red_wine

karthik

December 25, 2015

Overview of red wind data set

Dimention of red wind data set

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

Various variable involed in the data set

```
##  [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"
```

Data Structure

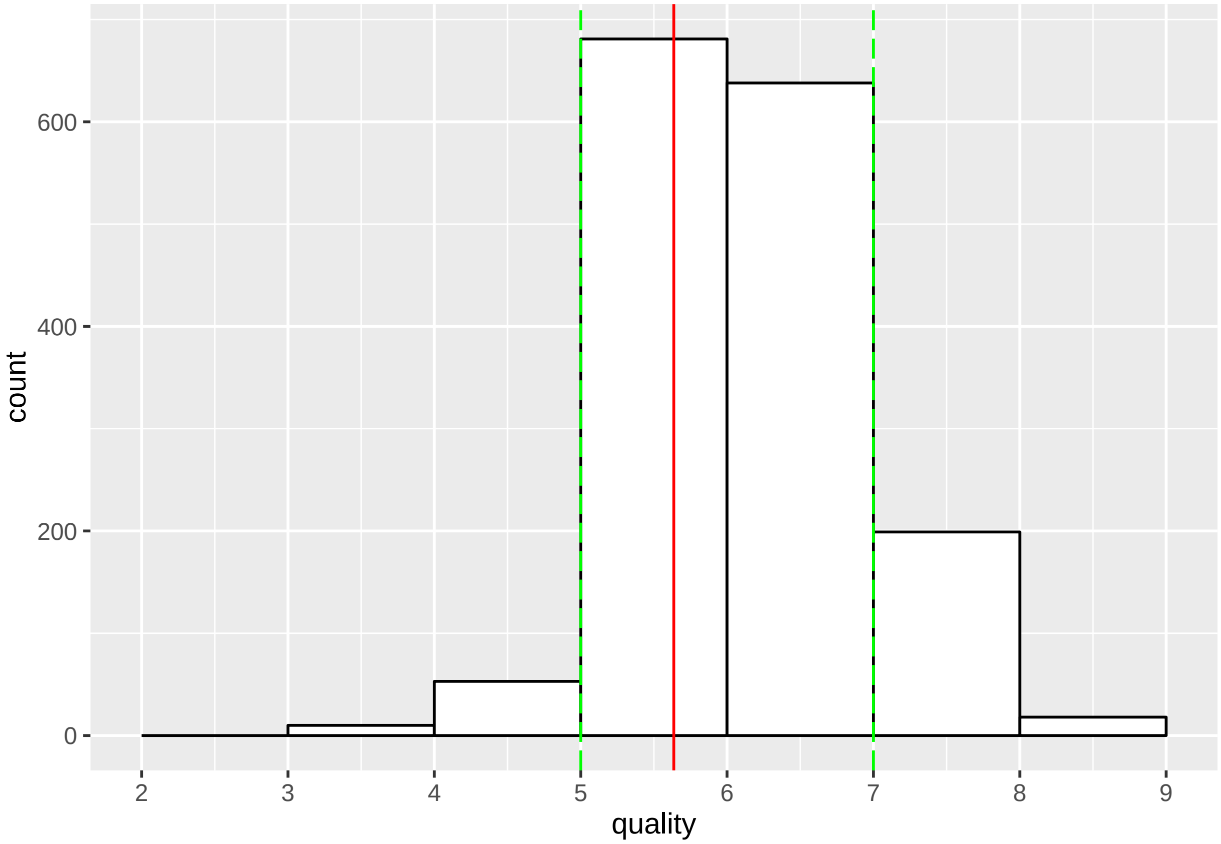
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  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 ...
```

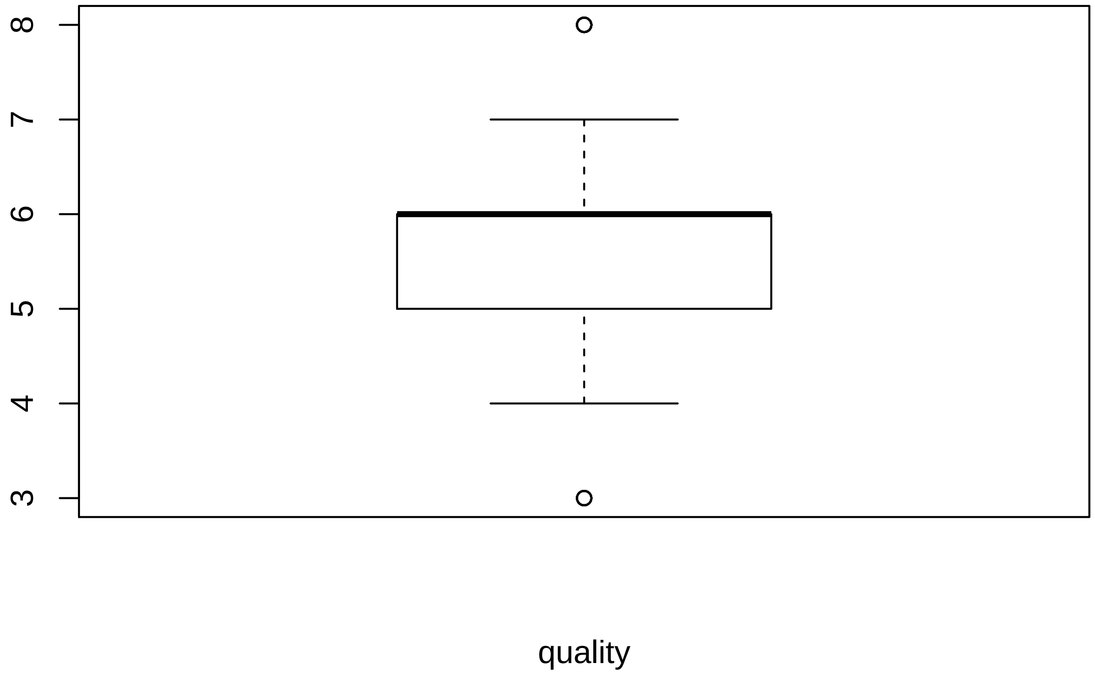
Dependent Variable's (qualtiy) statistics Summary

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

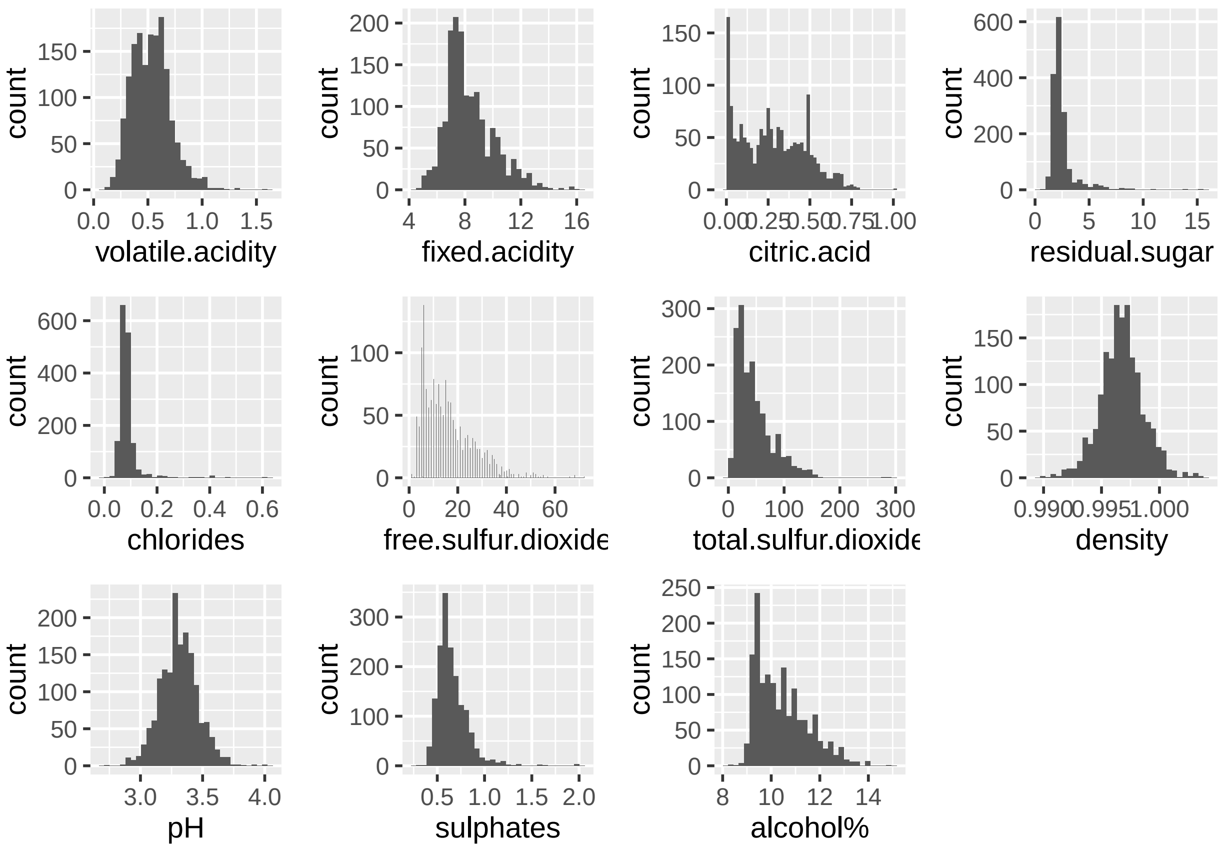
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

Histogram for Quality of Wine





Both from the histogram and stats summary the 50% of the wine are ranked as quality with 5 and 6 out of 10.



Initial Findings from Histogram

volatile acidity

- Most of the wine has a volatile acidity concentration between $0.4g/dm^3$ and $0.53g/dm^3$. More over the data is normaly distributed.

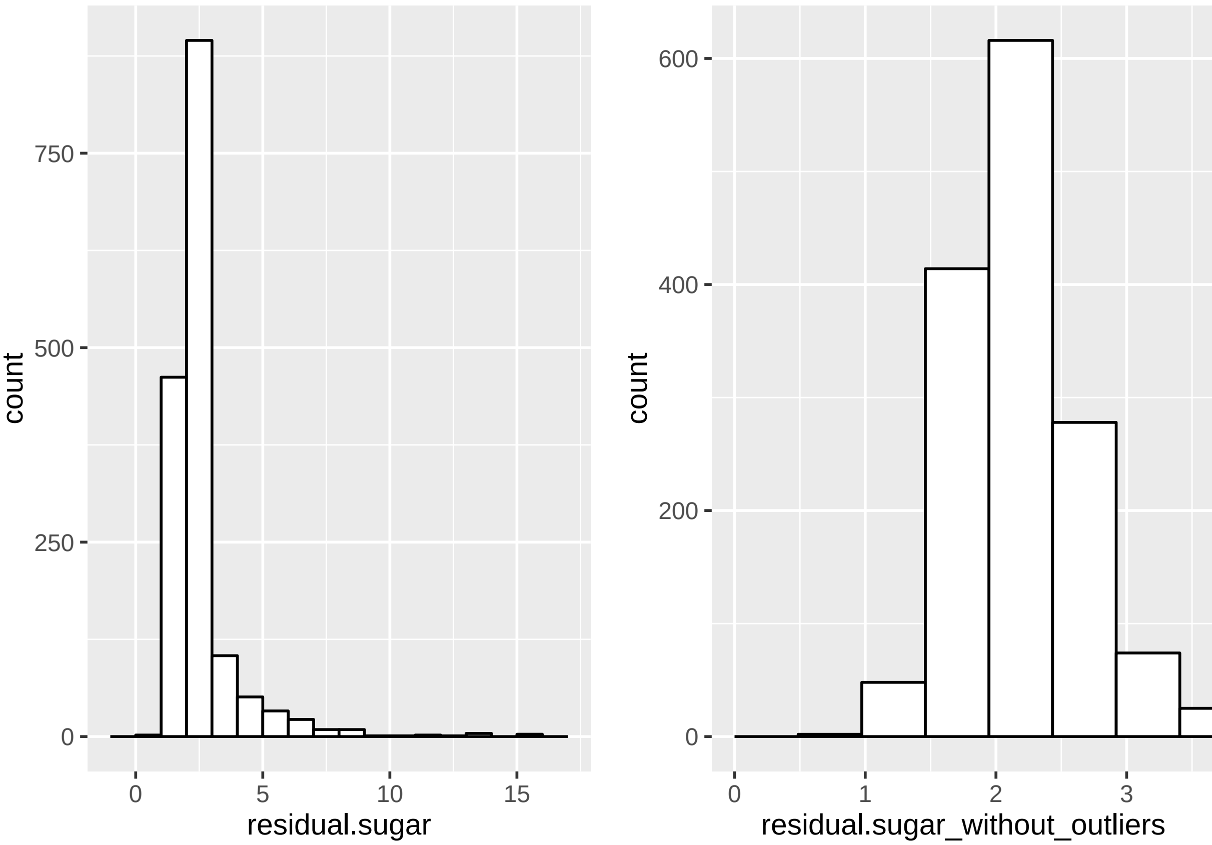
fixed acidity

- Most of the wine has fixed acidity concentration between $7g/dm^3$ and $9g/dm^3$.
- There are some outlies which are spread out to $16g/dm^3$.

citric.acid

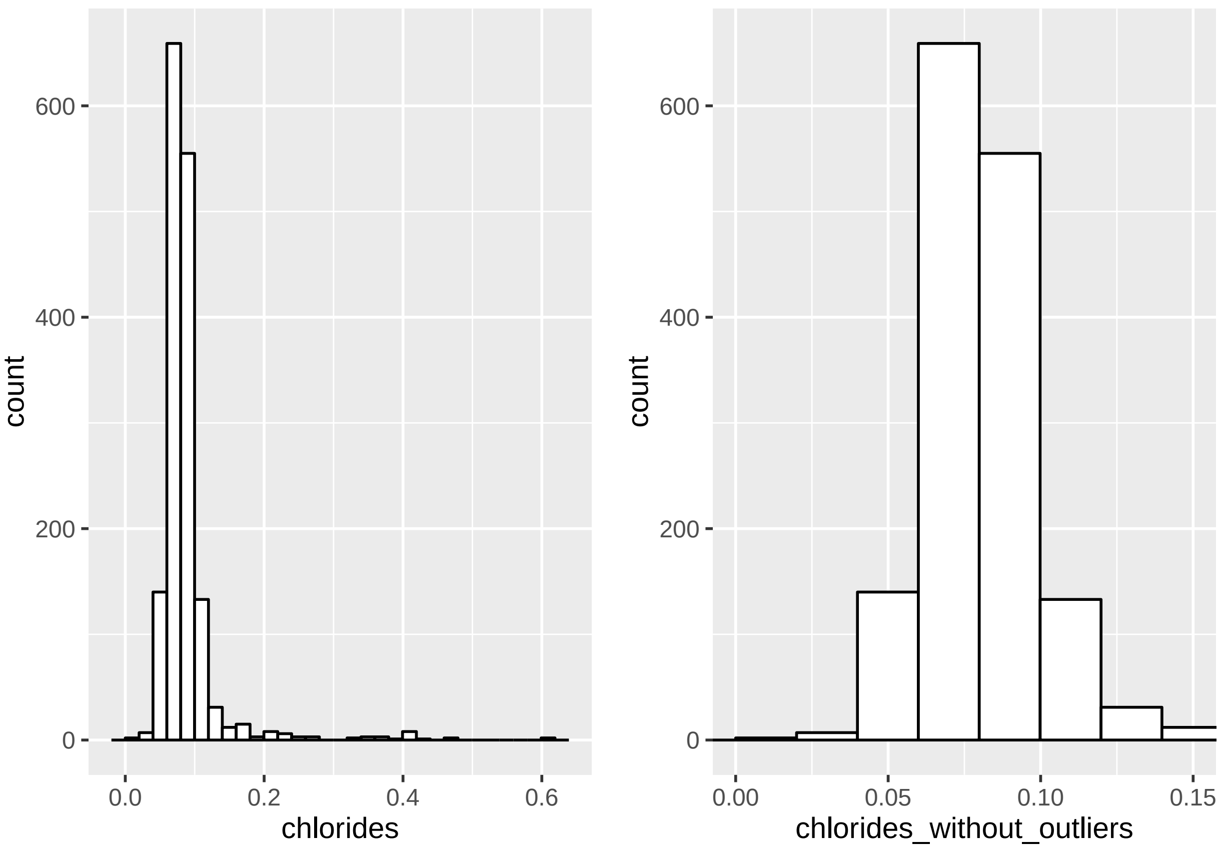
- There are alot of 0 which means a lot of wine don't has citric acid. It also make sense too because citric acid is added as a freshner or flavor to wines not as a key ingredients in cooking wine.

- I deffinitly like to investigate more on this lated as adding flavo make any changes to the quality of the wine.



residual sugar

- Almost most of the wine has sugar contanr less then $10g/dm^3$.
- The lookes alot like long tail but once the outlier is been removed the histogram looks normaly distribued.
- Interesing all wine sample has atleast $1g/dm^3$. i will investigate,is there any relations between the sugar contain and quality as i read in an
  artical apart from sugar from fermentation more sugar are added in wine making process so done adding more sugar inceases the wine quality?.



chlorides:

- Chloride is a salt. I don't like my wine to tase salty and others too i guess that is why the amount of salt contain is less. Almost 75% of sample have less then $0.09g/dm^3$.
- once the outliers are removed it very clear that the chloride are spread very well normaly spread across simple.

free and total sulfur dioxide:

- Free sulfur dioxide is added to prevent microbial growth and the oxidation of wine the amount of sulfur dioxide contain mean value $15.87mg/dm^3$ with some seious outlier striching out to $72mg/dm^3$.
- The total sulfur dioxide contain has a mean $72mg/dm^3$ which when compaied to only free sulfur dioxide's mean is much high. i will investigate does this high sulfur dioxide affect the quality in laster stage.

density

- Depending on the percent alcohol and sugar content the density of wine is close to water i will investigate how density of the wine affects the quality.
- One question does quality increses as the density of the wine moves close towords the density of the water?.

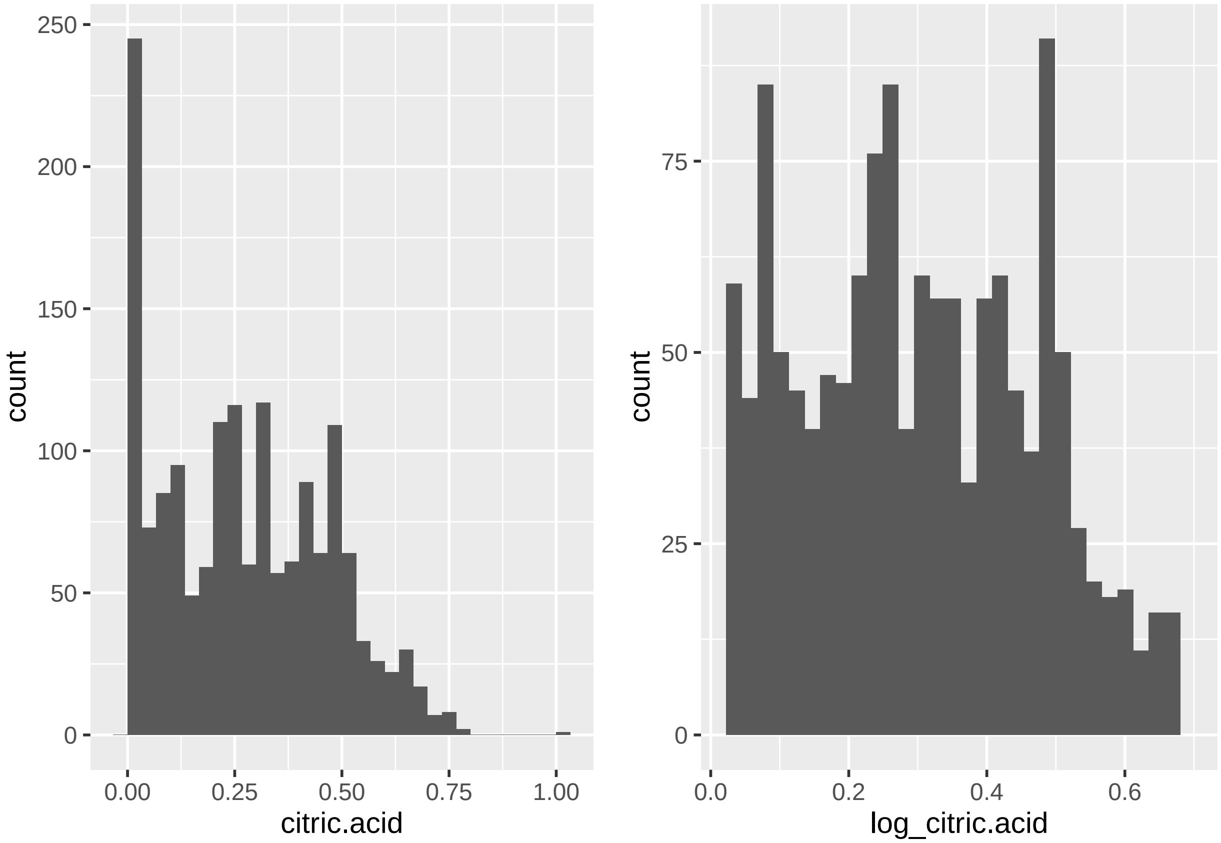
pH

- pH is the sacle is used to measure liquid is acid or base most of wine are have scale feom $3 - 4$ so as our sample.
- The question is there trend in the quality of wine as the pH level increase or deceases?.

Alcohol

- Alcohol is one of the importain ingredients thats why all wine bottles carry lable with percentage of alcohol contant but how much alcohol contain does a good quality wine has? let me do more anlaysis.

Transforming variable to check normality.



- dose wine sample has a lot of citric acid with 0 value ?*

Yes, the number of `citric.acid` with 0 value is 132*. So its not that data is missing but acutually the alot of wine doesn't have `citric.acid`.let me remove the 0 and do log transformation so that i can find any distribution from the histogram.* even after the transfomation there isn't any distibution visiable. i will investiage more with a box plot to check for trend.

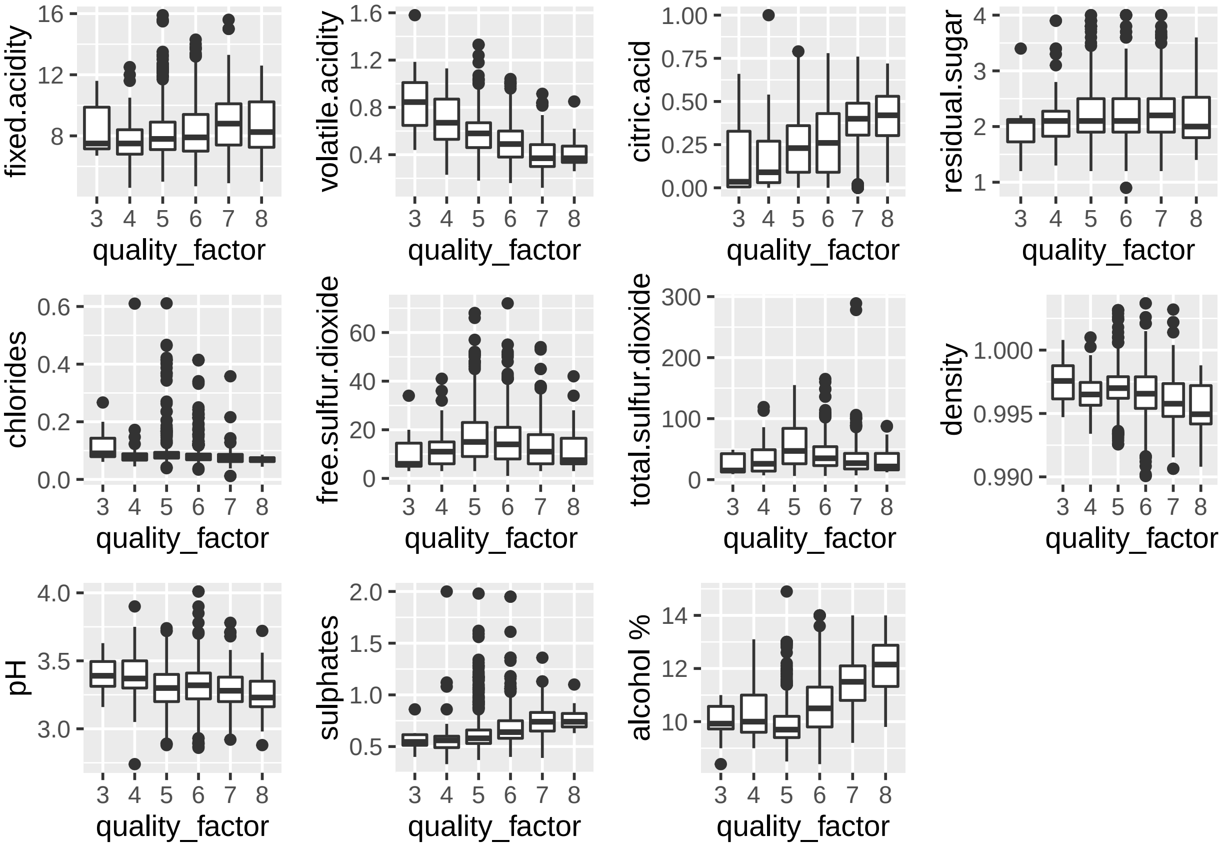
Corealtion between Variables and Quality

Positive correation	Negative correation
alcohol =0.476	volatile acidity = -0.390

Positive correation	Negative correation
sulphates=0.251	total sulpure di oxide = −0.185
citric acid=0.226	density = −0.174
fixed acidity=0.124	chlorides = −0.057
residual sugar=0.013	pH = −0.057
− − −	free sulfur dioxide = −0.050

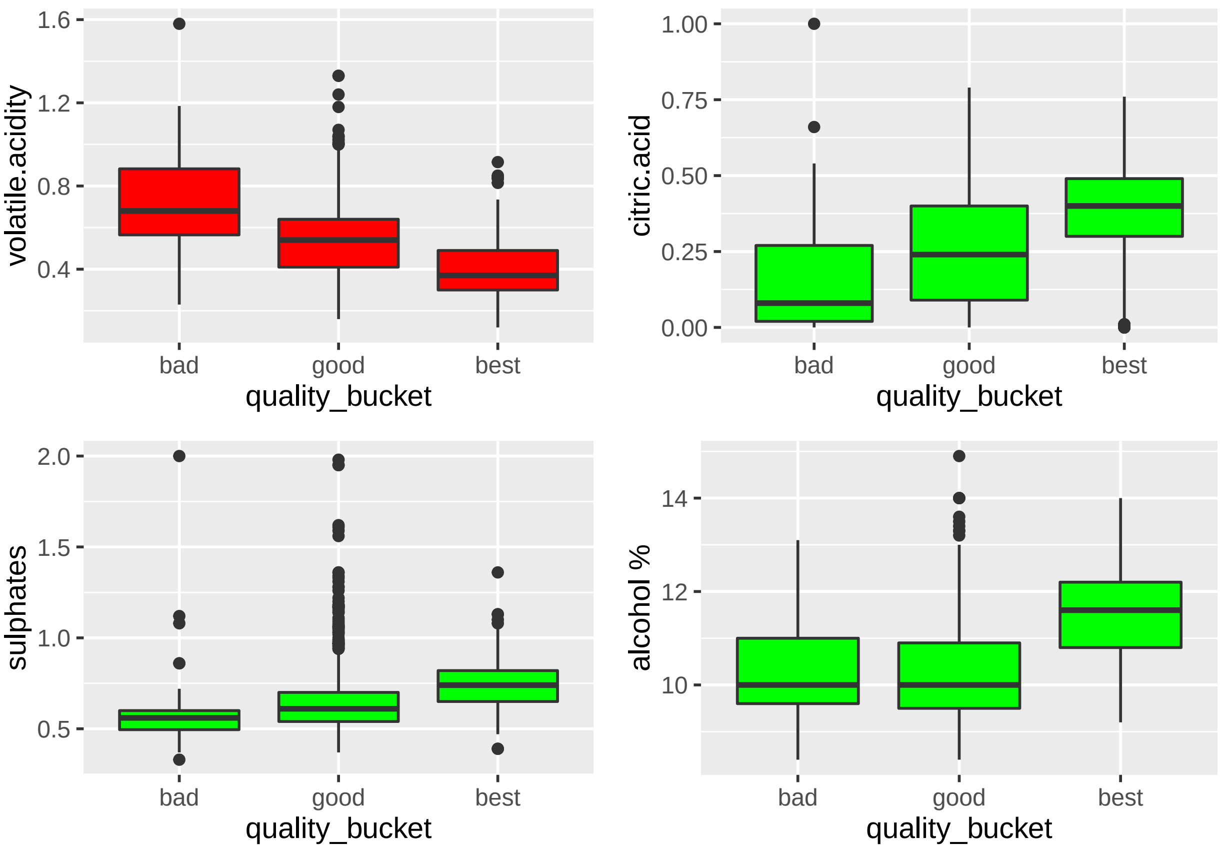
Bivariant Plot and Analysis

Boxpot of differnt Variable with Quality



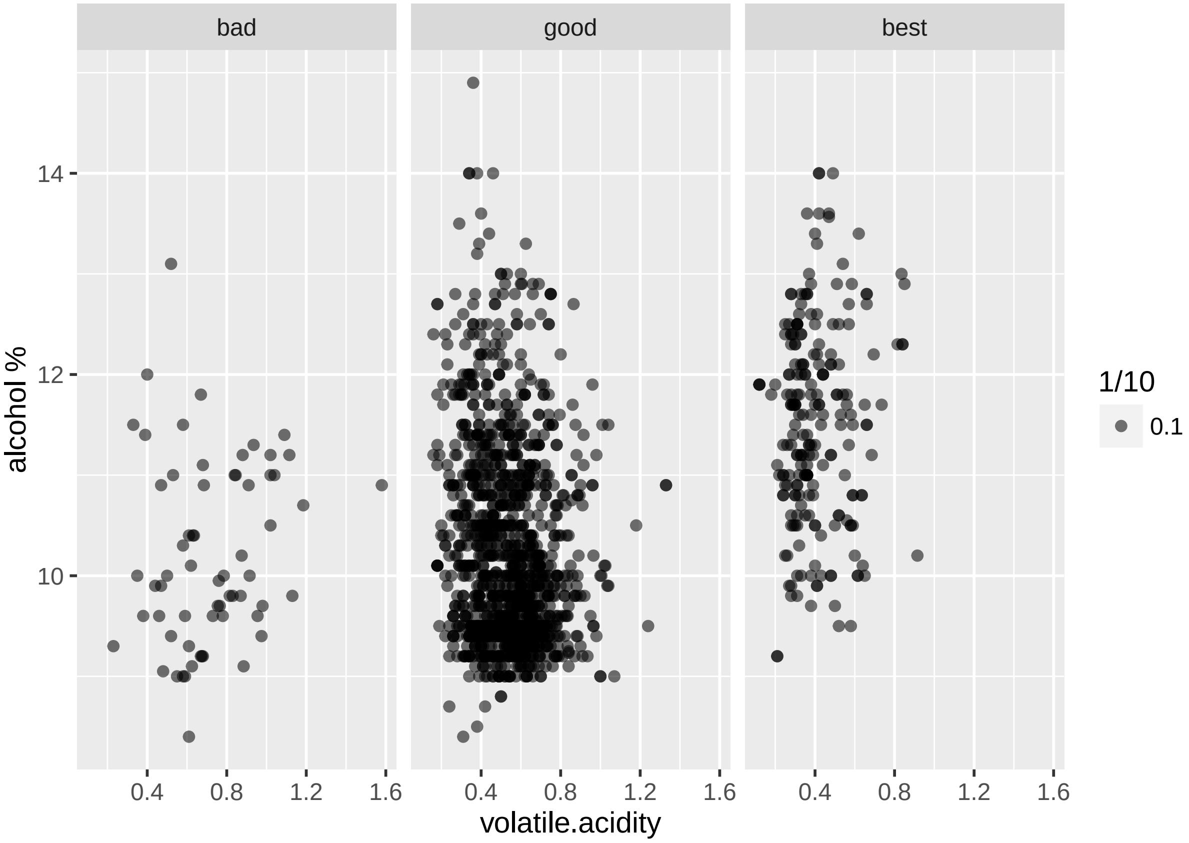
- If we follow the tread of the median bar in the box plot the we can find if any tread between the quality and the other variable.
- The variable `alcohol` ,`sulphates` , `citric.acid` all have a positive trend.
- The variable `volatile` has negative trend.
- All other variables doesn't have a good variation to have clear picture of there tread.

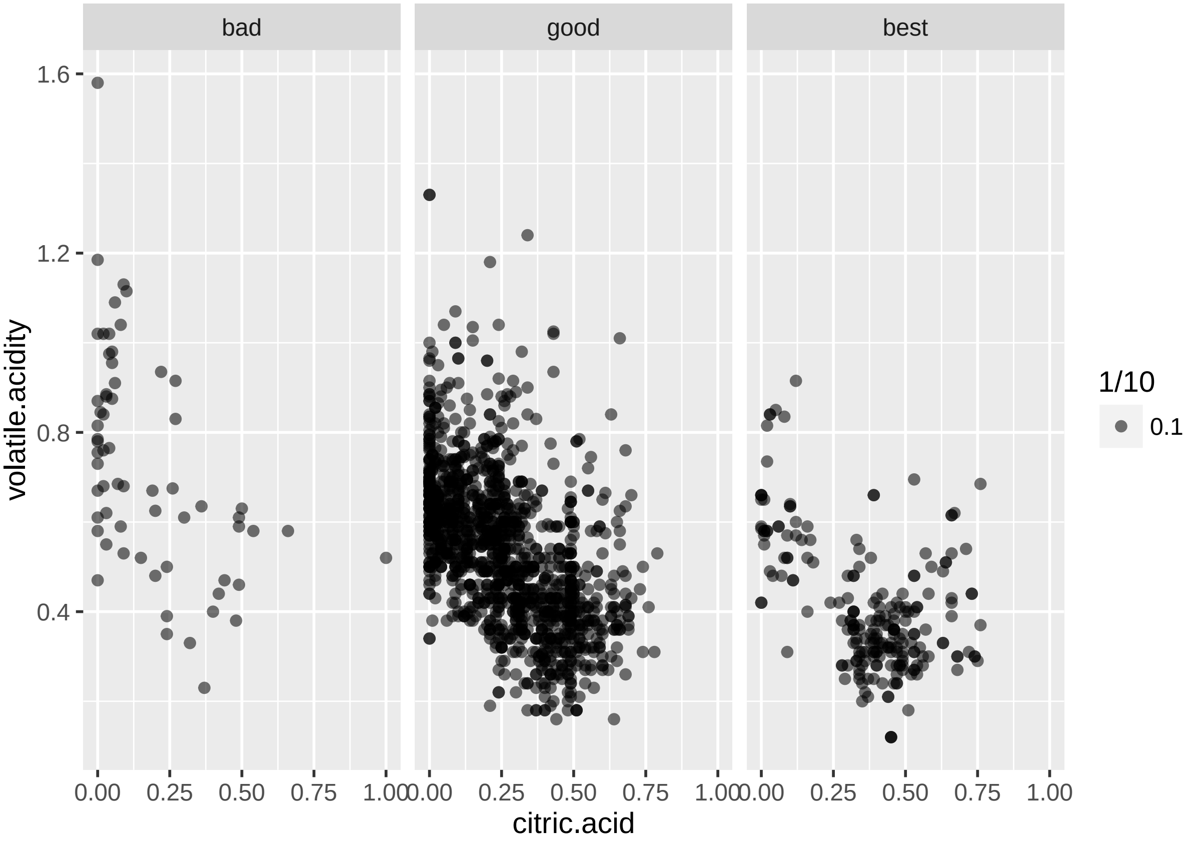
Investigate more on Alcohol ,Sulphates , Citric.acid, Volatile To have more clear understanding of quality variable the quality is splited into 3 categories **bad** (score [0-4]), **good** (score [5-6]), **best** (score[7-8])

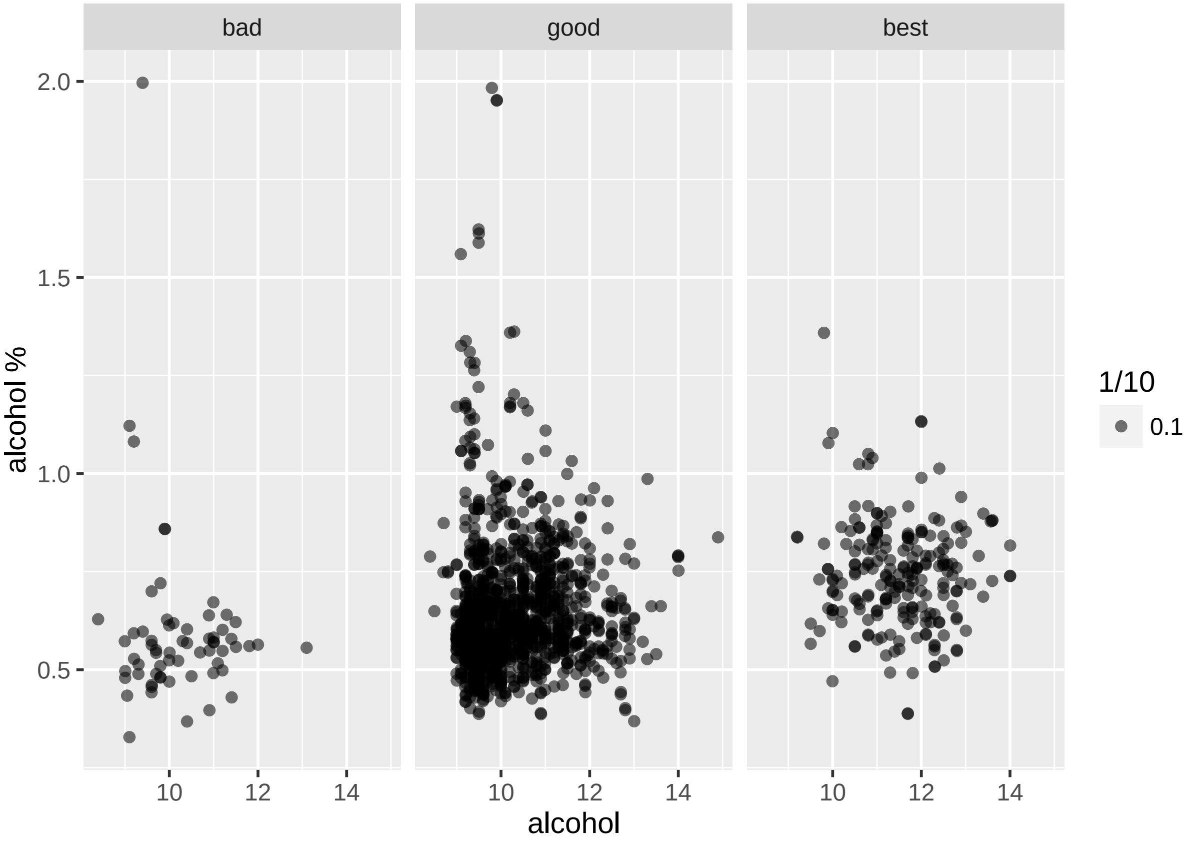


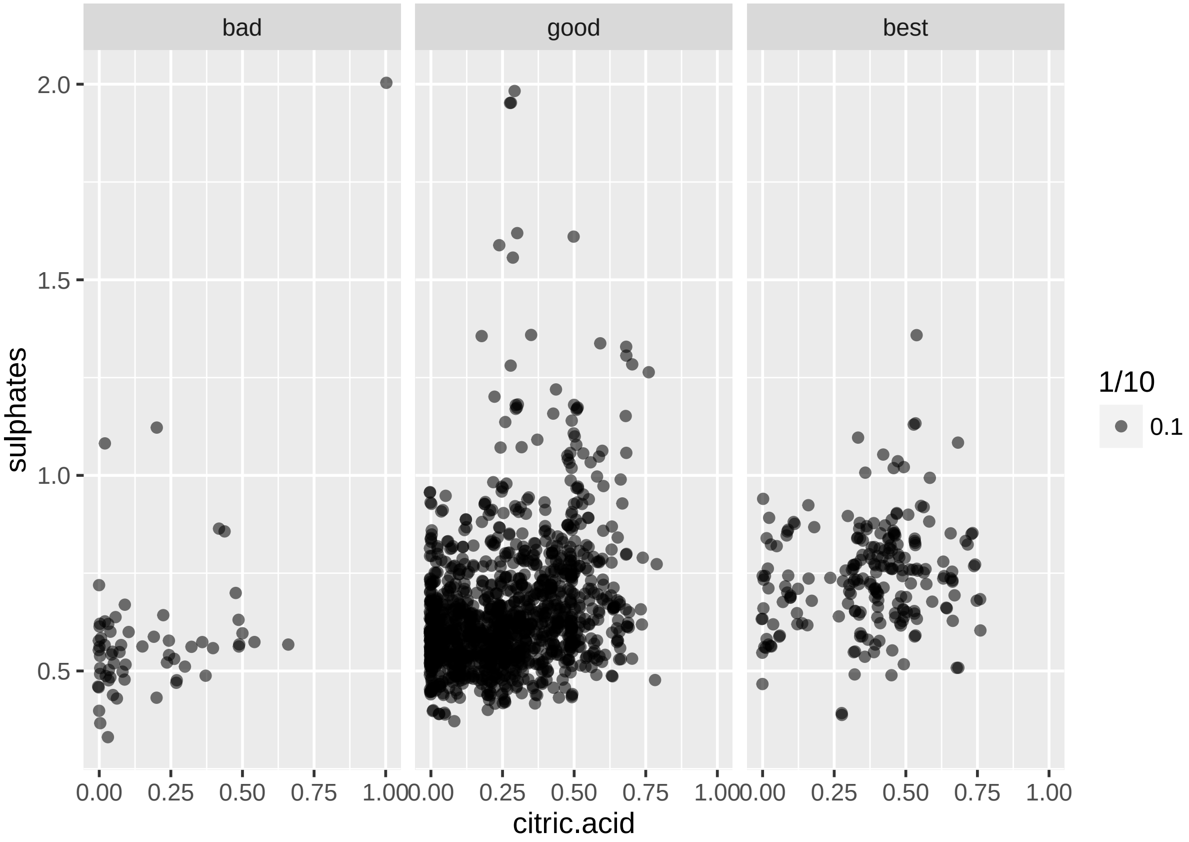
The trends are more clear and very well labeled for good visualisation. The **red color** of `volatile.acidity` shows it has negative trend while the **green color** of `citric.acid`, `sulphates`, `alcohol` shows the positive trend with *wine quality*

The volatile acid, sulphate, alchol, and citric acid have high correlation to wine quality. I like to check how these variable are corelated to each other with polted with a scatter plot.


bad
good
best
alcohol %
14
12
10
0.4
0.8
1.2
1.6
volatile.acidity
1/10
0.1


bad
good
best
1.6
1.2
0.8
0.4
volatile.acidity
0.00
0.25
0.50
0.75
1.00
citric.acid
1/10
0.1


bad
good
best
2.0
1.5
1.0
0.5
alcohol %
10
12
14
alcohol
1/10
● 0.1



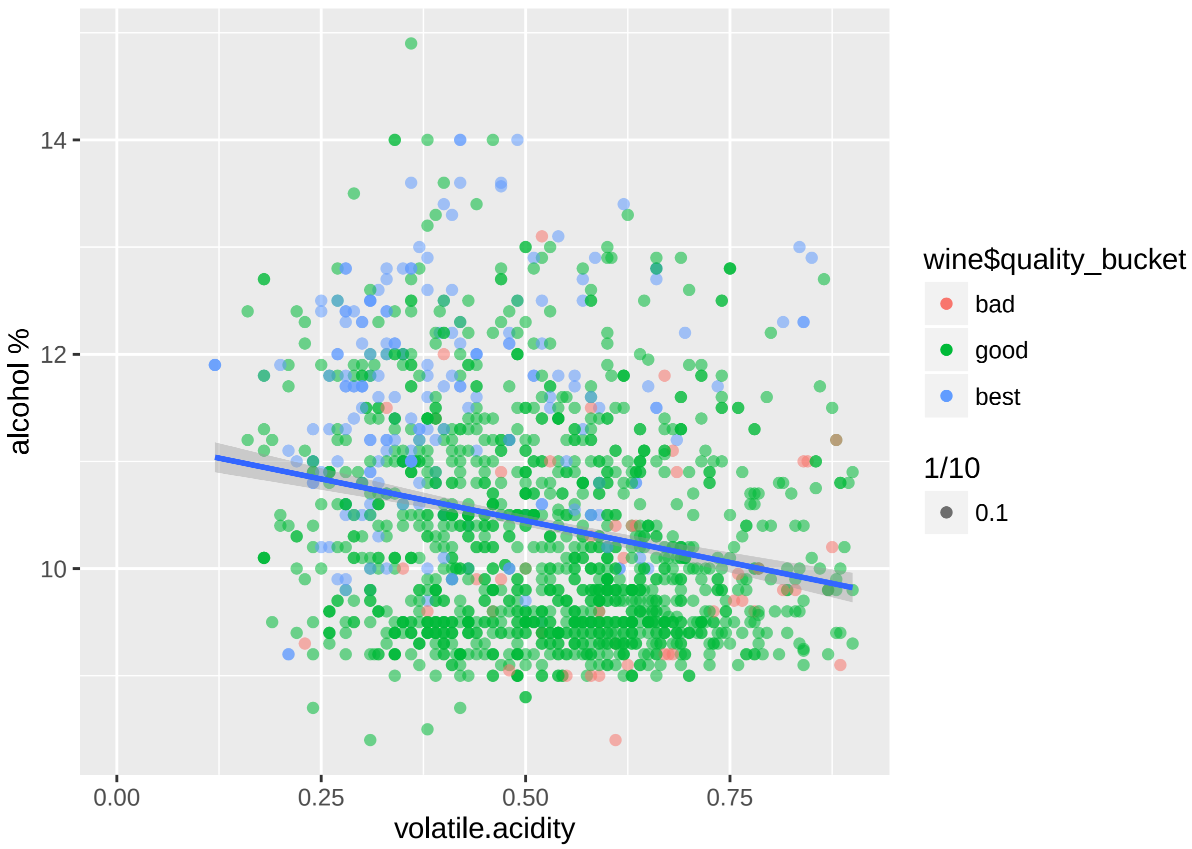
The plots revel very little relation between the variables and the scatter plot with citric acid and volatile acidity showed a clear negative trand. ## Building Models

```
##
## Calls:
## m1: lm(formula = I(quality) ~ I(alcohol), data = wine)
## m2: lm(formula = I(quality) ~ I(alcohol) + citric.acid, data = wine)
## m3: lm(formula = I(quality) ~ I(alcohol) + citric.acid + sulphates,
##     data = wine)
## m4: lm(formula = I(quality) ~ I(alcohol) + citric.acid + sulphates +
##     volatile.acidity, data = wine)
## m5: lm(formula = I(quality) ~ I(alcohol) + citric.acid + sulphates +
##     volatile.acidity + fixed.acidity + residual.sugar + chlorides +
##     total.sulfur.dioxide + density + pH, data = wine)
##
## ========================================================================
##                           m1         m2         m3         m4         m5
## ------------------------------------------------------------------------
## (Intercept)            1.875***   1.830***   1.434***   2.646***   25.493
##                       (0.175)    (0.171)    (0.176)    (0.201)    (21.142)
## I(alcohol)             0.361***   0.346***   0.338***   0.309***    0.275***
##                       (0.017)    (0.016)    (0.016)    (0.016)     (0.026)
## citric.acid                       0.730***   0.513***  -0.079      -0.231
##                                  (0.090)    (0.093)    (0.104)     (0.145)
## sulphates                                    0.814***   0.696***    0.929***
##                                             (0.107)    (0.103)     (0.114)
## volatile.acidity                                       -1.265***   -1.124***
##                                                        (0.113)     (0.120)
## fixed.acidity                                                       0.031
##                                                                    (0.026)
## residual.sugar                                                      0.020
##                                                                    (0.015)
## chlorides                                                          -1.825***
##                                                                    (0.419)
## total.sulfur.dioxide                                               -0.002***
##                                                                    (0.001)
## density                                                           -21.594
##                                                                   (21.575)
## pH                                                                 -0.361
##                                                                    (0.190)
## ----------------------------------------------------------------------------
## R-squared                 0.227      0.257      0.284      0.336      0.359
## adj. R-squared            0.226      0.256      0.282      0.334      0.355
## sigma                     0.710      0.696      0.684      0.659      0.649
## F                       468.267    276.595    210.501    201.777     88.909
## p                         0.000      0.000      0.000      0.000      0.000
## Log-likelihood        -1721.057  -1688.711  -1659.955  -1599.093  -1571.168
## Deviance                805.870    773.917    746.576    691.852    668.105
## AIC                    3448.114   3385.421   3329.910   3210.186   3166.337
## BIC                    3464.245   3406.930   3356.795   3242.448   3230.862
## N                      1599       1599       1599       1599       1599
## ============================================================================
```

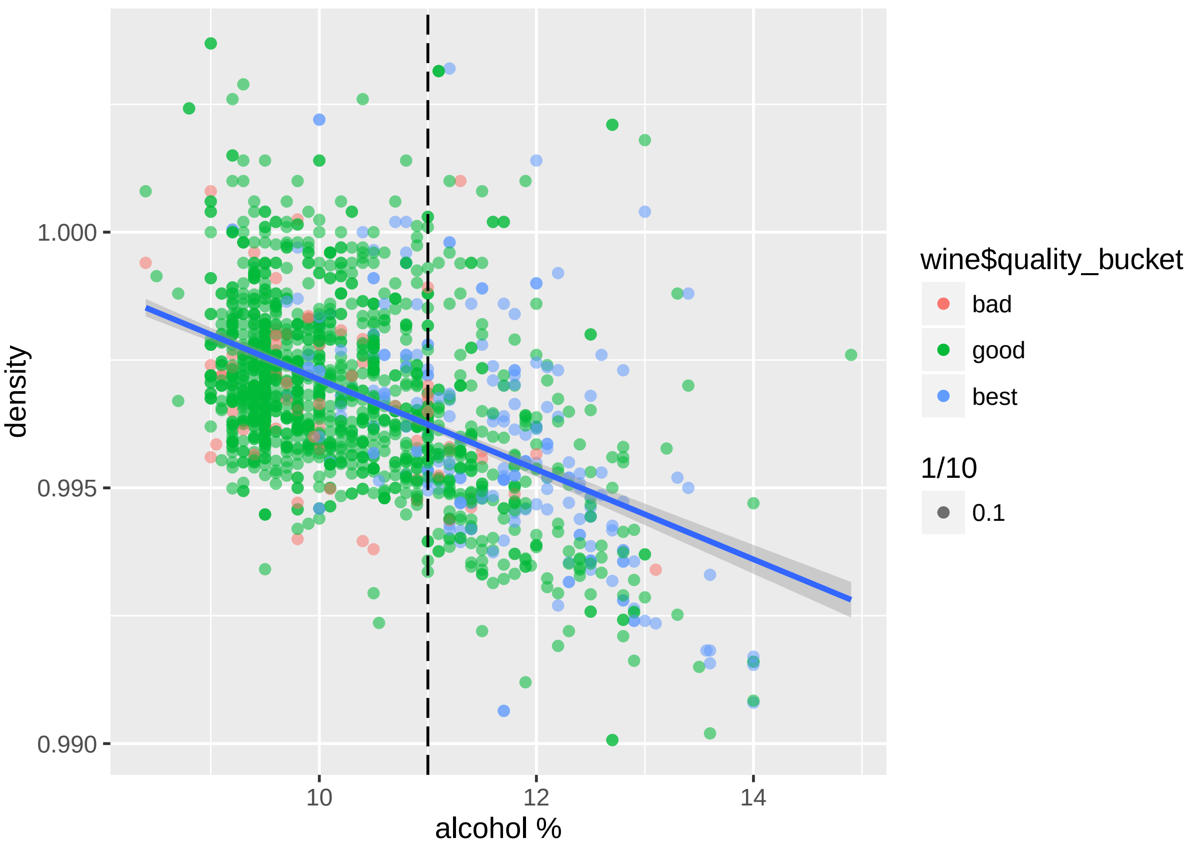
- Even after adding all the variable i couldn't find a good fix. There are 2 solution to this problem.
- Add two are more variable to get a new variable and add it to the model. I'm definitly skeptical in doing so since the full model $R - squared = .359$ which is a just a 2% high when compaired with a model have `alcohol`, `citric.acid`, `sulphates`, `volatile.acidity`.
- An other solution is to use a non-leaner model or do somthing call train and test your model to have a good prediction and of course a good quality wine.

Multi-Varient Plot:

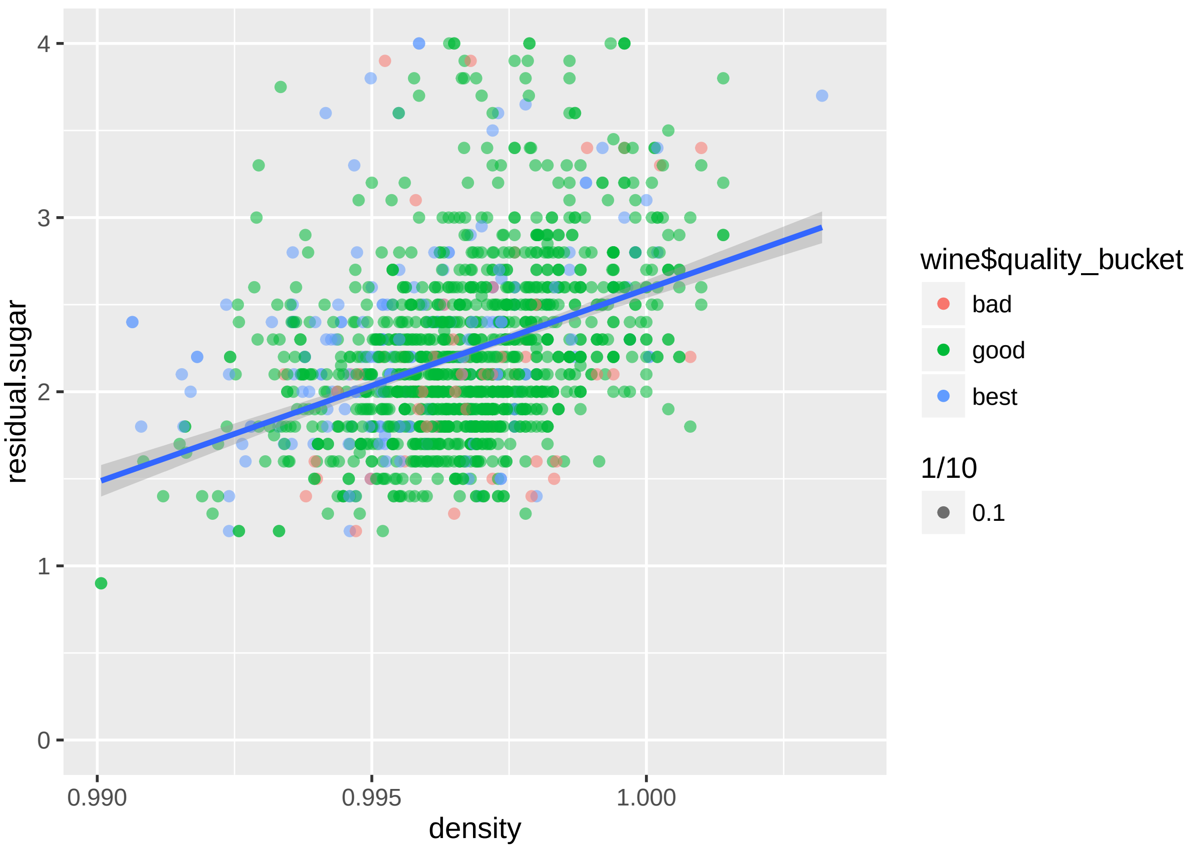
The variable such as `pH`, `density` may not directly corelated with the wine quality but they are a good measure of properties of wine. i will investigate more on them in coming analysis.



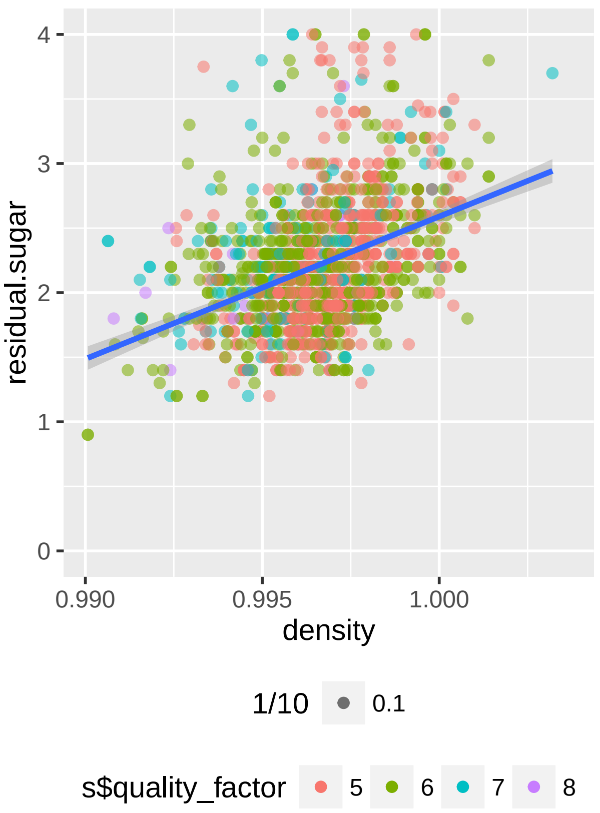
The scatter plot deffinitly separate the `best` Vs `good` wine. The best wine are the one with high alcohol with low volatile contain.

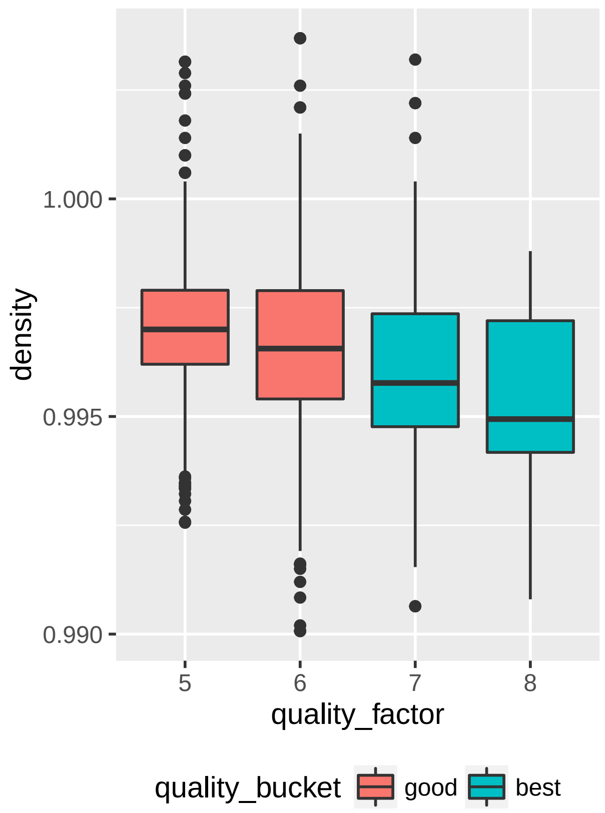


There is a trend that most of the best quality wine are from the graph it is clear that the wine with more alcohol contant and less density falls in the `best` quality and the `good` quality wine has comparetively less alcohol contain and dinsity to best quality.



I would see alot of `good` quality wine are below the line which means a `best` quality wine are with high sugar contant with lesser density but the visulisation is not as good so let me check just sample with best and good.

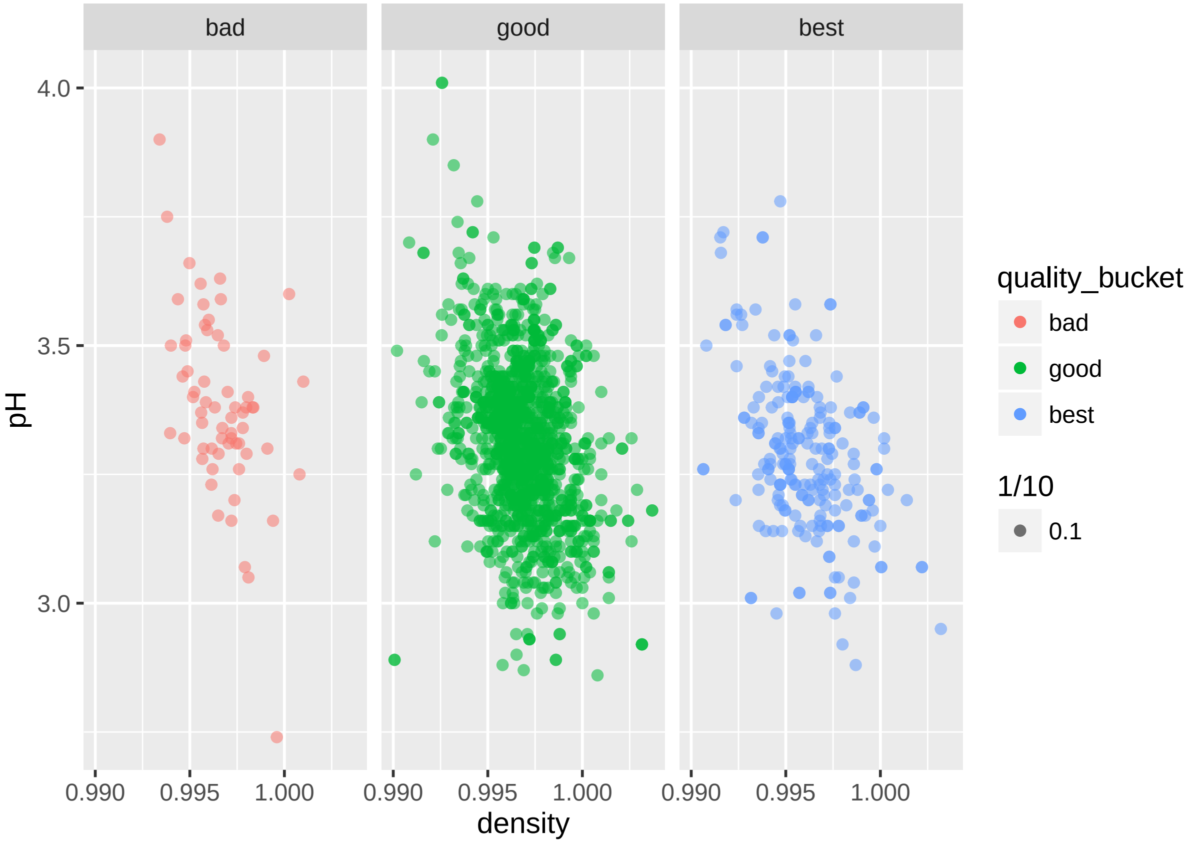




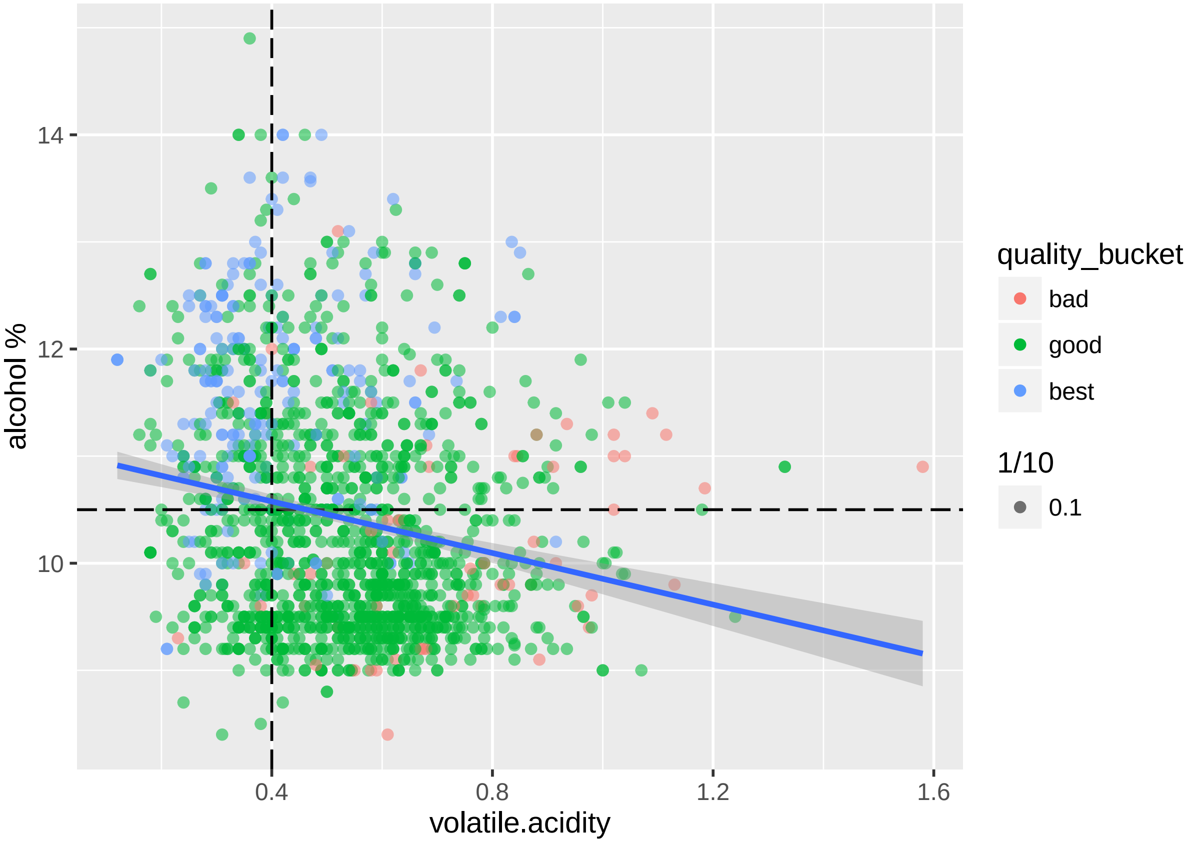
From the scatter plot there is deffinatly a trend as the density increases the residual sugar level increases this led to a simple question.

- Is a good quality wine has more density or lesser density? *

The box polt show as `density` decreases the `quality` of the wine increases. Thus it very much evident that a `best` quality wine has a property of *high suger contain with less density.* Finding the optimal range can we done in a future analysis.



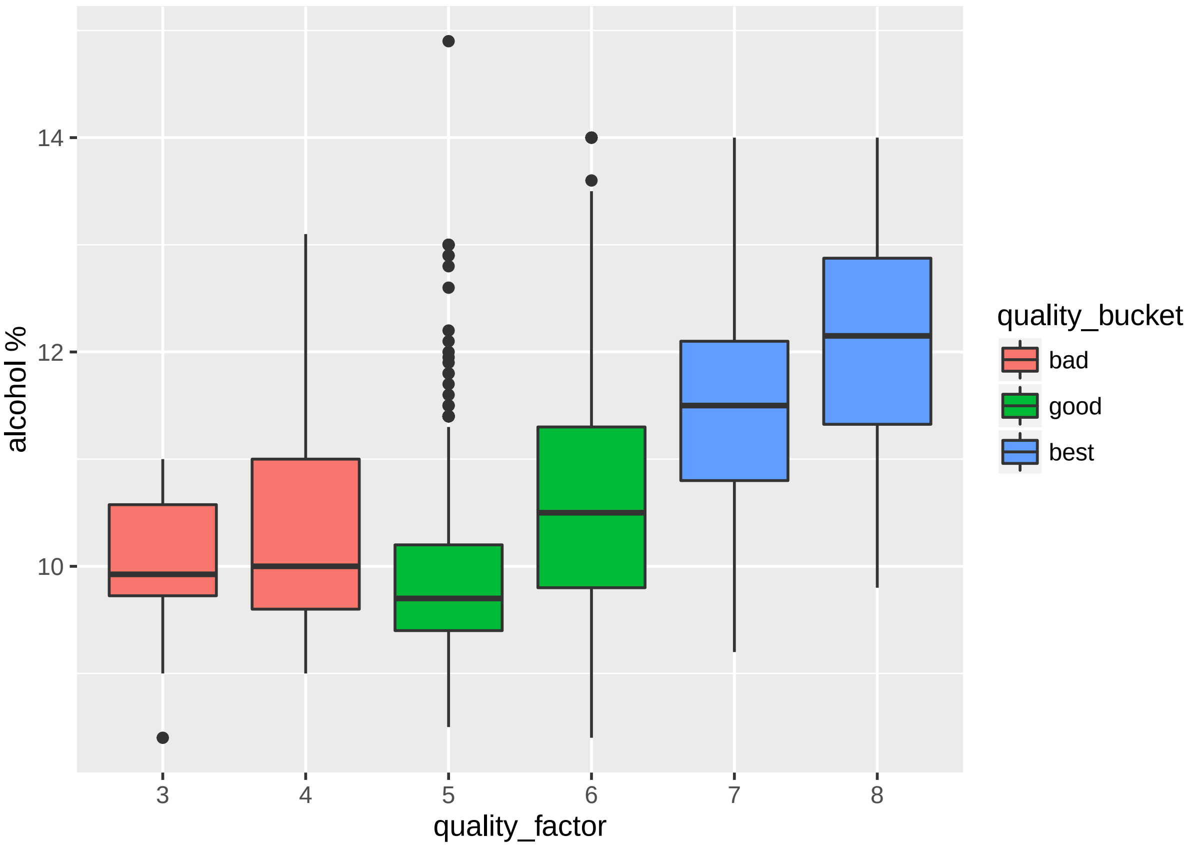
I though there might be some relationship between the density and pH but it seem there are not much to infer between them.



The scatter plot shows most of the best quality of the wine follow on high alcohol contain and low volatile acidity.

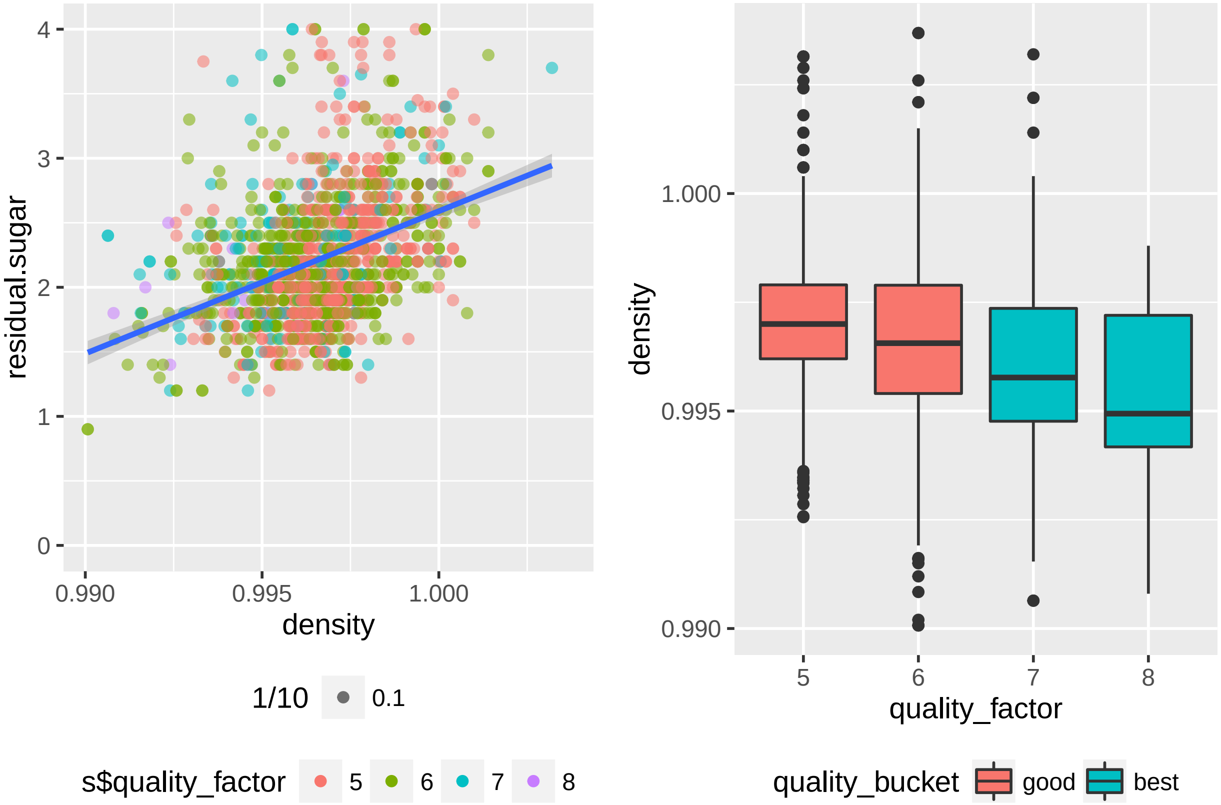
Final Plots

Plot One [box plot for *quality* vs *alcohol*]



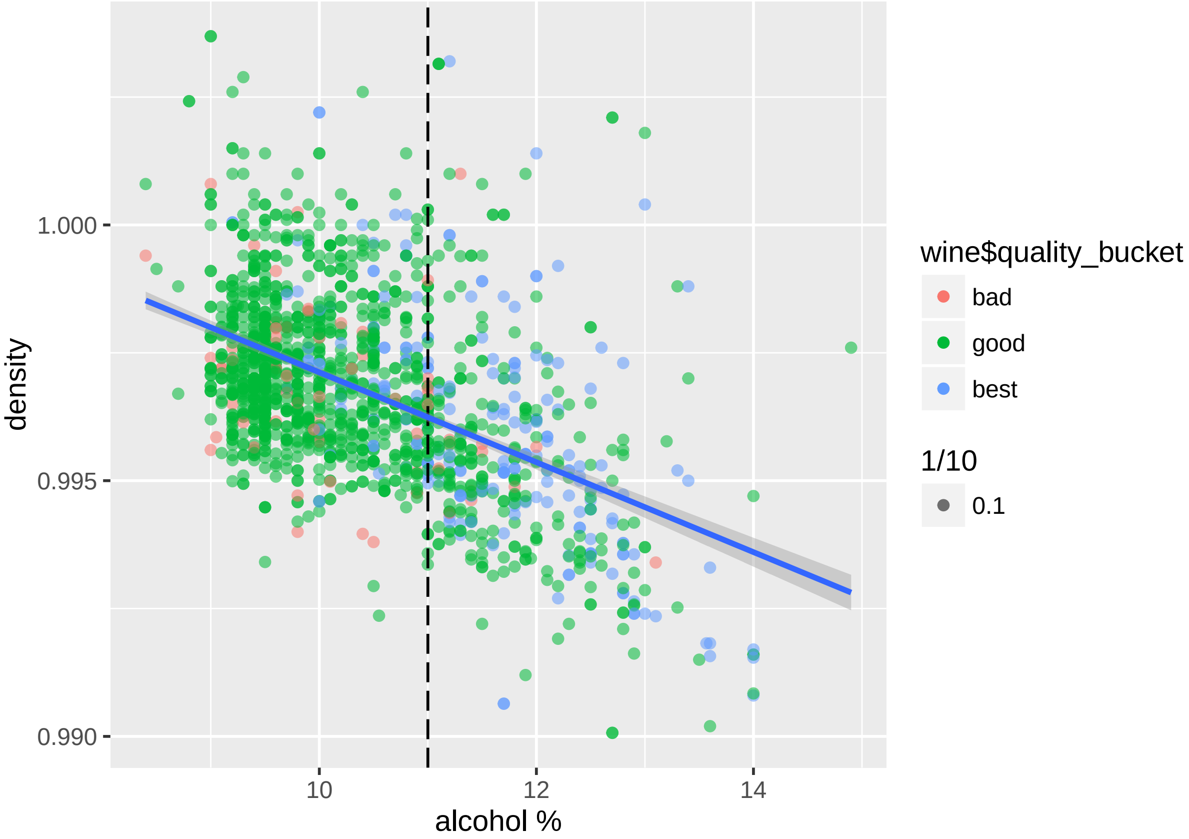
The alcohol% has a high positive correaltion with quality. A box plot with quality and alcohol% for different quality bucket will give a clear variation of alcohol% for each type of quality. The box plot show that alcohol % is one of the key ingredient in determinig the quality of the wine as the variations is very much seen as **higher** the *alcohol* contant the *quality* of the wine fall into `best` quality wine category.

Plot two [relationship between *residual_suger* with *density*]



The both scatter plot and box plot complement each other. Both plot together help in finding the properties of the `best` quality wine as sometime it need more then a single plot to explain a property of a variable in our case quality. The box polt show, as `density` decreases the `quality` of the wine increases. Thus it very much evident that a `best` quality wine has a property of *high suger contain with less density* which is infered from the scatter plot. Finding the optimal range can we done in a future analysis.

Plot three [box plot for *quality* vs *alcohol*]



A scatter plot of Alcohol % vs density had a clear trend but adding quality as color helped to find the key properties of the `best` quality wine. There is a trend that most of the best quality wine are from the graph it is clear that the wine with more alcohol contant and less density falls in the `best` quality and the `good` quality wine has comparetively less alcohol contain and dinsity to best quality.

Conclusion

From various plots i was able to summarise the property of the **best quality wine**. The following table contain the propreties *best quality wine*.

Higher contain	Lower contain
High alcohol contain	Low volatile acide
High alcohol contain	Lesser wine density
High residual sugar	lesser wine density

Reflection

The explorative analysis reveals the key factor affecting the quality of wine are alcohol,volatile acidity, sulphates, citric acid but the linear model which i have used for my prediction is not a good model as its coefficient of determination was .44. For the further study i would try diffrent modeling technic such as random forest modeling might give more accuracy.