PRED422 Assignment 1

Artur Mrozowski

April 2, 2017

PREDICT 422 Assignment 1

Excercise 8 (ISLR Section 2.4.

This exercise relates to the College data set, which can be found inthe file College.csv. It contains a number of variables for 777 differentuniversities and colleges in the US.

8(a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory setto the correct location for the data.

```
college=read.csv("C:\\playground\\Predict422\\R\\week1\\college.csv")
summary(college)
```

8(b) Look at the data using the fix() function.

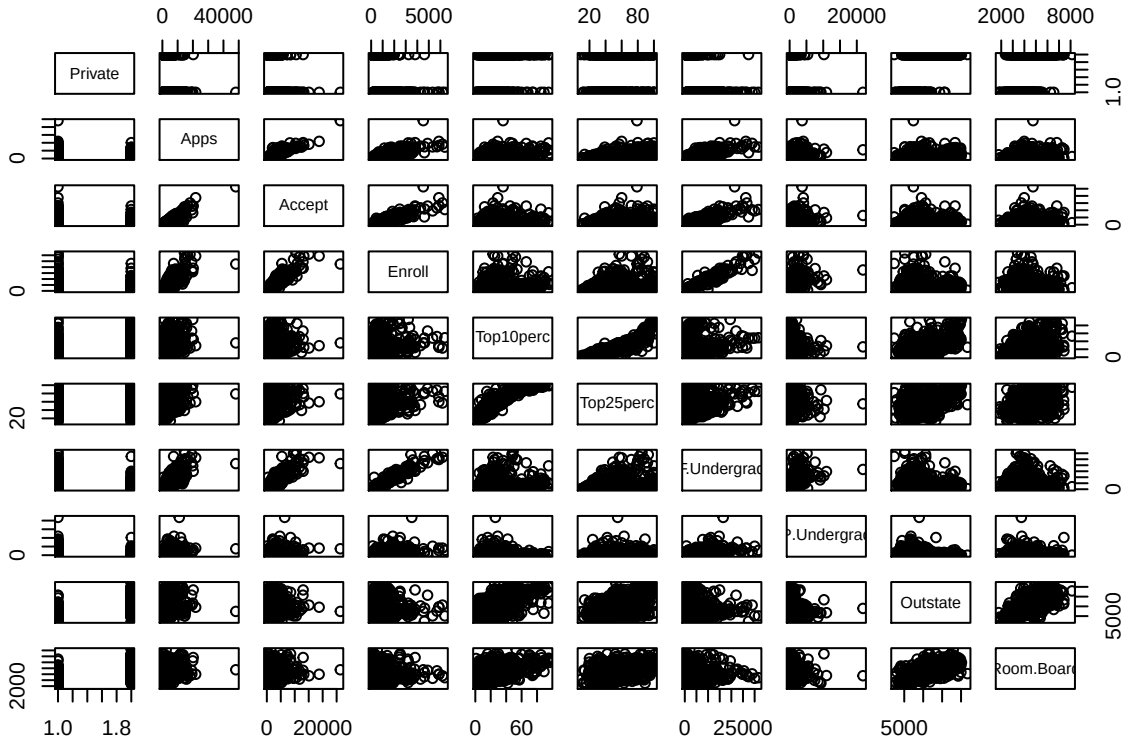
```
fix(college)
rownames(college) = college[,1]
fix(college)
college = college[,-1]
fix(college)
```

8c i. Use the summary() function to produce a numerical summaryof the variables in the data set.

```
summary(college)
```

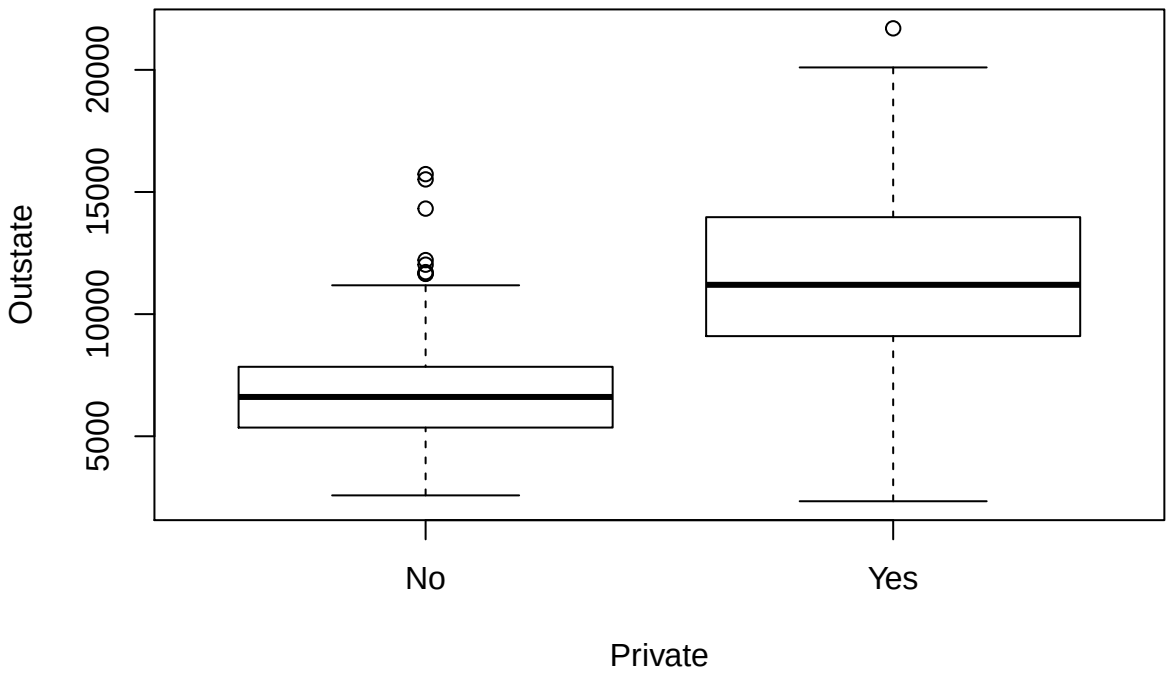
ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.

```
pairs(college[,1:10])
```



iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

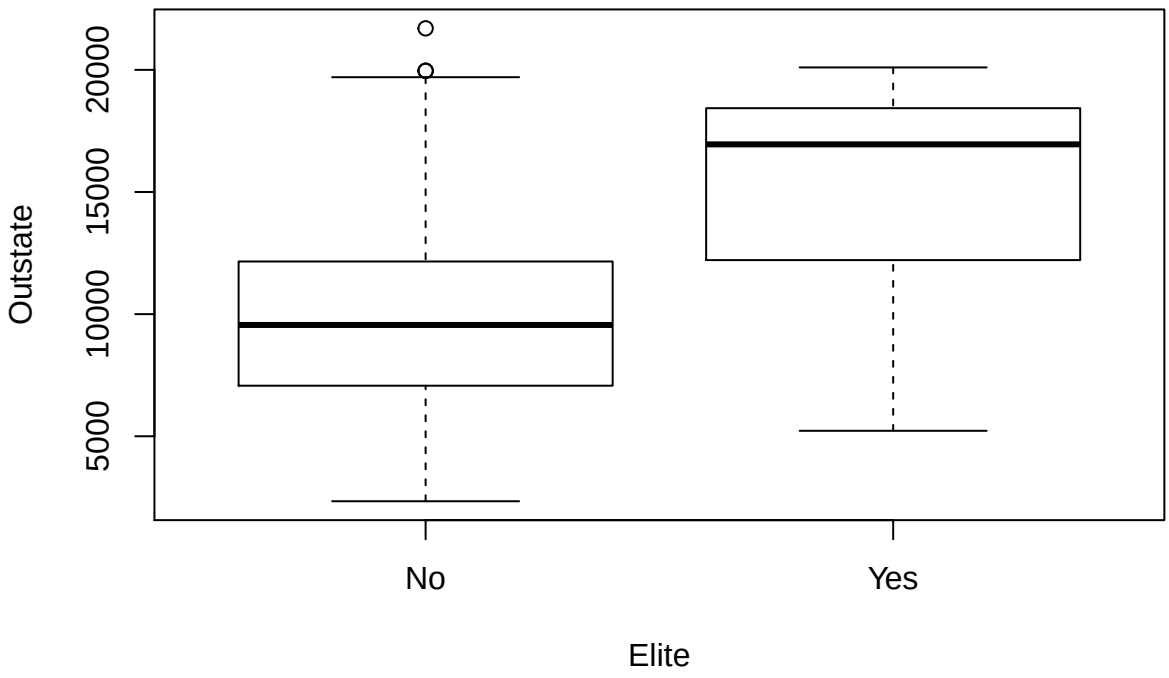
```
plot(college$Private,college$Outstate,xlab="Private",ylab="Outstate")
```



iv.Create a new qualitative variable, called Elite, by binning the Top10perc variable.

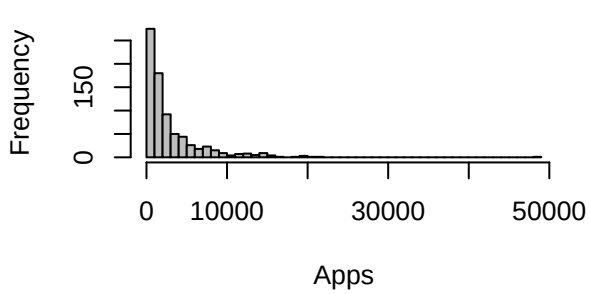
```
# 8 (c) iv.
Elite = rep("No",nrow(college))
Elite[college$Top10perc > 50] = "Yes"
Elite = as.factor(Elite)
college = data.frame(college,Elite)
summary(college$Elite)
```

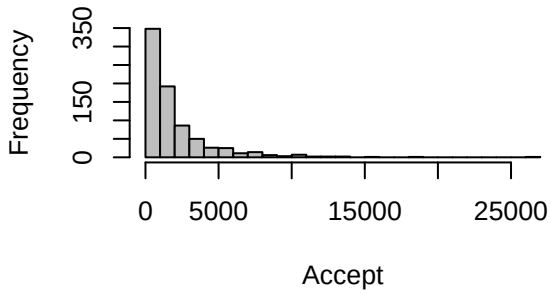
```
plot(college$Elite,college$Outstate,xlab="Elite",ylab="Outstate")
```

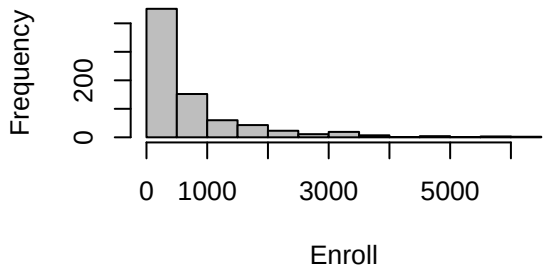


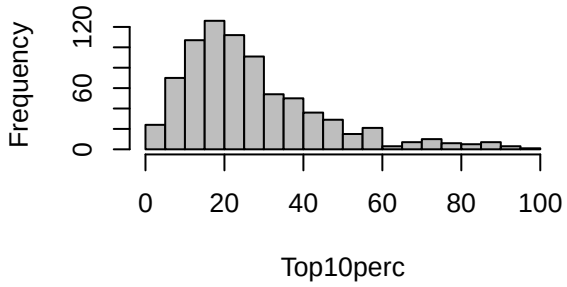
v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.

```
# 8 (c) v.
par(mfrow=c(2,2))
hist(college$Apps,breaks=50,col="gray",xlab="Apps",main="")
hist(college$Accept,breaks=20,col="gray",xlab="Accept",main="")
hist(college$Enroll,breaks=15,col="gray",xlab="Enroll",main="")
hist(college$Top10perc,breaks=15,col="gray",xlab="Top10perc",main="")
```









Exercise 9 (ISLR Section 2.4) This exercise involves the Auto data set studied in the lab

```
inPath = file.path("C:","playground","Predict422",
                   "R","week1")
# Load data and remove missing values per the lab (Section 2.3.4)
Auto = read.table(file.path(inPath,"Auto.data"),header=TRUE,na.strings="?")
dim(Auto)
Auto = na.omit(Auto)
dim(Auto)
```

a. Which of the predictors are quantitative, and which are qualitative?

```
# 9 (a)
str(Auto)
summary(Auto)
```

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
# 9 (b)
#sapply(Auto[,1:7],range)
for (ii in 1:7)
{
  rng = range(Auto[,ii])
  print(paste(names(Auto)[ii],": ",rng[1]," to ",rng[2],sep=""))
}
```

(c) What is the mean and standard deviation of each quantitative predictor?

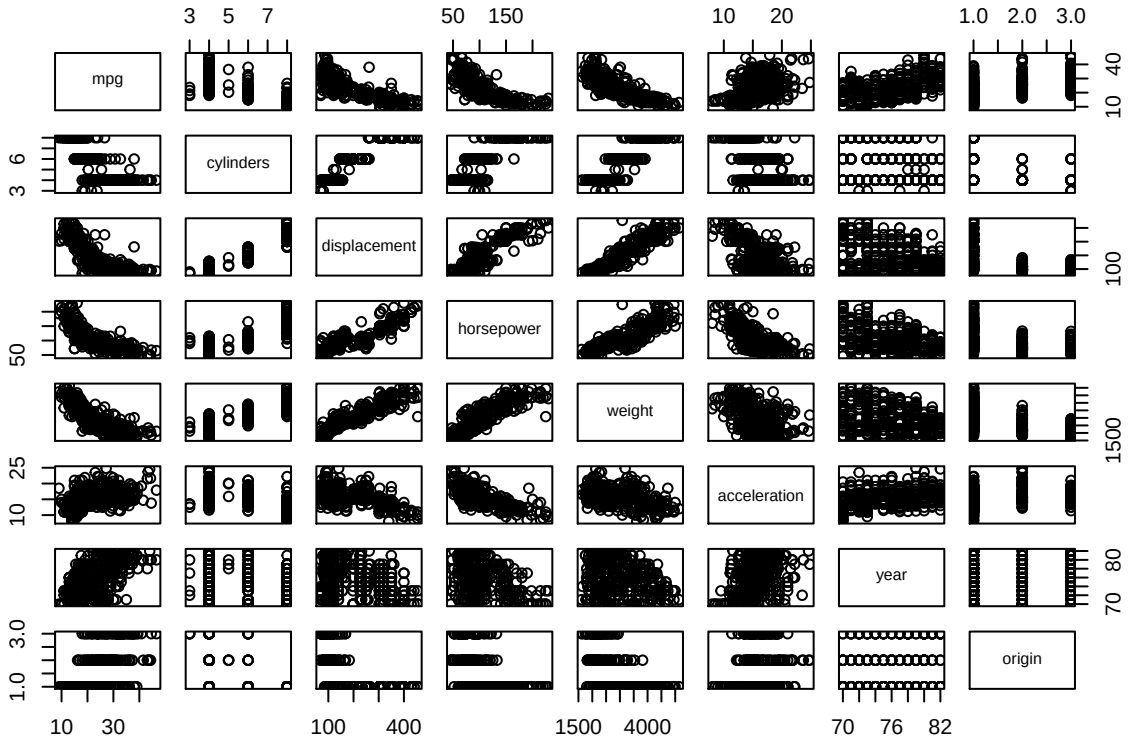
```
# 9 (c)
#apply(Auto[,1:7],2,mean)
#apply(Auto[,1:7],2,sd)
sapply(Auto[,1:7],mean)
sapply(Auto[,1:7],sd)
```

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

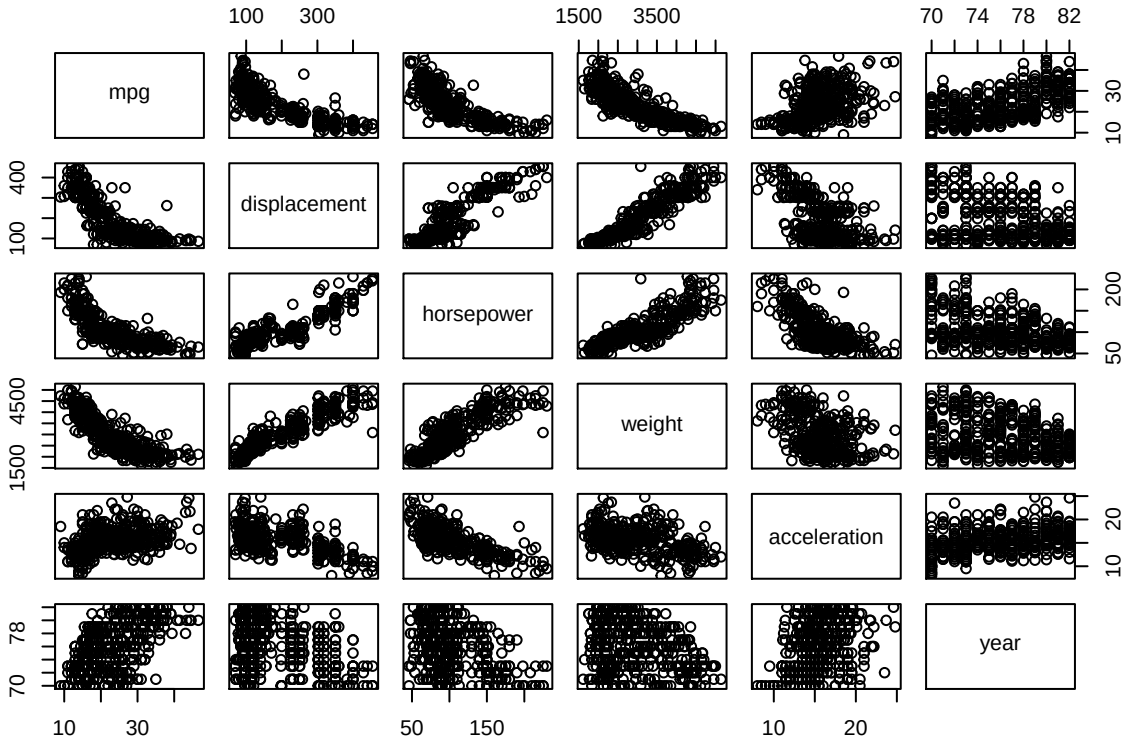
```
# 9 (d) Now remove the 10th through 85th observations.
AutoSubset = Auto[-(10:85),]
sapply(AutoSubset[,1:7],range)
sapply(AutoSubset[,1:7],mean)
sapply(AutoSubset[,1:7],sd)
```

(e) Using the full data set, investigate the predictors graphically

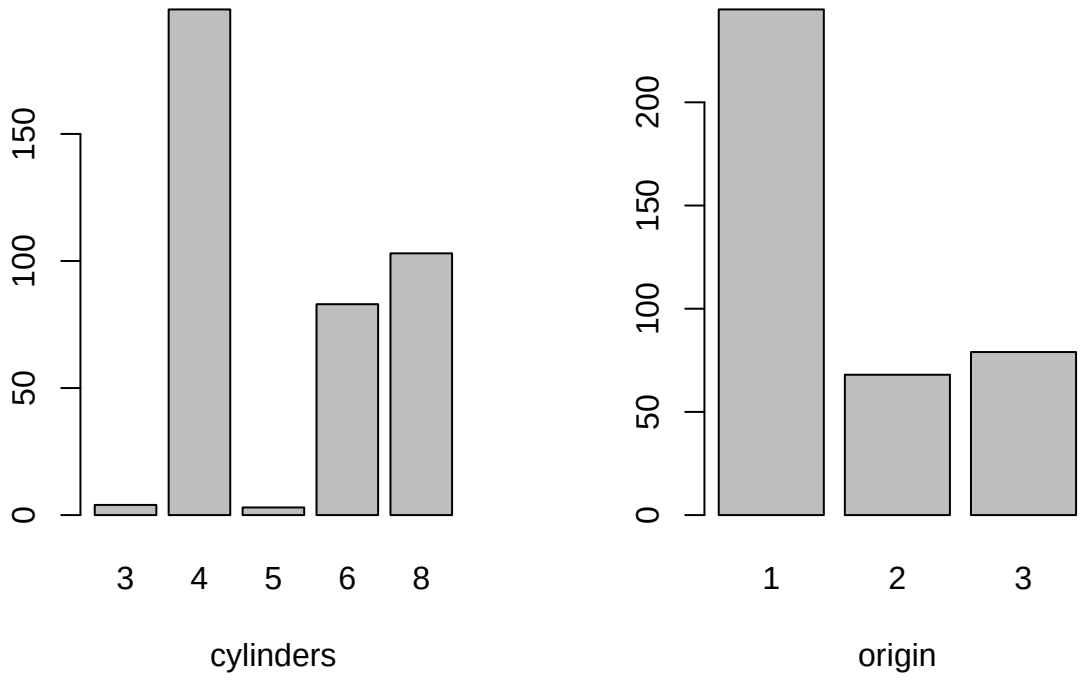
```
pairs(Auto[,1:8])
```




```
Auto$cylinders = as.factor(Auto$cylinders)
Auto$origin = as.factor(Auto$origin)
pairs(Auto[,c(1,3:7)])
```



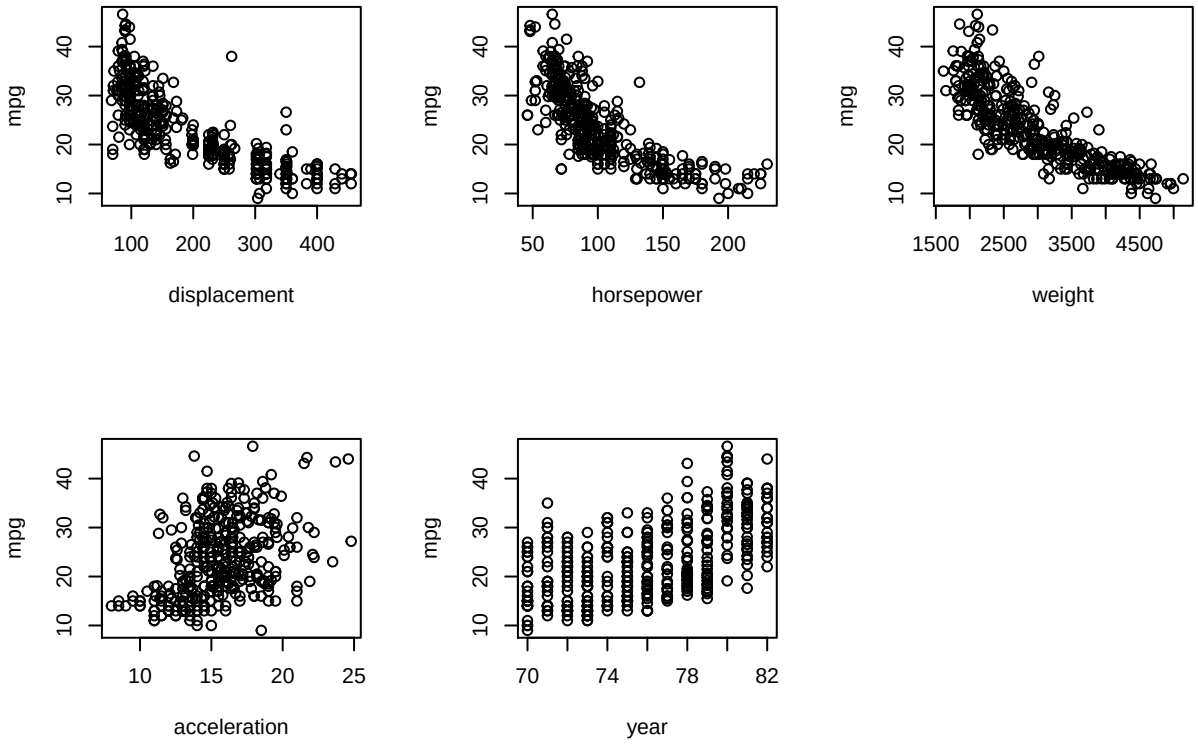

```
par(mfrow=c(1,2))
barplot(table(Auto$cylinders),xlab="cylinders")
barplot(table(Auto$origin),xlab="origin")
```



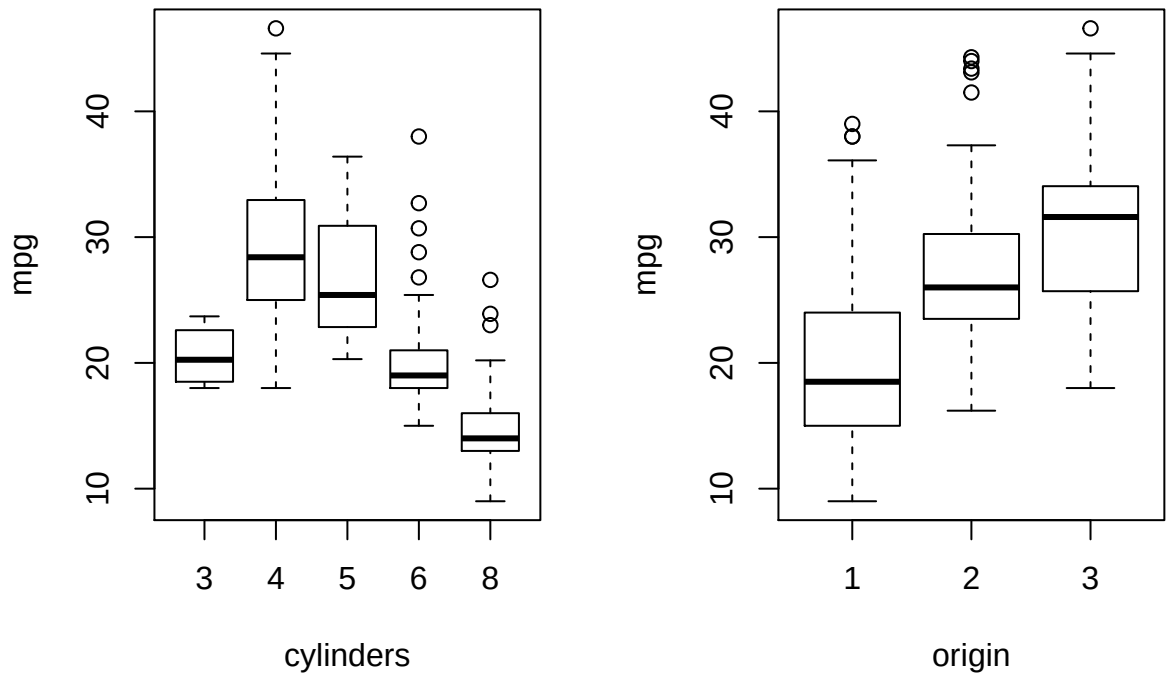
(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg?

```
par(mfrow=c(2,3))
plot(Auto$displacement,Auto$mpg,xlab="displacement",ylab="mpg")
# more horsepower correlates with lower mpg
plot(Auto$horsepower,Auto$mpg,xlab="horsepower",ylab="mpg")
#heavier weight correlates with lower mpg
plot(Auto$weight,Auto$mpg,xlab="weight",ylab="mpg")
plot(Auto$acceleration,Auto$mpg,xlab="acceleration",ylab="mpg")
plot(Auto$year,Auto$mpg,xlab="year",ylab="mpg")
#mpg increases over time

par(mfrow=c(1,2))
```




```
plot(Auto$cylinders,Auto$mpg,xlab="cylinders",ylab="mpg")
plot(Auto$origin,Auto$mpg,xlab="origin",ylab="mpg")
```



More horsepower correlates with lower mpg. Heavier weight correlates with lower mpg. Mpg increases over time as showed above.