Visualització de Dades (Enginyeria de Dades – EE - UAB)
Examen Segon Parcial – 16 Juny 2025
MODEL B

Nom i Cognom: David Morillo Massagué

NIU: 1666540

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

PART 1 (6 pts.)

Dataset: 25_noms_padro_any_sexe_1996_2019.csv

Agafarem aquest dataset que conté els noms de nens i nenes més freqüents dels nadons de Barcelona entre els anys 1996 i 2019. Podeu fer servir les llibreries R (plotly, gganimate, shiny, etc.) que creieu convenients i dibuixeu les gràfiques que us facin falta. Cal incloure les comandes R i una captura de pantalla de la gràfica que es demana.

Cada registre d'aquest data set conté informació d'un nom i any registrat. Conté les variables:

- **Ordre** → Número de ranking de nens o nenes d'un any, segons el nom
- **Nom** → Nom del nadó
- **Sexe** → Gènere. Té dos valors: "Dona", "Home"
- **Any** → Any de la dada
- **Nombre** → Nombre de nadons amb aquest nom i any.

1.1 (1 pt.) Feu una gràfica interactiva de línies sobre l'evolució del nom de CARLA al llarg dels anys i contesta les preguntes sobre la gràfica.

RESPOSTA:

Codi:

```
df <- read.csv("25_noms_padro_any_sexe_1996_2019.csv")

df_carla <- df %>% filter(Nom == "CARLA")

# Crear el gráfico
ggplot2_carla <- ggplot(df_carla, aes(x = Any, y = Nombre, color = Nom)) +
  geom_line() +
  labs(title = "Evolució temporal del nom CARLA a traves del anys",
       x = "Any", y = "Nombre") +
  theme_minimal()

# Interactivo con plotly
ggplotly(ggplot2_carla)
```

Sobre aquesta gràfica contesta les preguntes:

a) En quins anys hi ha el màxim nombre de nadons amb aquest nom i en quin any hi ha el mínim?. Dona l'any i el nombre.

   RESPOSTA:

   - Maxim: 2005 (171)

   - Mínim: 2019 (55)

b) En quin any es puja per primer cop el nombre de nadons amb aquest nom a més de 150 i en quin any es baixa per primer cop el nombre de nadons amb aquest nom a menys de 100 nadons?. Dona l'any i el nombre

   RESPOSTA:

   - Puja de 150 a l'any 2001 per primer cop amb 153 nadons

   - Baixa de 100 a l'any 2013 amb 81 nadons

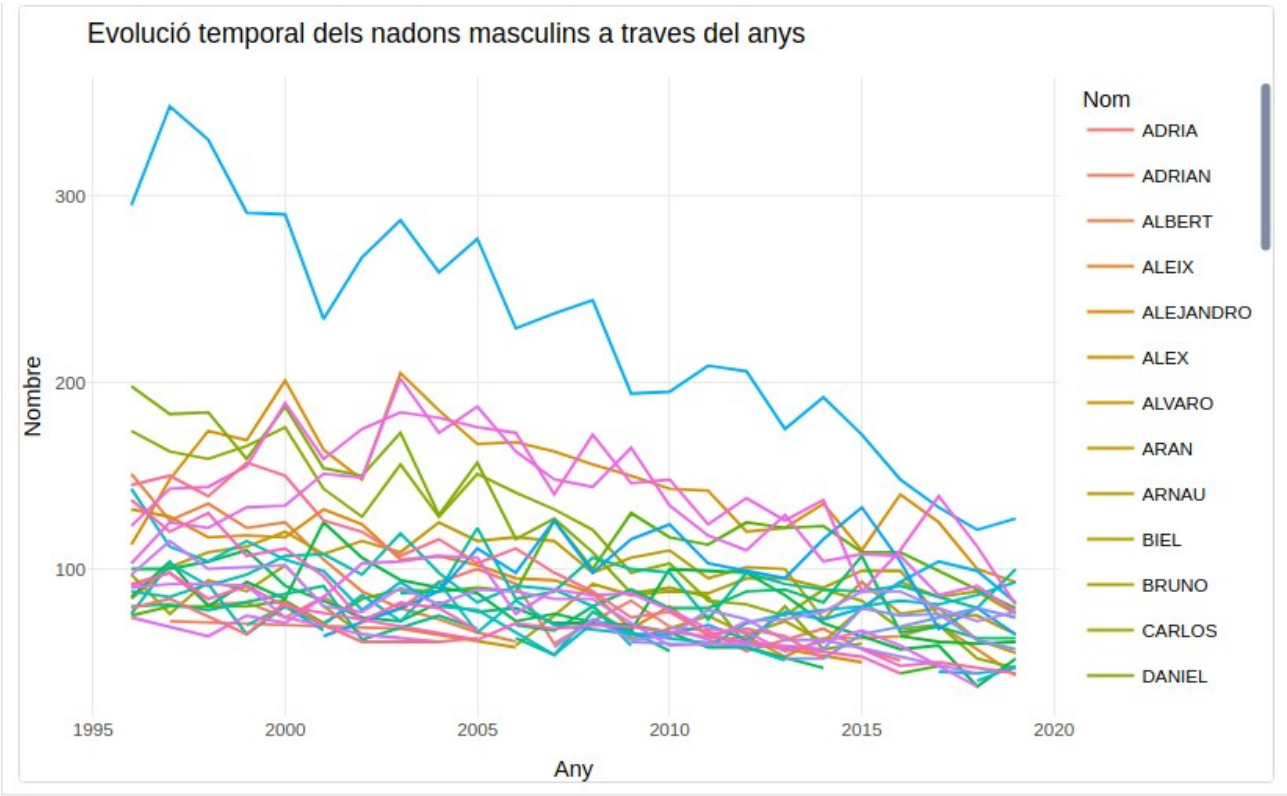
c) En quin parell d'anys s'assoleix el màxim descens de nadons amb aquest nom?.

   RESPOSTA:

   Del 2005 (171) al 2007 (125)

1.2. (1 pt.) Visualitza en gràfica de línies l'evolució temporal dels noms masculins (plotNomsMasculins).

RESPOSTA:




```
df_masc = df %>% filter(Sexe == "Home")

# Crear el gráfico
ggplot2_masc <- ggplot(df_masc, aes(x = Any, y = Nombre, color = Nom)) +
  geom_line() +
  labs(title = "Evolució temporal dels nadons masculins a traves del anys",
       x = "Any", y = "Nombre") +
  theme_minimal()

# Interactivo con plotly
ggplotly(ggplot2_masc)
```

Sobre aquesta gràfica interactiva, respon a les següents preguntes

a) Quins són els 3 noms masculins menys i més utilitzats els anys 1996 i 2016, especificant el nombre de nadons per a cada nom?.

RESPOSTA:

1996:

Menys utilitzats:

- PABLO (74)
- ERIC (75)
- JOAN (76)

Més utilitzats:

- MARC (295)
- DAVID (198)
- DANIEL (174)

2016:

Menys utilitzats:

- GABRIEL (44)
- ROGER (44)
- ROC (48)

Més utilitzats

- MARC (148)
- ALEX (140)
- POL (110)

Codi:

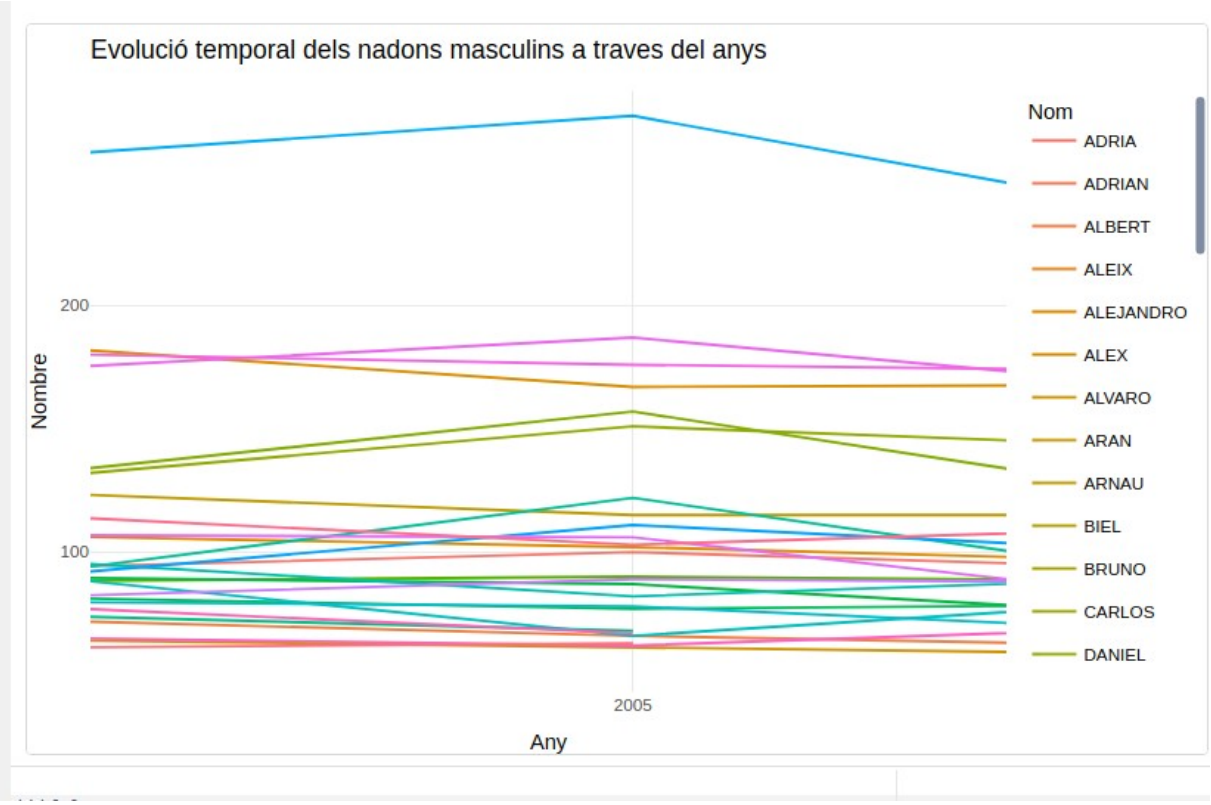
```
menys_populars_1996 <- df_masc %>%
  filter(Any == 1996) %>%
  arrange(Nombre) %>%
  slice_head(n = 3)

menys_populars_1996

mes_populars_1996 <- df_masc %>%
  filter(Any == 1996) %>%
  arrange(desc(Nombre)) %>%
  slice_head(n = 3)

mes_populars_1996

menys_populars_2016 <- df_masc %>%
```

```
  filter(Any == 2016) %>%
  arrange(Nombre) %>%
  slice_head(n = 3)

menys_populars_2016

mes_populars_2016 <- df_masc %>%
  filter(Any == 2016) %>%
  arrange(desc(Nombre)) %>%
  slice_head(n = 3)

mes_populars_2016
```

b) Quin és el nom o noms masculins del que es té la darrera referència l'any 2005?. Digues nom i nombre de nadons aquell any.

RESPOSTA:

- IVAN (68)
- SERGI (67)
- XAVIER (63)



1.3. (1 pt.) Reproduïu l'aplicació shiny amb entrada per desplegable que dibuixi la gràfica de línies amb tots els noms femenins de nadons del dataset, de forma que es puguin seleccionar els noms en el desplegable. Posa com a nom per defecte el nom 'SARA'. Utilitza la funció `plot_ly()` o `ggplot()`, la que vulguis.

RESPOSTA:

Selecciona els noms de CLAUDIA, CARLA, MARIA, PAULA i SARA. Mostra la gràfica.

RESPOSTA:

Sobre aquesta gràfica contesta a les preguntes:

a) Troba tres característiques de la gràfica.

   RESPOSTA:

b) Quins estan els tres primers l'any 2004?. Dona noms i nombre de nadons

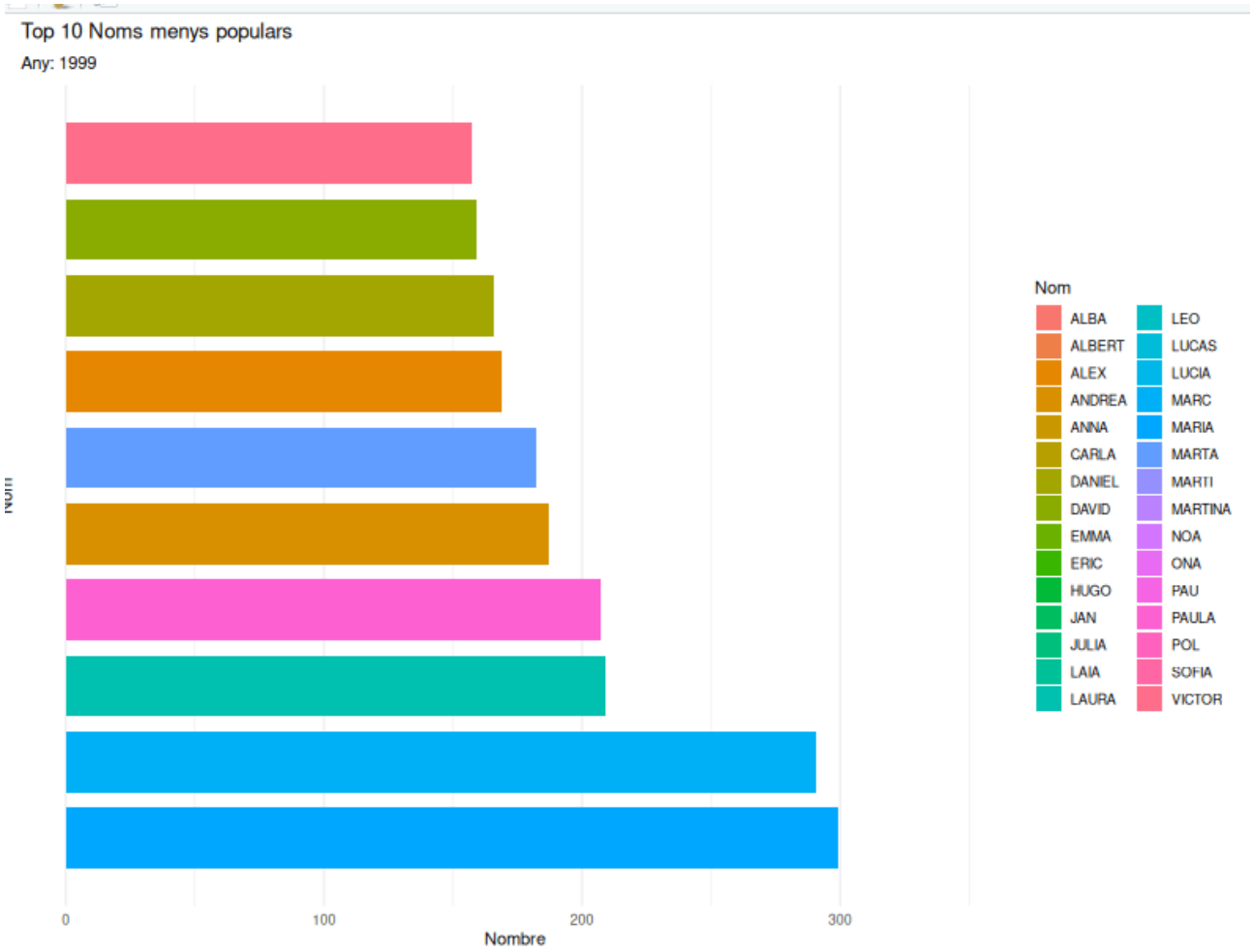
   RESPOSTA:

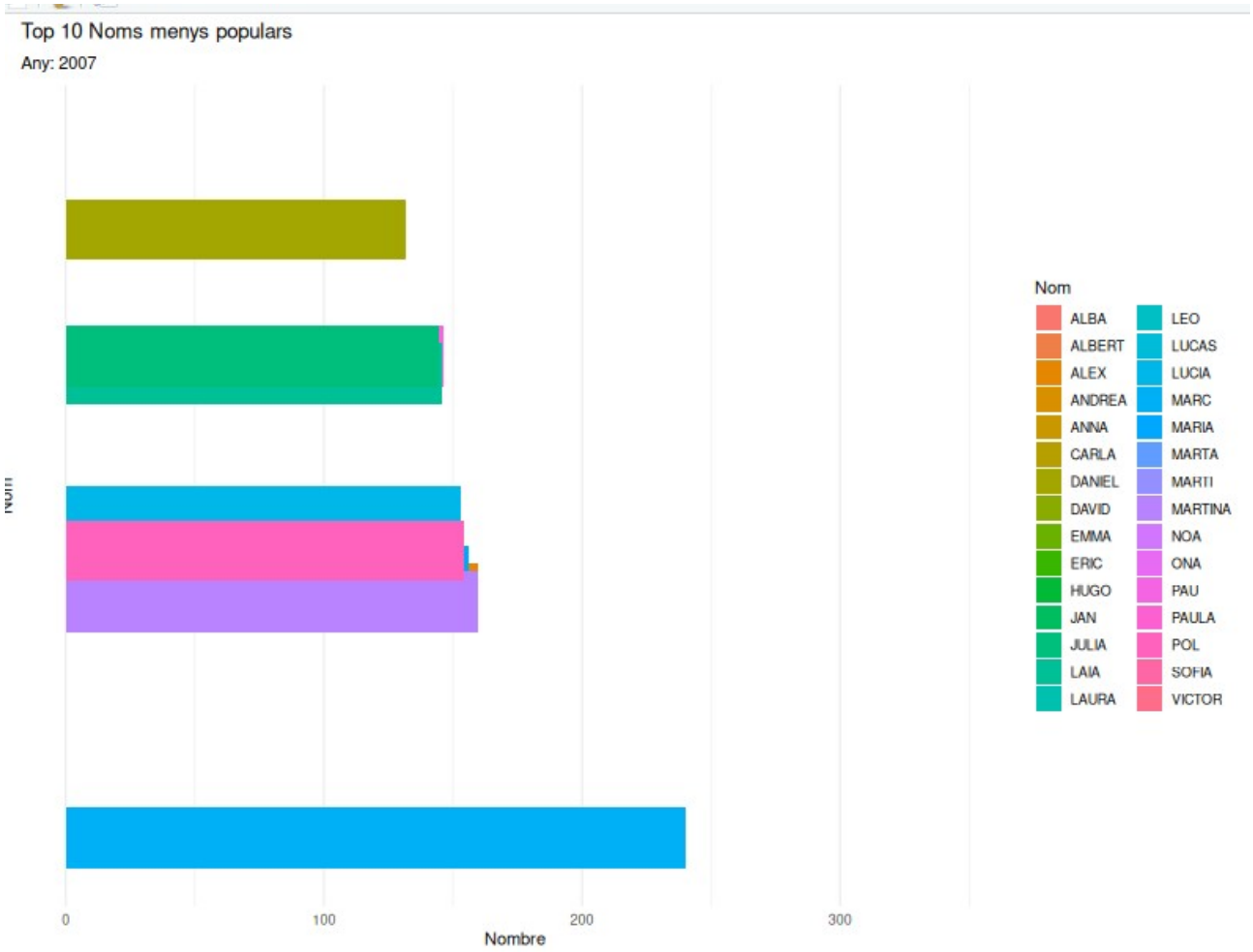
c) Quins estan els tres últims l'any 2017?.

   RESPOSTA:

1.4. (2 pts.) Mostra el codi per a generar el Ranking de Barres Animades (*Animated Bar Race Ranking*) sobre els 10 noms masculins o femenins menys utilitzats cada any. Fes un parell de captures de pantalla de l'animació.

RESPOSTA:


Top 10 Noms menys populars
Any: 1999
0
100
200
300
Nombre
Nom
ALBA
ALBERT
ALEX
ANDREA
ANNA
CARLA
DANIEL
DAVID
EMMA
ERIC
HUGO
JAN
JULIA
LAIA
LAURA
LEO
LUCAS
LUCIA
MARC
MARIA
MARTA
MARTI
MARTINA
NOA
ONA
PAU
PAULA
POL
SOFIA
VICTOR



Codi:

```
df_anim <- df %>%
  arrange(Any, desc(Nombre)) %>%
  group_by(Any) %>%
  mutate(Rank = row_number()) %>%
  filter(Rank <= 10)

# Creació de la gràfica animada (amb canvi dinàmic de posició)
anim_plot <- ggplot(df_anim, aes(x = Rank, y = Nombre, fill = Nom)) +
  geom_bar(stat = "identity", width = 0.8) +
  coord_flip() +
  scale_x_continuous(breaks = -1:-10, labels = df_anim$Nom[df_anim$Any == 1996]) +
  labs(title = 'Top 10 Noms menys populars',
       subtitle = 'Any: {frame_time}',
       x = 'Nom', y = 'Nombre') +
```

```
  theme_minimal() +
  theme(axis.text.y = element_text(hjust = 1)) +
  transition_time(Any) +
  ease_aes('linear') +
  labs(fill = "Nom")

# Generació i guardat del GIF
animated_gif <- animate(anim_plot, fps = 10, width = 800, height = 600, duration = 30, renderer = gifski_renderer())

animated_gif

anim_save("noms.gif", animation = animated_gif)
```

*No he aconseguit que sortissin els noms al eix y, em donava error i per això s'ha optat per una llegenda estàtica

1.5. (1 pt.) Defineix 4 dels següents conceptes en animació, interactivitat, usabilitat i Experiència d'Usuari en Visualització de Dades:

- *SUS:* Qüestionari curt de 10 ítems per mesurar la usabilitat percebuda d'un sistema de forma ràpida i general. Avalua facilitat d'ús, complexitat i confiança.
- *Utility*
- *UEQ*
- *Data linking:* Quan se seleccionen dades en un gràfic, la mateixa posició/element es resalta en altres gràfics.
- Event: Acció realitzada per l'usuari (clic amb el ratolí, passar el cursor per sobre d'un punt, etc.). Que en termes d'interactivitat és l'acció que permet canviar gràfiques, o donar informació a l'usuari de la informació concreta que vol consulta
- *Checkbox:* Objecte que permet la selecció d'un o varies categories de dades alhora, utilitzats en dades categòriques. Per exemple, per seleccionar la generació X, Y o Z en un gràfic.

PART 2 (4 pts.)

Dataset: filmdeathcounts_with_main_genre.csv

- Si utilitzeu Tableau, cal incloure una captura de pantalla de l'aplicació Tableu on es vegin les configuracions actives i la gràfica.
- Si utilitzeu R, cal incloure les comandes R i una captura de pantalla de la **gràfica.**

2.1 (Total: 2,5 pts.) Fes un *treemap* que mostri quins directors són els més "letalment productius" dins de cada gènere.

2.1.1 (0,75 pts.) Data Massage necessari per la visualització

a. Llegeix les dades.
b. Agrupa les dades pel director (Director) i el gènere principal (Main_Genre).
c. Calcula el total de morts (Body_Count) per a cada combinació de Director i Main_Genre.
d. Filtra directors amb almenys 300 morts acumulades.
e. Desa el resultat en un dataframe anomenat dataT.

```
df <- read.csv("filmdeathcounts_with_main_genre.csv")

dataT <- df %>%
  drop_na(Director) %>%
  drop_na(Main_Genre) %>%
  drop_na(Body_Count) %>%
  group_by(Director, Main_Genre) %>%
  reframe(MPAA_Rating, Count = sum(Body_Count, na.rm = TRUE), .groups = "drop") %>
    %
  mutate(mes300 = Count < 300)
```

2.1.2.a (0,75 pts.) Mostra un mapa d'arbre (*treemap*) que et permeti saber quins directors (*Director*) fan pel·lícules amb més de 300 morts totals per cada gènere principal (Main_Genre).

```
data2 <- df %>%
  drop_na(Director) %>%
  drop_na(Main_Genre) %>%
  drop_na(Body_Count) %>%
  group_by(Director, Main_Genre) %>%
  reframe(MPAA_Rating, Count = sum(Body_Count, na.rm = TRUE), .groups = "drop")
%>%
  mutate(mes300 = Count < 300)

ggplot(data2, aes(area=Count, fill=mes300 ,subgroup=Main_Genre))+
  geom_treemap()+
  geom_treemap_subgroup_border()+
  geom_treemap_subgroup_text(color='white')+
  geom_treemap_text(aes(label=Director),  reflow=TRUE)
```

2.1.2.b (0,75 pts.) Argumenta com és un *treemap* en general i detalla els passos que has de fer per construir la visualització d'aquest exercici (és a dir quina variable utilitzes per l'àrea de les graelles, variables d'agrupació, etc. i per què?).

Un treemap és un dibuix rectangular dividit en caselles, i cada casella representa una sola observació. És una bona manera de mostrar dades jeràrquiques mitjançant rectangles imbricats. I l'àrea relativa de cada casella expressava una variable contínua. Es defineix una variable que determina la mida de cada rectangle (variable continua Count), en el nostre cas uns subgrups (genere), i també una variable pel color dels rectangles (en la majoria de treemaps) que en el nostre cas es un booleà que representa si té >300 body counts en total en totes les pelicules del director i genere.

2.1.2.c (0,25 pts.) Quin director és més "letalment productiu" quan el gènere principal és aventura? I quan el gènere principal és crim?

Avengura: Wolfgang Petersen

Crim: King Hu

NOTA: En R, l'opció `reflow = TRUE` dins de `geom_treemap_text()` controla com es disposa el text dins de cada casella del *treemap*. D'aquesta manera el text s'adapta dinàmicament a l'espai disponible. És més llegible. (Podeu usar-lo, opcionalment en aquest exercici, per veure més informació).

RESPOSTES:

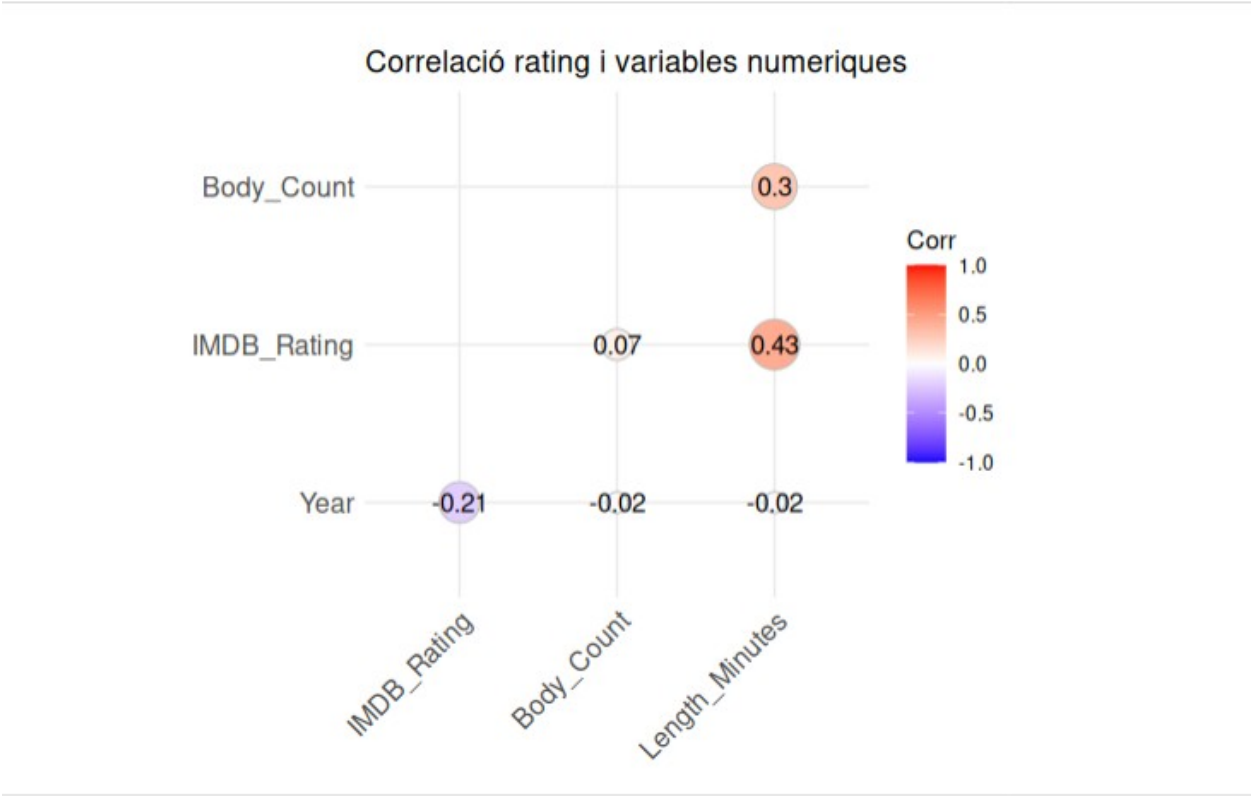
RESPOSTA 2.1.1:

RESPOSTA 2.1.2.a. i b. :

RESPOSTA 2.1.2.c:

2.2. (Total: 1,5 pts.)

a) Ajudat d'un gràfic per mostrar si la puntuació mitjana de la pel·lícula IMDb (IMDB_Rating) està correlacionada amb alguna de les variables numèriques d'aquest dataframe. (0,75 pt)



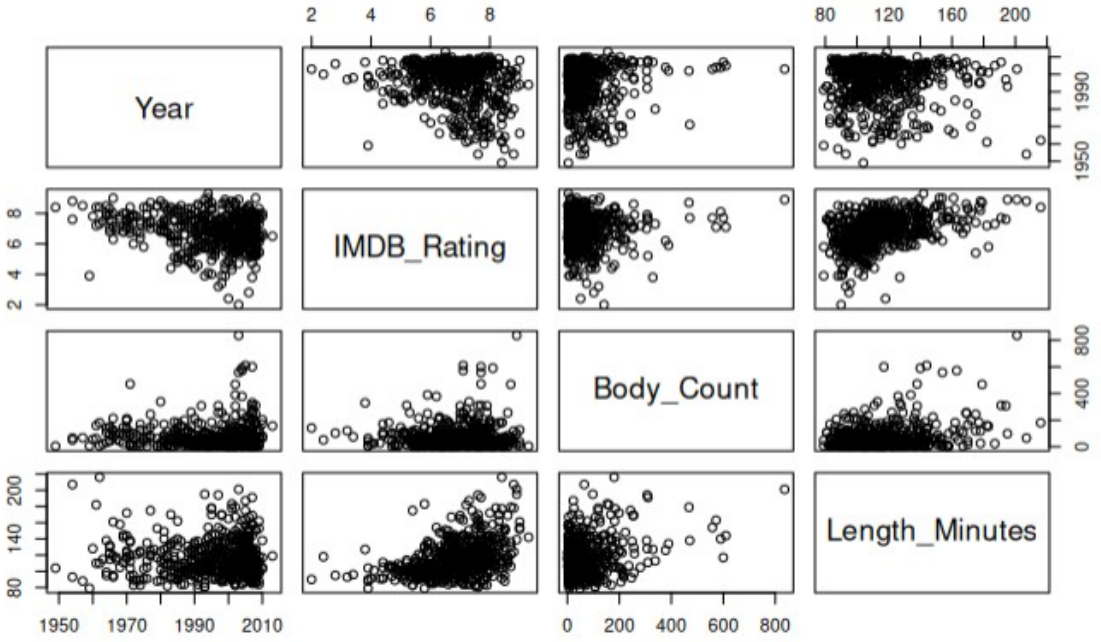

```
cols <- c("Year", "IMDB_Rating", "Body_Count", "Length_Minutes")

subset_data <- df[cols]


cor_matrix <- cor(subset_data, use = "complete.obs")


ggcorrplot(cor_matrix, method = "circle", type = "lower",
        title = "Correlació rating i variables numeriques",
        lab = TRUE, tl.col = "black", tl.srt = 45)
```

b) Mostra el gràfic, tot intentant que no hi hagi informació repetida, i argumenta l'elecció del gràfic. En pots extreure alguna conclusió? (0,5 pts.)

A la fila del mig veiem com hi ha una correlacio positiva de 0.43 entre la longitud de la pelicula i el rating IMDB

c) Descriu com faries un gràfic per veure com varia la valoració d'IMDB en funció d'aquestes variables numèriques, separant la informació per la classificació per edats establerta per la Motion Picture Association of America (MPAA_Rating). Nota: En aquest apartat, no es demana fer el gràfic, solament dir quin gràfic podria mostrar-ho i com. (0,25 pts.)

Fent servir un grafic de «pairs» com el següent:



Amb la variable MPAA_Rating es podria veure com els diferents valors categorics afecten al rating i de quina forma (linear, exponencial etc). No hi ha hagut temps d'implementar

RESPOSTES:

RESPOSTA 2.2.a):

RESPOSTA 2.2.b):

RESPOSTA 2.2.c):