

CMTH642 Assignment 3

Paul Ycay

25/07/2019

The following document was used to supplement this project

<https://rpubs.com/shradhit/winequality> The RMD file for Lab 10 and Lab 10 solutions was used to supplement this assignment as well Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

1. Check data characteristics. Is there missing data?

```
wine<-read.csv(file="http://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-white.csv", header = TRUE, sep= ";");
str(wine)

## 'data.frame': 4898 obs. of 12 variables:
## $ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
## $ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3
0.22 ...
## $ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34
0.43 ...
## $ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
## $ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045
0.045 0.049 0.044 ...
## $ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...
## $ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...
## $ density : num 1.001 0.994 0.995 0.996 0.996 ...
## $ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22
...
## $ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49
0.45 ...
## $ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

head(wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 7.0 0.27 0.36 20.7 0.045
## 2 6.3 0.30 0.34 1.6 0.049
## 3 8.1 0.28 0.40 6.9 0.050
## 4 7.2 0.23 0.32 8.5 0.058
## 5 7.2 0.23 0.32 8.5 0.058
## 6 8.1 0.28 0.40 6.9 0.050
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol
## 1 45 170 1.0010 3.00 0.45 8.8
## 2 14 132 0.9940 3.30 0.49 9.5
## 3 30 97 0.9951 3.26 0.44 10.1
```

```
## 4          47          186 0.9956 3.19          0.40          9.9
## 5          47          186 0.9956 3.19          0.40          9.9
## 6          30          97 0.9951 3.26          0.44         10.1
##  quality
## 1          6
## 2          6
## 3          6
## 4          6
## 5          6
## 6          6
```

tail(wine)

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 4893          6.5          0.23          0.38          1.3          0.032
## 4894          6.2          0.21          0.29          1.6          0.039
## 4895          6.6          0.32          0.36          8.0          0.047
## 4896          6.5          0.24          0.19          1.2          0.041
## 4897          5.5          0.29          0.30          1.1          0.022
## 4898          6.0          0.21          0.38          0.8          0.020
##      free.sulfur.dioxide total.sulfur.dioxide density    pH sulphates
## 4893                29                112 0.99298 3.29          0.54
## 4894                24                92 0.99114 3.27          0.50
## 4895                57                168 0.99490 3.15          0.46
## 4896                30                111 0.99254 2.99          0.46
## 4897                20                110 0.98869 3.34          0.38
## 4898                22                98 0.98941 3.26          0.32
##      alcohol quality
## 4893          9.7          5
## 4894         11.2          6
## 4895          9.6          5
## 4896          9.4          6
## 4897         12.8          7
## 4898         11.8          6
```

summary(wine)

```
##      fixed.acidity    volatile.acidity    citric.acid    residual.sugar
## Min.   : 3.800    Min.   :0.0800    Min.   :0.0000    Min.   : 0.600
## 1st Qu.: 6.300    1st Qu.:0.2100    1st Qu.:0.2700    1st Qu.: 1.700
## Median : 6.800    Median :0.2600    Median :0.3200    Median : 5.200
## Mean   : 6.855    Mean   :0.2782    Mean   :0.3342    Mean   : 6.391
## 3rd Qu.: 7.300    3rd Qu.:0.3200    3rd Qu.:0.3900    3rd Qu.: 9.900
## Max.   :14.200    Max.   :1.1000    Max.   :1.6600    Max.   :65.800
##      chlorides      free.sulfur.dioxide total.sulfur.dioxide
## Min.   :0.00900    Min.   : 2.00    Min.   : 9.0
## 1st Qu.:0.03600    1st Qu.: 23.00    1st Qu.:108.0
## Median :0.04300    Median : 34.00    Median :134.0
## Mean   :0.04577    Mean   : 35.31    Mean   :138.4
## 3rd Qu.:0.05000    3rd Qu.: 46.00    3rd Qu.:167.0
## Max.   :0.34600    Max.   :289.00    Max.   :440.0
```

```
##      density          pH      sulphates      alcohol
## Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.    : 8.00
## 1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
## Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
## Mean   :0.9940   Mean    :3.188   Mean    :0.4898   Mean    :10.51
## 3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
## Max.    :1.0390   Max.    :3.820   Max.    :1.0800   Max.    :14.20
##      quality
## Min.    :3.000
## 1st Qu.:5.000
## Median :6.000
## Mean    :5.878
## 3rd Qu.:6.000
## Max.    :9.000

sum(is.na(wine))

## [1] 0
```

There is no missing data

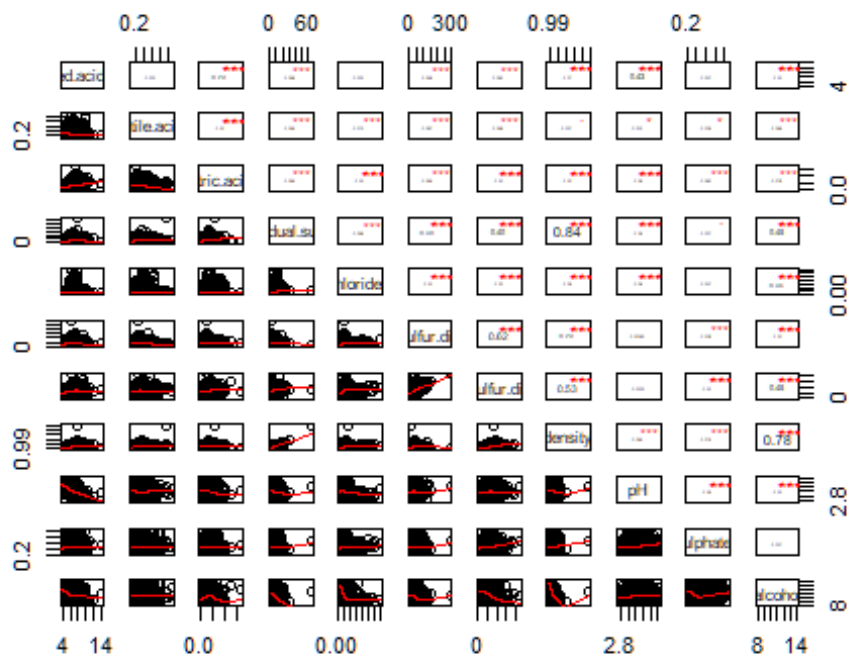
2.What is the correlation between the attributes other than wine quality?

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

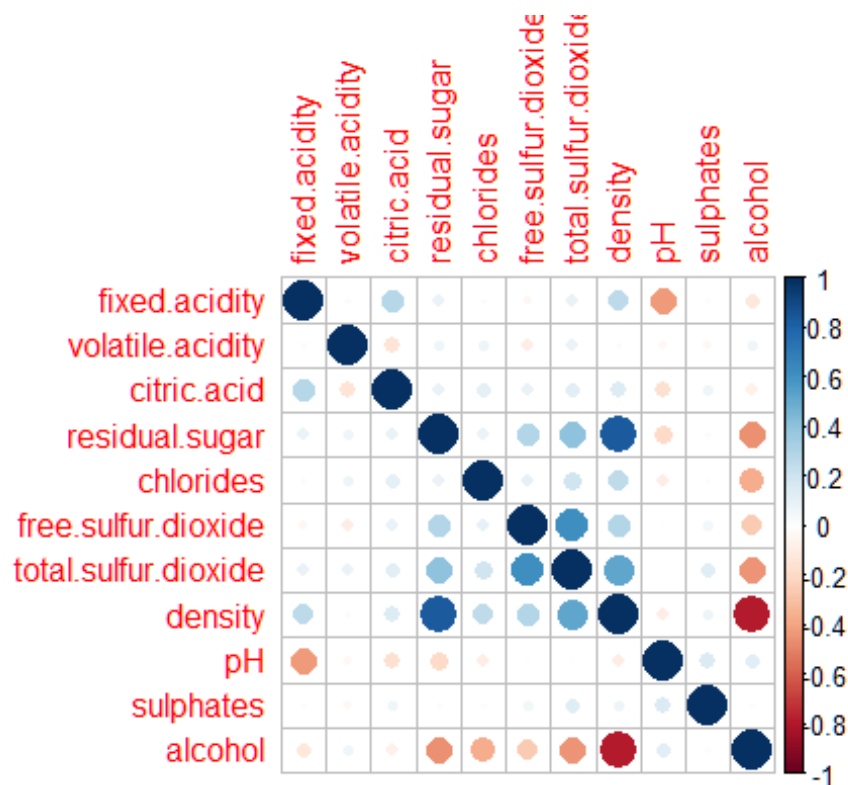
  test <- cor.test(x,y)
  # borrowed from printCoefmat
  Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
    cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
    symbols = c("***", "**", "*", ".", " "))

  text(0.5, 0.5, txt, cex = cex * r)
  text(.8, .8, Signif, cex=cex, col=2)
}

wine_cor<-subset(wine, select=c(1:11))
pairs(wine_cor, lower.panel=panel.smooth, upper.panel=panel.cor)
```

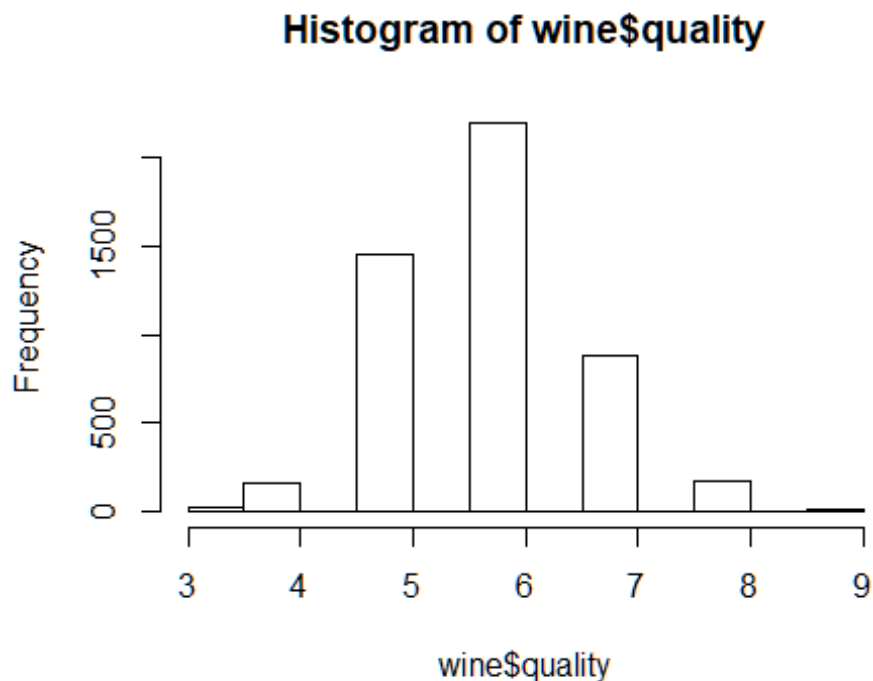


```
corrplot(cor(wine_cor))
```



3. Graph the frequency distribution of wine quality.

```
hist(wine$quality)
```



4. Reduce the levels of rating for quality to three levels as high, medium and low.

```
wine$quality = ifelse(wine$quality < 5, 'low', ifelse(wine$quality > 7,
'high', 'medium'))
wine$quality = ordered(wine$quality, c('low', 'medium', 'high'))
round(prop.table(table(wine$quality)) * 100, digits = 1)
```

```
##
##   low medium   high
##   3.7   92.6    3.7
```

```
head(wine$quality)
```

```
## [1] medium medium medium medium medium medium
## Levels: low < medium < high
```

```
tail(wine$quality)
```

```
## [1] medium medium medium medium medium medium
## Levels: low < medium < high
```

```
summary(wine$quality)
```

```
##   low medium   high
##   183   4535    180
```

5. Normalize the data set.

```
normalize <- function(x) {  
  return ((x - min(x)) / (max(x) - min(x))) }  
wine_n <- as.data.frame(lapply(wine[-12],normalize))  
wine_n <- cbind(wine_n,wine$quality)  
head(wine_n)  
  
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1    0.3076923    0.1862745    0.2168675    0.30828221 0.1068249  
## 2    0.2403846    0.2156863    0.2048193    0.01533742 0.1186944  
## 3    0.4134615    0.1960784    0.2409639    0.09662577 0.1216617  
## 4    0.3269231    0.1470588    0.1927711    0.12116564 0.1454006  
## 5    0.3269231    0.1470588    0.1927711    0.12116564 0.1454006  
## 6    0.4134615    0.1960784    0.2409639    0.09662577 0.1216617  
##   free.sulfur.dioxide total.sulfur.dioxide   density      pH sulphates  
## 1      0.14982578              0.3735499 0.2677848 0.2545455 0.2674419  
## 2      0.04181185              0.2853828 0.1328321 0.5272727 0.3139535  
## 3      0.09756098              0.2041763 0.1540389 0.4909091 0.2558140  
## 4      0.15679443              0.4106729 0.1636784 0.4272727 0.2093023  
## 5      0.15679443              0.4106729 0.1636784 0.4272727 0.2093023  
## 6      0.09756098              0.2041763 0.1540389 0.4909091 0.2558140  
##   alcohol wine$quality  
## 1 0.1290323      medium  
## 2 0.2419355      medium  
## 3 0.3387097      medium  
## 4 0.3064516      medium  
## 5 0.3064516      medium  
## 6 0.3387097      medium  
  
tail(wine_n)  
  
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 4893    0.2596154    0.1470588    0.2289157    0.010736196 0.06824926  
## 4894    0.2307692    0.1274510    0.1746988    0.015337423 0.08902077  
## 4895    0.2692308    0.2352941    0.2168675    0.113496933 0.11275964  
## 4896    0.2596154    0.1568627    0.1144578    0.009202454 0.09495549  
## 4897    0.1634615    0.2058824    0.1807229    0.007668712 0.03857567  
## 4898    0.2115385    0.1274510    0.2289157    0.003067485 0.03264095  
##   free.sulfur.dioxide total.sulfur.dioxide   density      pH  
## 4893      0.09407666              0.2389791 0.11316753 0.5181818  
## 4894      0.07665505              0.1925754 0.07769424 0.5000000  
## 4895      0.19163763              0.3689095 0.15018315 0.3909091  
## 4896      0.09756098              0.2366589 0.10468479 0.2454545  
## 4897      0.06271777              0.2343387 0.03046077 0.5636364  
## 4898      0.06968641              0.2064965 0.04434162 0.4909091  
##   sulphates alcohol wine$quality  
## 4893 0.3720930 0.2741935      medium  
## 4894 0.3255814 0.5161290      medium  
## 4895 0.2790698 0.2580645      medium  
## 4896 0.2790698 0.2258065      medium
```

```
## 4897 0.1860465 0.7741935      medium
## 4898 0.1162791 0.6129032      medium

summary(wine_n)

## fixed.acidity    volatile.acidity    citric.acid    residual.sugar
## Min.      :0.0000    Min.      :0.0000    Min.      :0.0000    Min.      :0.00000
## 1st Qu.:0.2404    1st Qu.:0.1275    1st Qu.:0.1627    1st Qu.:0.01687
## Median :0.2885    Median :0.1765    Median :0.1928    Median :0.07055
## Mean   :0.2937    Mean   :0.1944    Mean   :0.2013    Mean   :0.08883
## 3rd Qu.:0.3365    3rd Qu.:0.2353    3rd Qu.:0.2349    3rd Qu.:0.14264
## Max.   :1.0000    Max.   :1.0000    Max.   :1.0000    Max.   :1.00000
## chlorides      free.sulfur.dioxide    total.sulfur.dioxide
## Min.      :0.00000    Min.      :0.00000    Min.      :0.0000
## 1st Qu.:0.08012    1st Qu.:0.07317    1st Qu.:0.2297
## Median :0.10089    Median :0.11150    Median :0.2900
## Mean   :0.10912    Mean   :0.11606    Mean   :0.3001
## 3rd Qu.:0.12166    3rd Qu.:0.15331    3rd Qu.:0.3666
## Max.   :1.00000    Max.   :1.00000    Max.   :1.0000
## density        pH          sulphates        alcohol
## Min.      :0.00000    Min.      :0.0000    Min.      :0.0000    Min.      :0.0000
## 1st Qu.:0.08892    1st Qu.:0.3364    1st Qu.:0.2209    1st Qu.:0.2419
## Median :0.12782    Median :0.4182    Median :0.2907    Median :0.3871
## Mean   :0.13336    Mean   :0.4257    Mean   :0.3138    Mean   :0.4055
## 3rd Qu.:0.17332    3rd Qu.:0.5091    3rd Qu.:0.3837    3rd Qu.:0.5484
## Max.   :1.00000    Max.   :1.0000    Max.   :1.0000    Max.   :1.0000
## wine$quality
## low      : 183
## medium:4535
## high     : 180
##
##
##
```

6. Divide the data to training and testing groups.

```
set.seed(1)
index <- sample(1:nrow(wine_n), 0.65 *nrow(wine_n))
wine_train <- wine_n[index,]
wine_test  <- wine_n[-index,]
wine_train_labels <- wine_train[,12]
wine_test_labels  <- wine_test[,12]
summary(wine_train_labels)

##    low medium    high
##   124   2925    134

summary(wine_test_labels)

##    low medium    high
##    59   1610     46
```

7. Use the KNN algorithm to predict quality of wine using its attributes

```
wine_test_pred <- knn(train = wine_train[,1:11], test = wine_test[,1:11],cl =  
wine_train[,1], k=10)  
head(wine_test_pred)
```

```
## [1] 0.336538461538462 0.326923076923077 0.365384615384615  
0.394230769230769  
## [5] 0.442307692307692 0.25  
## 64 Levels: 0 0.00961538461538462 0.0384615384615385 ... 1
```

```
tail(wine_test_pred)
```

```
## [1] 0.221153846153846 0.25 0.259615384615385  
0.182692307692308  
## [5] 0.269230769230769 0.298076923076923  
## 64 Levels: 0 0.00961538461538462 0.0384615384615385 ... 1
```

```
summary(wine_test_pred)
```

```
##          0 0.00961538461538462 0.0384615384615385  
##          0 0 0  
## 0.0576923076923078 0.0865384615384616 0.0961538461538462  
##          0 0 0  
## 0.105769230769231 0.115384615384615 0.125  
##          0 3 4  
## 0.134615384615385 0.144230769230769 0.153846153846154  
##          3 5 8  
## 0.163461538461538 0.173076923076923 0.182692307692308  
##          6 12 24  
## 0.192307692307692 0.201923076923077 0.211538461538462  
##          30 22 49  
## 0.221153846153846 0.225961538461539 0.230769230769231  
##          51 0 92  
## 0.240384615384615 0.25 0.259615384615385  
##          62 105 74  
## 0.269230769230769 0.278846153846154 0.288461538461538  
##          155 121 118  
## 0.298076923076923 0.307692307692308 0.317307692307692  
##          102 88 66  
## 0.322115384615385 0.326923076923077 0.336538461538462  
##          0 94 88  
## 0.346153846153846 0.355769230769231 0.365384615384615  
##          72 48 34  
##          0.375 0.384615384615385 0.394230769230769  
##          35 26 16  
## 0.403846153846154 0.413461538461538 0.423076923076923  
##          15 12 15  
## 0.432692307692308 0.442307692307692 0.451923076923077  
##          11 9 7  
## 0.461538461538462 0.471153846153846 0.480769230769231  
##          7 3 2
```



```
## 0.490384615384616      0.5 0.509615384615385
##          5          1          0
## 0.519230769230769 0.528846153846154 0.538461538461539
##          12          0          3
## 0.557692307692308 0.567307692307692 0.576923076923077
##          0          0          0
## 0.586538461538462 0.596153846153846 0.615384615384615
##          0          0          0
##          0.625 0.663461538461539 0.769230769230769
##          0          0          0
##          1
##          0
```

8. Evaluate the model performance

```
CrossTable(x=wine_test_labels, y=wine_test_pred, prop.chisq=FALSE)
```

```
##
##
## Cell Contents
## |-----|
## |                      N |
## |      N / Row Total |
## |      N / Col Total |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table:  1715
##
##
##      | wine_test_pred
## wine_test_labels | 0.115384615384615 | 0.125 |
0.134615384615385 | 0.144230769230769 | 0.153846153846154 | 0.163461538461538
| 0.173076923076923 | 0.182692307692308 | 0.192307692307692 |
0.201923076923077 | 0.211538461538462 | 0.221153846153846 | 0.230769230769231
| 0.240384615384615 | 0.25 | 0.259615384615385 |
0.269230769230769 | 0.278846153846154 | 0.288461538461538 | 0.298076923076923
| 0.307692307692308 | 0.317307692307692 | 0.326923076923077 |
0.336538461538462 | 0.346153846153846 | 0.355769230769231 | 0.365384615384615
| 0.375 | 0.384615384615385 | 0.394230769230769 |
0.403846153846154 | 0.413461538461538 | 0.423076923076923 | 0.432692307692308
| 0.442307692307692 | 0.451923076923077 | 0.461538461538462 |
0.471153846153846 | 0.480769230769231 | 0.490384615384616 | 0.5
| 0.519230769230769 | 0.538461538461539 | Row Total |
## -----|-----|-----|-----|
---|-----|-----|-----|-----|
-----|-----|-----|-----|-----|
-----|-----|-----|-----|-----|
-----|-----|-----|-----|-----|
-----|-----|-----|-----|-----|
```


[illegible]

[illegible]

0.007	0.014	0.017	0.013
0.029	0.030	0.054	0.036
0.061	0.043	0.090	0.071
0.069	0.059	0.051	0.038
0.055	0.051	0.042	0.028
0.020	0.020	0.015	0.009
0.009	0.007	0.009	0.006
0.005	0.004	0.004	0.002
0.001	0.003	0.001	0.007
0.002			
##			
-- ----- ----- ----- -----			
--- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
- ----- ----- ----- -----			
--- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
----- ----- ----- ----- -----			
##			
##			

There were 1715 observations