BSc Coursework 1

Salik Tariq / Student ID: 12516369

1) Statistical learning methods

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible

statistical learning method to be better or worse than an inflexible method.

(a) The number of predictors p is extremely large, and the number of observations n is small.

Worse - If the number of predictors p is extremely large and the observations n is small then we will not have enough information about the effect and variation of each parameters considered, therefore flexible method will overfit the model due to small number of observations n.

(b) The sample size n is extremely large, and the number of predictors p is small.

Better - If the sample size n is extremely large, and the number of predictors p is small then we will have enough information about most predictors, so a flexible method would perform better.

(c) The relationship between the predictors and response is highly non-linear.

Better - a more flexible method will fit the data better due to the degree of freedom it provides, therefore felxible method would perform better.

(d) The standard deviation of the error terms, i.e. $\sigma = sd(\epsilon)$, is extremely high.

Worse - The flexible statistical learning method will incorporate the noise in error terms and thus will increase the SD of error and hence the model fitting would be poor in case of flexibel method.

2) Bayes' rule

Given a dataset including 20 samples (S__1, . . . , S__20) about the temperature (i.e. hot or cool) for playing golf

(i.e. yes or no), you are required to use the Bayes' rule to calculate the probability of playing golf according

to the temperature, i.e. P(Play Golf | Temperature).

P(Play Golf = Yes) = 10/20 = 0.5

P(Play Gold = No) = 10/20 = 0.5

P(Temperature = Hot) = 12/20 = 0.6

P(Temperature = Cool) = 8/20 = 0.4

P(Temperature = Hot | Play Golf = Yes) = 5/10 = 0.5

P(Temperature = Hot | Play Golf = No) = 7/10 = 0.7

P(Temperature = Cool | Play Golf = Yes) = 5/10 = 0.5

P(Temperature = Cool | Play Golf = No) = 3/10 = 0.3

P(Play Golf = Yes | Temperature = Hot) = (0.5 * 0.5)/0.6 = 5/12

P(Play Golf = No | Temperature = Hot) = (0.7 * 0.5)/0.6 = 7/12

P(Play Golf = Yes | Temperature = Cool) = (0.5 * 0.5)/0.4 = 5/8

P(Play Golf = No | Temperature = Cool) = (0.3 * 0.5)/0.4 = 3/8

3) Descriptive analysis

This exercise involves the Auto data set studied in the class.

```
#install.packages("ISLR")
library(ISLR) #this library contains Auto dataset
```

(a) Which of the predictors are quantitative, and which are qualitative?

```
str(Auto)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

Based on the output:

Quantitative predictors are: mpg, cylinders, displacement, horsepower, weight, acceleration, year and origin.
Qualitative predictors are: name

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
range(Auto$mpg)
```

```
## [1]  9.0 46.6
```

Range for mpg is 9.0 to 46.6

```
range(Auto$cylinders)
```

```
## [1] 3 8
```

Range for cylinders is from 3 to 8

```
range(Auto$displacement)
```

```
## [1]  68 455
```

Range for displacement is from 68 to 455

```
range(Auto$horsepower)
```

```
## [1]  46 230
```

Range for horsepower is from 46 to 230

```
range(Auto$weight)
```

```
## [1] 1613 5140
```

Range for weight is from 1613 to 5140

```
range(Auto$acceleration)
```

```
## [1]  8.0 24.8
```

Range for acceleration is from 8.0 to 24.8

```
range(Auto$year)
```

```
## [1] 70 82
```

Range for year is from 70 to 82

```
range(Auto$origin)
```

```
## [1] 1 3
```

Range for origin is from 1 to 3

(c) What is the median and variance of each quantitative predictor?

```
library(ISLR)
apply(Auto[,1:8],2,median)
```

```
##          mpg    cylinders displacement   horsepower       weight
##        22.75         4.00       151.00        93.50      2803.50
## acceleration         year       origin
##        15.50        76.00         1.00
```

Medians of all quantitative predictors

```
apply(Auto[,1:8],2,var)
```

```
##          mpg    cylinders displacement   horsepower       weight
## 6.091814e+01 2.909696e+00 1.095037e+04 1.481569e+03 7.214847e+05
## acceleration         year       origin
## 7.611331e+00 1.356991e+01 6.488595e-01
```

Variance of all quantitative predictors

(d) Now remove the 11th through 79th observations (inclusive) in the dataset. What is the range, median,

and variance of each predictor in the subset of the data that remains?

```
Auto2 <- Auto[-c(11:79),] #removing rows 11th to 79th inclusive

apply(Auto2[,1:8], 2, range) #2 indicates that the operation is applied by column
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
##      origin
## [1,]      1
```

```
## [2,]      3
```

Range of the remaining data after removing rows 11th through 79th.

```
apply(Auto2[,1:8], 2, median)
```

```
##          mpg    cylinders displacement   horsepower       weight
##         23.9          4.0        144.0         90.0       2789.0
## acceleration         year       origin
##         15.5         77.0          1.0
```

Median of the remaining data after removing rows 11th through 79th.

```
apply(Auto2[,1:8], 2, var)
```

```
##          mpg    cylinders displacement   horsepower       weight
## 6.135039e+01 2.748938e+00 1.003229e+04 1.291977e+03 6.574970e+05
## acceleration         year       origin
## 7.296234e+00 1.004873e+01 6.803261e-01
```

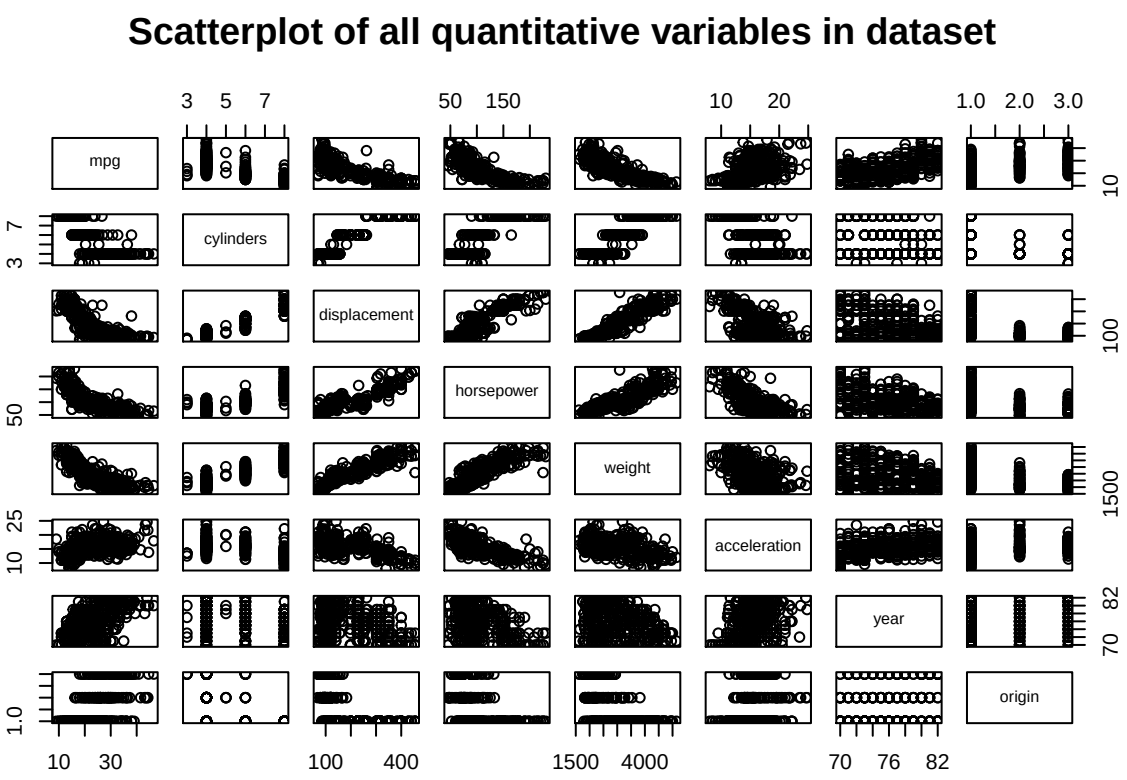
Variance of the remaining data after removing rows 11th through 79th.

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your

choice. Create some plots highlighting the relationships among the predictors. Comment on your

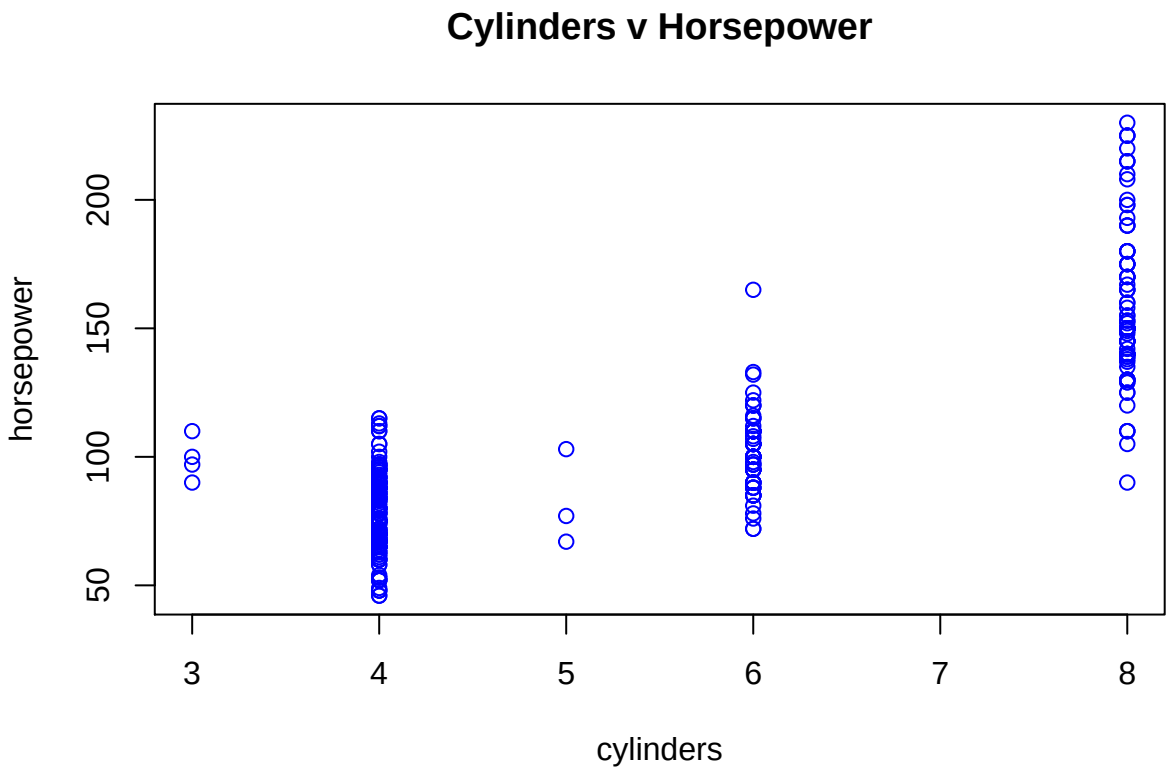
findings.

```
Auto3 <- Auto[, !sapply(Auto, is.factor)]
pairs(Auto3, main="Scatterplot of all quantitative variables in dataset")
```



Scatterplot of cylinders v horsepower

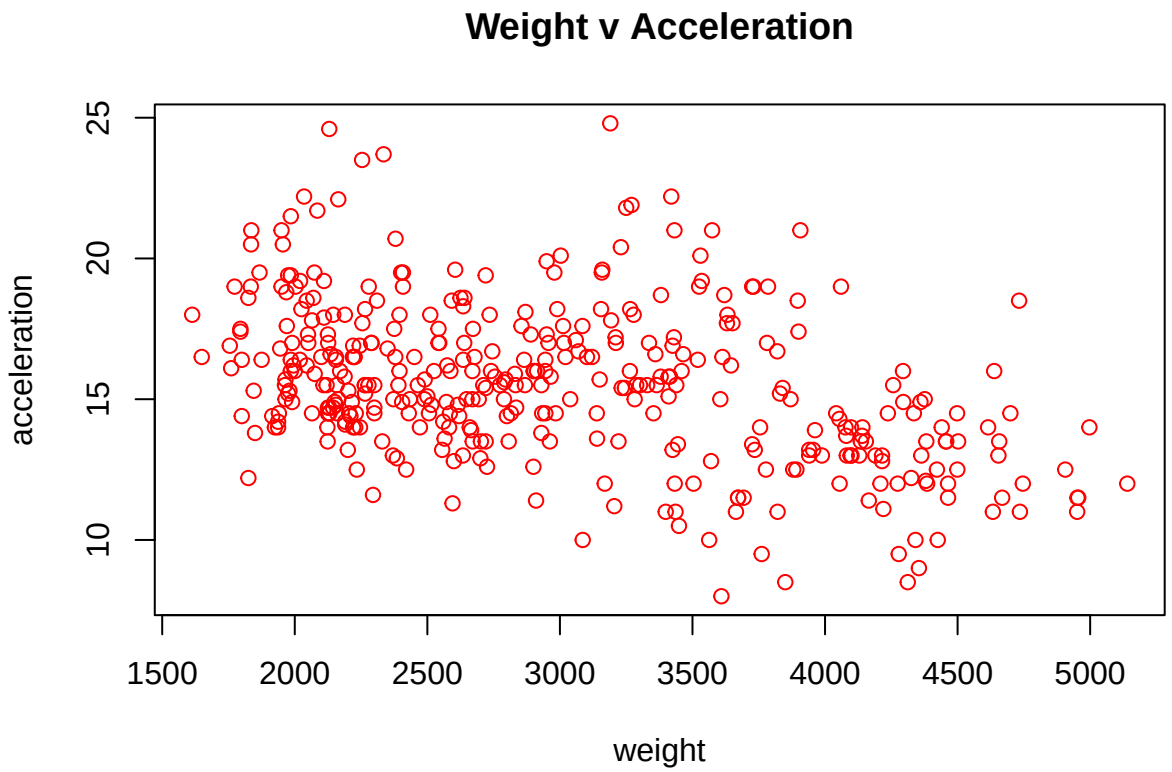
```r
attach(Auto3) # to access variables of a Auto3 without calling the Auto3.
plot(cylinders, horsepower, col = "blue", main="Cylinders v Horsepower")
```



We can see that by increasing the number of cylinders, the horsepower increases, therefore horsepower is directly proportional to the number of cylinders.

Scatterplot of weight v acceleration

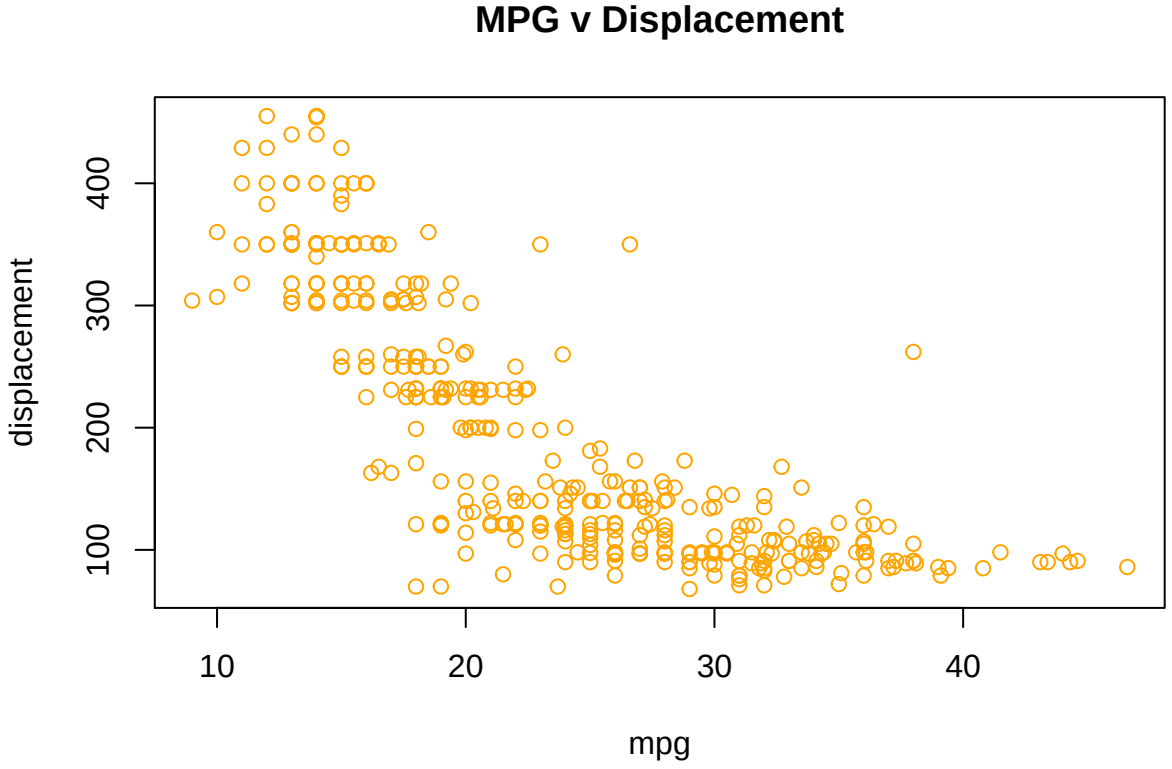
```r
plot(weight, acceleration, col = "red", main="Weight v Acceleration")
```



We can see that by increasing the weight, the acceleration decreases, therefore weight is inversely proportional to the acceleration.

Scatterplot of mpg v displacement

```
plot(mpg, displacement, col = "orange", main="MPG v Displacement")
```



We can

see that by increasing the mpg, the displacement decreases, therefore by increaseing the mpg(fuel efficiency), the displacement decreases.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots

suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

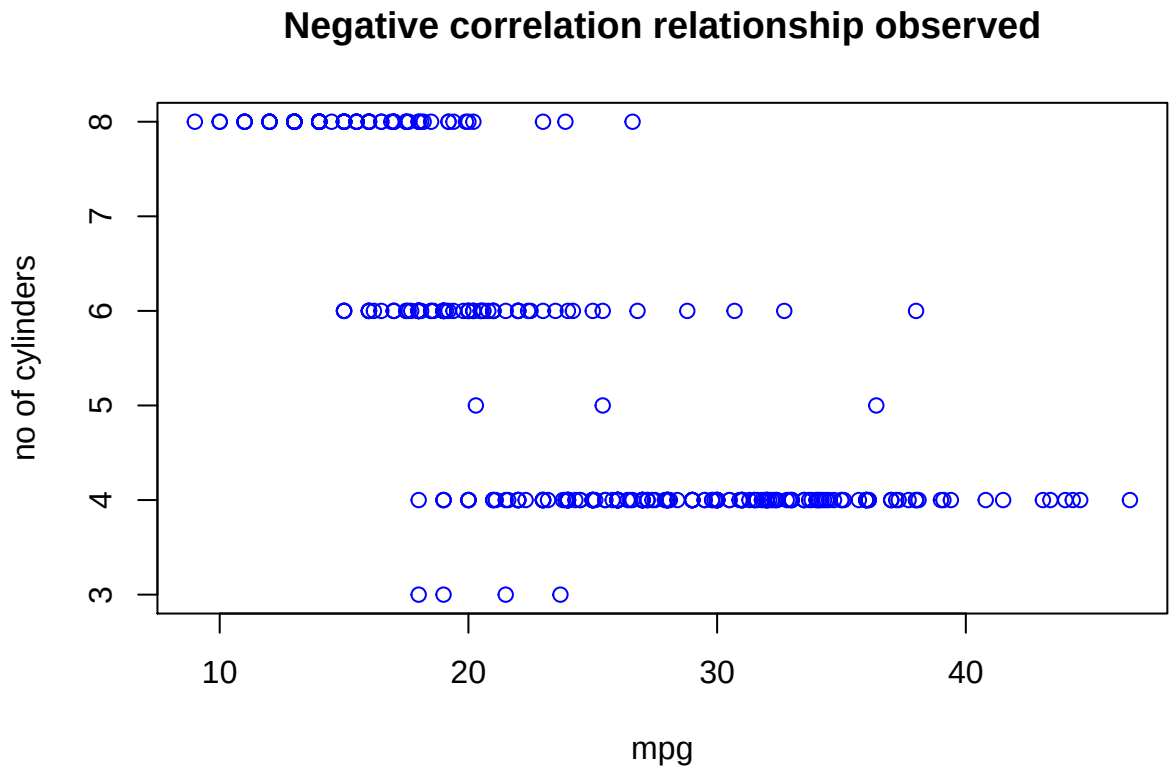
```
sapply(Auto3, function(x) cor(Auto$mpg, x))
```

```
##          mpg    cylinders displacement   horsepower       weight
##    1.0000000   -0.7776175   -0.8051269   -0.7784268   -0.8322442
## acceleration         year       origin
##    0.4233285    0.5805410    0.5652088
```

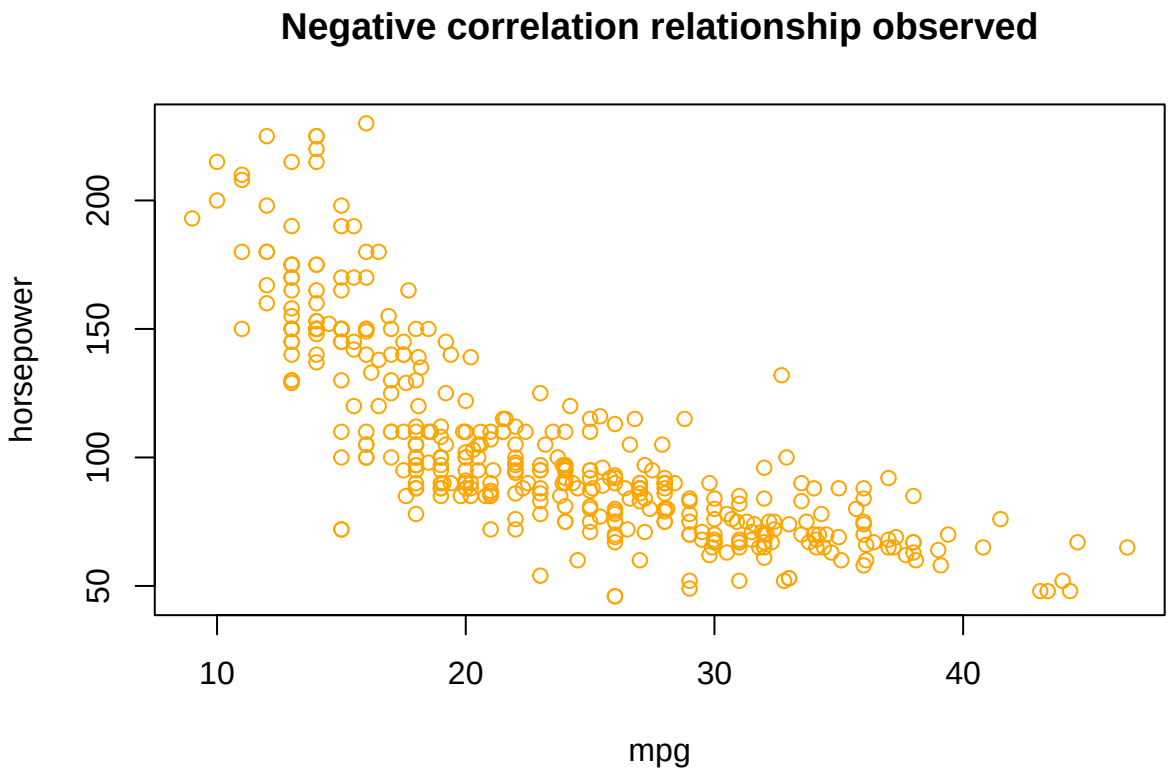
Looking at the above correlation data, we can see that cylinders, displacement, horsepower and weight have strong negative correlation relationship with mpg. Which means that any of these will need to be decreased in order to increase the gas milage. We have already observed the negative correlation relationship between mpg and displacement before in the scatterplot.

Plotting the respective plots:

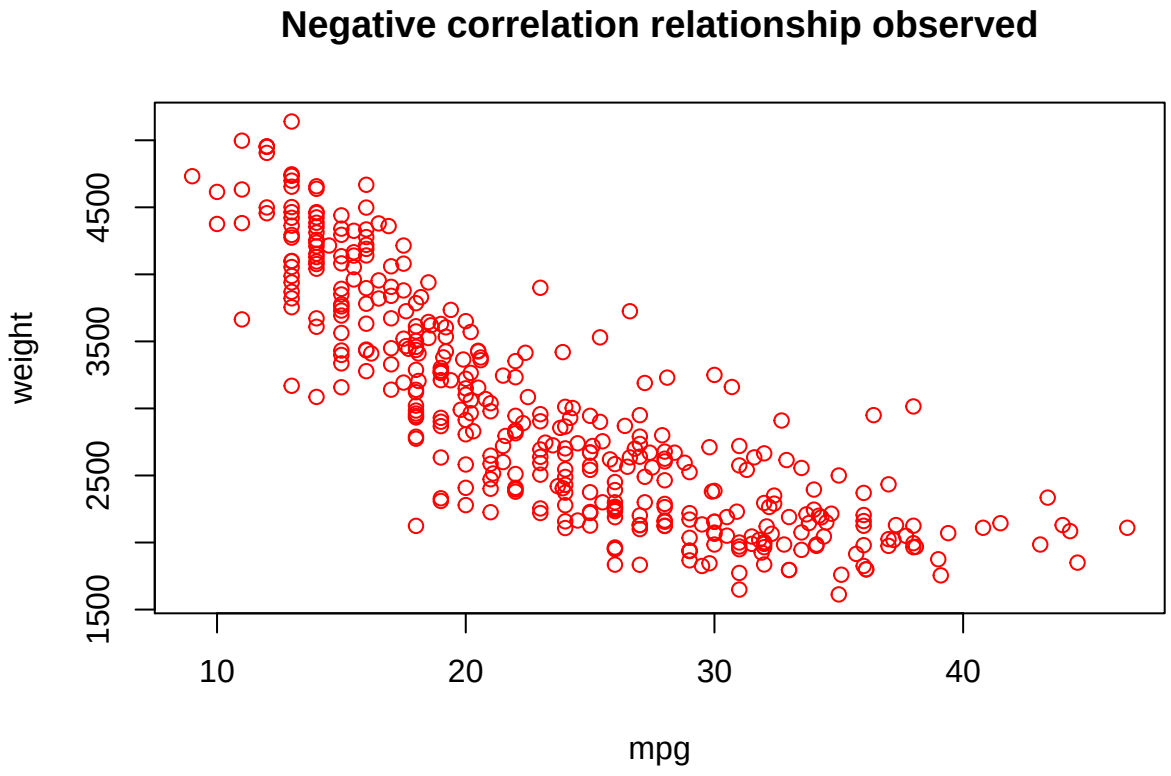
```
plot(mpg,cylinders, main="Negative correlation relationship observed", xlab="mpg", ylab="no of cylinder
```



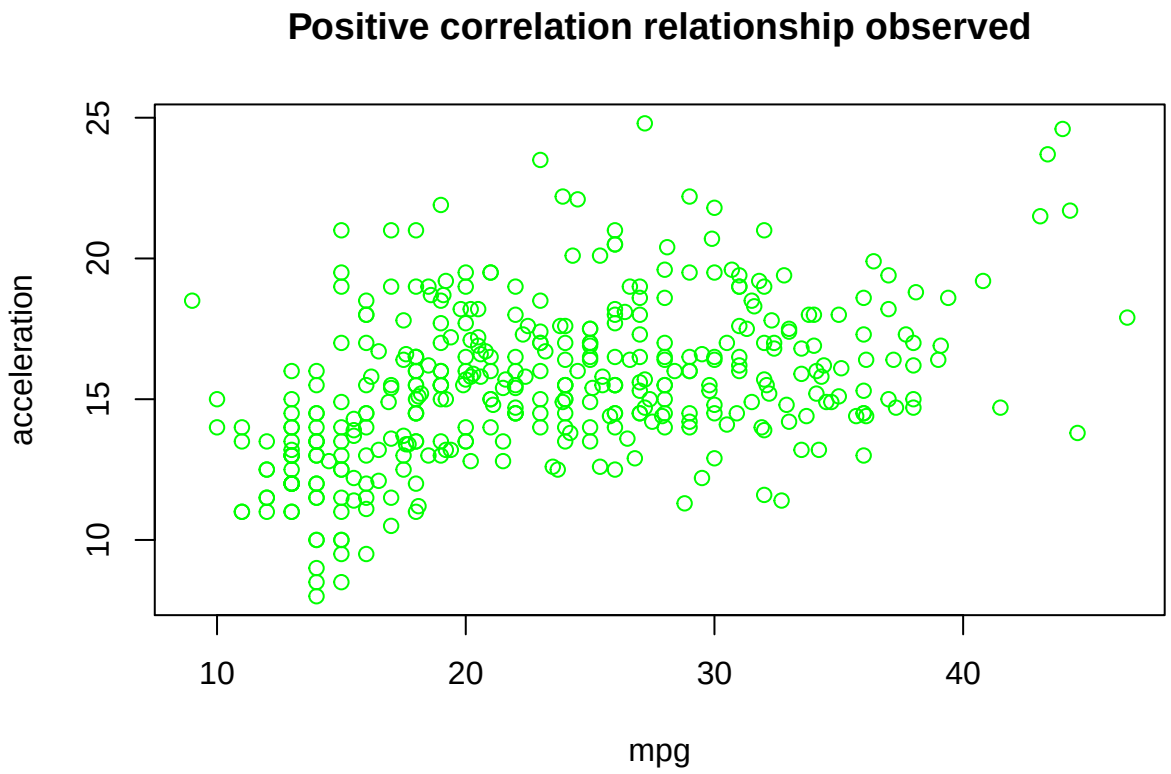

```
plot(mpg,horsepower, main="Negative correlation relationship observed", xlab="mpg", ylab="horsepower",
```




```
plot(mpg,weight, main="Negative correlation relationship observed", xlab="mpg", ylab="weight", col="red
```




```
plot(mpg,acceleration, main="Positive correlation relationship observed", xlab="mpg", ylab="acceleratio
```




```
detach(Auto3)
```

4) Linear regression

This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as

the predictor. Use the summary() function to print the results. Comment on the output. For example:

```
library(ISLR)
lm.fit <- lm(mpg ~ horsepower, data=Auto)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictor and the response?

The p-value for the coefficients is very close to zero, implying statistical significance which means that there is a relationship.

ii. How strong is the relationship between the predictor and the response?

The adjusted R square value indicated that 60.59% of variation in the response variable mpg is due to the horsepower. Also the coefficient value for horsepower is -0.157845 indicates an inverse relationship between the mpg and horsepower.

iii. Is the relationship between the predictor and the response positive or negative?

The regression coefficient of horsepower is negative; therefore the relationship is negative.

iv. What is the predicted mpg associated with a horsepower of 89? What are the associated 99% confidence

and prediction intervals?

```
predict(lm.fit, data.frame(horsepower = c(89)))
```

```
##        1
## 25.88768
```

The predicted mpg associated with horsepower of 89 is 25.88768.

```
p_conf <- predict(lm.fit, data.frame(horsepower = c(89)), interval = "confidence", conf.level=0.99)
print(p_conf)
```

```
##        fit      lwr      upr
## 1 25.88768 25.36257 26.41279
```

The upper 99% confidence interval is 26.41279 The lower 99% confidence interval is 25.36257

```
p_pred <- predict(lm.fit, data.frame(horsepower = c(89)), interval = "prediction", pred.level=0.99)
print(p_pred)
```

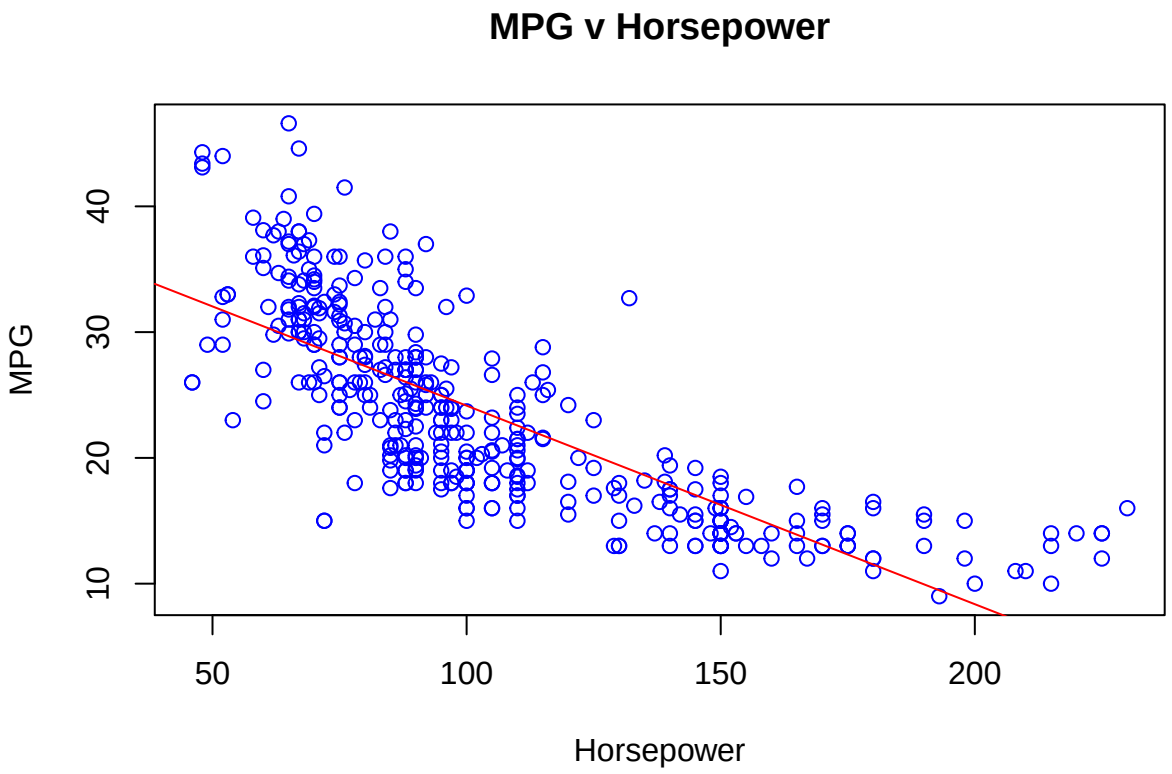
```
##        fit      lwr    upr
## 1 25.88768 16.22836 35.547
```

The upper 99% prediction interval is 35.547 The lower 99% prediction interval is 16.22836

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression

line.

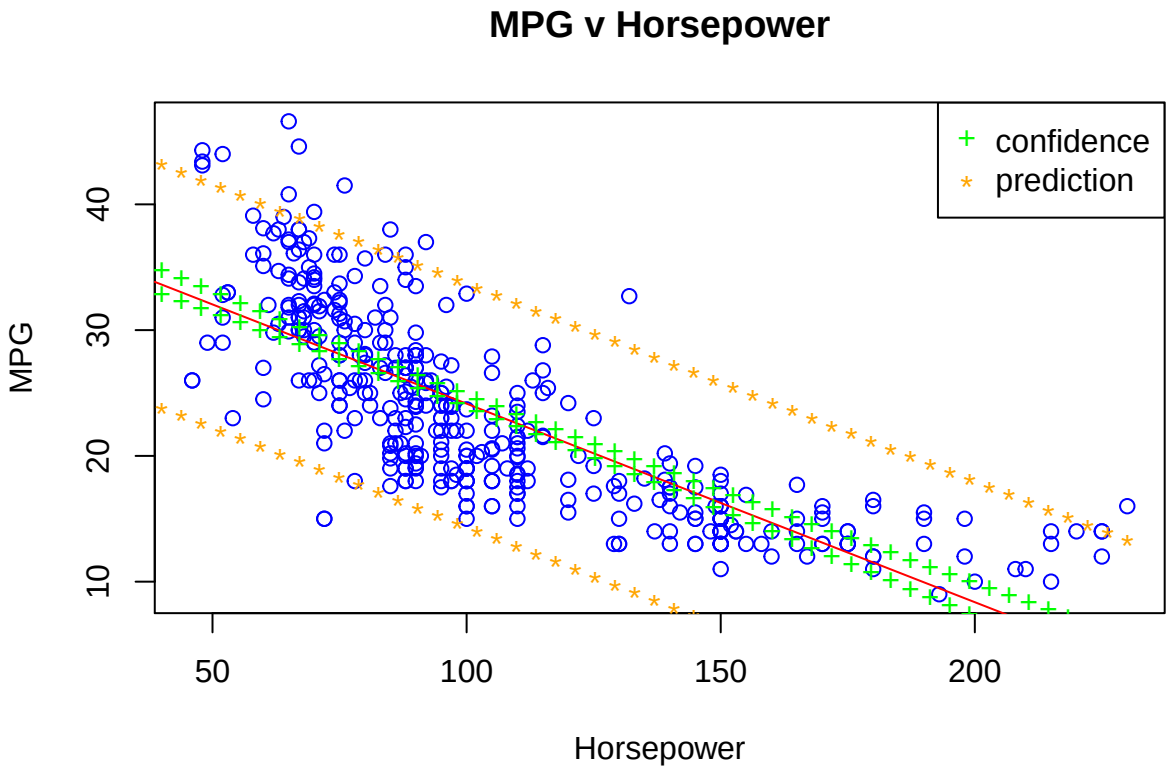
```
plot(Auto$mpg ~ Auto$horsepower, main="MPG v Horsepower", xlab="Horsepower", ylab="MPG", col="blue")
abline(coef = coef(lm.fit), col="red")
```



(c) Plot the 99% confidence interval and prediction interval in the same plot as (b) using different colours

and legends.

```
plot(Auto$mpg ~ Auto$horsepower, main="MPG v Horsepower", xlab="Horsepower", ylab="MPG", col="blue")
abline(coef = coef(lm.fit), col="red")
newHP <- data.frame(horsepower=seq(40,230,length=50)
                    )
p_conf <- predict(lm.fit, newHP,interval="confidence", conf.level=0.99)
p_pred <- predict(lm.fit, newHP,interval="prediction", conf.level=0.99)
lines(newHP$horsepower, p_conf[,"lwr"], col="green", type="b", pch="+")
lines(newHP$horsepower, p_conf[,"upr"], col="green", type="b", pch="+")
lines(newHP$horsepower, p_pred[,"upr"], col="orange", type="b", pch="*")
lines(newHP$horsepower, p_pred[,"lwr"], col="orange", type="b", pch="*")
legend("topright",
pch=c("+","*"),
col=c("green","orange"),
legend = c("confidence","prediction"))
```



5. Logistic regression

A recent study has shown that the accurate prediction of the office room occupancy leads to potential energy

savings of 30%. In this question, you are required to build logistic regression models by using different

environmental measurements as features, such as temperature, humidity, light, CO2 and humidity ratio, to

predict the office room occupancy. The provided training dataset consists of 2,000 samples, whilst the testing

dataset consists of 300 samples.

(a) Load the training and testing datasets from corresponding files, and display the statistics about different

features in the training dataset.

```
training_data <- read.table("Training_set for Q5.txt", header=T, sep=",")
testing_data <- read.table("Testing_set for Q5.txt", header = T, sep=",")
summary(training_data)
```

```
##   Temperature       Humidity         Light             CO2
##  Min.   :20.10   Min.   :18.96   Min.   :  0.0   Min.   : 426.0
##  1st Qu.:20.89   1st Qu.:21.82   1st Qu.:  0.0   1st Qu.: 448.0
##  Median :21.20   Median :25.00   Median :  0.0   Median : 485.5
##  Mean   :21.42   Mean   :24.22   Mean   :144.7   Mean   : 634.6
```

```
##  3rd Qu.:22.10   3rd Qu.:26.29   3rd Qu.:433.0   3rd Qu.: 845.8
##  Max.   :23.18   Max.   :28.50   Max.   :744.0   Max.   :1139.0
##  HumidityRatio        Occupancy
##  Min.   :0.002824   Min.   :0.0000
##  1st Qu.:0.003375   1st Qu.:0.0000
##  Median :0.003905   Median :0.0000
##  Mean   :0.003836   Mean   :0.2775
##  3rd Qu.:0.004343   3rd Qu.:1.0000
##  Max.   :0.004817   Max.   :1.0000
```

(b) Build a logistic regression model by only using the Temperature feature to predict the room occupancy.

Display the confusion matrix and the predictive accuracy obtained on the testing dataset.

```
glm.temp.fit <- glm(Occupancy ~ Temperature,data=training_data,family = binomial)
probability1 <- predict(glm.temp.fit, newdata= testing_data, type="response")
prediction1<- rep(0,nrow(testing_data))
prediction1[probability1>0.5]<-1
CM <- table(prediction = prediction1, truth = testing_data$Occupancy)
print(CM)
```

```
##           truth
## prediction   0   1
##          0 182  39
##          1  58  21
```

####(c) Build a logistic regression model by only using the Humidity feature to predict the room occupancy. ####D isplay the confusion matrix and the predictive accuracy obtained on the testing dataset.

```
glm.humid.fit <- glm(Occupancy ~ Humidity,data=training_data,family = binomial)
probability2 <- predict(glm.humid.fit, newdata= testing_data, type="response")
prediction2<- rep(0,nrow(testing_data))
prediction2[probability2>0.5]<-1
CM2 <- table(prediction = prediction2, truth = testing_data$Occupancy)
print(CM2)
```

```
##           truth
## prediction   0   1
##          0 133  22
##          1 107  38
```

(d) Build a logistic regression model by using all features to predict the room occupancy. Display the

confusion matrix and the predictive accuracy obtained on the testing dataset.

```
glm.all.fit <- glm(Occupancy ~ Temperature+Humidity+Light+CO2+HumidityRatio,data=training_data,family =

probability3 <- predict(glm.all.fit, newdata= testing_data, type="response")
prediction3<- rep(0,nrow(testing_data))
prediction3[probability3>0.5]<-1
CM3 <- table(prediction = prediction3, truth = testing_data$Occupancy)
print(CM3)
```

```
##           truth
```

```
## prediction   0   1
##          0 184  13
##          1  56  47
```

(e) Compare the predictive performance of three different models by drawing ROC curves and calculating

the AUROC values. Discuss the comparison results.

ROC Curve and AUROC values for prediction based on Temperature

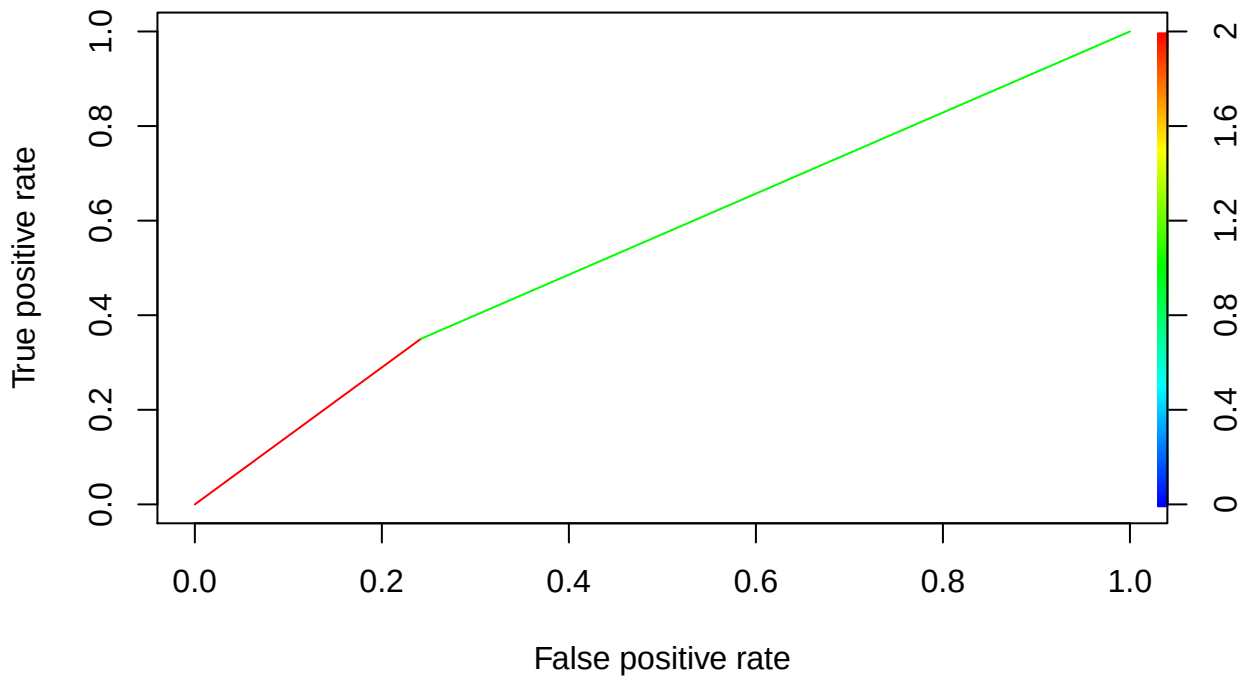
```
#install.packages("ROCR")
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

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

```
pred <- prediction(prediction1, testing_data$Occupancy)
perf <- performance(pred, measure="tpr", x.measure = "fpr")
plot(perf,colorize=TRUE)
```

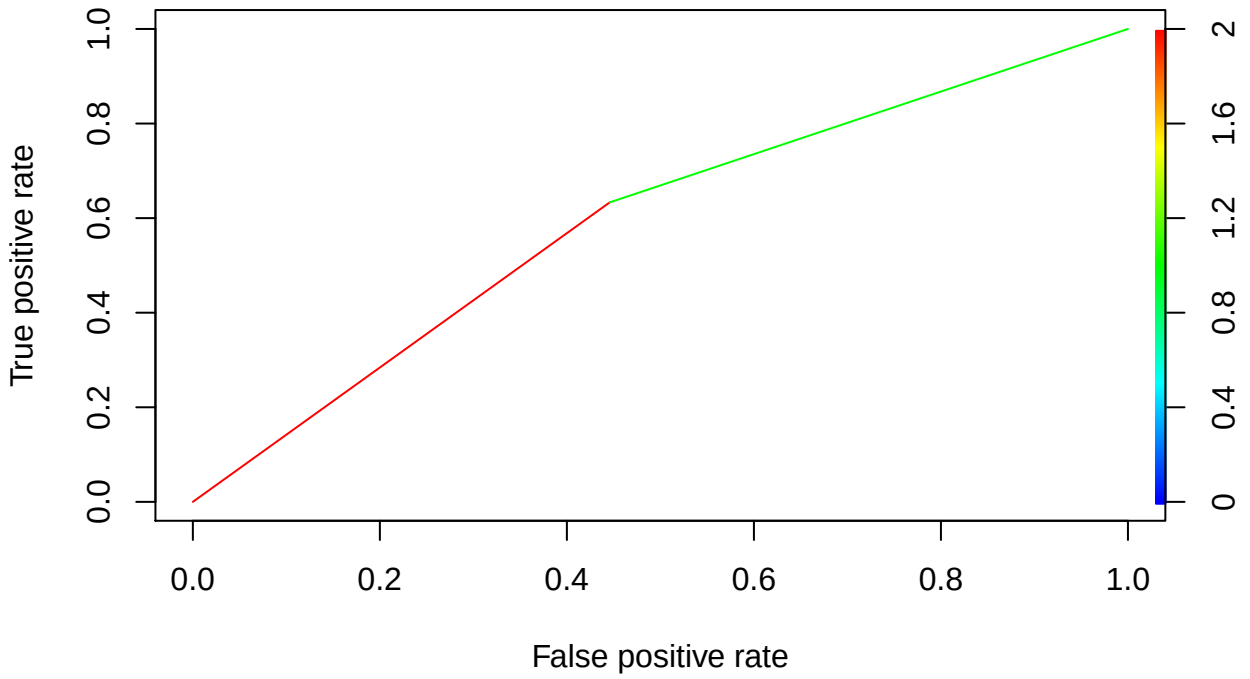



```
auroc <- performance(pred, measure="auc")
auroc_value <- auroc@y.values[[1]]
auroc_value
```

```
## [1] 0.5541667
```

ROC Curve and AUROC values for prediction based on Humidity

```
pred2 <- prediction(prediction2, testing_data$Occupancy)
perf2 <- performance(pred2, measure="tpr", x.measure = "fpr")
plot(perf2,colorize=TRUE)
```

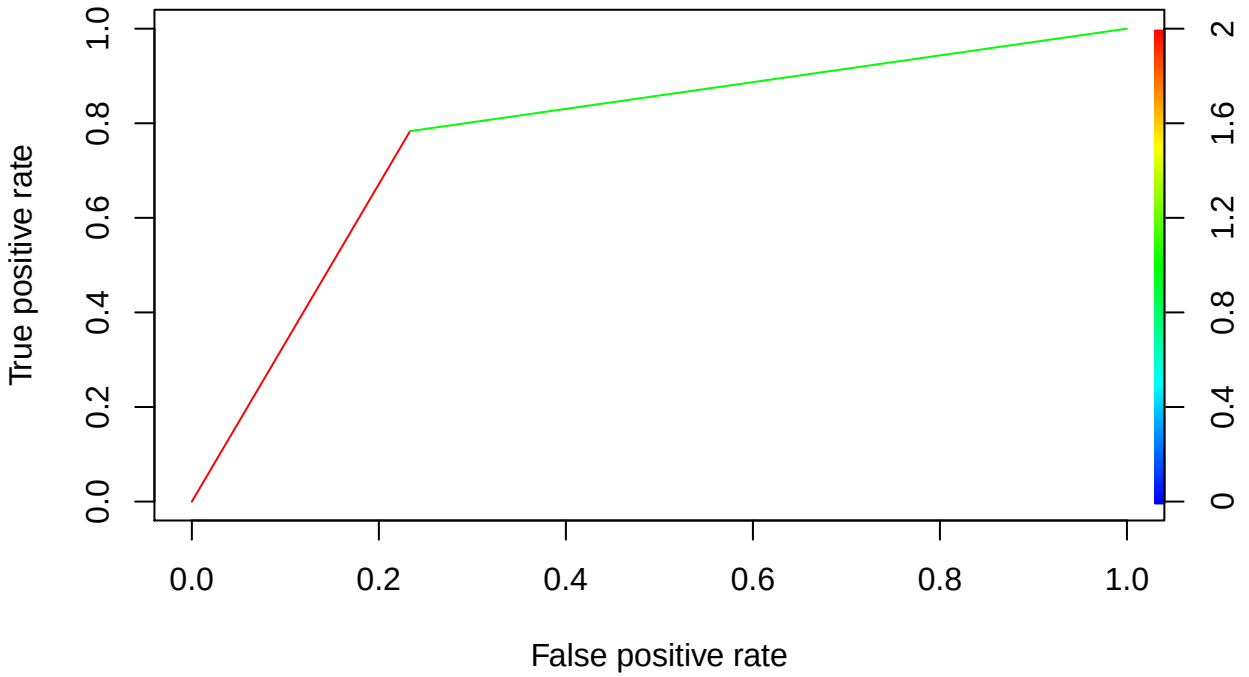



```
auroc2 <- performance(pred2, measure="auc")
auroc_value2 <- auroc2@y.values[[1]]
auroc_value2
```

```
## [1] 0.59375
```

ROC Curve and AUROC values for prediction based on all available variables

```
pred3 <- prediction(prediction3, testing_data$Occupancy)
perf3 <- performance(pred3, measure="tpr", x.measure = "fpr")
plot(perf3,colorize=TRUE)
```




```
auroc3 <- performance(pred3, measure="auc")
auroc_value3 <- auroc3@y.values[[1]]
auroc_value3
```

```
## [1] 0.775
```

From the values of AUROC and ROC curves, we can see that the predictive performance of logistic regression model to predict the occupancy is best when all features are assessed; therefore the AUROC value of the last model is largest at 0.775 and the area under the ROC curve is largest for that model hence it is the best predictive model amongst the 3 we observed here.