哪个化学成分影响白葡萄酒的质量?

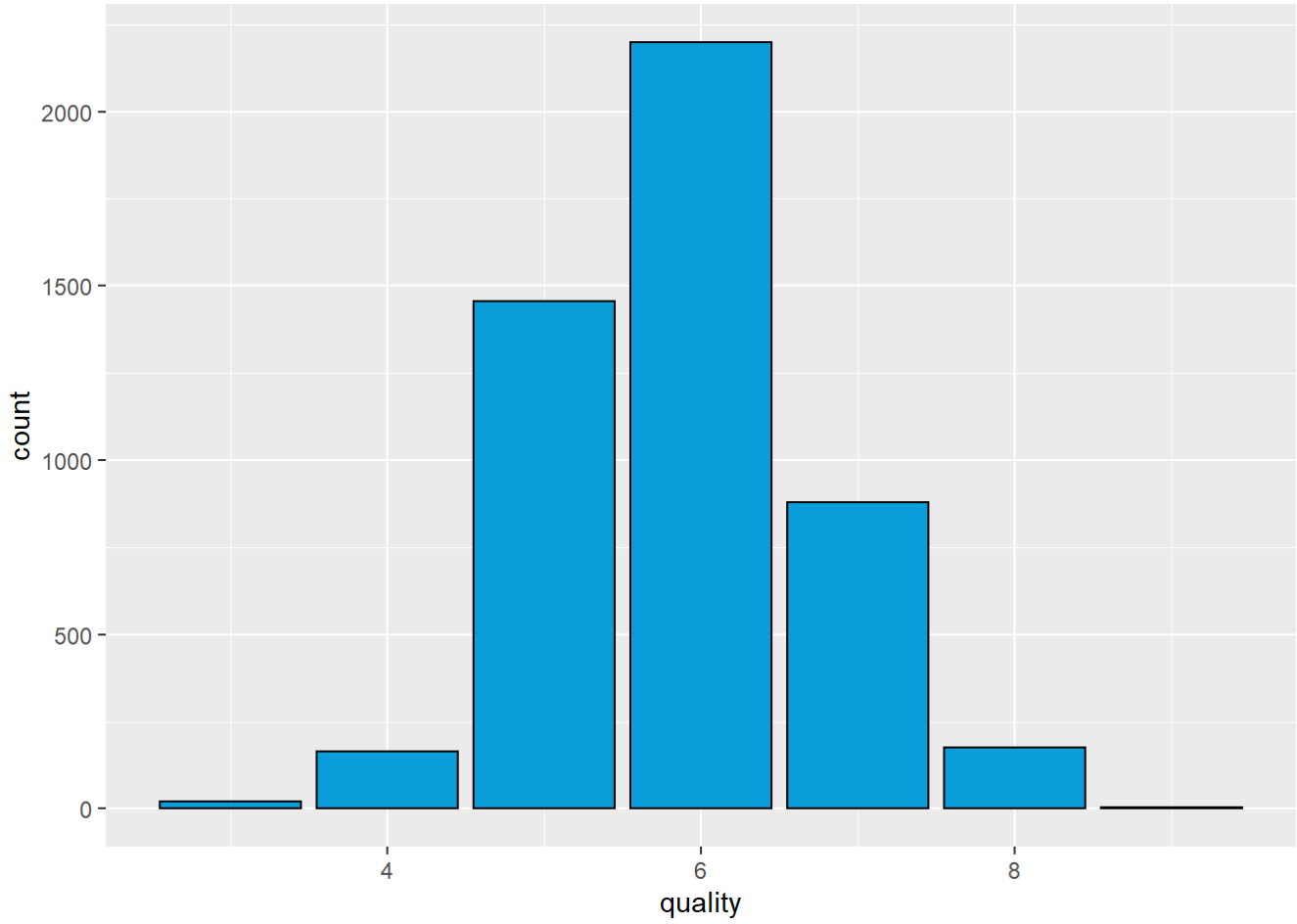
Image Name

==========

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

单变量绘图选择

quality特征



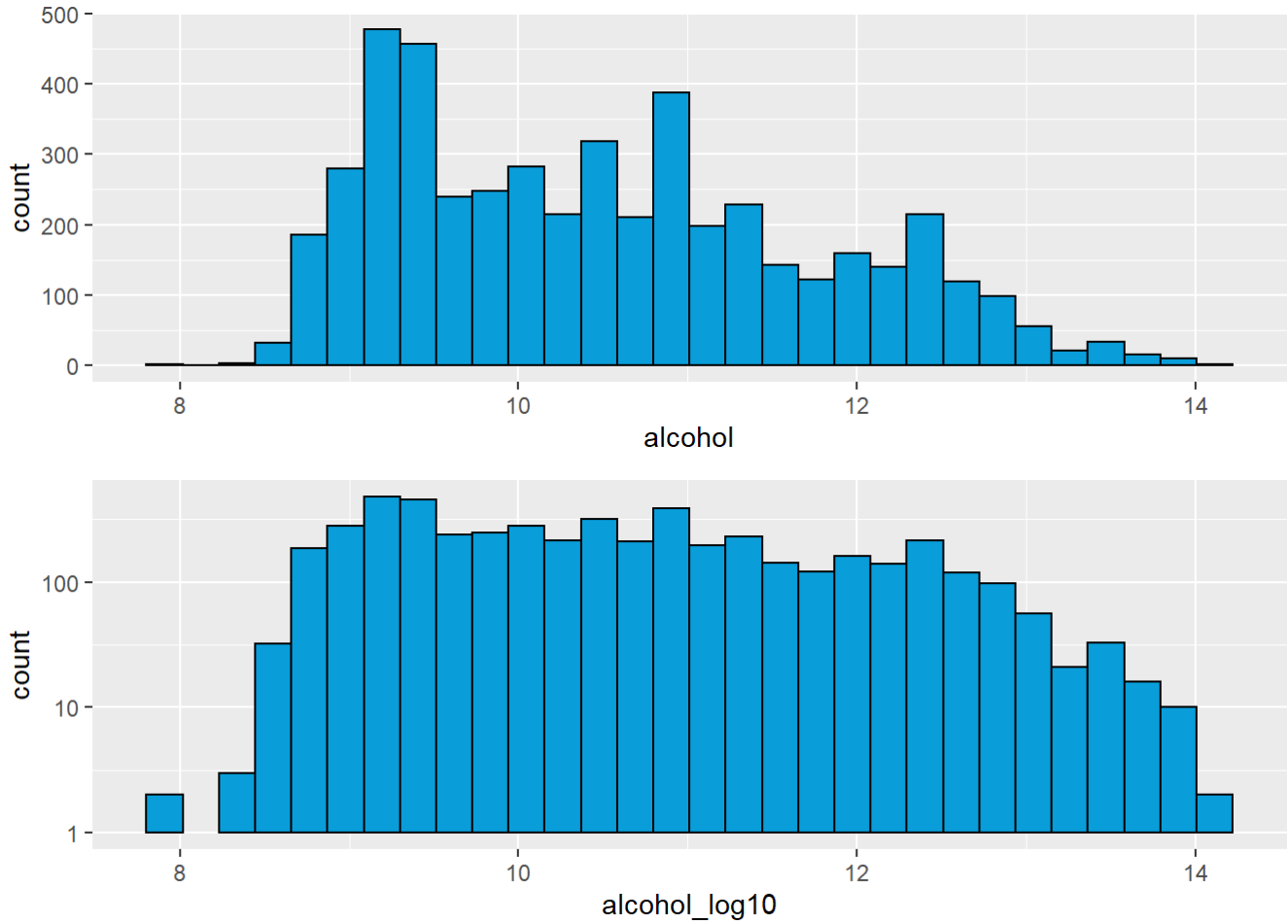
quality 变量也就是白葡萄酒的质量。 从质量中可以看出 6 分的质量的数量最多的。 9 分的极高质量的酒的数量最少!

alcohol特征

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

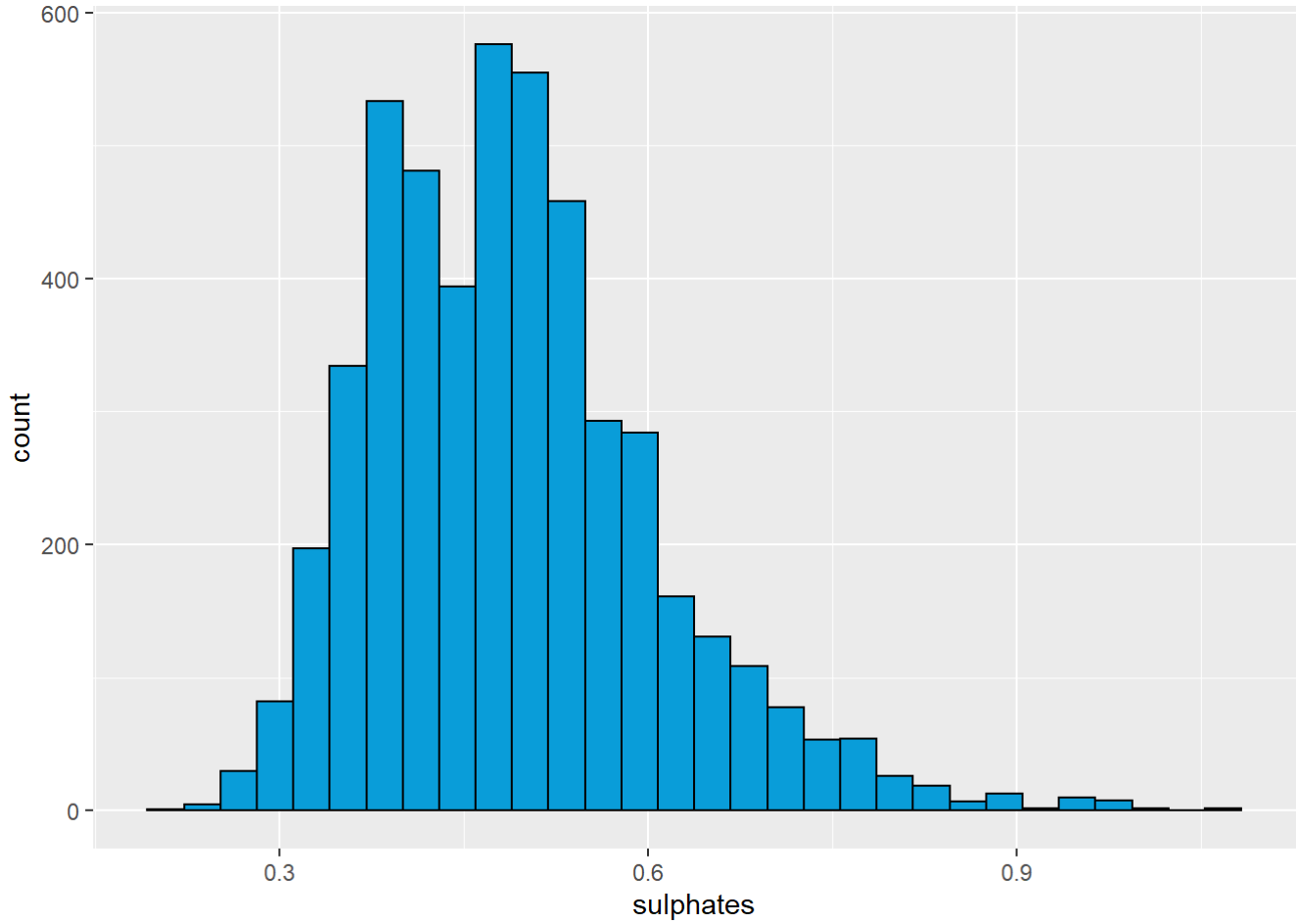
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



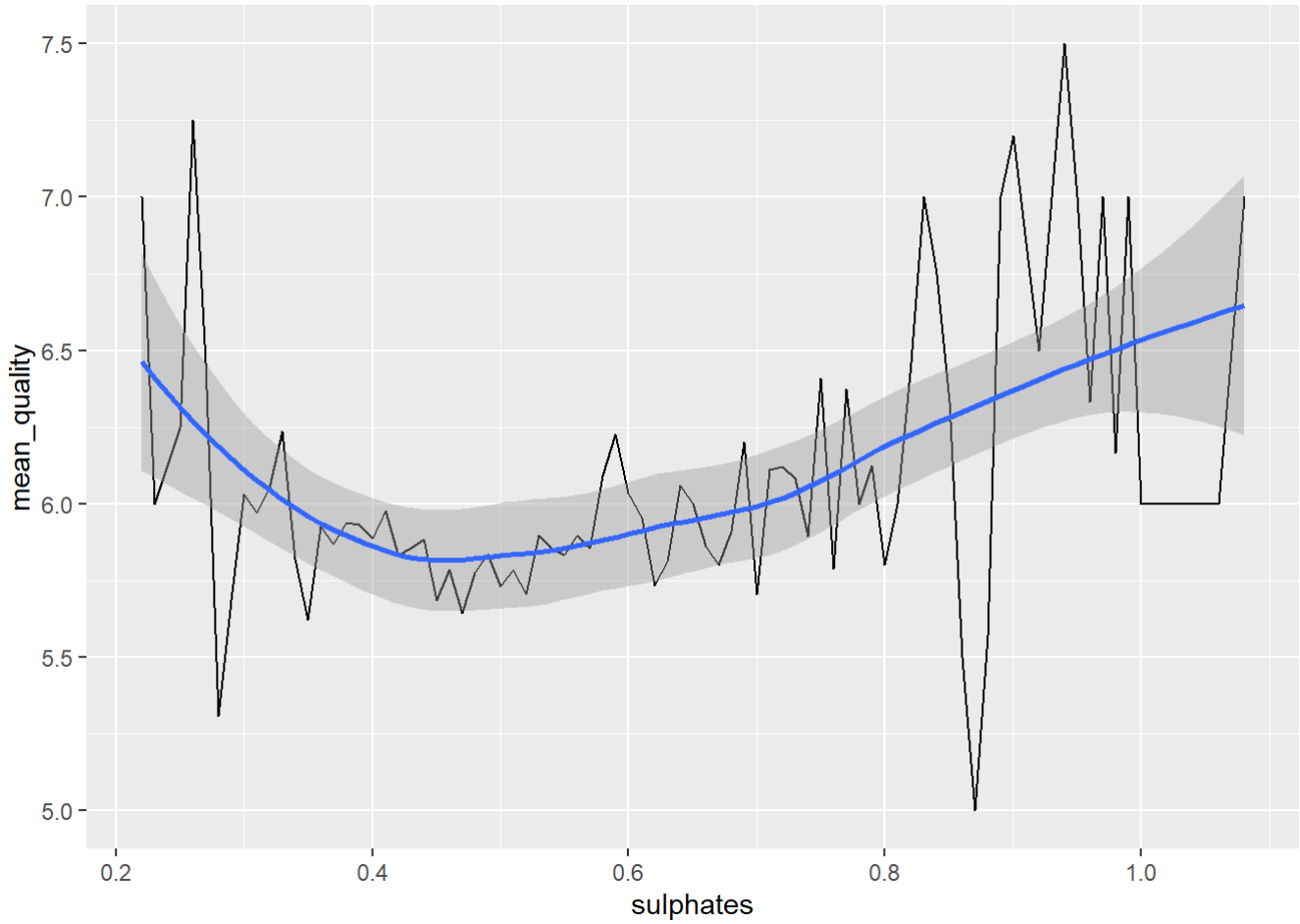
alcohol 也就是酒精 酒精的数量的原始分布有点参差不齐，我们对y轴log10一下，得到了下图。 从9~13之间大部分的酒精的数量是近似的。 也就是说大部分的白葡萄的放入的酒精的百分比是接近的。

sulphates特征

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



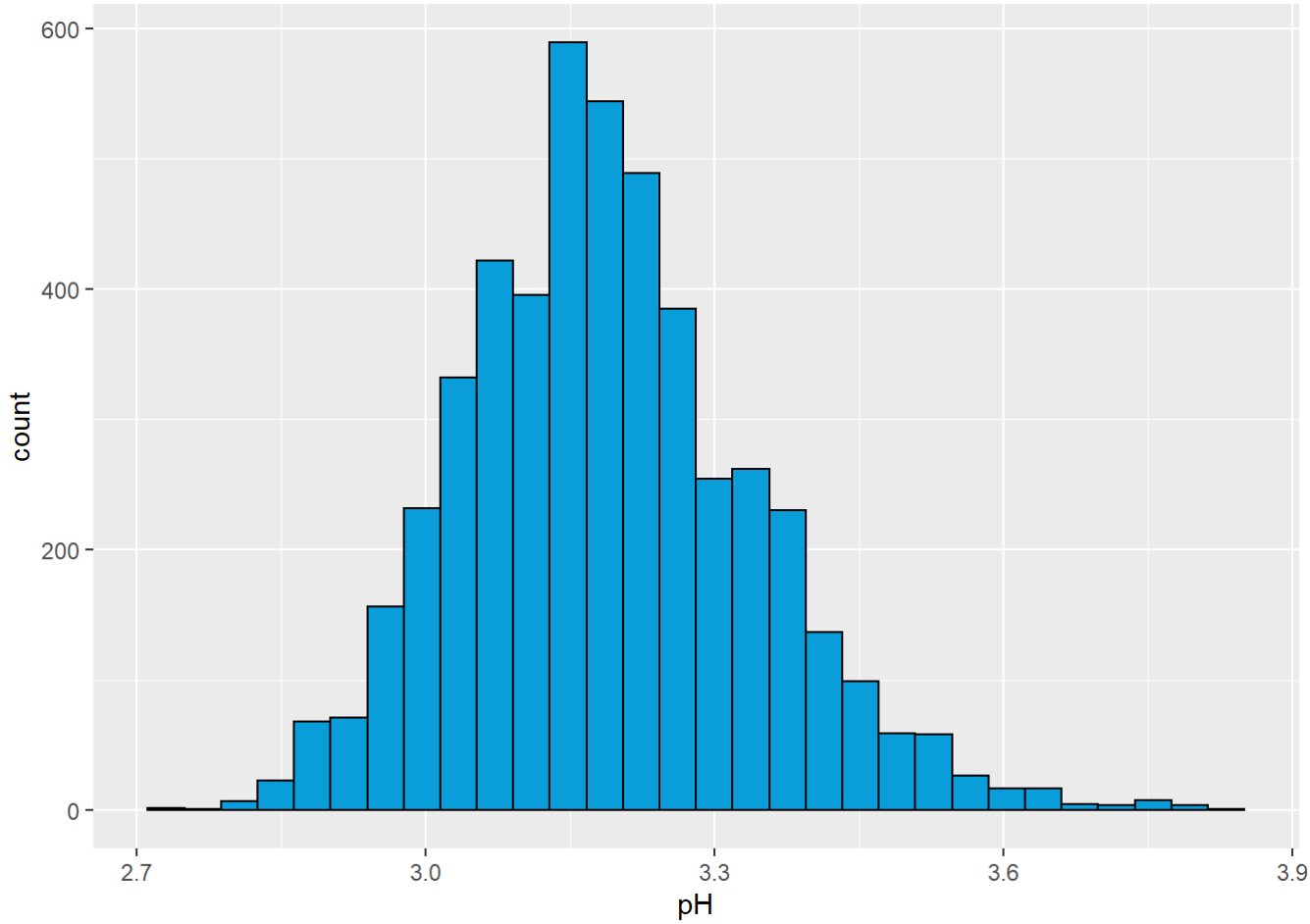

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



sulphates 就是硫酸盐 创建了新的变量 mean_quality （不同硫酸盐的平均分） 画出图形可以发现，硫酸盐平均质量分数是“微笑曲线”。 **0.2$_{0.4单调递减}$ 0.7最大值**单调递增

PH特征

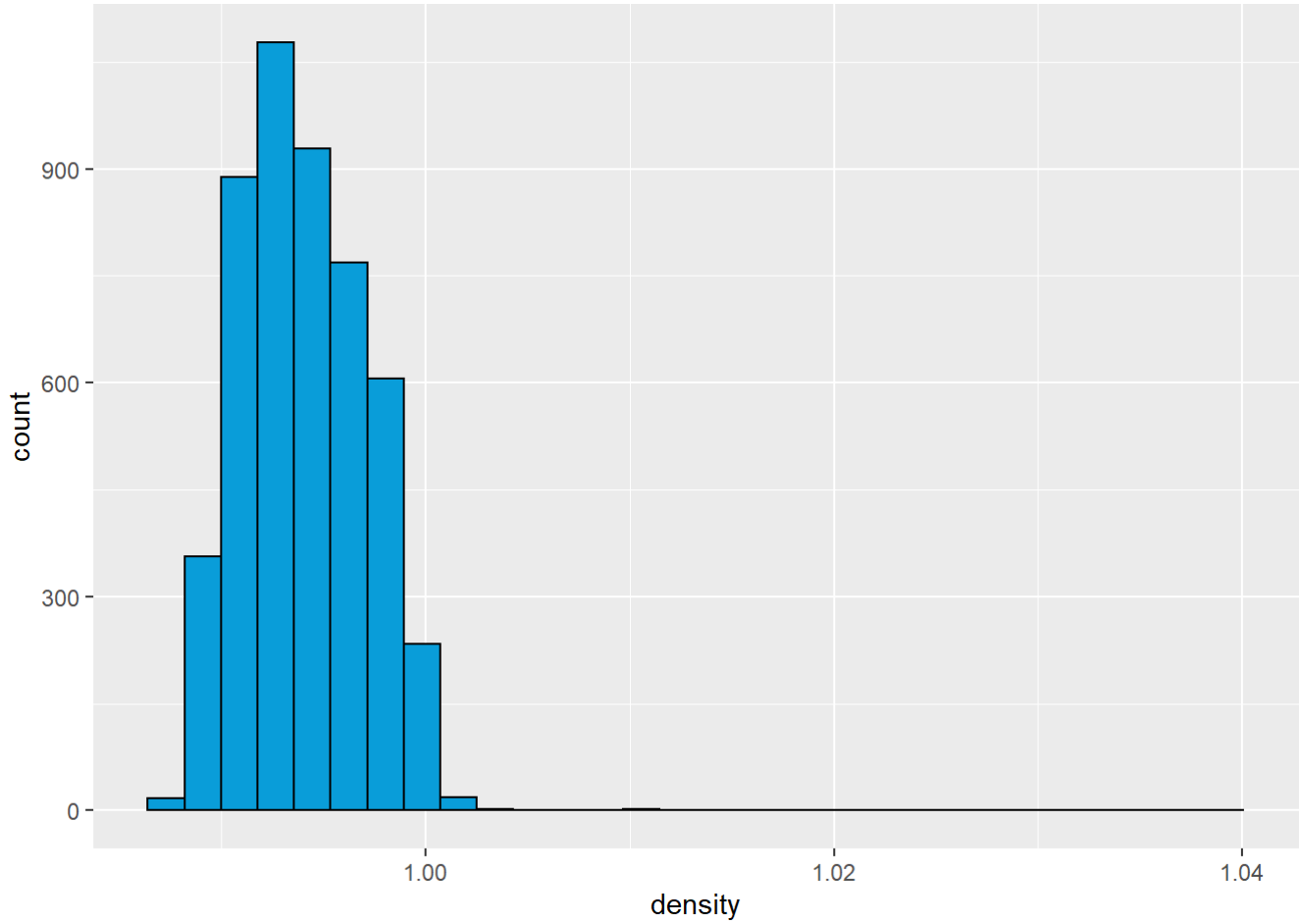
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



pH值的整体分布呈现较为理想的正态分布

density特征

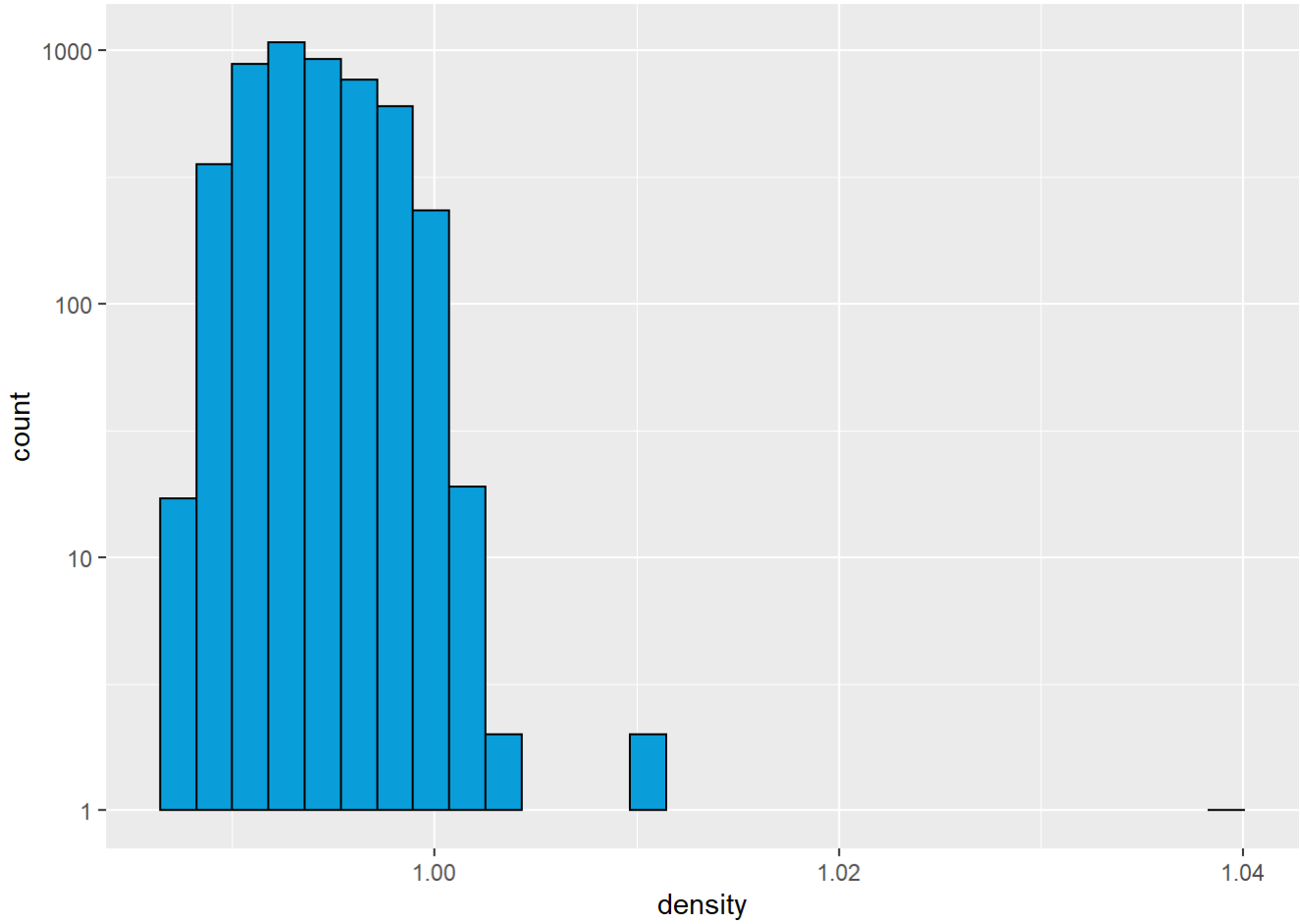
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

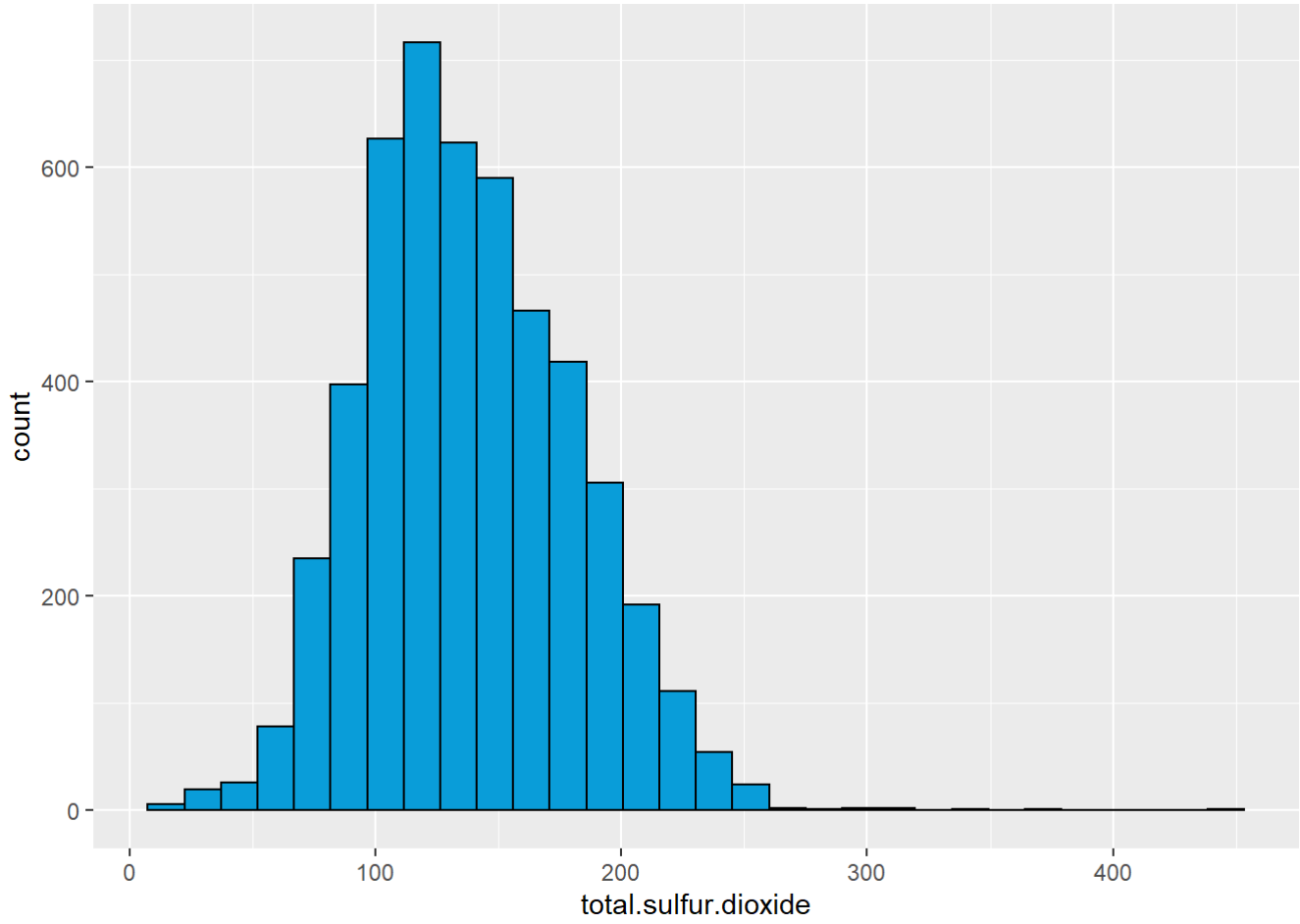
```
## Warning: Removed 18 rows containing missing values (geom_bar).
```



对密度这个特征的y轴log10转换。
密度分布呈现偏左分布，另外大部分的酒的密度是低于水的密度的。
查阅资料得知，酒精的密度是低于水的，那么我猜测当酒里面的酒精越多，
是不是密度越小。

total.sulfur.dioxide特征

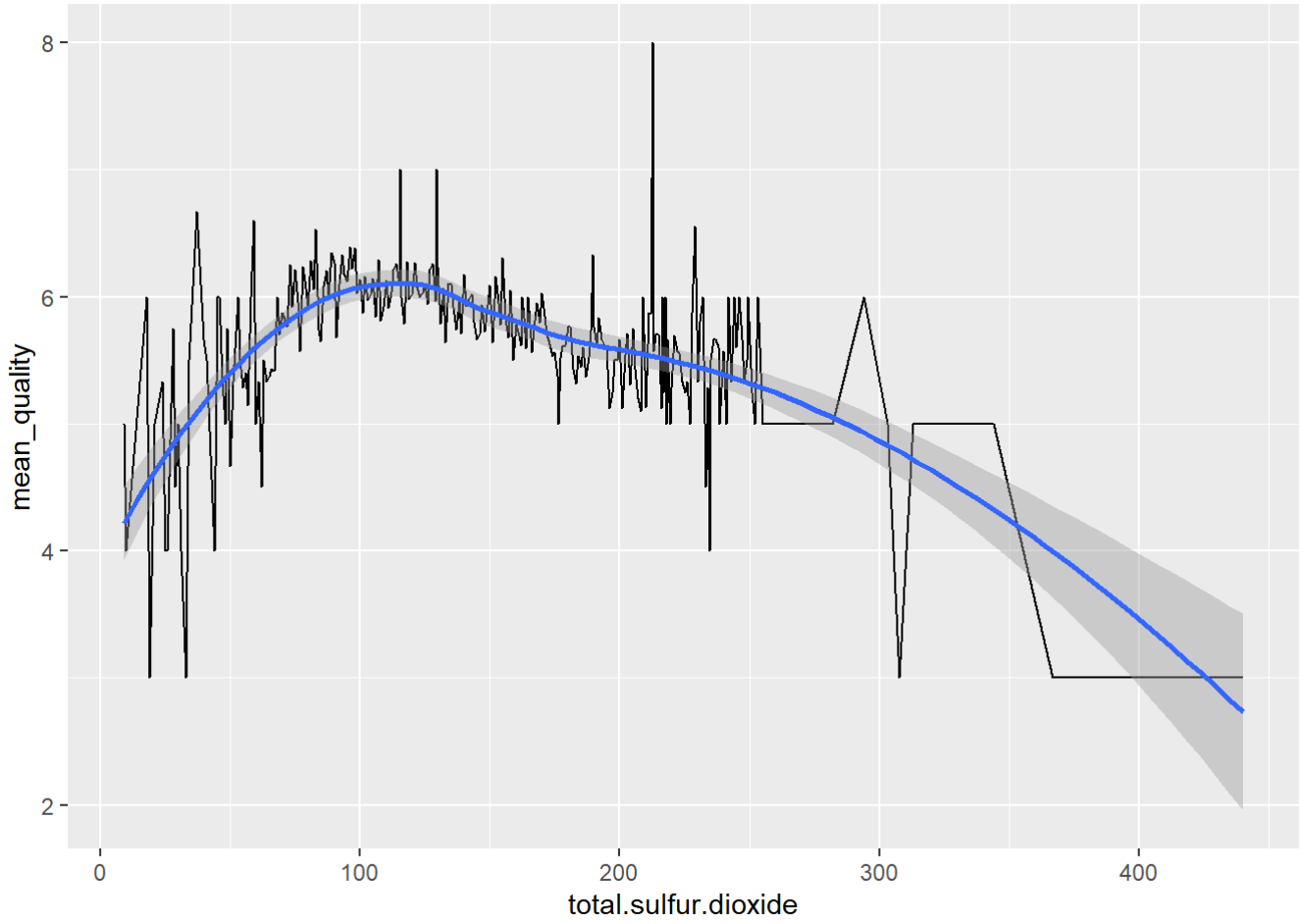
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



总二氧化硫。 通过查看数据文档得知，这种化学元素放多了会气味。 那么猜测是不是这总元素放的越多，气味越大。

考虑总二氧化硫和quality之间有没有什么关系，我绘制了它们之间的关系图

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



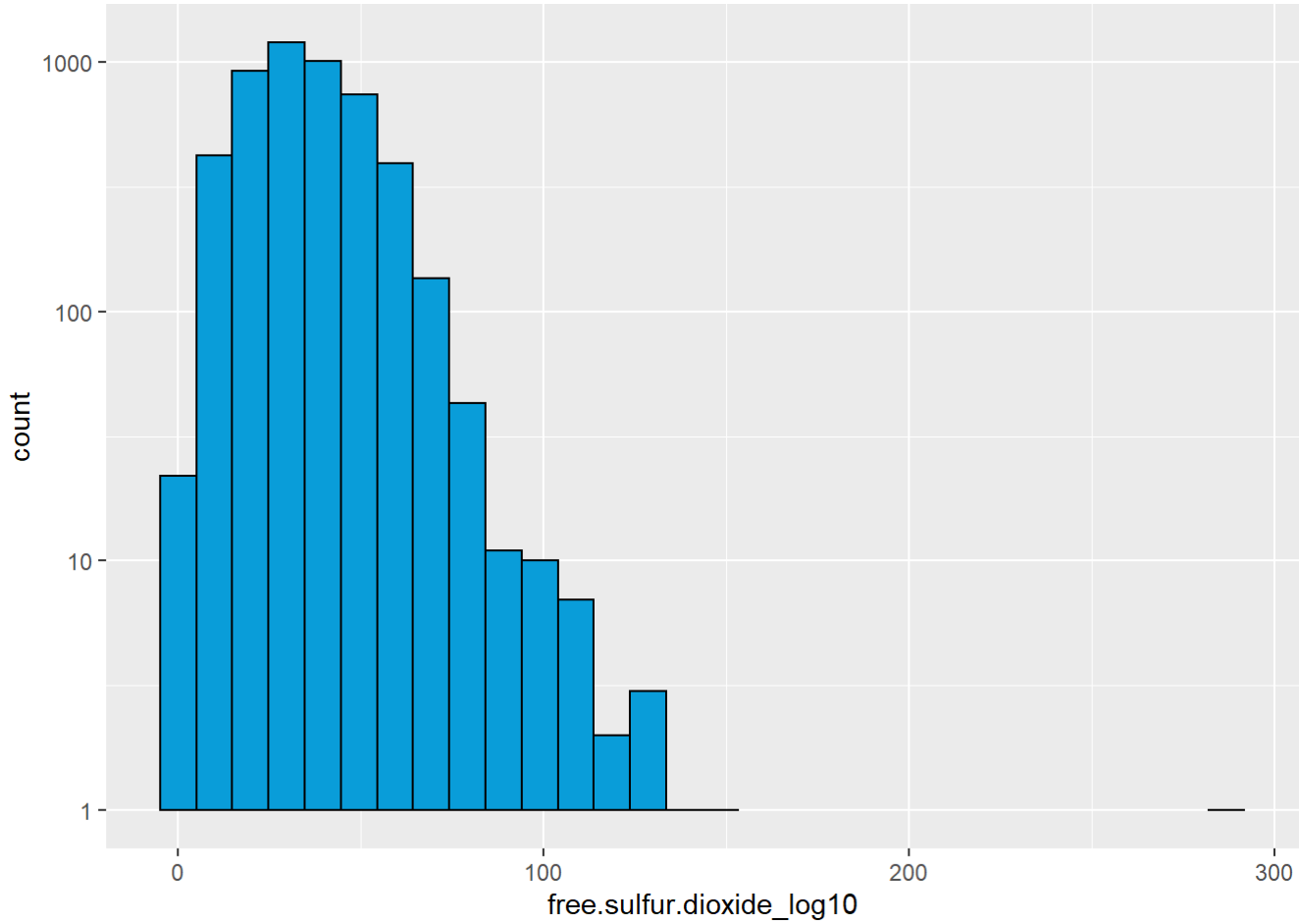
可以上图中看出，随着二氧化硫放的越多，气味越大，然后酒的质量得分也越来越低。

free.sulfur.dioxide

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 13 rows containing missing values (geom_bar).
```

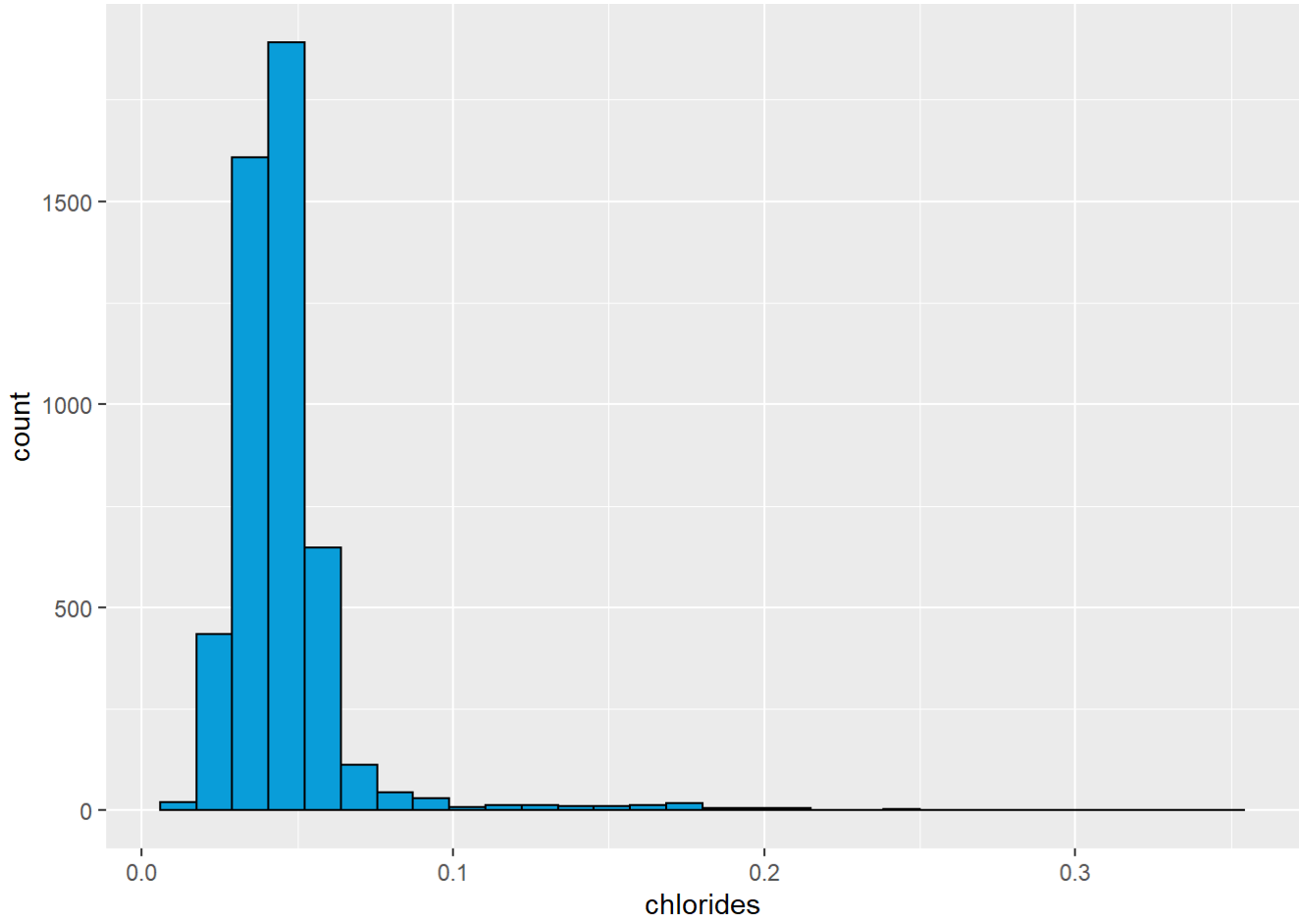


游离二氧化硫。整体呈左偏。

chlorides

```
ggplot(aes(x = chlorides), data = wineData) +
    geom_histogram(color = I("black"), fill = I("#099DD9"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



氯化物，整体数值偏向左边

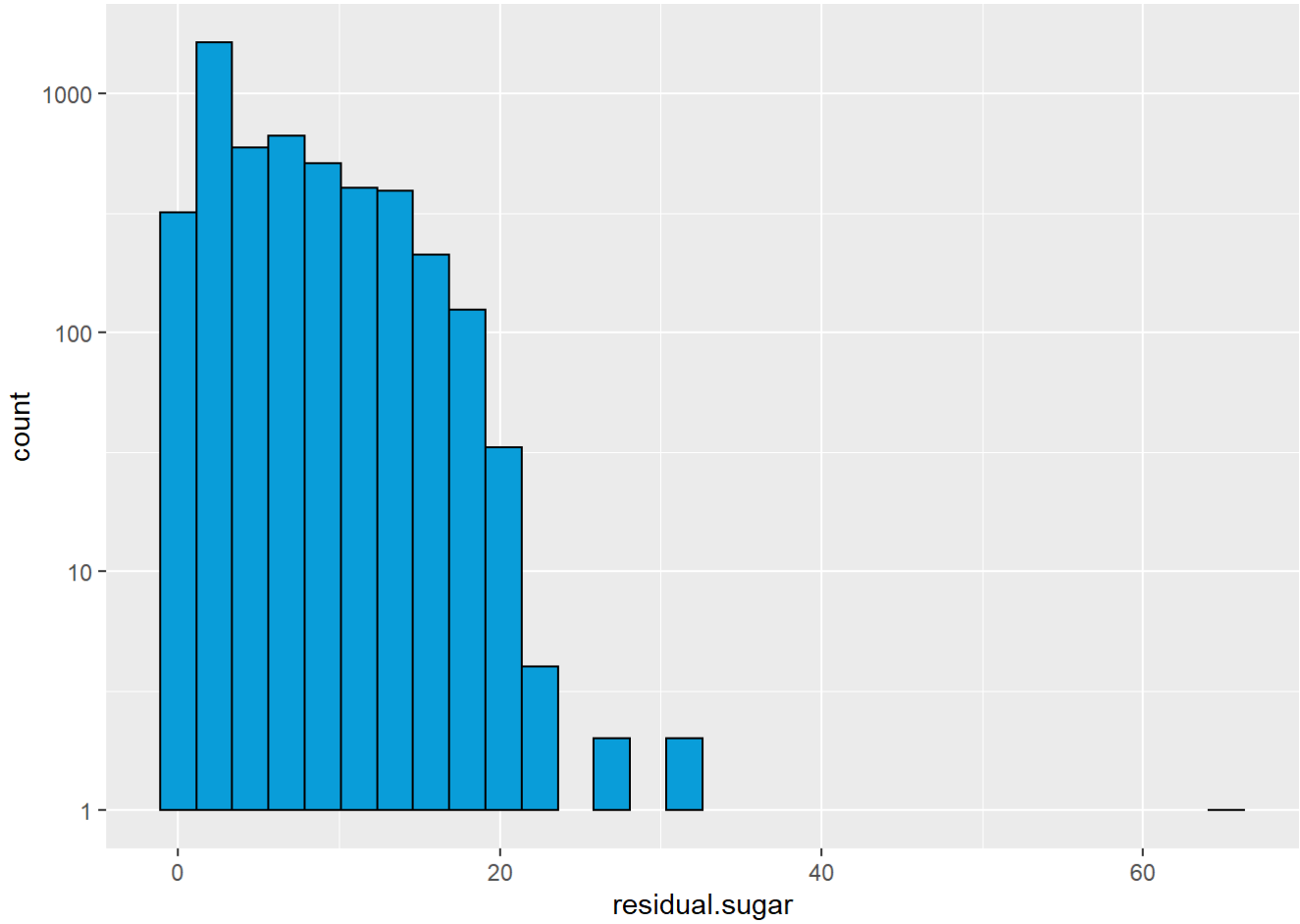
residual.sugar

```
ggplot(aes(x = residual.sugar), data = wineData) +
    geom_histogram(color = I("black"), fill = I("#099DD9")) +
    scale_y_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 16 rows containing missing values (geom_bar).
```

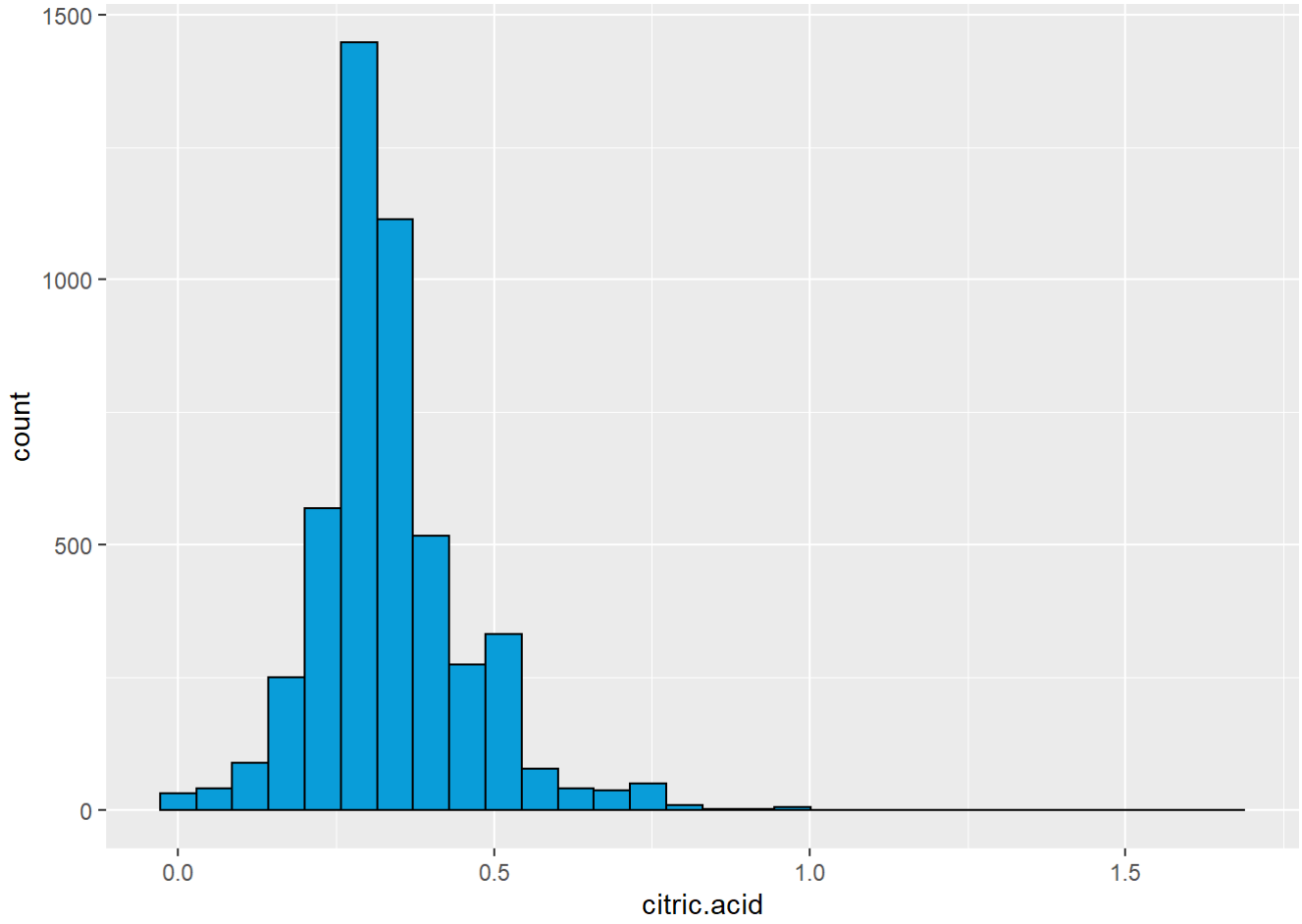


残糖，整体分布有点均匀，峰值偏向左边

citric.acid

```
ggplot(aes(x = citric.acid), data = wineData) +
    geom_histogram(color = I("black"), fill = I("#099DD9"))
```

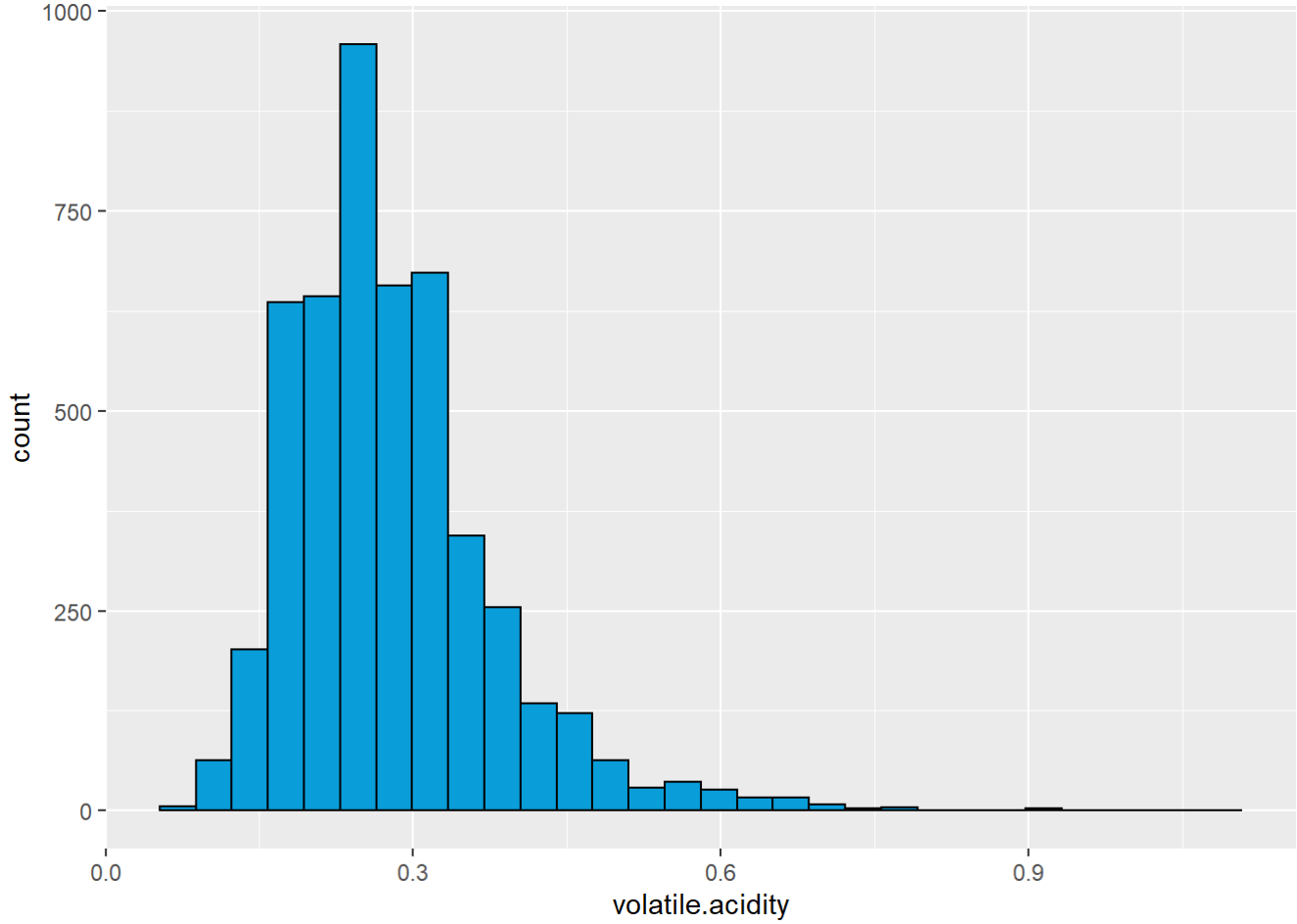
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



柠檬酸，整体较为对称，但是峰值还是偏向左边

volatile acidity

```
ggplot(aes(x = volatile.acidity), data = wineData) +
    geom_histogram(color = I("black"), fill = I("#099DD9"))
```



挥发性酸度。整体较为对称，但是峰值还是偏向左边

单变量分析

你的数据集结构是什么?

```
str(wineData)
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

化学成分特征:

alcohol 酒精

sulphates 硫酸盐

total.sulfur.dioxide 总二氧化硫（总 = 结合 + 游离）

free.sulfur.dioxide 游离二氧化硫

chlorides 氯化物

residual sugar 残糖（发酵过程中剩余的糖。百科中指出有一定的残糖酒的质量才好）

citric acid 柠檬酸（少量的这个玩意可以让人有“新鲜感”）

指标型特征：

fixed acidity 固定酸度

volatile acidity 挥发酸度

density 密度

pH

quality 质量指标（也是目标变量）

你的数据集内感兴趣的主要特性有哪些？

残糖。据说残糖可以提高酒的质量。 二氧化硫。过量的二氧化硫会让酒的质量下降。 酒精含量。到底你爱喝的是酒呢？还是水呢？

你认为数据集内哪些其他特征可以帮助你探索兴趣特点？

密度。是否密度接近水，口感越好。

根据数据集内已有变量，你是否创建了任何新变量？

新的变量“HuaXueYuanSuliang”（总的化学元素量，不包括酒精，因为酒精的单位不同）

```
# 新的变量“HuaXueYuanSuliang
wineData$HuaXueYuanSuliang = wineData$sulphates +
    wineData$total.sulfur.dioxide +
    wineData$chlorides + wineData$residual.sugar + wineData$citric.acid
```

新的变量“结合二氧化硫”（结合二氧化硫 = 总二氧化硫 - 游离二氧化硫）

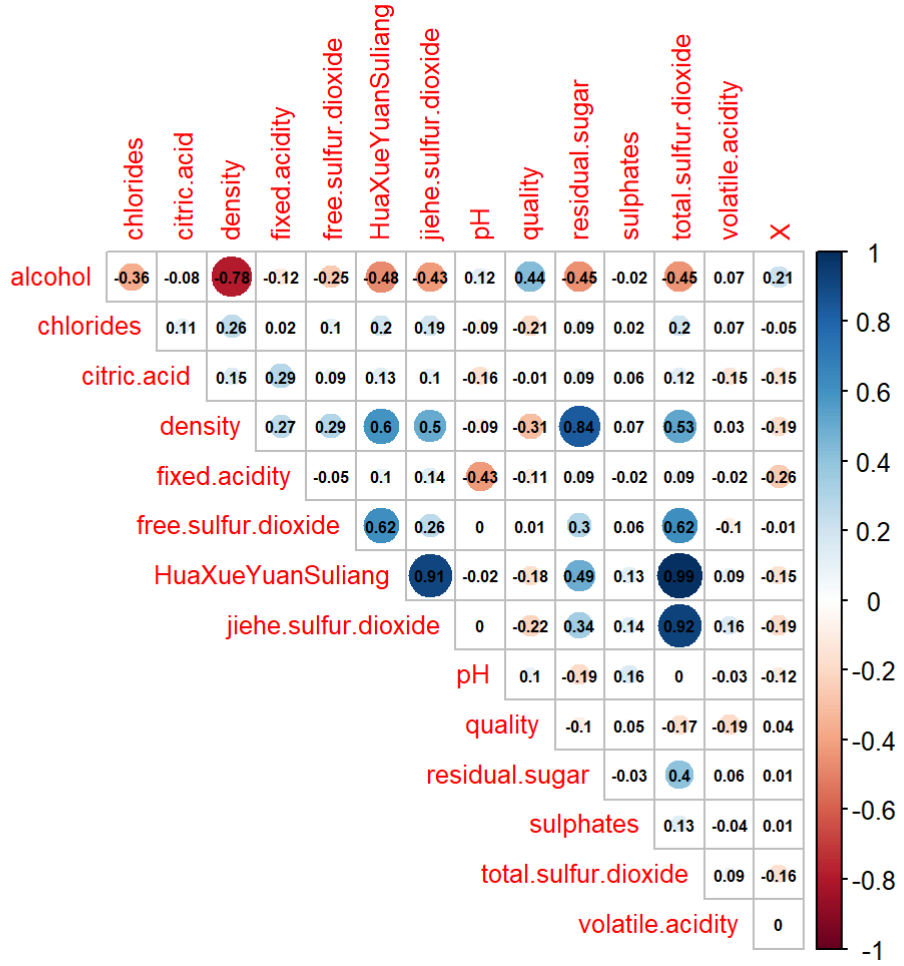
```
# 新的变量“结合二氧化硫”
wineData$jiehe.sulfur.dioxide = wineData$total.sulfur.dioxide - wineData$free.sulfur.dioxide
```

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，你为什么会这样做？

酒精的分布参差不齐，对齐进行log10转化。因为这样可以让数据的尺度缩小，最高的数据不会比最低的数据高太多。

双变量绘图选择

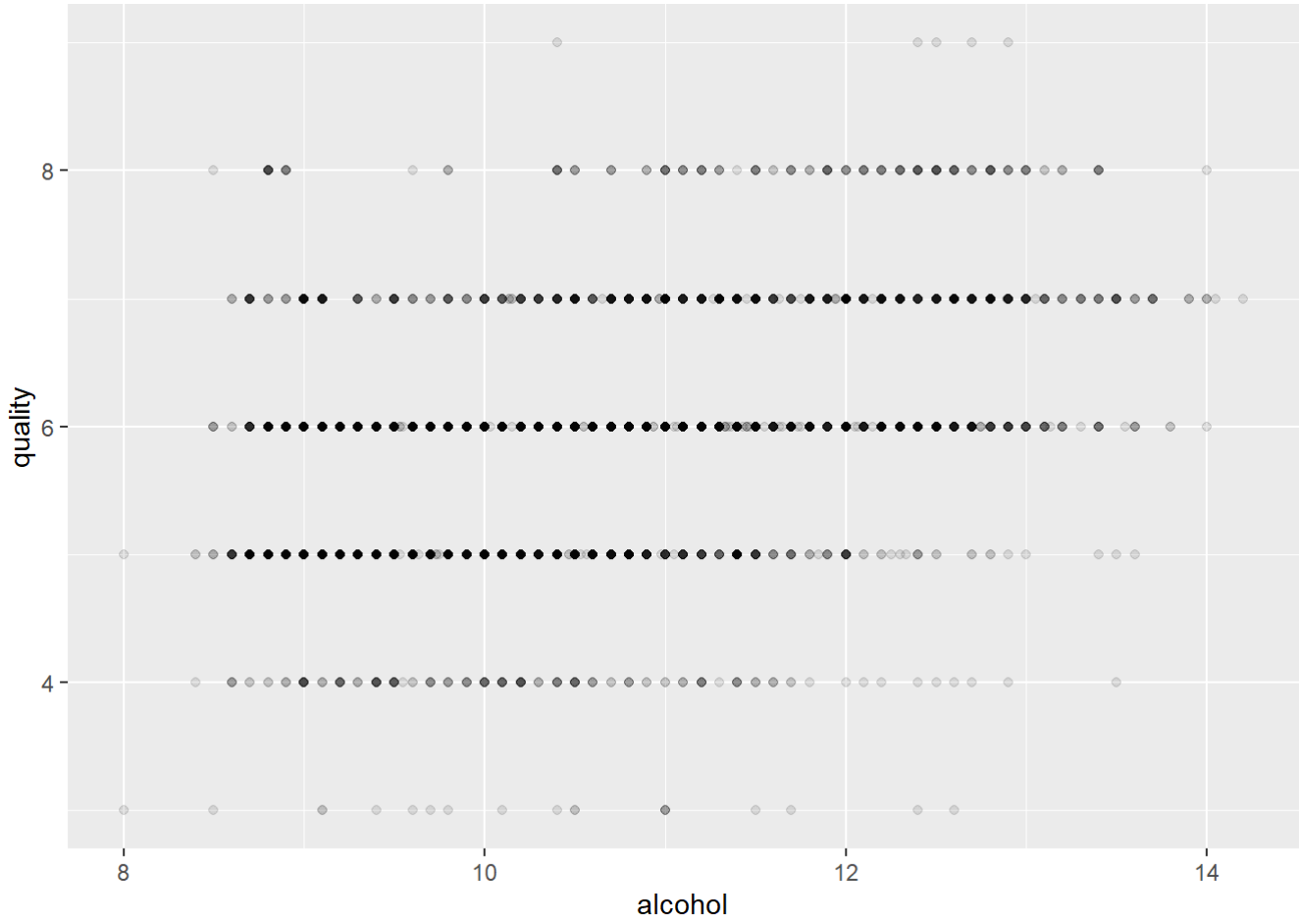
对于每两个变量之间的关系系数，绘制关系矩阵



从矩阵中可以看出，关系比较强烈的几个变量分别是： alcohol和quality、jiehe.sulfur.dioxide和quality、chlorides和qualityh、density和quality

alcohol和quality分析

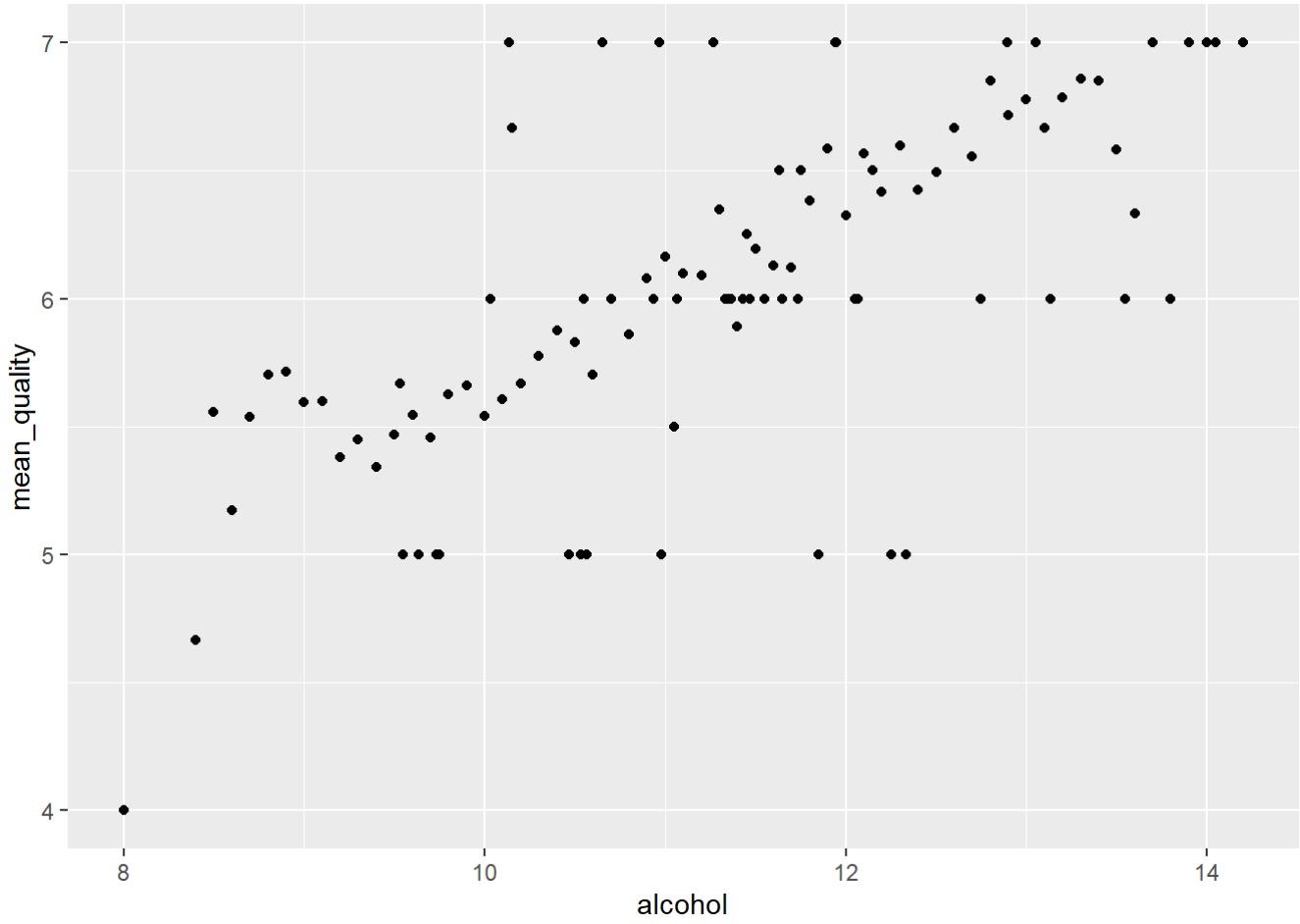
```
# 绘制alcohol vs quality 散点图
ggplot(aes(x = alcohol, y = quality), data = wineData) +
    geom_point(alpha = 10/100)
```



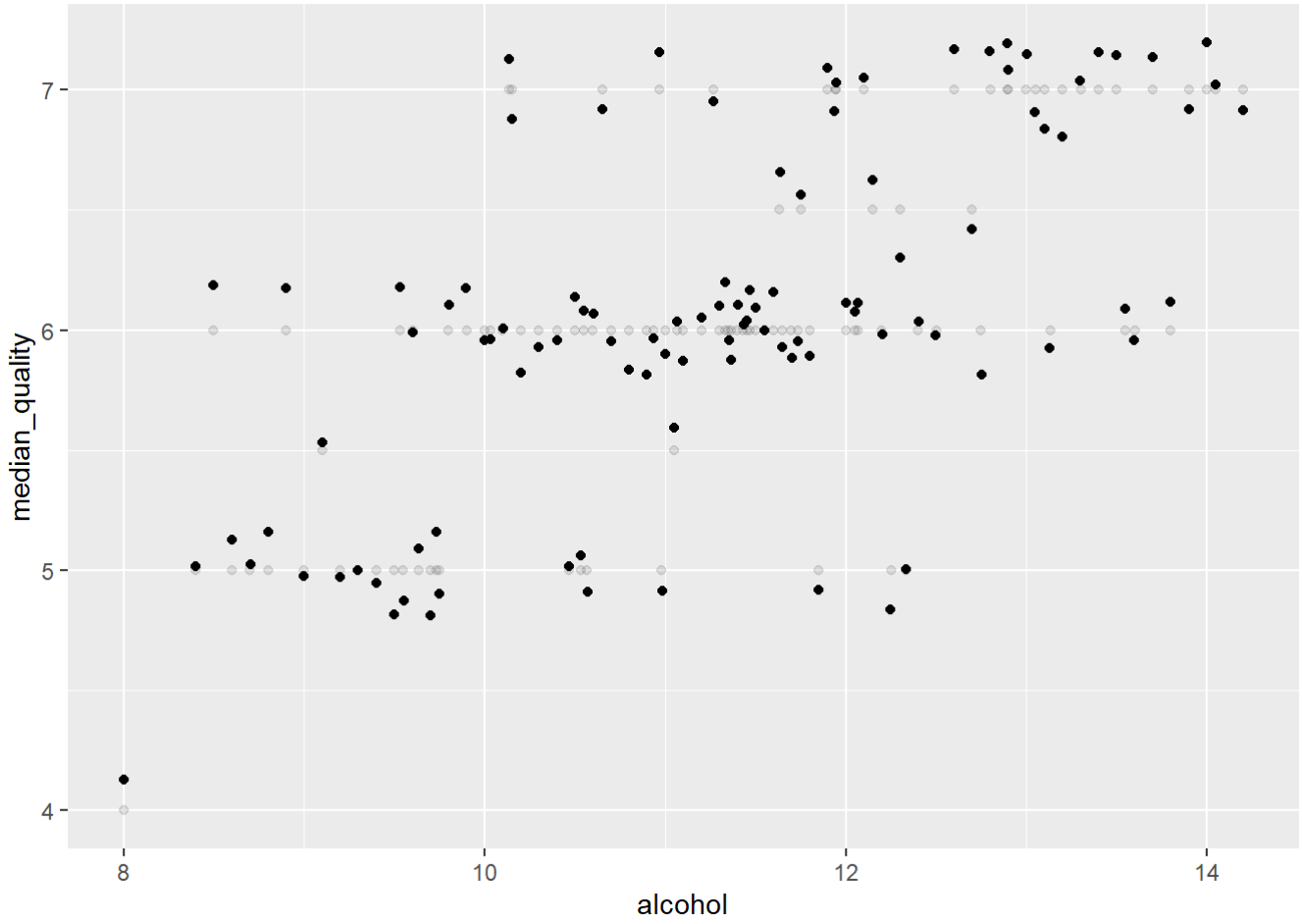

```
# 按照alcohol分组
wineDataByAlcohol = group_by(wineData, alcohol)

# 按照alcohol分组后，求mean_quality和median_quality
wineDataByAlcoholByAlcohol = summarise(wineDataByAlcohol,
                                       mean_quality = mean(quality),
                                        median_quality = median(quality))


ggplot(aes(x = alcohol, y = mean_quality),
       data = wineDataByAlcoholByAlcohol) +
    geom_point(color = I("black"), fill = I("#099DD9"))
```




```
ggplot(aes(x = alcohol, y = median_quality),
       data = wineDataByAlcoholByAlcohol) +
    geom_point(alpha = 1/10) +
    geom_jitter()
```



直接将酒精和quality对比会发现并没有什么太大的关系。 但是画出不同酒精的平均quality会发现酒精越多quality越高

jiehe.sulfur.dioxide和quality分析

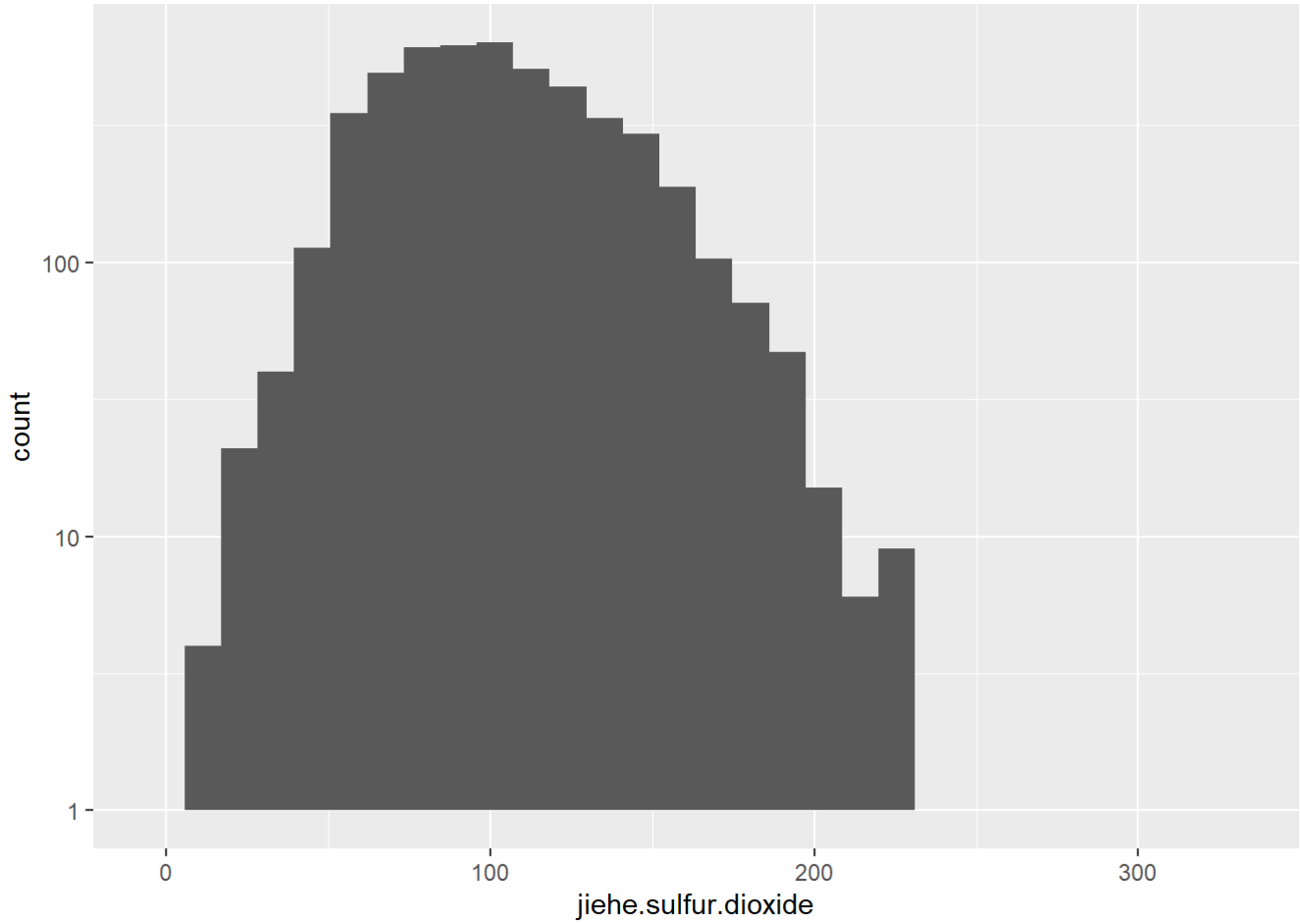
首先对新的变量，jiehe.sulfur.dioxide分析

```
ggplot(aes(x = jiehe.sulfur.dioxide), data = wineData) +
    geom_histogram() +
    scale_y_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



整体比较均匀，接近正态分布。

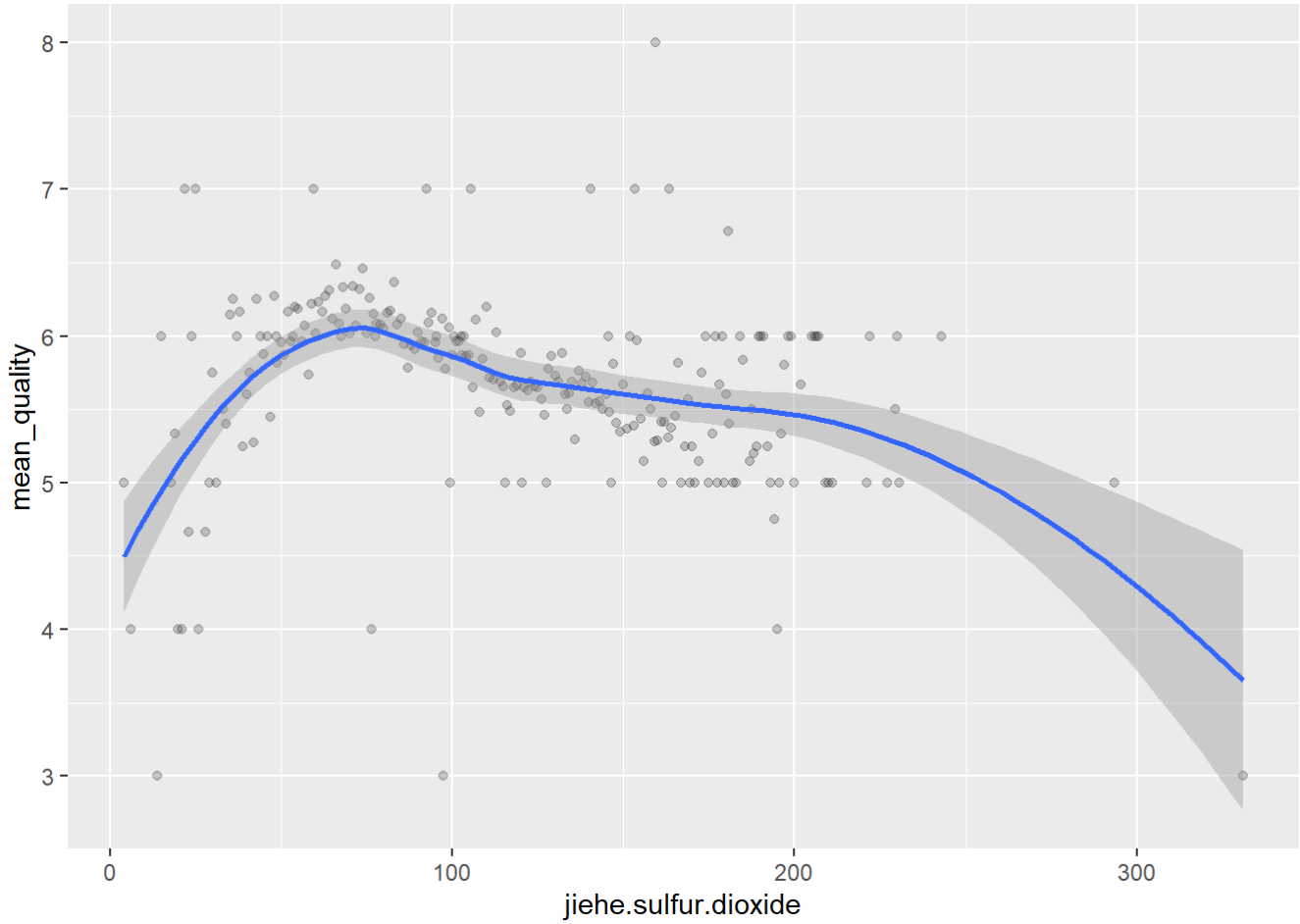
接着分析jiehe.sulfur.dioxide和平均quality的关系

```
# 按照jiehe.sulfur.dioxide分组
wineDataByAlcohol = group_by(wineData, jiehe.sulfur.dioxide)

# 按照jiehe.sulfur.dioxide分组后求mean_quality和median_quality
wineDataByAlcoholByAlcohol = summarise(wineDataByAlcohol,
                                       mean_quality = mean(quality),
                                        median_quality = median(quality))

ggplot(aes(x = jiehe.sulfur.dioxide, y = mean_quality),
       data = wineDataByAlcoholByAlcohol) +
    geom_point(color = I("black"), fill = I("#099DD9"), alpha = 20/100) +
    geom_smooth()
```

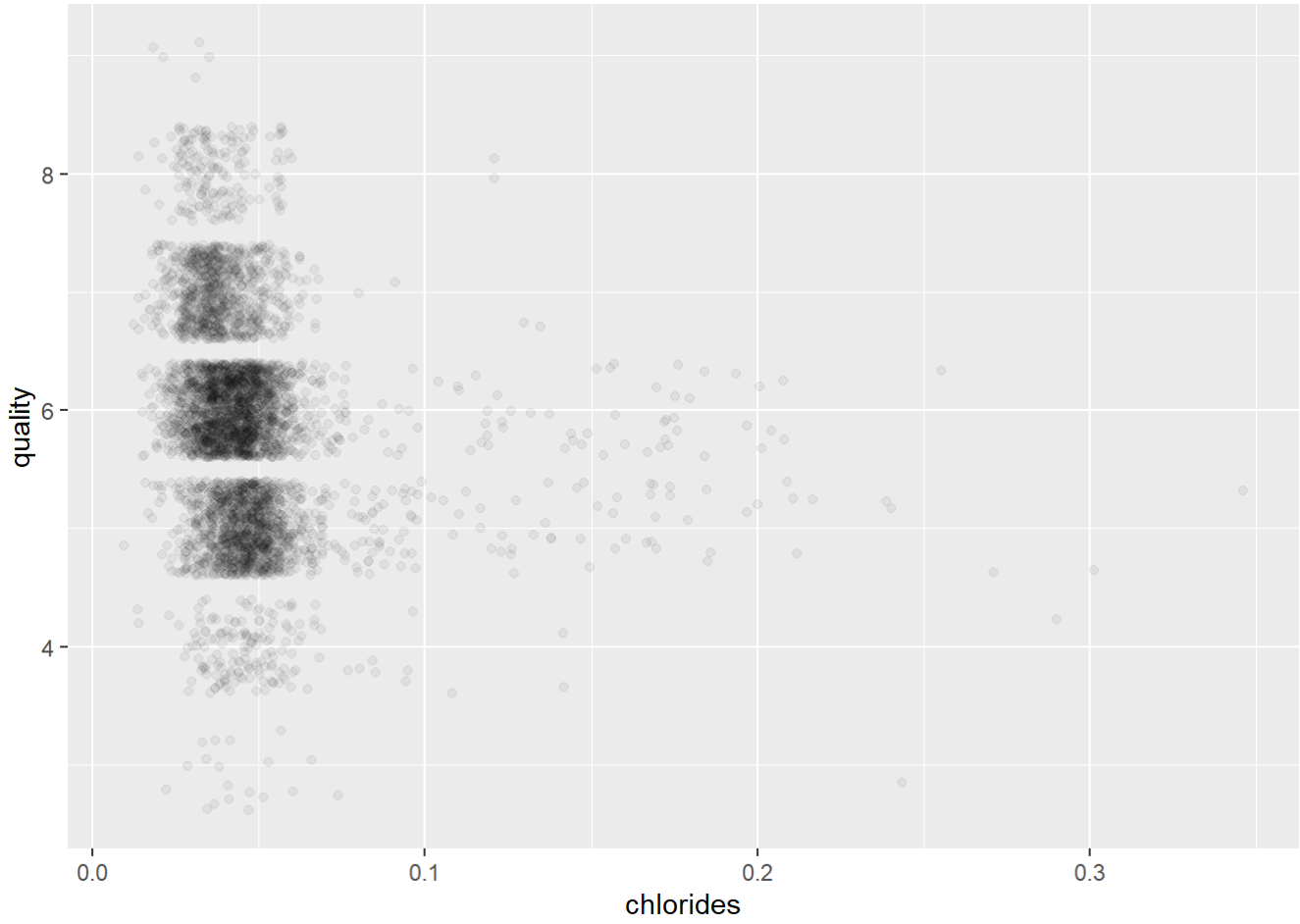
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



结合二氧化硫越高，平均quality得分越低。因为这个元素有刺激性臭味，所以说放的越多味道越差。但是少量的放入可以提高口感。

chlorides和quality分析

```
ggplot(aes(x = chlorides, y = quality), data = wineData) +
    geom_point(color = I("black"), fill = I("#099DD9"),
               position = position_jitter(), alpha = 0.05)
```

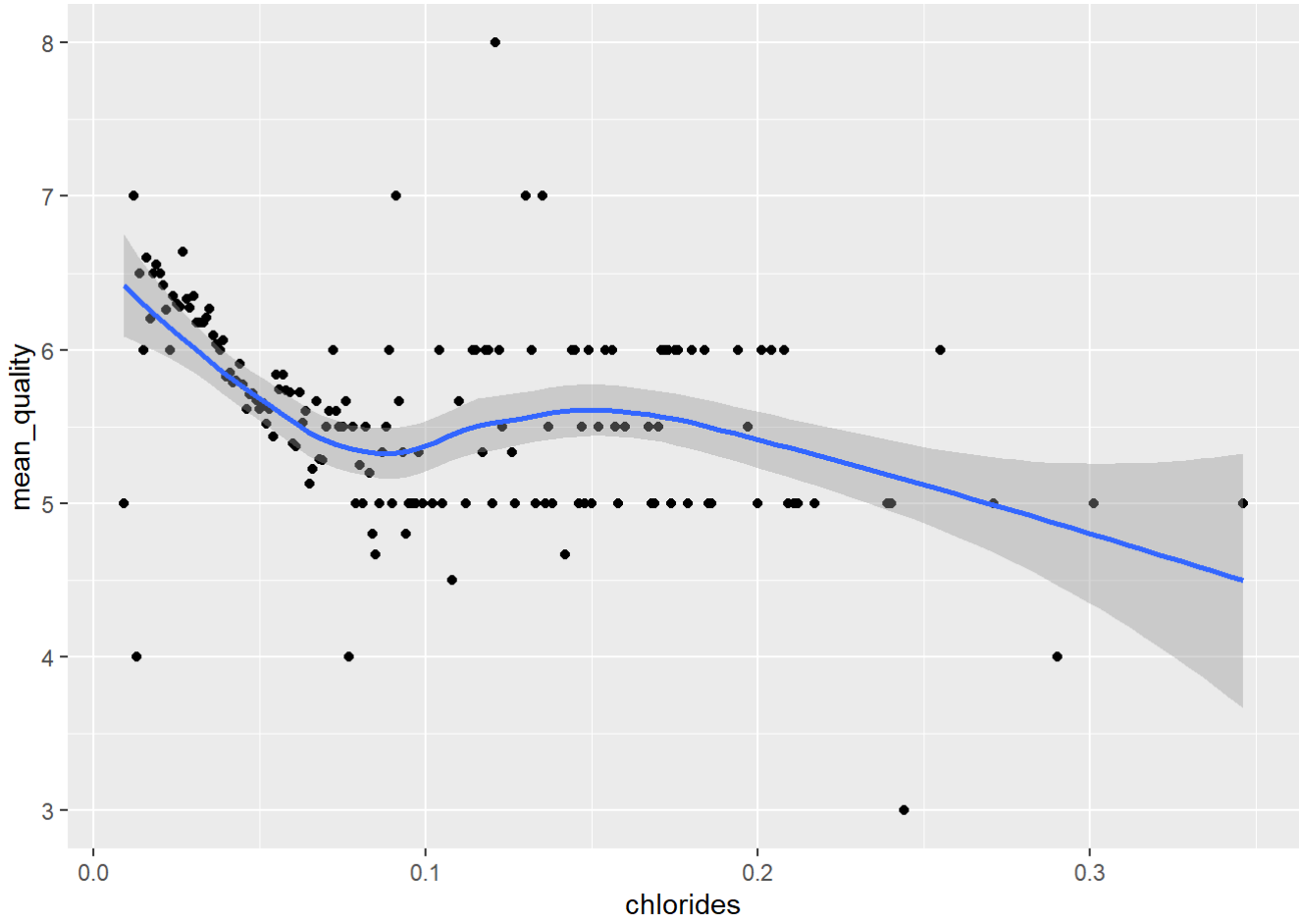


按照chlorides分组后求mean_quality和median_quality

```
# 按照chlorides分组后求mean_quality和median_quality
wineDataByAlcohol = group_by(wineData, chlorides)
wineDataByAlcoholByAlcohol = summarise(wineDataByAlcohol,
                                       mean_quality = mean(quality),
                                        median_quality = median(quality))

ggplot(aes(x = chlorides, y = mean_quality),
       data = wineDataByAlcoholByAlcohol) +
    geom_point(color = I("black"), fill = I("#099DD9")) +
    geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

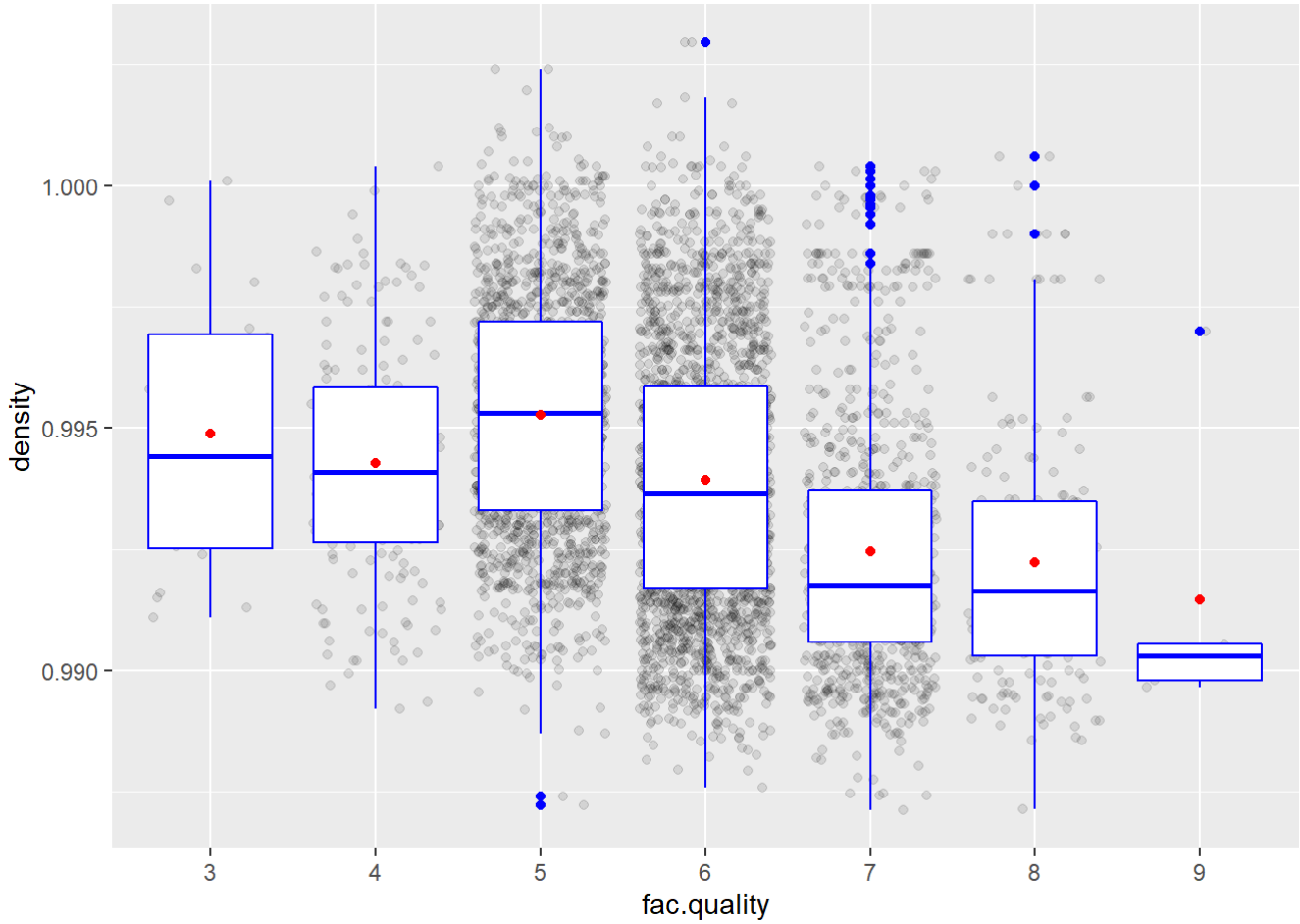


在区间[0~0.1]中，氯化物放的越多，平均quality越差

density和quality分析

```
# 因子化quality
wineData$fac.quality = factor(wineData$quality)


ggplot(aes(x = fac.quality, y = density),
       data = subset(wineData, density < 1.01)) +  # 选取密度小于1.01的值
    geom_jitter(alpha = 1/10) + # 画出抖动图
    geom_boxplot(color = "blue") +
    stat_summary(fun.y = mean, geom = "point", color = "red") # 标记均值点
```

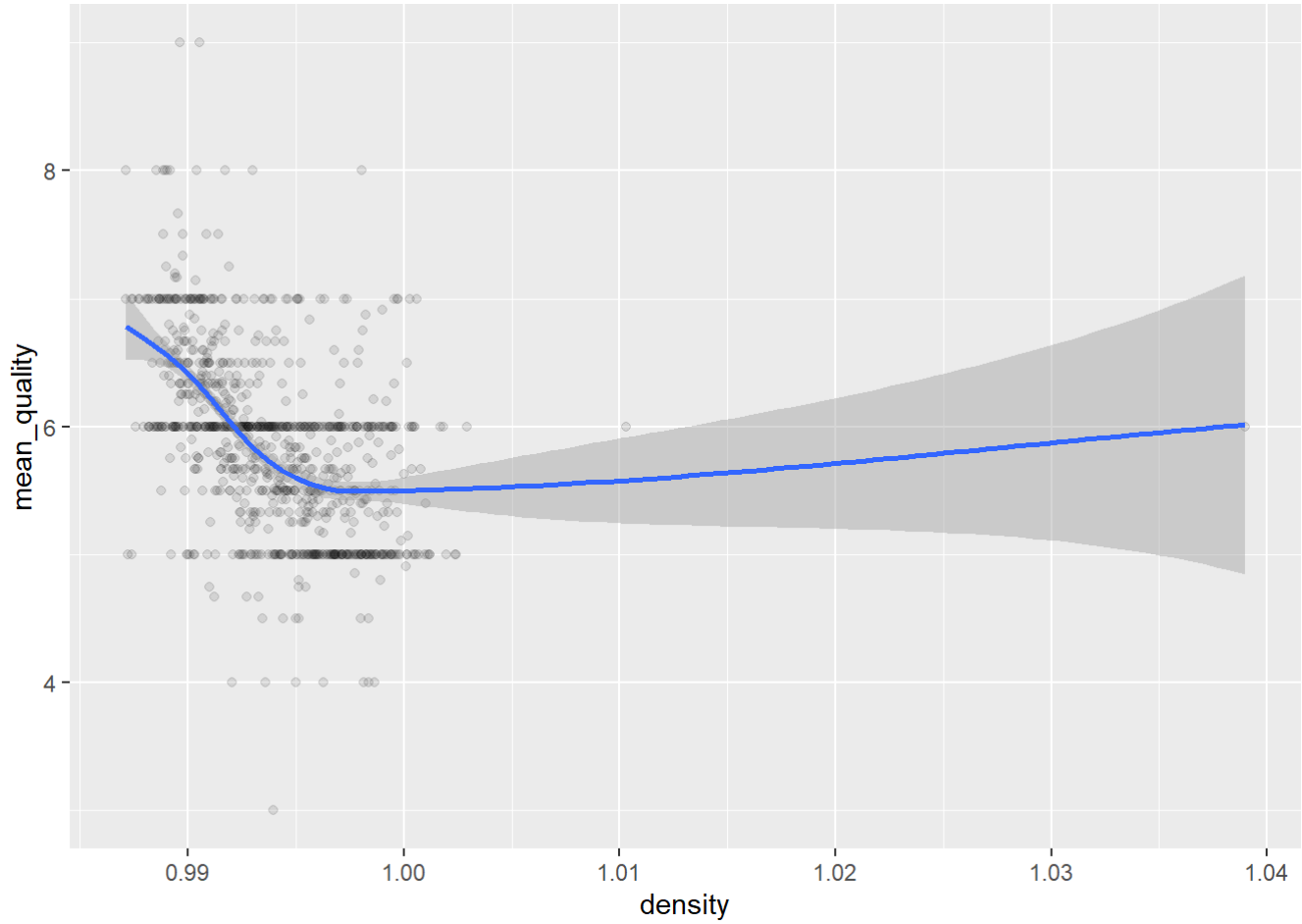


从上图可以看出，quality越高，density的值整体走低。也就是说密度越低的酒，quality越高。

```
# 按照density分组后求mean_quality和median_quality
wineDataByAlcohol = group_by(wineData, density)
wineDataByAlcoholByAlcohol = summarise(wineDataByAlcohol,
                                       mean_quality = mean(quality),
                                        median_quality = median(quality))

ggplot(aes(x = density, y = mean_quality),
       data = wineDataByAlcoholByAlcohol) +
    geom_point(color = I("black"), fill = I("#099DD9"), alpha = 10/100) +
    geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



在区间[0~1]之间，密度越高，平均quality越低。 也许酒的密度越是大于水的密度，口感越好。

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?

如下四个特征和quality的相关性比较强 alcohol 0.44 density -0.31 jiehe.sulfur.dioxide -0.22 chlorides -0.21

你是否观察到主要特性与其他特性之间的有趣关系?

总体上，酒精含量越高，quality越高。可能人们觉得喝了“上头”的酒才是好酒。

你发现最强的关系是什么?

最强的关系就是alcohol

多变量绘图选择

按照酒精分组

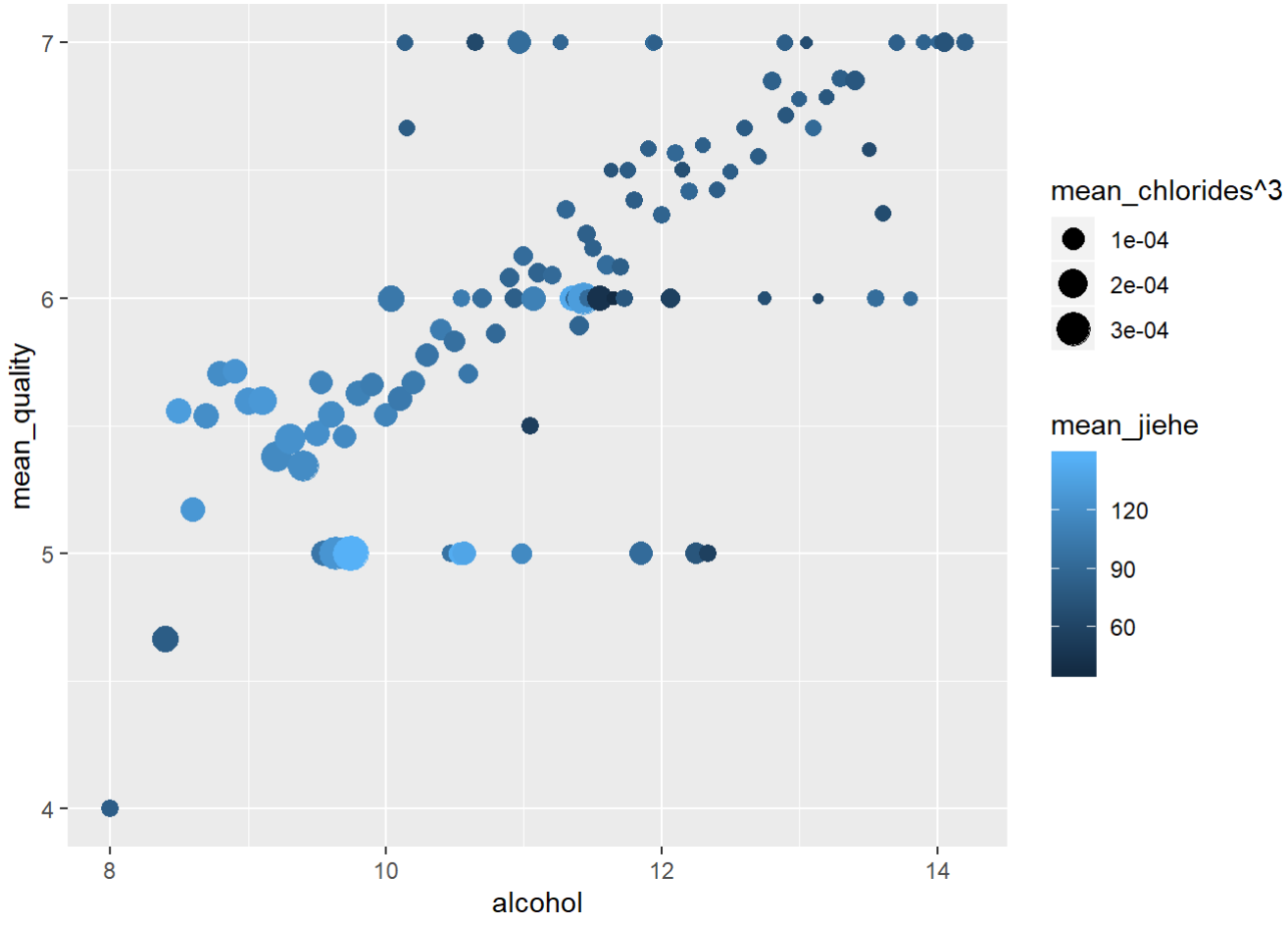
按照酒精分组，求出jiehe.sulfur.dioxide、chlorides、qualit的均值和中位数。

```
# 按照酒精分组，求出jiehe.sulfur.dioxide、chlorides、qualit的均值和中位数。
gbclco = group_by(wineData, alcohol)
wineData.group = summarise(gbclco,
                           mean_jiehe = mean(jiehe.sulfur.dioxide),
                           mean_chlorides = mean(chlorides),
                           mean_quality = mean(quality),
                           median_jiehe = median(jiehe.sulfur.dioxide),
                           median_chlorides = median(chlorides),
                           median_quality = median(quality))

head(wineData.group)
```

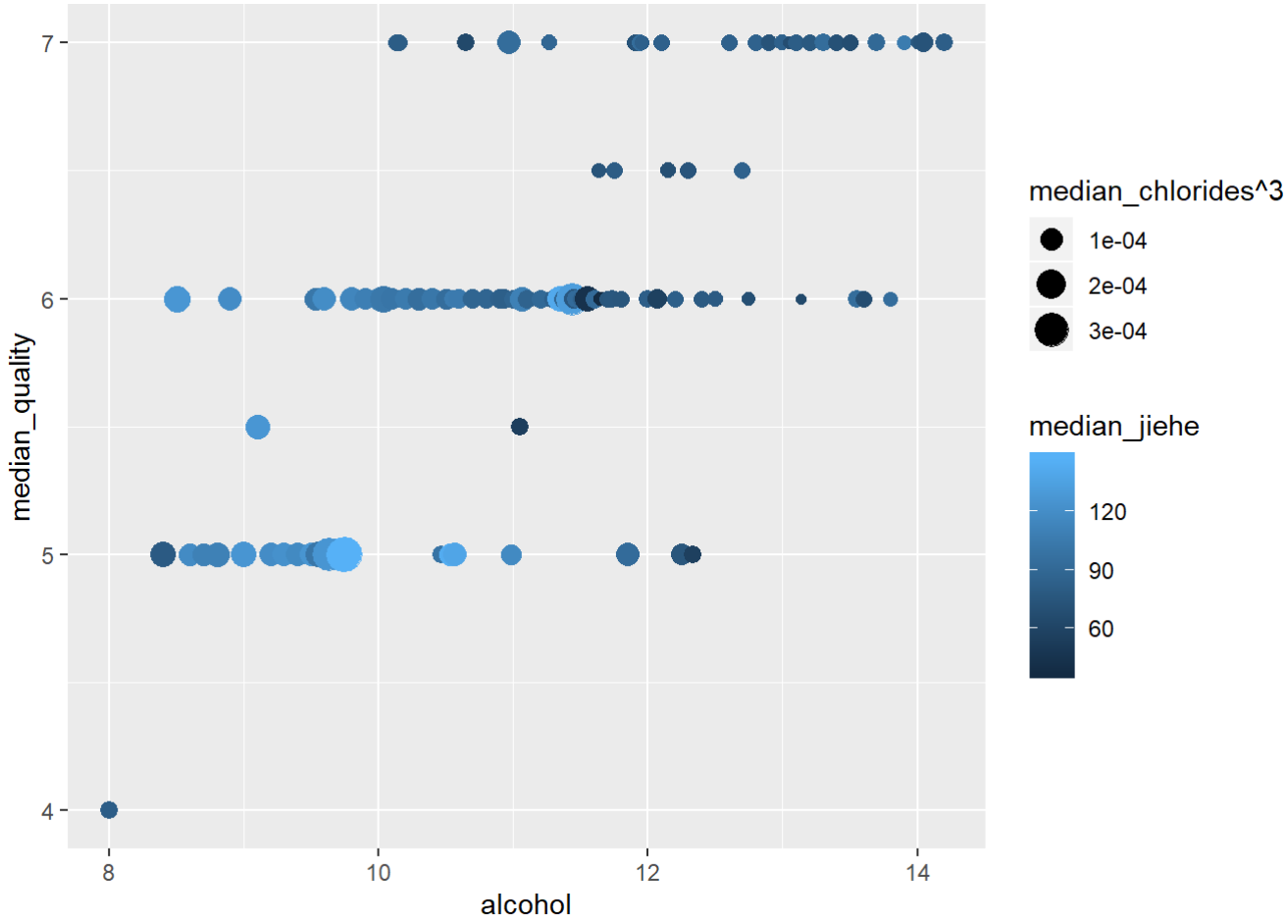
```
## # A tibble: 6 x 7
##   alcohol mean_jiehe mean_chlorides mean_quality median_jiehe
##     <dbl>      <dbl>          <dbl>        <dbl>        <dbl>
## 1     8          81.5         0.037          4           81.5
## 2     8.4        81           0.0533         4.67        79
## 3     8.5      132            0.052          5.56       127
## 4     8.6      127.           0.0507         5.17       118
## 5     8.7      121.           0.0519         5.54       110
## 6     8.8      120.           0.0511         5.70       110
## # ... with 2 more variables: median_chlorides <dbl>, median_quality <dbl>
```

```
ggplot(aes(x = alcohol, y = mean_quality), data = wineData.group) +
    geom_point(aes(color = mean_jiehe, size = mean_chlorides**3))
```



上图是qualit、jiehe、chorides和alcohol四个变量的均值的对比图。 可以看出随着alcoho的含量越多，quality整体越高 随着jiehe的含量越少（颜色越深），quality整体越高 随着chlorides的含量越少（形状越小），quality整体越高

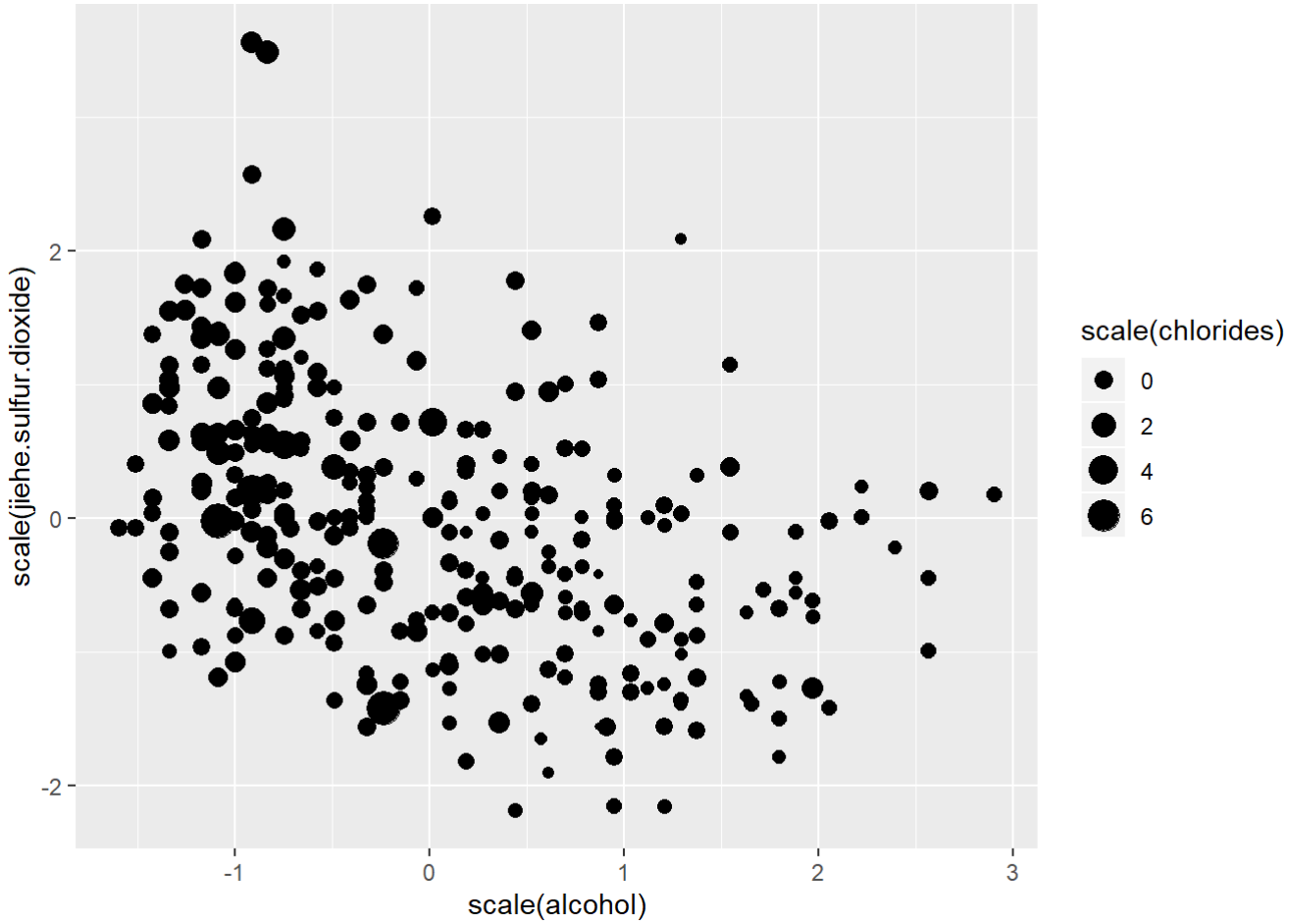
```
ggplot(aes(x = alcohol, y = median_quality), data = wineData.group) +
    geom_point(aes(color = median_jiehe, size = median_chlorides**3))
```



上图是qualit、jiehe、chorides和alcohol四个变量的中位数的对比图。 可以看出随着alcoho的含量越多，quality整体越高 随着jiehe的含量越少（颜色越深），quality整体越高 随着chlorides的含量越少（形状越小），quality整体越高

```
# 随机抽取300个数据，并且绘制散点图
x = sample(wineData$X, 300)

ggplot(aes(x = scale(alcohol), y = scale(jiehe.sulfur.dioxide)),
       data = wineData[x,]) +
    geom_point(aes(size = scale(chlorides)))
```

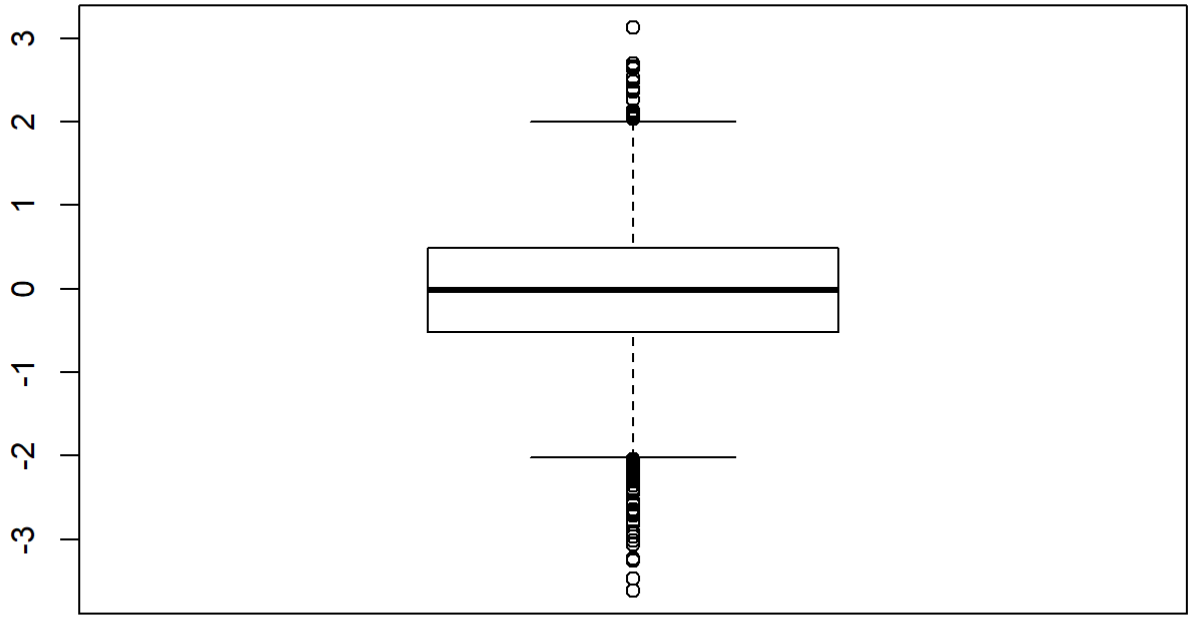


通过随机抽样的方法，在winedata数据中抽取200个数据，绘制alcohol、chlorides和jiehe.sulfur.dioxide三个变量的关系图。 因为考虑到这个三个变量的单位都是不一样的，所以我考虑用scale的方法，将它们归一化。 从上图看出，alcohol和chorides呈反比状态，和jiehe.sulfur.dioxide也呈反比状态。

模型建立

```
myLm = lm(quality ~ chlorides + alcohol + jiehe.sulfur.dioxide, data = wineData)

# 求模型的残差，并且绘制残差的boxplot和五数概括
resData = residuals(myLm)
boxplot(resData)
```




```
quantile(resData)
```

```
##          0%         25%         50%         75%        100%
## -3.61570281 -0.51995299 -0.01916194  0.49144872  3.12249649
```

使用quality作为目标变量，使用chlorides 、 alcohol 和 jiehe.sulfur.dioxide作为因变量。拟合出了一个多元的回归模型。该模型的残差中位数在0附近，这是一个好的预兆，我们的模型拟合的不错。但是残差中有很多极大的极小的并且偏离0太远的值，说明模型在处理某些值的时候不是特别的好。

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？

qualit、jiehe、chorides和alcohol四个变量相互促进。 可以看出随着alcoho的含量越多，quality整体越高 随着jiehe的含量越少（颜色越深），quality整体越高 随着chlorides的含量越少（形状越小），quality整体越高

这些特性之间是否存在有趣或惊人的联系呢？

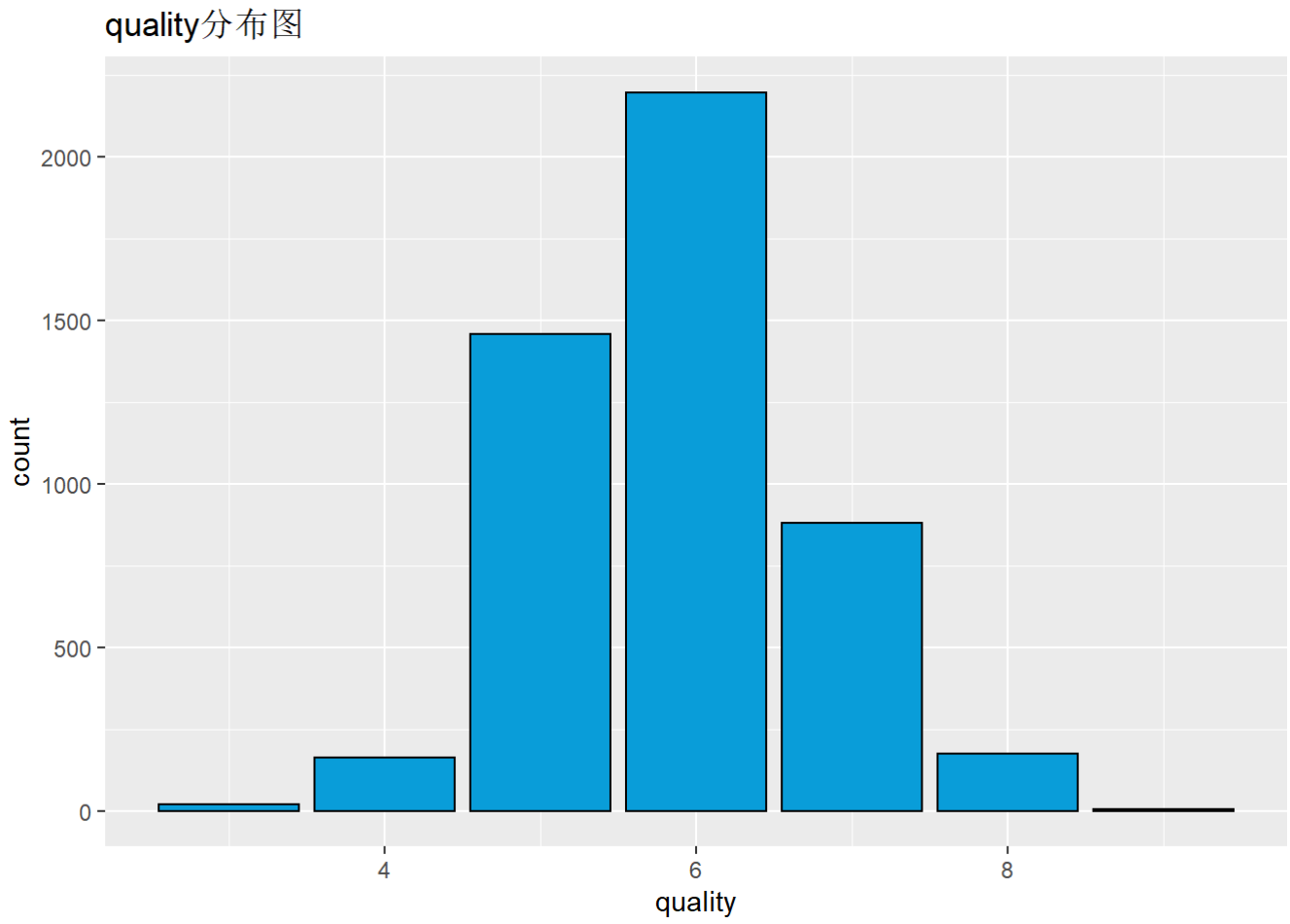
我发现酒精的含量和chorides呈反比状态，和jiehe.sulfur.dioxide也呈反比状态。

选项：你是否创建过数据集的任何模型？讨论你模型的优缺点。

使用quality作为目标变量，使用chlorides 、 alcohol 和 jiehe.sulfur.dioxide作为因变量。拟合出了一个多元的回归模型。该模型的残差中位数在0附近，这是一个好的预兆，我们的模型拟合的不错。但是残差中有很多极大的极小的并且偏离0太远的值，说明模型在处理某些值的时候不是特别的好。

定稿图与总结

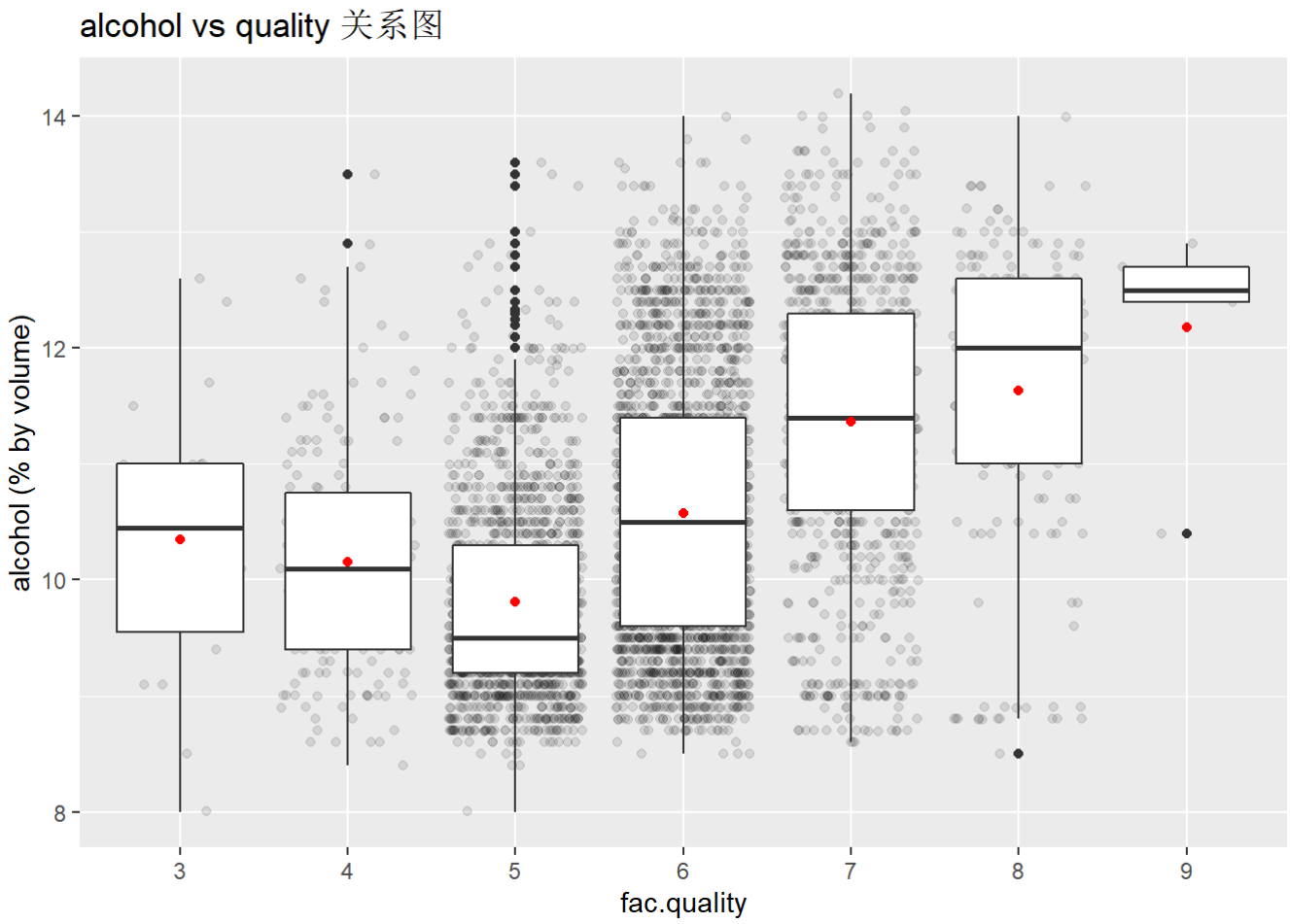
绘图一



描述一

大部分的白葡萄酒的quality为6，也就说大部分的白葡萄酒是中等的quality。 quality为0和9的白葡萄酒最少，几乎可以不计。

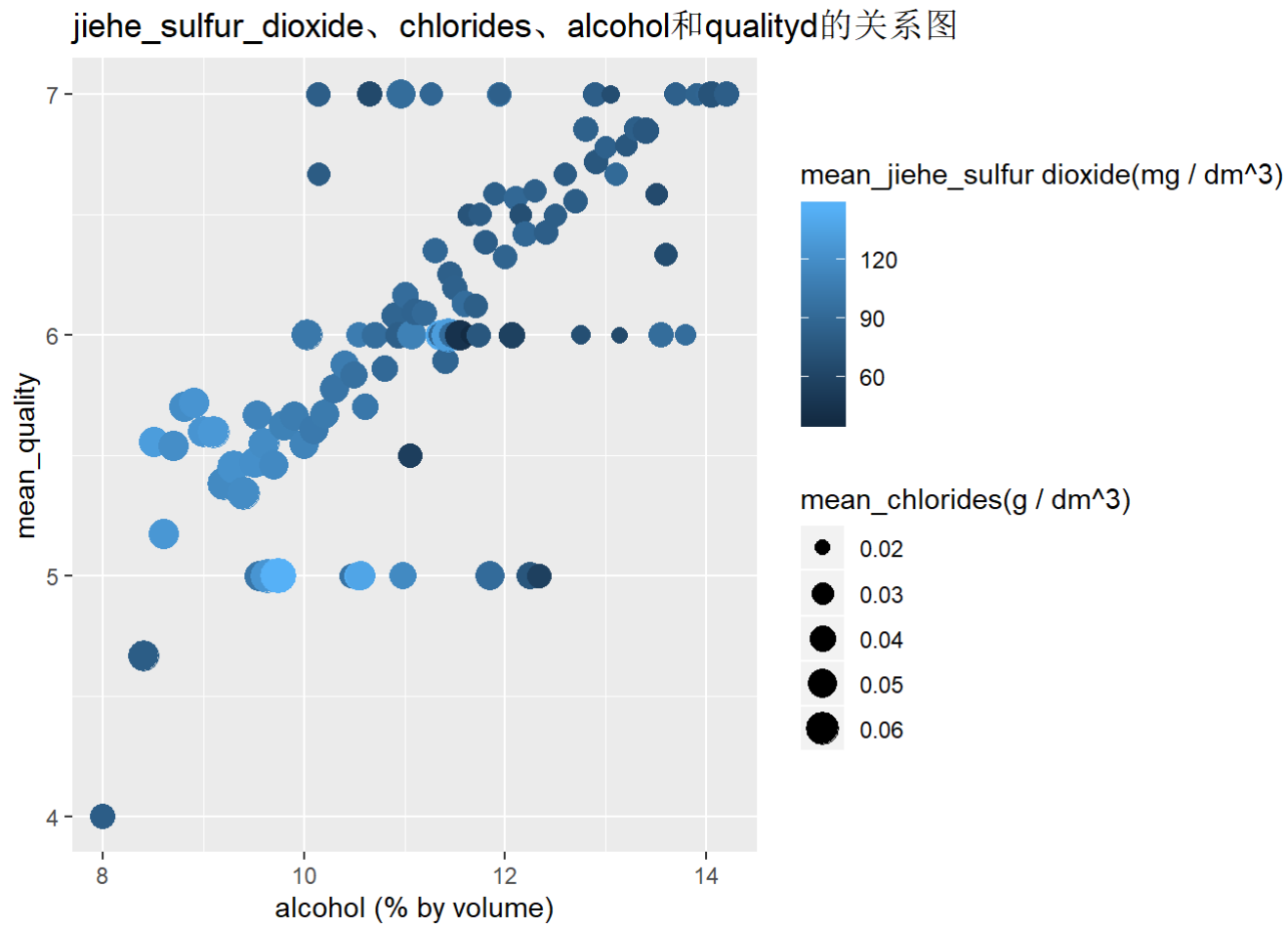
绘图二



描述二

从数据上来看，alcohol和quality的关系比较明显。随着alcohol的增加，quality也随之增加。也就说人们更爱喝酒精度数高的酒。

绘图三



描述三

上图是qualit、jiehe、chorides和alcohol四个变量的均值的对比图。 可以看出随着alcoho的含量越多，quality整体越高 随着jiehe的含量越少（颜色越深），quality整体越高 随着chlorides的含量越少（形状越小），quality整体越高

反思

1.难点和克服。一是英文，花了很多时间去翻译英文单词，理解每个变量名的意思；二是酒精对比quality的关系图中，因为没有什么发现，当时在干了很多咖啡，灵机一动，用均值来处理，然后得到不错的效果。三是模型，模型的建立是比较简单，但是如何衡量一个模型的好坏，花了很多时间去网上找资料，最后以我能能理解的残差来衡量模型。

2.成功的发现。最成功是酒精和quality的关系，一开始作者我认为酒精越低，越淡应该越好喝，但是实际上不是的，人们更爱喝烈酒。另外一个成功的发现“结合二氧化硫”。在某一篇百度知道上看到“总二氧化硫 等于 结合的 + 游离的”，我灵机一动，加了个新的变量，没想到新的变量和quality的关系系数绝对值居然很高。

3.未来如何进一步丰富分析内容和提高报告质量。一是增加数据二次清洗的部分，删除离群点，再去分析数据。二是，可以拿出80%的数据作为分析数据去建模，剩下的20%用来检测模型的残差，这样的结果更加有说服力，也让我更知道做了哪些事情可以提高模型的精准度。