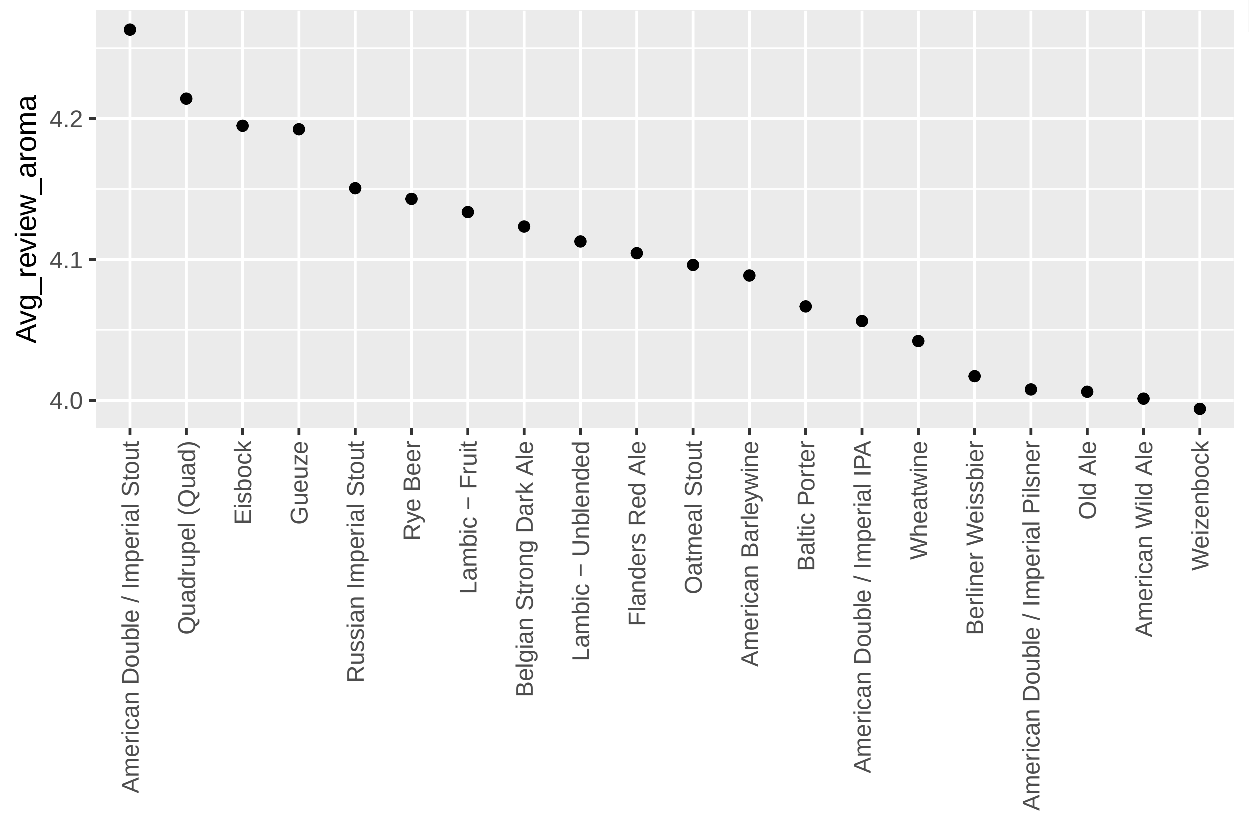


```
# Shows the average aroma taste by beer style. Outputs only the first 5 rows
review_aroma_df <- BP2 %>% group_by(beer_style) %>% summarize(mean(review_aroma))
names(review_aroma_df)[2] <-"Avg_review_aroma"
```

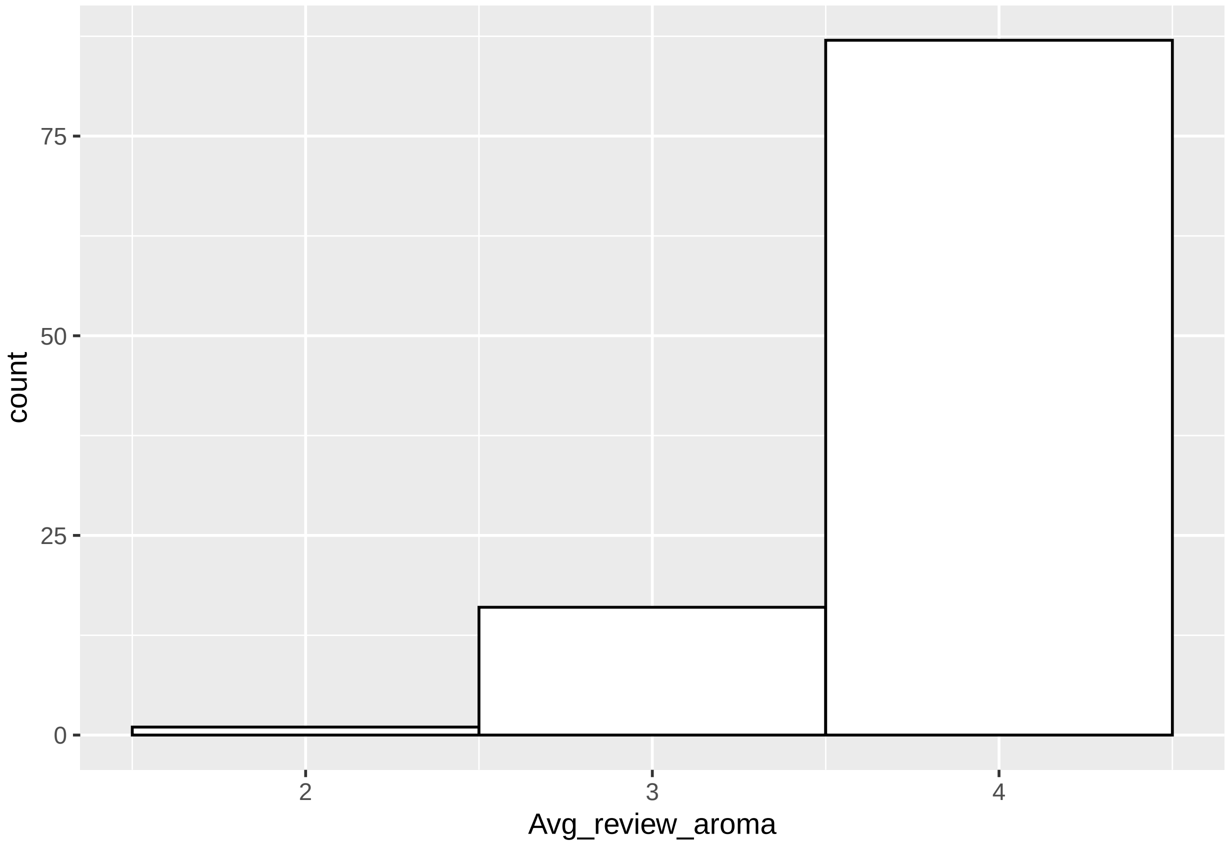
```
review_aroma_df <- review_aroma_df %>% arrange(desc(Avg_review_aroma))
head(review_aroma_df)
```

```
## # A tibble: 6 x 2
##   beer_style                           Avg_review_aroma
##   <chr>                                           <dbl>
## 1 American Double / Imperial Stout                 4.26
## 2 Quadrupel (Quad)                                 4.21
## 3 Eisbock                                          4.19
## 4 Gueuze                                           4.19
## 5 Russian Imperial Stout                           4.15
## 6 Rye Beer                                         4.14
```

```
# Plots the results from the table in a graph. Only outputs the top 20
review_aroma_df %>% top_n(n=20) %>%
mutate(beer_style =  fct_reorder(beer_style,Avg_review_aroma, .desc = TRUE)) %>%
ggplot(aes(beer_style,Avg_review_aroma)) + geom_point(stat ="identity" )+
```



```
# shows the frequency of each review score
review_aroma_df %>% ggplot(aes(x=Avg_review_aroma)) +
geom_histogram(color="black", fill="white",binwidth=1)
```



Correlation of the variables

The following code chunks were used to get the correlation of the variables that were reviewed for the beers. First I made a code that would only get the columns that had the scores regarding the review and put them in a correlation table showing the correlation the variables with each other.

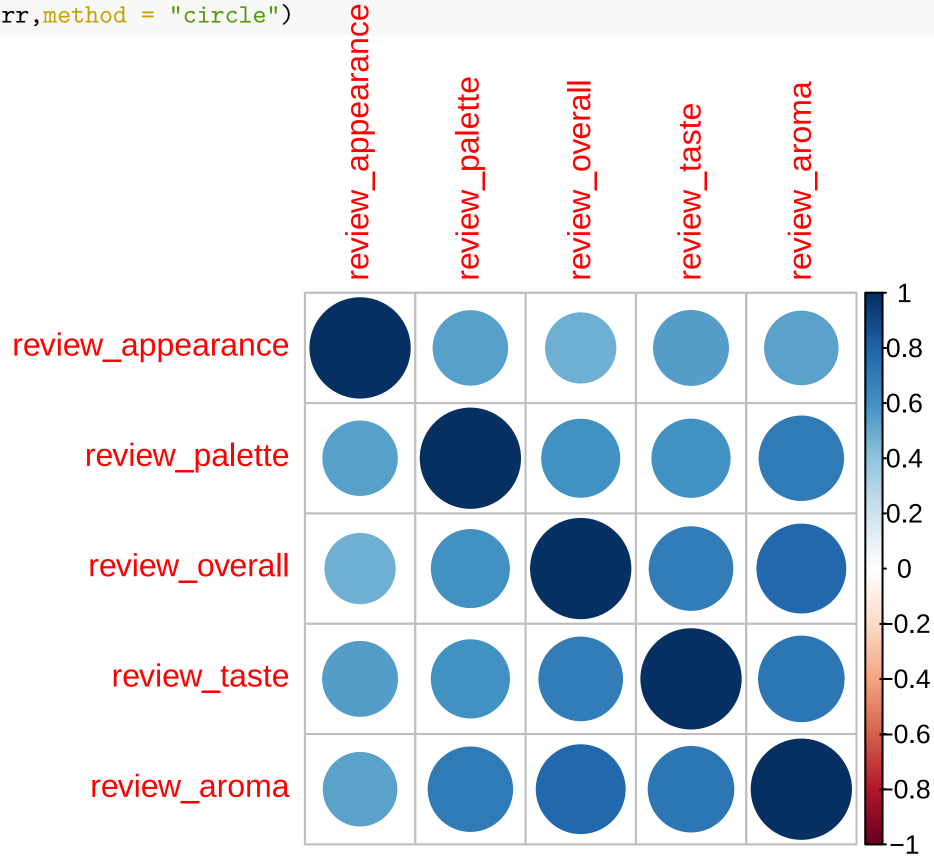
```
cordata = BP2[,c(6,7,8,9,10)]
corr <- cor(cordata)
corr
```

```
##                   review_appearance review_palette review_overall review_taste
## review_appearance         1.0000000      0.5476911      0.4866866    0.5547748
## review_palette            0.5476911      1.0000000      0.6019712    0.6042705
## review_overall            0.4866866      0.6019712      1.0000000    0.6924539
## review_taste              0.5547748      0.6042705      0.6924539    1.0000000
## review_aroma              0.5342441      0.7061559      0.7830024    0.7252735
##                   review_aroma
## review_appearance    0.5342441
## review_palette       0.7061559
## review_overall       0.7830024
## review_taste         0.7252735
## review_aroma         1.0000000
```

Correlation of the variables

Here I made correlation plot of the variables to visualize what the table had.

```
corrplot(corr,method = "circle")
```



Conclusion

With the data that has been collected we can see the highest correlation with the overall review was aroma, taste, and palette. A note to take from the analyzing the data set is that NOT all beers were revewived by the same amount which could have affected the correlation given here.