项目简介

都说 82 年的拉菲好，可我也不知道它好在哪啊，好不好是怎么定的呢？作为成分党，当然要看看葡萄酒的哪些因素影响它的质量了。

数据集

链接：

`https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%8E%A2%E7%B4%A2%E6%80%A7%E6%95%B0%E6%8D%AE%E5%88%86%E6%9E%90/%E9%A1%B9%E7%9B%AE/wineQualityReds.csv`

数据集包含 1,599 种红酒，以及 11 个关于酒的化学成分的变量。至少 3 名葡萄酒专家对每种酒的质量进行了评分，分数在 0（非常差）和 10（非常好）之间。

目的

看看哪些化学成分会影响葡萄酒的质量。

分析工具

RStudio

数据读取

```
library(ggplot2)
library(dplyr)
library(knitr)
library('GGally')

wine_red <- read.csv('wineQualityReds.csv')
wine_red <- wine_red[,-1]

str(wine_red)
```

```
'data.frame':   1599 obs. of  12 variables:
 $ 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 ...
```

可以看到，总共有 11 个化学成分变量和一个质量变量，共 12 个变量，化学成分变量都是数值型，质量是整数型。

```
summary(wine_red)
```

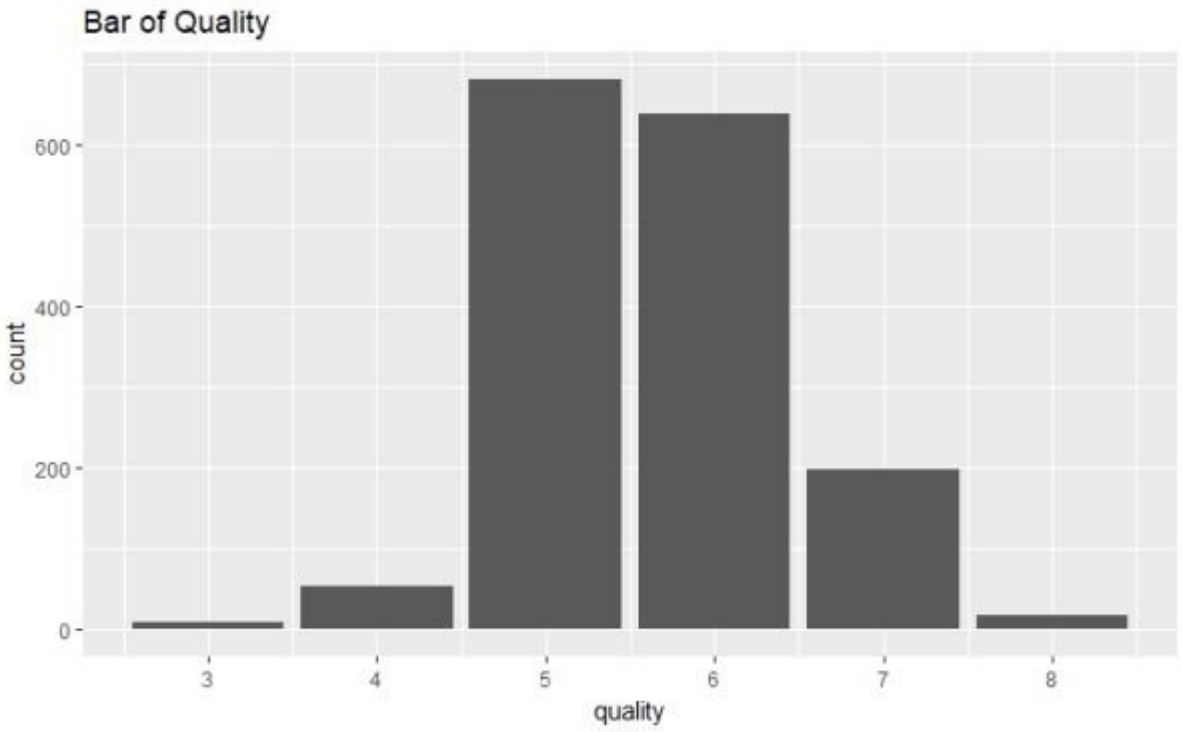
```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide
 Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900   Min.    :0.01200   Min.    : 1.00
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st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

探索分析

一、单变量分析

1.查看酒的质量的评分分布：

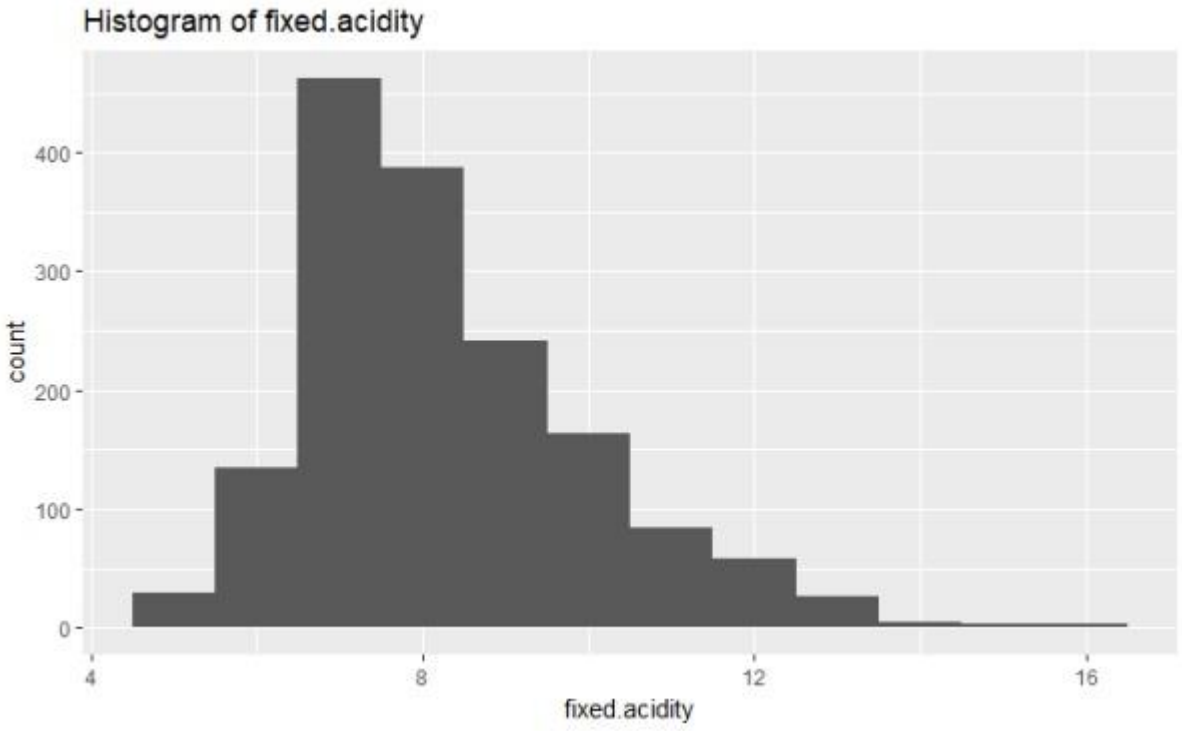
```
ggplot(wine_red,aes(x=quality)) +
  geom_bar() +
  scale_x_continuous(breaks=seq(3,8,1)) +
  ggtitle('Bar of Quality')
```



从上图可以看出，质量基本呈正态分布。

2.查看 fixed.acidity 的分布：

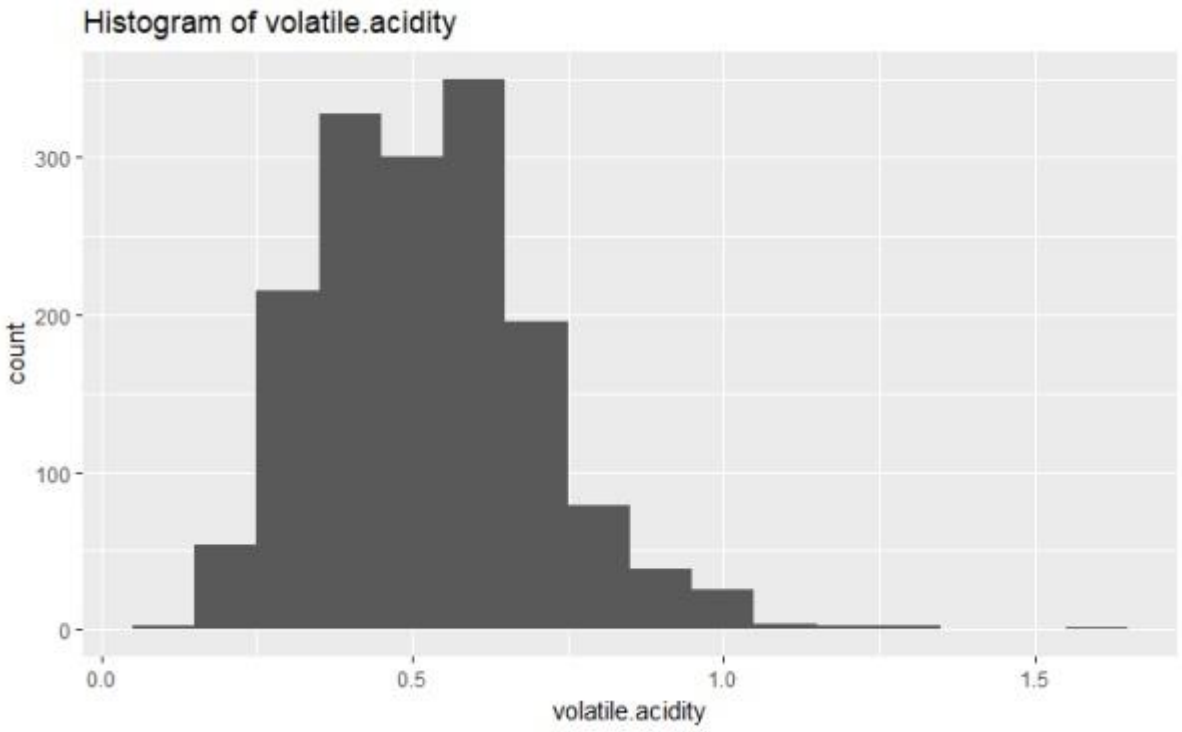
```
qplot(x=fixed.acidity,data=wine_red,binwidth=1) +
  ggtitle('Histogram of fixed.acidity')
```



上图表明 fixed.acidity 呈右偏分布,大多数在 6~12 之间。

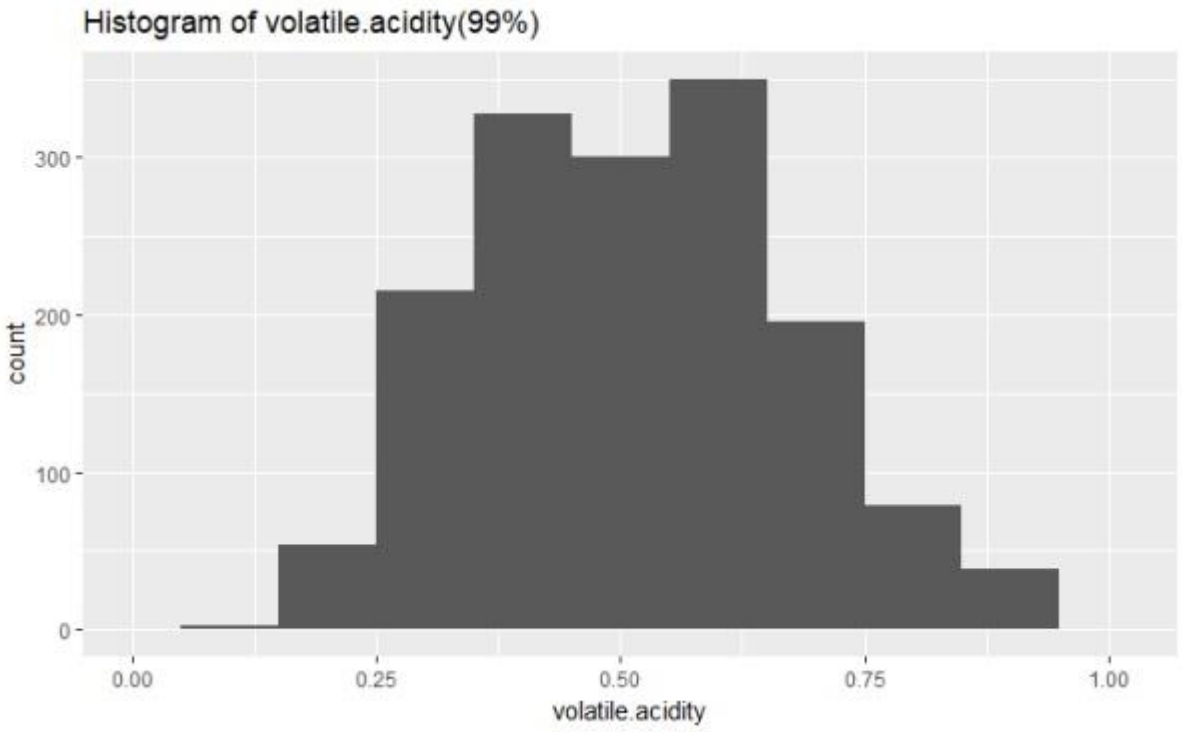
3.查看 volatile.acidity 的分布形态：

```
qplot(x=volatile.acidity,data=wine_red,binwidth=0.1) +
  ggtitle('Histogram of volatile.acidity')
```



volatile.acidity 也呈右偏分布，但其中有异常值，大多数值在 0.25~1 之间，我们取其前 99%的数绘图：

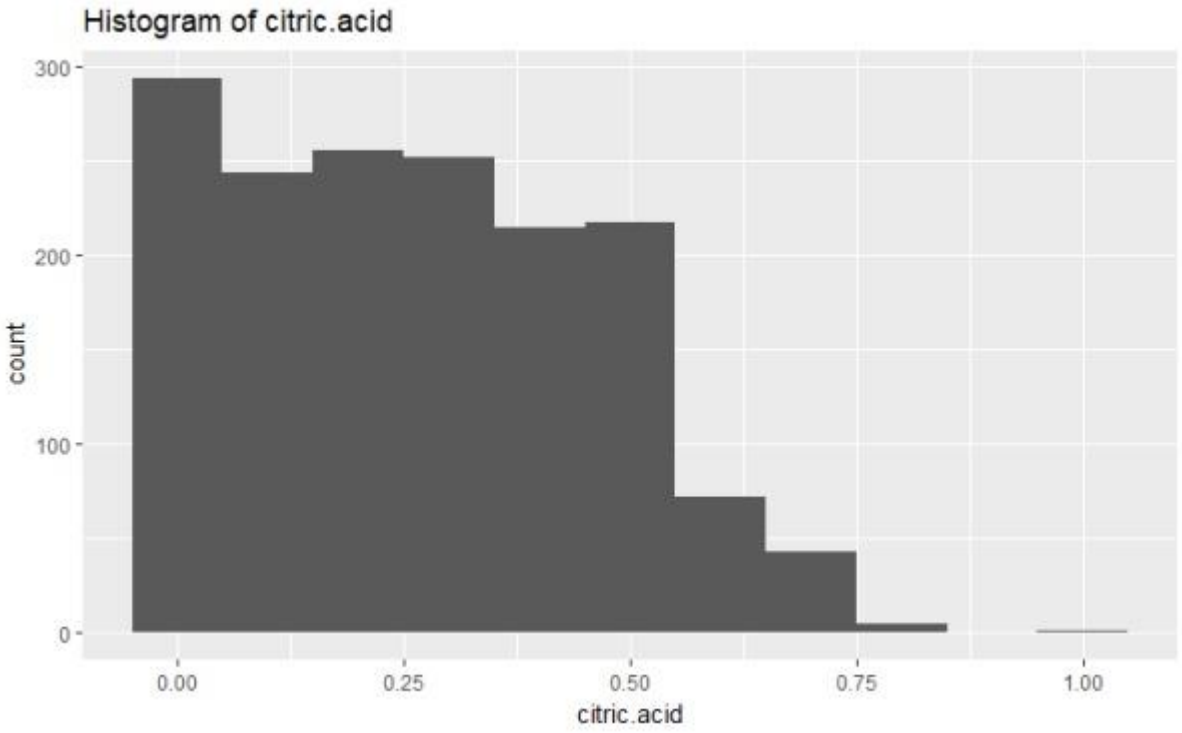
```
qplot(x=volatile.acidity,data=wine_red,binwidth=0.1) +
  ggtitle('Histogram of volatile.acidity(99%)') +
  xlim(0,quantile(wine_red$volatile.acidity,0.99))
```



上图可以看到，去掉异常值后，volatile.acidity 呈正态分布。

4.查看 citric.acid 的分布：

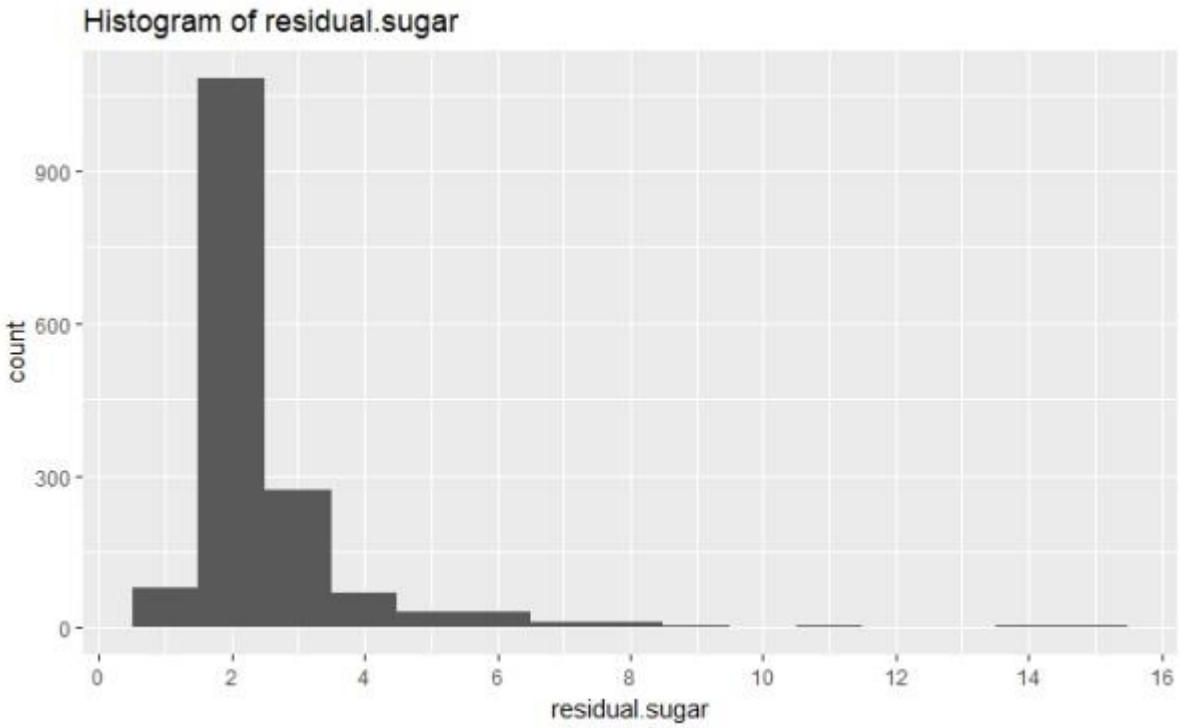
```
qplot(x=citric.acid,data=wine_red,binwidth=0.1) +
  ggtitle('Histogram of citric.acid')
```



上图可以看出，citric.acid 的分布非常集中，仅有少量值比其它略高，最高值可能为异常值。

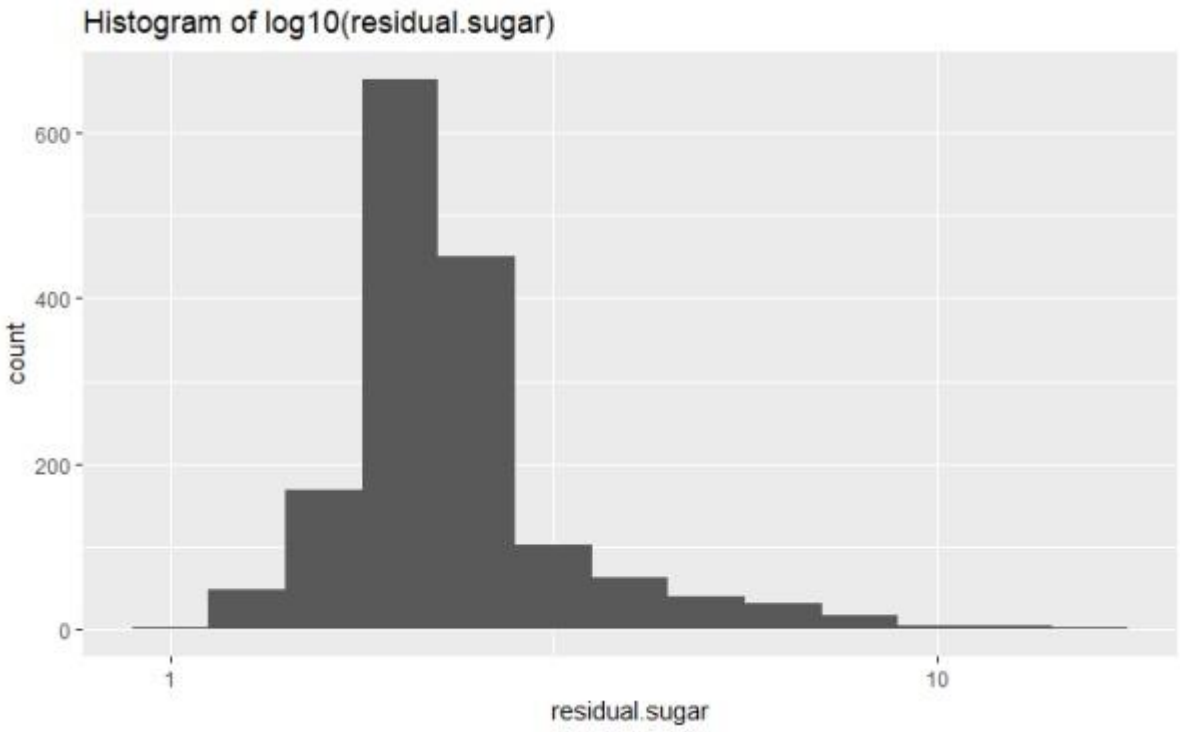
5.查看 residual.sugar 的分布：

```
qplot(x=residual.sugar,data=wine_red,binwidth=1) +
  ggtitle('Histogram of residual.sugar') +
  scale_x_continuous(breaks = seq(0,16,2))
```



上图中残糖呈长尾分布，对其取对数进行收敛：

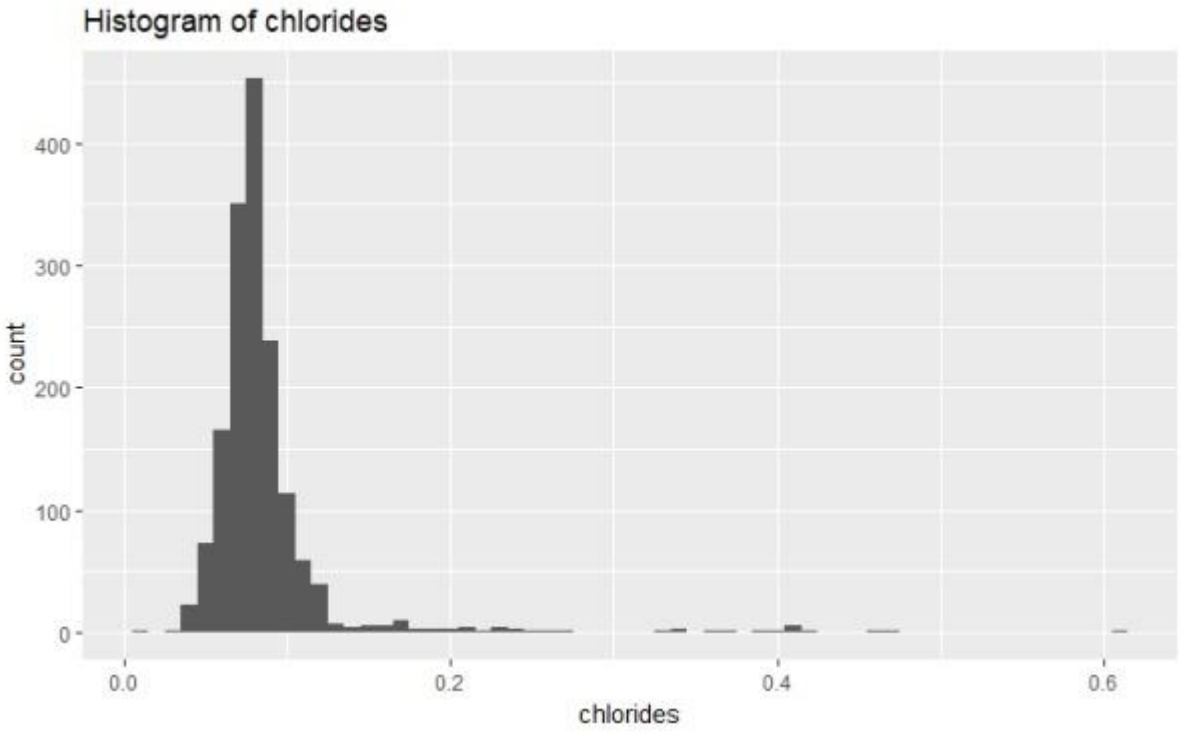
```
qplot(x=residual.sugar,data=wine_red,binwidth=0.1) +
  scale_x_log10() +
  ggtitle('Histogram of log10(residual.sugar)')
```



收敛后的图呈右偏的正态分布。

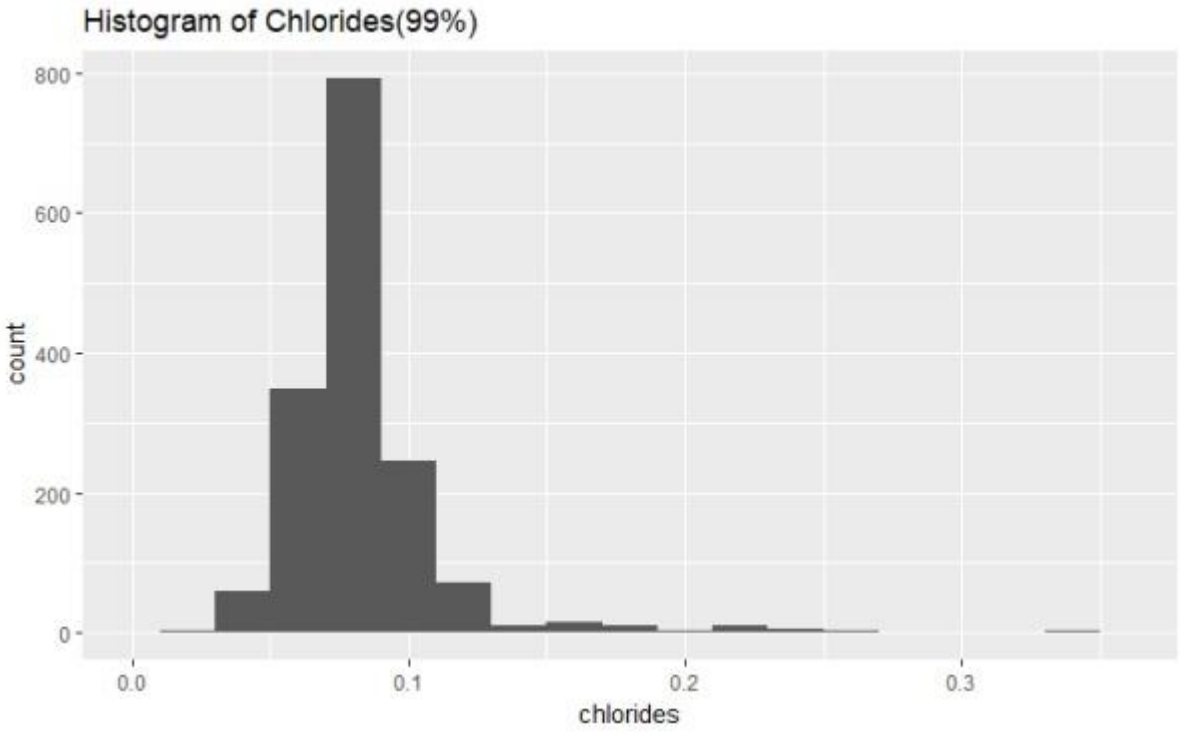
6.查看 chlorides 的分布：

```
qplot(x=chlorides,data=wine_red,binwidth=0.01) +
  ggtitle('Histogram of chlorides')
```



chlorides 的大多数值分布集中，少量值严重右偏，不知是不是数据错误，取其前 99%的值:

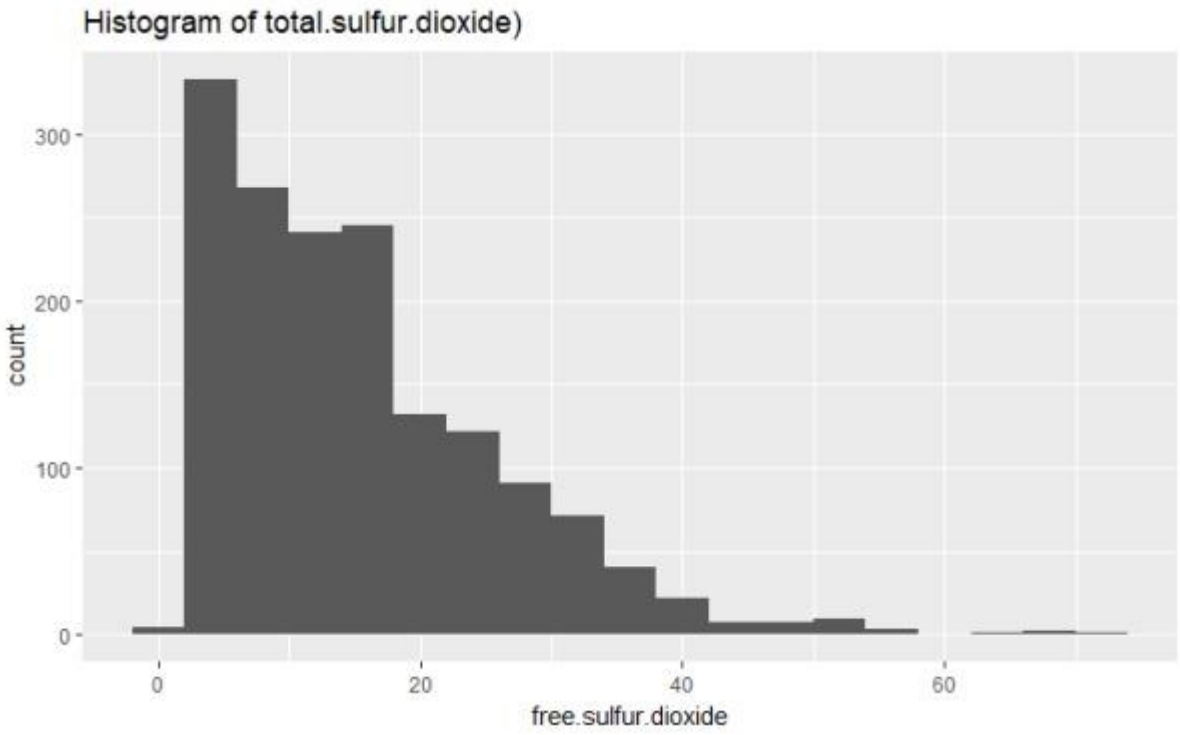
```
qplot(x=chlorides,data=wine_red,binwidth=0.02) +
  xlim(0,quantile(wine_red$chlorides, .99)) +
  ggtitle('Histogram of Chlorides(99%)')
```



转换后呈右偏正态分布。

7.查看 total.sulfur.dioxide 的分布形态：

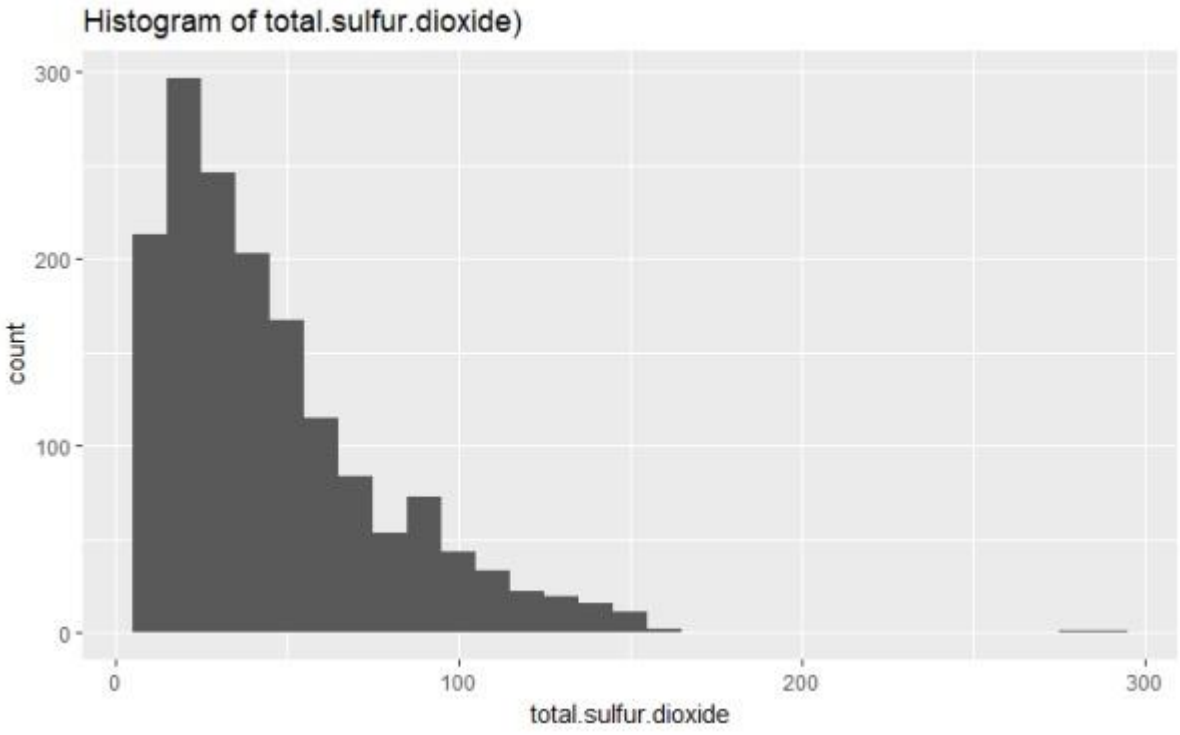
```
qplot(x=free.sulfur.dioxide,data=wine_red,binwidth=4) +
  ggtitle('Histogram of total.sulfur.dioxide)')
```



total.sulfur.dioxide 大部分值在 0~30,分布右偏，算不上钟形分布。

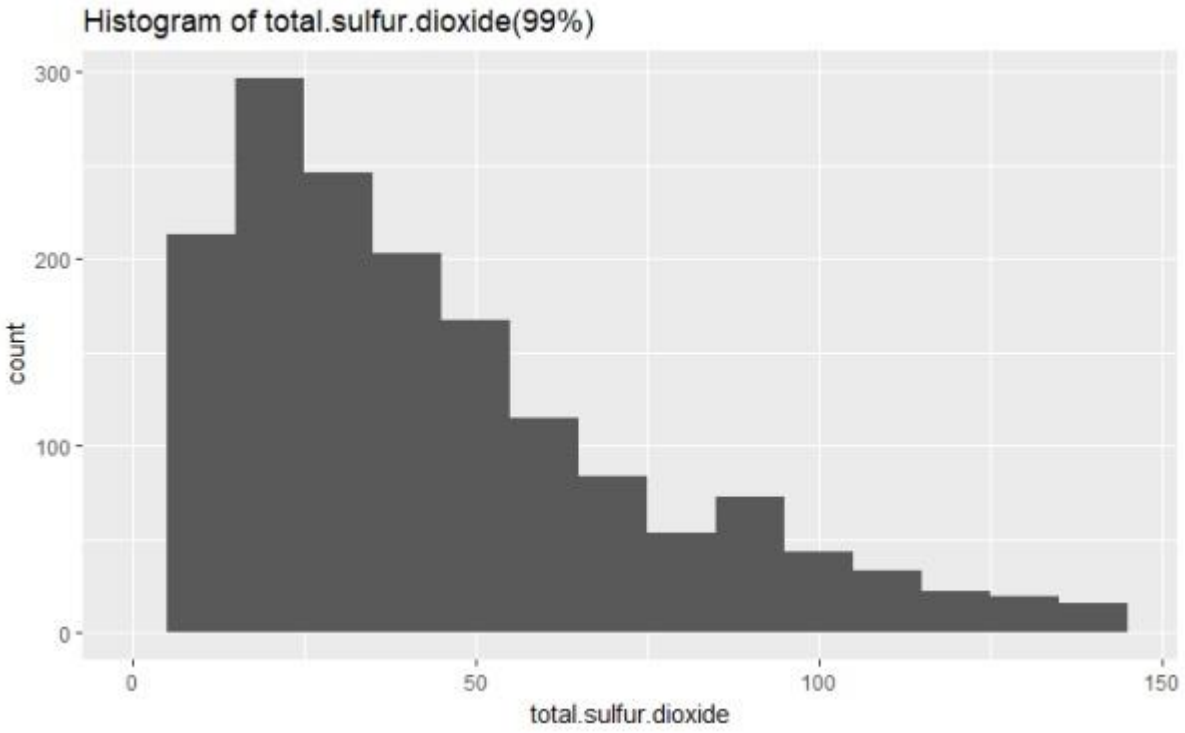
8.查看 total.sulfur.dioxide 的分布：

```
qplot(x=total.sulfur.dioxide,data=wine_red,binwidth=10) +
  ggtitle('Histogram of total.sulfur.dioxide)')
```



上面 total.sulfur.dioxide 存在异常值，取其前 99%的数绘制：

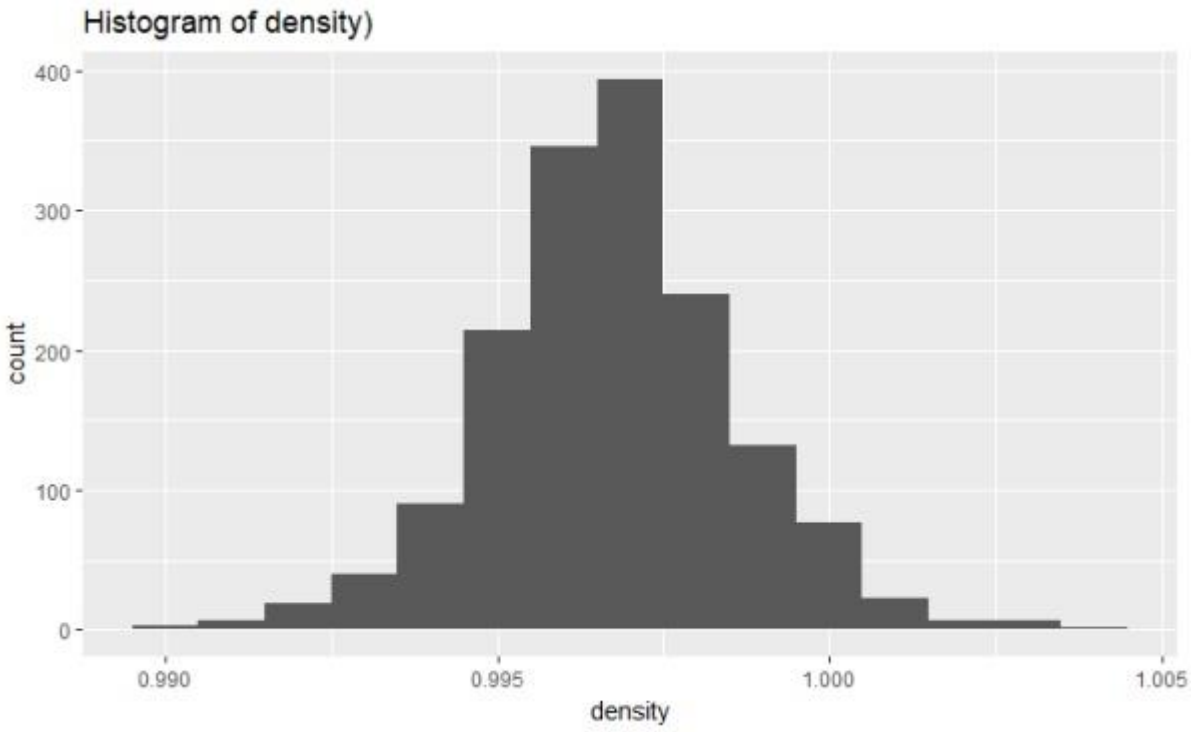
```
qplot(x=total.sulfur.dioxide,data=wine_red,binwidth=10) +
  xlim(0,quantile(wine_red$total.sulfur.dioxide,.99)) +
  ggtitle('Histogram of total.sulfur.dioxide(99%)')
```



去掉异常值的 total.sulfur.dioxide 呈长尾分布。

9.查看 density 的分布：

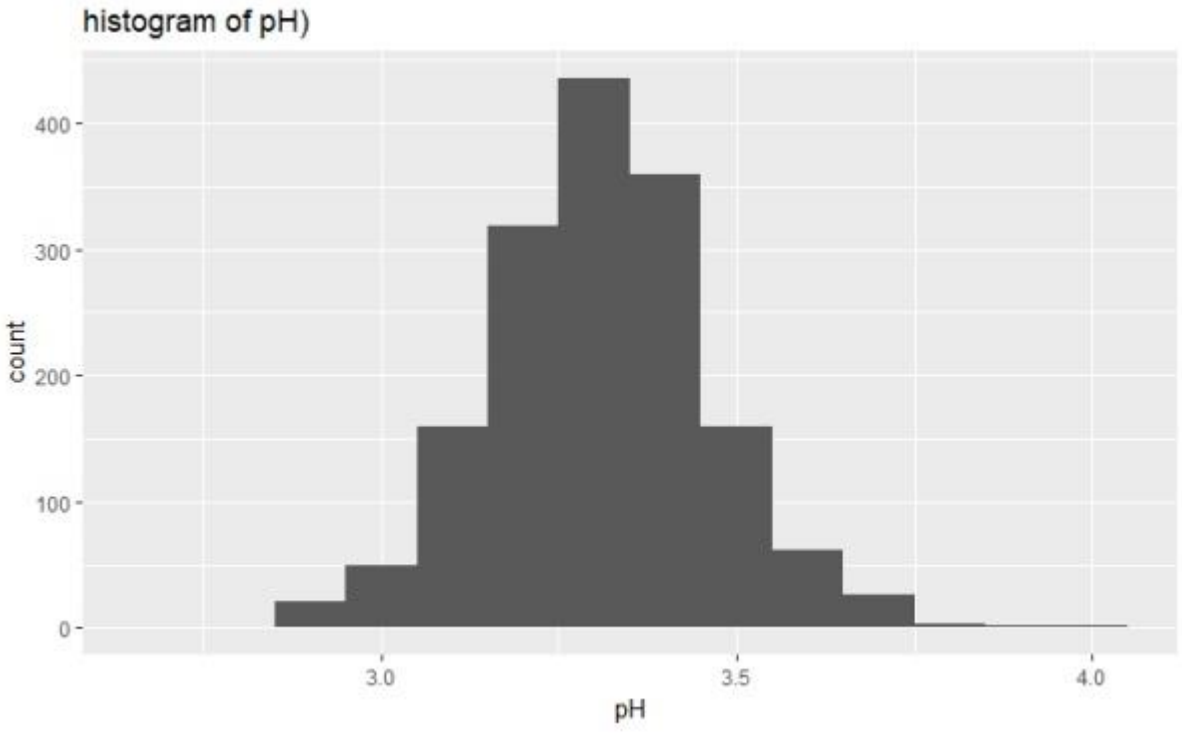
```
qplot(x=density,data=wine_red,binwidth=0.001) +
  ggtitle('Histogram of density)')
```



density 基本呈标准正态分布。

10.查看 PH 分布：

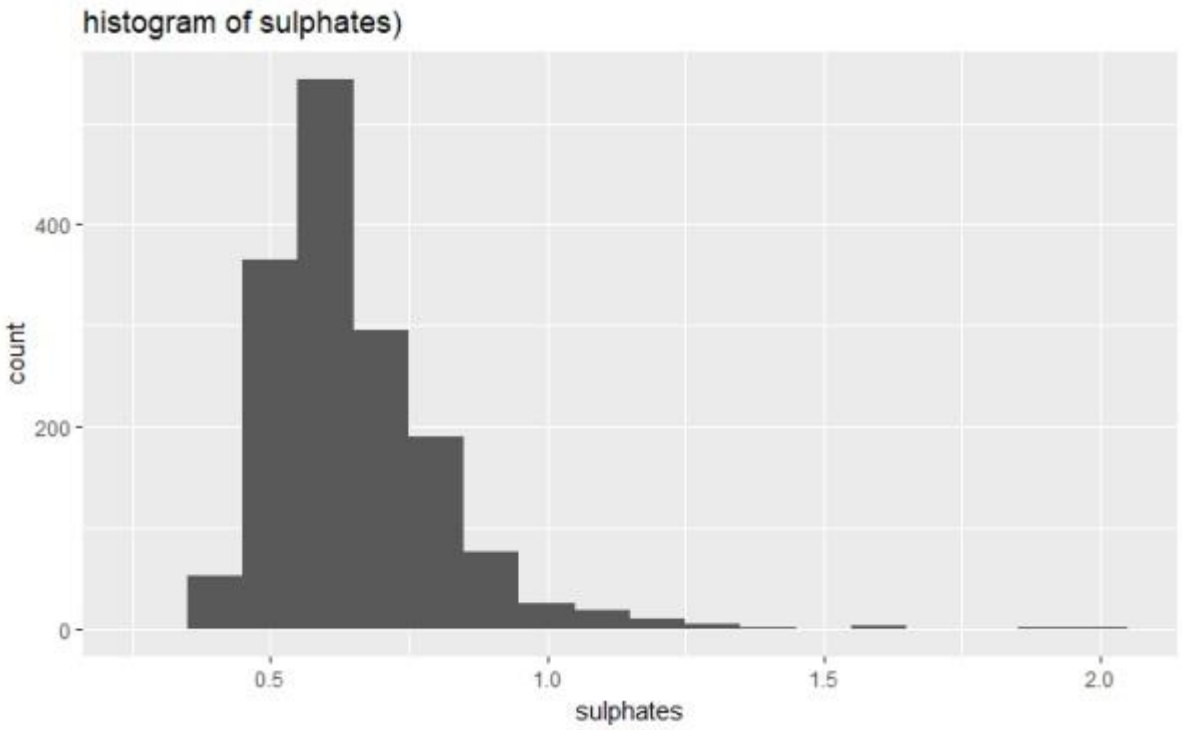
```
qplot(x=pH,data=wine_red,binwidth=0.1) +
  ggtitle('histogram of pH)')
```



PH 也基本呈标准正态分布。

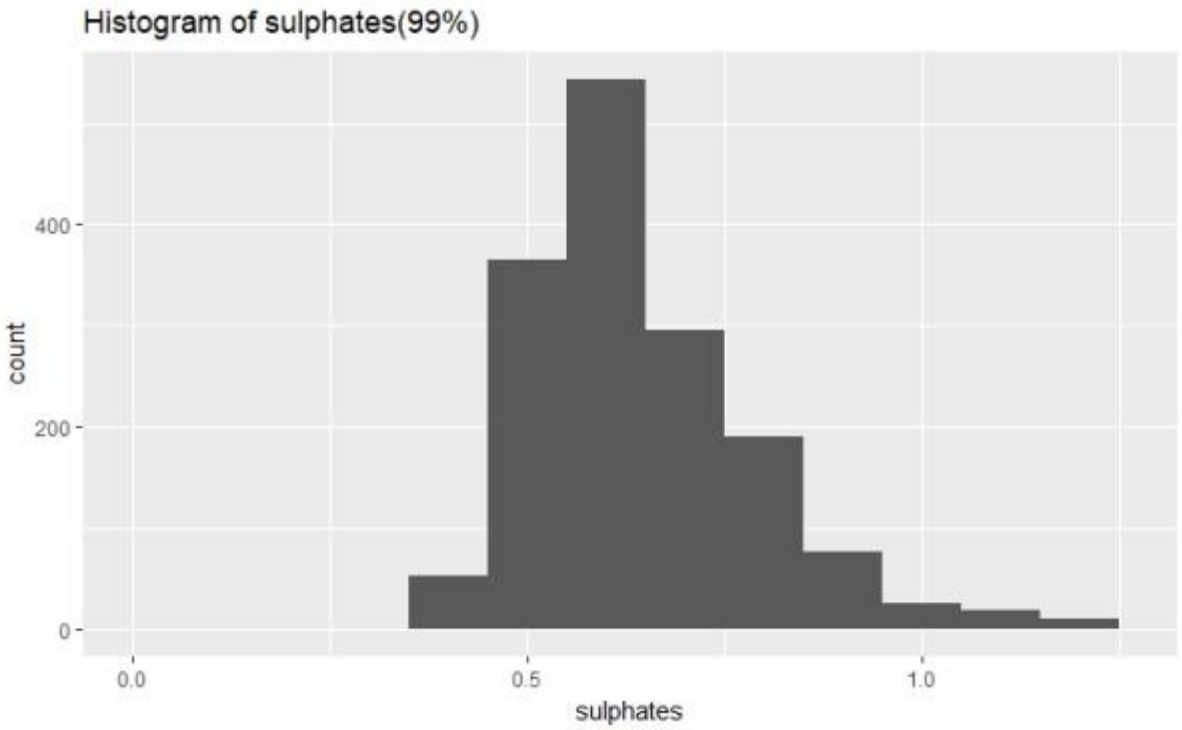
11.查看 sulphates 的分布：

```
qplot(x=sulphates,data=wine_red,binwidth=0.1) +
  ggtitle('histogram of sulphates)')
```



上图可以看出，sulphates 存在异常值，取 sulphates 前 99%的值绘图：

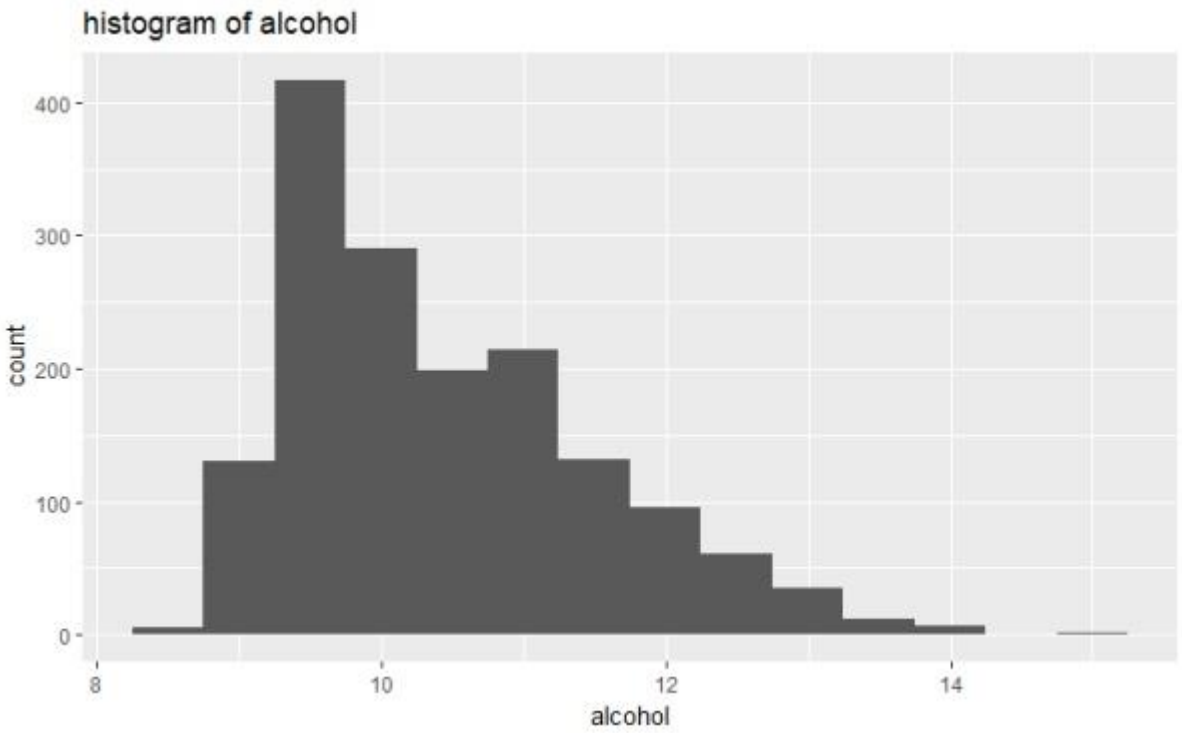
```
qplot(x=sulphates,data=wine_red,binwidth=0.1) +
  xlim(0,quantile(wine_red$sulphates,.99)) +
  ggtitle('Histogram of sulphates(99%)')
```



去掉异常值后 sulphates 呈右偏正态分布，大部分值在 0.5~1 之间。

12.查看 alcohol 的分布形态：

```
qplot(x=alcohol,data=wine_red,binwidth=0.5) +
  ggtitle('histogram of alcohol')
```



alcohol 呈右偏正态分布。

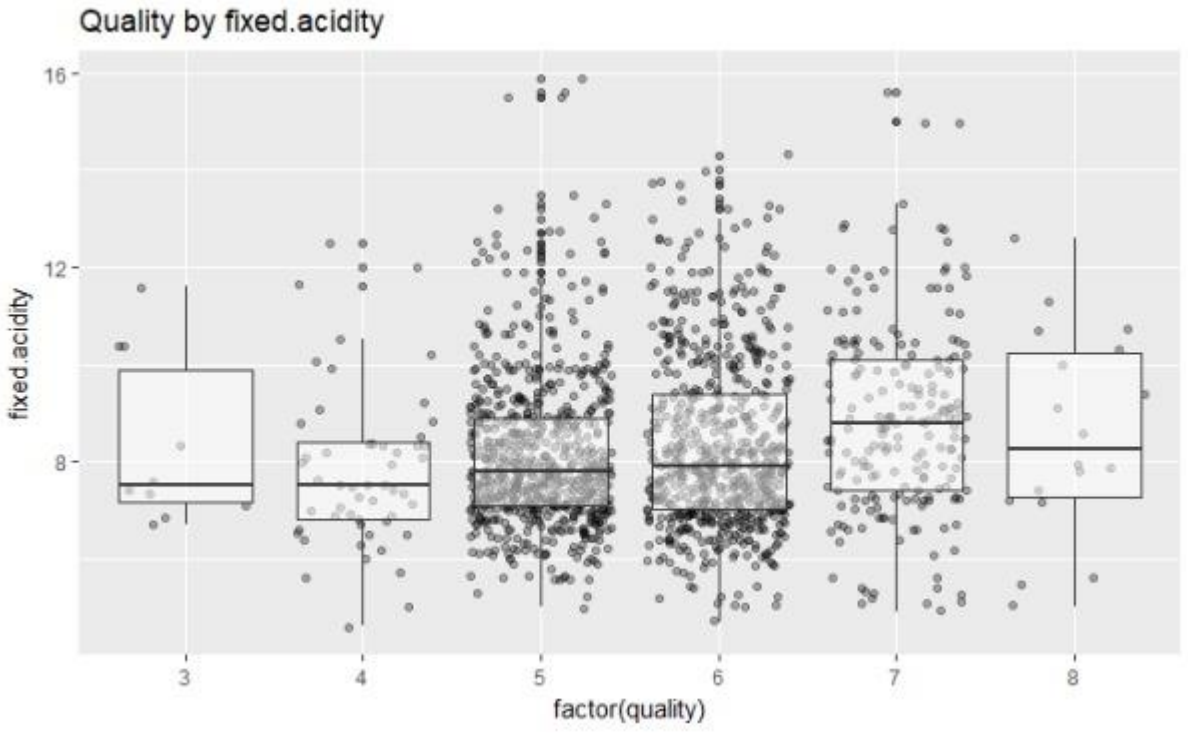
小结

游离酸度有异常值，对其取前 99%的数据；残糖呈现长尾分布，对其用对数转换；氯化物也是长尾分布，对其对数转换后效果不太好；总二氧化硫有异常值，取前 99%的数；硫酸盐也是取前 99%的值。

二、双变量分析

1.查看 fixed.acidity 和 quality 的关系：

```
ggplot(wine_red,aes(factor(quality), fixed.acidity)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality by fixed.acidity')
```



从分布看，quality 值的主要区间中 fixed.acidity 基本均匀分布，两者没有明显关系。

计算 fixed.acidity 与 quality 的相关性系数：

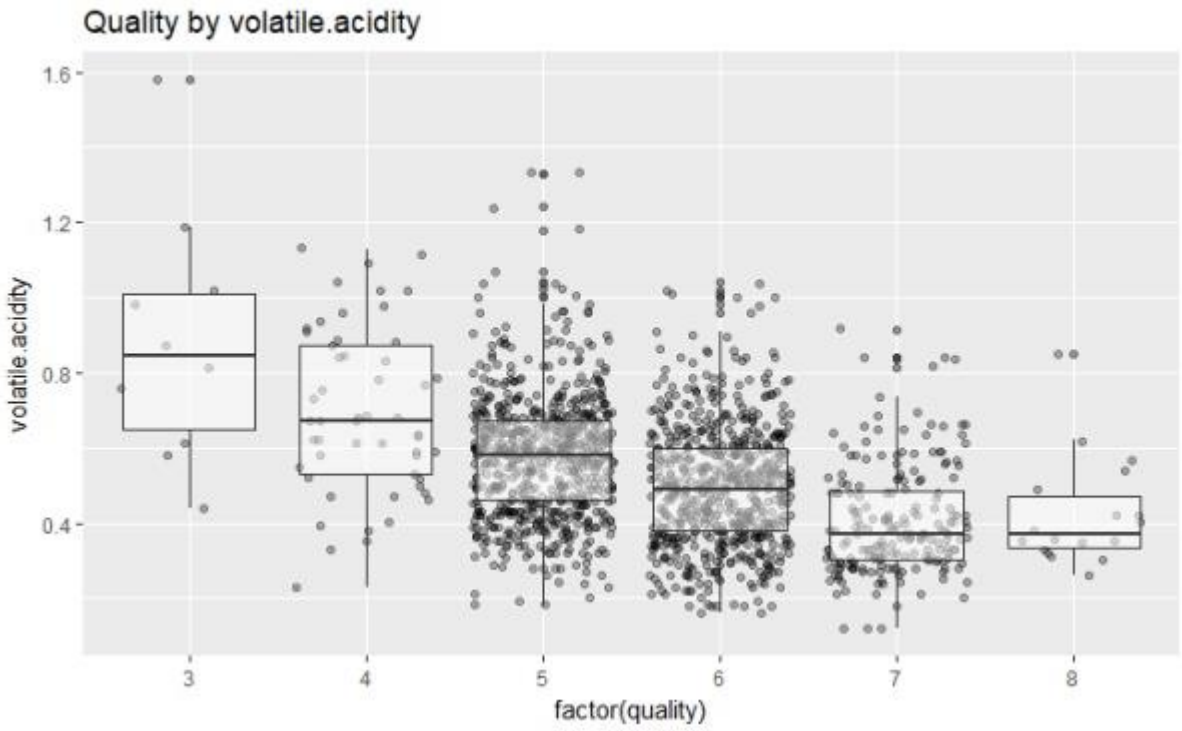
```
with(wine_red,cor.test(fixed.acidity,quality))
```

```
        Pearson's product-moment correlation

data:  fixed.acidity and quality
t = 4.996, df = 1597, p-value = 6.496e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07548957 0.17202667
sample estimates:
      cor
0.1240516
```

2.查看 volatile.acidity 和 quality 的关系：

```
ggplot(wine_red,aes(factor(quality), volatile.acidity)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality by volatile.acidity')
```



从分布看，貌似呈负相关，但不明显。

计算 volatile.acidity 和 quality 的相关性系数：

```
with(wine_red,cor.test(volatile.acidity,quality))
```

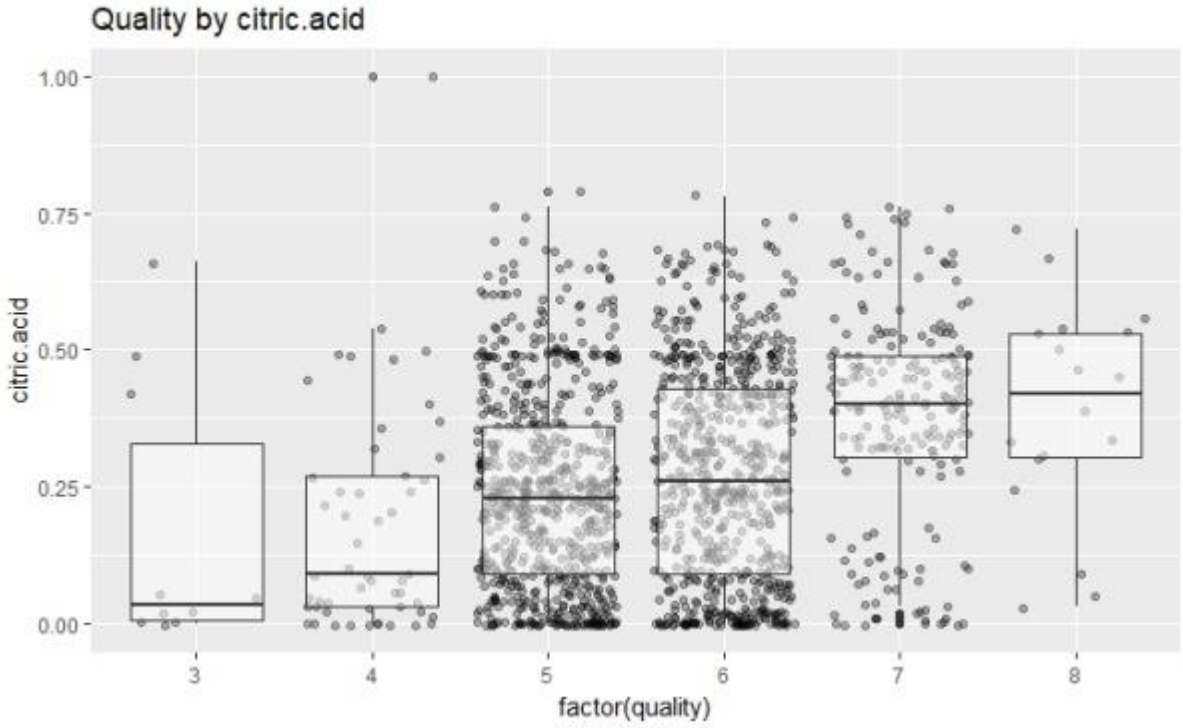
```
        Pearson's product-moment correlation

data:  volatile.acidity and quality
t = -16.954, df = 1597, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4313210 -0.3482032
sample estimates:
       cor
-0.3905578
```

volatile.acidity 和 quality 的相关系数为-0.39，与分布的观察结果一致。

3.查看 citric.acid 和质量的关系：

```
ggplot(wine_red,aes(factor(quality), citric.acid)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality by citric.acid')
```



看不出有明显的相关性。

计算柠檬酸与质量的相关性：

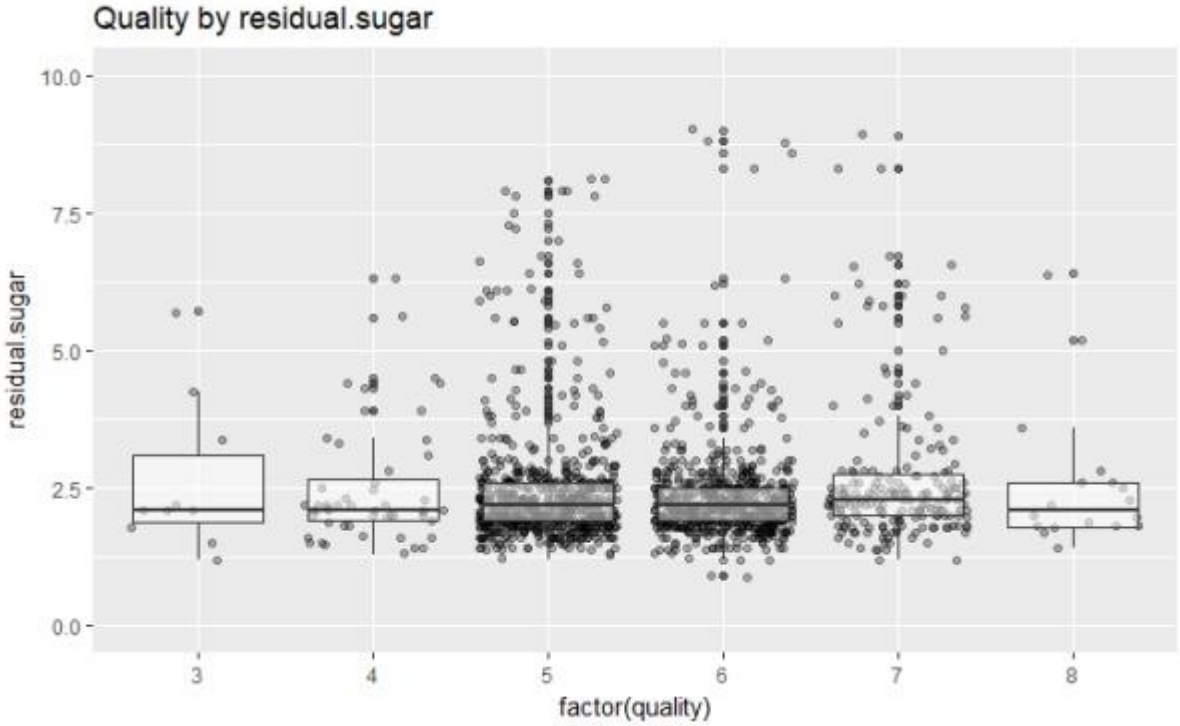
```
with(wine_red,cor.test(citric.acid,quality))
```

```
        Pearson's product-moment correlation

data:  citric.acid and quality
t = 9.2875, df = 1597, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1793415 0.2723711
sample estimates:
      cor
0.2263725
```

4.查看 residual.sugar 和 quality 的相关性：

```
ggplot(wine_red,aes(factor(quality), residual.sugar)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ylim(0,10) +
  ggtitle('Quality by residual.sugar')
```



残糖主要分布在质量 5/6 分上，没有相关性。

计算残糖与质量的相关性：

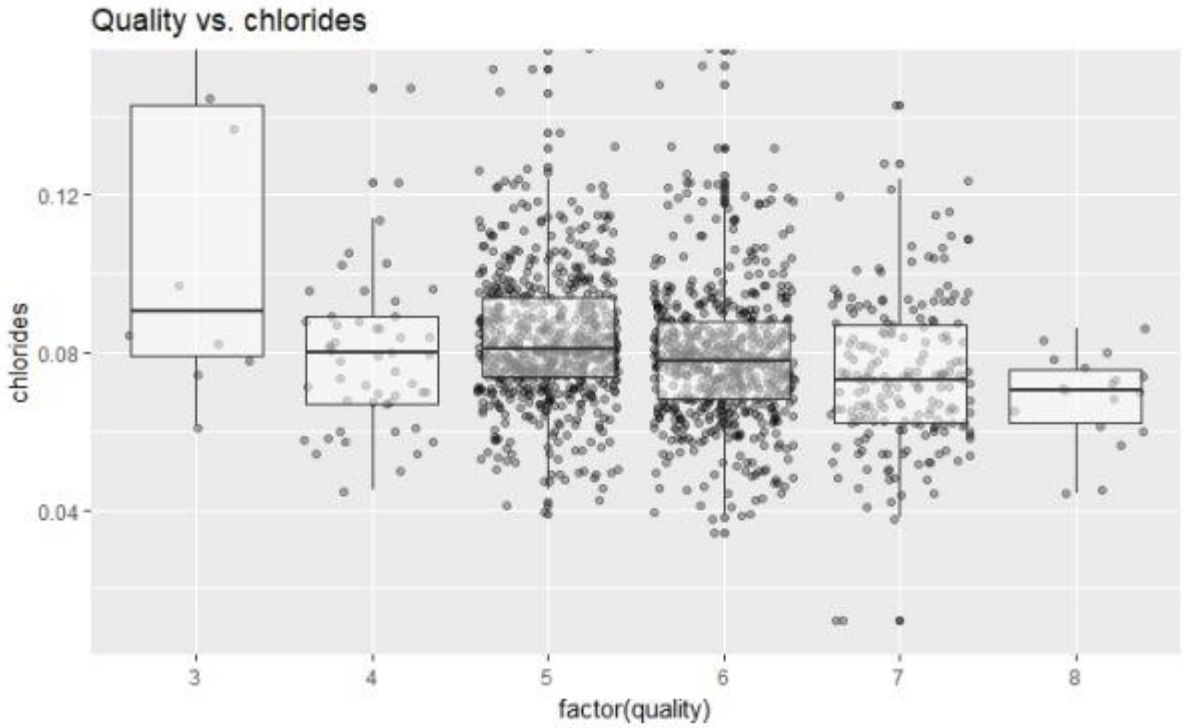
```
with(wine_red,cor.test(residual.sugar,quality))
```

```
        Pearson's product-moment correlation

data:  residual.sugar and quality
t = 0.5488, df = 1597, p-value = 0.5832
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03531327  0.06271056
sample estimates:
       cor
0.01373164
```

5.查看 Quality 和 chlorides 的相关性：

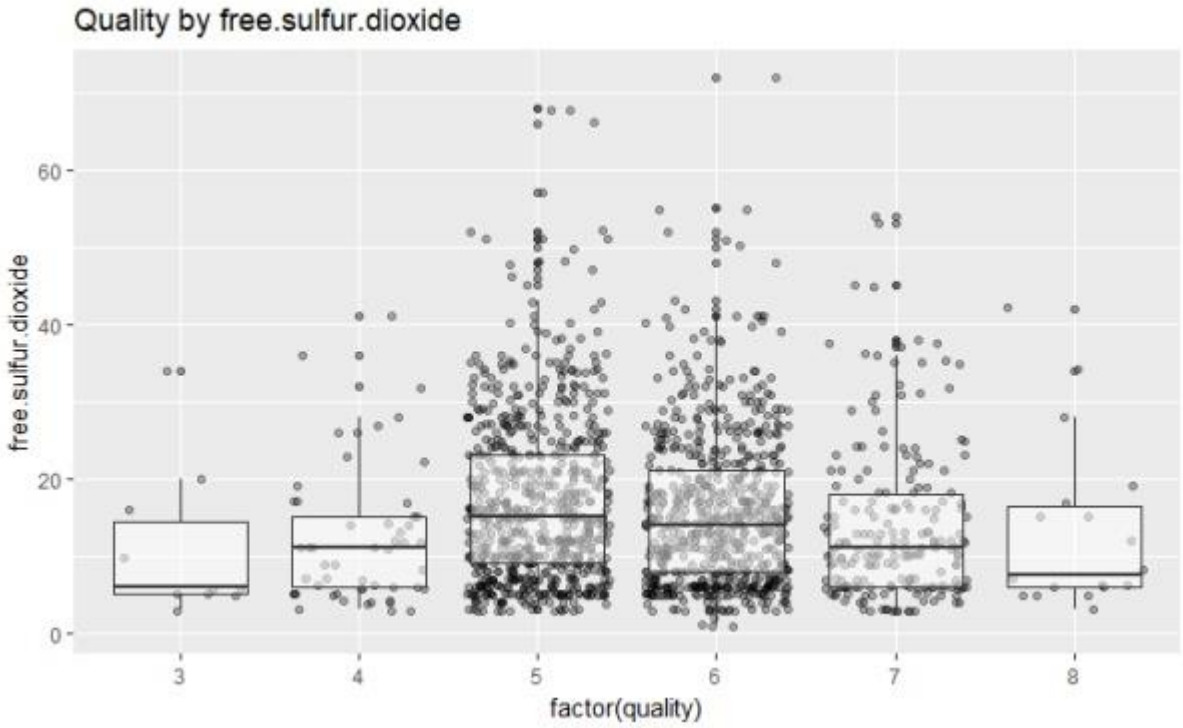
```
ggplot(wine_red,aes(factor(quality), chlorides)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  coord_cartesian(ylim = c(0.01, 0.15)) +
  ggtitle('Quality vs. chlorides')
```



看不出有相关性。

6.查看 free.sulfur.dioxide 和 quality 的相关性：

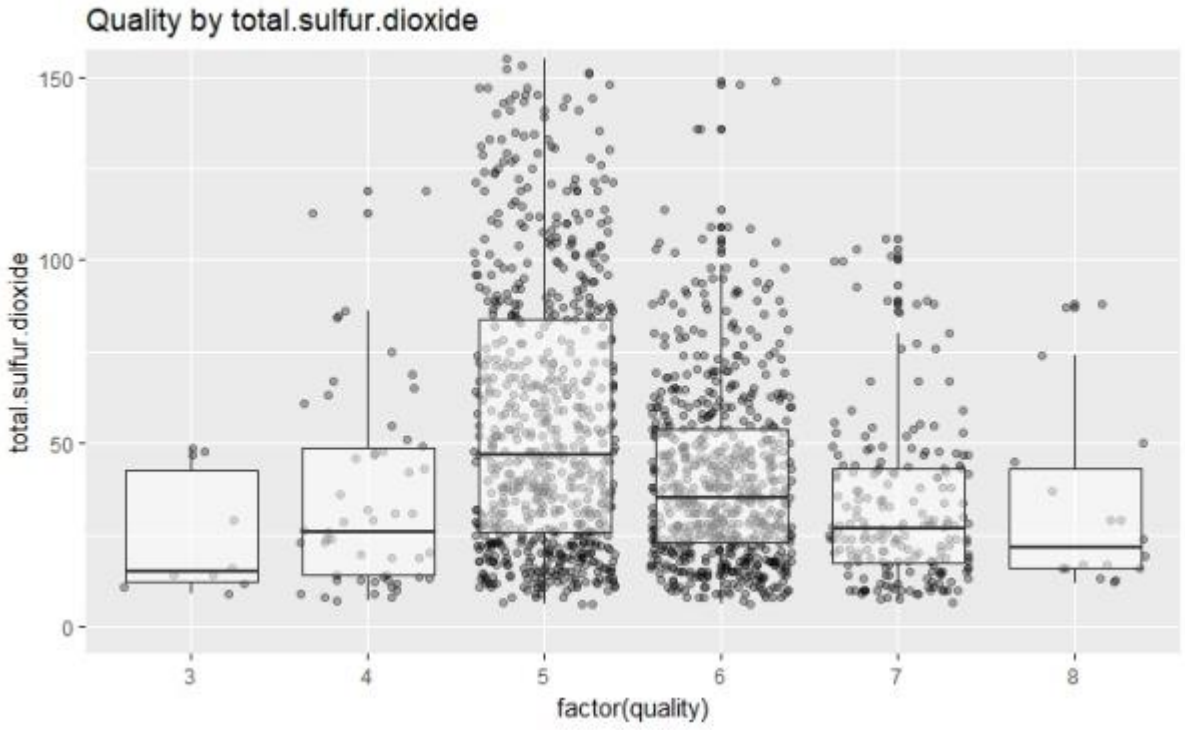
```
ggplot(wine_red,aes(factor(quality), free.sulfur.dioxide)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality by free.sulfur.dioxide')
```



明显没有相关性。

7.查看 total.sulfur.dioxide 和 quality 的相关性：

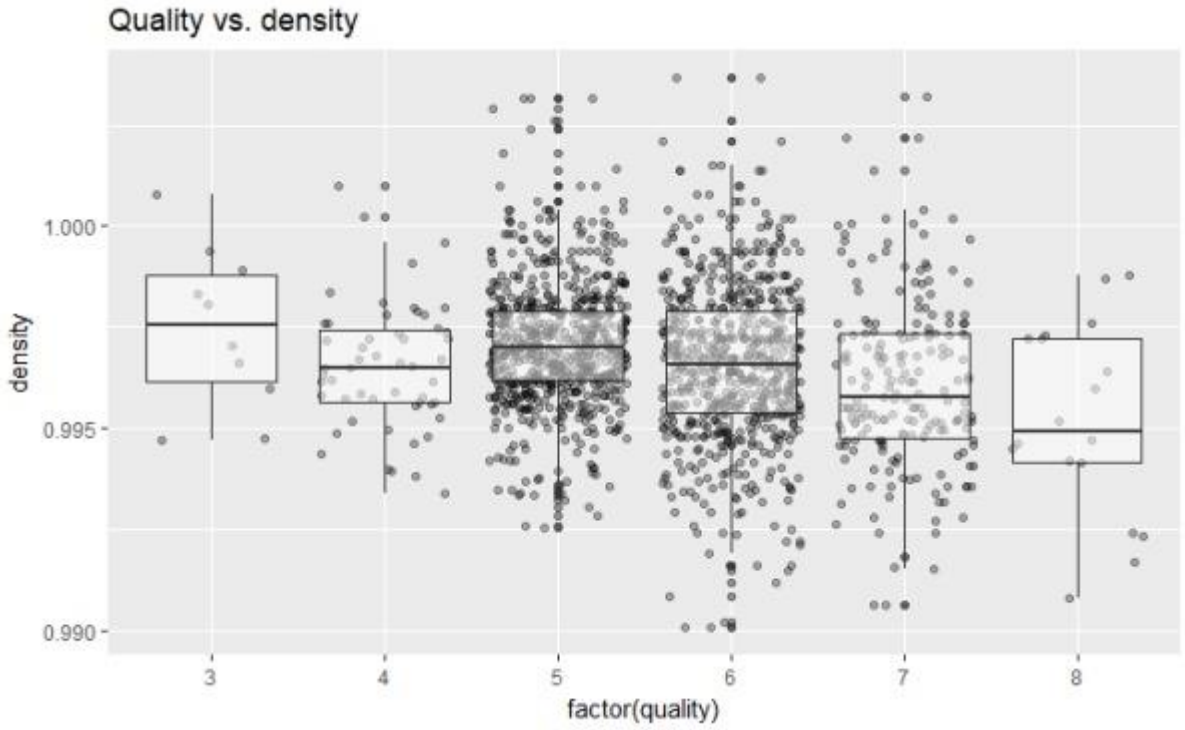
```
ggplot(wine_red,aes(factor(quality), total.sulfur.dioxide)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  coord_cartesian(ylim = c(0, 150)) +
  ggtitle('Quality by total.sulfur.dioxide')
```



total.sulfur.dioxide 主要分布在质量评分 5~7 分上面，与质量没有明显的相关性。

8.查看 density 和 quality 之间的关系：

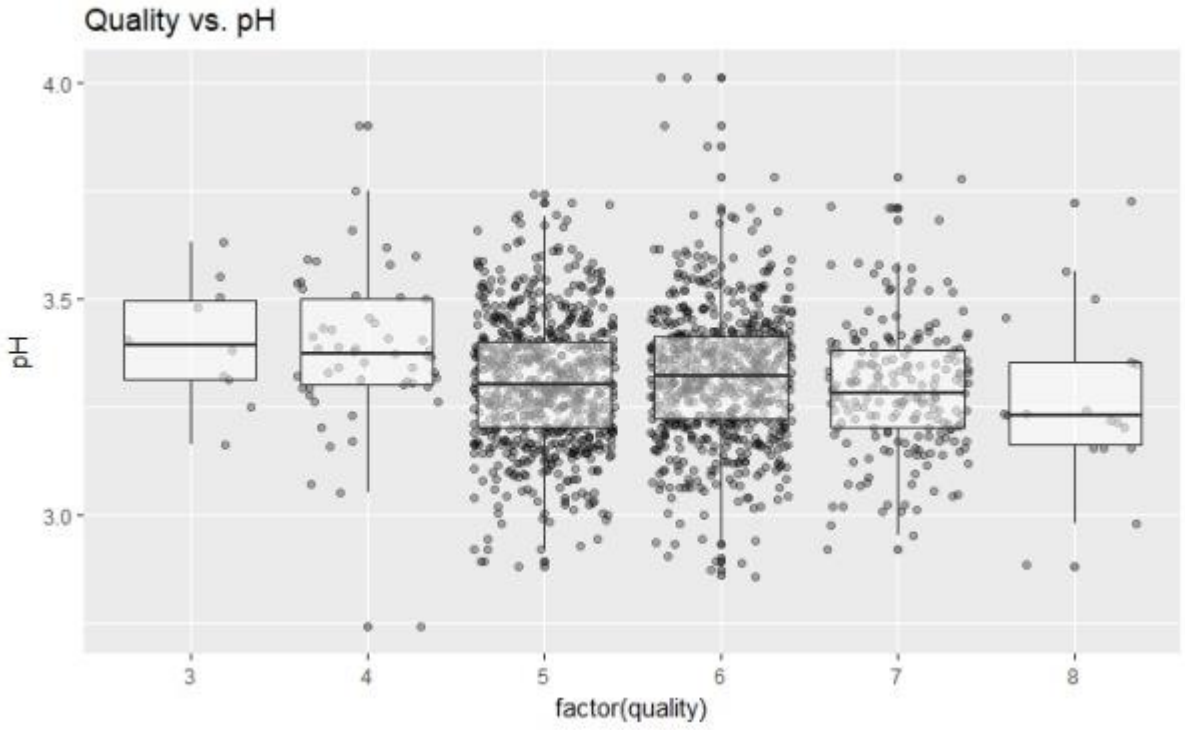
```
ggplot(wine_red,aes(factor(quality), density)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality vs. density')
```



密度与质量也没有明显的相关性，由于质量多集中于 5~6 分，散点也多位于此范围，密度基本分布均匀.

9.查看 pH 和 quality 的关系：

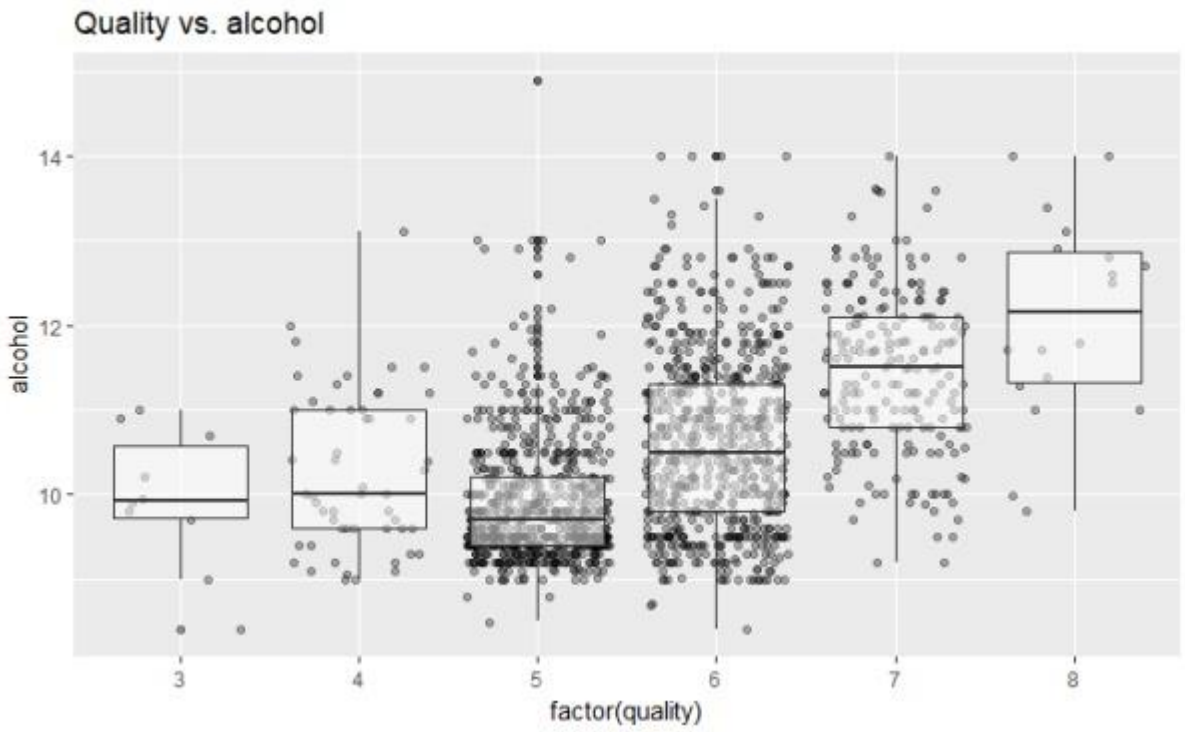
```
ggplot(wine_red,aes(factor(quality), pH)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality vs. pH')
```



看不出什么相关性.

10.查看 alcohol 和 quality 的相关性：

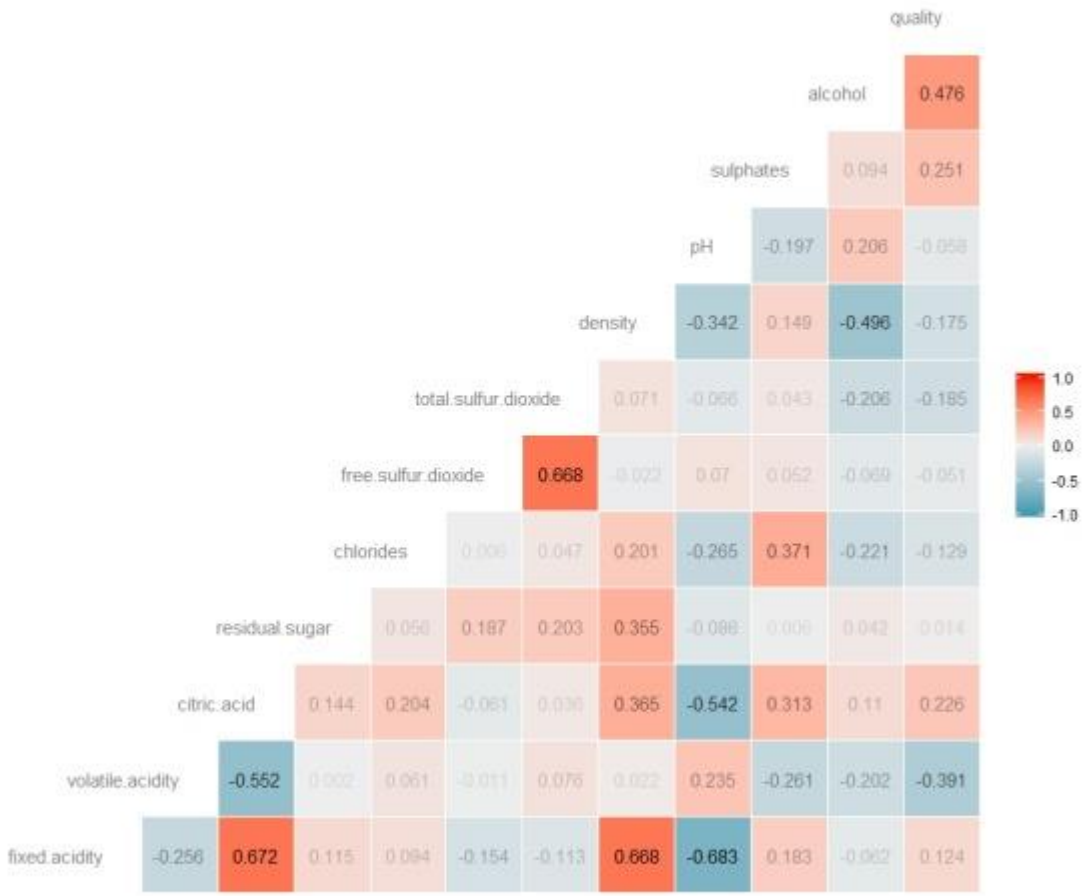
```
ggplot(wine_red,aes(factor(quality), alcohol)) +
  geom_jitter( alpha = 0.3)  +
  geom_boxplot( alpha = 0.5) +
  ggtitle('Quality vs. alcohol')
```



看起来像正相关。

11.绘制相关系数图：

```
ggcorr(data = wine_red, hjust = 1, size = 4, color = "grey50", layout.exp = 2,
       label = TRUE, label_size = 4, label_round = 3, label_alpha = TRUE)
```



小结

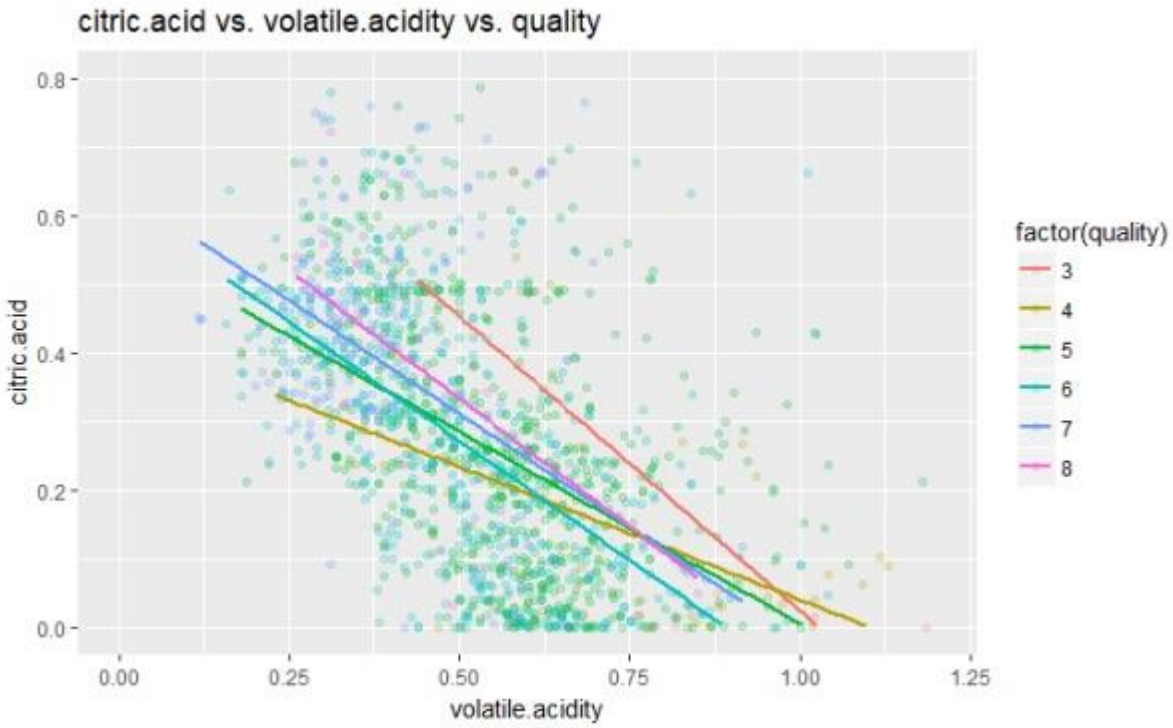
质量与游离酸度及酒精度相关，而游离酸度又与柠檬酸相关，酒精度和密度相关；固定酸度和 PH 相关性很强，达到-0.683。

三、多变量分析

1.查看 citric.acid， volatile.acidity， quality 间的关系：

```
ggplot(wine_red,aes(volatile.acidity, citric.acid,color=factor(quality))) +
  geom_jitter(alpha=0.2) +
```

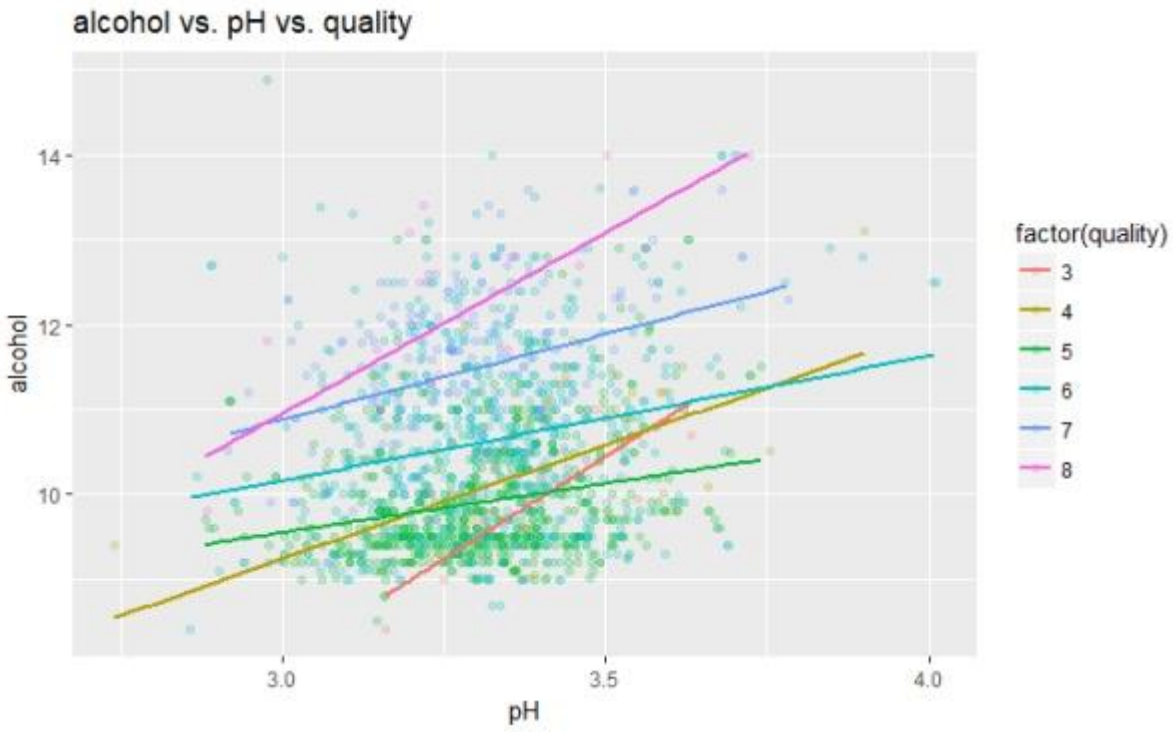
```
  ylim(0, 0.8) +
  xlim(0,1.2) +
  geom_smooth(method = "lm", se = FALSE,size=1) +
  ggtitle('citric.acid vs. volatile.acidity vs. quality')
```



不同的质量下，citric.acid 和 volatile.acidity 成负相关，相关度为-0.55。

2.查看 alcohol， pH， quality 变量间的关系：

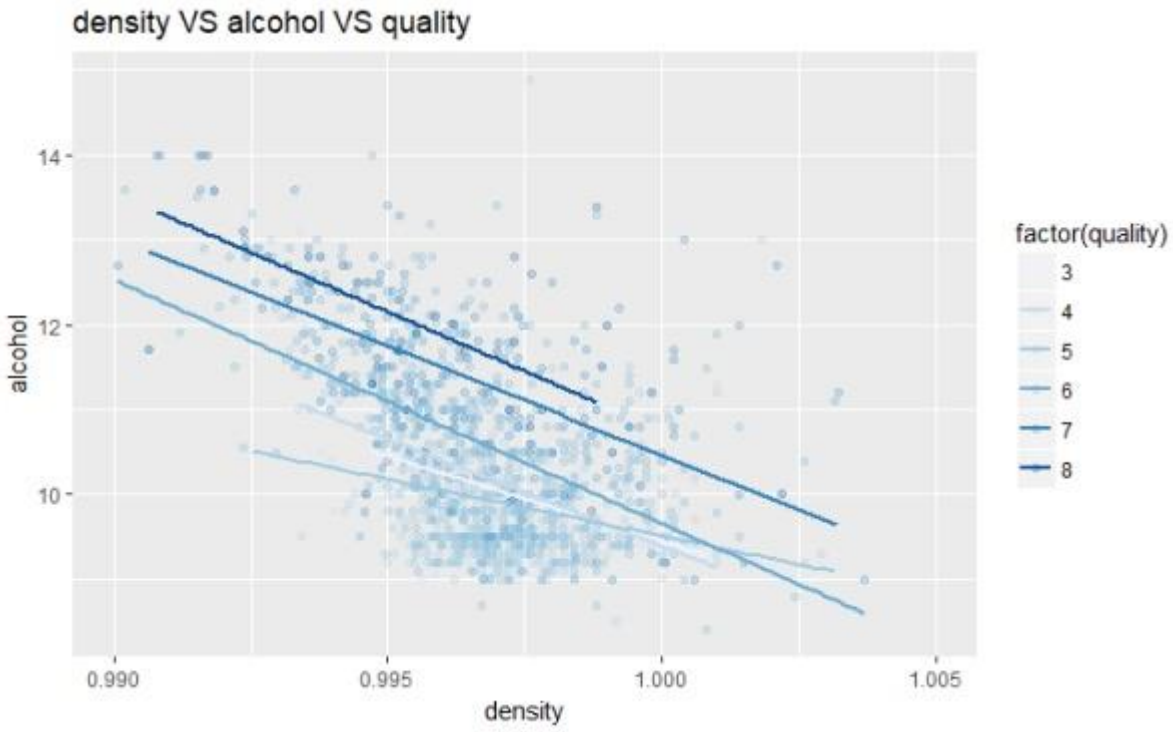
```
ggplot(wine_red,aes(pH, alcohol, color=factor(quality))) +
  geom_jitter(alpha=0.2) +
  geom_smooth(method = "lm",se = FALSE) +
  ggtitle('alcohol vs. pH vs. quality')
```



不同质量下，alcohol 和 pH 成正相关，相关度 0.2，在不同质量下的相关度变化较大。

3.查看 density，alcohol，quality 间的关系：

```
ggplot(aes(x = density, y = alcohol, color = factor(quality)), data = wine_red) +
  geom_jitter(alpha = 0.2) +
  scale_color_brewer(palette = "Blues") +
  geom_smooth(method = "lm", se = FALSE,size=1) +
  xlim(0.99, 1.005) +
  ggtitle("density VS alcohol VS quality")
```



不同质量下，density 和 alcohol 呈负相关，相关系数-0.5。

小结

酒精度高的质量高，与 pH 关系不大。同一游离酸度下，柠檬酸高的质量低。

建模

```
library(memisc)

m1 <- lm(I(quality) ~ I(alcohol), data=wine_red)
m2 <- update(m1, ~ . + volatile.acidity)
m3 <- update(m2, ~ . + density)
m4 <- update(m3, ~ . + citric.acid)
mtable(m1,m2,m3,m4)
```

```
Calls:
m1: lm(formula = I(quality) ~ I(alcohol), data = wine_red)
m2: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity, data = wine_red)
m3: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density,
    data = wine_red)
m4: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density +
    citric.acid, data = wine_red)

================================================================================
                          m1            m2            m3            m4
--------------------------------------------------------------------------------
  (Intercept)          1.875***      3.095***    -18.407       -21.552
                      (0.175)       (0.184)      (10.298)      (12.039)
  I(alcohol)           0.361***      0.314***      0.333***      0.336***
                      (0.017)       (0.016)       (0.018)       (0.019)
  volatile.acidity                  -1.384***     -1.365***     -1.399***
                                    (0.095)       (0.096)       (0.117)
  density                                         21.360*       24.520*
                                                  (10.228)      (11.994)
  citric.acid                                                    -0.061
                                                                 (0.121)
--------------------------------------------------------------------------------
  R-squared             0.227         0.317         0.319         0.319
  adj. R-squared        0.226         0.316         0.318         0.317
  sigma                 0.710         0.668         0.667         0.667
  F                   468.267       370.379       248.893       186.646
  p                     0.000         0.000         0.000         0.000
  Log-likelihood    -1721.057     -1621.814     -1619.631     -1619.503
  Deviance            805.870       711.796       709.855       709.742
  AIC                3448.114      3251.628      3249.261      3251.006
  BIC                3464.245      3273.136      3276.147      3283.269
  N                  1599          1599          1599          1599
```

结论

酒精度与质量成正相关；游离酸度与质量成明显负相关。

思考•总结

对于葡萄酒，个人不是很了解，不像钻石那样，大家都知道越重越漂亮的越贵，所以对于葡萄酒质量影响因素的分析，只能挨个全部分析，对于单变量的分析，有的变量存在异常值，无法判断其是否错误，或者只是反常而已，因此只能舍弃，另外由于要取不同的组宽，不方便用循环进行，因此无奈的做了大量重复工作。

对于双变量分析，我个人不喜欢酒，只喜欢奶茶奶糖，对于红酒，我觉得微酸，较甜，不苦是比较好的，可品酒师显然不这么认为，所以只能继续挨个探索，最终发现质量与游离酸度及酒精度有明显关系，而游

离酸度又与柠檬酸明显相关，酒精度和密度明显相关，综上，质量可能与游离酸度，柠檬酸，酒精度，及密度相关。另外发现，固定酸度和 PH 相关性是-0.683，算是强相关。

对于多变量分析，主要探索不同质量等级下其它强相关变量的相关度变化，发现酒精与 PH 在不同质量下的相关度变化较大。最后此数据集由于数据量略小，分析得出结论的过程不是很明显，另外变量都是数值型，没有分类变量，在多变量分析时数据略显得鸡肋。

对于此数据集，还可进一步探索存在异常值的变量在消除异常值或对变量进行对数转换后，再进行双变量和多变量的分析。