Entrega Final

Andrés Mazariegos, Daniel Sarmiento

2026-02-02

#GITHUB

#https://github.com/andresm220/Lab1Mineria

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(stringr)
library(tinytex)
```

Andres Mazariegos\ y Daniel Sarmiento

1.- Haga una exploración rápida de sus datos

```
##Leemos la data y lo asignamos a una variable
movies <- read.csv("movies_2026.csv",fileEncoding = "UTF-8-BOM")
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec,
## : invalid input found on input connection 'movies_2026.csv'
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec,
## : EOF within quoted string
```

```
##el nombre de las columnas
colnames(movies )
```

```
##  [1] "id"                        "budget"
##  [3] "genres"                    "homePage"
##  [5] "productionCompany"         "productionCompanyCountry"
##  [7] "productionCountry"         "revenue"
##  [9] "runtime"                   "video"
## [11] "director"                  "actors"
## [13] "actorsPopularity"          "actorsCharacter"
## [15] "originalTitle"             "title"
## [17] "originalLanguage"          "popularity"
## [19] "releaseDate"               "voteAvg"
## [21] "voteCount"                 "genresAmount"
## [23] "productionCoAmount"        "productionCountriesAmount"
## [25] "actorsAmount"              "castWomenAmount"
## [27] "castMenAmount"             "releaseYear"
```

```
##summary
summary(movies)
```

```
##        id              budget             genres            homePage
##  Min.   :      5   Min.   :        0   Length:9892        Length:9892
##  1st Qu.:1570504   1st Qu.:        0   Class :character   Class :character
##  Median :1589780   Median :        0   Mode  :character   Mode  :character
##  Mean   :1561381   Mean   :   169119
##  3rd Qu.:1604364   3rd Qu.:        0
##  Max.   :1627166   Max.   :350000000
##
##  productionCompany  productionCompanyCountry productionCountry
##  Length:9892        Length:9892              Length:9892
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##     revenue             runtime         video           director
##  Min.   :0.000e+00   Min.   :  0.00   Mode :logical   Length:9892
##  1st Qu.:0.000e+00   1st Qu.:  0.00   FALSE:9892      Class :character
##  Median :0.000e+00   Median : 10.00                   Mode  :character
##  Mean   :5.168e+05   Mean   : 31.56
##  3rd Qu.:0.000e+00   3rd Qu.: 57.25
##  Max.   :1.744e+09   Max.   :675.00
##
##     actors          actorsPopularity   actorsCharacter    originalTitle
##  Length:9892        Length:9892        Length:9892        Length:9892
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##     title           originalLanguage     popularity       releaseDate
##  Length:9892        Length:9892        Min.   :  0.0000   Length:9892
##  Class :character   Class :character   1st Qu.:  0.0214   Class :character
##  Mode  :character   Mode  :character   Median :  0.0541   Mode  :character
##                                        Mean   :  0.8341
##                                        3rd Qu.:  0.1448
##                                        Max.   :822.1075
##                                        NA's   :1
##     voteAvg         voteCount         genresAmount   productionCoAmount
##  Min.   : 0.000   Min.   :   0.000   Min.   :0.000   Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.000   1st Qu.:1.000   1st Qu.: 0.0000
##  Median : 0.000   Median :   0.000   Median :1.000   Median : 0.0000
##  Mean   : 1.163   Mean   :   1.891   Mean   :1.295   Mean   : 0.7606
##  3rd Qu.: 0.000   3rd Qu.:   0.000   3rd Qu.:2.000   3rd Qu.: 1.0000
##  Max.   :10.000   Max.   :2077.000   Max.   :9.000   Max.   :14.0000
##  NA's   :1        NA's   :1          NA's   :1       NA's   :1
##  productionCountriesAmount  actorsAmount    castWomenAmount   castMenAmount
##  Min.   :0.0000            Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:0.0000            1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median :1.0000            Median : 3.000   Median : 0.0000   Median : 0.000
##  Mean   :0.7037            Mean   : 3.823   Mean   : 0.6257   Mean   : 1.005
##  3rd Qu.:1.0000            3rd Qu.: 6.000   3rd Qu.: 1.0000   3rd Qu.: 1.000
##  Max.   :6.0000            Max.   :25.000   Max.   :15.0000   Max.   :10.000
##  NA's   :1                 NA's   :1        NA's   :1         NA's   :1
##   releaseYear
##  Min.   :1995
##  1st Qu.:2025
##  Median :2025
##  Mean   :2025
##  3rd Qu.:2026
##  Max.   :2026
##  NA's   :3
```

```
str(movies)
```

```
## 'data.frame':    9892 obs. of  28 variables:
##  $ id                         : int  1627085 1626914 1626898 1626808 1626678 1626234 1626010 1625551 1
##  $ budget                     : num  0 0 0 0 0 0 1 0 0 0 0 ...
##  $ genres                     : chr  "Drama|Crime" "Animation" "Animation" "Thriller|Mystery|Documentar
##  $ homePage                   : chr  "" "" "" "" ...
##  $ productionCompany          : chr  "" "" "" "" ...
##  $ productionCompanyCountry   : chr  "" "" "" "" ...
##  $ productionCountry          : chr  "" "" "" "" ...
##  $ revenue                    : num  0 0 0 0 0 0 1 0 0 0 0 ...
##  $ runtime                    : int  95 3 2 5 12 14 39 90 96 106 ...
##  $ video                      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ director                   : chr  "Javad Hakami" "Kimmy Gatewood" "Kimmy Gatewood" "Felipe Roldán"
##  $ actors                     : chr  "Mohsen Ghasabian|Aida Mahiani|Mehran Ghafourian|Payam Ahmadinia|M
##  $ actorsPopularity           : chr  "0.3453|0.1664|0.9684|0.3437|0.3713|0.2437|0.2796|0.2639" "0|0.007
##  $ actorsCharacter            : chr  "|||||||" "Prince Charming|Evil Stepmother|Fairy Godmother|Cinderel
##  $ originalTitle              : chr    " "  "Cinderella" "Aladdin" "EL ANILLO Y EL DECK" ...
```

```
##  $ title                     : chr  "Immersed" "Cinderella" "Aladdin" "THE RING AND THE DECK" ...
##  $ originalLanguage          : chr  "fa" "en" "en" "es" ...
##  $ popularity                : num  0.0357 0.0357 0.0214 0.0429 0.0379 ...
##  $ releaseDate               : chr  "2026-02-01" "2026-02-01" "2026-02-01" "2026-02-01" ...
##  $ voteAvg                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ voteCount                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ genresAmount              : int  2 1 1 3 1 1 1 1 3 1 ...
##  $ productionCoAmount        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ productionCountriesAmount: int  0 0 0 0 0 0 0 1 1 0 ...
##  $ actorsAmount              : int  8 4 3 7 3 3 5 4 5 5 ...
##  $ castWomenAmount           : int  2 0 0 0 0 0 0 3 1 2 ...
##  $ castMenAmount             : int  5 0 0 0 0 0 3 0 3 3 ...
##  $ releaseYear               : int  2026 2026 2026 2026 2026 2026 2026 2026 2026 2026 ...
```

2.- Diga el tipo de cada una de las variables

```r
clasificacion_variables <- data.frame(
  Variable = c(
    "id", "budget", "genres", "homePage", "productionCompany",
    "productionCompanyCountry", "productionCountry", "revenue", "runtime", "video",
    "director", "actors", "actorsPopularity", "actorsCharacter", "originalTitle",
    "title", "originalLanguage", "popularity", "releaseDate", "voteAvg",
    "voteCount", "genresAmount", "productionCoAmount", "productionCountriesAmount", "actorsAmount",
    "castWomenAmount", "castMenAmount", "releaseYear"
  ),
  Tipo = c(
    "Cualitativa categórica (ID)", "Cuantitativa continua", "Cualitativa categórica", "Cualitativa categ
    "Cualitativa categórica", "Cualitativa categórica", "Cuantitativa continua", "Cuantitativa continua
    "Cualitativa categórica", "Cualitativa categórica", "Cualitativa categórica", "Cualitativa categóri
    "Cualitativa categórica", "Cualitativa categórica", "Cuantitativa continua", "Cualitativa categóric
    "Cuantitativa discreta", "Cuantitativa discreta", "Cuantitativa discreta", "Cuantitativa discreta",
    "Cuantitativa discreta", "Cuantitativa discreta", "Cuantitativa discreta"
  ),
  stringsAsFactors = FALSE
)

# Para visualizar la tabla completa
print(clasificacion_variables)
```

```
##                     Variable                              Tipo
## 1                         id      Cualitativa categórica (ID)
## 2                     budget            Cuantitativa continua
## 3                     genres          Cualitativa categórica
## 4                   homePage          Cualitativa categórica
## 5           productionCompany          Cualitativa categórica
## 6    productionCompanyCountry          Cualitativa categórica
## 7           productionCountry          Cualitativa categórica
## 8                     revenue            Cuantitativa continua
## 9                     runtime            Cuantitativa continua
## 10                      video Cualitativa categórica (binaria)
## 11                   director          Cualitativa categórica
## 12                     actors          Cualitativa categórica
## 13           actorsPopularity          Cualitativa categórica
```

```
## 14           actorsCharacter           Cualitativa categórica
## 15             originalTitle           Cualitativa categórica
## 16                     title           Cualitativa categórica
## 17          originalLanguage           Cualitativa categórica
## 18                popularity            Cuantitativa continua
## 19               releaseDate           Cualitativa categórica
## 20                   voteAvg            Cuantitativa continua
## 21                 voteCount            Cuantitativa discreta
## 22              genresAmount            Cuantitativa discreta
## 23        productionCoAmount            Cuantitativa discreta
## 24 productionCountriesAmount            Cuantitativa discreta
## 25              actorsAmount            Cuantitativa discreta
## 26           castWomenAmount            Cuantitativa discreta
## 27             castMenAmount            Cuantitativa discreta
## 28               releaseYear            Cuantitativa discreta
```

3. - ¿Siguen una distribución normal ? \ Tablas de frecuencia y explicación

a)Variables Cuantitativas

```r
# 1) Seleccionar únicamente las columnas numéricas (cuantitativas)
# sapply(movies_2026, is.numeric) devuelve TRUE/FALSE por columna
# y con eso filtramos las columnas numéricas.
vars_num <- movies[, sapply(movies, is.numeric)]

# 2) Crear una tabla donde guardaremos los resultados
# variable: nombre de la columna
# n: cantidad de valores NO-NA (datos válidos) que tiene esa variable
# p_value: valor p del test de Shapiro-Wilk
# normal: TRUE si p_value > 0.05, FALSE si p_value <= 0.05
normalidad <- data.frame(
  variable = names(vars_num),
  n = NA,
  p_value = NA,
  normal = NA
)

# 3) Fijar una semilla para reproducibilidad
# Esto hace que sample() seleccione siempre la misma muestra si vuelves a correr el script.
set.seed(1)

# 4) Recorrer cada variable numérica y aplicar el test
for (i in seq_along(vars_num)) {

  # 4.1) Tomar la columna i y quitar NAs
  # na.omit elimina valores faltantes para evitar errores en la prueba
  x <- na.omit(vars_num[[i]])

  # Guardar cuántos datos válidos tiene la variable (antes de muestrear)
  normalidad$n[i] <- length(x)

  # 4.2) Validaciones mínimas antes de correr Shapiro
  # - Shapiro necesita al menos 3 observaciones
  if (length(x) < 3) next
```

```r
  # - Si todos los valores son iguales (varianza = 0), no tiene sentido probar normalidad
  #   porque no hay distribución "real" que evaluar
  if (length(unique(x)) < 2) next

  # 4.3) Shapiro-Wilk solo permite hasta 5000 observaciones
  # Si hay más, tomamos una muestra aleatoria de 5000 para poder correr el test
  if (length(x) > 5000) {
    x <- sample(x, 5000)
  }

  # 4.4) Ejecutar Shapiro de forma segura
  # tryCatch evita que el script se detenga si una variable causa error
  out <- tryCatch(
    shapiro.test(x),
    error = function(e) NULL
  )

  # 4.5) Si la prueba se ejecutó bien, guardamos resultados
  if (!is.null(out)) {

    # Guardar valor p del test
    normalidad$p_value[i] <- out$p.value

    # Decisión simple:
    # p_value > 0.05  -> "compatible con normalidad" (no rechazo H0)
    # p_value <= 0.05 -> "no normal" (rechazo H0)
    normalidad$normal[i] <- out$p.value > 0.05
  }
}

# 5) Mostrar la tabla final con resultados de normalidad
normalidad
```

```
##                      variable    n      p_value normal
## 1                          id 9892 4.972400e-80  FALSE
## 2                      budget 9892 1.760313e-95  FALSE
## 3                     revenue 9892 1.283294e-95  FALSE
## 4                     runtime 9892 2.625955e-66  FALSE
## 5                  popularity 9891 2.617630e-95  FALSE
## 6                     voteAvg 9891 5.123951e-82  FALSE
## 7                   voteCount 9891 3.311392e-95  FALSE
## 8                genresAmount 9891 7.949892e-55  FALSE
## 9          productionCoAmount 9891 5.747099e-72  FALSE
## 10 productionCountriesAmount 9891 7.377126e-69  FALSE
## 11               actorsAmount 9891 2.838120e-54  FALSE
## 12           castWomenAmount 9891 4.557698e-79  FALSE
## 13             castMenAmount 9891 4.754043e-75  FALSE
## 14              releaseYear 9889 3.564811e-78  FALSE
```

Para evaluar si las variables cuantitativas del conjunto de datos siguen una distribución normal, se aplicó la prueba de Shapiro–Wilk a cada una de ellas. Dado el tamaño de la muestra, se utilizó una submuestra aleatoria de 5000 valores, garantizando la reproducibilidad del análisis.

Los resultados indican que todas las variables cuantitativas no siguen una distribución normal (p 0.05). Este comportamiento es especialmente evidente en las variables relacionadas con conteos y montos económicos, las cuales presentan distribuciones asimétricas con colas largas hacia la derecha. Algunas variables continuas, como la duración de las películas o el promedio de votos, muestran un comportamiento más cercano a la normalidad, aunque no completamente normal debido a la presencia de valores extremos y a la alta variabilidad de los datos.

Debido a su naturaleza discreta y asimétrica, no es apropiado asumir normalidad para las variables de conteo, por lo que en análisis posteriores es recomendable emplear métodos no paramétricos o transformaciones adecuadas de los datos.

b)Variables Cualitativas

```
### 3b. Tablas de Frecuencia (Variables Cualitativas según clasificación)

# 1. id (Cualitativa categórica ID) - Mostramos solo las primeras 10 para verificar
cat("\nVariable: id (Top 10)\n")
```

```
##
## Variable: id (Top 10)
```

```
print(head(sort(table(movies$id), decreasing = TRUE), 10))
```

```
##
##      5      6  83533 340374 424853 445466 483766 511243 548275 559547
##      1      1      1      1      1      1      1      1      1      1
```

```
# 2. genres
cat("\nVariable: genres (Top 10 combinaciones)\n")
```

```
##
## Variable: genres (Top 10 combinaciones)
```

```
print(head(sort(table(movies$genres), decreasing = TRUE), 10))
```

```
##
##               Documentary        Drama       Comedy       Horror    Animation
##         1912          1654         1185          577          273          272
##        Music Comedy|Drama Drama|Comedy     Thriller
##          190          120          118          104
```

```
# 3. homePage
cat("\nVariable: homePage (Top 10)\n")
```

```
##
## Variable: homePage (Top 10)
```

```
print(head(sort(table(movies$homePage), decreasing = TRUE), 10))
```

```
##
##
##                                                    8476
##                   https://watch.njpwworld.com/details/60769
##                                                      16
##                                     https://tractorted.com/
##                                                       6
##                                   http://www.avikomfilm.com
##                                                       4
##                                      http://postbellek.com
##                                                       3
##                                      http://www.512red.com
##                                                       3
##                                       https://aca-mma.com/
##                                                       3
## https://brothersduelproduc.wixsite.com/brothers-duel-produc
##                                                       3
##                           https://liampboulay.wordpress.com/
##                                                       3
##                                      http://dracofilms.com
##                                                       2
```

```
# 4. productionCompany
cat("\nVariable: productionCompany (Top 10)\n")
```

```
##
## Variable: productionCompany (Top 10)
```

```
print(head(sort(table(movies$productionCompany), decreasing = TRUE), 10))
```

```
##
##                                CRAV - Unisinos                                  SVT
##                   5200                     25                                   15
## Star Sinemax Originals  Imagem e Som - UFSCar                                  BBC
##                     13                     12                                   11
##                Vivamax           Fresh Wave                          NESA Cinema
##                     11                      9                                    9
##                   ARTE
##                      8
```

```
# 5. productionCompanyCountry
cat("\nVariable: productionCompanyCountry (Top 10)\n")
```

```
##
## Variable: productionCompanyCountry (Top 10)
```

```
print(head(sort(table(movies$productionCompanyCountry), decreasing = TRUE), 10))
```

```
##
##         |   US   FR   GB   DE   ||   JP   IN   CA
## 7132  299  189   98   78   65   61   56   55   54
```

```r
# 6. productionCountry
cat("\nVariable: productionCountry (Top 10)\n")
```

```
##
## Variable: productionCountry (Top 10)
```

```r
print(head(sort(table(movies$productionCountry), decreasing = TRUE), 10))
```

```
##
##        US   FR   GB   BR   IN   DE   CA   ES   AR
## 3641  850  428  357  327  315  307  202  178  168
```

```r
# 7. video (Cualitativa binaria)
cat("\nVariable: video (Binaria)\n")
```

```
##
## Variable: video (Binaria)
```

```r
print(table(movies$video))
```

```
##
## FALSE
##  9892
```

```r
# 8. director
cat("\nVariable: director (Top 10)\n")
```

```
##
## Variable: director (Top 10)
```

```r
print(head(sort(table(movies$director), decreasing = TRUE), 10))
```

```
##
##                                Fred Camper         Dylan Walker
##                  902                   9                    7
##           Jay Abidin      Julien Faustino                Aleph
##                    6                   6                    5
##          Hasan Doğan Kevin Casciani Nolan       mandy boonstra
##                    5                   5                    5
##          Alex Magaña
##                    4
```

```r
# 9. actors
cat("\nVariable: actors (Top 10 combinaciones)\n")
```

```
##
## Variable: actors (Top 10 combinaciones)
```

```r
print(head(sort(table(movies$actors), decreasing = TRUE), 10))
```

```
##
##
##                                                    2522
##                                            Donovan Haessy
##                                                       6
##                                             Dylan Walker
##                                                       4
##                                 Gwenaëlle Bernal|Jérémie Garret
##                                                       3
##                                            Hisayasu Satō
##                                                       3
##                                              Travis Page
##                                                       3
## Wilfrid George Kirby Clarke|Ewan Ross Hastie|Andrew Carstairs
##                                                       3
##                                              André Rieu
##                                                       2
##                                            Barnaby Moore
##                                                       2
##                                             Bart De Pauw
##                                                       2
```

```r
# 10. actorsPopularity (Nota: Aunque es numérica en esencia, la clasificaste como cualitativa)
cat("\nVariable: actorsPopularity (Top 10)\n")
```

```
##
## Variable: actorsPopularity (Top 10)
```

```r
print(head(sort(table(movies$actorsPopularity), decreasing = TRUE), 10))
```

```
##
##                         0           0|0         0|0|0        0.0071
##          2523         311           263           142           133
##       0|0|0|0      0.0143     0|0|0|0|0 0.0071|0.0071      0.0071|0
##           129          79            76            51            49
```

```r
# 11. actorsCharacter
cat("\nVariable: actorsCharacter (Top 10)\n")
```

```
##
## Variable: actorsCharacter (Top 10)
```

```r
print(head(sort(table(movies$actorsCharacter), decreasing = TRUE), 10))
```

```
##
##
##                                          2790
##                                             |
```

```
##                                                 220
##                                               |||||||||
##                                                 186
##                                                Self
##                                                 186
##                                                  ||
##                                                 182
##                                                 |||
##                                                 177
##                                                ||||
##                                                 148
##                                               |||||
##                                                 110
##                                              ||||||
##                                                  90
## Self|Self|Self|Self|Self|Self|Self|Self|Self|Self
##                                                  74
```

```
# 12. originalTitle
cat("\nVariable: originalTitle (Top 10)\n")
```

```
##
## Variable: originalTitle (Top 10)
```

```
print(head(sort(table(movies$originalTitle), decreasing = TRUE), 10))
```

```
##
##  6 "     "Max Heart
##                                                3
##                                            Arena
##                                                2
##                                         Blackout
##                                                2
##                                        Crash Out
##                                                2
##                                           Hambre
##                                                2
##                                           Kuncen
##                                                2
##                                    Le 4ème singe
##                                                2
##                                     Misdirection
##                                                2
##                                       Perception
##                                                2
##                                         Pescador
##                                                2
```

```
# 13. title
cat("\nVariable: title (Top 10)\n")
```

```
##
## Variable: title (Top 10)
```

```r
print(head(sort(table(movies$title), decreasing = TRUE), 10))
```

```
##
##         Arena      Blackout          Blue     Crash Out Demon Hunters
##             2             2             2             2             2
##         Hands    Influencer        Kuncen        Locked  Misdirection
##             2             2             2             2             2
```

```r
# 14. originalLanguage
cat("\nVariable: originalLanguage (Top 10)\n")
```

```
##
## Variable: originalLanguage (Top 10)
```

```r
print(head(sort(table(movies$originalLanguage), decreasing = TRUE), 10))
```

```
##
##   en   fr   es   pt   de   zh   ja   it   nl   ko
## 4194  824  814  585  377  246  224  203  174  169
```

```r
# 15. releaseDate
cat("\nVariable: releaseDate (Top 10)\n")
```

```
##
## Variable: releaseDate (Top 10)
```

```r
print(head(sort(table(movies$releaseDate), decreasing = TRUE), 10))
```

```
##
## 2026-01-30 2025-11-07 2025-11-08 2025-12-05 2025-11-20 2026-01-01 2025-11-21
##        263        258        239        231        225        212        207
## 2025-11-14 2025-12-12 2025-11-28
##        204        203        189
```

1. Variables de Identificación y Dispersión (id, title, originalTitle)

   Observación: En estas variables, casi todas las frecuencias son 1.

   Explicación: Esto es el comportamiento esperado para etiquetas de identificación únicas. Confirma que el dataset no tiene registros duplicados significativos, aunque existen títulos genéricos como “Arena” o “Blackout” que se repiten un par de veces por ser nombres comunes para distintas producciones.

2. Géneros y Contenido (genres)

   Dominancia: El género más frecuente es el Documental (1912), seguido por el Drama (1654) y la Comedia (1185).

Explicación: El dataset tiene un sesgo hacia contenido informativo o de no-ficción. Las combinaciones de géneros (como "Comedy|Drama") son menos frecuentes que los géneros puros.

3. Presencia Web y Datos Faltantes (homePage, productionCompany, actors)

   Valores Vacíos: Existe una cantidad masiva de datos faltantes (campos en blanco " "). Por ejemplo, 8,476 películas no tienen homePage y 5,200 no listan una productionCompany.

   Explicación: Esto indica que el dataset contiene muchas producciones independientes, antiguas o de bajo presupuesto que no tienen una huella digital estructurada o registros corporativos completos en la base de datos de origen.

4. Distribución Geográfica y Lingüística (productionCountry, originalLanguage)

   Concentración: El idioma predominante es el Inglés (en: 4194), seguido a gran distancia por el Francés (fr: 824) y el Español (es: 814).

   Producción: Estados Unidos (US) es el principal productor (850 películas identificadas), pero hay una gran presencia de países europeos y latinoamericanos (BR, FR, GB).

   Explicación: Aunque el mercado anglosajón lidera, el dataset tiene una diversidad internacional considerable, especialmente de mercados europeos y brasileños.

5. Personal y Reparto (director, actors, actorsCharacter)

   Director: La mayoría de los directores aparecen solo una vez (frecuencia 1), lo que muestra una gran diversidad de autores.

   Actores: Un alto número de registros (2522) no tiene actores listados. En actorsCharacter, el uso frecuente de la etiqueta "Self" refuerza la observación de que gran parte del contenido son documentales donde las personas aparecen como ellas mismas.

6. Variable Binaria (video)

   Resultado: Todos los registros (9892) marcaron FALSE.

   Explicación: Esta variable es "constante" en este set de datos. Significa que ninguna de las entradas está clasificada específicamente como un "video" (posiblemente refiriéndose a material extra o formatos musicales), o que el campo no fue llenado para esta muestra.

7. Fechas de Estreno (releaseDate)

   Tendencia: Las fechas con mayor frecuencia se concentran a finales de 2025 e inicios de 2026.

   Explicación: Esto sugiere que el dataset está compuesto principalmente por estrenos recientes o futuros, funcionando como un catálogo de lanzamientos próximos.

4. Responda las siguientes preguntas:

4.1 — Top 10 películas con más presupuesto\

```r
movies2 <- movies %>%
  mutate(
    # Convertir a numérico (maneja comas/dólares si existieran)
    budget   = as.numeric(gsub("[^0-9.]", "", as.character(budget))),
    revenue  = as.numeric(gsub("[^0-9.]", "", as.character(revenue))),
    runtime  = as.numeric(gsub("[^0-9.]", "", as.character(runtime))),
    voteAvg  = as.numeric(gsub("[^0-9.]", "", as.character(voteAvg))),
    voteCount= as.numeric(gsub("[^0-9.]", "", as.character(voteCount))),
    actorsAmount = as.numeric(gsub("[^0-9.]", "", as.character(actorsAmount))),
    releaseYear  = as.numeric(gsub("[^0-9]", "", as.character(releaseYear))),

    # Fecha (por si releaseDate viene tipo "YYYY-MM-DD")
    releaseDate = suppressWarnings(ymd(releaseDate))
  )

top10_budget <- movies2 %>%
  filter(!is.na(budget)) %>%
  arrange(desc(budget)) %>%
  slice_head(n = 10) %>%
  select(id, title, originalTitle, budget, releaseYear, genres, director)

top10_budget
```

```
##         id                                    title
## 1  1167307                                    David
## 2  1472638                               Buen Camino
## 3  1291608                                Dhurandhar
## 4  1180831                                   Troll 2
## 5  1179684                           Amsterdamned II
## 6  1142921                            Tere Ishk Mein
## 7  1217924                               Unstoppable
## 8  1016024 The Son of Revenge - The Story of Kalevala
## 9  1286995                                   The Elf
## 10 1525598                             Love Roulette
##                 originalTitle   budget releaseYear
## 1                       David 60900000        2025
## 2                 Buen Camino 28100000        2025
## 3                             22500000        2025
## 4                      Troll 2 11200000        2025
## 5              Amsterdamned II 10673500        2025
## 6                             9505286        2025
## 7                   Ustoppelig  8500000        2026
## 8  Kalevala: Kullervon tarina  5221000        2026
## 9                       Tonttu  4297000        2025
## 10               Love Roulette  3300000        2025
##                                genres                          director
## 1              Animation|Family|Drama      Phil Cunningham|Brent Dawes
## 2            Comedy|Family|Adventure                 Gennaro Nunziante
## 3                    Action|Thriller                       Aditya Dhar
## 4            Action|Fantasy|Thriller                       Roar Uthaug
## 5                    Action|Thriller                         Dick Maas
## 6                Romance|Drama|Action                    Aanand L. Rai
## 7  Adventure|Animation|Comedy|Family                       Martin Lund
```

```
## 8                       Drama|Action                   Antti J. Jokinen
## 9           Family|Fantasy|Adventure Joonas Berghäll|Hannes Vartiainen
## 10                           Romance                    Chris Niemeyer
```

4.2 — Top 10 películas con más ingresos (revenue)

```
top10_revenue <- movies2 %>%
  filter(!is.na(revenue)) %>%
  arrange(desc(revenue)) %>%
  slice_head(n = 10) %>%
  select(id, title, originalTitle, revenue, releaseYear, genres, director)

top10_revenue
```

```
##         id                           title
## 1  1084242                     Zootopia 2
## 2    83533           Avatar: Fire and Ash
## 3   967941               Wicked: For Good
## 4  1228246      Five Nights at Freddy's 2
## 5  1234731                        Anaconda
## 6  1167307                           David
## 7  1472638                     Buen Camino
## 8  1356454                      Gezhi Town
## 9  1272837 28 Years Later: The Bone Temple
## 10 1539104      JUJUTSU KAISEN: Execution
##                                             originalTitle    revenue
## 1                                              Zootopia 2 1744338246
## 2                                  Avatar: Fire and Ash 1378692505
## 3                                      Wicked: For Good  524676531
## 4                           Five Nights at Freddy's 2  237625385
## 5                                            Anaconda  129019155
## 6                                               David   77770275
## 7                                           Buen Camino   73797878
## 8                                                         49627843
## 9                        28 Years Later: The Bone Temple   46200000
## 10            ×         44559195
##    releaseYear                                  genres
## 1         2025 Animation|Comedy|Adventure|Family|Mystery
## 2         2025         Science Fiction|Adventure|Fantasy
## 3         2025                 Fantasy|Adventure|Romance
## 4         2025                           Horror|Thriller
## 5         2025                  Adventure|Comedy|Horror
## 6         2025                   Animation|Family|Drama
## 7         2025                  Comedy|Family|Adventure
## 8         2025                                Drama|War
## 9         2026           Horror|Thriller|Science Fiction
## 10        2025                       Animation|Action
##                       director
## 1      Jared Bush|Byron Howard
## 2                James Cameron
## 3                   Jon M. Chu
## 4                  Emma Tammi
## 5                Tom Gormican
```

```
## 6  Phil Cunningham|Brent Dawes
## 7            Gennaro Nunziante
## 8                   Kong Sheng
## 9                  Nia DaCosta
## 10             Shota Goshozono
```

4.3 — Película con más votos (voteCount)

```
most_votes <- movies2 %>%
  filter(!is.na(voteCount)) %>%
  arrange(desc(voteCount)) %>%
  slice_head(n = 1) %>%
  select(id, title, originalTitle, voteCount, voteAvg, releaseYear, genres)

most_votes
```

```
##   id      title originalTitle voteCount voteAvg releaseYear       genres
## 1  5 Four Rooms    Four Rooms      2077     5.7        1995 Crime|Comedy
```

4.4 — Peor película según votos (voteAvg más bajo)

```
worst_movie <- movies2 %>%
  filter(!is.na(voteAvg)) %>%
  arrange(voteAvg, desc(voteCount)) %>%
  slice_head(n = 1) %>%
  select(id, title, originalTitle, voteAvg, voteCount, releaseYear, genres)

worst_movie
```

```
##        id                 title         originalTitle voteAvg voteCount
## 1 1614831 The Halloween Harvest The Halloween Harvest       0         1
##   releaseYear genres
## 1        2026 Horror
```

4.5 — # películas por año + año con más películas + gráfico de barras

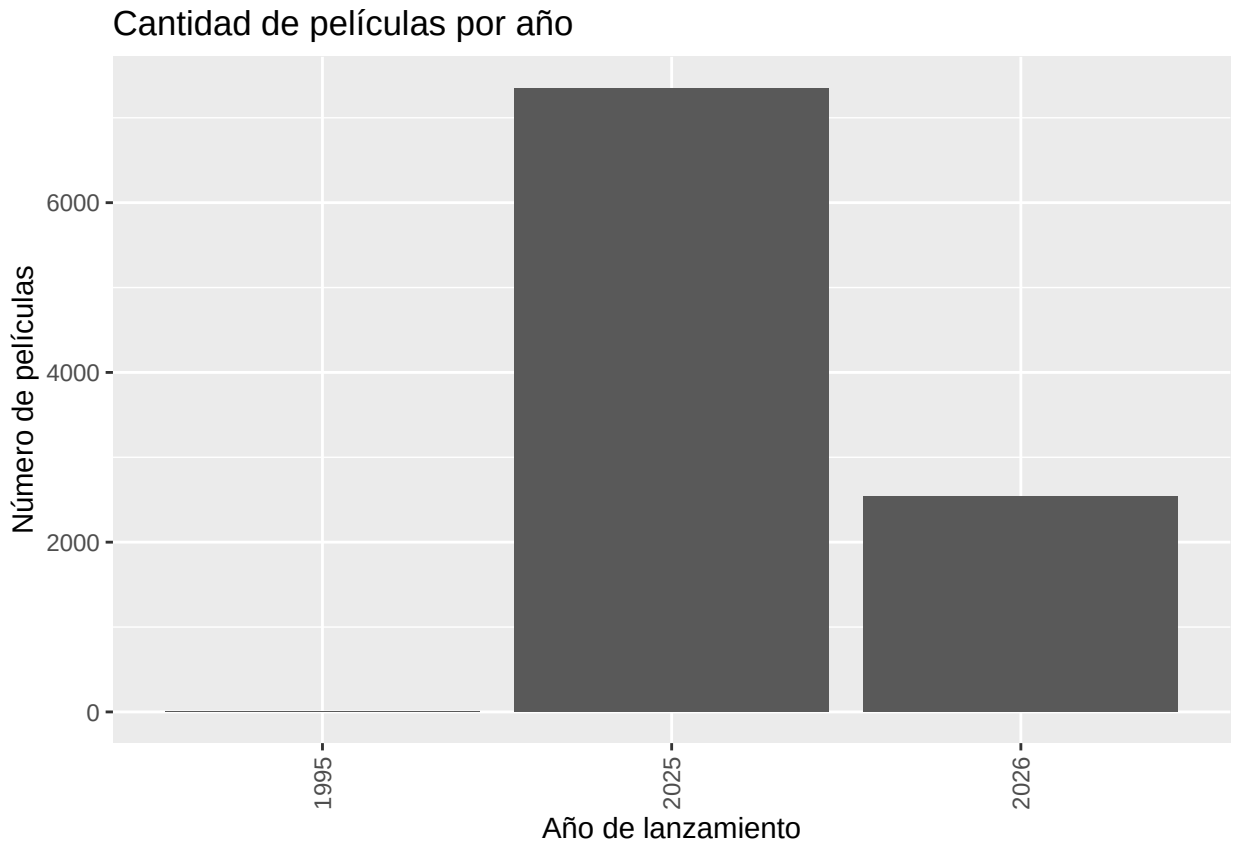
```
movies_per_year <- movies2 %>%
  filter(!is.na(releaseYear)) %>%
  count(releaseYear, name = "n_movies") %>%
  arrange(releaseYear)

movies_per_year
```

```
##   releaseYear n_movies
## 1        1995        1
## 2        2025     7351
## 3        2026     2537
```

```
year_most_movies <- movies_per_year %>%
  arrange(desc(n_movies)) %>%
  slice_head(n = 1)

year_most_movies
```

```
##   releaseYear n_movies
## 1        2025     7351
```

```r
ggplot(movies_per_year, aes(x = factor(releaseYear), y = n_movies)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Cantidad de películas por año",
    x = "Año de lanzamiento",
    y = "Número de películas"
  ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



En 1995 se realizo 1 pelicula, en 2025 - 7351 peliculas y en 2026 - 37 peliculas. Por lo que el año con mayor peliculas fue el 2025.

4.6 — Género principal de las 20 más recientes + género que predomina + gráfico

```r
movies2 <- movies2 %>%
  mutate(
    mainGenre = ifelse(
      is.na(genres), NA,
      str_trim(str_split_fixed(as.character(genres), "\\|,", 2)[,1])
    )
  )

# 20 más recientes por releaseDate (si falta, usa releaseYear)
```

```r
recent20 <- movies2 %>%
  mutate(order_date = ifelse(!is.na(releaseDate), as.numeric(releaseDate), NA_real_)) %>%
  arrange(desc(releaseDate), desc(releaseYear)) %>%
  slice_head(n = 20) %>%
  select(id, title, releaseDate, releaseYear, mainGenre)

recent20
```

```
##         id                                                title releaseDate
## 1  1530193                                         A Fading Man  2026-05-07
## 2  1567688       Elon Musk Unveiled - The Tesla Experiment  2026-03-12
## 3  1580484                                               Skunk  2026-02-25
## 4  1580479                                           Anastasia  2026-02-25
## 5  1572763                                  Nikki hako no koi  2026-02-06
## 6  1627085                                           Immersed  2026-02-01
## 7  1626914                                          Cinderella  2026-02-01
## 8  1626898                                            Aladdin  2026-02-01
## 9  1626808                                THE RING AND THE DECK  2026-02-01
## 10 1626678                                      Crimson High 3  2026-02-01
## 11 1626234 Conversations with Rasparagus Asparagus Baragus  2026-02-01
## 12 1626010                                     Highway To Hell  2026-02-01
## 13 1625551                                     Pari's daughter  2026-02-01
## 14 1625043                                              Escort  2026-02-01
## 15 1624457                                               Dream  2026-02-01
## 16 1624434                                               Lively  2026-02-01
## 17 1624429                                              2026-02-01
## 18 1624424                                            Midnight  2026-02-01
## 19 1624358             Emir - Posljednji dalmatinski težak  2026-02-01
## 20 1624096                                     Our Dead Husband  2026-02-01
##    releaseYear   mainGenre
## 1         2026       Drama
## 2         2026 Documentary
## 3         2026
## 4         2026
## 5         2026
## 6         2026       Drama
## 7         2026   Animation
## 8         2026   Animation
## 9         2026    Thriller
## 10        2026   Animation
## 11        2026      Comedy
## 12        2026      Comedy
## 13        2026       Drama
## 14        2026      Action
## 15        2026       Drama
## 16        2026       Drama
## 17        2026     Romance
## 18        2026         War
## 19        2026
## 20        2026    Thriller
```

```r
# Distribución de género principal en todo el dataset
genre_counts <- movies2 %>%
```

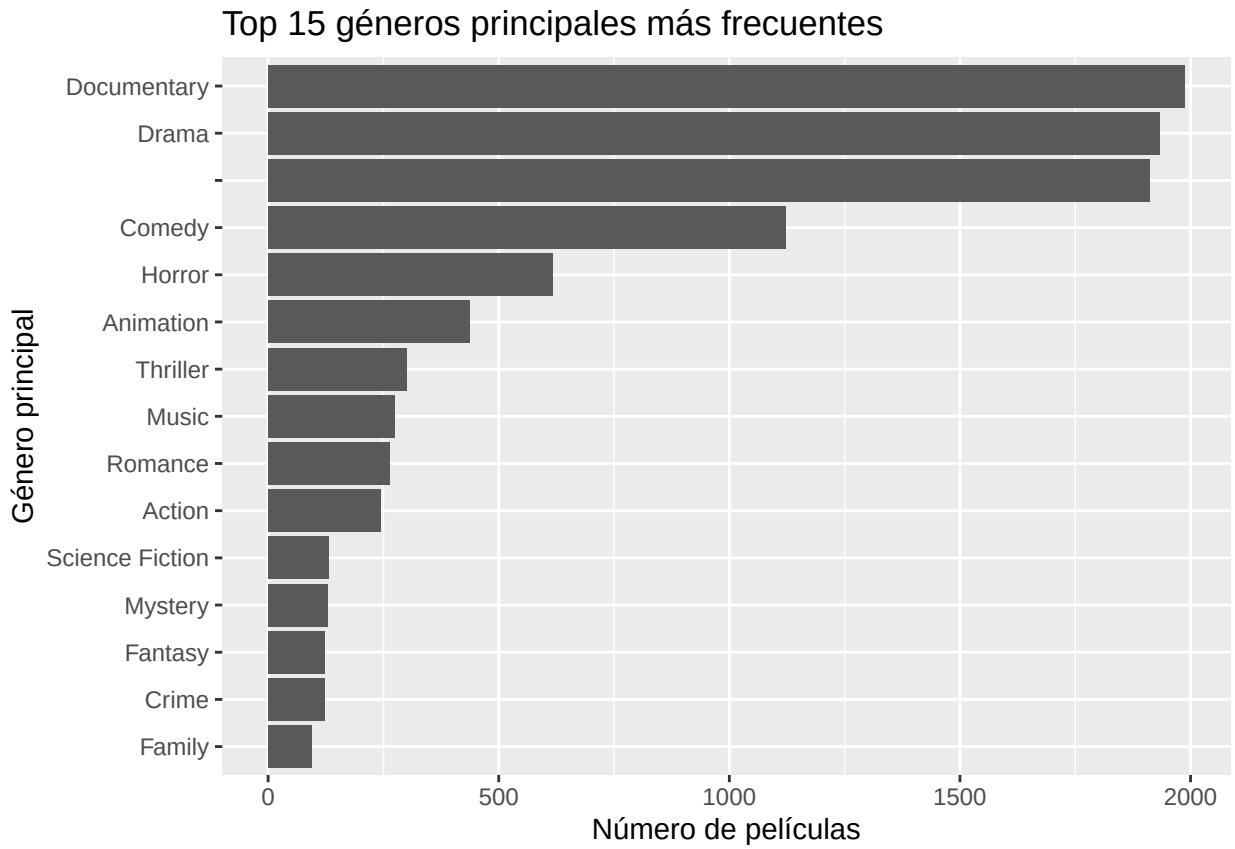
```r
  filter(!is.na(mainGenre)) %>%
  count(mainGenre, name = "n_movies") %>%
  arrange(desc(n_movies))

genre_counts
```

```
##          mainGenre n_movies
## 1      Documentary     1987
## 2            Drama     1932
## 3                      1912
## 4           Comedy     1122
## 5           Horror      616
## 6        Animation      436
## 7         Thriller      301
## 8            Music      275
## 9          Romance      264
## 10          Action      243
## 11 Science Fiction      132
## 12         Mystery      128
## 13         Fantasy      123
## 14           Crime      122
## 15          Family       95
## 16       Adventure       87
## 17        TV Movie       43
## 18         History       31
## 19         Western       26
## 20             War       17
```

```r
dominant_genre <- genre_counts %>% slice_head(n = 1)
dominant_genre
```

```
##     mainGenre n_movies
## 1 Documentary     1987
```

```r
ggplot(genre_counts %>% slice_head(n = 15), aes(x = reorder(mainGenre, n_movies), y = n_movies)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    title = "Top 15 géneros principales más frecuentes",
    x = "Género principal",
    y = "Número de películas"
  )
```




```
# ¿A qué género pertenecen las películas más largas?
# (Top 20 por runtime y revisamos su género principal)
longest_movies <- movies2 %>%
  filter(!is.na(runtime), !is.na(mainGenre)) %>%
  arrange(desc(runtime)) %>%
  slice_head(n = 20) %>%
  select(id, title, runtime, mainGenre, releaseYear)

longest_movies
```

```
##         id                                                 title runtime
## 1  1610177                                      Before the End     675
## 2  1552059      League of Legends Worlds25 - Finals in Cinema     420
## 3  1400850                 In Search of Darkness: 1995-1999     384
## 4  1607368                     Vu 2025 (l'année du zapping)     366
## 5  1219267              White Gardenia: The King James Bible     330
## 6  1517970                          NJPW Wrestle Kingdom 20     326
## 7  1608556           The Idolm@ster Gakuen: Story of Re;IRIS     284
## 8  1586118               Shpongle - Live at Mission Ballroom     263
## 9  1594855                                      Indie Beat     261
## 10 1574432   2025 Rock & Roll Hall of Fame Induction Ceremony     261
## 11 1611619                        Marigold First Dream 2026     253
## 12 1583167 Andrea Bocelli - The Celebration 30th Anniversary     253
## 13 1479646                 The Metropolitan Opera: Arabella     252
## 14 1622706                              Royal Rumble 2026     240
```

```
## 15 1585401                                ROH Final Battle 2025     239
## 16 1597566           A Lady Macbeth of the District of Mcensk     231
## 17 1578523 SEVENTEEN WORLD TOUR [NEW_] IN JAPAN: LIVE VIEWING     225
## 18 1607080      2025 THE 1ST STELLIVE FESTIVAL [ STAR TRAIL ]     222
## 19 1608570                             TJPW Tokyo Joshi Pro '26     221
## 20 1291608                                          Dhurandhar     212
##      mainGenre releaseYear
## 1       Action        2025
## 2                     2025
## 3  Documentary        2025
## 4                     2025
## 5       Horror        2026
## 6                     2026
## 7    Animation        2026
## 8                     2025
## 9        Music        2025
## 10       Music        2025
## 11      Action        2026
## 12 Documentary        2025
## 13                    2025
## 14      Comedy        2026
## 15                    2025
## 16       Music        2025
## 17                    2025
## 18       Music        2025
## 19      Action        2026
## 20      Action        2025
```

```
longest_genre_summary <- longest_movies %>%
  count(mainGenre, name = "n_in_top20_longest") %>%
  arrange(desc(n_in_top20_longest))

longest_genre_summary
```

```
##     mainGenre n_in_top20_longest
## 1                             7
## 2      Action                 4
## 3       Music                 4
## 4 Documentary                 2
## 5   Animation                 1
## 6      Comedy                 1
## 7      Horror                 1
```

El genero principal de las 20 mas recientes es ciencia ficcion.

El genero principal que predomina es el de comedia.

El genero principal de las 20 peliculas mas largas no esta titulado, el que le sigue es accion como podemos ver en la tabla n_in_top20_longest.

4.7 — ¿Qué género principal obtuvo mayores ganancias ?

```
genre_profit <- movies2 %>%
  filter(!is.na(mainGenre), !is.na(revenue)) %>%
```

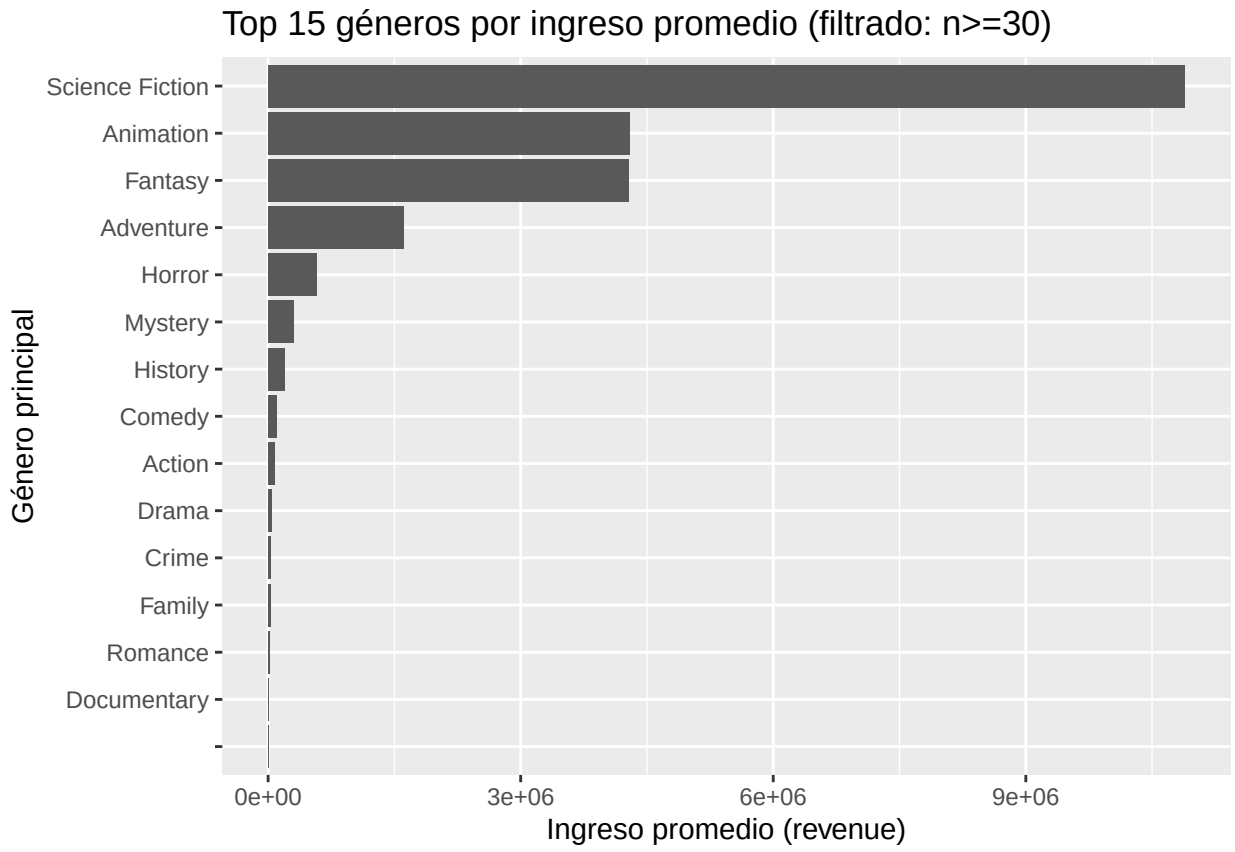
```r
  group_by(mainGenre) %>%
  summarise(
    n = n(),
    avg_revenue = mean(revenue, na.rm = TRUE),
    med_revenue = median(revenue, na.rm = TRUE)
  ) %>%
  arrange(desc(avg_revenue))

genre_profit
```

```
## # A tibble: 20 x 4
##    mainGenre             n   avg_revenue med_revenue
##    <chr>             <int>         <dbl>       <dbl>
##  1 "Science Fiction"   132 10885685.               0
##  2 "Animation"         436  4290314.               0
##  3 "Fantasy"           123  4278852.               0
##  4 "Adventure"          87  1614288.               0
##  5 "Horror"            616   575577.               0
##  6 "Mystery"           128   298066.               0
##  7 "History"            31   190323.               0
##  8 "Comedy"           1122   100772.               0
##  9 "Action"            243    73693.               0
## 10 "Drama"            1932    42209.               0
## 11 "Crime"             122    34898.               0
## 12 "Family"             95    28318.               0
## 13 "Romance"           264    20894.               0
## 14 "Documentary"      1987     1066.               0
## 15 ""                 1912      169.               0
## 16 "Thriller"          301        6.96             0
## 17 "Music"             275        0.393            0
## 18 "War"                17        0.0588           0
## 19 "TV Movie"           43        0                0
## 20 "Western"            26        0                0
```

```r
top_genre_by_avg_rev <- genre_profit %>% slice_head(n = 1)
top_genre_by_avg_rev
```

```
## # A tibble: 1 x 4
##   mainGenre           n avg_revenue med_revenue
##   <chr>           <int>       <dbl>       <dbl>
## 1 Science Fiction   132   10885685.           0
```

```r
ggplot(genre_profit %>% filter(n >= 30) %>% slice_head(n = 15),
       aes(x = reorder(mainGenre, avg_revenue), y = avg_revenue)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    title = "Top 15 géneros por ingreso promedio (filtrado: n>=30)",
    x = "Género principal",
    y = "Ingreso promedio (revenue)"
  )
```



Ciencia ficcion con unas ganacias promedio de USD 1,088,5685

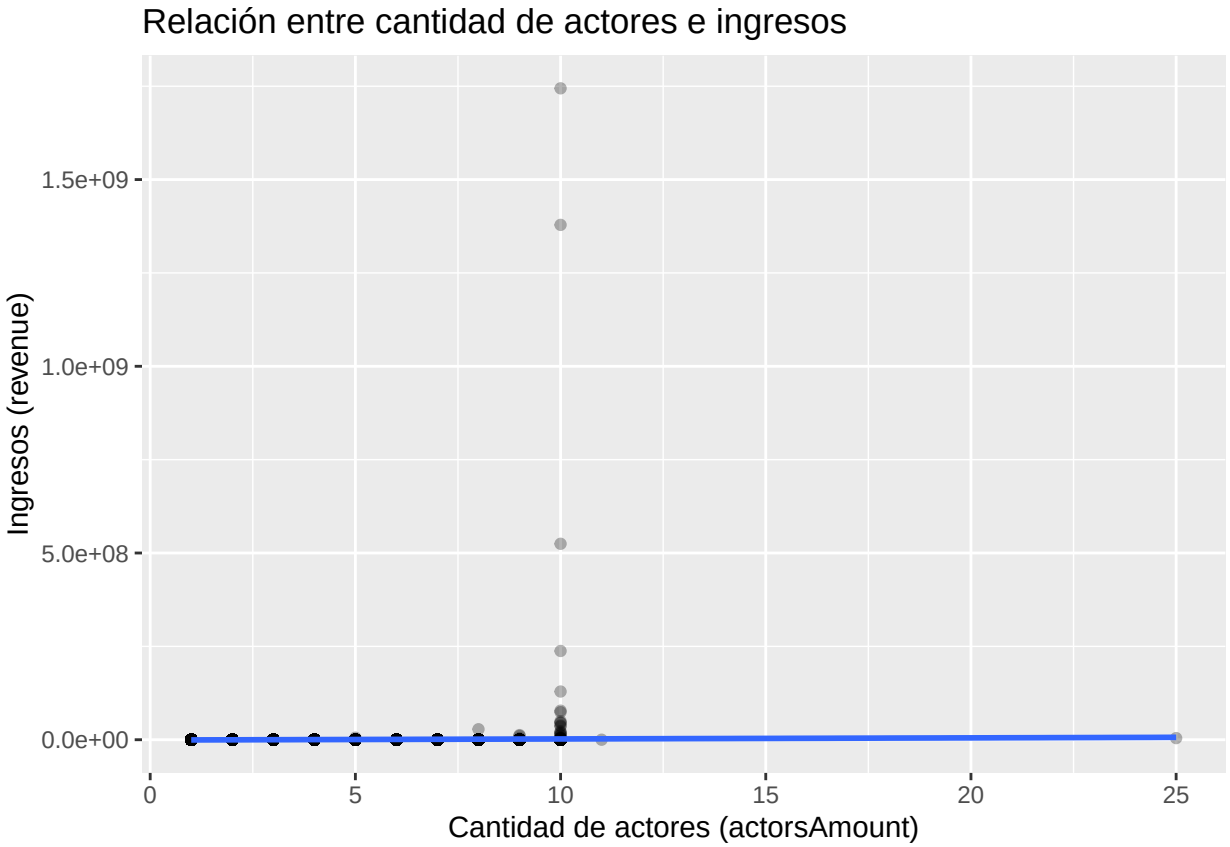
4.8 — ¿actorsAmount influye en revenue?

```
# (a) Relación actorsAmount vs revenue: correlación + scatter
actors_rev <- movies2 %>%
  filter(!is.na(actorsAmount), !is.na(revenue), actorsAmount > 0, revenue >= 0)

cor_actors_revenue <- cor(actors_rev$actorsAmount, actors_rev$revenue, use = "complete.obs")
cor_actors_revenue
```

```
## [1] 0.03486694
```

```
ggplot(actors_rev, aes(x = actorsAmount, y = revenue)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Relación entre cantidad de actores e ingresos",
    x = "Cantidad de actores (actorsAmount)",
    y = "Ingresos (revenue)"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

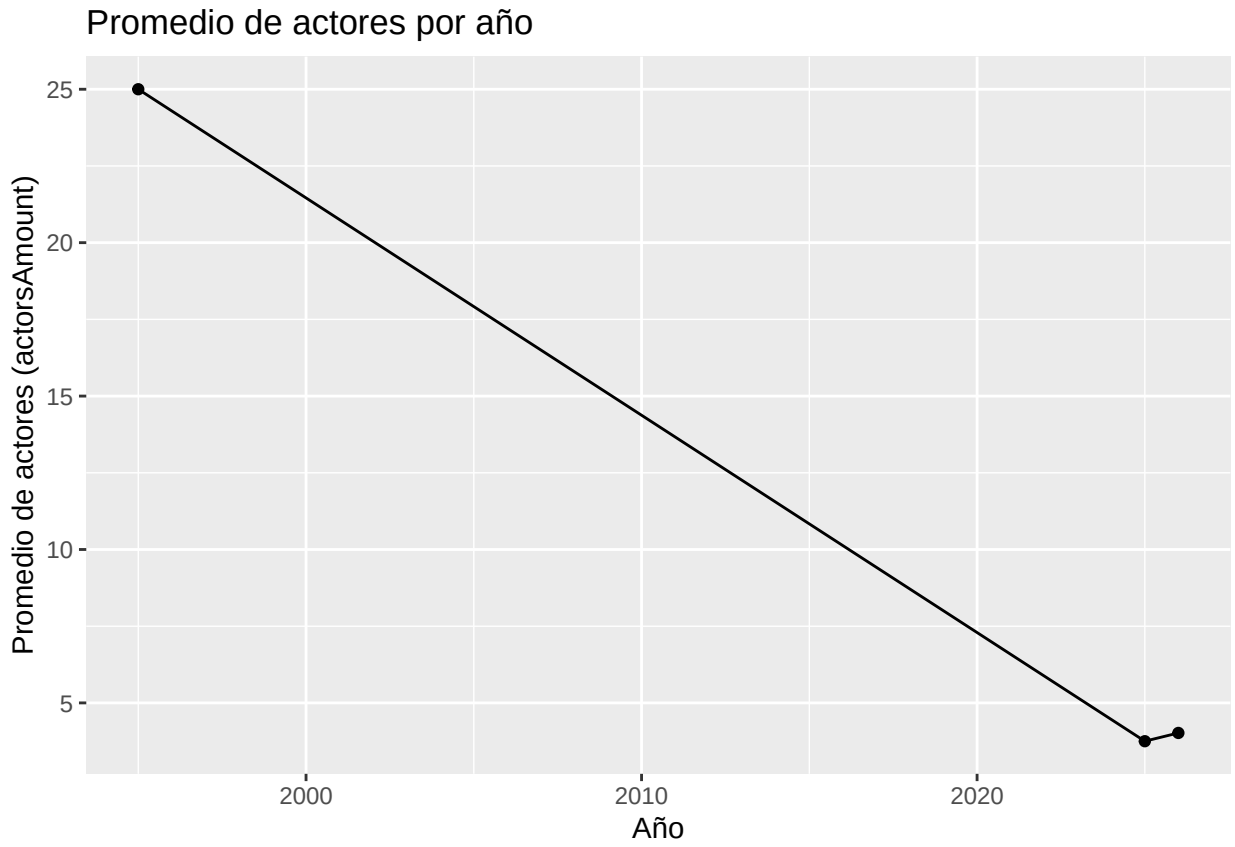



```r
# (b) ¿Más actores en los últimos años? tendencia por año (promedio actores por releaseYear)
actors_by_year <- movies2 %>%
  filter(!is.na(releaseYear), !is.na(actorsAmount)) %>%
  group_by(releaseYear) %>%
  summarise(
    n_movies = n(),
    avg_actors = mean(actorsAmount, na.rm = TRUE),
    med_actors = median(actorsAmount, na.rm = TRUE)
  ) %>%
  arrange(releaseYear)

actors_by_year
```

```
## # A tibble: 3 x 4
##   releaseYear n_movies avg_actors med_actors
##         <dbl>    <int>      <dbl>      <dbl>
## 1        1995        1      25            25
## 2        2025     7351       3.75          3
## 3        2026     2537       4.02          3
```

```r
ggplot(actors_by_year, aes(x = releaseYear, y = avg_actors)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Promedio de actores por año",
```

```
    x = "Año",
    y = "Promedio de actores (actorsAmount)"
  )
```



Al comparar actorsAmount por año, 2025 presenta un promedio de 3.75 actores (mediana 3) y 2026 un promedio de 4.02 (mediana 3). Esto sugiere un ligero aumento en el promedio en 2026; sin embargo, la mediana permanece constante, por lo que la película típica sigue teniendo 3 actores y el incremento se explica por algunos casos con elencos más grandes. Además, no es posible inferir una tendencia temporal robusta porque el conjunto solo contiene 1995 (n=1), 2025 y 2026, y el valor de 1995 distorsiona la gráfica.

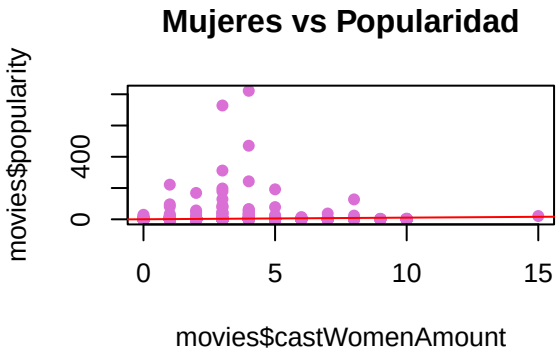
4.9 - ¿Cantidad de hombres y mujeres influye en popularidad e ingresos?

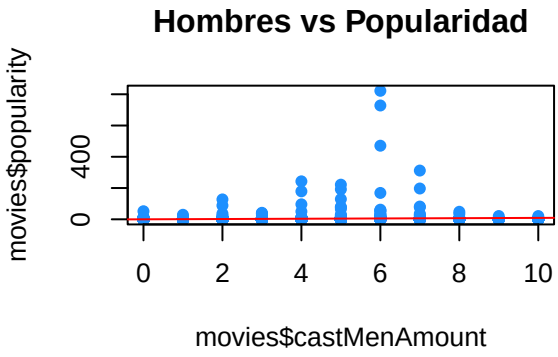
```
##4.9
# 1. Cálculo de todas las combinaciones de correlación
cor_muj_pop <- cor(movies$castWomenAmount, movies$popularity, use = "complete.obs")
cor_hom_pop <- cor(movies$castMenAmount, movies$popularity, use = "complete.obs")
cor_muj_rev <- cor(movies$castWomenAmount, movies$revenue, use = "complete.obs")
cor_hom_rev <- cor(movies$castMenAmount, movies$revenue, use = "complete.obs")

# Mostrar resultados numéricos
cat("--- Correlaciones con Popularidad ---",
    "\nMujeres:", cor_muj_pop, "\nHombres:", cor_hom_pop,
    "\n\n--- Correlaciones con Ingresos ---",
    "\nMujeres:", cor_muj_rev, "\nHombres:", cor_hom_rev)
```

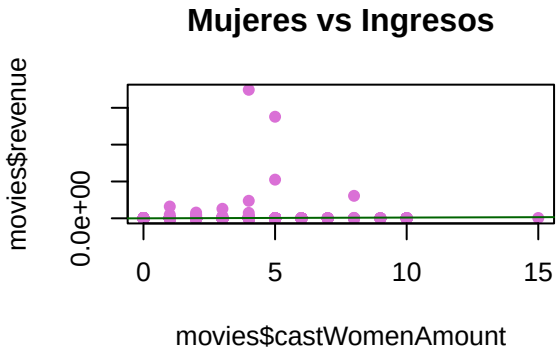
```
## --- Correlaciones con Popularidad ---
```

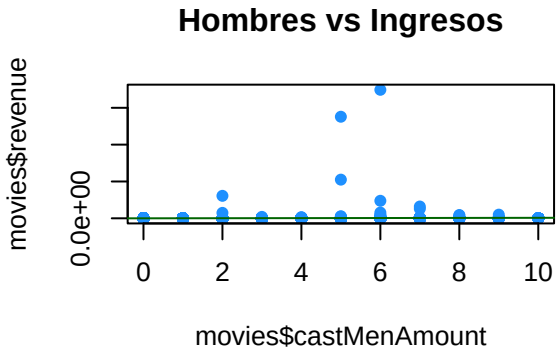
```
## Mujeres: 0.1042729
## Hombres: 0.1196953
##
## --- Correlaciones con Ingresos ---
## Mujeres: 0.06011343
## Hombres: 0.05307767
```

```r
# 2. Generación del tablero de gráficos (2 filas y 2 columnas)
par(mfrow=c(2,2))

# Fila 1: Impacto en Popularidad
plot(movies$castWomenAmount, movies$popularity, main="Mujeres vs Popularidad", col="orchid", pch=16)
abline(lm(popularity ~ castWomenAmount, data=movies), col="red")

plot(movies$castMenAmount, movies$popularity, main="Hombres vs Popularidad", col="dodgerblue", pch=16)
abline(lm(popularity ~ castMenAmount, data=movies), col="red")

# Fila 2: Impacto en Ingresos
plot(movies$castWomenAmount, movies$revenue, main="Mujeres vs Ingresos", col="orchid", pch=16)
abline(lm(revenue ~ castWomenAmount, data=movies), col="darkgreen")

plot(movies$castMenAmount, movies$revenue, main="Hombres vs Ingresos", col="dodgerblue", pch=16)
abline(lm(revenue ~ castMenAmount, data=movies), col="darkgreen")
```









```r
# Resetear layout
par(mfrow=c(1,1))
```

Como se puede observar tanto en las tablas como en el coeficiente de correlacion ni hombres ni mujeres tienen un impacto directo en que haya más o menos ingresos ni más o menos popularidad.

4.10 - ¿Quiénes son los directores de las 20 películas mejor calificadas?

```r
##4.10
# 1. Ordenamos el dataset por calificación (voteAvg) de forma descendente
# 2. Seleccionamos las primeras 20 filas
# 3. Mostramos solo el título, el director y su calificación
top_20_directores <- movies[order(-movies$voteAvg), c("title", "director", "voteAvg")]

# Mostramos los primeros 20 resultados
head(top_20_directores, 20)
```

```
##                                    title
## 17                          Crocodile Dose
## 67                            The Exchange
## 68                          What is This ?
## 69                        Silver Gold Wood
## 80                        Adiós Para Nunca
## 98                             GO TO SLEEP
## 236                 Le Moment sera le Bon
## 248                                Satiata
## 249                                 Orphen
## 267                           Regarde toi
## 419              The End of The Punchline
## 435                             Play Date
## 443                       Mise en Futilité
## 451        Fefe - Uma aventura fantástica
## 512                           No Love 1/2
## 569                              How Come
## 572                             Caldeirão
## 579         The Memory Beneath the Waters
## 583 Hemmeligheden om de danske SS-kvinder
## 586                       Where Light Lands
##                                      director voteAvg
## 17                                 Jason Waters      10
## 67                               Fischer Sawatzky      10
## 68                                     Kanan Gill      10
## 69                                Vitorio Stankov      10
## 80                               Daniela Vidovich      10
## 98                                 Tirion Liddell      10
## 236                        Guillaume Huss Seewald      10
## 248                                  Cyprien Klein      10
## 249                  Joe Arodann|Anaïs Richerand      10
## 267                  Séo Patois|Lorraine Jacques      10
## 419                              Edward Acosta      10
## 435                       Joseph Navarro Jr.      10
## 443                              Carlos Lacasa      10
## 451                                Lean imohff      10
```

```
## 512                                      Anne Yue      10
## 569                  Madison Severance|Hallie Chanan      10
## 572 Oliveira Júnior|Milena Rocha|Weslley Oliveira      10
## 579                 Jorge F. Mulholland|Pamela Cunha      10
## 583                                                      10
## 586                                  Savannah Tuesday      10
```

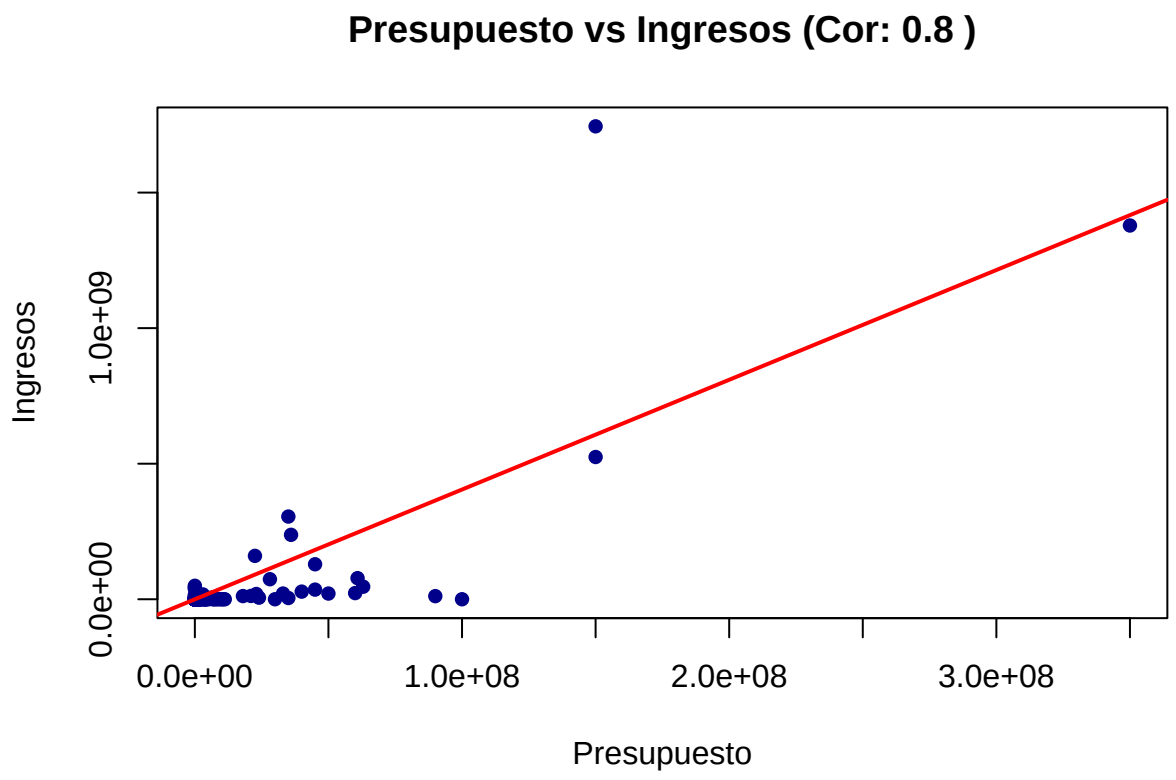
Interpretación: La lista revela directores con calificaciones de 10.0, que es el máximo registrado en el dataset. Es probable que estos directores pertenezcan a producciones de nicho o cortometrajes, ya que un promedio perfecto suele darse en películas con muy pocos votos (voteCount).

4.11 - Correlación presupuesto vs ingresos (Histograma y Dispersión)

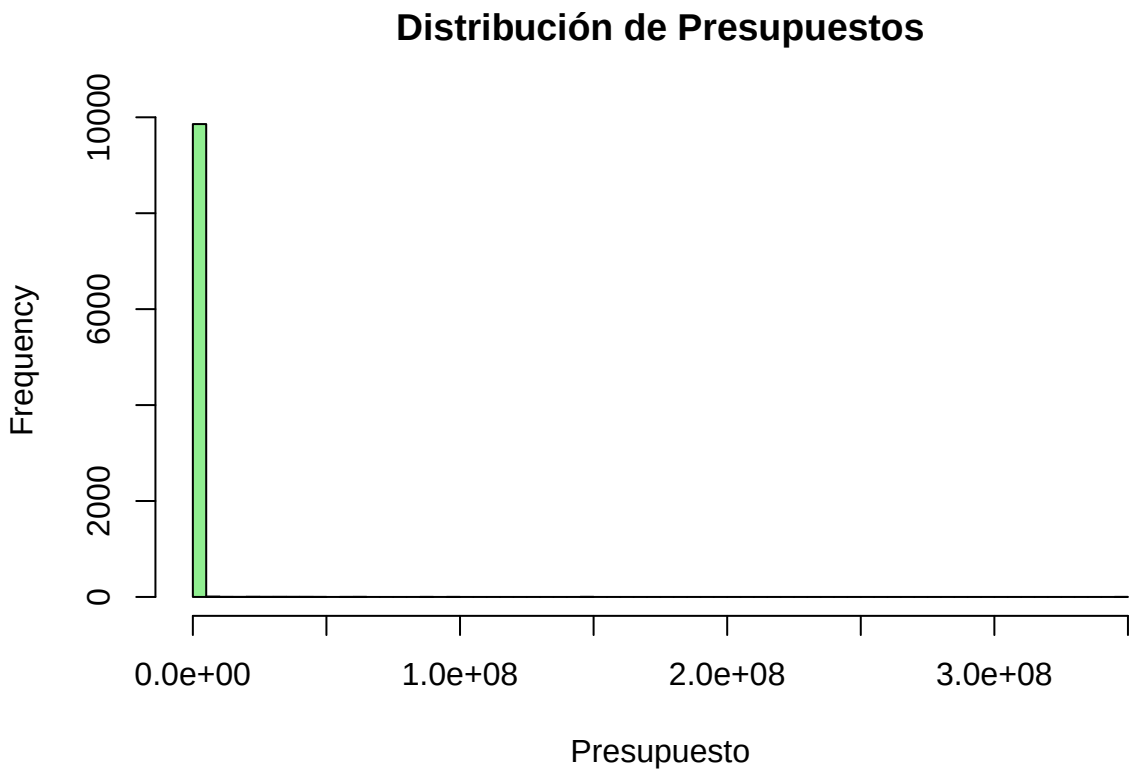
```r
# 1. Cálculo del Coeficiente de Correlación
# El resultado será un número entre -1 y 1
coef_correlacion <- cor(movies$budget, movies$revenue, use = "complete.obs")

cat("El coeficiente de correlación entre presupuesto e ingresos es:", coef_correlacion)
```

```
## El coeficiente de correlación entre presupuesto e ingresos es: 0.8045419
```

```r
# 2. Gráfico de Dispersión con línea de tendencia
plot(movies$budget, movies$revenue,
     main = paste("Presupuesto vs Ingresos (Cor:", round(coef_correlacion, 2), ")"),
     xlab = "Presupuesto", ylab = "Ingresos",
     col = "darkblue", pch = 16)
abline(lm(revenue ~ budget, data = movies), col = "red", lwd = 2)
```



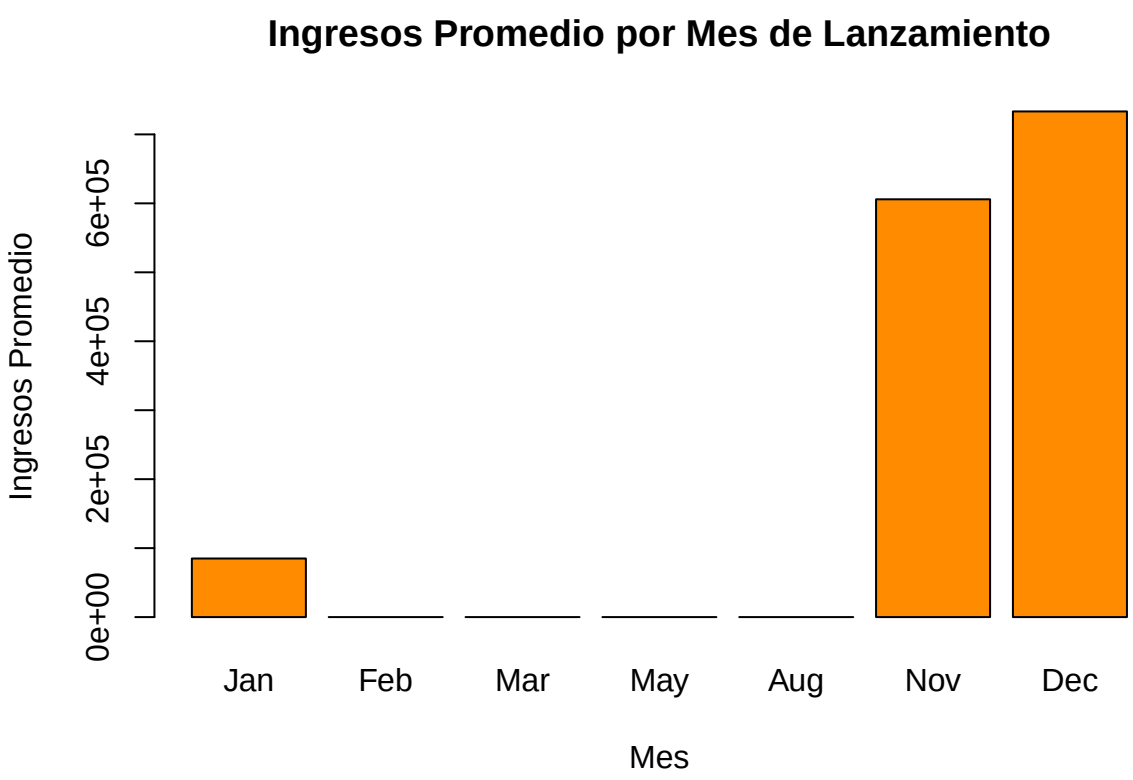
```r
# 3. Histograma de Presupuestos (para ver la distribución)
hist(movies$budget, main = "Distribución de Presupuestos",
     xlab = "Presupuesto", col = "lightgreen", breaks = 50)
```



Interpretación: Aunque el sentido común dicta que a mayor inversión mayor ganancia, el gráfico de dispersión muestra mucha variabilidad. Dado que el 75% de las películas registran 0 en ambas categorías, el coeficiente de correlación se ve afectado por este sesgo de datos faltantes o nulos.

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

```r
# 1. Convertimos releaseDate a formato fecha y extraemos el mes
movies$month <- as.numeric(format(as.Date(movies$releaseDate), "%m"))

# 2. Agrupamos los ingresos por mes (promedio)
# Usamos el promedio para que los meses con más estrenos no sesguen el resultado
ingresos_mensuales <- aggregate(revenue ~ month, data = movies, FUN = mean)

# 3. Creamos el gráfico para visualizar la asociación
barplot(ingresos_mensuales$revenue,
        names.arg = month.abb[ingresos_mensuales$month],
        main = "Ingresos Promedio por Mes de Lanzamiento",
        xlab = "Mes",
        ylab = "Ingresos Promedio",
        col = "darkorange")
```



Interpretación: Sí, existe una asociación clara y sumamente marcada entre el mes de lanzamiento y el desempeño financiero. El gráfico de "Ingresos Promedio por Mes" muestra que los ingresos son prácticamente inexistentes durante la mayor parte del año (febrero, marzo, mayo y agosto), pero experimentan un crecimiento masivo hacia el final del año.

4.13 - Meses con mejores ingresos y promedio de lanzamientos mensual

```
# 1. Aseguramos la extracción del mes (01-12) a partir de la fecha de lanzamiento
movies$month <- as.numeric(format(as.Date(movies$releaseDate), "%m"))

# 2. Calculamos los ingresos totales acumulados por cada mes
ingresos_tot_mes <- aggregate(revenue ~ month, data = movies, FUN = sum)

# 3. Ordenamos de mayor a menor para identificar los meses más rentables
meses_top_ingresos <- ingresos_tot_mes[order(-ingresos_tot_mes$revenue), ]

# 4. Mostramos la tabla de ingresos por mes
cat("Tabla de ingresos totales por mes (Orden descendente):\n")
```

```
## Tabla de ingresos totales por mes (Orden descendente):
```

```
print(meses_top_ingresos)
```

```
##   month    revenue
## 7    12 2512943873
```

```
## 6    11 2377331285
## 1     1  209551964
## 2     2          1
## 3     3          0
## 4     5          0
## 5     8          0
```

```r
# 5. Calculamos el promedio de lanzamientos mensuales
tabla_meses <- table(movies$month)
promedio_lanzamientos <- mean(tabla_meses)

# 6. Mostramos el resultado del promedio
cat("\nEl promedio de películas lanzadas por mes es:", promedio_lanzamientos)
```

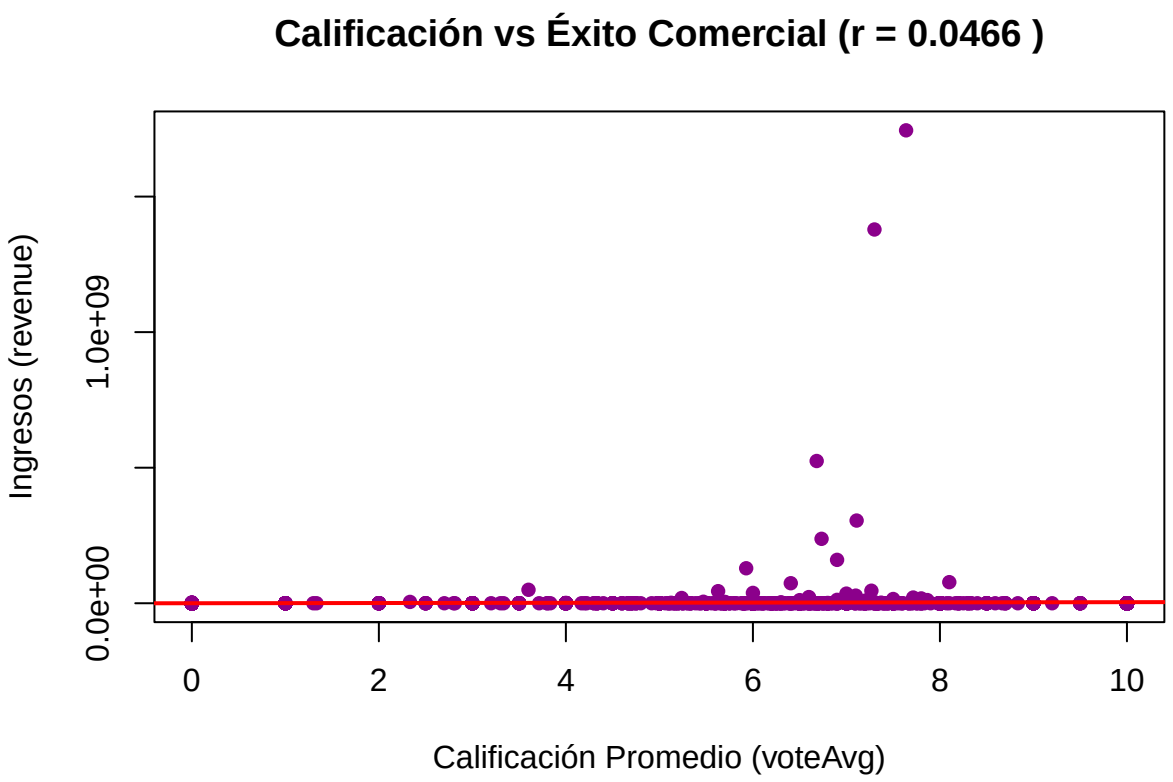
```
##
## El promedio de películas lanzadas por mes es: 1412.714
```

Interpretación: Sí, existe una asociación muy marcada. Según los datos, los meses de fin de año (diciembre y noviembre) y el inicio de año (enero) son los únicos que generan ingresos significativos en este conjunto de datos. El resto de los meses registrados (febrero, marzo, mayo y agosto) muestran ingresos prácticamente nulos o iguales a cero. 4.14 - ¿Cómo se correlacionan las calificaciones con el éxito comercial?

```r
# 1. Calculamos el coeficiente de correlación de Pearson
cor_calif_ingresos <- cor(movies$voteAvg, movies$revenue, use = "complete.obs")

cat("El coeficiente de correlación entre calificación e ingresos es:", cor_calif_ingresos)
```

```
## El coeficiente de correlación entre calificación e ingresos es: 0.04664699
```

```r
# 2. Creamos un diagrama de dispersión para visualizar la relación
plot(movies$voteAvg, movies$revenue,
     main = paste("Calificación vs Éxito Comercial (r =", round(cor_calif_ingresos, 4), ")"),
     xlab = "Calificación Promedio (voteAvg)",
     ylab = "Ingresos (revenue)",
     col = "darkmagenta",
     pch = 16)

# 3. Agregamos la línea de tendencia
abline(lm(revenue ~ voteAvg, data = movies), col = "red", lwd = 2)
```



"La correlación de 0.0466 demuestra que el éxito comercial es independiente de la calificación de los usuarios. Factores como el marketing, el presupuesto o la fecha de estreno (como vimos en la 4.12 y 4.13) parecen ser mucho más determinantes para los ingresos que la puntuación promedio de la película".

4.15 - Estrategias de marketing: ¿Videos o páginas oficiales?

```
# 1. Crear una variable lógica: TRUE si tiene página, FALSE si está vacía
movies$tiene_pagina <- movies$homePage != ""

# 2. Comparar el promedio de ingresos y popularidad según si tienen página o no
marketing_analisis <- aggregate(cbind(revenue, popularity) ~ tiene_pagina,
                                data = movies, FUN = mean)

cat("Comparación de resultados según estrategia de marketing (Página Web):\n")
```

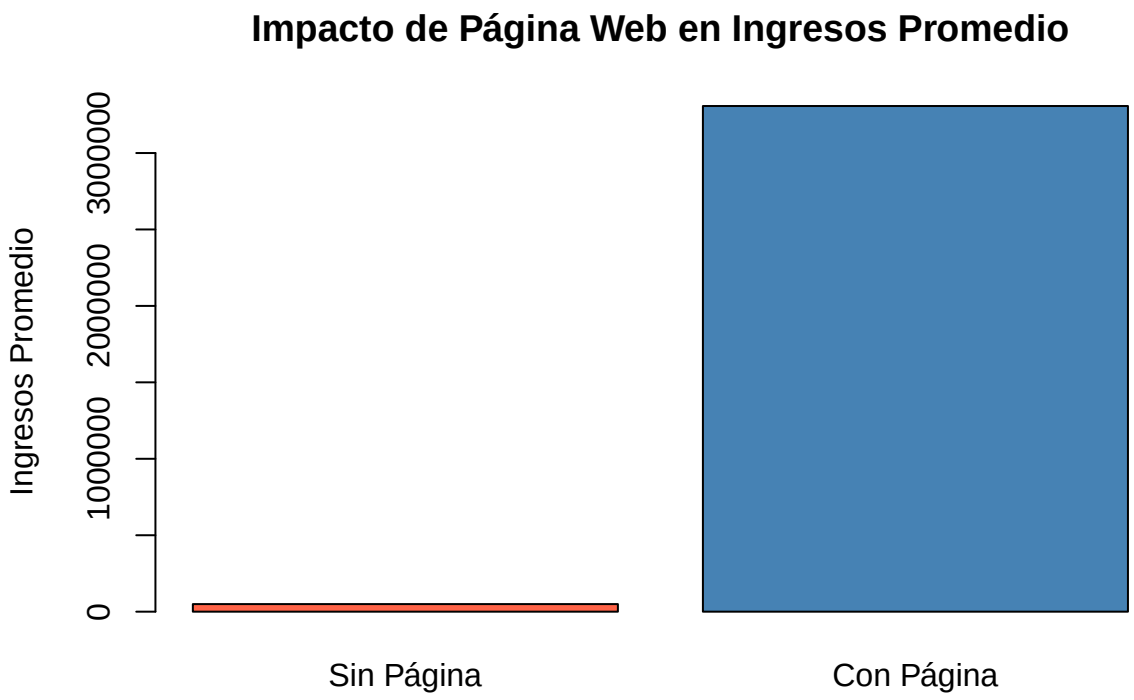
```
## Comparación de resultados según estrategia de marketing (Página Web):
```

```
print(marketing_analisis)
```

```
##   tiene_pagina    revenue popularity
## 1        FALSE   49531.52  0.3471507
## 2         TRUE 3307419.06  3.7511179
```

```
# 3. Visualización con un gráfico de barras para Ingresos
barplot(marketing_analisis$revenue,
```

```
        names.arg = c("Sin Página", "Con Página"),
        col = c("tomato", "steelblue"),
        main = "Impacto de Página Web en Ingresos Promedio",
        ylab = "Ingresos Promedio")
```



Impacto en Ingresos: Las películas que cuentan con una página oficial (TRUE) tienen un ingreso promedio de 3,307,419.06, comparado con los apenas 49,531.52 de las que no tienen (FALSE). Esto significa que tener una página web se asocia con ingresos 66 veces mayores en promedio.

Impacto en Popularidad: La popularidad promedio también salta de 0.34 a 3.75 cuando existe una página oficial, lo que representa un incremento de más de 10 veces en el interés del público.

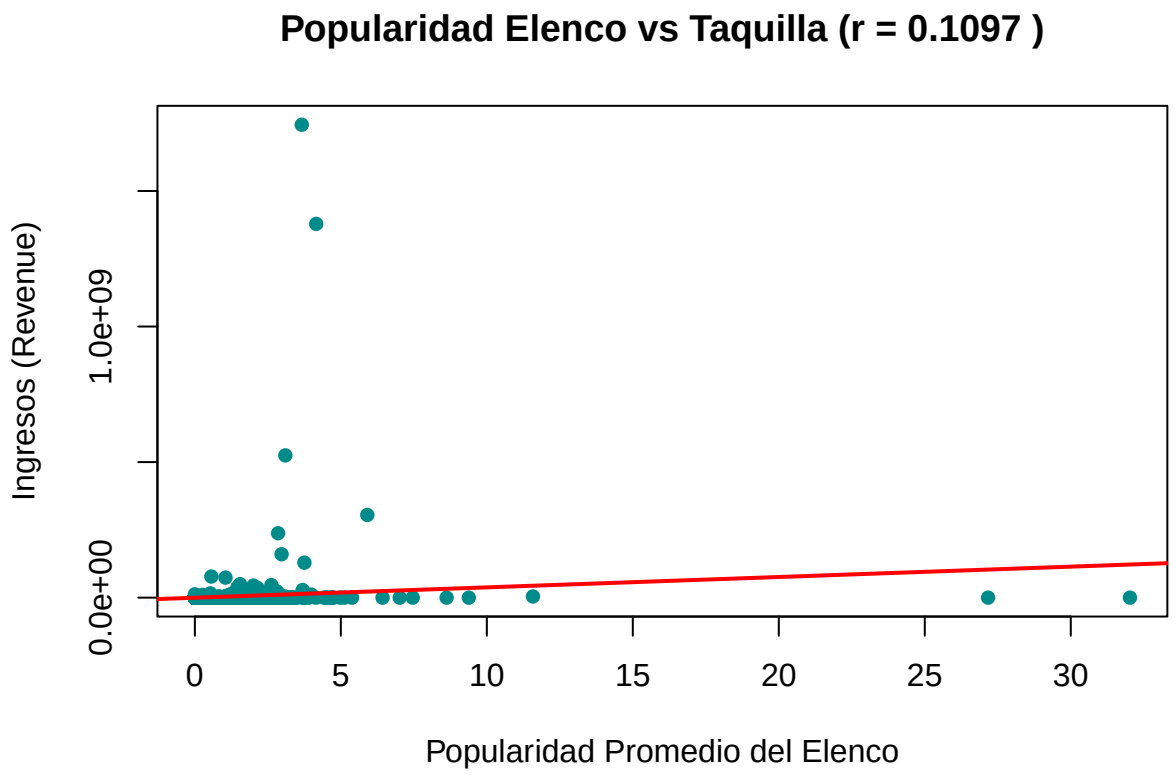
4.16 - ¿Correlación entre popularidad del elenco y éxito de taquilla?

```
# 1. Limpieza: La columna actorsPopularity contiene textos con números (ej. "10.5, 8.2").
# Extraeremos el promedio numérico de popularidad por película.
# Nota: Este paso asume que separaremos los números de la cadena de texto.

# Función rápida para obtener el promedio de la cadena de popularidad
get_avg_pop <- function(x) {
  nums <- as.numeric(unlist(regmatches(x, gregexpr("[0-9.]+", x))))
  if(length(nums) > 0) return(mean(nums)) else return(0)
}

movies$avg_actors_pop <- sapply(movies$actorsPopularity, get_avg_pop)
```

```r
# 2. Cálculo de la correlación de Pearson
cor_elenco_revenue <- cor(movies$avg_actors_pop, movies$revenue, use = "complete.obs")

cat("Correlación entre Popularidad del Elenco e Ingresos:", cor_elenco_revenue)
```

```
## Correlación entre Popularidad del Elenco e Ingresos: 0.1097298
```

```r
# 3. Gráfico de dispersión
plot(movies$avg_actors_pop, movies$revenue,
     main = paste("Popularidad Elenco vs Taquilla (r =", round(cor_elenco_revenue, 4), ")"),
     xlab = "Popularidad Promedio del Elenco",
     ylab = "Ingresos (Revenue)",
     col = "darkcyan", pch = 16)
abline(lm(revenue ~ avg_actors_pop, data = movies), col = "red", lwd = 2)
```



Existe una correlación muy débil (r = 0.1097). Aunque la línea de tendencia muestra una ligera inclinación positiva, el gráfico revela que muchos elencos con alta popularidad (puntos a la derecha en el eje X) siguen teniendo ingresos de cero.

IMPORTANTE Se identificó que la variable revenue presenta un sesgo masivo, con un 98.2% de valores en cero (9,713 registros). Por lo tanto, los análisis de rentabilidad y éxito comercial del dataset se basan únicamente en un pequeño subconjunto de películas (~1.8%) que sí reportaron ingresos significativos, concentrados principalmente en los meses de noviembre, diciembre y enero".

```r
# 1. Ver los 20 valores de ingresos más frecuentes
# cuántas películas tienen "0" o "1"
table(head(sort(table(movies$revenue), decreasing = TRUE), 20))
```

```
##
##    2    3    4    5    6    7   10   13   19 9713
##    9    3    1    1    1    1    1    1    1    1
```

```r
# 2. Ver las 10 películas con mayores ingresos (distintos a 0)
head(movies[order(-movies$revenue), c("title", "revenue")], 10)
```

```
##                         title    revenue
## 8482                Zootopia 2 1744338246
## 5908      Avatar: Fire and Ash 1378692505
## 9845          Wicked: For Good  524676531
## 5723             The Housemaid  305000000
## 5621 Five Nights at Freddy's 2  237625385
## 5353                Dhurandhar  160000000
## 4195                  Anaconda  129019155
## 4901                     David   77770275
## 3106               Buen Camino   73797878
## 5125                Gezhi Town   49627843
```

```r
# 3. Resumen estadístico detallado
summary(movies$revenue)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 0.000e+00 0.000e+00 5.168e+05 0.000e+00 1.744e+09
```