Homework #3

Justin Robinette

January 29, 2019

No collaborators for any problem

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 the logit representation for the logistic regression model are equivalent.

Results: *Logistic Function:*

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

Step 1:

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

Step 2:

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

Step 3:

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

Step 4:

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

Which gives us the same equation as 4.3

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

Question 4.7.10, pg 171: This question should be answered using teh **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, exceot that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

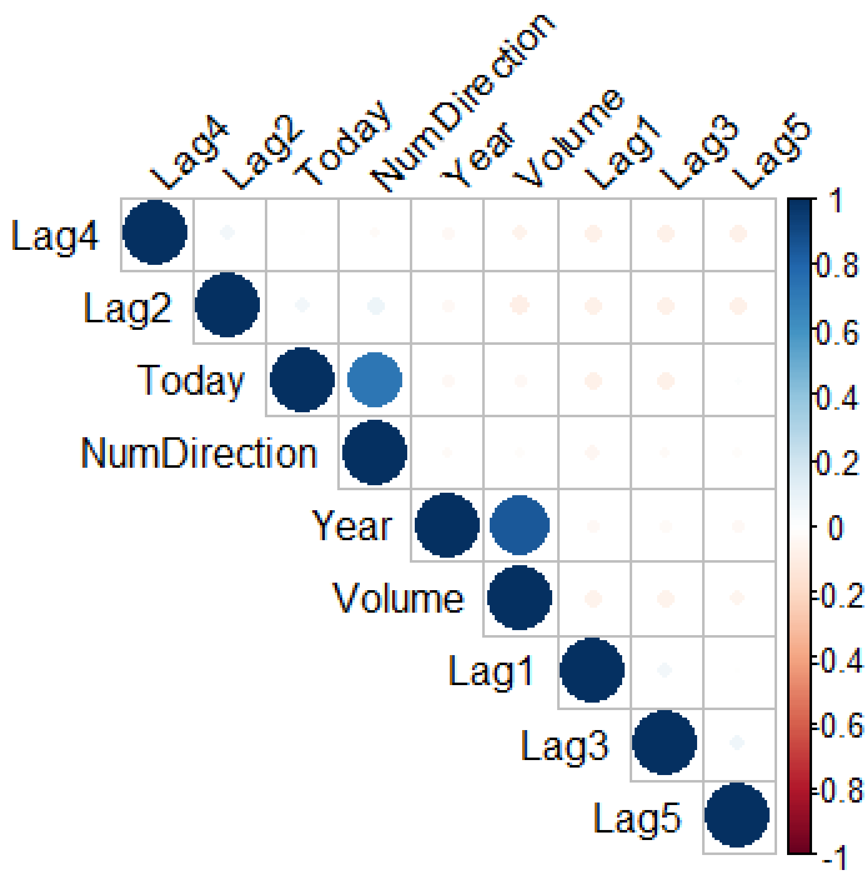
Part A: Produce some numerical and graphical summaries of the **Weekly** data. Do there appear to be any patterns?

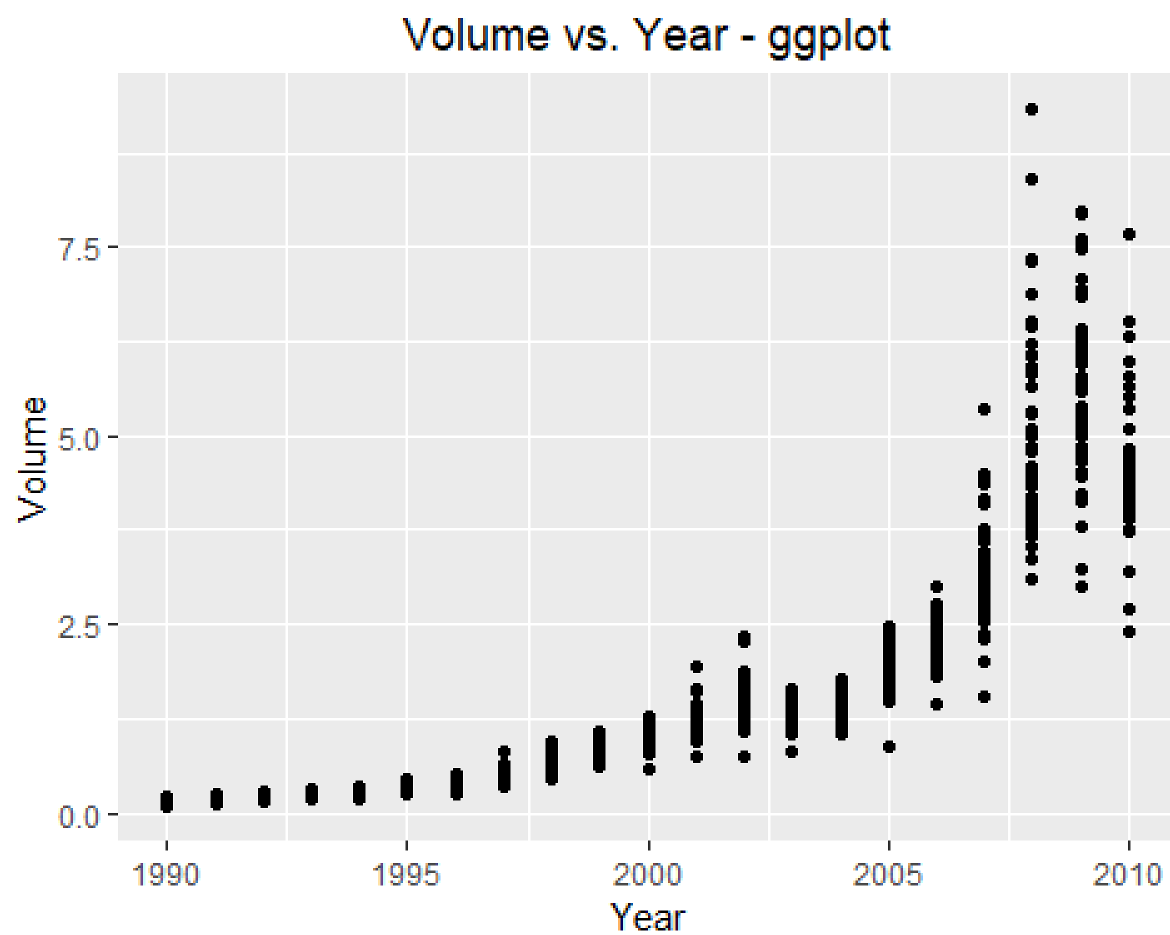
Results: First, I loaded the dataset and printed a summary to scan for NAs as well as examine the variables. Next, I created a numerical depiction of the Direction variable to better examine correlation between variables. I then printed a correlation matrix where we only see strong correlations between the Direction and Today, which is somewhat expected since we are measuring a the direction of the week, and between Volume and Year. A correlation plot was included which visually represents the data in the correlation matrix.

At this point, the only true pattern that I see is a correlation between the Volume and the Year. Therefore, a plot is done to further examine this relationship. The plot shows us that the Volume generally is increasing from year to year. An analogous base R plot is included.

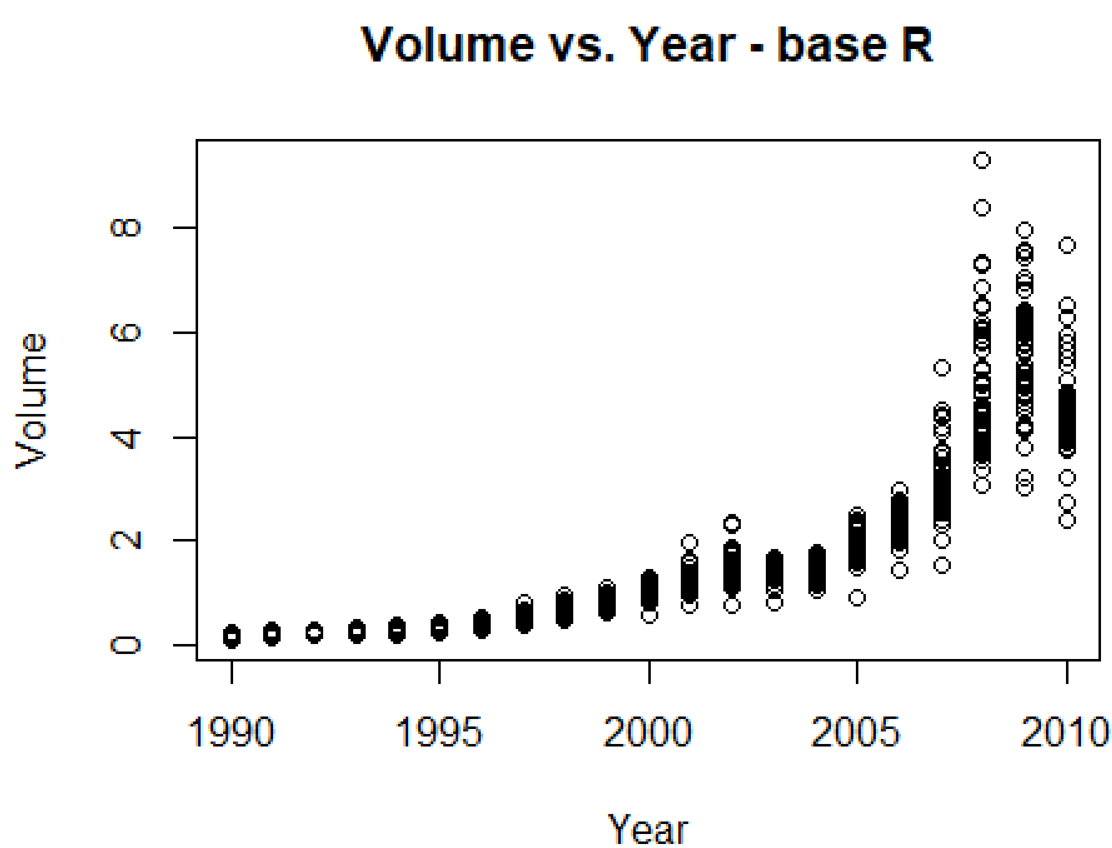
```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  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	Lag5	Volume	Today	NumDirection
Year	1.0000	-0.0323	-0.0334	-0.0300	-0.0311	-0.0305	0.8419	-0.0325	-0.0222
Lag1	-0.0323	1.0000	-0.0749	0.0586	-0.0713	-0.0082	-0.0650	-0.0750	-0.0500
Lag2	-0.0334	-0.0749	1.0000	-0.0757	0.0584	-0.0725	-0.0855	0.0592	0.0727
Lag3	-0.0300	0.0586	-0.0757	1.0000	-0.0754	0.0607	-0.0693	-0.0712	-0.0229
Lag4	-0.0311	-0.0713	0.0584	-0.0754	1.0000	-0.0757	-0.0611	-0.0078	-0.0205
Lag5	-0.0305	-0.0082	-0.0725	0.0607	-0.0757	1.0000	-0.0585	0.0110	-0.0182
Volume	0.8419	-0.0650	-0.0855	-0.0693	-0.0611	-0.0585	1.0000	-0.0331	-0.0180
Today	-0.0325	-0.0750	0.0592	-0.0712	-0.0078	0.0110	-0.0331	1.0000	0.7200
NumDirection	-0.0222	-0.0500	0.0727	-0.0229	-0.0205	-0.0182	-0.0180	0.7200	1.0000





Volume vs. Year - ggplot
Volume
0.0
2.5
5.0
7.5
1990
1995
2000
2005
2010
Year



Volume vs. Year - base R
Volume
0
2
4
6
8
1990
1995
2000
2005
2010
Year

Part B: Use the fully 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?

Results A model was fit using the instructions from the text, and a summary was printed per the same instructions. Additionally, for readability, I included a table of the p-values extracted from the summary.

Based on these outputs, aside from the Intercept, it appears that *"Lag2"* is the only predictor that is statistically significant at an alpha = 0.05. *Lag2*, per the ISLR documentation, represents the Percentage return for 2 weeks previous to the week being measured.

```
##
## 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
```

P-Values of Predictors for Direction

	P-Value
(Intercept)	0.0018988
Lag1	0.1181444
Lag2	0.0296014
Lag3	0.5469239
Lag4	0.2936533
Lag5	0.5833482
Volume	0.5376748

Part 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.

Results: First, I used the predict() function to predict Direction using the Weekly.glm model provided by the text book. Next, I added the predictions to the Weekly data set and printed a confusion matrix showing the breakdown of *Direction* predictions versus the observed *Direction*.

Next, per the instructions, I printed the overall fraction of correct predictions. I also included the overall accuracy as a percentage as I think that reads better. I rounded both outputs to 3 decimal places.

Lastly, per the instructions, I analyzed the types of the mistakes made by the model. The last table shows the percentage of correct predictions, by the model, based on whether the *Direction* was Up or Down in the given week. As we can see, in weeks whether the market was up, the model is ~92% accurate. In weeks when the market was down, the model is only ~11% accurate.

```
##         Predicted
## Observed Down  Up
##     Down   54 430
##     Up     48 557

## [1] "The percentage of accurate predictions is: 56.107 % (rounded to 3 decimals)"

## [1] "The overall fraction of correct predictions is: 611 / 1089"
```

Percentage Accuracy by Market Movement

Accuracy when Market is Up	Accuracy when Market is Down
92.066	11.157

Part 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).

Results: To complete this exercise, first I removed the prediction results from the **Weekly** data set from early exercises. Then I created subsets of training and test data sets using the years 2009 and 2010 as the test data and the prior years as the training set.

Then, I fit a model using only the **Lag2** variable as a predictor of **Direction**, using the training set to build the model.

I then used the model to predict the direction in the test data set and printed the confusion matrix and overall fraction of correct predictions, as instructed. I also included the percentage of accuracy for easier analysis.

Lastly, I included a comparison of *Model 1* (the model that uses all predictors of Direction from Parts B and C) and *Model 2* (the model that only uses "Lag2") as a predictor.

As we can see, despite using training/test data sets which often give less accurate predictions, the model that only uses "Lag2" to predict "Direction" is more accurate.

Per the homework instructions, I've skipped Parts E-I.

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly.training)
##
## 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
```

```
##         Predicted
## Observed Down Up
##     Down    9 34
##     Up      5 56
```

```
## [1] "The percentage of accurate predictions in test set is: 62.5 % (rounded to 3 decim
als)"
```

```
## [1] "The overall fraction of correct predictions in the test set is: 65 / 104"
```

Percentage Accuracy by Model

Accuracy of Model 1	Accuracy of Model 2
56.10652	62.5

Question 4.7.11, 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.

Part 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 the median.

Results: I created the variable "mpg01" per the instructions using the condition of whether or not the "mpg" value for each observation is above or below the median value of "mpg". I printed the header to confirm the creation of the variable.

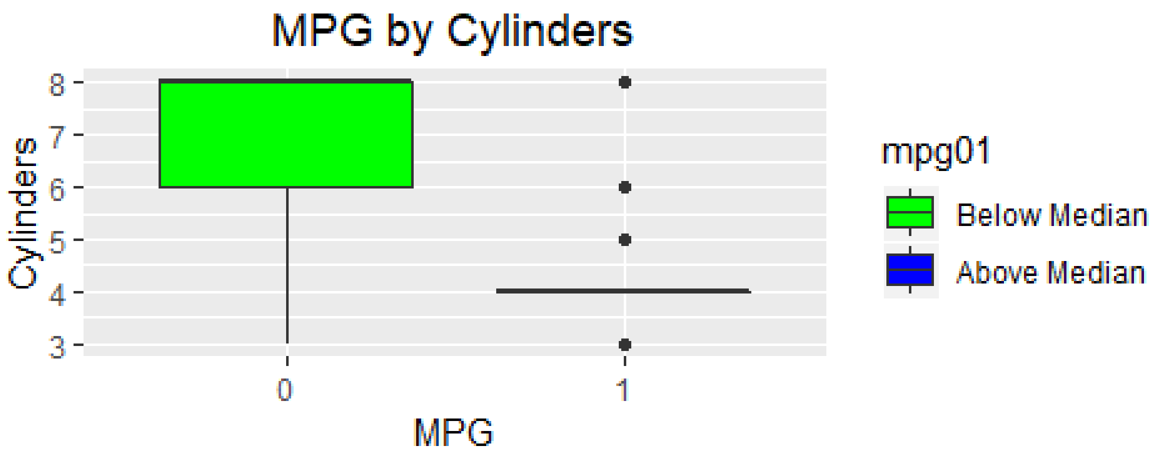
```
##   mpg01 mpg cylinders displacement horsepower weight acceleration year
## 1     0  18         8          307        130   3504         12.0   70
## 2     0  15         8          350        165   3693         11.5   70
## 3     0  18         8          318        150   3436         11.0   70
##   origin                      name
## 1      1 chevrolet chevelle malibu
## 2      1         buick skylark 320
## 3      1        plymouth satellite
```

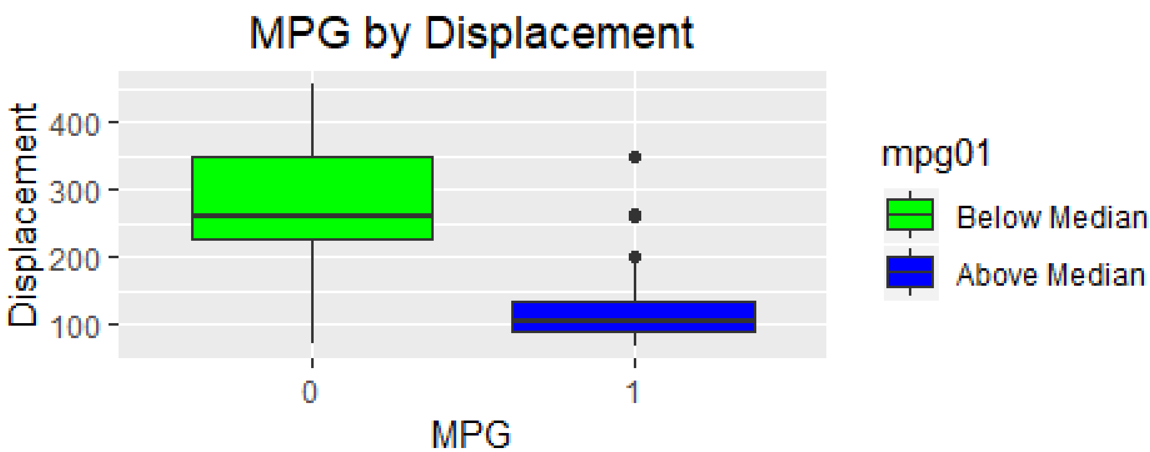
Part B: Explore the data set graphically in order to investigate the association between **mpg01** and the other features. Which of the other features seem most likely to be useful in predicting **mgp01**? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

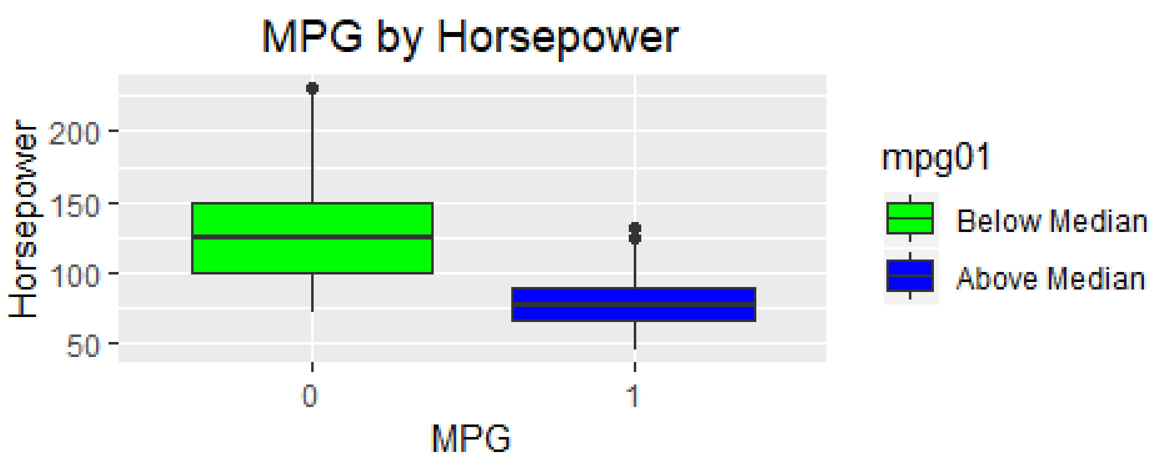
Results: First, I printed boxplots showing the relationship between the binary variable **mpg01** and the other predictors. From these plots, it appears to me that the most useful features are *Cylinders, Displacement, Horsepower and Weight.* There appears to be a possibly useful correlation between the dependent variable and *Year*, but I will examine that later in this exercise. Base R plots are included for reference.

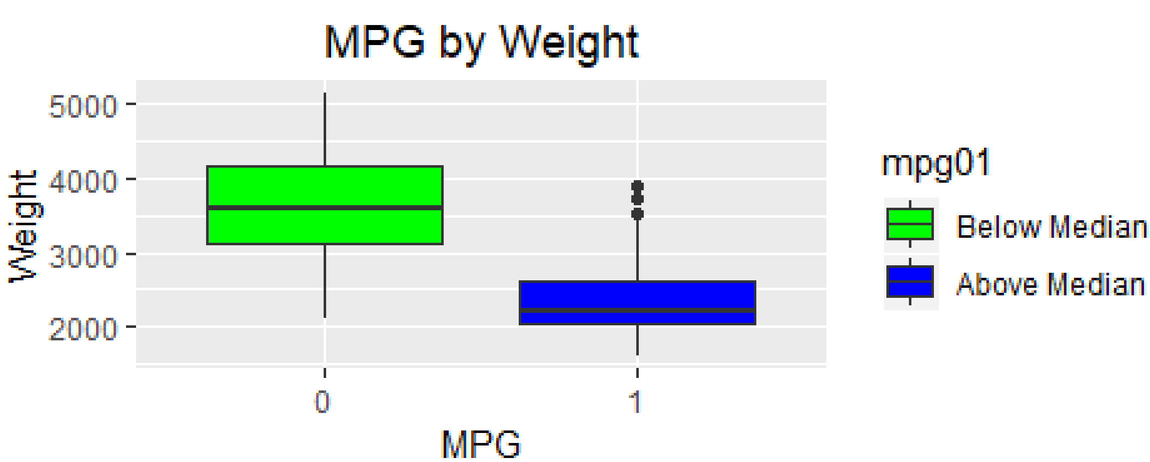
Next, I printed a table showing the correlation values of each variable as well as a corresponding correlation plot to visually depict the table. These two visuals confirm that *Cylinders, Displacement, Horsepower and Weight* will be useful predictors. The correlation between the response variable and *Year* still appears to be somewhat relevant but I am not sure if it will help the model. I will create two models, one with *Year* and one without to compare.

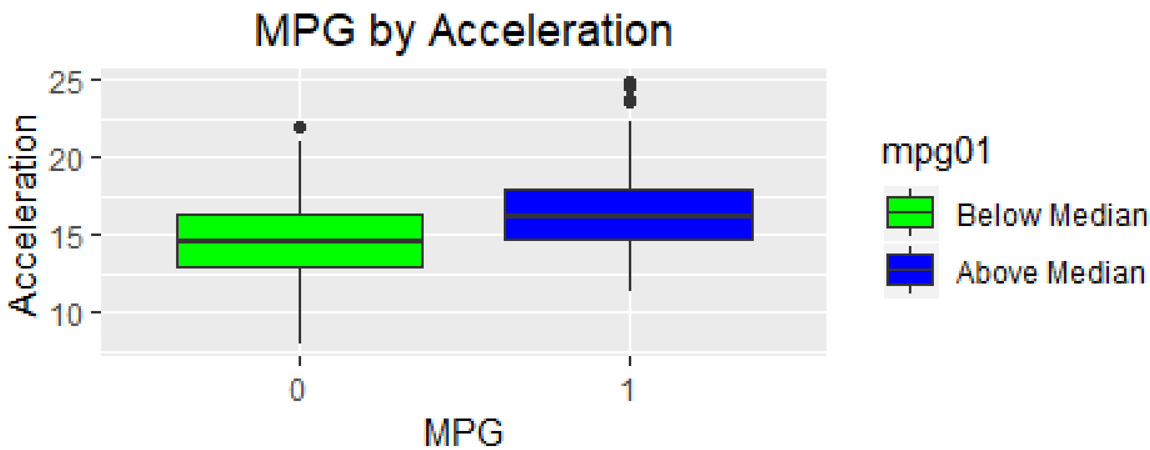
Lastly, to confirm my selections, I included scatterplot matrices and looked at the relationships again. These matrices confirm my decision in the prior paragraph. Analagous base R plots have been included per homework guidelines.

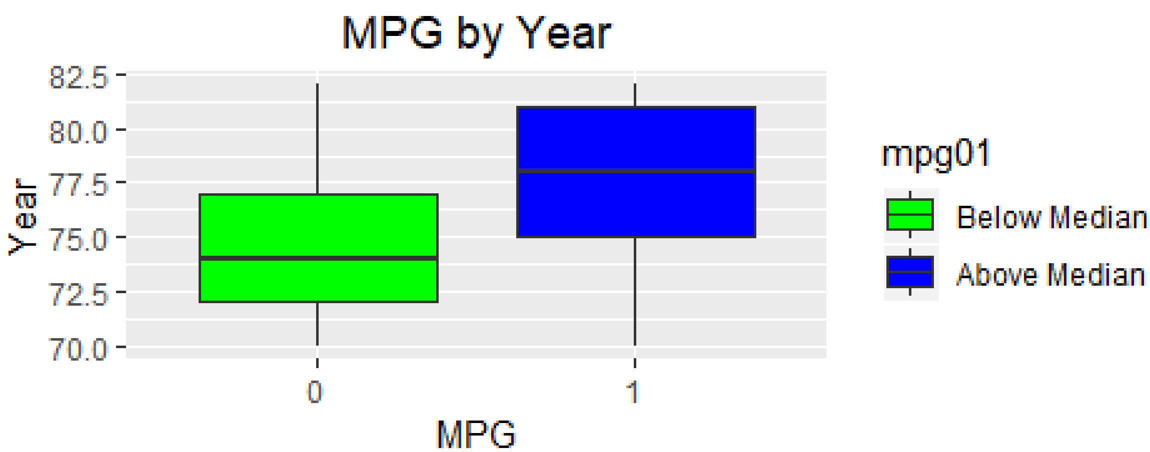

MPG by Cylinders
Cylinders
8
7
6
5
4
3
0
1
MPG
mpg01
Below Median
Above Median

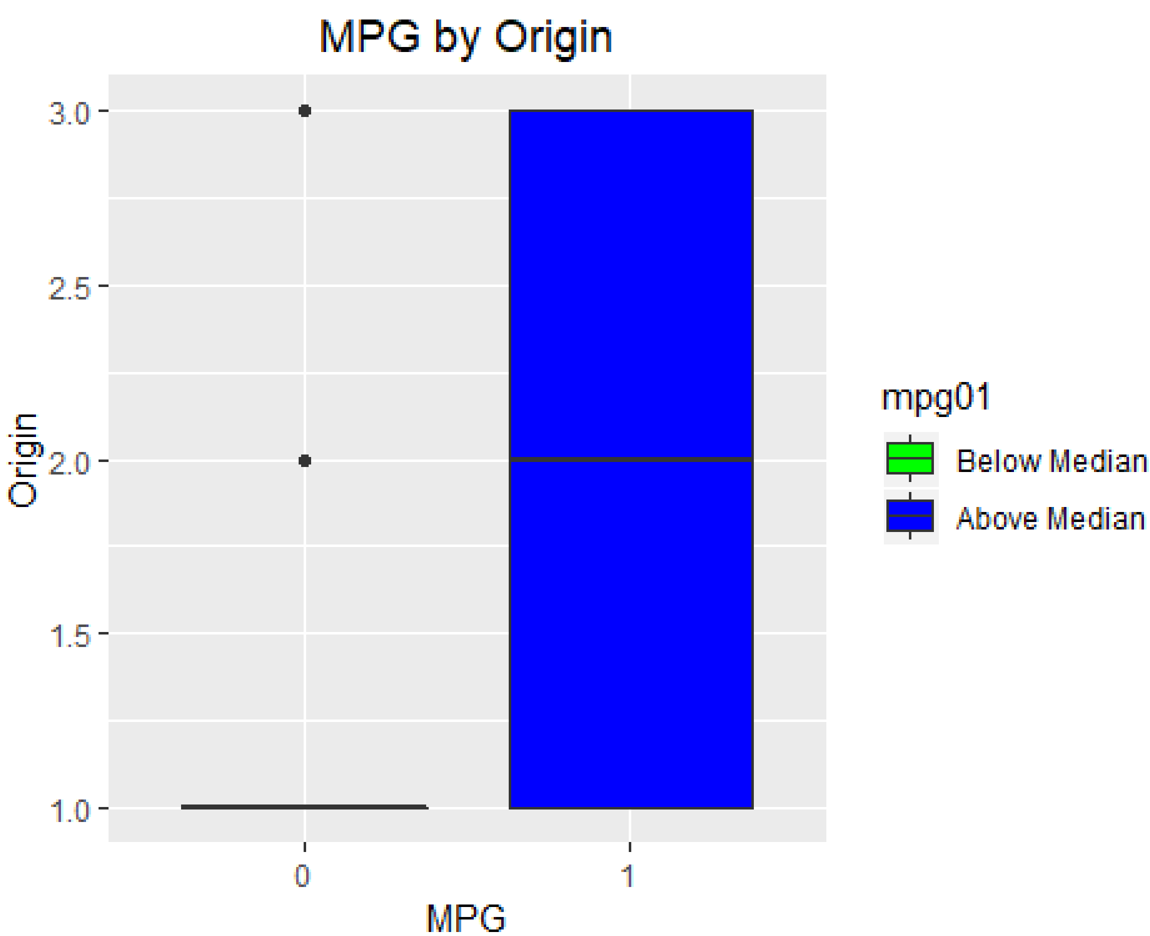

MPG by Displacement
Displacement
400
300
200
100
0
1
MPG
mpg01
Below Median
Above Median

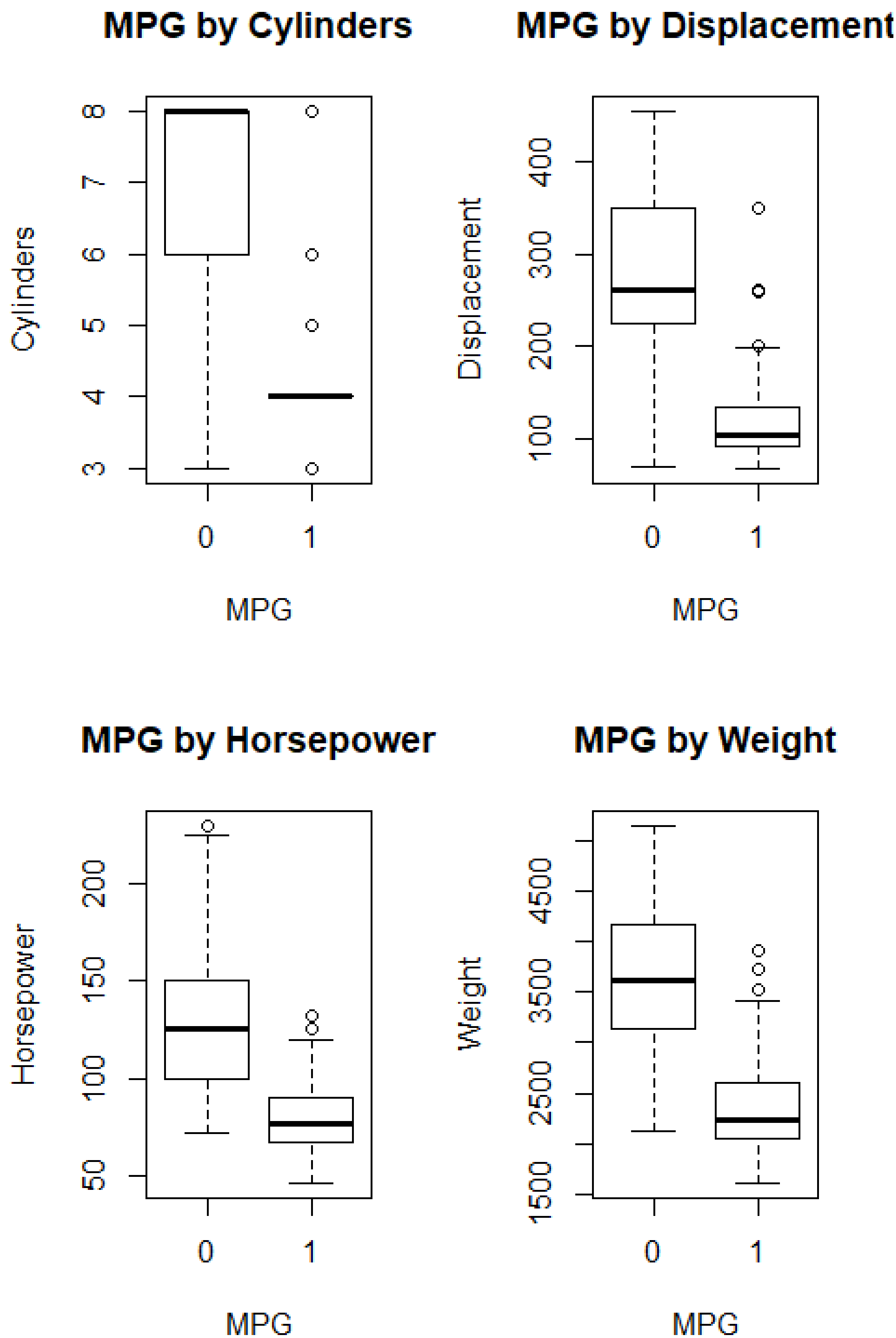

MPG by Horsepower
Horsepower
200
150
100
50
0
1
MPG
mpg01
Below Median
Above Median

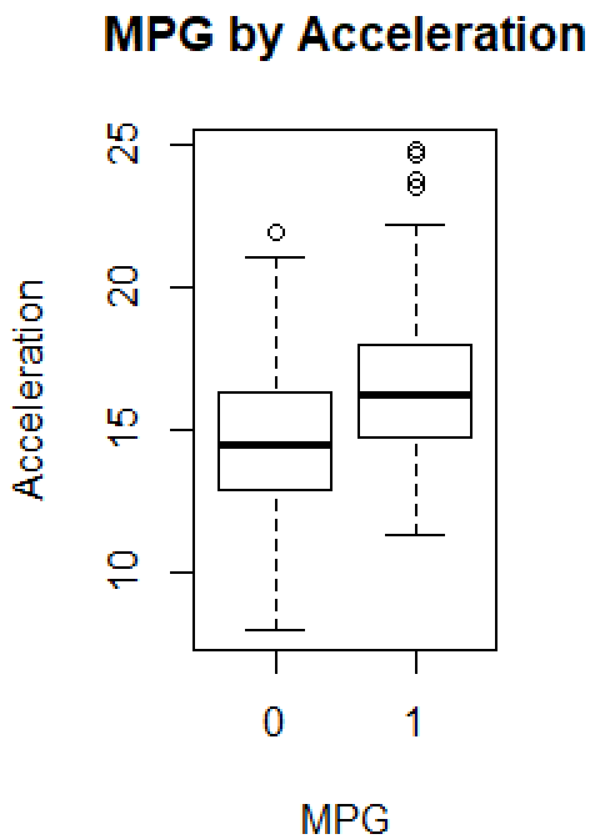

MPG by Weight
Weight
5000
4000
3000
2000
0
1
MPG
mpg01
Below Median
Above Median

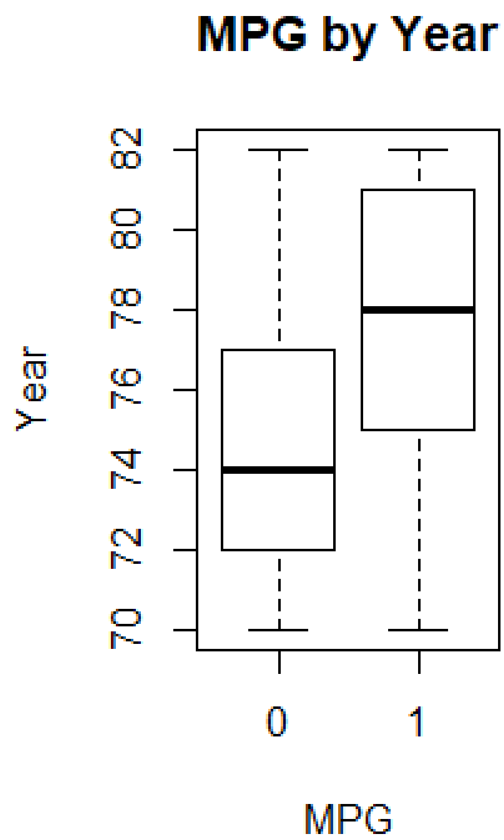

MPG by Acceleration
Acceleration
25
20
15
10
0
1
MPG
mpg01
Below Median
Above Median

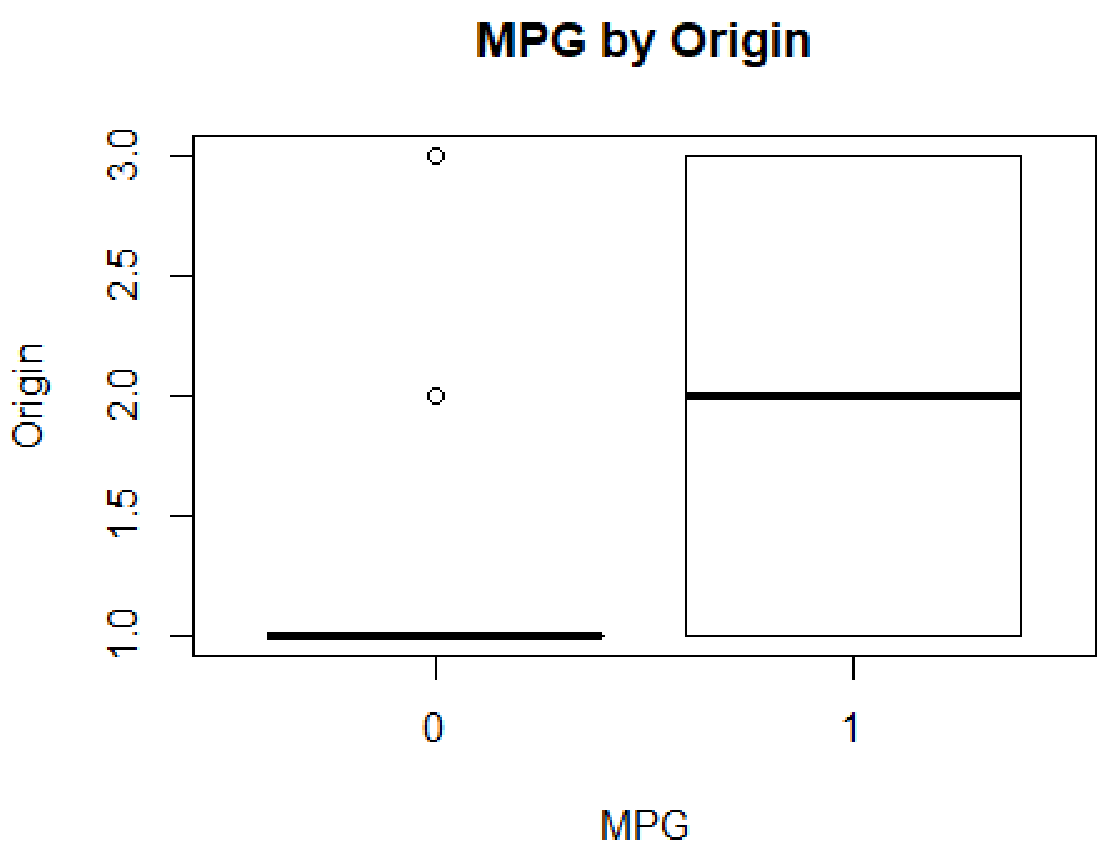

MPG by Year
Year
82.5
80.0
77.5
75.0
72.5
70.0
0
1
MPG
mpg01
Below Median
Above Median


MPG by Origin
Origin
3.0
2.5
2.0
1.5
1.0
0
1
MPG
mpg01
Below Median
Above Median

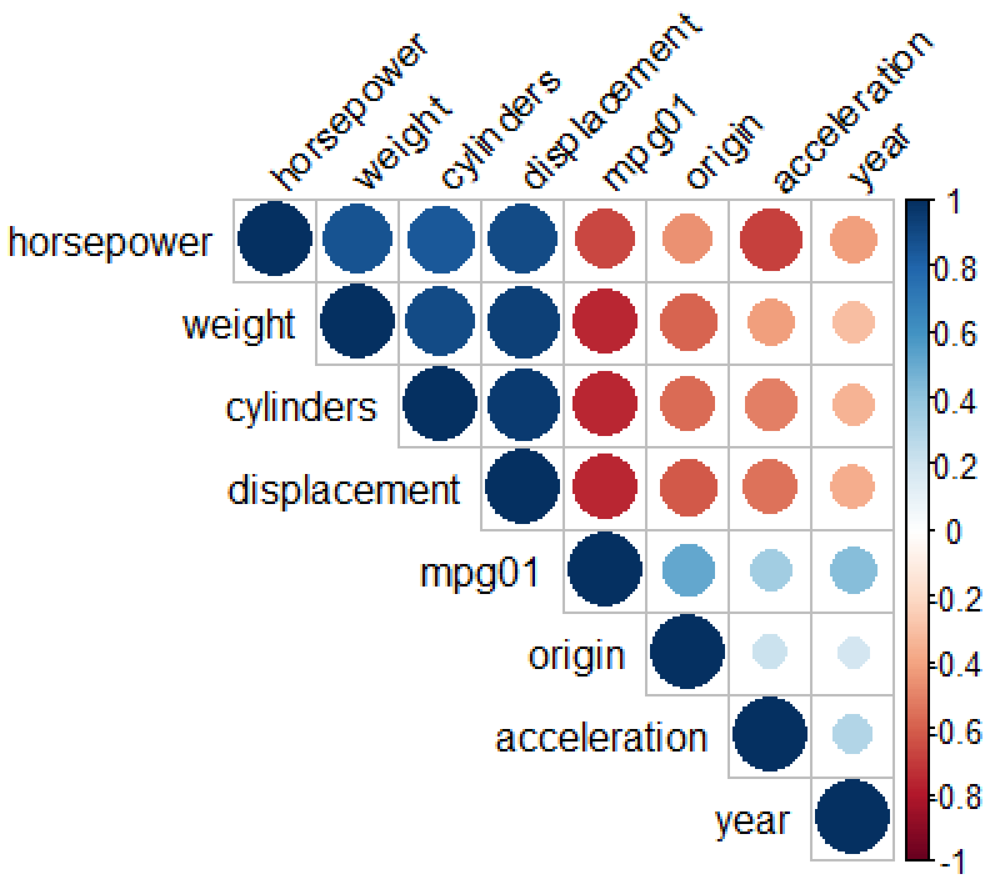

MPG by Cylinders
Cylinders
3
4
5
6
7
8
0
1
MPG
MPG by Displacement
Displacement
100
200
300
400
0
1
MPG
MPG by Horsepower
Horsepower
50
100
150
200
0
1
MPG
MPG by Weight
Weight
1500
2500
3500
4500
0
1
MPG

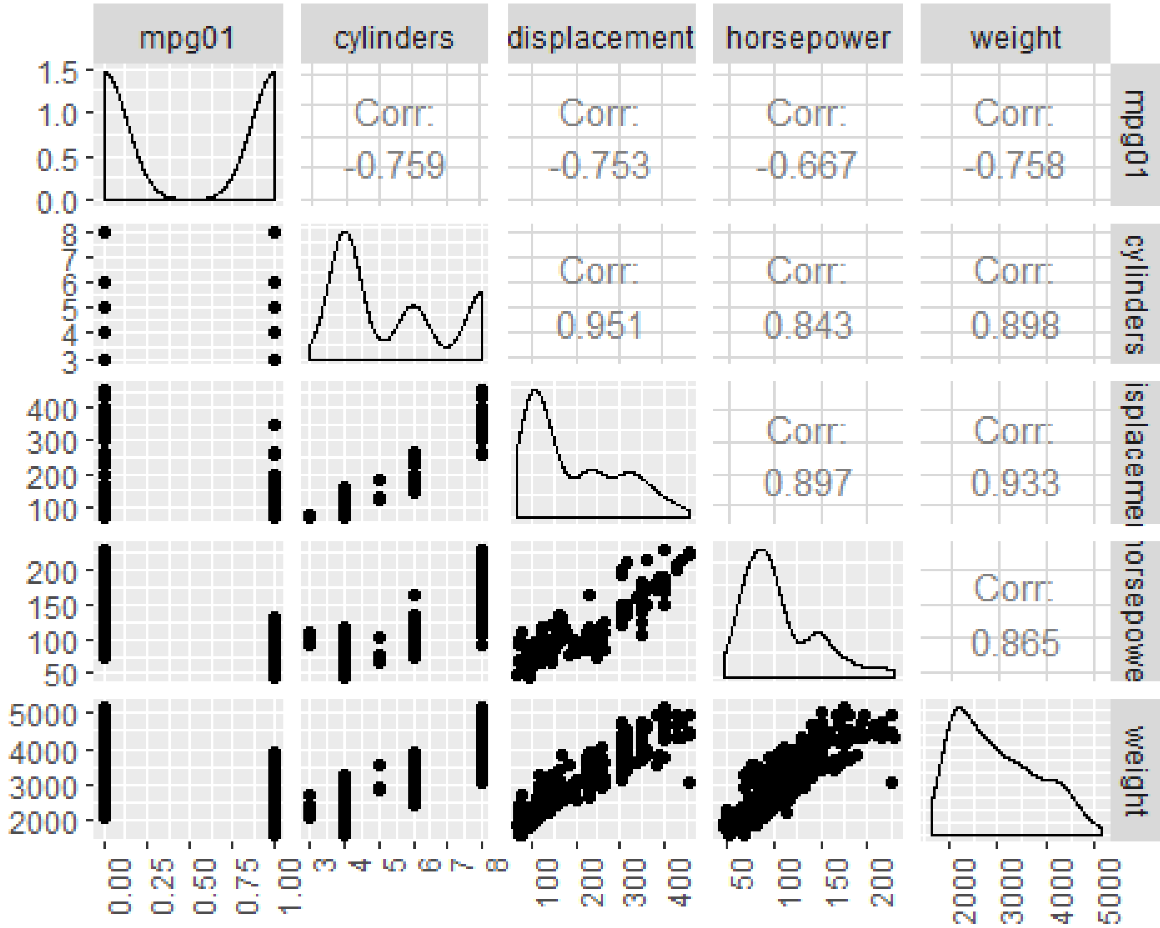

MPG by Acceleration
Acceleration
25
20
15
10
0
1
MPG

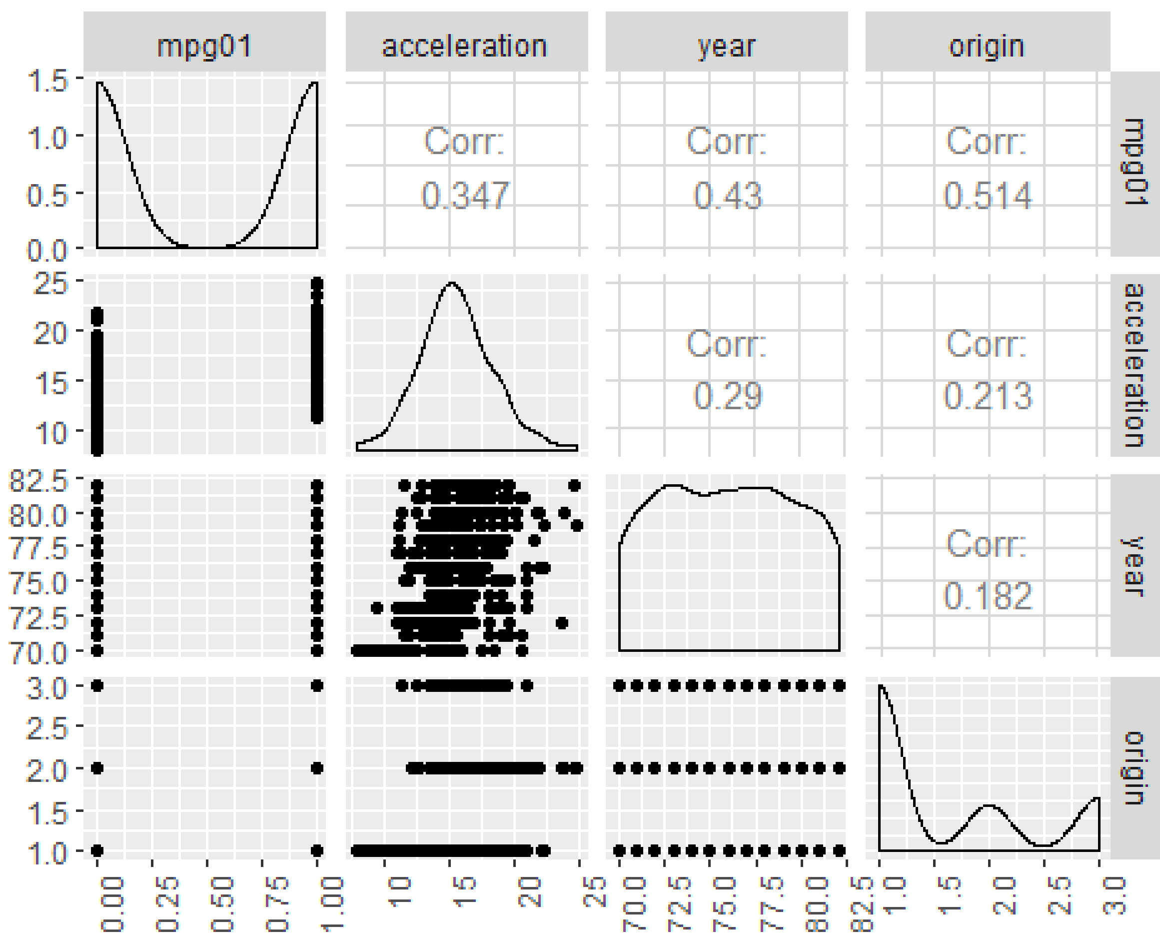

MPG by Year
Year
82
80
78
76
74
72
70
0
1
MPG

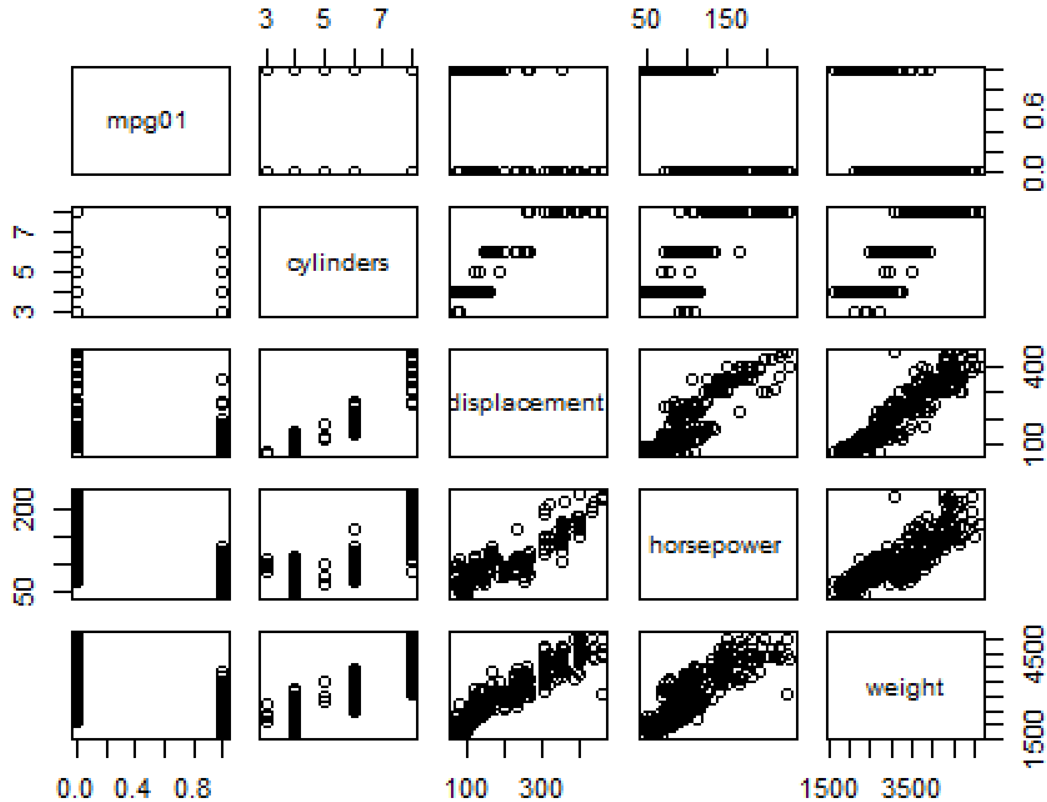

MPG by Origin
Origin
3.0
2.5
2.0
1.5
1.0
0
1
MPG

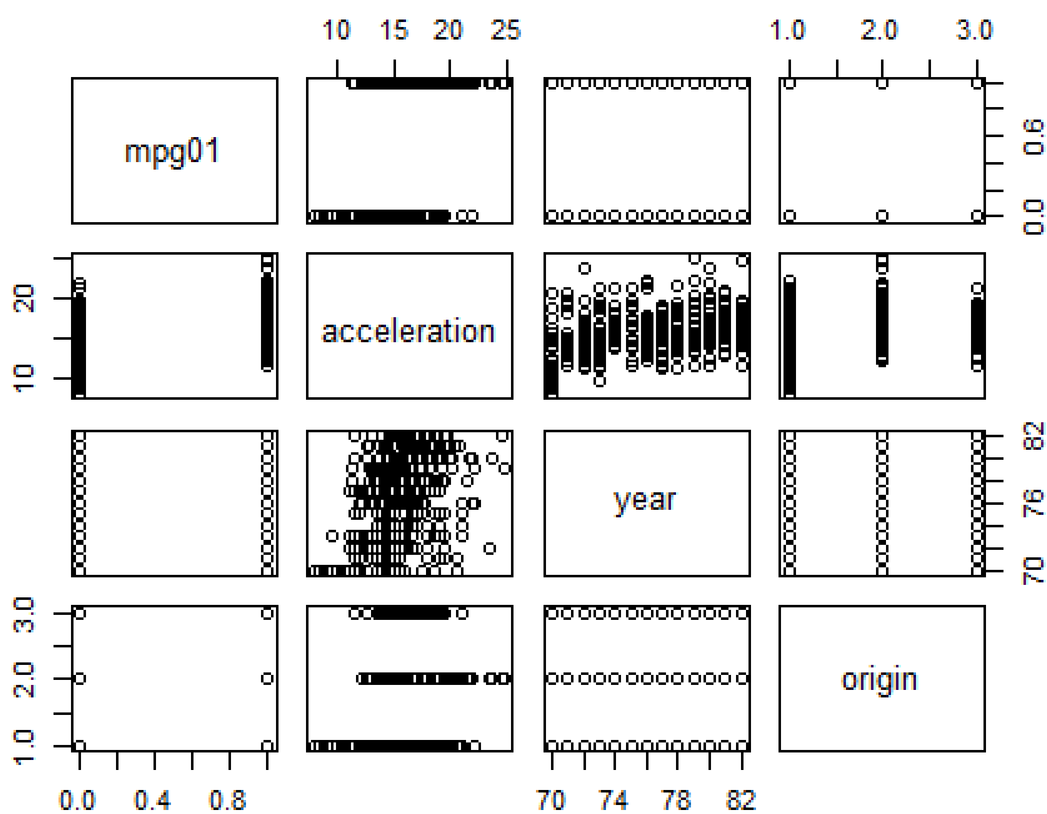
	mpg01	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg01	1.0000	-0.7592	-0.7535	-0.6671	-0.7578	0.3468	0.4299	0.5137
cylinders	-0.7592	1.0000	0.9508	0.8430	0.8975	-0.5047	-0.3456	-0.5689
displacement	-0.7535	0.9508	1.0000	0.8973	0.9330	-0.5438	-0.3699	-0.6145
horsepower	-0.6671	0.8430	0.8973	1.0000	0.8645	-0.6892	-0.4164	-0.4552
weight	-0.7578	0.8975	0.9330	0.8645	1.0000	-0.4168	-0.3091	-0.5850
acceleration	0.3468	-0.5047	-0.5438	-0.6892	-0.4168	1.0000	0.2903	0.2127
year	0.4299	-0.3456	-0.3699	-0.4164	-0.3091	0.2903	1.0000	0.1815
origin	0.5137	-0.5689	-0.6145	-0.4552	-0.5850	0.2127	0.1815	1.0000




mpg01
cylinders
displacement
horsepower
weight
Corr: -0.759
Corr: -0.753
Corr: -0.667
Corr: -0.758
Corr: 0.951
Corr: 0.843
Corr: 0.898
Corr: 0.897
Corr: 0.933
Corr: 0.865


mpg01
acceleration
year
origin
Corr: 0.347
Corr: 0.43
Corr: 0.514
Corr: 0.29
Corr: 0.213
Corr: 0.182


mpg01
cylinders
displacement
horsepower
weight
3
5
7
50
150
0.0
0.6
100
400
200
4500
1500
0.4
0.8
300
3500


mpg01
acceleration
year
origin
10
15
20
25
1.0
2.0
3.0
0.0
0.6
82
76
70
0.4
0.8
74
78

Part C: Split the data into training and test sets.

Results: Here I set a sample size to extract 75% of the data set as training data and 25% as testing data. I set a seed for reproducibility and split the sets.

To confirm, I printed the number of rows in each of the 3 data sets.

# of Rows in Each Data Set

# Rows Auto	# Rows Train	# Rows Test
392	294	98

Per the Homework PDF, I've skipped Parts D and E

Part 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?

Results: First, I fit both two models with the predictors discussed in (b). The first model included *Year* as a predictor, while the second model did not. I first compared the two models by comparing the p-values of the predictors. In the first model, **Weight, Horsepower and Year** are significant predictors. In the second model, we see that **Horsepower and Displacement** are the only significant predictors with p-values below our alpha of 0.05. The *Weight* variable is no longer significant when *Year* is removed from the modeling. This is interesting since *Year* was a predictor that did not appear to be as strong of a correlated variable as the others in part (b).

Next, I compared the AIC (Akaike Information Criterion) value is useful in comparing models to see which "fit" the data better. The lower AIC indicates a superior model. Here we see that the model with Year as a predictor is superior, according to AIC. Next we'll see if this superiority translates to better results when using them on our test data set.

Lastly, I used the two models to predict the test data set **mpg01** values. Then I printed the confusion matrices, accuracies, and fractions of accuracies for both models. Then, per assignment instructions, I also showed the error rate for each model.

As we can see, predictably (from the AIC discussion above), the model with *Year* included as a predictor is better at predicting the response variable in the test data.

P-Values of Predictors with Year Included

	P-Values
(Intercept)	0.0054592
cylinders	0.6565322
weight	0.0009514
displacement	0.1189136
horsepower	0.0339068
year	0.0000012

P-Values of Predictors without Year Included

	P-Values
(Intercept)	0.0000000
cylinders	0.7981999
weight	0.0561246
displacement	0.0494608
horsepower	0.0106244

Comparison of AIC Values

AIC of Model with Year	AIC of Model without Year
131.8796	163.036

```
##         Predicted #1
## Observed  0  1
##        0 45  5
##        1  2 46

##         Predicted #2
## Observed  0  1
##        0 42  8
##        1  3 45

## [1] "The percentage of accurate predictions in test set is: 92.857 % (rounded to 3 dec
imals)"

## [1] "The overall fraction of correct predictions in the test set is: 91 / 98"

## [1] "The percentage of accurate predictions in test set is: 88.776 % (rounded to 3 dec
imals)"

## [1] "The overall fraction of correct predictions in the test set is: 87 / 98"
```

Test Error by Model

Error Rate of Model 1 (with Year)	Error Rate of Model 2 (w/out Year)
7.142857	11.22449

Question 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. Test your function using the model from 4.7.10 b. (This needs to be an actual function using the function() command, not just a chunk of code). This will be something you will want to use throughout the semester, since we will be calculating these a lot! *Show the function code you wrote in your final write-up.*

```
class.function <-
  function(cutoff, probs, outcomes) {
    results <- list()
    predictions <- ifelse(probs > 0.5, 1, 0)
    confusion.matrix <- table(outcomes, predictions)
    names(dimnames(confusion.matrix)) <- c("Observed", "Predicted")
    results$misclassification.rate <- 1- ((confusion.matrix[1,1] +
                                               confusion.matrix[2,2])/(confusion.matrix[1,
1] +                                                                     confusion.matrix[
1,2]+confusion.matrix[2,1] +
                                                                         confusion.matrix[
2,2]))

     results$sensitivity <- confusion.matrix[2,2]/(confusion.matrix[2,2] + confusion.matr
ix[2,1])
    results$specificity <- confusion.matrix[1,1]/(confusion.matrix[1,1] + confusion.matri
x[1,2])
    return(as.data.frame(results))
  }

class.function(0.5, Weekly.probs, Weekly$Direction)

##   misclassification.rate sensitivity specificity
## 1              0.4389348   0.9206612   0.1115702
```