Assignment 3

Lorenzo Ausiello

2023-11-14

Problem 1

After downloading data from “Weekly” file, some statistics has been performed. Out of 1089 observations, 484 show a Down Direction of prices on Today, and 605 show a Up Direction. Box-plots and scatterplots, together with numerical summaries, demonstrate that both Down and Up today Directions of prices are preceded by returns on average close to zero. Therefore, for sure it will be difficult to try to explain the Direction feature in terms of lag returns (returns of previous days). As for volume, the same consideration applies: very little differences of the means in the two different levels. Below the numerical and graphical statistics/summaries.

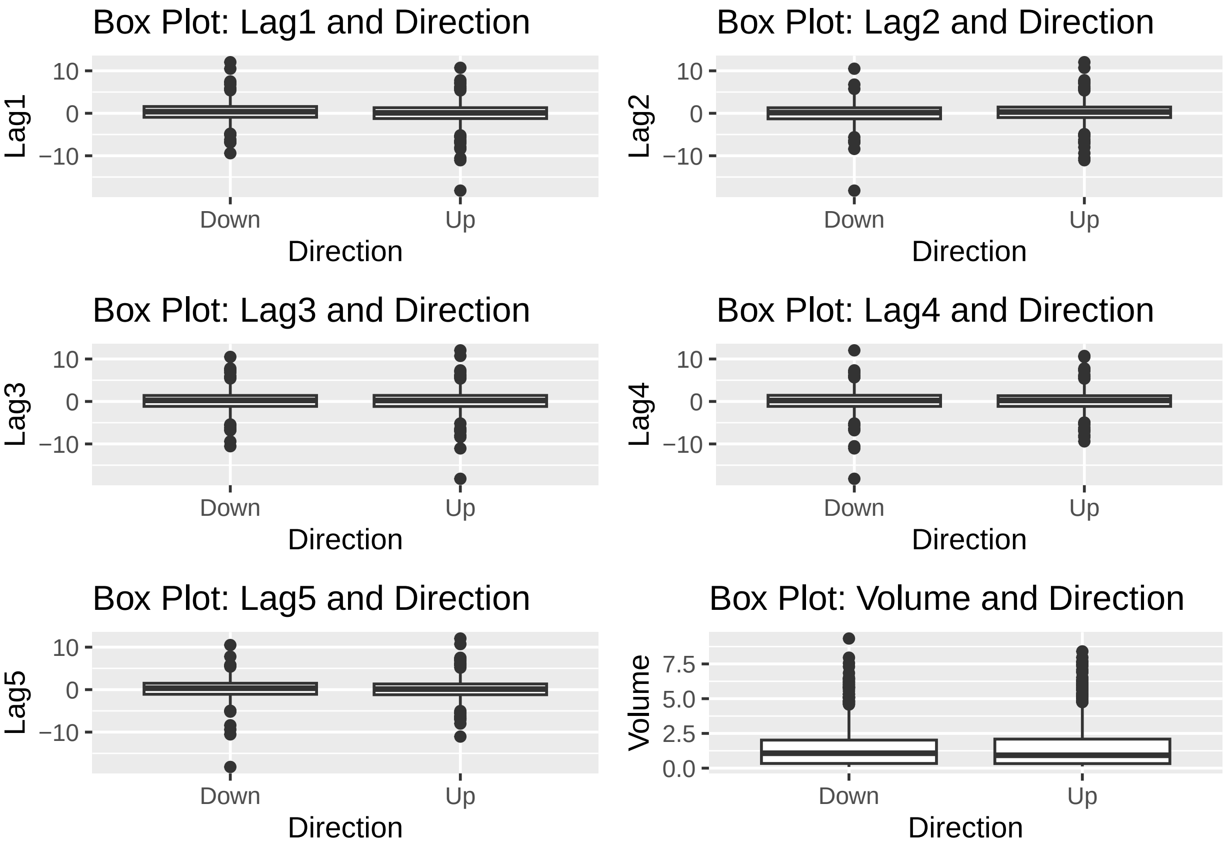
Table 1: Dataset

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1990	0.816	1.572	-3.936	-0.229	-3.484	0.1549760	-0.270	Down
1990	-0.270	0.816	1.572	-3.936	-0.229	0.1485740	-2.576	Down
1990	-2.576	-0.270	0.816	1.572	-3.936	0.1598375	3.514	Up
1990	3.514	-2.576	-0.270	0.816	1.572	0.1616300	0.712	Up
1990	0.712	3.514	-2.576	-0.270	0.816	0.1537280	1.178	Up
1990	1.178	0.712	3.514	-2.576	-0.270	0.1544440	-1.372	Down

Table 2: Summary

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
Min. :1990	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :0.08747	Min. :-18.1950	Down:484
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 Qu.: -1.1540	Up :605
Median :2000	Median : 0.2410	Median : 0.2410	Median : 0.2410	Median : 0.2380	Median : 0.2340	Median :1.00268	Median : 0.2410	NA
Mean :2000	Mean : 0.1506	Mean : 0.1511	Mean : 0.1472	Mean : 0.1458	Mean : 0.1399	Mean :1.57462	Mean : 0.1499	NA
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 Qu.: 1.4050	NA
Max. :2010	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. :9.32821	Max. : 12.0260	NA

```
##      Up
## Down  0
## Up    1
```


Box Plot: Lag1 and Direction
Lag1
10
0
−10
Down
Up
Direction
Box Plot: Lag2 and Direction
Lag2
10
0
−10
Down
Up
Direction
Box Plot: Lag3 and Direction
Lag3
10
0
−10
Down
Up
Direction
Box Plot: Lag4 and Direction
Lag4
10
0
−10
Down
Up
Direction
Box Plot: Lag5 and Direction
Lag5
10
0
−10
Down
Up
Direction
Box Plot: Volume and Direction
Volume
7.5
5.0
2.5
0.0
Down
Up
Direction

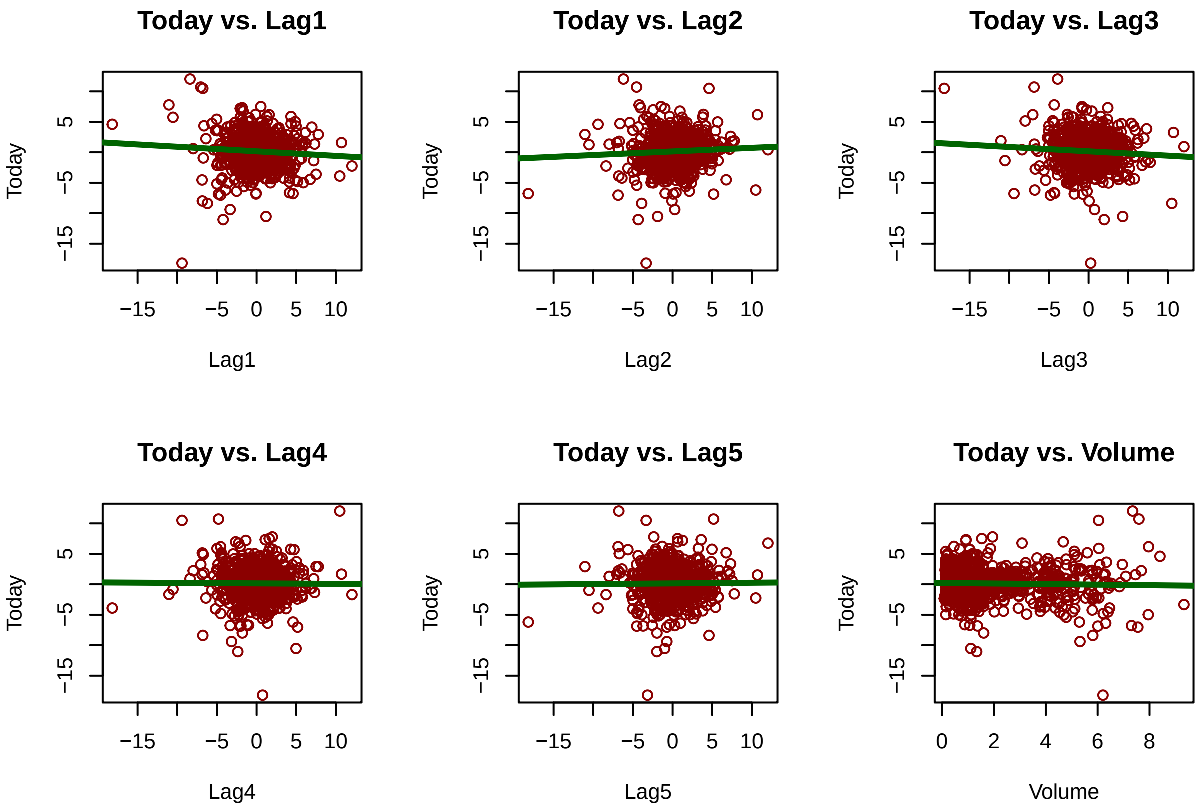


Table 3: Summary By: Lag1

Direction	Mean	Length	Min	Max
Down	0.2822955	484	-9.399	12.026
Up	0.0452165	605	-18.195	10.707

Table 4: Summary By: Lag2

Direction	Mean	Length	Min	Max
Down	-0.0404236	484	-18.195	10.491
Up	0.3042810	605	-11.050	12.026

Table 5: Summary By: Lag3

Direction	Mean	Length	Min	Max
Down	0.2076467	484	-10.538	10.491
Up	0.0988512	605	-18.195	12.026

Table 6: Summary By: Lag4

Direction	Mean	Length	Min	Max
Down	0.2000207	484	-18.195	12.026
Up	0.1024562	605	-9.399	10.707

Table 7: Summary By: Lag5

Direction	Mean	Length	Min	Max
Down	0.1878347	484	-18.195	10.491
Up	0.1015388	605	-11.050	12.026

Table 8: Summary By: Volume

Direction	Mean	Length	Min	Max
Down	1.608536	484	0.087465	9.328214
Up	1.547483	605	0.125075	8.403358

In this logistic regression analysis below, the entire dataset was utilized to model the relationship between the response variable “Direction” and the predictors, encompassing five lag variables (Lag 1, Lag 2, Lag 3, Lag 4, and Lag 5) along with “Volume.” As result, only Lag 2 appear to be statistically significant. It means that returns of the second day of the week show relationship with the likelihood of a specific direction. Particularly, the logistic regression performed result in a coefficient for Lag 2 positive and equal to 0.05844. The higher the Lag 2, the higher the probability to have an Up Direction. However, if we check the model on the same data used to build the regression, we notice that overall error rate is equal to 43,89%! This is mainly due to the fact that 88.84% of true down have not been correctly identified. Indeed, only 7.93% of true up have not been correctly predicted.

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

```
## In attesa che venga eseguita la profilazione...
```

Table 9: Logistic regression: confidence interval

	2.5 %	97.5 %
(Intercept)	0.0988087	0.4358010
Lag1	-0.0934771	0.0102927
Lag2	0.0061976	0.1116977
Lag3	-0.0686539	0.0360431
Lag4	-0.0799524	0.0240160
Lag5	-0.0664951	0.0371199
Volume	-0.0950519	0.0497934

Table 10: Direction prediction included in original dataset

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
1990	0.816	1.572	-3.936	-0.229	-3.484	0.1549760	-0.270	Down	Up
1990	-0.270	0.816	1.572	-3.936	-0.229	0.1485740	-2.576	Down	Up
1990	-2.576	-0.270	0.816	1.572	-3.936	0.1598375	3.514	Up	Up
1990	3.514	-2.576	-0.270	0.816	1.572	0.1616300	0.712	Up	Down
1990	0.712	3.514	-2.576	-0.270	0.816	0.1537280	1.178	Up	Up
1990	1.178	0.712	3.514	-2.576	-0.270	0.1544440	-1.372	Down	Up

Table 11: Logistic: Confusion matrix

	Down	Up
Down	54	48
Up	430	557

Table 12: % of predictions that are correct

x
0.5610652

Table 13: Overall error rate: logistic

x
0.4389348

Table 14: % of true down not identified

x
0.8884298

Table 15: % of true up not identified

x
0.0793388

Then, the logistic regression analysis was conducted using a training data period spanning from 1990 to 2008, with Lag2 as the sole predictor. Subsequently, the model was employed to predict the direction of the market for the held-out data from 2009 and 2010. Following this, three additional classification methods were employed and evaluated using the same training and testing data: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors (KNN) with K = 1, 2,3,4,5,6,10 and 20. Each method generated its own confusion matrix and overall fraction of correct predictions.

Logistic Regression and Linear Discriminant Analysis (LDA) exhibit identical error rates of 0.375, indicating comparable accuracy in predicting market direction. Quadratic Discriminant Analysis (QDA) shows a slightly higher error rate at 0.4134615. Moving beyond, Table 45 introduces additional K-Nearest Neighbors (KNN) models with different values of K. Notably, KNN1 and KNN2 display higher error rates of 0.5, while subsequent KNN models (KNN3 to KNN20) exhibit varying error rates between 0.4230769 and 0.4615385. This is due to the fact that the lower the K the higher the overfitting issue (for k=1, perfect overfitting on training data). The choice of the most effective method depends on specific analysis goals and considerations, with lower error rates generally indicating better predictive performance. Further evaluation, considering other metrics and study objectives, is essential for a comprehensive assessment.

Remember: Logistic regression, LDA, QDA, and KNN each have their strengths and assumptions. Logistic regression, for instance, is robust when linear relationships are present, while LDA and QDA assume different covariance structures. KNN, on the other hand, relies on proximity in feature space.

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

```
## In attesa che venga eseguita la profilazione...
```

Table 16: Logistic regression (training data): confidence interval

	2.5 %	97.5 %
(Intercept)	0.0774793	0.3295391
Lag2	0.0023008	0.1150942

Table 17: Direction prediction included in original test dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

Table 18: Confusion matrix: test data (logistic)

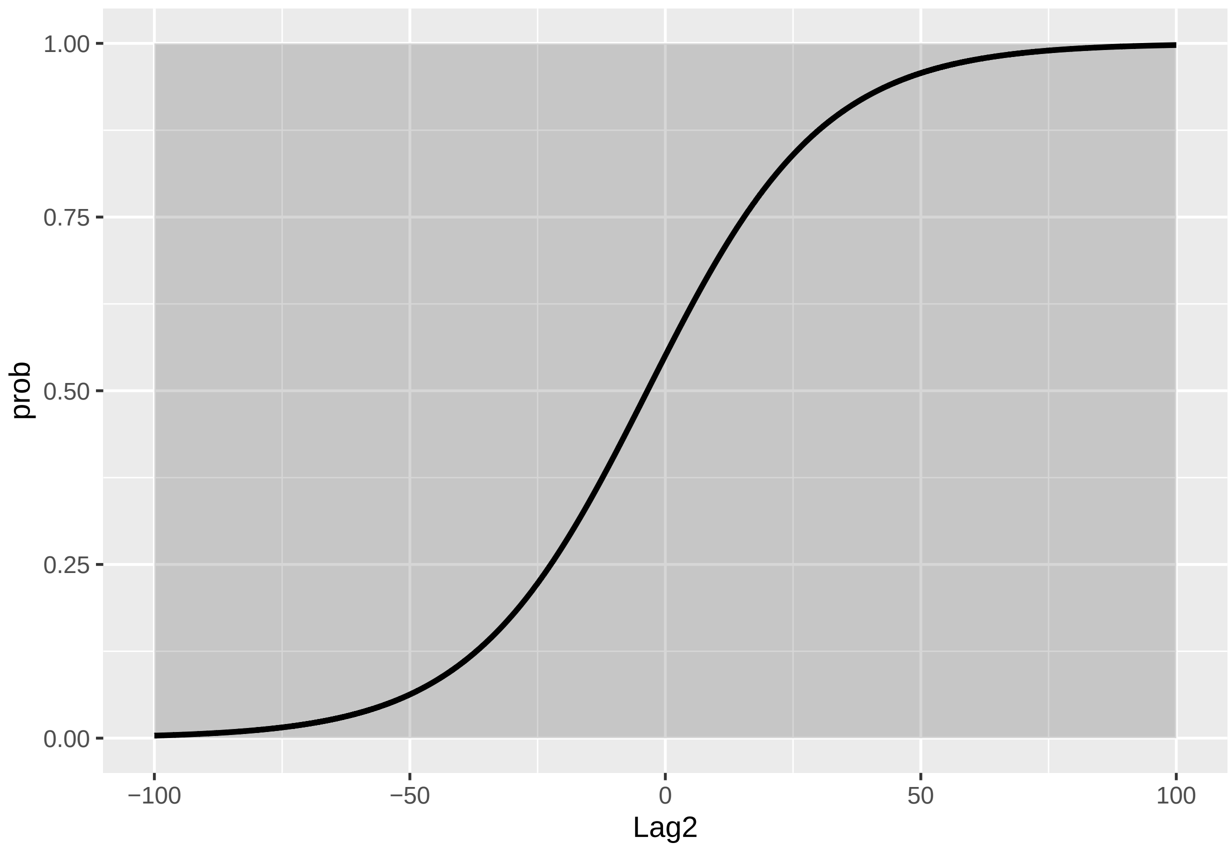
	Down	Up
Down	9	5
Up	34	56

Table 19: % of predictions that are correct

x
0.625

Table 20: Overall error rate on test data (logistic)

x
0.375



```
## Call:
## lda(Direction ~ Lag2, data = training)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4477157 0.5522843
##
## Group means:
##             Lag2
## Down -0.03568254
## Up    0.26036581
##
## Coefficients of linear discriminants:
##            LD1
## Lag2 0.4414162
```

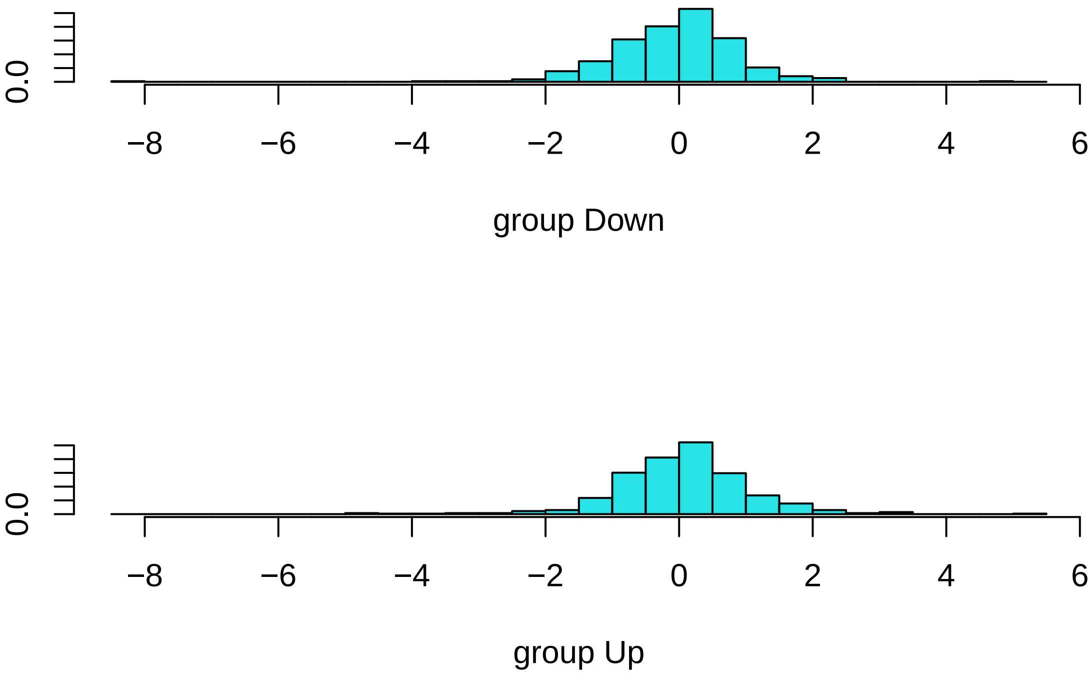


Table 21: LDA: Direction prediction included in original test dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

Table 22: LDA: Confusion Matrix

	Down	Up
Down	9	5
Up	34	56

Table 23: Overall Error rate: LDA

x
0.375

```
## Call:
## qda(Direction ~ Lag2, data = training)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4477157 0.5522843
##
## Group means:
##             Lag2
## Down -0.03568254
## Up    0.26036581
```

Table 24: QDA: Direction prediction included in original test dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Up
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Up
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

Table 25: QDA: Confusion Matrix

	Down	Up
Down	0	0
Up	43	61

Table 26: QDA: Overall error rate

x
0.4134615

Table 27: KNN K=1: Direction prediction included in original test dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Up
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Down
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Up

Table 28: KNN K=1: Confusion matrix

	Down	Up
Down	21	29
Up	22	32

Table 29: KNN K=1: Overall error rate

x
0.4903846

Table 30: KNN K=2: Confusion matrix

	Down	Up
Down	22	27
Up	21	34

Table 31: KNN K=2: Overall error rate

x
0.4615385

Table 32: KNN K=3: Confusion matrix

	Down	Up
Down	16	20
Up	27	41

Table 33: KNN K=3: Overall error rate

x
0.4519231

Table 34: KNN K=4: Confusion matrix

	Down	Up
Down	20	21
Up	23	40

Table 35: KNN K=4: Overall error rate

x
0.4230769

Table 36: KNN K=5: Confusion matrix

	Down	Up
Down	16	22
Up	27	39

Table 37: KNN K=5: Overall error rate

x
0.4711538

Table 38: KNN K=6: Confusion matrix

	Down	Up
Down	16	20
Up	27	41

Table 39: KNN K=6: Overall error rate

x
0.4519231

Table 40: KNN K=10: Confusion matrix

	Down	Up
Down	17	19
Up	26	42

Table 41: KNN K=10: Overall error rate

x
0.4326923

Table 42: KNN K=20: Confusion matrix

	Down	Up
Down	20	20

	Down	Up
Up	23	41

Table 43: KNN K=20: Overall error rate

x
0.4134615

Table 44: Compare Overall error rate for different predictions methods: pt.1

	Logistic	LDA	QDA	KNN1	KNN2
Overall error rate	0.375	0.375	0.4134615	0.4903846	0.4615385

Table 45: Compare Overall error rate for different predictions methods: pt.2

	KNN3	KNN4	KNN5	KNN6	KNN10	KNN20
Overall error rate	0.4519231	0.4230769	0.4711538	0.4519231	0.4326923	0.4134615

Below are performed more classification methods, such as multiple logistic regression and multiple LDA. Experiment with different combinations of predictors are so performed.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2, family = binomial, data = training)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.21109    0.06456   3.269  0.00108 **
## Lag1        -0.05421    0.02886  -1.878  0.06034 .
## Lag2         0.05384    0.02905   1.854  0.06379 .
## ---
## 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: 1347.0  on 982  degrees of freedom
## AIC: 1353
##
## Number of Fisher Scoring iterations: 4
```

```
## In attesa che venga eseguita la profilazione...
```

Table 46: Multiple logistic regression: confidence interval

	2.5 %	97.5 %
(Intercept)	0.0847865	0.3379648
Lag1	-0.1115531	0.0018931
Lag2	-0.0026436	0.1114614

Table 47: multiple logistic: Direction prediction included in original test dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Down
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Up
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Up
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Down

Table 48: multiple logistic: confusion matrix

	Down	Up
Down	7	8
Up	36	53

Table 49: multiple logistic: overall error rate

x
0.4230769

```
## Call:
## lda(Direction ~ Lag1 + Lag2, data = training)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4477157 0.5522843
##
## Group means:
##              Lag1        Lag2
## Down  0.289444444 -0.03568254
## Up   -0.009213235  0.26036581
##
## Coefficients of linear discriminants:
##             LD1
## Lag1 -0.3013148
## Lag2  0.2982579
```

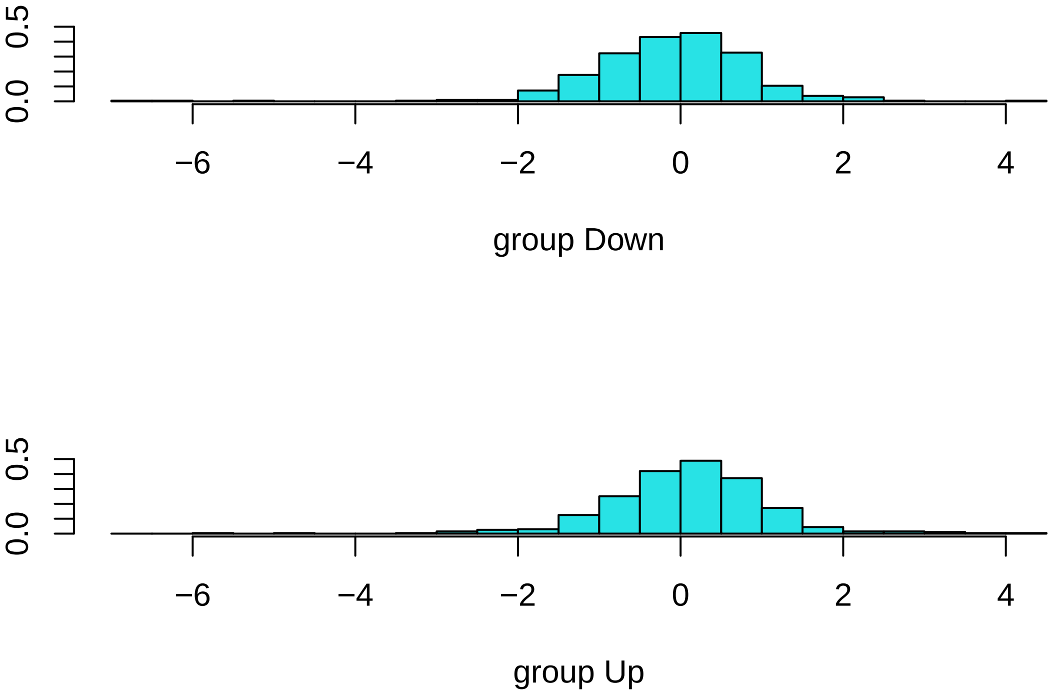


Table 50: multiple LDA: Direction prediction included in original test dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction	Direction prediction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down	Down
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down	Up
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down	Up
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down	Up
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down	Down

Table 51: multiple LDA: confusion matrix

	Down	Up
Down	7	8
Up	36	53

Table 52: multiple LDA: overall error rate

x
0.4230769

Problem 2

After downloading data from “Auto” file, and after creating a binary variable mpg01, the data has been explored graphically. Box-plots seem to show high differences between average weight, horsepower, displacement and cylinders in the two different level of mpg01 (miles per gas above the median (1) and below (0)). Scatterplots, instead, show a strong relationship between miles per gas and these quantitative variables. In cases like this, variables may be useful to explain the categorical variable. Therefore, they have been used to perform LDA, QDA, logistic regression and KNN.

The dataset has been divided into a training set and a test set for predictive modeling. LDA and QDA were employed to predict mpg01 on the training data, and the respective test errors were calculated. Logistic regression was also applied to predict mpg01 using the identified variables, and its test error was determined.

Additionally, K-Nearest Neighbors (KNN) was implemented on the training data with varying values of K (1,2,3,4,5,6,10 and 20). Test errors were computed for each K, and the performance of different K values was compared to identify the most effective.

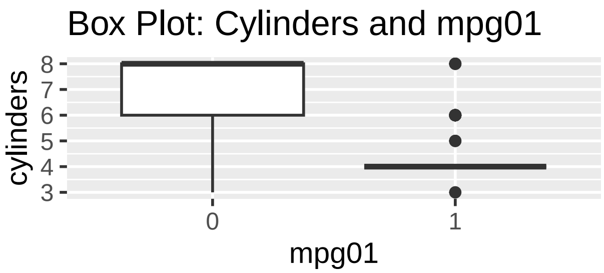
Below the results.

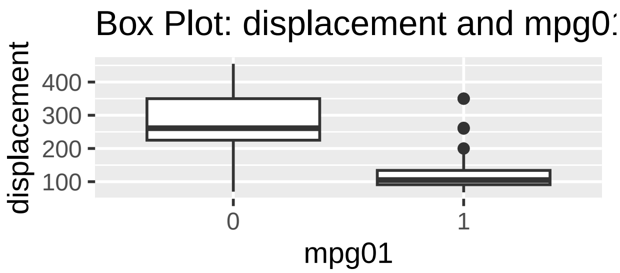
Table 53: Dataset

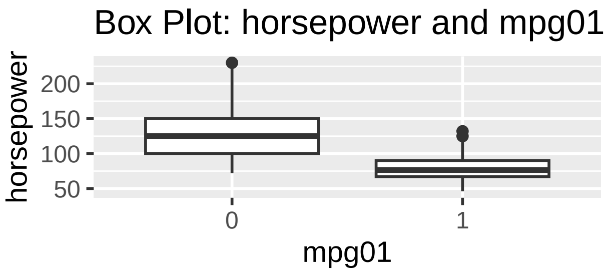
mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11.0	70	1	plymouth satellite
16	8	304	150	3433	12.0	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10.0	70	1	ford galaxie 500

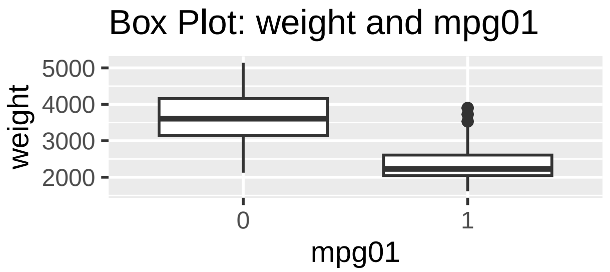
Table 54: Summary

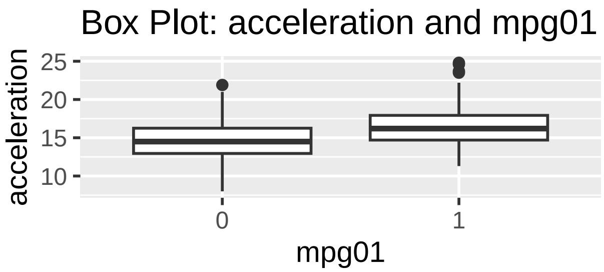
	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name	mpg01
	Min. : 9.00	Min. :3.000	Min. : 68.0	Min. : 46.0	Min. :1613	Min. : 8.00	Min. :70.00	Min. :1.000	Length:392	0:196
	1st Qu.:17.00	1st Qu.:4.000	1st Qu.:105.0	1st Qu.: 75.0	1st Qu.:2225	1st Qu.:13.78	1st Qu.:73.00	1st Qu.:1.000	Class :character	1:196
	Median :22.75	Median :4.000	Median :151.0	Median : 93.5	Median :2804	Median :15.50	Median :76.00	Median :1.000	Mode :character	NA
	Mean :23.45	Mean :5.472	Mean :194.4	Mean :104.5	Mean :2978	Mean :15.54	Mean :75.98	Mean :1.577	NA	NA
	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	3rd Qu.:2.000	NA	NA
	Max. :46.60	Max. :8.000	Max. :455.0	Max. :230.0	Max. :5140	Max. :24.80	Max. :82.00	Max. :3.000	NA	NA

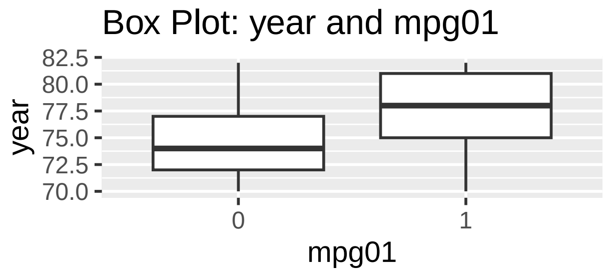

Box Plot: Cylinders and mpg01
cylinders
8
7
6
5
4
3
0
1
mpg01

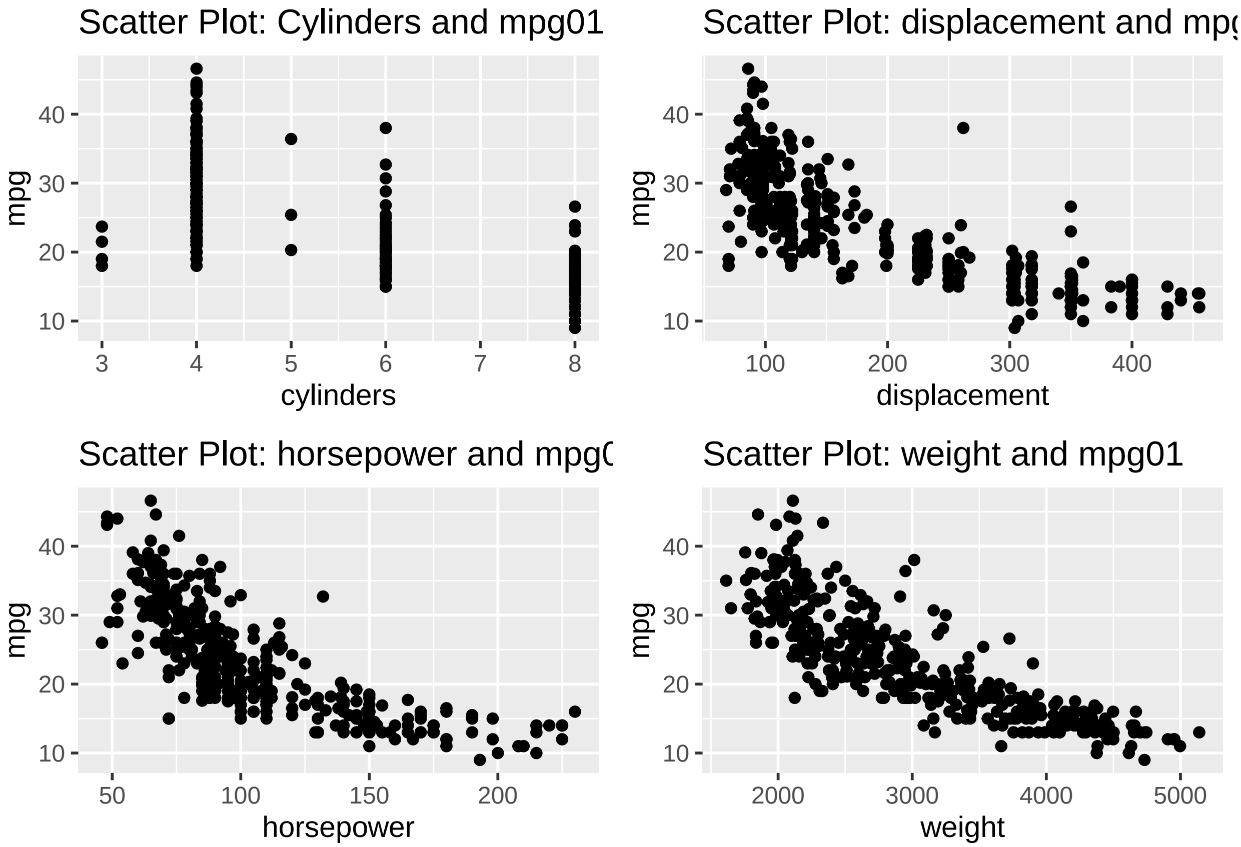

Box Plot: displacement and mpg0:
displacement
400
300
200
100
0
1
mpg01


Box Plot: horsepower and mpg01
horsepower
200
150
100
50
0
1
mpg01


Box Plot: weight and mpg01
weight
5000
4000
3000
2000
0
1
mpg01


Box Plot: acceleration and mpg01
acceleration
25
20
15
10
0
1
mpg01


Box Plot: year and mpg01
year
82.5
80.0
77.5
75.0
72.5
70.0
0
1
mpg01



Scatter Plot: Cylinders and mpg01

Scatter Plot: displacement and mp

Scatter Plot: horsepower and mpg

Scatter Plot: weight and mpg01

```
##
## Call:
## glm(formula = mpg01 ~ cylinders + weight + horsepower + displacement,
##     family = binomial, data = training)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  14.102661   2.853186   4.943  7.7e-07 ***
## cylinders    -0.041541   0.648295  -0.064   0.9489
## weight       -0.002673   0.001173  -2.279   0.0227 *
## horsepower   -0.048054   0.024353  -1.973   0.0485 *
## displacement -0.018039   0.015442  -1.168   0.2427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 326.709  on 249  degrees of freedom
## Residual deviance:  94.388  on 245  degrees of freedom
## AIC: 104.39
##
## Number of Fisher Scoring iterations: 8
```

```
## In attesa che venga eseguita la profilazione...
```

Table 55: Logistic regression: confidence interval

	2.5 %	97.5 %
(Intercept)	8.9900771	20.2886469
cylinders	-1.4295223	1.1781284
weight	-0.0051400	-0.0004715
horsepower	-0.0982084	-0.0017258
displacement	-0.0488756	0.0128038

Table 56: mpg01 prediction included in original test dataset

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name	mpg01	mpg01 prediction
251	19.2	6	231	105	3535	19.2	78	1	pontiac phoenix lj	0	0
252	20.5	6	200	95	3155	18.2	78	1	chevrolet malibu	0	0
253	20.2	6	200	85	2965	15.8	78	1	ford fairmont (auto)	0	0
254	25.1	4	140	88	2720	15.4	78	1	ford fairmont (man)	1	0
255	20.5	6	225	100	3430	17.2	78	1	plymouth volare	0	0
256	19.4	6	232	90	3210	17.2	78	1	amc concord	0	0

Table 57: Confusion matrix: test data (logistic)

	0	1
0	35	33
1	1	73

Table 58: Overall error rate on test data (logistic)

x
0.2394366

```
## Call:
## lda(mpg01 ~ cylinders + weight + horsepower + displacement, data = training)
##
## Prior probabilities of groups:
##    0    1
## 0.64 0.36
##
## Group means:
##   cylinders   weight horsepower displacement
## 0  6.843750 3673.525  133.56250     280.7750
## 1  4.044444 2203.733   77.08889     104.5611
##
```

```
## Coefficients of linear discriminants:
##                       LD1
## cylinders    -0.335566137
## weight       -0.001122609
## horsepower    0.011759201
## displacement -0.003896812
```

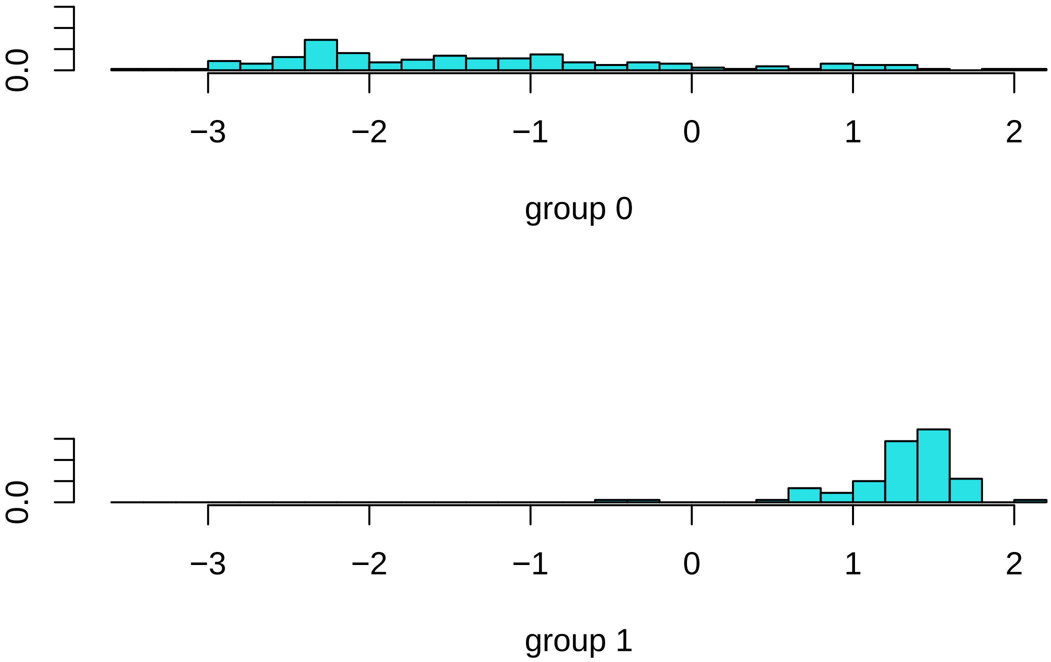


Table 59: LDA: mpg01 prediction included in original test dataset

											mpg01
	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name	mpg01	prediction
251	19.2	6	231	105	3535	19.2	78	1	pontiac phoenix lj	0	0
252	20.5	6	200	95	3155	18.2	78	1	chevrolet malibu	0	0
253	20.2	6	200	85	2965	15.8	78	1	ford fairmont (auto)	0	0
254	25.1	4	140	88	2720	15.4	78	1	ford fairmont (man)	1	1
255	20.5	6	225	100	3430	17.2	78	1	plymouth volare	0	0
256	19.4	6	232	90	3210	17.2	78	1	amc concord	0	0

Table 60: LDA: Confusion Matrix

	0	1
0	32	14
1	4	92

Table 61: Overall Error rate: LDA

x
0.1267606

```
## Call:
## qda(mpg01 ~ cylinders + weight + horsepower + displacement, data = training)
##
## Prior probabilities of groups:
##    0    1
## 0.64 0.36
##
## Group means:
##   cylinders   weight horsepower displacement
## 0  6.843750 3673.525  133.56250     280.7750
## 1  4.044444 2203.733   77.08889     104.5611
```

Table 62: QDA: mpg01 prediction included in original test dataset

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name	mpg01	mpg01 prediction
251	19.2	6	231	105	3535	19.2	78	1	pontiac phoenix lj	0	0
252	20.5	6	200	95	3155	18.2	78	1	chevrolet malibu	0	0
253	20.2	6	200	85	2965	15.8	78	1	ford fairmont (auto)	0	0
254	25.1	4	140	88	2720	15.4	78	1	ford fairmont (man)	1	1
255	20.5	6	225	100	3430	17.2	78	1	plymouth volare	0	0
256	19.4	6	232	90	3210	17.2	78	1	amc concord	0	0

Table 63: QDA: Confusion Matrix

	0	1
0	33	20
1	3	86

Table 64: Overall Error rate: QDA

x
0.1619718

Table 65: KNN K=1: mpg01 prediction included in original test dataset

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name	mpg01	mpg01 prediction
251	19.2	6	231	105	3535	19.2	78	1	pontiac phoenix lj	0	0
252	20.5	6	200	95	3155	18.2	78	1	chevrolet malibu	0	0
253	20.2	6	200	85	2965	15.8	78	1	ford fairmont (auto)	0	0
254	25.1	4	140	88	2720	15.4	78	1	ford fairmont (man)	1	1
255	20.5	6	225	100	3430	17.2	78	1	plymouth volare	0	0
256	19.4	6	232	90	3210	17.2	78	1	amc concord	0	0

Table 66: KNN K=1: Confusion matrix

	0	1
0	34	27
1	2	79

Table 67: KNN K=2: Confusion matrix

	0	1
0	34	26
1	2	80

Table 68: KNN K=3: Confusion matrix

	0	1
0	35	28
1	1	78

Table 69: KNN K=4: Confusion matrix

	0	1
0	35	26
1	1	80

Table 70: KNN K=5: Confusion matrix

	0	1
0	35	24
1	1	82

Table 71: KNN K=6: Confusion matrix

	0	1
0	35	25
1	1	81

Table 72: KNN K=10: Confusion matrix

	0	1
0	35	28
1	1	78

Table 73: KNN K=20: Confusion matrix

	0	1
0	35	26
1	1	80

Table 74: KNN: Compare Overall error rate for different value of K

	KNN1	KNN2	KNN3	KNN4	KNN5	KNN6	KNN10	KNN20
Overall error rate	0.2042254	0.1971831	0.2042254	0.1901408	0.1760563	0.1830986	0.2042254	0.1901408

The logistic regression shows that there is a negative relationship between all of these variables and the probability of mpg above the median (mpg01=1). The more the cylinders, the lower the probability of mpg01 qual to 1 (lower miles per gas). The more the horsepower, the lower the probability of mpg01 qual to 1 (lower miles per gas). The higher the weight, the lower the probability of mpg01 qual to 1 (lower miles per gas). The higher the displacement, the lower the probability of mpg01 qual to 1 (lower miles per gas). However, cylinders and displacement are not statistically significant.

The examination of classification methods on the test data unveils varying overall error rates. Logistic regression and Linear Discriminant Analysis (LDA) achieved overall error rates of approximately 23,94% and 12.68%, respectively, demonstrating LDA's superior accuracy. Quadratic Discriminant Analysis (QDA) exhibited a marginally higher overall error rate of around 16.20%. Notably, K-Nearest Neighbors (KNN) displayed a consistent overall error rate of about 20% from KNN1 to KNN3, then experienced a decrease to 1760% for KNN5 (lowest among KNN). This change may be attributed to the optimal balance between bias and variance in the model. The lower error rates of LDA compared to QDA may be attributed to the assumption of equal covariance matrices in QDA, making it more sensitive to variations in the data. LDA, with fewer assumptions, thus demonstrated superior performance in this particular analysis.