Analytics 512: Homework # 1

Arif Ali

January 28th, 2016

Exercise 2.4 #2

Part A

This is a regression problem because the objective is to figure out how the different explanatory variables affect CEO salary. Salary is not binned into categories, but rather is a number.

n is the 500 firms since they are the observations and *p* is profit, number of employees, and industry.

Part B

Since the response variable is binary, this is a classification problem.

n is the observation of the 20 previous products and *p* is price charged for the product, marketing budget, competition price, and ten other variables (not the success/failure variable).

Part C

This is a regression problem because the percent change is trying to determine to exact change in percentage.

n is the number of weeks during 2012 and *p* the % change in the dollar, the % change in the British market, and the % change in the German market.

Exercise 2.4 #4

Part A

1. **Predict**ing whether a person has cancer or not.
2.
3.

Part B

1.
2.
3.

Part C

1.
2.
3.

Exercise 2.4 #7

Part A

Observation	Euclidean Distance
1	3
2	2
3	$\sqrt{10}$
4	$\sqrt{5}$
5	$\sqrt{2}$
6	$\sqrt{3}$

Part B

Green, because observation 5 is the the 1^{st} nearest neighbor and the response for 5 is green.

Part C

Red, 2 out of the 3 closest neighbors has a response of red.

Part D

Small, it would be more flexible whereas a bigger value of K would result in trying to accomodate more points.

Exercise 2.4 #9

Part A

In [2]:
```
library("ISLR")
data(Auto)
Auto = na.omit(Auto)
Auto = Auto[!is.null(Auto),]
summary(Auto)
```

Out[2]:
```
      mpg          cylinders      displacement     horsepower        weight
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140

  acceleration        year           origin                      name
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
 1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
 Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
 Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
 3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
 Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
                                                 (Other)           :365
```

The quantitative predictors: mpg, displacement, horsepower, weight, and acceleration.

The qualitative predictors: cylinders, year, origin, and name.

Part B

In [3]:
```
quantitative = c("mpg", "displacement", "horsepower", "weight", "acceleratio
n")
for(i in quantitative){
    print(i)
    print(range(Auto[, i]))
    }
```

```
[1] "mpg"
[1]  9.0 46.6
[1] "displacement"
[1]  68 455
[1] "horsepower"
[1]  46 230
[1] "weight"
[1] 1613 5140
[1] "acceleration"
[1]  8.0 24.8
```

Part C

In [4]:
```
for(i in quantitative){
    print(i)
    print(mean(Auto[, i]))
    print(sd(Auto[, i]))
    }
```

```
[1] "mpg"
[1] 23.44592
[1] 7.805007
[1] "displacement"
[1] 194.412
[1] 104.644
[1] "horsepower"
[1] 104.4694
[1] 38.49116
[1] "weight"
[1] 2977.584
[1] 849.4026
[1] "acceleration"
[1] 15.54133
[1] 2.758864
```

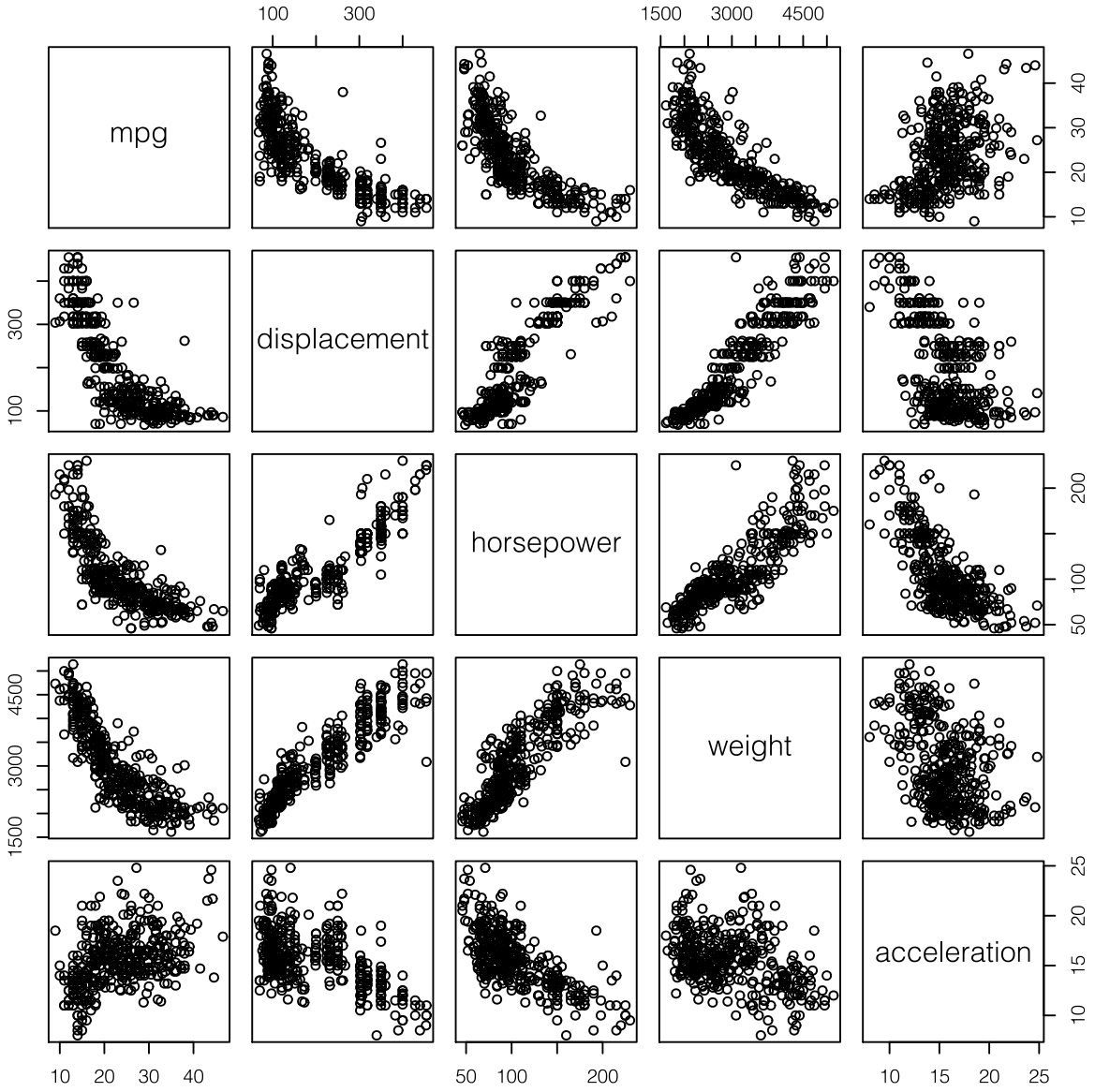
Part D

In [5]:
```
Auto.minus10.to85 = Auto[,-(10:85)]
for(i in quantitative){
    print(i)
    print(range(Auto.minus10.to85[, i]))
    print(mean(Auto.minus10.to85[, i]))
    print(sd(Auto.minus10.to85[, i]))
    }
```

```
[1] "mpg"
[1]  9.0 46.6
[1] 23.44592
[1] 7.805007
[1] "displacement"
[1]  68 455
[1] 194.412
[1] 104.644
[1] "horsepower"
[1]  46 230
[1] 104.4694
[1] 38.49116
[1] "weight"
[1] 1613 5140
[1] 2977.584
[1] 849.4026
[1] "acceleration"
[1]  8.0 24.8
[1] 15.54133
[1] 2.758864
```

Part E

```
In [27]: pairs(Auto[,quantitative])
```



For mpg, there seems to be a negative relation between it compared to displacement, horespower, and weight. The relation betwwen acceleration and mpg is not as well defined but seemingly more positive.

The is a positive relation between displacement compared to horsepower and weight. There seems to be a negative relationship between displacement and acceleration.

Horsepower and weight has a postive relationship and horsepower and acceleration have a negative relationship.

Relation weight and acceleration is less defined, but there seems to be a slightly negative relationship

Part F

The relation between mpg and horsepower seems to fit a line resembling $y = e^{-x}$. Displacement follows this line as well; however, the points in the middle are more spread out compared to mpg vs. horsepower.

Exercise 2.4 #10

Part A

In [2]:
```
library(MASS)
head(Boston)
```

Out[2]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

In [3]:
```
dim(Boston)
names(Boston)
```

Out[3]: 506 14

Out[3]: 'crim' 'zn' 'indus' 'chas' 'nox' 'rm' 'age' 'dis' 'rad' 'tax' 'ptratio' 'black' 'lstat' 'medv'

The rows represents observation in Bostonian suburbs; there are 506 observations. The columns are different attributes for the suburbs; there are 14 columns.

Part B

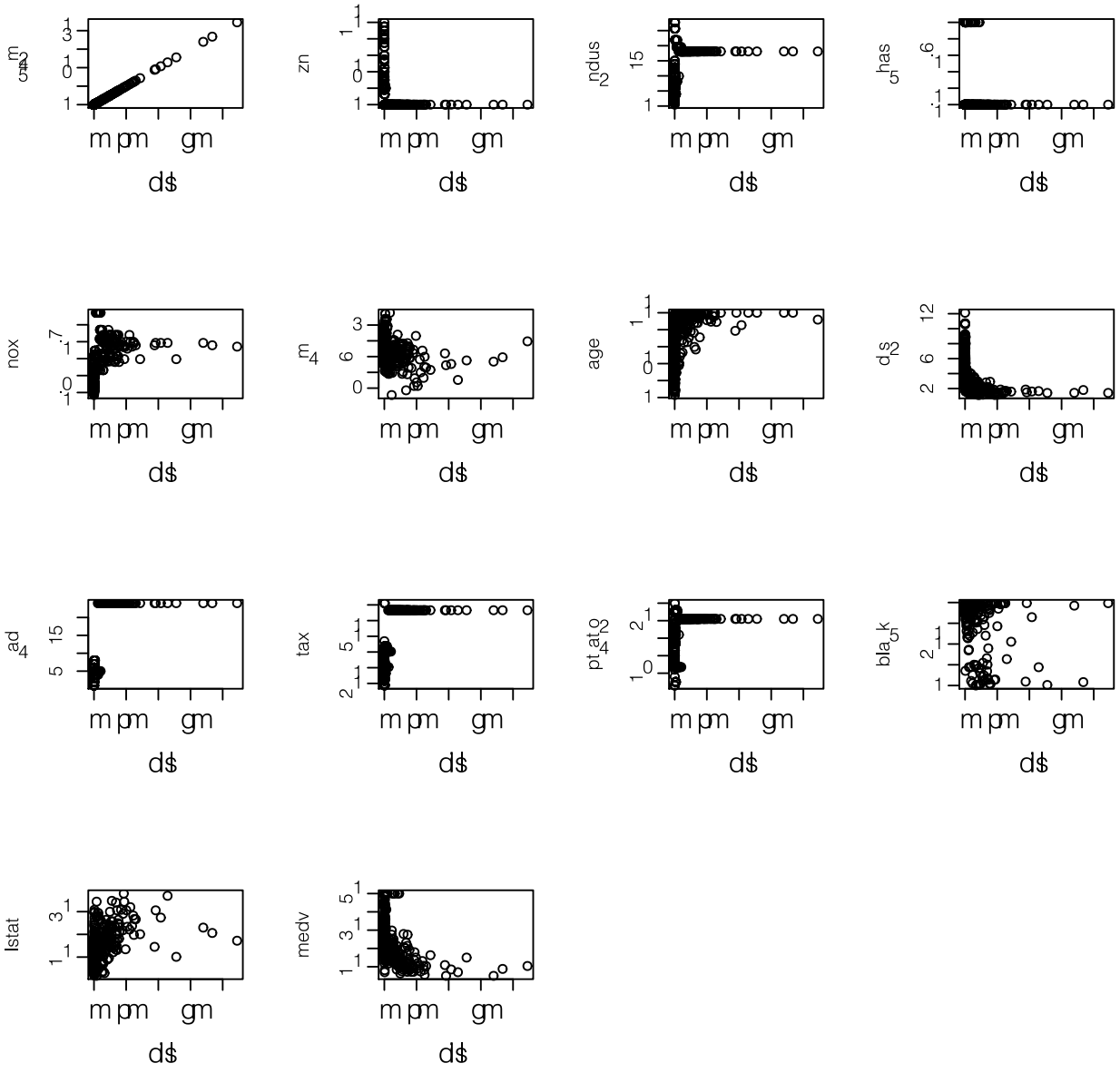
In [35]:
```
jpeg(filename = "ex10pb.jpeg")
pairs(Boston)
dev.off()
```

Out[35]: **pdf:** 2

Please see https://github.com/arifyali/Statistical-Learning-Analytics-512-Spring-2016/blob/master/HW1/ex10pb.jpeg (https://github.com/arifyali/Statistical-Learning-Analytics-512-Spring-2016/blob/master/HW1/ex10pb.jpeg) for plot

Part C

In [30]:
```
par(mfrow = c(4,4))
for(i in names(Boston)){
    plot(Boston$crim, Boston[, i], xlab="crim", ylab = i)
    }
```



It seems that the higher the age, the more crime. It could be that the older houses could be in more rundown neighborhoods; therefore, crime is more rampant. The lower the distance variable, the more crime, which is interesting because it seems counter-intuitive. In areas with higher tax rates, there is more crime. This could be because lower economic classes are taxed higher, and there is a relationship between poverty and crimes.

Part D

In [1]:
```
quantile(Boston$crim, seq(from = 0, to = 1, by = .1))
hist(Boston$crim, breaks = 25)
```

```
Error in quantile(Boston$crim, seq(from = 0, to = 1, by = 0.1)): object 'Bost
on' not found
```

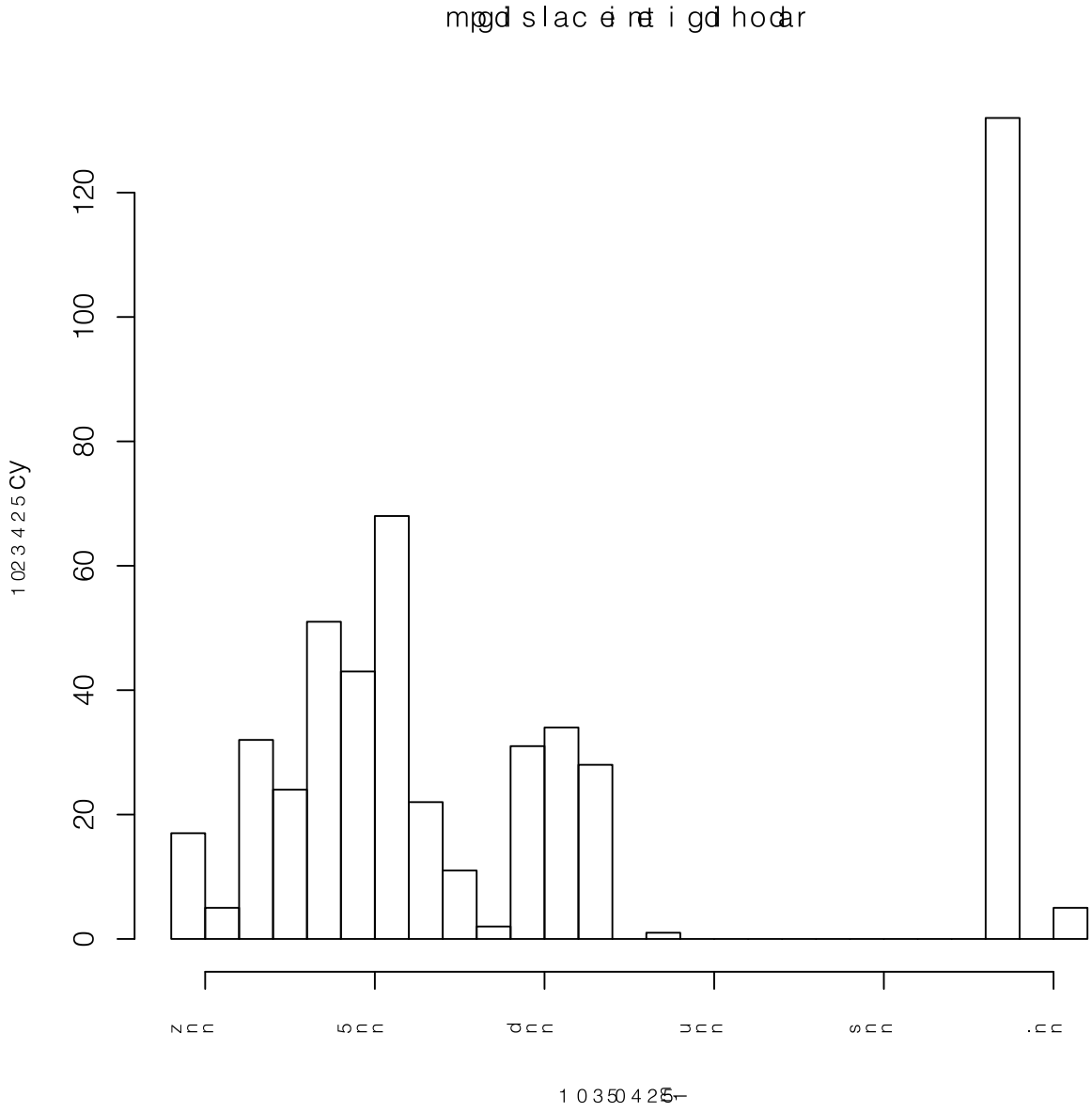
```
Error in hist(Boston$crim, breaks = 25): object 'Boston' not found
```

The crime rate is pretty low throughout boston, with less than 40% of the suburb having a crime rate of greater than one. Based on the quantile, there seems to be a right tail as indicated. Boston seems to be a pretty safe place.

In [15]:
```
quantile(Boston$tax)
hist(Boston$tax, breaks = 25)
```

Out[15]:

0%	187
25%	279
50%	330
75%	666
100%	711

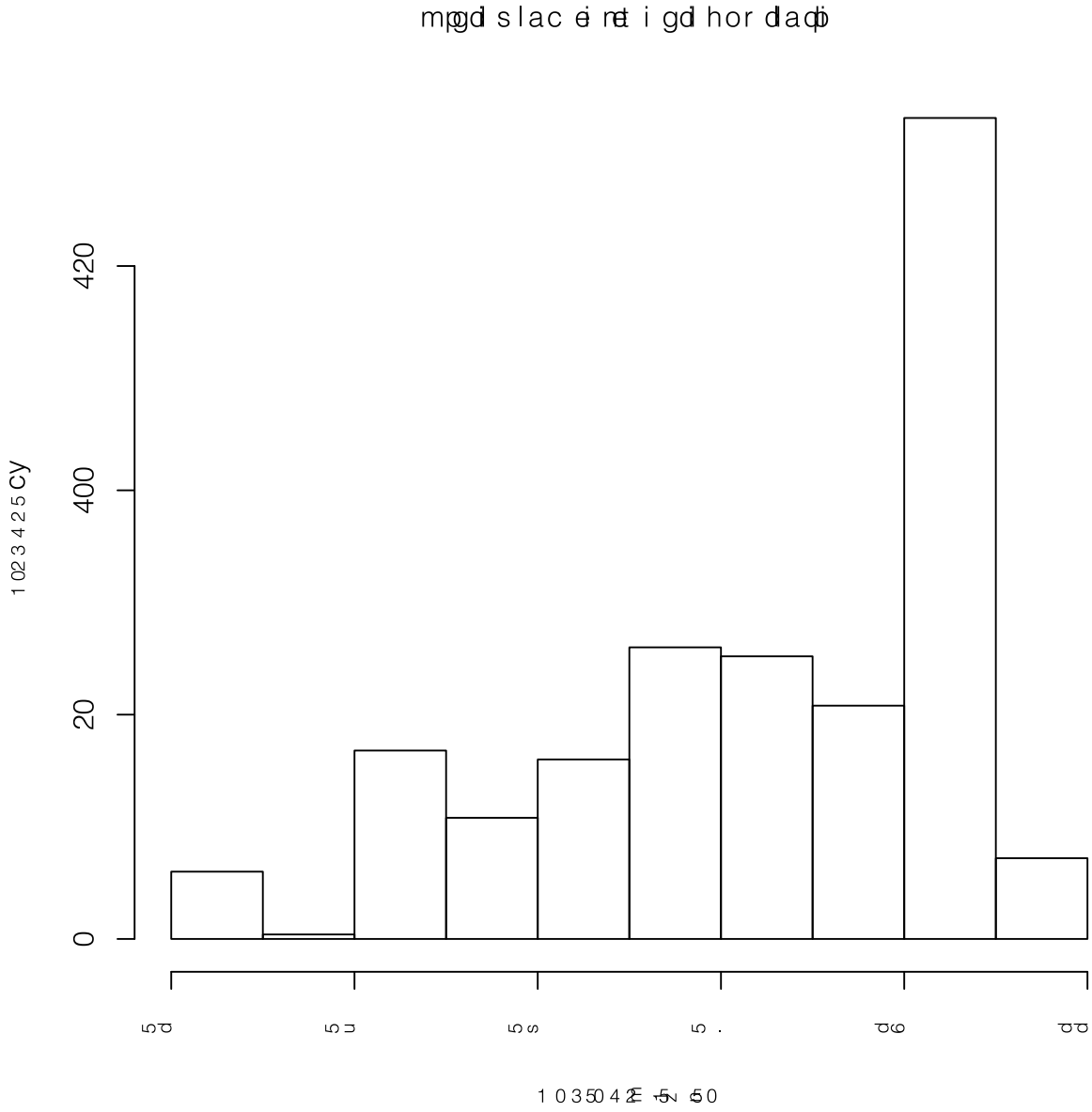


Two of the suburbs have very high Tax rates. I bet that Cambridge is one of those two. There is a significant gasp, which could indicate income inequality.

In [18]:
```
quantile(Boston$ptratio)
hist(Boston$ptratio, breaks = 10)
```

Out[18]:

0%	12.6
25%	17.4
50%	19.05
75%	20.2
100%	22



There are an unusual number of suburbs with a ratio between 20 and 21. However no indication that there is any skew.

Part E

In [6]:
```
sum(Boston$chas == 1)
```

Out[6]: 35

Part F

In [3]:
```
median(Boston$ptratio)
```

Out[3]: 19.05

Part G

In [5]:
```
Boston[Boston$medv == min(Boston$medv),]
```

Out[5]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.9	30.59	5
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24	666	20.2	384.97	22.98	5

Part H

In [23]:
```
sum(Boston$rm>7)
sum(Boston$rm>8)
summary(Boston[Boston$rm>8,])
```

Out[23]: 64

Out[23]: 13

Out[23]:

```
      crim                zn              indus             chas        
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
      nox               rm              age              dis        
 Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
 1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
 Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
 Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
 3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
 Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
      rad              tax           ptratio          black       
 Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
 1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
 Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
 Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
 3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
 Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
     lstat            medv      
 Min.   :2.47   Min.   :21.9  
 1st Qu.:3.32   1st Qu.:41.7  
 Median :4.14   Median :48.3  
 Mean   :4.31   Mean   :44.2  
 3rd Qu.:5.12   3rd Qu.:50.0  
 Max.   :7.44   Max.   :50.0  
```

The crime rate is significiantly lower than overall. Many of these suburbs aren't along the Charles river. Population of lower status is relatively lower