TITLE by WHITE WINE ANALYSIS.

Exploring Dataset:

Dimensions and Names of variables of Dataset are as below. As the sample of 1st 6 rows is as below

```
## [1] 4898   13
```

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

About Dataset :

This data set contains 4,898 white wines with 11 variables on quantifying the chemical properties of each wine.

Structure of Dataset:

```
## 'data.frame':    4898 obs. of  15 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 ...
##  $ qual_factor        : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ qual_levels        : Ord.factor w/ 3 levels "low"<"medium"<..: 2 2 2 2 2 2 2 2 2 2 ...
```

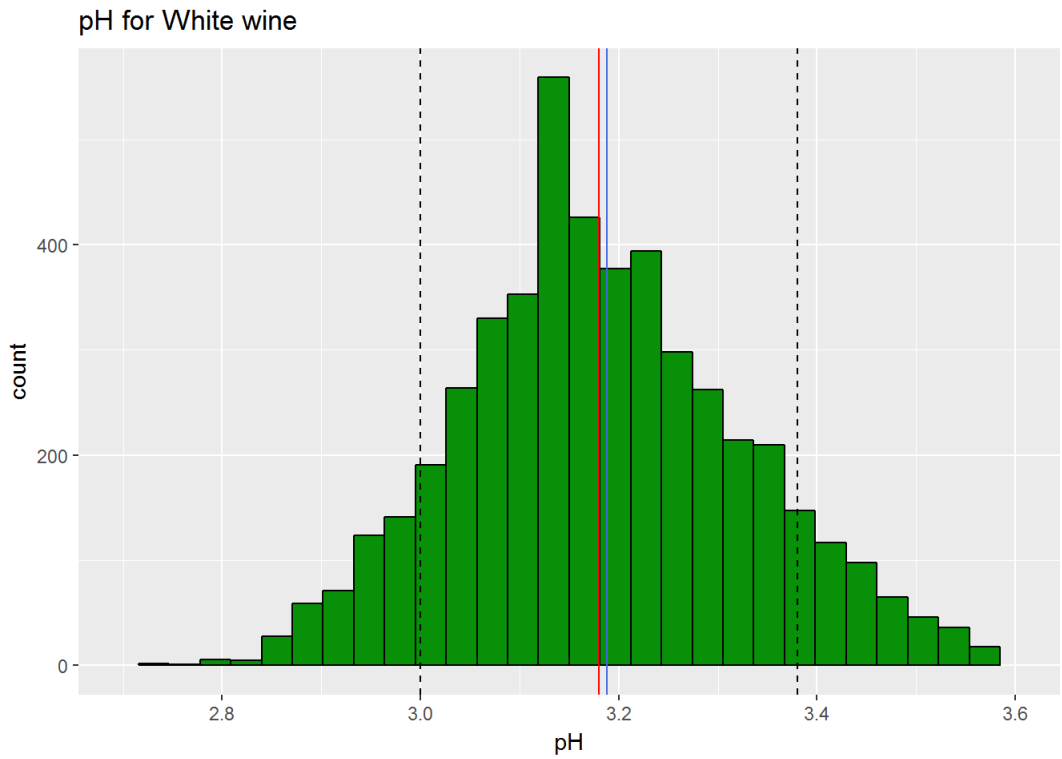
Summary Of Dataset.

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##
##     alcohol         quality      qual_factor qual_levels
##  Min.   : 8.00   Min.   :3.000   3:  20      low   : 183
##  1st Qu.: 9.50   1st Qu.:5.000   4: 163      medium:4535
##  Median :10.40   Median :6.000   5:1457      high  : 180
##  Mean   :10.51   Mean   :5.878   6:2198
##  3rd Qu.:11.40   3rd Qu.:6.000   7: 880
##  Max.   :14.20   Max.   :9.000   8: 175
##                                  9:   5
```

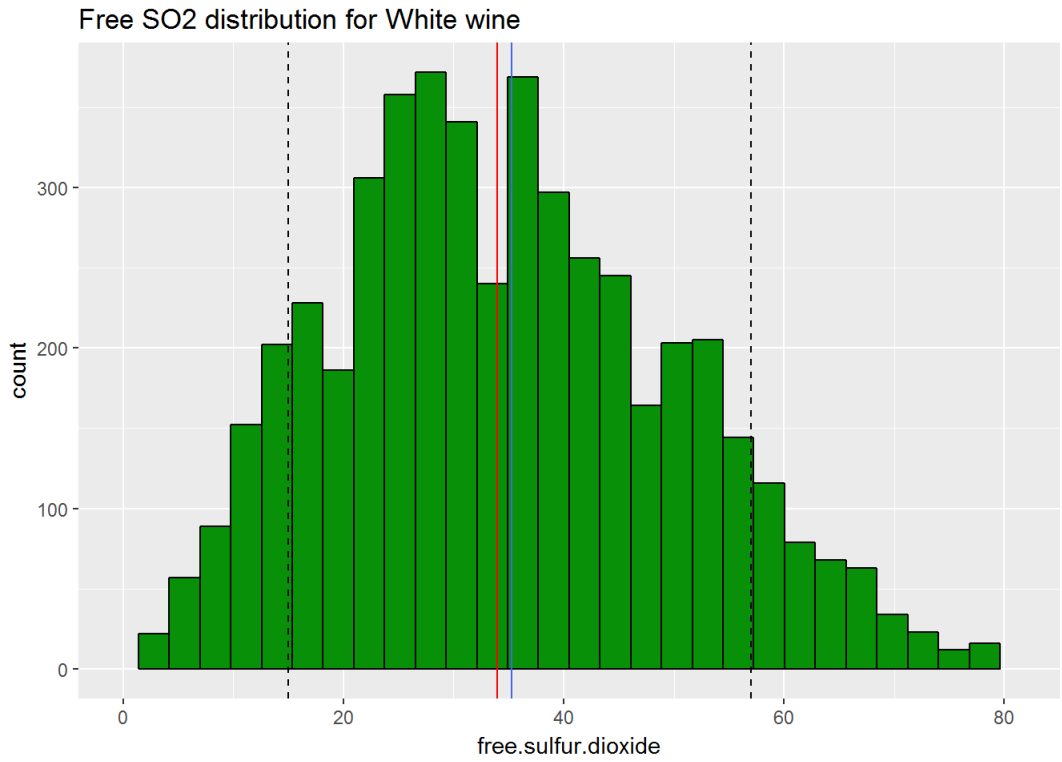
Univariate Plots Section

I have avoided using grid.arrange to grid all the below plots so that all the plots can be examined properly.

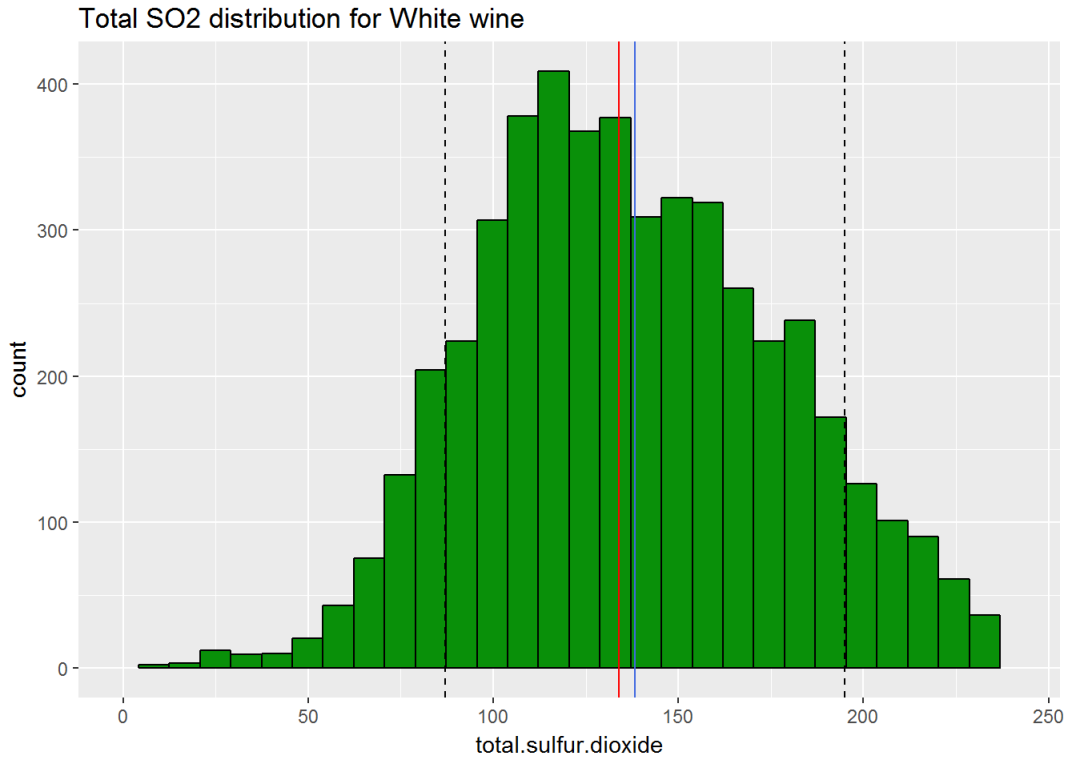
Mean is shown by blue vertical line and Median by red vertical line. Dotted vertical lines at beggining and end show the range of 10-90%. I have plotted the mean, median , and range of 10 - 90% in which the maximum items fit



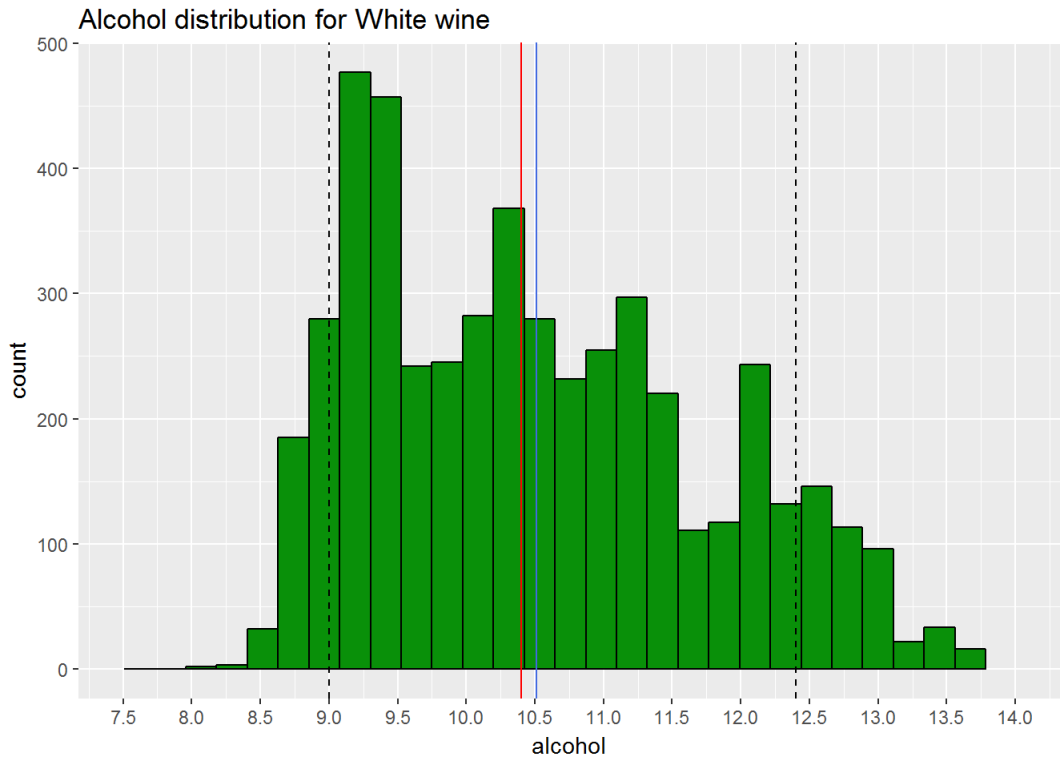
1. pH varies in the range of 2.7 - 3.8 , with maximum items falling in the range of 3 to 3.4. Mean and Median fall around 3.1



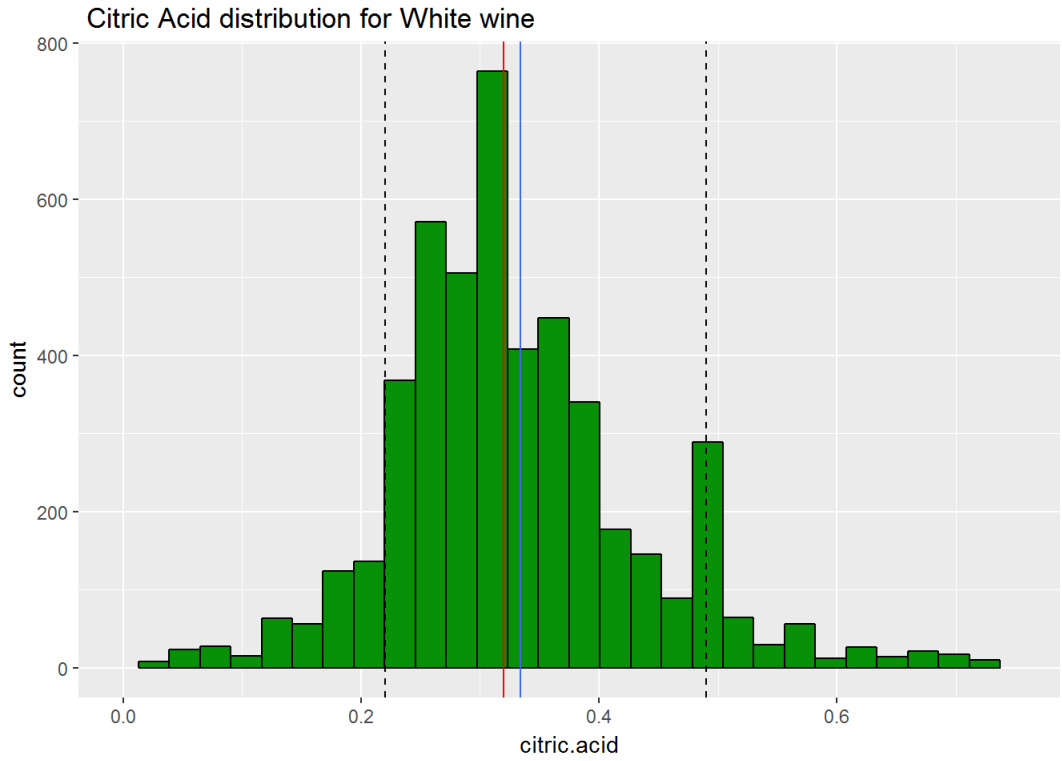
2)Free Sulphur Dioxide varies in the range of 0-280 with a long tail at left,after cutting off the tail we get, maximum items falling in the range of 15 to 55. There is a drop in between at 35 and before 50.Mean and Median fall around 35.



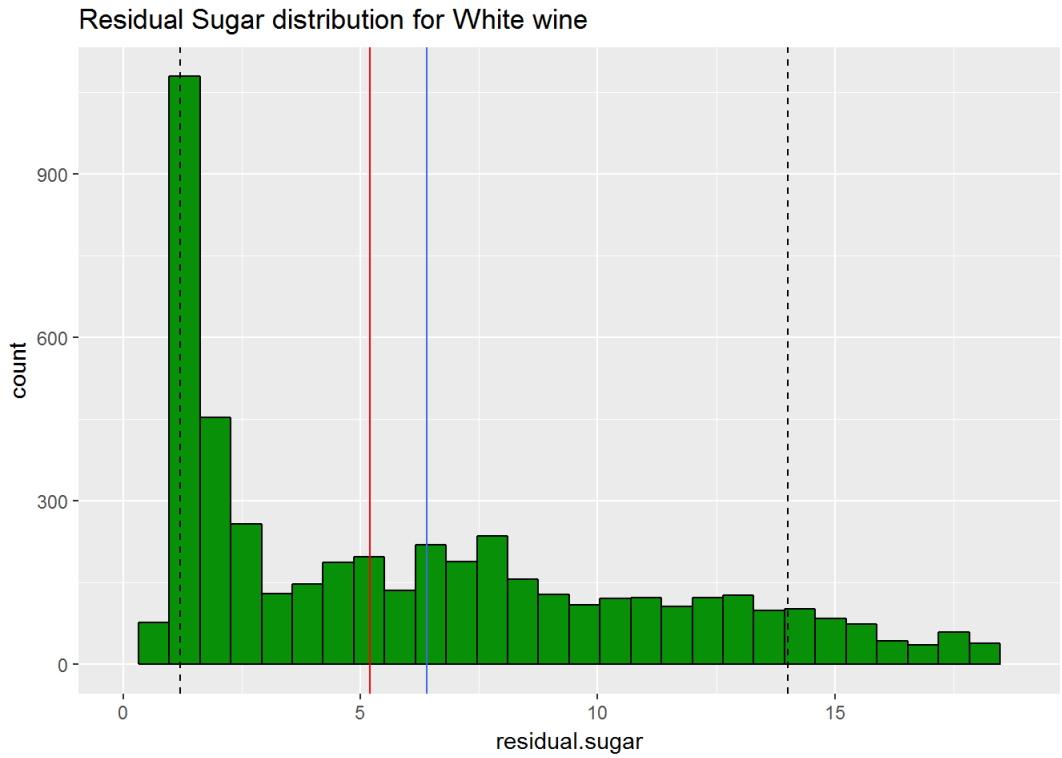
3)Total Sulphur Dioxide varies mostly in the range of 10-440 , with maximum items falling in the range of 80-190. Mean and Median fall around 135-140



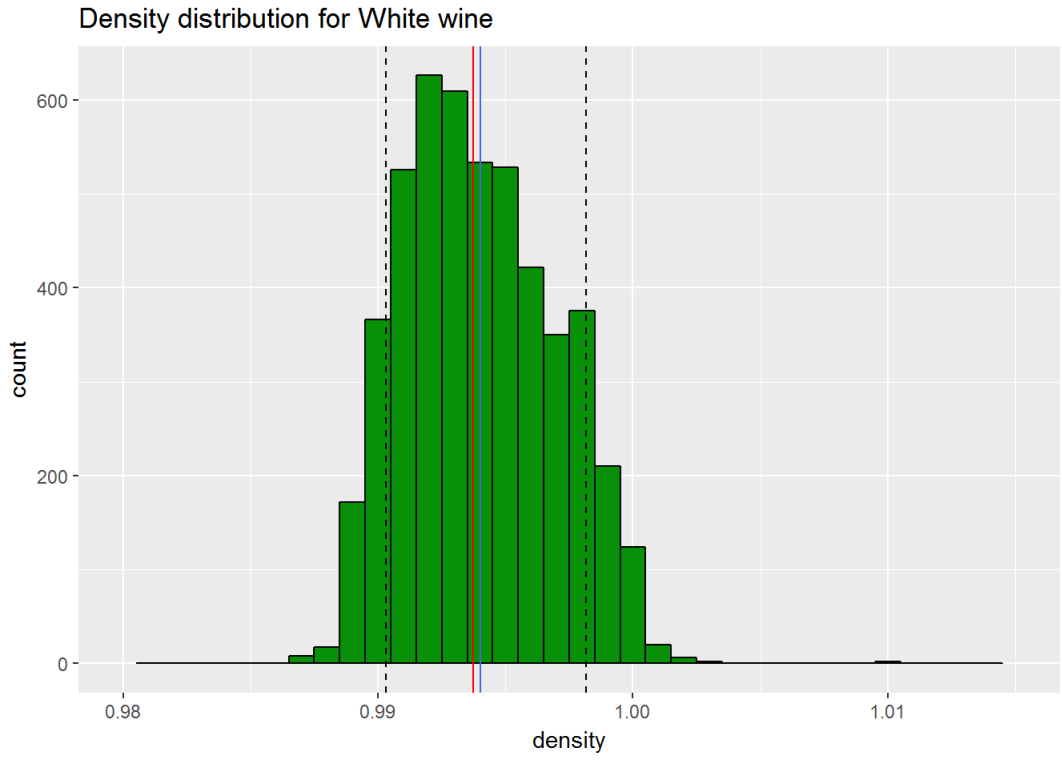
4)Alcohol in white wine varies mostly in the range of 8.5-13.5 , with maximum items falling in the range of 9 - 12.4. It has many drops in values around 9.5-10 , 10.5- 11 ,11.5-12 , also has a declining histogram. Seems that 9-9.5% is the standard range.Mean and Median fall around 10.5.



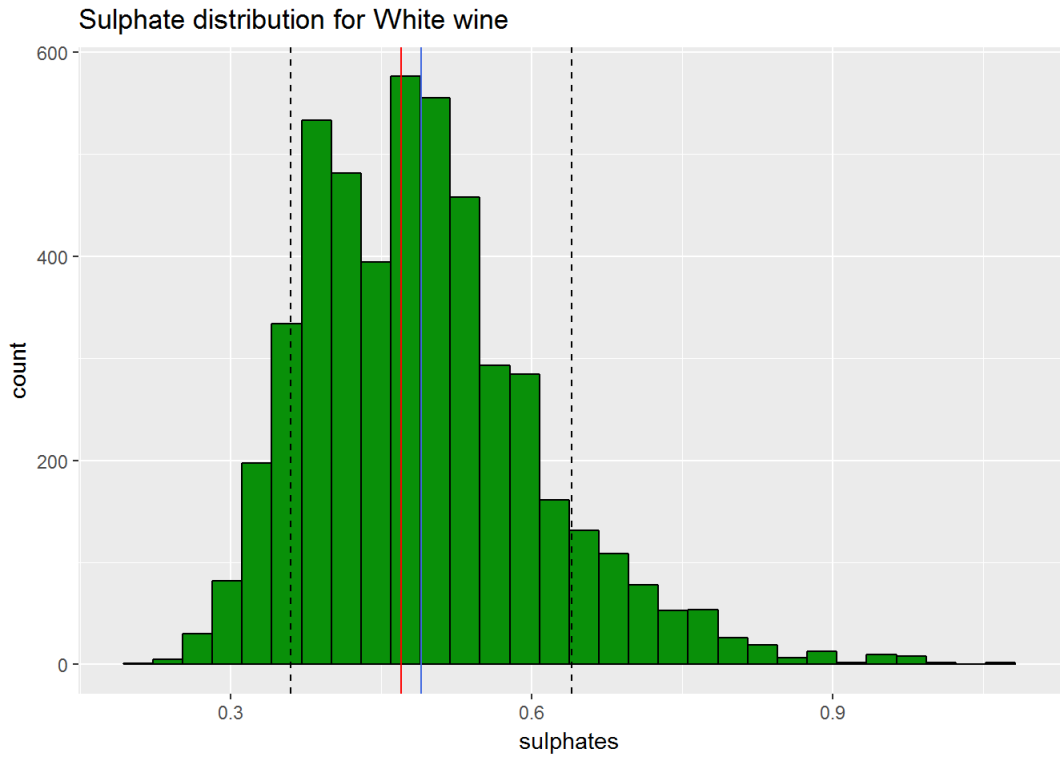
5)Citric Acid in white wine mostly varies in the range of 0-0.6 , with maximum items falling in the range of 0.24 to 0.4,we can see a peak nearly at 0.5 which inspite of fall after 0.4 , has included items in 0.4-0.5 into 10-90% percent range.There are again drop at both sides of 0.3.



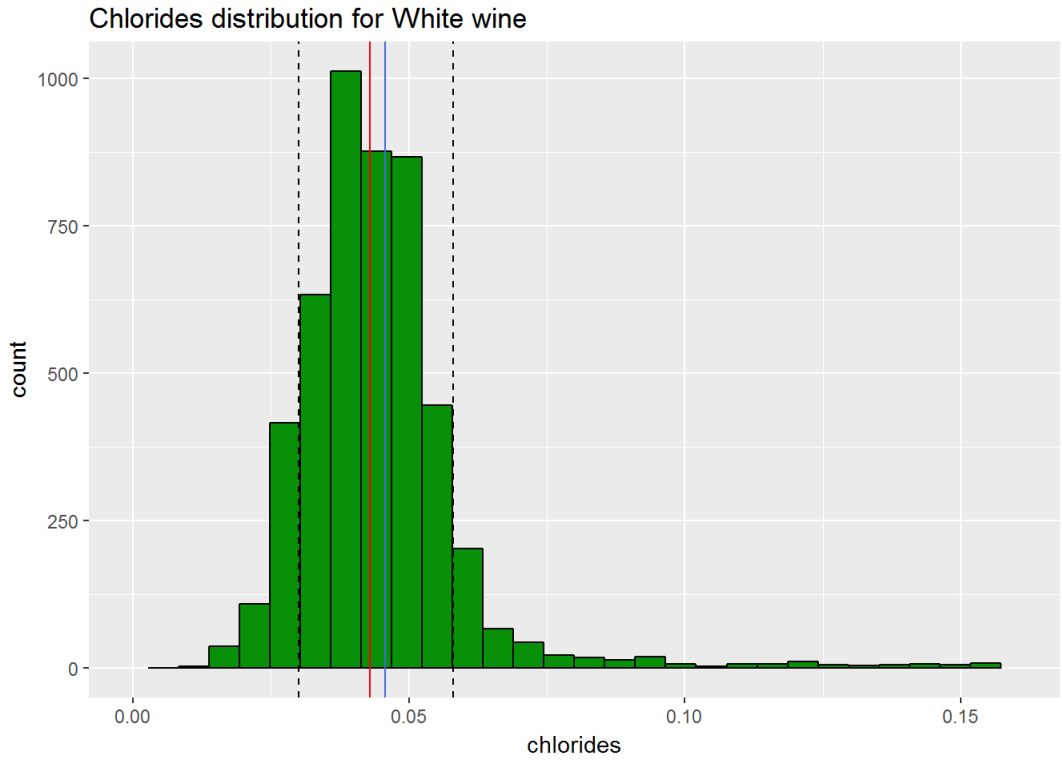
6)Residual Sugar in white wine mostly varies in the range of 0-17 , with maximum items falling in the range of 1.2 to 13,with median around 5 and mean around 6.It has an overshoot between 0-1.



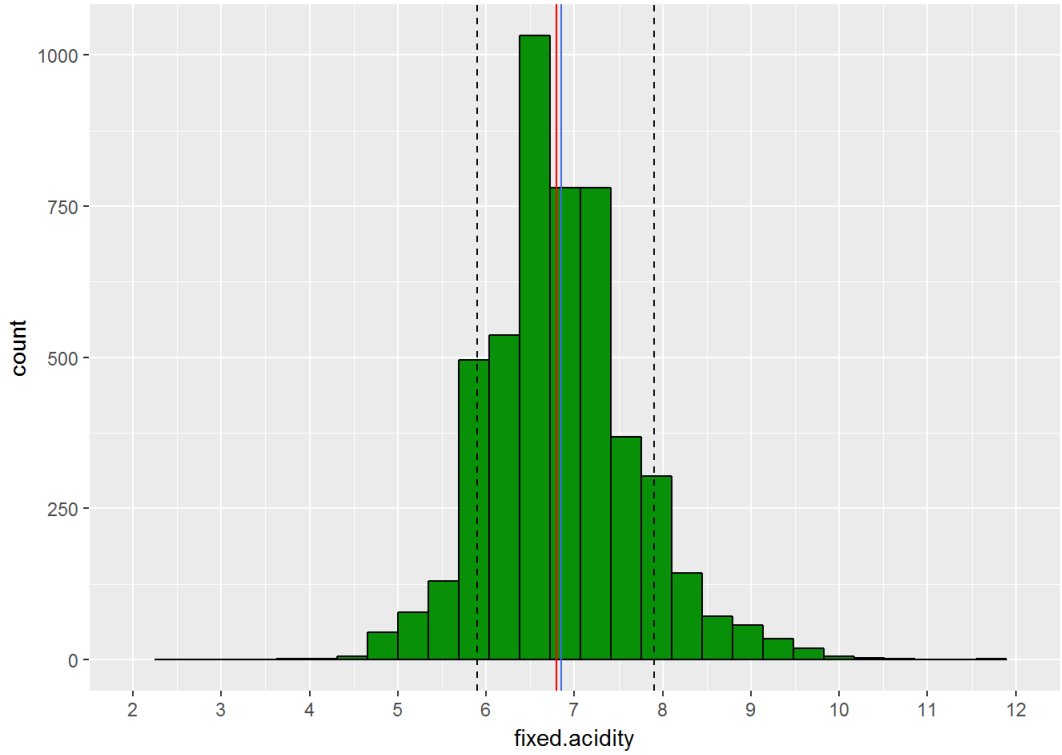
7)Density in white wine mostly varies in the range of 0.98-1.01 , with maximum items falling in the range of 0.99 - 1, It has mean and median very close exactly in between 0.99 and 1.



8)Sulphate in white wine mostly varies in the range of 0.3-0.9 , with maximum items falling in the range of 0.35 - 0.62.It has mean and median very close in between 0.45 and 0.5.It has a drop nearly at 0.4.

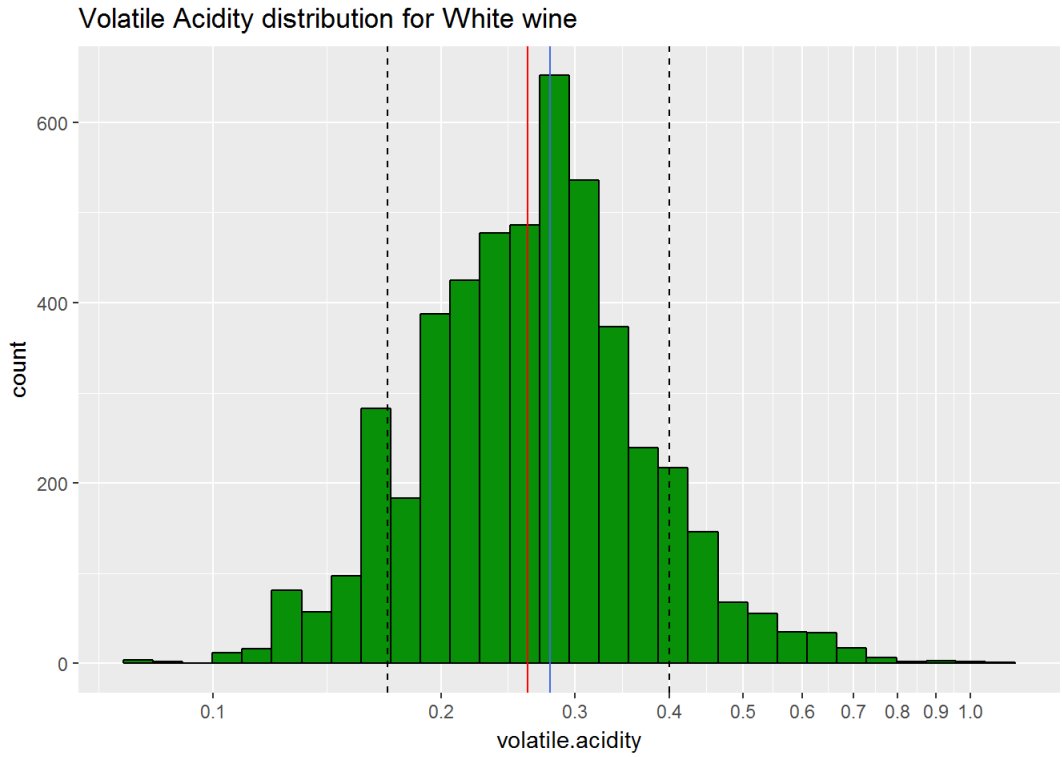


9)Chlorides in white wine mostly varies in the range of 0.3-0.9 , with maximum items falling in the range of 0.35 to 0.62.It has mean and median very close in between 0.45 and 0.5.It has a drop nearly at 0.4.




```
## $title
## [1] "Fixed Acidity distribution for White wine"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

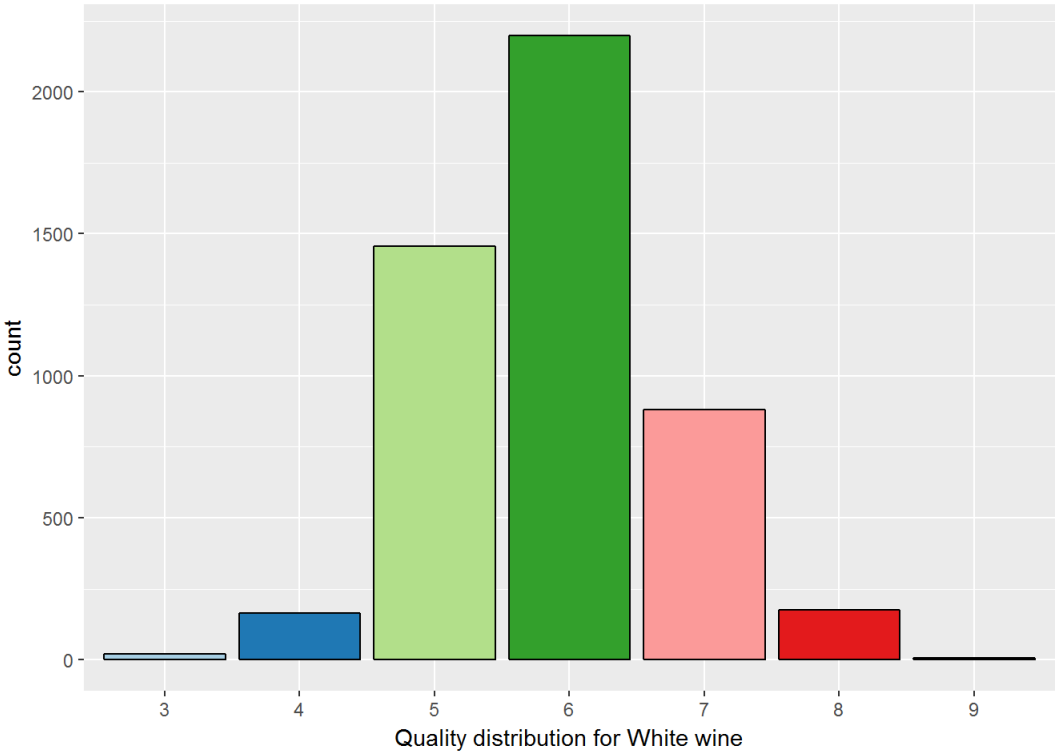
10. Fixed acity in white wine mostly varies in the range of 4 -10 , with maximum items falling in the range of 6-8. We see an overshoot at 6.5 ,seems that most of the wines are has the fixed acity of 6.5.



11. It can be seen that maximum volacity acidity is 0.2-0.35 , we see the mean and median are at 0.28 and 0.29.

```
ggplot(white_wine, aes_string(x = white_wine$qual_factor, fill = white_wine$qual_factor)) +
    geom_histogram(binwidth = 1, color = "black", show.legend  = FALSE, stat="count") +
    scale_fill_brewer(palette="Paired") +
    xlab("Quality distribution for White wine")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



12. We can see that most of the wines fall in quality range of 5 and 6. There are lesser numbers assiciated with 7 and 8 , and least with 3.

Univariate Analysis

What is the structure of your dataset?

It have a dataset of variables of composition of white wine with 11 variables .

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

Yes , I think pH , sulphur dioxide contents , quality, density , alcohol are few main featres

What other features in the dataset do you think will help support your Quality , density , pH and Sulphure dioxide .

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

Yes I created two variables: Quality which is an integer variable can be factored and also can be divided into levels as below: qual_factor containes factored quality data. Quality variables is divided in qual_levels as low, medium and high. In later section of Multivarient analysis I have also created a variable unfit , that indicates how pH levels and Total sulfurdioxide can affect the wine.

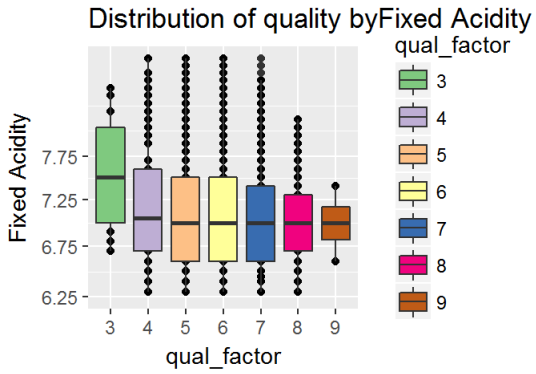
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

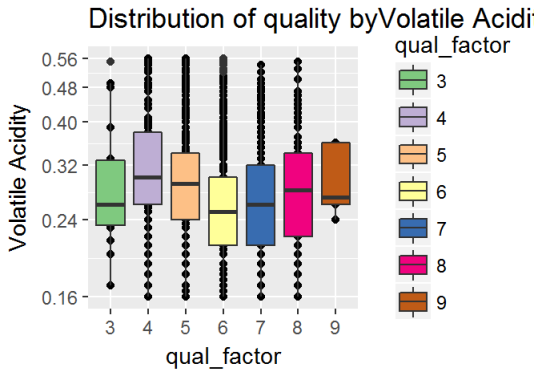
A histogram of all variables is plotted with proper scaling using xlim, scale_x_continuos , scale_x_log10 , or log10.

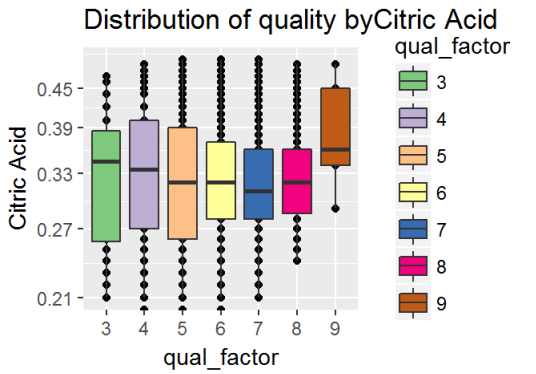
Bivariate Plots Section

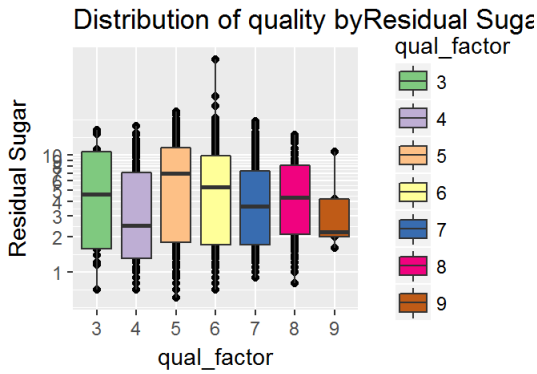
I am interested in looking at the distribution of quality on various parameters . Again I have avoided to apply plots to grid so that each plot

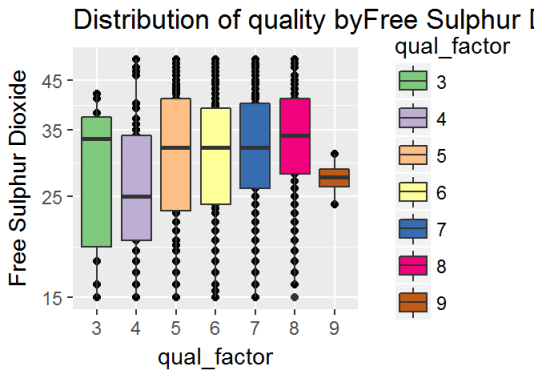
can be examined properly. Also I would like to see and plot correlations between different parameters.

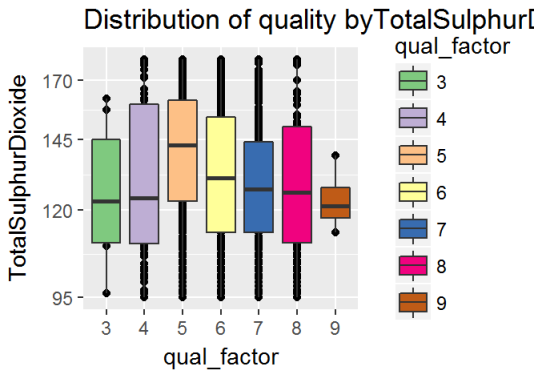


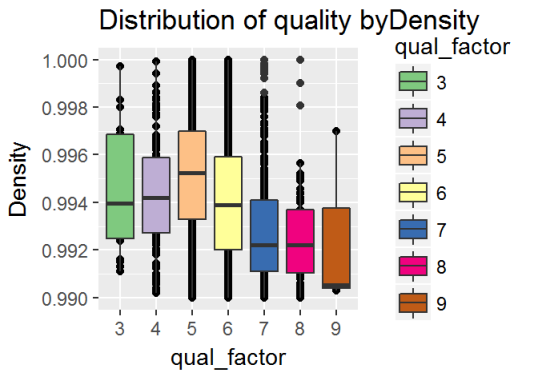


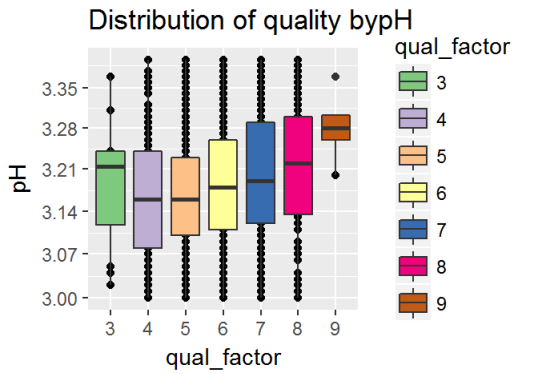


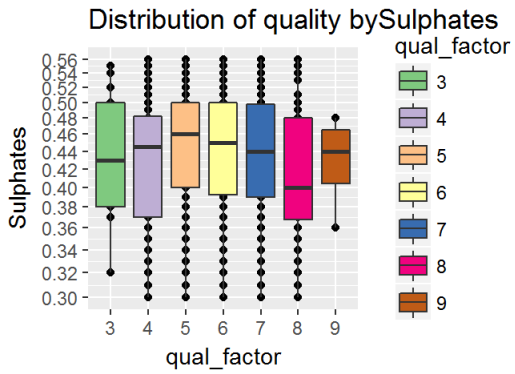


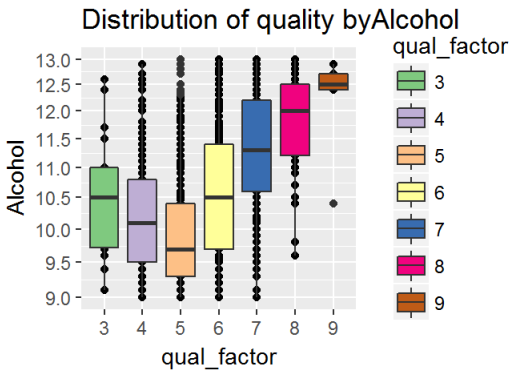


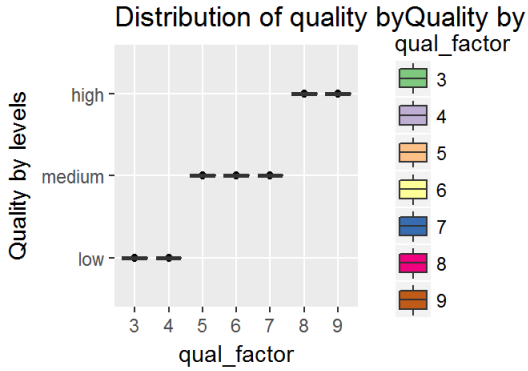






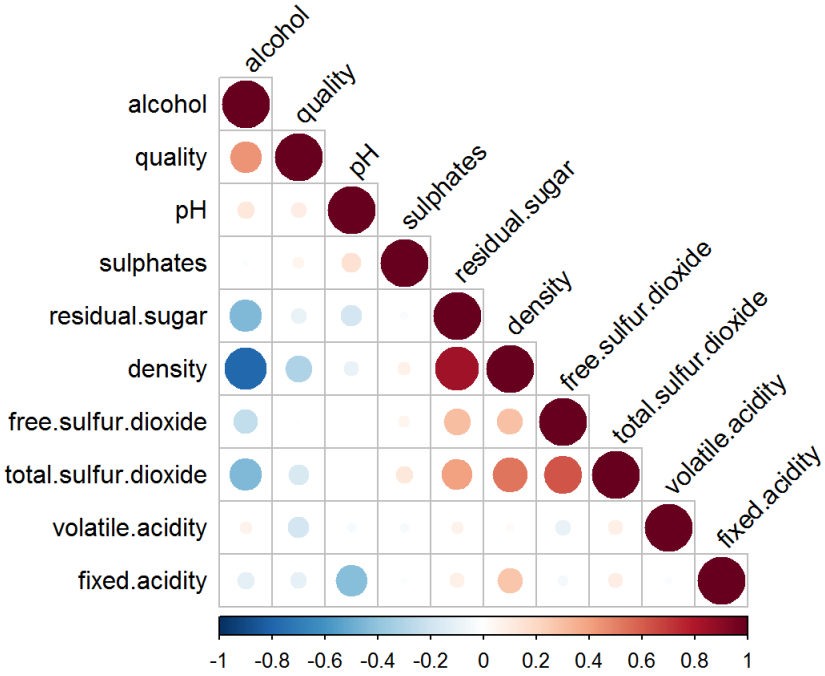






From above plots we can see that maximum quality factor of 6 is observed with each items . Then quality factor of 5 is observed . Quality factor of 7 and 8 are observed less than above two mentioned and quality factor of 4 is ver very small. We see that quality factor of 3 and 9 are negligible. Also it can seen that for higher fixed acidity, for citric acid range 0.27 - 0.39 also for higher sulphates we get lower quality wine . We can see from the “Distribution of quality by Quality levels” plot that qual_factor 3 & 4 are low quality wine, qual_factor 5, 6 & 7 are medium quality wine, qual_factor 5 & 6 are high quality wine

Below matrix shows a simple representation of effect of correlation between variables.




```
## $r
##                       alcohol quality       pH sulphates residual.sugar
## alcohol                     1
## quality                  0.44       1
## pH                       0.12   0.099        1
## sulphates              -0.017   0.054     0.16         1
## residual.sugar          -0.45  -0.098    -0.19    -0.027             1
## density                 -0.78   -0.31   -0.094     0.074          0.84
```

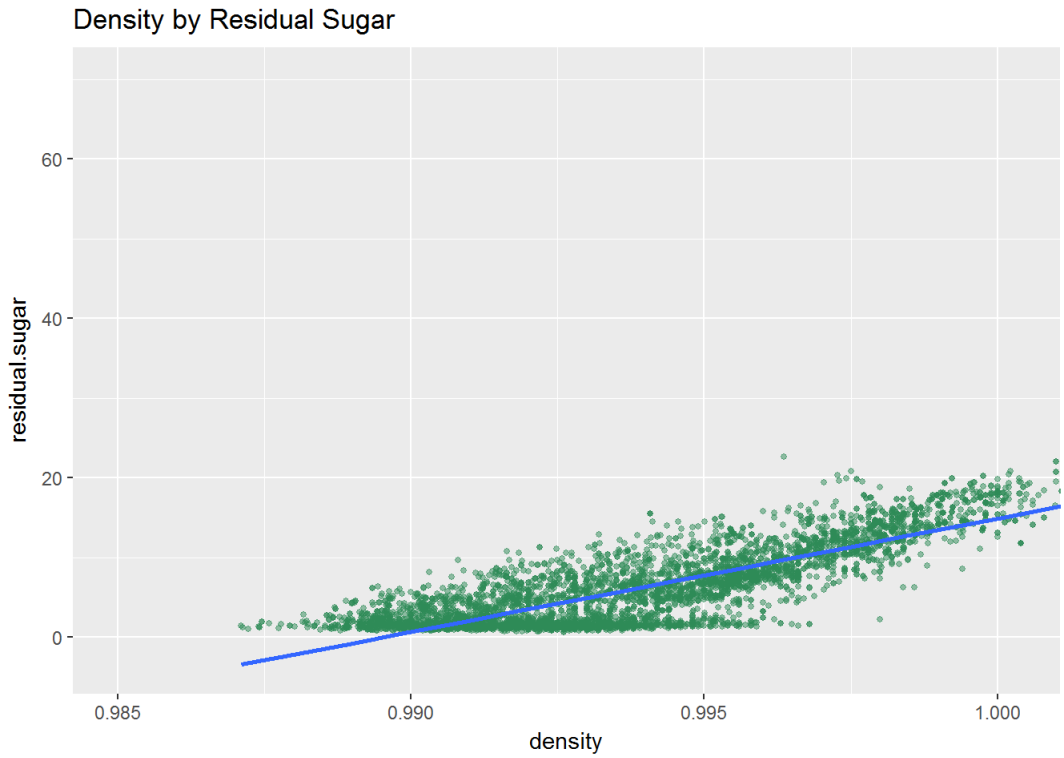
```
## density                 0.670    0.361    0.091     0.071           0.01
## free.sulfur.dioxide    -0.25  0.0082 -0.00062     0.059            0.3
## total.sulfur.dioxide   -0.45   -0.17   0.0023      0.13            0.4
## volatile.acidity       0.068   -0.19   -0.032    -0.036          0.064
## fixed.acidity          -0.12   -0.11    -0.43    -0.017          0.089
##                      density free.sulfur.dioxide total.sulfur.dioxide
## alcohol
## quality
## pH
## sulphates
## residual.sugar
## density                    1
## free.sulfur.dioxide     0.29                   1
## total.sulfur.dioxide    0.53                0.62                   1
## volatile.acidity       0.027              -0.097               0.089
## fixed.acidity           0.27              -0.049               0.091
##                      volatile.acidity fixed.acidity
## alcohol
## quality
## pH
## sulphates
## residual.sugar
## density
## free.sulfur.dioxide
## total.sulfur.dioxide
## volatile.acidity                   1
## fixed.acidity                 -0.023             1
##
## $p
##                       alcohol  quality       pH sulphates residual.sugar
## alcohol                     0
## quality              5.6e-226        0
## pH                    1.5e-17  3.1e-12         0
## sulphates                0.22  0.00017  4.8e-28         0
## residual.sugar       1.2e-243  7.7e-12  8.4e-43     0.062              0
## density                     0 1.7e-107  5.3e-11   1.8e-07              0
## free.sulfur.dioxide   9.6e-71     0.57     0.97   3.4e-05       8.8e-102
## total.sulfur.dioxide 1.5e-241    7e-35     0.87   3.1e-21       4.2e-189
## volatile.acidity      2.1e-06  4.7e-43    0.026     0.012        6.7e-06
## fixed.acidity         2.1e-17  1.5e-15 4.8e-215      0.23        4.3e-10
##                      density free.sulfur.dioxide total.sulfur.dioxide
## alcohol
## quality
## pH
## sulphates
## residual.sugar
## density                    0
## free.sulfur.dioxide  2.1e-98                   0
## total.sulfur.dioxide       0                   0                   0
## volatile.acidity       0.058               1e-11             3.9e-10
## fixed.acidity           1e-79             0.00054             1.7e-10
##                      volatile.acidity fixed.acidity
## alcohol
## quality
## pH
## sulphates
## residual.sugar
## density
## free.sulfur.dioxide
## total.sulfur.dioxide
## volatile.acidity                   0
## fixed.acidity                   0.11             0
##
## $sym
##                      alcohol quality pH sulphates residual.sugar density
## alcohol              1
## quality              .       1
## pH                                   1
## sulphates                               1
## residual.sugar       .                              1
## density              ,       .                      +              1
## free.sulfur.dioxide
## total.sulfur.dioxide .                              .              .
```

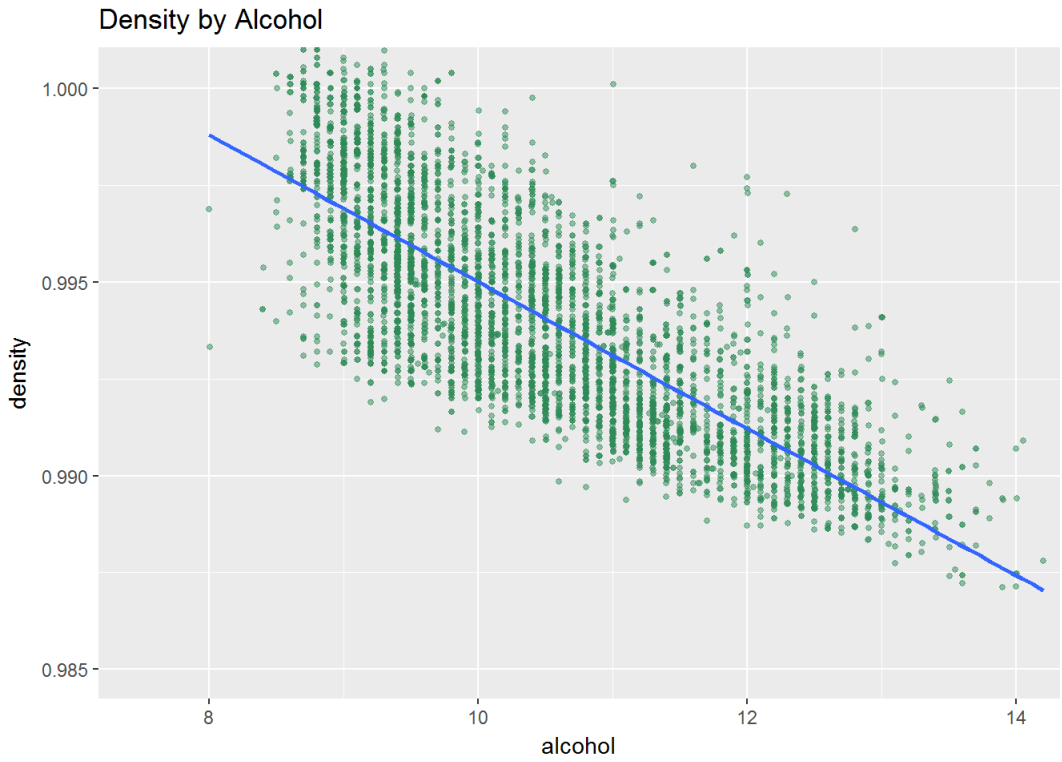
```
## volatile.acidity
## fixed.acidity                        .
##                      free.sulfur.dioxide total.sulfur.dioxide
## alcohol
## quality
## pH
## sulphates
## residual.sugar
## density
## free.sulfur.dioxide  1
## total.sulfur.dioxide ,                   1
## volatile.acidity
## fixed.acidity
##                      volatile.acidity fixed.acidity
## alcohol
## quality
## pH
## sulphates
## residual.sugar
## density
## free.sulfur.dioxide
## total.sulfur.dioxide
## volatile.acidity     1
## fixed.acidity                         1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

Conclusion drawn from correlation table and corelation matrix : Simple matrix representation above show that there is maximum correlation between density~Residual Sugar, minimum correlation between density ~ alcohol. Again maximum correlation between density~total sulfurdioxide, minimum correlation between total.sulfur.dioxide~ alcohol. Alcohol also shows minimum correlation with total sulfurdioxide, free sulfurdioxide, density, chlorides. There is no use of proving correlation between total sulfurdioxide and free sulfurdioxide as free sulfurdioxide is a part of total sulfurdioxide.

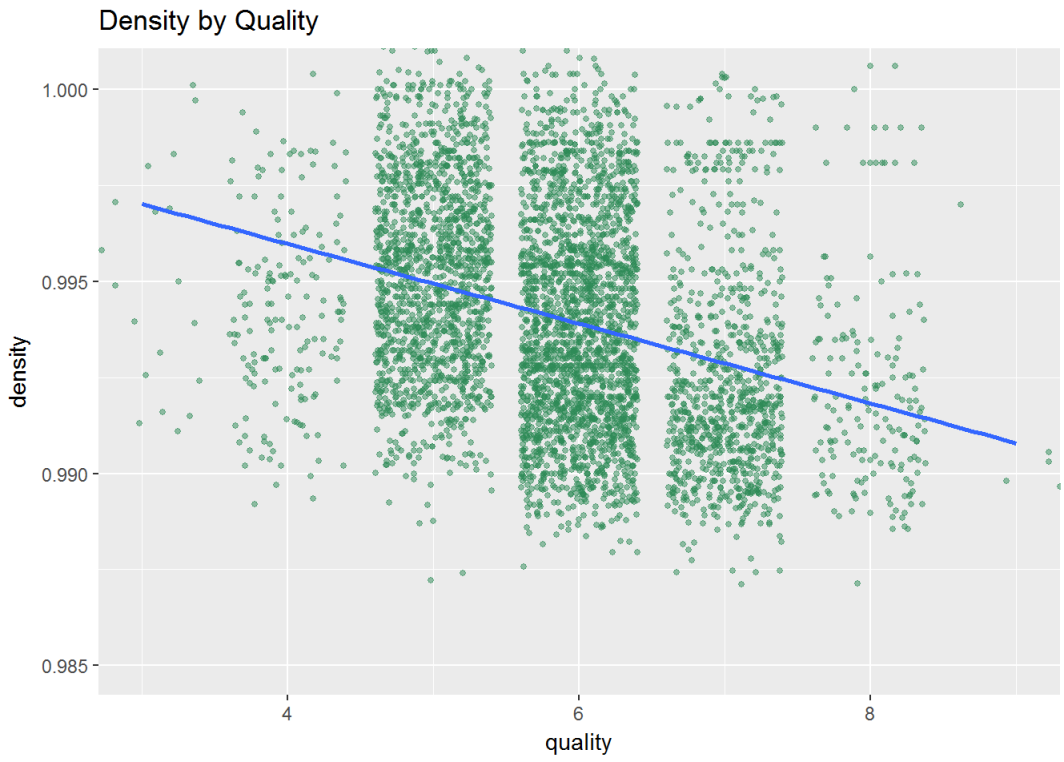
Below parameters show maximum correlation Density and Alcohol:-0.78 , Density and Residual Sugar: 0.84, Alcohol and Sugar : -0.45

Let us see the plots with maximum and minimum correlation.

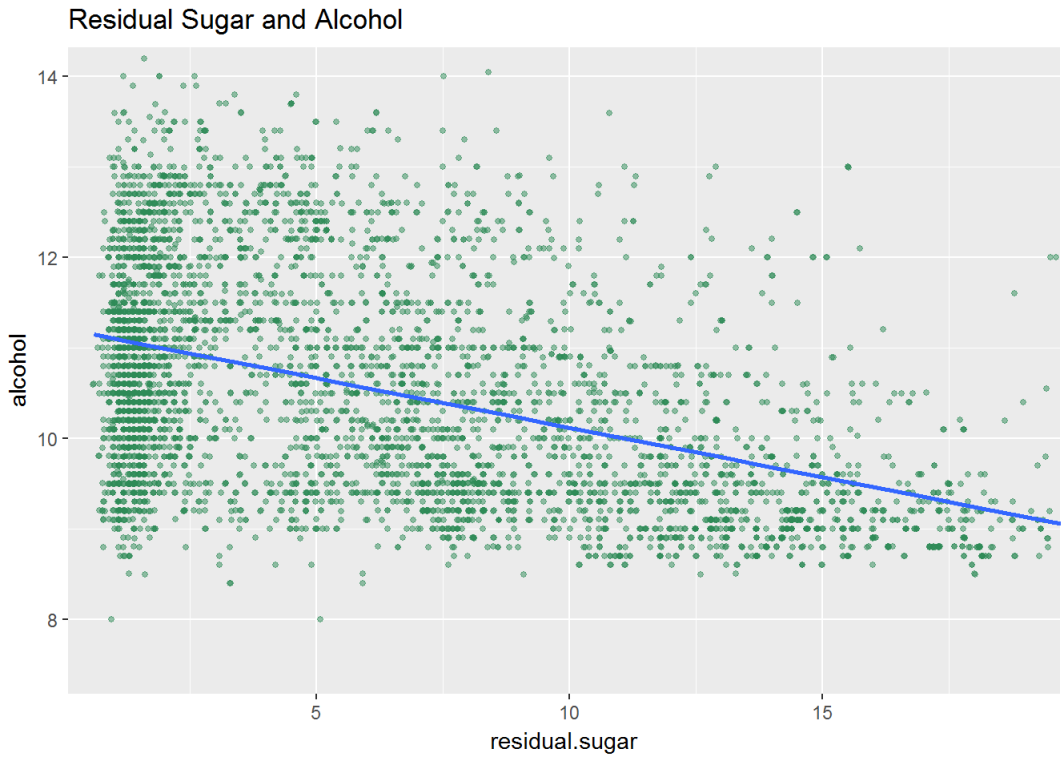




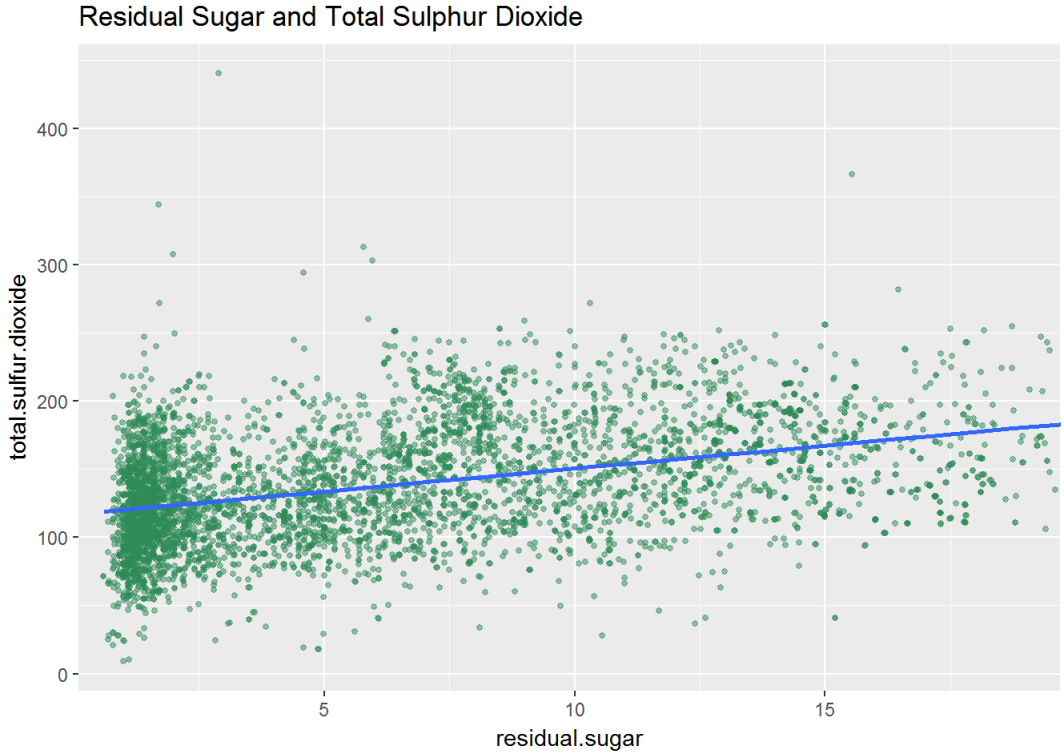
Hence density is affected by alcohol and residual sugar : It increases with Residual Sugar and lower the alcohol higher is the density which is very obvious.



Hence Low quality wines result from high density .



There is some negative correlation between alochol and sugar. Sugar is an essential component in the production of wine. During alcoholic fermentation, yeast feeds on the sugar found in grape juice and converts it to ethyl alcohol, or ethanol, and carbon dioxide. The amount of sugar fermented determines the wine's alcohol level and the amount of residual sugar left in the wine.



There is also quite a strong correlation between SO2 and sugar. That's because SO2, sulphur dioxide, plays a protective role in the wine against the phenomena of oxidation, oxidase enzyme action (enzymes that oxidize the polyphenols in wine), and the control of microbial populations in yeasts and bacteria (antiseptic effect).

Bivariate Analysis

I have selected to plot histograms of parametes of minimum and maximum correlation and quality wise all other parameters.

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?

I saw strong correlation between , density and sugar , and less correlation between density and quality which is surprising for me.

Did you observe any interesting relationships between the other features
Yes , relationship between SO2 and sugar, and relationship between alcohol and sugar .i.e That means sweeter the wine less alcohol.

What was the strongest relationship you found?

Density and Residual Sugar: 0.84

Multivariate Plots Section

Case no 1: Fit and Unfit Wine

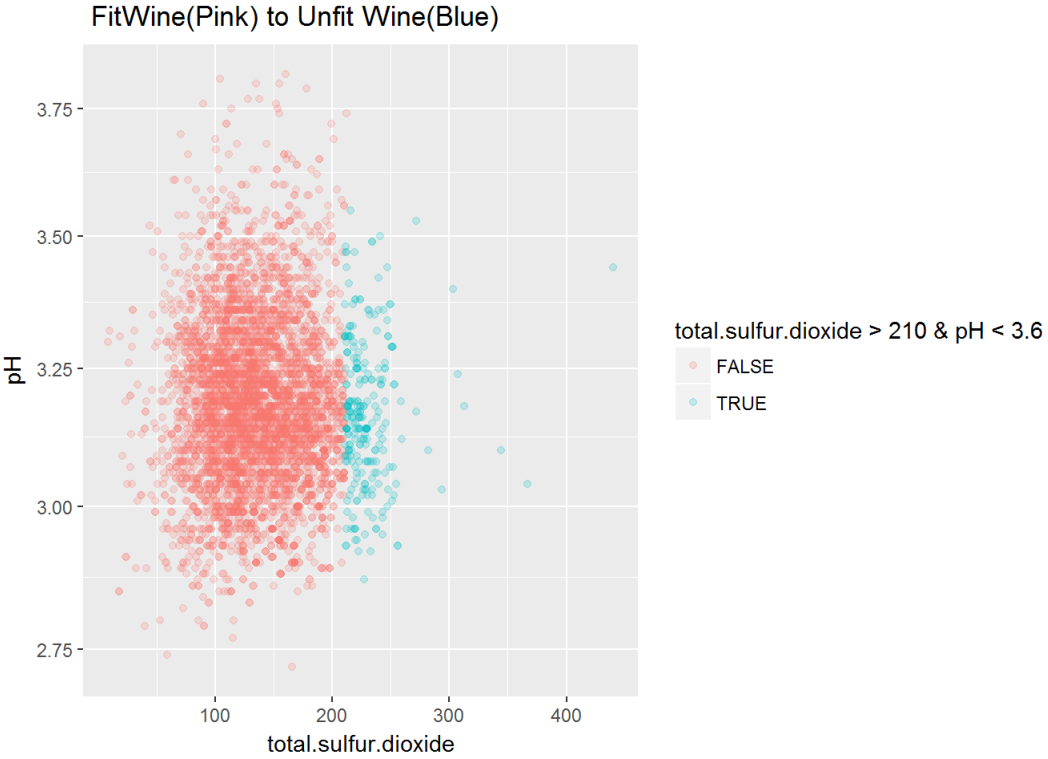
Considering below facts: As we know 0 (very acidic) to 14 (very basic) The EU limit for sulphur in dry white wine is 210mg/l.(mg/l = mg/dm^3) But in some cases like : Wines with lower acidity need more sulfur than higher acidity wines. At pH 3.6 and above, the sulfites needed is much higher because it's an exponential ratio. Hence for an unfit wine , pH is less than 3.6(becomes acidic) and sulphur is more than 210 mg/l. Sources of information :http://winefolly.com/tutorial/sulfites-in-wine/ http://www.scientistlive.com/content/total-sulphite-wine

Below is the count of unfit wines as per above information.

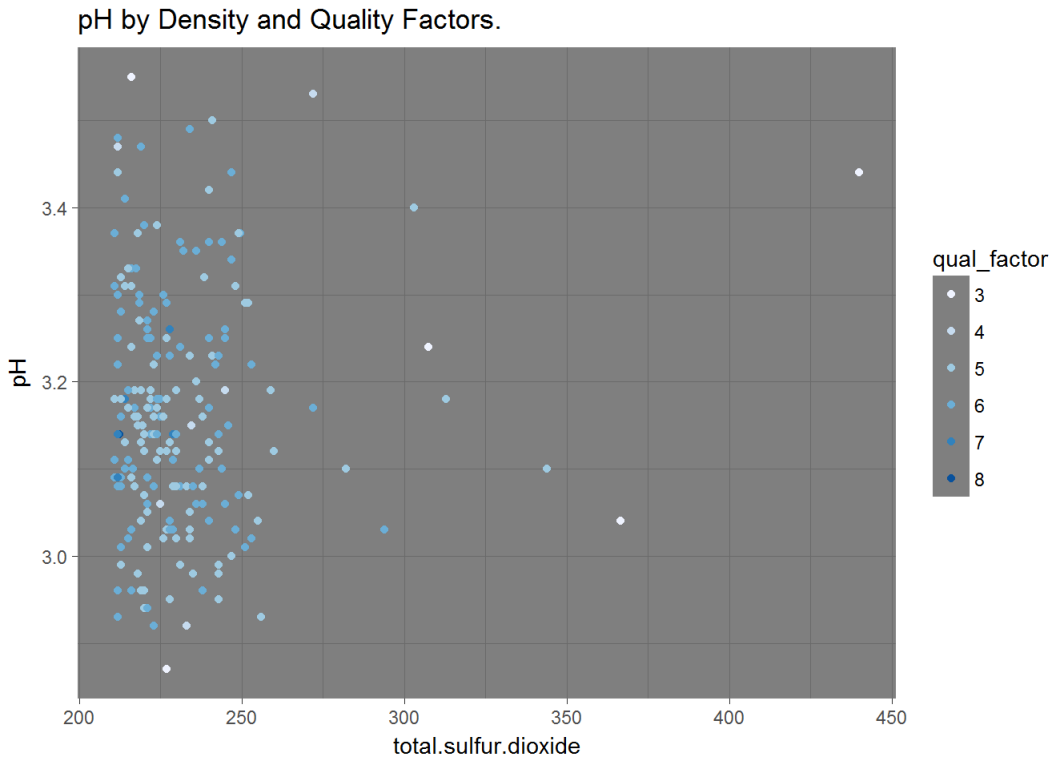
```
unfit_phwine <- white_wine[white_wine$total.sulfur.dioxide > 210 &
                           white_wine$pH < 3.6 ,]
dim(unfit_phwine)
```

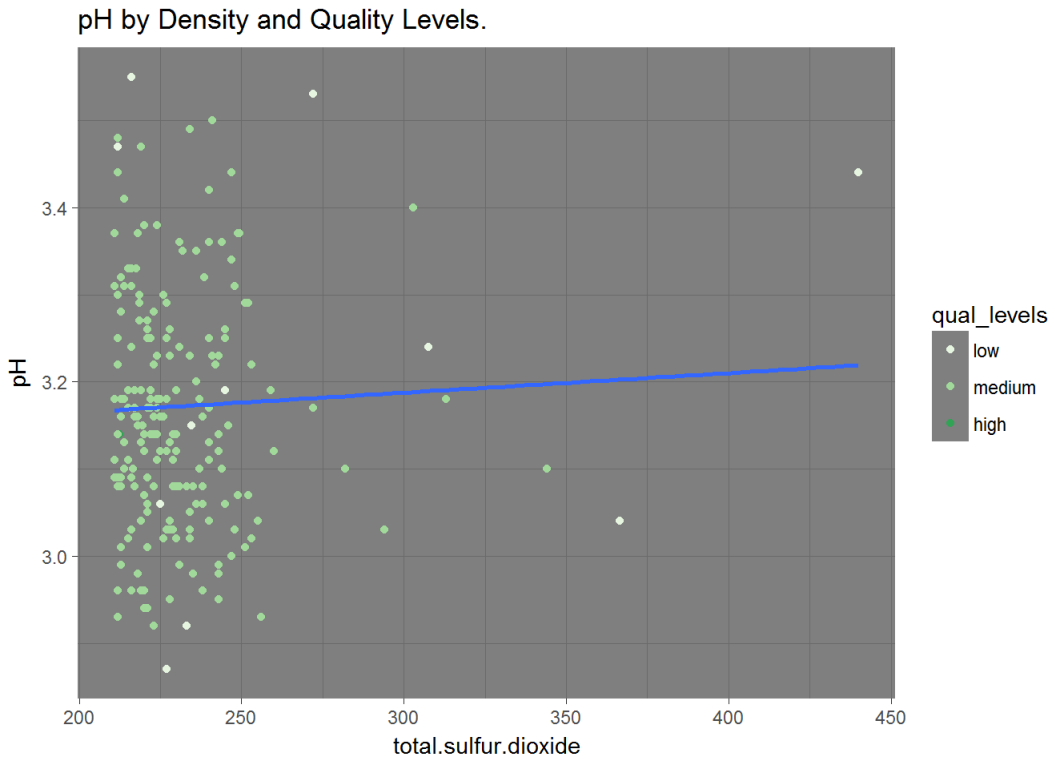
```
## [1] 261  15
```

Below is a plot that shows fit and unfit wine.



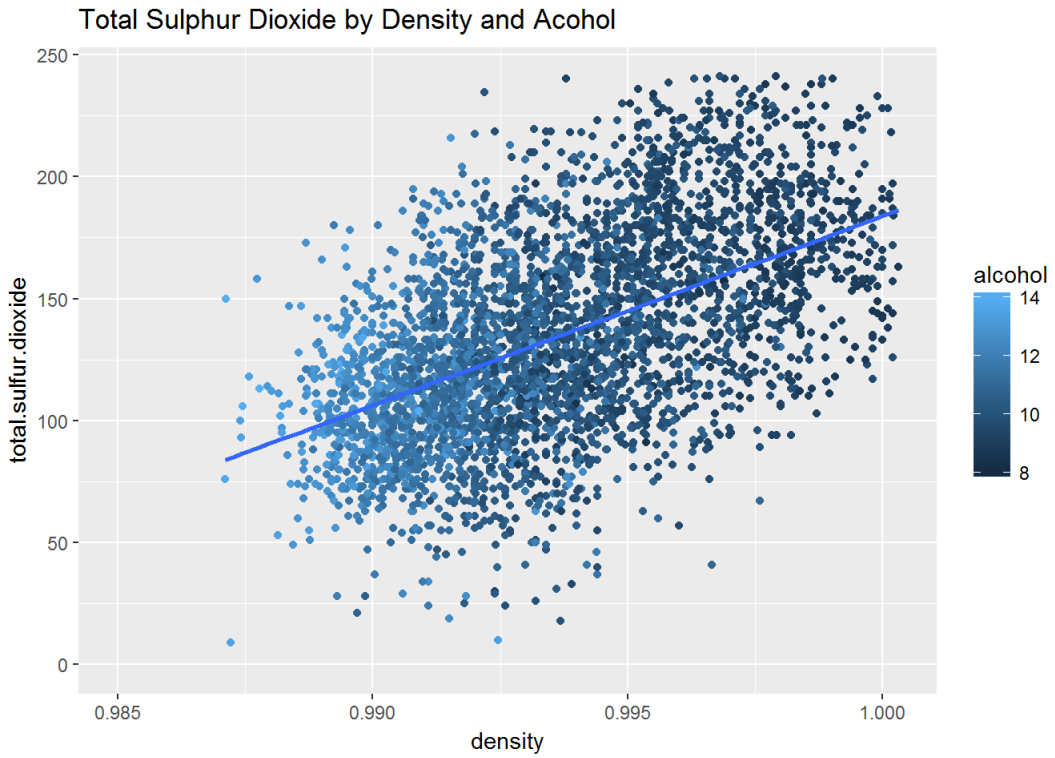
I wish to dig more into quality ratings of unfit wines, hence I plotted the two plots below.





Maximum ratings beyond 225 should have been of low category, but they are in medium category. This can be a drawback of the dataset.

Case no 2: Total Sulphur Dioxide by Density and Acohol



Hence we can see from above plot that total sulphur dioxide(T.D.S) can be one of the reason of increase in density but it is inversely true for (T.S.D) and alcohol.

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?

I observed the quality of an unfit wine.Unfit wine according to some sources of reseach is the the one with high sulphur contents with high pH . I looked how residual suagr and alcohol are realted. I looked how total sulphur dioxide is related to both density and dcohol.

Were there any interesting or surprising interactions between features?

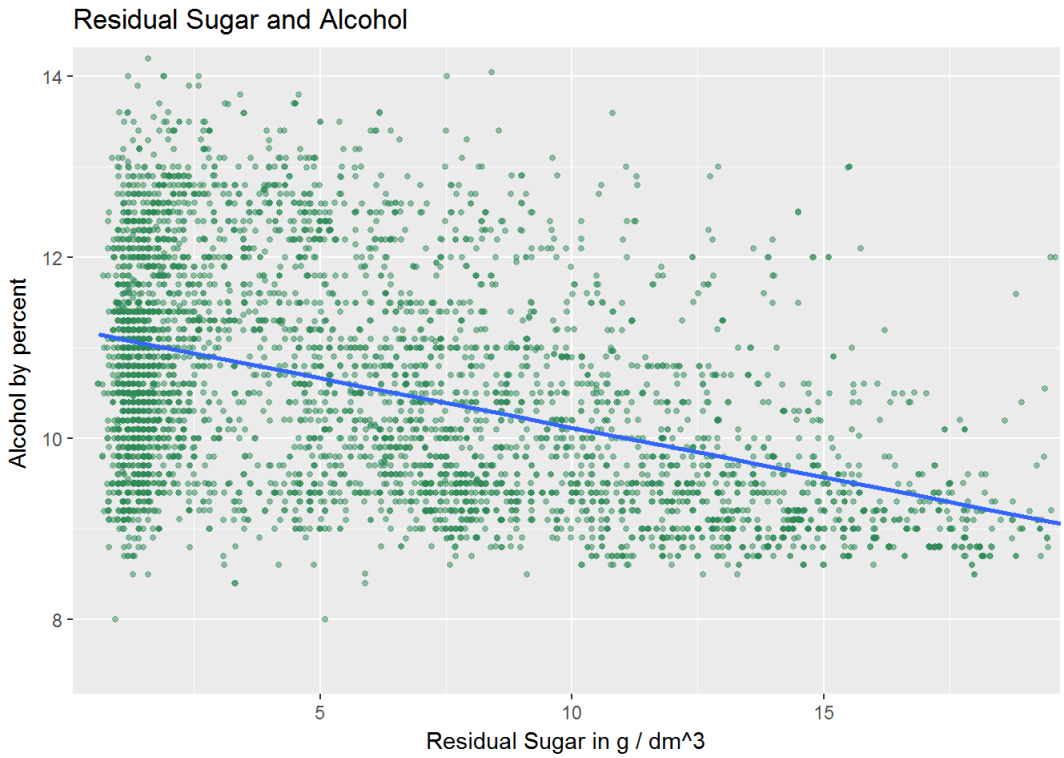
In the first analysis , it most of the wine samples beyond 225 shulfur content should have had low quality , but it shows medium quality of wine.

OPTIONAL: Did you create any models with your dataset? Discuss the

Based on the exploratory data analysis, the linear regression model doesn't provide any meaningful data.

Final Plots and Summary

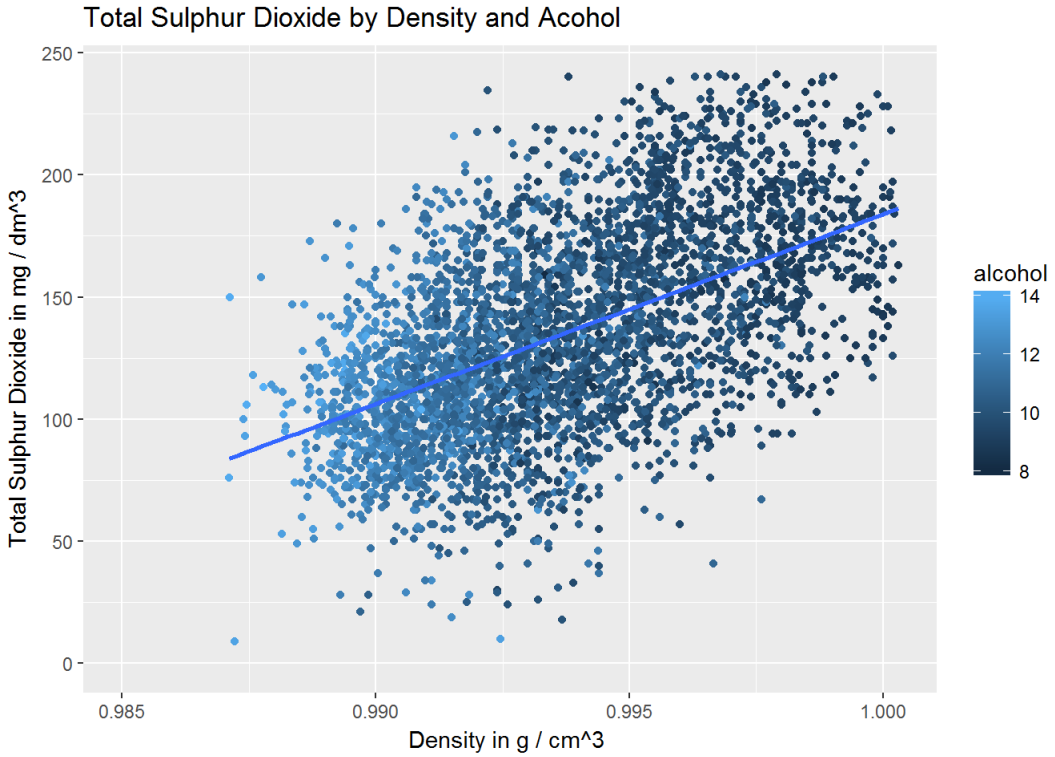
Plot One



Description One

There is some correlation between alochol and sugar. And that's fair enough: sweet wines, whether moelleux (Sweet: 12-45 g/l of sugar) or liquoreux (Fortified: >45 g/l sugar) wines are where the fermentation is interrupted before all the grape sugars are converted into alcohol: this is called Mutage or fortification[1].That means sweeter the wine (more sugar in the wine) - less alcohol.

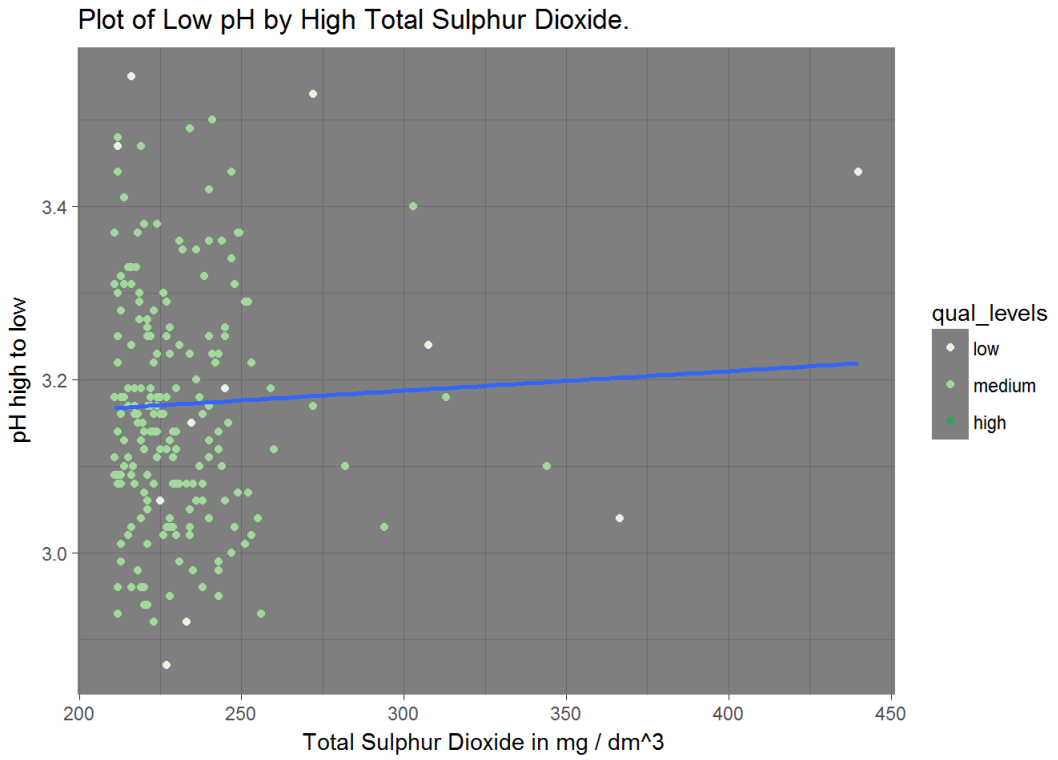
Plot Two



Description Two

Hence we can see from above plot that total sulphur dioxide(T.D.S) can be one of the reason of increase in density but it is inversely true for (T.S.D) and alcohol. This can be due to wines are formed when the fermentation is interrupted before all the grape sugars are converted into alcohol: this is called Mutage or fortification.And due to more sufur used for preservation of wine which increases density. Also with less alcohol the density has increased .

Plot Three



Description Three

As seen from above two plots the data set gives ratings to unfit wines of a low and medium “qual_levels”“, which is in the range of 3,4 and 5-7 of”qual_factor“respectively. But maximum ratings atleast beyond 225 should have been of low category. This can be a drawback of the dataset.

Reflection

This wine dataset helped me understand the basic characteristics of wine . Learning to identify wine characteristics helps to identify what you like about a wine.

1)Believe it or not, many dry wines can have a hint of sweetness to carry a larger impression of Body. If you find a wine you like has residual sugar, you may enjoy a hint (or a lot!) of sweetness in your wine.Sweetness is indicator of good wine.

2. Good wine tend to have more alcohol. Alcohol probably creates the flavor or sugar (as an alternative to alcohol) kills it Good wine tend to have lower density.

3)Sugar and SO2 increases density of wine , but higher the density lower the quality

4. We know from the description, everyone uses SO2 but too much of it might harm the wine and increase density

Limitations of the study : The case study represents data from particular region with, the data set had opinions from very less people from particular region , obviously there tastes concentrated. There should have data from more regions to get robust summary of white wines from many places.

Successes and Difficulties is problem set : I was successful in finding out fit and unfit wine count , also I successfully created quality by factor and levels. I had difficulty finding out why sugar and alcohol had negative correlation , finally with google search I got a useful piece of information. The link for the same is mentioned in the references.

References: http://winefolly.com/tutorial/sulfites-in-wine/ http://www.scientistlive.com/content/total-sulphite-wine https://winemakermag.com/501-measuring-residual-sugar-techniques