

探索性数据分析 - 红葡萄酒质量

Erick Zhang

2018-5-7

- 1 项目简介
 - 2 单变量分析
 - 3 双变量部分
 - 4 多变量部分
 - 5 定稿图与总结
 - 6 反思
-

1 项目简介

整个的数据集包含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。

- 分析的目的：通过EDA的分析方法解释：“哪个化学成分影响红葡萄酒的质量？”

数据集概述

```
## [1] 1599    13
```

```
##          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
```

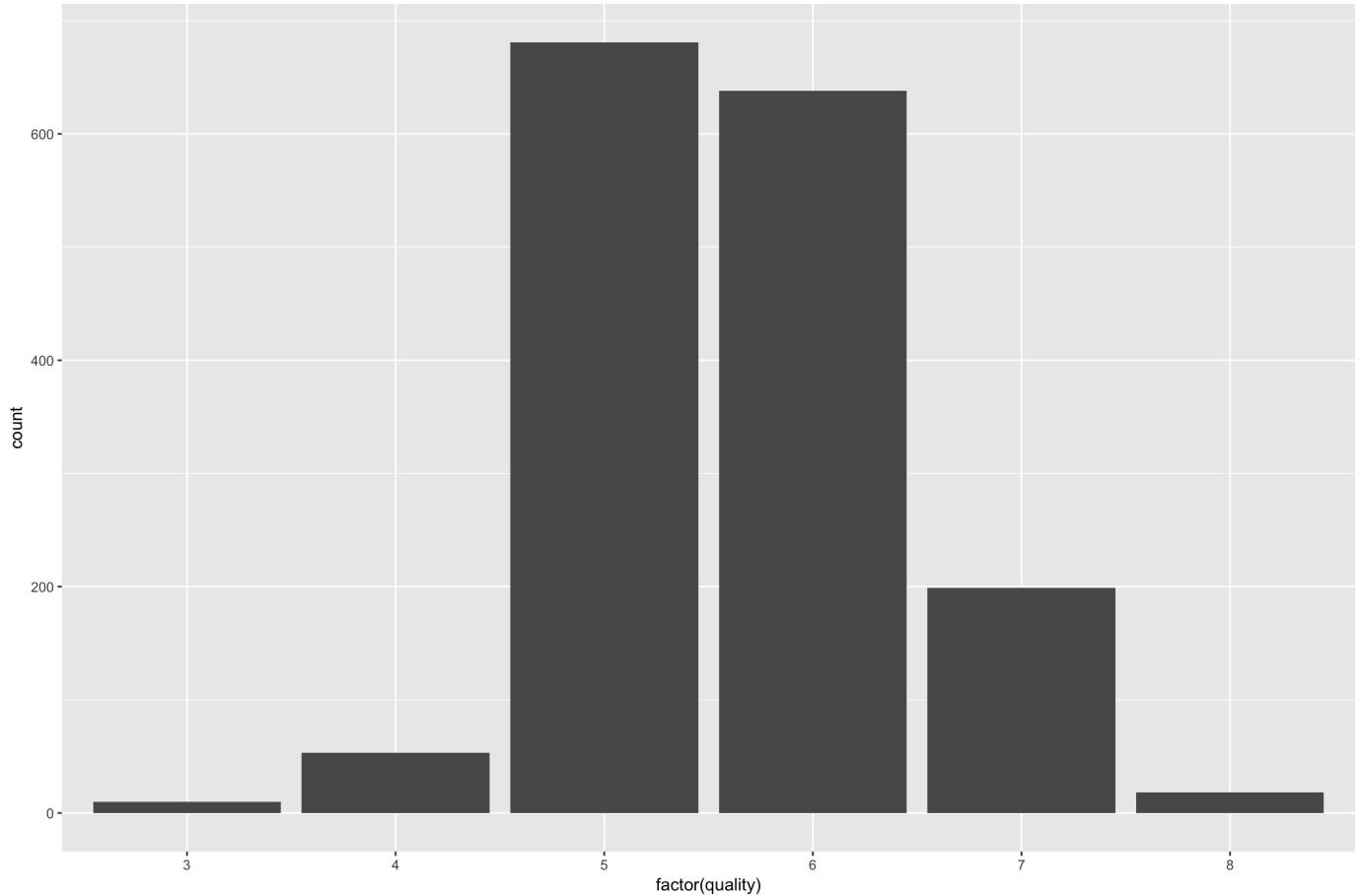
2 单变量分析

2.1 单变量绘图选择

葡萄酒质量

我们首先通过调查葡萄酒质量 `quality` 开始探索，测量得分范围在0到10之间。

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

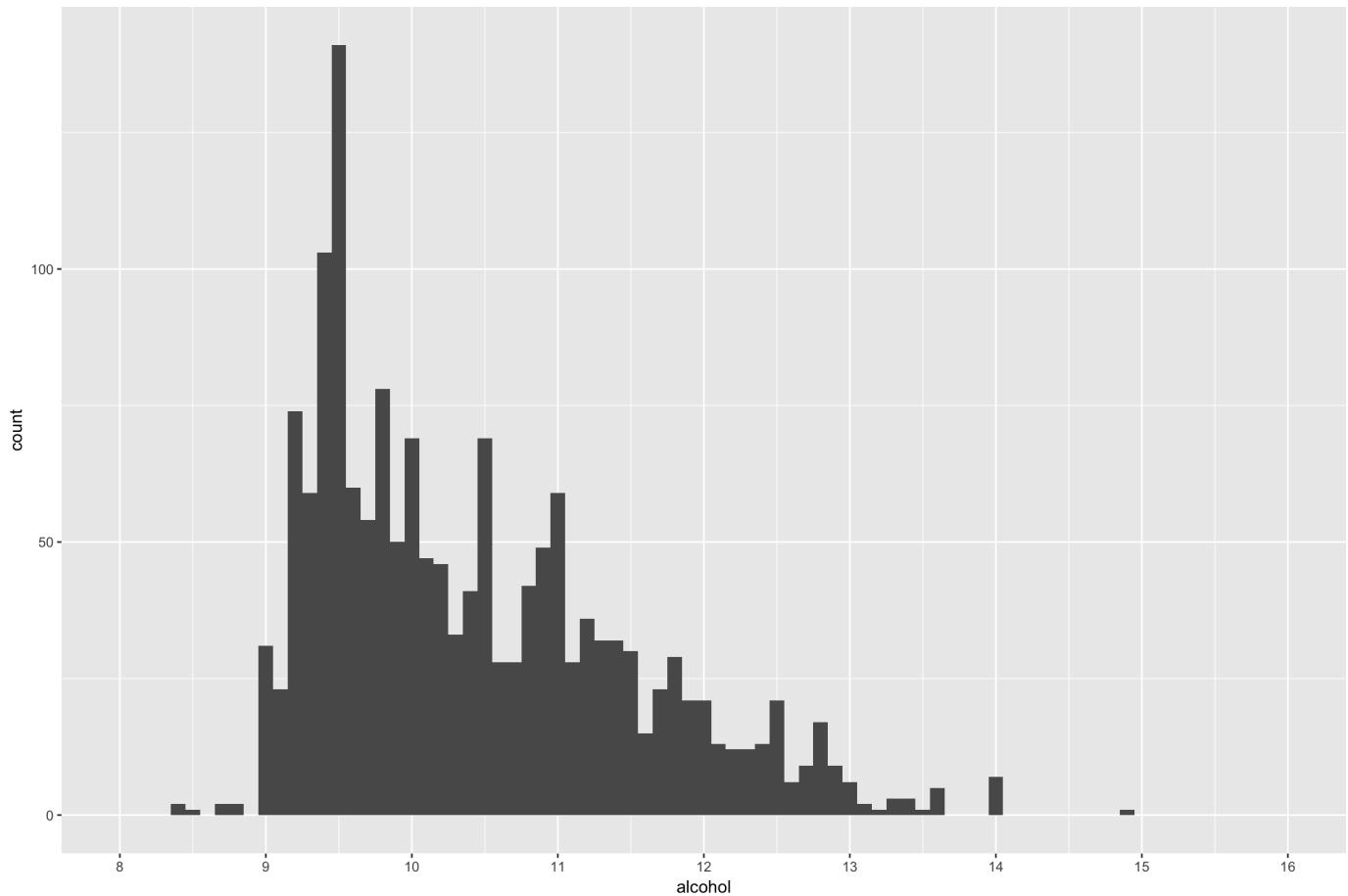


在给定的数据集中，葡萄酒得分在[3,8]范围内，大多数得分为5分到6分。

酒精率

之后我们来观察酒精率 `alcohol`

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 8.40    9.50   10.20  10.42   11.10  14.90
```

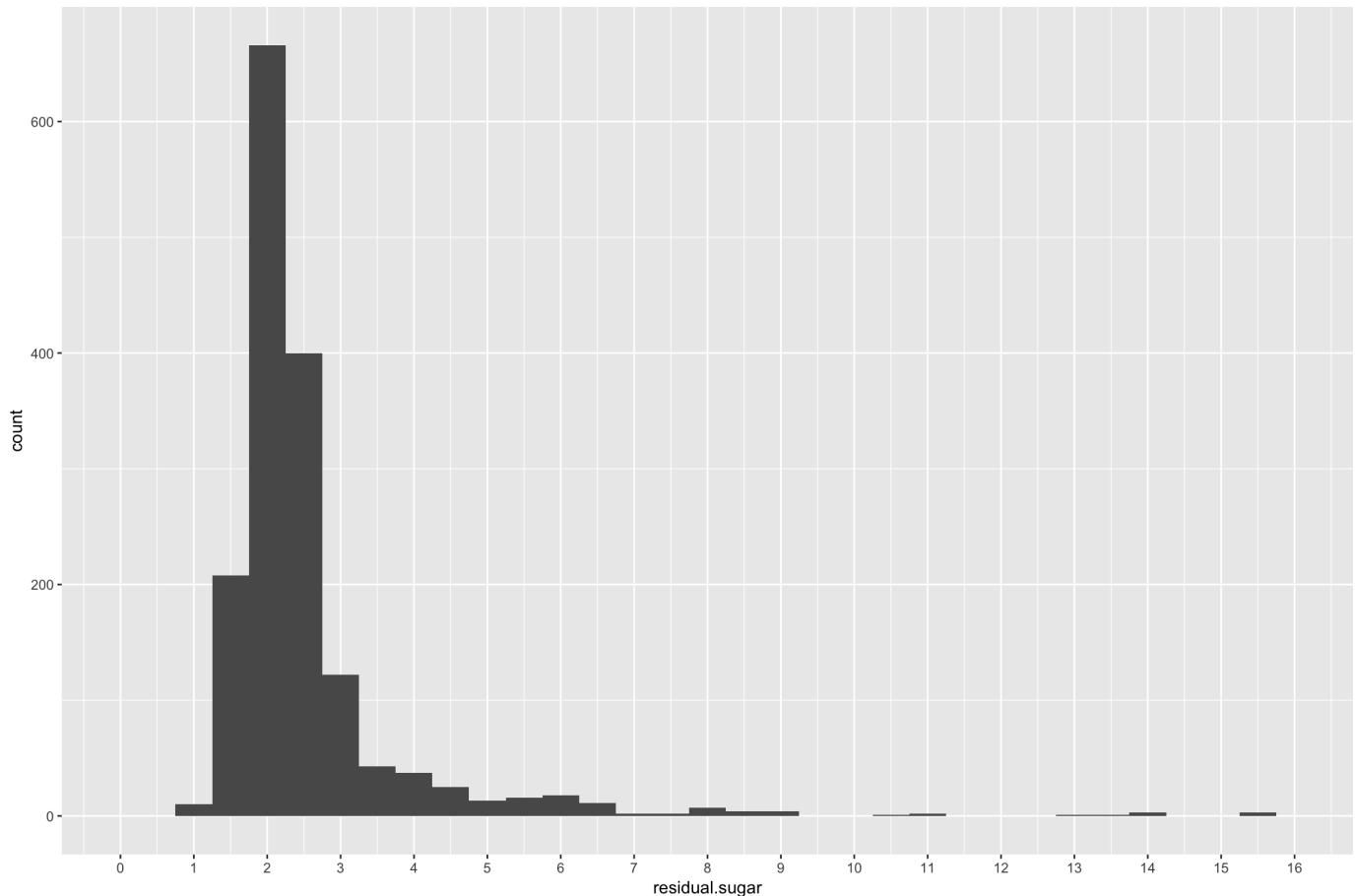


酒精中位数为10.2%，平均值为10.42%，第三四分位数为11.1%。如上图所示，酒精率图形右偏，这说明大多数葡萄酒的酒精率低于11.1%，只有25%的酒精率超过11.1%

残糖

下面试残糖量 residual.sugar 的对比，这个指标会影响红酒的甜度。

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.900   1.900   2.200   2.539   2.600 15.500
```

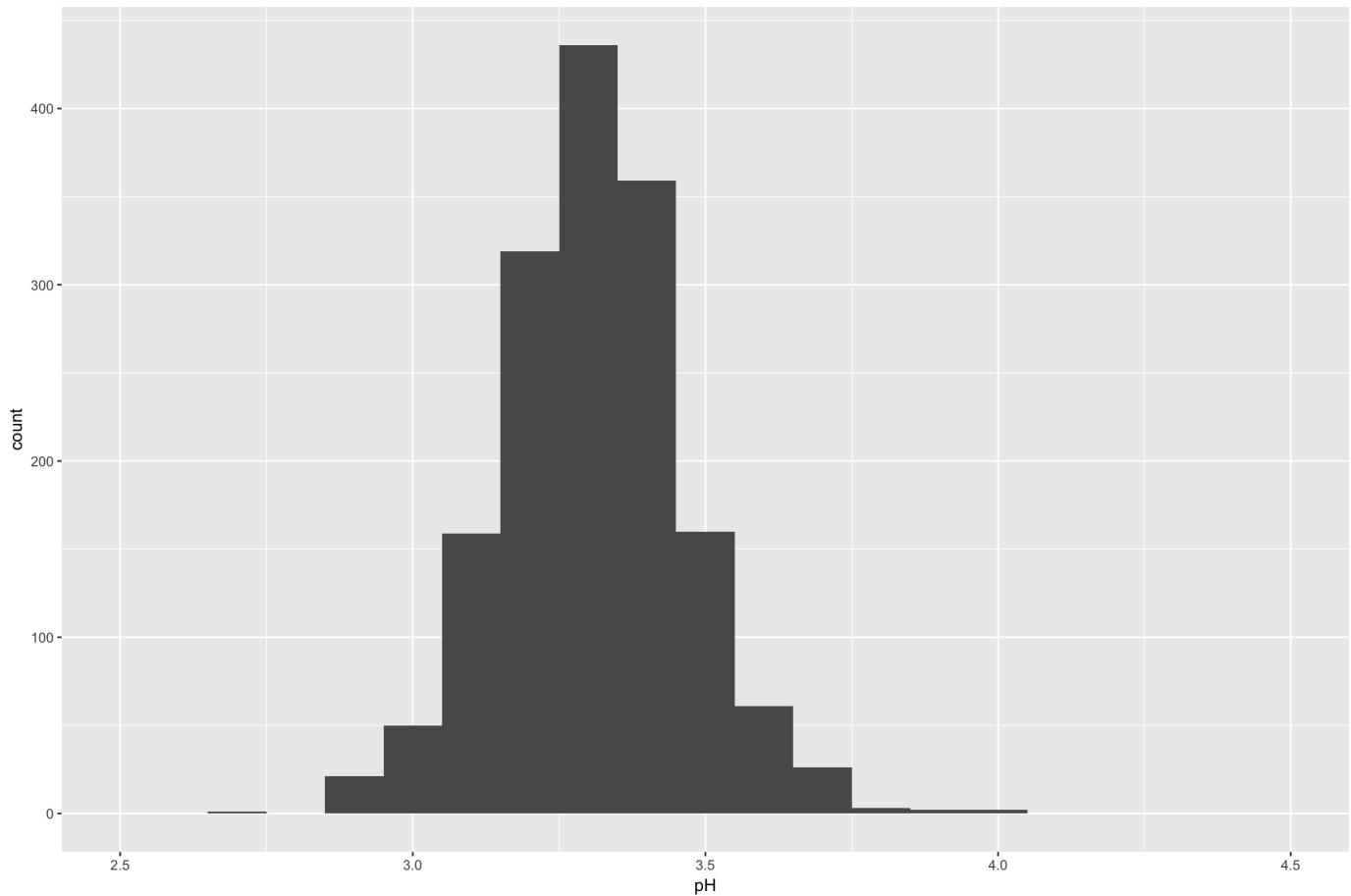


绝大多数的红酒残糖量在1.9~2.6之间，而残糖量的图形形成了一个很长的长尾，甚至有的酒残糖量达到了15以上

pH值

下面再就葡萄酒pH值 pH 的分布进行分析，葡萄酒酸度是其口感的重要指标。

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    2.740    3.210    3.310    3.311    3.400    4.010
```

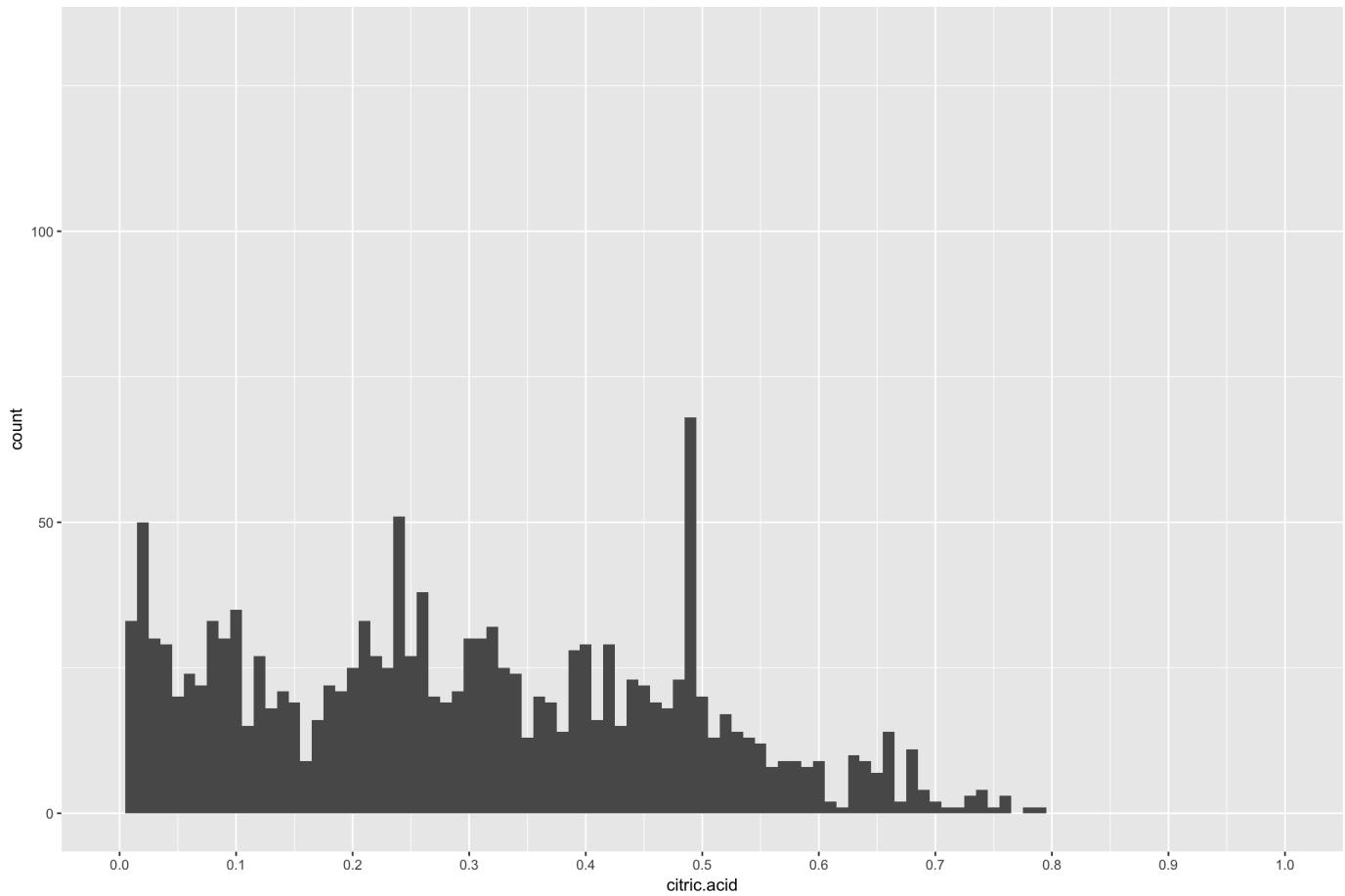


红酒的pH值平均在3.3左右，最高为4.01最低为2.74，分布形态比较正态，所有值都小于7，都呈现酸性口感。

柠檬酸

我们再来对柠檬酸 citric.acid 的分布进行讨论，柠檬酸对于葡萄酒的口感会起到调味的作用。

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
## 0.000   0.090   0.260   0.271   0.420   1.000
```

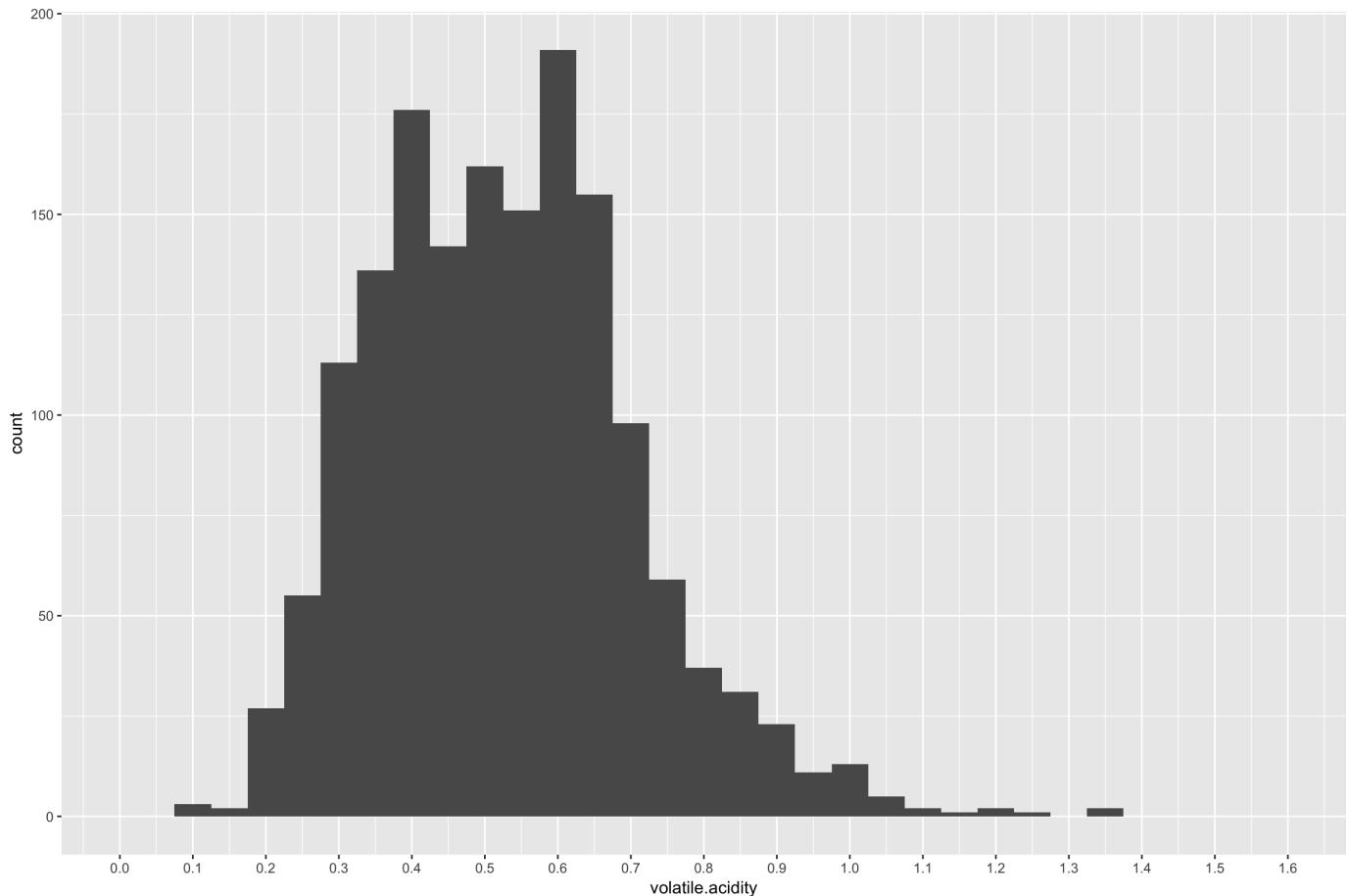


该图中有两个主峰。第一个在[0,0.02]之间，第二个在[0.48,0.5]的范围内。通过查看情节很难说它的分布。

挥发酸度

volatile.acidity 这个属性给出了葡萄酒中醋酸的含量，如果含量过高，会导致不希望得到的醋味

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

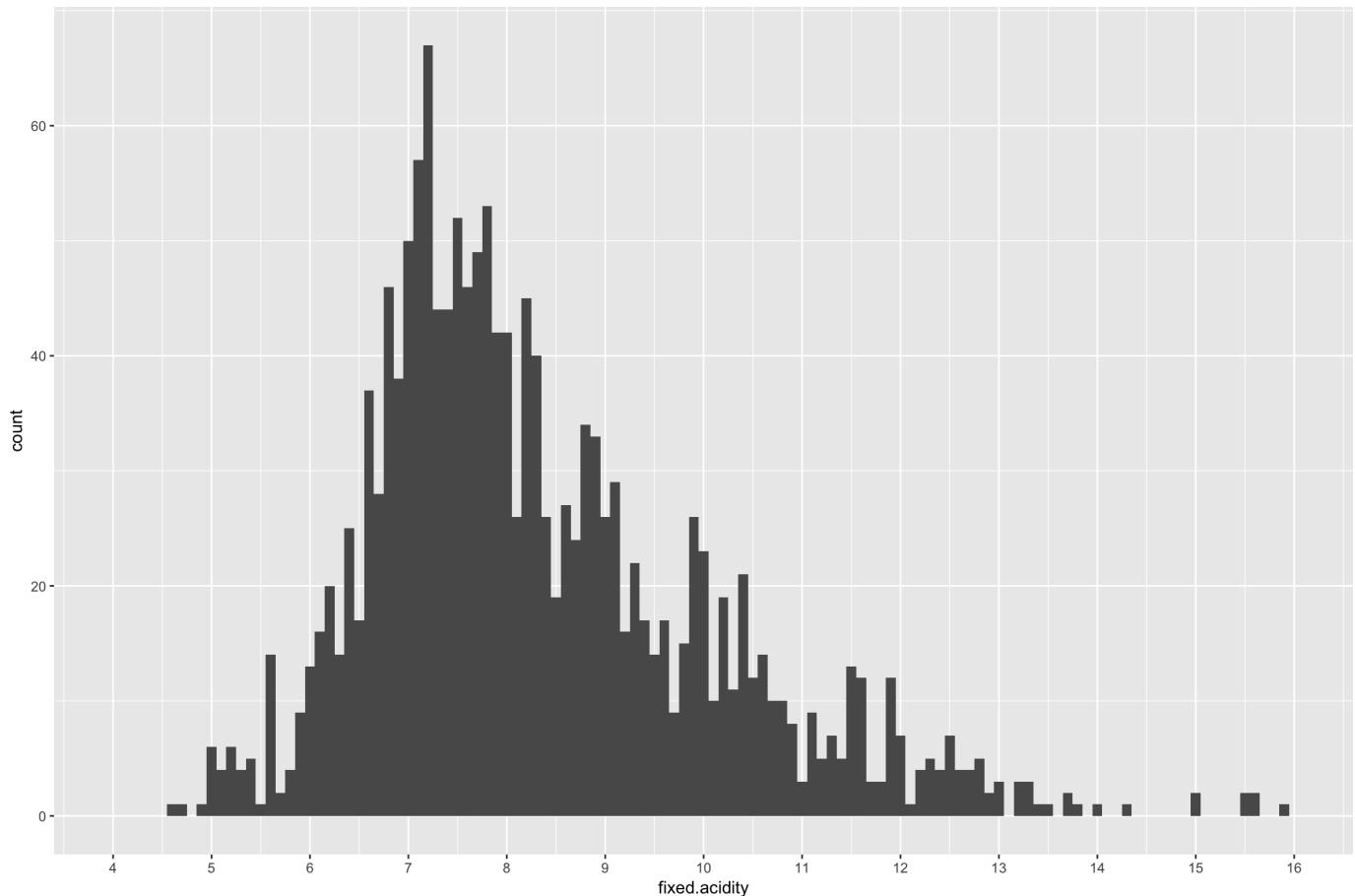


该属性的平均值和中位值几乎相等 (~ 0.52) , 似乎是具有正态分布的钟形曲线。但是图形的右侧有一个小尾巴

固定酸度

这是涉及葡萄酒的固定或非挥发性物质的酸性。现在我们来研究 `fixed.acidity` 这个属性:

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	4.60	7.10	7.90	8.32	9.20	15.90

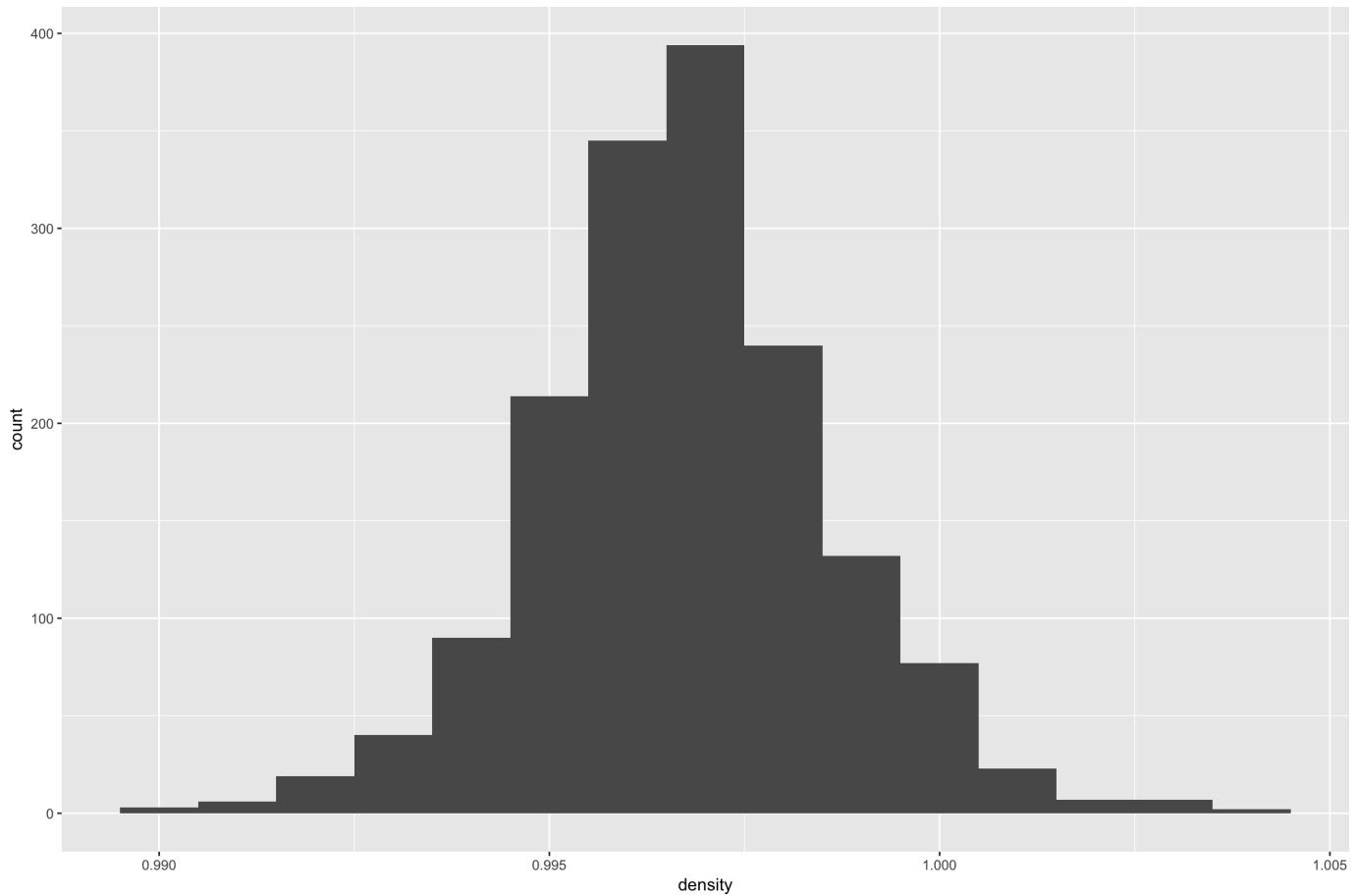


我们再次看到具有范围4.6~15.9的钟形图。中值为7.9，平均值为8.32

密度

接下来，我们来绘制密度属性 density

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.9901 0.9956 0.9968 0.9967 0.9978 1.0040
```

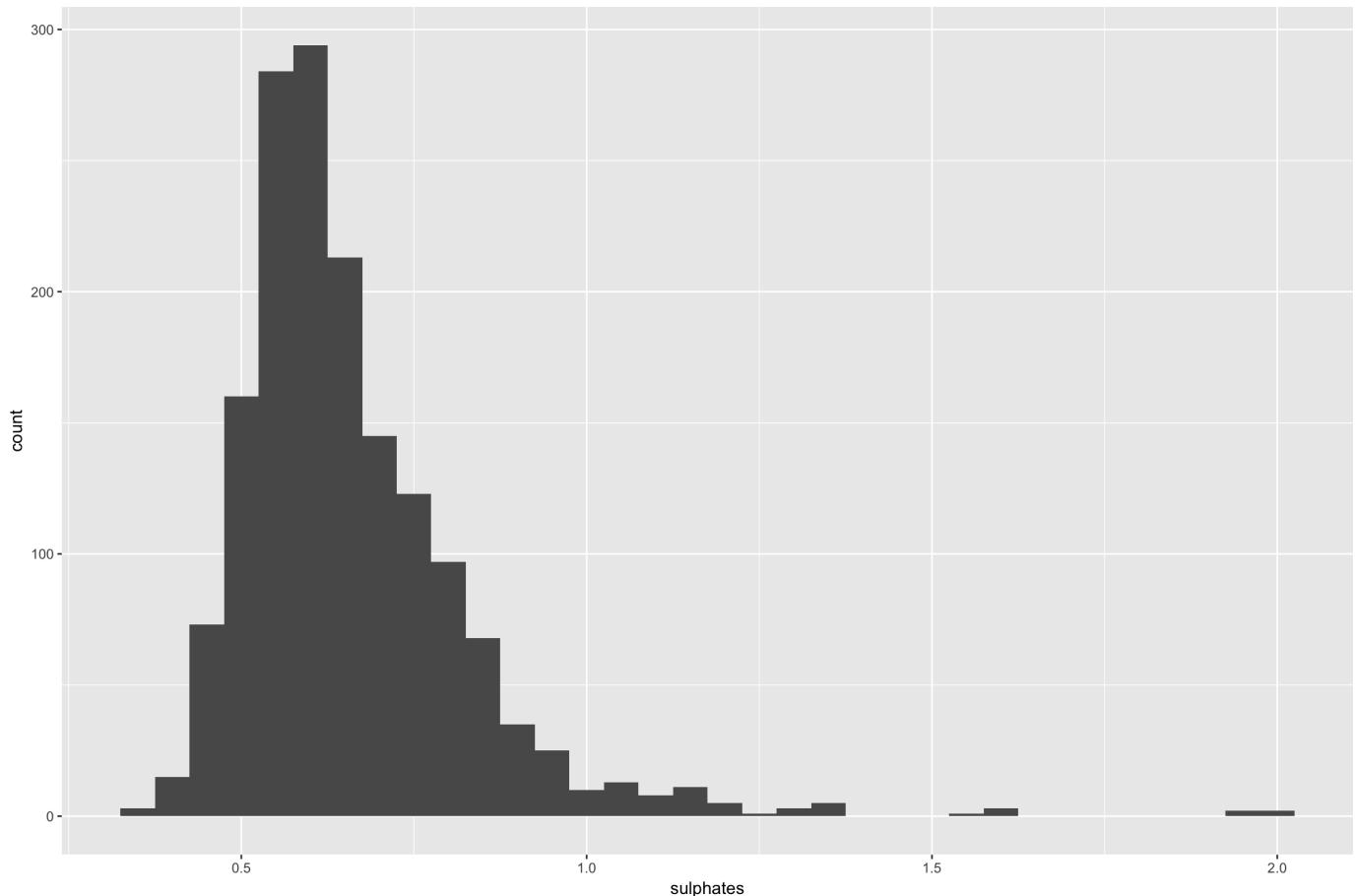


密度图看起来正态分布，平均等值为0.9967，中位数等于0.9968。

硫酸盐

sulphates 这是一种葡萄酒添加剂，可以促进二氧化硫气体的含量，并起到抗菌和抗氧化剂的作用。

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.  
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

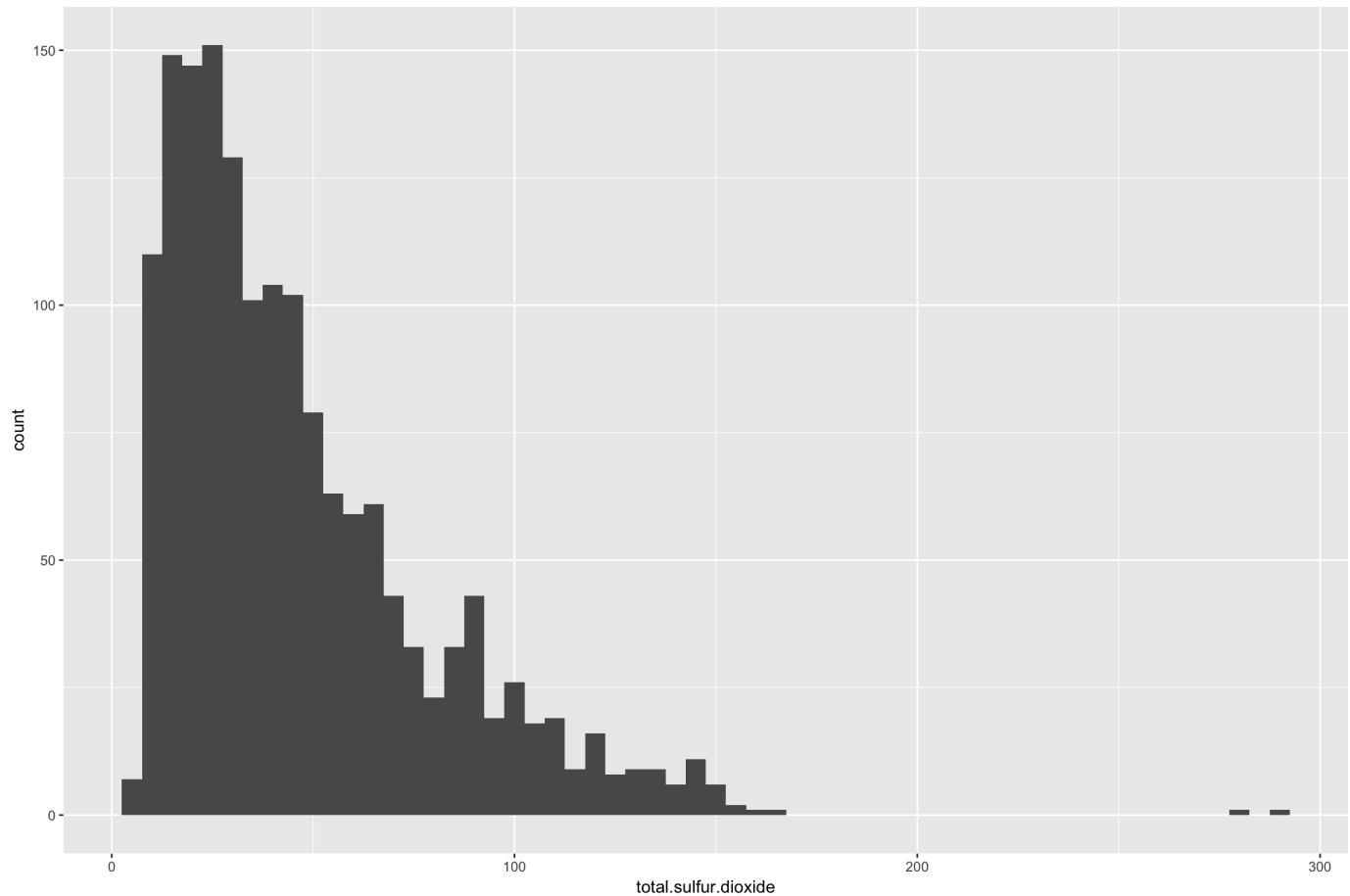


上面的图表，呈现出一个具有右侧长尾的钟形形态。硫酸盐浓度范围介于0.33至2之间，平均值为0.6581，中位值为0.6581，彼此非常接近。我们可以得出结论，在大多数葡萄酒中（在给定的数据集中），硫酸盐含量为0.62g / dm³

二氧化硫总量

total.sulfur.dioxide 表示SO2的游离和结合形式的量；在低浓度下，SO2在葡萄酒中几乎检测不到，但是当游离SO2浓度超过50mg / L时，SO2在葡萄酒的鼻子和味道中变得明显

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      6.00   22.00  38.00    46.47   62.00  289.00
```

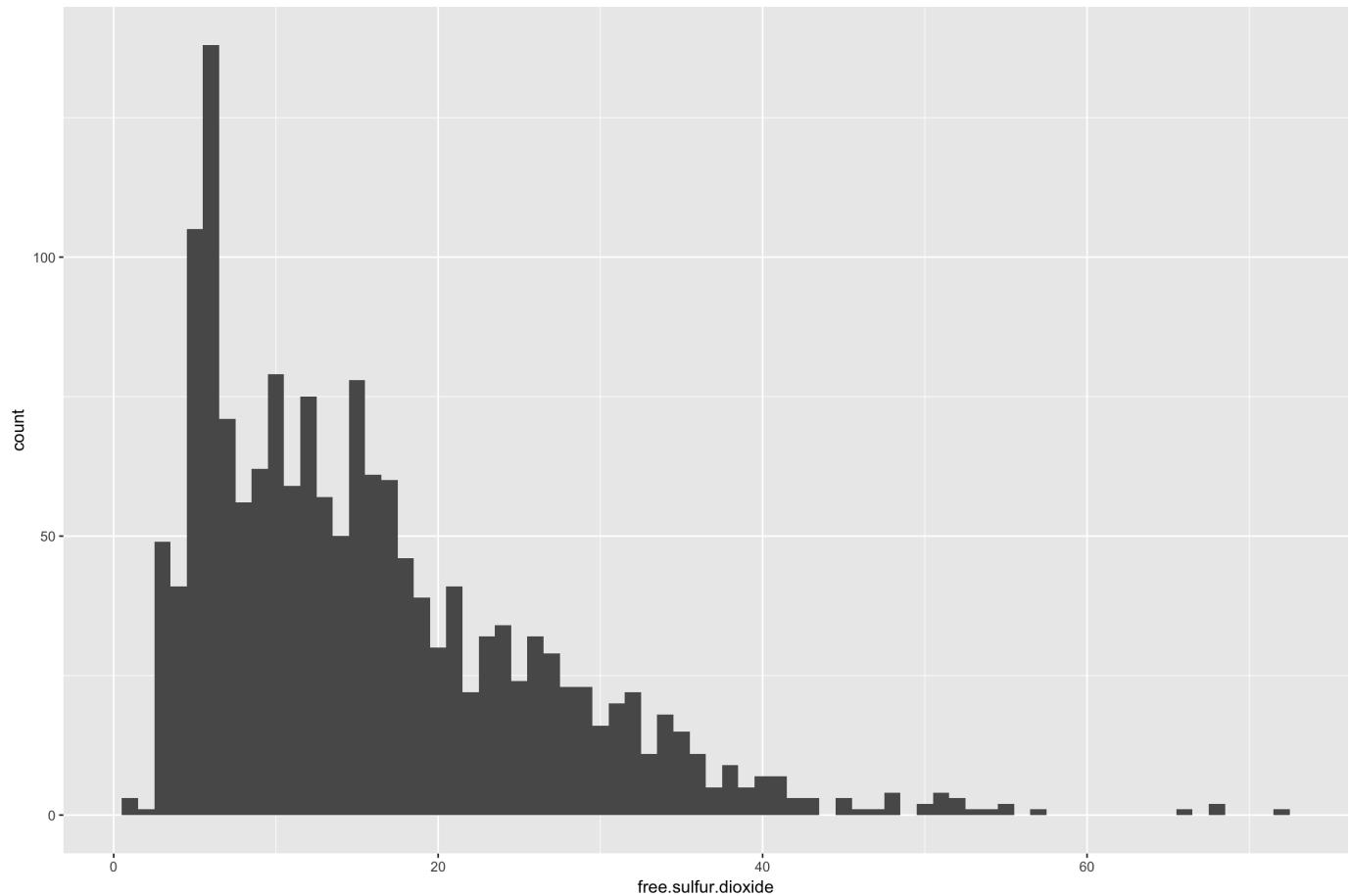


从上面的描述来看，我们看到这样低的二氧化硫水平并不令人惊讶。该数据集中75%的葡萄酒二氧化硫值低于62 mg / dm³

游离二氧化硫

在调查二氧化硫总量后，我们来研究游离二氧化硫属性 free.sulfur.dioxide，它可以防止微生物的生长和葡萄酒的氧化

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      1.00    7.00   14.00   15.87   21.00   72.00
```

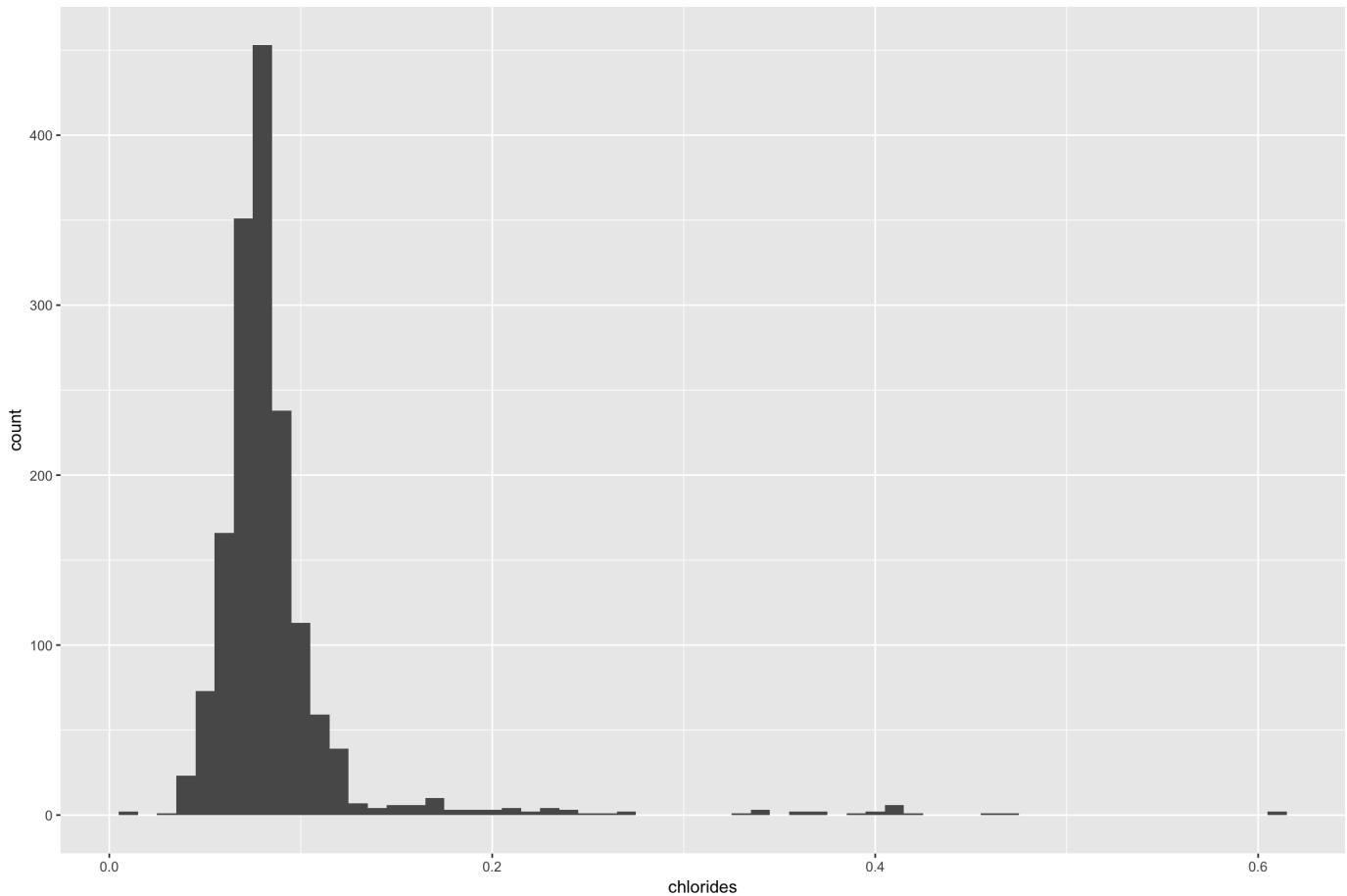


此图和二氧化硫的图形很相似，其中大部分值低于21 mg / dm³

氯化物

最后，我将研究氯化物的属性 chlorides，它给出了葡萄酒中盐的含量

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```



这个图也看起来像正态分布，但右侧有一个长尾。

2.2 单变量分析

2.2.1 你的数据集结构是什么？

这个整齐的数据集包含1599个红葡萄酒观测数据和12个属性。其中11个属性是葡萄酒的数字物理化学测试结果，1个属性（质量）由0到10的评价数据组成，这是一个分类变量，是葡萄酒专家至少3次评估的中位数。数据集中没有任何缺失值。

2.2.2 你的数据集内感兴趣的主要特性有哪些？

由于这个项目旨在找出哪些化学性质影响红葡萄酒的质量，其主要特点是质量。

2.2.3 你认为数据集内哪些其他特征可以帮助你探索兴趣特点？

根据我通过单变量分析的推测，有4个变量会影响葡萄酒的质量，故我下一步分析时会重点分析以下这些变量。

- 酒精：酒的酒精含量百分比
- 挥发性酸度：葡萄酒中醋酸的含量，如果含量过高，会导致不愉快的醋味
- 柠檬酸：少量发现，柠檬酸可以为葡萄酒添加“新鲜”和风味
- 二氧化硫：游离和结合形式的SO₂的量；在低浓度下，SO₂在葡萄酒中几乎检测不到，但当游离SO₂浓度超过50 ppm时，SO₂在葡萄酒的鼻子和口味中变得明显

2.2.4 根据数据集内已有变量，你是否创建了任何新变量？

我没有创建任何新变量。

2.2.5 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？

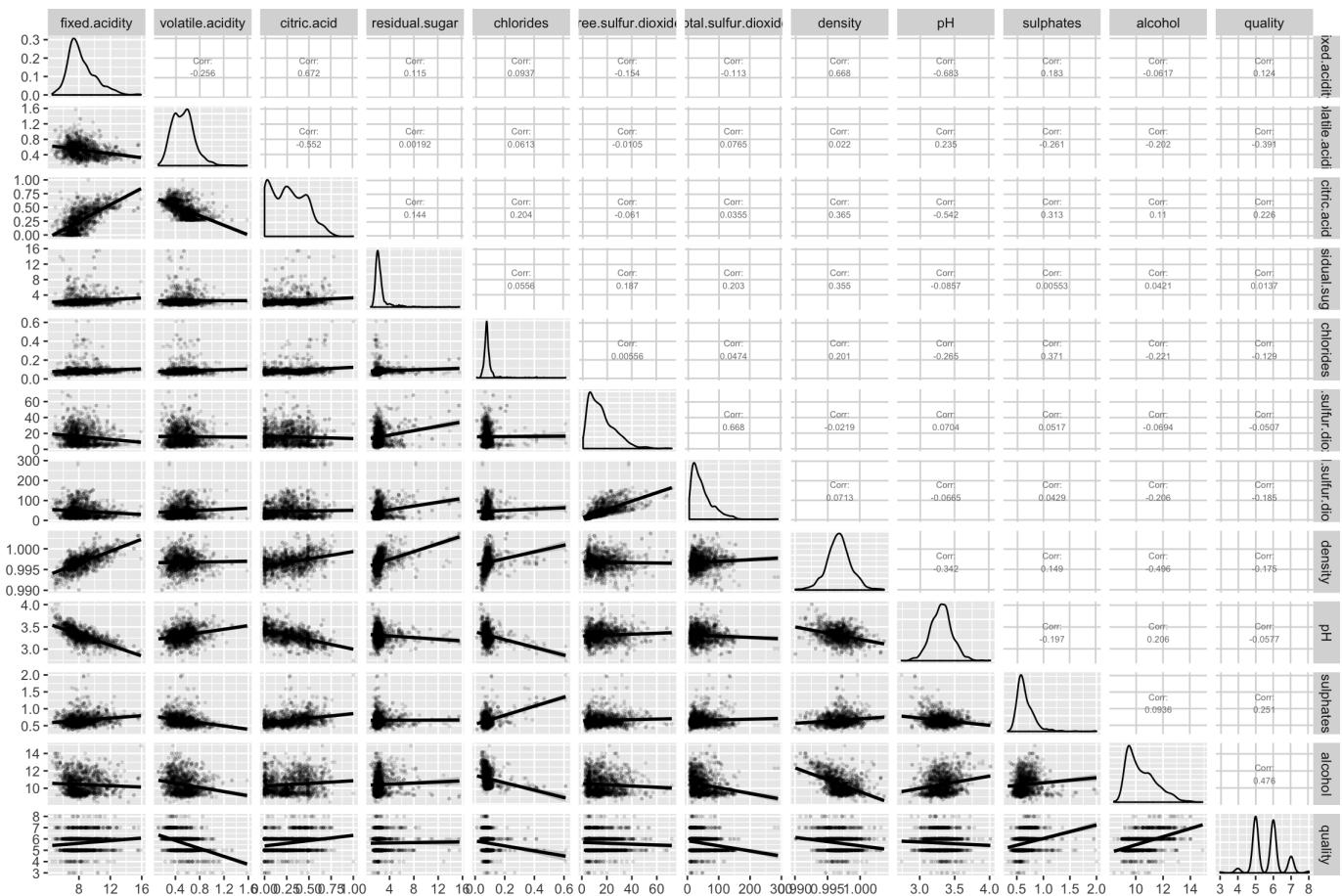
以下的这些图形有较长的长尾，可能是服从对数正态分布的：

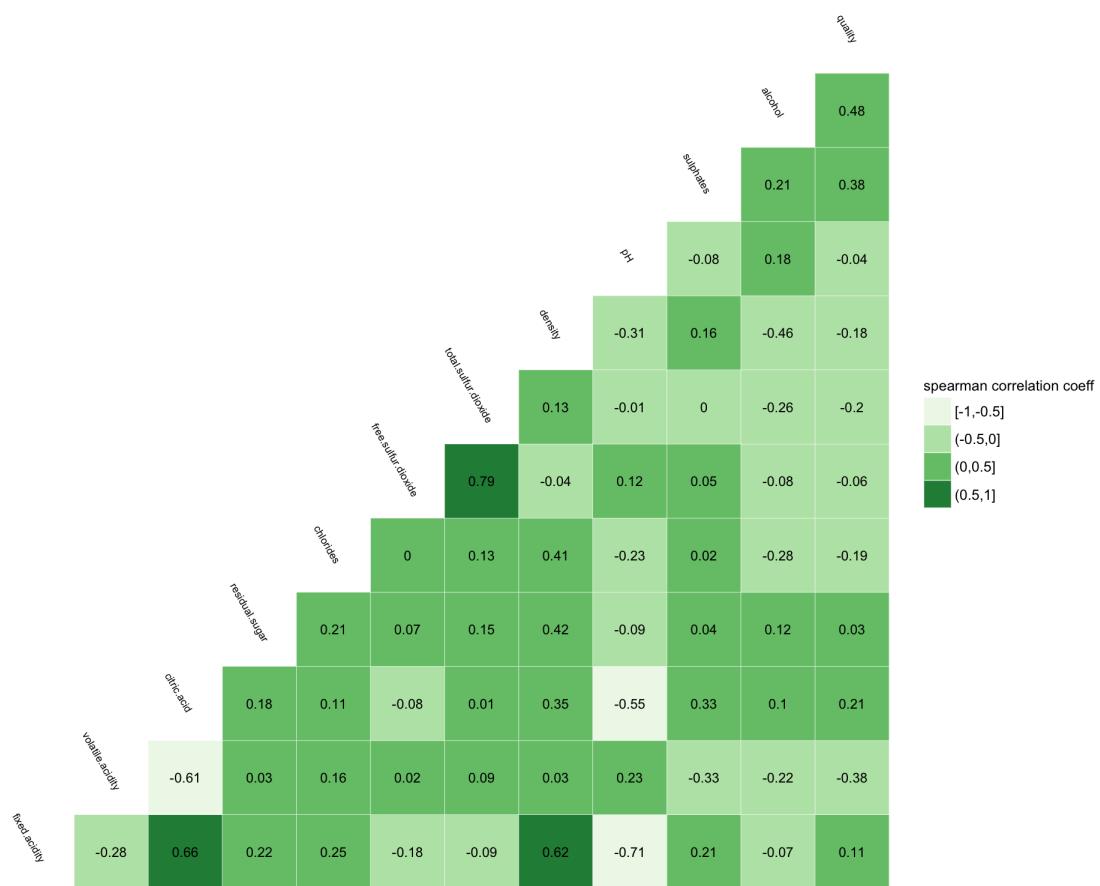
- 游离二氧化硫
- 二氧化硫总量
- 酒精含量
- 柠檬酸含量 由于给定的数据集非常干净，没有做额外的工作来整理数据集。

3 双变量部分

3.1 双变量绘图选择

首先，我们来研究变量之间的相关性。

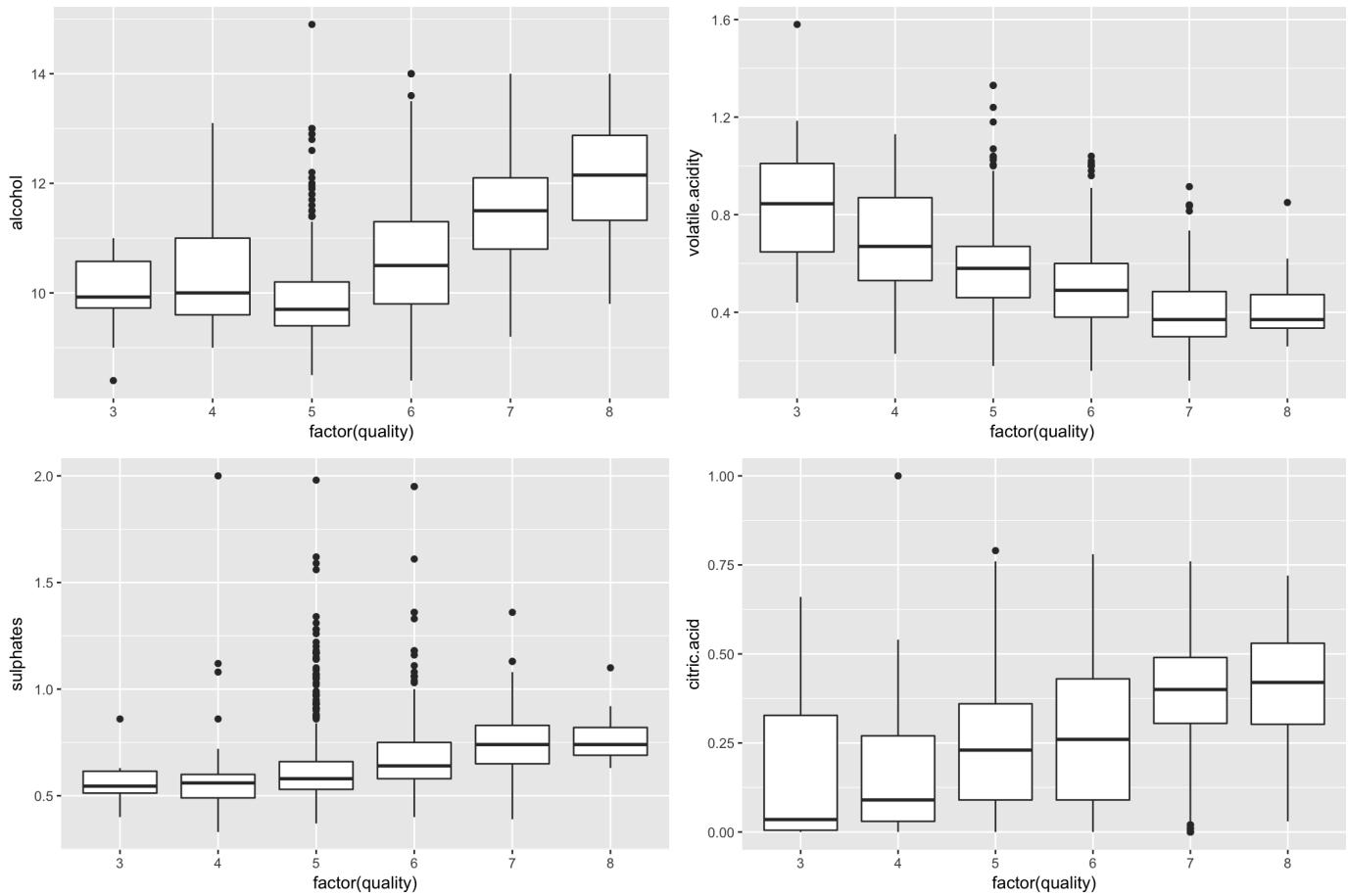




通过相关矩阵可以发现，数据集中多数自变量与因变量质量的相关性都非常弱。相关性最强的为酒精度与质量，相关系数0.48，可挥发酸与质量呈现负相关，相关系数为-0.39。其它变量与相关质量的相关系数绝对值都低于0.3。

酒精, 挥发酸, 柠檬酸和硫酸盐与质量

酒精, 挥发性酸度, 柠檬酸和硫酸盐是与质量最相关的属性。接下来, 我将挖掘这些变量以了解它们与质量的关系

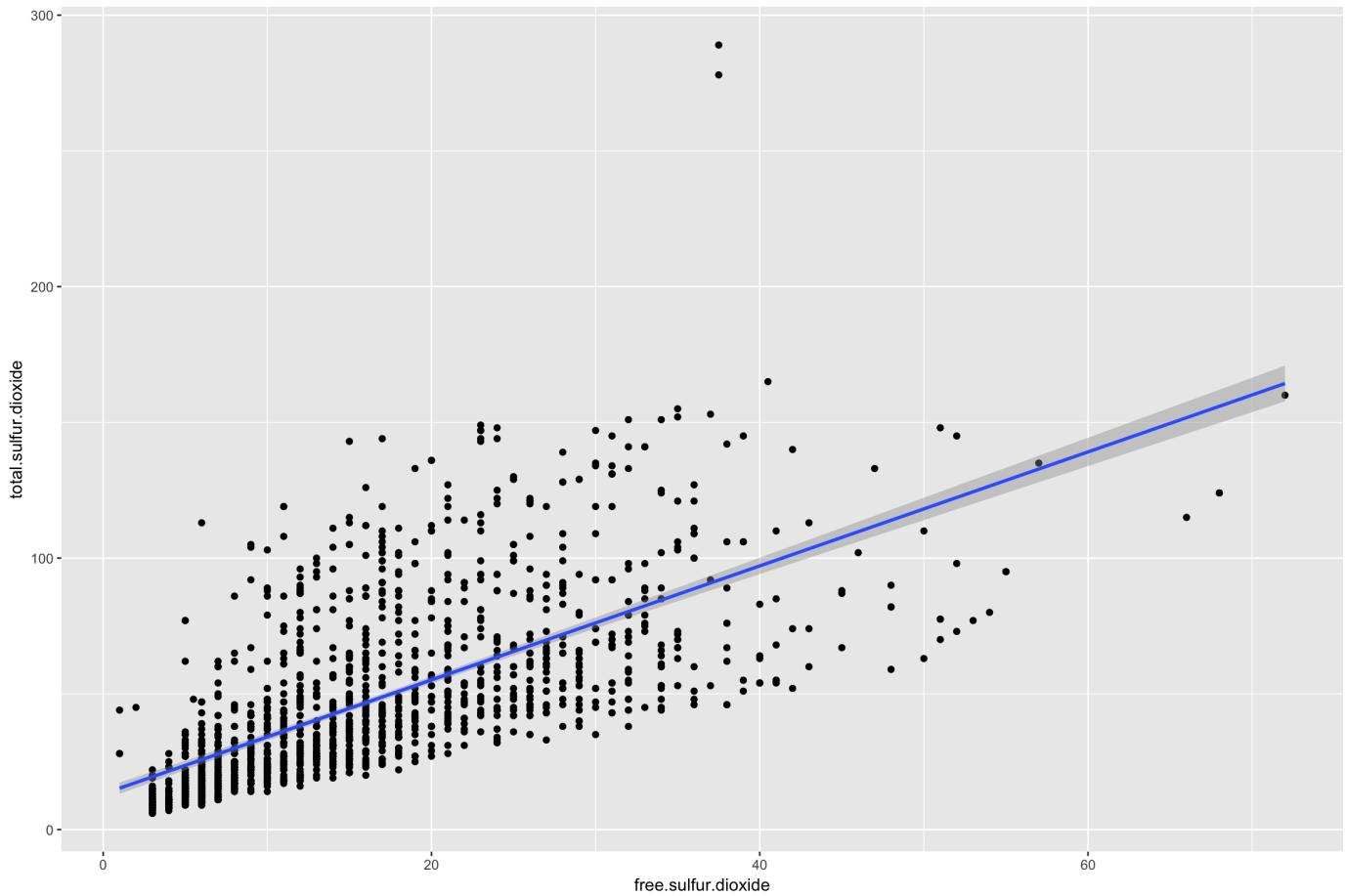


观察后我们发现质量与酒精, 柠檬酸和硫酸盐呈正相关, 与挥发性酸度呈负相关。

其他相关属性

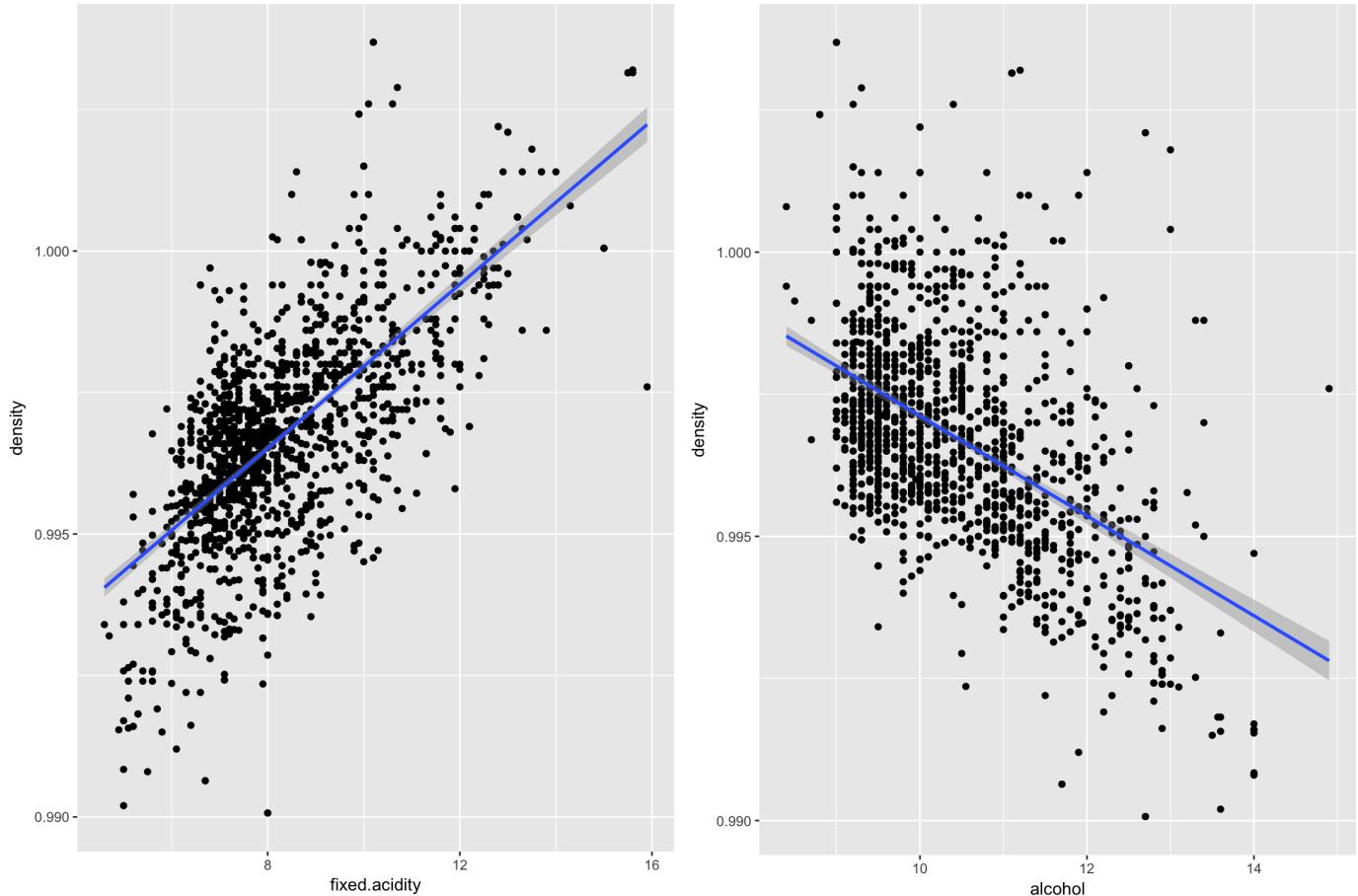
另外我们来查看下其他比较高的相关属性

二氧化硫总量 vs 游离二氧化硫



二氧化硫总量和游离二氧化硫属性彼此正相关，其值为0.668

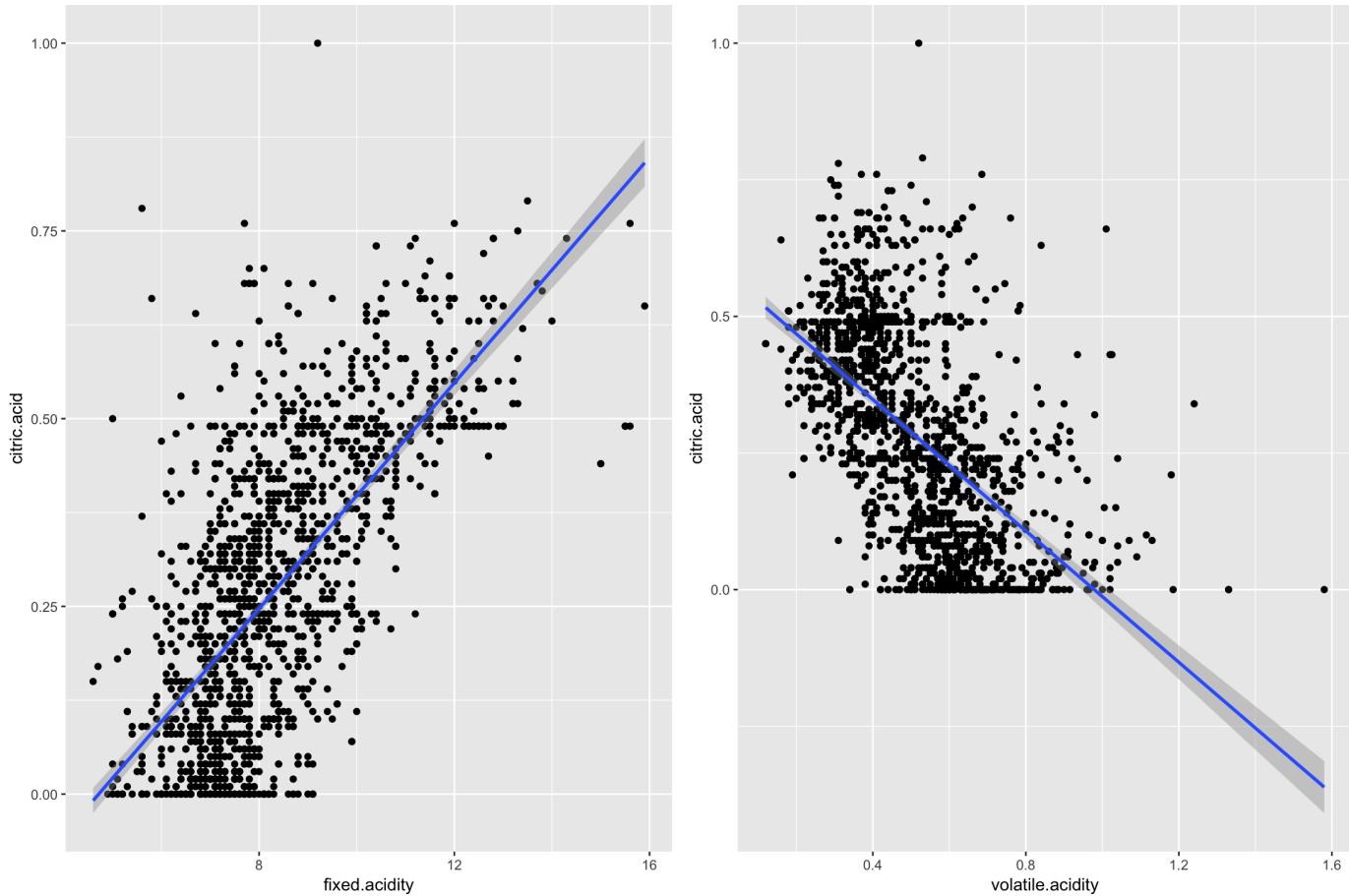
固定酸度与密度 & 酒精 VS 密度的关系



- 左侧的图显示了固定酸度和密度之间的强烈正相关关系，课件酒的固定酸度越高，其密度越高。

- 右侧的情节显示酒精和密度之间强烈的负相关。由于酒精的密度比水低，任何酒精百分比的增加都会降低密度。

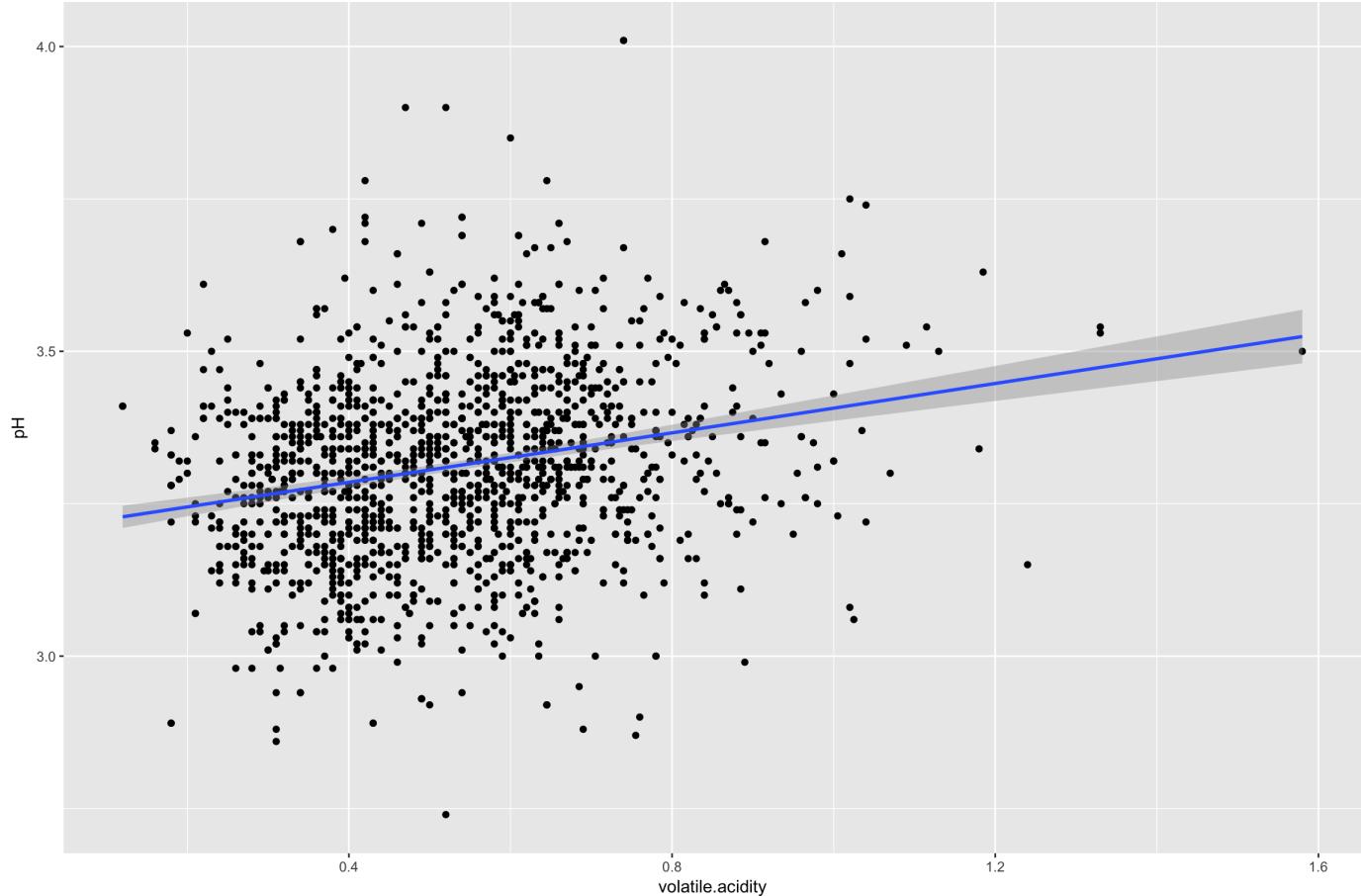
固定酸度 & 挥发酸度 vs 柠檬酸



柠檬酸和固定酸度具有正相关性，但挥发性酸度和柠檬酸之间存在负相关性。所以如果醋酸量（挥发性酸度）增加，我们预计柠檬酸减少，反之亦然。

挥发酸度 vs pH值

我们再来看下挥发酸度与pH值。



虽然这两个变量的相关系数为0.235，但看起来它们之间的相关性很弱。只是由于个别数据具有较高的挥发酸度，导致整个图形的相关性成正比例关系。

3.2 双变量分析

3.2.1 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中其他特性有什么区别？

在调查ggpair情节后，我发现4个属性与质量有关：

- 酒精百分比直接影响质量。大多数时候，更高的酒精率意味着更好的优质葡萄酒。
- 挥发性酸度与质量成反比。挥发酸度越低，葡萄酒的质量越好。
- 硫酸盐和质量有着弱正相关的关系，因此硫酸盐的增加可能会使葡萄酒变得更好
- 虽然柠檬酸与质量相关，但这两者之间的关系在统计学上并不显著
- 在去除柠檬酸属性后，其余3个变量解释了33.46%的质量变化。

3.2.2 你是否观察到主要特性与其他特性之间的有趣关系？

观察挥发性酸度和柠檬酸之间的负相关性是非常有趣的。

由于挥发酸度和pH值呈正相关，因此该图也非常有趣。

3.2.3 你发现最强的关系是什么？

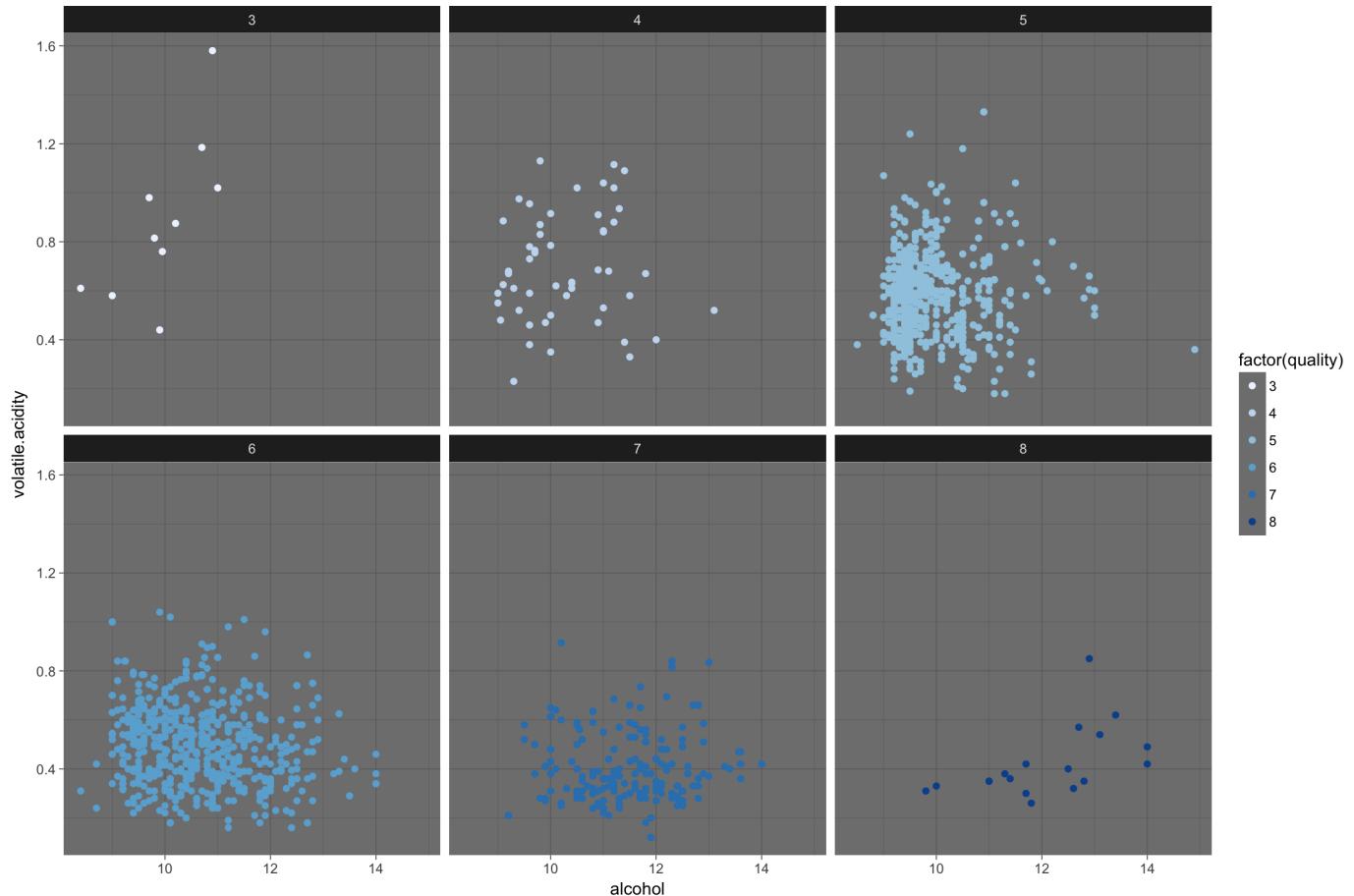
pH值与固定酸度之间最强的相关性为-0.683

4 多变量部分

4.1 多变量绘图选择

酒精 & 挥发酸度 VS 质量

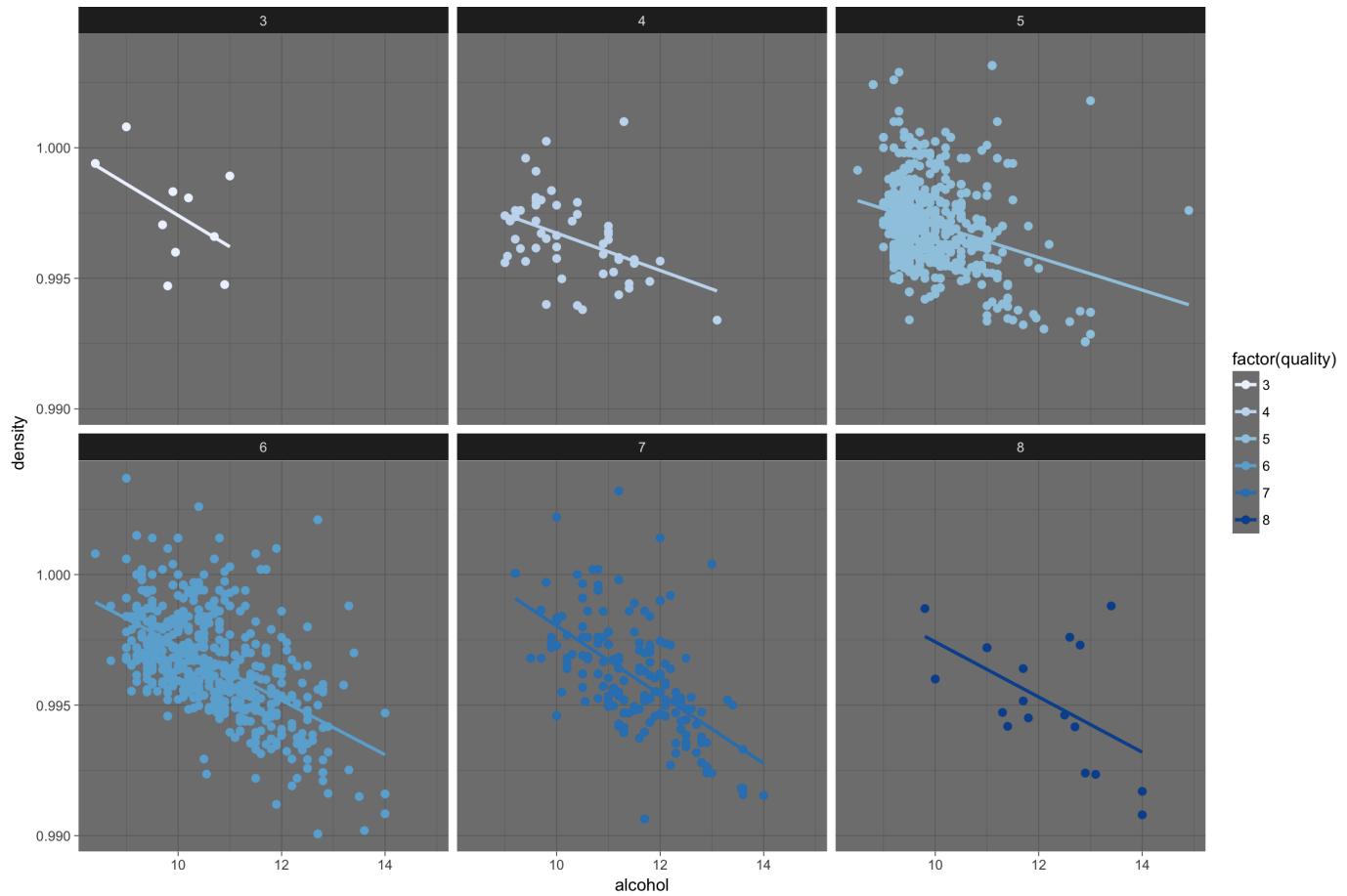
红酒的口感中酒精和挥发酸度的影响最为主要因素，图中蓝色颜色越深表示酒的品质越好



从图中可发现品质越好的酒越拥有更高的酒精度和更低的挥发酸度

密度 & 酒精 vs 质量

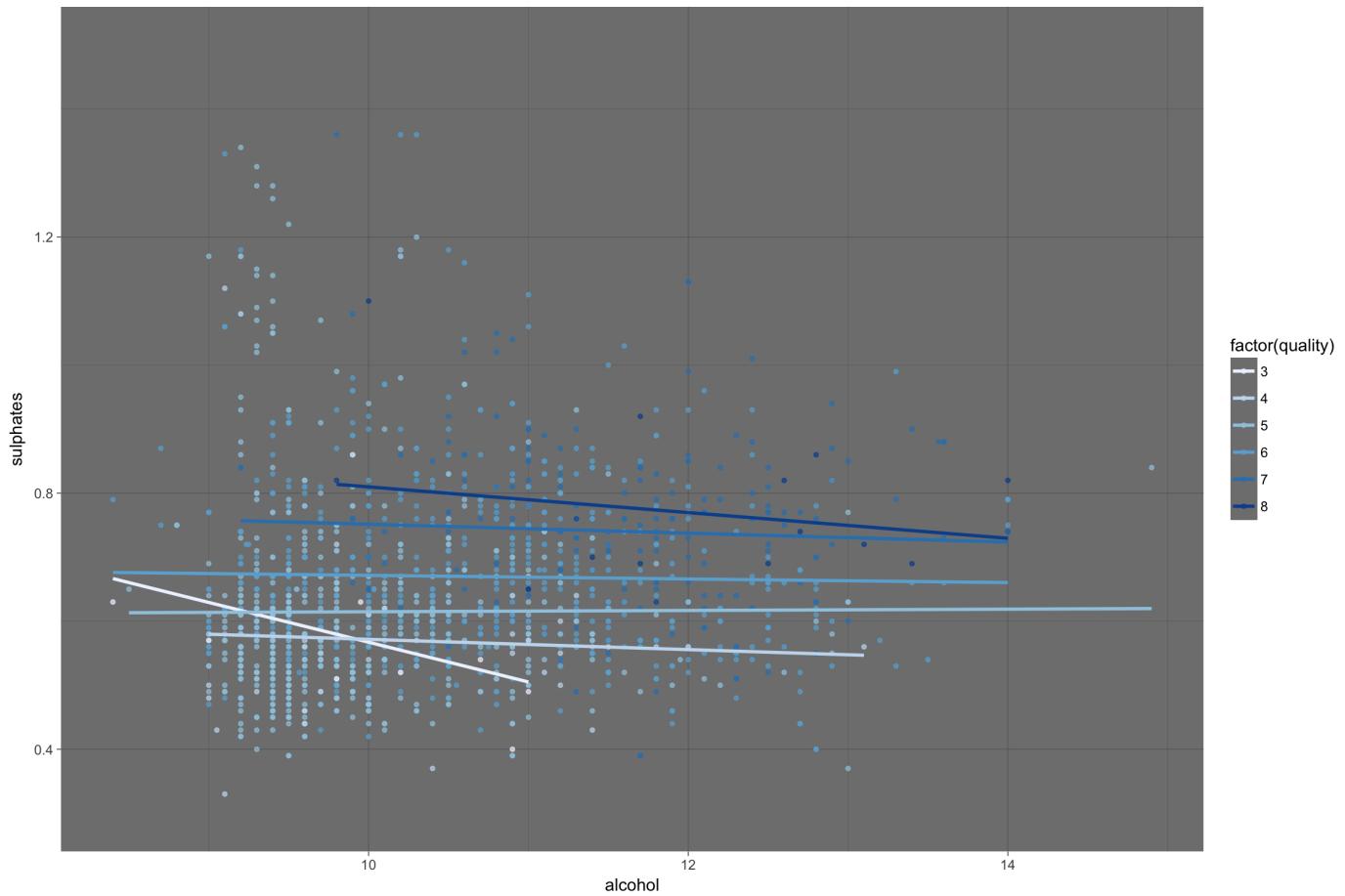
密度和酒精的关系



图形的状态也说明了我的猜测密度和酒精的含量是成反比的

硫酸盐 & 酒精 vs 质量

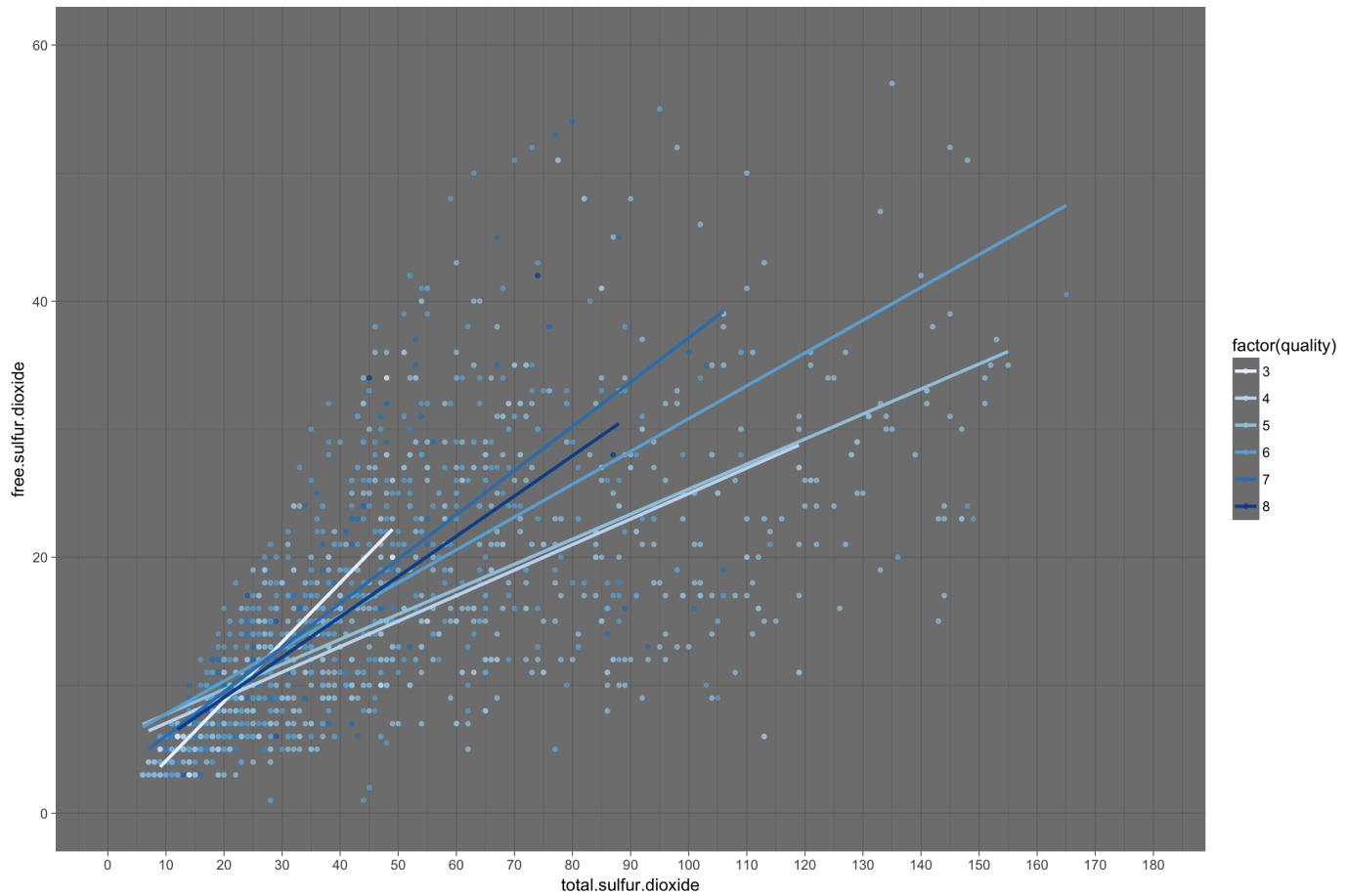
硫酸盐会影响红酒的口味，我们来看一下硫酸盐与酒精和红酒质量的关系



从图中可以看出，更高评分的红酒拥有更高的硫酸盐含量，说明硫酸盐是影响红酒品质的重要因素，其含量与红酒品质成正比

游离二氧化硫和二氧化硫总量对红酒品质

我们再来看一下二氧化硫对于红酒品质的影响



4.2 多变量分析

4.2.1 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？

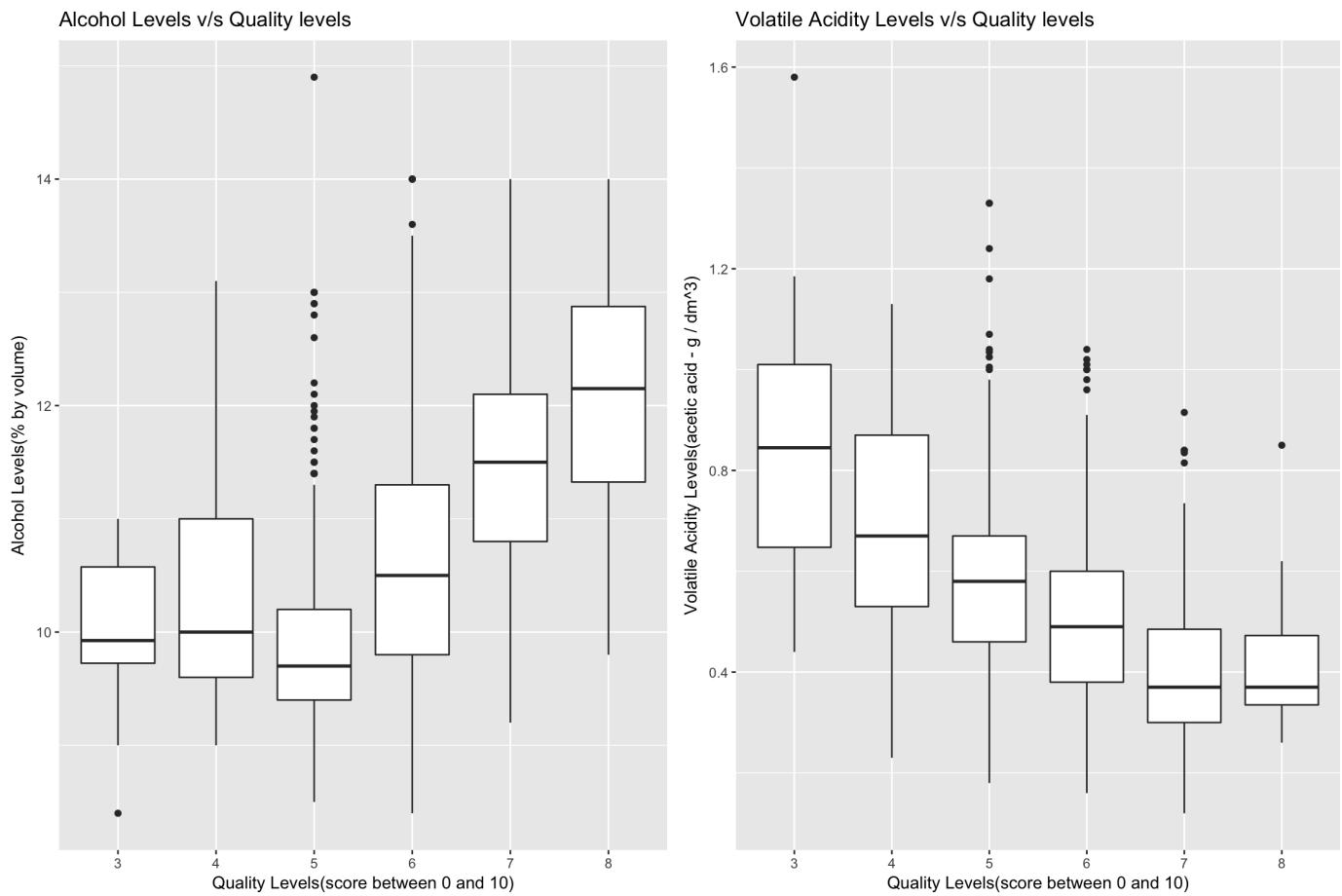
在这部分调查后我们发现，一下的因素对于制作高品质葡萄酒有所帮助： - 高酒精率 - 高硫酸盐量 - 低挥发酸度
- 低二氧化硫水平

4.2.2 这些特性之间是否存在有趣或惊人的联系呢？

酸度和酒精对于红酒的品质的关系十分有趣，这二者直接并显著影响红酒的质量，而硫酸盐和二氧化硫水平的高低对红酒的质量产生分层，我猜想这是不同红酒各有独特口味的重要原因。

5 定稿图与总结

5.0.1 绘图一



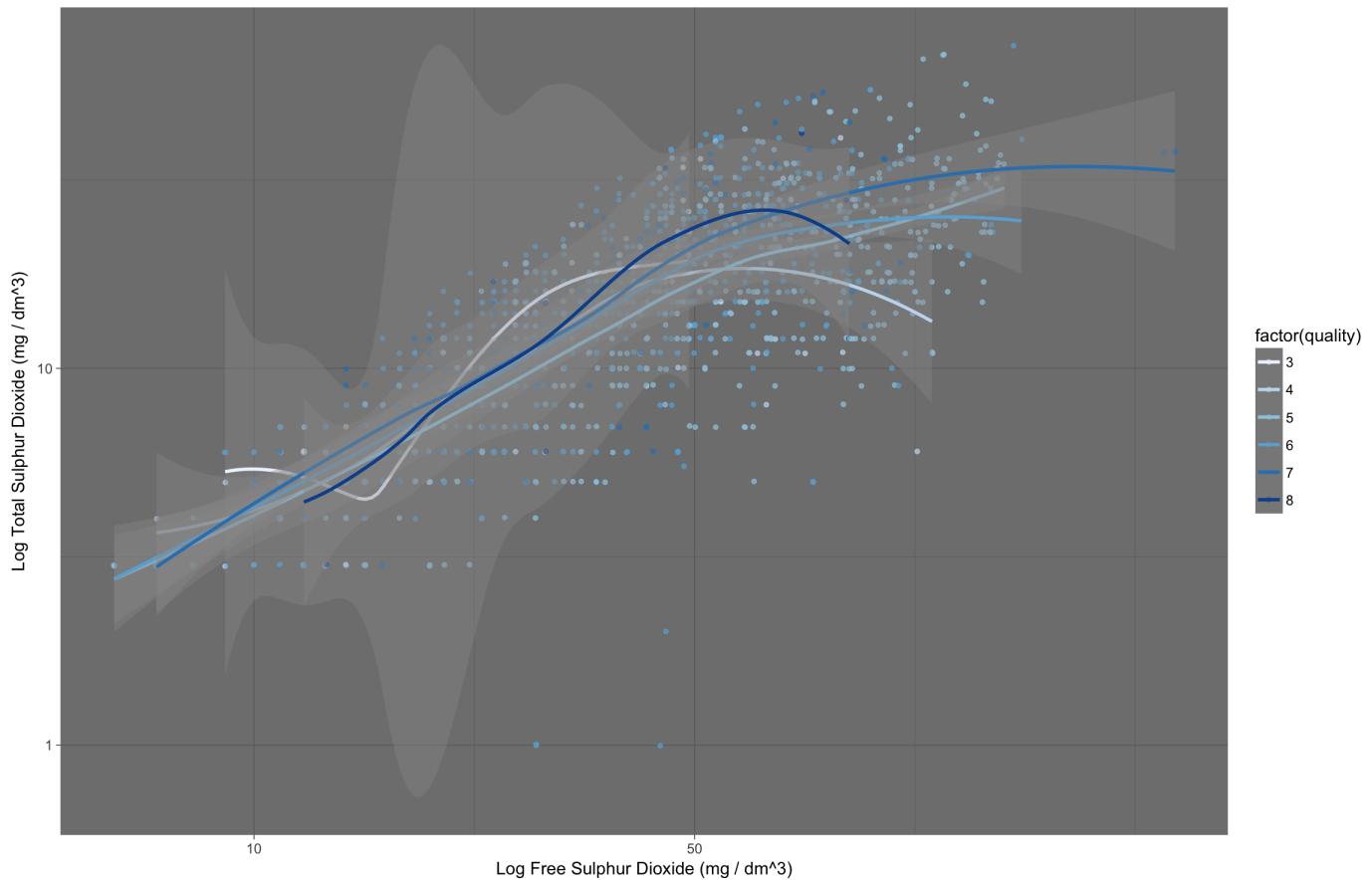
5.0.2 描述一

我们来检查酒精水平以及挥发酸度水平和质量水平之间的关系。

该图显示随着质量水平的提高，酒精含量不断增加。所以，酒量的增加可能意味着质量的提高。而随着质量水平的提高，挥发酸度不断减少。所以，挥发酸度的增加可能意味着质量的下降。

5.0.3 绘图二

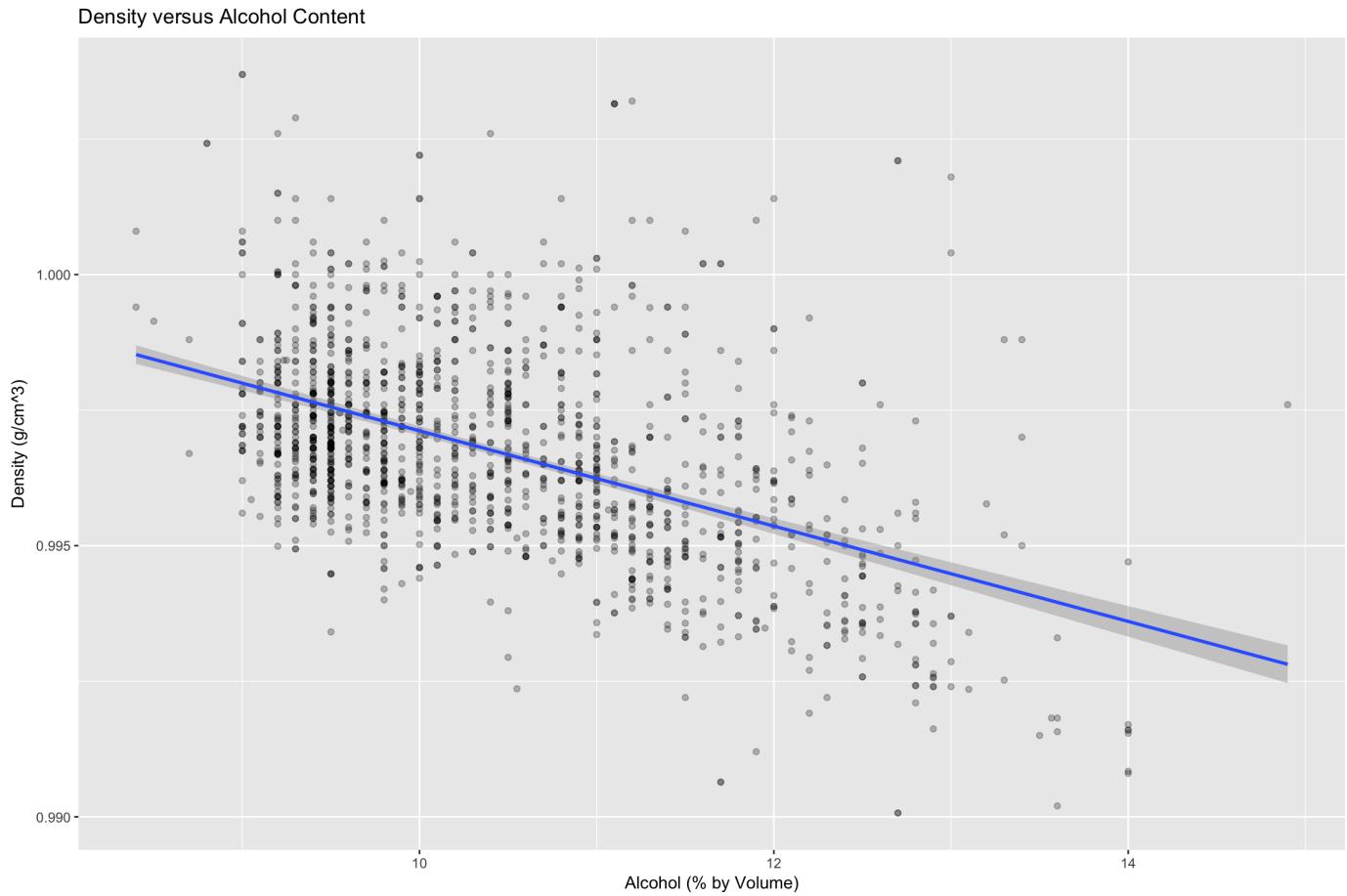
Total_SO2 Versus Free_SO2



5.0.4 描述二

二氧化硫总量与游离二氧化硫含量呈指数线性关系，从点的分布看，较低的二氧化硫含量更倾向于有助于提升葡萄酒质量。

5.0.5 绘图三



5.0.6 描述三

显而易见酒精的含量很大程度决定了红酒的密度，因为酒的密度小于水，所以酒精含量水平越高的样本其密度越低，这也与我们的常识一致。

6 反思

在整个数据探索和分析过程中，我了解到了影响红酒质量的因素。在进行分析之前只是感性的认识与猜想，而通过可视化的表达，有些猜想被证明了，有些结论让我重新认识和了解了各个要素之间的相关性。

我在进行三变量分析过程中，决定讲不同质量水平的红酒采用分块的形式展示，这使得我们更加清楚的能够看出酒精及酸度对于红酒的影响。

这次分析中主要分析了酒精、酸度、硫化物等权重较大的影响因素。对于残糖、氯化物等因素的分析说不定也会有令人惊喜的发现。而即使就相关性较高的部分而言，这些因素的成因又受哪些因素影响？是否还与原材料产地、气候、贮藏环境等相关？这些因素是如何影响红酒质量的？这也可以在后续的资料收集后进行详细研究与讨论。