Problem Set 6

1.

a.

```
library(MASS)
data(Boston)
```

b.

```
medianValue <- median(Boston$medv)
Boston$priceyHomes <- ifelse(Boston$medv >= medianValue, 1,0)
```

c.

```
doBy::summaryBy(crim + zn + indus + chas + dis ~ priceyHomes, data = Boston )
```

```
##   priceyHomes crim.mean   zn.mean indus.mean  chas.mean dis.mean
## 1           0 6.3879887  3.828685  14.387331 0.03585657 3.151489
## 2           1 0.8825795 18.780392   7.937216 0.10196078 4.428501
```

It is evidently clear that when homes are above the median housing price that the proportion of residential land zoned over 25,000 chas, and dis all increase. There is an increased association with the crime and housing prices that are below the median value. It has a mean of 6.38 where as the mean of crim in housing with median value above the mean is 0.8.

d.

```
set.seed(1861)
trainSize <- 0.5
trainInd <- sample(1:nrow(Boston), size = floor(nrow(Boston) * trainSize))
BostonTrain <- Boston[trainInd, ]
BostonTest <- Boston[-trainInd, ]
```

e.

A training set is our large collection of data that is used to compare to our testing set. Our testing set is what we use to compare to our training set. To avoid bias we may attempt to randomize our selection used for our testing set.

f.

```r
bostonLog <-glm(priceyHomes ~ crim + zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + bl

bostonLog
```

```
##
## Call:  glm(formula = priceyHomes ~ crim + zn + indus + chas + nox +
##     rm + age + dis + rad + tax + ptratio + black + lstat, family = binomial,
##     data = BostonTrain)
##
## Coefficients:
## (Intercept)         crim           zn        indus         chas
##   19.231957    -0.158387     0.016585     0.018933     1.149338
##         nox           rm          age          dis          rad
##   -5.254870     0.855528    -0.031419    -0.768119     0.345071
##         tax      ptratio        black        lstat
##   -0.014330    -0.517315     0.001156    -0.370261
##
## Degrees of Freedom: 252 Total (i.e. Null);  239 Residual
## Null Deviance:       348.3
## Residual Deviance: 126.7     AIC: 154.7
```

g.

```r
exp(bostonLog$coefficients)
```

```
##  (Intercept)         crim           zn        indus         chas
## 2.250778e+08 8.535191e-01 1.016723e+00 1.019113e+00 3.156102e+00
##          nox           rm          age          dis          rad
## 5.222023e-03 2.352617e+00 9.690698e-01 4.638849e-01 1.412091e+00
##          tax      ptratio        black        lstat
## 9.857723e-01 5.961191e-01 1.001157e+00 6.905538e-01
```

There is .8535 times more crime in non pricey homes than pricey homes, there is also .000522 times more nox in non pricey homes than pricey homes. Pricey homes are 3.156 times more likely to be closer to the Charles River than non pricey homes. There are 1.00157 more African Americans in non pricey homes than pricey homes.

2

a.

```r
scoresTest <- predict(bostonLog, newdata = BostonTest, type = "response")
scoresTrain <- predict(bostonLog, newdata = BostonTrain, type = "response")
```

b.

scoresTest

scoresTrain

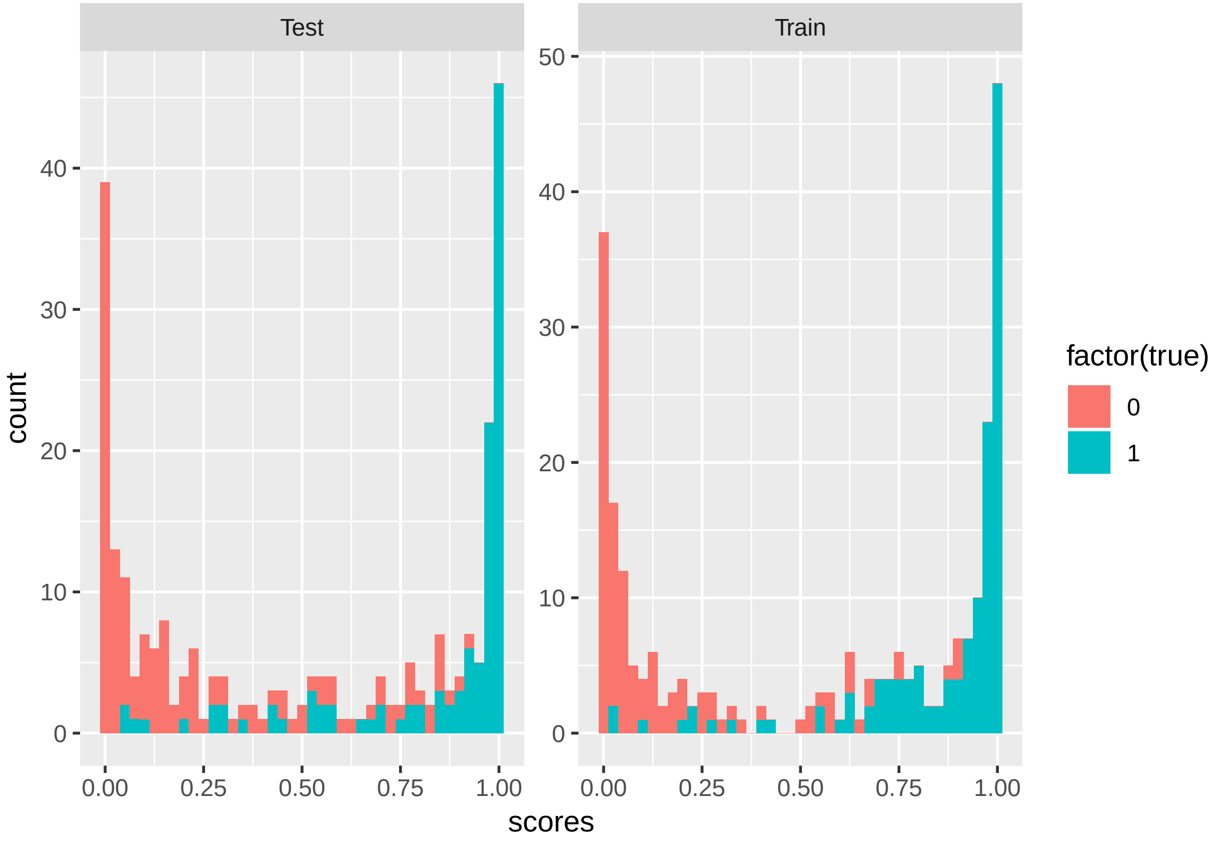
c.

```
scoresDF <- data.frame(scores = c(scoresTest, scoresTrain), type = c(rep("Test", times = length(scoresT
doBy::summaryBy(scores~ type, data = scoresDF,  fun = c(mean,sd))
```

```
##    type scores.FUN1
## 1  Test   0.5118312
## 2 Train   0.5494071
```

d.

```
library(ggplot2)
ggplot(scoresDF) + geom_histogram(aes ( x = scores, fill = factor(true) ), binwidth = 0.025) + facet_wr
```

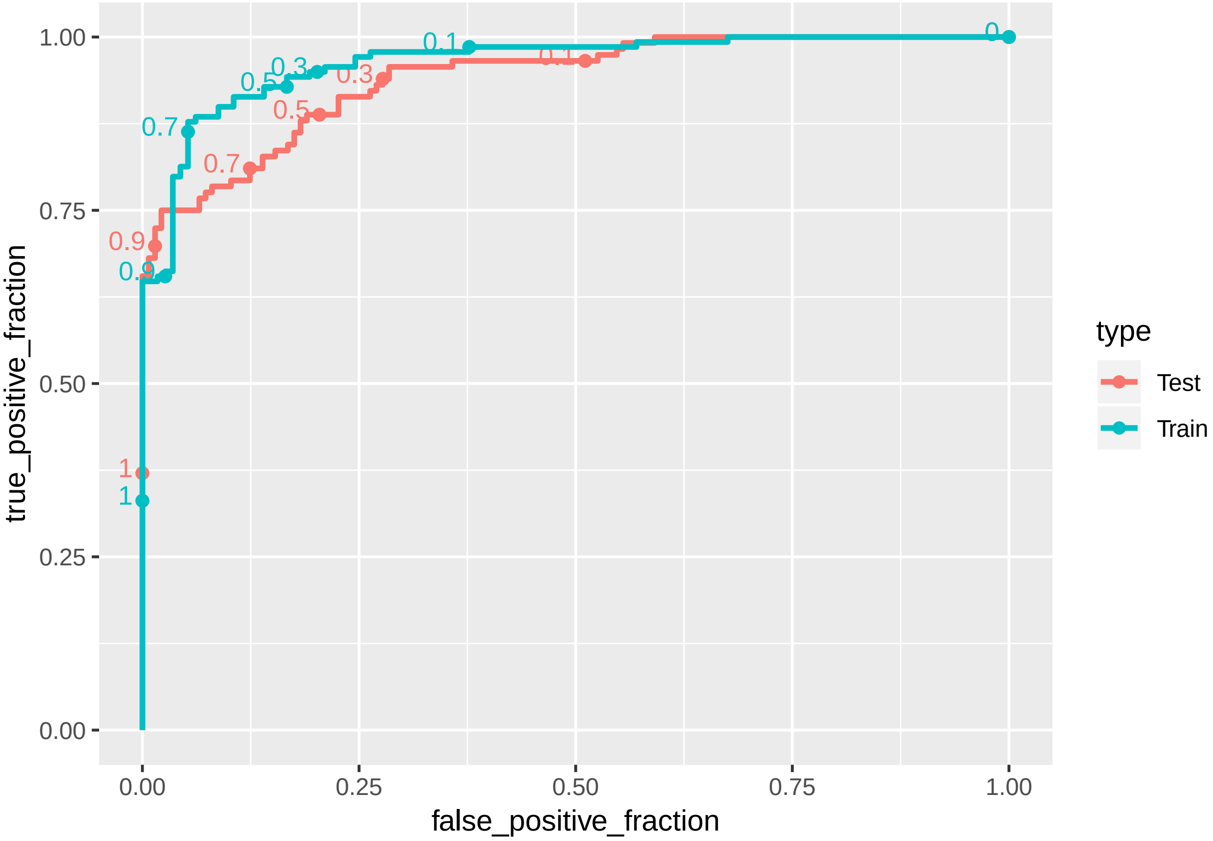


#e.

in the test vs the train, it appears to me that there is less overlap in the test graph. The train data set appears to have more bulk, if that makes sense. Especially around .75 to 1.00. Thats because there is more data present in the train data set. It is important to note though that they both look very identical.

f.

```
#install.packages('plotROC')
library("plotROC")
ggplot(scoresDF, aes(m = scores, d = true, color = type)) + geom_roc(show.legend = TRUE, labelsize = 3.5
```



An roc curve is a graphical plot that graphs the diagnostic capability of a binary classifier system as its discrimination threshold is varied. The aread under the curve is used to measure the usefulness of a test in general. We use ROC curves to visualize the performance of a multiple classification problem. In regards to our graph we can see that our test set did a pretty good job predicting the train set