白葡萄酒数据探索性分析 By Jerry Shi

```
## '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 0.04
5 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 ...
```

Univariate Plots Section

白葡萄酒数据集共包含 4898 个样本和 13 个变量，忽略其中的样本编号变量。以下列出各变量，并对各个变量做单变量 EDA：

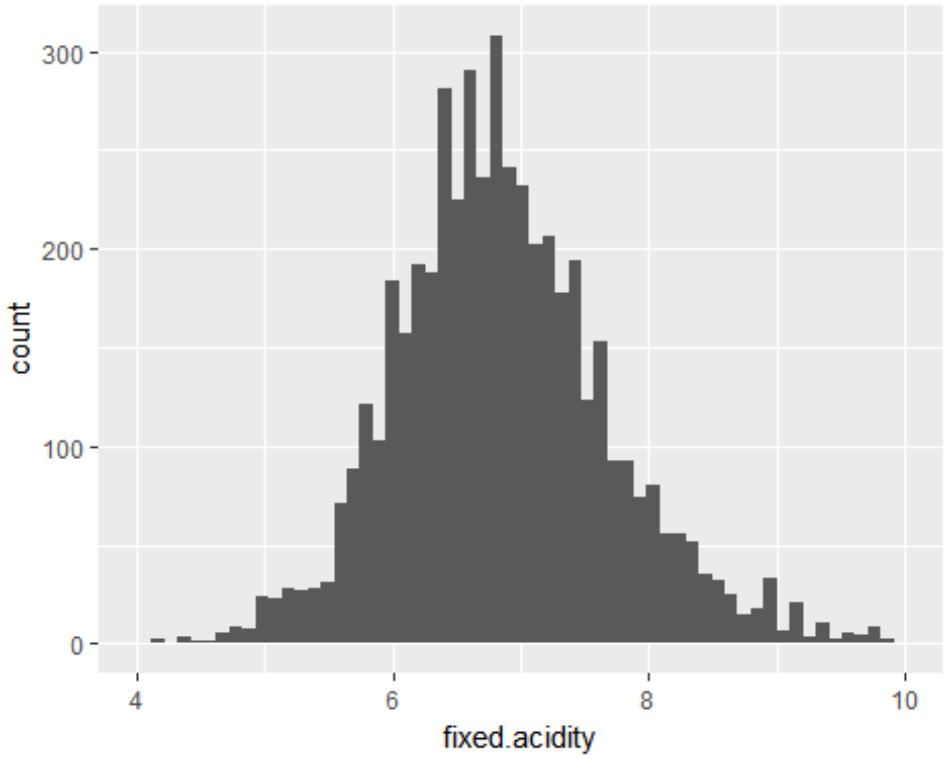
变量

- fixed acidity/非挥发性酸度: g/dm^3
- volatile acidity/挥发性酸度: g/dm^3
- citric acid/柠檬酸: g/dm^3
- residual sugar/残留糖分: g/dm^3
- chlorides/盐分: g/dm^3
- free sulfur dioxide/游离二氧化硫: mg/dm^3
- total sulfur dioxide/总二氧化硫: mg/dm^3
- density/密度: g/dm^3
- pH/酸碱度
- sulphates/硫酸盐浓度: g/dm^3
- alcohol/酒精度: %
- quality/评分：0~10

EDA

1. fixed acidity/非挥发性酸度: g/dm^3

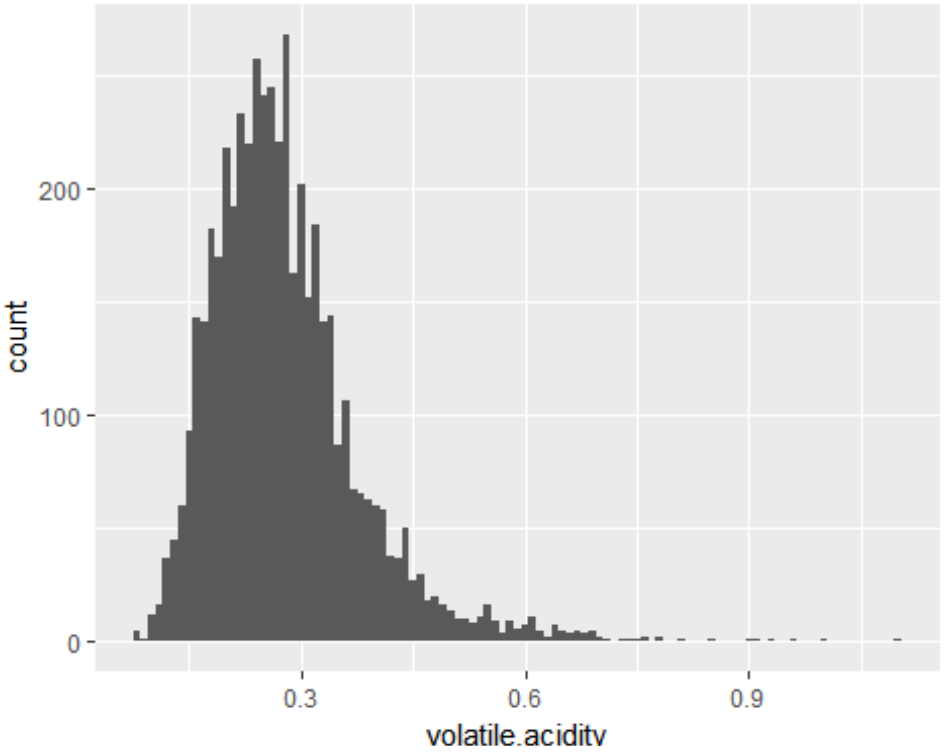
```
## Warning: Removed 9 rows containing non-finite values (stat_bin).

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

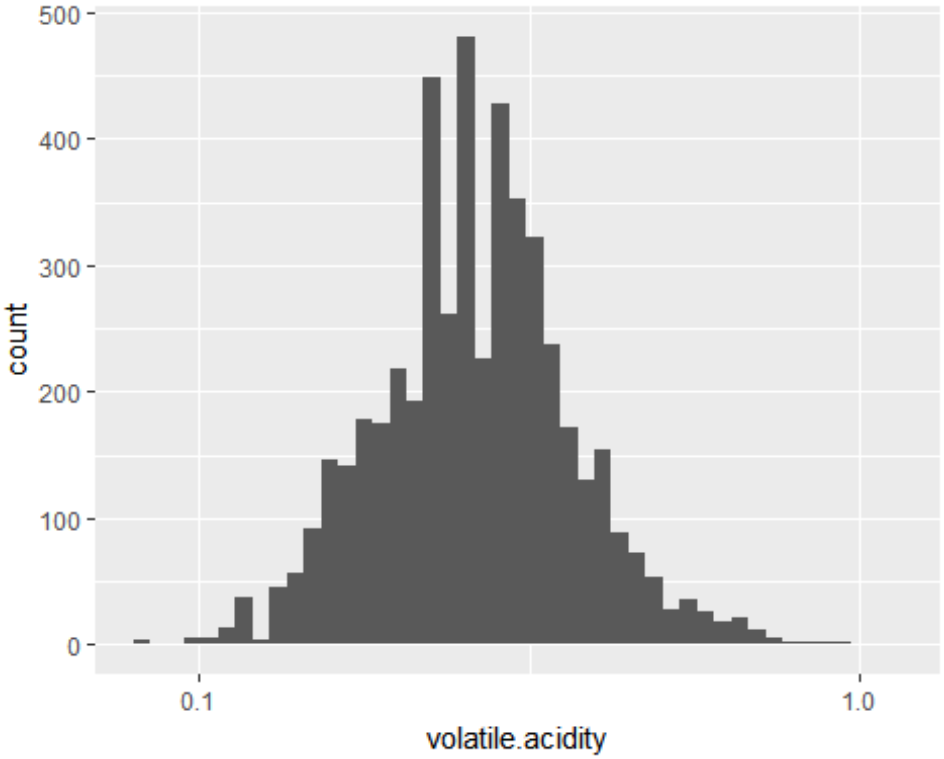


从直方图中可以看出，大部分葡萄酒的非挥发性酸度在 6-8 之间，大体呈正态分布，相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

2. volatile acidity/挥发性酸度: g/dm^3

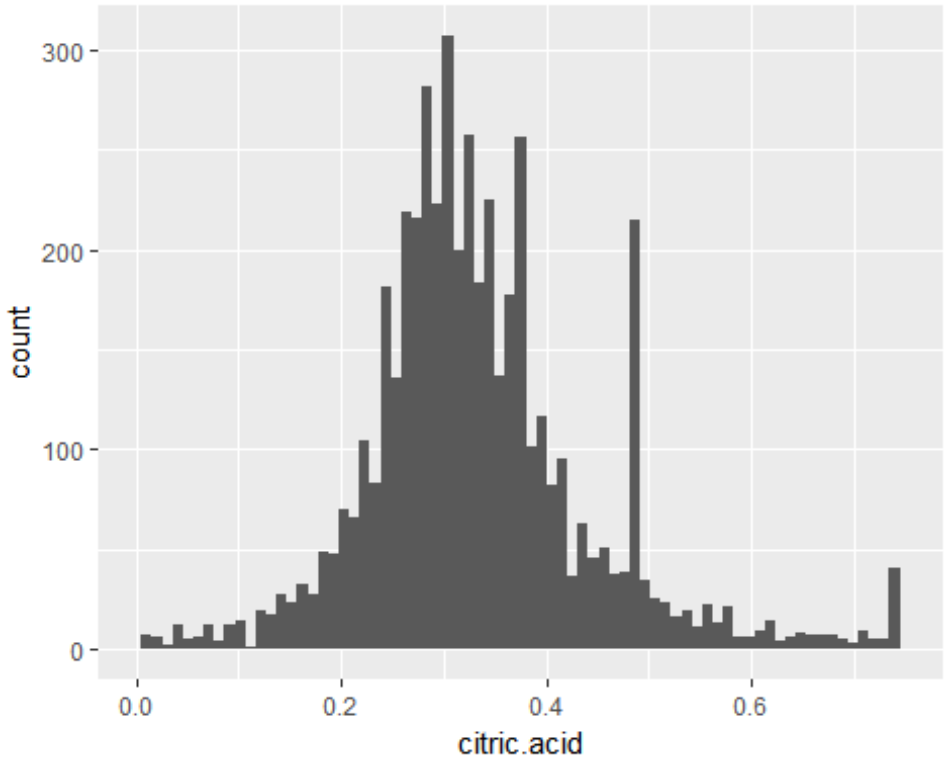




在原始直方图中分布稍有偏斜，因此进行对数转换。从转换后的直方图中，可以看到变量也是大体在 0.1-1.0 间呈正态分布的。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

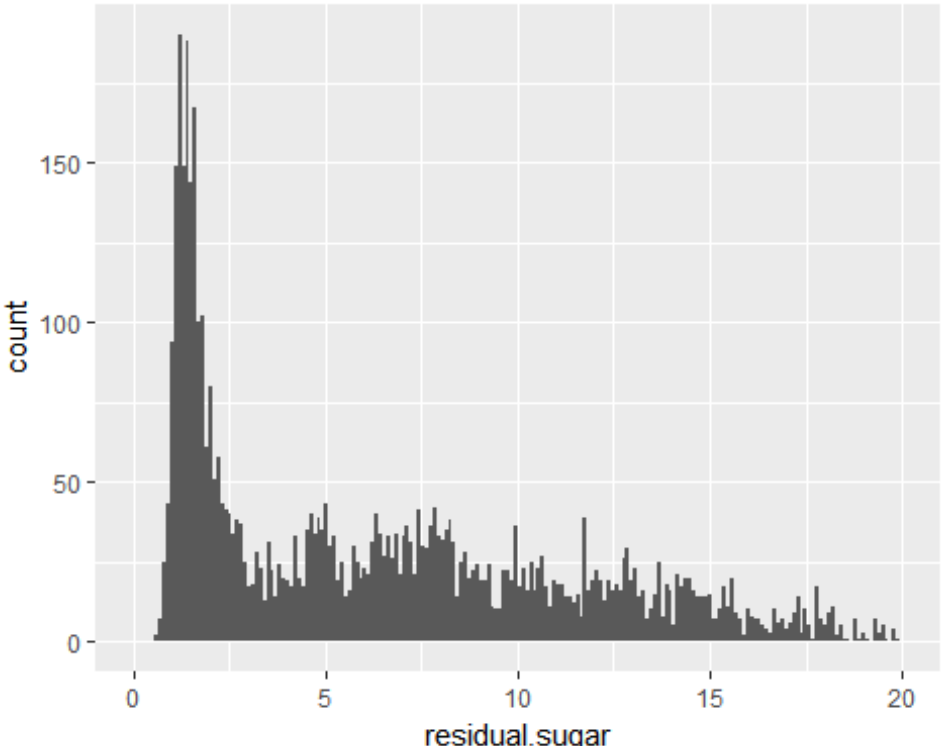
3. citric acid/柠檬酸: g/dm^3

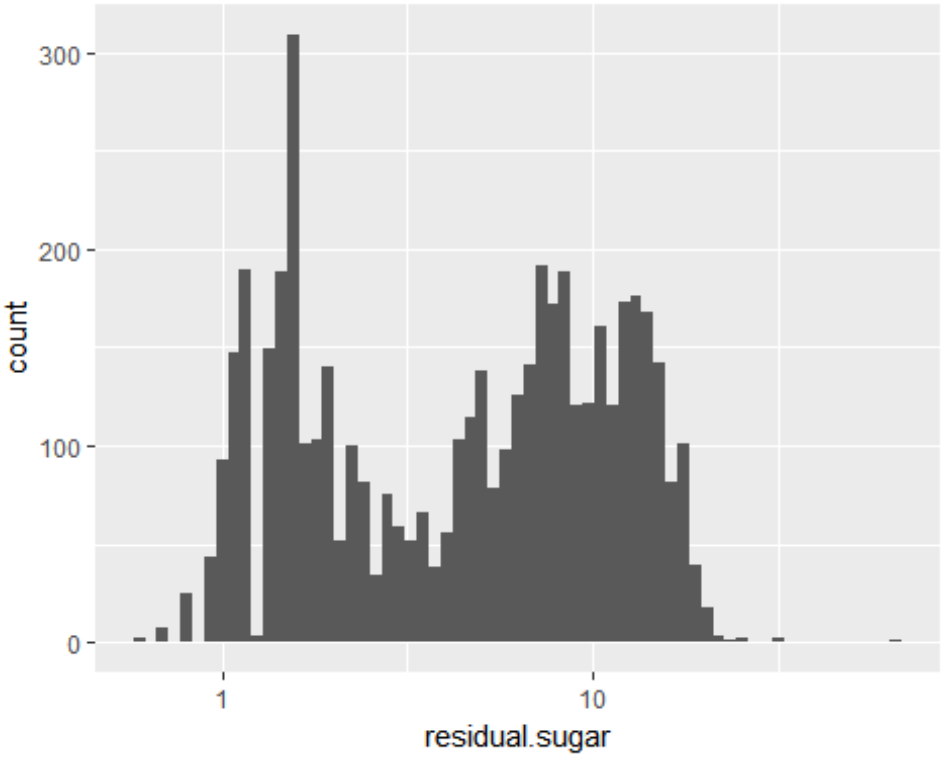


由上图可以看出，大部分葡萄酒的柠檬酸在 0.1-0.5g/dm^3 之间，整体呈正态分布，但在 0.5 附近形成了一个高峰。初步判断这可能和葡萄酒的规格有关。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

4. residual sugar/残留糖分: g/dm^3

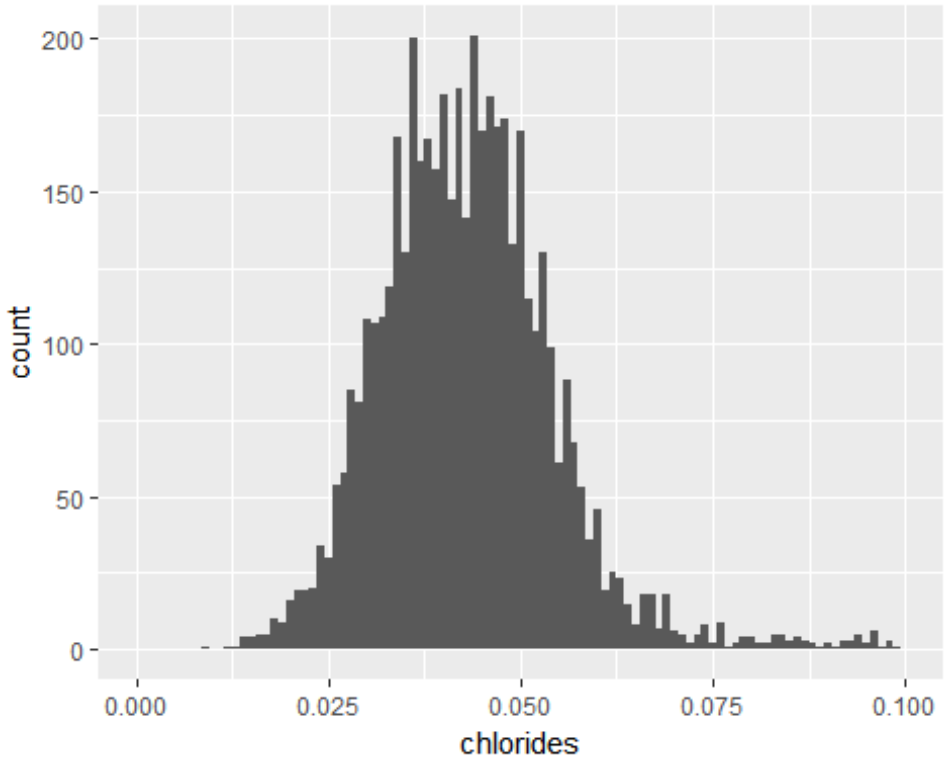




在原始直方图中，该变量分布偏斜，因此进行对数转换。从转换后的直方图中可以明显看出变量呈双峰分布。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

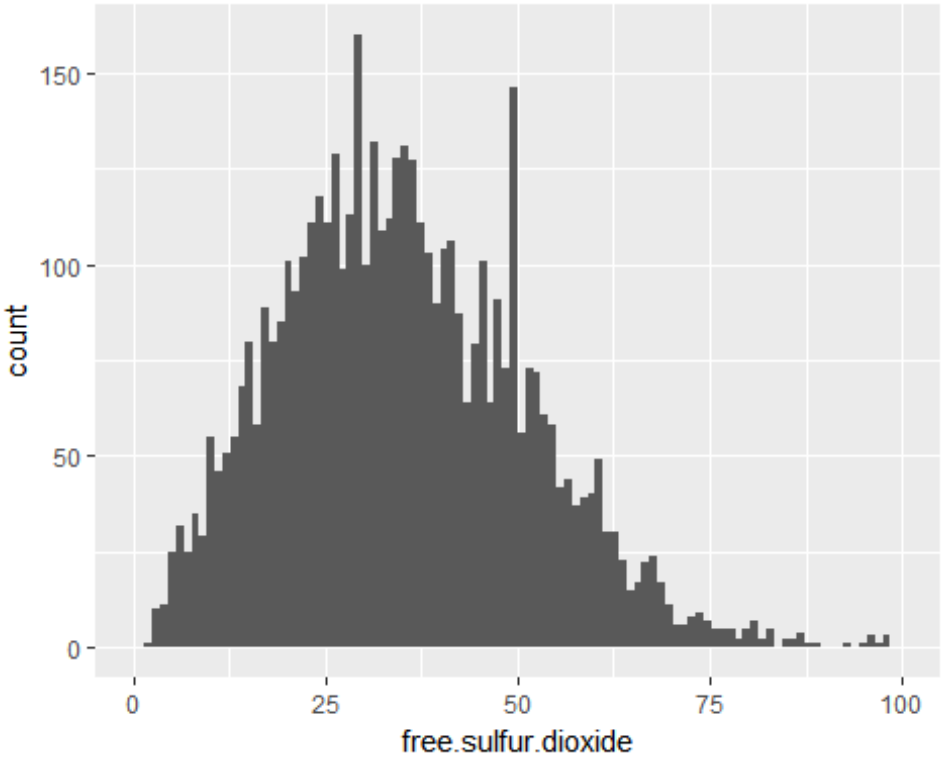
5. chlorides/盐分: g/dm^3



从直方图中可以看出，大部分葡萄酒的盐分都位于 0.025-0.075g/dm^3 的范围内，且呈正态分布。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

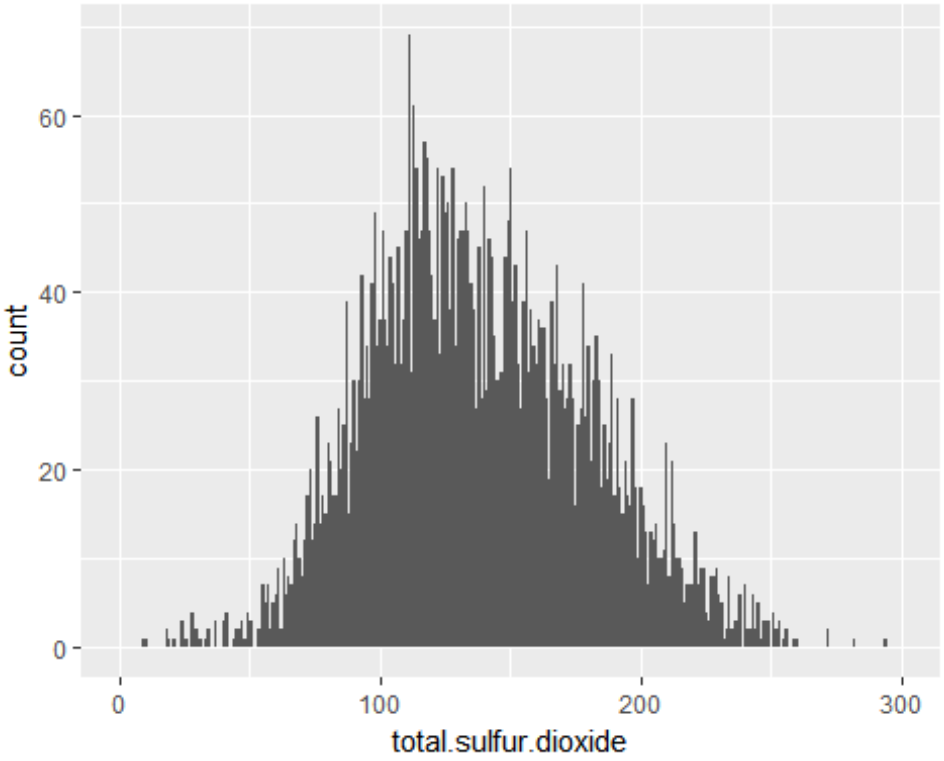
6. free sulfur dioxide/游离二氧化硫: mg/dm^3



由图可见，大部分葡萄酒的游离二氧化硫含量在 0-75mg/d^3 的范围内，呈正态分布。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

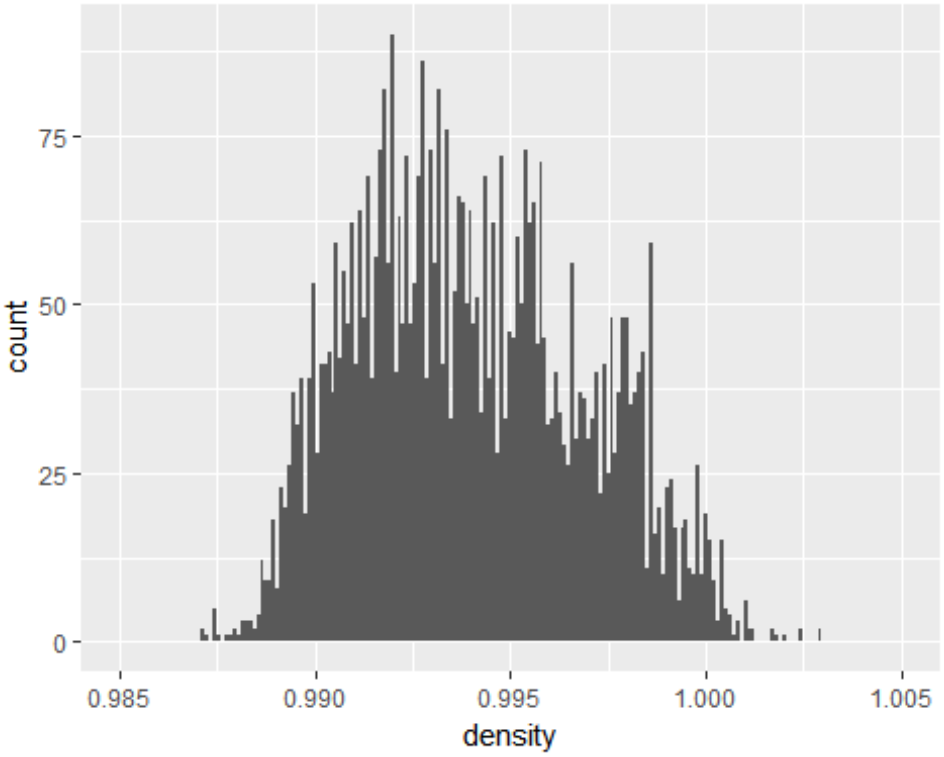
7. total sulfur dioxide/总二氧化硫: mg/dm^3



由图，大部分葡萄酒的总二氧化硫含量在 50-250mg/d^3 的范围内，呈正态分布。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

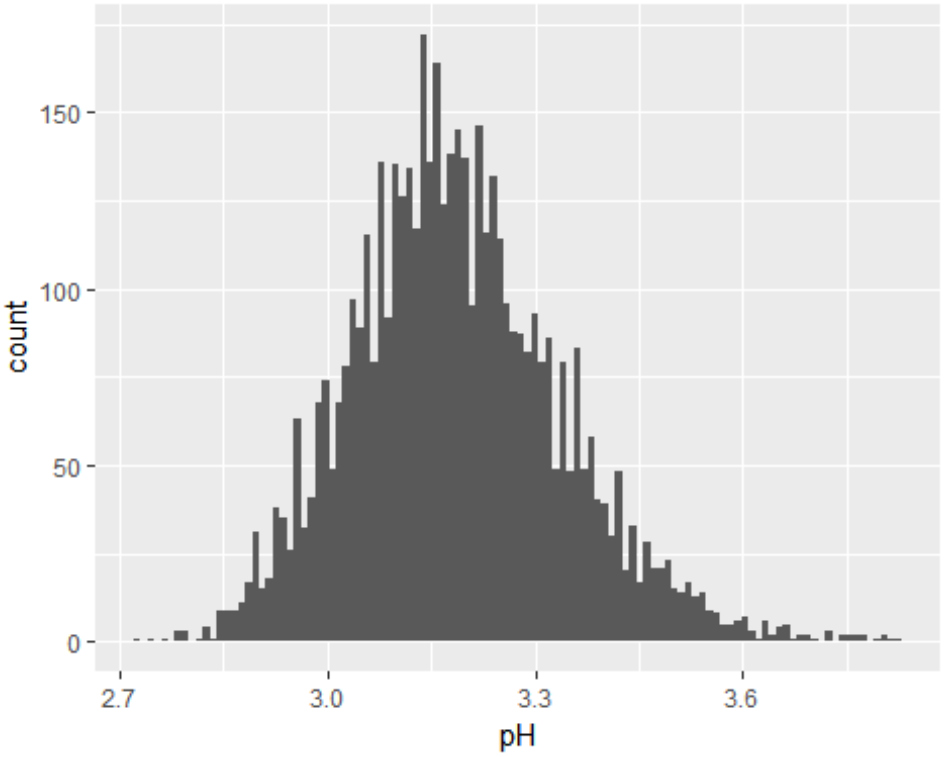
8. density/密度: g/dm^3



由图，大部分葡萄酒的密度在 0.99-1g/d^3 的范围内，呈正态分布，误差范围较小。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

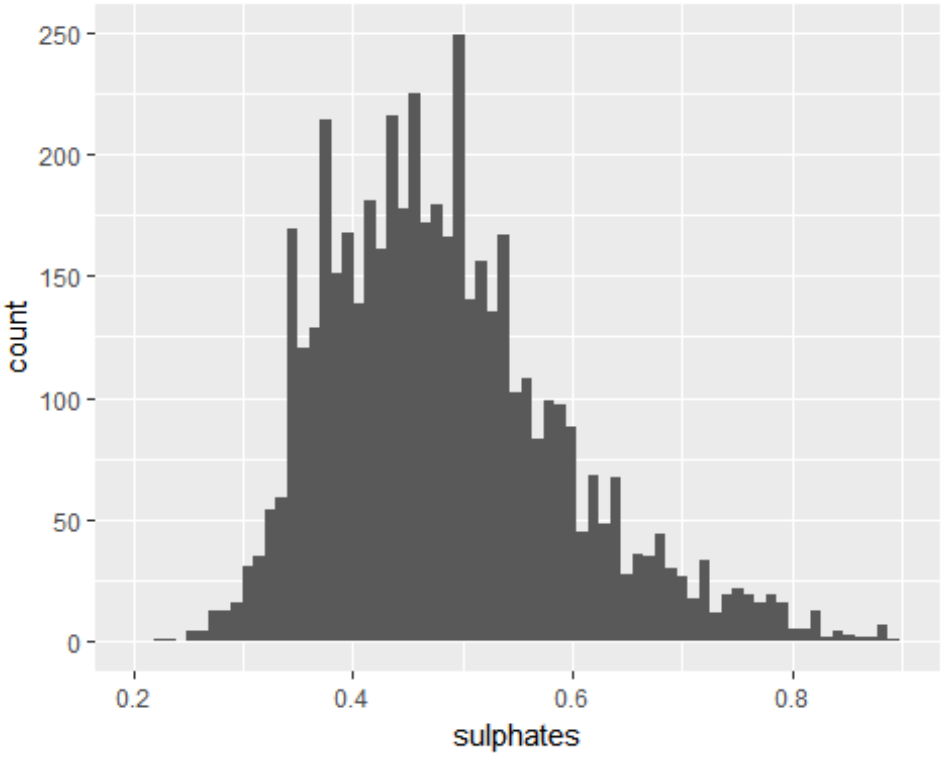
9. pH/酸碱度



由图，大部分葡萄酒的酸碱度在 2.7-3.6 之间，呈正态分布。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

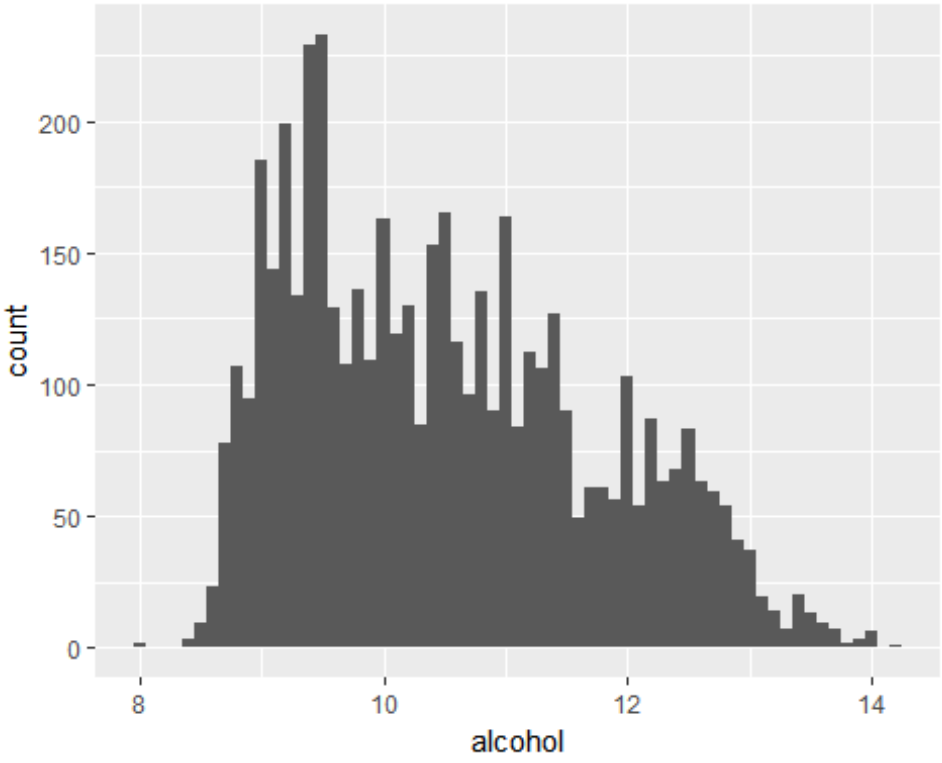
10. sulphates/硫酸盐浓度: g/dm^3



由图，大多数葡萄酒的硫酸盐浓度在 0.3-0.7g/dm^3 范围内，呈正态分布。相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```

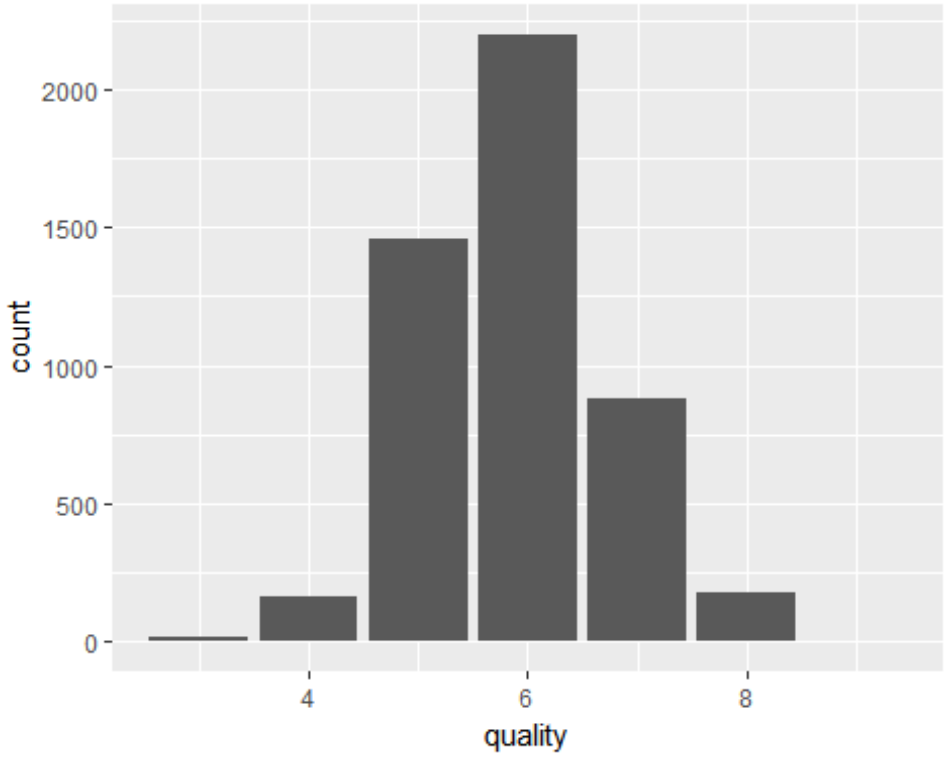
11. alcohol/酒精度: %



由图，酒精度变量呈长尾分布，相关统计值如下：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

12. quality/品质：0~10



将葡萄酒品质视为有序的因子变量。由图，葡萄酒的品质均在 3-9 之间。

Univariate Analysis

What is the structure of your dataset?

该数据集中有 4898 个种类的葡萄酒样本，包含 13 个变量。

其他的观测结果：

- 大部分葡萄酒的品质范围在 5-7 之间；
- 葡萄酒的密度差异较小

What is/are the main feature(s) of interest in your dataset?

葡萄酒的品质（评分）。通过探索其他变量和葡萄酒品质的关系，可以得出一系列统计推论。

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

由于还没有开始探索双变量及多变量关系，我认为除品质和序号外的其他 11 个变量都会有助于对葡萄酒品质的调查分析。

Did you create any new variables from existing variables in the dataset?

没有。因为除品质外各变量间相关性较小或者未知，所以没有生成新的变量。

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

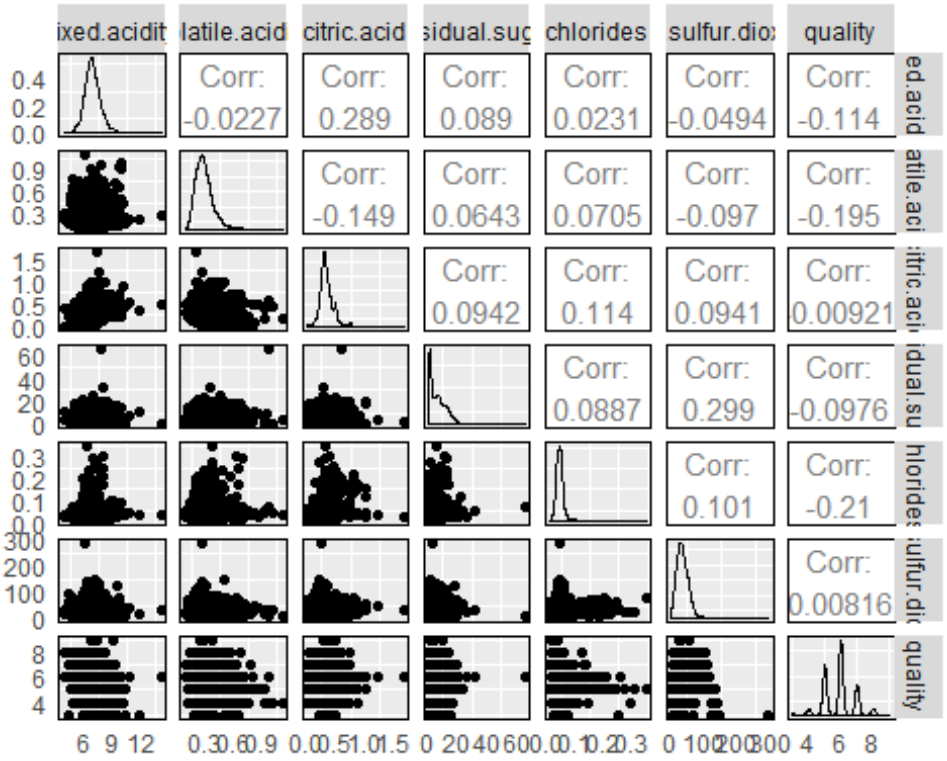
在探索过程中，以下几个变量在直方图中形成较为偏斜的分布，因此使用对数转换方式以形成正态分布的图形：

- volatile acidity/挥发性酸度: g/dm^3
- residual sugar/残留糖分: g/dm^3
- sulphates/硫酸盐浓度: g/dm^3

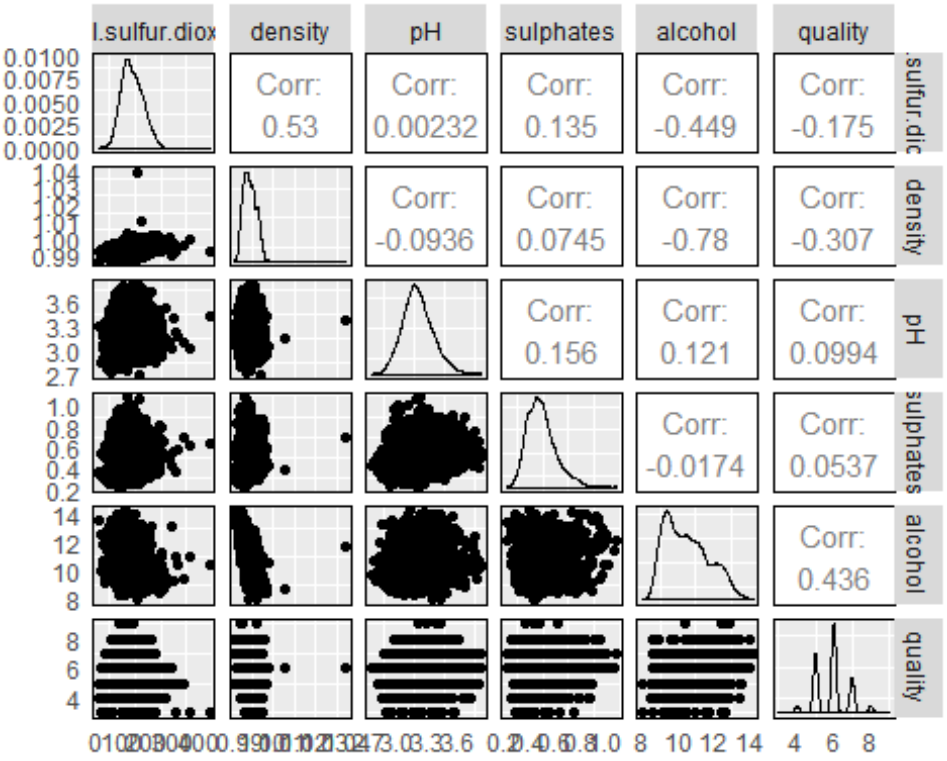
以下几个变量存在较大/小的异常值，因此通过限制范围获得更加合理的分布直方图：

- fixed acidity/非挥发性酸度: g/dm^3
- citric acid/柠檬酸: g/dm^3
- residual sugar/残留糖分: g/dm^3
- chlorides/盐分: g/dm^3
- free sulfur dioxide/游离二氧化硫: mg/dm^3
- total sulfur dioxide/总二氧化硫: mg/dm^3
- sulphates/硫酸盐浓度: g/dm^3

Bivariate Plots Section

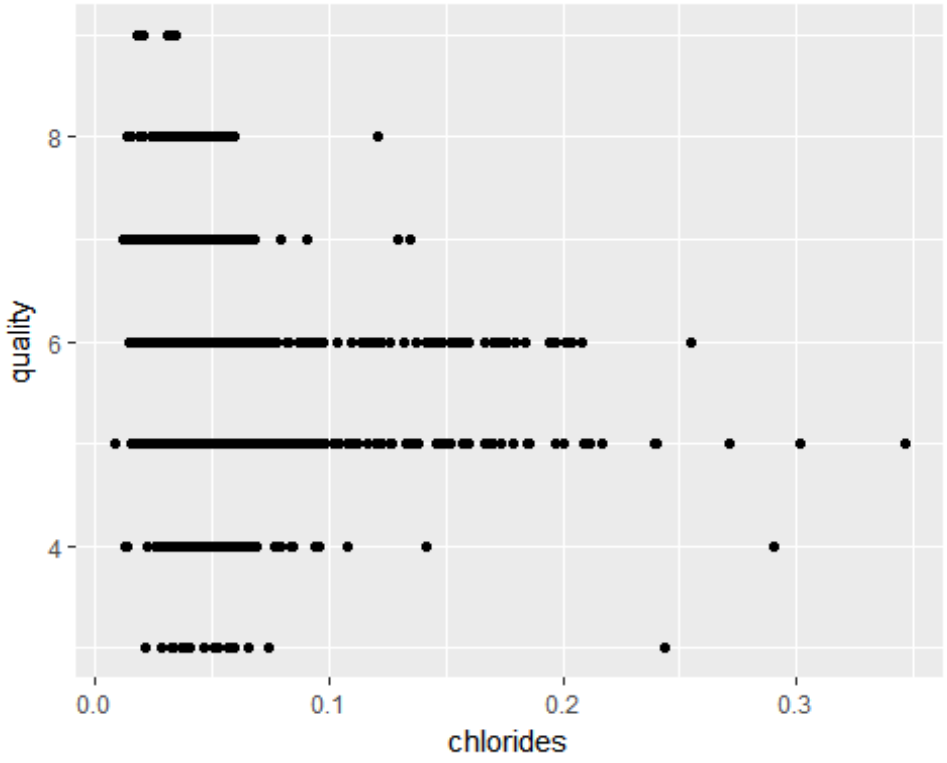


从上图可以看出，盐分与葡萄酒质量有一定相关性，相关系数达到-0.21，因此选择这两个变量进行分析。

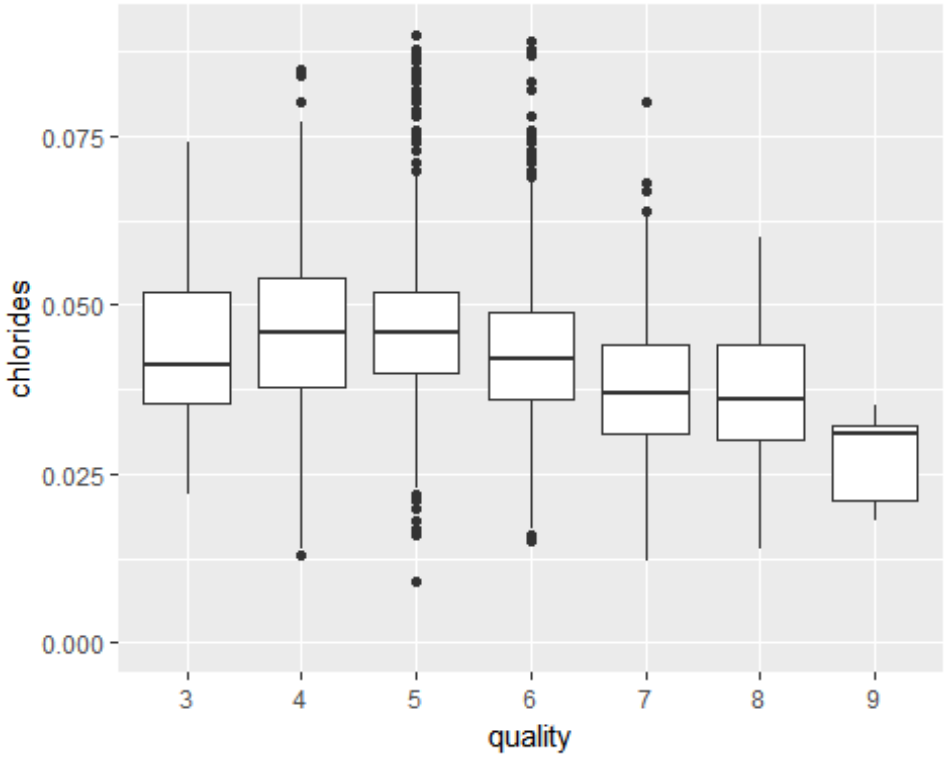


再由上图，密度和酒精度和葡萄酒品质有相关性，因此选择研究这两个变量。 接下来以盐分、密度和酒精度三个变量对葡萄酒品质变量依次做探索性分析。 同时可以发现，酒精度和密度相关性较强：根据常识，酒精的密度低于水，因此酒精度越高，酒的密度会较低，两个变量不相互独立，因此不作分析。

首先作散点图：



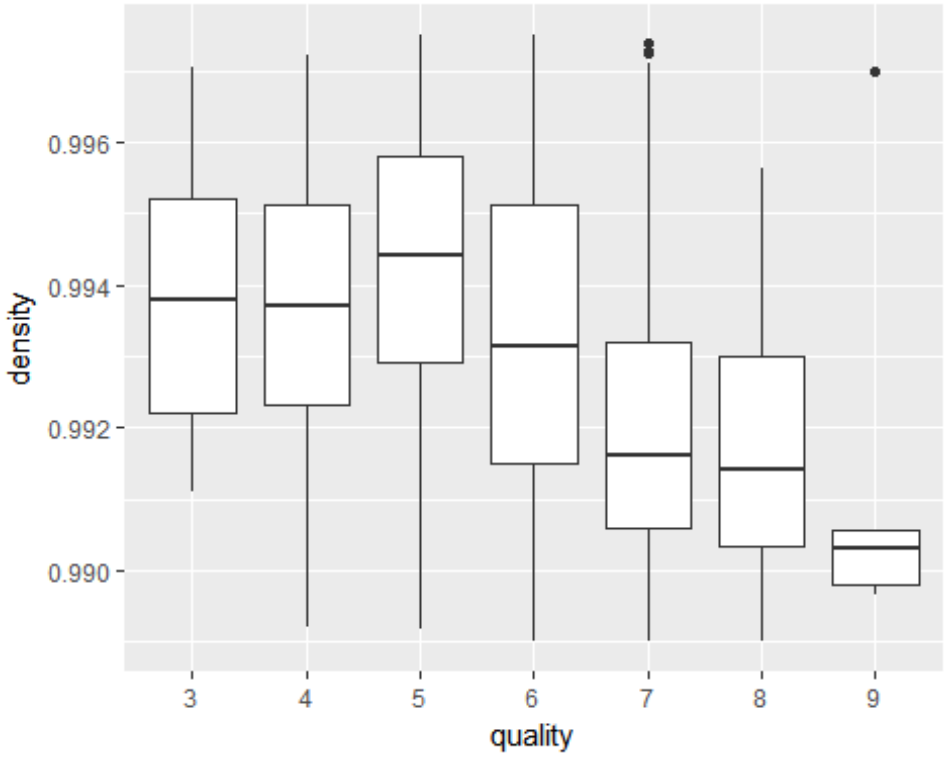
由于品质是整数并且有从低至高的顺序，因此可以将葡萄酒品质转化成因子变量，作箱线图分析：



由箱线图可以看出，品质较高的白葡萄酒盐度总体较低。相关统计数据如下：

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02200 0.03625 0.04100 0.05430 0.05400 0.24400
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0130  0.0380  0.0460  0.0501  0.0540  0.2900
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.04000 0.04700 0.05155 0.05300 0.34600
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01500 0.03600 0.04300 0.04522 0.04900 0.25500
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03819 0.04400 0.13500
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01400 0.03000 0.03600 0.03831 0.04400 0.12100
## ------------------------------------------------------------
## wine$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0180  0.0210  0.0310  0.0274  0.0320  0.0350
```

同理，对密度和葡萄酒品质作箱线图分析：

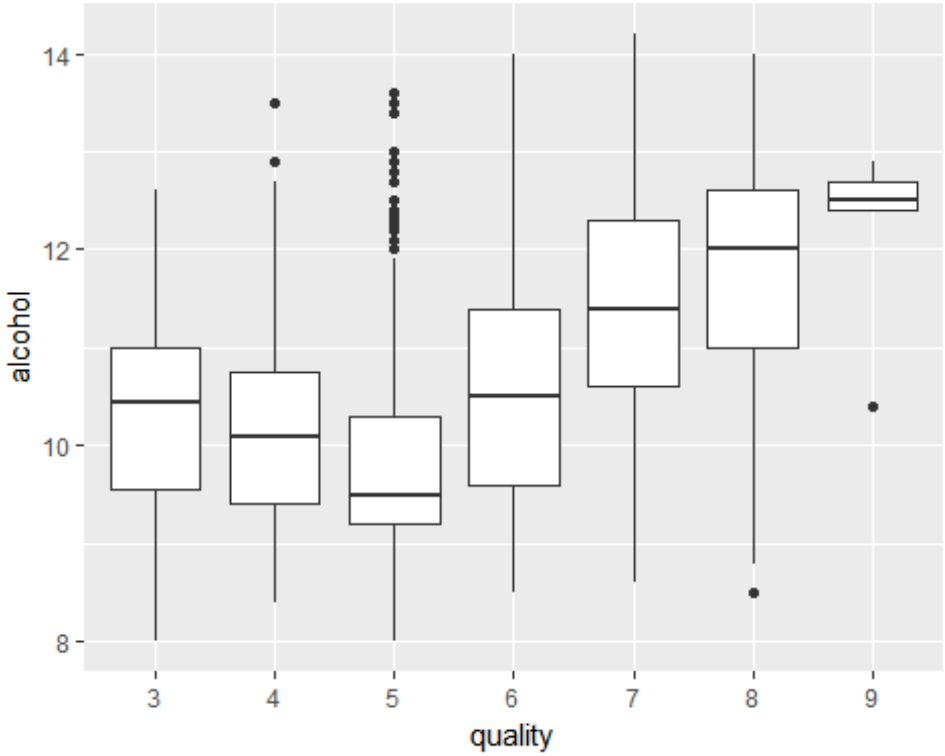


同样地，总体而言品质更高的葡萄酒密度会偏低。相关统计量如下：

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9911  0.9925  0.9944  0.9949  0.9969  1.0000
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9892  0.9926  0.9941  0.9943  0.9958  1.0000
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9933  0.9953  0.9953  0.9972  1.0020
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9906  0.9918  0.9925  0.9937  1.0000
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9903  0.9916  0.9922  0.9935  1.0010
## ------------------------------------------------------------
## wine$quality: 9
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9896  0.9898  0.9903  0.9915  0.9906  0.9970
```

对酒精度和葡萄酒作箱线图分析：



由图可以看出，品质好的葡萄酒酒精度大体也会偏高，相关统计量如下：

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.34   11.00   12.60
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
```

```
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
## wine$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   12.40   12.50   12.18   12.70   12.90
```

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

根据以上分析，可以初步得出以下推论：

- 葡萄酒的品质和盐度、酒精度和密度相关程度较高，与其他变量相关程度较低；
- 品质越好的葡萄酒总体而言盐度和密度会更低，酒精度会更高。

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

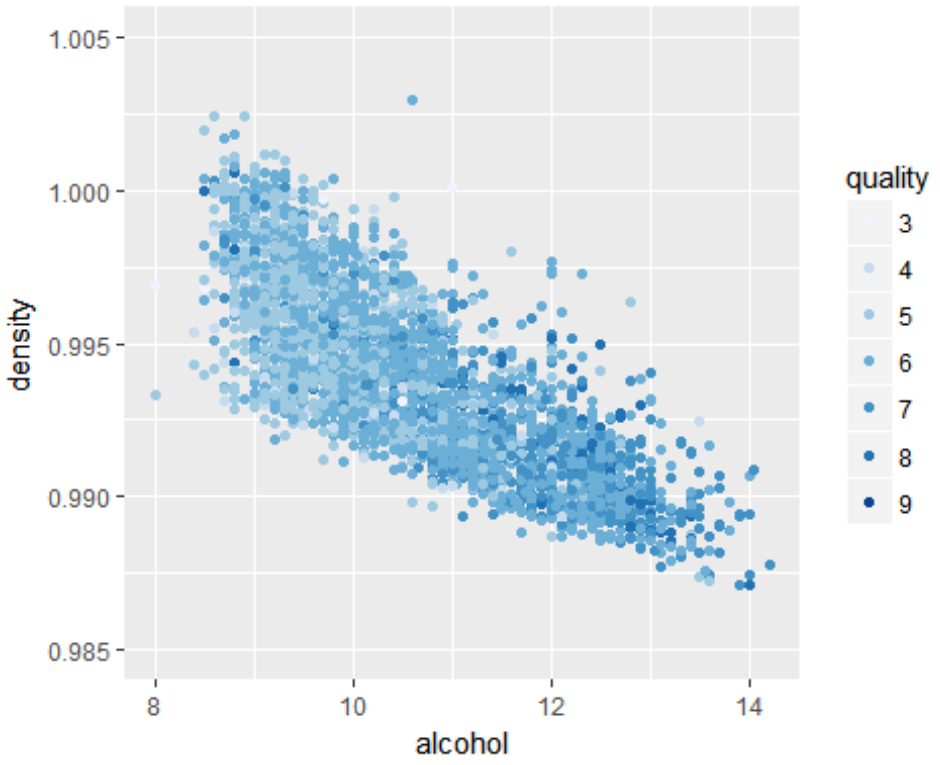
酒精度和密度相关程度高，其原因在于这两个变量是不相互独立的，某一方变化会导致另一方的变化。

What was the strongest relationship you found?

酒精度和密度的相关性是最强的。

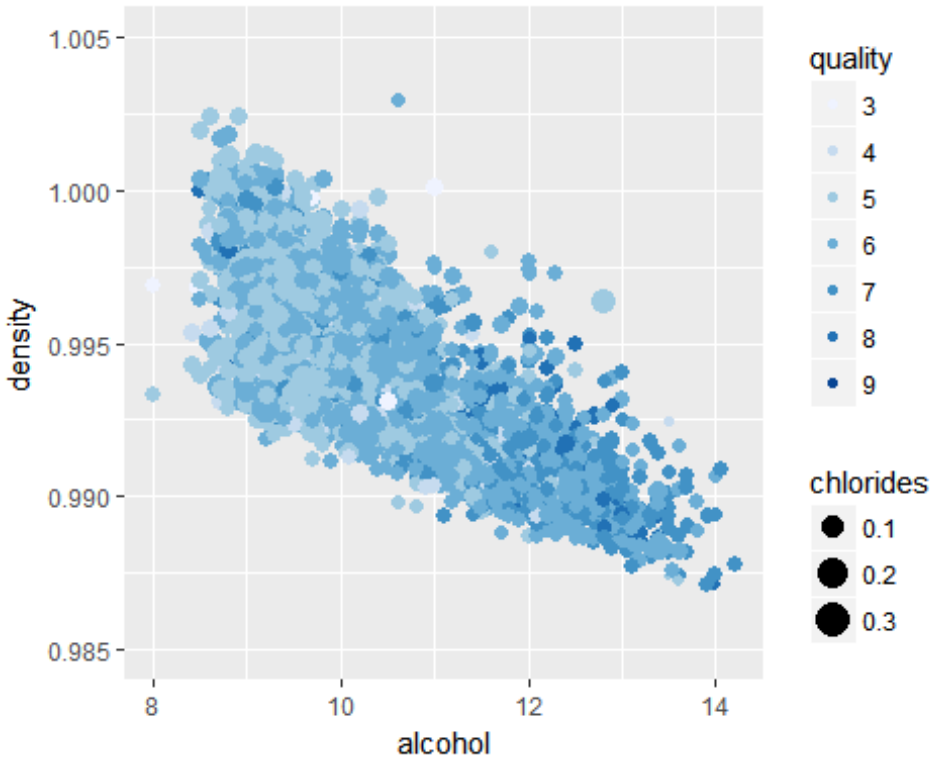
Multivariate Plots Section

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



观察散点图可以得知，品质较好的葡萄酒在散点图右下方聚集，因此可以显示出葡萄酒的品质和密度及酒精度的较强相关性。

进一步加入盐度变量：



可以看出，品质较高的葡萄酒，盐度相对偏小。

通过以上分析建立线性回归模型：

```
##
## Calls:
## m1: lm(formula = I(quality) ~ alcohol + chlorides, data = wine)
## m2: lm(formula = I(quality) ~ alcohol + chlorides + fixed.acidity +
##     log(volatile.acidity) + citric.acid + log(residual.sugar) +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates,
##     data = wine)
##
## ================================================
##                                m1         m2
## ------------------------------------------------
##   (Intercept)              2.861***   0.781*
##                           (0.116)    (0.347)
##   alcohol                  0.298***   0.361***
##                           (0.010)    (0.011)
##   chlorides               -2.471***  -0.984
##                           (0.558)    (0.539)
##   fixed.acidity                      -0.055***
##                                      (0.015)
##   log(volatile.acidity)              -0.615***
##                                      (0.034)
##   citric.acid                         0.016
##                                      (0.095)
```

```
##   log(residual.sugar)                0.155***
##                                     (0.014)
##   free.sulfur.dioxide                0.005***
##                                     (0.001)
##   total.sulfur.dioxide              -0.001*
##                                     (0.000)
##   pH                                 0.130
##                                     (0.082)
##   sulphates                          0.417***
##                                     (0.097)
## -----------------------------------------------
##   R-squared                  0.2        0.3
##   adj. R-squared             0.2        0.3
##   sigma                      0.8        0.8
##   F                        585.2      188.5
##   p                          0.0        0.0
##   Log-likelihood         -5829.6    -5555.8
##   Deviance                3099.8     2771.9
##   AIC                    11667.2    11135.5
##   BIC                    11693.2    11213.5
##   N                       4898       4898
## ===============================================
```

由图可见，酒精度和盐度可以解释 20%的葡萄酒品质，而加上其他的变量也仅能解释 30%。

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

变量间的关系： 品质好的葡萄酒密度更低，酒精度更高，盐度也更低。在多变量相关分析时可以明显看出这一点。

Were there any interesting or surprising interactions between features?

没有。

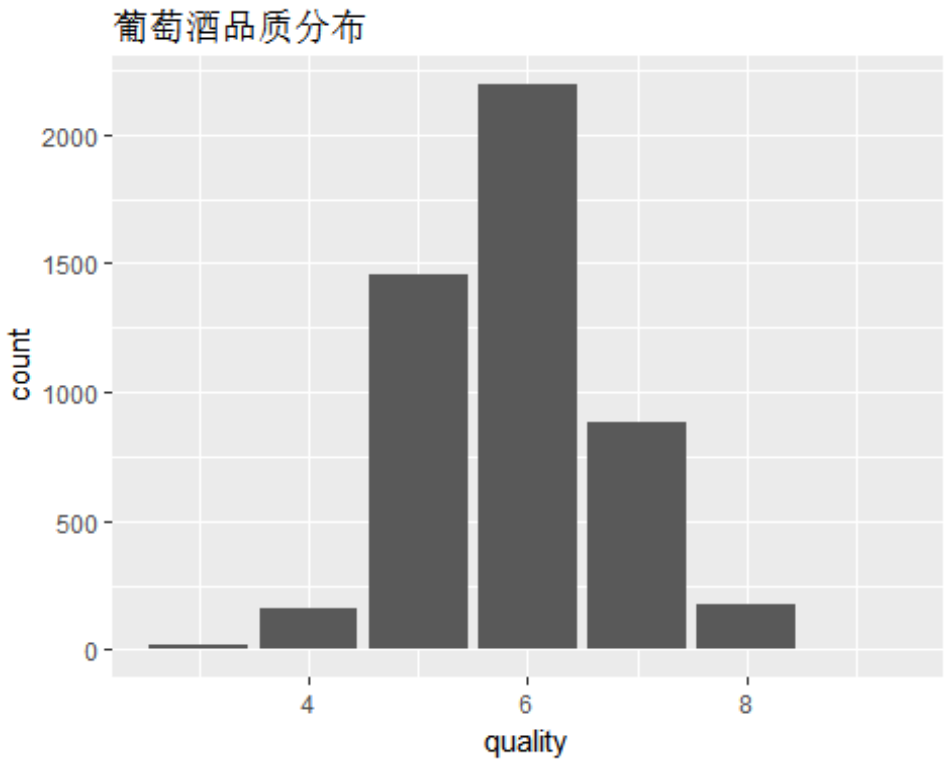
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

建立了多元线性回归模型，但该模型的拟合效果并不好。这可能与因变量为因子变量或者自变量数据处理得不够充分有关。但尽管如此，该模型的相关系数依然有 30%，具有一定的解释力。

Final Plots and Summary

Plot One

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

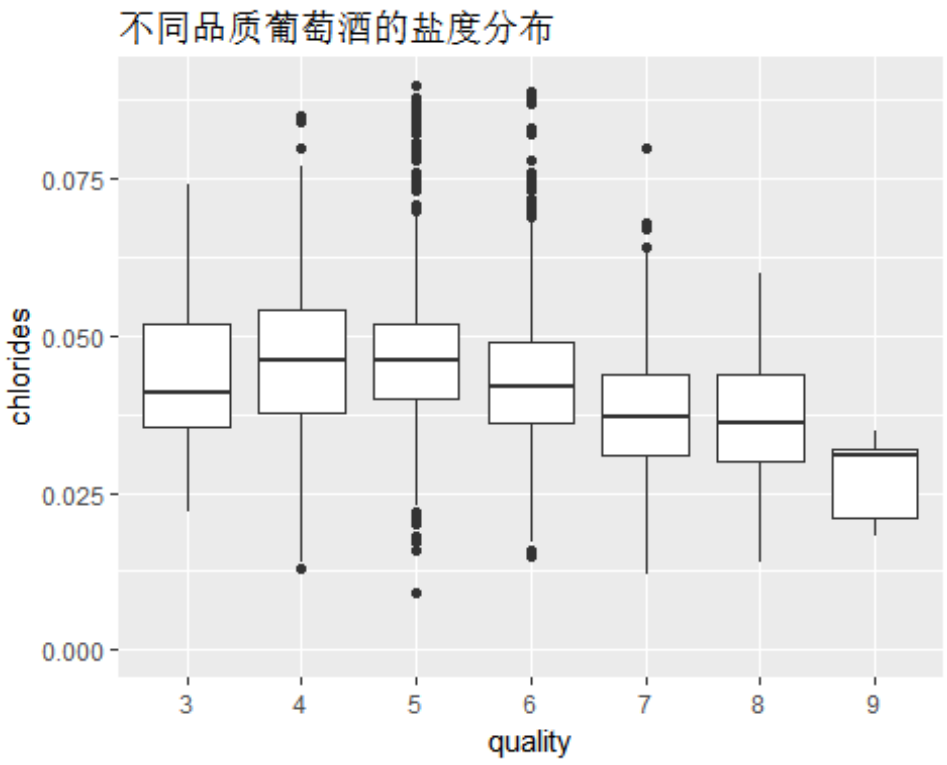


Description One

该图显示了葡萄酒的评分分布。可以看到，虽然评分范围为 0-10，但品质评分为 0、1、2、10 的葡萄酒是没有的，绝大部分都在 5-7 之间，形成明显的集中趋势。

Plot Two

```
## Warning: Removed 135 rows containing non-finite values (stat_boxplo
t).
```

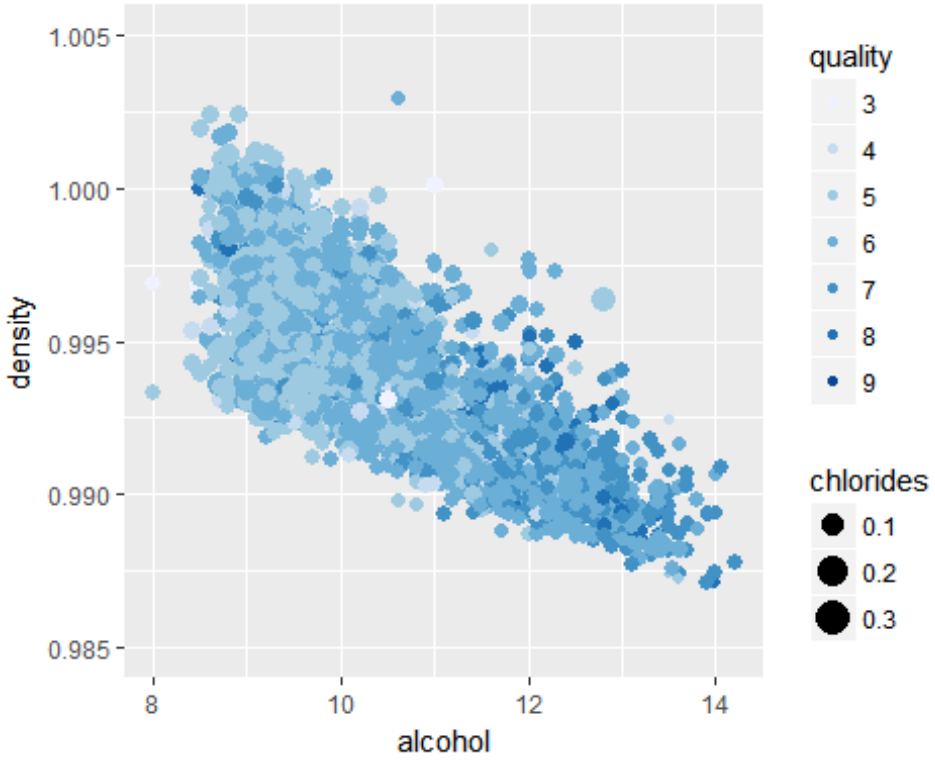


Description Two

该图显示出葡萄酒品质和盐度的关系。整体而言，盐度越低的葡萄酒品质越好。

Plot Three

```
## Warning: Ignoring unknown aesthetics: type

## Warning: Removed 3 rows containing missing values (geom_point).
```




```
## $title
## [1] "葡萄酒品质和盐度、密度、酒精度的关系"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

Description Three

这幅图探索了盐度、密度、酒精度和葡萄酒品质 4 个变量的关系。可以看到，品质为 5 的葡萄酒大部分集中于左上方，点较大，品质为 7 的葡萄酒则集中于右下方，点较小。这充分说明了葡萄酒品质的影响因素：盐度越低、酒精度越高、密度越小，则品质越好。

Reflection

本次探索性数据分析采用的是近 5000 份白葡萄酒样本数据。本数据的分析导向也十分明显，即“影响白葡萄酒品质的因素有哪些”，所呈现的变量几乎都是为这个目的而服务的。因此这也限制了一部分数据分析工作的开展，但同时有利于明确分析导向。

通过对单变量、双变量和多变量分析，确定了与葡萄酒品质相关性较强的三个因素即酒精度、密度和盐度，与其他变量的相关性则较小。通过绘制图表，进行描述性统计量的计算，并建立多元线性回归模型，初步分析了影响葡萄酒质量的各个因素。

这次分析的不足也是十分明显的。比如我将评分作为数值变量进行线性回归，方法并不科学，但由于统计学和机器学习基础薄弱，还没有办法通过更好的分类方法建立模型。同时探究的变量数量也比较少，分析还不够彻底。