2024Examen-1-Parcial-Solucions-M...

alucero

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 Primer Parcial – 08 Abril 2024
SOLUCIONS MODEL 2

Nom i Cognom:_____

NIU:_____ Grup de Matrícula:_____

PARTE 1 (3.5 pt)

Dataset: titanic.csv

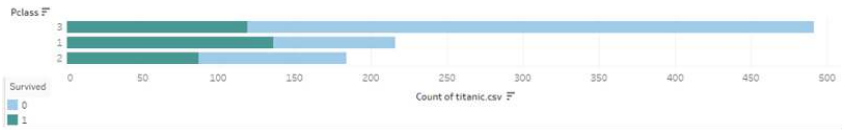
1.1. (0.5 pt) Abre el fichero. ¿Qué tipo de atributo son: ***Survived, Pclass, Sex, Age***, y ***Fare***? ¿Qué atributo es el key (clave primaria) del dataset?

RESPOSTA:

Survived=Categórico, Pclass=Ordinal, Sex=categórico, Age=Cuantitativo, Fare=Cuantitativo. El key es PassengerID

1.2. (1 pt) Haz una gráfica que muestre **en qué clase** (Pclass) hubo más sobrevivientes. Sube la gráfica y el código.

RESPOSTA:



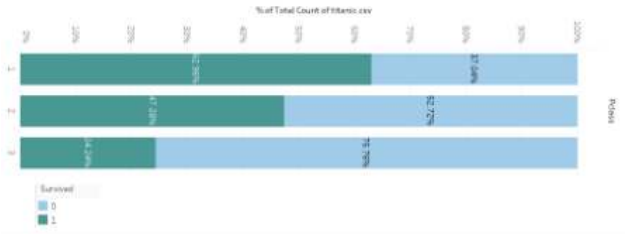
1.3 (2 pt) Queremos saber si la clase en la que viajaban los pasajeros (Pclass) influye en la **probabilidad** de sobrevivir (***Survived***). Haz una gráfica que permita visualizar esa relación y razona tu elección de acuerdo al framework Datos/Tareas/Codificación (Cuantos atributos usas y de qué tipo son; qué tarea ayuda a llevar a cabo la gráfica; marcas y canales empleados, etc).

Sube la imagen de la gráfica debidamente anotada junto al código y la respuesta.

RESPOSTA:

Survived y Pclass son de tipo categórico y categórico ordinal. Se puede visualizar como una gráfica de barras agrupadas o apiladas, donde un categórico se utiliza para determinar la posición en el eje horizontal y el otro para separar cada barra en dos secciones de colores distintos. Si se utilizan porcentajes seran barras normalizadas. Las barras normalizadas son mejores porque permiten comparar con más precisión entre clases a pesar de las diferencias en número de pasajeros.

(En R vimos en seminario 2 cómo hacer estos gráficos en R usando: geom_bar(position="dodge")



PARTE 2 (1.5 pt)

Dataset: 2017_accidents_vehicles_gu_bcn.csv

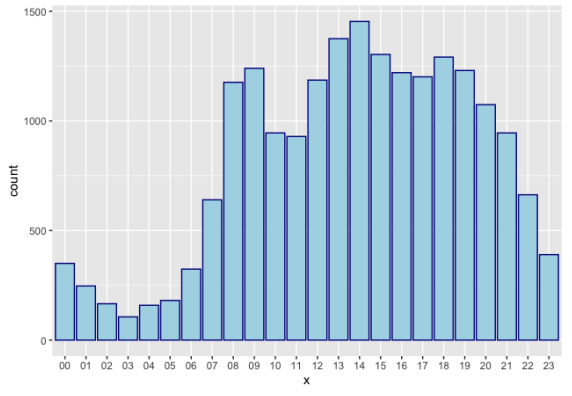
El dataset contiene registros de accidentes de tráfico en Barcelona durante 2017, con detalles como el día, hora,

2.1. (1.5 pt) Haz una gráfica que muestre a qué hora del día sucedió el mayor número de accidentes ese año. Razona brevemente el tipo de gráfica elegida y porqué la has usado en este caso en relación al framework Datos/Tareas/ Codificación. Sube la respuesta (0.5pt), la gráfica (0.5pt) y el código (0.5pt).

RESPOSTA:

En este caso se puede hacer un histograma con binSize=1 o una gráfica de barras usando "rep" para tener el conteo de cada valor. Se codifica la frecuancia en el tamaño de las barras y se ordenan por hora, de este modo es fácil ver las horas con mayor número de accidentes y compararlas entre ellas.

```
hour <- df$'Hora_de_dia'
# Histogram of Hour
ggplot(df, aes(x=as.numeric(hour))) +
  geom_histogram(binwidth=1, color="darkblue", fill="lightblue")
# Barplot of Hour
df <- data.frame(x = rep(hour))
ggplot(df, aes(x)) + geom_bar(color="darkblue", fill="lightblue")
```



PARTE 3 (5 pt)

Dataframe: 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*.

3.1 **(2 pt)** Queremos conocer la relación entre la fecha de emisión original (***original_air_date***) y los votos que ha obtenido en la Internet Movie Database (***imdb_votes***)

a) Hacer un gráfico que os permita ver la relación entre ambas variables y encontrar algún patrón que se le ajuste. Eliminar valores NAN y/o outliers. (0.75 pt)

b) Hacer una visualización multipanel que os permita analizar cómo varía la relación entre las variables (***original_air_date*** y ***imdb_votes***) a lo largo de las tres primeras temporadas. Para lograr esto, vamos a dividir nuestra visualización en tres paneles, uno para cada temporada ('1', '2', '3'). En cada panel, mostraréis, para cada temporada, la relación entre las variables y ajustaremos el patrón con el intervalo de confianza del 75%. ¿Os aporta algo colorear de un color distinto cada panel o obtenéis la misma información si dejáis los tres paneles sin colorear (razonad la respuesta)? (1.25 pt. Nota: si no se sabe hacer el multipanel mostrar cada gráfico por separado se puntuará 0.75 pt máximo)

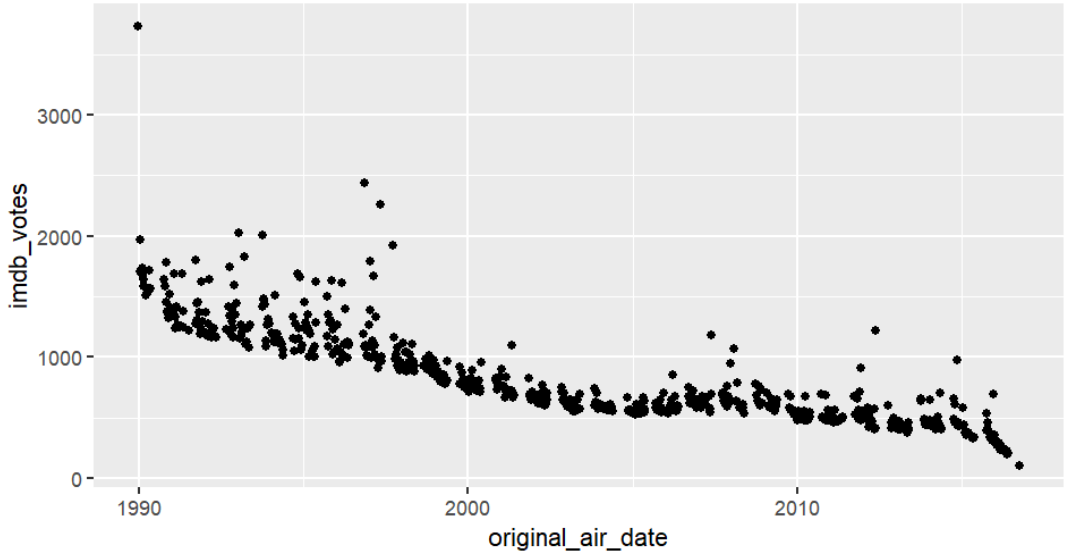
RESPOSTA:

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

```
simpsons <- read_csv('data/simpsons_episodes.csv') #o via el Environment
```

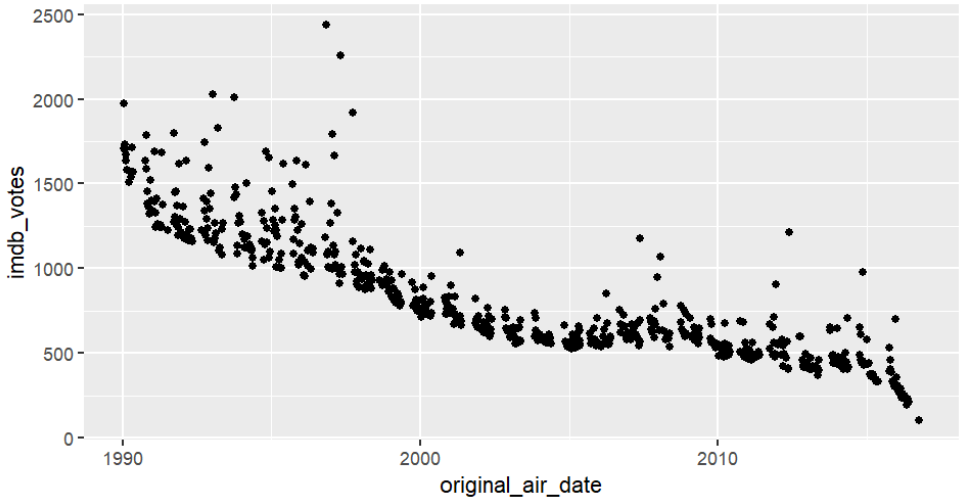
Abans de fer la gràfica de punts veiem que si hi ha NANs per treure's i fem el scatter plot que ens demanen a l'apartat (a):

```
simpsons %>% drop_na(imdb_votes)%>%ggplot(aes(x=original_air_date, y=imdb_votes)) +geom_point()
```



Traiem l'outlier que hi ha (vots per sobre 3000), que ens facilitarà el patró:

```
simpsons %>%filter(imdb_votes<3000)%>%ggplot(aes(x=original_air_date , y=imdb_votes)) +geom_point()
```



Traiem l'outlier que hi ha (vots per sobre 3000), que ens facilitarà el patró:

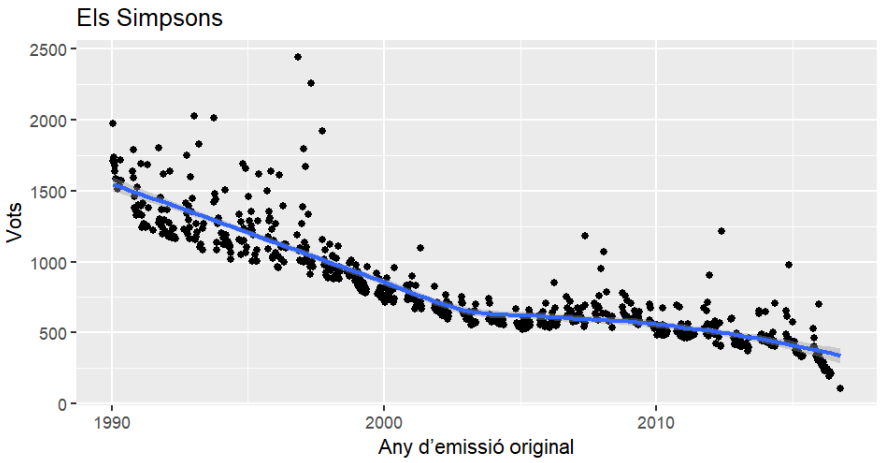
```
simpsons %>%filter(imdb_votes<3000)%>%ggplot(aes(x=original_air_date
, y=imdb_votes)) +geom_point()
```

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)

El patró per defecte ('loess') del geom_smooth ja s'ajusta prou bé:

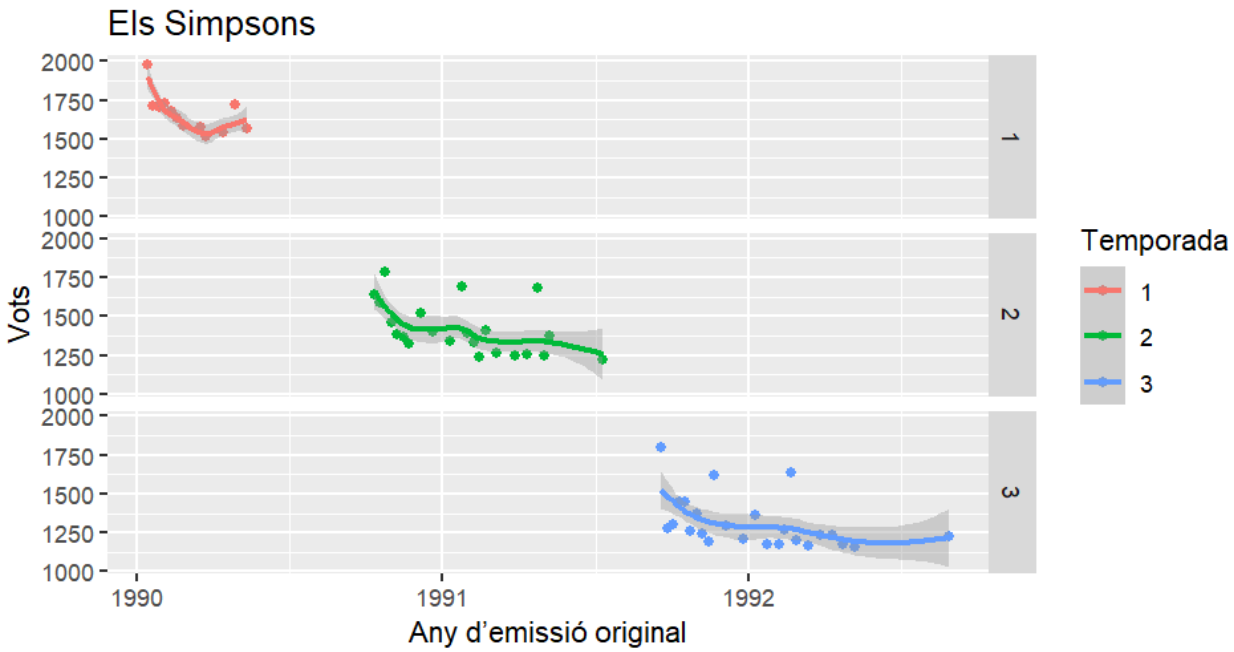
```
simpsons %>%filter(imdb_votes<3000)%>%ggplot(aes(x=original_air_dat
e, y=imdb_votes)) +geom_point()+ geom_smooth() +labs(title=paste("El
s Simpsons"),x="Any d'emissió original", y="Vots")
```



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)%>%filter(imdb_votes<3000)%>%ggplot(aes(x=ori
ginal_air_date, y=imdb_votes, color=factor(season))) +geom_point() +
geom_smooth(level=0.75) + xlab('Any d'emissió original') + ylab('Vots') +
facet_grid(season~ .) + ggtitle('Els Simpsons')+
scale_color_discrete(name = "Temporada", labels=c("1", "2", "3"))
```



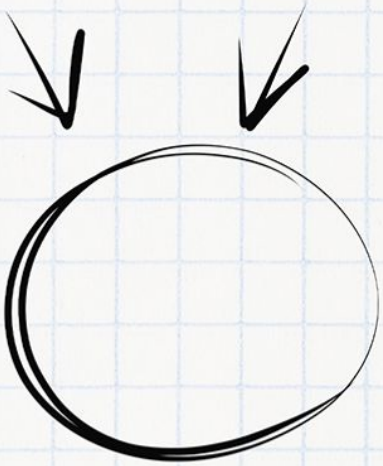
El color no ens aporta informació extra a no colorejar-lo ja que cada panel (fila de la visualització) ja mostra cada temporada

Imagínate aprobando el examen

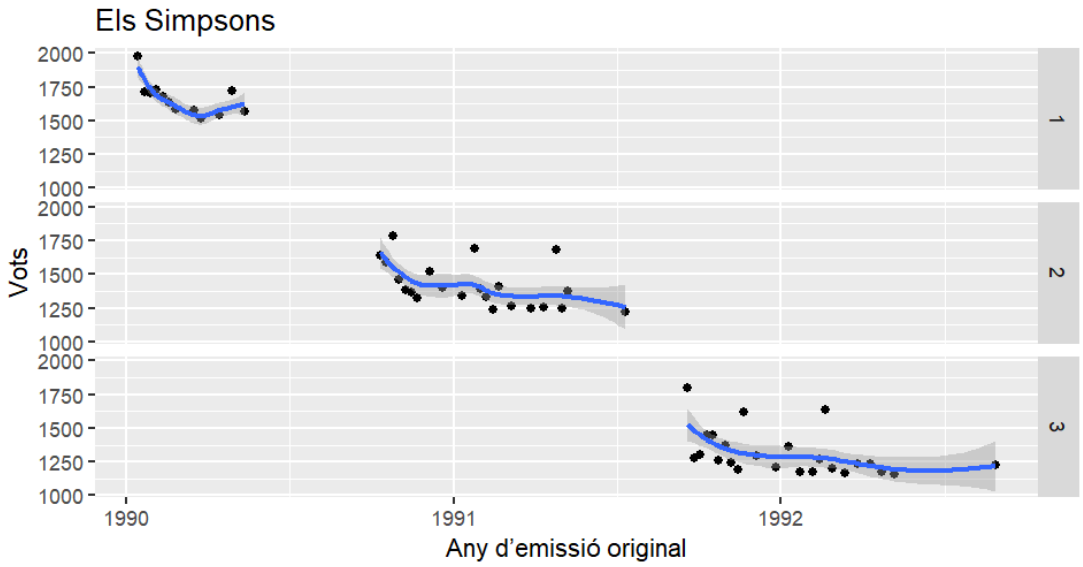
Necesitas tiempo y concentración

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?



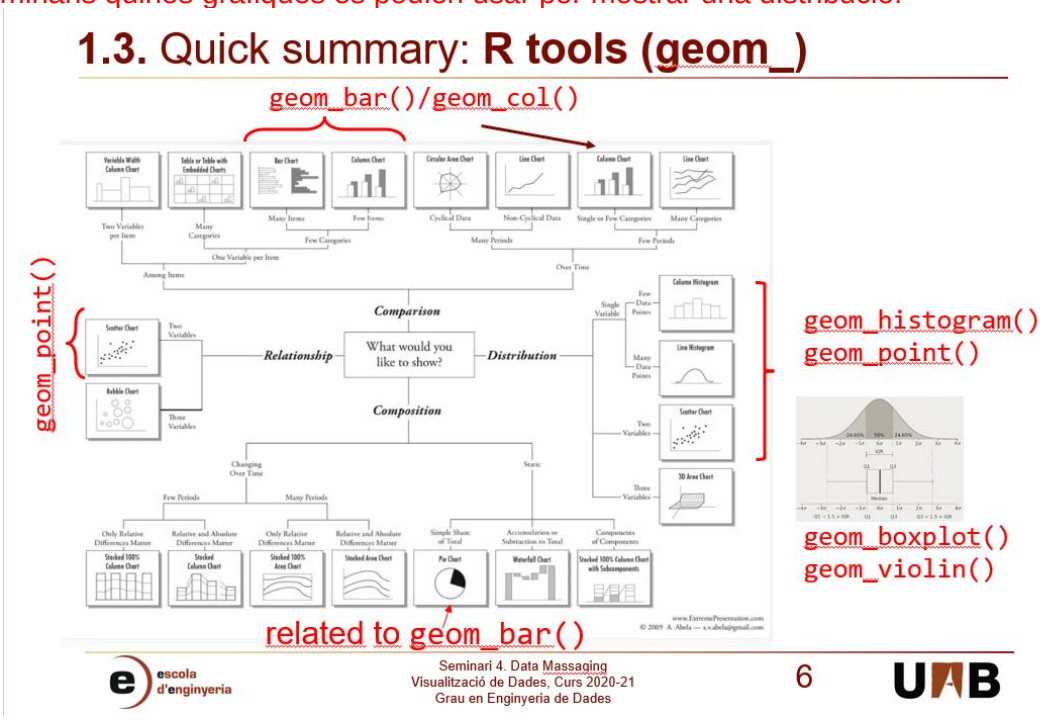
WUOLAH



3.2. (1.25 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:

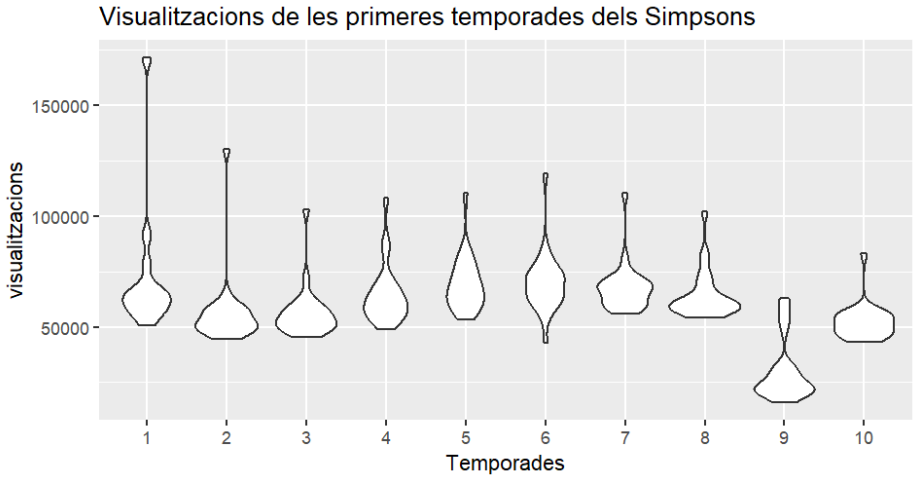
Primer necessitem filtrar les 10 primeres temporades. Sembla que no hi ha nans ni valors 'nulls' en aquestes pel que fa a les visualitzacions, per tant podem prosseguir.
En quant al gràfic, tenim una variable numèrica continua "views" i una variable 'season' que podem fer discreta categòrica amb factor 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:



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)%>%ggplot(aes(x=factor(season), y=views)) +geom_violin() +theme(legend.position='none')+xlab('Temporades')+ylab('visualitzacions')+ggtitle('Visualitzacions de les primeres temporades dels Simpsons')
```

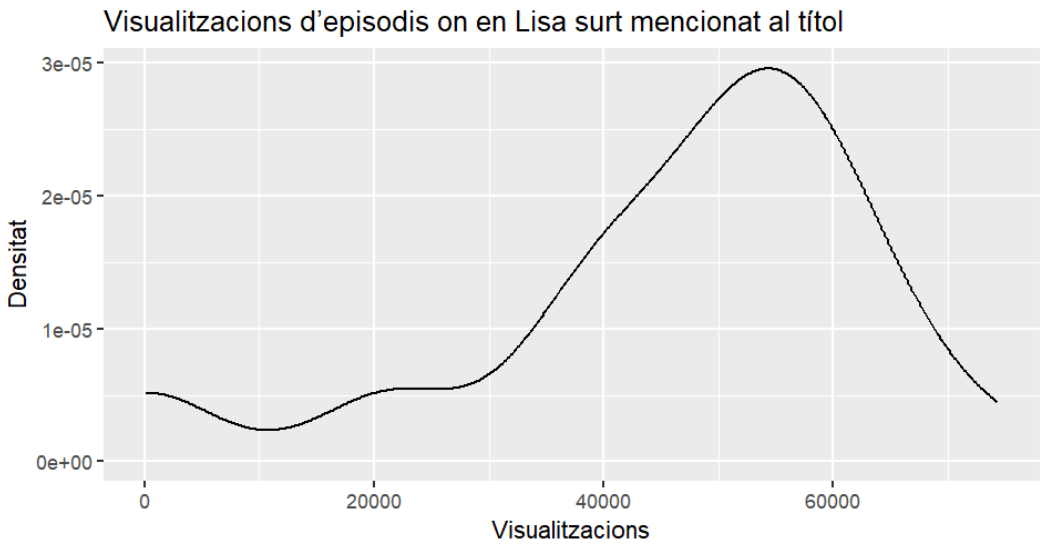


El gràfic ens mostra clarament una davallada de les visualitzacions durant la temporada 9, o que en la temporada 1 va haver algun episodi que va tenir més de 150000 visualitzacions, tot i que la majoria dels episodis van tenir unes 60000 visualitzacions. A més, els diagrames de violí de les tres primeres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

3.3. (1 pt) Graficar la distribución del número de visualizaciones ('views') para los episodios en que el nombre de Lisa parece en el título ('tittle'). ¿Cuántos episodios son? Extrae alguna conclusión del gráfico

```
simpsons%>% filter(grepl('Lisa',title))%>%ggplot(aes(x=views))+geom_
density()+ xlab('Visualitzacions')+ ylab('Densitat') +ggtitle('Visua
litzacions d'episodis on en Lisa surt mencionat al títol')
```

Fent el datamassage amb grepl veiem que són 38 episodis



El que es veu clarament aquí és que la major part dels episodis amb la Lisa apareixent en el títol de l'episodi es van veure entre 40000 i 60000 vegades. Tot i així cal puntualitzar que amb el filtratge ens hem quedat amb una mostra petita d'episodis (38 episodis)

3.4. (0.75 pt)

a) Di al menos dos tipos de gráficos que permitan reducir la dimensionalidad (0.25 pt).
Per exemple, LDA i PCA

b) Considera un conjunto de datos que contiene información sobre la composición química de muestras de diferentes tipos de vinos. Cada muestra puede tener docenas o incluso cientos de características que describen la presencia y la concentración de diferentes compuestos químicos en el vino. Sin embargo, no tenemos información sobre la clase de vino a la que pertenece cada muestra. ¿Qué gráfico sería apropiado para reducir la dimensionalidad de los datos al identificar las direcciones (o componentes) principales de variación en el conjunto de datos sin tener en cuenta las etiquetas de clase? (0.5 pt).
El PCA, de hecho, podría ayudar a simplificar la interpretación de los datos y a identificar patrones generales de variación en la composición química del vino.
LDA al no tener info de clases no es apropiado.