红葡萄酒质量特性探索

张然 2018-04-15

- 数据集来源 http://dx.doi.org/10.1016/j.dss.2009.05.016 (http://dx.doi.org/10.1016/j.dss.2009.05.016)
- 要探索的数据集包含1,599种红酒和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0非常差和10非常好之间
- 以期了解哪些成分影响了酒的质量等

> P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
> Modeling wine preferences by data mining from physicochemical properties.
> In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

数据集预览

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

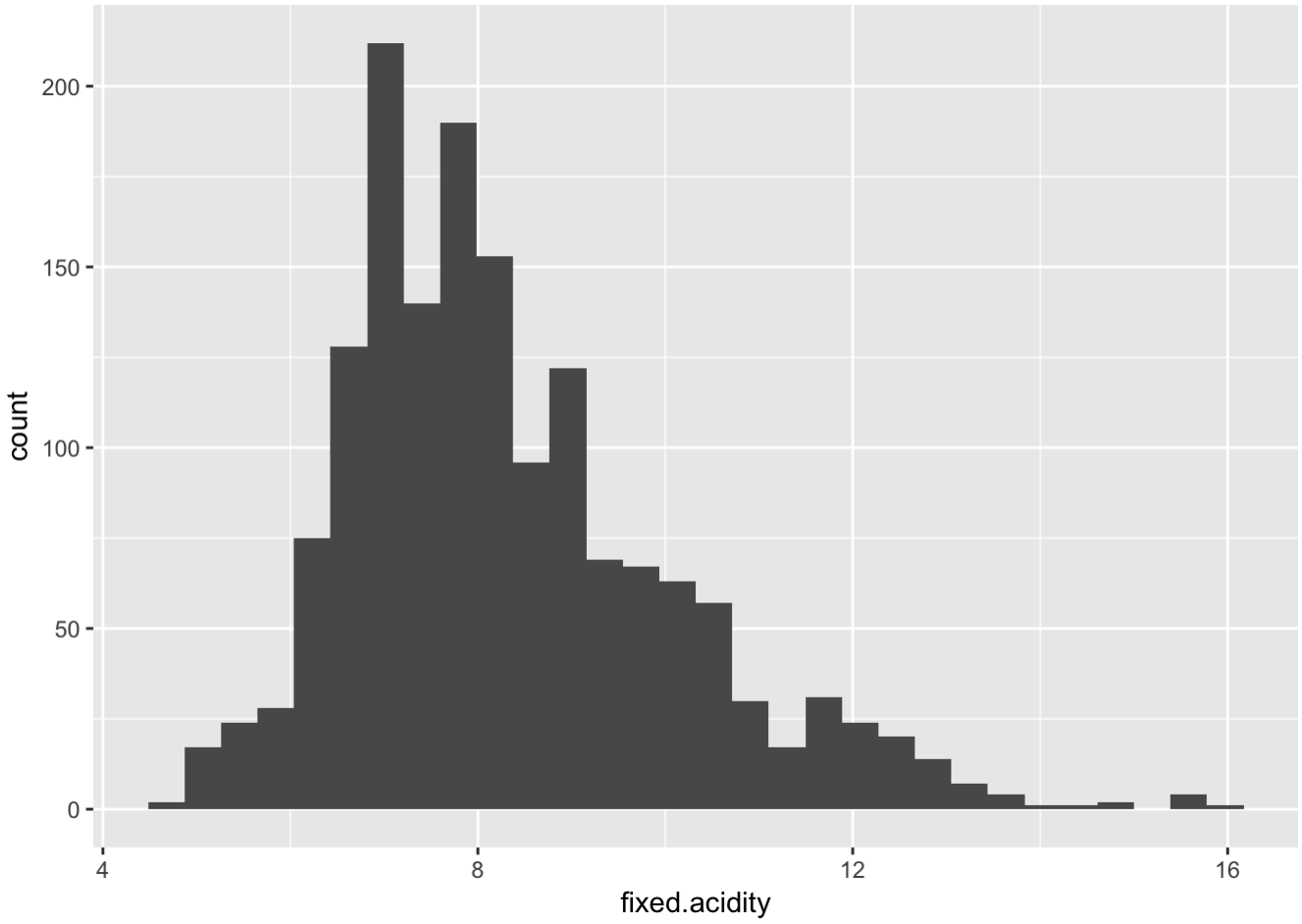
```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 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
```

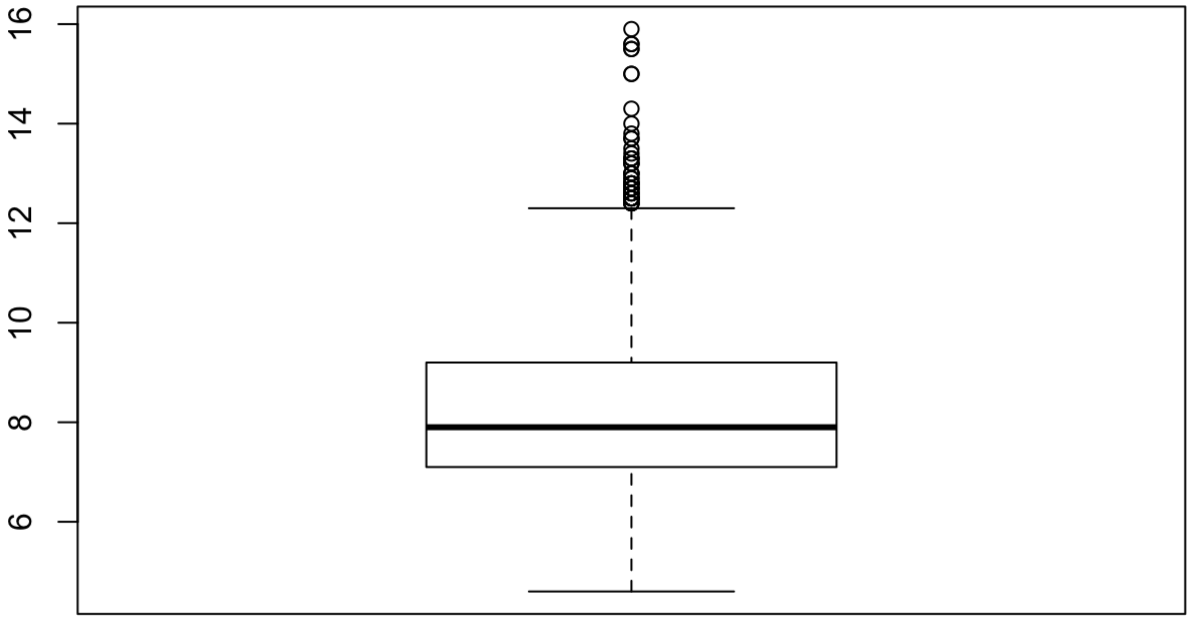
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  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 0.07
3 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 ...
```

这些变量的属性信息

- 0 这个数据集的结构有 1599 个观察和 13 个变量
- 1 固定酸度 - fixed acidity (tartaric acid - g / dm^3)
- 2 挥发性酸度 - volatile acidity (acetic acid - g / dm^3)
- 3 柠檬酸 - citric acid (g / dm^3)
- 4 残糖 - residual sugar (g / dm^3)
- 5 氯化物 - chlorides (sodium chloride - g / dm^3
- 6 游离二氧化硫 - free sulfur dioxide (mg / dm^3)
- 7 总二氧化硫 - total sulfur dioxide (mg / dm^3)
- 8 密度 - density (g / cm^3)
- 9 pH 值 - pH
- 10 硫酸盐 - sulphates (potassium sulphate - g / dm^3)
- 11 酒精度 - alcohol (% by volume)
- 12 质量评级 - quality (score between 0 and 10)

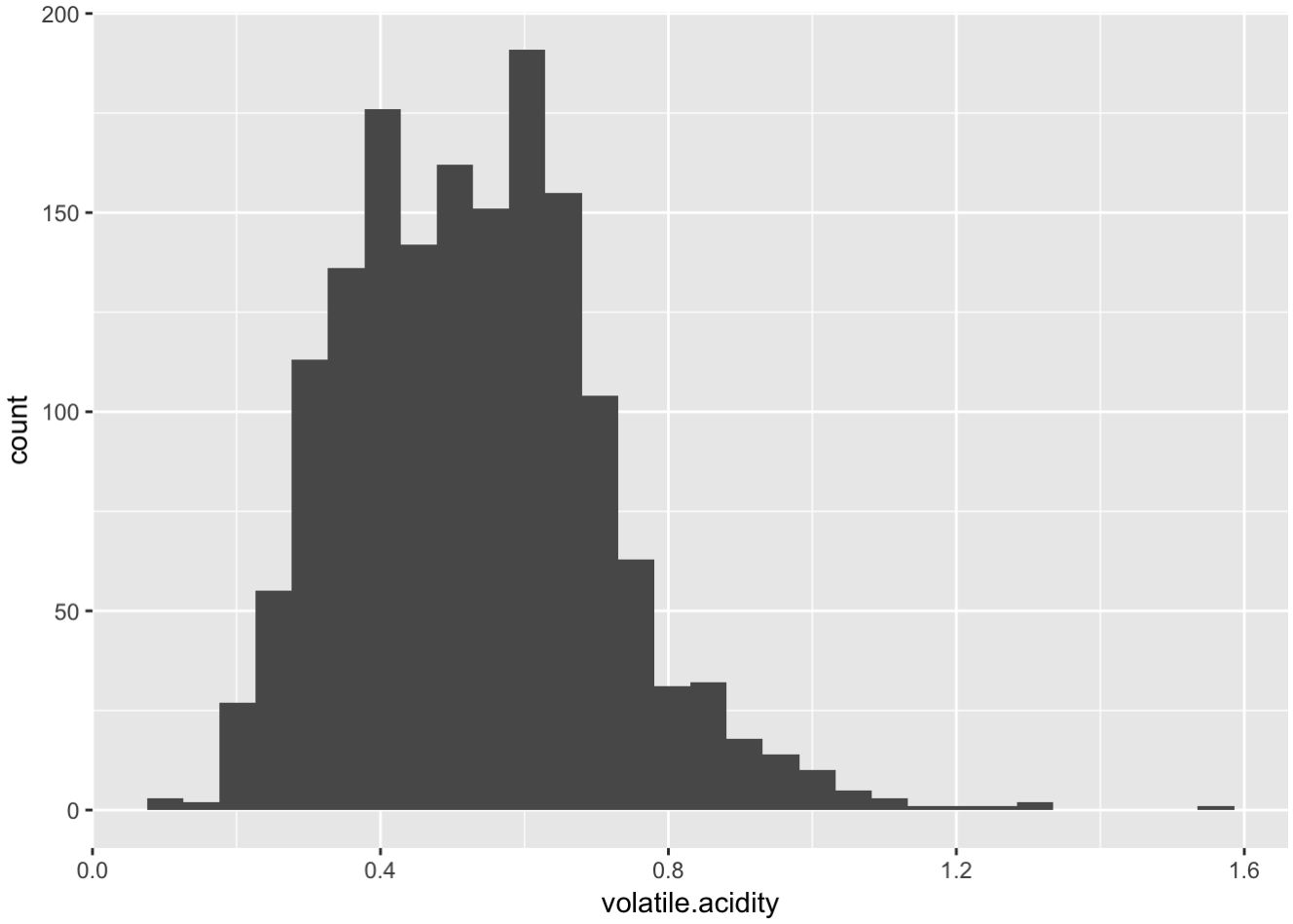
单变量绘图选择和分析

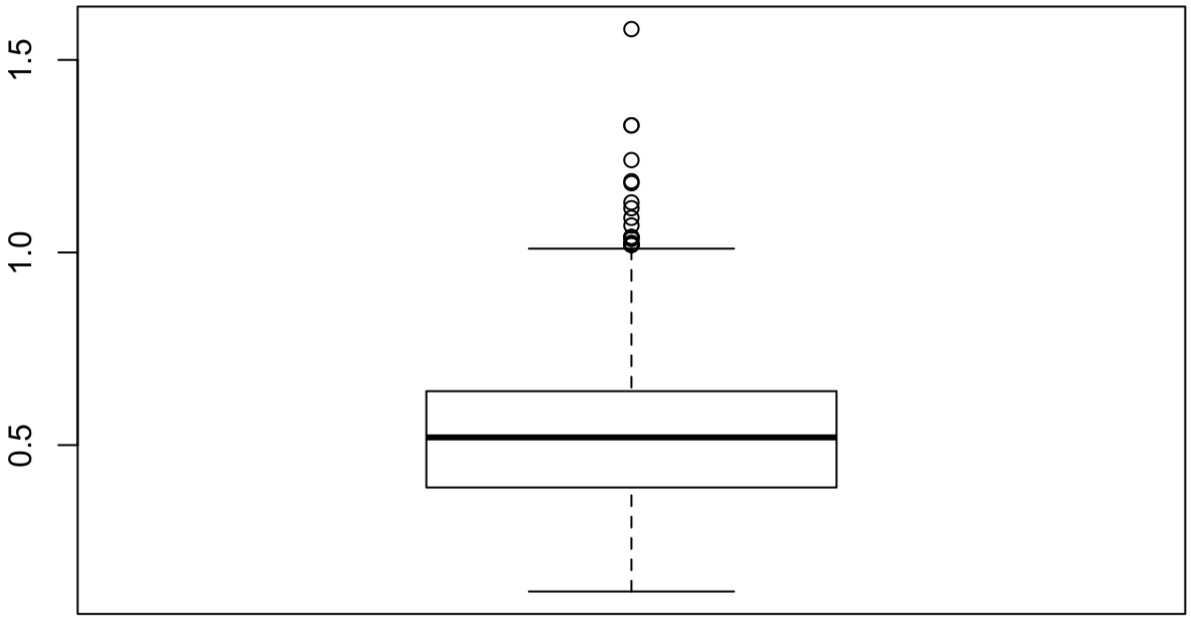





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

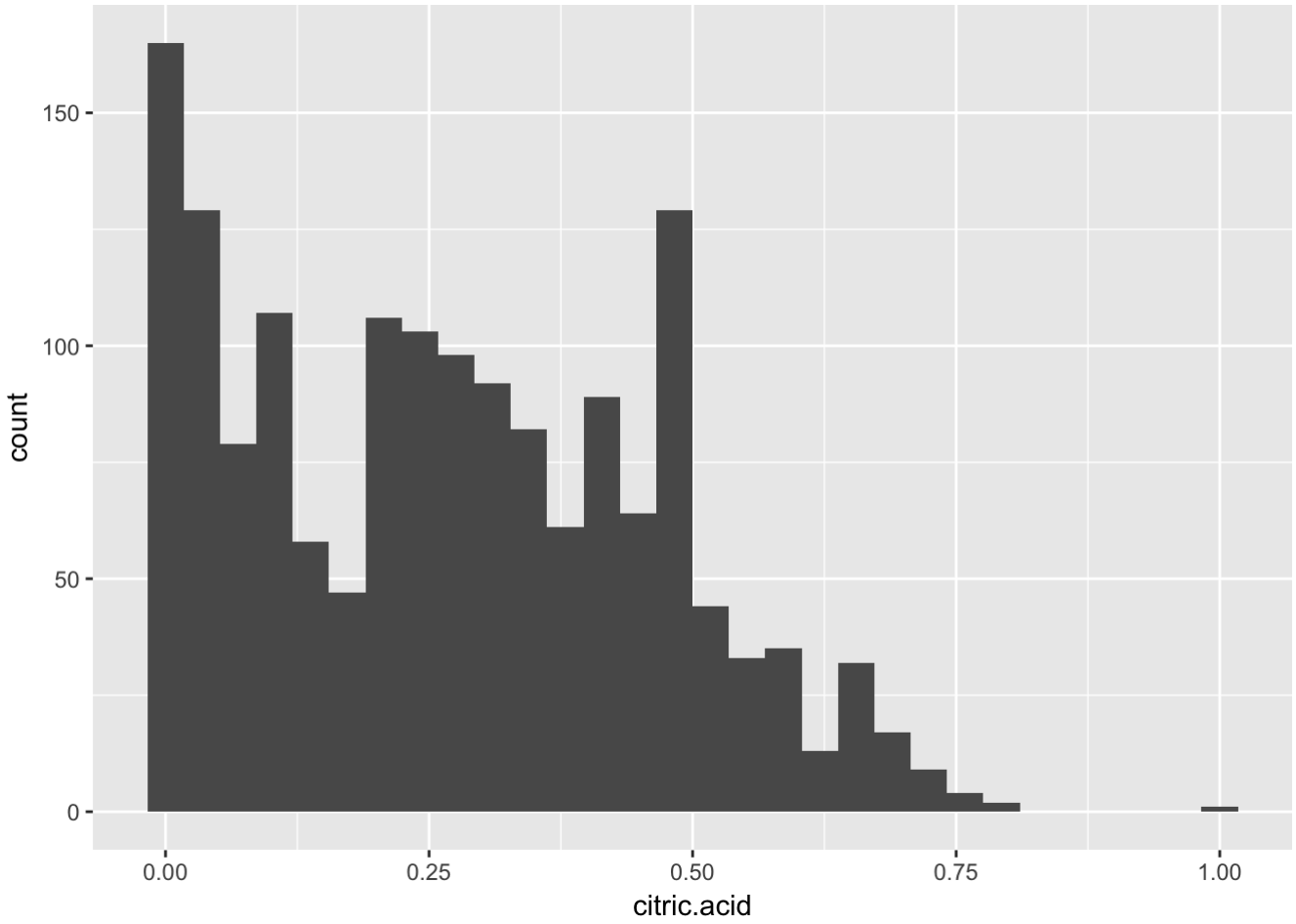
固定酸度呈现出正偏态分布，不考虑异常值的情况下数据主要分布在7到10之间

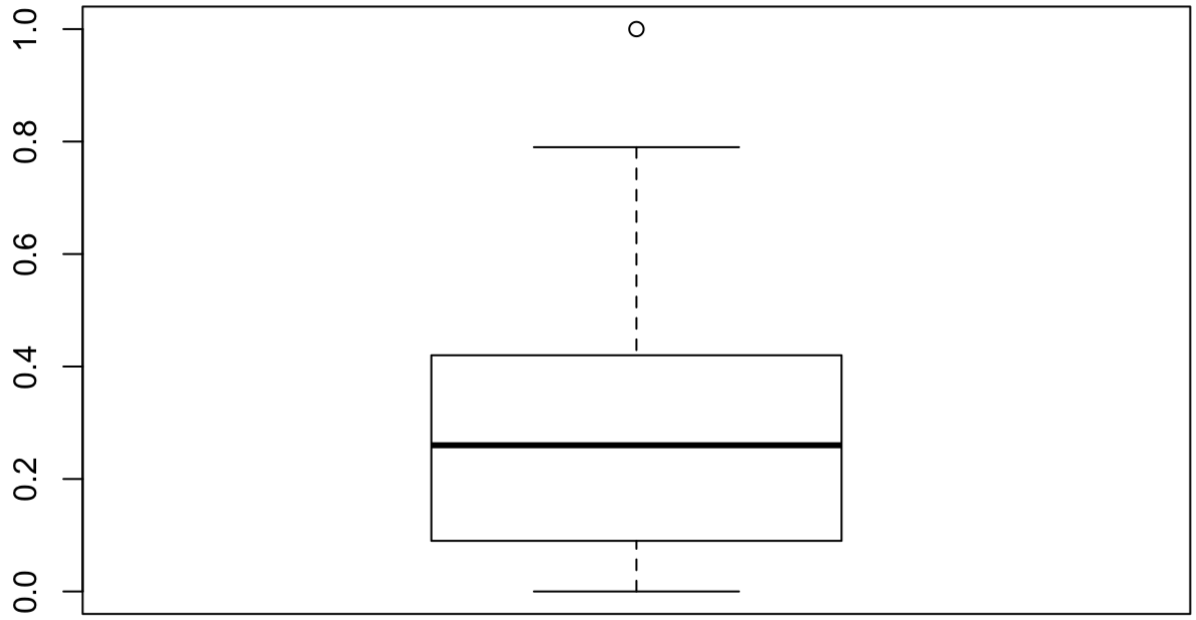





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

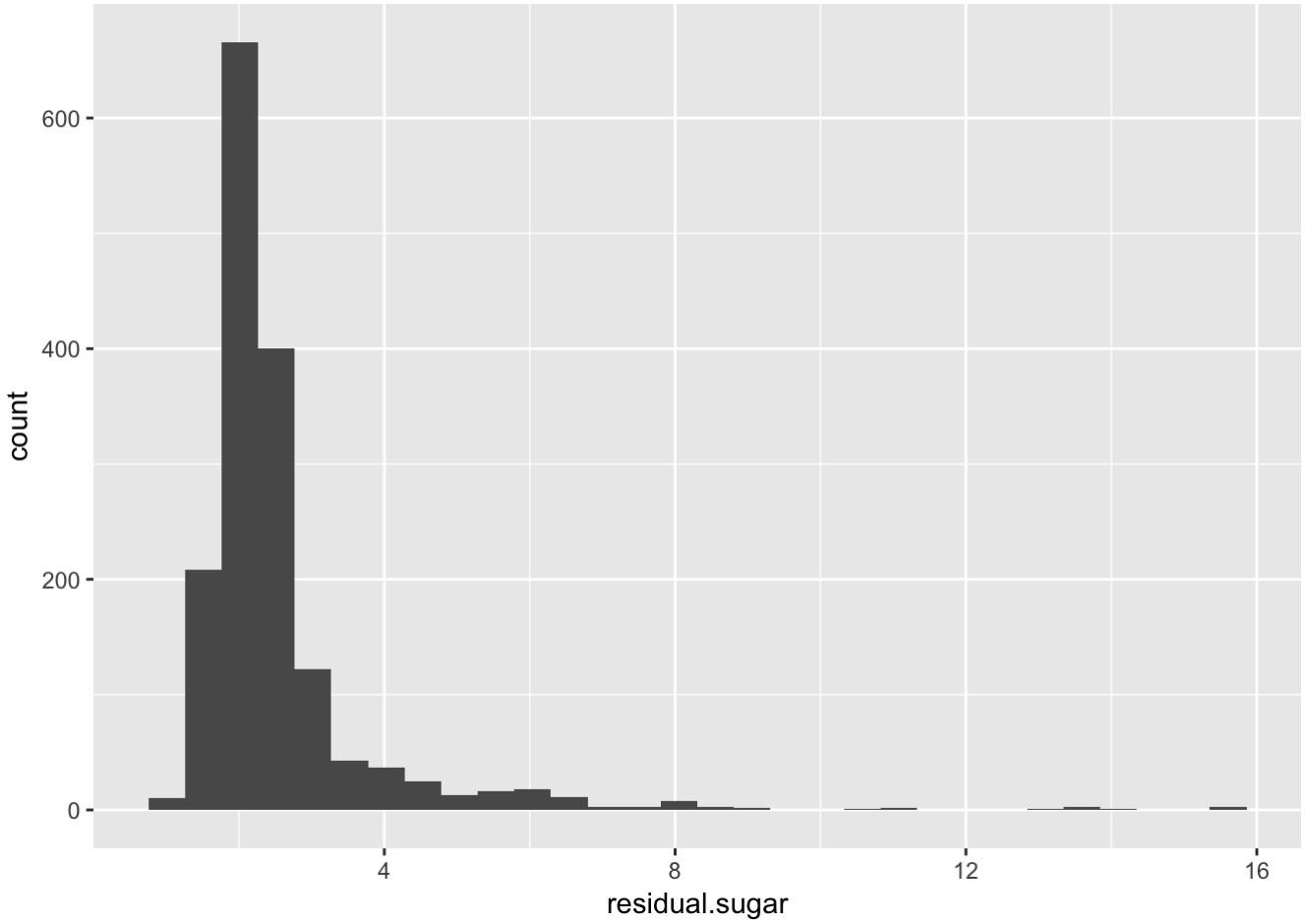
挥发性酸度的分布较为集中，主要在0.5附近，0.8到1.6之间的宽度大数据分布很少





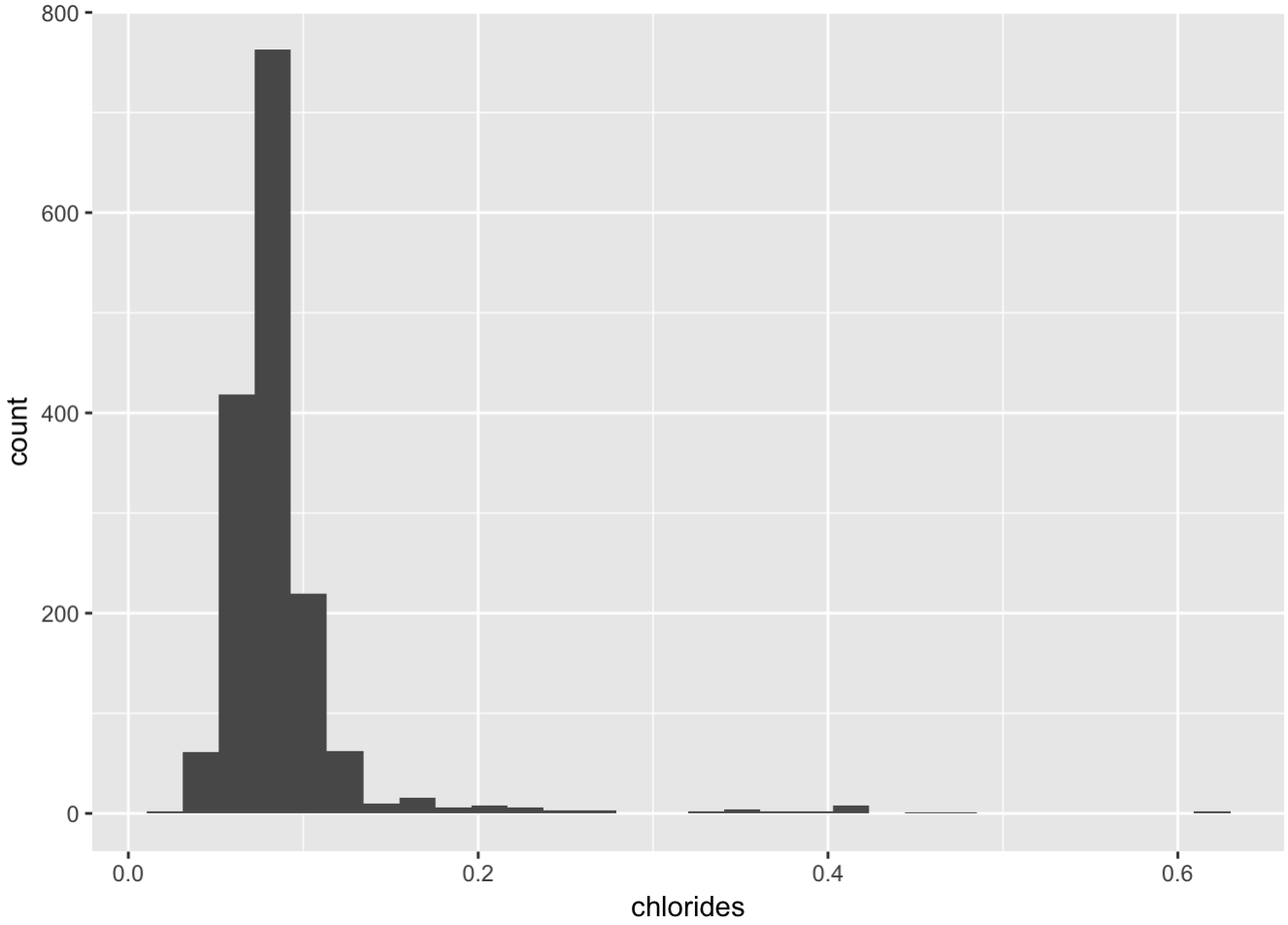

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

柠檬酸的分布主要在0.5以内，这个数据集的异常值很少都集中在1.0附近，方便做统计分析的时候排查数据噪音



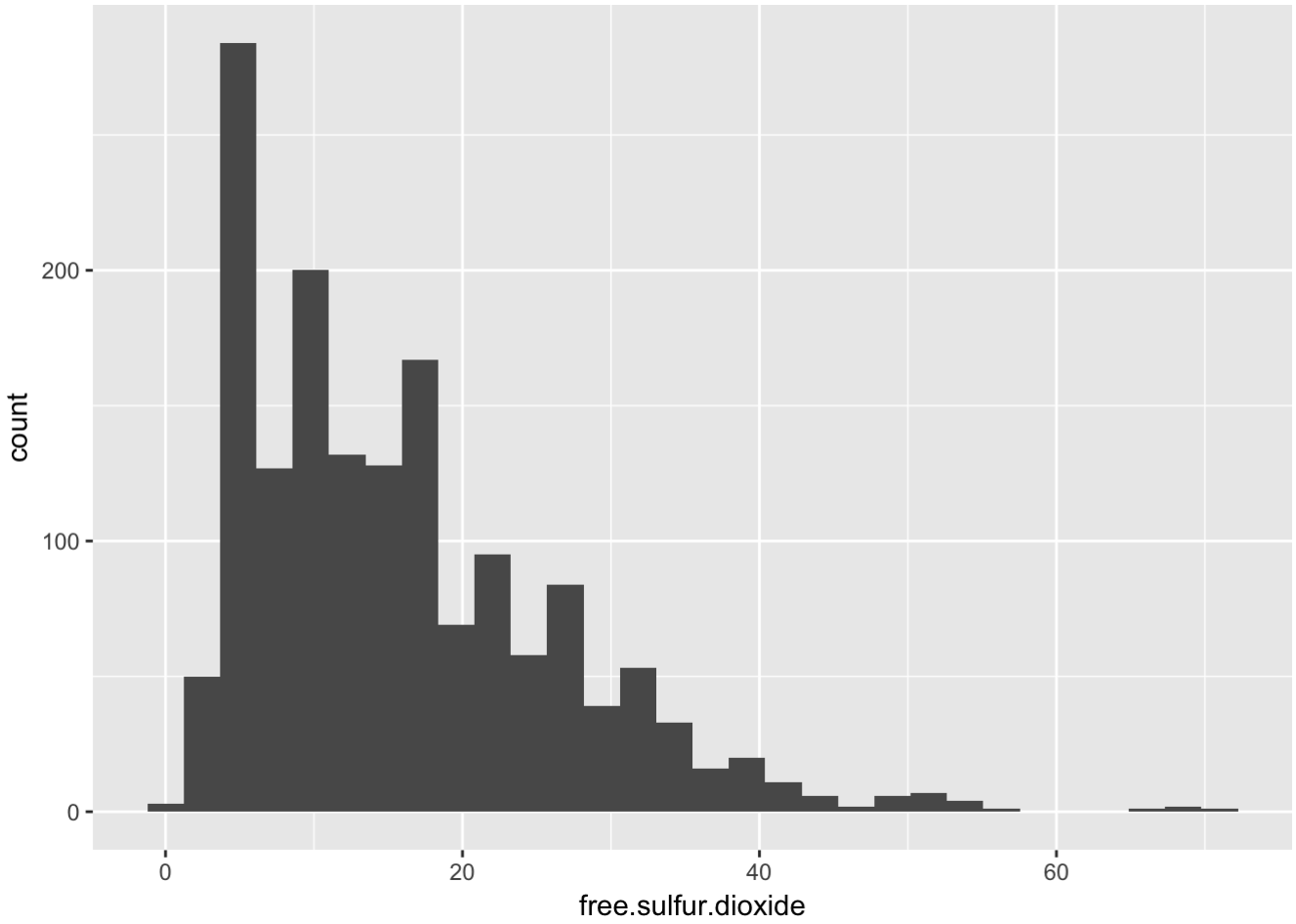

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

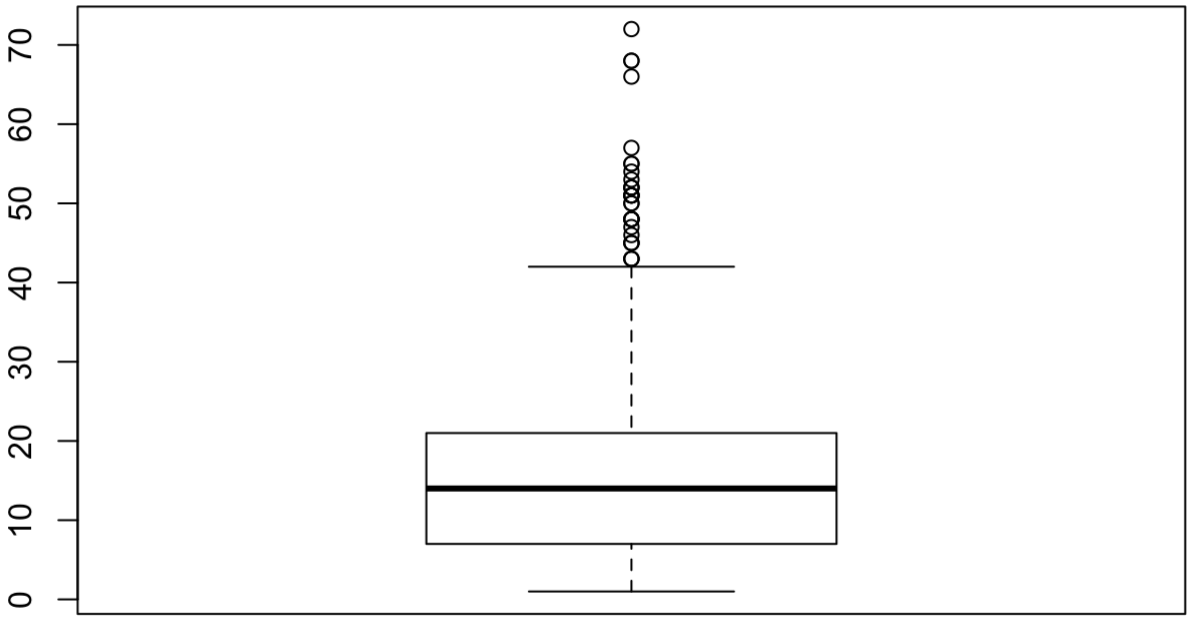
残糖的分布异常集中在2附近，4到16之间的数据量很少，因该是现实中大部分红酒的残糖含量比较稳定，个别的异常值不是主流




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

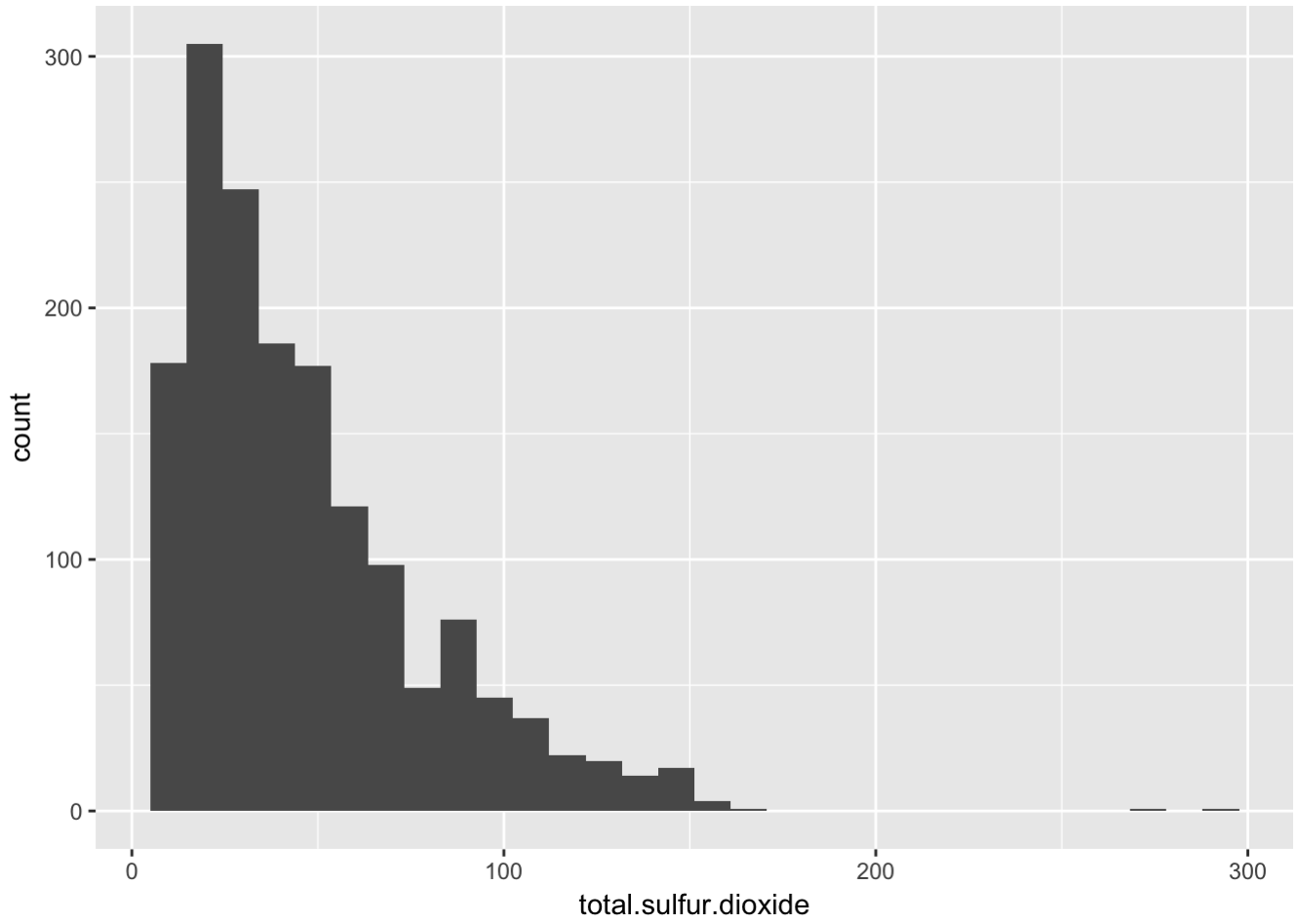
氯化物的分布基本都在0.1以下，没有趋势性的变化，所以在研究变量变化的时候可以忽略这个变量





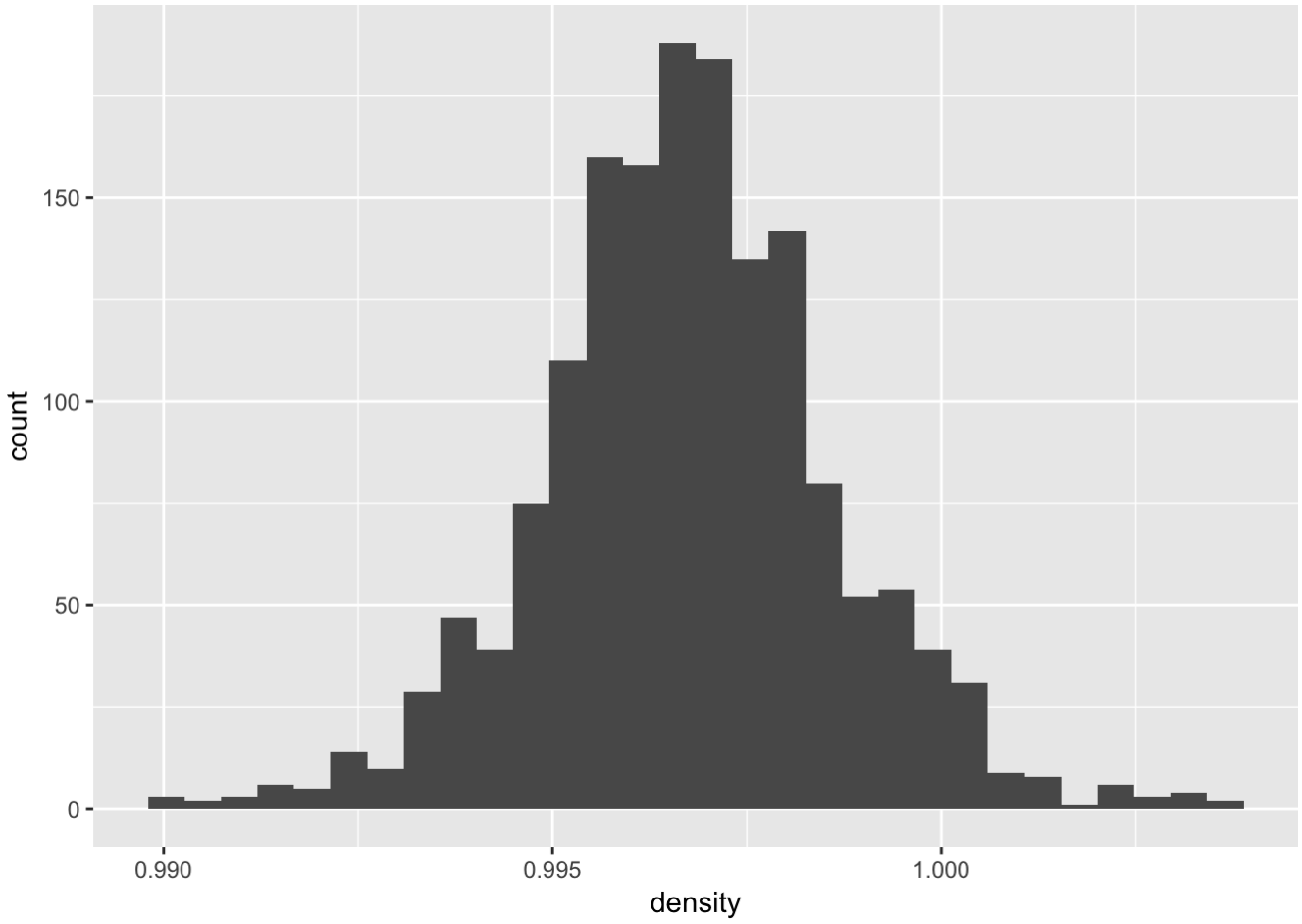

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

游离二氧化硫的分布大致分为1到20这个区间数据量众多数值稳定，20到40这个区间数值有递减趋势，40以上的长尾分布较少



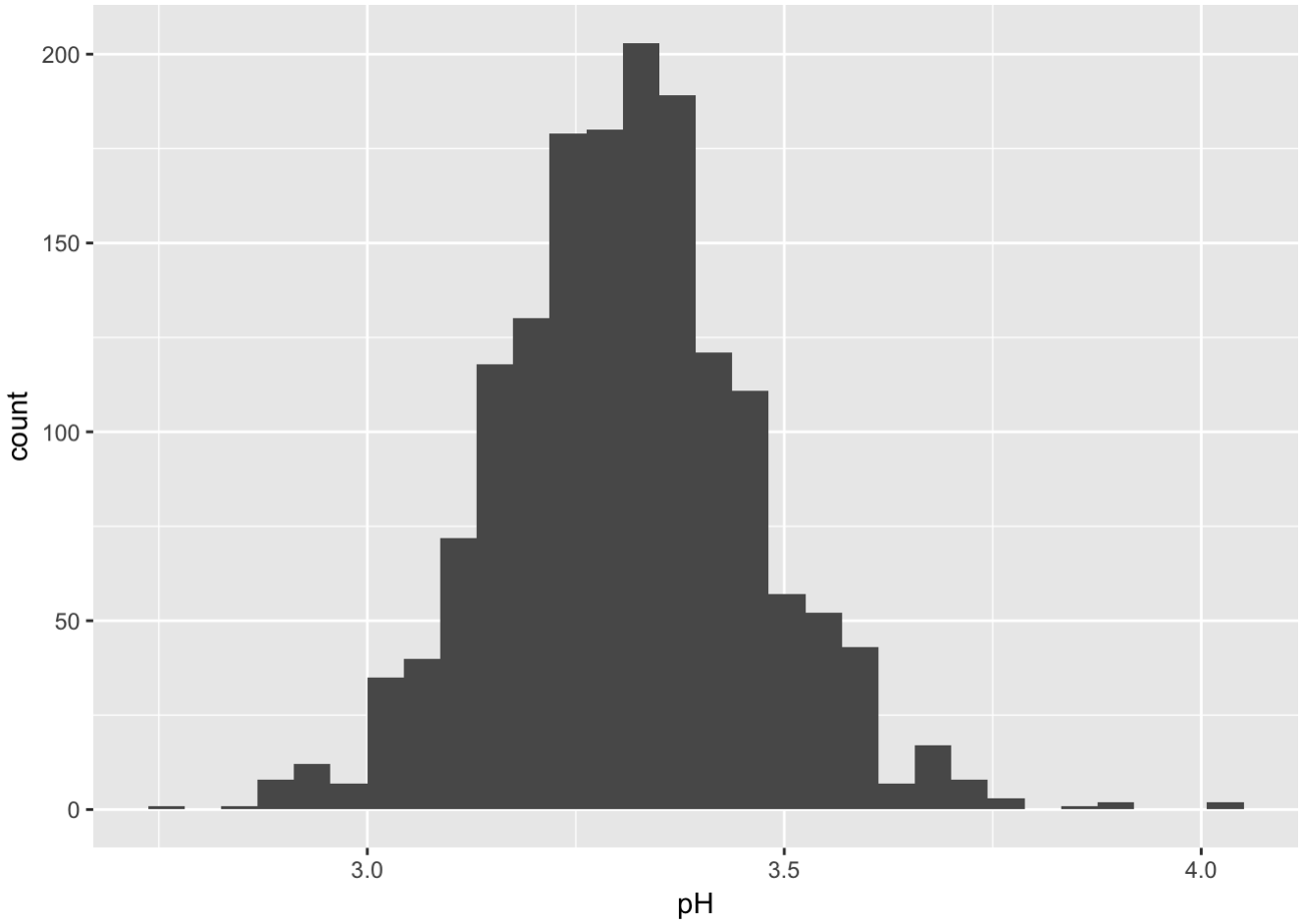

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

总二氧化硫在150以下的含量时候呈现正偏态分布



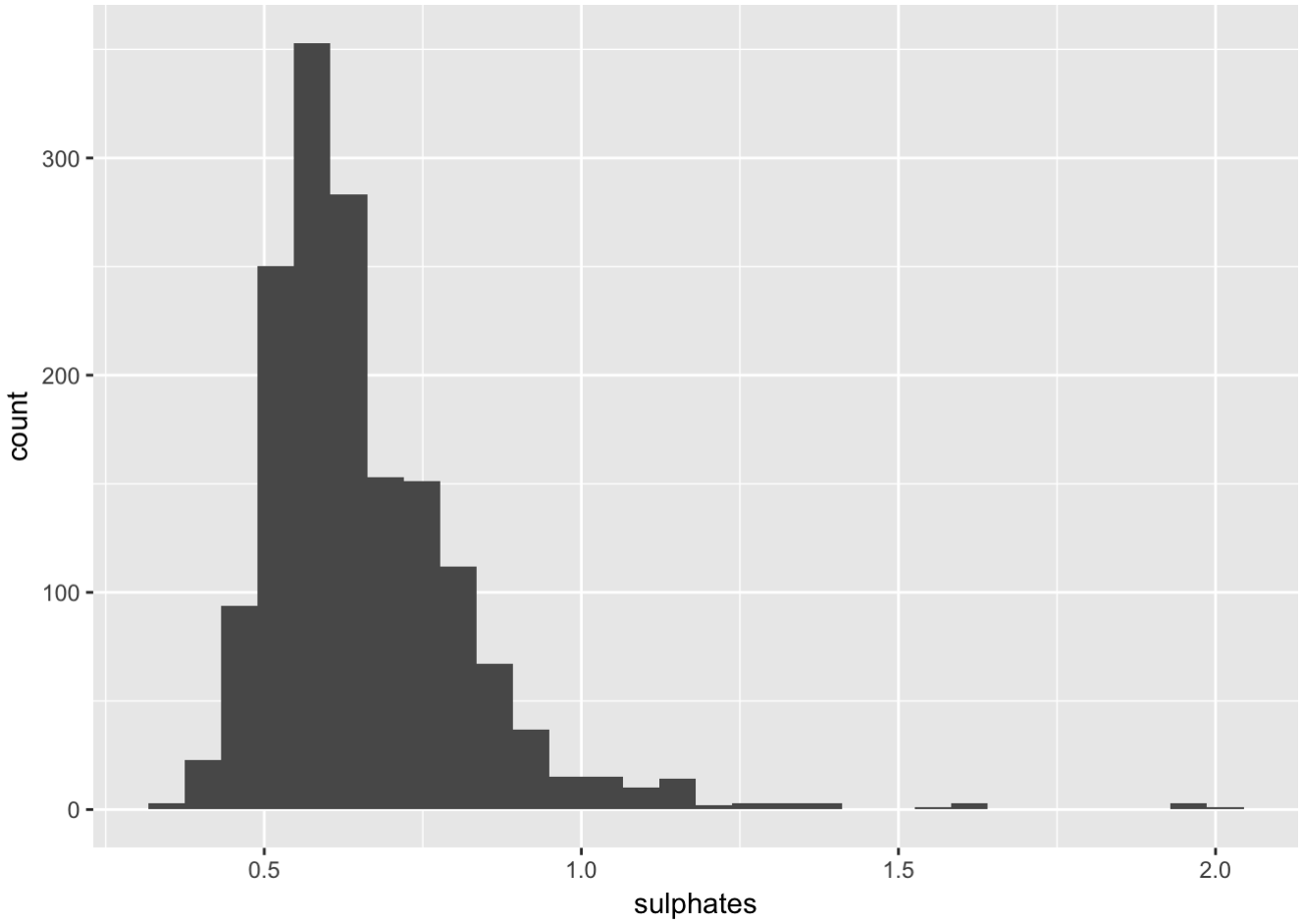

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

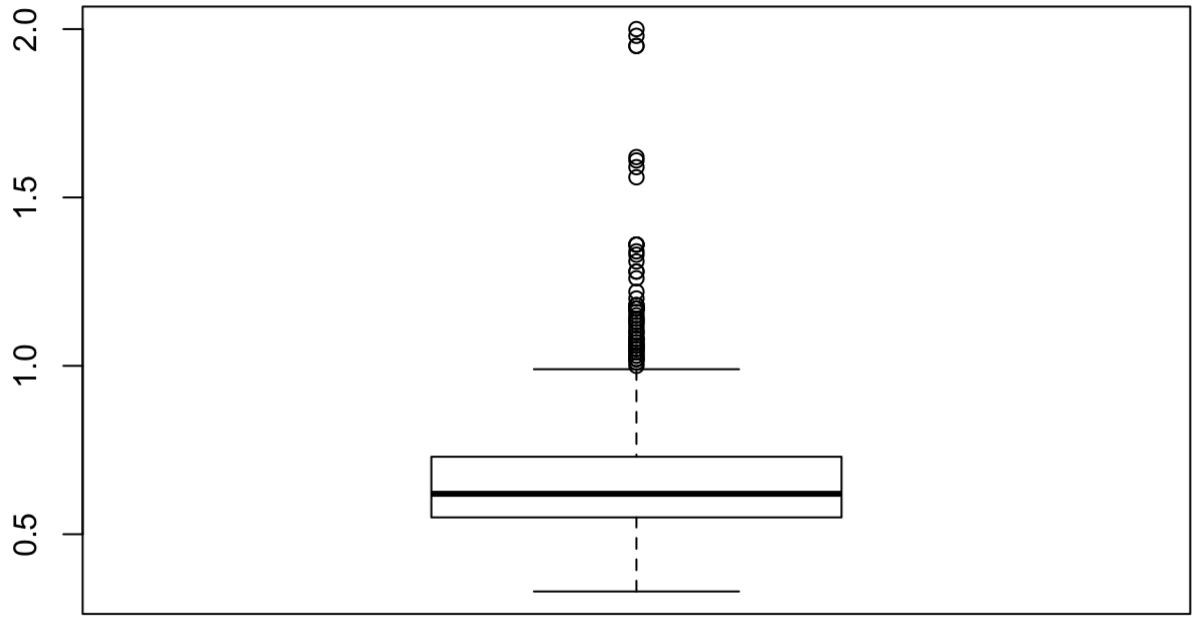
密度这个变量是正态分布，数据间距很小




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

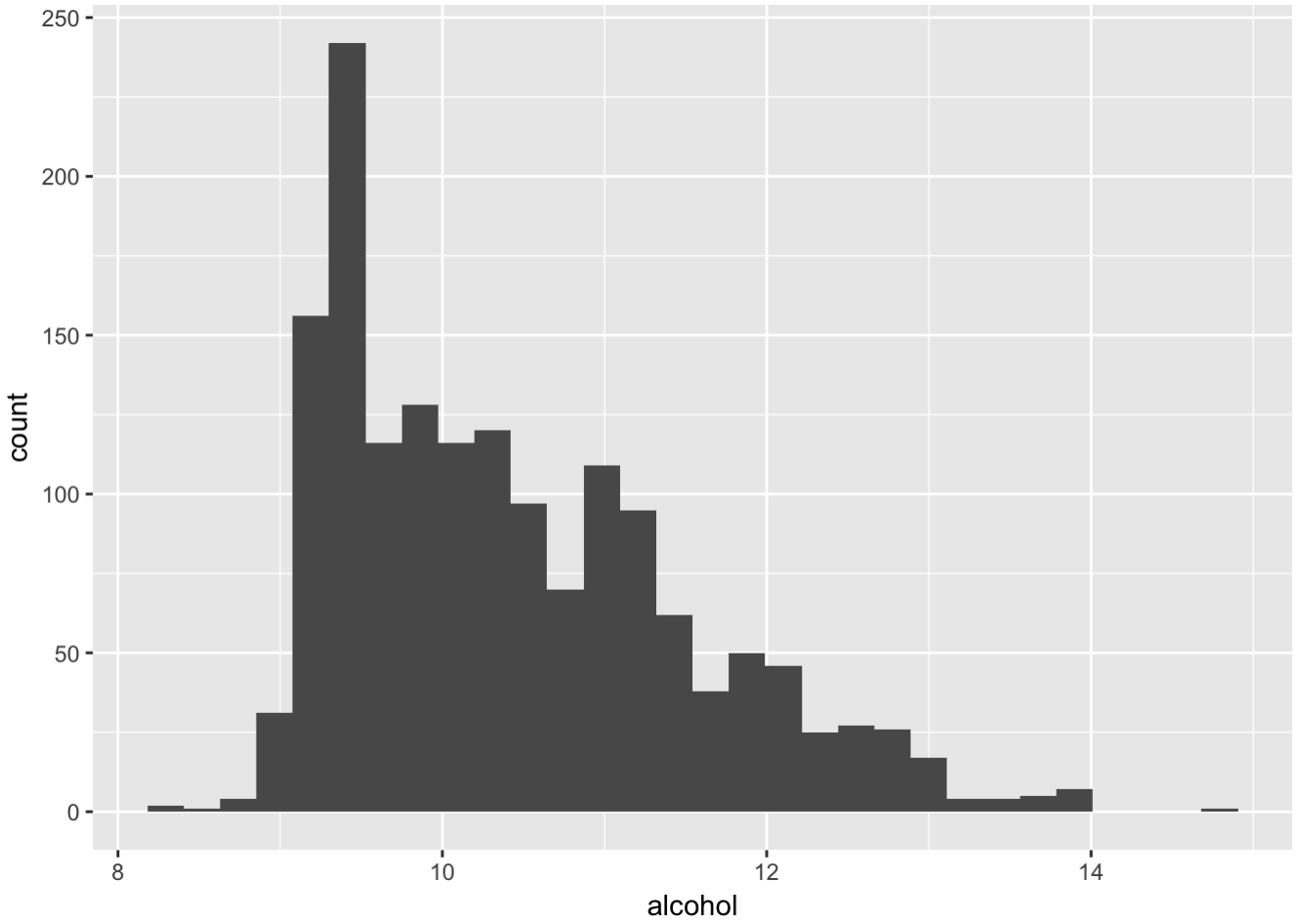
pH值也呈现出正态分布，但是两端的都有一定量的长尾数据

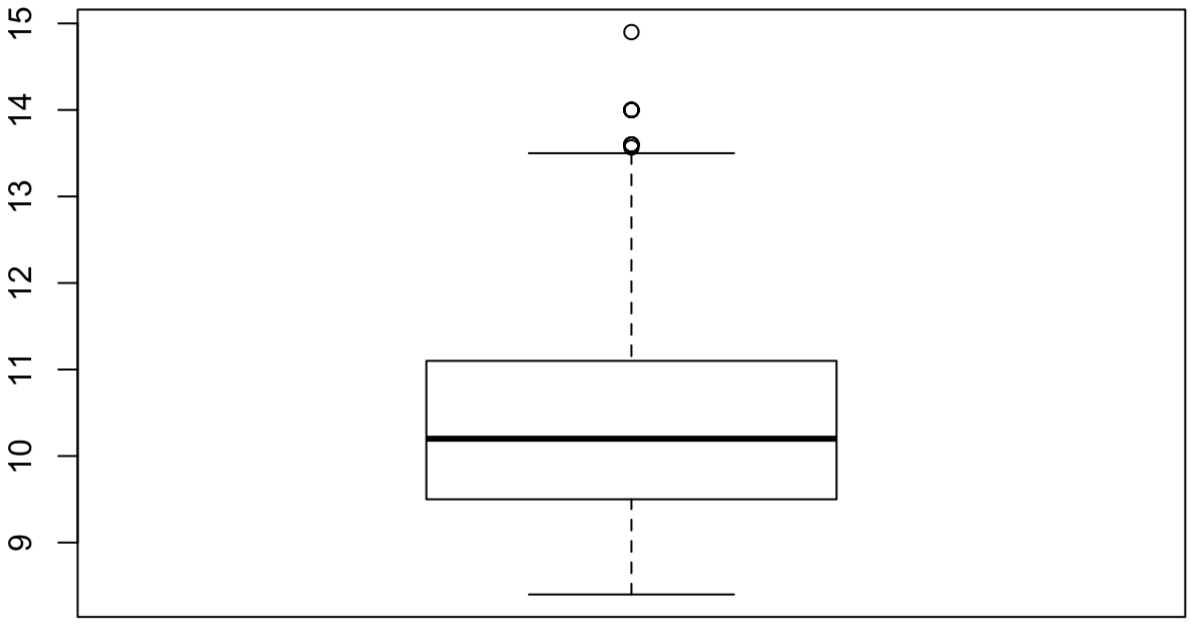





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

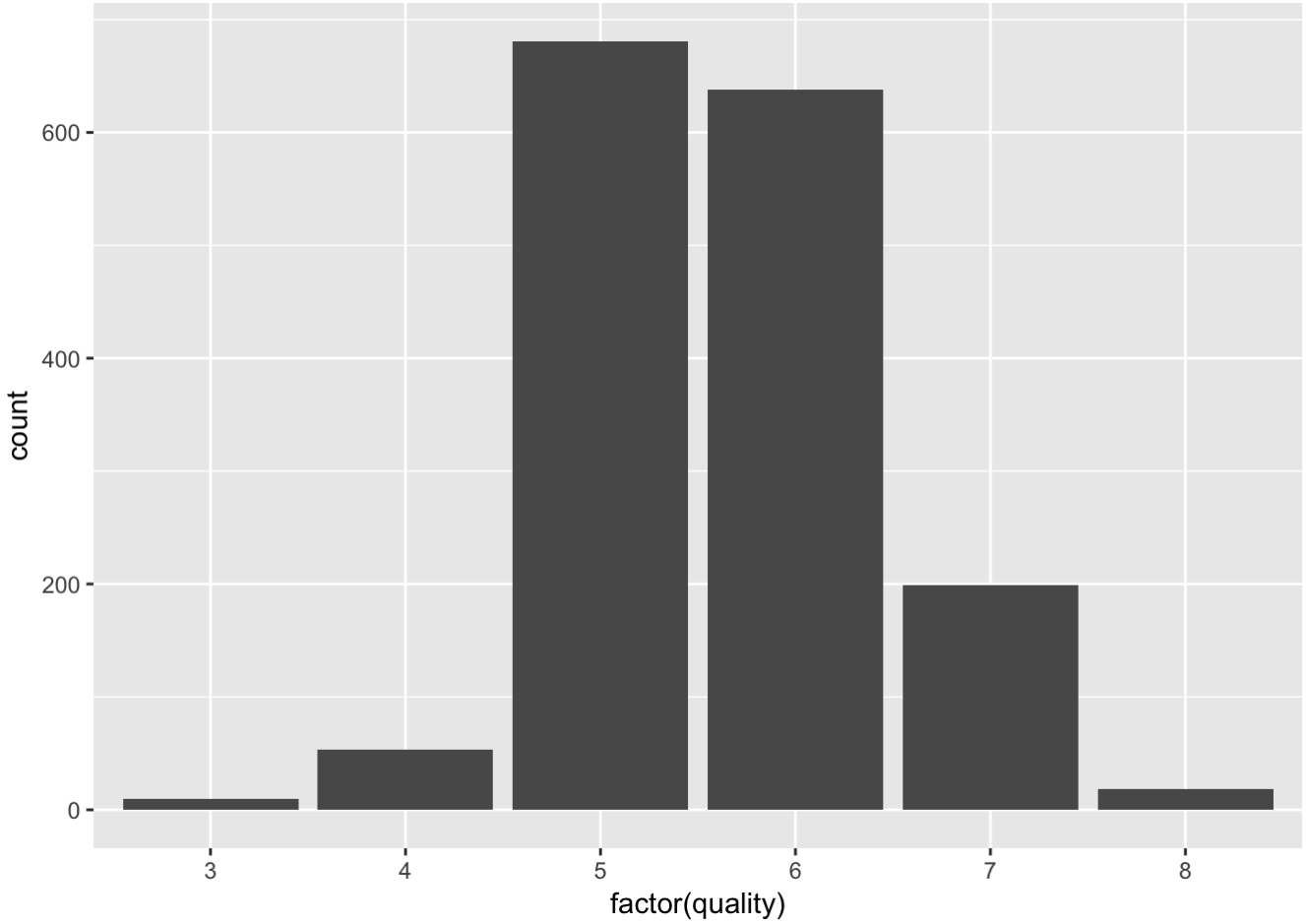
硫酸盐含量是正偏态分布，但是三分位以上有部分异常值






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

酒精度含量表现出规律的趋势分布，预计会与质量评级有强关联



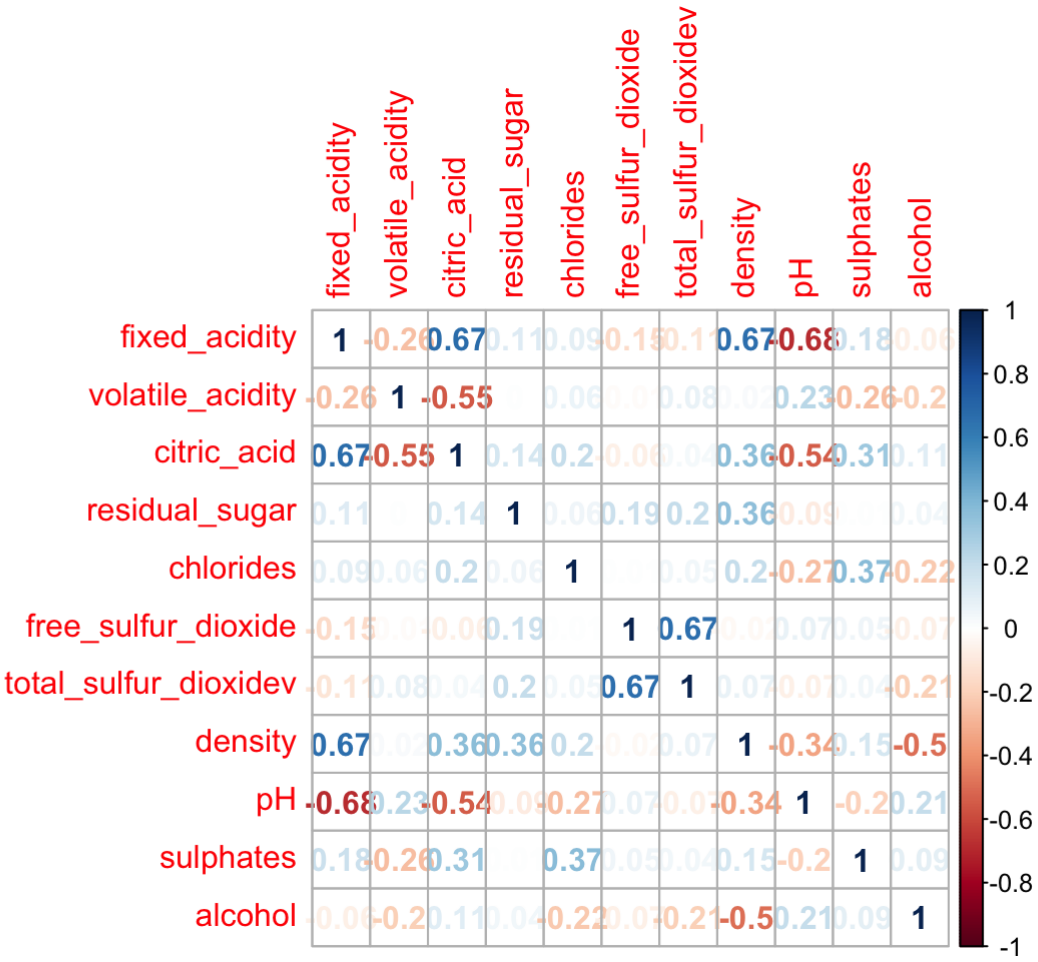
定序变量用柱状图且赋予其因子factor特性，质量评级以6为中心呈现正态分布

部分变量名称不直观和不方便，修改如下

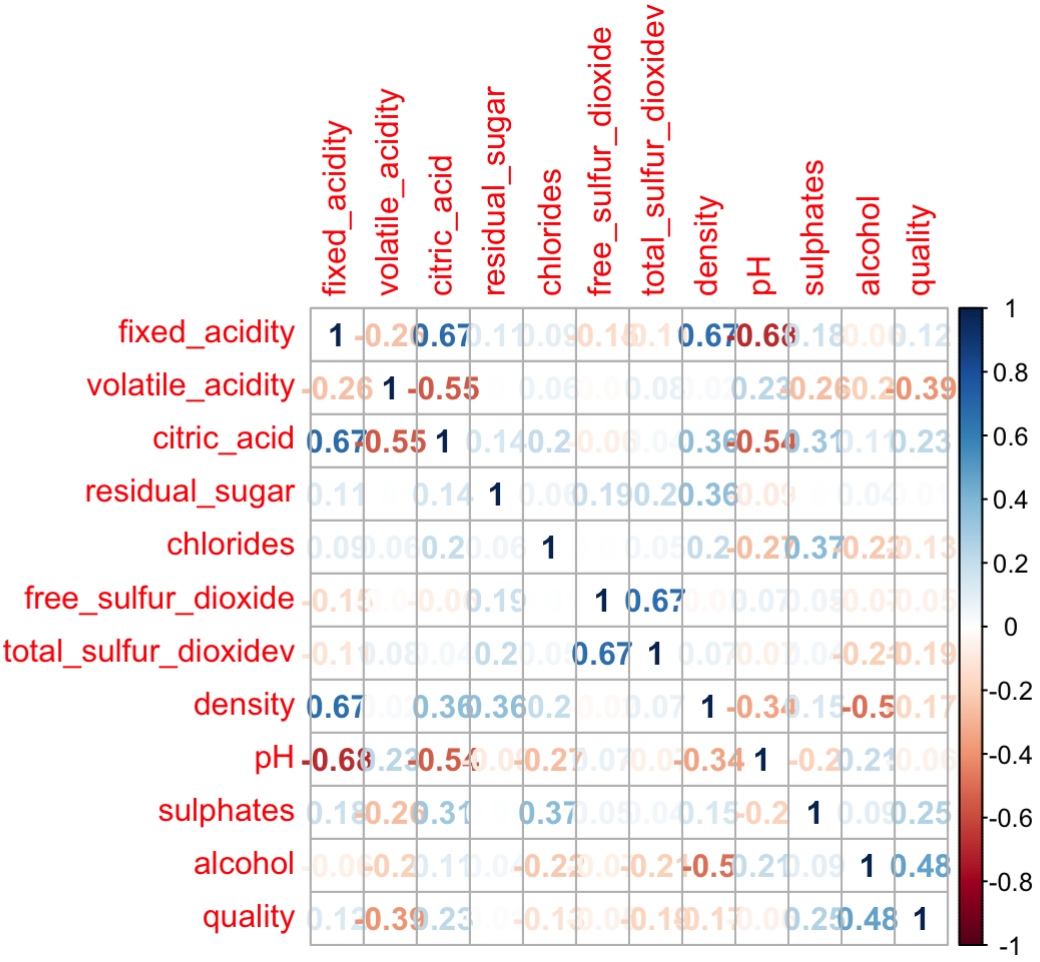
```
##  [1] "ID"                   "fixed_acidity"
##  [3] "volatile_acidity"     "citric_acid"
##  [5] "residual_sugar"       "chlorides"
##  [7] "free_sulfur_dioxide"  "total_sulfur_dioxidev"
##  [9] "density"              "pH"
## [11] "sulphates"            "alcohol"
## [13] "quality"
```

在进行两个以上变量分析之前首先宏观的观察下其相关性

皮尔逊相关系数矩阵（不含定序变量 quality）

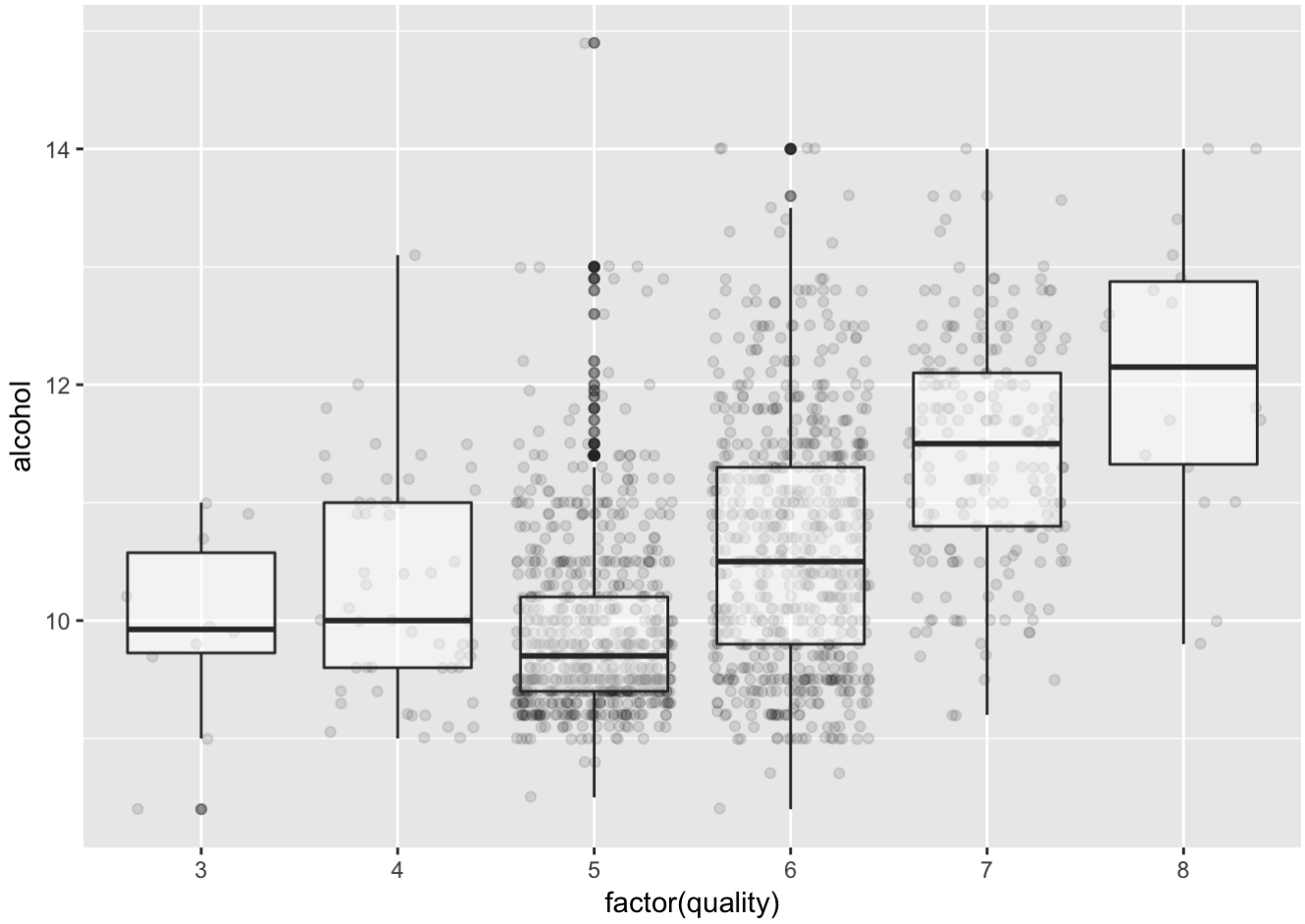


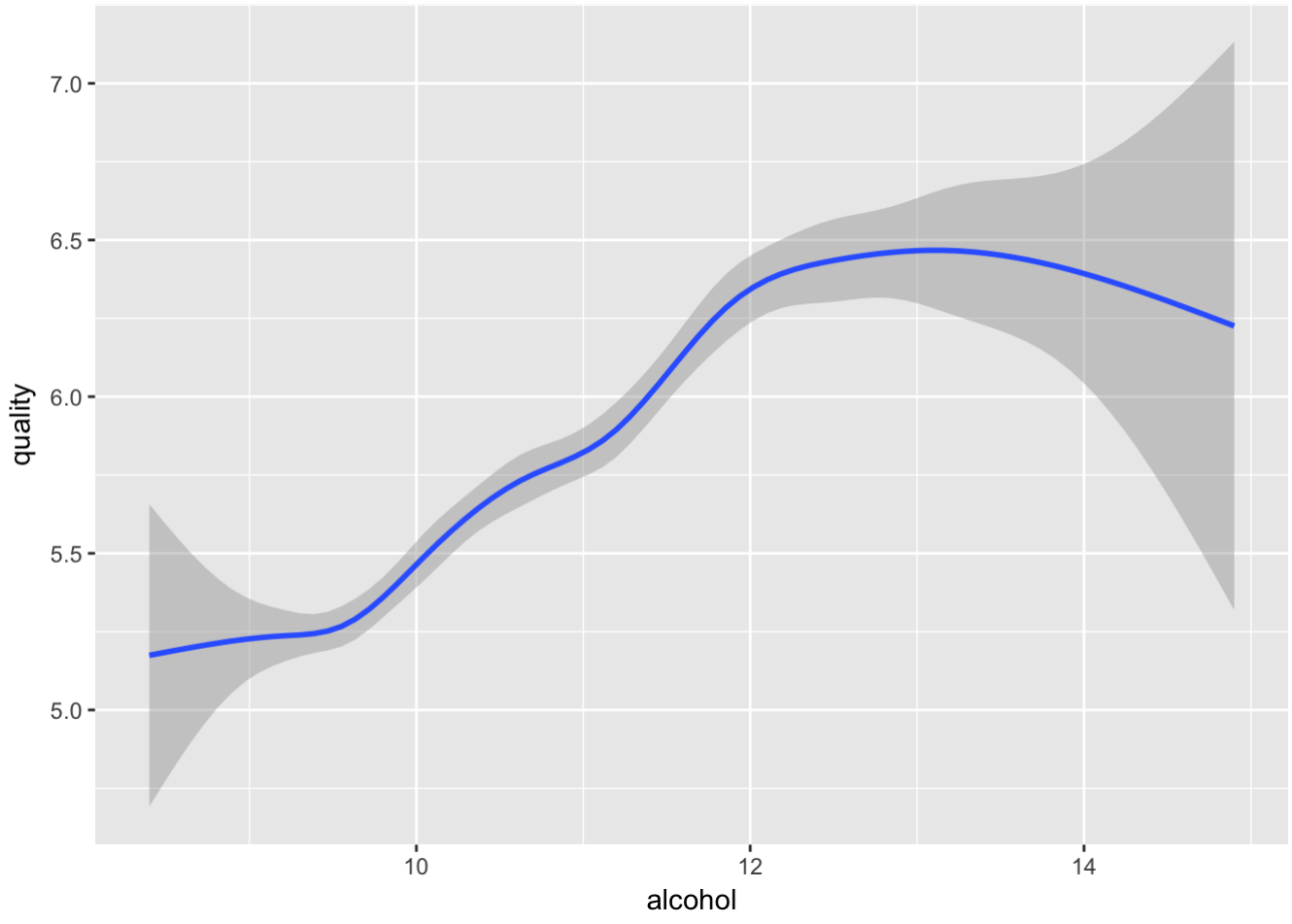
斯皮尔曼相关系数矩阵（含定序变量 quality）



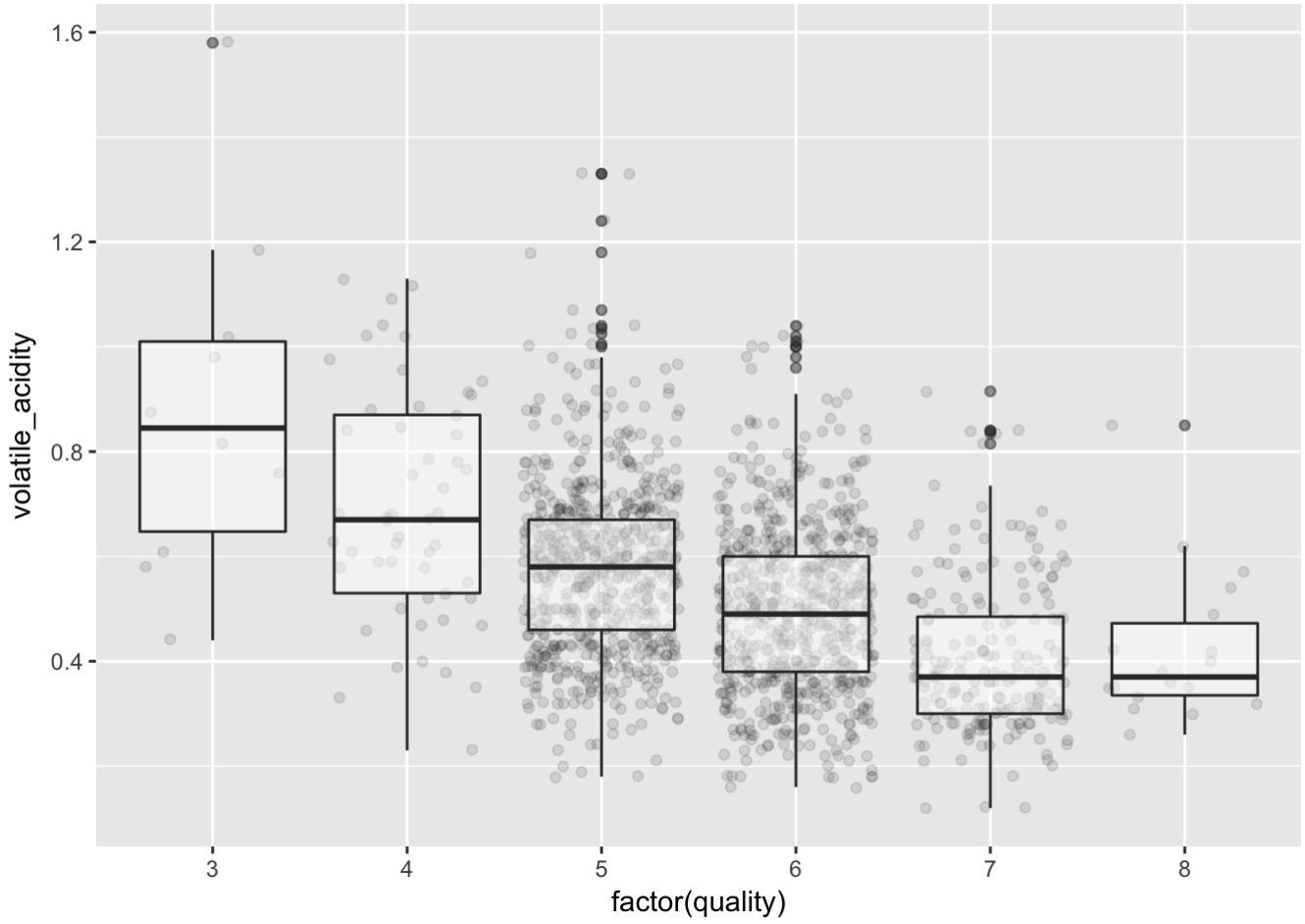
> 基于上述的图示在分析多变量关系的时候就有一定的参考，具体执行过程如下

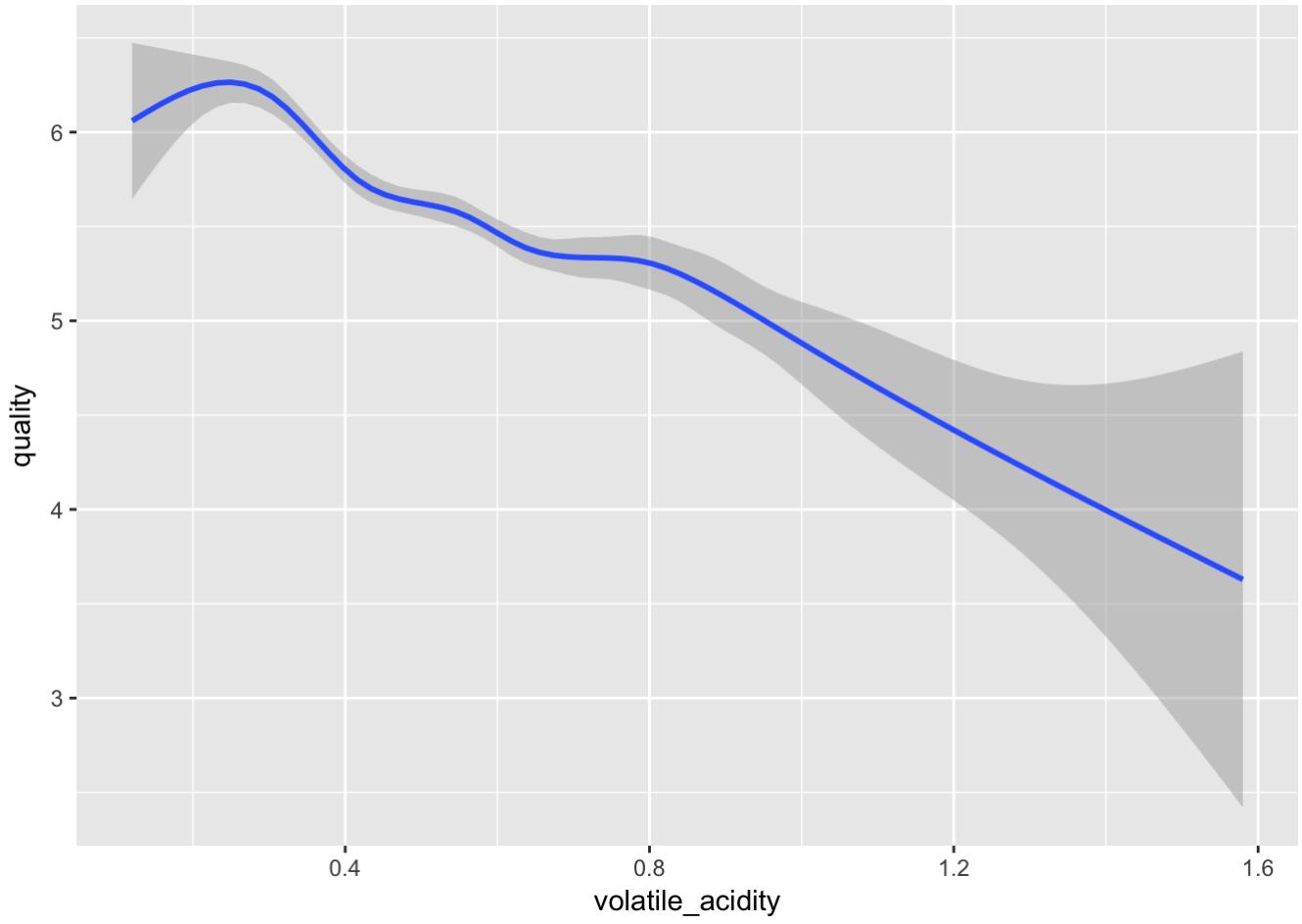
双变量绘图选择与分析



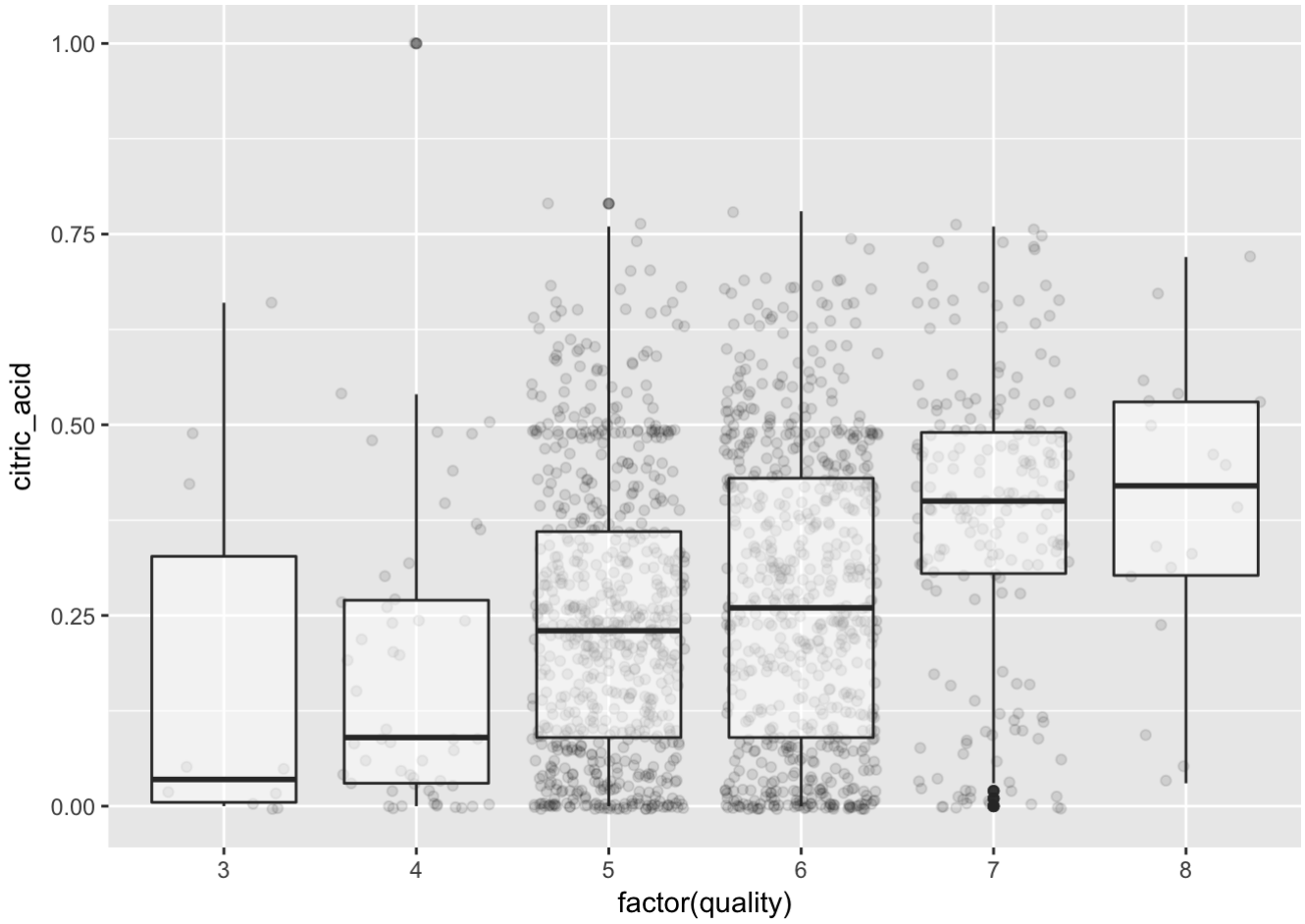


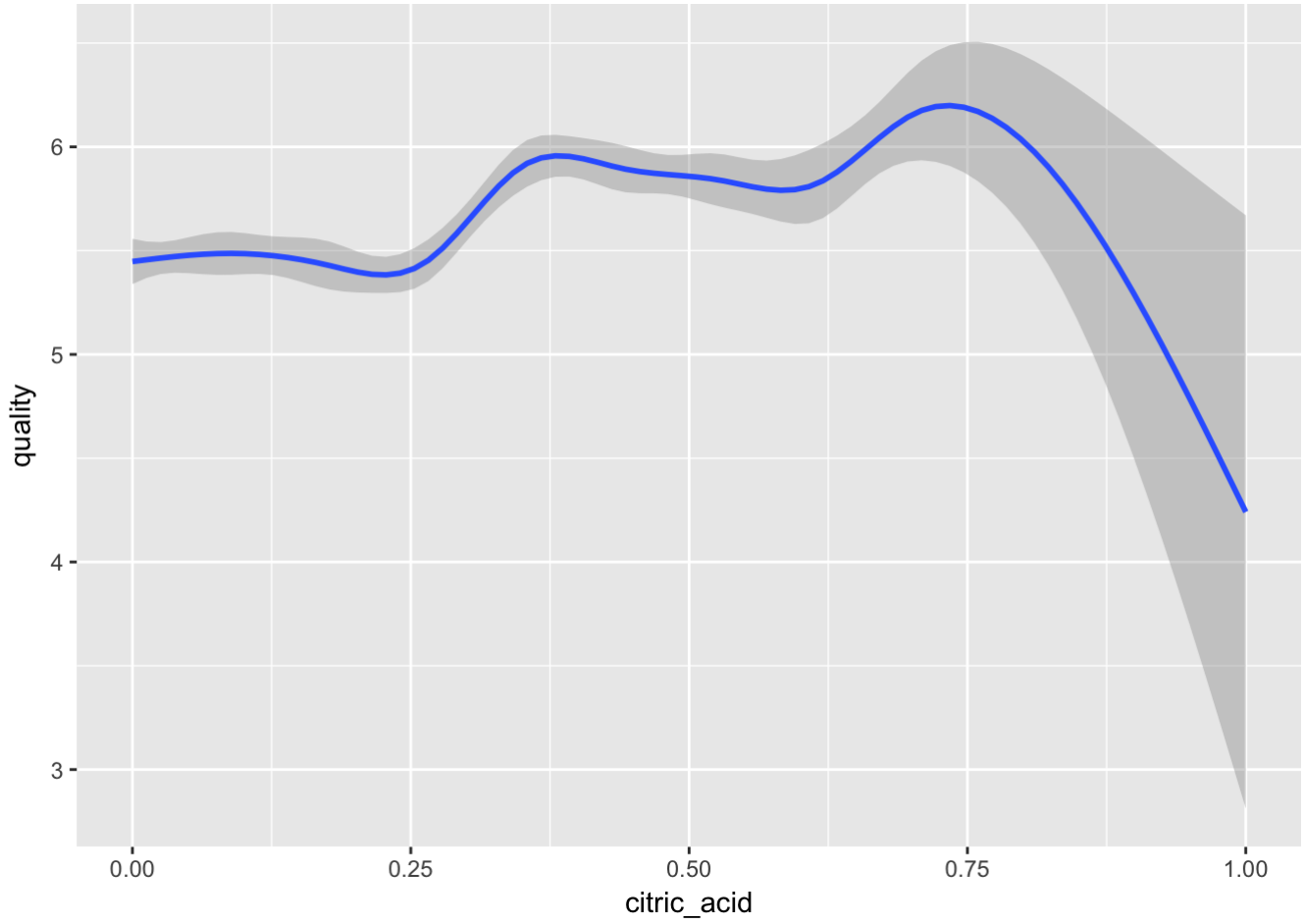
质量评级从5到8的酒精度有递增的趋势，而且从整体来看确实高度的酒的质量评级比低度数的就高一点


1.6
1.2
0.8
0.4
volatile_acidity
3
4
5
6
7
8
factor(quality)


6
5
4
3
quality
0.4
0.8
1.2
1.6
volatile_acidity

通过这两个图可以明显观察到随着挥发性酸度的降低其红酒的评级逐步提高，根据这个观察推断出，和质量评级相关性最强的变量可能就是挥发性酸度





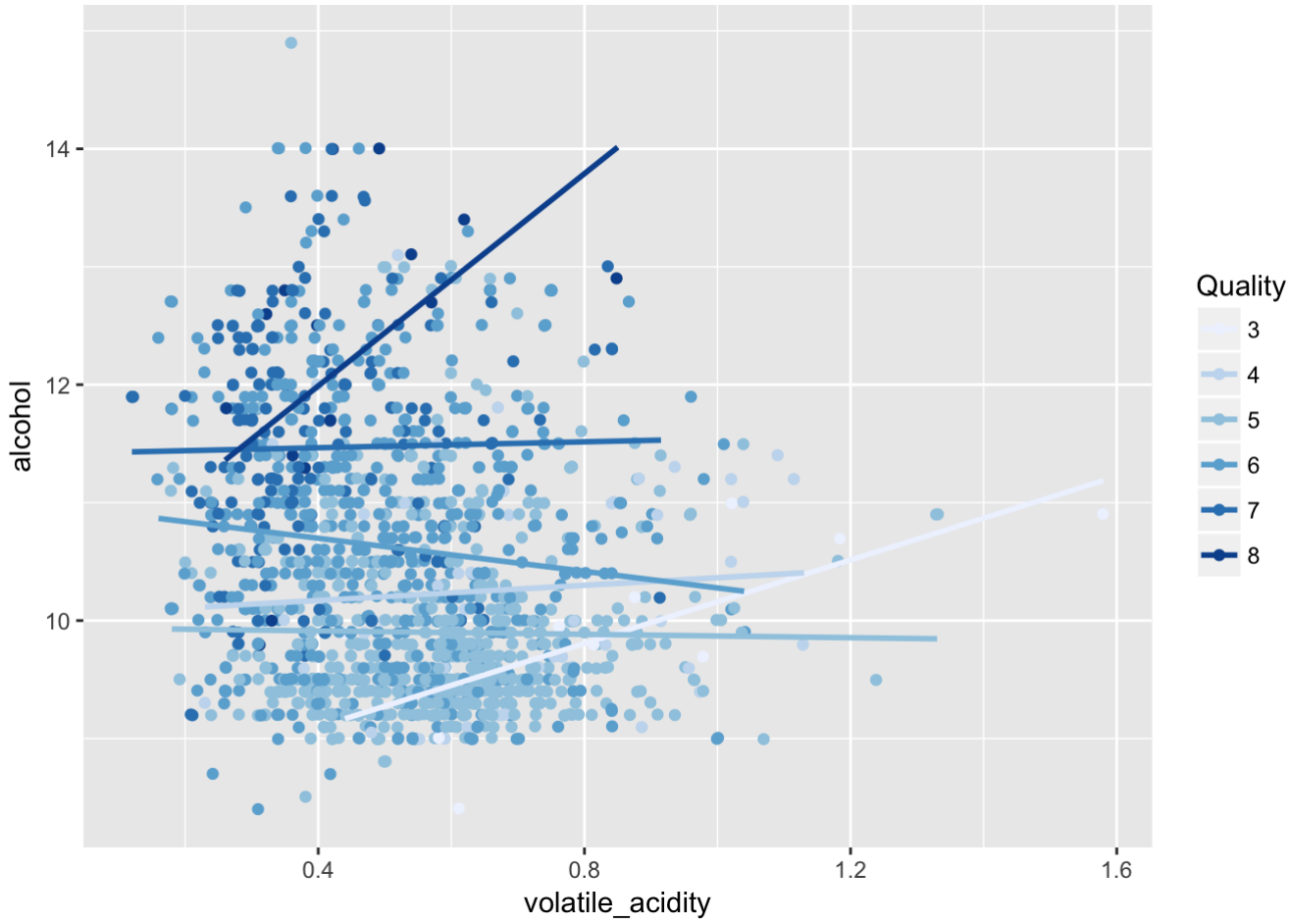
柠檬酸这个变量对质量评级的影响虽然没有挥发性酸度高，但是质量高的红酒的柠檬酸的含量确实比较高，只是各个等级间的差别比较小

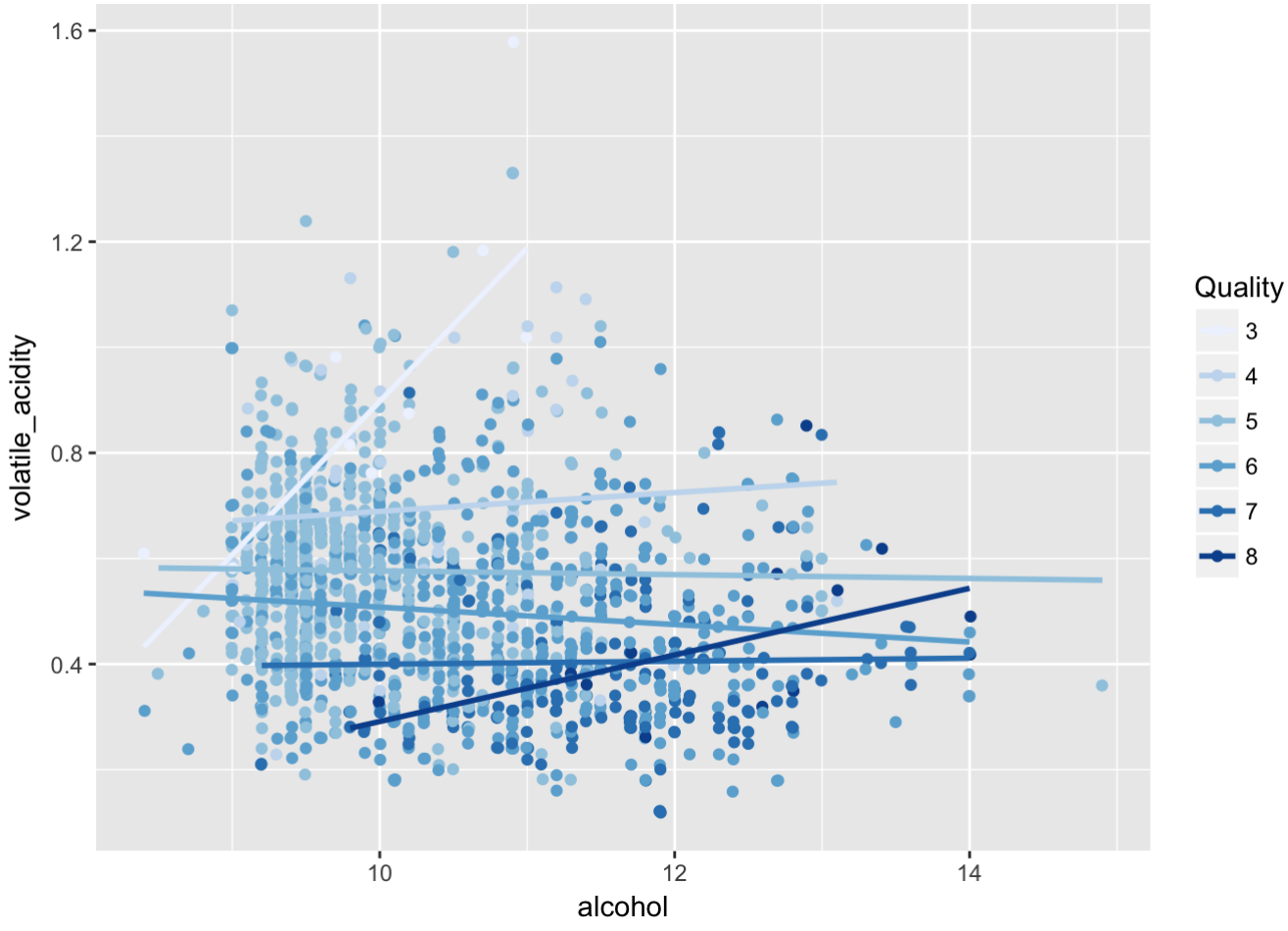
酒精度数和挥发性酸度的相关性系数：

```
## [1] -0.202288
```

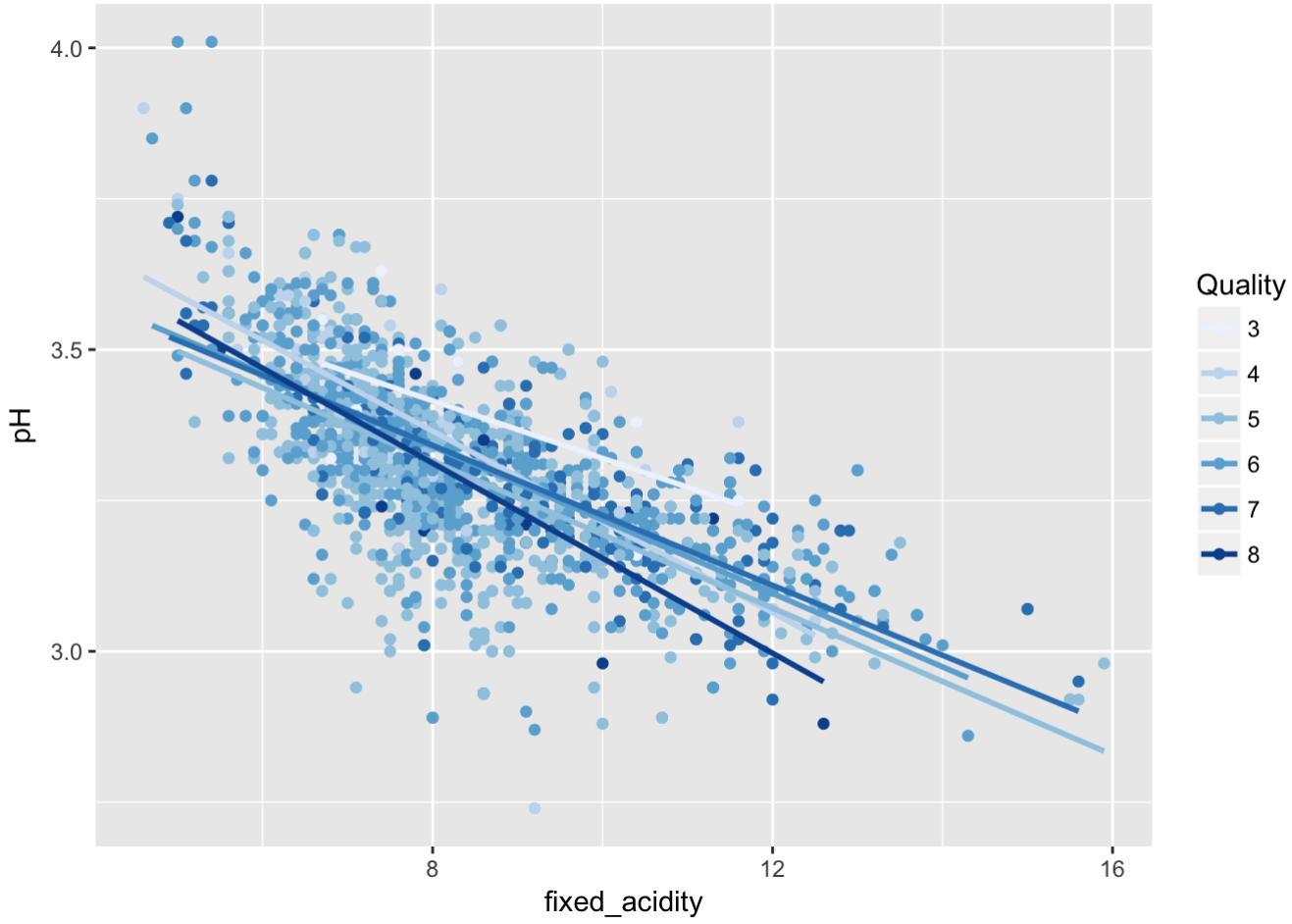
> 因为质量评级和酒精含量、挥发性酸度含量有联系，那么这两个变量之间就应该也有某种联系，以上的数据说明了他们确实有一定的负相关，印证了推测

多变量绘图选择和分析





为了更好的验证之前的分析，使用这一组两个图很好的说明了：酒精度数越高、挥发性酸度含量越低其红酒的质量评级越好



固定酸度和pH值的相关性系数：

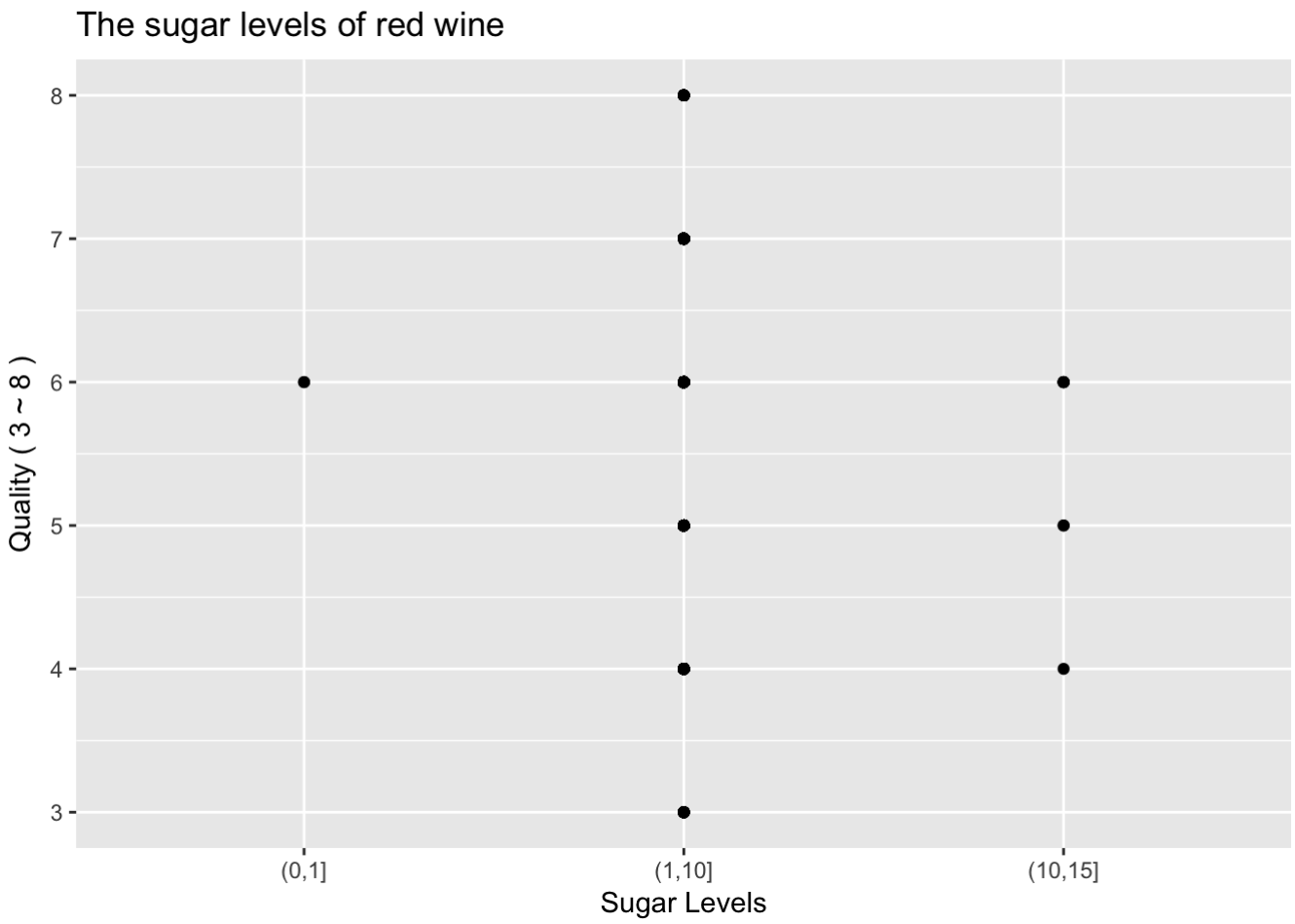
```
## [1] -0.6829782
```

> 虽然红酒的质量评级和固定酸度、pH值没有很直接的关联度，但是固定酸度和pH值却有强烈的负相关

定稿图与总结

绘图一

- 参考 http://winefolly.com/review/sugar-in-wine-chart/ (http://winefolly.com/review/sugar-in-wine-chart/) 对残糖的分级
- Bone Dry < 1 (g / dm^3)
- Dry 1 - 10 (g / dm^3)
- Off-Dry 10 – 35 (g / dm^3)
- 结合我们的数据对残糖分级后的分布图如下



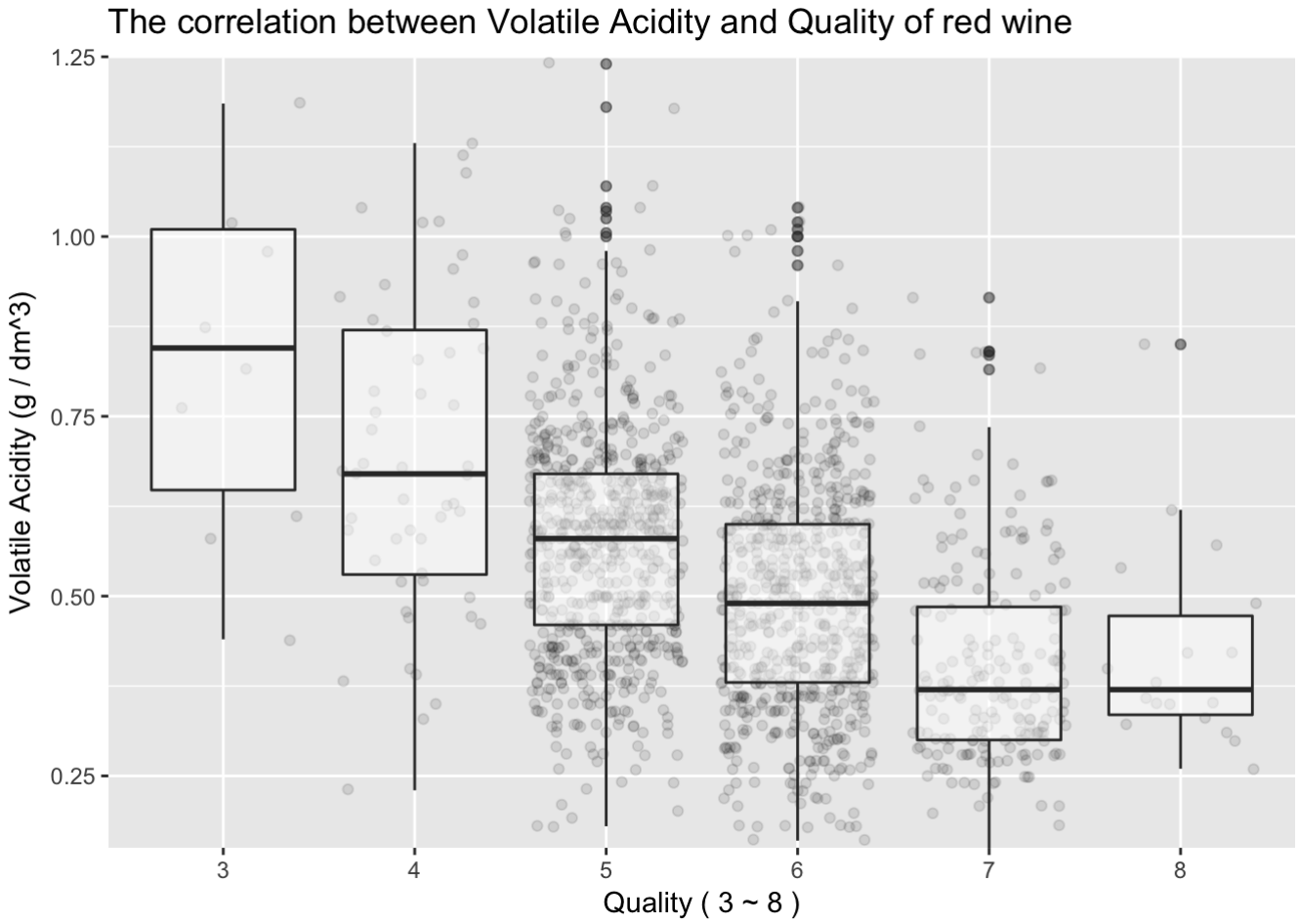
- 这些分级的统计数值如下

```
##
##   (0,1]  (1,10] (10,15]
##       2    1586       8
```

描述一

- 结合以上图示和数值，可以判断我们观察的红酒绝大部分都是“Dry”级别的，如果我的抽样就是主要针对干红葡萄酒的话，那么说明大部分的干红葡萄酒的残糖介于1到10克每升

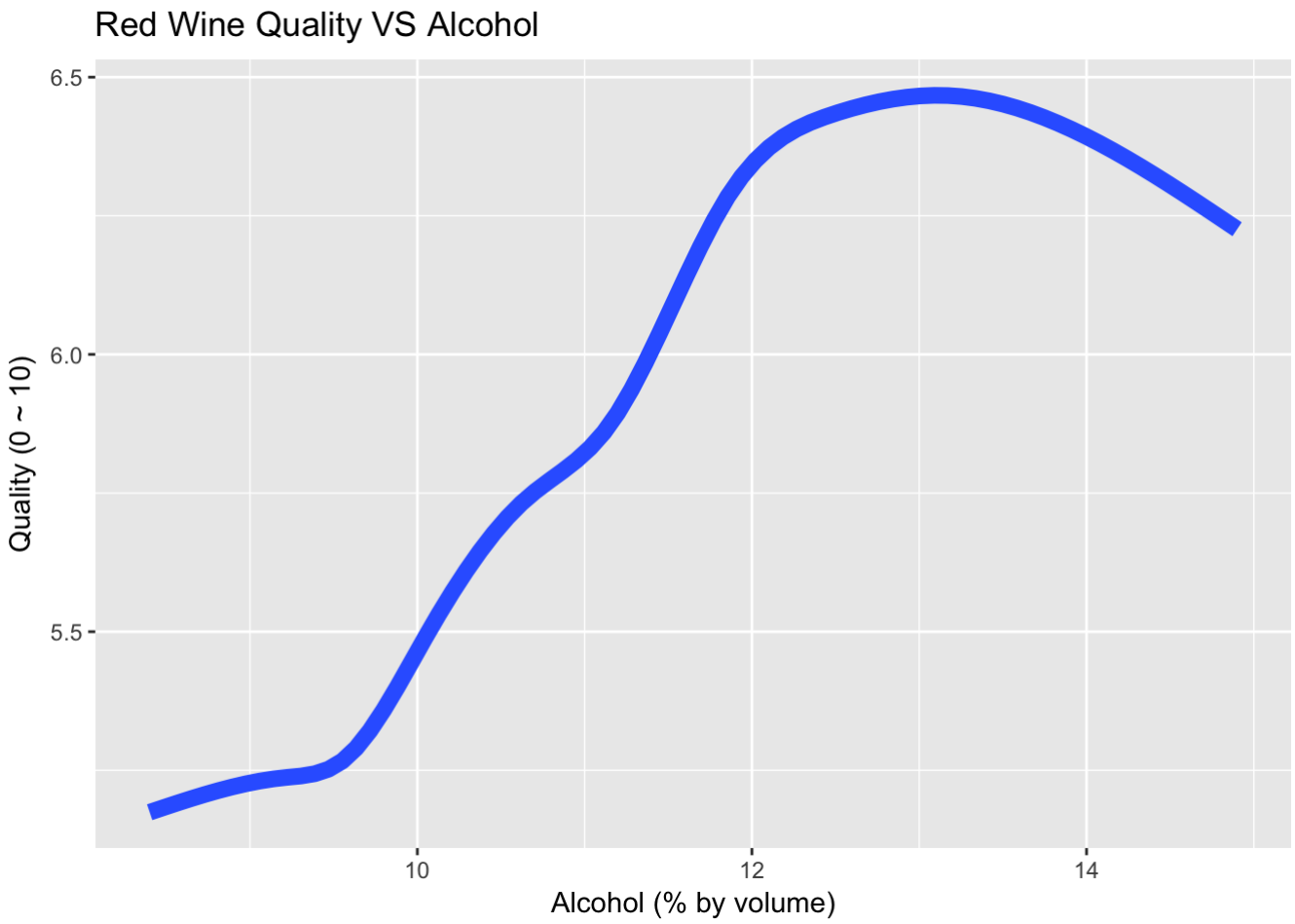
绘图二



描述二

- 通过修饰后的此图很清晰的展现了挥发性酸度越低其对应红酒的质量评级越高

绘图三



描述三

- 红酒评级随着酒精度数的提高而增加（13度以内）

反思

- 在进行两个以上变量分析之前首先宏观的观察下其相关性
- 对于部分变量可以尝试分组的方式分析其相关性
- 在分析过程中一度忽视了质量等级是定向变量这一属性，使得对应的图示不直观后来修改了这部分
- 在分析相关性的时候，可以使用 cor() 方便加以验证
- 分析两个变量的关系的时候适时的使用箱型图可以很直观的展现出对应的关系
- 对坐标系的调整也可以增进图示的效果
- 红酒评级随着酒精度数的提高而增加（13度以内）
- 红酒中挥发性酸度（Volatile Acidity）的含量越少其质量评级越高
- 但是数据中的噪音太多，可以从提高数据的量试着解决，另外可以尝试回归模型