Wine Quality

```
# removes all objects from the current workspace
rm(list = ls())

set.seed(2022)
# load all packages

library(plsRglm)
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(tidyverse)
```

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

```
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## v purrr   0.3.4
```

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

```
#load data
red <- read.csv('https://afs-wine-dataset.s3.amazonaws.com/winequality-red.csv', sep=';')
white <- read.csv('https://afs-wine-dataset.s3.amazonaws.com/winequality-white.csv', sep =';')
```

```
#look at the data
head(red)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            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             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
# merge red wine and white wine datasets
data <- rbind(red, white)
```

```
# First six rows
head(data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            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             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
# last six rows of dataset
tail(data)
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 6492           6.5             0.23        0.38            1.3     0.032
## 6493           6.2             0.21        0.29            1.6     0.039
## 6494           6.6             0.32        0.36            8.0     0.047
```

```
## 6495           6.5             0.24        0.19            1.2     0.041
## 6496           5.5             0.29        0.30            1.1     0.022
## 6497           6.0             0.21        0.38            0.8     0.020
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 6492                  29                  112 0.99298 3.29      0.54     9.7
## 6493                  24                   92 0.99114 3.27      0.50    11.2
## 6494                  57                  168 0.99490 3.15      0.46     9.6
## 6495                  30                  111 0.99254 2.99      0.46     9.4
## 6496                  20                  110 0.98869 3.34      0.38    12.8
## 6497                  22                   98 0.98941 3.26      0.32    11.8
##      quality
## 6492       5
## 6493       6
## 6494       5
## 6495       6
## 6496       7
## 6497       6
```

```
# structure of the data
str(data)
```

```
## 'data.frame':    6497 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
# number of missing values in each column
sapply(data, function(x) sum(is.na(x)))
```

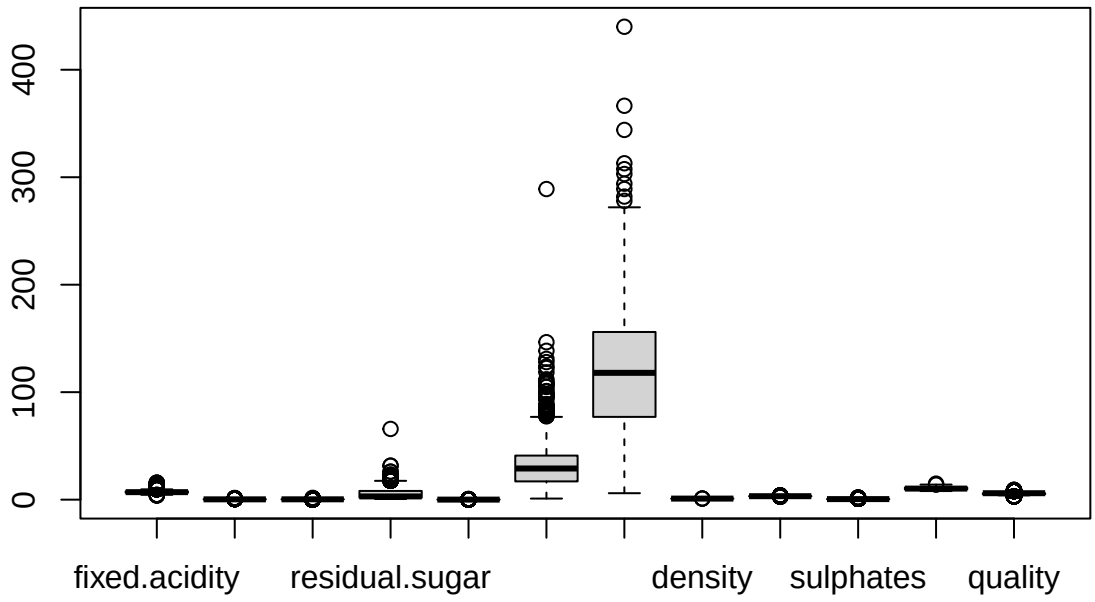
```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

```
# data summary
summary(data)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
```

```
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

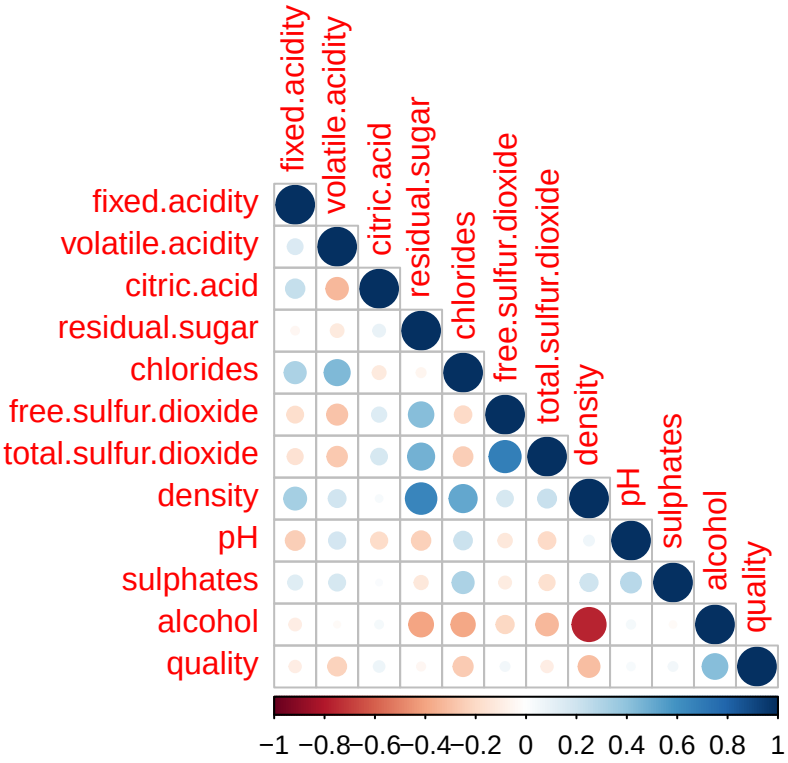
```
# look at outliers
boxplot(data)
```



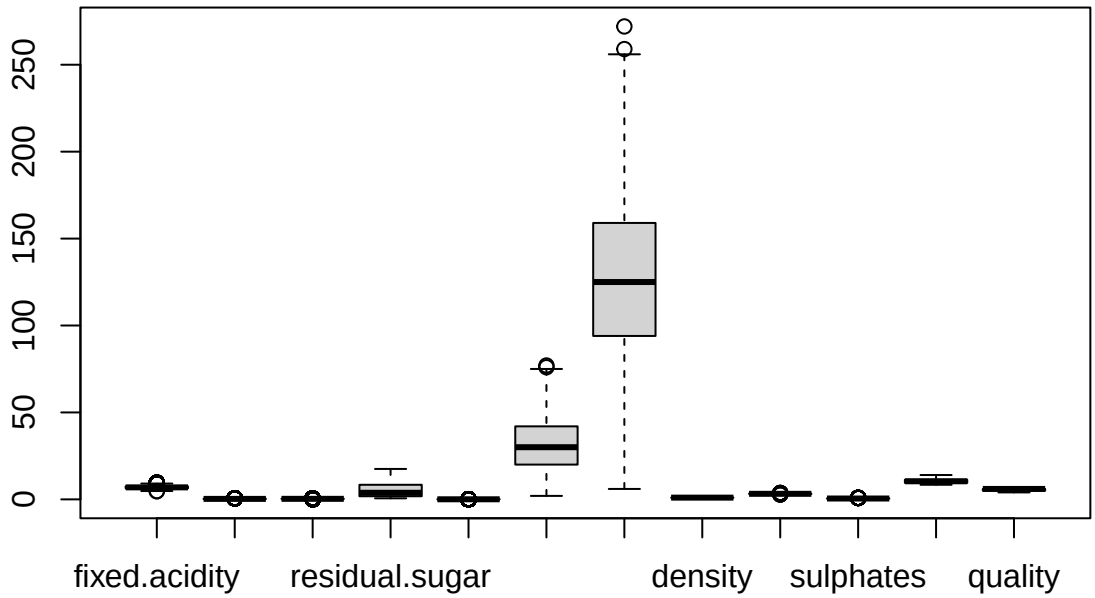

```
clear_data <- function(data){
```

```r
  # each feature of input data is analysed
  for (i in 1:ncol(data)){
    # particular feature observations
    vec <- data[, i]

    # values those are out of 1.5 * IQR
    vec_out <- boxplot.stats(vec)$out

    # all outlier values found in feature vector assigned as NA
    vec[vec %in% vec_out] <- NA

    # data feature is updated
    data[, i] <- vec
  }

  # only complete observation data subset is returned
  data[complete.cases(data), ]
}

# Creating new data without outliers using defined function
data <- clear_data(data)
```

```r
# Correlation between features
correlation <- cor(x=data%>%select_if(is.numeric))
corrplot(correlation, method = 'circle',type='lower')
```



```r
# looking at boxplot after removing outliers
boxplot(data)
```




```r
#Split the dataset to Train and Test
trainRowNumbers <- createDataPartition(data$quality, p=0.8, list=FALSE)

# Create the training  dataset
trainData <- data[trainRowNumbers,]

# Step 3: Create the test dataset
testData <- data[-trainRowNumbers,]
```

```r
set.seed(2022)

cv.modpls<-cv.plsR(quality~.,data=data,nt=10, verbose =F, NK=20)
```

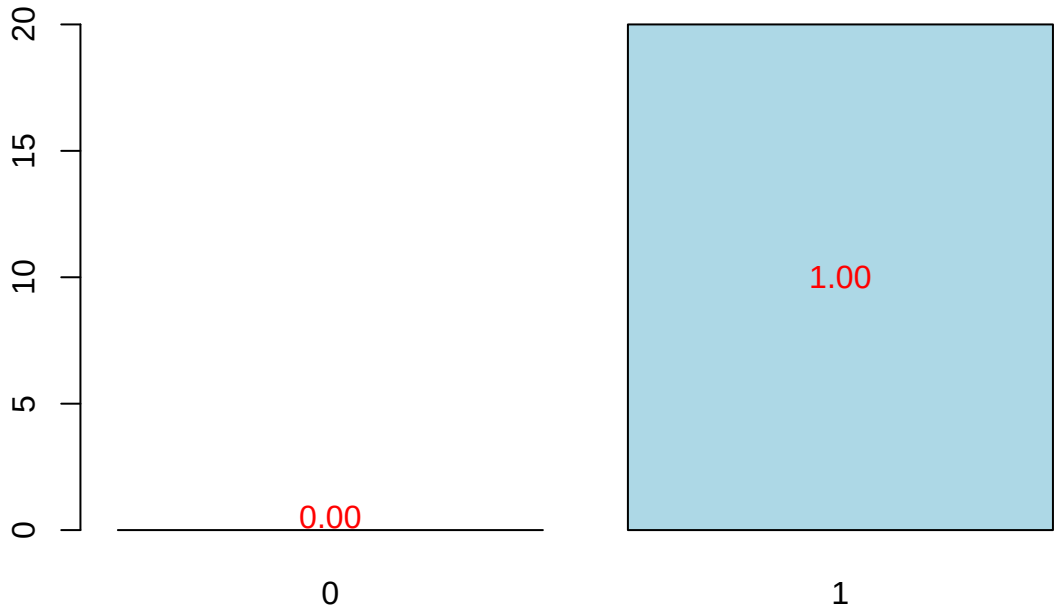
```r
cv.modpls
```

```
## Number of repeated crossvalidations:
## [1] 20
## Number of folds for each crossvalidation:
## [1] 5
```

```r
# We sum up the results in a single table using the summary.
res.cv.modpls=cvtable(summary(cv.modpls))
```

```
## ____**********************************************____
## ____Component____ 1 ____
## ____Component____ 2 ____
## ____Component____ 3 ____
## ____Component____ 4 ____
## ____Component____ 5 ____
## ____Component____ 6 ____
## ____Component____ 7 ____
## ____Component____ 8 ____
## ____Component____ 9 ____
## ____Component____ 10 ____
## ____Predicting X without NA neither in X nor in Y____

## Loading required namespace: plsdof

## ****__________________________________________________****
##
##
## NK: 1,  2,  3,  4,  5,  6,  7,  8,  9,  10
## NK: 11,  12,  13,  14,  15,  16,  17,  18,  19,  20
##
## CV Q2 criterion:
##  0  1
##  0 20
##
## CV Press criterion:
##  1  2  3  4  5  6  7  8  9 10
##  0  0  0  0  0  1  1  0  3 15
```

```r
# The results, based on the use of the Q2 criterion to find the number of components
plot(res.cv.modpls)
```




```
# nt: number of components is set to 1
res<-plsR(quality~.,data=trainData,nt=1,pvals.expli=TRUE)
```

```
## ____**********************************************____
## ____Component____ 1 ____
## ____Predicting X without NA neither in X nor in Y____
## ****______________________________________________****
```

```
# Model's Descriptive Statistics
res
```

```
## Number of required components:
## [1] 1
## Number of successfully computed components:
## [1] 1
## Coefficients:
##                                [,1]
## Intercept              4.182153e+01
## fixed.acidity         -3.429035e-02
## volatile.acidity      -6.878919e-01
## citric.acid            3.173985e-01
## residual.sugar        -3.470784e-03
## chlorides             -5.579842e+00
## free.sulfur.dioxide    1.362143e-03
## total.sulfur.dioxide -5.712616e-04
```

```
## density              -3.728231e+01
## pH                    9.880713e-02
## sulphates             1.658911e-01
## alcohol               1.251309e-01
## Information criteria and Fit statistics:
##                AIC    RSS_Y      R2_Y R2_residY RSS_residY  AIC.std DoF.dof
## Nb_Comp_0 8946.717 2282.205        NA        NA   3872.000 10994.10 1.00000
## Nb_Comp_1 8122.882 1843.956 0.1920288 0.1920288   3128.464 10170.26 5.66823
##           sigmahat.dof   AIC.dof   BIC.dof  GMDL.dof DoF.naive sigmahat.naive
## Nb_Comp_0    0.7677320 0.5895646 0.5905176 -1013.403         1      0.7677320
## Nb_Comp_1    0.6904201 0.4775006 0.4818690 -1400.051         2      0.6901821
##           AIC.naive BIC.naive GMDL.naive
## Nb_Comp_0 0.5895646 0.5905176  -1013.403
## Nb_Comp_1 0.4765973 0.4781376  -1420.141
```

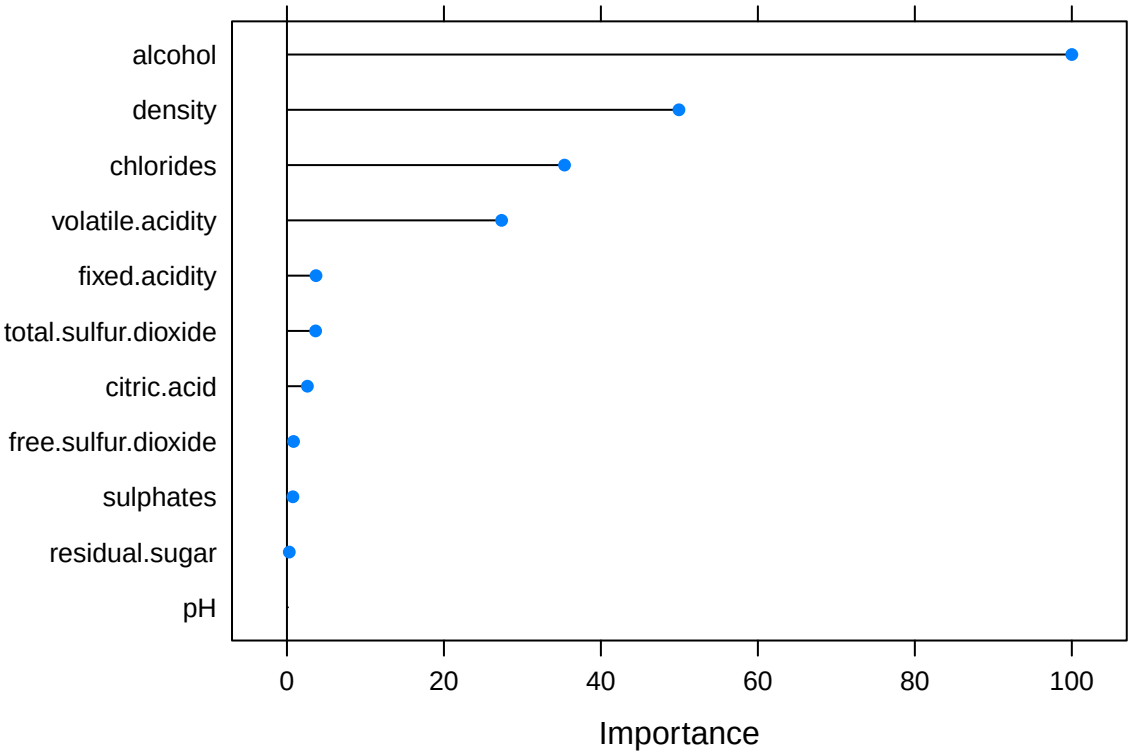
```
#Predictions
# Mean Absolute Error
predict <- testData$quality
predictions <- predict(res, testData[,1:11])
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
MAE(predictions,predict)
```

```
## [1] 0.548803
```

```
# Plot feature importance
# Grab a coffee, this one takes some time
train(data[,1:11], data[,12], method='plsRglm', verbose =F, preProcess = c("center","scale")) %>% varIm
```


alcohol
density
chlorides
volatile.acidity
fixed.acidity
total.sulfur.dioxide
citric.acid
free.sulfur.dioxide
sulphates
residual.sugar
pH
0
20
40
60
80
100
Importance