

红酒数据集探索 by 张逸松

这个报告探索了一个包含大约1600种红酒和相关化学成分的属性，并相应被专家评分的数据集。

单变量绘图选择

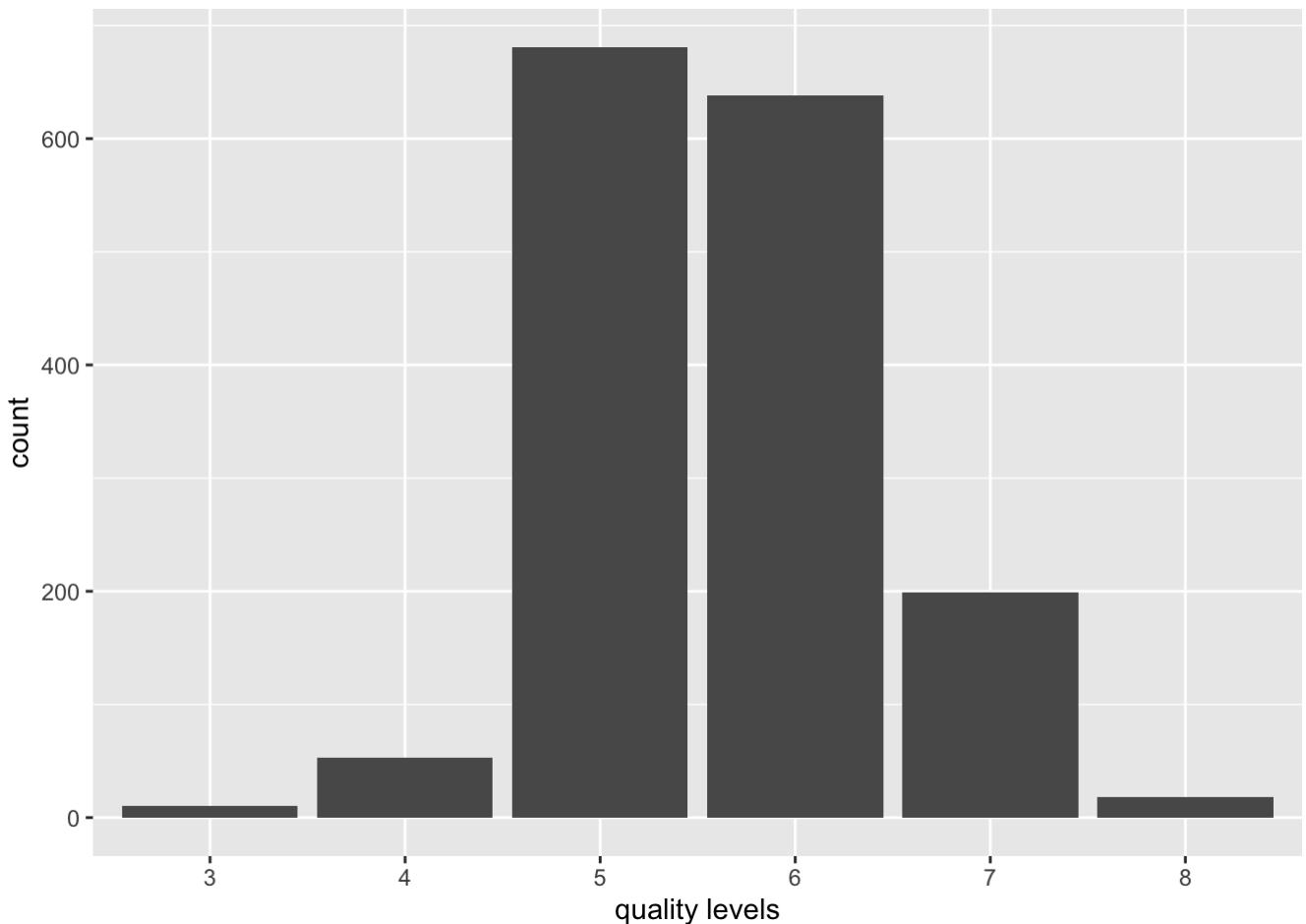
```
## [1] 1599 12
```

```
## '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 0.07
## $ 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 ...
```

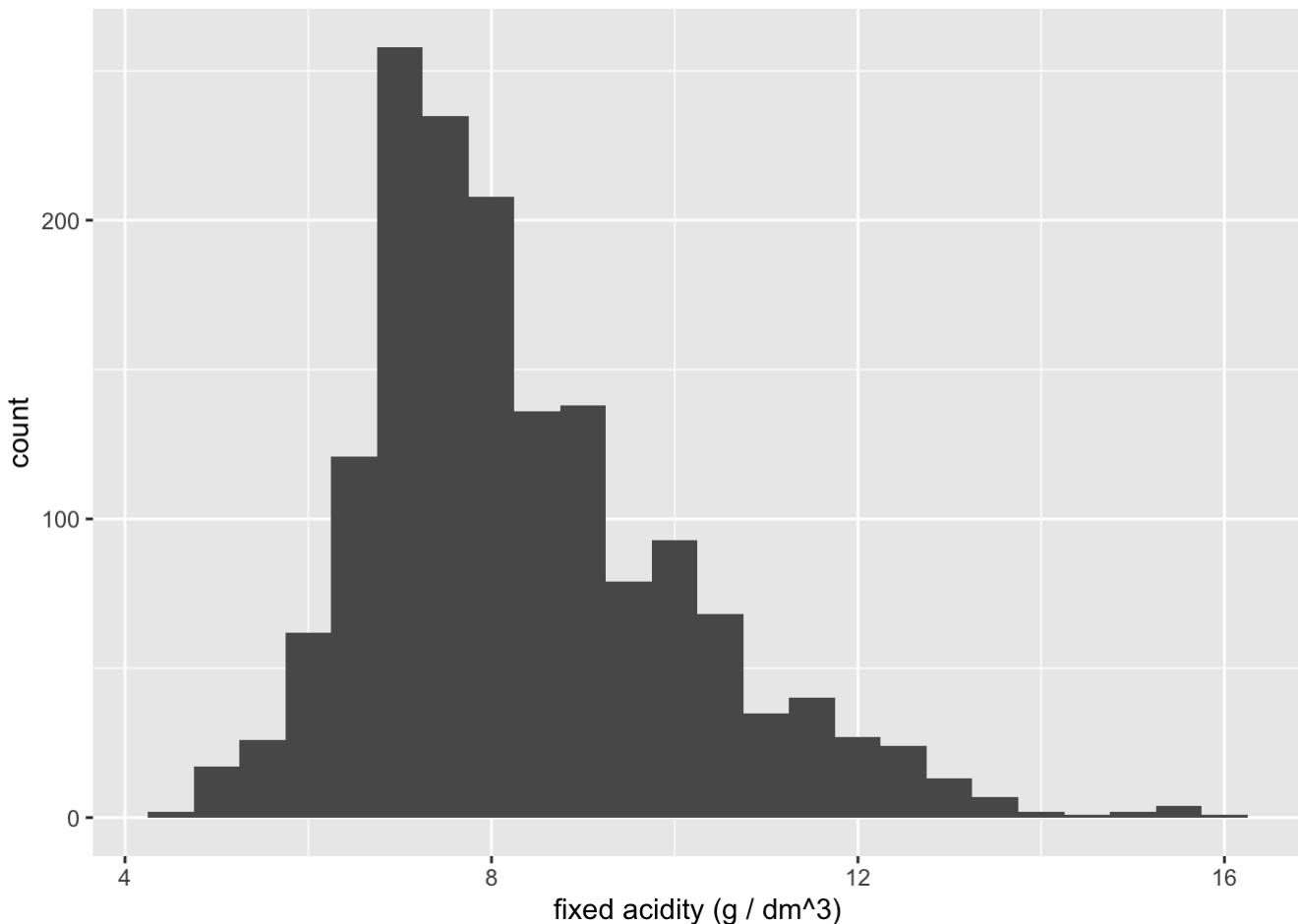
```
## fixed.acidity    volatile.acidity    citric.acid    residual.sugar
## Min. : 4.60      Min. :0.1200      Min. :0.000      Min. : 0.900
## 1st Qu.: 7.10    1st Qu.:0.3900    1st Qu.:0.090      1st Qu.: 1.900
## Median : 7.90    Median :0.5200      Median :0.260      Median : 2.200
## Mean   : 8.32    Mean   :0.5278      Mean   :0.271      Mean   : 2.539
## 3rd Qu.: 9.20    3rd Qu.:0.6400      3rd Qu.:0.420      3rd Qu.: 2.600
## Max.   :15.90    Max.   :1.5800      Max.   :1.000      Max.   :15.500
## chlorides       free.sulfur.dioxide total.sulfur.dioxide
## Min. :0.01200    Min. : 1.00      Min. : 6.00
## 1st Qu.:0.07000   1st Qu.: 7.00      1st Qu.: 22.00
## Median :0.07900   Median :14.00      Median : 38.00
## Mean   :0.08747   Mean   :15.87      Mean   : 46.47
## 3rd Qu.:0.09000   3rd Qu.:21.00      3rd Qu.: 62.00
## 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
```

```
## '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 0.07
## $ 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              : Factor w/ 6 levels "3","4","5","6",...: 3 3 3 4 3 3 3 5 5
## $ ...
```

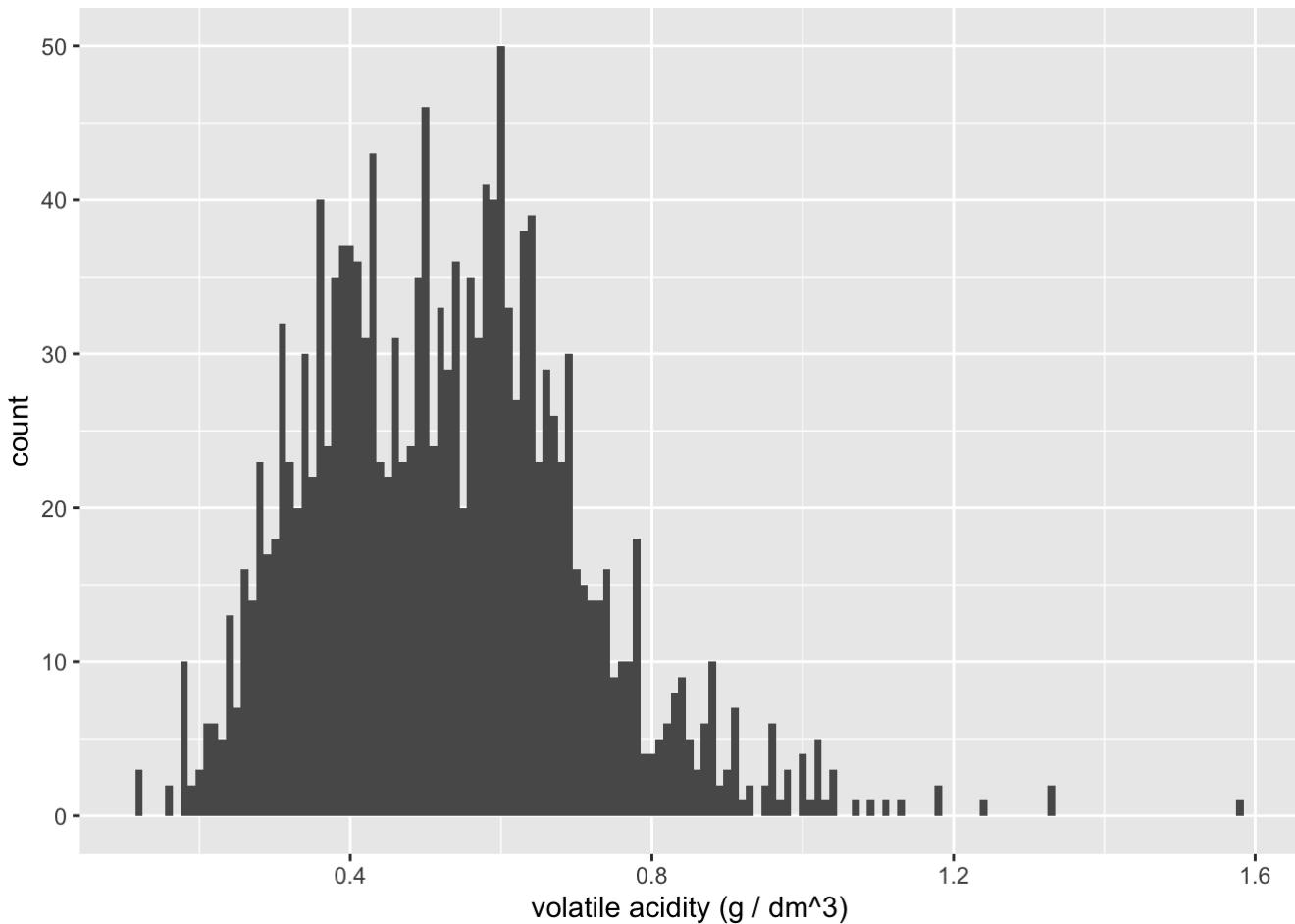
这个数据集有1599个观察对象，12个变量。



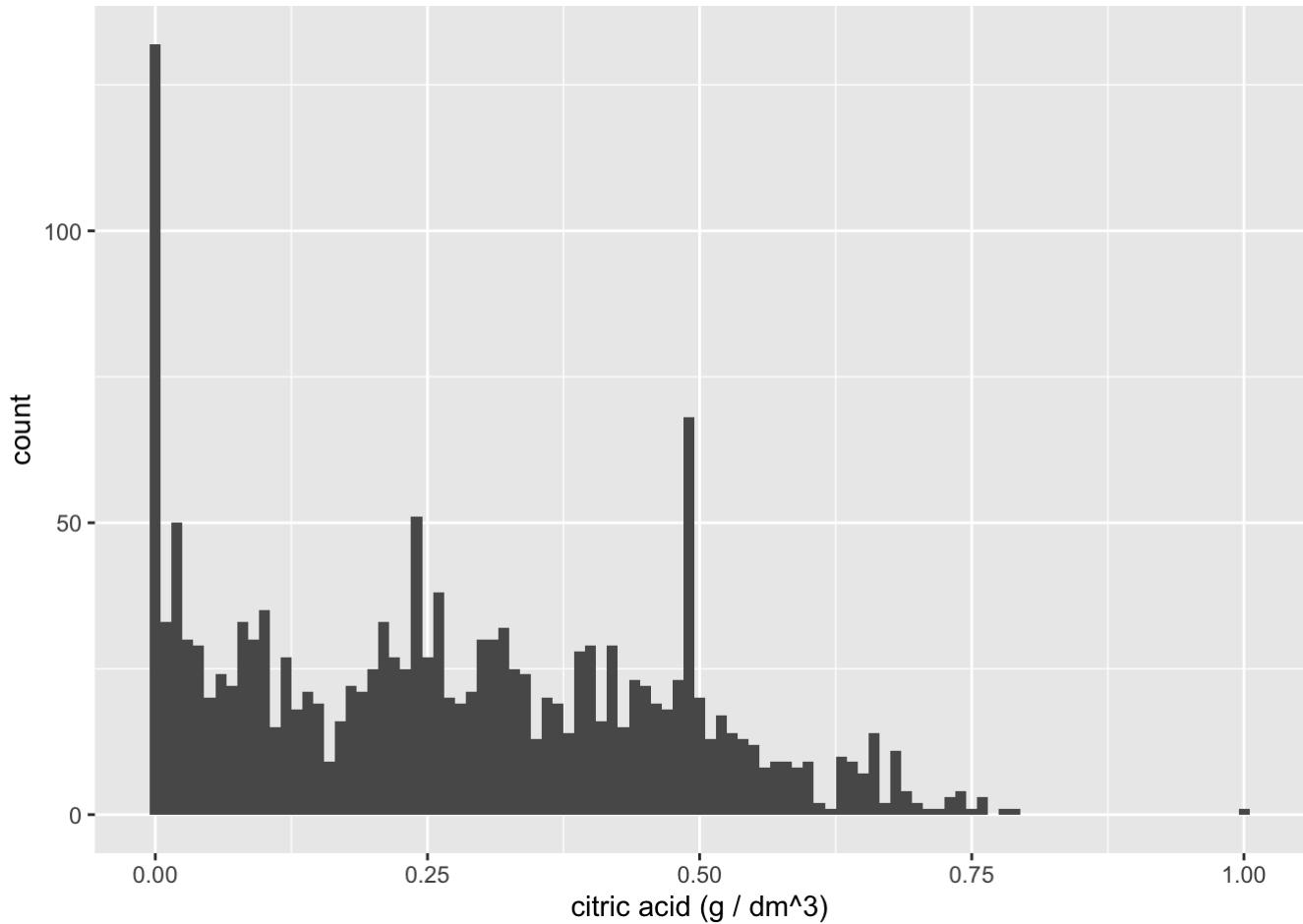
quality的直方图是正态分布，大多数红酒的评分都在5, 6分。



fixed acidity的分布接近正态分布。



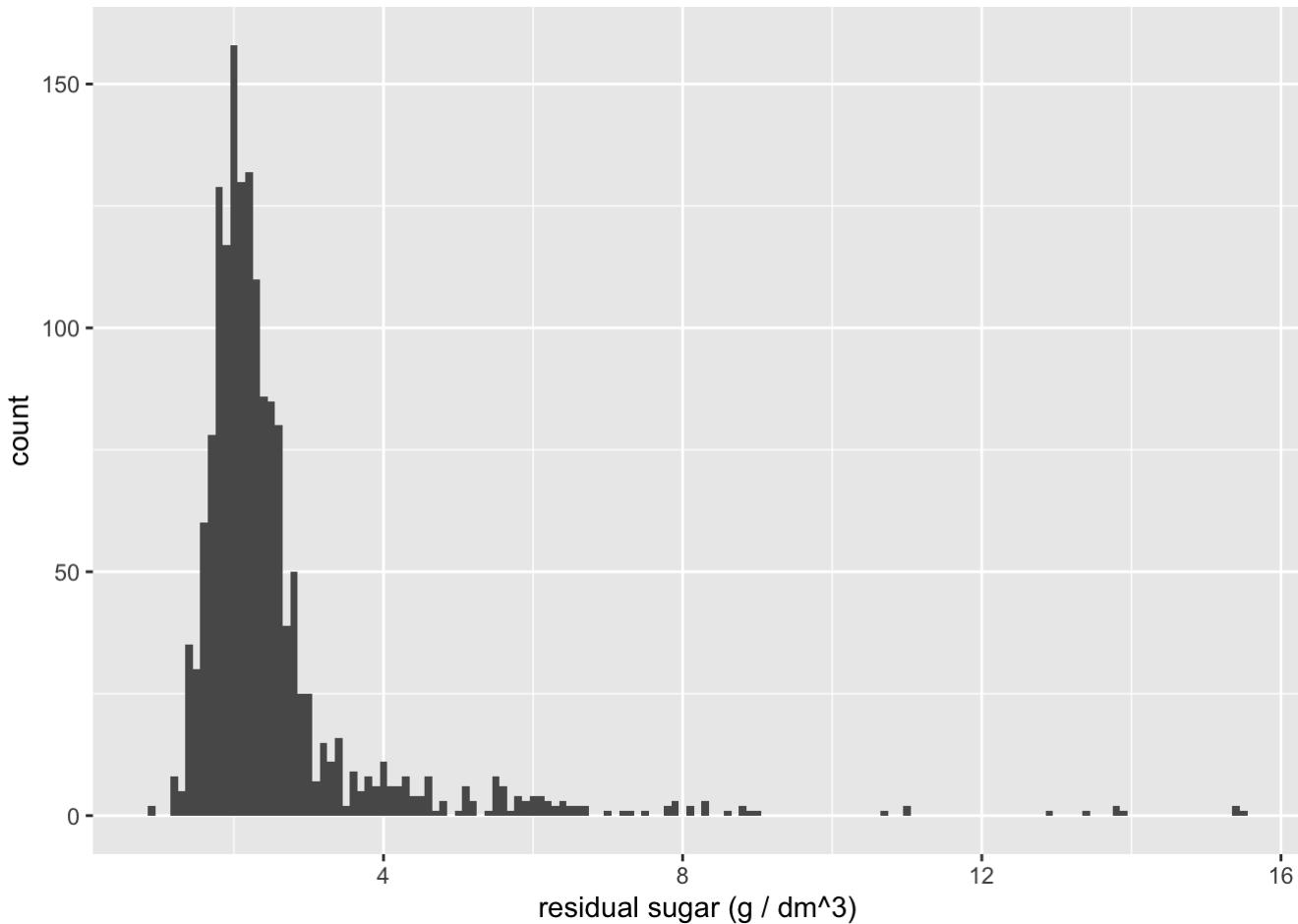
volatile acidity太高的话会导致葡萄酒有不好的酸的口感，可以从图中看出来大多数的葡萄酒都处于正常的范围，只有一小部分的值偏高。



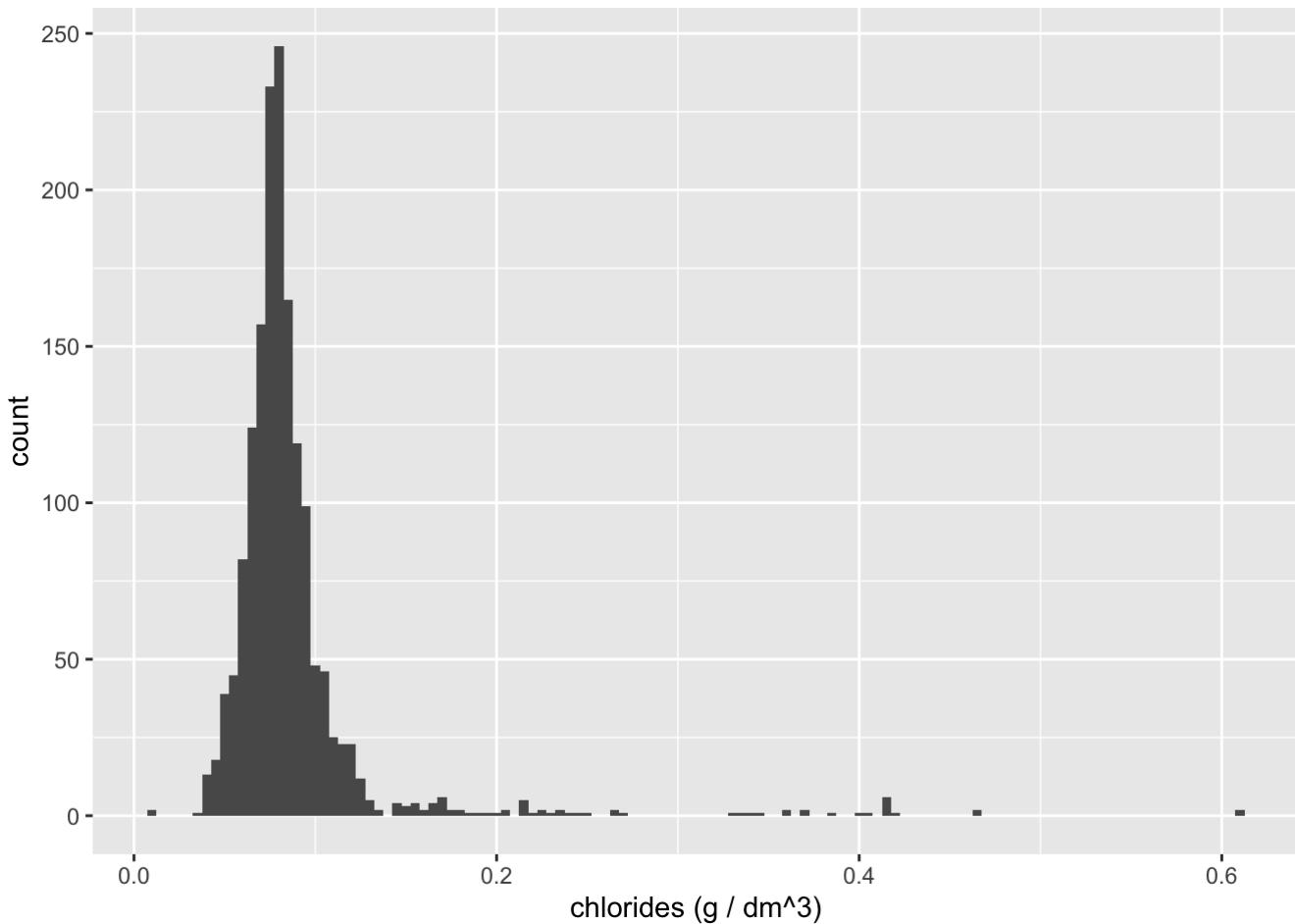
citric acid 可以增加葡萄酒的新鲜口感，可以看到有两个双峰结构，第一个在0左右，第二个在0.5左右，我猜测是因为葡萄酒的档次问题，分布在0左右的可能价格较低，在0.5左右的葡萄酒属于高档酒，价格较高。

```
## [1] 0.9 0.9
```

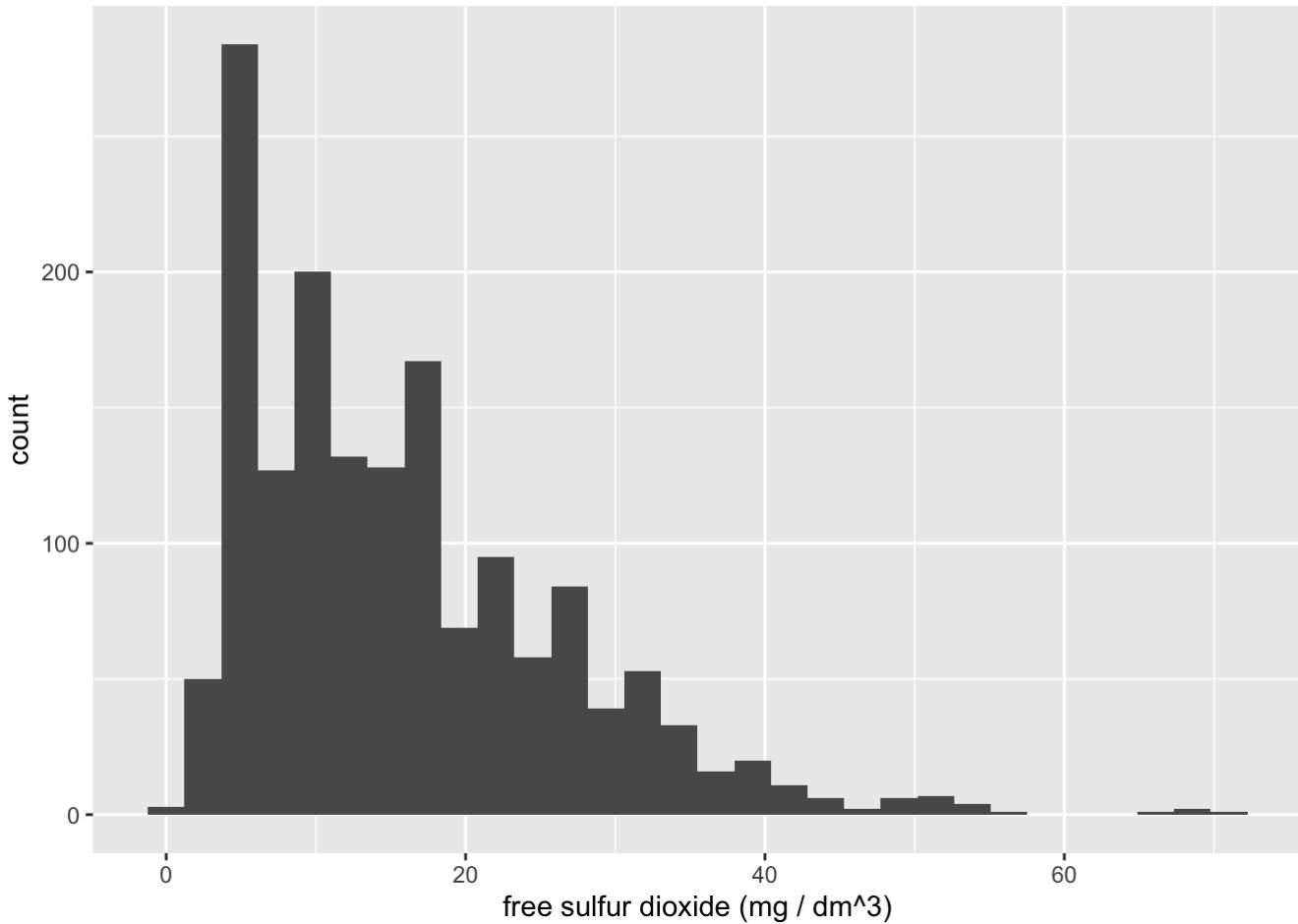
```
## numeric(0)
```



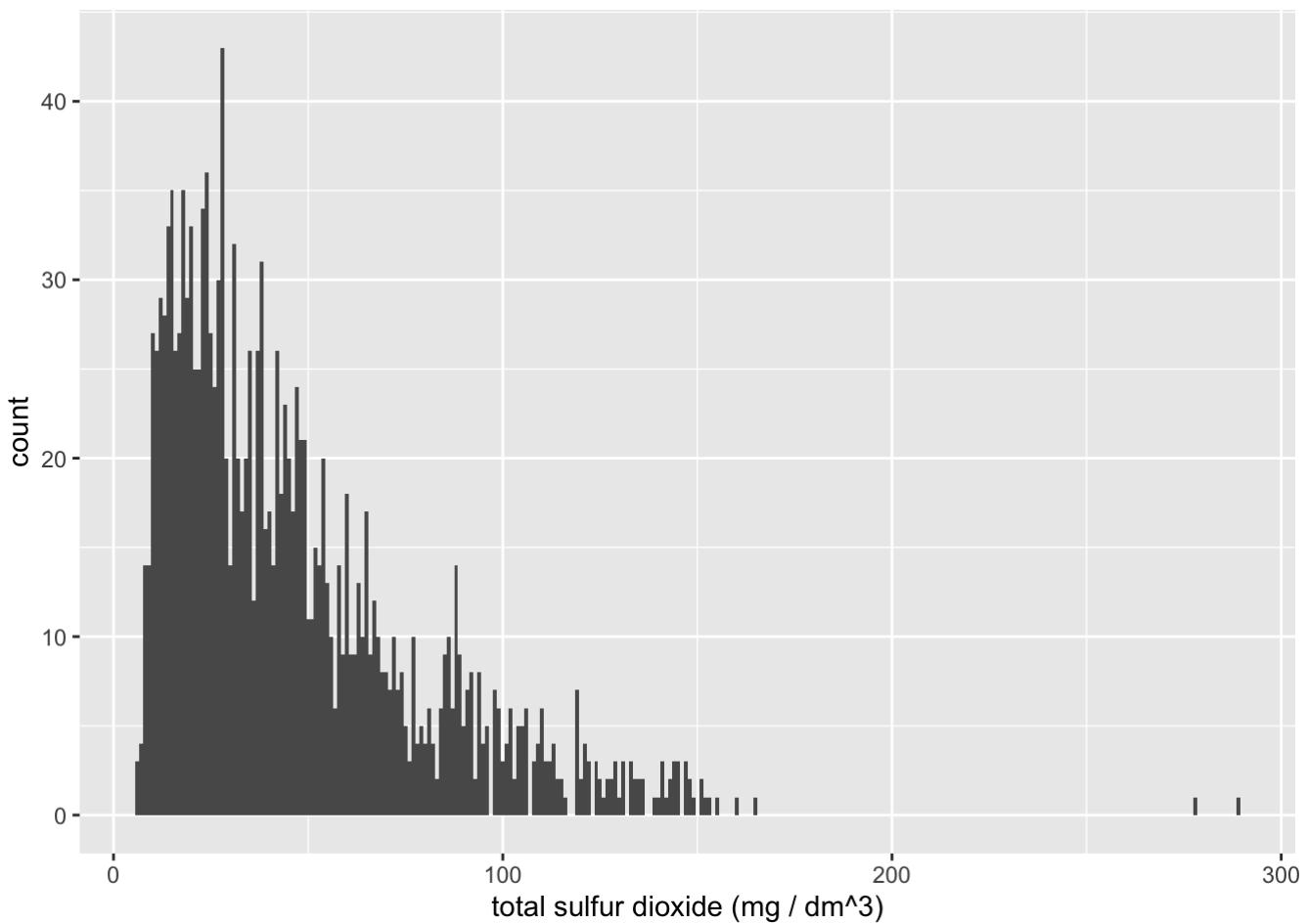
可以看到residual sugar在0到4的分布近似正太分布，后面的数据比较稀少，零散的分布在长尾上面。只有2种红酒的residual sugar小于1，没有大于45的红酒，也就是说在这个数据集里面没有红酒是甜的。



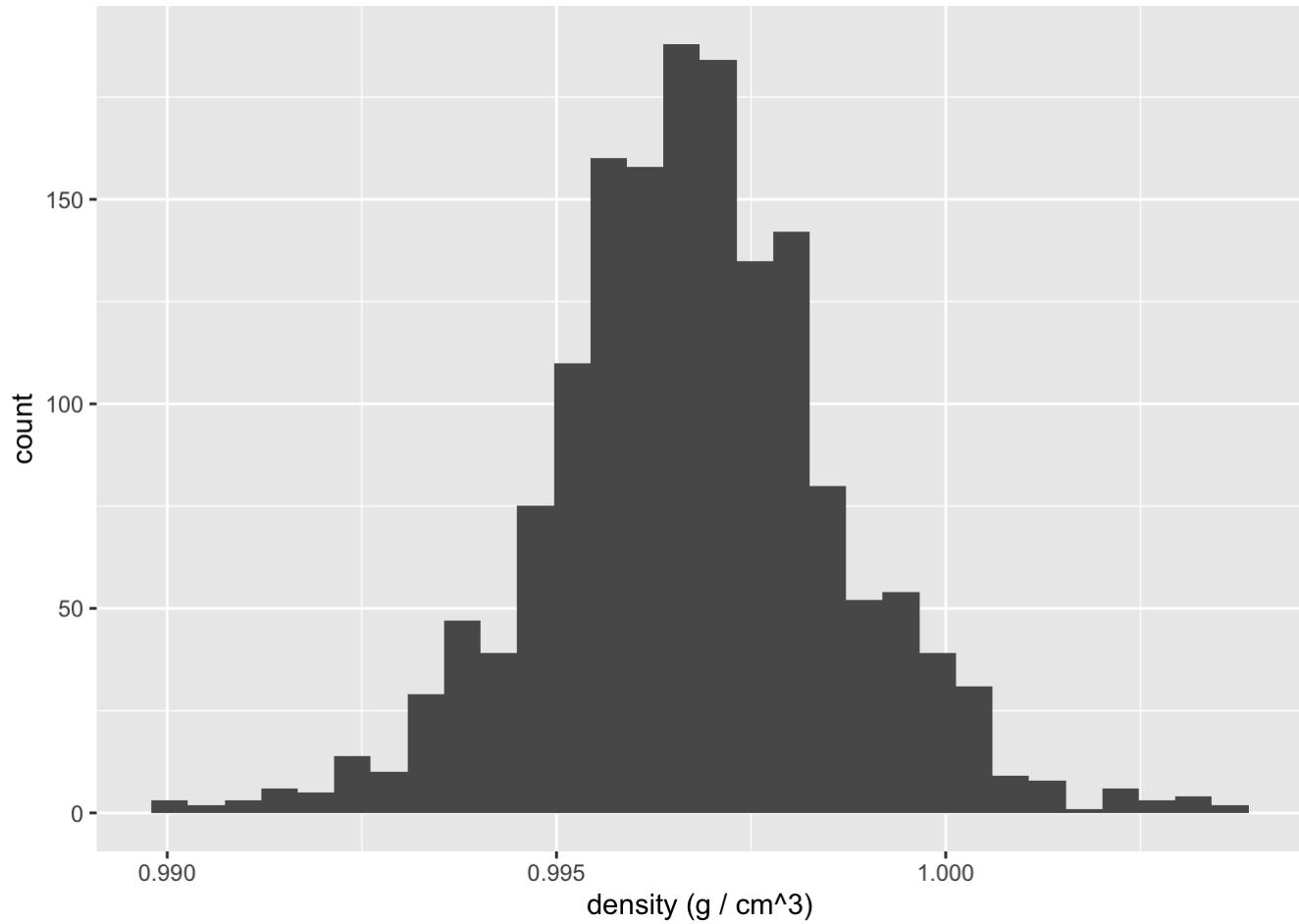
可以看到chlorids在0到0.2的分布近似正太分布，后面的数据比较稀少，零散的分布在长尾上面。



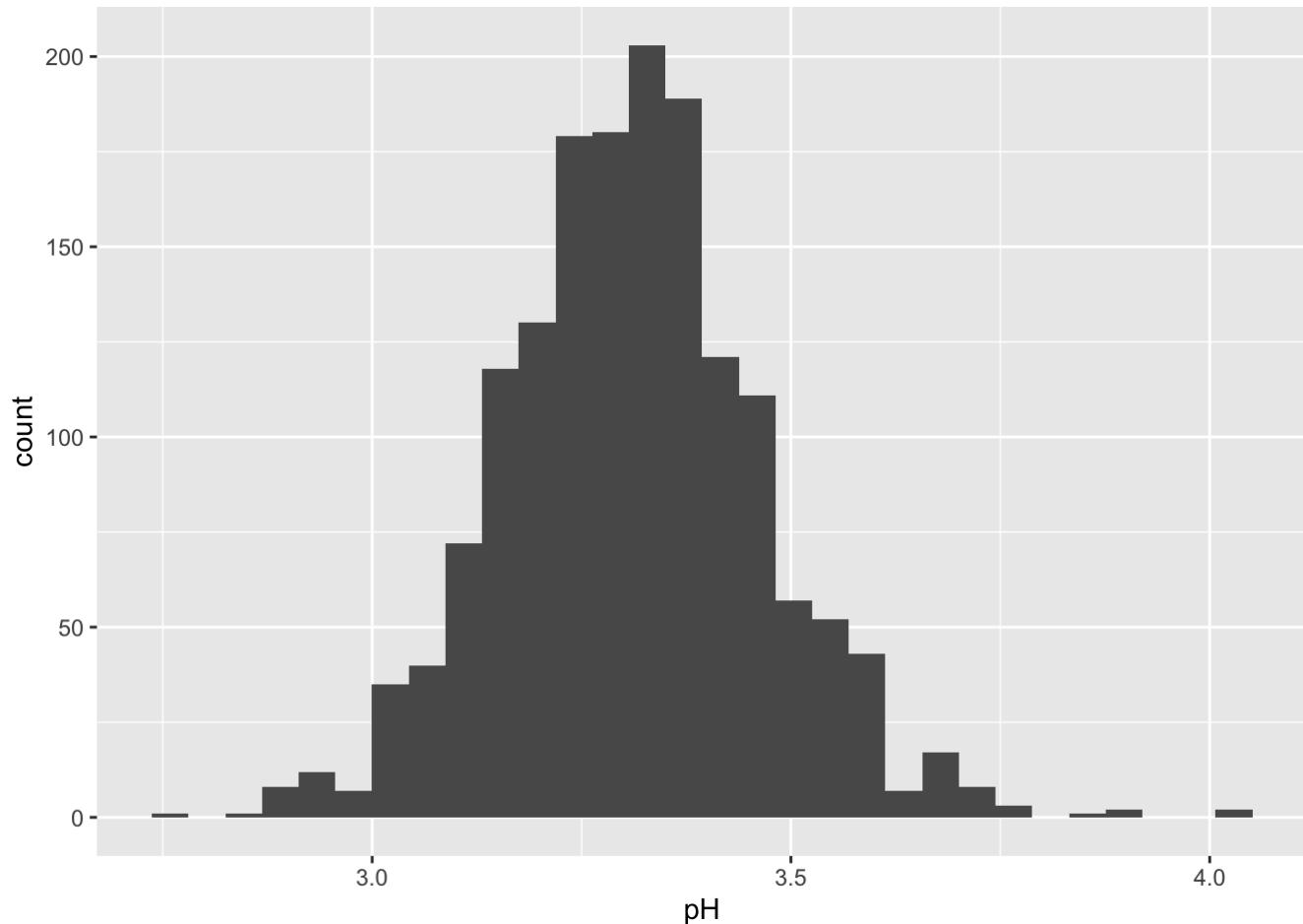
可以看到这幅图是偏向右偏态的， free sulfur dioxide可以防止微生物的生长和葡萄酒的氧化。



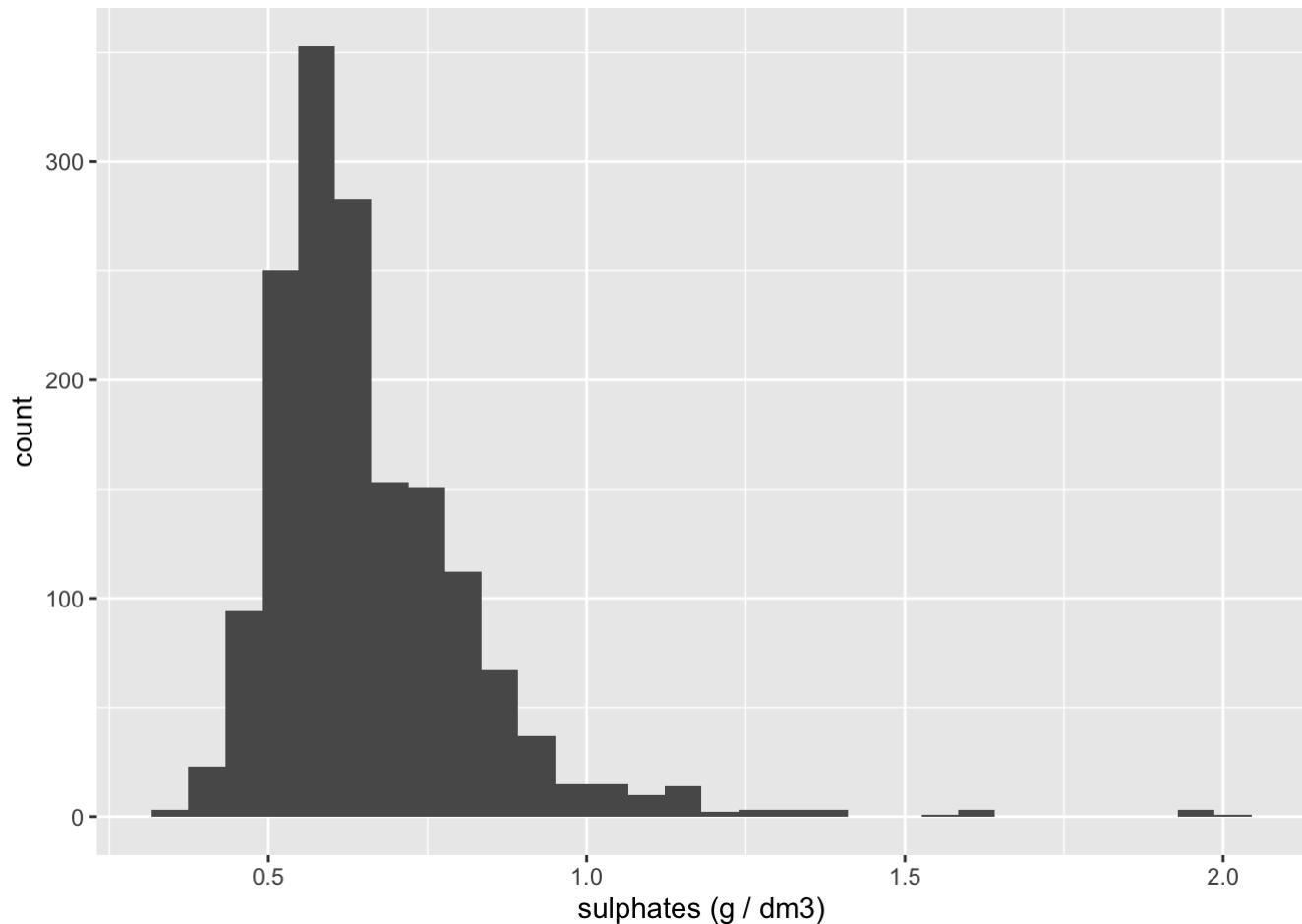
total sulfur dioxide的分布偏右偏态，有个别的值分散在很远处。



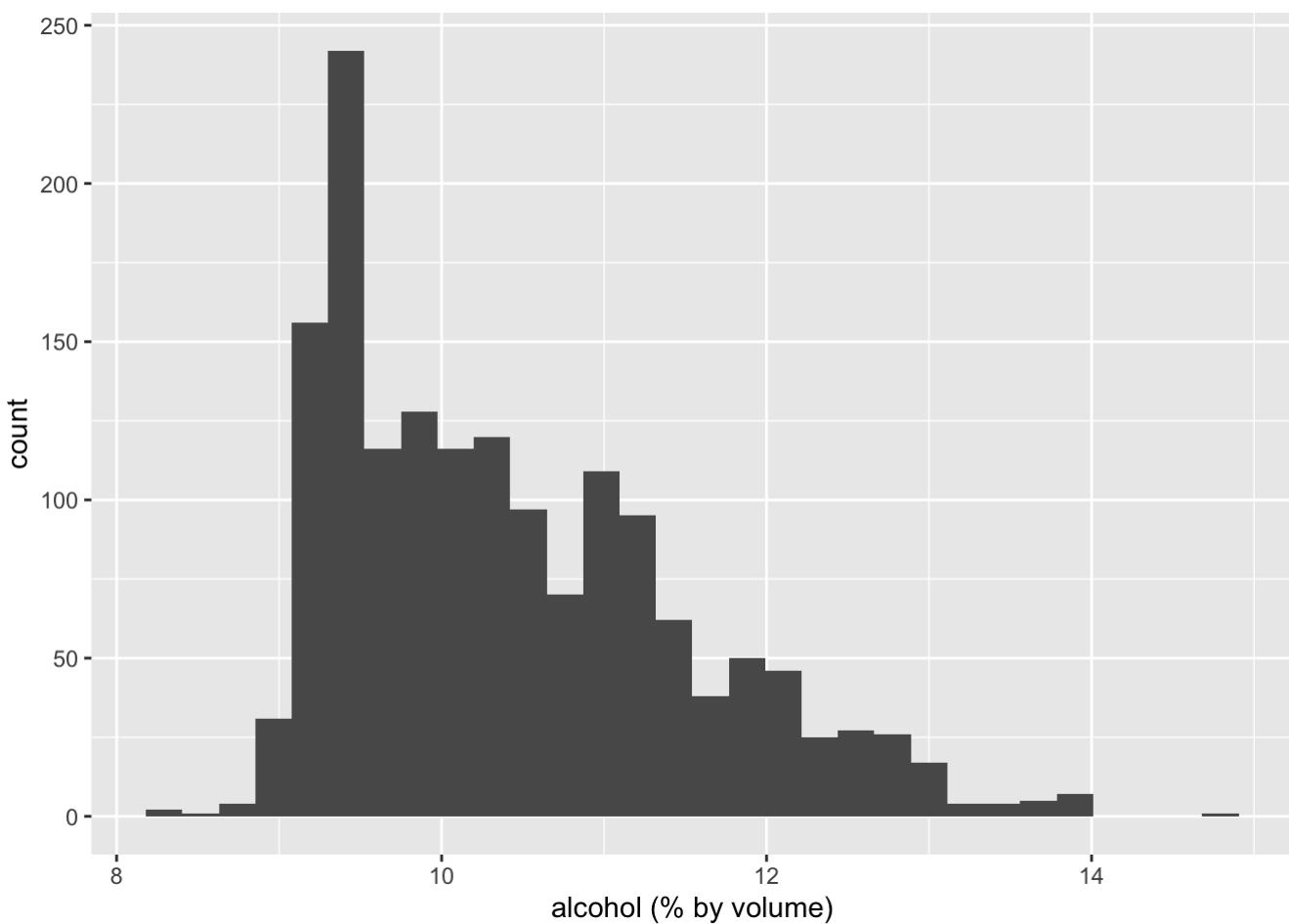
density的分布偏正态分布。



pH属于正态分布。虽然说大多数葡萄酒的酸度都处在3到4之间，但是我们的数据集整体酸度明显偏酸一些。接近4的酸度数量的葡萄酒明显较少。



这个属于右偏态。看似觉得sulphates的含量都处在较少的含量，有少部分葡萄酒会把含量提高。



alcohol的分布也是右偏态。酒精含量越高的看似越少。

单变量分析

你的数据集结构是什么？

这个数据集有1599种红酒，有12个特性。

你的数据集中感兴趣的主要特性有哪些？

我对quality这个特性比较感兴趣，我想探究哪些特性与quality存在相关性。

你认为数据集中哪些其他特征可以帮助你探索兴趣特点？

volatile acidity, citric acid, residual sugar, free sulfur dioxide, sulphates, pH, density, 这几个特性都能帮助我进入深度探索。这些特性都有着影响葡萄酒口感和质量的关联。

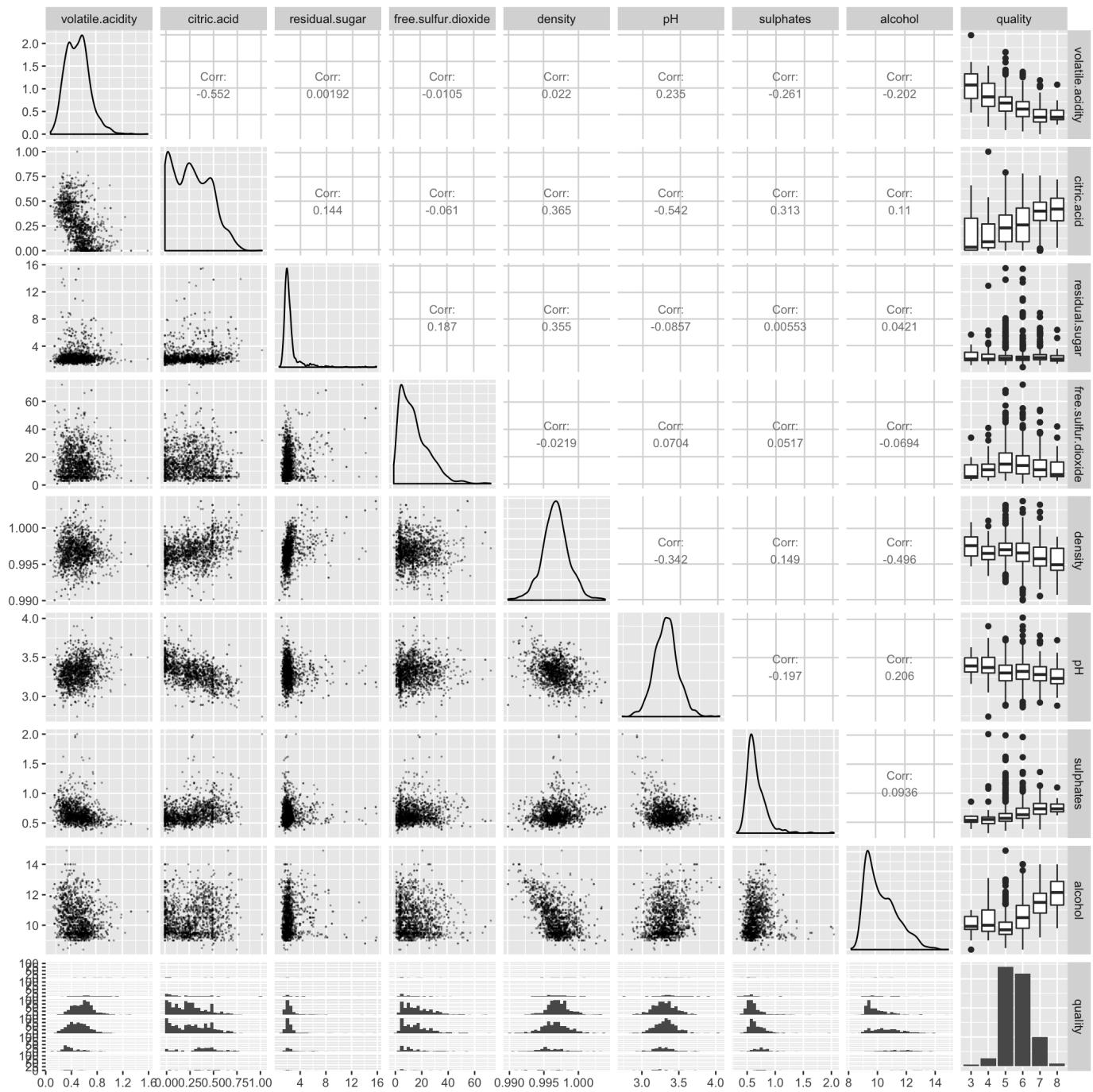
根据数据集中已有变量，你是否创建了任何新变量？

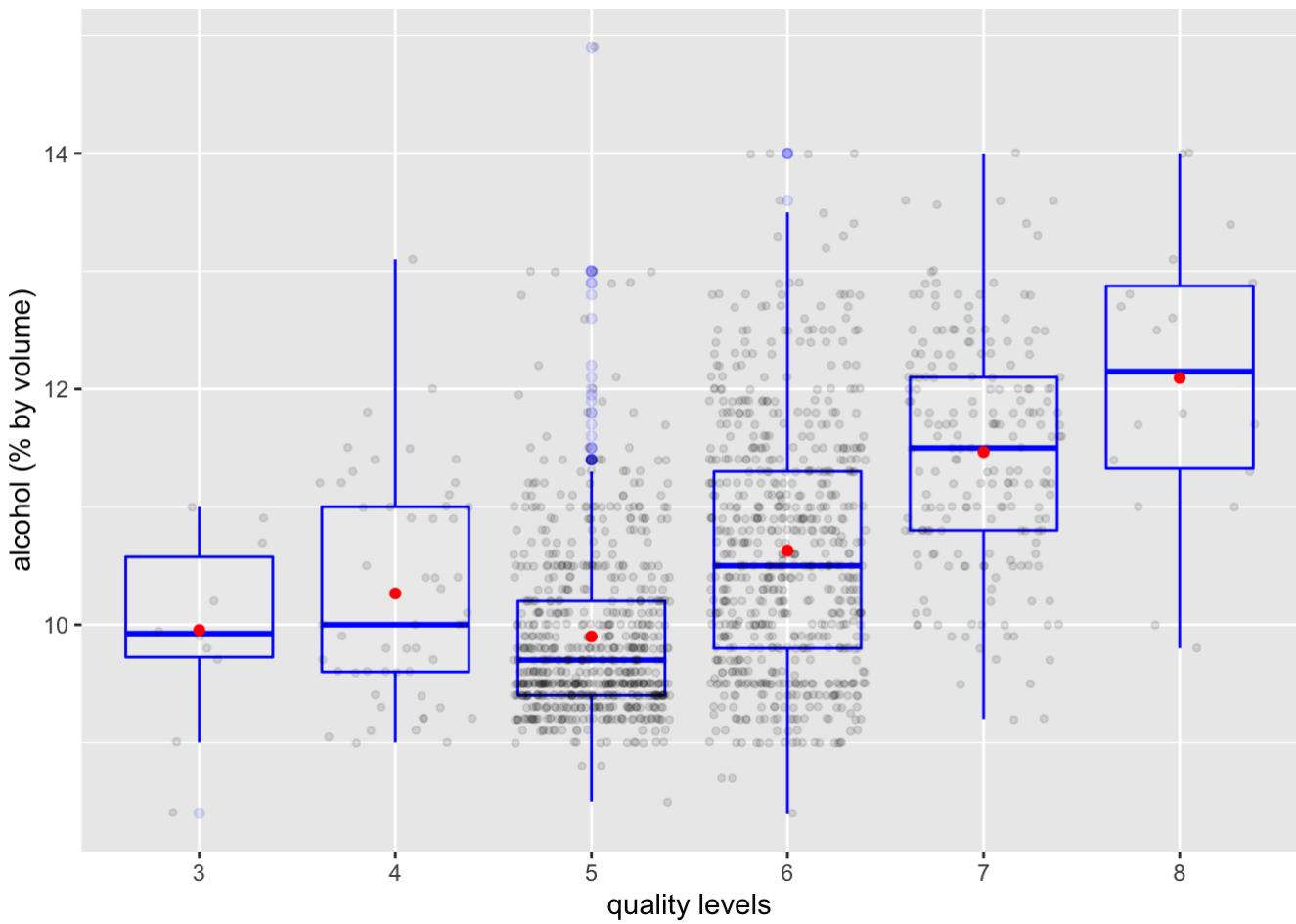
我没有创建新的变量。

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，如清洁、调整或改变数据的形式？如果是，你为什么这样做？

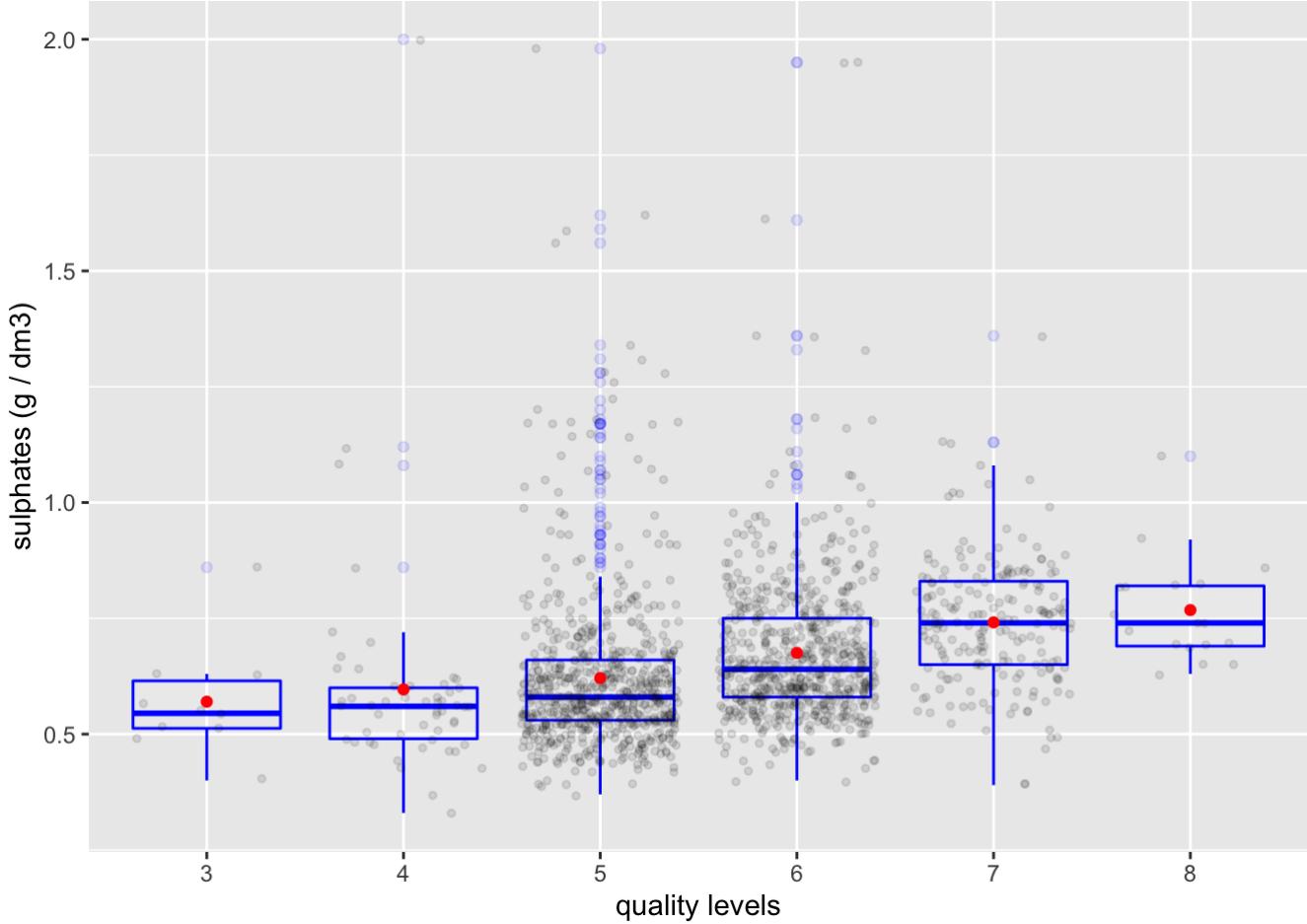
在特性citric acid中我发现了一个双峰结构的分布，我没有操作和修改数据，我猜测是因为葡萄酒的档次问题，分布在0左右的可能价格较低，在0.5左右的葡萄酒属于高档酒，价格较高。在这个数据集里面，quality应该是分类变量，我把它转换成分类变量的属性了。

双变量绘图选择

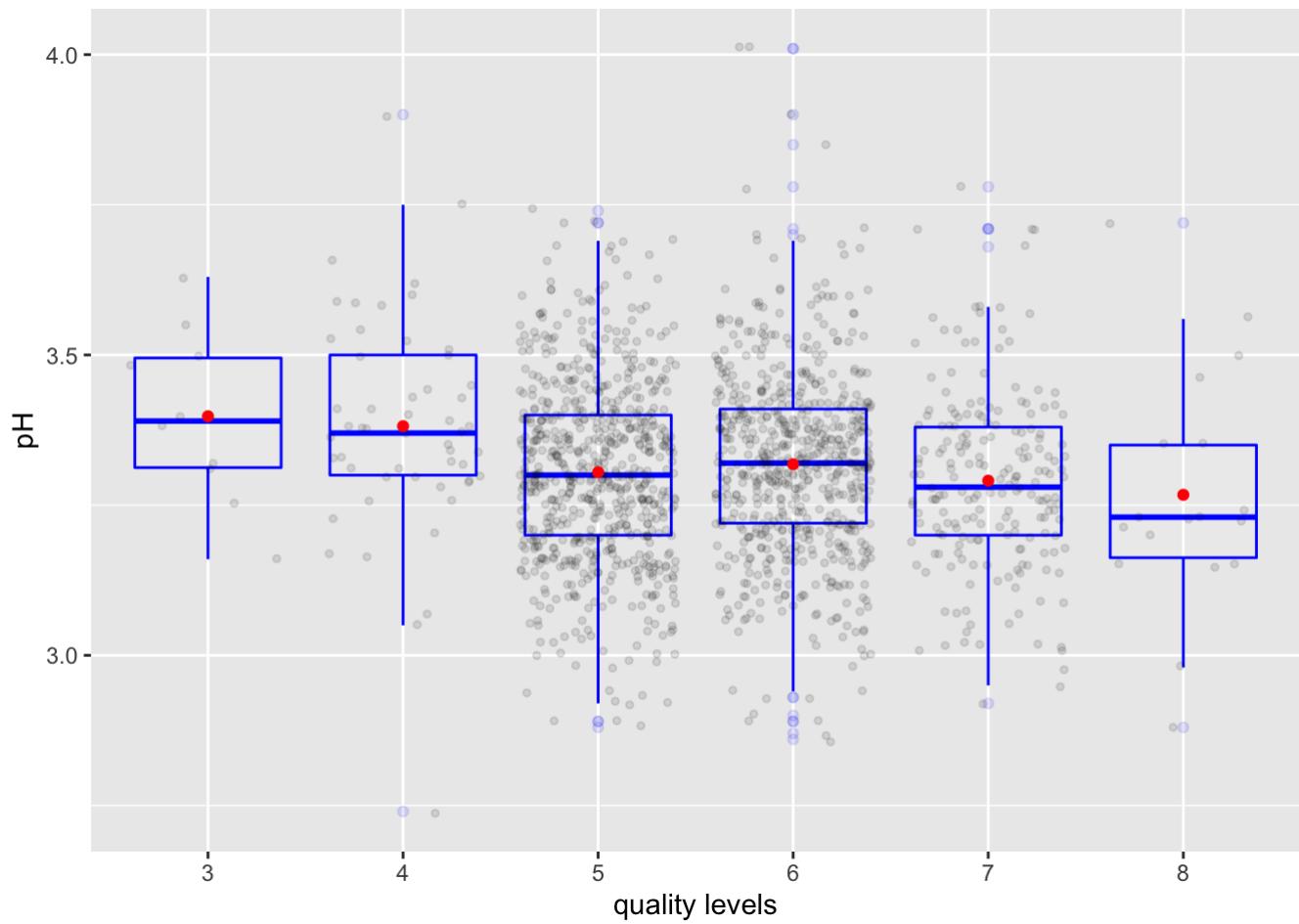




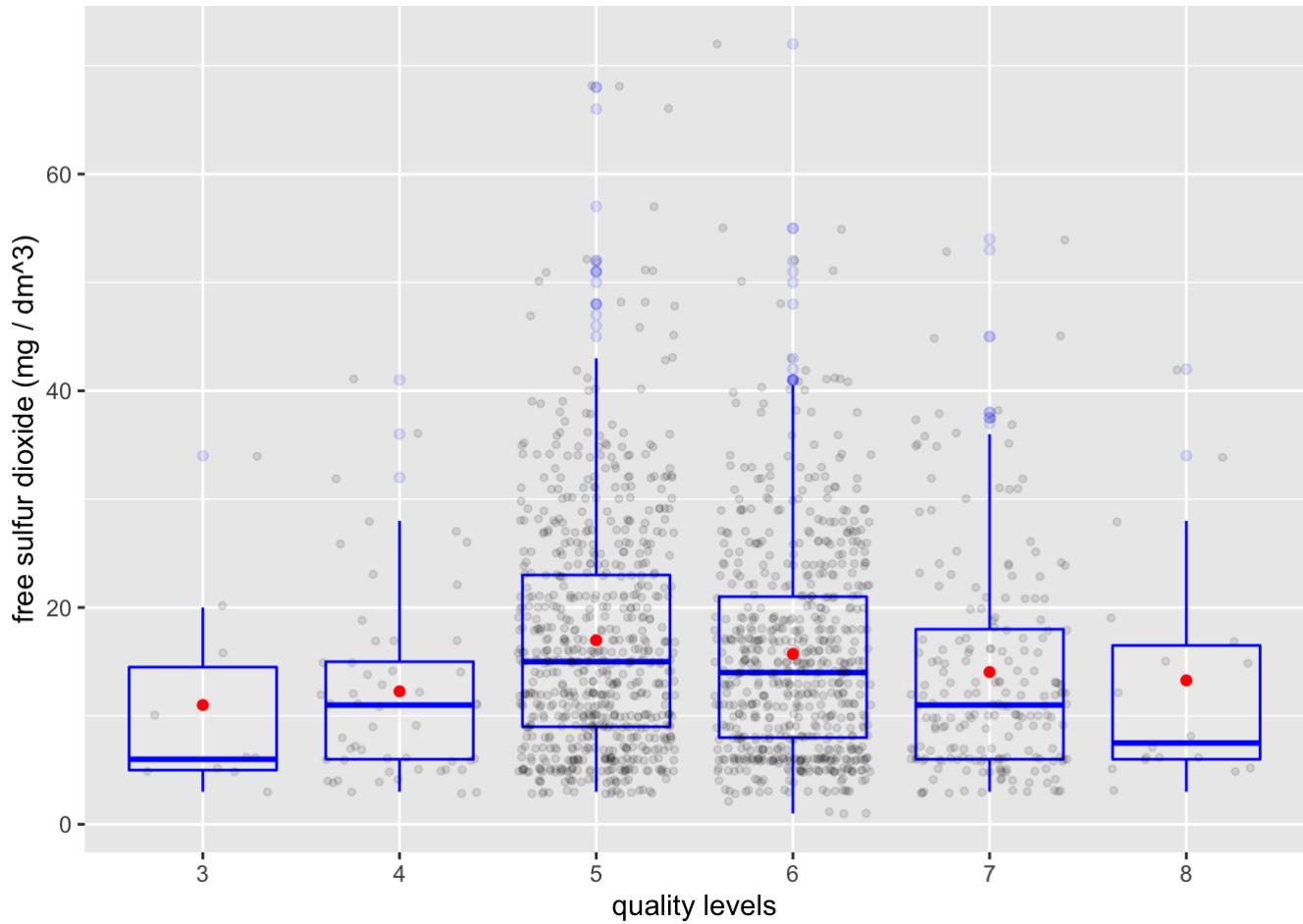
可以看到，酒的质量越高，酒精含量度的均值水平也在显著上升。



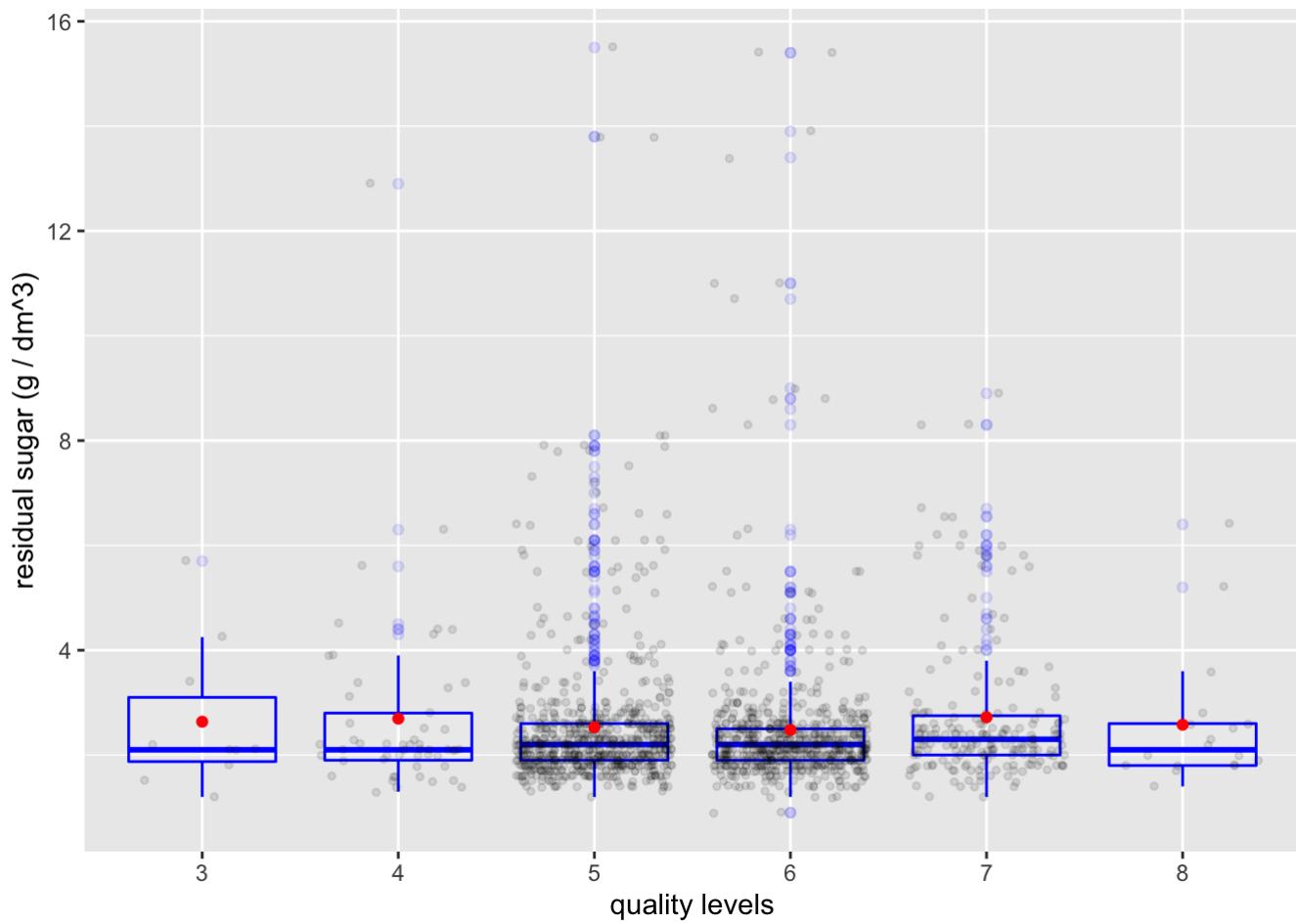
可以看到，随着酒质量的提高，硫酸盐也有轻微的上升趋势。



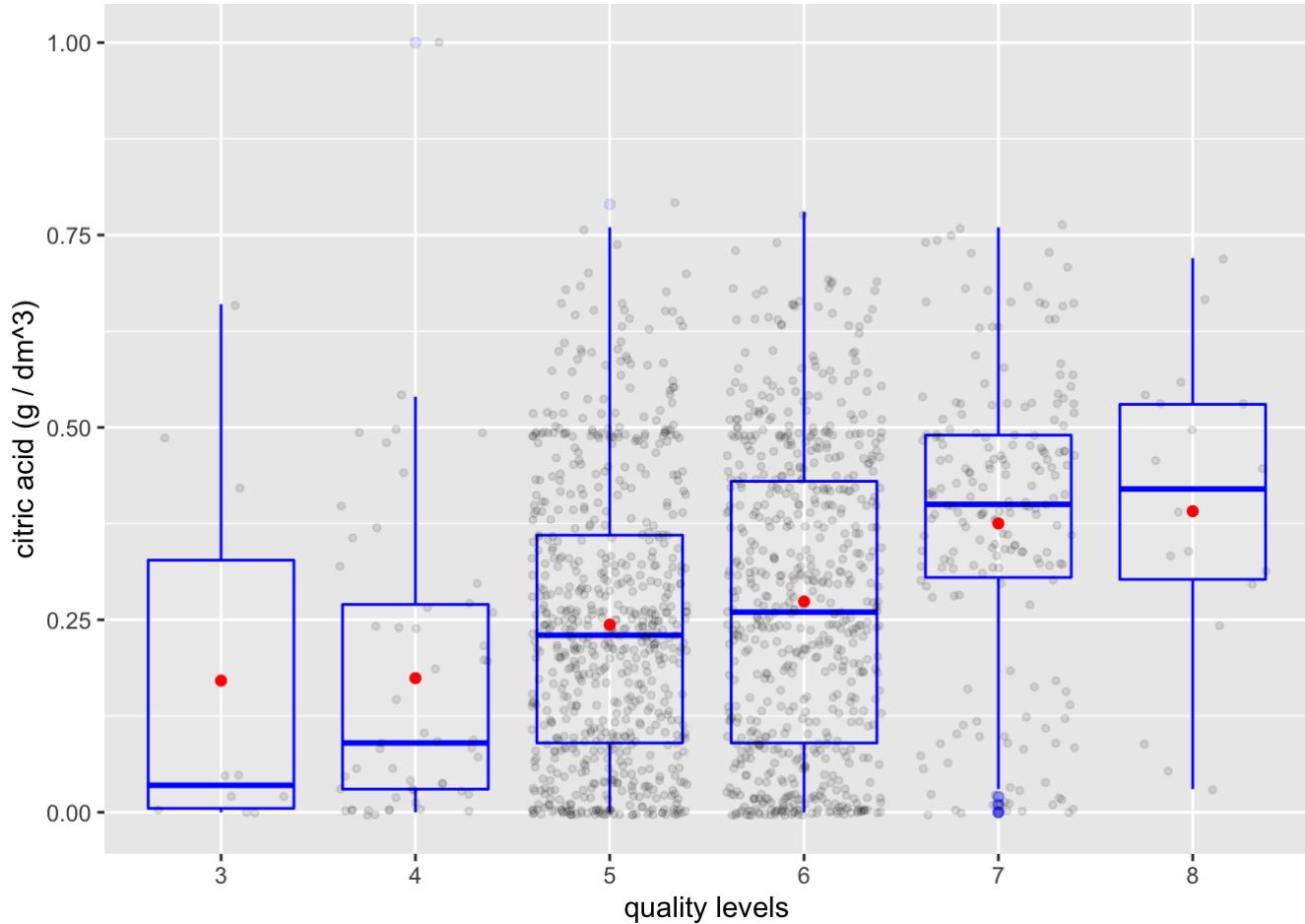
从图中可以看出，pH值和quality几乎没有相关性。



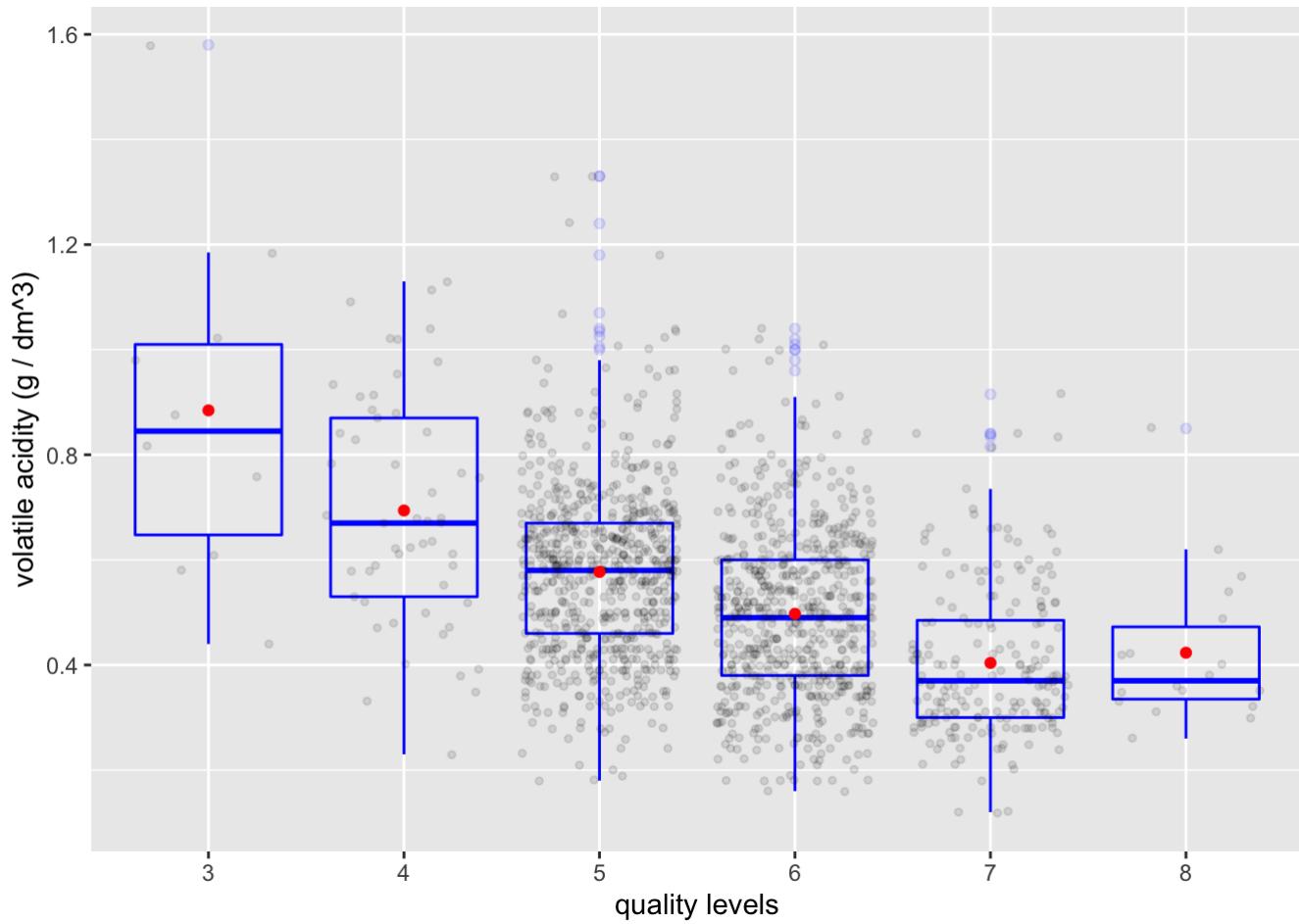
可以看出，free sulfur dioxide与quality几乎没有相关性。



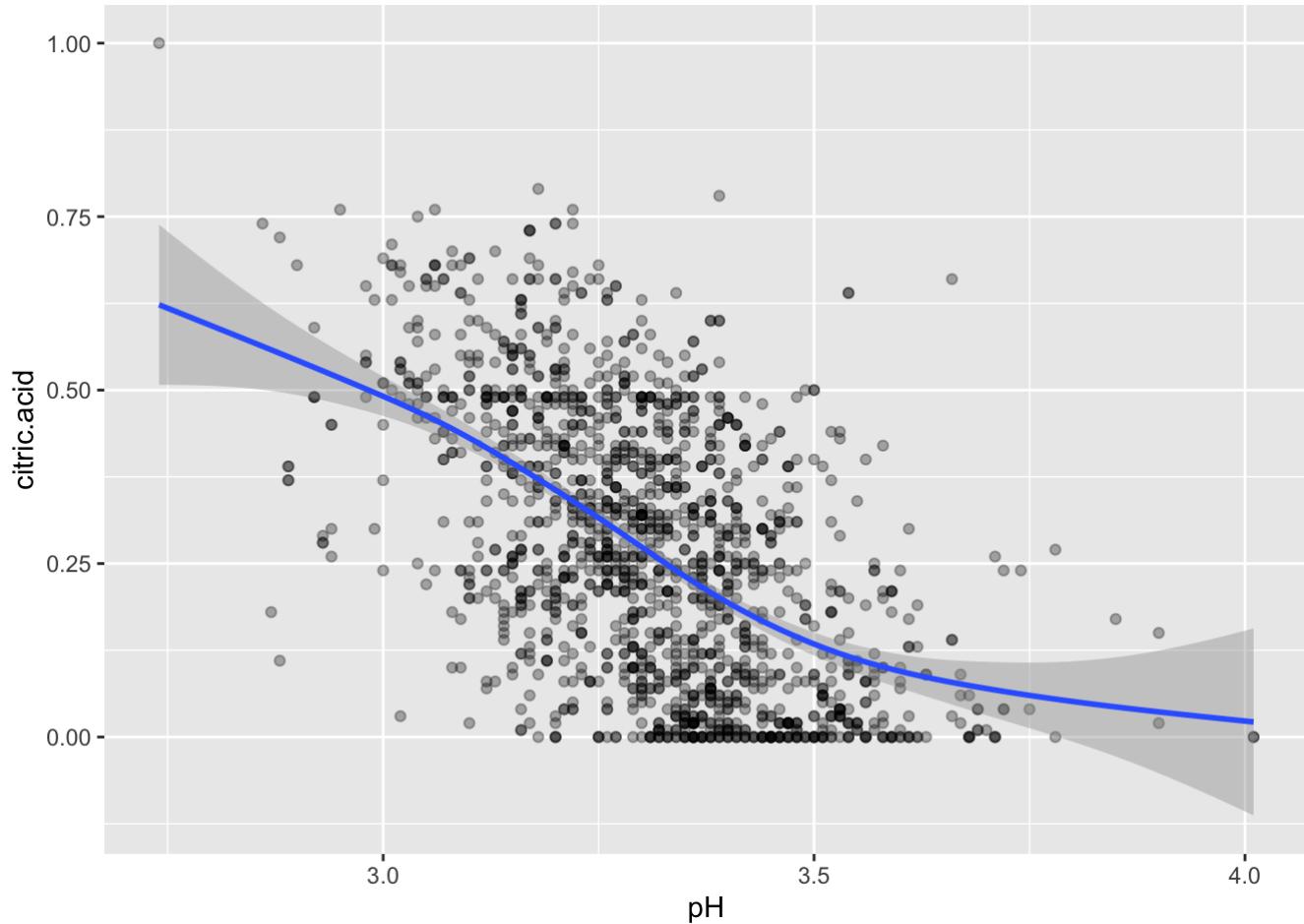
可以看出，residual sugar与quality几乎没有相关性。



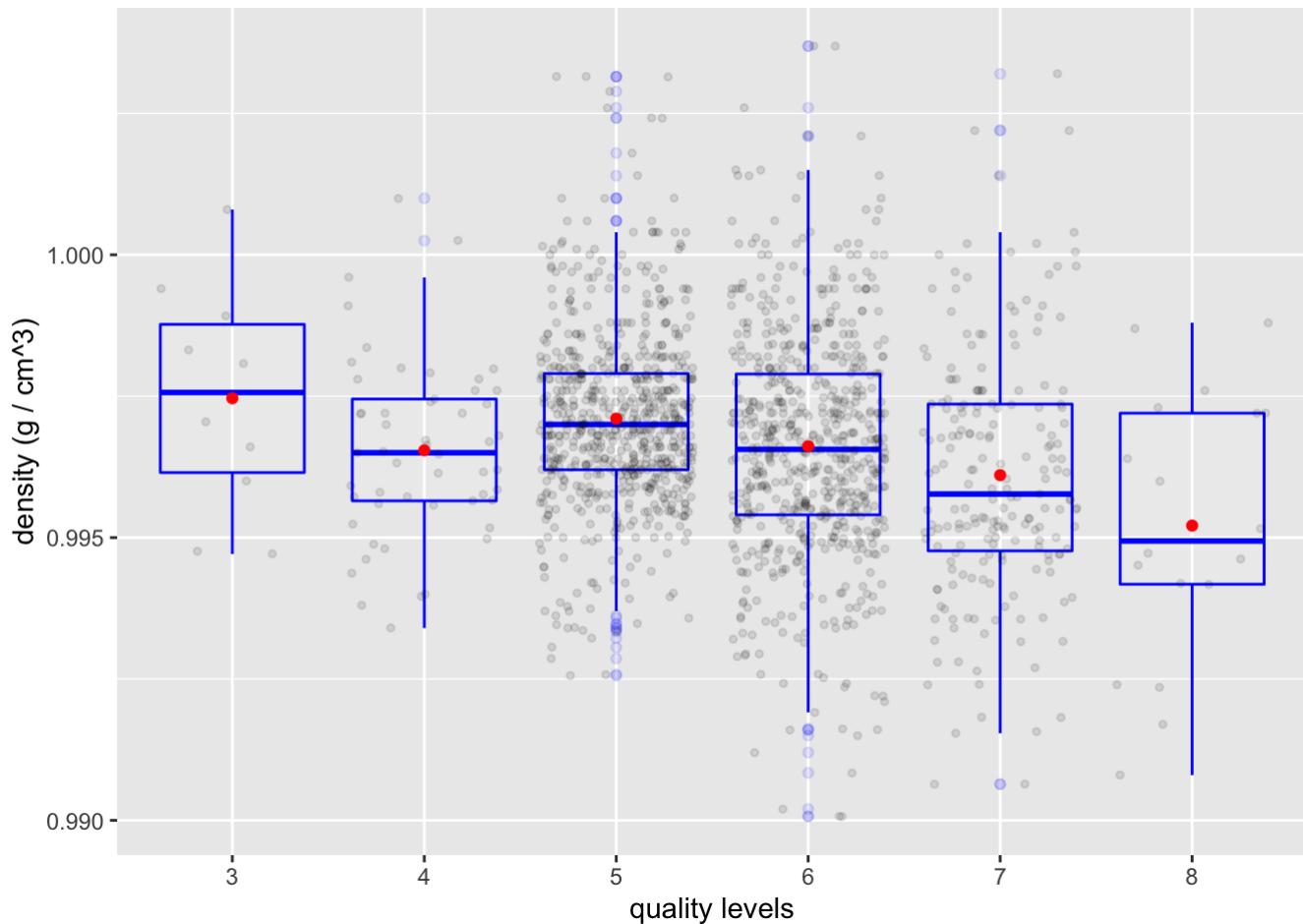
可以从曲线看出，柠檬酸越高，质量也有提高，这可能与柠檬酸可提高酒的口味新鲜感有关。



可以看出曲线呈下降趋势，也就是说volatile acidity越高，酒的质量越差，这应该与volatile acidity过高会让酒的酸味口感变差有关。



可以看到，随着pH值越来越高，柠檬酸的含量是越来越少的，也就是说，酸度越高，柠檬酸一般就越多。



可以看到密度和质量呈轻微负相关性。

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？

我发现了酒精含量越高，酒的评分也越高一些。硫酸盐的含量越高，酒的评分也越高。相关性较小。我想是因为硫酸盐起到了抗菌和抗氧化的作用，对酒的质量有提高效果。柠檬酸含量越高，酒的评分也呈微微上升趋势。相关性很小。我觉得是因为柠檬酸会提高酒新鲜口感的性质发挥了作用。挥发性酸度越高，酒的质量越低。它们的负相关性一般。我觉得是因为挥发性酸度变高会让酒的口感变差，所有是有负相关性。

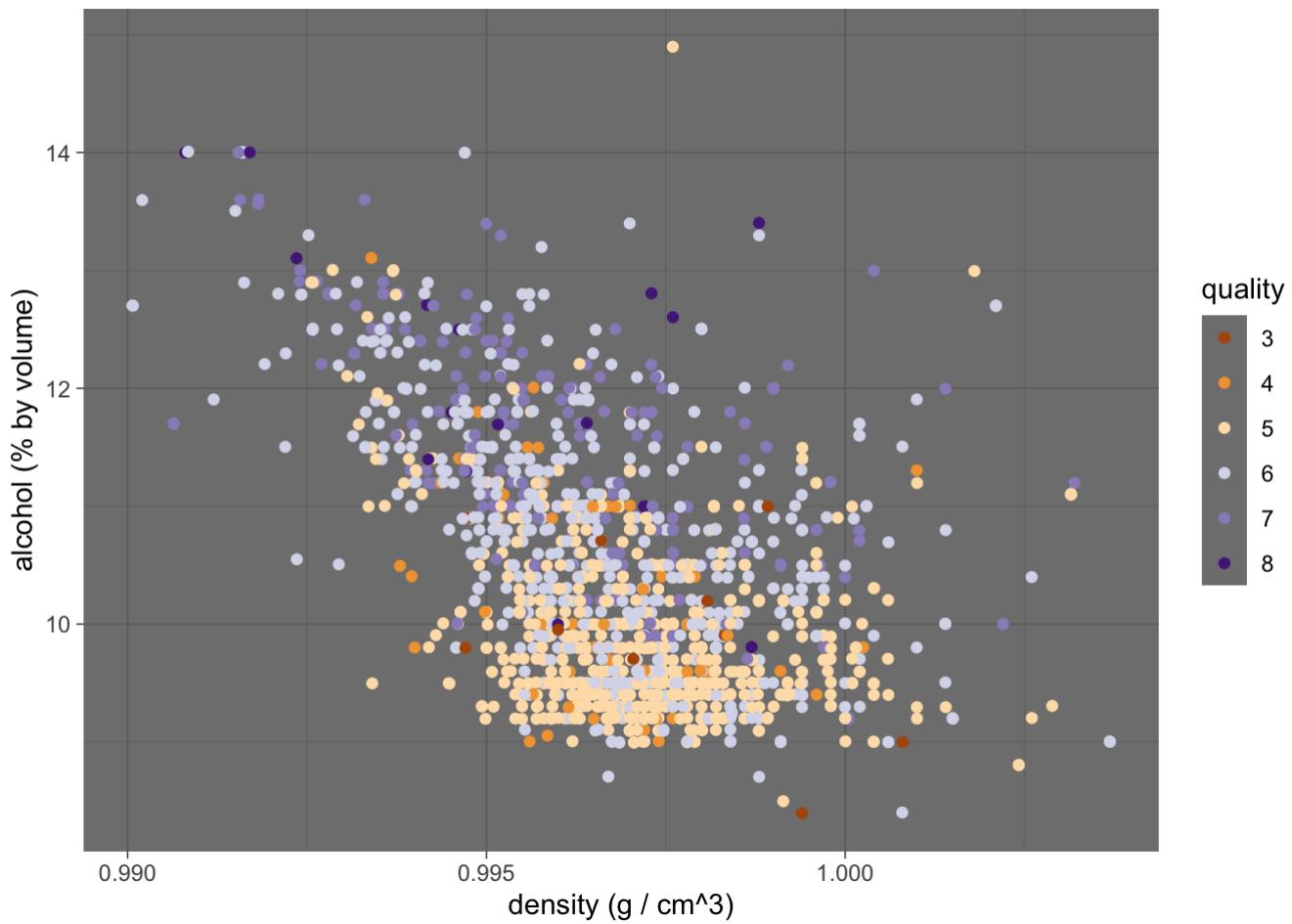
你是否观察到主要特性与其他特性之间的有趣关系？

随着pH值越来越高，柠檬酸的含量是越来越少的，也就是说，酸度越高，柠檬酸一般就越多。我觉得可能是因为它们都是酸的化学属性这一原因。

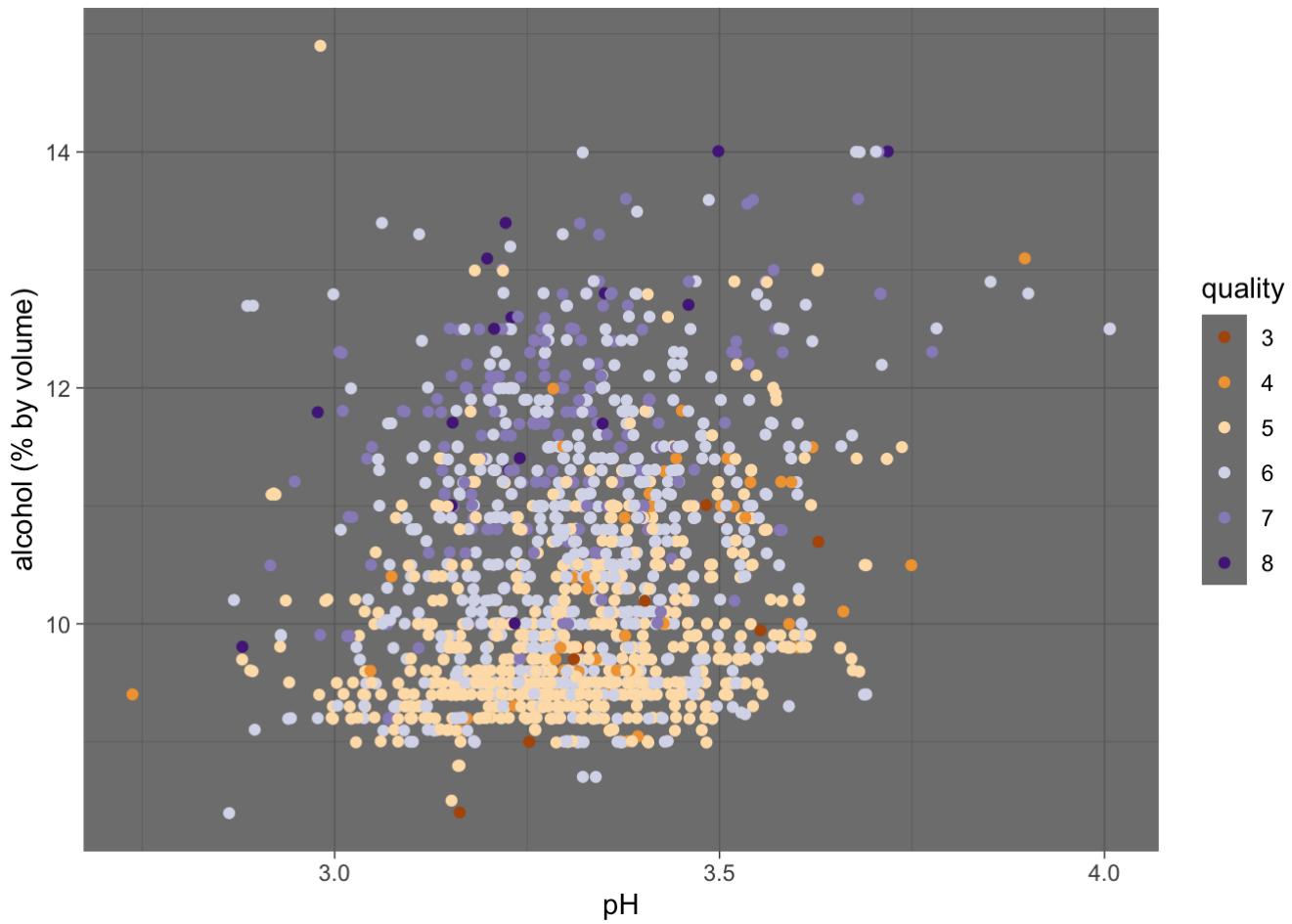
你发现最强的关系是什么？

我发现的最强关系是酒精和质量。

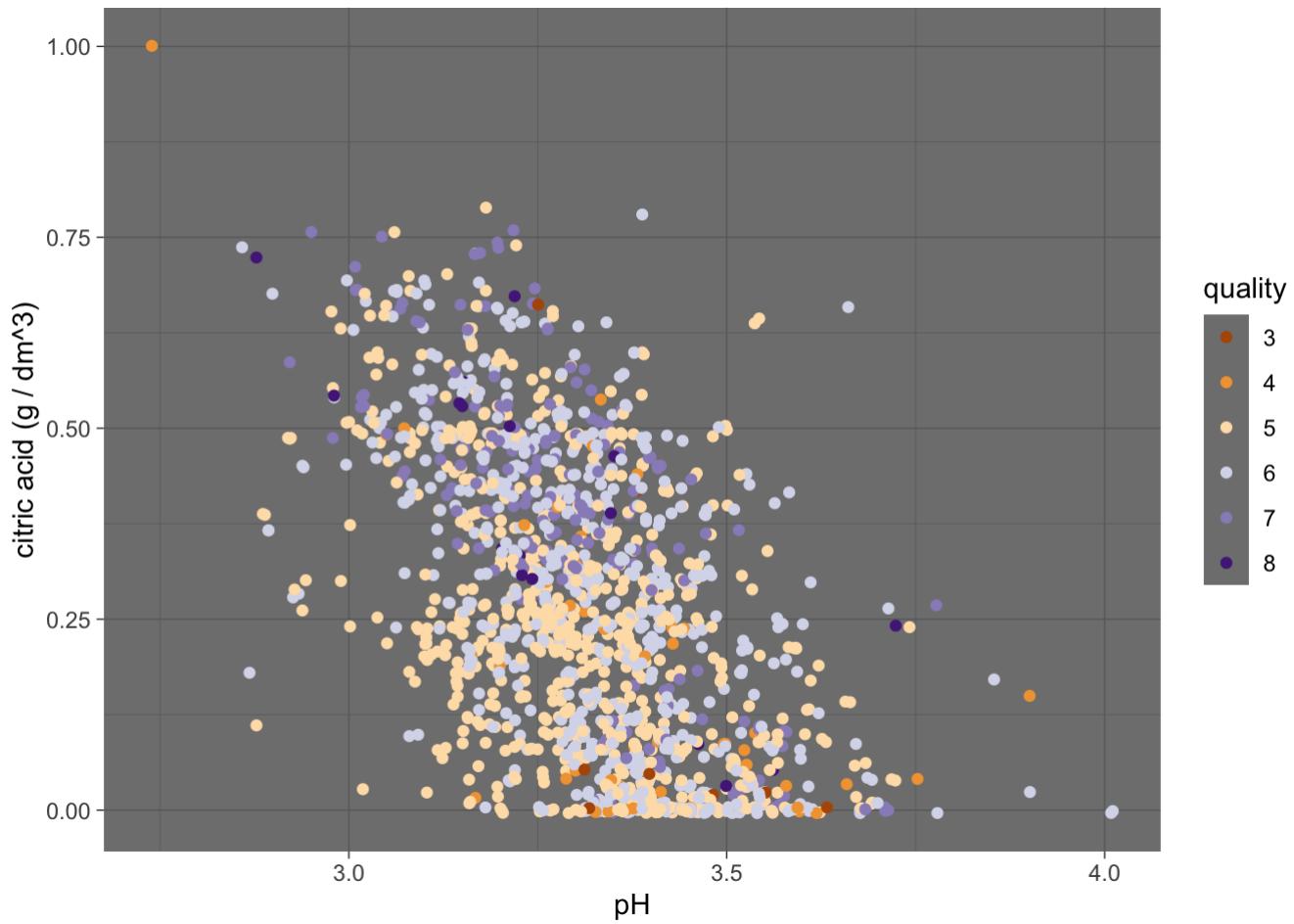
多变量绘图选择



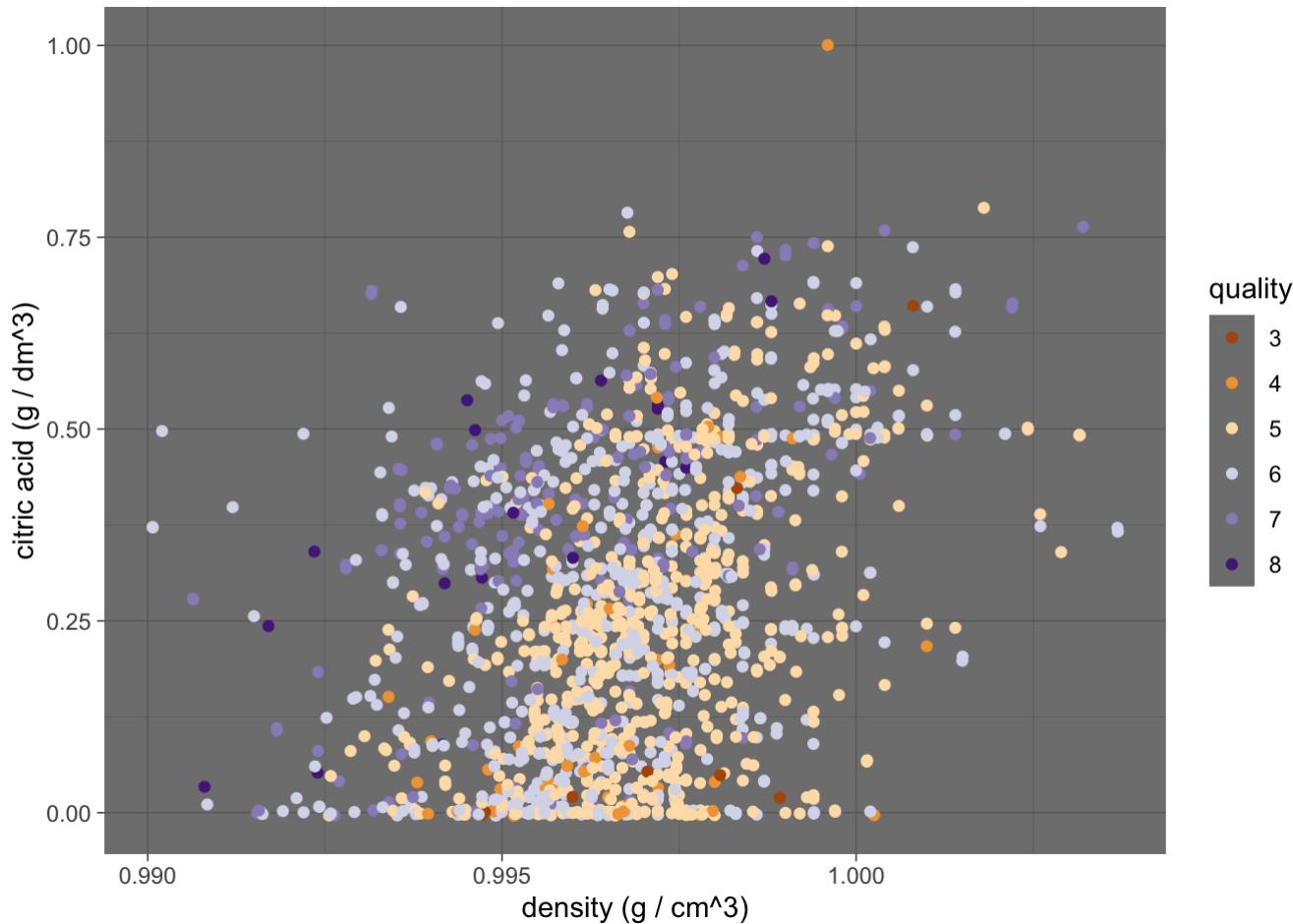
可以看出，在所有酒里面，密度越高酒精越少的趋势，大多数评分较低的酒的酒精含量都偏少，而质量较高的酒酒精含量都偏高一些。



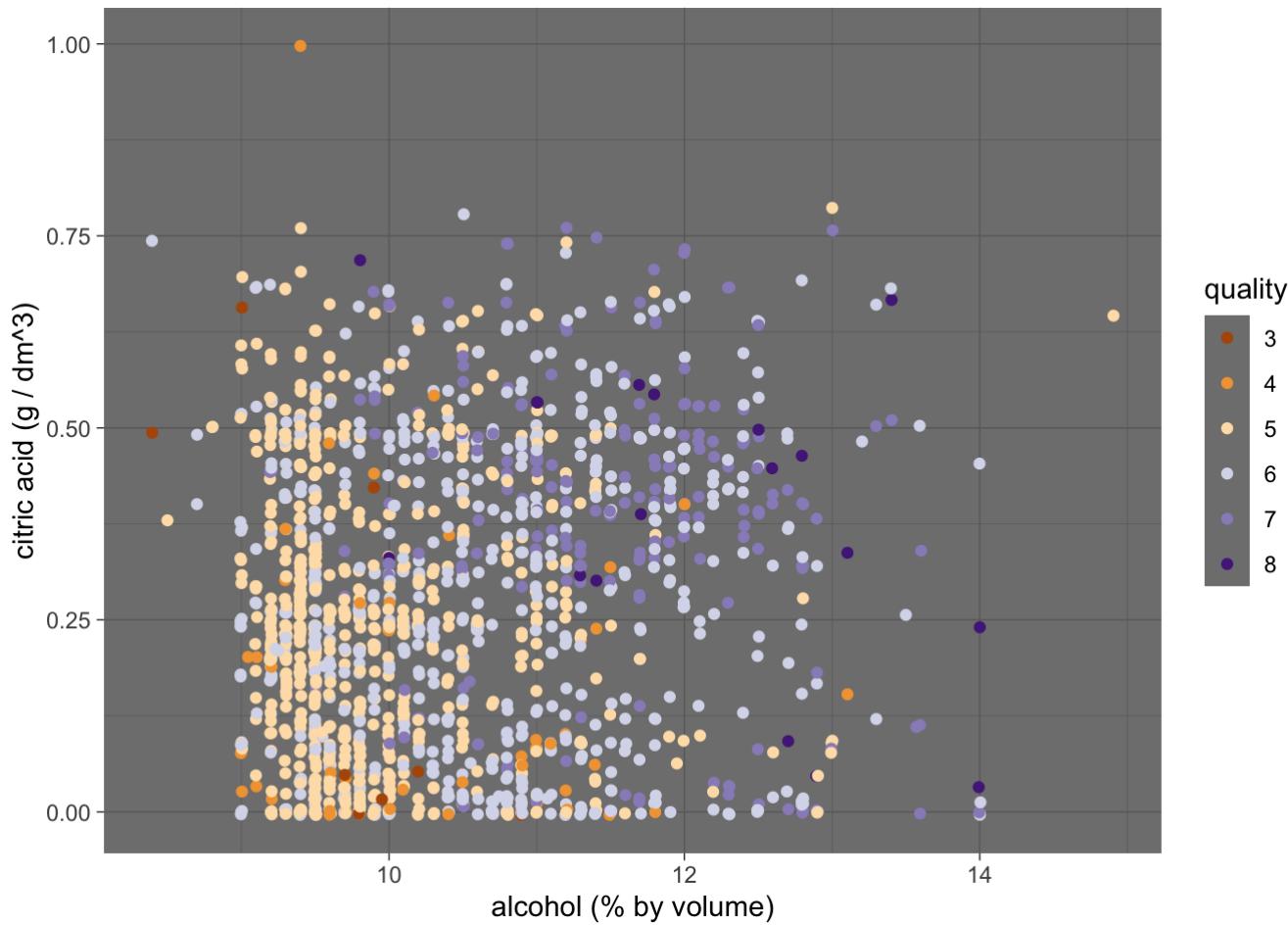
可以看到，评分较低的酒酒精含量偏低。而评分高的酒酒精都偏高，并且它们的酸度与酒精的负相关性要比评分低的酒强。



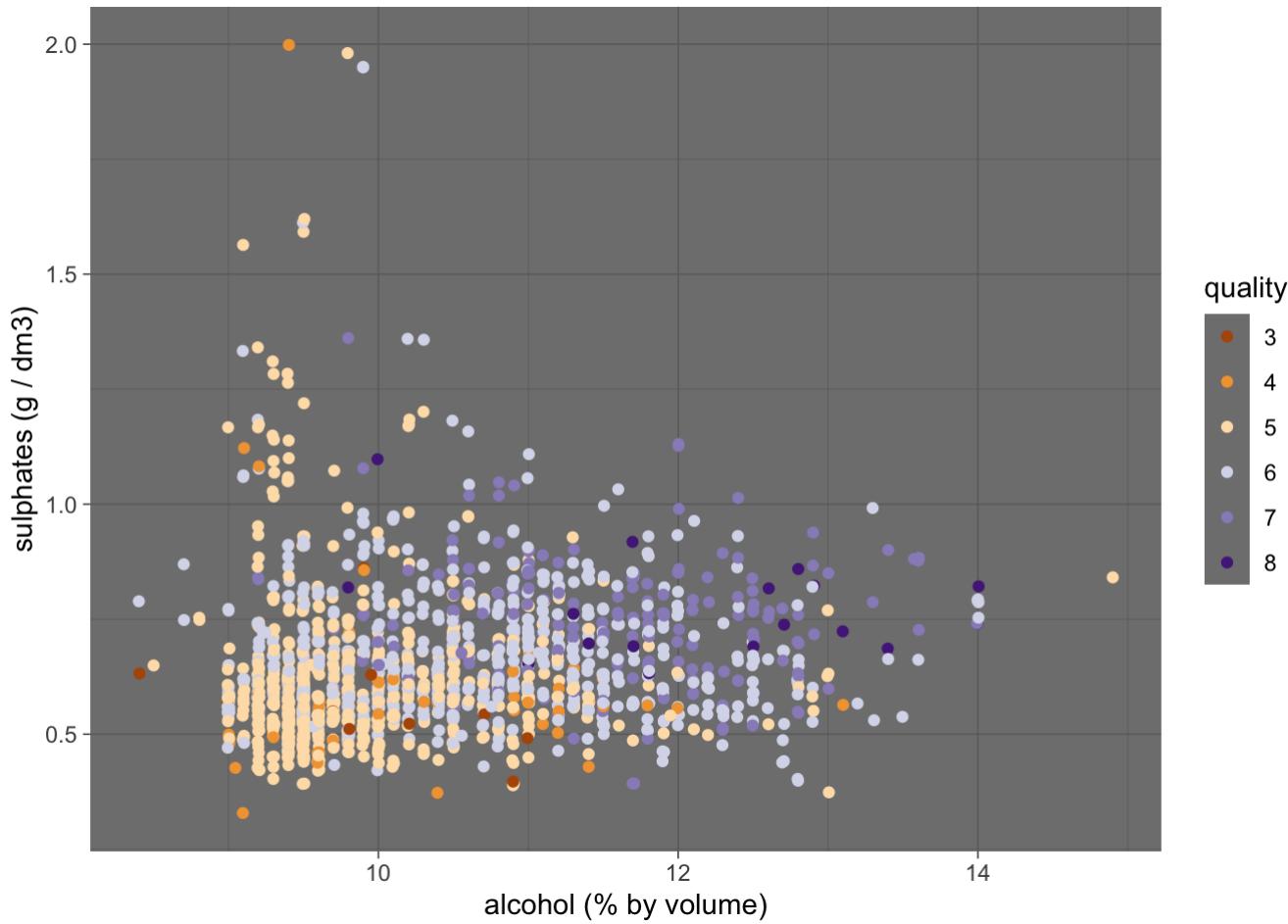
这幅图可以看出，所有评分的酒中，酸度越低柠檬酸就越少。



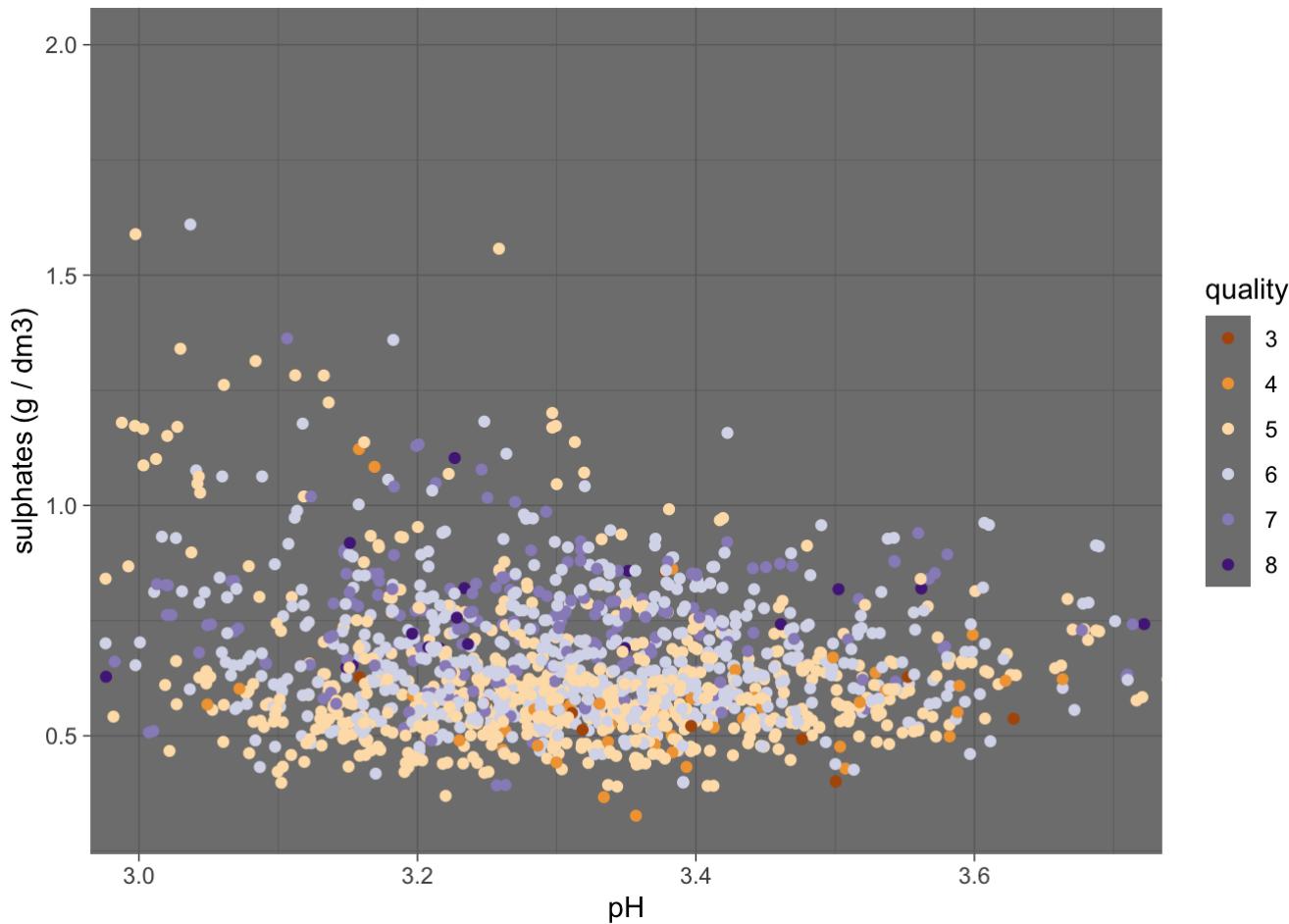
可以看到，在所有酒里面，密度和柠檬酸呈正相关性，但是评分较低的酒的密度相比评分较高的酒密度要高。



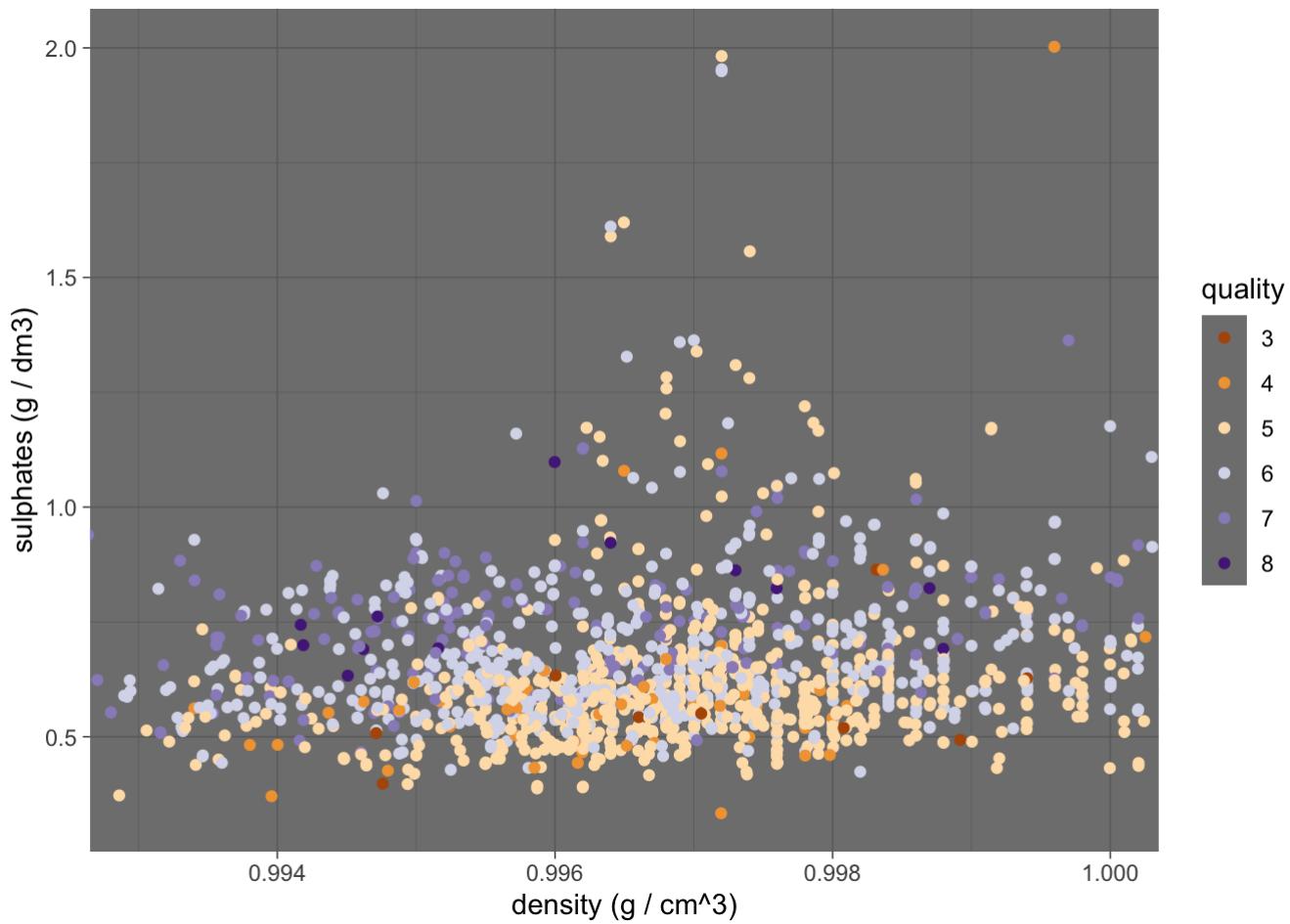
可以看出，酒精与柠檬酸没有什么相关性，但是评分较高的酒，明显酒精含量比评分较低的酒要高。



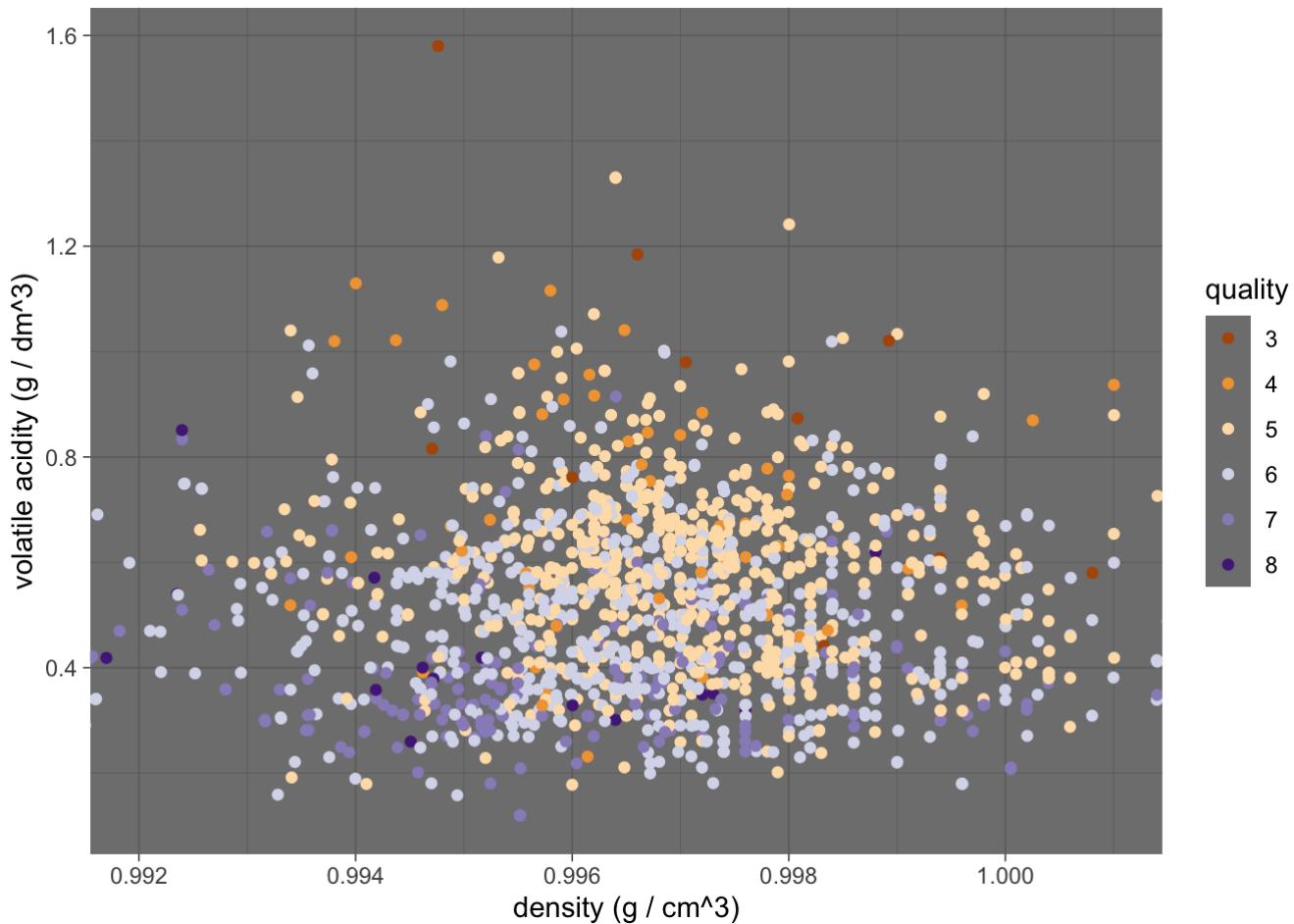
这幅图看不出酒精与硫酸盐的相关性，但是可以看出质量较高的酒酒精含量明显较评分低的酒要分布的更高。



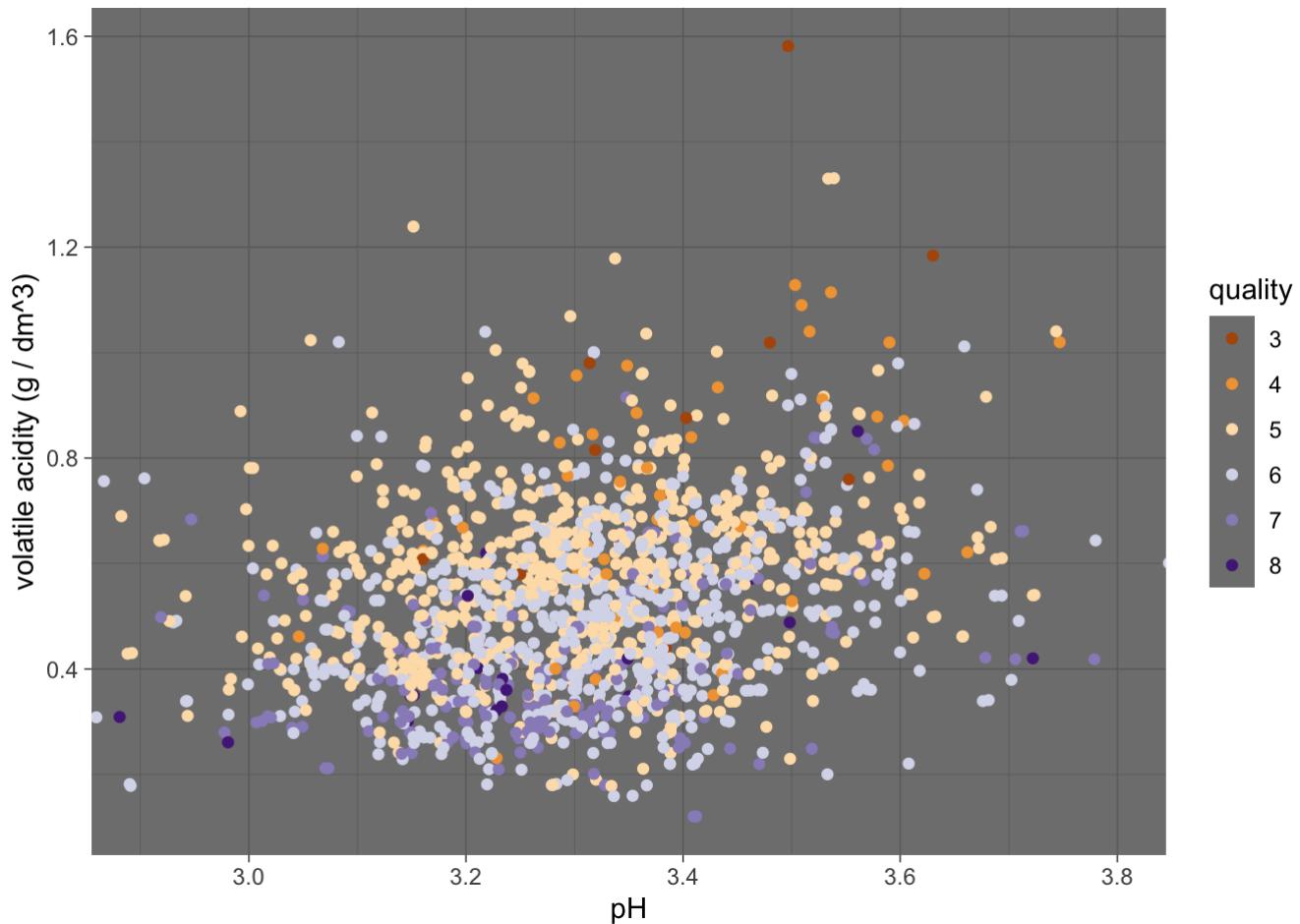
看不出酸度和硫酸盐的相关性，可以看出评分较低的酒的硫酸盐含量较评分高的酒分布的要低。



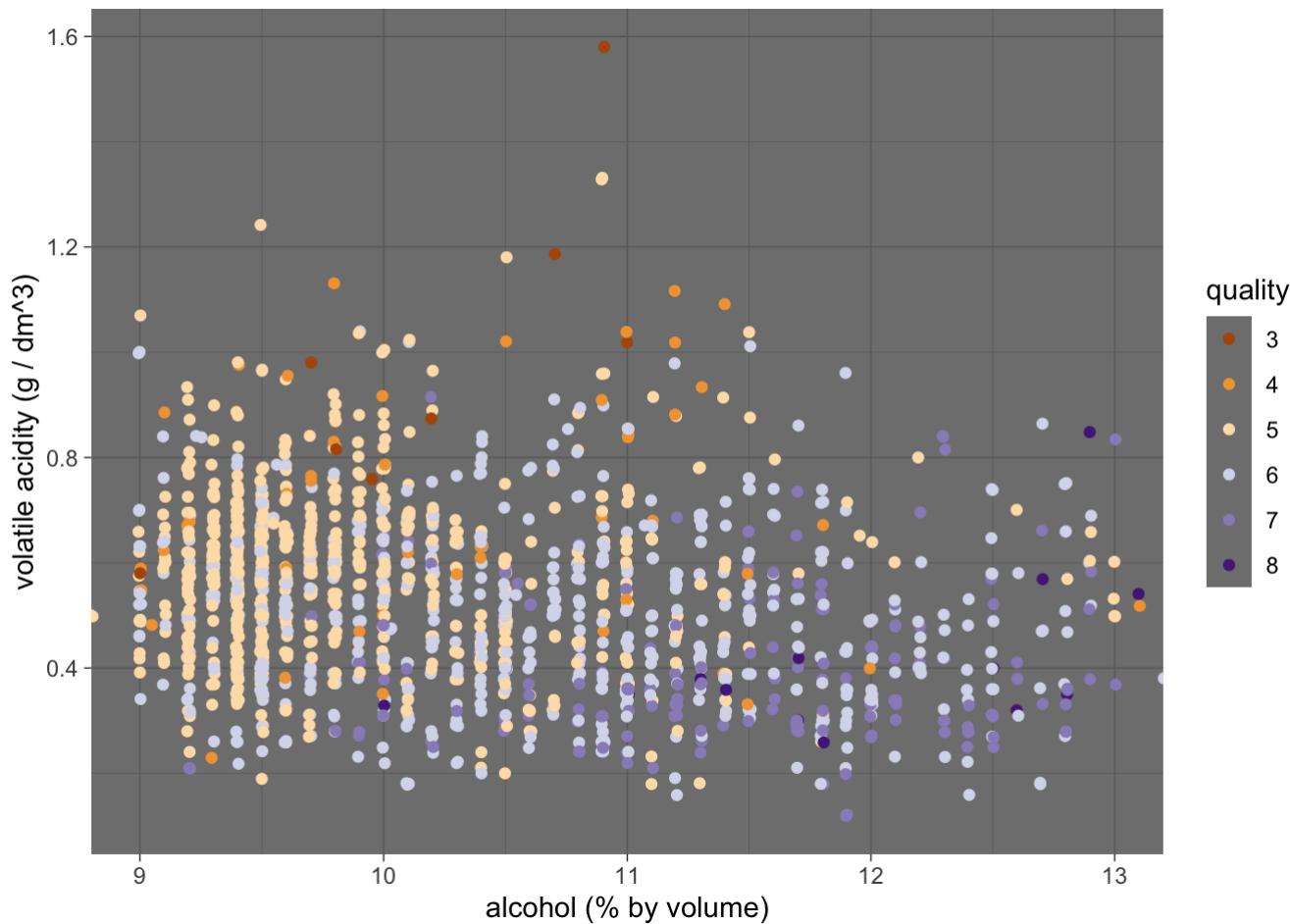
看不出密度和硫酸盐的相关性，但是可以看出评分较低的酒硫酸盐含量分布较评分高的酒要低。



看不太出密度和挥发性酸度相关性，但是可以轻微的看出评分较低的酒里挥发性酸度值要比评分高的酒要高一些。



可以看到，酸度和挥发性酸度呈轻微负相关性，质量较差的酒挥发性酸度相比质量较好的酒要高一些。



从图上面看不出酒精和挥发性酸度的相关性，但是可以看出质量较低的酒中酒精含量要比质量较高的酒要低。

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？

我发现了密度越高酒精越少的趋势，大多数评分较低的酒的酒精含量都偏少，而质量较高的酒酒精含量都偏高一些。评分高的酒的酸度与酒精的负相关性要比评分低的酒强。

所有评分的酒中，酸度越低柠檬酸就越少。密度和柠檬酸呈正相关性，但是评分较低的酒的密度相比评分较高的酒密度要高。

评分较低的酒的硫酸盐含量较评分高的酒分布的要低。我想是因为硫酸盐起到了抗微生物和抗氧化的作用。

但是可以轻微的看出评分较低的酒里挥发性酸度值要比评分高的酒要高一些。我想是因为挥发性酸度会造成一种不好的口感，所以会降低评分。酸度和挥发性酸度呈轻微负相关性。

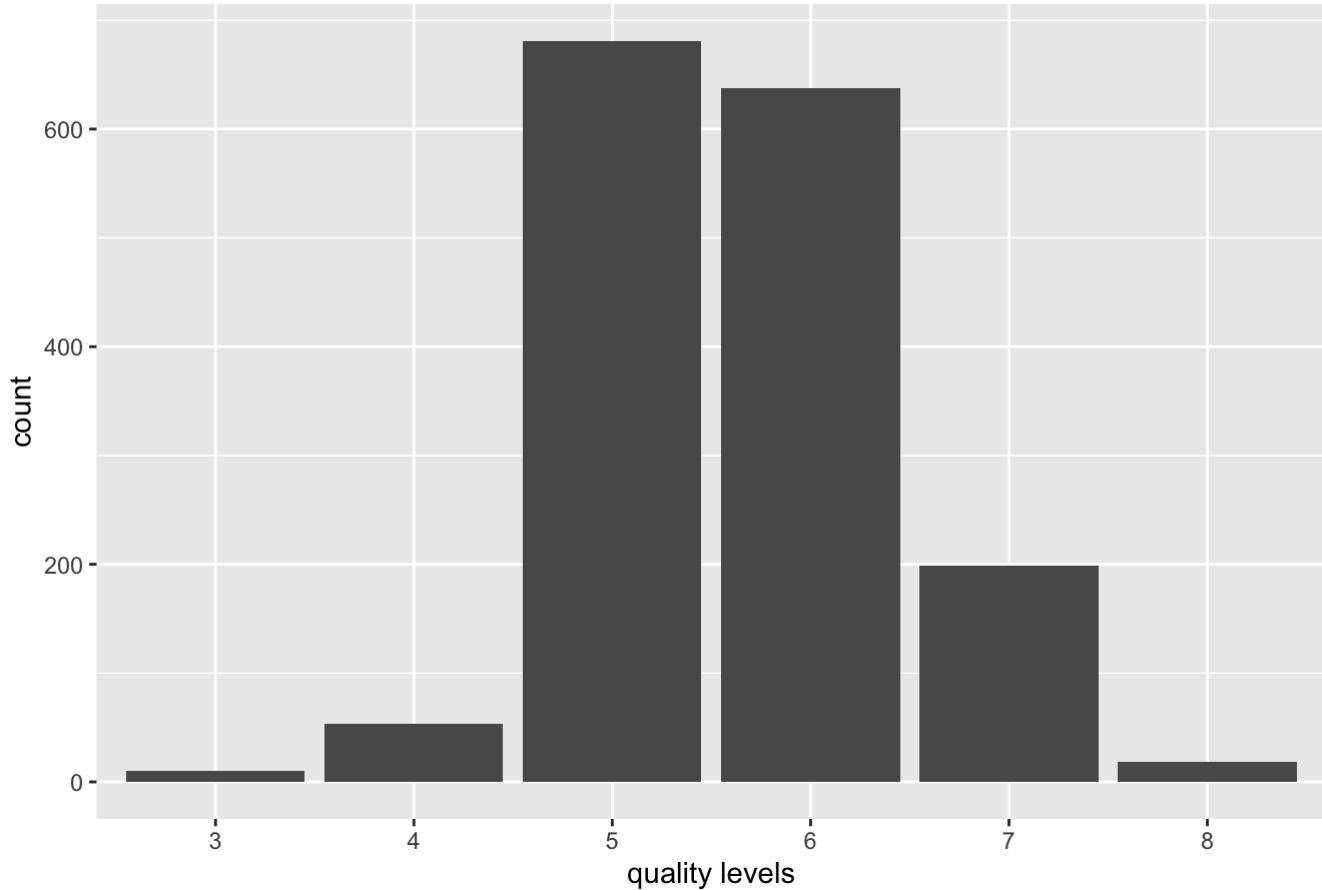
这些特性之间是否存在有趣或惊人的联系呢？

我发现不管在密度，酸度还是酒精含量分析里面，所有评分的酒都呈现了评分较低的酒的酒精含量都偏少，而质量较高的酒酒精含量都偏高一些的现象。这说明了不管因为什么原因，酒精含量高很可能会提高酒的质量。

定稿图与总结

绘图一

The distribution of Quality levels

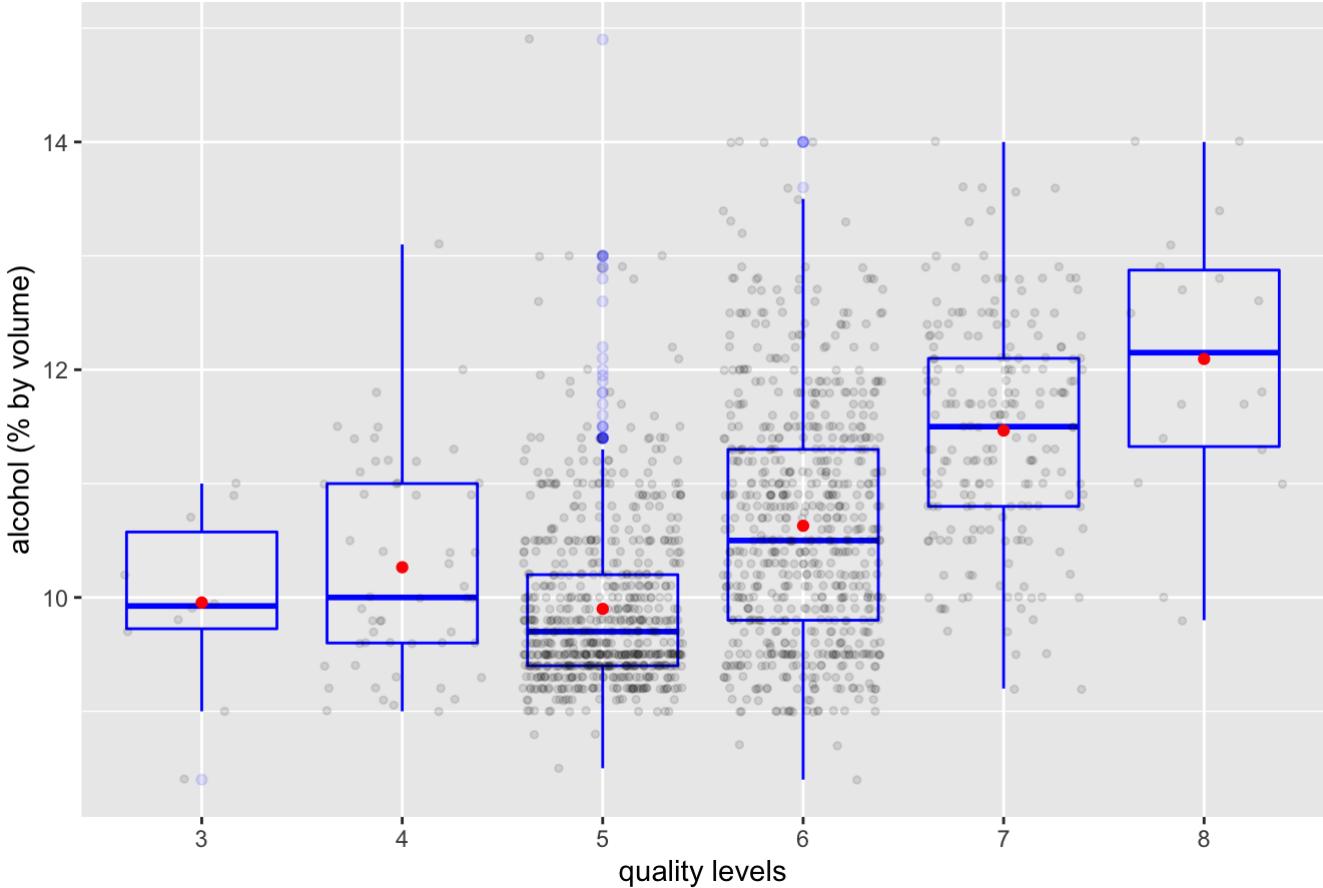


描述一

可以看到，大多数红酒的评分都在5和6分之中。

绘图二

Relationship of Quality levels and Alcohol

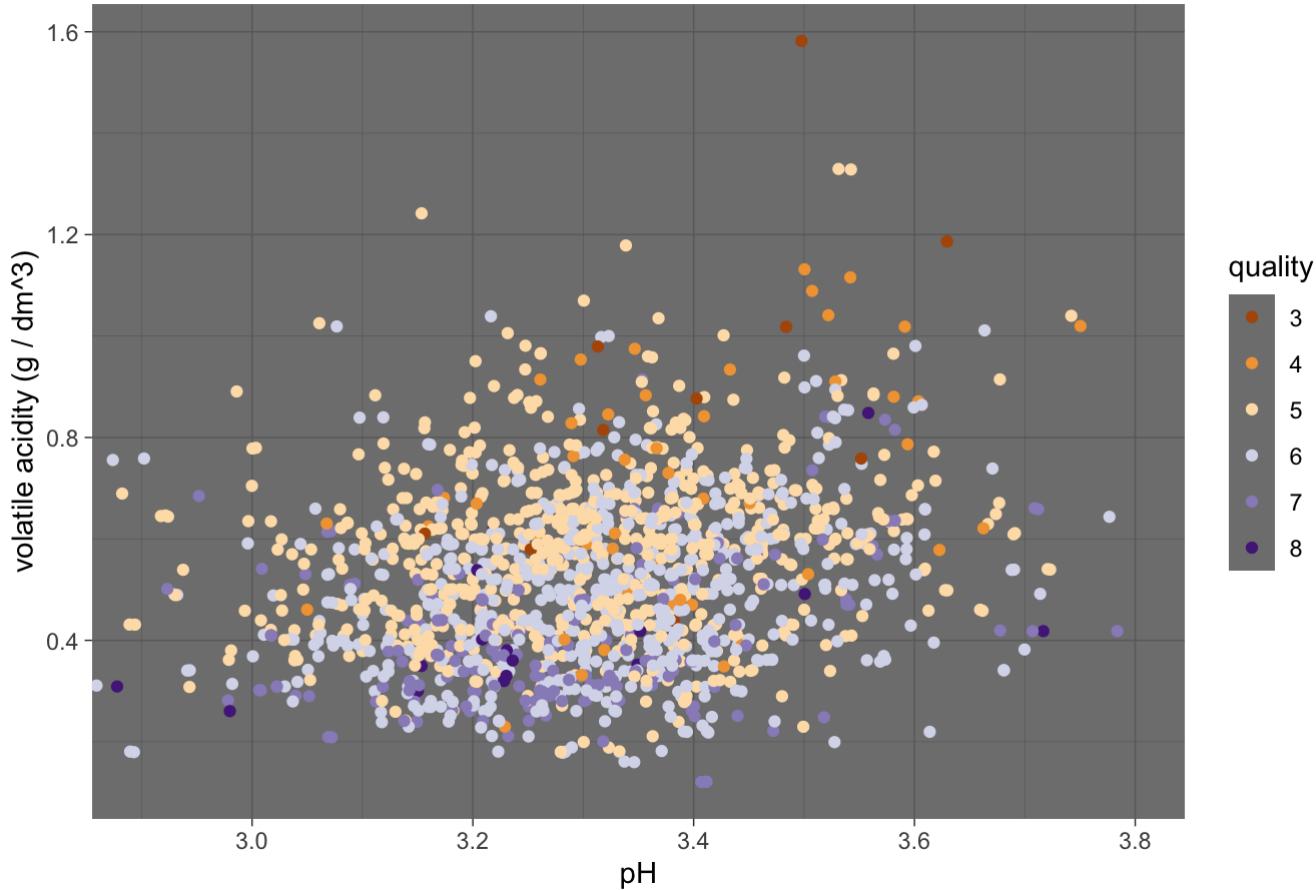


描述二

可以看到，酒的质量越高，酒精含量度的均值水平也在显著上升。

绘图三

Relationship of pH and Volatile acidity by Quality



描述三

可以看到，酸度越低，挥发性酸度越高的轻微趋势，质量较差的酒挥发性酸度相比质量较好的酒要高一些。我想是因为挥发性酸度会造成一种不好的口感，所以会降低评分。

反思

在做这个项目期间，我忘记了一些作图和函数的用法，不过后来用？查了一些官方文档就会做了，说明在以后的工作当中学习看文档是很重要的技能，因为我不可能记住所有的编程知识。还有就是理解数据集，包括变量与变量之间的关系。对分析问题的思考很有帮助。刚开始我没有设置margin线，这个习惯影响了代码可读性，以后的编程工作中不光要注意代码的有效性也要注意可读性。在思考酸度这个变量的时候，我没有考虑到酸度数值越高其实是酸度越小这一个特性，后来我还要重新检查和更改一遍所有关于酸度的分析内容。以后的工作里要提前考虑到变量特殊性质的方面。不过在分析过程中我也成功的发现了影响了红酒质量的变量。这要归功于正确运用图表的分析。在第三幅总结图中，我把酒按质量做了分类，并且用散点图做出了酸度和挥发性酸度的关系。正确分析出了质量较差的酒挥发性酸度相比质量较好的酒要高一些，并且符合这一化学特性的现实意义。