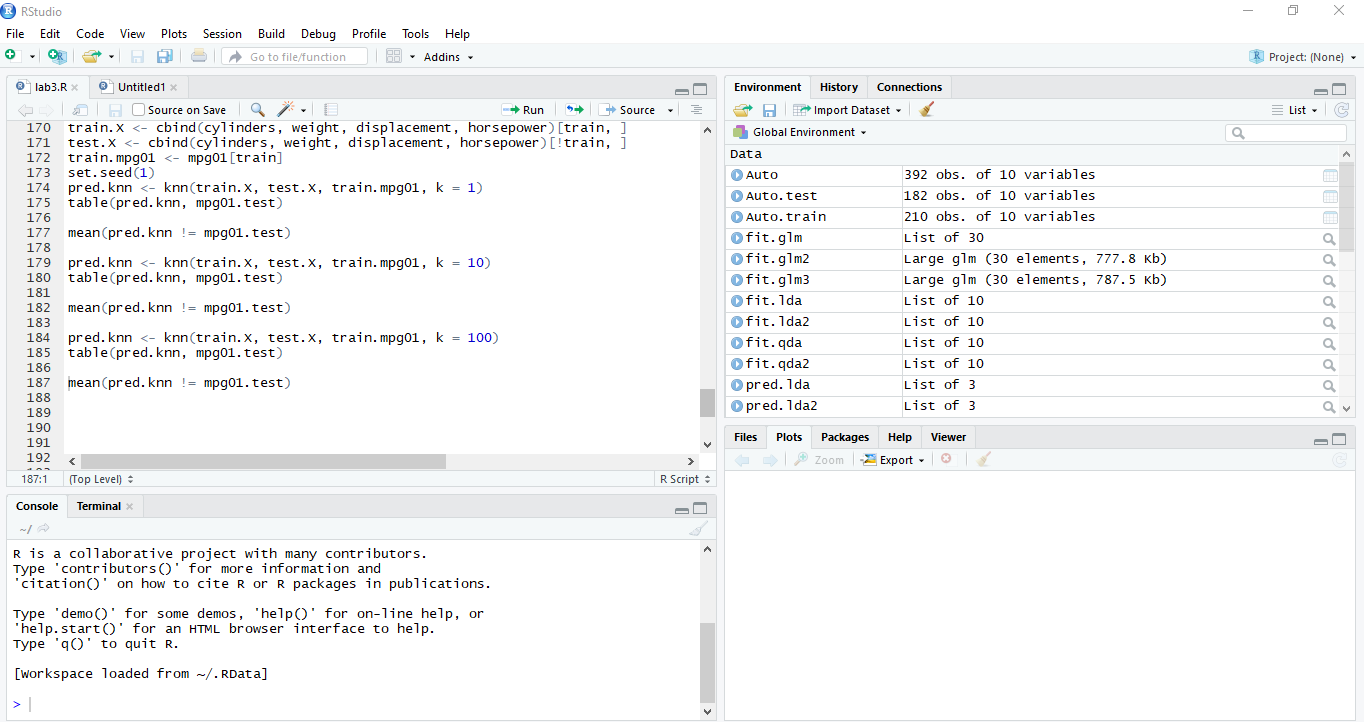
# **Introduction to Statistical Learning Lab3**

**Name: Sandeep Reddy Salkuti**

**Id: 16296868**

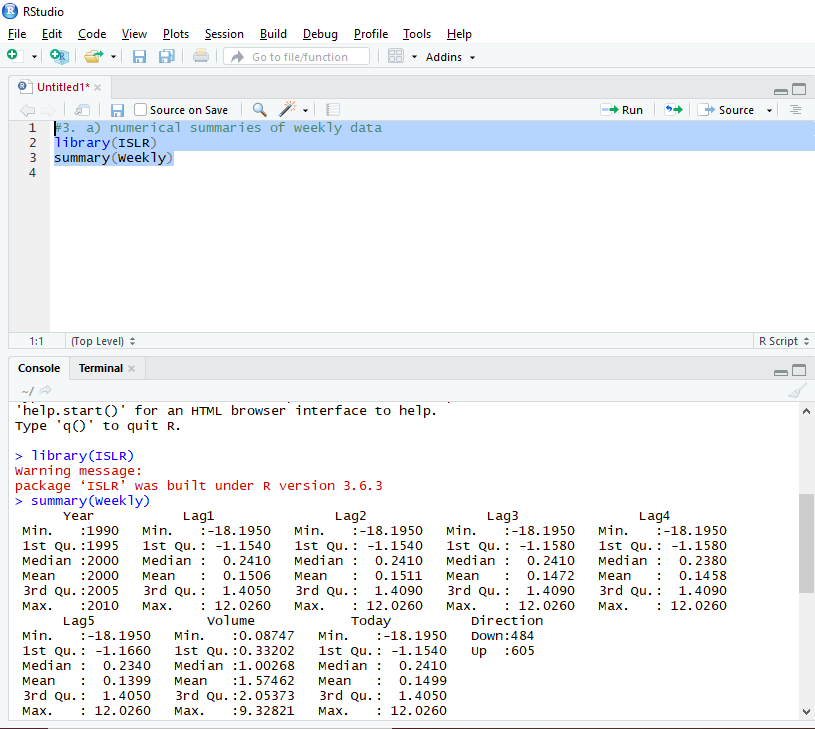
**Email:** [**sswf7@umsystem.edu**](mailto:sswf7@umsystem.edu)

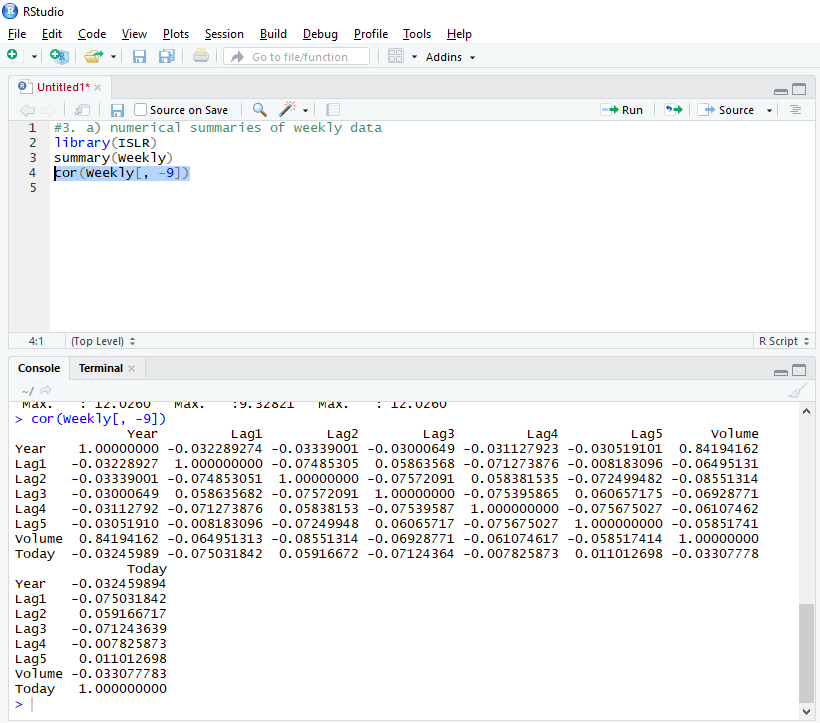
1. **You may download the R Code for Labs and the Data Sets to use from the textbook website.**

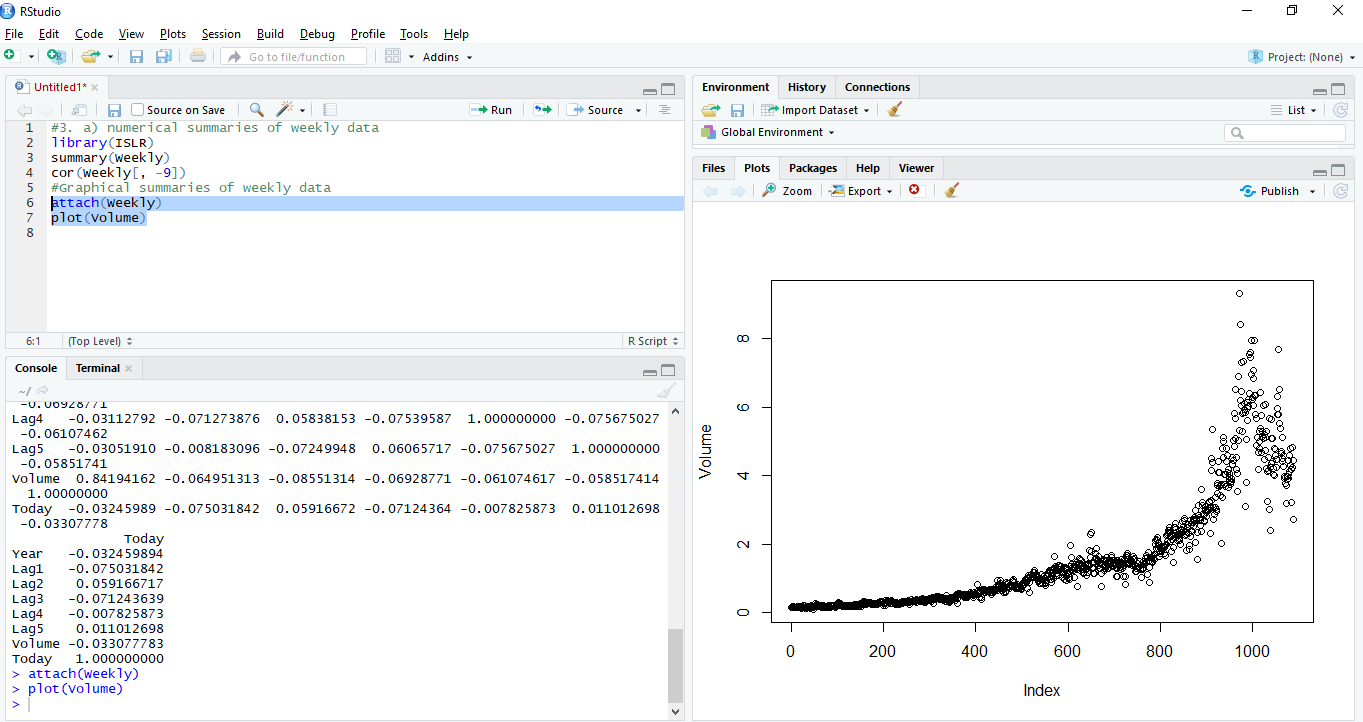


**2. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapters lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.**

**(a) (5 points) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?**

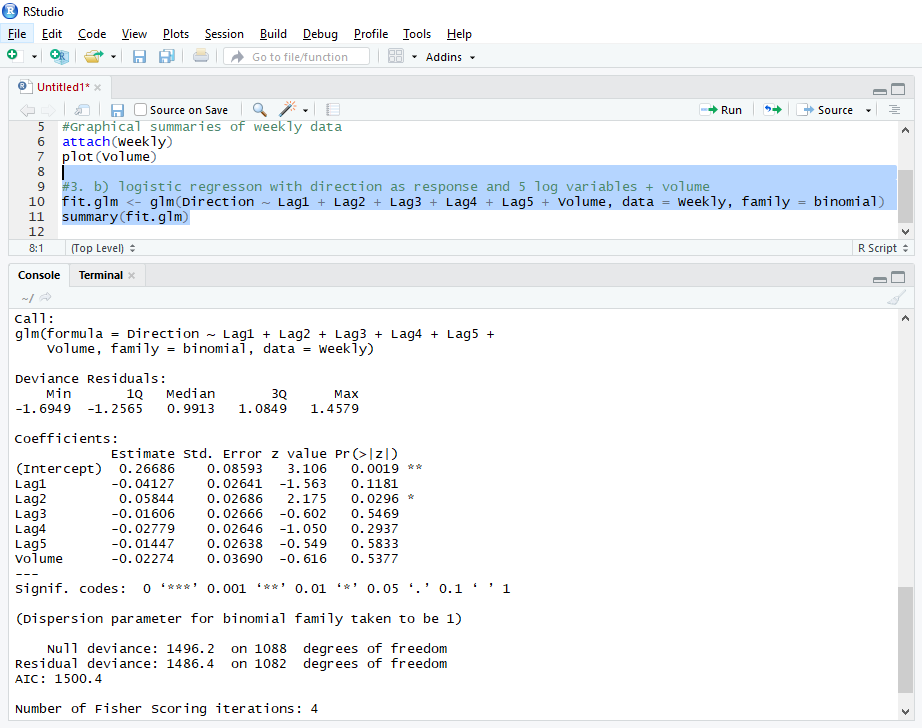






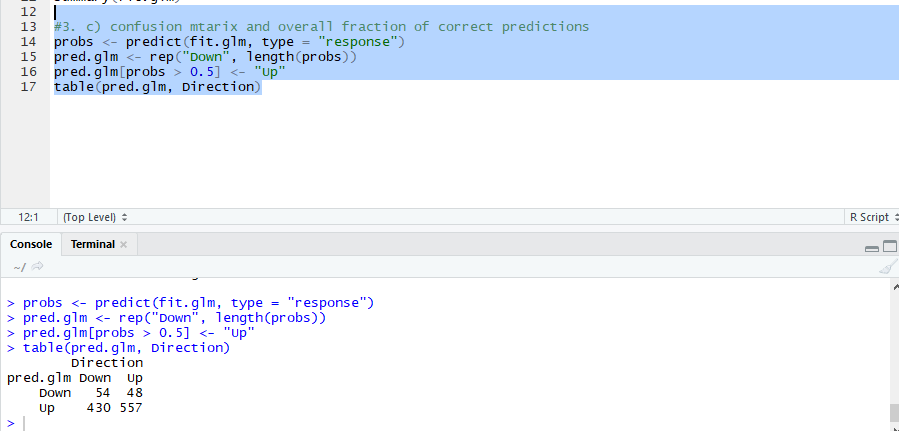
From numerical summaries correlations between lag variables and today’s returns are close to zero. Their exists substantial correlation between and “year” and “volume”. From the above plot it is clear that volume increases for index values and

**(b) (5 points)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?**



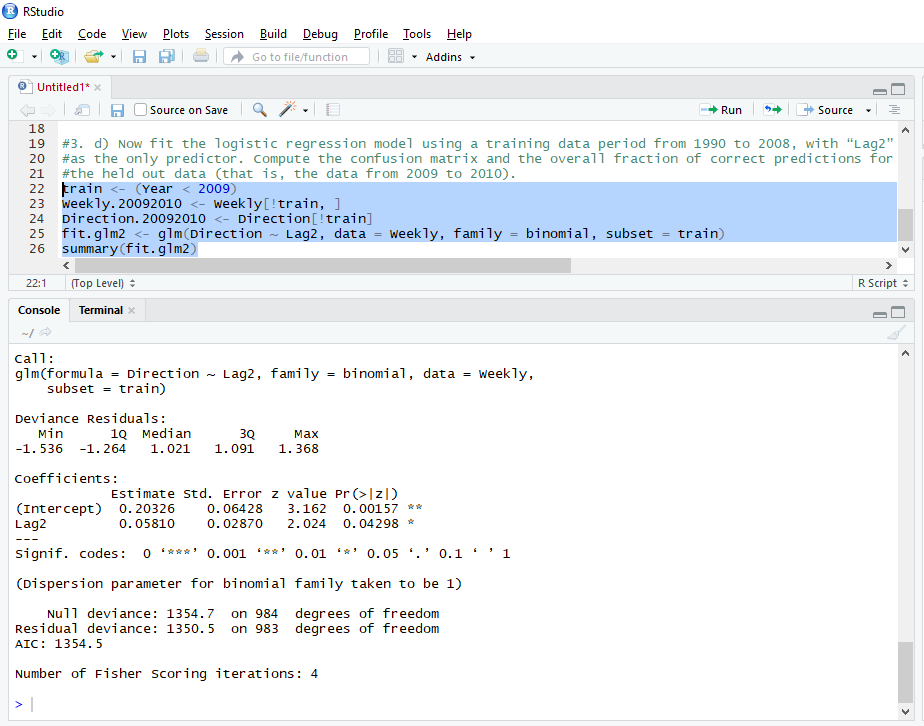
From the above it is clear that “**Lag2”** appears to be only predictor that is statistically significant. It is because of its p value is less than 0.05 ( Pr(>|z|) = 3%.)

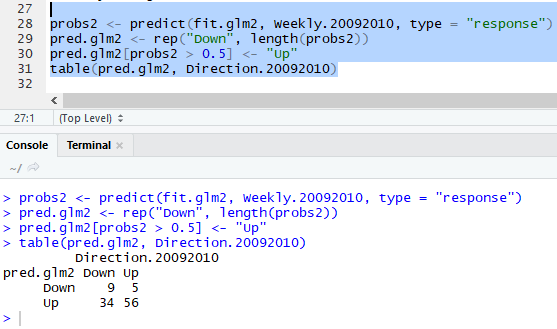
**(c) (5 points) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.**



From above it is clear that percentage of correct predictions on the training data is (54 + 557)/1089= 0.5610 which is equal to 56.1% In contrast training error rate is 43.89% which is often overly optimistic. We can say that for weeks when the market goes up, the model is 92.0661% of the time (557/(48+557)). For weeks when the market goes down, the model is only 11.15702% of the time (54/(54+430)).

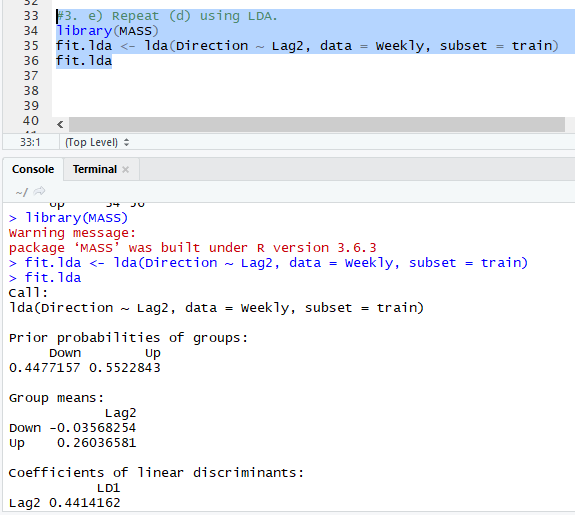
**(d) (5 points) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).**

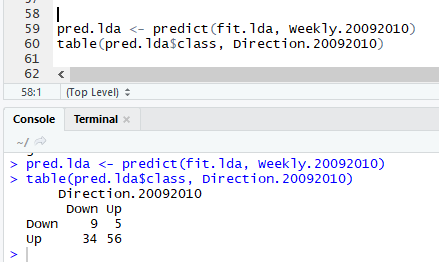




From the above it is clear that the percentage of correct predictions on the test data is (9+56)/104 which equals to 62.5%. In other words test error rate is 37.5%. Also for weeks when the market goes up, the model is 91.803% of the time (56/(56+5)). For weeks when the market goes down, the model is only 20.930% of the time (9/(9+34)).

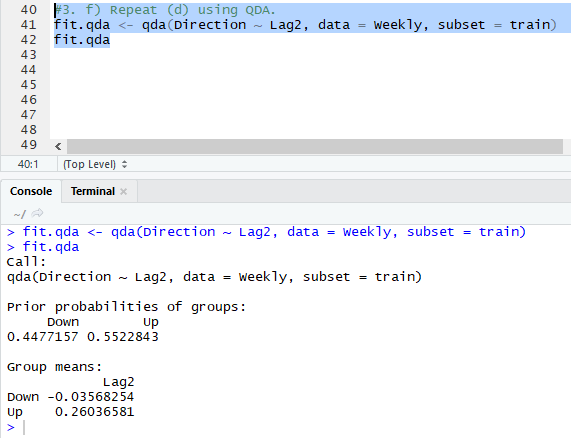
**(e) (5 points) Repeat (d) using LDA.**

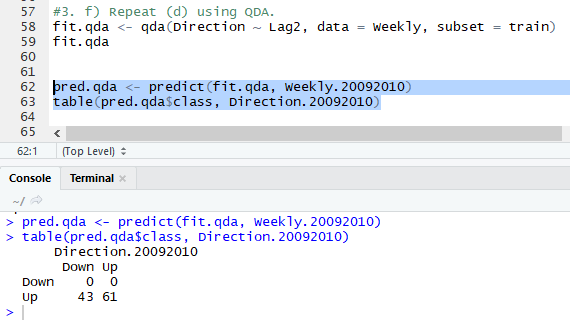




From above it is clear that percentage of correct predictions on the test data is 62.5%. In other words test error rate is 37.5%. We can say that for weeks when the market goes up, the model is 91.803% of the time. For weeks when the market goes down, the model is only 20.9302% of the time. These results are very close to those obtained with the logistic regression model which is not surprising.

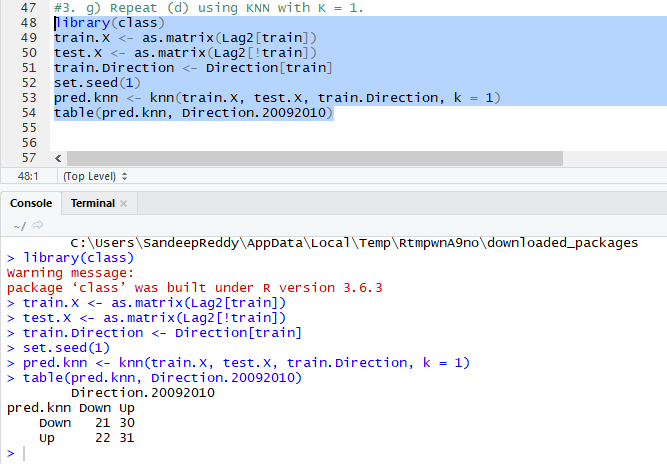
**(f) (5 points) Repeat (d) using QDA.**





From above it is clear that the percentage of correct predictions on the test data is 58.6538%. In other words test error rate is 41.3461%. We could also say that for weeks when the market goes up, the model is 100% of the time. For weeks when the market goes down, the model is only 0% of the time. It is clear that QDA achieves a correctness of 58.6538% even though the model chooses “Up” the whole time.

**(g) (5 points) Repeat (d) using KNN with K = 1.**



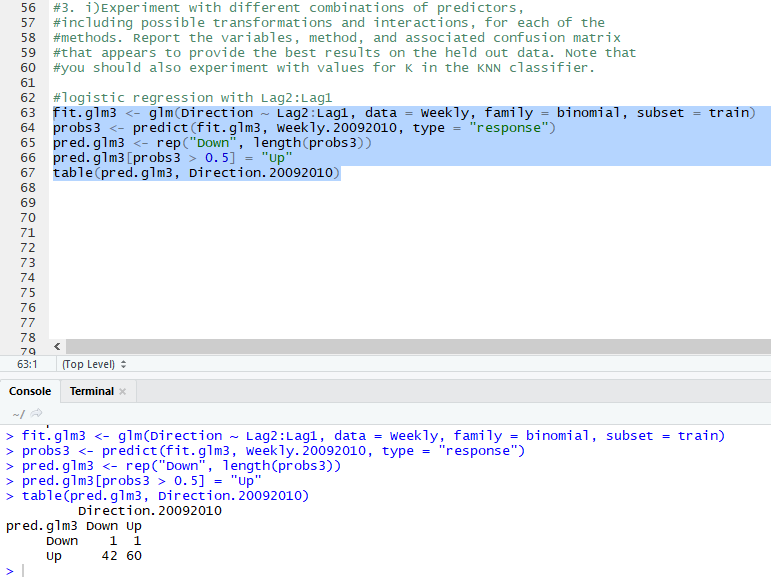
From above it is clear that the percentage of correct predictions on the test data is 50%. In other words test error rate is 50%. We can also say that for weeks when the market goes up, the model is 50.8196% of the time. For weeks when the market goes down, the model is only 48.8372% of the time.

**(h) (5 points) Which of these methods appears to provide the best results on this data?**

After comparison of all test error rates, It is clear that logistic regression and LDA have the minimum error rates, followed by QDA and KNN.

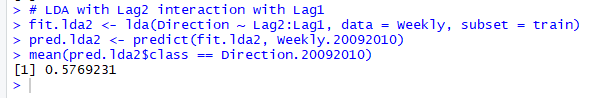
1. **(5 points) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.**

**Logistic regression with Lag2:Lag1**

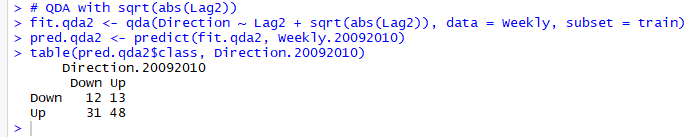




**LDA with Lag2 interaction with Lag1**

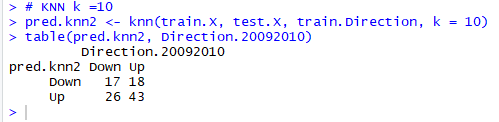


**QDA with sqrt(abs(Lag2))**



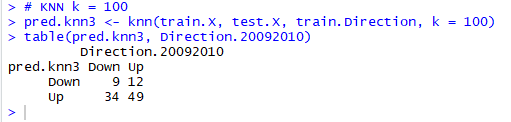


**KNN with K=10**





**KNN with K=100**

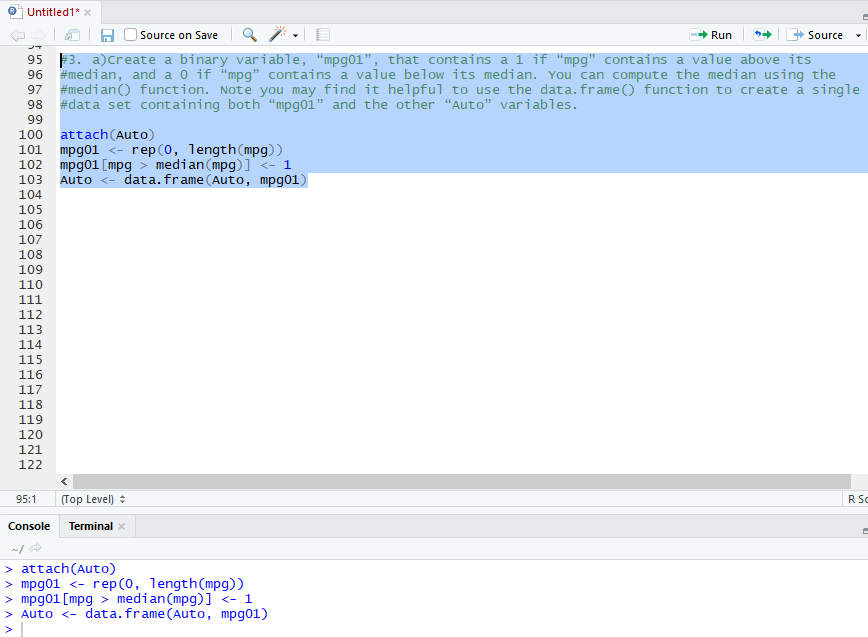




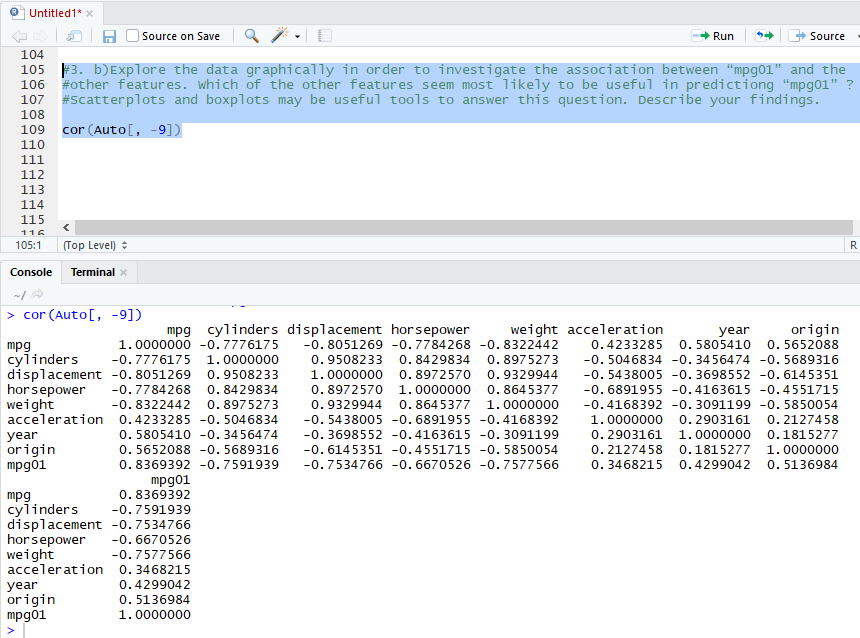
Out of all of these combinations, the original logistic regression and LDA have the best performance in terms of test error rates.

**3. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.**

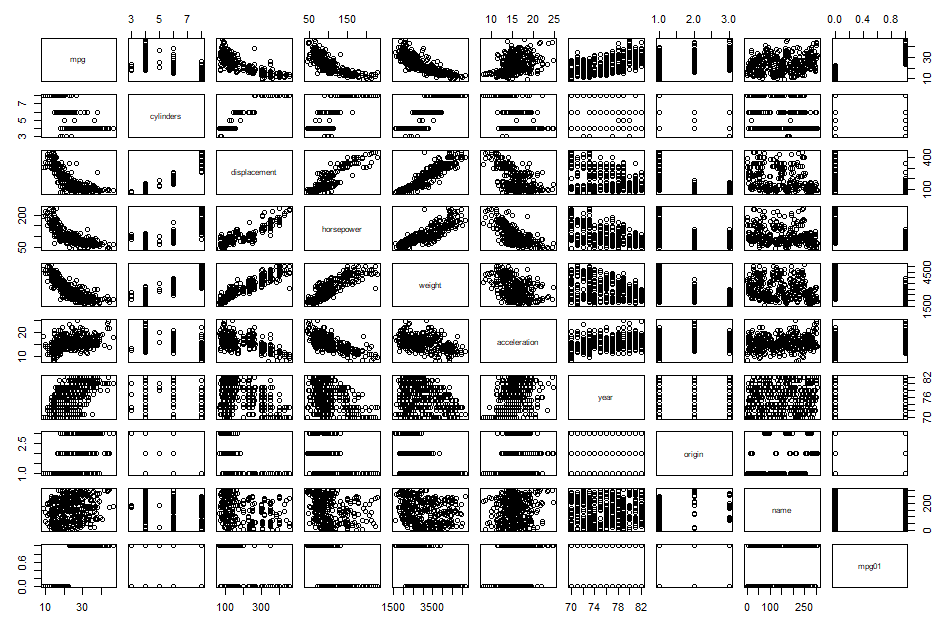
**(a) (5 points) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can 2 compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.**



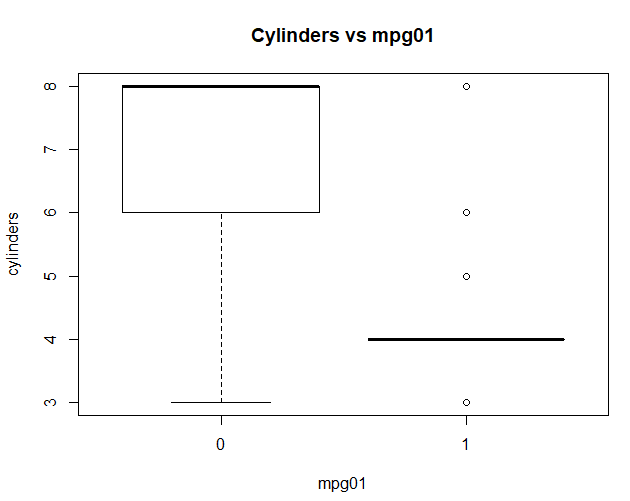
**(b) (5 points) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.**



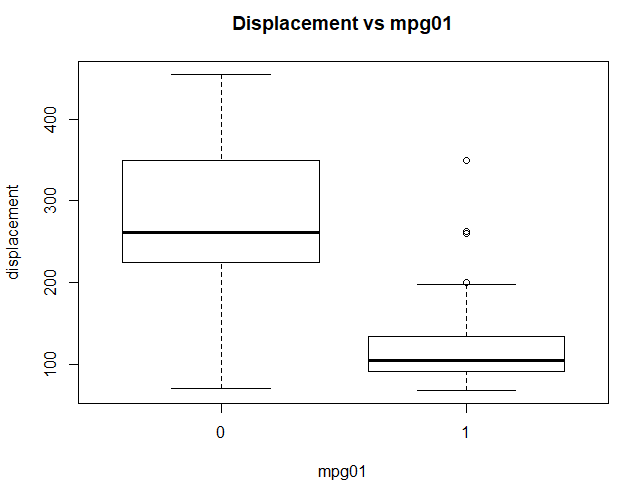




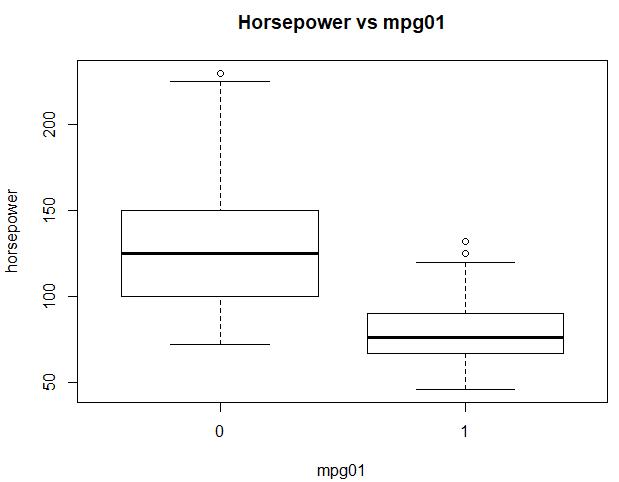




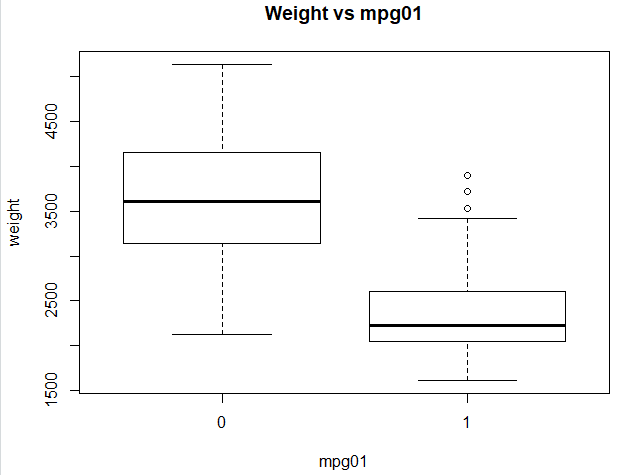




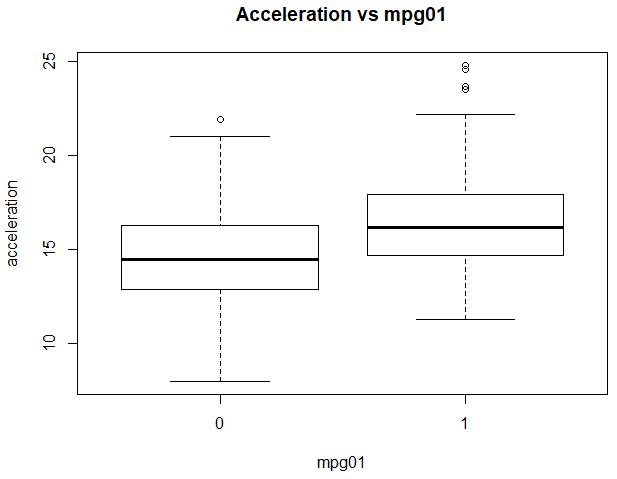




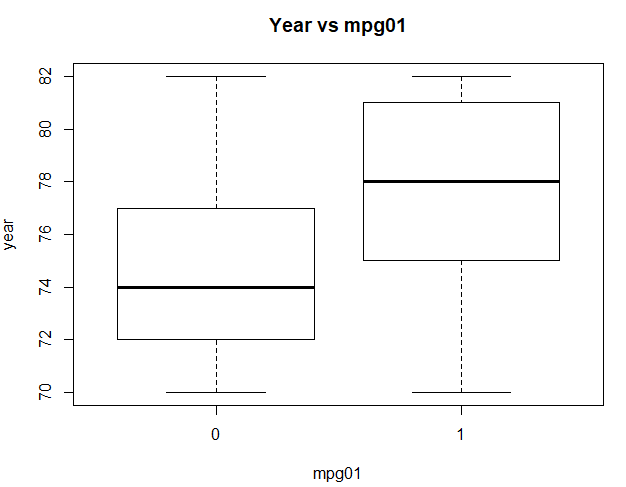






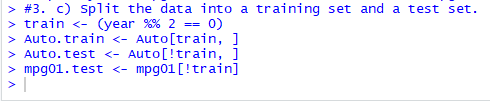




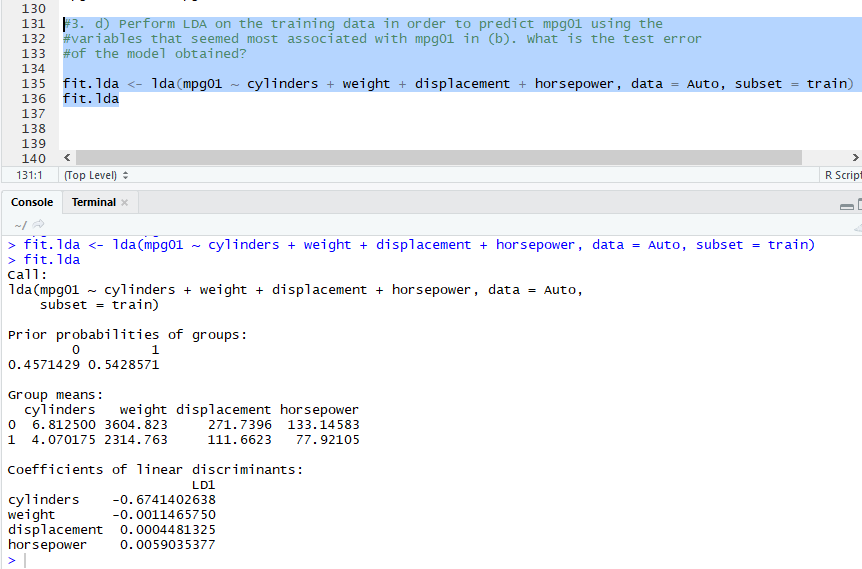


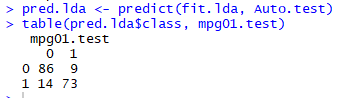
For predicting mpg01 I used box plot for all data against mpg01 and found that there exists some association between “mpg01” and “cylinders”, “weight”, “displacement” and “horsepower”.

**(c) Split the data into a training set and a test set.**



**(d) (5 points) 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?**

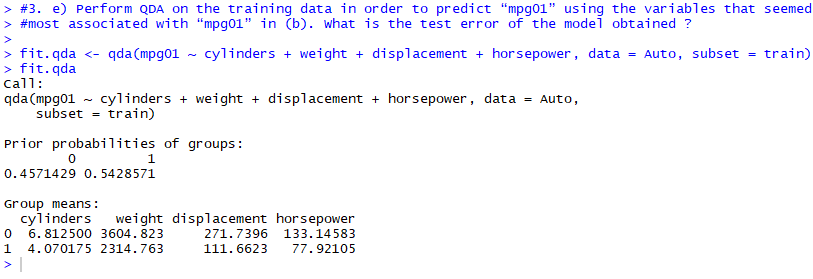


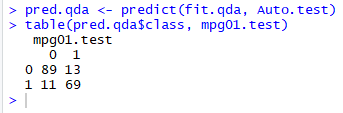




From above it is clear that the test error rate is 12.637%

**(e) (5 points) 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?**

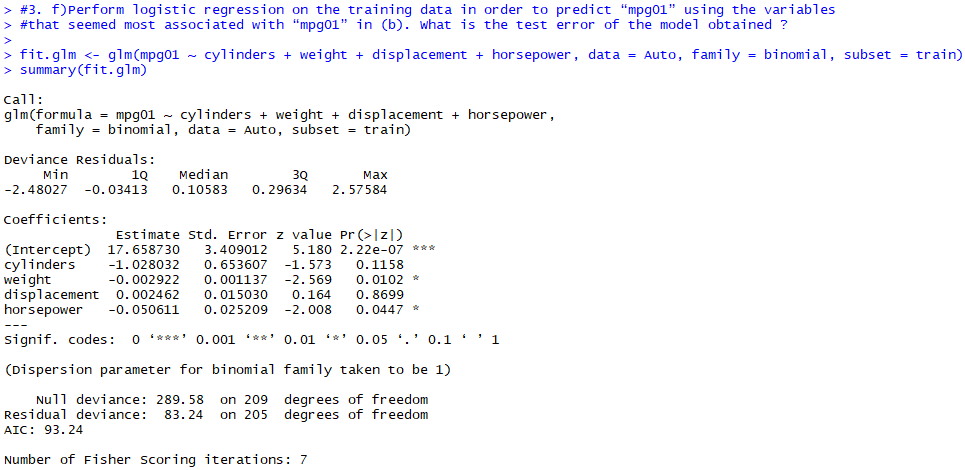


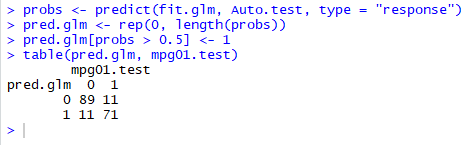




From above it is clear that test error rate is 13.186%

**(f) (5 points) 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?**

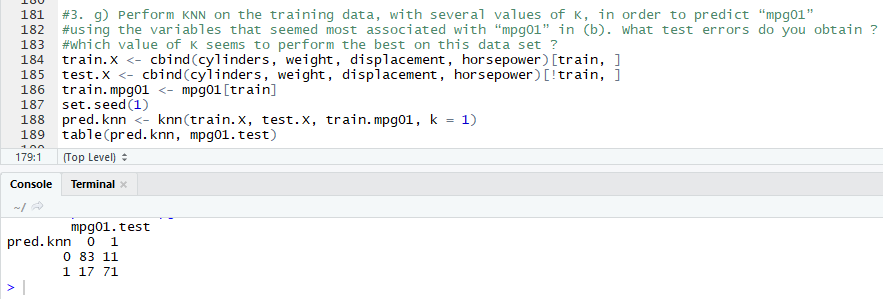






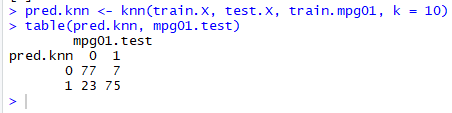
From above it is clear that test error rate is 12.08791%

**(g) (5 points) 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?**



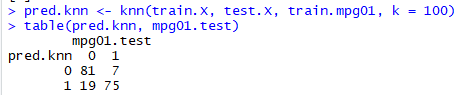


For K=1 the test error rate is 15.38462%





Foe K=10 the test error rate is 16.48352%





From above for K=100 the test error rate is 14.28571%