Project 4 : Explartion of Red Wine Quality

Isabel María Villalba Jiménez

December 6th, 2016

Analysis

In this work it will be analyzed the impact in quality of several parameters describing red wine. The dataset is curated by Udacity and comes from UCI repository https://archive.ics.uci.edu/ml/datasets/Wine+Quality and consists of 1599 sample data for Red wine https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub?embedded=true.

In Cortez et al. (2009) it is shown that the most imporant features for assessing Red Wine quality are:

- sulphates
- pH
- total sulfur dioxide

Variable summary

```
# Load the Data
redwines <- read.csv('wineQualityReds.csv')

dim(redwines)
```

```
## [1] 1599   13
```

```
#names(redwines)
summary(redwines)
```

```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
```

```
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

```
# New variables
#redwines$quality.factor <- factor(redwines$quality)

redwines$quality.cat <- NA
redwines$quality.cat <- ifelse(redwines$quality>=7, 'good','medium')
redwines$quality.cat <- ifelse(redwines$quality<=4, 'bad',redwines$quality.cat) # if not, leave the pre
redwines$quality.cat <- factor(redwines$quality.cat, levels = list('bad', 'medium','good')) # set the o

print("Variables after dividing into quality groups")
```

```
## [1] "Variables after dividing into quality groups"
```

```
str(redwines) #summary of values for each variable
```

```
## 'data.frame':    1599 obs. of  14 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
##  $ quality.cat         : Factor w/ 3 levels "bad","medium",..: 2 2 2 2 2 2 2 3 3 2 ...
```

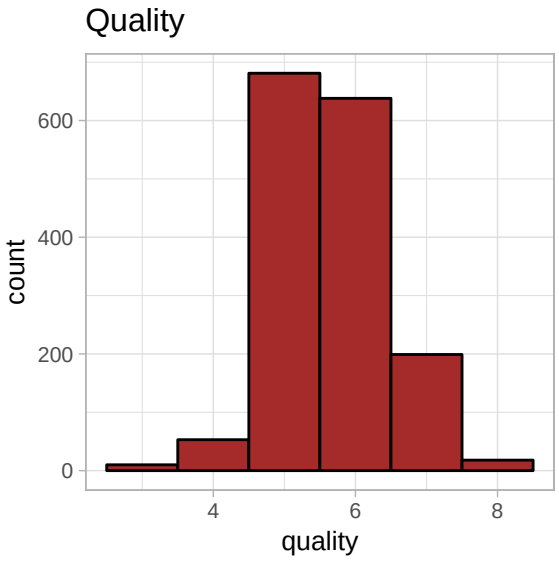
```
#unique(redwines$quality.cat)
```

Univariate Plots Section

In this section it will be analyzed each of the variables describing the wines.

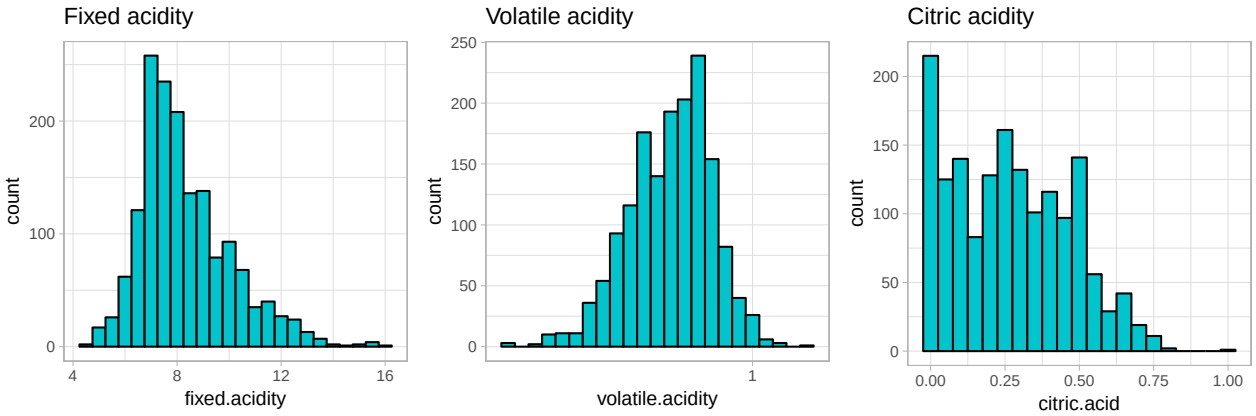
Quality

The distribution of wine shows that most of wines have a quality between 5-6 points.



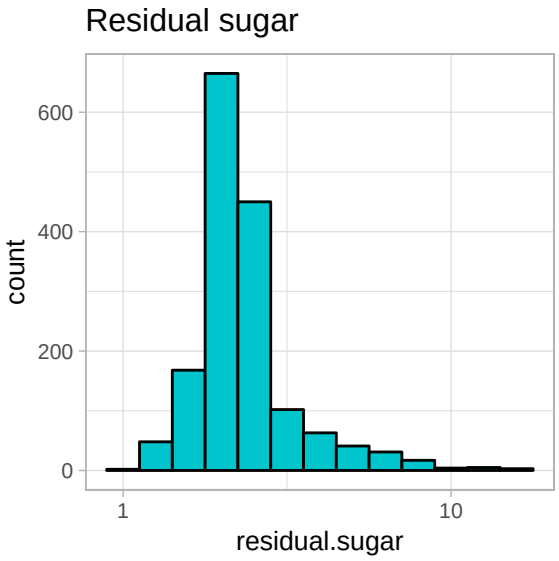
Fixed and volatile acidity

In the next plot it is deduced that the quality of the wine is directly proportional to the fixed acidity and acid levels and inversely proportional to volatile acidity.



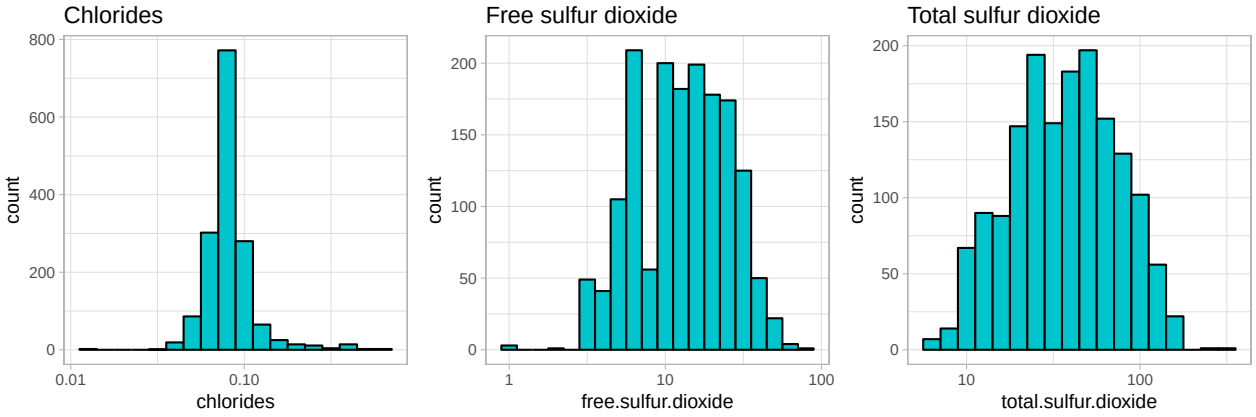
Residual sugar

The plot of residual suggar shows that the better the wine the higher the residual sugar levels.



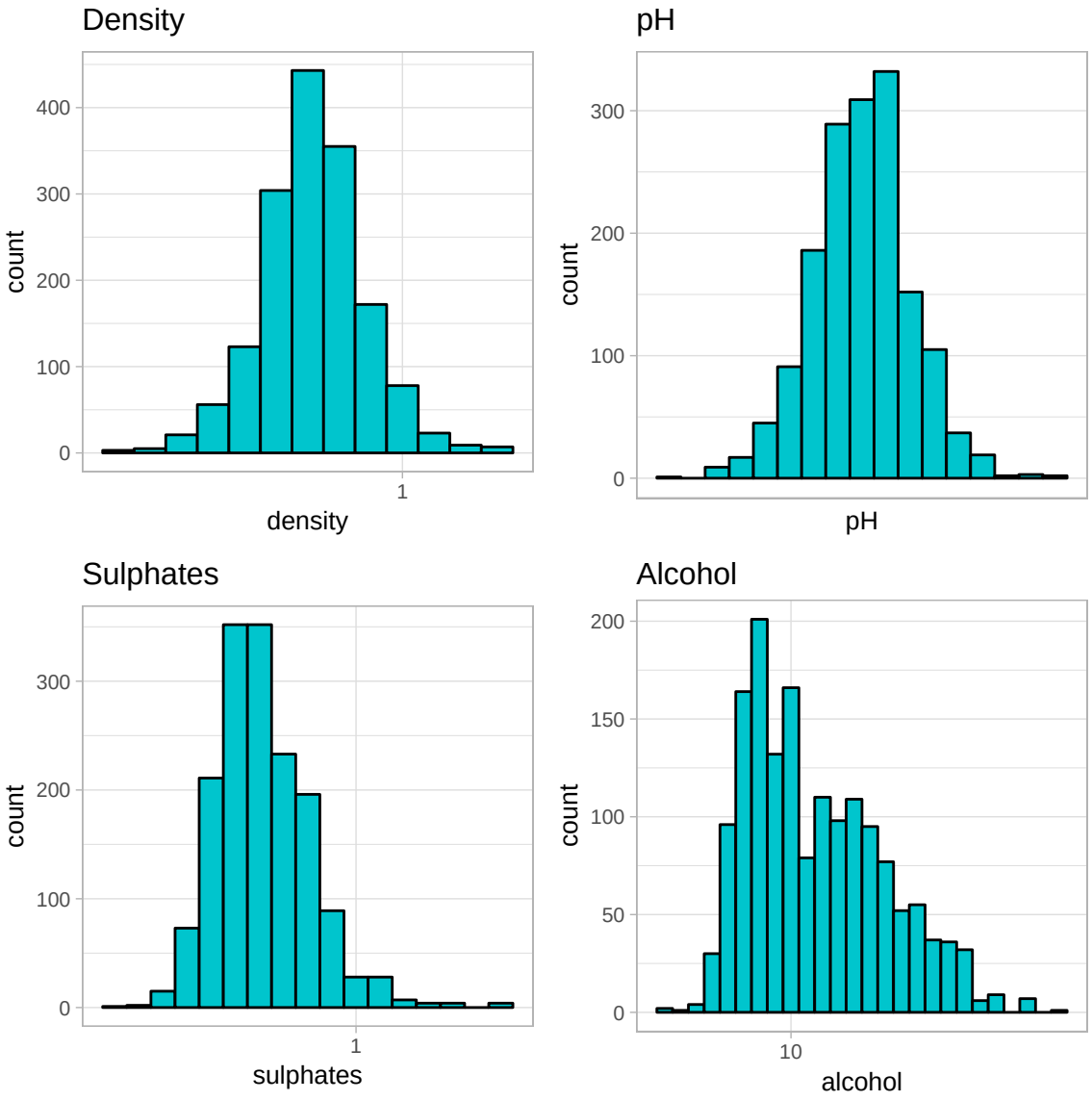
Chlorides and sulfur dioxide

In the plot of the chlorides, it can be onserved that the better the wine, the lower the chloride levels. For the sulfur dioxide, either the free or the total sulfur dioxide, high levels are indicator of medium quality, whereas bad and good wine have the same low amount of sulfure dioxide.



Density, pH, sulphates and alcohol

The next plot shows that, generally, the lower the density, the better the quality of the wine. ALso, low pH levels are sign of beter quality. The higher the sulphates level, the better quality of the wine and also, good wines have higher amount of alcohol.



Univariate Analysis

What is the structure of your dataset?

There are 1599 red wines with 12 features (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol and quality). The variables quality is converted in a factor variable (adding a new variable named `quality.cat`) with the following levels:

(worst) ————————-> (best)

`quality.cat`: BAD (quality [0,4]), MEDIUM (quality (4,7)), GOOD (quality [7,10]),

Other observations:

The mean quality of the red wines is 5.636 and the median is 6. Q1 corresponds to 5 and Q3 to 6, hence, 50% of the data lays within 5-6 range of quality, this is the level MEDIUM.

What is/are the main feature(s) of interest in your dataset?

The main feature of interest in this dataset is the **quality** of the wine.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

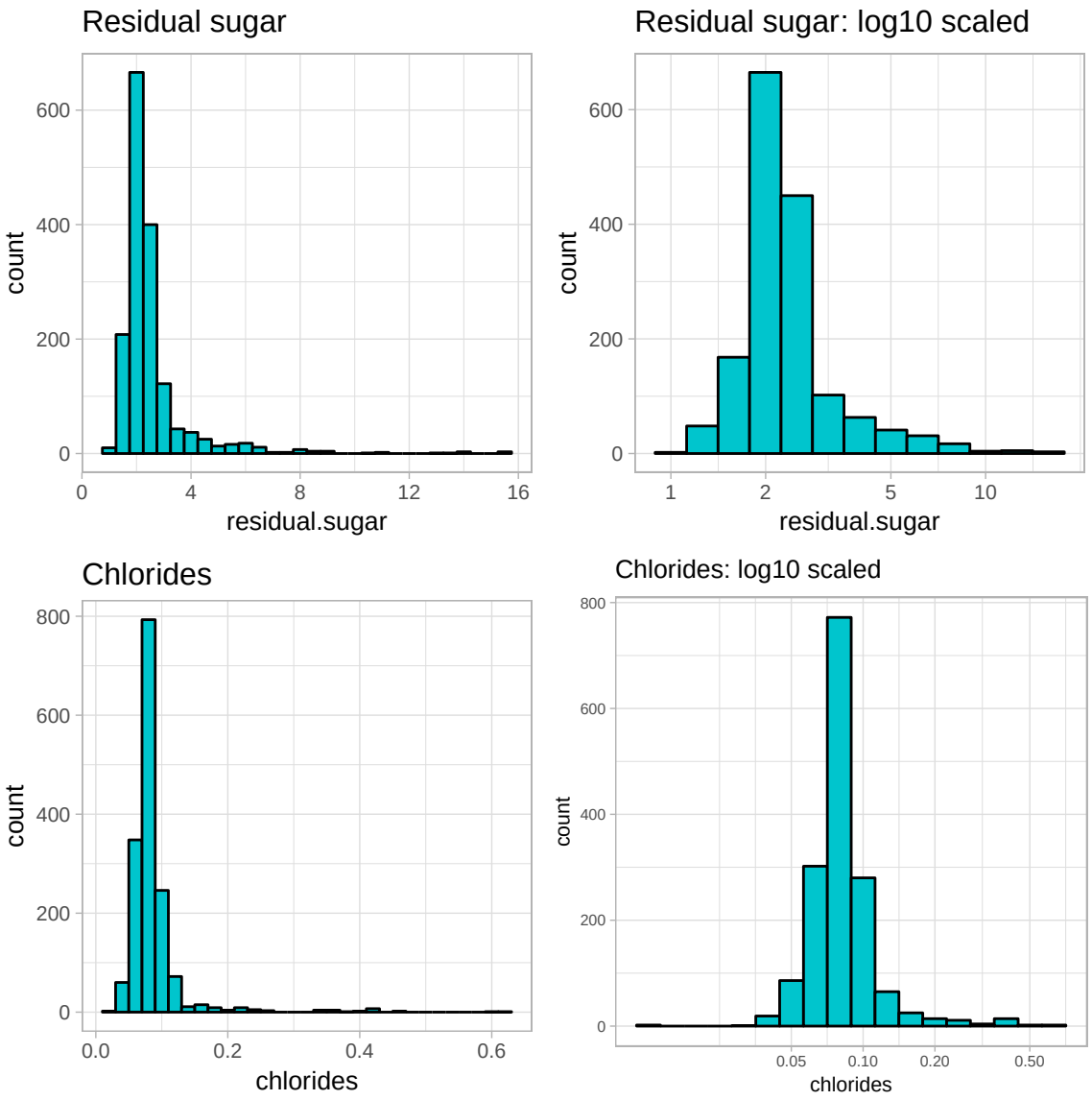
In Cortez et al. (2009) it is shown that the most imporant features for assessing Red Wine quality are: **sulphates**, **pH** and **total sulfur dioxide**.

Did you create any new variables from existing variables in the dataset?

The amount of information available is enough to assess the quality of the wine and I did not create any new variables to support the analysis.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

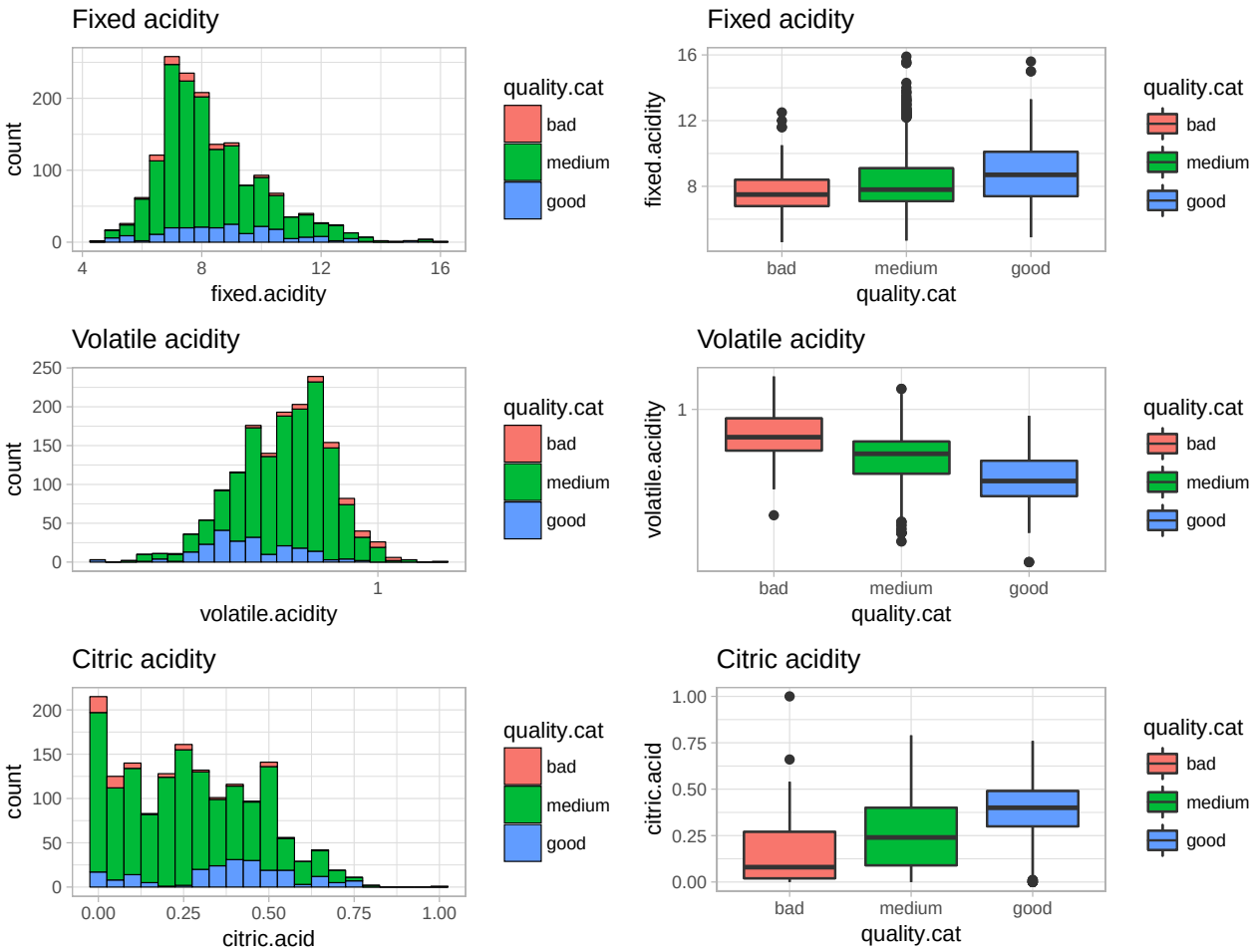
Most of the features had a normal distribution. Some of the features had quite skewed distributions and many oultiers and I performed a log10 tranformation in order to have a better view.



Bivariate Plots Section

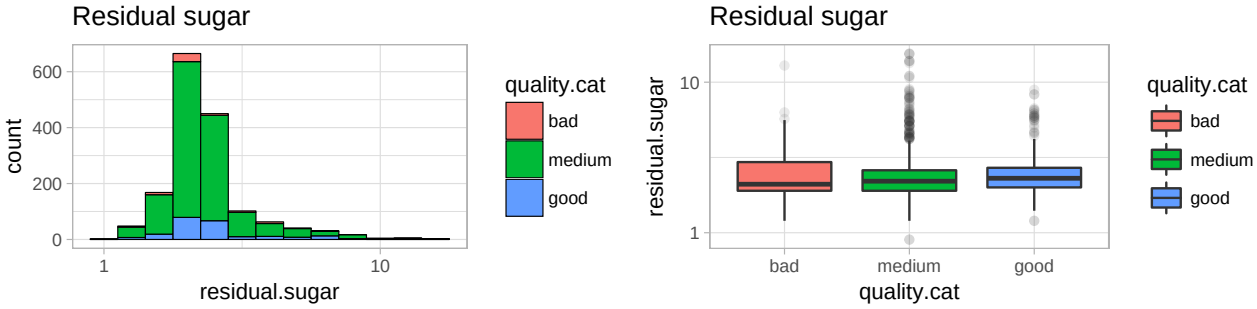
Fixed and volatile acidity

In the next plot it is deduced that the quality of the wine is directly proportional to the fixed acidity and acid levels and inversely proportional to volatile acidity.



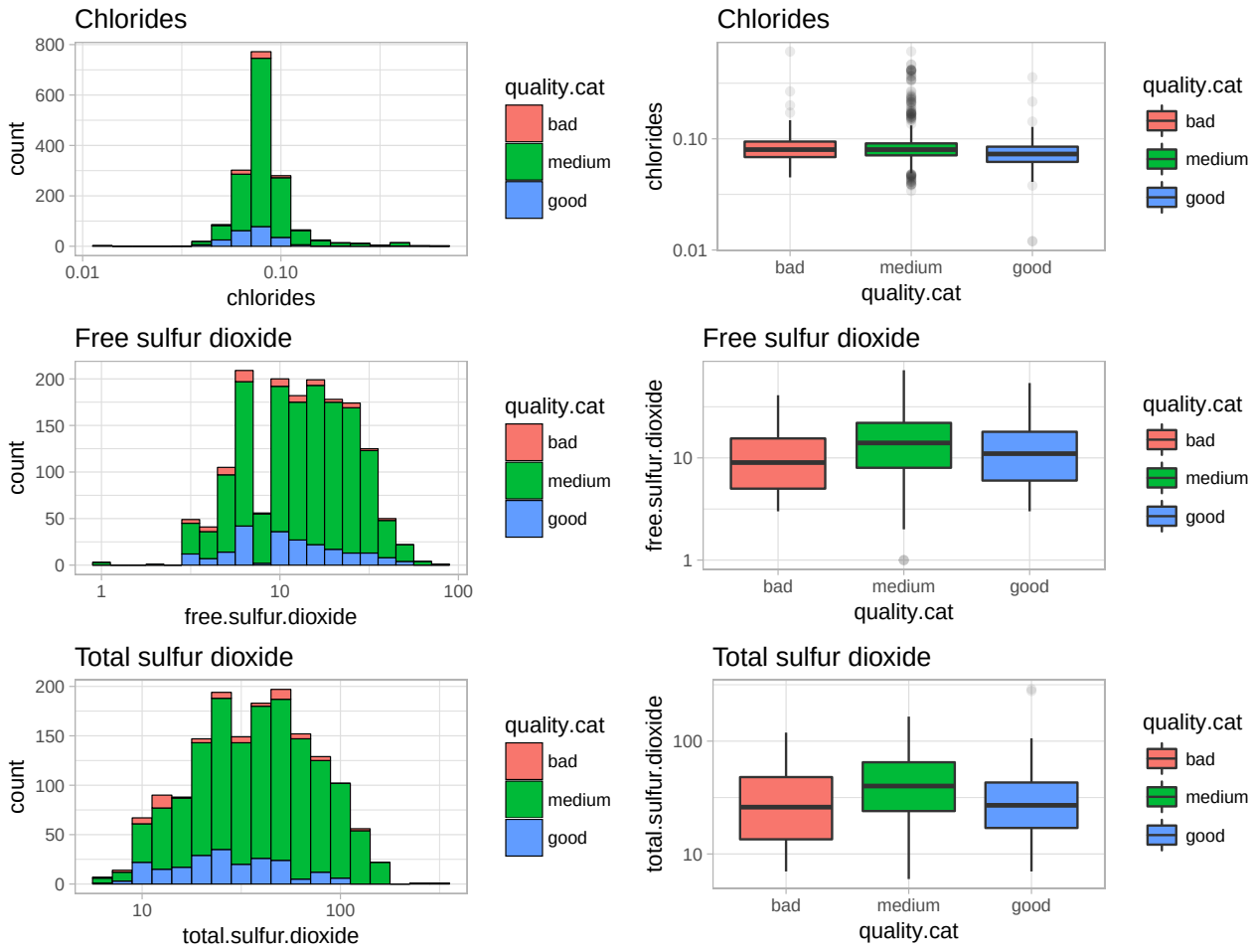
Residual sugar

The plot of residual suggar shows that the better the wine the higher the residual sugar levels.



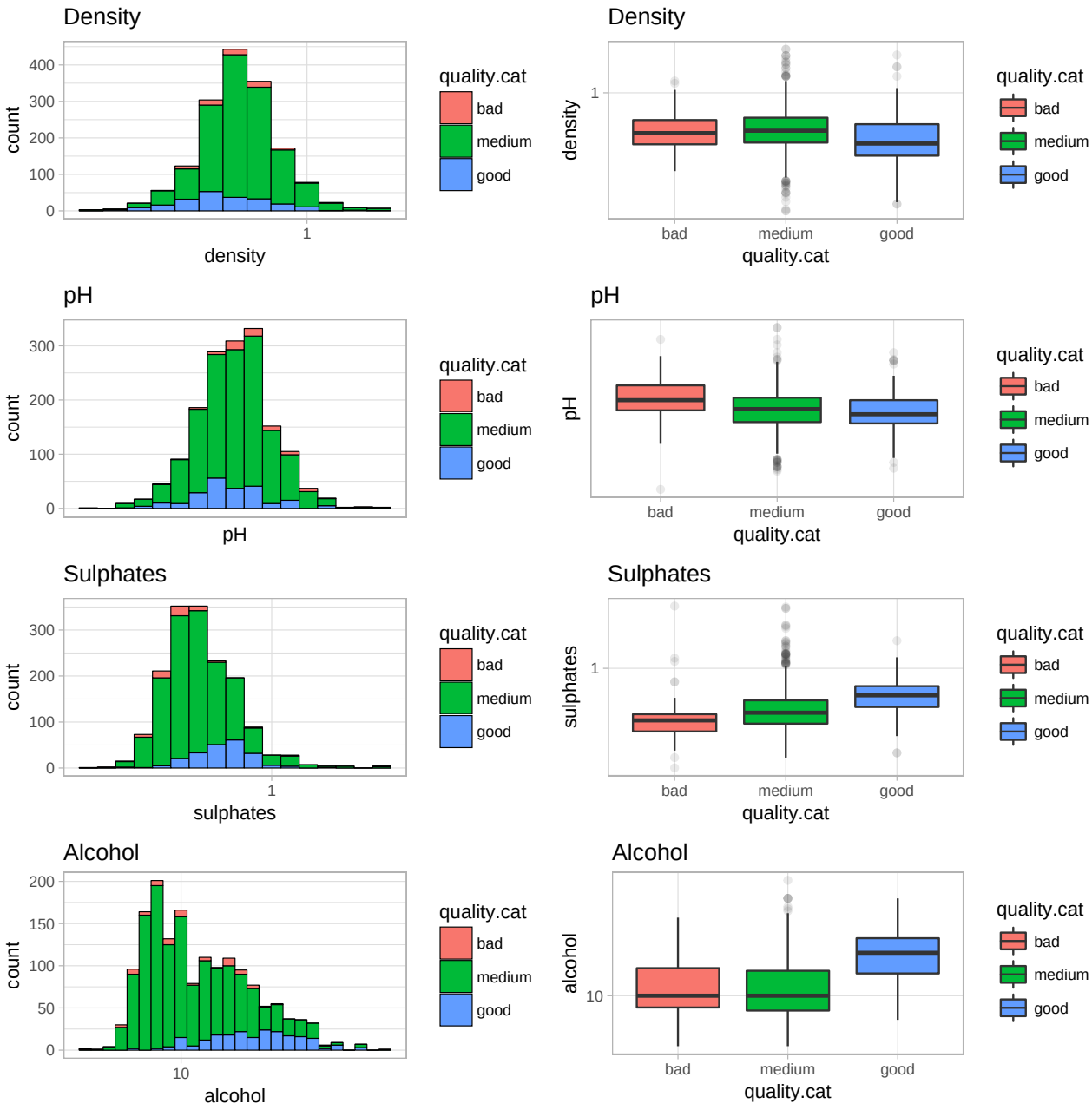
Chlorides and sulfur dioxide

In the plot of the chlorides, it can be onserved that the better the wine, the lower the chloride levels. For the sulfur dioxide, either the free or the total sulfur dioxide, high levels are indicator of medium quality, whereas bad and good wine have the same low amount of sulfure dioxide.



Density, pH, sulphates and alcohol

The next plot shows that, generally, the lower the density, the better the quality of the wine. ALso, low pH levels are sign of beter quality. The higher the sulphates level, the better quality of the wine and also, good wines have higher amount of alcohol.


Density
count
density
quality.cat
bad
medium
good
Density
density
quality.cat
pH
count
pH
quality.cat
pH
Sulphates
count
sulphates
quality.cat
Sulphates
sulphates
quality.cat
Alcohol
count
alcohol
quality.cat
Alcohol
alcohol
quality.cat

Bivariate Analysis

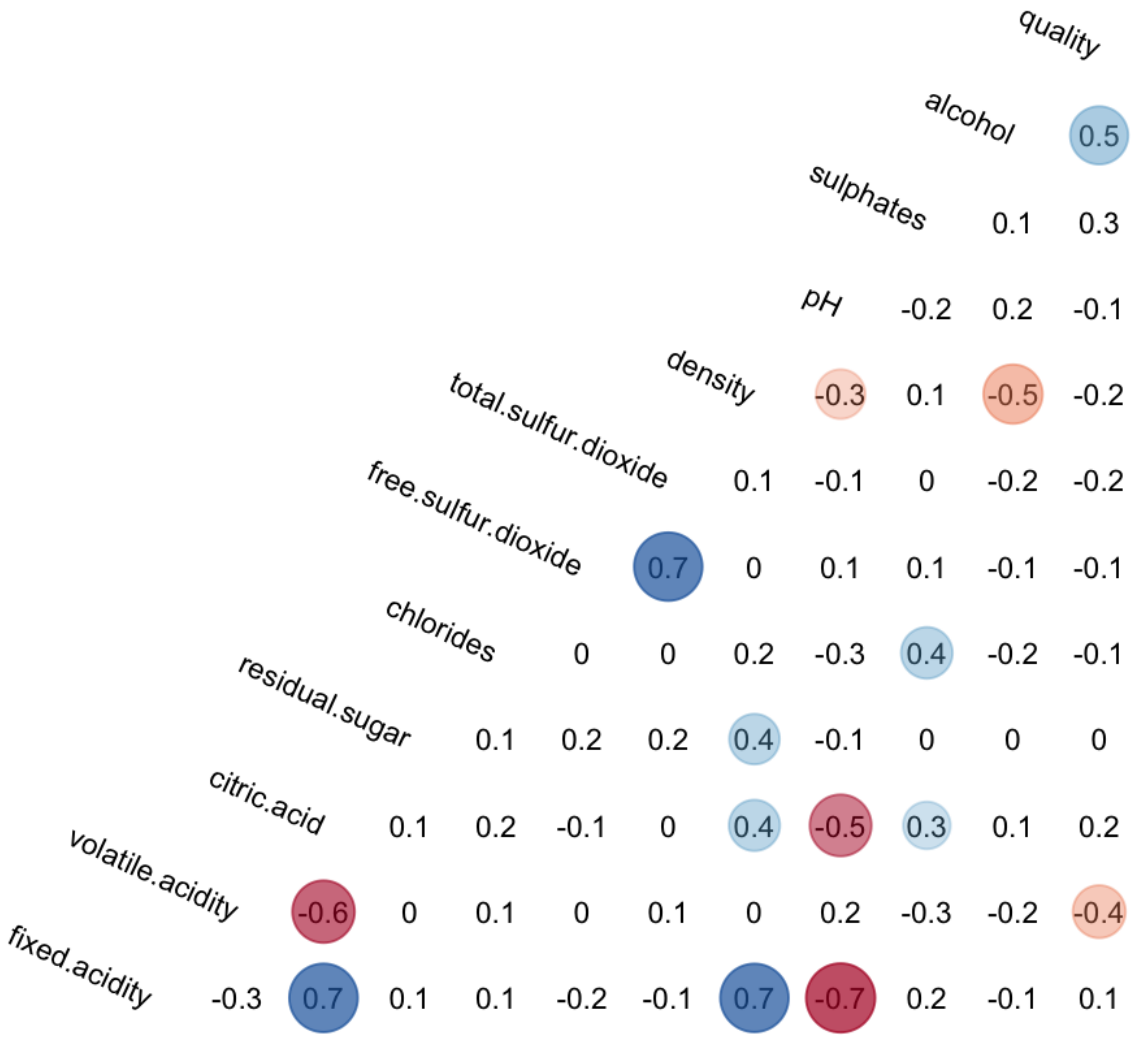
Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

What was the strongest relationship you found?

Multivariate Plots Section

Correlation betweeen variables


quality
alcohol
sulphates
pH
density
total.sulfur.dioxide
free.sulfur.dioxide
chlorides
residual.sugar
citric.acid
volatile.acidity
fixed.acidity
0.5
0.1 0.3
-0.2 0.2 -0.1
-0.3 0.1 -0.5 -0.2
0.1 -0.1 0 -0.2 -0.2
0.7 0 0.1 0.1 -0.1 -0.1
0 0 0.2 -0.3 0.4 -0.2 -0.1
0.1 0.2 0.2 0.4 -0.1 0 0 0
0.1 0.2 -0.1 0 0.4 -0.5 0.3 0.1 0.2
-0.6 0 0.1 0 0.1 0 0.2 -0.3 -0.2 -0.4
-0.3 0.7 0.1 0.1 -0.2 -0.1 0.7 -0.7 0.2 -0.1 0.1

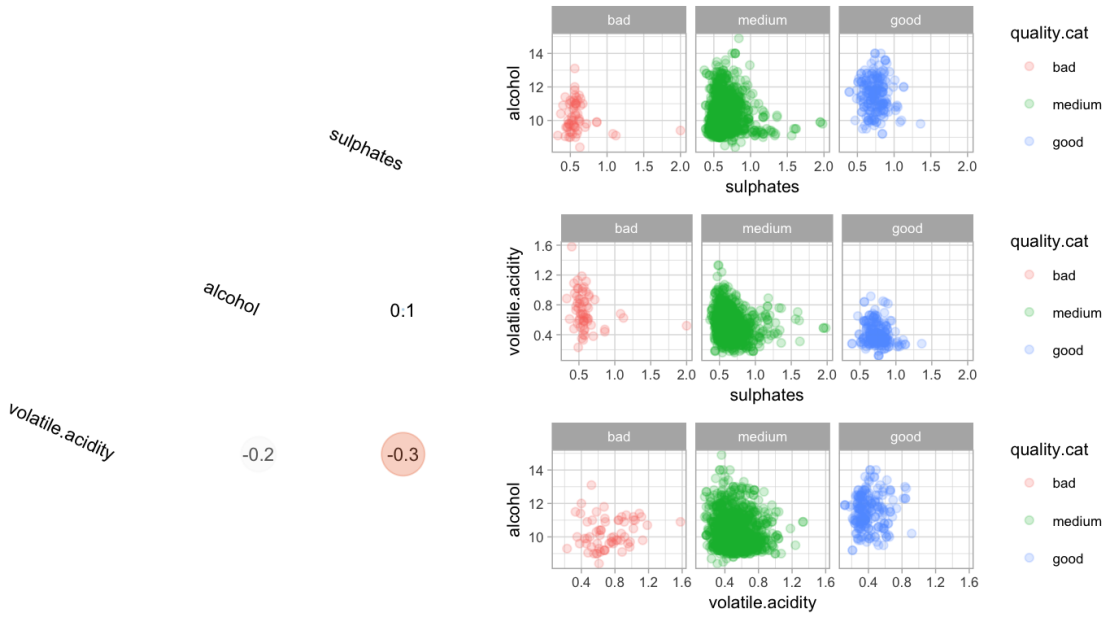
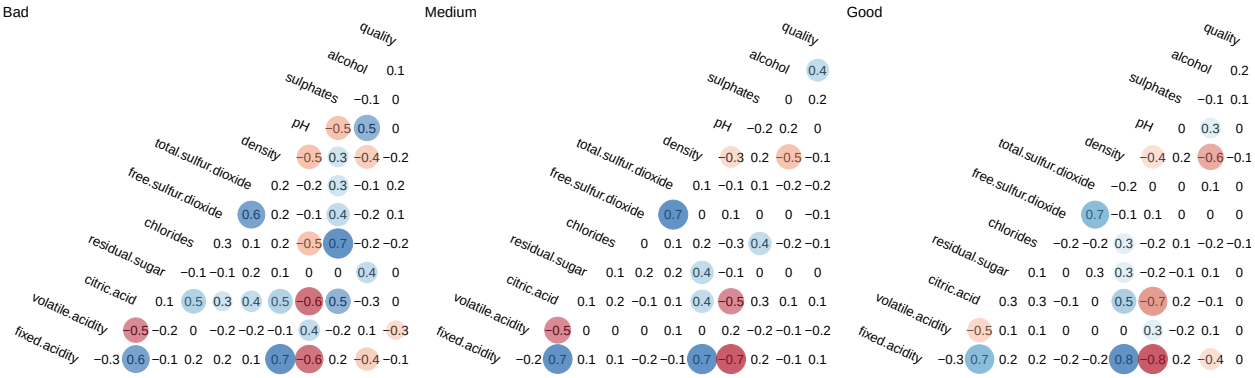
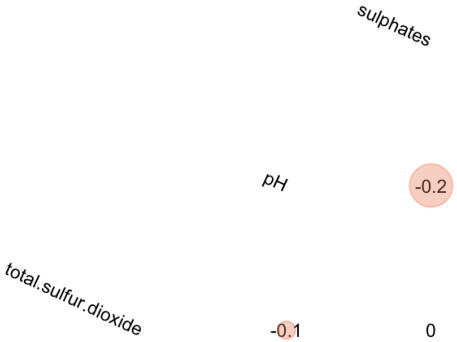


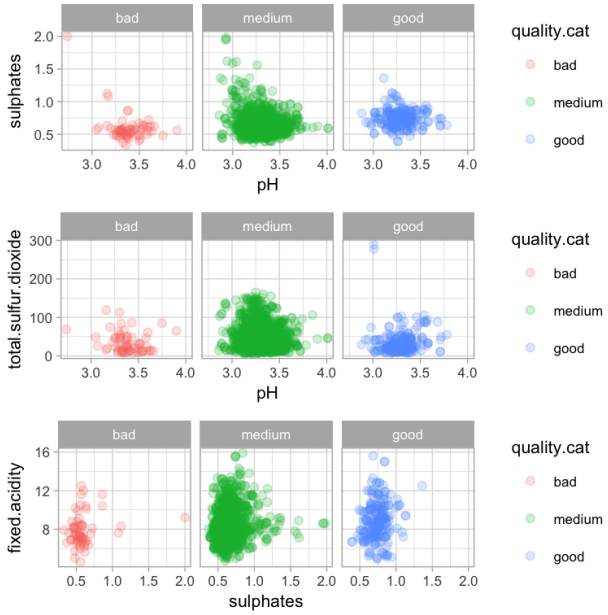
Figure 1: Correlation between alcohol, volatile acidity and sulphates

Correlation by quality



From the general correlation matrix, three main variables can be selected due to their high correlation with quality: **alcohol** (R=0.5), **volatile.acidity** (R=0.4) and **sulphates** (R=0.3). In order to see if they are suitable to perform an analysis let us explore the correlation between them and also the distribution of samples in bivariates plots in the next graphics 1.





Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Were there any interesting or surprising interactions between features?

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Final Plots and Summary

Plot One

Description One

Plot Two

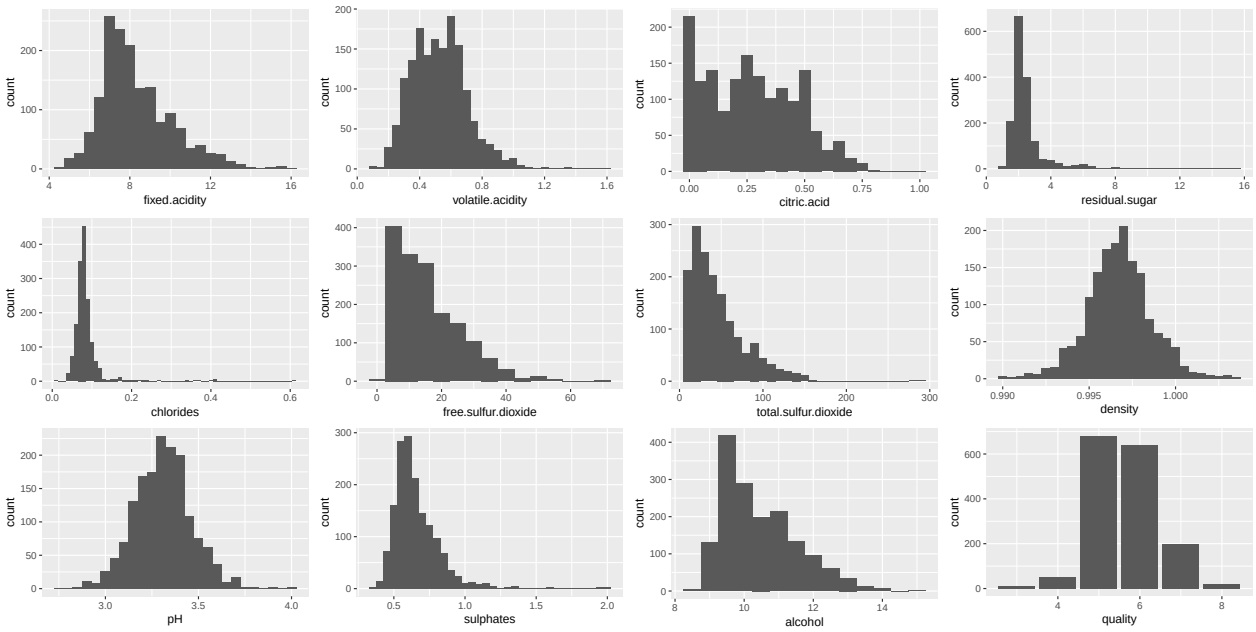
Description Two

Plot Three

Description Three

Reflection

References



References

Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos and José Reis. 2009. "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems* 47(4):547–553.