WHITE WINE QUALITY by ABIR PATTNAIK

INTRODUCTION

This dataset has been collected from: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. The data set is about the White wine and its quality.Here **Quality** is the target variable and it has various factors that affect the White Wine quality. This site (http://www.vinhoverde.pt/en/) gives details of the palce where it was collected.

About White Wine

White wine is a wine whose colour can be straw-yellow, yellow-green, or yellow-gold. It is produced by the alcoholic fermentation of the non-coloured pulp of grapes,which may have a skin of any colour. White wine has existed for at least 2500 years.Source (https://en.wikipedia.org/wiki/White_wine)

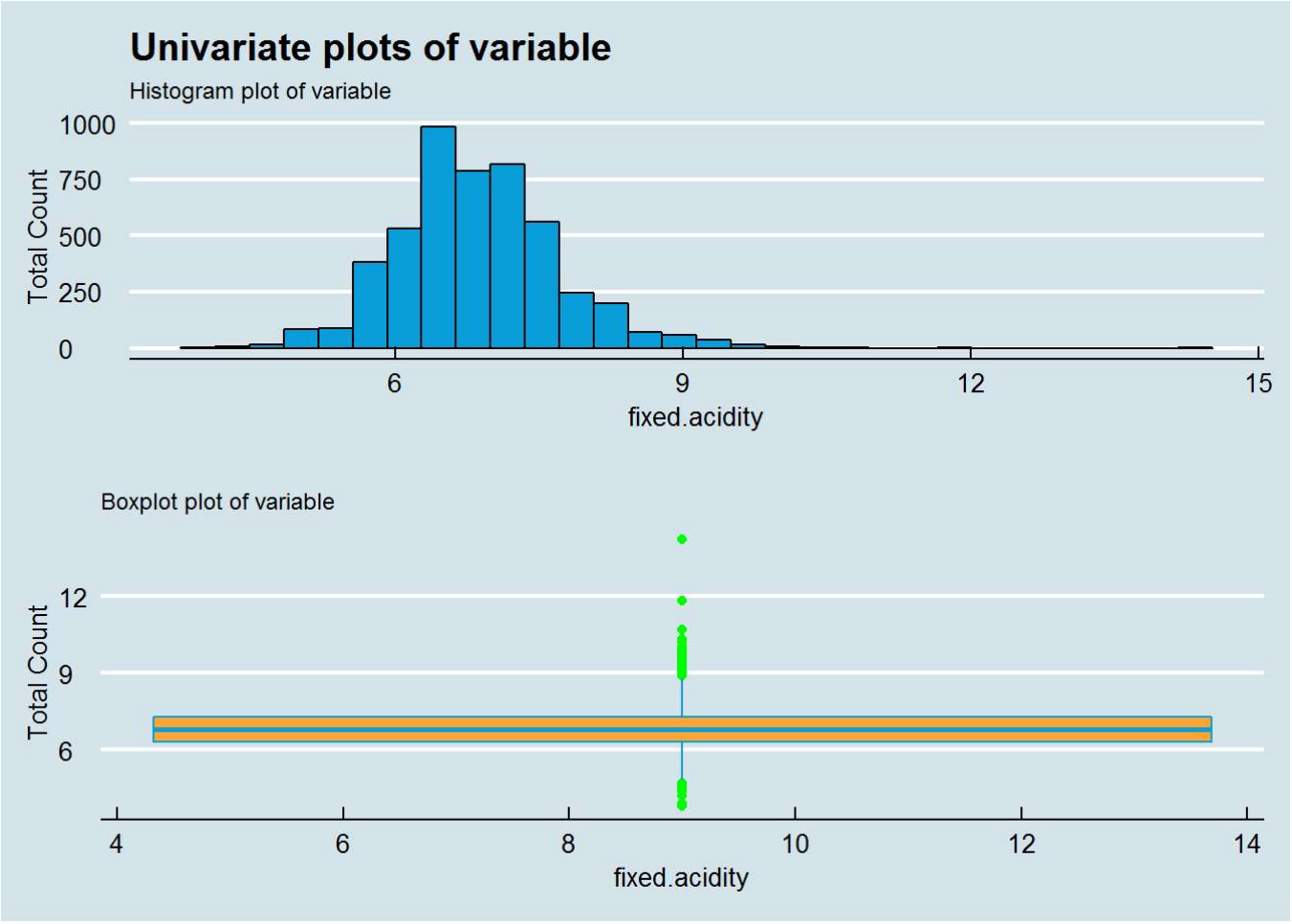
Univariate Plots Section

For Univariate analysis, help of box plots and histograms were taken.Although histograms demonstrate distributions better but box plots help in identifying the outliers.

Both of these plots demonstrate their unique features hence can help in making one variate plots a little easier.Here is a link (http://www.brighthubpm.com%20/six-sigma/58254-box-plots-vs-histograms-in-project-management/)

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044
...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
UnivariatePlots(fixed.acidity)
```



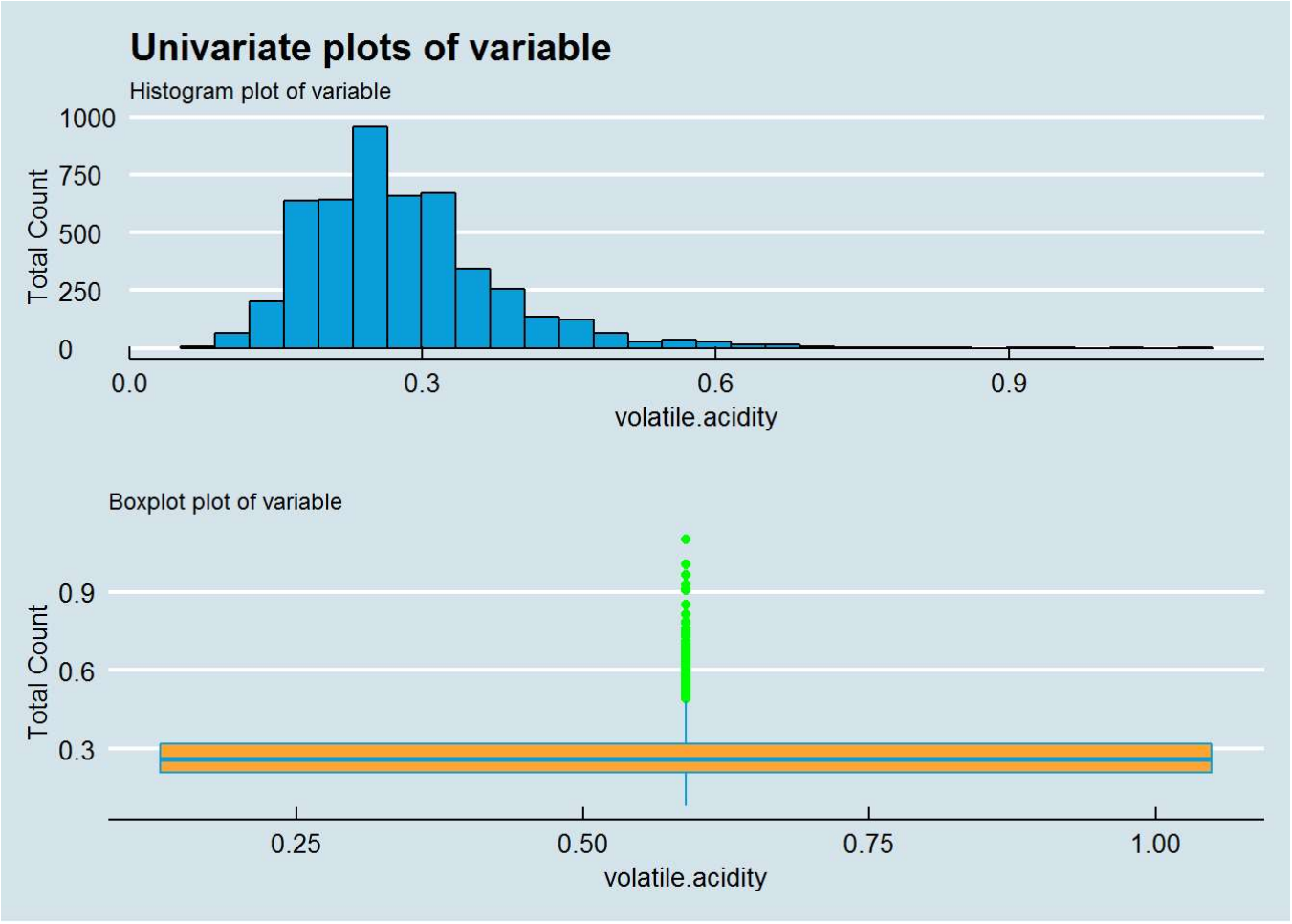

```
describe(fixed.acidity)
```

```
## fixed.acidity
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0       68    0.998    6.855    0.925      5.6      5.9
##      .25      .50      .75      .90      .95
##      6.3      6.8      7.3      7.9      8.3
##
## lowest :  3.8  3.9  4.2  4.4  4.5, highest: 10.2 10.3 10.7 11.8 14.2
```

Fixed Acidity

1. The fixed acidity is coming at a mean of 6.85 g/dm^3
2. Most of the wine has fixed acidity is between 6.3 g/dm^3.
3. The hishest fixed acidity is 14.2 g/dm^3 and is an outlier.

```
UnivariatePlots(volatile.acidity)
```




```
describe(volatile.acidity)
```

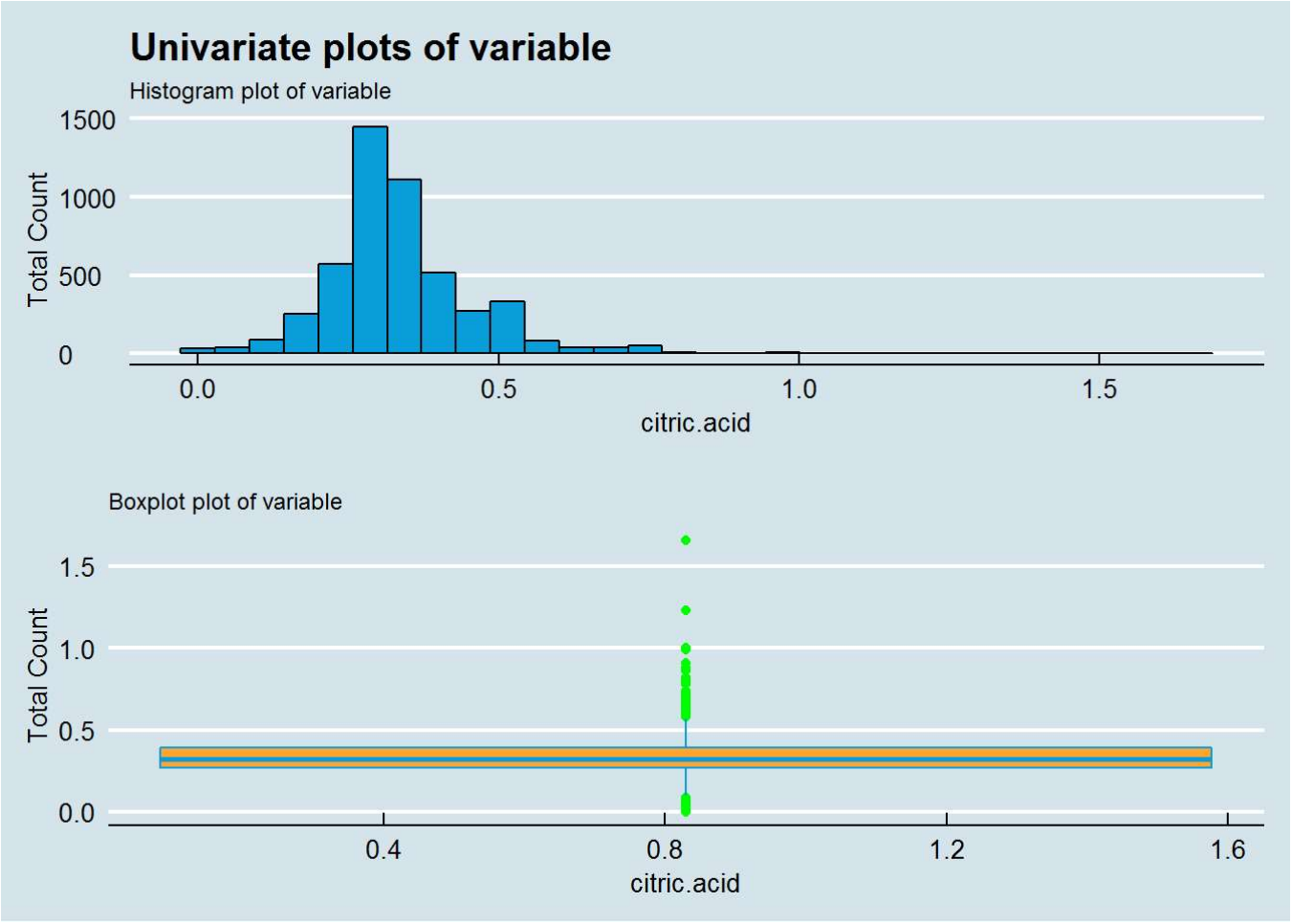
```
## volatile.acidity
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      125    0.999   0.2782   0.1055     0.15     0.17
##      .25      .50      .75      .90      .95
##     0.21     0.26     0.32     0.40     0.46
##
## lowest : 0.080 0.085 0.090 0.100 0.105, highest: 0.910 0.930 0.965 1.005 1.100
```

Volatile Acidity

The volatile acidity is the amount of acetic acid in wine.**High** values of lead to unpleasant smell.The histogram drawn is a left skewed graph with 0.26 g/dm^3 being the most quantity used in wine.

The last five outliers 0.91-1.10 g/dm^3 lead in unpleasantness and needs to be investigated further.The mean(0.278) is slightly higher than the median(0.26) due to the graph being left skewed.

```
UnivariatePlots(citric.acid)
```



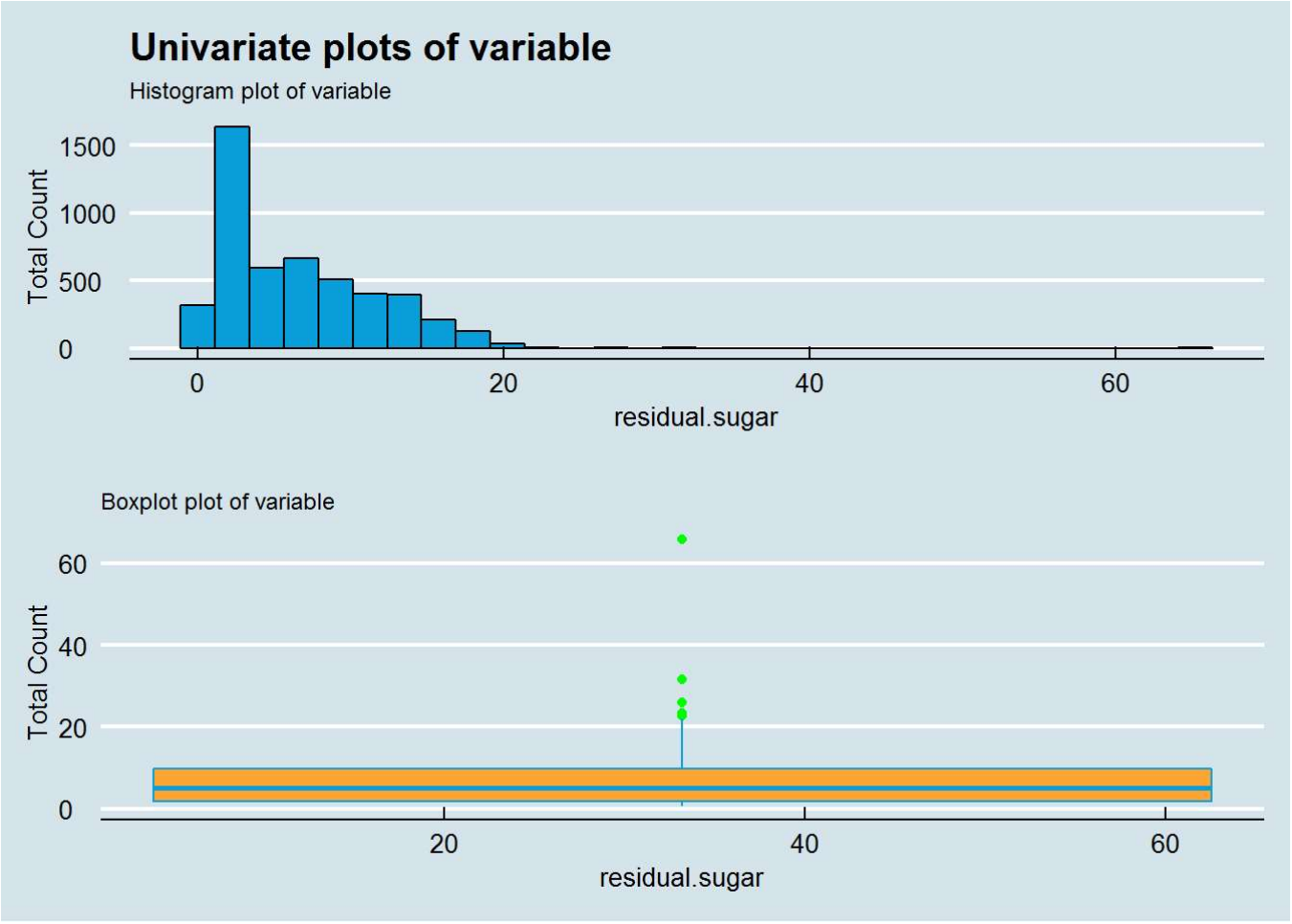

```
describe(citric.acid)
```

```
## citric.acid
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0       87    0.999   0.3342   0.1258     0.17     0.22
##      .25      .50      .75      .90      .95
##     0.27     0.32     0.39     0.49     0.54
##
## lowest : 0.00 0.01 0.02 0.03 0.04, highest: 0.91 0.99 1.00 1.23 1.66
```

Citric Acid

Citric Acid is found in small quantities but it adds ‘freshness’ and flavor to wines.Thus most of them are in the range of 0.27 g/dm^3 to 0.39 g/dm^3.It is possible having 0 g/dm^3 while it can also go high as 1.66 g/dm^3.

```
UnivariatePlots(residual.sugar)
```




```
describe(residual.sugar)
```

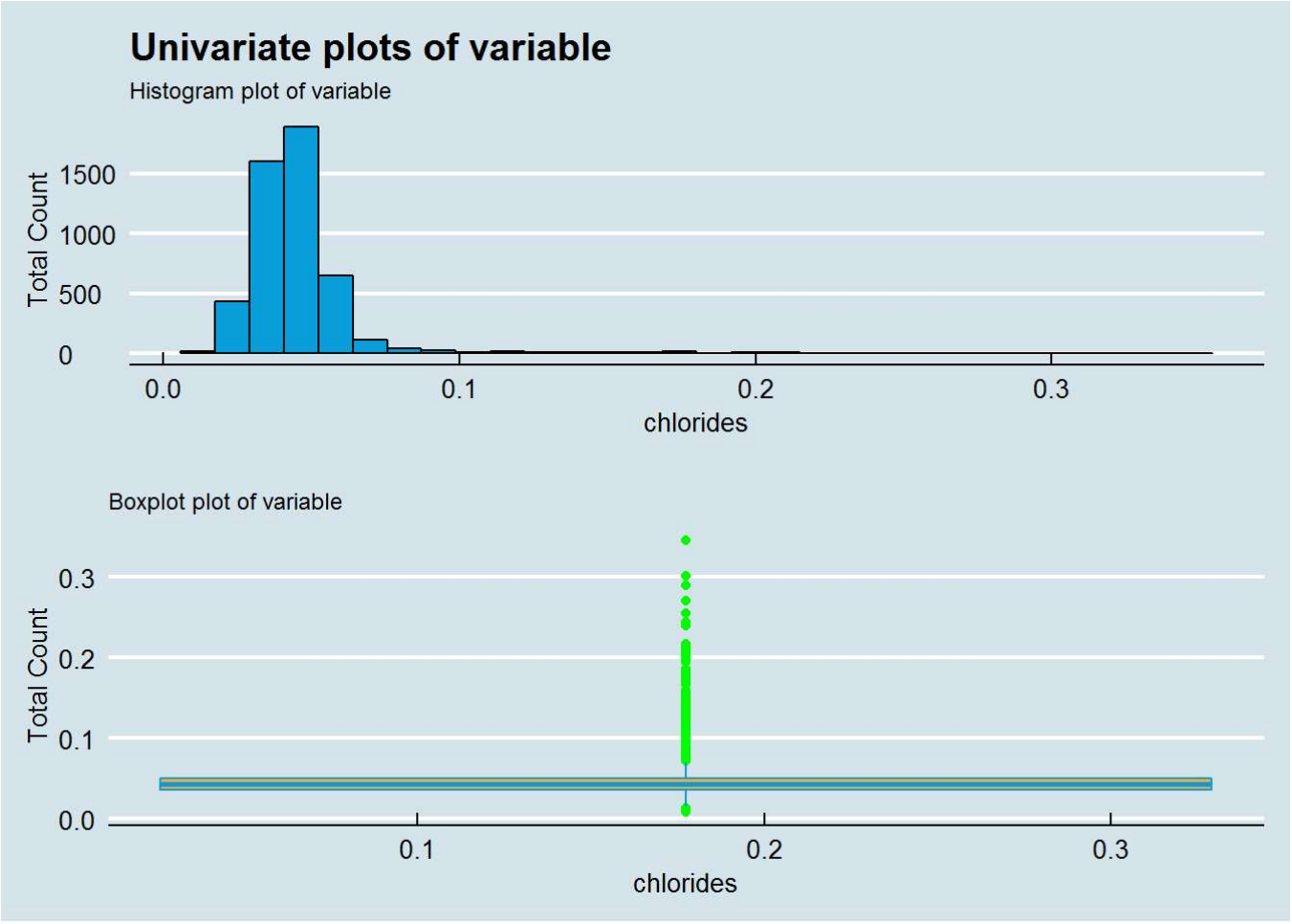
```
## residual.sugar
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      310        1    6.391    5.548      1.1      1.2
##      .25      .50      .75      .90      .95
##      1.7      5.2      9.9     14.0     15.7
##
## lowest :  0.60  0.70  0.80  0.90  0.95, highest: 22.60 23.50 26.05 31.60 65.80
```

Residual Sugar

Its the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.

Only 5% have 1.1g/dm^3. while 95% is within 15.7 g/dm^3.The outliers such as 65.80 g/dm^3 would be too sweet and its quality should be known.

```
UnivariatePlots(chlorides)
```



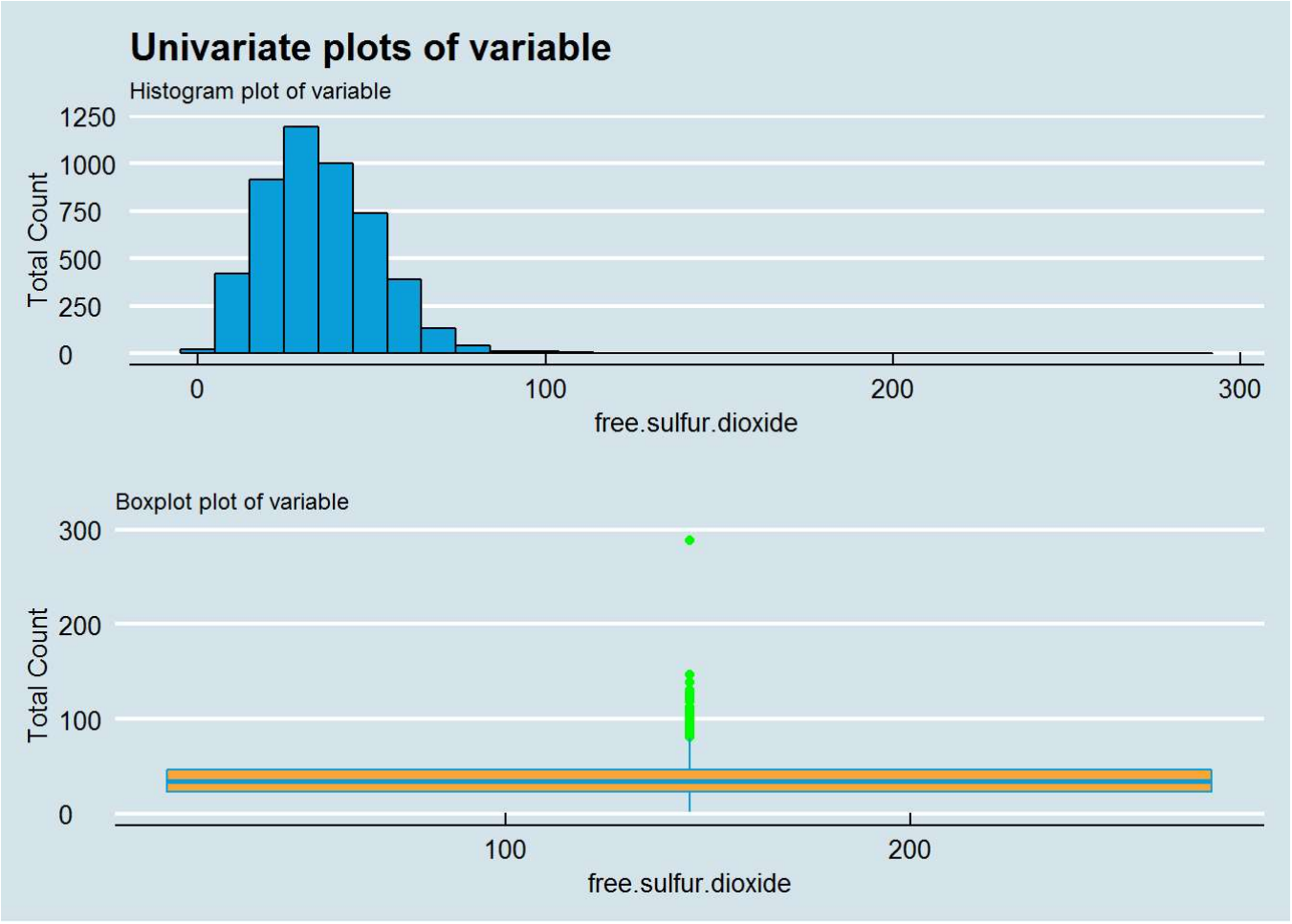

```
describe(chlorides)
```

```
## chlorides
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      160    0.999  0.04577  0.01708    0.027    0.030
##      .25      .50      .75      .90      .95
##    0.036    0.043    0.050    0.058    0.067
##
## lowest : 0.009 0.012 0.013 0.014 0.015, highest: 0.255 0.271 0.290 0.301 0.346
```

Chlorides

Chlorides is the amount of salt in wine.Both the plots give interesting insight on the chlorides column. 95% of the values of chloride are within 0.067 g/dm^3. Thus high values *may* not be suitable.If the outliers are removed the graph would then be a normal distribution curve.

```
UnivariatePlots(free.sulfur.dioxide)
```



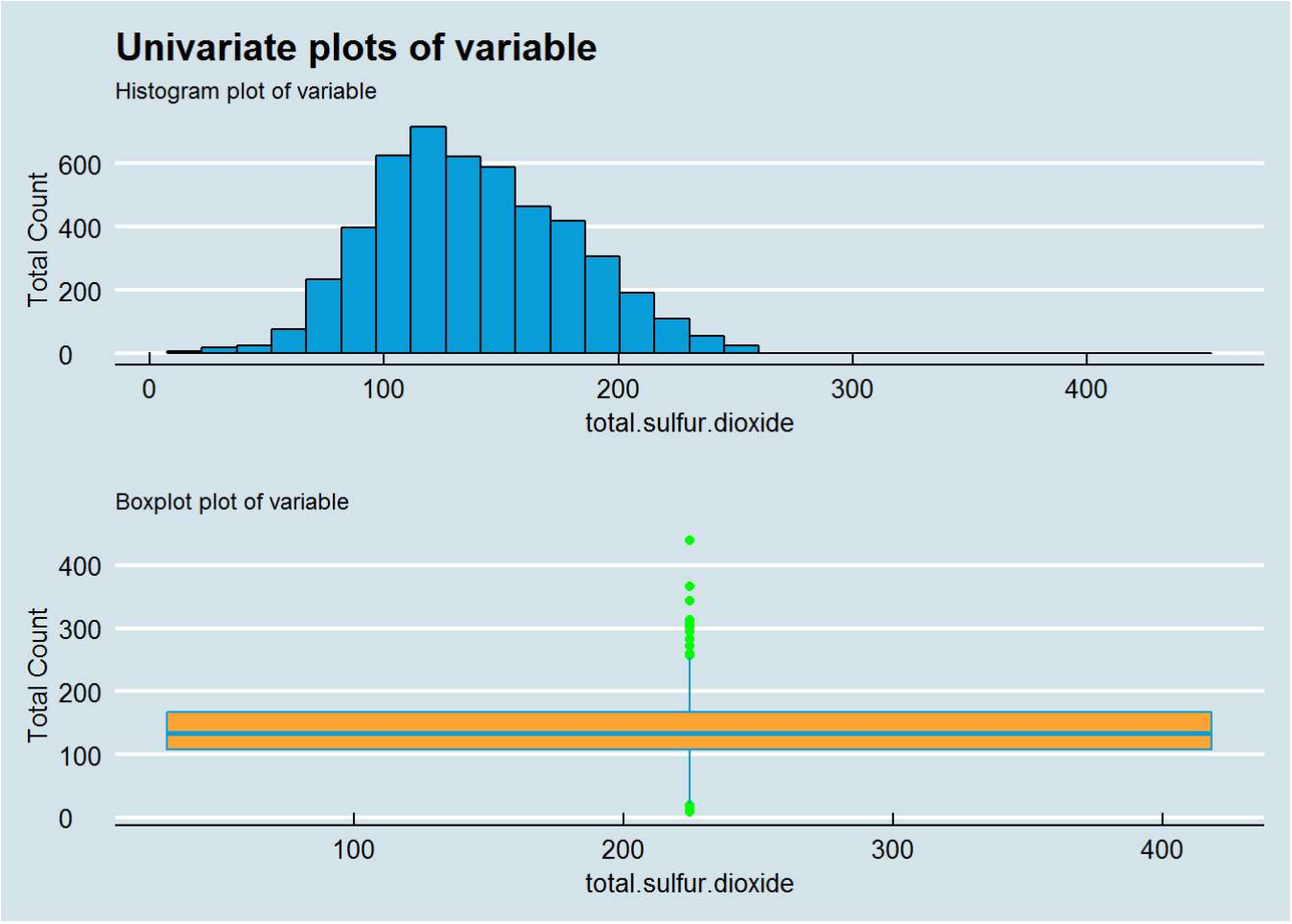

```
describe(free.sulfur.dioxide)
```

```
## free.sulfur.dioxide
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      132        1    35.31     18.5       11       15
##      .25      .50      .75      .90      .95
##       23       34       46       57       63
##
## lowest :   2.0   3.0   4.0   5.0   6.0, highest: 128.0 131.0 138.5 146.5 289.0
```

Free Sulphur Dioxide

The free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.Most of them are in range of 23 mg/dm^3 to 46 mg/dm^3 but it goes as high as 289 g/dm^3 and needs to investigated further.

```
UnivariatePlots(total.sulfur.dioxide)
```




```
describe(total.sulfur.dioxide)
```

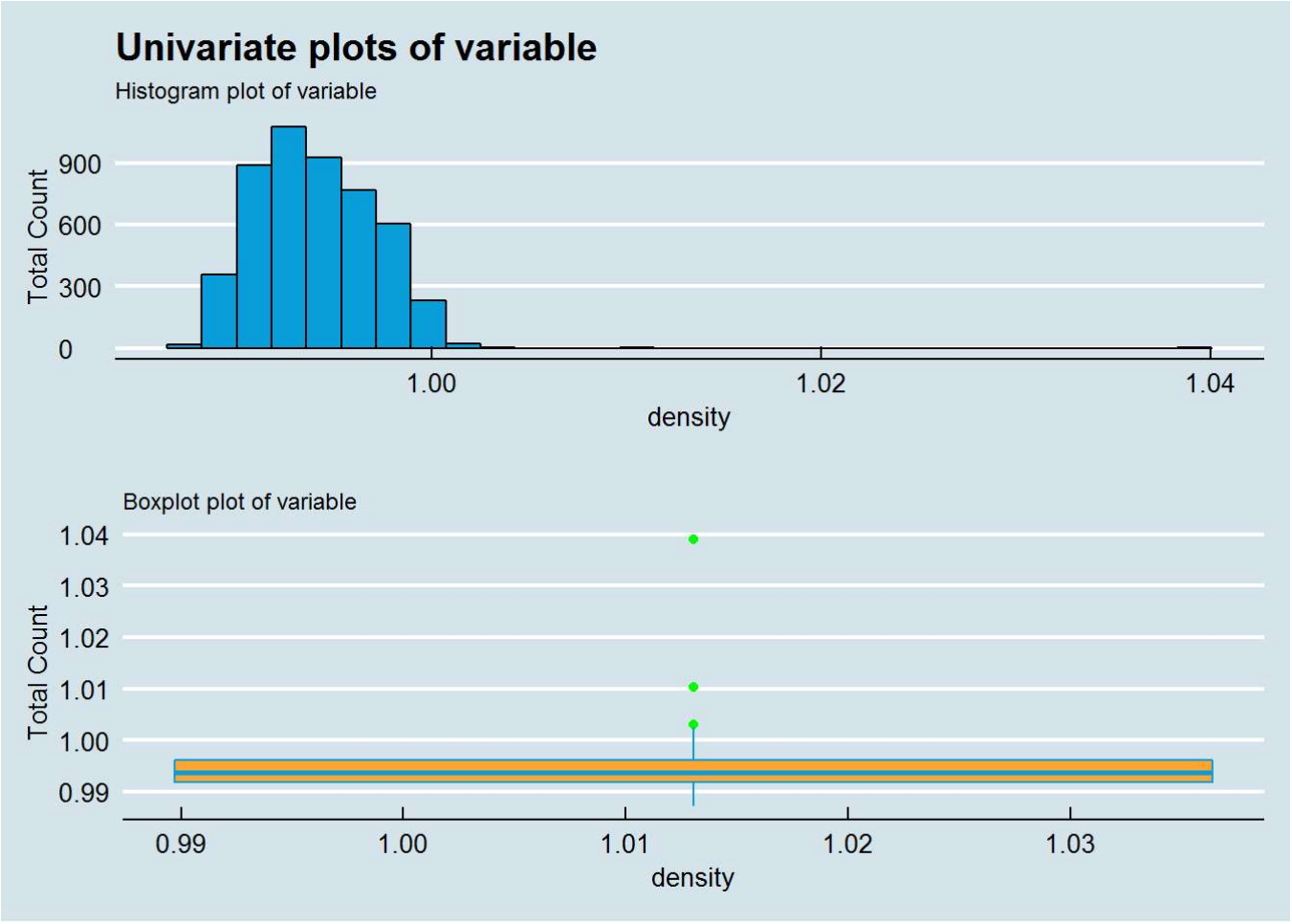
```
## total.sulfur.dioxide
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      251        1    138.4    47.79       75       87
##      .25      .50      .75      .90      .95
##      108      134      167      195      212
##
## lowest :   9.0  10.0  18.0  19.0  21.0, highest: 307.5 313.0 344.0 366.5 440.0
```

Total Sulphur Dioxide

It is the amount of free and bound forms of $S0_2$; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine.

Information needs to be clearer as most of the wines are in range of 108 mg/dm^3 to 167 mg/dm^3.

```
UnivariatePlots(density)
```




```
describe(density)
```

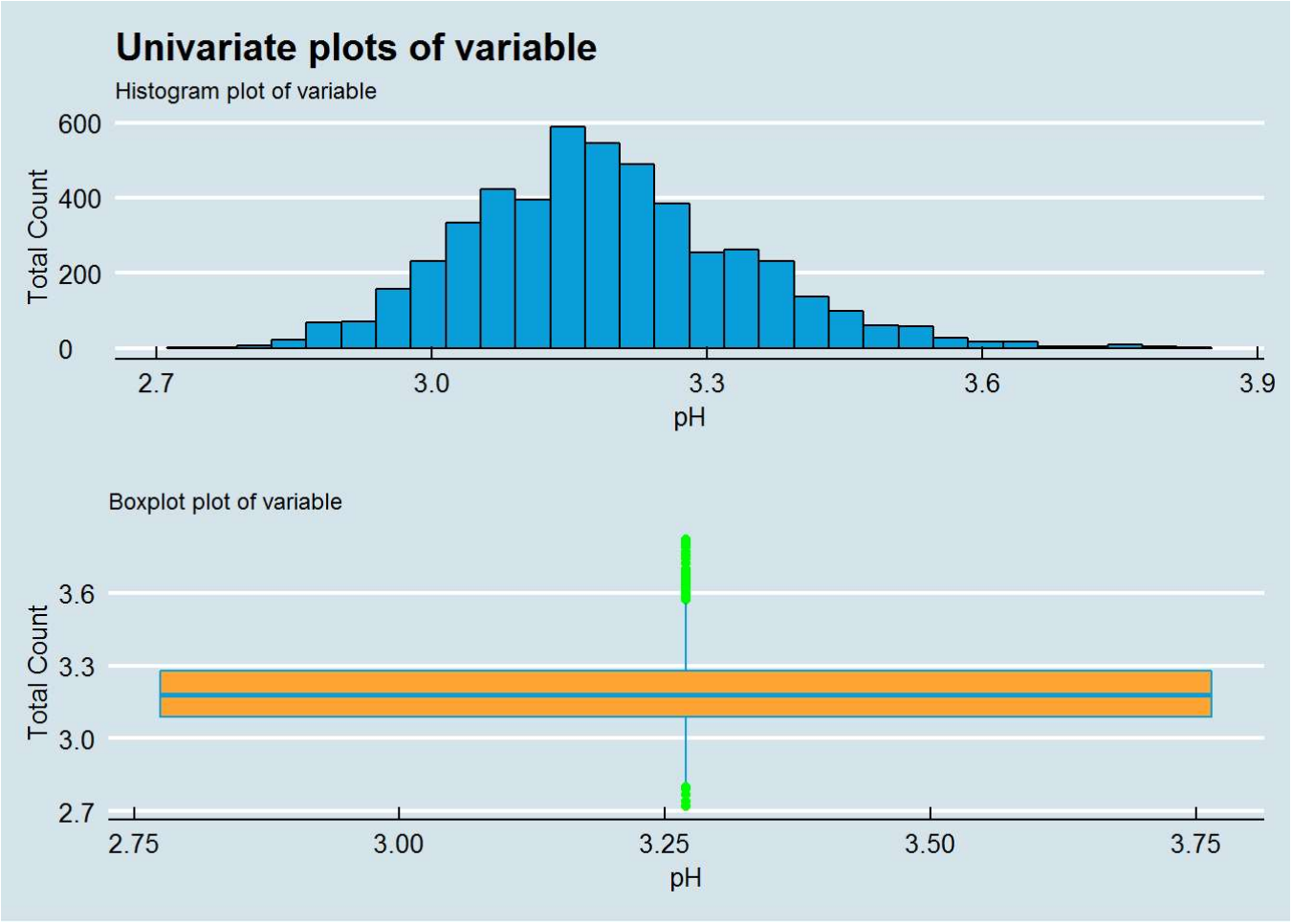
```
## density
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      890        1    0.994 0.003351   0.9896   0.9903
##      .25      .50      .75      .90      .95
##   0.9917   0.9937   0.9961   0.9981   0.9990
##
## lowest : 0.98711 0.98713 0.98722 0.98740 0.98742
## highest: 1.00240 1.00241 1.00295 1.01030 1.03898
```

Density

As the mean and median are quite close together the outliers 1.03898 g/cm^3 doesn't change the graph of histogram and most of them are in range of 0.9917-0.9961.

The density of water is close to that of water depending on the percent alcohol and sugar content.

```
UnivariatePlots(pH)
```



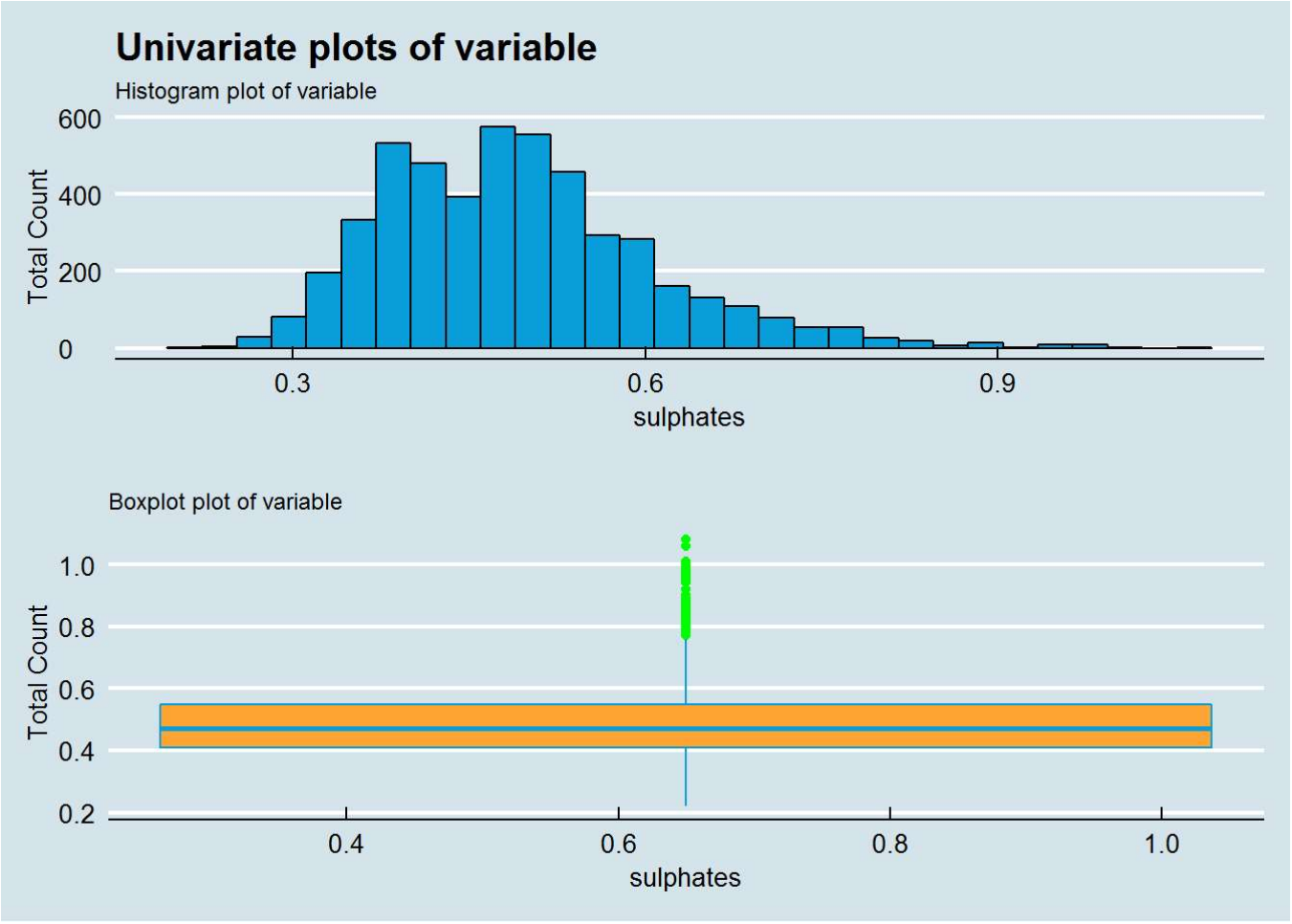

```
describe(pH)
```

```
## pH
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      103        1    3.188   0.1684     2.96     3.00
##      .25      .50      .75      .90      .95
##     3.09     3.18     3.28     3.38     3.46
##
## lowest : 2.72 2.74 2.77 2.79 2.80, highest: 3.77 3.79 3.80 3.81 3.82
```

pH

This is a fairly normal distribution curve. 95% pf values are below 3.46.pH describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic).Most of them are between 3-4.

```
UnivariatePlots(sulphates)
```



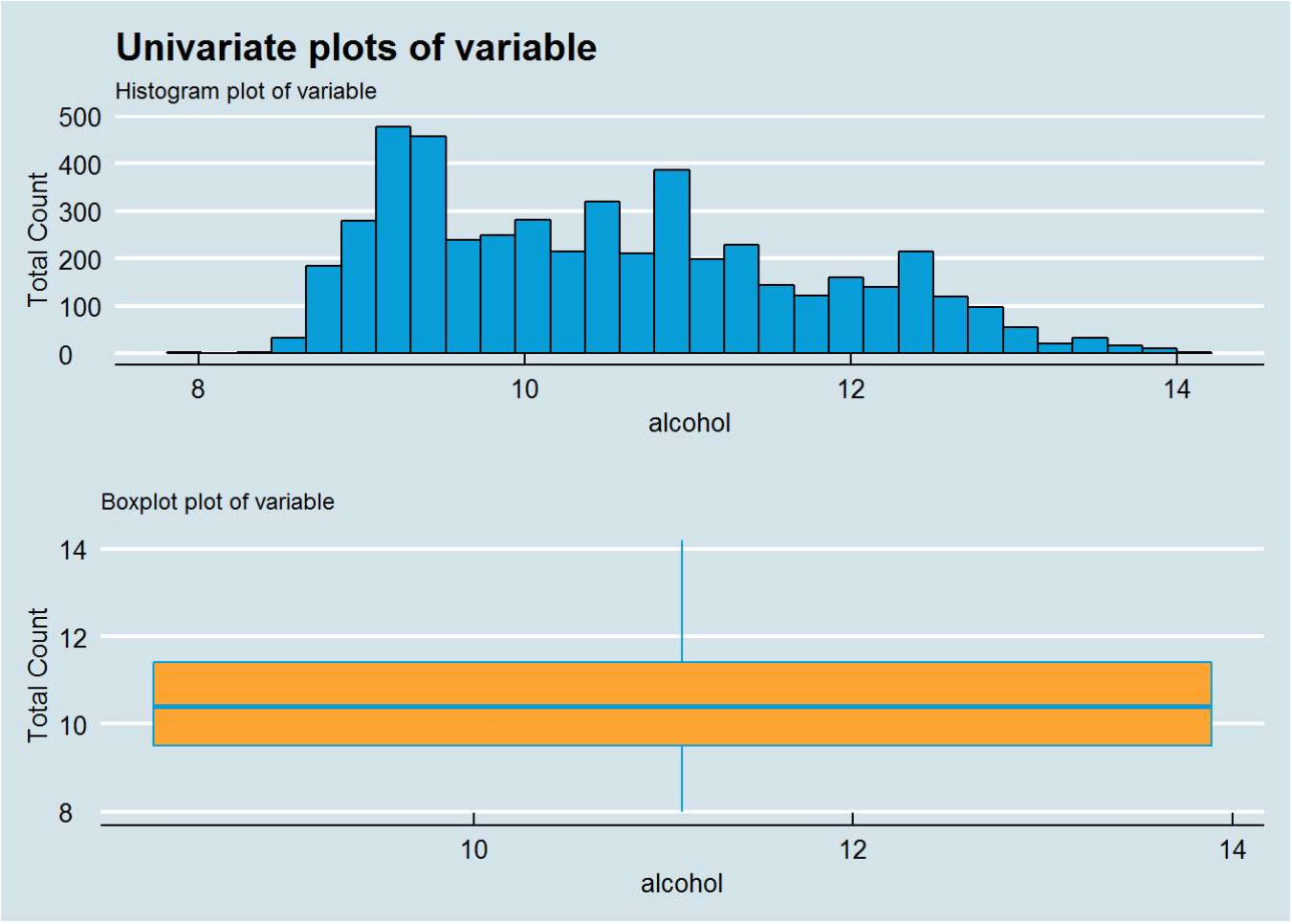

```
describe(sulphates)
```

```
## sulphates
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0       79    0.999   0.4898   0.1243     0.34     0.36
##      .25      .50      .75      .90      .95
##     0.41     0.47     0.55     0.64     0.71
##
## lowest : 0.22 0.23 0.25 0.26 0.27, highest: 0.99 1.00 1.01 1.06 1.08
```

Sulphates

Although they aren't any extreme outliers most of sulphates are between 0.41 and 0.55 g/dm^3.Sulphate is a wine additive which can contribute to sulfur dioxide gas ($S0_2$) levels, wich acts as an antimicrobial and antioxidant.

```
UnivariatePlots(alcohol)
```



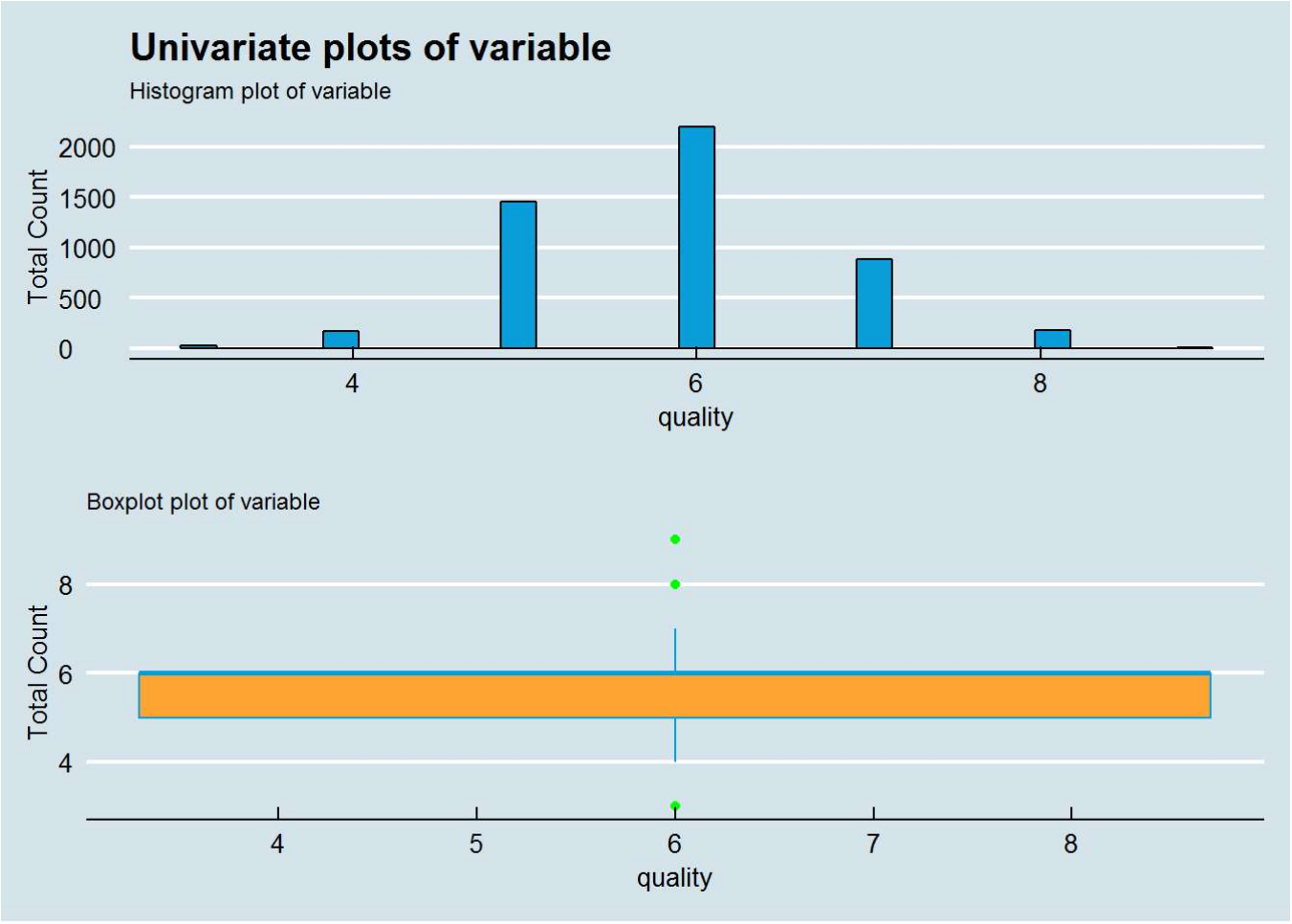

```
describe(alcohol)
```

```
## alcohol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4898        0      103    0.999    10.51    1.398      8.9      9.0
##      .25      .50      .75      .90      .95
##      9.5     10.4     11.4     12.4     12.7
##
## lowest :  8.00  8.40  8.50  8.60  8.70, highest: 13.80 13.90 14.00 14.05 14.20
```

Alcohol

The Histogram for alcohol is nothing like the plots of the previous variables as most of the values are spreaded over values hence it is inconclusive on whether it could be considered as a factor for quality or not. It gives the the percent alcohol content of the wine.

```
UnivariatePlots(quality)
```




```
describe(quality)
```

```
## quality
##        n  missing distinct     Info     Mean      Gmd
##     4898        0        7    0.877    5.878   0.9377
##
## Value          3     4     5     6     7     8     9
## Frequency     20   163  1457  2198   880   175     5
## Proportion 0.004 0.033 0.297 0.449 0.180 0.036 0.001
```

Quality

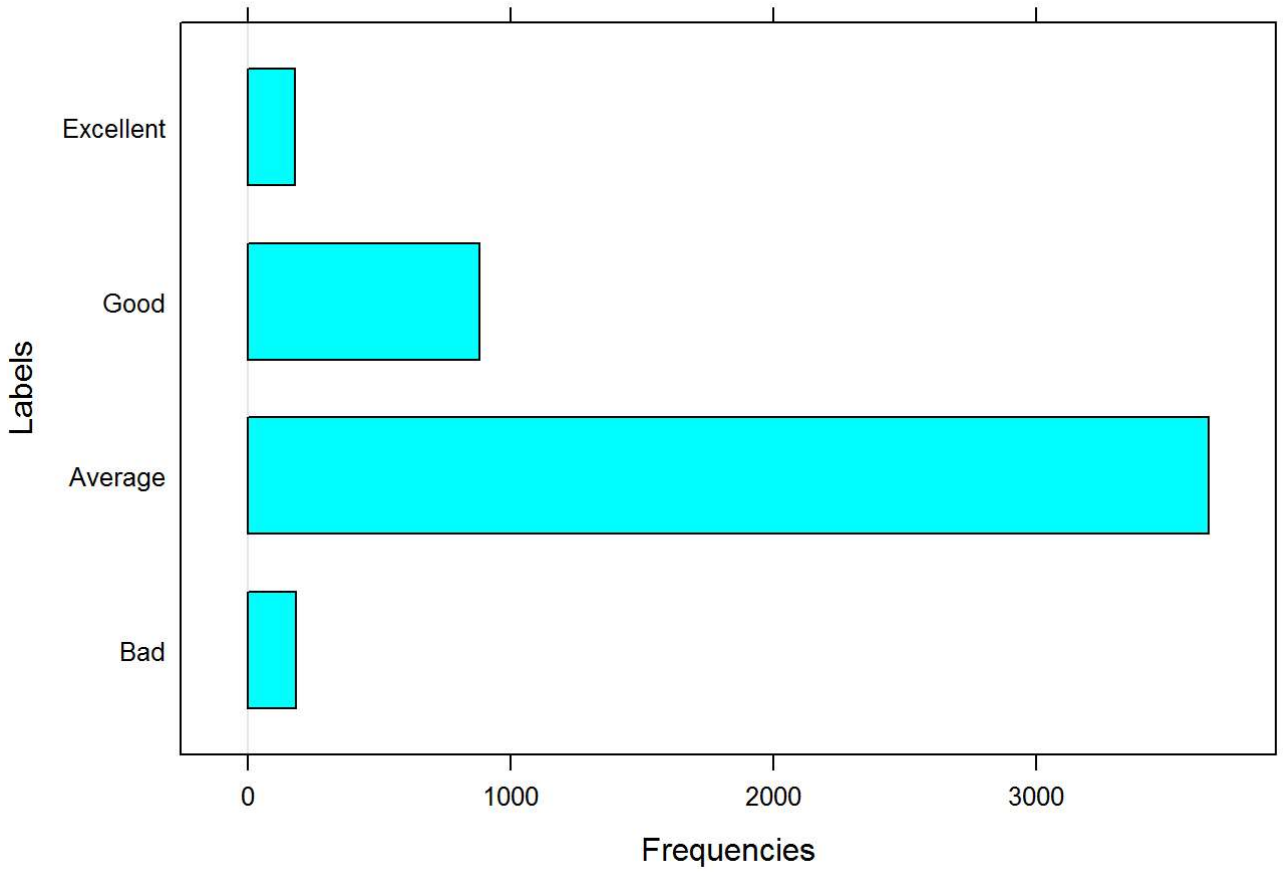
Quality here is the target variable.We need to know what factors make the quality of wine 8 or higher.FOr these,univariate plots were created for each factor.The plots of Quality is a normal distribution curve with mean at 5.878. We need to find relations with them and make an analysis further.

Categorising Quality variable

```
# Creating variables to categorise the quality of wine into 4 groups namely,
# Bad,Average,Good,Excellent
category=cut(quality,4,labels=c('Bad','Average','Good','Excellent'))
summary(category)
```

```
##       Bad   Average      Good Excellent
##       183      3655       880       180
```

```
barchart(table(category), freq=TRUE,
         xlab = "Frequencies",
         ylab = "Labels")
```



Average has the highest quantity of wines as compared to Bad and Excellent Wines. Hence, we need to see what all qualities does these labels for bad and excellent wines are made.

Univariate Analysis

What is the structure of your dataset?

The structure of dataset is the white wine dataset in which there are 11 variables that can help in determining the quality variable that is the target variable.There are in total of 4989 observations and has has been collected from a winery in Western Portugal.The Quality can be converted into categorical variable while others are all continuos variables.

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

1.Our main objective is to determine which factors can **cause** the quality of wine. 2.The alcohol column is inconclusive and needs to be researched upon that. 3.Another feature was the outliers in the data set and it needs to be researched upon with the Quality. Quality here is a target variable and is of interest.

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

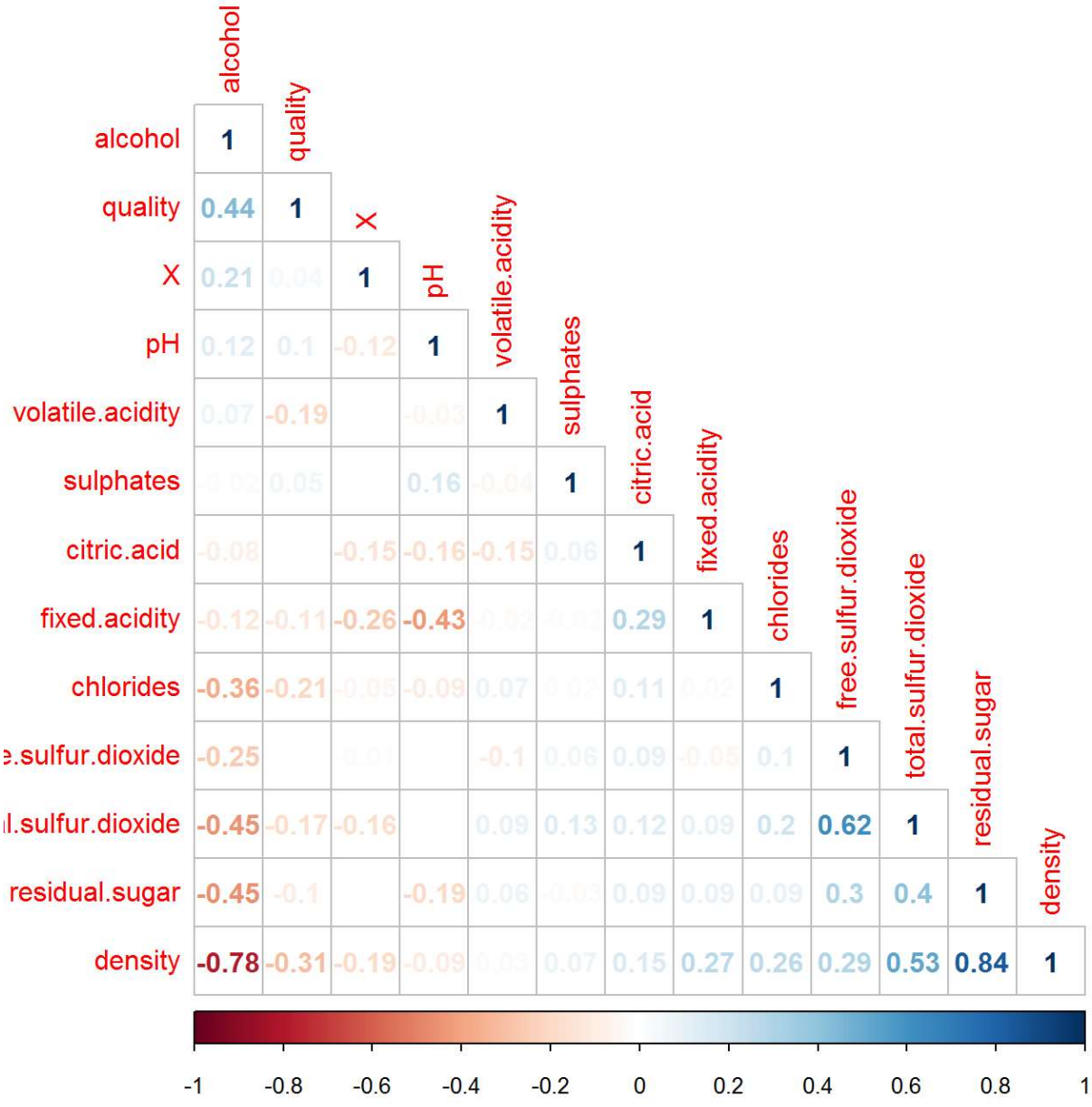
Features like which would be better for the test of wine such as volatile acidity ,citric acid, the amount of residual sugar may help in the investigation.

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

Since the quality variable contains only integers I can convert them to categories like 'bad','average','good','great'.

Bivariate Plots Section

```
corrplot(cor(White_wine),method="number"
         ,type = "lower"
         ,order = "FPC"
         ,number.cex=1)
```

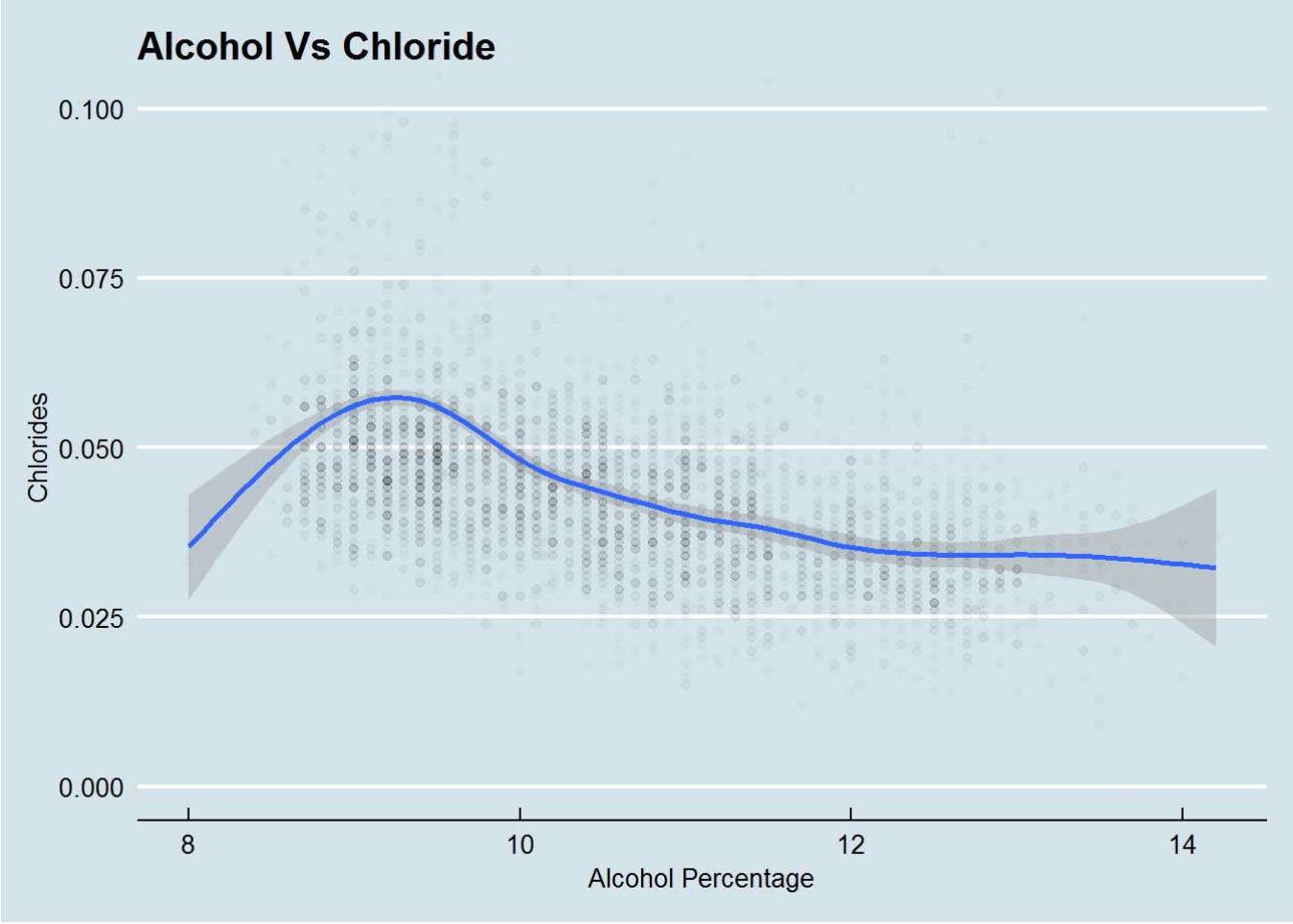


Bivariate Analysis

Corrplot packages gives correlation between the variables and it does give interesting insights on how to variables give us features of interest.However it should be noted that **CORRELATION IS NOT CAUSATION**. This means that even though variables may be correlated but it isn't necessary that it might be the cause.More understanding can be found by clicking this link (https://goo.gl/Trxp3W)

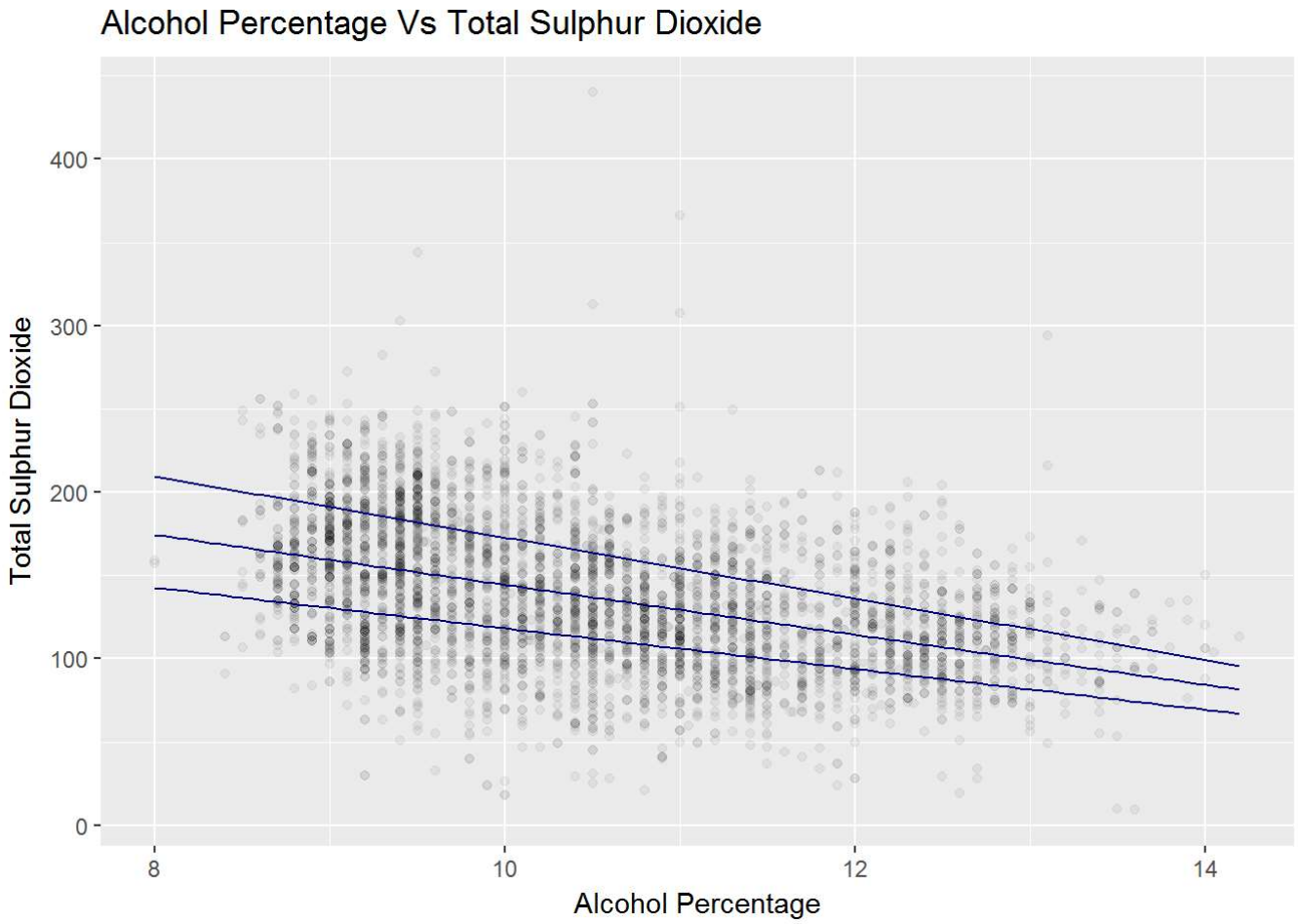
But there is a need of exploration on variables where they are correlated. Therefore plots were created of the following graphs on the basis of correlation. The **Rule of Thumb** says if two variables are correlated with >0.3 and <-0.3 then it is meaningful. A correlation of |0.5| is moderate. A correlation of |0.7| is large. Keeping that in mind following variables were investigated. 1.Alcohol v/s Chloride 2.Alcohol v/s Total Sulphuric Dioxide 3.Alcohol v/s residual sugar 4.Alcohol v/s Density 5.Density v/s Residual Sugar 6.Total Sulphuric Dioxide v/s Free Sulphuric Dioxide 7.Density v/s Total Sulphuric DIoxide 8.Quality v/s Alcohol

```
ggplot(data=White_wine,aes(x=alcohol,y=chlorides))+
  geom_point(alpha=1/50)+
  geom_smooth()+
  coord_cartesian(ylim=c(0,0.1))+
  theme_economist()+
  xlab('Alcohol Percentage') +
  ylab('Chlorides') +
  labs(title = 'Alcohol Vs Chloride')
```



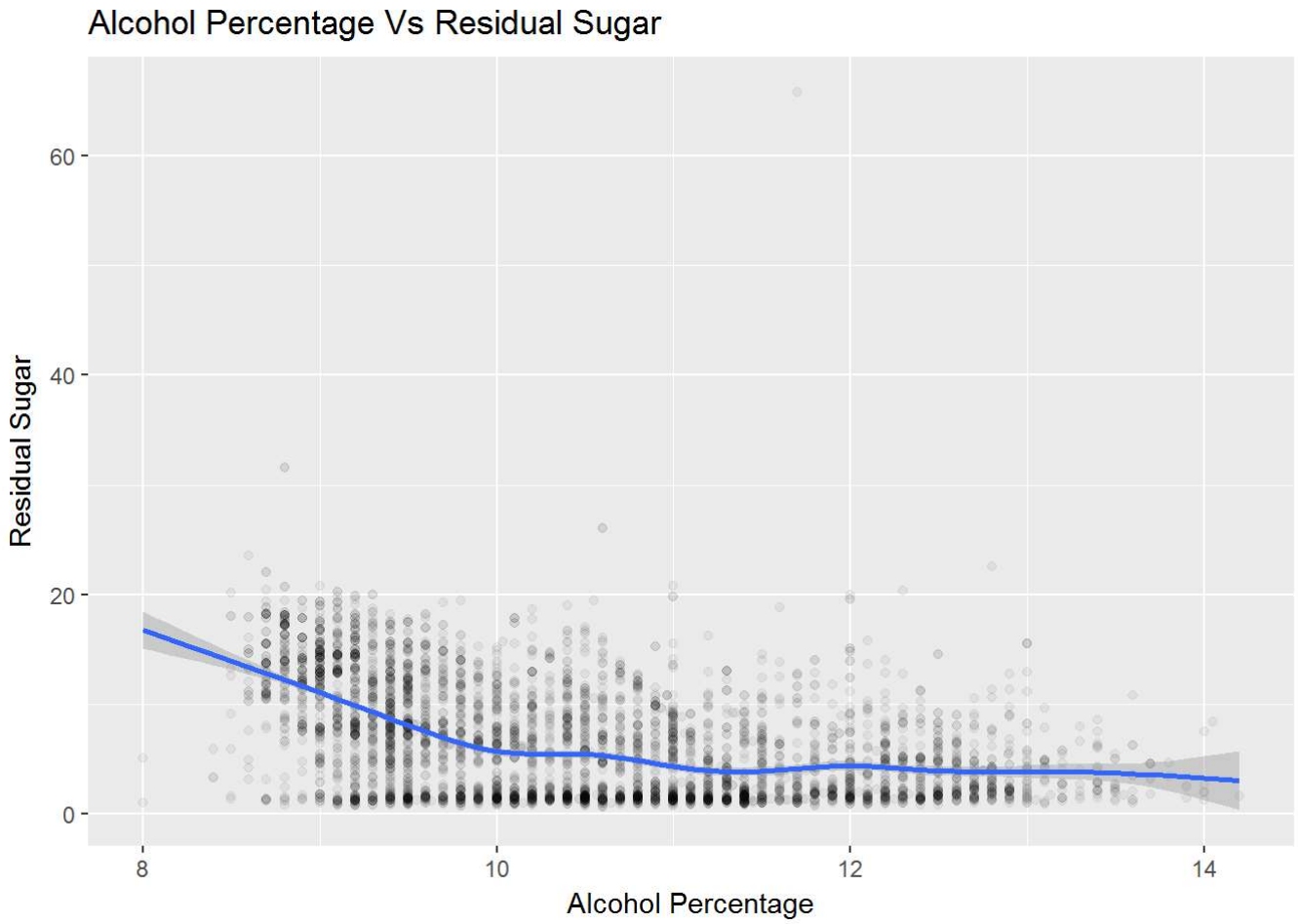
The alcohol content with 9% has chloride of approx 0.56 g/dm^3.Then the graph has a fairly linear decrease as the alchol is increased.

```
ggplot(data=White_wine,aes(x=alcohol,y=total.sulfur.dioxide))+
  geom_point(alpha=1/20)+
  geom_quantile(color = 'navyblue')+
  xlab('Alcohol Percentage') +
  ylab('Total Sulphur Dioxide') +
  labs(title = 'Alcohol Percentage Vs Total Sulphur Dioxide')
```



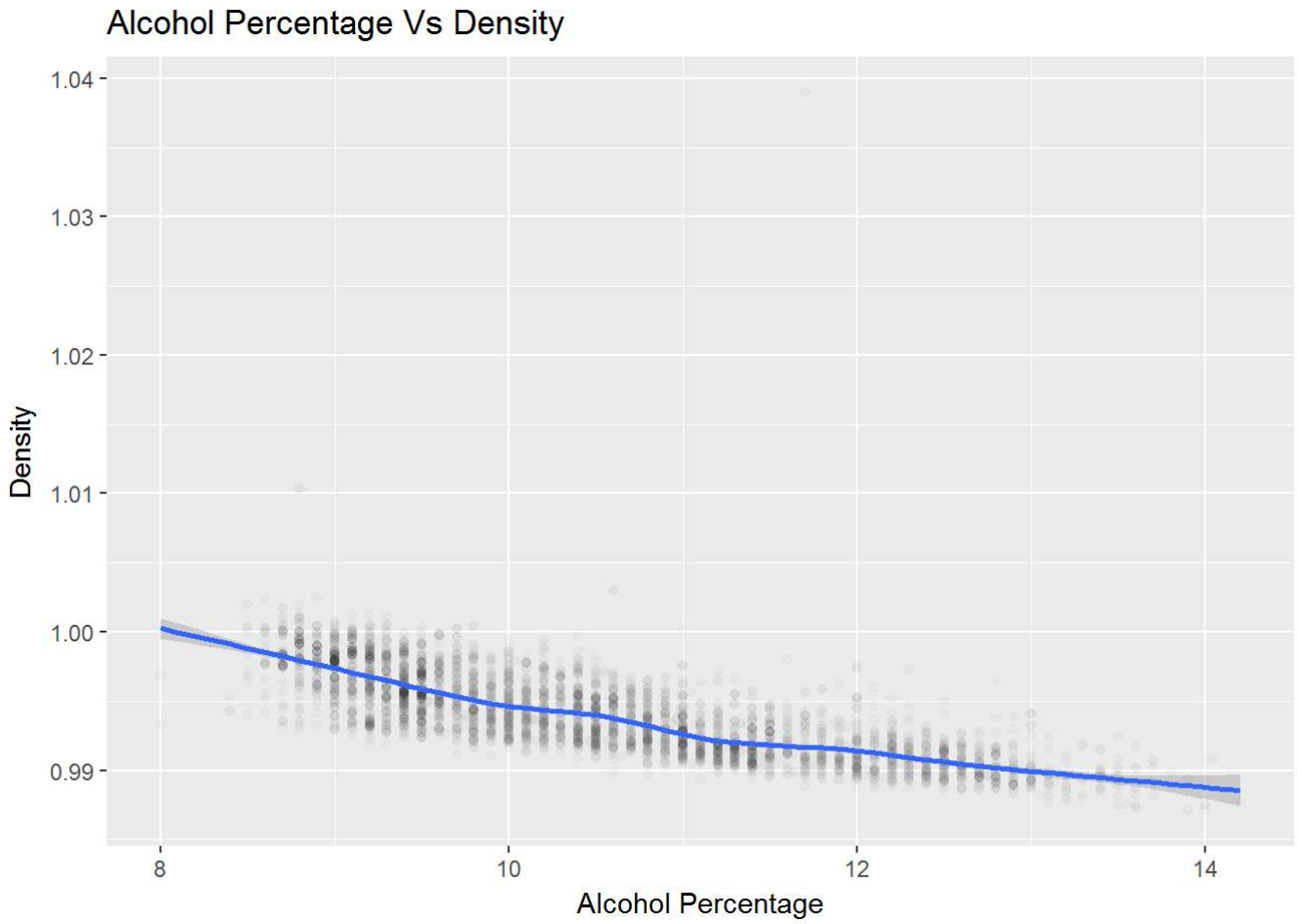
The Quantreg package plots quantiles onto the graph as a layer onto the graph and displayed that 9% alcohol has a wide range of quantile and most of the wines is present in this range only.

```
ggplot(data=White_wine,aes(x=alcohol,y=residual.sugar))+
  geom_point(alpha=1/20)+
  geom_smooth()+
  xlab('Alcohol Percentage') +
  ylab('Residual Sugar') +
  labs(title = 'Alcohol Percentage Vs Residual Sugar')
```



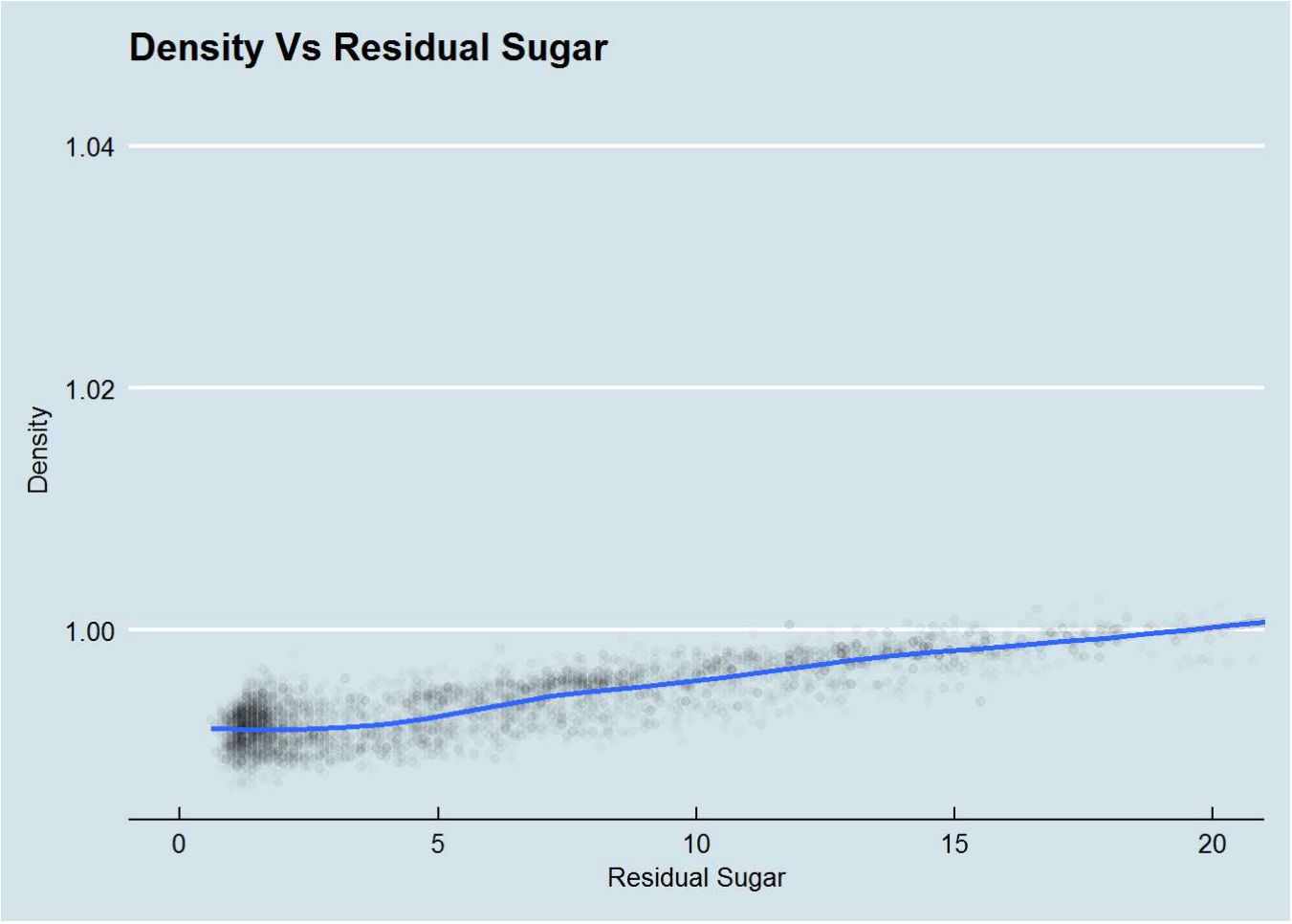
The residual sugar displays steep decline till it reaches 10% alcohol then there is a stagnant decrease.

```
ggplot(data=White_wine,aes(x=alcohol,y=density))+
  geom_point(alpha=1/50)+
  geom_smooth()+
  xlab('Alcohol Percentage') +
  ylab('Density') +
  labs(title = 'Alcohol Percentage Vs Density')
```



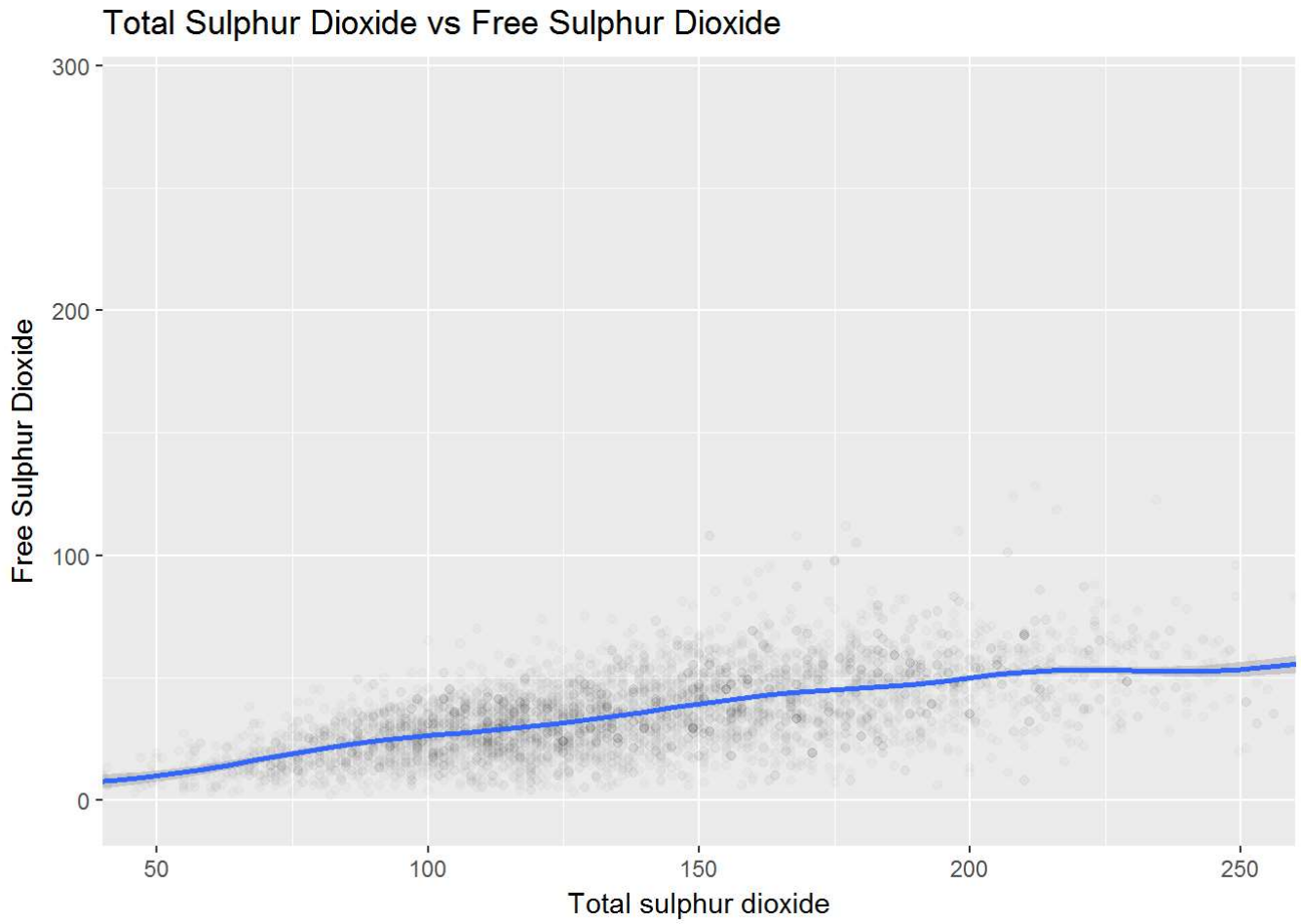
Drawing a line over a graph provides a fair relationship between variables.In this there is a downward a linear relationship between density and alcohol. Denser points are created between 9.6-9.8% alcohol.

```
ggplot(data=White_wine,aes(x=residual.sugar,y=density))+
  geom_point(alpha=1/50)+
  geom_smooth()+
  coord_cartesian(xlim = c(0,20))+
  theme_economist()+
  xlab('Residual Sugar') +
  ylab('Density') +
  labs(title = 'Density Vs Residual Sugar')
```



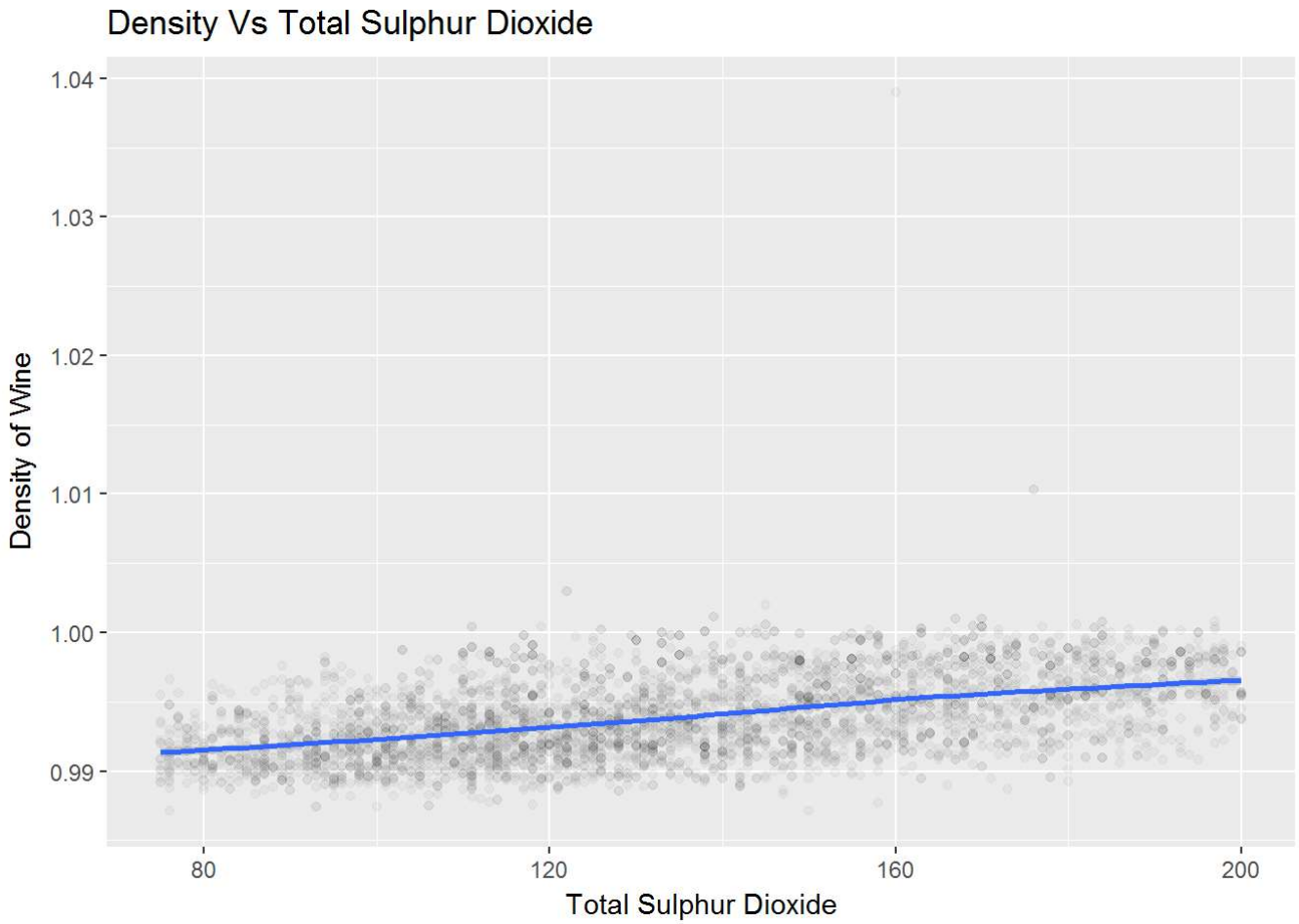
The graph displays some interesting features such as the residual sugar is increased the density also increases thus making the wine sweeter.Very few have residual sugar as g/dm^3.However between many wines are made between 1 and 3 gm/litre.

```
ggplot(data=White_wine,aes(x=total.sulfur.dioxide,y=free.sulfur.dioxide))+
  geom_point(alpha=1/50)+
  geom_smooth()+
  coord_cartesian(xlim = c(50,250))+
  xlab('Total sulphur dioxide') +
  ylab('Free Sulphur Dioxide') +
  labs(title = 'Total Sulphur Dioxide vs Free Sulphur Dioxide')
```



Having correlation having 0.62 it has a moderate correlation and it is possible since Free sulphuric dioxide and Total sulphuric dioxide are related hence have a high correlation.

```
ggplot(data=White_wine,aes(x=total.sulfur.dioxide,y=density))+
  geom_point(alpha=1/30)+
  geom_smooth()+
  scale_x_log10()+
  scale_x_continuous(limits = c(75,200))+
  coord_cartesian(xlim = c(75,200))+
  xlab('Total Sulphur Dioxide') +
  ylab('Density of Wine') +
  labs(title = 'Density Vs Total Sulphur Dioxide')
```

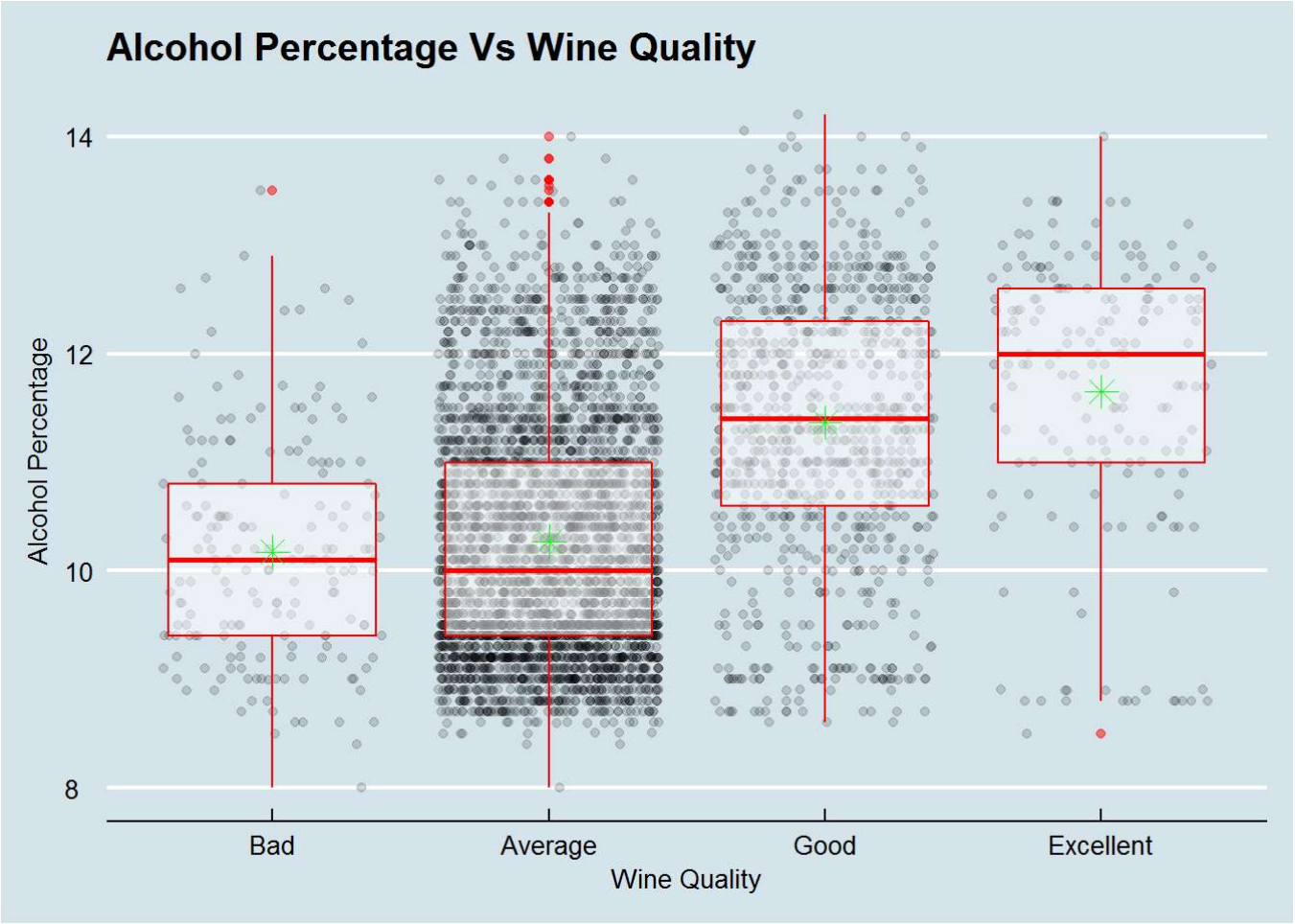


Points are denser between 80-120 g/dm^3 and although there is linear increase it keeps getting lesser when density is reaching towards 1 g/cm^3.

Inference

Alcohol can be considered as a factor that can affect the quality of wine. Thus creating a graph between Wine Quality and Alcohol is important.

```
ggplot(aes(category, alcohol),
       data = White_wine) +
  geom_jitter(alpha=.15)+
  geom_boxplot( alpha = .5,color = 'red')+
  stat_summary(fun.y = "mean",
               geom = "point",
              color = "green",
             shape = 8,
            size = 4) +
  xlab('Wine Quality') +
  ylab('Alcohol Percentage') +
  labs(title = 'Alcohol Percentage Vs Wine Quality')+
  theme_economist()
```



The Wine Quality was divided into 4 classes i.e. 'Bad','Average','Good', 'Excellent'.A layer of boxplot was placed over point graph. link (goo.gl/KBHik6) It shows that mean alcohol percentage for Excellent wines is approx 11.8% while the bad and average wines have approx 10.2 %.

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?

Although correlation is not causation but it can help us to determine the relationships between variables easily hence the following variables were taken and corresponding values of correlation was noted.

```
Comparison            | Correlation
----------------------|-----------------
Alcohol/Chlorides     | -0.36
Alcohol/TSD           | -0.45
Alcohol/Residual Sugar| -0.45
Alcohol/Density       | -0.78
Density/Residual Sugar| 0.84
TSD/FSD               | 0.62
Density/TSD           | 0.53
Alcohol/Quality       | 0.44
```

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

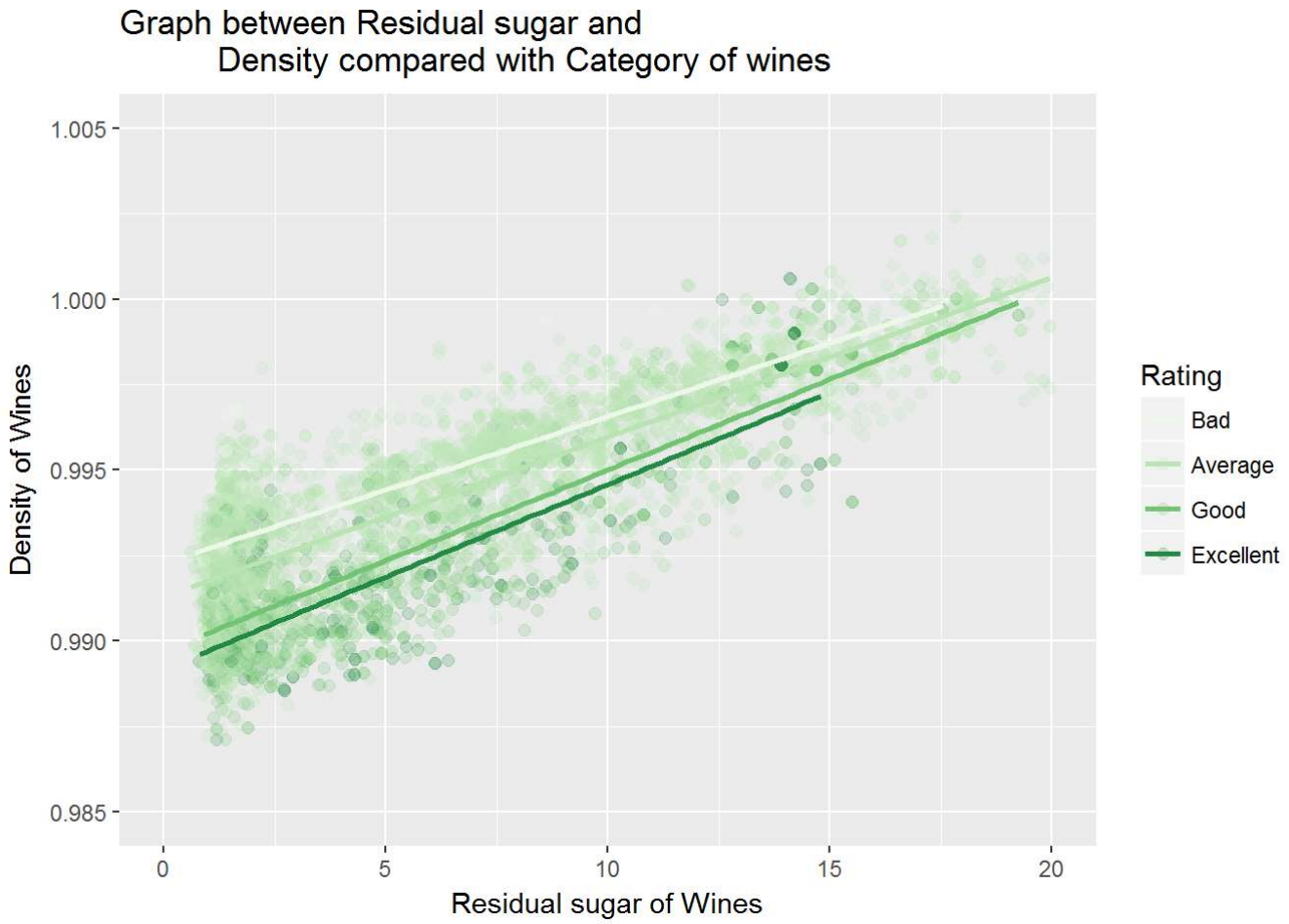
The relationship between density and TSD shows a correlation of 0.53.There is a linear relationship between 0.99 to 1.00 as the TSD increases.

What was the strongest relationship you found?

Strongest relationship was found between Density and Residual Sugar but still needs to be investigated upon.

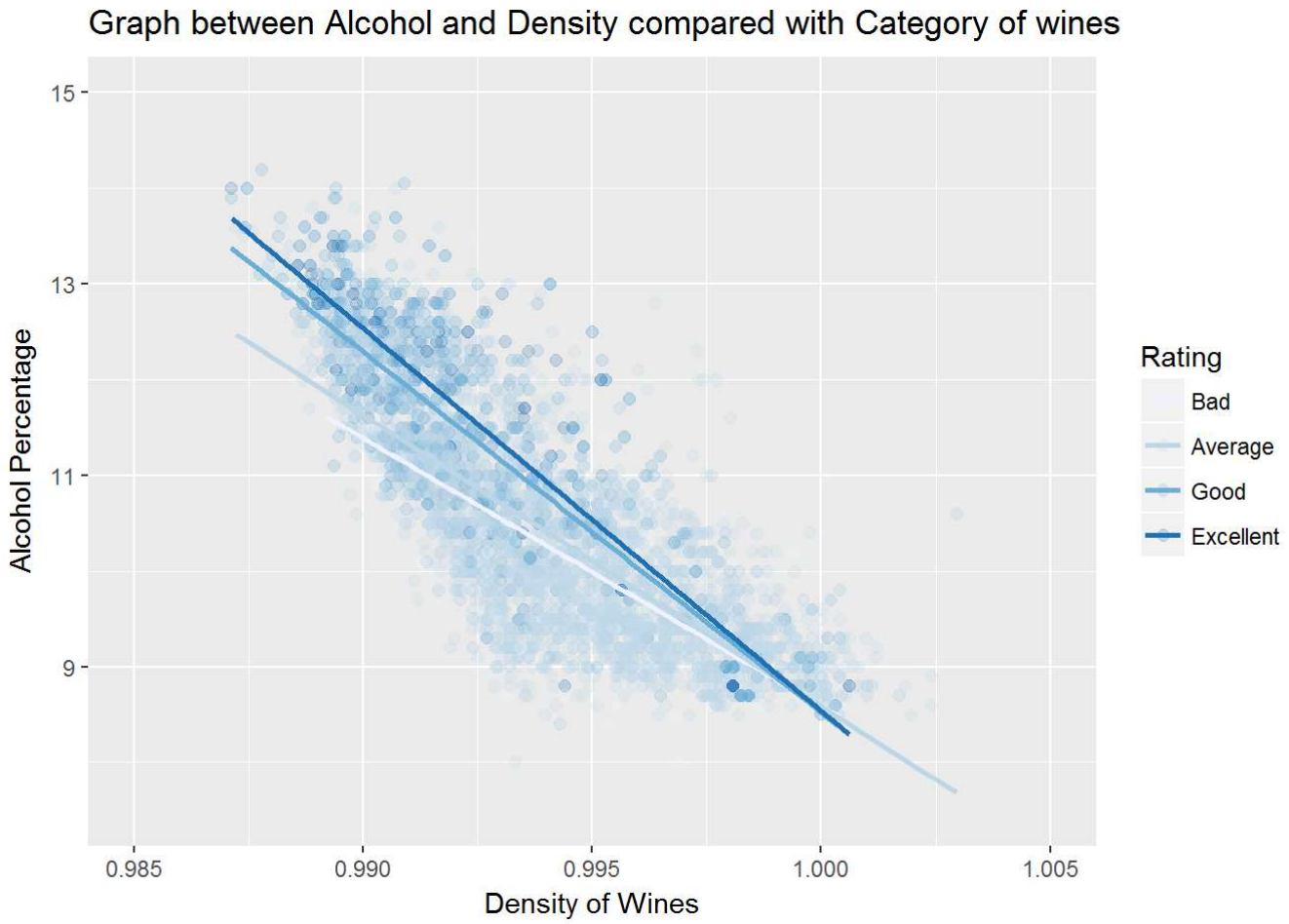
Multivariate Plots Section

```
ggplot(aes(x = residual.sugar, y = density, color = category),
               data = White_wine) +
     scale_color_brewer(type = 'seq',
                        guide = guide_legend(title = 'Rating'),
                        palette = 5) +
     geom_jitter(size = 2,alpha=1/5) +
     geom_smooth(method = "lm",
                 se = FALSE,
                 size=1)+
     xlim(0,20)+
     ylim(0.985,1.005)+
     xlab('Residual sugar of Wines')+
     ylab('Density of Wines')+
     labs(title='Graph between Residual sugar and
          Density compared with Category of wines')
```



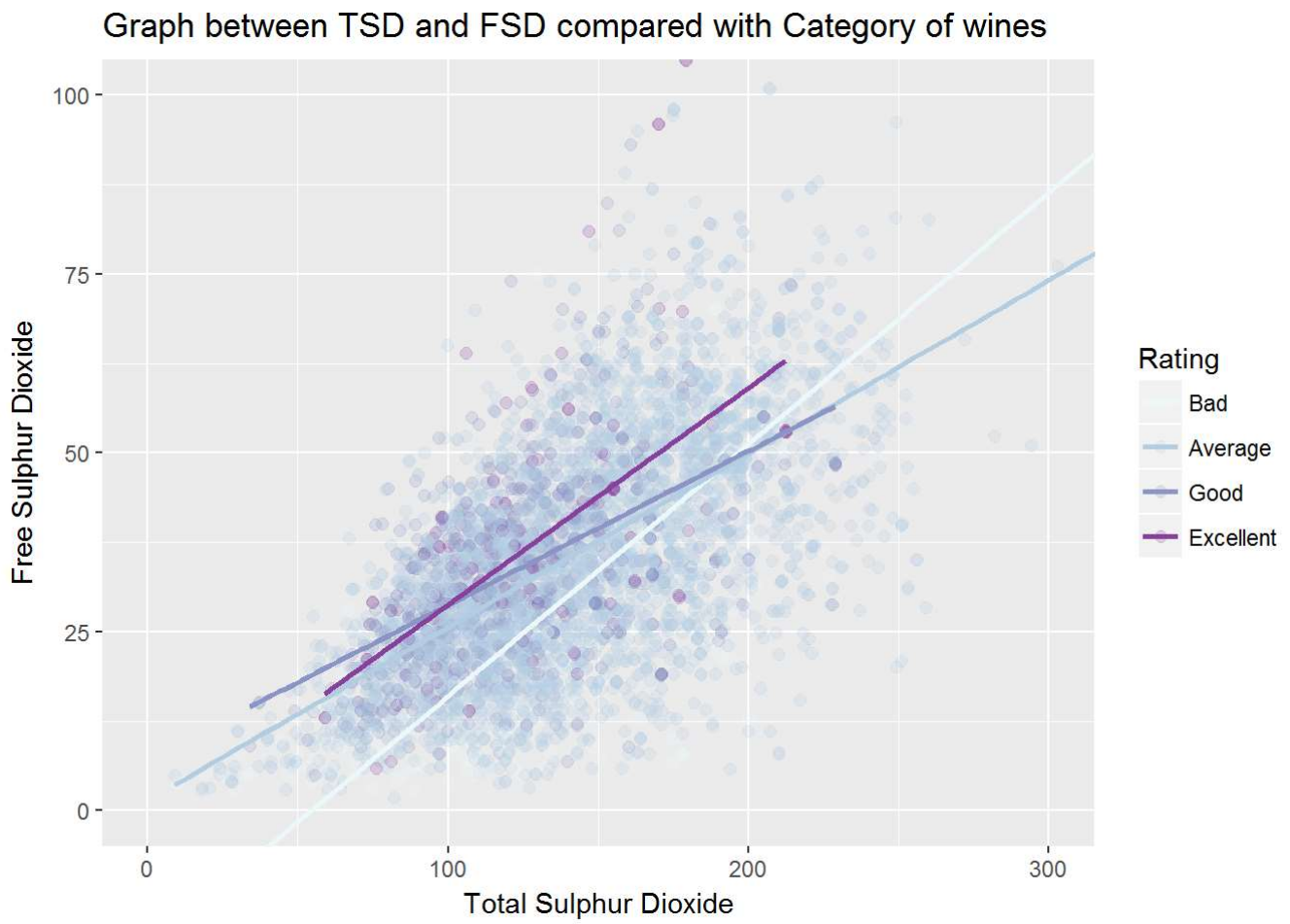
The previous graphs showed only the linear relationship between Residual Sugar and Density of wines.In this graph,the points were separated with respected to their rating and thus gives interesting points.These can help us by creating the perfect wine.Any residual sugar above 15 is not good for making wine.And density should be between the ranges of 0.990-0.995.

```
ggplot(aes(x = density, y = alcohol, color = category),
       data = White_wine) +
  scale_color_brewer(type = 'seq',
                     guide = guide_legend(title = 'Rating')) +
  geom_jitter(size = 2,
              alpha=1/5) +
  geom_smooth(method = "lm",
              se = FALSE,
              size=1)+
  xlim(0.985,1.005)+
  ylim(7.5,15)+
  xlab('Density of Wines')+
  ylab('Alcohol Percentage')+
  labs(title='Graph between Alcohol and Density compared with Category of wines')
```



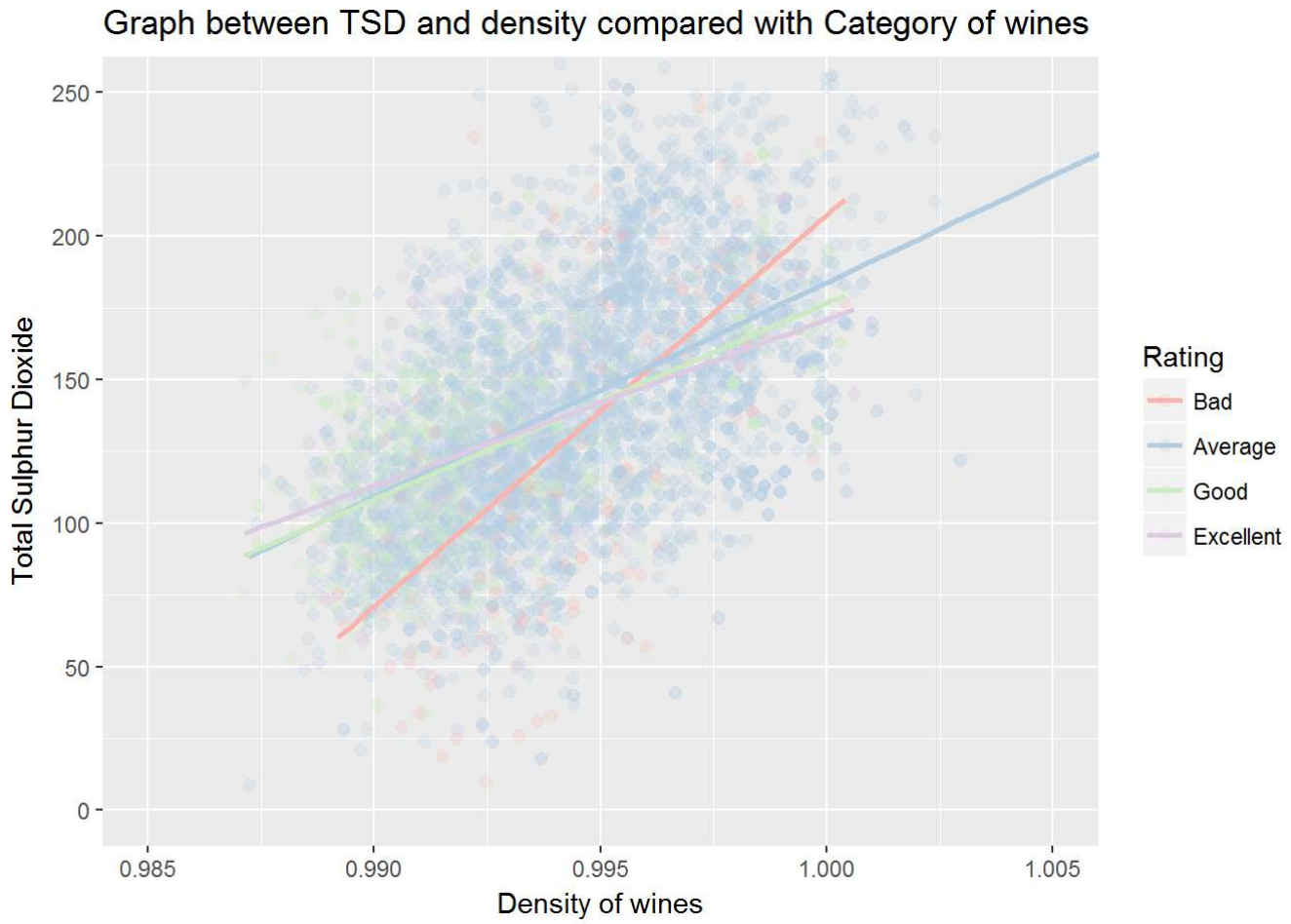
Bad Wines have comparatively low precentage of alcohol as compared to Excellent wines as displayed by the four lines.As the density is increased the alcohol percentage also decreases thus helping us understand the fact that alcohol percentage between 11-13% is quite good for making wine.

```
ggplot(aes(x = total.sulfur.dioxide, y = free.sulfur.dioxide, color = category),
       data = White_wine) +
  scale_color_brewer(type = 'seq',
                     guide = guide_legend(title = 'Rating'),
                     palette = 3) +
  geom_jitter(size = 2,
              alpha=1/5) +
  geom_smooth(method = "lm",
              se = FALSE,
              size=1)+
  coord_cartesian(xlim = c(0,300),
                  ylim = c(0,100))+
  xlab('Total Sulphur Dioxide')+
  ylab('Free Sulphur Dioxide')+
  labs(title='Graph between TSD and FSD compared with Category of wines')
```



The relationship between TSD(Total Sulphur Dioxide) and FSD (Free Sulphur Dioxide) has some interesting features as slope of Bad Wine is quite higher than avaerage wines as TSD and FSD increases.

```
ggplot(aes(x = density, y = total.sulfur.dioxide, color = category),
       data = White_wine) +
  scale_color_brewer(type = 'qual',
                     guide = guide_legend(title = 'Rating'),
                     palette = 4) +
  geom_jitter(size = 2,
              alpha=1/5) +
  geom_smooth(method = "lm",
              se = FALSE,
              size=1)+
  coord_cartesian(xlim = c(0.985,1.005),ylim = c(0,250))+
  xlab('Density of wines')+
  ylab('Total Sulphur Dioxide')+
  labs(title='Graph between TSD and density compared with Category of wines')
```



Similar graphs were achieved when they were compared with Bad Wine crosses at 0.995 density and slope has a tremendous increases.Through this graph,we can infer that 100-150 g/dm^3 can help in preparation of excellent wine ranges.

Multivariate Analysis

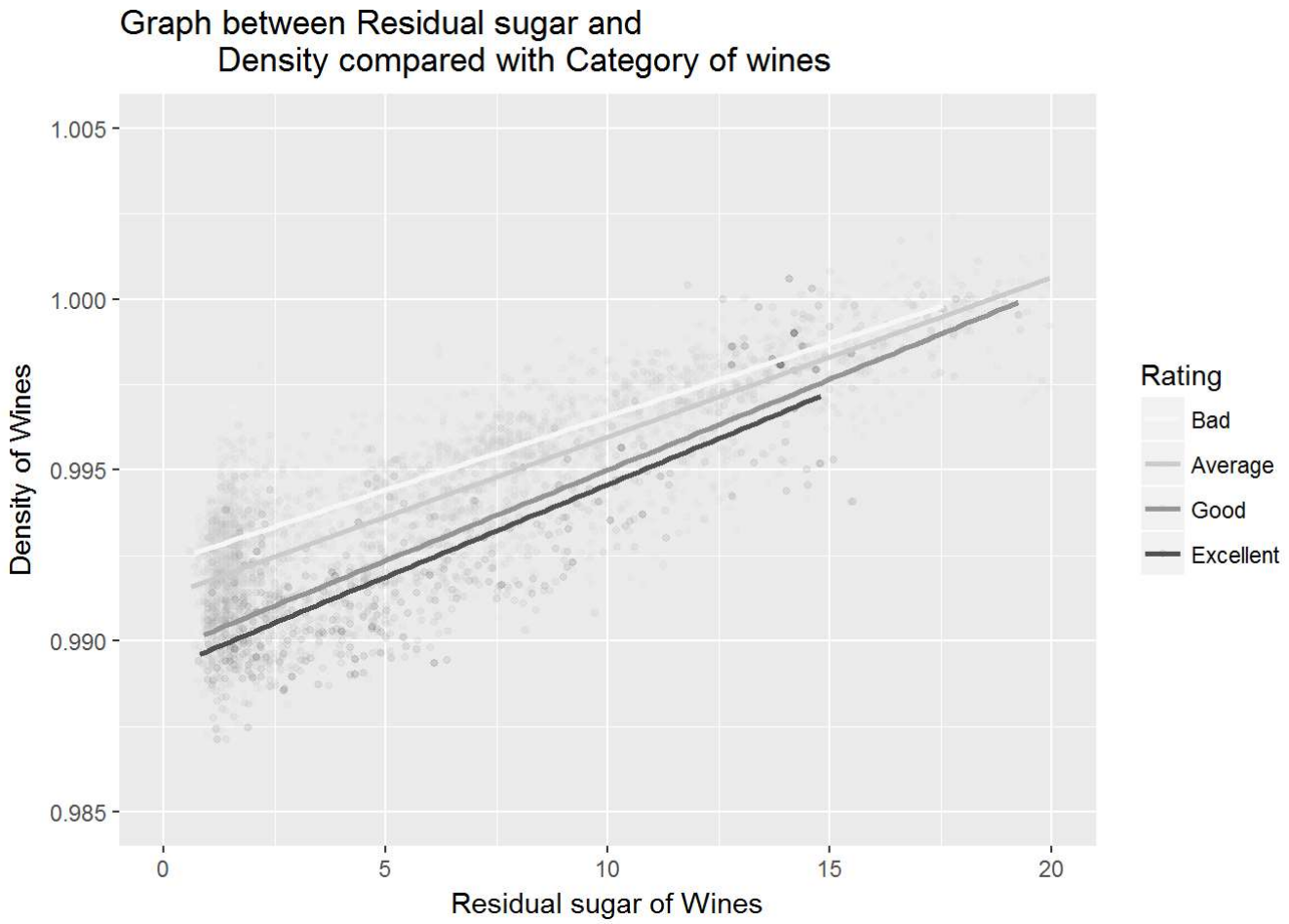
Talk about some of the relationships you observed in this part of the

investigation. Were there features that strengthened each other in terms of
looking at your feature(s) of interest?

Multivariate analysis gives interesting relationships when we enter the 3rd variable.Interesting features on bad wines were found as compared to excellent wines due to steep slopes of bad wines as they are increased.

Final Plots and Summary

Plot One

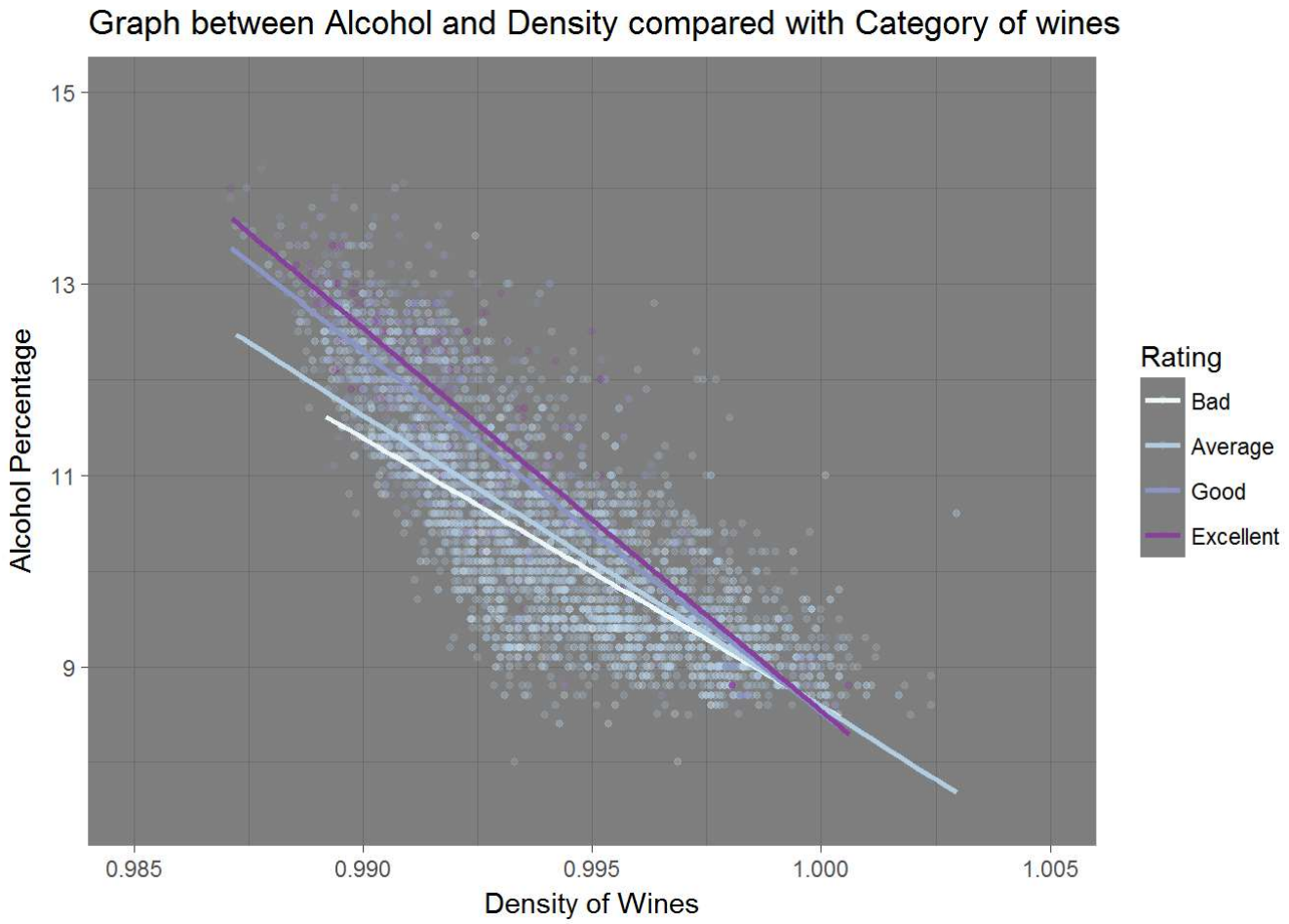


Description One

The graph between Residual Sugar and Density has similar trend for every categoryof rating made.The Density range was between 0.99-0.996 for excellent wine.

Secondly,All the four categories are converging towards one area thus reaching towards 1 gm/dm^3.

Plot Two



Description Two

The alcohol can be considered as main feature for selection of high quality wine. It has correlation with most of the factors hence for this case it is the graph between Density and Alcohol Percentage.It's the percentage between 10-13.2%.

Plot Three

Description Three

The graph between Alcohol percentage and Wine Quality shows the final trend between the graphs.Most of the Wine quality is between Average and Good wines while very less wines are present in bad and excellent ans based on that the approximate mean of excellent wine is 11.8%.Secondly,the mean of bad wine is 10.2%.Good wines and Excellent wines have alcohol percentage between 10.5% and 12.4%.

Reflection

This analysis was done on White wine dataset.White wine is a wine whose colour can be straw-yellow, yellow-green, or yellow-gold. It is produced by the alcoholic fermentation of the non-coloured pulp of grapes,which may have a skin of any colour. White wine has existed for at least 2500 years.

There are 4898 rows and 13 variables in this dataset to workupon hence require a mixture of univariate,bivariate and multivariate analysis on it.One of the problems faced in this dataset is that for Wine Quality the Average had the highest no. of data points hence when plotting with geom_point() function it was overcrowded with the points.Alcohol showed the highest no. of correlation and hence can be considered as a factor for creating a linear model.

However,this dataset only takes in data from one particular area and external factors like climate change may also affect the quality of wine and hence next time maybe it would be better to have a controlled experiment with a minimum of 10k data points.Collecting data from various regions may also help with the analysis.