```
library("ggplot2")
library("dplyr")
library("gridExtra")
library(Simpsons)
library(GGally)
library(memisc)
library(pander)
library(corrplot)

#Loading the csv file
wine <- read.csv('wineQualityReds.csv')

#Transforming Quality from an Integer to a Factor
wine$quality <- factor(wine$quality, ordered = T)

#Creating a new Factored Variable called 'Rating'

wine$rating <- ifelse(wine$quality < 5, 'bad', ifelse(
  wine$quality < 7, 'average', 'good'))

wine$rating <- ordered(wine$rating,
                       levels = c('bad', 'average', 'good'))

str(wine)

summary(wine)

ggplot(data = wine, aes(x = quality)) +
  geom_bar(width = 1, color = 'black',fill = I('orange'))

ggplot(data = wine, aes(x = rating)) +
  geom_bar(width = 1, color = 'black',fill = I('blue'))

grid.arrange(ggplot(wine, aes( x = 1, y = fixed.acidity ) ) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(4,14)),
             ggplot(data = wine, aes(x = fixed.acidity)) +
               geom_histogram(binwidth = 1, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(4,14)),ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = volatile.acidity ) ) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(0,1)),
             ggplot(data = wine, aes(x = volatile.acidity)) +
               geom_histogram(binwidth = 0.05, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(0,1)), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = citric.acid )) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ),
             ggplot(data = wine, aes(x = citric.acid)) +
               geom_histogram(binwidth = 0.08, color = 'black',fill =
I('orange')) +
               scale_x_continuous(breaks = seq(0,1,0.1), lim = c(0,1)),
ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = residual.sugar )) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(1,8)),
             ggplot(data = wine, aes(x = residual.sugar)) +
               geom_histogram(binwidth = 0.1, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(1,8)), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = chlorides )) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(0,0.25)),
             ggplot(data = wine, aes(x = chlorides)) +
               geom_histogram(binwidth = 0.01, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(0,0.25)), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = free.sulfur.dioxide )) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(0,45)),
             ggplot(data = wine, aes(x = free.sulfur.dioxide)) +
               geom_histogram(binwidth = 1, color = 'black',fill =
I('orange')) +
               scale_x_continuous(breaks = seq(0,80,5), lim = c(0,45)),
ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = total.sulfur.dioxide )) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(0,180)),
             ggplot(data = wine, aes(x = total.sulfur.dioxide)) +
               geom_histogram(binwidth = 5, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(0,180)), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = density)) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ),
             ggplot(data = wine, aes(x = density)) +
               geom_histogram(binwidth = 0.001, color = 'black',fill =
I('orange')), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = pH)) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ),
             ggplot(data = wine, aes(x = pH)) +
               geom_histogram(binwidth = 0.1, color = 'black',fill =
I('orange')), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = sulphates)) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(0.3,1.6)),
             ggplot(data = wine, aes(x = sulphates)) +
               geom_histogram(binwidth = 0.1, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(0.3,1.6)), ncol = 2)

grid.arrange(ggplot(wine, aes( x = 1, y = alcohol)) +
               geom_jitter(alpha = 0.1 ) +
               geom_boxplot(alpha = 0.2, color = 'red' ) +
               scale_y_continuous(lim = c(8,14)),
             ggplot(data = wine, aes(x = alcohol)) +
               geom_histogram(binwidth = 0.1, color = 'black',fill =
I('orange')) +
               scale_x_continuous(lim = c(8,14)), ncol = 2)

c <- cor(
  wine %>%
    # first we remove unwanted columns
    dplyr::select(-X) %>%
    dplyr::select(-rating) %>%
    mutate(
      # now we translate quality to a number
      quality = as.numeric(quality)
    )
)
emphasize.strong.cells(which(abs(c) > .3 & c != 1, arr.ind = TRUE))
pandoc.table(c)

ggplot(data = wine, aes(x = quality, y = fixed.acidity)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x = quality, y = volatile.acidity)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=citric.acid)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=residual.sugar)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  scale_y_continuous(lim = c(0,5)) +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=chlorides)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  scale_y_continuous(lim = c(0,0.2)) +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=free.sulfur.dioxide)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  scale_y_continuous(lim = c(0,40)) +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=total.sulfur.dioxide)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  scale_y_continuous(lim = c(0,150)) +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=density)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=pH)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data = wine, aes(x = fixed.acidity, y = pH)) +
  geom_point(alpha = 0.3) +
  scale_x_log10(breaks=seq(5,15,1)) +
  xlab("Fixed Acidity in Log Scale") +
  geom_smooth(method="lm")

ggplot(data = wine, aes(x = volatile.acidity, y = pH)) +
  geom_point(alpha = 0.3) +
  scale_x_log10(breaks=seq(.1,1,.1)) +
  xlab("Volatile Acidity in Log Scale") +
  geom_smooth(method="lm")

ggplot(data = subset(wine, citric.acid > 0), aes(x = citric.acid, y = pH))
+
  geom_point(alpha = 0.3) +
  scale_x_log10() +
  xlab("Citric Acid in Log Scale") +
  geom_smooth(method="lm")

simpsons <- Simpsons(volatile.acidity, pH, data=wine)
plot(simpsons)


ggplot(data=wine, aes(x=quality, y=sulphates)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  scale_y_continuous(lim = c(0.25,1)) +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

ggplot(data=wine, aes(x=quality, y=alcohol)) +
  geom_jitter( alpha = .3) +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)

alcoholQualityLinearModel <- lm(as.numeric(quality) ~ alcohol,
                                data = wine)
summary(alcoholQualityLinearModel)


simple_cor_test <- function(x, y) {
  return(cor.test(x, as.numeric(y))$estimate)
}

correlations <- c(
  simple_cor_test(wine$fixed.acidity, wine$quality),
  simple_cor_test(wine$volatile.acidity, wine$quality),
  simple_cor_test(wine$citric.acid, wine$quality),
  simple_cor_test(log10(wine$residual.sugar), wine$quality),
  simple_cor_test(log10(wine$chlorides), wine$quality),
  simple_cor_test(wine$free.sulfur.dioxide, wine$quality),
  simple_cor_test(wine$total.sulfur.dioxide, wine$quality),
  simple_cor_test(wine$density, wine$quality),
  simple_cor_test(wine$pH, wine$quality),
  simple_cor_test(log10(wine$sulphates), wine$quality),
  simple_cor_test(wine$alcohol, wine$quality))
names(correlations) <- c('fixed.acidity', 'volatile.acidity',
'citric.acid',
                         'log10.residual.sugar',
                         'log10.chlordies', 'free.sulfur.dioxide',
                         'total.sulfur.dioxide', 'density', 'pH',
                         'log10.sulphates', 'alcohol')

correlations

ggplot(data = wine,
       aes(y = density, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))

ggplot(data = wine,
       aes(y = sulphates, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  scale_y_continuous(limits=c(0.3,1.5)) +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))

ggplot(data = wine,
       aes(y = volatile.acidity, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))

ggplot(data = wine,
       aes(y = pH, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))

ggplot(data = wine,
       aes(y = residual.sugar, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))

ggplot(data = wine,
       aes(y = total.sulfur.dioxide, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))


ggplot(data = wine,
       aes(y = citric.acid, x = volatile.acidity,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))


ggplot(data = wine,
       aes(y = citric.acid, x = fixed.acidity,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))


ggplot(data = wine,
       aes(y = fixed.acidity, x = volatile.acidity,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality'))

set.seed(1221)
training_data <- sample_frac(wine, .6)
test_data <- wine[ !wine$X %in% training_data$X, ]
m1 <- lm(as.numeric(quality) ~ alcohol, data = training_data)
m2 <- update(m1, ~ . + sulphates)
m3 <- update(m2, ~ . + volatile.acidity)
m4 <- update(m3, ~ . + citric.acid)
m5 <- update(m4, ~ . + fixed.acidity)
m6 <- update(m2, ~ . + pH)
mtable(m1,m2,m3,m4,m5,m6)


wine_predict <- data.frame(
  test_data$quality,
  predict(m5, test_data) - as.numeric(test_data$quality)
)
names(wine_predict) <- c("quality", "error")
ggplot(data=wine_predict, aes(x=quality,y=error)) +
  geom_jitter(alpha = 0.3)

ggplot(data=wine, aes(y=alcohol, x=quality)) +
  geom_jitter(alpha = .3)  +
  geom_boxplot(alpha = .5,color = 'blue') +
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4) +
  xlab("Quality") +
  ggtitle("Influence of alcohol on wine quality")


ggplot(data = wine,
       aes(y = sulphates, x = alcohol,
           color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  scale_y_continuous(limits=c(0.3,1.5)) +
  ylab("potassium sulphate (g/dm3)") +
  xlab("Alcohol Percentage") +
  scale_color_brewer(type='seq',
                     guide=guide_legend(title='Quality')) +
  ggtitle("Alcohol and sulphates over wine quality")


df <- data.frame(
  test_data$quality,
  predict(m5, test_data) - as.numeric(test_data$quality)
)
names(df) <- c("quality", "error")
ggplot(data=df, aes(x=quality,y=error)) +
  geom_jitter(alpha = 0.3) +
  ggtitle("Linear model errors vs expected quality")
```