STAT240 Group3 Data and Analysis Plan

Tasheena Narraidoo, Michael Shi, Reynaldo Pena

12/04/16

Contents


```
require(mosaic)
require(mosaicData)
require(MVA)
require(aplpack)
require(scatterplot3d)
require(MASS)
require(tourr)
require(plyr)
library(caTools)
options(digits=3)
```

```
whitewine <- read.csv("winequalitywhite.csv", sep = ";", header = TRUE)
redwine <- read.csv("winequalityred.csv", sep = ";", header = TRUE)
```

```
names(whitewine)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

```r
names(redwine)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

We have 2 datasets, one for red wine and one for white wine. They have the same variable names. The datasets have each 12 variables.

Data Codebook

variables

white wine

We have 4,898 observations for white wine.

```r
nrow(whitewine)
```

```
## [1] 4898
```

```r
summary(whitewine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 3.80   Min.   :0.080    Min.   :0.000   Min.   : 0.6
##  1st Qu.: 6.30   1st Qu.:0.210    1st Qu.:0.270   1st Qu.: 1.7
##  Median : 6.80   Median :0.260    Median :0.320   Median : 5.2
##  Mean   : 6.85   Mean   :0.278    Mean   :0.334   Mean   : 6.4
##  3rd Qu.: 7.30   3rd Qu.:0.320    3rd Qu.:0.390   3rd Qu.: 9.9
##  Max.   :14.20   Max.   :1.100    Max.   :1.660   Max.   :65.8
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.009   Min.   :  2.0       Min.   :  9          Min.   :0.987
##  1st Qu.:0.036   1st Qu.: 23.0       1st Qu.:108          1st Qu.:0.992
##  Median :0.043   Median : 34.0       Median :134          Median :0.994
##  Mean   :0.046   Mean   : 35.3       Mean   :138          Mean   :0.994
##  3rd Qu.:0.050   3rd Qu.: 46.0       3rd Qu.:167          3rd Qu.:0.996
##  Max.   :0.346   Max.   :289.0       Max.   :440          Max.   :1.039
##        pH         sulphates       alcohol        quality
##  Min.   :2.72   Min.   :0.22   Min.   : 8.0   Min.   :3.00
##  1st Qu.:3.09   1st Qu.:0.41   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.18   Median :0.47   Median :10.4   Median :6.00
##  Mean   :3.19   Mean   :0.49   Mean   :10.5   Mean   :5.88
##  3rd Qu.:3.28   3rd Qu.:0.55   3rd Qu.:11.4   3rd Qu.:6.00
##  Max.   :3.82   Max.   :1.08   Max.   :14.2   Max.   :9.00
```

fixed.acidity

fixed.acidity takes values from 3.80 to 14.20. Its unit of measurement is g(tartaric acid)/dm3. The mean fixed.acidity is 6.85. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$fixed.acidity)
```

```
##  min  Q1 median  Q3  max mean    sd    n missing
##  3.8 6.3    6.8 7.3 14.2 6.85 0.844 4898       0
```

volatile.acidity

volatile.acidity takes values from 0.080 to 1.1. Its unit of measurement is g(acetic acid)/dm3. The mean volatile.acidity is 0.278. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$volatile.acidity)
```

```
##   min   Q1 median   Q3 max  mean    sd    n missing
##  0.08 0.21   0.26 0.32 1.1 0.278 0.101 4898       0
```

citric.acid

citric.acid takes values from 0 to 1.66. Its unit of measurement is g/dm3. The mean citric.acid is 0.334. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$citric.acid)
```

```
##  min   Q1 median   Q3  max  mean    sd    n missing
##    0 0.27   0.32 0.39 1.66 0.334 0.121 4898       0
```

residual.sugar

residual.sugar takes values from 0.6 to 65.8. Its unit of measurement is g/dm3. The mean residual.sugar is 6.39. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$residual.sugar)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  0.6 1.7    5.2 9.9 65.8 6.39 5.07 4898       0
```

chlorides

chlorides takes values from 0.009 to 0.346. Its unit of measurement is g(sodium chloride)/dm3. The mean chlorides is 0.0458. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$chlorides)
```

```
##    min    Q1 median   Q3   max   mean     sd    n missing
##  0.009 0.036  0.043 0.05 0.346 0.0458 0.0218 4898       0
```

free.sulfur.dioxide

free.sulfur.dioxide takes value from 2.0 to 289.0. Its unit of measurement is mg/dm3. The mean free.sulfur.dioxide is 35.3. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$free.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean sd    n missing
##    2 23     34 46 289 35.3 17 4898       0
```

total.sulfur.dioxide

total.sulfur.dioxide takes values from 9 to 440. Its unit of measurement is mg/dm3. The mean total.sulfur.dioxide is 138. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$total.sulfur.dioxide)
```

```
##  min  Q1 median  Q3 max mean   sd    n missing
##    9 108    134 167 440  138 42.5 4898       0
```

density

density takes values from .987 to 1.039. Its unit of measurement is g/cm3. The mean density is 0.994. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$density)
```

```
##    min    Q1 median    Q3  max  mean      sd    n missing
##  0.987 0.992  0.994 0.996 1.04 0.994 0.00299 4898       0
```

pH

pH takes values form 2.72 to 3.82. The mean pH is 3.19. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$pH)
```

```
##   min   Q1 median   Q3  max mean    sd    n missing
##  2.72 3.09   3.18 3.28 3.82 3.19 0.151 4898       0
```

sulphates

sulphates take values from .22 to 1.08. Its unit of measurement is g(potassium sulphate)/dm3. The mean sulphates is 0.49. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$sulphates)
```

```
##   min   Q1 median   Q3  max mean    sd    n missing
##  0.22 0.41   0.47 0.55 1.08 0.49 0.114 4898       0
```

alcohol

alcohol takes values from 8 to 14.2. Its unit of measurement is vol.%. The mean alcohol is 10.5 %. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$alcohol)
```

```
##  min  Q1 median   Q3  max mean   sd    n missing
##    8 9.5   10.4 11.4 14.2 10.5 1.23 4898       0
```

quality

quality (which is within the [0,10] range) takes value from 3 to 9. It is the value attributed to the quality of wine, 0 being the lowest quality and 10, the highest.

```
tally(~quality, data=whitewine)
```

```
## quality
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

red wine

```
nrow(redwine)
```

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

```
summary(redwine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.120    Min.   :0.000   Min.   : 0.90
##  1st Qu.: 7.10   1st Qu.:0.390    1st Qu.:0.090   1st Qu.: 1.90
##  Median : 7.90   Median :0.520    Median :0.260   Median : 2.20
##  Mean   : 8.32   Mean   :0.528    Mean   :0.271   Mean   : 2.54
##  3rd Qu.: 9.20   3rd Qu.:0.640    3rd Qu.:0.420   3rd Qu.: 2.60
##  Max.   :15.90   Max.   :1.580    Max.   :1.000   Max.   :15.50
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.012   Min.   : 1.0        Min.   :  6.0        Min.   :0.990
##  1st Qu.:0.070   1st Qu.: 7.0        1st Qu.: 22.0        1st Qu.:0.996
##  Median :0.079   Median :14.0        Median : 38.0        Median :0.997
##  Mean   :0.087   Mean   :15.9        Mean   : 46.5        Mean   :0.997
##  3rd Qu.:0.090   3rd Qu.:21.0        3rd Qu.: 62.0        3rd Qu.:0.998
##  Max.   :0.611   Max.   :72.0        Max.   :289.0        Max.   :1.004
##        pH          sulphates        alcohol        quality
##  Min.   :2.74   Min.   :0.330   Min.   : 8.4   Min.   :3.00
##  1st Qu.:3.21   1st Qu.:0.550   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.31   Median :0.620   Median :10.2   Median :6.00
##  Mean   :3.31   Mean   :0.658   Mean   :10.4   Mean   :5.64
##  3rd Qu.:3.40   3rd Qu.:0.730   3rd Qu.:11.1   3rd Qu.:6.00
##  Max.   :4.01   Max.   :2.000   Max.   :14.9   Max.   :8.00
```

fixed.acidity

fixed.acidity takes values from 4.6 to 15.9. Its unit of measurement is g(tartaric acid)/dm3. The mean fixed.acidity is 8.32. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$fixed.acidity)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  4.6 7.1    7.9 9.2 15.9 8.32 1.74 1599       0
```

volatile.acidity

volatile.acidity takes values from 0.12 to 1.58. Its unit of measurement is g(acetic acid)/dm3. The mean volatile.acidity is 0.528. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$volatile.acidity)
```

```
##   min   Q1 median   Q3  max  mean    sd    n missing
##  0.12 0.39   0.52 0.64 1.58 0.528 0.179 1599       0
```

citric.acid

citric.acid takes values from 0.09 to 1. Its unit of measurement is g/dm3. The mean citric.acid is 0.271. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$citric.acid)
```

```
##  min   Q1 median   Q3 max  mean    sd    n missing
##    0 0.09   0.26 0.42   1 0.271 0.195 1599       0
```

residual.sugar

residual.sugar takes values from 0.9 to 15.5. Its unit of measurement is g/dm3. The mean residual.sugar is 2.54. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$residual.sugar)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  0.9 1.9    2.2 2.6 15.5 2.54 1.41 1599       0
```

chlorides

chlorides takes values from 0.012 to 0.611. Its unit of measurement is g(sodium chloride)/dm3. The mean chlorides is 0.0875. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$chlorides)
```

```
##    min   Q1 median   Q3   max   mean     sd    n missing
##  0.012 0.07  0.079 0.09 0.611 0.0875 0.0471 1599       0
```

free.sulfur.dioxide

free.sulfur.dioxide takes value from 1 to 72. Its unit of measurement is mg/dm3. The mean free.sulfur.dioxide is 15.9. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$free.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean   sd    n missing
##    1  7     14 21  72 15.9 10.5 1599       0
```

total.sulfur.dioxide

total.sulfur.dioxide takes values from 6 to 289. Its unit of measurement is mg/dm3. The mean total.sulfur.dioxide is 46.5. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$total.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean   sd    n missing
##    6 22     38 62 289 46.5 32.9 1599       0
```

density

density takes values from 0.99 to 1. Its unit of measurement is g/cm3. The mean density is 0.997. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$density)
```

```
##   min    Q1 median    Q3 max  mean      sd    n missing
##  0.99 0.996  0.997 0.998   1 0.997 0.00189 1599       0
```

pH

pH takes values form 2.74 to 4.01. The mean pH is 3.31. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$pH)
```

```
##   min   Q1 median  Q3  max mean    sd    n missing
##  2.74 3.21   3.31 3.4 4.01 3.31 0.154 1599       0
```

sulphates

sulphates take values from 0.33 to 2. Its unit of measurement is g(potassium sulphate)/dm3. The mean sulphates is 0.658. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$sulphates)
```

```
##   min   Q1 median   Q3 max  mean   sd    n missing
##  0.33 0.55   0.62 0.73   2 0.658 0.17 1599       0
```

alcohol

alcohol takes values from 8.4 to 14.9. Its unit of measurement is vol.%. The mean alcohol is 10.4 %. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$alcohol)
```

```
##  min  Q1 median   Q3  max mean   sd    n missing
##  8.4 9.5   10.2 11.1 14.9 10.4 1.07 1599       0
```

quality

quality (which is within the [0,10] range) takes value from 3 to 8. It is the value attributed to the quality of wine, 0 being the lowest quality and 10, the highest.

```
tally(~quality, data=redwine)
```

```
## quality
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

combined dataset

```
library(plyr)
nrow(redwine) #1599
```

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

```
nrow(whitewine) # 4898
```

```
## [1] 4898
```

```
redwine[,"type"]  <- c("red")
whitewine[,"type"]  <- c("white")

wine <- join(redwine, whitewine,type = "full")
```

```
## Joining by: fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dio
```

We have a total of 6497 observations for the combined dataset.

```
nrow(whitewine)
```

```
## [1] 4898
```

```
summary(whitewine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 3.80   Min.   :0.080    Min.   :0.000   Min.   : 0.6
##  1st Qu.: 6.30   1st Qu.:0.210    1st Qu.:0.270   1st Qu.: 1.7
##  Median : 6.80   Median :0.260    Median :0.320   Median : 5.2
##  Mean   : 6.85   Mean   :0.278    Mean   :0.334   Mean   : 6.4
```

```
##  3rd Qu.: 7.30   3rd Qu.:0.320    3rd Qu.:0.390   3rd Qu.: 9.9
##  Max.   :14.20   Max.   :1.100    Max.   :1.660   Max.   :65.8
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.009   Min.   :  2.0       Min.   :  9          Min.   :0.987
##  1st Qu.:0.036   1st Qu.: 23.0       1st Qu.:108          1st Qu.:0.992
##  Median :0.043   Median : 34.0       Median :134          Median :0.994
##  Mean   :0.046   Mean   : 35.3       Mean   :138          Mean   :0.994
##  3rd Qu.:0.050   3rd Qu.: 46.0       3rd Qu.:167          3rd Qu.:0.996
##  Max.   :0.346   Max.   :289.0       Max.   :440          Max.   :1.039
##        pH          sulphates       alcohol        quality
##  Min.   :2.72   Min.   :0.22   Min.   : 8.0   Min.   :3.00
##  1st Qu.:3.09   1st Qu.:0.41   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.18   Median :0.47   Median :10.4   Median :6.00
##  Mean   :3.19   Mean   :0.49   Mean   :10.5   Mean   :5.88
##  3rd Qu.:3.28   3rd Qu.:0.55   3rd Qu.:11.4   3rd Qu.:6.00
##  Max.   :3.82   Max.   :1.08   Max.   :14.2   Max.   :9.00
##      type
##  Length:4898
##  Class :character
##  Mode  :character
##
##
##
```

fixed.acidity

fixed.acidity takes values from 3.80 to 15.9. Its unit of measurement is g(tartaric acid)/dm3. The mean fixed.acidity is 7.22. It is a continuous input variable in assessing wine quality.

```
favstats(wine$fixed.acidity)
```

```
##  min  Q1 median  Q3  max mean  sd    n missing
##  3.8 6.4      7 7.7 15.9 7.22 1.3 6497       0
```

volatile.acidity

volatile.acidity takes values from 0.080 to 1.58. Its unit of measurement is g(acetic acid)/dm3. The mean volatile.acidity is 0.34. It is a continuous input variable in assessing wine quality.

```
favstats(wine$volatile.acidity)
```

```
##   min   Q1 median  Q3  max mean    sd    n missing
##  0.08 0.23   0.29 0.4 1.58 0.34 0.165 6497       0
```

citric.acid

citric.acid takes values from 0 to 1.66. Its unit of measurement is g/dm3. The mean citric.acid is 0.319. It is a continuous input variable in assessing wine quality.

```
favstats(wine$citric.acid)
```

```
##  min   Q1 median   Q3  max  mean    sd    n missing
##    0 0.25   0.31 0.39 1.66 0.319 0.145 6497       0
```

residual.sugar

residual.sugar takes values from 0.6 to 65.8. Its unit of measurement is g/dm3. The mean residual.sugar is 5.44. It is a continuous input variable in assessing wine quality.

```
favstats(wine$residual.sugar)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  0.6 1.8      3 8.1 65.8 5.44 4.76 6497       0
```

chlorides

chlorides takes values from 0.009 to 0.611. Its unit of measurement is g(sodium chloride)/dm3. The mean chlorides is 0.056. It is a continuous input variable in assessing wine quality.

```
favstats(wine$chlorides)
```

```
##    min    Q1 median    Q3   max  mean    sd    n missing
##  0.009 0.038  0.047 0.065 0.611 0.056 0.035 6497       0
```

free.sulfur.dioxide

free.sulfur.dioxide takes value from 1 to 289.0. Its unit of measurement is mg/dm3. The mean free.sulfur.dioxide is 30.5. It is a continuous input variable in assessing wine quality.

```
favstats(wine$free.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean   sd    n missing
##    1 17     29 41 289 30.5 17.7 6497       0
```

total.sulfur.dioxide

total.sulfur.dioxide takes values from 6 to 440. Its unit of measurement is mg/dm3. The mean total.sulfur.dioxide is 116. It is a continuous input variable in assessing wine quality.

```
favstats(wine$total.sulfur.dioxide)
```

```
##  min Q1 median  Q3 max mean   sd    n missing
##    6 77    118 156 440  116 56.5 6497       0
```

density

density takes values from .987 to 1.04. Its unit of measurement is g/cm3. The mean density is 0.995. It is a continuous input variable in assessing wine quality.

```
favstats(wine$density)
```

```
##    min    Q1 median    Q3  max  mean    sd    n missing
##  0.987 0.992  0.995 0.997 1.04 0.995 0.003 6497       0
```

pH

pH takes values form 2.72 to 4.01. The mean pH is 3.22. It is a continuous input variable in assessing wine quality.

```
favstats(wine$pH)
```

```
##   min   Q1 median   Q3  max mean    sd    n missing
##  2.72 3.11   3.21 3.32 4.01 3.22 0.161 6497       0
```

sulphates

sulphates take values from .22 to 2. Its unit of measurement is g(potassium sulphate)/dm3. The mean sulphates is 0.531. It is a continuous input variable in assessing wine quality.

```
favstats(wine$sulphates)
```

```
##   min   Q1 median  Q3 max  mean    sd    n missing
##  0.22 0.43   0.51 0.6   2 0.531 0.149 6497       0
```

alcohol

alcohol takes values from 8 to 14.9. Its unit of measurement is vol.%. The mean alcohol is 10.5 %. It is a continuous input variable in assessing wine quality.

```
favstats(wine$alcohol)
```

```
##  min  Q1 median   Q3  max mean   sd    n missing
##    8 9.5   10.3 11.3 14.9 10.5 1.19 6497       0
```

quality

quality (which is within the [0,10] range) takes value from 3 to 9. It is the value attributed to the quality of wine, 0 being the lowest quality and 10, the highest.

```
tally(~quality, data=wine)
```

```
## quality
##    3    4    5    6    7    8    9
##   30  216 2138 2836 1079  193    5
```

Analysis

Assumptions

Here is our analysis plan for the combined dataset. We first look at a pairsplot and since we have too many data points to observe any possible pattern, we have removed data points to get a big picture.

```
simpall <- wine[-(1:1500),]
simpall2 <- simpall[-(400:6000),]
#pairs(simpall2[-13])
```

Looking at the scatterplot matrix of a subset of the wine data in order to make it less cluttered, we can get an overview of some of the relationships between different variables. Residual sugar and density, for example, have a fairly strong positive correlation.

We will be performing both PCA and Exploratory Factor Analysis on our data set. We will be using PCA as a means of dimension reduction on our 12 numerical variables.

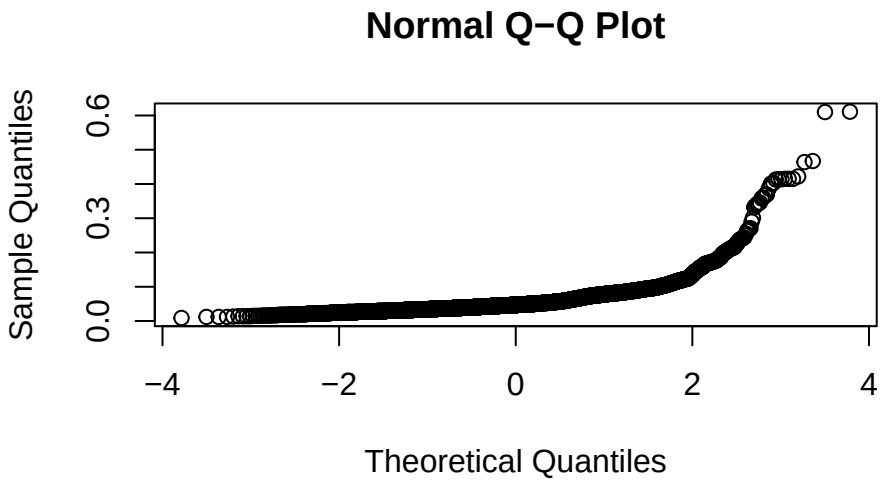
```r
cor(wine[,-13])
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                1.0000           0.2190      0.3244
## volatile.acidity             0.2190           1.0000     -0.3780
## citric.acid                  0.3244          -0.3780      1.0000
## residual.sugar              -0.1120          -0.1960      0.1425
## chlorides                    0.2982           0.3771      0.0390
## free.sulfur.dioxide         -0.2827          -0.3526      0.1331
## total.sulfur.dioxide        -0.3291          -0.4145      0.1952
## density                      0.4589           0.2713      0.0962
## pH                          -0.2527           0.2615     -0.3298
## sulphates                    0.2996           0.2260      0.0562
## alcohol                     -0.0955          -0.0376     -0.0105
## quality                     -0.0767          -0.2657      0.0855
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                 -0.112    0.2982             -0.2827
## volatile.acidity              -0.196    0.3771             -0.3526
## citric.acid                    0.142    0.0390              0.1331
## residual.sugar                 1.000   -0.1289              0.4029
## chlorides                     -0.129    1.0000             -0.1950
## free.sulfur.dioxide            0.403   -0.1950              1.0000
## total.sulfur.dioxide           0.495   -0.2796              0.7209
## density                        0.553    0.3626              0.0257
## pH                            -0.267    0.0447             -0.1459
## sulphates                     -0.186    0.3956             -0.1885
## alcohol                       -0.359   -0.2569             -0.1798
## quality                       -0.037   -0.2007              0.0555
##                       total.sulfur.dioxide density      pH sulphates
## fixed.acidity                      -0.3291  0.4589 -0.2527   0.29957
## volatile.acidity                   -0.4145  0.2713  0.2615   0.22598
## citric.acid                         0.1952  0.0962 -0.3298   0.05620
## residual.sugar                      0.4955  0.5525 -0.2673  -0.18593
## chlorides                          -0.2796  0.3626  0.0447   0.39559
## free.sulfur.dioxide                 0.7209  0.0257 -0.1459  -0.18846
## total.sulfur.dioxide                1.0000  0.0324 -0.2384  -0.27573
## density                             0.0324  1.0000  0.0117   0.25948
## pH                                 -0.2384  0.0117  1.0000   0.19212
## sulphates                          -0.2757  0.2595  0.1921   1.00000
## alcohol                            -0.2657 -0.6867  0.1212  -0.00303
## quality                            -0.0414 -0.3059  0.0195   0.03849
##                        alcohol quality
## fixed.acidity        -0.09545 -0.0767
## volatile.acidity     -0.03764 -0.2657
## citric.acid          -0.01049  0.0855
## residual.sugar       -0.35941 -0.0370
## chlorides            -0.25692 -0.2007
## free.sulfur.dioxide  -0.17984  0.0555
## total.sulfur.dioxide -0.26574 -0.0414
## density              -0.68675 -0.3059
## pH                    0.12125  0.0195
## sulphates            -0.00303  0.0385
## alcohol               1.00000  0.4443
```

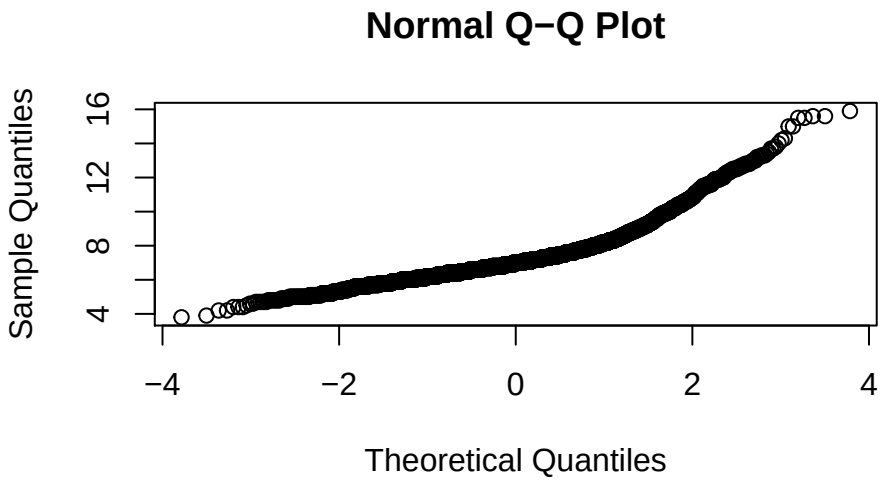
```
## quality               0.44432  1.0000
```

There is some correlation in this data set, suggesting that there is some sort of underlying structure to the data and PCA could be useful in reducing the number of dimensions.

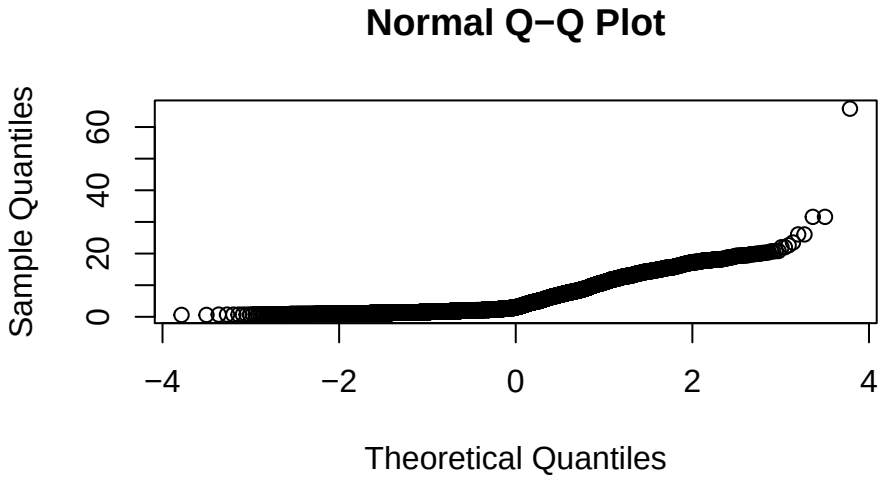
```
qqnorm(wine$chlorides)
```




```
qqnorm(wine$fixed.acidity)
```




```
qqnorm(wine$residual.sugar)
```



Our variables are not univariate normally distributed, so we cannot assume multivariate normality in our data set. Therefore, we will be using Principal Factor Analysis over Maximum Likelihood Factor Analysis.

```
cor(wine[,-13])
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                1.0000           0.2190      0.3244
## volatile.acidity             0.2190           1.0000     -0.3780
## citric.acid                  0.3244          -0.3780      1.0000
## residual.sugar              -0.1120          -0.1960      0.1425
## chlorides                    0.2982           0.3771      0.0390
## free.sulfur.dioxide         -0.2827          -0.3526      0.1331
## total.sulfur.dioxide        -0.3291          -0.4145      0.1952
## density                      0.4589           0.2713      0.0962
## pH                          -0.2527           0.2615     -0.3298
## sulphates                    0.2996           0.2260      0.0562
## alcohol                     -0.0955          -0.0376     -0.0105
## quality                     -0.0767          -0.2657      0.0855
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                 -0.112    0.2982             -0.2827
## volatile.acidity              -0.196    0.3771             -0.3526
## citric.acid                    0.142    0.0390              0.1331
## residual.sugar                 1.000   -0.1289              0.4029
## chlorides                     -0.129    1.0000             -0.1950
## free.sulfur.dioxide            0.403   -0.1950              1.0000
## total.sulfur.dioxide           0.495   -0.2796              0.7209
## density                        0.553    0.3626              0.0257
## pH                            -0.267    0.0447             -0.1459
## sulphates                     -0.186    0.3956             -0.1885
## alcohol                       -0.359   -0.2569             -0.1798
## quality                       -0.037   -0.2007              0.0555
##                       total.sulfur.dioxide density      pH sulphates
## fixed.acidity                      -0.3291  0.4589 -0.2527   0.29957
## volatile.acidity                   -0.4145  0.2713  0.2615   0.22598
## citric.acid                         0.1952  0.0962 -0.3298   0.05620
## residual.sugar                      0.4955  0.5525 -0.2673  -0.18593
## chlorides                          -0.2796  0.3626  0.0447   0.39559
## free.sulfur.dioxide                 0.7209  0.0257 -0.1459  -0.18846
```

```
## total.sulfur.dioxide               1.0000  0.0324 -0.2384  -0.27573
## density                            0.0324  1.0000  0.0117   0.25948
## pH                                -0.2384  0.0117  1.0000   0.19212
## sulphates                         -0.2757  0.2595  0.1921   1.00000
## alcohol                           -0.2657 -0.6867  0.1212  -0.00303
## quality                           -0.0414 -0.3059  0.0195   0.03849
##                         alcohol quality
## fixed.acidity          -0.09545 -0.0767
## volatile.acidity       -0.03764 -0.2657
## citric.acid            -0.01049  0.0855
## residual.sugar         -0.35941 -0.0370
## chlorides              -0.25692 -0.2007
## free.sulfur.dioxide    -0.17984  0.0555
## total.sulfur.dioxide   -0.26574 -0.0414
## density                -0.68675 -0.3059
## pH                      0.12125  0.0195
## sulphates              -0.00303  0.0385
## alcohol                 1.00000  0.4443
## quality                 0.44432  1.0000
```

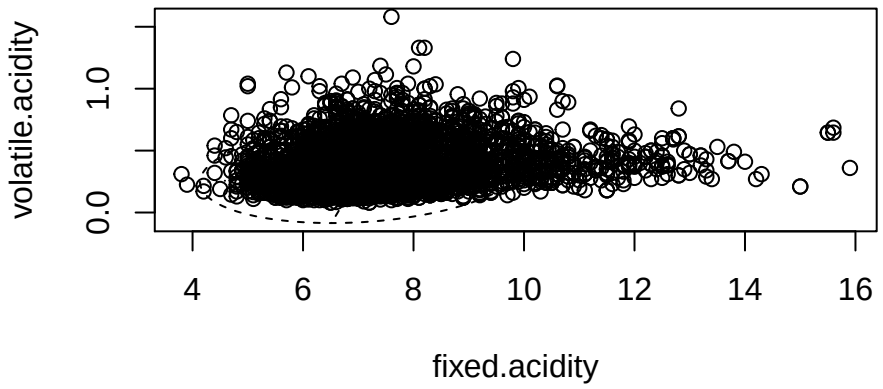
Looking at the correlation matrices again, there is a .72 correlation between "free.sulfur.dioxide" and "total.sulfur.dioxide." This brings up a problem with multicollinearity, so we will only be using "free.sulfur.dioxide" in our cluster analysis.

```
wine2 <- wine[,-13][,-7] # remove type and total.sulfur.dioxide
```

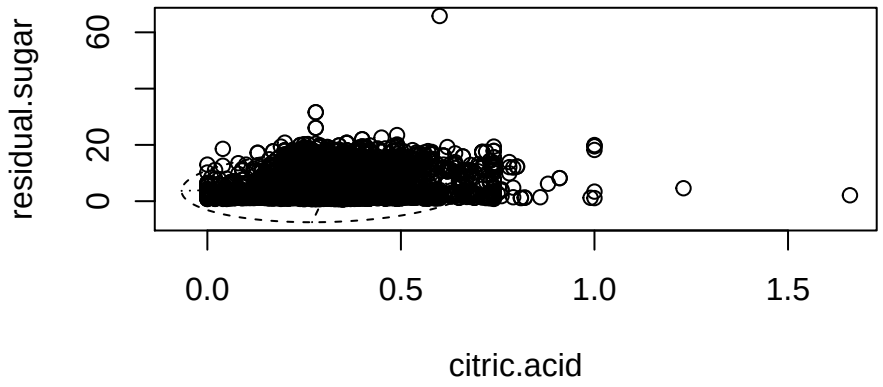
Outlier analysis

Now we look at bivariate boxplots for some pairs of variables. We do this in an attempt to eliminate the clear outliers which will bolster the statistical methods we are using. We still keep some outliers, though, since it is important to keep as many points as possible for cluster analysis.

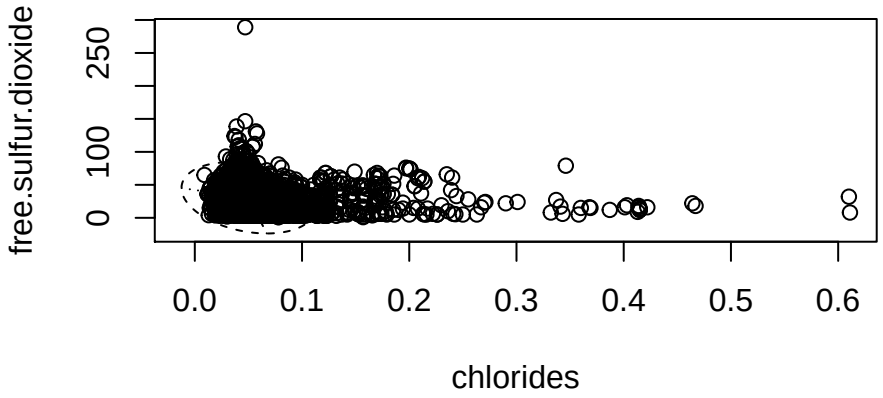
```
#bivariate boxplots for some pairs of variables
#RedWine = `winequality.red.(1)`
twoVars = wine2[c("fixed.acidity", "volatile.acidity")]
bvbox(twoVars, mtitle = "", xlab = "fixed.acidity", ylab = "volatile.acidity")
```



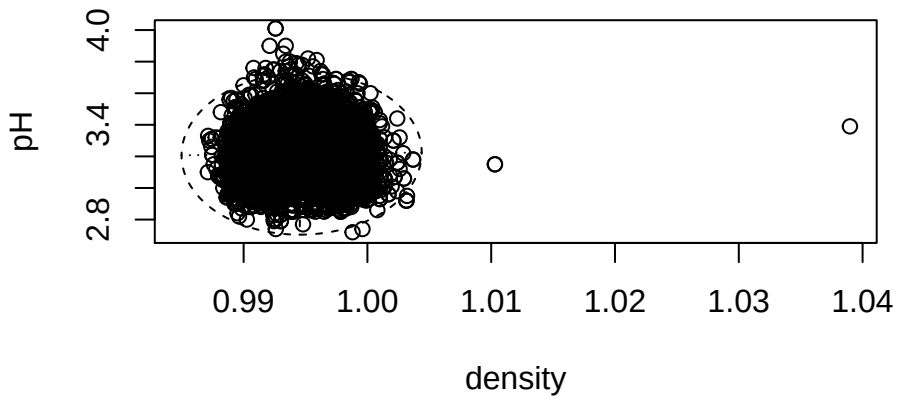

```
twoVars = wine2[c("citric.acid", "residual.sugar")]
bvbox(twoVars, mtitle = "", xlab = "citric.acid", ylab = "residual.sugar")
```



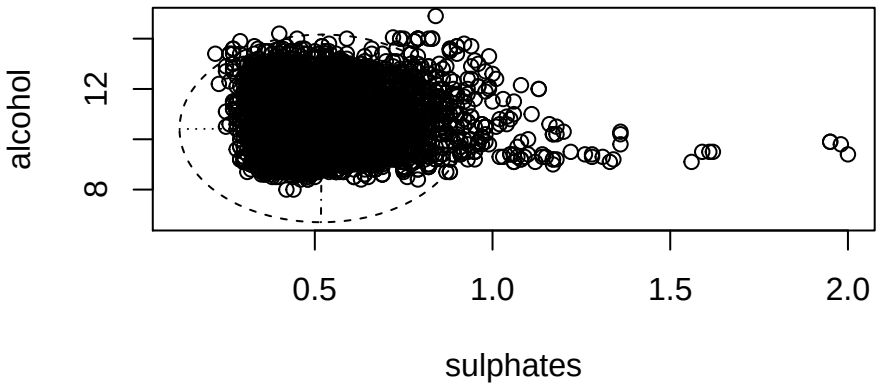

```
twoVars = wine2[c("chlorides", "free.sulfur.dioxide")]
bvbox(twoVars, mtitle = "", xlab = "chlorides", ylab = "free.sulfur.dioxide")
```




```
twoVars = wine2[c("density", "pH")]
bvbox(twoVars, mtitle = "", xlab = "density", ylab = "pH")
```




```
twoVars = wine2[c("sulphates", "alcohol")]
bvbox(twoVars, mtitle = "", xlab = "sulphates", ylab = "alcohol")
```



Given the above boxplots, we eliminate major outliers below. Our new dataset has now 4540 observations.

```
outliers = with(wine2, c(which(volatile.acidity>1.4), which(fixed.acidity>14.5), which(citric.acid>1), w

wine3 <- wine2[-outliers,]

nrow(wine3)
```

```
## [1] 4540
```

Next steps

We will now go ahead and proceed with our PCA, EFA and cluster analysis. To summarize our methods, we are performing PCA to reduce the dimensionality of our dataset to look at wine quality; we are performing EFA to see how much of total variation is captured by our variables and our cluster analysis simply attempts to identify clusters in the data.

EFA

All our variables are numerical except wine type, so we will omit this from our exploratory factor analysis.

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:tourr':
##
##     rescale

## The following objects are masked from 'package:mosaic':
##
##     logit, rescale

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
sapply(1:10, function(f) fa(wine[,-13], nfactors=f,rotate="varimax")$PVAL)
```

```
##  [1]  0.00e+00  0.00e+00  0.00e+00  0.00e+00  0.00e+00 9.28e-264  1.58e-35
##  [8]         NA         NA         NA
```

Now we run the factor analysis on 1-10 factors to see which would work best for our purposes. Unfortunately, it looks as though the P value of any number of factors is below .05. The maximum number of factors according to the degrees of freedom for our data set is 7. Even at 7 factors, the p value of the factor analysis is below .05, which means that we can reject the null hypothesis that the model is a good fit to the data.

The factor analysis also does not yield a p value below .05 for either red wine alone or white wine alone.

PCA

```
cor(wine[,-13])
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                1.0000           0.2190      0.3244
## volatile.acidity             0.2190           1.0000     -0.3780
## citric.acid                  0.3244          -0.3780      1.0000
## residual.sugar              -0.1120          -0.1960      0.1425
## chlorides                    0.2982           0.3771      0.0390
## free.sulfur.dioxide         -0.2827          -0.3526      0.1331
## total.sulfur.dioxide        -0.3291          -0.4145      0.1952
## density                      0.4589           0.2713      0.0962
## pH                          -0.2527           0.2615     -0.3298
## sulphates                    0.2996           0.2260      0.0562
## alcohol                     -0.0955          -0.0376     -0.0105
## quality                     -0.0767          -0.2657      0.0855
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                 -0.112    0.2982             -0.2827
## volatile.acidity              -0.196    0.3771             -0.3526
## citric.acid                    0.142    0.0390              0.1331
## residual.sugar                 1.000   -0.1289              0.4029
## chlorides                     -0.129    1.0000             -0.1950
## free.sulfur.dioxide            0.403   -0.1950              1.0000
## total.sulfur.dioxide           0.495   -0.2796              0.7209
## density                        0.553    0.3626              0.0257
## pH                            -0.267    0.0447             -0.1459
## sulphates                     -0.186    0.3956             -0.1885
## alcohol                       -0.359   -0.2569             -0.1798
## quality                       -0.037   -0.2007              0.0555
##                       total.sulfur.dioxide density      pH sulphates
## fixed.acidity                      -0.3291  0.4589 -0.2527   0.29957
## volatile.acidity                   -0.4145  0.2713  0.2615   0.22598
## citric.acid                         0.1952  0.0962 -0.3298   0.05620
## residual.sugar                      0.4955  0.5525 -0.2673  -0.18593
## chlorides                          -0.2796  0.3626  0.0447   0.39559
## free.sulfur.dioxide                 0.7209  0.0257 -0.1459  -0.18846
## total.sulfur.dioxide                1.0000  0.0324 -0.2384  -0.27573
## density                             0.0324  1.0000  0.0117   0.25948
```

```
## pH                                -0.2384  0.0117  1.0000   0.19212
## sulphates                         -0.2757  0.2595  0.1921   1.00000
## alcohol                           -0.2657 -0.6867  0.1212  -0.00303
## quality                           -0.0414 -0.3059  0.0195   0.03849
##                         alcohol quality
## fixed.acidity          -0.09545 -0.0767
## volatile.acidity       -0.03764 -0.2657
## citric.acid            -0.01049  0.0855
## residual.sugar         -0.35941 -0.0370
## chlorides              -0.25692 -0.2007
## free.sulfur.dioxide    -0.17984  0.0555
## total.sulfur.dioxide   -0.26574 -0.0414
## density                -0.68675 -0.3059
## pH                      0.12125  0.0195
## sulphates              -0.00303  0.0385
## alcohol                 1.00000  0.4443
## quality                 0.44432  1.0000
```

Looking at the correlations between variables in our data, it looks like there are a number of correlations above .3, indicating that there is some sort of underlying structure in the data and fulfills the assumptions required for principal components analysis.

```
PCWine=prcomp(wine[,-13],scale=TRUE)
PCWine$rotation
```

```
##                            PC1    PC2     PC3     PC4      PC5     PC6
## fixed.acidity          -0.2569  0.262 -0.4675  0.1440 -0.16536  0.0300
## volatile.acidity       -0.3949  0.105  0.2797  0.0801 -0.14777 -0.3827
## citric.acid             0.1465  0.144 -0.5881 -0.0555  0.23462  0.3622
## residual.sugar          0.3189  0.343  0.0755 -0.1125 -0.50792 -0.0633
## chlorides              -0.3134  0.270 -0.0468 -0.1653  0.39390 -0.4254
## free.sulfur.dioxide     0.4227  0.111  0.0990 -0.3033  0.24845 -0.2832
## total.sulfur.dioxide    0.4744  0.144  0.1013 -0.1322  0.22397 -0.1068
## density                -0.0924  0.555  0.0516 -0.1506 -0.33036  0.1546
## pH                     -0.2081 -0.153  0.4068 -0.4715  0.00146  0.5609
## sulphates              -0.2999  0.120 -0.1687 -0.5880  0.19325 -0.0201
## alcohol                -0.0589 -0.493 -0.2129 -0.0800 -0.11602 -0.1695
## quality                 0.0875 -0.297 -0.2958 -0.4724 -0.45913 -0.2779
##                            PC7      PC8      PC9    PC10     PC11      PC12
## fixed.acidity          -0.3934  0.00116  0.42417 -0.2724 -0.27693 -0.335093
## volatile.acidity       -0.4451  0.31008 -0.12323  0.4939  0.14080 -0.082421
## citric.acid            -0.0477  0.44496 -0.24623  0.3304  0.22928  0.001347
## residual.sugar          0.0958  0.08194 -0.48802 -0.2072  0.00514 -0.451215
## chlorides               0.4733  0.37553 -0.04405 -0.2389 -0.19340 -0.043278
## free.sulfur.dioxide    -0.3627  0.12010  0.30140 -0.3034  0.48616 -0.000905
## total.sulfur.dioxide   -0.2348  0.01128  0.00181  0.2948 -0.72016  0.064063
## density                -0.0133  0.04294  0.07108 -0.0768 -0.00332  0.715667
## pH                     -0.0793  0.36228  0.13666 -0.1124 -0.13908 -0.206763
## sulphates              -0.1702 -0.59222 -0.29740  0.0855  0.04722 -0.078200
## alcohol                -0.3389  0.22604 -0.41706 -0.4161 -0.19129  0.332012
## quality                 0.2732  0.09305  0.35665  0.3078 -0.01808  0.008288
```

We run our PCA on correlations over covariances because our variables are on different scales and we do not want to weight variables with higher covariances differently.

```r
summary(PCWine) #prop variance explained by each component
```

```
## Importance of components:
##                           PC1   PC2   PC3    PC4   PC5   PC6   PC7   PC8
## Standard deviation     1.744 1.628 1.281 1.0337 0.917 0.813 0.751 0.718
## Proportion of Variance 0.253 0.221 0.137 0.0891 0.070 0.055 0.047 0.043
## Cumulative Proportion  0.253 0.474 0.611 0.7001 0.770 0.825 0.872 0.915
##                           PC9   PC10  PC11    PC12
## Standard deviation     0.6770 0.5468 0.477 0.18107
## Proportion of Variance 0.0382 0.0249 0.019 0.00273
## Cumulative Proportion  0.9534 0.9783 0.997 1.00000
```

Looking at the proportion of variance explained by each principal component

```r
cor(wine[,-13],PCWine$x) #loadings
```

```
##                          PC1    PC2     PC3     PC4      PC5     PC6
## fixed.acidity        -0.448  0.426 -0.5989  0.1488 -0.15160  0.0244
## volatile.acidity     -0.689  0.171  0.3583  0.0828 -0.13548 -0.3110
## citric.acid           0.255  0.235 -0.7535 -0.0574  0.21510  0.2944
## residual.sugar        0.556  0.558  0.0967 -0.1163 -0.46566 -0.0515
## chlorides            -0.547  0.439 -0.0599 -0.1709  0.36112 -0.3457
## free.sulfur.dioxide   0.737  0.181  0.1268 -0.3135  0.22778 -0.2301
## total.sulfur.dioxide  0.827  0.234  0.1298 -0.1367  0.20533 -0.0868
## density              -0.161  0.903  0.0661 -0.1557 -0.30287  0.1256
## pH                   -0.363 -0.249  0.5212 -0.4874  0.00134  0.4558
## sulphates            -0.523  0.195 -0.2161 -0.6079  0.17717 -0.0164
## alcohol              -0.103 -0.802 -0.2728 -0.0827 -0.10637 -0.1377
## quality               0.153 -0.483 -0.3790 -0.4884 -0.42092 -0.2258
##                          PC7      PC8      PC9    PC10     PC11      PC12
## fixed.acidity        -0.29542  0.00083  0.28718 -0.1490 -0.13211 -0.060674
## volatile.acidity     -0.33423  0.22273 -0.08343  0.2701  0.06717 -0.014924
## citric.acid          -0.03582  0.31963 -0.16671  0.1806  0.10938  0.000244
## residual.sugar        0.07191  0.05886 -0.33041 -0.1133  0.00245 -0.081700
## chlorides             0.35539  0.26975 -0.02982 -0.1306 -0.09226 -0.007836
## free.sulfur.dioxide  -0.27236  0.08627  0.20406 -0.1659  0.23193 -0.000164
## total.sulfur.dioxide -0.17632  0.00810  0.00123  0.1612 -0.34356  0.011600
## density              -0.00998  0.03085  0.04812 -0.0420 -0.00159  0.129583
## pH                   -0.05956  0.26023  0.09252 -0.0615 -0.06635 -0.037438
## sulphates            -0.12783 -0.42540 -0.20135  0.0467  0.02253 -0.014159
## alcohol              -0.25448  0.16237 -0.28236 -0.2275 -0.09126  0.060116
## quality               0.20512  0.06684  0.24146  0.1683 -0.00863  0.001501
```

Cluster Analysis

Now we perform cluster analysis on our wine3 data set. We examine the correlation between the varibles to eliminate any highly correlated variables.

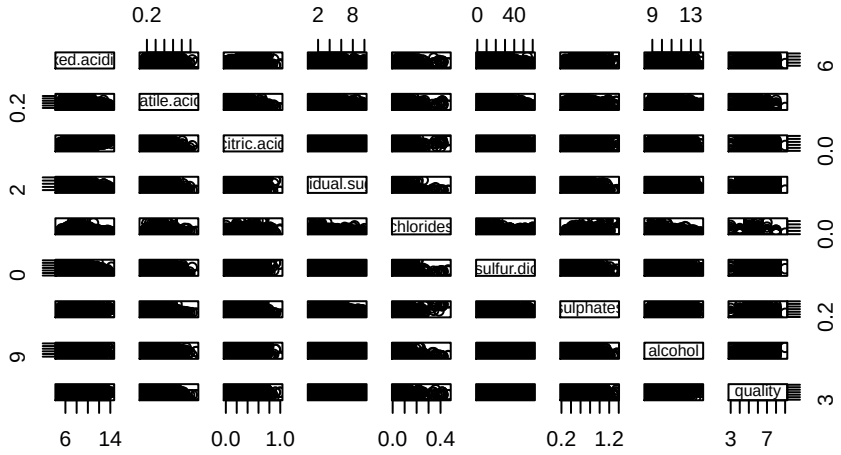
```r
cor(wine3)
```

```
##                       fixed.acidity volatile.acidity citric.acid
```

```
## fixed.acidity              1.0000            0.200      0.3761
## volatile.acidity           0.2003            1.000     -0.4387
## citric.acid                0.3761           -0.439      1.0000
## residual.sugar            -0.1675           -0.170      0.0784
## chlorides                  0.2961            0.397      0.0314
## free.sulfur.dioxide       -0.3355           -0.391      0.0923
## density                    0.5836            0.472      0.0118
## pH                        -0.2938            0.267     -0.3547
## sulphates                  0.3066            0.219      0.0632
## alcohol                   -0.0325           -0.105      0.0988
## quality                   -0.0294           -0.283      0.1561
##                      residual.sugar chlorides free.sulfur.dioxide density
## fixed.acidity               -0.1675    0.2961             -0.3355  0.5836
## volatile.acidity            -0.1698    0.3969             -0.3910  0.4715
## citric.acid                  0.0784    0.0314              0.0923  0.0118
## residual.sugar               1.0000   -0.1966              0.3742  0.1225
## chlorides                   -0.1966    1.0000             -0.2622  0.4908
## free.sulfur.dioxide          0.3742   -0.2622              1.0000 -0.2548
## density                      0.1225    0.4908             -0.2548  1.0000
## pH                          -0.1855    0.0545             -0.0991  0.1659
## sulphates                   -0.2369    0.4037             -0.2463  0.3925
## alcohol                     -0.0187   -0.2626             -0.0563 -0.5058
## quality                      0.0535   -0.1795              0.1382 -0.2839
##                           pH sulphates alcohol quality
## fixed.acidity       -0.2938    0.3066 -0.0325 -0.0294
## volatile.acidity     0.2673    0.2191 -0.1049 -0.2827
## citric.acid         -0.3547    0.0632  0.0988  0.1561
## residual.sugar      -0.1855   -0.2369 -0.0187  0.0535
## chlorides            0.0545    0.4037 -0.2626 -0.1795
## free.sulfur.dioxide -0.0991   -0.2463 -0.0563  0.1382
## density              0.1659    0.3925 -0.5058 -0.2839
## pH                   1.0000    0.2120  0.0615  0.0269
## sulphates            0.2120    1.0000  0.0435  0.0862
## alcohol              0.0615    0.0435  1.0000  0.4613
## quality              0.0269    0.0862  0.4613  1.0000
```

There are only two pairs of highly correlate variables (>0.7): free.sulfur.dioxide and total.sulfur.dioxid, alcohol and density. We will choose to eliminate total.sulfur.dioxide and density.

```
wine4 <- wine3[-c(7,8)]
pairs(wine4)
```



It looks like there are no outliers present in our data.

Now we need to determine whether to standarize our variables.

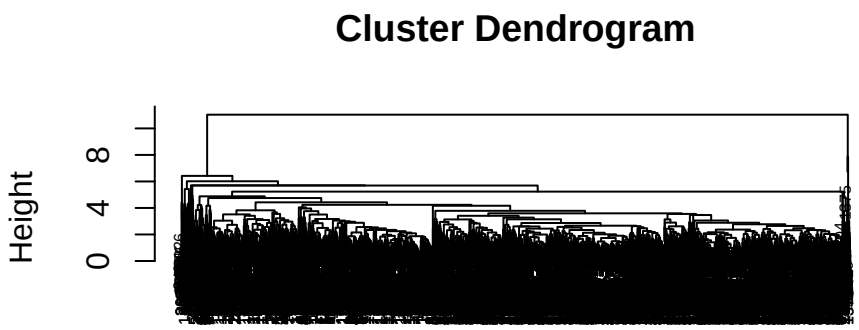
```
summary(wine4)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.080    Min.   :0.000   Min.   : 0.60
##  1st Qu.: 6.50   1st Qu.:0.230    1st Qu.:0.230   1st Qu.: 1.70
##  Median : 7.10   Median :0.310    Median :0.310   Median : 2.40
##  Mean   : 7.39   Mean   :0.362    Mean   :0.308   Mean   : 3.65
##  3rd Qu.: 7.90   3rd Qu.:0.450    3rd Qu.:0.390   3rd Qu.: 5.50
##  Max.   :14.30   Max.   :1.330    Max.   :1.000   Max.   :10.00
##    chlorides     free.sulfur.dioxide   sulphates        alcohol
##  Min.   :0.012   Min.   : 1.0        Min.   :0.220   Min.   : 8.4
##  1st Qu.:0.039   1st Qu.:14.0        1st Qu.:0.450   1st Qu.: 9.6
##  Median :0.050   Median :24.0        Median :0.530   Median :10.4
##  Mean   :0.060   Mean   :25.6        Mean   :0.549   Mean   :10.5
##  3rd Qu.:0.074   3rd Qu.:36.0        3rd Qu.:0.620   3rd Qu.:11.2
##  Max.   :0.467   Max.   :60.0        Max.   :1.360   Max.   :14.1
##     quality
##  Min.   :3.00
##  1st Qu.:5.00
##  Median :6.00
##  Mean   :5.77
##  3rd Qu.:6.00
##  Max.   :9.00
```

Our summary reveals that our variables have very different scales and should be standarized.

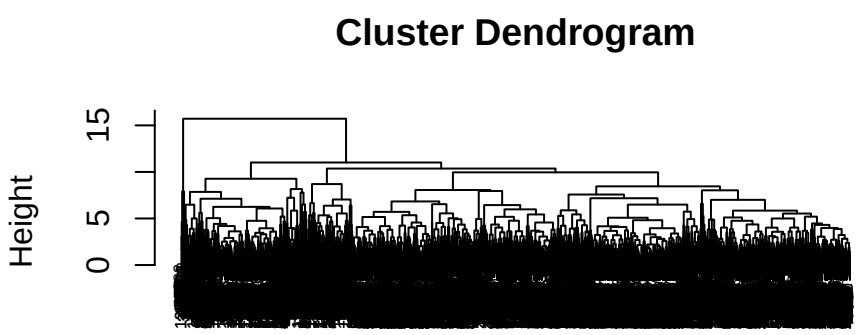
```
WinesScaled = dist(scale(wine4))
```

```
avg = hclust (WinesScaled, method = "average")
plot(avg, cex=0.5)
```



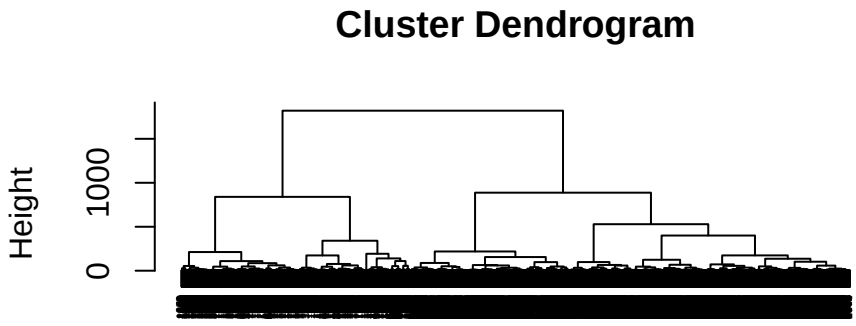
The average solution is very bad so we try another one

```
comp = hclust (WinesScaled, method = "complete")
plot(comp, cex=0.5)
```



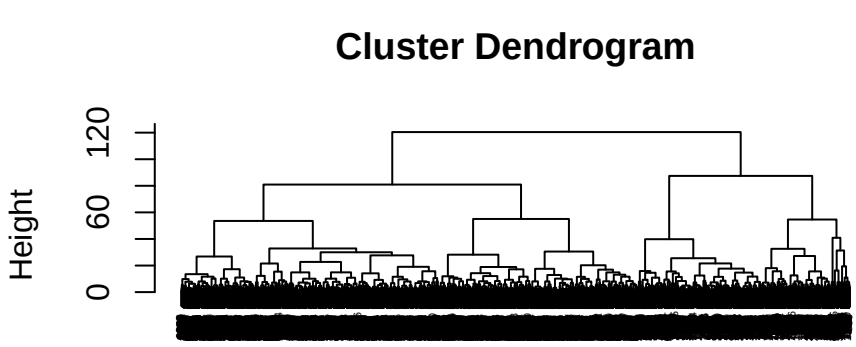
The complete is also very bad. Let's try ward.D

```
ward = hclust (WinesScaled, method = "ward.D")
plot(ward, cex= 0.3)
```



The ward.D solution is much better, let's see if the D2 solution is best

```
ward2 = hclust (WinesScaled, method = "ward.D2")
plot(ward2, cex=0.3)
```



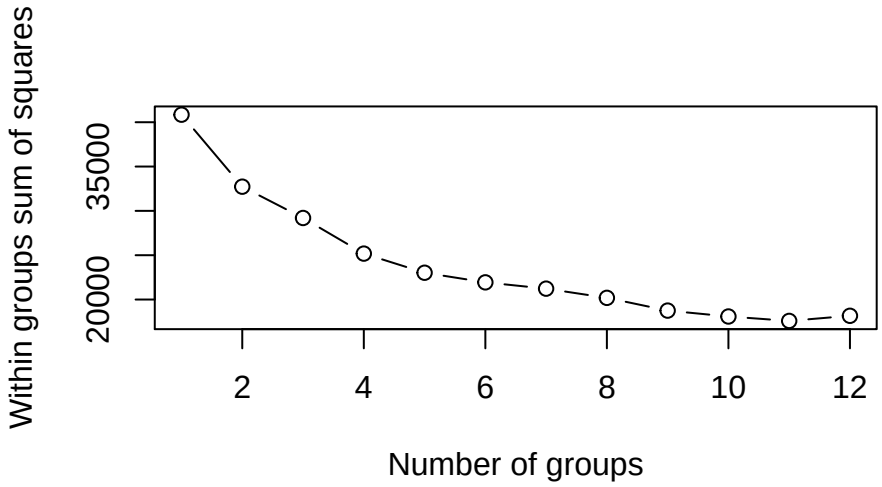
We liked the ward.D solution best because it gives clear cutoffs and is widely spread among the data points. We will cut the ward solution

```
wardSol = cutree(ward, k = 5)
summary(as.factor(wardSol))
```

```
##    1    2    3    4    5
##  799 1471 1080  753  437
```

Before we decide on the final number of clusters, we will carry on a K-means solution.

```
set.seed(100)
wss = rep(0, 12)
for(i in 1:12){ wss[i] <- sum(kmeans(scale(wine4),centers= i)$withinss)}
plot(1:12, wss, type = "b", xlab = "Number of groups", ylab = "Within groups sum of squares")
```



An elbow happens at around k = 5, so our cluster group number is about right. We stick with 5 clusters.

Now we want to see the overlap of both the k-means and the hierarchical solutions.

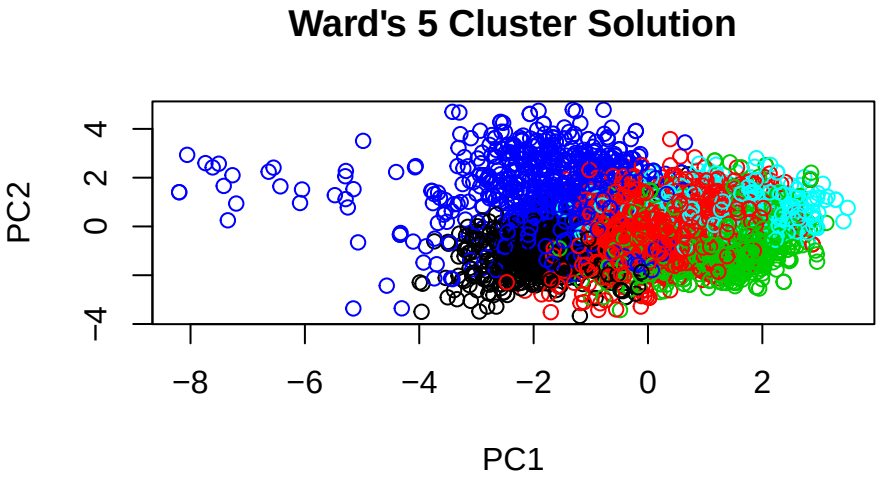
```
kSol = kmeans(scale(wine4), centers =5)
tally(kSol$cluster~wardSol, format = "count")
```

```
##              wardSol
## kSol$cluster   1   2   3   4   5
##            1 734  65  25  92   3
##            2   9  20   4 514   0
##            3  27 461 108  53 379
##            4  29 902  90  68  15
##            5   0  23 853  26  40
```

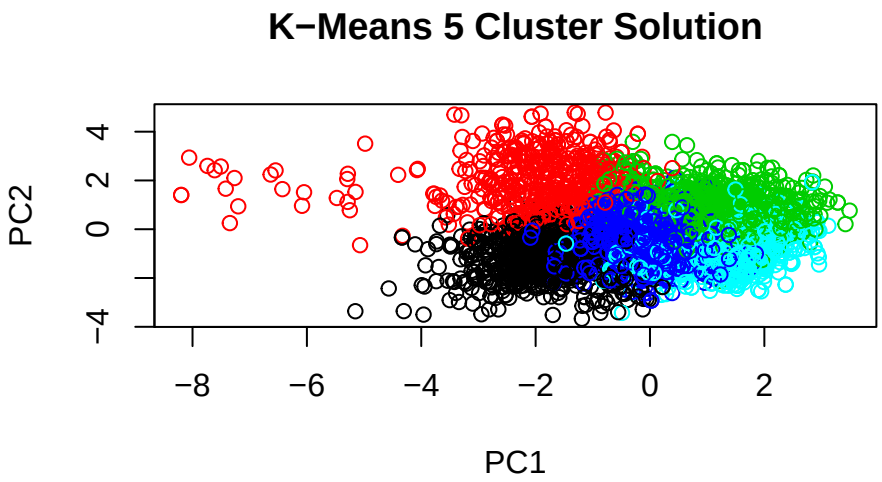
We see some overlap in our solutions. Cluster 5 from wardSol overlaps mostly with KSol cluster 1, wardSol cluster 4 with kSol cluster 2 and 3, wardSol 3 with kSol3, wardSol2 with kSol 2, and wardSol 1 with kSol5. The trend is not perfect but it is clear.

To make the decision between the two solutions, we will plot them.

```
wardPCA = prcomp(wine4, scale = TRUE)
plot(wardPCA$x[,1:2], xlab = "PC1", ylab = "PC2", main = "Ward's 5 Cluster Solution", col= wardSol)
```



```
plot(wardPCA$x[,1:2], xlab = "PC1", ylab = "PC2", main = "K-Means 5 Cluster Solution", col= kSol$cluste
```



Our K-Means 5 Cluster solution has a clearer distinction between the clusters. Therefore, we will proceed with K-Means 5-Cluster solution in my analysis.

Expository components

Overview

So we are starting from our original combined dataset and we remove the 'total.sulfur.dioxide' variable following our correlation analysis and we remove the outliers followign our outlier analysis. We are randomly splitting the data by type (i.e. red and white wine) to have a representative sample in each set. Our training set will have 70% of our oriniginal data and our test set will have the remaining 30% of the data. We will use cross-validation. Cross-validation is a model validation technique to assess how the results of a statistical analysis will generalize to an independent data set.

```
wine4 <- wine[,-7] # remove total.sulfur.dioxide

outliers = with(wine4, c(which(volatile.acidity>1.4), which(fixed.acidity>14.5), which(citric.acid>1), w

wine5 <- wine4[-outliers,]

nrow(wine5)
```

```
## [1] 4540
```

```
set.seed(240)
split <-  sample.split(wine5$type, SplitRatio = 0.70)
train <- subset(wine5, split == TRUE)
test <- subset(wine5, split == FALSE)
```