FA582 Assignment 3 Report

Naveen Mathews Renji - 20016323

Problem 1

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

I started with the **summary** of the Weekly.csv dataset

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

Numerical data analysis.

	means	medians	modes	sds	variances	skewness	kurtoses
Lag1	0.1505849	0.24100	0.241000	2.357013	5.555508	-0.4806399	5.668866
Lag2	0.1510790	0.24100	0.241000	2.357254	5.556647	-0.4809673	5.665817
Lag3	0.1472048	0.24100	0.241000	2.360502	5.571970	-0.4788321	5.623316
Lag4	0.1458182	0.23800	0.241000	2.360279	5.570916	-0.4773127	5.625406
Lag5	0.1398926	0.23400	0.241000	2.361285	5.575665	-0.4740865	5.609173
Volume	1.5746176	1.00268	0.154976	1.686636	2.844742	1.6159582	2.062587

This allowed me to find the mean, median, mode, standard deviation, variance, skewness and kurtoses of the 5 lag variables and the volume variable. The weekly returns for Lag1 to Lag5 averaged slightly above zero with consistent median values, hinting at modest gains. All lag variables shared a mode of 0.241, which is noteworthy and might reflect data collection methods. High standard deviations and variances across these variables indicate market volatility, and negative skewness suggests fewer extreme gains than losses. The mean trading volume was much higher than the median, with a lower mode, indicating a right-skewed distribution with occasional spikes in trading activity. Elevated kurtosis in the lag variables implies the presence of significant outliers, while the volume's kurtosis above normal suggests outliers in trading volumes too.

Correlation Analysis:

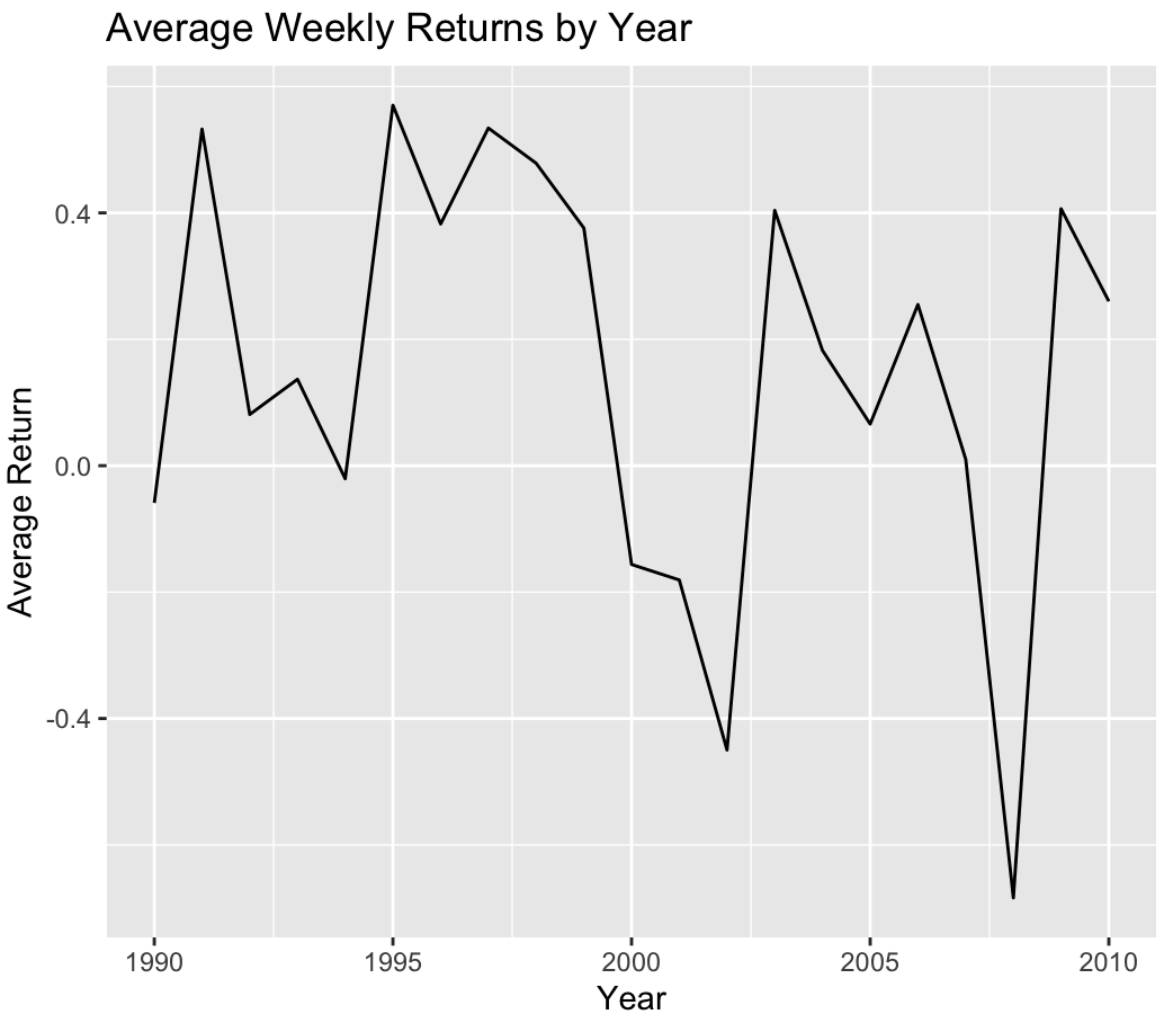
	Lag1	Lag2	Lag3	Lag4	Lag5	Volume
Lag1	1.000000000	-0.07485305	0.05863568	-0.07127388	-0.008183096	-0.06495131
Lag2	-0.074853051	1.00000000	-0.07572091	0.05838153	-0.072499482	-0.08551314
Lag3	0.058635682	-0.07572091	1.00000000	-0.07539587	0.060657175	-0.06928771
Lag4	-0.071273876	0.05838153	-0.07539587	1.00000000	-0.075675027	-0.06107462
Lag5	-0.008183096	-0.07249948	0.06065717	-0.07567503	1.000000000	-0.05851741
Volume	-0.064951313	-0.08551314	-0.06928771	-0.06107462	-0.058517414	1.00000000

The correlation matrix revealed very low correlations between all lag variables and between lag variables and volume. All correlations were below 0.1 in absolute value, indicating a very weak linear relationship. This suggests that past weekly returns and volume do not strongly linearly predict each other in the subsequent week.

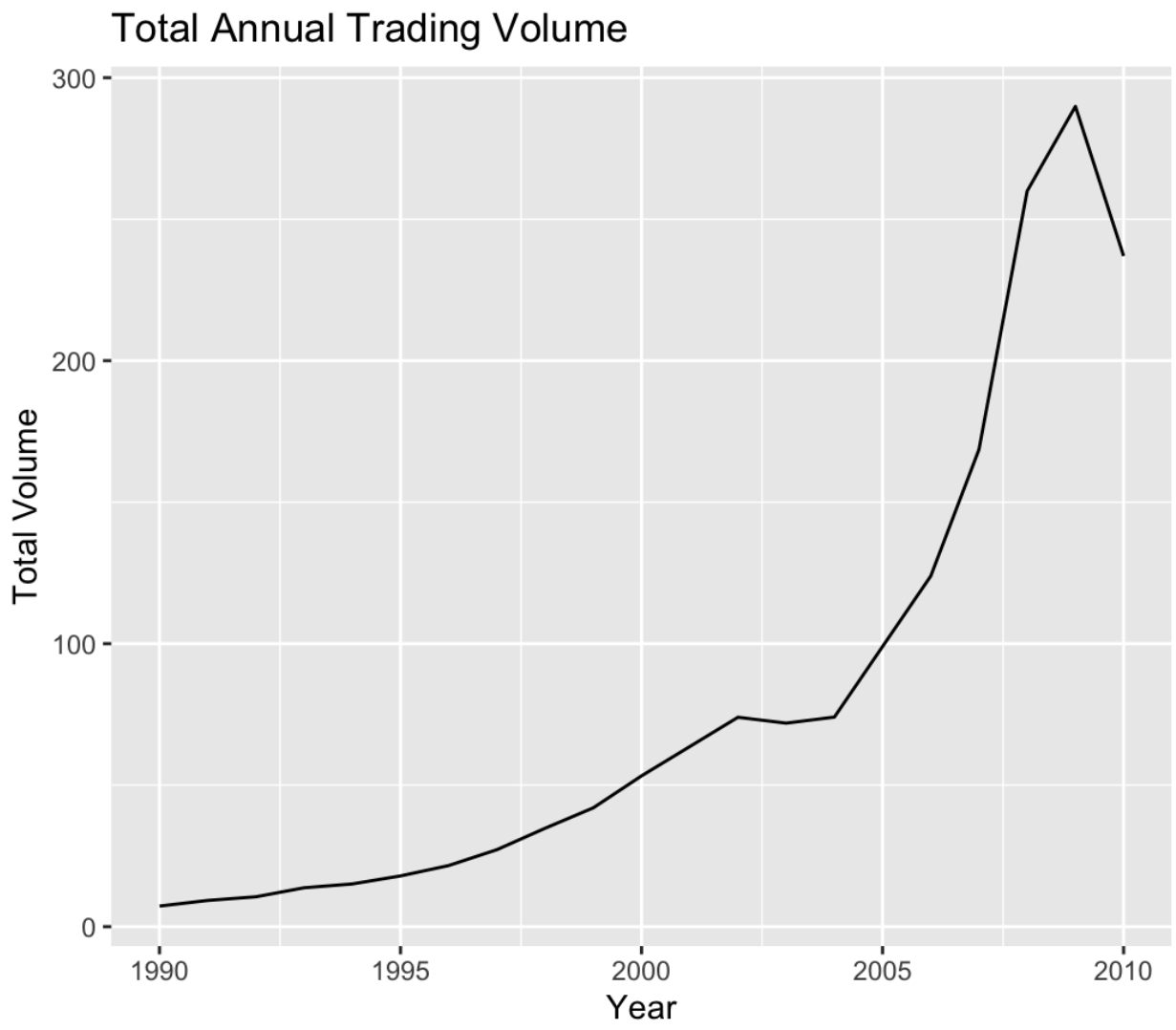
Autocorrelation of Today variable:
The autocorrelation of the Today variable was slightly negative at -0.07499, which suggests that if the market moved in one direction in one week, it had a very slight tendency to move in the opposite direction the following week. However, the strength of this relationship is very weak.

Frequency Counts:
The direction of the market was 'Up' in 605 weeks and 'Down' in 484 weeks. This indicates that the market was more often positive than negative in this period.

Time Series Analysis:
I first aggregated the data by year to get average 'Today' values, then I plotted the average weekly returns and the total trading volume over the years.



This plot shows us the sharp decline in the average return during the years between 1998 and 2002 and again between 2007 and 2008. We are also able to see a sharp rise right after the two declines, this could indicate that even though the market can collapse due to certain circumstances, there is a certainty that it will climb back soon.



This plot is a clear depiction of the increase in trend of trading in the stock market. This is a sign of a positively developing economy and more companies and shares being bought and sold. We are able to notice a plateau in the volume rise similar to the first dip in the weekly gains and we can see a steep decrease in the volume corresponding to the second dip in weekly gains mentioned above.

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?

First I converted the Direction variable as a factor and then I used the glm logistic regression model from R using the 5 Lag variables and Volume as predictors

```
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
```

Lag2 has a p-value of 0.0296, which is below the 0.05 threshold typically used for statistical significance. This suggests that Lag2 is a statistically significant predictor of Direction.

The (Intercept) also appears to be statistically significant, with a p-value of 0.0019, indicating that the model predicts a significant effect on the response variable even when all predictors are zero.

The other variables (Lag1, Lag3, Lag4, Lag5, and Volume) have p-values greater than 0.05, suggesting that they are not statistically significant at the 5% level. In other words, they do not have a statistically significant relationship with the Direction of the market, based on this model and the data provided.

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.

To do this I first predicted the probabilities and then classified them as Up or Down, then I made the confusion matrix and calculated the Accuracy.

	Actual	
Predicted	Down	Up
Down	54	48
Up	430	557

The confusion matrix for my logistic regression model reveals that it's better at predicting 'Up' weeks than 'Down' weeks. It correctly identified 557 out of 605 'Up' weeks, but only correctly predicted 54 out of 484 'Down' weeks. This suggests a conservative tendency in the model, often mislabeling actual 'Up' weeks as 'Down'. False negatives are quite high, with the model missing many 'Up' weeks. Overall, the model's **accuracy stands at about 56.1%,** which indicates it's performing better than random chance. However, this level of accuracy still leaves room for improvement, especially in correctly identifying 'Down' weeks.

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).

To do this, we split the data into training (1990 and 2008) and test (2009 and 2010).

	Actual	
Predicted	Down	Up
Down	9	5
Up	34	56

The accuracy this time is higher, up by around 6% to 62.5%. This is a positive indication that the model is performing better when Lag2 is the sole predictor which also aligns with our analysis of Lag2 being the most significant predictor.

Repeat the above using LDA, QDA and KNN with K = 1 and Which of these methods appears to provide the best results on this data?

```
> lda_confusion_matrix
          Actual
Predicted Down Up
     Down    9  5
     Up     34 56
> qda_confusion_matrix
          Actual
Predicted Down Up
     Down    0  0
     Up     43 61
> knn_confusion_matrix
          Actual
Predicted Down Up
     Down   21 30
     Up     22 31
```

```
> lda_accuracy
[1] 0.625
> qda_accuracy
[1] 0.5865385
> knn_accuracy
[1] 0.5096154
```

These results indicate that the LDA method appears to give us the best results on this data with a higher accuracy than QDA and KNN with K = 1.

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.

To do this I divided the dataset into a training set from the years 1990 to 2008 and a test set from 2009 to 2010. Using LDA and QDA, I tested different combinations of predictors, including Lag2 and its interaction with Volume. For KNN, I tried various numbers of neighbors (k-values) with similar predictor sets. For each model, I computed the accuracy by comparing the predictions against the actual market directions in the test set. Finally, I identified the best-performing model based on the highest accuracy, giving me the most effective method and predictors for this data. The below depicts the variables used, the method and the associated accuracy which is derived from the confusion matrix.

	Method	Variables	Accuracy
1	LDA	Direction ~ Lag2	0.6250000
2	LDA	Direction ~ Lag2 + Lag1	0.5769231
3	LDA	Direction ~ Lag2 * Volume	0.5384615
4	QDA	Direction ~ Lag2	0.5865385
5	QDA	Direction ~ Lag2 + Lag1	0.5576923
6	QDA	Direction ~ Lag2 * Volume	0.4711538
7	KNN 1	~Lag2	0.5096154
8	KNN 1	~Lag2 + Lag1	0.4807692
9	KNN 1	~Lag2 * Volume	0.5288462
10	KNN 2	~Lag2	0.5288462
11	KNN 2	~Lag2 + Lag1	0.4903846
12	KNN 2	~Lag2 * Volume	0.4903846
13	KNN 3	~Lag2	0.5480769
14	KNN 3	~Lag2 + Lag1	0.5192308
15	KNN 3	~Lag2 * Volume	0.4711538
16	KNN 4	~Lag2	0.5480769
17	KNN 4	~Lag2 + Lag1	0.5000000
18	KNN 4	~Lag2 * Volume	0.4711538
19	KNN 5	~Lag2	0.5480769
20	KNN 5	~Lag2 + Lag1	0.4903846
21	KNN 5	~Lag2 * Volume	0.5096154
22	KNN 6	~Lag2	0.5769231
23	KNN 6	~Lag2 + Lag1	0.5865385
24	KNN 6	~Lag2 * Volume	0.5000000
25	KNN 7	~Lag2	0.5576923
26	KNN 7	~Lag2 + Lag1	0.5288462
27	KNN 7	~Lag2 * Volume	0.5000000
28	KNN 8	~Lag2	0.5576923
29	KNN 8	~Lag2 + Lag1	0.5096154
30	KNN 8	~Lag2 * Volume	0.4903846
31	KNN 9	~Lag2	0.5576923
32	KNN 9	~Lag2 + Lag1	0.5384615
33	KNN 9	~Lag2 * Volume	0.4615385
34	KNN 10	~Lag2	0.5673077
35	KNN 10	~Lag2 + Lag1	0.4711538
36	KNN 10	~Lag2 * Volume	0.5192308
37	KNN 11	~Lag2	0.5480769
38	KNN 11	~Lag2 + Lag1	0.4903846
39	KNN 11	~Lag2 * Volume	0.4711538
40	KNN 12	~Lag2	0.5865385
41	KNN 12	~Lag2 + Lag1	0.5096154
42	KNN 12	~Lag2 * Volume	0.4807692
43	KNN 13	~Lag2	0.5769231

```
44 KNN 13             ~Lag2 + Lag1 0.5096154
45 KNN 13           ~Lag2 * Volume 0.5288462
46 KNN 14                    ~Lag2 0.5961538
47 KNN 14             ~Lag2 + Lag1 0.5000000
48 KNN 14           ~Lag2 * Volume 0.5480769
49 KNN 15                    ~Lag2 0.5865385
50 KNN 15             ~Lag2 + Lag1 0.5769231
51 KNN 15           ~Lag2 * Volume 0.5384615
52 KNN 16                    ~Lag2 0.5769231
53 KNN 16             ~Lag2 + Lag1 0.5288462
54 KNN 16           ~Lag2 * Volume 0.5000000
55 KNN 17                    ~Lag2 0.5961538
56 KNN 17             ~Lag2 + Lag1 0.5384615
57 KNN 17           ~Lag2 * Volume 0.5000000
58 KNN 18                    ~Lag2 0.5769231
59 KNN 18             ~Lag2 + Lag1 0.5384615
60 KNN 18           ~Lag2 * Volume 0.5096154
61 KNN 19                    ~Lag2 0.5576923
62 KNN 19             ~Lag2 + Lag1 0.5384615
63 KNN 19           ~Lag2 * Volume 0.5384615
64 KNN 20                    ~Lag2 0.5865385
65 KNN 20             ~Lag2 + Lag1 0.5480769
66 KNN 20           ~Lag2 * Volume 0.5192308
```

```
> print(best_model)
  Method         Variables Accuracy
1    LDA Direction ~ Lag2    0.625
```

```
Direction ~ Lag2
          Actual
Predicted Down Up
     Down    9  5
     Up     34 56
```

This LDA model with Lag2 as the predictor gave us the best accuracy and its confusion matrix shows us that is was able to predict 56 Up accurately and misclassified on 5 Up and 9 actual Down as Down and misclassified 34 down as up indicating that it is more likely to correctly classify up and misclassify down more often.

Conclusion

I conducted a series of analyses to predict market direction using the Weekly dataset, exploring different statistical methods and predictor combinations. The dataset was split into a training set (1990-2008) and a test set (2009-2010). I applied Logistic Regression, LDA, QDA, and KNN models, initially using multiple predictors, then narrowed down to Lag2 based on its statistical significance.

The Logistic Regression model revealed Lag2 as a significant predictor. Using this insight, I built LDA and QDA models with Lag2 and experimented with KNN, adjusting the number of neighbors. Each model's performance was evaluated on the test set, and accuracy was calculated using confusion matrices.

The findings showed varying degrees of accuracy, with none of the models performing exceptionally well, indicating the challenging and noisy nature of stock market prediction. The models did better than random chance but highlighted the conservative bias of these methods, often misclassifying 'Up' weeks.

In conclusion, while Lag2 emerged as a valuable predictor and the LDA with Lag2 predictor emerged as the most accurate of all the experimented models,, the search for a robust predictive model for the market direction suggested a need for more nuanced approaches, possibly incorporating additional data and sophisticated modeling techniques.

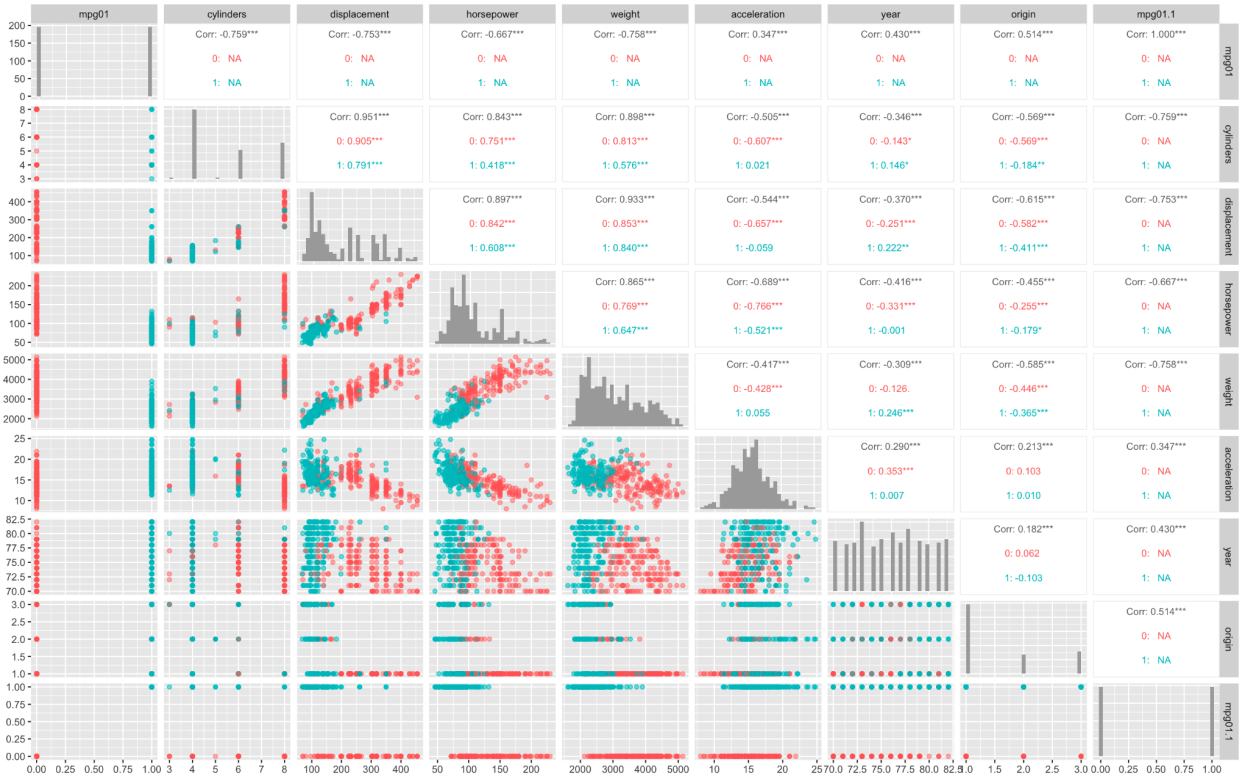
Problem 2

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.

This was done pretty straightforwardly just as described in the question.

```
      mpg          cylinders      displacement     horsepower        weight      acceleration        year      
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00  
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00  
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50   Median :76.00  
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54   Mean   :75.98  
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00  
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00  
     origin          name                mpg01       
 Min.   :1.000   Length:392         Min.   :0.0  
 1st Qu.:1.000   Class :character   1st Qu.:0.0  
 Median :1.000   Mode  :character   Median :0.5  
 Mean   :1.577                      Mean   :0.5  
 3rd Qu.:2.000                      3rd Qu.:1.0  
 Max.   :3.000                      Max.   :1.0  
```

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? Scatter Plots and boxplots may be useful tools to answer this question. Describe your findings.



Upon reviewing the scatter plot matrix, several insights emerge regarding how various vehicle features relate to `mpg01`. The diagonal of the matrix clearly shows the binary nature of `mpg01`, with distinct distributions for each category.

The lower triangle of the matrix, which compares each feature against `mpg01`, suggests a strong inverse relationship between vehicle weight and fuel efficiency, as the two color groups are clearly separated. Similarly, a significant negative correlation is observed with features such as `cylinders`, `displacement`, and `horsepower`, indicating that these factors are likely to decrease the mpg of a vehicle.

The upper triangle provides correlation coefficients, with a negative value such as -0.753 for `cylinders` suggesting a higher number of cylinders is generally associated with lower mpg.

In essence, the matrix suggests that `weight`, `displacement`, `horsepower`, and `cylinders` are key features negatively associated with high mpg, making them important factors for predicting fuel efficiency. Conversely, `year` and `origin` show little to no correlation, implying they might be less useful as predictors in this context.

Split the data into a training set and a test set

I used the Caret library in R to do this, I set seed to 123 to produce reproducibility of training and testing data split over different executions. And then I split the data with a 75:25 split between the training and testing data. I created a variable most_associated_vars <- c("weight", "displacement", "horsepower", "cylinders") and these variables are going to be used as predictors for the below models.

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?

```
$lda_accuracy
[1] 0.9081633

$lda_confusion
          Actual
Predicted  0  1
        0 41  1
        1  8 48
```

The LDA model achieved an accuracy of approximately 90.82%. The confusion matrix indicated that the model correctly predicted 41 instances of the mpg01 category '0' and 48 instances of category '1'. There were 8 instances where category '1' was incorrectly predicted as '0', and 1 instance where category '0' was incorrectly predicted as '1'.

```
$lda_test_error
[1] 0.09183673
```

The test error for the LDA model was approximately 9.18%.

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?

```
$qda_accuracy
[1] 0.9081633

$qda_confusion
          Actual
Predicted  0  1
        0 41  1
        1  8 48

$qda_test_error
[1] 0.09183673
```

I applied Quadratic Discriminant Analysis (QDA) to the same training dataset and features. The test error for the QDA model was also about 9.18%, identical to that of the LDA model.The QDA model resulted in an accuracy identical to that of the LDA, at around 90.82%. The confusion matrix was the same as LDA, with 41 correct predictions for category '0', 48 correct for '1', and the same number of incorrect predictions for both categories.

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?

```
$logit_accuracy
[1] 0.9081633

$logit_confusion
          Actual
Predicted  0  1
        0 42  2
        1  7 47

$logit_test_error
[1] 0.09183673
```

This model also posted an accuracy of 90.82%. The confusion matrix showed a slightly different distribution of predictions compared to LDA and QDA, with 42 correct predictions for category '0' and 47 for '1'. The model incorrectly predicted 7 instances of category '1' as '0', and 2 instances of category '0' as '1'. The logistic regression model yielded a test error of roughly 9.18%, which was consistent with the LDA and QDA models.

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?

```
$knn_confusion
          Actual
Predicted  0  1
        0 43  1
        1  6 48
```

```
$best_k
[1] 9

$best_k_error
[1] 0.07142857
```

I utilized the K-Nearest Neighbors (KNN) algorithm for the prediction task, experimenting with various values of K (1, 3, 5, 7, 9, 11, 13, 15). After evaluating multiple models, the KNN with K=9 provided the best performance, with a test error of about 7.14%, which was lower than the test errors from LDA, QDA, and logistic regression models. The KNN model with K=9 outperformed the other models, achieving an accuracy of approximately 92.86%. The confusion matrix revealed that the model correctly predicted 43 instances of category '0' and 48 of category '1', with fewer misclassifications: 6 instances of category '1' were predicted as '0', and only 1 instance of category '0' was predicted as '1'. This suggests that for this particular dataset, KNN with K=9 is the most effective model for predicting mpg01 based on the features selected.

Conclusion

After conducting a thorough analysis using various classification methods to predict fuel efficiency (mpg01), we observed the following:

LDA, QDA, and Logistic Regression: These models showed equivalent performance, with an accuracy rate of approximately 90.82%. The consistency across these methods suggests that the variables selected have a linear relationship with the mpg01 outcome.
KNN: The KNN classifier with K=9 outperformed the other models, achieving an accuracy of 92.86%. This indicates that a non-linear approach that captures local similarities offers a slight improvement in predictive power for this dataset.
The test errors and confusion matrices reinforced these findings, showing that while all models are robust, the KNN model with K=9 is slightly more reliable for our dataset. The insights gained from this analysis will be instrumental in developing predictive models for fuel efficiency, leveraging the identified key features such as weight, displacement, horsepower, and cylinders.