Data Visualization Assignment 4 (Data Analysis)

Jingyu Wang 7701969

April 19, 2019

Two most common wine, red and white wine, data are included with eleven variables which can affect the quality of both wine. We want to know which variables are most important and really affecr the quality of wine.

```
library(ggplot2)
#library(GGally)
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------- tidyverse 1.2.1 --
```

```
## v tibble  2.0.1       v purrr   0.2.5
## v tidyr   0.8.2       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.3.1
## v tibble  2.0.1       v forcats 0.3.0
```

```
## -- Conflicts -------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Users/summe/Downloads/Stat-7350-class
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

Reading data

```
#here()
red_wine<-read_csv(here("Assignemnt 4_Final project","Data","red wine.csv"),
                   col_types = cols())
white_wine<-read_csv(here("Assignemnt 4_Final project","Data","white wine.csv"),
                     col_types = cols())
```

Combining white and red wine data and expressed by group.

```
# Group 1 is red wine and group 2 is white wine.
comb_wines<-rbind(red_wine,white_wine)
group<-c(rep("red",times=1599),rep("white",times=4898))
wines_new<-cbind(comb_wines,group)
```

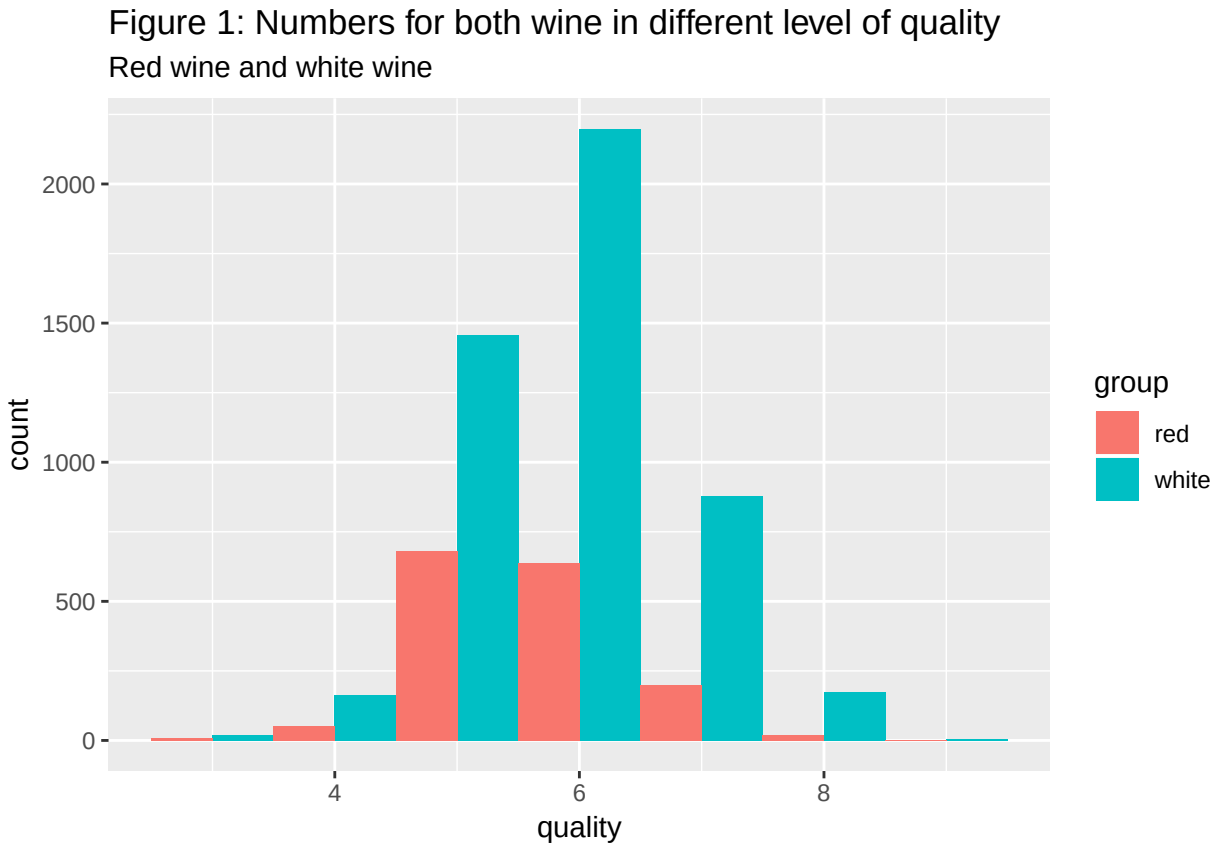
Reshape dataset for both wines

```
# red wine
red_wine_m<- melt(red_wine,id.vars='quality',
                  measure.vars=c("fixed acidity","volatile acidity",
                                 "citric acid","residual sugar","chlorides",
                                 "free sulfur dioxide","total sulfur dioxide",
                                 "density", "pH", "sulphates","alcohol"))


# white wine
white_wine_m<- melt(white_wine,id.vars='quality',
                  measure.vars=c("fixed acidity","volatile acidity",
                                 "citric acid","residual sugar","chlorides",
                                 "free sulfur dioxide","total sulfur dioxide",
                                 "density", "pH", "sulphates","alcohol"))
```

Counting of both wine in different level of quality

```r
ggplot(wines_new, aes(quality, fill = group))+
  geom_histogram(binwidth = 1,position = "dodge")+
  ggtitle("Figure 1: Numbers for both wine in different level of quality",
          subtitle = "Red wine and white wine")
```



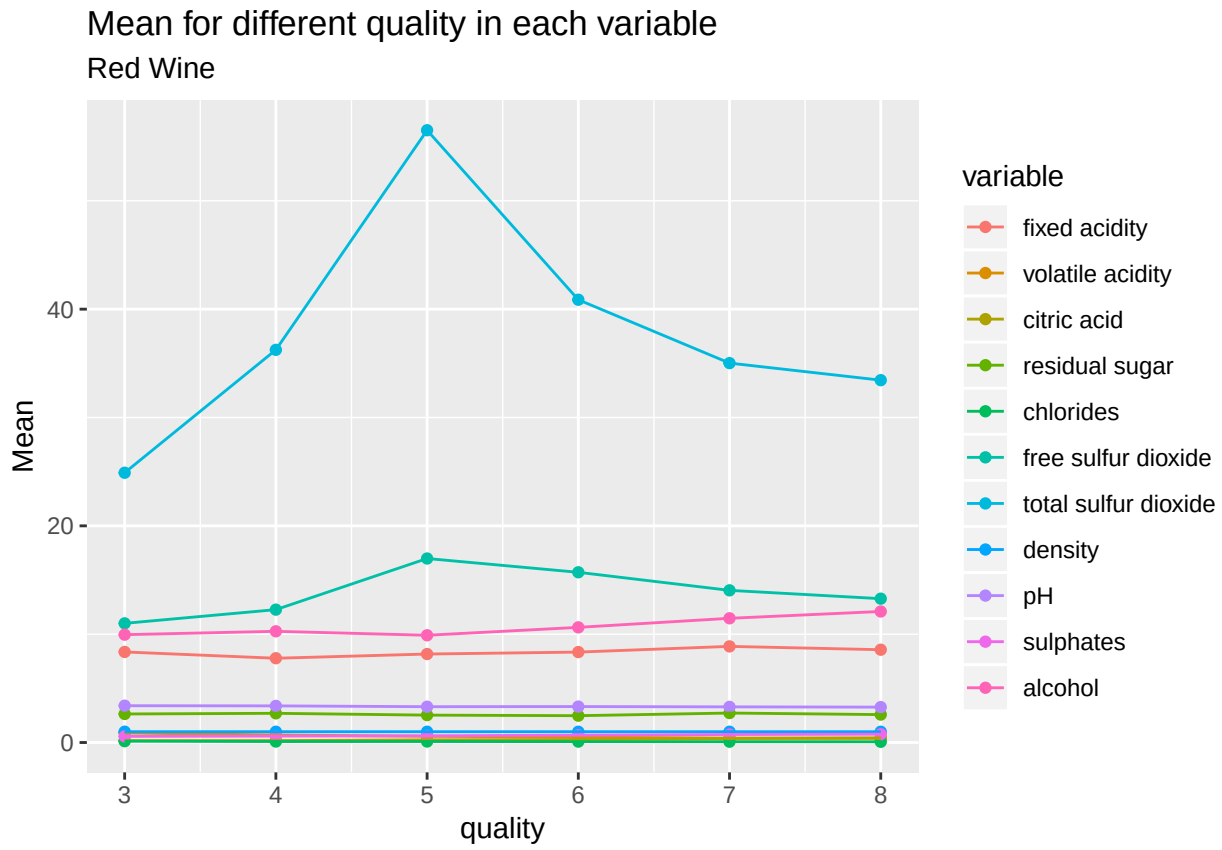

```r
# com.dat<-rbind(red_wine_m,white_wine_m)
# dim(com.dat)
# group2<-c(rep("red",times=17589),rep("white",times=53878))
# wines_new_m<-cbind(com.dat,group2)
#
# t.w.m<-wines_new_m%>%
#   filter(variable%in%c("fixed acidity","volatile acidity",
#                                "citric acid","residual sugar","chlorides",
#                                "free sulfur dioxide","total sulfur dioxide",
#                                "density", "pH", "sulphates","alcohol"))%>%
#   group_by(quality,variable,group2)%>%
#   summarise(Mean=mean(value))
```

```r
# ggplot(t.r, mapping = aes(x=quality, y=Mean, fill = group2))+
#   geom_point()+
#   geom_line()+
#   facet_wrap(~variable, scales = "free_y")
```

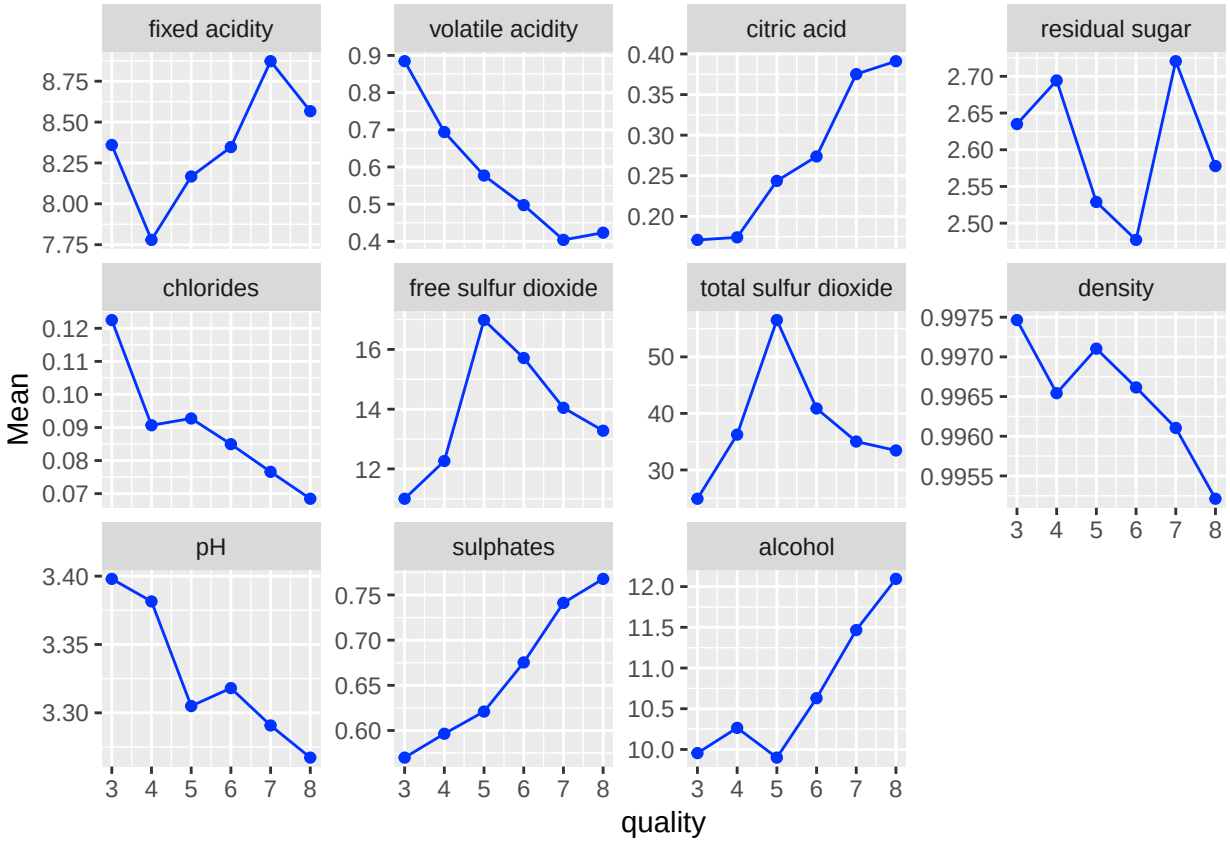
Mean for different quality in each variable(red wine)

```
t.r<-red_wine_m%>%
  filter(variable%in%c("fixed acidity","volatile acidity",
                                  "citric acid","residual sugar","chlorides",
                                  "free sulfur dioxide","total sulfur dioxide",
                                  "density", "pH", "sulphates","alcohol"))%>%
  group_by(quality,variable)%>%
  summarise(Mean=mean(value))


ggplot(t.r, mapping = aes(x=quality, y=Mean, colour= variable))+
  geom_point()+
  geom_line()+
  ggtitle("Mean for different quality in each variable", subtitle = "Red Wine")
```



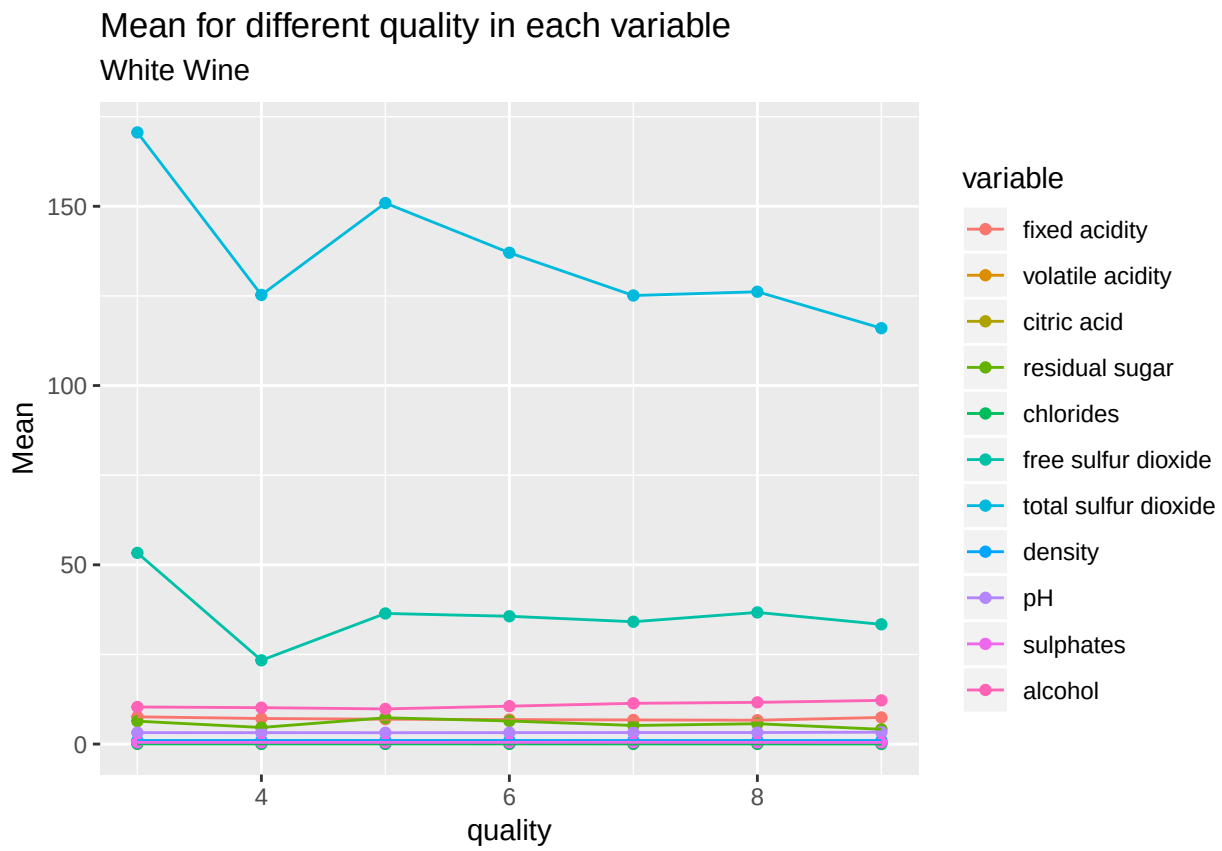

```
ggplot(t.r, mapping = aes(x=quality, y=Mean))+
  geom_point(colour = "#0033FF")+
  geom_line(colour = "#0033FF")+
  facet_wrap(~variable, scales = "free_y")
```



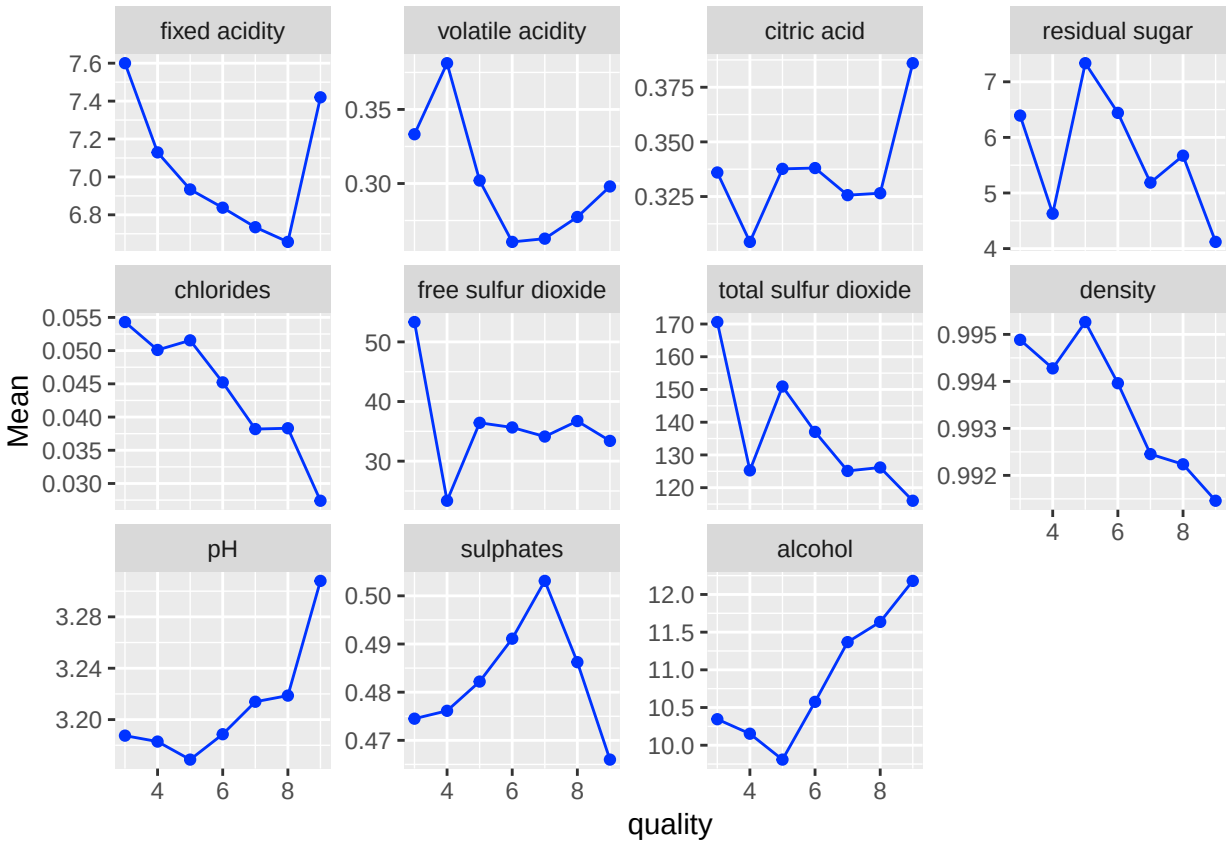
Mean changing of each variable in different quality(white wine)

```
t.w<-white_wine_m%>%
  filter(variable%in%c("fixed acidity","volatile acidity",
                                "citric acid","residual sugar","chlorides",
                                "free sulfur dioxide","total sulfur dioxide",
                                "density", "pH", "sulphates","alcohol"))%>%
  group_by(quality,variable)%>%
  summarise(Mean=mean(value))

ggplot(t.w, mapping = aes(x=quality, y=Mean, colour= variable))+
  geom_point()+
  geom_line()+
  ggtitle("Mean for different quality in each variable", subtitle = "White Wine")
```




```
ggplot(t.w, mapping = aes(x=quality, y=Mean))+
  geom_point(colour = "#0033FF")+
  geom_line(colour = "#0033FF")+
  facet_wrap(~variable, scales = "free_y")
```



Multiple Regression

Red wine

```
#red_wine$quality<-as.factor(red_wine$quality)
model.red1<-lm(quality~`fixed acidity` + `volatile acidity` +
                `citric acid` + `residual sugar` + `chlorides` +
                `free sulfur dioxide` + `total sulfur dioxide` +
                `density` + `pH` + `sulphates`+
                `alcohol`, data = red_wine)
summary(model.red1)
```

```
##
## Call:
## lm(formula = quality ~ `fixed acidity` + `volatile acidity` +
##     `citric acid` + `residual sugar` + chlorides + `free sulfur dioxide` +
##     `total sulfur dioxide` + density + pH + sulphates + alcohol,
##     data = red_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.197e+01  2.119e+01   1.036   0.3002
## `fixed acidity`         2.499e-02  2.595e-02   0.963   0.3357
## `volatile acidity`     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## `citric acid`          -1.826e-01  1.472e-01  -1.240   0.2150
## `residual sugar`        1.633e-02  1.500e-02   1.089   0.2765
## chlorides              -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## `free sulfur dioxide`   4.361e-03  2.171e-03   2.009   0.0447 *
## `total sulfur dioxide` -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density                -1.788e+01  2.163e+01  -0.827   0.4086
## pH                     -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates               9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol                 2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
# step(model.red1, direction= "both")

model.red2<-lm(quality ~ `volatile acidity` + `chlorides` +
                `free sulfur dioxide`+ `total sulfur dioxide` +
                `pH`+ `sulphates`+ `alcohol`, data=red_wine)
summary(model.red2)
```

```
##
## Call:
## lm(formula = quality ~ `volatile acidity` + chlorides + `free sulfur dioxide` +
##     `total sulfur dioxide` + pH + sulphates + alcohol, data = red_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.4300987  0.4029168  10.995  < 2e-16 ***
## `volatile acidity`     -1.0127527  0.1008429 -10.043  < 2e-16 ***
## chlorides              -2.0178138  0.3975417  -5.076 4.31e-07 ***
## `free sulfur dioxide`   0.0050774  0.0021255   2.389    0.017 *
## `total sulfur dioxide` -0.0034822  0.0006868  -5.070 4.43e-07 ***
## pH                     -0.4826614  0.1175581  -4.106 4.23e-05 ***
## sulphates               0.8826651  0.1099084   8.031 1.86e-15 ***
## alcohol                 0.2893028  0.0167958  17.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

```r
anova(model.red1,model.red2)
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ `fixed acidity` + `volatile acidity` + `citric acid` +
##     `residual sugar` + chlorides + `free sulfur dioxide` + `total sulfur dioxide` +
##     density + pH + sulphates + alcohol
## Model 2: quality ~ `volatile acidity` + chlorides + `free sulfur dioxide` +
##     `total sulfur dioxide` + pH + sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1587 666.41
## 2   1591 667.54 -4   -1.1264 0.6706 0.6124
```

White wine

```r
#white_wine$quality<-as.factor(white_wine$quality)
model.white1<-lm(quality~`fixed acidity` + `volatile acidity` +
                `citric acid` + `residual sugar` + `chlorides` +
                `free sulfur dioxide` + `total sulfur dioxide` +
                `density` + `pH` + `sulphates`+
                `alcohol`, data = white_wine)
summary(model.white1)
```

```
##
## Call:
## lm(formula = quality ~ `fixed acidity` + `volatile acidity` +
##     `citric acid` + `residual sugar` + chlorides + `free sulfur dioxide` +
##     `total sulfur dioxide` + density + pH + sulphates + alcohol,
##     data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.502e+02  1.880e+01   7.987 1.71e-15 ***
## `fixed acidity`        6.552e-02  2.087e-02   3.139  0.00171 **
## `volatile acidity`    -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## `citric acid`          2.209e-02  9.577e-02   0.231  0.81759
## `residual sugar`       8.148e-02  7.527e-03  10.825  < 2e-16 ***
## chlorides             -2.473e-01  5.465e-01  -0.452  0.65097
## `free sulfur dioxide`  3.733e-03  8.441e-04   4.422 9.99e-06 ***
## `total sulfur dioxide` -2.857e-04  3.781e-04  -0.756  0.44979
## density               -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                     6.863e-01  1.054e-01   6.513 8.10e-11 ***
## sulphates              6.315e-01  1.004e-01   6.291 3.44e-10 ***
## alcohol                1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
```

```
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

```
model.white2<-lm(quality ~ `fixed acidity` + `volatile acidity` +
                   `residual sugar` + `free sulfur dioxide` +
                   `density` + `pH` + `sulphates` + `alcohol`,
                 data=white_wine)
summary(model.white2)
```

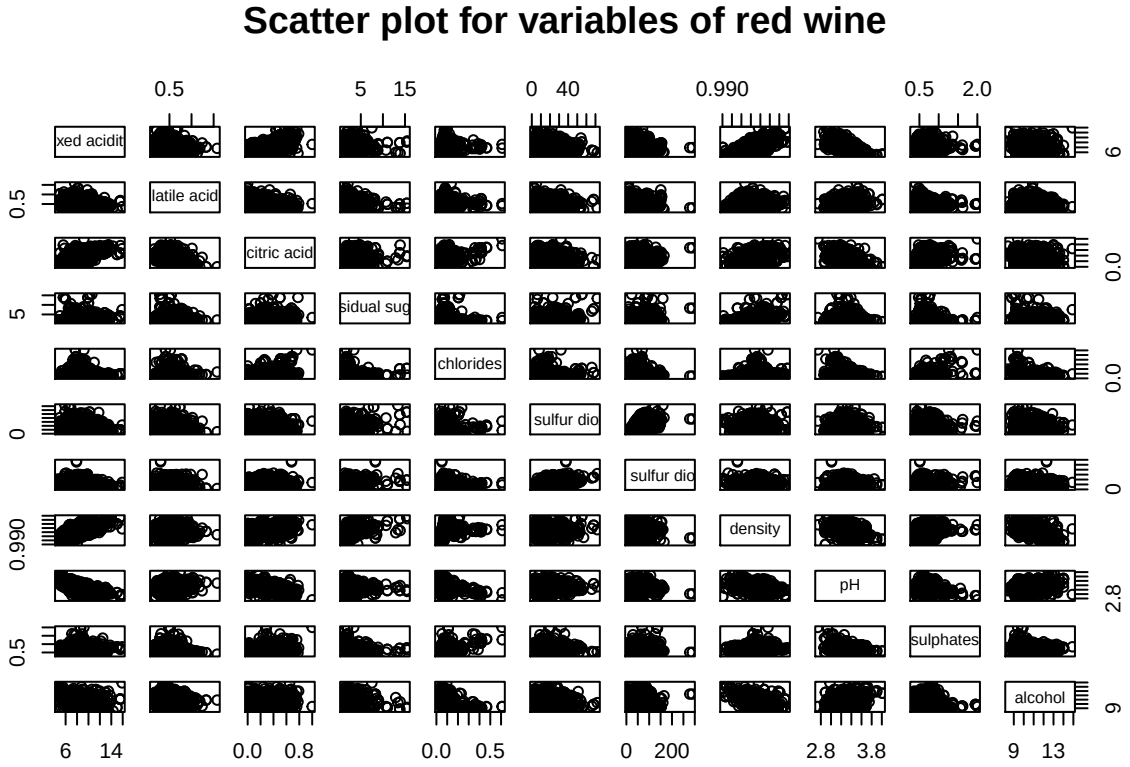
```
##
## Call:
## lm(formula = quality ~ `fixed acidity` + `volatile acidity` +
##     `residual sugar` + `free sulfur dioxide` + density + pH +
##     sulphates + alcohol, data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8246 -0.4938 -0.0396  0.4660  3.1208
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.541e+02  1.810e+01   8.514  < 2e-16 ***
## `fixed acidity`        6.810e-02  2.043e-02   3.333 0.000864 ***
## `volatile acidity`    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
## `residual sugar`       8.285e-02  7.287e-03  11.370  < 2e-16 ***
## `free sulfur dioxide`  3.349e-03  6.766e-04   4.950 7.67e-07 ***
## density               -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
## pH                     6.942e-01  1.034e-01   6.717 2.07e-11 ***
## sulphates              6.285e-01  9.997e-02   6.287 3.52e-10 ***
## alcohol                1.932e-01  2.408e-02   8.021 1.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7512 on 4889 degrees of freedom
## Multiple R-squared:  0.2818, Adjusted R-squared:  0.2806
## F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

```
anova(model.white1,model.white2)
```

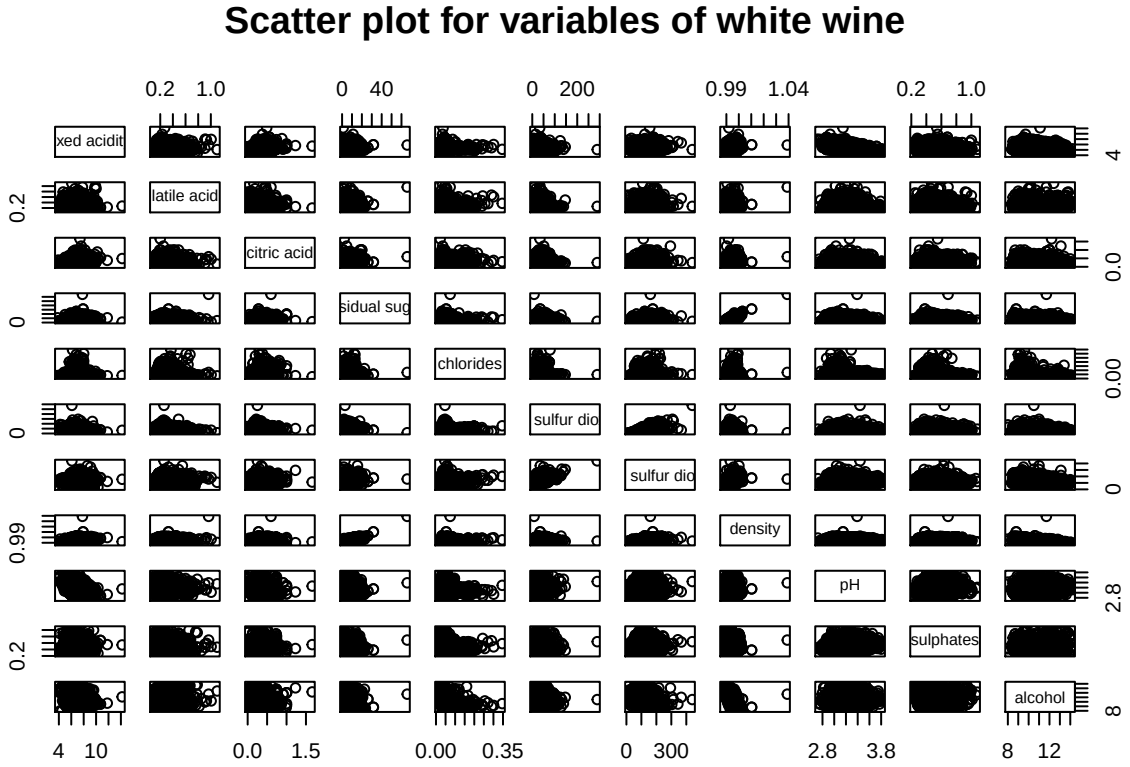
```
## Analysis of Variance Table
##
## Model 1: quality ~ `fixed acidity` + `volatile acidity` + `citric acid` +
##     `residual sugar` + chlorides + `free sulfur dioxide` + `total sulfur dioxide` +
##     density + pH + sulphates + alcohol
## Model 2: quality ~ `fixed acidity` + `volatile acidity` + `residual sugar` +
##     `free sulfur dioxide` + density + pH + sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4886 2758.3
## 2   4889 2758.8 -3  -0.45477 0.2685 0.8481
```

Scatter plot for multivariable

```
pairs(red_wine[,1:11], main = "Scatter plot for variables of red wine")
```



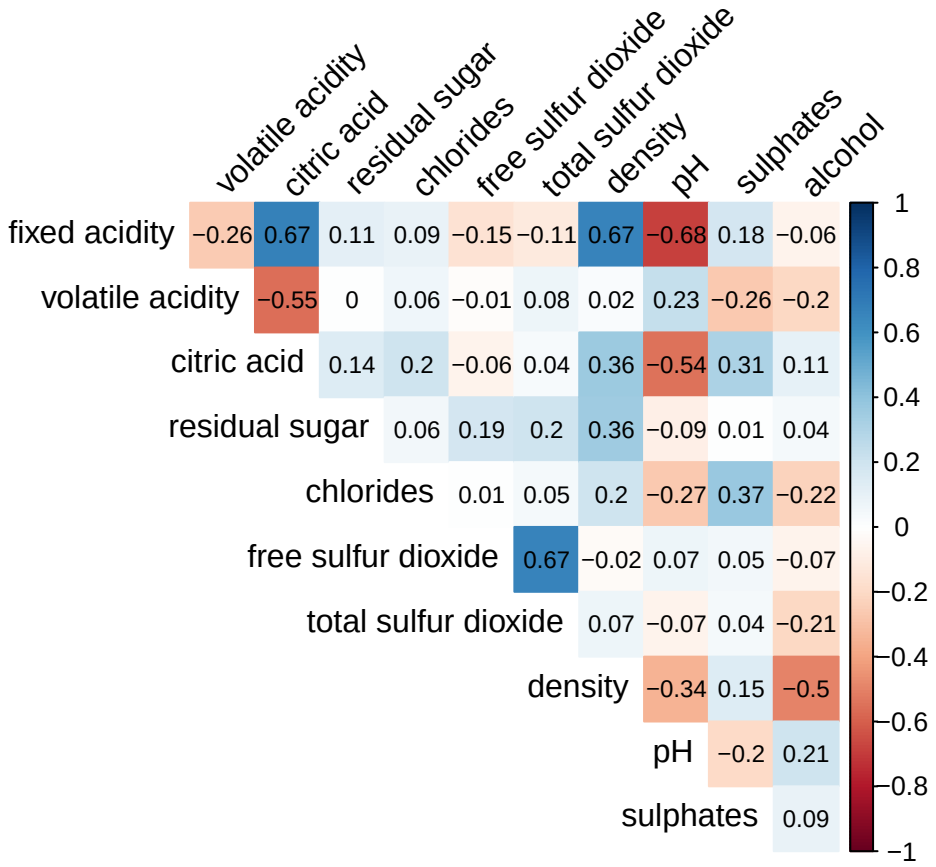

```
pairs(white_wine[,1:11],main = "Scatter plot for variables of white wine")
```



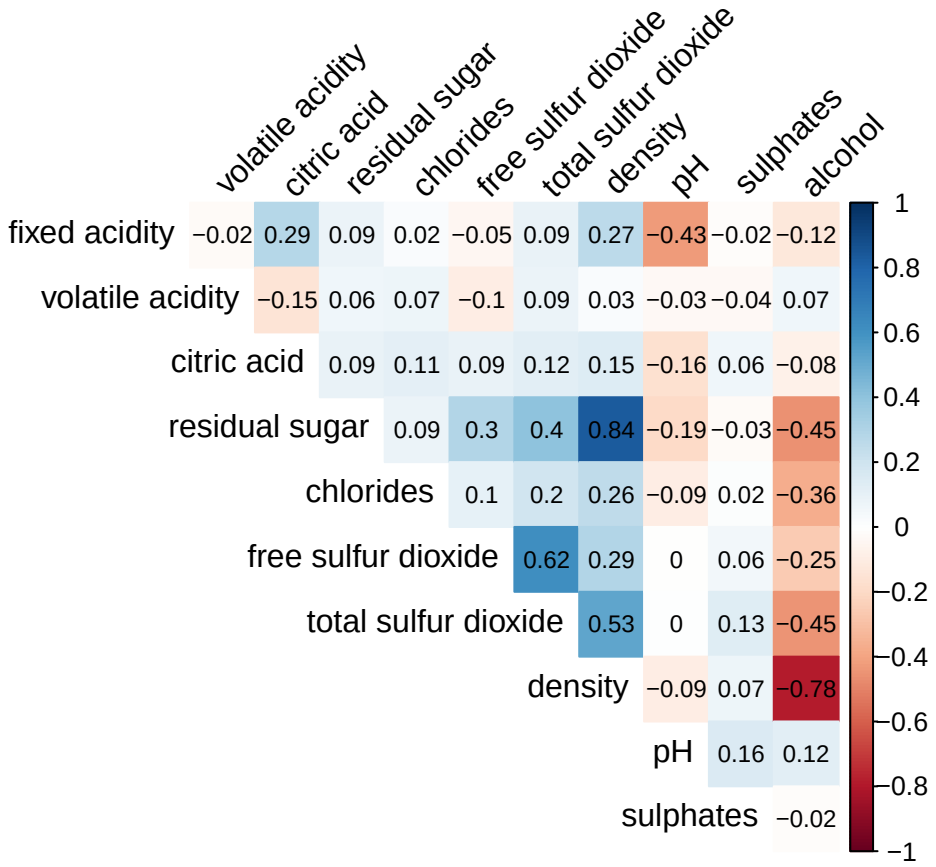

```
# ggpairs(red_wine,columns =,c("fixed acidity","volatile acidity",
#                              "citric acid","residual sugar","chlorides",
#                                   "free sulfur dioxide","total sulfur dioxide",
#                                   "density", "pH", "sulphates","alcohol"),
#         lower = list(continuous = "points"),
#         upper = list(continuous = "blank"),
#         axisLabels = "none")
```

Correlation and its plot

```
varib_cor_R<-cor(red_wine[,1:11])
corrplot(varib_cor_R,method="color",type="upper",
         addCoef.col = "black", tl.col="black",
         tl.srt=45,insig = "blank", diag=FALSE,
         number.font = 6,number.cex = 0.75)
```




```
varib_cor_W<-cor(white_wine[,1:11])
corrplot(varib_cor_W,method="color",type="upper",
         addCoef.col = "black", tl.col="black",
         tl.srt=45,insig = "blank", diag=FALSE,
         number.font = 6,number.cex = 0.75)
```




```
# cor.test(red_wine$`fixed acidity`,red_wine$density)
# cor.test(red_wine$alcohol,red_wine$pH)
# cor.test(red_wine$`volatile acidity`,red_wine$`residual sugar`)
#
# ##assumption
# shapiro.test(red_wine$alcohol)
# shapiro.test(red_wine$pH)
# ggqqplot(red_wine$pH, ylab = "pH")
# ggqqplot(red_wine$alcohol, ylab = "alcohol")


# library("ggpubr")
# ggscatter(red_wine, x = "alcohol", y = "density",
#           add = "reg.line",cor.method = "pearson")
```

Principal component Analysis

Checking the fitted regression model for multicolinearity (full model)

```
vif(model.red1)
```

```
##        `fixed acidity`     `volatile acidity`           `citric acid`
##               7.767512               1.789390               3.128022
##       `residual sugar`              chlorides  `free sulfur dioxide`
##               1.702588               1.481932               1.963019
```

```
## `total sulfur dioxide`                density                     pH
##               2.186813               6.343760               3.329732
##              sulphates                alcohol
##               1.429434               3.031160
```

```
vif(model.white1)
```

```
##        `fixed acidity`     `volatile acidity`          `citric acid`
##               2.691435               1.141156               1.165215
##       `residual sugar`              chlorides  `free sulfur dioxide`
##              12.644064               1.236822               1.787880
## `total sulfur dioxide`                density                     pH
##               2.239233              28.232546               2.196362
##              sulphates                alcohol
##               1.138540               7.706957
```

Checking the fitted regression model for multicolinearity (reduced model)

```
vif(model.red2)
```

```
##     `volatile acidity`              chlorides  `free sulfur dioxide`
##               1.241819               1.333333               1.882706
## `total sulfur dioxide`                     pH              sulphates
##               1.943920               1.254570               1.321931
##                alcohol
##               1.220157
```

```
vif(model.white2)
```

```
##       `fixed acidity`    `volatile acidity`      `residual sugar`
##              2.579640              1.057310             11.854253
## `free sulfur dioxide`               density                    pH
##              1.149027             26.123154              2.113597
##             sulphates               alcohol
##              1.129688              7.622843
```

Some values of variance inflation factors for both wines are quite large, so multicolinearity exists in the full model for red wine as well as the full and reduced model for white wine. Thus, our multiple regression results may not be accurate.

PCA for Red Wine

To do principle component analysis, we need to calculate covariance of dataset firstly.

```
cov.red<-cov(red_wine[,1:11])
cov.red
```

```
##                       fixed acidity volatile acidity   citric acid
## fixed acidity          3.031416389    -7.985142e-02  0.2278200037
## volatile acidity      -0.079851417     3.206238e-02 -0.0192716208
## citric acid            0.227820004    -1.927162e-02  0.0379474831
## residual sugar         0.281756262     4.841910e-04  0.0394342700
## chlorides              0.007678692     5.165869e-04  0.0018687248
## free sulfur dioxide   -2.800921493    -1.967359e-02 -0.1242521139
## total sulfur dioxide  -6.482345858     4.504257e-01  0.2276972740
## density                0.002195224     7.443665e-06  0.0001341746
```

```
## pH                    -0.183585704     6.494699e-03 -0.0162975823
## sulphates              0.054010092    -7.921434e-03  0.0103277145
## alcohol               -0.114421153    -3.860022e-02  0.0228151729
##                      residual sugar     chlorides free sulfur dioxide
## fixed acidity          0.2817562623  7.678692e-03       -2.800921e+00
## volatile acidity       0.0004841910  5.165869e-04       -1.967359e-02
## citric acid            0.0394342700  1.868725e-03       -1.242521e-01
## residual sugar         1.9878971330  3.690176e-03        2.758611e+00
## chlorides              0.0036901759  2.215143e-03        2.738303e-03
## free sulfur dioxide    2.7586114522  2.738303e-03        1.094149e+02
## total sulfur dioxide   9.4164414790  7.338675e-02        2.297375e+02
## density                0.0009454109  1.782176e-05       -4.332504e-04
## pH                    -0.0186442890 -1.925745e-03        1.136531e-01
## sulphates              0.0013209414  2.961878e-03        9.159247e-02
## alcohol                0.0632189598 -1.109152e-02       -7.736984e-01
##                      total sulfur dioxide       density            pH
## fixed acidity               -6.482346e+00  2.195224e-03 -1.835857e-01
## volatile acidity             4.504257e-01  7.443665e-06  6.494699e-03
## citric acid                  2.276973e-01  1.341746e-04 -1.629758e-02
## residual sugar               9.416441e+00  9.454109e-04 -1.864429e-02
## chlorides                    7.338675e-02  1.782176e-05 -1.925745e-03
## free sulfur dioxide          2.297375e+02 -4.332504e-04  1.136531e-01
## total sulfur dioxide         1.082102e+03  4.424727e-03 -3.376988e-01
## density                      4.424727e-03  3.562029e-06 -9.956395e-05
## pH                          -3.376988e-01 -9.956395e-05  2.383518e-02
## sulphates                    2.394710e-01  4.750962e-05 -5.146186e-03
## alcohol                     -7.209298e+00 -9.979518e-04  3.383162e-02
##                         sulphates       alcohol
## fixed acidity         5.401009e-02 -0.1144211534
## volatile acidity     -7.921434e-03 -0.0386002214
## citric acid           1.032771e-02  0.0228151729
## residual sugar        1.320941e-03  0.0632189598
## chlorides             2.961878e-03 -0.0110915178
## free sulfur dioxide   9.159247e-02 -0.7736984003
## total sulfur dioxide  2.394710e-01 -7.2092978948
## density               4.750962e-05 -0.0009979518
## pH                   -5.146186e-03  0.0338316166
## sulphates             2.873262e-02  0.0169067772
## alcohol               1.690678e-02  1.1356473950
```

Because the scales of covariance for both wines are quite different, the correlation matrix should be used for principle component analysis in our case.

```
cor.red<-round(cor(red_wine[,1:11]),3)
```

We need to obtain the eigenvalues and eigenvectors of correlation matrix.

```
eig.val.red<-eigen(cor.red)$values # Pick the eigenvalues from result
round(eig.val.red,3)
```

```
##  [1] 3.100 1.925 1.551 1.212 0.960 0.659 0.584 0.423 0.345 0.181 0.060
```

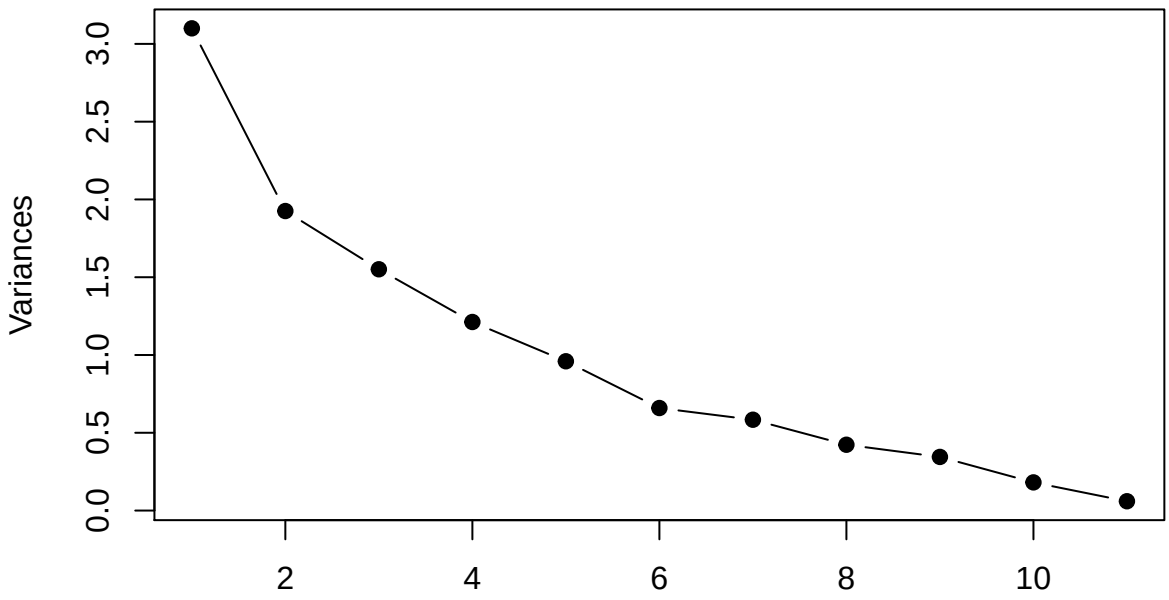
```
eig.vec.red<-eigen(cor.red)$vectors # Pick eigenvectors from result
rownames(eig.vec.red)<-colnames(red_wine[,1:11])
colnames(eig.vec.red)<-c("PC1","PC2","PC3","PC4","PC5","PC6",
                         "PC7","PC8","PC9","PC10","PC11")
```

```
round(eig.vec.red,2)
```

```
##                         PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
## fixed acidity          0.49 -0.11 -0.12  0.23  0.08  0.10  0.35  0.18 -0.19
## volatile acidity      -0.24  0.27 -0.45 -0.08 -0.22  0.41  0.53  0.08  0.13
## citric acid            0.46 -0.15  0.24  0.08  0.06  0.07 -0.10  0.38  0.38
## residual sugar         0.15  0.27  0.10  0.37 -0.73  0.05 -0.29 -0.30 -0.01
## chlorides              0.21  0.15 -0.09 -0.67 -0.25  0.31 -0.37  0.36 -0.11
## free sulfur dioxide   -0.04  0.51  0.43  0.04  0.16 -0.01  0.12  0.20 -0.64
## total sulfur dioxide   0.02  0.57  0.32  0.04  0.22  0.14  0.09 -0.02  0.59
## density                0.40  0.23 -0.34  0.17 -0.16 -0.39  0.17  0.24 -0.02
## pH                    -0.44  0.01  0.06  0.00 -0.27 -0.52  0.02  0.56  0.17
## sulphates              0.24 -0.04  0.28 -0.55 -0.23 -0.38  0.45 -0.37  0.06
## alcohol               -0.11 -0.39  0.47  0.12 -0.35  0.36  0.33  0.22 -0.04
##                         PC10  PC11
## fixed acidity          -0.25  0.64
## volatile acidity        0.36  0.00
## citric acid             0.62 -0.07
## residual sugar          0.09  0.18
## chlorides              -0.22  0.05
## free sulfur dioxide     0.25 -0.05
## total sulfur dioxide   -0.37  0.07
## density                -0.24 -0.57
## pH                     -0.01  0.34
## sulphates               0.11  0.07
## alcohol                -0.30 -0.31
```

Principle component analysis helps us reduce variables which are not important in the model. The variables we kept which really affects the response variable. In our case, we are trying to keep those kinds of variables which affects the quality of wine and fit the linear model. Thus, the following method shows the number of components whih should be kept.

```
plot(eig.val.red, type = "b", pch=19, xlab = "", ylab = "Variances")
```




```
# Proportion of variation explained by each PCs
round(eig.val.red/sum(eig.val.red),3)
```

```
##  [1] 0.282 0.175 0.141 0.110 0.087 0.060 0.053 0.038 0.031 0.016 0.005
```

```
cumsum(round(eig.val.red/sum(eig.val.red),3))
```

```
##  [1] 0.282 0.457 0.598 0.708 0.795 0.855 0.908 0.946 0.977 0.993 0.998
```

Also, we need to standardize all PCs as the explanatory variables.

```
x<-red_wine[,1:11]
x.mean<-apply(x, 2, mean)
x.mean
```

```
##        fixed acidity     volatile acidity          citric acid
##           8.31963727           0.52782051           0.27097561
##       residual sugar            chlorides  free sulfur dioxide
##           2.53880550           0.08746654          15.87492183
## total sulfur dioxide              density                   pH
##          46.46779237           0.99674668           3.31111320
##            sulphates              alcohol
##           0.65814884          10.42298311
```

```
z<-x
for (i in 1:11) {
  z[,i]<-(x[,i]-x.mean[i])/sqrt(diag(cov.red)[i])
}
```

```
red.new<-x
for (i in 1:11){ # number of pcs
  for (j in 1:1599){
    red.new[j,i]<-t(eig.vec.red[,i]) %*% t(z[j,])
  }
}

# To check the correlation in the new dataset
colnames(red.new) = c("PC1", "PC2", "PC3", "PC4","PC5",
                      "PC6","PC7","PC8","PC9","PC10","PC11")
round(cor(red.new),2)
```

```
##      PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11
## PC1    1   0   0   0   0   0   0   0   0    0    0
## PC2    0   1   0   0   0   0   0   0   0    0    0
## PC3    0   0   1   0   0   0   0   0   0    0    0
## PC4    0   0   0   1   0   0   0   0   0    0    0
## PC5    0   0   0   0   1   0   0   0   0    0    0
## PC6    0   0   0   0   0   1   0   0   0    0    0
## PC7    0   0   0   0   0   0   1   0   0    0    0
## PC8    0   0   0   0   0   0   0   1   0    0    0
## PC9    0   0   0   0   0   0   0   0   1    0    0
## PC10   0   0   0   0   0   0   0   0   0    1    0
## PC11   0   0   0   0   0   0   0   0   0    0    1
```

Regression analysis with all standardized PCs

```
dim(red.new)
```

```
## [1] 1599   11
```

```
red.new.dat<-cbind(red_wine$quality,red.new)

pc.model<-lm(red_wine$quality~PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10+PC11,
             data=red.new.dat)
summary(pc.model)
```

```
##
## Call:
## lm(formula = red_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 +
##     PC6 + PC7 + PC8 + PC9 + PC10 + PC11, data = red.new.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.636023   0.016205 347.788  < 2e-16 ***
## PC1          0.050552   0.009208   5.490 4.67e-08 ***
## PC2         -0.224746   0.011681 -19.240  < 2e-16 ***
## PC3          0.259318   0.013018  19.920  < 2e-16 ***
## PC4          0.032300   0.014717   2.195  0.02833 *
## PC5         -0.083681   0.016551  -5.056 4.78e-07 ***
## PC6         -0.025450   0.019960  -1.275  0.20248
```

```
## PC7          0.094912   0.021216   4.474 8.24e-06 ***
## PC8         -0.086300   0.024926  -3.462  0.00055 ***
## PC9         -0.142225   0.027613  -5.151 2.92e-07 ***
## PC10        -0.094222   0.038068  -2.475  0.01342 *
## PC11        -0.064332   0.066424  -0.969  0.33293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
pc.mode2<-lm(red_wine$quality~PC1+PC2+PC3+PC4+PC5+PC7+PC8+PC9+PC10,
             data=red.new.dat)
summary(pc.mode2)
```

```
##
## Call:
## lm(formula = red_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 +
##     PC7 + PC8 + PC9 + PC10, data = red.new.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69650 -0.35961 -0.05619  0.45087  1.99028
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.63602    0.01621 347.727  < 2e-16 ***
## PC1          0.05056    0.00921   5.489 4.69e-08 ***
## PC2         -0.22474    0.01168 -19.237  < 2e-16 ***
## PC3          0.25931    0.01302  19.916  < 2e-16 ***
## PC4          0.03232    0.01472   2.196 0.028236 *
## PC5         -0.08368    0.01655  -5.055 4.80e-07 ***
## PC7          0.09487    0.02122   4.471 8.34e-06 ***
## PC8         -0.08628    0.02493  -3.461 0.000553 ***
## PC9         -0.14215    0.02762  -5.147 2.98e-07 ***
## PC10        -0.09416    0.03807  -2.473 0.013499 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6481 on 1589 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3559
## F-statistic: 99.11 on 9 and 1589 DF,  p-value: < 2.2e-16
```

```
anova(pc.model,pc.mode2)
```

```
## Analysis of Variance Table
##
## Model 1: red_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 +
##     PC8 + PC9 + PC10 + PC11
## Model 2: red_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC7 + PC8 +
##     PC9 + PC10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1587 666.41
## 2   1589 667.49 -2   -1.0752 1.2802 0.2783
```

PCA for White wine

```
cov.white<-cov(white_wine[,1:11])
cov.white
```

```
##                       fixed acidity volatile acidity   citric acid
## fixed acidity         0.7121135857    -1.930571e-03  0.0295325116
## volatile acidity     -0.0019305706     1.015954e-02 -0.0018232776
## citric acid           0.0295325116    -1.823278e-03  0.0146457930
## residual sugar        0.3810218137     3.286533e-02  0.0578289265
## chlorides             0.0004256255     1.552775e-04  0.0003023838
## free sulfur dioxide  -0.7089186424    -1.663005e-01  0.1936297767
## total sulfur dioxide  3.2660133926     3.823539e-01  0.6229887081
## density               0.0006696773     8.173933e-06  0.0000541138
## pH                   -0.0542648260    -4.857531e-04 -0.0029923451
## sulphates            -0.0016509923    -4.109902e-04  0.0008608829
## alcohol              -0.1255328219     8.399723e-03 -0.0112782389
##                       residual sugar     chlorides free sulfur dioxide
## fixed acidity            0.381021814  4.256255e-04        -0.708918642
## volatile acidity         0.032865334  1.552775e-04        -0.166300459
## citric acid              0.057828926  3.023838e-04         0.193629777
## residual sugar          25.725770164  9.827502e-03        25.800577899
## chlorides                0.009827502  4.773337e-04         0.037674498
## free sulfur dioxide     25.800577899  3.767450e-02       289.242719999
## total sulfur dioxide    86.531302970  1.846875e-01       444.865890947
## density                  0.012727165  1.680754e-05         0.014965532
## pH                      -0.148683661 -2.983649e-04        -0.001586555
## sulphates               -0.015434743  4.179687e-05         0.114937934
## alcohol                 -2.812740332 -9.684235e-03        -5.234508674
##                       total sulfur dioxide       density            pH
## fixed acidity                   3.26601339  6.696773e-04 -5.426483e-02
## volatile acidity                0.38235390  8.173933e-06 -4.857531e-04
## citric acid                     0.62298871  5.411380e-05 -2.992345e-03
## residual sugar                 86.53130297  1.272717e-02 -1.486837e-01
## chlorides                       0.18468749  1.680754e-05 -2.983649e-04
## free sulfur dioxide           444.86589095  1.496553e-02 -1.586555e-03
## total sulfur dioxide         1806.08549085  6.735203e-02  1.489422e-02
## density                         0.06735203  8.945524e-06 -4.226861e-05
## pH                              0.01489422 -4.226861e-05  2.280118e-02
## sulphates                       0.65264458  2.542747e-05  2.687523e-03
## alcohol                       -23.47660460 -2.871430e-03  2.256505e-02
##                          sulphates       alcohol
## fixed acidity        -1.650992e-03  -0.125532822
## volatile acidity     -4.109902e-04   0.008399723
## citric acid           8.608829e-04  -0.011278239
## residual sugar       -1.543474e-02  -2.812740332
## chlorides             4.179687e-05  -0.009684235
## free sulfur dioxide   1.149379e-01  -5.234508674
## total sulfur dioxide  6.526446e-01 -23.476604603
## density               2.542747e-05  -0.002871430
## pH                    2.687523e-03   0.022565052
## sulphates             1.302471e-02  -0.002448356
## alcohol              -2.448356e-03   1.514426982
```

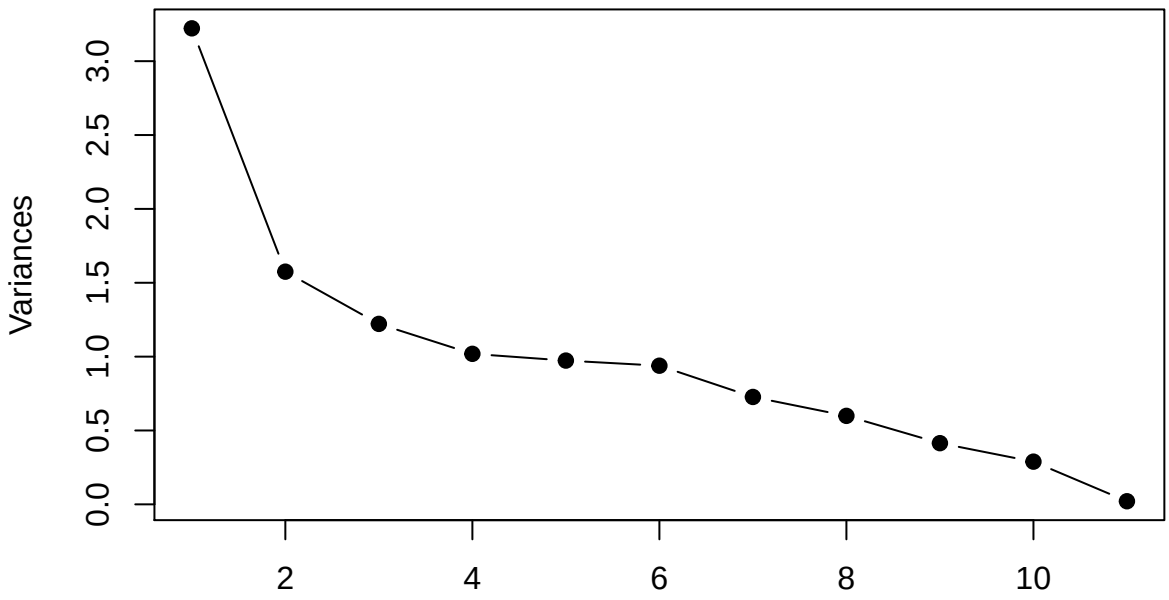
```r
cor.white<-round(cor(white_wine[,1:11]),3)
eig.val.white<-eigen(cor.white)$values # Pick the eigenvalues from result
round(eig.val.white,3)
```

```
##  [1] 3.222 1.575 1.222 1.019 0.973 0.939 0.727 0.599 0.414 0.289 0.021
```

```r
eig.vec.white<-eigen(cor.white)$vectors # Pick eigenvectors from result
rownames(eig.vec.white)<-colnames(white_wine[,1:11])
colnames(eig.vec.white)<-c("PC1","PC2","PC3","PC4","PC5","PC6",
                           "PC7","PC8","PC9","PC10","PC11")
round(eig.vec.white,2)
```

```
##                        PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
## fixed acidity         0.16  0.59 -0.12 -0.02  0.25 -0.11  0.20  0.59 -0.33
## volatile acidity      0.00 -0.05  0.59 -0.28  0.64  0.11 -0.27  0.03  0.15
## citric acid           0.14  0.35 -0.50 -0.15  0.05  0.13 -0.71 -0.15  0.20
## residual sugar        0.43 -0.01  0.21  0.27  0.01 -0.29 -0.21 -0.39 -0.41
## chlorides             0.21  0.01  0.10 -0.71 -0.32  0.40  0.08 -0.10 -0.39
## free sulfur dioxide   0.30 -0.29 -0.28  0.31  0.19  0.49  0.17 -0.08 -0.14
## total sulfur dioxide  0.41 -0.24 -0.13  0.06  0.30  0.27  0.07  0.25  0.15
## density               0.51 -0.01  0.13  0.02 -0.09 -0.33 -0.11  0.07 -0.09
## pH                   -0.13 -0.58 -0.13 -0.10 -0.13 -0.19 -0.43  0.53 -0.26
## sulphates             0.04 -0.22 -0.43 -0.44  0.39 -0.49  0.31 -0.27  0.01
## alcohol              -0.44  0.04 -0.11  0.14  0.34  0.13 -0.13 -0.20 -0.62
##                        PC10  PC11
## fixed acidity          0.13 -0.17
## volatile acidity       0.22 -0.02
## citric acid            0.04 -0.01
## residual sugar        -0.09 -0.49
## chlorides             -0.05 -0.02
## free sulfur dioxide    0.57  0.03
## total sulfur dioxide -0.71 -0.04
## density                0.07  0.76
## pH                     0.11 -0.14
## sulphates              0.06 -0.04
## alcohol               -0.27  0.36
```

```r
plot(eig.val.white, type = "b", pch=19, xlab = "", ylab = "Variances")
```




```
# Proportion of variation explained by each PCs
round(eig.val.white/sum(eig.val.white),3)
```

```
##  [1] 0.293 0.143 0.111 0.093 0.088 0.085 0.066 0.054 0.038 0.026 0.002
```

```
cumsum(round(eig.val.white/sum(eig.val.white),3))
```

```
##  [1] 0.293 0.436 0.547 0.640 0.728 0.813 0.879 0.933 0.971 0.997 0.999
```

```
x<-white_wine[,1:11]
x.mean.w<-apply(x, 2, mean)
x.mean.w
```

```
##        fixed acidity     volatile acidity          citric acid
##           6.85478767           0.27824112           0.33419151
##       residual sugar            chlorides  free sulfur dioxide
##           6.39141486           0.04577236          35.30808493
## total sulfur dioxide              density                   pH
##         138.36065741           0.99402738           3.18826664
##            sulphates              alcohol
##           0.48984688          10.51426705
```

```
# Standardize white wine data
z<-x
for (i in 1:11) {
  z[,i]<-(x[,i]-x.mean.w[i])/sqrt(diag(cov.white)[i])
}

white.new<-x
```

```r
for (i in 1:11){ # number of pcs
  for (j in 1:4898){
    white.new[j,i]<-t(eig.vec.white[,i]) %*% t(z[j,])
  }
}

# To check the correlation in the new dataset
colnames(white.new) = c("PC1", "PC2", "PC3", "PC4","PC5",
                        "PC6","PC7","PC8","PC9","PC10","PC11")
round(cor(white.new),2)
```

```
##      PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11
## PC1    1   0   0   0   0   0   0   0   0    0    0
## PC2    0   1   0   0   0   0   0   0   0    0    0
## PC3    0   0   1   0   0   0   0   0   0    0    0
## PC4    0   0   0   1   0   0   0   0   0    0    0
## PC5    0   0   0   0   1   0   0   0   0    0    0
## PC6    0   0   0   0   0   1   0   0   0    0    0
## PC7    0   0   0   0   0   0   1   0   0    0    0
## PC8    0   0   0   0   0   0   0   1   0    0    0
## PC9    0   0   0   0   0   0   0   0   1    0    0
## PC10   0   0   0   0   0   0   0   0   0    1    0
## PC11   0   0   0   0   0   0   0   0   0    0    1
```

```r
# Principle Component Regression
white.new.dat<-cbind(white_wine$quality,white.new)

pc.mode3<-lm(white_wine$quality~PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10+PC11,
             data=white.new.dat)
summary(pc.mode3)
```

```
##
## Call:
## lm(formula = white_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 +
##     PC6 + PC7 + PC8 + PC9 + PC10 + PC11, data = white.new.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.877909   0.010736 547.502  < 2e-16 ***
## PC1         -0.146500   0.005981 -24.493  < 2e-16 ***
## PC2         -0.041014   0.008555  -4.794 1.68e-06 ***
## PC3         -0.174933   0.009714 -18.008  < 2e-16 ***
## PC4          0.166690   0.010639  15.668  < 2e-16 ***
## PC5          0.042560   0.010883   3.911 9.33e-05 ***
## PC6          0.002490   0.011082   0.225   0.8222
## PC7         -0.022020   0.012596  -1.748   0.0805 .
## PC8         -0.183116   0.013869 -13.203  < 2e-16 ***
## PC9         -0.358122   0.016684 -21.465  < 2e-16 ***
## PC10        -0.108122   0.019956  -5.418 6.31e-08 ***
## PC11        -0.480311   0.074719  -6.428 1.41e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

```
pc.mode4<-lm(white_wine$quality~PC1+PC2+PC3+PC4+PC5+PC8+PC9+PC10+PC11,
             data=white.new.dat)
summary(pc.mode4)
```

```
##
## Call:
## lm(formula = white_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 +
##     PC8 + PC9 + PC10 + PC11, data = white.new.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8835 -0.4900 -0.0336  0.4647  3.1256
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.877909   0.010737 547.440  < 2e-16 ***
## PC1         -0.146501   0.005982 -24.490  < 2e-16 ***
## PC2         -0.041017   0.008556  -4.794 1.68e-06 ***
## PC3         -0.174944   0.009715 -18.007  < 2e-16 ***
## PC4          0.166695   0.010640  15.667  < 2e-16 ***
## PC5          0.042568   0.010884   3.911 9.32e-05 ***
## PC8         -0.183107   0.013870 -13.201  < 2e-16 ***
## PC9         -0.358149   0.016686 -21.464  < 2e-16 ***
## PC10        -0.108157   0.019958  -5.419 6.28e-08 ***
## PC11        -0.480669   0.074727  -6.432 1.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4888 degrees of freedom
## Multiple R-squared:  0.2814, Adjusted R-squared:  0.2801
## F-statistic: 212.7 on 9 and 4888 DF,  p-value: < 2.2e-16
```

```
anova(pc.mode3,pc.mode4)
```

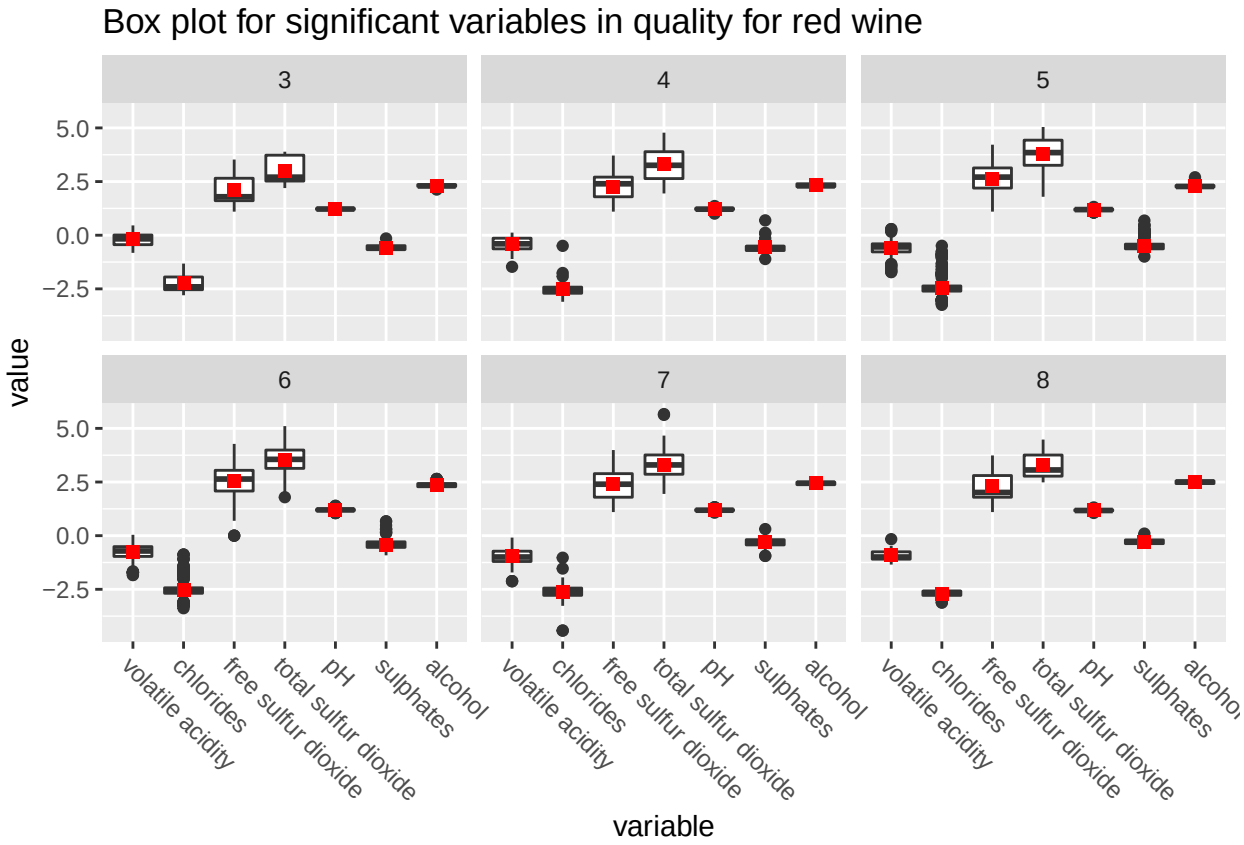
```
## Analysis of Variance Table
##
## Model 1: white_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 +
##     PC8 + PC9 + PC10 + PC11
## Model 2: white_wine$quality ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC8 + PC9 +
##     PC10 + PC11
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4886 2758.3
## 2   4888 2760.1 -2   -1.7538 1.5533 0.2116
```

Plot for each variables in different quality

```
# red_wine$`volatile acidity`<-log(red_wine$`volatile acidity`)
# red_wine$`residual sugar`<-log(red_wine$`residual sugar`)
# red_wine$chlorides<-log(red_wine$chlorides)
# red_wine$`free sulfur dioxide`<-log(red_wine$`free sulfur dioxide`)
# red_wine$`total sulfur dioxide`<-log(red_wine$`total sulfur dioxide`)
# red_wine$sulphates<-log(red_wine$sulphates)
# red_wine$alcohol<-log(red_wine$alcohol)
# red_wine_m<- melt(red_wine,id.vars='quality',
#                   measure.vars=c("fixed acidity","volatile acidity",
#                                  "citric acid","residual sugar","chlorides",
#                                  "free sulfur dioxide","total sulfur dioxide",
#                                  "density", "pH", "sulphates","alcohol"))

red_wine_m<- melt(red_wine,id.vars='quality',
                  measure.vars=c("volatile acidity","chlorides",
                                 "free sulfur dioxide","total sulfur dioxide",
                                 "pH", "sulphates","alcohol"))
red_wine_m[,3]<-log(red_wine_m[,3])


ggplot(red_wine_m,aes(x=variable, y=value)) +
  geom_boxplot()+
  facet_wrap(~quality)+
  stat_summary(fun.y = "mean", geom = "point",
               size = 2, color = "red", shape = 15)+
  ggtitle("Box plot for significant variables in quality for red wine")+
  theme(axis.text.x=element_text(angle = -45, hjust = 0))
```


Box plot for significant variables in quality for red wine
3
4
5
6
7
8
value
5.0
2.5
0.0
−2.5
volatile acidity
chlorides
free sulfur dioxide
total sulfur dioxide
pH
sulphates
alcohol
variable