WineQuality Study(Analysis using R)

Nandini Singh

2022-09-03

About the Project

In this project, we will be looking at physicochemical properties of red wine, we will find out how various factors influence the quality of the wine and also find out if there is any realtionship between the factors present.

DataSet:

The dataset used for this study can be found at : https://docs.google.com/spreadsheets/d/1zfoQsw_t8_EW-muJyJ7YN-NkCAomlt1d/edit?usp=sharing&ouid=106789331917540530199&rtpof=true&sd=true (https://docs.google.com/spreadsheets/d/1zfoQsw_t8_EW-muJyJ7YN-NkCAomlt1d/edit?usp=sharing&ouid=106789331917540530199&rtpof=true&sd=true) . This dataset contains physicochemical properties of red wine and its respective sensory qualities as assessed by wine experts.

Loading the required libraries:

```
library(tidyverse)
```

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

```
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.4
## ✔ tibble  3.1.7     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(ggplot2)
library(tidyr)
```

Importing the Data:

```
winequality <- read.csv("winequality.csv")
```

Now taht we have imported out data, let's take a look at it's structure.

```
str(winequality)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Now that we have looked at the structure, we find out its summary.

```
summary(winequality)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Looking at the summary we find that there are some outliers in the data, where the maximum value of the attribute is much larger than the 75 percentile value.

Cleaning the Data:

Firstly, we need to find out if there are any null entries.

```
sum(is.null(winequality))
```

```
## [1] 0
```

The above result shows that we have no null entries in this dataframe.

Secondly, we need to find out if there are any duplicate entries.

```
sum(duplicated(winequality))
```

```
## [1] 240
```

The above result shows that we have 240 duplicate entries so we need to remove them.

```
winequality <- winequality %>%
  distinct() %>%
  drop_na()
```

Now we verify that the duplicate entries are removed from the dataset.

```
sum(duplicated(winequality))
```

```
## [1] 0
```

The above result shows that we have no duplicate entries so now we can move to cleaning and renaming the column headers. For this, we need to load packages skimr and janitor

```
install.packages("skimr",repos = "http://cran.us.r-project.org")
```

```
## package 'skimr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\NANDINI\AppData\Local\Temp\RtmpWQM4hb\downloaded_packages
```

```
library(skimr)
install.packages("janitor",repos = "http://cran.us.r-project.org")
```

```
## package 'janitor' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\NANDINI\AppData\Local\Temp\RtmpWQM4hb\downloaded_packages
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

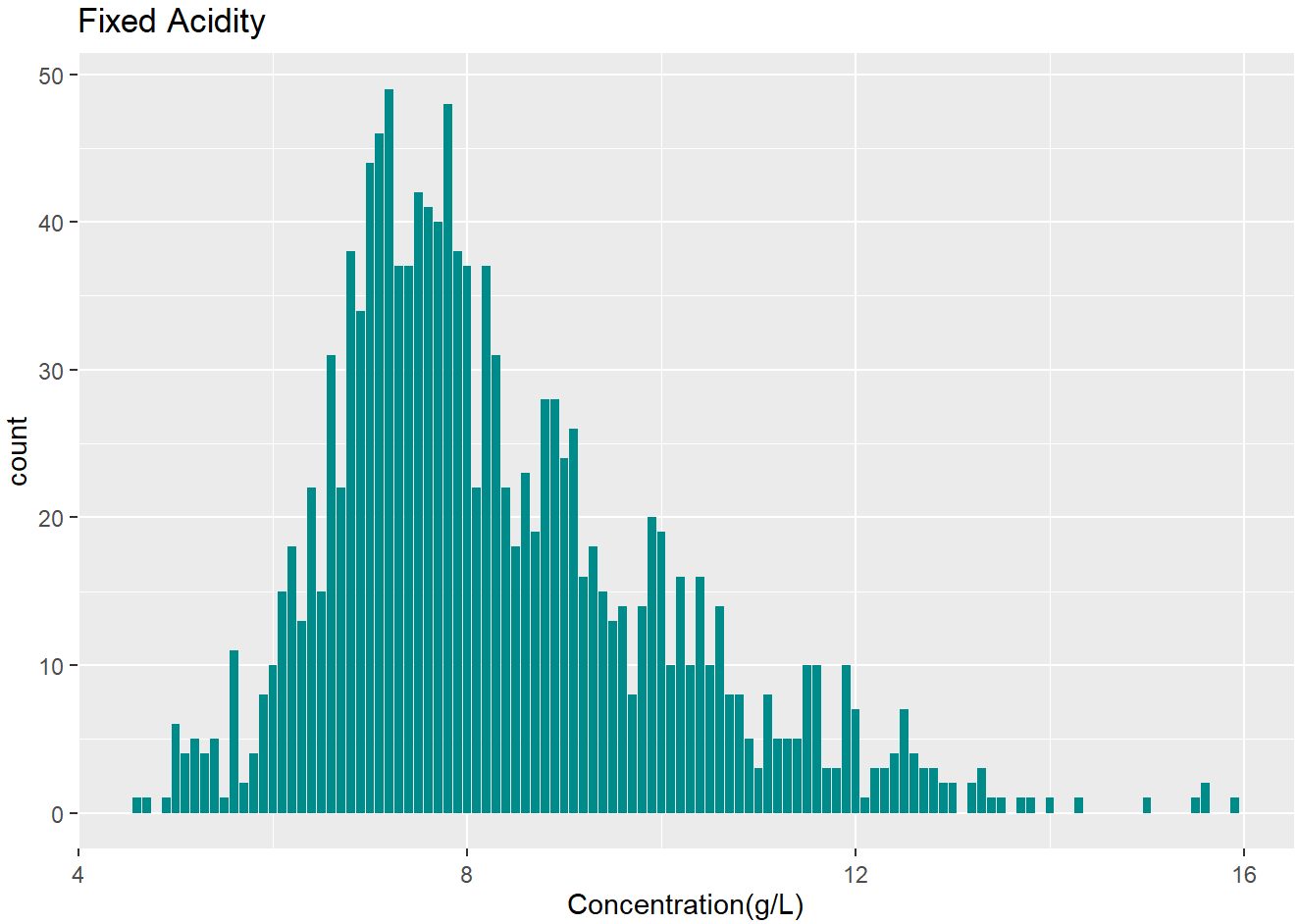
```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Analysis:

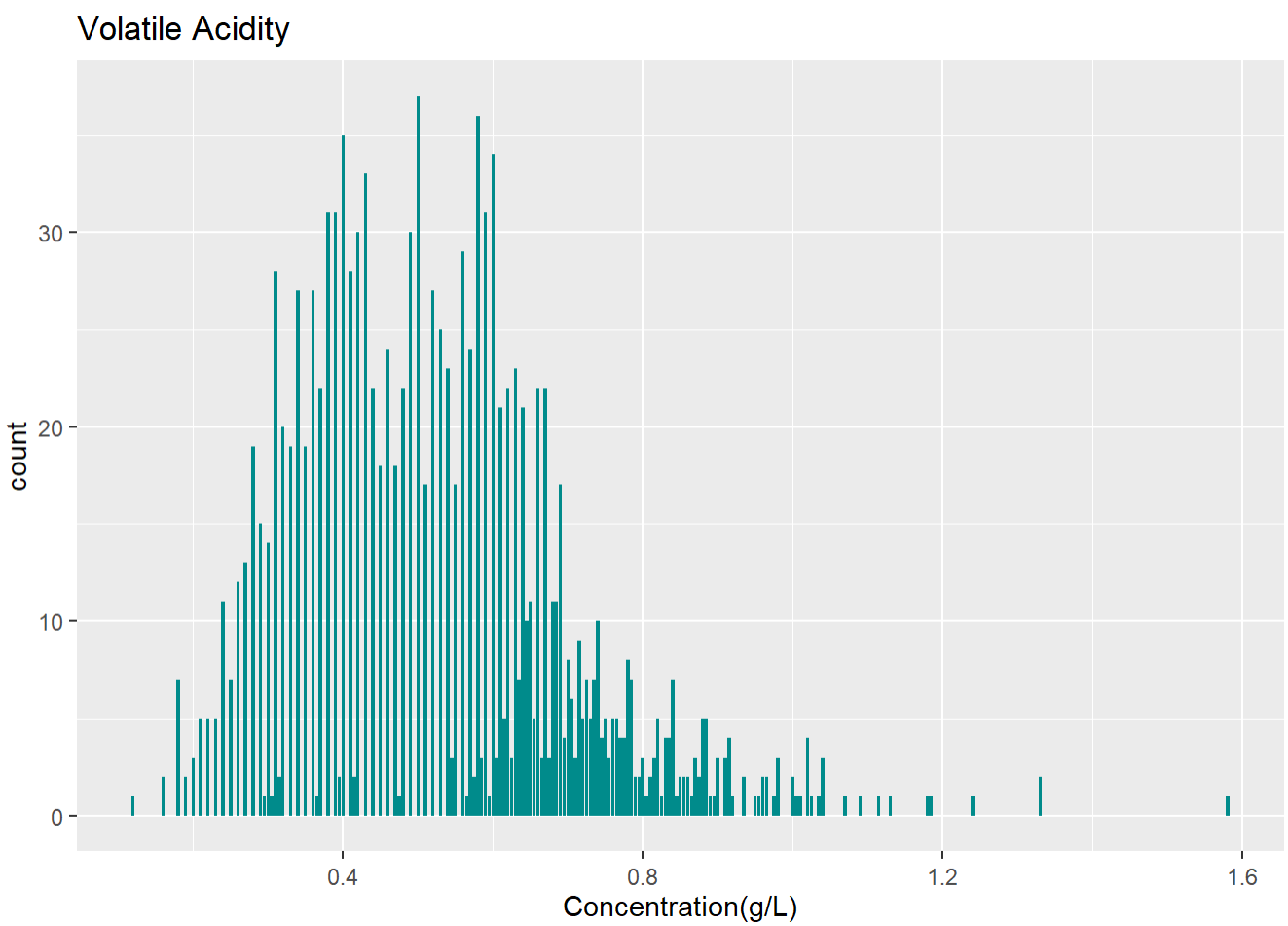
1.ACIDITY

Lets plot the graphs of fixed acidity, volatile acidity and citric acid.

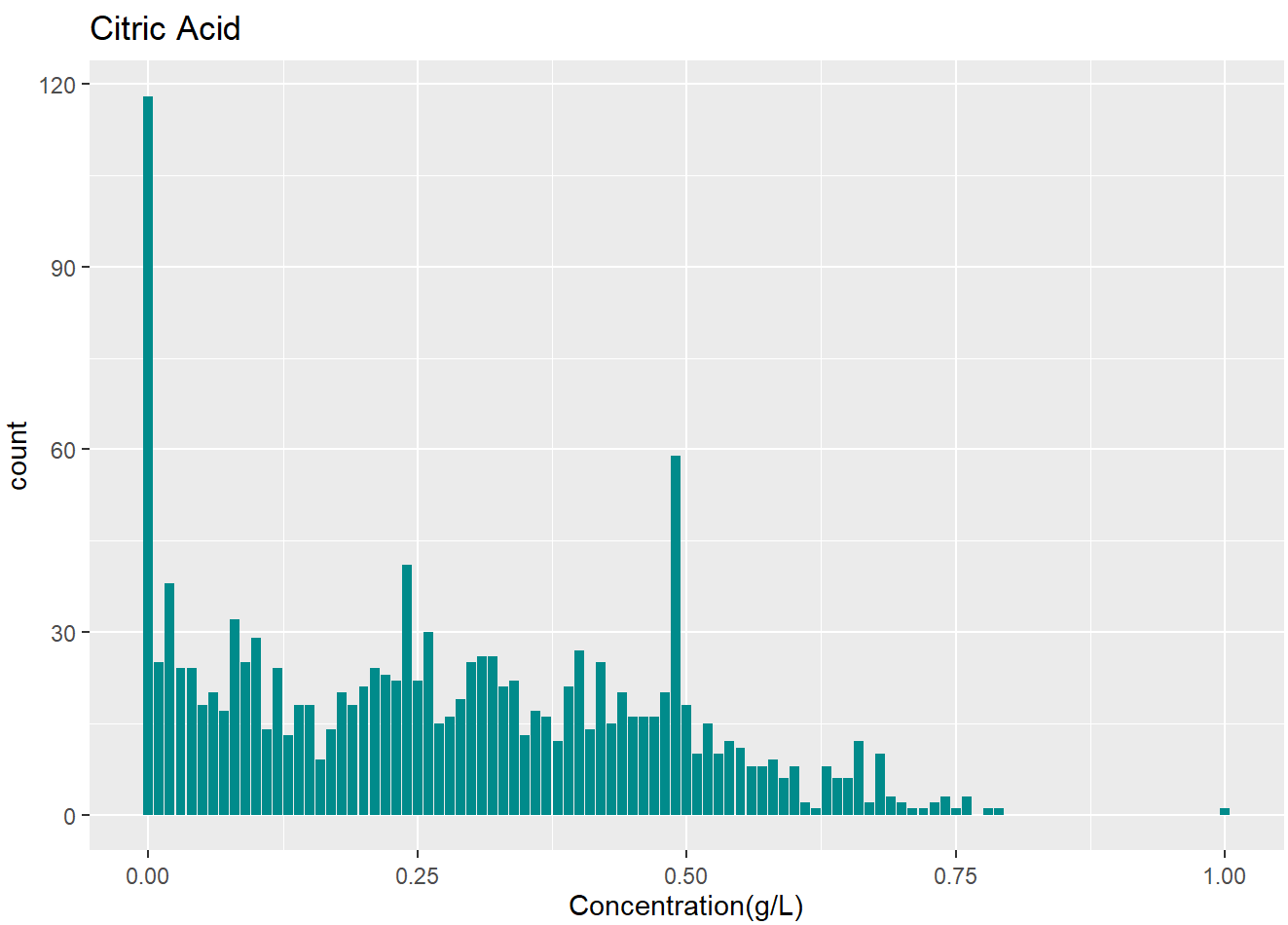
```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = fixed.acidity), fill = "cyan4")+
 labs(title = "Fixed Acidity", x= "Concentration(g/L)", y = "count") +
 theme(axis.text.x = element_text(vjust = 0.5,hjust = 0.5))
```




```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = volatile.acidity), fill = "cyan4")+
 labs(title = "Volatile Acidity", x= "Concentration(g/L)", y = "count") +
 theme(axis.text.x = element_text(vjust = 0.5,hjust = 0.5))
```




```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = citric.acid), fill = "cyan4")+
 labs(title = "Citric Acid", x= "Concentration(g/L)", y = "count") +
 theme(axis.text.x = element_text(vjust = 0.5,hjust = 0.5))
```

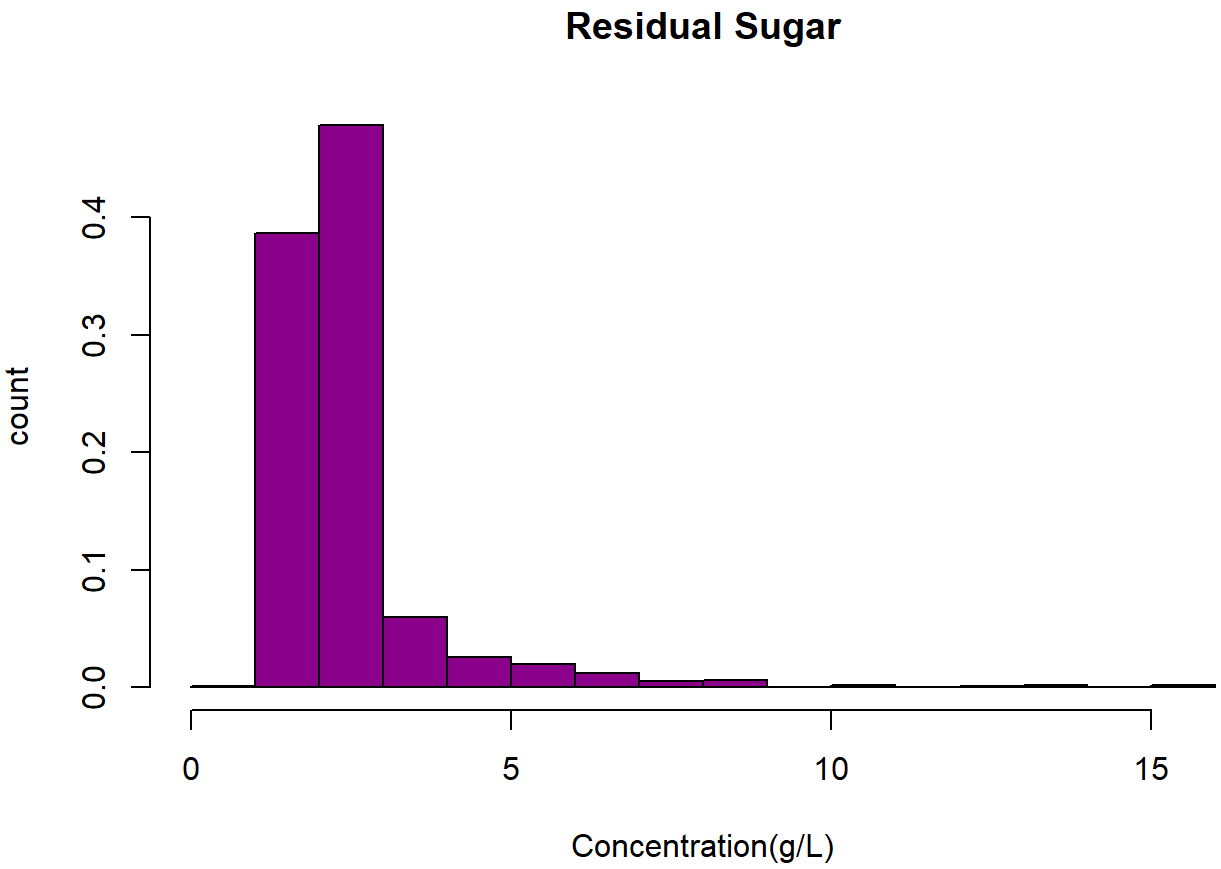


The above three plots show that both Fixed acidity and Volatile acidity are positively skewed distributions but Citric Acid gives an edge peak distribution because a large group of wines seems to have citric acid concentration close to zero. Also we can see there are few outliers present for the three attributes.

2.Residual Sugar

We plot a histogram to find out the amount of residual sugar in wine.

```
hist(winequality$residual.sugar,
main="Residual Sugar",
xlab="Concentration(g/L)",
ylab="count",
col="darkmagenta",
xlim=c(0,16),
freq=FALSE
)
```

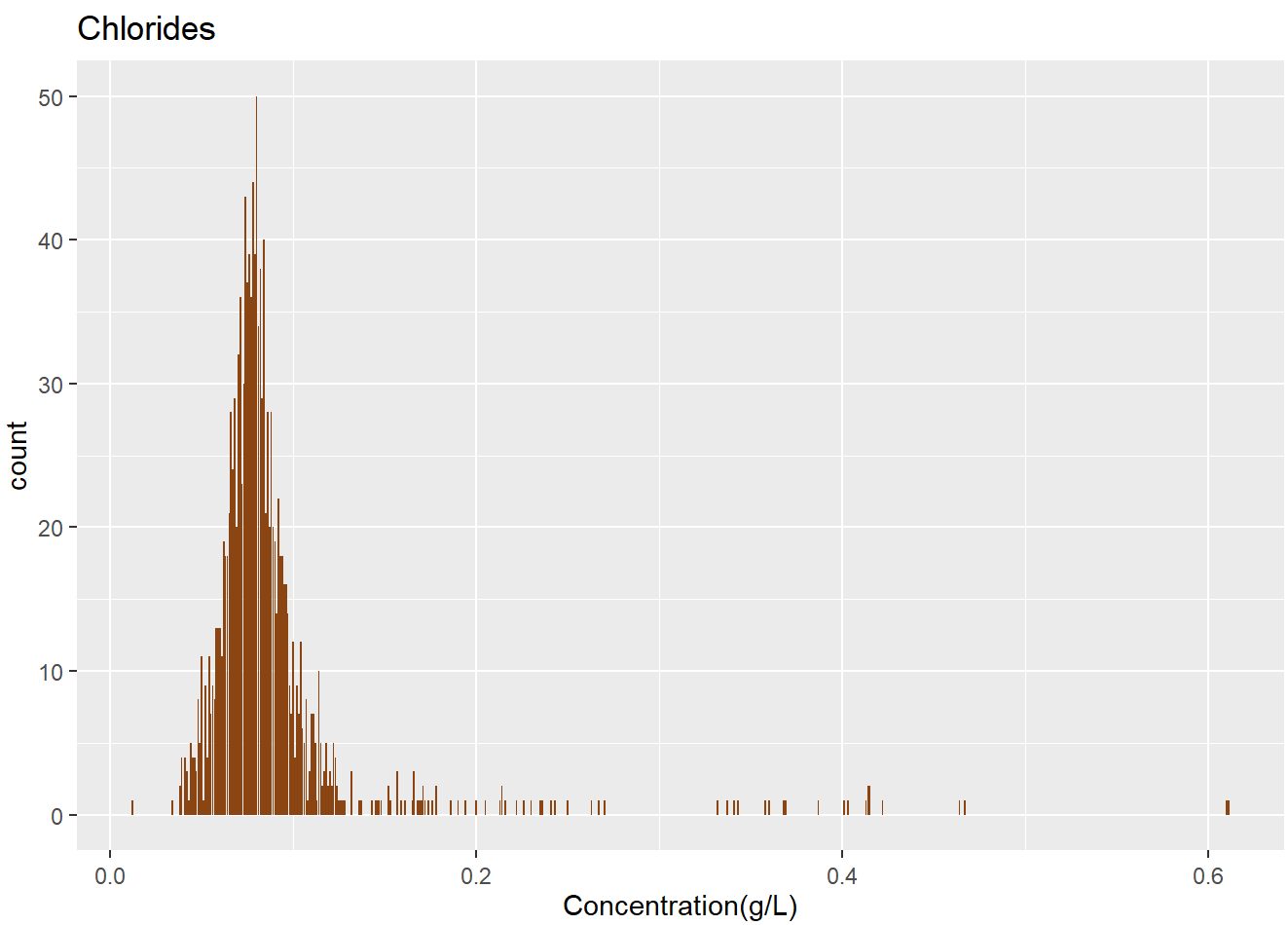


The above histogram shows that mostly wines have a low residual sugar concentration. This also has a positively skewed distribution and also an outlier to the right tail.

3.Chlorine:

Now lets see how is the distribution of chlorine using a plot.

```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = chlorides), fill = "chocolate4")+
 labs(title = "Chlorides", x= "Concentration(g/L)", y = "count") +
 theme(axis.text.x = element_text(vjust = 0.5,hjust = 0.5))
```



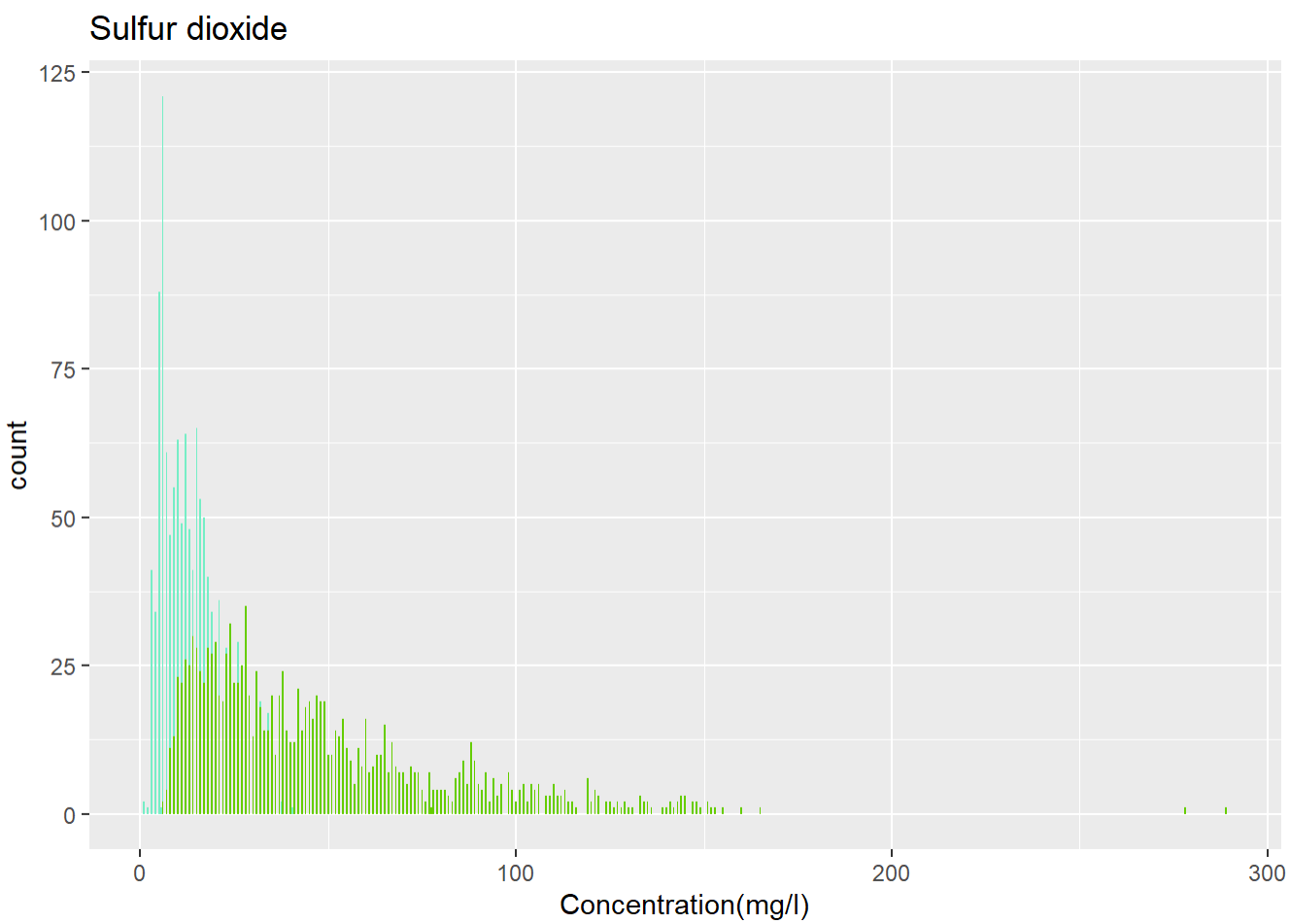
The above plot shows that the most frequent chlorine concentration can be found around 0.04-0.05 g/L.This distribution has a very long right tail with outliers 0.6g/L as we can see from the graph.

4.Sulfur:

In this part of the analysis we will be looking at three parameters, free sulfur dioxide, total sulfur dioxide and sulphates.

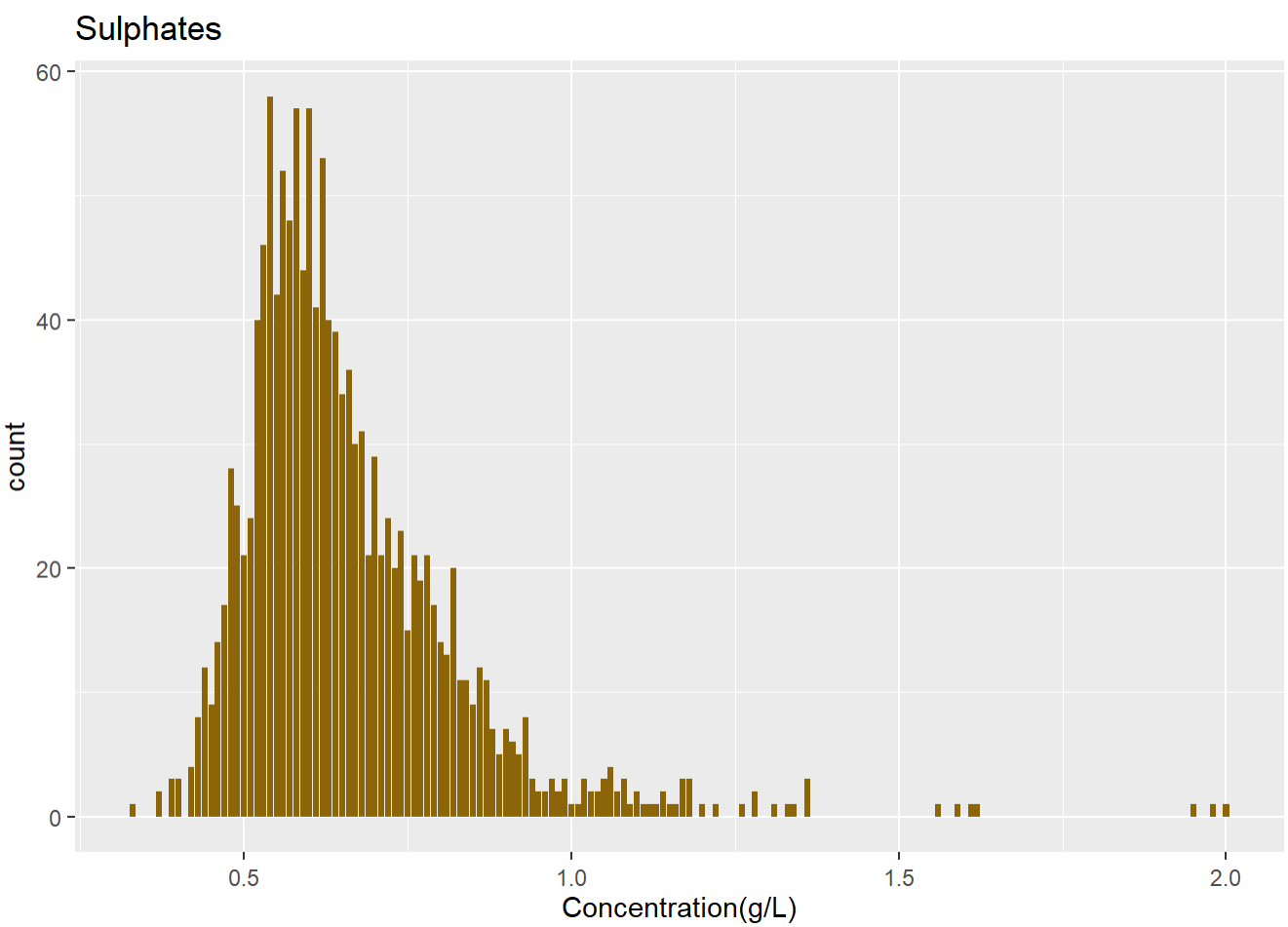
```
sulfur <- c("free.sulfur.dioxide", "total.sulfur.dioxide")
```

```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = free.sulfur.dioxide), fill = "aquamarine2")+
 geom_bar(mapping = aes(x = total.sulfur.dioxide), fill = "chartreuse3")+
labs(title = "Sulfur dioxide", x= "Concentration(mg/l)", y = "count")
```



The above plot shows the distribution of free sulfur dioxide in light blue and the distribution of total sulfur dioxide in green color. From the plot above we understand that about one-fourth of the total sulfur dioxide occurs in the form of free sulfur dioxide. Now let's look at sulphates in the wine:

```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = sulphates), fill = "darkgoldenrod4")+
 labs(title = "Sulphates", x= "Concentration(g/L)", y = "count") +
 theme(axis.text.x = element_text(vjust = 0.5,hjust = 0.5))
```

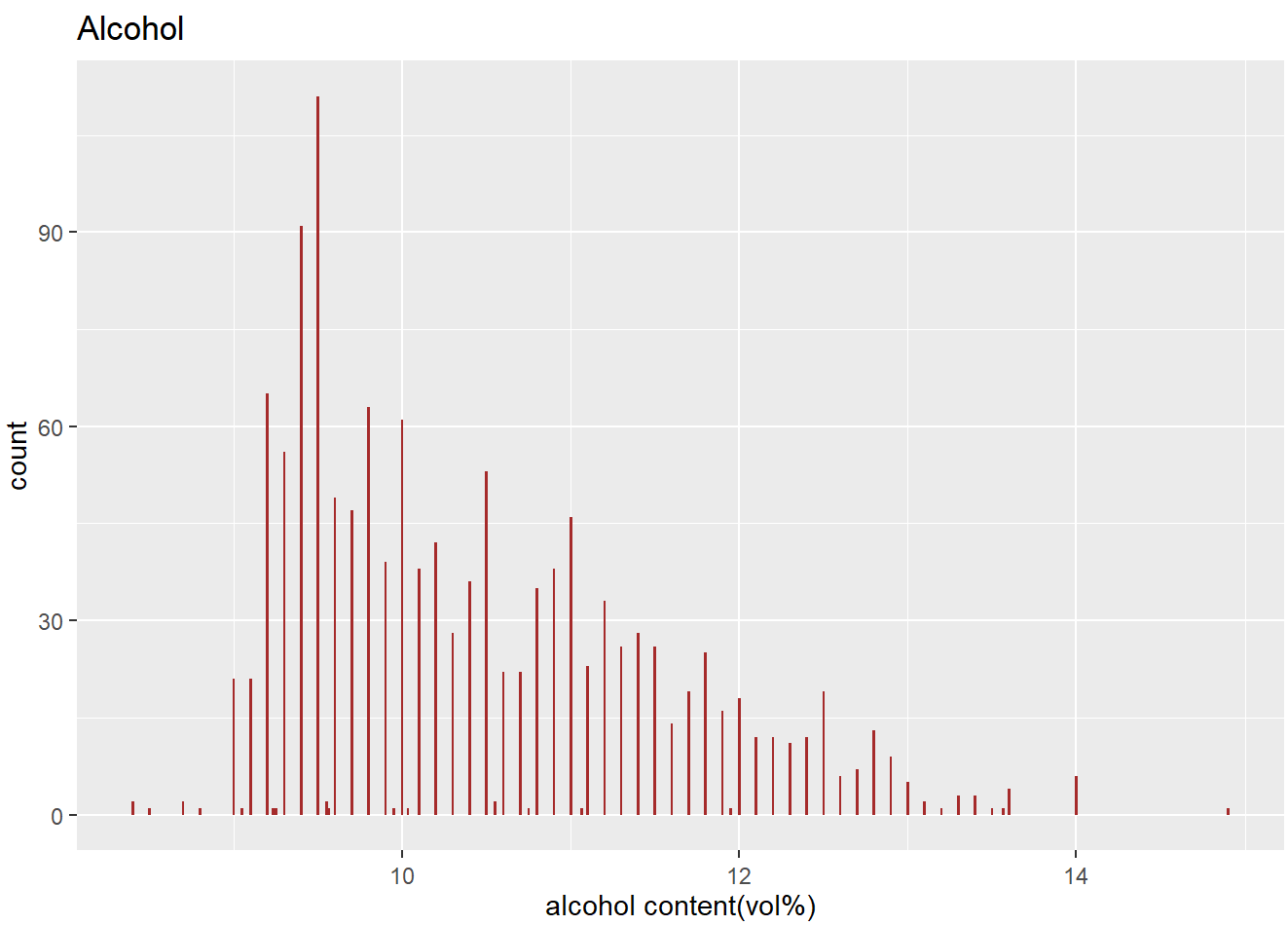


We found that most of the wines have sulfate concentration near 0.5 g/L and that this parameter has two outlier groups one around 1.6-1.7 g/L and one around 1.9-2.0 g/L.

5.Alcohol:

Plotting alcohol content in red wine.

```
ggplot(data = winequality) +
 geom_bar(mapping = aes(x = alcohol), fill = "brown")+
 labs(title = "Alcohol", x= "alcohol content(vol%)", y = "count") +
 theme(axis.text.x = element_text(vjust = 0.5,hjust = 0.5))
```



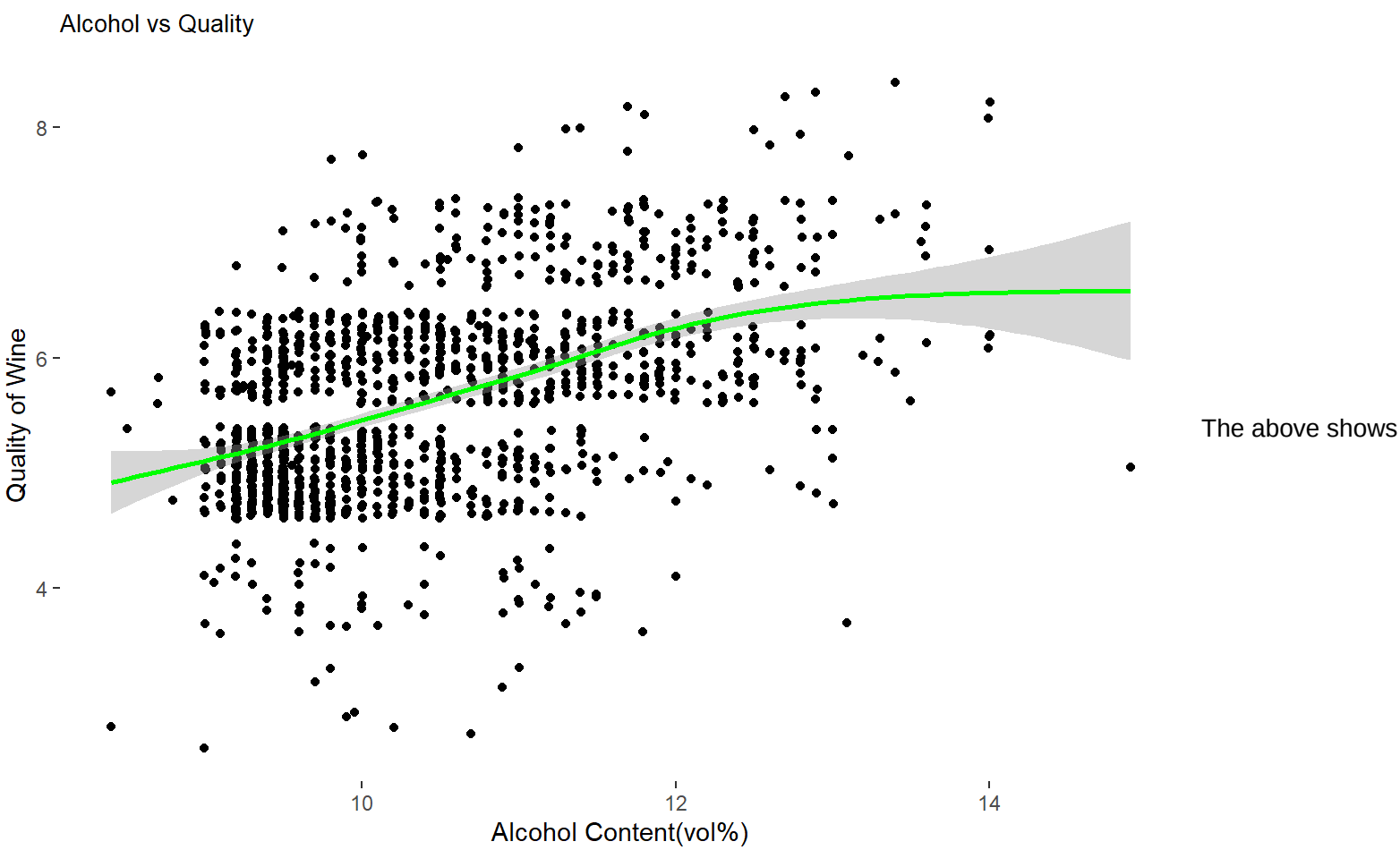
We observe that most of the wines have alcohol content between 8-10 vol%. We also find that alcohol content of wines ranges between 8-15 vol% plus the distribution is positively skewed. In this analysis we will not be doing univariant analysis of density parameter because it shows really less variations throughout the data so its effect of quality determination is minimum.

CORRELATIONS:

6. Alcohol vs Quality

```
ggplot(winequality, aes(x=alcohol, y=quality))+
 geom_jitter() +
 geom_smooth(color = "green") +
 labs(title = "Alcohol vs Quality", x = "Alcohol Content(vol%)", y= "Quality of Wine") +
 theme(panel.background = element_blank(),
 plot.title = element_text( size=10))
```

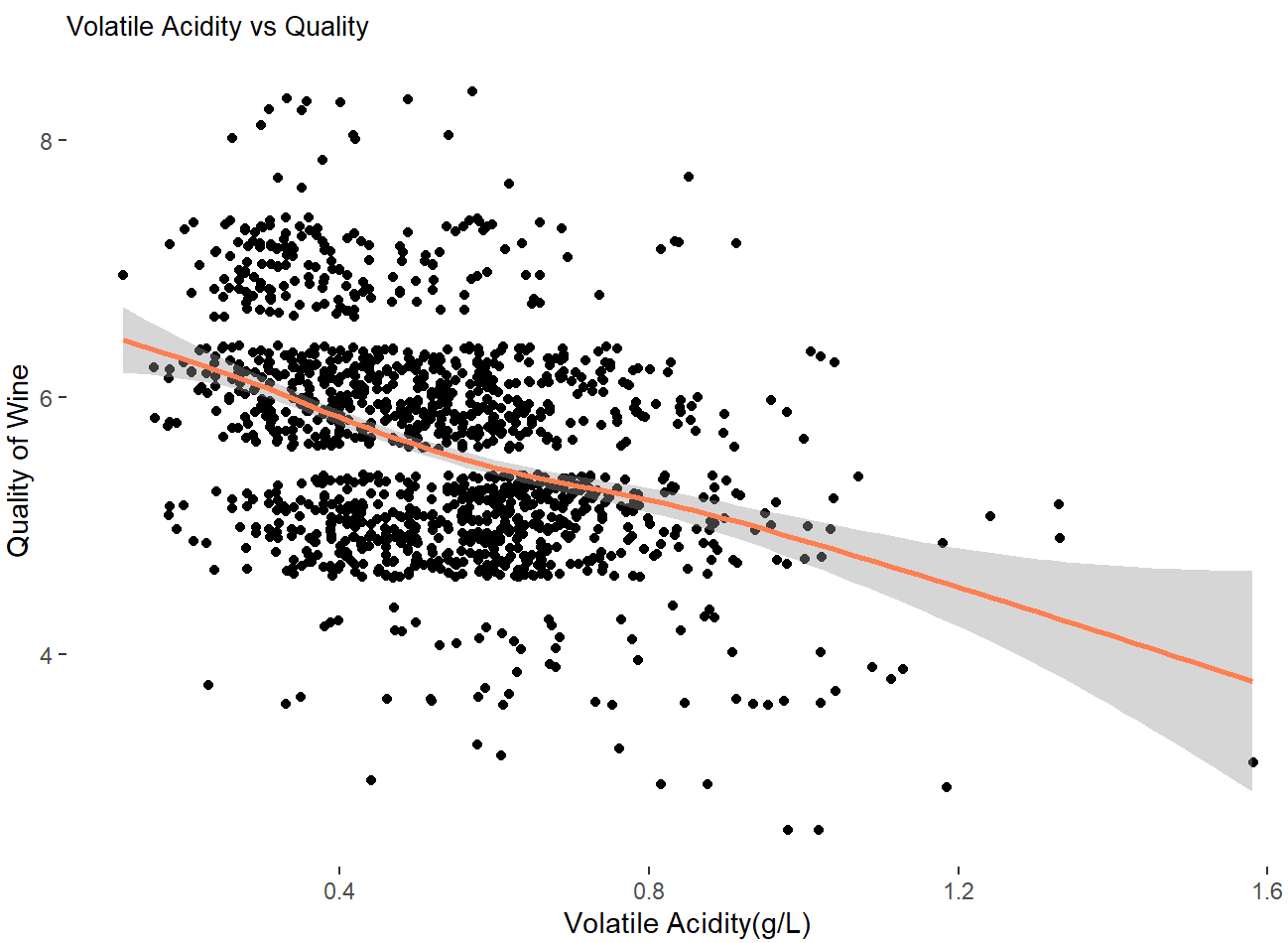
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The above shows that wines with higher alcohol percentage tend to have higher quality score, but the relationship is not very significant as the line in the plot seems very smooth with very less slope. ### 7. Volatile Acidity vs Quality

```
ggplot(winequality, aes(x=volatile.acidity, y=quality))+
 geom_jitter() +
 geom_smooth(color = "coral") +
 labs(title = "Volatile Acidity vs Quality", x = "Volatile Acidity(g/L)", y= "Quality of Wine") +
 theme(panel.background = element_blank(),
 plot.title = element_text( size=10))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

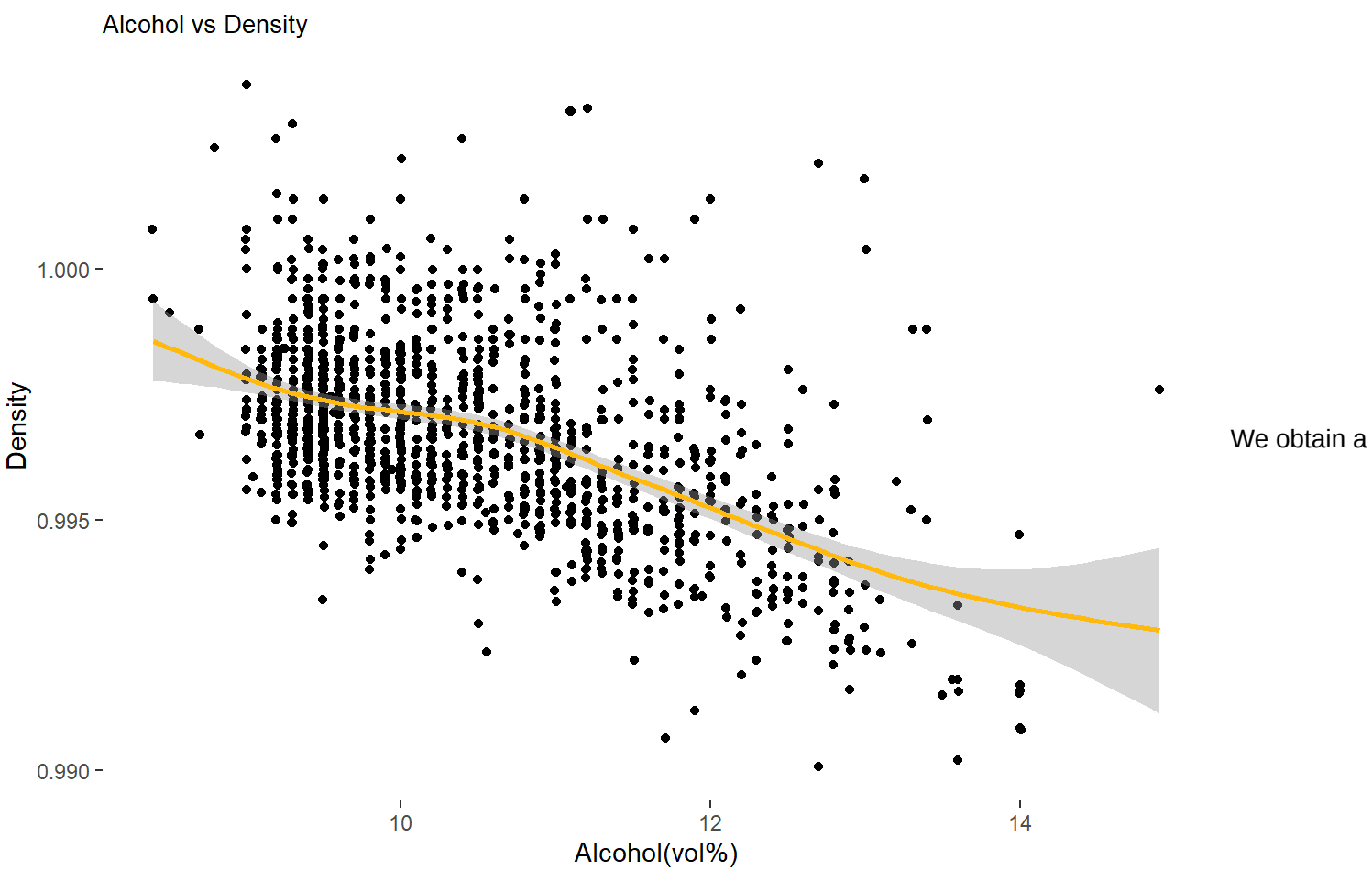


Here we obtain a negative correlation between Volatile Acidity and Quality of the wine, i.e. with increasing volatile Acidity the Quality of wines keeps on decreasing.

8. Alcohol vs Density

```
ggplot(winequality, aes(x=alcohol, y=density))+
 geom_jitter() +
 geom_smooth(color = "darkgoldenrod1") +
 labs(title = "Alcohol vs Density", x = "Alcohol(vol%)", y= "Density") +
 theme(panel.background = element_blank(),
 plot.title = element_text( size=10))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

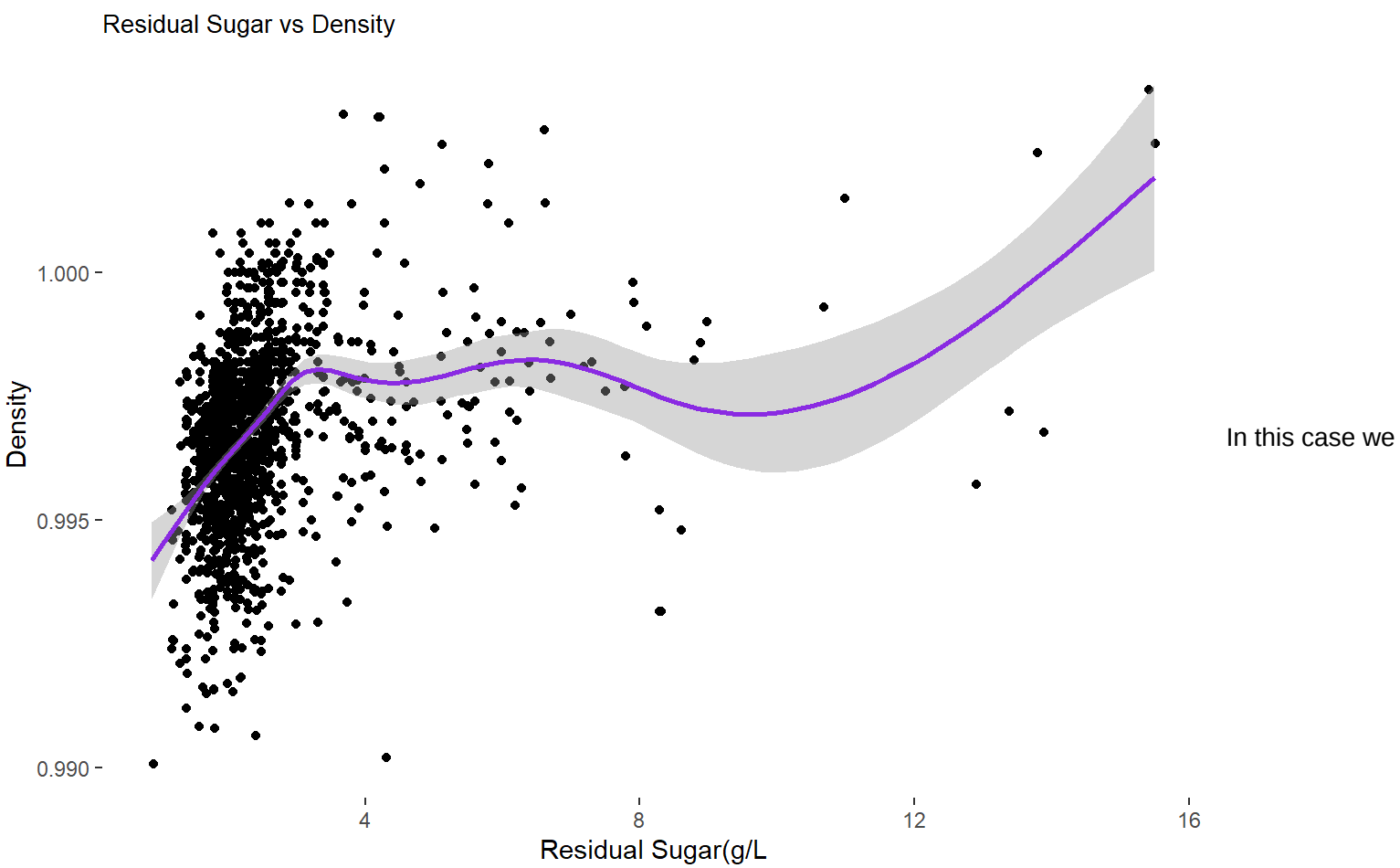


We obtain a negative correlation between alcohol content and density of the wine. With increasing alcohol content, the density of the wine decreases.

9. Residual Sugar vs Density

```
ggplot(winequality, aes(x=residual.sugar, y=density))+
 geom_jitter() +
 geom_smooth(color = "blueviolet") +
 labs(title = "Residual Sugar vs Density", x = "Residual Sugar(g/L", y= "Density") +
 theme(panel.background = element_blank(),
 plot.title = element_text( size=10))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



In this case we obtain a positive correlation between residual sugar and density of the wine, with increasing sugar, the density of the wine also increases.Sugar has a higher density than water and thus increases the density of the mixture while alcohol does the opposite.

INFERENCES:

During this analysis project we found the following things: A. Univariate Analysis: 1. Fixed acidity and Volatile acidity are positively skewed distributions but Citric Acid gives an edge peak distribution. 2. Most wines have a low residual sugar concentration(also positively skewed). 3. The most frequent chlorine concentration can be found around 0.04-0.05 g/L. 4. About one-fourth of the total sulfur dioxide occurs in the form of free sulfur dioxide. 5. Most wines have sulfate concentration near 0.5g/L. 6. Generally, wines have alcohol content between 8-10 vol%.

B.Bivariate Analysis: 1. Wines with higher alcohol percentage tend to have higher quality score. 2. There is a negative correlation between Volatile Acidity and Quality. 3. There is a negative correlation between Alcohol Content and Density. 4. There is a positive correlation between Residual Sugar and Density.

CONCLUSIONS

In this analysis task, we did two type of analysis, univariate and bivariate analysis, in the first we looked at variations in parameters one by one and in the later we looked at how these parameters were affecting each other plus how they determined the wine quality. It is astonishing that wine quality is not that strongly affected by any of the given parameters. We found medium to weak correlations between quality and density, alcohol content, volatile acidity and chloride concentration. In my opinion, we need to find more parameters that significantly influence the quality of wine, because according to this dataset it was more dependent on the personal taste of the person rating the wine. We should explore things like the type of grape used, the time when the wine was made, the taste of the wine based on environmental conditions like heat and humidity.