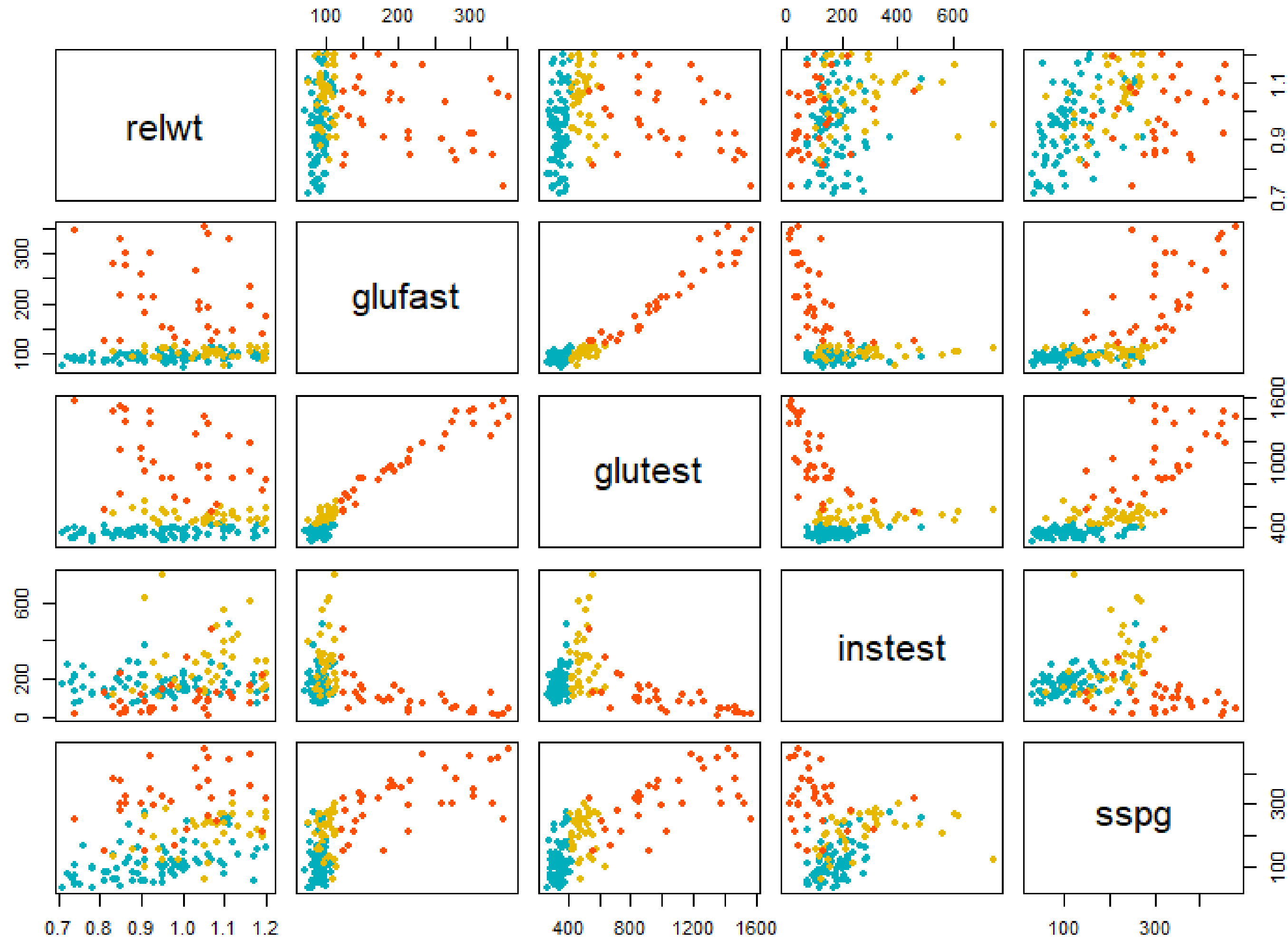
HOMEWORK 4

Rohith Reddy Kolla | rkolla@buffalo.edu

1)

a) Pairwise scatterplots for all five variables with the normal group represented in blue, the chemical diabetics in yellow and the overt diabetics in red.



In most of the scatter plots above, we can see a clear separation between the 3 groups indicating that they may have different covariance matrices. In particular, the overly diabetic group has a significantly different covariance matrix from the rest whereas the normal group and chemical diabetic group do not do not show as much difference.

Using the MVN package to check if they are multivariate normal shows that they are not according to both Mardia's and Henze-Zirkler's MVN test.

b) Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) is done using the lda and qda functions from the MASS package.

```
> library(MASS)
> ldax = lda(group~., data = Diabetes)
> ldap = predict(ldax, Diabetes)
> mean(ldap$class == Diabetes$group)
[1] 0.9034483
> qdax = qda(group~., data = Diabetes)
> qdap = predict(qdax, Diabetes)
> mean(qdap$class == Diabetes$group)
[1] 0.9517241
```

With the given data, QDA performs better than LDA as shown above by about 5% accuracy. The confusion matrix for lda predictions and qda predictions respectively are shown below. Y axis are actual classes and X axis are predictions.

```
> cfml <- table(Diabetes$group,ldap$class)
> cfml

                    Normal Chemical_Diabetic Overt_Diabetic
  Normal                73                 3              0
  Chemical_Diabetic      5                31              0
  Overt_Diabetic         1                 5             27
> cfmq <- table(Diabetes$group,qdap$class)
> cfmq

                    Normal Chemical_Diabetic Overt_Diabetic
  Normal                75                 1              0
  Chemical_Diabetic      3                33              0
  Overt_Diabetic         0                 3             30
```

c) Given the details of the individual, LDA assigns him in the Normal class whereas QDA assigns him in the Overt Diabetic class.

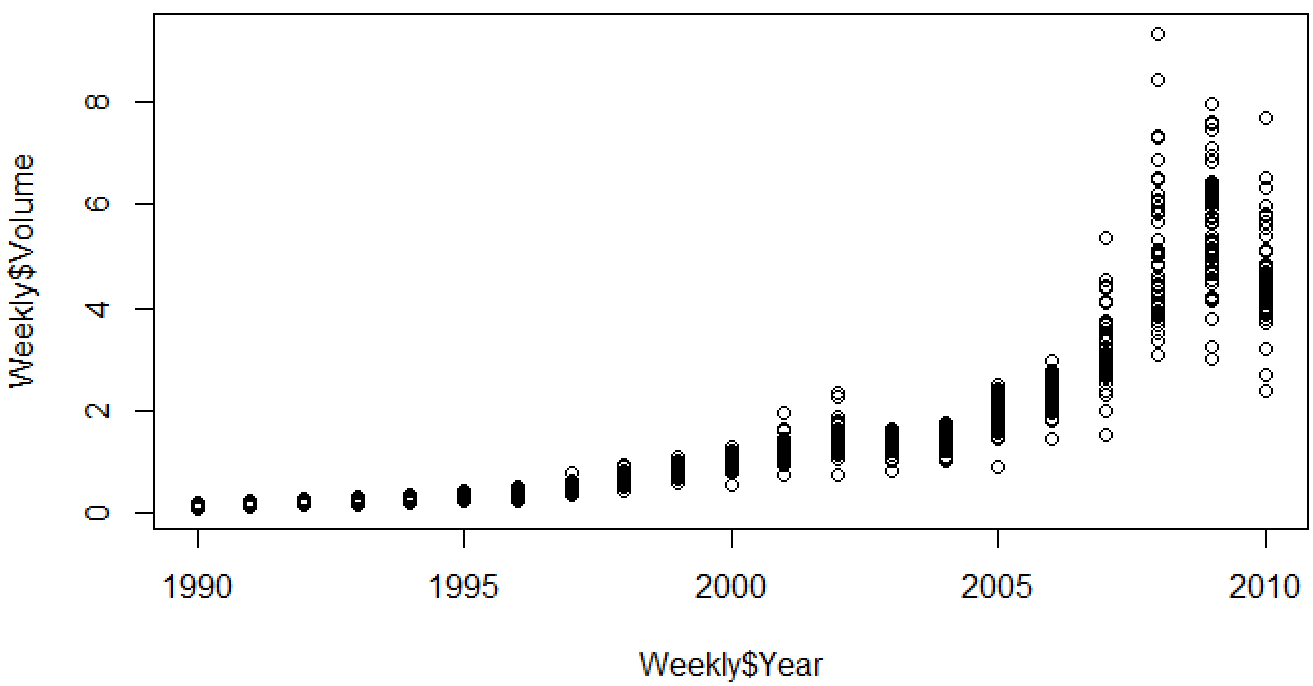
```
> relwt = 1.86
> glufast = 184
> glutest = 68
> instest = 122
> sspg = 544
> indiv <- data.frame(relwt,glufast,glutest,instest,sspg)
> ldaip <- predict(ldax, indiv)
> ldaip$class
[1] Normal
Levels: Normal Chemical_Diabetic Overt_Diabetic
> qdaip <- predict(qdax, indiv)
> qdaip$class
[1] Overt_Diabetic
Levels: Normal Chemical_Diabetic Overt_Diabetic
```

2)

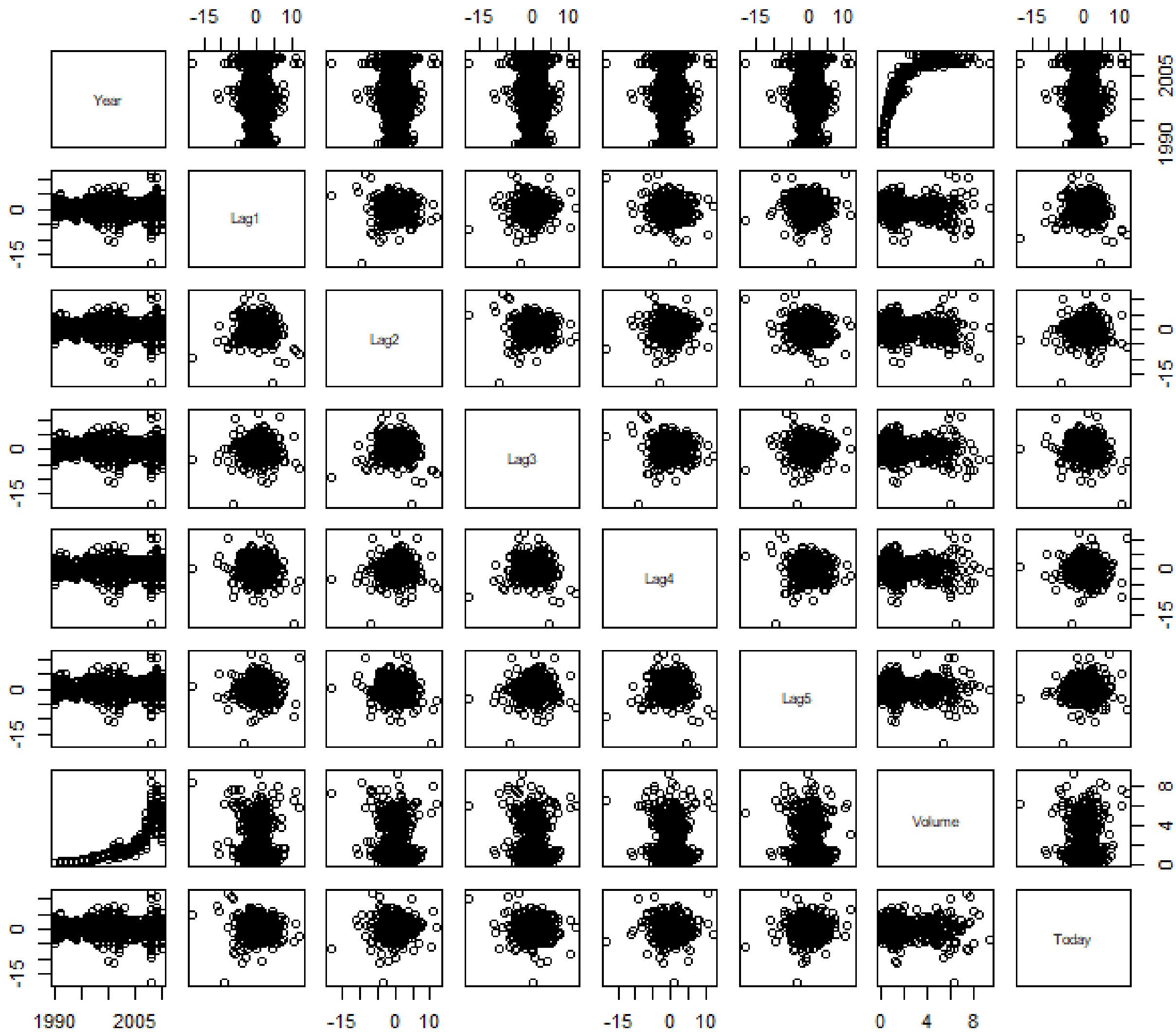
a) Summaries of the "Weekly" data –

```
> summary(Weekly)
      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          Direction
 Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950   Down:484 
 1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605 
 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            
```

From the above information, the "Weekly" data has 9 variables and 1089 observations. The Year variable is the year of the observation ranging from 1990 to 2010. The Lag variables indicate the percentage return for the previous number of weeks ex – Lag1 for previous week, Lag2 for previous 2 weeks. Volume indicates the average number of daily shares traded in billions. Today indicates the percentage return for the current week and Direction indicates whether the market had a positive or negative return.



The above plot of volume and year shows a clear increase in the number of daily shares traded with increasing year. The remaining basic plots do not provide any significant visual insight into the data.



b) Applying logistic regression with Direction as the response variable using the glm() function.

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = "binomial", data = Weekly[, 2:9])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6949  -1.2565   0.9913   1.0849   1.4579

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
```

From the information shown above, Lag2 seems to be the only predictor that is statistically significant.

c) The confusion matrix with true classes on Y axis and predicted classes on X axis shown below indicates that most of the predictions are “Up” leading to 430 wrong “Up” predictions out of 1089 predictions. As a result, only a relative few of “Downs” are being predicted out of which a further few are correct. 1 and 0 are the predictions which indicate “Up” and “Down” respectively.

	p	
	0	1
Down	54	430
Up	48	557

d) Fitting a training data period from 1990 – 2008 with Lag2 as the only predictor and testing on data from 2009 and 2010 gives successful prediction rate of 62.5%. 1 and 0 are the predictions which indicate “Up” and “Down” respectively.

	p2	
	0	1
Down	9	34
Up	5	56

e) Repeating d) using LDA provides essentially identical results with the same over prediction of “Up”.

	p3	
	Down	Up
Down	9	34
Up	5	56

f) Repeating d) using KNN with K = 1 provides better results than all other predictions before with a successful prediction rate of 50.96%. 2 and 1 are the predictions which indicate “Up” and “Down” respectively.

	pred4	
	1	2
Down	21	22
Up	29	32

g) Clearly from all the prediction methods used above, KNN with K = 1 appears to have provided the best results.