Examen-Recuperacio-1-Parcial-Sol...

annahidalgo

Visualització de Dades

3º Grado en Ingeniería de Datos

Escuela de Ingeniería
Universidad Autónoma de Barcelona

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

Nom i Cognom:_____

NIU:_____ Grup de Matrícula:_____

PARTE 1 (2 pt)

Dataset: filmdeathcounts.csv

1.1. (0.5pt) Abre el fichero. ¿Qué tipo de atributo son: *Film, Year, Body_Count y IMDB_Rating*? ¿Qué atributo usaríamos como key del dataset (valor único)?

RESPOSTA:

Film=Categórico,Year=Cuantitativo temporal, Body_Count=cuantitativo, IMDB_Rating=Cuantitativo

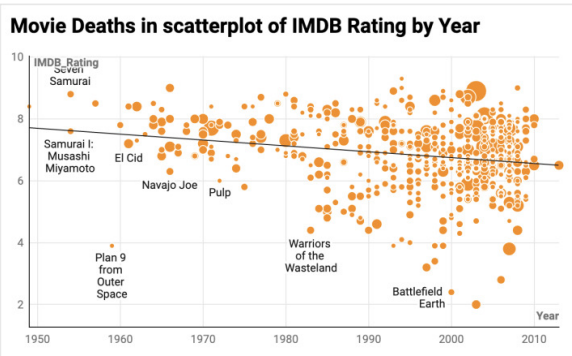
Ninguno de los atributos tiene valores únicos. El key tendría que ser un ID basado en el número de fila.

1.2. (1.5pt) Teniendo en cuenta el tipo de atributos, ¿cuál sería la gráfica óptima para explorar la relación entre *Year, Body_Count y IMDB_Rating* de todas las películas? Justifica tu respuesta.

Haz la gráfica con los ejes correctamente nombrados y numerados.

RESPOSTA:

La gràfica más adecuada es un bubble Chart (un scatter plot con un tercer atributo cuantitativo codificado en el tamaño de cada punto) porque tenemos 3 attrs. cuantitativos y queremos ver correlaciones entre ellos. Lo hicimos como pràctica en la clase de teoria 2.



PARTE 2 (3 pt)

Dataset: life_expectancy.csv

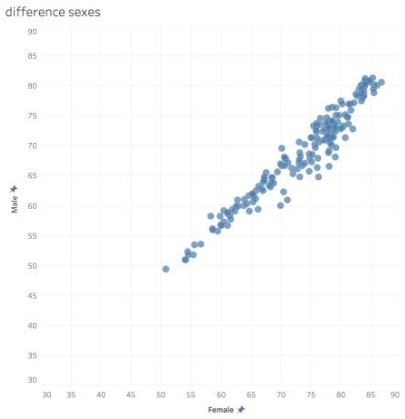
2.1. (1.5pt) Crea un subset del año 2015, con los valores de *Male*, *Female* y Both_sexes de todos los países.

¿Cual sería la forma óptima de visualizar la correlación entre las columnas Male y Female? (todos los valores). **Razona tu respuesta con relación al framework Datos/Tareas/Codificación.** (tipo de atributos, tarea, etc).

Haz la gráfica con los ejes nombrados y numerados correctamente, y adjunta la imagen.

RESPOSTA:

La forma óptima es un scatter plot porque: Utiliza dos variables cuantitativas; Es bueno para correlaciones; y permite graficar cientos de atributos. Male iría en un eje y Female en el otro.



2.2. (1.5 pt) En la gráfica que has hecho, identifica el país con la esperanza de vida de Male y Female más baja en ese año.

Vuelve al dataset completo *life_expectancy.csv*. y haz un subset que contenga Male, Female y Both_sexes de todos los años para ese país. Haz una gráfica que permita COMPARAR la esperanza de vida de Male, Female y Both_sexes de ese país a través del tiempo.

¿Qué gráfica utilizas y porqué?

RESPOSTA:

El pais es Sierra Leone. La gráfica más adecuada es una gràfica de líneas. Cada atributo se representa con una linea de un color distinto. El eje Y se puede cortar para ver major las variaciones de las lineas.

PARTE 3 (5 pt)

Dataset: simpsons_episodes.csv. Dataset con los detalles de aproximadamente 600 episodios de los Simpson

NOTA: En los ejercicios de esta parte, hacer uso de las *pipes* y añadir al gráfico un título y etiquetas personalizadas en los ejes x e y.

3.1 (2 pt) Queremos conocer la relación entre la fecha de emisión original (original_air_date) y su respectiva clasificación en Internet Movie Database (imdb_rating)

a) **Hacer un *scatter plot* y encontrar algún patrón que se ajuste** (0.75)

b) **Repetir el ejercicio separando la información para cada una de las 3 primeras temporadas ('season') y ajustando vuestro patrón con un intervalo de confianza del 75%. Hacer la visualización separando cada una de las tres temporadas ('1', '2', '3') en una fila de un *facet* respectivamente y asignando un color lo bastante diferente para cada temporada.** (1.25)

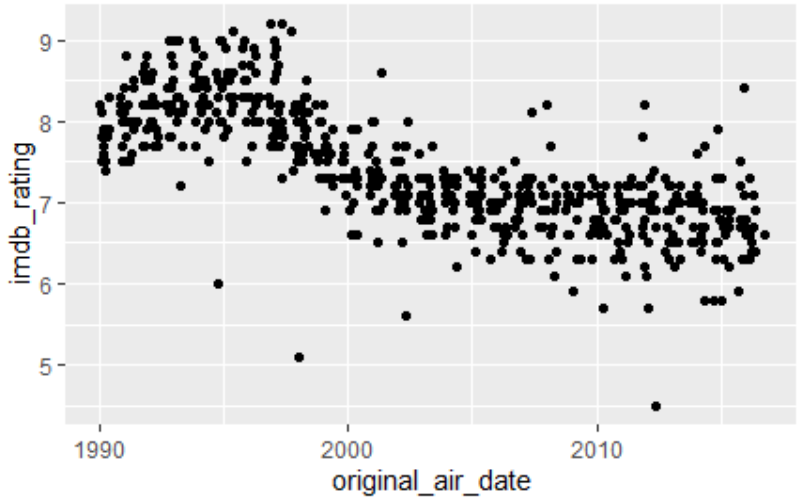
RESPOSTA:

a) Carreguem les llibreries tidyverse, dplyr i ggplot2 com sempre. I llegim el fitxer:

```
Simpsons <- read_csv('data/simpsons_episodes.csv')
```

Abans de trobar el patró fem una gràfica de punts traient els valors nans de les classificacions ('imdb_rating') i fem el scatter plot que ens demanen a l'apartat (a):

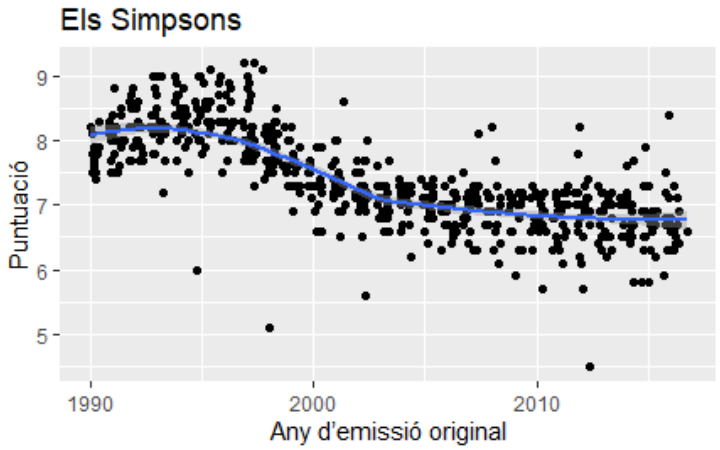
```
Simpsons %>% drop_na(imdb_rating)%>% ggplot(aes(x=original_air_date, y=imdb_rating)) +geom_point()
```



Ara provaríem diferent tipus de regressions amb geom_smooth () i canviant 'method'. Però es veu fàcilment amb el gràfic de punts que no s'ajustarà a una regressió lineal, i serà millor un ajust de regressió polinòmica local ('loess' en R) que és també la per defecte de geom_smooth

Finalment posem etiquetes als eixos i títol amb labs, (o amb `xlab` i `ylab`, i un títol amb `ggtitle`) i ajuntem totes les comandes necessàries amb pipes per evitar crear variables temporals. (Això últim s'aplica a tots els exercicis de la part 3)


```
>Simpsons %>% drop_na(imdb_rating)%>% ggplot(aes(x=original_air_date,
y=imdb_rating)) +geom_point() + geom_smooth() +labs(title=paste("Els
Simpsons"),x="Any d'emissió original", y="Puntuació")
```



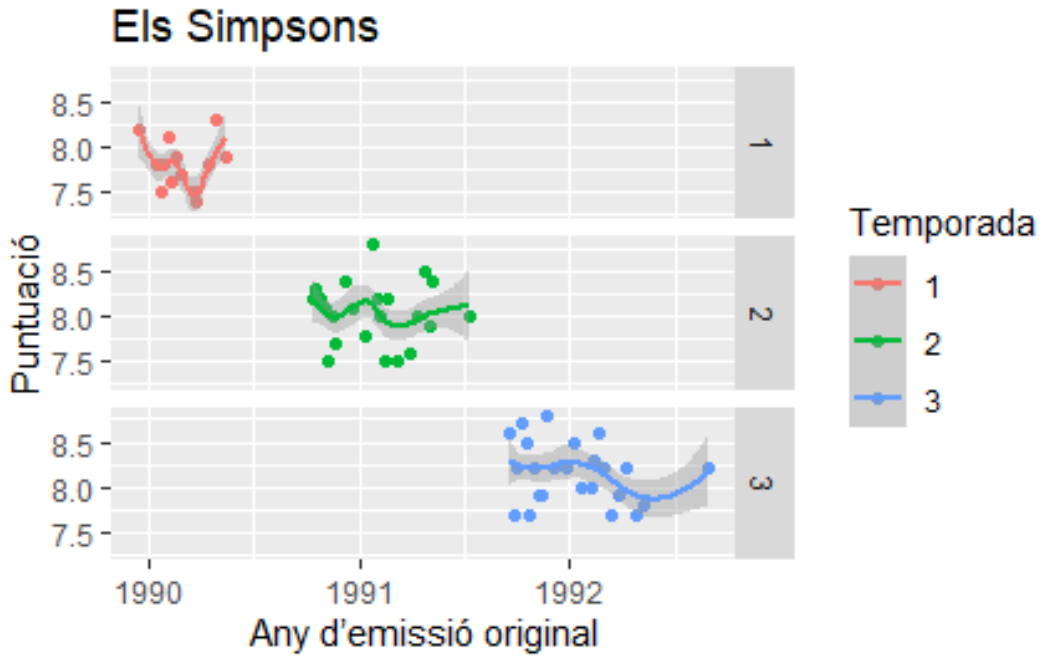
O també:

```
> Simpsons %>% drop_na(imdb_rating)%>% ggplot(aes(x=original_air_date,
y=imdb_rating)) +geom_point() + geom_smooth()+xlab('Any d'emissió
original')+ylab('Puntuació')+ggtitle('Els Simpsons')
```

B)

Molt similar a l'apartat (a), ara però necessitem: i) filtrar les temporades que ens interessen (les tres primeres); ii) posar en *aes* el mapeig del color segons la temporada que ens demanen - però tenint en compte que l'escala de color ha de ser discreta assignant un color 'lo suficientment' diferent, i sabem que per això últim necessitem factoritzar 'season' que ara és una variable numèrica contínua amb valors d'1 a 10 ; iii) especificar el % de l'interval de confiança que ens demanen fent us de 'level=0.75' en geom_smooth; iv) fer un *facet* on cada fila correspongui a una de les tres temporades, per tant ho fem amb `facet_grid` o `facet_wrap` però especificant-li que volem visualitzar el patró d'una temporada per cada fila, o en un *facet* amb sol una columna total:

```
>Simpsons%>%filter(season<4)%>%drop_na(imdb_rating)%>%
drop_na(imdb_rating)%>%ggplot(aes(x=original_air_date,   y=imdb_rating,
color=factor(season)))   +geom_point()   +   geom_smooth(level=0.75)   +
xlab('Any     d'emissió     original')     +     ylab('Puntuació')     +
facet_grid(season~      .)      +      ggtitle('Els      Simpsons')+
scale_color_discrete(name = "Temporada", labels=c("1", "2", "3"))
```



3.2. (1.5 pt) Graficar la distribución de las 10 primeras temporadas ('season') respecto al número de visualizaciones ('views'). Explicar la elección del gráfico y las conclusiones que podéis extraer del mismo.

RESPOSTA:

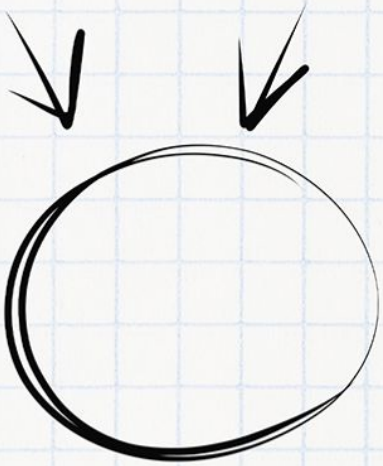
Primer necessitem filtrar les 10 primeres temporades. Sembla que no hi ha nans ni valors 'nulls' en aquestes pel que fa a les visualitzacions, per tant podem prosseguir.
En quant al gràfic, tenim una variable numèrica continua "views" i una variable 'season' que podem fer discreta categòrica amb factor i se'ns demana una distribució. En la 1a part de l'assignatura vam veure varies vegades a classe teòrica amb Guillermo i en seminaris quines gràfiques es podien usar per mostrar una distribució:

Imagínate aprobando el examen

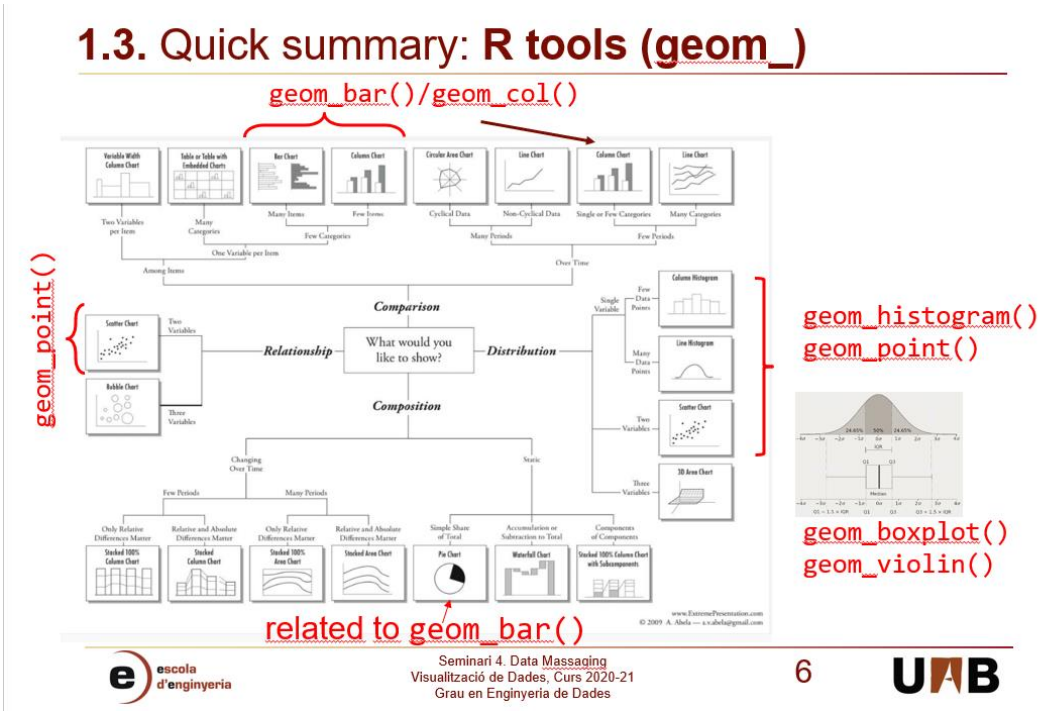
Necesitas tiempo y concentración

Planes	PLAN TURBO	PLAN PRO	PLAN PRO+
Descargas sin publi al mes	10	40	80
Elimina el video entre descargas	✓	✓	✓
Descarga carpetas	✗	✓	✓
Descarga archivos grandes	✗	✓	✓
Visualiza apuntes online sin publi	✗	✓	✓
Elimina toda la publi web	✗	✗	✓
Precios Anual	0,99 € / mes	3,99 € / mes	7,99 € / mes

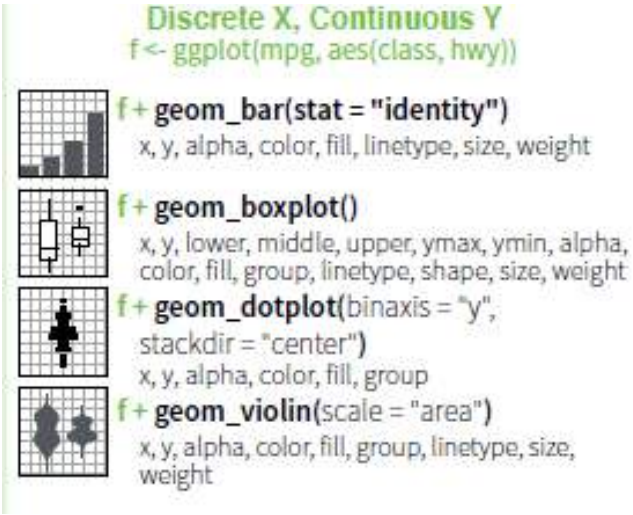
Ahora que puedes conseguirlo, ¿Qué nota vas a sacar?



WUOLAH



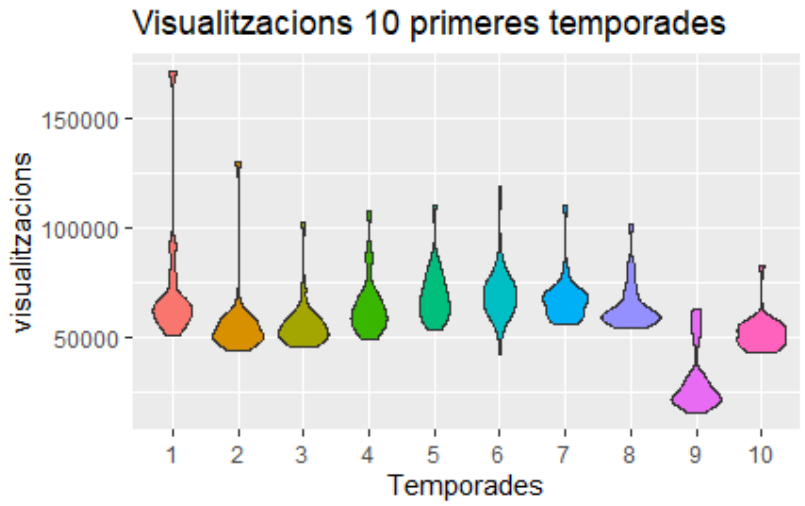
Donat el tipus de variables que tenim, descartem el histograma, ja que el histograma ens mostra la distribució d'una variable contínua (i l'única manera seria fer un *face*t amb un histograma per cadascuna de les 10 temporades, complicant el gràfic).
Entre les gràfiques de distribucions possibles, ja deuríem saber triar segons el tipus de variables, però, a més, ens podem ajudar del xuletari de ggplot de R que se'ns va donar en seminari 2. Un cop identificades els gràfic òptims per mostrar distribucions (*column histogram, line histogram, scatter chart, 3D area chart, boxplots, diagrama de violins, per exemple*), només hem de buscar quins d'ells serveixen pel nostre tipus de variables, fent servir el xuletari:



Podem doncs fer per exemple, un geom_boxplot() o un geom_violin(). Ara bé, l'enunciat s'interessa per la distribució, no ens demana ni outliers, ni quartils, ni medianes, per tant

ambdós són òptims, i si volem per exemple saber la 'forma' de la distribució geom_violin () pot ser bona elecció:

```
> Simpsons%>%filter(season<=10)%>%drop_na(views)%>%
drop_na(imdb_rating)%>%ggplot(aes(x=factor(season), y=views,
fill=factor(season))) +geom_violin()
+theme(legend.position='none')+xlab('Temporades')+ylab('visualitzacion
s')+ggtitle('Visualitzacions de les primeres temporades dels
Simpsons')
```



El gràfic ens mostra clarament una davallada de les visualitzacions durant la temporada 9, o que en la temporada 1 va haver algun episodi que va tenir més de 150000 visualitzacions, tot i que la majoria dels episodis van tenir unes 60000 visualitzacions. A més, els diagrames de violí de les tres primeres sessions tenen cues més llargues, indicant que alguns dels episodis tenien puntualment més visualitzacions que la resta de la mateixa temporada.

3.3. (1.5pt)

Graficar la distribución de:

(a) la audiencia en millones en US (us_viewers_in_millions) de todos los episodios (0.5 pts)

(b) Las visualizaciones ('views') de los episodios teniendo como personaje principal 'Homer' (es decir, que en el título aparezca 'Homer'). (1 pt)

Para ambos apartados, explicar la elección de los gráficos y las conclusiones que podéis extraer del mismo.

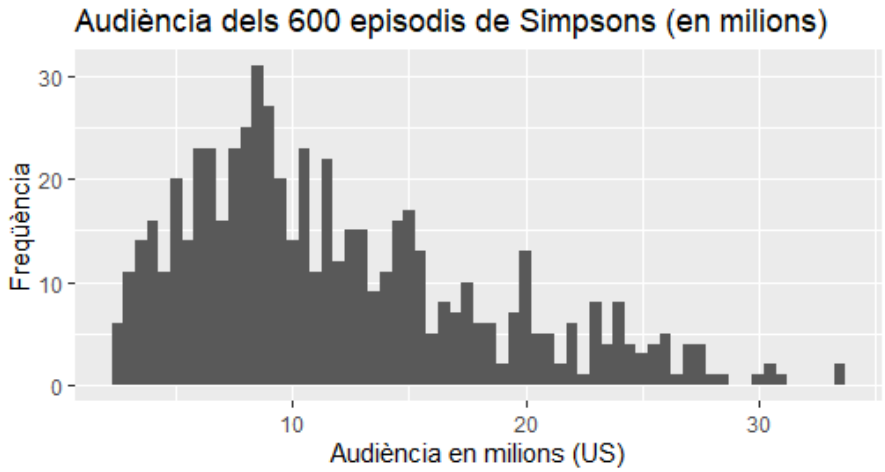
RESPOSTA:

(a i b) Altra vegada ens demanen una distribució, i ja sabem quines gràfiques ens mostren una distribució. També sabem que tant 'us_viewers_in_millions', com 'views' són variables numèriques contínues. Si encara no tenim clar quina gràfica (d'entre les de

distribucions) ens mostra la distribució d'una variable contínua, fem us del xuletari. Podem veure que per exemple geom_density() o geom_histogram() serien òptimes.

(a) Fem per exemple geom_histogram i triem un binwidth que ens sembli adequat:

```
>Simpsons%>%drop_na(us_viewers_in_millions)%>%ggplot(aes(x=us_viewers_
in_millions))+geom_histogram(binwidth = 0.5)+ xlab('Audiència en milions
(US)')+ ylab('Freqüència') +ggtitle('Audiència dels 600 episodis de
Simpsons (en milions)')
```

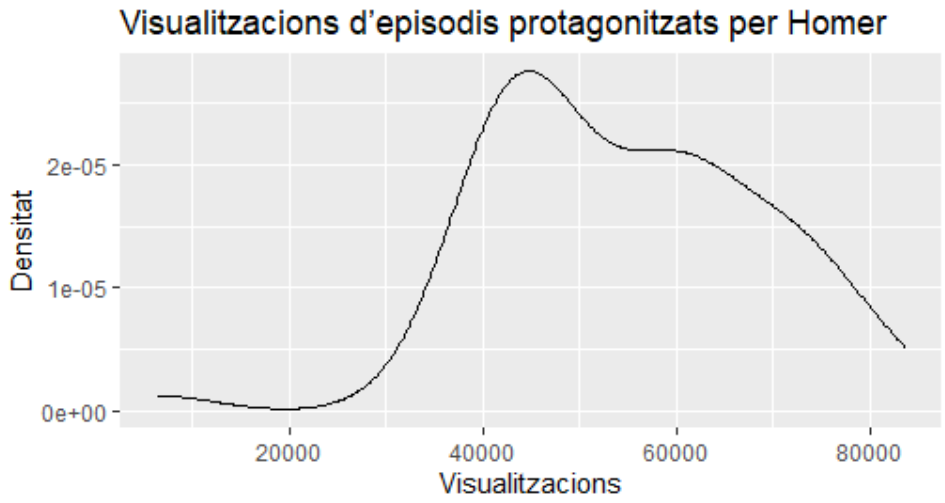


Podem veure que l'audiència va superar els 30 milions i mai va ser inferior a 2 milions, però majorment es va trobar entre els 5 i 15 milions.

(b) En aquest cas primer filtrem les files del dataframe que continguin 'Homer' en el títol de la pel·lícula. Podem utilitzar un filtre amb la funció `grepl` que vam veure.

Per variar de (a), un exemple pot ser doncs geom_density ():

```
> Simpsons%>%
filter(grepl('Homer',title))%>%ggplot(aes(x=views))+geom_density()+
xlab('Visualitzacions')+ ylab('Densitat') +ggtitle('Visualitzacions
d'episodis protagonitzats per Homer')
```



Sembla que tot i que algun no es va veure, cap d'ells es va veure unes 20000 vegades (en un histograma de fet veuríem clarament que no hi ha episodis que s'hagin vist més de 10000 vegades i menys de 30000). El que es veu clarament aquí és que la major part dels episodis amb en Homer com protagonista es van veure entre 40000 i 60000 vegades. Tot i així la densitat es baixa, però també cal puntualitzar que amb el filtratge ens hem quedat amb una mostra petita d'episodis (55 episodis)