hw4

```
head(Auto)
```

```
##   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
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

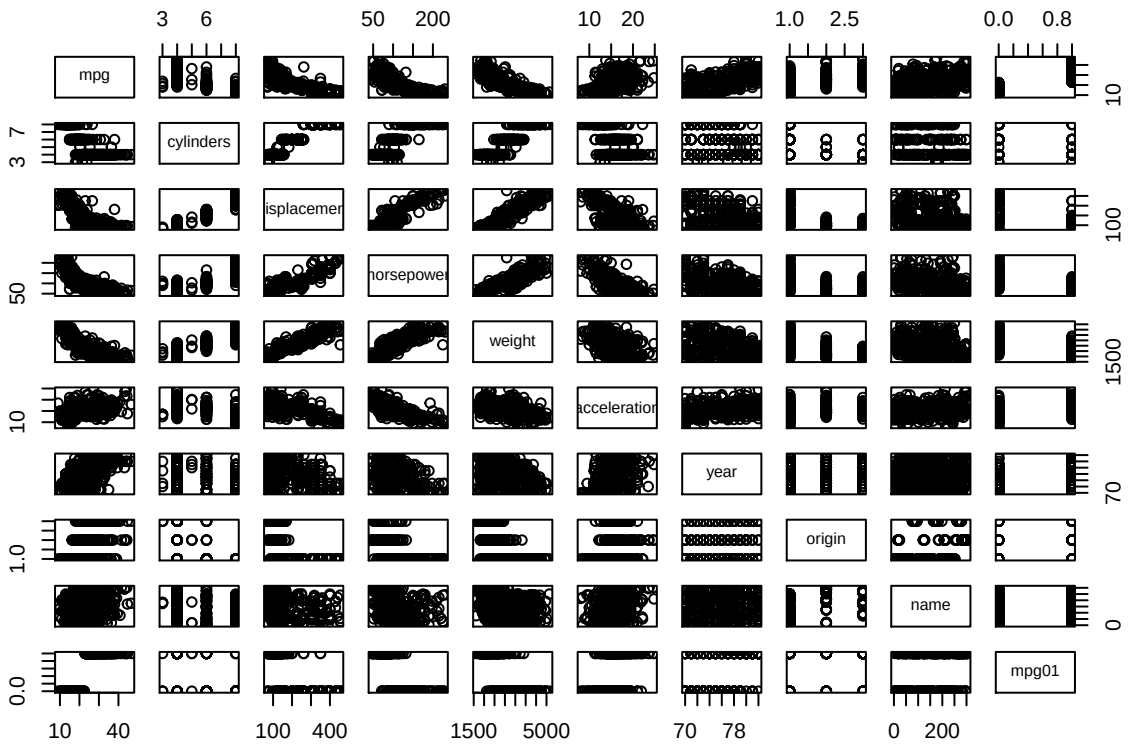
Question 1

```
cars = transform(Auto, mpg01= ifelse(mpg>median(Auto$mpg), 1, 0))
head(cars)
```

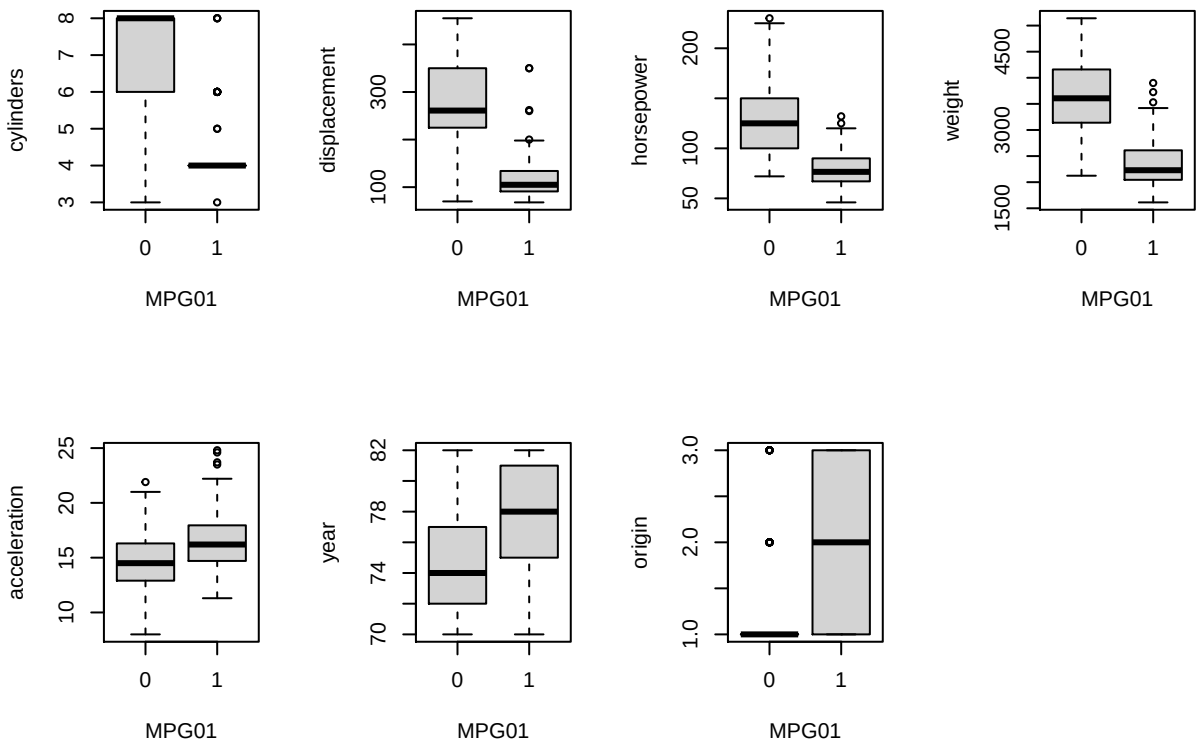
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name mpg01
## 1 chevrolet chevelle malibu     0
## 2         buick skylark 320     0
## 3        plymouth satellite     0
## 4             amc rebel sst     0
## 5               ford torino     0
## 6          ford galaxie 500     0
```

Question 2: From the pairs of scatter plots we can compare an attribute to the MPG column in order to find the most useful feature for predicting mpg01.The cylinders attribute does not give us much information, because it is a discreet variable. Comparing the displacement column to the MPG shows us that there is a negative correlation with the two, which makes sense as a bigger engine will be heavier and require more gas to move it. Similarly with horsepower, the more a car has the lower MPG it will most likely have. Weight also seems to be an important attribute when compared to MPG: there is a clear negative correlation between the two. Acceleration does not seem to correlate with MPG in anyway. From the year column, we can tell that newer cars to tend have slightly better MPG than older cars. Origin and Name attributes are discreet variables and are not very helpful in predicting the MPG of a vehicle.

```
pairs(cars)
```




```
par(mfrow=c(2,4))
for (i in c(2,3,4,5,6,7,8)){
  boxplot(cars[,i] ~ cars$mpg01, xlab = "MPG01", ylab = names(cars)[i])
}
```



Question #3

```
train = cars[1:300,]
test = cars[301:392,]
```

Question #4

```
n_0 <- length(which(train$mpg01 == 0))
n_1 <- length(which(train$mpg01 == 1))
p_0 <- n_0/300
p_1 <- n_1/300
X0 <- train[train$mpg01 == 0, 3:5]
X1 <- train[train$mpg01 == 1, 3:5]
mean_0 <- colMeans(X0)
mean_1 <- colMeans(X1)
s_0 <- cov(X0)
s_1 <- cov(X1)
s_pooled <- ((n_0-1)*s_0 + (n_1-1)*s_1) / (n_0+n_1-2)
s_inv <- solve(s_pooled)
alpha_0 <- -0.5* t(mean_0) %*% s_inv %*% mean_0
alpha_1 <- -0.5* t(mean_1) %*% s_inv %*% mean_1
beta_0 <- s_inv %*% mean_0
beta_1 <- s_inv %*% mean_1

prediction <- c()
d_0vec <- c()
```

```
d_1vec <- c()
label <- c("0", "1")

for (i in 1:nrow(test)){
  y <- t(test[i, 3:5])
  d_0 <- alpha_0 + t(beta_0) %*% y
  d_1 <- alpha_1 + t(beta_1) %*% y
  d_vec <- c(d_0, d_1)
  prediction <- append(prediction, label[which.max( d_vec )])
  d_0vec <- append(d_0vec, d_0)
  d_1vec <- append(d_1vec, d_1)
}

test$prediction <- prediction

error <- length(which(test$mpg01 != test$prediction)) / 92
```

The test error rate is 0.0543.

Question #5

```
true.val = test$mpg01
cars.qda = qda(mpg01 ~ .- name - origin, data = train)
cars.predict.qda = predict(cars.qda, test)
pred.val.qda = cars.predict.qda$class
test.error.QDA = mean(pred.val.qda != true.val)
test.error.QDA
```

```
## [1] 0.06521739
```