Decision Tree

shorveer

October 30, 2017

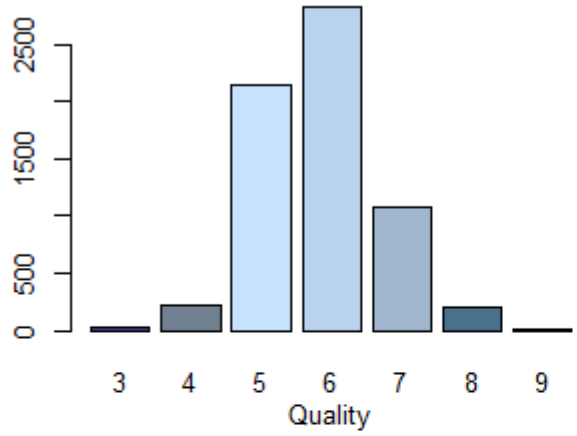
Data Exploration

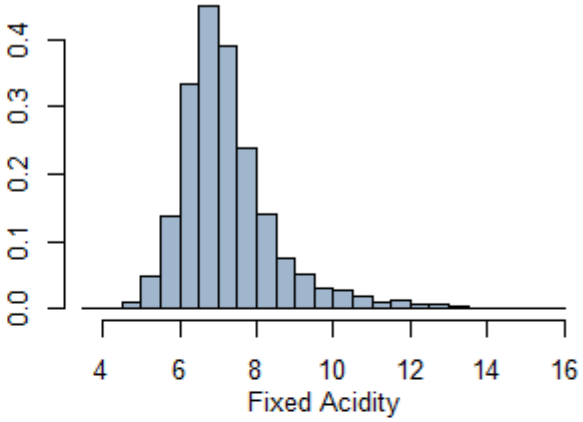
```
setwd("C:/Users/SS186102.TD/Desktop/exercise/Exercise_Case_Onsite_Modeling_Wine/Case_Onsite_Mode
ling_Wine")
getwd()
```

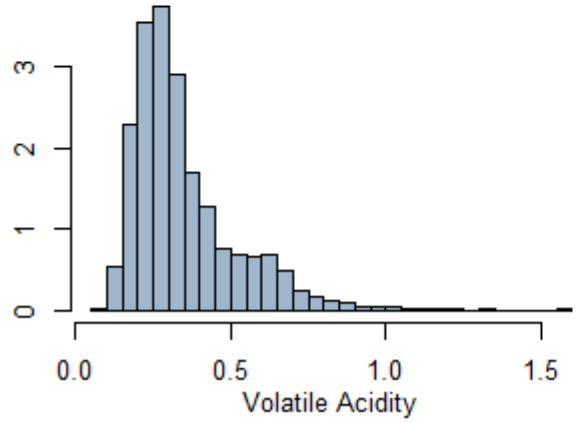
```
## [1] "C:/Users/SS186102.TD/Desktop/exercise/Exercise_Case_Onsite_Modeling_Wine/Case_Onsite_Mod
eling_Wine"
```

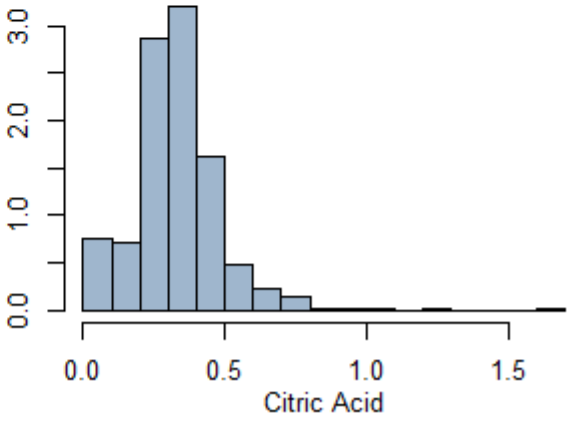
```
library(MASS)

wine<-read.csv("wine_dataset.csv")
attach(wine)
par(mfrow=c(2,2), oma = c(1,1,0,0) + 0.1, mar = c(3,3,1,1) + 0.1)
barplot((table(quality)), col=c("slateblue4", "slategray", "slategray1", "slategray2", "slategra
y3", "skyblue4"))
mtext("Quality", side=1, outer=F, line=2, cex=0.8)
truehist(fixed_acidity, h = 0.5, col="slategray3")
mtext("Fixed Acidity", side=1, outer=F, line=2, cex=0.8)
truehist(volatile_acidity, h = 0.05, col="slategray3")
mtext("Volatile Acidity", side=1, outer=F, line=2, cex=0.8)
truehist(citric_acid, h = 0.1, col="slategray3")
mtext("Citric Acid", side=1, outer=F, line=2, cex=0.8)
```

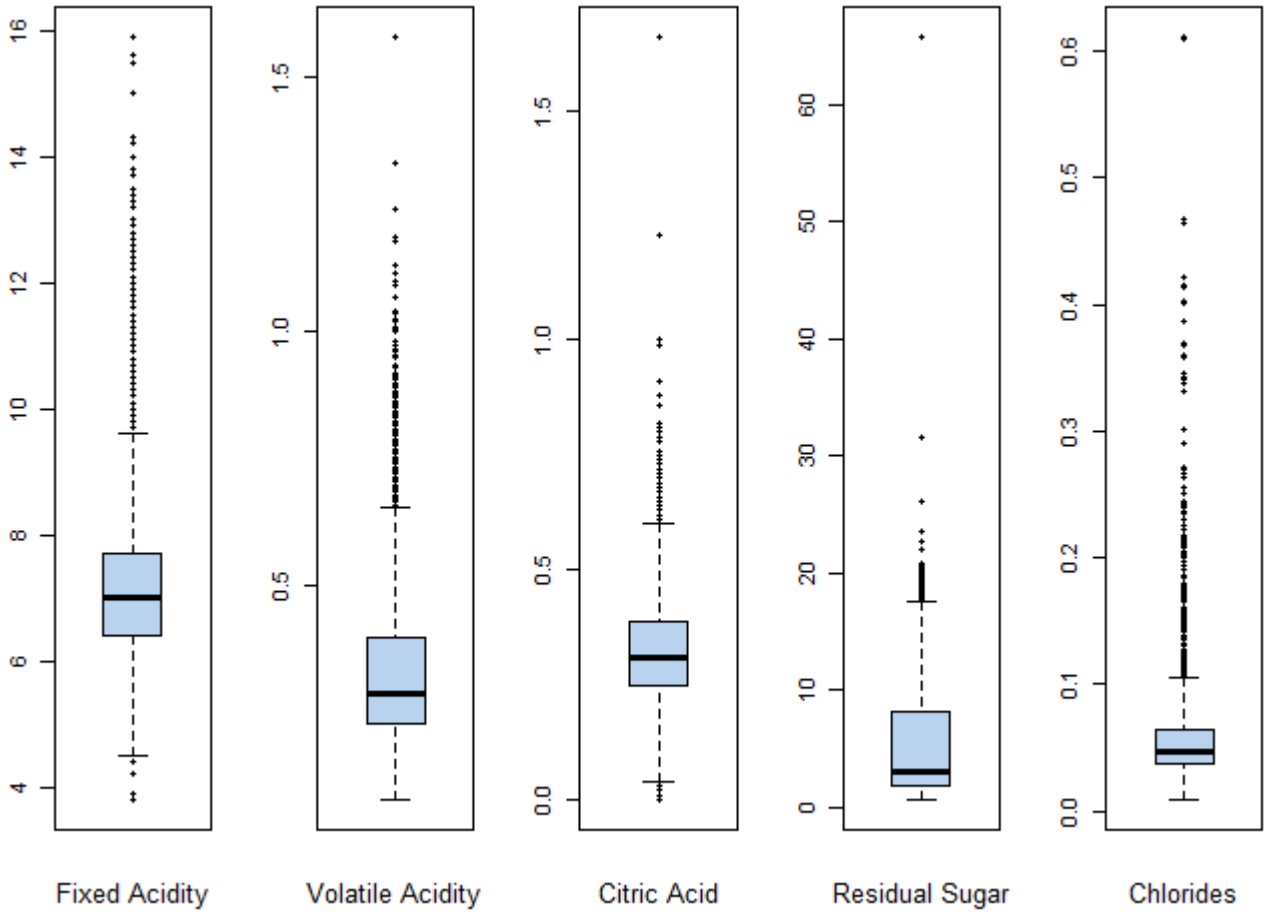









```
par(mfrow=c(1,5), oma = c(1,1,0,0) + 0.1,  mar = c(3,3,1,1) + 0.1)
boxplot(fixed_acidity, col="slategray2", pch=19)
mtext("Fixed Acidity", cex=0.8, side=1, line=2)
boxplot(volatile_acidity, col="slategray2", pch=19)
mtext("Volatile Acidity", cex=0.8, side=1, line=2)
boxplot(citric_acid, col="slategray2", pch=19)
mtext("Citric Acid", cex=0.8, side=1, line=2)
boxplot(residual_sugar, col="slategray2", pch=19)
mtext("Residual Sugar", cex=0.8, side=1, line=2)
boxplot(chlorides, col="slategray2", pch=19)
mtext("Chlorides", cex=0.8, side=1, line=2)
```



Dividing data for training and testing

```
ind<-sample(2,nrow(wine),replace = TRUE,prob=c(0.8,0.2))
training<-wine[ind==1,]
test<-wine[ind==2,]
```

Plotting

```
#install.packages("plotly")
library(plotly)
```

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

```
##
## Attaching package: 'plotly'
```

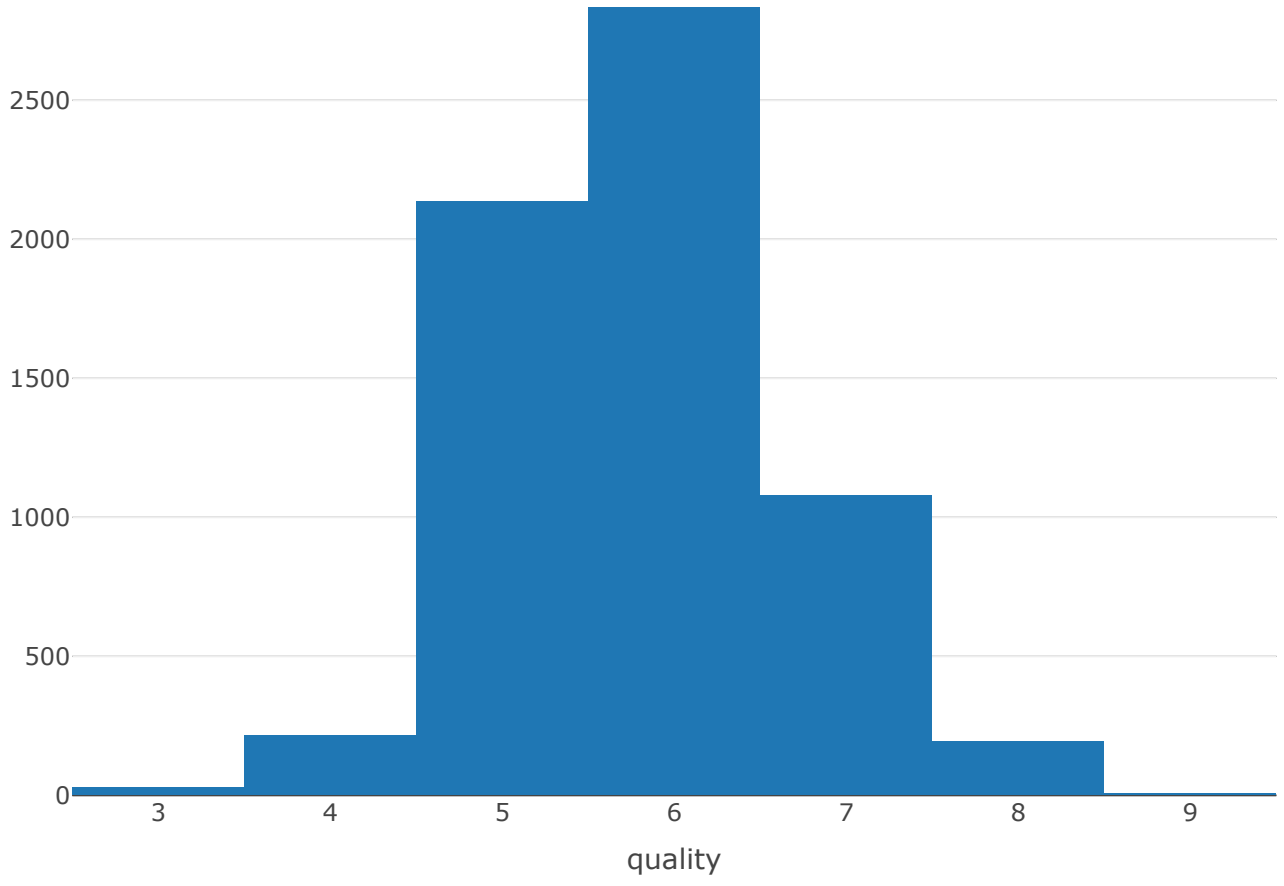
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

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

```
## The following object is masked from 'package:stats':
##
##     filter
```

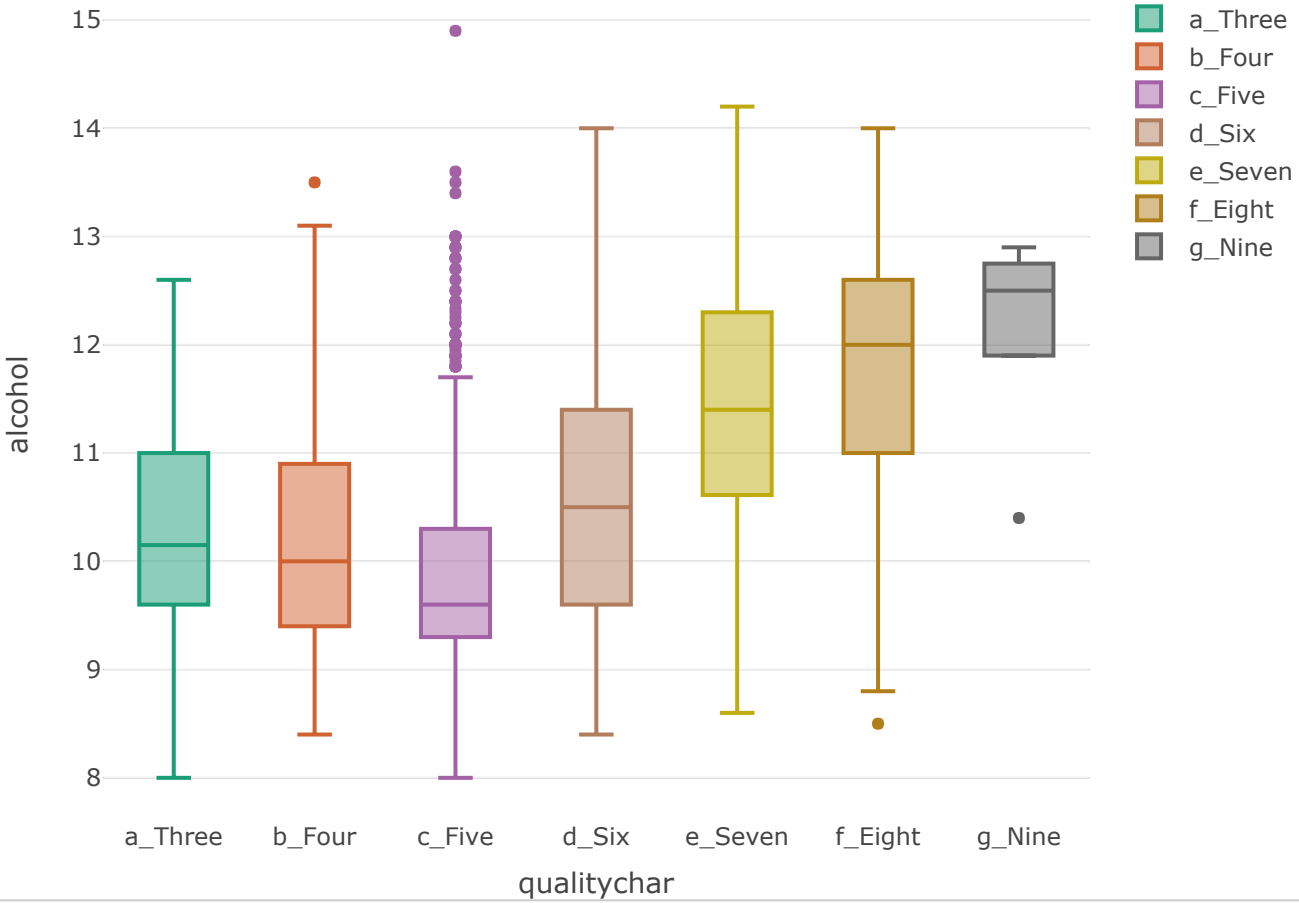
```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
plot_ly(data = wine, x =~quality, type = "histogram")
```

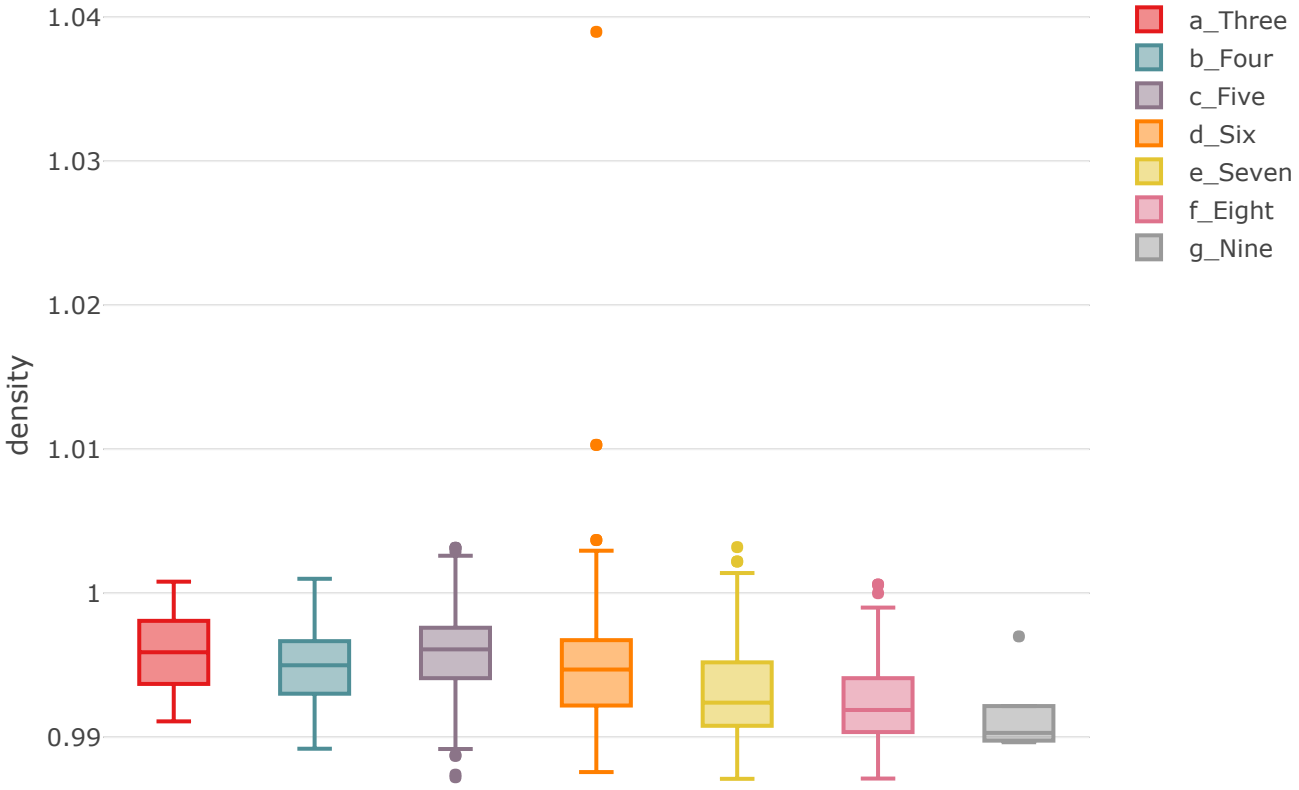


Classifying Quality into Varchar

```
wine2<-wine
wine2$qualitychar <- ifelse(wine2$quality == 3, "a_Three", ifelse(wine2$quality == 4, "b_Four",
ifelse(wine2$quality == 5, "c_Five", ifelse(wine2$quality == 6, "d_Six", ifelse(wine2$quality ==
 7, "e_Seven", ifelse(wine2$quality == 8, "f_Eight", "g_Nine"))) )))
plot_ly(data = wine2, x = ~qualitychar, y = ~alcohol, color = ~qualitychar, type = "box", colors
 = "Dark2")
```




```
plot_ly(data = wine2, x = ~qualitychar, y = ~density, color = ~qualitychar, type = "box", colors
 = "Set1")
```



a_Three b_Four c_Five d_Six e_Seven f_Eight g_Nine

qualitychar

ceating decision tree model

```
library(rpart)
m.rpart <- rpart(quality ~. , data = training)
summary(m.rpart)
```

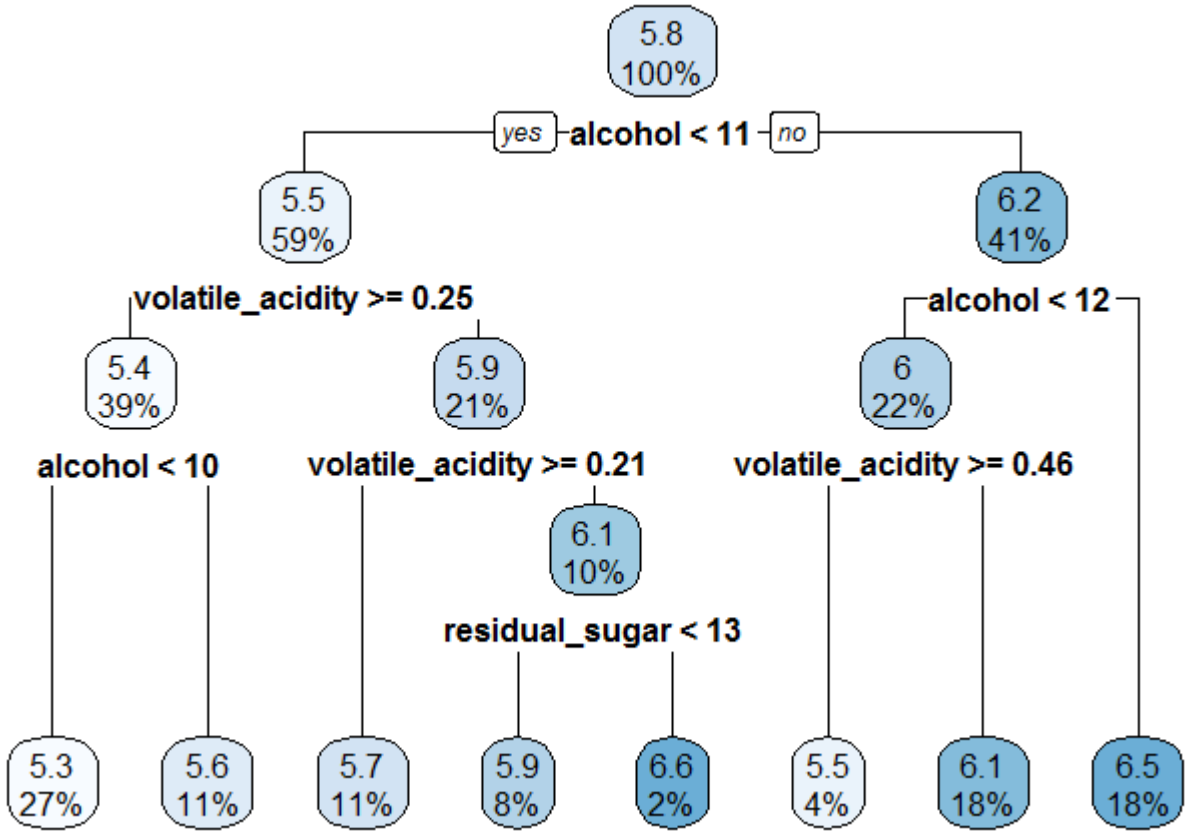
```
## Call:
## rpart(formula = quality ~ ., data = training)
##   n= 5148
##
##           CP nsplit rel error    xerror       xstd
## 1 0.15685615      0 1.0000000 1.0003991 0.02091585
## 2 0.04618368      1 0.8431439 0.8435833 0.01979467
## 3 0.03136367      2 0.7969602 0.8022446 0.01922041
## 4 0.01416425      3 0.7655965 0.7733910 0.01834768
## 5 0.01061926      4 0.7514323 0.7648303 0.01818130
## 6 0.01039343      6 0.7301937 0.7562525 0.01794797
## 7 0.01000000      7 0.7198003 0.7541135 0.01787880
##
## Variable importance
##              alcohol              density     volatile_acidity
##                   41                   21                   15
##            chlorides       residual_sugar total_sulfur_dioxide
##                   11                    3                    3
##        fixed_acidity            sulphates          citric_acid
##                    2                    2                    1
##                   pH                style
##                    1                    1
##
## Node number 1: 5148 observations,    complexity param=0.1568561
##   mean=5.820319, MSE=0.7507371
##   left son=2 (3058 obs) right son=3 (2090 obs)
##   Primary splits:
##       alcohol          < 10.625   to the left,  improve=0.15685610, (0 missing)
##       density          < 0.992355 to the right, improve=0.10238200, (0 missing)
##       chlorides        < 0.0395   to the right, improve=0.07559757, (0 missing)
##       volatile_acidity < 0.535    to the right, improve=0.04893905, (0 missing)
##       citric_acid      < 0.235    to the left,  improve=0.04461683, (0 missing)
##   Surrogate splits:
##       density              < 0.992845 to the right, agree=0.796, adj=0.499, (0 split)
##       chlorides            < 0.0375   to the right, agree=0.714, adj=0.296, (0 split)
##       total_sulfur_dioxide < 123.5    to the right, agree=0.620, adj=0.064, (0 split)
##       fixed_acidity        < 6.05     to the right, agree=0.614, adj=0.049, (0 split)
##       sulphates            < 0.375    to the right, agree=0.609, adj=0.036, (0 split)
##
## Node number 2: 3058 observations,    complexity param=0.04618368
##   mean=5.536625, MSE=0.5632433
##   left son=4 (1984 obs) right son=5 (1074 obs)
##   Primary splits:
##       volatile_acidity    < 0.2525   to the right, improve=0.10362900, (0 missing)
##       citric_acid         < 0.265    to the left,  improve=0.03389501, (0 missing)
##       alcohol             < 9.85     to the left,  improve=0.02881660, (0 missing)
##       chlorides           < 0.0595   to the right, improve=0.02815980, (0 missing)
##       free_sulfur_dioxide < 16.5     to the left,  improve=0.02175889, (0 missing)
##   Surrogate splits:
##       residual_sugar < 13.675   to the left,  agree=0.663, adj=0.041, (0 split)
##       sulphates      < 0.395    to the right, agree=0.663, adj=0.041, (0 split)
##       pH             < 3.005    to the right, agree=0.663, adj=0.039, (0 split)
##       density        < 0.99423  to the right, agree=0.661, adj=0.035, (0 split)
```

```
##       chlorides      < 0.0305   to the right, agree=0.652, adj=0.008, (0 split)
##
## Node number 3: 2090 observations,    complexity param=0.03136367
##   mean=6.235407, MSE=0.7350143
##   left son=6 (1141 obs) right son=7 (949 obs)
##   Primary splits:
##       alcohol             < 11.61667 to the left,  improve=0.07890621, (0 missing)
##       volatile_acidity    < 0.665    to the right, improve=0.04048507, (0 missing)
##       citric_acid         < 0.235    to the left,  improve=0.04044102, (0 missing)
##       free_sulfur_dioxide < 11.5     to the left,  improve=0.03754727, (0 missing)
##       density             < 0.991235 to the right, improve=0.03666693, (0 missing)
##   Surrogate splits:
##       density          < 0.991015 to the right, agree=0.717, adj=0.376, (0 split)
##       chlorides        < 0.0365   to the right, agree=0.641, adj=0.210, (0 split)
##       volatile_acidity < 0.2975   to the left,  agree=0.578, adj=0.072, (0 split)
##       fixed_acidity    < 5.85     to the right, agree=0.568, adj=0.048, (0 split)
##       sulphates        < 0.365    to the right, agree=0.565, adj=0.042, (0 split)
##
## Node number 4: 1984 observations,    complexity param=0.01039343
##   mean=5.358871, MSE=0.4619374
##   left son=8 (1410 obs) right son=9 (574 obs)
##   Primary splits:
##       alcohol          < 9.975    to the left,  improve=0.04382889, (0 missing)
##       volatile_acidity < 0.5925   to the right, improve=0.02397443, (0 missing)
##       sulphates        < 0.545    to the left,  improve=0.02121005, (0 missing)
##       chlorides        < 0.0405   to the right, improve=0.01479430, (0 missing)
##       fixed_acidity    < 10.75    to the left,  improve=0.01005238, (0 missing)
##   Surrogate splits:
##       density              < 0.99283  to the right, agree=0.738, adj=0.094, (0 split)
##       chlorides            < 0.0315   to the right, agree=0.716, adj=0.019, (0 split)
##       fixed_acidity        < 5.05     to the right, agree=0.713, adj=0.009, (0 split)
##       total_sulfur_dioxide < 9.5      to the right, agree=0.713, adj=0.009, (0 split)
##       free_sulfur_dioxide  < 86.5     to the left,  agree=0.713, adj=0.007, (0 split)
##
## Node number 5: 1074 observations,    complexity param=0.01061926
##   mean=5.864991, MSE=0.5841933
##   left son=10 (559 obs) right son=11 (515 obs)
##   Primary splits:
##       volatile_acidity    < 0.2075   to the right, improve=0.06129307, (0 missing)
##       residual_sugar      < 12.55    to the left,  improve=0.03048954, (0 missing)
##       density             < 0.99781  to the left,  improve=0.02472688, (0 missing)
##       free_sulfur_dioxide < 13.5     to the left,  improve=0.02365689, (0 missing)
##       citric_acid         < 0.265    to the left,  improve=0.02247904, (0 missing)
##   Surrogate splits:
##       residual_sugar       < 6.875    to the right, agree=0.614, adj=0.194, (0 split)
##       density              < 0.99503  to the right, agree=0.597, adj=0.159, (0 split)
##       total_sulfur_dioxide < 131.5    to the right, agree=0.591, adj=0.148, (0 split)
##       free_sulfur_dioxide  < 34.5     to the right, agree=0.579, adj=0.122, (0 split)
##       pH                   < 3.275    to the left,  agree=0.574, adj=0.111, (0 split)
##
## Node number 6: 1141 observations,    complexity param=0.01416425
##   mean=6.015776, MSE=0.7166661
##   left son=12 (195 obs) right son=13 (946 obs)
##   Primary splits:
```

```
##       volatile_acidity     < 0.455    to the right, improve=0.06694488, (0 missing)
##       citric_acid          < 0.235    to the left,  improve=0.06410912, (0 missing)
##       free_sulfur_dioxide  < 11.5     to the left,  improve=0.05172412, (0 missing)
##       total_sulfur_dioxide < 74.5     to the left,  improve=0.02208169, (0 missing)
##       pH                   < 3.495    to the right, improve=0.01990531, (0 missing)
##   Surrogate splits:
##       citric_acid          < 0.145    to the left,  agree=0.916, adj=0.508, (0 split)
##       style                splits as  LR,           agree=0.871, adj=0.246, (0 split)
##       chlorides            < 0.0705   to the right, agree=0.866, adj=0.215, (0 split)
##       total_sulfur_dioxide < 54.5     to the left,  agree=0.857, adj=0.164, (0 split)
##       pH                   < 3.495    to the right, agree=0.847, adj=0.103, (0 split)
##
## Node number 7: 949 observations
##   mean=6.499473, MSE=0.6293464
##
## Node number 8: 1410 observations
##   mean=5.268085, MSE=0.3735204
##
## Node number 9: 574 observations
##   mean=5.581882, MSE=0.6091491
##
## Node number 10: 559 observations
##   mean=5.683363, MSE=0.4596695
##
## Node number 11: 515 observations,    complexity param=0.01061926
##   mean=6.062136, MSE=0.6446828
##   left son=22 (406 obs) right son=23 (109 obs)
##   Primary splits:
##       residual_sugar      < 12.55    to the left,  improve=0.13139840, (0 missing)
##       density             < 0.998045 to the left,  improve=0.12379120, (0 missing)
##       alcohol             < 9.05     to the right, improve=0.11928560, (0 missing)
##       free_sulfur_dioxide < 10.5     to the left,  improve=0.04841303, (0 missing)
##       fixed_acidity       < 8.25     to the right, improve=0.03655032, (0 missing)
##   Surrogate splits:
##       density              < 0.99701  to the left,  agree=0.940, adj=0.716, (0 split)
##       alcohol              < 9.15     to the right, agree=0.850, adj=0.294, (0 split)
##       pH                   < 2.945    to the right, agree=0.804, adj=0.073, (0 split)
##       total_sulfur_dioxide < 212.25   to the left,  agree=0.792, adj=0.018, (0 split)
##
## Node number 12: 195 observations
##   mean=5.533333, MSE=0.7001709
##
## Node number 13: 946 observations
##   mean=6.115222, MSE=0.6621996
##
## Node number 22: 406 observations
##   mean=5.91133, MSE=0.5389357
##
## Node number 23: 109 observations
##   mean=6.623853, MSE=0.6383301
```

ploting of decision tree

```
library(rpart.plot)
rpart.plot(m.rpart)
```



Prediction on Test Datset

```
p.rpart <- predict(m.rpart,test)

MAE <- function(actual, predicted){
mean(abs(actual - predicted))
}
MAE(test$quality, p.rpart)
```

```
## [1] 0.6175123
```

Testing a new entry of Wine

```
test <- data.frame(fixed_acidity = 8.5, volatile_acidity = 0.33, citric_acid = 0.42, residual_su
gar = 10.5, chlorides = 0.065, free_sulfur_dioxide = 47, total_sulfur_dioxide = 186, density =
0.9955, pH = 3.10, sulphates = 0.40, alcohol = 9.9,style='red')
test_pred <- predict(m.rpart, test)
test_pred
```

```
##        1
## 5.268085
```

Another approaches

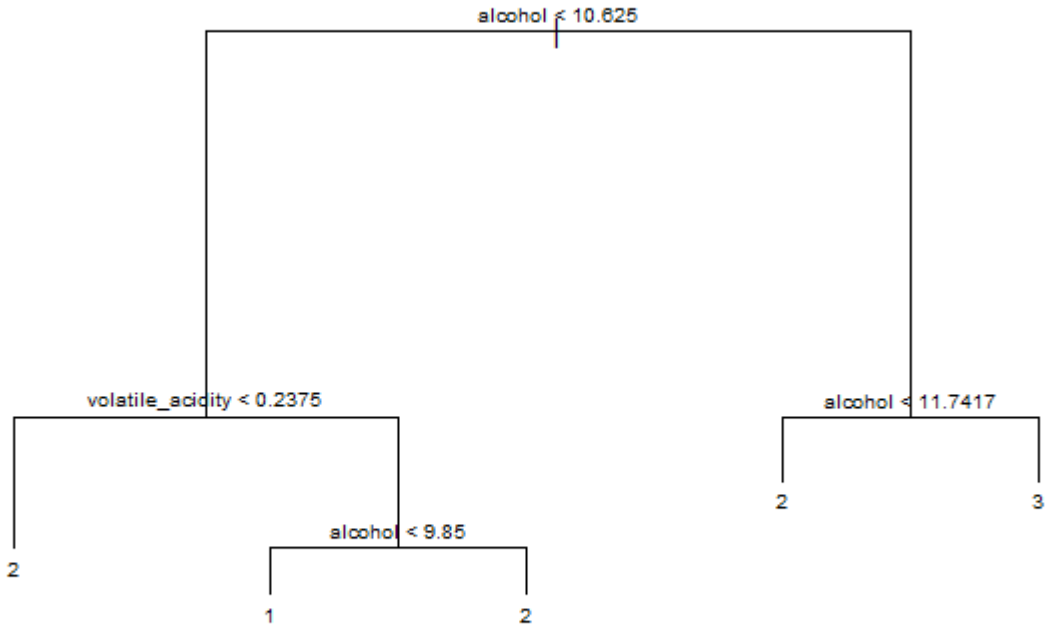
```
wine1<-wine
wine1$FactQ <- ifelse(wine$quality<=5 ,1,ifelse(wine2$quality==6,2,3))
wine1$FactQ<-as.factor(wine1$FactQ)
prop.table(table(wine1$FactQ))
```

```
##
##         1         2         3
## 0.3669386 0.4365092 0.1965523
```

```
library(tree)
attach(wine1)
```

```
## The following objects are masked from wine:
##
##     alcohol, chlorides, citric_acid, density, fixed_acidity,
##     free_sulfur_dioxide, pH, quality, residual_sugar, style,
##     sulphates, total_sulfur_dioxide, volatile_acidity
```

```
WhiteWineTree <- tree(FactQ ~ fixed_acidity+volatile_acidity+citric_acid+
residual_sugar+chlorides+free_sulfur_dioxide+total_sulfur_dioxide+pH+sulphates+alcohol+density,
data=wine, method="class")
plot(WhiteWineTree)
text(WhiteWineTree, pretty=0, cex=0.6)
```




```
misclass.tree(WhiteWineTree, detail=T)
```

```
##    1    2    4    5   10   11    3    6    7
## 3661 1892  464 1112  594  517 1381  740  578
```

```
Treefit1 <- predict(WhiteWineTree, wine1, type="class")
table(Treefit1, wine1$FactQ)
```

```
##
## Treefit1    1    2    3
##        1 1261  556   38
##        2 1048 1777  673
##        3   75  503  566
```

```
ind<-sample(2,nrow(wine1),replace = TRUE,prob=c(0.8,0.2))
training_wine<-wine1[ind==1,]
test_wine<-wine1[ind==2,]
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

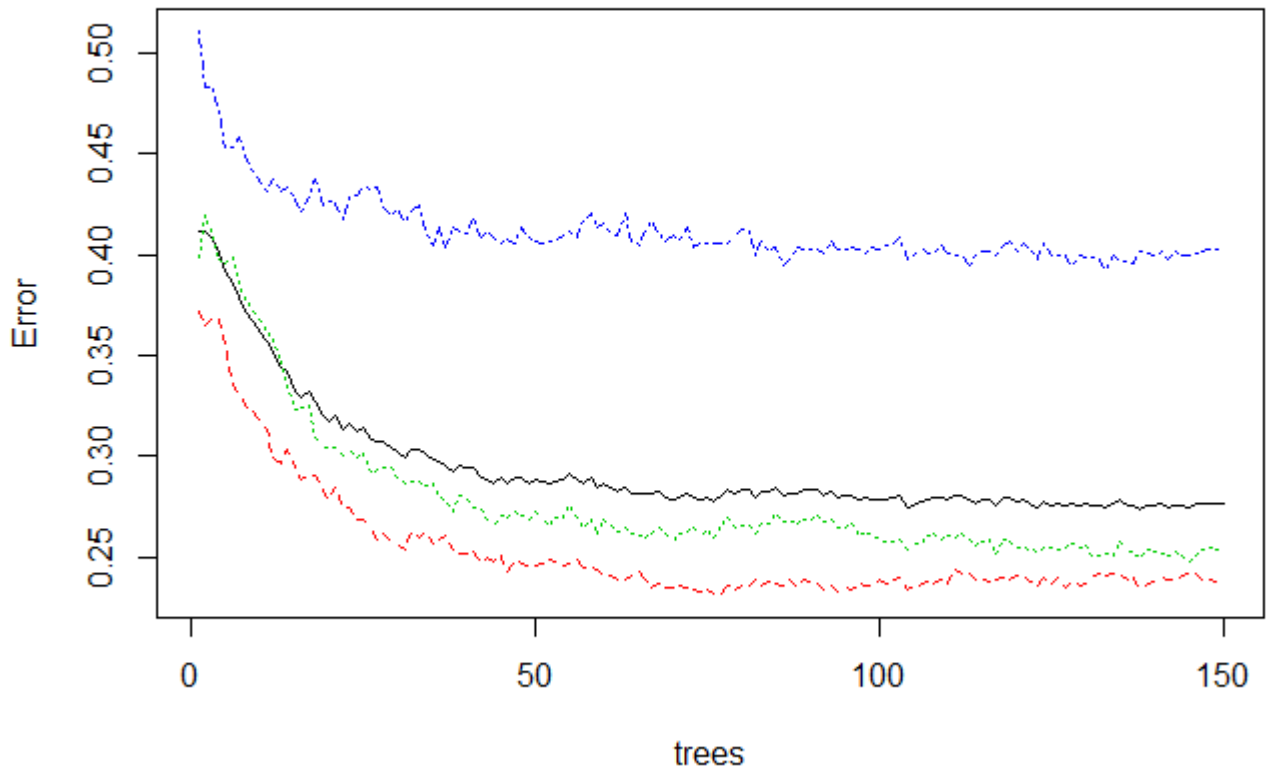
```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
WRF_model <- randomForest(FactQ ~ . , data=training_wine[,-12], ntree=150, importance=T, proximi
ty=T)
WRF_model_pred <- predict(WRF_model, test_wine, type="class")
ac<-table(WRF_model_pred, test_wine$FactQ)
accuracy<- sum(diag(ac))/sum(ac)
accuracy
```

```
## [1] 0.7215385
```

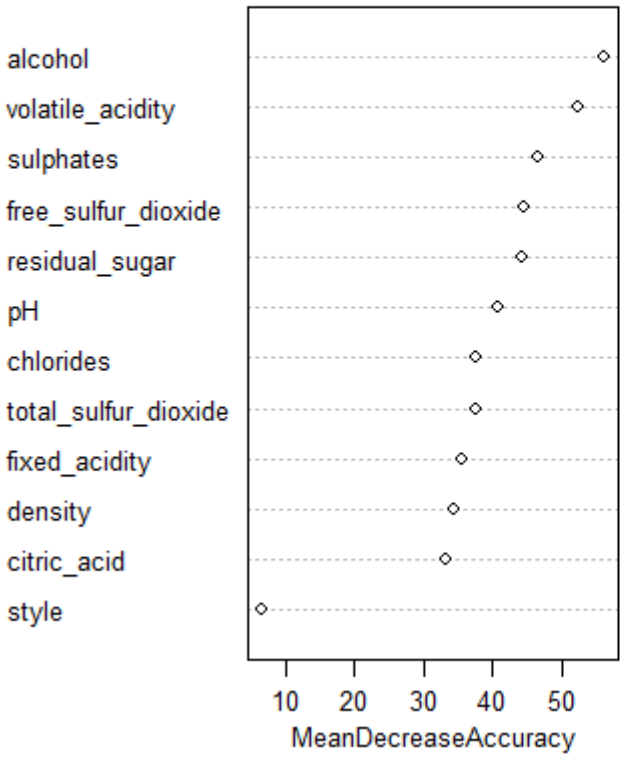
Completely unsupervised random forest method on Training data with ntree = 150 leads to the following error plot:

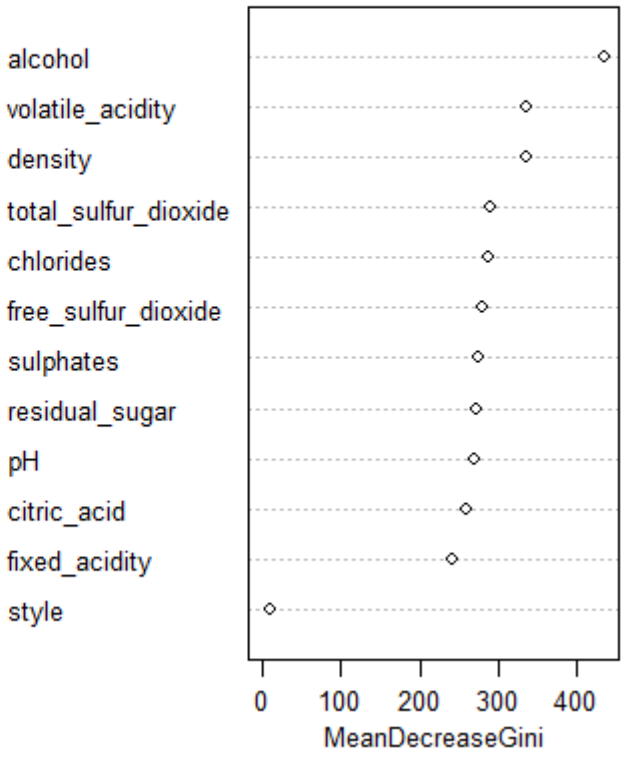
```
plot(WRF_model, main="")
```



Importance of predictors are given in the following dotplot:

```
varImpPlot(WRF_model,  main="", cex=0.8)
```


alcohol
volatile_acidity
sulphates
free_sulfur_dioxide
residual_sugar
pH
chlorides
total_sulfur_dioxide
fixed_acidity
density
citric_acid
style
10
20
30
40
50
MeanDecreaseAccuracy


alcohol
volatile_acidity
density
total_sulfur_dioxide
chlorides
free_sulfur_dioxide
sulphates
residual_sugar
pH
citric_acid
fixed_acidity
style
0
100
200
300
400
MeanDecreaseGini