Universidad del Valle de Guatemala
Minería de Datos sección 10
Grupo # 8

Hoja de trabajo 1
Análisis exploratorio

Autores:
Pablo Noack 17596
Axel López 20768
Kevin Macario 17369

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

Solución:

```
summary(data)
```

```
##        id            imdb_id            popularity           budget
##  Min.   :     5   Length:10866       Min.   : 0.00006   Min.   :        0
##  1st Qu.: 10596   Class :character   1st Qu.: 0.20758   1st Qu.:        0
##  Median : 20669   Mode  :character   Median : 0.38386   Median :        0
##  Mean   : 66064                      Mean   : 0.64644   Mean   : 14625701
##  3rd Qu.: 75610                      3rd Qu.: 0.71382   3rd Qu.: 15000000
##  Max.   :417859                      Max.   :32.98576   Max.   :425000000
##     revenue          original_title         cast              homepage
##  Min.   :0.000e+00   Length:10866       Length:10866       Length:10866
##  1st Qu.:0.000e+00   Class :character   Class :character   Class
:character
##  Median :0.000e+00   Mode  :character   Mode  :character   Mode
:character
##  Mean   :3.982e+07
##  3rd Qu.:2.400e+07
##  Max.   :2.782e+09
##    director           tagline            keywords           overview
##  Length:10866       Length:10866       Length:10866       Length:10866
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     runtime         genres          production_companies release_date
##  Min.   :  0.0   Length:10866       Length:10866         Length:10866
##  1st Qu.: 90.0   Class :character   Class :character     Class :character
##  Median : 99.0   Mode  :character   Mode  :character     Mode  :character
##  Mean   :102.1
##  3rd Qu.:111.0
##  Max.   :900.0
##    vote_count      vote_average    release_year    budget_adj
##  Min.   :  10.0   Min.   :1.500   Min.   :1960   Min.   :        0
##  1st Qu.:  17.0   1st Qu.:5.400   1st Qu.:1995   1st Qu.:        0
##  Median :  38.0   Median :6.000   Median :2006   Median :        0
##  Mean   : 217.4   Mean   :5.975   Mean   :2001   Mean   : 17551040
##  3rd Qu.: 145.8   3rd Qu.:6.600   3rd Qu.:2011   3rd Qu.: 20853251
##  Max.   :9767.0   Max.   :9.200   Max.   :2015   Max.   :425000000
##   revenue_adj
##  Min.   :0.000e+00
##  1st Qu.:0.000e+00
```

```
##  Median :0.000e+00
##  Mean   :5.136e+07
##  3rd Qu.:3.370e+07
##  Max.   :2.827e+09
```

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

Solucion

```
str(data)
```

```
## 'data.frame':    10866 obs. of  21 variables:
##  $ id                    : int  135397 76341 262500 140607 168259 281957
87101 286217 211672 150540 ...
##  $ imdb_id               : chr  "tt0369610" "tt1392190" "tt2908446"
"tt2488496" ...
##  $ popularity            : num  32.99 28.42 13.11 11.17 9.34 ...
##  $ budget                : int  150000000 150000000 110000000 200000000
190000000 135000000 155000000 108000000 74000000 175000000 ...
##  $ revenue               : num  1.51e+09 3.78e+08 2.95e+08 2.07e+09 1.51e+09
...
##  $ original_title        : chr  "Jurassic World" "Mad Max: Fury Road"
"Insurgent" "Star Wars: The Force Awakens" ...
##  $ cast                  : chr  "Chris Pratt|Bryce Dallas Howard|Irrfan
Khan|Vincent D'Onofrio|Nick Robinson" "Tom Hardy|Charlize Theron|Hugh
Keays-Byrne|Nicholas Hoult|Josh Helman" "Shailene Woodley|Theo James|Kate
Winslet|Ansel Elgort|Miles Teller" "Harrison Ford|Mark Hamill|Carrie
Fisher|Adam Driver|Daisy Ridley" ...
##  $ homepage              : chr  "http://www.jurassicworld.com/"
"http://www.madmaxmovie.com/"
"http://www.thedivergentseries.movie/#insurgent"
"http://www.starwars.com/films/star-wars-episode-vii" ...
##  $ director              : chr  "Colin Trevorrow" "George Miller" "Robert
Schwentke" "J.J. Abrams" ...
##  $ tagline               : chr  "The park is open." "What a Lovely Day."
"One Choice Can Destroy You" "Every generation has a story." ...
##  $ keywords              : chr  "monster|dna|tyrannosaurus
rex|velociraptor|island" "future|chase|post-apocalyptic|dystopia|australia"
"based on novel|revolution|dystopia|sequel|dystopic future"
"android|spaceship|jedi|space opera|3d" ...
##  $ overview              : chr  "Twenty-two years after the events of
Jurassic Park, Isla Nublar now features a fully functioning dinosaur theme"|
__truncated__ "An apocalyptic story set in the furthest reaches of our
planet, in a stark desert landscape where humanity is b"| __truncated__
"Beatrice Prior must confront her inner demons and continue her fight against
a powerful alliance which threaten"| __truncated__ "Thirty years after
defeating the Galactic Empire, Han Solo and his allies face a new threat from
the evil Kylo "| __truncated__ ...
```

```
##  $ runtime             : int  124 120 119 136 137 156 125 141 91 94 ...
##  $ genres              : chr  "Action|Adventure|Science Fiction|Thriller"
"Action|Adventure|Science Fiction|Thriller" "Adventure|Science
Fiction|Thriller" "Action|Adventure|Science Fiction|Fantasy" ...
##  $ production_companies: chr  "Universal Studios|Amblin
Entertainment|Legendary Pictures|Fuji Television Network|Dentsu" "Village
Roadshow Pictures|Kennedy Miller Productions" "Summit
Entertainment|Mandeville Films|Red Wagon Entertainment|NeoReel"
"Lucasfilm|Truenorth Productions|Bad Robot" ...
##  $ release_date        : chr  "6/9/15" "5/13/15" "3/18/15" "12/15/15" ...
##  $ vote_count          : int  5562 6185 2480 5292 2947 3929 2598 4572 2893
3935 ...
##  $ vote_average        : num  6.5 7.1 6.3 7.5 7.3 7.2 5.8 7.6 6.5 8 ...
##  $ release_year        : int  2015 2015 2015 2015 2015 2015 2015 2015 2015
2015 ...
##  $ budget_adj          : num  1.38e+08 1.38e+08 1.01e+08 1.84e+08 1.75e+08
...
##  $ revenue_adj         : num  1.39e+09 3.48e+08 2.72e+08 1.90e+09 1.39e+09

...
```

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

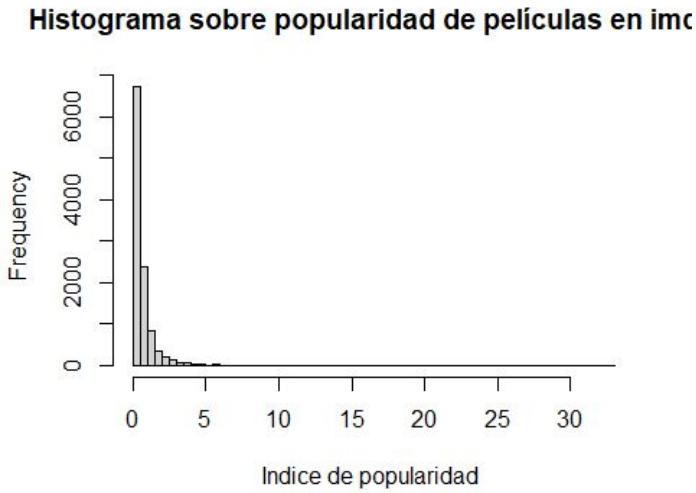
Solución:

Antes que todo, es importante remarcar que las variables id, imbd_id, original_title, cast, tagline y overview, son meras variables cuantitativas que no exhiben ningún comportamiento estadístico. Por ejemplo, si hiciéramos una tabla de frecuencias o un test de normalidad estas variables no exhibieron ningún comportamiento debido a que son etiquetas para cada película.

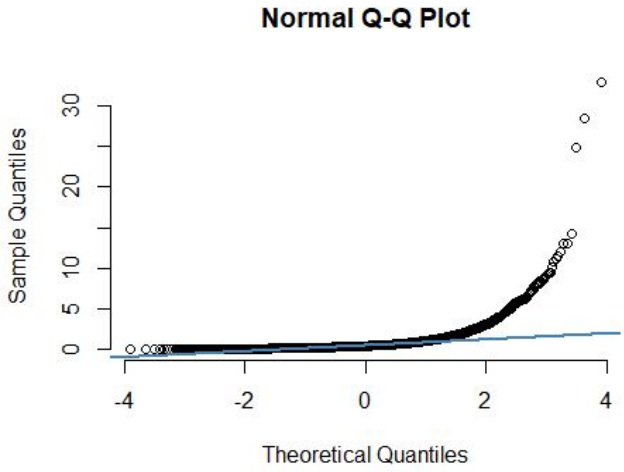
Variables cuantitativas.

Popularidad:

- Histograma: muestra un sesgo hacia el 0, esto implica que la mayoría de películas que salen al cine no son reconocidas y son muy pocas las que triunfan. Por otro lado, no muestra la simetría que cabría esperar de una distribución normal.



- QQnormal: dado ese comportamiento y el sesgo que tienen los datos, se aplicó una gráfica de qqnorm.En la que se observa que los extremos no se comportan de manera normal.

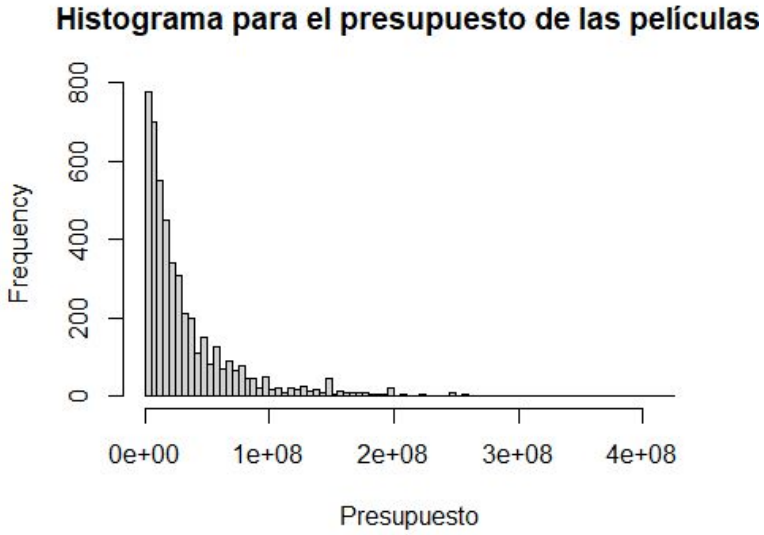


- Test de Lilliefors: tenemos que el valor p es menor a 0.05 con lo que se confirma que la popularidad **No exhibe un comportamiento normal**

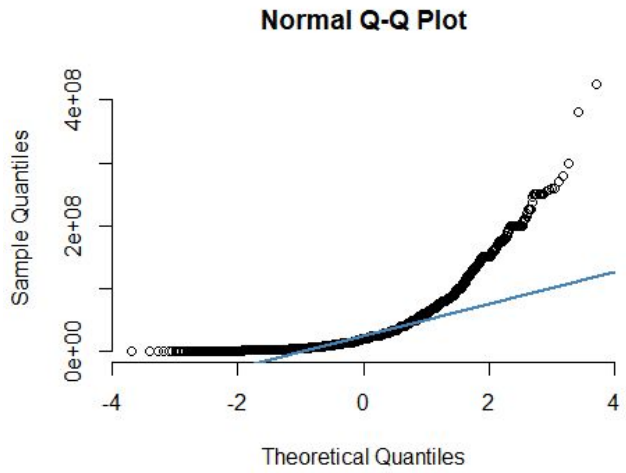
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data$popularity)
## D = 0.25907, p-value < 2.2e-16
```

Presupuesto

Para el estudio de estos datos, se analizaron películas que hayan costado más de $1,000,000 USD debido a que se detectaron 2 escalas distintas en el set de datos. Además de esto no se sabe si la escala fue cambiada alguna otra vez. * Histograma: En el histograma nos damos cuenta que esta no sigue ninguna distribución de normalidad, nos damos cuenta que no posee simetría.



- QQnormal: dado ese comportamiento y el sesgo que tienen los datos, se aplicó una gráfica de qqnorm.En la que se observa que los extremos no se comportan de manera normal.



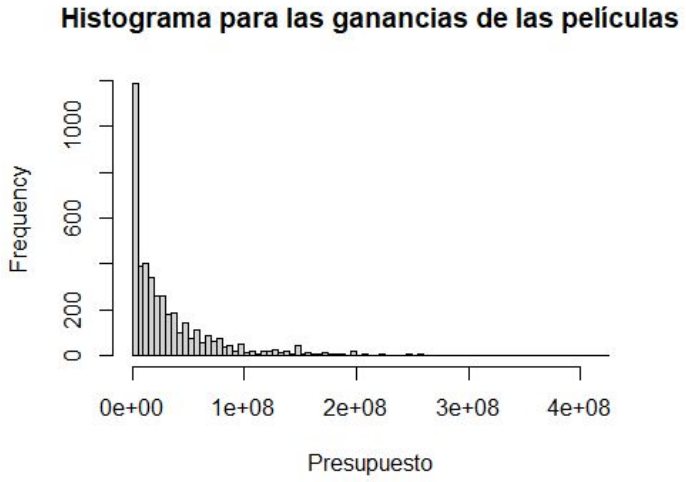
- Test de Lilliefors: tenemos que el valor p es menor a 0.05 con lo que se confirma que la popularidad **No exhibe un comportamiento normal**

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data2$budget)
## D = 0.20528, p-value < 2.2e-16
```

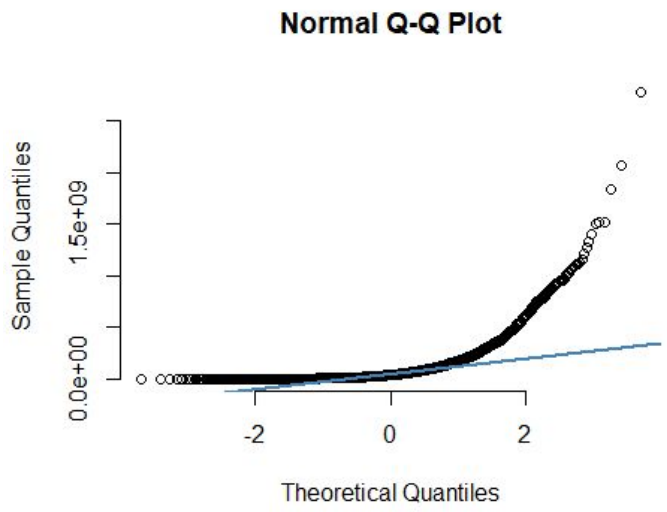
Ganancias

Para el estudio de estos datos, se analizaron películas que hayan ganado más de $1,000,000 USD debido a que se detectaron 2 escalas distintas en el set de datos. Además de esto no se sabe si la escala fue cambiada alguna otra vez.

- Histograma: En el histograma nos damos cuenta que esta no sigue ninguna distribución de normalidad, nos damos cuenta que no posee simetría.

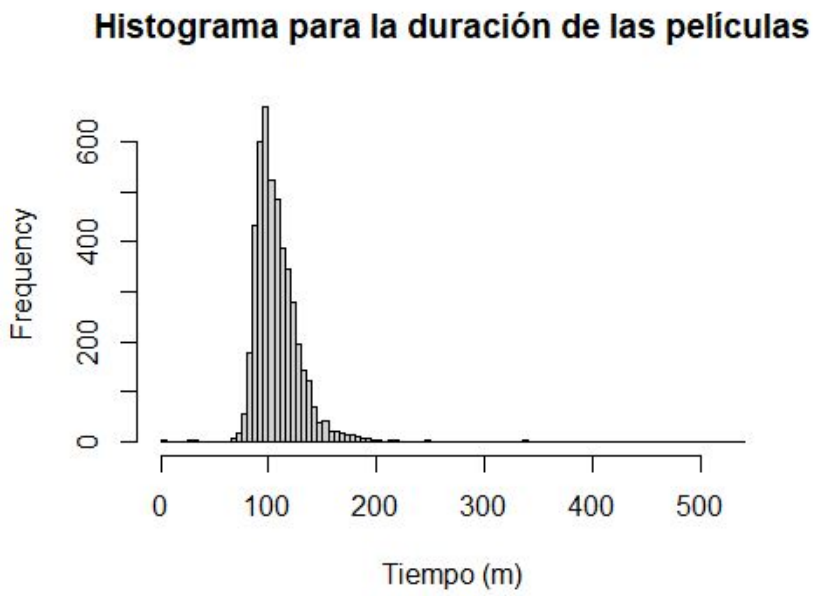


- QQnormal: dado ese comportamiento y el sesgo que tienen los datos, se aplicó una gráfica de qqnorm.En la que se observa que la distribución no se comporta de manera normal.



Duración Para estos datos se utilizaron solamente las películas que duran más de 30 minutos, esto debido a que se observó que la base de datos no se solo películas, sino que también cortos y en ciertos casos se toma como la duración del corto como 0.Es por esto que, con el fin de evitar utilizar datos de películas con longitud de 0 minutos, se utilizó la cota inferior de 30 minutos. De cualquier forma, esto hace que los datos pasen de 10866 a 10625. Por lo que se considera que no afecta al estudio de la población.

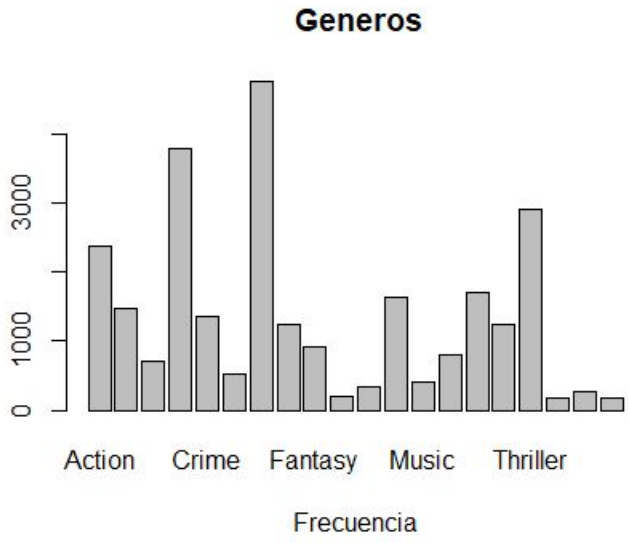
*Histograma: Podemos darnos cuenta que en este caso el tiempo parece comportarse con normalidad, como uno cabría esperar ya que, según nuestra experiencia la mayoría de películas rondan la hora y media. Esto nos lo comprueba el análisis exploratorio del punto 1. Donde el tiempo promedio es de 102 minutos (1 hr y 42 minutos).



Variables cualitativas:

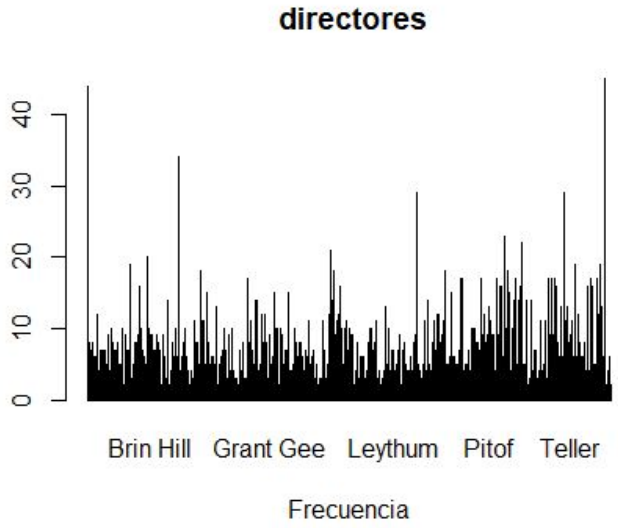
Generos:

- Tabla de frecuencias; en la tabla nos podemos dar cuenta que el programa no es capaz de organizar todos los géneros debido a la gran cantidad que hay. Sin embargo es suficiente para mostrar que los géneros más populares son el drama,acción y thriller.



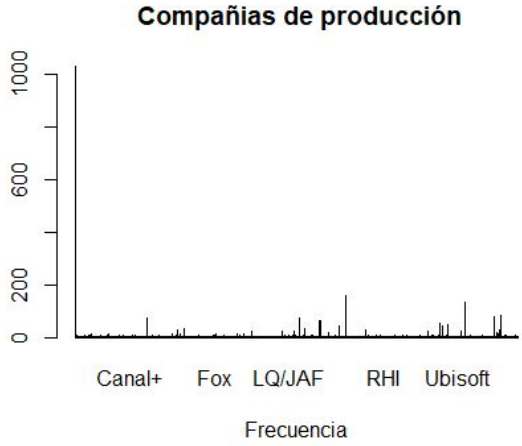
Directores.

- Tabla de frecuencias; en la tabla nos podemos dar cuenta que el programa no es capaz de organizar todos los géneros debido a la gran cantidad que hay. Sin embargo, podemos darnos cuenta que la mayoría de directores no pasan de las 10 películas, además, solamente 3 directores han hecho más de 30 películas y solamente 2



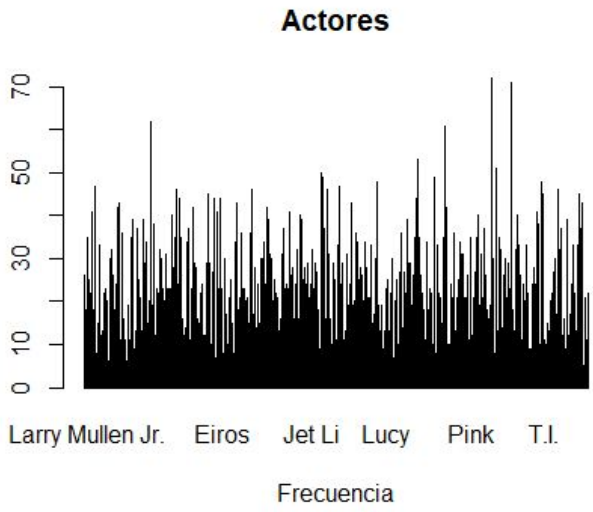
Compañas de producción

- Tabla de frecuencias; en la tabla nos podemos dar cuenta que el programa no es capaz de organizar todas las compañías debido a la gran cantidad que hay. Sin embargo, podemos darnos cuenta que todas las compañías de películas no pasan de las 100 y solamente una llega a superar la marca de las 1000 películas.



Elenco

- Tabla de frecuencias; en la tabla nos podemos dar cuenta que el programa no es capaz de organizar todas las compañías debido a la gran cantidad que hay. Sin embargo, podemos darnos cuenta que son muy pocos los actores que superan las 50 películas. Este es un dato entendible tomando en cuenta que muchos actores hacen una película por año, o bien, hacen varias al mismo tiempo con fechas de estreno en años a futuro.



4. Responda las siguientes preguntas .

4. 1 ¿Cuáles son las películas que costaron más presupuesto?

Las películas con más presupuesto fueron:

```
data5 <- data[order(-data$budget),]
head(data5$original_title,10)
```

```
##  [1] "The Warrior's Way"
##  [2] "Pirates of the Caribbean: On Stranger Tides"
##  [3] "Pirates of the Caribbean: At World's End"
##  [4] "Avengers: Age of Ultron"
##  [5] "Superman Returns"
##  [6] "Tangled"
##  [7] "John Carter"
##  [8] "Spider-Man 3"
##  [9] "The Lone Ranger"
## [10] "The Hobbit: The Battle of the Five Armies"
```

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

Las películas que más recaudaron fueron:

```
data6 <- data[order(-data$revenue),]
head(data6$original_title,10)
```

```
##  [1] "Avatar"
##  [2] "Star Wars: The Force Awakens"
##  [3] "Titanic"
##  [4] "The Avengers"
##  [5] "Jurassic World"
##  [6] "Furious 7"
##  [7] "Avengers: Age of Ultron"
##  [8] "Harry Potter and the Deathly Hallows: Part 2"
##  [9] "Frozen"
## [10] "Iron Man 3"
```

4.3 ¿Cuál es la película que más votos tuvo?

```
maxVote <- movies[order(-movies$vote_count),]
maxVote <- head(maxVote,1)
tMaxVote <- maxVote[c("original_title","vote_count")]
names(tMaxVote) <- c('Pelicula', 'Votos')
tMaxVote
```

```
      Pelicula Votos
1920 Inception  9767
```

4.4 ¿Cuál es la película peor calificada?

```
worstMovie <- movies[order(-movies$vote_average),]
worstMovie <- head(worstMovie,1)
tWorstMovie <- worstMovie[c("original_title","vote_average")]
names(tWorstMovie) <- c('Pelicula', 'Calificacion Promedio')
tWorstMovie
```

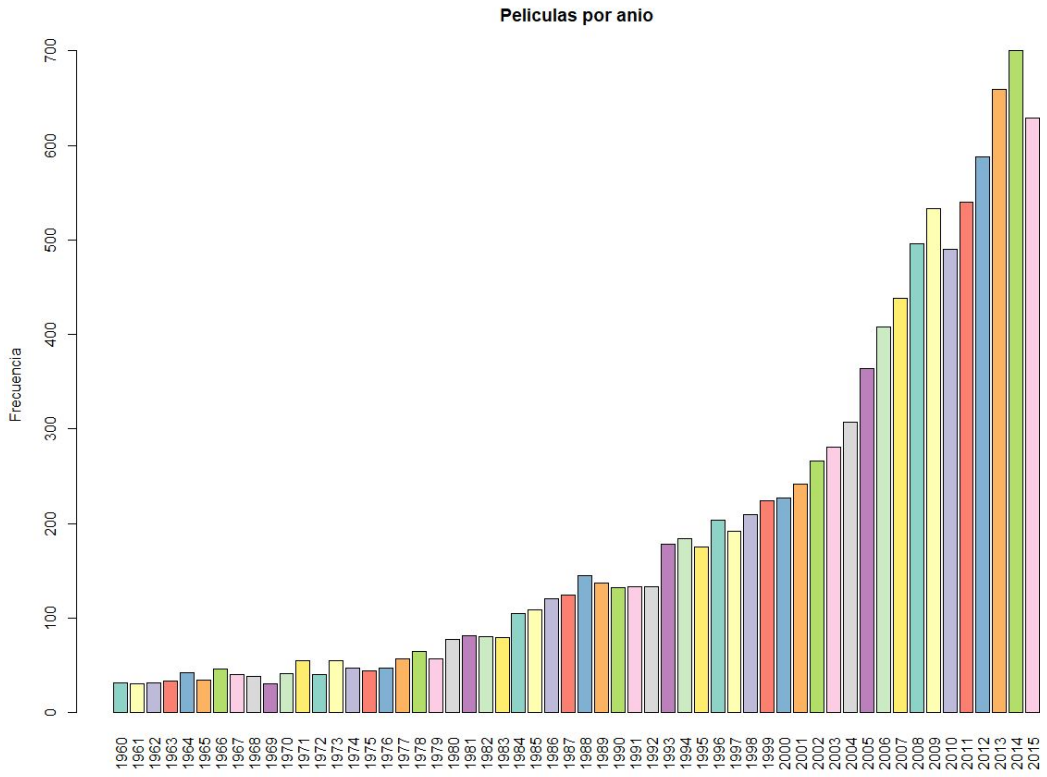
```
                          Pelicula Calificacion Promedio
3895 The Story of Film: An Odyssey                    9.2
```

4.5 ¿Cuántas películas se hicieron cada año?

```
library(RColorBrewer)
col1 <- brewer.pal(12,'Set3')
```

```
realiseYear <- table(movies$release_year)
barplot(realiseYear, ylab = 'Frecuencia', main = 'Peliculas por anio',
col=col1, las=3)
```

Como se puede ver en la siguiente gráfica el año en el que se produjeron más películas fue el año 2014, con 700 películas producidas



4.6 ¿Cuál es el género principal de las 20 películas más populares?

```
popMovies <- mov[order(-mov$popularity ),]
popMovies <- head(popMovies, 20)
tPopMovies <- popMovies[c("popularity", "original_title","primary_genre")]
names(tPopMovies) <- c('Popularidad', 'Pelicuala', 'Genero principal')
tPopMovies
```

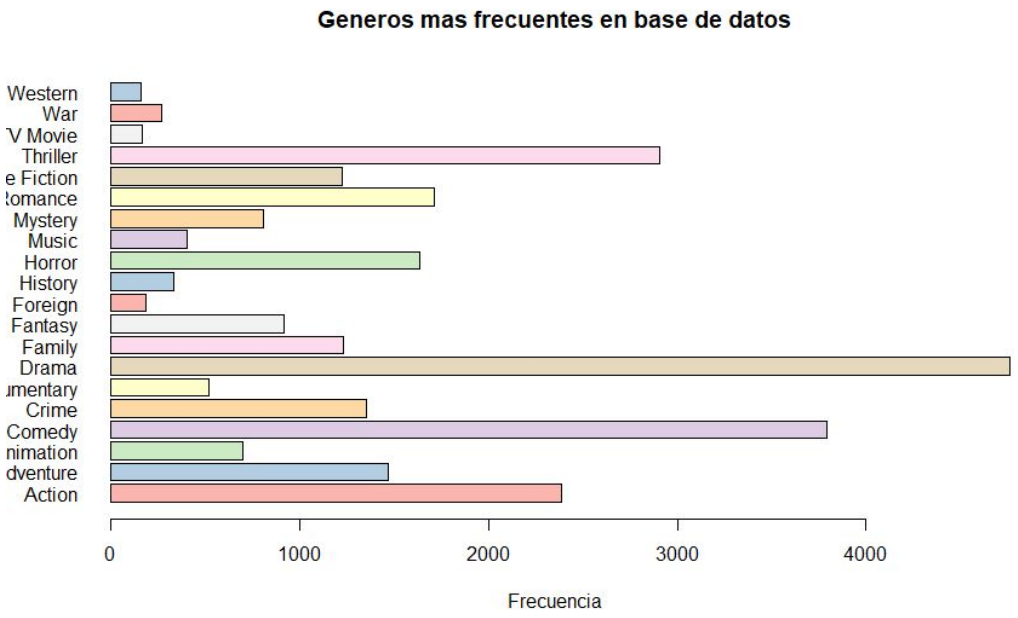
	Popularidad	Peliculala	Genero principal
1	32.985763	Jurassic World	Action
2	28.419936	Mad Max: Fury Road	Action
630	24.949134	Interstellar	Adventure
631	14.311205	Guardians of the Galaxy	Action
3	13.112507	Insurgent	Adventure
632	12.971027	Captain America: The Winter Soldier	Action
1330	12.037933	Star Wars	Adventure
633	11.422751	John Wick	Action
4	11.173104	Star Wars: The Force Awakens	Action
634	10.739009	The Hunger Games: Mockingjay - Part 1	Science
635	10.174599	The Hobbit: The Battle of the Five Armies	Adventure
1387	9.432768	Avatar	Action
1920	9.363643	Inception	Action
5	9.335014	Furious 7	Action
6	9.110700	The Revenant	Western
2410	8.947905	Fight Club	Drama
636	8.691294	Big Hero 6	Adventure
7	8.654359	Terminator Genisys	Science
2634	8.575419	The Lord of the Rings: The Fellowship of the Ring	Adventure
2876	8.466668	The Dark Knight	Drama

4.7 ¿Cuál es el género que predomina en el conjunto de datos?

```
splitPopGenres <- c(movies$genres)
splitPopGenres <- strsplit(splitPopGenres, "|", fixed = TRUE)
splitPopGenres <- unlist(splitPopGenres)
splitPopGenres <- table(splitPopGenres)
splitPopGenres

col3 <- brewer.pal(9,'Pastel1')
barplot(splitPopGenres, horiz = TRUE, xlab = 'Frecuencia', main = 'Generos
mas frecuentes en base de datos', col=col3, las = 1)
```

El género que predomina en la base de datos es el Drama con 4761, seguido de Comedia, con 3792 películas.




```
splitPopGenres
        Action       Adventure       Animation          Comedy           Crime     Documentary           Drama
          2385            1471             699            3793            1355             520            4761
        Family         Fantasy         Foreign         History         Horror           Music         Mystery
          1231             916             188             334            1637             408             810
       Romance Science Fiction        Thriller      TV Movie             War         Western
          1712            1230            2908             167             270             165
```

4.8 ¿Las películas de que genero principal obtuvieron mayores ganancias?

```
mostRevenue <- mov[order(-mov$revenue),]
mostRevenue <- head(mostRevenue, 20)
select(mostRevenue, original_title, revenue, primary_genre)
```

```
                                  original_title    revenue primary_genre
1387                                      Avatar 2781505847        Action
4                      Star Wars: The Force Awakens 2068178225        Action
5232                                     Titanic 1845034188         Drama
4362                                 The Avengers 1519557910       Science
1                                  Jurassic World 1513528810        Action
5                                      Furious 7 1506249360        Action
15                         Avengers: Age of Ultron 1405035767        Action
3375  Harry Potter and the Deathly Hallows: Part 2 1327817822     Adventure
5423                                      Frozen 1274219009     Animation
5426                                    Iron Man 3 1215439994        Action
9                                        Minions 1156730962        Family
3523                  Transformers: Dark of the Moon 1123746996        Action
4950 The Lord of the Rings: The Return of the King 1118888979     Adventure
4366                                     Skyfall 1108561013        Action
8095                                     The Net 1106279658         Crime
4364                           The Dark Knight Rises 1081041287        Action
6556    Pirates of the Caribbean: Dead Man's Chest 1065659812     Adventure
1931                                  Toy Story 3 1063171911     Animation
1922                            Alice in Wonderland 1025467110        Family
3376   Pirates of the Caribbean: On Stranger Tides 1021683000     Adventure
```

4.9 ¿Las películas de que genero principal necesitaron más presupuesto?

```
budget <- mov[order(-mov$budget),]
budget <- head(budget, 20)
select(budget, original_title, primary_genre, budget)
```

```
                                  original_title primary_genre    budget
2245                             The Warrior's Way     Adventure 425000000
3376  Pirates of the Caribbean: On Stranger Tides     Adventure 380000000
7388     Pirates of the Caribbean: At World's End     Adventure 300000000
15                         Avengers: Age of Ultron        Action 280000000
6571                              Superman Returns     Adventure 270000000
1930                                       Tangled     Animation 260000000
4412                                    John Carter        Action 260000000
7395                                  Spider-Man 3       Fantasy 258000000
5509                                The Lone Ranger        Action 255000000
635     The Hobbit: The Battle of the Five Armies     Adventure 250000000
644                     X-Men: Days of Future Past        Action 250000000
1390       Harry Potter and the Half-Blood Prince     Adventure 250000000
1924 Harry Potter and the Deathly Hallows: Part 1     Adventure 250000000
4364                           The Dark Knight Rises        Action 250000000
4368             The Hobbit: An Unexpected Journey     Adventure 250000000
5432          The Hobbit: The Desolation of Smaug     Adventure 250000000
11                                         Spectre        Action 245000000
1387                                        Avatar        Action 237000000
2903     The Chronicles of Narnia: Prince Caspian     Adventure 225000000
4382                                Men in Black 3        Action 225000000
```

4.10 ¿Quiénes son los 20 mejores directores que hicieron películas altamente calificadas?

Los 20 directores con películas altamente calificadas son:

```
directores <- movies[!is.na(movies$director), ]
directores <-  directores[!directores$director == "",]
directores <- directores[order(-directores$vote_average), ]
directores <- head(directores, 20)
tabla_directores <- directores[c("director", "vote_average")]
names(tabla_directores) <- c("Director", "Calificación")
```

```
                                    Director Calificación
3895                            Mark Cousins          9.2
539                      Jennifer Siebel Newsom          8.9
1201                              Carl Tibbetts          8.8
2270                           Derek Frankowski          8.8
6912                               David Mallet          8.7
3691                                Curt Morgan          8.5
5831                                James Payne          8.5
8222       Martin Scorsese|Michael Henry Wilson          8.5
8412                               Saul Swimmer          8.5
8840                                Stan Lathan          8.5
```

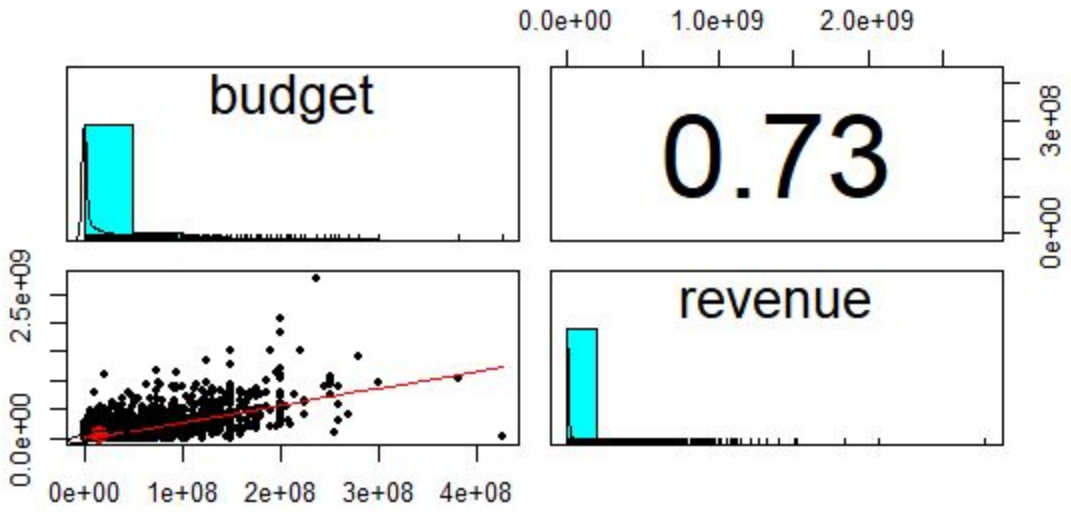
```
610                                  Andrew Jarecki          8.4
2335                       Sam Dunn|Scot McFadyen          8.4
4179                               Frank Darabont          8.4
5924                              Anthony Mandler          8.4
5987                    Jorge RamÃƒÂrez SuÃƒÂ¡rez         8.4
7949                               Jonathan Demme          8.4
8371                                  Chris Bould          8.4
9291 D.A. Pennebaker|David Dawkins|Chris Hegedus          8.4
1323                                Paul Dugdale          8.3
1865                       Sam Dunn|Scot McFadyen          8.3
```

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

La relación entre las variables Budget(presupuesto) y Revenue(Ingresos) es significativa ya que tiene un coeficiente de correlación de 0.73, lo que quiere decir que no necesariamente un gran presupuesto supone grandes ingresos, pero es común que esto suceda.

```
correlacion <- movies[c("budget", "revenue")]
library(psych)
pairs.panels(correlacion)
```

Diagrama de dispersión y correlación entre las variables Presupuesto e Ingresos



4.12 ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?

La relación entre las variables Month(mes) y Revenue(Ingresos) es nula. Significa que no existe ninguna relación entre las variables, el coeficiente de correlación de 0.01, lo que confirma con más exactitud la conclusión.

```
library(lubridate)
movies$month <- vapply(strsplit(movies$release_date, "/"), `[`, 1,
FUN.VALUE=character(1))
asociacion <- movies[c("month", "revenue")]
pairs.panels(asociacion)
```

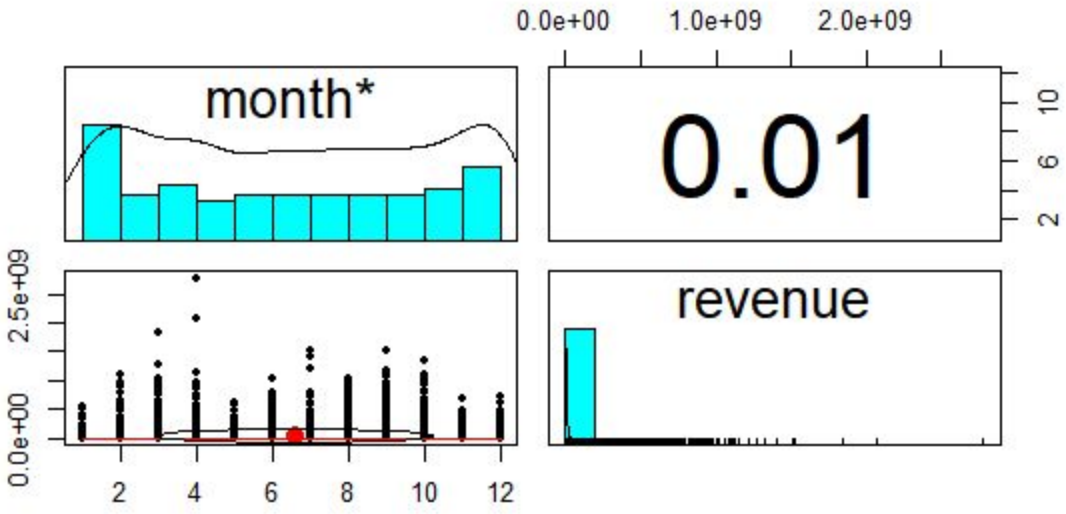


Diagrama de dispersión y correlación entre las variables Mes e Ingresos

4.13 ¿En qué meses se han visto los lanzamientos máximos?

Se puede observar que el mes con más lanzamientos es septiembre, seguido de octubre y diciembre.

```
library(plyr)
lanzamientos <- count(movies, "month")
lanzamientos <- lanzamientos[order(-lanzamientos$freq), ]
names(lanzamientos) <- c("Mes", "Lanzamientos")
```

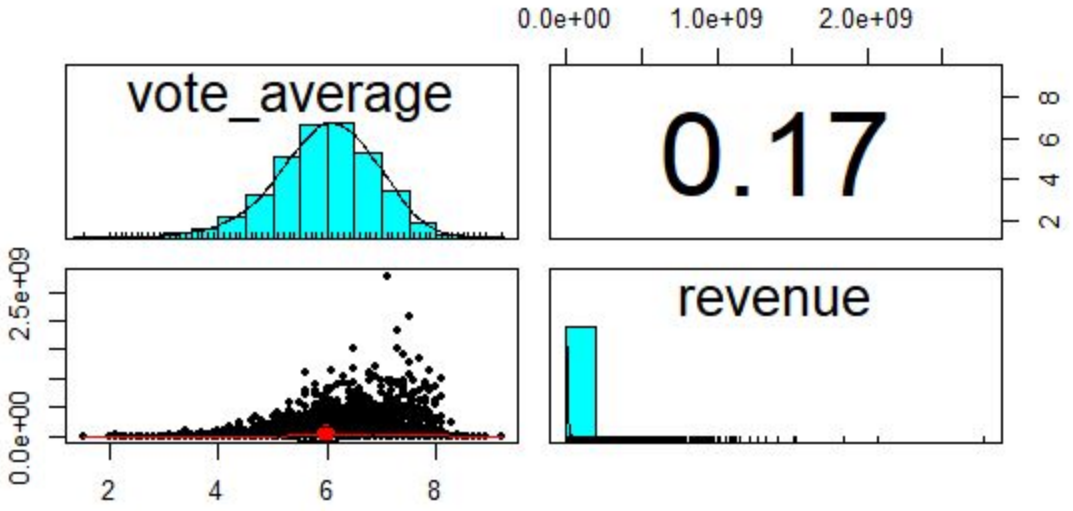
```
   Mes Lanzamientos
12   9         1331
2   10         1153
4   12          985
1    1          919
11   8          918
9    6          827
6    3          823
3   11          814
8    5          809
10   7          799
7    4          797
5    2          691
```

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

En este caso existe una relación muy débil, ya que como se puede observar el coeficiente es de 0.17 lo que quiere decir que en muy contadas ocasiones estas variables se tienen relación con la otra.

```
correlacion1 <- movies[c("vote_average", "revenue")]
library(psych)
pairs.panels(correlacion1)
```

Diagrama de dispersión y correlación entre las variables Votos de la Plataforma e Ingresos



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

Se puede concluir que el género con más duración es el Drama, ya que en general todas es el Drama, con una duración 79537 minutos.

```
duracion <- aggregate(movies$runtime, by=list(Genero=movies$genres), sum)
duracion <- duracion[order(-duracion$x), ]
names(duracion) <- c("Genero", "Duración")
```

```
                                          Genero Duración
986                                        Drama    79537
688                                       Comedy    66346
1197                                Drama|Romance    32671
959                                   Documentary    32628
734                                 Comedy|Drama    28727
849                                Comedy|Romance    26918
1548                               Horror|Thriller    24054
768                         Comedy|Drama|Romance    23227
1480                                       Horror    22550
1228                               Drama|Thriller    14694
329                                Action|Thriller    10250
779                                 Comedy|Family     9543
1040                                Drama|Comedy     9543
1801                                    Thriller     9216
 [ reached 'max' / getOption("max.print") -- omitted 1540 rows ]
```