# A2_Q1

Chia-Ying Chao

17/02/2023

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.4      v dplyr   1.0.10
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.0.2      v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

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

```
# Be sure the data is in your current working directory
wine_data<-read_csv("red_wine_data.csv", col_types = cols())
glimpse(wine_data)
```

```
## Rows: 1,599
## Columns: 12
## $ `fixed acidity`        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7~
## $ `volatile acidity`     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600~
## $ `citric acid`          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00,~
## $ `residual sugar`       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.~
## $ chlorides              <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069~
## $ `free sulfur dioxide`  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, ~
## $ `total sulfur dioxide` <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 10~
## $ density                <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978,~
## $ pH                     <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39,~
## $ sulphates              <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47,~
## $ alcohol                <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 1~
## $ quality                <dbl> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5,~
```

```
wine_data_chem <- wine_data %>% select(-quality)
head(wine_data_chem)
```

```
## # A tibble: 6 x 11
##   `fixed acidity` `volatile acidity` `citric acid` `residual sugar` chlorides
##             <dbl>              <dbl>         <dbl>            <dbl>     <dbl>
## 1             7.4               0.7           0                 1.9     0.076
## 2             7.8               0.88          0                 2.6     0.098
## 3             7.8               0.76          0.04              2.3     0.092
## 4            11.2               0.28          0.56              1.9     0.075
## 5             7.4               0.7           0                 1.9     0.076
## 6             7.4               0.66          0                 1.8     0.075
## # ... with 6 more variables: free sulfur dioxide <dbl>,
## #   total sulfur dioxide <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
## #   alcohol <dbl>
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
library(GGally)
```

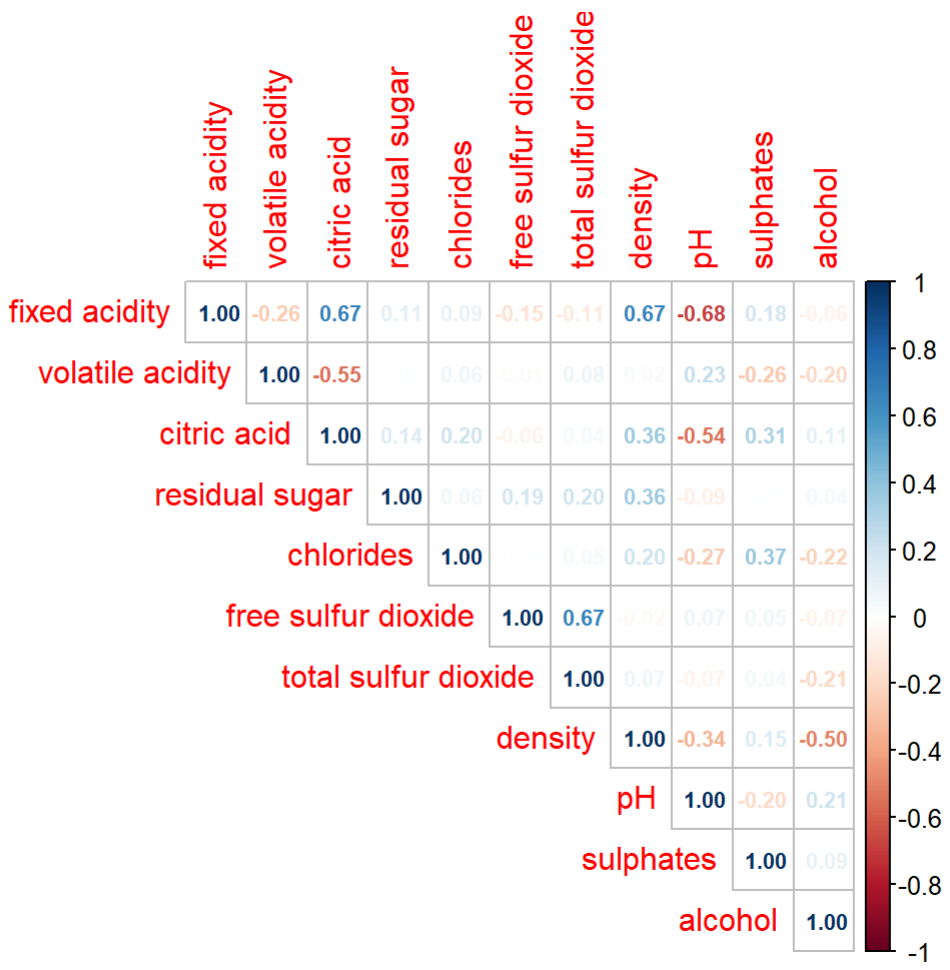
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(corrplot)
```

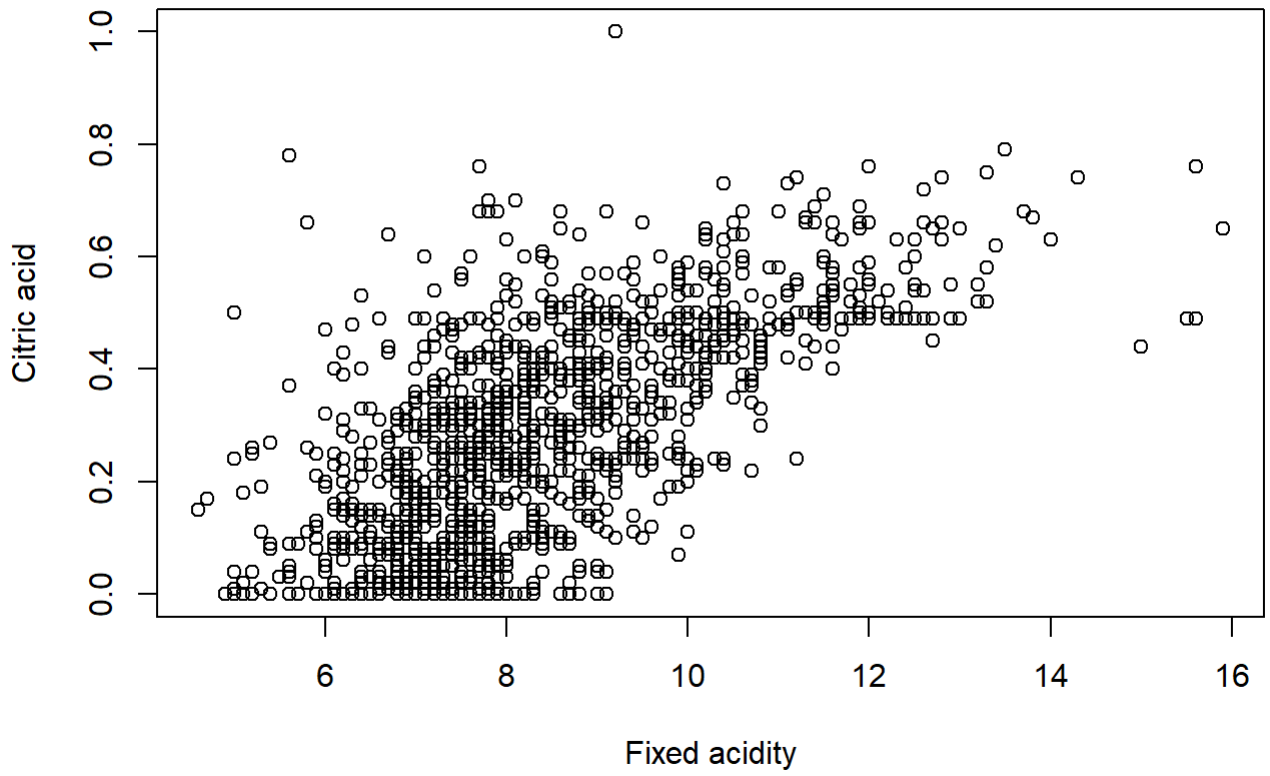
```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

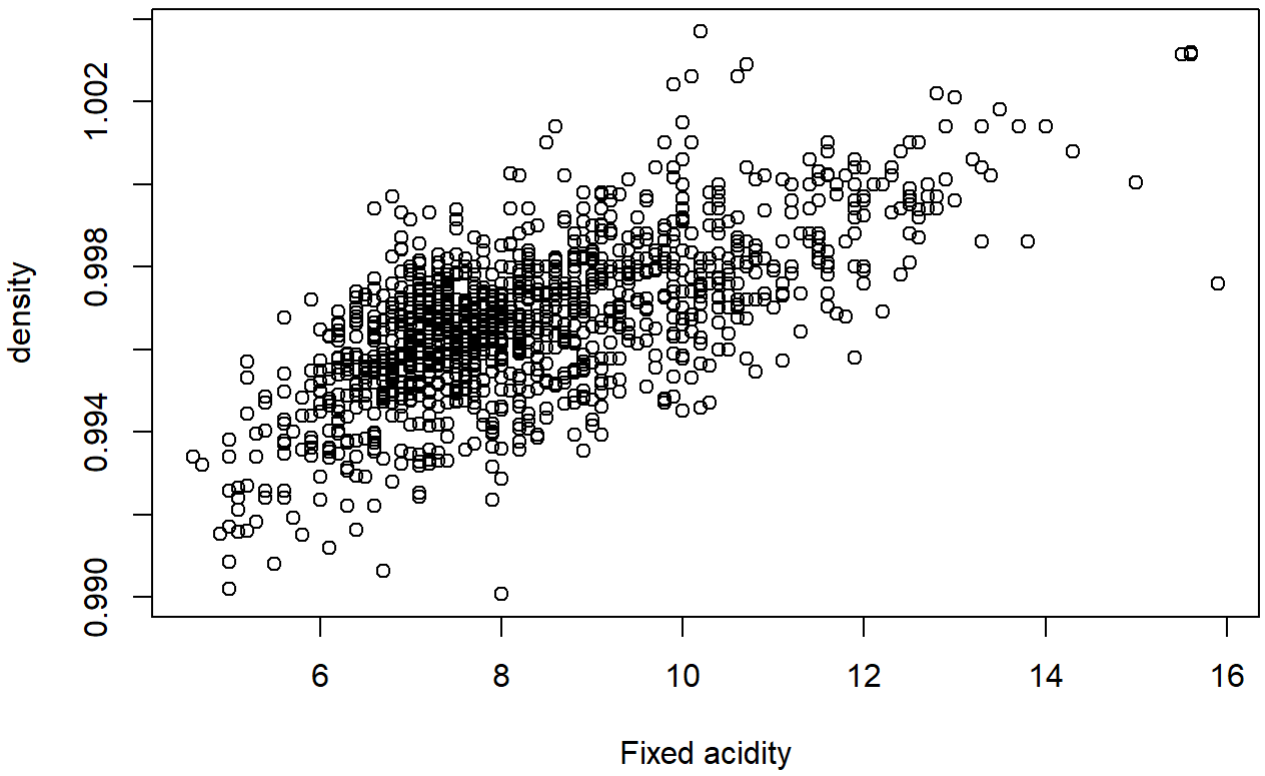
```
#Sample correlation matrice
corrplot(cor(wine_data_chem), type="upper", method="number",number.cex = 0.7)
```



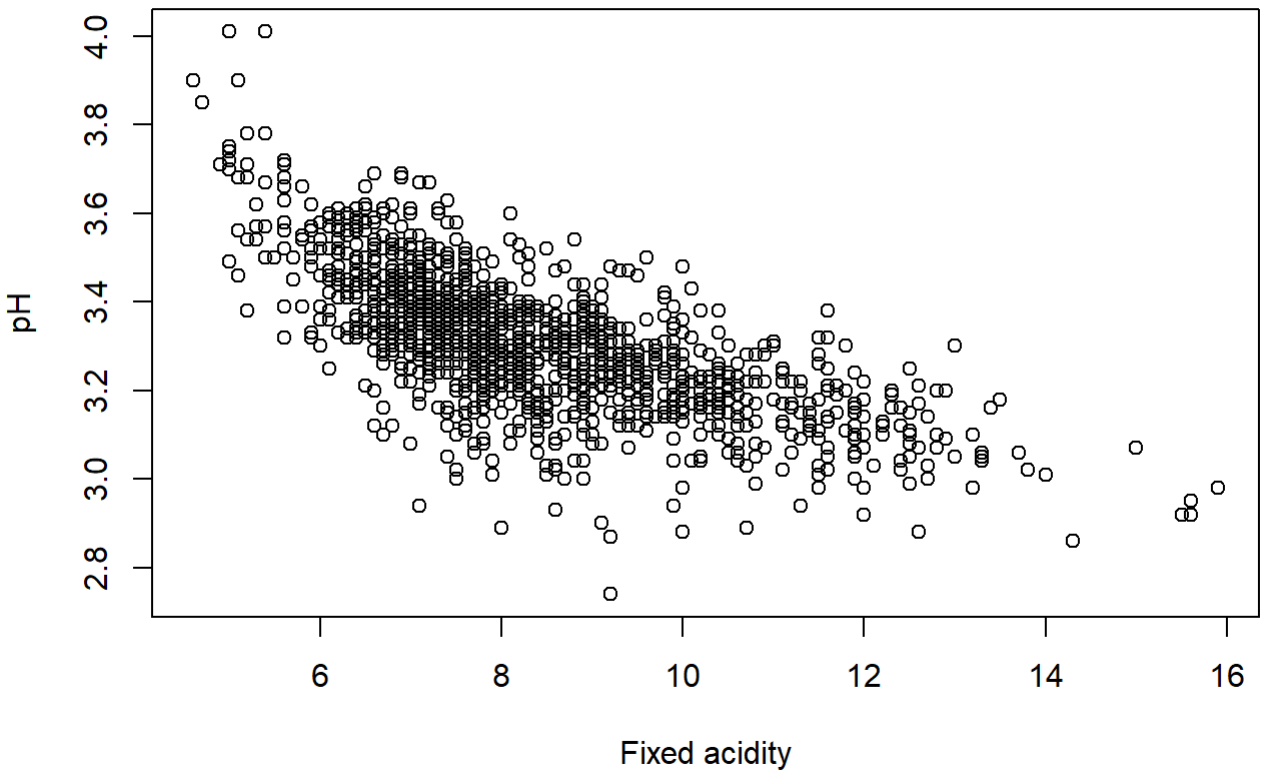

```
#scatterplots for those that have stronger correlation with fixed acidity
plot(wine_data_chem$'fixed acidity', wine_data_chem$'citric acid',
   xlab="Fixed acidity", ylab="Citric acid")
```



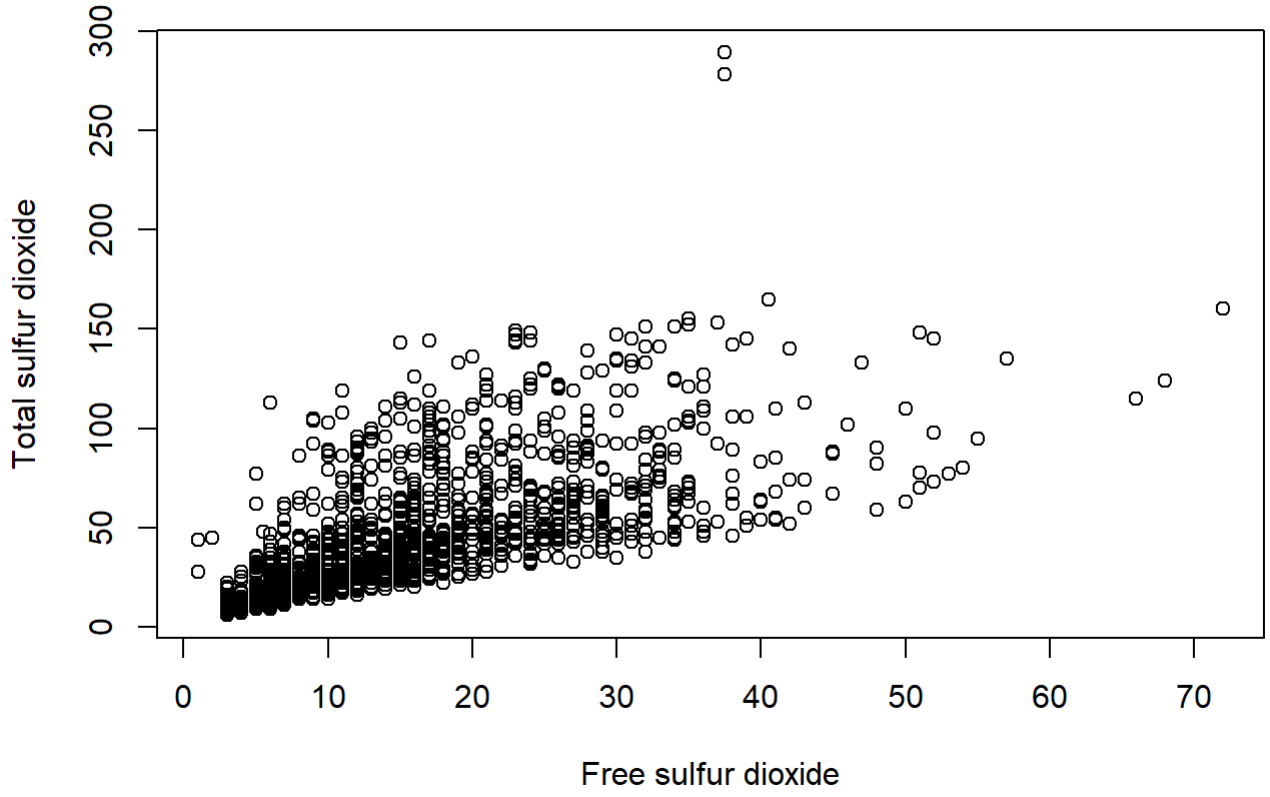

```
plot(wine_data_chem$'fixed acidity', wine_data_chem$'density',
   xlab="Fixed acidity", ylab="density")
```




```
plot(wine_data_chem$'fixed acidity', wine_data_chem$'pH',
   xlab="Fixed acidity", ylab="pH")
```




```
plot(wine_data_chem$'free sulfur dioxide', wine_data_chem$'total sulfur dioxide',
   xlab="Free sulfur dioxide", ylab="Total sulfur dioxide")
```




The sample correlation matrices shows the correlation between variables, we can observe that there are four correlation values exceed 0.6 indicating that they have a strong correlation between the variables. The correlation coefficients between citric acid and density and fixed acidity are 0.67. Total sulfur dioxide versus free sulfur dioxide also has a correlation coefficient of 0.67 and pH versus fixed acidity is -0.68. The correlation relationship can also be seen in the scatterplots, i.e. positive or negative.

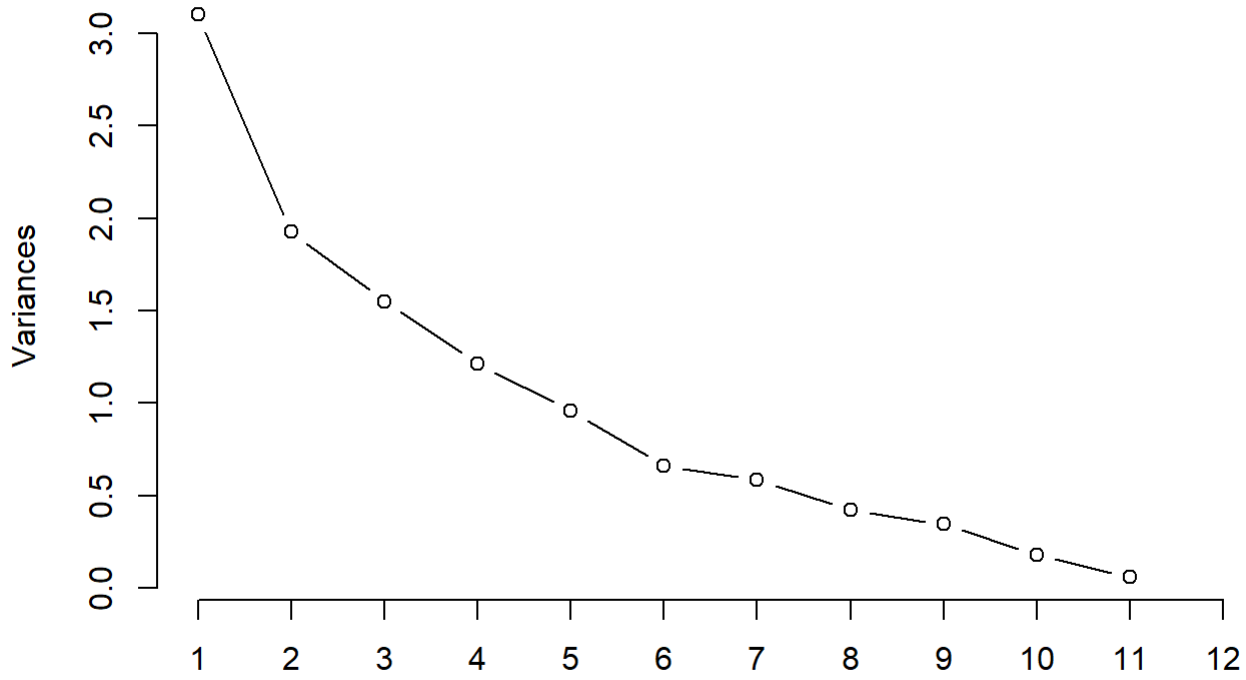
```
###ii)
PCA <- prcomp(wine_data_chem, center=TRUE,scale = TRUE )
PCA
```

```
## Standard deviations (1, .., p=11):
##  [1] 1.7604353 1.3877715 1.2452082 1.1014684 0.9794346 0.8121627 0.7640623
##  [8] 0.6503512 0.5870623 0.4258323 0.2440457
##
## Rotation (n x k) = (11 x 11):
##                              PC1          PC2         PC3          PC4
## fixed acidity         0.48931422 -0.110502738  0.12330157 -0.229617370
## volatile acidity     -0.23858436  0.274930480  0.44996253  0.078959783
## citric acid           0.46363166 -0.151791356 -0.23824707 -0.079418256
## residual sugar        0.14610715  0.272080238 -0.10128338 -0.372792562
## chlorides             0.21224658  0.148051555  0.09261383  0.666194756
## free sulfur dioxide  -0.03615752  0.513566812 -0.42879287 -0.043537818
## total sulfur dioxide  0.02357485  0.569486959 -0.32241450 -0.034577115
## density               0.39535301  0.233575490  0.33887135 -0.174499758
## pH                   -0.43851962  0.006710793 -0.05769735 -0.003787746
## sulphates             0.24292133 -0.037553916 -0.27978615  0.550872362
## alcohol              -0.11323207 -0.386180959 -0.47167322 -0.122181088
##                             PC5         PC6         PC7         PC8
## fixed acidity         0.08261366 -0.10147858  0.35022736 -0.17759545
## volatile acidity     -0.21873452 -0.41144893  0.53373510 -0.07877531
## citric acid           0.05857268 -0.06959338 -0.10549701 -0.37751558
## residual sugar       -0.73214429 -0.04915555 -0.29066341  0.29984469
## chlorides            -0.24650090 -0.30433857 -0.37041337 -0.35700936
## free sulfur dioxide   0.15915198  0.01400021  0.11659611 -0.20478050
## total sulfur dioxide  0.22246456 -0.13630755  0.09366237  0.01903597
## density              -0.15707671  0.39115230  0.17048116 -0.23922267
## pH                   -0.26752977  0.52211645  0.02513762 -0.56139075
## sulphates            -0.22596222  0.38126343  0.44746911  0.37460432
## alcohol              -0.35068141 -0.36164504  0.32765090 -0.21762556
##                              PC9        PC10         PC11
## fixed acidity         0.194020908  0.24952314  0.639691452
## volatile acidity     -0.129110301 -0.36592473  0.002388597
## citric acid          -0.381449669 -0.62167708 -0.070910304
## residual sugar        0.007522949 -0.09287208  0.184029964
## chlorides             0.111338666  0.21767112  0.053065322
## free sulfur dioxide   0.635405218 -0.24848326 -0.051420865
## total sulfur dioxide -0.592115893  0.37075027  0.068701598
## density               0.020718675  0.23999012 -0.567331898
## pH                   -0.167745886  0.01096960  0.340710903
## sulphates            -0.058367062 -0.11232046  0.069555381
## alcohol               0.037603106  0.30301450 -0.314525906
```

```
###a)Find the eigenvalues
PCA$sdev^2
```

```
##  [1] 3.09913244 1.92590969 1.55054349 1.21323253 0.95929207 0.65960826
##  [7] 0.58379122 0.42295670 0.34464212 0.18133317 0.05955831
```

```
###b) choose the number of the principle component to retain
screeplot(PCA, npcs=12, type = "lines", main="")
```




To decide which principle components to retain, we can look at the screeplot, we select the ones that have eigenvalues that are at least 1 by Kaiser's rule where the average of all eigenvalues is p/p = 1 where p = 11. Additionally, we select the principle component that satisfies the 80% rule since they are able to describe at least 80% of the variance. In this case, it is the first five principle components that we have to retain.

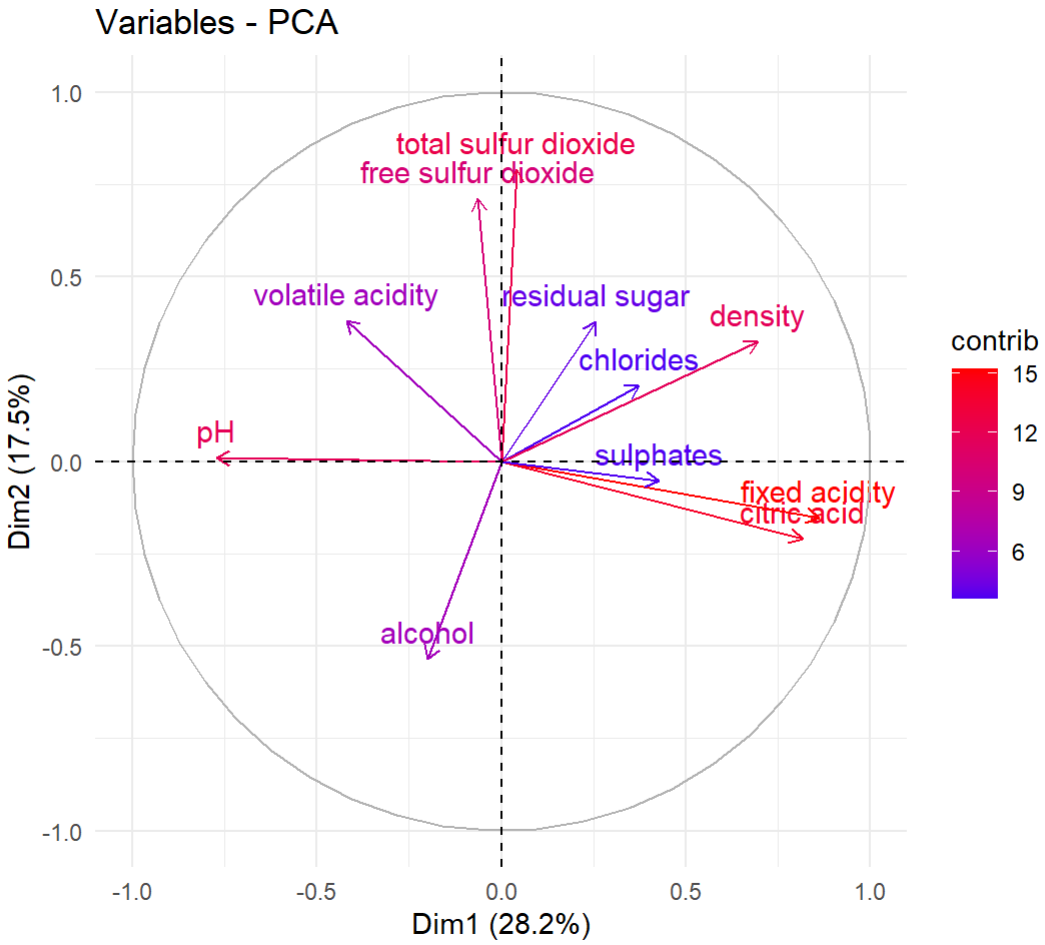
p.s.the eigenvalue of PC5 is extreme close to 1, so we count that as well.

```
### c)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_pca_var(PCA, col.var="contrib", axes=c(1,2))+scale_color_gradient2(low = "white", mid="blu
e", high="red", midpoint = 3)+xlim(c(-1, 1))+ylim(c(-1,1)) +theme_minimal()
```




In the plot, x-axes represent PC2 and y-axis represent PC1. Here,we can see that total sulfur dioxide and free sulfur dioxide strongly and positively influence PC2 but alcohol gives a negative loading on PC2. Sulphates, fixed acidity, citric acid ,and pH strongly influence PC1. Furthermore, Sulphates and pH diverge and form a large angle that is almost 180 degrees meaning that they are negatively correlated.

```
library(FactoMineR)
```

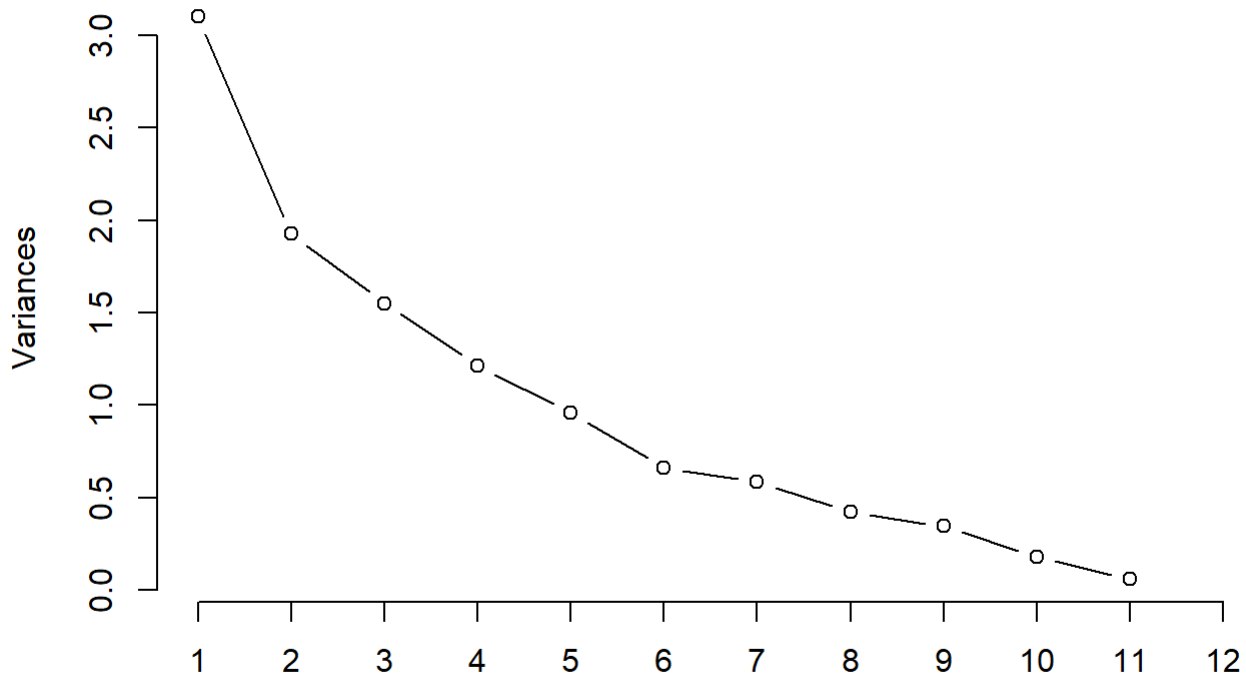
```
## Warning: package 'FactoMineR' was built under R version 4.1.3
```

```
PCA_summary<- summary(PCA(wine_data_chem, graph=FALSE))
```

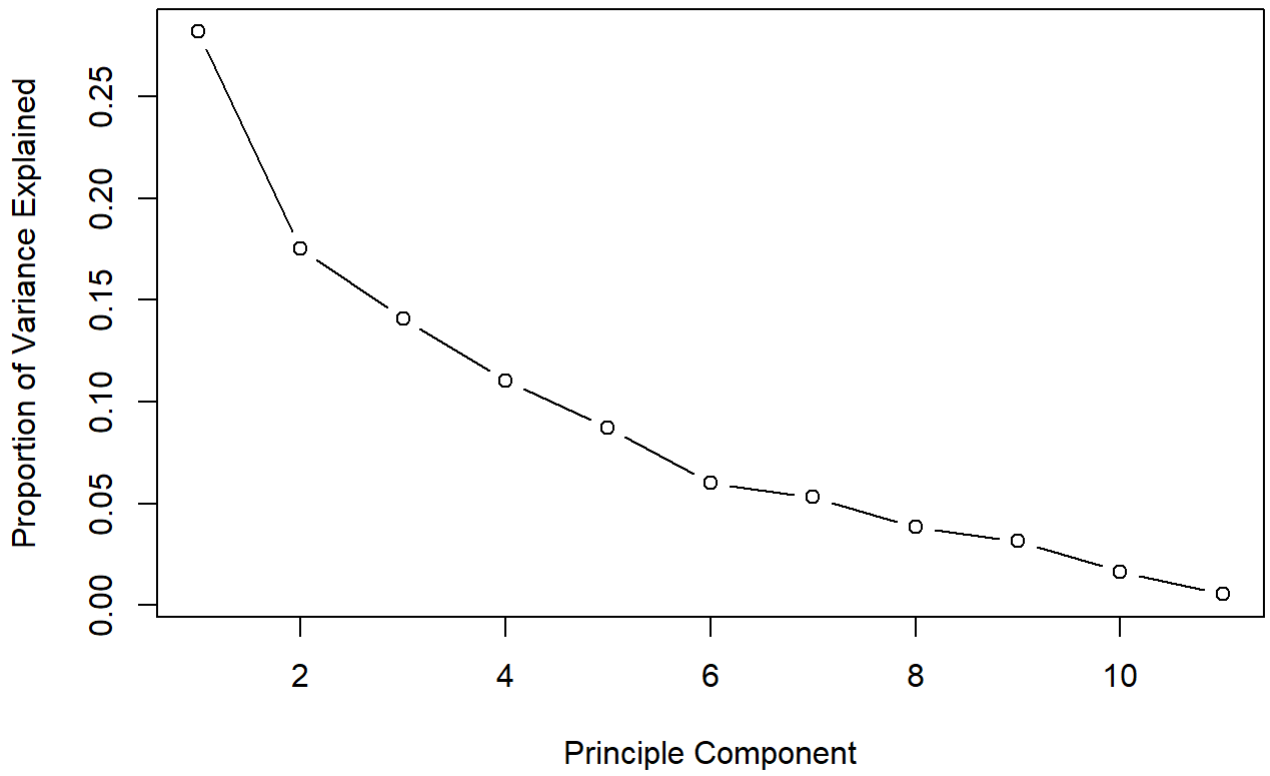
```
##
## Call:
## PCA(X = wine_data_chem, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               3.099   1.926   1.551   1.213   0.959   0.660   0.584
## % of var.             28.174  17.508  14.096  11.029   8.721   5.996   5.307
## Cumulative % of var.  28.174  45.682  59.778  70.807  79.528  85.525  90.832
##                        Dim.8   Dim.9  Dim.10  Dim.11
## Variance               0.423   0.345   0.181   0.060
## % of var.              3.845   3.133   1.648   0.541
## Cumulative % of var.  94.677  97.810  99.459 100.000
##
## Individuals (the 10 first)
##                          Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## 1                     |  2.645 | -1.620  0.053  0.375 |  0.451  0.007  0.029 |
## 2                     |  2.824 | -0.799  0.013  0.080 |  1.857  0.112  0.432 |
## 3                     |  1.936 | -0.748  0.011  0.149 |  0.882  0.025  0.208 |
## 4                     |  3.045 |  2.358  0.112  0.600 | -0.270  0.002  0.008 |
## 5                     |  2.645 | -1.620  0.053  0.375 |  0.451  0.007  0.029 |
## 6                     |  2.540 | -1.584  0.051  0.389 |  0.569  0.011  0.050 |
## 7                     |  2.115 | -1.101  0.024  0.271 |  0.608  0.012  0.083 |
## 8                     |  2.726 | -2.249  0.102  0.681 | -0.417  0.006  0.023 |
## 9                     |  2.093 | -1.087  0.024  0.270 | -0.309  0.003  0.022 |
## 10                    |  3.302 |  0.655  0.009  0.039 |  1.665  0.090  0.254 |
##                        Dim.3    ctr   cos2
## 1                     -1.774  0.127  0.450 |
## 2                     -0.912  0.034  0.104 |
## 3                     -1.171  0.055  0.366 |
## 4                      0.243  0.002  0.006 |
## 5                     -1.774  0.127  0.450 |
## 6                     -1.538  0.095  0.367 |
## 7                     -1.076  0.047  0.259 |
## 8                     -0.987  0.039  0.131 |
## 9                     -1.518  0.093  0.526 |
## 10                     1.209  0.059  0.134 |
##
## Variables (the 10 first)
##                         Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## fixed acidity         |  0.861 23.943  0.742 | -0.153  1.221  0.024 | -0.154
## volatile acidity      | -0.420  5.692  0.176 |  0.382  7.559  0.146 | -0.560
## citric acid           |  0.816 21.495  0.666 | -0.211  2.304  0.044 |  0.297
## residual sugar        |  0.257  2.135  0.066 |  0.378  7.403  0.143 |  0.126
## chlorides             |  0.374  4.505  0.140 |  0.205  2.192  0.042 | -0.115
## free sulfur dioxide   | -0.064  0.131  0.004 |  0.713 26.375  0.508 |  0.534
## total sulfur dioxide  |  0.042  0.056  0.002 |  0.790 32.432  0.625 |  0.401
## density               |  0.696 15.630  0.484 |  0.324  5.456  0.105 | -0.422
## pH                    | -0.772 19.230  0.596 |  0.009  0.005  0.000 |  0.072
## sulphates             |  0.428  5.901  0.183 | -0.052  0.141  0.003 |  0.348
##                         ctr   cos2
```

```
## fixed acidity         1.520  0.024 |
## volatile acidity     20.247  0.314 |
## citric acid           5.676  0.088 |
## residual sugar        1.026  0.016 |
## chlorides             0.858  0.013 |
## free sulfur dioxide  18.386  0.285 |
## total sulfur dioxide 10.395  0.161 |
## density              11.483  0.178 |
## pH                    0.333  0.005 |
## sulphates             7.828  0.121 |
```

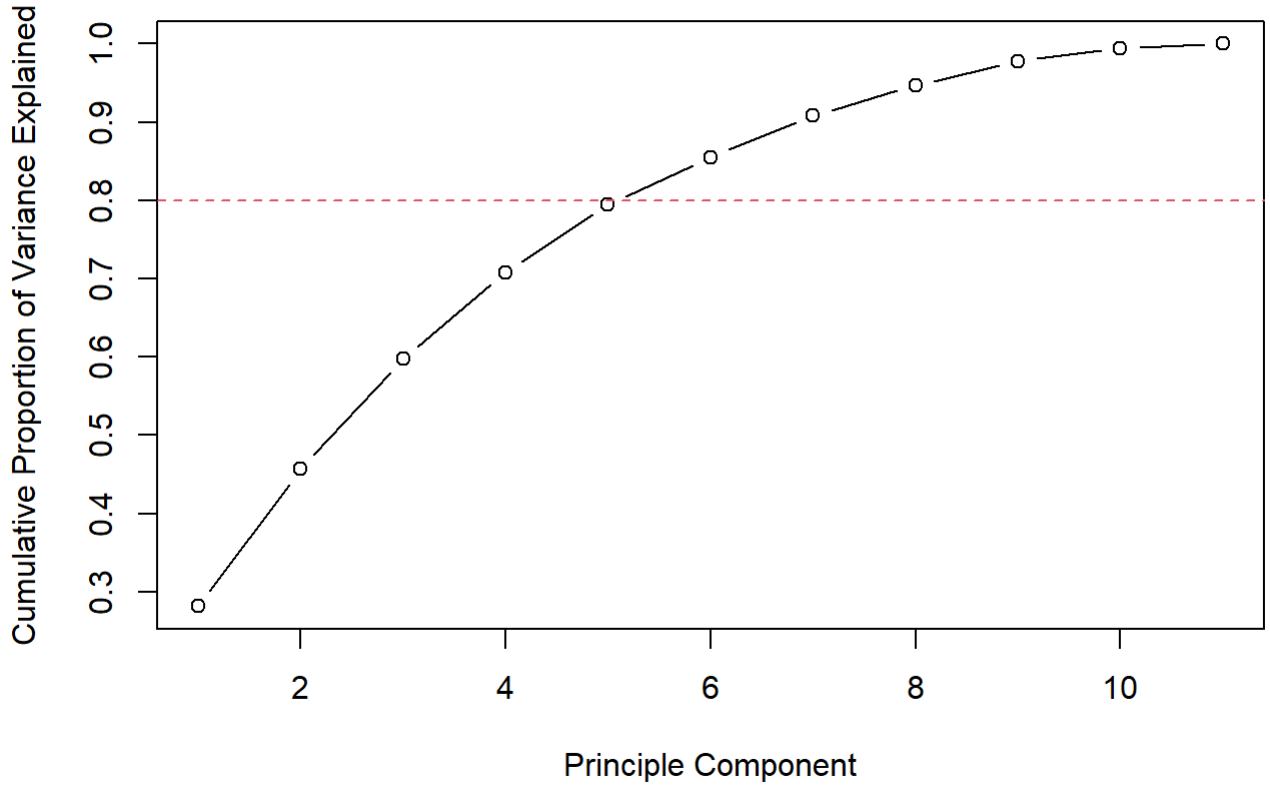
```
###d)
eigens<-PCA$sdev^2
screeplot(PCA, npcs=12, type = "lines", main="")
```




```
#Proportion of variability
plot((eigens/sum(eigens)), xlab="Principle Component", ylab="Proportion of Variance Explained",
type="b")
```

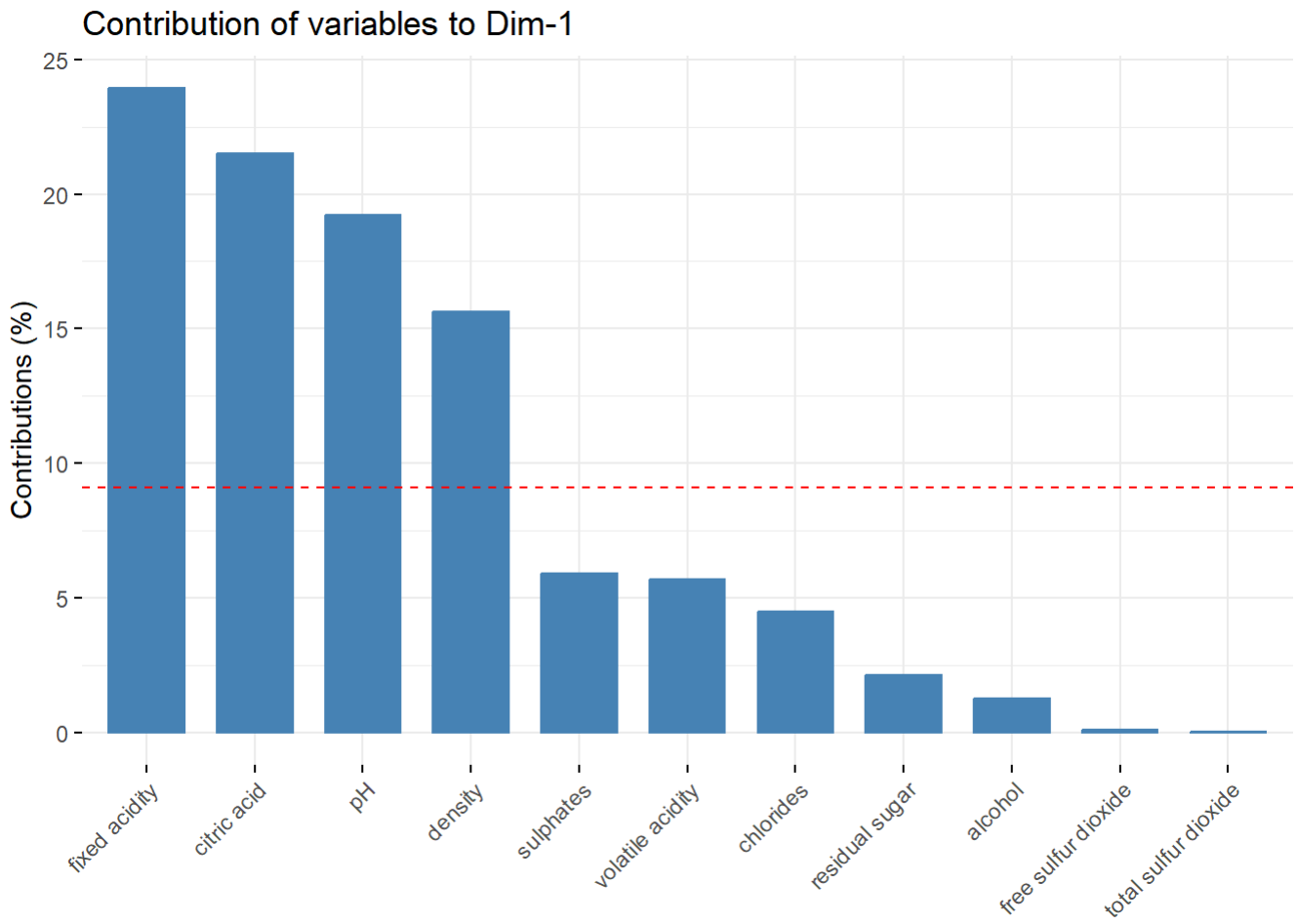



```
#Cumulative proportion of variability
plot(cumsum(eigens/sum(eigens)), xlab="Principle Component", ylab="Cumulative Proportion of Vari
ance Explained", type="b")
abline(h=0.8, col=2, lty=2)
```

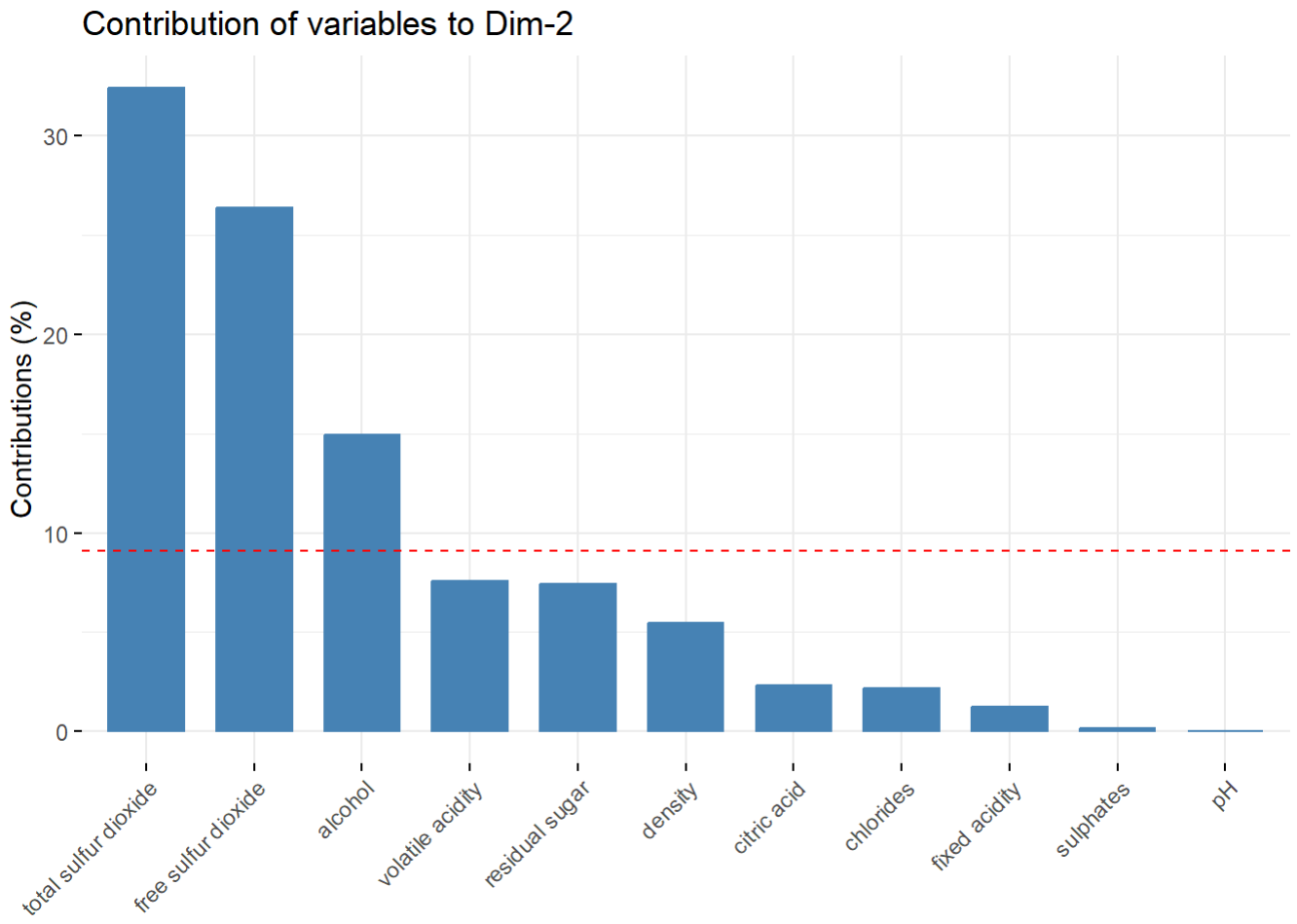



From the summary and the graphs, we can observe that the first principle component explains 28.1% of the variability, PC2 explains 17.5%, PC3 explains 14.1%, PC4 explains 11.0% and finally PC5 explains 8.7% of the variability. In the cumulative proportion of variability plot, we can see that there are five dots representing the first five principle components. Therefore, we can conclude that 79.5% of the entire variability which is approximately 80% in the data set is explained by the first five principle components according to 80% rule.

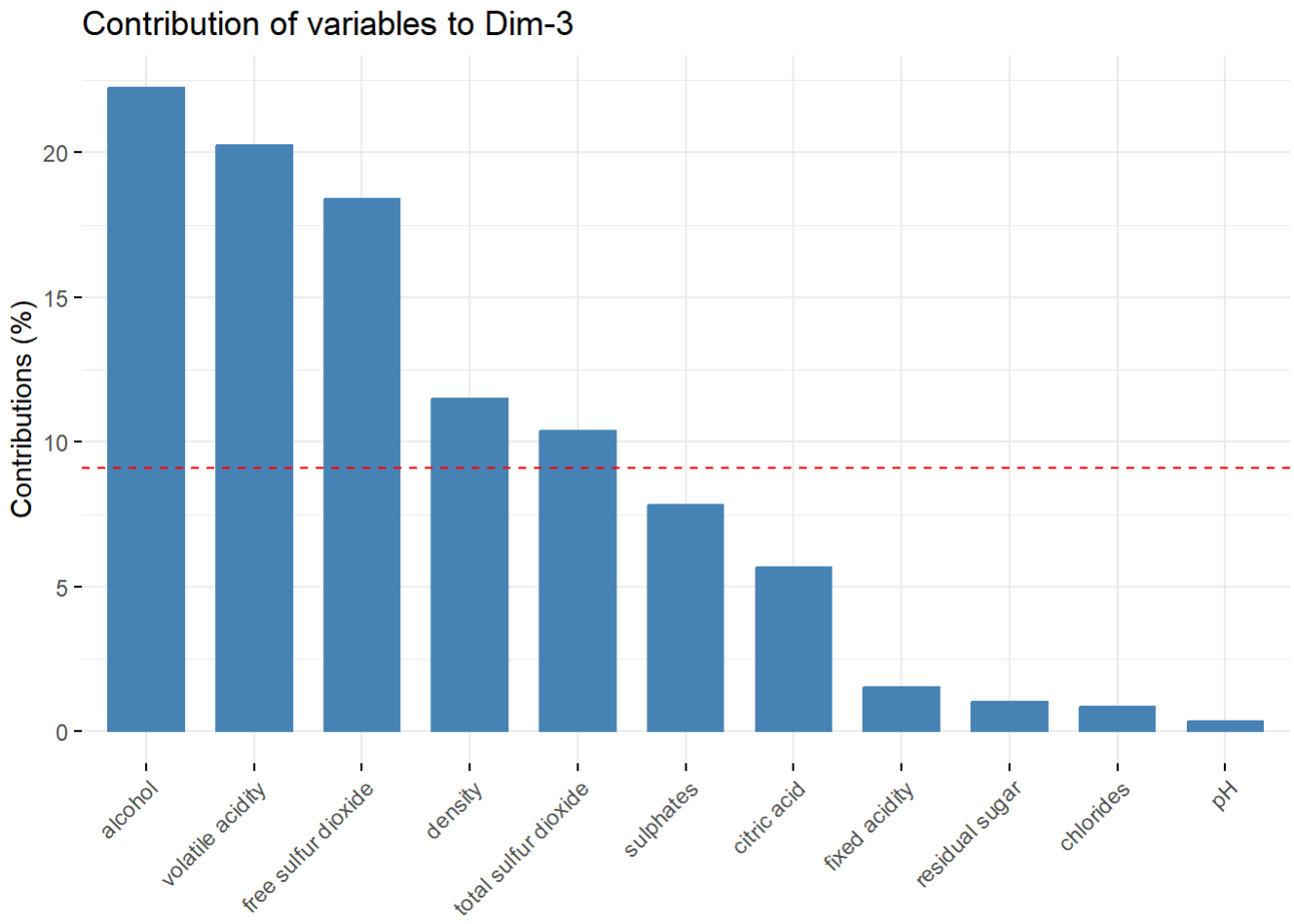
```
###e)  plot the graph of their contributions
fviz_contrib(PCA, choice="var", axes= 1)
```



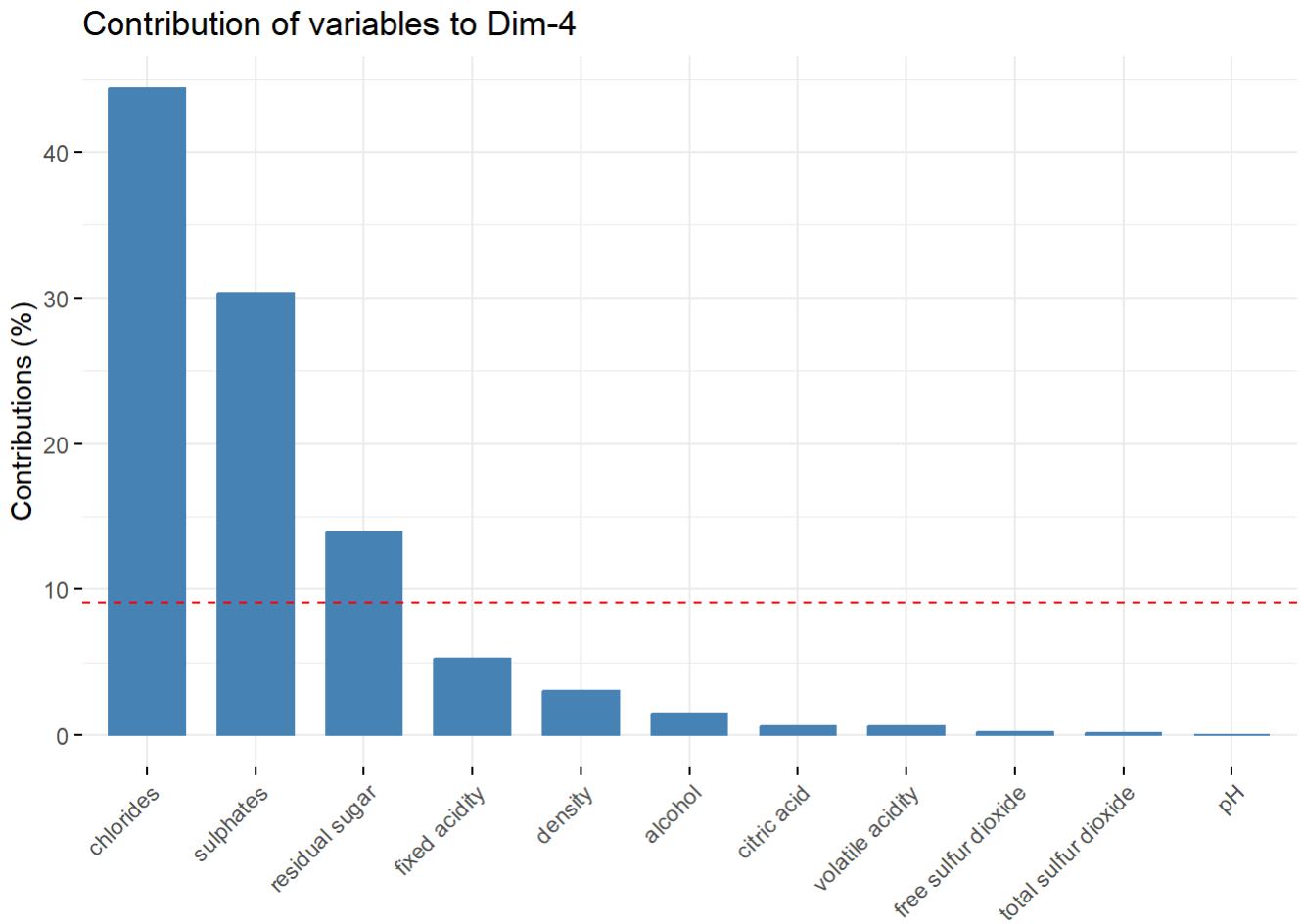

```
fviz_contrib(PCA, choice="var", axes=c(2))
```



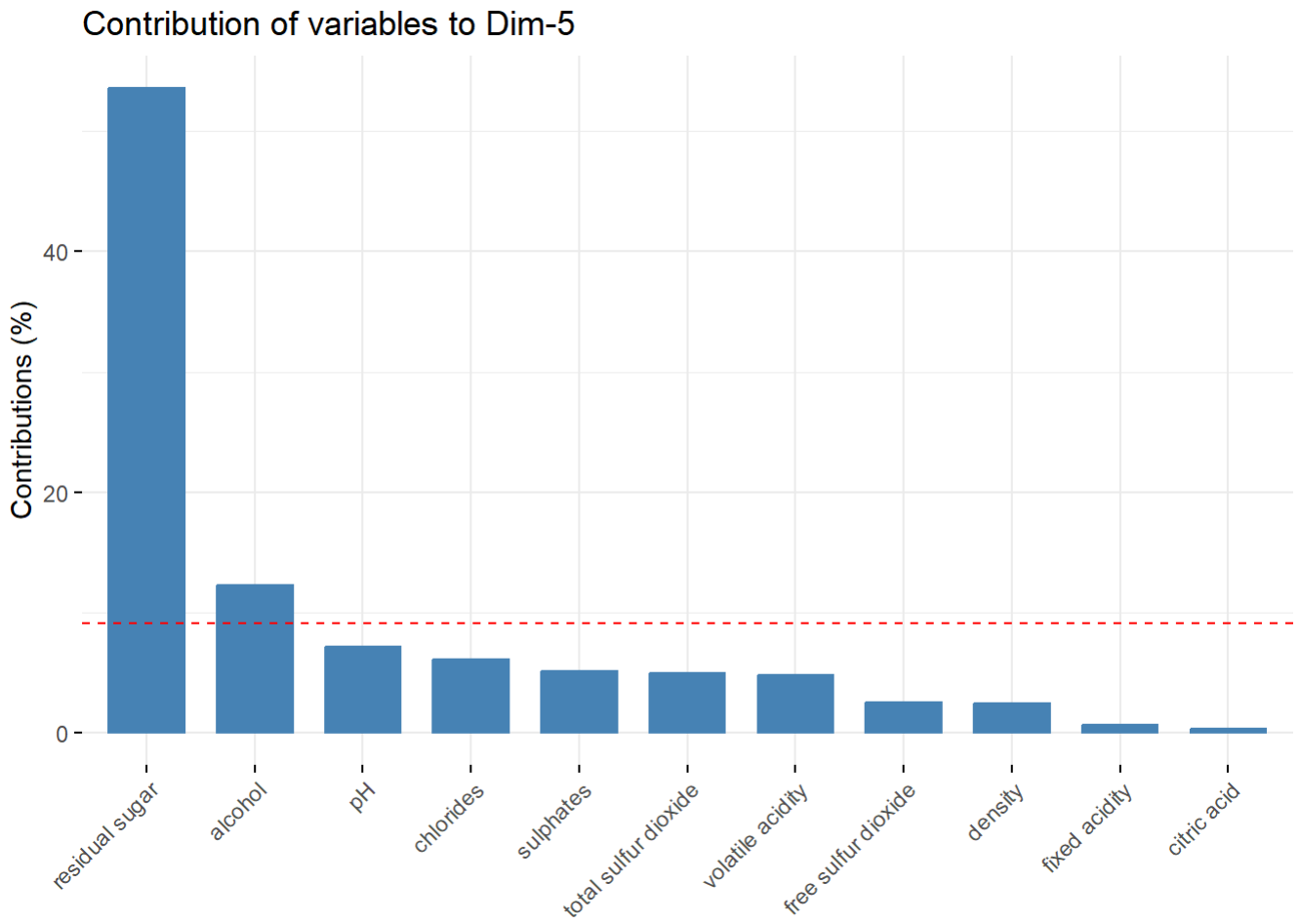

```
fviz_contrib(PCA, choice="var", axes=c(3))
```



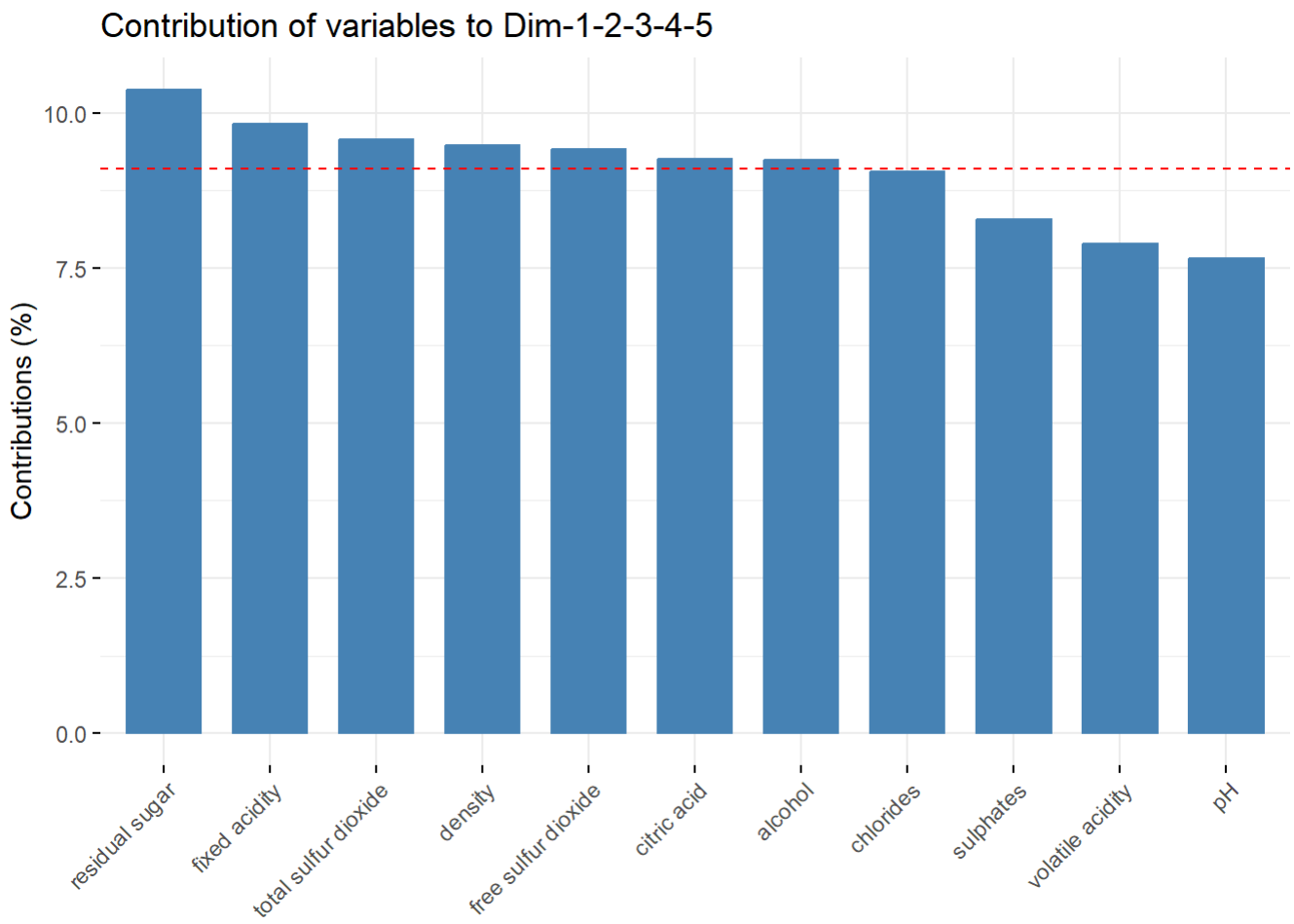

```
fviz_contrib(PCA, choice="var", axes=c(4))
```




```
fviz_contrib(PCA, choice="var", axes=c(5))
```




```
fviz_contrib(PCA, choice="var", axes=c(1,2,3,4,5))
```




We can observe fixed acidity, citric acid, pH, and density are strongly correlated to the first principle component. Free sulfur dioxide and total sulfur dioxide have extreme low contribution and it is possible to remove them to simplify the overall analysis.However, in the second principle component, total sulfur dioxide, free sulfur dioxide and alcohol passes the expected average contribution of all the variables, meaning that they have strong correlation with PC2, but the sulphates and pH have really low contribution in this case. Alcohol, volatile acidity, free sulfur dioxide, density, and total sulfur dioxide highly influence the third principle component, and chlorides and pH have less influence in this component. In fourth principle component, chlorides, sulphates, and residual sugar have high contribution, but again pH has the least correlation in this case. In the fifth principle component, residual sugar and alcohol has stronger correlation but fixed acidity and citric acid has the least correlation.

Overall, the contribution of all the variables except sulphates, volatile acidity, and pH have passed the expected average contribution for all the variables, meaning that they have a stronger correlation with the first five principle components. After analyzing contribution of variables in each component, we know that pH has the least contribution (correlation) to most of the components ,therefore, we can see that pH has the least contribution in the last plot.

f. It is important to standardize the variables so that the covariance are easily comparable for each pair of features. If the variables are not standardized, the variance changes, and features with larger ranges of numbers will have higher covariance, therefore, the values of principle components will become different. For instance, sulfur dioxide and citric acid would have a larger influence on the principle components and the other variables with less variance would have less influence. Moreover, after standardizing the data, we know that the variables are scaled equally and they are directly comparable to each other in principle component analysis since standardization prevents any variable that has a larger variance from dominating the principle components.