Homework 3

Snigdha Peddi

Question 1:Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

Logistic function from 4.2

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Logit function from 4.3

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

Below equations shows that both the logistic fucntion representation and logit representation are equal,

Subtracting both side of logistic function equation by 1,

$$1 - p(X) = 1 - (\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}})$$

$$1 - p(X) = \frac{1+e^{\beta_0+\beta_1 X} - e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Cancelling out the $e^{\beta_0+\beta_1 X}$,

$$1 - p(X) = \frac{1}{1+e^{\beta_0+\beta_1 X}}$$

Rearranging the equation,

$$\frac{1}{1-p(X)} = 1 + e^{\beta_0+\beta_1 X}$$

Multiplying with the $p(X)$ from logistic function (4.2) on both sides,

$$\frac{p(X)}{1-p(X)} = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}(1+e^{\beta_0+\beta_1 X})$$

Cancelling out $(1+e^{\beta_0+\beta_1 X})$ will equal to,

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

Hence, proving both logistic function (4.2) representation and logit (4.3) representation of logistic regression are equal.

Question 2:This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

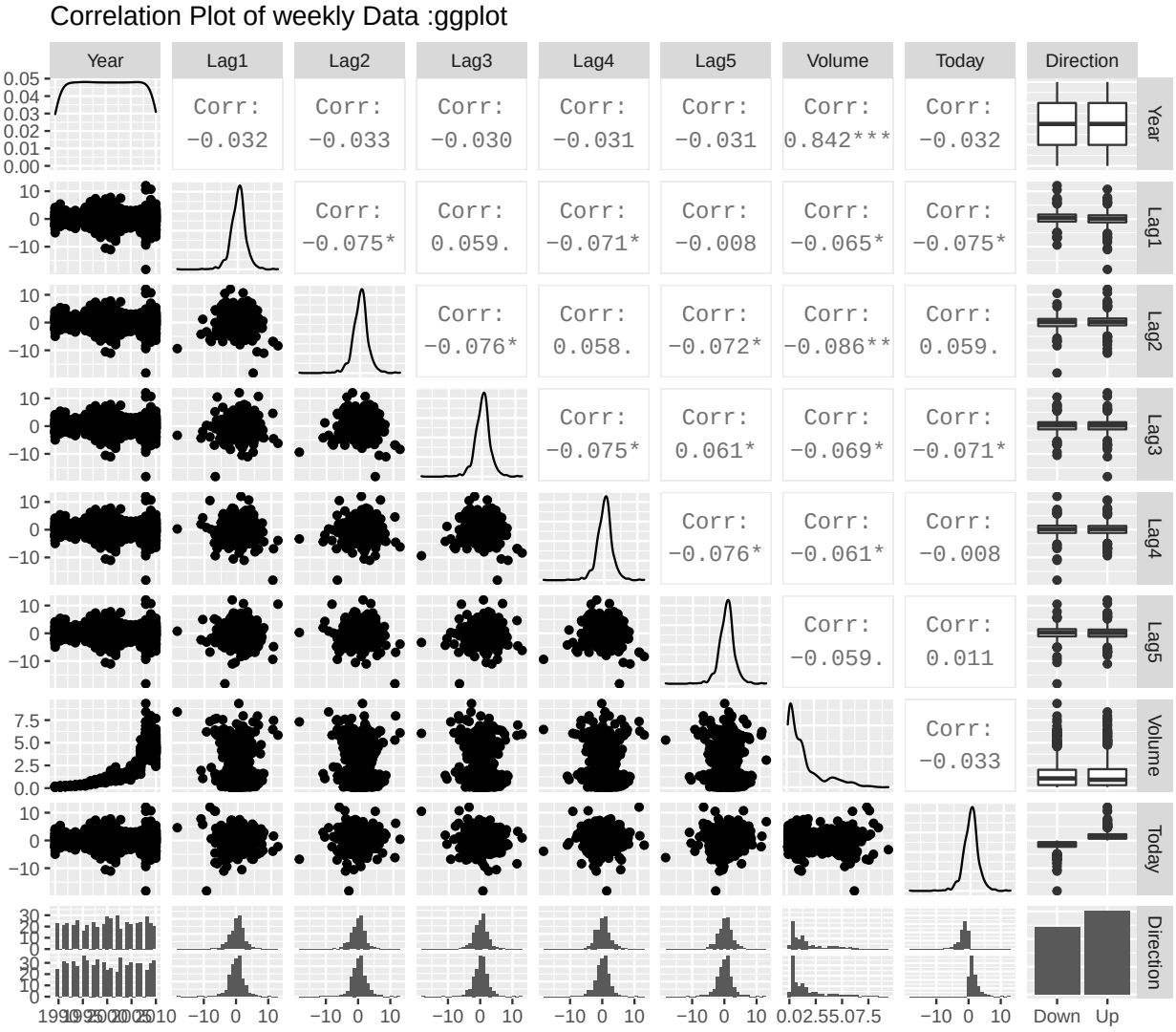
```
##
##  Dimensions of Weekly dataset: 1089 9
```

```
##
##  Number of missing values in Weekly dataset: 0
```

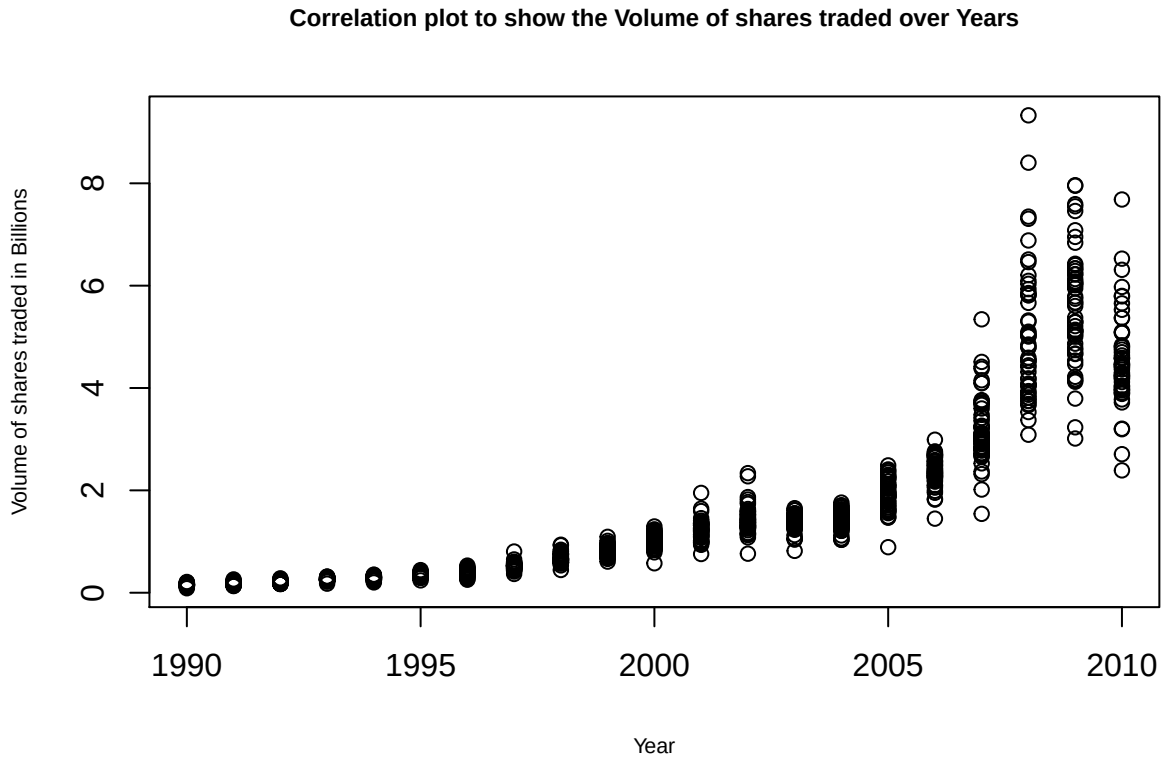
2.a Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

```
##
##  Summary of weekly Data:
```

```
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
##       Lag4               Lag5              Volume            Today
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```



The correlation plot shows that there is approximately 0 correlation between the Year and all other features except Volume variable which is 0.84.The correlation between the Volume of stocks traded (in Billions) over years can be clearly seen in the plot below.



2.b Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

The below model is fit between Direction as response variable and 5 lag variables and volume variable as predictors.

```
reg.mod1<-glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,
          data=weekly,family=binomial)
```

The summary of the model indicates that only Lag2 variable is statistically significant with a lower p-value of 0.0296.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
```

```
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

2.c Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

The predictions for the weekly data are made using the logistic regression model created using Direction as response variable and 5 Lag variables and Volume variable as predictors.The probability values predicted corresponds to the market going. This can be explained by the *contrasts()* function which shows that R has created a dummy variable with a 1 for UP Direction.Then a vector is created where the values with all the probabilities greater than 0.5 are named as "UP" and less than 0.5 are named as "Down". A *table()* function is used to create a confusion matrix to determine the accuracy of the prediction.

```
##
##  Contrasts of Direction Variable:

##      Up
## Down  0
## Up    1

##
## pred.mod1 Down  Up
##      Down   54  48
##      UP    430 557

##
##  Accuracy of prediction: 56.11 %
```

The model has correctly predicted that the market would go down 54 days and it would go up 557 days. It gave an accuracy of 56.11%.In other words there is an 43.89% Training error.

2.d Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

A new Logistic Regression Model is fit with Lag2 as predictor variable and Direction as Response variable using the weekly data from Years 1990 to 2008(train data). The weekly data for Years 2009 and 2010 is used as a test data.

```
reg.mod2<-glm(Direction~Lag2,data=weekly,
          family=binomial,subset=train)
```

Summary of the model shows that Lag2 variable is statistically significant with a p value of 0.04298.

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = weekly,
##     subset = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

Predictions were made on test data using the above logistic regression model.The confusion matrix indicate that there is an 62.5% accuracy in predictions where the model correctly predicts that the market goes Up 56 days and goes Down 9 days out of 61 days and 43 days respectively. However, there is still a test error of 37.5%.

```
##
## pred.mod2 Down Up
##      Down    9  5
##      UP     34 56
```

```
##
##  Accuracy of prediction: 62.5 %
```

```
##
##  Test Error: 37.5 %
```

REFERENCES

- Chapter 4,Classification,*An Introduction to Statistical Learning with Applications in R* by Gareth James.

Question 3: In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

```
##
##  Dimensions of dataset: 392 9
```

```
##
##  Number of missing values in dataset: 0
```

```
##
##  Summary of Auto Data:
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##
##   acceleration        year           origin                      name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
##                                                  (Other)           :365
```

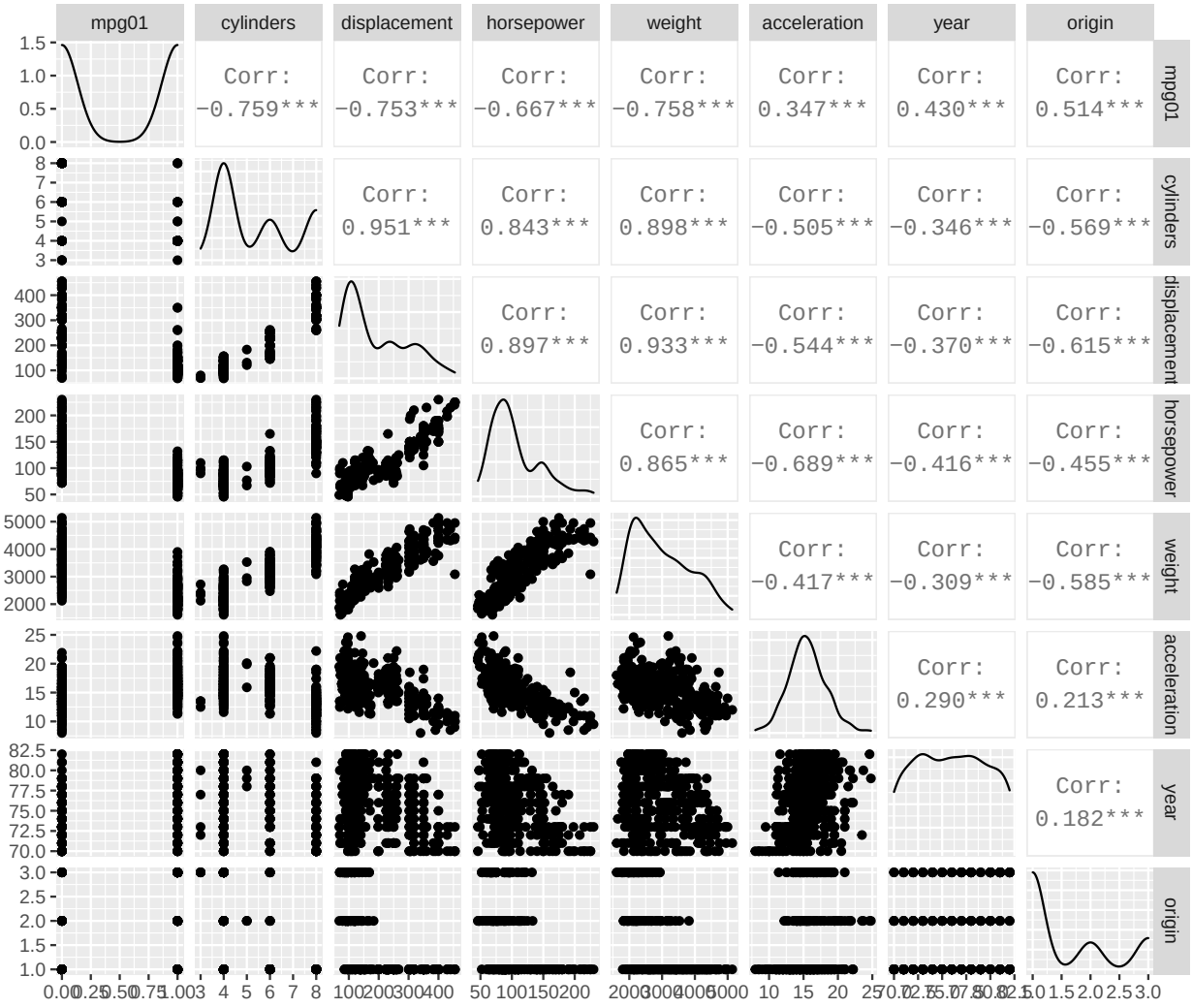
3.a:Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

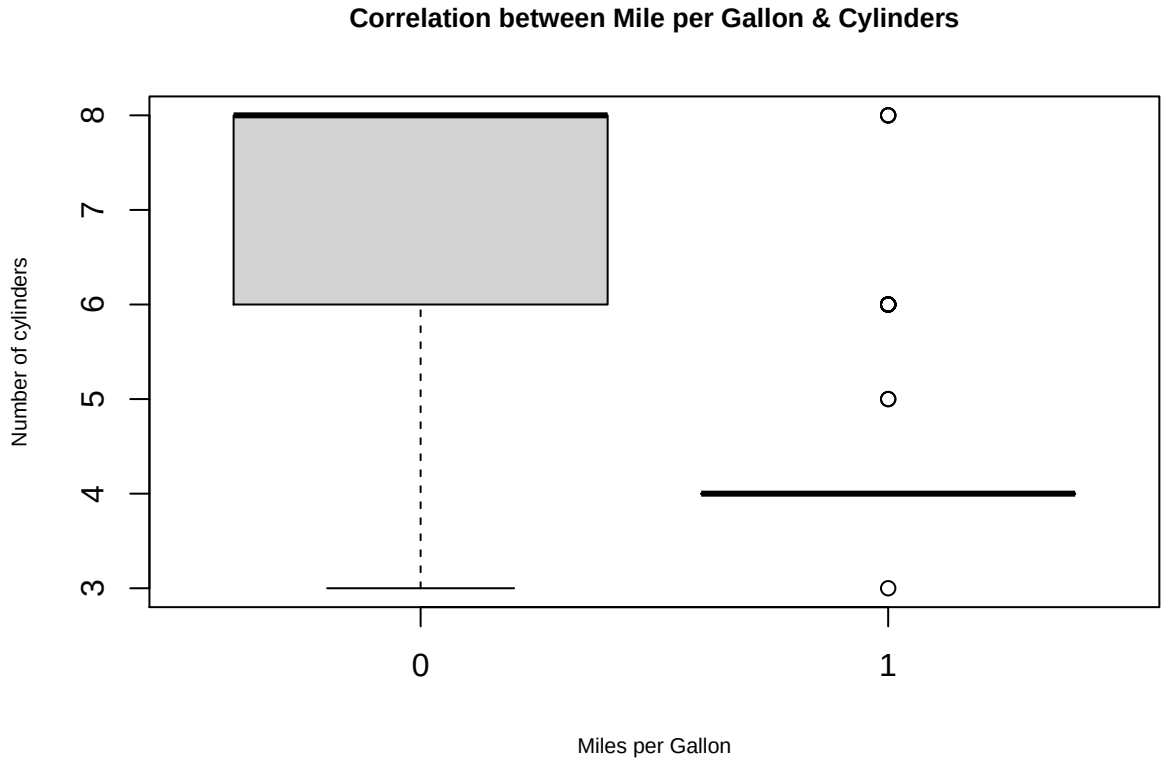
```
##
##  Summary of New Auto Data:
```

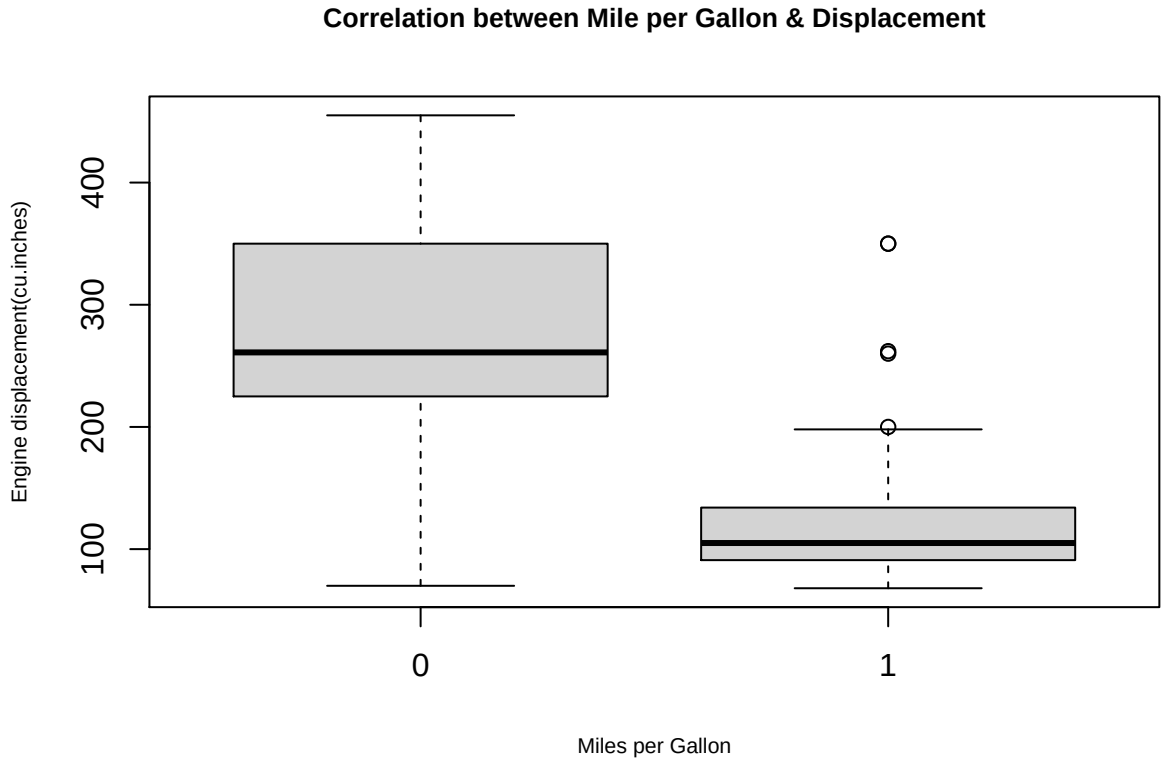
```
##      mpg01       cylinders      displacement     horsepower        weight
##  Min.   :0.0   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:0.0   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :0.5   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :0.5   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:1.0   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :1.0   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##
##   acceleration        year           origin                      name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
##                                                  (Other)           :365
```

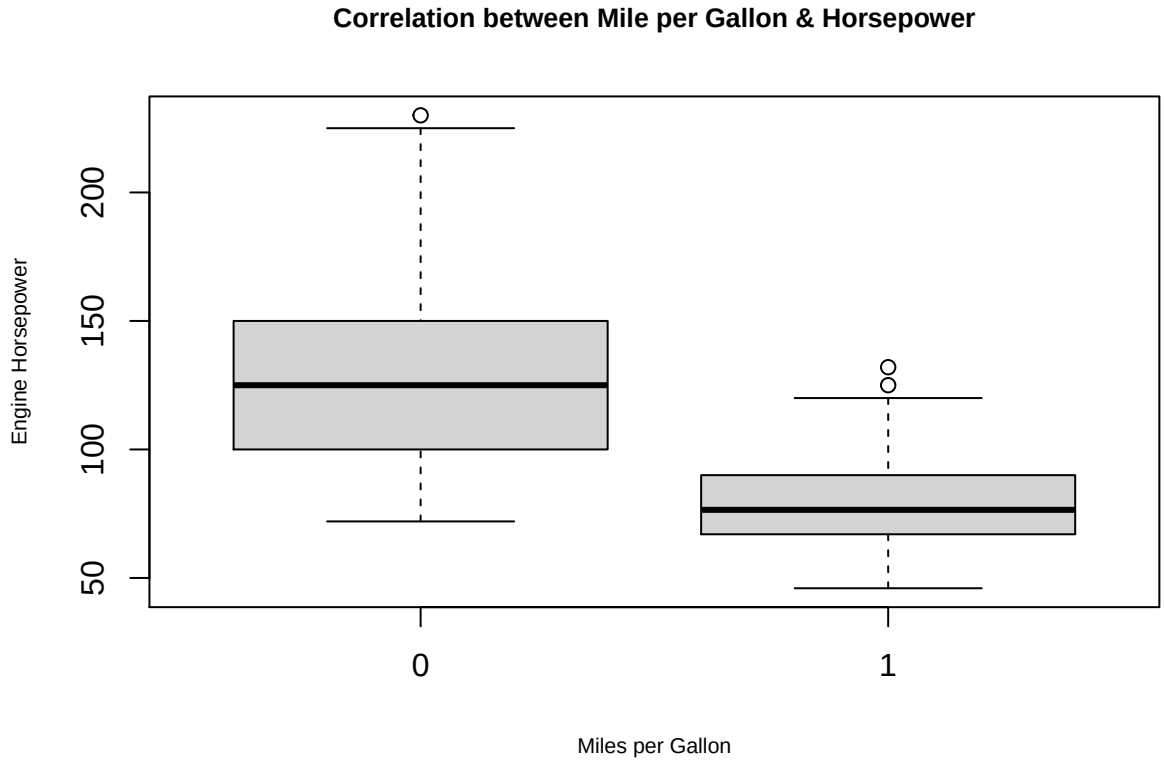
3.b:Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

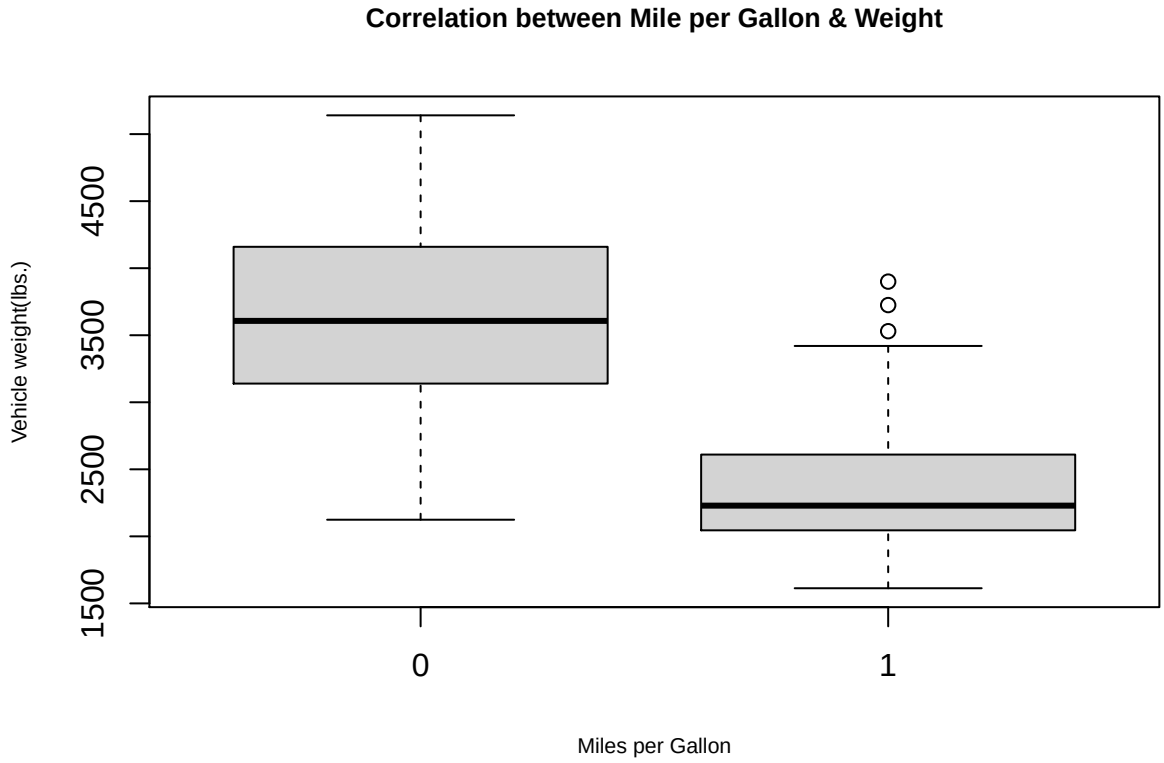
Correlation Plot of New Auto Data :ggplot

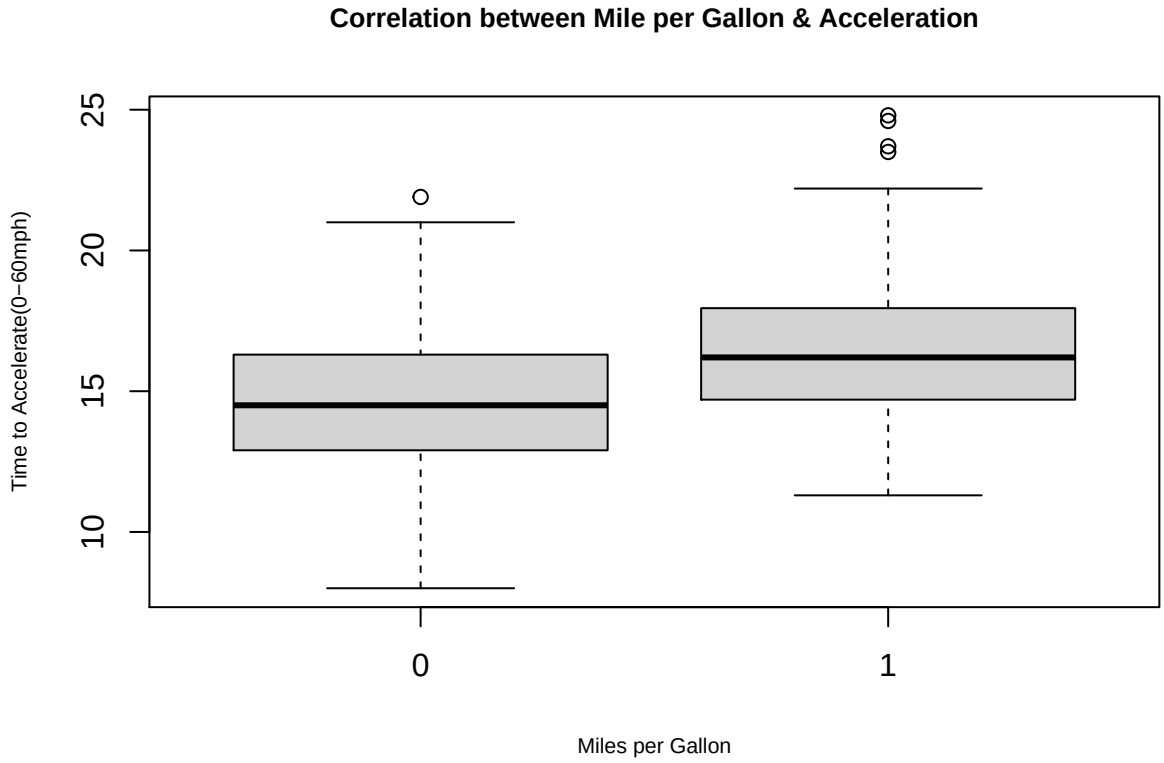


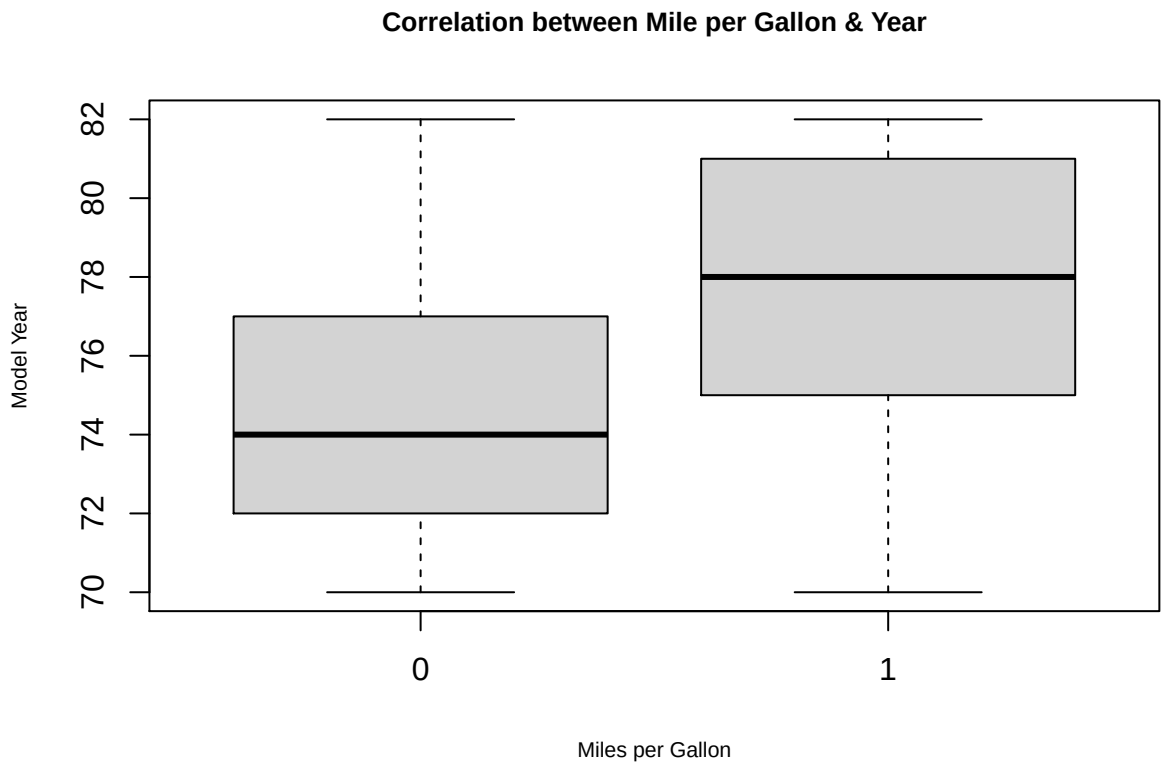

Correlation between Mile per Gallon & Cylinders
Number of cylinders
3
4
5
6
7
8
0
1
Miles per Gallon


Correlation between Mile per Gallon & Displacement
Engine displacement(cu.inches)
100
200
300
400
0
1
Miles per Gallon


Correlation between Mile per Gallon & Horsepower
Engine Horsepower
50
100
150
200
0
1
Miles per Gallon


Correlation between Mile per Gallon & Weight
Vehicle weight(lbs.)
1500
2500
3500
4500
0
1
Miles per Gallon


Correlation between Mile per Gallon & Acceleration
Time to Accelerate(0−60mph)
10
15
20
25
0
1
Miles per Gallon



From the correlation plot and the box plot it is clear that there is a correlation between miles per gallon and number of cylinders(4 Cylinders have higher mpg compared to 8 cylinder cars that have lower mpg),displacement(lower displacement-higher mpg and higher displacement-lower mpg),horsepower(lower horsepower-higher mpg and higher horsepower-lower mpg),weight(lower weight-higher mpg and higher weight-lower mpg).

3.c:Split the data into a training set and a test set.

A 70:30 split was made for training and test data.

```
##
## Size of Training Data: 274
```

```
## Size of Test Data: 118
```

3.f:Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

A Logistic regression model is fit for training data using the variables cylinders,displacement,horsepower,weight as predictors and mpg01 variable as response variable.

```
reg.mod3<-glm(mpg01~cylinders+displacement+horsepower+weight,
          data=train1, family=binomial)
```

```
##
```

```
## Call:
## glm(formula = mpg01 ~ cylinders + displacement + horsepower +
##     weight, family = binomial, data = train1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.34637  -0.20995  -0.00221   0.31924   2.99393
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.6438919  2.1114523   5.515  3.5e-08 ***
## cylinders     0.2248937  0.4051382   0.555  0.57882
## displacement -0.0171093  0.0098817  -1.731  0.08338 .
## horsepower   -0.0542435  0.0173610  -3.124  0.00178 **
## weight       -0.0016724  0.0009096  -1.839  0.06599 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.32  on 273  degrees of freedom
## Residual deviance: 137.61  on 269  degrees of freedom
## AIC: 147.61
##
## Number of Fisher Scoring iterations: 7
```

```
##          mpg01.test
## pred.mod3  0  1
##         0 45  5
##         1  8 60
```

```
##
##  Accuracy of prediction: 88.98 %
```

```
##
##  Test Error of the model: 11.02 %
```

Predictions were made on test data using logistic regression model.The confusion matrix indicate that there is an 88.98% accuracy in predictions and a test error of 11.02%.

Question 4. Write a reusable function in RMD that calculates the misclassification rate, sensitivity, and specificity, and return a table similar to `Table 4.7`. Call this function `misclass.fun.*`, replacing `*` with your initials. The arguments for this function are a threshold, predicted probabilities, and original binary response data. Test your function using the data and model from 4.7.10 b) with threshold values of `c(0.25, 0.5, 0.75)`.

Definitions:

- *True Positive:*When a true value is positive and predicted positive, its a True Positive.
- *True Negative:*When a true value is negative and predicted negative, its a True Negative.
- *False Positive:*When a true value is negative and predicted positive its a False Positive.
- *False Negative:*When a true value is positive and predicted negative its a False Negative.
- *Misclassification Rate:*The rate of incorrectly identified predictions.

- *Sensitivity:*It is proportion of samples that test Positive using the test in question that are genuinely positive. It is also called as True Positive Rate.It is given by ratio of True Positive values to True Positive and False Negative values.
- *Specificity:* It is proportion of samples that test Negative using the test in question that are genuinely Negative. It is also called as True Negative Rate.It is given by ratio of True Negative values to True Negative and False Positive values.

Below is the reusable Function created for Misclassification Rate,Specificity and Sensitivity

```
misclass.fun.SP <- function(predicted,actual,threshold=0.5){
  predictied.values<- ifelse(predicted >= threshold,1,0)

  TP <- sum(ifelse(actual == 1 & predictied.values == 1,1,0))
  TN <- sum(ifelse(actual == 0 & predictied.values == 0,1,0))
  FP <- sum(ifelse(actual == 0 & predictied.values == 1,1,0))
  FN <- sum(ifelse(actual == 1 & predictied.values == 0,1,0))

  misclassfication.rate <- ((FP+FN)/(TP+TN+FP+FN))*100
  sensitivity <- TP/(TP+FN)
  specificity <- TN/(TN+FP)

  Info.Table<- c('Misclassification Rate'=misclassfication.rate,                    'Sensitivity'=sensiti
                 'Specificity'=specificity)
  return(Info.Table)
}
```

The test data is predicted at different thresholds for Weekly data(2.d)and the comparison of Misclassification rate(%),Specificity,Sensitivity ware presented in below table.

Table 1: Comparision at Different Thresholds

	0.25	0.5	0.75
Misclassification Rate	41.346	37.500	58.654
Sensitivity	1.000	0.918	0.000
Specificity	0.000	0.209	1.000

REFERENCES

- Blog post by Karen Steward PhD,*Sensitivity vs Specificity* April 16,2019.
- Blog post by Stephanie Glen, Statistics How To,*Sensitivity vs Specificity and Predictive Value* .