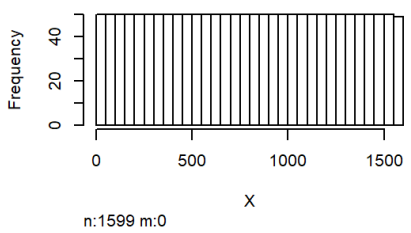
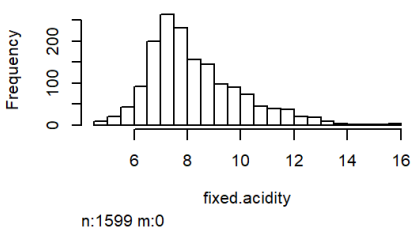
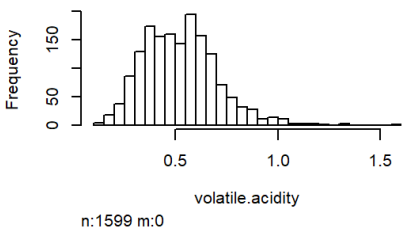
红酒数据集探索性数据分析

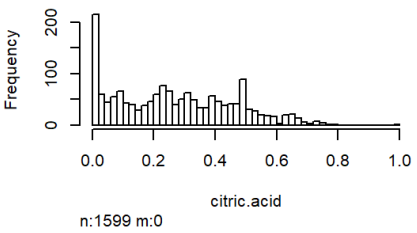
单变量绘图选择

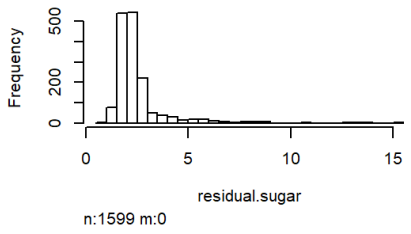
数据可视化

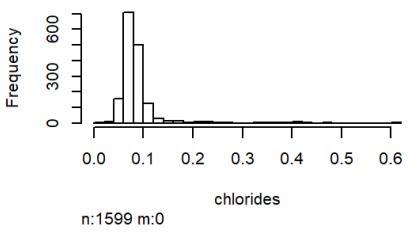

Frequency
0
20
40
0
500
1000
1500
X
n:1599 m:0

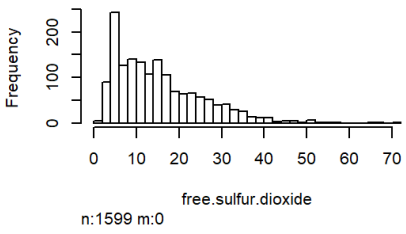

Frequency
0
100
200
6
8
10
12
14
16
fixed.acidity
n:1599 m:0

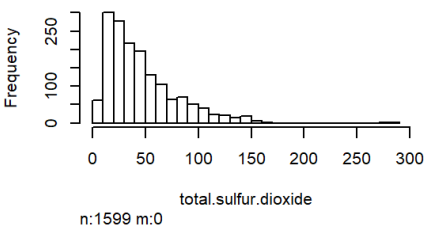

Frequency
0
50
150
0.5
1.0
1.5
volatile.acidity
n:1599 m:0

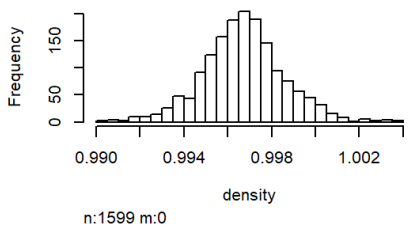

Frequency
0
100
200
0.0
0.2
0.4
0.6
0.8
1.0
citric.acid
n:1599 m:0

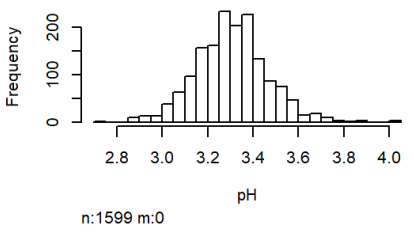

Frequency
0
200
500
0
5
10
15
residual.sugar
n:1599 m:0

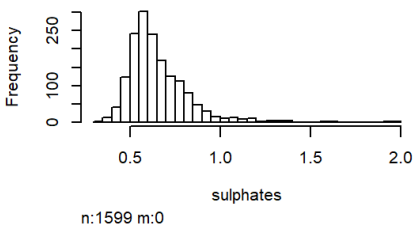

Frequency
0
300
600
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides
n:1599 m:0

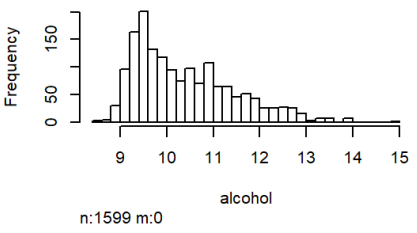

Frequency
0
100
200
0
10
20
30
40
50
60
70
free.sulfur.dioxide
n:1599 m:0

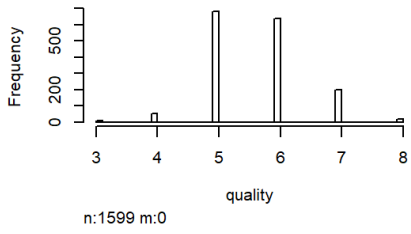

Frequency
0
100
250
0
50
100
150
200
250
300
total.sulfur.dioxide
n:1599 m:0


Frequency
0
50
150
0.990
0.994
0.998
1.002
density
n:1599 m:0


Frequency
0
100
200
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH
n:1599 m:0


Frequency
0
100
250
0.5
1.0
1.5
2.0
sulphates
n:1599 m:0

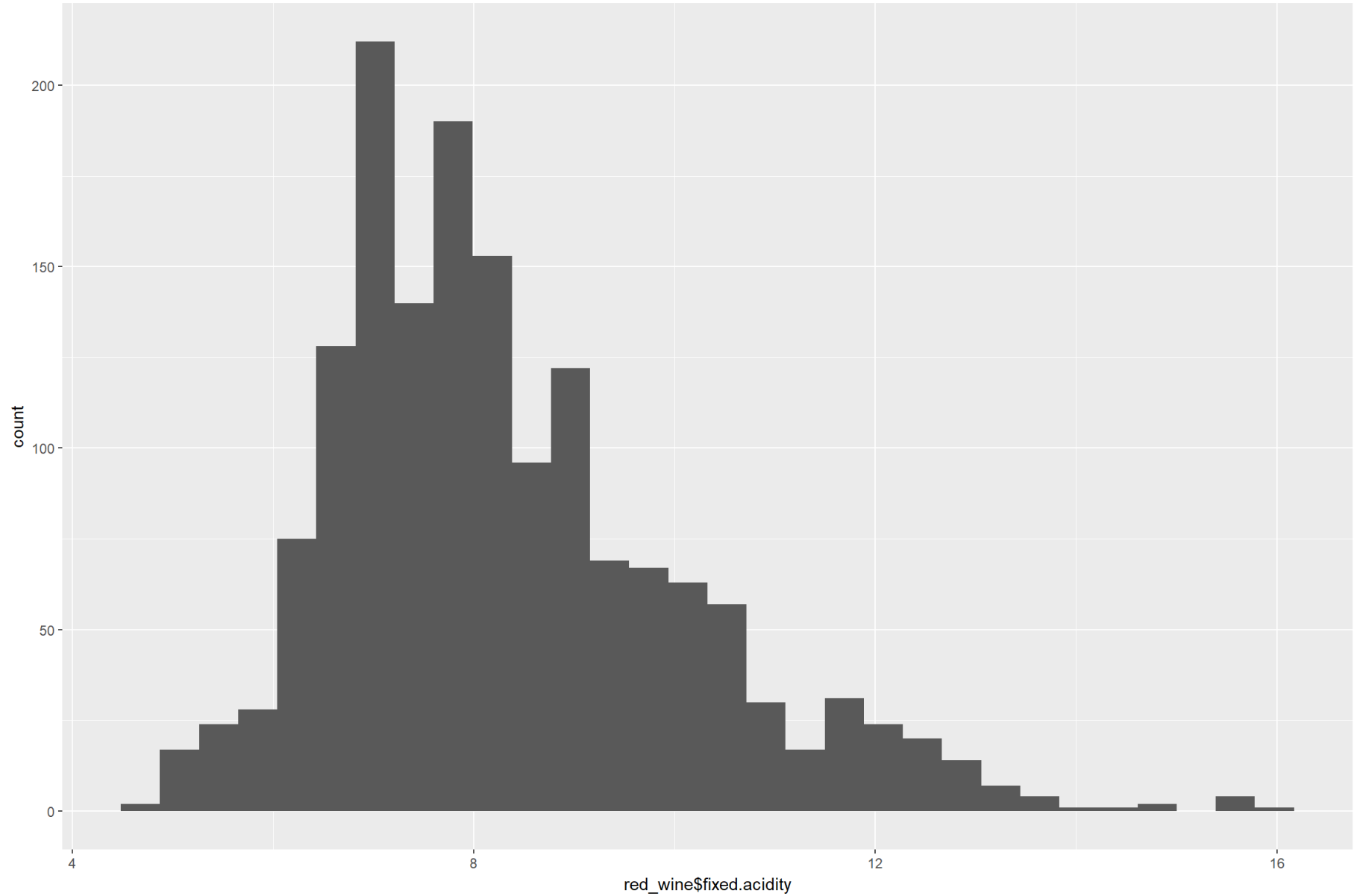

Frequency
0
50
150
9
10
11
12
13
14
15
alcohol
n:1599 m:0


Frequency
0
200
500
3
4
5
6
7
8
quality
n:1599 m:0

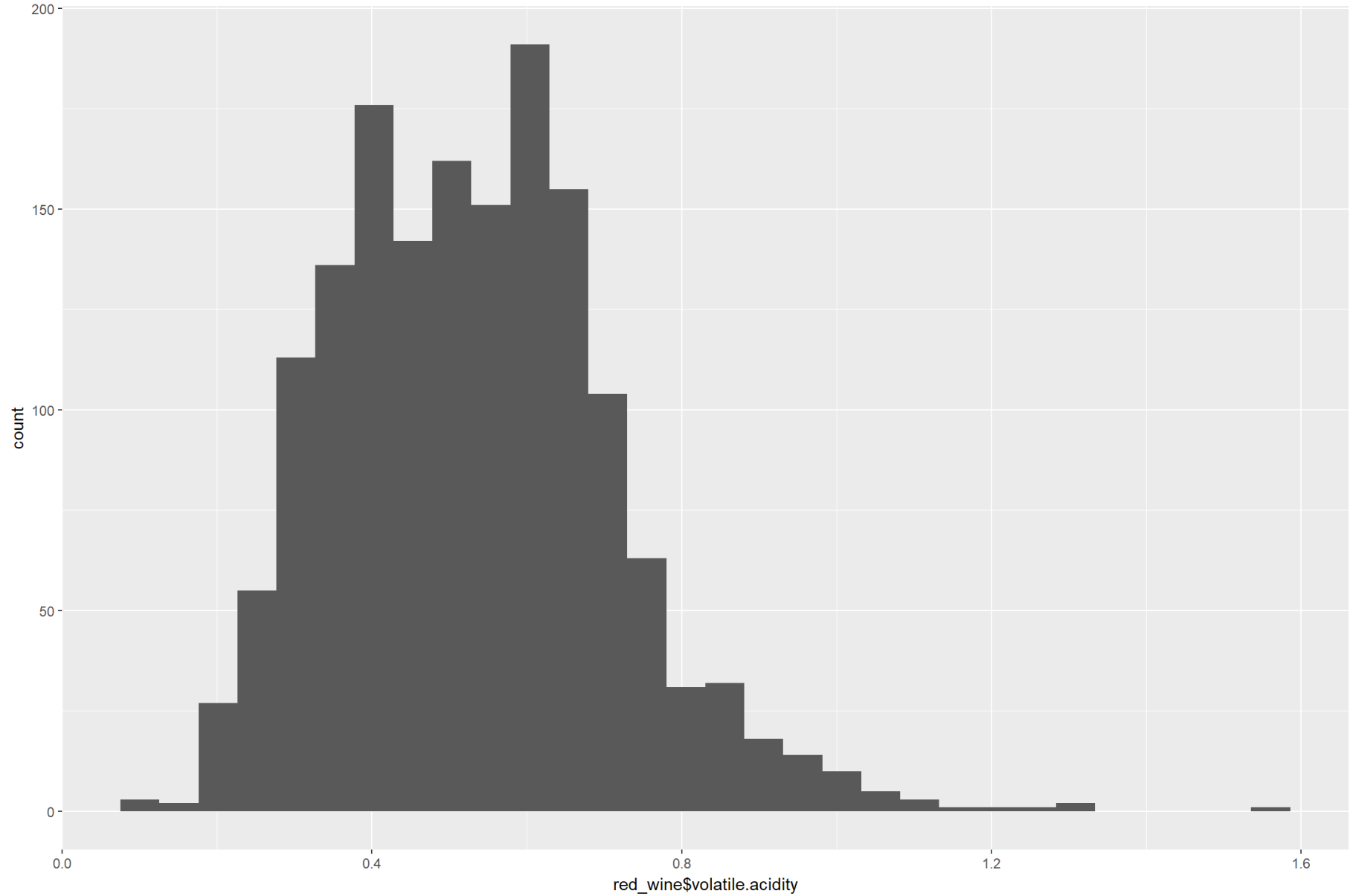
数据总览

```
##        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
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

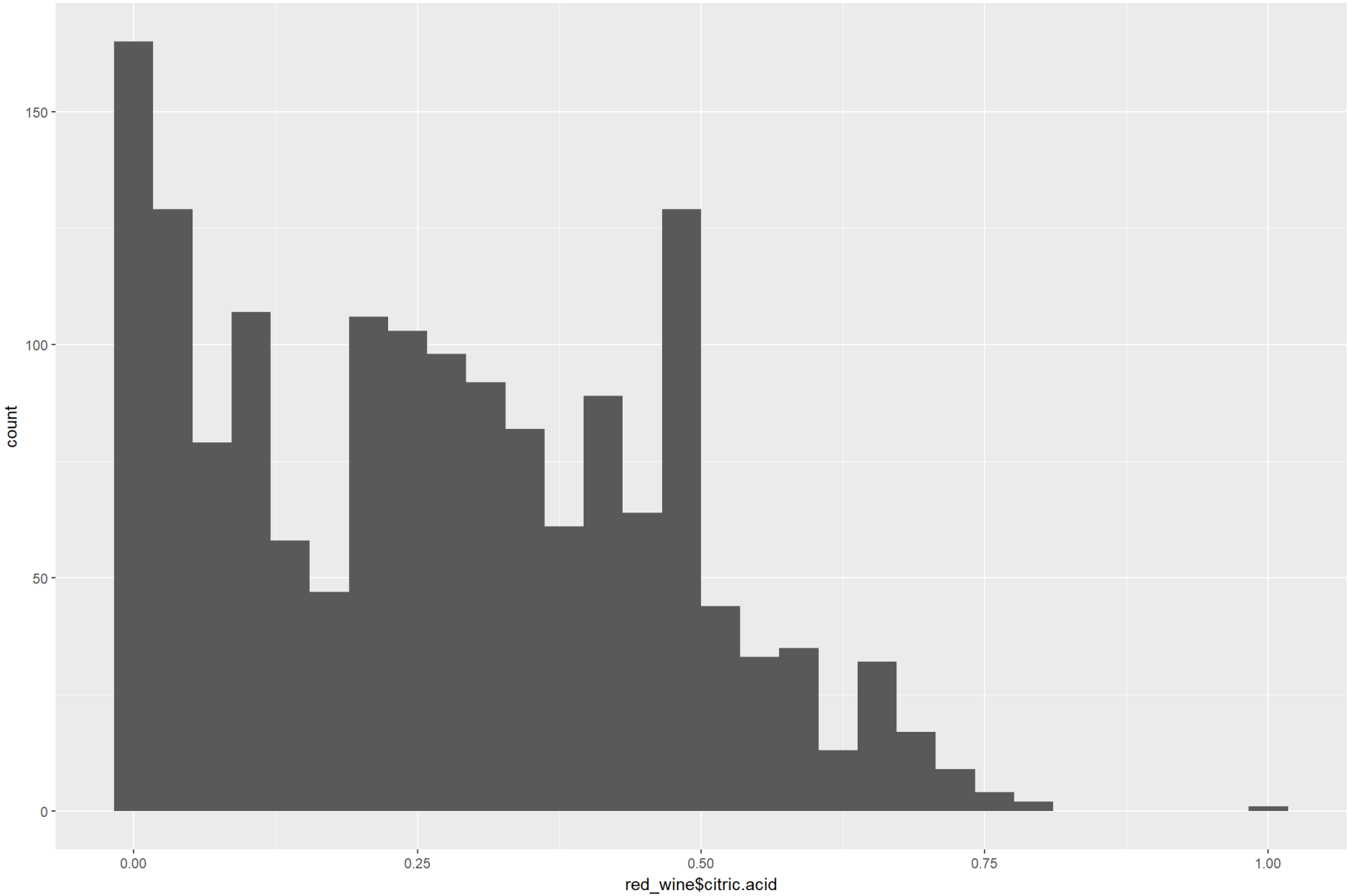
固定酸



挥发酸

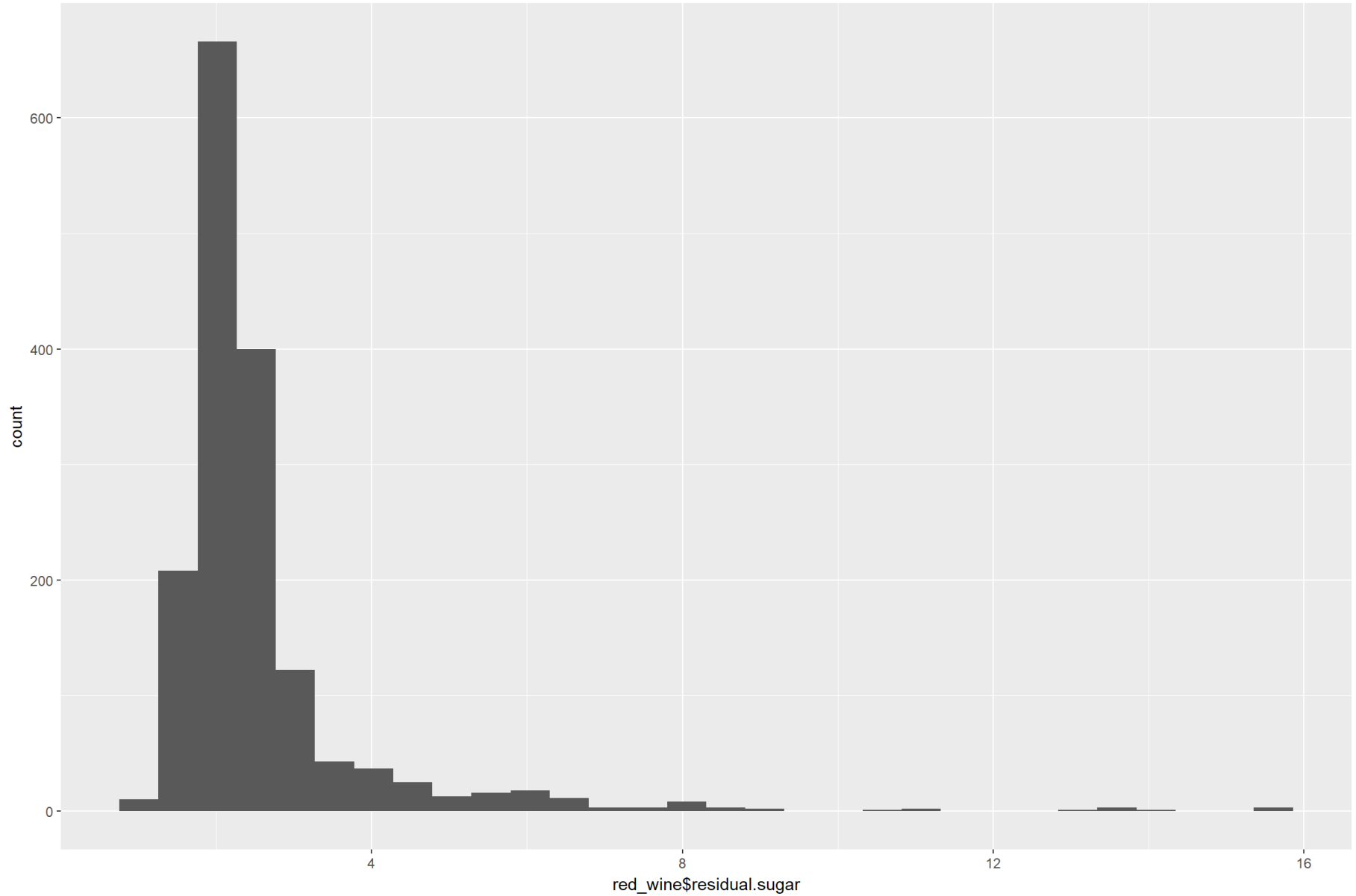


柠檬酸



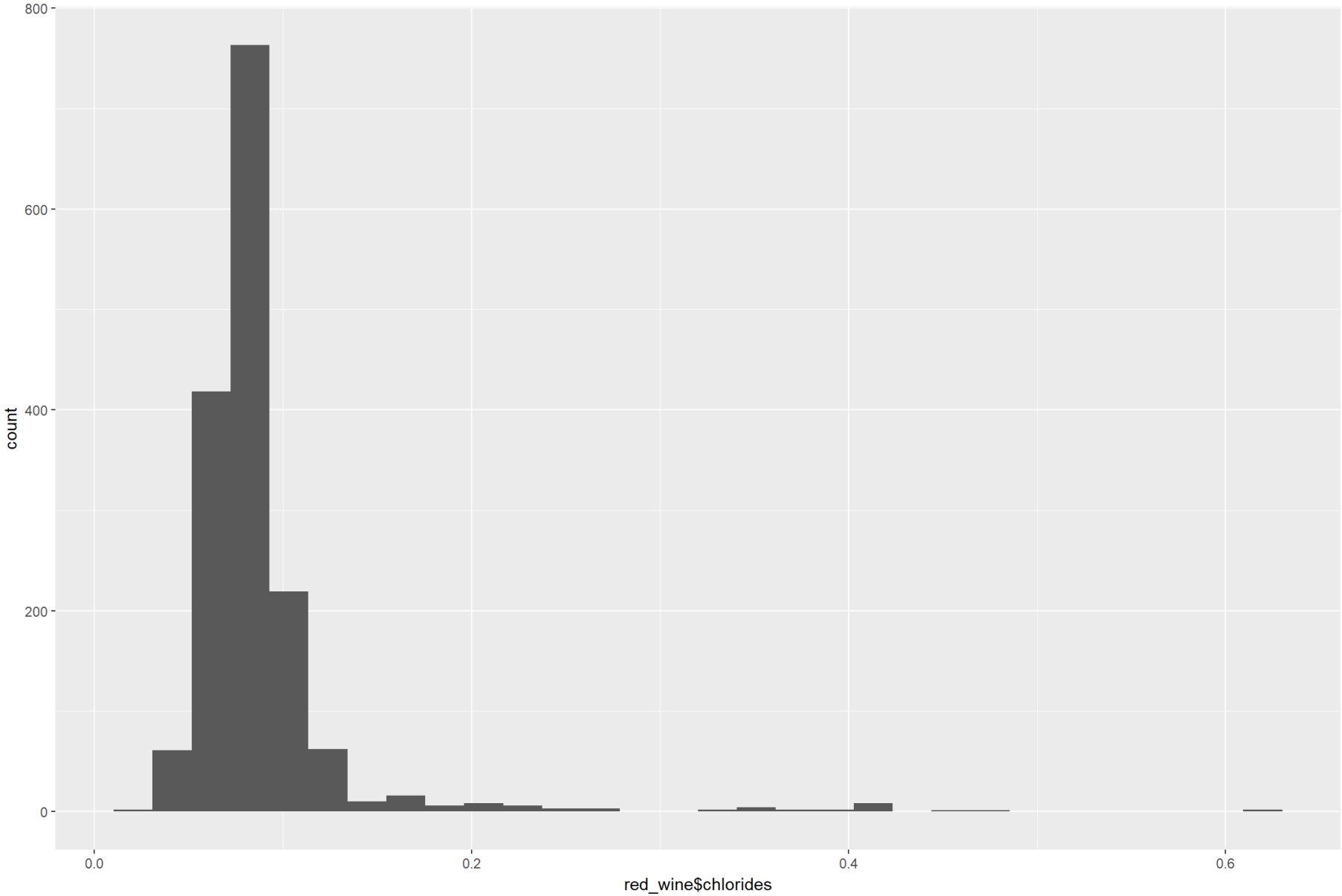
柠檬酸分布有些特殊

残糖



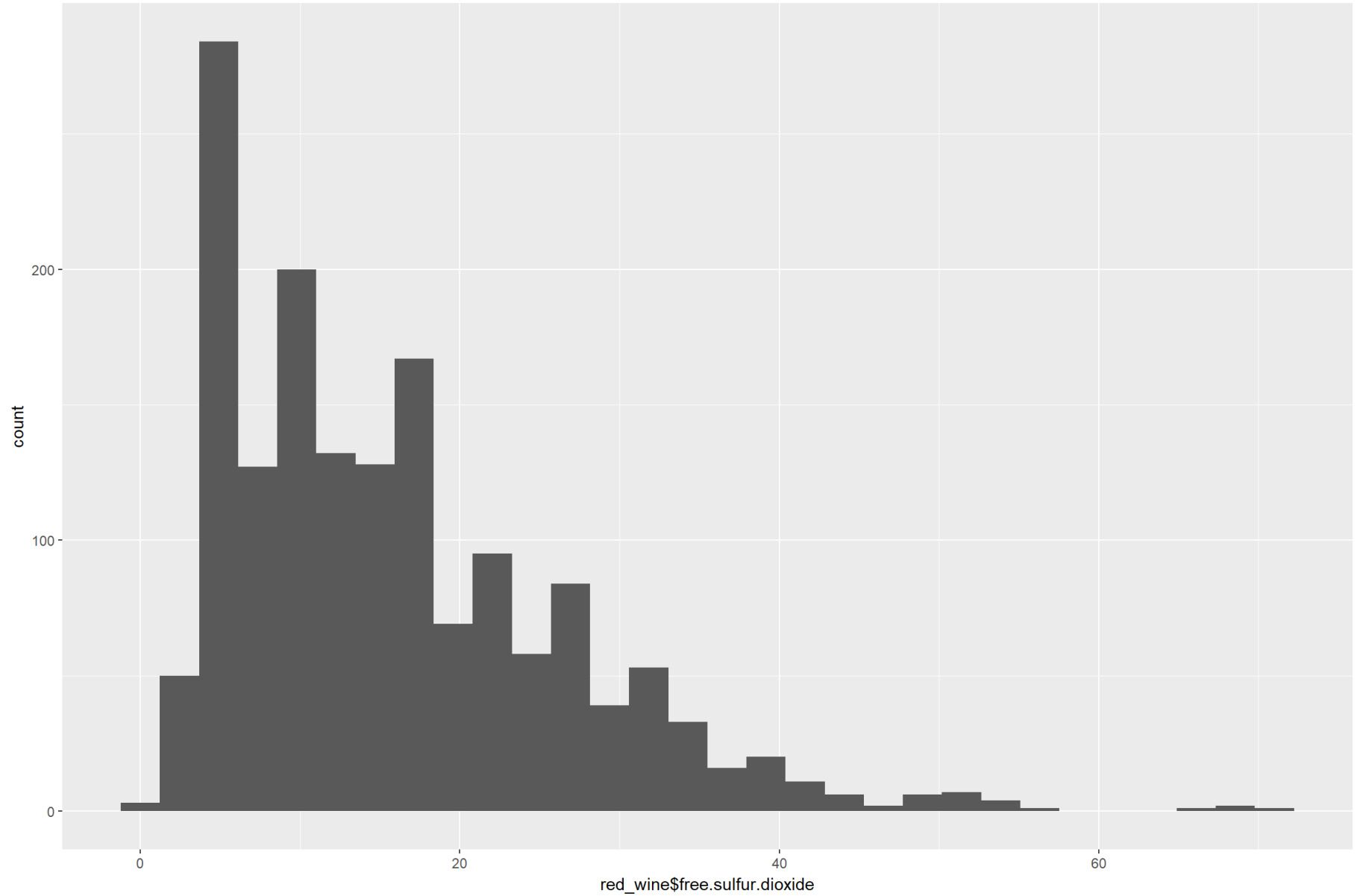
甜度差别很大,但没有一款是甜酒.

氯化物

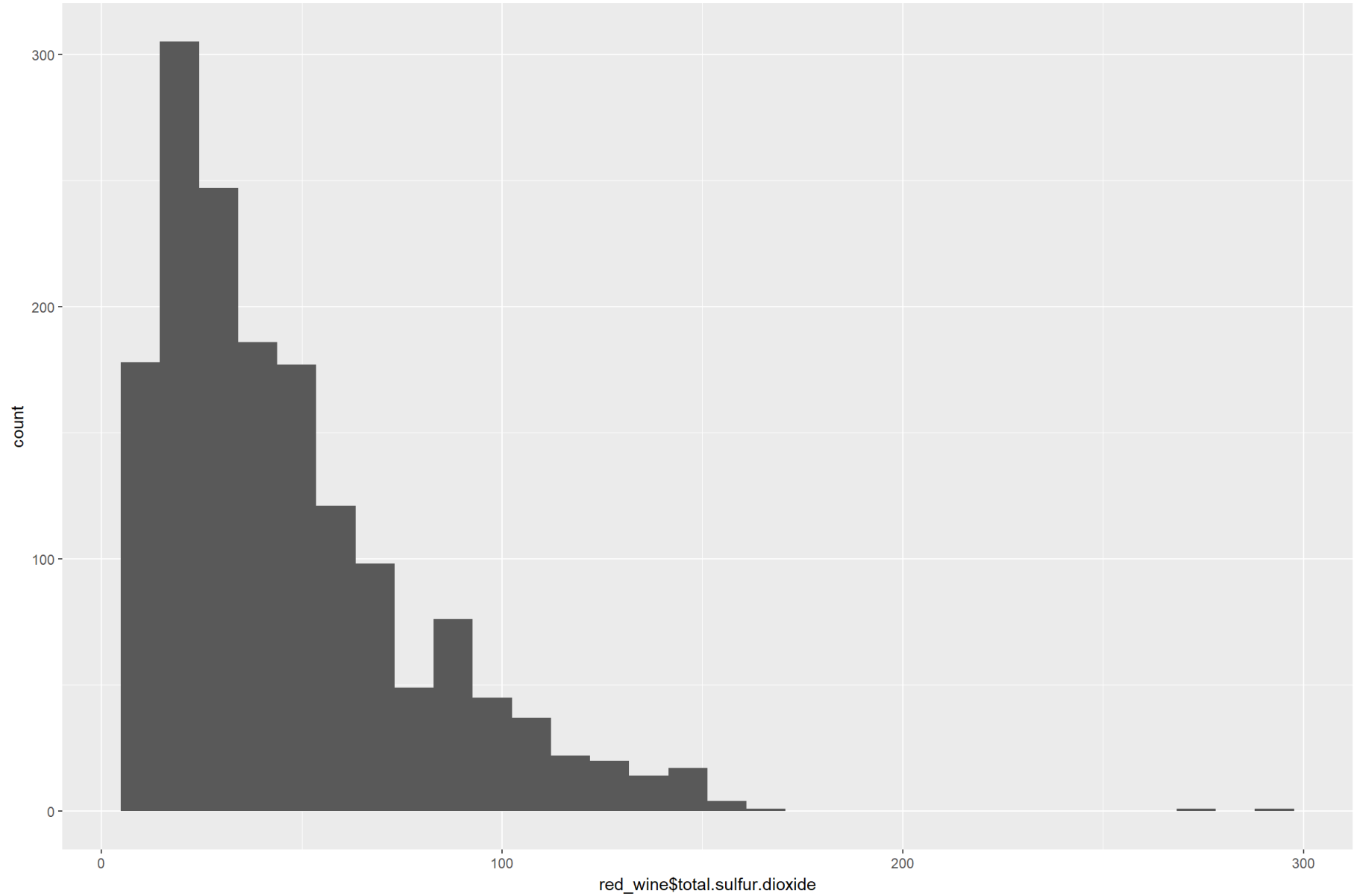


氯化物差别很大

游离二氧化硫

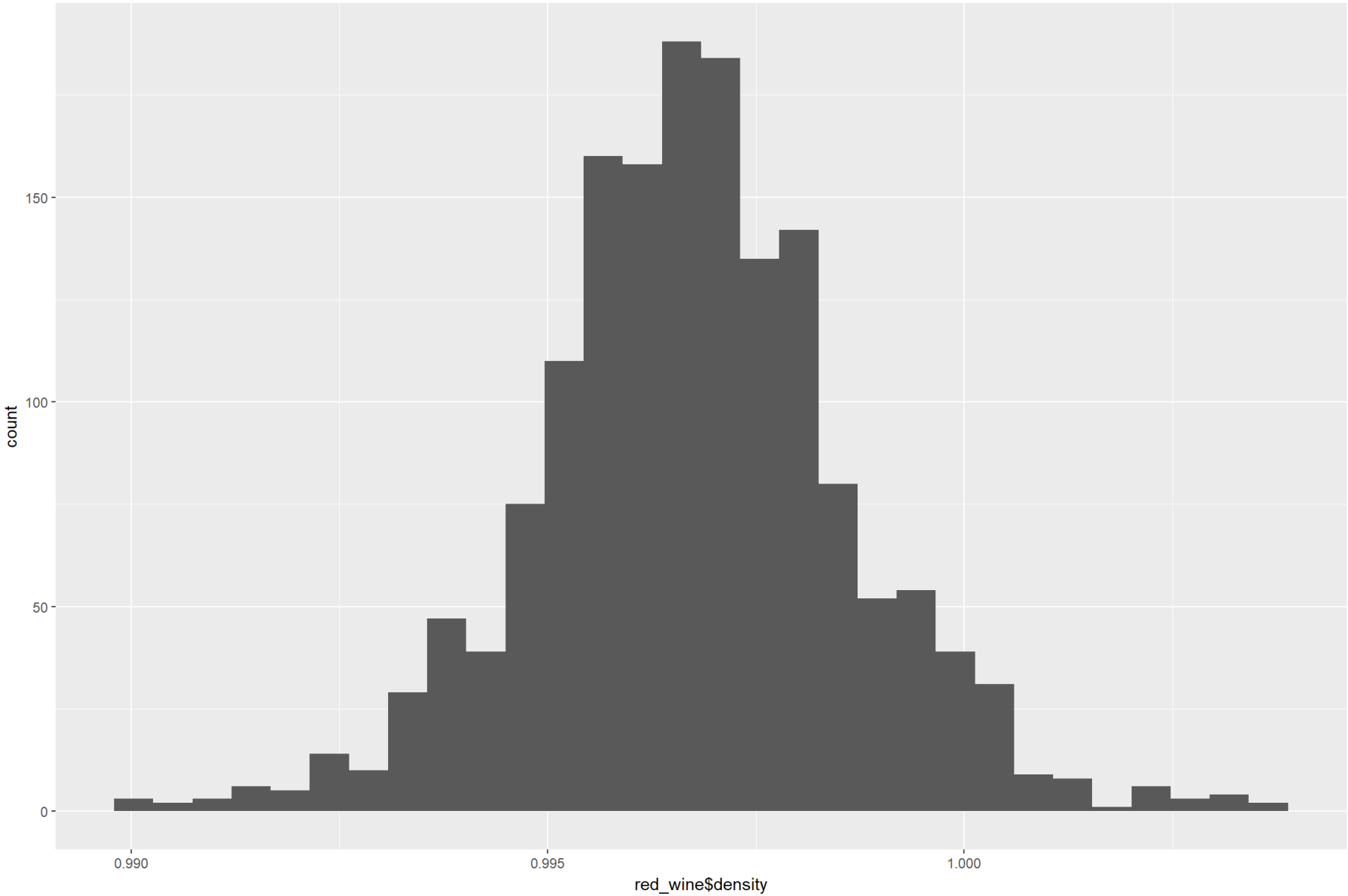


总二氧化硫



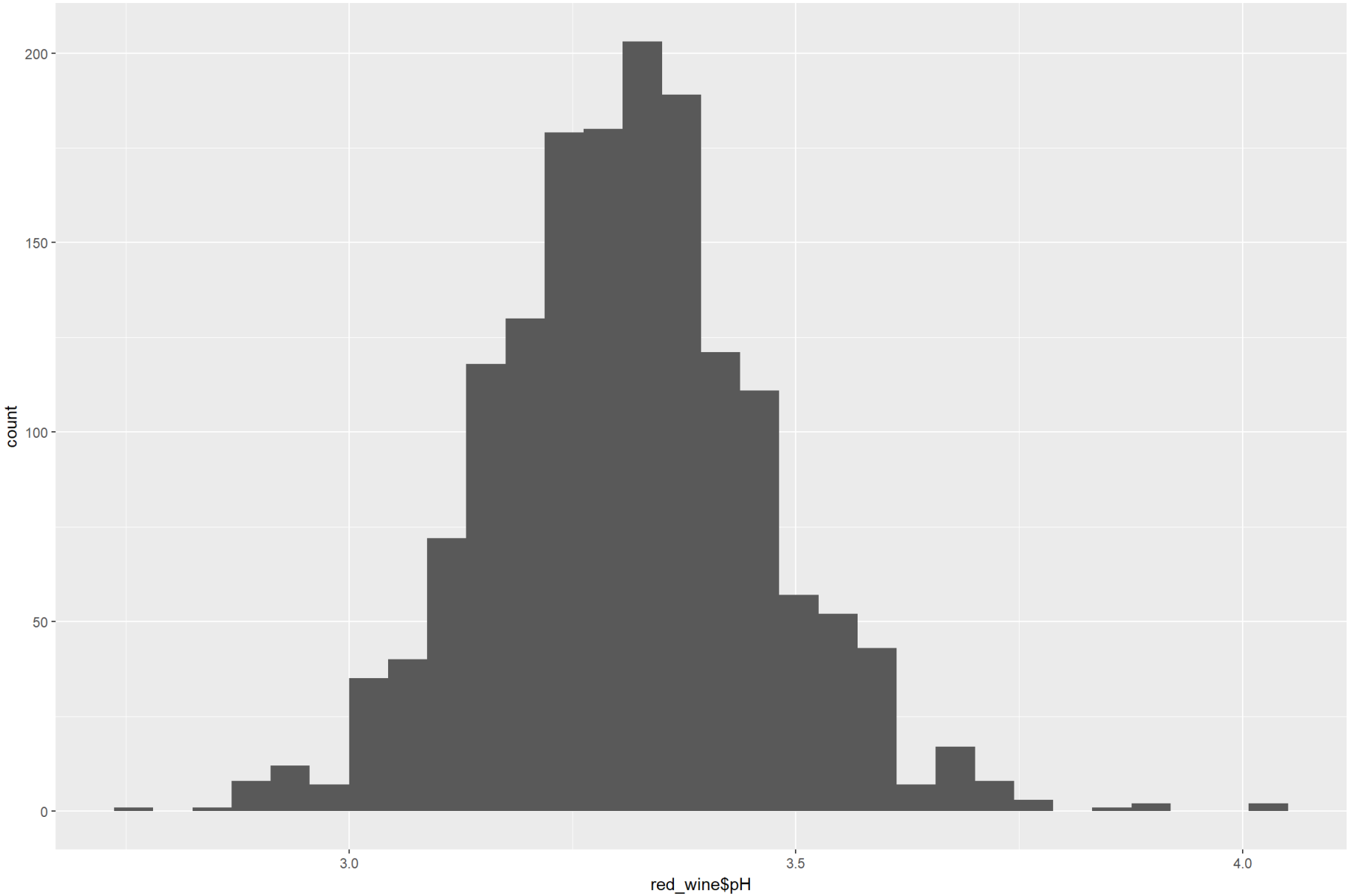
总二氧化硫浓度高于游离二氧化硫,符合常识.

密度



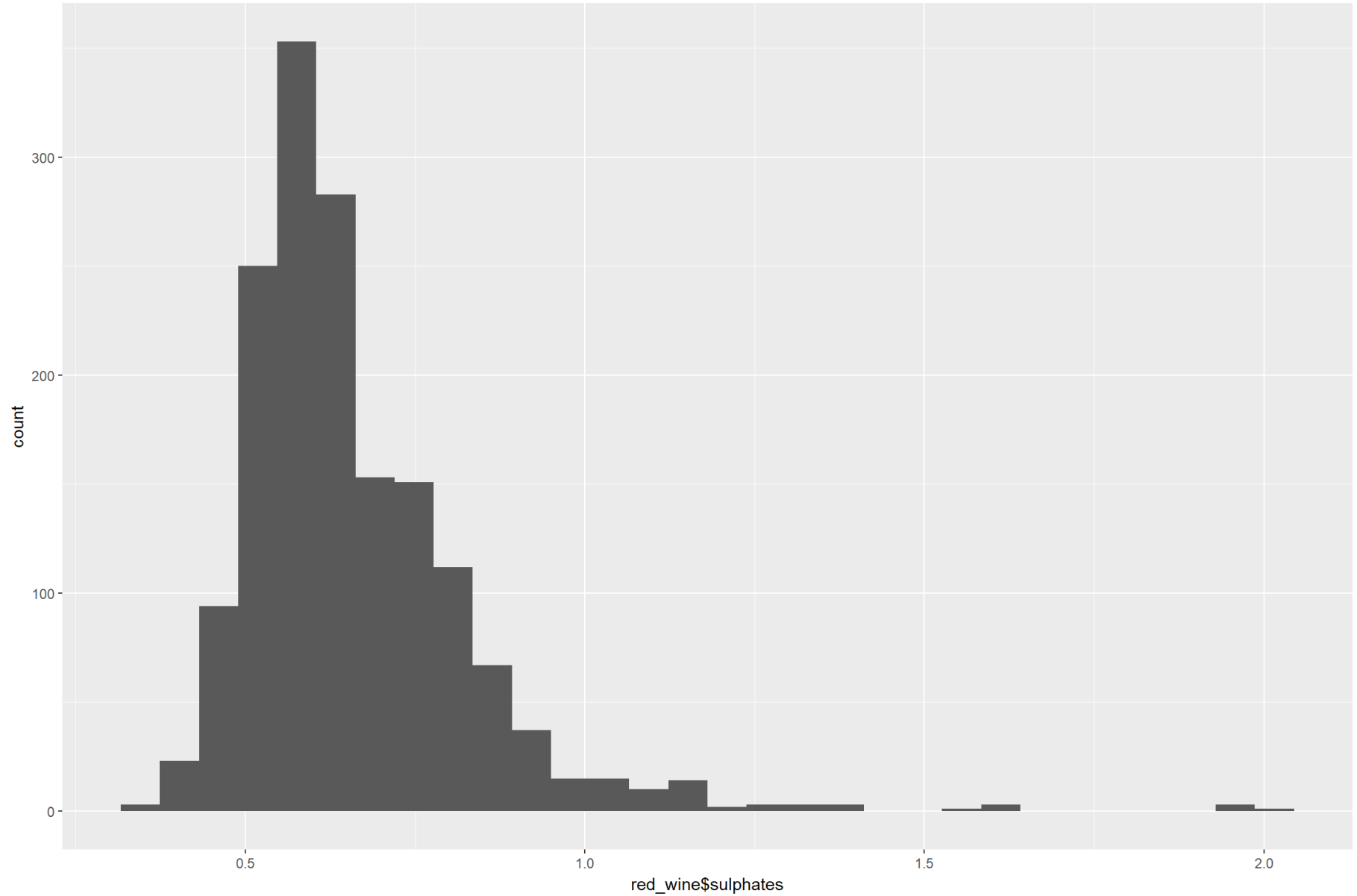
密度因素的分布非常接近于正态分布

PH值

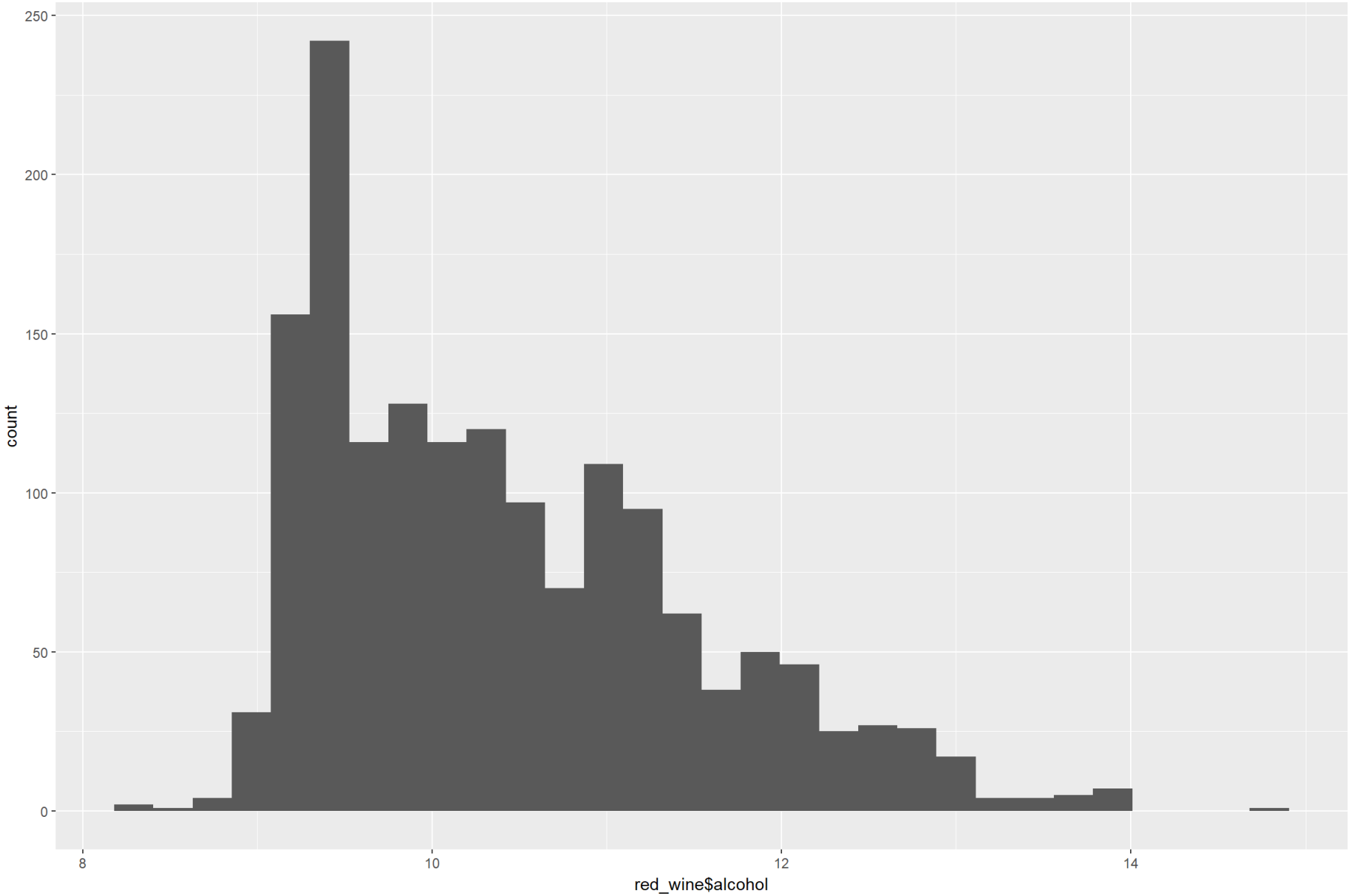


PH值也是非常接近正态分布

硫化物

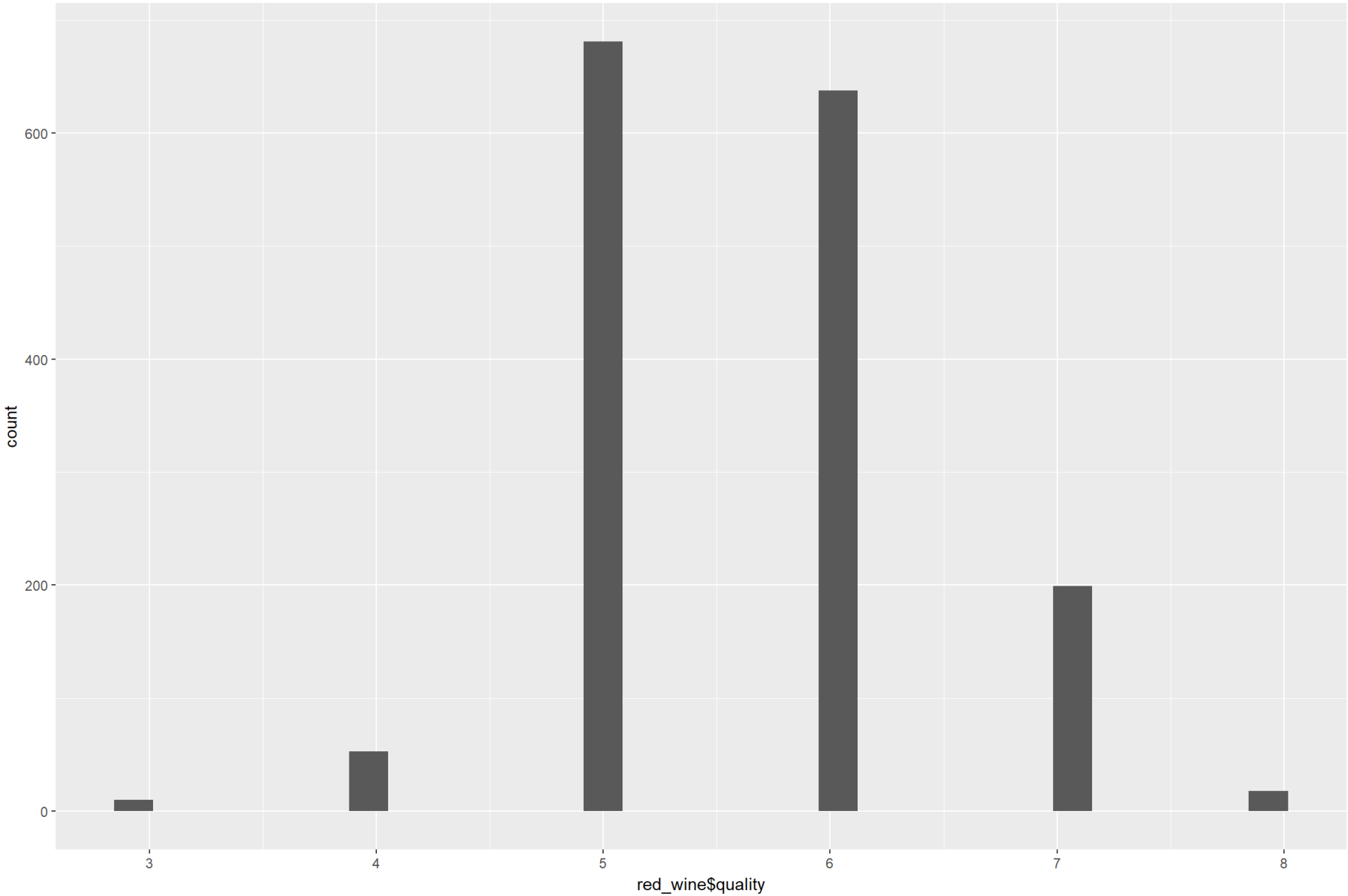


酒精



红酒属于低度酒

质量



大多数红酒的质量为5~6分,符合市场规律.

单变量分析

1你的数据集结构是?

数据集中共有1599种红酒,具有11种化学成分,一个主观质量评分(3~8).

2你对数据集内感兴趣的特征?

探究影响红酒质量的因素和柠檬酸的特性.

3你认为数据集内哪些其他特征可以帮助你探索兴趣特点?

与红酒质量相关的因素可能有酒精度,糖类,酸类,二氧化硫.

与柠檬酸相关的因素可能有固定酸,挥发酸,pH值.

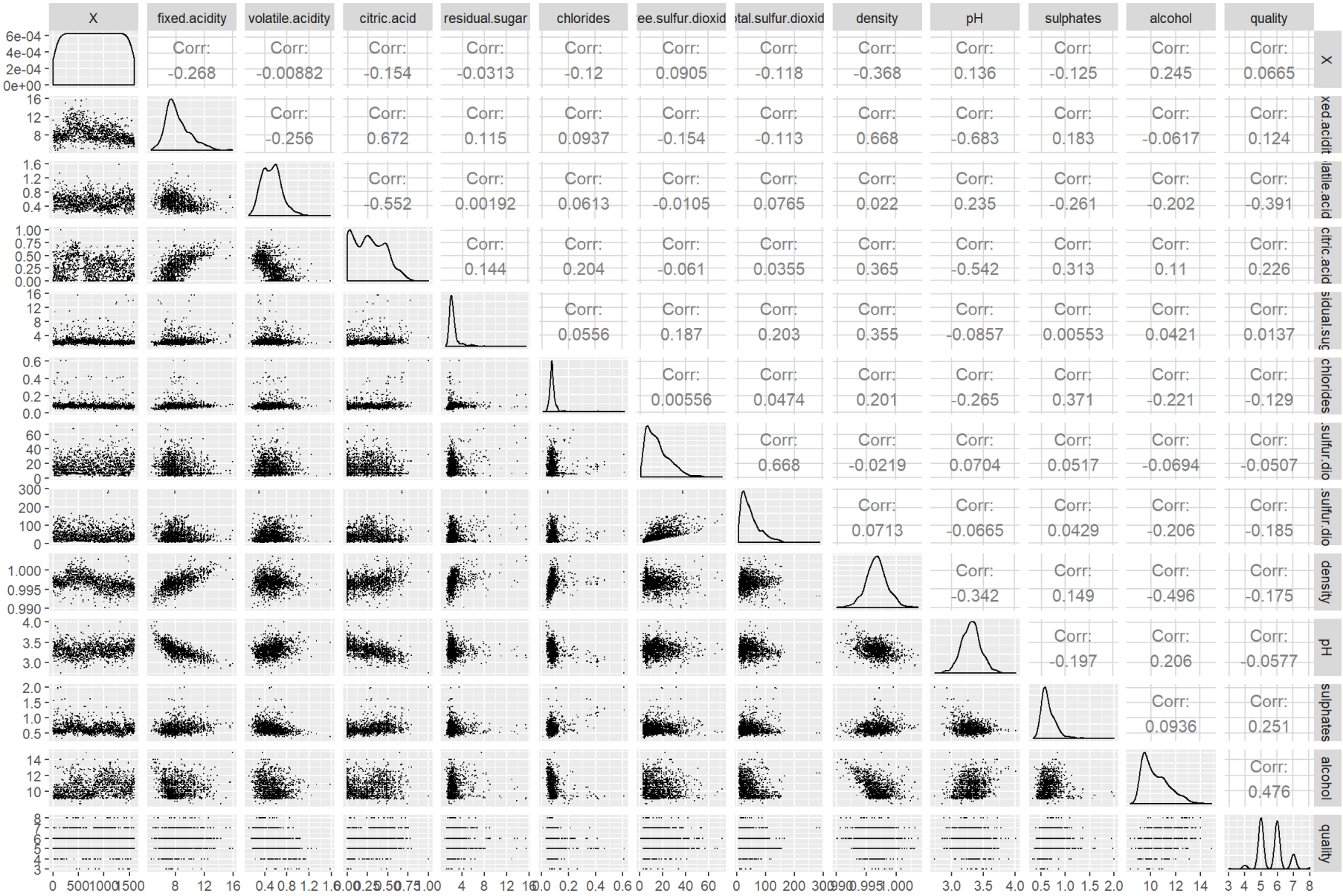
4在已经探究的特性中,是否存在任何异常分布?你是否对数据进行一些操作?

通过简单绘图发现,多数直方图都是近似正态分布,并且左偏.

柠檬酸的直方图比较特殊,呈现双峰结构.

双变量绘图选择

散点图矩阵



1影响质量的主要因素有:

挥发酸,柠檬酸,硫酸盐,酒精.

2高度相关的因素有:

柠檬酸和固定酸,挥发酸

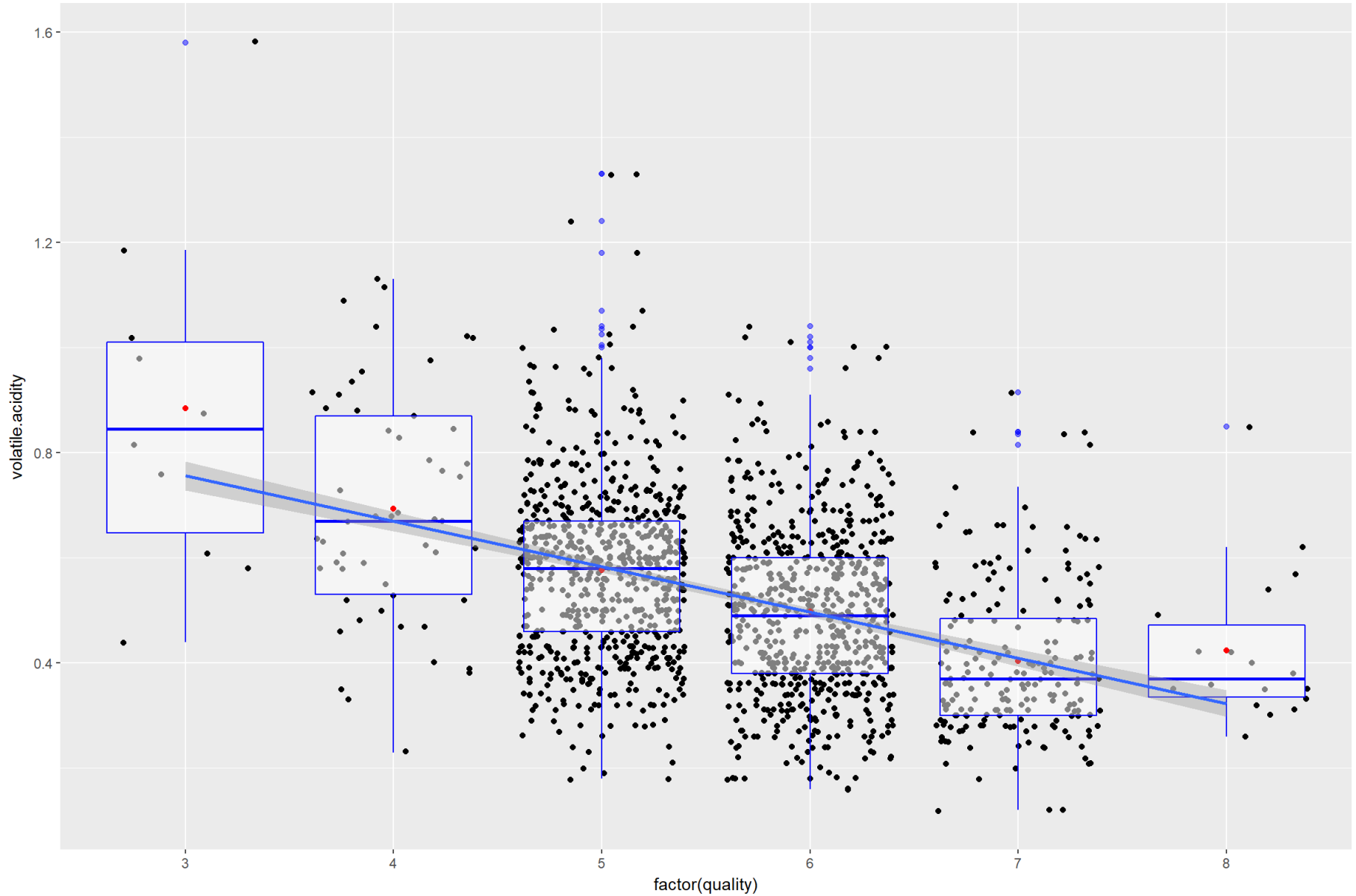
柠檬酸和PH值

总二氧化硫和游离二氧化硫

酒精和密度

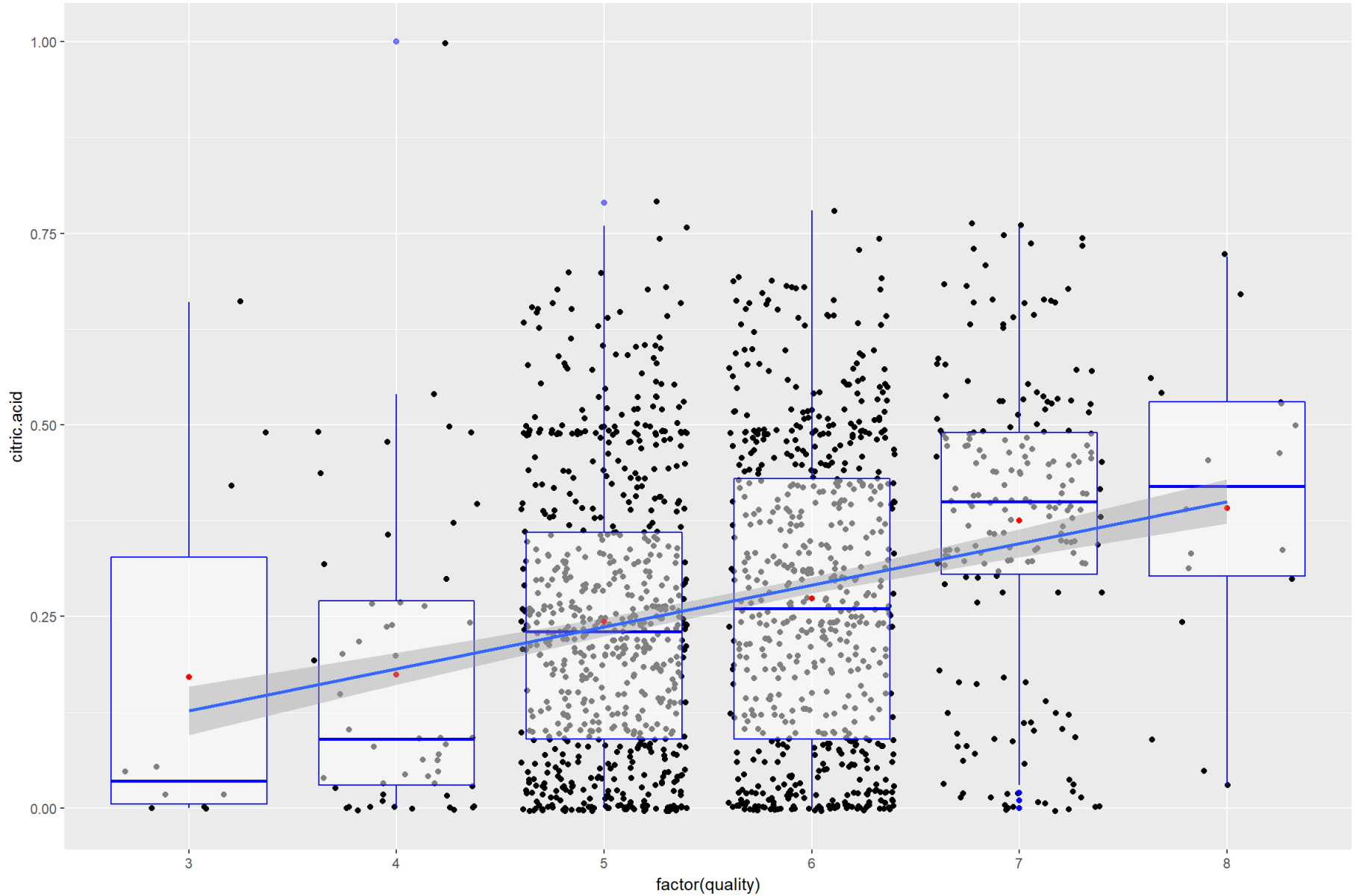
酒精和质量

质量和挥发酸的相关关系:



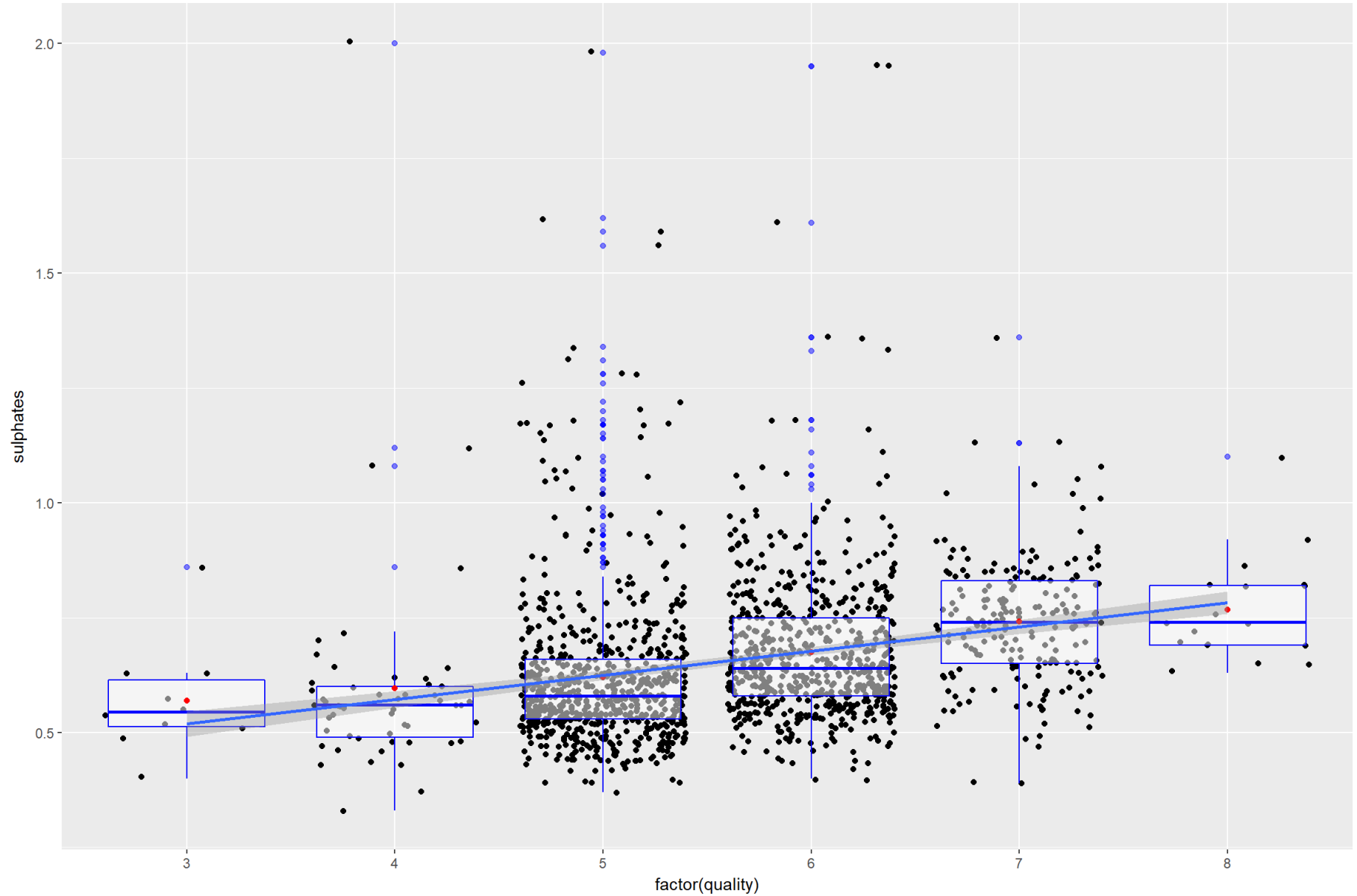
质量和挥发酸不是严格的线性相关, 除8级外呈负相关.

质量和柠檬酸的相关关系:



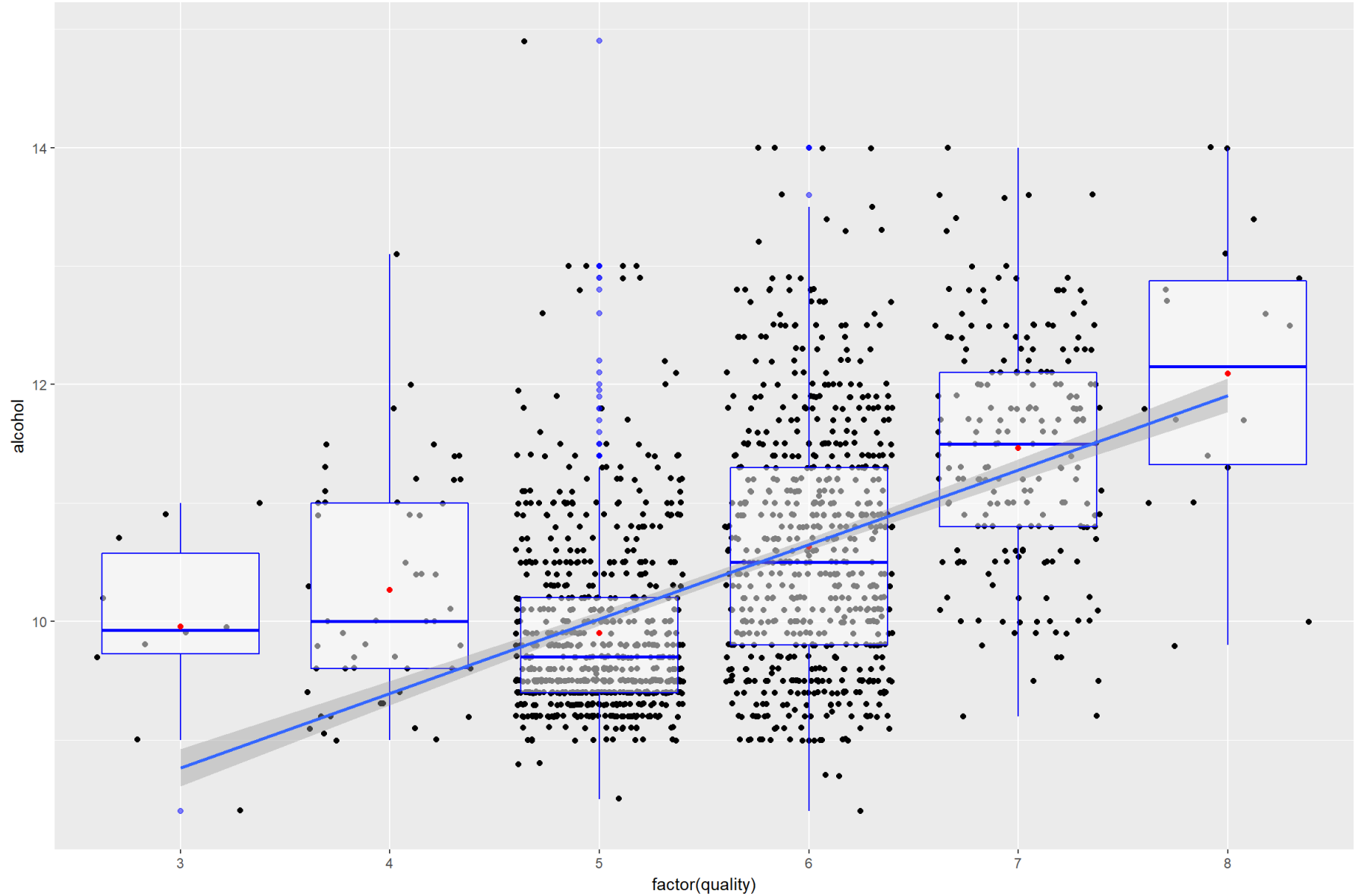
质量和柠檬酸呈正相关

质量和硫化物的相关关系:



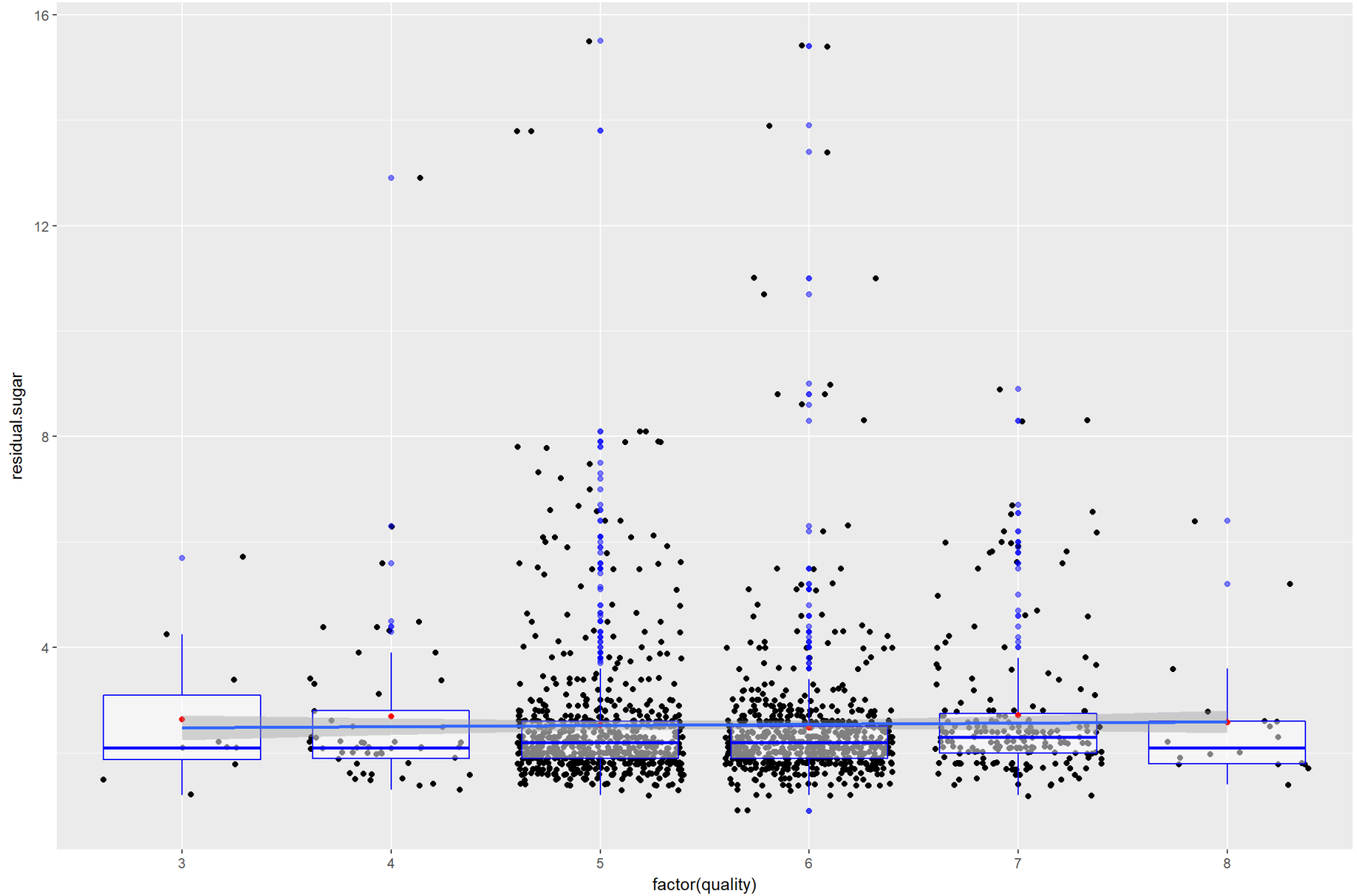
质量和硫化物呈正相关

质量和酒精的相关关系:



质量和酒精不是线性相关,5级之后呈正相关

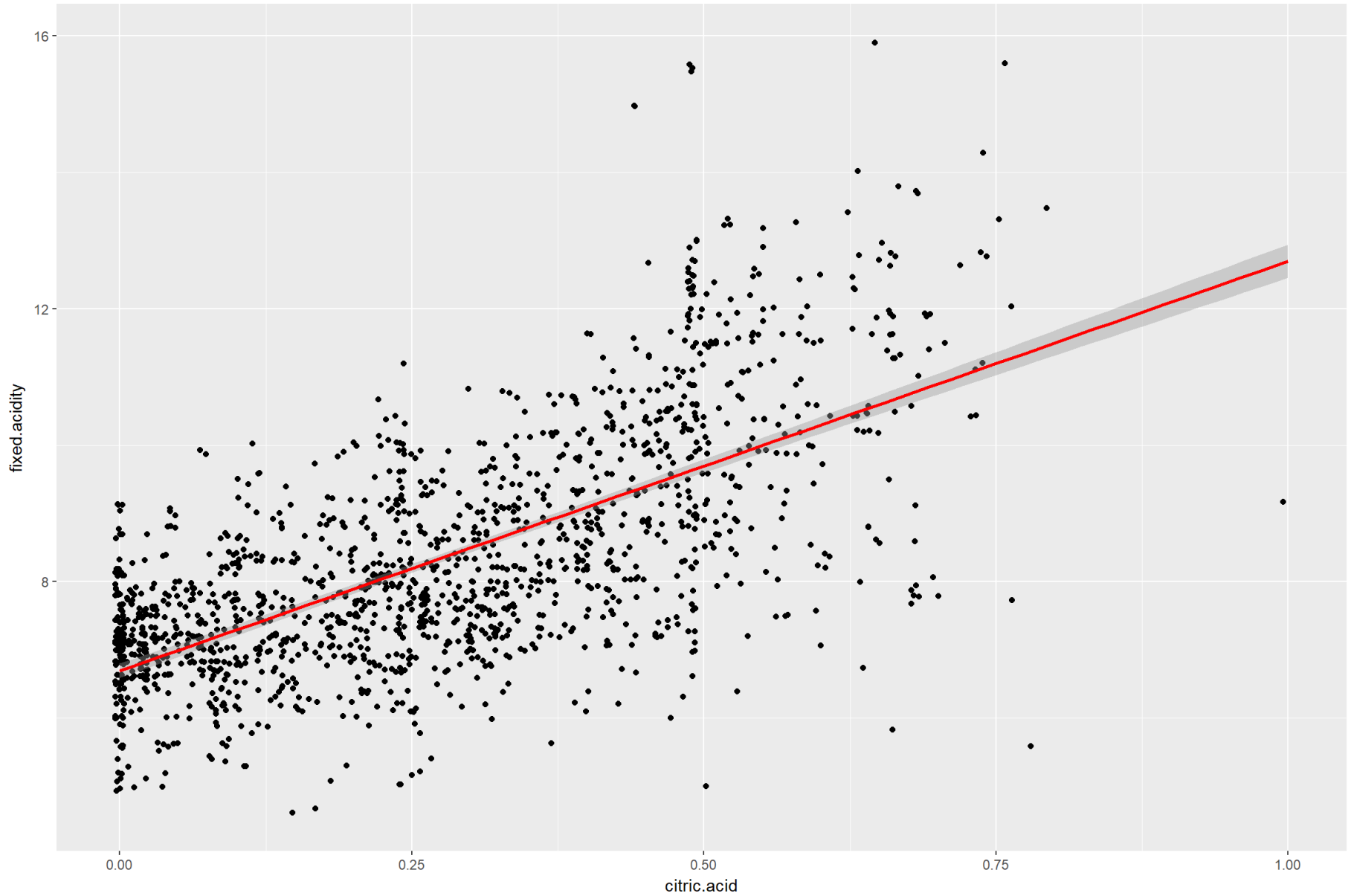
质量和残糖的相关关系:



散点图矩阵中显示质量和残糖无关,发现确实如此.

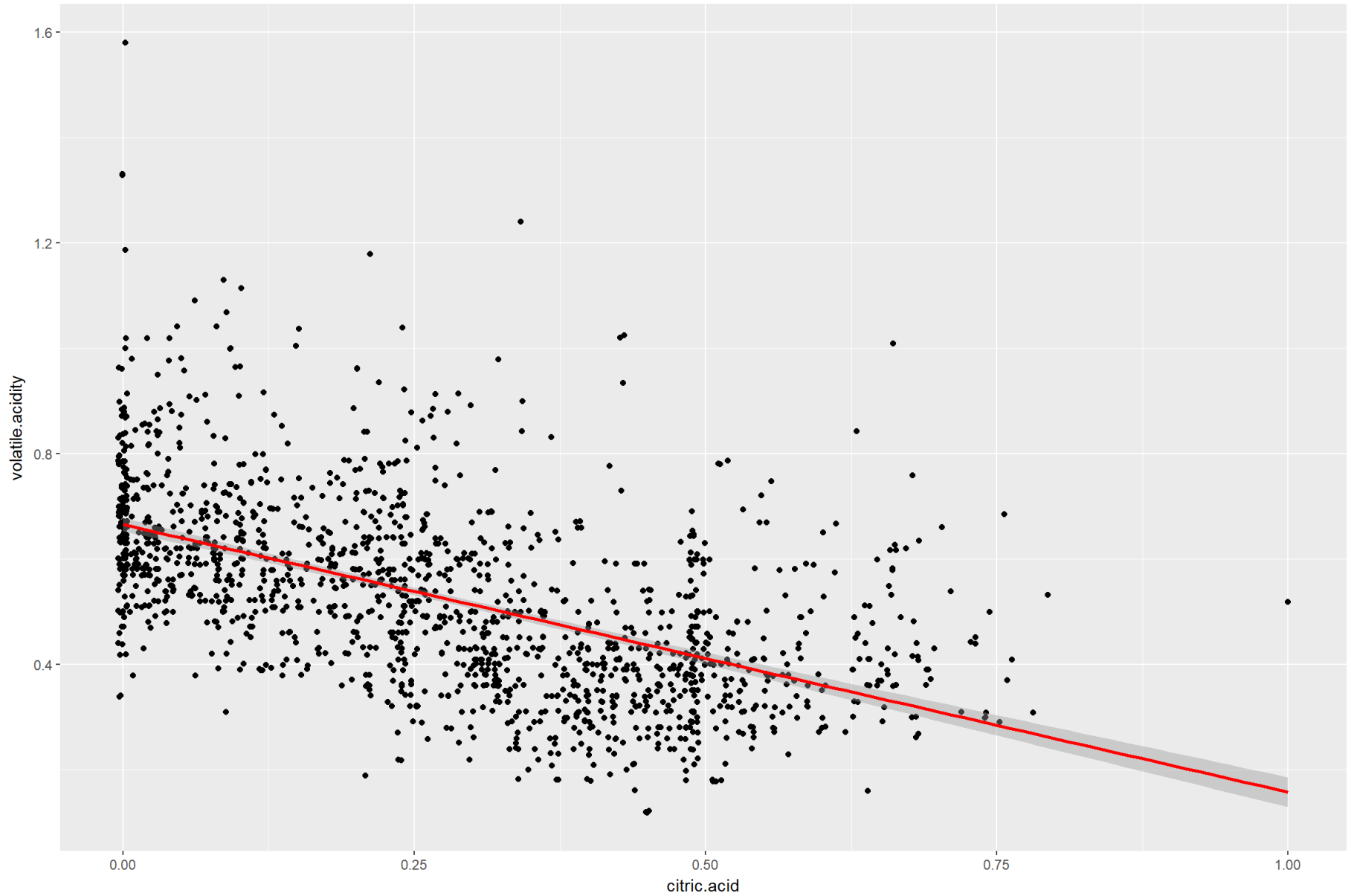
其他相关关系:

柠檬酸和固定酸的相关关系:



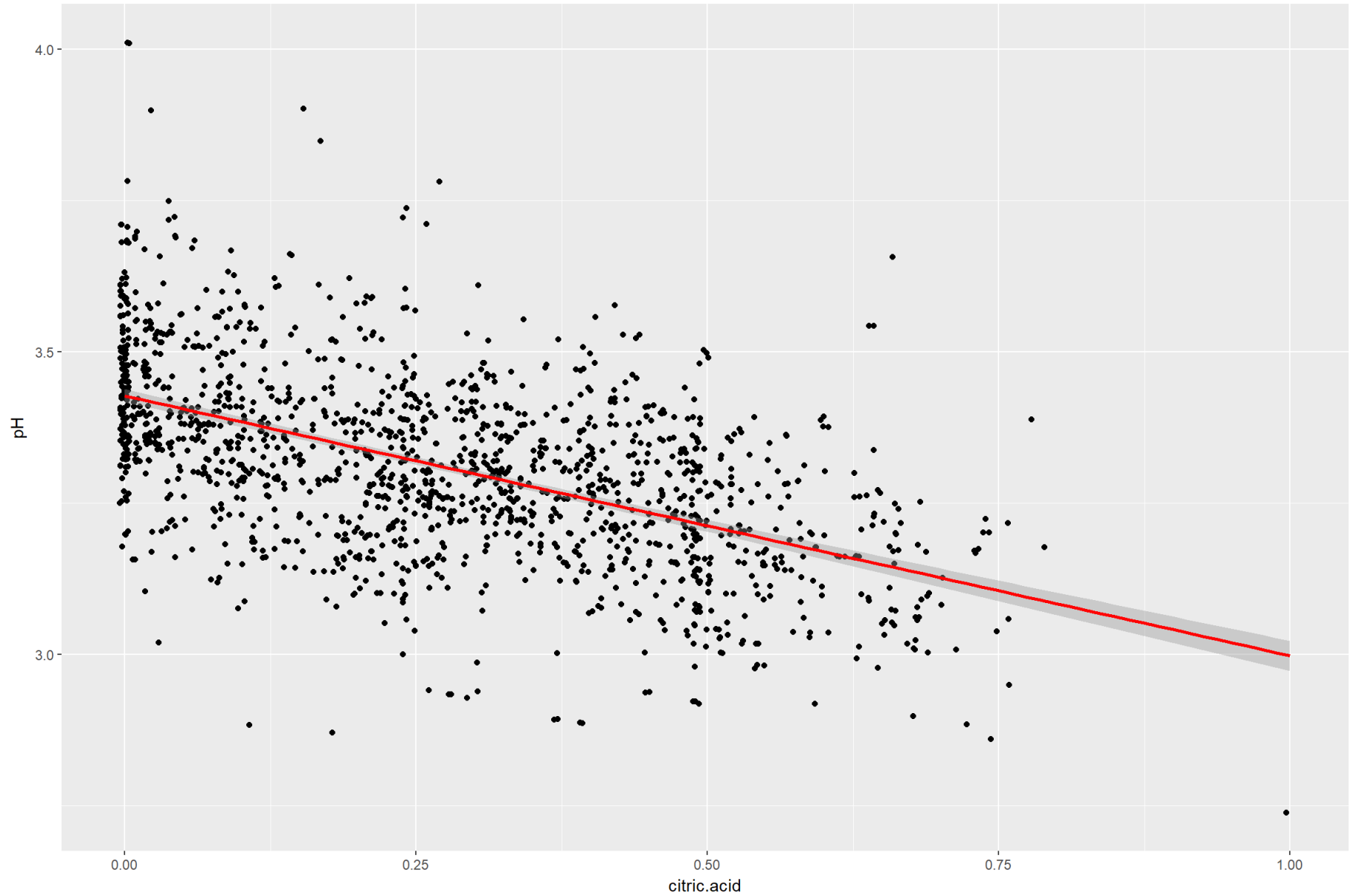
柠檬酸和固定酸呈正相关

柠檬酸和挥发酸的相关关系:



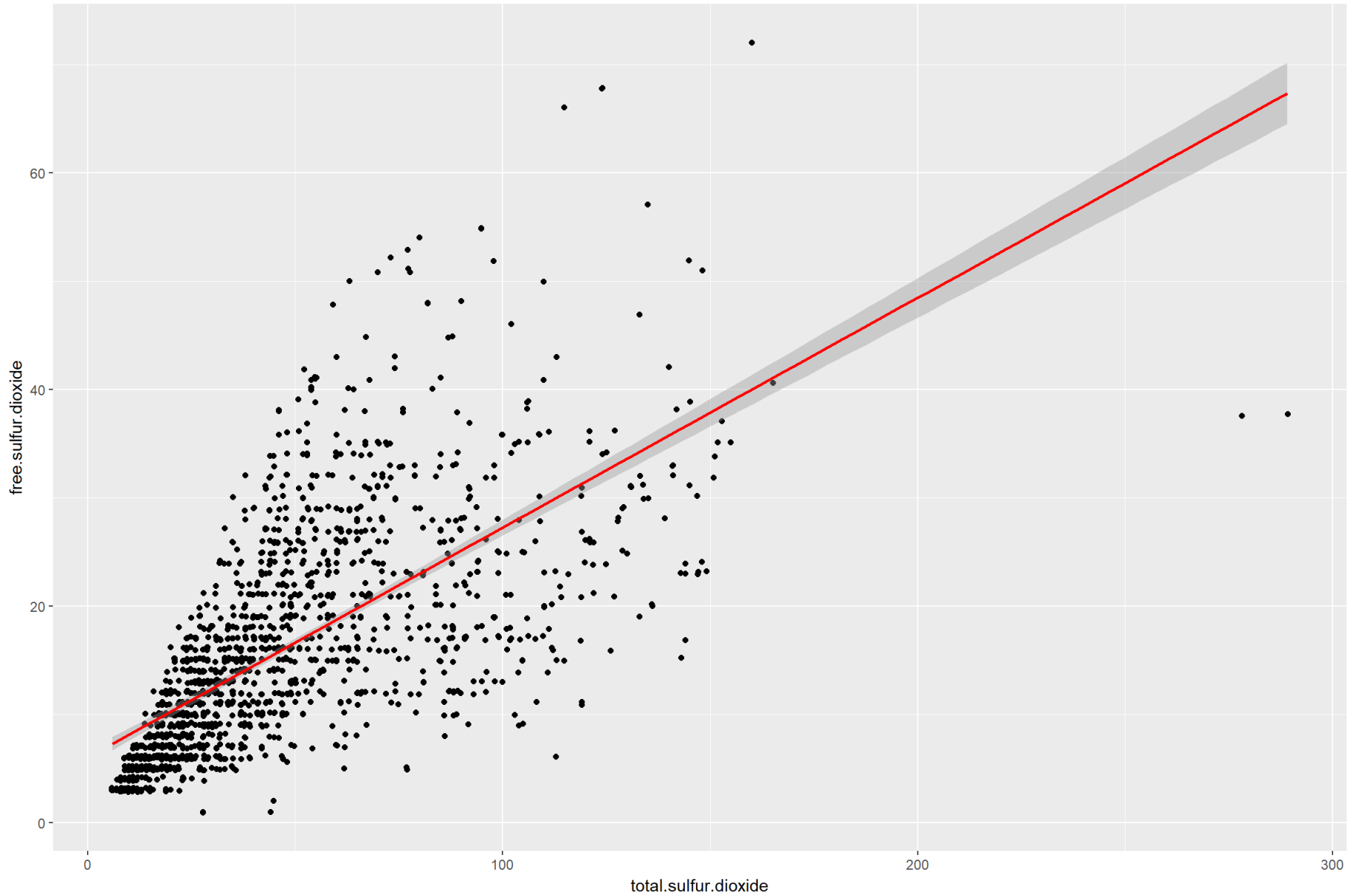
柠檬酸和挥发酸呈负相关

柠檬酸和PH值的相关关系:



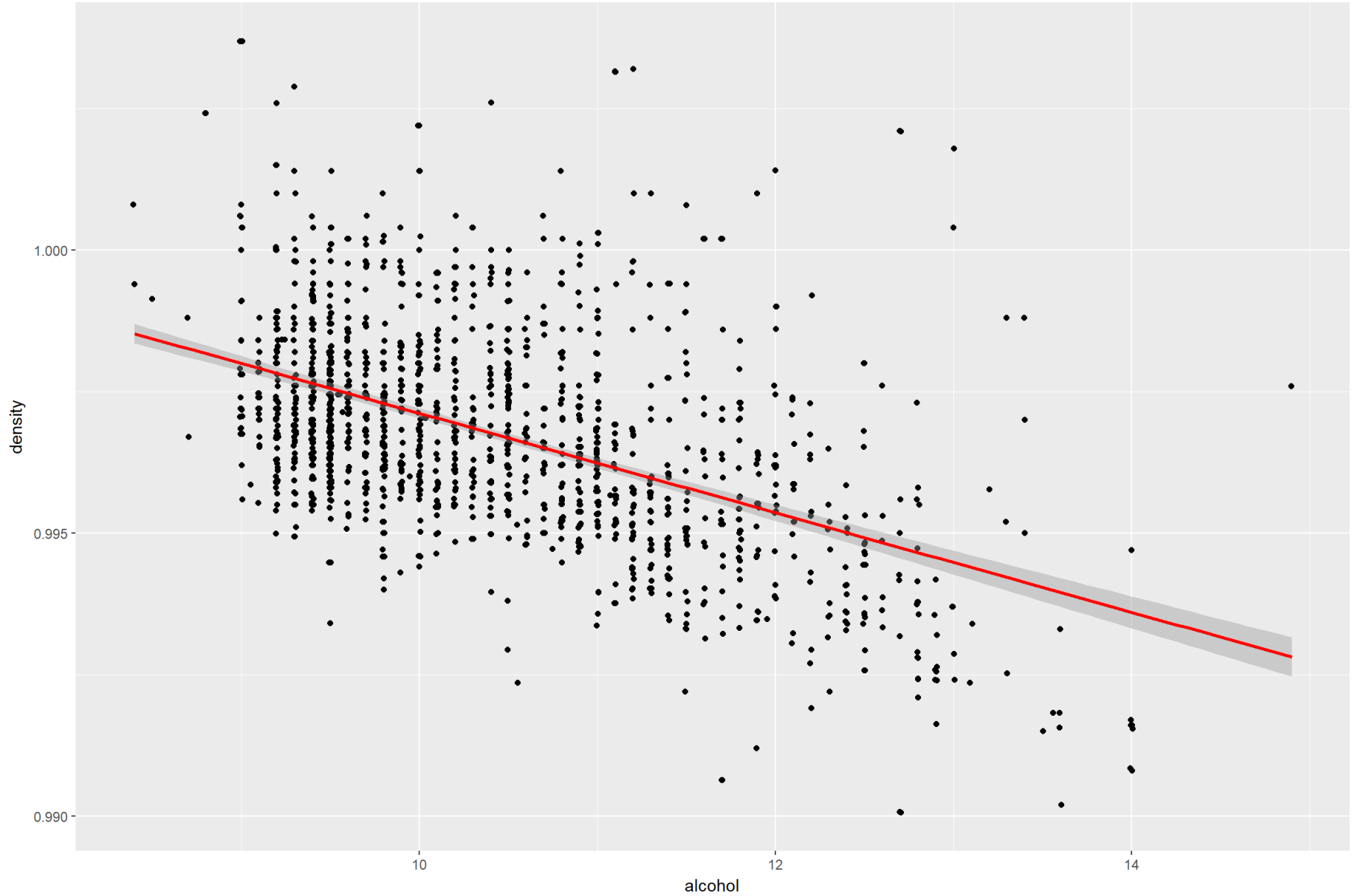
柠檬酸和PH值呈负相关

总二氧化硫和游离二氧化硫的相关关系:



总二氧化硫和游离二氧化硫呈正相关

酒精和密度的相关关系:



酒精和密度呈负相关

双变量分析

探讨你在这部分探究中观察到的一些关系,这些感兴趣的特性与数据集内其他特性有什么区别?

基于散点图矩阵,探究了影响质量的最重要的四个因素,根据相关程度依次为酒精,挥发酸,硫化物和柠檬酸.

质量与柠檬酸呈正相关,与硫化物呈正相关.

质量和挥发酸不是严格的线性相关, 除8级外呈负相关.

质量和酒精不是线性相关,5级之后呈正相关

你是否观察到主要特征与其它特征之间有趣的关系?

柠檬酸和固定酸呈正相关

柠檬酸和挥发酸呈负相关

柠檬酸和PH值呈负相关

总二氧化硫和游离二氧化硫呈正相关

酒精和密度呈负相关

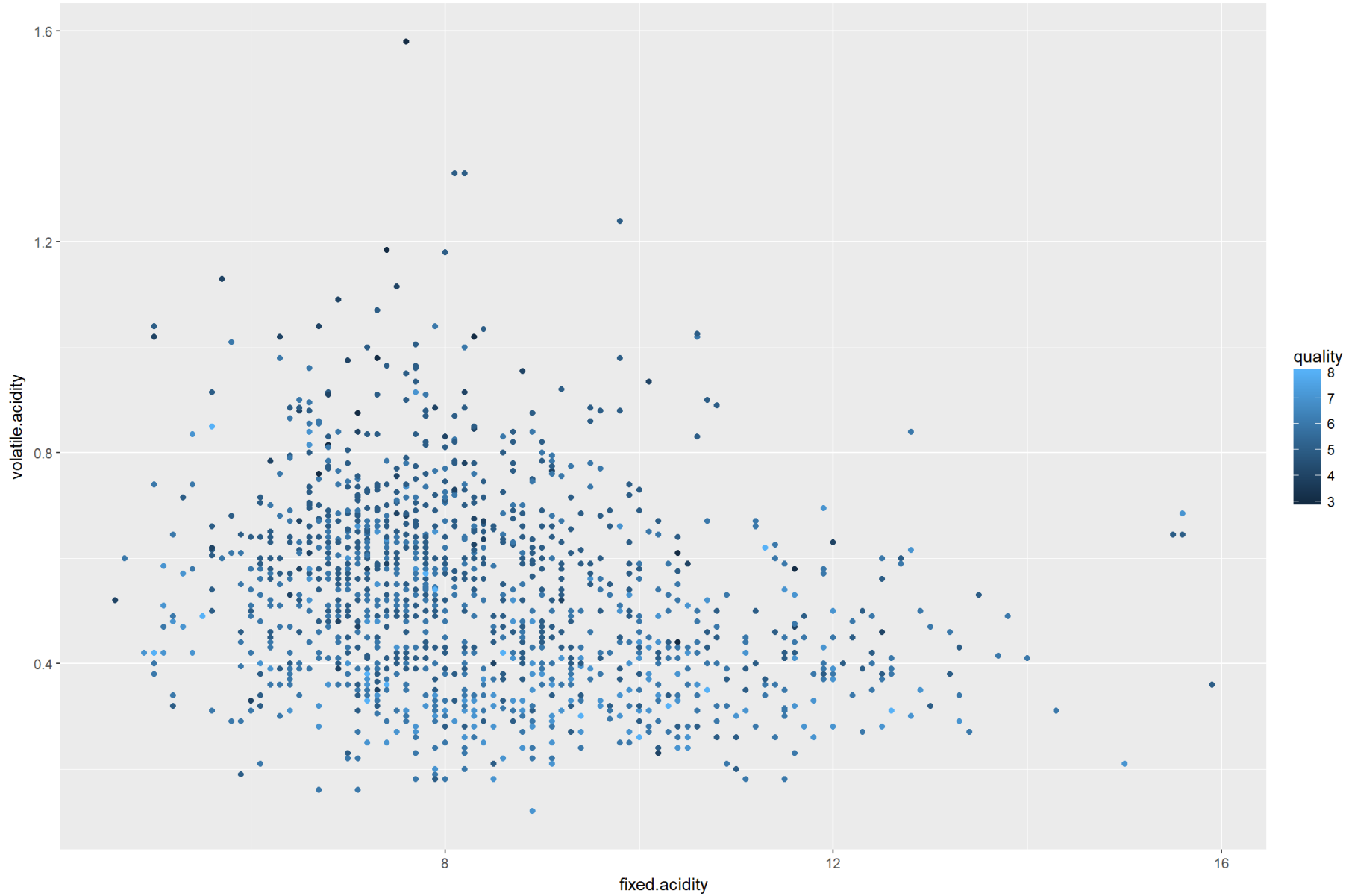
你发现的最强关系是什么?

```
##
##  Pearson's product-moment correlation
##
## data:  red_wine$fixed.acidity and red_wine$pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##       cor
## -0.6829782
```

最强相关为固定酸和PH值,皮尔森相关系数约为-0.68.

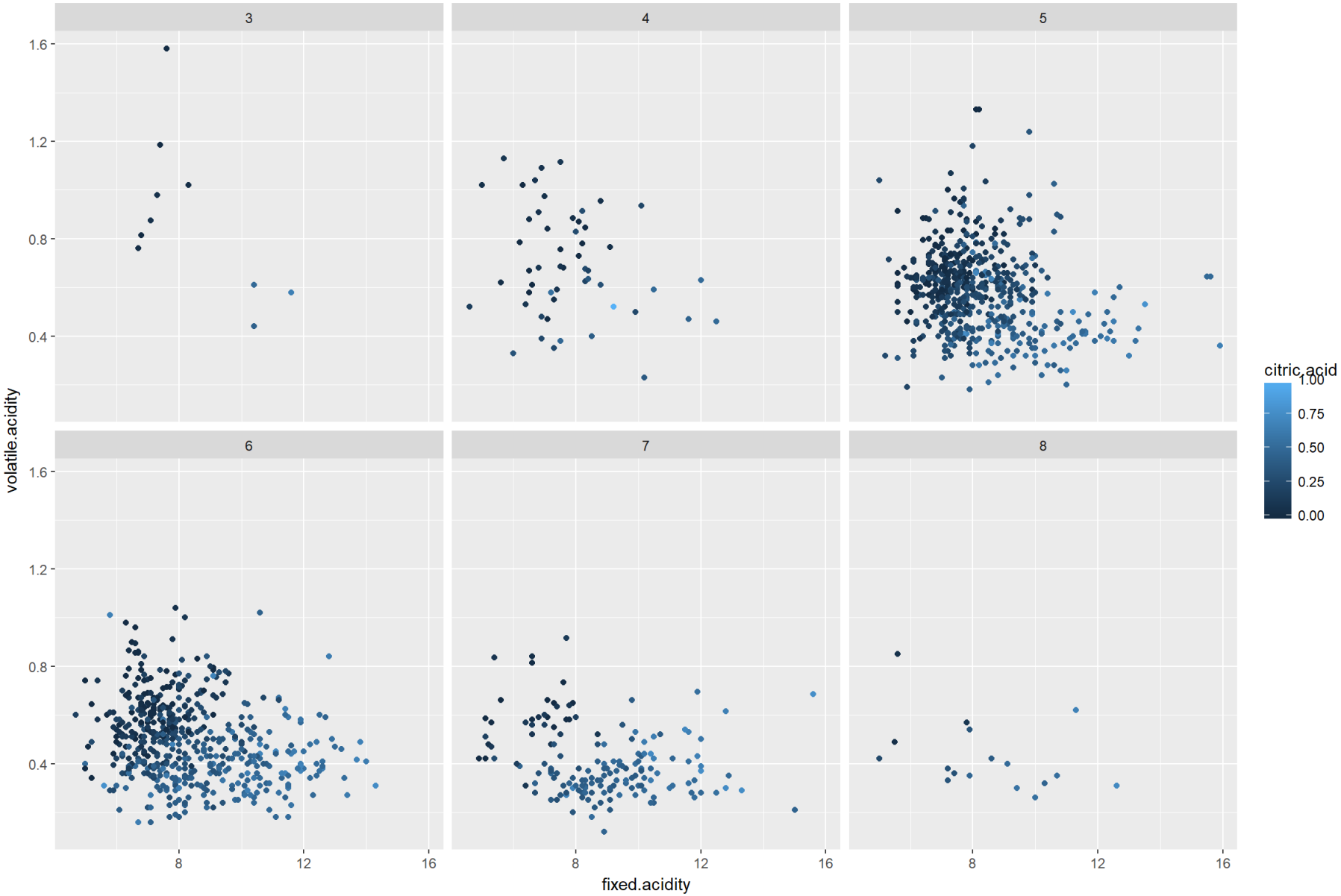
多变量绘图选择

固定酸和挥发酸在不同质量下的相关关系:



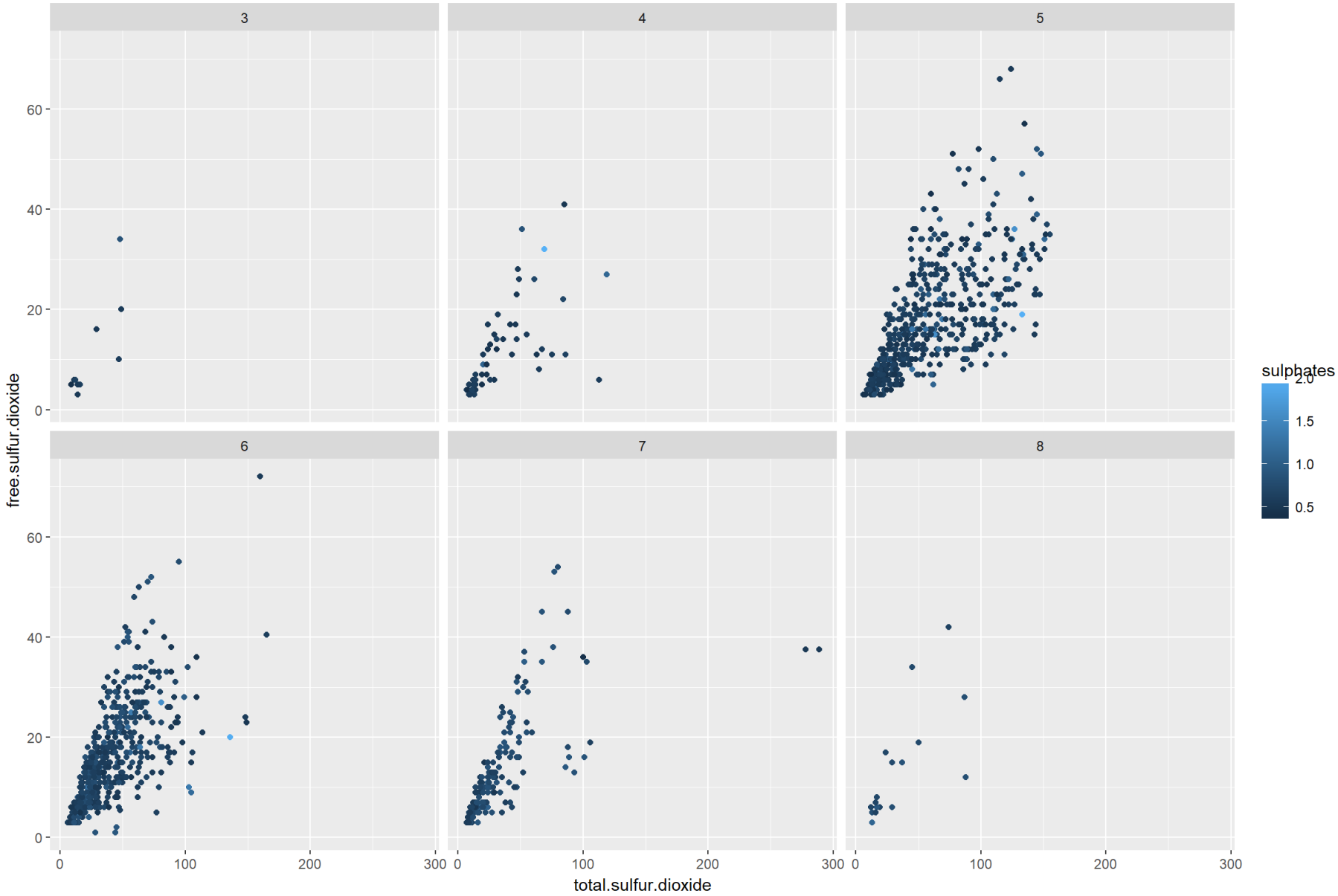
高品质红酒多集中在挥发酸低的区域的区域.

柠檬酸和酸类在不同质量下的相关关系:



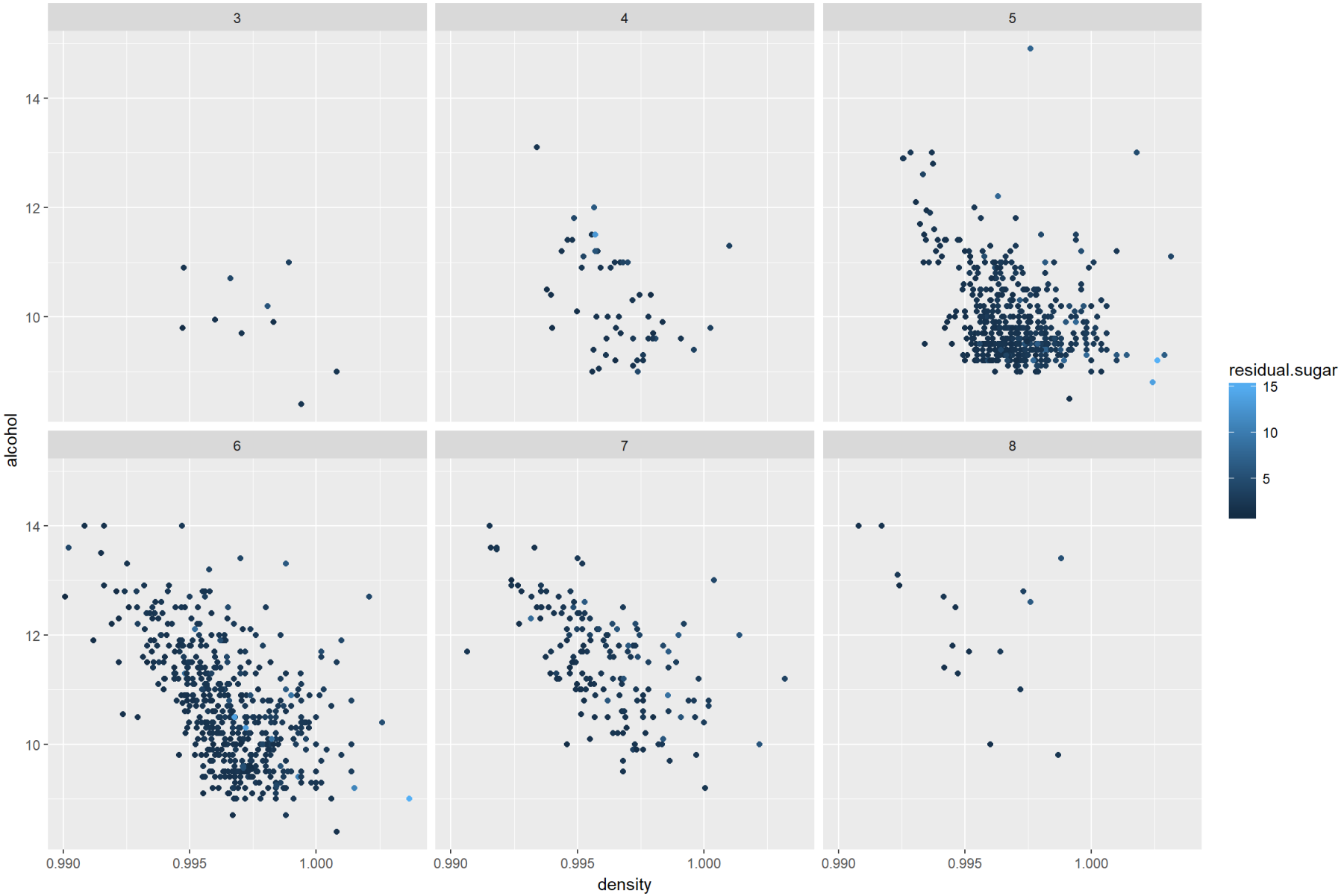
高柠檬酸总出现在挥发酸少的区域,并倾向于出现在高固定酸区域.

总二氧化硫,游离二氧化硫和硫化物在不同质量下的相关关系:



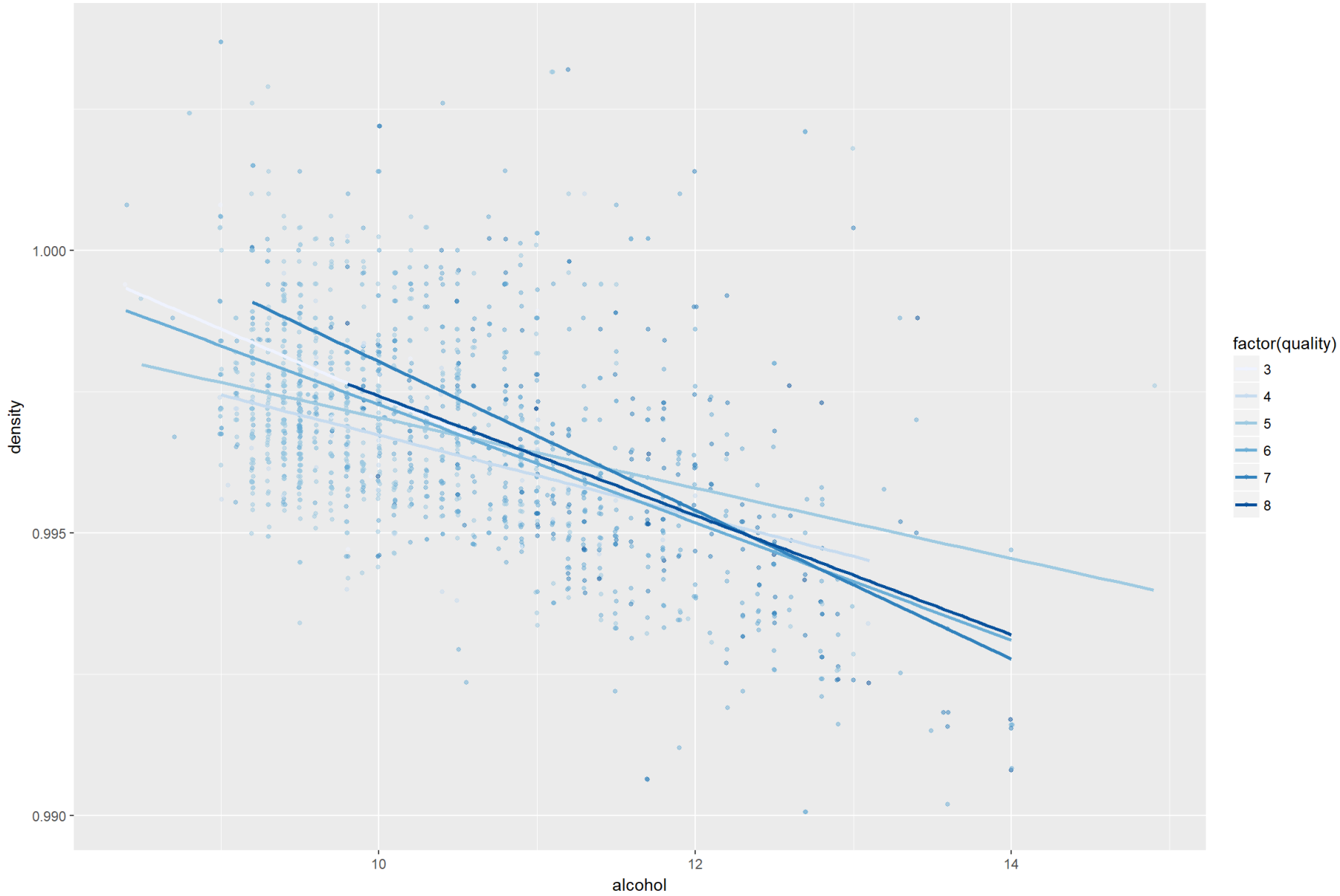
总二氧化硫和游离二氧化硫正相关

密度酒精和残糖在不同质量下的相关关系:



高质量酒出现在酒精度高区域,高残糖出现在高密度区域.

酒精和密度在不同质量下的相关关系:



不同质量的红酒,都是酒精和密度呈负相关.

多变量分析

探讨你在这部分探究中观察到的一些关系,通过观察感兴趣的特性,是否存在相互促进的特性?

高品质红酒多集中在挥发酸低的区域的区域.

高柠檬酸总出现在挥发酸少的区域,并倾向于出现在高固定酸区域.

总二氧化硫和游离二氧化硫正相关,未发现和硫化物的关系.

高质量酒出现在酒精度高区域,高残糖出现在高密度区域.

不同质量的红酒,都是酒精和密度呈负相关.

猜测柠檬酸和固定酸有促进作用,总二氧化硫和游离二氧化硫有促进作用.

猜测高残糖和低酒精对高密度有促进作用.

这些特性之间是否存在有趣或惊人的联系呢?

残糖不会影响酒精度.

没有挥发酸和固定酸同时特别高的红酒,也许二者可以单向转化.

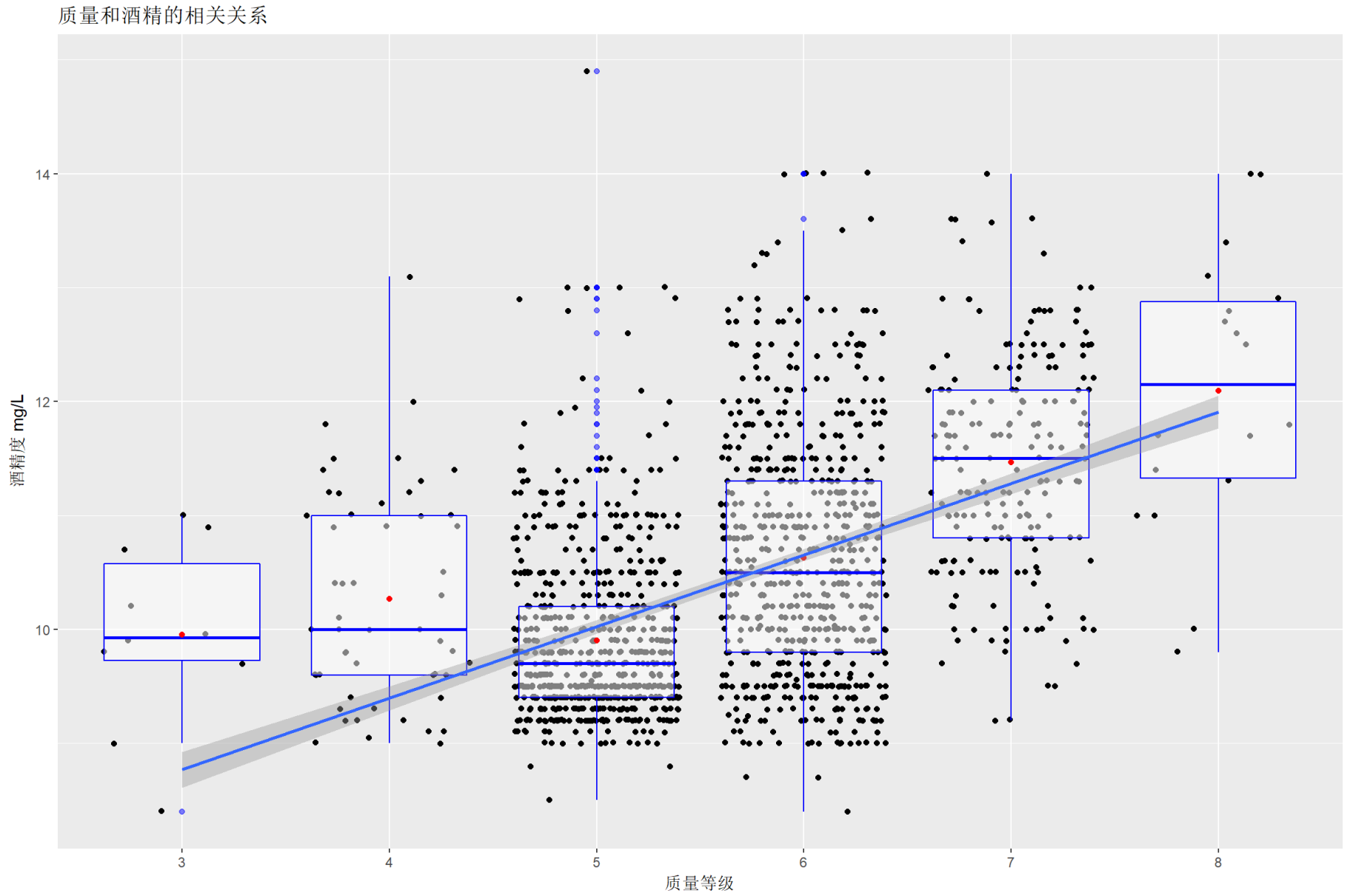
柠檬酸可能是固定酸的一种.

总二氧化硫可能生成游离二氧化硫.

密度也许由残糖和酒精决定.

定稿图和总结

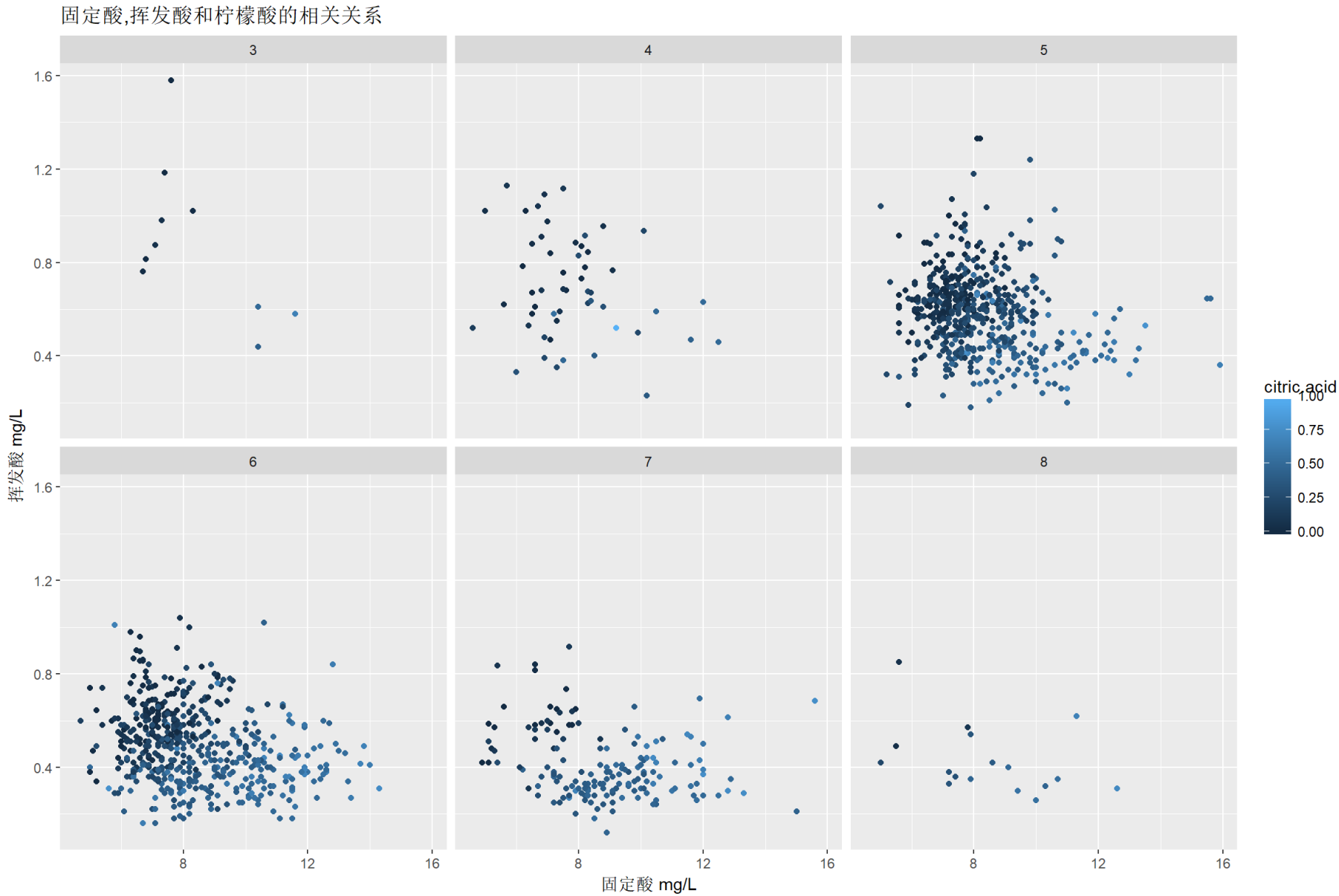
绘图一:



描述一:

质量和酒精度呈正相关

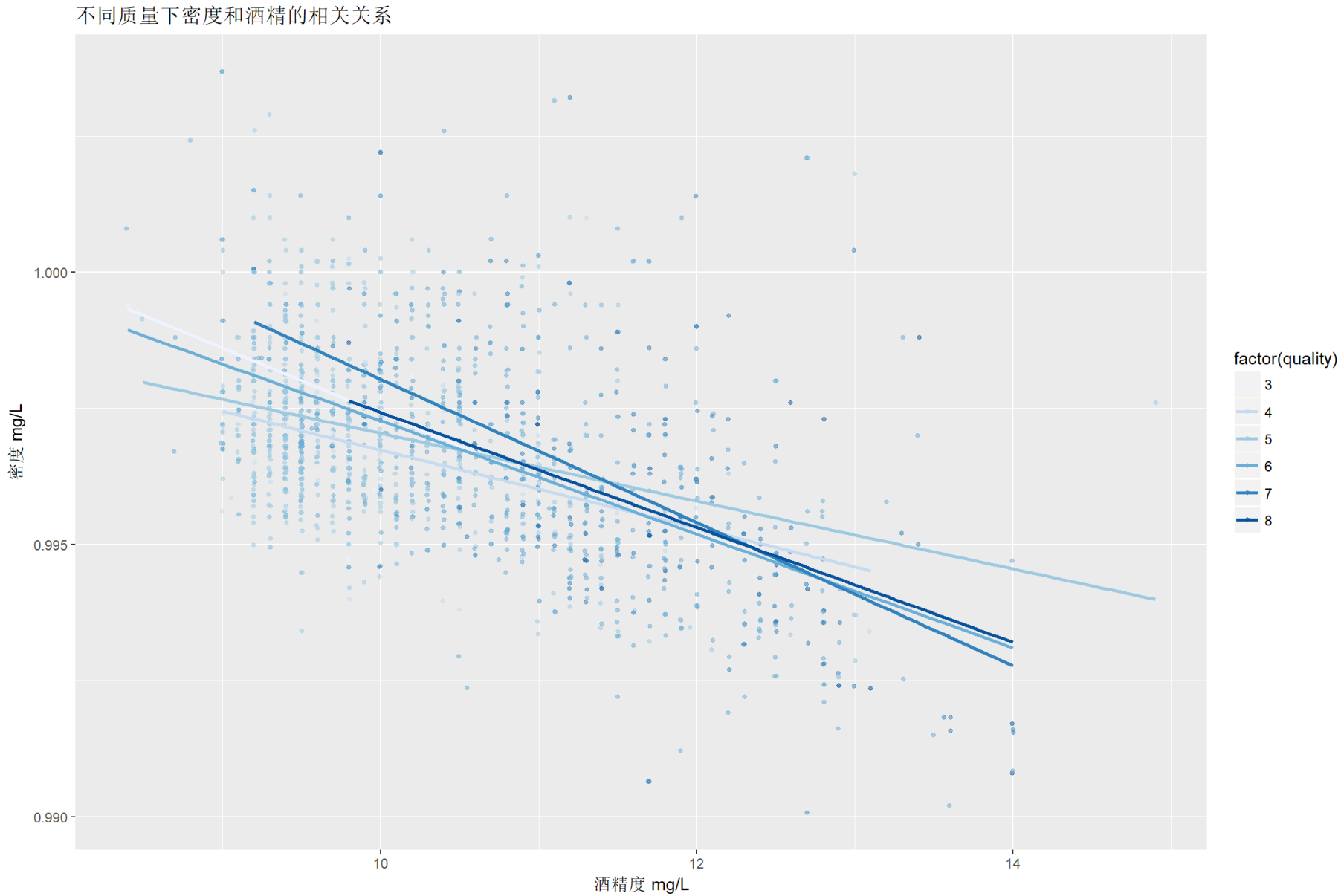
绘图二:



描述二:

高柠檬酸总出现在挥发酸少的区域,并倾向于出现在高固定酸区域.

绘图三:


不同质量下密度和酒精的相关关系
密度 mg/L
1.000
0.995
0.990
10
12
14
酒精度 mg/L
factor(quality)
3
4
5
6
7
8

描述三:

不同质量的红酒,都是酒精和密度呈负相关.

反思:

分析过程中,特别依赖于散点图矩阵和相关系数.思路不够开阔.