Homework 2 Part 2

Will Scheib

9/22/2021

6

```
Auto$hilow <- ifelse(Auto$mpg > median(Auto$mpg), 1, 0)
```

a

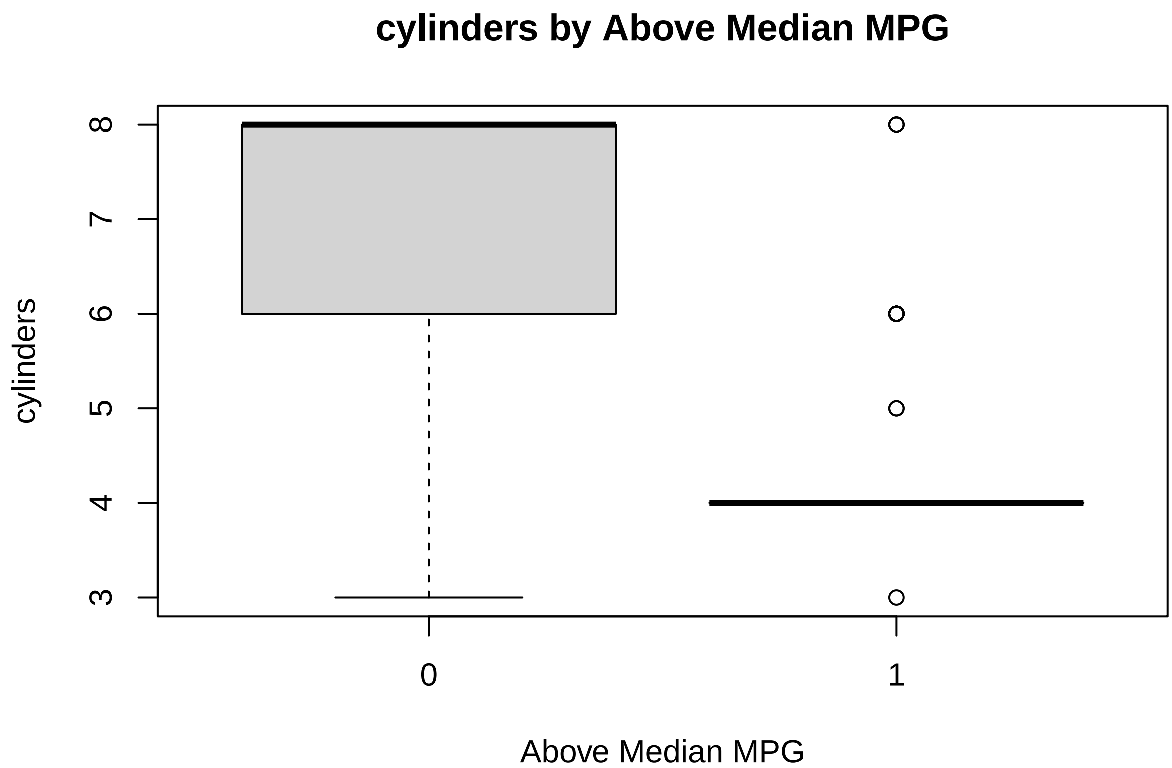
We should remove mpg because it was used to create hilow and thus would break the model.

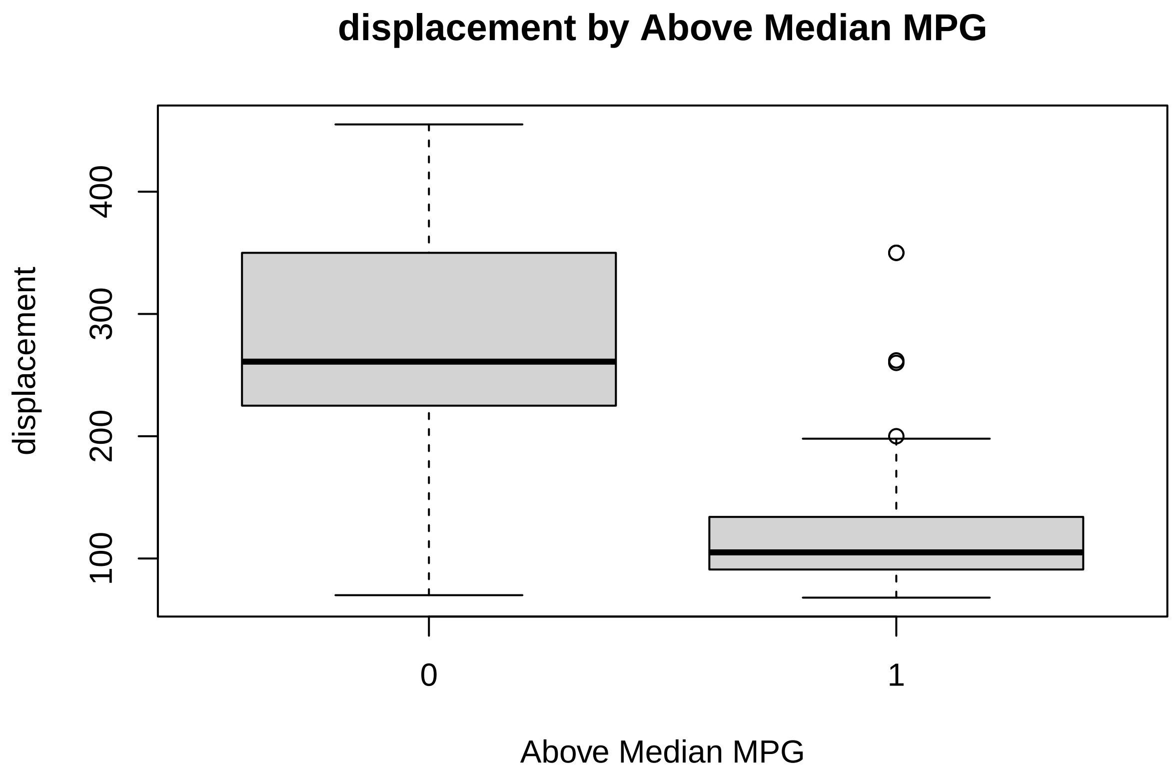
b

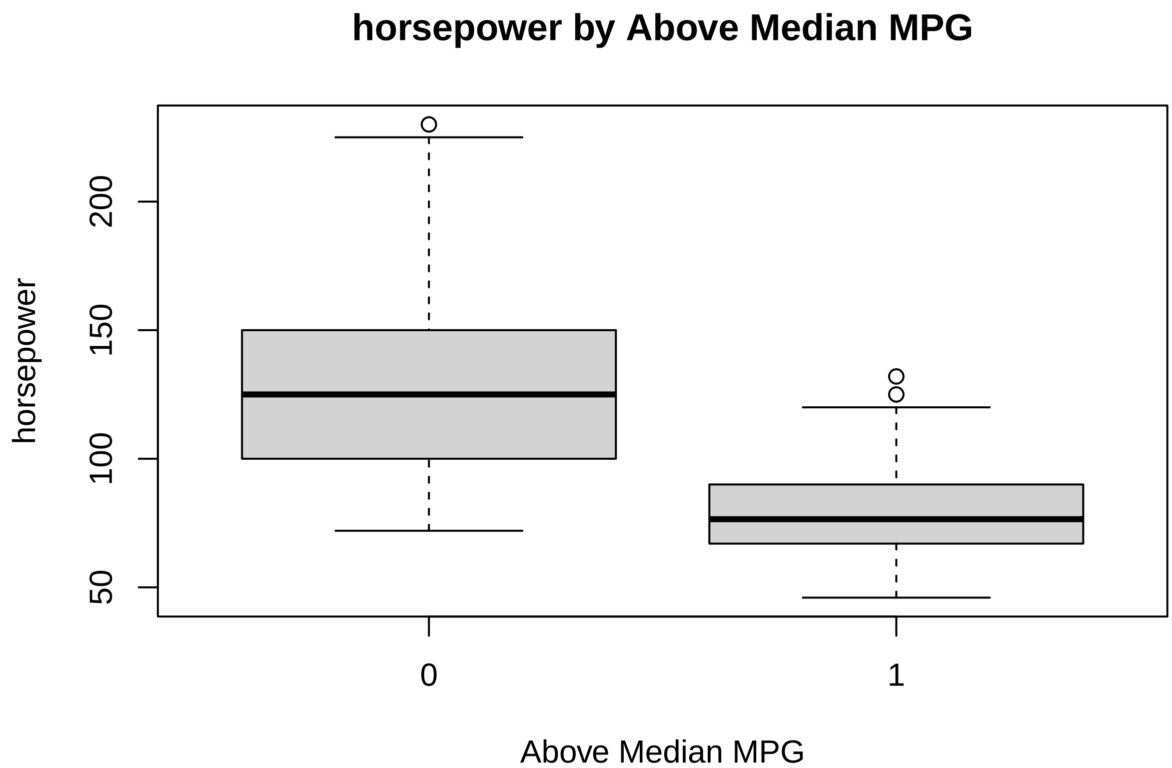
Yes, because the assumptions are less important when classifying.

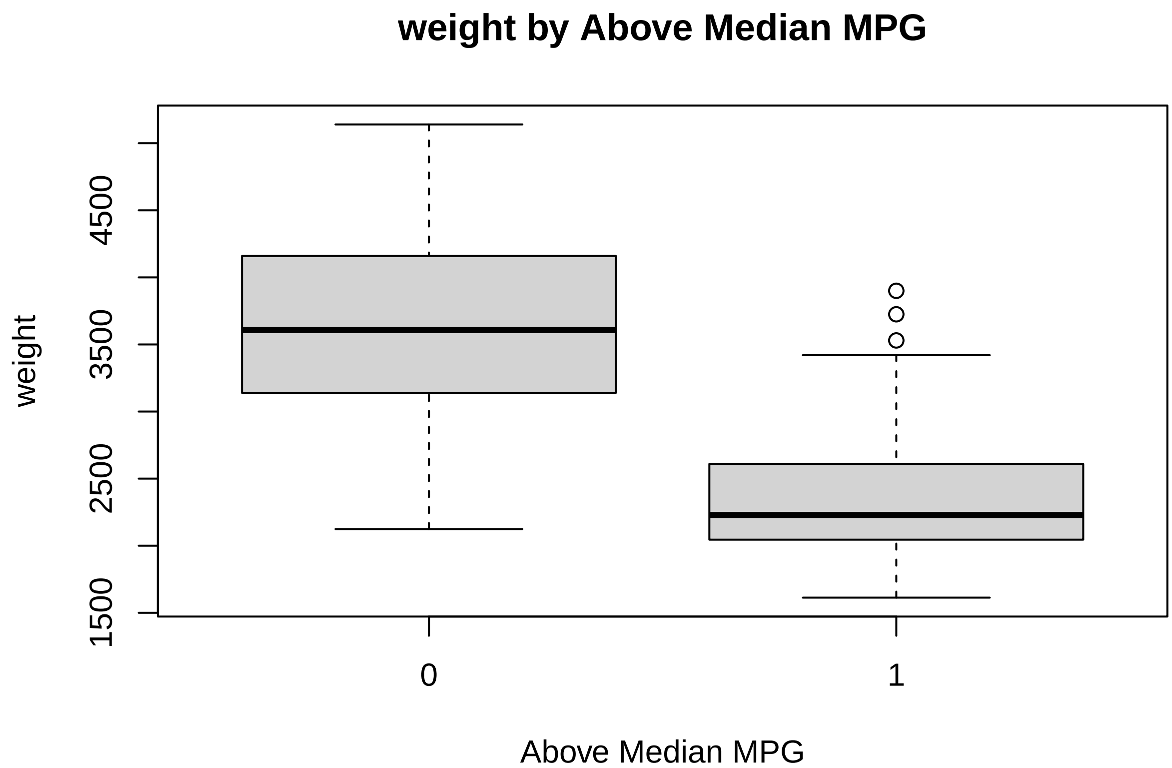
c

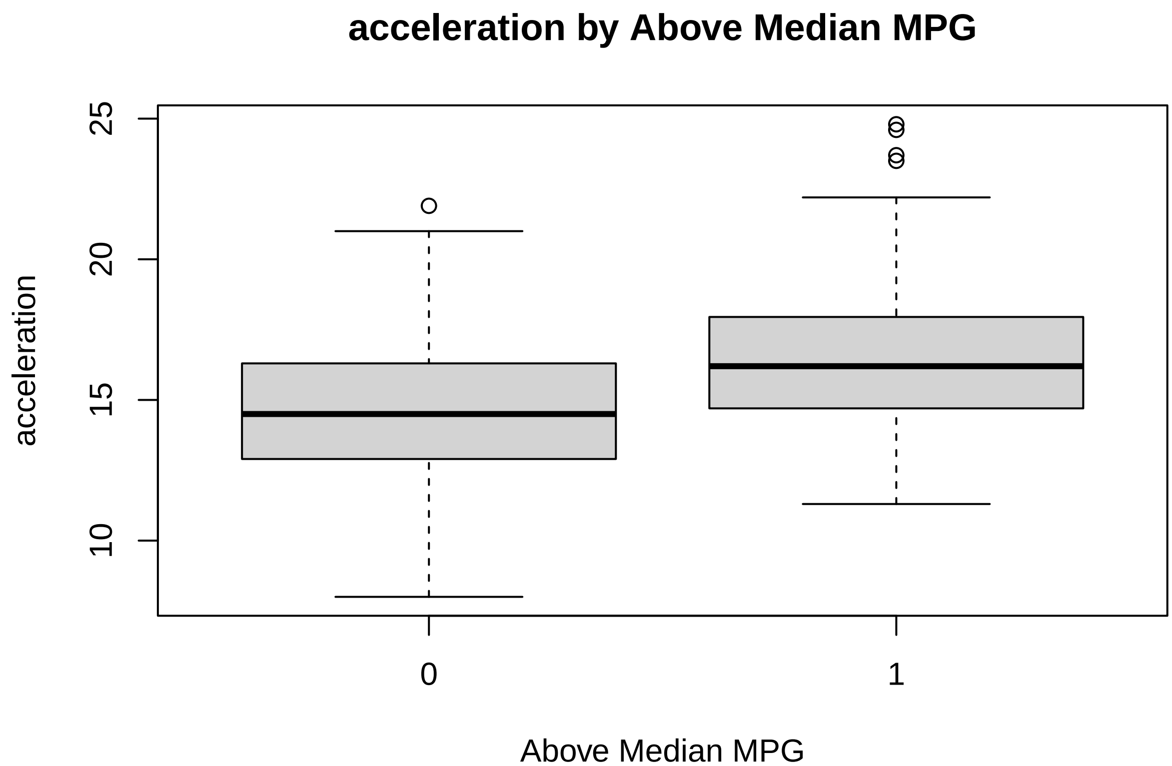
```
for(col in c("cylinders","displacement","horsepower","weight","acceleration","year")){
  is.list(col)
  boxplot(
    Auto[,col]~Auto$hilow,
    xlab="Above Median MPG",
    ylab=col,
    main=paste(col,"by Above Median MPG")
  )
}
```

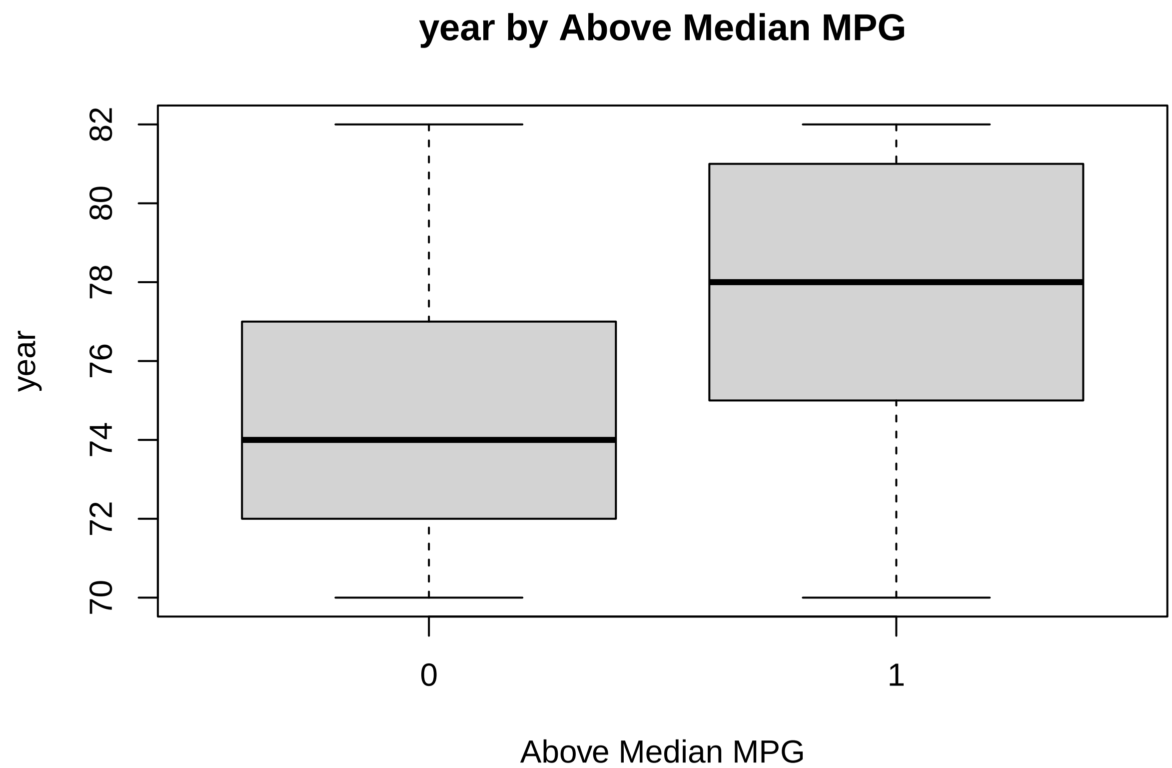

cylinders by Above Median MPG
cylinders
8
7
6
5
4
3
0
1
Above Median MPG


displacement by Above Median MPG
displacement
100
200
300
400
0
1
Above Median MPG


horsepower by Above Median MPG
horsepower
50
100
150
200
0
1
Above Median MPG


weight by Above Median MPG
weight
1500
2500
3500
4500
0
1
Above Median MPG


acceleration by Above Median MPG
acceleration
25
20
15
10
0
1
Above Median MPG



All predictors except acceleration and maybe year look to have significantly differing distributions depending on hilow.

d

```r
set.seed(99)

lda_v_qda <- function(trial){
  if(trial%%1000 == 0){print(trial)}

  # i
  sample.data <- sample(nrow(Auto),floor(.50*nrow(Auto)))
  train <- Auto[sample.data,]
  test <- Auto[-sample.data,]

  # ii
  lda.hilow <- lda(
    hilow ~ cylinders + displacement + horsepower + weight + acceleration + year,
    data=train
  )

  # iii
  lda.test <- predict(lda.hilow,test)
  conf_mat_lda <- table(test$hilow,lda.test$class)
  error_rate_lda <- (conf_mat_lda["0","1"] + conf_mat_lda["1","0"]) / sum(conf_mat_lda)

  # iv
  qda.hilow <- qda(
    hilow ~ cylinders + displacement + horsepower + weight + acceleration + year,
    data=train
  )
  qda.test <- predict(qda.hilow,test)
  conf_mat_qda <- table(test$hilow,qda.test$class)
  error_rate_qda <- (conf_mat_qda["0","1"] + conf_mat_qda["1","0"]) / sum(conf_mat_qda)

  c(error_rate_lda, error_rate_qda)
}

results <- sapply(1:10000, lda_v_qda)
```

```
## [1] 1000
## [1] 2000
## [1] 3000
## [1] 4000
## [1] 5000
## [1] 6000
## [1] 7000
## [1] 8000
## [1] 9000
## [1] 10000
```

e

```r
mean_lda <- mean(results[1,])
mean_qda <- mean(results[2,])
```

```r
c(mean_lda=mean_lda, mean_qda=mean_qda)
```

```
##   mean_lda   mean_qda
## 0.09524337 0.10031224
```

The LDA appears to work marginally better than the QDA over the 10,000 training and test splits.