Untitled

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

wine<-read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv",sep=";",header=T)  
#install.packages("class")  
#install.packages("gmodels")  
library(class)

## Warning: package 'class' was built under R version 3.4.1

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.4.1

#1. Check data characteristics. Is there missing data? Prepare the data for analysis.  
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 ...

sum(is.na(wine))

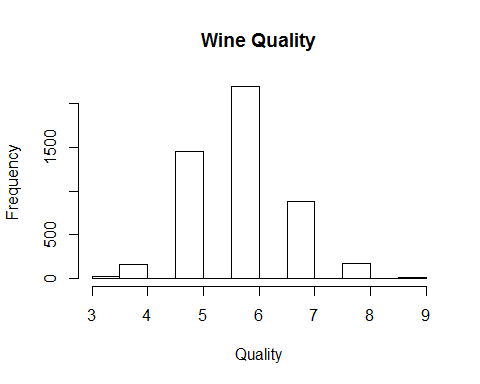
## [1] 0

# There is no missing data as we see from the output of sum(is.na(wine))

# 2. What is the correlation between the attributes other than wine quality?  
cor(wine[-12])

## fixed.acidity volatile.acidity citric.acid  
## fixed.acidity 1.00000000 -0.02269729 0.28918070  
## volatile.acidity -0.02269729 1.00000000 -0.14947181  
## citric.acid 0.28918070 -0.14947181 1.00000000  
## residual.sugar 0.08902070 0.06428606 0.09421162  
## chlorides 0.02308564 0.07051157 0.11436445  
## free.sulfur.dioxide -0.04939586 -0.09701194 0.09407722  
## total.sulfur.dioxide 0.09106976 0.08926050 0.12113080  
## density 0.26533101 0.02711385 0.14950257  
## pH -0.42585829 -0.03191537 -0.16374821  
## sulphates -0.01714299 -0.03572815 0.06233094  
## alcohol -0.12088112 0.06771794 -0.07572873  
## residual.sugar chlorides free.sulfur.dioxide  
## fixed.acidity 0.08902070 0.02308564 -0.0493958591  
## volatile.acidity 0.06428606 0.07051157 -0.0970119393  
## citric.acid 0.09421162 0.11436445 0.0940772210  
## residual.sugar 1.00000000 0.08868454 0.2990983537  
## chlorides 0.08868454 1.00000000 0.1013923521  
## free.sulfur.dioxide 0.29909835 0.10139235 1.0000000000  
## total.sulfur.dioxide 0.40143931 0.19891030 0.6155009650  
## density 0.83896645 0.25721132 0.2942104109  
## pH -0.19413345 -0.09043946 -0.0006177961  
## sulphates -0.02666437 0.01676288 0.0592172458  
## alcohol -0.45063122 -0.36018871 -0.2501039415  
## total.sulfur.dioxide density pH  
## fixed.acidity 0.091069756 0.26533101 -0.4258582910  
## volatile.acidity 0.089260504 0.02711385 -0.0319153683  
## citric.acid 0.121130798 0.14950257 -0.1637482114  
## residual.sugar 0.401439311 0.83896645 -0.1941334540  
## chlorides 0.198910300 0.25721132 -0.0904394560  
## free.sulfur.dioxide 0.615500965 0.29421041 -0.0006177961  
## total.sulfur.dioxide 1.000000000 0.52988132 0.0023209718  
## density 0.529881324 1.00000000 -0.0935914935  
## pH 0.002320972 -0.09359149 1.0000000000  
## sulphates 0.134562367 0.07449315 0.1559514973  
## alcohol -0.448892102 -0.78013762 0.1214320987  
## sulphates alcohol  
## fixed.acidity -0.01714299 -0.12088112  
## volatile.acidity -0.03572815 0.06771794  
## citric.acid 0.06233094 -0.07572873  
## residual.sugar -0.02666437 -0.45063122  
## chlorides 0.01676288 -0.36018871  
## free.sulfur.dioxide 0.05921725 -0.25010394  
## total.sulfur.dioxide 0.13456237 -0.44889210  
## density 0.07449315 -0.78013762  
## pH 0.15595150 0.12143210  
## sulphates 1.00000000 -0.01743277  
## alcohol -0.01743277 1.00000000

# 3. Graph the frequency distribution of wine quality.  
hist(wine$quality,main="Wine Quality",xlab="Quality")



# 4. Reduce the levels of rating for quality to three levels as high, medium and low.  
# .the range function, range(quality) shows the range of quality as 3 to 9. We will allocate levels as explained .low to ratings 3 and 4 , medium to ratings 5,6 and 7, and high to ratings 8 and 9.  
wine$quality<-cut(wine$quality,breaks=c(3,5,8,10),labels=c("low","medium","high"),right=FALSE)

# 5. Normalize the data set.  
# We use the following function to normalize data .  
Normalize\_function <- function(x) {(x - min(x))/(max(x) - min(x))}  
wine\_normalized<- as.data.frame(lapply(wine[-12], Normalize\_function))  
wine<-data.frame(wine\_normalized,wine$quality)

# 6. Divide the data to training and testing groups.  
train\_index <- sample(nrow(wine), floor(nrow(wine)\*.7))  
train\_set<-wine[train\_index,]  
test\_set<-wine[-train\_index,]  
train\_set\_new<-train\_set[-12]  
test\_set\_new<-test\_set[-12]  
train\_class<-train\_set$wine.quality  
test\_class <- test\_set$wine.quality

# 7. Use the KNN algorithm to predict the quality of wine using its attributes.  
wine\_knn\_prediction <- knn(train = train\_set\_new, test = test\_set\_new, cl= train\_class, k = 3)

# 8. Evaluate the model performance.  
CrossTable(x=test\_class, y=wine\_knn\_prediction, prop.chisq=FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1470   
##   
##   
## | wine\_knn\_prediction   
## test\_class | low | medium | high | Row Total |   
## -------------|-----------|-----------|-----------|-----------|  
## low | 3 | 56 | 1 | 60 |   
## | 0.050 | 0.933 | 0.017 | 0.041 |   
## | 0.273 | 0.039 | 0.040 | |   
## | 0.002 | 0.038 | 0.001 | |   
## -------------|-----------|-----------|-----------|-----------|  
## medium | 8 | 1323 | 16 | 1347 |   
## | 0.006 | 0.982 | 0.012 | 0.916 |   
## | 0.727 | 0.923 | 0.640 | |   
## | 0.005 | 0.900 | 0.011 | |   
## -------------|-----------|-----------|-----------|-----------|  
## high | 0 | 55 | 8 | 63 |   
## | 0.000 | 0.873 | 0.127 | 0.043 |   
## | 0.000 | 0.038 | 0.320 | |   
## | 0.000 | 0.037 | 0.005 | |   
## -------------|-----------|-----------|-----------|-----------|  
## Column Total | 11 | 1434 | 25 | 1470 |   
## | 0.007 | 0.976 | 0.017 | |   
## -------------|-----------|-----------|-----------|-----------|  
##   
##