红葡萄酒质量探索分析

Chen Pan

2019/1/5

==========

1. 加载并初步观察数据

1.1 观察数据结构

```
## '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 0.073 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 ...
```

1.2 观察数据集描述性统计

```
##        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
```

通过对数据集的初步观察发现：该数据集一共有13个变量，分别是：X, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlrides, free.sulfur.dioxide, total.sulfur.doxide, density, pH, sulphates, alcohl, quality。其中变量X和quality的数据类型是int型，其他均为num型，进一步观察发现变量X为观测值的编号，在进行EDA时可以忽略。同时，该数据集一共有1599个观测值。

2. Univariate Plots Section 单变量分析

通过查阅网上资料，葡萄酒的主要质量指标分为感官指标和理化指标两大类，其中感官指标主要指色泽、香气、滋味和典型性方面的要求，理化指标主要指酒精度、酸度和糖分指标，因此优先对这三类指标进行分析。

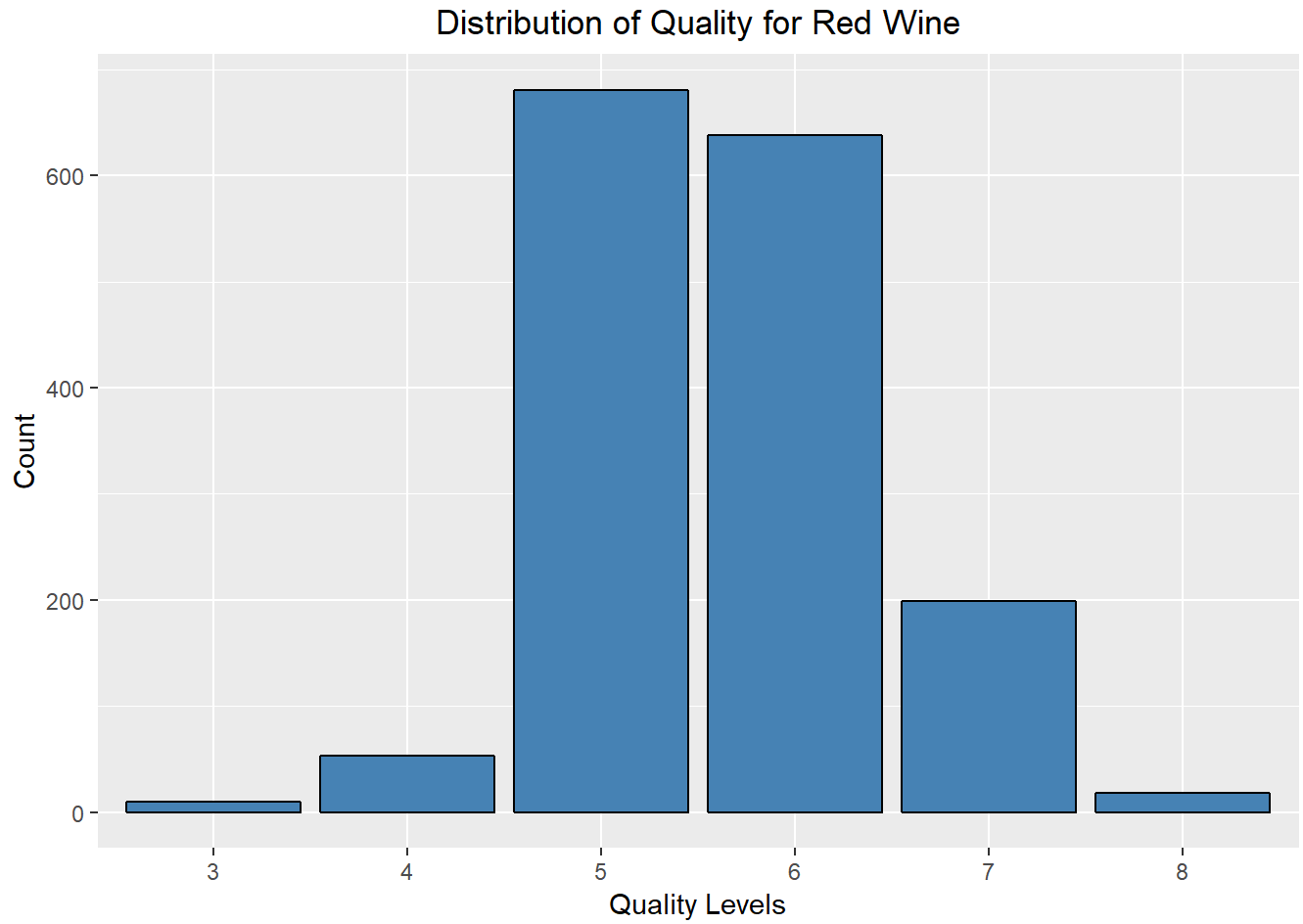
绘图说明：定量型数据一般使用直方图，分类型数据一般使用柱状图，在本数据集中，变量质量属于定序变量，适合使用柱状图来绘制，其他变量属于定量型数据，适合使用直方图绘制。

2.1 Quality 质量

2.1.1 查看质量的描述性统计信息

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

2.1.2 绘制质量的柱状图



通过观察可以发现，质量的分布基本上是一个正态分布，最小值为3，最大值为8，平均值为5.6，中位数为6。

2.1.3 查看整个数据集是否存在缺失值

```
## [1] 0
```

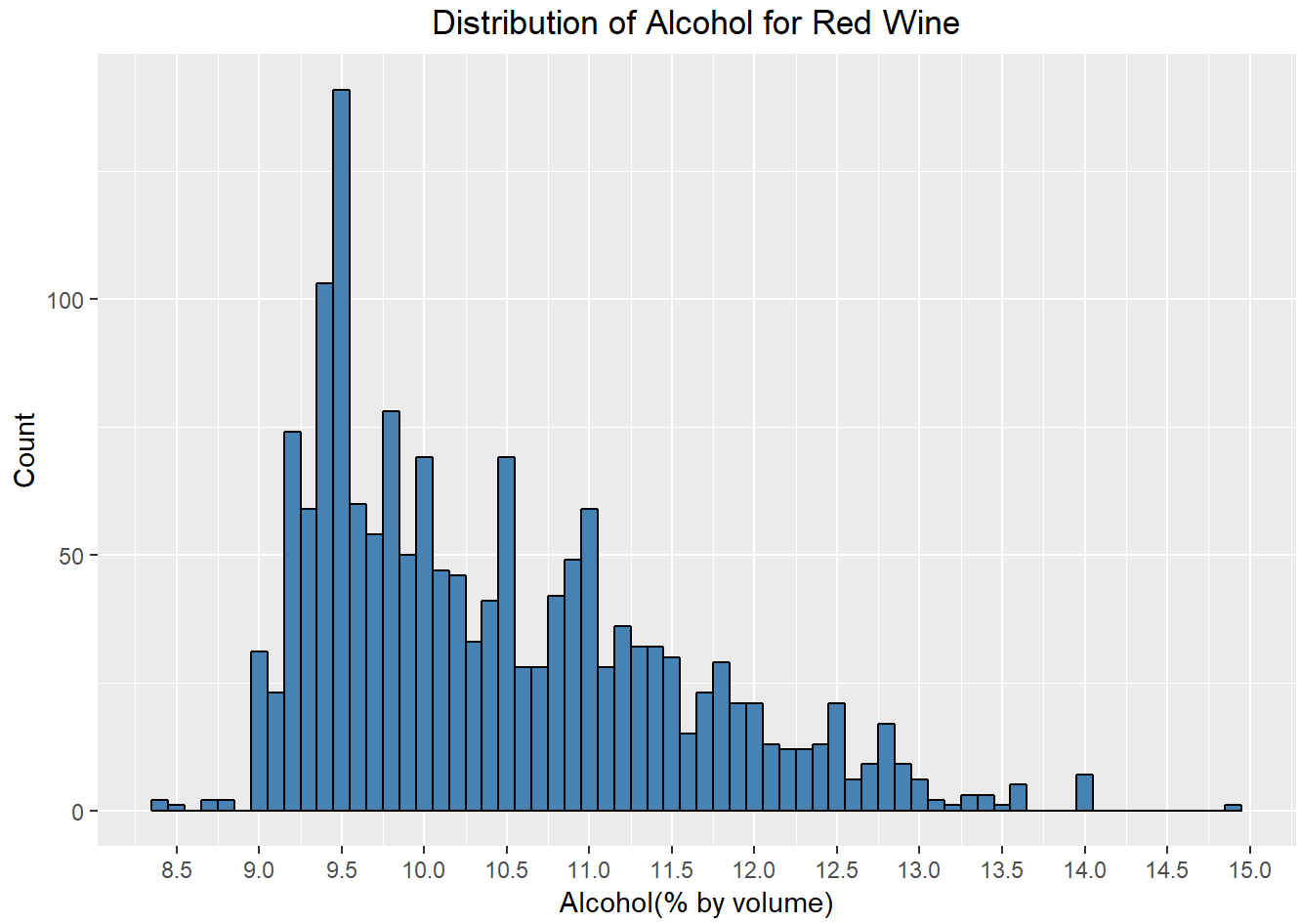
结果显示整个数据集不存在缺失值，因此在后续的分析中暂时不用再考虑NA值的影响。

2.2 Alcohol 酒精度

2.2.1 查看酒精度的描述性统计信息

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

2.2.2 绘制酒精度的直方图



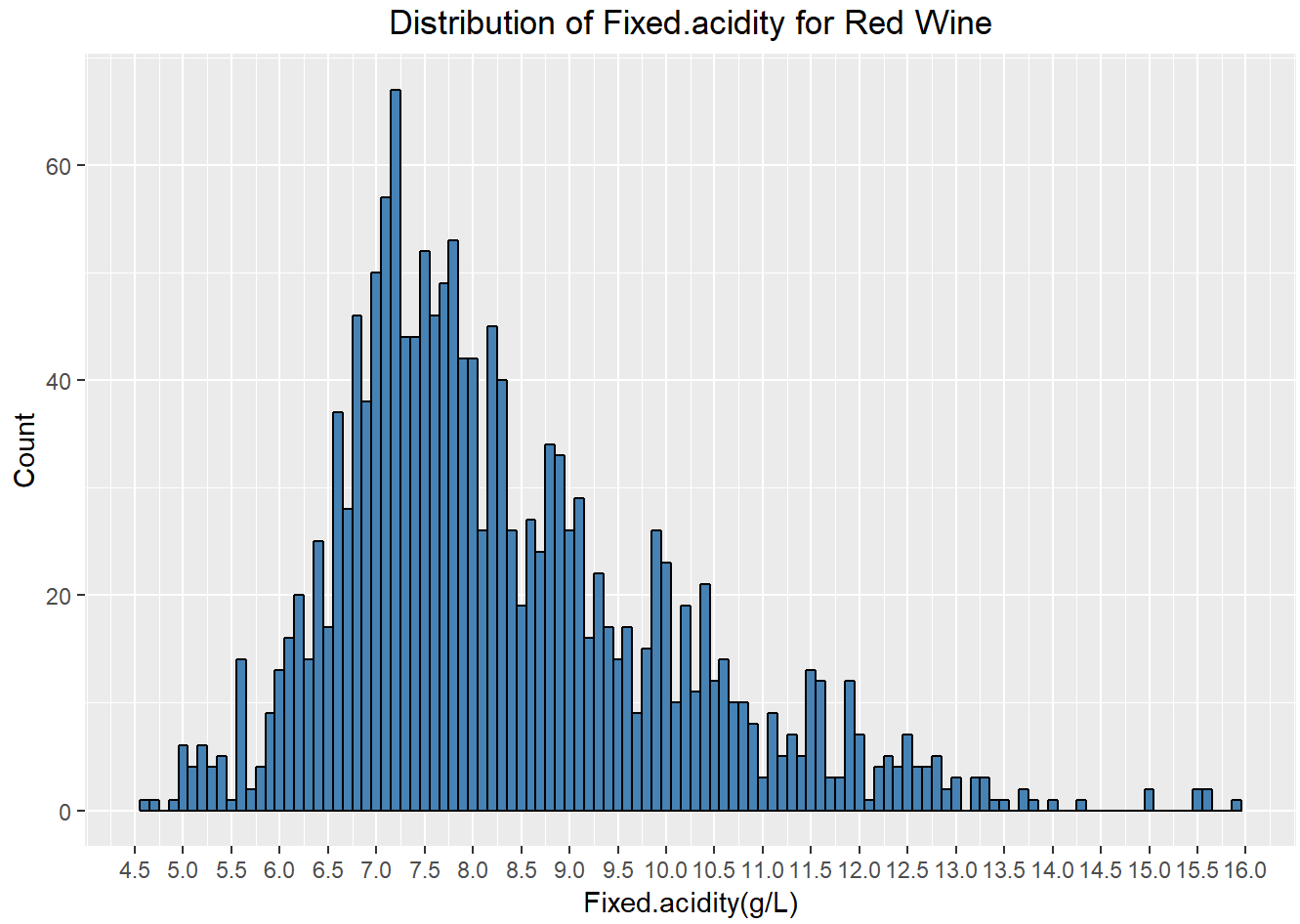
通过观察发现，酒精度主要集中在9.5到11.1之间，平均值为10.42，中位数为10.20。

2.3 Fixed.acidity 非挥发性酸

2.3.1 查看非挥发性酸的描述性统计信息

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

2.3.2 绘制非挥发性酸的直方图



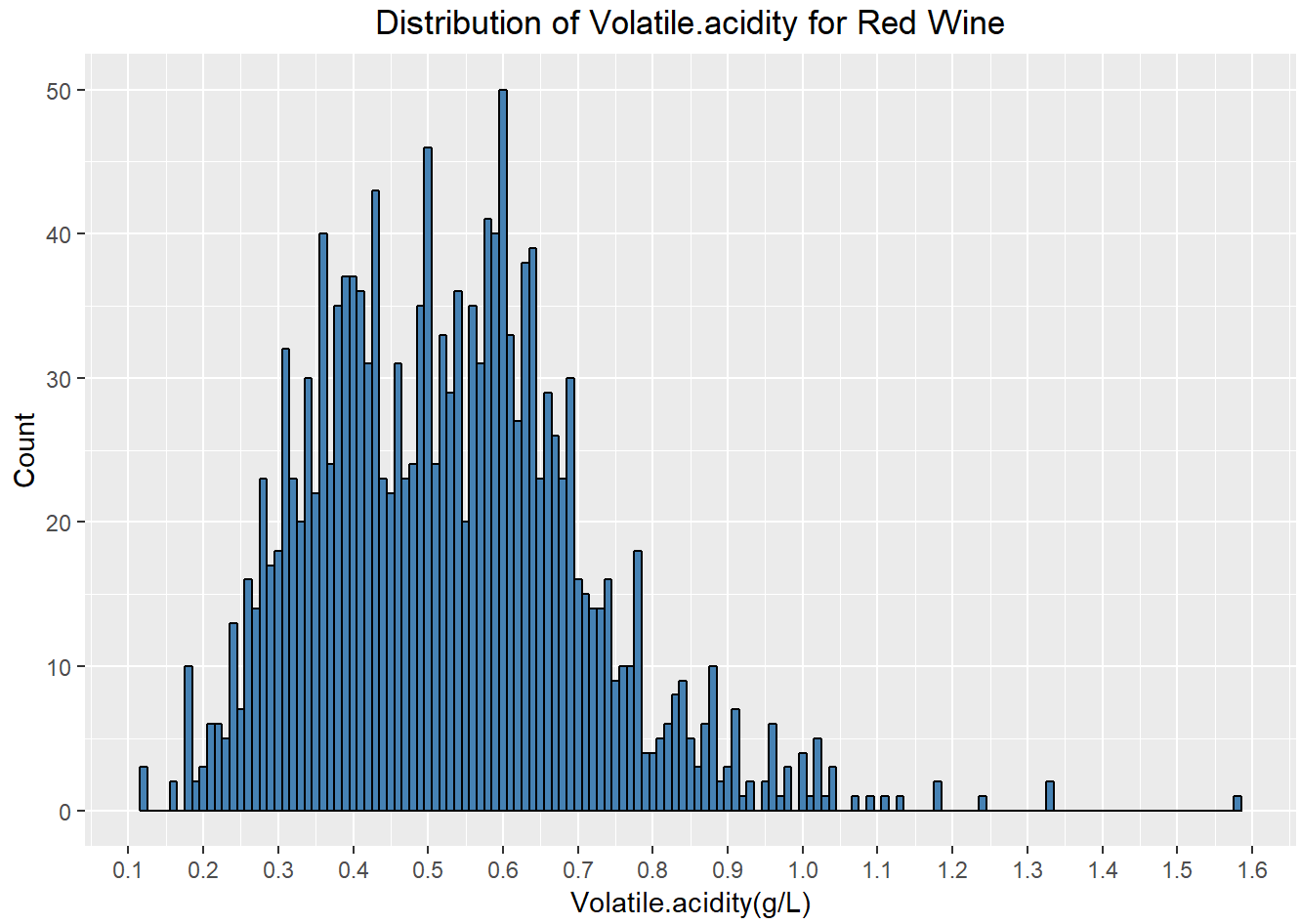
通过观察发现，非挥发性酸的分布呈现近似正态分布，主要集中在7.1到9.2之间，平均值为8.32，中位数为7.9。

2.4 Volatile.acidity 挥发性酸

2.4.1 查看挥发性酸的描述性统计信息

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

2.4.2 绘制挥发性酸的直方图



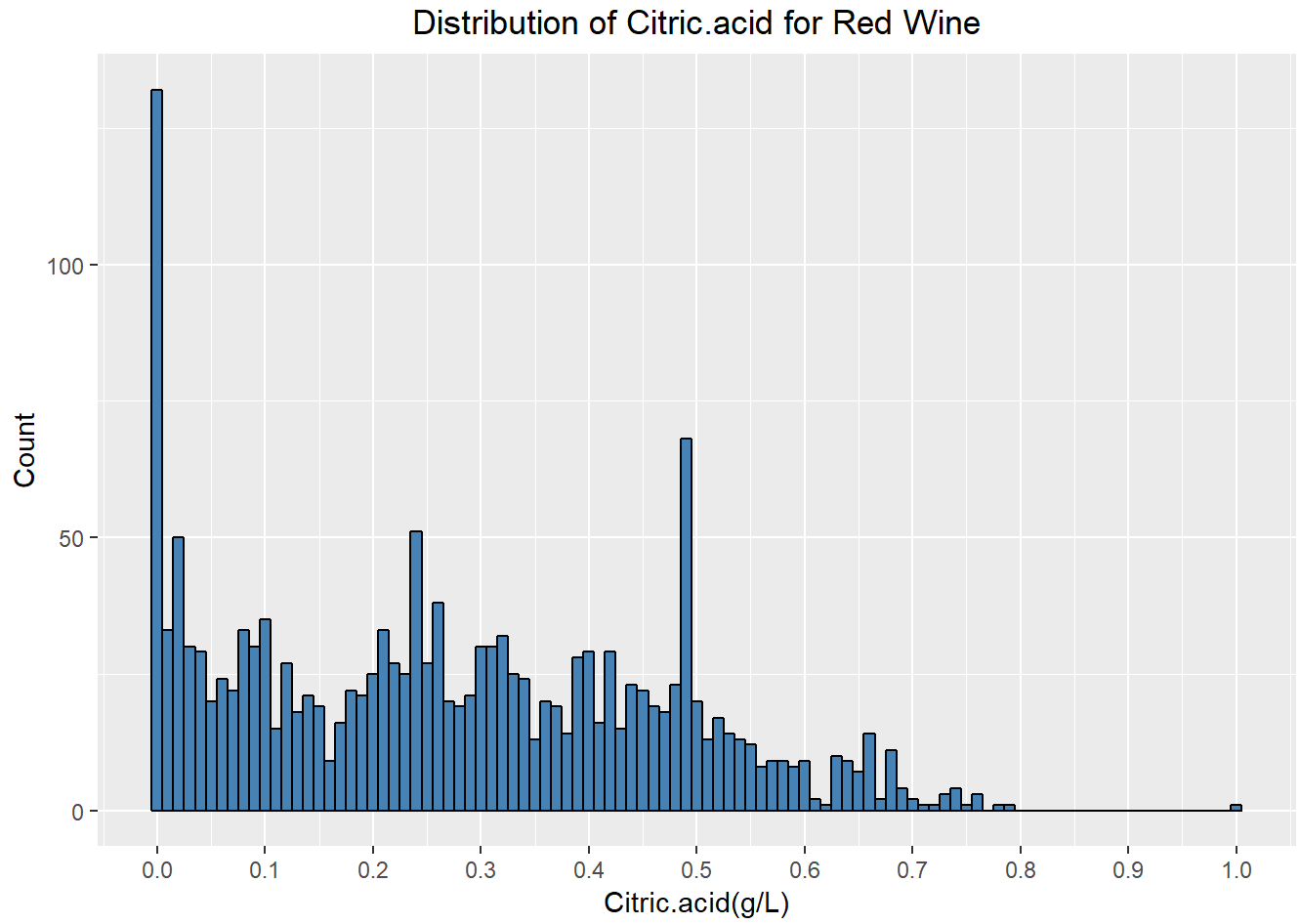
通过观察发现，挥发性酸的分布呈现近似正态分布，主要集中在0.39到0.64之间，平均值为0.53，中位数为0.52。

2.5 Citric.acid 柠檬酸

2.5.1 查看柠檬酸的描述性统计信息

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

2.5.2 绘制柠檬酸的直方图



通过观察发现，柠檬酸主要集中在0.09到0.42之间，平均值为0.27，中位数为0.26。

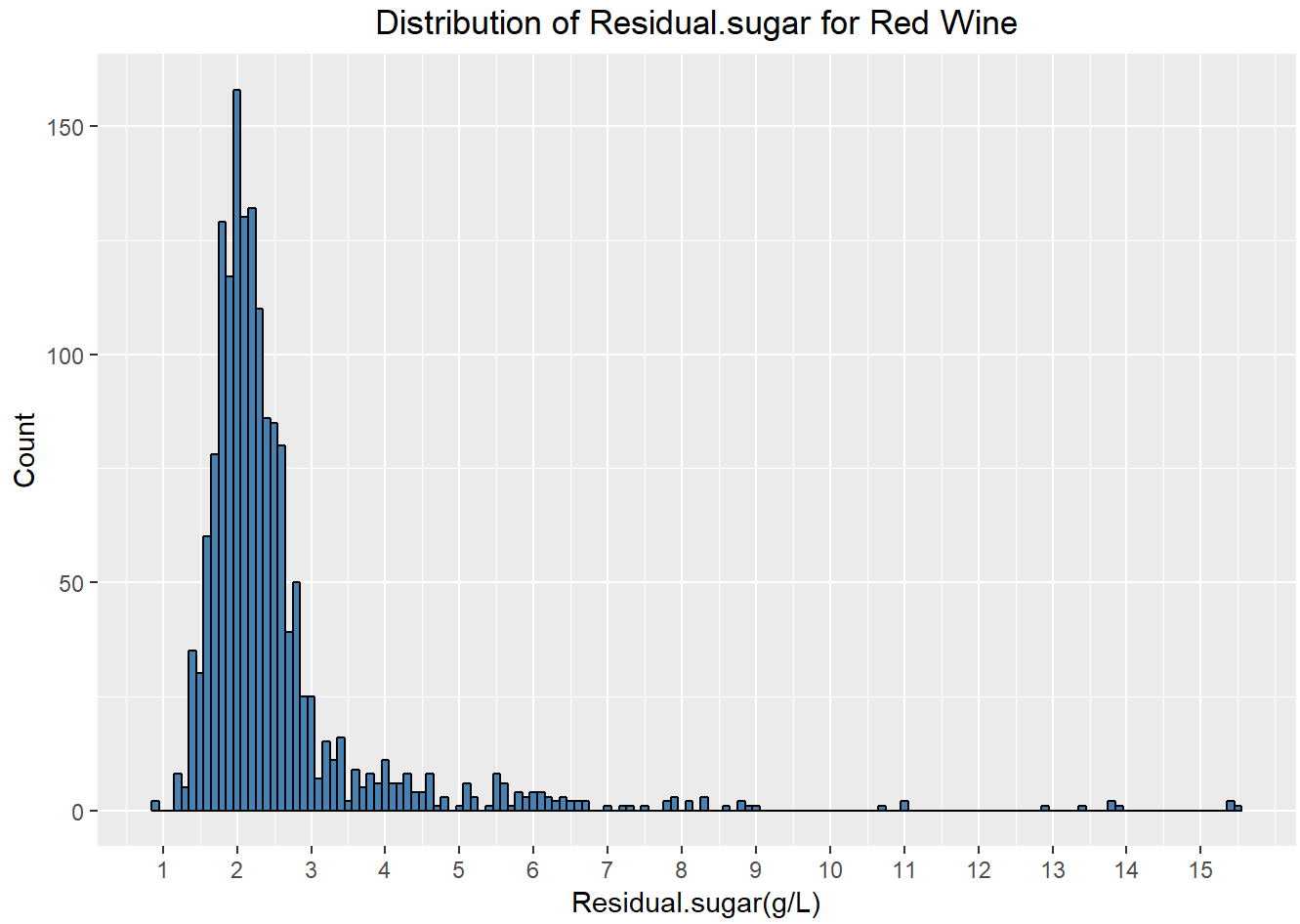
通过对比观察上面三种类型的酸可以发现，三种酸在红葡萄酒中的含量从高到低依次为：非挥发性酸 > 挥发性酸 > 柠檬酸，且三种类型的酸的数据均存在一定的异常值。

2.6 Residual.sugar 残留糖分

2.6.1 查看糖分的描述性统计信息

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

2.6.2 绘制糖分的直方图



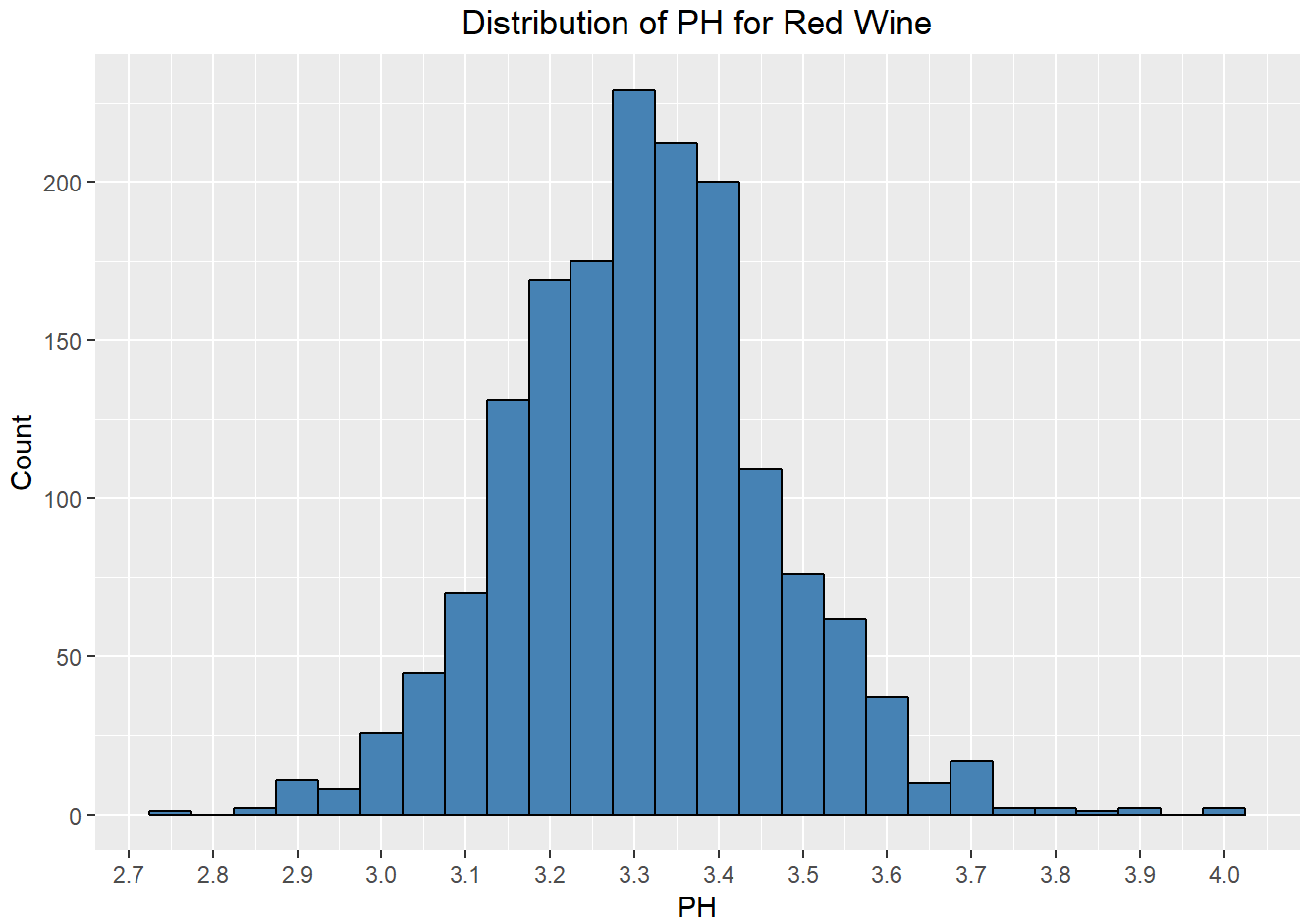
通过观察发现，糖分的分布呈现近似正态分布，主要集中在1.9到2.6之间，平均值为2.5，中位数为2.2，且存在大量的异常值。

2.7 PH 酸碱度

2.7.1 查看酸碱度的描述性统计信息

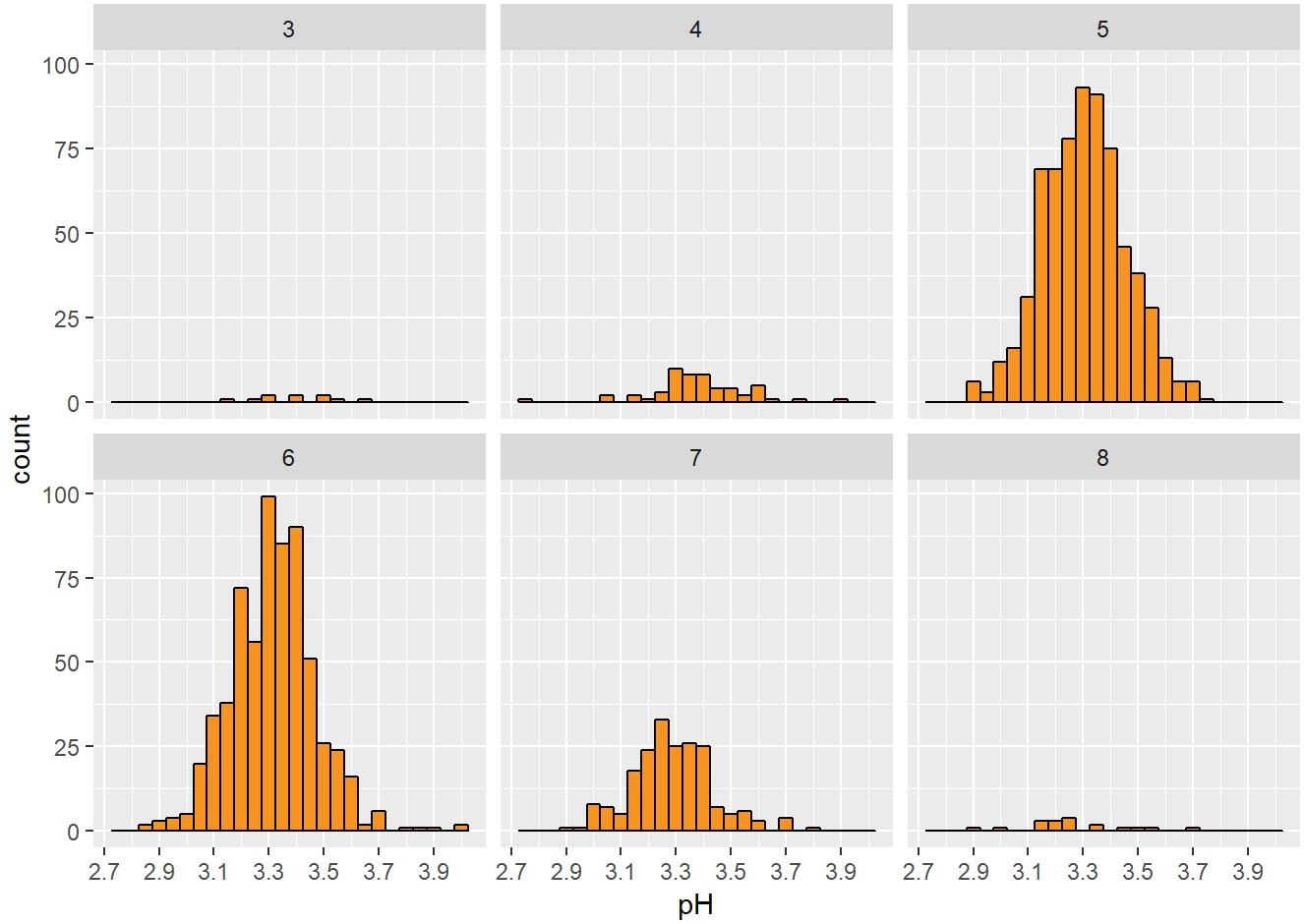
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

2.7.2 绘制酸碱度的直方图



通过观察发现，酸碱度的分布呈现近似正态分布，主要集中在3.2到3.4之间，平均值为3.3，中位数为3.3，变化范围很小，且均为酸性，在口感上不会有明显的差异，初步认为PH值对红葡萄酒的质量等级影响不大。

2.7.3 使用变量“质量”将PH值进行分面展示



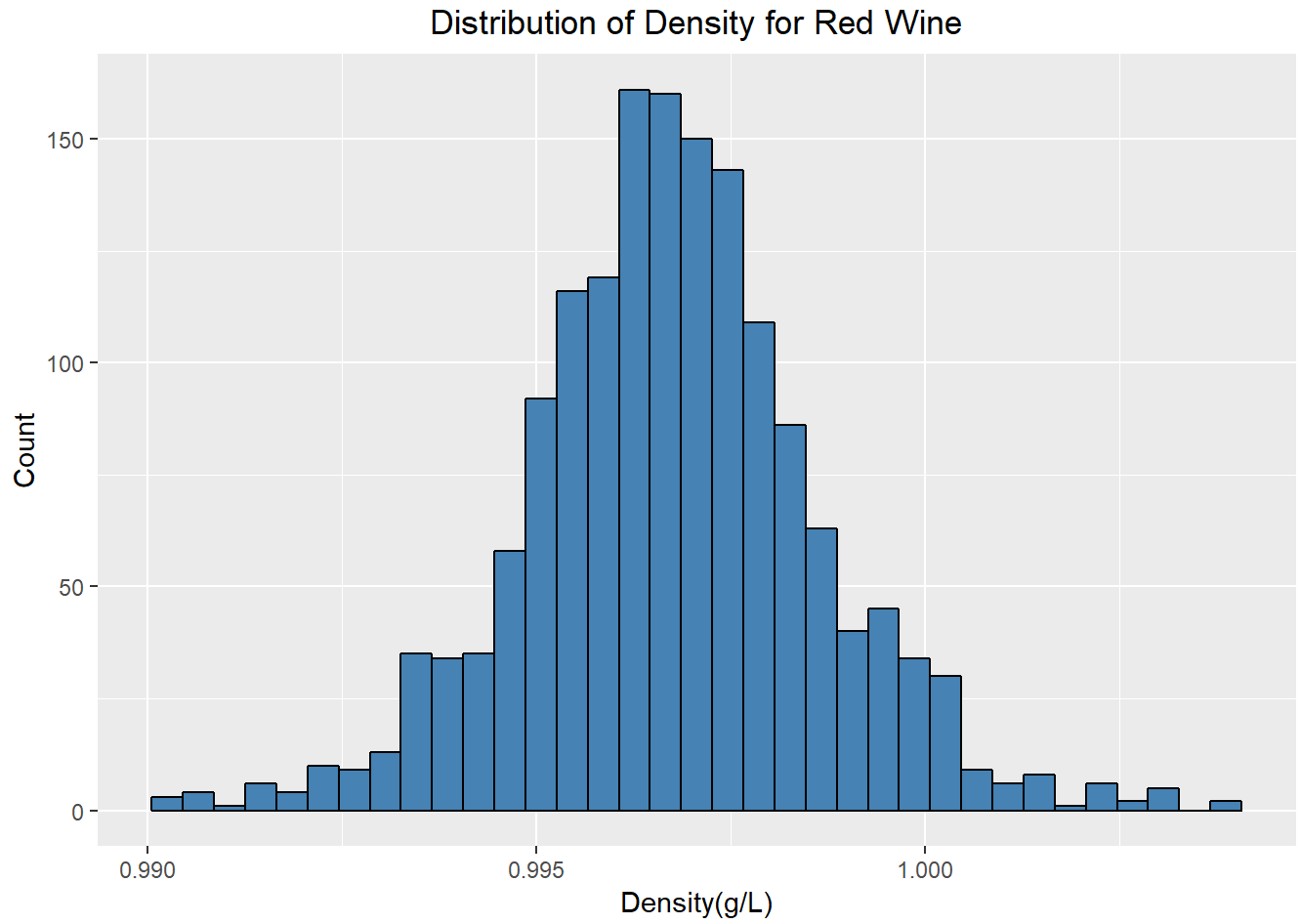
通过观察发现，PH值在质量等级为5、6、7三个分面中均呈现近似正态分布，进一步证实了PH值对红葡萄酒的质量等级关系不大。

2.8 Density 密度

2.8.1 查看密度的描述性统计信息

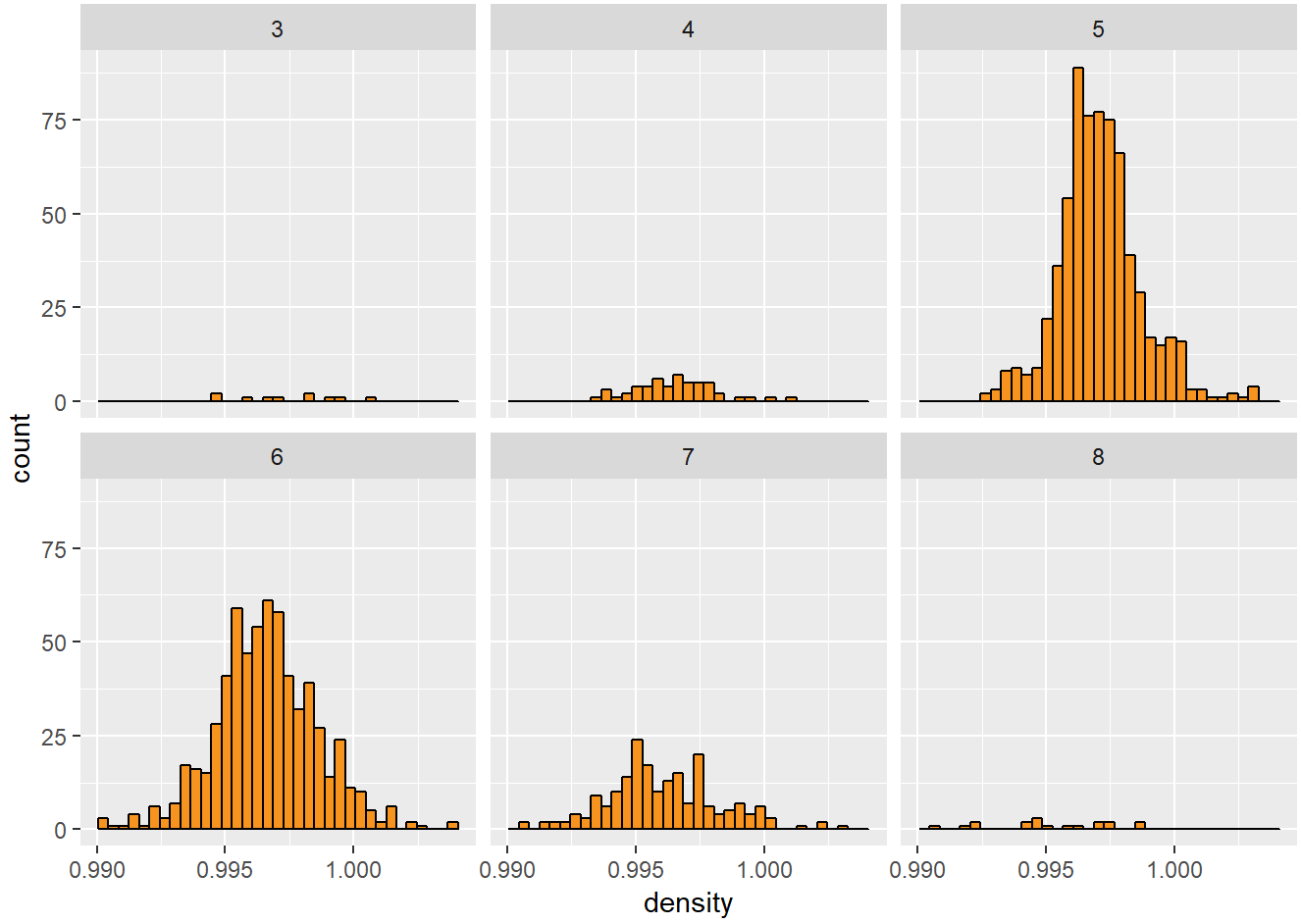
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

2.8.2 绘制密度的直方图



通过观察发现，密度的分布呈现近似正态分布，主要集中在0.9956到0.9978之间，平均值为0.9967，中位数为0.9968，变化范围很小，初步认为密度对红葡萄酒的质量等级影响不大。

2.8.3 使用变量“质量”将密度进行分面展示



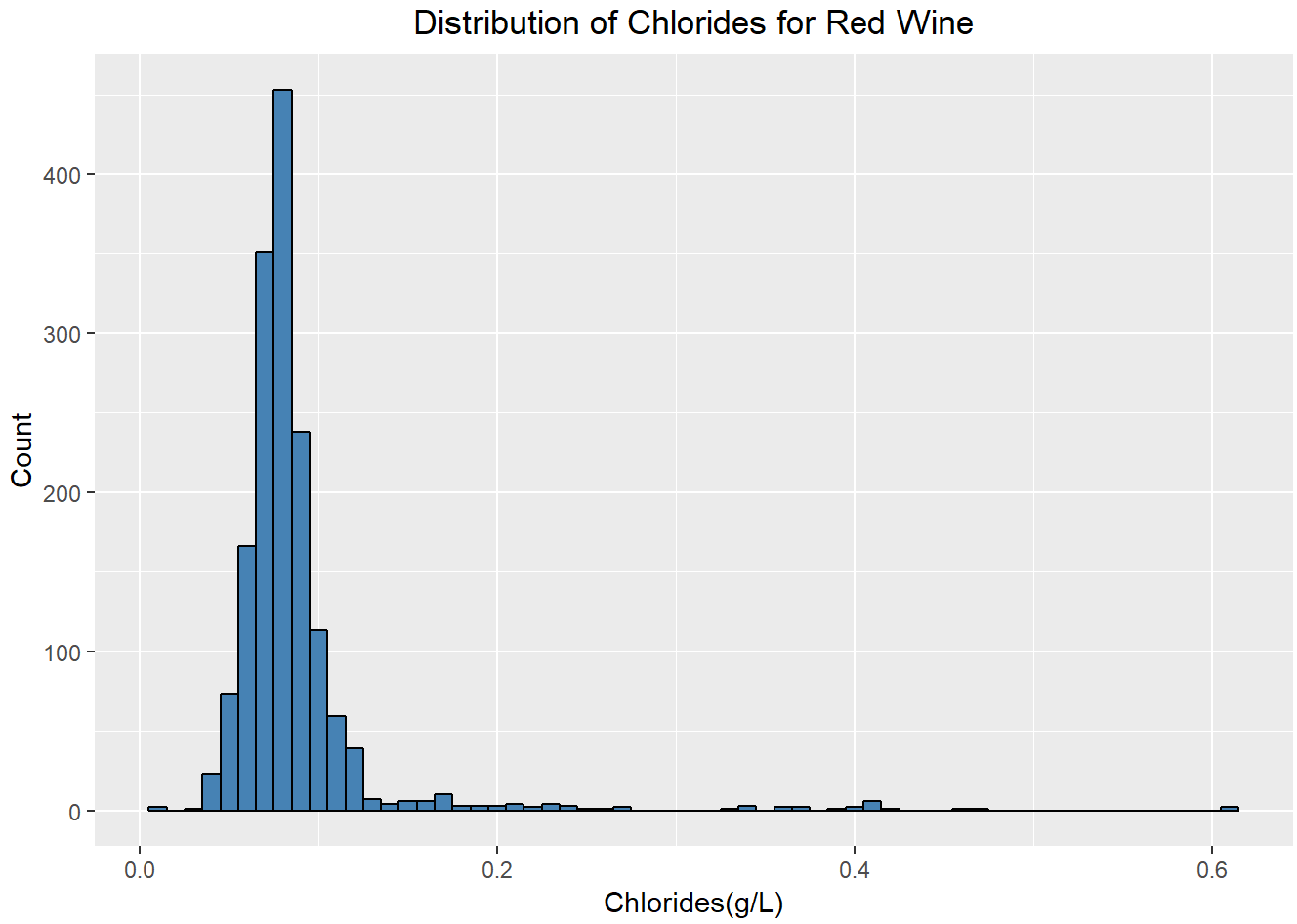
通过观察发现，密度在质量等级为5、6、7三个分面中均呈现近似正态分布，进一步证实了密度对红葡萄酒的质量等级关系不大。

2.9 Chlorides 氯化物

2.9.1 查看氯化物的描述性统计信息

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

2.9.2 绘制氯化物的直方图



通过观察发现，氯化物的分布出现了长尾情况，初步认为是大量的异常值造成的，除开异常值其分布基本是正态分布，主要集中在0.07到0.09之间，平均值为0.087，中位数为0.079，变化范围很小，且在红葡萄酒中含量很低。

通过网上查阅相关信息，氯化物的主要作用是红葡萄酒生产过程中杀菌用的，因此初步考虑可以忽略氯化物对红葡萄酒的质量影响。

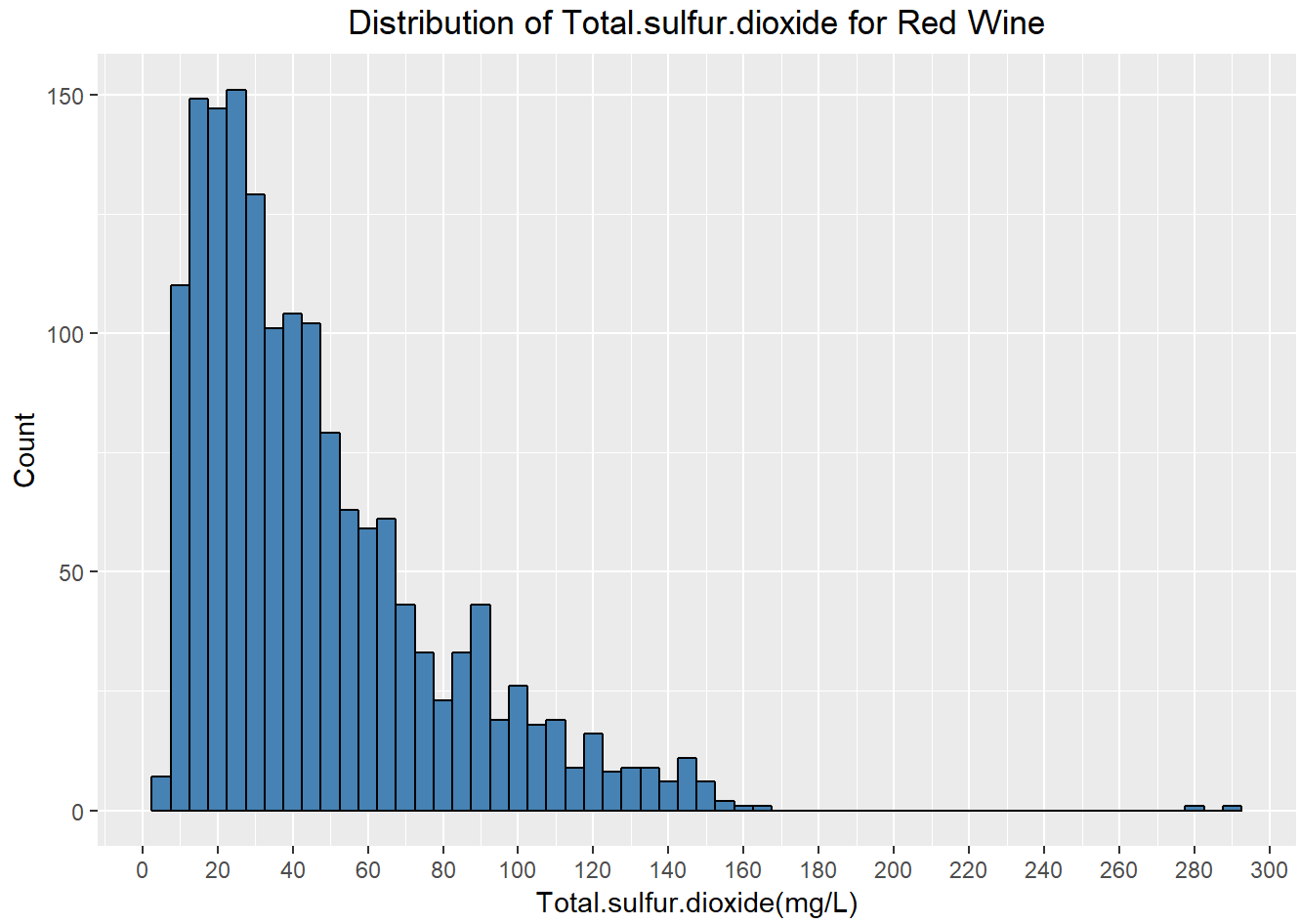
2.10 Total.sulfur.dioxide 二氧化硫总量

因为考虑到二氧化硫总量已经包含了游离二氧化硫，两者是完全的正相关关系，因此这里只分析二氧化硫总量。

2.10.1 查看二氧化硫总量的描述性统计信息

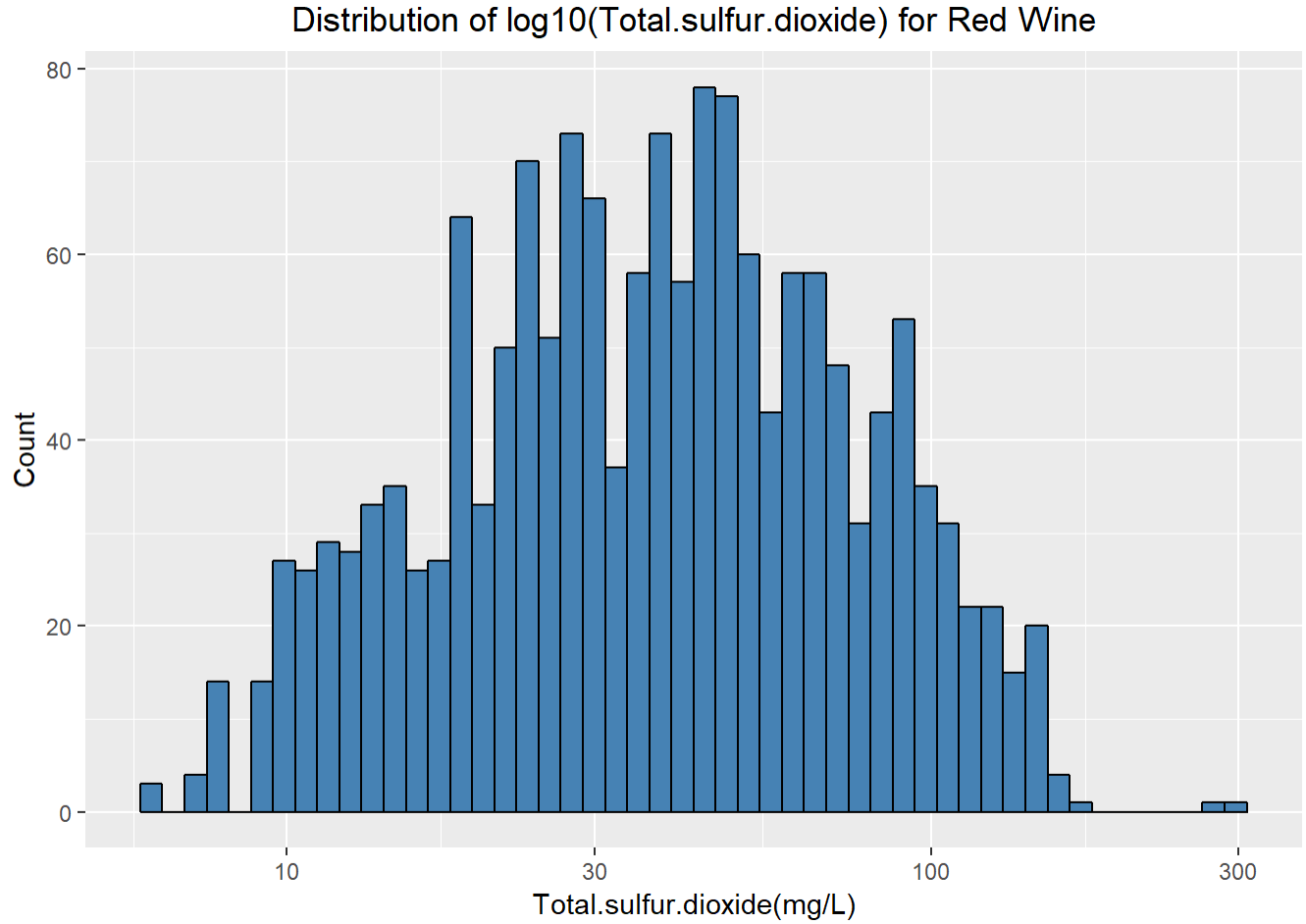
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

2.10.2 绘制二氧化硫总量的直方图



通过观察发现，二氧化硫总量呈现左偏的分布，对其使用对数转换处理。

2.10.3 将二氧化硫总量进行数据变换后绘制直方图



通过观察发现，经过对数转换后，二氧化硫总量基本呈现正态分布，主要集中在22到62之间，平均值为46，中位数为38。

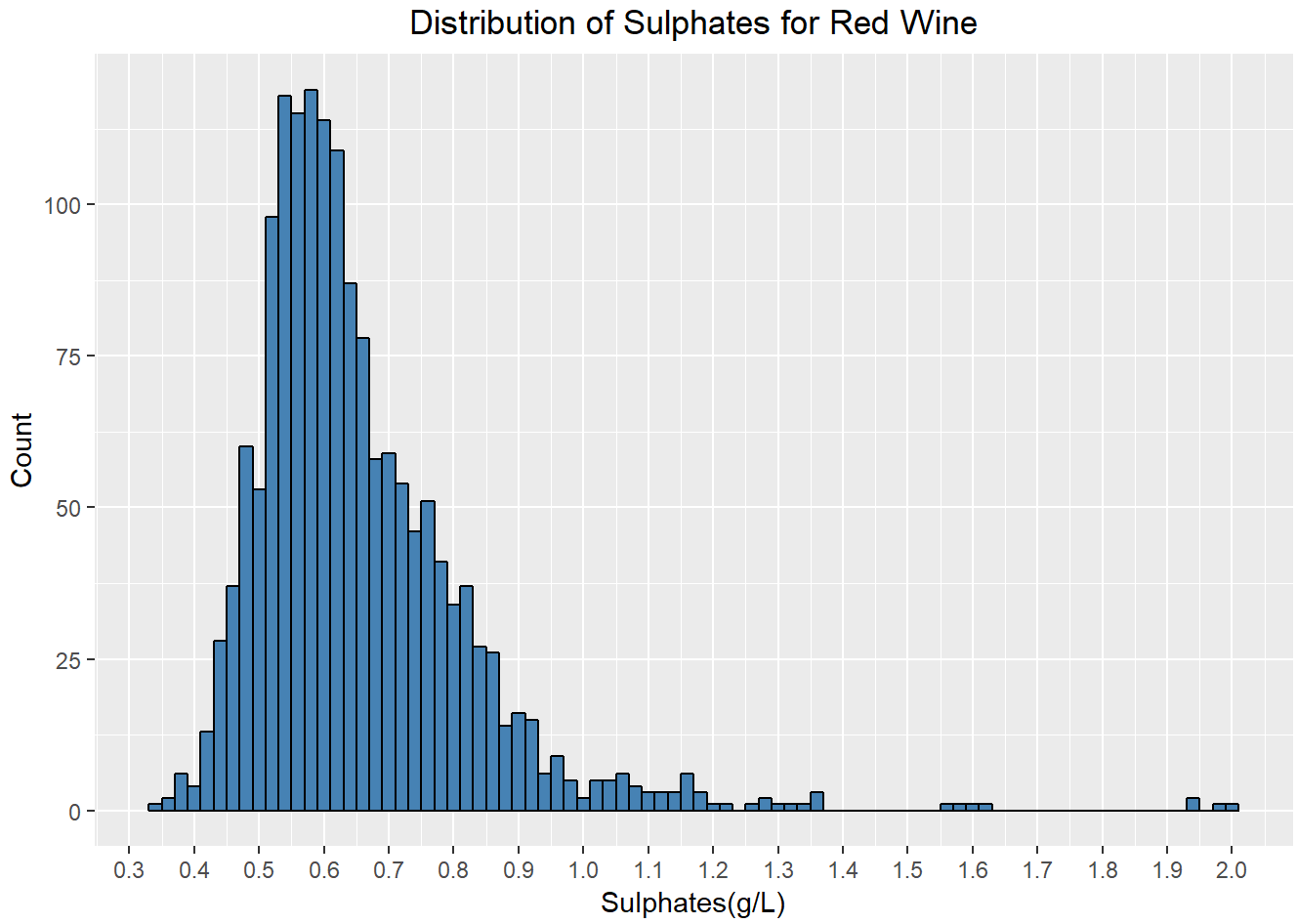
通过网上查阅相关资料，发现红葡萄酒中加入二氧化硫是为了防止葡萄酒杯氧化成醋，在酿造过程中具有抗氧化和杀菌的作用，就整体而言，其在酒中的含量非常少，其适当的摇杯或者醒酒可以令其挥发掉，因此二氧化硫总量对质量的影响可以不考虑。

2.11 Sulphates 硫酸盐

2.11.1 查看硫酸盐的描述性统计信息

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

2.11.2 绘制硫酸盐的直方图



通过观察发现，硫酸盐的分布出现了长尾情况，初步认为是异常值造成的，除开异常值其分布基本是正态分布，主要集中在0.55到0.73之间，平均值为0.66，中位数为0.62。

通过网上查阅相关资料发现，红葡萄酒中的硫酸盐是一种防腐剂，它能起到杀菌和抗氧化的作用，其在酒中的含量非常小，基本不会产生任何味道，因此考虑忽略其对质量的影响。

2.12 进一步分析

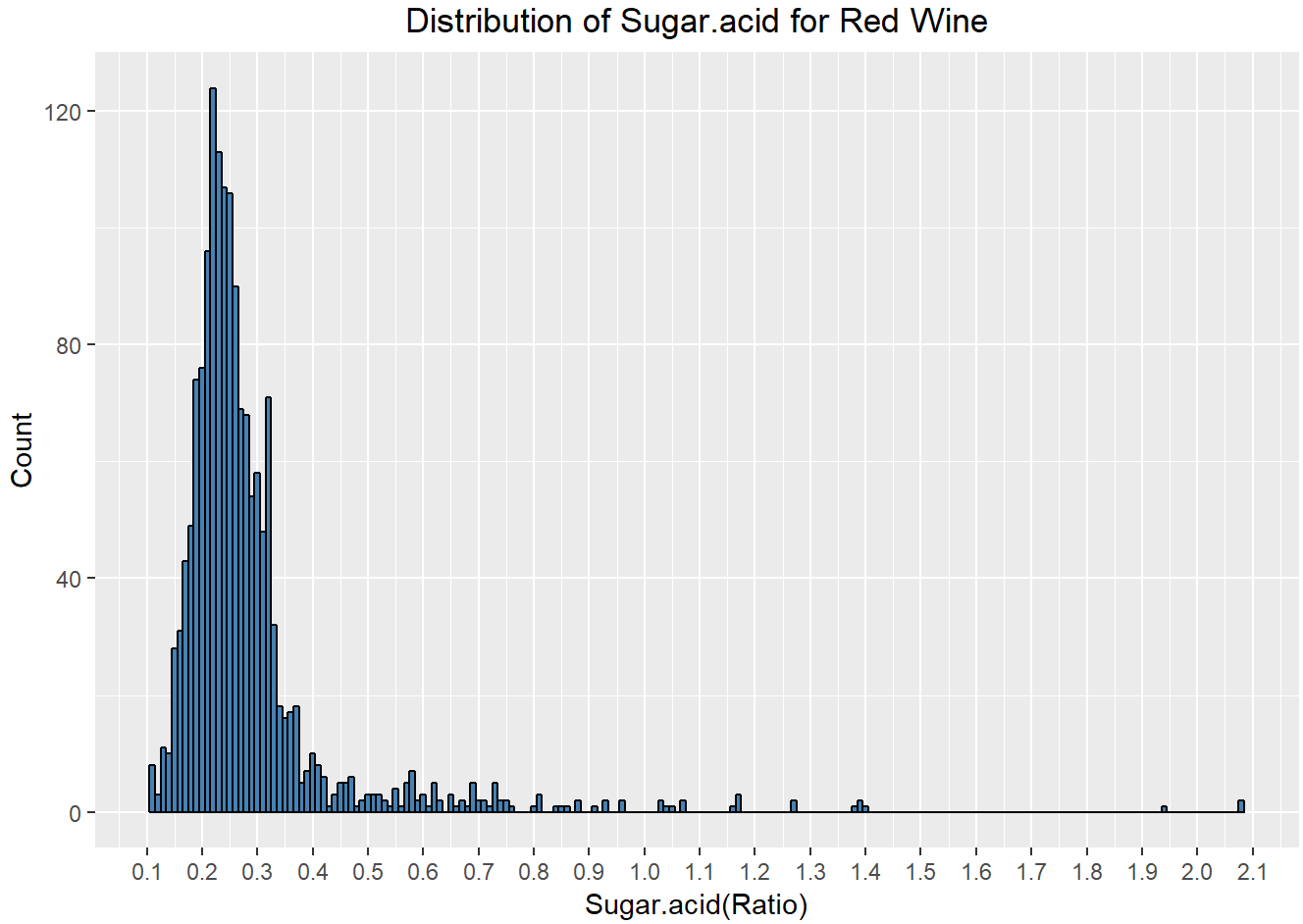
2.12.1 新增变量sugar.acid

从红葡萄酒的口感上来分析，人的味觉主要感受到的是甜和酸，而酒精度的影响作为一个单独的变量分析。因此，这里新增一个变量糖分和酸的比值sugar.acid来进行分析，其中酸的含量由非挥发性酸、挥发性酸和柠檬酸三部分组成。

2.12.2 查看糖酸比的描述性统计信息

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1053  0.2117  0.2482  0.2854  0.3008  2.0807
```

2.12.3 绘制糖酸比的直方图



通过观察发现，糖酸比的分布存在大量的异常值，需要进一步处理。

2.12.4 处理糖酸比的异常值

这里使用箱线图分析法，通过设置自定义函数boxout来识别异常值，并设定超出1.5倍IQR为异常值

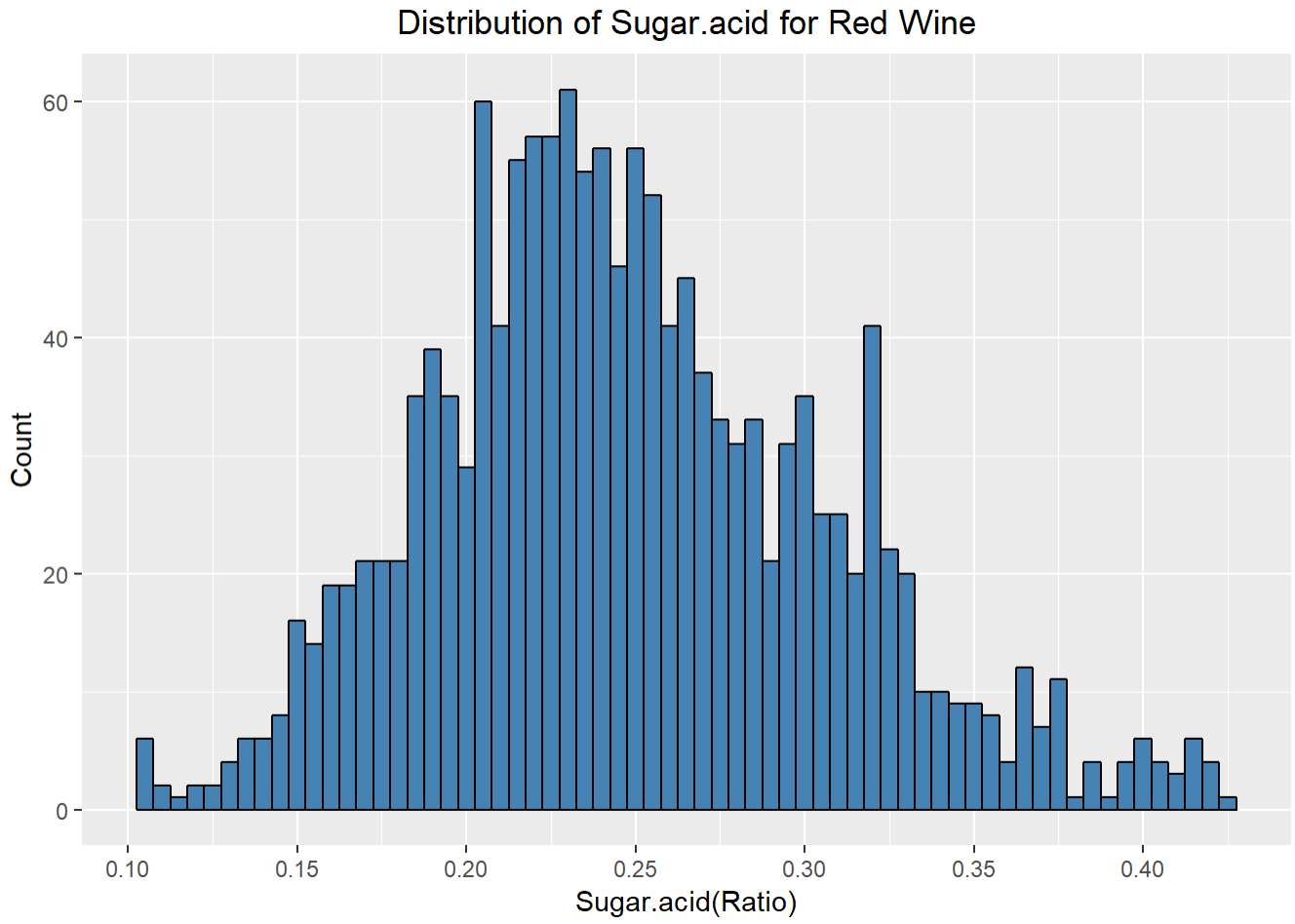
```
##   [1] 0.7297 0.7297 0.5452 1.4033 0.6513 0.7275 0.7275 0.4545 0.6182 0.5831
##  [11] 0.5831 0.6918 0.6918 0.6918 0.6918 0.8838 0.8834 0.5033 0.7018 0.4703
##  [21] 0.8403 0.4703 0.5781 0.5821 0.5821 1.0290 1.0290 0.4888 1.0741 1.0741
##  [31] 0.6187 0.6900 1.3753 1.1657 0.5661 0.5661 0.6325 0.6029 0.7976 0.4470
##  [41] 0.4586 0.4672 0.7477 0.7477 0.7026 0.4489 1.1637 0.4435 0.5528 0.4950
##  [51] 0.6260 0.5026 0.4387 0.4492 0.5070 0.5528 0.8054 0.8054 0.9146 0.4486
##  [61] 1.1702 1.1702 0.4690 0.6184 0.4663 0.4596 0.6513 0.8641 0.4635 0.4716
##  [71] 0.9563 0.9563 0.6653 0.6653 0.9347 0.9347 0.5087 0.7417 0.7417 0.7288
##  [81] 0.7097 0.4591 0.5709 0.7557 1.0498 0.4568 0.5355 0.5055 1.9398 0.5997
##  [91] 2.0807 0.6107 0.6774 0.5696 0.5696 0.5947 0.5947 0.5152 0.6577 0.4358
## [101] 0.8079 0.5485 0.5809 0.5809 0.5556 1.3861 1.3861 0.5996 0.6209 0.6209
## [111] 0.5013 1.2661 1.2661 0.7103 0.4826 0.5157 0.5157 0.6471 0.5293 0.8524
## [121] 2.0777 0.7234 1.0365 0.5310
```

通过观察可以发现，一共有124个异常值，占所有观测值的7.75%左右，这里将异常值做删除处理。

2.12.5 查看处理异常值后的糖酸比的描述性统计信息

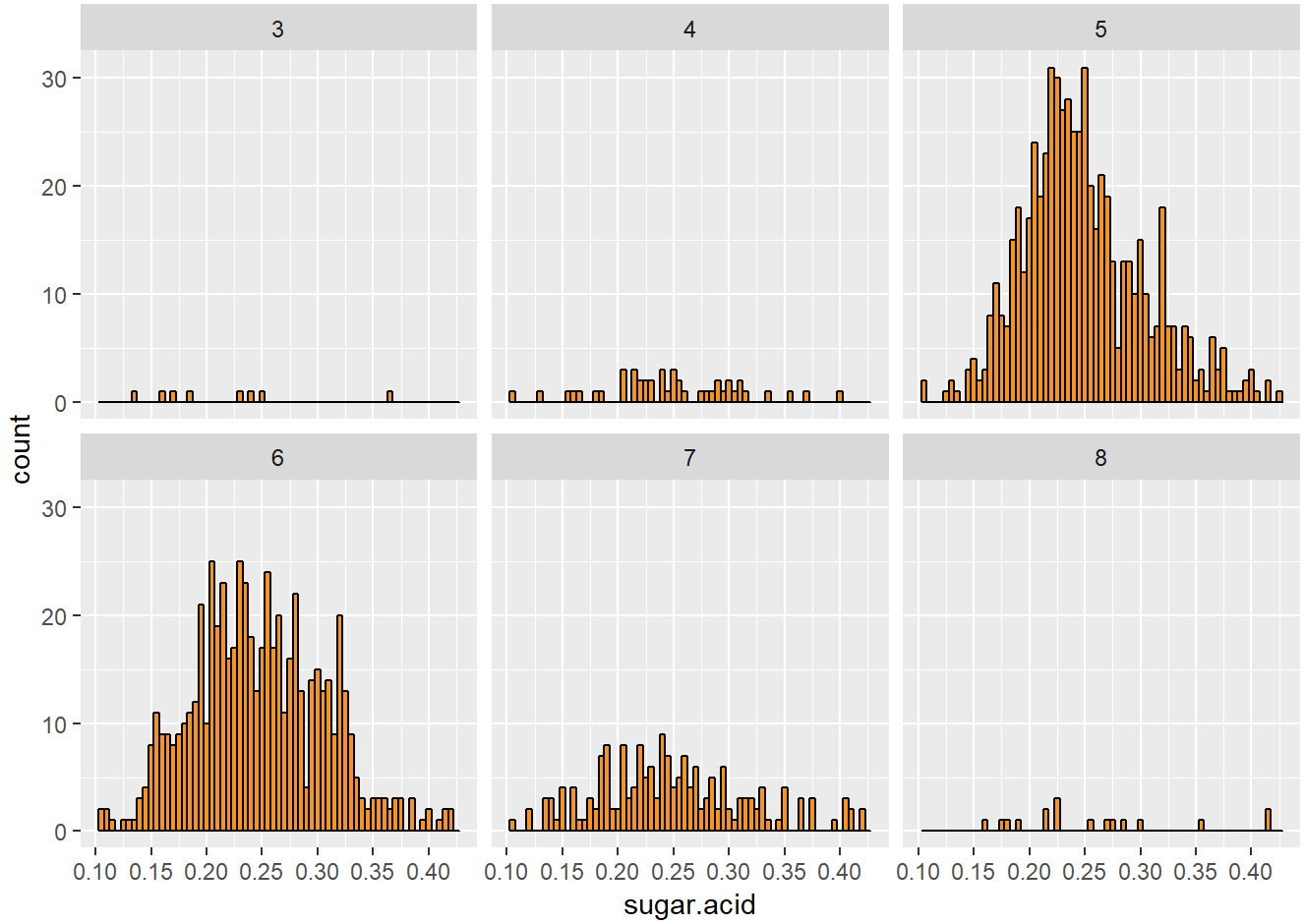
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1053  0.2083  0.2418  0.2482  0.2852  0.4255
```

2.12.6 绘制处理异常值后的糖酸比的直方图



通过观察发现，去掉异常值后，糖酸比的分布基本是一个正态分布，主要集中在0.20到0.29之间，平均值为0.25，中位数为0.24。

2.12.7 使用变量“质量”将糖酸度进行分面展示



通过观察发现，在质量为5、6、7三级的分面图中，糖酸度的分布均基本呈正态分布，初步认为糖酸度对红葡萄酒的质量没有很大影响。而在质量为3、4、8三级的分面图中，由于数据较少，不具备统计学意义，故不做分析。

2.13 单变量分析小结

- 针对原数据集中的1599个观测值和13个变量，主要选取了质量、酒精度、非挥发性酸、挥发性酸、柠檬酸、残留糖分、PH值、密度、氯化物、二氧化硫总量、硫酸盐这11个变量进行单变量分析。
- 其中PH值主要集中在3.2到3.4之间，对葡萄酒的口感影响不大，因此后续分析中可以忽略PH值的影响。
- 同时，其他几个变量氯化物、游离二氧化硫、二氧化硫总量、密度、硫酸盐对红葡萄酒的品质影响不大，故后续分析时可以忽略。
- 此外，从红葡萄酒的口感方面入手，新增了变量糖酸比进行单变量分析，发现在除去124个异常值之后，糖酸比的分布接近正态分布，主要集中在0.20到0.29之间。且通过质量将糖酸度进行分面后发现，糖酸度对红葡萄酒的质量影响不大。

3. Bivariate Plots Section 双变量分析

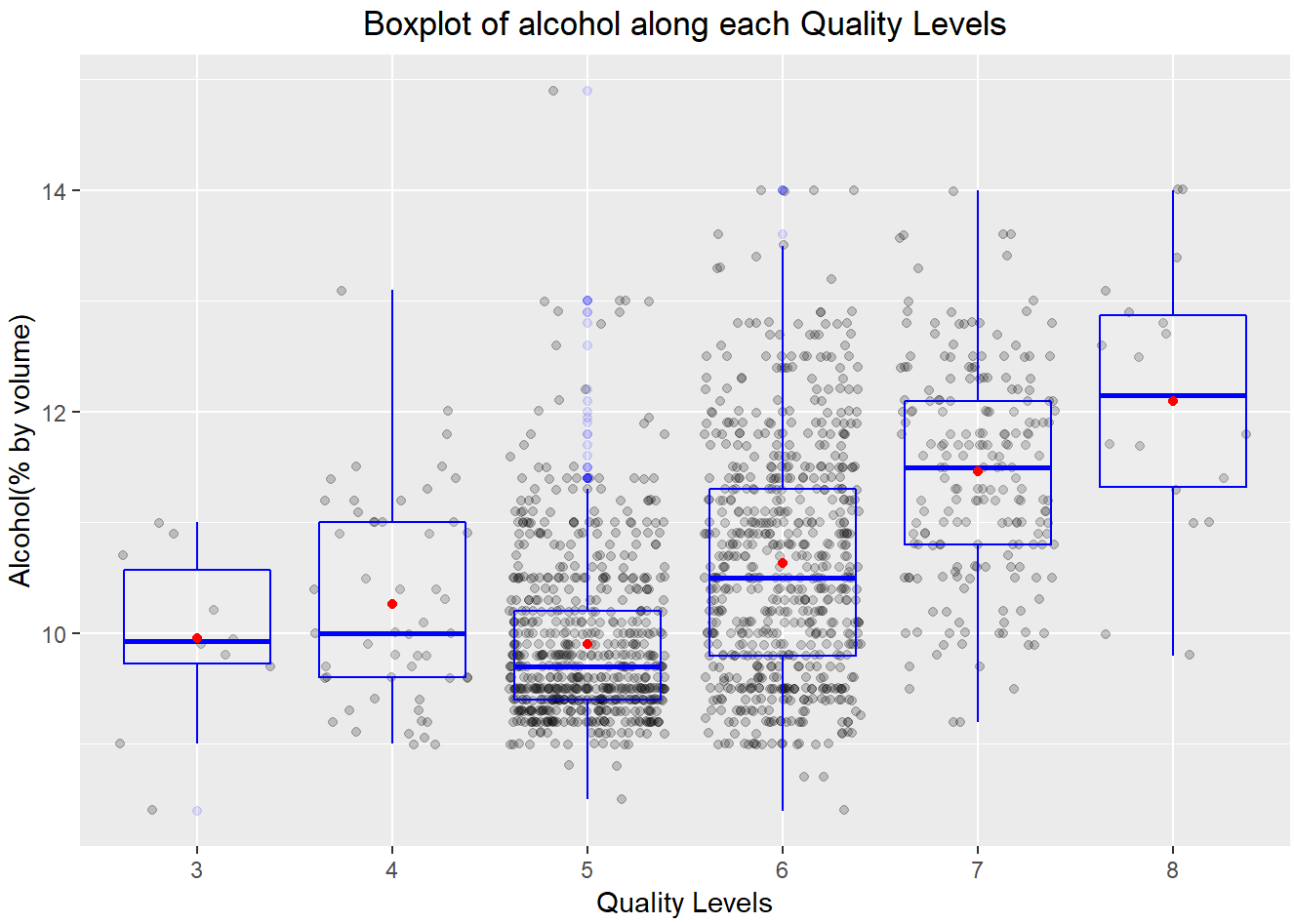
根据单变量分析的结果，双变量分析时将主要考虑酒精度、非挥发性酸、挥发性酸、柠檬酸、残留糖分和糖酸比与质量之间的相互关系。

绘图说明：对于分类变量和数值型变量之间关系的探索，最常用的分析图像是箱线图，该数据集中，quality属于定序变量，因此，分析质量和其他变量之间的关系时，采用boxplot绘图。

因此，为了方便根据不同质量等级绘制箱线图，同时考虑到后面需要进行多变量分析，新增一列变量quality_factor将quality的数据类型由int转换为factor存入其中。

3.1 酒精度和质量

3.1.1 绘制酒精度和质量的箱线图



3.1.2 计算酒精度和质量的相关性系数

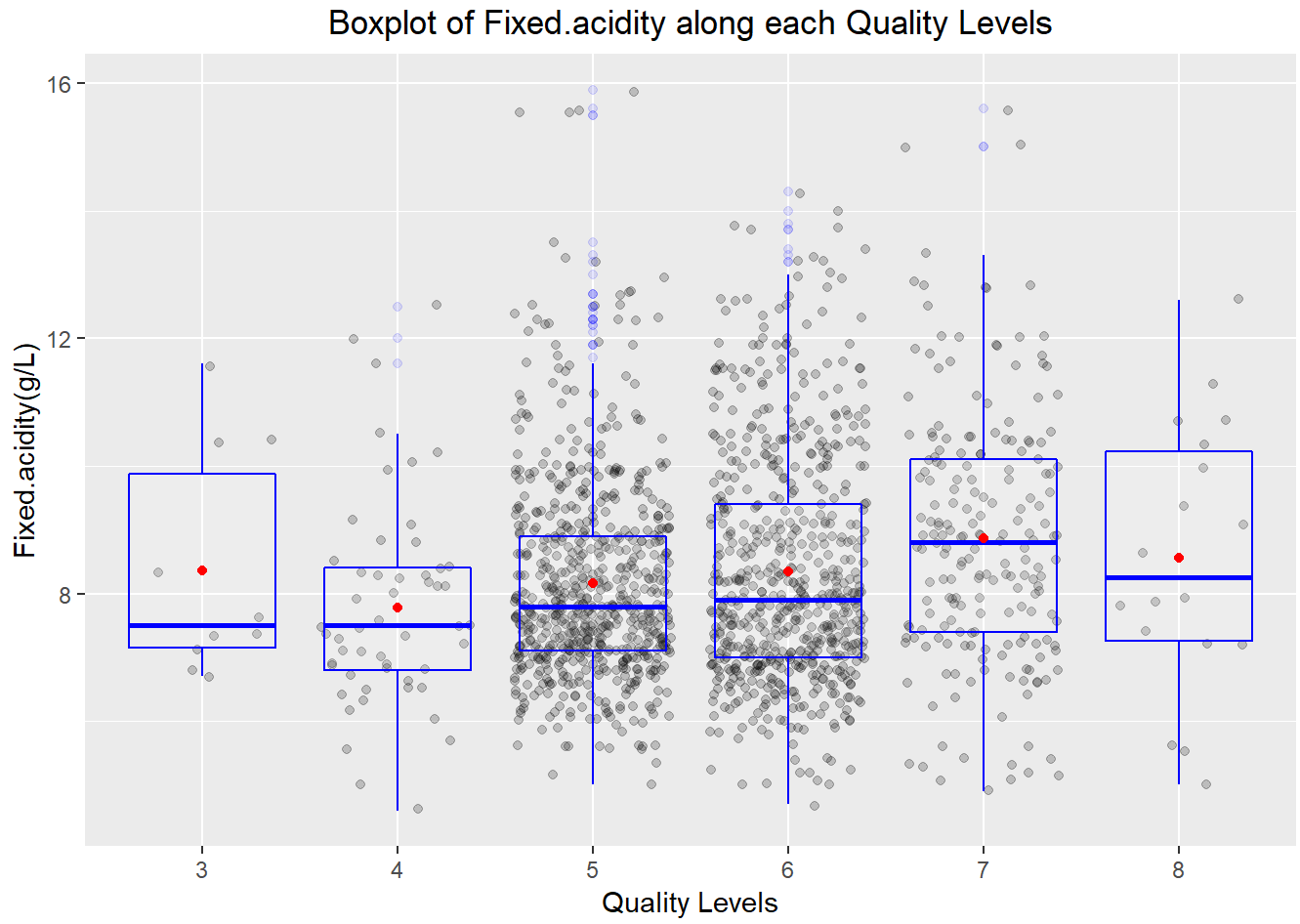
说明：计算相关性系数时均采用pearson方法进行计算。

```
##
##  Pearson's product-moment correlation
##
## data:  alcohol and quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

通过以上过程可以发现，酒精度和红葡萄酒的质量呈明显的正相关性，即在一定范围内，随着酒精度的增加，红葡萄酒的质量等级相应缓慢提升，相关性系数为0.4762，相关性强度为Moderate中等。

3.2 非挥发性酸和质量

3.2.1 绘制非挥发性酸和质量的箱线图



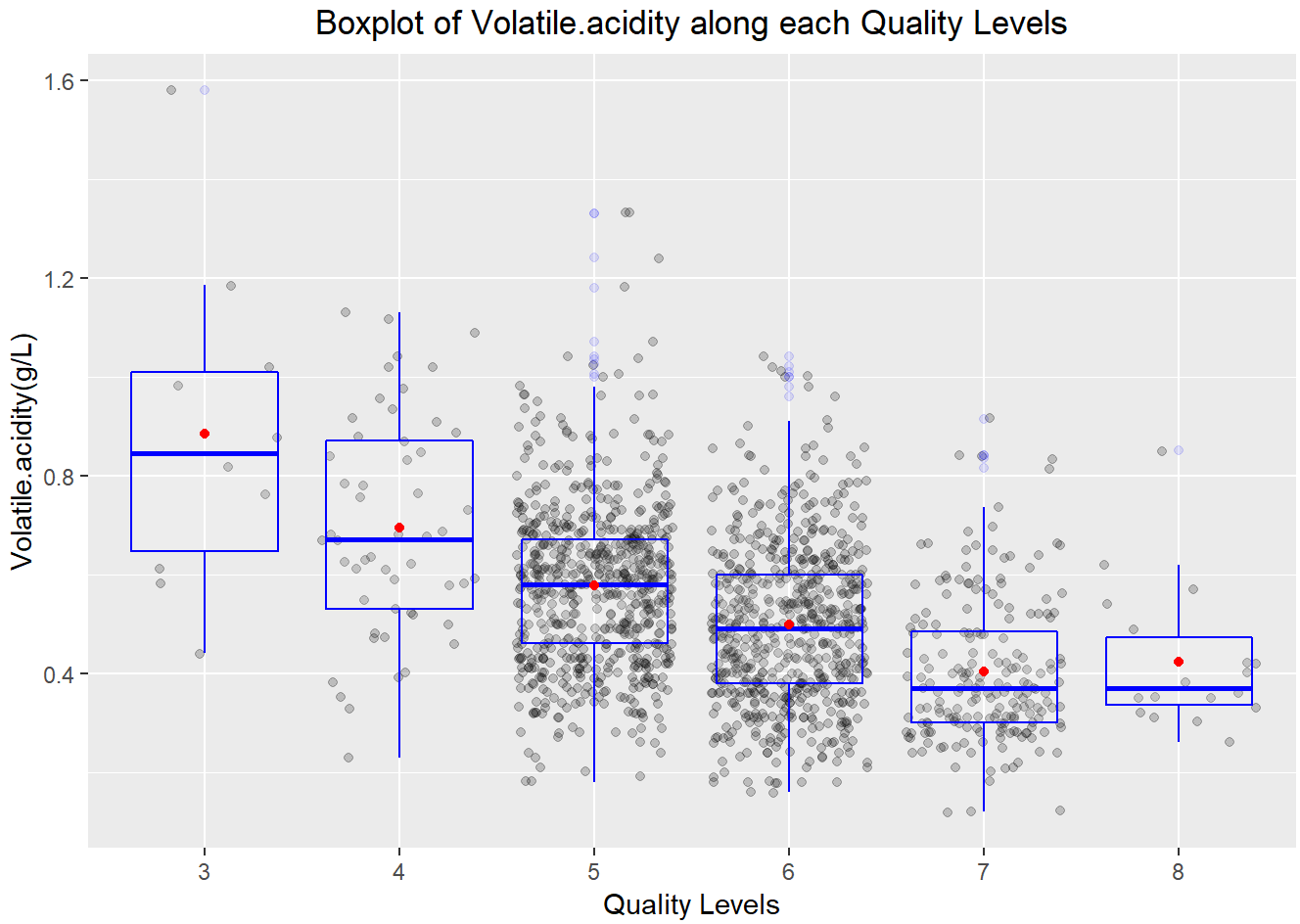
3.2.2 计算非挥发性酸和质量的相关性系数

```
##
##  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
```

通过以上过程可以发现，非挥发性酸和红葡萄酒的质量呈非常弱的正相关性，相关性系数为0.1241，相关性强度为very low非常弱，因此可以考虑忽略挥发性酸对红葡萄酒质量等级的影响。

3.3 挥发性酸和质量

3.3.1 绘制挥发性酸和质量的箱线图



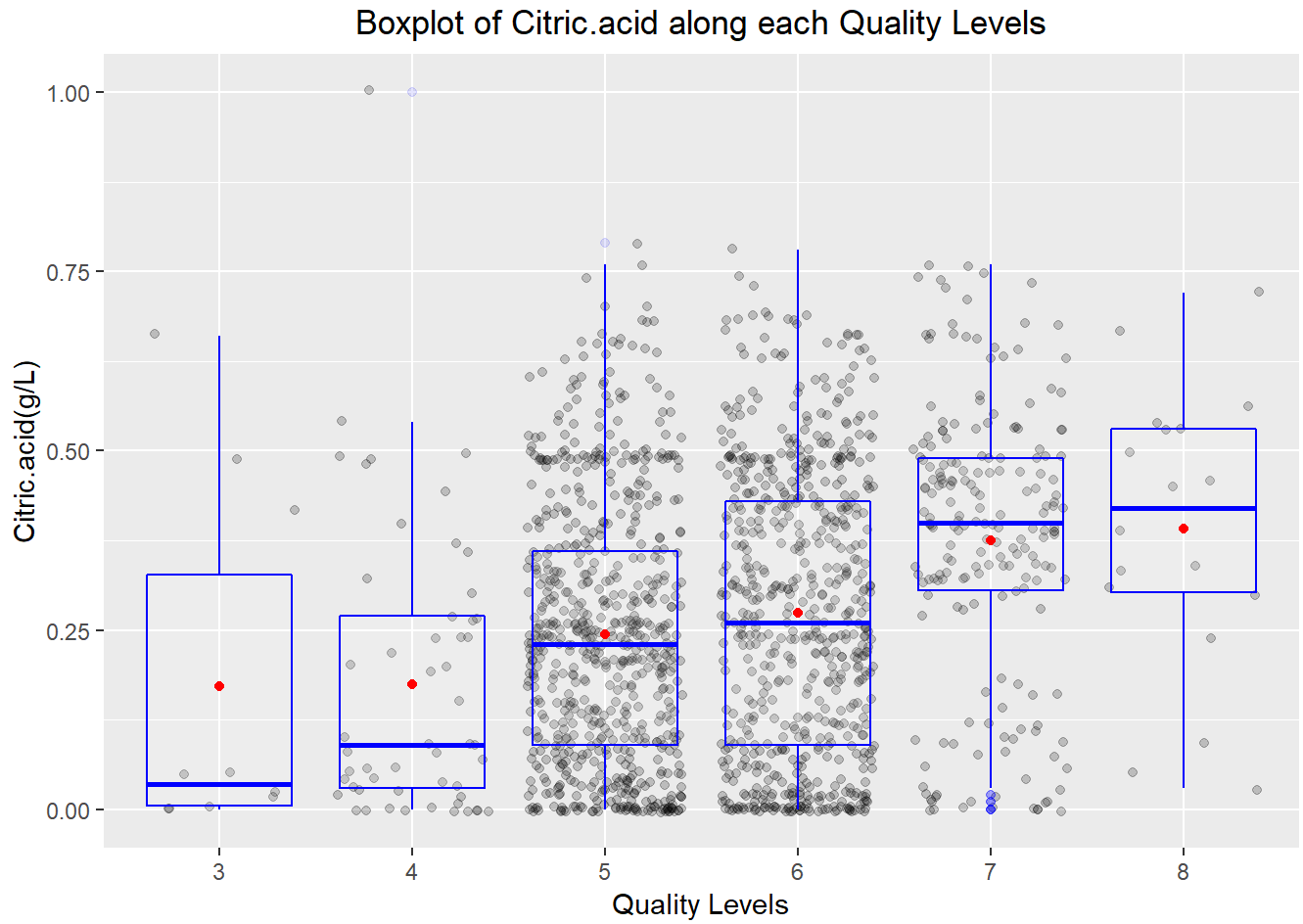
3.3.2 计算挥发性酸和质量的相关性系数

```
##
##  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
```

通过以上过程可以发现，挥发性酸和红葡萄酒的质量呈明显的负相关性，即在一定范围内，随着挥发性酸的增加，红葡萄酒的质量等级相应缓慢下降，相关性系数为0.3906，相关性强度为Moderate中等。

3.4 柠檬酸和质量

3.4.1 绘制柠檬酸和质量的箱线图



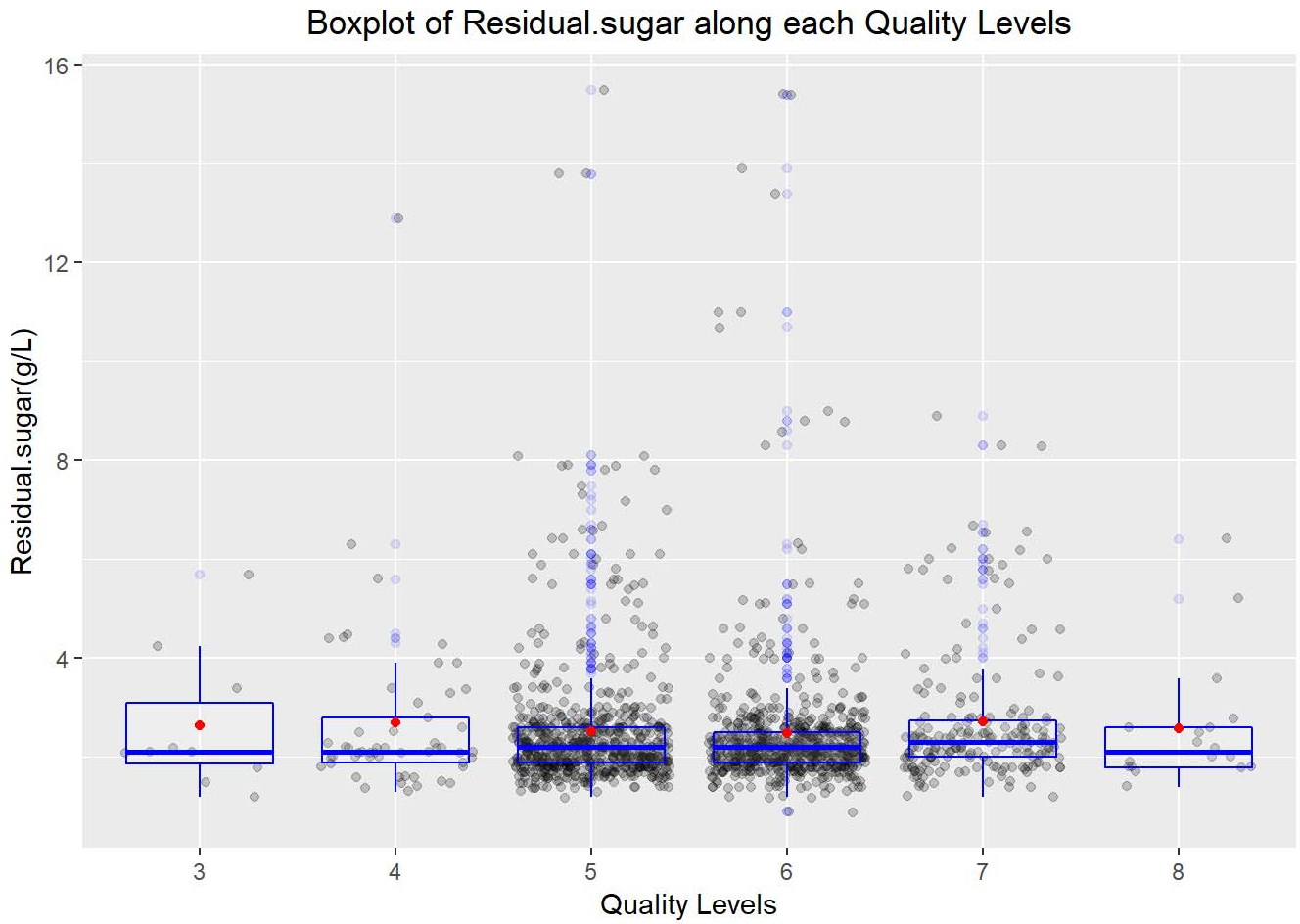
3.4.2 计算柠檬酸和质量的相关性系数

```
##
##  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
```

通过以上过程可以发现，柠檬酸和红葡萄酒的质量呈较弱的正相关性，即在一定范围内，随着柠檬酸的增加，红葡萄酒的质量等级相应微弱提升，相关性系数为0.2264，相关性强度为Low弱。

3.5 残留糖分和质量

3.5.1 绘制残留糖分和质量的箱线图



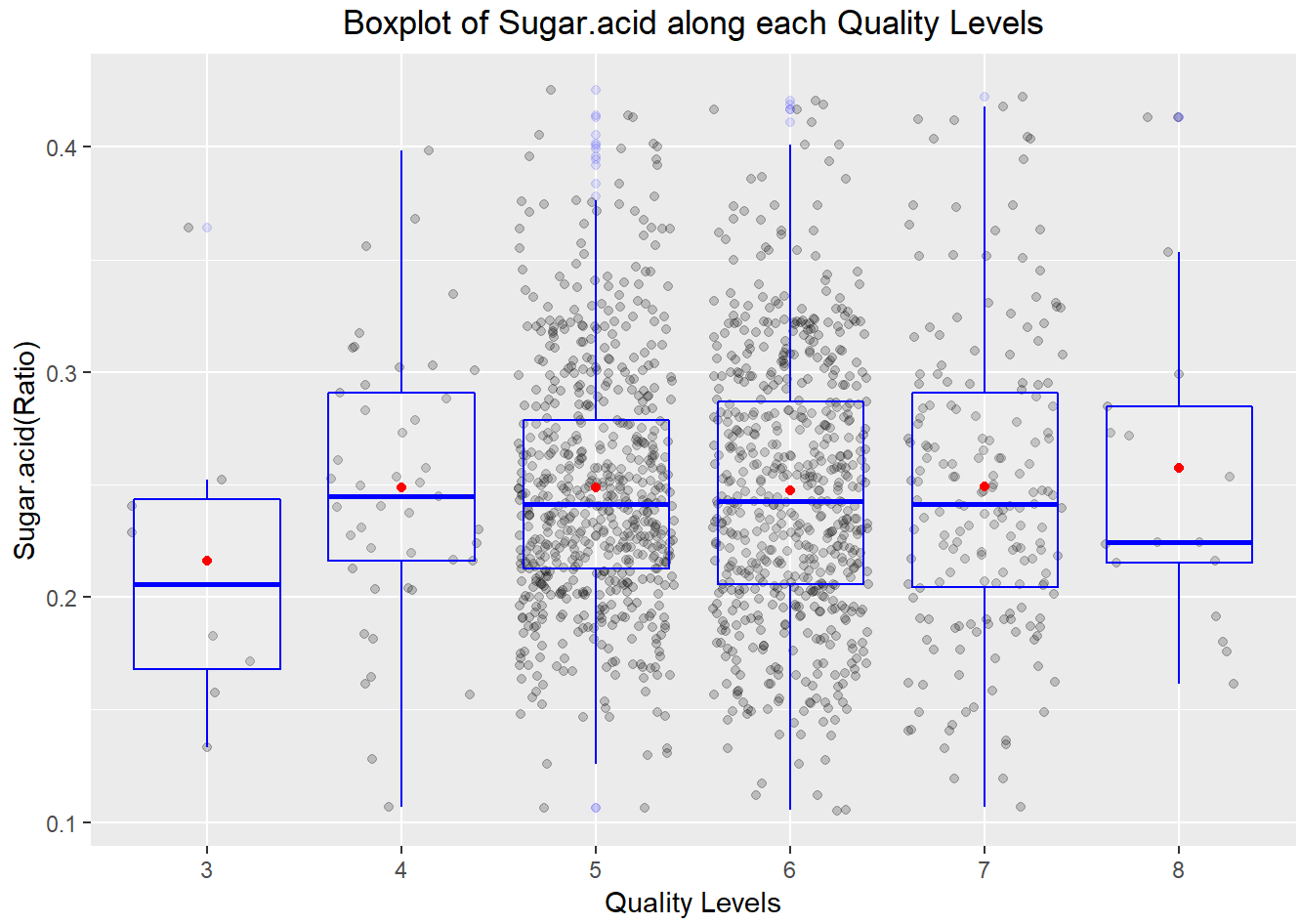
3.5.2 计算残留糖分和质量的相关性系数

```
##
##  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
```

通过以上过程可以发现，残留糖分酸和红葡萄酒的质量呈非常弱的正相关性，相关性系数为0.0137，相关性强度为very low非常弱，因此可以忽略残留糖分对红葡萄酒质量等级的影响。

3.6 糖酸比和质量

3.6.1 绘制糖酸比和质量的箱线图



3.6.2 计算糖酸比和质量的相关性系数

```
##
##  Pearson's product-moment correlation
##
## data:  sugar.acid and quality
## t = 0.53406, df = 1473, p-value = 0.5934
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03715320  0.06490849
## sample estimates:
##        cor
## 0.01391389
```

通过以上过程可以发现，糖酸比和红葡萄酒的质量呈非常弱的正相关性，相关性系数为0.0139，相关性强度为very low非常弱，进一步证明了在单变量分析时初步认定的糖酸比对红葡萄酒基本没有影响，因此可以忽略糖酸比对红葡萄酒质量等级的影响。

3.7 双变量分析小结

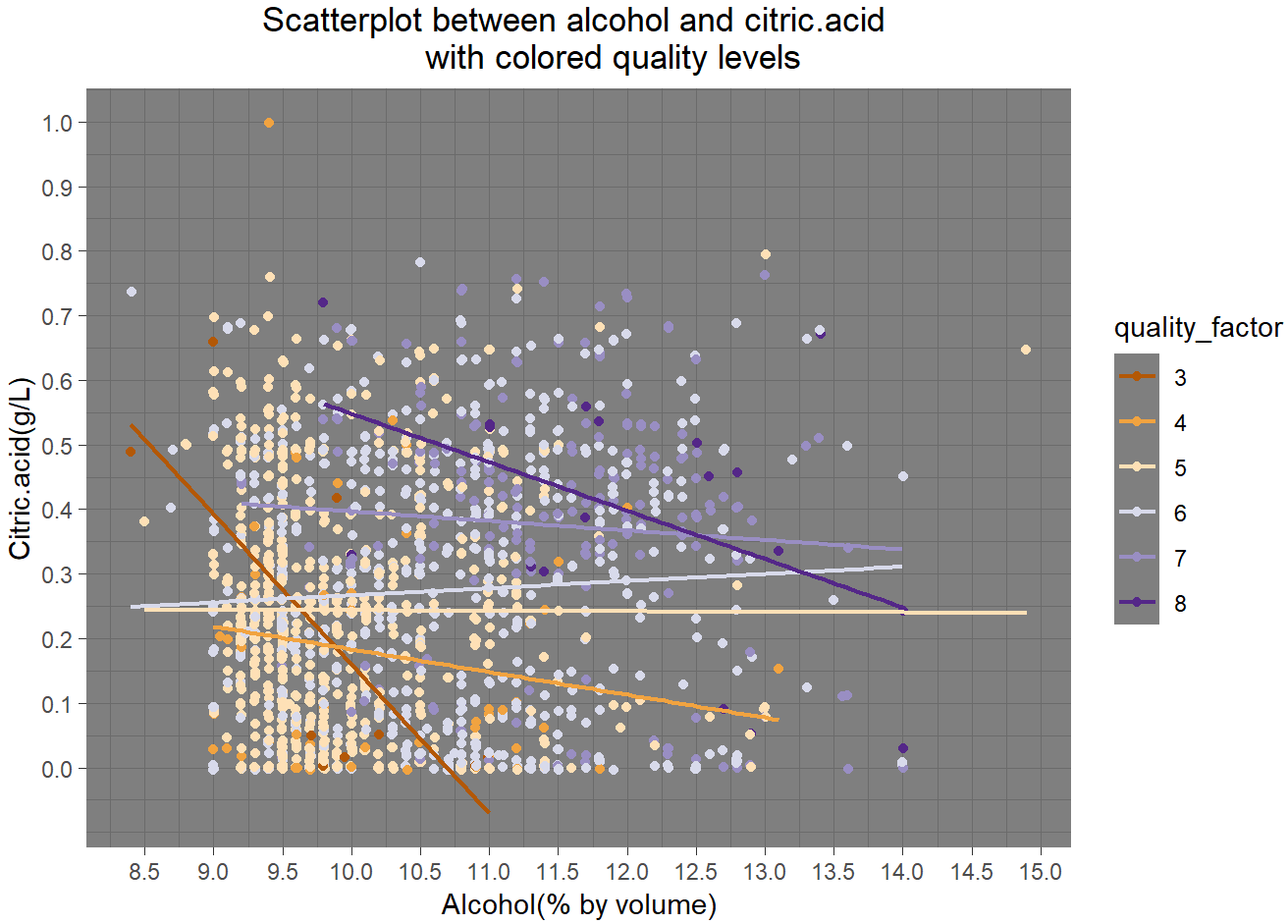
- 在单变量分析的基础上，分别针对酒精度-质量、非挥发性酸-质量、挥发性酸-质量、柠檬酸-质量、残留糖分-质量、糖酸比-质量进行了双变量分析。
- 其中，酒精度对质量有明显的正相关性，柠檬酸对质量有中等的正相关性。
- 其次，挥发性酸对质量有明显的负相关性。
- 此外，剩下的非挥发性酸、残留糖分、糖酸比和质量的相关性非常弱，可以忽略其影响。

4. Multivariate Plots Section 多变量分析

根据双变量分析时得出的结论，在进行多变量分析，进一步对变量进行筛选，将主要考虑酒精度-柠檬酸、酒精度-挥发性酸、柠檬酸-挥发性酸分别对质量的影响。

4.1 酒精度-柠檬酸-质量

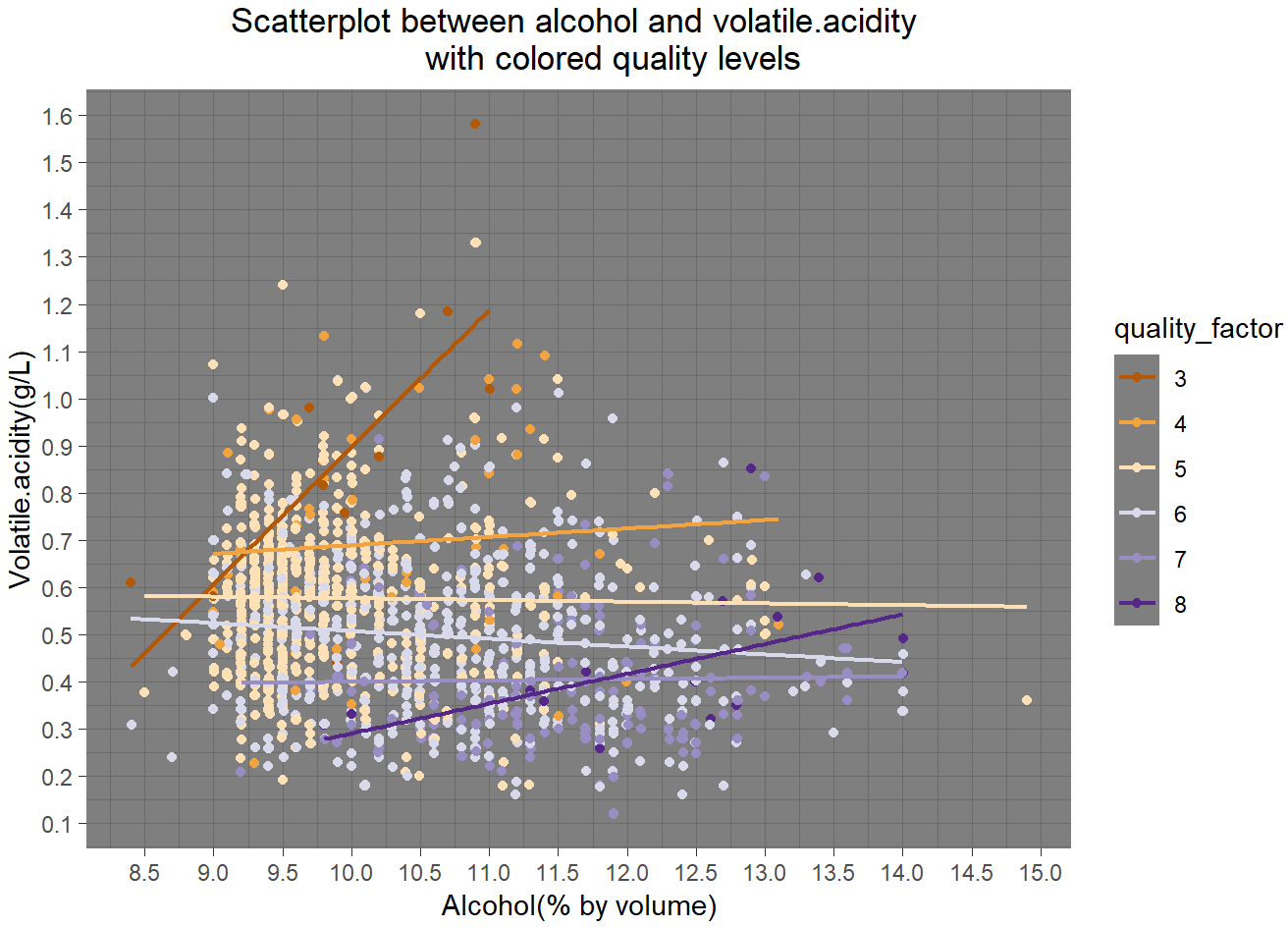
根据质量分类，绘制酒精度-柠檬酸散点图



通过观察，并没有发现这三个变量之间具有明显的关联性。 但是在一定范围类，在同一酒精度的情况下，柠檬酸含量越高红葡萄酒的质量等级越高。 同理，在一定范围类，在同一柠檬酸含量的情况下，酒精度越高红葡萄酒的质量等级越高。

4.2 酒精度-挥发性酸-质量

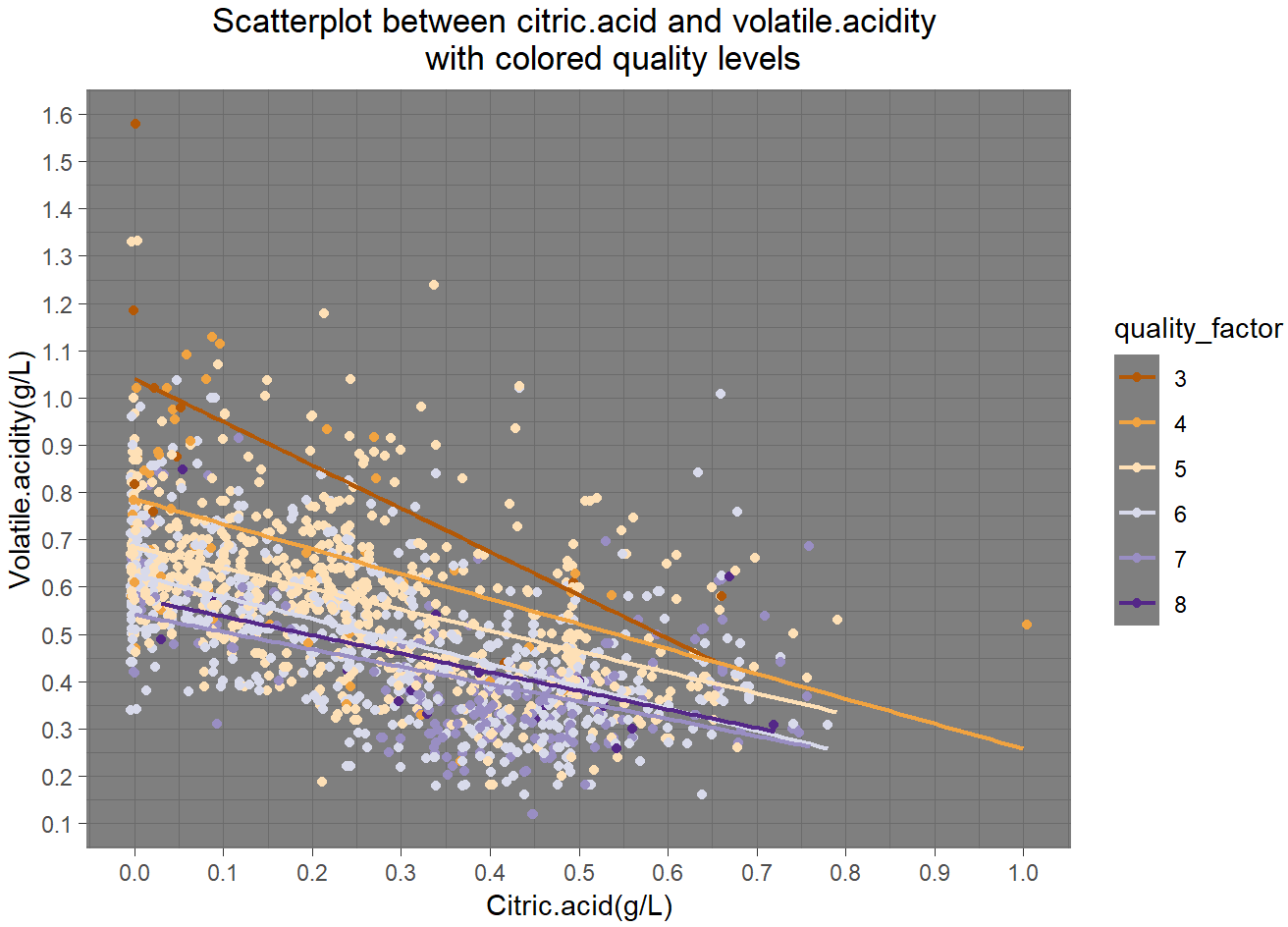
根据质量分类，绘制酒精度-挥发性酸散点图



通过观察，并没有发现这三个变量之间具有明显的关联性。 但是在一定范围类，在同一酒精度的情况下，挥发性酸含量越低红葡萄酒的质量等级越高。 同理，在一定范围类，在同一挥发性酸含量下，酒精度越高红葡萄酒的质量等级越高。

4.3 柠檬度-挥发性酸-质量

根据质量分类，绘制柠檬酸-挥发性酸散点图



通过观察，并没有发现这三个变量之间具有明显的关联性。 但是在一定范围类，在同一柠檬酸含量下，挥发性酸含量越低红葡萄酒的质量等级越高。

4.4 多变量分析小结

- 三个变量之间的关联性不大。
- 三个变量两两组合后（酒精度-柠檬酸、酒精度-挥发性酸、柠檬酸-挥发性酸）分别作用于质量，并没有发现很大影响。
- 只有在三个变量：酒精度、柠檬酸、挥发性酸固定其中任意一个，分析另外一个变量对质量的影响时，才会发现有比较明显的影响，其影响的正反性和强度与双变量分析时保持一致。

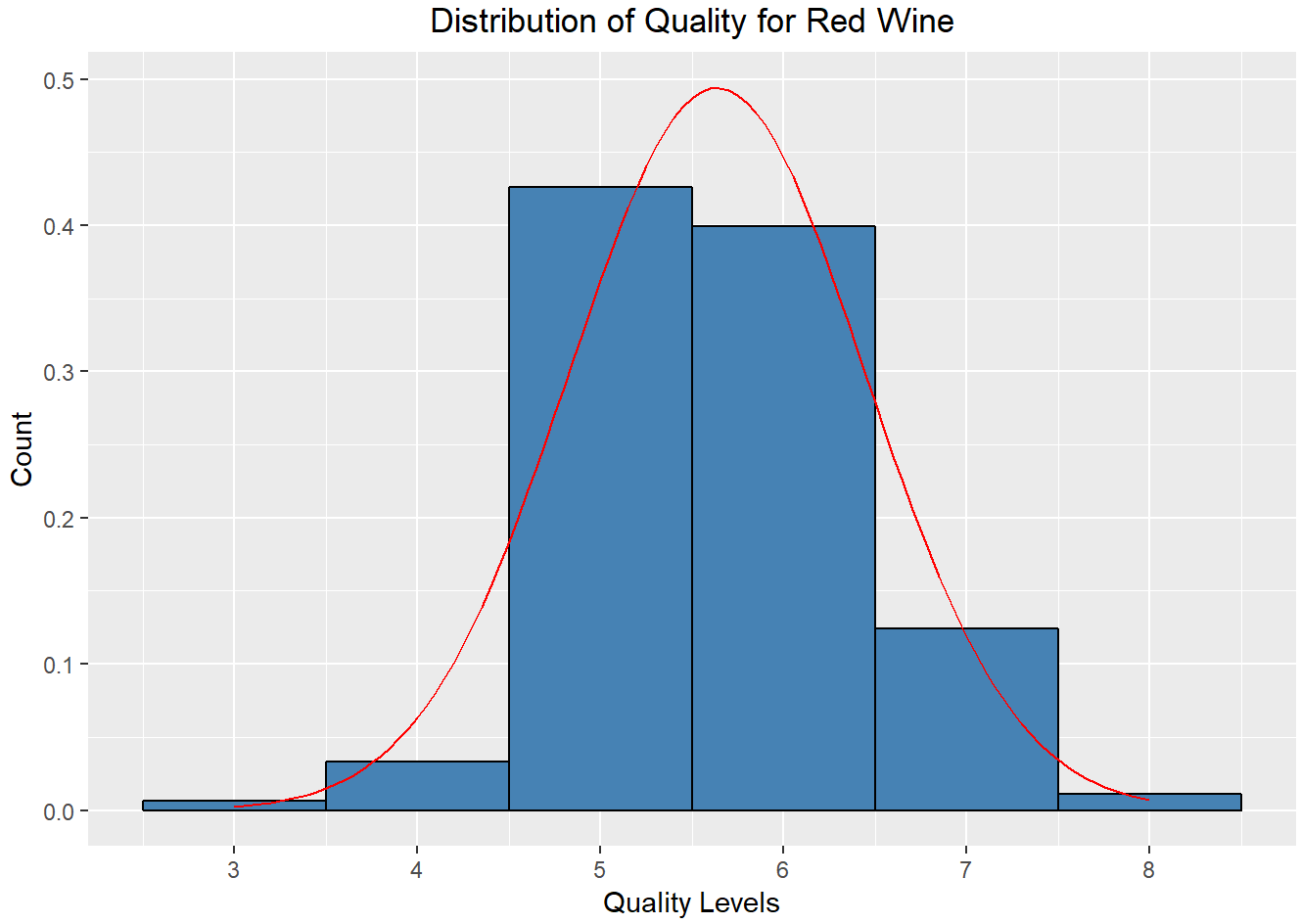
---

5 Final Plots and Summary 最终图形和总结

- 由于本项目的主要目的是为了分析红葡萄酒的质量影响因素，因此在该数据集中只分析了其他变量对质量的影响，而没有分析其他变量之间的相互关联性。
- 该数据集除去变量X和quality还剩下11个变量，通过单变量分析、双变量分析以及多变量分析，层层递进，逐步排除非主要影响变量，最终发现只有挥发性酸volatile.acidity、柠檬酸citric.acid、酒精度alcohol三个变量对质量具有比较明显的影响。
- 其中，挥发性酸对质量具有反向的中等影响，酒精度对质量具有正向的中等影响，柠檬酸对质量具有正向的较弱影响。
- 总的来说，只要在有效范围内，重点控制好红葡萄酒中酒精度、挥发性酸和柠檬酸这三个指标的含量，就能生产出品质不错的红葡萄酒。

5.1 Plot One 红葡萄酒的质量等级分布图

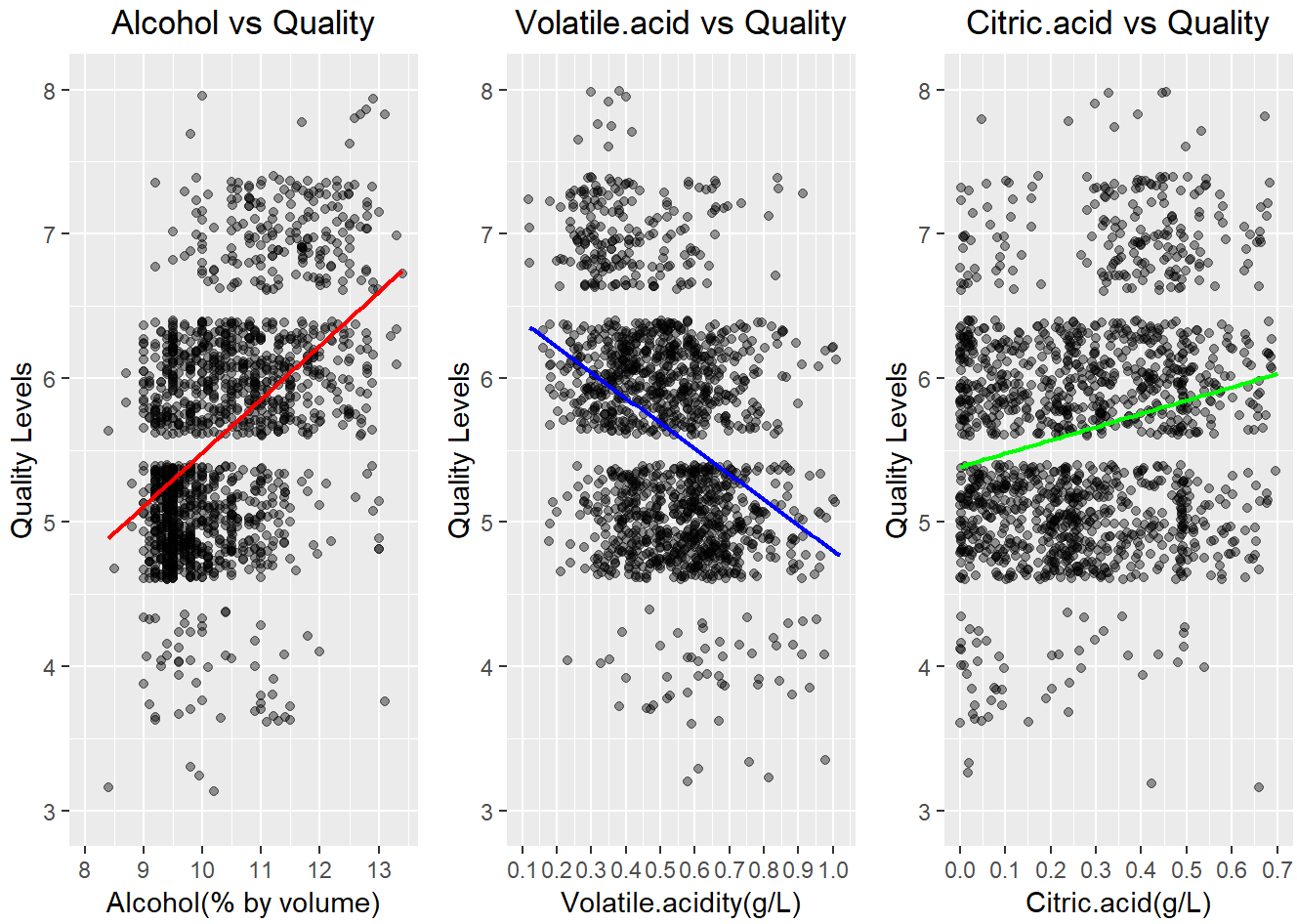
说明：在质量等级分布图中增加了一条正态分布曲线。



Description One

- 红葡萄酒的质量等级分布基本呈现正态分布，大部分的红葡萄酒质量等级为5和6，其次是等级7和4，特别低的等级3和特别高的等级8占比非常低。其中，等级5占比最多，大约43%，等级6其次，大约40%，等级7占比排第三，大约14%，剩下的三个等级总和只占3%左右。说明红葡萄酒的品质还是比较集中的，品质特别高和特别低的红葡萄酒非常少。

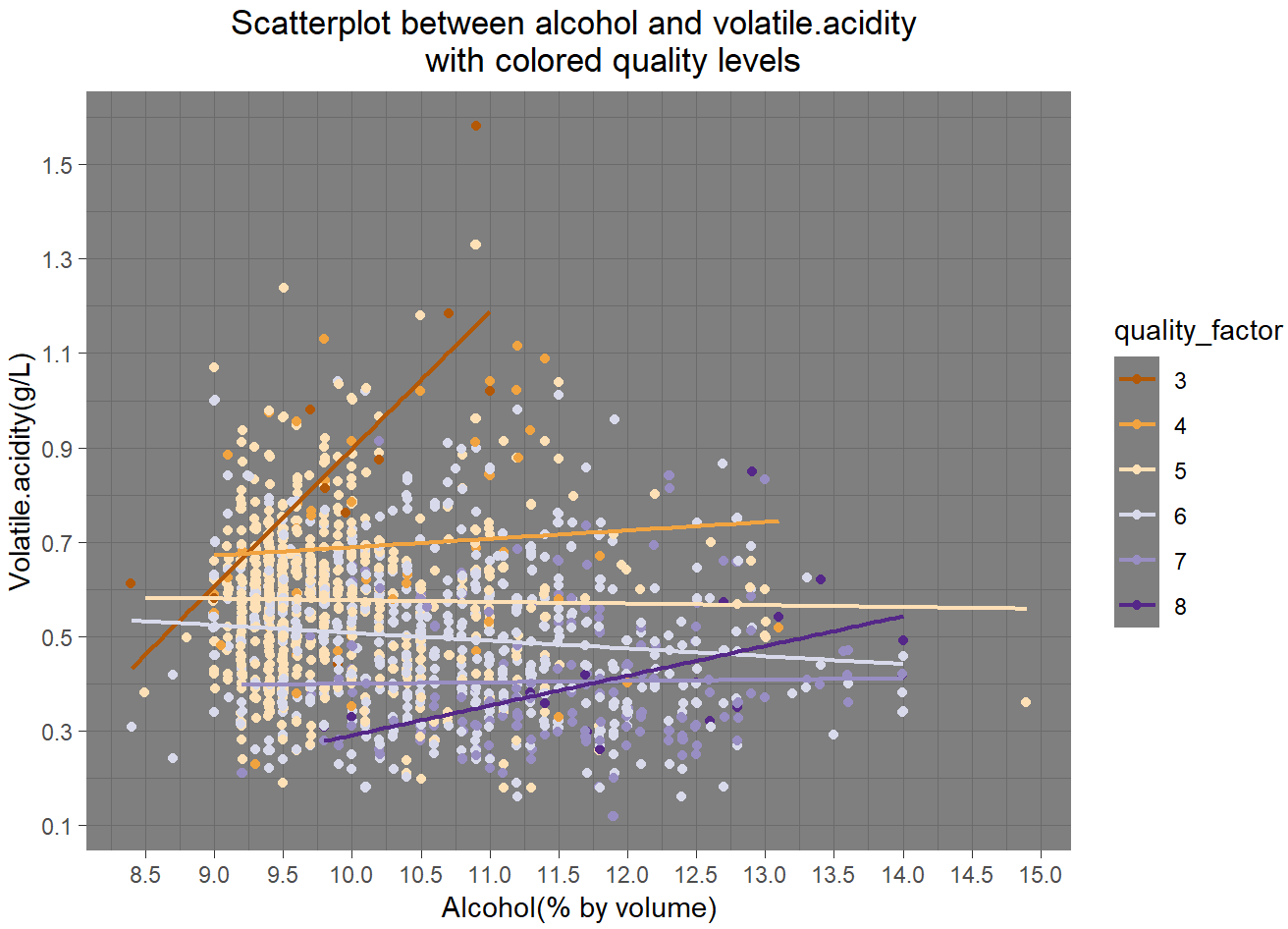
5.2 Plot Two



Description Two

- 酒精度对红葡萄酒质量等级具有正向的中等影响，挥发性酸对红葡萄酒质量等级具有反向的中等影响，柠檬酸对红葡萄酒质量等级具有正向的较弱影响。

5.3 Plot Three



Description Three

- 两个变量的共同作用并没有对质量等级造成很大的影响，反而是单一变量对质量等级的影响比较明显，如图所示的在酒精度保持不变的情况下挥发性酸对红葡萄酒质量等级具明显的反向影响，相反的，在挥发性酸保持不变的情况下酒精度对红葡萄酒质量等级具有明显的正向影响。

6 Reflection 反思

- 在进行本项目时，只考虑了分析其他变量对质量等级的影响，其他变量之间的相互影响可以在后续进行分析。
- 尝试对多变量线性回归进行建模，期望可以得出模型：红葡萄酒质量等级 = X1·酒精度 + X2·挥发性酸 + X3·柠檬酸 + 常数，但是受限于知识水平，目前只会计算单变量模型： y = bx + a 。待后续补充建模知识后，争取能够得到期望的模型。
- 看似容易的一个项目，真正做起来才暴露出很多问题，也明白了为什么开始建议的时间估计达20小时之多。