Guió de les classes de laboratori sobre anàlisi descriptiva

1. Començar a treballar amb R

1. Comencem fent una ullada a l'script d'R que es troba a `http://www-eio.upc.es/teaching/pe/read-data` i l'executem:

   ```
   > source(url("http://www-eio.upc.es/teaching/pe/read-data"))
   ```

   Fixem-nos que, entre d'altres, s'ha creat un conjunt de dades, un *data frame*, que es diu `DAT`. Abans de començar es recomanable mirar-nos aquestes dades, per exemple executant els següents comandes:

   ```
   > DAT
   > View(DAT)

   > head(DAT)

         op        mysql        post
   1 INSERT 0.0003011227 0.036271811
   2 INSERT 0.1224038601 0.069139004
   3 INSERT 0.0002570152 0.001724005
   4 INSERT 0.0016028881 0.020658970
   5 INSERT 0.0480089188 0.098726988
   6 INSERT 0.0015101433 0.055121899

   > str(DAT)

   'data.frame':        130 obs. of  3 variables:
    $ op   : Factor w/ 4 levels "DELETE","INSERT",..: 2 2 2 2 2 2 2 2 2 2 ...
    $ mysql: num  0.000301 0.122404 0.000257 0.001603 0.048009 ...
    $ post : num  0.03627 0.06914 0.00172 0.02066 0.09873 ...
   ```

2. El que ens interessa és fer una descripció d'aquestes dades. Com ho podem fer en cas de les variables numèriques?
   $\rightsquigarrow$ Transparències 4 i 7 a 10 de `ED.ppt`

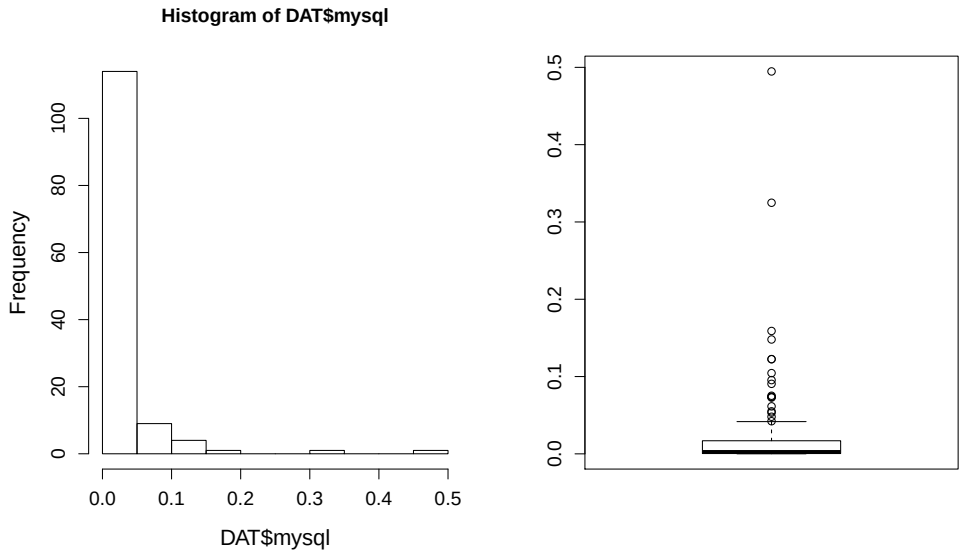
   Amb R:

   ```
   > mean(DAT$mysql)

   [1] 0.02258645

   > median(DAT$mysql)

   [1] 0.002939939

   > summary(DAT$mysql)

        Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   0.0000980 0.0004218 0.0029400 0.0225900 0.0166600 0.4947000

   > var(DAT$mysql)

   [1] 0.003413237

   > sd(DAT$mysql)

   [1] 0.05842291
   ```

   Com s'interpreten aquests valors?

3. Apart de calcular els indicadors numèrics de la tendència central i de la dispersió, es recomanable fer una representació gràfica de la distribució d'aquesta variable.
   ⇝ Transparències 12 a 16 de `ED.ppt`

```
> hist(DAT$mysql)
> boxplot(DAT$mysql)
```



   Què s'hi observa? Quin dels dos gràfics es preferible en aquest cas?

4. En canvi, en cas de la variable categòrica `op`, què ens interessa saber? Com la podem descriure?
   ⇝ Transparències 12, 17 i 19 de `ED.ppt`

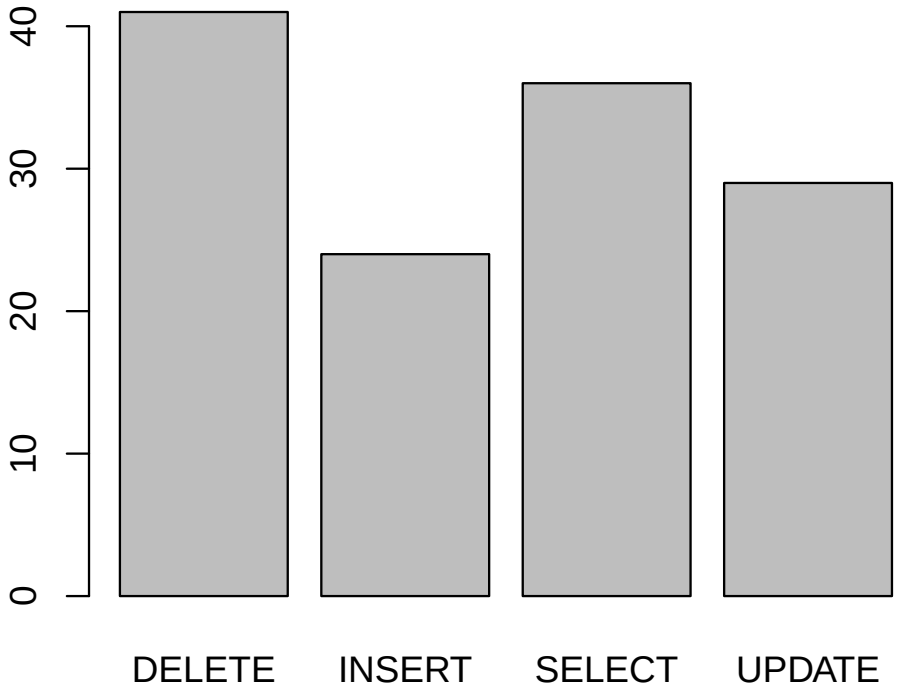
```
> table(DAT$op)

DELETE INSERT SELECT UPDATE
    41     24     36     29

> prop.table(table(DAT$op))

   DELETE    INSERT    SELECT    UPDATE
0.3153846 0.1846154 0.2769231 0.2230769
```

5. Seria desitjable que la taula de freqüència contingués tant les freqüències absolutes com les relatives. Existeixen varies funcions en diferents paquets d'R que es poden instal·lar. Per exemple podem utilitzar la funció `freq` del paquet `descr`:

```
> install.packages("descr")    # Instal·lació del paquet
> library(descr)               # Es carrega el paquet
> freq(DAT$op, plot=F)

DAT$op
       Frequency Percent
DELETE        41   31.54
INSERT        24   18.46
SELECT        36   27.69
UPDATE        29   22.31
Total        130  100.00
```

   Important: Un cop instal·lat un paquet d'R en un ordinador, ja no cal fer-ho a les pròximes sesions d'R. En canvi, se'l ha de cargar en cada nova sesió d'R.

```
> barplot(table(DAT$op))
```



6. Si volem exportar les dades podem utilitzar la funció `write.table` i per guardar el contingut d'una sessió d'R podem fer servir la funció `save.image`:

```
> write.table(DAT, file="DadesDAT.txt", quote=F, row.names=F)
> save.image("DadesDAT.RData")
```

2. Exercicis

`Nota:` Copieu tant les preguntes com les instruccions d'R següents i pegueu-les a un document WORD. A continuació completeu el document WORD amb les instruccions completes, els resultats i vostres comentaris.

1. Feu una descripció de la segona variable numèrica `post`. Com aquesta variable te alguna dada mancant (*missing*), que es denota amb `NA` en R, les funcions `mean`, `var`, etc. tanmateix tornen un `NA`. Per resoldre aquest problema podeu mirar l'ajuda de la funció `mean` (executant `?mean` en R) o mirar l'apartat 3.1.1 del tutorial d'R a `http://www-eio.upc.es/teaching/pe/B1/`:

```
> mean(DAT$post)

[1] NA

> mean(DAT$post, na.rm=T)

[1] 0.09212431

> median(DAT$post, na.rm=T)

[1] 0.05184507

> summary(DAT$post, na.rm=T)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
0.0007269 0.0207600 0.0518500 0.0921200 0.0987300 1.2010000         1
```

```
> var(DAT$post, na.rm=T)
```

```
[1] 0.02284356
```

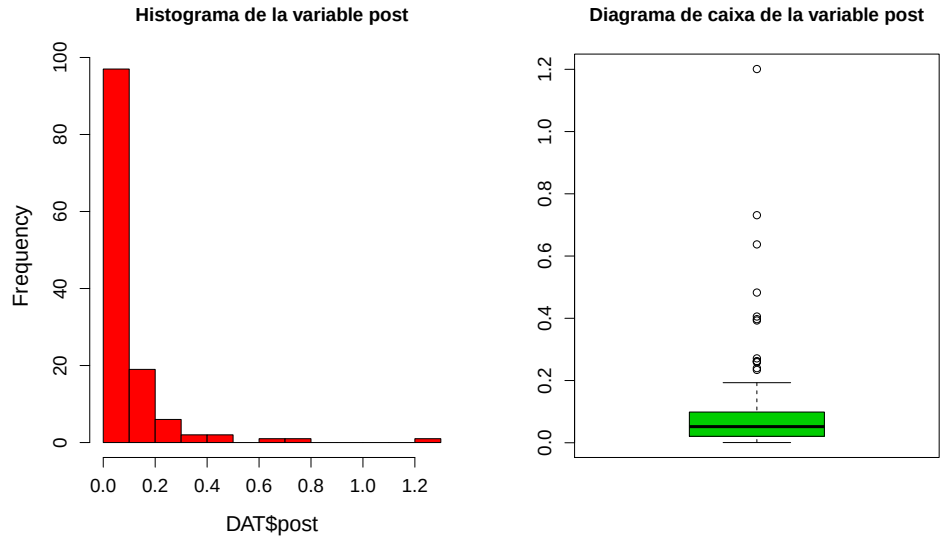
```
> sd(DAT$post, na.rm=T)
```

```
[1] 0.1511409
```

Un cop resolt el petit problema i fet els càlculs, interpreteu els resultats obtinguts.

2. Feu també una representació gràfica de la variable `post`:

```
> windows(width=10)
> par(mfrow=1:2, cex.lab=1.3, cex.axis=1.2)
> hist(DAT$post, col=2, main="Histograma de la variable post")
> boxplot(DAT$post, col=3, main="Diagrama de caixa de la variable post")
```



Com es pot descriure aquesta distribució?

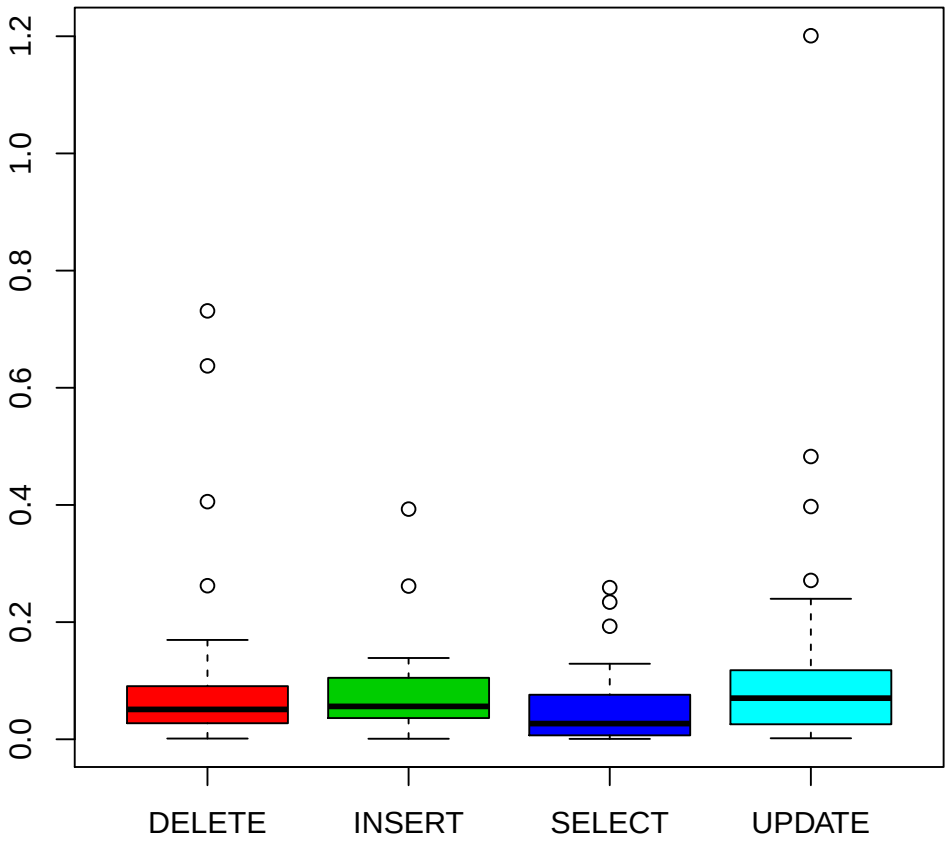
3. Per saber si el comportament d'aquesta variable varia d'un `op` a un altre, s'han de fer els càlculs dels indicadors numèrics per cada grup. En R ho podem fer amb la funció `tapply`.
   ⇝ Transparència 11 de `ED.ppt` o pàgina 39 del tutorial

```
> with(DAT, tapply(post, op, summary))
```

```
$DELETE
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001259 0.027390 0.050980 0.096240 0.090700 0.731300

$INSERT
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0009499 0.0362600 0.0561600 0.0808900 0.1018000 0.3930000

$SELECT
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
0.0007269 0.0067450 0.0268300 0.0551600 0.0760100 0.2589000         1

$UPDATE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00173 0.02565 0.07020 0.14020 0.11790 1.20100
```

4. Dibuixeu un diagrama de caixa de la variable `post` en funció de `op`:
   ⇝ Transparència 14 de `ED.ppt` o pàgines 52 i 53 del tutorial

```
> boxplot(post~op, DAT, col=2:5)
```



Comenteu les diferències que hi podeu observar.

3. Anàlisi descriptiva bivariant

1. A continuació utilitzarem un dels conjunts de dades de la llibreria `datasets` d'R. Es tracta de dades dels 50 estats dels Estats Units:

```
> ?state
> View(state.x77)

> str(state.x77)

 num [1:50, 1:8] 3615 365 2212 2110 21198 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
  ..$ : chr [1:8] "Population" "Income" "Illiteracy" "Life Exp" ...

> str(as.data.frame(state.x77))

'data.frame':        50 obs. of  8 variables:
 $ Population: num  3615 365 2212 2110 21198 ...
 $ Income    : num  3624 6315 4530 3378 5114 ...
 $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
 $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
 $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
 $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
 $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
 $ Area      : num  50708 566432 113417 51945 156361 ...

> state.region

 [1] South         West          West          South         West
 [6] West          Northeast     South         South         South
[11] West          West          North Central North Central North Central
[16] North Central South         South         Northeast     South
[21] Northeast     North Central North Central South         North Central
[26] West          North Central West          Northeast     Northeast
[31] West          Northeast     South         North Central North Central
[36] South         West          Northeast     Northeast     South
[41] North Central South         South         West          Northeast
[46] South         West          South         North Central West
Levels: Northeast South North Central West

> state77 <- cbind(as.data.frame(state.x77), state.region)
> head(state77)

           Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
           state.region
Alabama           South
Alaska             West
Arizona            West
Arkansas          South
California         West
Colorado           West
```

```
> summary(state77)
```

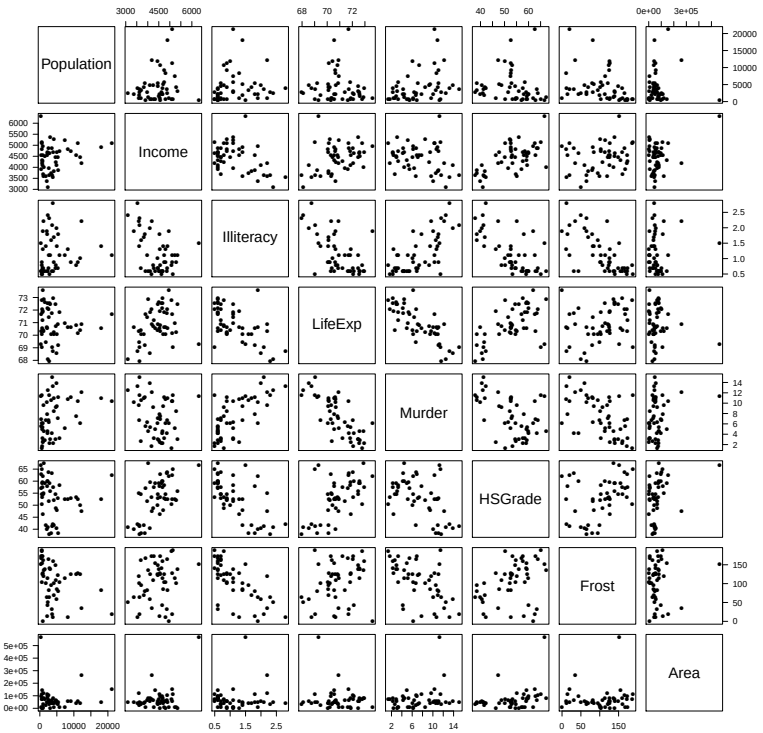
```
   Population        Income       Illiteracy       Life Exp        Murder
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96  Min.   : 1.400
 1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12  1st Qu.: 4.350
 Median : 2838   Median :4519   Median :0.950   Median :70.67  Median : 6.850
 Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88  Mean   : 7.378
 3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89  3rd Qu.:10.675
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60  Max.   :15.100
    HS Grad          Frost             Area                state.region
 Min.   :37.80   Min.   :  0.00   Min.   :  1049   Northeast    : 9
 1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985   South        :16
 Median :53.25   Median :114.50   Median : 54277   North Central:12
 Mean   :53.11   Mean   :104.46   Mean   : 70736   West         :13
 3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
 Max.   :67.30   Max.   :188.00   Max.   :566432
```

```
> names(state77)[c(4, 6, 9)] <- c("LifeExp", "HSGrade", "Region")
> names(state77)
```

```
[1] "Population" "Income"     "Illiteracy" "LifeExp"    "Murder"
[6] "HSGrade"    "Frost"      "Area"       "Region"
```

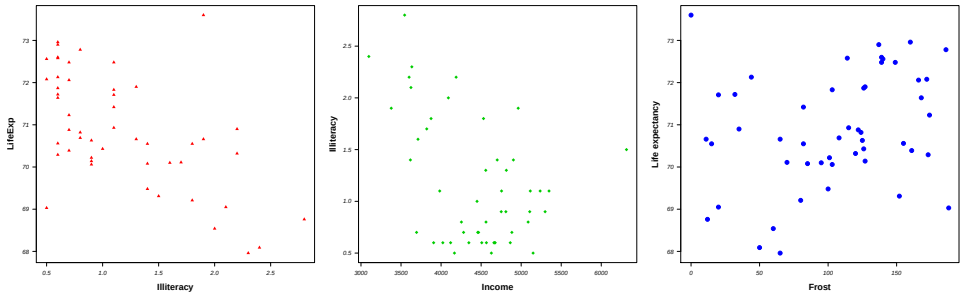
2. Ens interessa ara fer una anàlisi descriptiva bivariant, tant de dues variables numèriques com d'un parell de variables categòriques. Per al primer cas és molt recomanable fer un diagrama de dispersió (*Scatterplot*), que ens dona una idea de la relació entre ambdues variables.
   ⇝ Transparències 12, 20, 21 i 24 de `ED.ppt`

```
> windows(height=10, width=10)
> par(las=1)
> pairs(state77[, 1:8], pch=16)
```



Entre quines variables sembla haver-hi més relació? Mirem amb més detall les relacions entre algunes de les variables:

```
> windows(height=5, width=15)
> par(mfrow=c(1, 3), las=1, font.lab=2, font.axis=3)
> with(state77, plot(Illiteracy, LifeExp, pch=17, col=2))
> with(state77, plot(Illiteracy~Income, pch=18, col=3, cex=1.3))
> plot(LifeExp~Frost, data=state77, pch=19, col=4, ylab="Life expectancy", cex=1.5)
```



Què hi podem observar? Com podem descriure les relacions?

3. En cas de que podem suposar una relació lineal entre dues variables numèriques, es pot calcular el coefficient de correlació (lineal), que quantifica el grau de relació lineal:
   ⇝ Transparències 22 i 23 de `ED.ppt`

```
> cor(state77)

Error en cor(state77) : "x" must be numeric

> round(cor(state77[, 1:8]), 3)

           Population Income Illiteracy LifeExp Murder HSGrade  Frost   Area
Population      1.000  0.208      0.108  -0.068  0.344  -0.098 -0.332  0.023
Income          0.208  1.000     -0.437   0.340 -0.230   0.620  0.226  0.363
Illiteracy      0.108 -0.437      1.000  -0.588  0.703  -0.657 -0.672  0.077
LifeExp        -0.068  0.340     -0.588   1.000 -0.781   0.582  0.262 -0.107
Murder          0.344 -0.230      0.703  -0.781  1.000  -0.488 -0.539  0.228
HSGrade        -0.098  0.620     -0.657   0.582 -0.488   1.000  0.367  0.334
Frost          -0.332  0.226     -0.672   0.262 -0.539   0.367  1.000  0.059
Area            0.023  0.363      0.077  -0.107  0.228   0.334  0.059  1.000

> with(state77, round(cor(Area, Illiteracy), 3))

[1] 0.077

> with(state77, round(cor(LifeExp, Illiteracy), 3))

[1] -0.588
```

   Interpreteu aquests valors.

4. Per categoritzar una variable numèrica per tal de crear una variable ordinal podem usar la funció `cut`:
   ⇝ Pàgina 54 del tutorial

```
> cut(state77$Income, c(0, 4000, 4500, 5000, 10000))

 [1] (0,4e+03]       (5e+03,1e+04]   (4.5e+03,5e+03] (0,4e+03]
 [5] (5e+03,1e+04]   (4.5e+03,5e+03] (5e+03,1e+04]   (4.5e+03,5e+03]
 [9] (4.5e+03,5e+03] (4e+03,4.5e+03] (4.5e+03,5e+03] (4e+03,4.5e+03]
  :
 [49] (4e+03,4.5e+03] (4.5e+03,5e+03]
Levels: (0,4e+03] (4e+03,4.5e+03] (4.5e+03,5e+03] (5e+03,1e+04]
```

```
> state77$Income.cat <- cut(state77$Income, c(0, 4000, 4500, 5000, 10000),
+ labels=c("<= 4000", "4001--4500", "4501--5000", ">5000"))
> head(state77, 10)
```

```
            Population Income Illiteracy LifeExp Murder HSGrade Frost   Area
Alabama           3615   3624        2.1   69.05   15.1    41.3    20  50708
Alaska             365   6315        1.5   69.31   11.3    66.7   152 566432
Arizona           2212   4530        1.8   70.55    7.8    58.1    15 113417
Arkansas          2110   3378        1.9   70.66   10.1    39.9    65  51945
California       21198   5114        1.1   71.71   10.3    62.6    20 156361
Colorado          2541   4884        0.7   72.06    6.8    63.9   166 103766
Connecticut       3100   5348        1.1   72.48    3.1    56.0   139   4862
Delaware           579   4809        0.9   70.06    6.2    54.6   103   1982
Florida           8277   4815        1.3   70.66   10.7    52.6    11  54090
Georgia           4931   4091        2.0   68.54   13.9    40.6    60  58073
               Region Income.cat
Alabama         South    <= 4000
Alaska           West      >5000
Arizona          West 4501--5000
Arkansas        South    <= 4000
California       West      >5000
Colorado         West 4501--5000
Connecticut Northeast      >5000
Delaware        South 4501--5000
Florida         South 4501--5000
Georgia         South 4001--4500
```

5. La relació entre dues variables categòriques es pot presentar mitjançant taules de contingència. Aquestes poden mostrar la distribució conjunta o la distribució condicional d'una de les dues variables en funció de l'altra:

```
> with(state77, table(Region, Income.cat))
```

```
               Income.cat
Region          <= 4000 4001--4500 4501--5000 >5000
  Northeast           2          2          3     2
  South              10          2          3     1
  North Central       0          4          6     2
  West                1          3          6     3
```

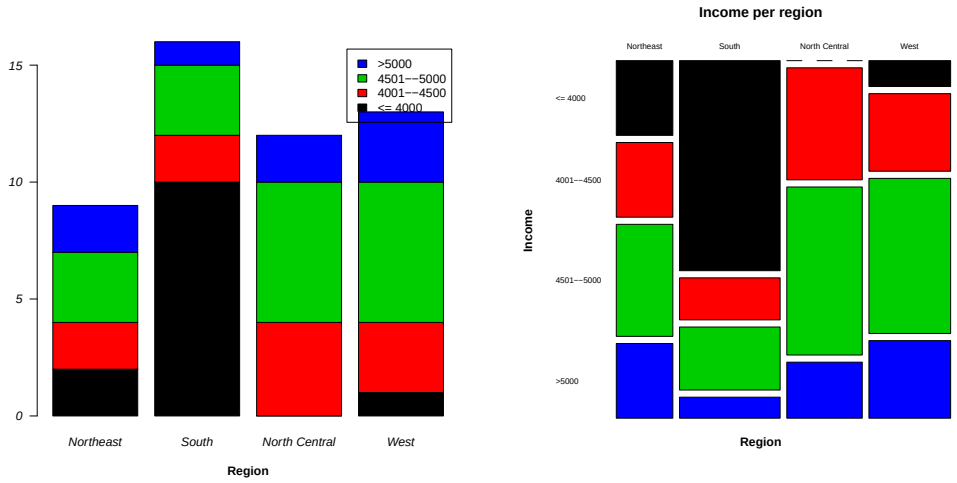
```
> library(descr)
> with(state77, CrossTable(Region, Income.cat, prop.c = F, prop.t = F,
+ prop.chisq = F, format='SPSS'))
```

```
=================================================================
                Income.cat
Region          <= 4000   4001--4500   4501--5000    >5000    Total
-----------------------------------------------------------------
Northeast             2            2            3        2        9
                 22.222       22.222       33.333   22.222   18.000
-----------------------------------------------------------------
South                10            2            3        1       16
                 62.500       12.500       18.750    6.250   32.000
-----------------------------------------------------------------
North Central         0            4            6        2       12
                  0.000       33.333       50.000   16.667   24.000
-----------------------------------------------------------------
West                  1            3            6        3       13
                  7.692       23.077       46.154   23.077   26.000
-----------------------------------------------------------------
Total                13           11           18        8       50
=================================================================
```

Sembla existir una relació entre les dues variables?

6. A més a més es poden fer diferents diagrames –un diagrama de barres o un diagrama de mosaic– per visualitzar aquesta relació:
   ⇝ Transparències 18, 19 i 25 i 28 de `ED.ppt`

```
> windows(height=9, width=18)
> par(mfrow=c(1, 2), las=1, font.lab=2, font.axis=3)
> with(state77, barplot(table(Income.cat, Region), col=1:4, legend=T, xlab="Region"))
> mosaicplot(Region~Income.cat, data=state77, col=1:4, ylab="Income", main="Income per region")
```



   Interpreteu els dos diagrames.

4. Exercicis

`Nota:` Copieu tant les preguntes com les instruccions d'R següents i pegueu-les a un document WORD. A continuació completeu el document WORD amb les instruccions completes, els resultats i vostres comentaris.

1. Tornem a treballar amb les dades dels dos gestors de bases de dades:

```
> source(url("http://www-eio.upc.es/teaching/pe/read-data"))
```

```
> head(DAT)
```

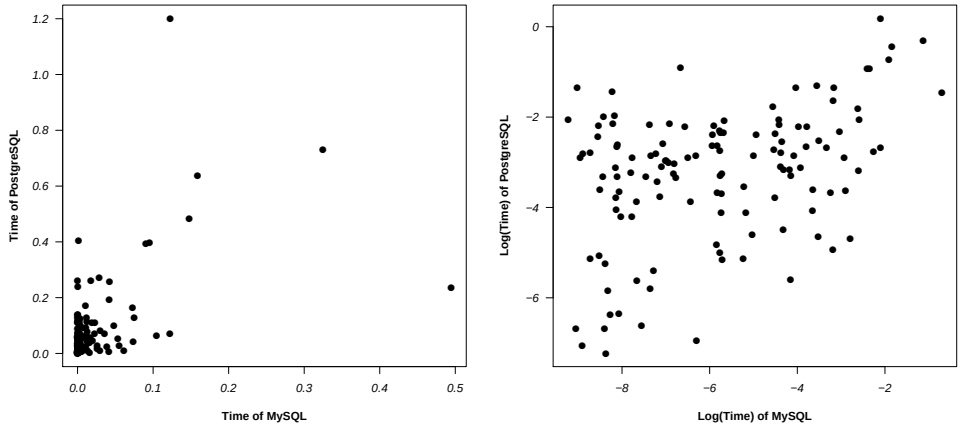
```
     op        mysql        post
1 INSERT 0.0003011227 0.036271811
2 INSERT 0.1224038601 0.069139004
3 INSERT 0.0002570152 0.001724005
4 INSERT 0.0016028881 0.020658970
5 INSERT 0.0480089188 0.098726988
6 INSERT 0.0015101433 0.055121899
```

```
> summary(DAT)
```

```
     op         mysql               post          
 DELETE:41   Min.   :0.0000980   Min.   :0.0007269  
 INSERT:24   1st Qu.:0.0004218   1st Qu.:0.0207570  
 SELECT:36   Median :0.0029399   Median :0.0518451  
 UPDATE:29   Mean   :0.0225864   Mean   :0.0921243  
             3rd Qu.:0.0166588   3rd Qu.:0.0987270  
             Max.   :0.4947410   Max.   :1.2009261  
                                 NA's   :1          
```

2. Estudieu la relació entre les dues variables numèriques fent un diagrama de dispersió i feu un altre per a les variables transformades amb logaritme.

```
> windows(width=10, height=5)
> par(mfrow=c(1, 2), las=1, font.lab=2, font.axis=3)
> plot(post~mysql, DAT, xlab="Time of MySQL", ylab="Time of PostgreSQL", pch=16, cex=1.3)
> plot(log(post)~log(mysql), DAT, xlab="Log(Time) of MySQL",
+ ylab="Log(Time) of PostgreSQL", pch=16, cex=1.3)
```



Què hi podeu observar?

3. Calculeu la correlació de les dues parelles de variables i n'interpreteu els seus valors:

```
> with(DAT, round(cor(mysql, post), 3))

[1] NA

> with(DAT, round(cor(mysql, post, use="c"), 3))

[1] 0.528

> with(DAT, round(cor(log(mysql), log(post), use="c"), 3))

[1] 0.4
```

4. Creeu una variable ordinal amb els temps de `post` utilitzant com punts de tall els tres quartils. A continuació feu una taula de contingència amb la nova variable i la variable `op`.

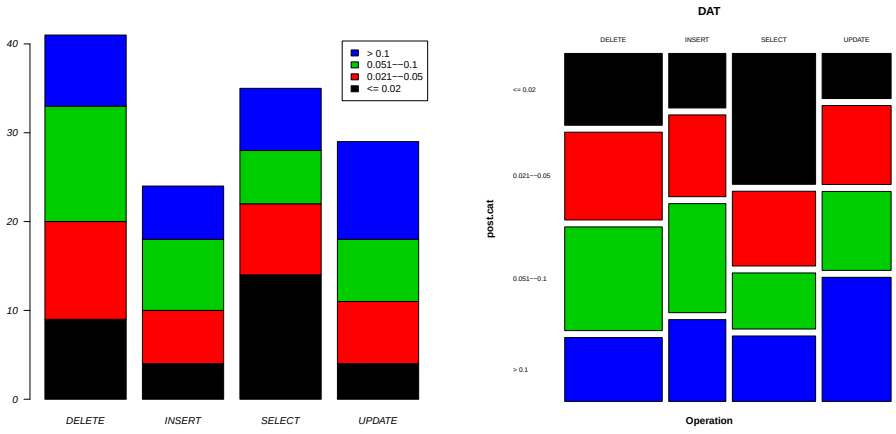
```
> summary(DAT$post)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
0.0007269 0.0207600 0.0518500 0.0921200 0.0987300 1.2010000         1

> DAT$post.cat <- cut(DAT$post, c(0, 0.02, 0.05, 0.1, 2),
+ labels=c("<= 0.02", "0.021--0.05", "0.051--0.1", "> 0.1"))
> summary(DAT)

      op         mysql               post                post.cat
 DELETE:41   Min.   :0.0000980   Min.   :0.0007269   <= 0.02    :31
 INSERT:24   1st Qu.:0.0004218   1st Qu.:0.0207570   0.021--0.05:32
 SELECT:36   Median :0.0029399   Median :0.0518451   0.051--0.1 :34
 UPDATE:29   Mean   :0.0225864   Mean   :0.0921243   > 0.1      :32
             3rd Qu.:0.0166588   3rd Qu.:0.0987270   NA's       : 1
             Max.   :0.4947410   Max.   :1.2009261
                                 NA's   :1
```

```
> with(DAT, CrossTable(op, post.cat, prop.c = F, prop.t = F, prop.chisq = F, format='SPSS'))

   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

============================================================
          post.cat
op        <= 0.02   0.021--0.05   0.051--0.1    > 0.1    Total
------------------------------------------------------------
DELETE          9             11           13        8       41
          21.951        26.829       31.707   19.512   31.783
------------------------------------------------------------
INSERT          4              6            8        6       24
          16.667        25.000       33.333   25.000   18.605
------------------------------------------------------------
SELECT         14              8            6        7       35
          40.000        22.857       17.143   20.000   27.132
------------------------------------------------------------
UPDATE          4              7            7       11       29
          13.793        24.138       24.138   37.931   22.481
------------------------------------------------------------
Total          31             32           34       32      129
============================================================
```

Sembla haver-hi diferències entre les operacions pel que fa als temps que triga `post`?

5. Feu una representació gràfica de la taula de contingència anterior amb un diagrama de barres i també amb un diagrama de mosaic.
   ⇝ Veure també les pàgines 56 i 57 del tutorial

```
> windows(width=10, height=5)
> par(mfrow=c(1, 2), las=1, font.lab=2, font.axis=3)
> with(DAT, barplot(table(post.cat, op), legend=T, col=1:4))
> mosaicplot(op~post.cat, DAT, xlab="Operation", col=1:4)
```



6. Executeu l'script `letsmakeadeal.R` i comenteu el resultat. És el que heu esperat?

```
> source("Letsmakeadeal.R")

Success proportion without changing the door
[1] 0.3387
Success proportion changing the door
[1] 0.6613
```