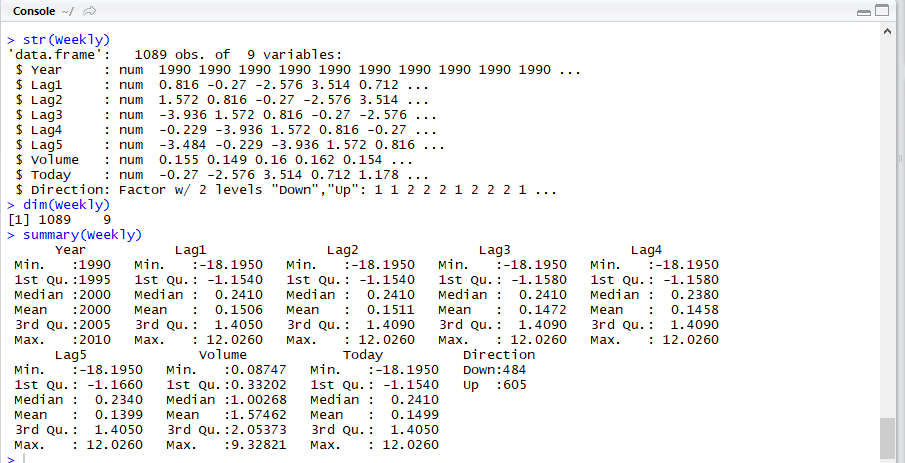
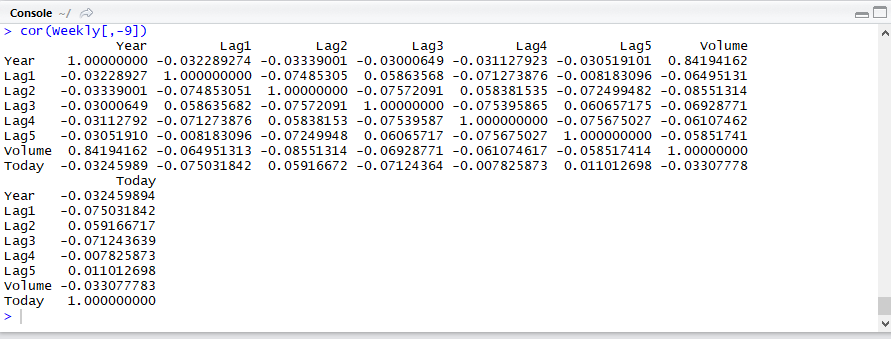
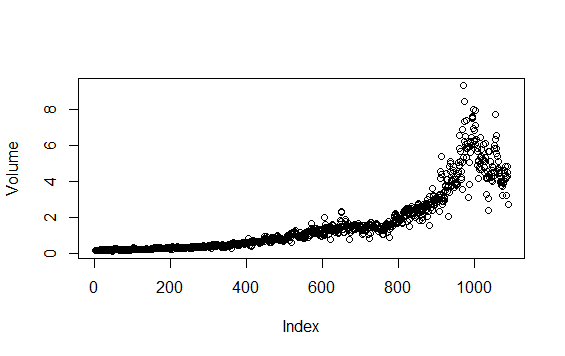
1 a)

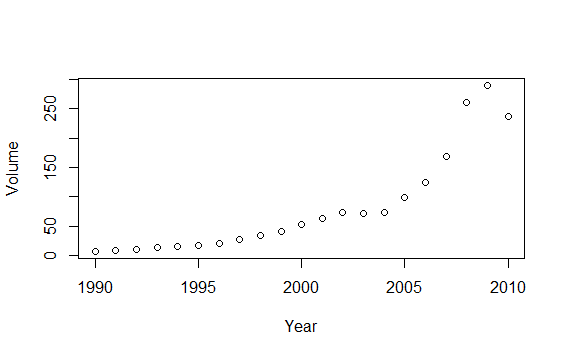




The correlation matrix suggests that there is no significant relationship/correlation between Lag1,2,3,4,5 variables, but there is a 0.842 correlation between ‘Year’ and ‘Volume’ variable. Hence plot between year and volume is plotted.

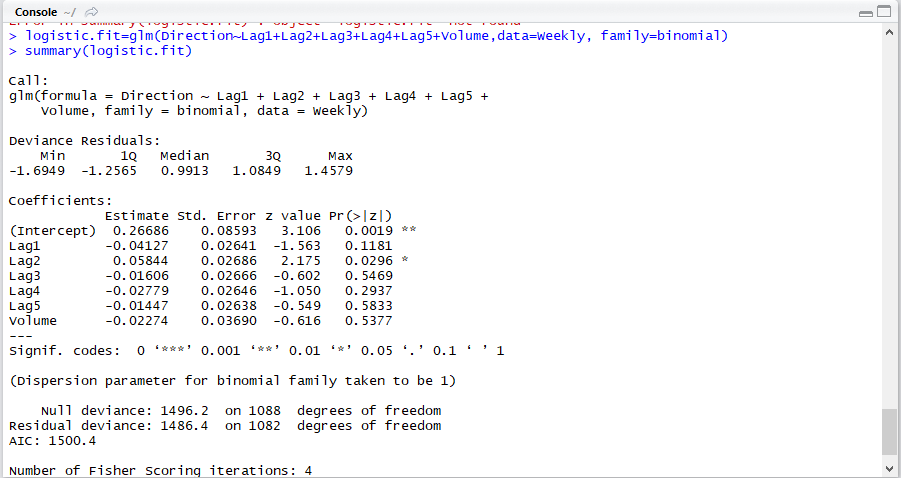
.

PLOT(VOLUME)



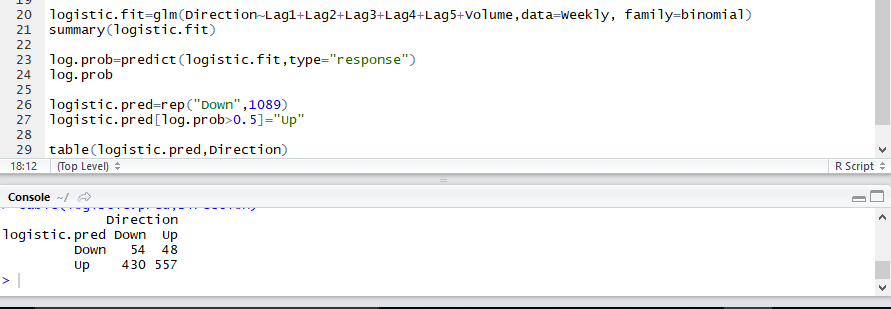
YEAR WISE MEAN OF VOLUME

1 b)



Predictor Lag 2 seems to be significant with a p value less than 0.05.

1c)



Accuracy= (True Positive+True Negative)/(Negative+Positive) = (557+54)/(430+557+48+54) =0.561=56.1%

Error Rate= 1- Accuracy = 0.4389=43.89%

Sensitivity =True Positive / Positive = 557/(557+48)= 0.92=92%

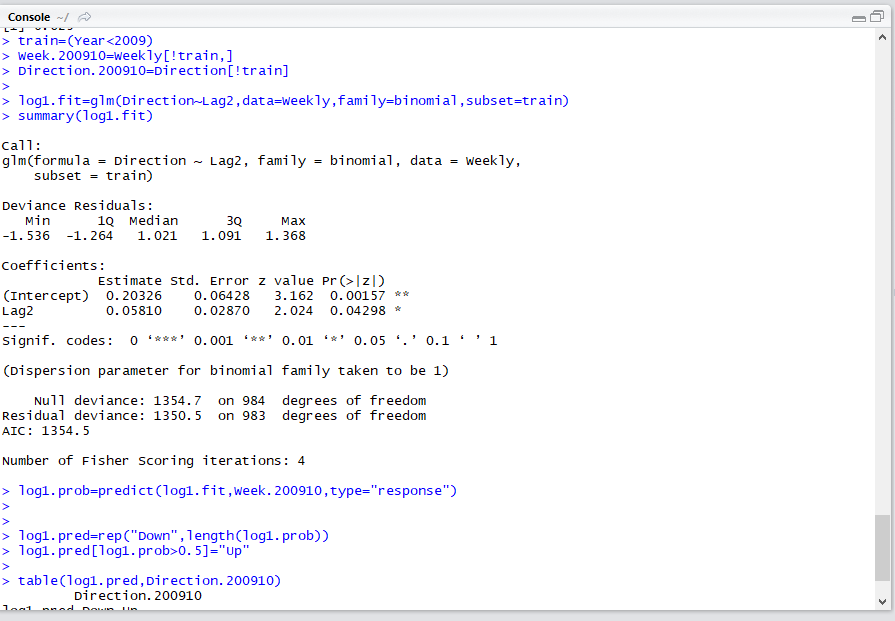
Specificity=True Negative/Negative=54/(54+430)=0.1115=11.15%

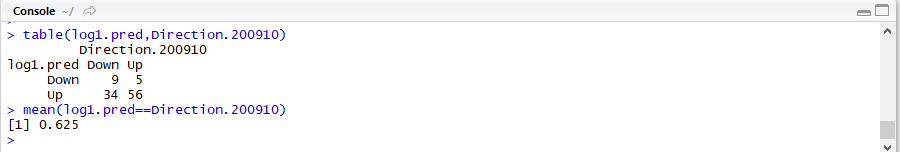
False Positive Rate= 1-Specificity= 88.84% {The confusion matrix suggests that the model predicts the Direction to be Up 88.84% when it actually is Down according to the given data.]

The model accurately predicts 92% of the times when the market goes Up and only 11.15% when the market goes Down. The overall accuracy though is 56.1%

The error rate does not represent the performance of logistic regression in prediction as the error rate calculated here is the training error rate as the data being compared for accuracy , error and other metrics is between the model predicted using the training data using which the logistic model was made and not the test data.

1)d)





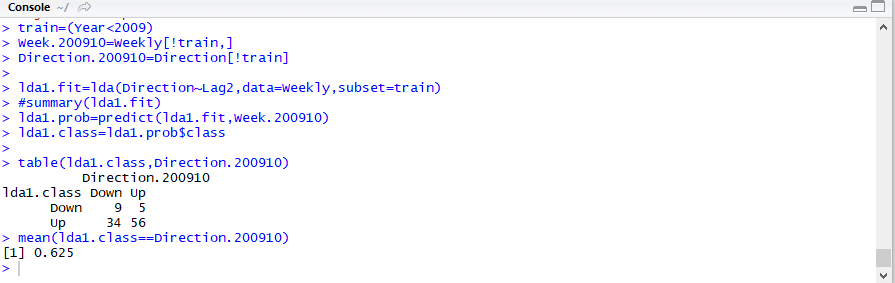
Accuracy= (True Positive+True Negative)/(Negative+Positive) = (56+9)/(56+9+5+34) =0.625=62.5%

Error Rate= 1- Accuracy = 0.375=37.5%

Sensitivity =True Positive / Positive = 56/(56+5)= 0.918=91.8%

Specificity=True Negative/Negative=9/(9+34)=0.209=20.93%

1)e) LDA



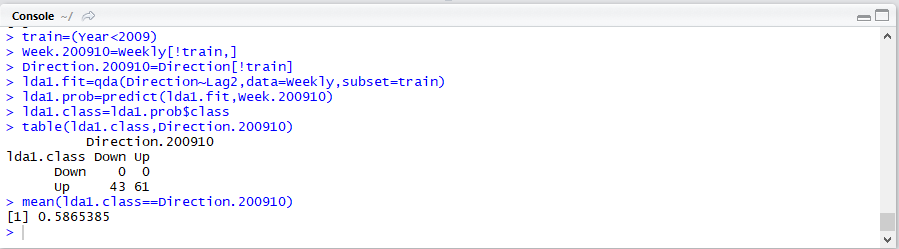
Accuracy= (True Positive+True Negative)/(Negative+Positive) = (56+9)/(56+9+5+34) =0.625=62.5%

Error Rate= 1- Accuracy = 0.375=37.5%

Sensitivity =True Positive / Positive = 56/(56+5)= 0.918=91.8%

Specificity=True Negative/Negative=9/(9+34)=0.209=20.93%

1)f) QDA



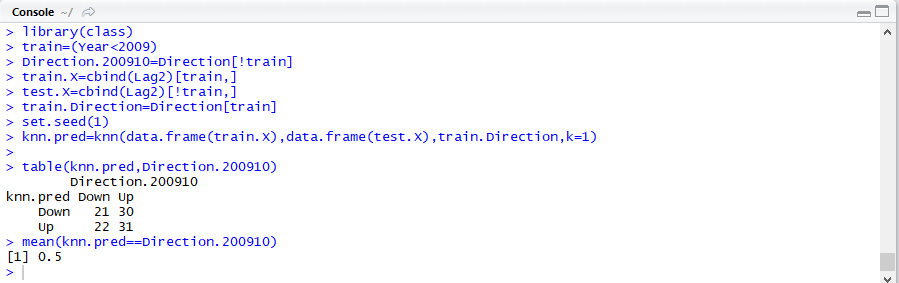
Accuracy= (True Positive+True Negative)/(Negative+Positive) = (61)/(61+43) =0.5865=58.65%

Error Rate= 1- Accuracy = 0.4134=41.34%

Sensitivity =True Positive / Positive = 61/(61+0)= 1=100%

Specificity=True Negative/Negative=0/(0+43)=0=0%

1)g) KNN



Accuracy= (True Positive+True Negative)/(Negative+Positive) = (31+21)/(31+21+30+22) =0.5=50%

Error Rate= 1- Accuracy = 0.5=50%

Sensitivity =True Positive / Positive = 31/(31+30)= 0.5082=50.82%

Specificity=True Negative/Negative=21/(21+22)=0.4883=48.83%

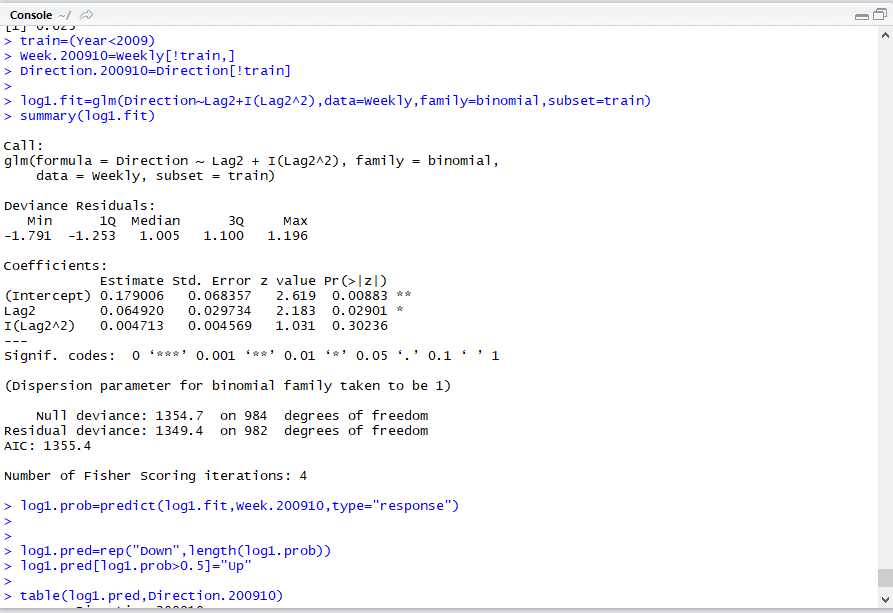
1)h)

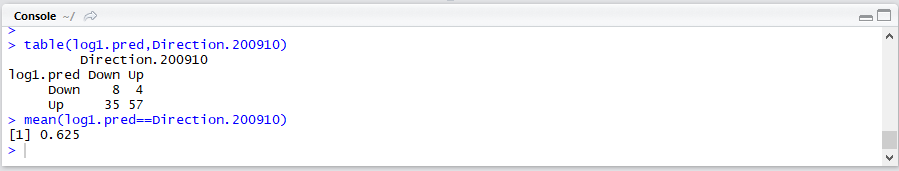
|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy(%)** | **Error Rate(%)** |
| Logistic | 62.5 | 37.5 |
| LDA | 62.5 | 37.5 |
| QDA | 58.65 | 41.34 |
| KNN | 50 | 50 |

According to the above table, Logistic and the LDA models have the highest accuracy and the lowest error rate and seem to be the methods with the best results for this data.

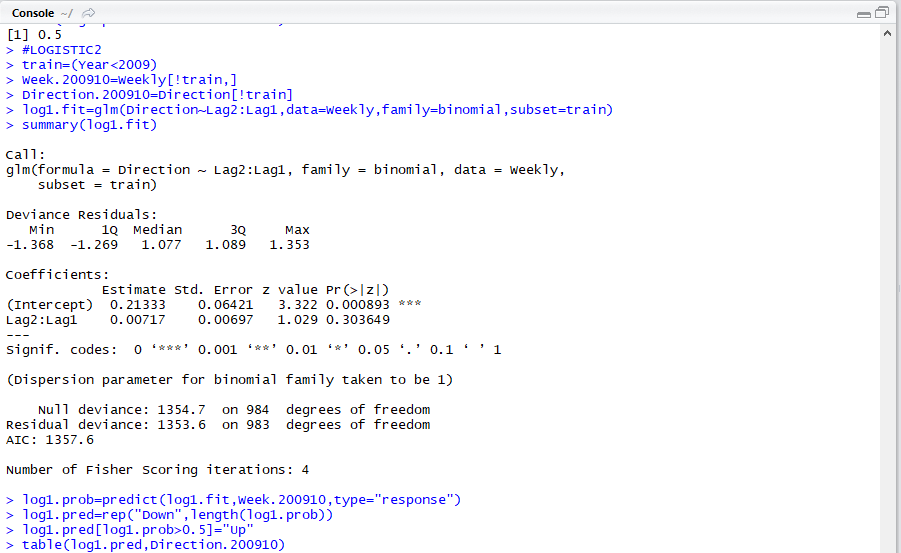
1. i)

#LOGISTIC1





#LOGISTIC2





Of the above 2,

The basic model with Direction calculated using

1. **Direction~Lag2 (62.5%)**
2. Lag2+(Lag^2) (62.5%) **–** Cannot use this as (Lag2^2) does not have significance wrt p values
3. Lag2:Lag1 (58.65%)

As the basic model (1) itself is having the higher Accuracy and hence a high error rate, it can be chosen over model(2) as the accuracy is the same for additional terms and effort. With respect to interaction Lag2:Lag1 interaction gives the maximum accuracy of all the interaction possibilities.

Lag 2 is the central terms in all the models considered as it is the most significant variable in the basic model.

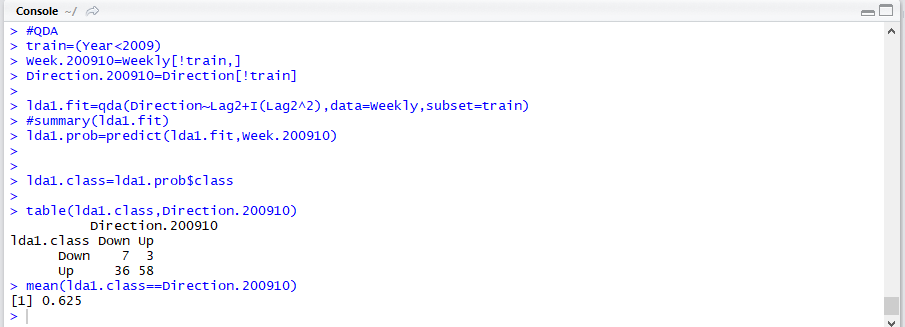
#LDA



Adding interaction terms and transformations only makes the basic model worse(at best 58.65%). Hence,

**Direction ~ Lag2** **is the best model** for the given data with an error rate of 37.5 and an accuracy of 62.5%.

#QDA

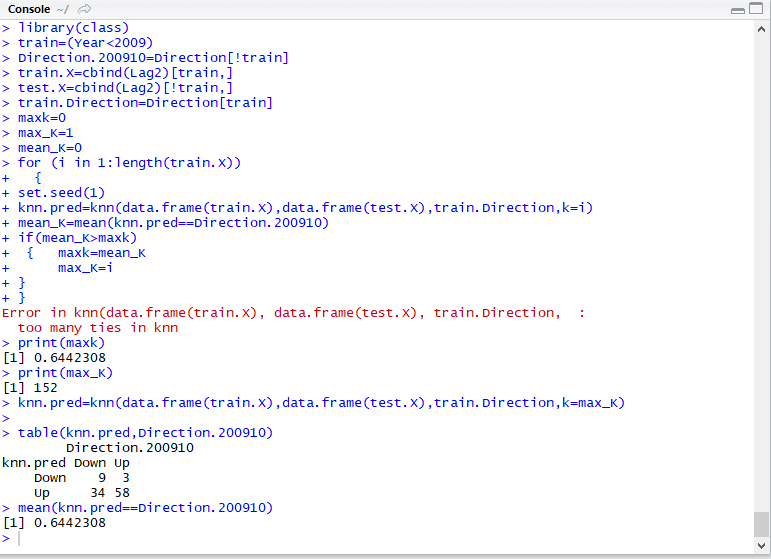


Taking a square transformation increases the accuracy from 58.65%(basic model) to 62.5%. Hence

**Direction~Lag2+(Lag2^2)**

Is a good model for QDA

#KNN

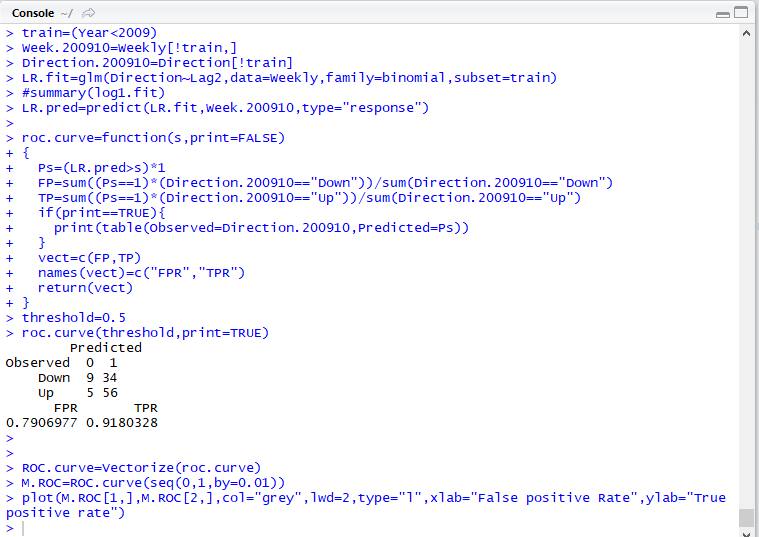


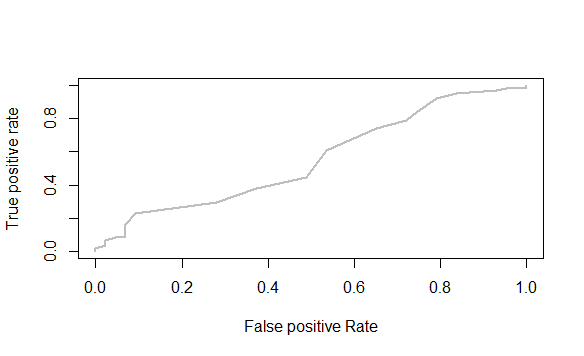
It can be seen on running a for loop for all values of K possible till the length of the train data set, it can be seen that with a K value of 152, accuracy obtained which is the maximum is 0.644=64.4%

Though K values of around 10 are preferred because, the test error rate might be really high at such high values(152) of K in general, though here it turns out better here.

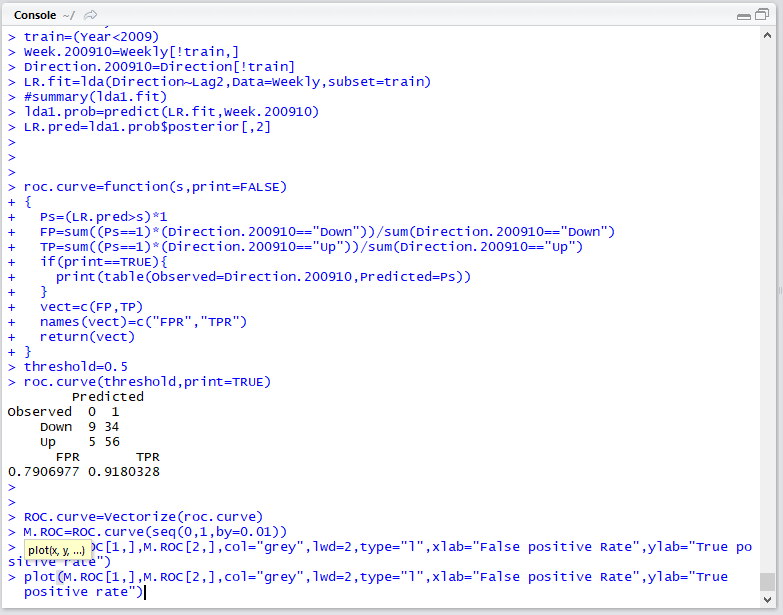
2)

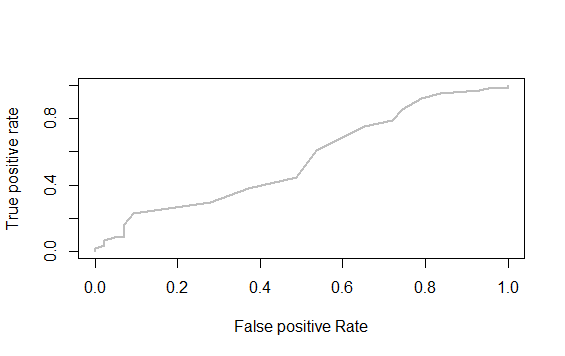
The best model chosen in Question 1(i) is Direction~Lag2 .





The ROC Curve graces along the 45degree line(AUC around 0.5) and slightly above , hence the model can be assumed to be more or less like a random guessing model. The model is not a good one.

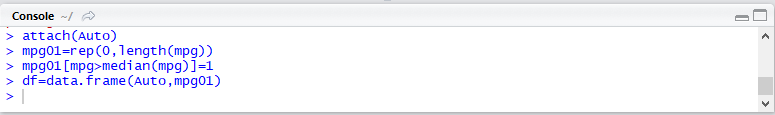




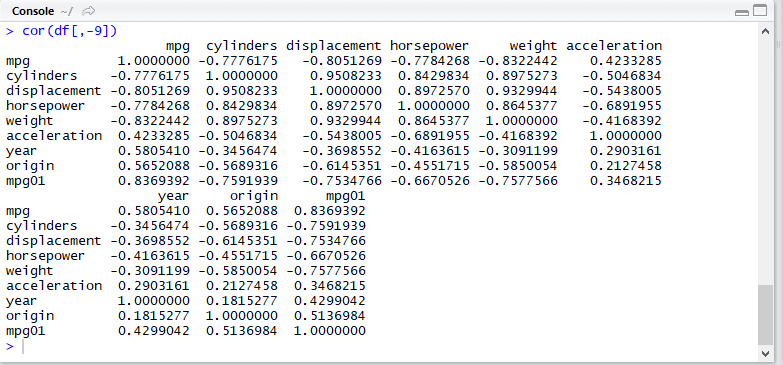
The ROC Curve graces along the 45degree line and slightly above , hence the model can be assumed to be more or less like a random guessing model. The model is not a good one.

3)

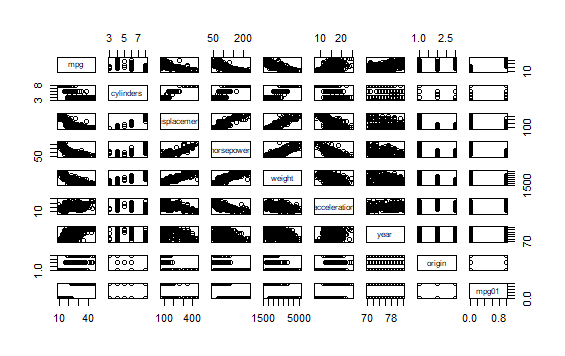
a)



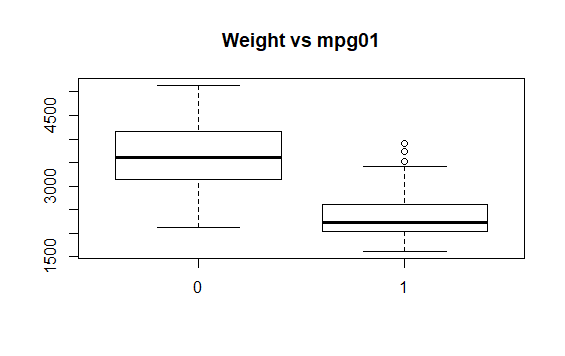
b)

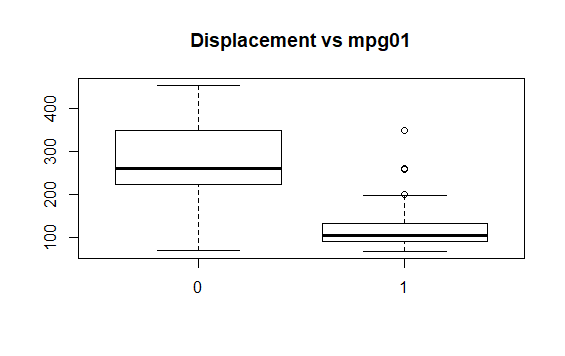


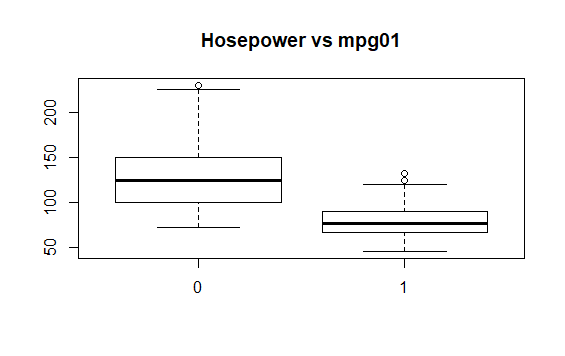
pairs(df)

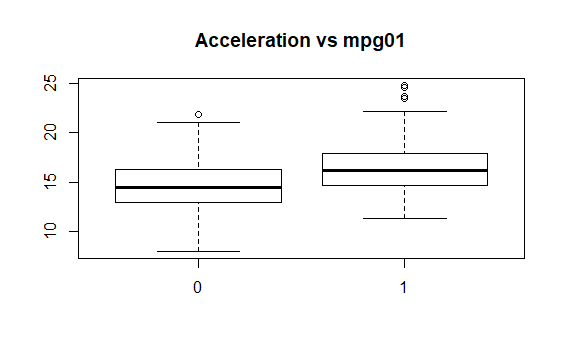


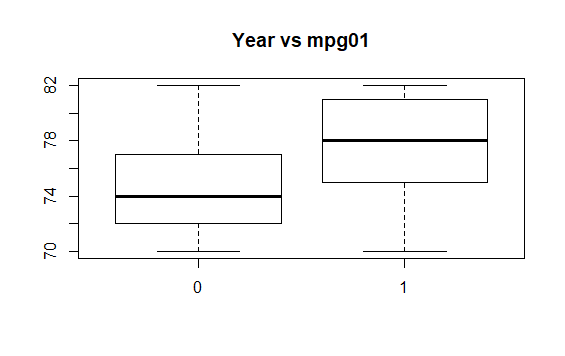
Mpg01 has some relation with displacement, horsepower, weight and acceleration according to the above boxplot.

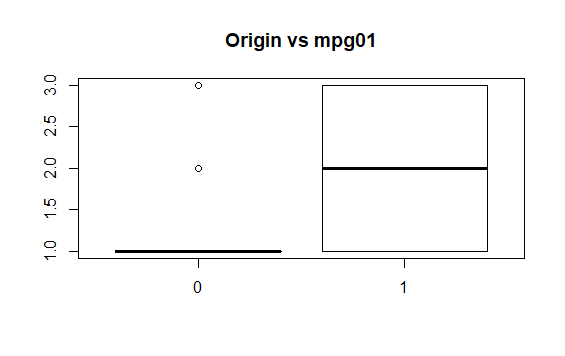


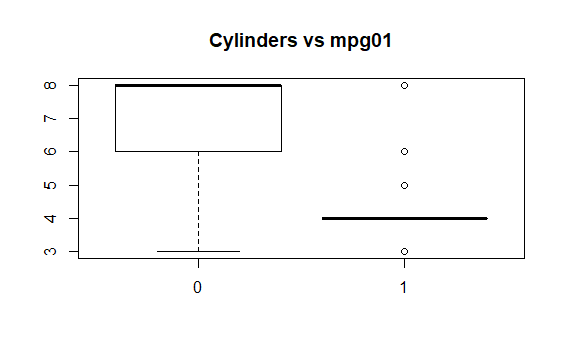










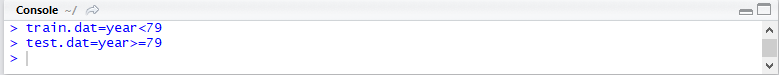


It can be concluded that there is some relationship between mpg01 and weight, displacement, horsepower, cylinders.

c)

#According to the summary, 75 percentile of data is above year=79, hence assuming 75 % of

data to bet rain data and 25% data as test data, data is split accordingly.

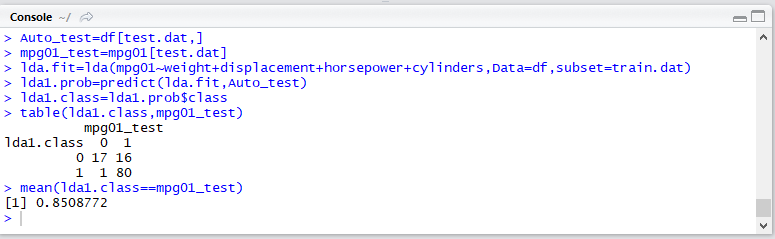


train.dat is considered as the train data

test.dat is considered as the test data

d)

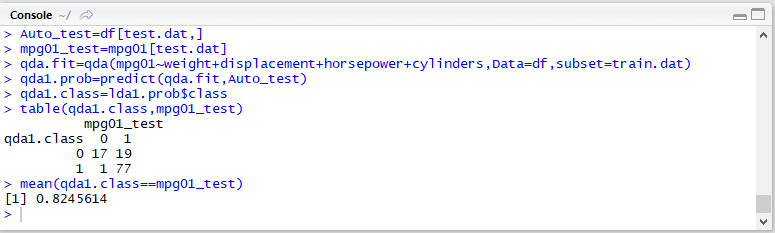
LDA



Error Rate: 14.92%

e)

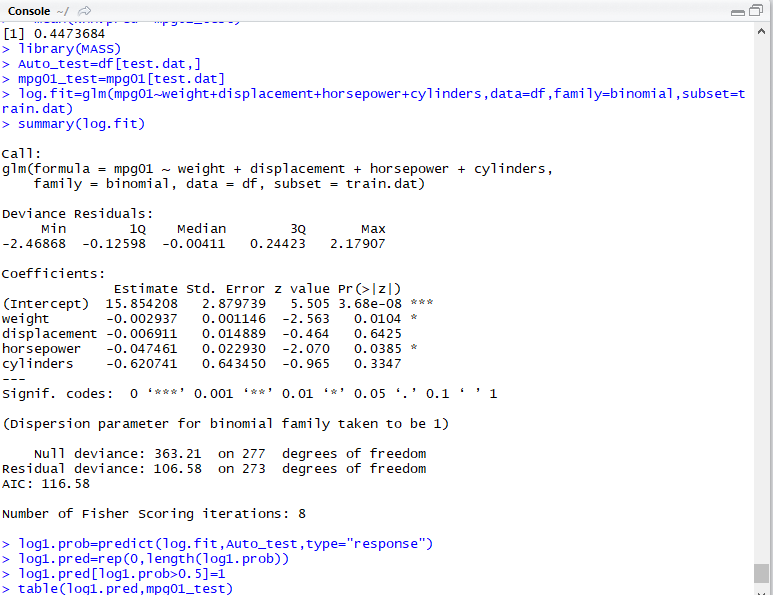
QDA

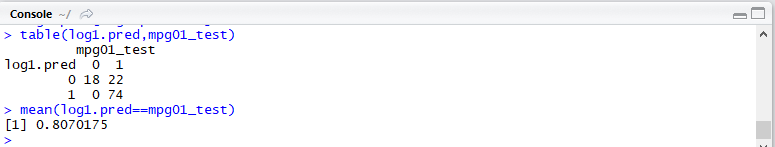


Error Rate: 17.56%

f)

Logistic

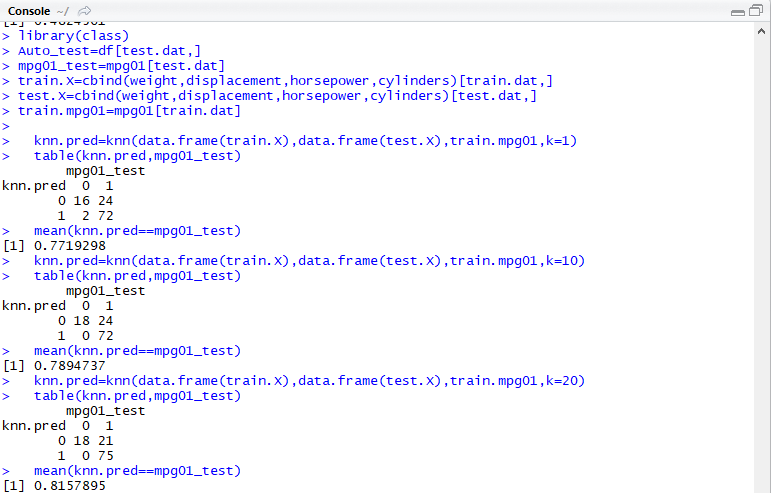


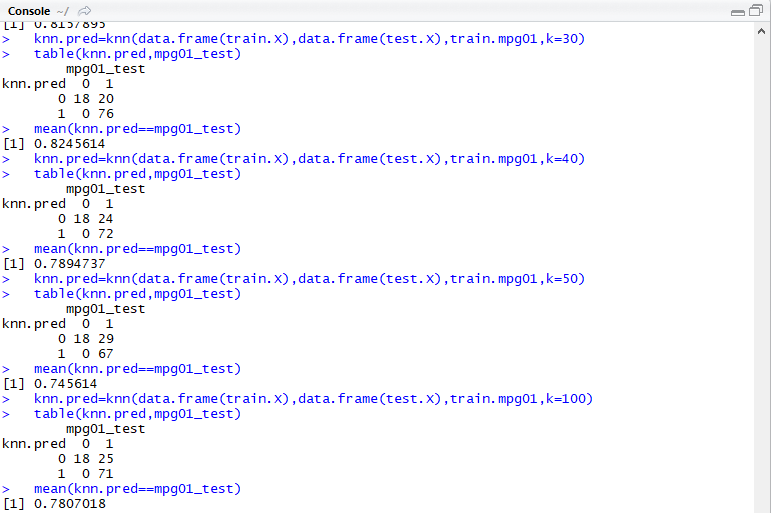


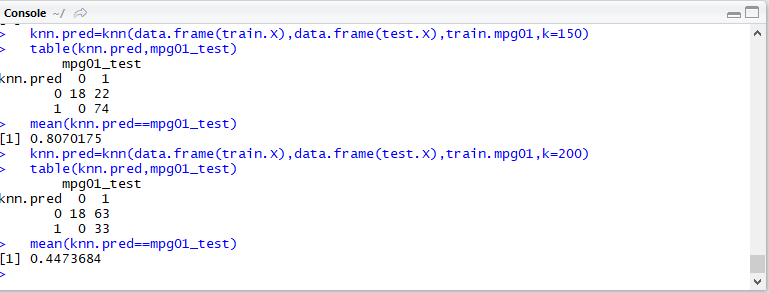
Error Rate=19.3%

g)

KNN







|  |  |
| --- | --- |
| K Value | Error Rate |
| 1 | 22.8% |
| 10 | 19.3% |
| 20 | 18.42% |
| 30 | 17.54% |
| 40 | 21.05% |
| 50 | 25.44% |
| 100 | 21.93% |
| 150 | 19.3% |
| 200 | 55.3% |

K Value of 30 seems to perform better for this KNN model.