Capítol 4

Data Frames

Objectius: Introduir l'estructura `dataframe` de R i el format i la seva utilització en R.

4.1 Exercicis

1. Tenim un dataframe amb dades dels estudiants de Programació del Grau d'Estadística. El dataframe `D` té, per a cada estudiant, les dades següents:

   ```
   D: nom  niub  nota_prog  nota_mitjana
   ```

   Fes una funció que, donat aquest dataframe `D` i la nota mitjana històrica de programació a l'escola `MeanProg`, retorni un vector amb els noms dels estudiants que compleixin que la seva nota de programació sigui més gran que la seva pròpia mitjana i també que la mitjana històrica de programació.

2. Tenim un data frame amb els resultats de les 5 últimes eleccions al Parlament de Catalunya on la primera columna conté els anys en que hi va haver eleccions i les columnes següents són els resultats obtinguts pels 5 primers partits (per aquest ordre, CIU, ERC, PSC, PP i ICV) en cada un dels anys. Sempre rebent com a entrada el dataframe `eleccions`, dissenyeu:

   (a) Una funció que retorni quin any el PP va obtenir el seu pitjor resultat i quin resultat va ser. Podeu retornar les dues dades en un vector.

   (b) Una funció que retorni quin any la diferència d'escons entre CIU i PSC va ser més gran i quina va ser aquesta diferència.

   (c) Una funció que retorni quin partit va obtenir més escons l'any 2003 i quants escons van ser.

   (d) Una funció que retorni un nou dataframe amb els anys de les eleccions i el partit que va obtenir més escons en cada una.

3. Dissenyeu una funció que rebi un dataframe i el nom d'una de les seves variables i retorni la mitjana d'aquesta variable (amb arrodoniment a 4 xifres decimals).

4. Dissenyeu una funció que rebi un dataframe i el nom d'una de les seves variables i retorni la mediana d'aquesta variable.

5. Dissenyeu una funció que rebi dues variables d'un dataframe i en retorni la seva correlació. Podeu usar la funció `mean()`.

6. *Segons mitjana (Examen de recuperació 13/14).* Feu una funció que donats un data frame `dades` i el nom de dues variables (columnes) `var1` i `var2` modifiqui el data frame `dades` afegint una columna `Result` de forma que el valor corresponent en aquesta nova columna es calculi de la següent manera:

   > Si en la fila corresponent el valor de `var1` és superior a la mitjana d'aquesta mateixa variable (mitjana de la columna `var1`), el resultat serà el valor de `var1` + el valor de `var2`. Si per contra, en la fila corresponent el valor de `var1` no és superior a la mitjana d'aquesta mateixa variable (mitjana de la columna `var1`), el resultat serà el valor de `var1` – el valor de `var2`.

   Podeu usar qualsevol de les funcions que ofereix R que necessiteu.

   Per exemple, si tenim el data frame

```
  v1 v2 v3 v4
1  1  2  3  4
2  2  2  3  2
3  2  4  2  3
4  4  2  4  2
```

   i les variables són `"v4"` i `"v1"`, la mitjana de la columna amb nom `"v4"` és 2.75 i per tant el data frame resultant ha de ser:

```
  v1 v2 v3 v4 Result
1  1  2  3  4      5
2  2  2  3  2      0
3  2  4  2  3      5
4  4  2  4  2     -2
```

7. *Sobre mitjana (Examen fianal 13/14).* Feu una funció que donats un data frame `dades` i el nom d'una variable (columna) `var` retoni un altra data frame que contingui totes les files de `dades` tals que el valor corresponent a la variable `var` sigui superior a la mitjana d'aquesta mateixa variable (columna).

   Podeu usar qualsevol de les funcions que ofereix R que necessiteu.

   Per exemple, si tenim el data frame

```
   v1  v2  v3  v4
1   1   2   3   4
2   2   2   3   2
3   2   4   2   3
4   4   2   4   2
```

i la variable és "v4", la mitjana de la columna amb nom "v4"és 2.75 i per tant el data frame resultant ha de ser:

```
   v1  v2  v3  v4
1   1   2   3   4
3   2   4   2   3
```

8. El dataframe `trees` conté dades sobre l'alçada, el diàmetre i el volum de 31 cirerers talats. Rebent com a entrada `trees`,

```
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
3    8.8     63   10.2
    ...
30  18.0     80   51.0
31  20.6     87   77.0
```

   (a) Feu una funció que retorni quants arbres tenen alçada (Height) més gran que 85 i diàmetre (Girth) més gran que 10.

   (b) Afegiu al dataframe les dades de 2 arbres nous amb variables:

```
# Arbre 1: Girth=9.2, Height=80 i Volume=11.2
# Arbre 2: Girth=14, Height=69 i Volume=33.6
```

   (c) Feu una funció que retorni quin volum té l'arbre més alt. Utilitzeu-la per saber si és el volum més alt de tots els arbres.

9. El dataframe `iris` conté dades sobre les dimensions (longitud i amplada) dels sèpals i pètals, i l'espècie de 150 flors. Rebent com a entrada `iris`:

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
     ...
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
     ...
149          6.2         3.4          5.4         2.3  virginica
150          5.9         3.0          5.1         1.8  virginica
```

(a) Dissenyeu una funció que rebi el dataframe `iris`, el nom d'una de les seves variables i el nom d'una de les seves espècies i retorni la mitjana de la variable per a l'espècie indicada.

(b) Dissenyeu una funció que rebi el dataframe `iris` i retorni un altra dataframe amb les mitjanes de cada variable de flor en cada espècie, és a dir es busca un dataframe amb les dades següents: (pots considerar sabut que només hi ha 3 espècies diferents)

```
       Specie  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
1      setosa    meanSL-set   meanSW-set    meanPL-set   meanPW-set
2  versicolor    meanSL-ver   meanSW-ver    meanPL-ver   meanPW-ver
3   virginica    meanSL-vir   meanSW-vir    meanPL-vir   menaPW-vir
```

10. El dataframe `Traffic` conté dades d'accidents de trànsit en els anys 1961 i 1962 a Suècia.

    L'heu de carregar de la llibreria MASS amb el codi següent:

```
library(MASS)
data(Traffic)
```

    I les dades són com segueix:

```
     year day limit  y
1    1961   1    no  9
2    1961   2    no 11
3    1961   3    no  9
    ...
183 1962  91   yes 15
184 1962  92   yes  9
```

    (a) Feu una funció que rebi el dataframe `Traffic` i retorni la diferència entre la mitjana dels dies en què hi ha límit de velocitat i la dels dies que no.

    (b) Feu una funció que rebi l'any (1961 o 1962) i retorni el màxim d'accidents d'aquest any i si el dia en que es va donar aquest màxim hi havia límit de velocitat o no.

11. Tenim un dataframe amb dades sobre els desplaçaments d'estudiants universitaris per comarques (fitxer desplacaments_uni.txt). Tenint en compte que el total d'universitaris atrets per una comarca és la suma dels interns, dels des de fora i dels que no consta i que el total d'estudiants generats són els interns i els que marxen a fora, es demana

    (a) Feu una funció que retorni quina comarca atreu més estudiants.

    (b) Feu una funció que retorni quina comarca genera menys estudiants.

(c) Dissenyeu una funció que retorni quants desplaçaments interns té la comarca que genera més estudiants.

(d) Dissenyeu una funció que retorni quina és la diferència més gran entre universitaris atrets i generats.

(e) Feu una funció que retorni quantes comarques hi ha que no atreguin cap estudiant.

12. *Analfabetisme vs. salari (Examen final única 13/14).* Fes una funció que donat el data frame `estats`, descrit a continuació, retorni cert si i només si l'estat que té el mínim índex d'analfabetisme (variable `Analfabetisme`) és el mateix que té el màxim salari (variable `Salari`).

    Podeu considerar que no hi han valors iguals.

    El data frame `estats` té les dades en el format següent:

```
          Poblacio  Salari  Analfabetisme  Esp_vida          Regio
Alabama       3615    3624            2.1     69.05          South
Alaska         365    6315            1.5     69.31           West
    .....
Wisconsin     4589    4468            0.7     72.48  North_Central
Wyoming        376    4566            0.6     70.29           West
```

13. Disposem de dades sobre els matrimonis entre dones, entre homes i heterosexuals en cada mes de l'any 2012 (fitxer matrimonis.txt). Es demana:

    (a) Dissenyeu una funció que retorni si en algun mes el nombre de matrimonis entre dones va ser més gran que el de matrimonis entre homes.

    (b) Dissenyeu una funció que retorni quin mes la proporció de matrimonis entre parelles del mateix sexe va ser més gran respecte el total de matrimonis.

    (c) Dissenyeu una funció que retorni quantes persones es van casar el mes que es va casar més gent.

14. Es té un `dataframe` amb els resultats de la Lliga d'un cert any. A cada fila hi ha els resultats d'un equip (amb els equips ordenats per ordre alfabètic) i la informació que tenim per a cada equip és: Partits Jugats (PJ), Partits Guanyats (PG), Partits Empatats (PE), Gols a Favor (GF) i Gols en Contra (GC). Fent sempre funcions que rebin com a entrada el `dataframe` amb les dades de la lliga, feu

    (a) Una funció que retorni el nom de l'equip que ha guanyat la Lliga.

(b) Una funció que retorni quins equips han baixat a Segona Divisió. Recordeu que els equips que baixen a Segona Divisió són els tres últims de la classificació.

(c) Una funció que retorni quin equip té la diferència més petita entre gols marcats i gols rebuts.

(d) Una funció que retorni `TRUE` o `FALSE` segons si en la classificació final hi ha hagut equips empatats a punts o no.

(e) Una funció que rebi els noms de dos equips i retorni la diferència de punts entre ells.

(f) Una funció que rebi els noms de dos equips i retorni quantes posicions els van separar.

15. Disposem d'un dataframe amb dades sobre el nombre d'adopcions a Catalunya segons el país d'origen.

(a) Una funció que rebi el dataframe i retorni de quin pais hi va haver més adopcions entre els anys 2005 i 2010.

(b) Dissenyeu una funció que tingui com a entrada el dataframe, dos anys (entre 2005 i 2012) i un país i retorni la diferència entre les adopcions en aquell país entre els dos anys en valor absolut.

(c) Dissenyeu una funció que tingui com a entrada el dataframe i un any i retorni quantes adopcions d'Àfrica hi va haver aquell any.

(d) Feu una funció que rebi com a entrada el dataframe i retorni un altre dataframe on cada fila sigui un any i cada columna un continent, obtenint així quantes adopcions hi va haver en cada any de cada continent.

(e) Utilitzeu la funció definida en l'apartat anterior per construir-ne una altra que rebi el dataframe original i un any i retorni de quin continent va haver-hi més adopcions aquell any.

16. *Guanys esportius (Examen final 14/15).* Tenim dos dataframes amb informació dels tornejos guanyats enguany. En el primer dataframe `D1` tenim informació de l'equip que ha guanyat cada torneig:

```
D1:  equip  torneig
```

I en el segon dataframe `D2` tenim la informació dels diners que es guanyen amb cada torneig (el campió del torneig):

```
D2:  torneig recompensa
```

Es demana implementar una funció que, donats els dos dataframes `D1` i `D2` generi un tercer dataframe `D3` que contingui la classificació dels equips pels seus guanys, de manera que per a cada equip tindrem la suma de tots els seus guanys (segons els diferents tornejos que hagi guanyat) i a més els tindrem ordenats de més guanys a menys guanys.

```
D3:  equip  guanys
```

Per exemple, si tenim:

```
D1:       equip     torneig          D2:     torneig   recompensa
    1     Barça    Lliga-1a              1  Lliga-1a         1500
    2    Madrid   Lliga-ACB              2 Lliga-ACB          450
    3    Girona    Lliga-2a              3   CopaRey          500
    4    Girona      Gamper              4  Lliga-2a          150
    5     Barça     CopaRey              5 Champions         3000
    6     Barça   Champions              6 FinalFour         1000
    7  Juventud   FinalFour              7    Gamper          200
```

El resultat hauria de ser:

```
D3:       equip    guanys
    1     Barça      5000
    2  Juventud      1000
    3    Madrid       450
    4    Girona       350
```

17. Disposem de dos dataframes que contenen informació sobre els diferents estats dels Estats Units. Cada fila correspon a un estat i cada columna a una variable. Per més informació de les dades podeu consultar la help de R mitjançant

    Després de llegir les dades

    es demana

    (a) Dissenyeu una funció que rebi com a entrada un dels dataframes i retorni quina és la regió de l'estat amb l'àrea més gran.

    (b) Amb la informació obtinguda en l'apartat anterior (i, per tant, utilitzant la funció que heu fet), dissenyeu una funció que rebi els dos dataframes i calculi la mitjana de població per als estats de la regió obtinguda en l'apartat anterior.

18. Disposem de tres dataframes. En el fitxer `estudiants.txt` hi ha informació sobre estudiants i el curs al qual estan matriculats, en `cap_aules.txt` hi trobem informació sobre aules i la seva capacitat i `aules_curs.txt` relaciona els cursos amb les aules on s'imparteixen.

    (a) Feu una funció que retorni els noms dels estudiants matriculats en els cursos que es fan a l'aula A1.

    (b) Tenint en compte el nombre d'estudiants matriculats a cada curs, feu una funció que retorni quins cursos estan programats en aules que no tenen prou capacitat.

        Tenim ara un altre dataframe `curs_coord` amb la informació sobre els coordinadors de cada curs i la seva nacionalitat.

    (c) Dissenyeu una funció que retorni quina és la nacionalitat del professor del curs amb menys alumnes matriculats.

    (d) Feu una funció que donada una aula retorni els noms dels professors que fan classe en aquella aula.

    (e) Utilitzant la funció de l'apartat b), feu una funció que retorni els noms dels professors que hauran de buscar aules amb més capacitat pels seus cursos.

19. *Despeses de la compra (Examen de recuperació 14/15)* Tenim dos dataframes amb informació dels productes que ha comprat cada persona i el preu d'aquests productes. En el primer dataframe `D1` tenim informació de la persona i el que ha comprat (quantitat):

    ```
    D1:  nom  producte  Kg
    ```

    I en el segon dataframe `D2` tenim la informació del que costa cada producte per Kg:

    ```
    D2:  producte  preuKg
    ```

    Es demana implementar una funció que, donats els dos dataframes `D1` i `D2` generi un tercer dataframe `D3` que contingui la despesa de la compra, de manera que per a cada persona tindrem la suma de la despesa de tots els productes que ha comprat. Aquest dataframe resultat ha d'estar ordenat per nom.

    ```
    D3:  nom  despesa
    ```

    Per exemple, si tenim:

```
D1:        nom    producte   Kg     D2:    producte   preuKg
    1      Pep       pomes    5         1     pomes      2.5
    2      Kim       peres    3         2     peres      3.0
    3      Pep       peres  1.5         3   platans      4.5
    4    Maria       pomes    3         4   cireres      6.0
    5    Maria     platans    2         5    figues      5.5
    6   Albert     cireres  3.5
    7      Pep      figues    5
```

El resultat hauria de ser:

```
D3:        nom   despesa
    1   Albert        21
    2      Kim         9
    3    Maria      16.5
    4      Pep      44.5
```

20. Disposem de dos dataframes amb informació sobre clients en un supermercat. El primer dataframe, que està en el fitxer `cl_prod.txt` conté informació sobre els productes de cada client. El segon, `caixes`, conté les caixes, el nombre de productes que té assignat cada caixa en aquest moment i el màxim de productes que pot tenir.

    (a) Donat un número de client, feu una funció que retorni a quines caixes podria anar.

    (b) Tenint en compte que les caixes sempre s'han d'omplir per ordre (és a dir, un client no pot anar a la caixa 3 si pot anar a la caixa 2), feu una funció que actualitzi el dataframe `caixes` amb els clients que tenim. La funció ha de retornar el dataframe actualitzat i un vector amb els clients que no han pogut ser col·locats en cap caixa.

21. Per tal d'estudiar l'efecte que causa el tabac sobre l'agregació de les plaquetes en la sang, Levine (1973) va disenyar un estudi per mesurar el grau al qual les plaquetes es van agregar en individus abans i després de fumar. Les plaquetes estan implicades en la formació de coàguls de sang, i se sap que els fumadors pateixen més sovint de desordres que impliquen la seva formació que no pas les persones no fumadores.

    Es van fer dos estudis de l'agregació de plaquetes, amb els mateixos individus, abans i després de fumar tabac i abans i després de fumar enciam. Les dades de totes dues taules les teniu als arxius `tabac.txt` i `enciam.txt` respectivament.

    Les dades d'aquests dos arxius tenen el format següent (tots dos iguals):

Persona	Abans	Despres	Diferencia
1	25	24	-1
2	25	30	5
...	...	...	...
10	60	62	2
11	68	70	2

Si denotem per `X` i `Y` les variables aleatòries que mesuren el percentatge màxim d'agregació de plaquetes a la sang abans i després de fumar un cigarret d'enciam, respectivament columnes `Abans` i `Despres` en l'arxiu, es demana

(a) Feu una funció que donat un dels data frames anteriors, comprovi si la mitjana de la diferència és la diferència de les mitjanes de les variables `X` i `Y` i retorni un List amb la següent informació: Un booleà que serà cert o fals responent a la pregunta anterior i la mitjana de la diferència que s'ha calculat. No podeu usar la funció `mean` de R.

(b) Feu una funció que donat un dels data frames anteriors, calculi i retorni les variàncies de `X`, de `Y` i de la diferència. Podeu retornar els tres valors en un únic vector.

Recordeu que la fórmula per calcular la variància d'una variable `X` és

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Es recomana que definiu una funció que us retorni la variància d'una certa variable donada (variància d'un vector de valors).

(c) Feu una funció que donat un dels data frames anteriors i usant la crida a la funció de l'apartat b), comprovi la següent fórmula i retorni cert o fals depenent de si la fórmula es compleix o no.

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$

Recordeu que la fórmula per calcular la covariància entre dues variables `X` i `Y` és

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

(d) Feu una funció que donat un dels data frames anteriors i usant les funcions anteriors que us siguin necessàries, retorni el valor de la mesura de discrepància.

$$d = \frac{\bar{D}}{S_D/\sqrt{n}}$$
$$S_D = \sqrt{Var(D)}$$

22. Basant-nos en els mateixos arxius que l'exercici anterior, `tabac` i `enciam`, es demana fer una funció que donats tots dos data frames i usant totes les funcions necessàries de l'exercici anterior, calculi i retorni el valor de la mesura de discrepància entre la diferència d'abans i després de fumar tabac i la diferència d'abans i després de fumar enciam (cal calcular la diferència d'interès $D = D_{Tabac} - D_{Enciam}$).