Programming R Final

Matthias Rathbun

2022-10-30

# Set up the work space

First I want to reduce the number of displayed digits.

options(  
 digits = 4  
)

# Import Data

Import dataset as a dataframe

Wine <- read.csv(  
 file = "winequality-red.csv",  
 header = TRUE  
)

## Check Dataset Structure for tidyness

head(Wine[,-1],10)

## volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide  
## 1 0.70 0.00 1.9 0.076 11  
## 2 0.88 0.00 2.6 0.098 25  
## 3 0.76 0.04 2.3 0.092 15  
## 4 0.28 0.56 1.9 0.075 17  
## 5 0.70 0.00 1.9 0.076 11  
## 6 0.66 0.00 1.8 0.075 13  
## 7 0.60 0.06 1.6 0.069 15  
## 8 0.65 0.00 1.2 0.065 15  
## 9 0.58 0.02 2.0 0.073 9  
## 10 0.50 0.36 6.1 0.071 17  
## total.sulfur.dioxide density pH sulphates alcohol quality  
## 1 34 0.9978 3.51 0.56 9.4 5  
## 2 67 0.9968 3.20 0.68 9.8 5  
## 3 54 0.9970 3.26 0.65 9.8 5  
## 4 60 0.9980 3.16 0.58 9.8 6  
## 5 34 0.9978 3.51 0.56 9.4 5  
## 6 40 0.9978 3.51 0.56 9.4 5  
## 7 59 0.9964 3.30 0.46 9.4 5  
## 8 21 0.9946 3.39 0.47 10.0 7  
## 9 18 0.9968 3.36 0.57 9.5 7  
## 10 102 0.9978 3.35 0.80 10.5 5

Data is Clean and Tidy, each column is a feature, each row is a unique observation, and every cell has a single value.

# Visualize Data

## Corr Plot

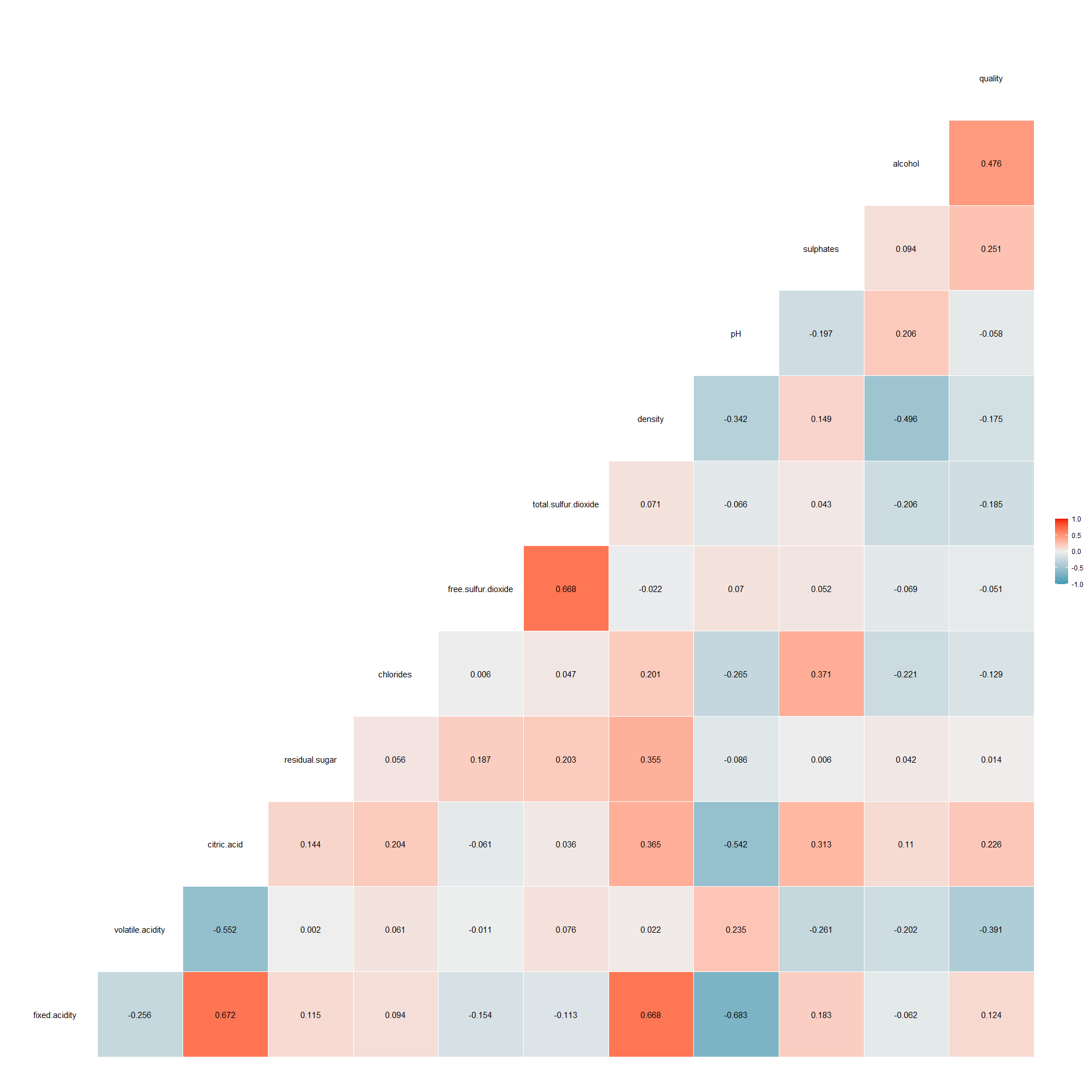
I would Like to See the correlations between features as this is a regression problem

library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

ggcorr(Wine, label = TRUE, label\_round = 3)

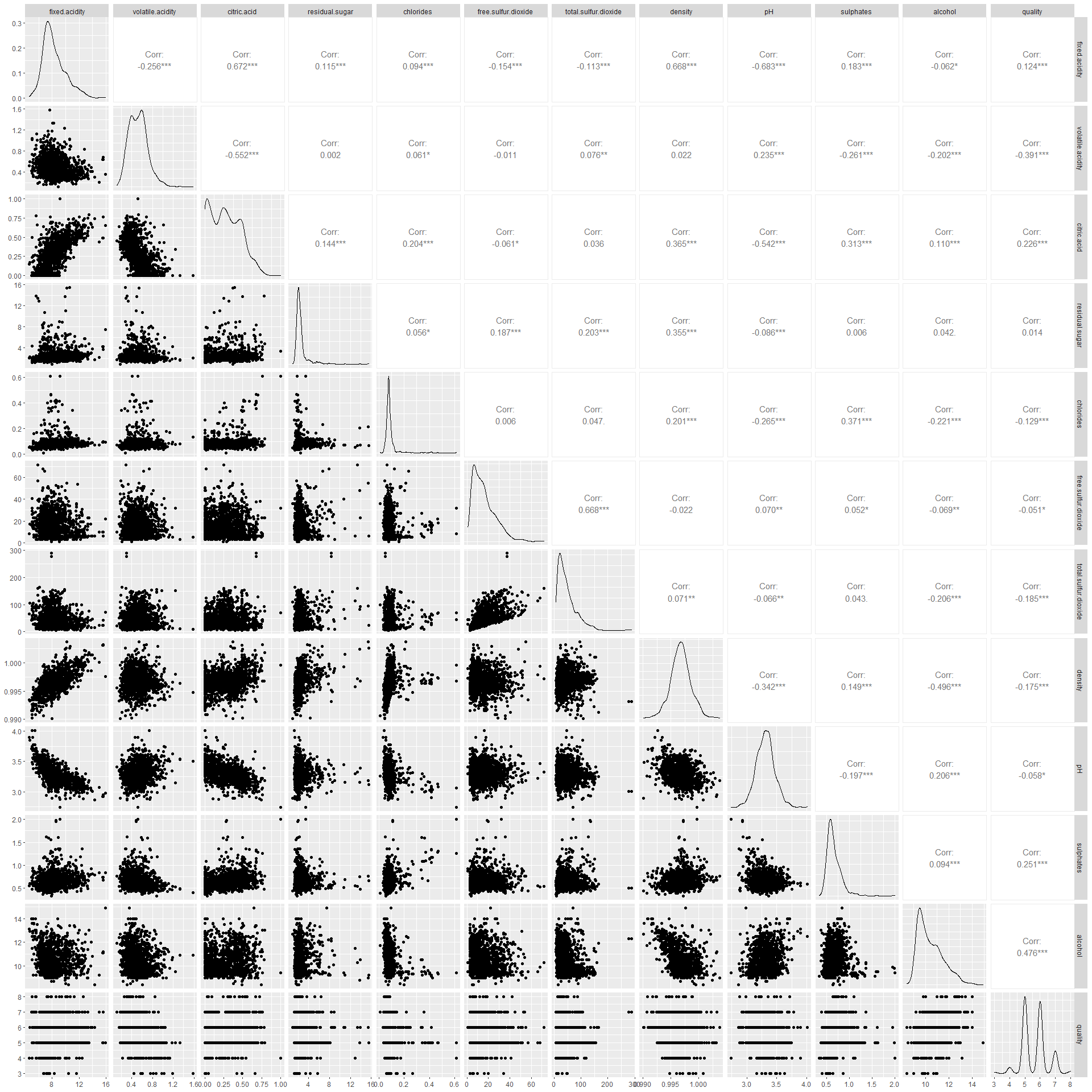


ggsave("Corrplot.pdf")

## Saving 20 x 20 in image

## Pairplot

ggpairs(Wine)



ggsave("pairplot.pdf")

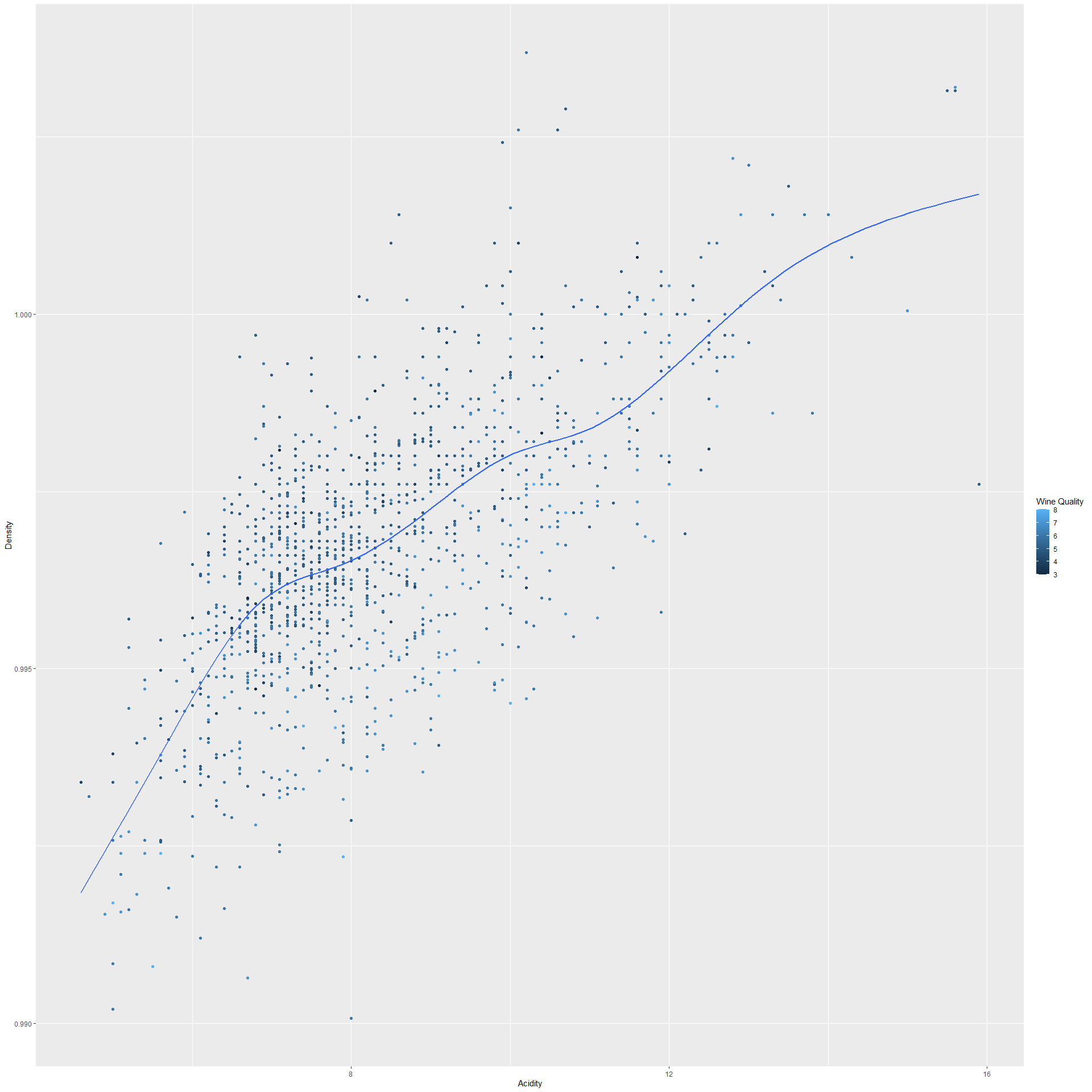
## Saving 20 x 20 in image

Usefull for seing distributions and for correlations between variables.

# Scatter Plot with Loess

ggplot(Wine, aes(fixed.acidity, density)) +  
 geom\_point(aes(colour = quality)) +  
 geom\_smooth(se = FALSE) +  
 labs(  
 x = "Acidity",  
 y = "Density",  
 colour = "Wine Quality"  
 )

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



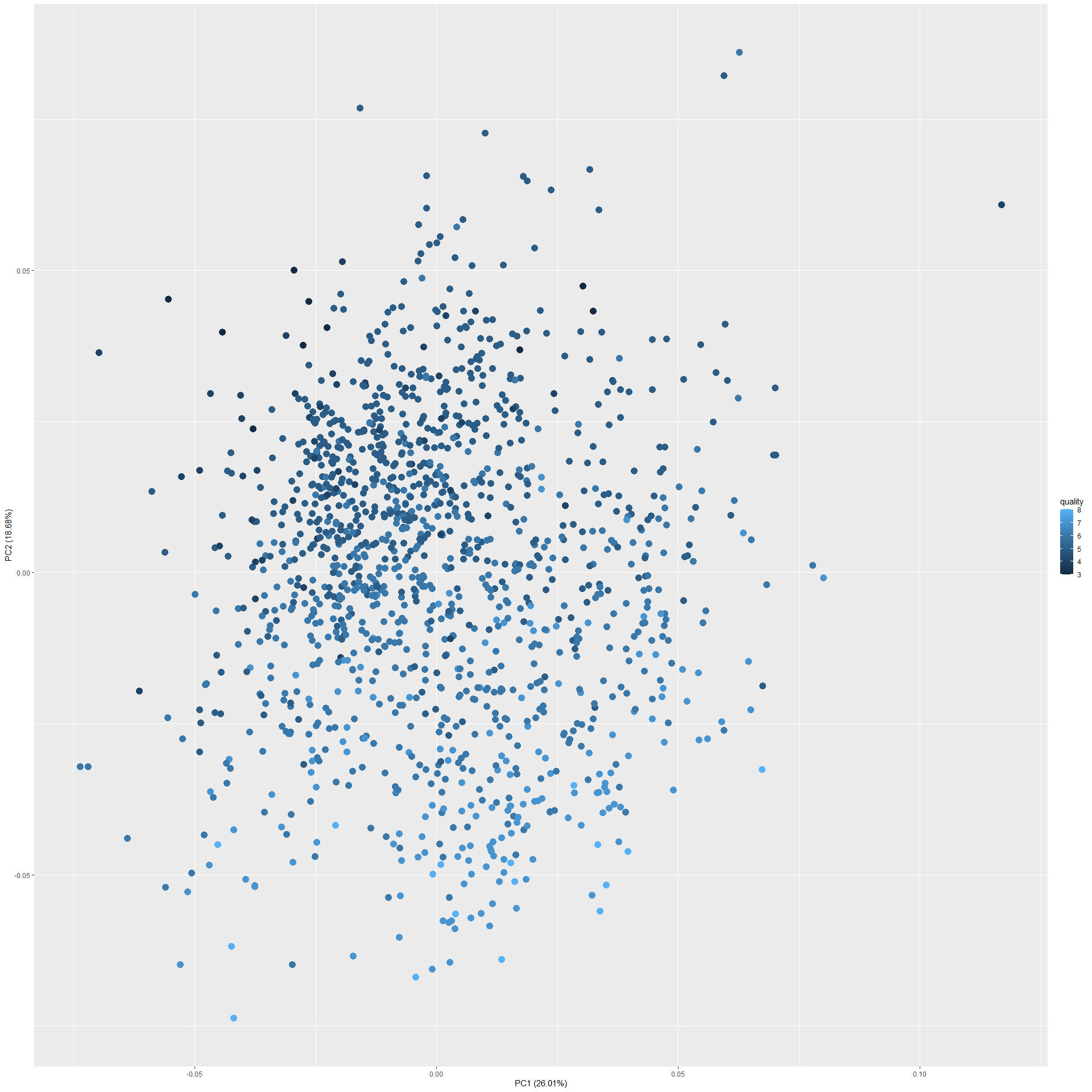
ggsave("scatter.pdf")

## Saving 20 x 20 in image  
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

This shows some trend between Acidity and density of wine. These were chosen as they already had high correlation based on the corrplot

# Principle Component Analysis

library(ggfortify)  
pca\_res <- prcomp(Wine, scale. = TRUE)  
  
autoplot(pca\_res, data = Wine , colour = 'quality', size = 4)



ggsave("pca.pdf")

## Saving 20 x 20 in image

Reduce 11 Features to 2 axes the higher quality wines seem to be at the bottom left part of the plot. Lower quality wines are clustered together.