评分：_____

SHANGHAI UNIVERSITY

COURSE PAPER

Homework 2

Information	**student1**	**student2**
Academy	Management	Management
Specialty	Management Science	Information Science
Student ID	18121216	19121255
Name	Lulu Shi	Jingwen Zhang
Course	Machine Learning in Business	
Teacher	Si Zhang | Liyang Xiao	

Listed are the **Answers** to the **Questions.**

PART 1_data_Stock_Data.csv

Using the data set , Stock Data.csv, as in Homework 1, carry out a linear regression analysis on the data with following instructions:

1) Define a variable calling dat_AC which only contains the adjusted closing prices of every stock and S&P 500.

```
>setwd("D:/ ML_in_Business/homework 1")
>dat=read.csv('Stock_Data.csv')
>Indicator=seq(3,dim(dat)[2],2)
```

Reminder:

After importing the data, define a vector Indicator=seq(3,dim(dat)[2],2)to indicate the number of columns you want to pick out from the data set.

```
>dat_AC=(dat[,Indicator])
```

(the output of dat_AC is in **H2_dat_AC.csv**)

2) Plugging dat AC into the procedure we used to calculate daily returns of Ford, calculate daily returns of all the stocks and S&P 500.

```
>dat$Date=as.Date(dat$Date,format='%d/%m/%Y')
>result=data.frame(index=(1:4962))
>name=names(dat_AC)
>for (o in name){
  ###
  start1= which(dat$Date=='1987-01-02')
  end1= which(dat$Date=='2006-08-31')
  prePrice=(dat[o][,1])[start1:end1]
  ###
  start2= which(dat$Date=='1987-01-05')
  end2= which(dat$Date=='2006-09-01')
  curPrice=(dat[o][,1])[start2:end2]
  ###
  Return=NULL
  for(i in 1:4962){
    Return=c(Return,curPrice[i]/prePrice[i]-1)
  }
  AC_rtn=paste(o,'_Return',sep='')
  result[AC_rtn]>-(Return)
}
```

(the output of returns is in **result.csv**)

3) Conduct a linear regression analysis on the returns with returns of S&P 500 being the response variable and all the other returns being predict variables.

```
>lm.fit=lm(S.P_AC_Return~.,data=result)
>sink('H2_returnsRegression.csv')
>summary(lm.fit)
>sink(NULL)
```

(the output of returns_Regression is in **H2_returnsRegression.csv**)

4) What can you conclude from the regression result, are these predictors significant and can you explain why?

Coefficients:	
	Estimate Std. Error t value Pr(>|t|)
(Intercept)	-2.005e-04 6.408e-05 -3.129 0.00177 **
GM_AC_Return	5.533e-02 4.087e-03 13.538 < 2e-16 ***
F_AC_Return	3.778e-02 4.007e-03 9.429 < 2e-16 ***
UTX_AC_Return	7.825e-02 4.337e-03 18.042 < 2e-16 ***
CAT_AC_Return	5.611e-02 3.810e-03 14.727 < 2e-16 ***
MRK_AC_Return	6.705e-02 4.502e-03 14.891 < 2e-16 ***
PFE_AC_Return	6.916e-02 4.272e-03 16.190 < 2e-16 ***
IBM_AC_Return	1.044e-01 3.890e-03 26.834 < 2e-16 ***
MSFT_AC_Return	8.998e-02 3.059e-03 29.412 < 2e-16 ***
C_AC_Return	1.121e-01 3.519e-03 31.847 < 2e-16 ***
XOM_AC_Return	1.415e-01 4.819e-03 29.357 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*'	
Residual standard error: 0.004502 on 4951 degrees of freedom	
Multiple R-squared: 0.8245	Adjusted R-squared: 0.8241
F-statistic: 2325 on 10 and 4951 DF	p-value: < 2.2e-16

TABLE 2.1 *For the result data, least squares coefficient estimates of the multiple linear regression of* ***Returns*** *of S&P_AP on those predictors.*

From the regression result, consider the predictors. The ***RSE = 0.004502*** , which is quite small , indicating that the model fits the data well. The ***Adjusted R-squared= 0.8241***, which is close to 1, indicating that a large proportion of the variability in the response has been explained by the regression, and about 82.41% variability in $S.P_{AC_{Return}}$ is explained by those predictors. The ***p-values*** are smaller than the default value ***5% or 1%***, then we can infer that there is an association between the predictors and the response. We reject the null hypothesis—that is, we declare a relationship to exist between X and Y.

In general, suppose that we have p distinct predictors. .Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p + \varepsilon \qquad (2.1)$$

From **TABLE 2.1**, we get the multiple linear regression model

$$\begin{aligned} \mathrm{S.P}_{\mathrm{AC_{Return}}} = & -2.005e^{-04} + 5.533e^{-02} \times GM_{\mathrm{AC_{Return}}} + 3.778e^{-02} \times F_{\mathrm{AC_{Return}}} \\ & + 7.825e^{-02} \times UTX_{\mathrm{AC_{Return}}} + 5.611e^{-02} \times CAT_{\mathrm{AC_{Return}}} \\ & + 6.705e^{-02} \times MRK_{\mathrm{AC_{Return}}} + 6.916e^{-02} \times PFE_{\mathrm{AC_{Return}}} \\ & + 1.044e^{-01} \times IBM_{\mathrm{AC_{Return}}} + 8.998e^{-02} \times MSFT_{\mathrm{AC_{Return}}} \\ & + 1.121e^{-01} \times C_{\mathrm{AC_{Return}}} + 1.415e^{-01} \times XOM_{\mathrm{AC_{Return}}} + \varepsilon \end{aligned} \qquad (2.2)$$

The result function is (2.2). Also, we drew **FIGURE 2.2** to visualize the regression result of the response on the predictors. Apparently, the model fits the data well.
(source code : **ML_h2_StockPlot.R**)

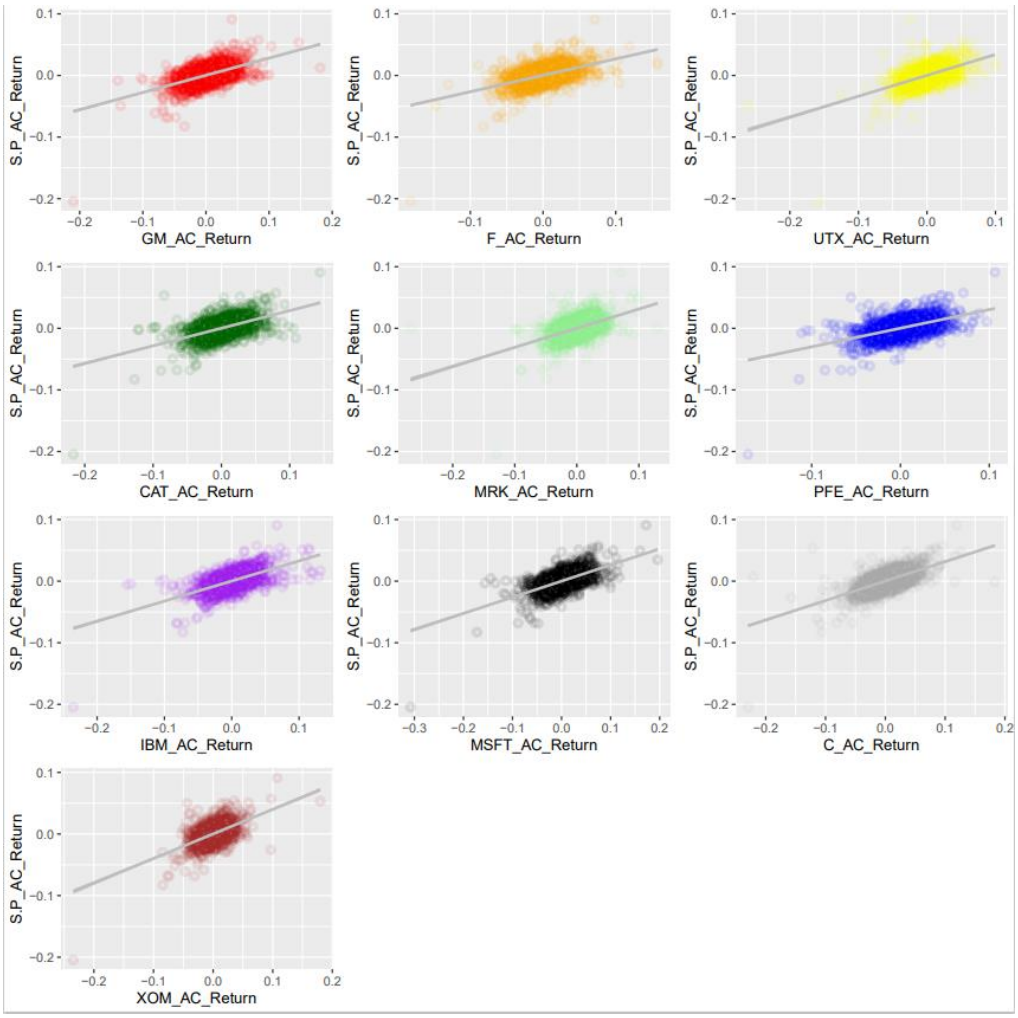

FIGURE 2.2

Code Parts (PART 1_data_Stock_Data.csv) (source code : ML_h2_stock.R)

```
#define a variable named dat_AC only contains AC
>setwd("D:/ML_in_Business/homework 1")
>dat=read.csv('Stock_Data.csv')
>Indicator=seq(3,dim(dat)[2],2)
>dat_AC=(dat[,Indicator])
# output the dat_AC to a csv file.
>write.csv(dat_AC,"H2_dat_AC.csv",row.names = FALSE)
# to format the Date column.
>dat$Date=as.Date(dat$Date,format='%d/%m/%Y')
>result=data.frame(index=(1:4962))
>name=names(dat_AC)
# Create a loop to repeatedly calculate the returns of each predictor.
# Create a dataframe then a csv file named result to hold the data of returns.
>for (o in name){
  ###
  start1= which(dat$Date=='1987-01-02')
  end1= which(dat$Date=='2006-08-31')
  prePrice=(dat[o][,1])[start1:end1]
  ###
  start2= which(dat$Date=='1987-01-05')
  end2= which(dat$Date=='2006-09-01')
  curPrice=(dat[o][,1])[start2:end2]
  ###
  Return=NULL
  for(i in 1:4962){
    Return=c(Return,curPrice[i]/prePrice[i]-1)
  }
  AC_rtn=paste(o,'_Return',sep='')
  result[AC_rtn]>-(Return)
}
>result=result[,-1]
# Analyze the returns of S&P on the predictors.
>summary(result)
>lm.fit=lm(S.P_AC_Return~.,data=result)
>sink('H2_returnsRegression.csv')
>summary(lm.fit)
>sink(NULL) # export the summary(lm.fit) to a csv file.
```

PART 2_data_wine.csv

In the data set wine.csv, it contains the results of chemical analysis of different wines. Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). The rank of each individual wine is given by a group of tasters:

1) Split the data set into two based on the color of the wines and name the two data sets wine red and wine white respectively.

```
>setwd("C:/Users/啦啦露/Desktop/homework 2")
>dat2=read.csv('wine.csv')
>wine_red=dat2[dat2$color=="red",]#wine_red data set
>wine_white=dat2[dat2$color=="white",]#wine_white data set
```

2) Conduct linear regressions for these two data sets separately with quality being the response variable and any other variables(except color) you think might be relevant as predict variables.

```
>corTRed=cor(wine_red)# calculate regression data
>corTWhite=cor(wine_white)
>install.packages("corrplot")
>library(corrplot)
>corrplot(corTRed,method='color',type='upper',order='hclust',addCoef.col='b
lack')
# draw the matrix diagram of correlation coefficient
>corrplot(corTWhite,method='color',type='upper',order='hclust',addCoef.col=
'black')
# calculate LR data
>lm.fit1=lm(quality~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+c
hlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcoh
ol,data=wine_red)
>summary(lm.fit1)
# use stepwise regression to eliminate variables
>lm.fit11>-step(lm.fit1)
summary(lm.fit11)
#get wine red regression
>lm.fit111=lm(quality~volatile.acidity+sulphates*alcohol,data=wine_red)
>summary(lm.fit111)

>lm.fit2=lm(quality~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+c
hlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcoh
ol,data=wine_white)
>summary(lm.fit2)
```

```
>lm.fit21>-step(lm.fit2)# use stepwise regression to eliminate variables
>summary(lm.fit21)
#get wine white regression
>lm.fit22=lm(quality ~ alcohol*volatile.acidity+density,data=wine_white)
>Summary(lm.fit22)
```

2) From your regression results, what factors affect the quality of red wines and white wines respectively?

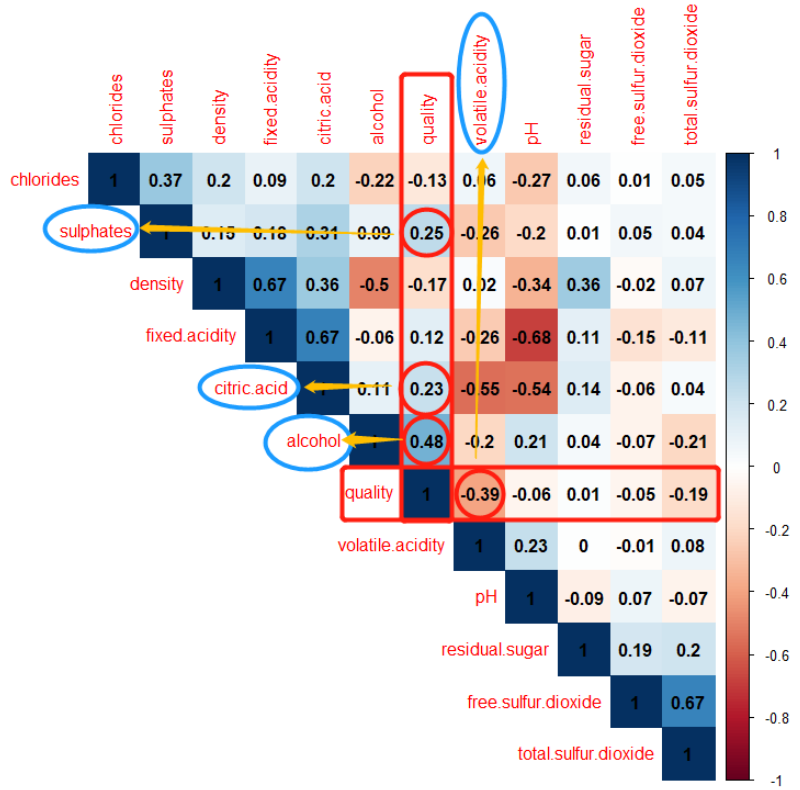

FIGURE 2.3

```
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
```

FIGURE 2.4

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.43406    0.70988   9.064  < 2e-16 ***
volatile.acidity  -1.17957    0.09639 -12.238  < 2e-16 ***
sulphates         -5.03825    1.02609  -4.910 1.00e-06 ***
alcohol           -0.07294    0.07003  -1.041    0.298
sulphates:alcohol  0.56723    0.10132   5.599 2.54e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6525 on 1594 degrees of freedom
Multiple R-squared:  0.3487,    Adjusted R-squared:  0.3471
F-statistic: 213.4 on 4 and 1594 DF,  p-value: < 2.2e-16
```

FIGURE 2.5

From the wine red regression result, we choose alcohol , volatile.acidity, sulphates, and sulphates*alcohol as our variables. From **FIGURE 2.5**, we get our regression model(2.3):

$$Quality = 6.43406 - 1.17957 \times volatile.acidity - 5.03825 \times sulphates - 0.07294 \times alcohol + 0.56723 \times sulphates \times alcohol \qquad (2.3)$$

First, we use stepwise regression model to obtain 7 variables. After comparing their p-value and correlation with quality , we choose 3 variables: alcohol , volatile.acidity, sulphates. By comparing the direct correlation of the three variables, we choose to analyze sulphates*alcohol in our model. We find that this has the highest Adjusted R-squared= 0.3471.

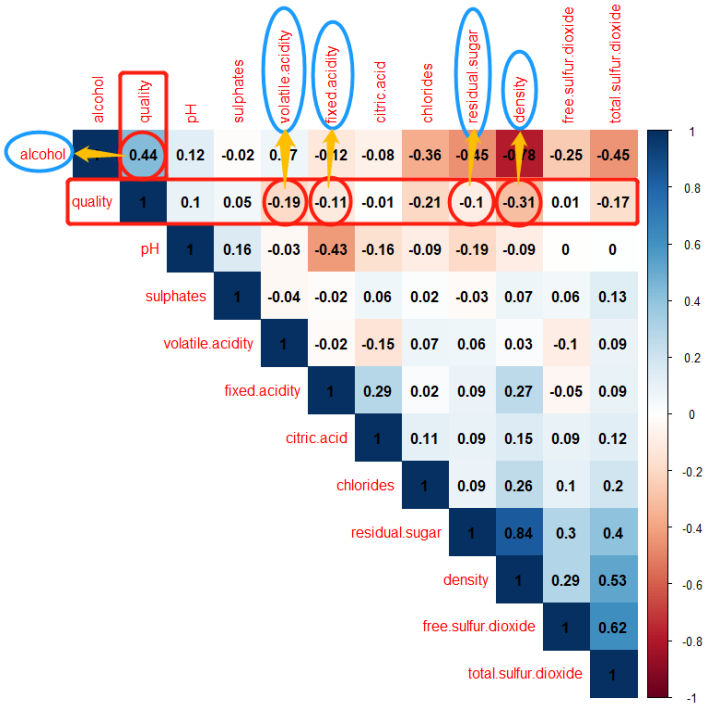

FIGURE 2.6

```
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
```

FIGURE 2.7

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -27.19249    6.08006  -4.472 7.91e-06 ***
alcohol                    0.18568    0.03068   6.052 1.54e-09 ***
volatile.acidity          -9.02433    0.89138 -10.124  < 2e-16 ***
density                   31.88575    5.95285   5.356 8.88e-08 ***
alcohol:volatile.acidity   0.65911    0.08387   7.859 4.73e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7641 on 4893 degrees of freedom
Multiple R-squared:  0.2563,    Adjusted R-squared:  0.2557
F-statistic: 421.6 on 4 and 4893 DF,  p-value: < 2.2e-16
```

FIGURE 2.8

From the wine white regression result, we choose alcohol, volatile.acidity, density, and volatile.acidity *alcohol as our variables. From **FIGURE 2.8**, we get our regression model(2.4):

$$Quality = -27.19249 + 0.18568 \times alcohol - 9.02433 \times volatile.acidity + 31.88575 \times density + 0.65911 \times alcohol \times volatile.acidity \quad (2.4)$$

First, we use stepwise regression model to obtain 8 variables. After comparing their p-value and correlation with quality, we choose 3 variables: alcohol, volatile.acidity, density. By comparing the direct correlation of the three variables, we choose to analyze volatile.acidity *alcohol in our model. We find that this has the highest Adjusted R-squared= 0.2557.

Also, we drew **FIGURE 2.9 and 2.10** to visualize the regression result of the response ***quality*** on the predictors. Apparently, the distribution of each predictor concerns to ***quality*** is distinguishing. (source code : **ML_h2_WinePlot.R**)

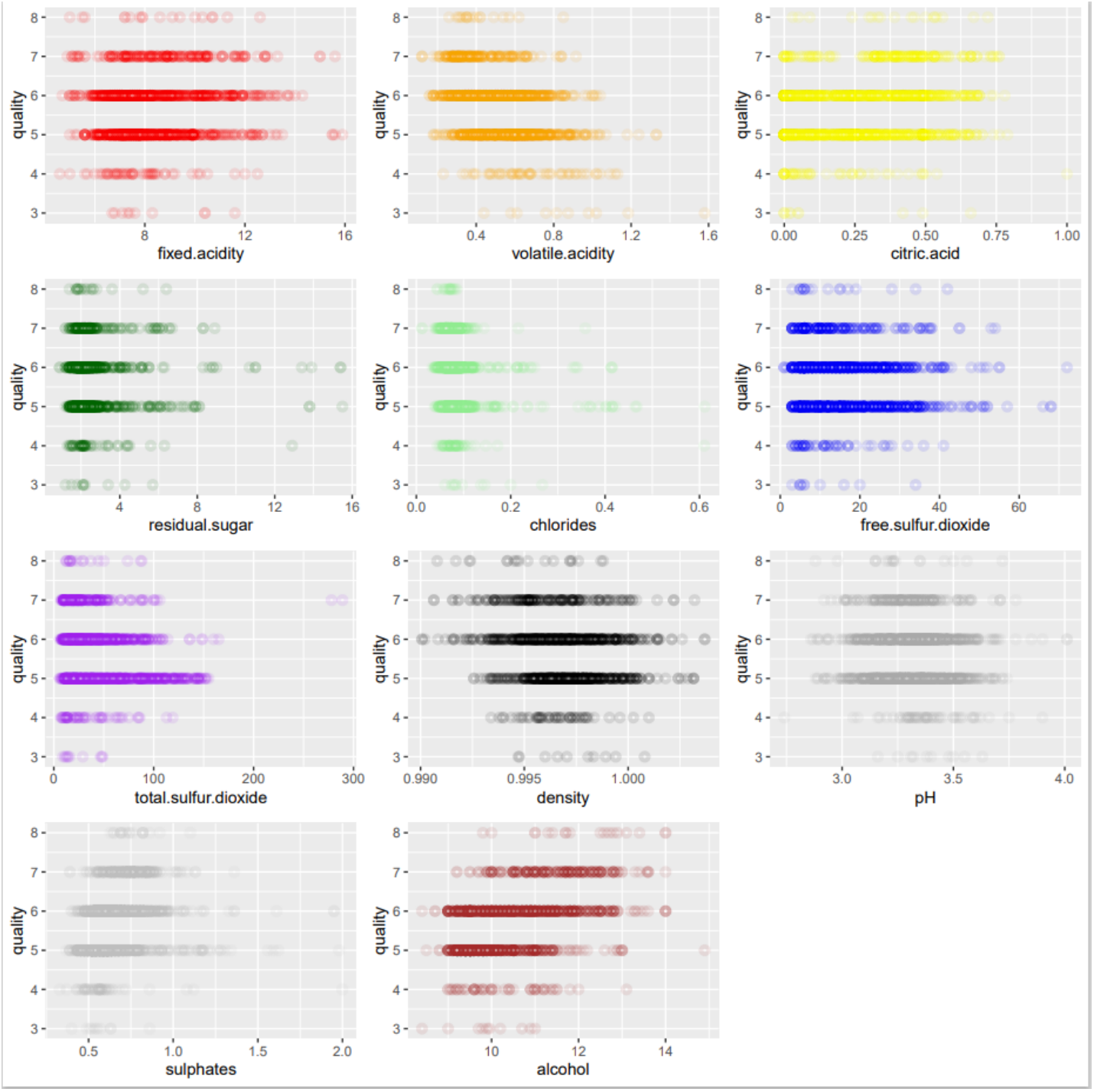

FIGURE 2.9 red_wine

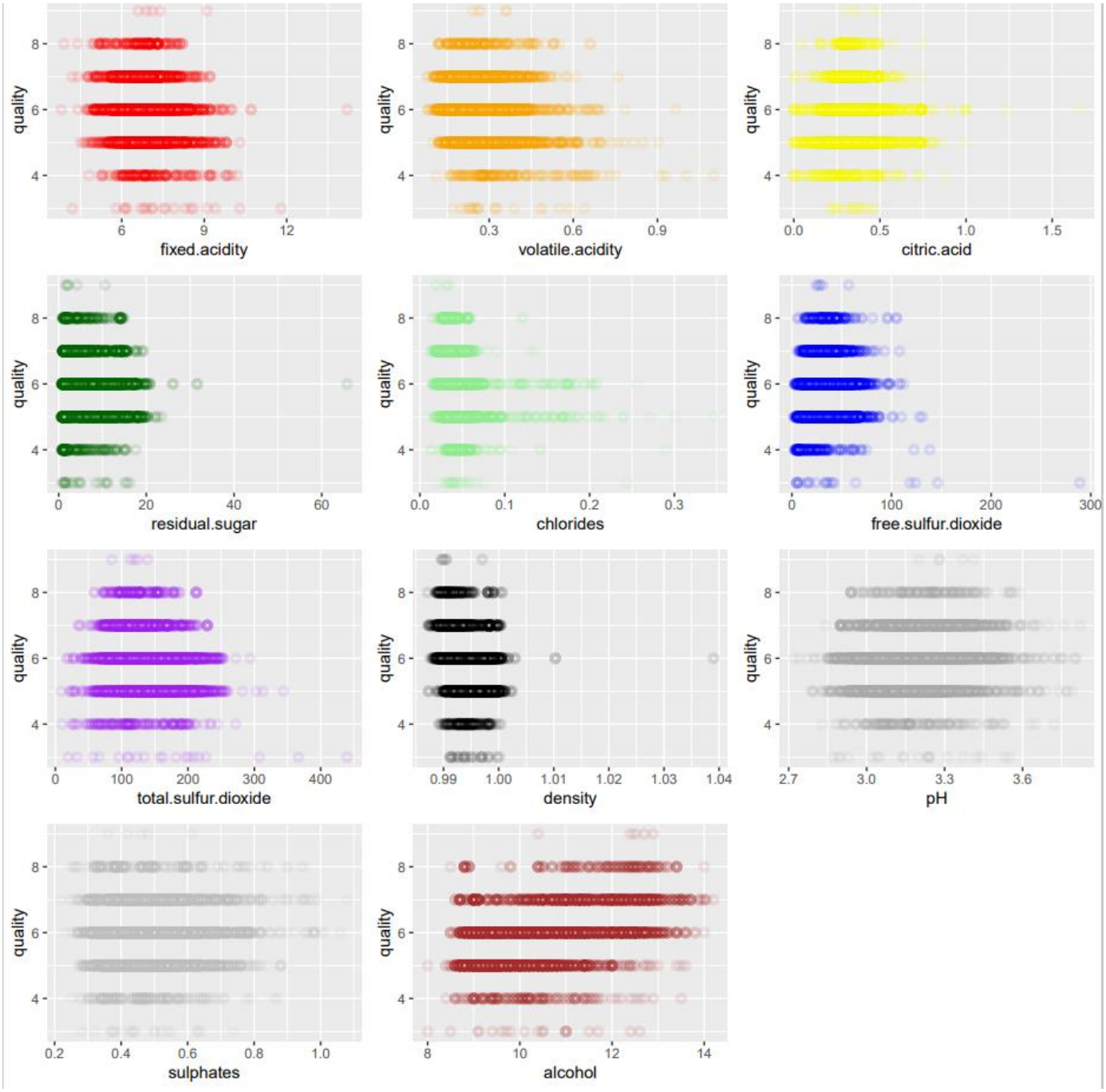

FIGURE 2.10 white_wine

Code Parts (PART 2_data_wine.csv) (source code : ML_h2_wine.R)

```
#2.1
>setwd("C:/Users/啦啦露/Desktop/homework 2")
>dat2=read.csv('wine.csv')
>wine_red=dat2[dat2$color=="red",]
>wine_white=dat2[dat2$color=="white",]
#2.2
#calculate
>corTRed=cor(wine_red)
>corTWhite=cor(wine_white)
#corrplot
#install.packages("corrplot")
>library(corrplot)
# draw figures
>corrplot(corTRed,method='color',type='upper',order='hclust',addCoef.col='black'
)
>corrplot(corTWhite,method='color',type='upper',order='hclust',addCoef.col='bla
ck')
#2.3
#wine_red
>lm.fit1=lm(quality~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlori
des+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol,data=
wine_red)
>summary(lm.fit1)
>lm.fit11<-step(lm.fit1)
>summary(lm.fit11)
>lm.fit111=lm(quality~volatile.acidity+sulphates*alcohol,data=wine_red)
>summary(lm.fit111)

#wine_white
>lm.fit2=lm(quality~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlori
des+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol,data=
wine_white)
>summary(lm.fit2)
>lm.fit21<-step(lm.fit2)
>summary(lm.fit21)
>lm.fit22=lm(quality ~ alcohol*volatile.acidity +density,data=wine_white)
>summary(lm.fit22)
```