Assignment 3

2024-11-13

Overview of the Weekly Financial Data

The weekly_data CSV contains financial information spanning from 1990 to 2010, featuring various variables.

Year: This column indicates the year of each observation. Lag1 to Lag5: These columns represent the returns from the previous five weeks, with Lag1 corresponding to one week ago, Lag2 to two weeks ago, and so forth. Volume: This measures the number of shares traded in billions during that week, providing insight into market activity. Today: This column shows the percentage return for the current week. Direction: A categorical variable that indicates whether the market moved "Up" (positive return) or "Down" (negative return) for that week.

This dataset is useful for analyzing the relationship between past returns and trading volume in relation to the current week's market performance.

PROBLEM 1

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

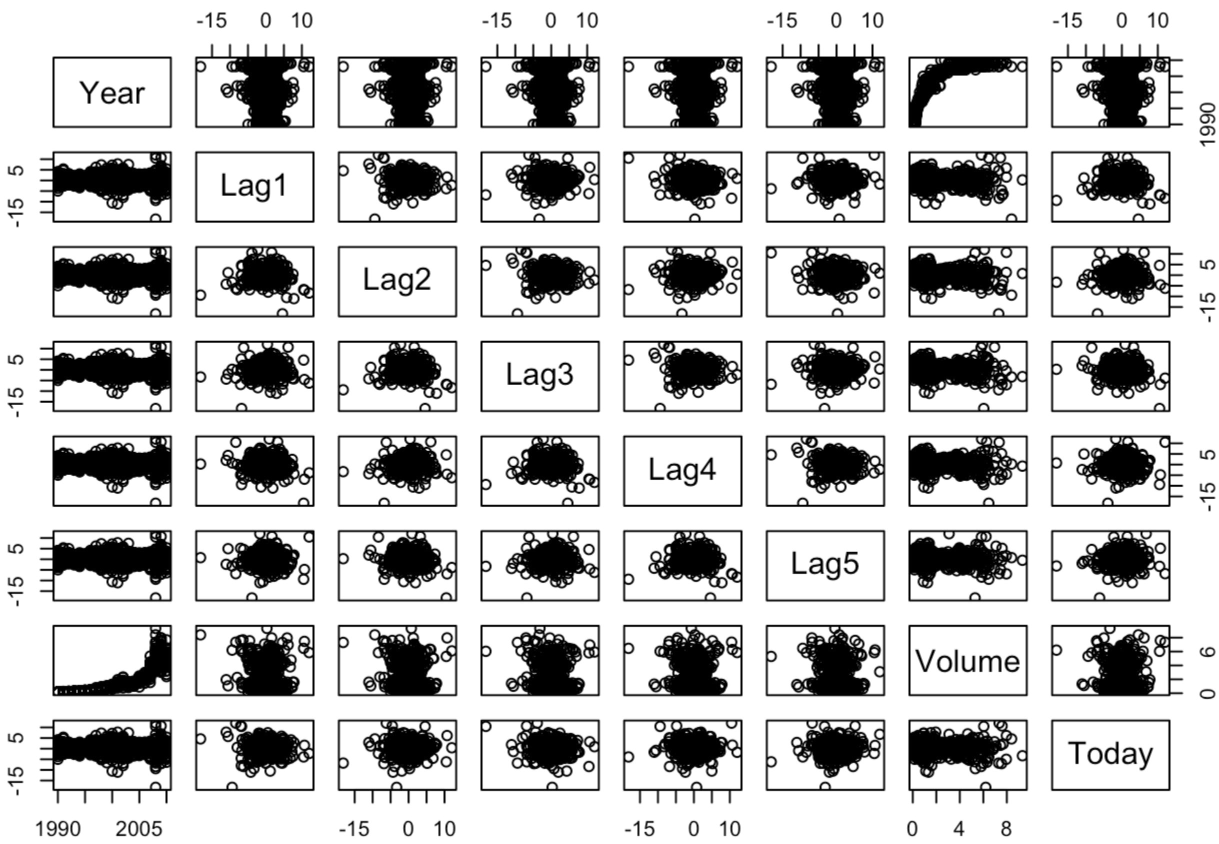
```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

```
##      Year   Lag1   Lag2   Lag3   Lag4   Lag5   Volume  Today Direction
## 1084 2010  0.043 -2.173  3.599  0.015  0.586 4.177436 -0.861      Down
## 1085 2010 -0.861  0.043 -2.173  3.599  0.015 3.205160  2.969        Up
## 1086 2010  2.969 -0.861  0.043 -2.173  3.599 4.242568  1.281        Up
## 1087 2010  1.281  2.969 -0.861  0.043 -2.173 4.835082  0.283        Up
## 1088 2010  0.283  1.281  2.969 -0.861  0.043 4.454044  1.034        Up
## 1089 2010  1.034  0.283  1.281  2.969 -0.861 2.707105  0.069        Up
```

```
##       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        
##  Length:1089       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

```
##               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
```

```
##               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
```



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?

```
##
## Call:
## glm(formula = DirectionBinary ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = weekly_data)
##
## 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
```

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.

```
##          actual
## predicted Down  Up
##      Down   54  48
##      Up    430 557
```

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).

```
## [1] 0.625
```

e. Repeat d) using LDA.

```
##          actual
## predicted Down Up
##      Down    9  5
##      Up     34 56
```

```
## [1] 0.625
```

f. Repeat d) using QDA

```
##          actual
## predicted Down Up
##        Up   43 61
```

```
## [1] 0.5865385
```

g. KNN for N = 1

```
##          actual
## predicted Down Up
##      Down   21 29
##      Up     22 32
```

```
## [1] 0.5096154
```

i. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

```
##          actual
## predicted Down Up
##      Down   27 32
##      Up     16 29
```

```
## [1] 0.5384615
```

```
##          actual
## predicted Down Up
##      Down   33 42
##      Up     10 19
```

```
## [1] 0.5
```

```
##          actual
## predicted Down Up
##      Down   33 42
##      Up     10 19
```

```
## [1] 0.5
```

```
##          actual
## predicted Down Up
##      Down   27 32
##      Up     16 29
```

```
## [1] 0.5384615
```

```
##          actual
## predicted Down Up
##      Down   32 46
##      Up     11 15
```

```
## [1] 0.4519231
```

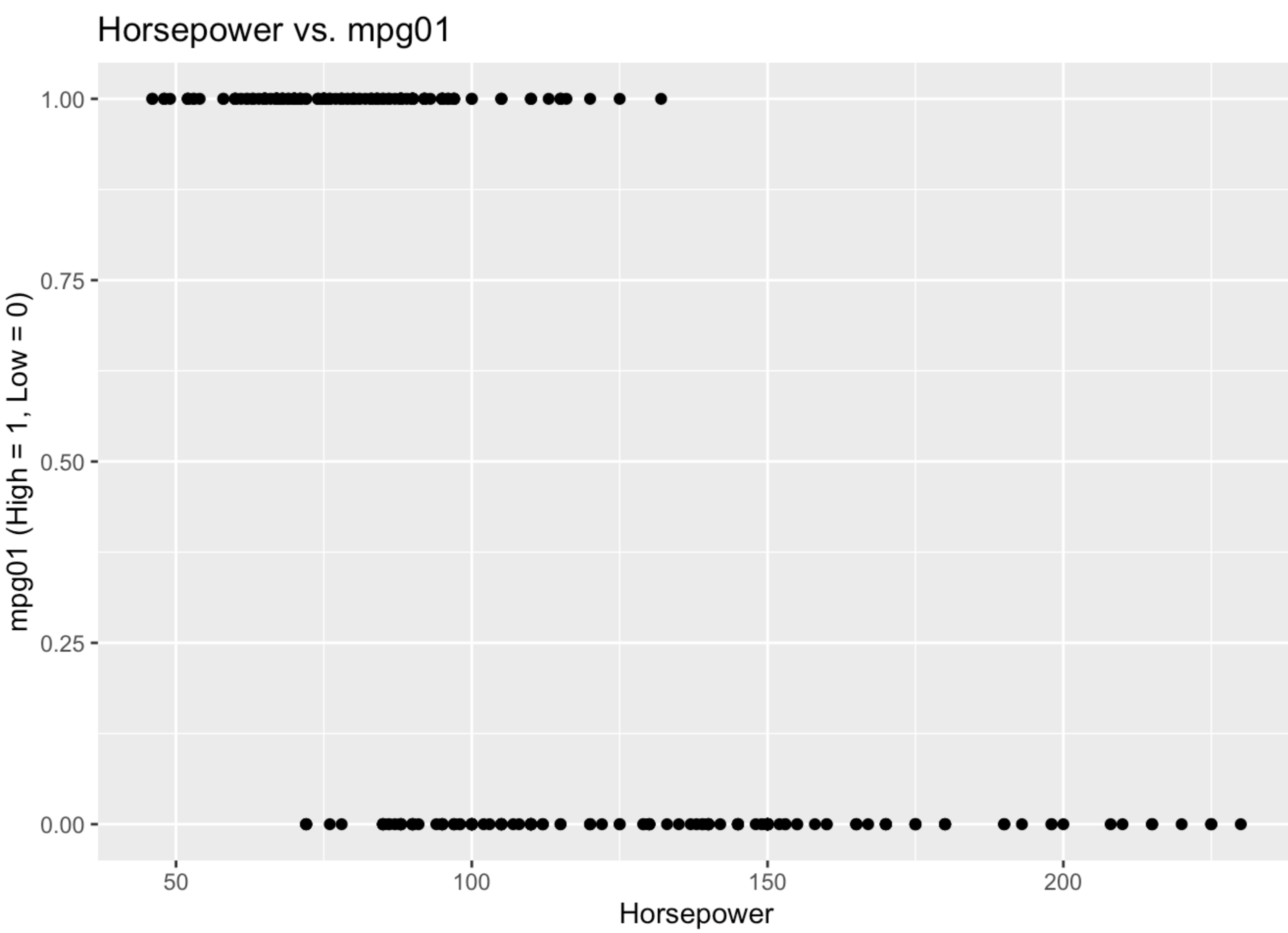
```
## [1] "Confusion Matrix for K = 3"
##          actual
## predicted Down Up
##      Down   27 38
##      Up     16 23
## [1] "Accuracy for K = 3 :"
## [1] 0.4807692
## [1] "Confusion Matrix for K = 5"
##          actual
## predicted Down Up
##      Down   33 36
##      Up     10 25
## [1] "Accuracy for K = 5 :"
## [1] 0.5576923
## [1] "Confusion Matrix for K = 7"
##          actual
## predicted Down Up
##      Down   31 35
##      Up     12 26
## [1] "Accuracy for K = 7 :"
## [1] 0.5480769
```

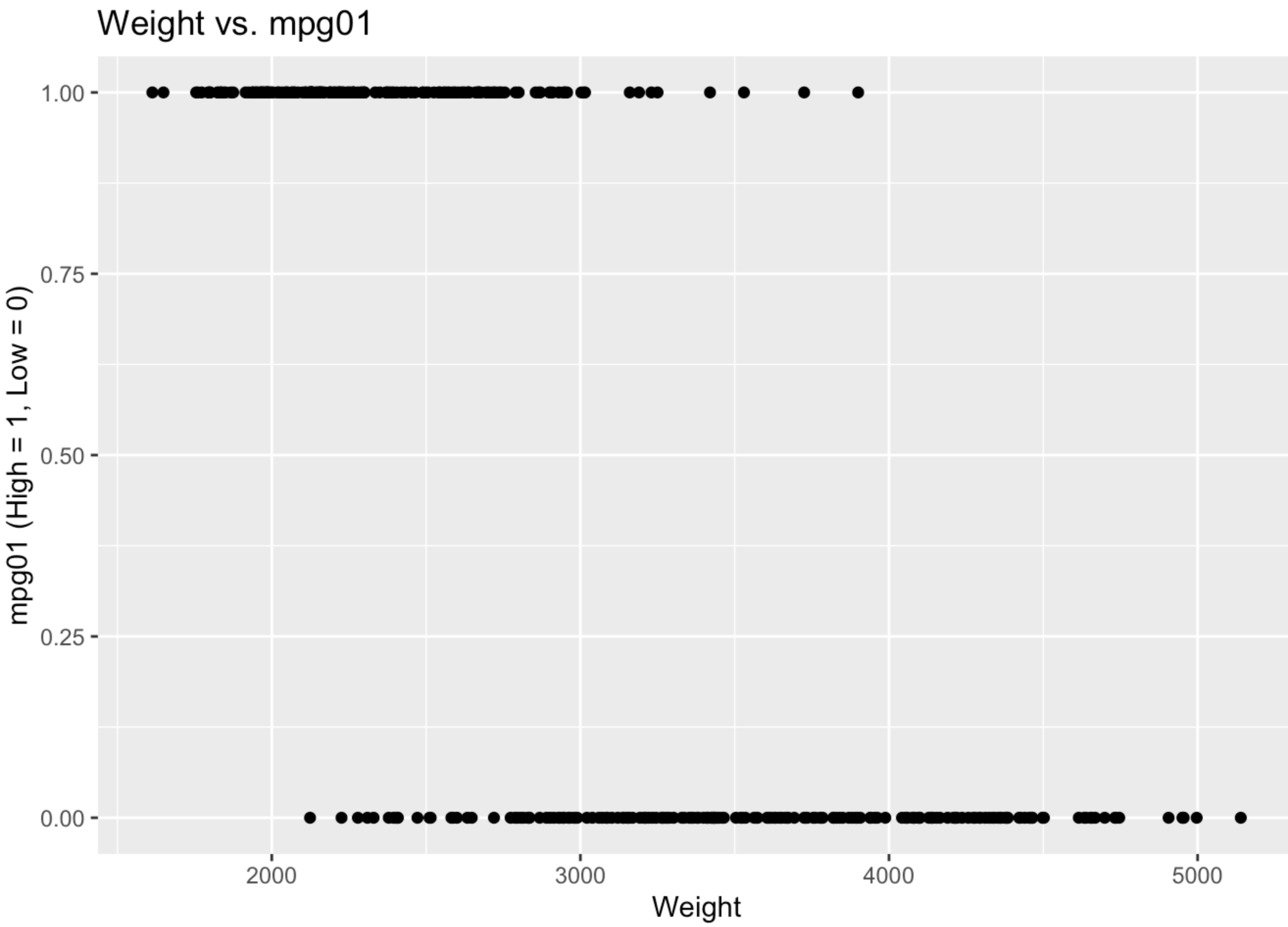
PROBLEM 2

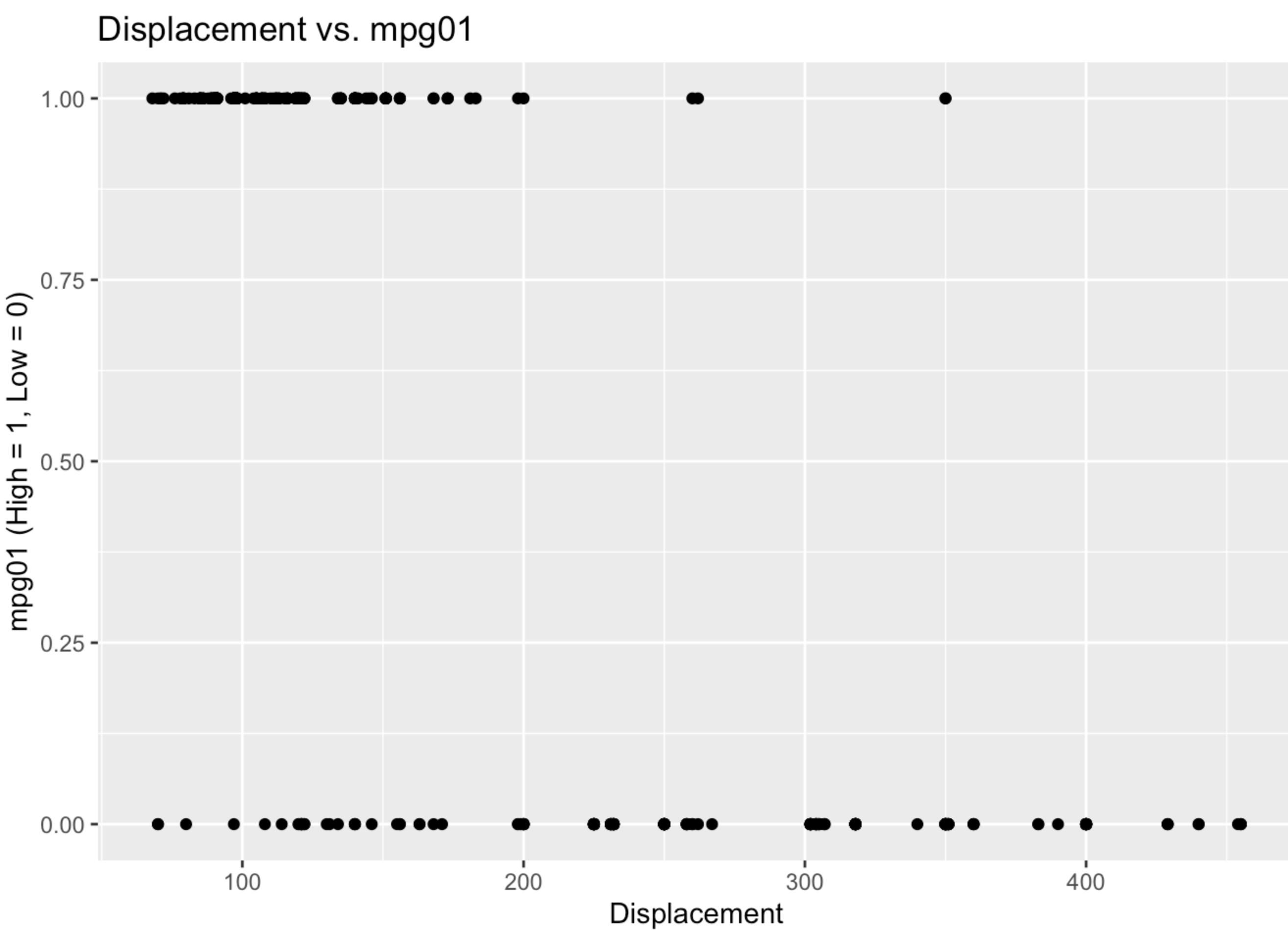
a. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing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

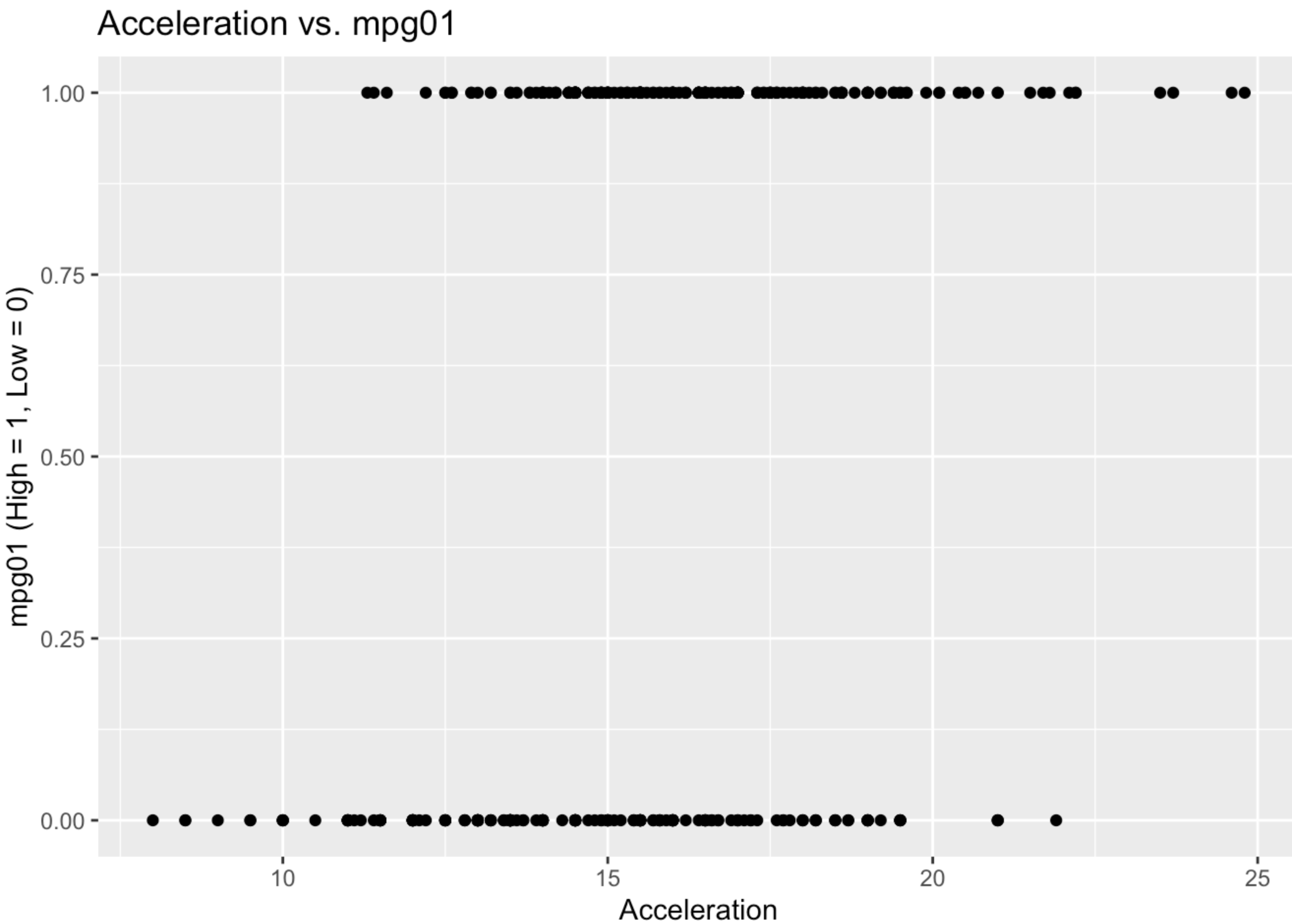
```
##   mpg01 mpg cylinders displacement horsepower weight acceleration year origin
## 1     0  18         8          307        130   3504         12.0   70      1
## 2     0  15         8          350        165   3693         11.5   70      1
## 3     0  18         8          318        150   3436         11.0   70      1
## 4     0  16         8          304        150   3433         12.0   70      1
## 5     0  17         8          302        140   3449         10.5   70      1
## 6     0  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
```

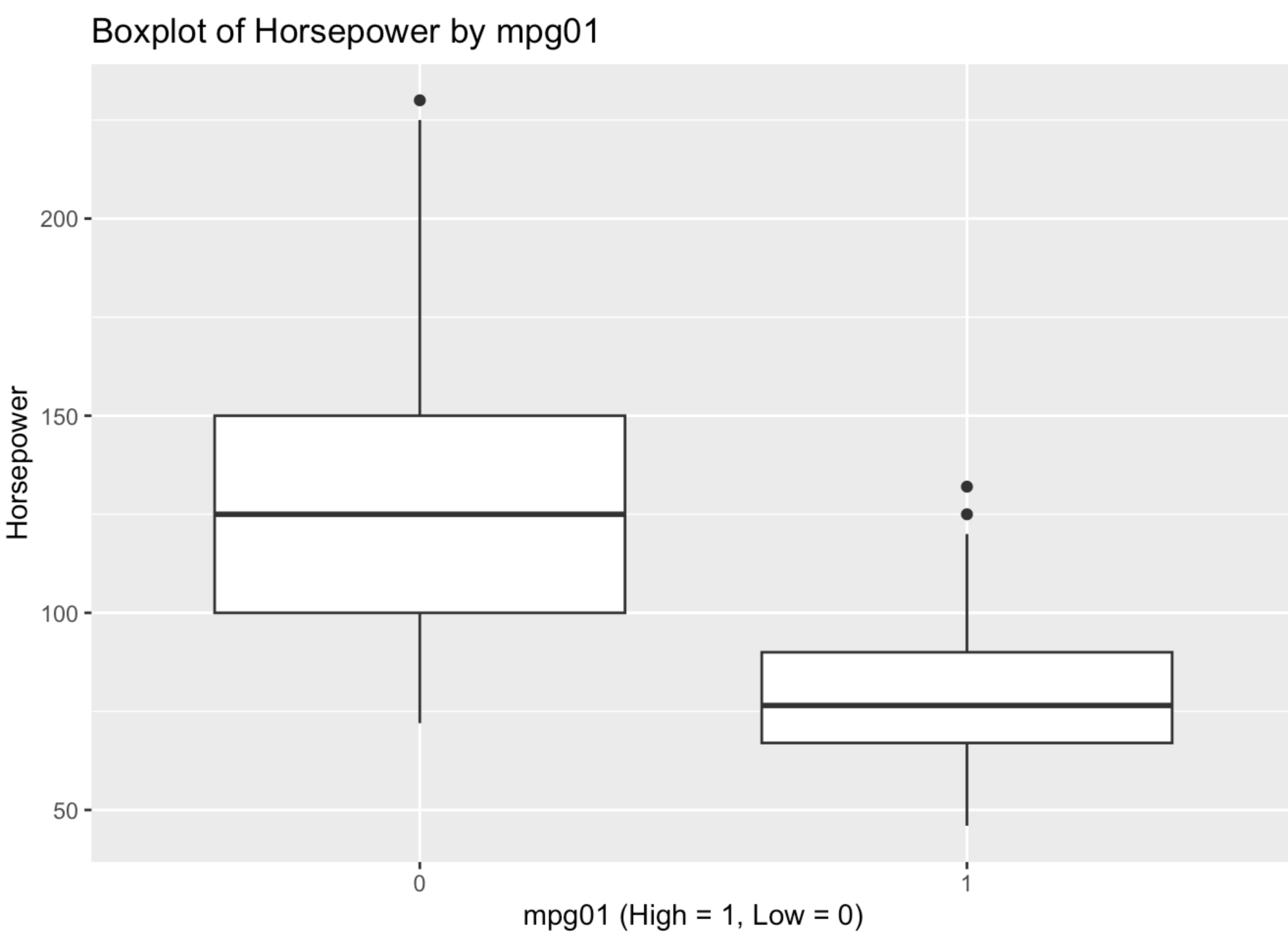
b. Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

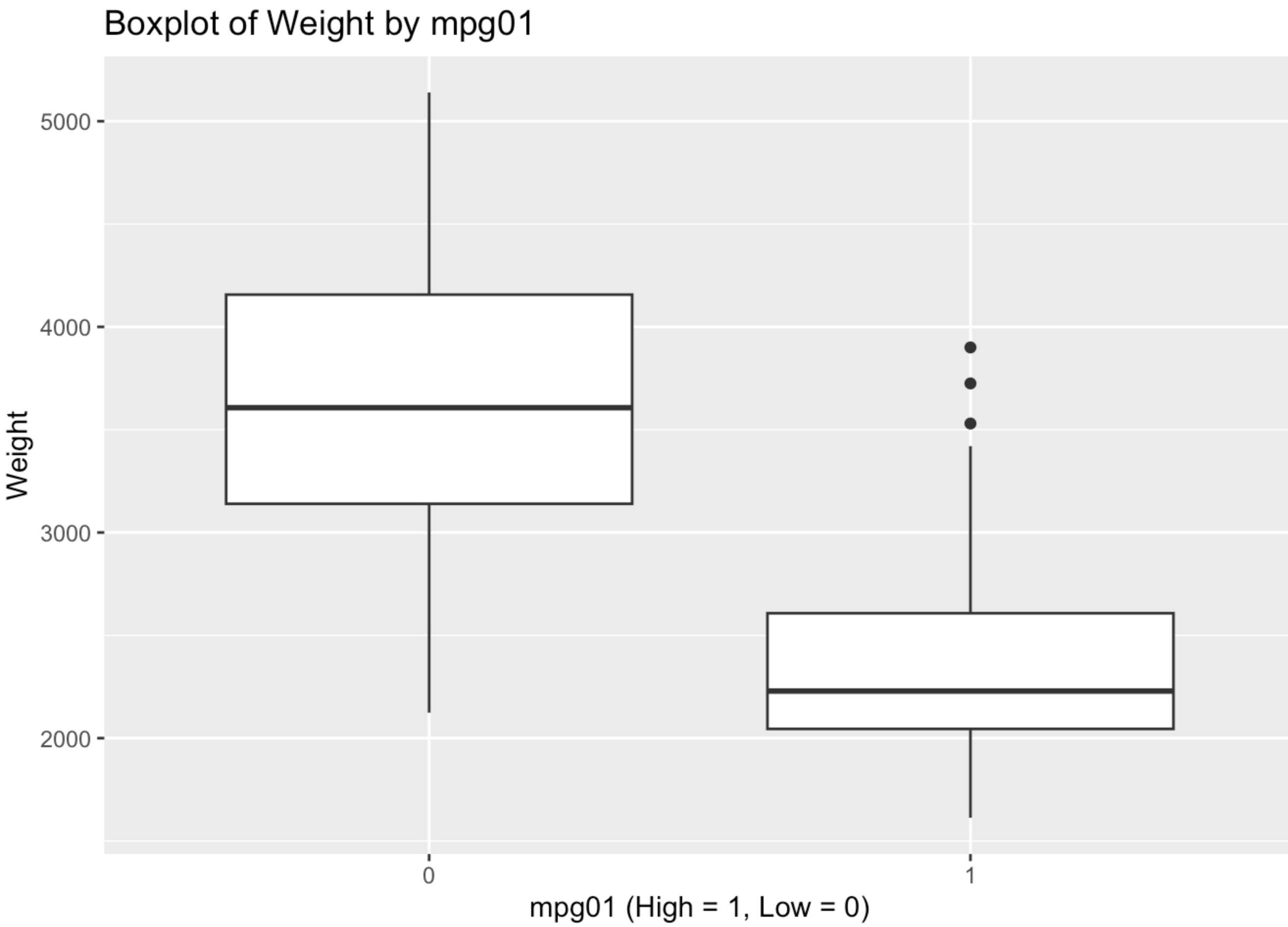


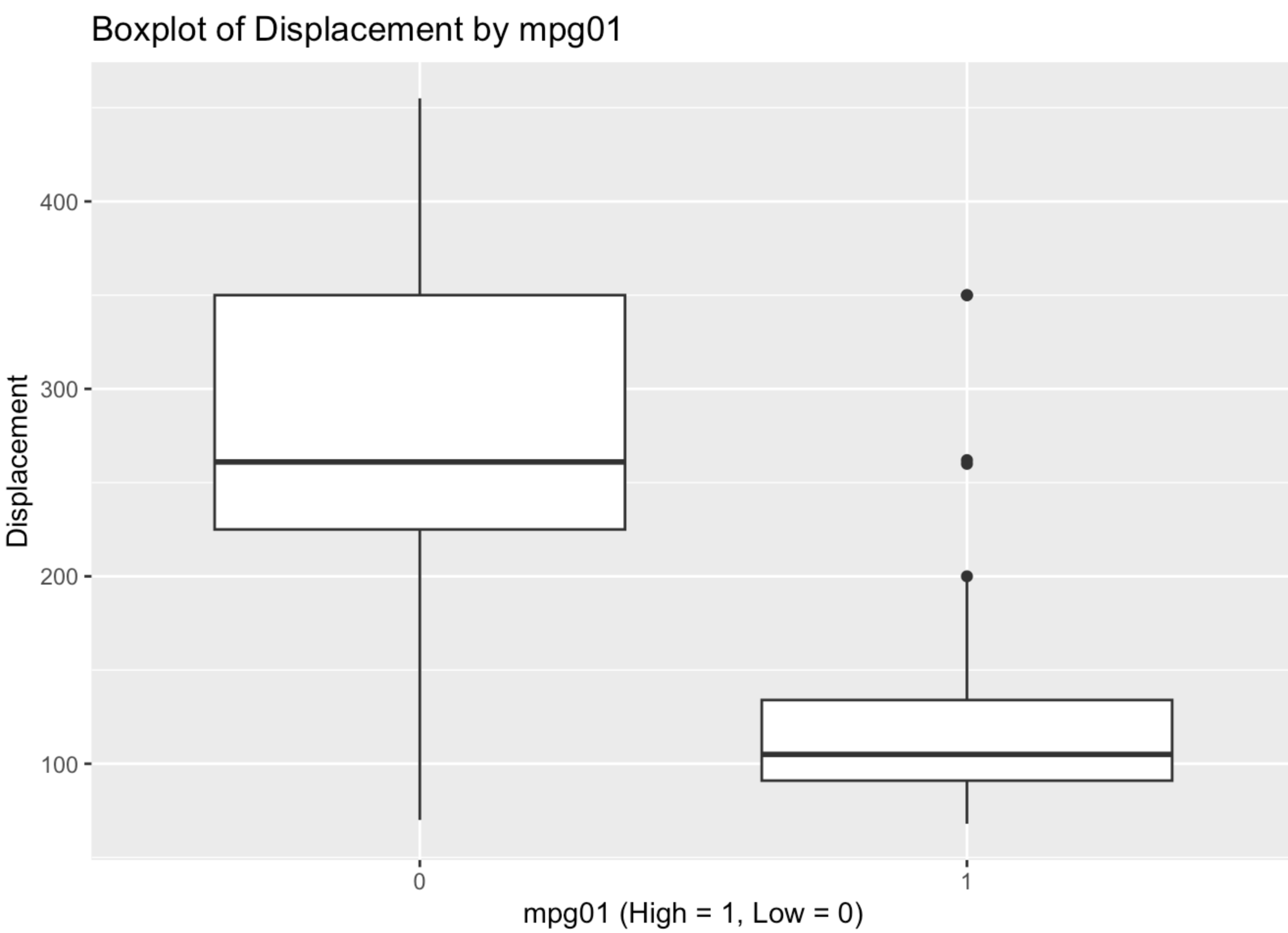

Weight vs. mpg01
mpg01 (High = 1, Low = 0)
1.00
0.75
0.50
0.25
0.00
2000
3000
4000
5000
Weight

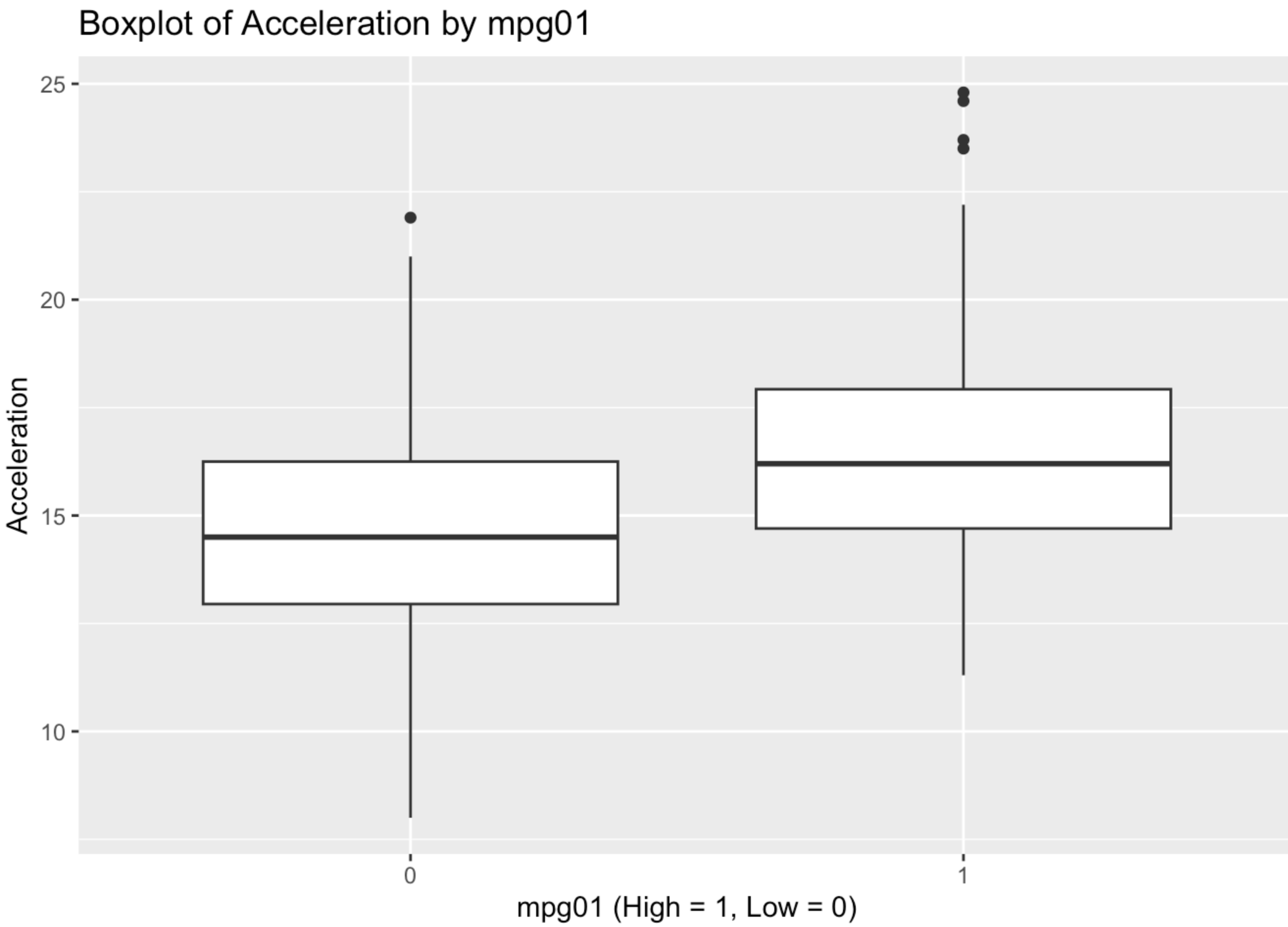

Displacement vs. mpg01
mpg01 (High = 1, Low = 0)
1.00
0.75
0.50
0.25
0.00
100
200
300
400
Displacement


Acceleration vs. mpg01
mpg01 (High = 1, Low = 0)
1.00
0.75
0.50
0.25
0.00
10
15
20
25
Acceleration


Boxplot of Horsepower by mpg01
Horsepower
200
150
100
50
0
1
mpg01 (High = 1, Low = 0)


Boxplot of Weight by mpg01
5000
4000
3000
2000
Weight
0
1
mpg01 (High = 1, Low = 0)


Boxplot of Displacement by mpg01
400
300
200
100
Displacement
0
1
mpg01 (High = 1, Low = 0)



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

d. Perform LDA 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?

```
##  [1] mpg01        mpg          cylinders    displacement horsepower  
##  [6] weight       acceleration year         origin       name        
## <0 rows> (or 0-length row.names)
```

```
## [1] 0.1326531
```

e. Perform QDA 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?

```
## [1] 0.122449
```

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?

```
## [1] 0.1122449
```

g. Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

```
##  [1] 0.1632653 0.1479592 0.1428571 0.1275510 0.1122449 0.1173469 0.1071429
##  [8] 0.1224490 0.1071429 0.1173469
```

```
## [1] 7
```