Visualització de dades (Enginyeria de Dades – EE - UAB)
Examen Recuperació Segon Parcial – 5 Juliol 2021
SOLUCIONS MODEL A

Nom i Cognom:_____

NIU:_____ Grup de Matrícula:_____

Només es permet l'ús d'internet per l'accés al campus virtual en el moment de descarregar el full d'enunciats y d'entregar l'examen.

Sólo se permite el uso de internet para el acceso al campus virtual en el momento de descargar la hoja de enunciados y de entregar el examen.

PARTE 1 (2.5 pt)

1.1. (0.5 pt) Explica brevemente qué es una escala de color perceptualmente correcta y qué ventajas tiene.

RESPOSTA:

Son escalas que tienen en cuenta cómo nuestro cerebro procesa el color y en las que los colores varían de forma gradual e uniforme. Una ventaja es que estas variaciones entre colores se asocian mejor a las variaciones graduales e uniformes entre los valores del dataset.

1.2. (1 pt) Imagina que tenemos un dataset de precipitaciones (lluvia) en Europa y queremos visualizarlo en un mapa. La precipitación es un atributo cuantitativo continuo con valores min=0 y max=10 mm/h. ¿Qué escala de color usarías?

Di qué tipo de escala es, justifica la elección de colores, y hazla en R.

RESPOSTA:

Una escala secuencial continua. Los colores más adecuados por semántica y significado podrían ser una gama de azules.

1.3. (1 pt) ¿Qué escala de color usarías para representar la velocidad en X del viento, teniendo en cuenta que tiene valores min=-15 y max=50 m/s?

Di qué tipo de escala es, justifica la elección de colores, y hazla en R o sube una imagen.

RESPOSTA:

Escala divergente o cuantitativa divergente. El color neutro (blanco) debería corresponder al 0 en los datos y habrá dos colores distintos en cada extremo.

PART 2 (3.5 pt) *Dataset:*

beers.csv

Aquest dataset té 7 atributs per 44 cerveses de Minnesota i el farem servir en l'exercici 2.1. Els atributs són: Nom de la cerveseria, cervesa, descripció, estil, alcohol en volum (ABV), unitats internacionals d'amargor (IBU), puntuació que han rebut

Dataset: statsNBA2008.csv

Aquest dataset té 21 atributs de 50 jugadors de la NBA. Recull les estadístiques de la NBA del 2008. I estudiarem la relació entre alguns dels atributs en l'exercici 2.2

2.1 (2.25 pt) És estiu, esteu a Minnesota i voleu anar a prendre una cervesa. Utilitzant el dataset *beers.csv*.

a) Feu un mapa d'arbre (*treemap*) que us permeti contestar les preguntes de l'apartat b. Podeu fer-lo en R o en qualsevol altre llenguatge, utilitzant les llibreries que us semblin convenients. Ara bé, heu d'explicar-ho bé: comenceu narrant com és un *treemap* en general i per quin tipus de dades serveix i redacteu els passos que heu de fer per construir la visualització triada. (1.5 pt)

b) Esteu amb una persona que li puja molt l'alcohol i vol una cervesa amb poc alcohol per volum. Si esteu a la cerveseria Bauhaus, quin tipus de cervesa (Lager, Ale o IPA) li demaneu?. I si esteu en la cerveseria Summit, entre quines tres cerveses podeu triar per demanar-li (doneu el nom de les 3 cerveses)? (0.25 pt)

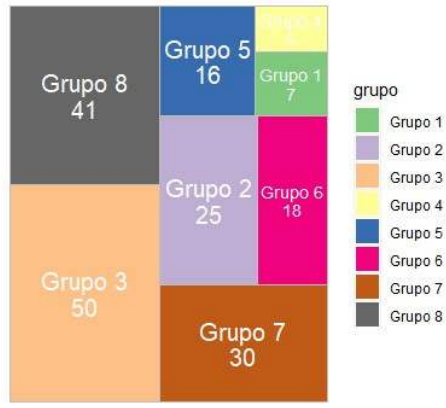
c) Expliqueu quan usem un mosaic enlloc d'un gràfic de barres apilades, i quan no podem utilitzar un mosaic i usem un treemap. (0.25 pt)

d) Quin tipus d'escala de color, de les de la pàgina següent, seria la més òptima i perquè? (0.25 pt) Paleta 1

Paleta 2

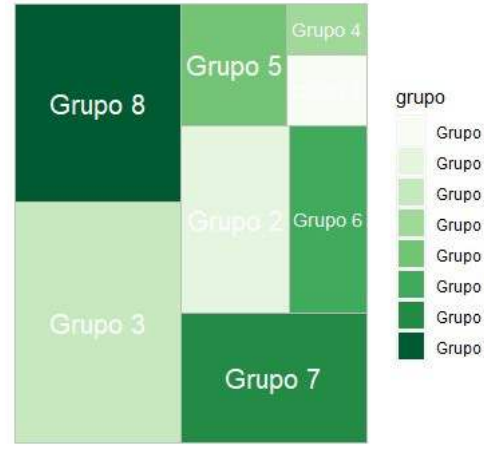
Paleta 3

Raona la resposta:



Paleta 1

Paleta 2



Paleta 3

RESPOSTA:

a) Com vam veure a classe, un mapa d'arbre és un dibuix rectangular dividit en caselles, i cada casella representa una sola observació. Vam veure que era una bona manera de mostrar dades jeràrquiques mitjançant rectangles imbricats. I l'àrea relativa de cada casella expressava una variable contínua. També vam veure que era òptim quan hi ha com a màxim dues variables d'agrupació, per tant no en definirem més.

Quan s'especifiquen les proporcions d'acord amb múltiples variables d'agrupació, les representacions els mapes d'arbres són aproximacions de visualització útils. Per tant, una possibilitat amb aquest dataset seria:

- *Per definir el color i actuar doncs com un 'grup pare' utilitzaríem l'estil /style (Agafem aquest com grup de color, ja que sol hi ha 3 estils – Lager, Ale, IPA- i serà més fàcil entendre la llegenda de la visualització que si posem el nom de la cervesa, que té molts més nivells).*
- *Com a 'subgrup' utilitzaríem la cerveseria (variable amb 8 nivells).*
- *Com a variable que descrigui l'àrea de les caselles triarem per exemple ABV. Aquesta tria ve donada perquè ens estant preguntant sobre el alcohol en volum de les cerveses i l'àrea de les caselles necessita d'una variable numèrica continua.*
- *Finalment, com a 'label'/nivell, escollirem el nom de la cervesa.*

Si ho fem amb R, el primer que hem de fer és carregar les llibreries necessàries. La llibreria especifica aquí és treemapify (abans necessitarem tenir instal·lat el paquet com vam veure al seminari 7). També haurem de fer us de geom_treemap. Per fer un treemap basic doncs:

```
> library(ggplot2)
> library(treemapify)
> Beers <-read_csv('C:/DATA/beers.csv')
> ggplot(Beers, aes(area=ABV, fill=Style, label=Beer, subgroup=Brewery))+geom_treemap()+ geom_treemap_subgroup_border(colour = "black",size=3) +geom_treemap_subgroup_text(alpha=0.5,colour = "white")+geom_treemap_text (aes(label=Beer))
```

b)

En la cerveseria Bauhaus, la cervesa IPA sembla tenir més ABV (àrea major) que les Lager, per tant demanaria una Lager.

En la cerveseria Summit, les tres cerveses amb menys ABV són la Pilsener, la Oatmeal Stout i la Extra Pale Ale

c) Els mosaics són similars als stacked bar o gràfic de barres apilades, però en el cas dels mosaics, l'amplada i l'alçada de les àrees individuals poden variar. En el gràfic de barres apilades només varia una d'aquestes, normalment l'alçada.

Ara bé, els mosaics assumeixen que tots els nivells d'una variable d'agrupament es poden combinar amb tots els nivells d'una altra variable d'agrupació. El treemap no fa aquesta assumpció.

(d) Estem agrupant variables categòriques. La paleta 2 és per variables continues, la paleta 3 és per variables contínues discretitzades. La paleta 1 en canvi, presenta colors diferents per cada grup/categoria, i és per tant la paleta òptima

2.2. (1.25 pt) Voleu estudiar algunes relacions entre el tipus de jugades de basquet i extreure conclusions de si això porta als equips a guanyar o perdre. Per això teniu el dataset *statsNBA2008.csv*.

(a) Feu una tibble amb les 10 variables detallades a sota i estudieu la relació entre elles sense reduir-ne el número, tot fent una visualització. Un cop tingueu la visualització mostreu només aquella part que us mostra la informació d'interès. Podeu fer-ho en qualsevol llenguatge, però expliqueu perquè heu escollit aquesta visualització, si la vostra elecció està relacionada amb el tipus de dades que teniu, els passos que heu seguit per fer-la i extrèieu alguna/es conclusió/ns del vostre gràfic.

- **PTS (percentatge de punts de l'equip, punts per joc)**
- **FGM (cistelles de camp realitzats, percentatge de cistelles de camp)**
- **FGA (tirs de camp intentats, número d'intencions de cistelles de camp)**
- **FTP (percentatge de tirs lliures)**
- **DRB (rebots defensius)**
- **ORB (rebots ofensius)**
- **TRB (rebots totals)**
- **3XPM /3PM (cistelles de camp de 3 punts anotats, tirs de 3 punts)**
- **3XPA/3PA (rebots defensius, tirs de tres punts)**
- **3XPP/3PA (percentatge de tirs de camp de 3 punts)**

Nota: Segons les versions us sortirà la X o no en les últimes 3 variables (és a dir, per exemple, alguns tindreu 3XPM i altres 3PM) (1 pt)

(b) Quin gràfic podríeu fer si volguéssiu reduir el nombre de variables però no fos possible identificar les variables que es podrien eliminar a simple vista, i volguéssiu a la vegada, assegurar que les vostres variables són independents unes de les altres? (no cal fer la visualització, només contestar la pregunta) (0.25pts)

RESPOSTA: *Les matrius de correlació mostren els coeficients de correlació/relacions entre un nombre relativament gran de variables contínues. Com tenim 10 variables, podem fer una matriu de correlació per veure les relacions entre totes elles.*

Primer ens preparem el dataframe que ens interessa, vam veure al seminari 7 les tibles de R i com utilitzar-les. Una manera que es podria fer servir, seria

```
> NBA_tb <- as_tibble(NBA)
> NBA_sel<-NBA_tb[,c('PTS','FGM','FGA', 'FTP', 'DRB','ORB','TRB',
'3PM','3PA','3PP' )]
```

També ho podríem fer utilitzant select de dplyr com havíem vist en classes anteriors.

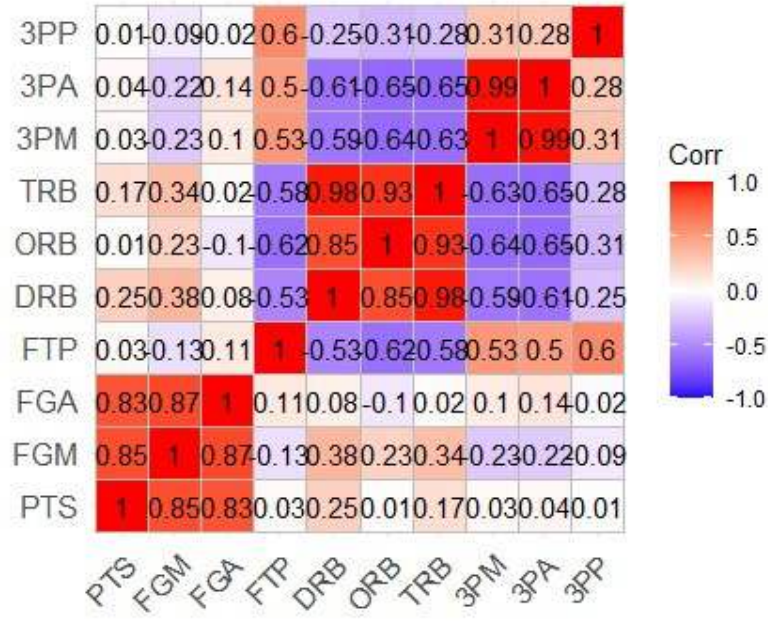
Necessitarem a més de la tidyverse, la llibreria ggcorrplot

```
> library(ggcorrplot)
```

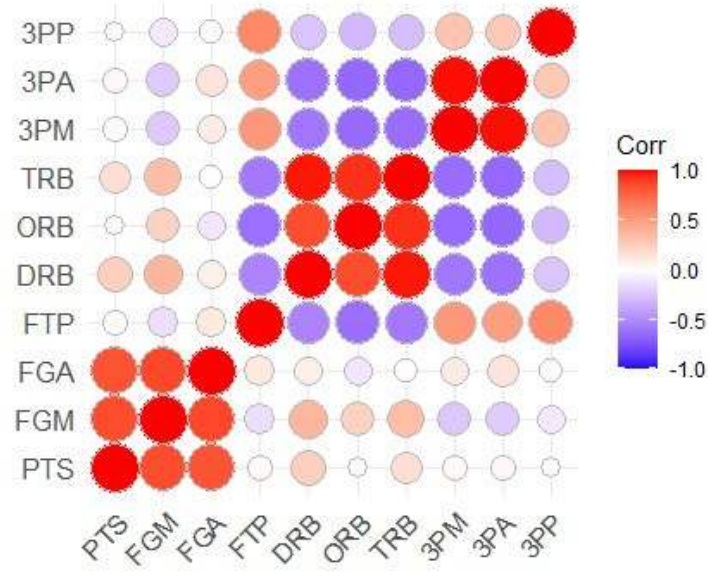
Fem la matriu de correlació com vam veure al seminari 7, arrodonint a 2 decimals, per exemple

```
> cormat <- round(cor(NBA_sel),2)

> ggcorrplot(cormat, lab=TRUE)
```

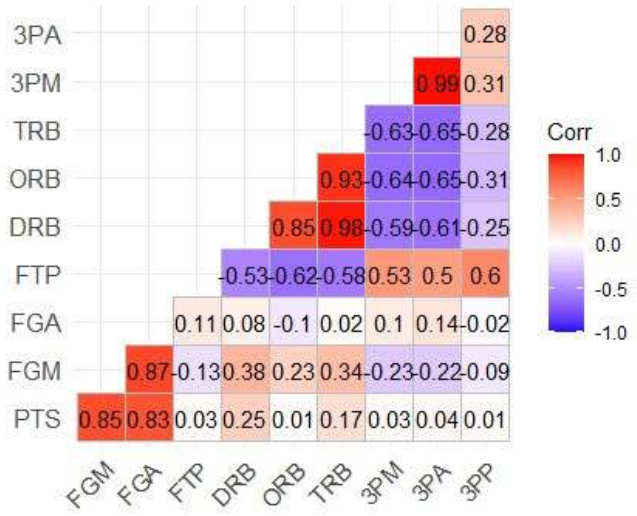


O amb cercles: >
```
ggcorrplot(cormat,method= 'circle')
```

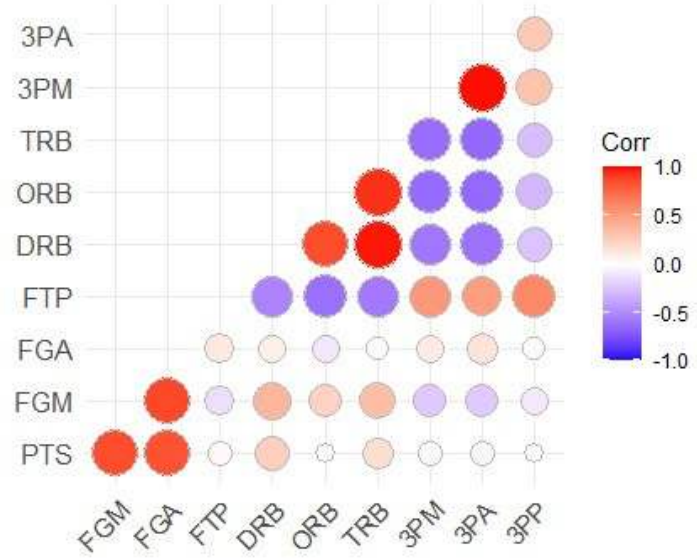


En qualsevol dels casos al ser una matriu simètrica, només mostrem, la part de sobre o sota la diagonal

```
> ggcorrplot(cormat, lab=TRUE, type = "lower")
```




```
> ggcorrplot(cormat, method='circle', type = "lower")
```



CONCLUSIONS POSIBLES (n'hi ha moltes)

- *Les variables: PTS (percentatge de punts de l'equip, punts per joc), FGM (cistelles de camp realitzats, percentatge de cistelles de camp) i FGA (tirs de camp intentats, número*

d'intencions de cistelles de camp) , tenen una correlació positiva entre ells és forta, sembla que fer més intents resulta en obtenir una puntuació més alta.

- *Sembla que els DRB (Rebots defensius), ORB (rebots ofensius), també van estar molt correlacionats (negativament) en la NBA 2008.*
- *Hi ha bastanta correlació també entre FTP i 3PM, 3PA i 3PP*
- *No hi ha cap correlació entre ORB i PTS. Fer més rebots ofensius no ajudava a guanyar. B) Clarament si volem reduir el nombre de variables, assegurant la seva independència, usarem un PCA com vam veure en Teoria 7 i seminari següent*

PART 3 (4 pt)

Dataset: 25_noms_padro_any_sexe_1996_2019.csv

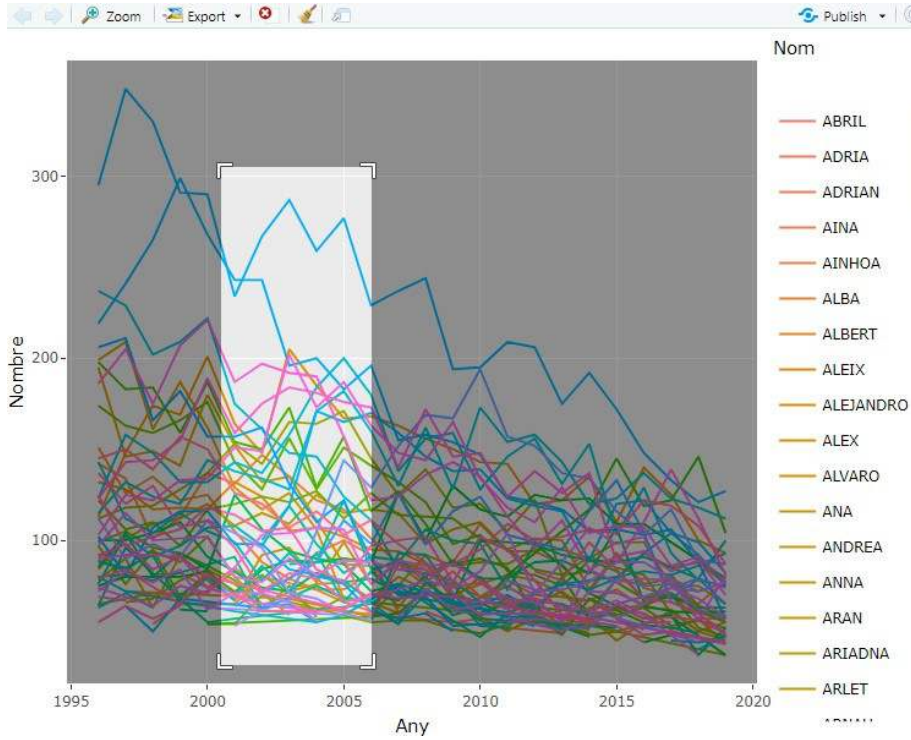
Agafarem el dataset de noms del padró de naixements de Barcelona. Utilitzeu les llibreries d'interactivitat o animació (plotly, gganimate, shiny, gifski, etc.) que creieu convenients i dibuixeu les gràfiques que us facin falta. RESPOSTA: *(afegiu en aquesta secció les llibreries que utilitzareu)*

```
> library(tidyverse)
> library(dplyr)
> library(plotly)
> getwd()
> setwd("C:/Users/enric/Documents/R")
> NadonsBCN <- read.csv('./25_noms_padro_any_sexe_1996_2019.csv')
```

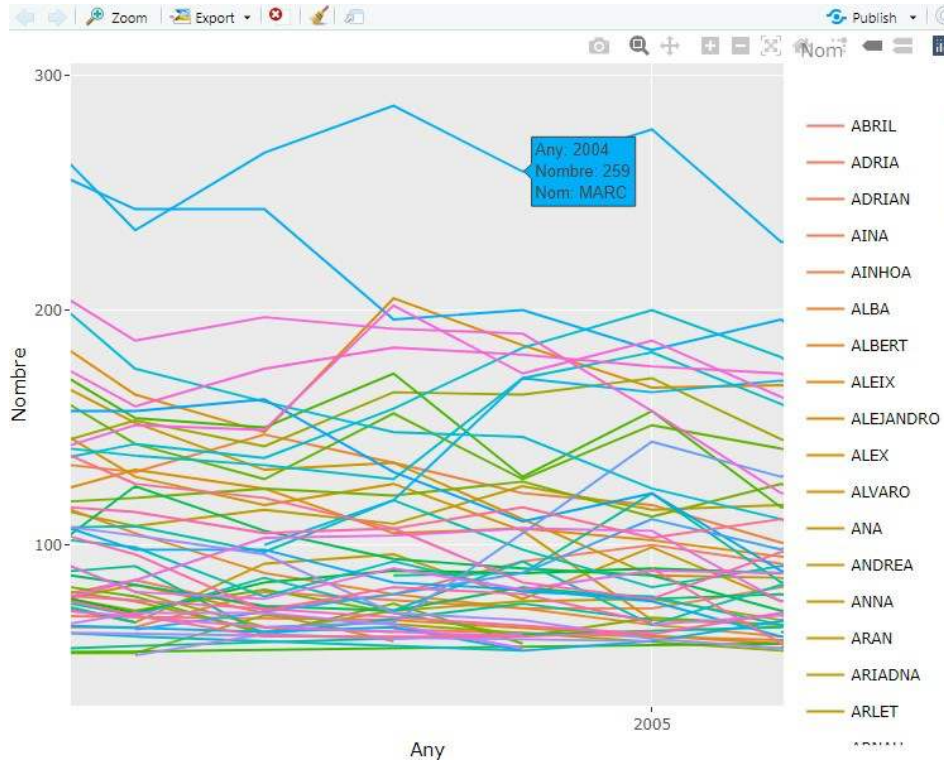
3.1 (1 pt) Mostra el codi i la gràfica de l'evolució temporal de tots els noms del data set (gràfica 1) i enumera els 10 noms femenins o masculins més posats l'any 2004. Digues quin són aquests noms i el nombre de nadons, per ordre decreixent i mostra el codi i la gràfica de l'evolució temporal d'aquests 10 noms (gràfica 2).

RESPOSTA: *En primer lloc, llistar tots els noms del dataset (gràfica 1):*

```
> plotNoms<- ggplot(NadonsBCN,aes(x=Any, y=Nombre,color=Nom))+geom_path()
> ggplotly(plotNoms)
```



CERCA INTERACTIVA: fer un zoom en la zona propera a l'any 2004.



CERCA AMB DATA MASSAGING (igual de bé):

Filtrar per l'any 2004, definir un ranking,per a obtenir els 10 primers i (opcionalment) ordenar-los en decreixent:

```
> NadonsBCN_2004O <- NadonsBCN %>% filter(Any==2004) %>% mutate(rank =
rank(-Nombre)) %>% group_by(Nom) %>% filter(rank <=10) %>% ungroup() %>%
arrange(-Nombre) > NadonsBCN_2004O
# A tibble: 10 x 6
   Ordre Nom   Sexe    Any Nombre  rank
   <int> <chr> <chr> <int>  <int> <dbl>
1      1 MARC  Home   2004    259   1
2      1 MARIA Dona   2004    200   2
3      2 PAULA Dona   2004    190   3
4      2 ALEX  Home   2004    185   4
5      3 LAIA  Dona   2004    184   5
6      3 POL   Home   2004    181   6
7      4 PAU   Home   2004    173   7
8      4 JULIA Dona   2004    171   8.5
9      5 LUCIA Dona   2004    171   8.5
10     6 CARLA Dona   2004    164  10
```

*Els noms són **MARC, MARIA, PAULA, ALEX, LAIA, POL, PAU, JULIA, LUCIA i CARLA.***

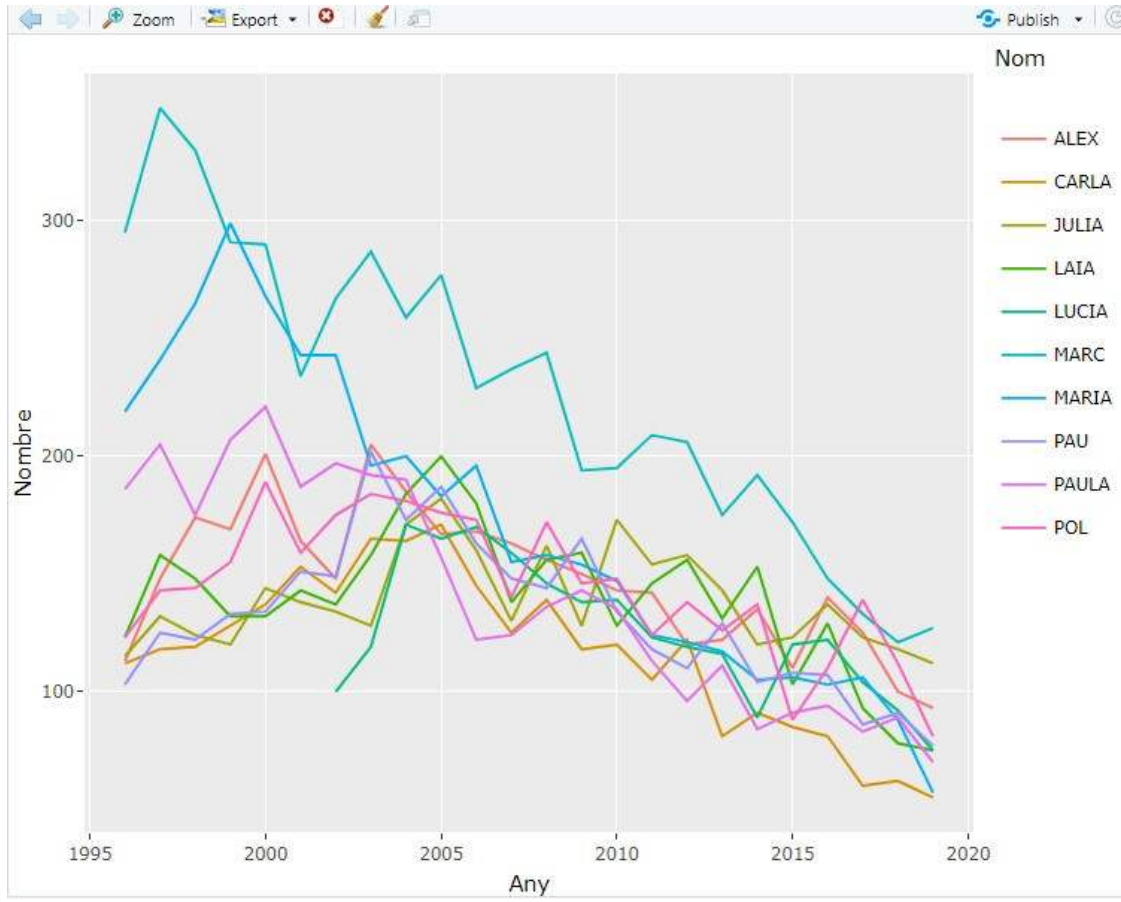
GRÀFICA 2:

Després llistem les 10 gràfiques, bé per ggplotly:

```
> NadonsBCN_Noms2004 <- NadonsBCN %>% filter(Nom=="MARC"|Nom=="MARIA"
|Nom=="PAULA"|Nom=="ALEX"|Nom=="LAIA"|Nom=="POL"|Nom=="PAU"|Nom=="JULIA"|
Nom=="LUCIA"|Nom=="CARLA")
> plotNoms2004 <- ggplot(NadonsBCN_Noms2004,aes(x=Any, y=Nombre,
color=Nom)) + geom_path() > ggplotly(plotNoms2004)
```

 # O bé per plot_ly:

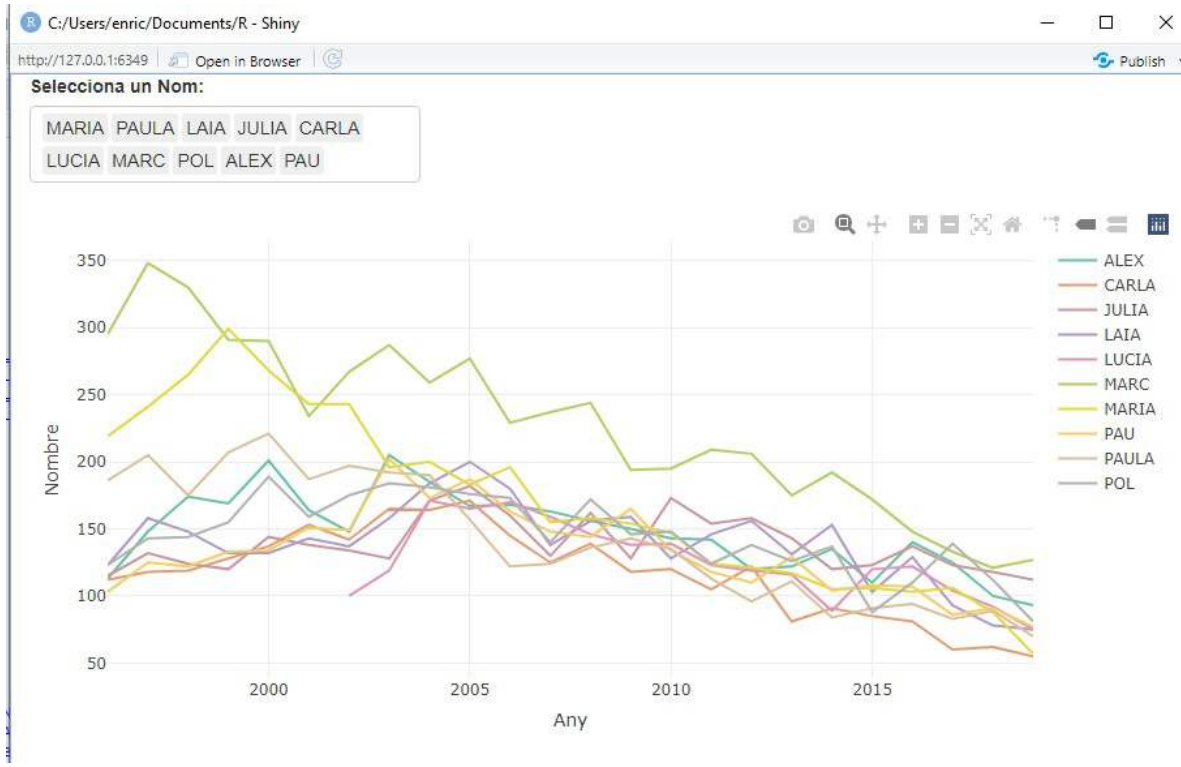
```
> plot_ly(NadonsBCN_Noms2004,x=~Any, y=~Nombre,color=~Nom) %>%
add_lines()
```



O bé per shiny:

```
> ui <- fluidPage( selectizeInput( inputId = "Noms",
label = "Selecciona un Nom:",                    choices =
unique(NadonsBCN$Nom),                 selected =
c("MARC","MARIA","PAULA","ALEX","LAIA",
                                      "POL","PAU","JULIA","LUCIA","CARLA"),
                  multiple = TRUE
                      ),
    plotlyOutput(outputId = "p")
      )

server <- function(input, output, ...)
{ output$p <- renderPlotly (
  { plot_ly(NadonsBCN, x = ~Any, y = ~Nombre, color=~Nom) %>%
                  filter(Nom %in% input$Noms) %>%
                group_by(Nom) %>%
     add_lines()     })
 }
shinyApp(ui, server)
```



3.2. (0,5 pt) Sobre gràfica 2 anterior, quin és l'ordre decreixent d'aquests noms l'any 2010?.

RESPOSTA:

CERCA INTERACTIVA: Mirar la gràfica 2 i cercar en ordre decreixent els noms l'any 2010.

```
   Ordre   Nom Sexe  Any Nombre
1      1  MARC Home 2010    195
2      2 JULIA Dona 2010    173
3      2   POL Home 2010    148
4      3 MARIA Dona 2010    147
5      3  ALEX Home 2010    143
6      4 LUCIA Dona 2010    139
7      5 PAULA Dona 2010    135
8      4   PAU Home 2010    134
9      6  LAIA Dona 2010    128
10     7 CARLA Dona 2010    120
```

CERCA AMB DATA MASSAGING (igual de bé):

Sobre el data set inicial (NadonsBCN) filtrar pels 10 noms i per l'any 2010, i ordenar en decreixent:

```
> NadonsBCN_2004_20100 <- NadonsBCN %>% filter(Nom=="MARC"|Nom=="MARIA"
  |Nom=="PAULA"|Nom=="ALEX"|Nom=="LAIA"|Nom=="POL"|Nom=="PAU"|
```

```
axis.text.x=element_blank(),
axis.text.y=element_blank(),
axis.ticks=element_blank(),
axis.title.x=element_blank(),
axis.title.y=element_blank(),         legend.position="none",
panel.background=element_blank(),
panel.border=element_blank(),
panel.grid.major=element_blank(),
panel.grid.minor=element_blank(),
        panel.grid.major.x = element_line( size=.1, color="grey" ),
panel.grid.minor.x = element_line( size=.1, color="grey" ),
plot.title=element_text(size=25,      hjust=0.5,      face="bold",
colour="grey", vjust=-1),
        plot.subtitle=element_text(size=18, hjust=0.5, face="italic",
color="grey"),
        plot.caption =element_text(size=8, hjust=0.5, face="italic",
color="grey"),
        plot.background=element_blank(),
plot.margin = margin(2,2, 2, 4, "cm")) +
  transition_states(Any, transition_length = 4, state_length = 1, wrap
= FALSE) +
  view_follow(fixed_x = TRUE) +
  labs(title = 'Noms de Nens a BCN Any: {closest_state}',
       subtitle  =  "Top 10 Noms del 2004",
       caption  = "Nombre de nadons | Data Source: Ajuntament de BCN")
> anim
>
```

PAS 3a: EXPORTAR FRAMES A FITXER GIF

```
> animate(anim, 200, fps = 20,  width = 1200, height = 1000,  renderer
        = gifski_renderer("nadonsBCN_2004.gif"), end_pause =
        15, start_pause =  15)
>
```

PAS 3b: EXPORTAR FRAMES A FITXER AVI

```
>  animate(anim,  200,  fps  =  20,    width  =  1200,  height  =  1000,
renderer   =   av_renderer("nadonsBCN_2004.avi"),   end_pause   =   15,
start_pause =  15)  >
```

4 (1 pt) Defineix 4 dels següents conceptes en animació, interactivitat, usabilitat i Experiència d'Usuari en Visualització de Dades:

- **SUS**
- **Control**
- ***CheckBox***

- **Participar / Col·laborar**
- ***Useful***
- **Qüestionari Attrakdiff**
- **UEQ**
- **Qüestionari SAM**