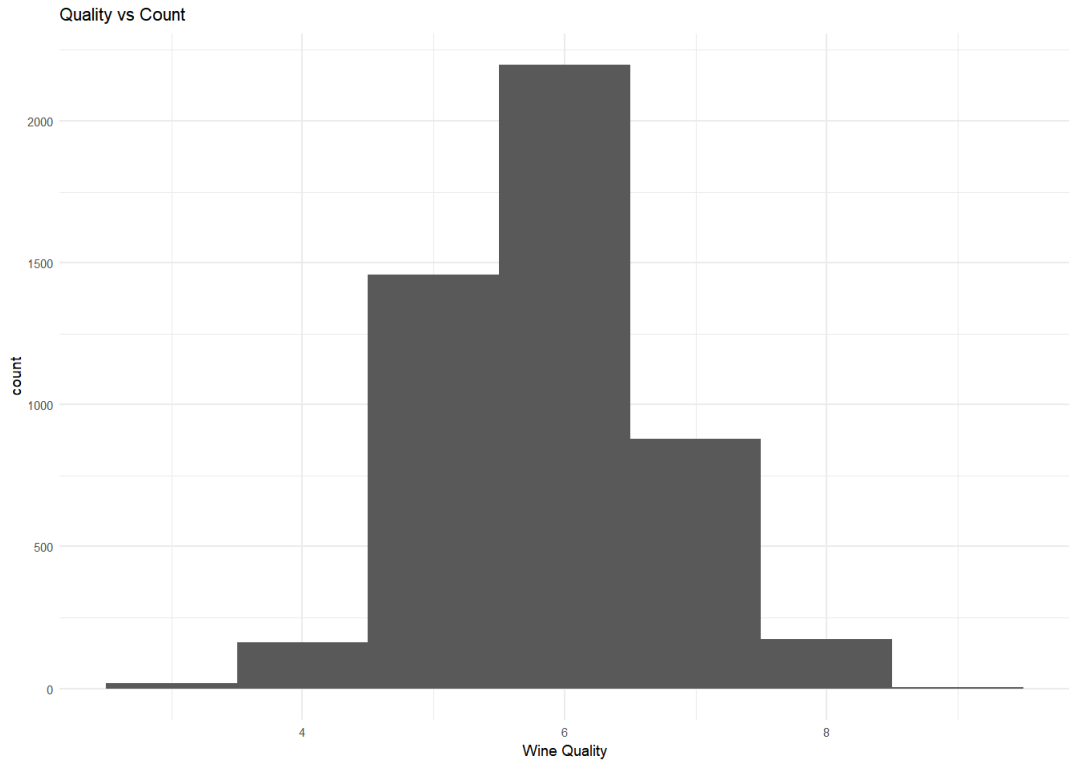
EXPLORATORY DATA ANALYSIS OF WHITE WINE QUALITY(DATA SET) by ANIRUDDHA AGARWAL

The data set that I am exploring here is : White Wine Quality, which is publicly available for research. The author of this data set are : Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV). It is related to white variants of the Portuguese “Vinho Verde” wine. No. of instances in it are - 4898 with 12 useful variables a general summary of data set has been given below.

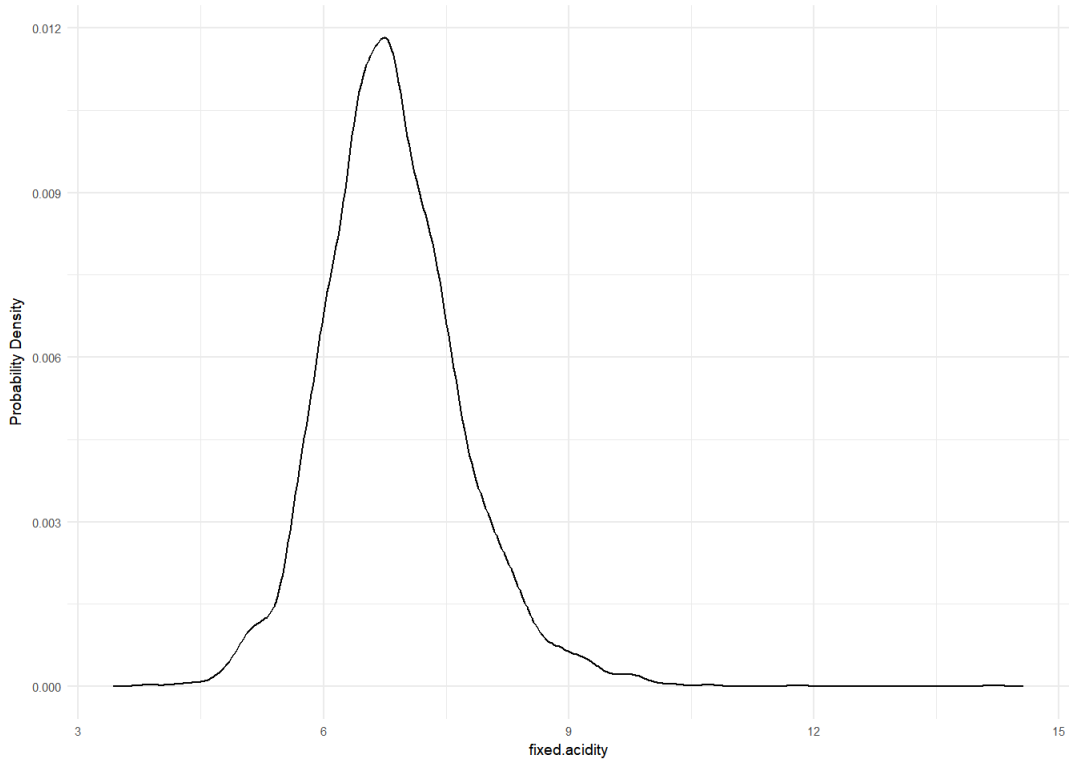
Univariate Plots Section

```
## [1] "Frequency distribution of quality"
```

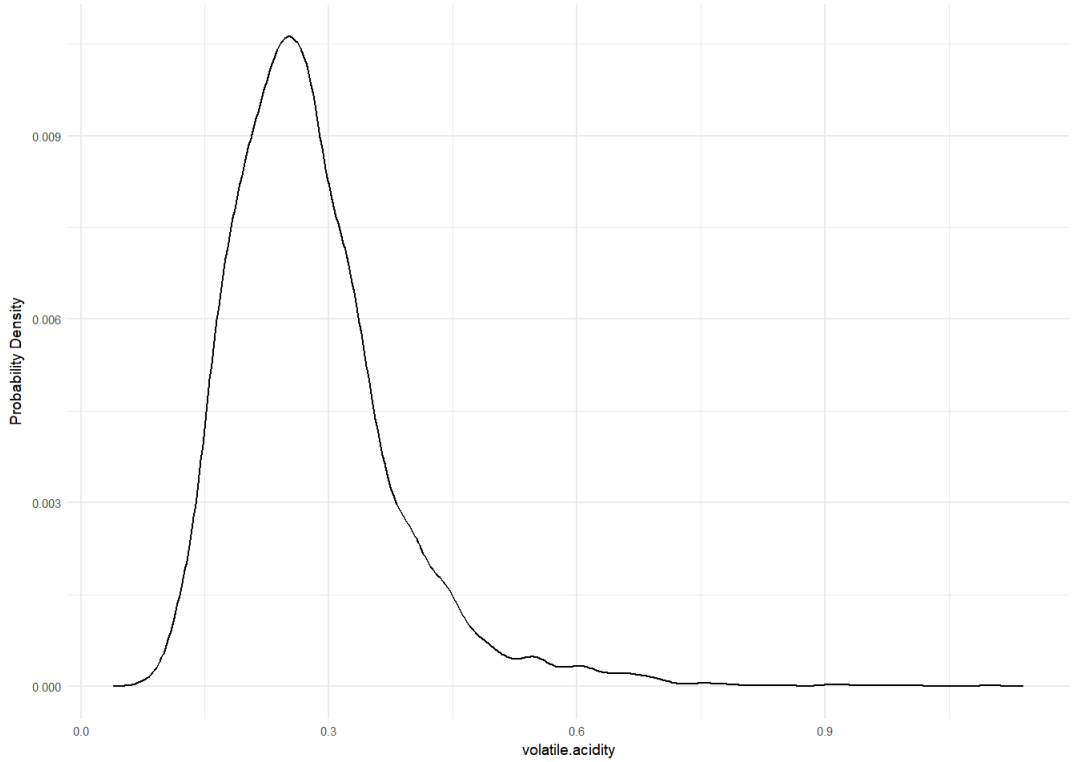
```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```



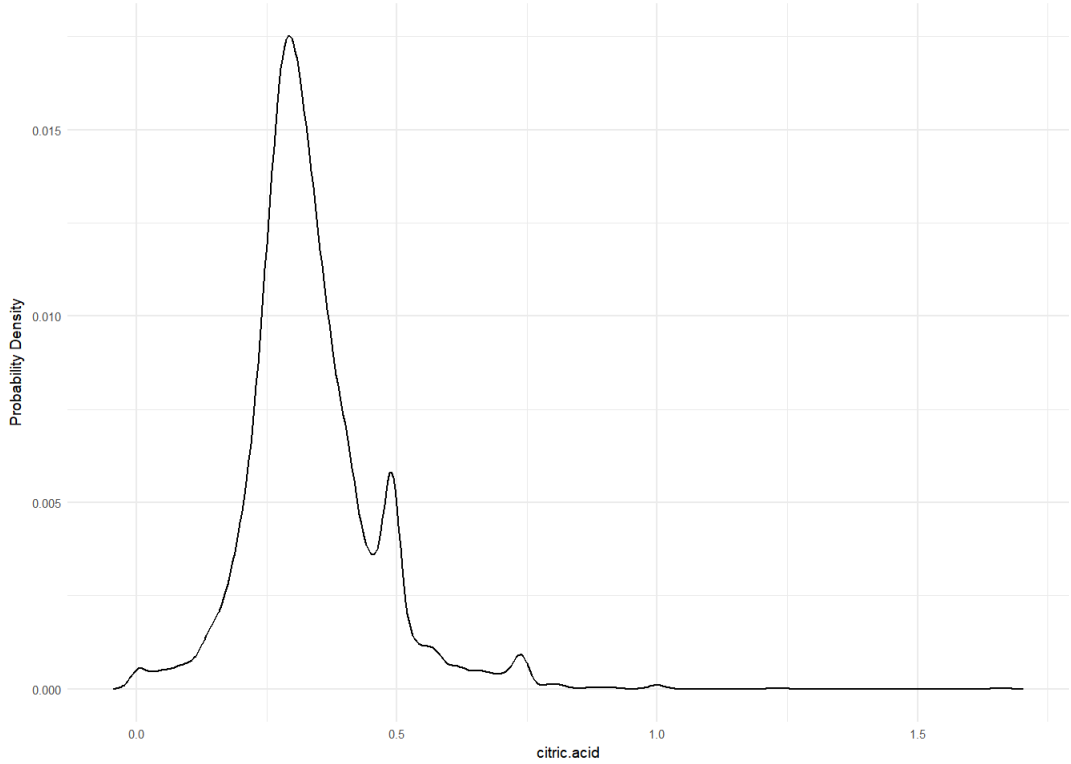

```
## [1] "Variable:  fixed.acidity"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```



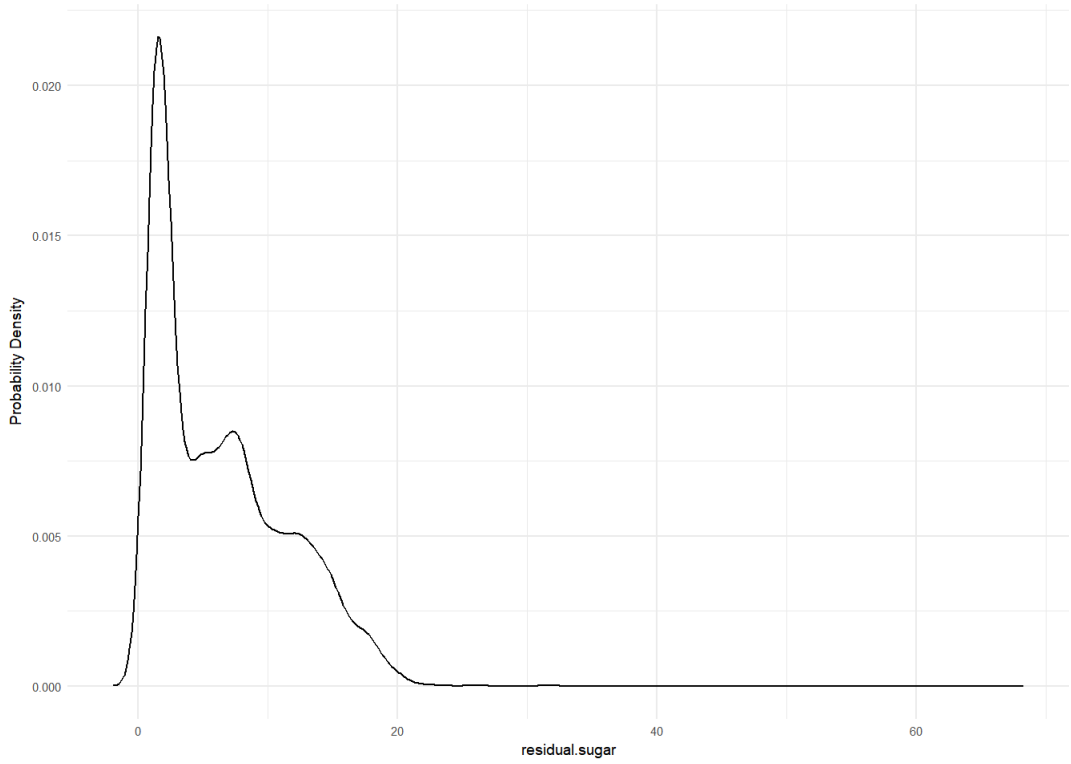

```
## [1] "Variable:  volatile.acidity"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```



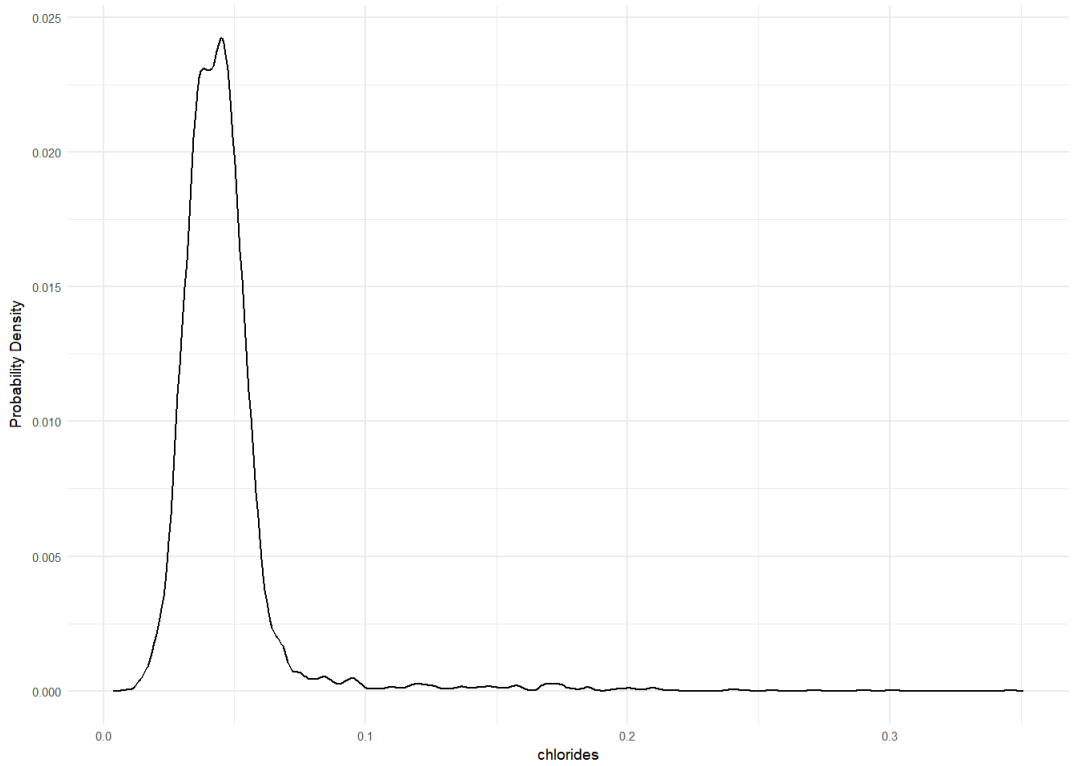

```
## [1] "Variable:  citric.acid"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```



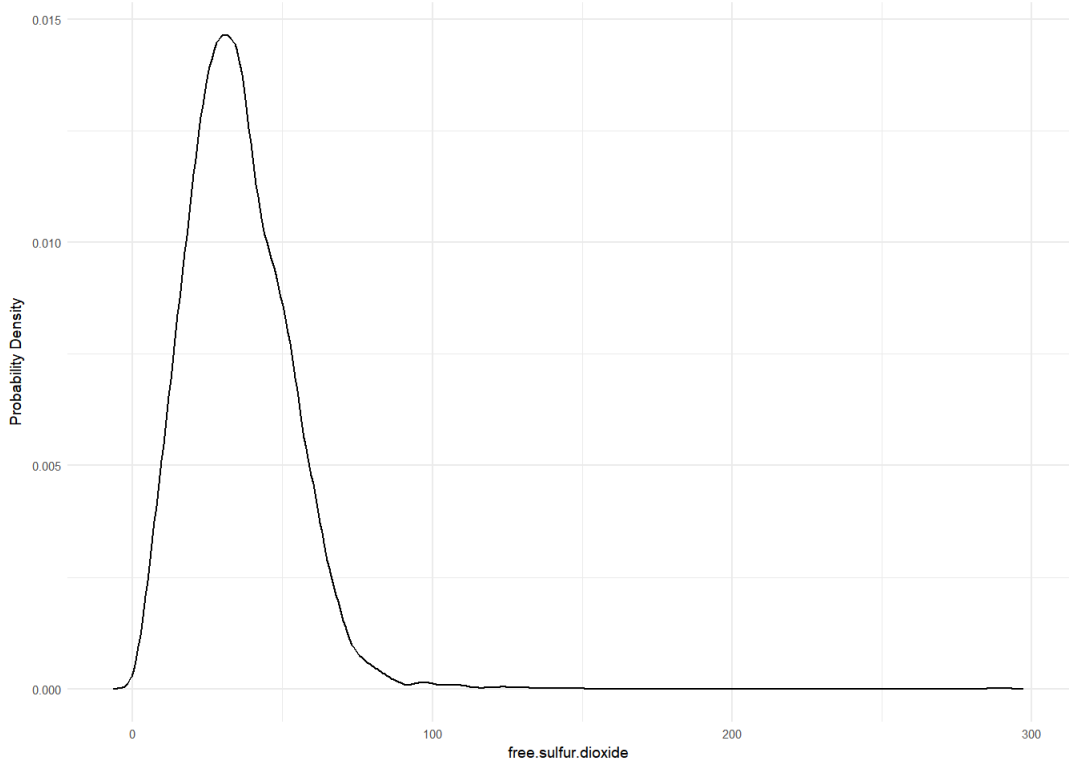

```
## [1] "Variable:  residual.sugar"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```



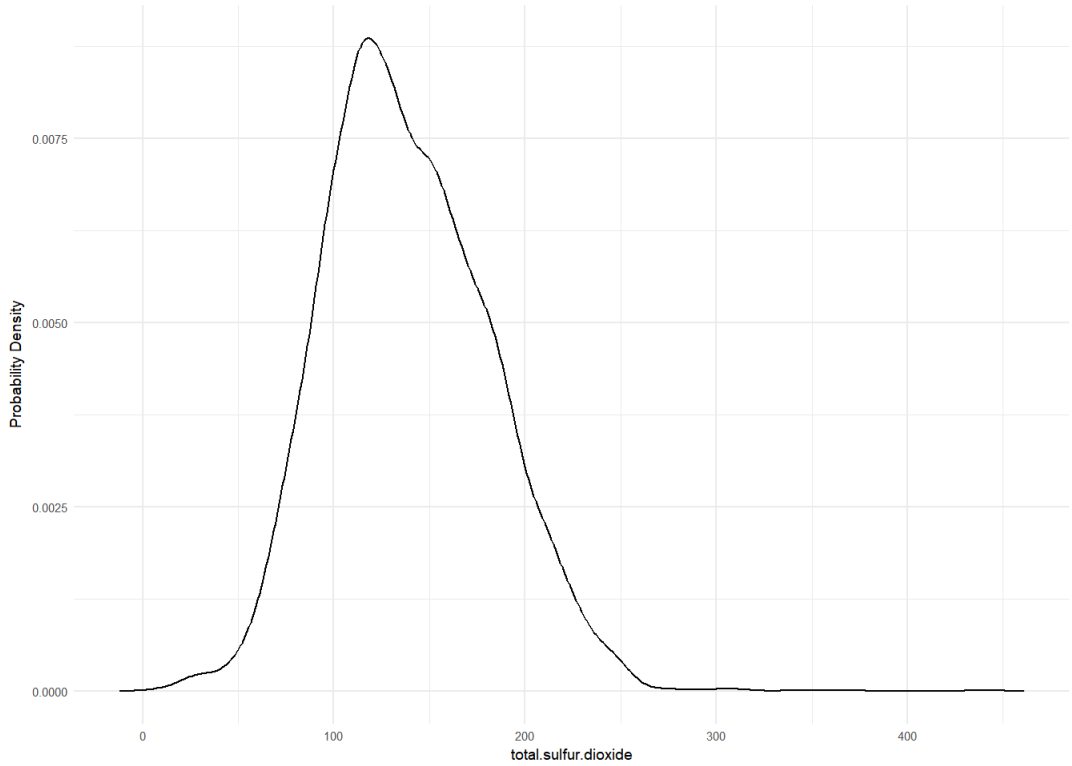

```
## [1] "Variable:  chlorides"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```



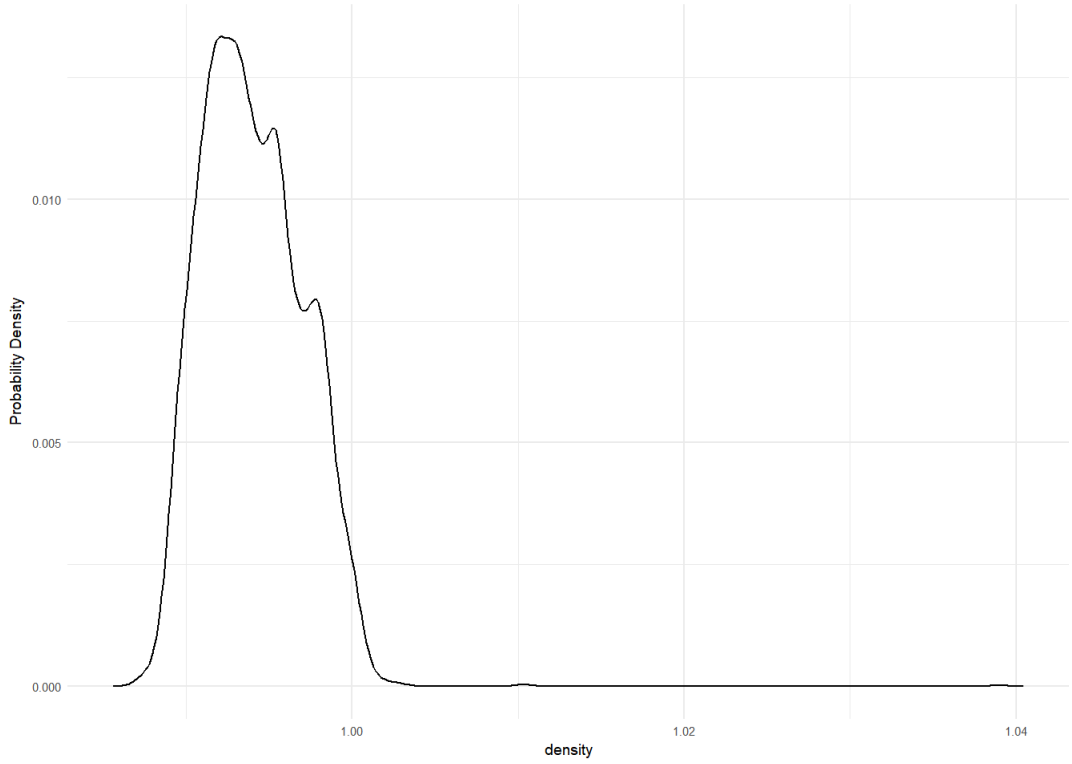

```
## [1] "Variable:  free.sulfur.dioxide"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```



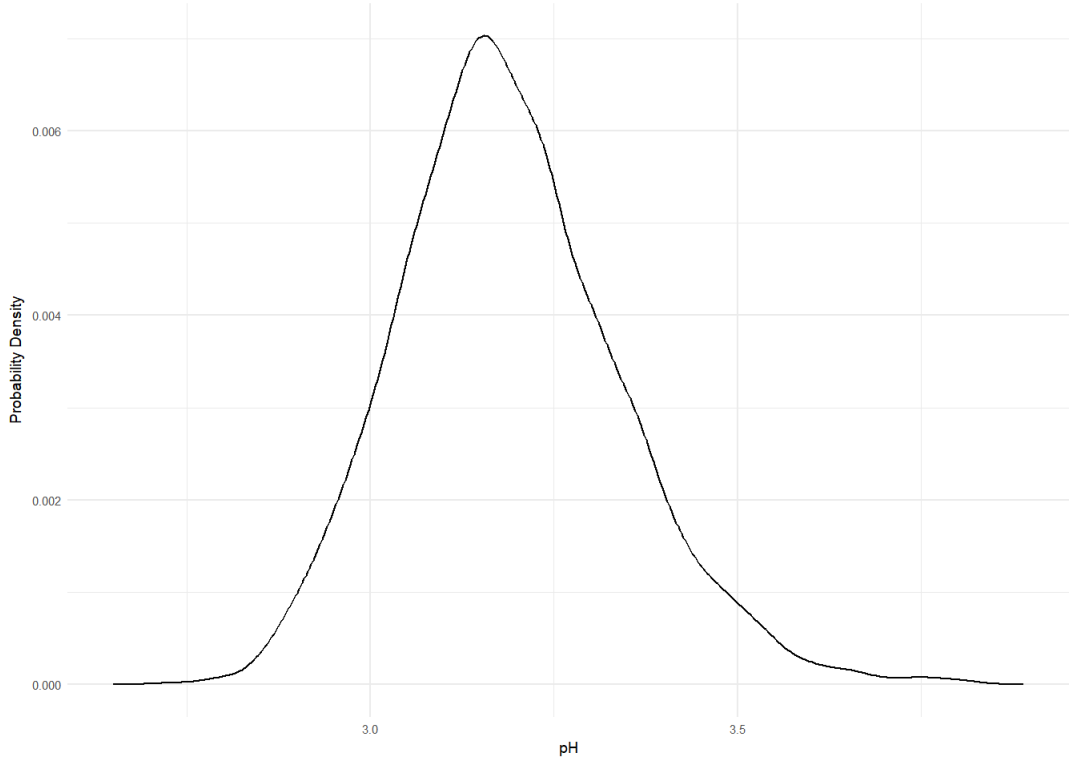

```
## [1] "Variable:  total.sulfur.dioxide"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```



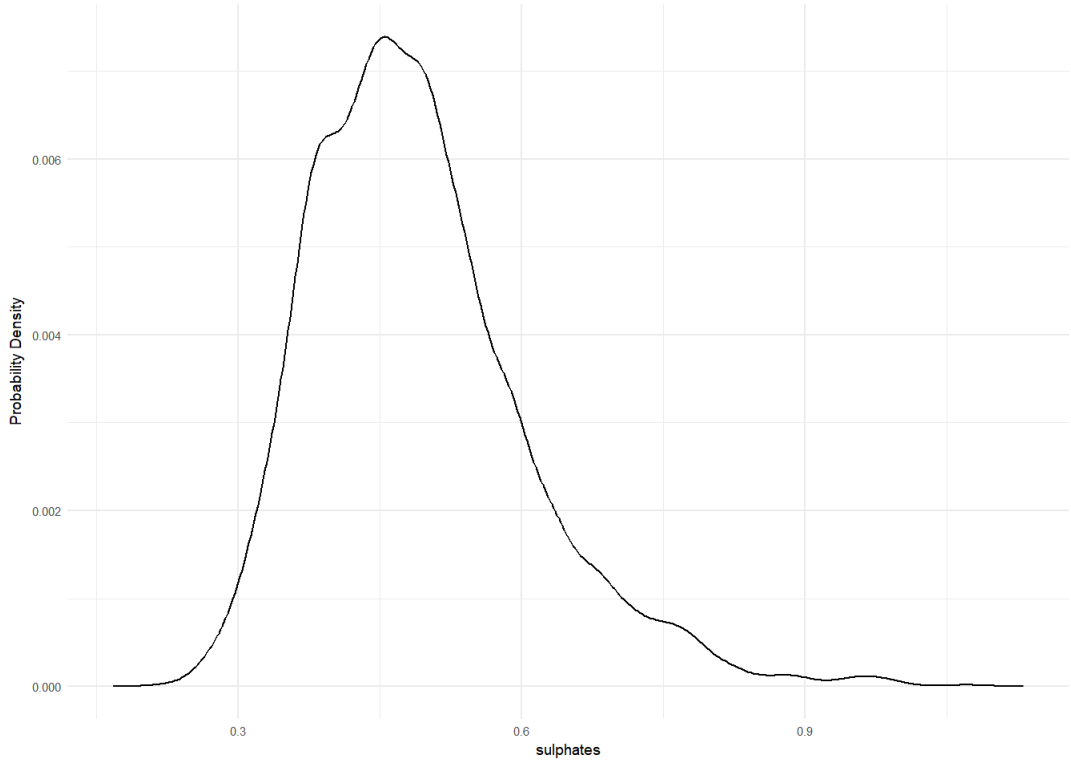

```
## [1] "Variable:  density"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```



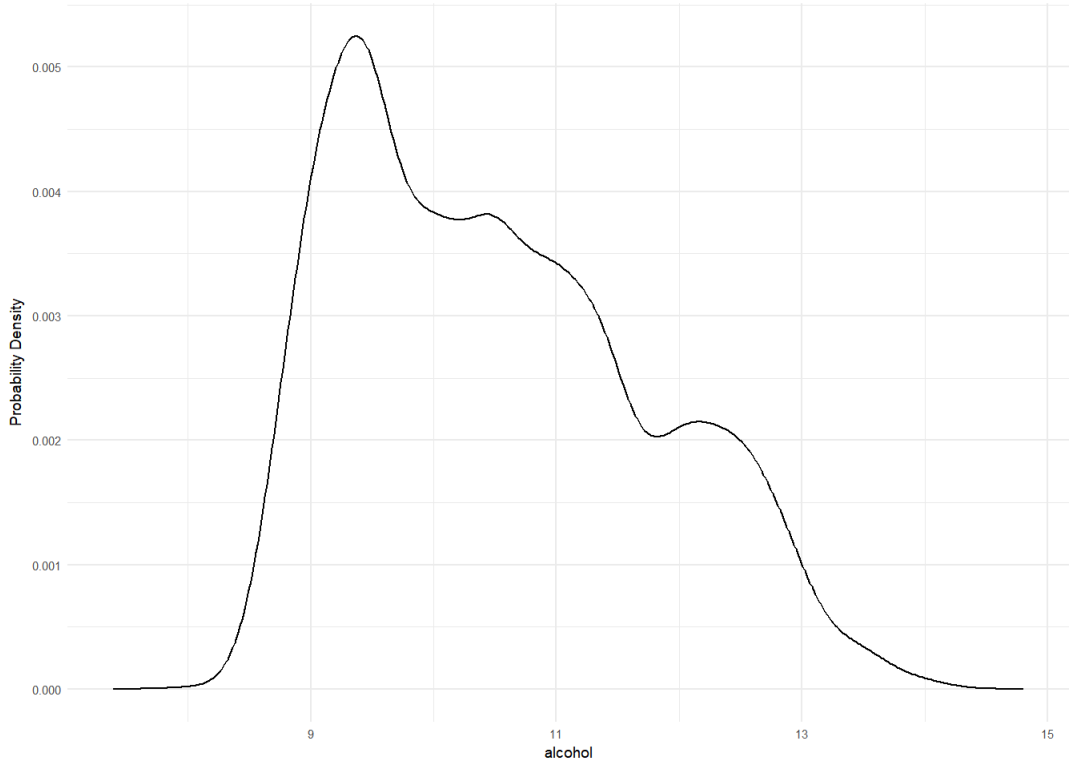

```
## [1] "Variable:  pH"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```




```
## [1] "Variable:  sulphates"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```




```
## [1] "Variable:  alcohol"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```



1. In a quality scale of 0 (very bad) to 10 (very excellent) we can see from the bar plot that the most wine lie in a range of 5 to 7.

2.Variables like fixed.acidity, volatile.acidity, citric.acid, free.sulfur.dioxide,total.sulfur.dioxide and alcohol seems to be follow a poisson distribution but they still have a long tail associated with them on the right side except in case of variable : alcohol.

3.The variables like pH and sulphates have a rough normal distribution.

4.The variable chrolides and residual.sugar have really long tails on the positive side.

Univariate Analysis

What is the structure of your dataset?

```
str(white_wine)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ 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 ...
```

This data-set have information about 4898 variants of wine of same brand. Each wine have 12 varaibles associated with it.

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

The main feature of interest to me in this data set is the quality assocoated with the wine. I want to understand how the other independent variables are related to the Quality of a wine.

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

```
print("Correlation between variables and quality")
```

```
## [1] "Correlation between variables and quality"
```

```
abs(round(cor(white_wine),3))[-12,"quality"]
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                0.114                0.195                0.009
##       residual.sugar            chlorides  free.sulfur.dioxide
##                0.098                0.210                0.008
## total.sulfur.dioxide              density                   pH
##                0.175                0.307                0.099
##            sulphates              alcohol
##                0.054                0.436
```

As we can see that individually the correlation of variables with the quality is weak so it will be required to use them at the same time in predicting the quality.

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

Yes, one

```
white_wine$qlesseqFive <- white_wine$quality <= 5
```

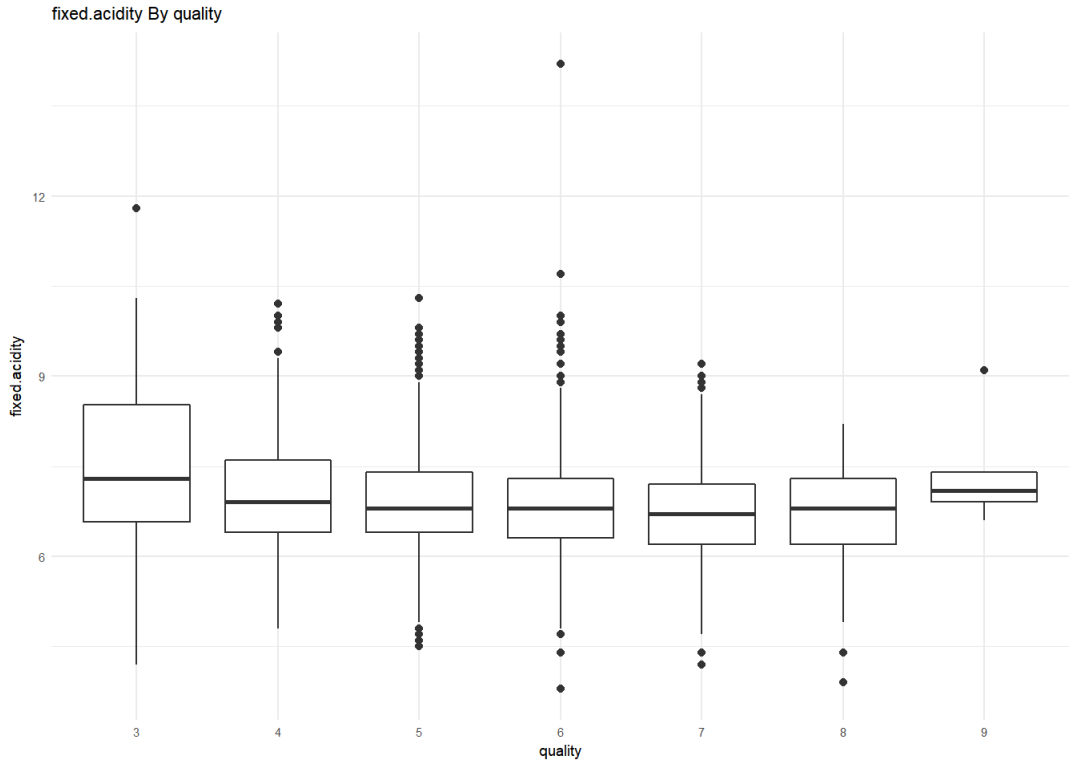
Of the features you investigated, were there any unusual distributions? you perform any operations on the data to tidy, adjust, or change the form the data? If so, why did you do this?

For ease of prediction I am converting the quality variable in to categorical feature by converting it into a factor varaible.

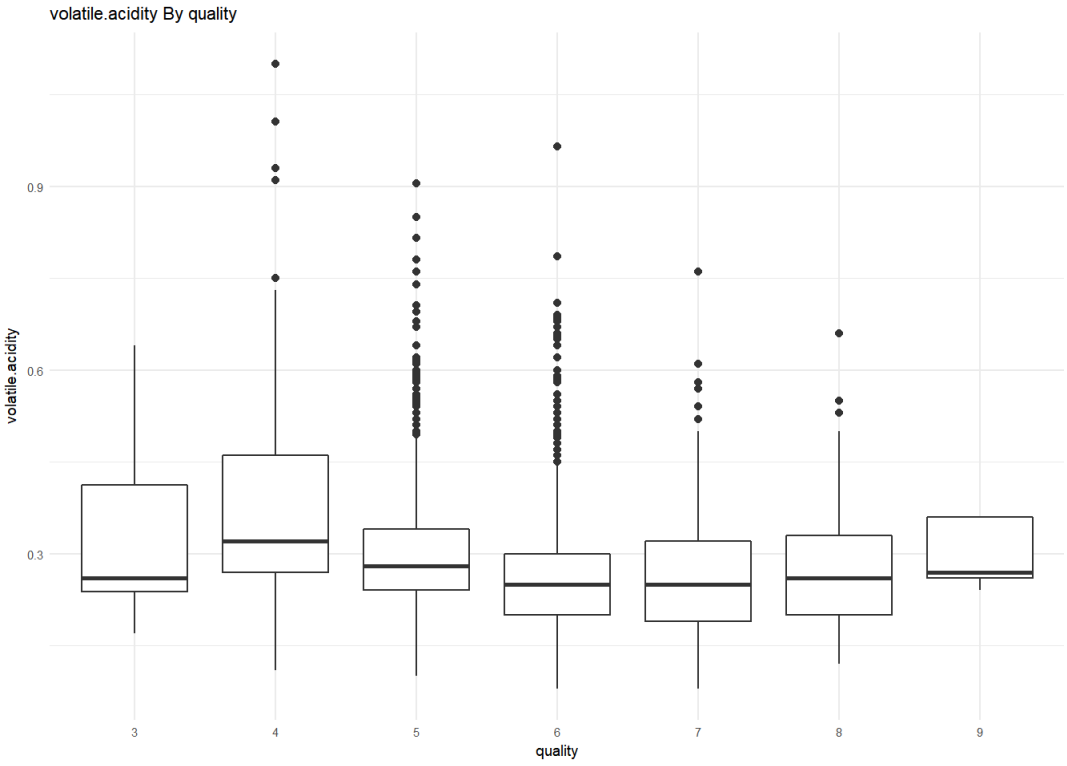
```
white_wine$qualety <- factor(white_wine$quality)
```

Bivariate Plots Section

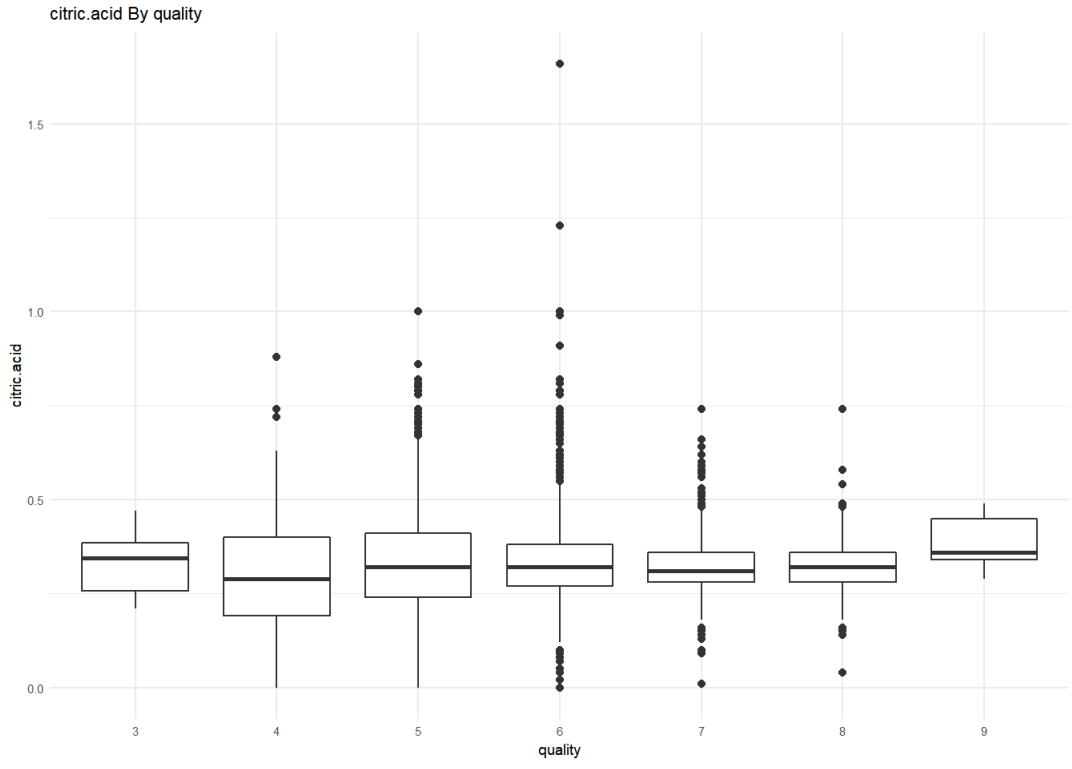
```
## [1] "Summary of  fixed.acidity By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.200   6.575   7.300   7.600   8.525  11.800
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.800   6.400   6.900   7.129   7.600  10.200
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.500   6.400   6.800   6.934   7.400  10.300
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.838   7.300  14.200
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.200   6.200   6.700   6.735   7.200   9.200
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.900   6.200   6.800   6.657   7.300   8.200
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.60    6.90    7.10    7.42    7.40    9.10
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## quality        1     45   45.05   64.08 1.48e-15 ***
## Residuals   4896   3442    0.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



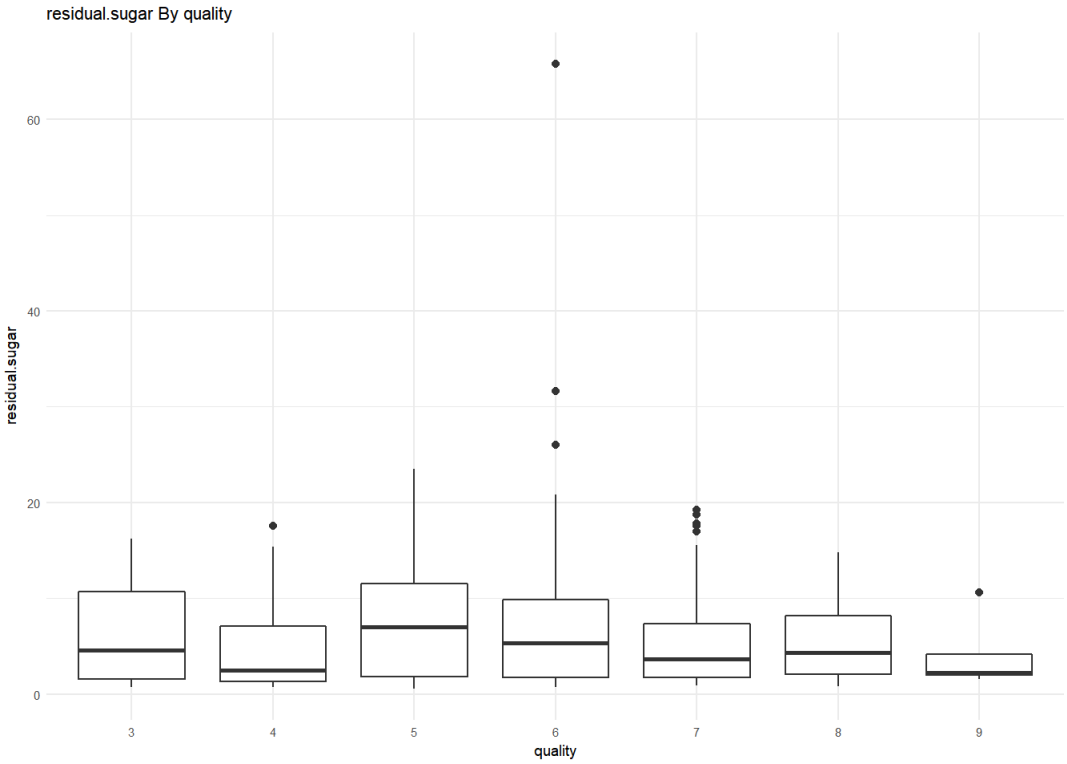
```
## [1] "Summary of  volatile.acidity By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1700  0.2375  0.2600  0.3332  0.4125  0.6400
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1100  0.2700  0.3200  0.3812  0.4600  1.1000
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.240   0.280   0.302   0.340   0.905
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2000  0.2500  0.2606  0.3000  0.9650
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.1900  0.2500  0.2628  0.3200  0.7600
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.2000  0.2600  0.2774  0.3300  0.6600
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.240   0.260   0.270   0.298   0.360   0.360
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value Pr(>F)
## quality        1   1.89  1.8864     193 <2e-16 ***
## Residuals   4896  47.86  0.0098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



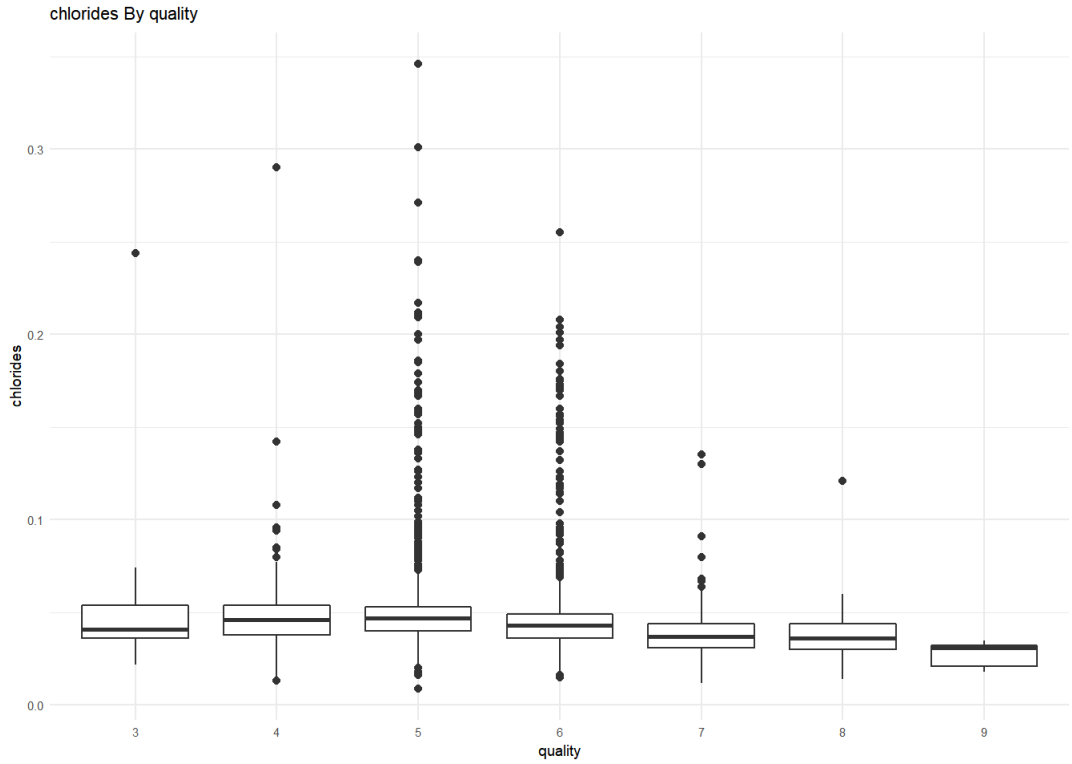
```
## [1] "Summary of  citric.acid By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2100  0.2575  0.3450  0.3360  0.3850  0.4700
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1900  0.2900  0.3042  0.4000  0.8800
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2400  0.3200  0.3377  0.4100  1.0000
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.270   0.320   0.338   0.380   1.660
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.2800  0.3100  0.3256  0.3600  0.7400
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.2800  0.3200  0.3265  0.3600  0.7400
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.290   0.340   0.360   0.386   0.450   0.490
## [1] "One-way ANOVA test"
##              Df Sum Sq  Mean Sq F value Pr(>F)
## quality       1   0.01 0.006082   0.415  0.519
## Residuals  4896  71.71 0.014648
```



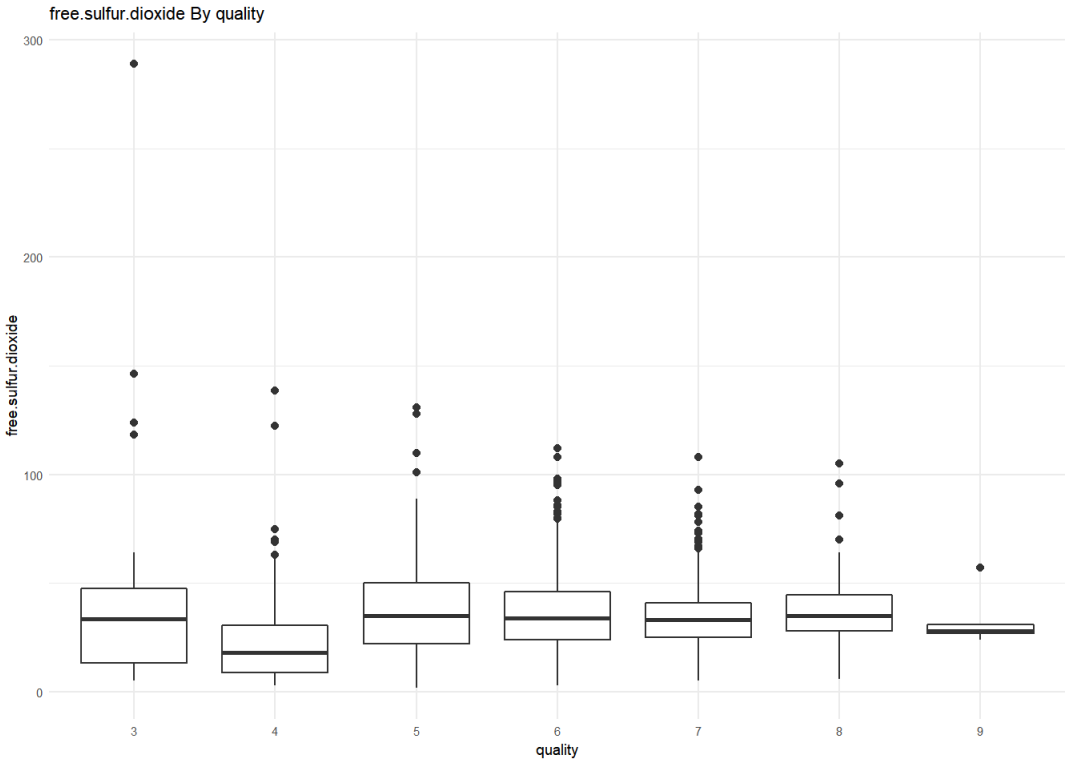
```
## [1] "Summary of  residual.sugar By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.588   4.600   6.392  10.700  16.200
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.300   2.500   4.628   7.100  17.550
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.800   7.000   7.335  11.500  23.500
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.700   5.300   6.442   9.900  65.800
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.700   3.650   5.186   7.325  19.250
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   2.100   4.300   5.671   8.200  14.800
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.60    2.00    2.20    4.12    4.20   10.60
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## quality        1   1199  1199.5   47.06 7.72e-12 ***
## Residuals   4896 124780    25.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



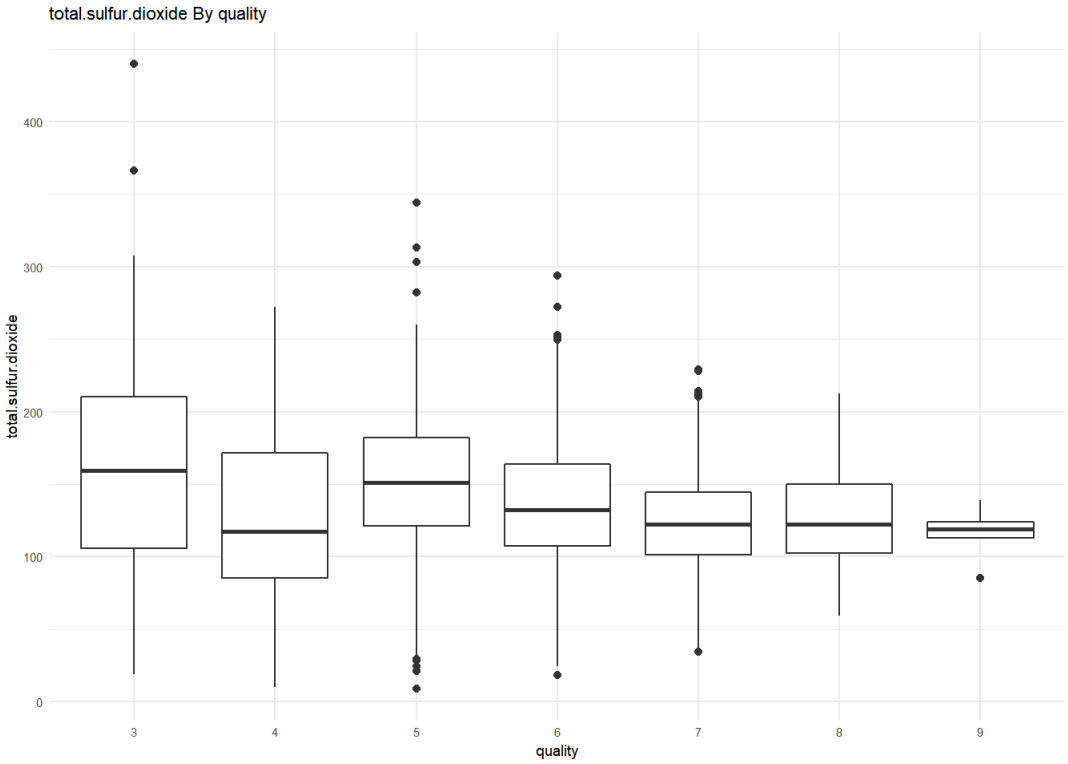
```
## [1] "Summary of  chlorides By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02200 0.03625 0.04100 0.05430 0.05400 0.24400
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0130  0.0380  0.0460  0.0501  0.0540  0.2900
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.04000 0.04700 0.05155 0.05300 0.34600
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01500 0.03600 0.04300 0.04522 0.04900 0.25500
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03819 0.04400 0.13500
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01400 0.03000 0.03600 0.03831 0.04400 0.12100
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0180  0.0210  0.0310  0.0274  0.0320  0.0350
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value Pr(>F)
## quality        1  0.103 0.10302   225.7 <2e-16 ***
## Residuals   4896  2.235 0.00046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



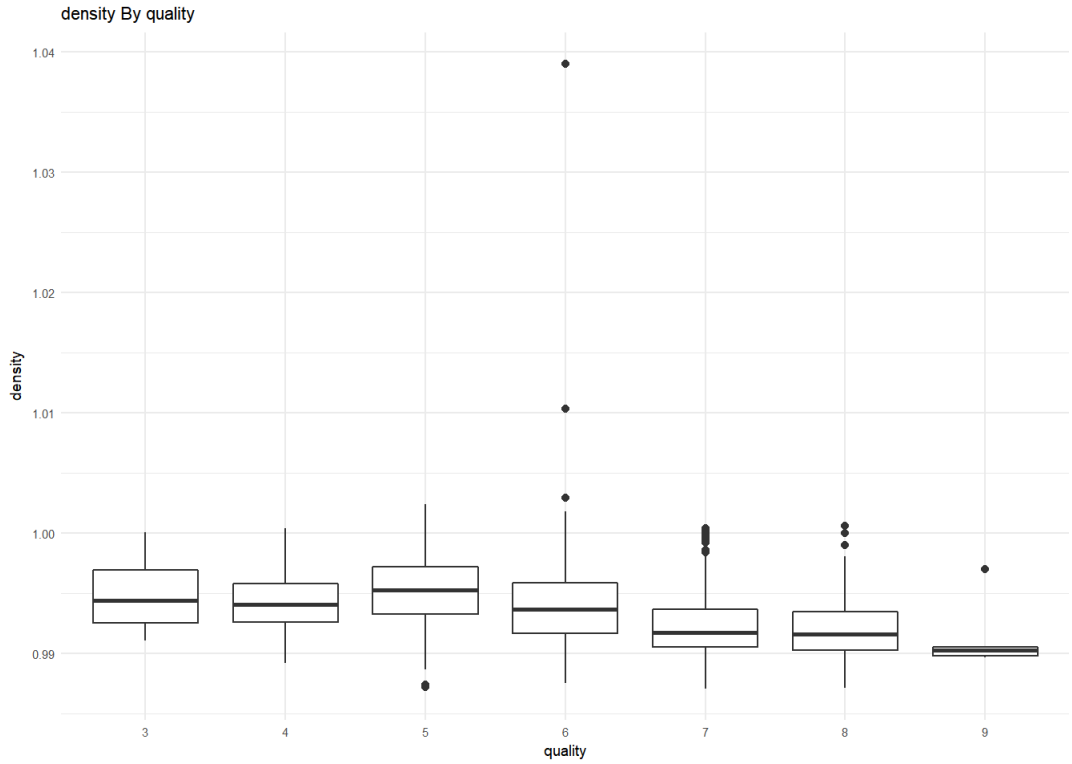
```
## [1] "Summary of  free.sulfur.dioxide By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.00   13.25   33.50   53.32   47.50  289.00 
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    9.00   18.00   23.36   30.50  138.50 
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2.00   22.00   35.00   36.43   50.00  131.00 
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00   24.00   34.00   35.65   46.00  112.00 
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.00   25.00   33.00   34.13   41.00  108.00 
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   28.00   35.00   36.72   44.50  105.00 
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    24.0    27.0    28.0    33.4    31.0    57.0 
## [1] "One-way ANOVA test"
##               Df  Sum Sq Mean Sq F value Pr(>F)
## quality        1      94   94.27   0.326  0.568
## Residuals   4896 1416327  289.28
```



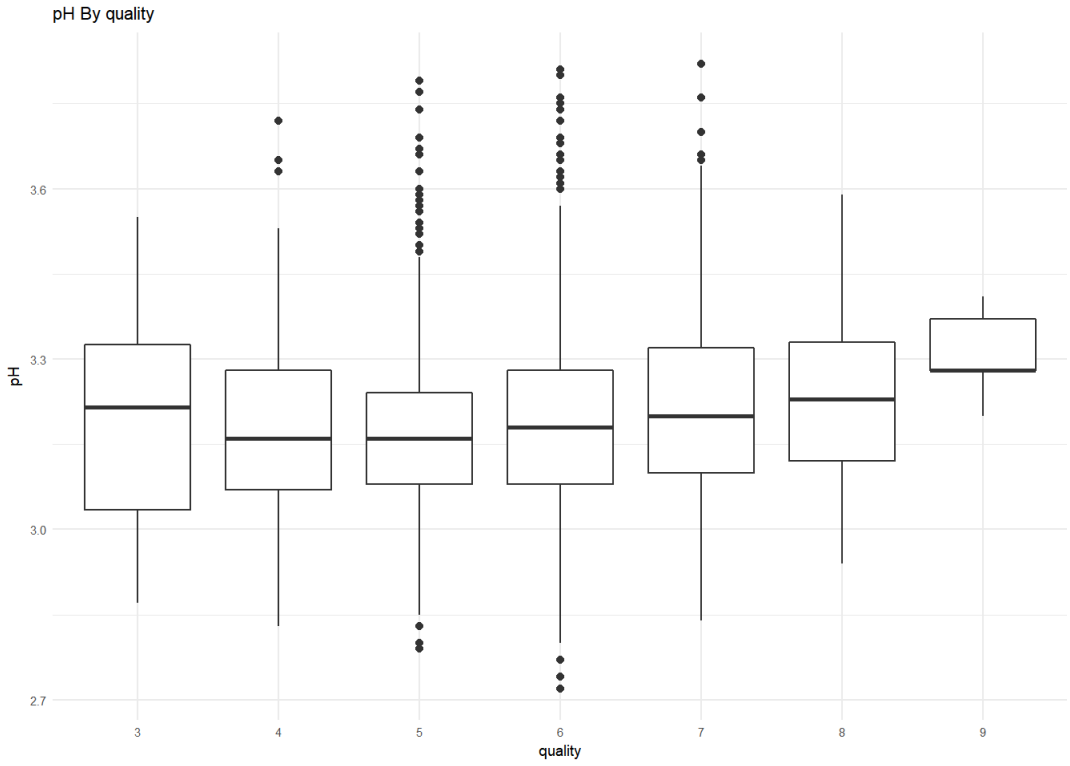
```
## [1] "Summary of  total.sulfur.dioxide By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    19.0   105.8   159.5   170.6   210.0   440.0
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    85.0   117.0   125.3   171.5   272.0
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   121.0   151.0   150.9   182.0   344.0
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0   107.2   132.0   137.0   164.0   294.0
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    34.0   101.0   122.0   125.1   144.2   229.0
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    59.0   102.5   122.0   126.2   150.0   212.5
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      85     113     119     116     124     139
## [1] "One-way ANOVA test"
##               Df  Sum Sq Mean Sq F value Pr(>F)
## quality        1  270047  270047   154.2 <2e-16 ***
## Residuals   4896 8574354    1751
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



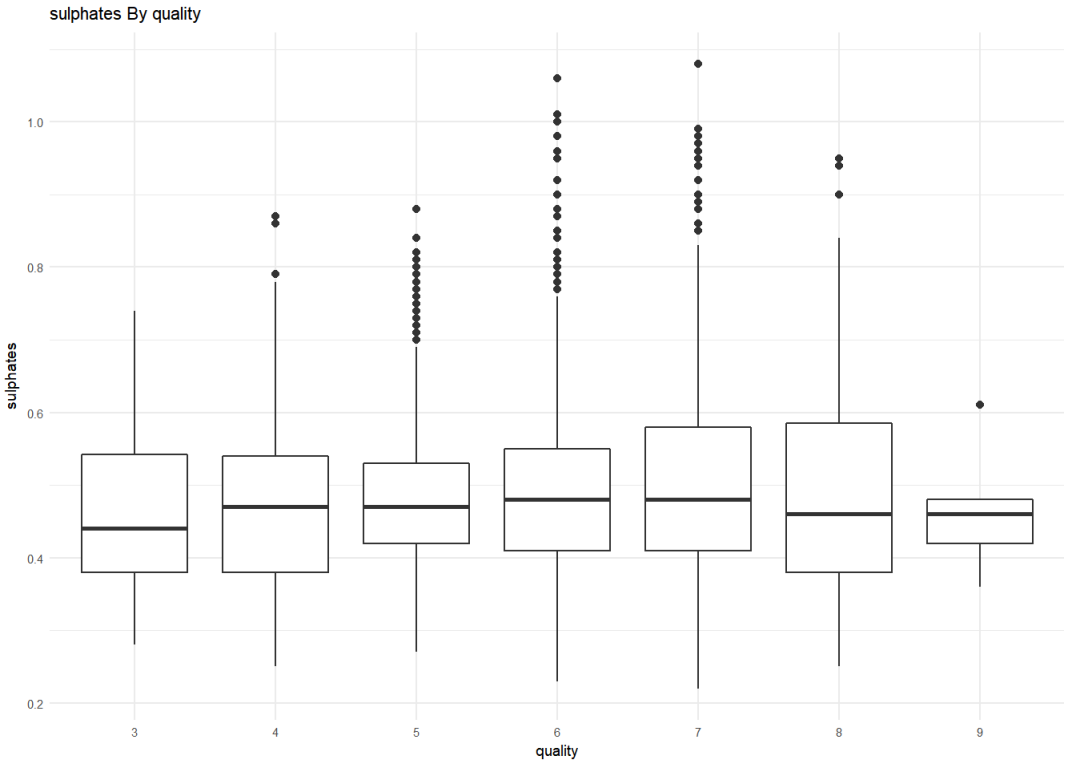
```
## [1] "Summary of  density By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9911  0.9925  0.9944  0.9949  0.9969  1.0000
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9892  0.9926  0.9941  0.9943  0.9958  1.0000
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9933  0.9953  0.9953  0.9972  1.0020
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9906  0.9918  0.9925  0.9937  1.0000
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9903  0.9916  0.9922  0.9935  1.0010
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9896  0.9898  0.9903  0.9915  0.9906  0.9970
## [1] "One-way ANOVA test"
##               Df  Sum Sq  Mean Sq F value Pr(>F)
## quality        1 0.00413 0.004132   509.9 <2e-16 ***
## Residuals   4896 0.03967 0.000008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



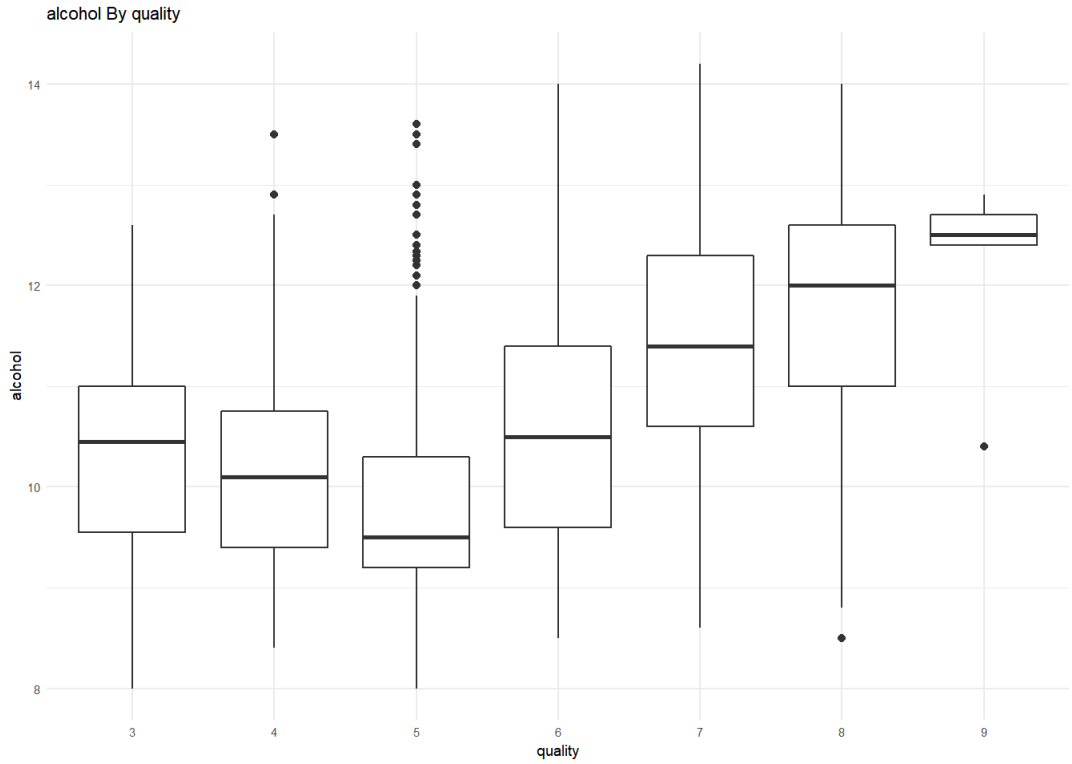
```
## [1] "Summary of  pH By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.870   3.035   3.215   3.188   3.325   3.550
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.830   3.070   3.160   3.183   3.280   3.720
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.790   3.080   3.160   3.169   3.240   3.790
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.080   3.180   3.189   3.280   3.810
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.840   3.100   3.200   3.214   3.320   3.820
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.940   3.120   3.230   3.219   3.330   3.590
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.200   3.280   3.280   3.308   3.370   3.410
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## quality        1    1.1  1.1038   48.88 3.08e-12 ***
## Residuals   4896  110.5  0.0226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
## [1] "Summary of  sulphates By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2800  0.3800  0.4400  0.4745  0.5425  0.7400
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.3800  0.4700  0.4761  0.5400  0.8700
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2700  0.4200  0.4700  0.4822  0.5300  0.8800
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.4100  0.4800  0.4911  0.5500  1.0600
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4800  0.5031  0.5800  1.0800
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.3800  0.4600  0.4862  0.5850  0.9500
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.360   0.420   0.460   0.466   0.480   0.610
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## quality        1   0.18 0.18378   14.15 0.000171 ***
## Residuals   4896  63.60 0.01299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
## [1] "Summary of  alcohol By quality"
## white_wine$qualety: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.34   11.00   12.60
## ------------------------------------------------------------
## white_wine$qualety: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## white_wine$qualety: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## white_wine$qualety: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## white_wine$qualety: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
## white_wine$qualety: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
## white_wine$qualety: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   12.40   12.50   12.18   12.70   12.90
## [1] "One-way ANOVA test"
##               Df Sum Sq Mean Sq F value Pr(>F)
## quality        1   1407  1407.0    1146 <2e-16 ***
## Residuals   4896   6009     1.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
## [1] "Correlation matrix for independent variables"
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.000            0.023       0.289
## volatile.acidity              0.023            1.000       0.149
## citric.acid                   0.289            0.149       1.000
## residual.sugar                0.089            0.064       0.094
## chlorides                     0.023            0.071       0.114
## free.sulfur.dioxide           0.049            0.097       0.094
## total.sulfur.dioxide          0.091            0.089       0.121
## density                       0.265            0.027       0.150
## pH                            0.426            0.032       0.164
## sulphates                     0.017            0.036       0.062
## alcohol                       0.121            0.068       0.076
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                  0.089     0.023                0.049
## volatile.acidity               0.064     0.071                0.097
## citric.acid                    0.094     0.114                0.094
## residual.sugar                 1.000     0.089                0.299
## chlorides                      0.089     1.000                0.101
## free.sulfur.dioxide            0.299     0.101                1.000
## total.sulfur.dioxide           0.401     0.199                0.616
## density                        0.839     0.257                0.294
## pH                             0.194     0.090                0.001
## sulphates                      0.027     0.017                0.059
## alcohol                        0.451     0.360                0.250
##                       total.sulfur.dioxide density    pH sulphates alcohol
## fixed.acidity                        0.091   0.265 0.426     0.017   0.121
## volatile.acidity                     0.089   0.027 0.032     0.036   0.068
## citric.acid                          0.121   0.150 0.164     0.062   0.076
## residual.sugar                       0.401   0.839 0.194     0.027   0.451
## chlorides                            0.199   0.257 0.090     0.017   0.360
## free.sulfur.dioxide                  0.616   0.294 0.001     0.059   0.250
## total.sulfur.dioxide                 1.000   0.530 0.002     0.135   0.449
## density                              0.530   1.000 0.094     0.074   0.780
## pH                                   0.002   0.094 1.000     0.156   0.121
## sulphates                            0.135   0.074 0.156     1.000   0.017
## alcohol                              0.449   0.780 0.121     0.017   1.000
```

1.volatile.acidity, density and pH tends to decrease as the quality of wine increases.

2.citric.acid, sulphates and alcohol increases with the quality of wine.

3.fixed.acidity, residual.sugar and chlorides seems to be stagnat incomparison with the change in wine quality.

4.free.sulfur.dioxide and total.sulfur.dioxid is lower in case of both low and high quality wines. But in case of wines which fall in between the these variables shows higher values.

5. As the correlaion coefficients between free.sulfur.dioxide and total.sulfur dioxide is greater than 0.6 so we can use one of the related variables to build a model.

Bivariate Analysis

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

On aiming for a higher quality of wine we can observe that the variables like volatile.acidity, density, pH, citric.acid, sulphates, alcohol changes.

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

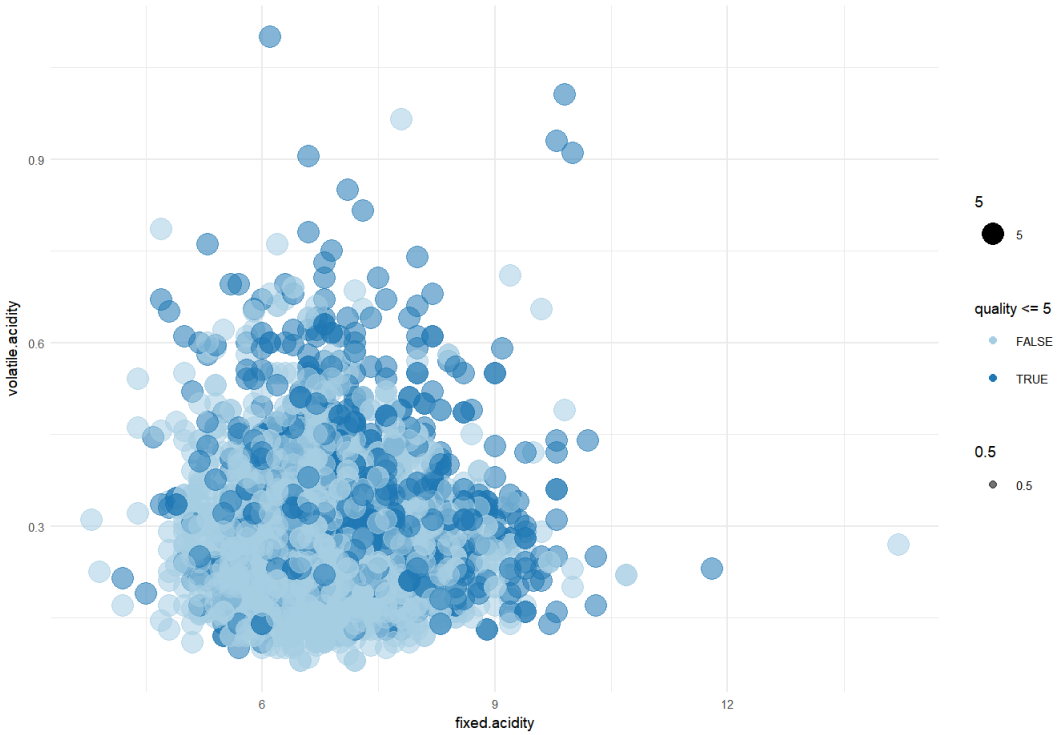
The most interesting relationship that I seem to find is that of density and residual.sugar with a correlation of 0.839

What was the strongest relationship you found?

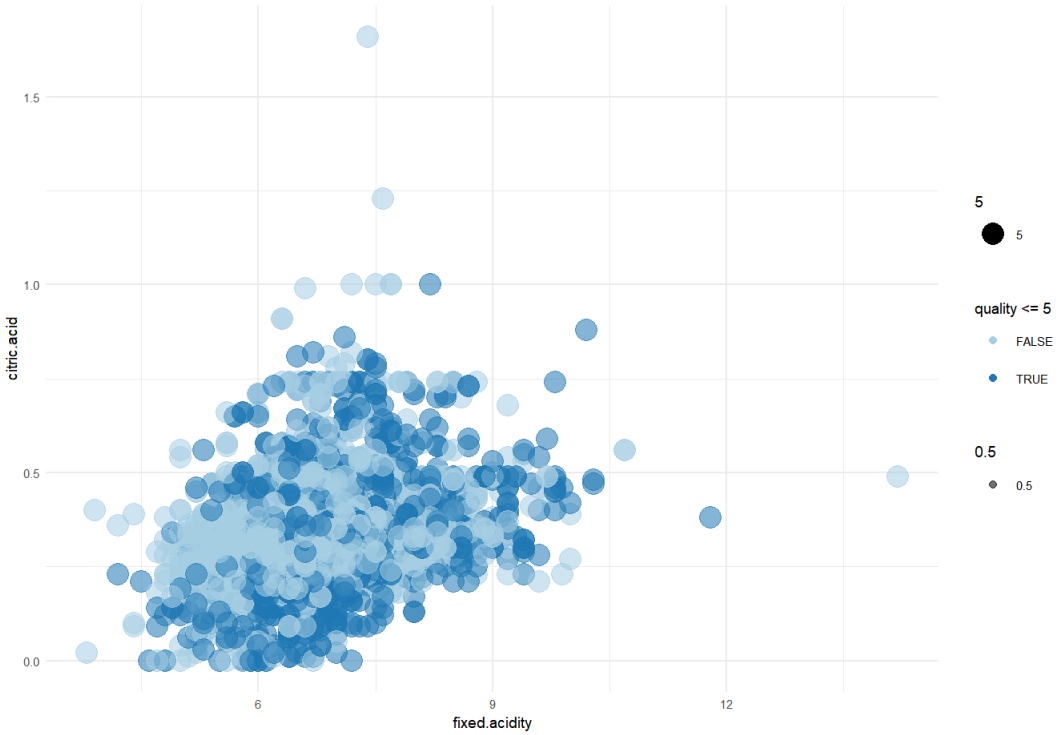
The strongest relationship that I found was that of residual.sugar and density with the correlation coefficient being 0.839 which is the highest among any pssible pair of variables.

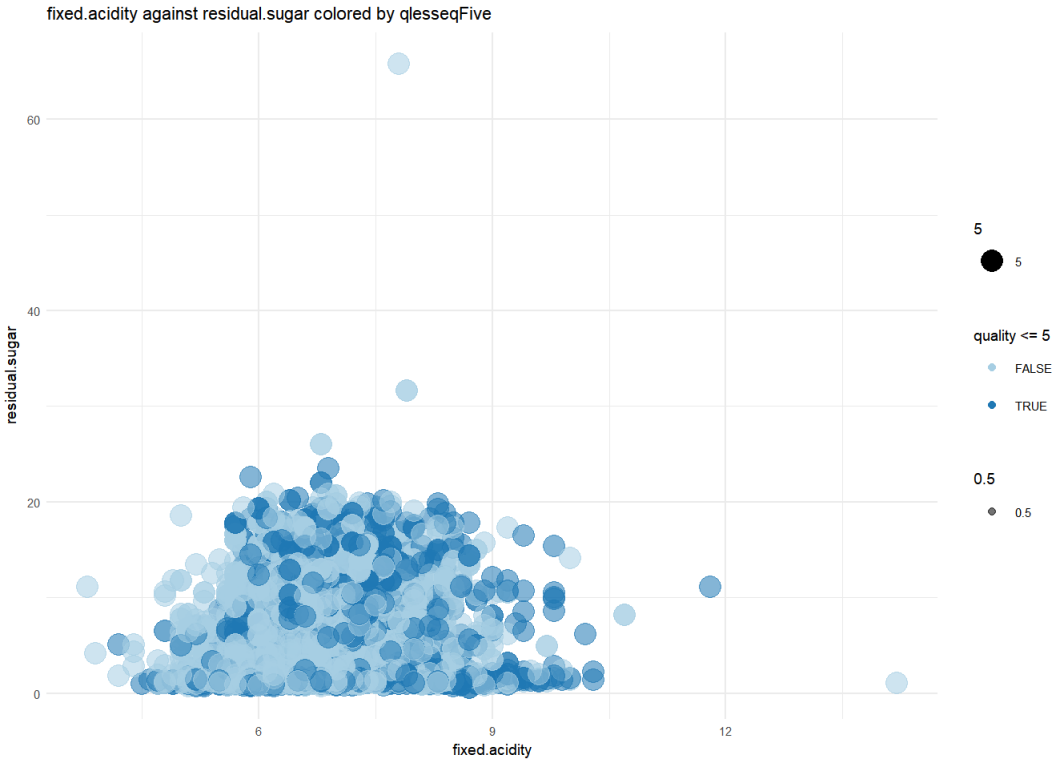
Multivariate Plots Section

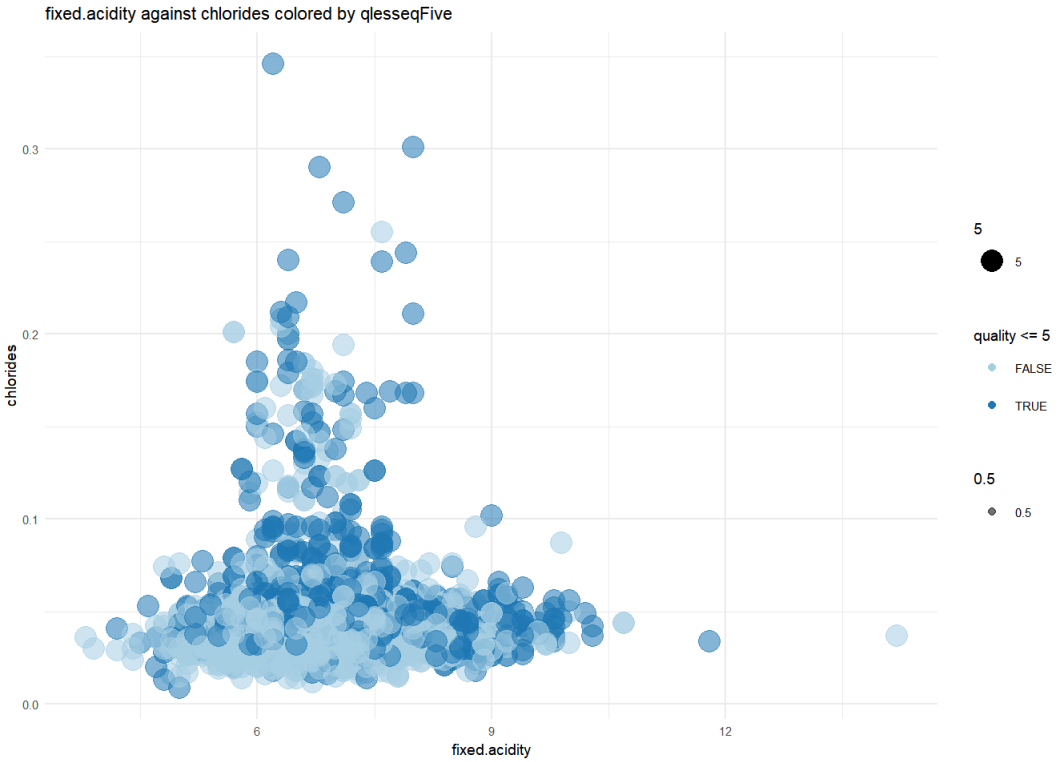
fixed.acidity against volatile.acidity colored by qlesseqFive



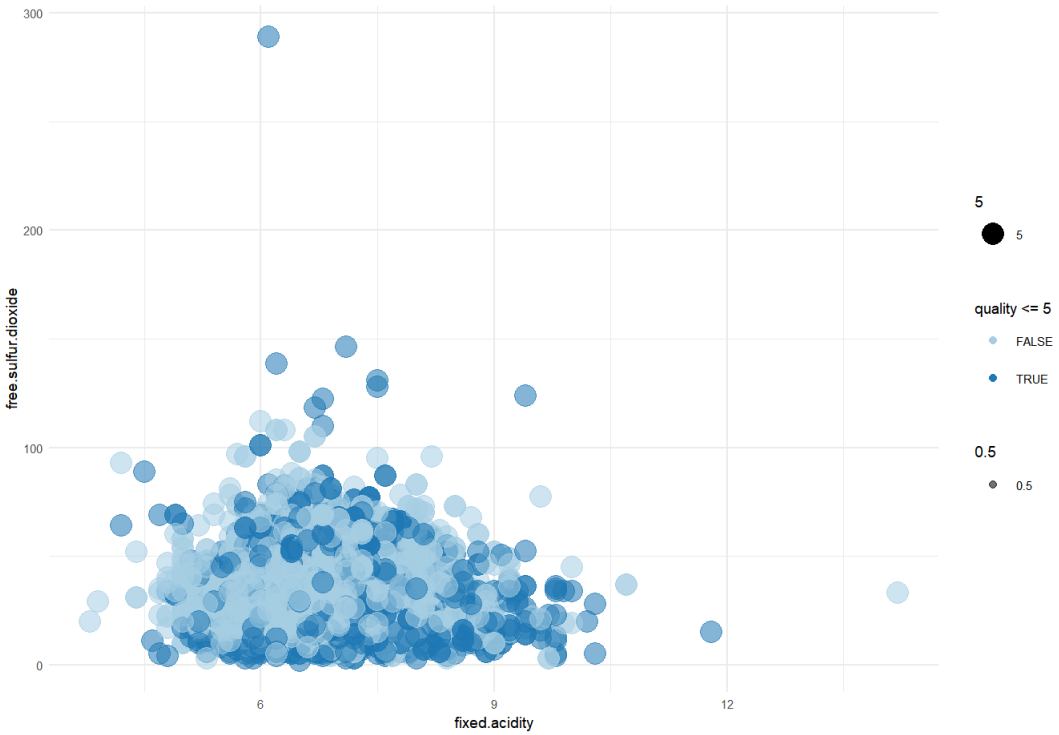
fixed.acidity against citric.acid colored by qlesseqFive



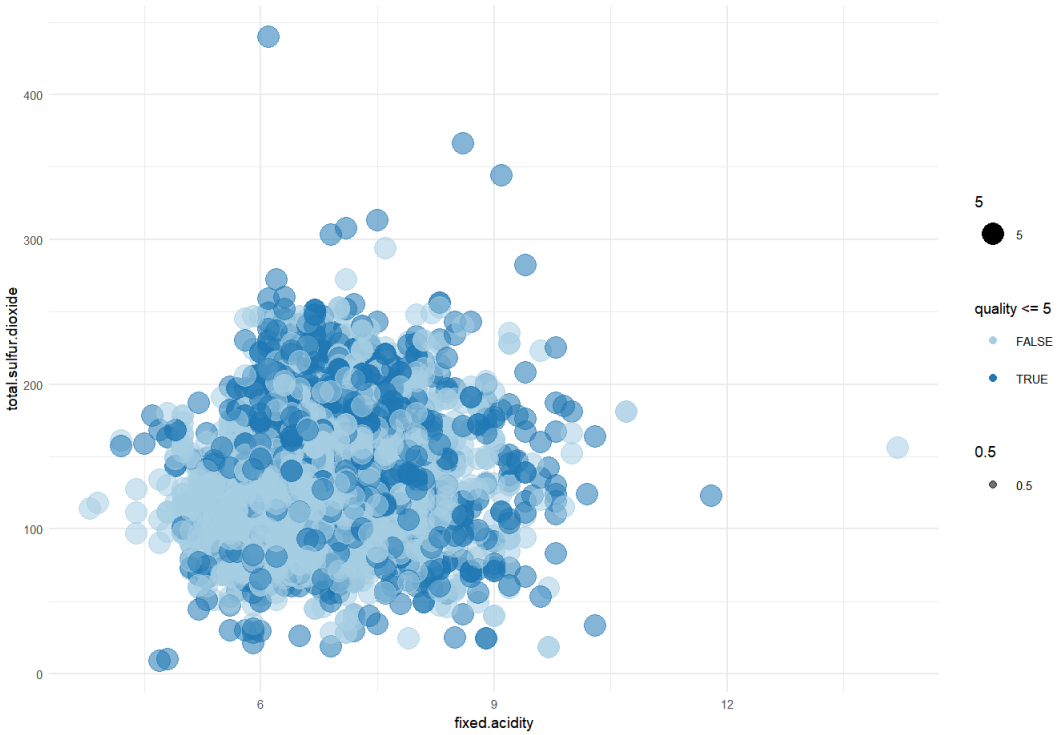

fixed.acidity against residual.sugar colored by qlesseqFive
residual.sugar
fixed.acidity
0
20
40
60
6
9
12
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


fixed.acidity against chlorides colored by qlesseqFive
chlorides
fixed.acidity
0.0
0.1
0.2
0.3
6
9
12
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

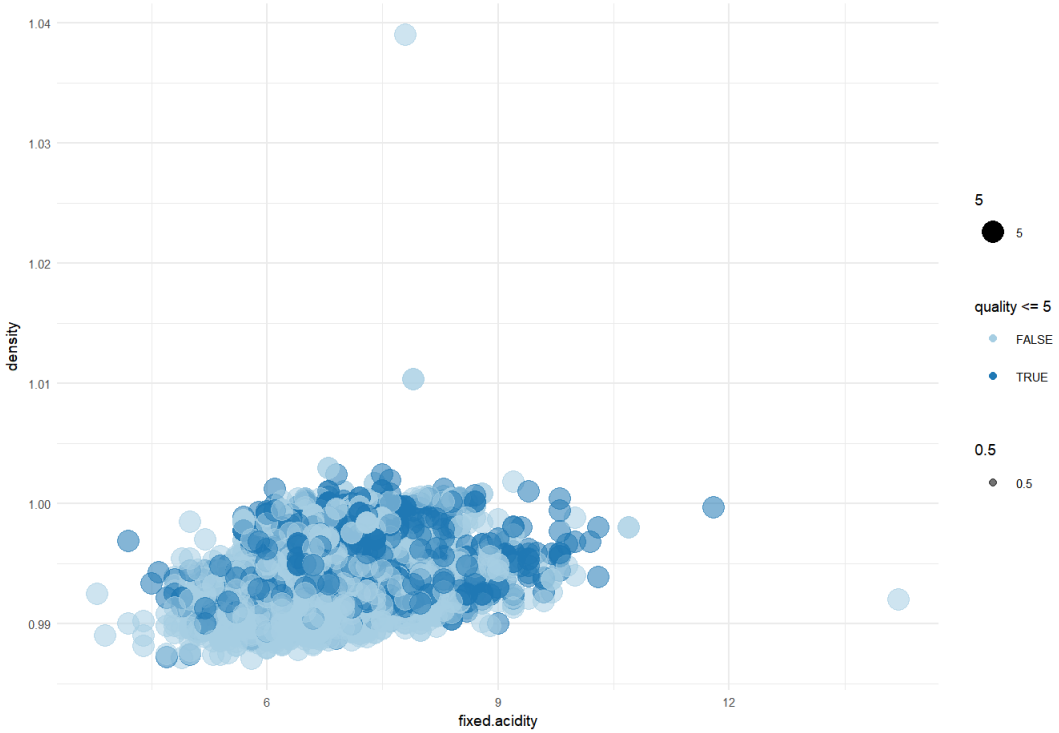
fixed.acidity against free.sulfur.dioxide colored by qlesseqFive



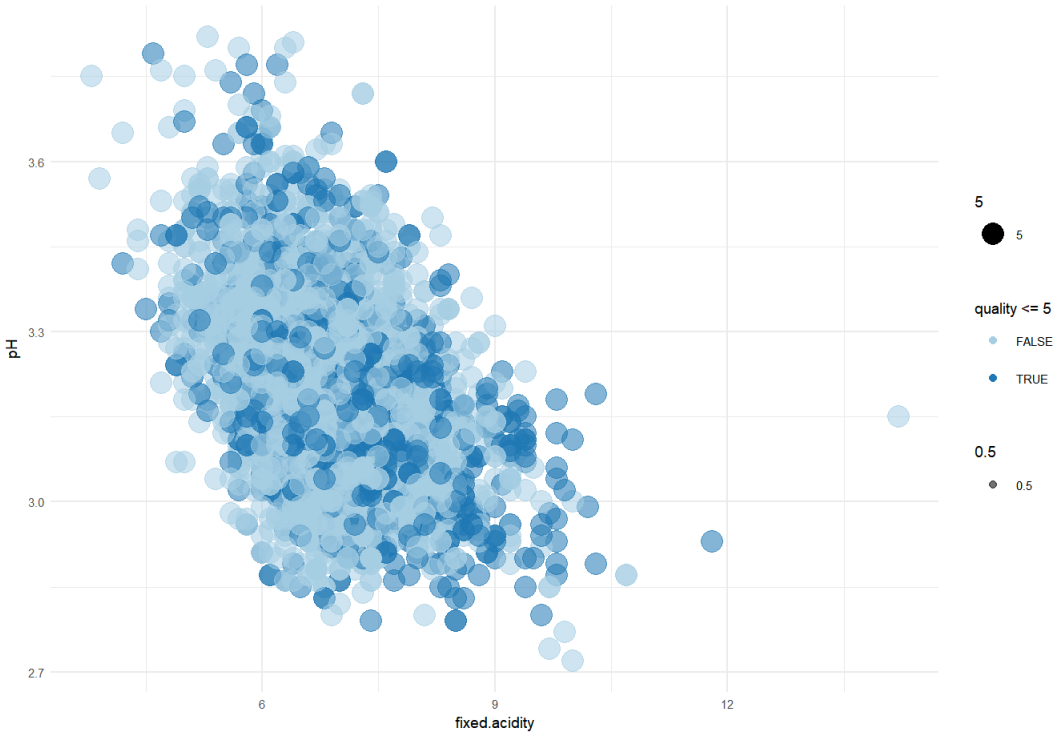
fixed.acidity against total.sulfur.dioxide colored by qlesseqFive

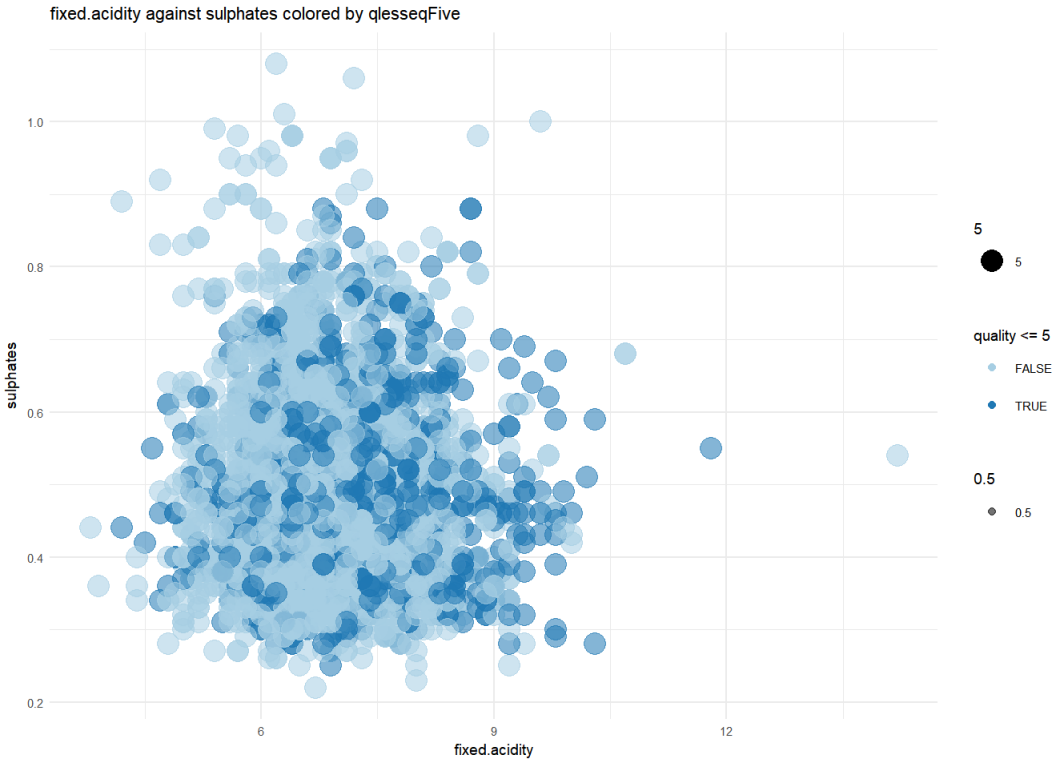


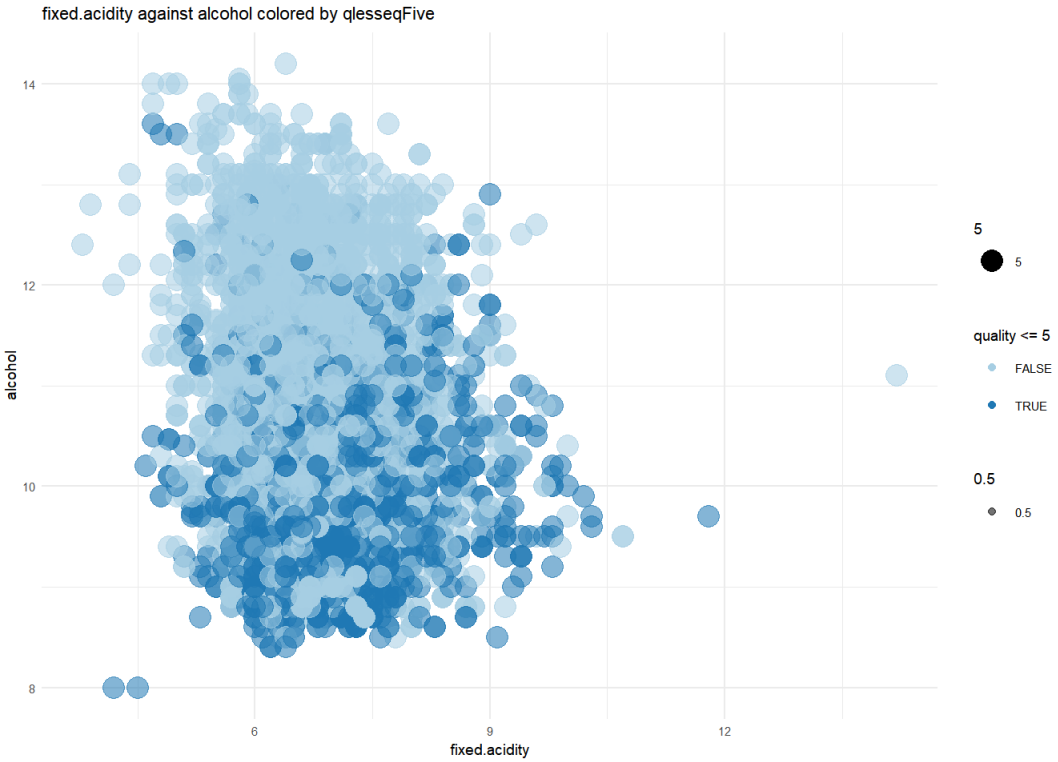
fixed.acidity against density colored by qlesseqFive



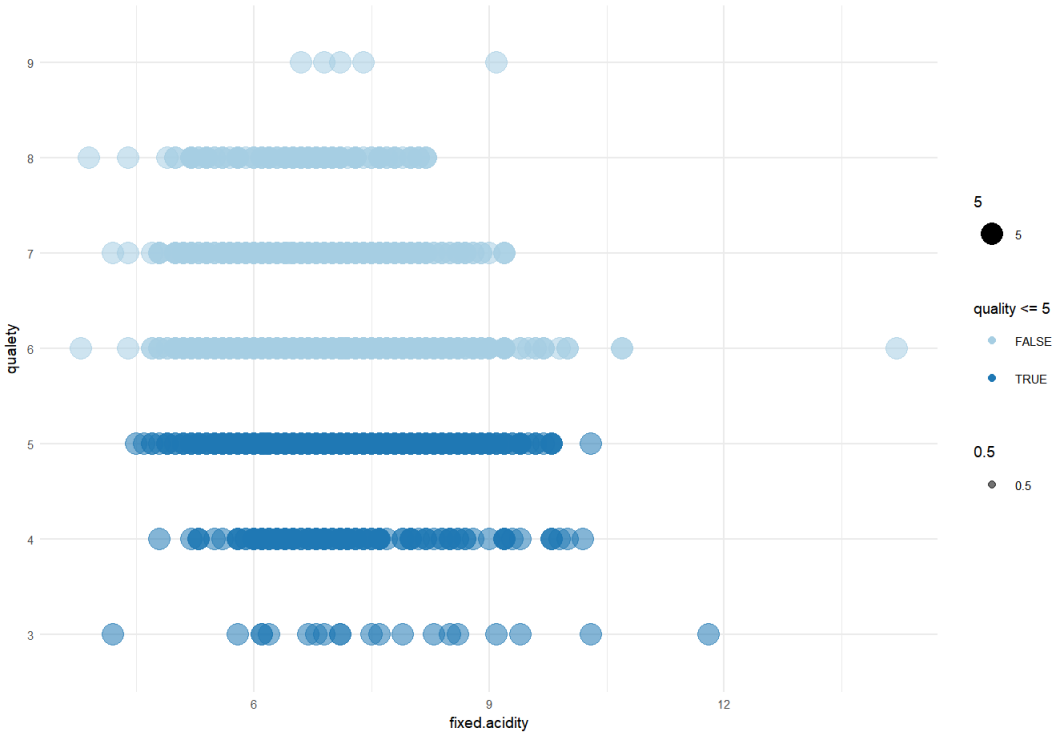
fixed.acidity against pH colored by qlesseqFive



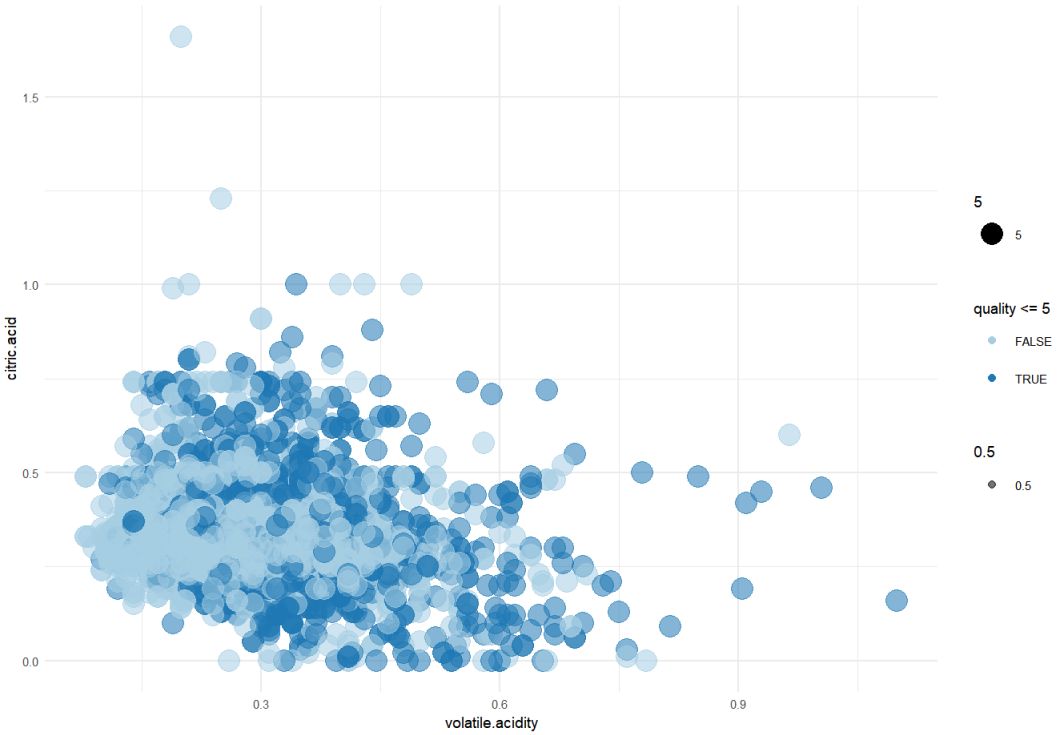

fixed.acidity against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
6
9
12
fixed.acidity
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


fixed.acidity against alcohol colored by qlesseqFive
alcohol
14
12
10
8
6
9
12
fixed.acidity
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

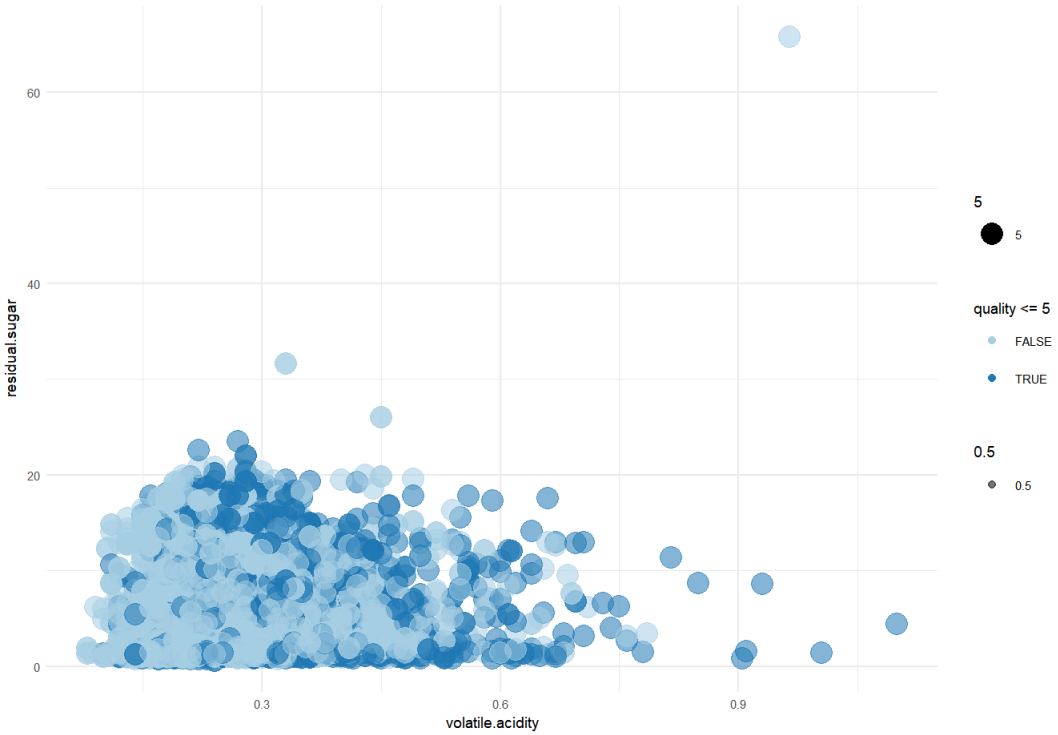
fixed.acidity against qualety colored by qlesseqFive



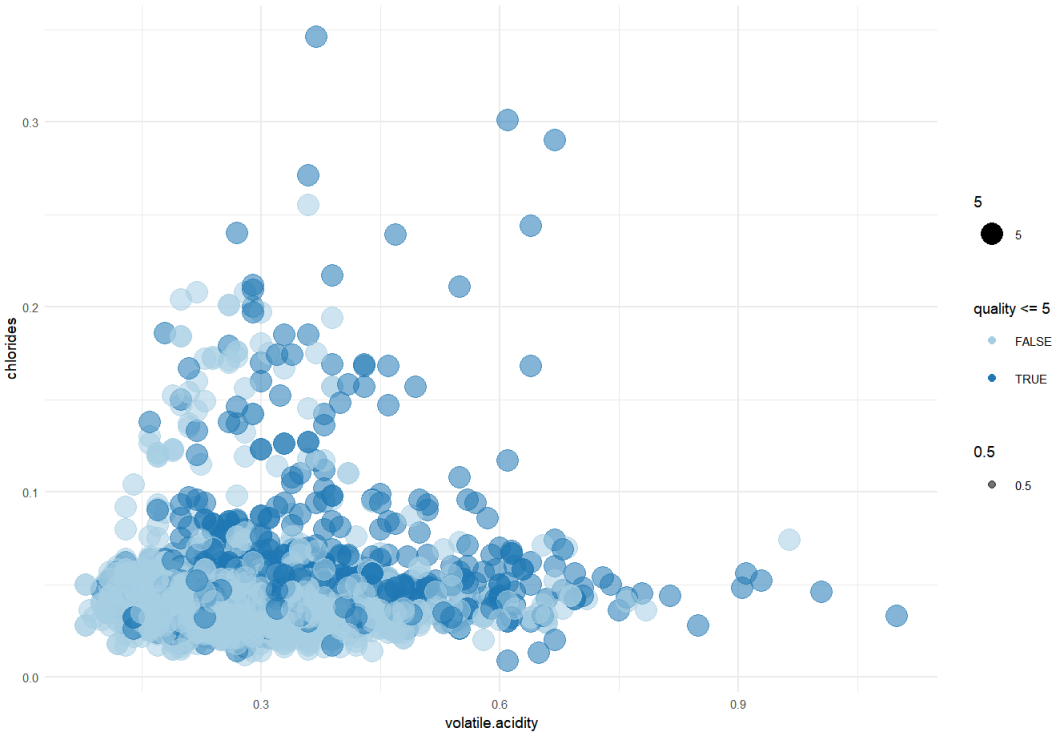
volatile.acidity against citric.acid colored by qlesseqFive



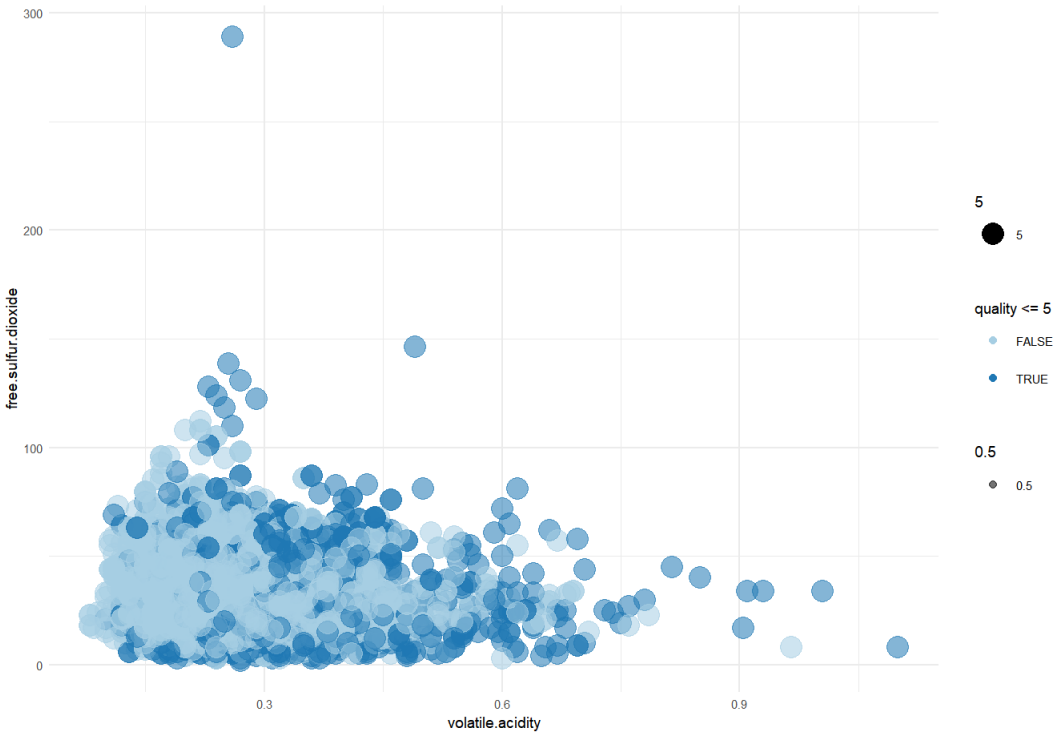
volatile.acidity against residual.sugar colored by qlesseqFive



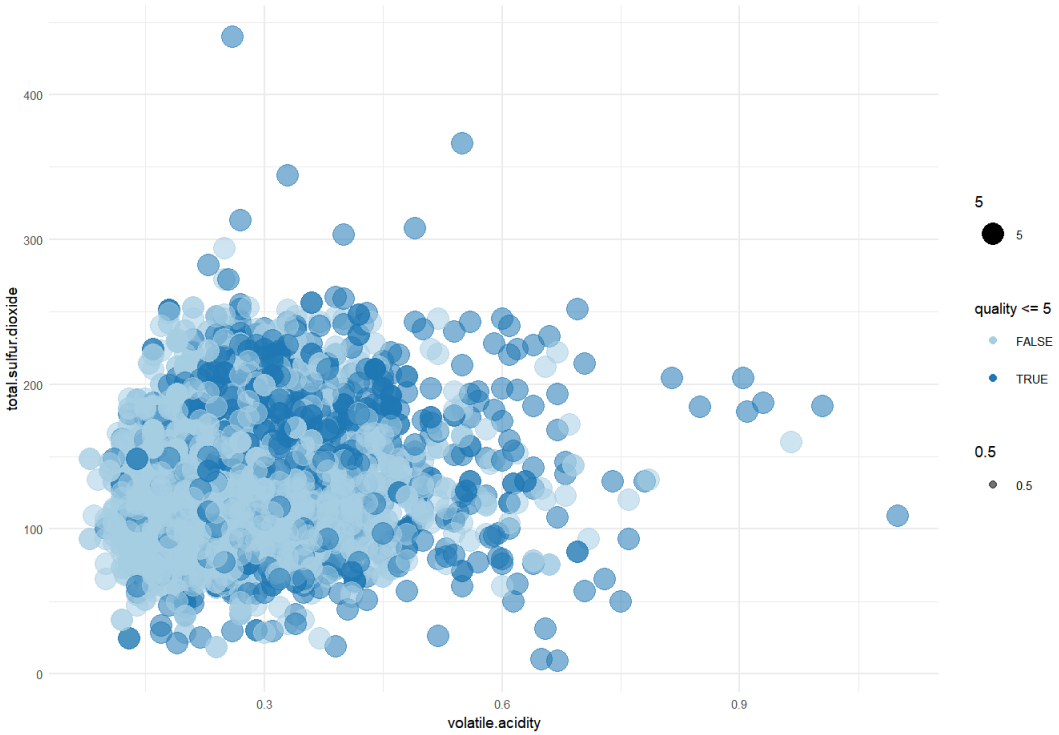
volatile.acidity against chlorides colored by qlesseqFive

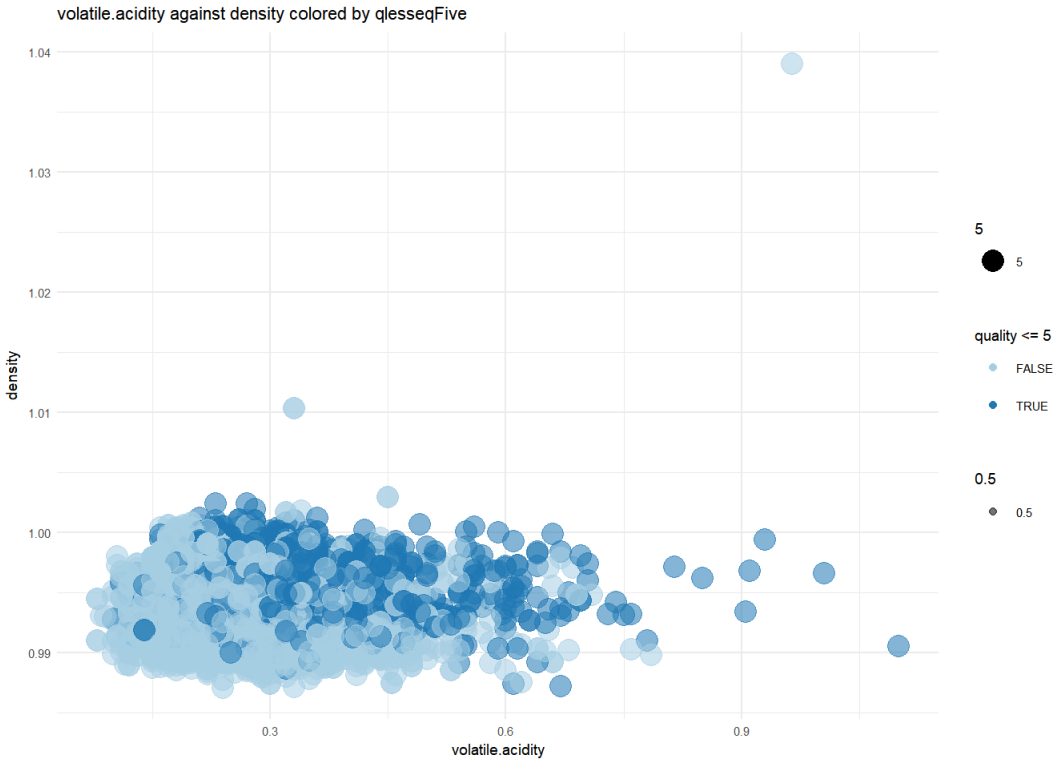


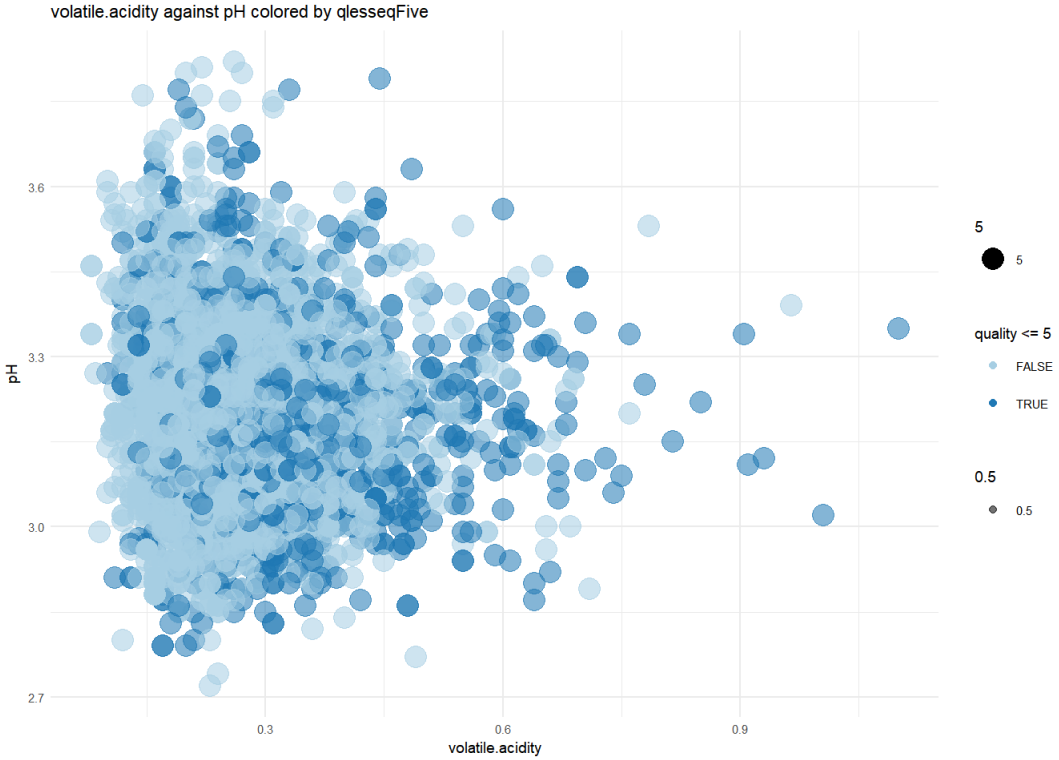
volatile.acidity against free.sulfur.dioxide colored by qlesseqFive



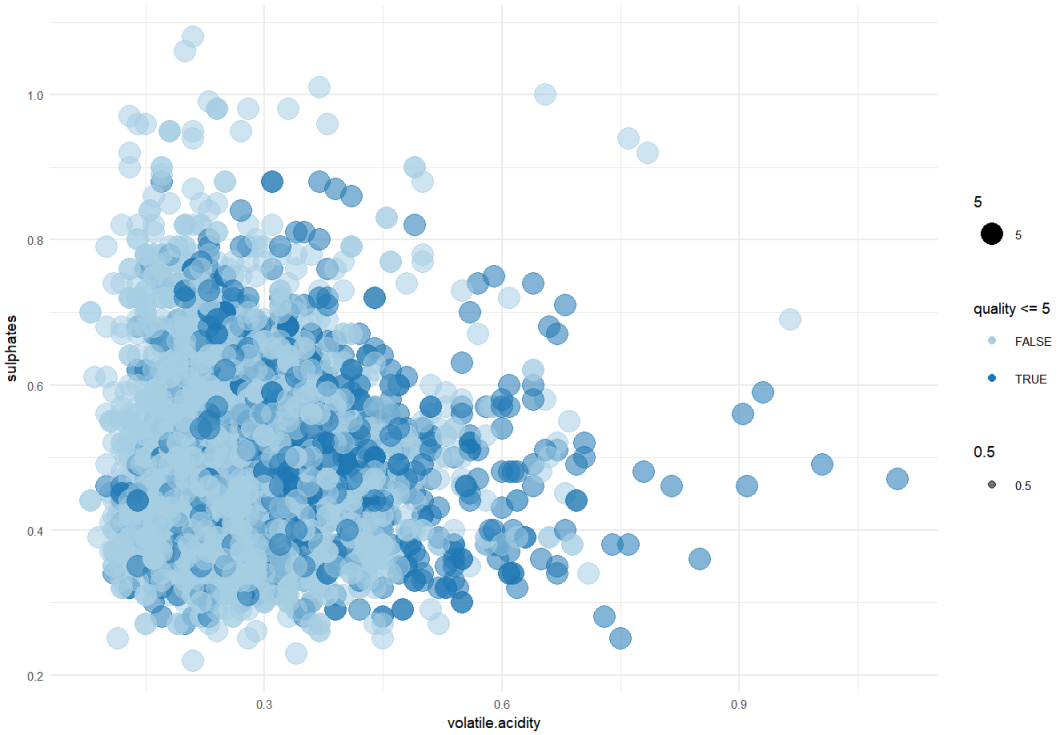
volatile.acidity against total.sulfur.dioxide colored by qlesseqFive



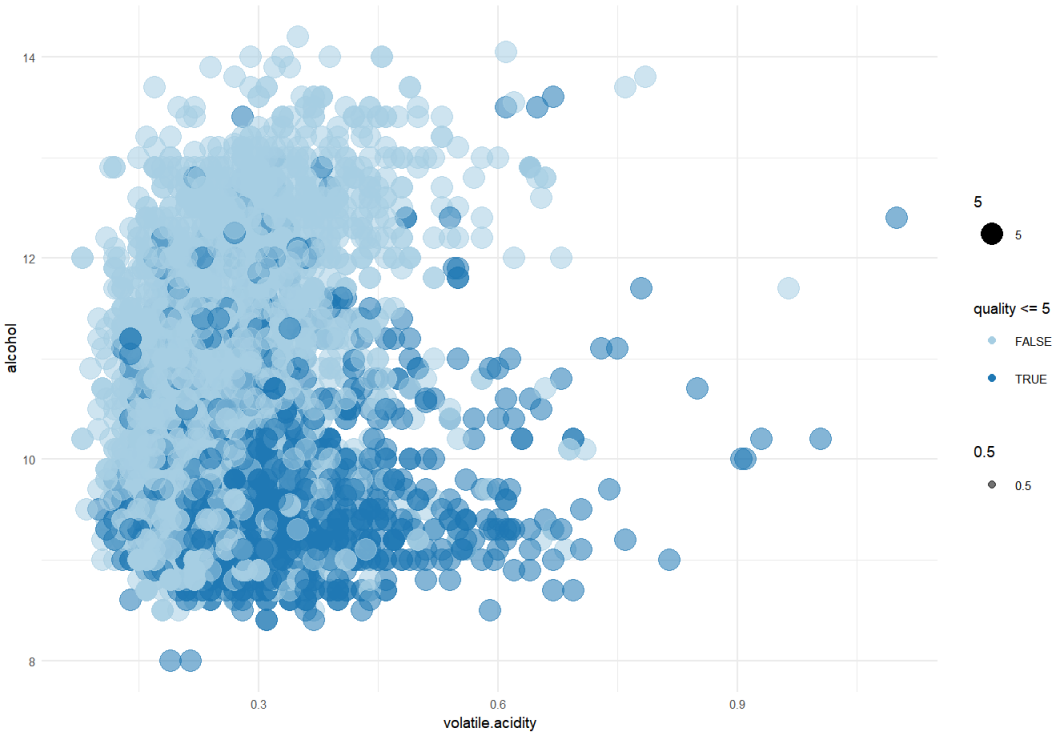

volatile.acidity against density colored by qlesseqFive
density
1.04
1.03
1.02
1.01
1.00
0.99
0.3
0.6
0.9
volatile.acidity
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

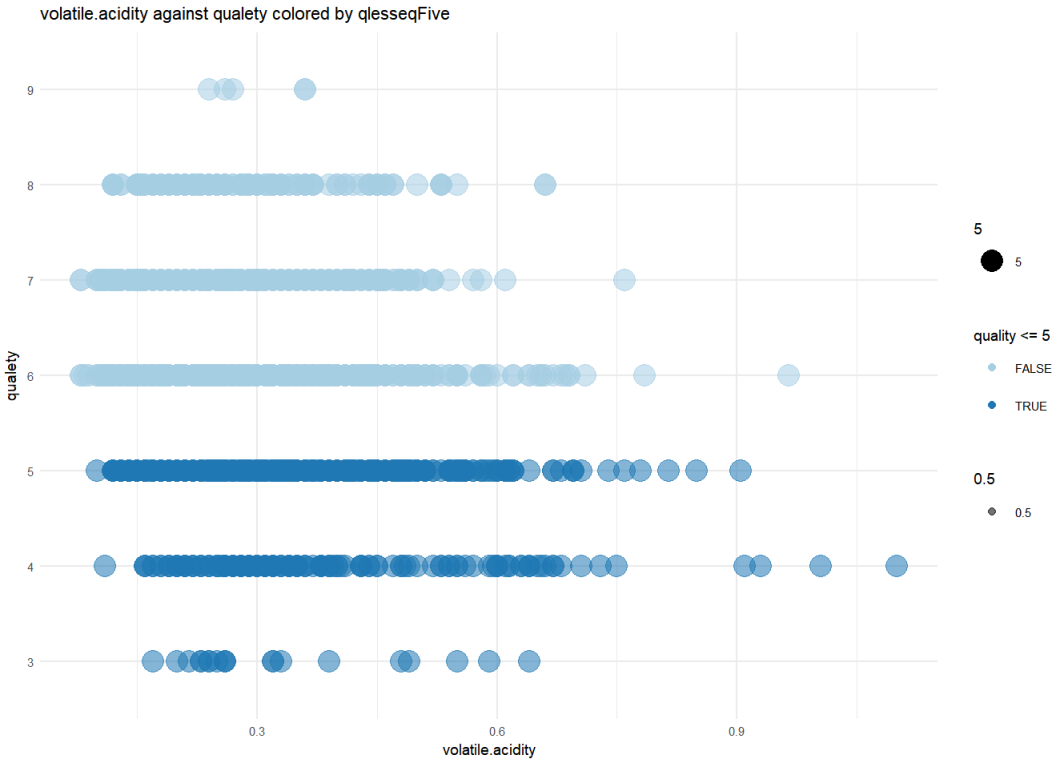

volatile.acidity against pH colored by qlesseqFive
pH
3.6
3.3
3.0
2.7
0.3
0.6
0.9
volatile.acidity
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

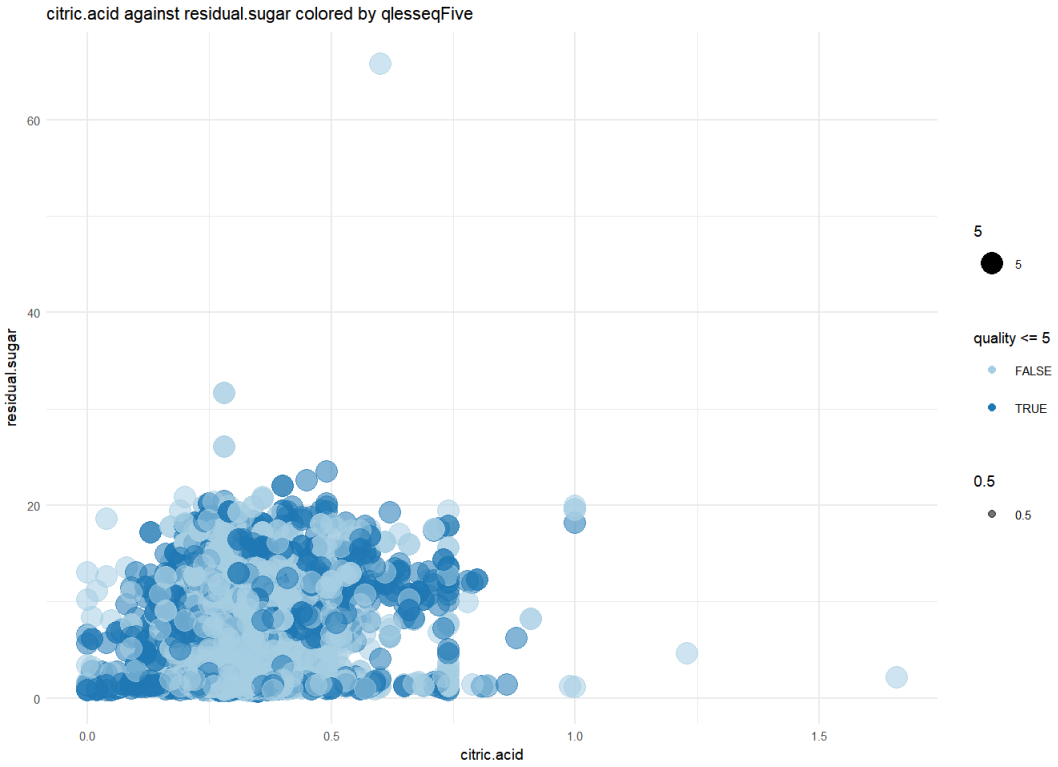
volatile.acidity against sulphates colored by qlesseqFive

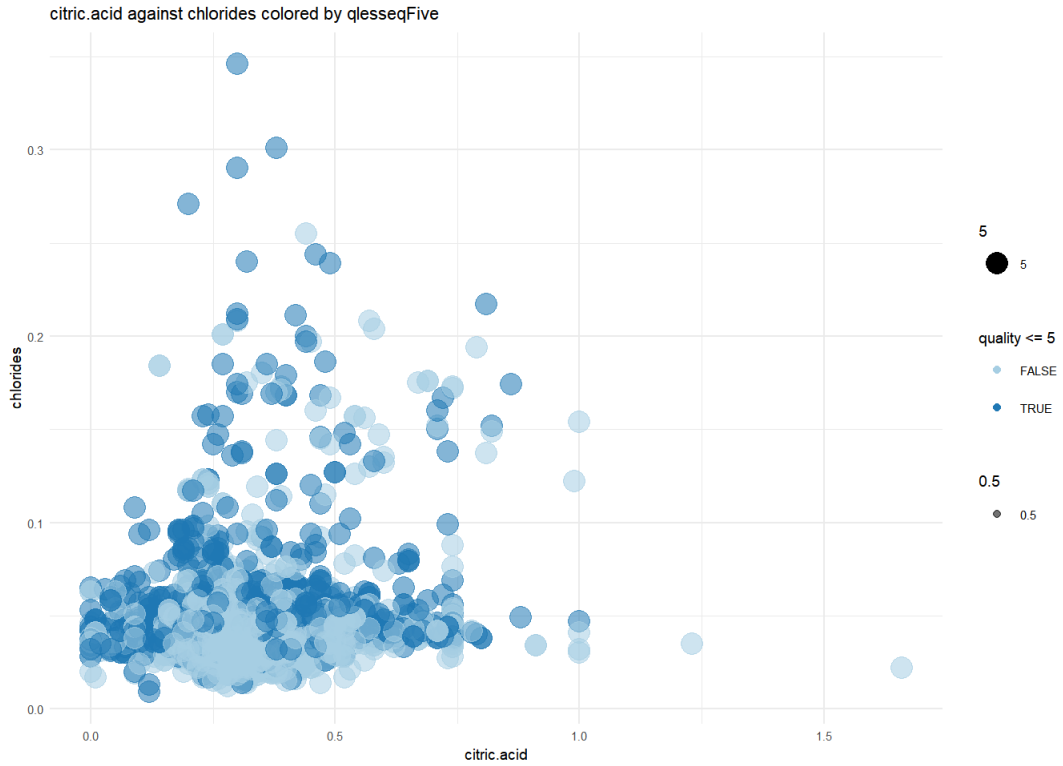


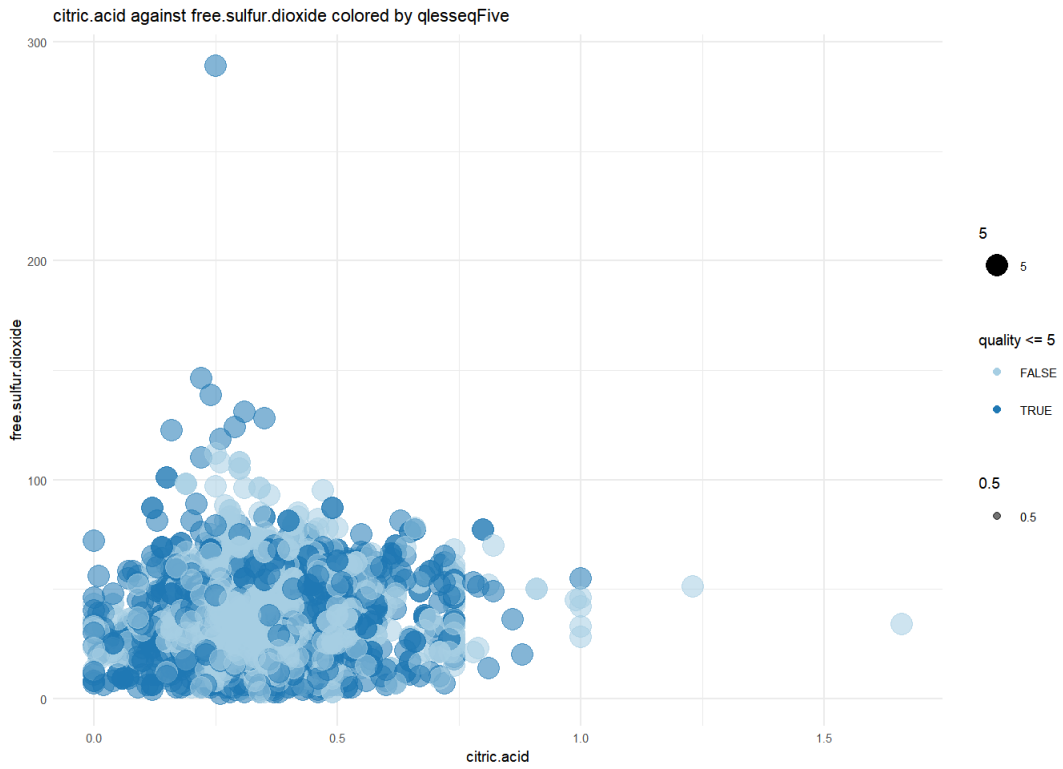
volatile.acidity against alcohol colored by qlesseqFive

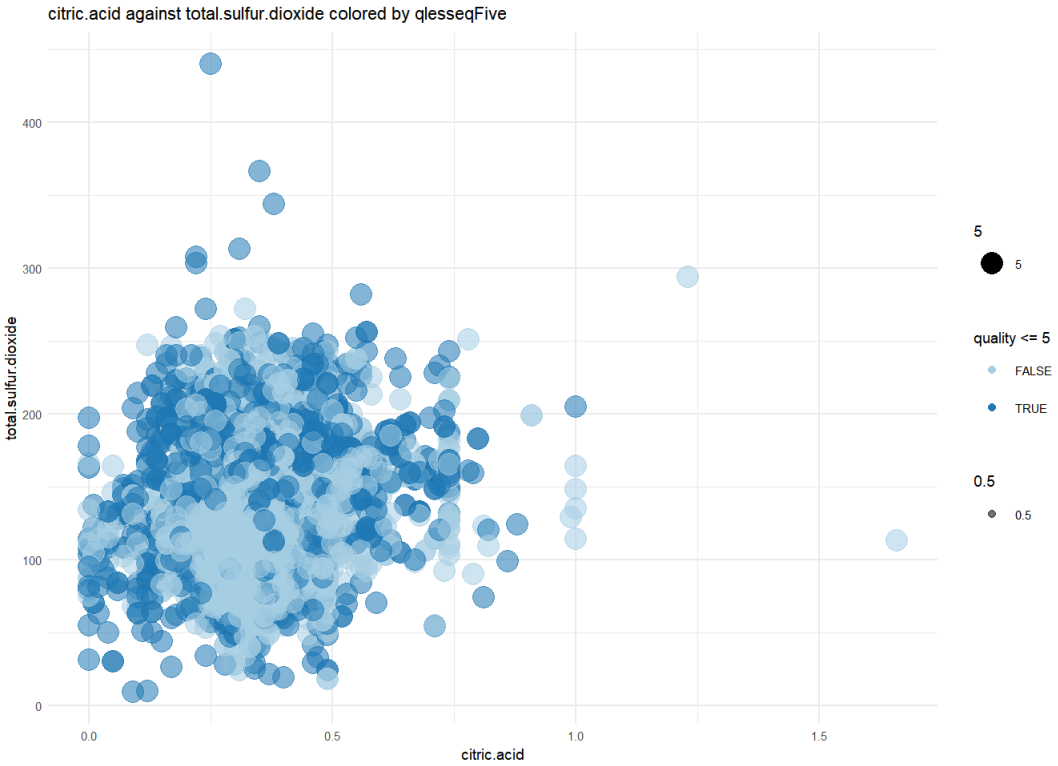


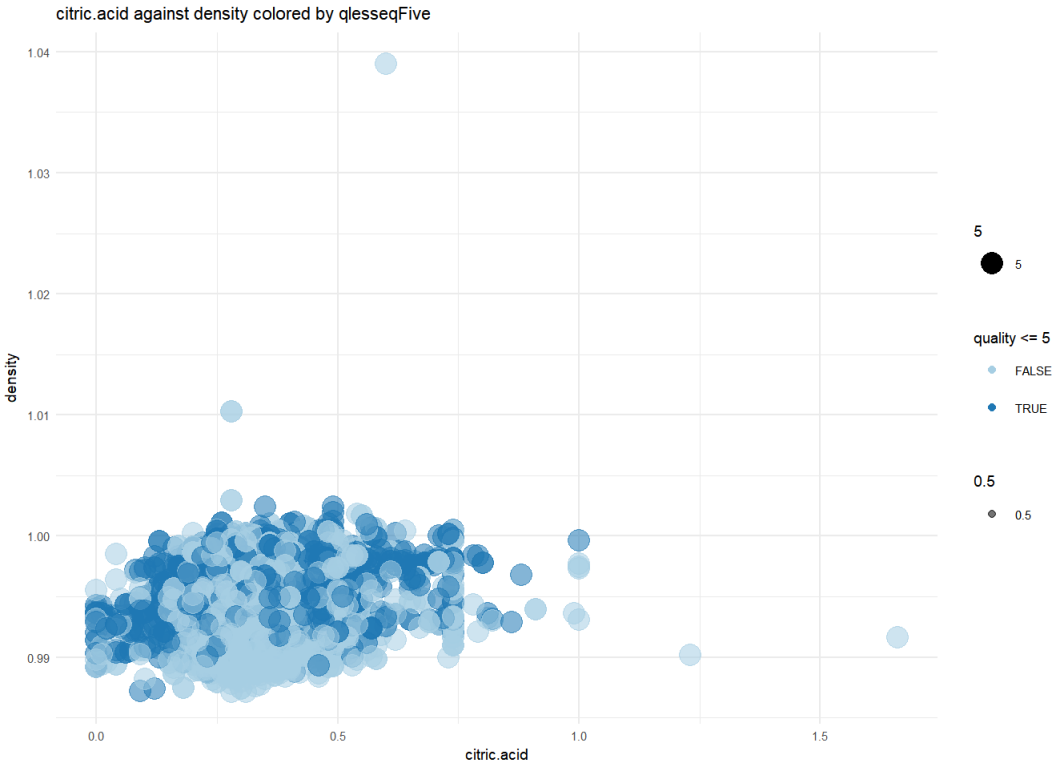

volatile.acidity against quality colored by qlesseqFive
qualety
volatile.acidity
9
8
7
6
5
4
3
0.3
0.6
0.9
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

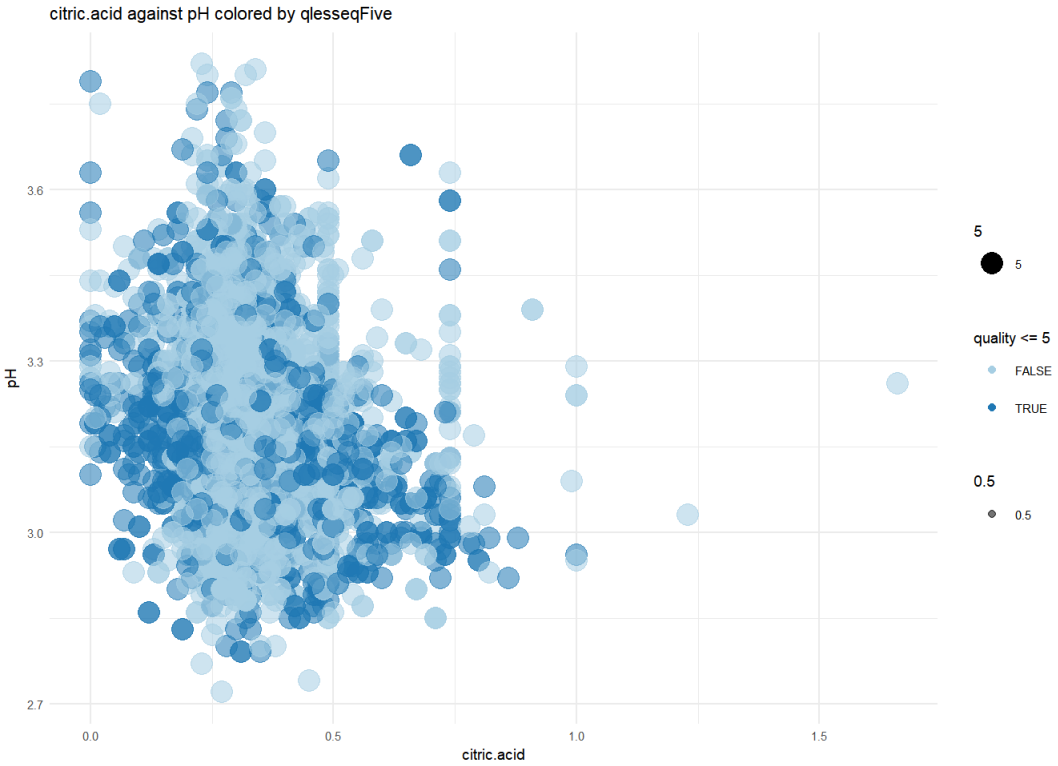

citric.acid against residual.sugar colored by qlesseqFive
residual.sugar
citric.acid
60
40
20
0
0.0
0.5
1.0
1.5
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

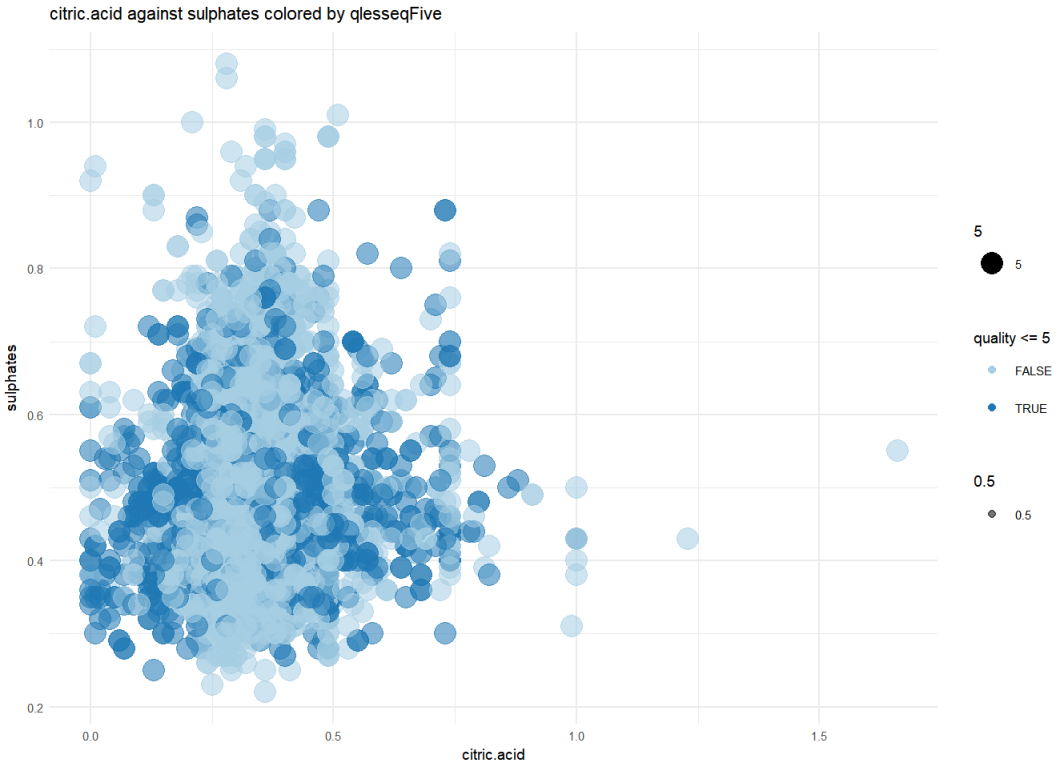

citric.acid against chlorides colored by qlesseqFive
chlorides
0.0
0.1
0.2
0.3
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

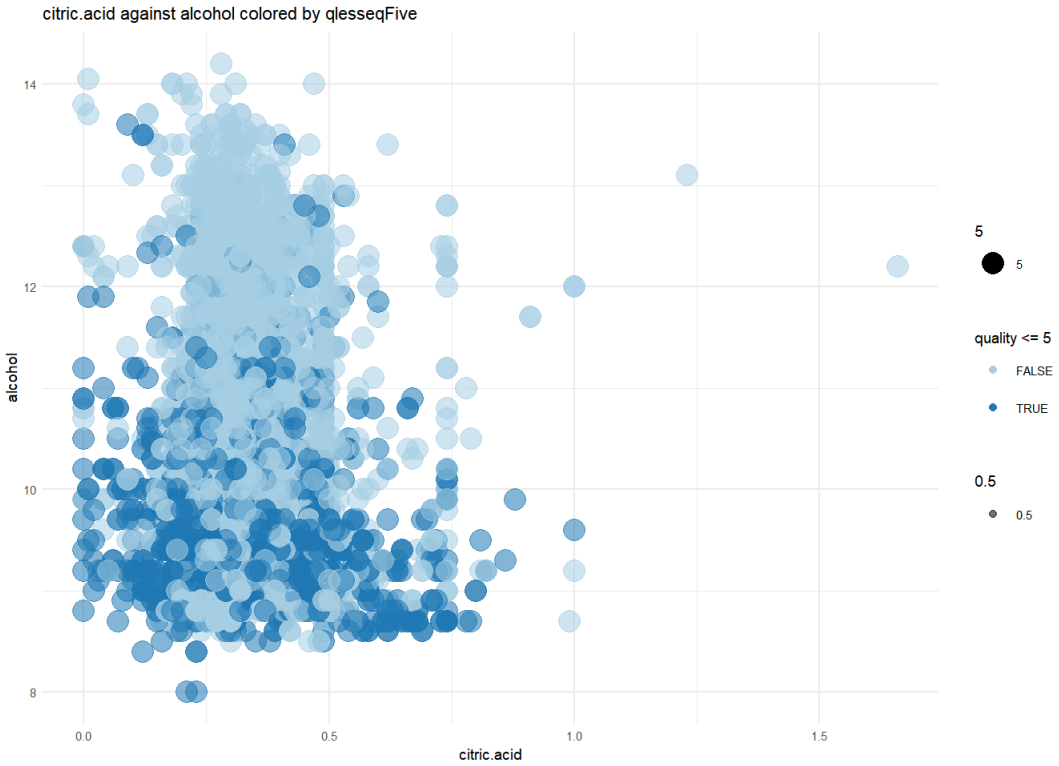

citric.acid against free.sulfur.dioxide colored by qlesseqFive
free.sulfur.dioxide
0
100
200
300
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

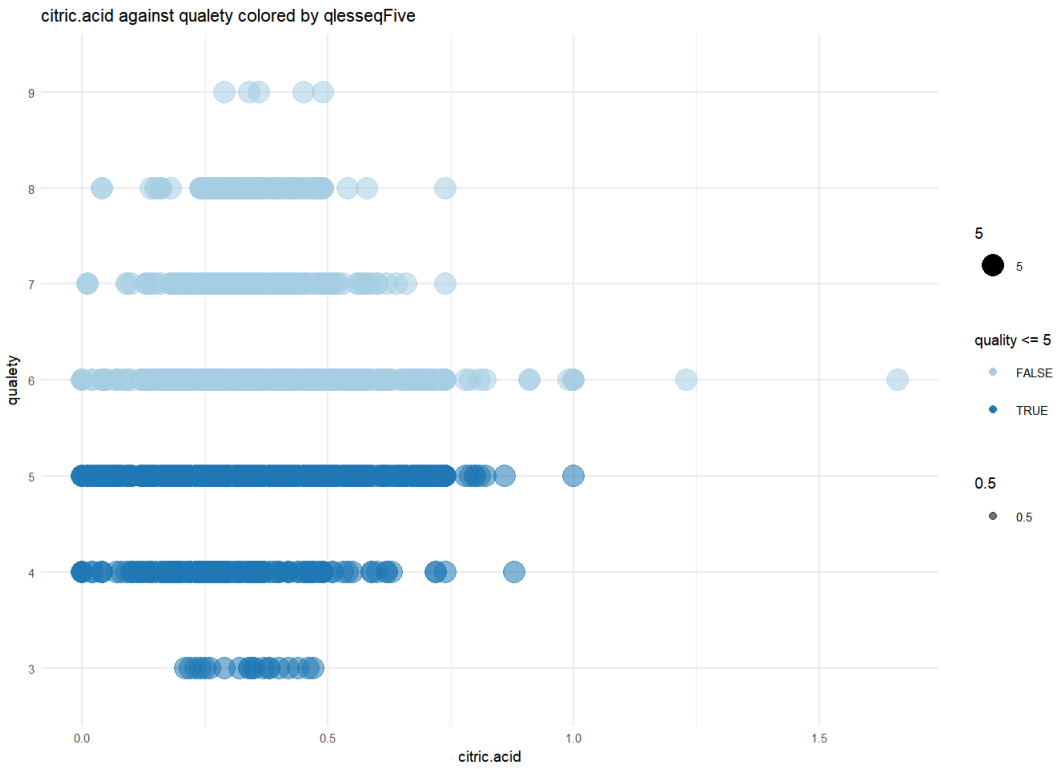

citric.acid against total.sulfur.dioxide colored by qlesseqFive
total.sulfur.dioxide
400
300
200
100
0
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

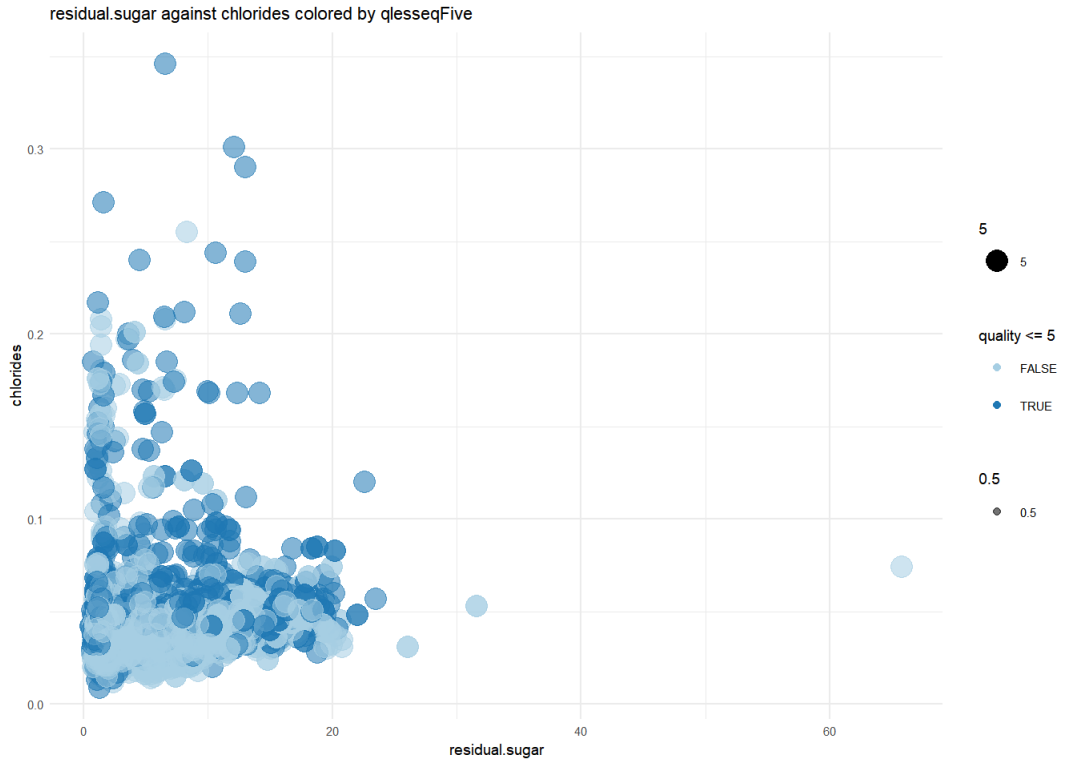

citric.acid against density colored by qlesseqFive
density
1.04
1.03
1.02
1.01
1.00
0.99
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

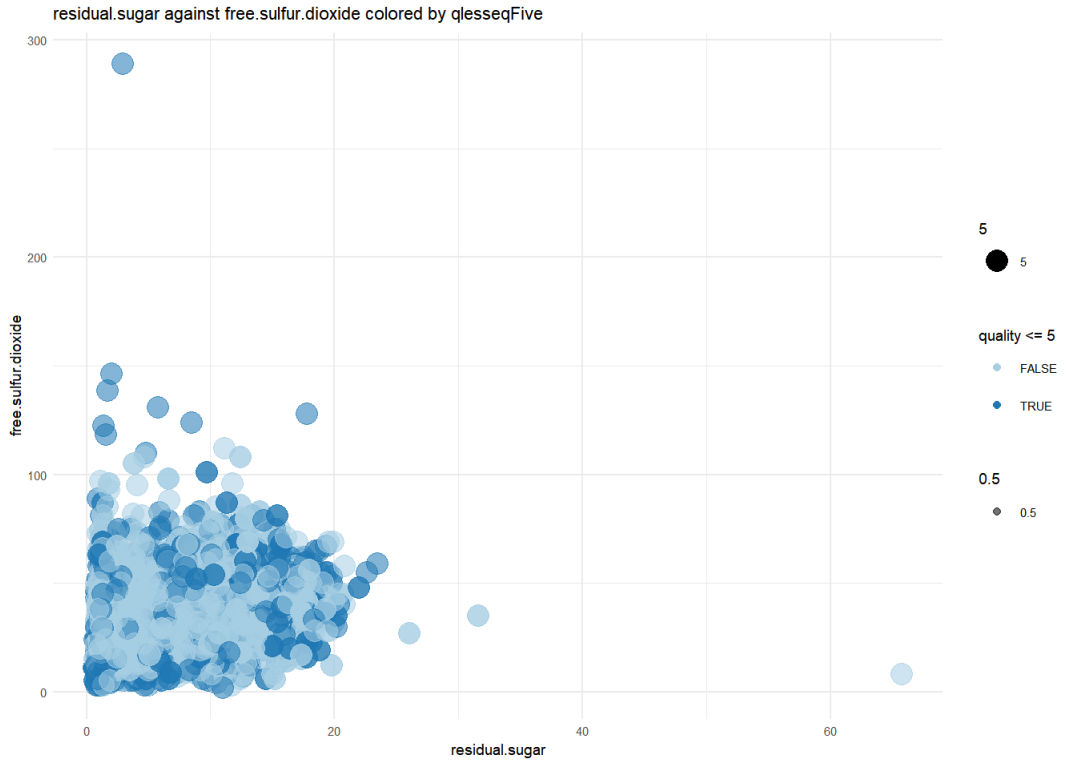

citric.acid against pH colored by qlesseqFive
pH
3.6
3.3
3.0
2.7
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

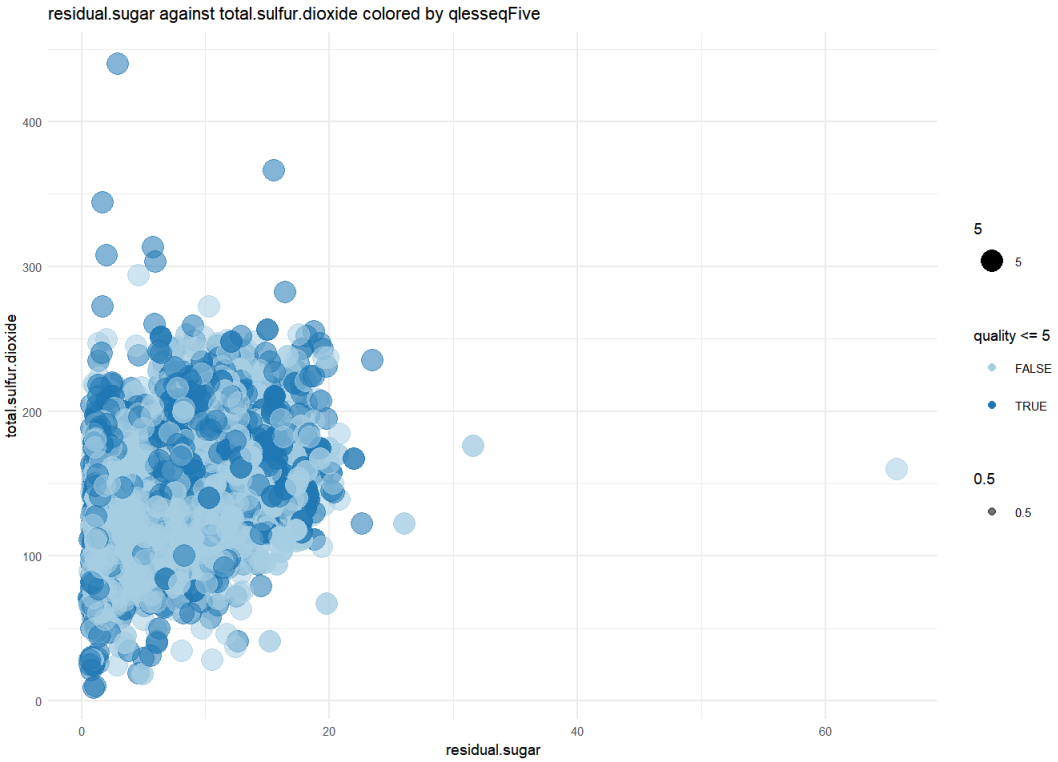

citric.acid against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

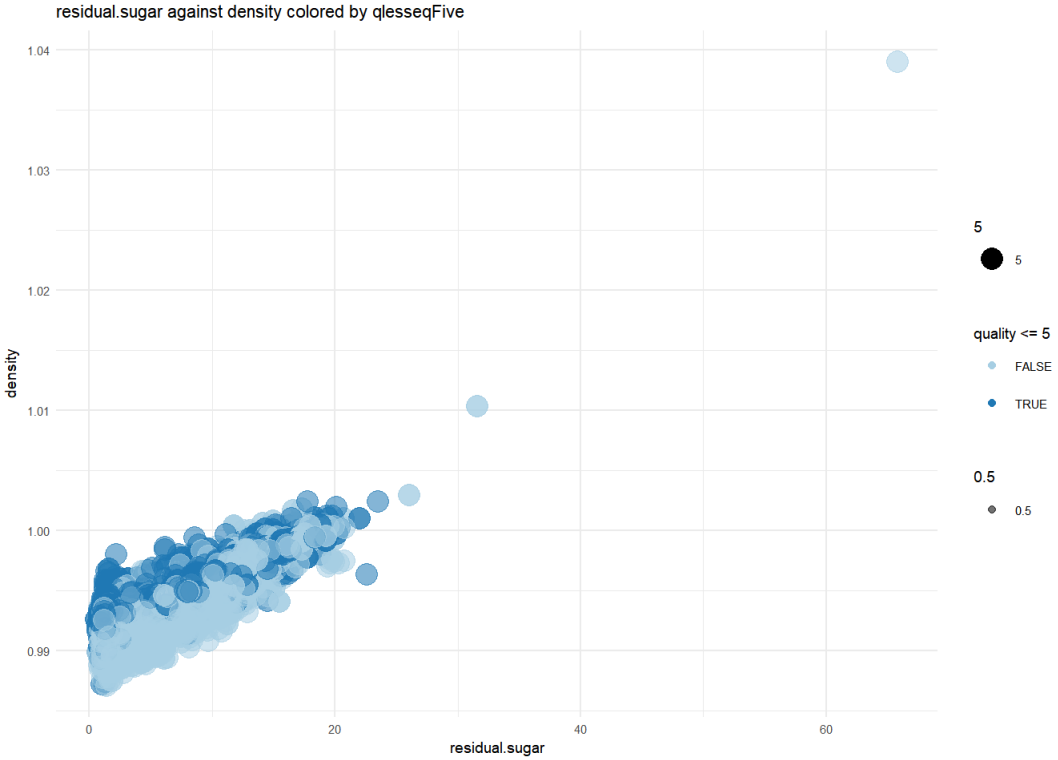

citric.acid against alcohol colored by qlesseqFive
alcohol
14
12
10
8
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

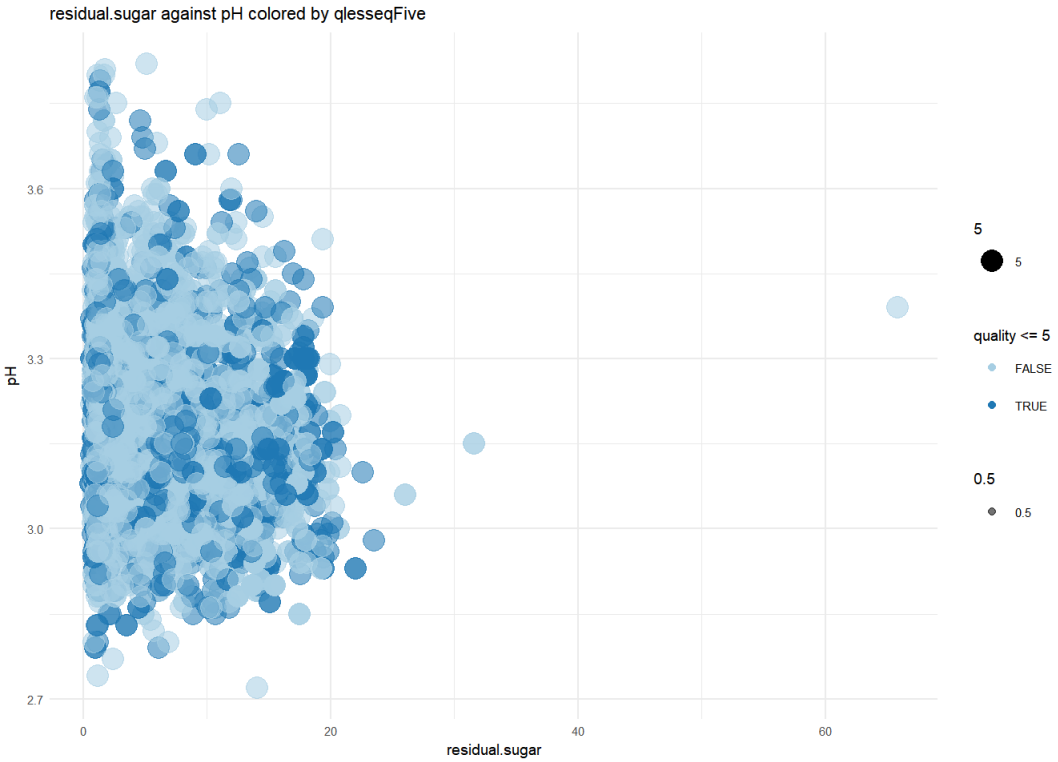

citric.acid against quality colored by qlesseqFive
qualety
9
8
7
6
5
4
3
0.0
0.5
1.0
1.5
citric.acid
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

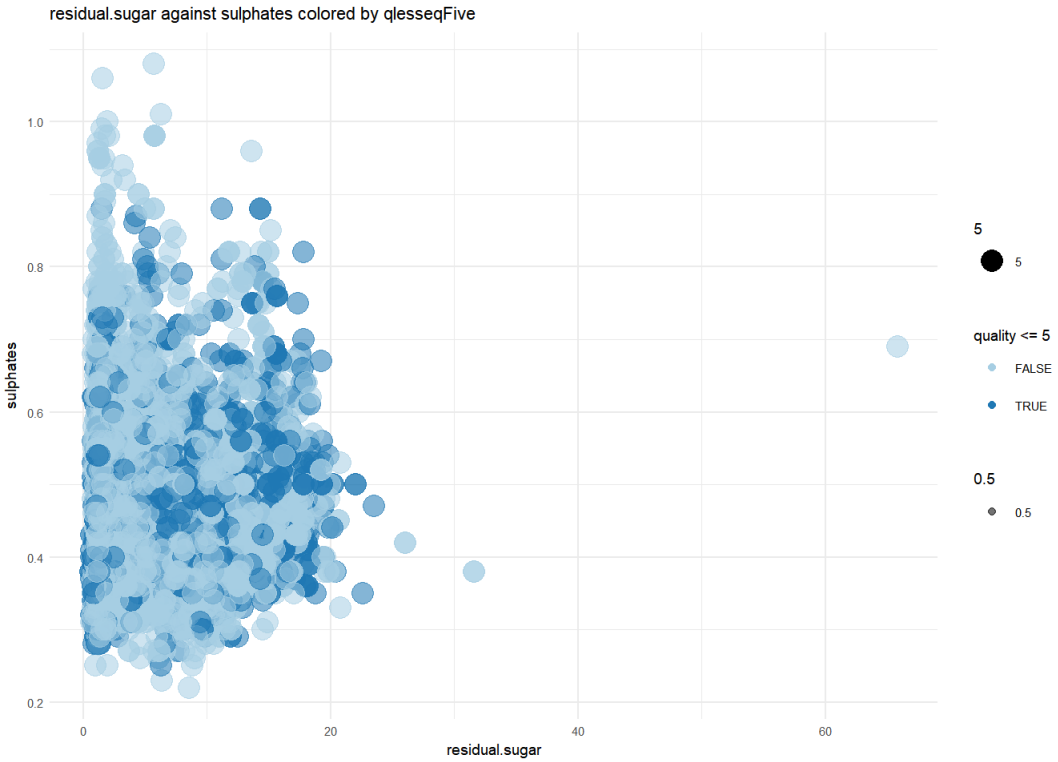

residual.sugar against chlorides colored by qlesseqFive
chlorides
0.0
0.1
0.2
0.3
0
20
40
60
residual.sugar
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

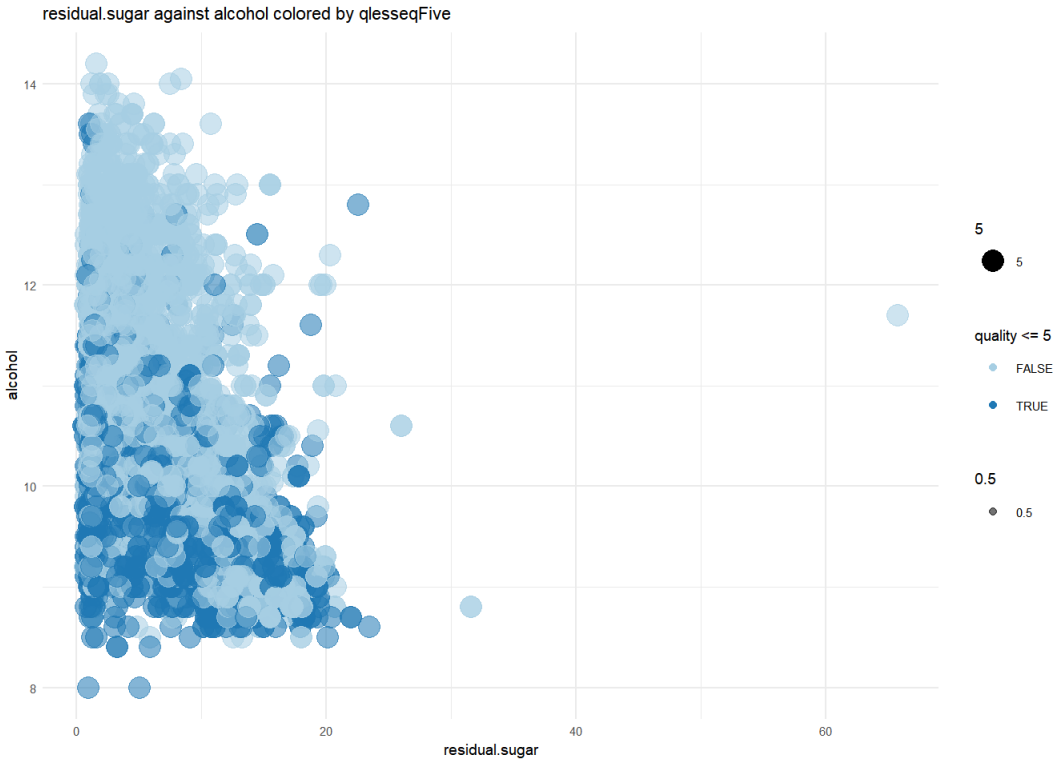

residual.sugar against free.sulfur.dioxide colored by qlesseqFive
free.sulfur.dioxide
0
100
200
300
0
20
40
60
residual.sugar
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

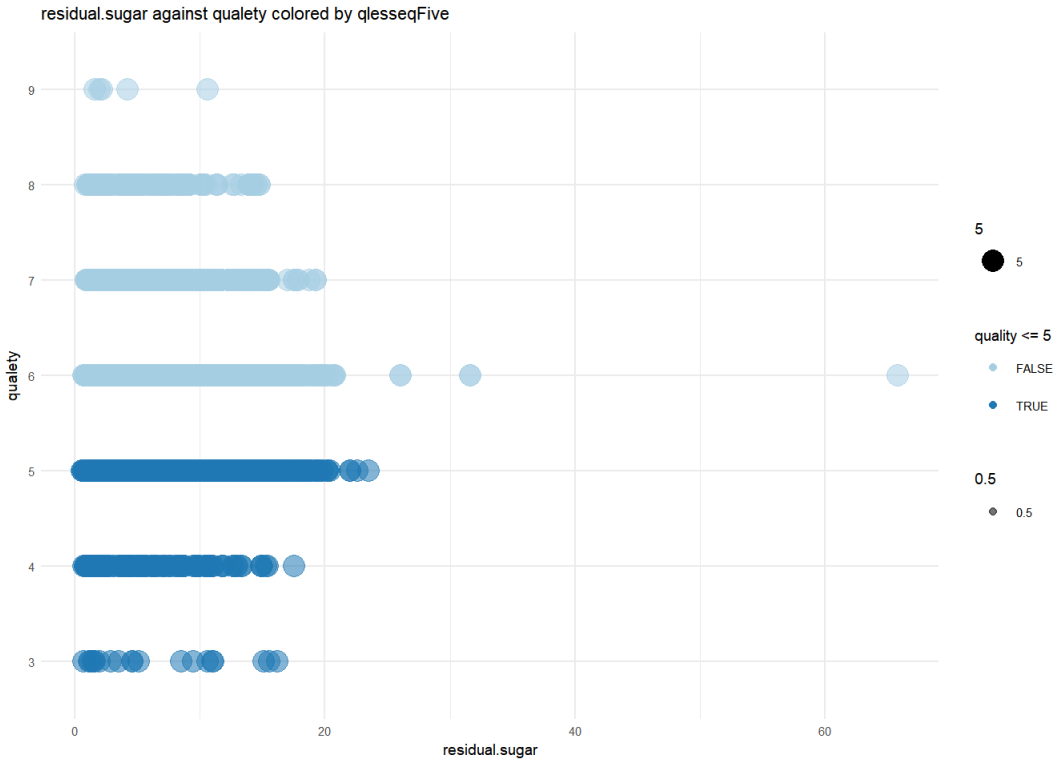

residual.sugar against total.sulfur.dioxide colored by qlesseqFive
total.sulfur.dioxide
residual.sugar
0
100
200
300
400
0
20
40
60
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


residual.sugar against density colored by qlesseqFive
density
residual.sugar
0.99
1.00
1.01
1.02
1.03
1.04
0
20
40
60
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

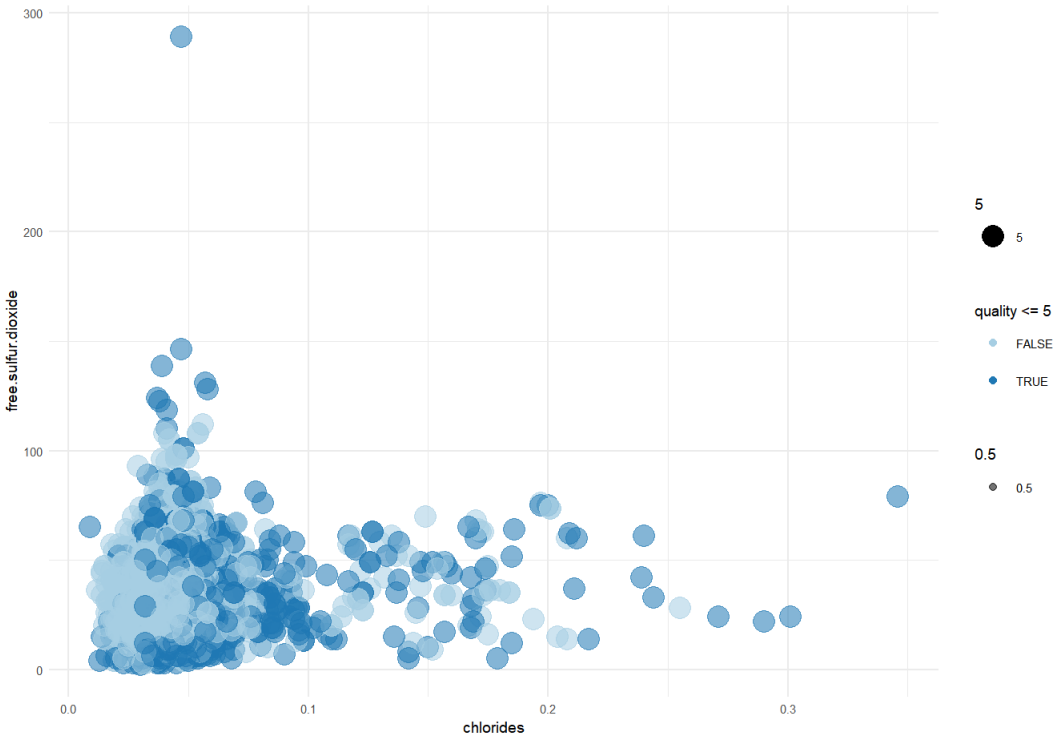

residual.sugar against pH colored by qlesseqFive
pH
3.6
3.3
3.0
2.7
0
20
40
60
residual.sugar
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


residual.sugar against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
0
20
40
60
residual.sugar
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

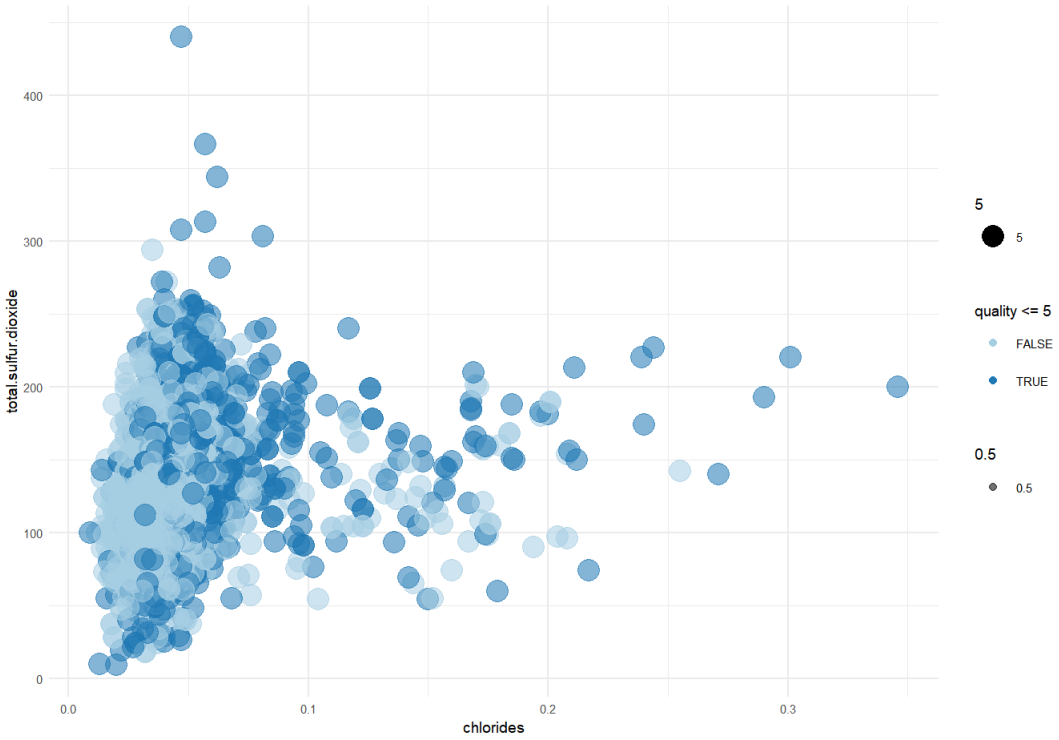

residual.sugar against alcohol colored by qlesseqFive
alcohol
residual.sugar
14
12
10
8
0
20
40
60
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

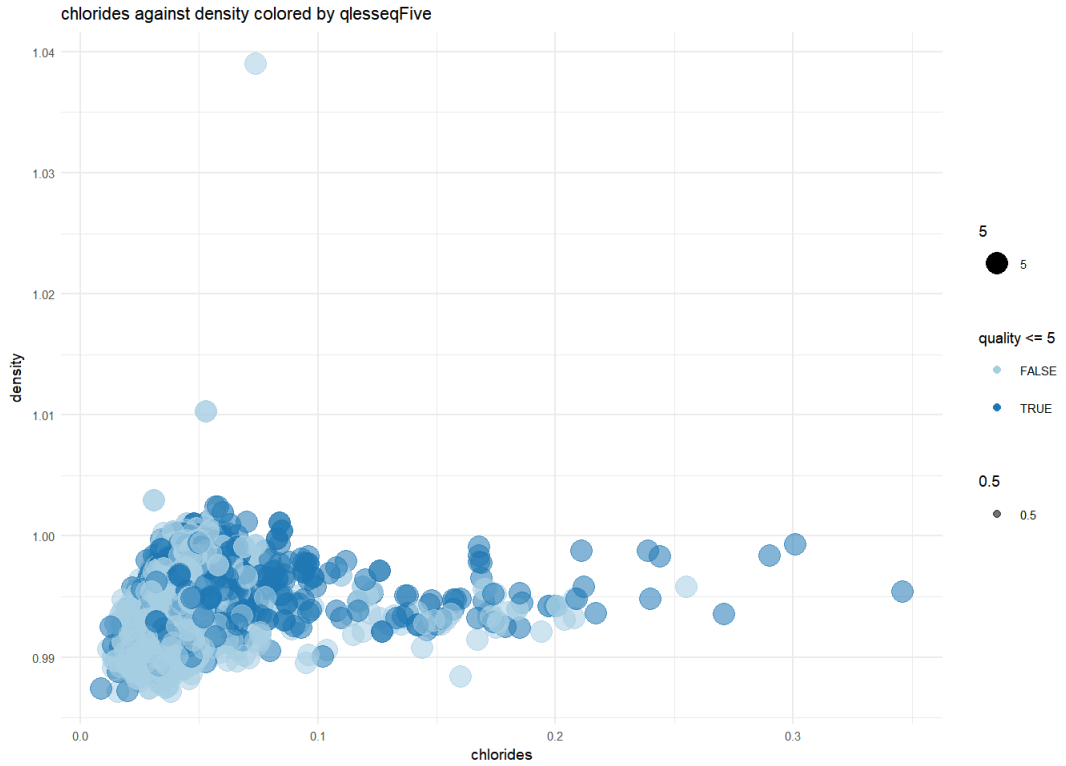

residual.sugar against quality colored by qlesseqFive
qualety
residual.sugar
9
8
7
6
5
4
3
0
20
40
60
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

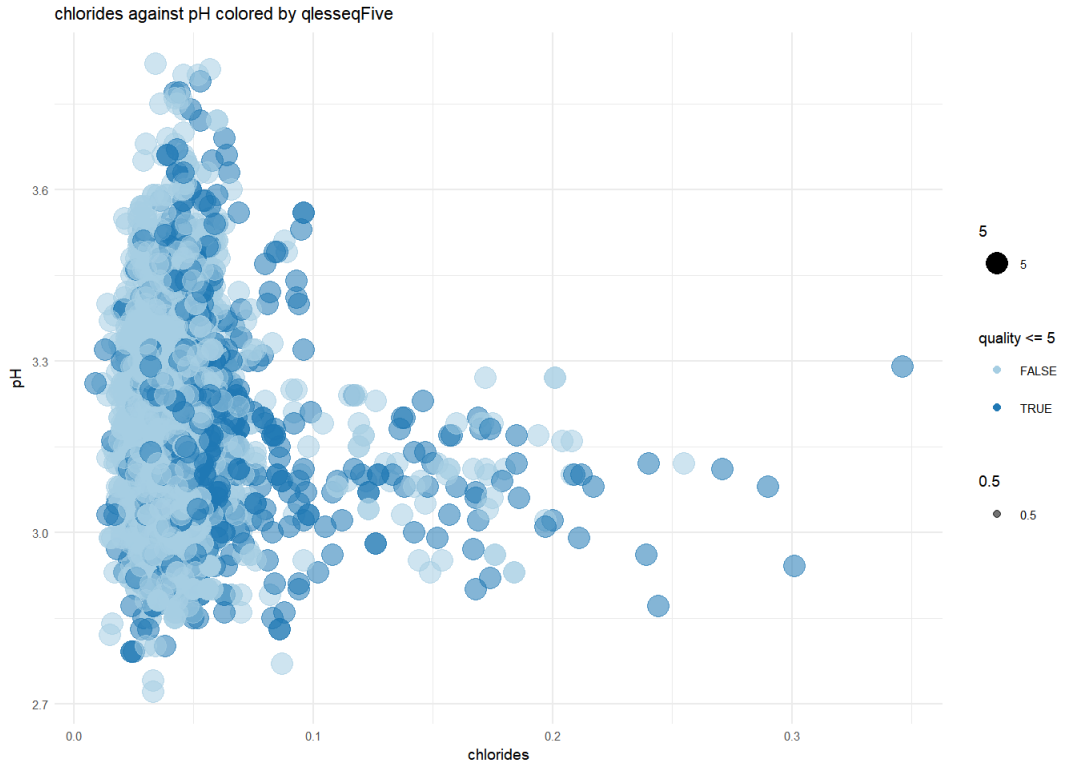
chlorides against free.sulfur.dioxide colored by qlesseqFive

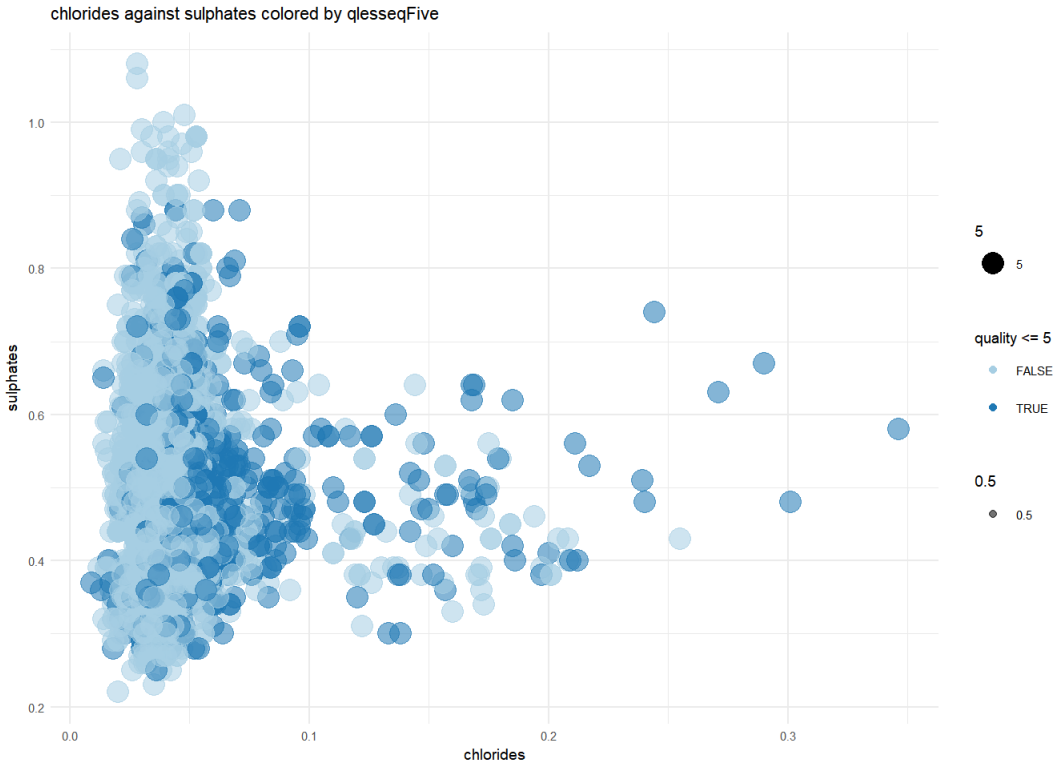


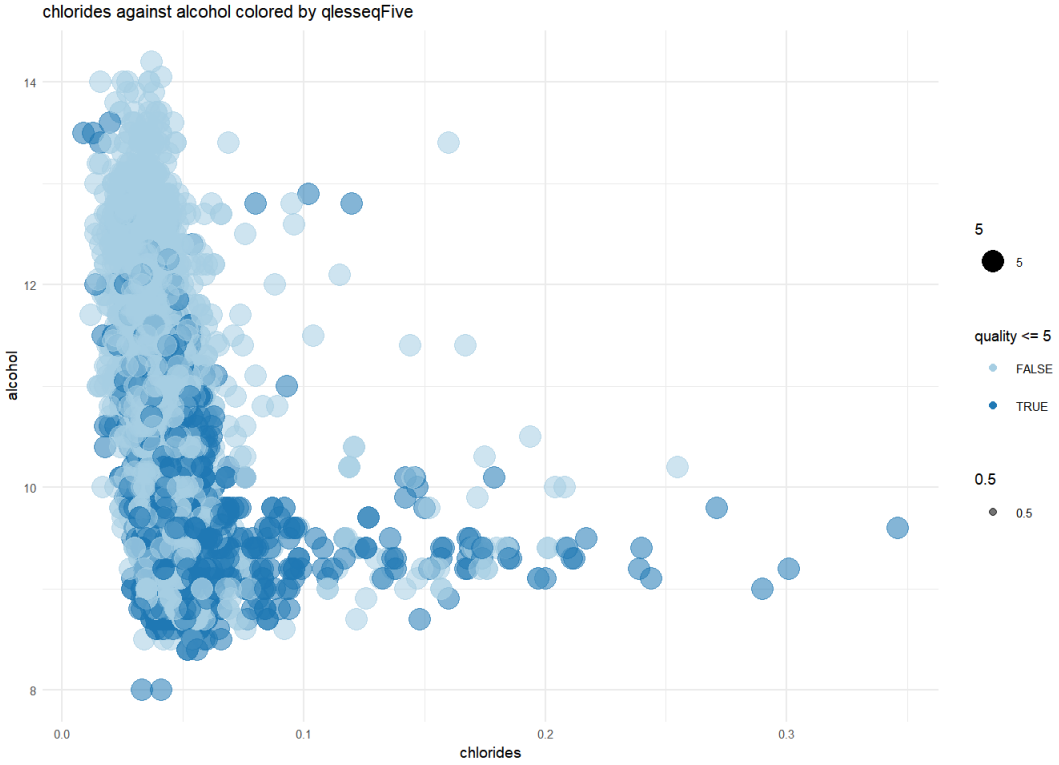
chlorides against total.sulfur.dioxide colored by qlesseqFive

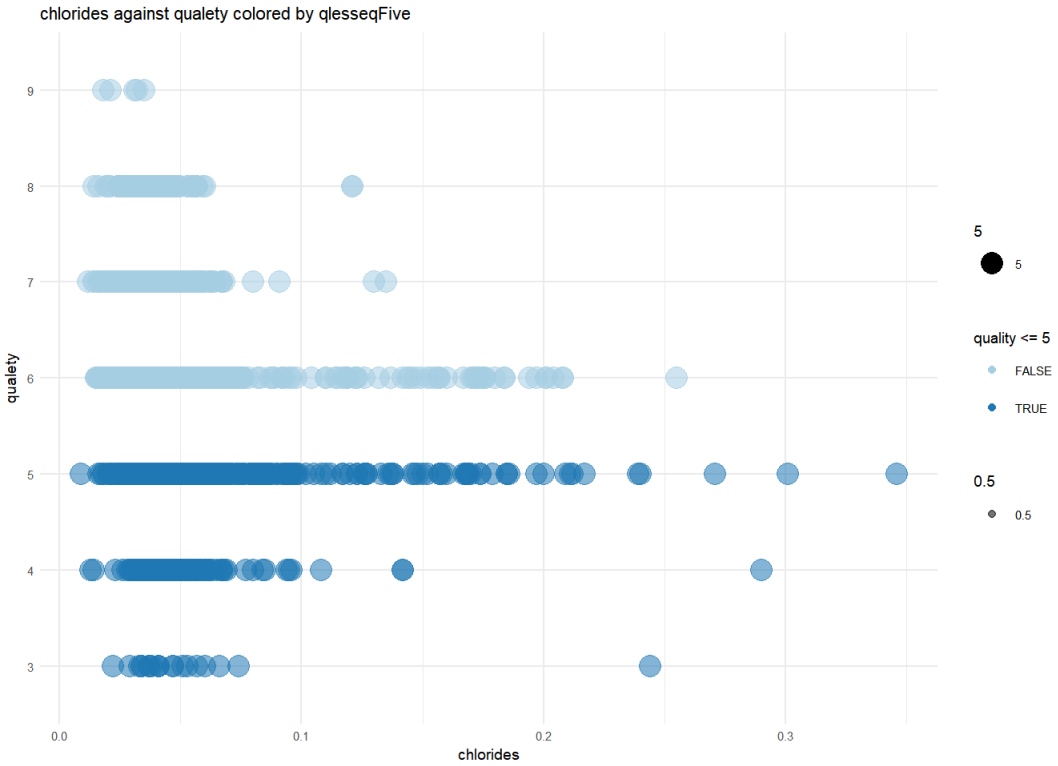


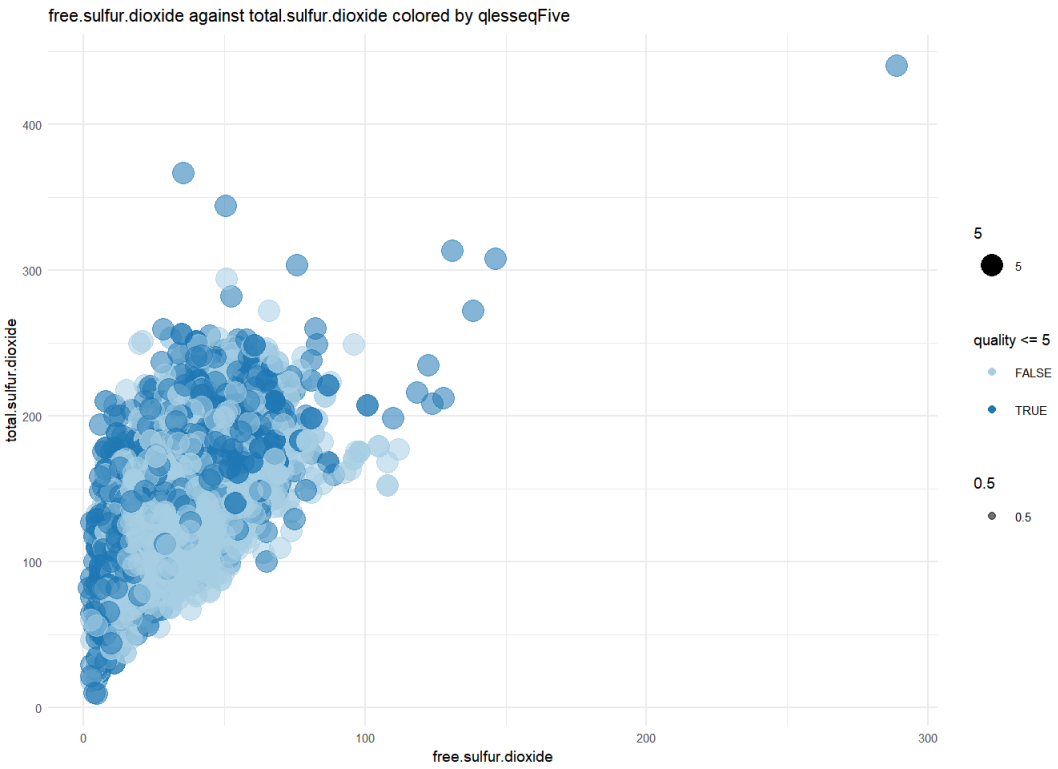

chlorides against density colored by qlesseqFive
density
1.04
1.03
1.02
1.01
1.00
0.99
0.0
0.1
0.2
0.3
chlorides
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

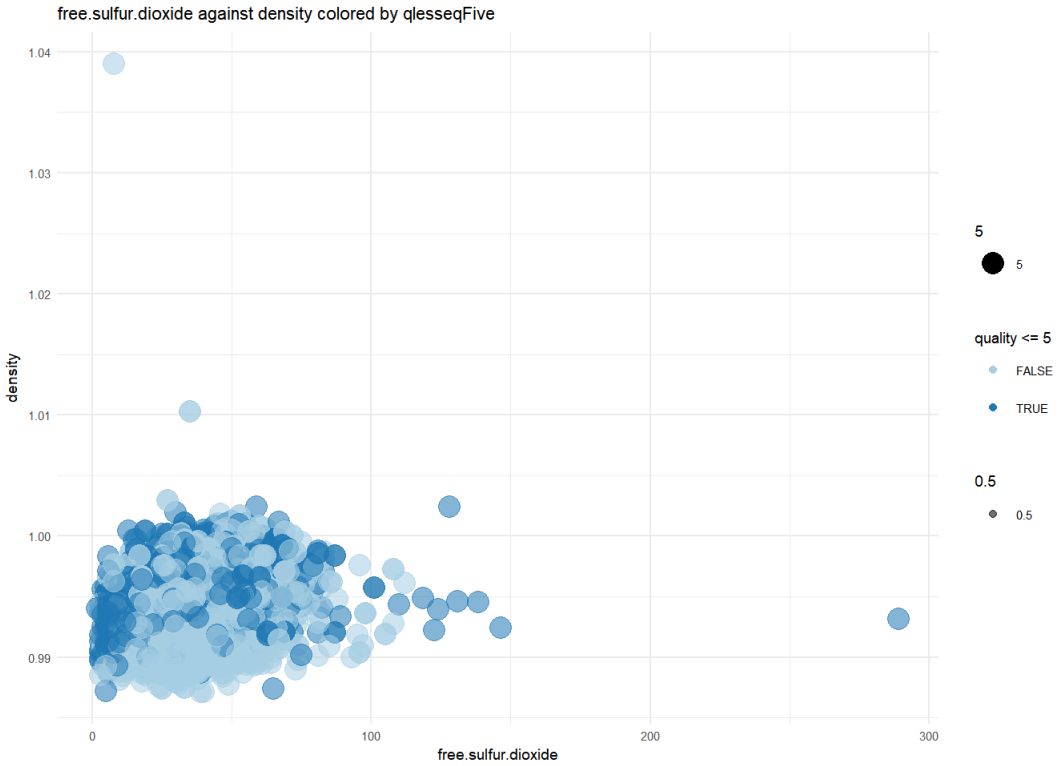

chlorides against pH colored by qlesseqFive
pH
3.6
3.3
3.0
2.7
0.0
0.1
0.2
0.3
chlorides
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

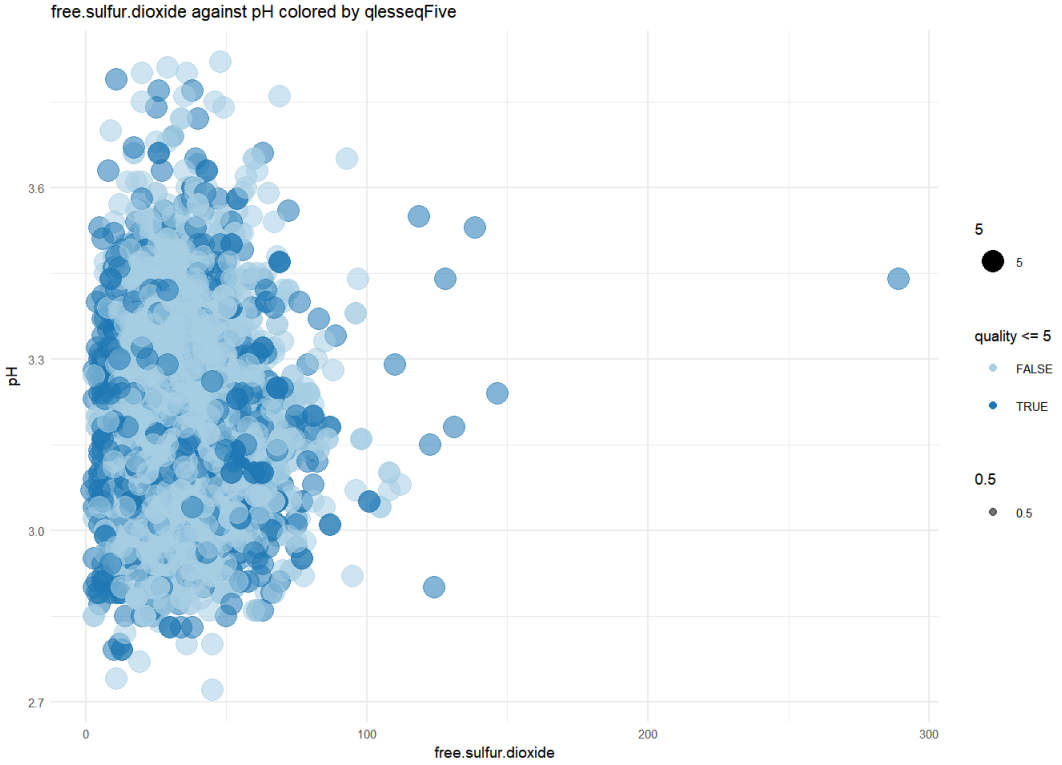

chlorides against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
chlorides
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

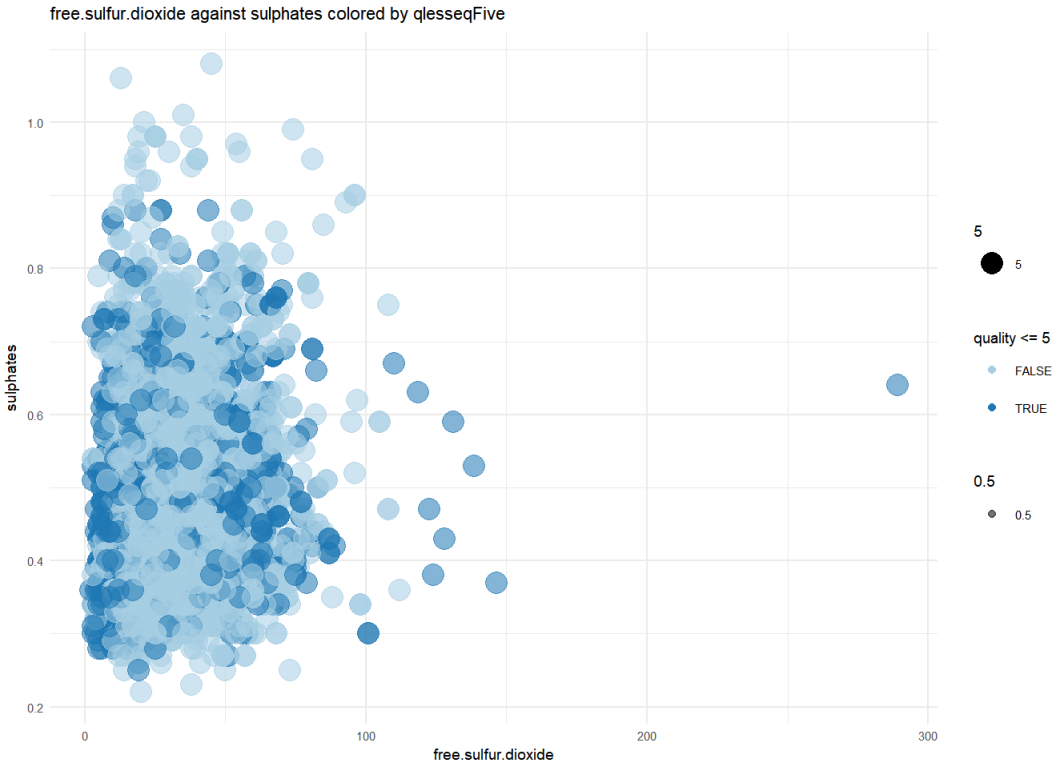

chlorides against alcohol colored by qlesseqFive
alcohol
14
12
10
8
0.0
0.1
0.2
0.3
chlorides
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

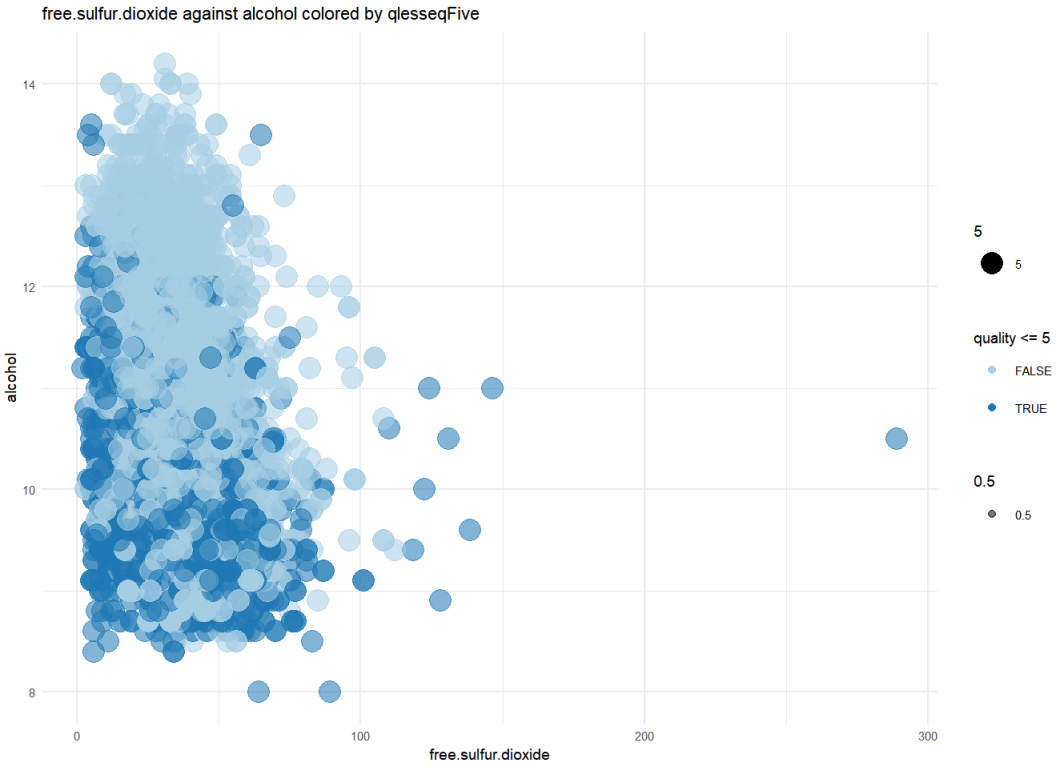

chlorides against quality colored by qlesseqFive
qualety
chlorides
9
8
7
6
5
4
3
0.0
0.1
0.2
0.3
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

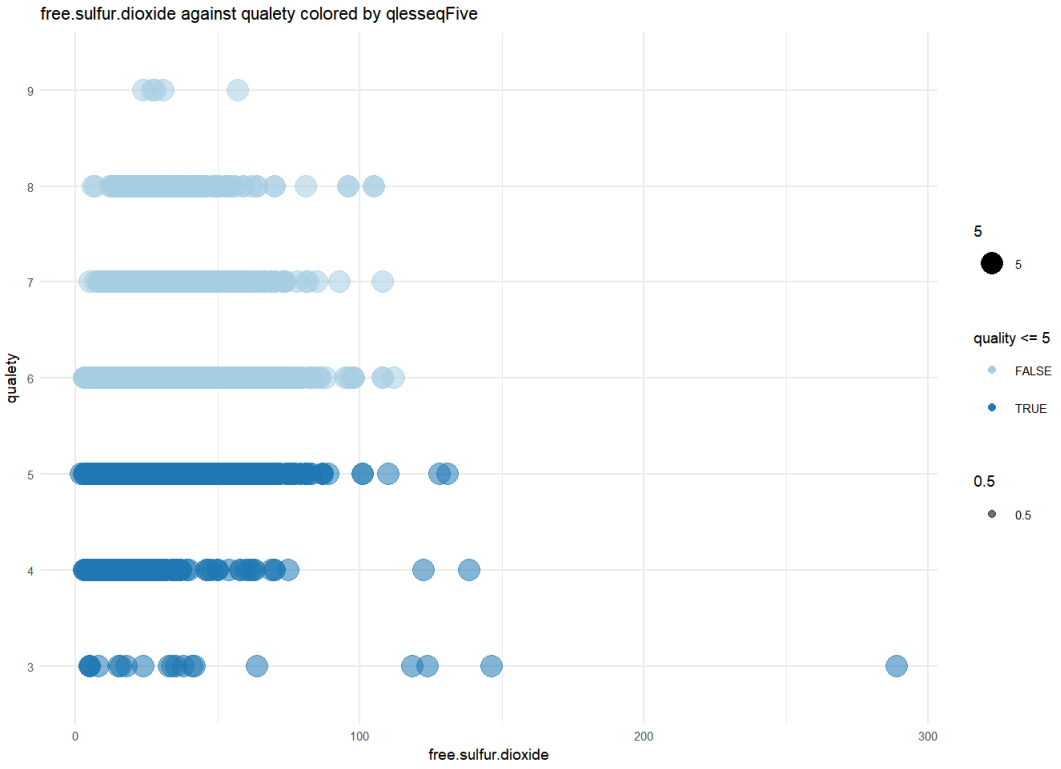

free.sulfur.dioxide against total.sulfur.dioxide colored by qlesseqFive
total.sulfur.dioxide
free.sulfur.dioxide
400
300
200
100
0
0
100
200
300
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

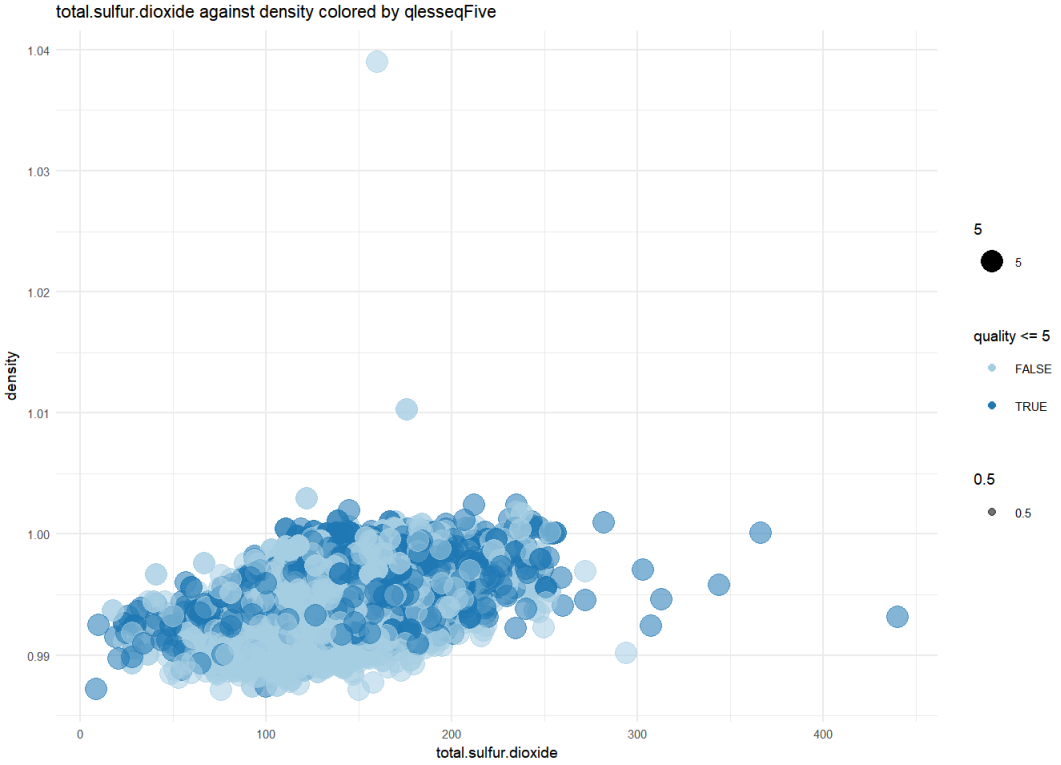

free.sulfur.dioxide against density colored by qlesseqFive
density
free.sulfur.dioxide
1.04
1.03
1.02
1.01
1.00
0.99
0
100
200
300
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

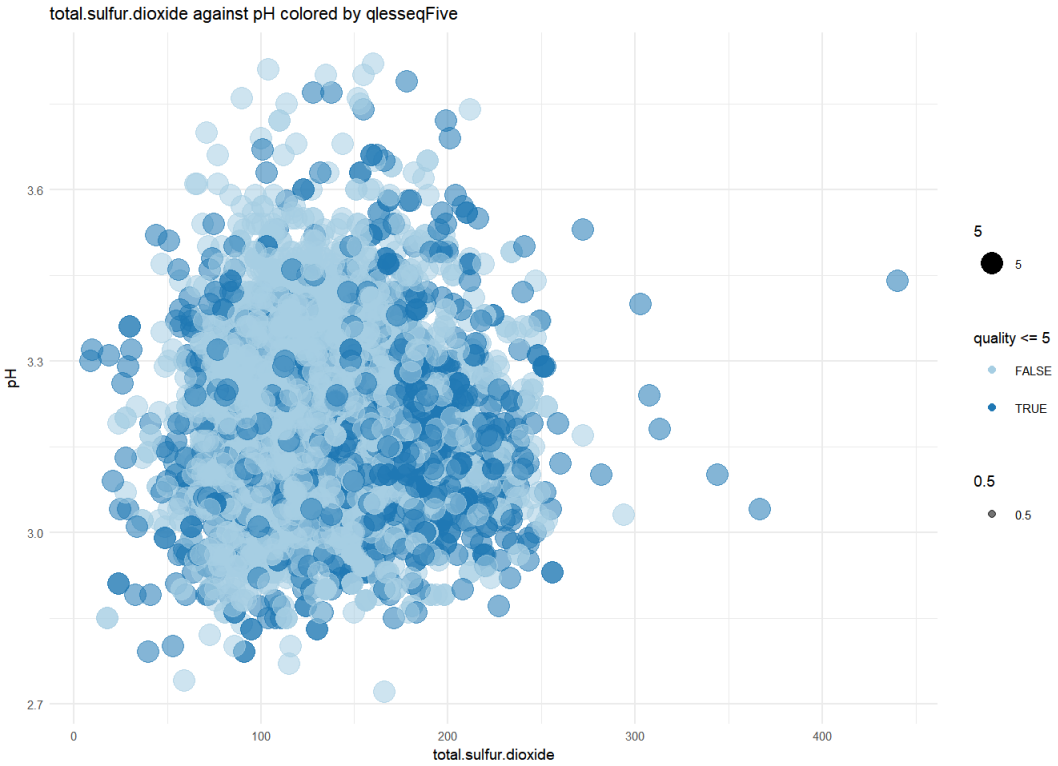

free.sulfur.dioxide against pH colored by qlesseqFive
pH
free.sulfur.dioxide
3.6
3.3
3.0
2.7
0
100
200
300
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

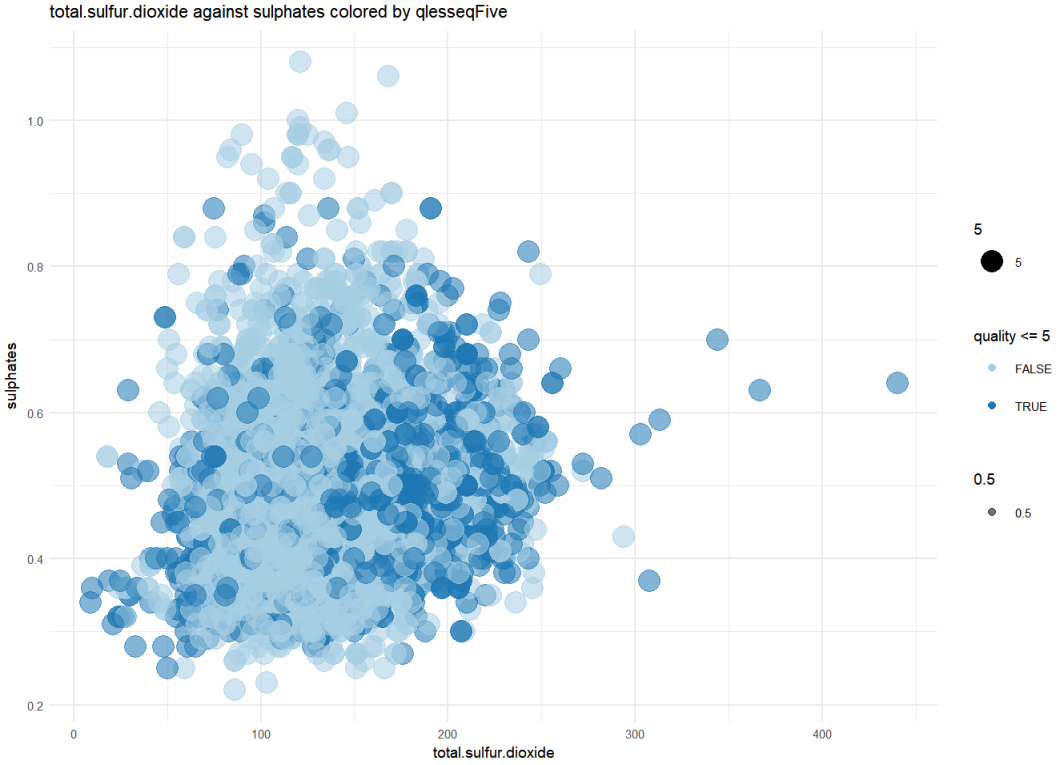

free.sulfur.dioxide against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
0
100
200
300
free.sulfur.dioxide
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

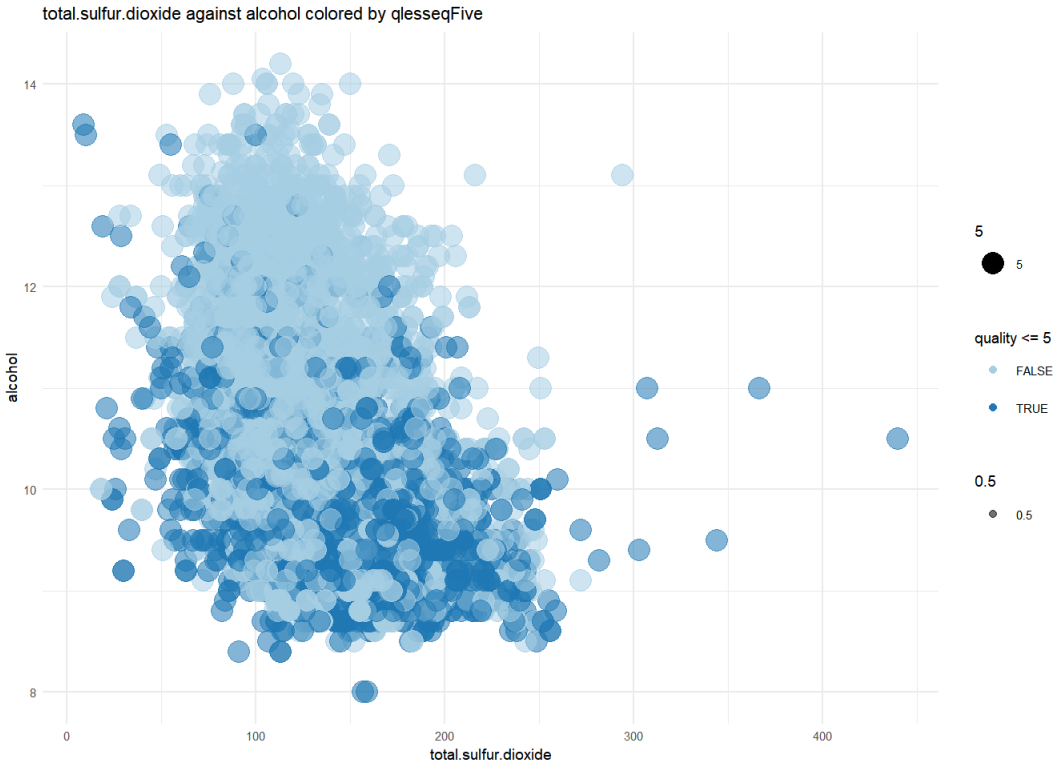

free.sulfur.dioxide against alcohol colored by qlesseqFive
alcohol
14
12
10
8
0
100
200
300
free.sulfur.dioxide
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

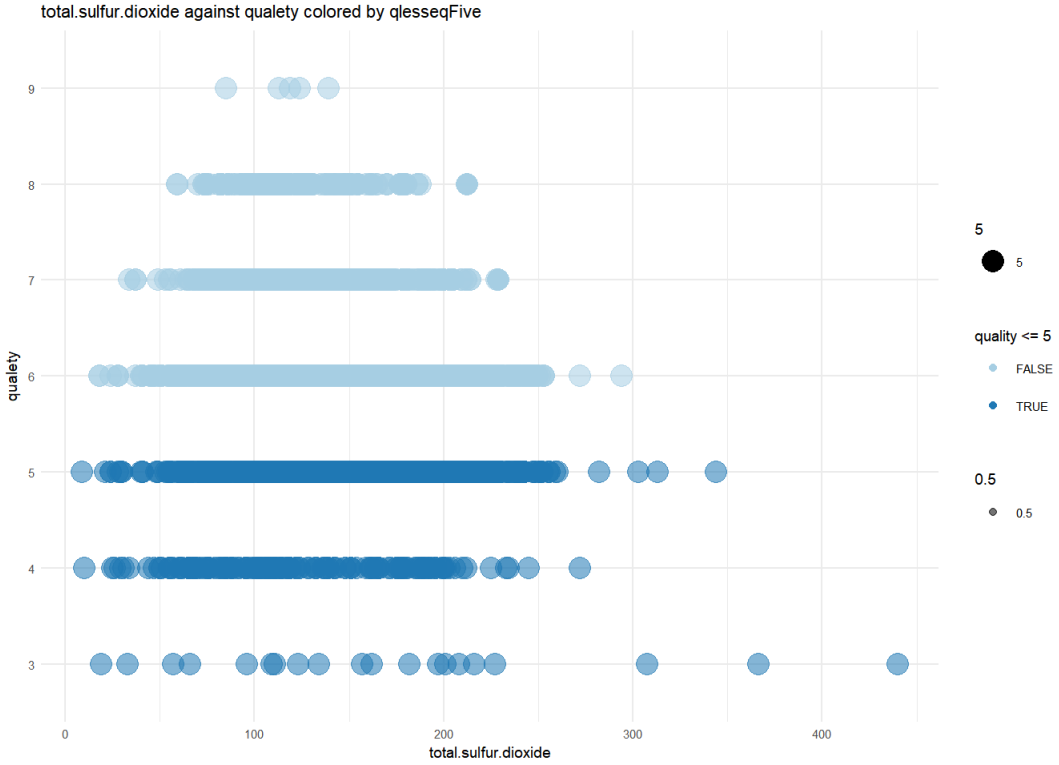

free.sulfur.dioxide against qualety colored by qlesseqFive
qualety
free.sulfur.dioxide
0
100
200
300
3
4
5
6
7
8
9
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

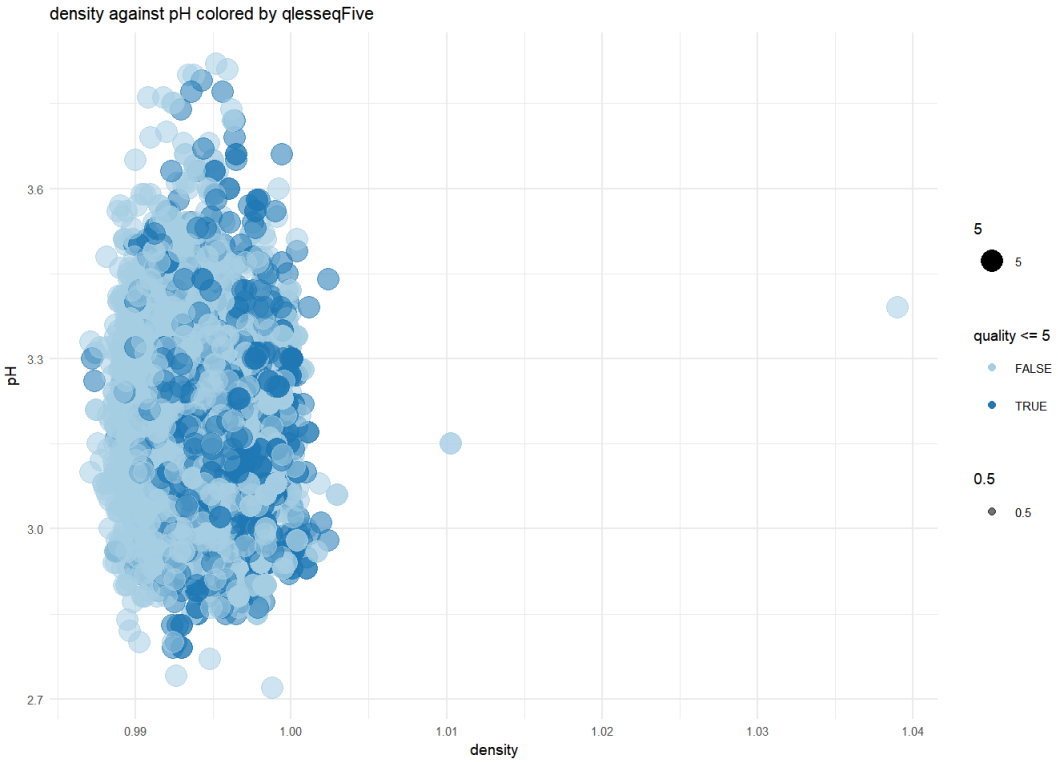

total.sulfur.dioxide against density colored by qlesseqFive
density
total.sulfur.dioxide
0
100
200
300
400
0.99
1.00
1.01
1.02
1.03
1.04
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

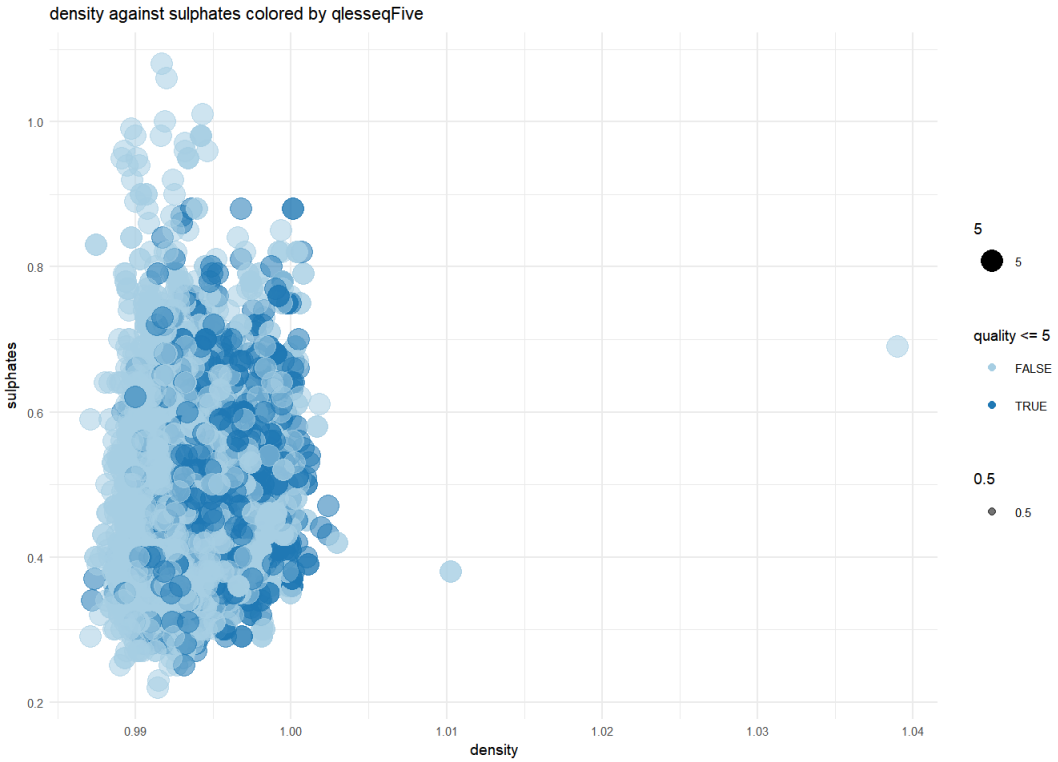

total.sulfur.dioxide against pH colored by qlesseqFive
pH
3.6
3.3
3.0
2.7
0
100
200
300
400
total.sulfur.dioxide
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

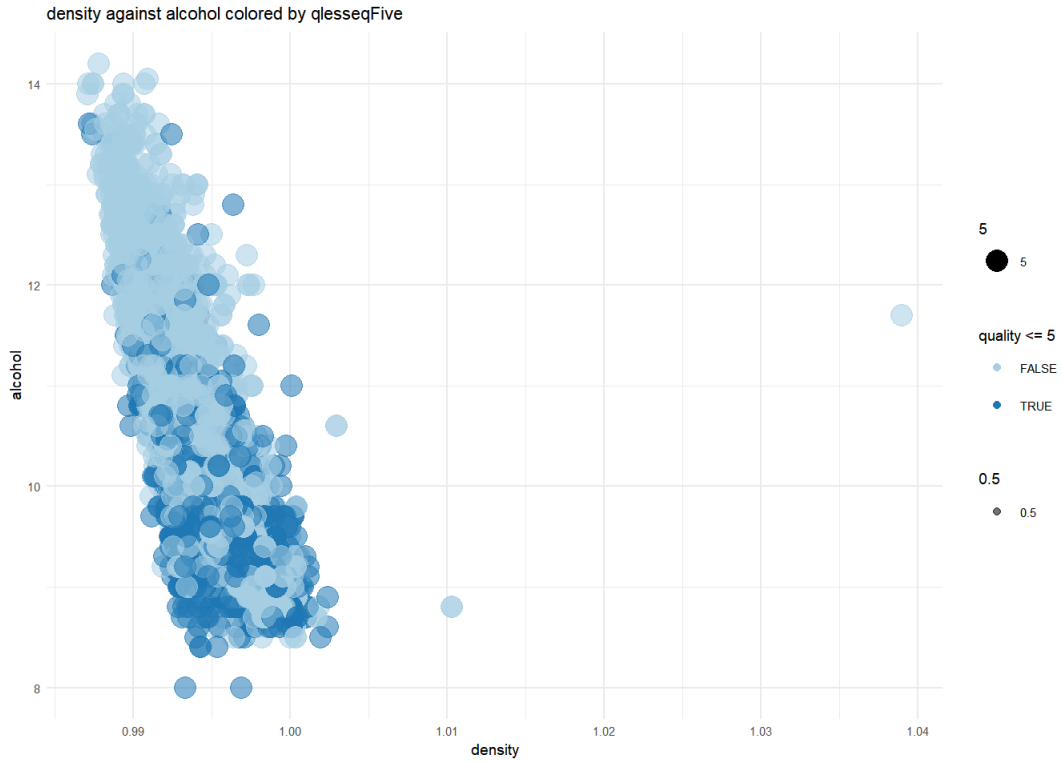

total.sulfur.dioxide against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
0
100
200
300
400
total.sulfur.dioxide
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

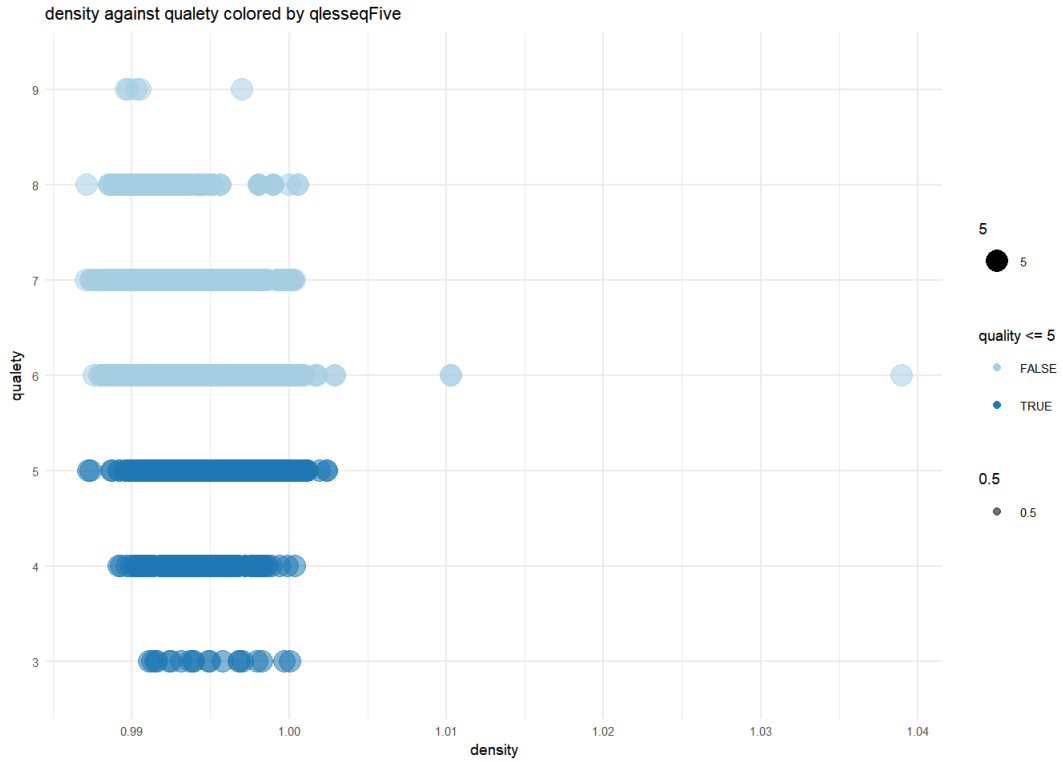

total.sulfur.dioxide against alcohol colored by qlesseqFive
alcohol
total.sulfur.dioxide
14
12
10
8
0
100
200
300
400
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

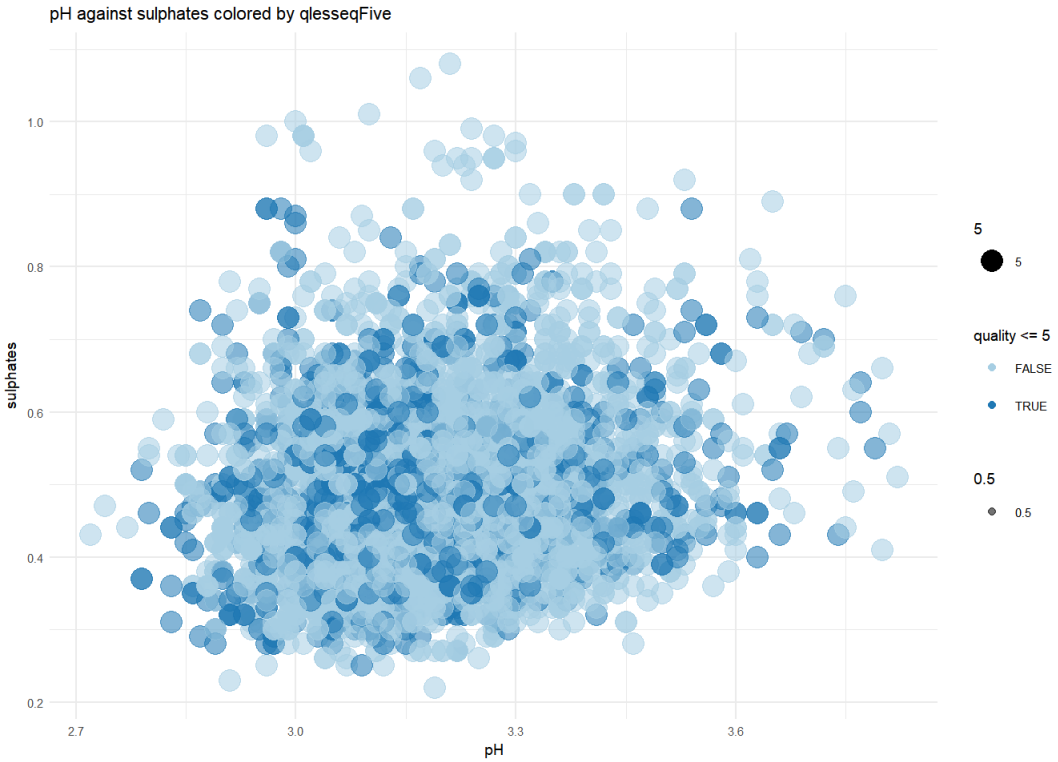

total.sulfur.dioxide against quality colored by qlesseqFive
qualety
total.sulfur.dioxide
9
8
7
6
5
4
3
0
100
200
300
400
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

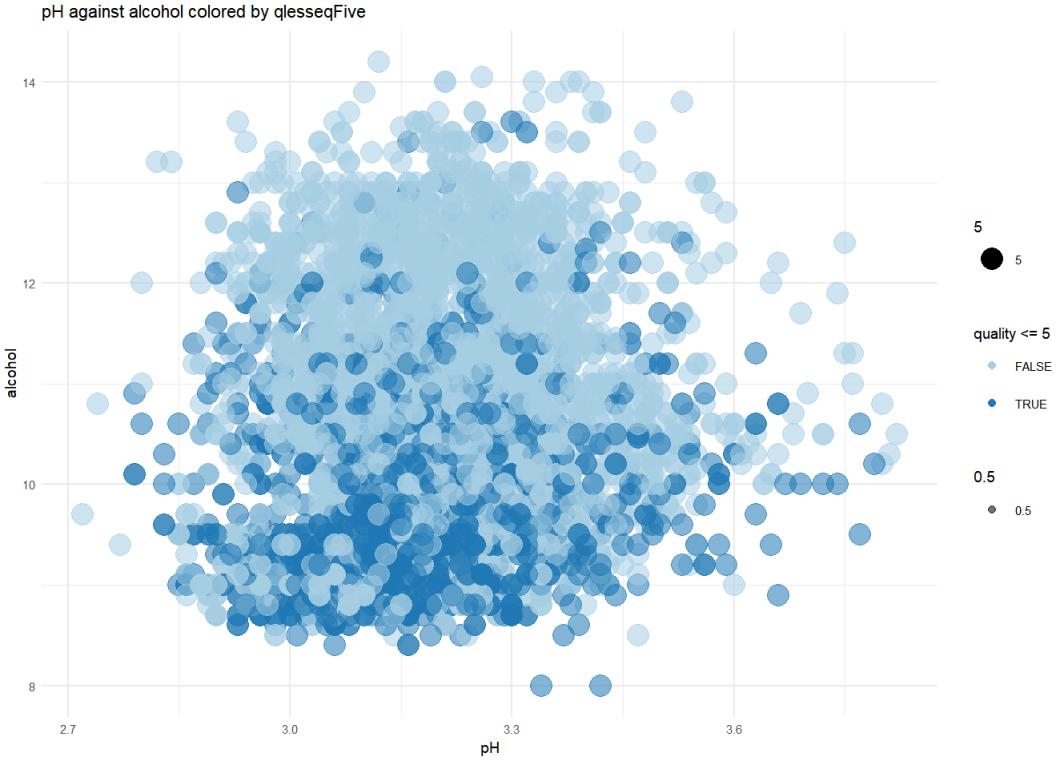

density against pH colored by qlesseqFive
pH
density
3.6
3.3
3.0
2.7
0.99
1.00
1.01
1.02
1.03
1.04
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

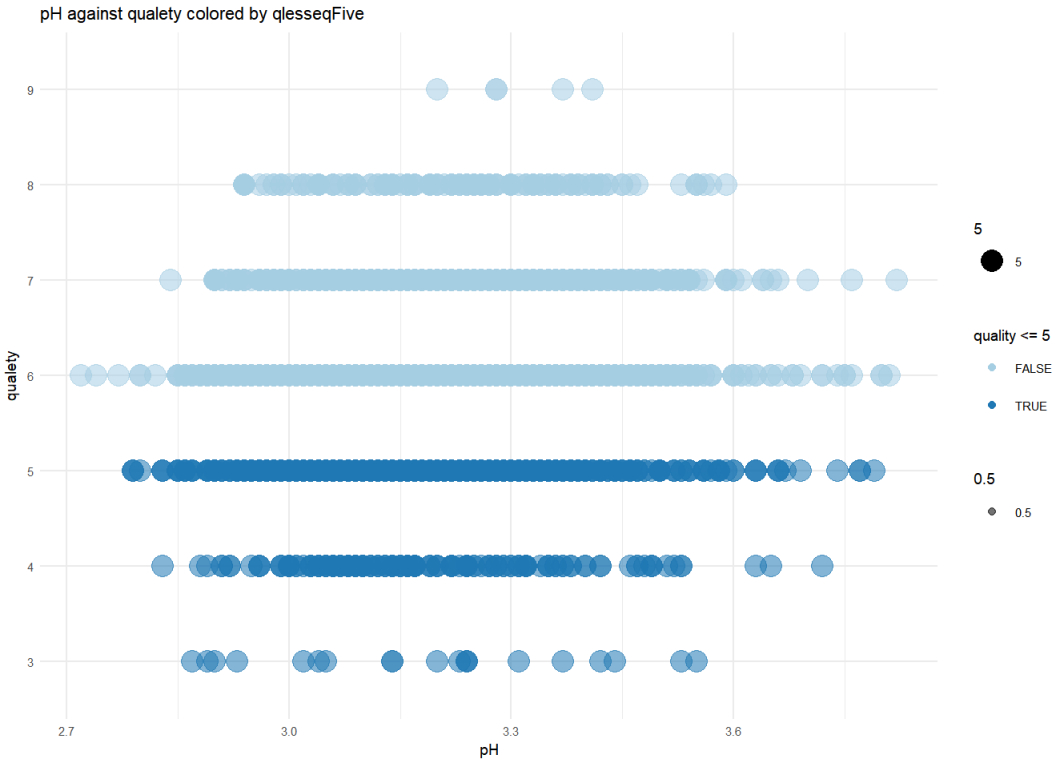

density against sulphates colored by qlesseqFive
sulphates
density
1.0
0.8
0.6
0.4
0.2
0.99
1.00
1.01
1.02
1.03
1.04
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

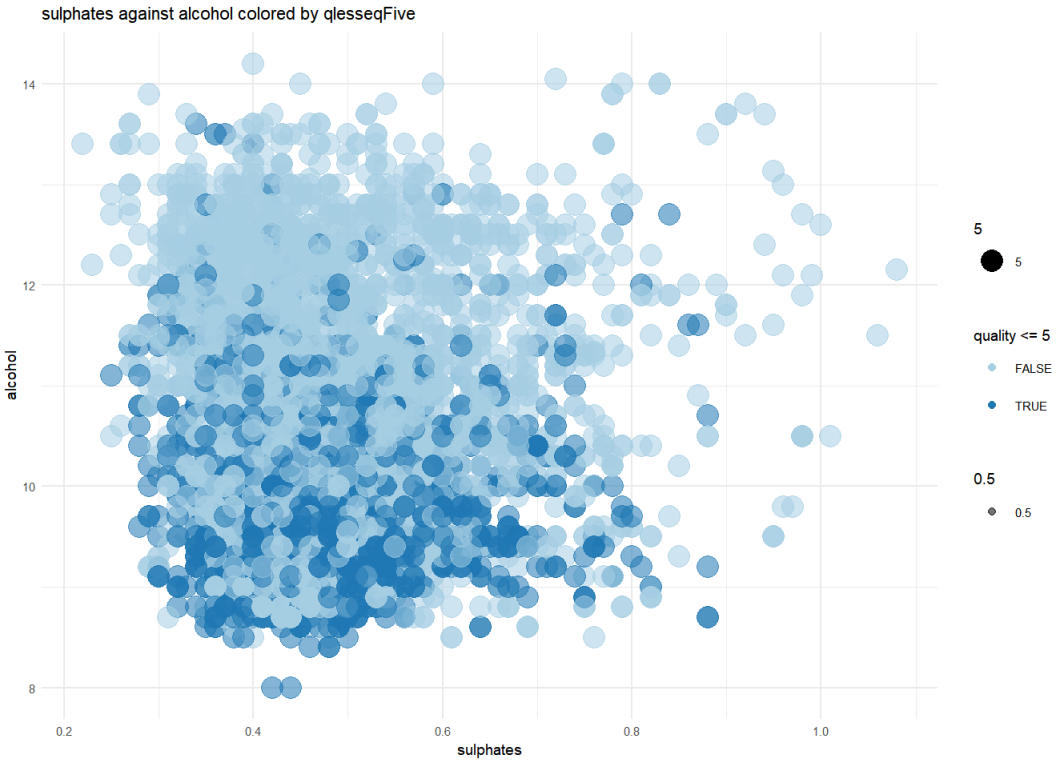

density against alcohol colored by qlesseqFive
alcohol
density
14
12
10
8
0.99
1.00
1.01
1.02
1.03
1.04
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

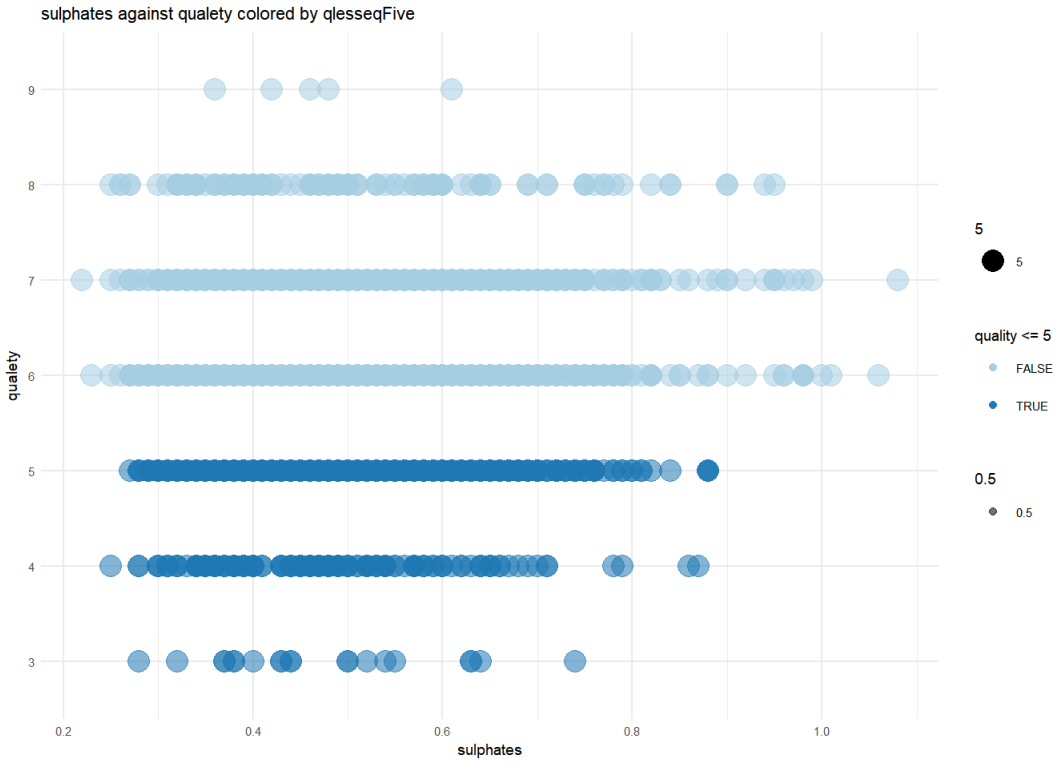

density against quality colored by qlesseqFive
quality
density
9
8
7
6
5
4
3
0.99
1.00
1.01
1.02
1.03
1.04
5
5
quality <= 5
FALSE
TRUE
0.5
0.5

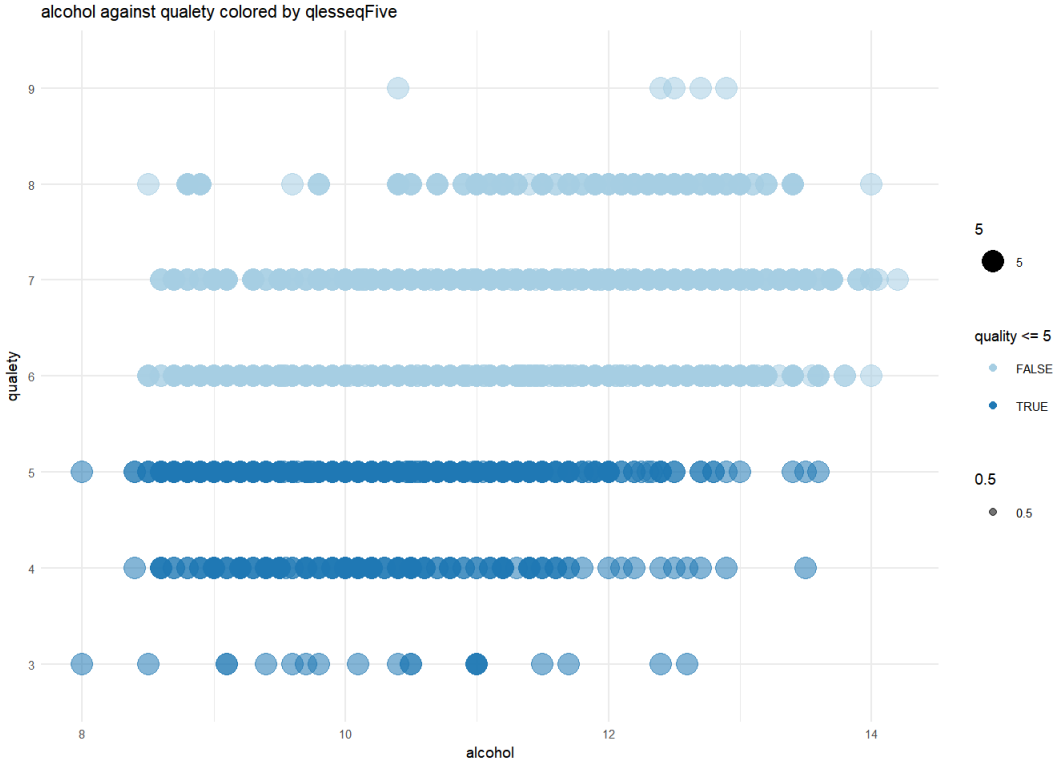

pH against sulphates colored by qlesseqFive
sulphates
1.0
0.8
0.6
0.4
0.2
2.7
3.0
3.3
3.6
pH
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


pH against alcohol colored by qlesseqFive
alcohol
14
12
10
8
2.7
3.0
3.3
3.6
pH
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


pH against qualety colored by qlesseqFive
qualety
9
8
7
6
5
4
3
2.7
3.0
3.3
3.6
pH
5
5
quality <= 5
FALSE
TRUE
0.5
0.5


sulphates against alcohol colored by qlesseqFive
alcohol
14
12
10
8
0.2
0.4
0.6
0.8
1.0
sulphates
5
5
quality <= 5
FALSE
TRUE
0.5
0.5





Aim of generating these plots is to find the right combination of variables that taht can help us in distinguishing wines with quality less than or equal to five with thos having higher quality.

Multivariate Analysis

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

From the plots we can infer that the combination of sulphates and alcohol, the combination of chlorides and alcohol, the combination of volate.acidity and alcohol, and the combination of volatile.acidity and sulphates seem to able to help us distinguish wines with higher quality and wines with lower quality less than or equal to five.

Were there any interesting or surprising interactions between features?

Even though free.sulfur.dioxide and total.sulfur.dioxide are moderately correlated with each other, from the plots we can see that , many low quality wine tend to have higher value of total.sulfur.dioxide for a given value of free.sulfur.dioxide. So, these two variables can together

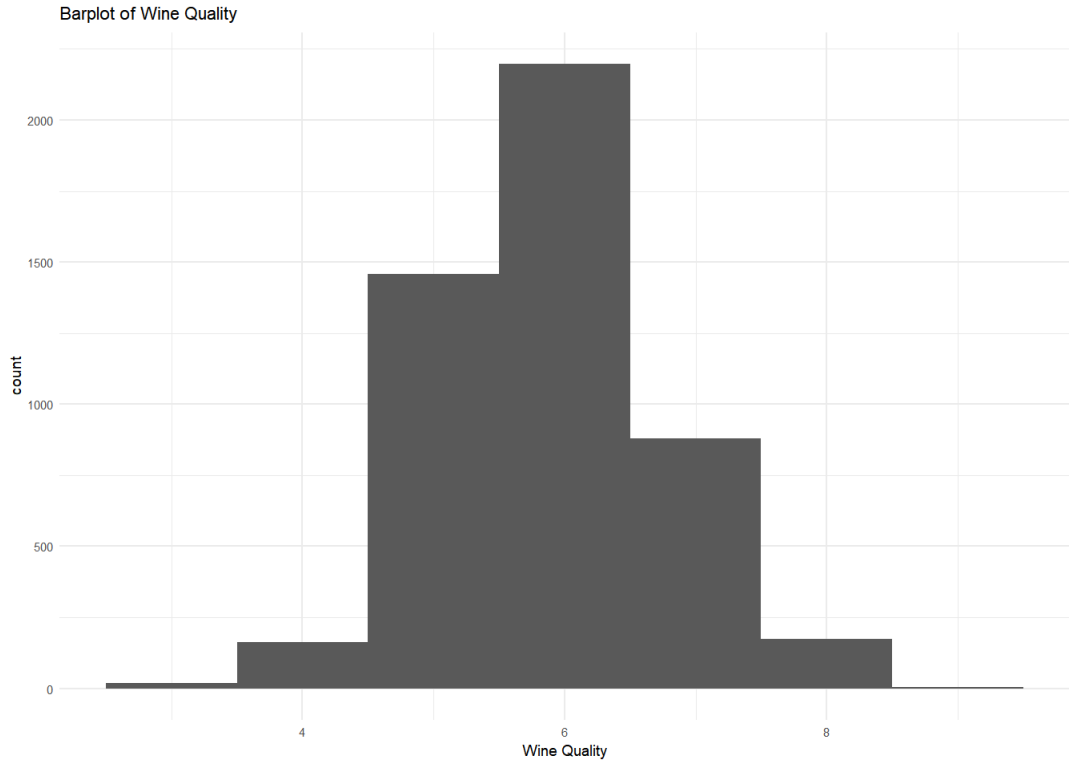
provide some resoning for wine quality.

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

No

Final Plots and Summary

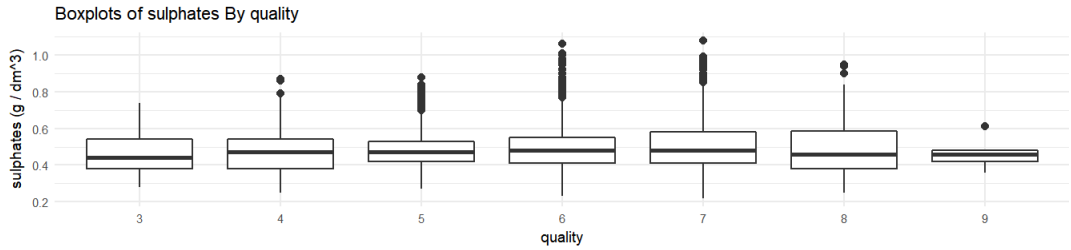
Plot One

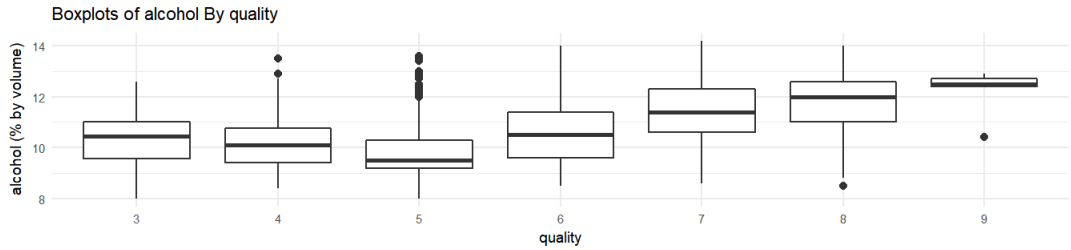


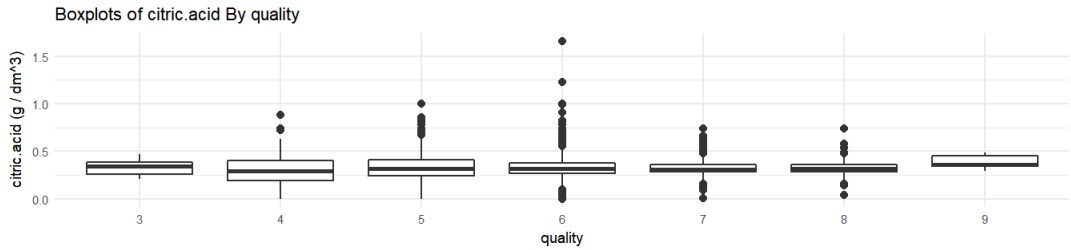
Description One

The range of quality from 0 to 5 , 6 and above divides the data set in two chunks of roughly same size and the most no. of wines lying in the qujality region of 5 and 6

Plot Two



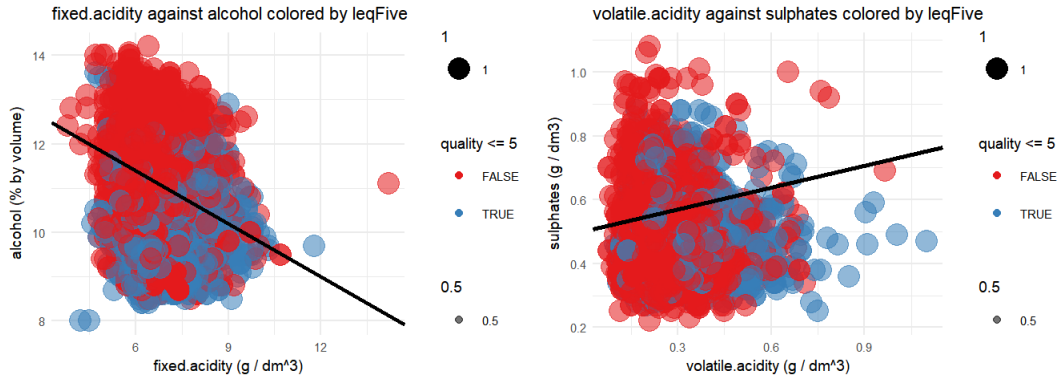


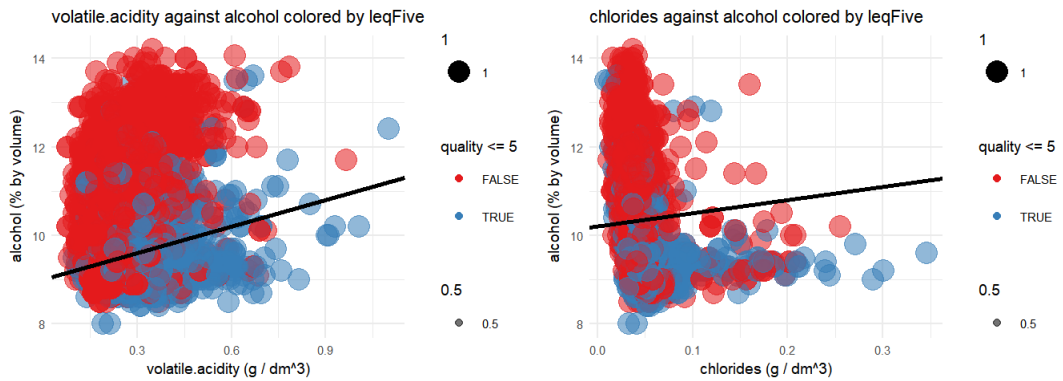


Description Two

These plots shows us that the median values of variableS: sulphates, alcohol, citric.acid increases as the quality of wine increases.

Plot Three





Description Three

We use two varaibles though individually they have weak correlation in combinations of two to have a line that can separate wine with quality less tha or equal to five with those having higher quality.

Reflection

In this data set I tried to find out how the quality of a wine is related to it’s differnet properties, but it was frustrating for a moment to see that

a variable individually cannot do justice in telling about the wine quality.

We can use different variables in combination to have a better grasp of the wine quality.

One thing that I did not like about this data set was that quality was the only variable that seemed to me somewhat explorable. It would have been great if a factor of price or other economical values were also present in this data set.