Taller3

May 12, 2022

#A

[ ]:
```
library(mlbench)
data(BostonHousing)
```

[ ]:
```
BostonHousing
```

A data.frame: 506 × 14

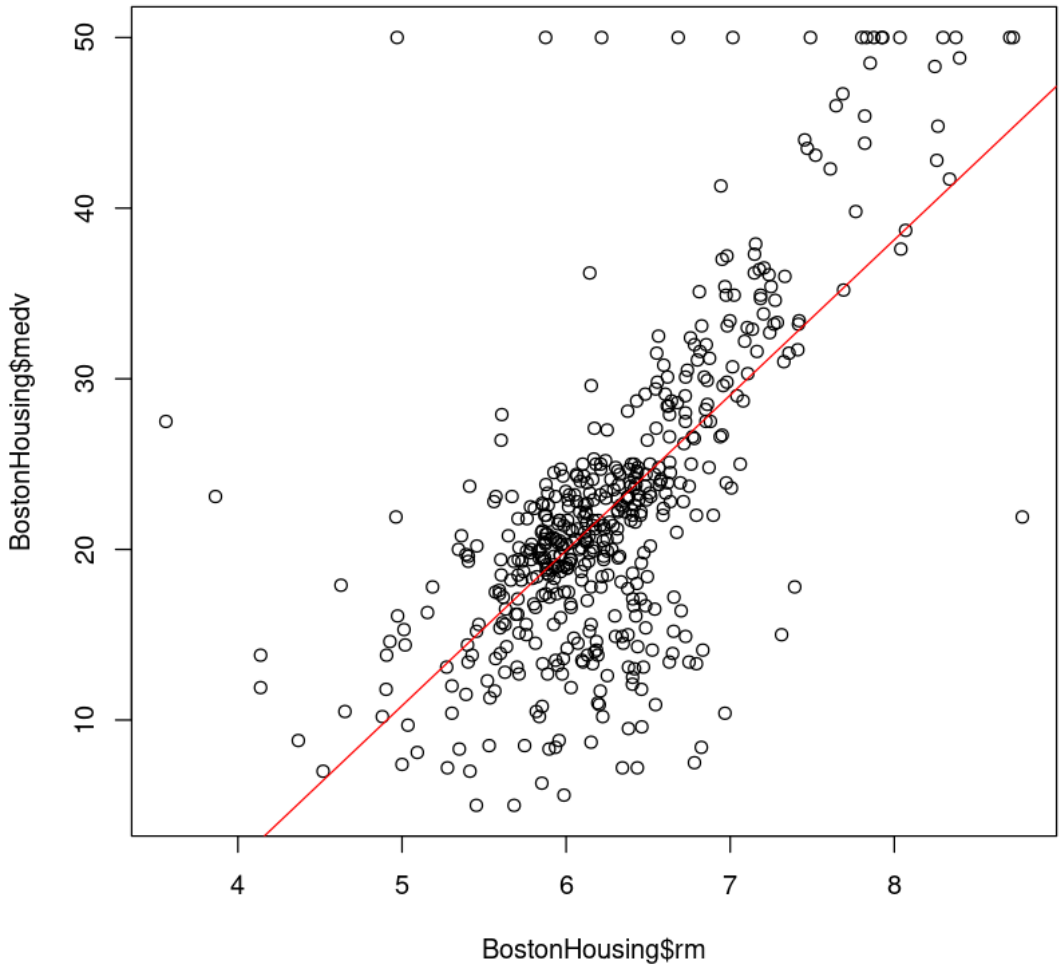
	crim	zn	indus	chas	nox	rm	age	dis	rad
	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1
2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2
3	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2
4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3
5	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3
6	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5
9	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5
13	0.09378	12.5	7.87	0	0.524	5.889	39.0	5.4509	5
14	0.62976	0.0	8.14	0	0.538	5.949	61.8	4.7075	4
15	0.63796	0.0	8.14	0	0.538	6.096	84.5	4.4619	4
16	0.62739	0.0	8.14	0	0.538	5.834	56.5	4.4986	4
17	1.05393	0.0	8.14	0	0.538	5.935	29.3	4.4986	4
18	0.78420	0.0	8.14	0	0.538	5.990	81.7	4.2579	4
19	0.80271	0.0	8.14	0	0.538	5.456	36.6	3.7965	4
20	0.72580	0.0	8.14	0	0.538	5.727	69.5	3.7965	4
21	1.25179	0.0	8.14	0	0.538	5.570	98.1	3.7979	4
22	0.85204	0.0	8.14	0	0.538	5.965	89.2	4.0123	4
23	1.23247	0.0	8.14	0	0.538	6.142	91.7	3.9769	4
24	0.98843	0.0	8.14	0	0.538	5.813	100.0	4.0952	4
25	0.75026	0.0	8.14	0	0.538	5.924	94.1	4.3996	4
26	0.84054	0.0	8.14	0	0.538	5.599	85.7	4.4546	4
27	0.67191	0.0	8.14	0	0.538	5.813	90.3	4.6820	4
28	0.95577	0.0	8.14	0	0.538	6.047	88.8	4.4534	4
29	0.77299	0.0	8.14	0	0.538	6.495	94.4	4.4547	4
30	1.00245	0.0	8.14	0	0.538	6.674	87.3	4.2390	4
477	4.87141	0	18.10	0	0.614	6.484	93.6	2.3053	24
478	15.02340	0	18.10	0	0.614	5.304	97.3	2.1007	24
479	10.23300	0	18.10	0	0.614	6.185	96.7	2.1705	24
480	14.33370	0	18.10	0	0.614	6.229	88.0	1.9512	24
481	5.82401	0	18.10	0	0.532	6.242	64.7	3.4242	24
482	5.70818	0	18.10	0	0.532	6.750	74.9	3.3317	24
483	5.73116	0	18.10	0	0.532	7.061	77.0	3.4106	24
484	2.81838	0	18.10	0	0.532	5.762	40.3	4.0983	24
485	2.37857	0	18.10	0	0.583	5.871	41.9	3.7240	24
486	3.67367	0	18.10	0	0.583	6.312	51.9	3.9917	24
487	5.69175	0	18.10	0	0.583	6.114	79.8	3.5459	24
488	4.83567	0	18.10	0	0.583	5.905	53.2	3.1523	24
489	0.15086	0	27.74	0	0.609	5.454	92.7	1.8209	4
490	0.18337	0	27.74	0	0.609	5.414	98.3	1.7554	4
491	0.20746	0	27.74	0	0.609	5.093	98.0	1.8226	4
492	0.10574	0	27.74	0	0.609	5.983	98.8	1.8681	4
493	0.11132	0	27.74	0	0.609	5.983	83.5	2.1099	4
494	0.17331	0	9.69	0	0.585	5.707	54.0	2.3817	6
495	0.27957	0	9.69	0	0.585	5.926	42.6	2.3817	6
496	0.17899	0	9.69	0	0.585	5.670	28.8	2.7986	6

[ ]:
```r
fit <- lm(medv ~ rm, data = BostonHousing)

summary(fit)
```

```
Call:
lm(formula = medv ~ rm, data = BostonHousing)

Residuals:
    Min      1Q  Median      3Q     Max
-23.346  -2.547   0.090   2.986  39.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,        Adjusted R-squared:  0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

B1, intercepto en -34.671, dado que una vivienda no tiene habitaciones se estima que la vivienda cuesta $-34.671. lo cual es logico, ya que no existen viviendas que no tienen habitaciones. B2, 9.102, por cada habitacion adicional se estima que el precio de la vivienda aumenta en $9.102.

[ ]:
```r
plot(BostonHousing$medv ~ BostonHousing$rm)
abline(fit, col = "red")
```



[ ]:
```
fit <- lm(medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + tax +␣
↪ptratio + b + lstat, data = BostonHousing)
summary(fit)
```

```
Call:
lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
    dis + rad + tax + ptratio + b + lstat, data = BostonHousing)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199
```

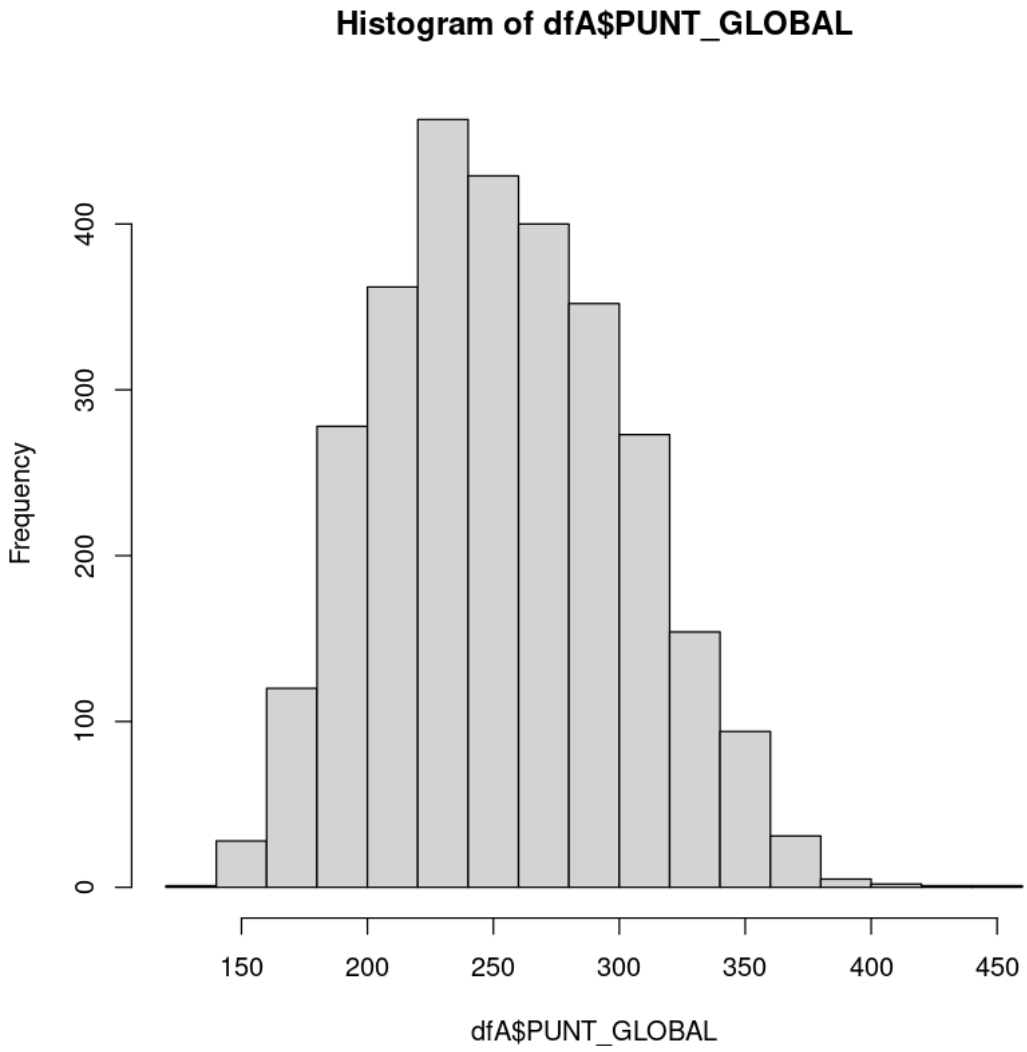
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas1        2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
b            9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,        Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

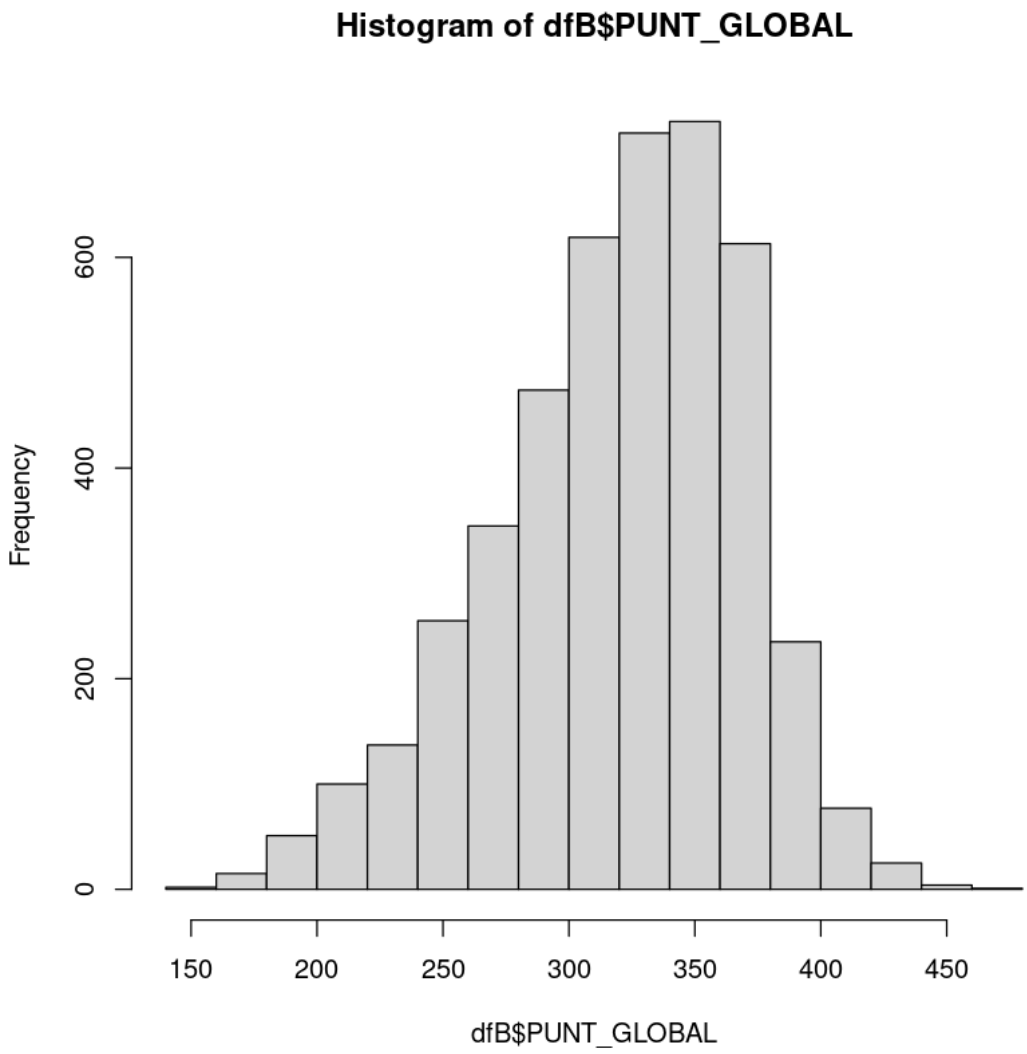
crim, por cada unidad que aumente el crimen se estima que la vivienda pierde -1.080e-01 de precio. nox, por cada unidad que aumenta la concentracion de oxido nitrico en la zona se estima que la vivienda pierde -1.777e+01 de precio. b, por cada unidad que aumenta la cantidad de personas de poblacion negra en el sector se estima que la vivienda pierde 9.312e-03 de precio.

#B

[ ]:
```r
library(readr)
dfA <- read_csv("saber18A.csv")
dfB <- read_csv("saber18B.csv")
```

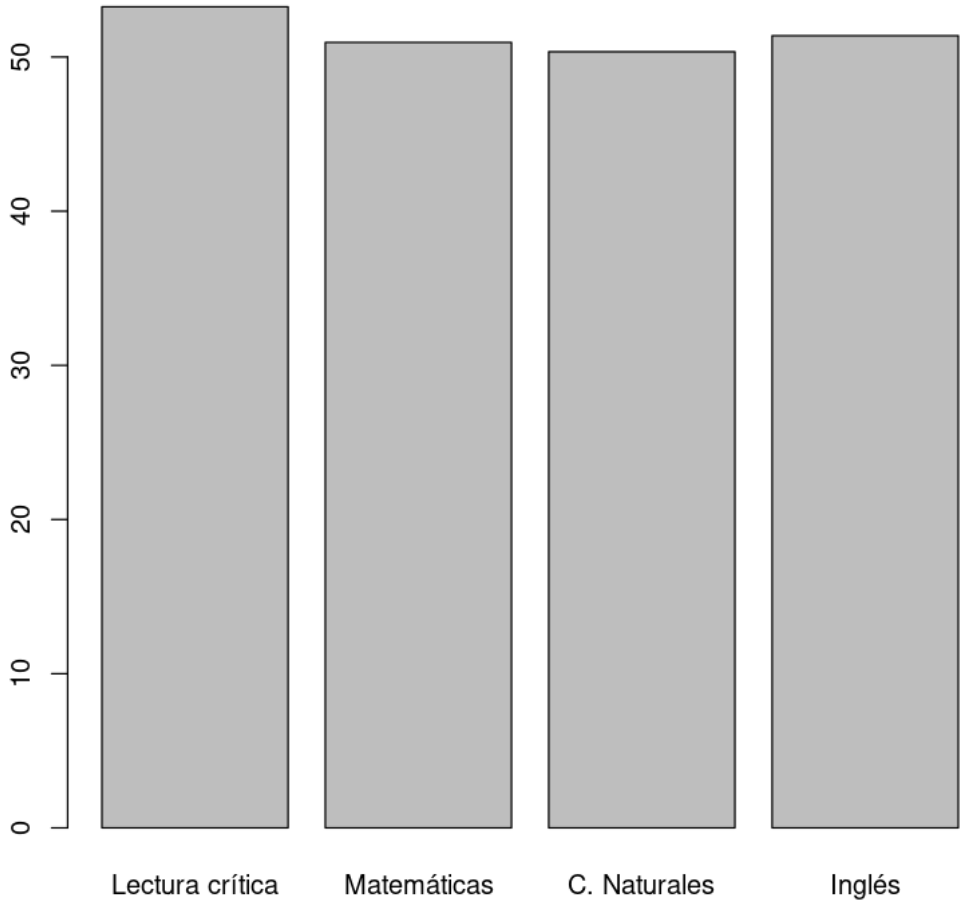
[ ]:
```r
hist(dfA$PUNT_GLOBAL)
hist(dfB$PUNT_GLOBAL)
```

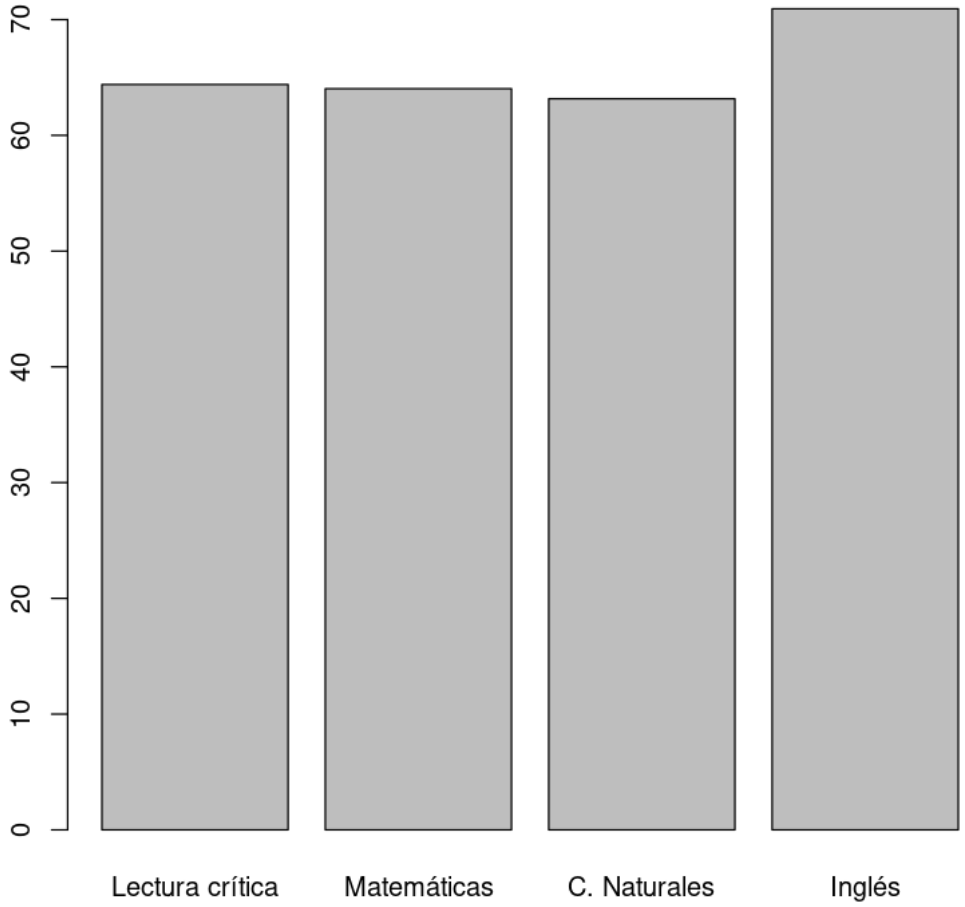

Histogram of dfA$PUNT_GLOBAL
Frequency
0
100
200
300
400
150
200
250
300
350
400
450
dfA$PUNT_GLOBAL



los datos del calendario B tienen una mayor acumulacion de datos mayores a 300 puntos quen los estudiantes de calendario A los cuales tienen los datos mayormente distribuidos a la izquierda.

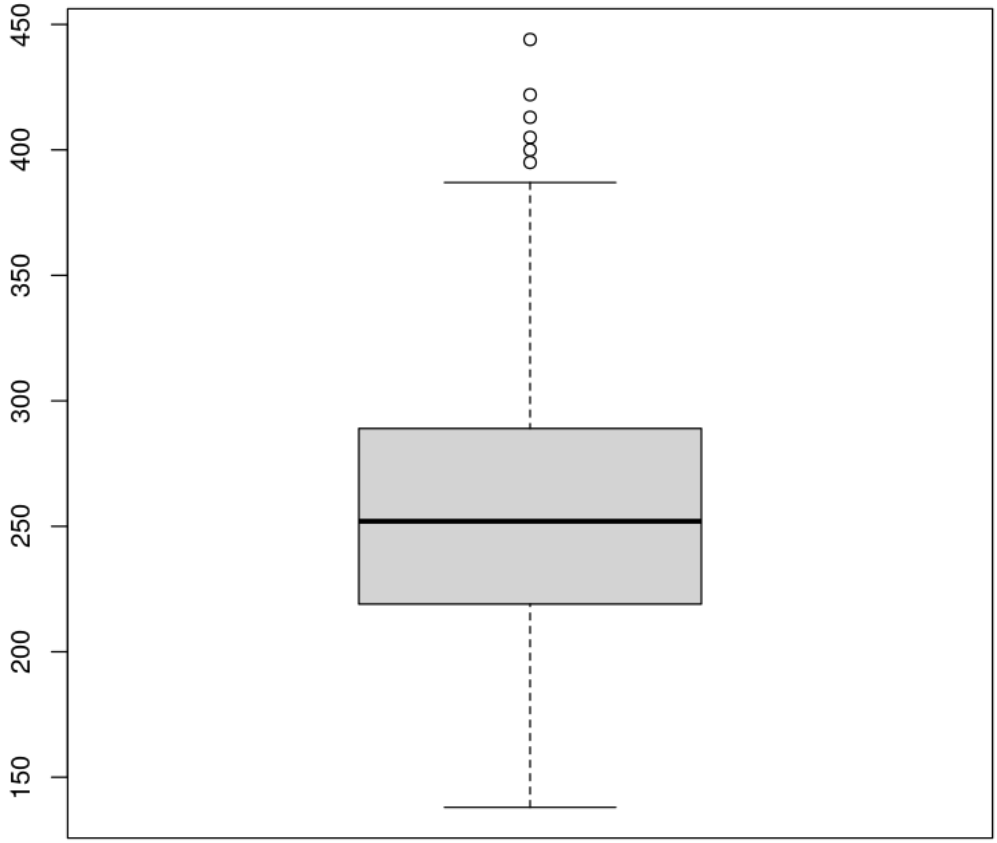
[ ]:
```
puntA_mean <- c(mean(dfA$PUNT_LECTURA_CRITICA), mean(dfA$PUNT_MATEMATICAS),␣
 ↪mean(dfA$PUNT_C_NATURALES), mean(dfA$PUNT_INGLES))
puntB_mean <- c(mean(dfB$PUNT_LECTURA_CRITICA), mean(dfB$PUNT_MATEMATICAS),␣
 ↪mean(dfB$PUNT_C_NATURALES), mean(dfB$PUNT_INGLES))
barplot(puntA_mean, names = c("Lectura crítica", "Matemáticas", "C. Naturales",␣
 ↪"Inglés"))
barplot(puntB_mean, names = c("Lectura crítica", "Matemáticas", "C. Naturales",␣
 ↪"Inglés"))
```

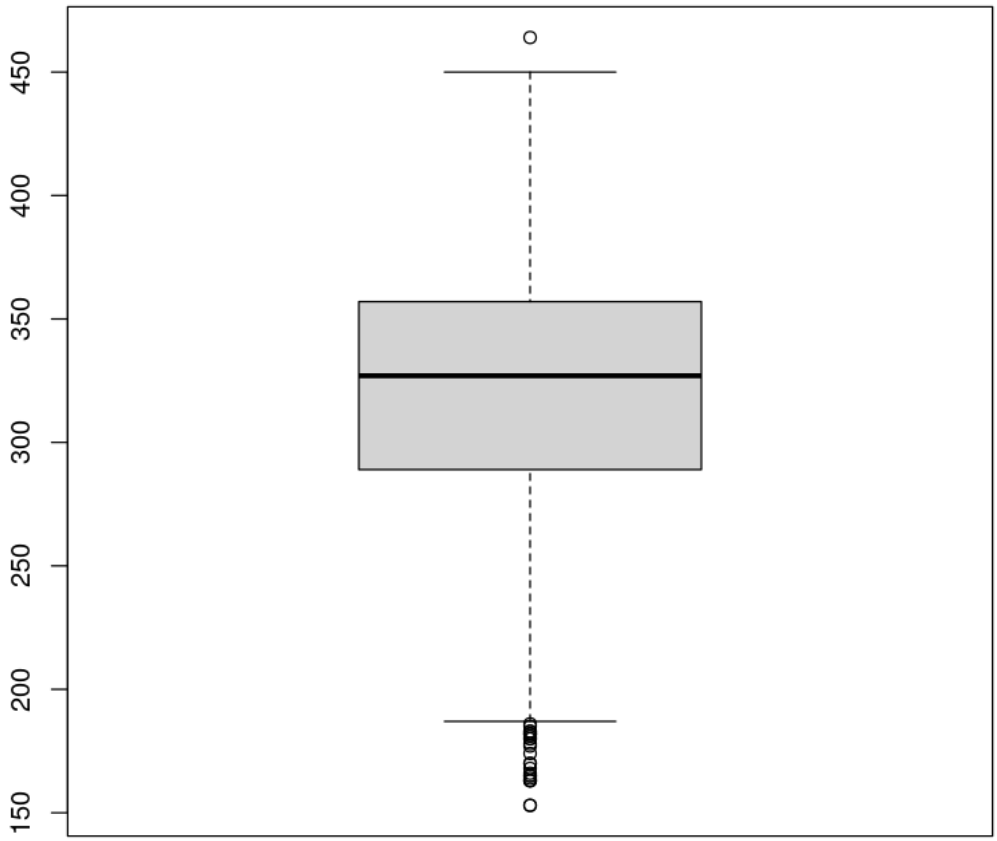

0
10
20
30
40
50
Lectura crítica
Matemáticas
C. Naturales
Inglés



los estudiantes de calendario B tienen mayores notas en todas las asignaturas que los estudiantes de calendario A, los de calendario B destacan principalmente en ingles y los de calendario A en lectura critica no obstante tienen menor nota de media en lectura critica que los de calendario B

[ ]:
```
boxplot(dfA$PUNT_GLOBAL)
boxplot(dfB$PUNT_GLOBAL)
```


450
400
350
300
250
200
150



como era de esperarse por una mayor media los estudiantes de calendario B tienen sus datos atipicos en los puntajes bajos y por el contrario tienen pocos datos atipicos altos, y los estudiantes de calendario A tienen datos atipicos en los puntajes altos y no tienen datos atipicos bajos.

[ ]:
```
table(dfA$FAMI_ESTRATOVIVIENDA, dfA$FAMI_TIENEINTERNET)

chisq <- chisq.test(dfA$FAMI_ESTRATOVIVIENDA, dfA$FAMI_TIENEINTERNET)

round(chisq$stdres,2)
```

```
            No  Si
  Estrato 1 761 406
  Estrato 2 342 666
```

```
  Estrato 3 106 627
  Estrato 4  11  62
  Estrato 5   3   7
  Estrato 6   1   0
```

```
Warning message in chisq.test(dfA$FAMI_ESTRATOVIVIENDA, dfA$FAMI_TIENEINTERNET):
"Chi-squared approximation may be incorrect"
```

```
                        dfA$FAMI_TIENEINTERNET
dfA$FAMI_ESTRATOVIVIENDA     No     Si
               Estrato 1  21.62 -21.62
               Estrato 2  -5.54   5.54
               Estrato 3 -16.76  16.76
               Estrato 4  -4.55   4.55
               Estrato 5  -0.70   0.70
               Estrato 6   1.20  -1.20
```

por la baja cantidad de datos de personas estrato 6 en calendario A podria descartarse este valor, y se encuentra que de alguna forma si hay asociacion al menos entre ser de estrato 1 y no tener internet, se presenta una gran desviacion por parte del Estrato 3 que puede estar relacionada con mayor cantidad de ingresos.

[ ]:
```
var.test(dfA$PUNT_GLOBAL, dfB$PUNT_GLOBAL)
```

```
        F test to compare two variances

data:  dfA$PUNT_GLOBAL and dfB$PUNT_GLOBAL
F = 0.98028, num df = 2993, denom df = 4399, p-value = 0.554
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9181487 1.0470532
sample estimates:
ratio of variances
         0.9802809
```

[ ]:
```
t.test(dfA$PUNT_GLOBAL, dfB$PUNT_GLOBAL, var.equal = FALSE, alternative = "two.
  ↪sided", conf.level = 0.95)
```

```
        Welch Two Sample t-test

data:  dfA$PUNT_GLOBAL and dfB$PUNT_GLOBAL
t = -57.199, df = 6469.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -67.98049 -63.47520
sample estimates:
mean of x mean of y
```

```
 254.5167  320.2445
```

```
[ ]: t.test(dfA$PUNT_GLOBAL, dfB$PUNT_GLOBAL, var.equal = FALSE, alternative =␣
      ↪"greater", conf.level = 0.95)
```

```
        Welch Two Sample t-test

data:  dfA$PUNT_GLOBAL and dfB$PUNT_GLOBAL
t = -57.199, df = 6469.4, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -67.61824       Inf
sample estimates:
mean of x mean of y
 254.5167  320.2445
```

la media de nota global de los estudiantes de Calendario B es mayor a la de los estudiantes de Calendario A