Visualització de dades (Enginyeria de Dades – EE - UAB)
Examen Primer Parcial – 04 Abril 2025
MODEL A

Nom i Cognom: David Morillo Massagué

NIU: 1666540 Grup de Matrícula:

PARTE 1 (5 pt)

IMPORTANT: INSTALA LA LIBRERÍA: `install.packages("colorspace")`, cárgala `library(colorspace)` y prueba: `hcl_palettes(plot = TRUE)`

Dataset*: **2022_poblacio_gini_barris_renda.csv***

*Nota: Las variables 'Total * ' son porcentaje normalizado (0,1).*

Hay nombres de variables que no tienen el símbolo '_' *(por ejemplo:* `Poblacio total`*). En estos casos poner comillas al principio y final del nombre de variables (* 'Poblacio total'*).*

1.0. *Carga las librerías necesarias y el dataframe.*

library(colorspace)

library(ggplot2)

library(dplyr)

library(tidyverse)

df1 <- read.csv("/home/dakur/Desktop/examen/2022_poblacio_gini_barris_renda.csv")

1.1. (0.4 pt) Abre el fichero. Según los tipos de datos vistos en clase, escribe qué tipo de atributo son: Codi_Districte, Nom_Districte, Total Dona, Total Espanya, Index_Gini

RESPOSTA:

Codi_Districte: Categoric

Nom_Districte: Categoric

Total Dona: Quantitatiu

Total Espanya: Quantitatiu

Index_Gini: Quantitatiu

1.2. (1.6 pt)

A. Haz una gráfica que permita comparar el **Index_Gini** de los diferentes **distritos (Nom_Districte)**. Sube la gráfica y el código.

B. Haz otra igual para **Promedi_Renda_Bruta**. Sube la gráfica y el código.

Para ambas: Presta atención a la métrica del cálculo que usas para agregar por distrito. Para ello, añade la opción:

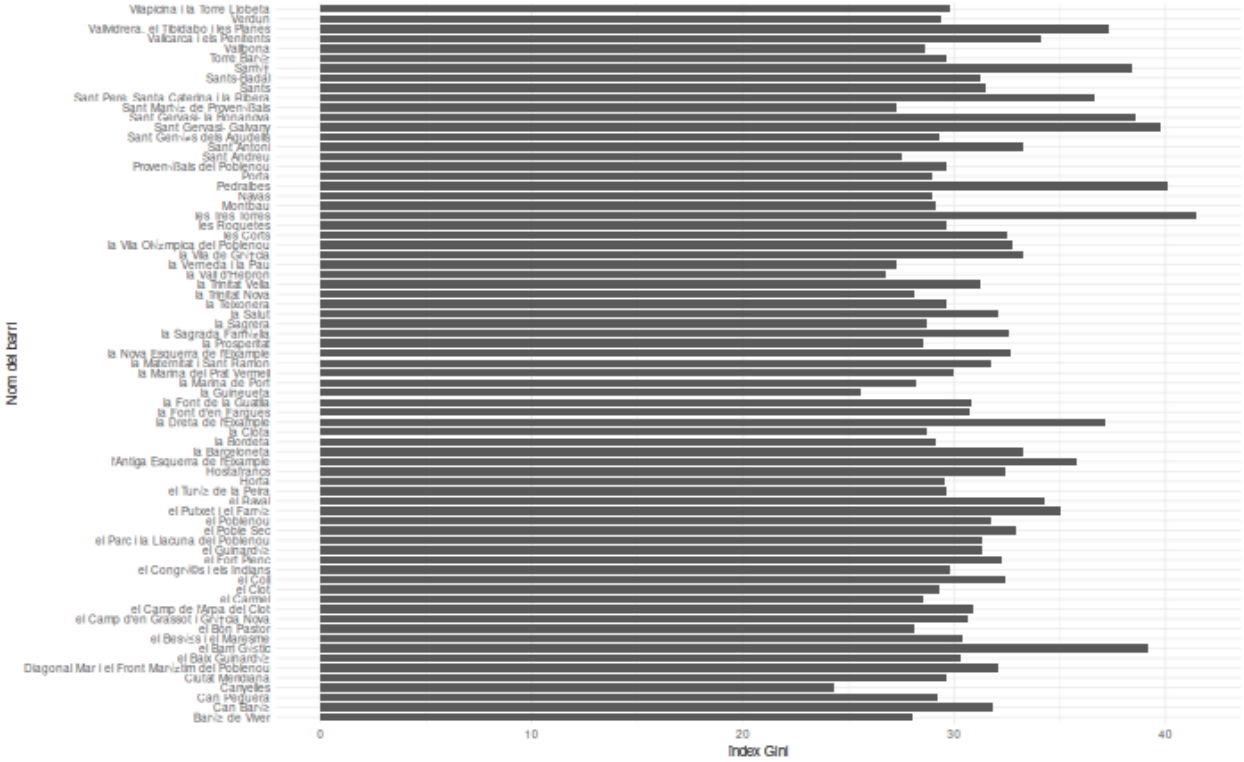
stat='summary': Realiza un cálculo (como suma, promedio, etc.) antes de crear las barras. En este caso**, ignora el warning** 'No summary function supplied,…'

Sube las gráficas y el código.

C. Argumenta muy brevemente: la gráfica que has usado y porqué es la más adecuada en este caso.

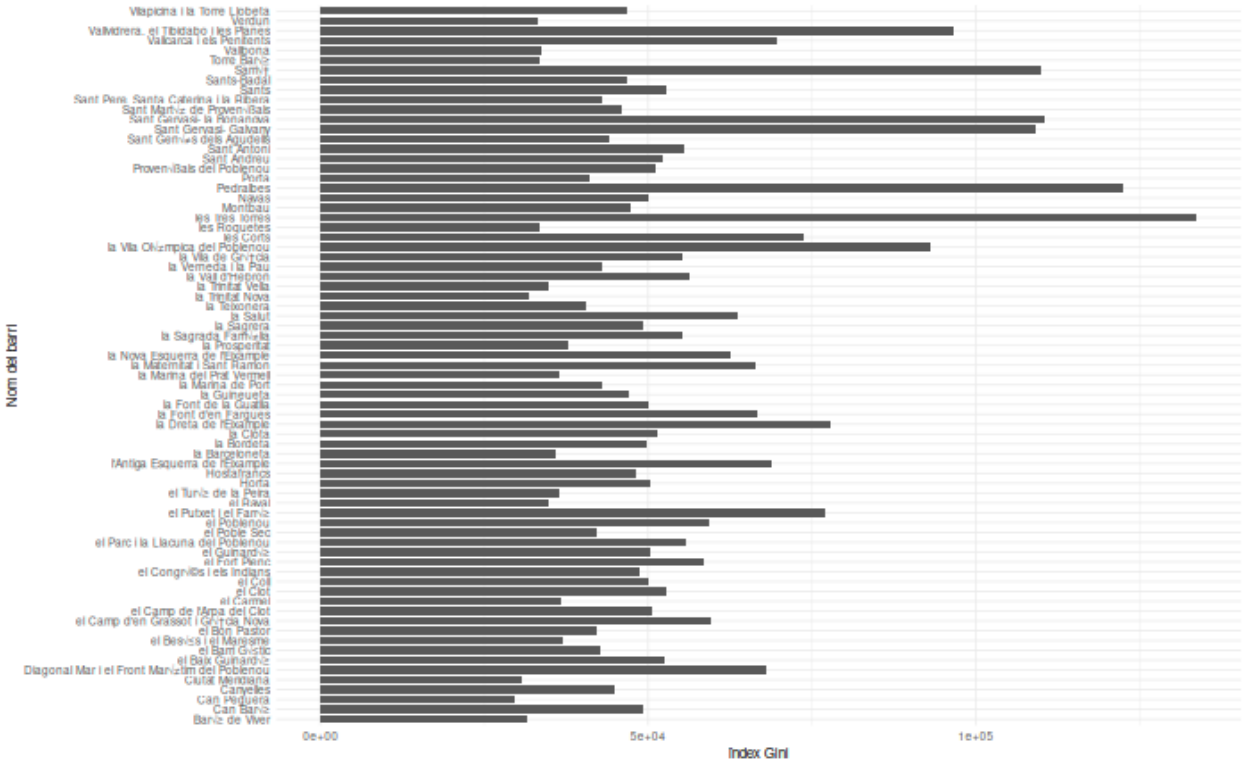
RESPOSTA:

A.




```
ggplot(df1, aes(x=Nom_Barri, y=Index_Gini)) +
  geom_col(width = 0.8, stat='summary') + coord_flip() +
  xlab("Nom del barri") + ylab("ïndex Gini") +
  theme_minimal(base_size=6)
```

B.




```
ggplot(df1, aes(x=Nom_Barri, y=Promedi_Renda_Bruta)) +
  geom_col(width = 0.8, stat='summary') + coord_flip() +
  xlab("Nom del barri") + ylab("ïndex Gini") +
  theme_minimal(base_size=6)
```

C. He fet servir un gràfic de tipus columna (horitzontal) ja que és indicada per a representar valors continus per a cada columna. En aquest cas, tenim els barris com a columna, i la longitud de la barra representa l'índex gini. Altres tipus de gràfica no s'adequerien tan bé a taules que representin un valor continu i un de categòric per eix.

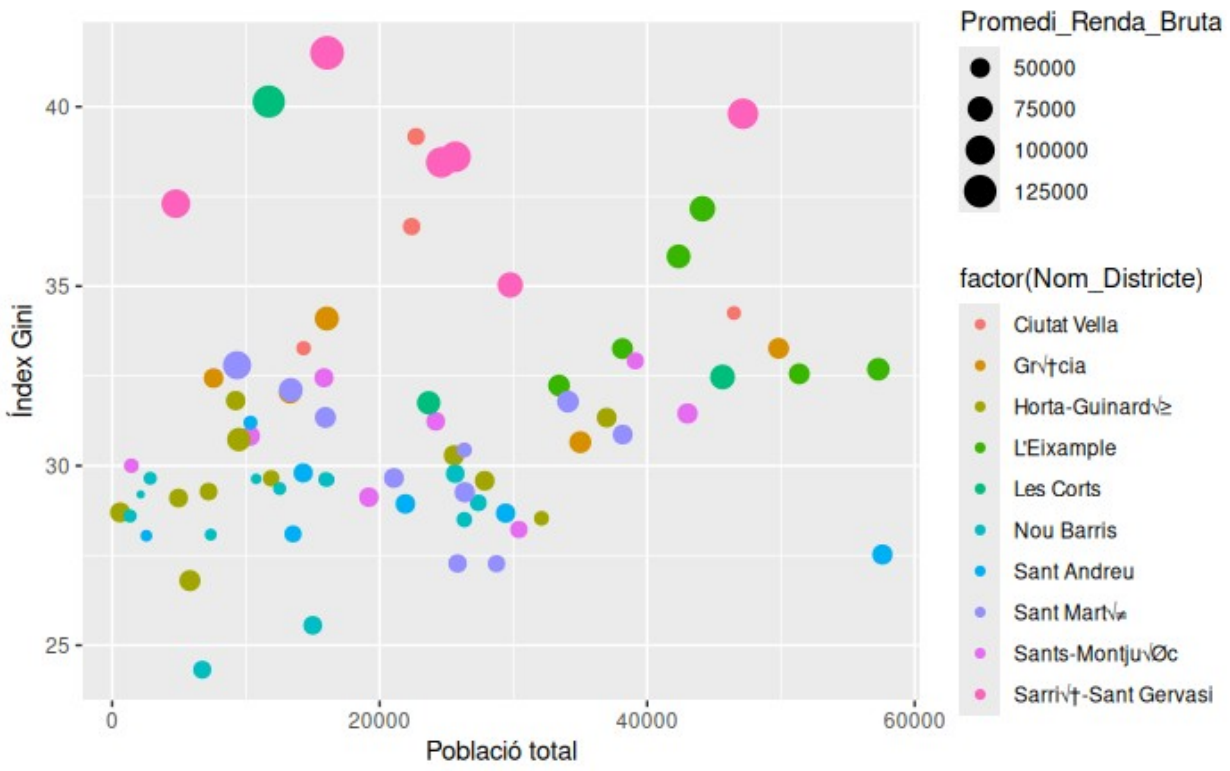
1.3 (1.4 pt)

A. Haz una gráfica que permita identificar outliers y explorar la correlación entre **Index_Gini**, **Poblacio total**, **Promedi_Renda_Bruta** y **Districte**. Elige la gráfica más adecuada según la tarea a realizar, y según el número y tipo de variables. Sube las gráficas y el código.

B. Argumenta muy brevemente: la gráfica que has usado y porqué es la más adecuada en este caso.

RESPOSTA:

A.



ggplot(df1, aes(Poblacio.total, Index_Gini, color=factor(Nom_Districte), size = Promedi_Renda_Bruta)) +

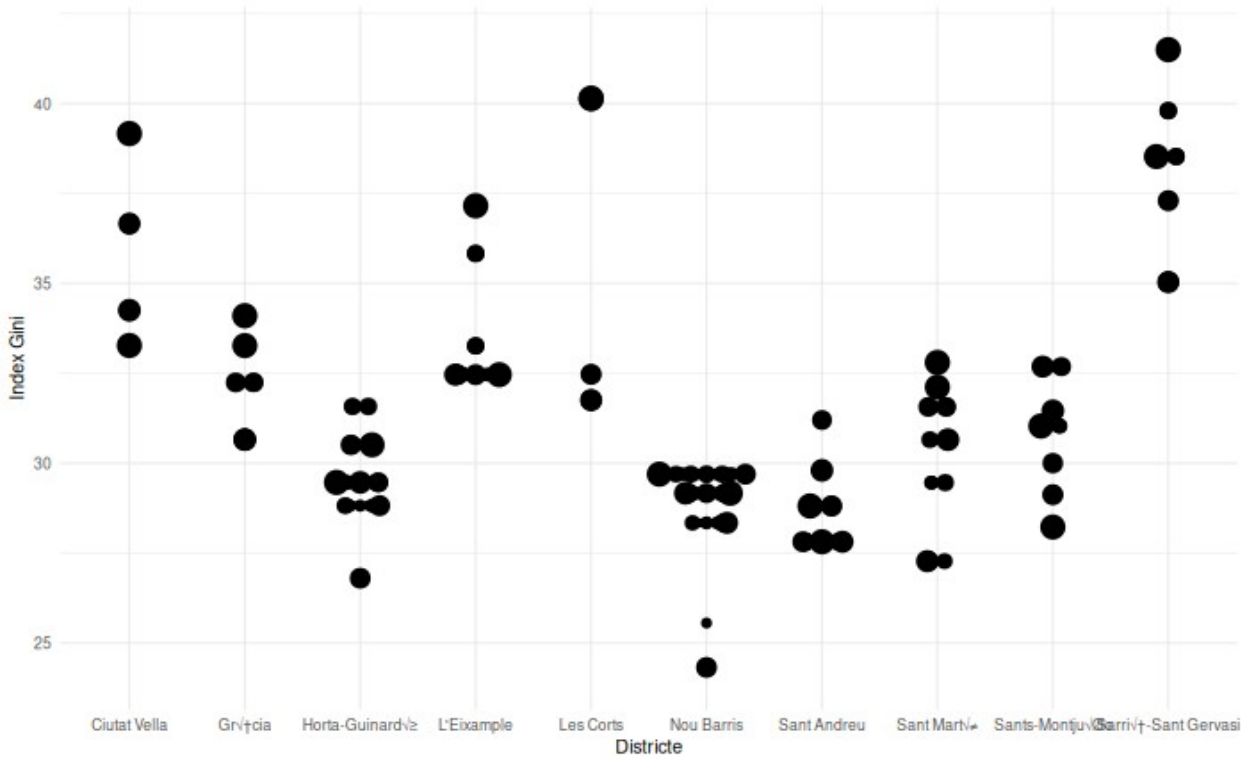
geom_point() + xlab("Població total") + ylab("Índex Gini")

B. He fet servir un gràfic de punts, on els eixos representen dos valors continus (població i índex gini), i podem fer servir dos més atributs com el color, per els barris (categòric), i la mida dels punts pel valor de la renda bruta en cada cas. Al ser menys de 12 colors, podem identificar fàcilment els barris del gràfic. Les variables continues estan correctament representades, amb l'ajuda dels eixos i llegendes.

1.4 (1 pt) Haz un dot plot que muestre la distribución de todo el dataset de **Index_Gini** por **Distrito**. Codifica **Poblacio total** en el tamaño de los círculos. Calcula la diferencia entre proporción de hombres y mujeres, y usa esa nueva métrica para colorear los puntos escogiendo el tipo de escala de color más adecuada entre *continuous/discrete _diverging, _qualitative, o _sequential*.

Sube las gráficas y el código de cada una.

RESPOSTA:



df1 <- df1%>%mutate(dif = Total.Dona - Total.Home) %>% na.omit()

ggplot(df1, aes(x = factor(Nom_Districte), y = Index_Gini)) +

geom_dotplot(binaxis = "y", stackdir = "center", dotsize = df1$Poblacio.total/70000, binwidth = 1) +

theme_minimal(base_size = 8) +

xlab("Districte") +

ylab("Index Gini")+

scale_color_gradient(low = "lightblue", high = "darkblue")

1.5 (0.6 pt) Según las gráficas que has hecho, responde:

¿Cuáles son los dos distritos con mayor desigualdad?

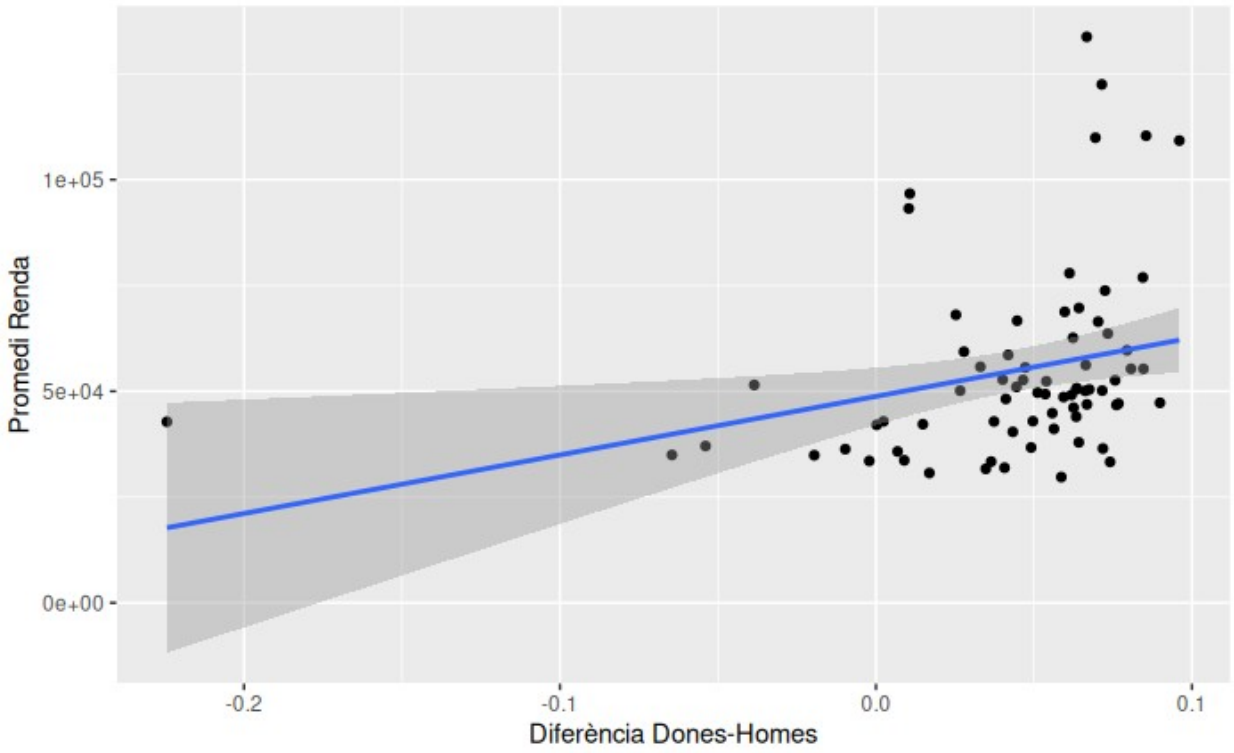
¿Hay correlación entre renta y desigualdad? ¿Positiva o negativa?

¿Viven más hombres que mujeres en los barrios más desiguales?

RESPOSTA:

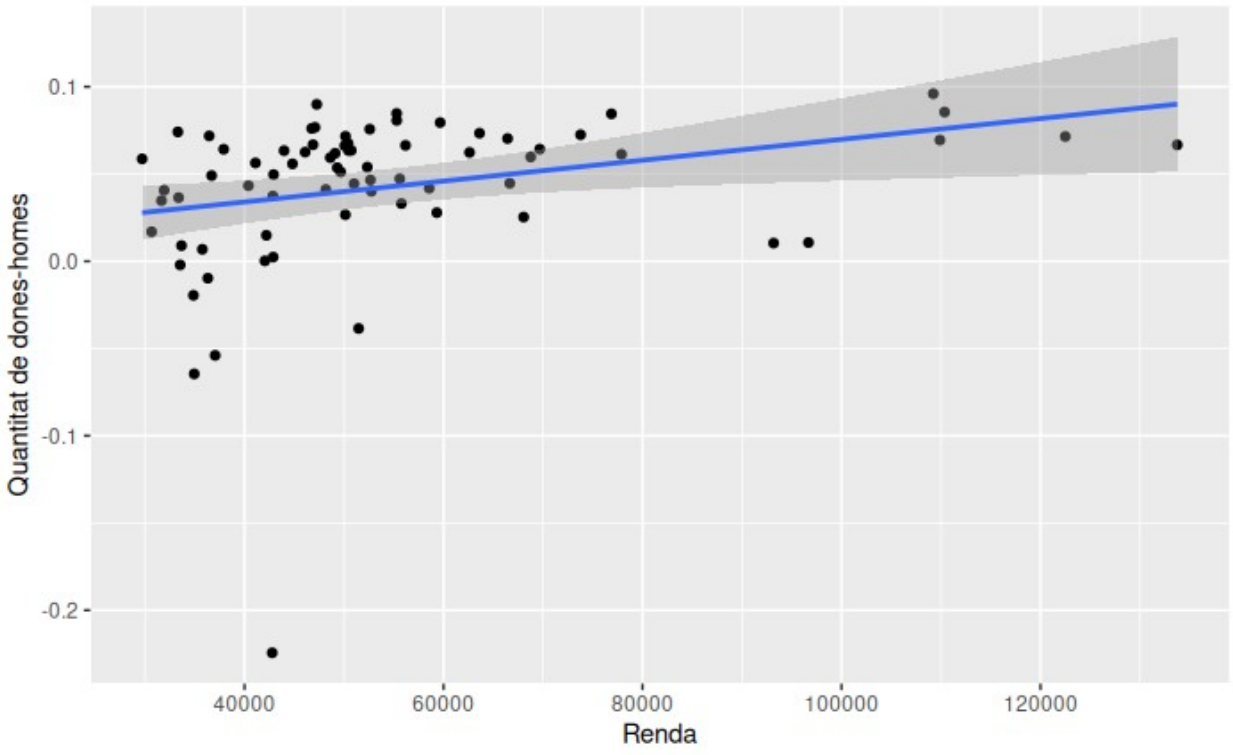
- Major desigualtat: Les Corts, i Sarrià/Sant Gervasi

- No s'ha pogut colorejar els punts segons la diferència calculada, pero s'ha fet un gràfic comparant la renta amb aquest valor:




```
ggplot(df1, aes(x=dif, y=Promedi_Renda_Bruta)) + geom_point() +
  xlab("Diferència Dones-Homes") + ylab("Promedi Renda") +
  geom_smooth(method = lm)
```

Es pot veure com hi ha una petita correlació entre la diferència i el promedi de renda, fent servir regressió lineal



ggplot(df1, aes(x=Promedi_Renda_Bruta, y=dif)) + geom_point()+

xlab("Renda") + ylab("Quantitat de dones-homes") +

geom_smooth(method = lm)

S'observa com hi ha correlació, encara que no molt forta, entre les dos variables renda i la diferència de sexes

PART 2 (2.5 pts)

Dataframe: BoxOffice_Performance.csv del conjunt de dades de *Marvel Cinematic Universe* de Kaggle***.*** Dataframe centrat em les pel·lícules de superherois produïdes per Marvel Studios.

Tot i que no és necessari llegir aquest requadre per fer els exercicis, per si us ajuda, el dataframe inclou variables com:

- **movie_id (Primary Key, Foreign Key)**: un **identificador únic** per a cada pel·lícula
- **movie_name**: Nom de la pel·lícula.
- **Phase_id (Foreign Key)**:Fa referència a la fase o la fase de l'univers cinematogràfic (com "Fase 1", "Fase 2", etc., en el cas de l'Univers Cinematogràfic Marvel). Gràcies a un altre dataframe saben que *Phase 1: Avengers Assembled, Phase 2: Ultron Revolution, Phase 3: Infinity Saga, Phase 4: Multiverse Saga, Phase 5: Multiverse Saga Continues*
- **worldwide_box_office**: Ingressos totals generats a nivell mundial per la pel·lícula (taquilla global).
- **domestic_box_office**: Ingressos generats només dins del mercat domèstic, sovint el país d'origen de la pel·lícula (per exemple, als EUA en el cas de moltes pel·lícules).
- **international_box_office**: Ingressos generats als mercats internacionals fora del país d'origen.
- **opening_weekend**:Ingressos generats durant el **cap de setmana d'estrena** de la pel·lícula.
- **production_budget**: Pressupost de producció de la pel·lícula, és a dir, els diners invertits en la realització de la pel·lícula.
- **Ràtios Financers Calculats:** A més de les mètriques directes de taquilla i pressupost, la taula també calcula diversos **ràtios financers** per proporcionar més informació sobre el rendiment financer de les pel·lícules. Alguns d'aquests ràtios inclouen:

  Opening Weekend Gross Percentage: Aquest ràtio calcula el **percentatge de la taquilla del cap de setmana d'estrena** en comparació amb els ingressos totals de la pel·lícula. Per exemple, si la pel·lícula guanya 50 milions el cap de setmana d'estrena i finalment genera 500 milions, aquest ràtio serà del 10%.

  International Collection Percentage: Aquest ràtio indica **quina part dels ingressos mundials prové dels mercats internacionals**. Si una pel·lícula guanya 100 milions a nivell internacional i 200 milions mundialment, aquest ràtio serà del 50%.

NOTA: En els exercicis d'aquesta part feu ús de les *pipes.*

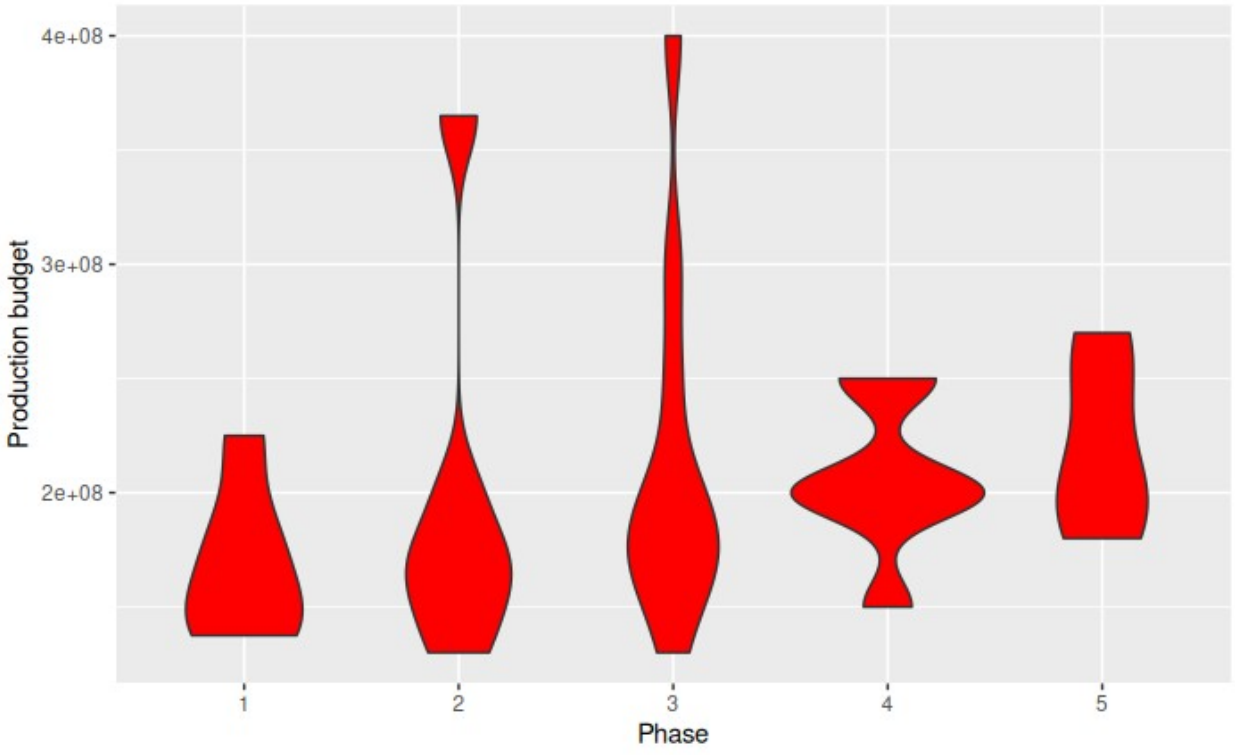
2.0 *Carregueu les llibreries necessàries i el dataframe:*

RESPOSTA:

```
library(dplyr)
df2 <- read.csv("/home/dakur/Desktop/examen/BoxOffice_Performance.csv")
```

2.1 **(1 pt)** Mostra en un mateix gràfic la distribució del pressupost de producció ('production budget') per cadascuna de les diferents fases ('Phase_id'). Justifica l'elecció del gràfic i indica quina variable és la variable de resposta i quina la d'agrupament.

RESPOSTA:




```
df2 <- read.csv("/home/dakur/Desktop/examen/BoxOffice_Performance.csv")
ggplot(df2, aes(x=factor(Phase_id), y=production_budget)) + geom_violin(fill="red") +
  xlab("Phase") + ylab("Production budget")
```

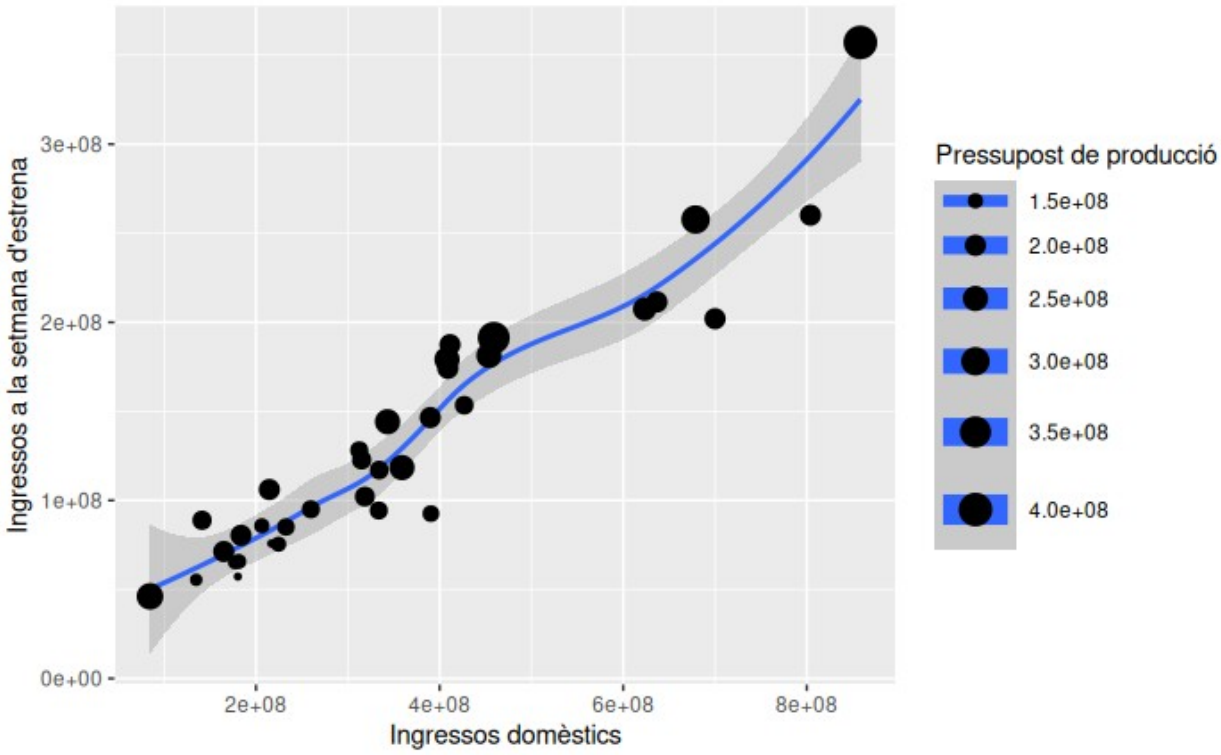
He escollit el violin plot ja que permet mostrar de forma intuitiva la distribució del pressupost de les películes segons cada fase (cada violí).

2.2. **(1.5 pts)** Ajuda't d'un gràfic per respondre les següents preguntes:

a) Hi ha una relació entre els ingressos totals generats a nivell domèstic ('domestic_box_office') i els ingressos del cap de setmana d'estrena ('opening_weekend')? (0.5 pt)

b) I entre les variables anteriors i el pressupost de producció ('production_budget')? (mostra-ho amb una visualització on es vegin les tres variables en un mateix gràfic) (1 pt)

RESPOSTA:




```
ggplot(df2, aes(x=domestic_box_office, y=opening_weekend, size=production_budget)) +
  geom_smooth() +
  geom_point() +
  xlab("Ingressos domèstics")+ ylab("Ingressos a la setmana d'estrena")+
  labs(size="Pressupost de producció")
```

Veiem que hi ha una forta relació entre els ingressos domèstics i els de la setmana d'estrena, amb la regressió de la gràfica. Visualment podem observar com el pressupost de producció (mida dels punts) no sembla estar correlacionat amb cap de les dues variables (ingressos domèstics o ingressos al cap de setmana)

PART 3 (2.5 pts)

Continuem amb el mateix dataframe que en la part 3: BoxOffice_Performance.csv del conjunt de dades de *Marvel Cinematic Universe* de Kaggle***.*** Dataframe centrat em les pel·lícules de superherois produïdes per Marvel Studios.

3.1. **(2.5 pt)**

a) Fes un codi que et permeti saber quines pel·lícules de la fase 4 i 5 van tenir uns ingressos generats el primer cap de setmana ('opening_weekend') superior a un cinquè del preu de producció ('production_budget') (0.5 pt)

b) Fent servir les dades de l'apartat (a). En un gràfic multipanell, compara el percentatge de la taquilla del cap de setmana d'estrena ('Opening_Weekend_Gross_Percentage') per cada pel·lícula, tot separant les pel·lícules de la fase 4 i de la fase 5 (1 pt)

c) Fes un gràfic amb la mateixa informació que a l'apartat (b), sense usar el multipanell (0.5 pt)

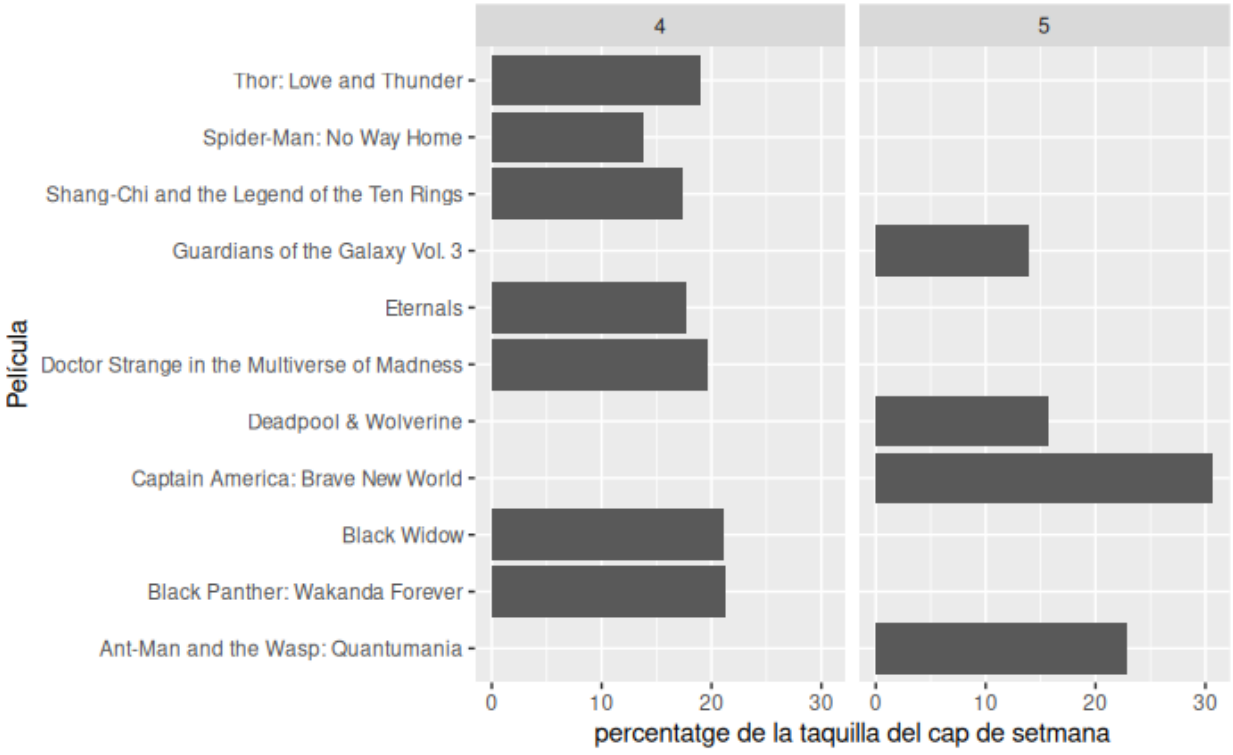
d) Fent servir les dades de l'apartat (a). Quantes d'aquestes pel·lícules tenen a 'Avengers' en el títol de la pel·lícula ('movie_name')? (Dona la resposta fent servir data massatge i mostra el codi que fas servir) (0.5 pt)

RESPOSTA:

a)

```
df3 <- df2%>%filter(Phase_id>=4) %>%filter(opening_weekend > 1/5*production_budget)
```

b)

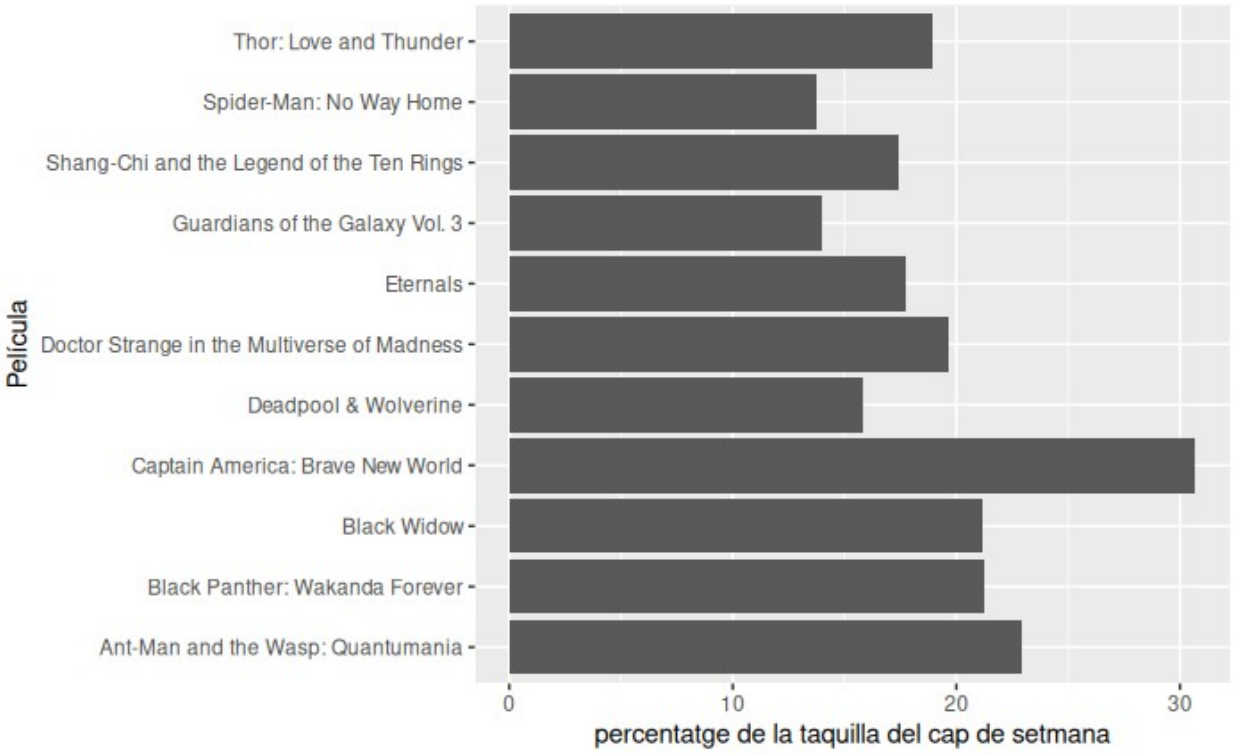


```
df3 <- df2%>%filter(Phase_id>=4) %>%filter(opening_weekend >
1/5*production_budget)


xfase <- df3%>%group_by(Phase_id)


ggplot(xfase, aes(x=movie_name,
y=Opening_Weekend_Gross_Percentage)) + geom_col()+
  facet_wrap(Phase_id ~ .) +
  ylab("percentatge de la taquilla del cap de setmana") +
  xlab("Película") + coord_flip()
```

c)




```
df3 <- df2%>%filter(Phase_id>=4) %>%filter(opening_weekend >
1/5*production_budget)


xfase <- df3%>%group_by(Phase_id)


ggplot(xfase, aes(x=movie_name,
y=Opening_Weekend_Gross_Percentage)) + geom_col()+
  #facet_wrap(Phase_id ~ .) +
  ylab("percentatge de la taquilla del cap de setmana") +
  xlab("Película") + coord_flip()
```

d)

filter(df3, grepl("Avengers", movie_name))

dona 0 resultats.