Homework 3

Prakash Paudyal

Jan 4 , 2018

Due February 6th

Please do the following problems from the text book ISLR.

$$\mathbf{1.}\ \boldsymbol{Question 4.7.1 pg 168}$$

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.

$$We\ have\ given\ equation\ 4.2\ is\ [p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \Leftrightarrow e^{\beta_0+\beta_1 X}(1-p(X)) = p(X),]which\ is\ equivalent\ to\ [\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.]that\ is\ equation\ 4.3$$

$$**\ \mathbf{2.}\ \boldsymbol{Question 4.7.10(a-d) pg 171}\ **$$

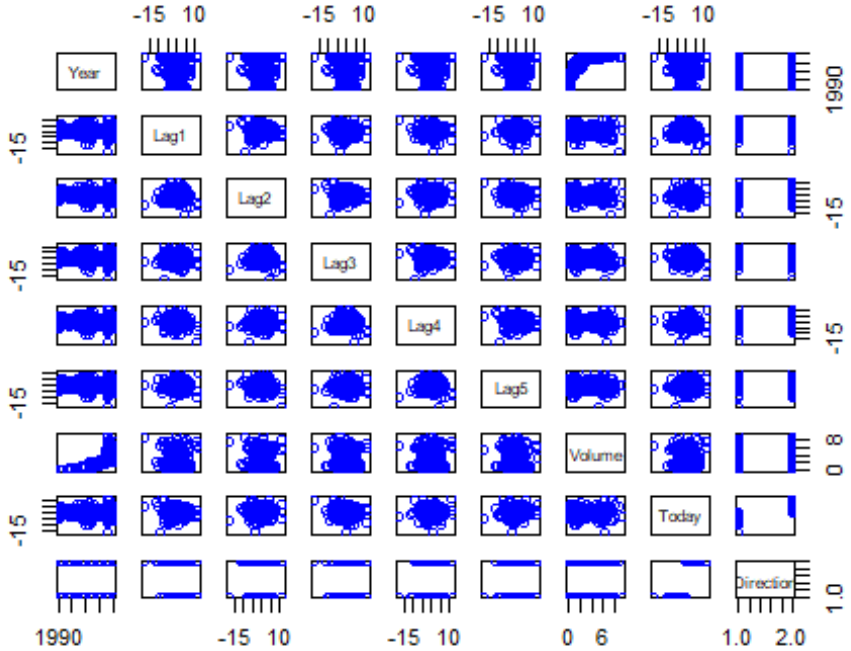
10. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

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

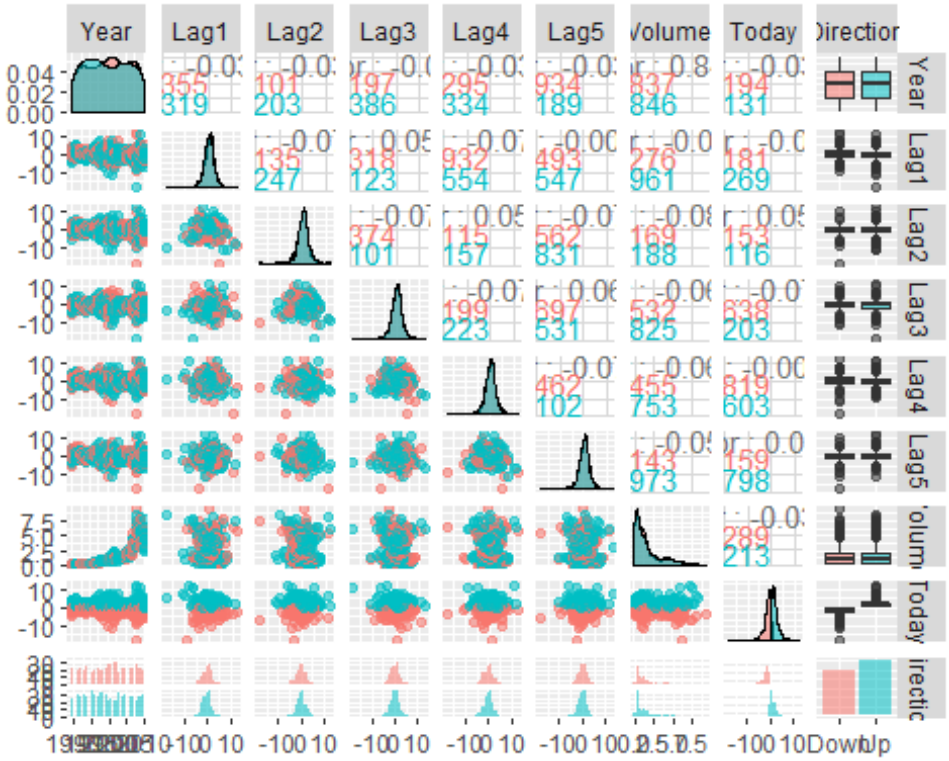
```
##       Year           Lag1                Lag2                Lag3          
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  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       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
```

```
##  Min.   :-18.1950   Down:484
##  1st Qu.: -1.1540   Up  :605
##  Median :  0.2410
##  Mean   :  0.1499
##  3rd Qu.:  1.4050
##  Max.   : 12.0260
```

```
##                Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```



ggplot



Comment:

From the correlation table and scaterplot matrix, we can see that variable "year" has correlation with volume. Volume is increasing over the year.Pirwaise plot and correlaton table indicate that all previous week returns (lags) do not show any correlation with other variables.Today's rate of return also does not show any correlations

(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?

Fit a model

```
##
## 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
```

Comment:

lag2 is the only significant predictor for direction with less p-value at 5% confidence level.

(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.

```
##
## p_class Down  Up
##    Down   54  48
##    Up    430 557

## [1] 0.5610652
```

Comment:

The confusion matrix tells us,model's correct prediction fraction **(54+557) / (total no of obs)** that is 56.11%. The table reveals that Logestic model predicted (430+557)=987 times rate of return will go up.Of those, 557 went up but 430 of them went down. Hence 430 out of 484 were incorrectly labled (88.8%).

when actual value is going up ,**557/ (48+557)557/(48+557))=92%** of times prediction was correct.

when actual value is going down, **(54/(54+430)54/(54+430))=11.1%** of times prediction was correct.

(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).

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9  5
##       Up     34 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.2439
##
##                   Kappa : 0.1414
##  Mcnemar's Test P-Value : 7.34e-06
##
##             Sensitivity : 0.20930
##             Specificity : 0.91803
##          Pos Pred Value : 0.64286
##          Neg Pred Value : 0.62222
##              Prevalence : 0.41346
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.13462
##       Balanced Accuracy : 0.56367
##
##        'Positive' Class : Down
##
```

correct predection fraction , Accuracy : 0.625

3. Question 4.7.11(a,b,c,f) pg 172

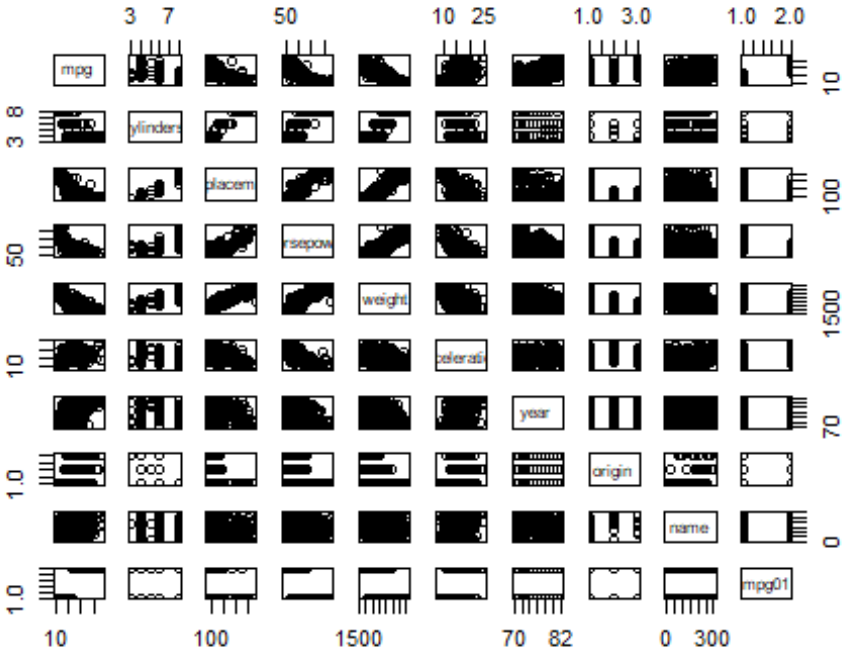
In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

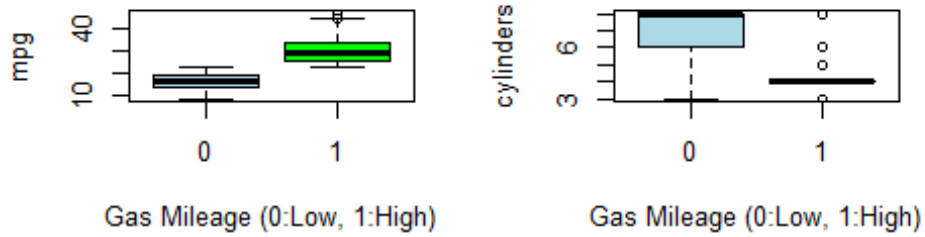
(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame()function to create a single data set containing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

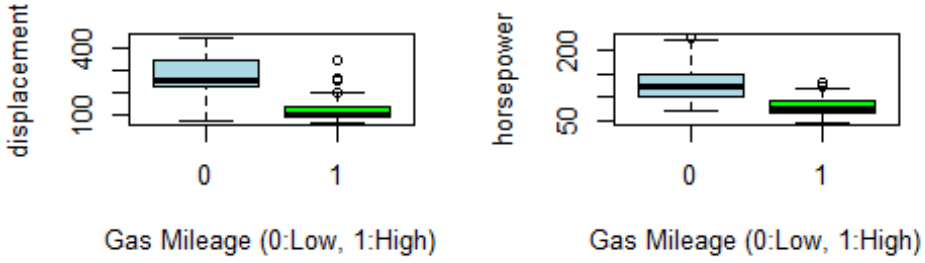
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
```

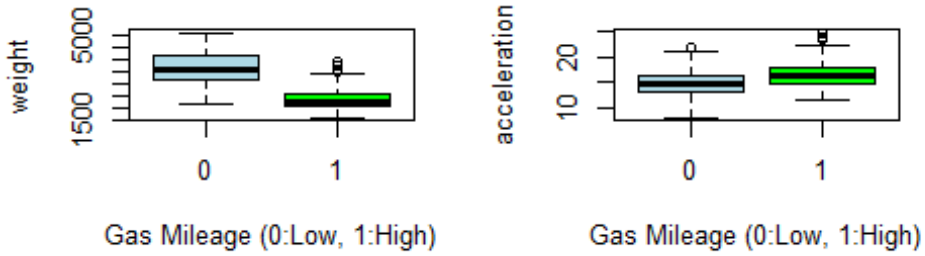
```
## 5                ford torino
## 6           ford galaxie 500
```

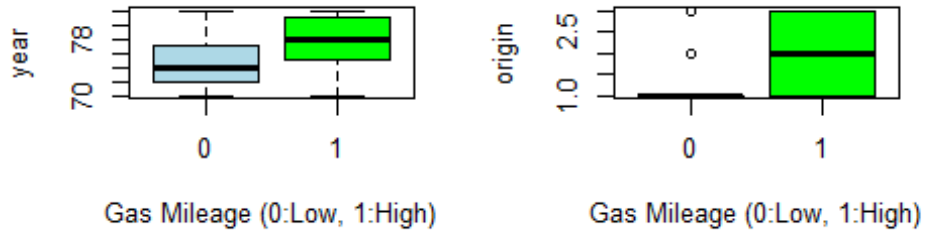
(b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.


mpg
cylinders
displacem
horsepow
weight
acceleratio
year
origin
name
mpg01


mpg
Gas Mileage (0:Low, 1:High)
cylinders
Gas Mileage (0:Low, 1:High)

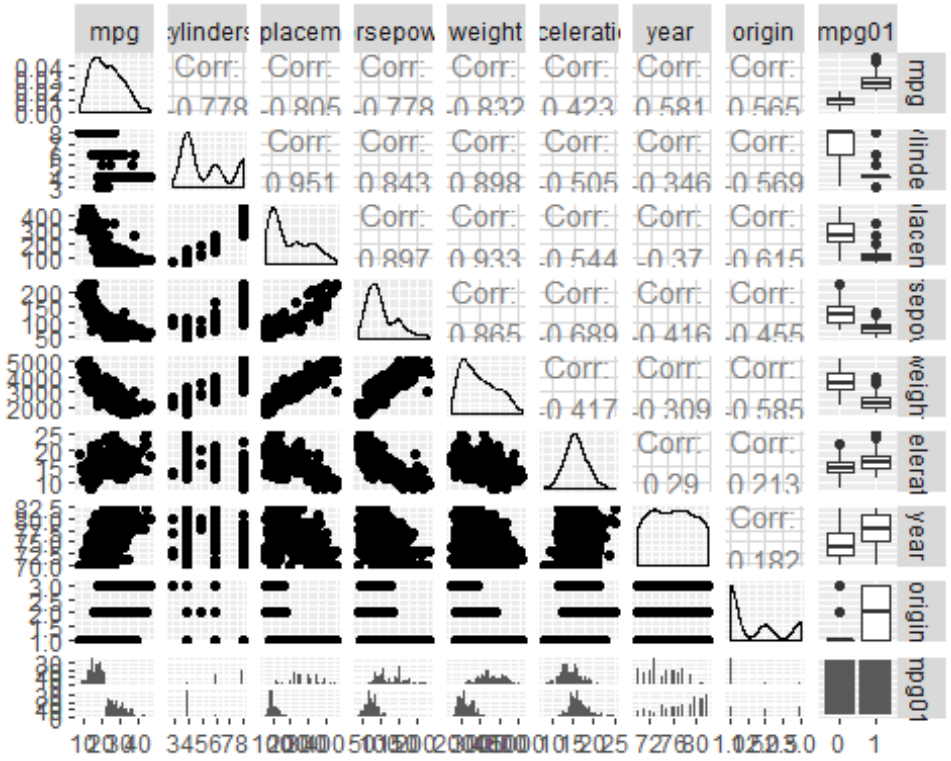

displacement
Gas Mileage (0:Low, 1:High)
horsepower
Gas Mileage (0:Low, 1:High)

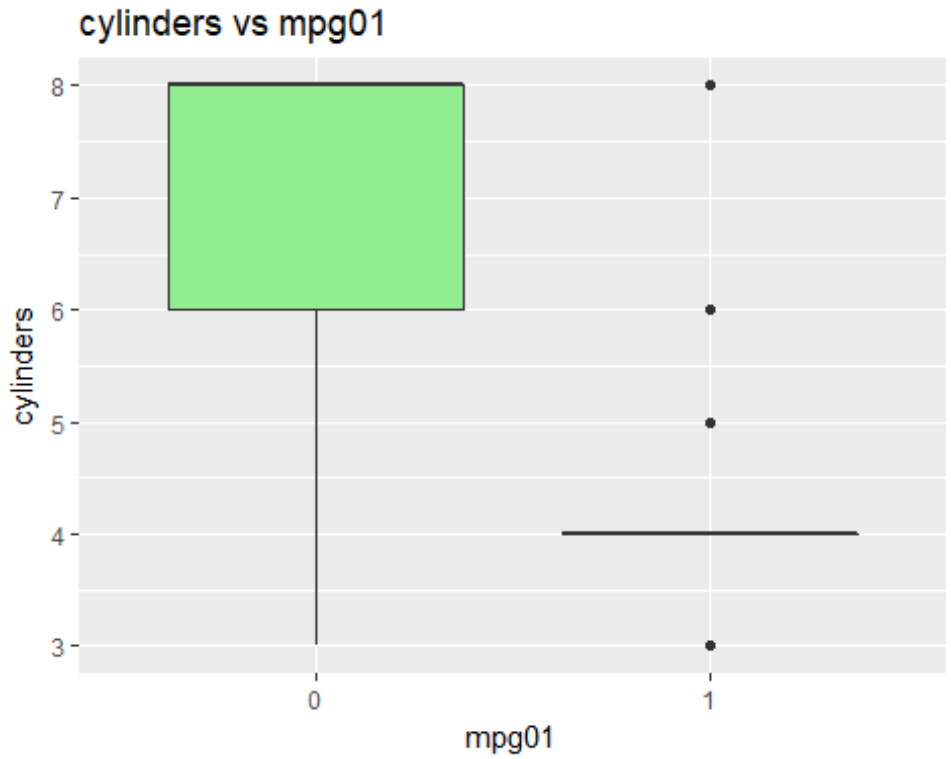


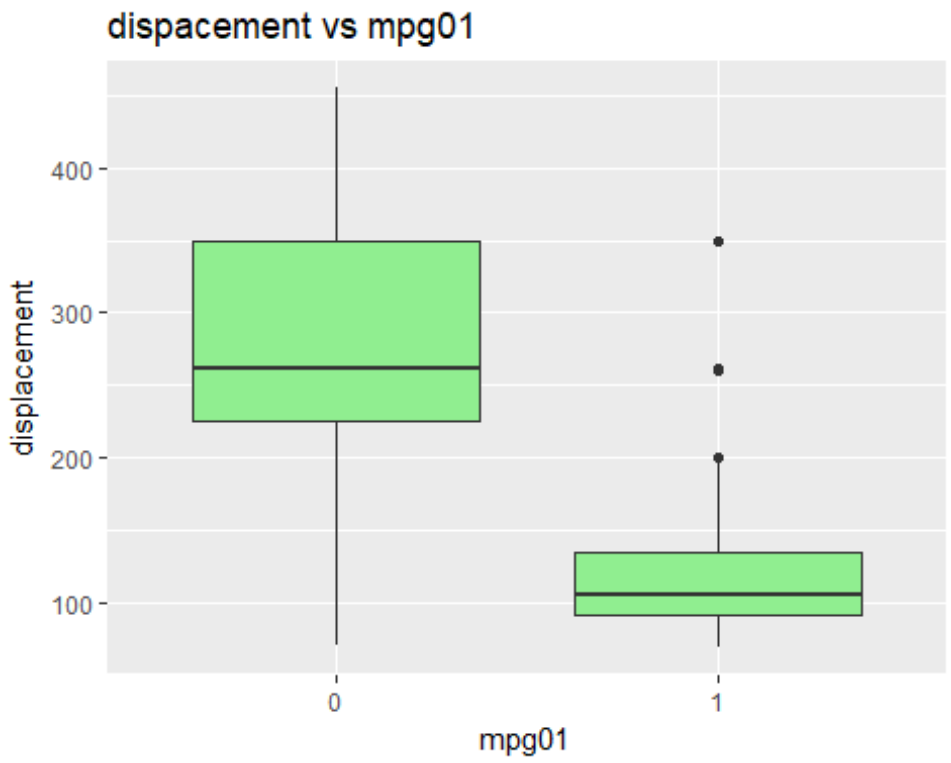


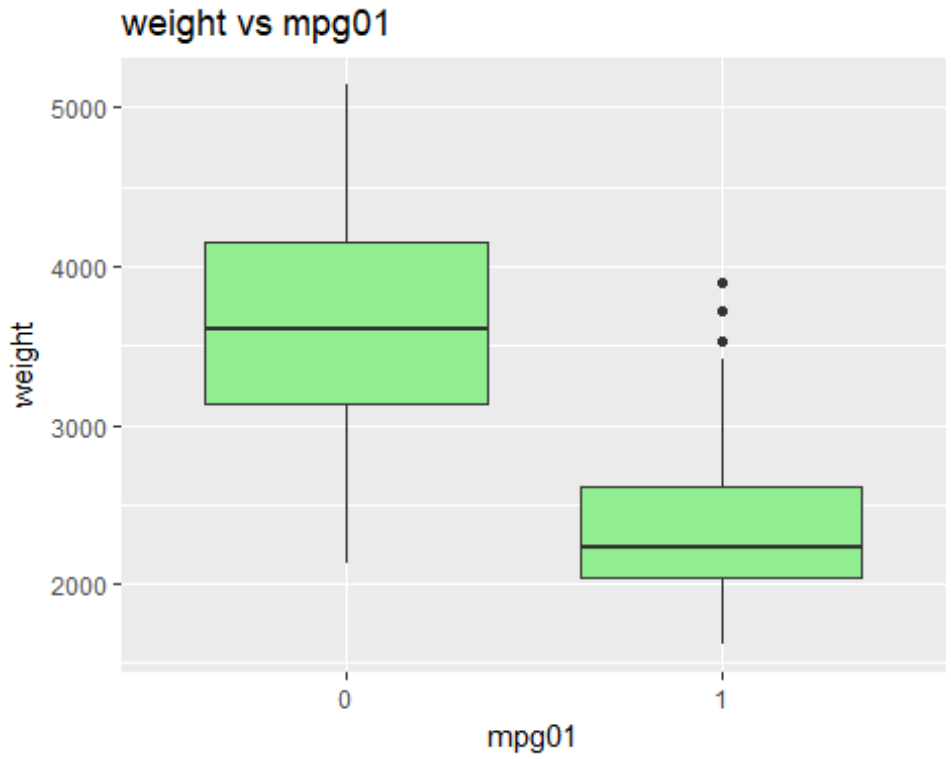
From the bar plot and correlation matrix, we can see gas mileage hase positive and negetive correlation with most of the variables. I would choose year, weight, displacement and horsepower.

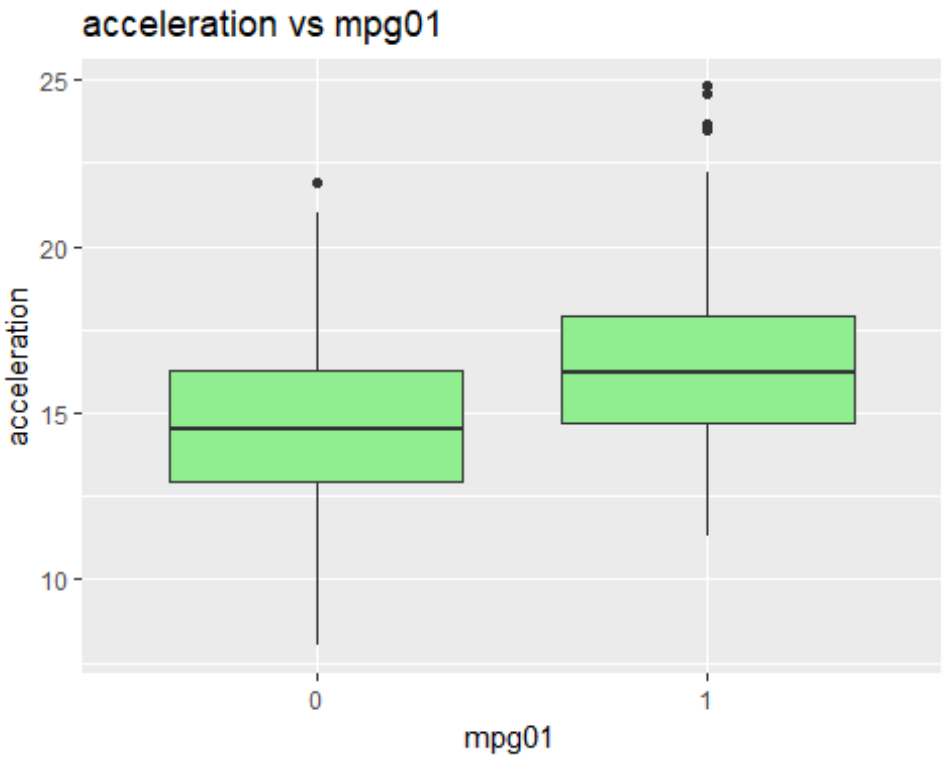
ggplot

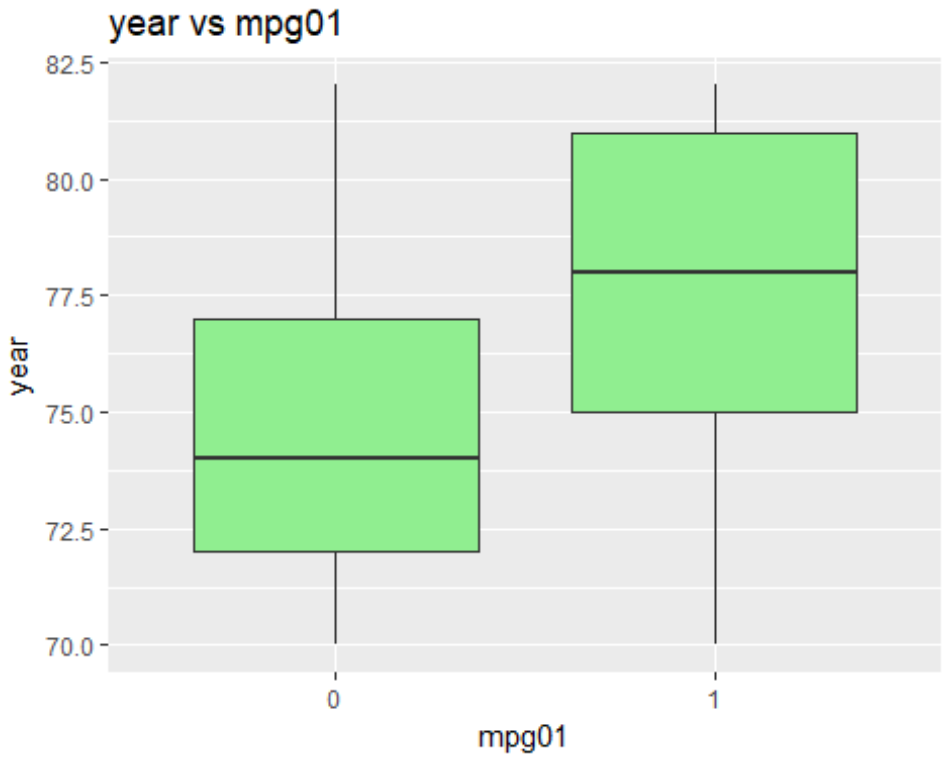

mpg
ylinders
placem
rsepow
weight
celerati
year
origin
mpg01
Corr:
-0.778
-0.805
-0.778
-0.832
0.423
0.581
0.565
0.951
0.843
0.898
-0.505
-0.346
-0.569
0.897
0.933
-0.544
-0.37
-0.615
0.865
-0.689
-0.416
-0.455
-0.417
-0.309
-0.585
0.29
0.213
0.182

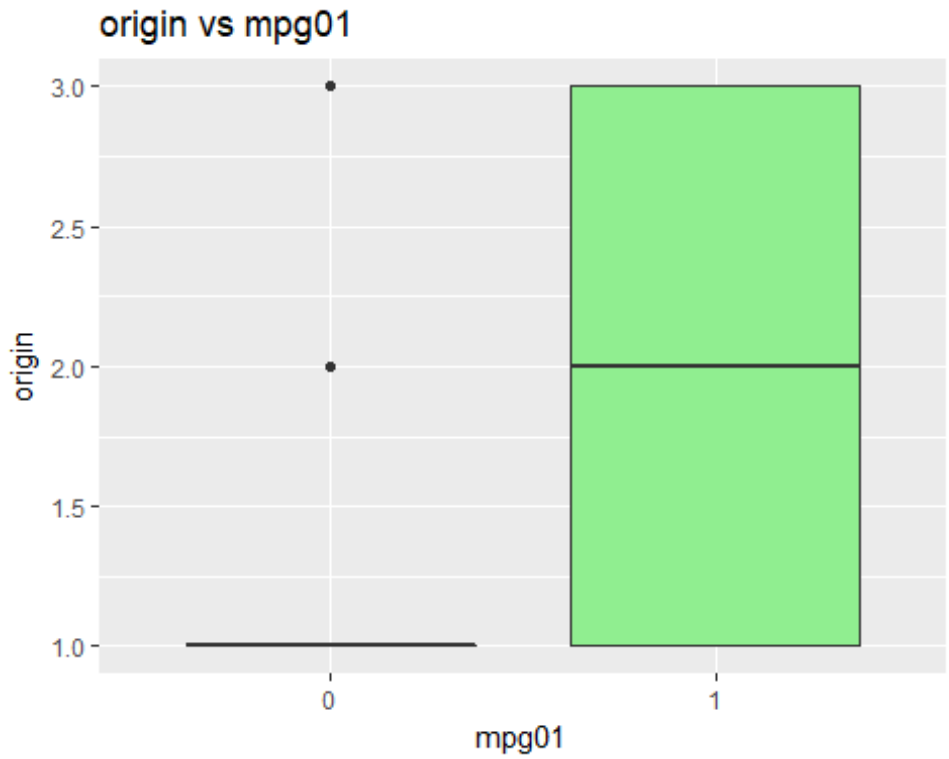

cylinders vs mpg01
cylinders
mpg01
8
7
6
5
4
3
0
1


dispacement vs mpg01
displacement
400
300
200
100
0
1
mpg01


weight vs mpg01
weight
5000
4000
3000
2000
0
1
mpg01


acceleration vs mpg01
25
20
15
10
acceleration
0
1
mpg01


year vs mpg01
82.5
80.0
77.5
75.0
72.5
70.0
year
0
1
mpg01



(c) Split the data into a training set and a test set.

```
## [1] 0.75
```

I split the data into 75:25 ratio.

(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?

```
##
## p_class2  0  1
##        0 46  2
##        1  9 41

## [1] 0.1122449
```

Test Error for the model is `0.1122449`

4. Write a function in RMD that calculates the misclassification rate, sensitivity, and specificity. The inputs for this function are a cutoff point, predicted probabilities, and original binary response. (Post any questions you might have regarding this on the discussion board.) (Needs to be an actual function you create, using the function() command, not just a chunk of code.)

```r
error_function<-function(p,o,cutoff=0.5)
{
  pred_class <- ifelse(p > .50, "1", "0")
  # Make simple 2-way frequency table
 confusion1<-table(pred_class, o)

 TP <- confusion1[2, 2]
 TN <-confusion1[1, 1]
 FP <- confusion1[2, 1]
 FN <- confusion1[1, 2]
 misclassificationrate<-mean(pred_class!=o)
 #misclassificationrate<-(FP + FN) / (TP + TN + FP + FN)
 #glm.sensitivity1 = round(length(which(pred_class == "1" & o == "1"))/length
(which(o == "1"))*100,2)
 sensitivity <-TP / (FN + TP)
 specificity <-TN / (TN + FP)
return(list(misclassificationrate=misclassificationrate,sensitivity=sensitivi
ty,specificity=specificity,confusion1))

}
error_function(prob2,test$mpg01, 0.5)
```

```
## $misclassificationrate
## [1] 0.1122449
##
## $sensitivity
## [1] 0.9534884
##
## $specificity
## [1] 0.8363636
##
## [[4]]
##           o
## pred_class  0  1
##          0 46  2
##          1  9 41
```