Informe HDT1

Marco Ramirez 19588, Alfredo Quezada 191002, Estuardo Hernandez 19202

04/02/2022

Hoja de Trabajo 1: Analisis exploratorio

El objetivo de esta hoja de trabajo es realizar un analisis exploratorio a la base de datos de peliculas, extraido de IMDB. Dicha base cuenta de datos con 10000 filas y con 27 de columnas.
Para verificar nuestro codigo y lo que hicimos paso por paso puede consultar en nuestro repositorio. Exactamente en el archivo ScriptsFinal podra ver la documentacion de cada pregunta respondida, ademas, dentro del repositorio podra ver el historial de cambios de este informe.

LINK DE GITHUB (https://github.com/MarcoRamirezGT/HT1_Mineria)

1. Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

```
summary(datos)
```

```
##        id             budget             genres            homePage
##  Min.   :     5   Min.   :        0   Length:10000       Length:10000
##  1st Qu.: 12286   1st Qu.:        0   Class :character   Class :character
##  Median :152558   Median :   500000   Mode  :character   Mode  :character
##  Mean   :249877   Mean   : 18551632
##  3rd Qu.:452022   3rd Qu.: 20000000
##  Max.   :922260   Max.   :380000000
##  productionCompany  productionCompanyCountry productionCountry
##  Length:10000       Length:10000             Length:10000
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##     revenue             runtime        video           director
##  Min.   :0.000e+00   Min.   :  0.0   Mode :logical   Length:10000
##  1st Qu.:0.000e+00   1st Qu.: 90.0   FALSE:9430      Class :character

##  Median :1.631e+05   Median :100.0   TRUE :84        Mode  :character
##  Mean   :5.674e+07   Mean   :100.3   NA's :486
##  3rd Qu.:4.480e+07   3rd Qu.:113.0
##  Max.   :2.847e+09   Max.   :750.0
##     actors          actorsPopularity   actorsCharacter    originalTitle
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     title           originalLanguage     popularity        releaseDate
##  Length:10000       Length:10000       Min.   :    4.258   Length:10000
##  Class :character   Class :character   1st Qu.:   14.578   Class :character
##  Mode  :character   Mode  :character   Median :   21.906   Mode  :character
##                                        Mean   :   51.394
##                                        3rd Qu.:   40.654
##                                        Max.   :11474.647
##     voteAvg         voteCount      genresAmount    productionCoAmount
##  Min.   : 1.300   Min.   :    1   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 5.900   1st Qu.:  120   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 6.500   Median :  415   Median : 3.000   Median : 3.000
##  Mean   : 6.483   Mean   : 1342   Mean   : 2.596   Mean   : 3.171
##  3rd Qu.: 7.200   3rd Qu.: 1316   3rd Qu.: 3.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :30788   Max.   :16.000   Max.   :89.000
##  productionCountriesAmount  actorsAmount    castWomenAmount
##  Min.   :  0.000           Min.   :     0   Length:10000
##  1st Qu.:  1.000           1st Qu.:    13   Class :character
##  Median :  1.000           Median :    21   Mode  :character
##  Mean   :  1.751           Mean   :  2148
##  3rd Qu.:  2.000           3rd Qu.:    36
##  Max.   :155.000           Max.   :919590
##  castMenAmount
##  Length:10000
##  Class :character
```

```
##  Class :character
##  Mode  :character
##
##
##
```

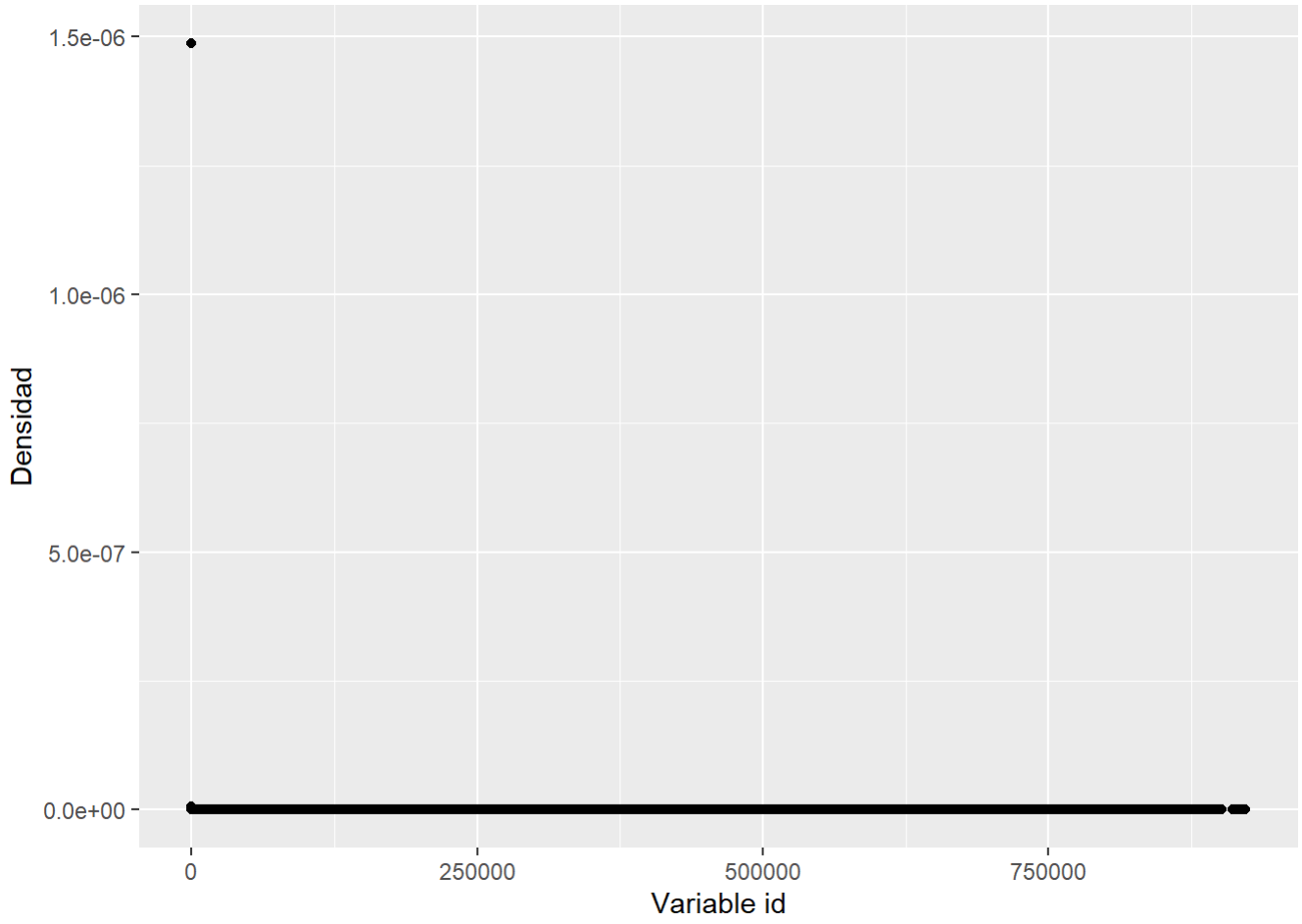
2. Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)

VARIABLES	TIPO DE VARIABLES
Id de la película	Cualitativa Nominal
- popularity: Índice de popularidad de la película calculado semanalmente	Cuantitativa Continua
- budget: El presupuesto para la película.	Cuantitativa Discreta
- revenue: El ingreso de la película.	Cuantitativa Discreta
- original_title: El título original de la película, en su idioma original.	Cualitativa Nominal
- originalLanguage: Idioma original en que se encuentra la película	Cualitativa Nominal
- title: El título de la película traducido al inglés	Cualitativa Nominal
- homePage: La página de inicio de la película	Cualitativa Ordinal
- video: Si tiene videos promocionales o no	Cualitativa Nominal
- director: Director de la película	Cualitativa Nominal
- runtime: La duración de la película.	Cuantitativa Discreta
- genres: El género de la película.	Cualitativa Nominal
- genresAmount: Cantidad de géneros que representan la película	Cualitativa Ordinal
- productionCompany: Las compañías productoras de la película.	Cualitativa Nominal
- productionCoAmount: Cantidad de compañías productoras que participaron en la película	Cuantitativa Discreta
- productionCompanyCountry: Países de las compañías productoras de la película	Cualitativa Ordinal
- productionCountry: Países en los que se llevó a cabo la producción de la película	Cuantitavia Discreta
- productionCountriesAmount: Cantidad de países en los que se rodó la película	Cuantitavia Discreta
- releaseDate: Fecha de lanzamiento de la película	Cualitativa Ordinal
- voteCount: El número de votos en la plataforma para la película.	Cuantitativa Discreta
- voteAvg: El promedio de los votos en la plataforma para la película	Cuantitativa Continua
- actors: Actores que participan en la película (Elenco)	Cualitativa Nominal
- actorsPopularity: Índice de popularidad del elenco de la película.	Cualitativa Ordinal
- actorsCharacter: Personaje que interpreta cada actor en la película	Cualitativa Nominal
- actorsAmount: Cantidad de personas que actúan en la película	Cuantitativa Discreta
- castWomenAmount: Cantidad de actrices en el elenco de la película	Cuantitativa Discreta
- castMenAmount: Cantidad de actores en el elenco de la película	Cuantitativa Discreta

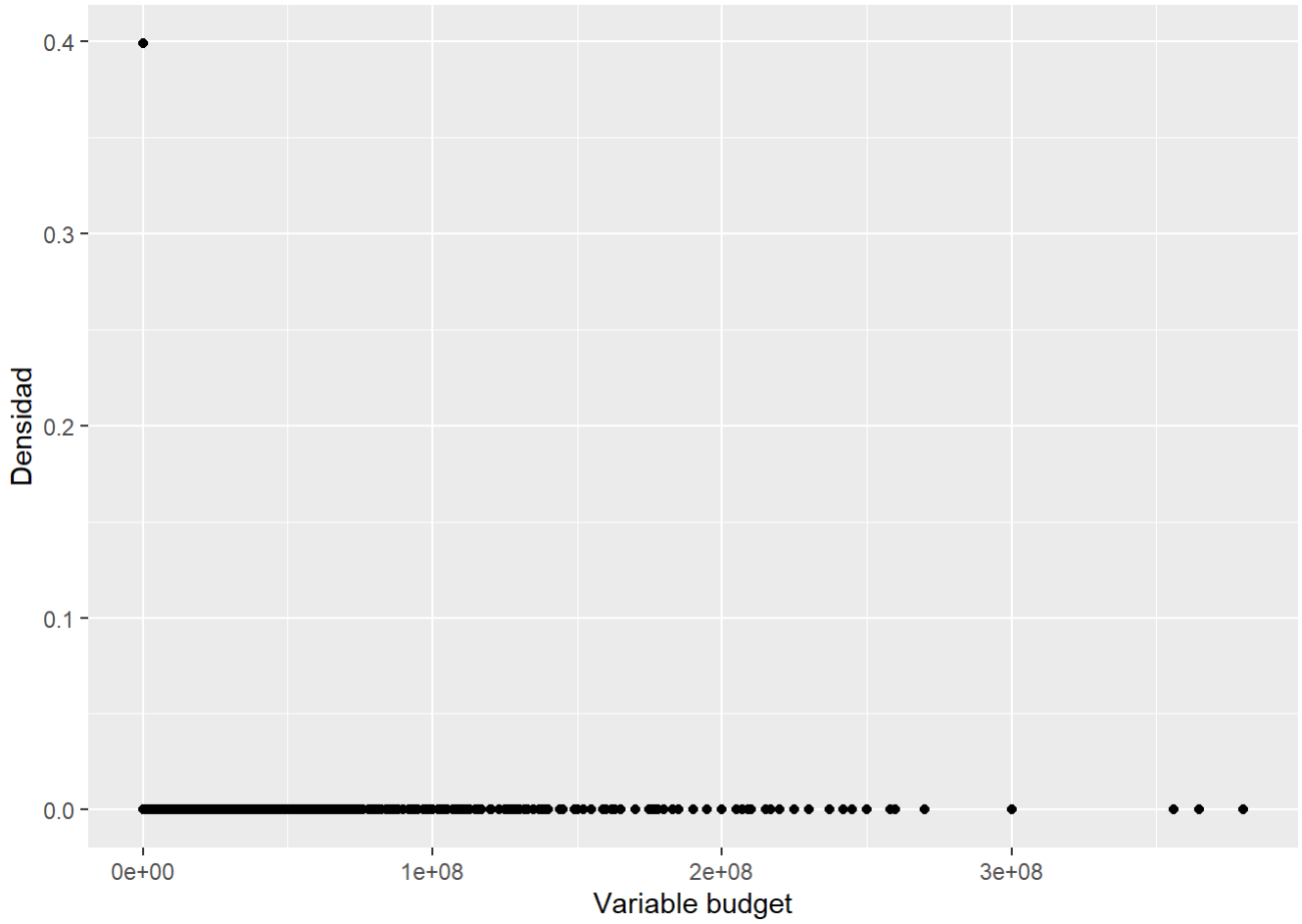
Caption for the picture.

3. Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados.

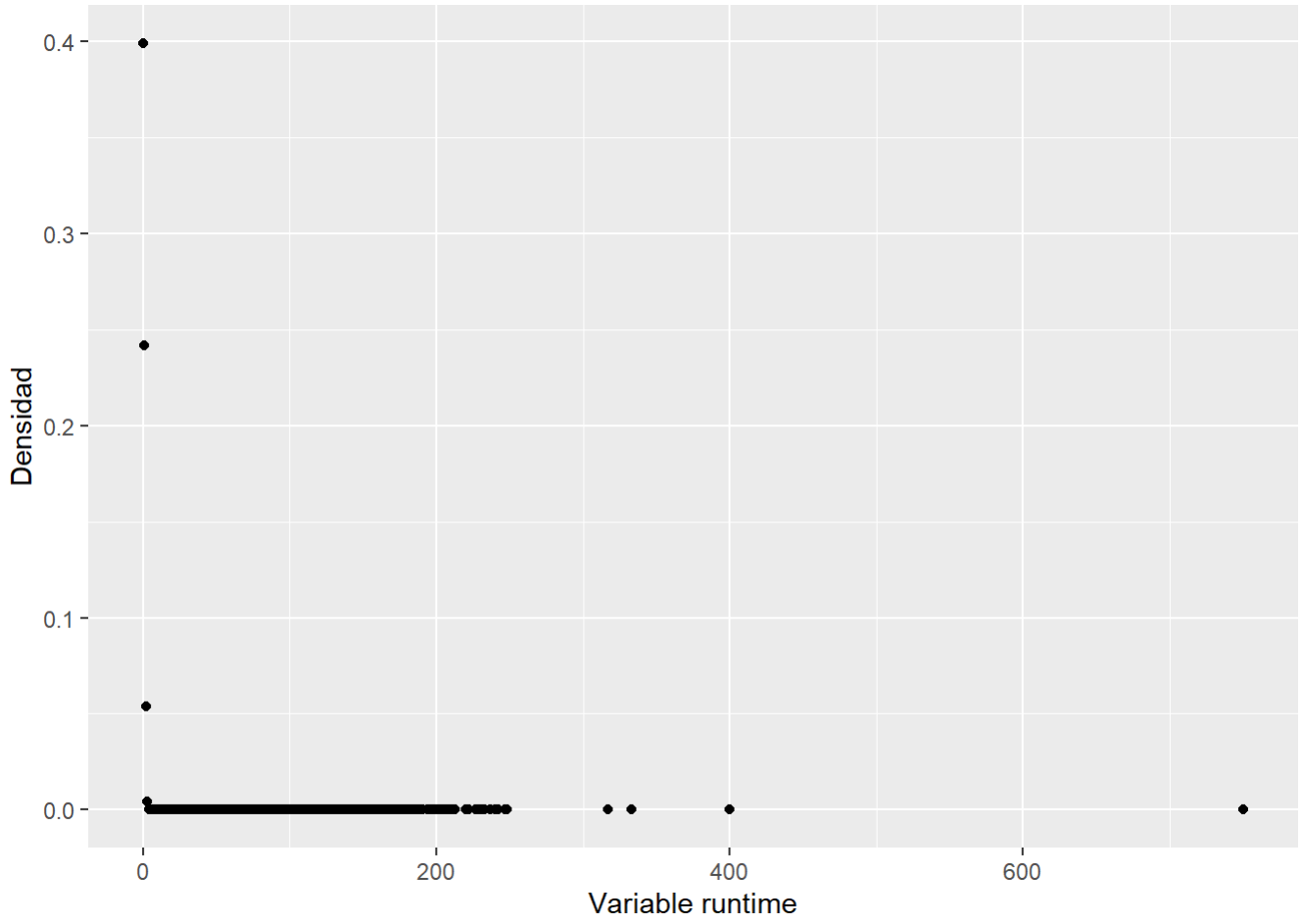
```
#Distribucion para la variable ID
j
```



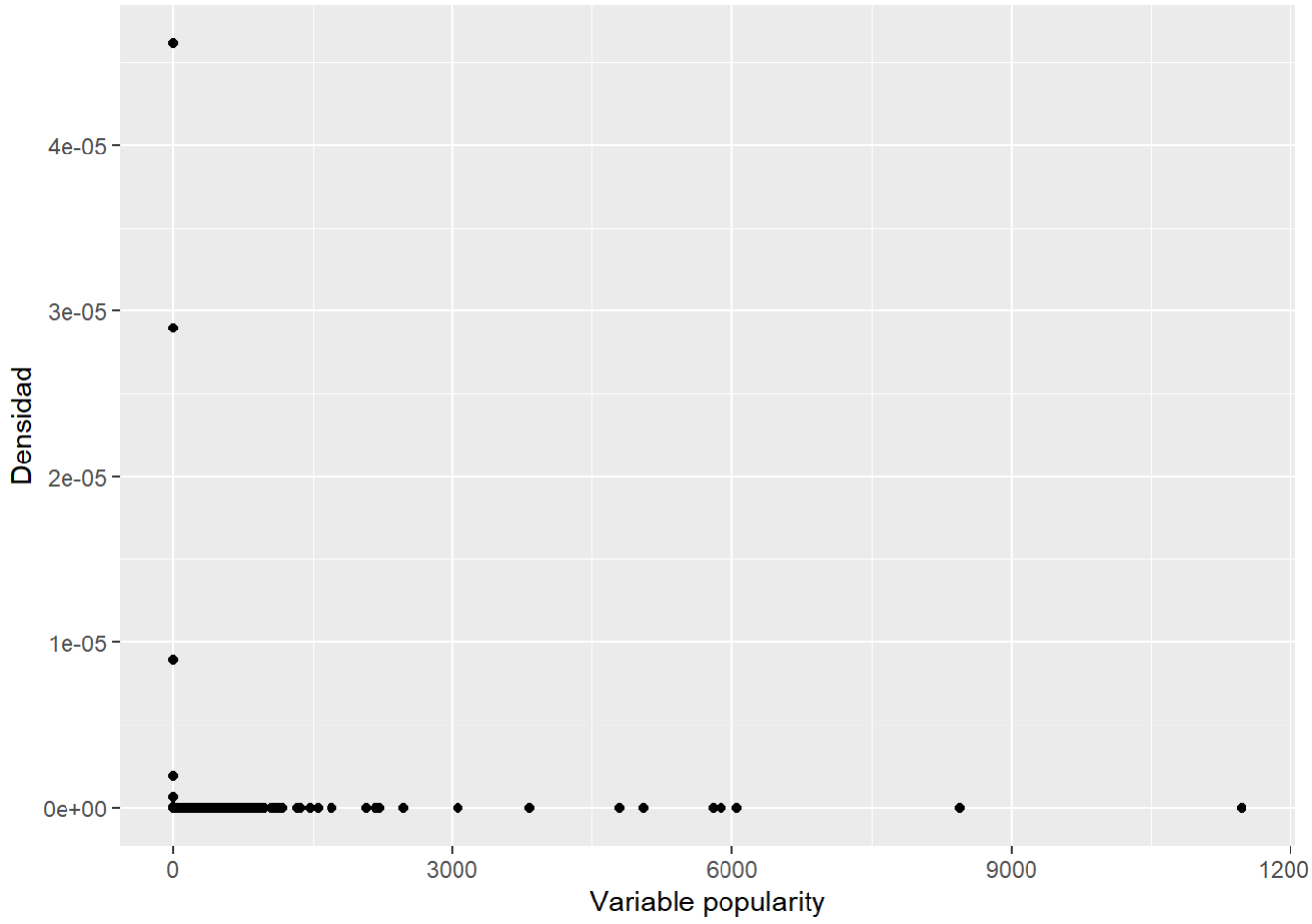

```
#Distribucion para la variable BUDGET
a
```



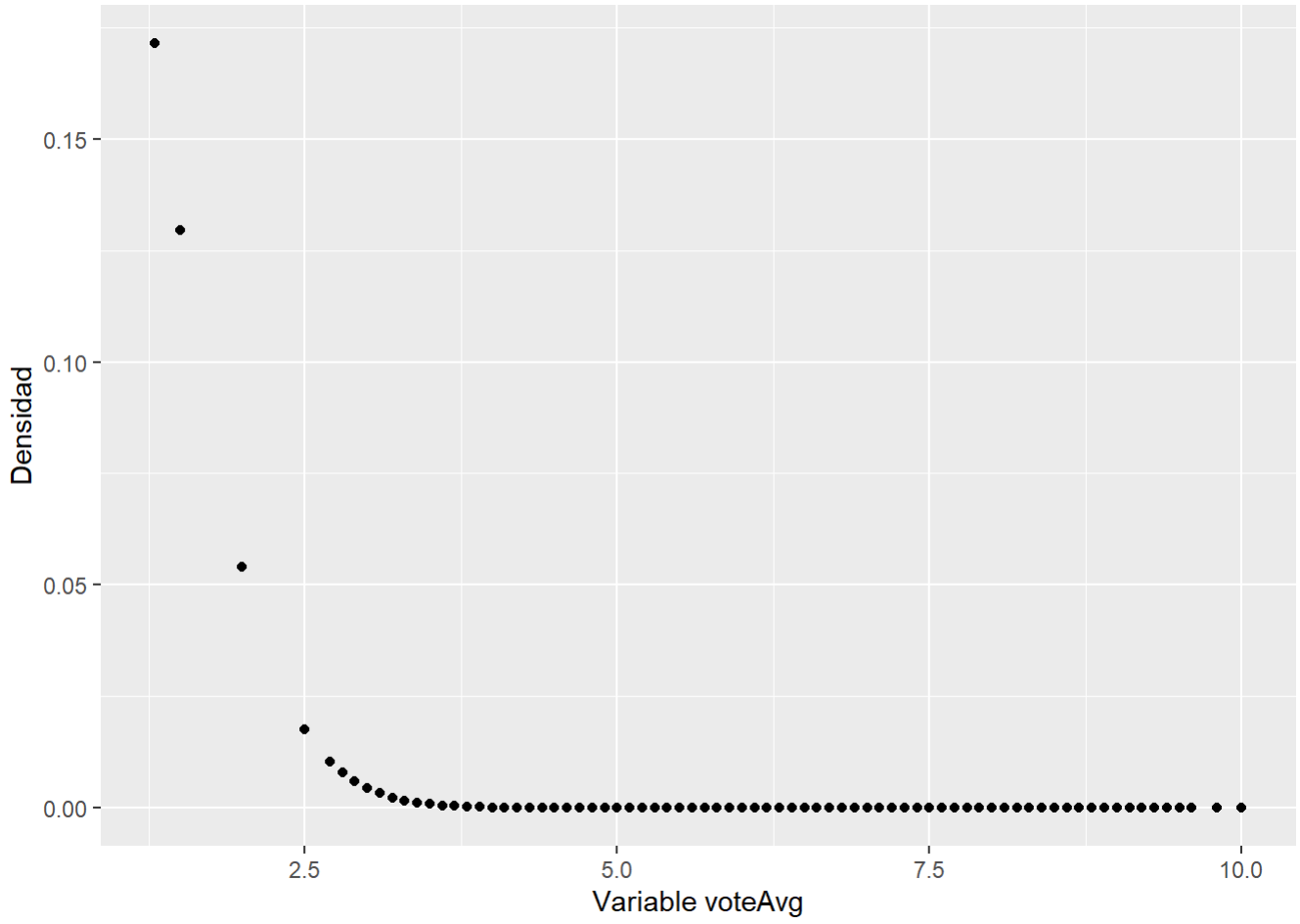

```
#Distribucion para la variable RUNTIME
b
```



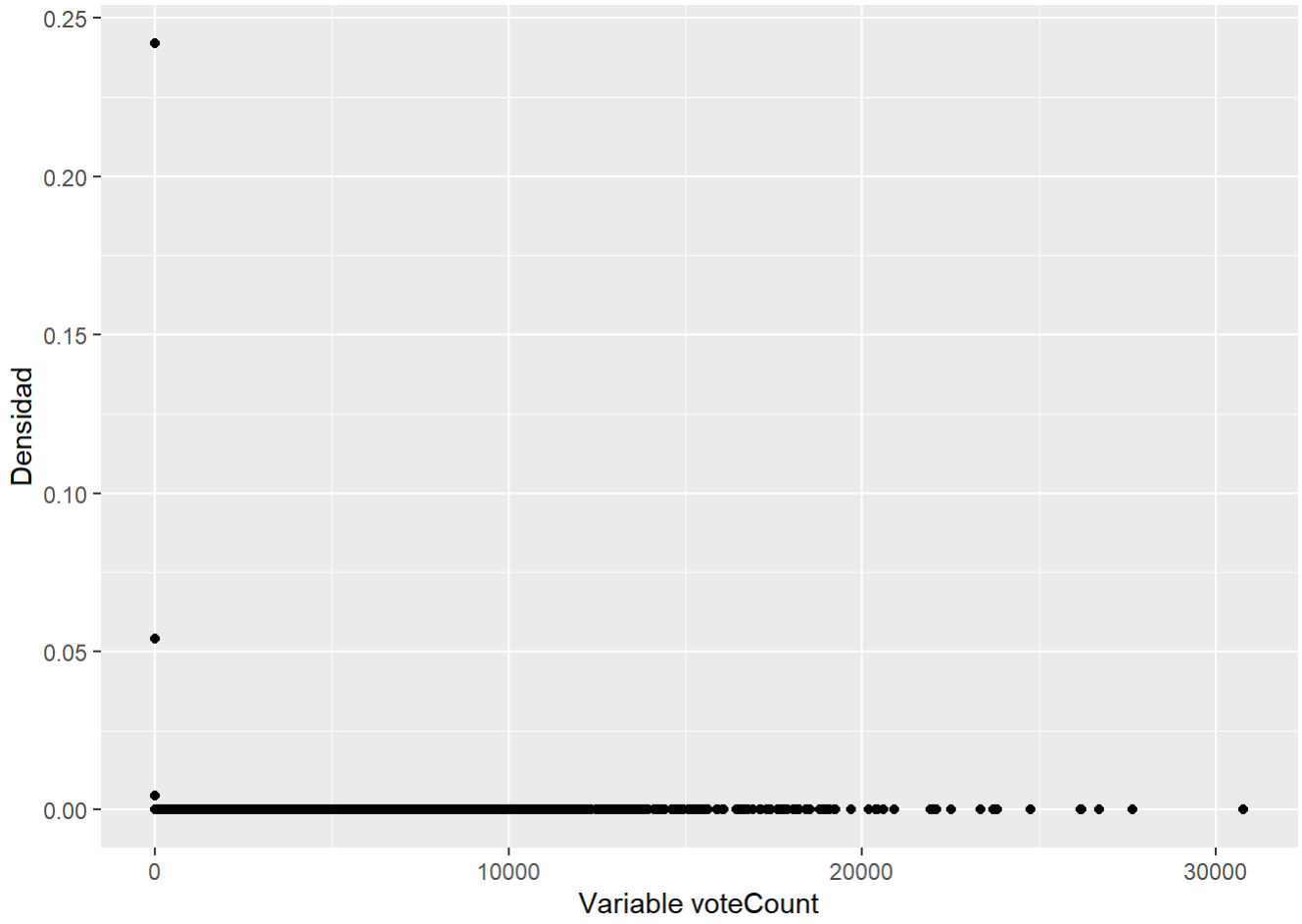

```
#Distribucion para la variable POPULARITY
c
```



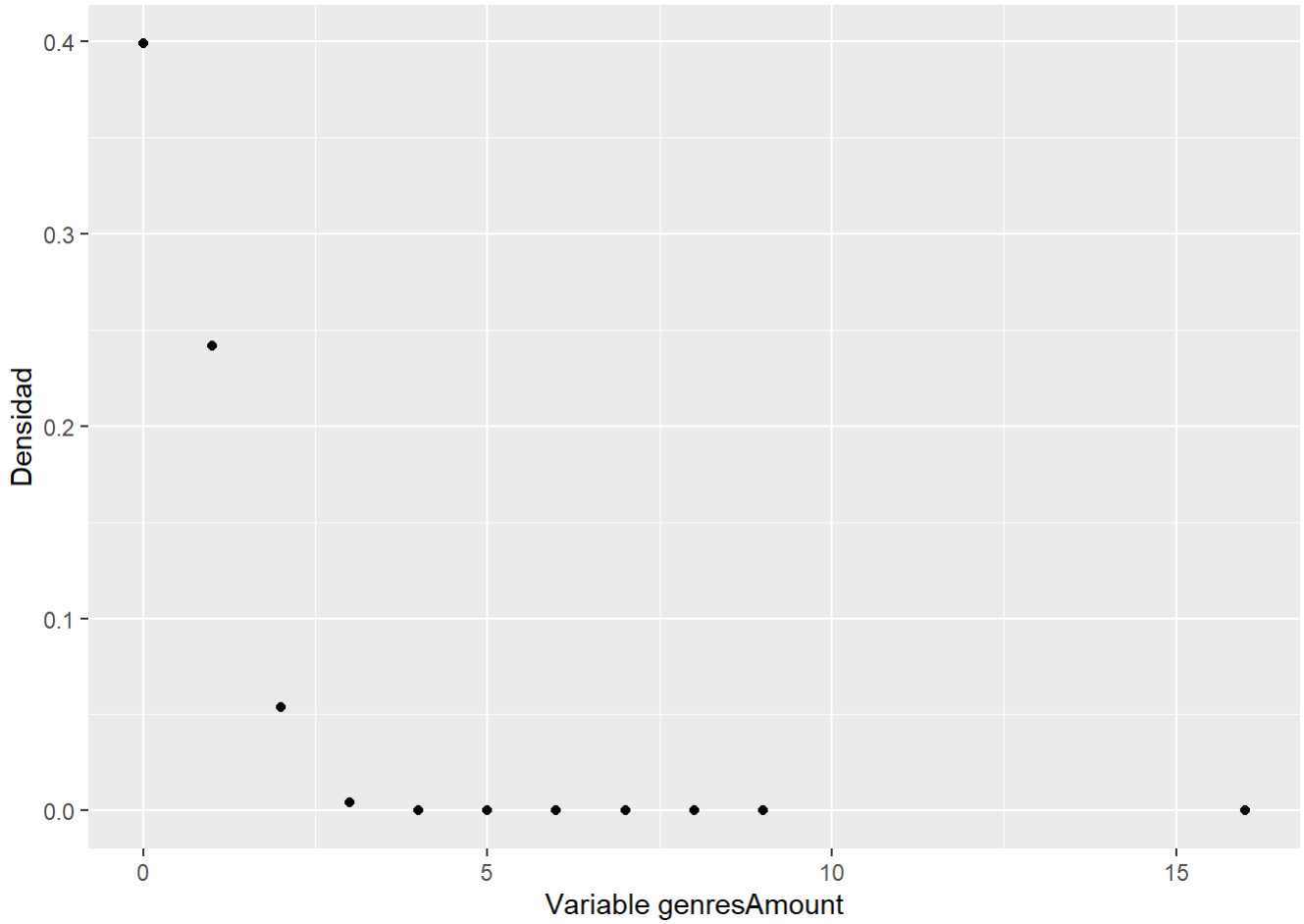

```
#Distribucion para la variable VOTEAVG
d
```



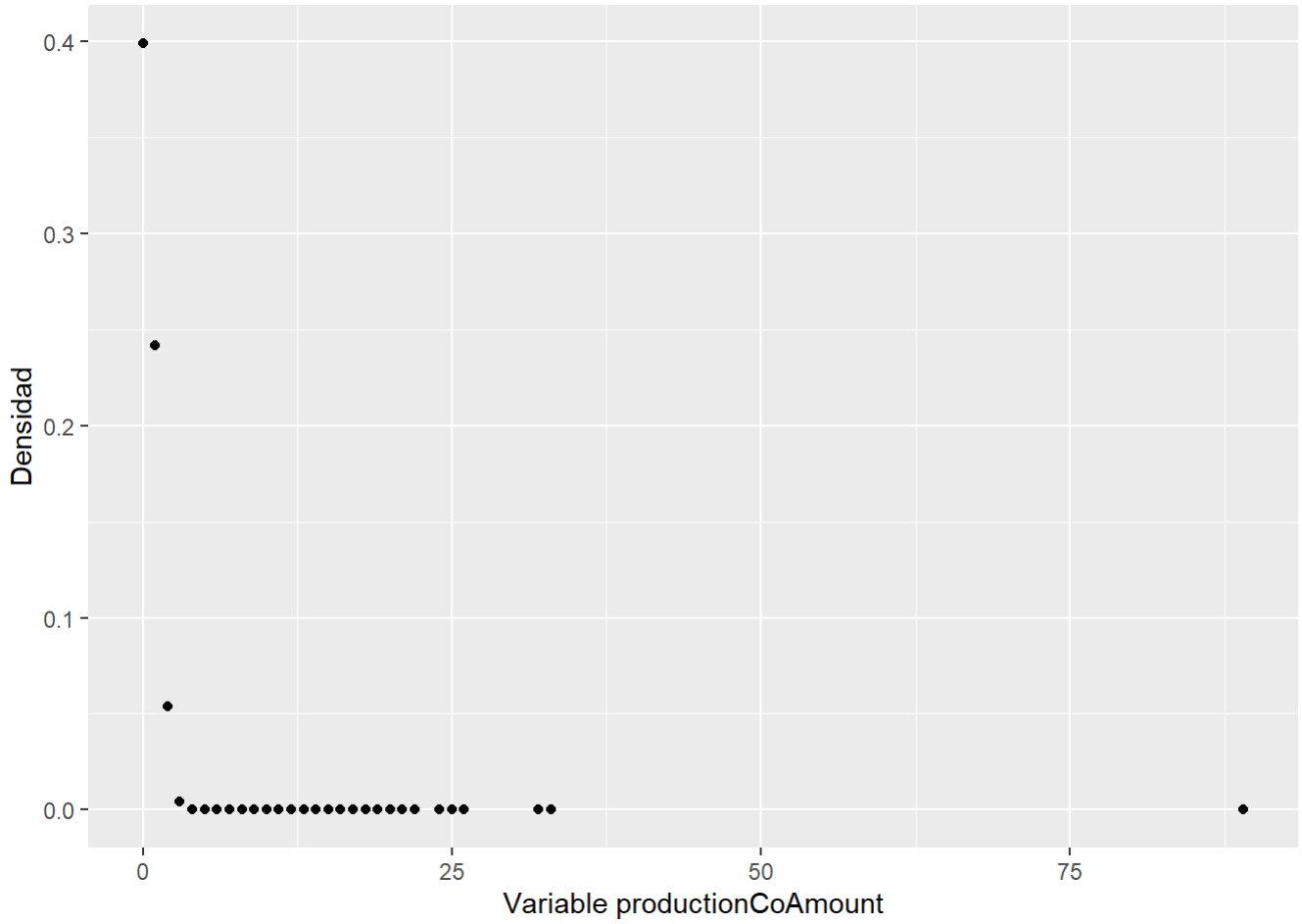

```
#Distribucion para la variable VOTECOUNT
e
```



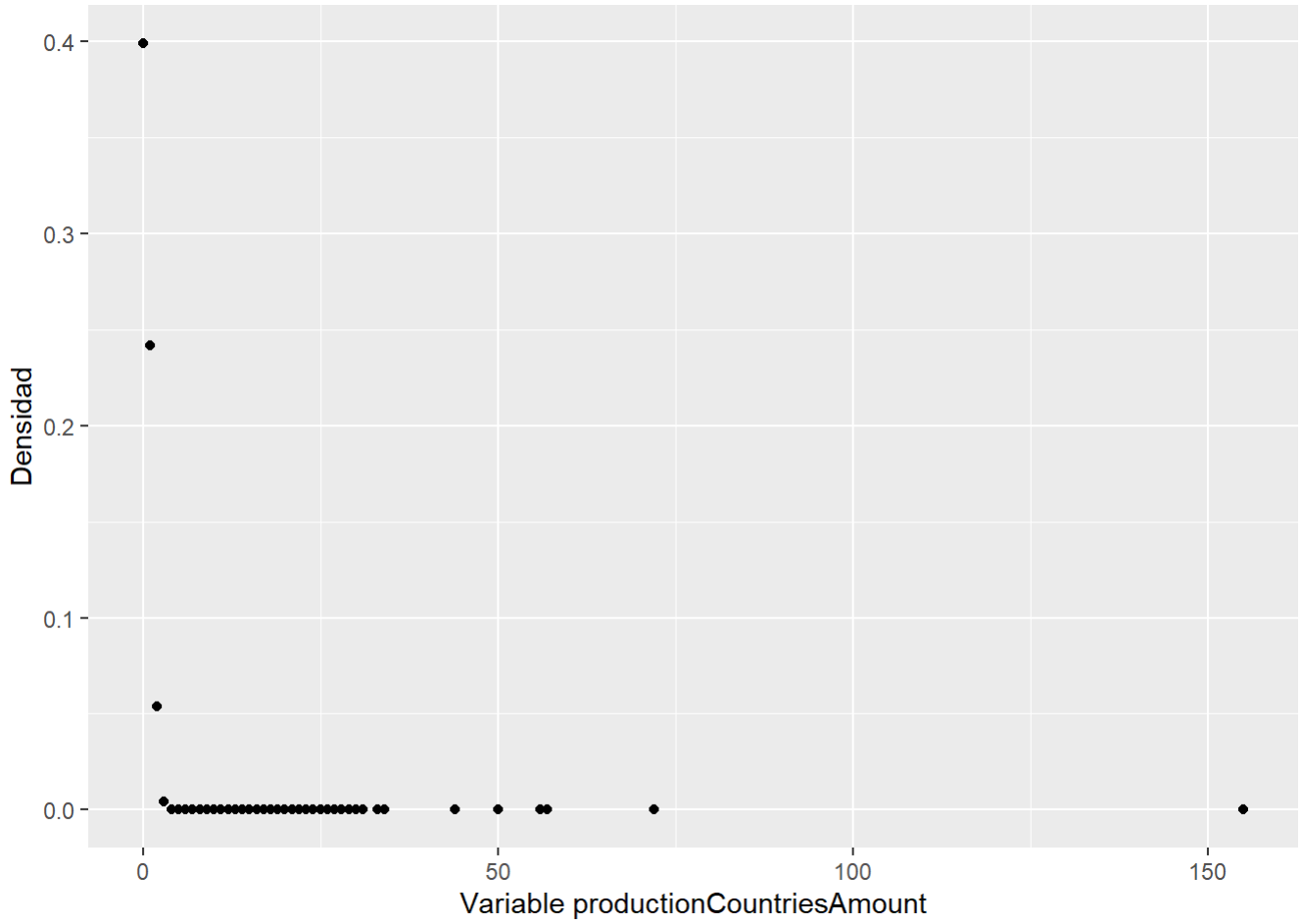

```
#Distribucion para la variable GENRESAMOUNT
f
```



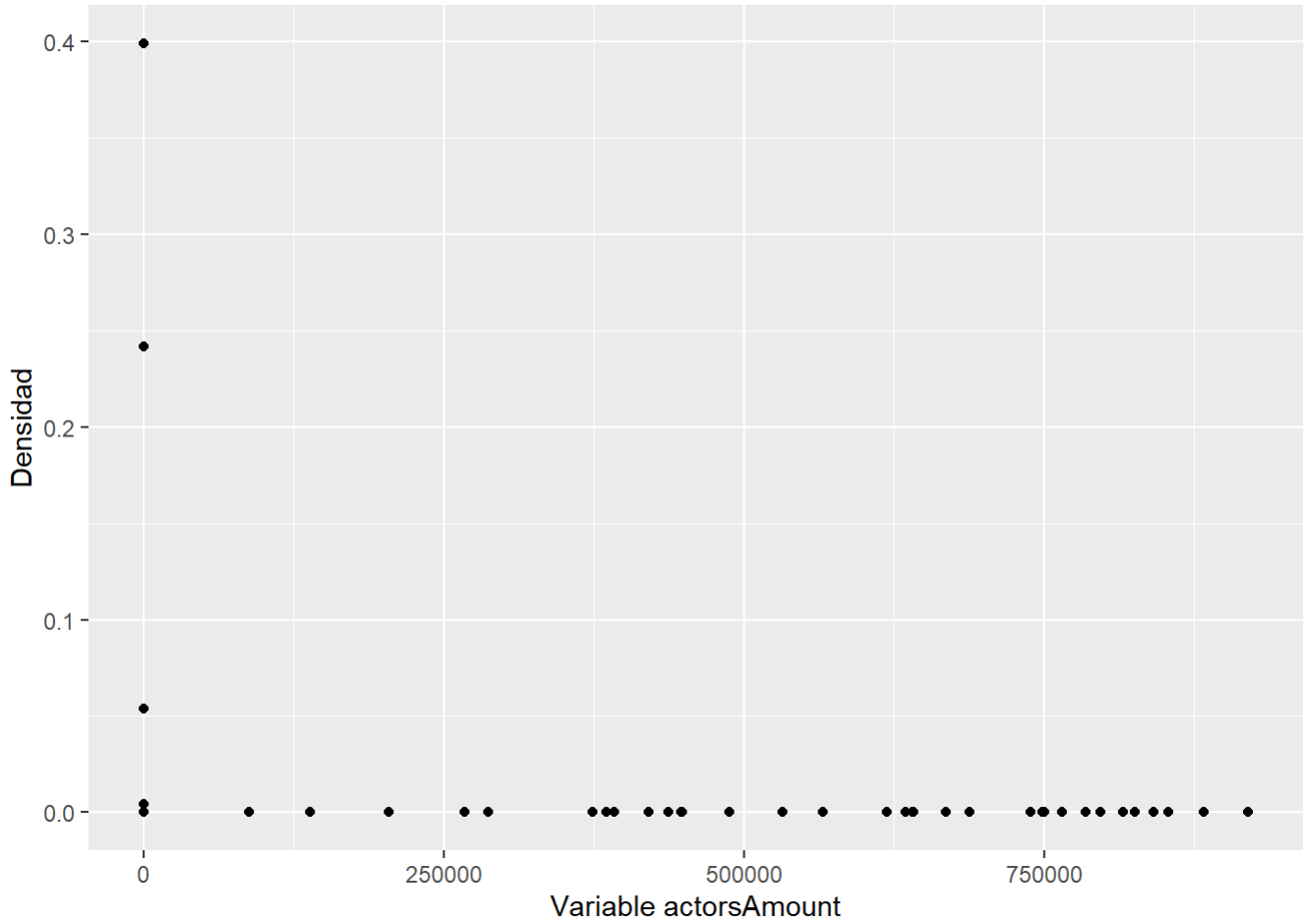

```
#Distribucion para la variable PRODUCTIONAMOUNT
g
```




```
#Distribucion para la variable PRODUCTION_COUNTRY_AMOUNT
h
```




```
#Distribucion para la variable ACTORS_AMOUNT
i
```

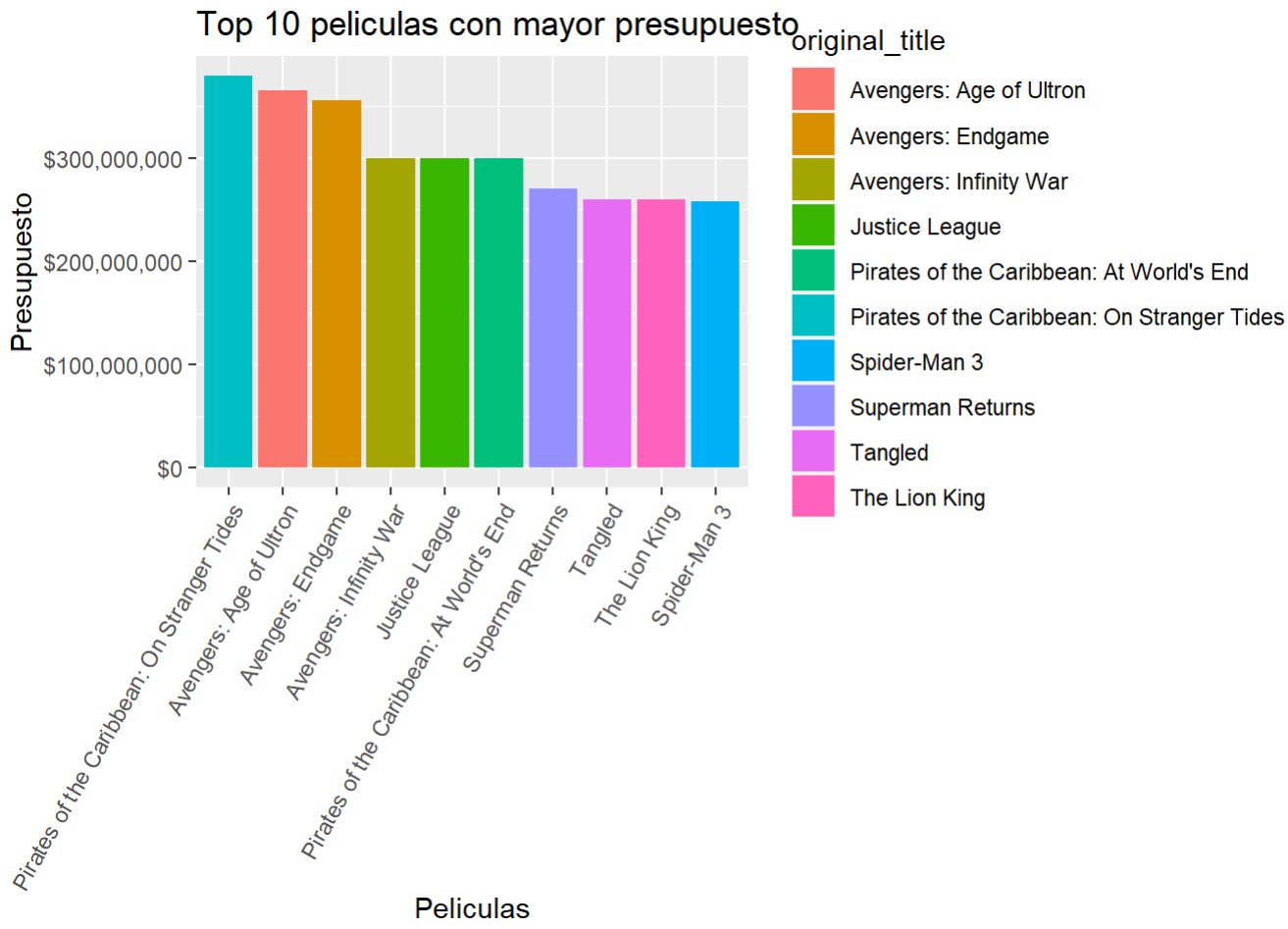


###

4. Responda las siguientes preguntas

4.1. ¿Cuáles son las 10 películas que contaron con más presupuesto?

```
pregunta1
```

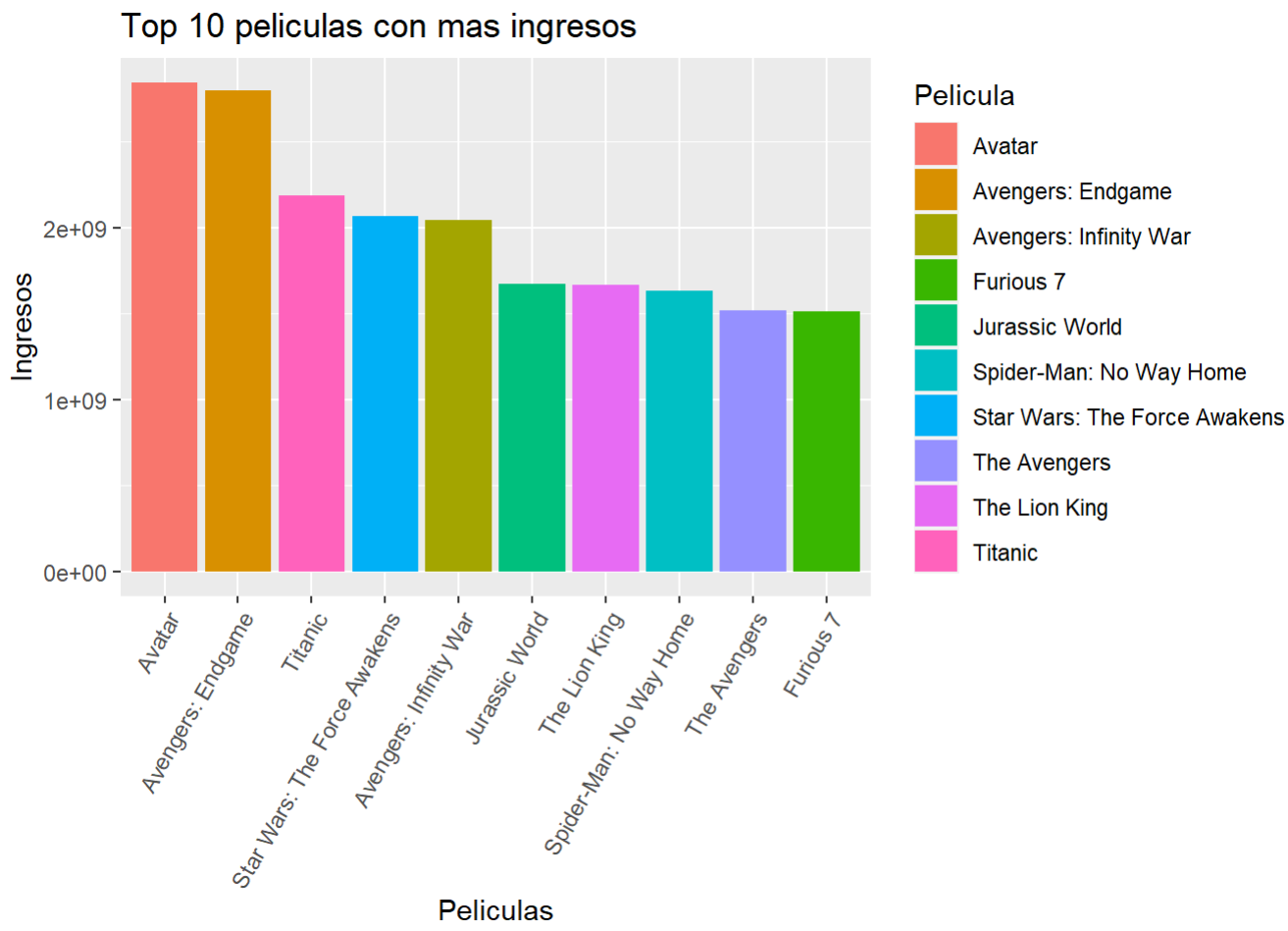


Como se observa en la grafica anterior las 10 peliculas con mayor presupuesto son:

1. **Pirates of the Caribbean: On Stranger Tides** con un presupuesto de **380000000** dolares.
2. **Avengers: Age of Ultron** con un presupuesto de **365000000** dolares.
3. **Avengers: Endgame** con un presupuesto de **356000000** dolares.
4. **Pirates of the Caribbean: At World's End** con un presupuesto de **300000000** dolares.
5. **Justice League** con un presupuesto de **300000000** dolares.
6. **Avengers: Infinity War** con un presupuesto de **300000000** dolares.
7. **Superman Returns** con un presupuesto de **270000000** dolares.
8. **Tangled** con un presupuesto de **260000000** dolares.
9. **The Lion King** con un presupuesto de **260000000** dolares.
10. **Spider-Man 3** con un presupuesto de **258000000** dolares.

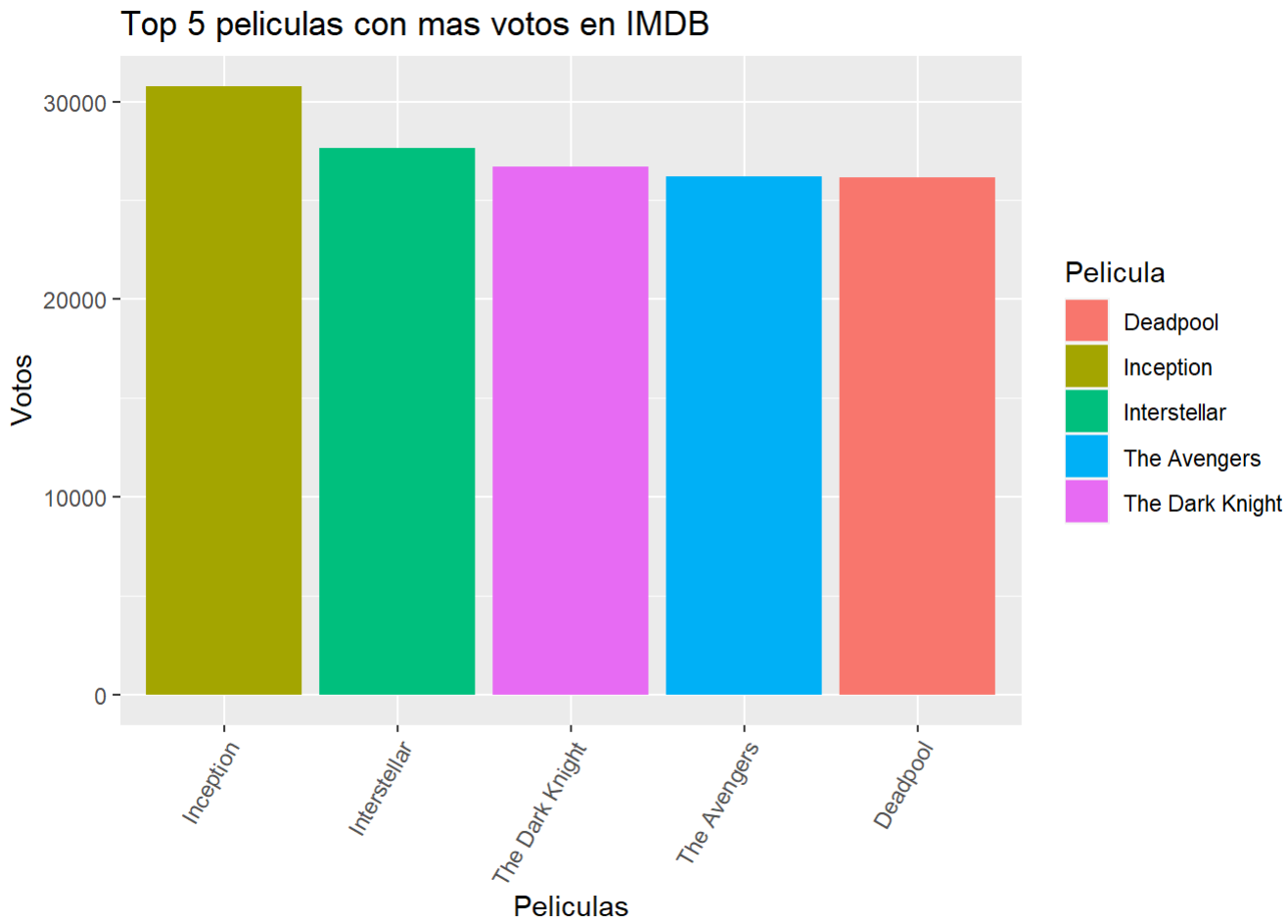
4.2. ¿Cuáles son las 10 películas que más ingresos tuvieron?

```
pregunta4.2
```



4.3. ¿Cuál es la película que más votos tuvo?

```
pregunta4.3
```

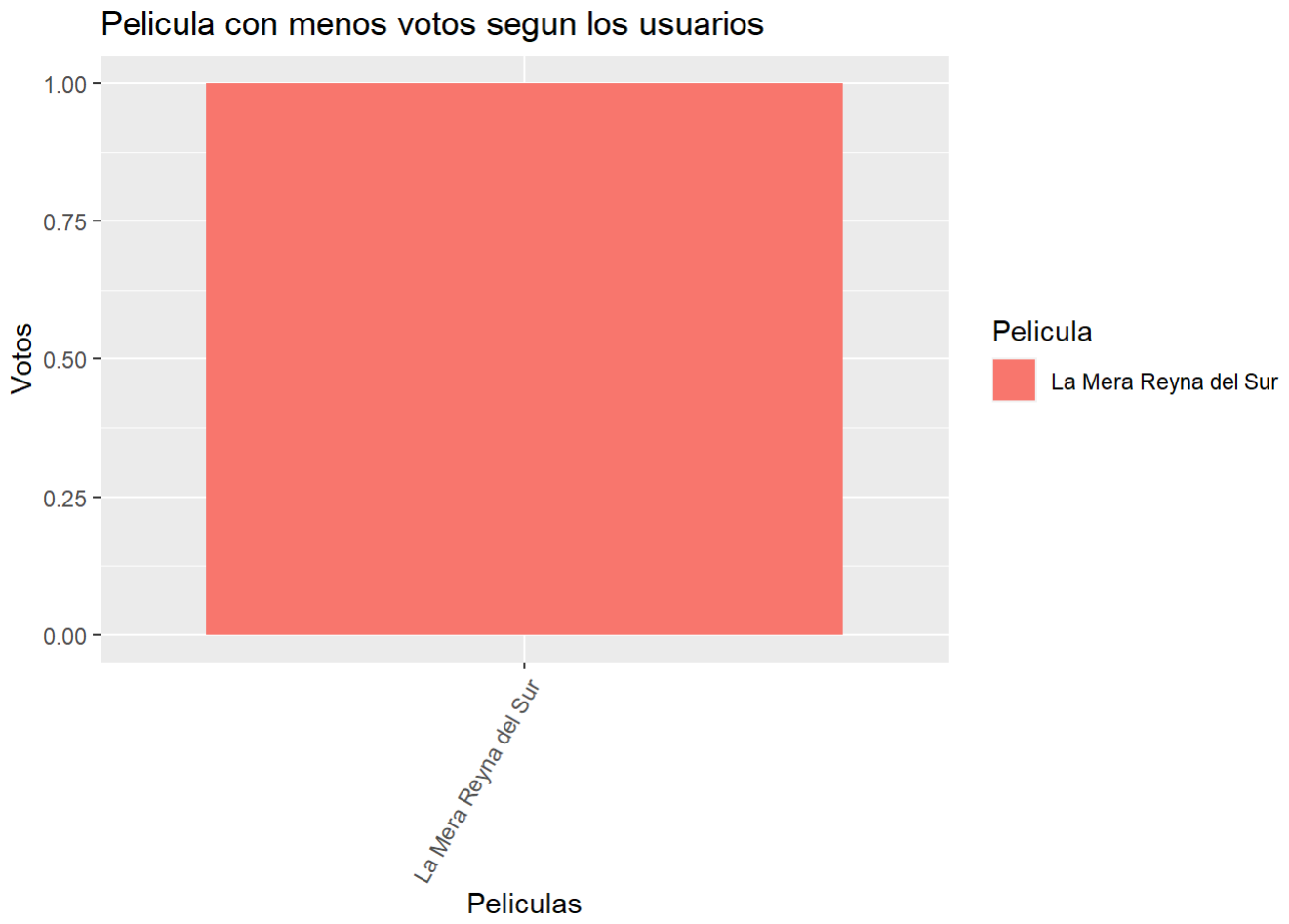


Como se observa en la grafica anterior las 5 peliculas con mas votos en IMDB son:

1. **Inception** con **30788** de votos.
2. **Interstellar** con **27644** de votos.
3. **The Dark Knight** con **26690** de votos.
4. **The Avengers** con **26215** de votos.
5. **Deadpool** con **26178** de votos.

4.4. ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios?

```
pregunta4.4
```

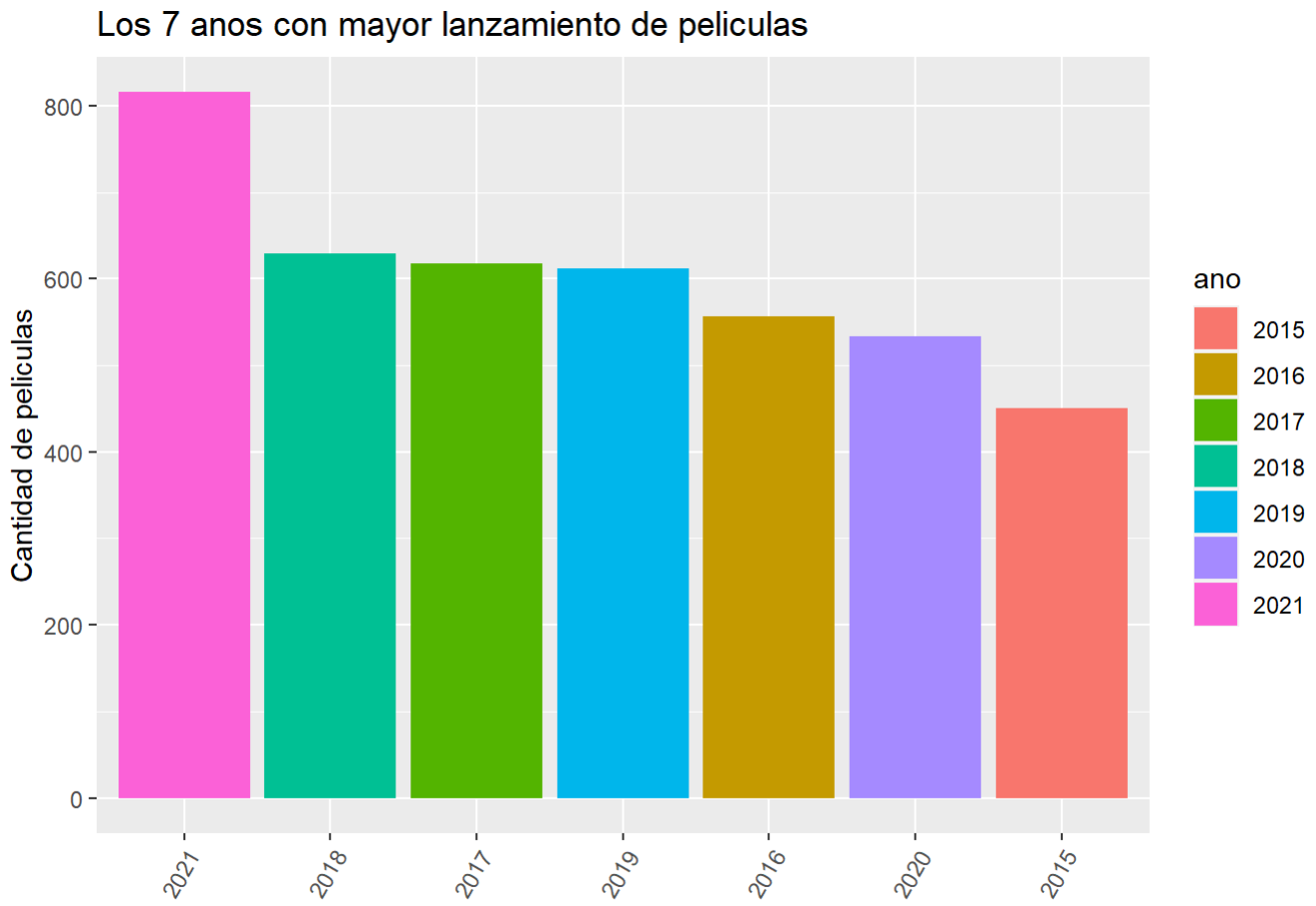


Como se observa en la grafica anterior, la pelicula que presenta menor cantidad de votos es:

1. **La Mera Reyna del Sur** con **1** voto.

4.5. ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras

```
pregunta4_5
```

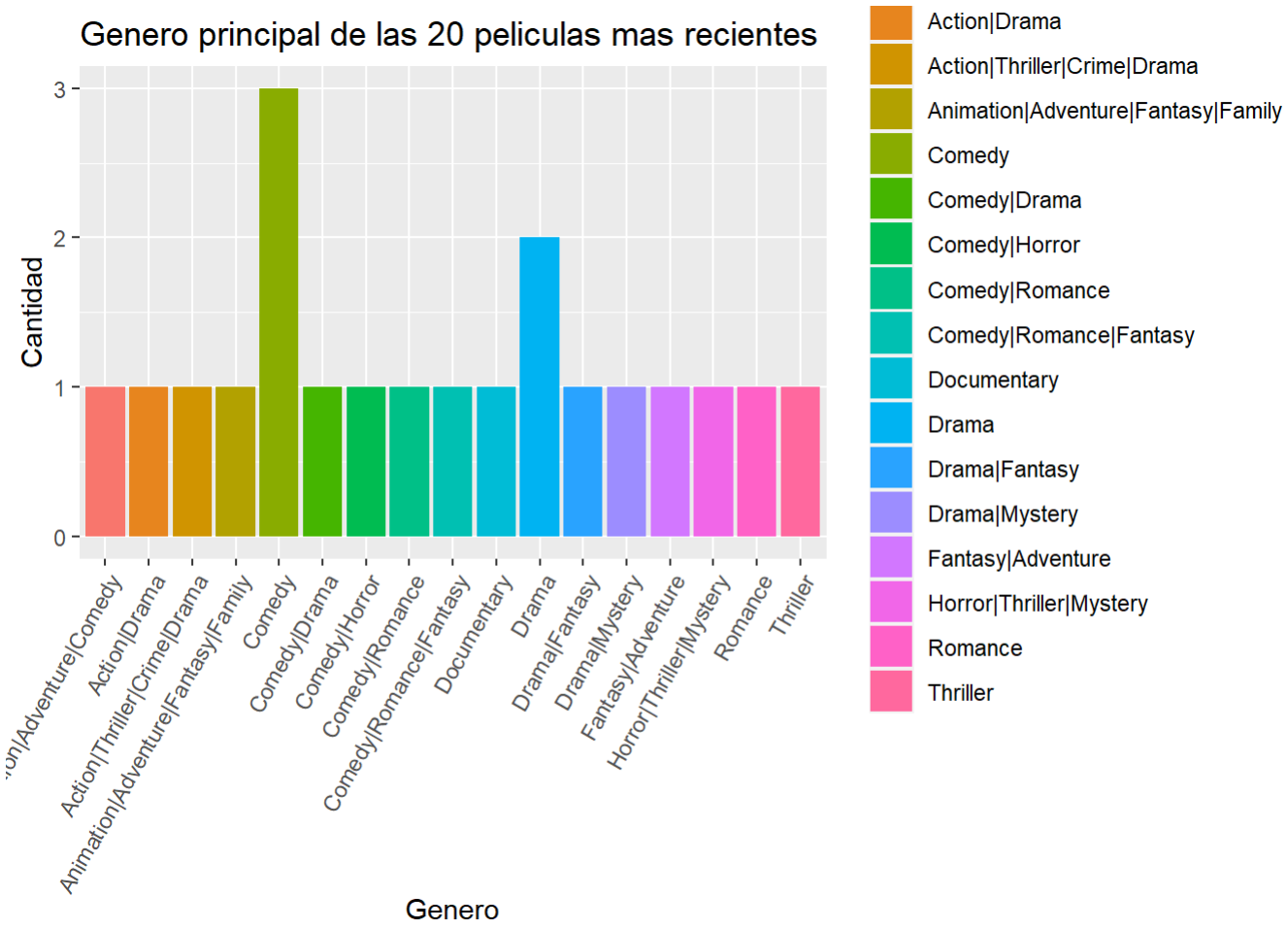


Como se observa en la grafica anterior los años con mayor cantidad de peliculas lanzadas son:

1. **2021** ano con una cantidad de peliculas lanzadas de **816**
2. **2018** ano con una cantidad de peliculas lanzadas de **629**
3. **2017** ano con una cantidad de peliculas lanzadas de **618**
4. **2019** ano con una cantidad de peliculas lanzadas de **612**
5. **2016** ano con una cantidad de peliculas lanzadas de **557**
6. **2020** ano con una cantidad de peliculas lanzadas de **533**
7. **2015** ano con una cantidad de peliculas lanzadas de **450**

4.6. ¿Cuál es el género principal de las 20 películas más recientes? ¿Cuál es el género principal que predomina en el conjunto de datos? Represéntelo usando un gráfico

```
pregunta4.6
```

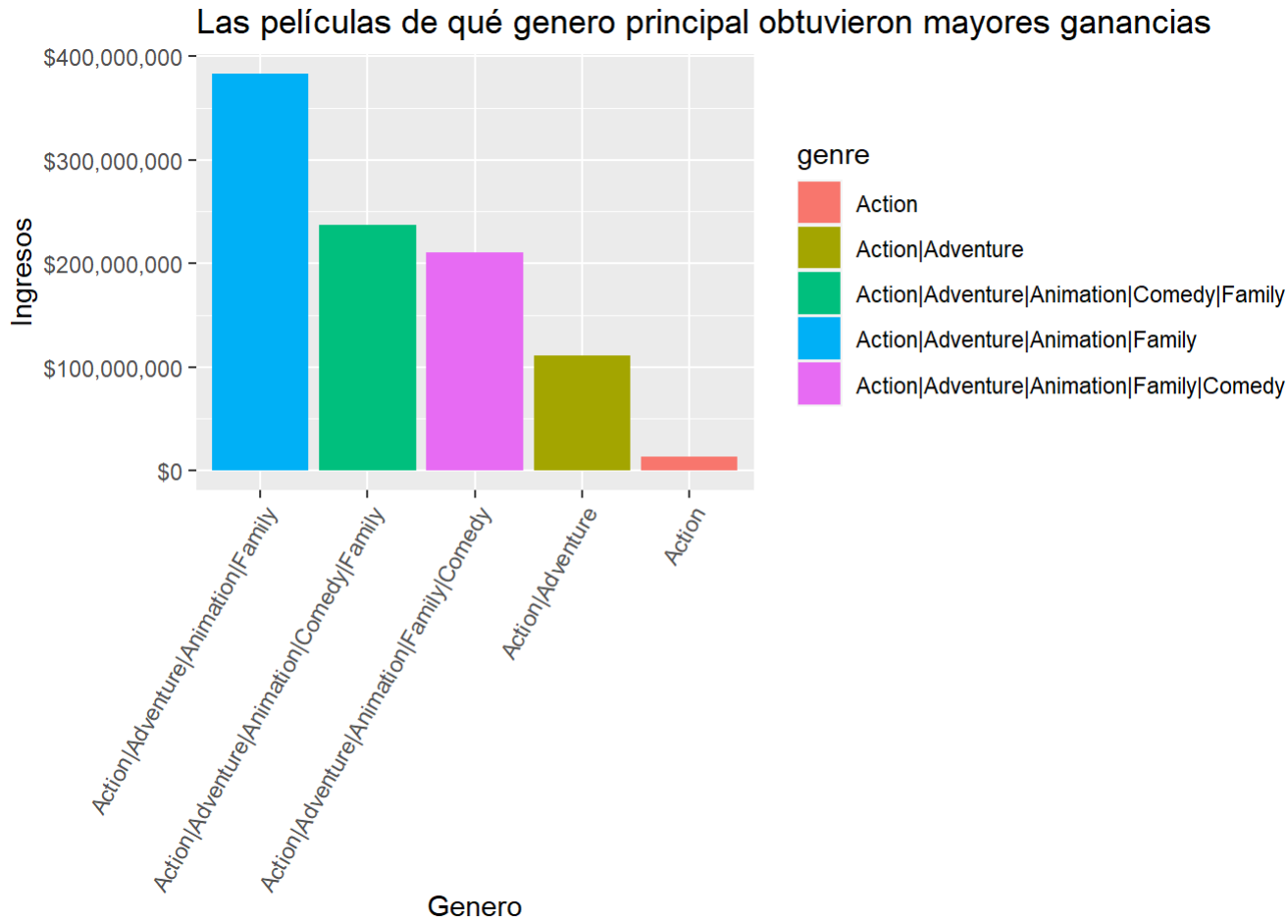


Como se observa en la grafica anterior los generos principales de las 20 peliculas mas recientes son:

1. **Comedy** repitiendose **3** veces.
2. **Drama** repitiendose **2** veces.
3. **El resto de generos** repitiendose **1** veces.

4.7 ¿Las películas de qué genero principal obtuvieron mayores ganancias?

```
pregunta4.7
```

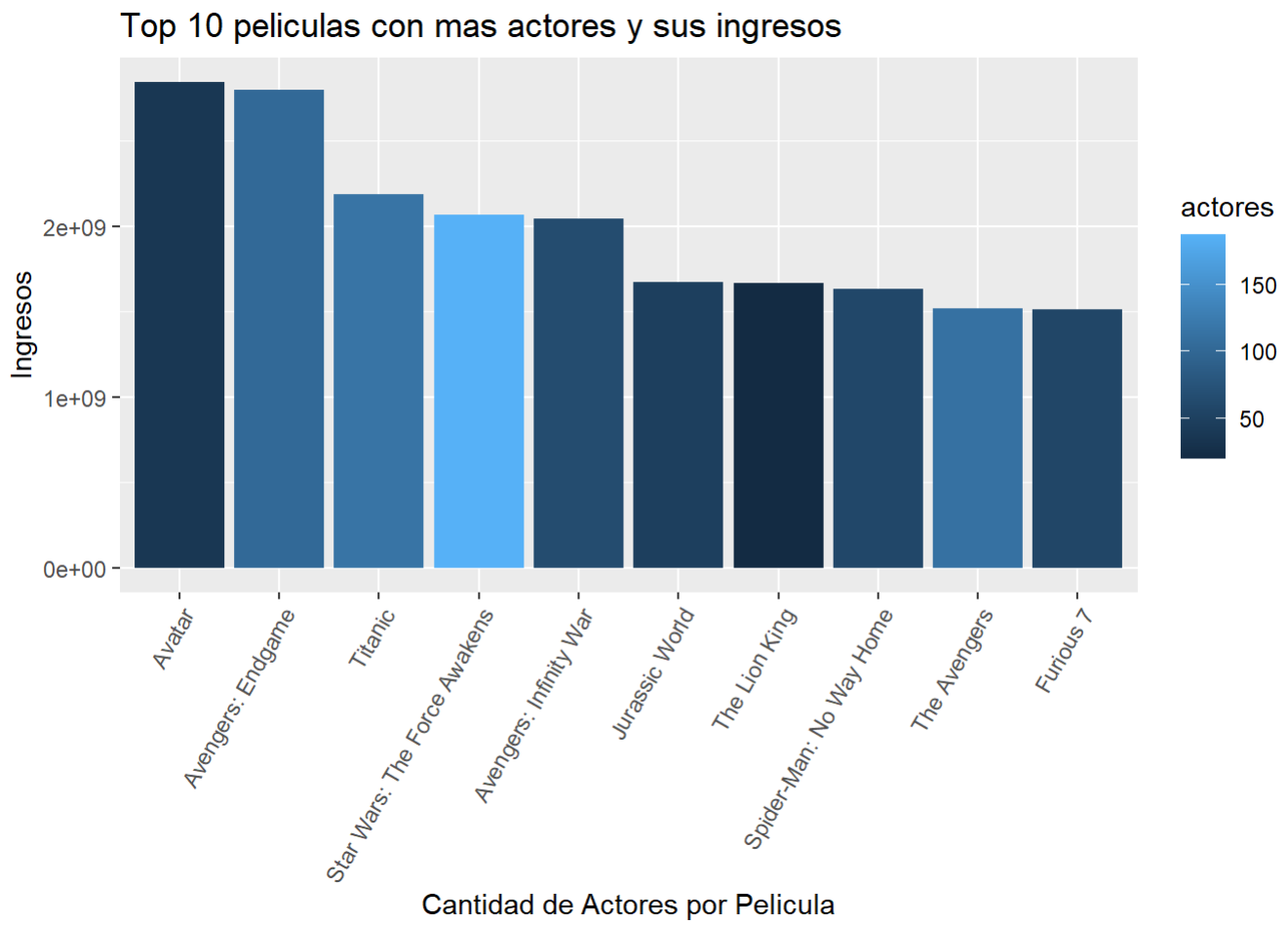


Como se logra apreciar de manera formal en la grafica, los 5 generos que mas ganancias obtuvieron fueron:

1. **Action|Adventure|Animation|Family|Comedy** con **2.1058152^{8}**
2. **Action|Adventure|Animation|Family** con **3.8307376^{8}**
3. **Action|Adventure|Animation|Comedy|Family** con **2.372284^{8}**
4. **Action|Adventure** con **1.1153593^{8}**
5. **Action** con **1.3928637^{7}**

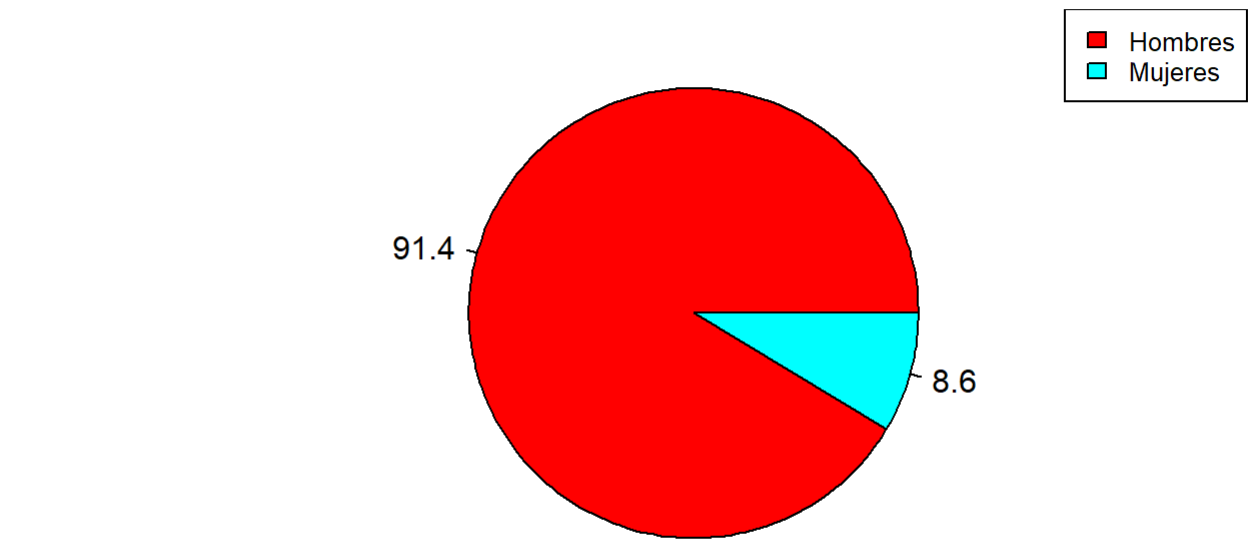
4.8 ¿La cantidad de actores influye en los ingresos de las películas?¿se han hecho películas con más actores en los últimos años?

```
pregunta4.8
```



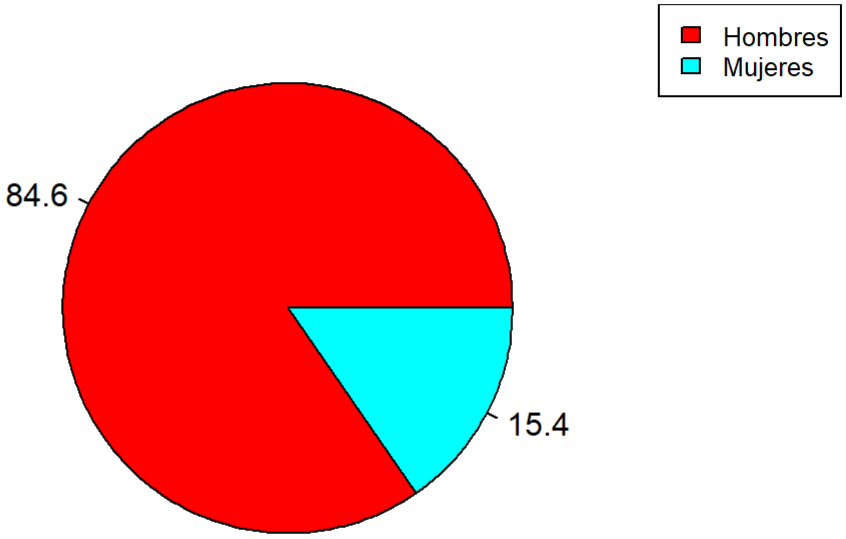
4.9. ¿Es posible que la cantidad de hombres y mujeres en el reparto influya en la popularidad y los ingresos de las películas?

```
pie(x, labels=piepercent, main = "Porcentaje de ganancias de las peliculas cuando hay un sexo pr
edominante", col = rainbow(length(x)))
                legend("topright", c("Hombres","Mujeres"), cex = 0.8,
                       fill = rainbow(length(x)))
```



Como se observa en la grafica, del 100% de los ingresos de las peliculas en la base de datos, se demuestra que el 91.4% de los ingresos se debe cuando hay mas actores que actrices. Evidenciando que si influye que hayan mas actores que actrices.

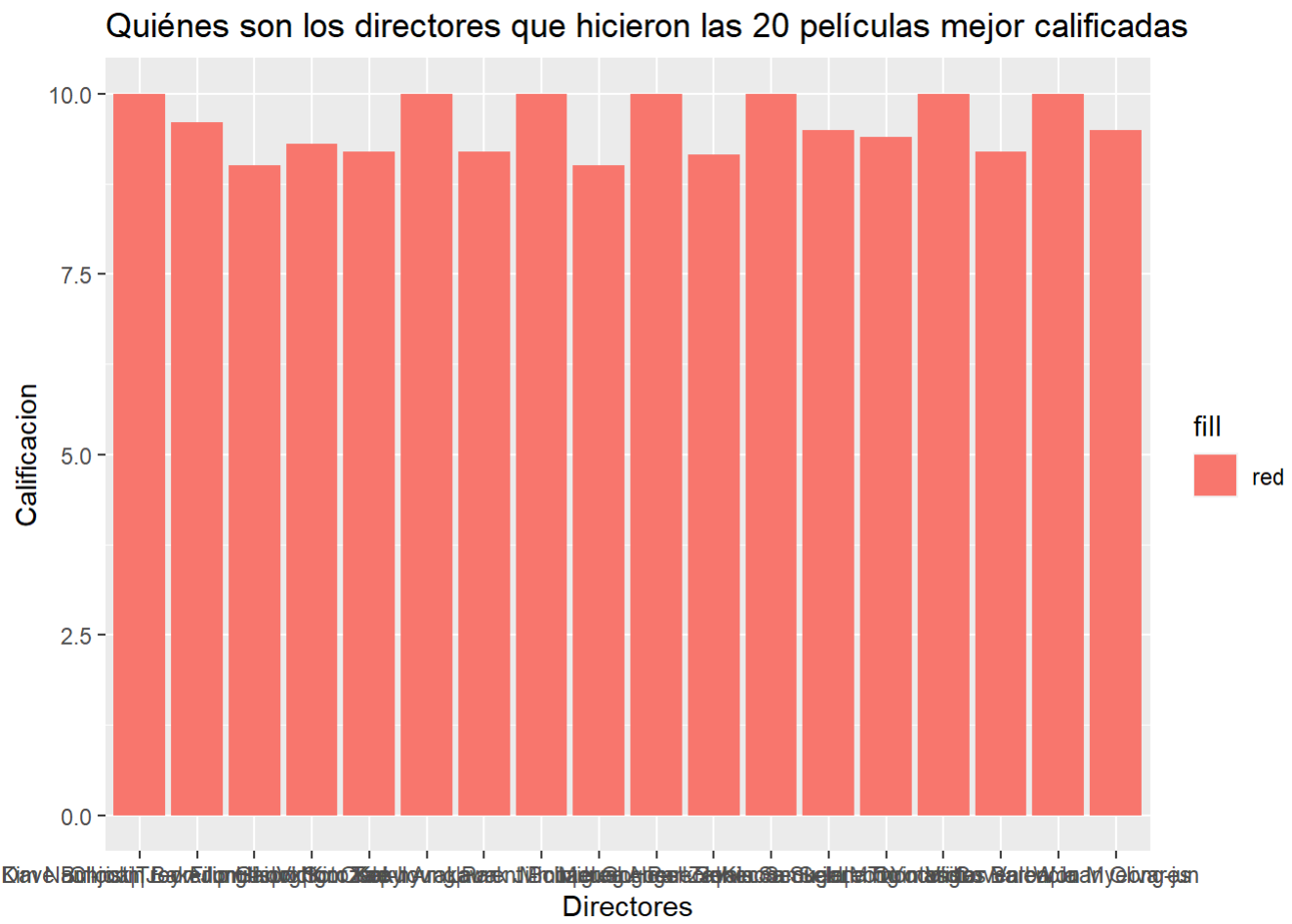
```
pie(xPopu, labels=piepercentPopu, main = "Porcentaje de popularidad de las peliculas cuando hay
 un sexo predominante", col = rainbow(length(x)))
legend("topright", c("Hombres","Mujeres"), cex = 0.8,
       fill = rainbow(length(xPopu)))
```



Y como se ve en la grafica de pie, se observa que las peliculas con mas hombres que mujeres obtienen mayor popularidad. Sin embargo, estos porcentajes son menores a diferencia de los porcentajes de ingresos.

4.10. ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

```
pregunta4.10
```



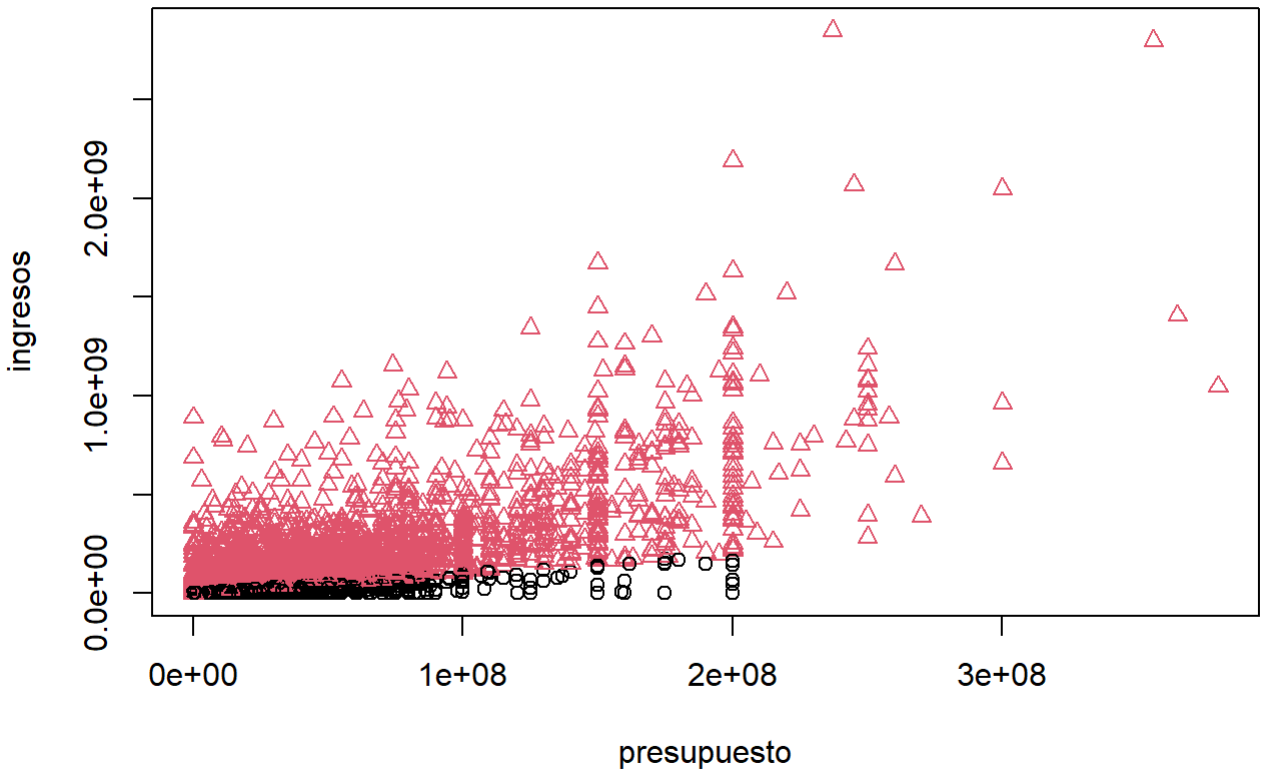
Como se puede ver en la grafica, los directores con mejor calificaciones son:

1. **Hot Naked Sex & the City** dirigida por **Thomas Coven** obteniendo una puntuacion de **10**
2. **Vacaciones** dirigida por **Víctor Barba|Juan Olivares** obteniendo una puntuacion de **10**
3. **Steven Universe: The Movie: Behind the Curtain** dirigida por **Rebecca Sugar** obteniendo una puntuacion de **10**
4. **Spirit of Vengeance: The Making of 'Ghost Rider'** dirigida por **Laurent Bouzereau** obteniendo una puntuacion de **10**
5. **<U+30DD><U+30CB><U+30E7><U+306F><U+3053><U+3046><U+3057><U+3066><U+751F><U+307E><U+308C><U+305F> <U+301C> <U+5BAE><U+FA11><U+99FF><U+306E><U+601D><U+8003><U+904E><U+7A0B> <U+301C>** dirigida por **Kaku Arakawa** obteniendo una puntuacion de **10**
6. **Christmas at the Ranch** dirigida por **Christin Baker** obteniendo una puntuacion de **10**
7. **Los Vengadores Chiflados** dirigida por **Miguel Angel Zavala** obteniendo una puntuacion de **10**
8. **The Spectacular Spider-Man Attack of the Lizard** dirigida por **Dave Bullock|Troy Adomitis|Victor Cook** obteniendo una puntuacion de **9.6**
9. **Ebola Zombies** dirigida por **Samuel Leong** obteniendo una puntuacion de **9.5**
10. **<U+C774><U+BAA8><U+C758> <U+C720><U+D639> 3** dirigida por **Won Myeong-jun** obteniendo una puntuacion de **9.5**
11. **Selena - Live: The Last Concert** dirigida por **Selena Quintanilla** obteniendo una puntuacion de **9.4**
12. **<U+9B3C><U+6EC5><U+306E><U+5203> <U+67F1><U+5408><U+4F1A><U+8B70>·<U+8776><U+5C4B><U+6577><U+7DE8>** dirigida por **Haruo Sotozaki** obteniendo una puntuacion de **9.3**
13. **<U+9B3C><U+6EC5><U+306E><U+5203> <U+5144><U+59B9><U+306E><U+7D46>** dirigida por **Haruo Sotozaki** obteniendo una puntuacion de **9.3**
14. **Franco Escamilla: Por La Anécdota** dirigida por **Ulises Valencia** obteniendo una puntuacion de **9.2**

15. **BTS World Tour: Love Yourself - Japan Edition** dirigida por **Kim Nam-joon|Jeon Jung-kook|Kim Tae-hyung|Park Ji-min|Jung Ho-seok|Kim Seok-jin|Min Yoon-gi** obteniendo una puntuacion de **9.2**
16. **<U+BE0C><U+B808><U+C774><U+D06C> <U+B354> <U+C0AC><U+C77C><U+B7F0><U+C2A4>: <U+B354> <U+BB34><U+BE44>** dirigida por **Park Jun-soo** obteniendo una puntuacion de **9.2**
17. **<U+041D><U+0435><U+0431><U+043E>** dirigida por **Igor Kopylov** obteniendo una puntuacion de **9.2**
18. **<U+BE0C><U+B9C1> <U+B354> <U+C18C><U+C6B8>: <U+B354> <U+BB34><U+BE44>** dirigida por **Park Jun-soo** obteniendo una puntuacion de **9.1**
19. **Scooby-Doo! and the Spooky Scarecrow** dirigida por **Michael Goguen** obteniendo una puntuacion de **9**
20. **Three Preludes for Solo Piano By Adam Sherkin** dirigida por **Filip Ghiorghi** obteniendo una puntuacion de **9**

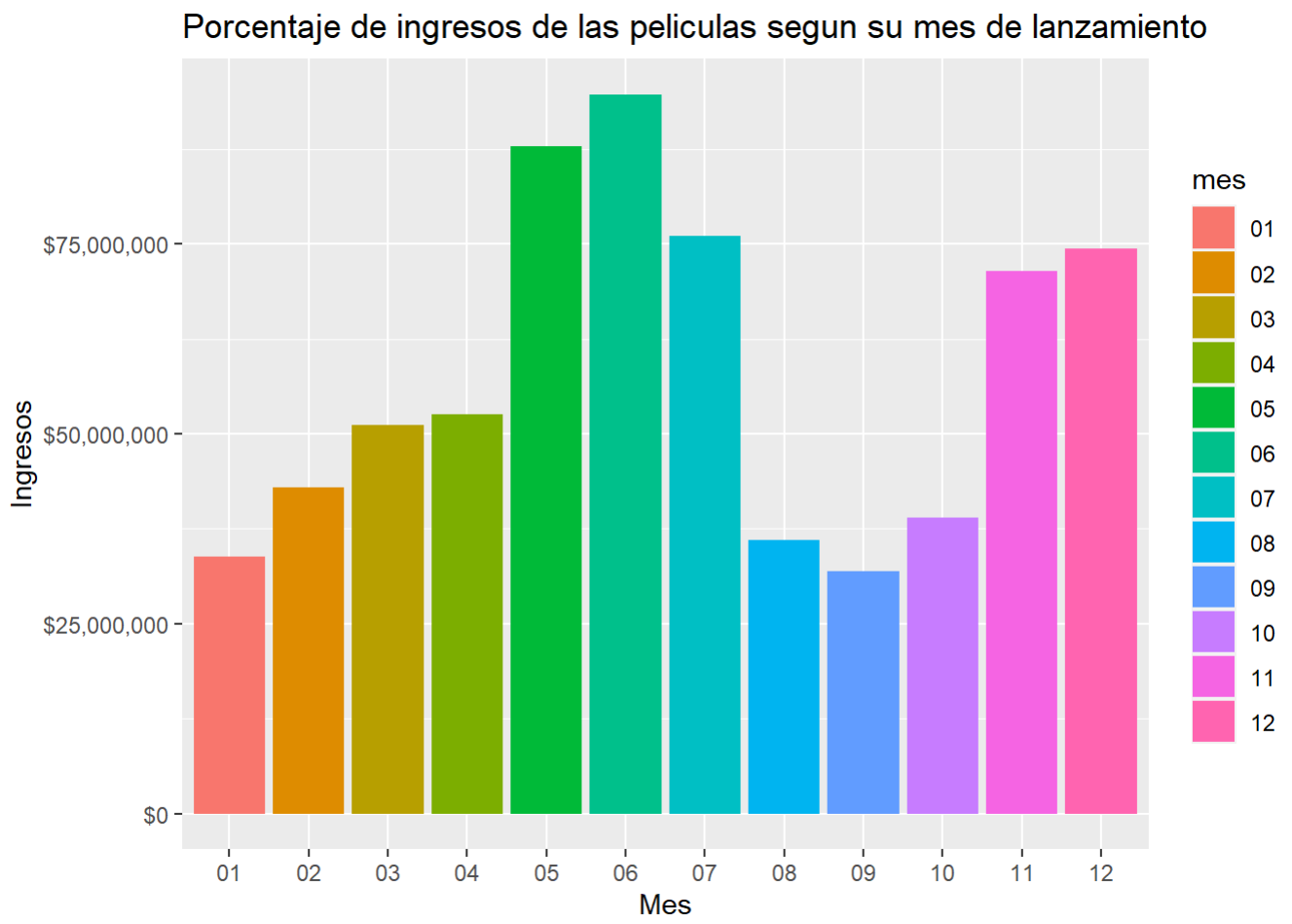
4.11. ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión

```
plot(x=presupuesto, y=ingresos, pch = as.numeric(grupo), col = grupo)
```



Como se observa los puntos rojos equivalen al ingreso de las peliculas, y los puntos negros equivalen al presupuesto de la pelicula, evidenciando que son pocos los casos en donde se llegan a generar perdidas, ya que todo aque, punto rojo que sobrepase el punto negro representan una perdida, caso contrario, entre mas lejos este el punto rojo del punto negro, representa una ganancia excesiva.
4.12. ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?

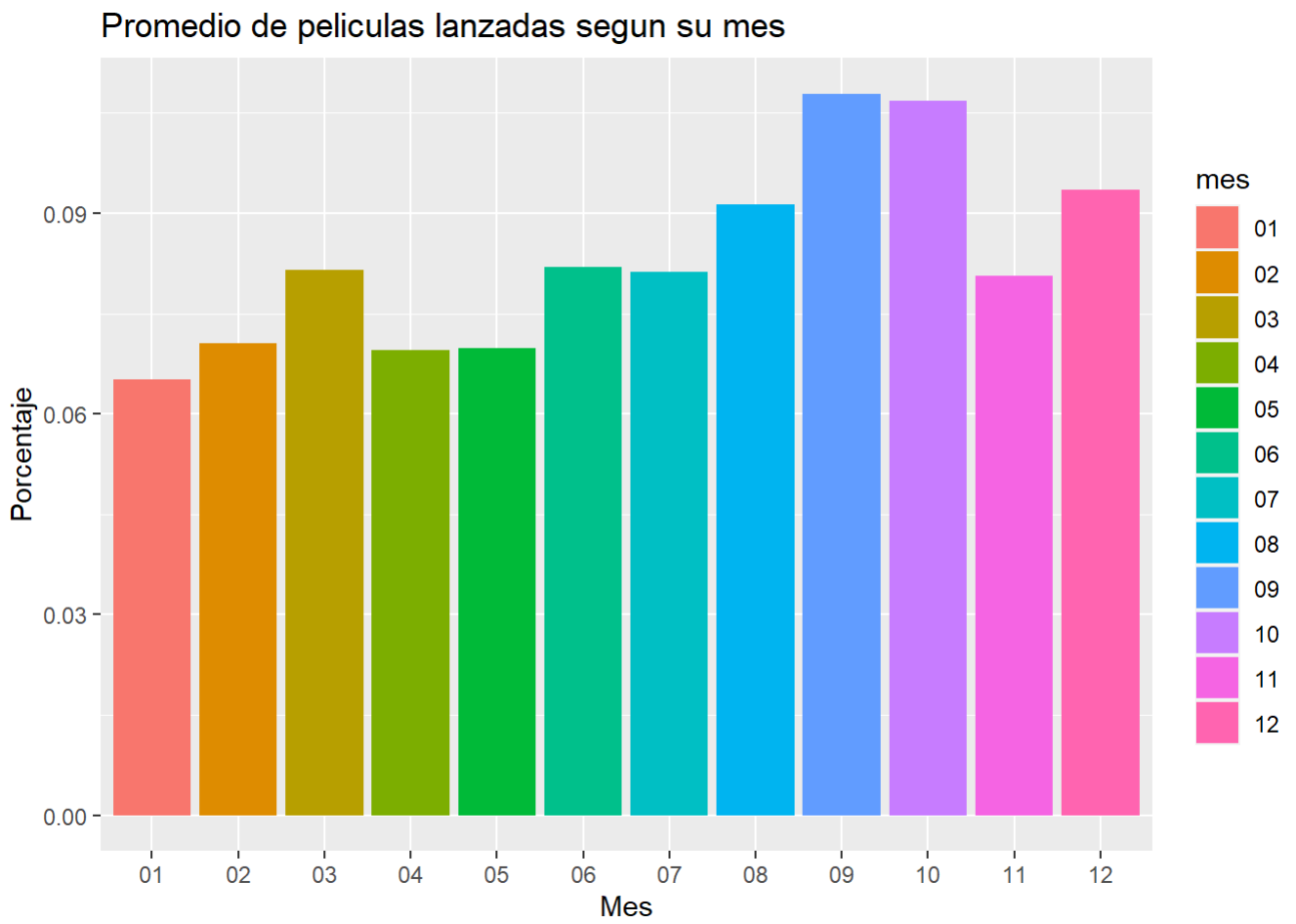
```
pregunta4.12
```



Para la elaboracion de esta grafica fue necesario promediar el ingreso de las peliculas respecto a su mes de lanzamiento, evidenciando que el mes con mayor ingresos segun su fecha de lanzamiento es **Junio** , siguiendo **Mayo** y como tercer lugar a **Julio**. Demostrando que el peor mes para lanzar una pelicula es **Septiembre**

4.13. ¿cuantas películas, en promedio, se han lanzado por mes?

```
pregunta4.13
```

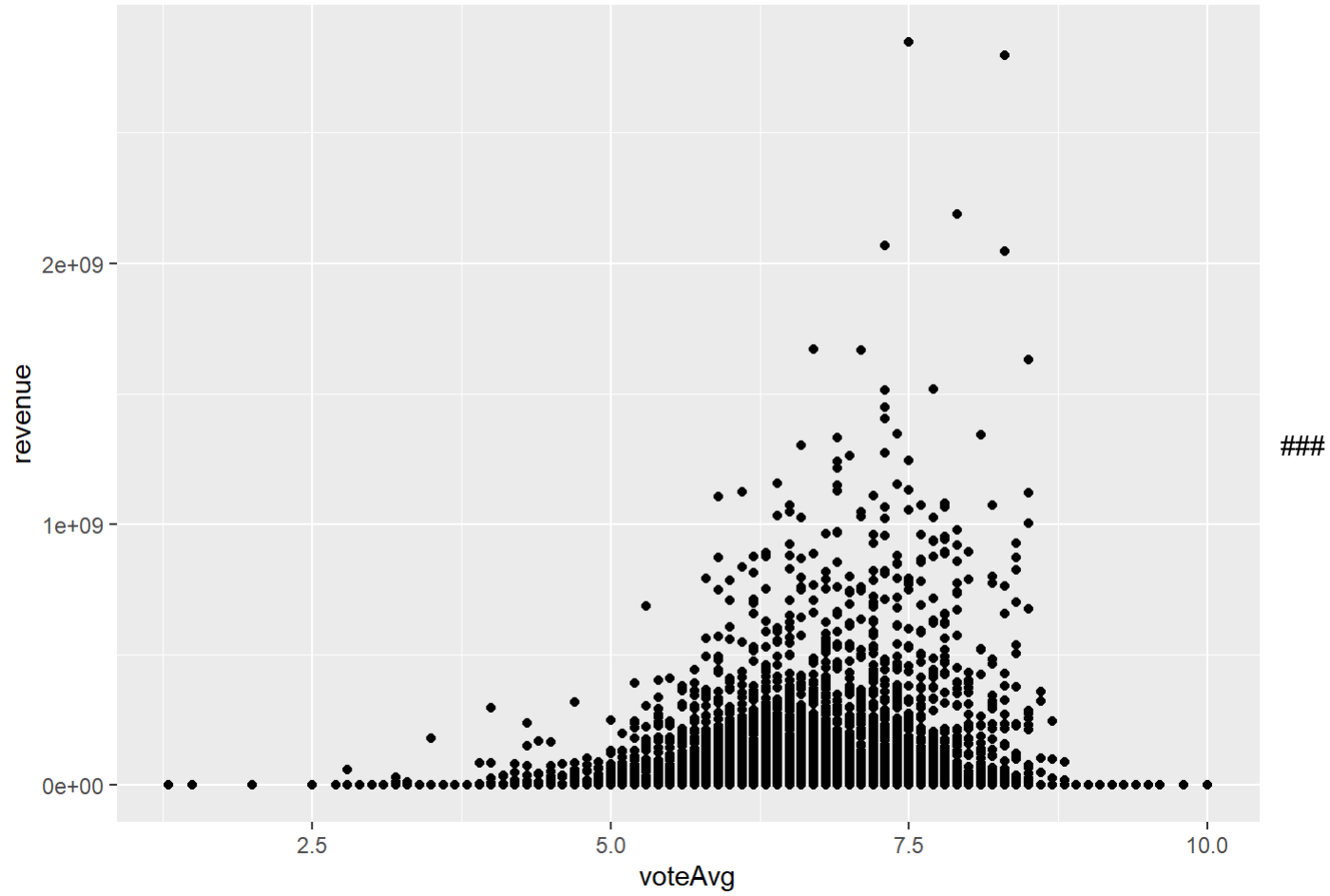


Como se puede ver en la grafica, son 3 los meses que representan mejor ingresos, los cuales son:

El mes que se encuentra con mayor lanzamiento de peliculas es **septiembre**, en segundo lugar **octubre** y seguido por **diciembre**

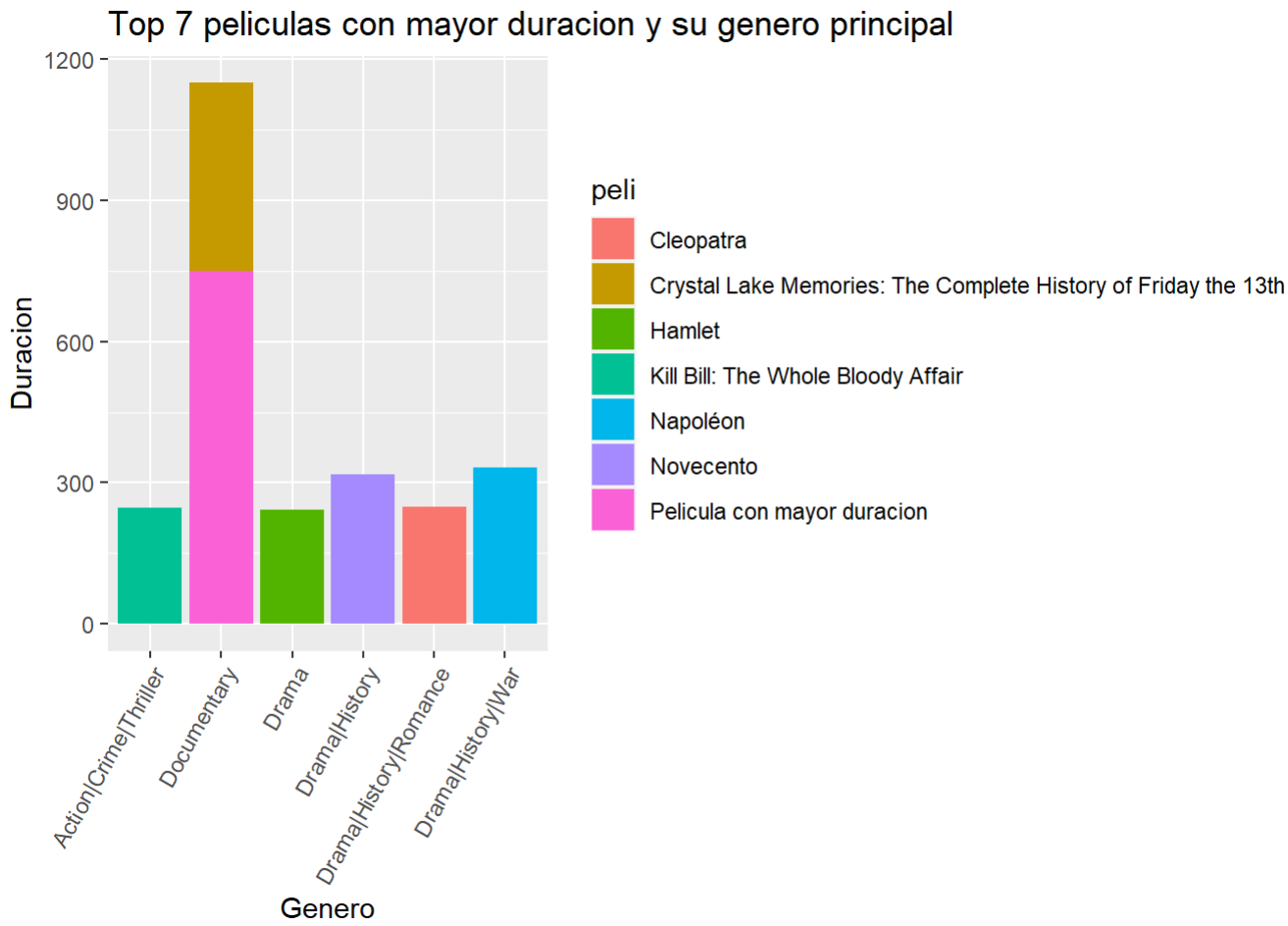
4.14. ¿Cómo se correlacionan las calificaciones con el éxito comercial?

```
pregunta4.14
```



4.15. ¿A qué género principal pertenecen las películas más largas?

```
pregunta4.15
```



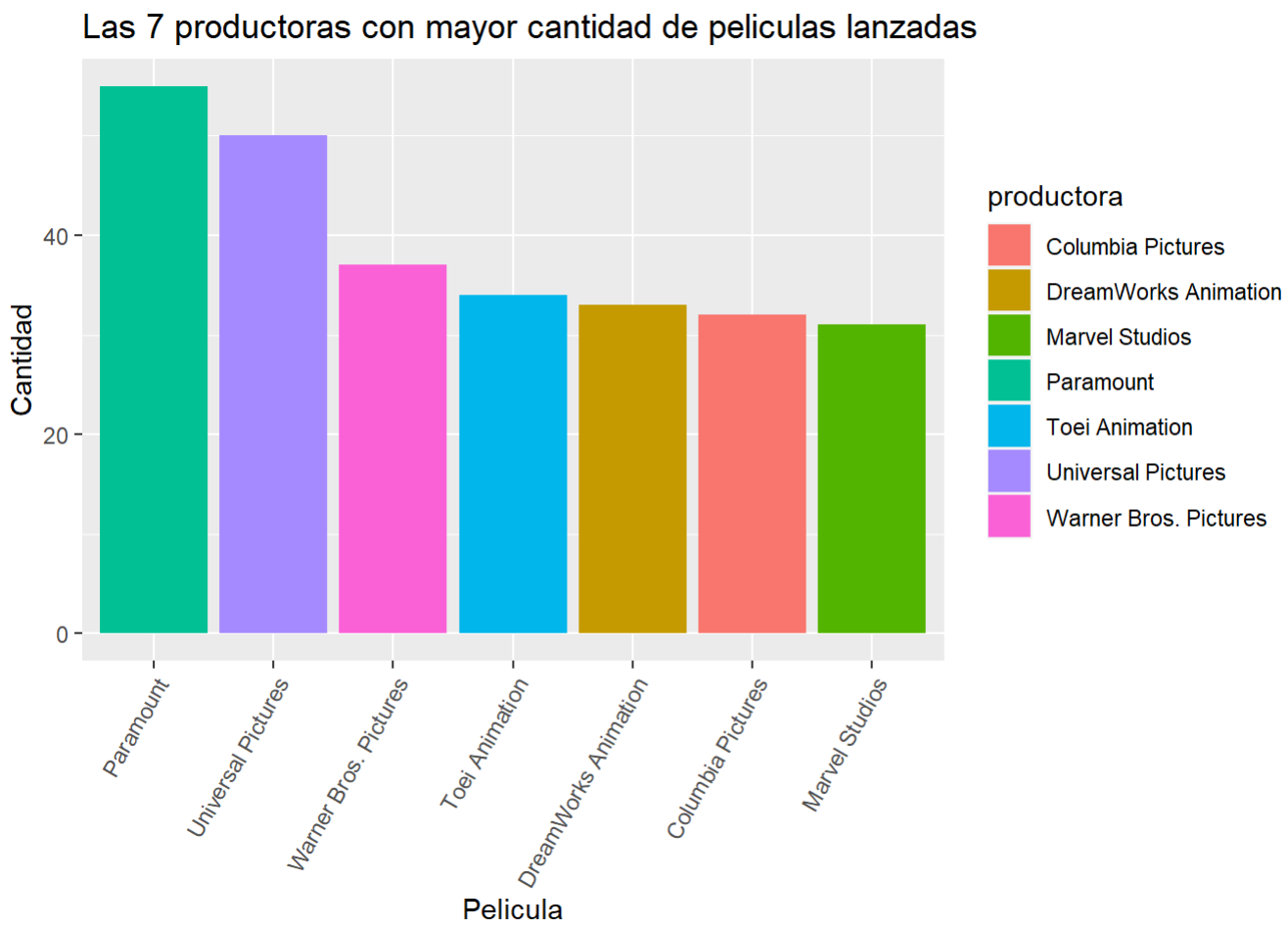
Como se observa en la grafica anterior , las 5 peliculas con mayor duracion y su respectivo genero fueron:

1. **<U+30DD><U+30CB><U+30E7><U+306F><U+3053><U+3046><U+3057><U+3066><U+751F><U+307E><U+308C><U+305F> <U+301C> <U+5BAE><U+FA11><U+99FF><U+306E><U+601D><U+8003><U+904E><U+7A0B> <U+301C>** con una duracion de **750** minutos y su genero es **Documentary** (cabe mencionar que para fines estadisticos fue necesario cambiar el nombre de esta pelicula por “Pelicula con mayor duracion” esto debido que su nombre era demasiado largo para mostrar graficamente).
2. **Crystal Lake Memories: The Complete History of Friday the 13th** con una duracion de **400** minutos y su genero es **Documentary**
3. **Napoléon** con una duracion de **333** minutos y su genero es **Drama|History|War**
4. **Novecento** con una duracion de **317** minutos y su genero es **Drama|History**
5. **Cleopatra** con una duracion de **248** minutos y su genero es **Drama|History|Romance**
6. **Kill Bill: The Whole Bloody Affair** con una duracion de **247** minutos y su genero es **Action|Crime|Thriller**
7. **Hamlet** con una duracion de **242** minutos y su genero es **Drama**

5. Preguntas extras

5.1 ¿Cuáles son las productoras con mayores películas lanzadas?

```
preguntaEx1
```

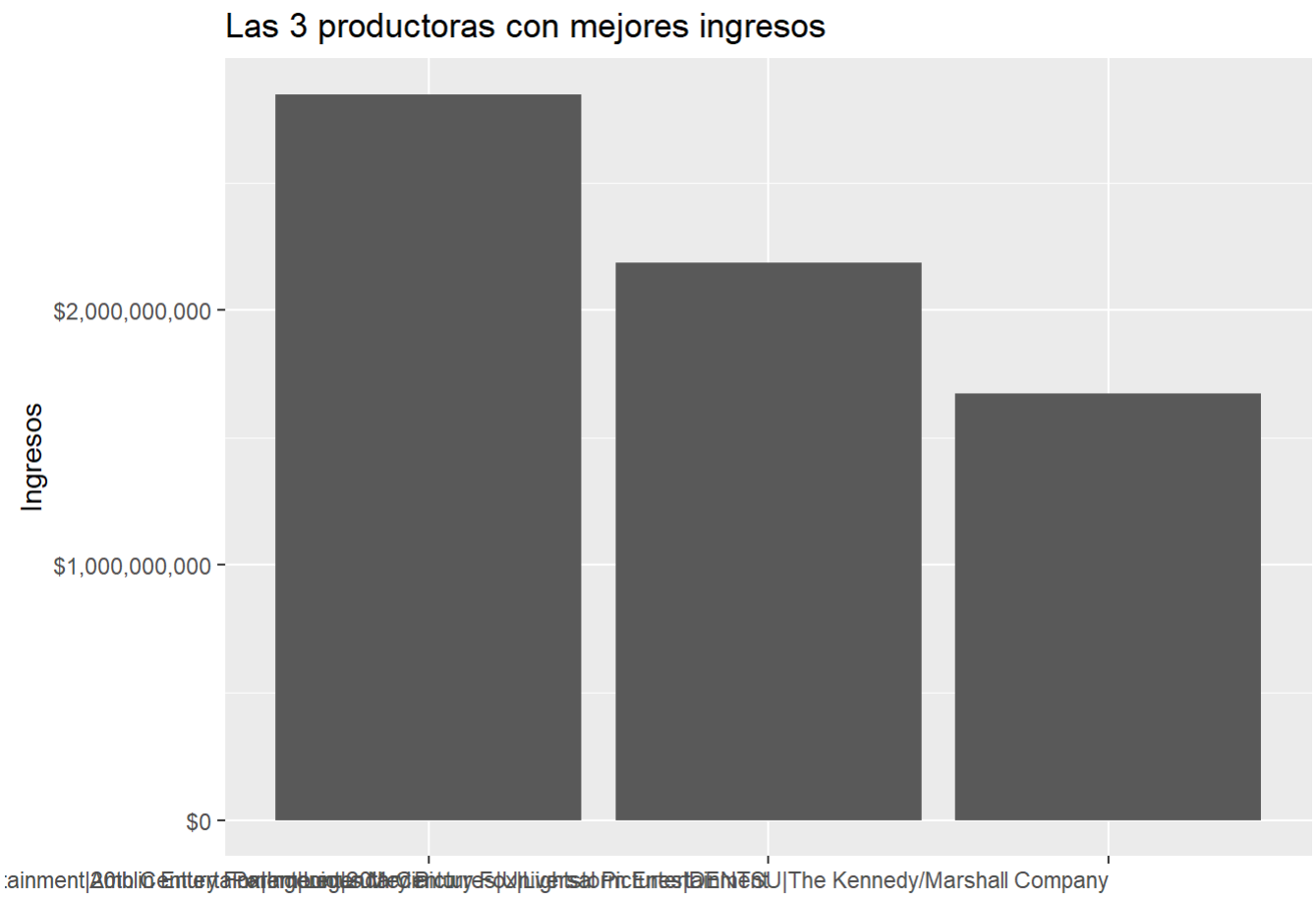


Como se observa en la grafica anterior las 7 productoras con mayor lanzamiento de peliculas son las siguientes:

1. **Paramount** con una cantidad de **55**
2. **Universal Pictures** con una cantidad de **50**
3. **Warner Bros. Pictures** con una cantidad de **37**
4. **Toei Animation** con una cantidad de **34**
5. **DreamWorks Animation** con una cantidad de **33**
6. **Columbia Pictures** con una cantidad de **32**
7. **Marvel Studios** con una cantidad de **31**

5.2 ¿Cuáles son las productoras con mejores ingresos?

```
preguntaEx2
```

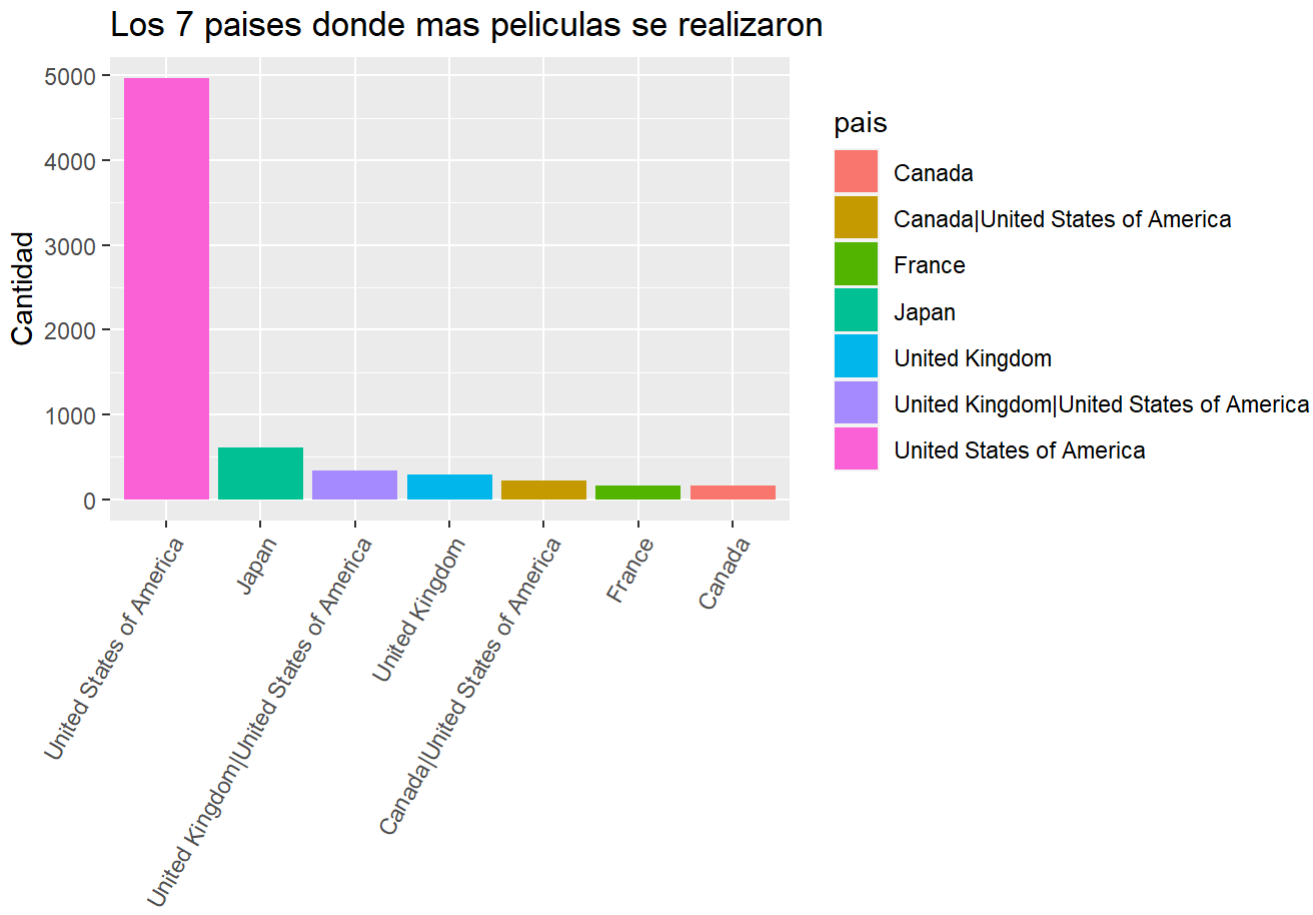


Como se observa en la grafica anterior las 3 productores con mejores ingresos son:

1. **Dune Entertainment|Lightstorm Entertainment|20th Century Fox|Ingenious Media** con una cantidad de **2.8472462^{9}** dolares
2. **Paramount|20th Century Fox|Lightstorm Entertainment** con una cantidad de **2.1874639^{9}** dolares
3. **Amblin Entertainment|Legendary Pictures|Universal Pictures|DENTSU|The Kennedy/Marshall Company** con una cantidad de **1.6717132^{9}** dolares

5.3 Paises donde se llevaron a cabo mas peliculas

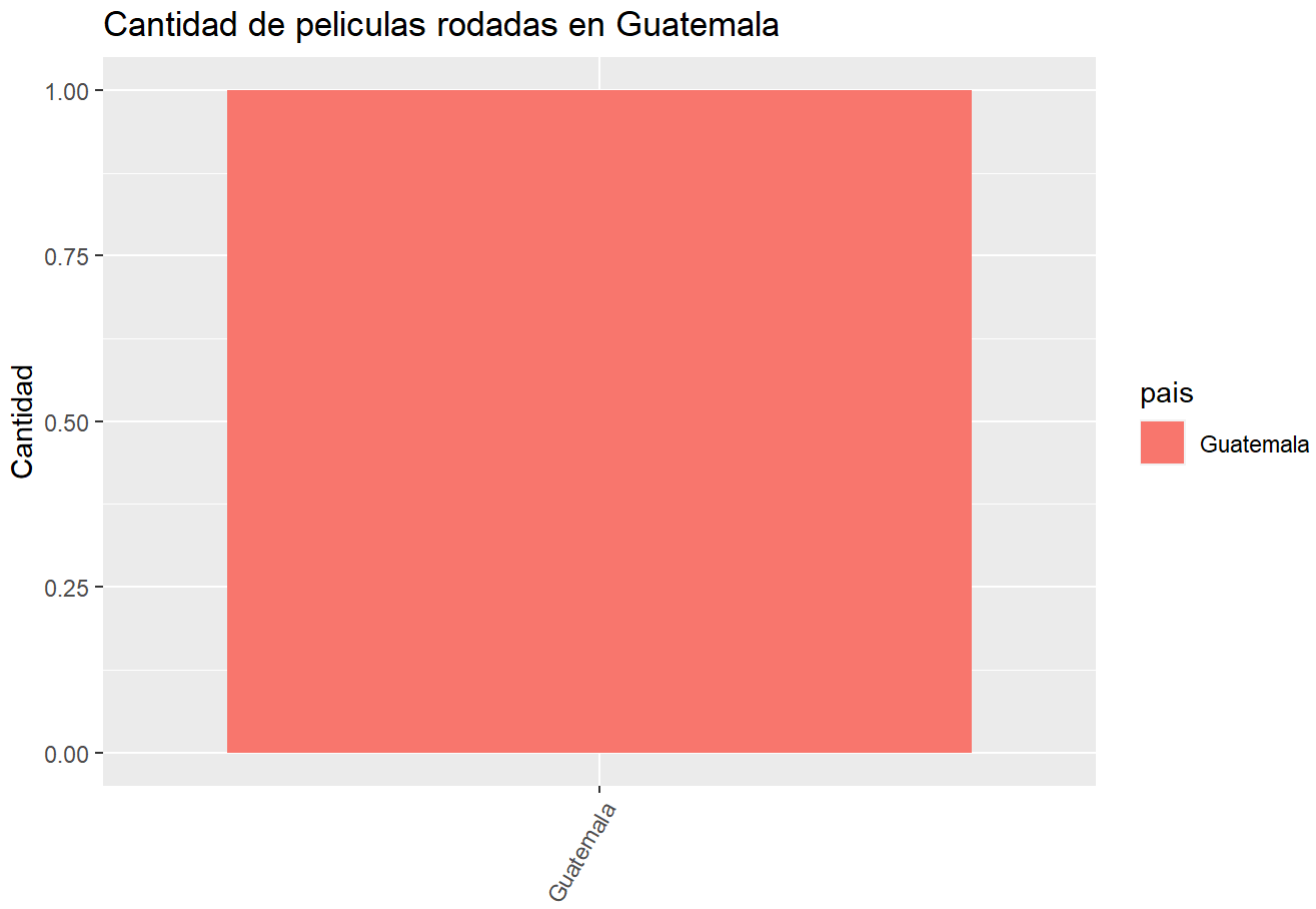
```
preguntaEx3
```



Como se observa en la grafica anterior los paises donde se llevaron a cabo mas peliculas son:
1. **United States of America** con una cantidad de **4971**
2. **Japan** con una cantidad de **613**
3. **United Kingdom|United States of America** con una cantidad de **339**
4. **United Kingdom** con una cantidad de **294**
5. **Canada|United States of America** con una cantidad de **223**
6. **France** con una cantidad de **164**
7. **Canada** con una cantidad de **157**

5.4 ¿Cuántas películas se han rodado en Guatemala?

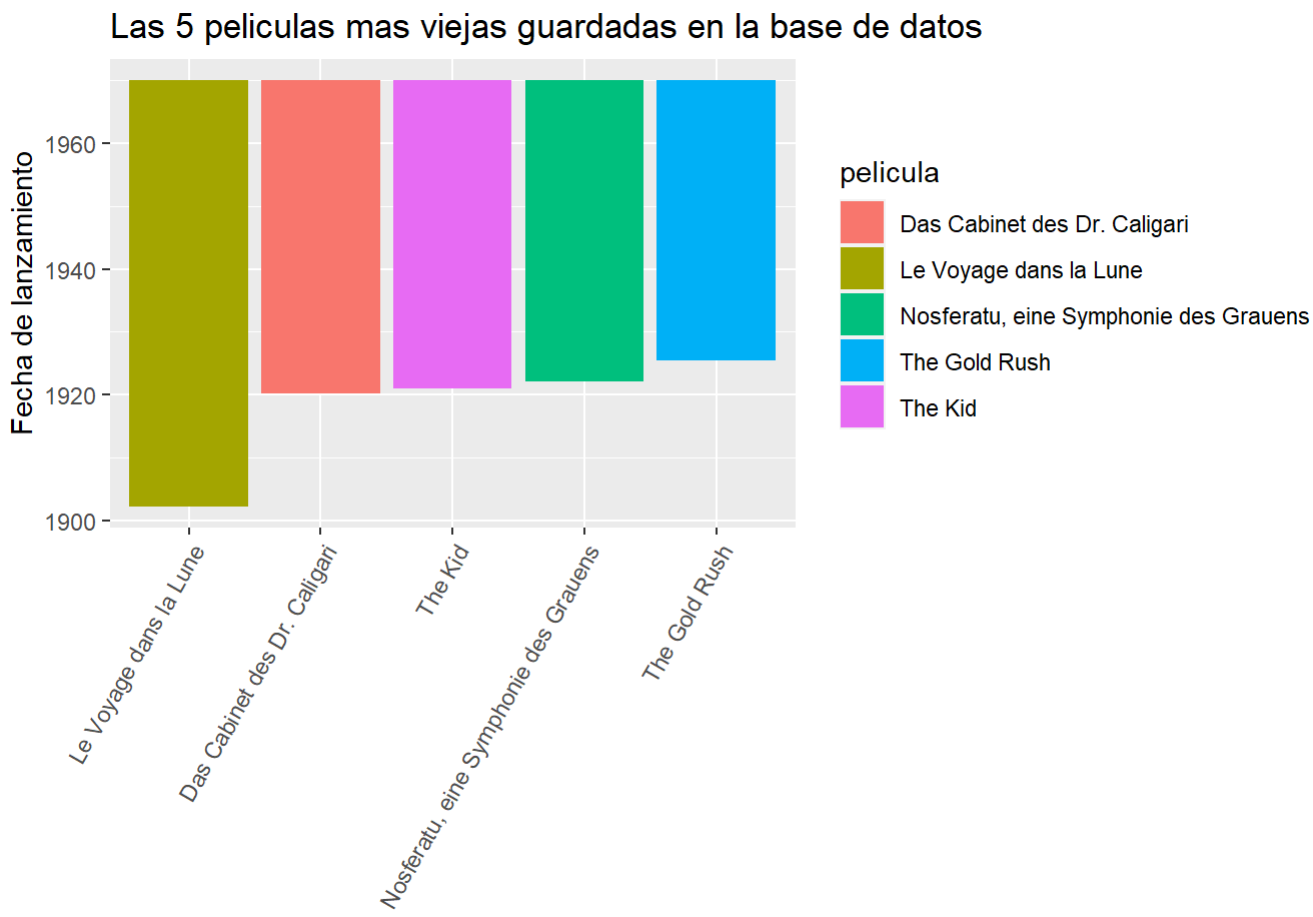
```
preguntaEx4
```



Como se observa la cantidad de peliculas rodadas en Guatemala son 1, siendo la pelicula **Exorcismo Documentado**

5.5 ¿Cuáles son las películas más viejas?

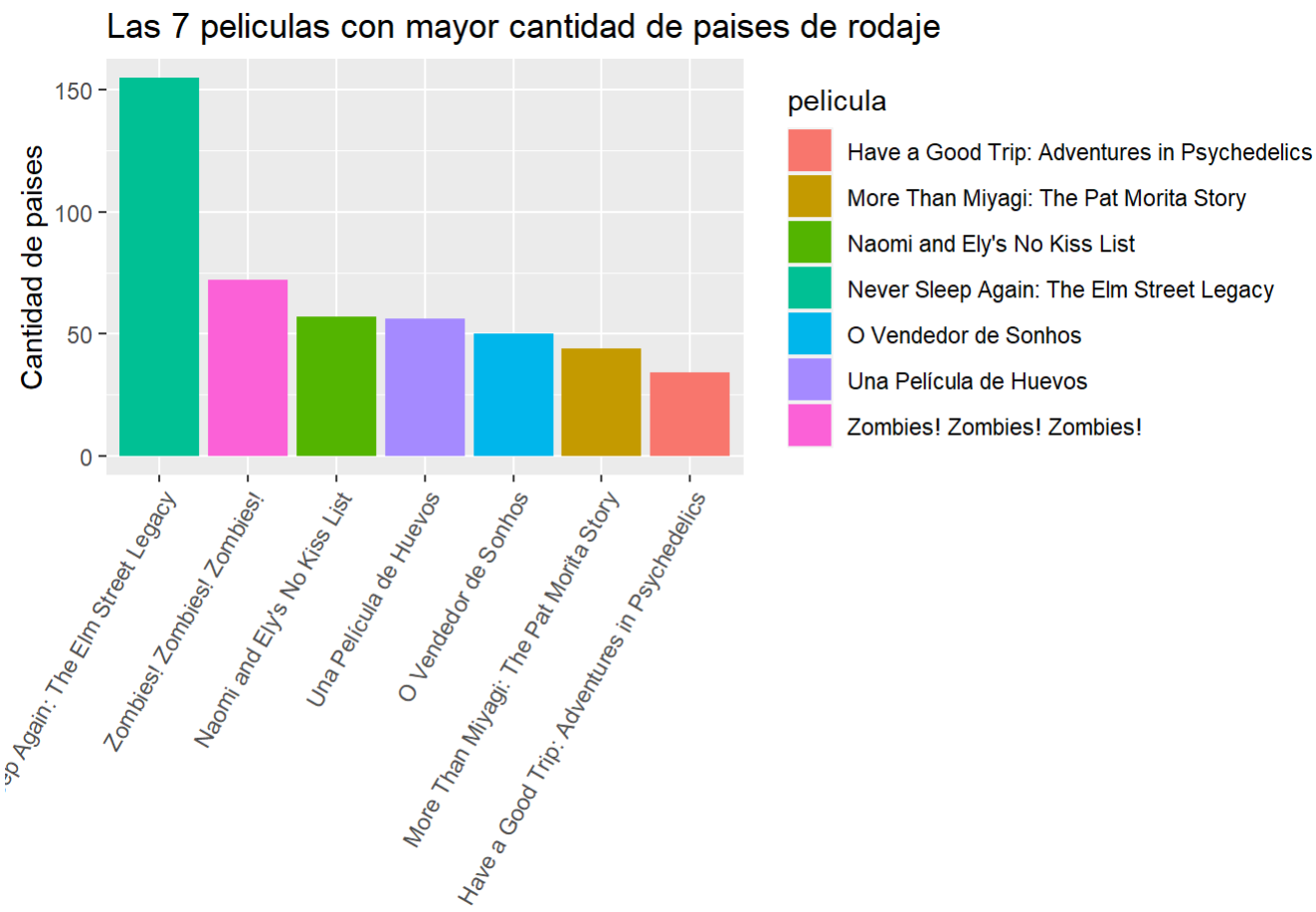
```
preguntaEx5
```



Como se observa en la grafica anterior las peliculas mas viejas guardadas en el sistem son:
1. **Le Voyage dans la Lune** lanzada el **1902-04-17**
2. **Das Cabinet des Dr. Caligari** lanzada el **1920-02-27**
3. **The Kid** lanzada el **1921-01-21**
4. **Nosferatu, eine Symphonie des Grauens** lanzada el **1922-02-17**
5. **The Gold Rush** lanzada el **1925-07-12**

5.6 Peliculas con mayores paises visitados para su rodaje

```
preguntaEx6
```



Como se observa en la grafica anterior las peliculas con mayor cantidad de paises de rodaje son:

1. **Never Sleep Again: The Elm Street Legacy** con una cantidad de **155** paises para llevarse a cabo
2. **Zombies! Zombies! Zombies!** con una cantidad de **72** paises para llevarse a cabo
3. **Naomi and Ely's No Kiss List** con una cantidad de **57** paises para llevarse a cabo
4. **Una Película de Huevos** con una cantidad de **56** paises para llevarse a cabo
5. **O Vendedor de Sonhos** con una cantidad de **50** paises para llevarse a cabo
6. **More Than Miyagi: The Pat Morita Story** con una cantidad de **44** paises para llevarse a cabo
7. **Have a Good Trip: Adventures in Psychedelics** con una cantidad de **34** paises para llevarse a cabo