Homework 3

Kris Hanus, Laura Glathar, Arkya Rakshit, Jace Crist, Brandon Dlugosz
University of Nebraska at Omaha

BSAD 8700 - Business Analytics
Due: February 2, 2015

ANSWER FOR 10:

(a)
```
tail(Weekly, 1)
```

```
##      Year  Lag1  Lag2  Lag3  Lag4   Lag5   Volume Today Direction
## 1089 2010 1.034 0.283 1.281 2.969 -0.861 2.707105 0.069        Up
```

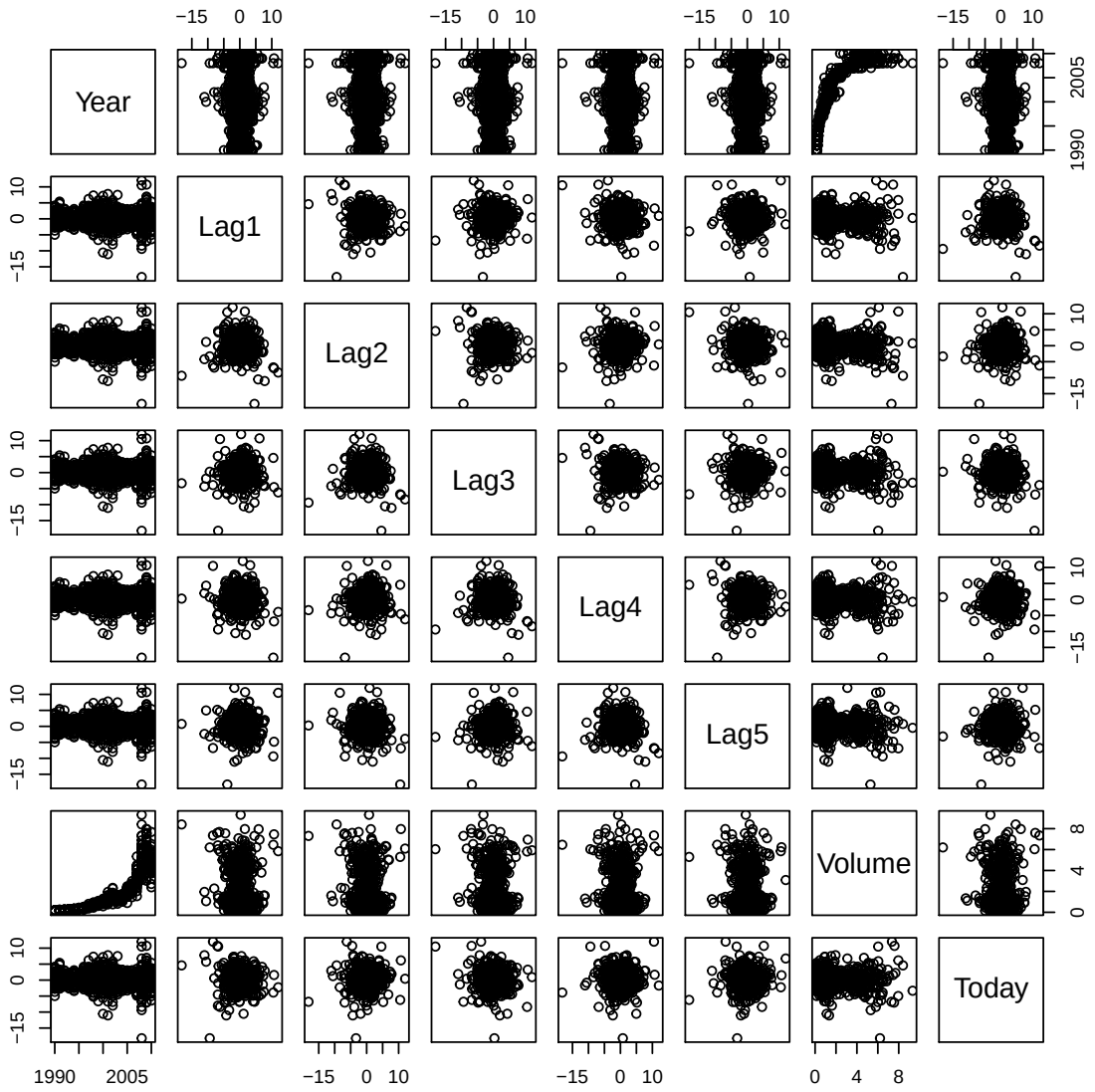
```
summary(Weekly)
```

```
##       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       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
##  Min.   :-18.1950   Down:484  
##  1st Qu.: -1.1540   Up  :605  
##  Median :  0.2410             
##  Mean   :  0.1499             
##  3rd Qu.:  1.4050             
##  Max.   : 12.0260
```

```
data1<-Weekly[,1:8]
attach(Weekly)
cor(data1)
```

```
##                Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
```

```
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000

pairs(data1)
```



There are a few interesting places of correlation. Primarily, with Volume and Year. Other wise, it is observable that each of the Lags are clustered, but it is difficult to observe other relationships.

(b)
```
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Weekly, family = binomial)
summary(glm.fit)
```

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

coef(glm.fit)

## (Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5
##  0.26686414 -0.04126894  0.05844168 -0.01606114 -0.02779021 -0.01447206
##      Volume
## -0.02274153
```

The only predictors which have significance are the intercept and Lag2. Lag2 is between 95% and 99% significant. The Intercept is 99% and 99.9%.

(c)
```
contrasts(Direction)

##      Up
## Down  0
## Up    1

glm.pred=rep("Down", 1089)
glm.probs=predict(glm.fit,type="response")
glm.probs[1:10]

##         1         2         3         4         5         6         7
## 0.6086249 0.6010314 0.5875699 0.4816416 0.6169013 0.5684190 0.5786097
##         8         9        10
## 0.5151972 0.5715200 0.5554287

glm.pred[glm.probs>0.5]<-"Up"
table(glm.pred,Direction)

##         Direction
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
```

```
557/(557+430)

## [1] 0.5643364

430/(557+430)

## [1] 0.4356636

48/(48+54)

## [1] 0.4705882
```

The confusion matrix shows that on days when logistic regression predicts an increase in the market, it has a 56.4% accuracy rate. The error rate is 43.6% for predicting Up and is actually Down. The error rate is 47.1% for predicting Down and is actually Up.

(d)
```
glm.fit2=glm(Direction~Lag2,data=Weekly, family = binomial)
summary(glm.fit2)

##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.564  -1.267   1.008   1.086   1.386
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.21473    0.06123   3.507 0.000453 ***
## Lag2         0.06279    0.02636   2.382 0.017230 *
## ---
## 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: 1490.4  on 1087  degrees of freedom
## AIC: 1494.4
##
## Number of Fisher Scoring iterations: 4

coef(glm.fit2)

## (Intercept)        Lag2
##  0.21473151  0.06279058

contrasts(Direction)

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

```
glm.pred2=rep("Down", 1089)
glm.probs2=predict(glm.fit2,type="response")
glm.probs2[1:10]

##         1         2         3         4         5         6         7
## 0.5777243 0.5661029 0.5492840 0.5132426 0.6071571 0.5644982 0.5716776
##         8         9        10
## 0.5321015 0.5659641 0.5541137

glm.pred2[glm.probs2>0.5]<-"Up"
table(glm.pred2,Direction)

##          Direction
## glm.pred2 Down  Up
##      Down   33  26
##      Up    451 579

579/(579+451)

## [1] 0.5621359

451/(579+451)

## [1] 0.4378641

26/(26+33)

## [1] 0.440678
```

The confusion matrix shows that on days when logistic regression predicts an increase in the market, it has a 56.2% accuracy rate. The error rate is 43.8% for predicting Up and is actually Down. The error rate is 44.1% for predicting Down and is actually Up.

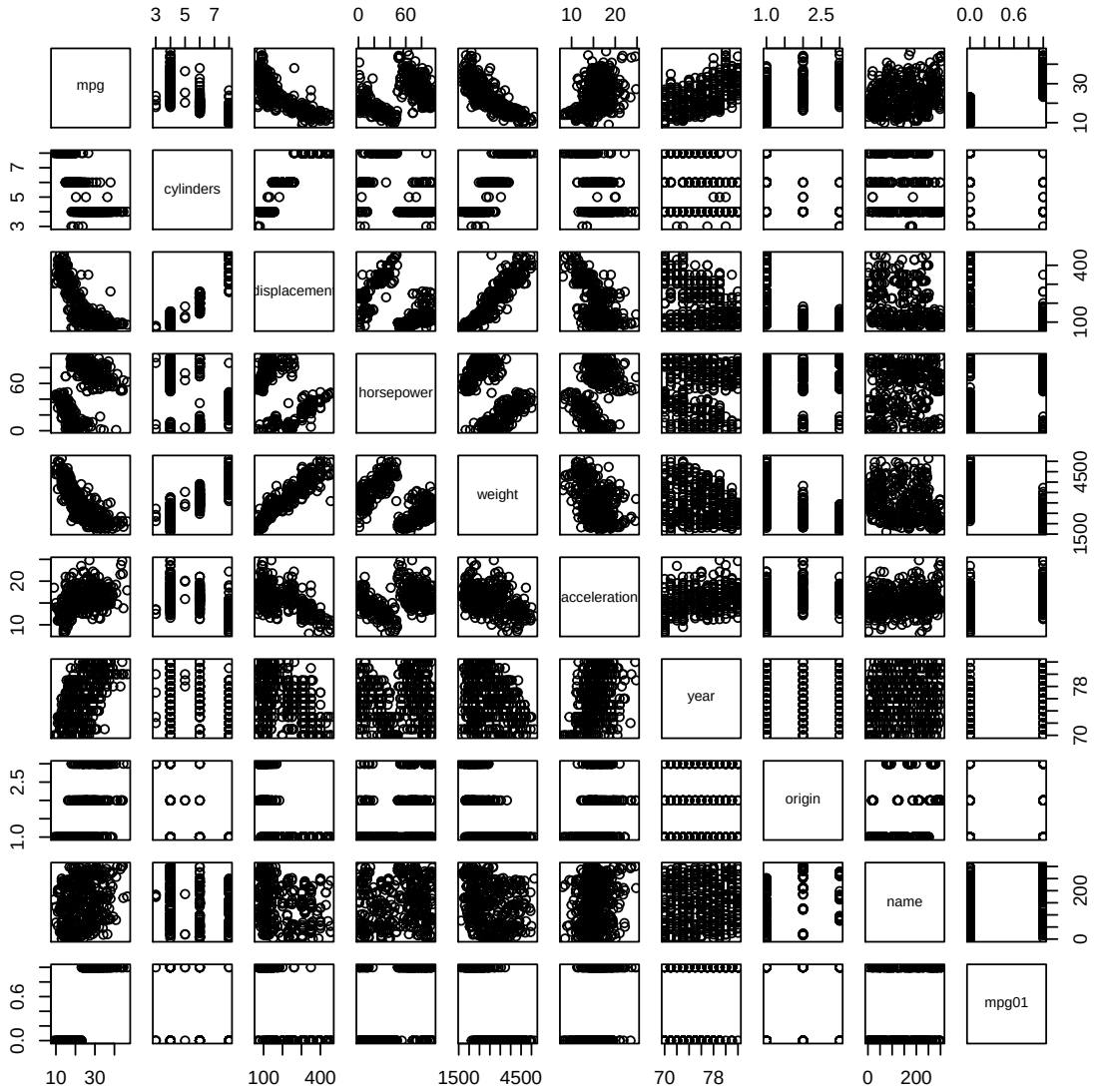
ANSWER FOR 11:

(a)
```
#install.packages("RCurl")
library(RCurl)
## Loading required package:  bitops

dataset<-getURL(
    'https://raw.githubusercontent.com/Jwcrist/BusA/master/homeworks/Assignment%204/Auto.csv',
    ssl.verifypeer=0L, followlocation=1L)
dataset1<-read.csv(text=dataset)
head(dataset1,3)

##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
##                        name mpg01
```

```
## 1 chevrolet chevelle malibu     0
## 2         buick skylark 320     0
## 3        plymouth satellite     0
```

As can be shown above we have created our binary data and a median column

(b)
```
pairs(dataset1)
```



(c)

(d) Do Not Do

(e) Do Not Do

(f)