White Wine Quality

Baskaran Viswanathan

October 01, 2015

Abstract

In this project I"m going to investigate white wine quality. The final result will be predictive model and patterns discovery of wine quality based on chemical properties. In the first section presented data exploration. In the second part building predictive model.

First Part Section

```
setwd("~/GitHub/Udacity/Project3/Project")
df <- read.csv('wineQualityWhites.csv')
dim(df)
```

```
## [1] 4898   13
```

```
str(df)
```

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

```
summary(df)
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

More precise look at quality column.

```
table(df$quality)
```

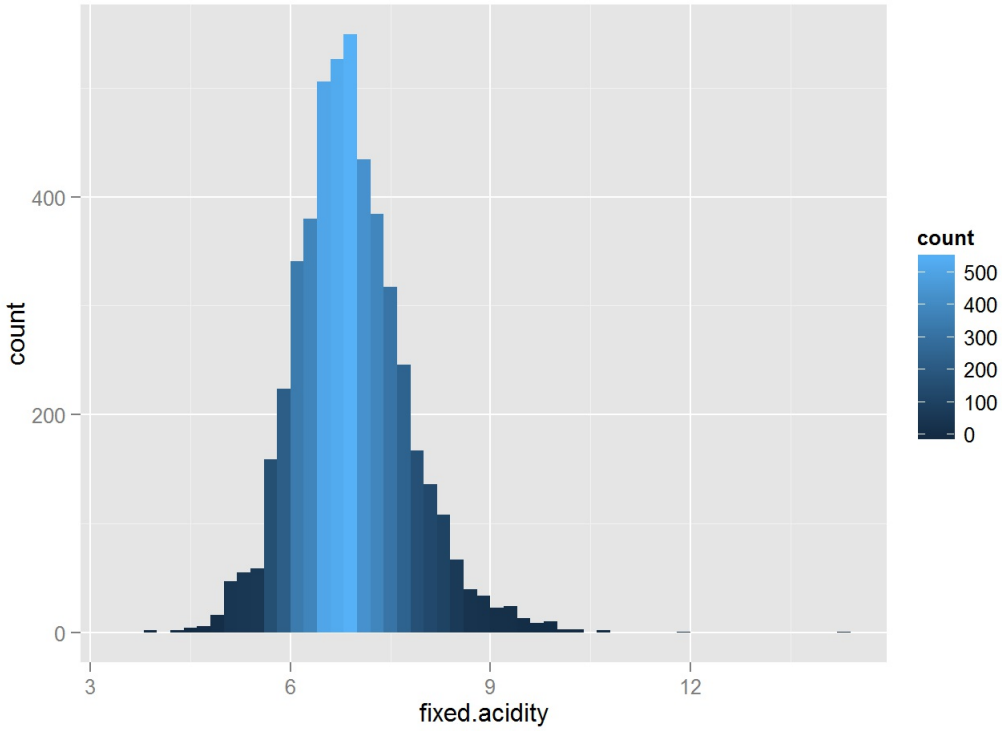
```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

So it's more useful and suitable to create ordered factor.

```
df$quality.factor <- factor(df$quality, ordered=TRUE)
df$X <- NULL
```

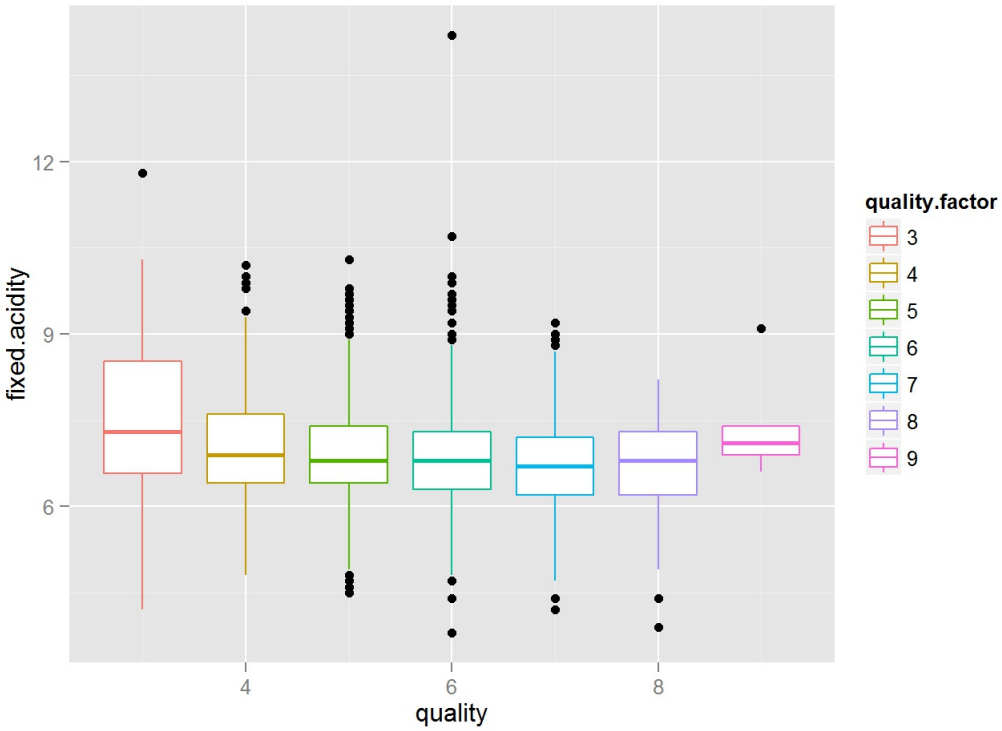
Data Exploration

```
library(ggplot2)
library(GGally)
library(gridExtra)
ggplot(data=df, aes(x=fixed.acidity)) + geom_histogram(aes(fill=..count..), binwidth = 0.2)
```



Looks very normal, let's add boxplot.

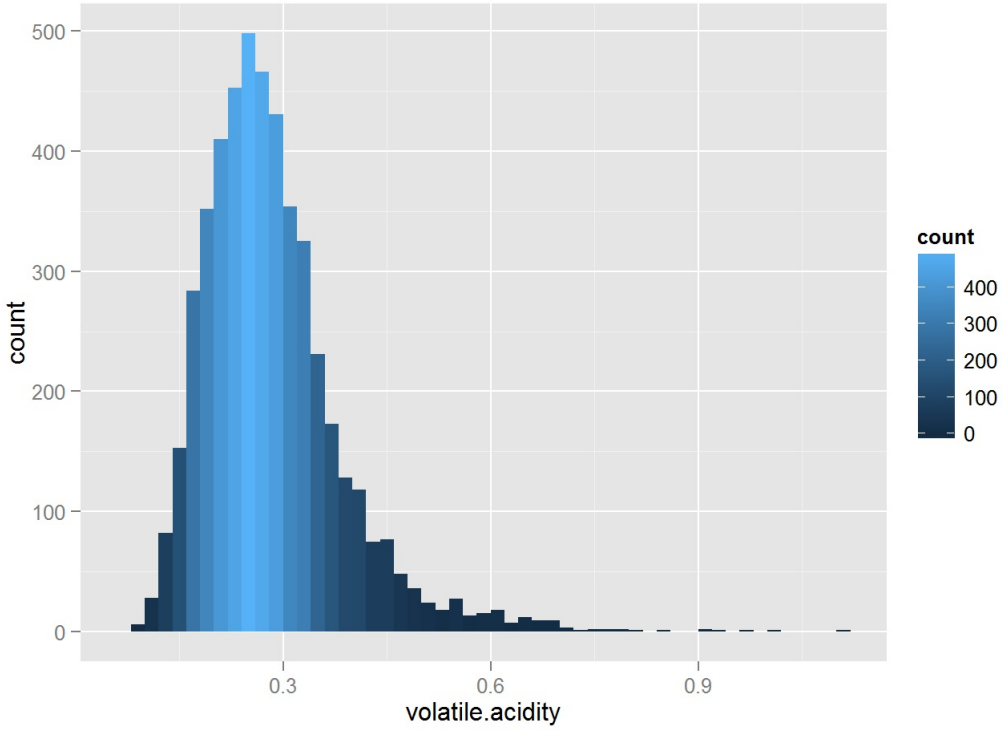
```
ggplot(data=df, aes(y=fixed.acidity, x = quality)) + geom_boxplot(aes(color=quality.factor))
```



There is no some significant difference between quality and fixed acidity. Remind that fixed acidity means value of the most acids involved with wine.

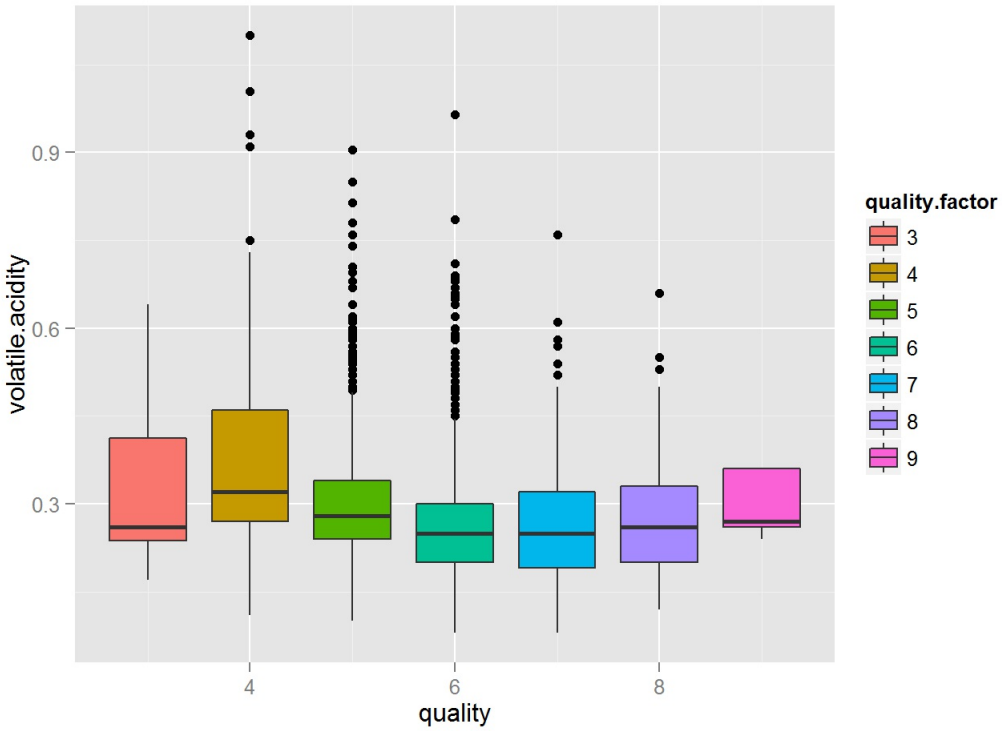
Move to next variable. It's volatile acidity (the amount of acetic acid in wine).

```
ggplot(data=df, aes(x=volatile.acidity)) + geom_histogram(aes(fill=..count..), binwidth=0.02)
```



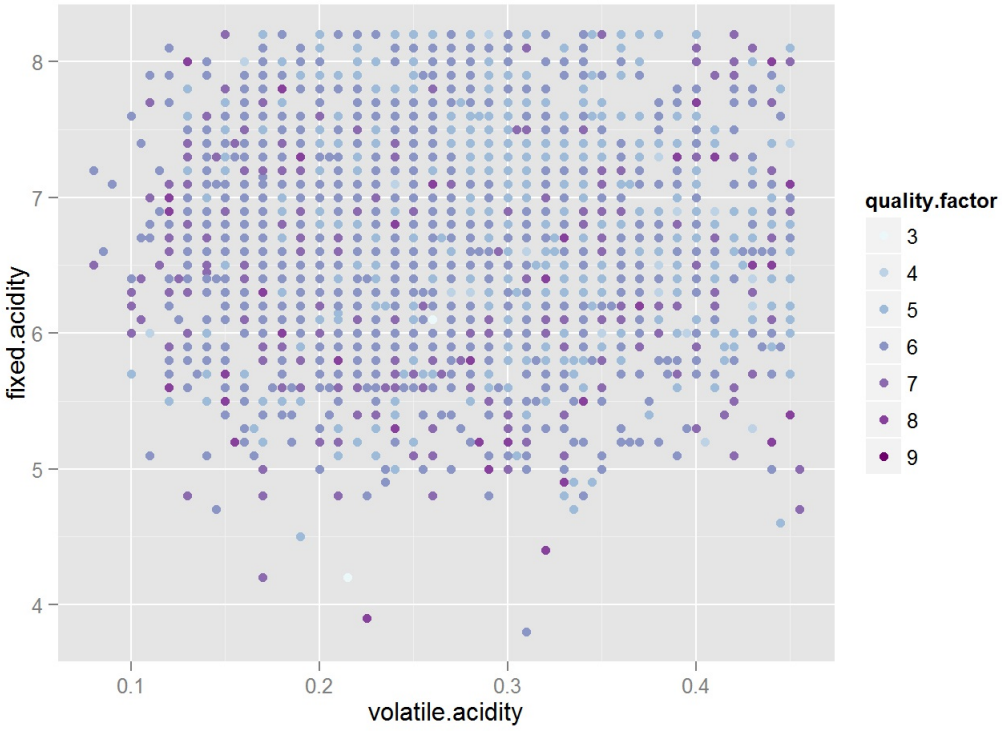
Look at relationship with quality.

```
ggplot(data=df, aes(y=volatile.acidity, x = quality)) + geom_boxplot(aes(fill=quality.factor))
```



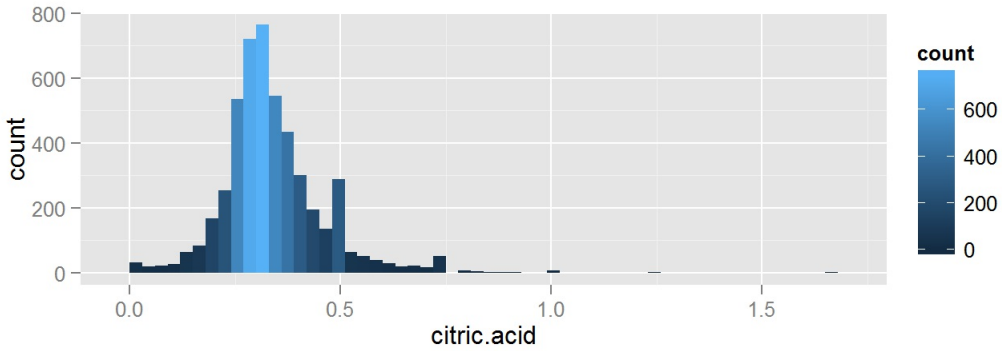
There is no visible separation based on this plot. So I'm going to combine acidity variables with quality.

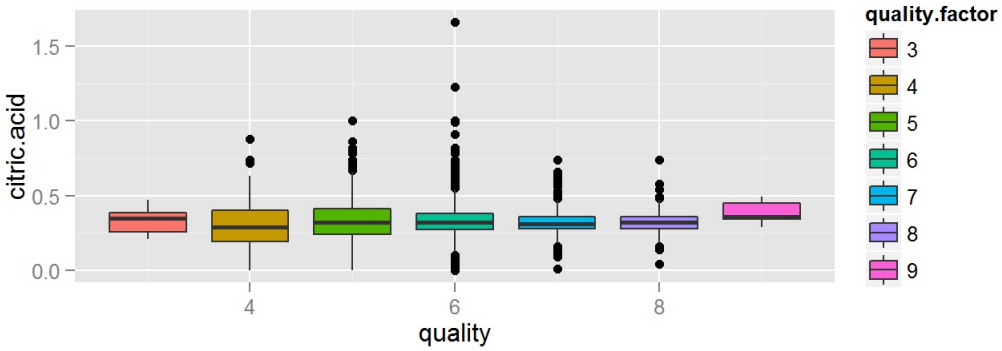
```
ggplot(data=subset(df, fixed.acidity < quantile(fixed.acidity, 0.95) & volatile.acidity < quantile(volatile.aci
dity, .95)), aes(y=fixed.acidity, x = volatile.acidity)) + geom_point(aes(color=quality.factor)) + scale_colour
_brewer(type="seq", palette=3)
```



Unfortunately, no visual understandable separation using this two features. Go forward to next feature from white wine dataset.

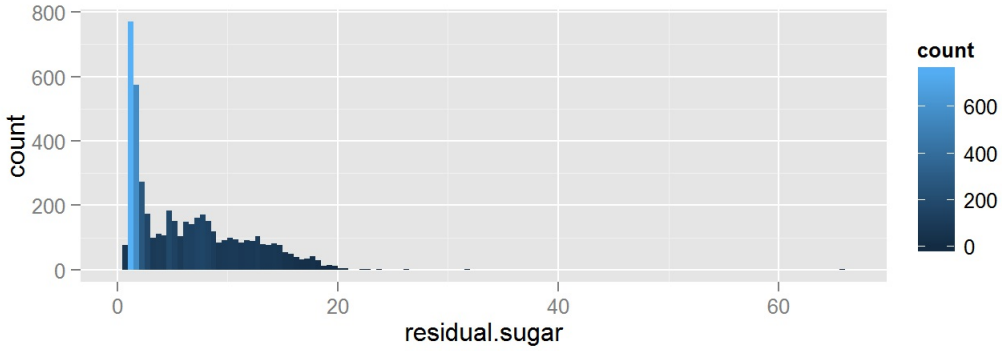
```
g1 <- ggplot(data=df, aes(x=citric.acid)) + geom_histogram(aes(fill=..count..), binwidth=0.03)
g2 <- ggplot(data=df, aes(y=citric.acid, x = quality)) + geom_boxplot(aes(fill=quality.factor))
grid.arrange(g1,g2, ncol=1)
```

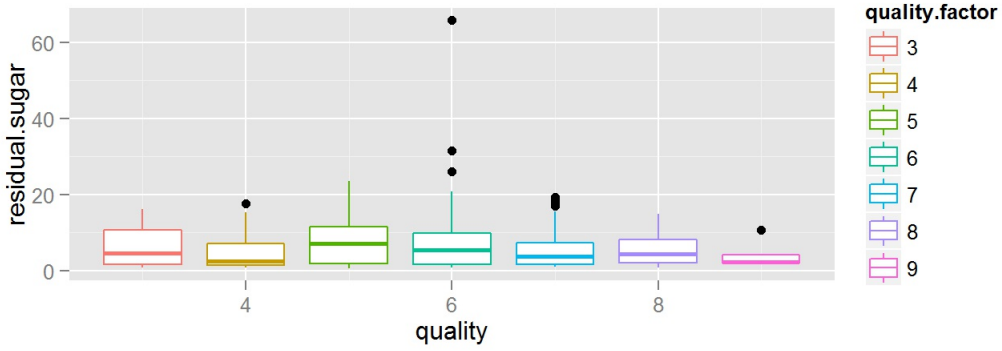




Plots above don't show any clear patterns. But we can note that there are much more citric.acid outliers with quality equal 6. Look at residual.sugar variable. In histogram below we can observe unusual peaks with count near 600~800, but at the same time boxplots gives us no additional information about patterns.

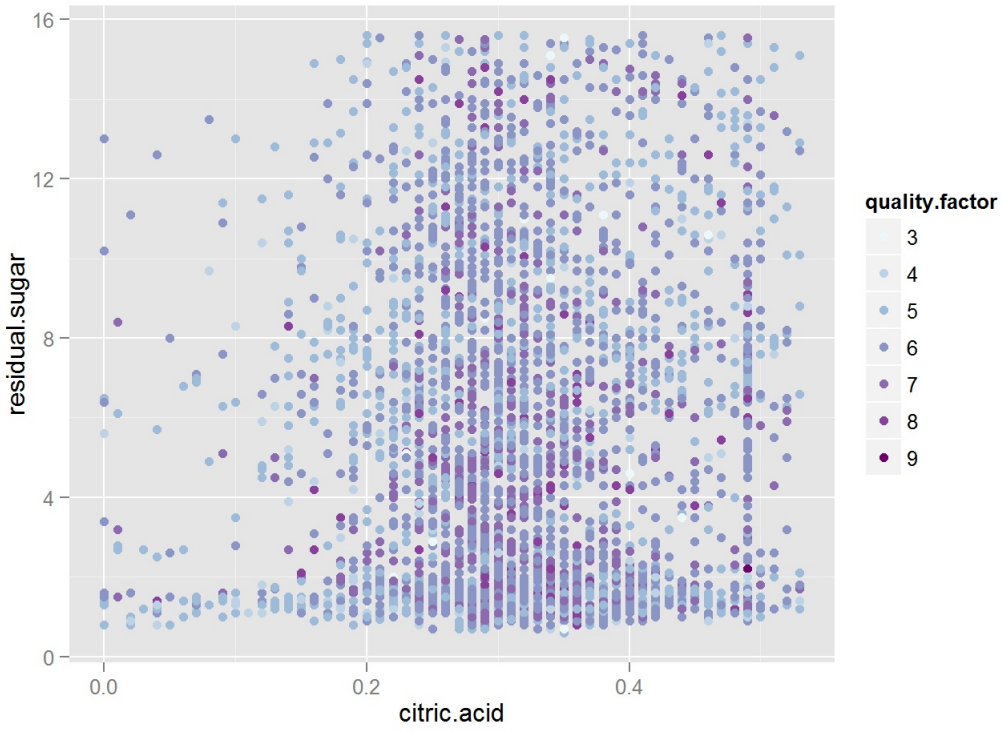
```
g1 <- ggplot(data=df, aes(x=residual.sugar)) + geom_histogram(aes(fill=..count..), binwidth=0.5)
g2 <- ggplot(data=df, aes(y=residual.sugar, x = quality)) + geom_boxplot(aes(color=quality.factor))
grid.arrange(g1,g2, ncol=1)
```





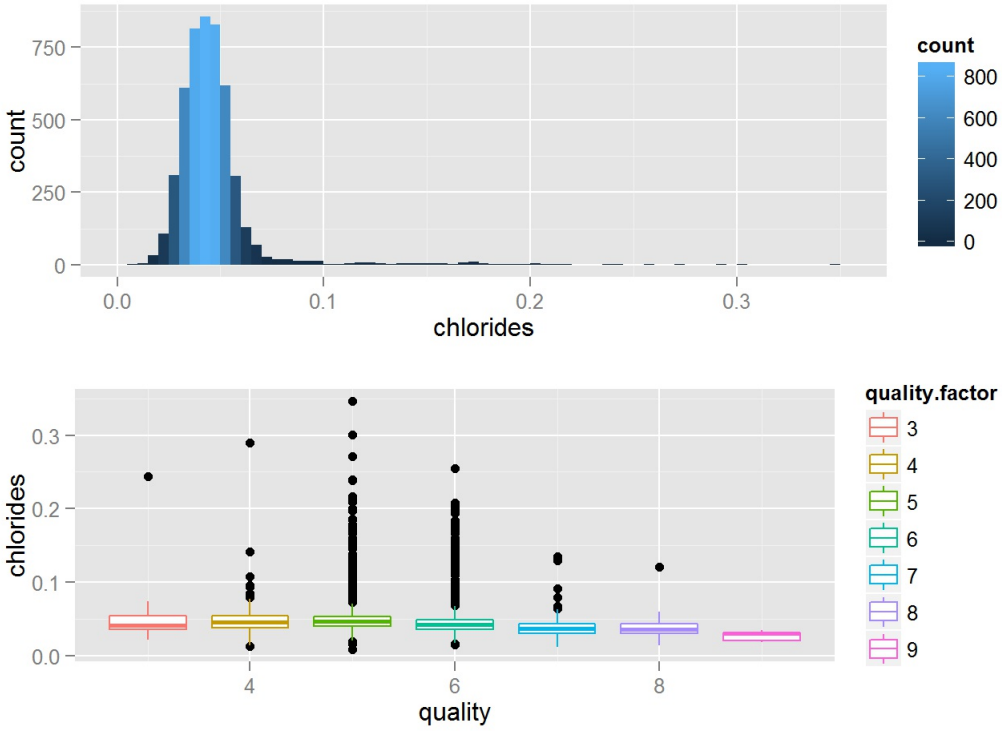
Combining residual.sugar and citric.acid variables to determine some linear separation.

```
ggplot(data=subset(df, residual.sugar < quantile(residual.sugar, .95) & citric.acid < quantile(citric.acid, .95
)), aes(y=residual.sugar, x = citric.acid)) + geom_point(aes(color=quality.factor)) + scale_colour_brewer(type=
"seq", palette=3)
```



No clearn patterns, but we can observe quite nice relationship between this variables with different modes(citric.acid - centered, residual.sugar near zero) and plot looks nice. Move to next variable - chlorides.

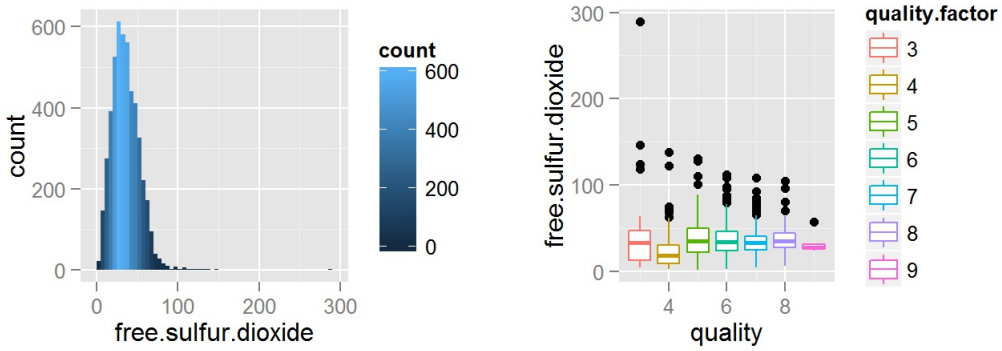
```
g1 <- ggplot(data=df, aes(x=chlorides)) + geom_histogram(aes(fill=..count..), binwidth=0.005)
g2 <- ggplot(data=df, aes(y=chlorides, x = quality)) + geom_boxplot(aes(color=quality.factor))
grid.arrange(g1,g2, ncol=1)
```

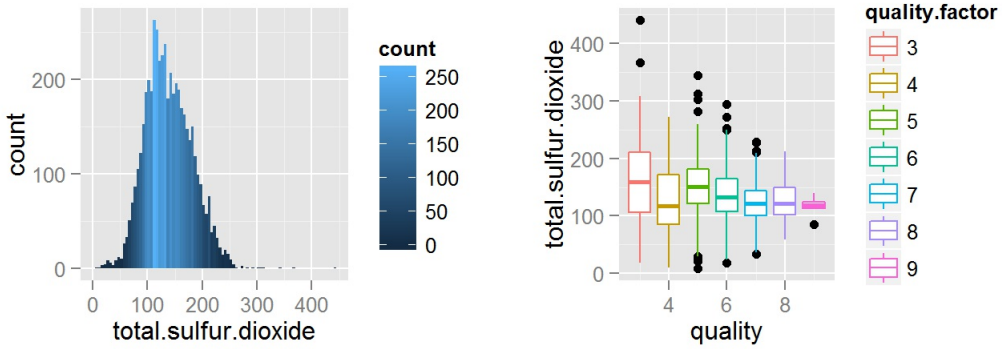


More meaningful variable. Easily can see to many outliers in wine with quality 5 and 6. But medians is still near the same value for all wine quality.

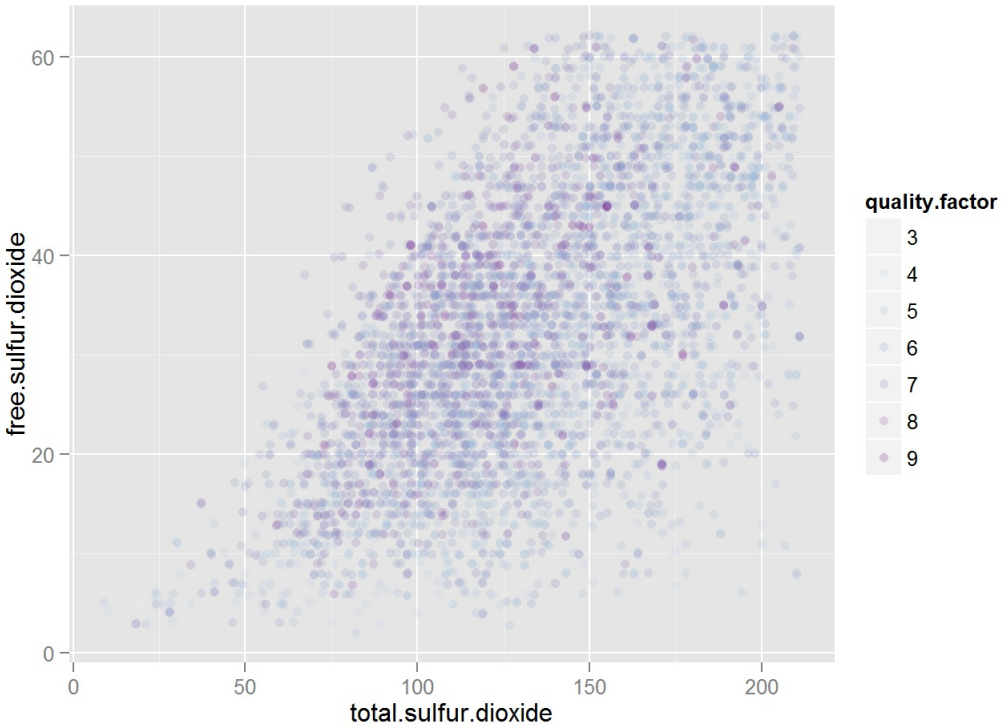
Exploring together free silfur and total sulfur dioxides. From plots below there is no signifficant information for patterns discovery except few outliers for wine with quality 3.

```
g1 <- ggplot(data=df, aes(x=free.sulfur.dioxide)) + geom_histogram(aes(fill=..count..), binwidth=5)
g2 <- ggplot(data=df, aes(y=free.sulfur.dioxide, x = quality)) + geom_boxplot(aes(color=quality.factor))
g3 <- ggplot(data=df, aes(x=total.sulfur.dioxide)) + geom_histogram(aes(fill=..count..), binwidth=5)
g4 <- ggplot(data=df, aes(y=total.sulfur.dioxide, x = quality)) + geom_boxplot(aes(color=quality.factor))
grid.arrange(g1,g2,g3,g4, ncol=2)
```





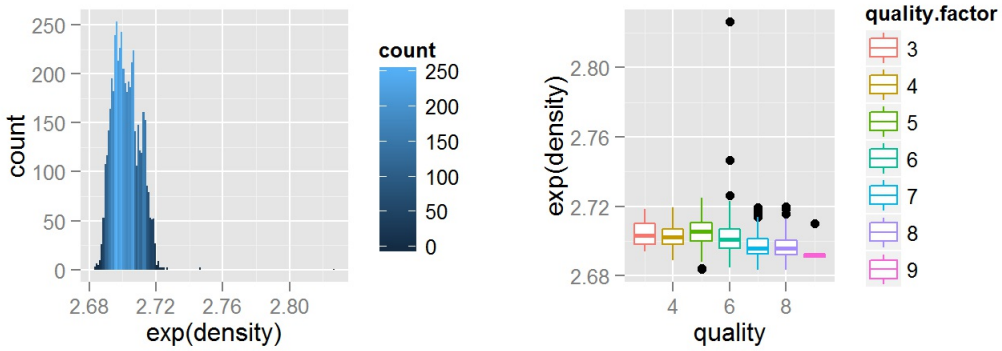

```
ggplot(data=subset(df, free.sulfur.dioxide < quantile(free.sulfur.dioxide, .95) & total.sulfur.dioxide < quanti
le(total.sulfur.dioxide, .95)), aes(y=free.sulfur.dioxide, x = total.sulfur.dioxide)) + geom_jitter(alpha=1/5,a
es(color=quality.factor)) + scale_colour_brewer(type="seq", palette=3)
```

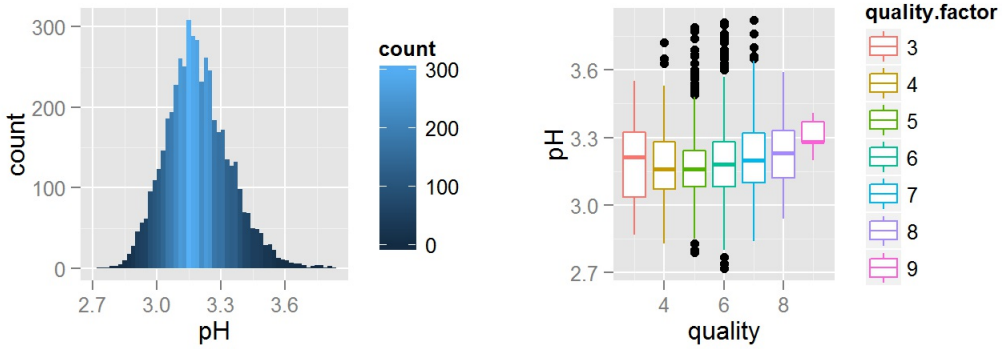


Unfortunately no meaningful separation yet, but insteresting plot above is one of the examples of regression to the mean i think. Going to next variables density and Ph. Density is too simillar for all kinds of wines and hard to investigate due to some outliers, so i decided to use 0.95 quantile to filter data.

```
g1 <- ggplot(data=df, aes(x=exp(density))) + geom_histogram(aes(fill=..count..), binwidth=0.001)
g2 <- ggplot(data=df, aes(y=exp(density), x = quality)) + geom_boxplot(aes(color=quality.factor))
g3 <- ggplot(data=df, aes(x=pH)) + geom_histogram(aes(fill=..count..), binwidth=0.02)
g4 <- ggplot(data=df, aes(y=pH, x = quality)) + geom_boxplot(aes(color=quality.factor))
grid.arrange(g1,g2,g3,g4, ncol=2)
```

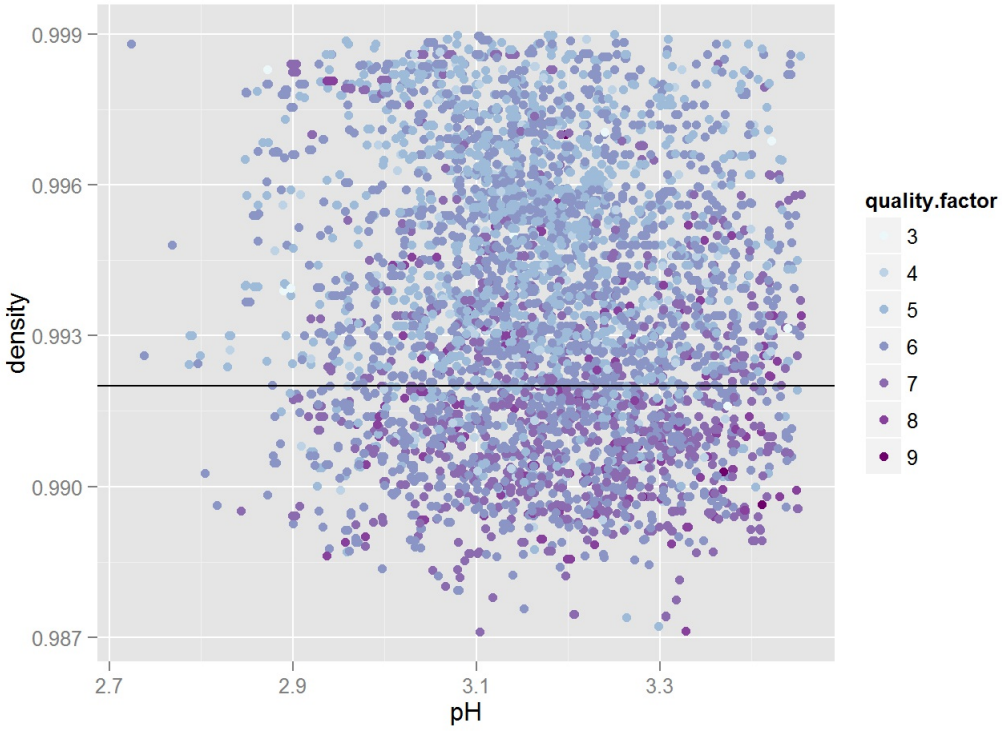
```
## Warning: position_stack requires constant width: output may be incorrect
```





We can see some interesting trends from this plots, like with less density -> quality higher. The same is for pH, but in median thinking. Combine this two features to look at this data.

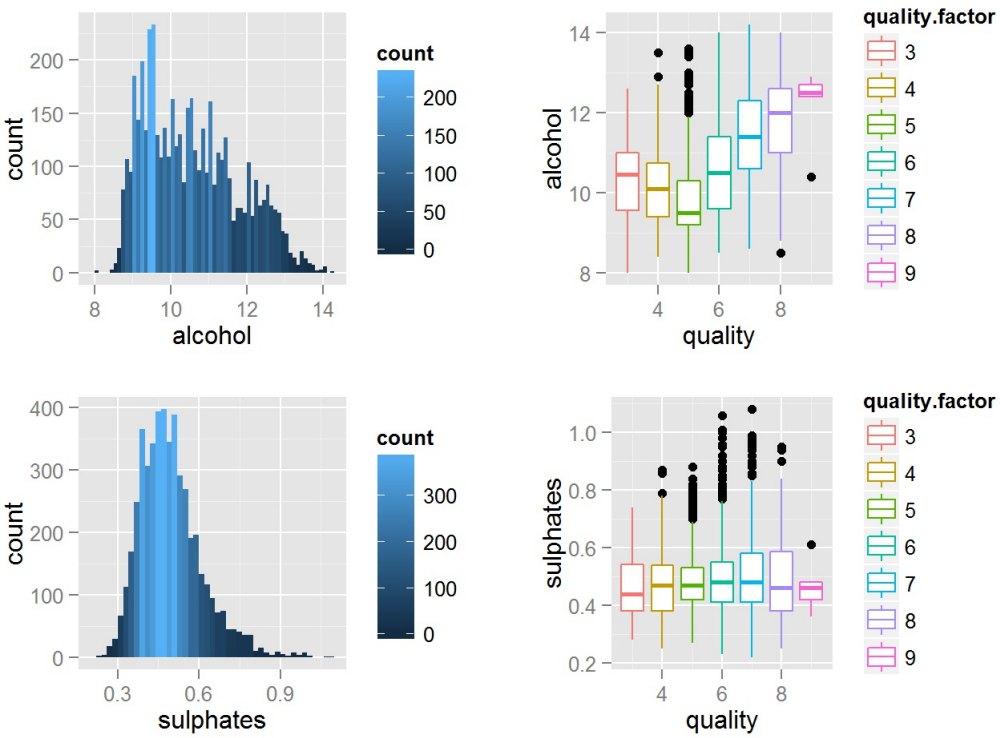
```
ggplot(data=subset(df, density < quantile(density, .95) & pH < quantile(pH, .95)), aes(y=density, x = pH)) + ge
om_jitter(aes(color=quality.factor)) + scale_colour_brewer(type="seq", palette=3) + geom_abline(intercept = .99
2, slope = 0)
```



Look like under black line there is more chance that wine quality is high. In someway first result.

Okey. And the last two variables alcohol and sulphates.

```
g1 <- ggplot(data=df, aes(x=alcohol)) + geom_histogram(aes(fill=..count..), binwidth=0.1)
g2 <- ggplot(data=df, aes(y=alcohol, x = quality)) + geom_boxplot(aes(color=quality.factor))
g3 <- ggplot(data=df, aes(x=sulphates)) + geom_histogram(aes(fill=..count..), binwidth=0.02)
g4 <- ggplot(data=df, aes(y=sulphates, x = quality)) + geom_boxplot(aes(color=quality.factor))
grid.arrange(g1,g2,g3,g4, ncol=2)
```



In median thinking more alcohol in wine implies higher quality, based on this plots, unfortunately sulphates is not so separable for different wine qualities. Combining isn't suitable due to low variability of sulphates variables.

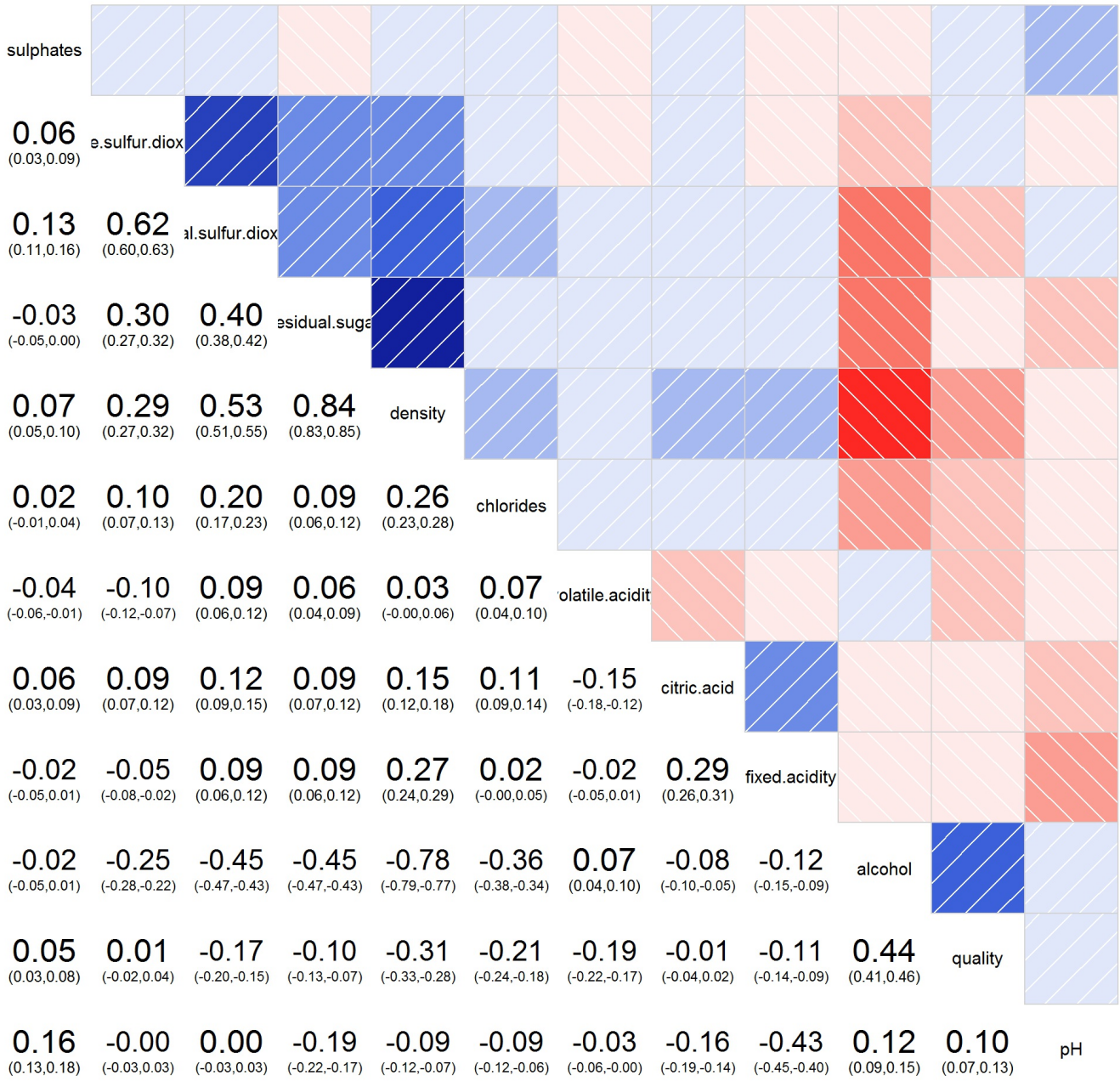
Correlation exploration

Let's look at correlations between all variables and numeric analogue of quality.

```
library(corrgram)

corrgram(df, type="data", lower.panel=panel.conf,
         upper.panel=panel.shade, main= "Corrgram for wine quality dataset", order=T, cex.labels=1.4)
```

Corrgram for wine quality dataset



Notes:

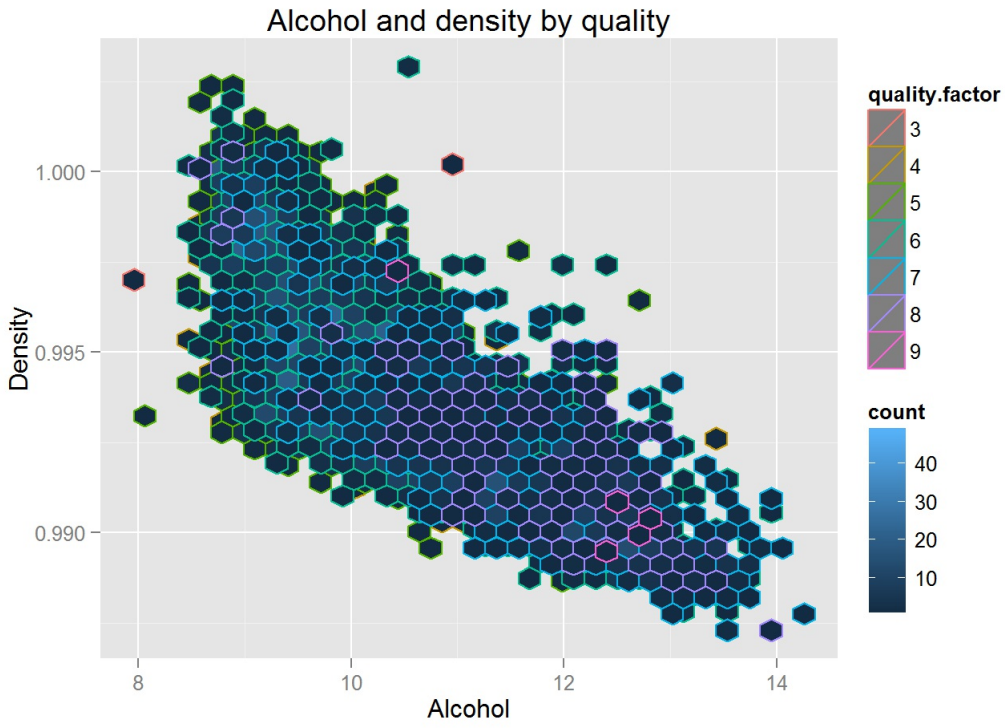
1. Highest correlation for quality with alcohol 0.44. I've note this before during exploration.
2. Highest correaltion between residual sugar and density 0.84.
3. A lot of correlations are meaningless due to confidence interval.
4. Lowest correlation for quality with density -0.31.
5. Lowest correlation between alcohol and density -0.78 .

From corrgram we can conclude next important variables for quality prediction (decision is made using confidence intervals):

1. pH (0.1)
2. alcohol (0.44)
3. fixed.acidity (-0.11)
4. volatile.acidity (-0.19)
5. chlorides (-0.21)
6. density (-0.31)
7. residual.sugar (-0.10)
8. total.sulfur.dioxide (-0.17)

Multivariate plots for understanding patterns between features

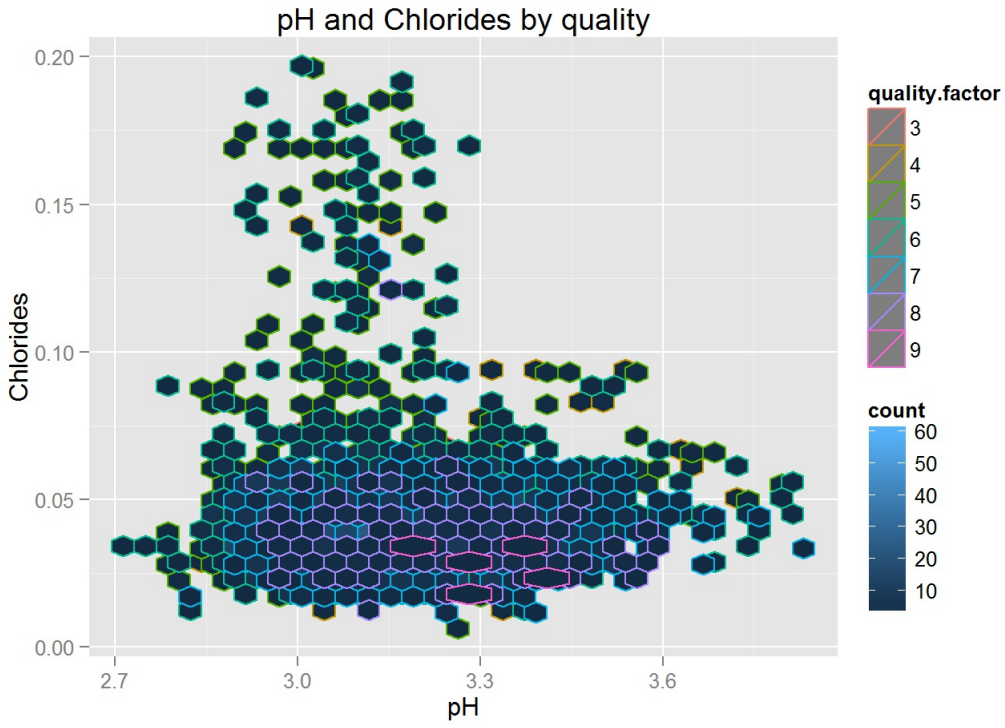
```
ggplot(data=subset(df, density < 1.005), aes(x=alcohol, y = density, color = quality.factor)) + xlab("Alcohol")
 +
ylab("Density") + ggtitle("Alcohol and density by quality") +
stat_binhex()
```



We easily can see some patterns here. This patterns is small clusters where quality is the same. This plot is awesome, it shows quality, density, alcohol relationship. With low alcohol or high density it's more usual to be low quality wine.

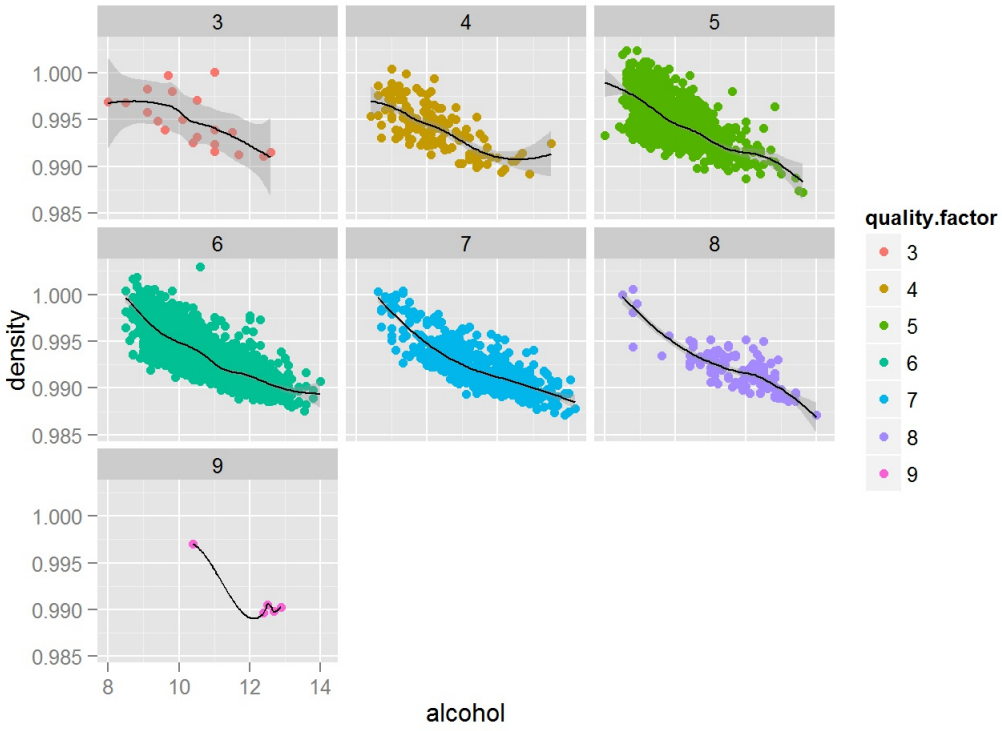
Next plot is about pH and chlorides. They both has high absolute correlation among others variables.

```
ggplot(data=subset(df, chlorides < 0.2), aes(x=pH, y = chlorides, color = quality.factor)) + xlab("pH") +
  ylab("Chlorides") + ggtitle("pH and Chlorides by quality") +
  stat_binhex()
```



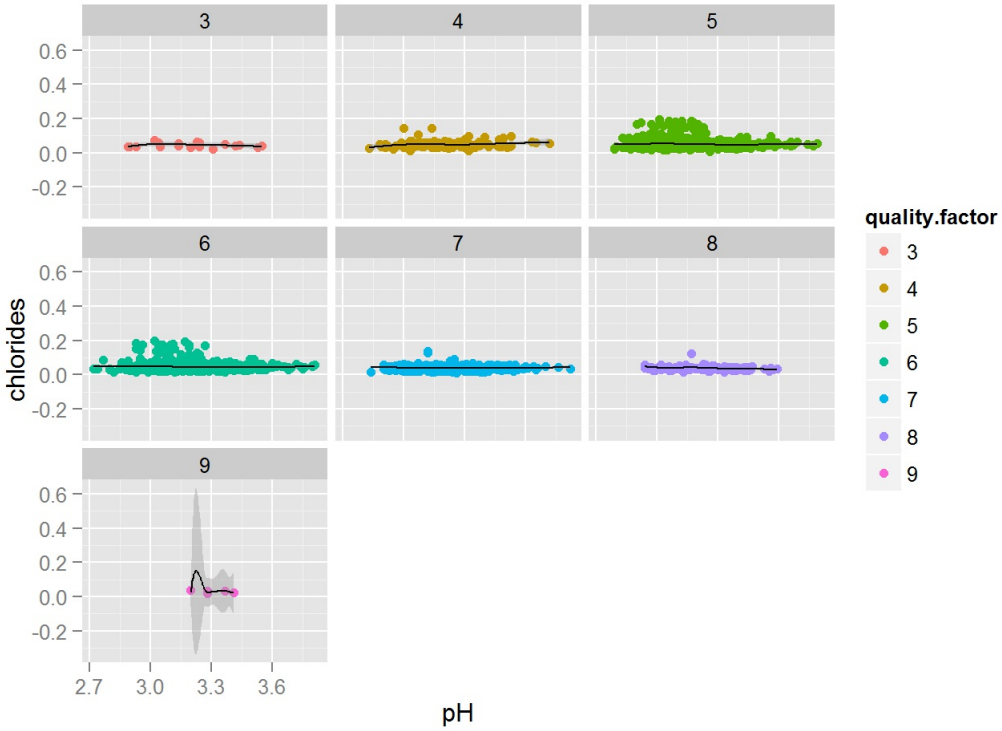
We got more patterns. High chlorides means low quality. Based on this two plots we can easily predict whether wine is low or high quality, but this is not our case, so we move to prediction.

```
ggplot(data=subset(df,density < 1.005) , aes(x=alcohol, y=density, color=quality.factor)) + geom_point() + face
t_wrap(~quality.factor) + geom_smooth(colour='black')
```



We can see some bound trends between this variables across different wine qualities.

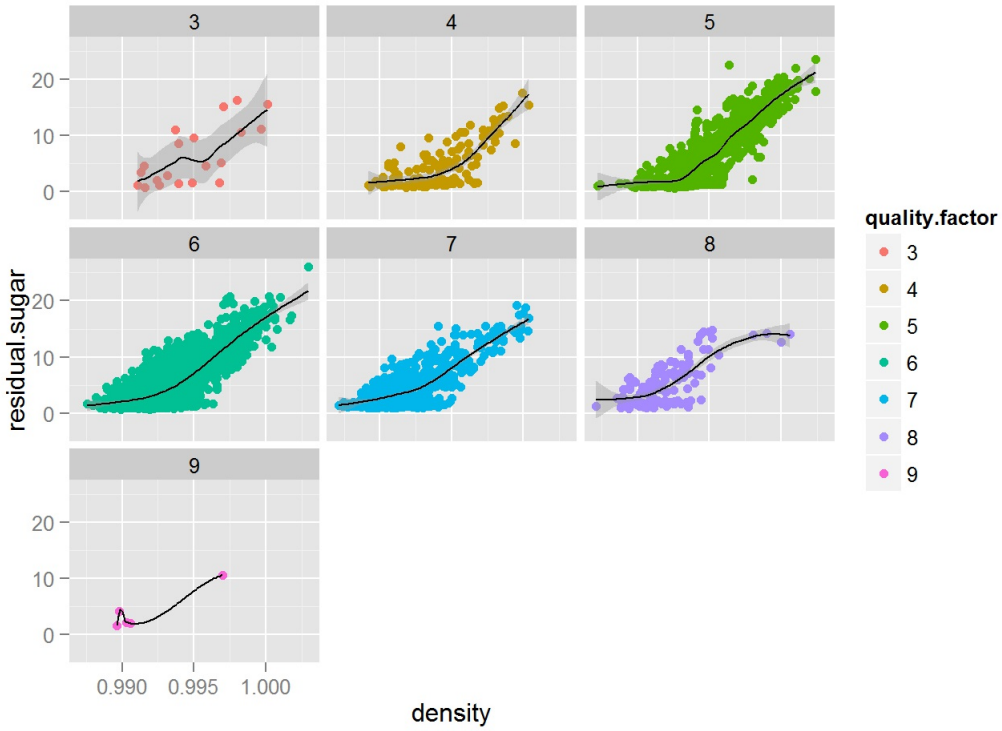
```
ggplot(data=subset(df, chlorides < 0.2), aes(x=pH, y = chlorides, color = quality.factor)) + geom_point() + fac
et_wrap(~quality.factor) + geom_smooth(colour='black')
```



Unfortunately here we can observate some stability between this variables, trend is the same, no unusual things.

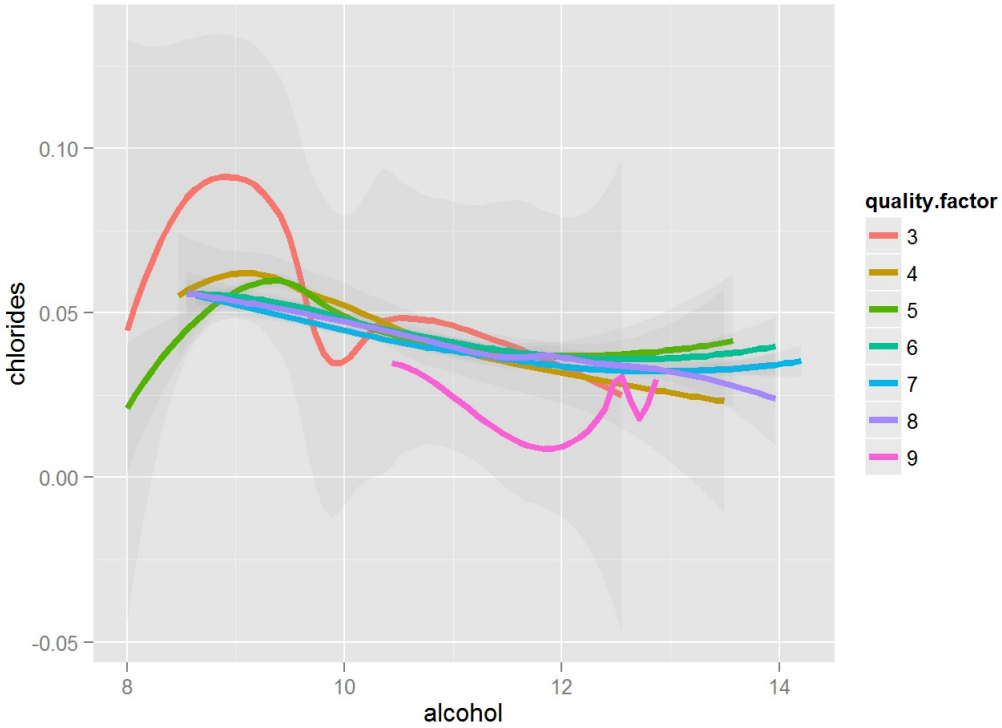
Let's explore the highest correlation variable by quality.

```
ggplot(data=subset(df, density < 1.005), aes(x=density, y = residual.sugar, color = quality.factor)) + geom_poi
nt() + facet_wrap(~quality.factor) + geom_smooth(colour='black')
```



Investigate next relationship between alcohol and free.sulfur.dioxide by quality.

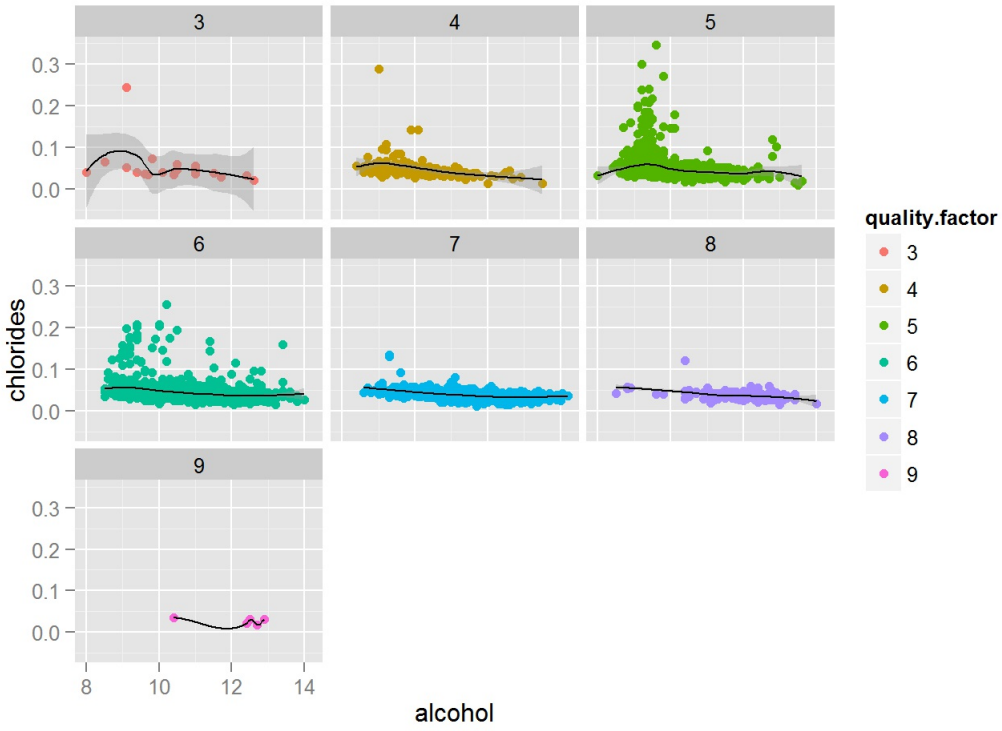
```
ggplot(aes(x=alcohol,y=chlorides, colour=quality.factor), data = df) +
stat_smooth(method=loess, fullrange=TRUE, alpha = 0.1, size =1.5 )# +
```



This plot gives us more visual separation between low and high quality wine. High quality wine have the lowest chlorides and alcohol more than 10%. And the lowest quality wine have more chlorides and low alcohol.

Let's make a scatter plot of chlorides versus alcohol by quality.

```
ggplot(aes(x=alcohol, y=chlorides, color = quality.factor), data = df) + geom_point() +
  facet_wrap(~quality.factor) + geom_smooth(colour='black')
```



Very interesting peaks for quality 5 and 6.

Prediction white wine quality based on features.

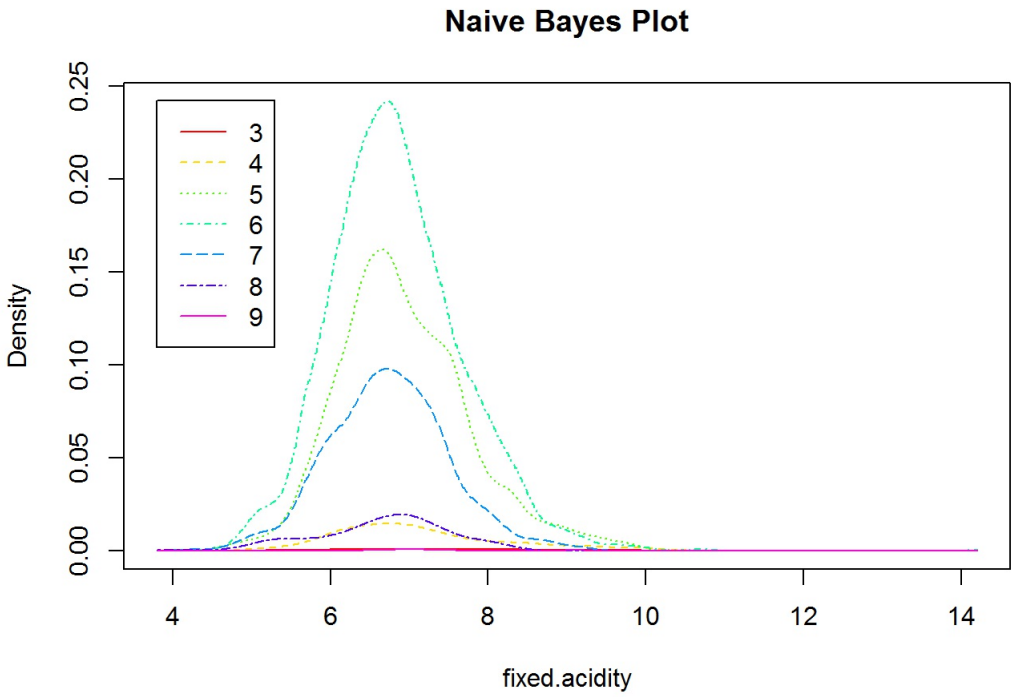
I want to use naive bayes model as my main model.

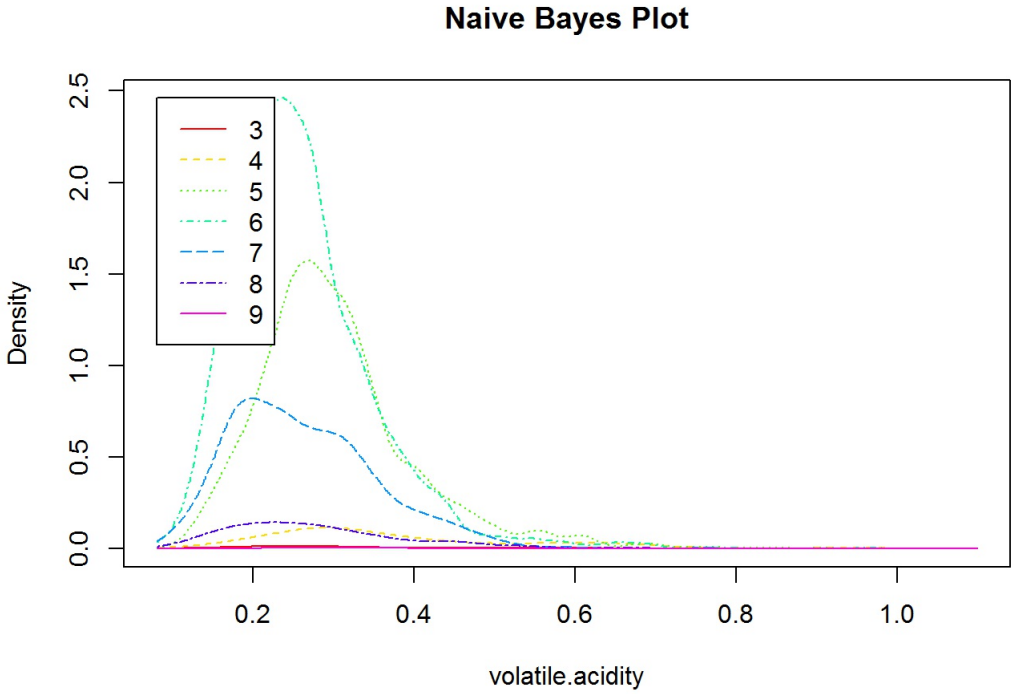
```
library(caret)
library(klaR)
df_pred <- df
df_pred$quality <- NULL
ctrl <- trainControl(method = "repeatedcv",  repeats = 3)
fit <- train(quality.factor ~ ., data = df_pred, method = "nb", trControl = ctrl)
fit
```

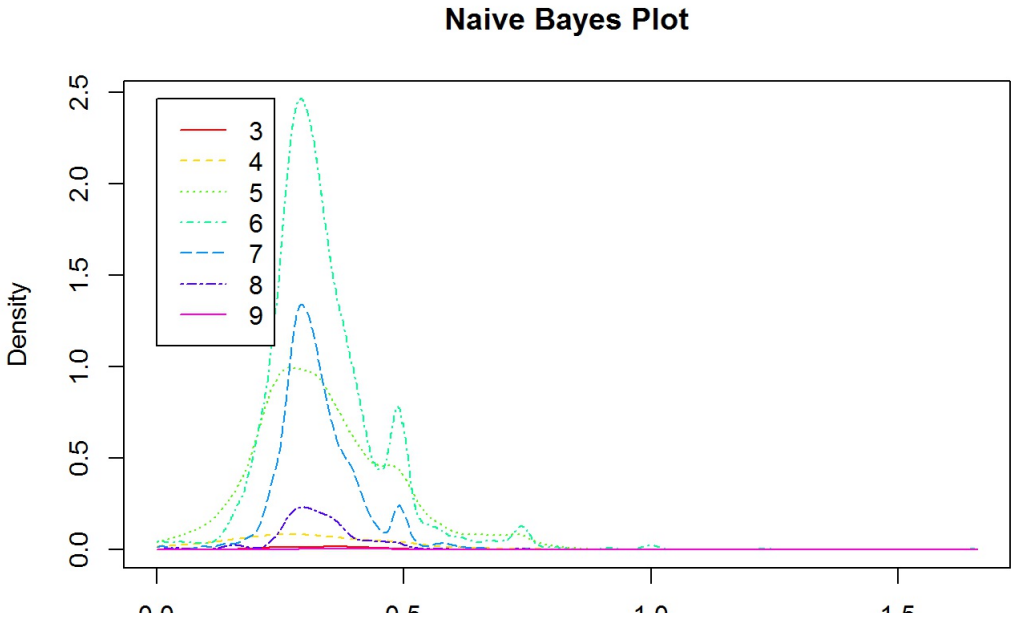
```
## Naive Bayes
##
## 4898 samples
##   11 predictor
##    7 classes: '3', '4', '5', '6', '7', '8', '9'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 4407, 4409, 4408, 4409, 4406, 4409, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa      Accuracy SD  Kappa SD
##   FALSE      0.4429628  0.2170360  0.01842355   0.02577427
##    TRUE      0.4877461  0.2535171  0.01764613   0.02640659
##
## Tuning parameter 'fL' was held constant at a value of 0
## Accuracy was used to select the optimal model using  the largest value.
## The final values used for the model were fL = 0 and usekernel = TRUE.
```

Not good not bad, but acceptable as for initial solution. Let's visualize density estimation. Accuracy is the ratio of true classified divided by all true classified plus false classified wines. Train control is 10-fold cross validation repeated 3-times. For highest accuracy is used kernel estimation (not gaussian distributed parameters). Accuracy equal 0.4888546.

```
plot(fit$finalModel)
```

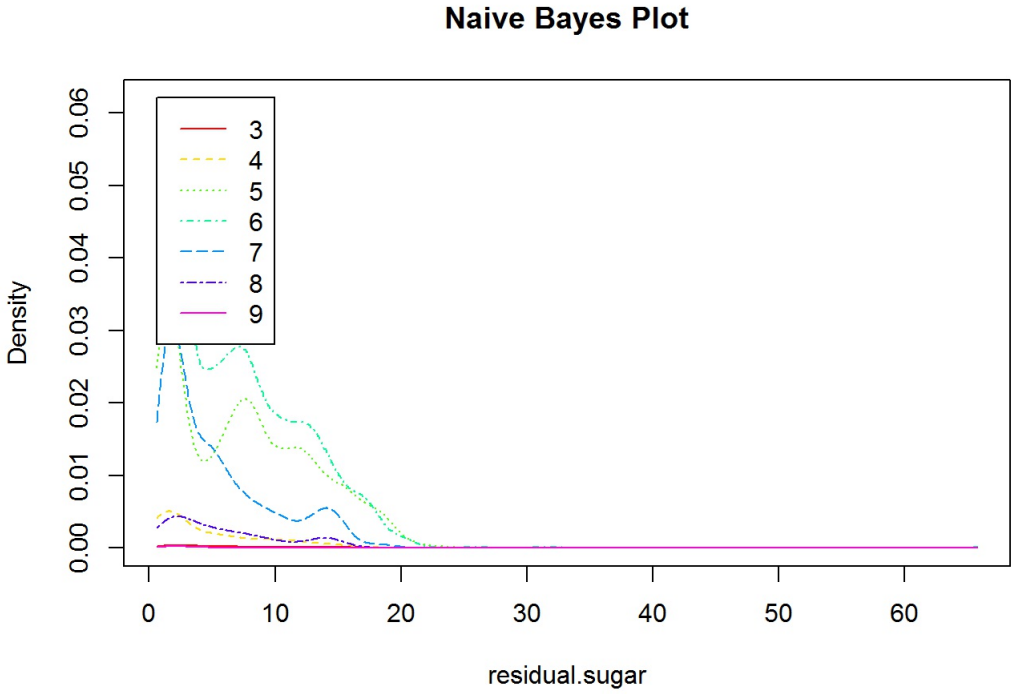

Naive Bayes Plot
Density
fixed.acidity
3
4
5
6
7
8
9
0.00
0.05
0.10
0.15
0.20
0.25
4
6
8
10
12
14

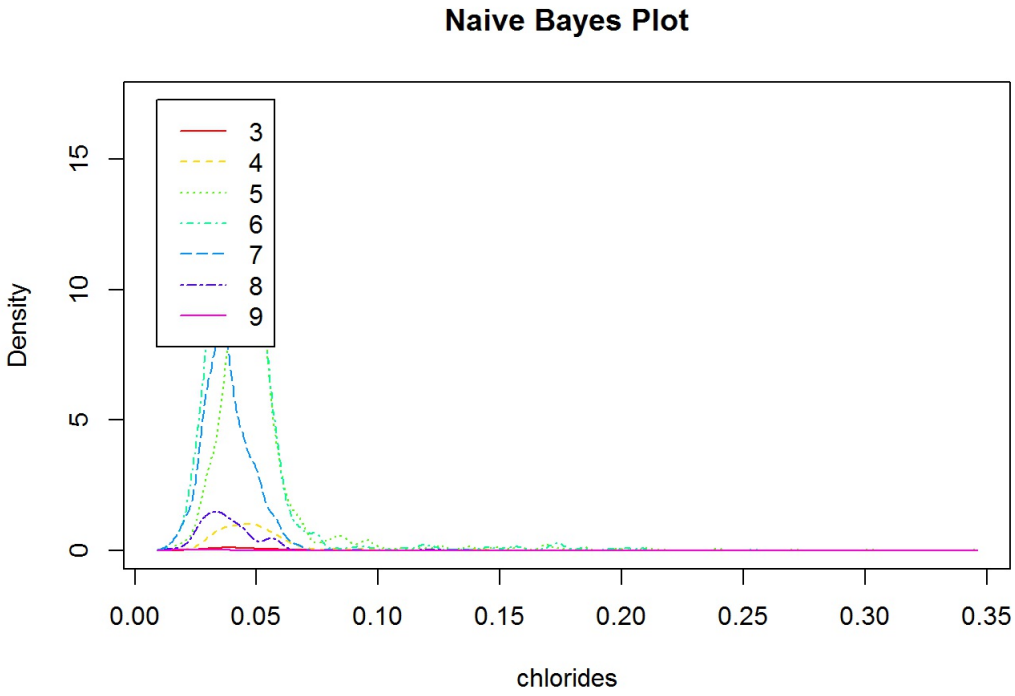

Naive Bayes Plot
Density
volatile.acidity
3
4
5
6
7
8
9
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.4
0.6
0.8
1.0

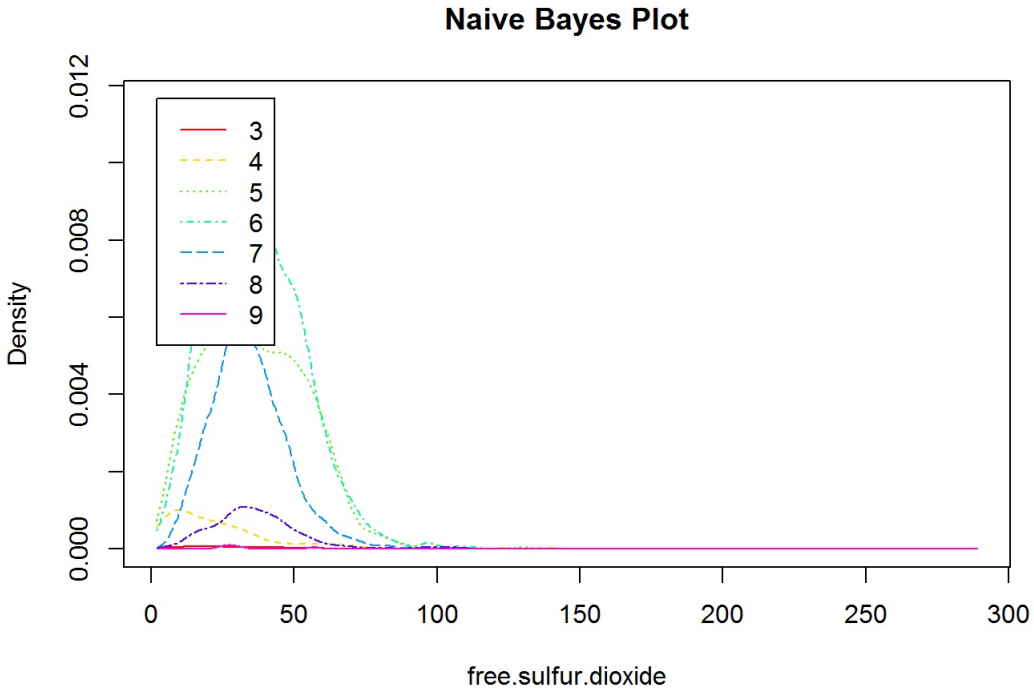

Naive Bayes Plot
Density
3
4
5
6
7
8
9
0.0
0.5
1.0
1.5
2.0
2.5

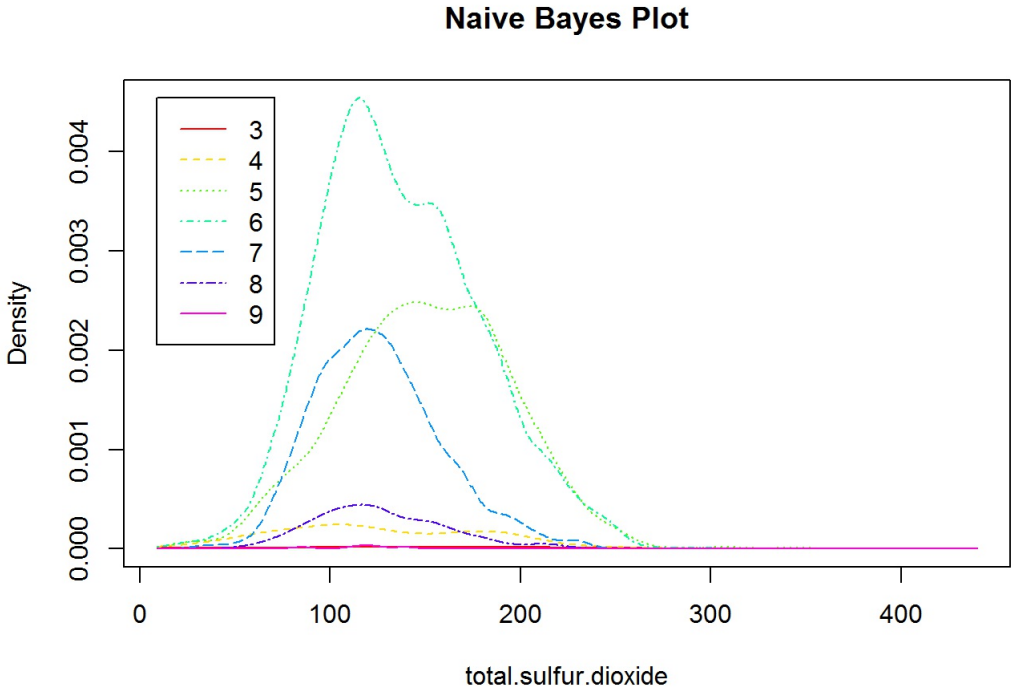

0.0
0.5
1.0
1.5

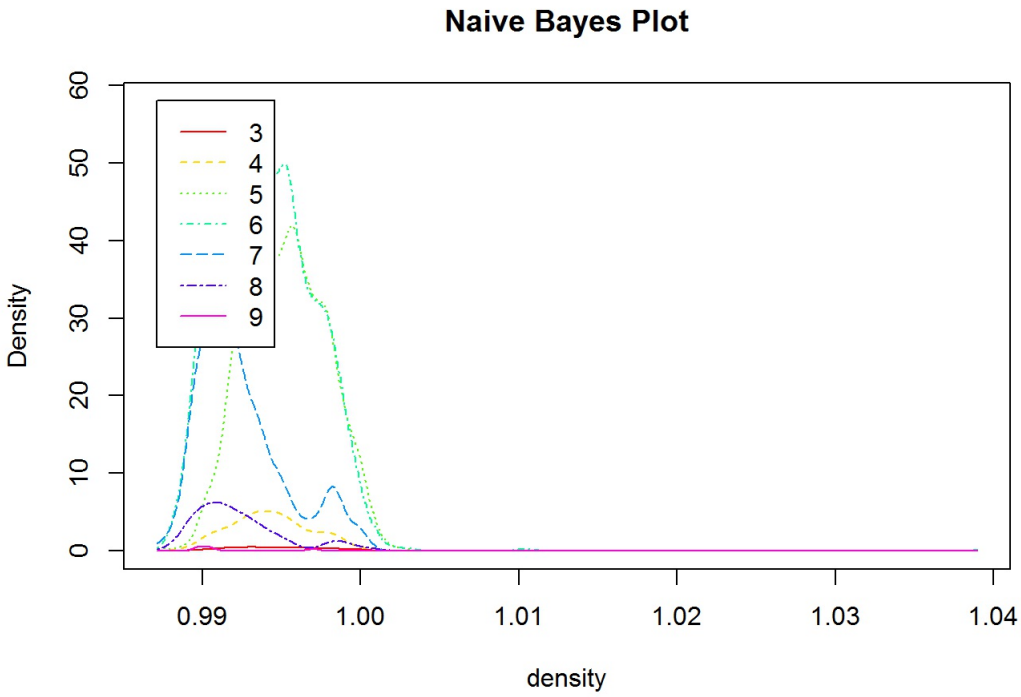
citric.acid

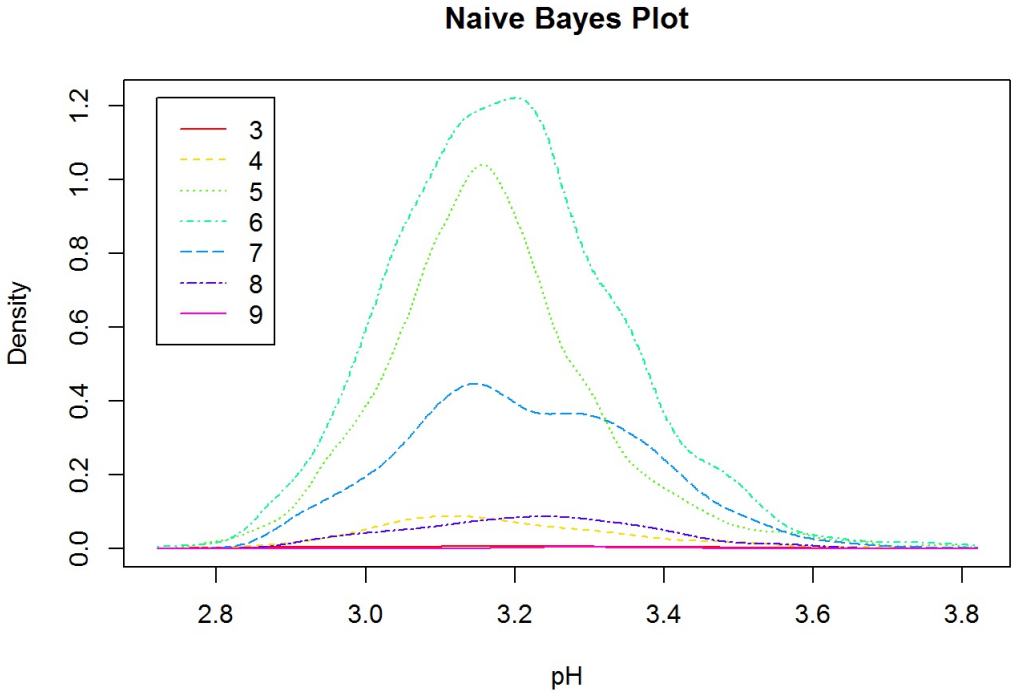

Naive Bayes Plot
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
3
4
5
6
7
8
9
0
10
20
30
40
50
60
residual.sugar

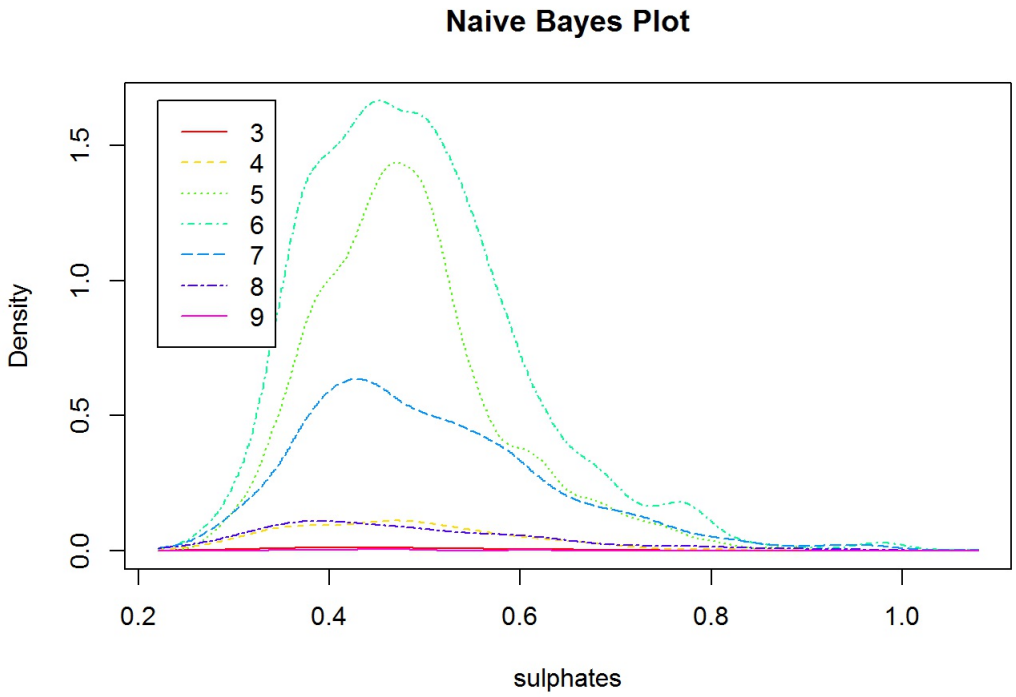

Naive Bayes Plot
Density
0
5
10
15
3
4
5
6
7
8
9
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
chlorides

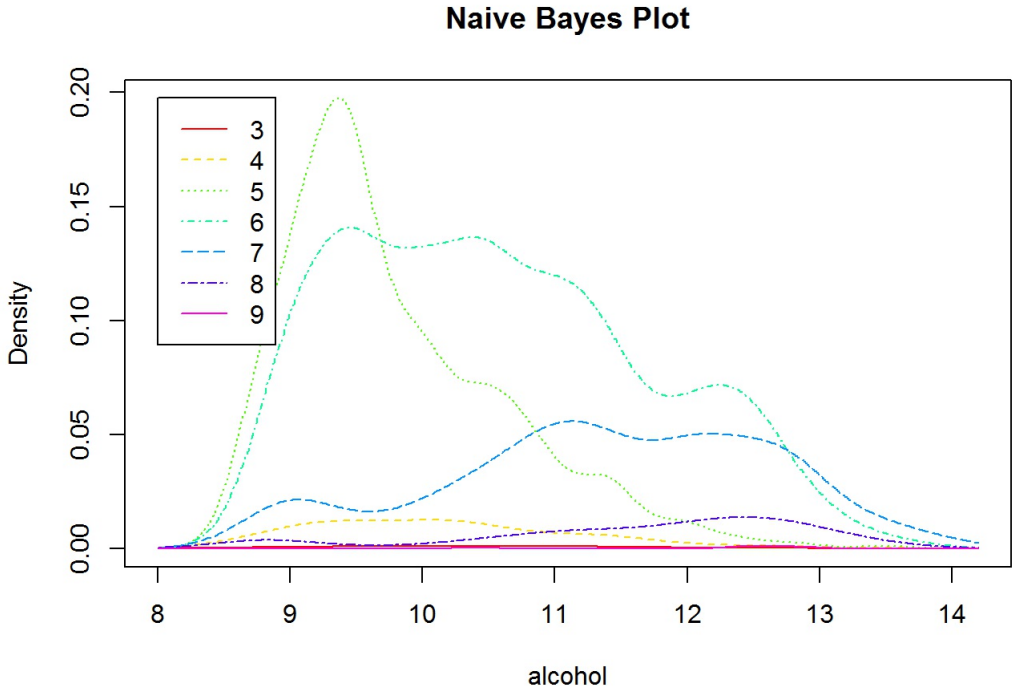

Naive Bayes Plot
Density
0.000
0.004
0.008
0.012
0
50
100
150
200
250
300
free.sulfur.dioxide
3
4
5
6
7
8
9


Naive Bayes Plot
Density
0.000
0.001
0.002
0.003
0.004
0
100
200
300
400
total.sulfur.dioxide
3
4
5
6
7
8
9


Naive Bayes Plot
Density
density
3
4
5
6
7
8
9
0
10
20
30
40
50
60
0.99
1.00
1.01
1.02
1.03
1.04


Naive Bayes Plot
Density
pH
3
4
5
6
7
8
9
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2.8
3.0
3.2
3.4
3.6
3.8

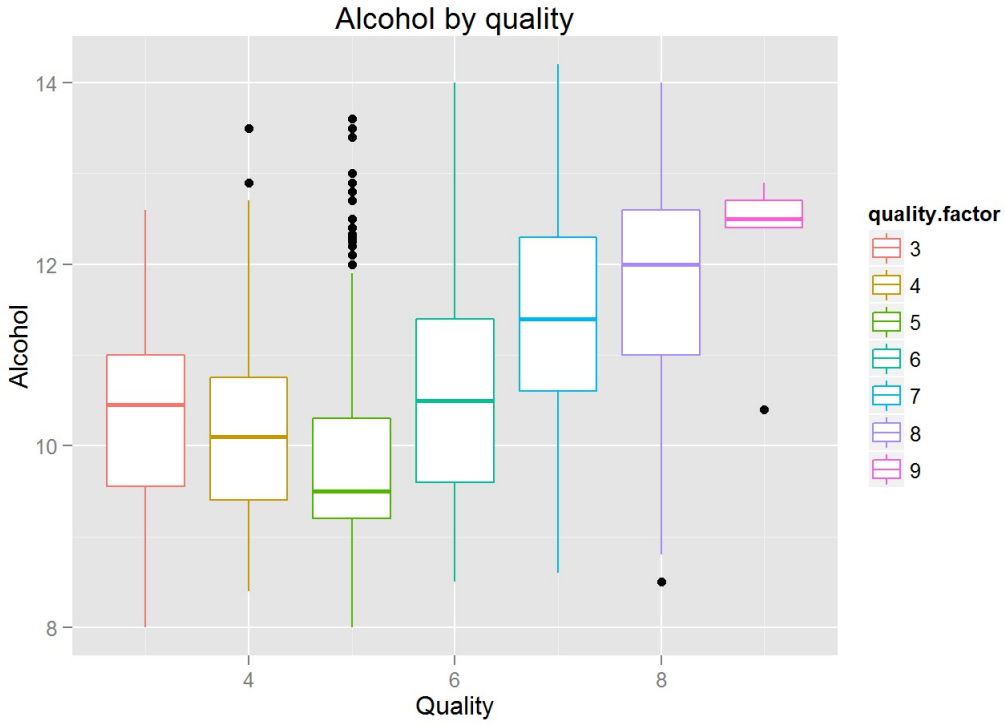




From densities we can easily observe that the main variable for wine quality separation is alcohol.

Final plots and Summary

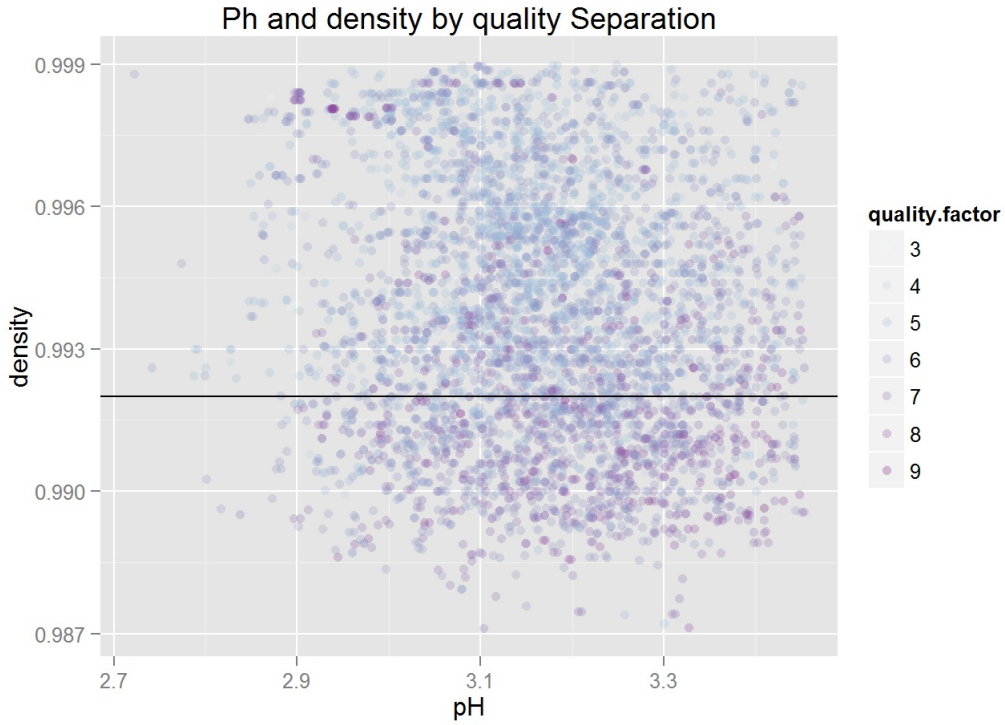
Plot One



Description One

The median of alcohol variable by quality is likely to be higher for higher quality white wine. This follows that one of the main features of high quality wine is highthe percent alcohol content of this wine.

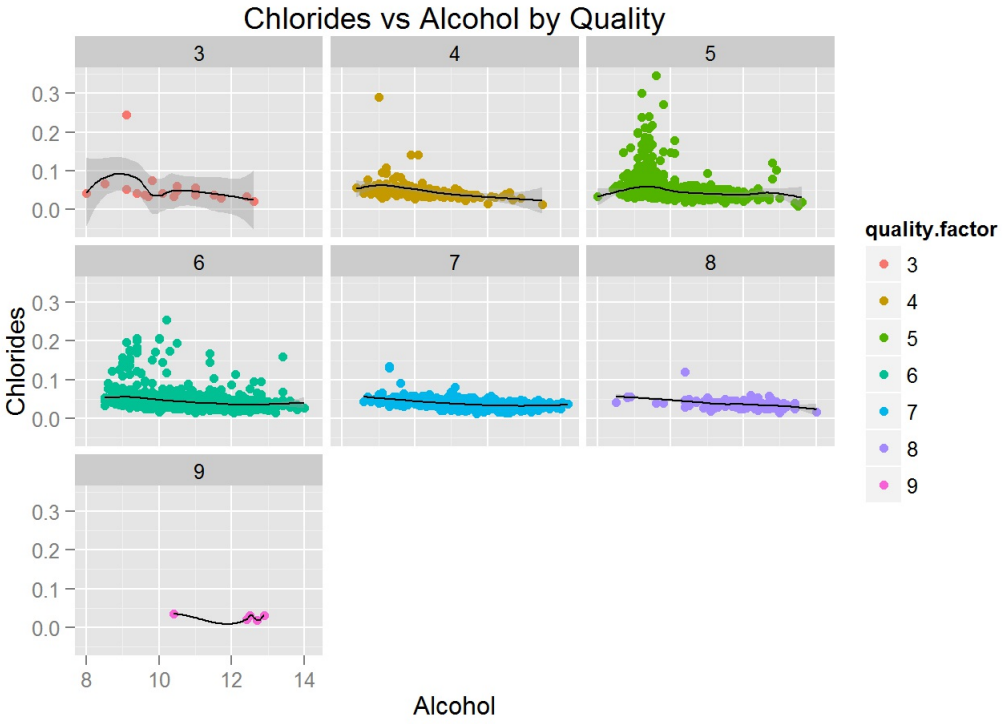
Plot Two



Description Two

There is exists some soft separation line between higher and lower quality wines. One of this lines can be straight line for density equal .992. Below .992 pH the most of wines with high quality, otherwise low quality.

Plot Three



Description Three

This plot gives us more specific visual separation between low and high quality wine. High quality wine in general have the lowest chlorides and alcohol more than 10%. And the lowest quality wine have more chlorides and low alcohol. Also we can see unusual and interesting peaks for 5 and six quality wines. Also we can see nearly identical smooth lines for quality 6, 7, 8.

Reflection

The white wine quality data set contains information on almost 4898 wines, their chemical properties and wine quality from best experts (i believe). I've asked a question how we can predict wine quality using only information about chemical properties of this wine. Quality measures from 0 (worst) to 10 (best). I started by understanding the individual variables in the data set and their influence on wine quality. I've transformed quality from numeric to ordered factor. During exploration I've found some linear patterns how to separate low and high quality wines. The highest influence on wine quality is alcohol content in wine, it's has the highest correlation and density separation. During correlation analysis I've found four important variables for this task are Alcohol, pH, density and chlorides. This variables I've included in simple naive bayes model for predicting wine quality. I've obtained 0.4888546 accuracy. It's quite high, but for initial result is ok. From multivariate plots i can conclude, that there are non linear patterns in this data set. More better model for prediction is SVM, it's gives high accuracy as described in this article http://www3.dsi.uminho.pt/pcortez/white.pdf (http://www3.dsi.uminho.pt/pcortez/white.pdf).

List of Resources

1. Chunk Options. Url: http://yihui.name/knitr/options/#chunk_options (http://yihui.name/knitr/options/#chunk_options).
2. Wine quality Prediction. Url: http://www3.dsi.uminho.pt/pcortez/white.pdf (http://www3.dsi.uminho.pt/pcortez/white.pdf).