Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
Minería de Datos

Hoja de Trabajo 1

Andrés Paíz 191142
Eduardo Ramírez 19946
Rene Ventura 19554

1. (3 puntos) Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

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


   popularity        releaseDate          voteAvg         voteCount      genresAmount    productionCoAmount
 Min.   :    4.258   Length:10000       Min.   : 1.300   Min.   :    1   Min.   : 0.000   Min.   : 0.000
 1st Qu.:   14.578   Class :character   1st Qu.: 5.900   1st Qu.:  120   1st Qu.: 2.000   1st Qu.: 2.000
 Median :   21.906   Mode  :character   Median : 6.500   Median :  415   Median : 3.000   Median : 3.000
 Mean   :   51.394                      Mean   : 6.483   Mean   : 1342   Mean   : 2.596   Mean   : 3.171
 3rd Qu.:   40.654                      3rd Qu.: 7.200   3rd Qu.: 1316   3rd Qu.: 3.000   3rd Qu.: 4.000
 Max.   :11474.647                      Max.   :10.000   Max.   :30788   Max.   :16.000   Max.   :89.000
 productionCountriesAmount  actorsAmount    castWomenAmount    castMenAmount
 Min.   :  0.000           Min.   :     0   Length:10000       Length:10000
 1st Qu.:  1.000           1st Qu.:    13   Class :character   Class :character
 Median :  1.000           Median :    21   Mode  :character   Mode  :character
 Mean   :  1.751           Mean   :  2148
 3rd Qu.:  2.000           3rd Qu.:    36
 Max.   :155.000           Max.   :919590
>
```

2. (5 puntos) Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)

id: cuantitativa continua

Budget: cuantitativa continua

genres: cualitativa nominal

homePage: cualitativa nominal

productionCompany:cualitativa nominal

productionCompanyCountry: cualitativa nominal

productionCountry: cualitativa nominal

revenue: cuantitativa continua

Runtime: cuantitativa continua

video: cualitativa nominal

actors: cualitativa nominal

actorsPopularity: cuantitativa continua

actorsCharacter: cualitativa nominal

originalTitle: cualitativa nominal

title: cualitativa nominal

OriginalLanguage: cualitativa nominal

popularity: cuantitativa continua

releaseDate: cuantitativa discreta

voteAvg: cuantitativa continua

voteCount: cuantitativa discreta

genresAmount: cuantitativa discreta

productionCoAmount: cuantitativa discreta

productionCountriesAmount: cuantitativa discreta

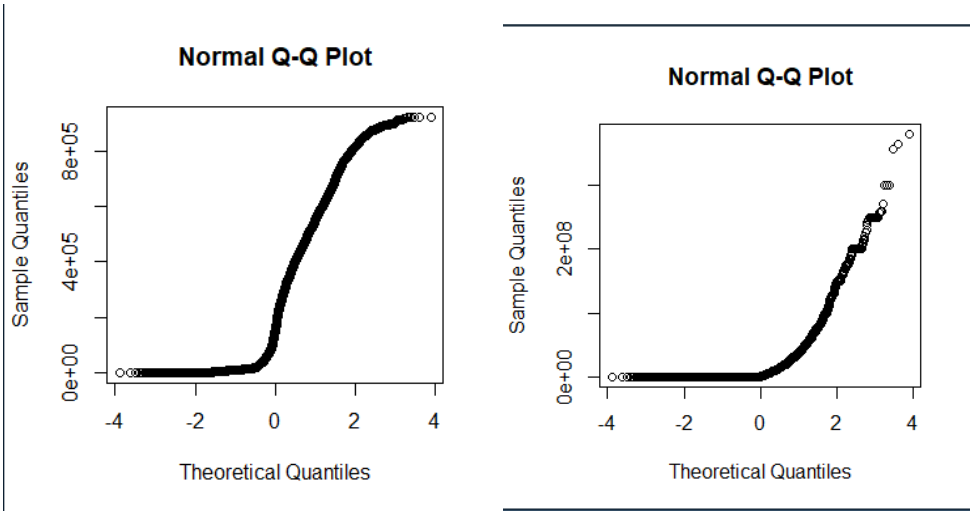
actorsAmount: cuantitativa discreta

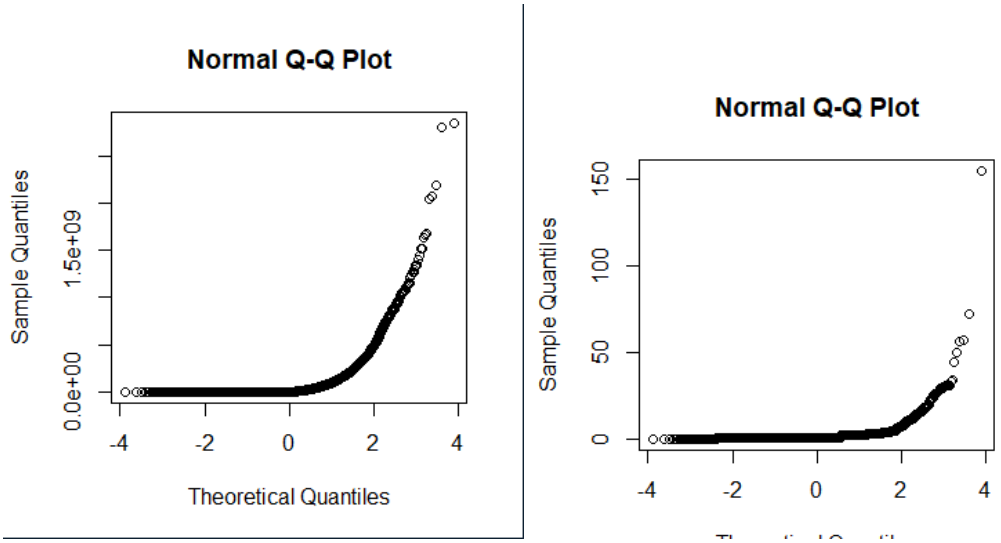
castWomenAmount: cuantitativa discreta

castMenAmount: cuantitativa discreta

3. (6 puntos) Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados.

Se puede observar que ninguna variable cuantitativa sigue una distribución normal y esto puede ser debido al tipo de datos que se estan analizando los cuales reciben muchisimos datos y son muy variados.

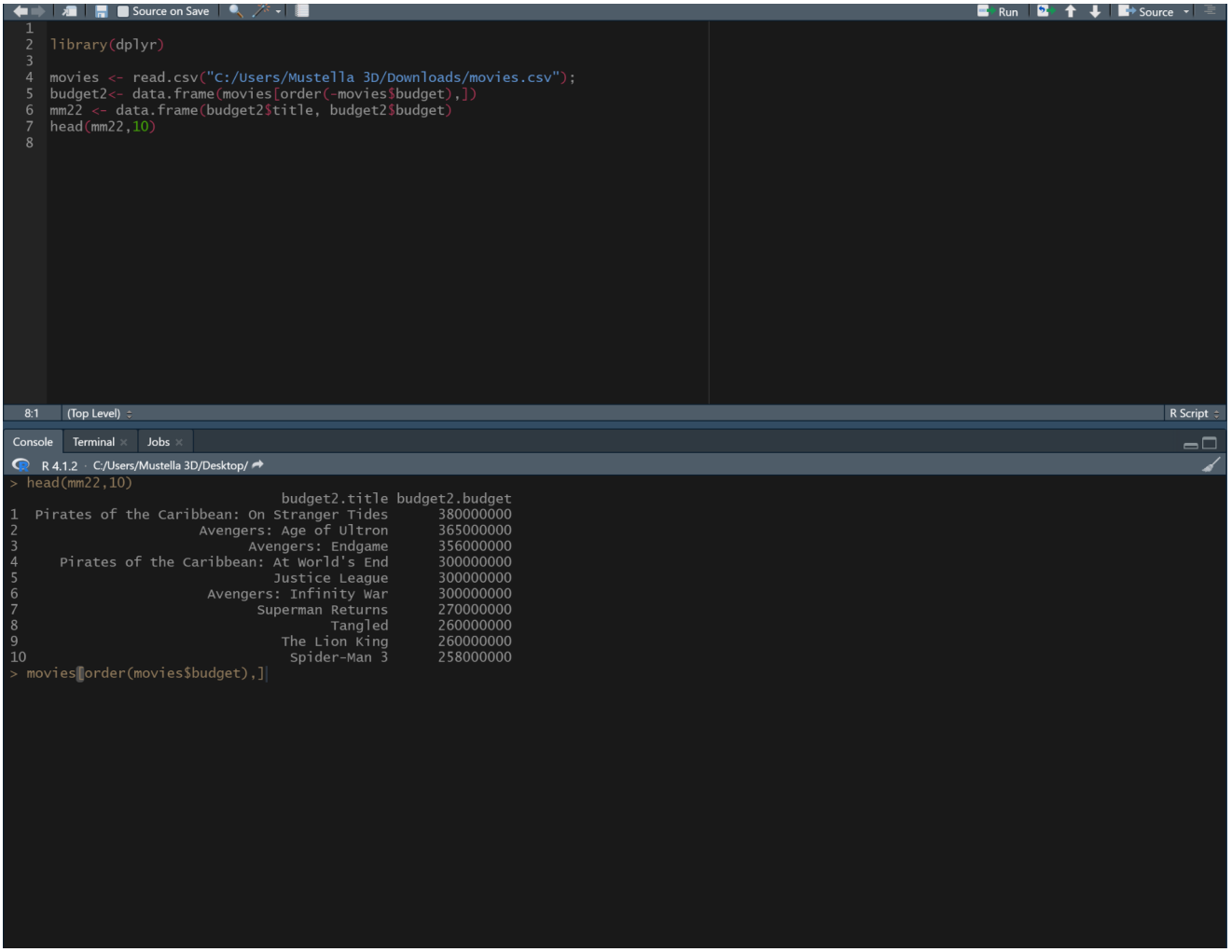




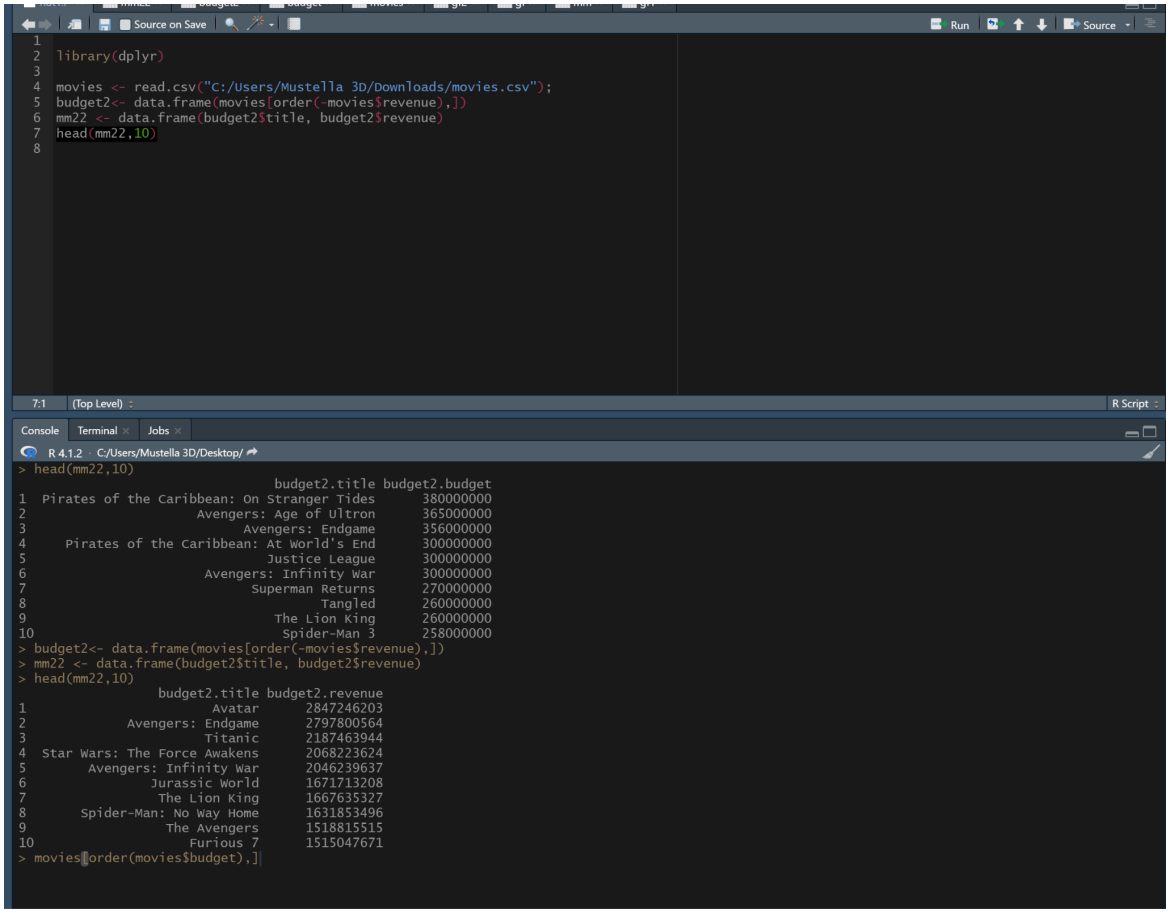

```r
#Ejercicio 3

library(scales)
datos<-read.csv('movies.csv')
qqnorm(datos$id) #no es distribucion normal
qqnorm(datos$budget) #no es distribucion normal
qqnorm(datos$revenue) #no es distribucion normal
shapiro.test(datos$Runtime) #no es distribucion normal
qqnorm(datos$popularity) #no es distribucion normal
qqnorm(datos$releaseDate) #no es distribucion normal
qqnorm(datos$voteAvg)#no es distribucion normal
qqnorm(scale(datos$voteCount))#no es distribucion normal
qqnorm(datos$genresAmount)#no es distribucion normal
qqnorm(datos$productionCoAmount)#no es distribucion normal
qqnorm(datos$productionCountriesAmount)#no es distribucion normal
qqnorm(datos$actorsAmount)#no es distribucion normal
qqnorm(datos$castWomenAmount)#no es distribucion normal
qqnorm(datos$castMenAmount)#no es distribucion normal
```

4. Responda las siguientes preguntas:

4.1. (3 puntos) ¿Cuáles son las 10 películas que contaron con más presupuesto?


Source on Save
Run
Source
1 library(dplyr)
2
3 movies <- read.csv("C:/Users/Mustella 3D/Downloads/movies.csv");
4 budget2<- data.frame(movies[order(-movies$budget),])
5 mm22 <- data.frame(budget2$title, budget2$budget)
6 head(mm22,10)
7
8:1
(Top Level)
R Script
Console
Terminal
Jobs
R 4.1.2 · C:/Users/Mustella 3D/Desktop/
> head(mm22,10)
budget2.title budget2.budget
1 Pirates of the Caribbean: On Stranger Tides 380000000
2 Avengers: Age of Ultron 365000000
3 Avengers: Endgame 356000000
4 Pirates of the Caribbean: At World's End 300000000
5 Justice League 300000000
6 Avengers: Infinity War 300000000
7 Superman Returns 270000000
8 Tangled 260000000
9 The Lion King 260000000
10 Spider-Man 3 258000000
> movies[order(movies$budget),]

4.2. (3 puntos) ¿Cuáles son las 10 películas que más ingresos tuvieron?



```
library(dplyr)

movies <- read.csv("C:/Users/Mustella 3D/Downloads/movies.csv");
budget2<- data.frame(movies[order(-movies$revenue),])
mm22 <- data.frame(budget2$title, budget2$revenue)
head(mm22,10)
```

```
> head(mm22,10)
                                budget2.title budget2.budget
1  Pirates of the Caribbean: On Stranger Tides      380000000
2                      Avengers: Age of Ultron      365000000
3                             Avengers: Endgame      356000000
4     Pirates of the Caribbean: At World's End      300000000
5                               Justice League      300000000
6                       Avengers: Infinity War      300000000
7                            Superman Returns      270000000
8                                     Tangled      260000000
9                                  The Lion King      260000000
10                                  Spider-Man 3      258000000
> budget2<- data.frame(movies[order(-movies$revenue),])
> mm22 <- data.frame(budget2$title, budget2$revenue)
> head(mm22,10)
                  budget2.title budget2.revenue
1                        Avatar      2847246203
2             Avengers: Endgame      2797800564
3                       Titanic      2187463944
4  Star Wars: The Force Awakens      2068223624
5         Avengers: Infinity War      2046239637
6                 Jurassic World      1671713208
7                  The Lion King      1667635327
8        Spider-Man: No Way Home      1631853496
9                   The Avengers      1518815515
10                     Furious 7      1515047671
> movies[order(movies$budget),]
```

4.3. (3 puntos) ¿Cuál es la película que más votos tuvo?

```
13
14  aa <- data.frame(movies$title, movies$voteCount)
15  vote<- data.frame(aa[order(-aa$movies.voteCount),])
16  head(vote,1)
17
```

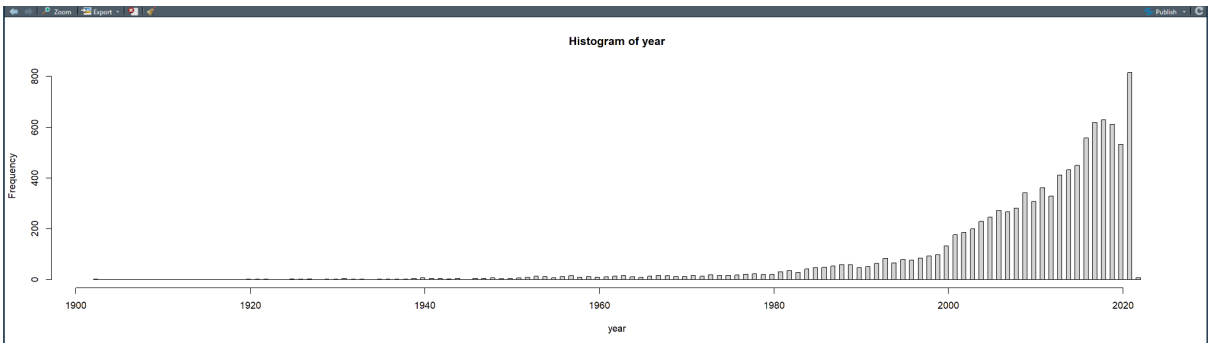
```
6:1  (Top Level)
nsole  Terminal  Jobs
 R 4.1.2 · C:/Users/Mustella 3D/Desktop/
head(vote,1)
   movies.title movies.voteCount
12    Inception           30788
votes$voteCount
```

4.4. (3 puntos) ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios?

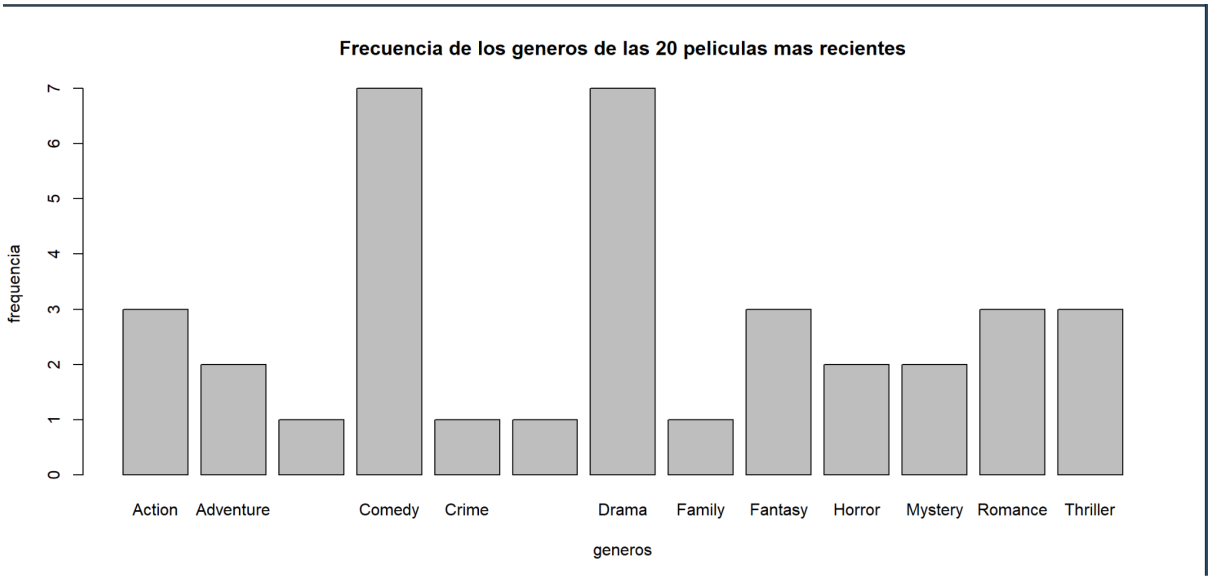
```
17
18  aa <- data.frame(movies$title, movies$voteAvg)
19  vote<- data.frame(aa[order(aa$movies.voteAvg),])
20  head(vote,1)
```

```
20:1  (Top Level)  R Script
Console  Terminal  Jobs
R 4.1.2 · C:/Users/Mustella 3D/Desktop/
> head(vote,1)
                                                  movies.title movies.voteAvg
9787 DAKAICHI -I'm Being Harassed by the Sexiest Man of the Year- The Movie: In Spain           1.3
>
```

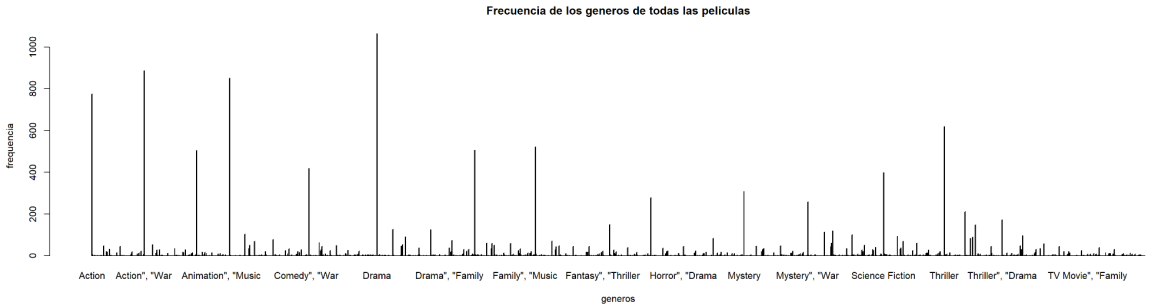
4.5. (8 puntos) ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras



4.6. (9 puntos) ¿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

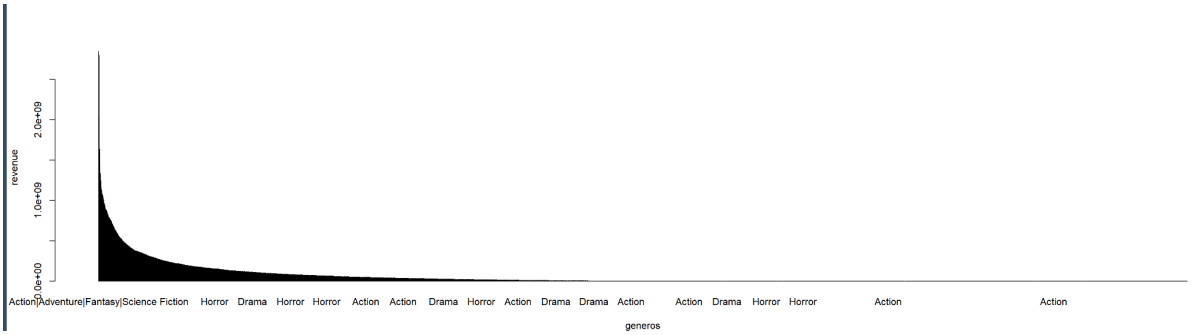


Para este gráfico se separaron los géneros de las películas, ya que habían películas con más de un género, al momento de generar el gráfico se puede observar que drama y comedia tienen un ligero empate. Estos dos géneros estuvieron de modo individual o con algún otro género asignado para las películas.

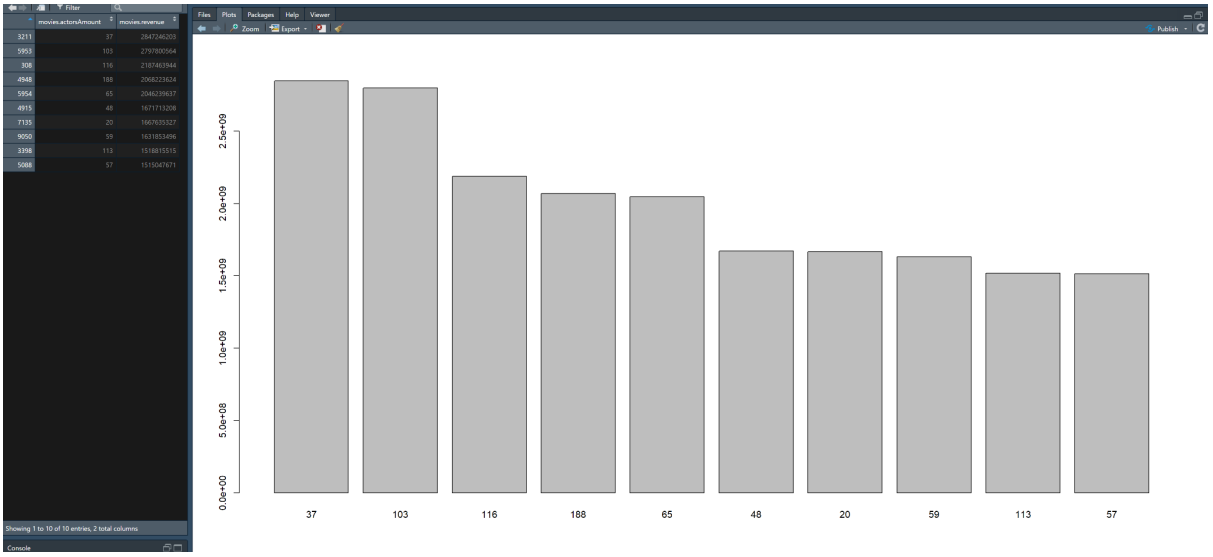


Este gráfico representa los géneros de las peliculas, pero esta vez no fueron separados en los “subgeneros”.

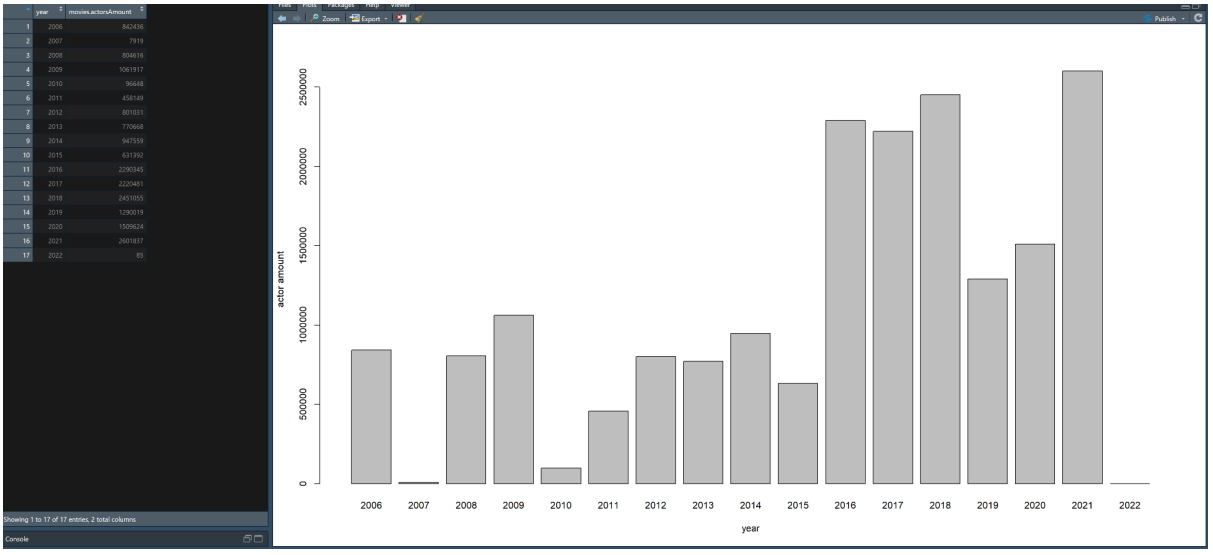
4.7. (8 puntos) ¿Las películas de qué género principal obtuvieron mayores ganancias?



4.8. (3 puntos) ¿La cantidad de actores influye en los ingresos de las películas?¿se han hecho
películas con más actores en los últimos años?

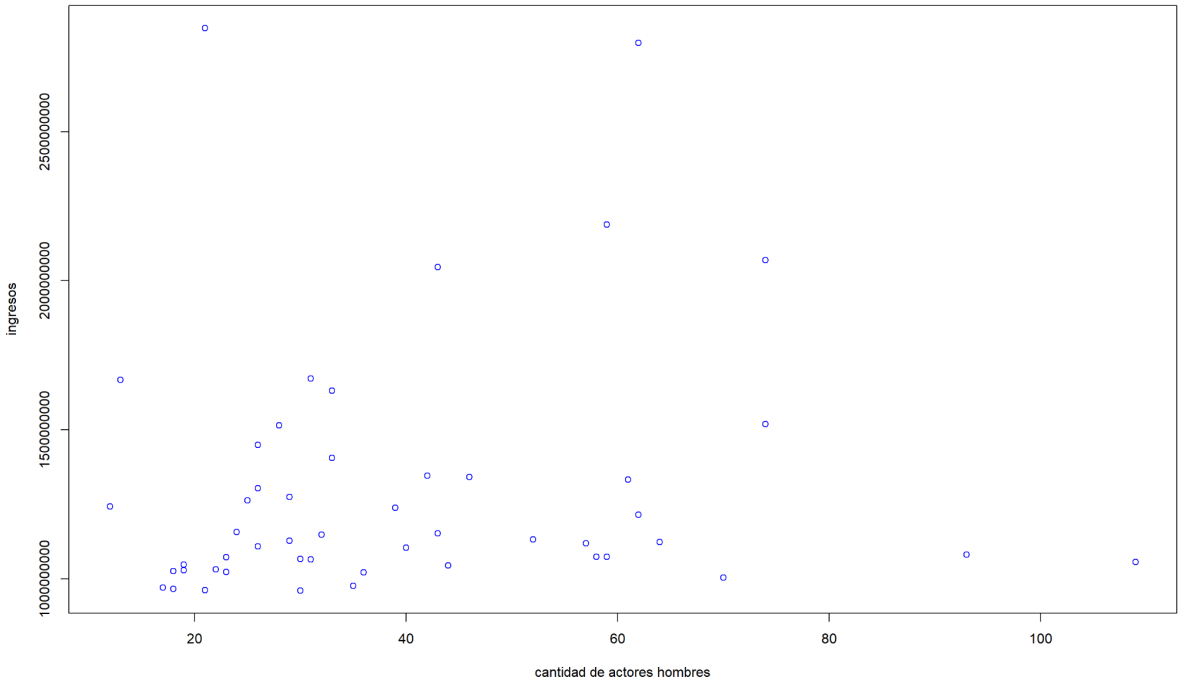


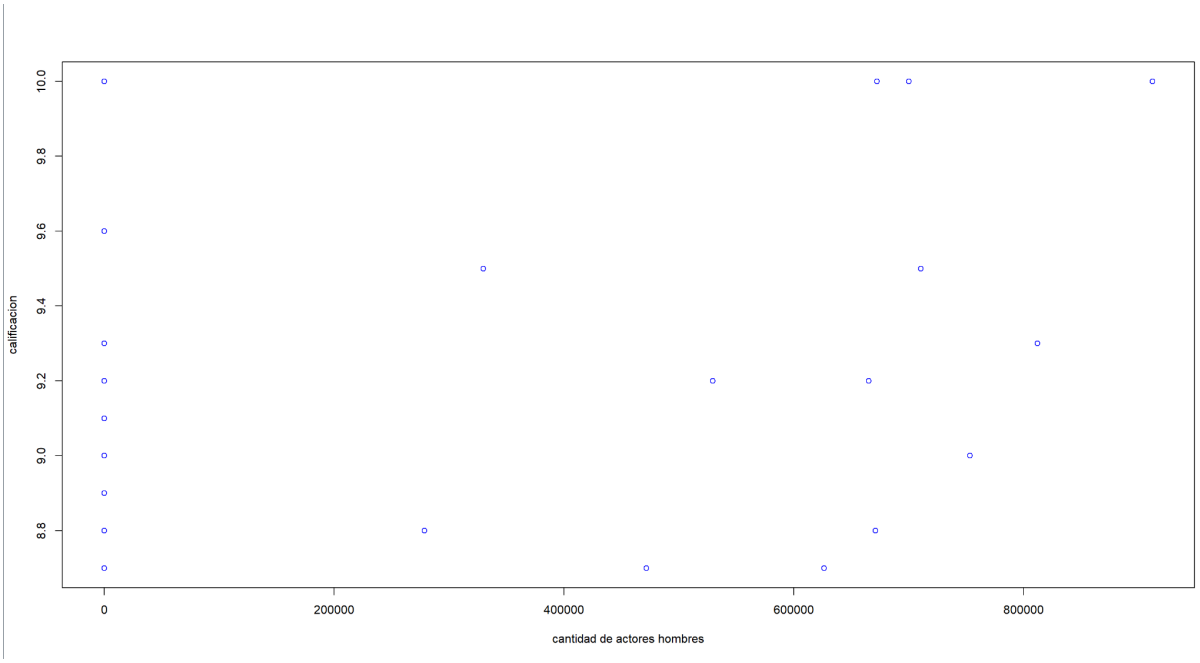
Como se observa en la gráfica anterior, la cantidad de ingresos no depende de la cantidad de actores ya que la película con más cantidad de ingresos solo tiene 37 actores y se puede observar que hay cantidades de actores que no pasan los 100 que generaron una gran cantidad de ingresos.

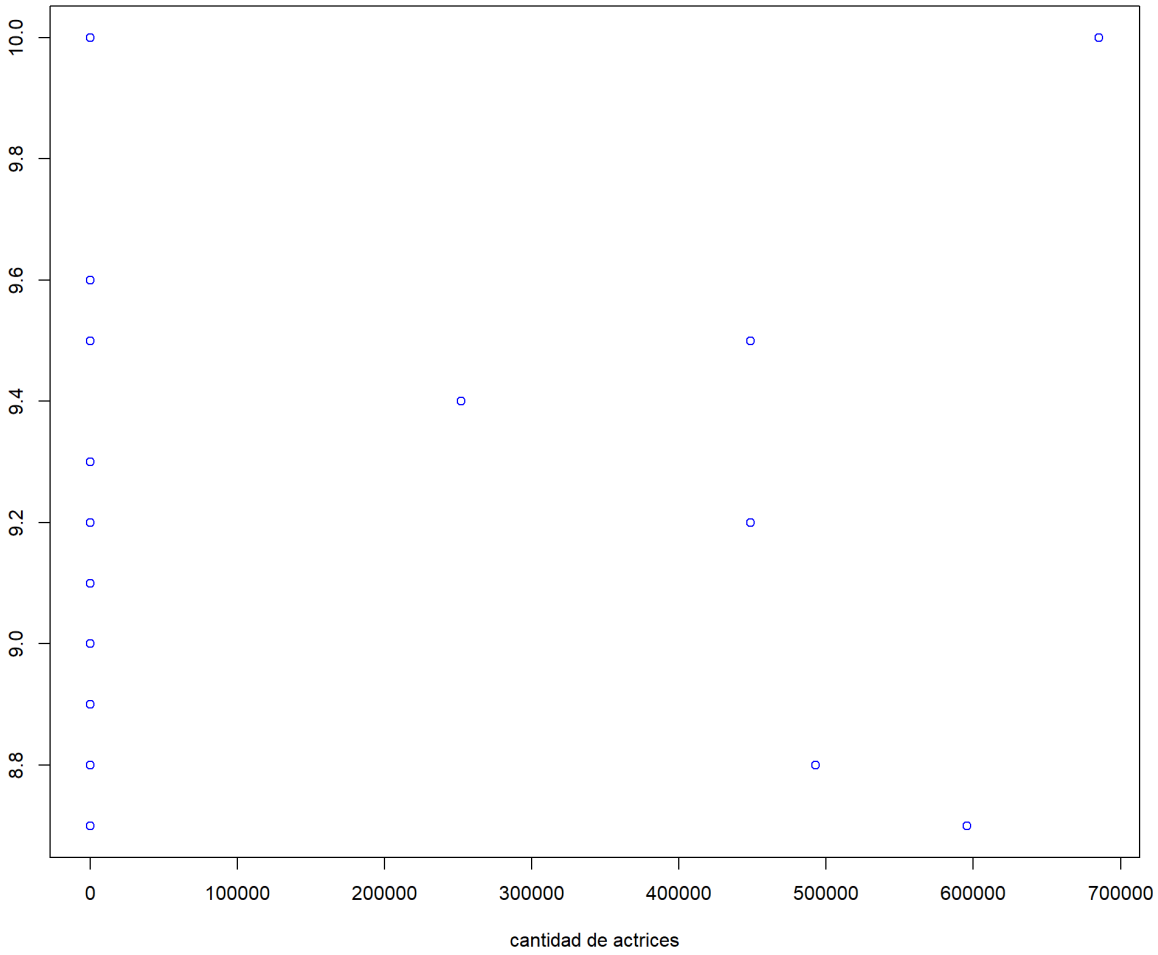


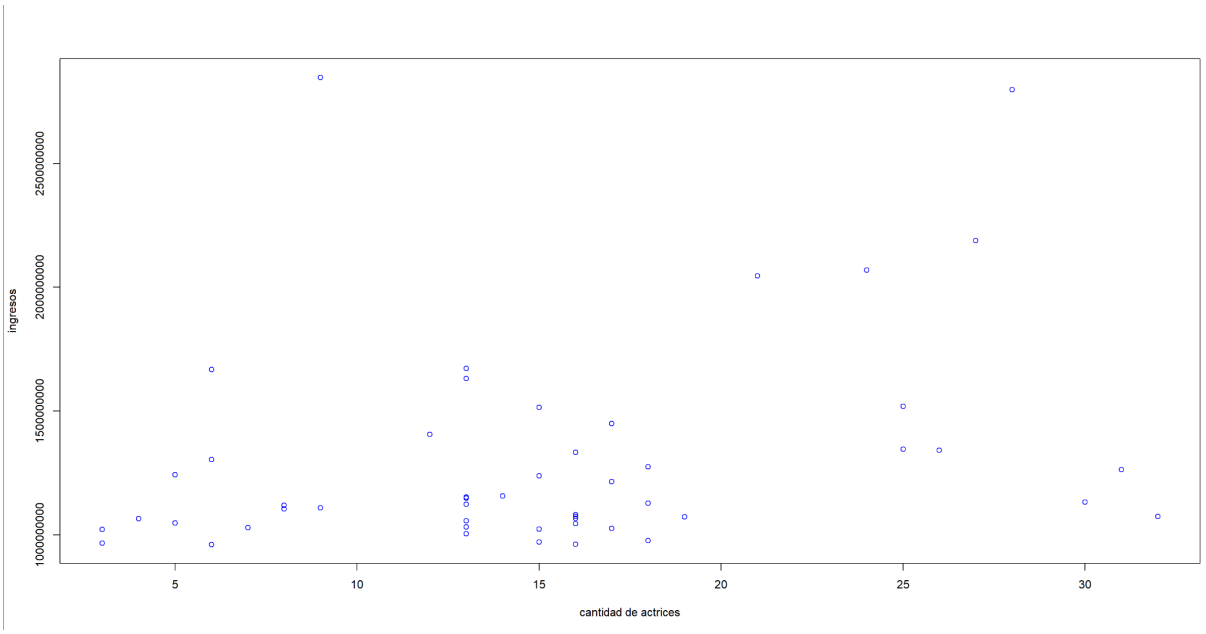
Se puede observar que a partir del 2016 la cantidad de actores incrementó casi el doble a excepción del 2020 debido a la pandemia que se está viviendo actualmente.

4.9. (3 puntos) ¿Es posible que la cantidad de hombres y mujeres en el reparto influya en la
popularidad y los ingresos de las películas?





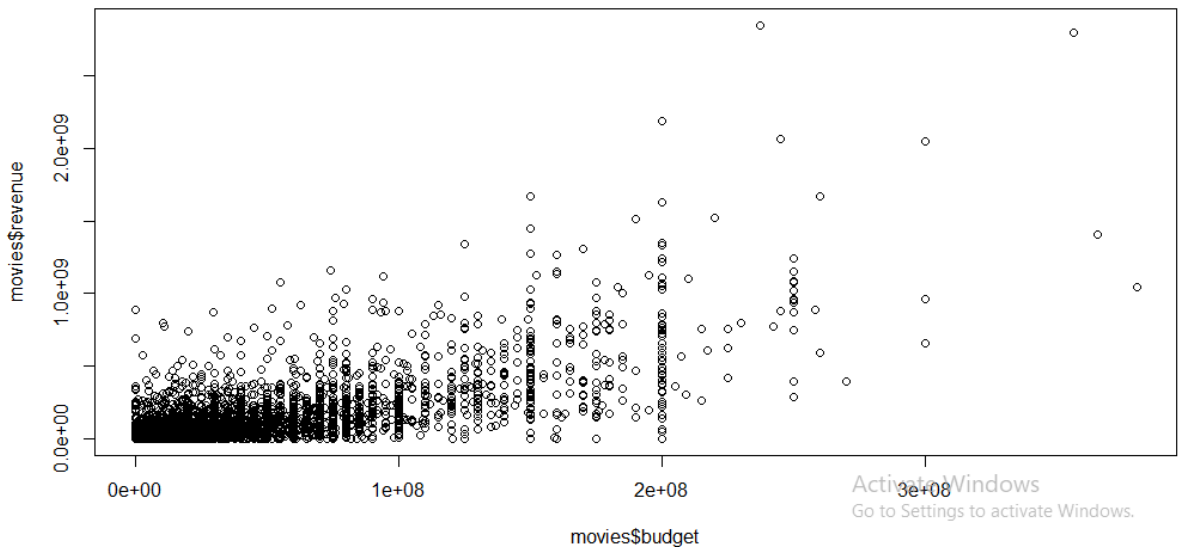




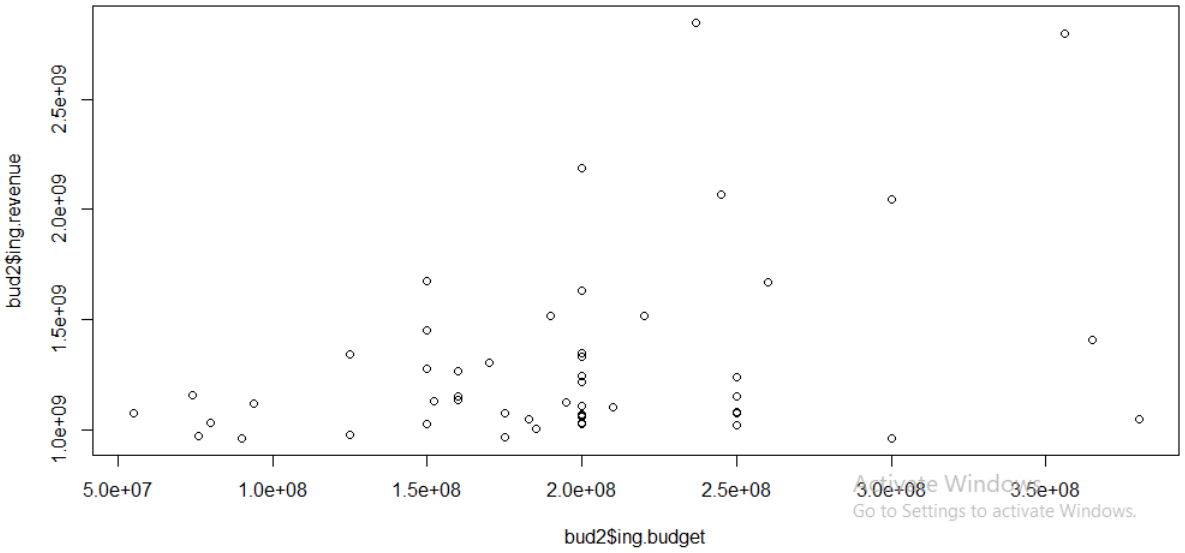
4.10. (8 puntos) ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

	pop.director	pop.voteAvg
1	Thomas Coven	10.0
2	Víctor Barba|Juan Olivares	10.0
3	Rebecca Sugar	10.0
4	Laurent Bouzereau	10.0
5	Kaku Arakawa	10.0
6	Christin Baker	10.0
7		10.0
8	Miguel Angel Zavala	10.0
9		9.8
10	Dave Bullock|Troy Adomitis|Victor Cook	9.6
11	Samuel Leong	9.5
12		9.5
13		9.5
14	Won Myeong-jun	9.5
15	Selena Quintanilla	9.4
16	Haruo Sotozaki	9.3
17	Haruo Sotozaki	9.3
18		9.2
19	Ulises Valencia	9.2

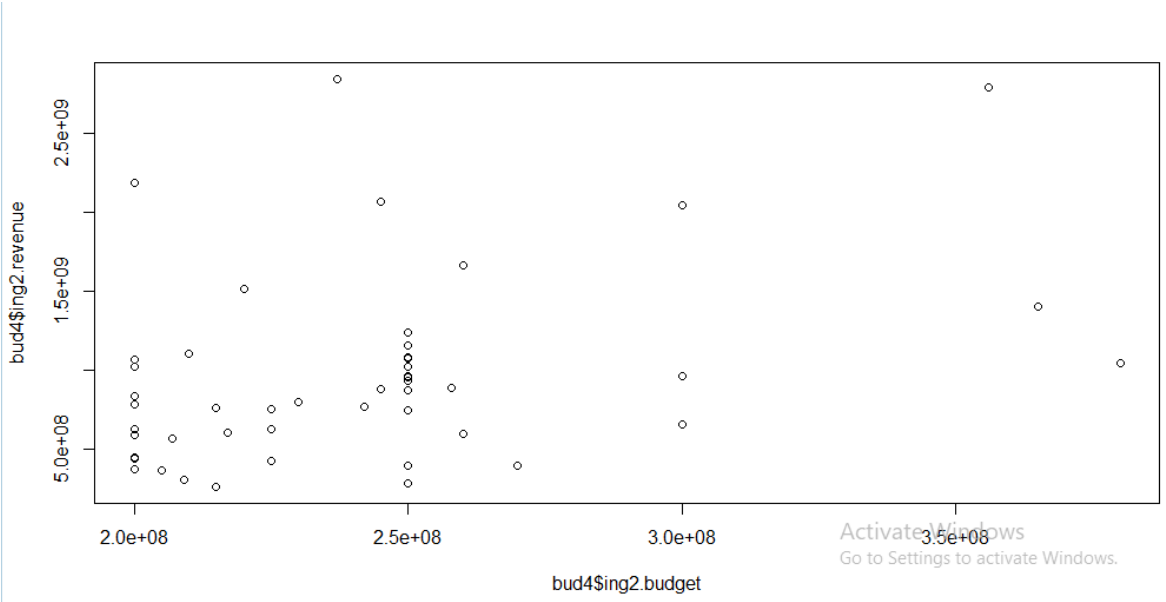
4.11. (8 puntos) ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión



Utilizando este diagrama de dispersión no se puede observar una relación clara entre la variable de revenue con la de budget para todas las películas.

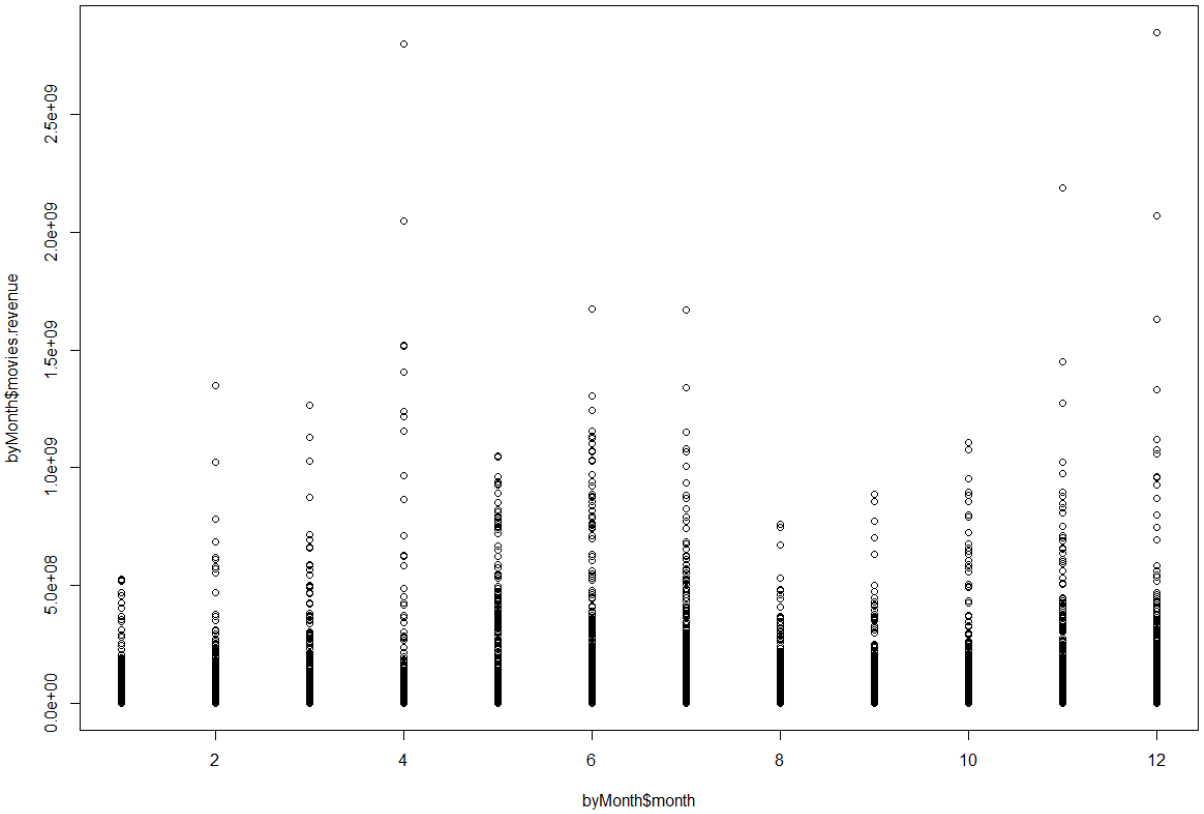


Al limitar el número de películas a 50, con el fin de observar las 50 películas con mayores ingresos junto con su presupuesto e igualmente no se observa una relación significativa entre las dos variables.



Al limitar el número de películas a 50, con el fin de observar las 50 películas con mayores presupuestos junto con sus ingresos e igualmente no se observa una relación significativa entre las dos variables.

4.12. (7 puntos) ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?




```
> mean(byMonth[byMonth$month == 1, 1])
[1] 33773691
> mean(byMonth[byMonth$month == 2, 1])
[1] 42908353
> mean(byMonth[byMonth$month == 3, 1])
[1] 51115942
> mean(byMonth[byMonth$month == 4, 1])
[1] 52595654
> mean(byMonth[byMonth$month == 5, 1])
[1] 87845442
> mean(byMonth[byMonth$month == 6, 1])
[1] 94747108
> mean(byMonth[byMonth$month == 7, 1])
[1] 76028696
> mean(byMonth[byMonth$month == 8, 1])
[1] 35970079
> mean(byMonth[byMonth$month == 9, 1])
[1] 31928917
> mean(byMonth[byMonth$month == 10, 1])
[1] 38987332
> mean(byMonth[byMonth$month == 11, 1])
[1] 71492112
> mean(byMonth[byMonth$month == 12, 1])
[1] 74358880
>
```

Aquí se puede observar la diferencia entre los ingresos de cada mes y la forma en que están distribuidos los ingresos. Podemos afirmar que existe una gran diferencia entre ciertos meses, la diferencia entre el peor mes (Septiembre) y el mejor (Junio) es de 62,818,191. Se puede observar como los ingresos de los meses situados a la mitad y al final del año son mayores. Mientras que los que están situados entre esas temporadas son menores.

4.13. (8 puntos) ¿En qué meses se han visto los lanzamientos con mejores ingresos?¿Cuántas películas, en promedio, se han lanzado por mes?

Utilizando el diagrama anterior podemos observar que los lanzamientos con mayores ingresos (los mejores 5) pertenecen a los meses de diciembre, noviembre y abril.

No logramos comprender claramente la segunda parte de la pregunta, por lo cual respondimos lo que pensamos que se preguntaba, disculpa si no es correcto.

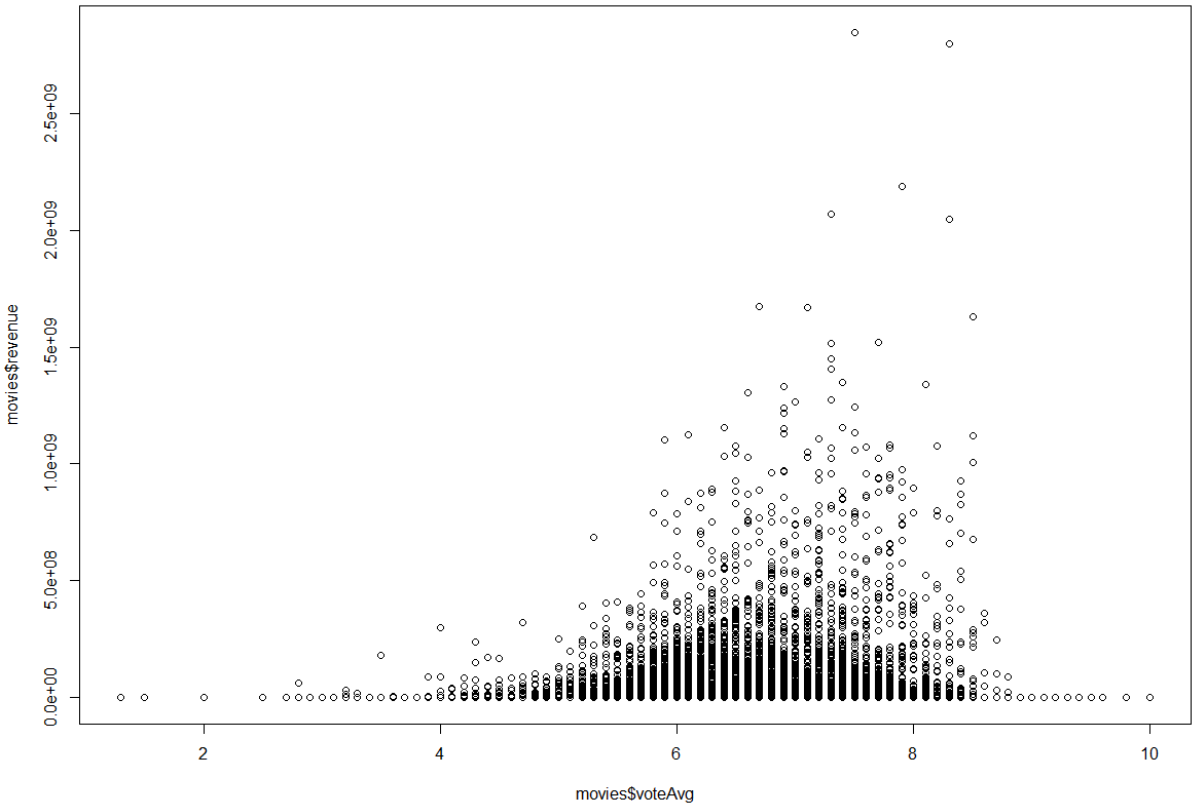
```
> sum(byMonth$month == 1)
[1] 652
> sum(byMonth$month == 2)
[1] 706
> sum(byMonth$month == 3)
[1] 815
> sum(byMonth$month == 4)
[1] 696
> sum(byMonth$month == 5)
[1] 698
> sum(byMonth$month == 6)
[1] 819
> sum(byMonth$month == 7)
[1] 812
> sum(byMonth$month == 8)
[1] 913
> sum(byMonth$month == 9)
[1] 1079
> sum(byMonth$month == 10)
[1] 1068
> sum(byMonth$month == 11)
[1] 807
> sum(byMonth$month == 12)
[1] 935
```

```
> sum(byMonth$month == 1)/12
[1] 54.33333
> sum(byMonth$month == 2)/12
[1] 58.83333
> sum(byMonth$month == 3)/12
[1] 67.91667
> sum(byMonth$month == 4)/12
[1] 58
> sum(byMonth$month == 5)/12
[1] 58.16667
> sum(byMonth$month == 6)/12
[1] 68.25
> sum(byMonth$month == 7)/12
[1] 67.66667
> sum(byMonth$month == 8)/12
[1] 76.08333
> sum(byMonth$month == 9)/12
[1] 89.91667
> sum(byMonth$month == 10)/12
[1] 89
> sum(byMonth$month == 11)/12
[1] 67.25
> sum(byMonth$month == 12)/12
[1] 77.91667
```

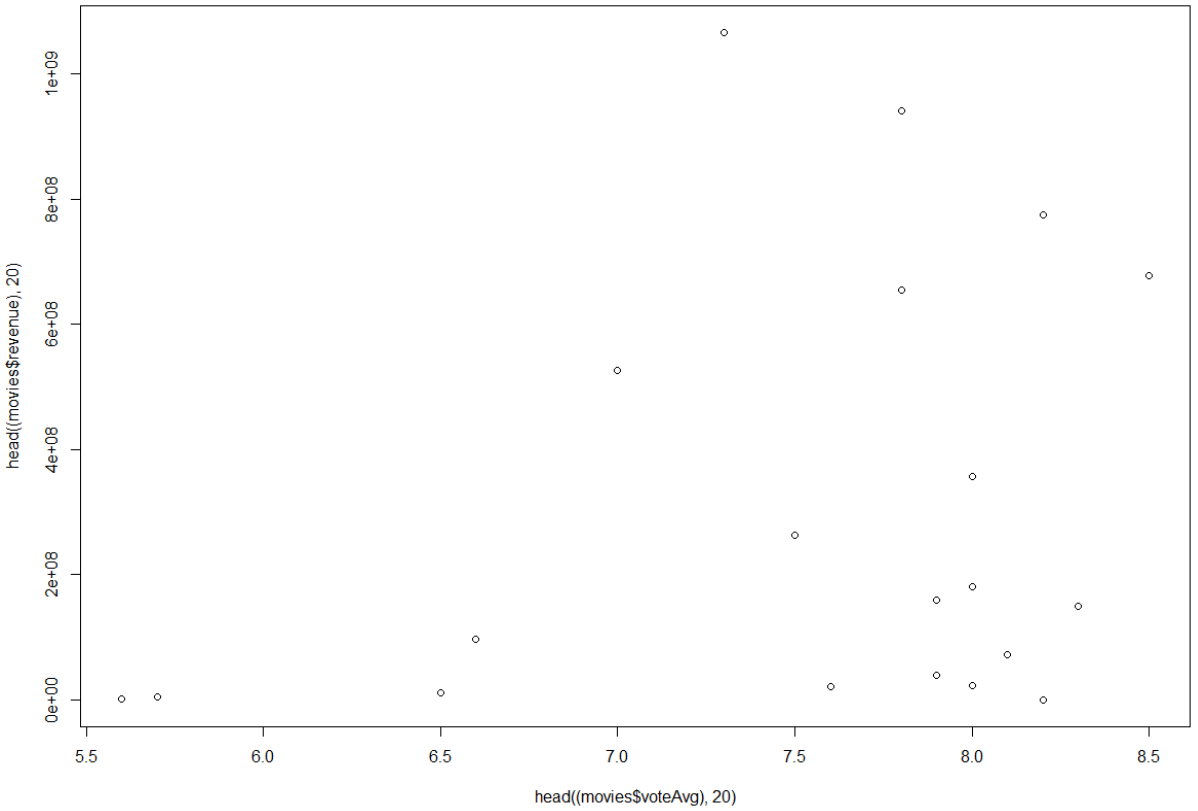
El promedio esperado de películas por cada mes es observado en la división del total de películas de cada mes por 12.

La cantidad de películas que salen por promedio en un mes son 833.3333333333333.

4.14. (7 puntos) ¿Cómo se correlacionan las calificaciones con el éxito comercial?



Utilizando este diagrama de dispersión podemos observar que conforme el voteAvg (es decir la recepción de la película) aumenta, igualmente aumenta el revenue (éxito comercial). Podemos afirmar una relación directa entre las calificaciones con el éxito comercial.



Al observar las mayores 20 películas podemos observar la relación que se afirmó más claramente.

4.15. (5 puntos) ¿A qué género principal pertenecen las películas más largas?

En la base de datos existen dos géneros predominantes que son los que más tiempo en pantalla tienen, estos son documentales y drama. El más largo es el de documentales y esto puede ser debido a que el contenido de los documentales tiende a ser muchisimo mas largo que el de una película normal.

9347	Documentary
5358	Documentary
3885	Drama|History|War
962	Drama|History
1263	Drama|History|Romance
7065	Action|Crime|Thriller
1948	Drama
9686	Action|Adventure|Fantasy|Science Fiction
3740	Documentary
5592	Action|Drama

5. (¡10 puntos extras!) Genere usted otras seis preguntas que le parezcan interesantes porque le
permitan realizar otras exploraciones y respóndalas. No puede repetir ninguna de las instrucciones anteriores.