Cap 2. Exercises

#### Applied

1. This exercise relates to the College data set, which can be found in the file **College.csv**. It contains a number of variables for 777 different universities and colleges in the US.
2. Use the **read.csv()** function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

college <- read.csv("https://uclspp.github.io/datasets/data/College.csv")

1. Look at the data using the **fix()** function. You should notice that the first column is just the name of each university.We don’t really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

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

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

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

Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

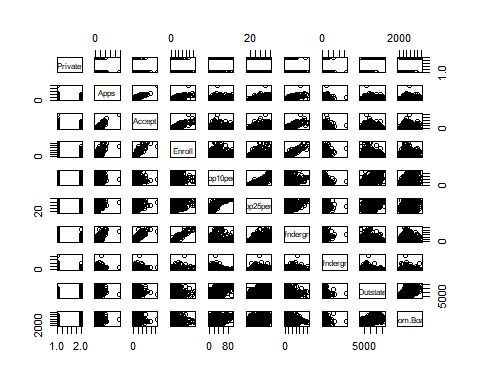
* 1. Use the summary() function to produce a numerical summary of the variables in the data set.

college$Private <- as.factor(college$Private)  
summary(college)

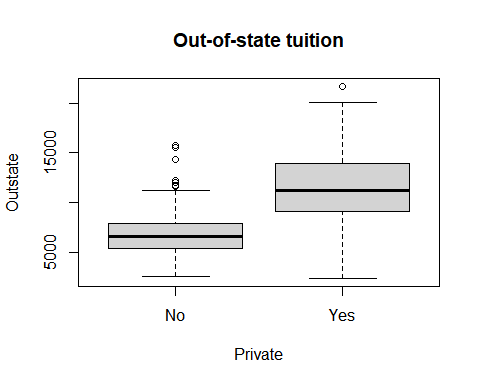
## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate   
## Min. : 10.00   
## 1st Qu.: 53.00   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

1. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

pairs(college[,1:10])

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

plot(college$Private,college$Outstate, main= "Out-of-state tuition", ylab= "Outstate", xlab= "Private")



Create a new qualitative variable, called **Elite**, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

Elite =rep ("No",nrow(college ))  
Elite [college$Top10perc >50]=" Yes"  
Elite =as.factor (Elite)  
college =data.frame(college ,Elite)

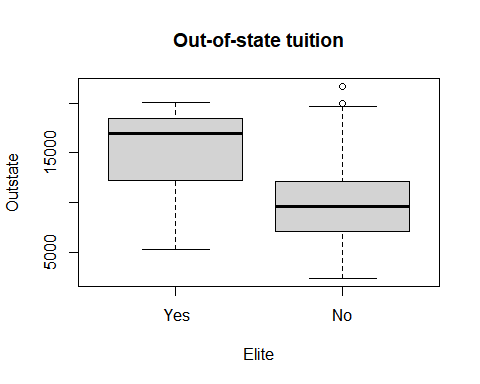
Use the summary() function to see how many elite universities there are.

summary(college$Elite)

## Yes No   
## 78 699

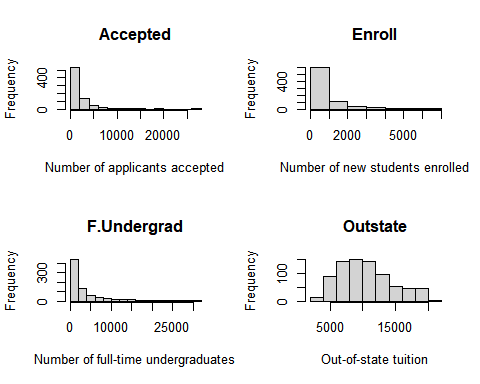
Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

plot(college$Elite,college$Outstate, main= "Out-of-state tuition", ylab= "Outstate", xlab= "Elite")

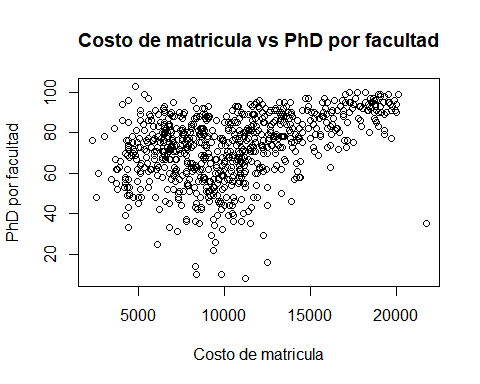


1. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

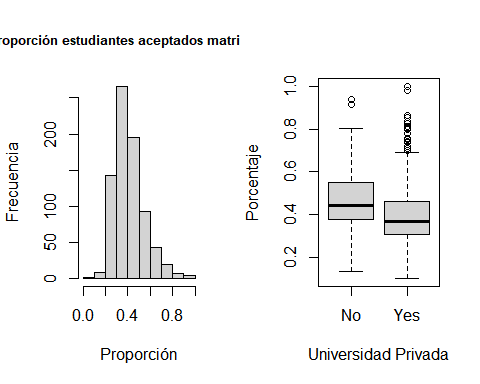
par(mfrow=c(2,2))  
hist(college$Accept, main = "Accepted", breaks = 12, xlab = "Number of applicants accepted")  
hist(college$Enroll, main = "Enroll", breaks = 8, xlab = "Number of new students enrolled")  
hist(college$F.Undergrad, main = "F.Undergrad", breaks = 15, xlab = "Number of full-time undergraduates")  
hist(college$Outstate, main = "Outstate", breaks = 10, xlab = "Out-of-state tuition")

 vi. Continue exploring the data, and provide a brief summary of what you discover.

plot(college$Outstate, college$PhD, main = "Costo de matricula vs PhD por facultad", xlab= "Costo de matricula", ylab = "PhD por facultad")

 Se puede observar que existe una menor variación entre la cantidad de profesores que poseen un PhD en aquellos colleges que tienen una matricula alta, por ende se puede decir que el personal edicativo de estos establecimientos tienen en promedio un nivel educativo más alto.

par(mfrow=c(1,2))  
hist(college$Enroll/college$Accept, main = "Proporción estudiantes aceptados matriculados", xlab="Proporción", ylab="Frecuencia", cex.main=0.8)  
plot(college$Private,(college$Enroll/college$Accept),xlab="Universidad Privada", ylab="Porcentaje")

 Vemos de los estudiantes se fueron aceptados y se matricularon el **71%** de estos, se encuentra entre el **30%** al **60%** de matriculados por universidad. Dentro de los cuales el promedio de admitimos que pagan su matricula es mucho mayor en las universidades publicas que en las privadas, sin embargo estas presentan una mayor cantidad de datos atipicos, en los cuales incluso el **100%** de los admitidos se matriculan.

1. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

library(ISLR)  
data(Auto)  
str(Auto)

## 'data.frame': 392 obs. of 9 variables:  
## $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...  
## $ cylinders : num 8 8 8 8 8 8 8 8 8 8 ...  
## $ displacement: num 307 350 318 304 302 429 454 440 455 390 ...  
## $ horsepower : num 130 165 150 150 140 198 220 215 225 190 ...  
## $ weight : num 3504 3693 3436 3433 3449 ...  
## $ acceleration: num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...  
## $ year : num 70 70 70 70 70 70 70 70 70 70 ...  
## $ origin : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ name : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2 ...

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

Gracias al comando **str()**, podemos ver que todas las variables con excepción de **name** son variables cuantitativas y por ende **name** es un predictor cuantitativo

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

sapply(Auto[ ,(1:8)], range)

## mpg cylinders displacement horsepower weight acceleration year origin  
## [1,] 9.0 3 68 46 1613 8.0 70 1  
## [2,] 46.6 8 455 230 5140 24.8 82 3

1. What is the mean and standard deviation of each quantitative predictor?

La media de cada variable predictora esta dada por:

sapply(Auto[ ,-9], mean)

## mpg cylinders displacement horsepower weight acceleration   
## 23.445918 5.471939 194.411990 104.469388 2977.584184 15.541327   
## year origin   
## 75.979592 1.576531

La desviación estándar de cada variable predictora esta dada por:

sapply(Auto[ ,-9], sd)

## mpg cylinders displacement horsepower weight acceleration   
## 7.8050075 1.7057832 104.6440039 38.4911599 849.4025600 2.7588641   
## year origin   
## 3.6837365 0.8055182

1. 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?

# Se ejecuta el subset de los datos  
  
Auto\_subset <- subset(Auto[-10:-85,-9])

El rango de dado para los datos restantes es:

sapply(Auto\_subset,range)

## mpg cylinders displacement horsepower weight acceleration year origin  
## [1,] 11.0 3 68 46 1649 8.5 70 1  
## [2,] 46.6 8 455 230 4997 24.8 82 3

La media para los datos restantes es:

sapply(Auto\_subset,mean)

## mpg cylinders displacement horsepower weight acceleration   
## 24.404430 5.373418 187.240506 100.721519 2935.971519 15.726899   
## year origin   
## 77.145570 1.601266

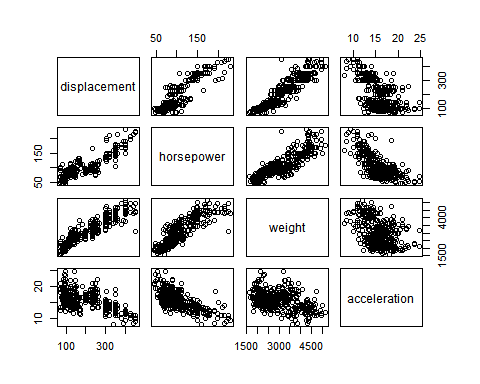
La desviación estándar para los datos restantes es:

sapply(Auto\_subset,sd)

## mpg cylinders displacement horsepower weight acceleration   
## 7.867283 1.654179 99.678367 35.708853 811.300208 2.693721   
## year origin   
## 3.106217 0.819910

1. Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

pairs(Auto[,3:6])

 1. Parece existir una relación lineal positiva entre los caballos de fuerza y el desplazamiento del motor, es decir que a mayor caballaje más desplazamiento del pistón dentro del cilindro. 2. Puede haber una relación lineal positiva que indica un mayor caballaje a un mayor masa del vehículo. 3. De igual manera que la anterior a mayor masa del vehículo, mayor desplazamiento. 4. Se presume una relación lineal negativa entre desplazamiento y la aceleración, es decir que a un mayor recorrido del pistón, una menor aceleración. 5. Se identifica una aparente relación lineal negativa entre los caballos de fuerza y la aceleración, a mayor cantidad de caballos de fuerza menor aceleración.

1. 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? Justify your answer.

Las otras variables que sugieren una relación de predicción de la variable mpg son displacement, horsepower y weight. Tenemos entonces que al incrementar cualquiera de estas variables las millas realizadas por galón decrecen, pudiendo incluso ajustarse a una recta con pendiente negativa.

1. This exercise involves the Boston housing data set.
2. To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

library(MASS)  
data(Boston)  
str(Boston)

## 'data.frame': 506 obs. of 14 variables:  
## $ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
## $ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  
## $ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
## $ chas : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ rm : num 6.58 6.42 7.18 7 7.15 ...  
## $ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
## $ dis : num 4.09 4.97 4.97 6.06 6.06 ...  
## $ rad : int 1 2 2 3 3 3 5 5 5 5 ...  
## $ tax : num 296 242 242 222 222 222 311 311 311 311 ...  
## $ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
## $ black : num 397 397 393 395 397 ...  
## $ lstat : num 4.98 9.14 4.03 2.94 5.33 ...  
## $ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

How many rows are in this data set?

nrow(Boston)

## [1] 506

How many columns?

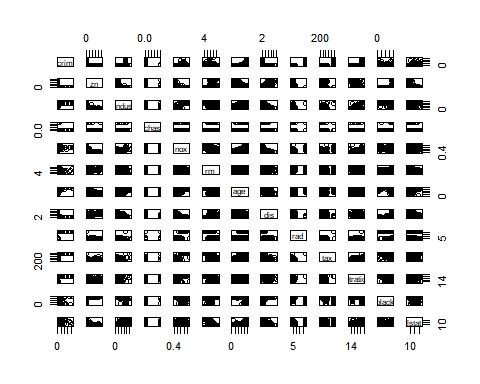
ncol(Boston)

## [1] 14

Whatdo the rows and columns represent?

1. crim: per capita crime rate by town
2. zn: proportion of residential land zoned for lots over 25,000 sq.ft
3. indus: proportion of non-retail business acres per town
4. chas: Charles River dummy variable (1 if tract bounds river; else 0)
5. nox: nitric oxides concentration (parts per 10 million)
6. rm: average number of rooms per dwelling
7. age: proportion of owner-occupied units built prior to 1940
8. dis: weighted distances to five Boston employment centres
9. rad: index of accessibility to radial highways
10. tax: full-value property-tax rate per $10,000
11. ptratio: pupil-teacher ratio by town
12. black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. lstat: % lower status of the population
14. medv: Median value of owner-occupied homes in $1000’s

pairs(Boston[,-14])

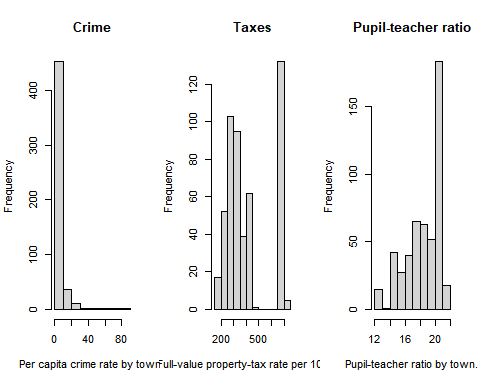


1. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Parece haber algunos predictores asociados a la tasa de crimen por distritos sin embargo no se observa una relación clara que determine una función **f** entre ellos y las variables predictoras siendo ellas: nox, rm, age, dis y lstat.

1. nox: Aparenta tener una relación linear con la tasa de crimen, es decir que tiene a haber más crimenes en zonas más contaminadas,
2. rm: Parece que la gran parte de los crimes esta asociados a áreas residenciales que poseen entre 5 a 7 hábitantes por vivienda
3. aga: Esta dada una rleacón linear positiva entre estas variables con lo que se puede decir que, a una mayor ocupación de casa antiguas ocupadas mayor nivel de crimen.
4. dist: Parte de una relación lineal aparentemente negativa, es decir que a mayor distancia de sus habitantes y los centro de empleo exsite una menor criminalidad.
5. lstat: tiene una relación lineal con la tasa de crimen es decir que a mayores niveles de probreza existe una memor ciminalidad.
6. Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios?

par(mfrow=c(1,3))  
hist(Boston$crim, main = "Crime", xlab = "Per capita crime rate by town.")  
hist(Boston$tax, main = "Taxes", xlab = "Full-value property-tax rate per 10,000.")  
hist(Boston$ptratio, main = "Pupil-teacher ratio", xlab = "Pupil-teacher ratio by town.")

 Como se puede ver gran parte de los suburbios tiene una tasa de crimen inferior al **20%**, la cantidad de suburbios valores superiores a este son:

sum(Boston$crim > 20)

## [1] 18

Los suburbios que tienen una tasa particularmente alta, es decir superior al **40%** son:

|  |  |
| --- | --- |
| Suburbio | Tasa |
| 381 | 88.97 |
| 419 | 73.53 |
| 406 | 67.92 |
| 411 | 51.13 |
| 415 | 45.74 |
| 405 | 41.53 |

Tax rates?

La cantidad de suburbios con un impuesto superior a **$600 USD**, sobre el total de bienes inmuebles por valor total de 10.000 dólares es de:

sum(Boston$tax > 600)

## [1] 137

Dentro de los cuales los siguientes distritos superan un impuesto de **$700 USD** son:

|  |  |
| --- | --- |
| Suburbio | Tasa |
| 491 | 711 |
| 490 | 711 |
| 489 | 711 |
| 493 | 711 |
| 492 | 711 |

Pupil-teacher ratios?

La cantidad de suburbios que tienen una tasa de Profesor-alumno superior a 20 alumnos por profesor son:

sum(Boston$ptratio > 20)

## [1] 201

Dentro de los cuales los siguientes 2 distritos tiene una tasa de 22 alumnos por profesor:

|  |  |
| --- | --- |
| Suburbio | Tasa |
| 356 | 22 |
| 355 | 22 |

Comment on the range of each predictor.

sapply(Boston[,-14],range)

## crim zn indus chas nox rm age dis rad tax ptratio black  
## [1,] 0.00632 0 0.46 0 0.385 3.561 2.9 1.1296 1 187 12.6 0.32  
## [2,] 88.97620 100 27.74 1 0.871 8.780 100.0 12.1265 24 711 22.0 396.90  
## lstat  
## [1,] 1.73  
## [2,] 37.97

1. crim: Vemos que la tasa de crimenes en todo Boston no supera el **90%**, y en algunos casos en cercana a **0%**, tambien se aprecia como el **96%** de los suburbios de Boston, posee una tasa inferior al **20%** de criminalidad.
2. zn: Se aprecia que existen áreas de de Boston que no poseen zonas residenciales por cada 25000 sq.ft, mientras otras se encuentran completamente llenas de áreas residenciales siendo estas apenas el **2%**.
3. indus: El rango nos muestra que el grado de áreas industriales máximo por cada Acre en Boston es apenas el **28%**. No obstante el **32%** de estas áreas esta industrializadas en alrededor del **18%**.
4. chas: Esta en una variable logica donde se representa que si los distritos de Boston conectan con el río Charles vemos que apenas el **7%** de ellos no lo hacen.
5. nox: Es apreciable que el rango de concentración de Nitrogeno de oxigeno por cada 10 millones se encuentra entre 0.385 y 0.871. con el **50%** de los distritos con un nivel de contaminazión que va entre **0.45 - 0.62**
6. rm: Vemos que el numero de habitantes por vivienda se encentra entre el 3 a 9 personas, de las cuales el **84%** se encuentran habitadas entre 5 a 7 personas por vivienda.
7. age: Se aprecia que la proporción de viviendas ocupadas por distrito que fueron construidas antes del 1940, se encuentra entre el **2** y el **100**, lo que parece ser un indicativo que grandes áreas de la ciudad poseen edificios activos con más de 80 años de antiguedad para la fecha actual.
8. dis: Se puede observar que la media de la distancia a los centro de empleo no supera los 12 km, con lo cual se puede decir que Boston es un a ciudad relativamente pequeña.
9. rad: Se aprecia que el indice de accesibilidad a autopistas radiales es va desde **0** a **24**, de los cuales el **74%** tiene un indice inferior a **10**.
10. tax: Se evidencia que para ningún distrito la tasas de impuesto por inmueble por cada **$1000 USD** es inferior a **$187 USD** y que en solo el **27%** de los suburbios es superior a **$600 USD**.
11. ptratio: Es claro que la tasa de profesores por alumno, no supera **22** alumnos por profesor es decir que se puede decir que en el peor de los casos los salones apenas superan la mitad de los estudiantes máximos por profesor en Bogotá.
12. black: podemos ver que la propoción de habitantes de color Negro en algunas áreas es casi nula, mientras otras ronda los 400.
13. lstat: Vemos que el porcentaje de ciudadanos de bajos ingresos, por cada distrito de boston, no supera el **38%**, y que el **73%**, de los distritos registra niveles de probreza entre el **5-20%**
14. How many of the suburbs in this data set bound the Charles river?

La cantidad de suburbios en Boston que limitan con el río charles son:

sum(Boston$chas == 1)

## [1] 35