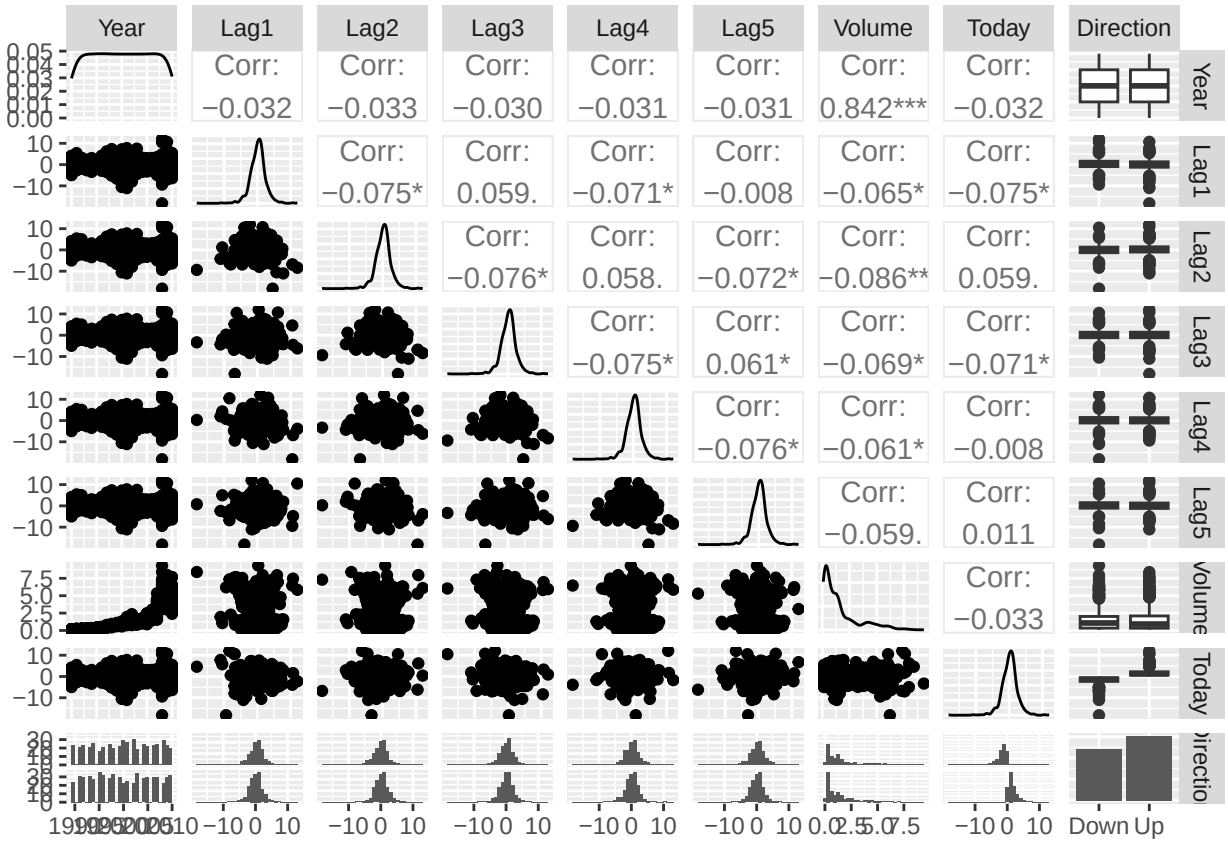
# Modern Applied Statistics exercises from ISLR

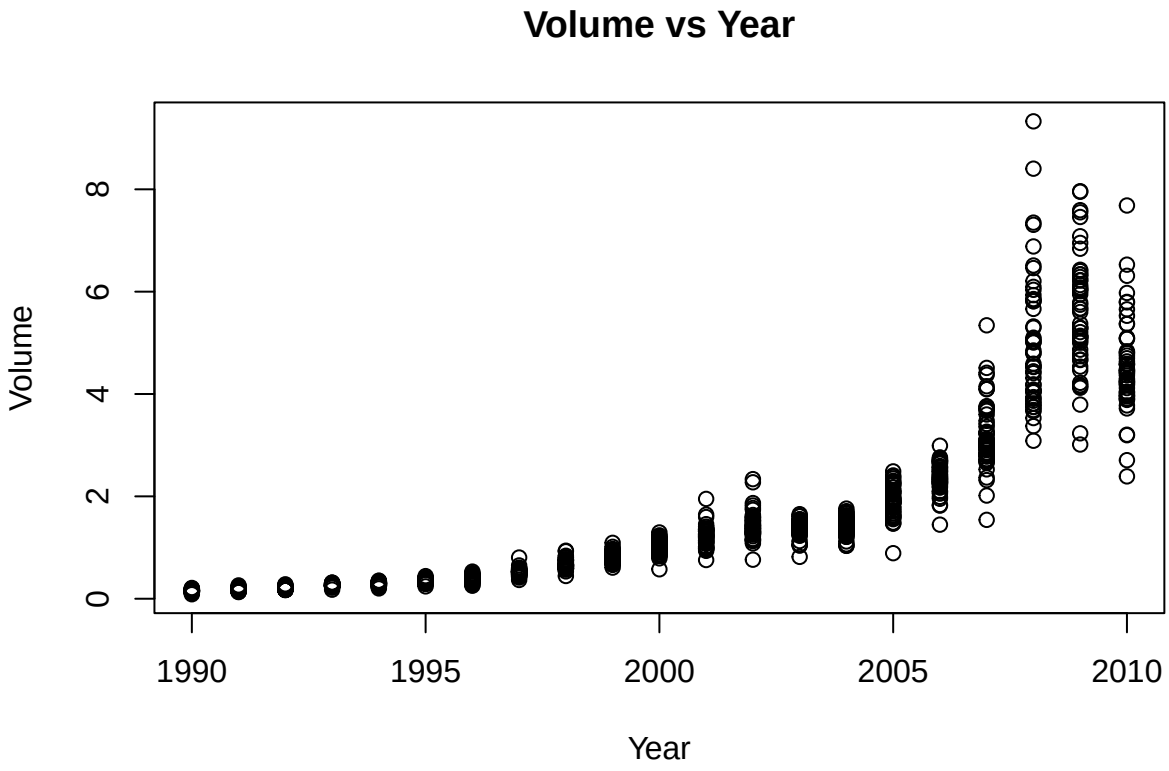
Yamuna Dhungana

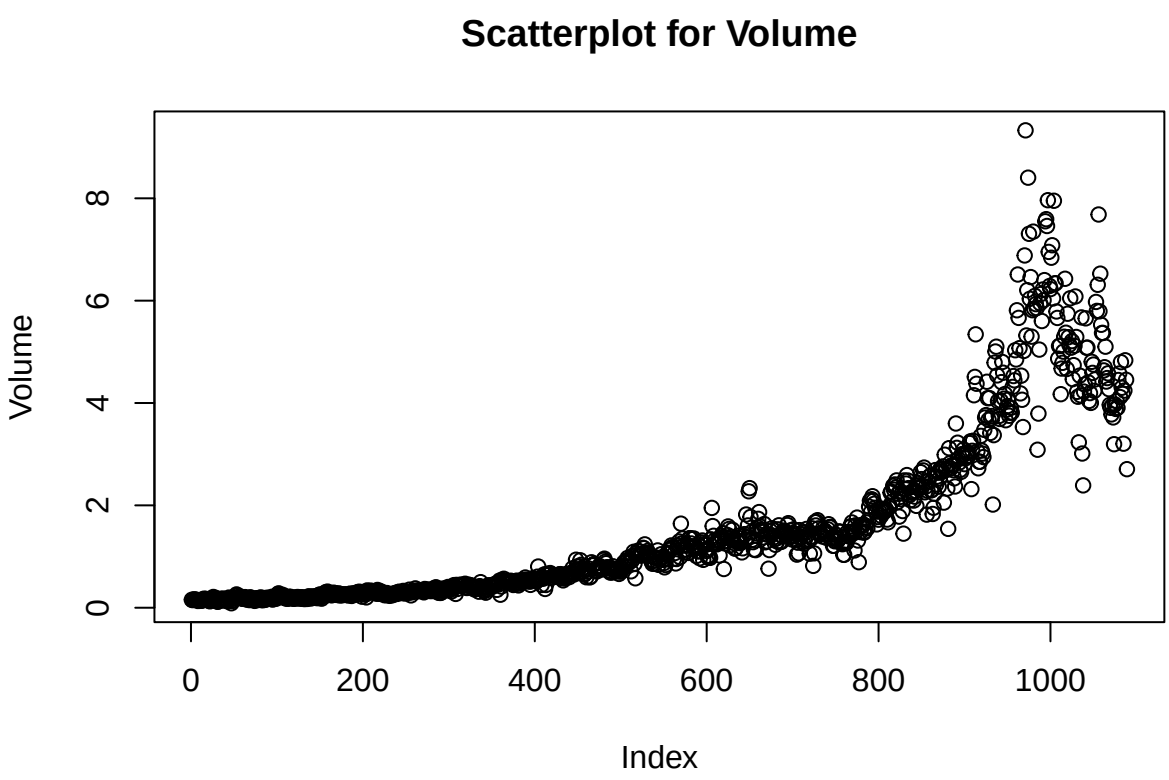
This analysis employs the Weekly dataset, which covers the percentage returns for the S&P 500 stock index spanning from 1990 to 2010. The dataset is organized as a data frame with 1089 observations related to nine variables: Year, Lag1, Lag2, Lag3, Lag4, Lag5, Volume, Today, and Direction. The following numerical and graphical summaries are presented to identify any discernible patterns.

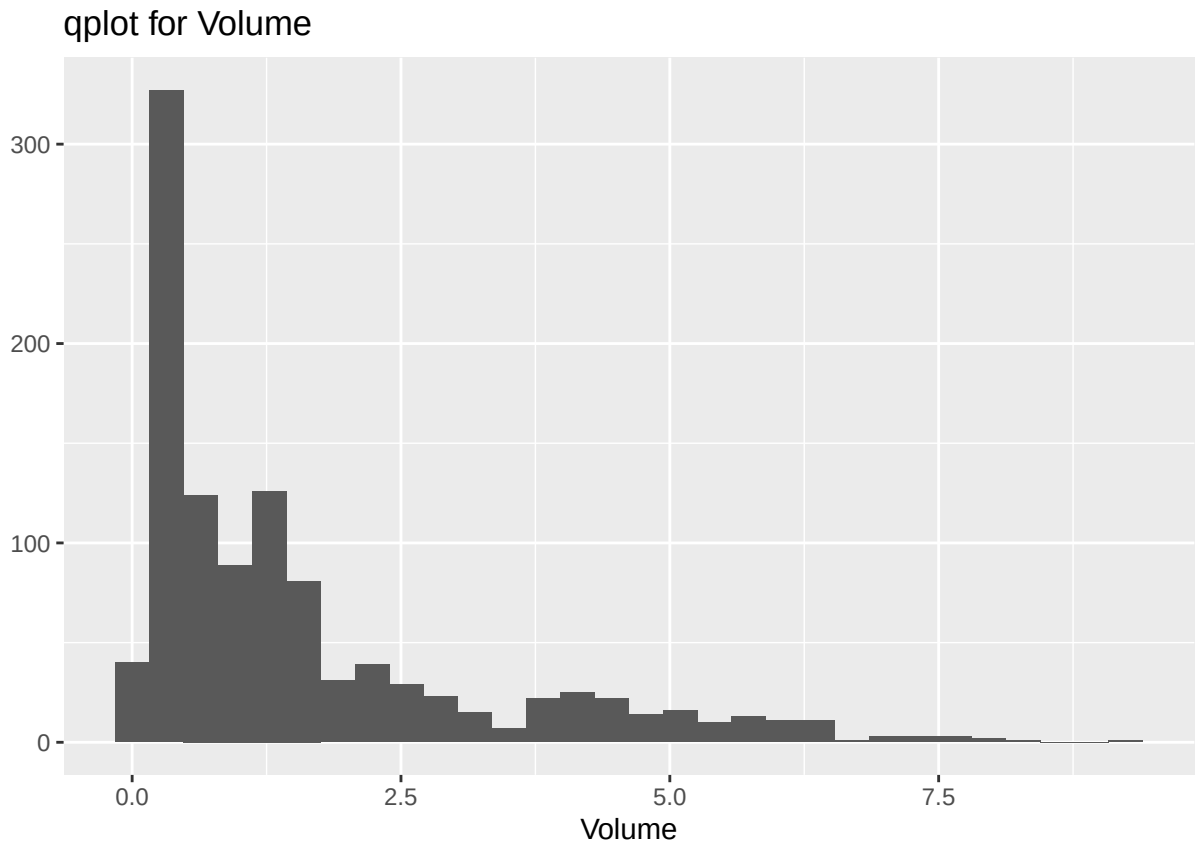
```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  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            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
##
##
##
##
```

```
##                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
Lag5
Volume
Today
Direction
Corr: −0.032
Corr: −0.033
Corr: −0.030
Corr: −0.031
Corr: −0.031
Corr: 0.842***
Corr: −0.032
Corr: −0.075*
Corr: 0.059.
Corr: −0.071*
Corr: −0.008
Corr: −0.065*
Corr: −0.075*
Corr: −0.076*
Corr: 0.058.
Corr: −0.072*
Corr: −0.086**
Corr: 0.059.
Corr: −0.075*
Corr: 0.061*
Corr: −0.069*
Corr: −0.071*
Corr: −0.076*
Corr: −0.061*
Corr: −0.008
Corr: −0.059.
Corr: 0.011
Corr: −0.033
Down
Up


Volume vs Year
Volume
0
2
4
6
8
1990
1995
2000
2005
2010
Year


Scatterplot for Volume
Volume
0
2
4
6
8
0
200
400
600
800
1000
Index




The correlation analysis of the 'weekly' dataset reveals a robust association between volume and year. In contrast, other variables do not exhibit a similarly pronounced correlation. Additionally, a visualization of the year and volume variables suggests a gradual exponential rise from 1995 to 2004. Subsequently, for the subsequent years, there appears to be a consistent increase in volume, with a slight decline noted in 2010. Conducting logistic regression on the entire dataset involves employing Direction as the response variable and utilizing the five lag variables along with Volume as predictors. Additionally, an examination will be conducted to assess the statistical significance of the regression results.

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

According to the model summary, it is evident that only lag2 exhibits statistical significance, with a p-value of 0.0296, meeting the criteria of $P < 0.05$. The estimated coefficient for lag2 is 0.05844, signifying that, holding the other predictors constant, there is an anticipated mean increase in log odds as the stock market rises by a unit increase in lag2. Beyond this, the deviance residual of the model indicates a positive skewness in the data. I am calculating the confusion matrix and the overall fraction of correct predictions. Additionally, I aim to identify the specific types of errors made by the logistic regression model.

```
## [1] "Confusion Matrix:"
##
## preds  Down  Up
##   Down   54  48
##   Up    430 557
```

The confusion matrix delineates correct and erroneous predictions made by the model. It comprises four distinct factors: True Positive, True Negative, False Positive, and False Negative. True Positive and True Negative signify correct predictions, while False Positive and False Negative denote incorrect ones. In our matrix, the model accurately predicted the direction as up and down in 557 and 54 instances, respectively. The value 48 represents false positives, where the model predicted an upward direction, but the actual direction was down. The value 430 indicates false negatives, signifying instances where the model predicted a downward direction, but the actual direction was up.

Furthermore, we can calculate the test error from the matrix using the formula `(54 + 48) / 1089`, yielding a percentage of correct predictions at 56.10%. Additionally, if the model predicts an upward direction, it will be correct 92.06% of the time (`557 / (48 + 557)`), while for a downward direction, the correctness rate is 11.15% (`54 / (54 + 430)`).

Now, I am fitting the logistic regression model using training data spanning from 1990 to 2008, where Lag2 serves as the sole predictor. Following this, I will compute the confusion matrix and determine the overall fraction of correct predictions for the held-out data, specifically the data from 2009 and 2010.

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

```
## Down   Up
##   43   61
```

```
## [1] "Confusion Matrix:"
##
## preds  Down  Up
##   Down   32  25
##   Up    452 580
```

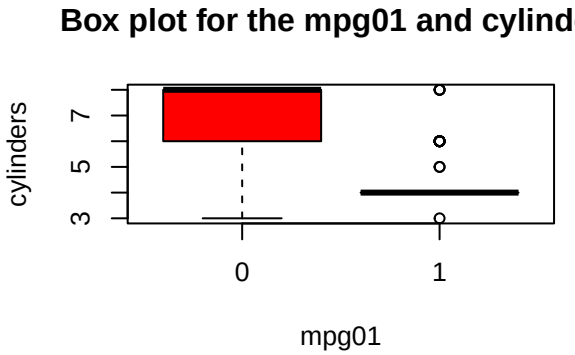
In our model, there are 43 instances of the total data being down and 61 instances being up. Within the confusion matrix, our accurate predictions for the upward and downward directions are 580 and 32, respectively. The value 25 represents false positives, indicating instances where the model predicted an upward direction, but the actual direction was down. The value 452 represents false negatives, signifying cases where the model predicted a downward direction, but the actual direction was up.

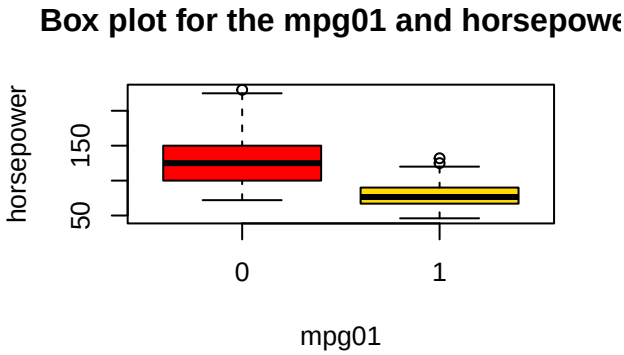
Furthermore, the test error can be computed from the matrix using the formula `(32 + 25) / 1089`, resulting in a percentage of correct predictions at 56.19%. Specifically, when the model predicts an upward direction, it is correct 95.86% of the time (`580 / (25 + 580)`), while for a downward direction, the correctness rate is 5.22% (`32 / (32 + 580)`).

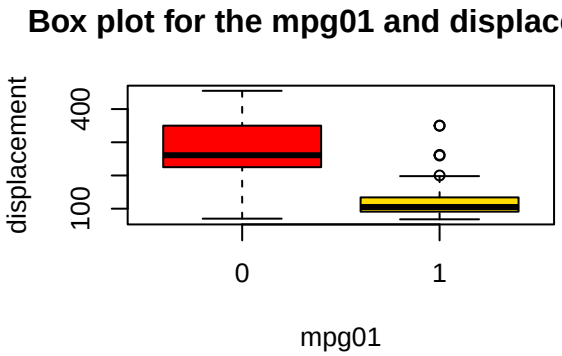
In an attempt to develop a model for predicting whether a given car has high or low gas mileage using the Auto dataset, I am creating a binary variable named 'mpg01.' This variable takes the value 1 if the 'mpg' variable contains a value above its median and 0 if 'mpg' contains a value below its median. The median can be computed using the median() function.

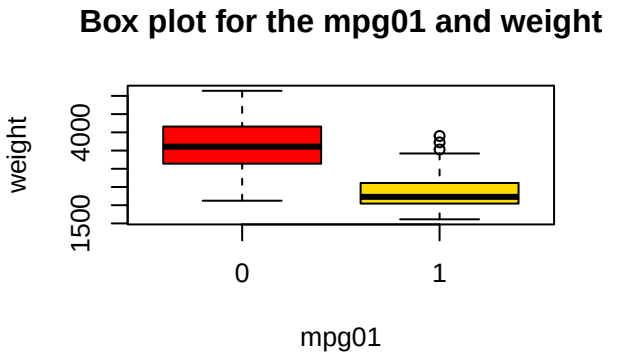
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      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 mpg01
## 1 chevrolet chevelle malibu     0
## 2         buick skylark 320     0
## 3        plymouth satellite     0
## 4             amc rebel sst     0
## 5               ford torino     0
## 6          ford galaxie 500     0
```

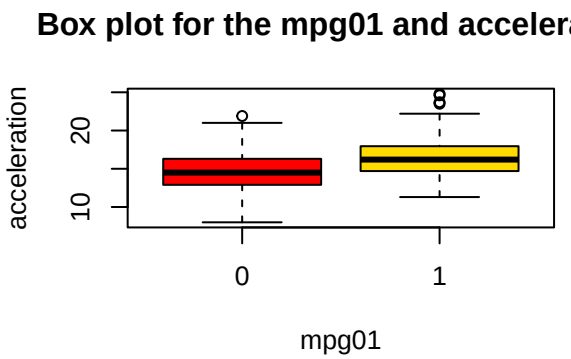
Examining the data graphically to explore the relationship between 'mpg01' and the remaining features. Identifying which of the other features appear to be most relevant for predicting 'mpg01' by plotting scatterplots and boxplots.

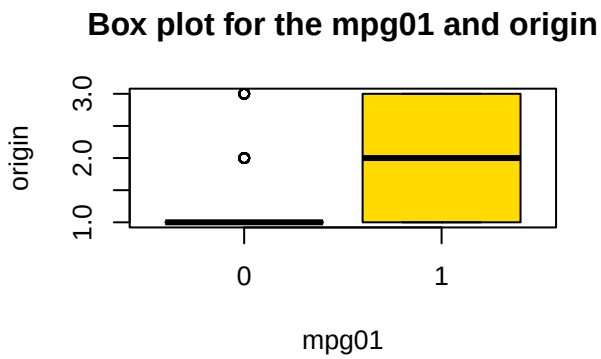

Box plot for the mpg01 and cylinders
cylinders
3
5
7
0
1
mpg01

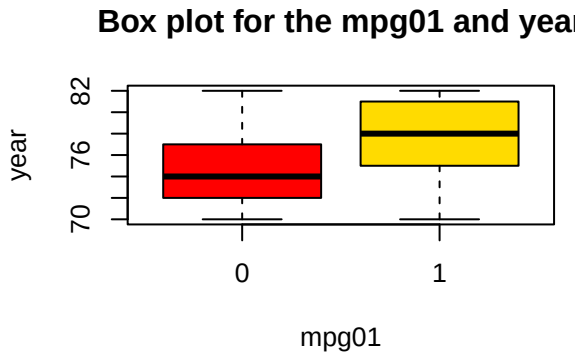

Box plot for the mpg01 and horsepowe
horsepower
50
150
0
1
mpg01

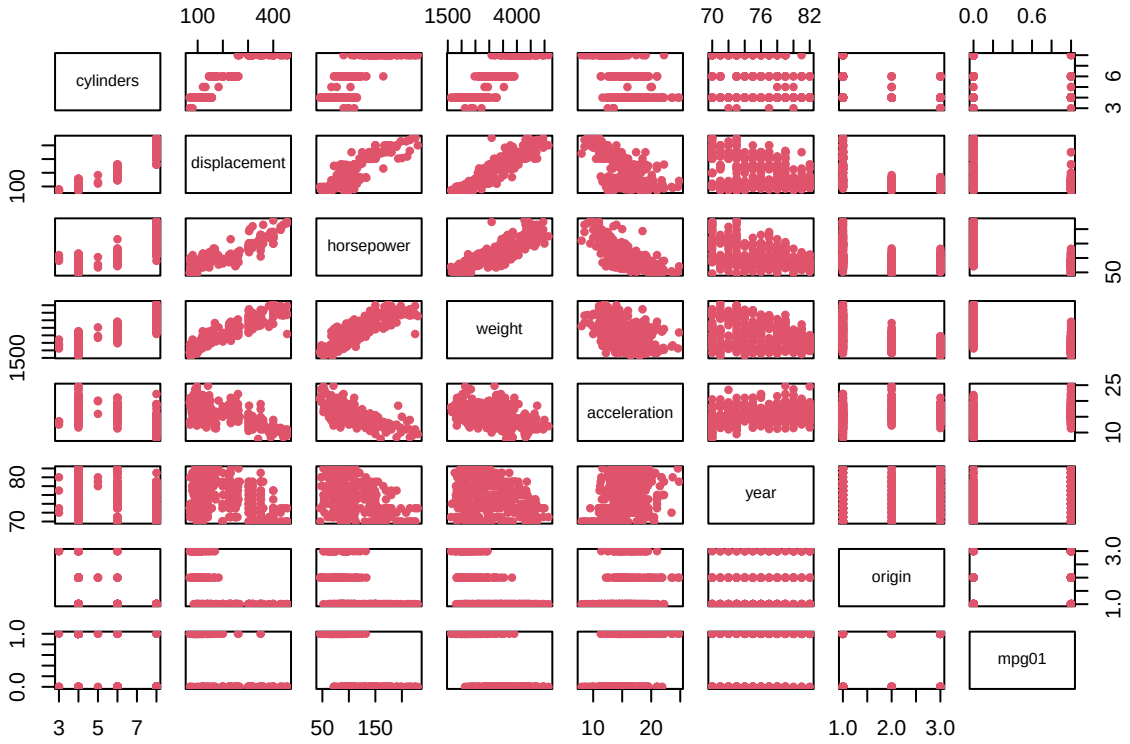

Box plot for the mpg01 and displaceme
displacement
100
400
0
1
mpg01


Box plot for the mpg01 and weight
weight
1500
4000
0
1
mpg01

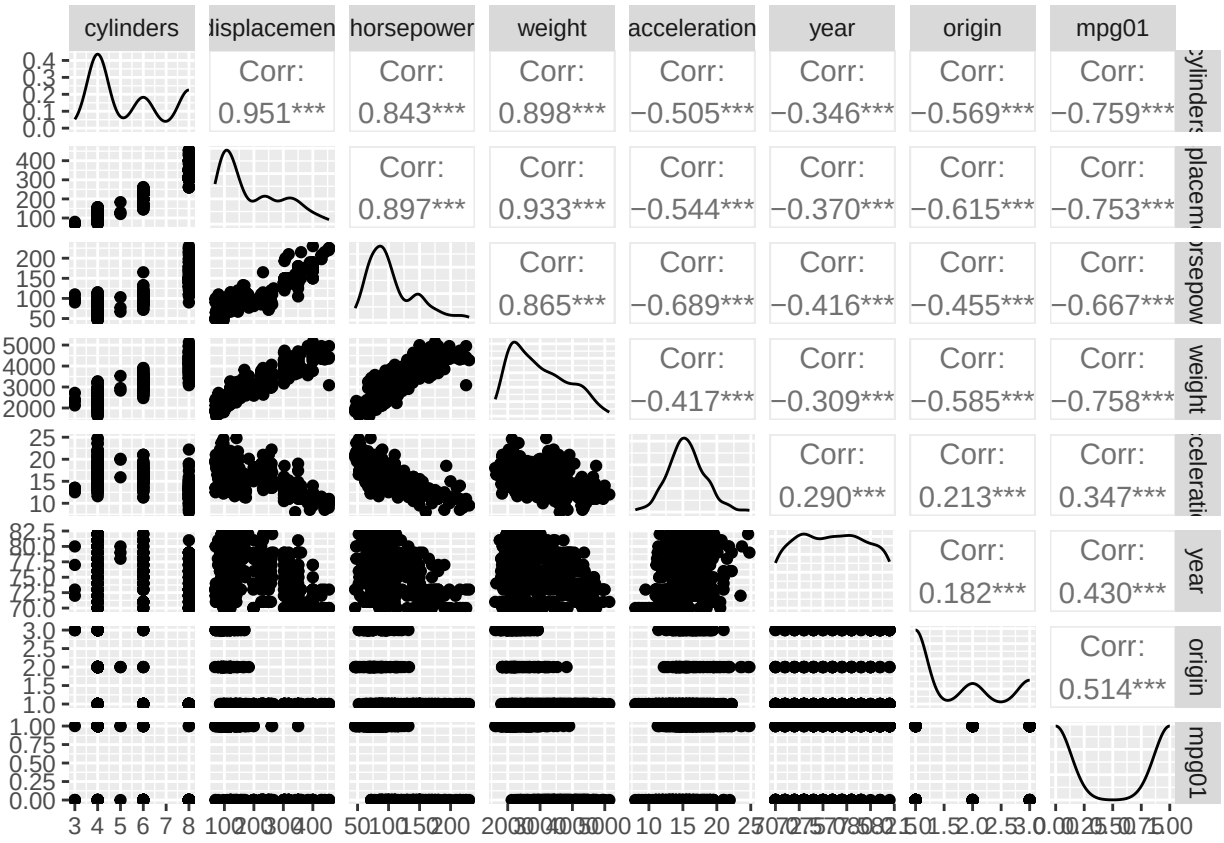

Box plot for the mpg01 and acceleratio
acceleration
10
20
0
1
mpg01


Box plot for the mpg01 and origin
origin
1.0
2.0
3.0
0
1
mpg01


Box plot for the mpg01 and year
year
70
76
82
0
1
mpg01




```
##                cylinders displacement horsepower     weight acceleration
## cylinders      1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834
## displacement   0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005
## horsepower     0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955
## weight         0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392
## acceleration  -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000
## year          -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161
## origin        -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458
## mpg01         -0.7591939   -0.7534766 -0.6670526 -0.7577566    0.3468215
##                      year     origin      mpg01
## cylinders     -0.3456474 -0.5689316 -0.7591939
## displacement  -0.3698552 -0.6145351 -0.7534766
## horsepower    -0.4163615 -0.4551715 -0.6670526
## weight        -0.3091199 -0.5850054 -0.7577566
## acceleration   0.2903161  0.2127458  0.3468215
## year           1.0000000  0.1815277  0.4299042
## origin         0.1815277  1.0000000  0.5136984
## mpg01          0.4299042  0.5136984  1.0000000
```




The box plot clearly indicates a discernible distinction in the distribution between two groups for the variables cylinders, horsepower, displacement, weight, origin, and year. Notably, a majority of the automobiles originated in Japan. Cars from the United States are predominantly concentrated at lower mpg, whereas European and Japanese cars exhibit a more even distribution. Additionally, older cars generally tend to have lower mpg, while modern cars tend to have higher mpg.

Moreover, the correlation plot reveals significant correlations among the physical attributes of the car. Notably, there appears to be a high correlation between displacement and horsepower, suggesting an exponential relationship between the two. Splitted the data in the ration of 70% and 30% .

Conducting logistic regression on the training data to predict 'mpg01' using the variables that exhibited the strongest associations with 'mpg01.'

```
##
## Call:
## glm(formula = mpg01 ~ cylinders + weight + displacement + horsepower,
##     family = binomial, data = train)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.0170868  1.9377971   5.685 1.31e-08 ***
## cylinders     0.3252329  0.3959873   0.821  0.41146
## weight       -0.0017252  0.0008048  -2.144  0.03205 *
## displacement -0.0197255  0.0097760  -2.018  0.04362 *
## horsepower   -0.0490093  0.0169575  -2.890  0.00385 **
## ---
## 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: 149.91  on 269  degrees of freedom
## AIC: 159.91
##
## Number of Fisher Scoring iterations: 7
```

```
##
## preds  0  1
##     0 50  8
##     1  3 57
```

```
## [1] "Test error (percantage): 9.32"
```

Based on the findings from question b, where cylinders, weight, displacement, and horsepower were identified as variables most associated with 'mpg01,' logistic regression was performed using these variables. In the computed model, it was determined that weight and horsepower are statistically significant predictors. Additionally, the model exhibited a negative skewness in the data.

For evaluating test accuracy, a confusion matrix was generated, revealing that 88.14% of the data was correctly predicted, while 11.86% was predicted incorrectly. Consequently, the test error for the model stands at 11.86%.