RedWine_202211

Stephan

2022-11-24

Install all needed libraries and load them

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.2.2
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.2.2

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:reshape2':
##
##     smiths
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.2
```

```
##
## Attaching package: ’gridExtra’

## The following object is masked from ’package:dplyr’:
##
##     combine
```

```
library(ggplot2)
library(alr4)
```

```
## Warning: package ’alr4’ was built under R version 4.2.2

## Loading required package: car

## Warning: package ’car’ was built under R version 4.2.2

## Loading required package: carData

## Warning: package ’carData’ was built under R version 4.2.2

##
## Attaching package: ’car’

## The following object is masked from ’package:dplyr’:
##
##     recode

## Loading required package: effects

## Warning: package ’effects’ was built under R version 4.2.2

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(GGally)
```

```
## Warning: package ’GGally’ was built under R version 4.2.2

## Registered S3 method overwritten by ’GGally’:
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
library(GGally)
library(scales)
```

```
## Warning: package ’scales’ was built under R version 4.2.2
```

```
library(memisc)
```

```
## Warning: package 'memisc' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'memisc'
```

```
## The following object is masked from 'package:scales':
##
##     percent
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:dplyr':
##
##     collect, recode, rename, syms
```

```
## The following object is masked from 'package:ggplot2':
##
##     syms
```

```
## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts
```

```
## The following object is masked from 'package:base':
##
##     as.array
```

Import the data set

```
## [1] "C:/Users/steph/Documents/RedWineAnalysis2022/docs"
```

Uni-variate Plots Section

Determine the class, structure of data set as well as a summary.

Print a head of the data to see what it looks like.

```
## [1] "data.frame"
```

```
## ’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 ...
```

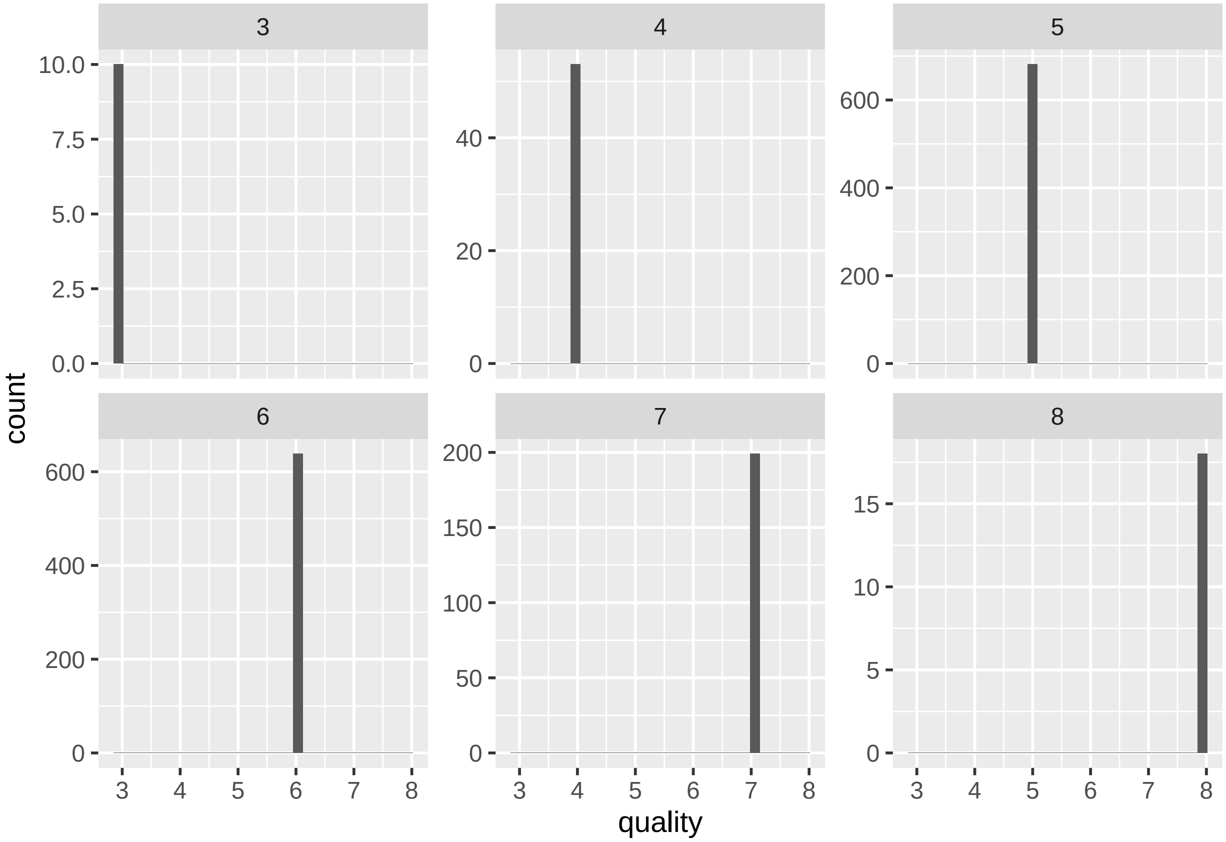
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00             1.9     0.076
## 2 2           7.8             0.88        0.00             2.6     0.098
## 3 3           7.8             0.76        0.04             2.3     0.092
## 4 4          11.2             0.28        0.56             1.9     0.075
## 5 5           7.4             0.70        0.00             1.9     0.076
## 6 6           7.4             0.66        0.00             1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

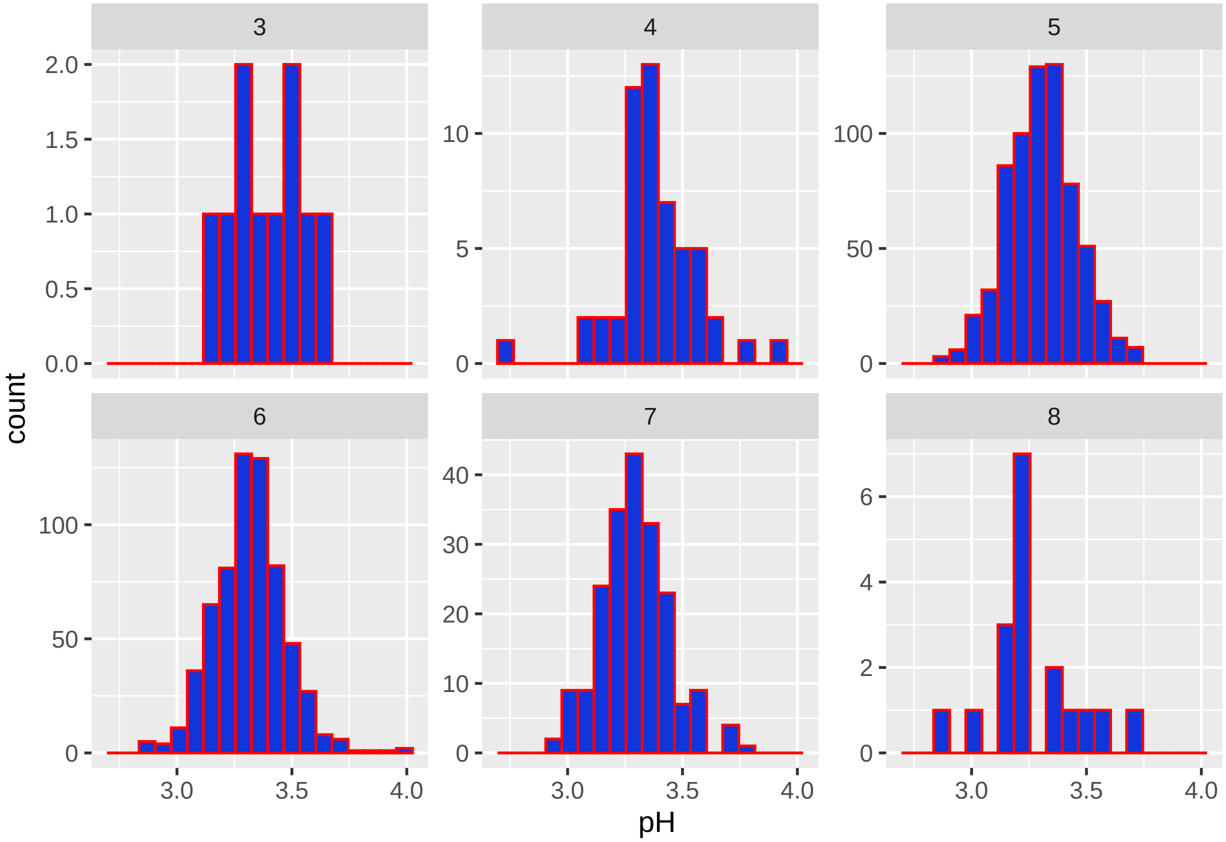
Draw different graphs to show the relationship of all the determining factors to the quality of the wine.

```
ggplot(red_wine,aes(x = quality))+
  geom_histogram()+
  facet_wrap(~quality, scales = 'free_y')
```

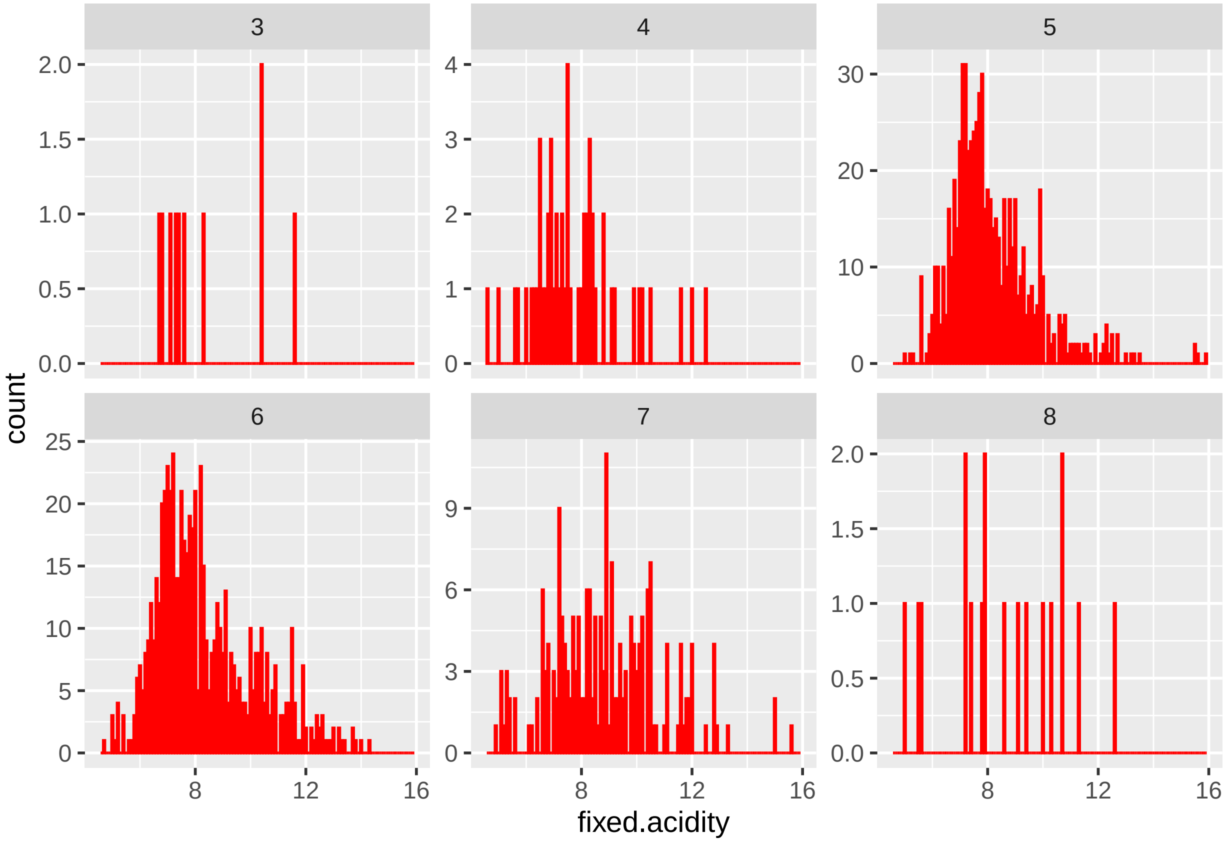
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



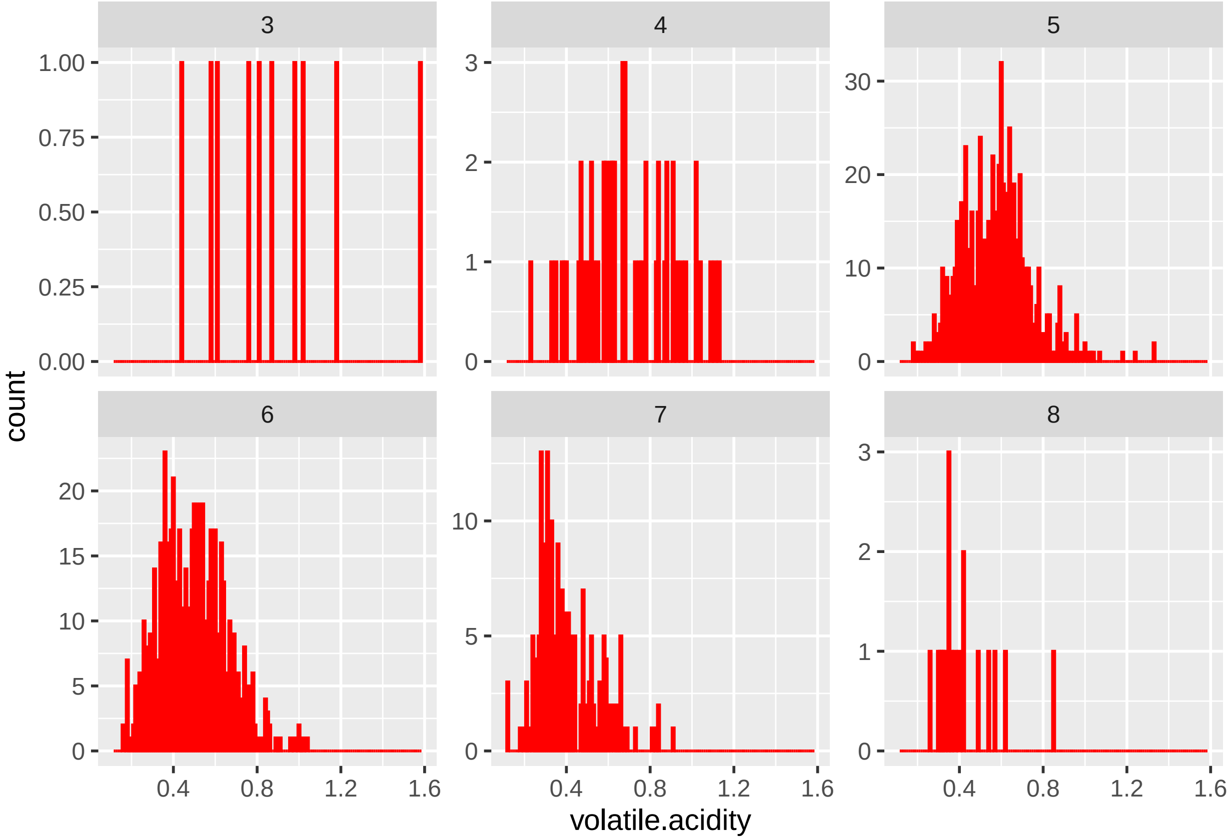
```
ggplot(red_wine, aes(x = pH))+
  geom_histogram(binwidth = 0.07, fill = '#1234DC', color = 'red')+
  facet_wrap(~quality, scales = 'free_y')
```



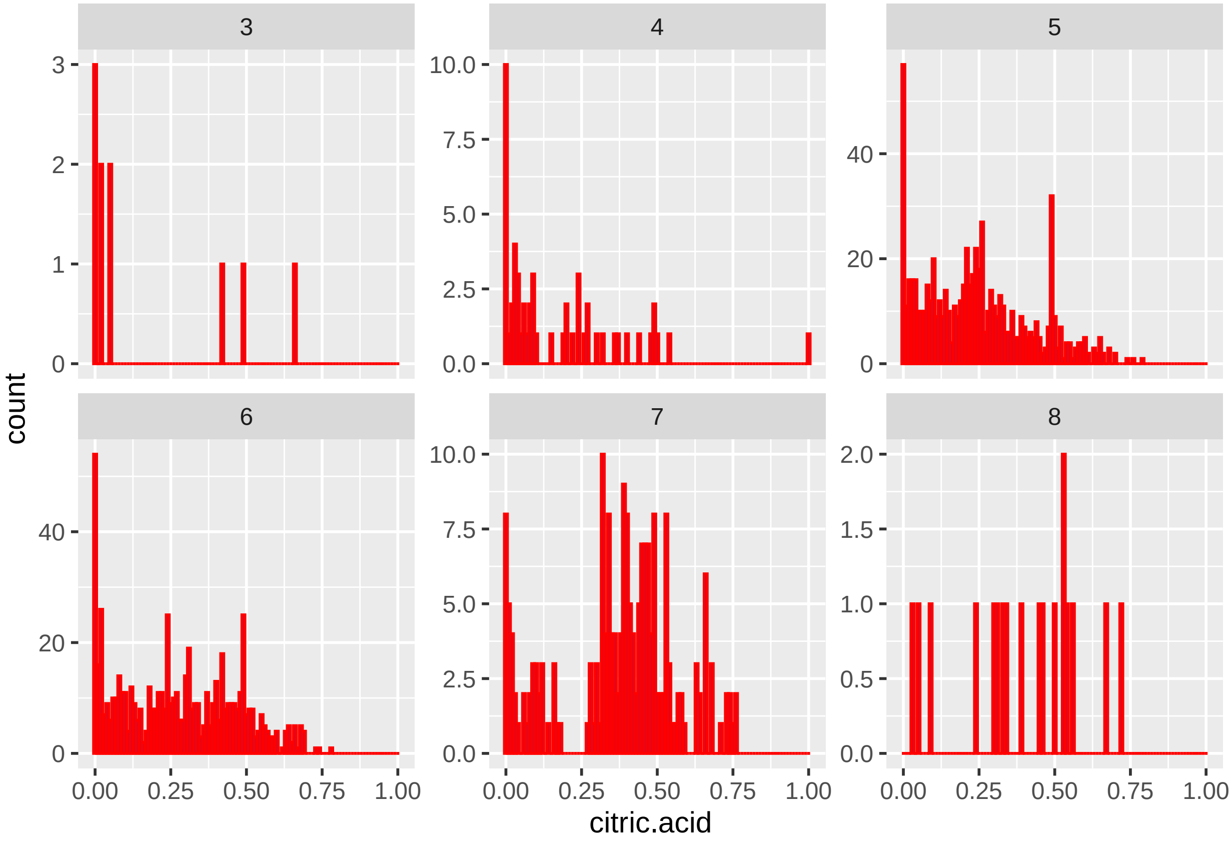
```
ggplot(red_wine, aes(x = fixed.acidity))+
  geom_histogram(binwidth = 0.05, fill = '#1234DC', color = 'red')+
  facet_wrap(~quality, scales = 'free_y')
```



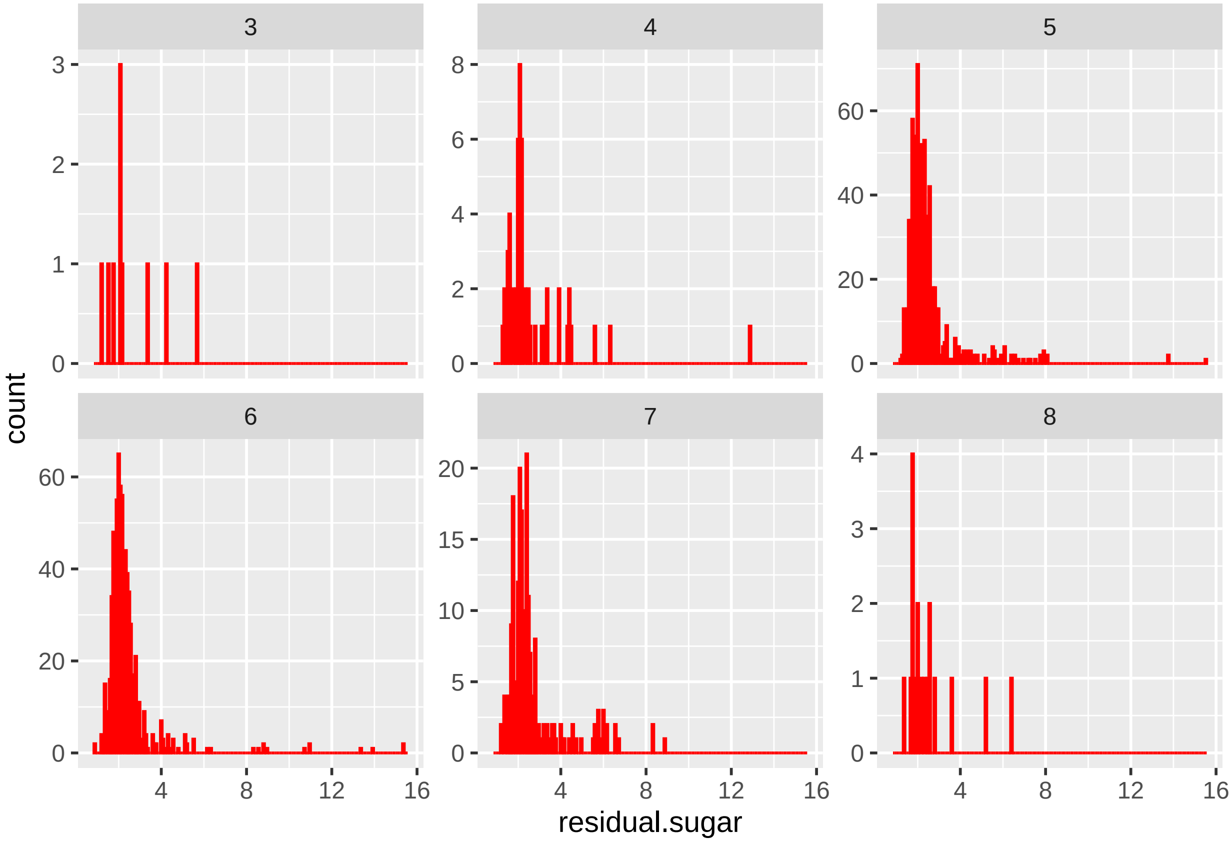
```
ggplot(red_wine, aes(x = volatile.acidity))+
  geom_histogram(binwidth = 0.01, fill = '#1234DC', color = 'red')+
  facet_wrap(~quality, scales = 'free_y')
```



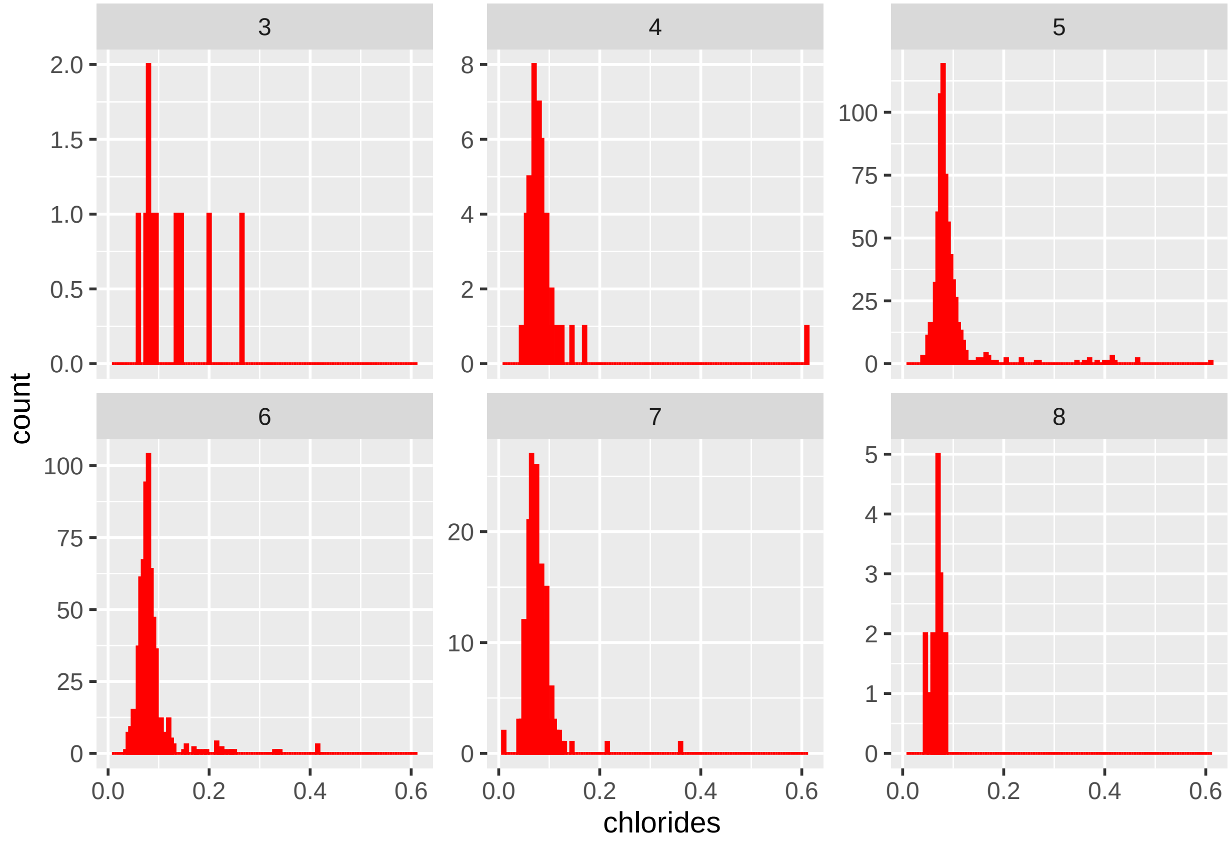
```
ggplot(red_wine, aes(x = citric.acid))+
  geom_histogram(binwidth = 0.01, fill = '#1234DC', color = 'red')+
  facet_wrap(~quality, scales = 'free_y')
```



```
ggplot(red_wine, aes(x = residual.sugar))+
  geom_histogram(binwidth = 0.08, fill = '#1234DC', color = 'red')+
  facet_wrap(~quality, scales = 'free_y')
```

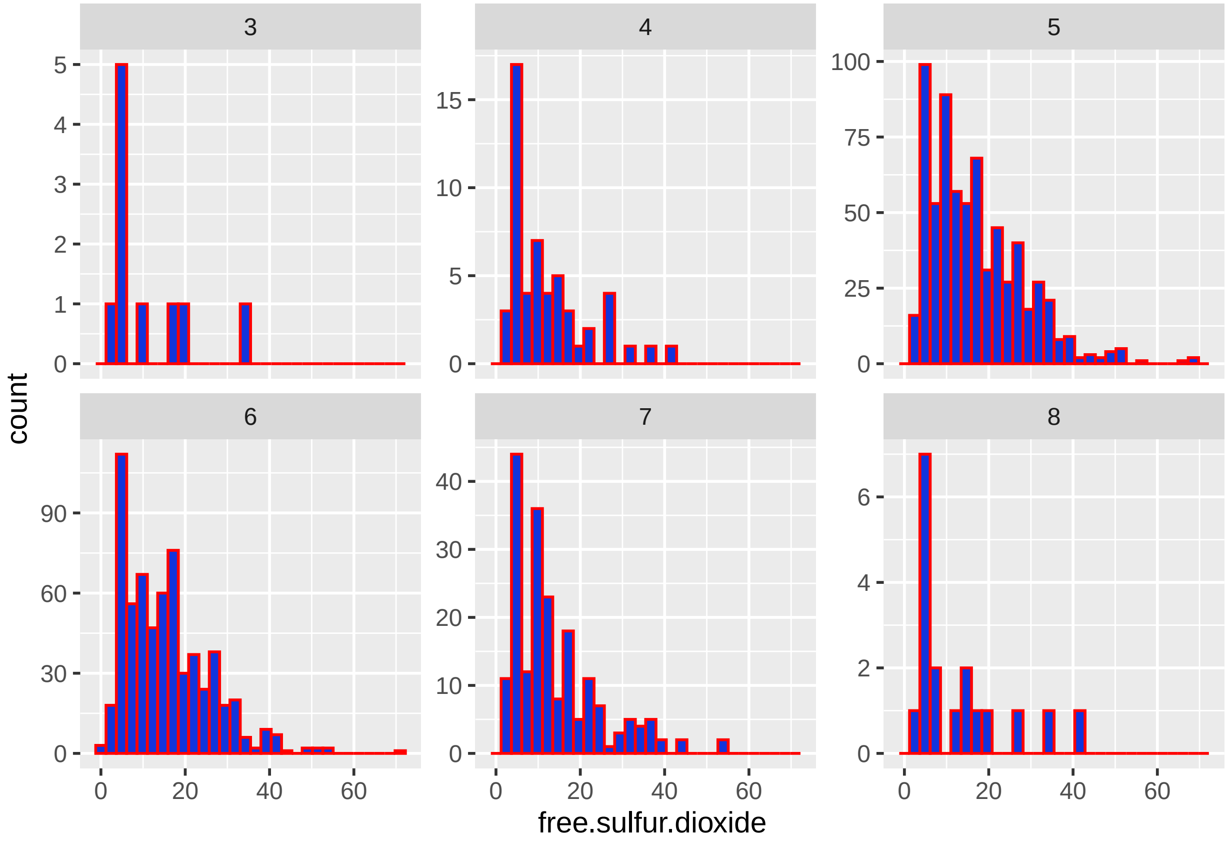


```
ggplot(red_wine, aes(x = chlorides))+
 geom_histogram(binwidth = 0.005, fill = '#1234DC', color = 'red')+
 facet_wrap(~quality, scales = 'free_y')
```



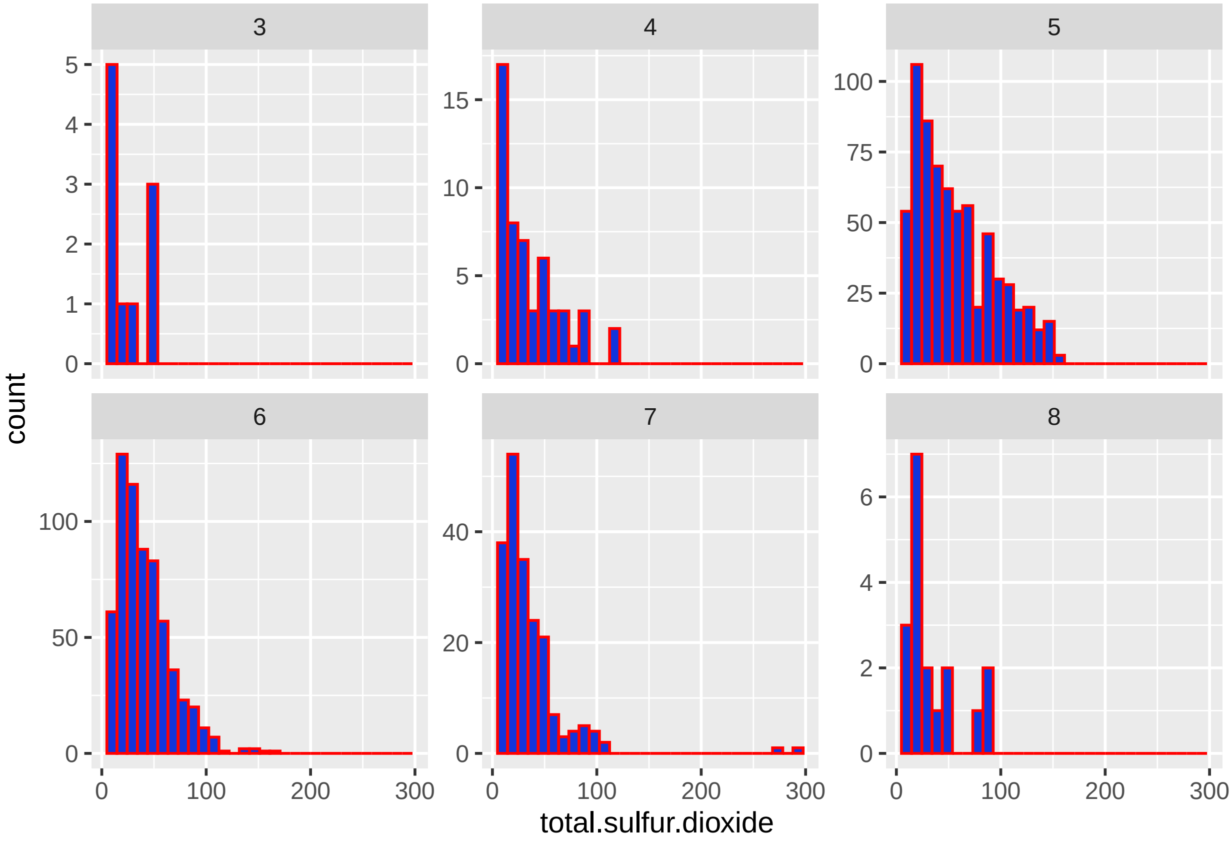
```
ggplot(red_wine, aes(x = free.sulfur.dioxide))+
  geom_histogram(fill = '#1234DC', color = 'red')+
  facet_wrap(~quality, scales = 'free_y')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

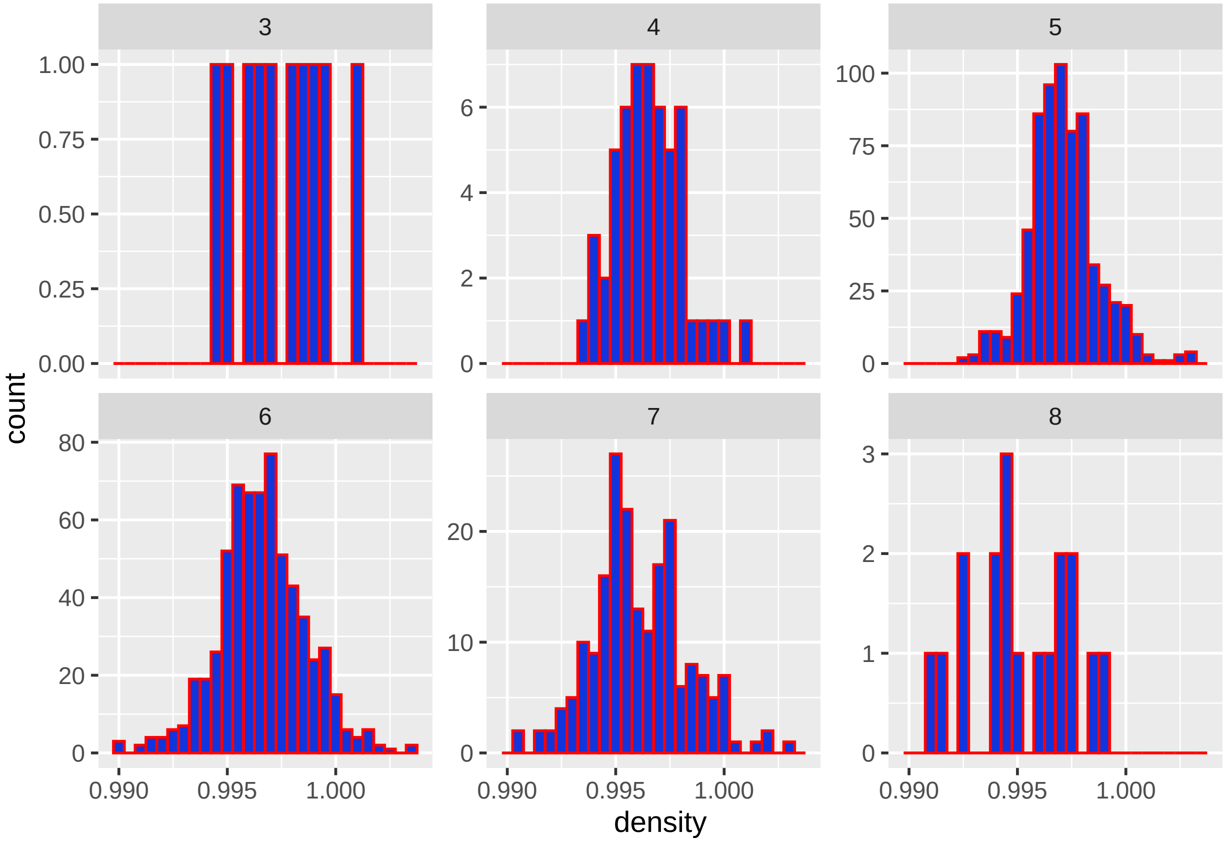


```r
ggplot(red_wine, aes(x = total.sulfur.dioxide))+
 geom_histogram(fill = '#1234DC', color = 'red')+
 facet_wrap(~quality, scales = 'free_y')
```

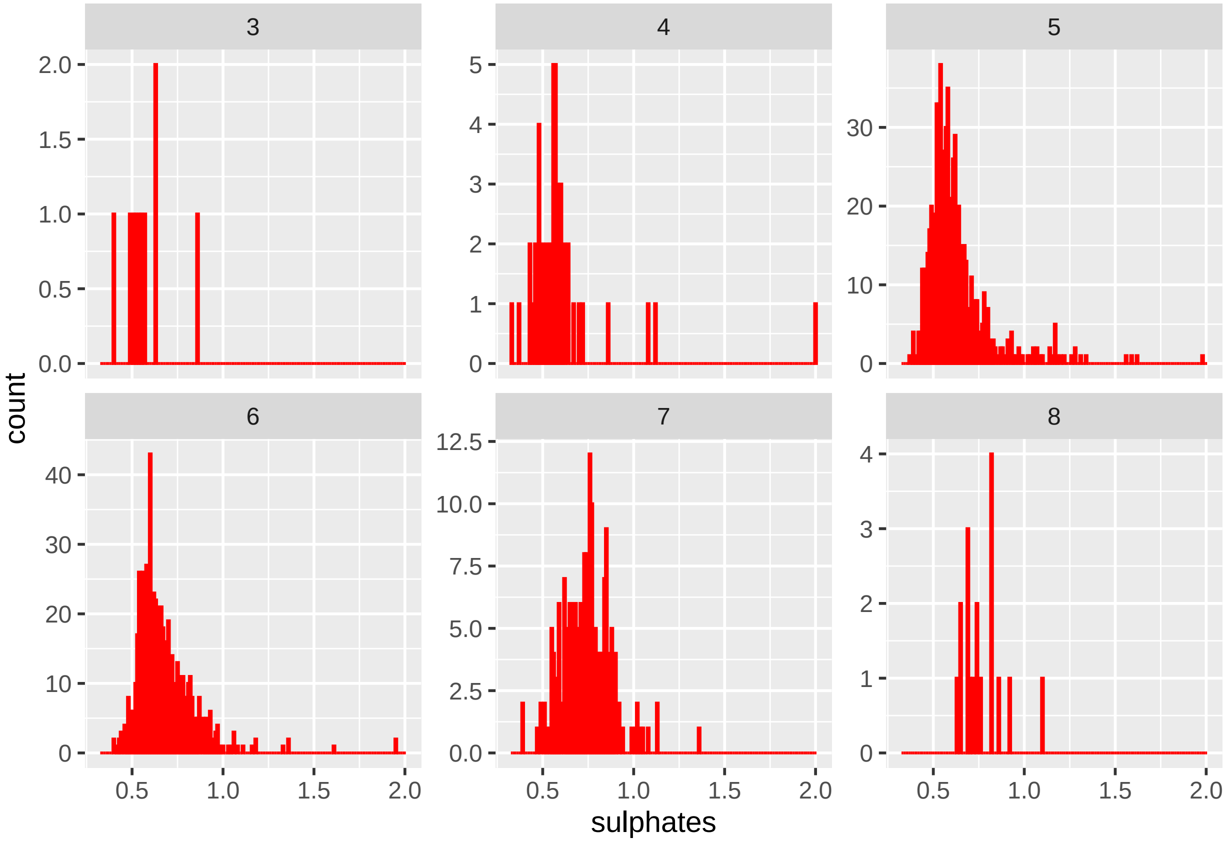
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



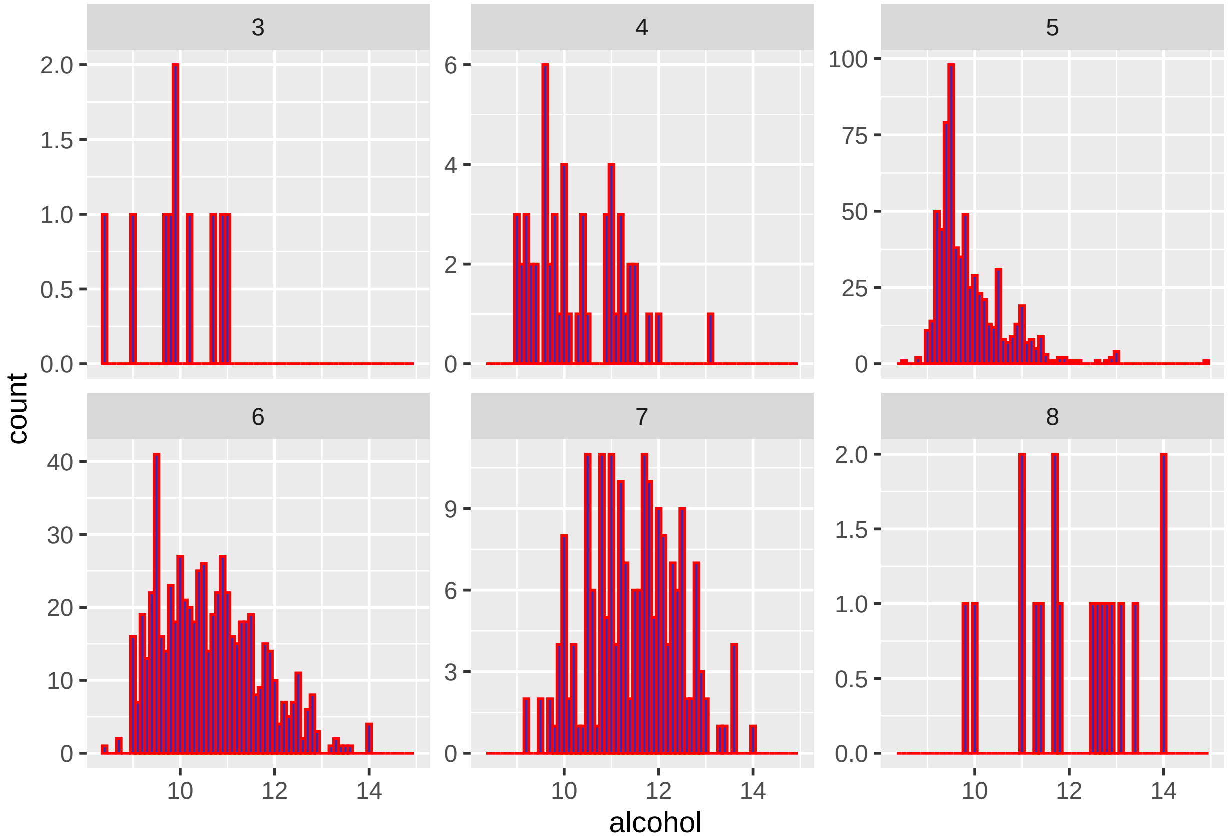
```
ggplot(red_wine, aes(x = density))+
 geom_histogram(binwidth = 0.0005, fill = '#1234DC', color = 'red')+
 facet_wrap(~quality, scales = 'free_y')
```



```
ggplot(red_wine, aes(x = sulphates))+
 geom_histogram(binwidth = 0.01, fill = '#1234DC', color = 'red')+
 facet_wrap(~quality, scales = 'free_y')
```



```
ggplot(red_wine, aes(x = alcohol))+
 geom_histogram(binwidth = 0.1, fill = '#1234DC', color = 'red')+
 facet_wrap(~quality, scales = 'free_y')
```



```
summary(red_wine$quality)
```

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

Calculate the amount of wines above and below the mean and median

```
###Sum above and below Mean and Medain Quality
sum(red_wine$quality)
```

```
## [1] 9012
```

```
sum(red_wine$quality> mean(red_wine$quality))
```

```
## [1] 855
```

```
sum(red_wine$quality< mean(red_wine$quality))
```

```
## [1] 744
```

```
sum(red_wine$quality> median(red_wine$quality))
```

```
## [1] 217
```

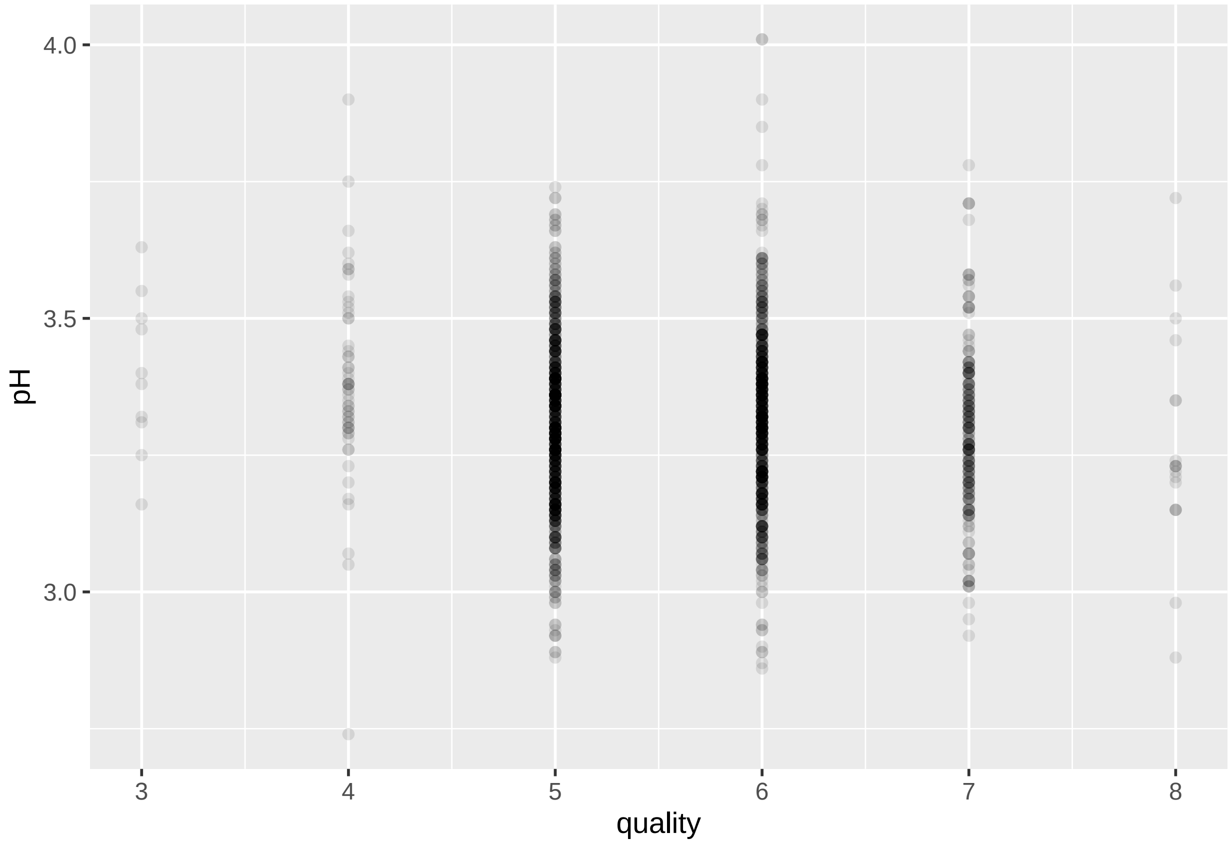
```
sum(red_wine$quality< median(red_wine$quality))
```

```
## [1] 744
```

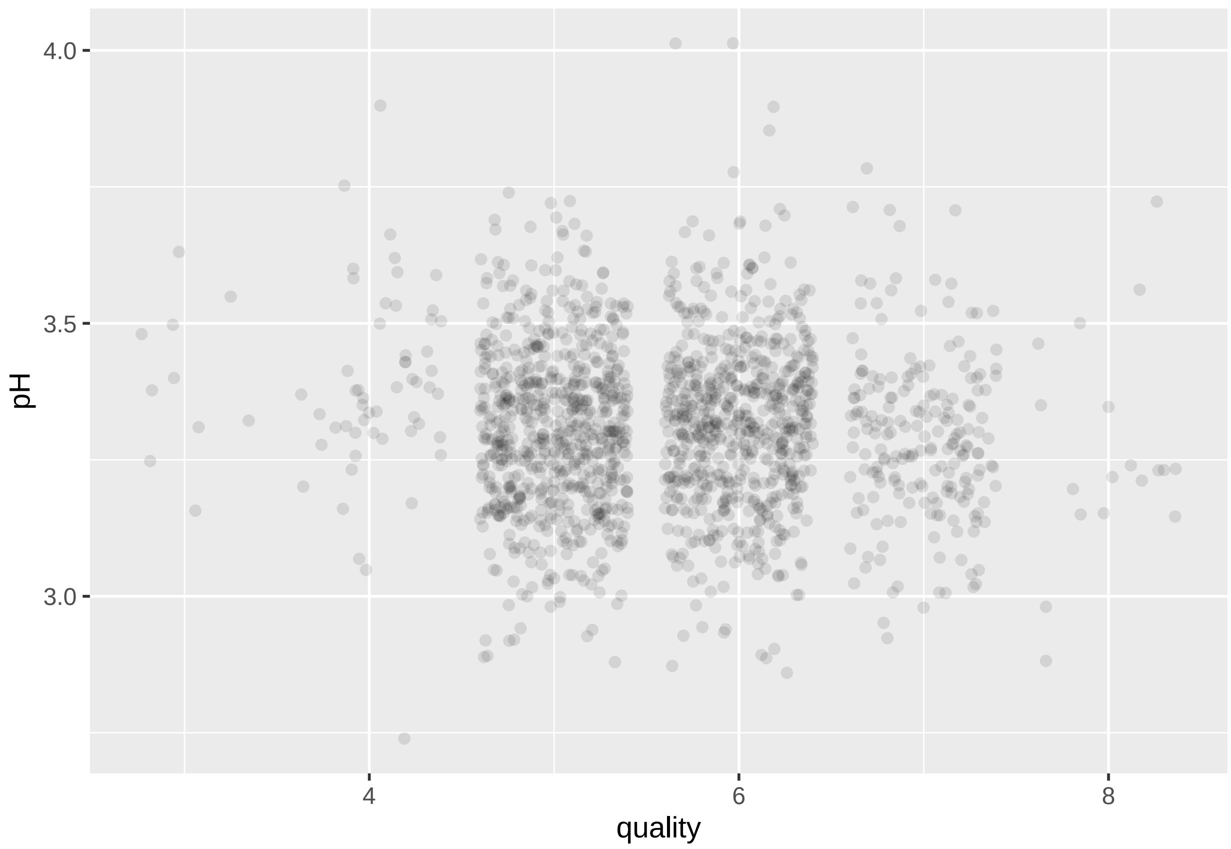
3 different graphs:

1 - quality vs pH with geom point / scatter plot
2 - quality vs pH with geom jitter plot
3 - quality vs alcohol with jitter plot and line chart

```
ggplot(aes(x = quality, y = pH),data = red_wine)+
  geom_point(alpha = 1/10)
```

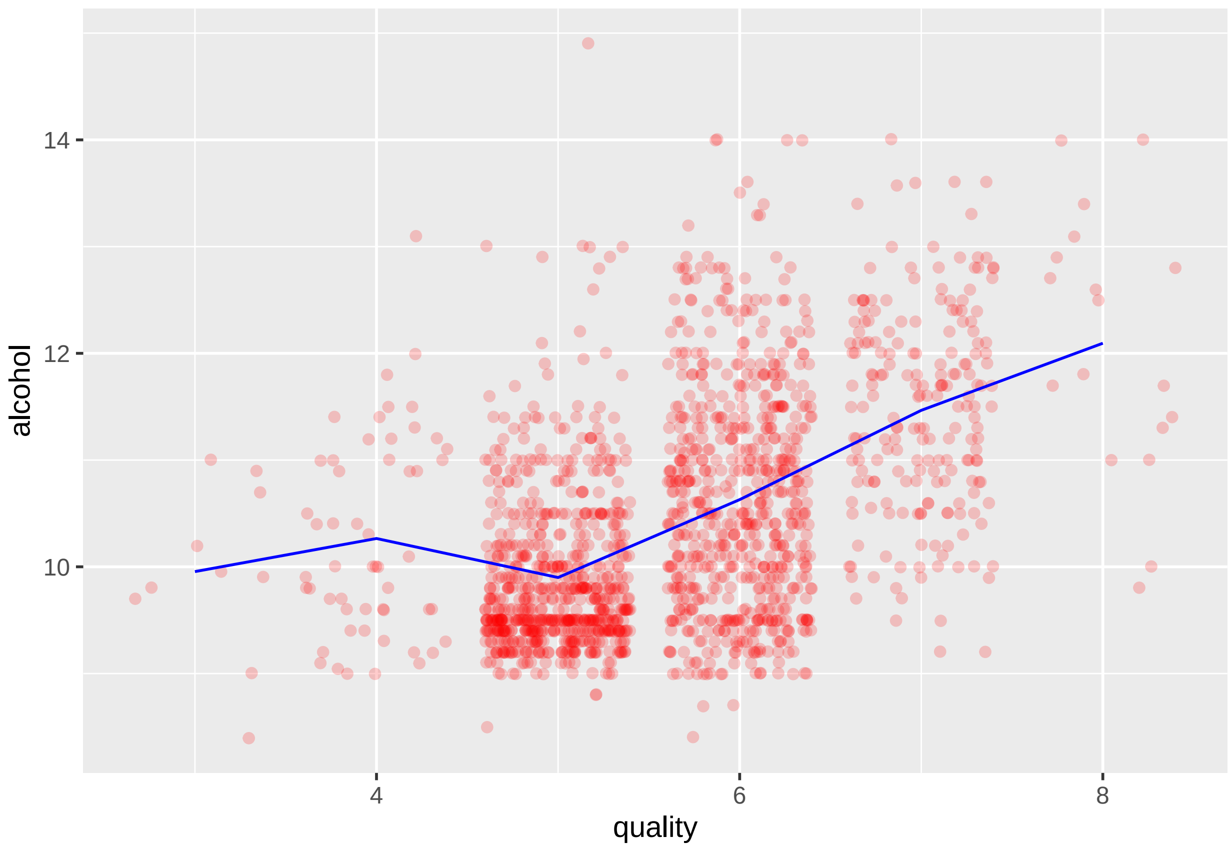


```
ggplot(aes(x = quality, y = pH),data = red_wine)+
  geom_jitter(alpha = 1/10)
```



```
ggplot(aes(x = quality, y = alcohol),data = red_wine)+
  geom_jitter(alpha = 1/5, color = "red")+
  geom_line(stat = "summary", color = "blue")
```

```
## No summary function supplied, defaulting to ‘mean_se()‘
```

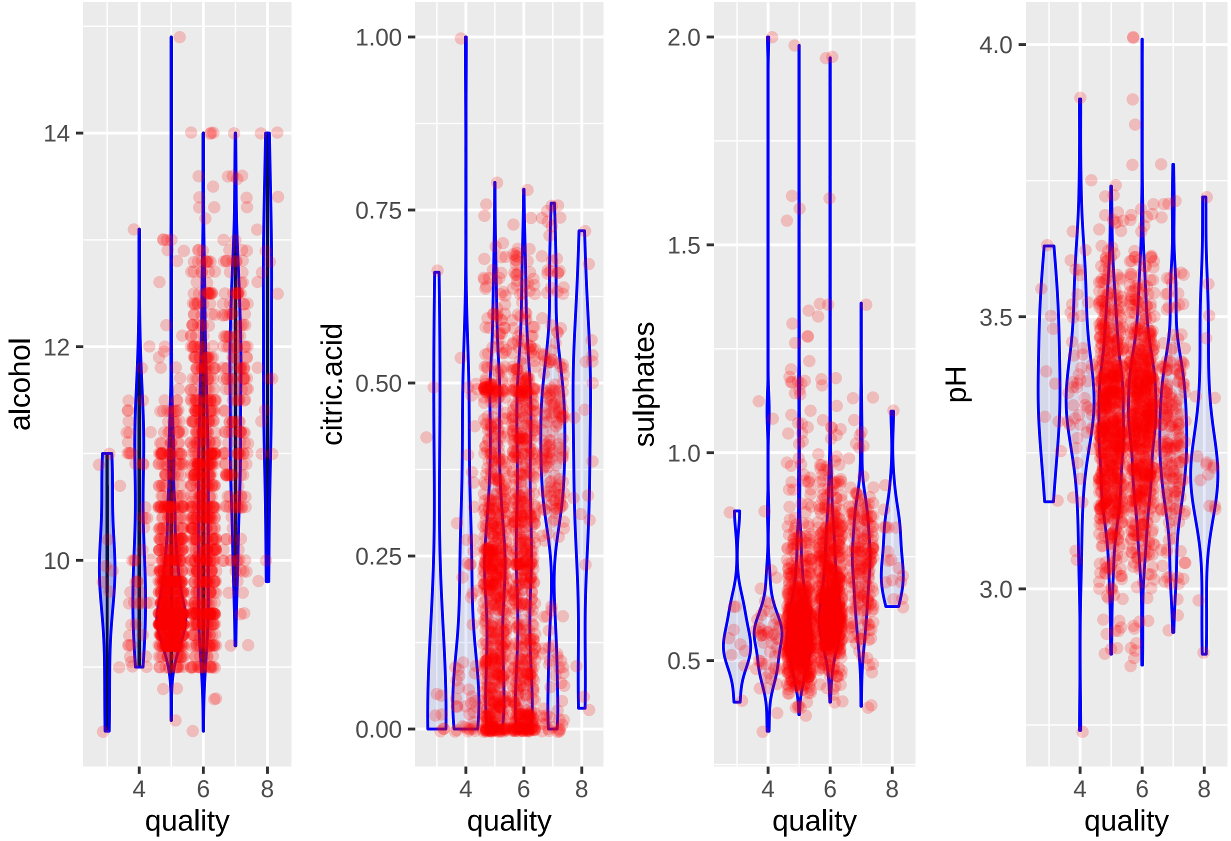


4 x violin plots between quality and 4 influencing elements:

-alcohol
-citric.acid
-sulphates
-pH

```
library(gridExtra)
library(grid)


plot1 <- ggplot(aes(x = quality, y = alcohol, group = quality),data = red_wine)+
            geom_line()+
            geom_violin(alpha = 1/3, color = "blue", fill = '#1234DC' )+
            geom_jitter(alpha = 1/5, color = "red")




plot2 <- ggplot(aes(x = quality, y = citric.acid, group = quality),data = red_wine)+
            geom_violin(alpha = 1/10, color = "blue", fill = '#1234DC' )+
            geom_jitter(alpha = 1/5, color = "red")


plot3 <- ggplot(aes(x = quality, y = sulphates, group = quality),data = red_wine)+
            geom_violin(alpha = 1/10, color = "blue", fill = '#1234DC' )+
            geom_jitter(alpha = 1/5, color = "red")
```

```
plot4 <- ggplot(aes(x = quality, y = pH, group = quality),data = red_wine)+
        geom_violin(alpha = 1/10, color = "blue", fill = '#1234DC' )+
        geom_jitter(alpha = 1/5, color = "red")


grid.arrange(plot1, plot2 ,plot3, plot4, ncol = 4)
```

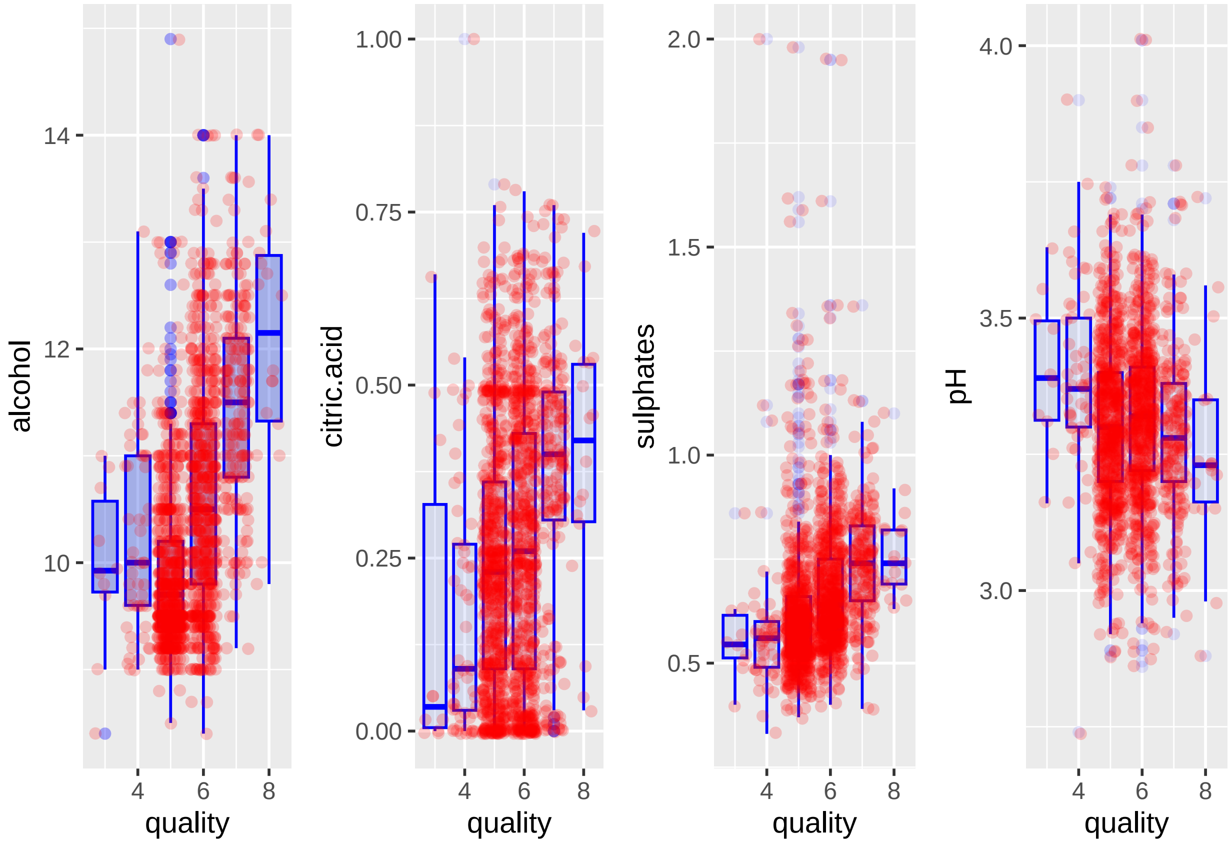


4 x box plots between quality and 4 influencing elements:

-alcohol
-citric.acid
-sulphates
-pH

```
library(gridExtra)
library(grid)


plot5 <- ggplot(aes(x = quality, y = alcohol, group = quality),data = red_wine)+
        geom_boxplot(alpha = 1/3, color = "blue", fill = '#1234DC' )+
        geom_jitter(alpha = 1/5, color = "red")

plot6 <- ggplot(aes(x = quality, y = citric.acid, group = quality),data = red_wine)+
        geom_boxplot(alpha = 1/10, color = "blue", fill = '#1234DC' )+
        geom_jitter(alpha = 1/5, color = "red")
```

```
plot7 <- ggplot(aes(x = quality, y = sulphates, group = quality),data = red_wine)+
            geom_boxplot(alpha = 1/10, color = "blue", fill = '#1234DC' )+
            geom_jitter(alpha = 1/5, color = "red")

plot8 <- ggplot(aes(x = quality, y = pH, group = quality),data = red_wine)+
            geom_boxplot(alpha = 1/10, color = "blue", fill = '#1234DC' )+
            geom_jitter(alpha = 1/5, color = "red")


grid.arrange(plot5, plot6 ,plot7, plot8, ncol = 4)
```

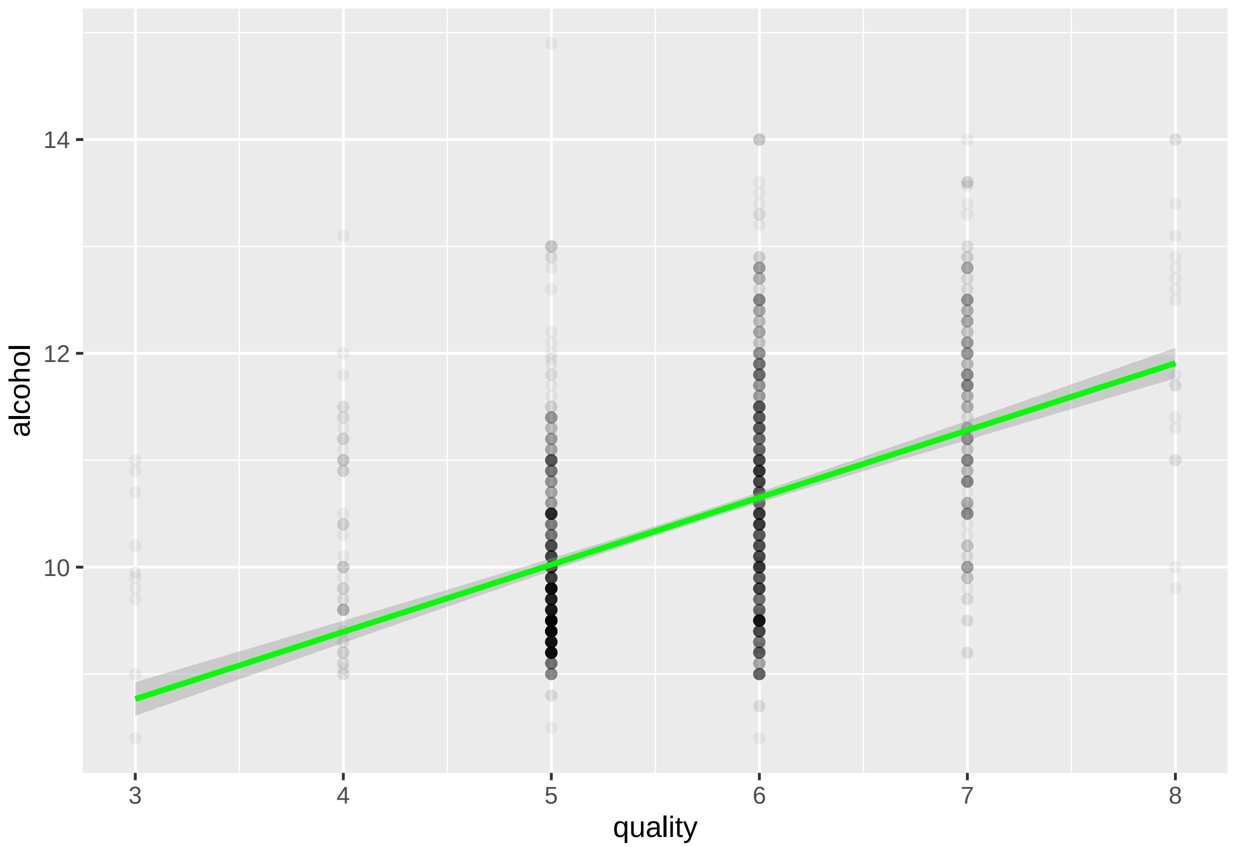


Two graphs showing the relationship between:

1- Quality vs Alcohol
2- Quality vs Volatile Acidity

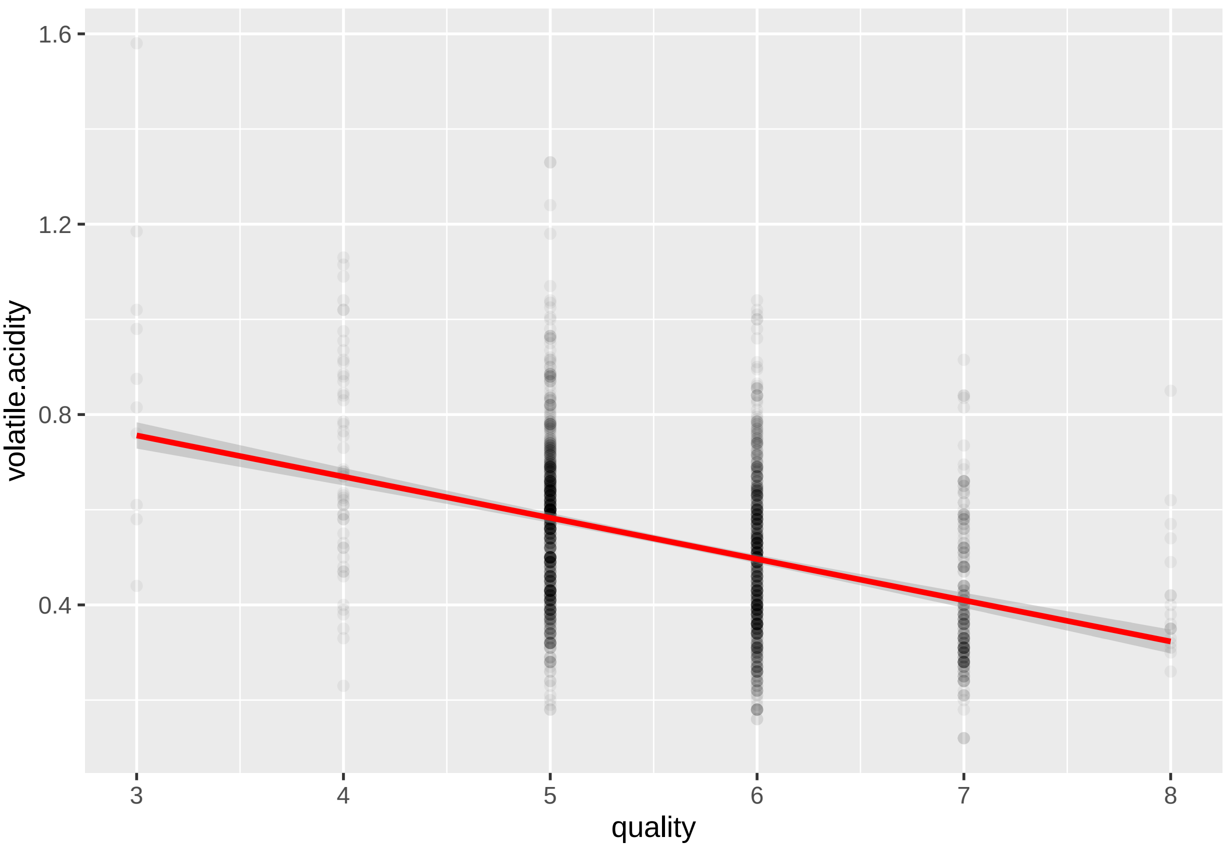
```
ggplot(aes(y = alcohol, x = quality), data = red_wine) +
  geom_point(alpha = 1/20)+
  geom_smooth(method = 'lm', color = 'green')
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
ggplot(aes(y = volatile.acidity, x = quality), data = red_wine) +
  geom_point(alpha = 1/20)+
  geom_smooth(method = 'lm', color = 'red')
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


1.6
1.2
0.8
0.4
volatile.acidity
3
4
5
6
7
8
quality