# EDA with R: Red Wine Quality Dataset

by Benjamin Soukiassian

Exploratory Data Analysis with R on a Red Wine quality dataset.

Part of the Data Analysis Nanodegree by Udacity.

# 1.Exploration

Let's load our csv file and have a look at the dimension of the dataset

```
library(ggplot2)
wine <- read.csv('wineQualityReds.csv')
dim(wine)
```

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

We have 1599 observations of 13 variables. Let's now have a look at the structure and variables types of the dataset.

```
str(wine)
```

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

All of the relevant variables are numerical except the quality variable which are integers. By looking at the dataset reference file, we learn that the quality score is graded within a 0 to 10 interval. We can deduce that quality score is a discrete variable, given its integer type.

Let's take a look at the summary of the data for the quality variable.

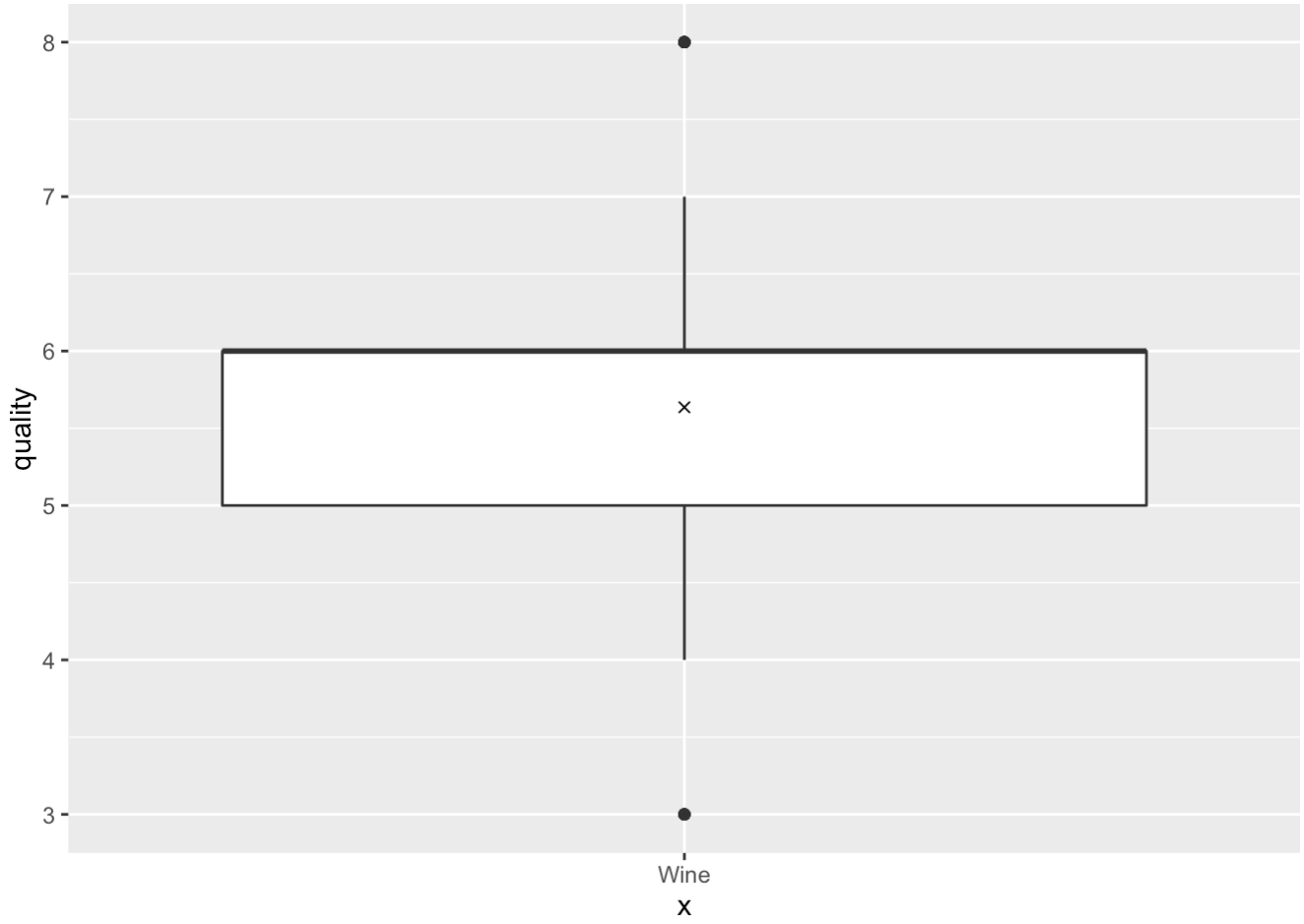
```
summary(wine$quality)
```

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

Our min quality score is 3 while our max is 8, which is a bit suprising given the 1 to 10 interval. we have a mean of 5.636 and median at 6.

Let take a look at a box plot:

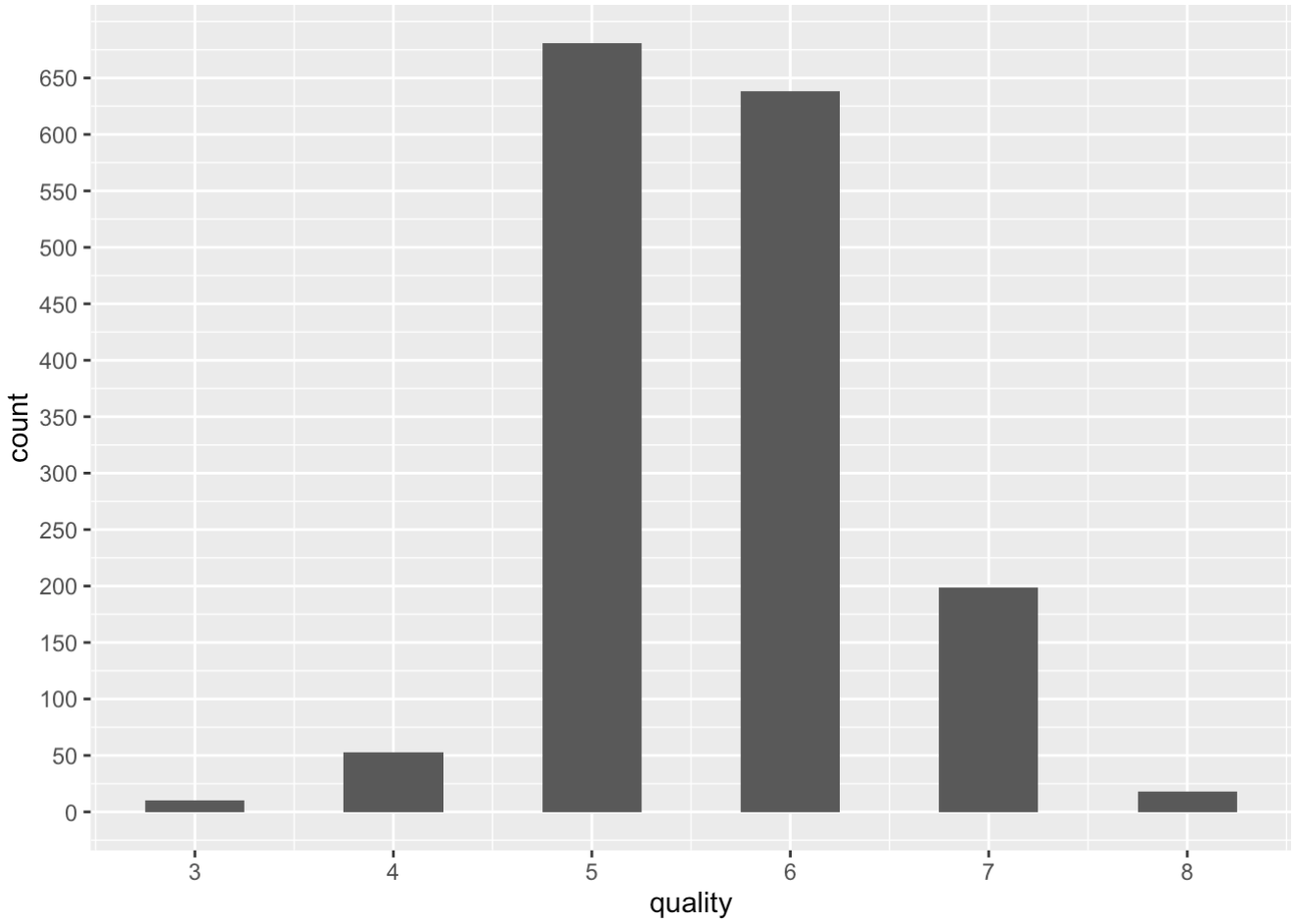
```
ggplot(aes(x="Wine", y=quality), data=wine) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 4)
```




50% of the bottles are graded between 5 and 6, which considering the 1 to 10 scale seem quite concentrated. We have a few outliers at 3 and 8 (min and max values) but most of the values fall between 4 and 7.

To get a better sense of this concentration Let's plot the data as a histogram of frequencies.

```
ggplot(aes(x=quality), data=wine) +
  geom_histogram(binwidth = 0.5) +
  scale_x_continuous(breaks = seq(1,10,1)) +
  scale_y_continuous(breaks = seq(0, 650, 50))
```




The frequency distribution is indeed concentrated around 5 and 6. It also appear to be a little skewed toward the right (upper quality).

The obvious interesting question regarding this dataset is going to be to try to find which variables are correlated with the quality notation. This question imply the assumption that this notation is trustworthy.

Let's go back to our dataset attributes.

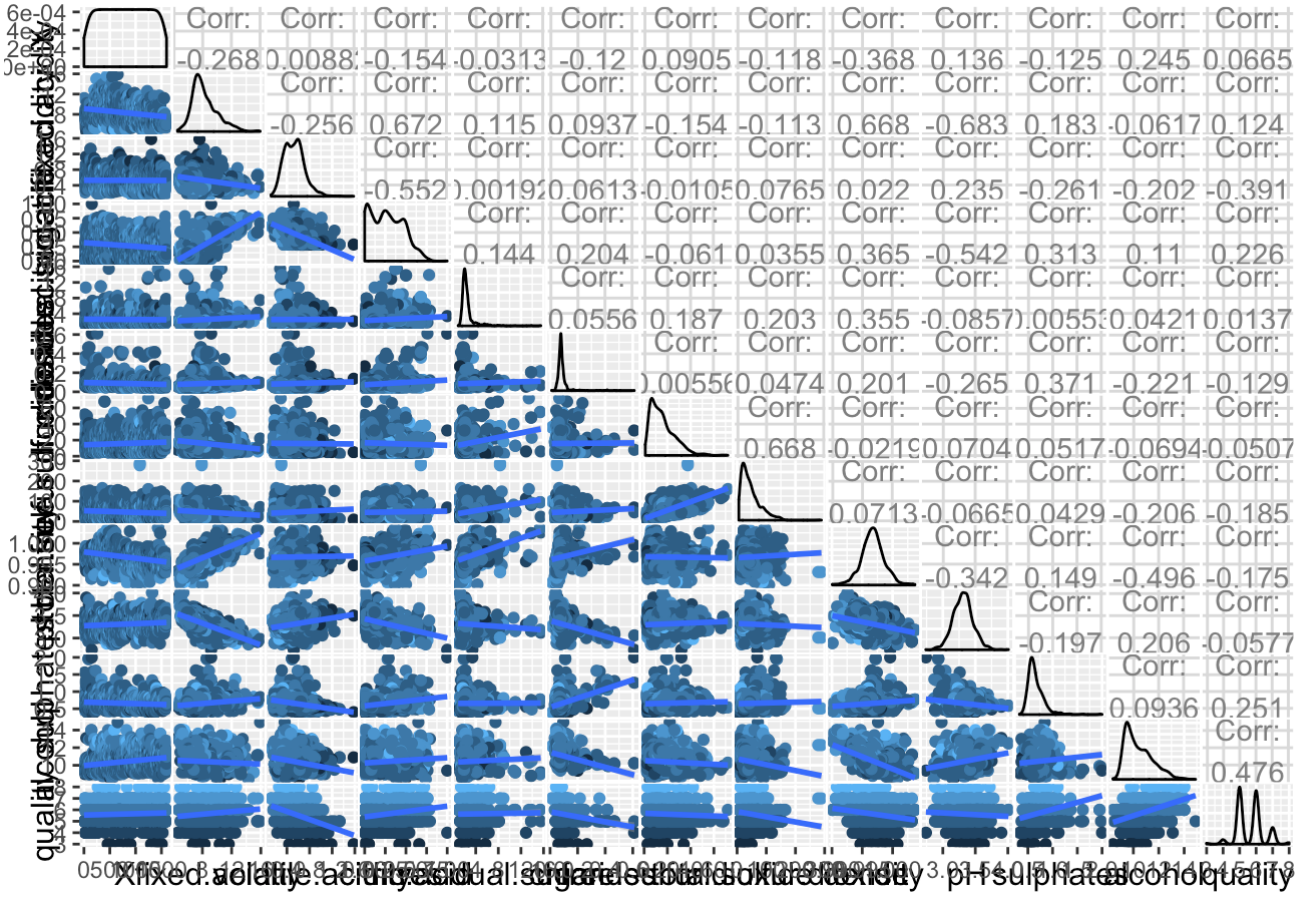
```
str(wine)
```

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

The dataset provides us with a number of chemical elements which were obtained through physicochemical tests. We have for instance the percent of alcohol, the density, the pH or different measures of acidity.

Instead of computing each pair of variables separately, let's create a scatterplot matrix to get a sense of which attributes are worth exploring further.

```
library(GGally)
ggpairs(wine,
        lower = list(
          continuous = "smooth",
          combo = "facetdensity",
          mapping = aes(color=quality)
          )
        )
```



**Bigger version of the image (ggpair-matrix.png)**

Thanks to the scatterplot matrix, we are able to get a sense of the relationship between each pair of variables. One of the first thing that visually jumps out is that there do not seem to have any "very strong" relationship between pairs of variables. But a few pairs appear to have interesting relationships nonetheless. For instance, by looking at the lower part of the plots, we see that citric.acid appear to have a positive relationship with fixed.acidity, which we can assert by looking at the corresponding correlation coefficient of 0.672 (Strong relationship). However, citric.acid seems to have a negative relationship with volatile acidity. Corr: -0.552 (Moderate).

We also see that pH has a negative relationship with both fixed.acidity (-0.683) and citric.acid (-0.542) which makes sense because a low pH means more acidity.

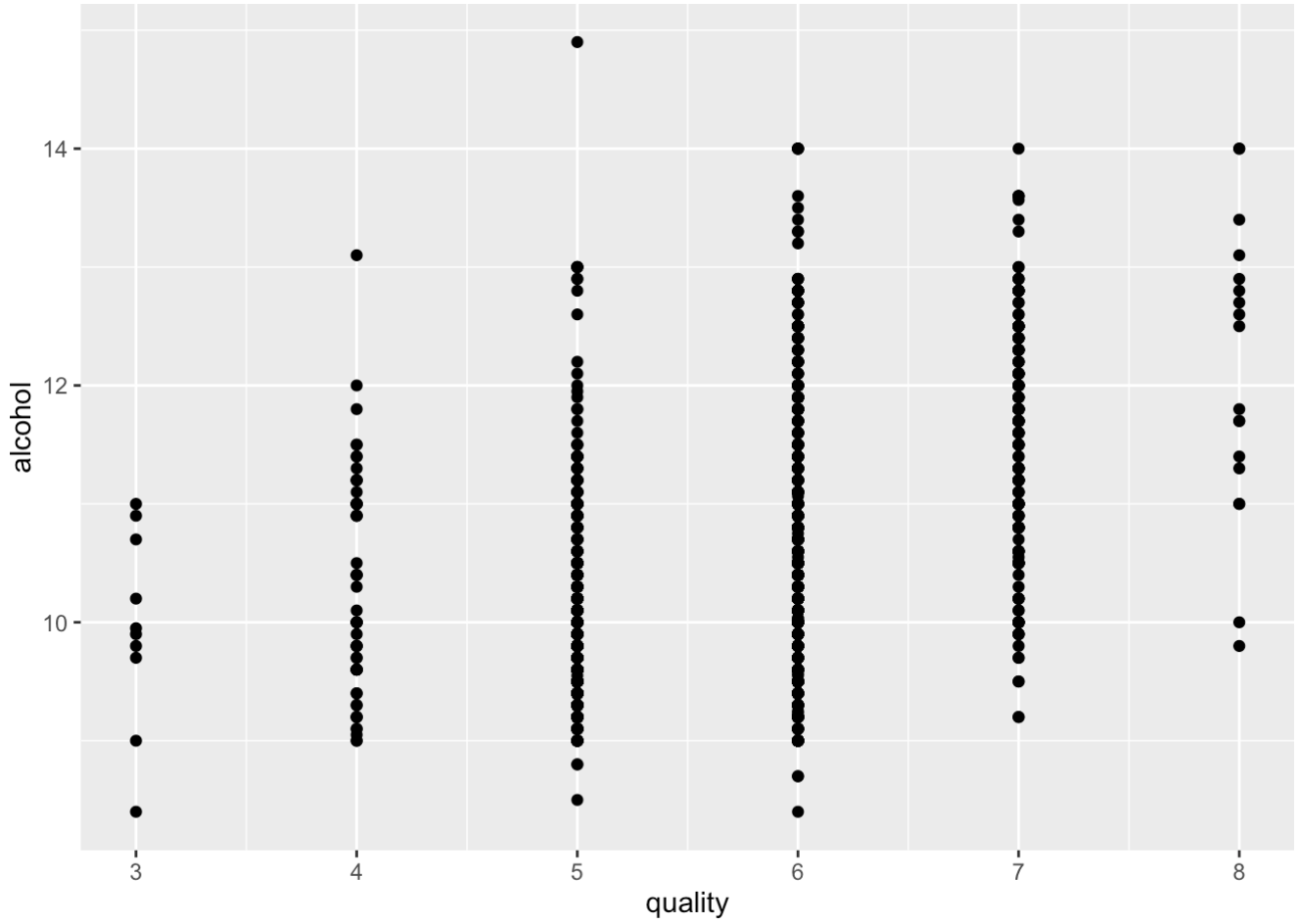
We can see thanks to the diagonal frequency polygons plots that both density and pH values seem to follow an approximatively normal distribution while most of the other chemical attributes are positively skewed.

But let's focus on the plots between the different chemical attributes and quality. First, by looking at the correlation coefficients at the far right, we see that alcohol degree is the most attribute most correlated with quality, which I must admit I find quite surprising. The second is volatile acidity. with a negative correlation of -0.391

# Alcohol

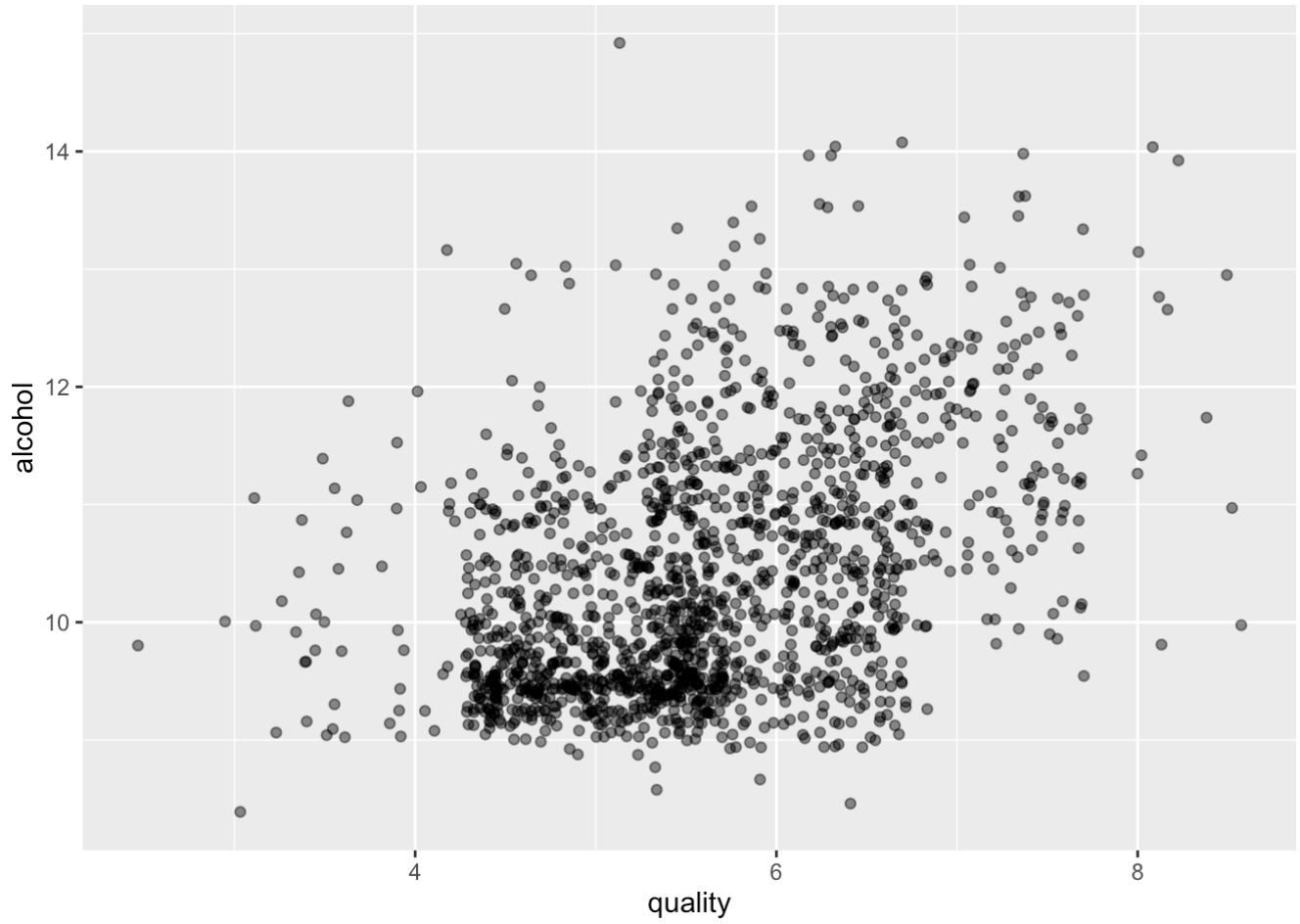
Let's start by plotting percent of alcohol against quality.

```
ggplot(aes(quality, alcohol), data = wine) +
  geom_point()
```



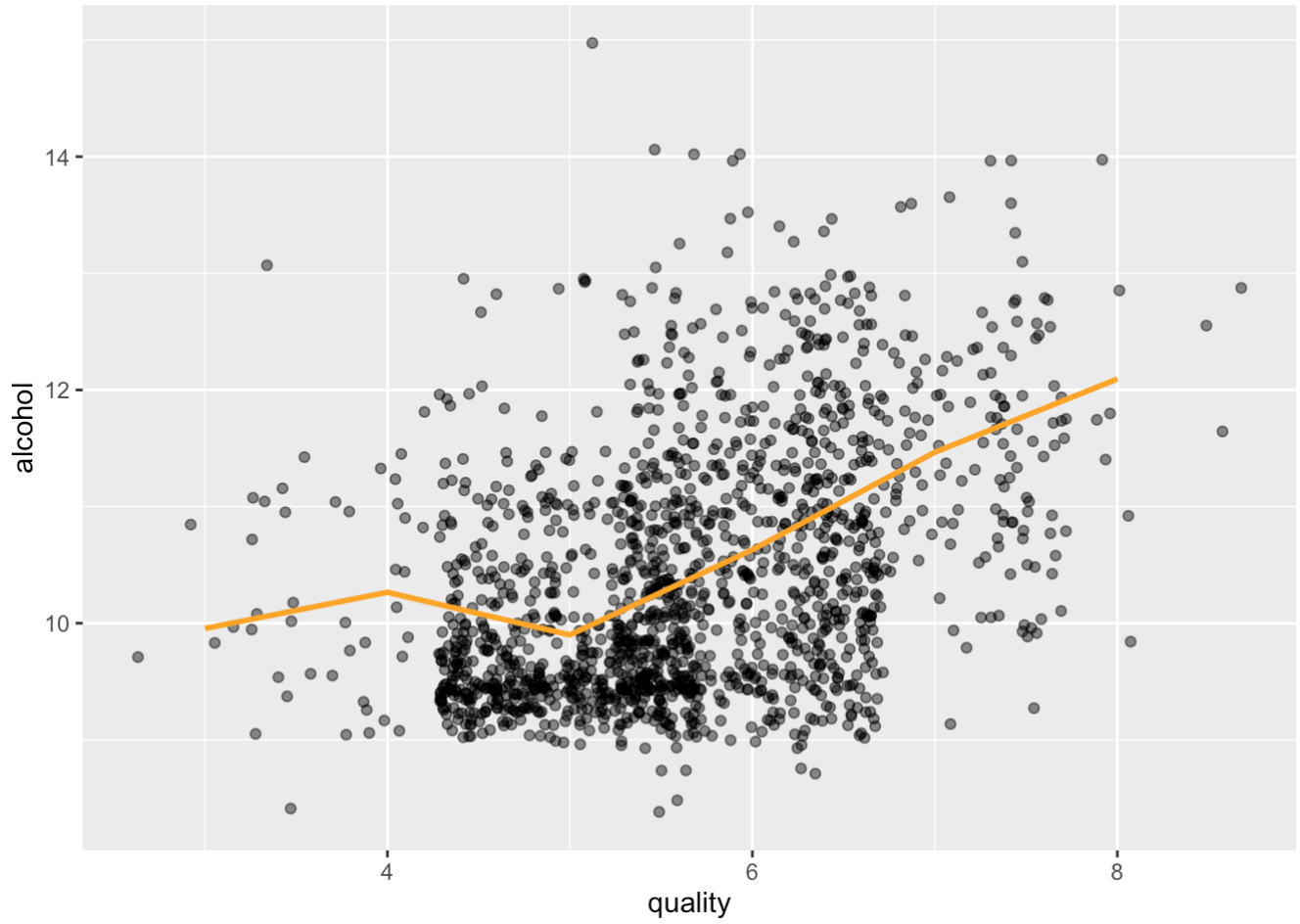

It is quite hard to see anything with the discrete quality data. Let's add some noise to the data with jitter.

```
ggplot(aes(quality, alcohol), data = wine) +
  geom_jitter(alpha=1/2, width = 1.8, height = 0.2)
```




There is definitely not a negative relationship between both variables, There could be a light positive relationship but it's still hard to grasp. Let's add a mean line.

```
ggplot(aes(quality, alcohol), data = wine) +
  geom_jitter(alpha=1/2,  width = 1.8, height = 0.2) +
  stat_summary(fun.y = mean, color="orange", geom="line", size = 1)
```




The mean variable shows a positive relationship, but by looking at the raw data the relationship do not seem very strong. We can compute a Pearson correlation coefficient between the two variables to quantify this strength.

```
cor.test(wine$quality, wine$alcohol, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$alcohol
## 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
```

The correlation coefficient is 0.4762. Using the guide that Evans (1996) suggest for different r intervals this value is considered to be moderate.

So the percent of alcohol is moderately correlated to quality, but it probably does not alone account for the total variation in wine quality. Thankfully, we still have other attributes to look at.

# Volatile acidity

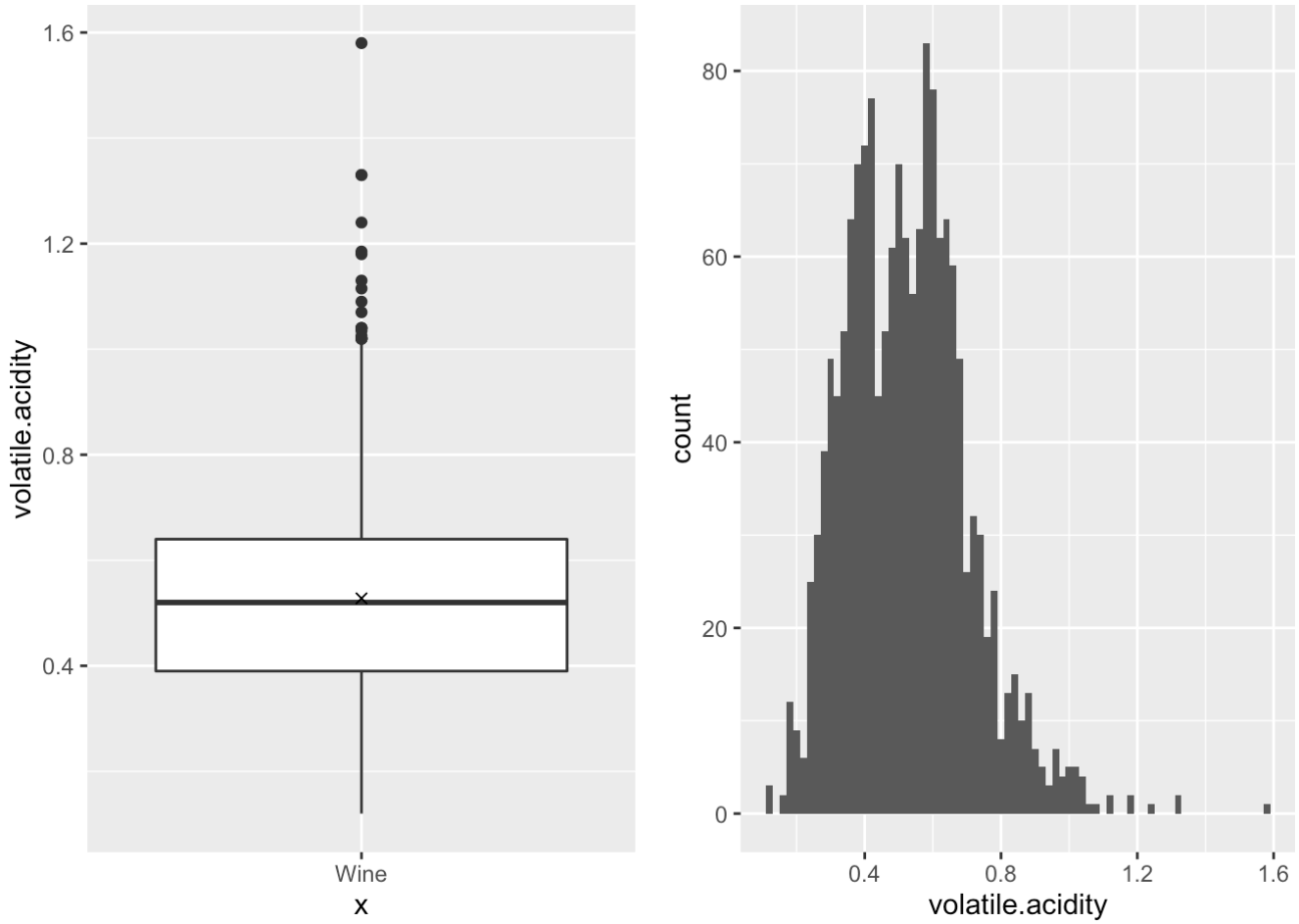
Volatile acidity (VA) is a measure of the wine's volatile (or gaseous) acids. The primary volatile acid in wine is acetic acid, which is also the primary acid associated with the smell and taste of vinegar. So it is is generally considered to be a spoilage product but apparently some winemakers seek a low or barely detectible level to

add to the perceived complexity of a wine.

```
library(gridExtra)
summary(wine$volatile.acidity)
```

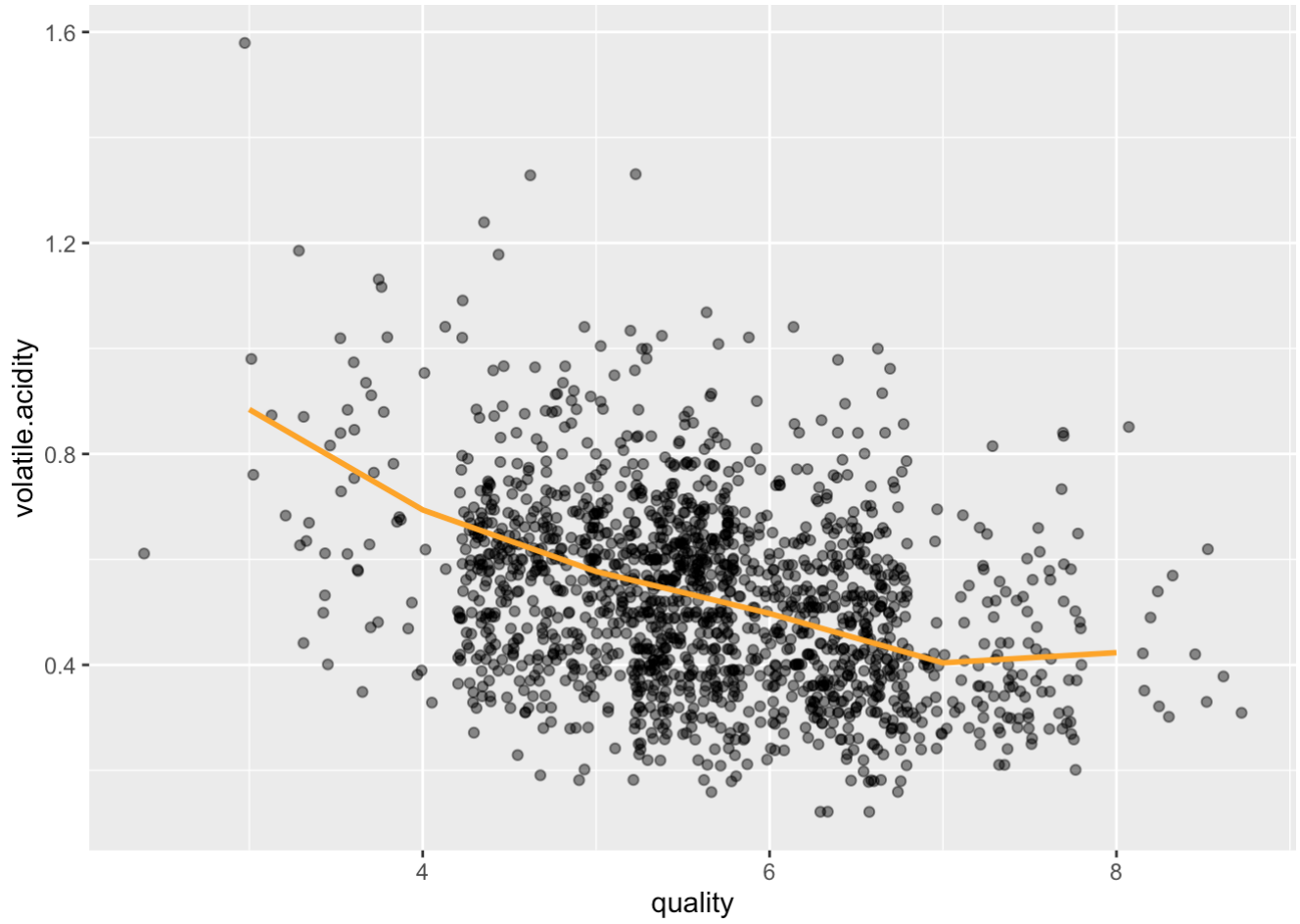
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

```
plot1 <- ggplot(aes(x="Wine", y=volatile.acidity), data=wine) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 4)

plot2 <- ggplot(aes(volatile.acidity), data=wine) +
  geom_histogram(binwidth = 0.02)

grid.arrange(plot1, plot2, ncol=2)
```




The distribution is positively skewed and has a few outliers. Let's take a look at the relationship with quality, also adding the same mean line previously used.

```
ggplot(aes(quality, volatile.acidity), data = wine) +
  geom_jitter(alpha=1/2, width = 2) +
  stat_summary(fun.y=mean, colour="orange", geom="line", size = 1)
```




We can definitely see a negative relationship between the variables, but as with alcohol, it does not seem really strong.

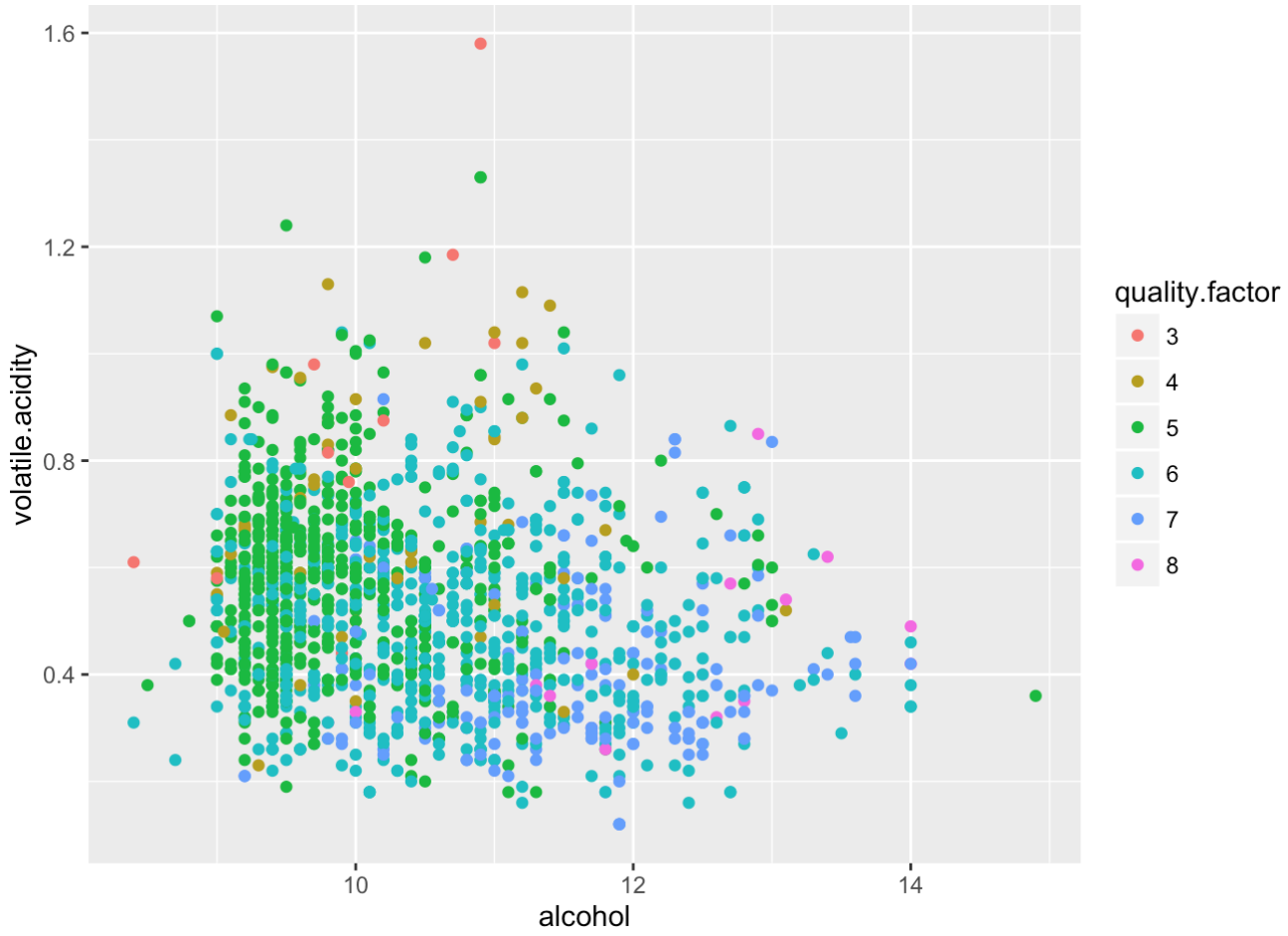
```
cor.test(wine$quality, wine$volatile.acidity, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$volatile.acidity
## 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
```

And indeed the -0.39 Pearson Correlation coefficient is considered weak, nearly moderate.

But since alcohol and volatile acidity are still the variables with the biggest relationship to quality, let's try to plot them againts each other while adding quality as a third variable.

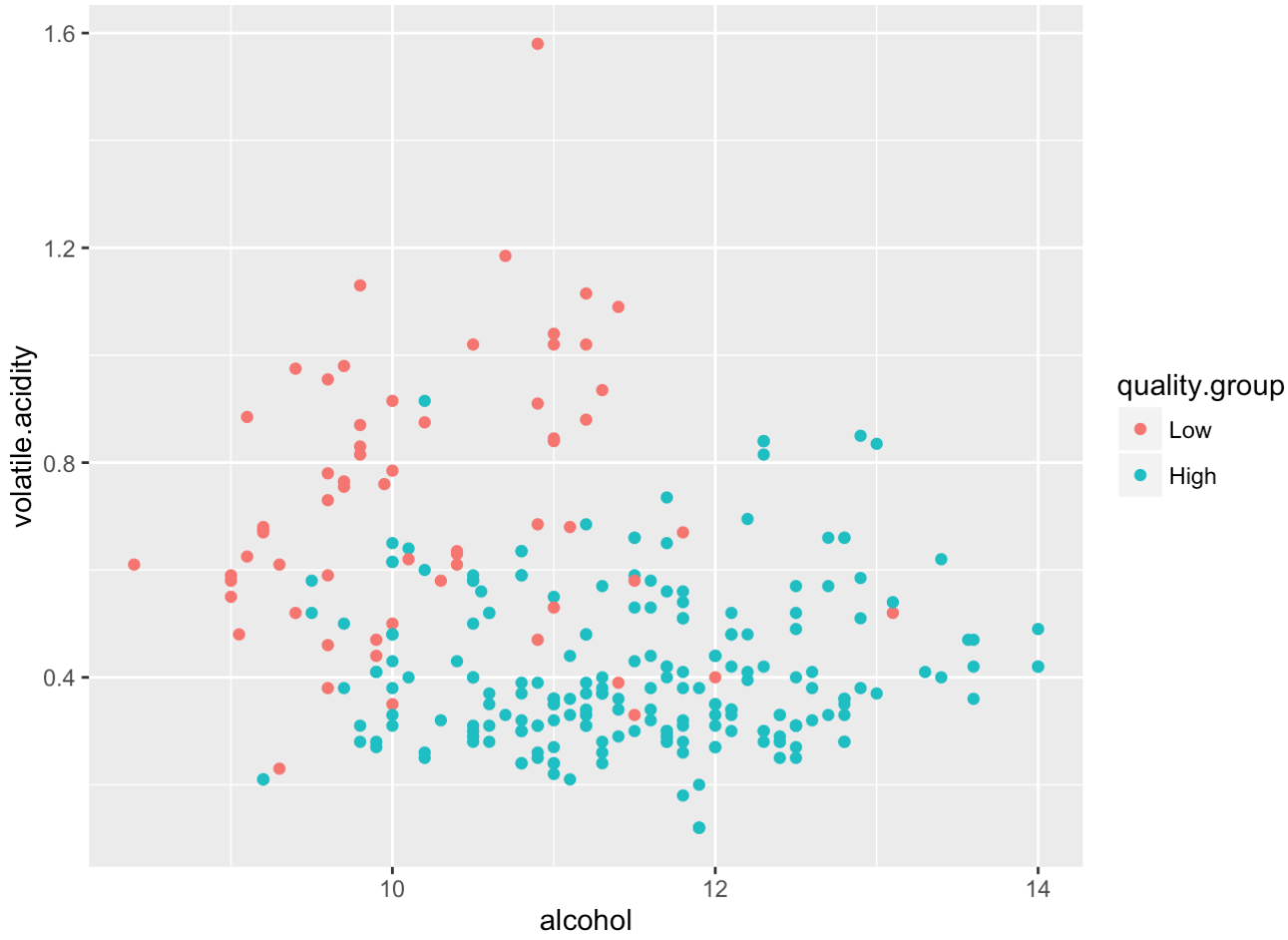
```
wine$quality.factor <- as.factor(wine$quality)
ggplot(aes(alcohol, volatile.acidity), data = wine) +
  geom_point(aes(color=quality.factor))
```




The best wines appear to be mostly located in the lower right of the plot and the worst higher on the left. But there are so many middle range wine (graded 5 and 6, colored green and blue) with quite a important dispersion that it's hard to see clearly the wines graded higher or lower on the quality scale.

Let's divide our quality notation data into three groups : Low, Medium and High and plot only the Low and High groups.

```
wine$quality.group <- cut(wine$quality,  breaks=c(3,4,6,8),
                          labels=c("Low","Medium","High"),
                          include.lowest=TRUE)

low_high_wines <- subset(wine, quality.group %in% c('Low','High'))

ggplot(aes(alcohol, volatile.acidity),
       data = low_high_wines) +
  geom_point(aes(color=quality.group))
```

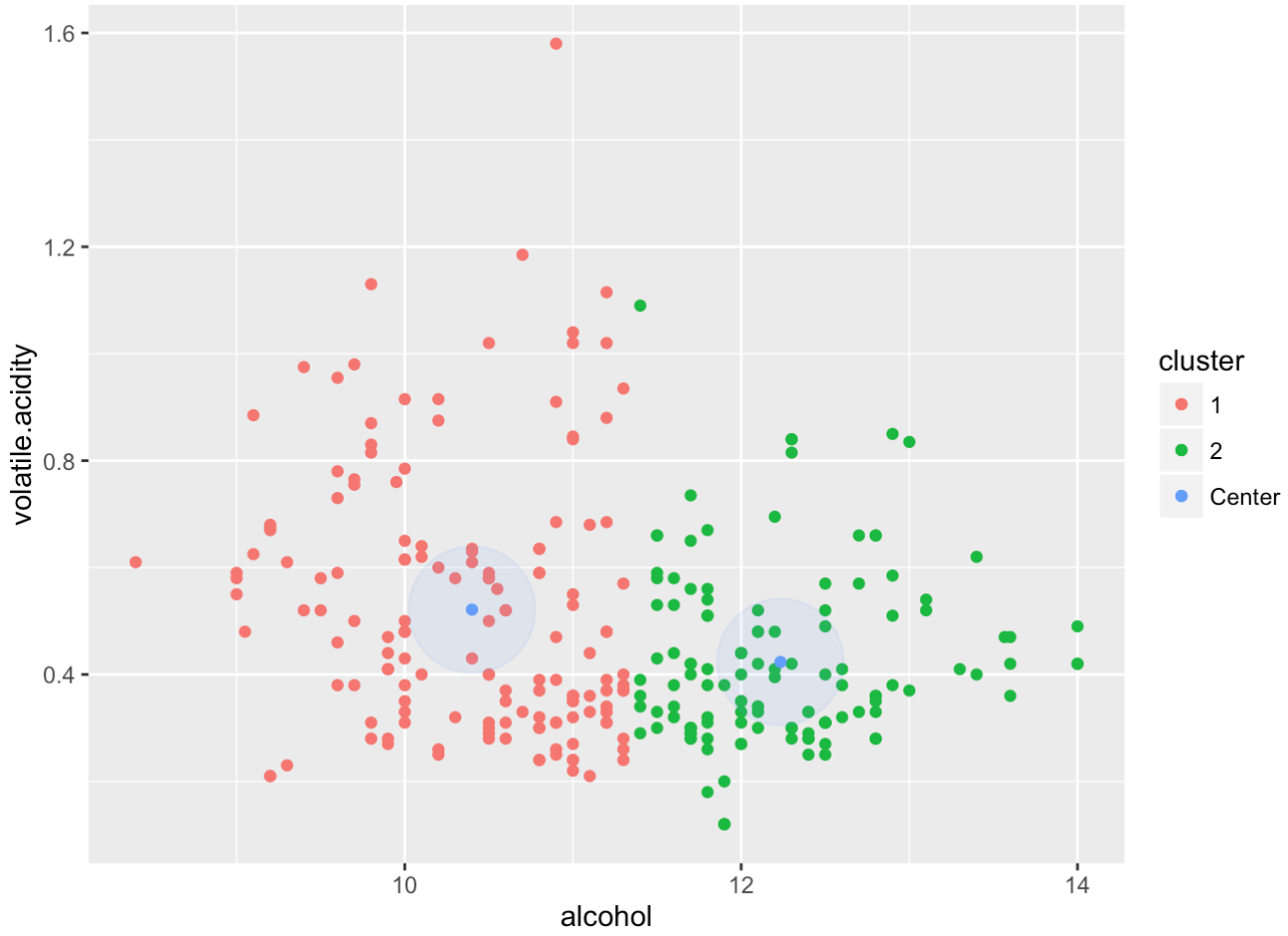



The difference is much more visible but the two data groups are clearly overlapping.

## Volatile acidity: Bonus

Althought in practice it doesn't make much sense since the data is already labelled and overlapping, I wonder how an unsupervised clustering algorithm like K-mean would cluster the Low-High data.

```
m = as.matrix(cbind(low_high_wines$alcohol, low_high_wines$volatile.acidity), ncol=2)

cl=(kmeans(m, 2, iter.max=1, nstart=1))
low_high_wines$cluster=factor(cl$cluster)
centers=as.data.frame(cl$centers)

ggplot(data=low_high_wines, aes(alcohol, volatile.acidity)) +
 geom_point(aes(color=cluster)) +
 geom_point(data=centers, aes(x=V1,y=V2, color="Center")) +
 geom_point(data=centers, aes(x=V1,y=V2, color="Center"), size=22, alpha=.1,
            show.legend = FALSE)
```

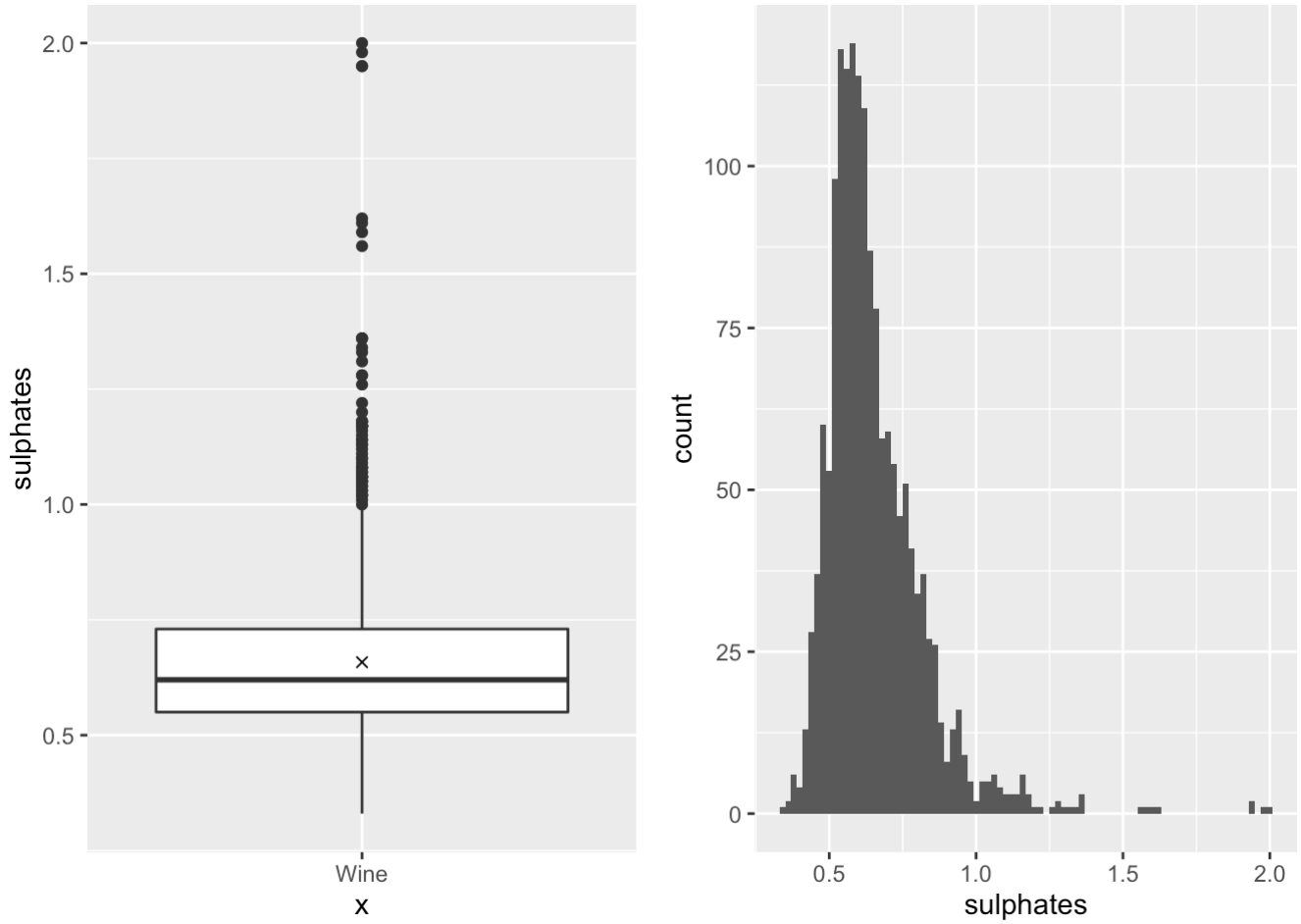



It splits the data vertically at about 11.5% on the alcohol scale.

# Sulphates

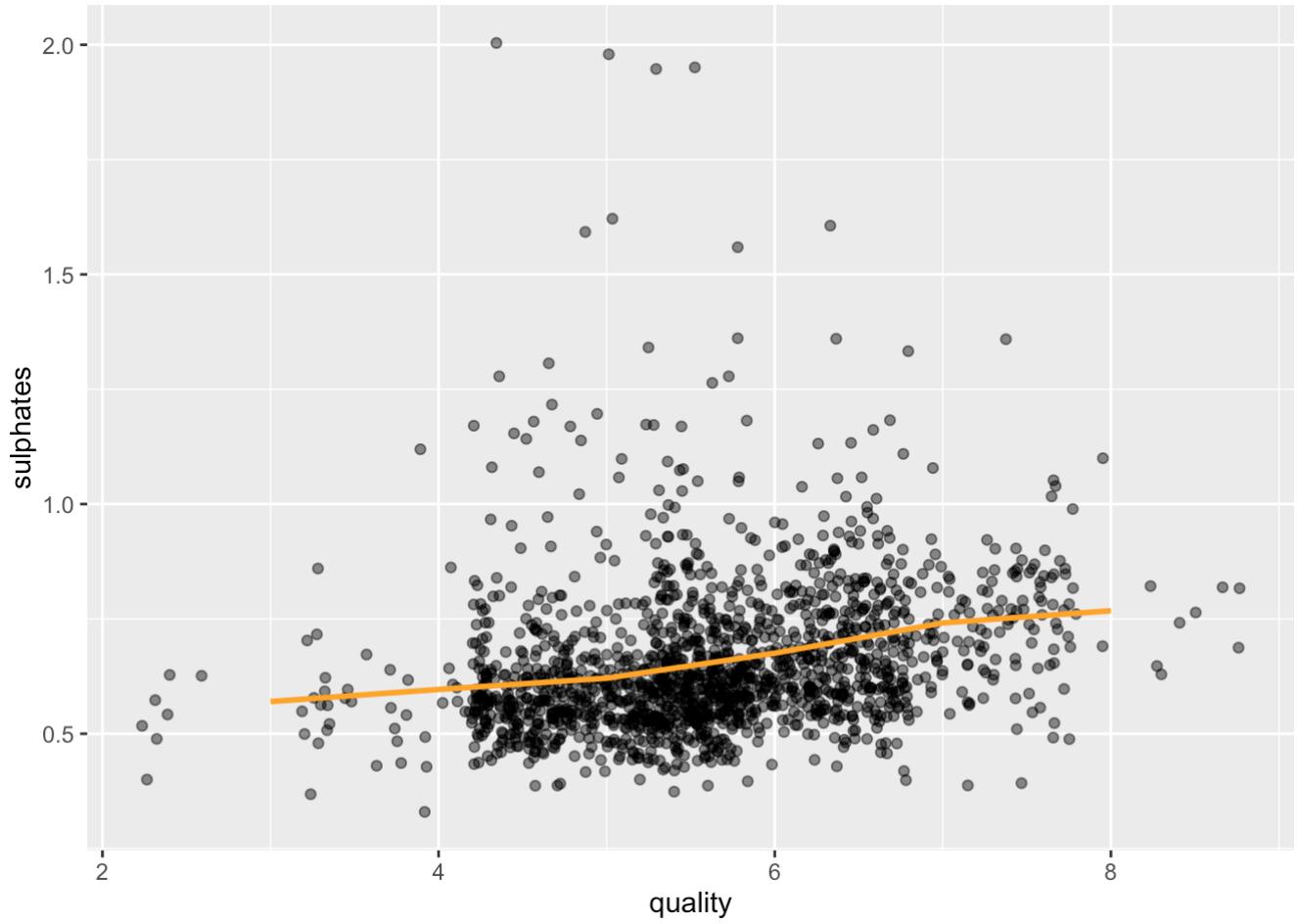
Sulphates is the third variable most correlated with quality.

```
library(gridExtra)
plot1 <- ggplot(aes(x="Wine", y=sulphates), data=wine) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 4)

plot2 <- ggplot(aes(sulphates), data=wine) +
  geom_histogram(binwidth = 0.02)

grid.arrange(plot1, plot2, ncol=2)
```




The sulphates distribution is positively skewed and even more concerned by outliers than Volatile Acidity or Alcohol.

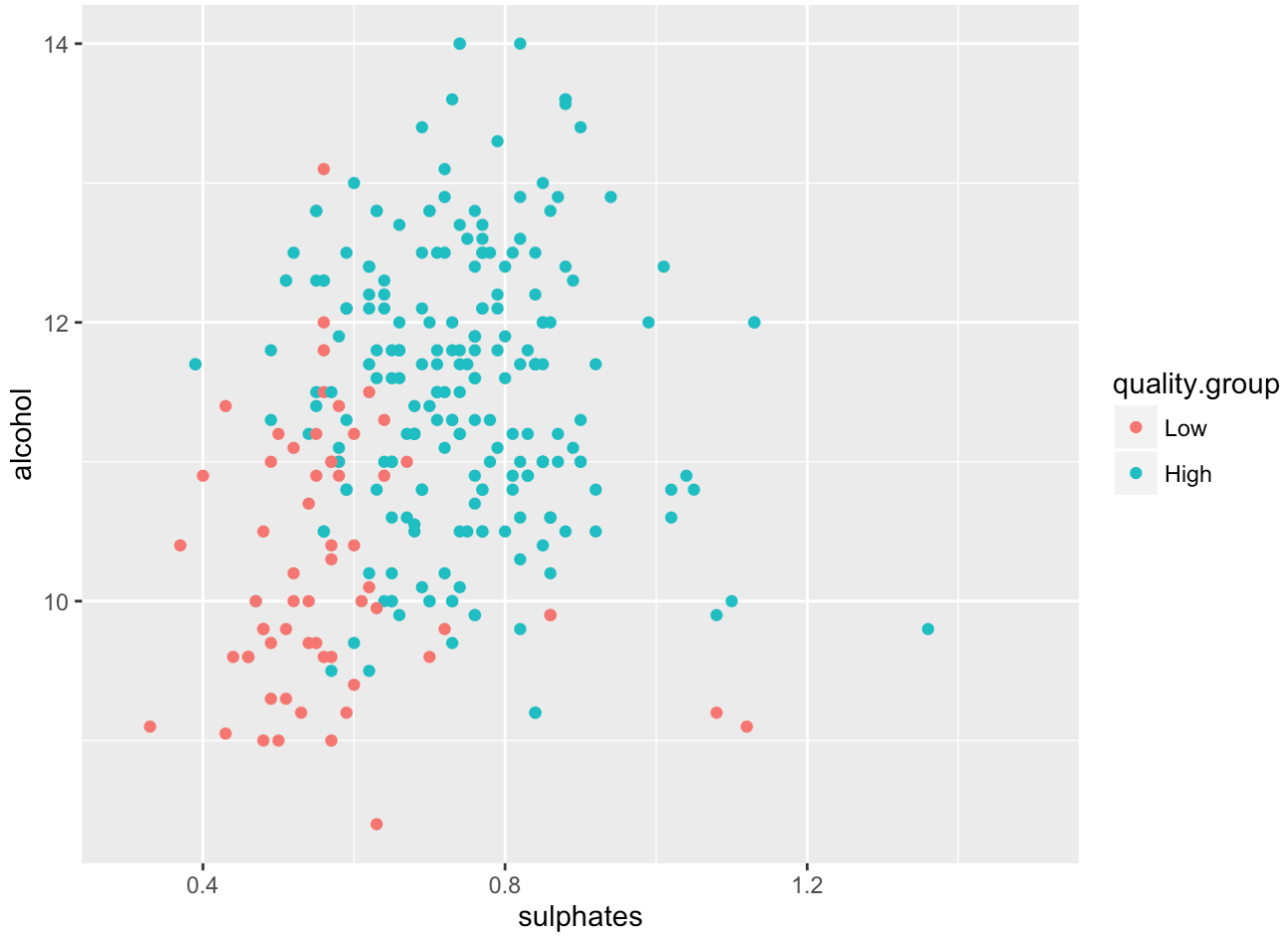
```
ggplot(aes(quality, sulphates), data=wine) +
    geom_jitter(alpha=1/2, width = 2) +
    stat_summary(fun.y=mean, colour="orange", geom="line", size = 1)
```




We can see a very light positive relationship between sulphates and quality, relationship confirmed by the correlation coefficient of 0.251.

Let's plot sulphates against alcohol and color by quality group, using our low_high_wines subset.

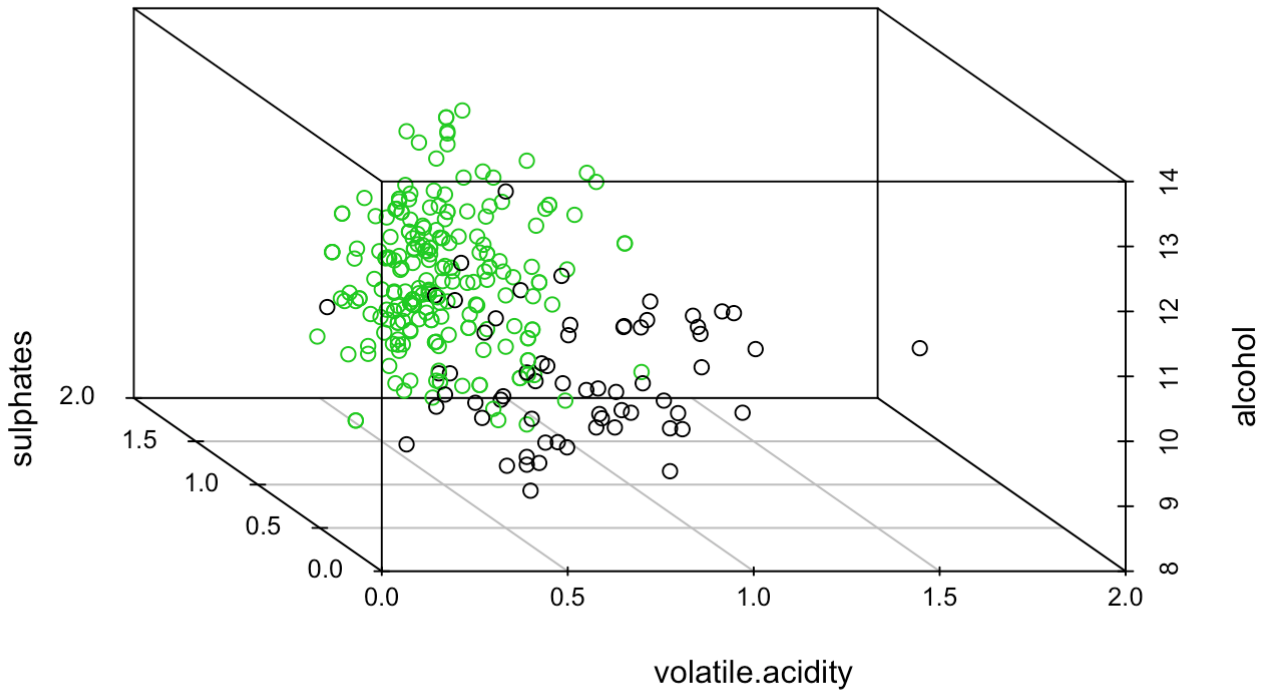
```
ggplot(aes(sulphates, alcohol), data=low_high_wines) +
  coord_cartesian(xlim=c(0.3, 1.5)) +
  geom_point(aes(color=quality.group))
```




Again, our groups are overlapping.

Let's try something else by plotting a 3D scatterplot with the three variables studied so far (volatile.acidity, sulphates, alcohol).

```
library(scatterplot3d)
attach(low_high_wines)
scatterplot3d(volatile.acidity, sulphates, alcohol,
              angle=120, color=as.numeric(quality.group))
```

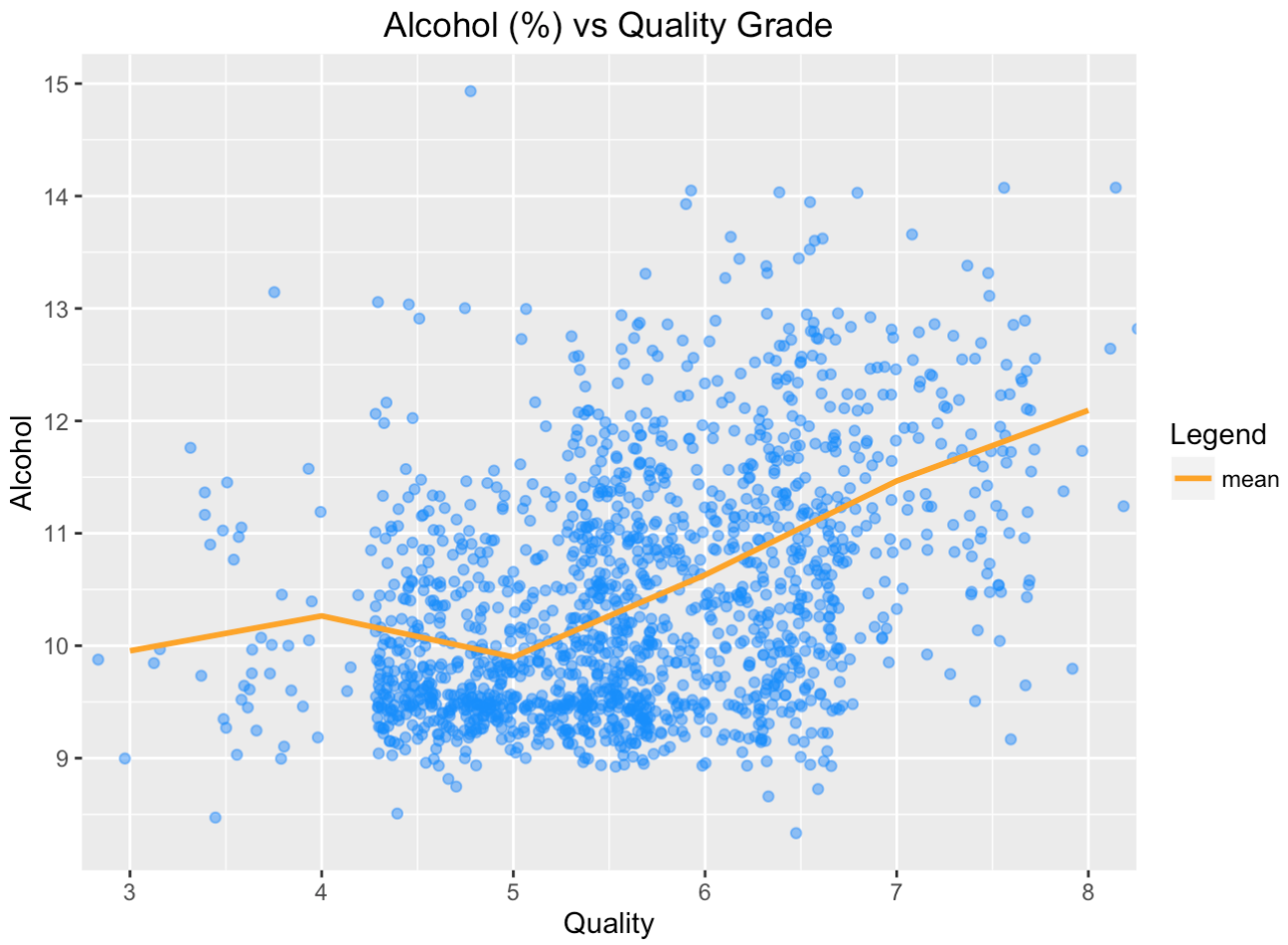



While we can't seem to avoid some overlapping, the two clusters are definitely visible.

# 2.Final plots

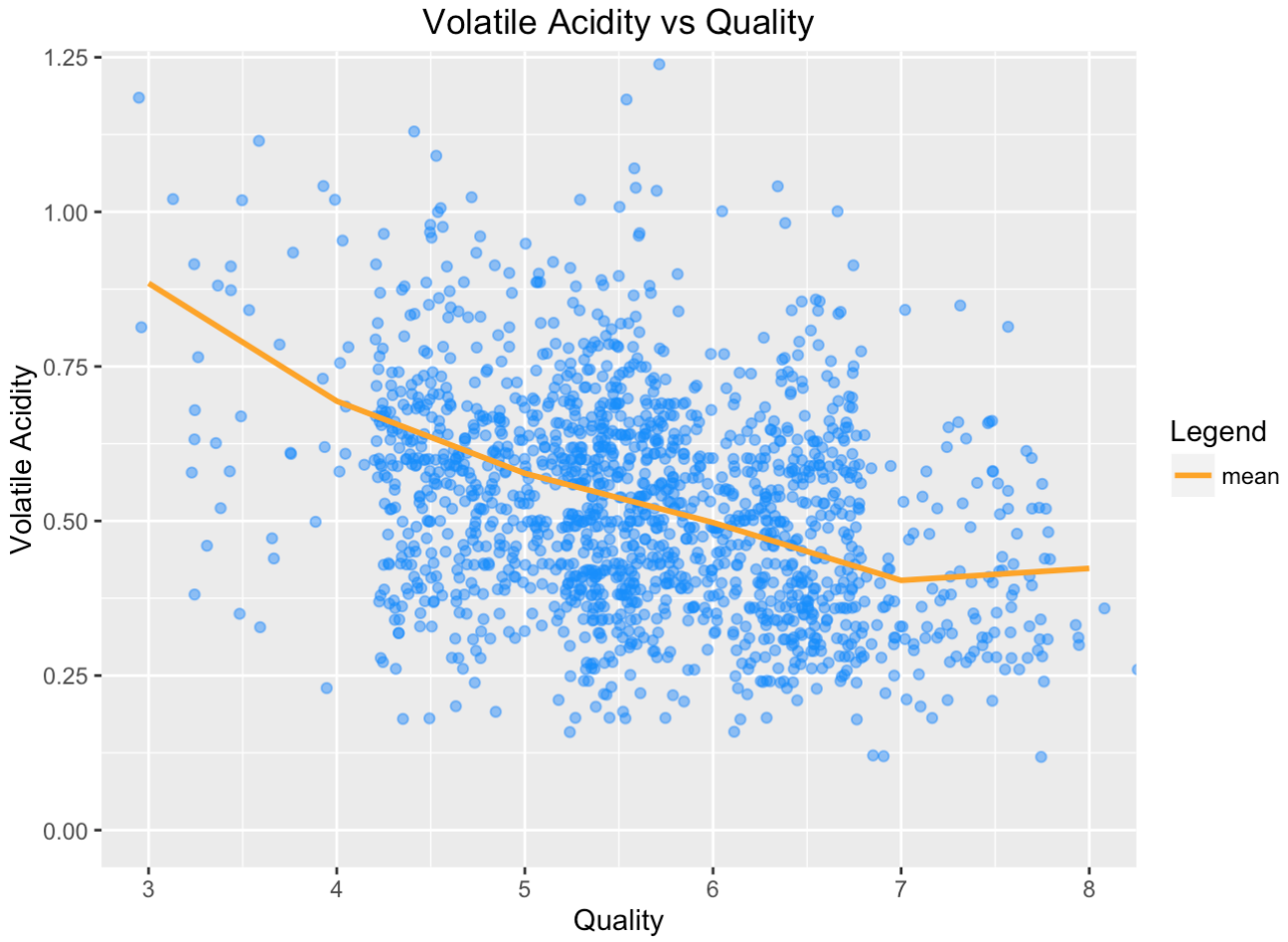
**Alcohol vs Quality**

```
ggplot(aes(quality, alcohol), data = wine) +
  geom_jitter(alpha=1/2,  width = 1.8, height = 0.2, color="#008cff") +
  coord_cartesian(xlim=c(3, 8)) +
  scale_y_continuous(breaks=seq(8,15,1)) +
  stat_summary(fun.y = mean, aes(shape="mean", color="mean"), geom="line", size = 1)
+
  scale_colour_manual("Legend", values=c("mean"="orange")) +
  labs(x="Quality", y="Alcohol", title="Alcohol (%) vs Quality Grade")
```



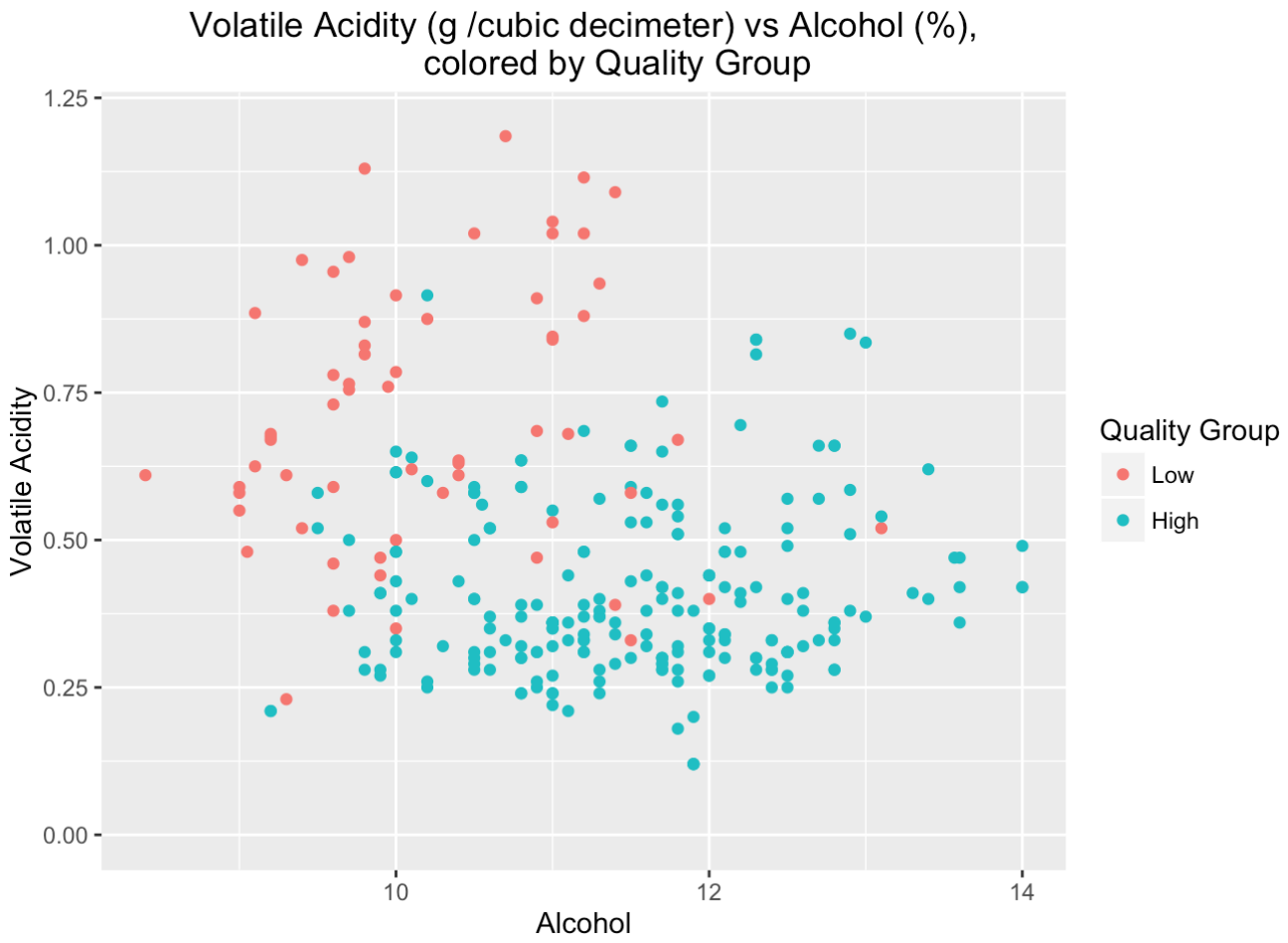

### Volatile Acidity vs Quality

```
ggplot(aes(quality, volatile.acidity), data = wine) +
  geom_jitter(alpha=1/2, width = 2, color="#008cff") +
  coord_cartesian(xlim=c(3, 8), ylim=c(0, 1.2)) +
  stat_summary(fun.y=mean, aes(shape="mean", color="mean"), geom="line", size=1) +
  scale_colour_manual("Legend", values=c("mean"="orange")) +
  labs(x="Quality", y="Volatile Acidity", title="Volatile Acidity vs Quality")
```




### Volatile Acidity vs Alcohol

```
ggplot(aes(alcohol, volatile.acidity), data = low_high_wines) +
  geom_point(aes(color=quality.group)) +
  coord_cartesian(ylim=c(0, 1.2)) +
  labs(x="Alcohol",
       y="Volatile Acidity",
       title="Volatile Acidity (g /cubic decimeter) vs Alcohol (%),
       colored by Quality Group",
       colour="Quality Group")
```

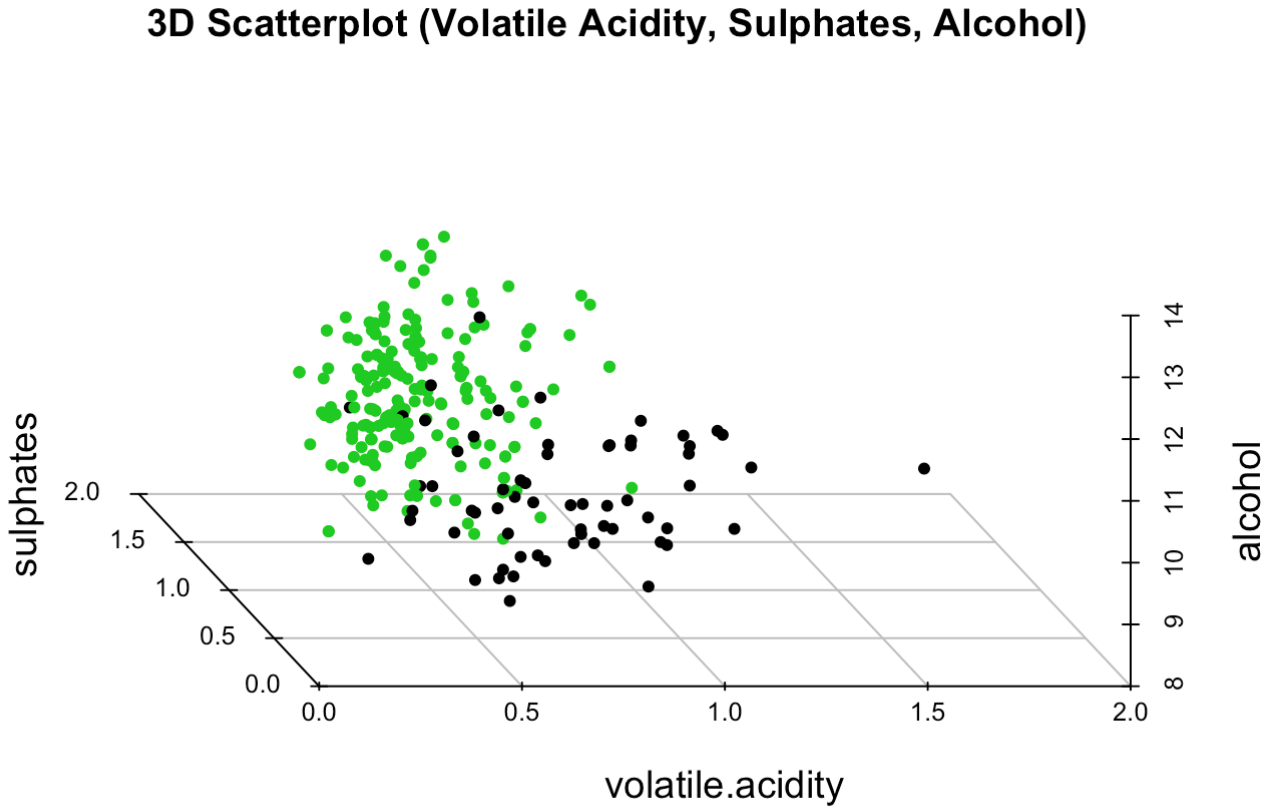



### 3D Scatterplot (Volatile Acidity, Sulphates, Alcohol)

```
attach(low_high_wines)
```

```
## The following objects are masked from low_high_wines (pos = 3):
##
##     alcohol, chlorides, citric.acid, cluster, density,
##     fixed.acidity, free.sulfur.dioxide, pH, quality,
##     quality.factor, quality.group, residual.sugar, sulphates,
##     total.sulfur.dioxide, volatile.acidity, X
```

```
scatterplot3d(volatile.acidity, sulphates, alcohol, pch=19,
              main="3D Scatterplot (Volatile Acidity, Sulphates, Alcohol)",
              color=as.numeric(quality.group), box=FALSE, cex.lab   =1.2, cex.symbols=0
7,
              angle=110)
```




**Legend:** green: High Quality black: Low Quality

# 3.Reflection

We've seen that a few chemical elements have relationships with wine quality but haven't found any strong correlation. All of our modelling is based on the assumption the quality notation is trustworthy that there is a direct relationship between the grades and the physicochemical properties.

Regarding the dataset, I found the information gap between the accucary of physicochemical data and the quality scale (graded on a discrete scale from 1 to 10, in practice only from 3 to 8) to be too important. I wished the experts who graded the wines had given more variables to study: e.g. perceived acidity, color..

Even if accurately graded, "Quality" is quite an abstract concept. Are we talking about quality against some sort of standard ? Given that wines differ from region to region and all of our dataset wines are coming from Portugal, are we talking about the quality for a Portuguese wine of that region, or of quality for all red wines?

On a regional scale and If our assumptions are true, we may have missed relationships that more complex modeling techniques could put forward. But I do have concerns regarding a theorical global wine grading using physicochemical data, it seems that people who buy Bordeaux are probably not looking for the same experience as people buying Argentine wine. A great argentine wine taste different than a great Bordeaux. So even assuming that it is possible to accurately represent quality or taste through physicochemical data, it seems likely that what is considered "great" could differ in physicochemical terms depending on the region.

---

## Ressources used:

- https://en.wikipedia.org/wiki/PH (https://en.wikipedia.org/wiki/PH)
- http://extension.psu.edu/food/enology/wine-production/wine-made-easy-fact-sheets/volatile-acidity-in-wine (http://extension.psu.edu/food/enology/wine-production/wine-made-easy-fact-sheets/volatile-

acidity-in-wine)
- https://winemakermag.com/676-the-perils-of-volatile-acidity (https://winemakermag.com/676-the-perils-of-volatile-acidity)
- http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity (http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity)
- http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf (http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf)
- https://ggobi.github.io/ggally/ggpairs.html (https://ggobi.github.io/ggally/ggpairs.html)
- https://stat.ethz.ch/R-manual/R-devel/library/base/html/cut.html (https://stat.ethz.ch/R-manual/R-devel/library/base/html/cut.html)