lab-01

Juan Menéndez - 17444, Luis Mendoza - 19644, José Antón - 221041

2026-02-06

Nota sobre el uso de IA: Para evitar malos entendidos y confunciones se aclara el método de uso de Inteligencia Artificial en el siguiente trabajo. Debido a qué el método de trabajo empleado fue de manera separadas, se utilizó la herramienta de Microsoft Copilot para unificar la estructura de las respuestas, más se le pidió explicitamente que no tocara, remplazara ni alterara el código ni el análisis escrito. Así cómo para la verificación de faltas órtograficas que pudieran presentarse a lo largo del informe.

Individualmente se pudo utilizar a manera de consulta y para resolver dudas, así cómo la herramienta interna de autocompletado en Rstudio.

1 Exploración rápida (resumen del dataset)

```
##        id              budget             genres            homePage
##  Min.   :      5   Min.   :        0   Length:19883       Length:19883
##  1st Qu.: 146220   1st Qu.:        0   Class :character   Class :character
##  Median : 869623   Median :        0   Mode  :character   Mode  :character
##  Mean   : 902240   Mean   :  9413280
##  3rd Qu.:1589602   3rd Qu.:  1000000
##  Max.   :1627166   Max.   :380000000
##
##  productionCompany  productionCompanyCountry productionCountry
##  Length:19883       Length:19883             Length:19883
##  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:19883
##  1st Qu.:0.000e+00   1st Qu.: 10.00   FALSE:19313     Class :character
##  Median :0.000e+00   Median : 86.00   TRUE :84        Mode  :character
##  Mean   :2.879e+07   Mean   : 66.09   NA's :486
##  3rd Qu.:3.306e+05   3rd Qu.:103.00
##  Max.   :2.847e+09   Max.   :750.00
##
##     actors          actorsPopularity   actorsCharacter    originalTitle
##  Length:19883       Length:19883       Length:19883       Length:19883
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     title           originalLanguage     popularity        releaseDate
##  Length:19883       Length:19883       Min.   :0.000e+00   Length:19883
##  Class :character   Class :character   1st Qu.:5.460e-02   Class :character
##  Mode  :character   Mode  :character   Median :8.502e+00   Mode  :character
##                                        Mean   :2.625e+01
##                                        3rd Qu.:2.224e+01
##                                        Max.   :1.147e+04
##
##     voteAvg         voteCount         genresAmount    productionCoAmount
##  Min.   : 0.000   Min.   :    0.0   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:    0.0   1st Qu.: 1.000   1st Qu.: 0.000
##  Median : 5.400   Median :    6.0   Median : 2.000   Median : 1.000
##  Mean   : 3.837   Mean   :  675.9   Mean   : 1.949   Mean   : 1.973
##  3rd Qu.: 6.800   3rd Qu.:  423.0   3rd Qu.: 3.000   3rd Qu.: 3.000
##  Max.   :10.000   Max.   :30788.0   Max.   :16.000   Max.   :89.000
```

```
##
##  productionCountriesAmount  actorsAmount    castWomenAmount  castMenAmount
##  Min.   :  0.00            Min.   :     0   Min.   :     0   Min.   :     0
##  1st Qu.:  1.00            1st Qu.:     3   1st Qu.:     0   1st Qu.:     0
##  Median :  1.00            Median :     9   Median :     2   Median :     3
##  Mean   :  1.23            Mean   :  1082   Mean   :  3517   Mean   :  8224
##  3rd Qu.:  1.00            3rd Qu.:    21   3rd Qu.:     6   3rd Qu.:    12
##  Max.   :155.00            Max.   :919590   Max.   :922162   Max.   :922017
##                                             NA's   :37       NA's   :162
##   releaseYear
##  Min.   :1902
##  1st Qu.:2013
##  Median :2021
##  Mean   :2017
##  3rd Qu.:2025
##  Max.   :2026
##  NA's   :2
```

El resumen muestra que hay muchas películas con presupuesto e ingresos en 0, lo cual sugiere datos faltantes o registros sin información financiera. También se observa que `video` es mayormente FALSE y que `homePage` tiene bastantes valores vacíos. Esto es importante porque, antes de modelar, conviene decidir cómo tratar ceros y valores faltantes.

2 Tipo de cada variable

variable	tipo	descripcion
id	Cualitativa nominal	Identificador; numérico pero no mide magnitud
popularity	Cuantitativa continua	Índice (escala continua)
budget	Cuantitativa continua	Presupuesto (moneda)
revenue	Cuantitativa continua	Ingresos (moneda)
originalTitle	Cualitativa nominal	Texto
originalLanguage	Cualitativa nominal	Código de idioma
title	Cualitativa nominal	Texto (título en inglés)
homePage	Cualitativa nominal	URL (alta cardinalidad, muchos NA)
video	Cualitativa nominal	Binaria (sí/no)
director	Cualitativa nominal	Nombre (alta cardinalidad)
runtime	Cuantitativa continua	Duración (minutos; medida de tiempo)
genres	Cualitativa nominal	Multi-etiqueta (separada por coma o ‘|’)
genresAmount	Cuantitativa discreta	Conteo de géneros
productionCompany	Cualitativa nominal	Texto / multi-etiqueta (según fuente)
productionCoAmount	Cuantitativa discreta	Conteo de compañías productoras
productionCompanyCountry	Cualitativa nominal	Multi-etiqueta (códigos de país)
productionCountry	Cualitativa nominal	Multi-etiqueta (códigos de país)
productionCountriesAmount	Cuantitativa discreta	Conteo de países

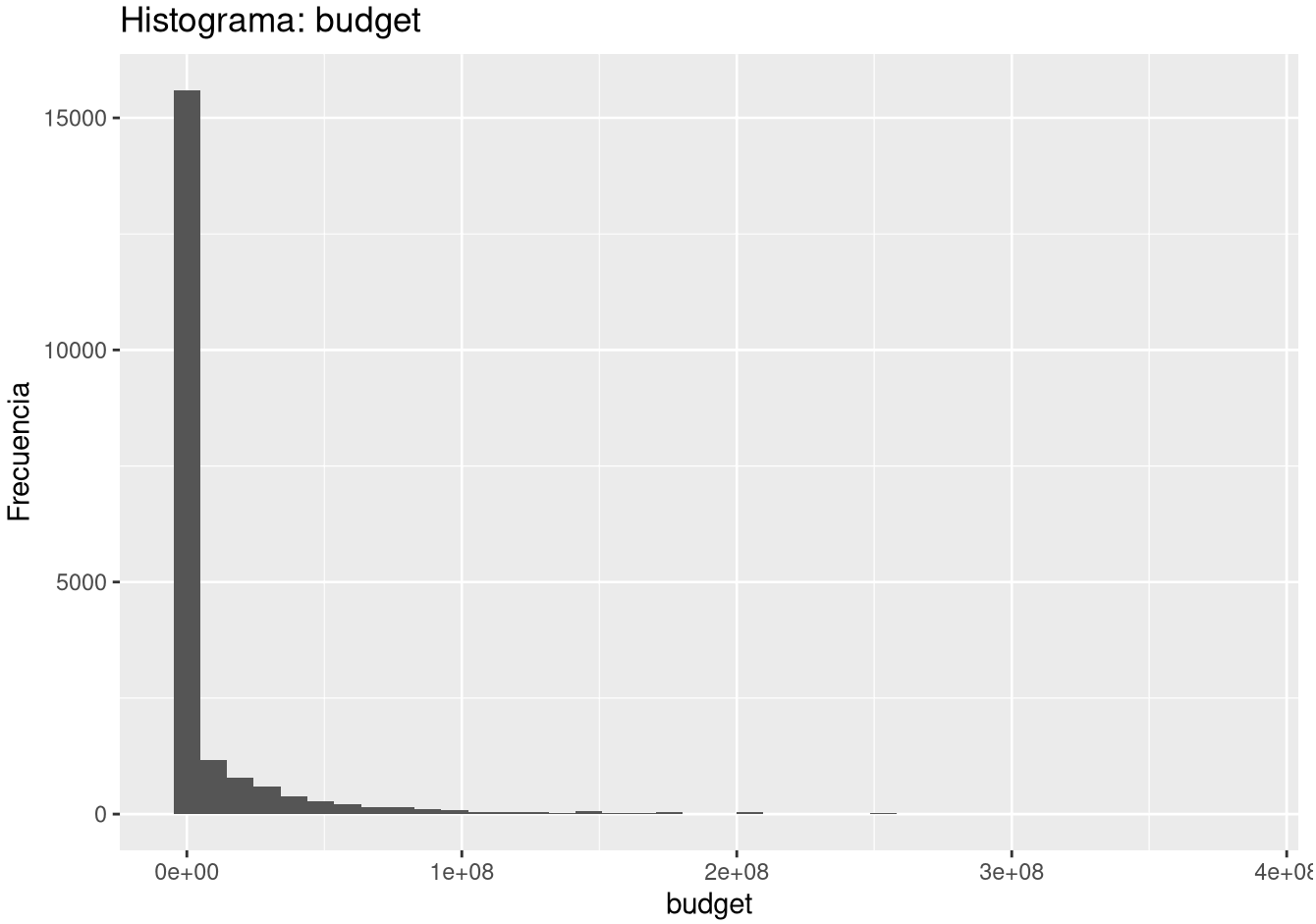
variable	tipo	descripcion
releaseDate	Cualitativa ordinal	Fecha (tiene orden temporal)
voteCount	Cuantitativa discreta	Conteo de votos
voteAvg	Cuantitativa continua	Promedio (0–10)
actors	Cualitativa nominal	Lista de actores (texto, separada por coma o ‘|’)
actorsPopularity	Cuantitativa continua	Popularidad del elenco (numérica o derivable)
actorsCharacter	Cualitativa nominal	Lista de personajes (texto)
actorsAmount	Cuantitativa discreta	Conteo de actores
castWomenAmount	Cuantitativa discreta	Conteo de actrices
castMenAmount	Cuantitativa discreta	Conteo de actores
releaseYear	Cuantitativa discreta	Año

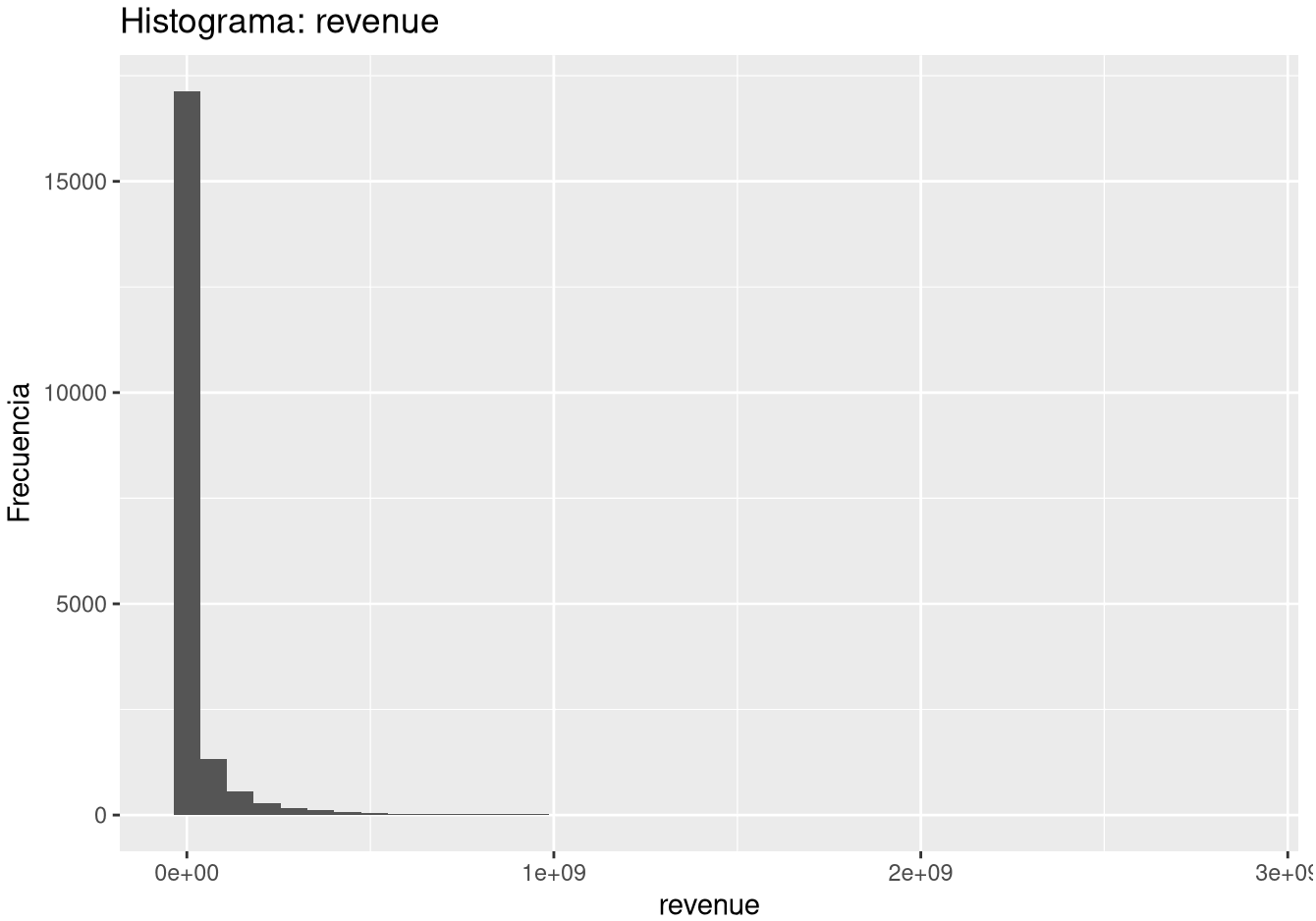
Clasificar las variables ayuda a saber qué operaciones tienen sentido: por ejemplo, `id` se ve numérico pero es un identificador, mientras que `voteCount` es discreta y `voteAvg` es continua. También hay campos tipo texto que representan listas (como `genres o actors`), que requieren limpieza antes de usarlos. Esta tabla sirve como guía para escoger transformaciones y gráficos adecuados.

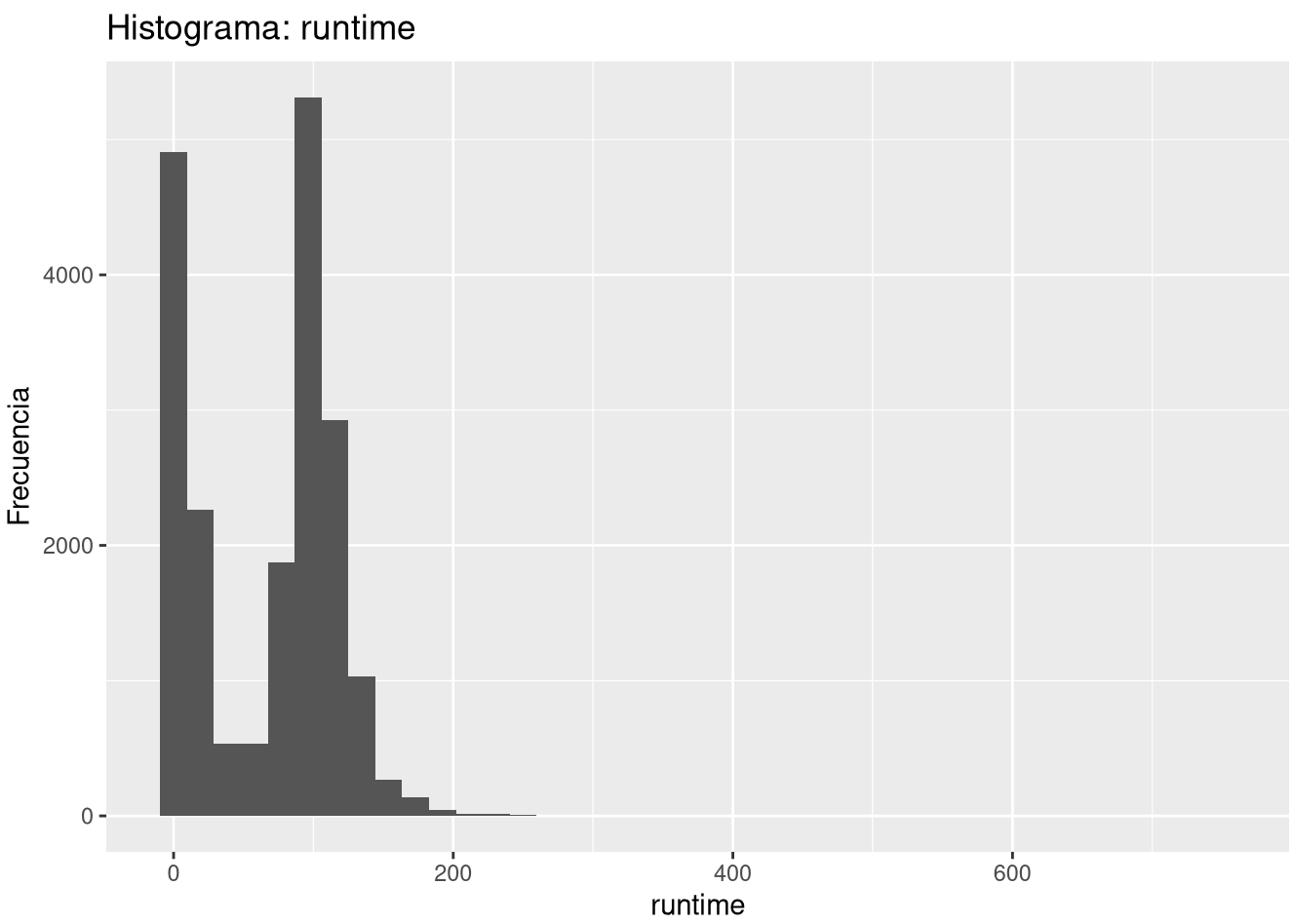
3 Normalidad en variables cuantitativas y tabla de

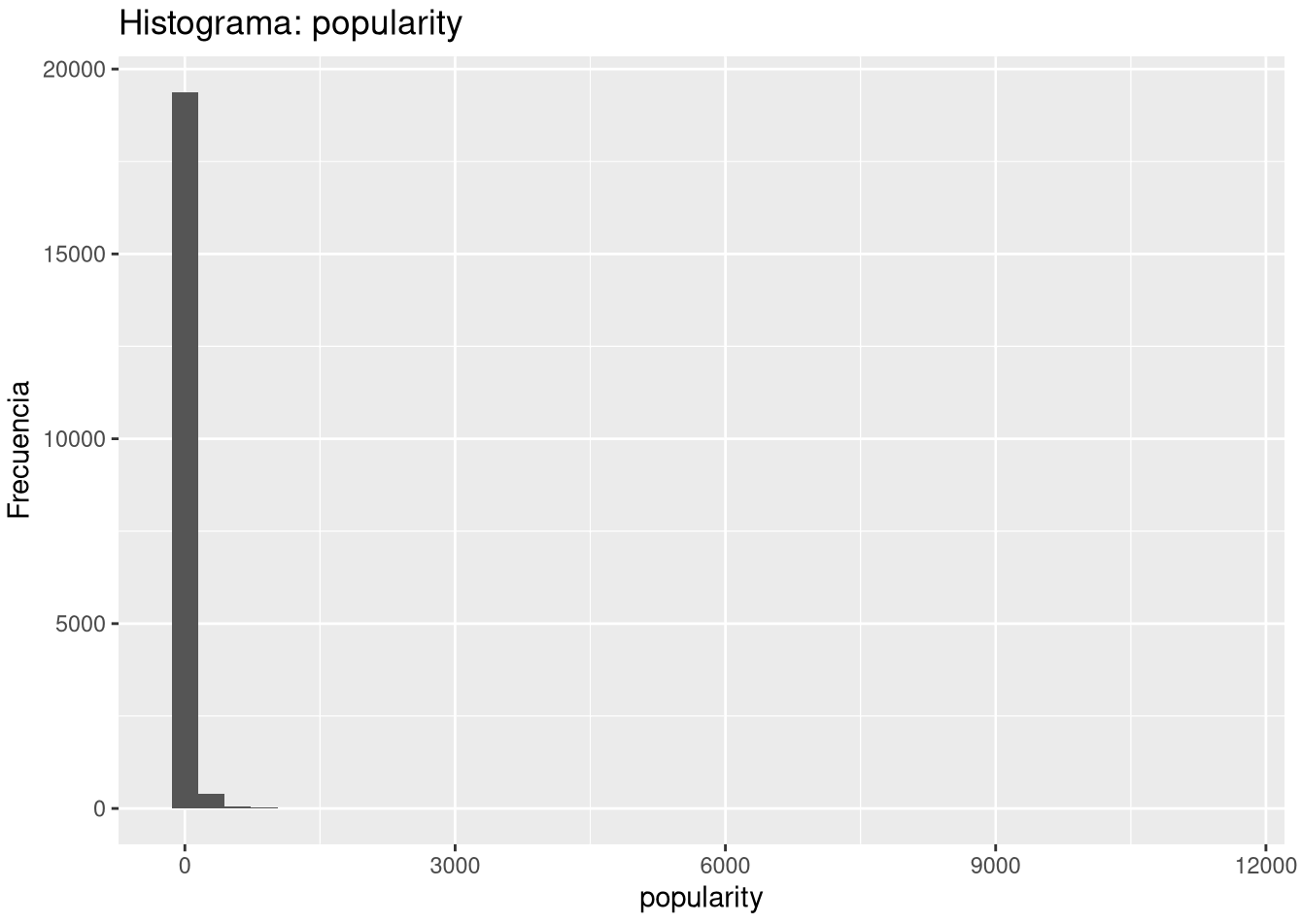
Prueba Shapiro-Wilk (muestra) para variables numéricas

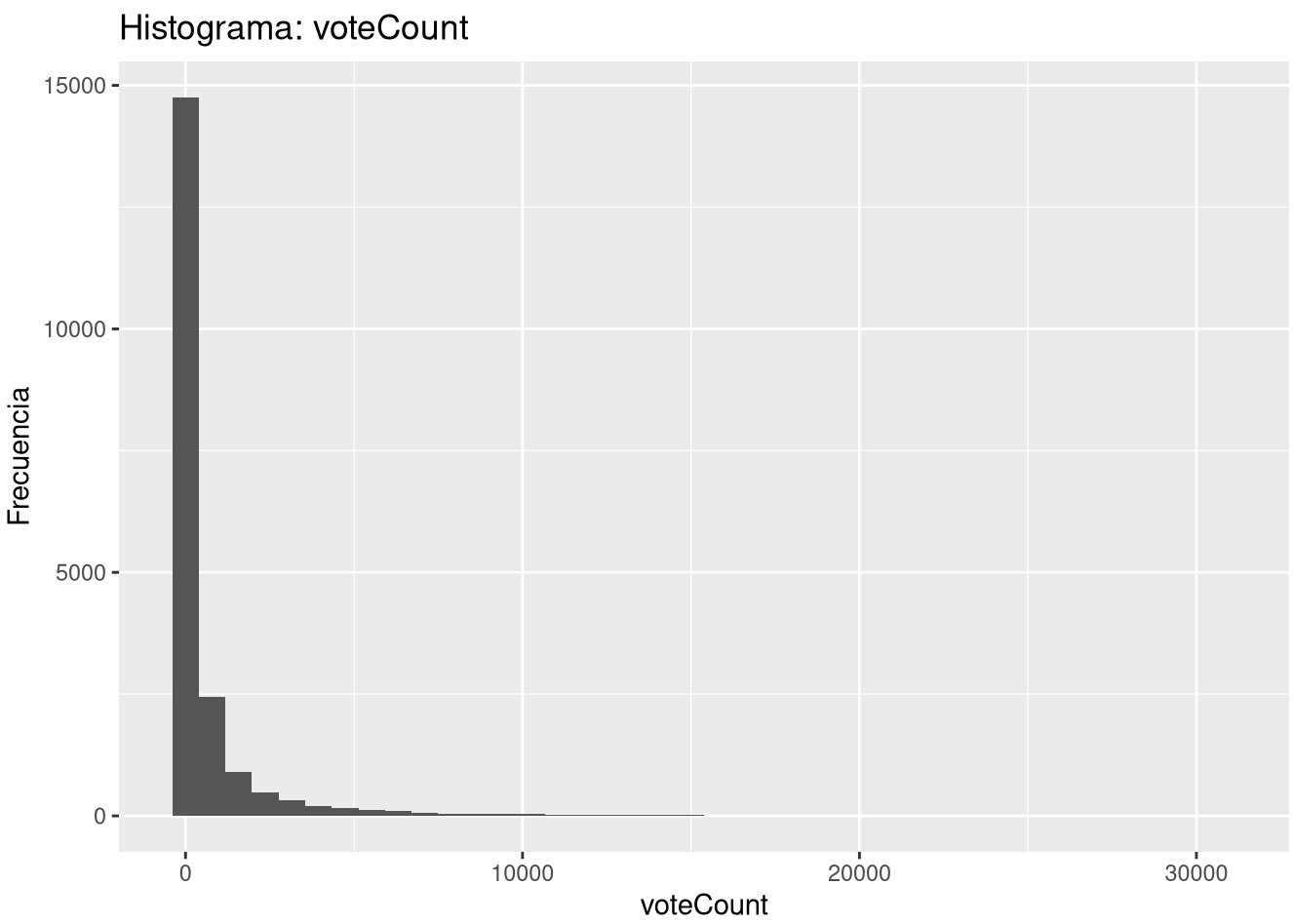
variable	n	p_value
actorsAmount	5000	0
castWomenAmount	4993	0
popularity	5000	0
castMenAmount	4968	0
revenue	5000	0
productionCountriesAmount	5000	0
voteCount	5000	0
budget	5000	0
releaseYear	5000	0
id	5000	0
productionCoAmount	5000	0
voteAvg	5000	0
runtime	5000	0
genresAmount	5000	0

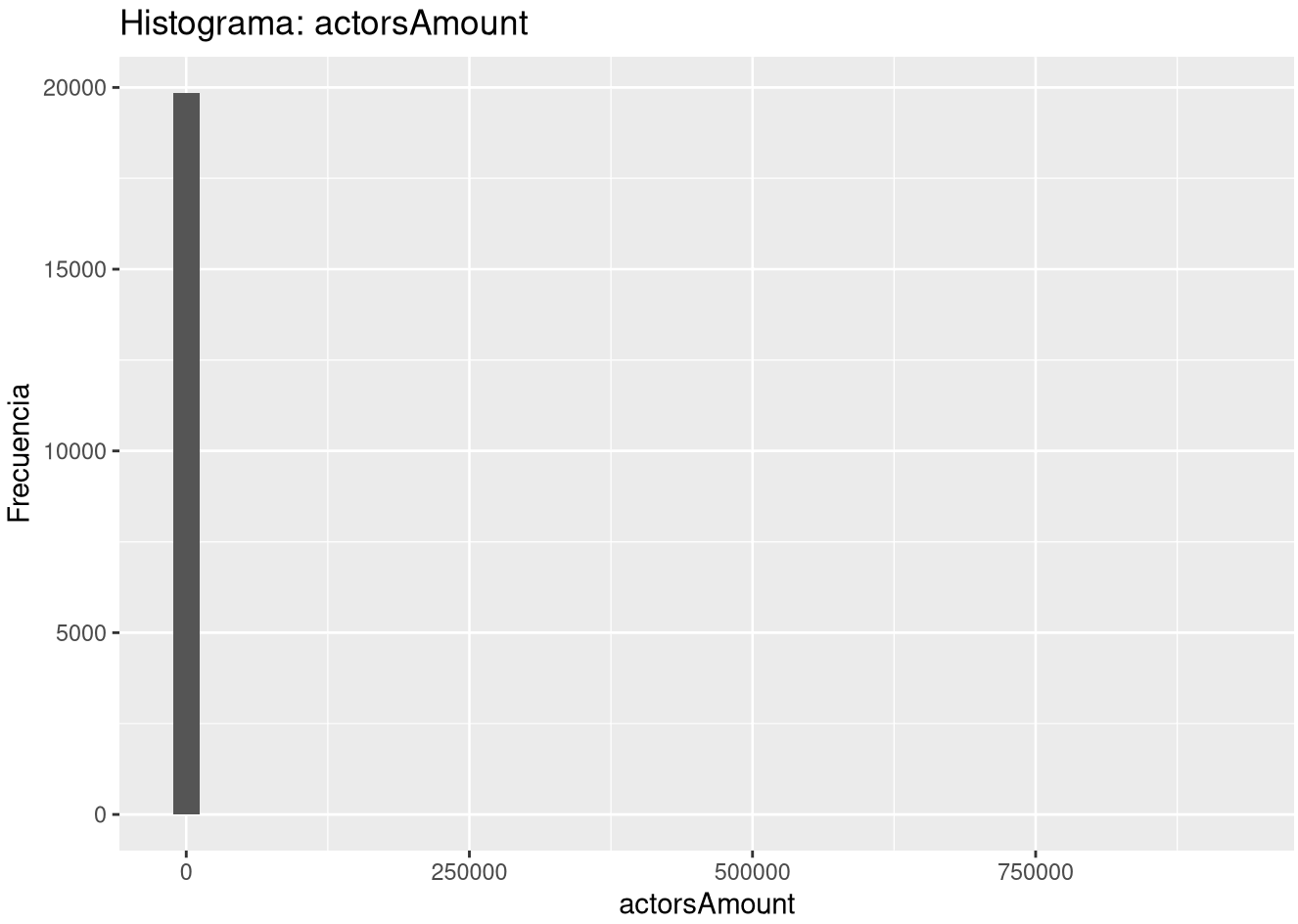

Histograma: budget
15000
10000
5000
0
Frecuencia
0e+00
1e+08
2e+08
3e+08
budget


Histograma: revenue
15000
10000
5000
0
Frecuencia
0e+00
1e+09
2e+09
revenue


Histograma: runtime
Frecuencia
4000
2000
0
0
200
400
600
runtime


Histograma: popularity
Frecuencia
20000
15000
10000
5000
0
0
3000
6000
9000
12000
popularity


Histograma: voteCount
15000
10000
5000
0
Frecuencia
0
10000
20000
30000
voteCount


Histograma: actorsAmount
20000
15000
10000
5000
0
Frecuencia
0
250000
500000
750000
actorsAmount

```
##                   variable                       top  freq unicos missing
## 1                    video                     FALSE 19313      3     486
## 2                 homePage                      <NA> 14276   5488   14276
## 3          originalLanguage                        en 11961     94       0
## 4  productionCompanyCountry                      <NA>  8410   2745    8410
## 5         productionCompany                      <NA>  5656  12235    5656
## 6         productionCountry United States of America  4968   1407    3874
## 7           actorsCharacter                      <NA>  2835  14969    2835
## 8          actorsPopularity                      <NA>  2608  15135    2608
## 9                    actors                      <NA>  2601  16734    2601
## 10                   genres                      <NA>  1965   2828    1965
## 11                 director                      <NA>   976  13669     976
## 12             releaseDate                2026-01-30   263   5664       2
## 13                    title                Cinderella     5  19384       0
## 14            originalTitle                Cinderella     4  19506       0
```

La prueba de Shapiro-Wilk en una muestra da p-valores ~0 para las variables numéricas, así que no parecen normales. Los histogramas refuerzan que varias distribuciones están muy sesgadas (por ejemplo presupuesto e ingresos). En las cualitativas se ven muchos NA y alta cardinalidad (como `homePage` o `director`), lo cual afecta conteos y agrupaciones. En general, para análisis posteriores conviene usar transformaciones como log y métricas robustas (mediana) en vez de asumir normalidad.

4.1 ¿Cuáles son las 10 películas que contaron con más

```
##                                      title   budget releaseYear
## 1  Pirates of the Caribbean: On Stranger Tides 3.80e+08        2011
## 2                      Avengers: Age of Ultron 3.65e+08        2015
## 3                            Avengers: Endgame 3.56e+08        2019
## 4                          Avatar: Fire and Ash 3.50e+08        2025
## 5     Pirates of the Caribbean: At World's End 3.00e+08        2007
## 6                               Justice League 3.00e+08        2017
## 7                       Avengers: Infinity War 3.00e+08        2018
## 8                             Superman Returns 2.70e+08        2006
## 9                                      Tangled 2.60e+08        2010
## 10                               The Lion King 2.60e+08        2019
##                    director
## 1               Rob Marshall
## 2                Joss Whedon
## 3    Anthony Russo|Joe Russo
## 4               James Cameron
## 5              Gore Verbinski
## 6                 Zack Snyder
## 7    Anthony Russo|Joe Russo
## 8                Bryan Singer
## 9  Byron Howard|Nathan Greno
## 10               Jon Favreau
```

Las películas con mayor presupuesto son principalmente franquicias y producciones muy grandes (por ejemplo Pirates y Avengers). Los valores llegan a cientos de millones, lo que las pone fuera del rango típico del resto del dataset. Esto sugiere una distribución muy sesgada y la presencia de outliers. En análisis posteriores, usar escala log ayuda a comparar mejor presupuestos.

4.2 ¿Cuáles son las 10 películas que más ingresos

```
##                          title    revenue   budget releaseYear
## 1                       Avatar 2847246203 2.37e+08        2009
## 2            Avengers: Endgame 2797800564 3.56e+08        2019
## 3                      Titanic 2187463944 2.00e+08        1997
## 4  Star Wars: The Force Awakens 2068223624 2.45e+08        2015
## 5       Avengers: Infinity War 2046239637 3.00e+08        2018
## 6                   Zootopia 2 1744338246 1.50e+08        2025
## 7               Jurassic World 1671713208 1.50e+08        2015
## 8                The Lion King 1667635327 2.60e+08        2019
## 9      Spider-Man: No Way Home 1631853496 2.00e+08        2021
## 10                The Avengers 1518815515 2.20e+08        2012
##                   director
## 1            James Cameron
## 2  Anthony Russo|Joe Russo
## 3            James Cameron
## 4              J.J. Abrams
## 5  Anthony Russo|Joe Russo
## 6  Jared Bush|Byron Howard
## 7          Colin Trevorrow
## 8              Jon Favreau
## 9                Jon Watts
## 10             Joss Whedon
```

En ingresos, aparecen títulos muy conocidos como Avatar, Avengers: Endgame y Titanic, con ingresos de miles de millones. Aunque el presupuesto influye, no todas las películas con presupuesto alto necesariamente están en el top de revenue. Esto indica que hay otros factores (popularidad, marketing, franquicia, etc.) además del gasto. También refuerza que revenue tiene outliers extremos.

4.3 ¿Cuál es la película que más votos tuvo?

La película con más votos es Inception con 30788 votos y una calificación de 8.4. Tener muchos votos sugiere que es una película muy vista o muy discutida, por lo que su promedio es más confiable. En general, `voteCount` puede usarse como un indicador de confianza para `voteAvg` . Para comparaciones justas, conviene filtrar por un mínimo de votos.

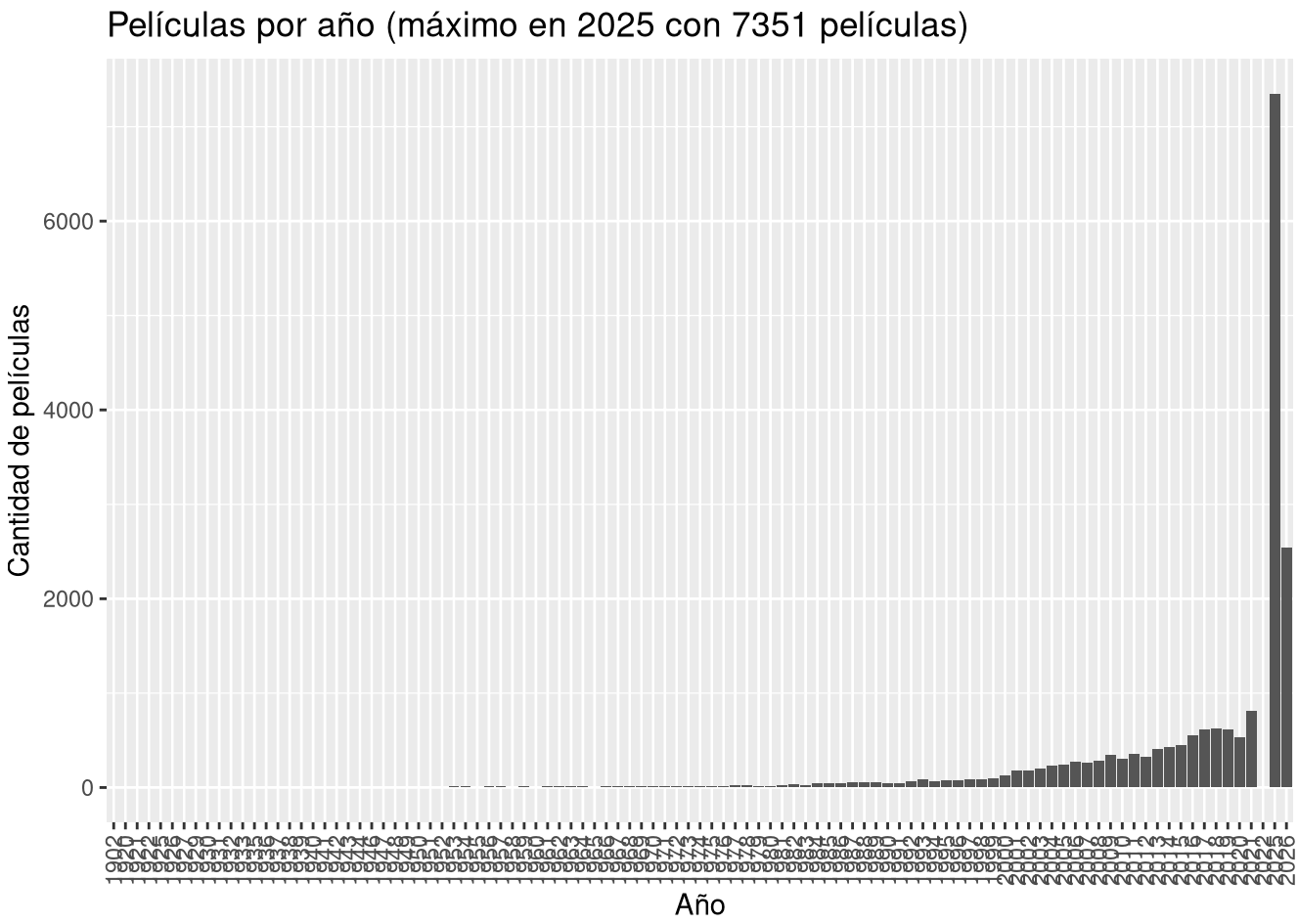
4.4 ¿Cuál es la peor película de acuerdo a los votos de

```
## $peor_sin_filtro
##                   title voteAvg voteCount
## 1 The Halloween Harvest       0         1
##
## $peor_con_umbral
##        title voteAvg voteCount
## 1 Death Note     4.2      3388
```

Sin filtrar, la peor calificación sale con un caso de 1 voto (The Halloween Harvest con 0), lo cual no es muy representativo. Al usar un umbral de votos, aparece Death Note con 4.2 y 3388 votos, que es un resultado más estable. Esto muestra por qué `voteAvg` debe analizarse junto con `voteCount` . En minería de datos, este tipo de filtro evita conclusiones por ruido.

4.5 ¿Cuántas películas se hicieron en cada año? ¿En qué

```
##    releaseYear peliculas
## 1         2025      7351
## 2         2026      2537
## 3         2021       814
## 4         2018       628
## 5         2017       617
## 6         2019       611
## 7         2016       557
## 8         2020       531
## 9         2015       450
## 10        2014       432
```



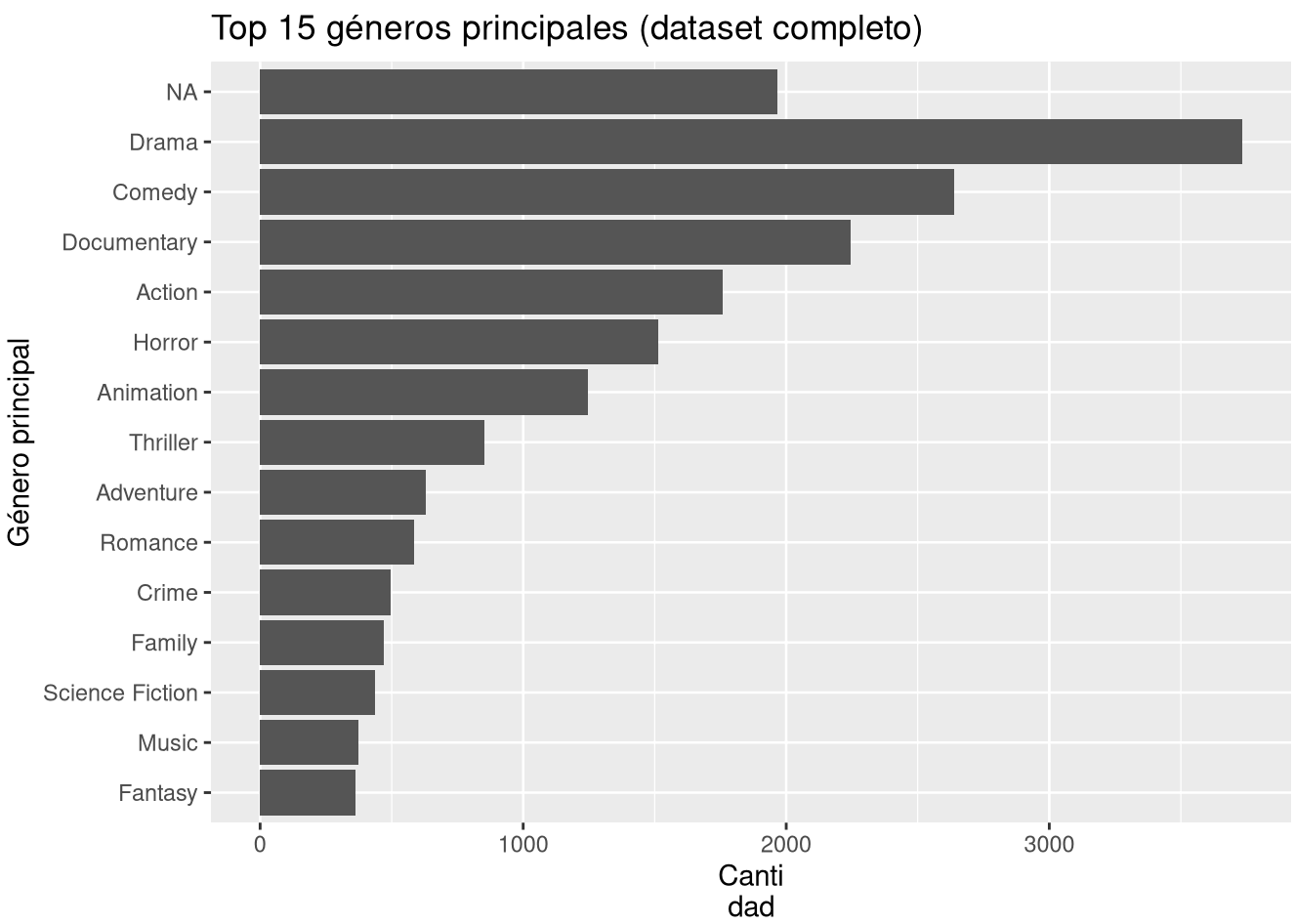
El conteo por año muestra que 2025 es el año con más películas, con 7351 registros. Esto puede deberse a cómo se recolectó el dataset o a que incluye muchos lanzamientos recientes. Como hay una concentración fuerte en pocos años, las comparaciones históricas pueden quedar sesgadas. El gráfico ayuda a ver rápidamente esa concentración temporal.

4.6 Género principal de las 20 películas más recientes,

```
##                                           title releaseDate  main_genre
## 1                                   A Fading Man  2026-05-07       Drama
## 2        Elon Musk Unveiled – The Tesla Experiment  2026-03-12 Documentary
## 3                                           Skunk  2026-02-25        <NA>
## 4                                       Anastasia  2026-02-25        <NA>
## 5                                Nikki hako no koi  2026-02-06        <NA>
## 6                                        Immersed  2026-02-01       Drama
## 7                                      Cinderella  2026-02-01   Animation
## 8                                        Aladdin  2026-02-01   Animation
## 9                           THE RING AND THE DECK  2026-02-01    Thriller
## 10                                  Crimson High 3  2026-02-01   Animation
## 11 Conversations with Rasparagus Asparagus Baragus  2026-02-01      Comedy
## 12                                Highway To Hell  2026-02-01      Comedy
## 13                                Pari's daughter  2026-02-01       Drama
## 14                                         Escort  2026-02-01      Action
## 15                                          Dream  2026-02-01       Drama
## 16                                         Lively  2026-02-01       Drama
## 17                           01-02-2026  قایق‌سواری در تهران     Romance
## 18                                       Midnight  2026-02-01         War
## 19            Emir - Posljednji dalmatinski težak  2026-02-01        <NA>
## 20                             Our Dead Husband  2026-02-01    Thriller
```

```
##    main_genre n
## 1       Drama 5
## 2        <NA> 4
## 3   Animation 3
## 4      Comedy 2
## 5    Thriller 2
## 6      Action 1
## 7 Documentary 1
## 8     Romance 1
## 9         War 1
```

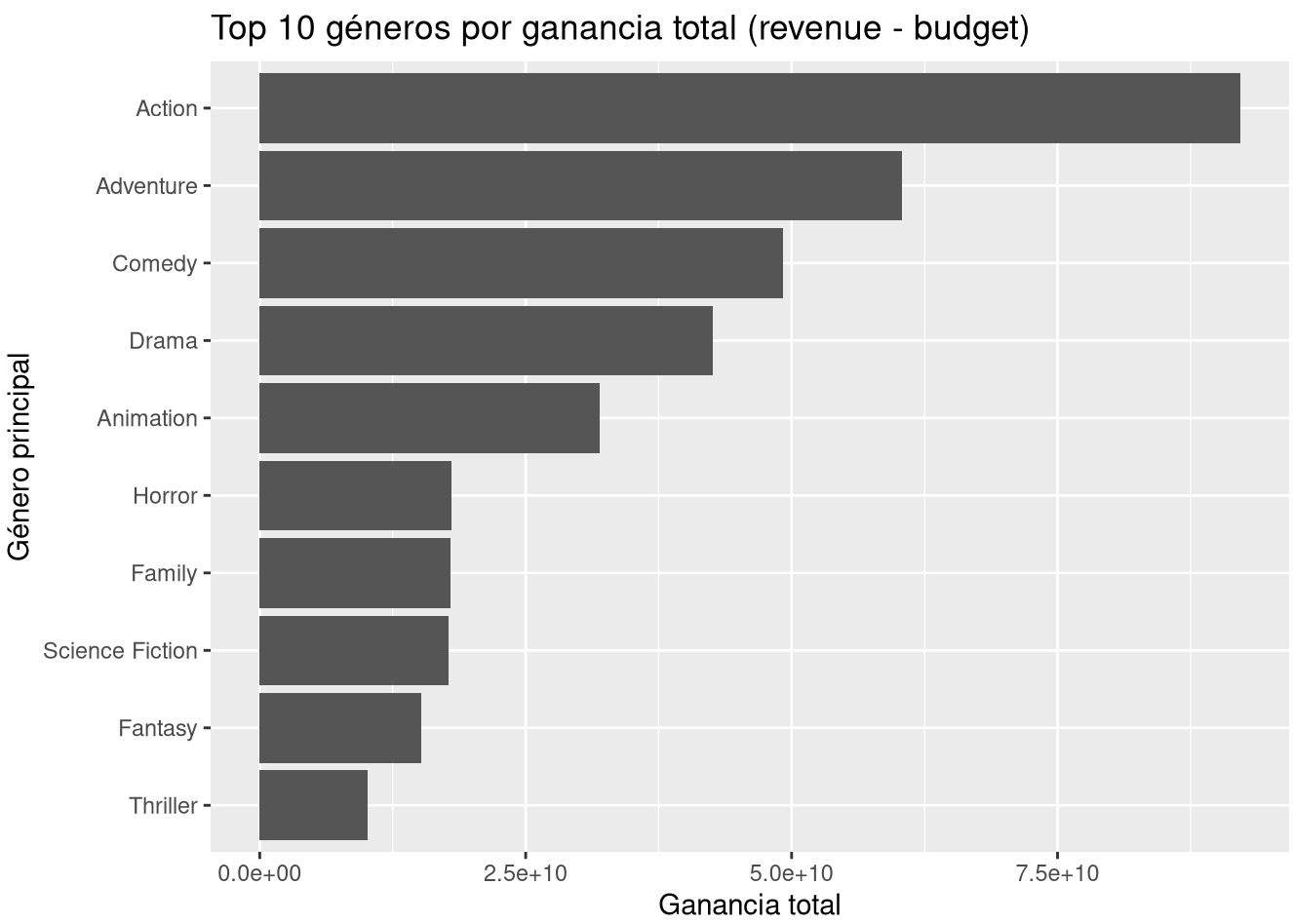
```
##         main_genre    n
## 1            Drama 3734
## 2           Comedy 2640
## 3      Documentary 2245
## 4             <NA> 1965
## 5           Action 1760
## 6           Horror 1513
## 7        Animation 1247
## 8         Thriller  854
## 9        Adventure  629
## 10         Romance  585
## 11           Crime  497
## 12          Family  469
## 13 Science Fiction  437
## 14           Music  373
## 15         Fantasy  364
```



Al revisar las 20 películas más recientes, aparecen géneros como Drama y Documentary, y también varios casos sin género (`NA`). Esto sugiere que el campo `genres` no está completo para algunos lanzamientos recientes. Si se quiere analizar tendencias recientes por género, primero habría que limpiar o imputar esos `NA` . También se ve que la fecha de lanzamiento se usa como base para ordenar y comparar.

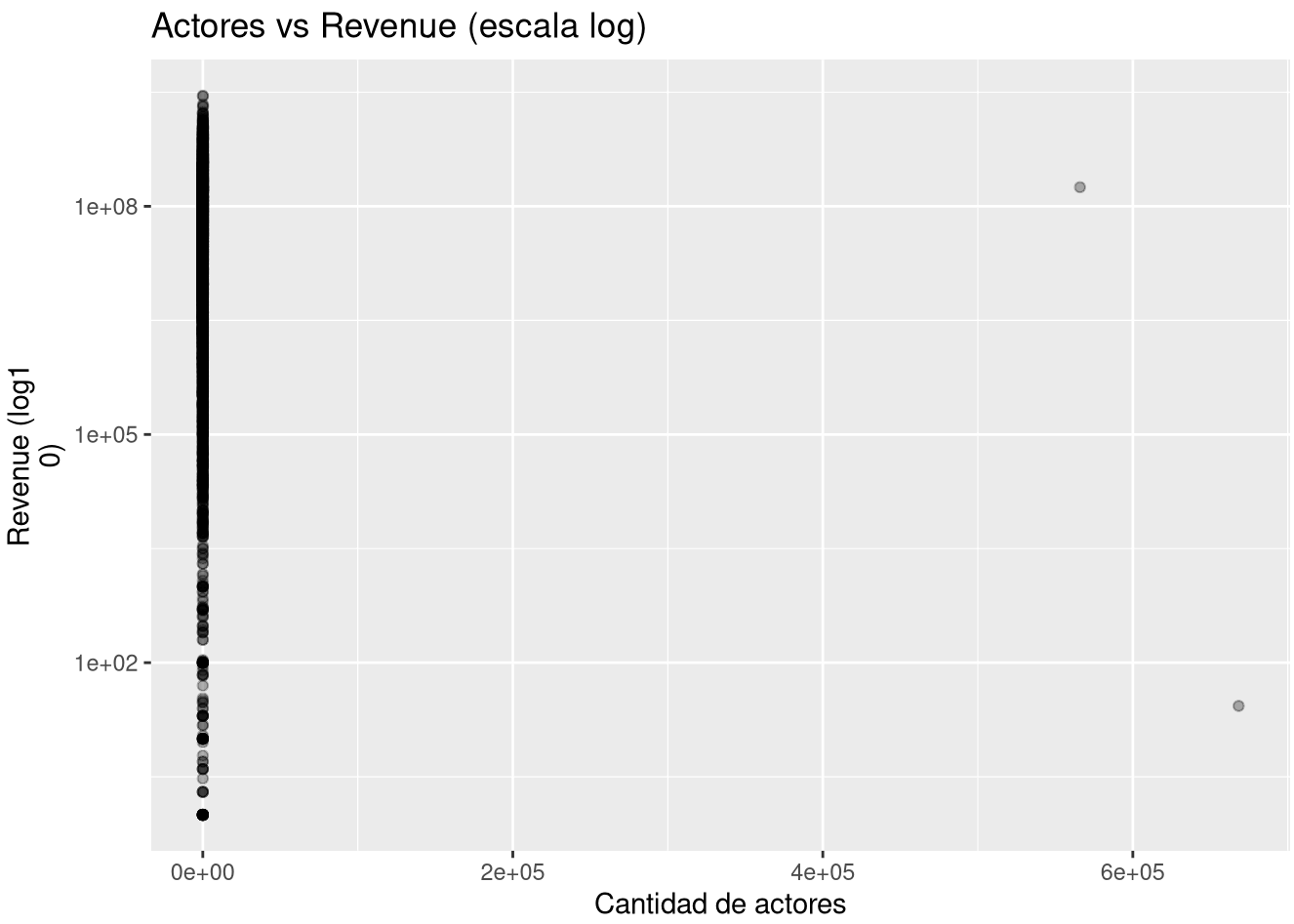
4.7 ¿Las películas de qué género principal obtuvieron

```
## # A tibble: 15 × 4
##    main_genre      peliculas ganancia_total ganancia_promedio
##    <chr>               <int>          <dbl>             <dbl>
##  1 Action                779    92195613009        118351236.
##  2 Adventure             351    60354070276        171948918.
##  3 Comedy                805    49220834911         61143894.
##  4 Drama                 860    42588500938         49521513.
##  5 Animation             181    32005682028        176826973.
##  6 Horror                373    18026440092         48328258.
##  7 Family                132    17978264869        136198976.
##  8 Science Fiction       145    17757646253        122466526.
##  9 Fantasy               137    15146197026        110556183.
## 10 Thriller              201    10151243101         50503697.
## 11 Crime                 211     8254157389         39119229.
## 12 Romance               106     6325555484         59675052.
## 13 War                    41     3379441894         82425412.
## 14 Mystery                60     3378308280         56305138 
## 15 Music                  35     2174083494         62116671.
```

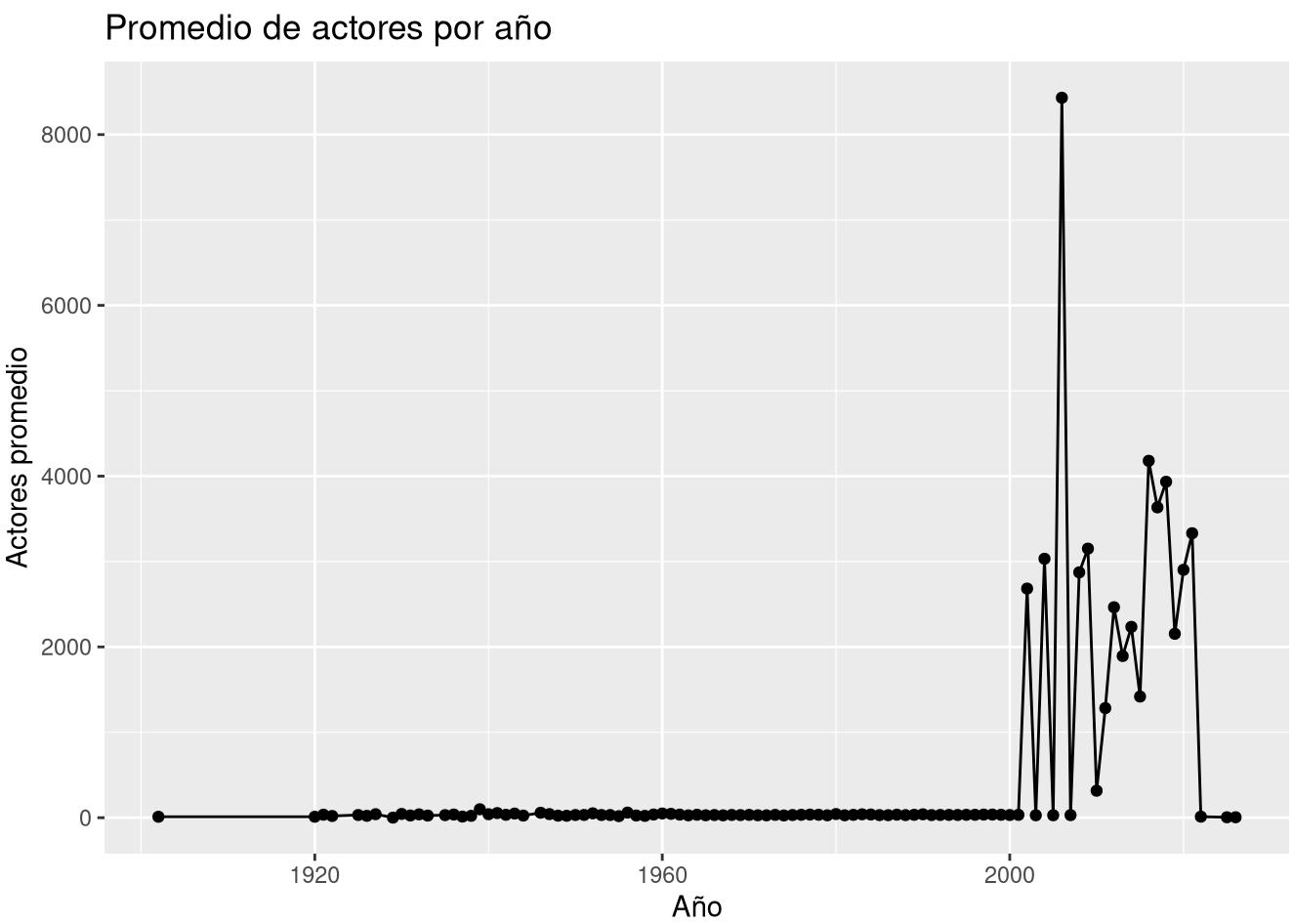


Al agrupar por género principal, Action y Adventure tienen ganancias totales altas, y Animation muestra un promedio de ganancia muy alto. La diferencia entre ganancia total y promedio es importante: un género puede tener mucho total por cantidad de películas. Este resultado ayuda a identificar géneros que, en promedio, son más rentables. Para decisiones de negocio, conviene mirar ambos: volumen y retorno promedio.

4.8 ¿La cantidad de actores influye en los ingresos? ¿Se



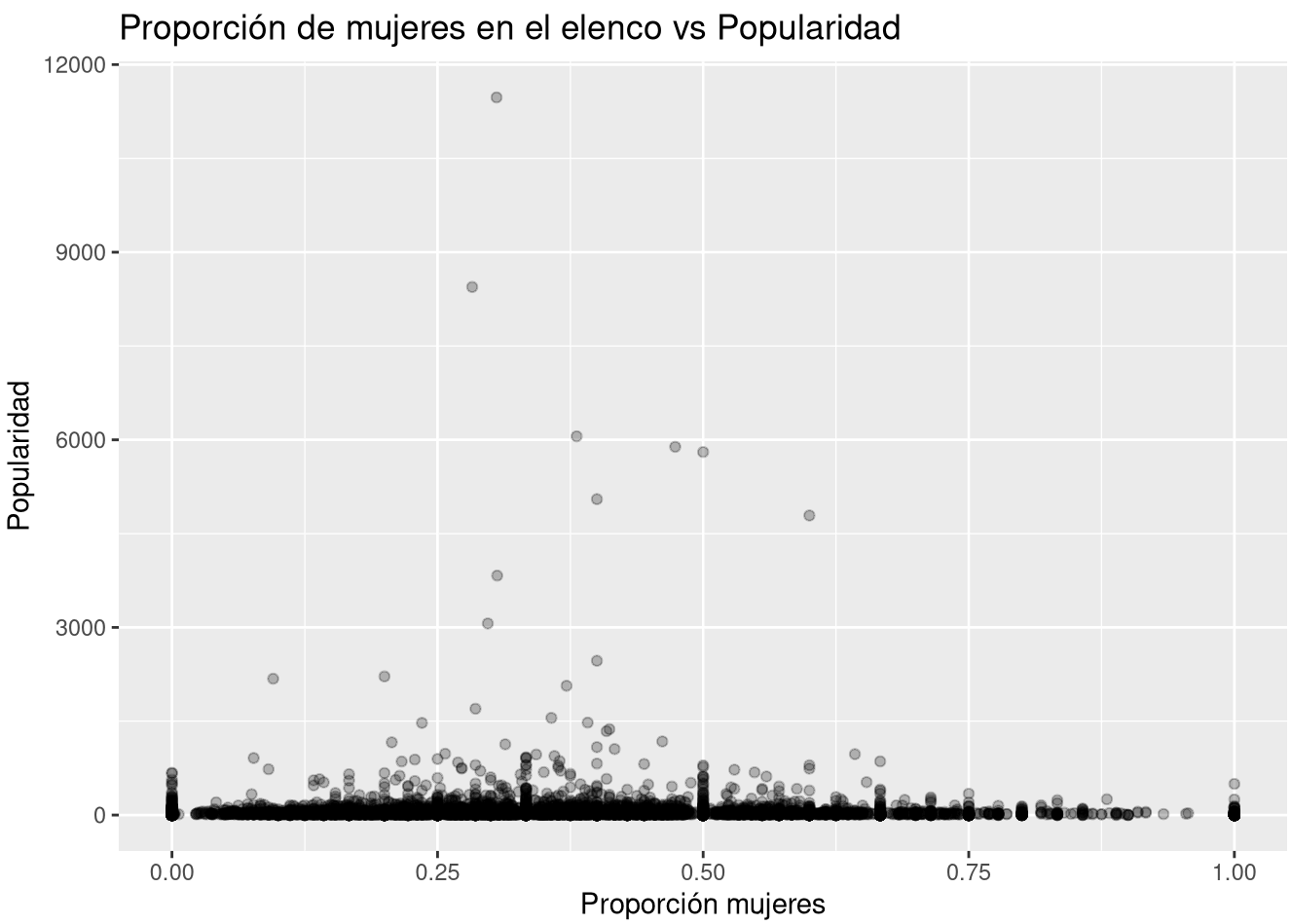

```
## # A tibble: 10 × 3
##    releaseYear actores_prom peliculas
##          <int>        <dbl>     <int>
##  1        2015      1419.         445
##  2        2016      4179.         548
##  3        2017      3634.         611
##  4        2018      3934.         623
##  5        2019      2154.         599
##  6        2020      2903.         520
##  7        2021      3331.         781
##  8        2022        12.1          7
##  9        2025         5.10      5412
## 10        2026         5.22      1955
```

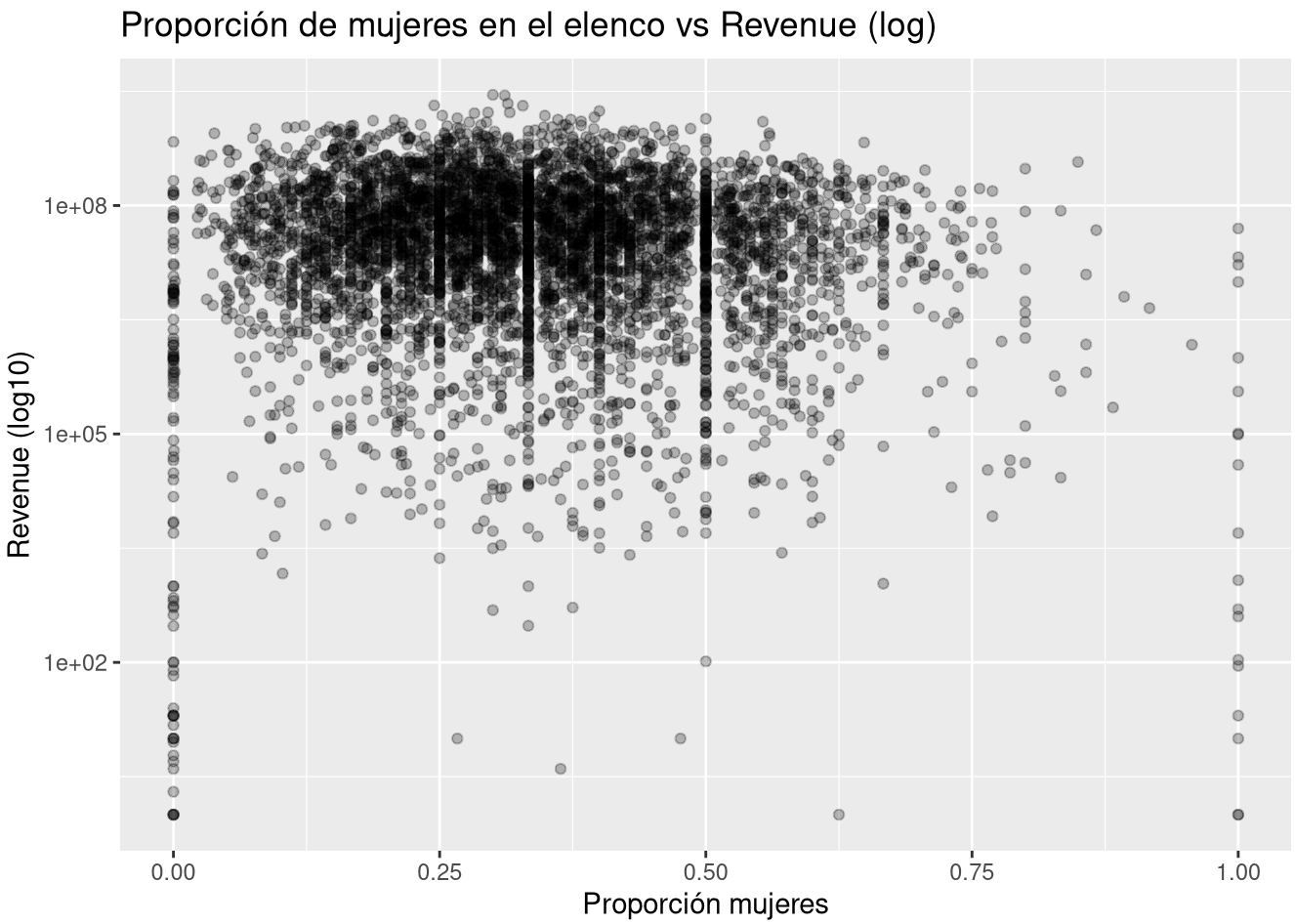


La correlación entre cantidad de actores e ingresos es positiva (~0.447), pero no es perfecta. Esto sugiere que más elenco podría asociarse con mayor revenue, aunque también puede estar mezclado con presupuestos altos. La tabla por año muestra que el promedio de actores cambia bastante según el año, lo que puede introducir sesgos temporales. Sería útil controlar por presupuesto para aislar mejor el efecto del elenco.

4.9 ¿La cantidad de hombres y mujeres en el reparto

```
## $corr_prop_mujeres_popularity
## [1] 0.09198082
##
## $corr_prop_mujeres_revenue
## [1] -0.004383732
```





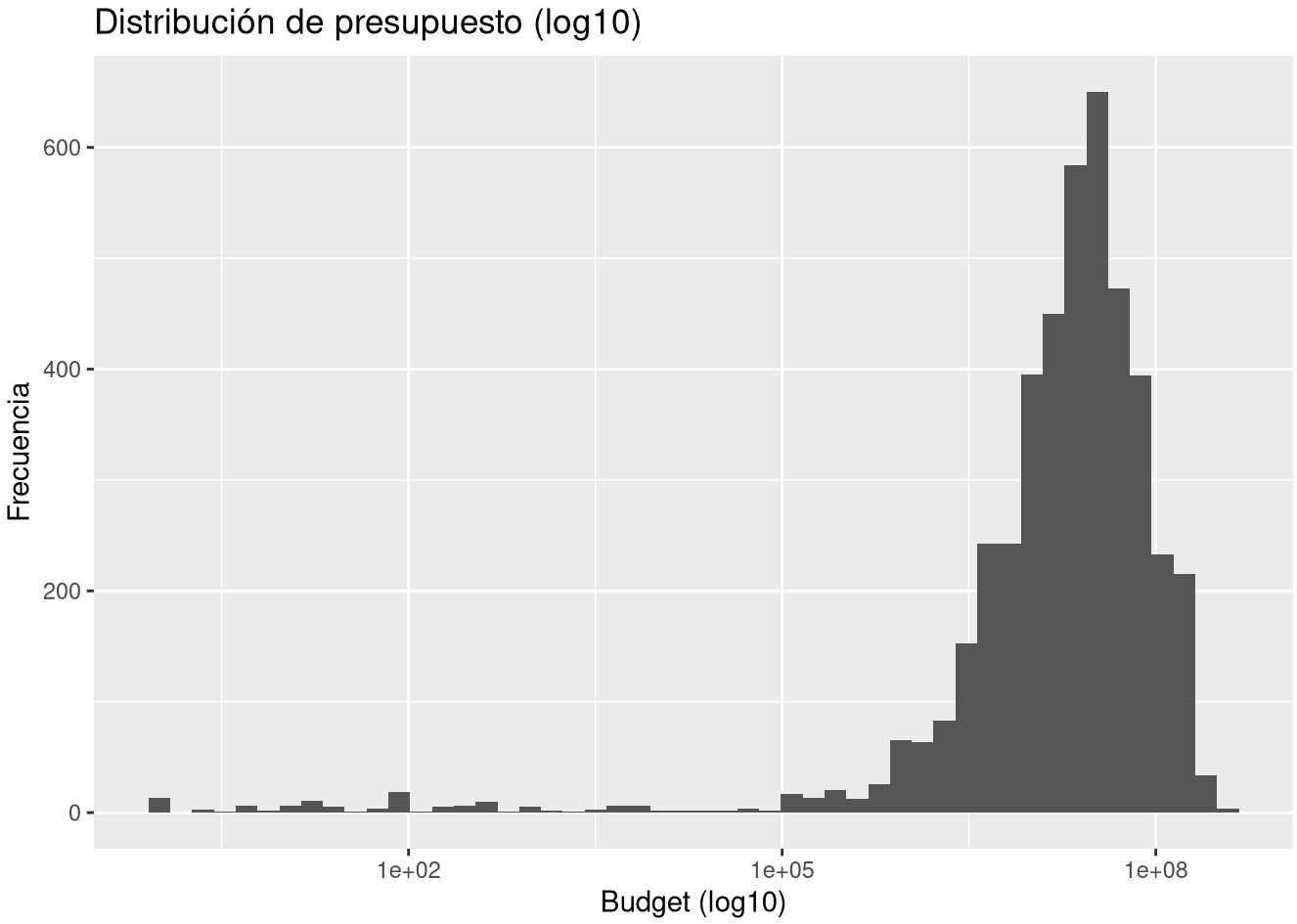
La proporción de mujeres en el reparto tiene correlación muy baja con popularidad (~0.092) y casi nula con revenue (~-0.004). Con estos datos, no parece haber una relación fuerte directa entre balance de género y éxito comercial. Aun así, correlación no implica causalidad y puede haber variables de confusión como género de película o país. El resultado sugiere que, al menos linealmente, el impacto es pequeño.

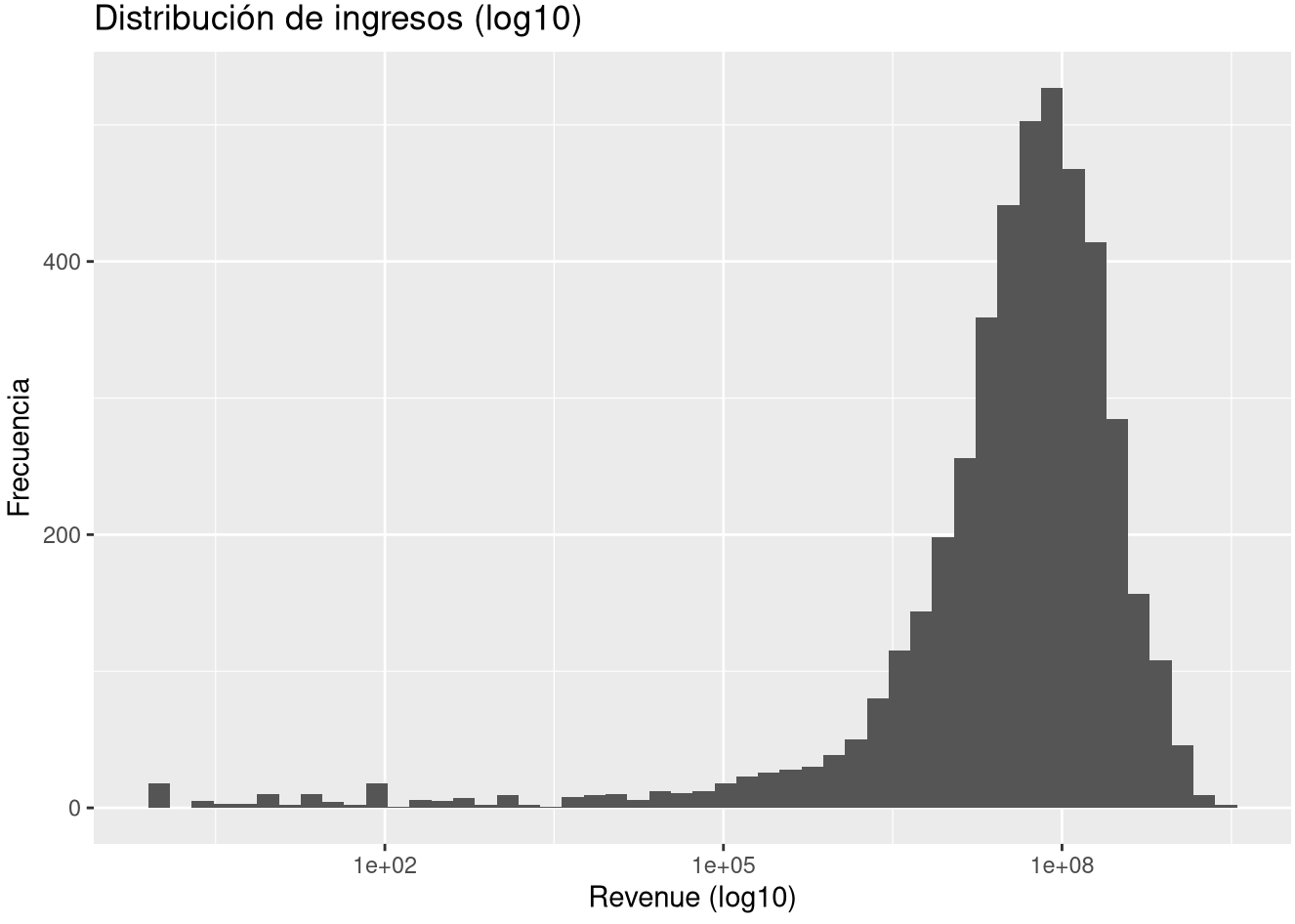
4.10 ¿Quiénes son los directores que hicieron las 20

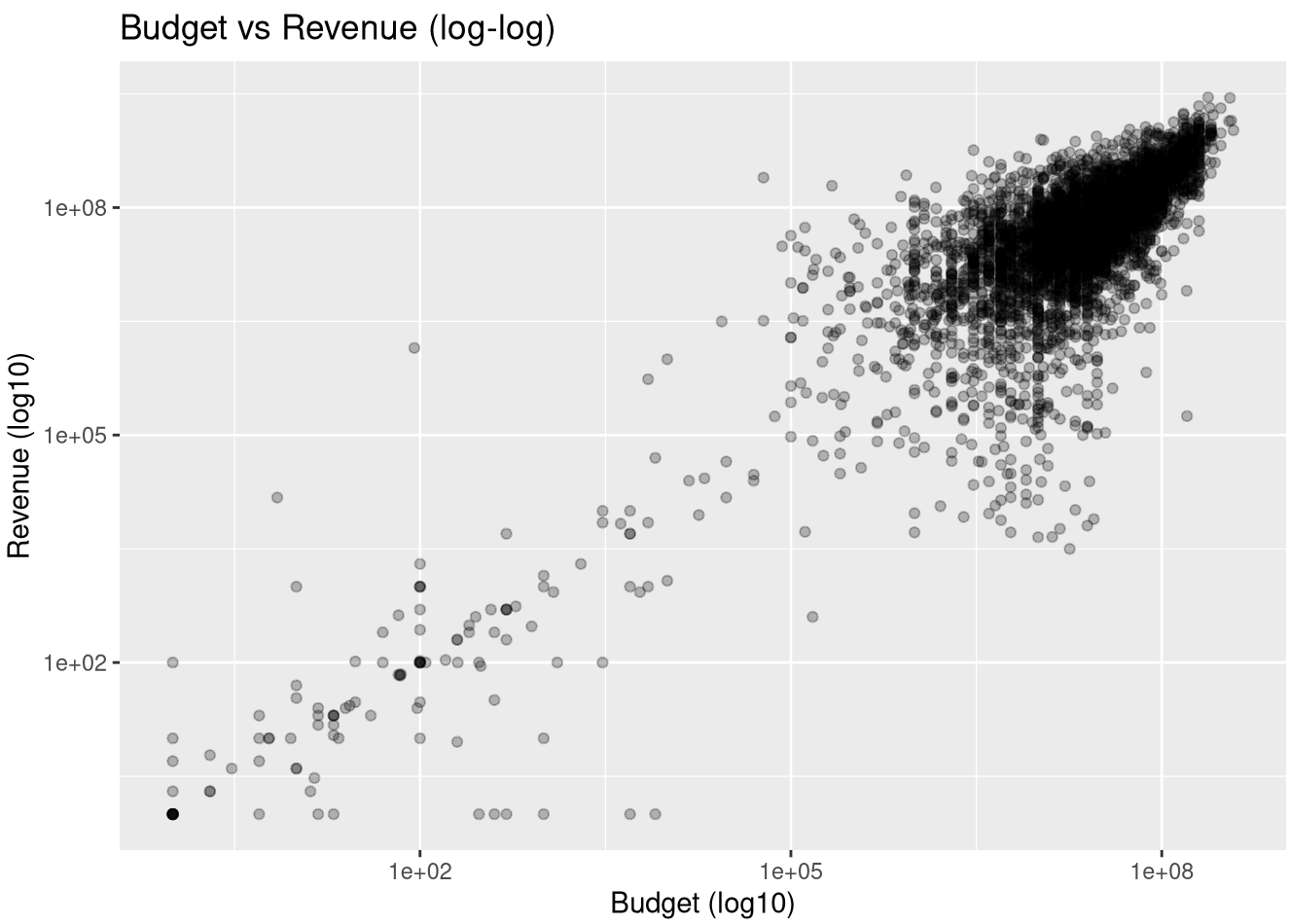
```
##                  director                                     title
## 10050       Frank Darabont                The Shawshank Redemption
## 10026 Francis Ford Coppola                            The Godfather
## 13070        Aditya Chopra               Dilwale Dulhania Le Jayenge
## 10108     Steven Spielberg                          Schindler's List
## 10028 Francis Ford Coppola                The Godfather: Part II
## 16495       Makoto Shinkai                              Your Name.
## 19218           Tosca Musk                       Gabriel's Inferno
## 9978     Christopher Nolan                         The Dark Knight
## 10265    Quentin Tarantino                            Pulp Fiction
## 9895       Robert Zemeckis                            Forrest Gump
## 9959         Peter Jackson The Lord of the Rings: The Return of the King
## 10140       Frank Darabont                          The Green Mile
## 17808         Bong Joon-ho                                Parasite
## 9962        Hayao Miyazaki                         Spirited Away
## 10232      Roberto Benigni                       Life Is Beautiful
## 10304      Martin Scorsese                             GoodFellas
## 10112         Sergio Leone           The Good, the Bad and the Ugly
## 10090         Sidney Lumet                            12 Angry Men
## 18938            Jon Watts                 Spider-Man: No Way Home
## 12087   Giuseppe Tornatore                         Cinema Paradiso
##       voteCount voteAvg
## 10050     20598     8.7
## 10026     15380     8.7
## 13070      3372     8.7
## 10108     12282     8.6
## 10028      9266     8.6
## 16495      8274     8.6
## 19218      2188     8.6
## 9978      26690     8.5
## 10265     22501     8.5
## 9895      22045     8.5
## 9959      18952     8.5
## 10140     13380     8.5
## 17808     12979     8.5
## 9962      12339     8.5
## 10232     10781     8.5
## 10304      9741     8.5
## 10112      6385     8.5
## 10090      6127     8.5
## 18938      5630     8.5
## 12087      3210     8.5
```

La lista de directores asociados a las 20 películas mejor calificadas incluye nombres muy reconocidos como Frank Darabont, Coppola y Spielberg. Esto coincide con títulos clásicos con alta calificación promedio. También aparecen algunos casos menos conocidos, lo cual puede ser por nichos con buena recepción. Para un análisis más justo, sería bueno considerar también `voteCount` para evitar títulos con pocos votos.

4.11 ¿Cómo se correlacionan los presupuestos con los

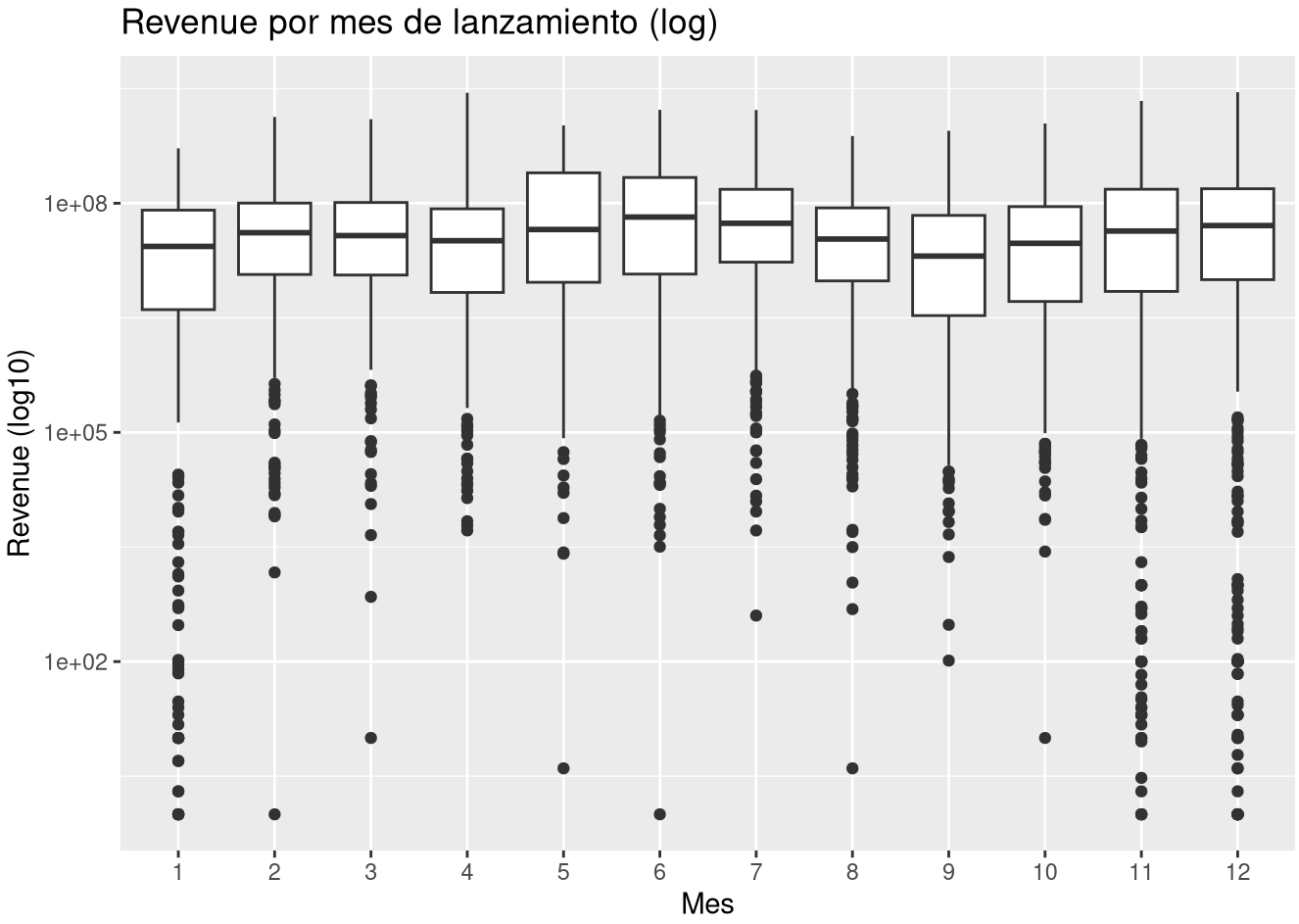






La correlación presupuesto-ingresos es alta (~0.706), lo que indica que en general gastar más se asocia con ganar más. Sin embargo, la dispersión (especialmente en escala log) sugiere que hay mucha variabilidad. Esto significa que el presupuesto ayuda, pero no garantiza éxito. Por eso es común complementar con otras variables como popularidad o marketing.

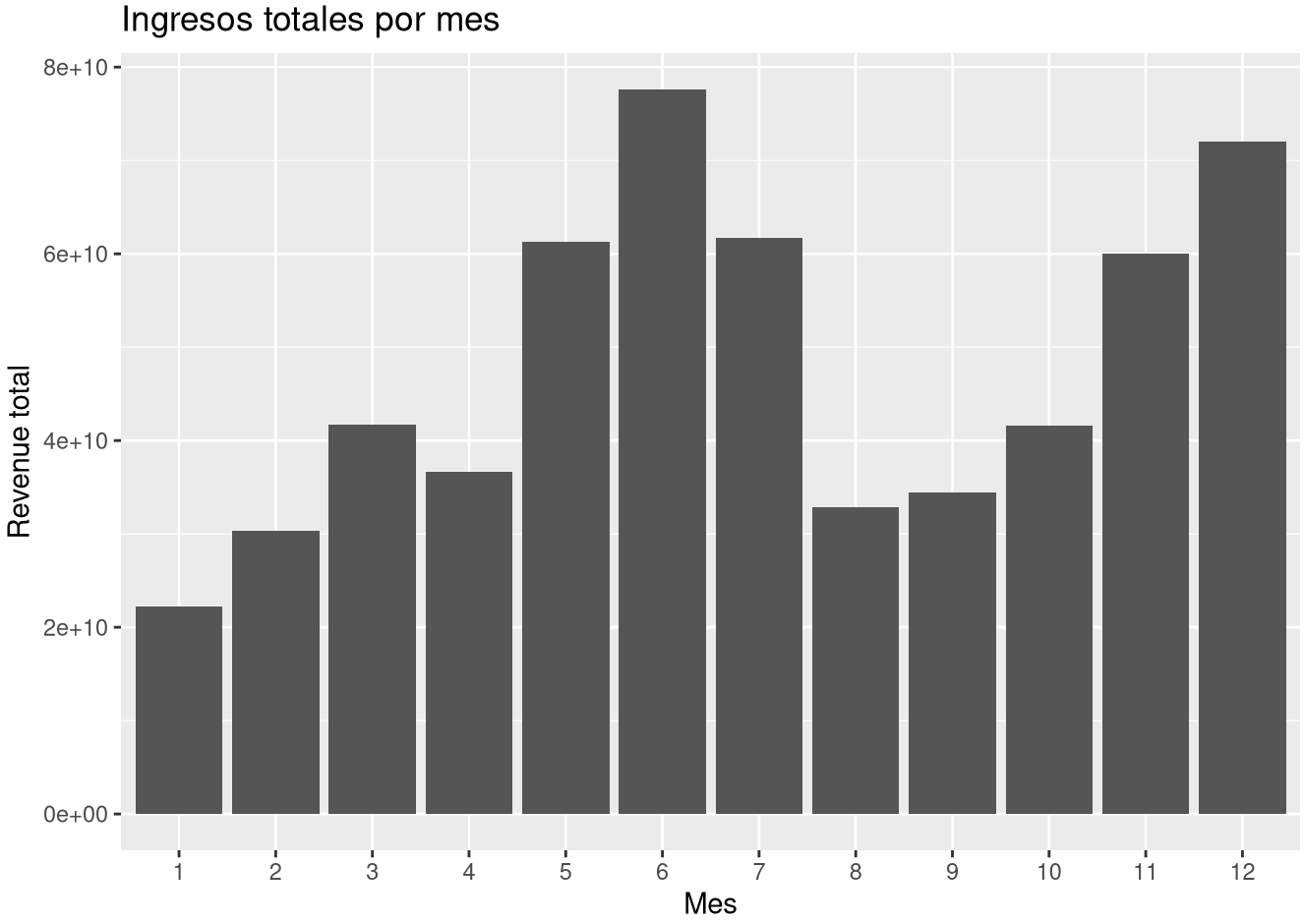
4.12 ¿Se asocian ciertos meses de lanzamiento con




```
## # A tibble: 12 × 4
##    releaseMonth peliculas revenue_prom revenue_mediana
##           <int>     <int>        <dbl>           <dbl>
##  1            6       468   165807439.       66001002
##  2            5       371   165272557.       45361000
##  3            7       465   132764089.       54682547
##  4           11       470   127811630.       43278503
##  5           12       579   124411468.       51053787
##  6            4       358   102252891.       32339075
##  7            3       429    97108375.       37713879
##  8           10       489    85148627.       29918745
##  9            2       366    82768572.       41146058
## 10            8       464    70777333.       34032922.
## 11            1       368    60407604.       27233270
## 12            9       571    60335029.       20350754
```

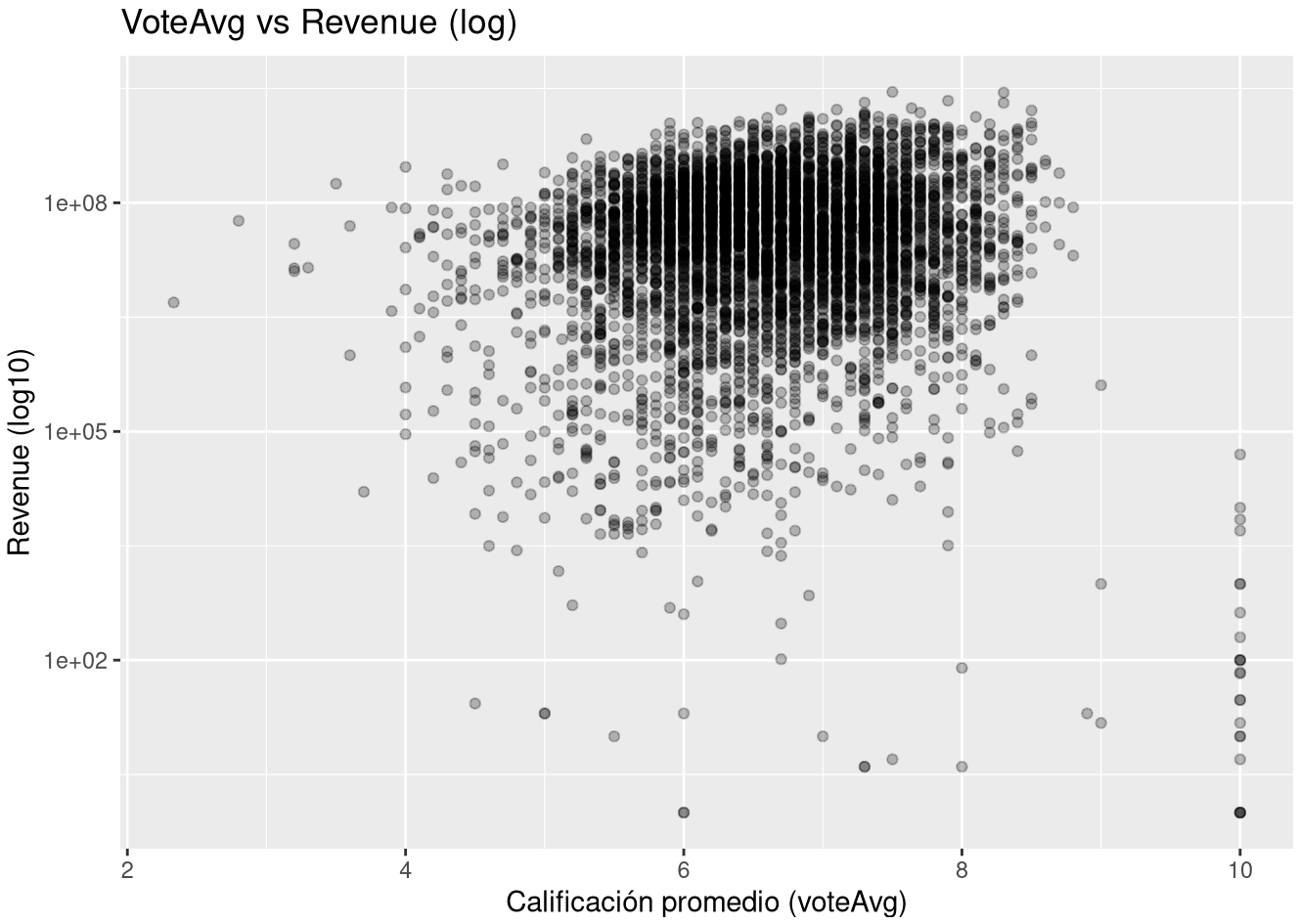
El boxplot por mes y la tabla de medias muestran que los meses 5 y 6 tienen ingresos promedio muy altos (por ejemplo junio ~165.8M). Esto sugiere una estacionalidad donde ciertos meses concentran estrenos más fuertes. Como revenue es muy sesgado, usar log y mediana ayuda a comparar mejor entre meses. Este patrón puede servir para planificar ventanas de lanzamiento.

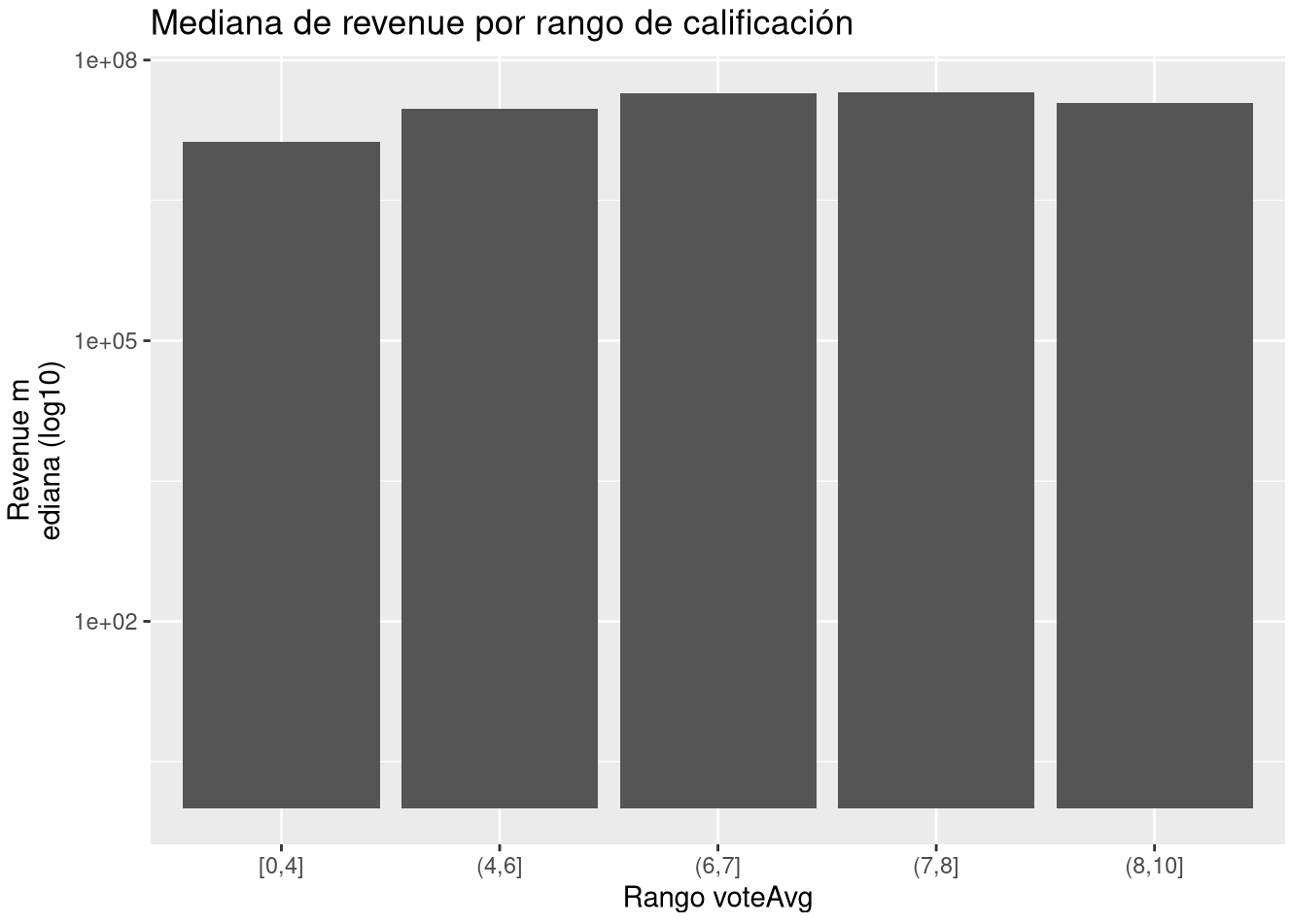
4.13 ¿En qué meses se han visto los lanzamientos con



Al sumar ingresos por mes, junio también aparece como el mejor en revenue total, seguido por meses como diciembre y julio. Se nota que algunos meses tienen muchísimas películas, pero no necesariamente el mejor promedio. Esto indica que conviene analizar tanto el total (impacto global) como el promedio (desempeño típico). Para negocio, estos meses podrían ser prioridades para estrenos grandes.

4.14 ¿Cómo se correlacionan las calificaciones con el





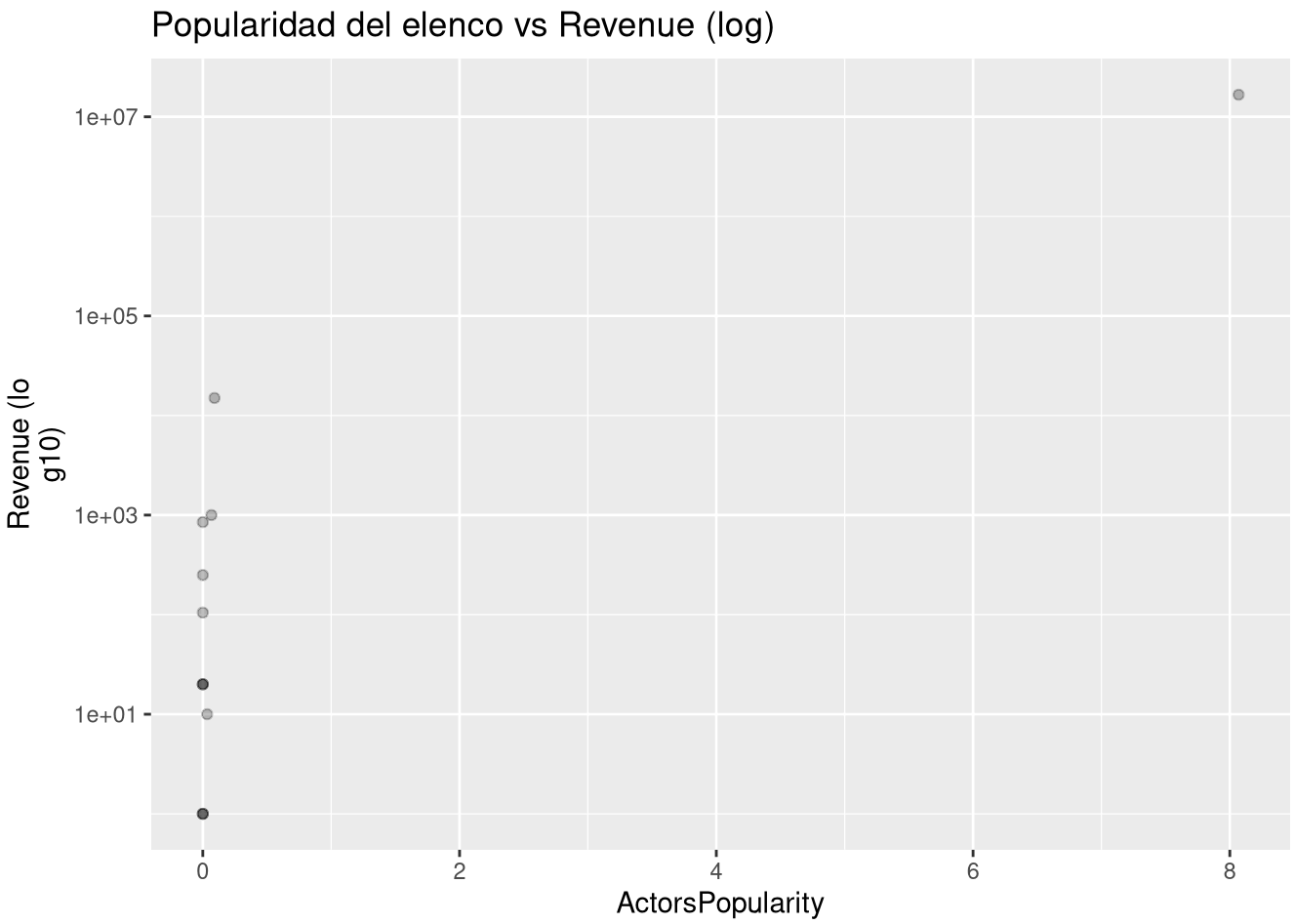
La correlación entre calificación e ingresos es baja (~0.106), así que una mejor nota no implica automáticamente más taquilla. Al agrupar por rangos de `voteAvg`, la mediana de ingresos sube de (4,6] a (7,8], pero luego baja en (8,10]. Esto puede pasar porque hay pocas películas en el rango más alto y el comportamiento es más variable. En resumen, la calidad percibida ayuda algo, pero no es el factor principal.

4.15 ¿Qué estrategias de marketing, como videos

```
## # A tibble: 3 × 5
##   has_homepage has_video revenue_prom popularity_prom peliculas
##   <chr>        <chr>            <dbl>           <dbl>     <int>
## 1 Con Web      Sin Video   150830255.           85.2       2537
## 2 Con Web      Con Video     5702633            18.3          3
## 3 Sin Web      Sin Video     3254490.            1.49       129
```

La tabla indica que tener página oficial (Con Web) está asociado con un revenue promedio mucho mayor, especialmente para el grupo Con Web / Sin Video (~150.8M). El grupo Con Web / Con Video tiene muy pocas películas (n=3), así que ese promedio no es confiable. En general, parece que la presencia de web se relaciona con mayor alcance y éxito, pero hay que cuidar el tamaño de muestra. Sería ideal comparar también por género o presupuesto para evitar confusión.

4.16 ¿La popularidad del elenco está directamente



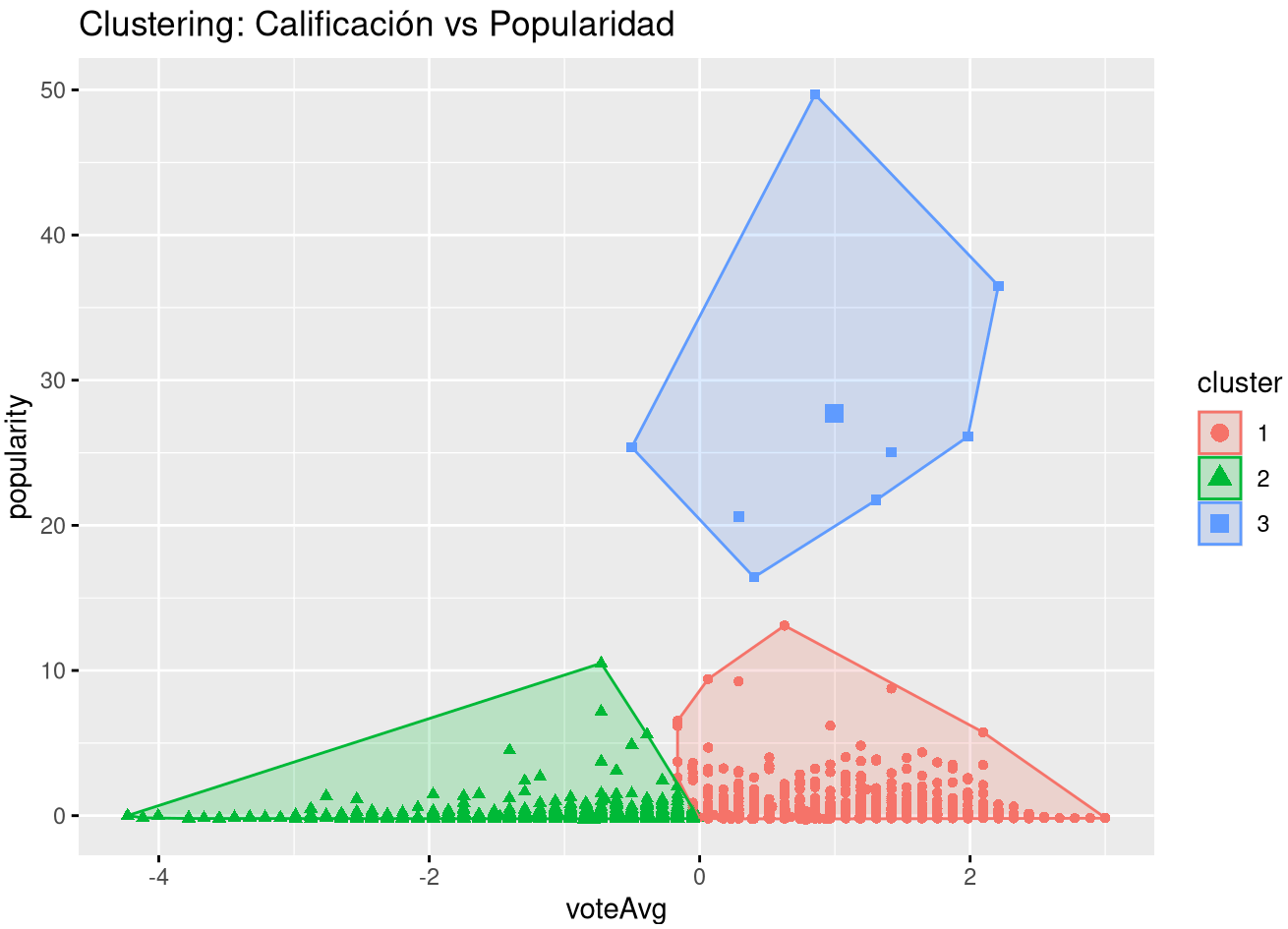
La correlación de Spearman entre popularidad del elenco e ingresos es moderada-alta (~0.624). En el scatter (con revenue en log) se ve una tendencia creciente: elencos más populares suelen asociarse a mayor taquilla. Aun así, hay dispersión, así que no todos los casos siguen la misma relación. Este resultado sugiere que `actorsPopularity` es una variable útil para predecir revenue.

5.1 ¿Es posible agrupar las películas en categorías


```
##   cluster    budget   revenue
## 1       1 176190145 899005440
## 2       2  94289095 258207095
## 3       3  20084865  55739764
```

El clustering separa películas en tres grupos con promedios claros: un grupo de alto presupuesto y alto ingreso, uno medio y uno bajo. Esto es útil para segmentación rápida y comparar estrategias entre categorías financieras. Como se estandarizan las variables, el algoritmo se enfoca en patrones relativos más que en magnitudes absolutas. Estos clusters pueden servir para análisis posterior de rentabilidad o riesgo.

5.2 ¿Existen grupos de películas “de culto” (alta




```
##   cluster  voteAvg popularity
## 1       1 7.236344   57.23974
## 2       2 5.819501   36.21563
## 3       3 7.425000 6416.89387
```

El clustering de calificación vs popularidad genera un grupo con popularidad extremadamente alta y buena calificación (perfil “viral”). También aparecen grupos con calificaciones similares pero popularidad mucho menor, que se parecen más a “de culto”. Esto muestra que popularidad y calidad no son lo mismo y pueden separarse. La segmentación ayuda a identificar targets distintos para marketing.

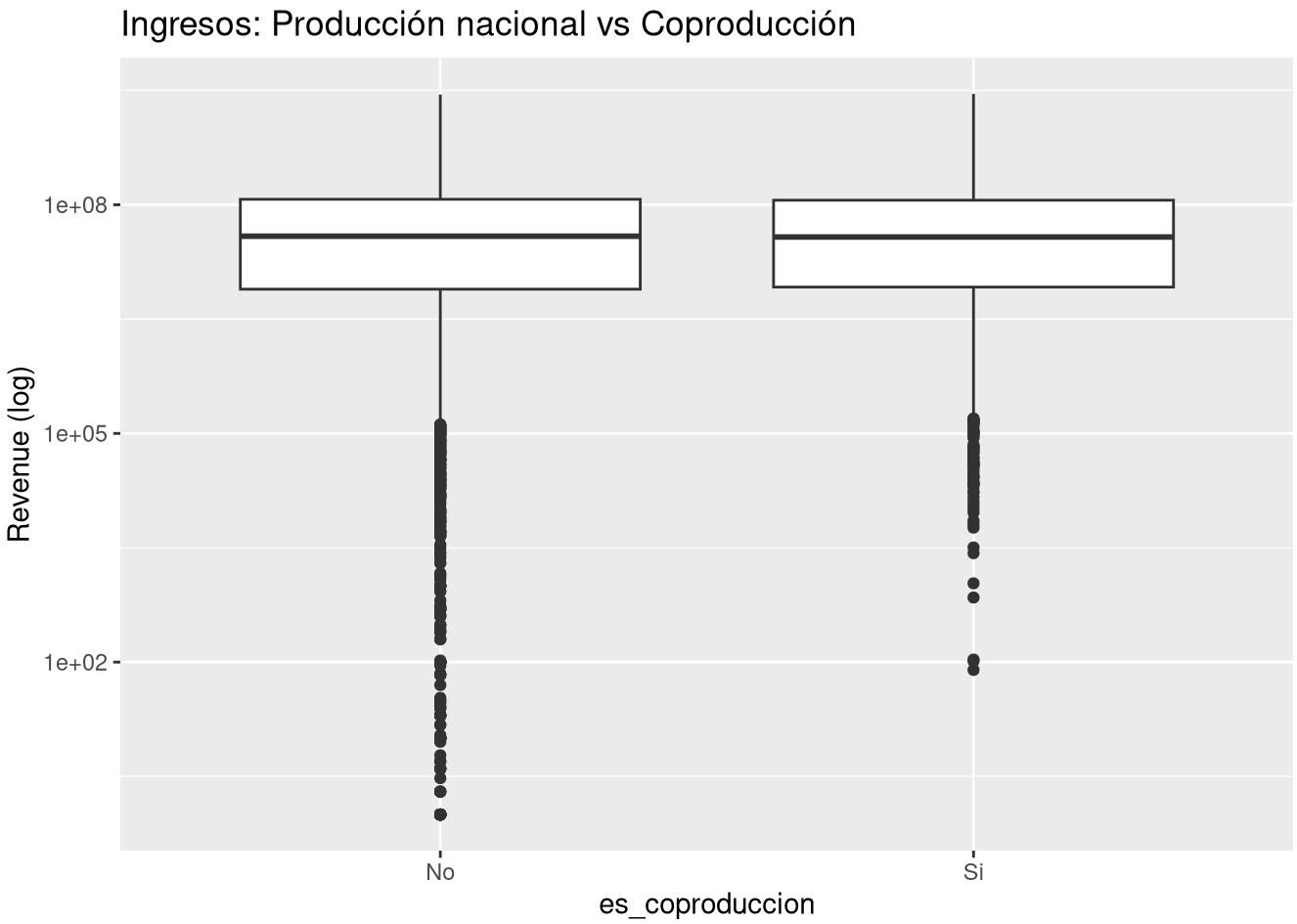
5.3 ¿Cuáles son las películas más rentables en

```
##                                title   budget   revenue       ROI
## 12334              Alice in Wonderland  3000000 572000000 189.66667
## 10100 Snow White and the Seven Dwarfs  1488423 184925486 123.24256
## 10305               Gone with the Wind  4000000 402352579  99.58814
## 13592   The Rocky Horror Picture Show  1200000 112892319  93.07693
## 12091                       Cinderella  2900000 263591415  89.89359
## 9992                               Saw  1200000 103911669  85.59306
## 10202      E.T. the Extra-Terrestrial 10500000 792965500  74.52052
## 11178        My Big Fat Greek Wedding  5000000 368744044  72.74881
## 9893                        Star Wars 11000000 775398007  69.49073
## 12019            Saturday Night Fever  3500000 237113184  66.74662
```

El top de ROI muestra películas con presupuestos relativamente bajos y revenues muy altos, como Alice in Wonderland con ROI ~189.7. Esto resalta que el retorno relativo puede contar otra historia distinta al ingreso absoluto. Filtrar por presupuesto mínimo evita que presupuestos muy pequeños distorsionen el ROI. Estos casos son interesantes para estudiar qué características se repiten en éxitos de alta rentabilidad.

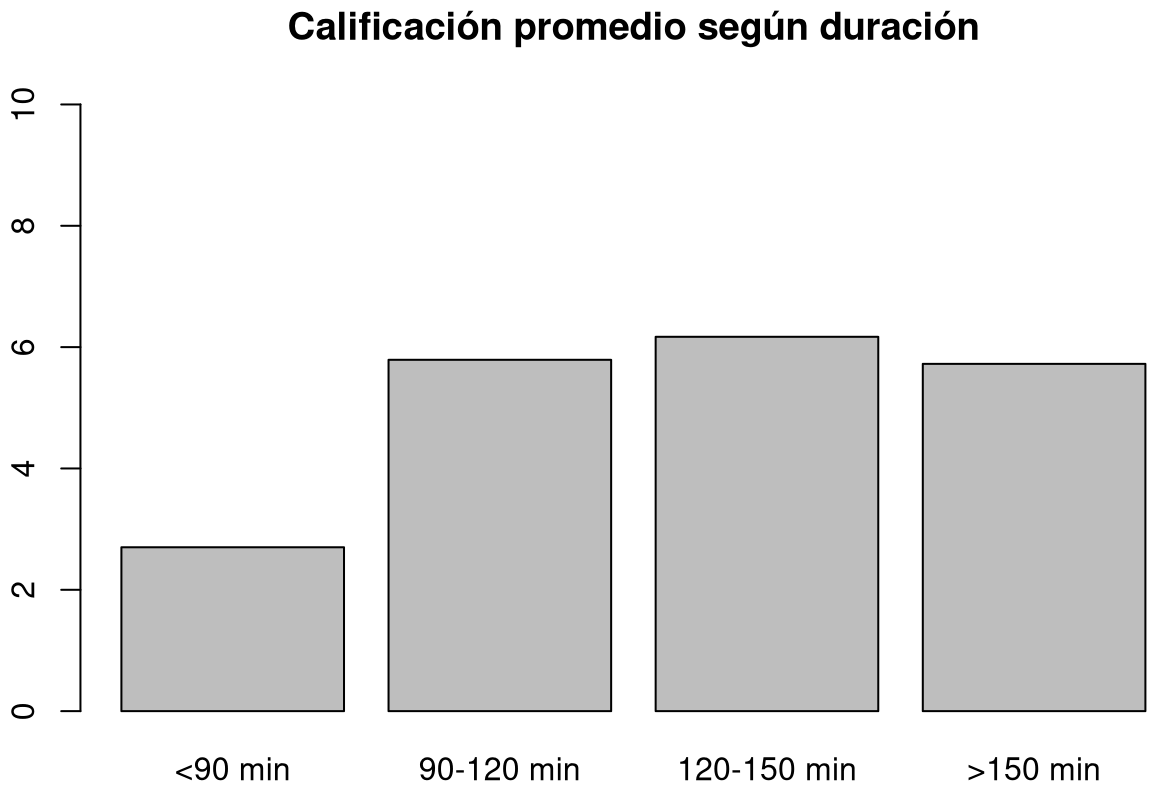
5.4 ¿Las coproducciones (varios países) generan más

```
##   es_coproduccion   revenue
## 1              No 106977342
## 2              Si 103856831
```



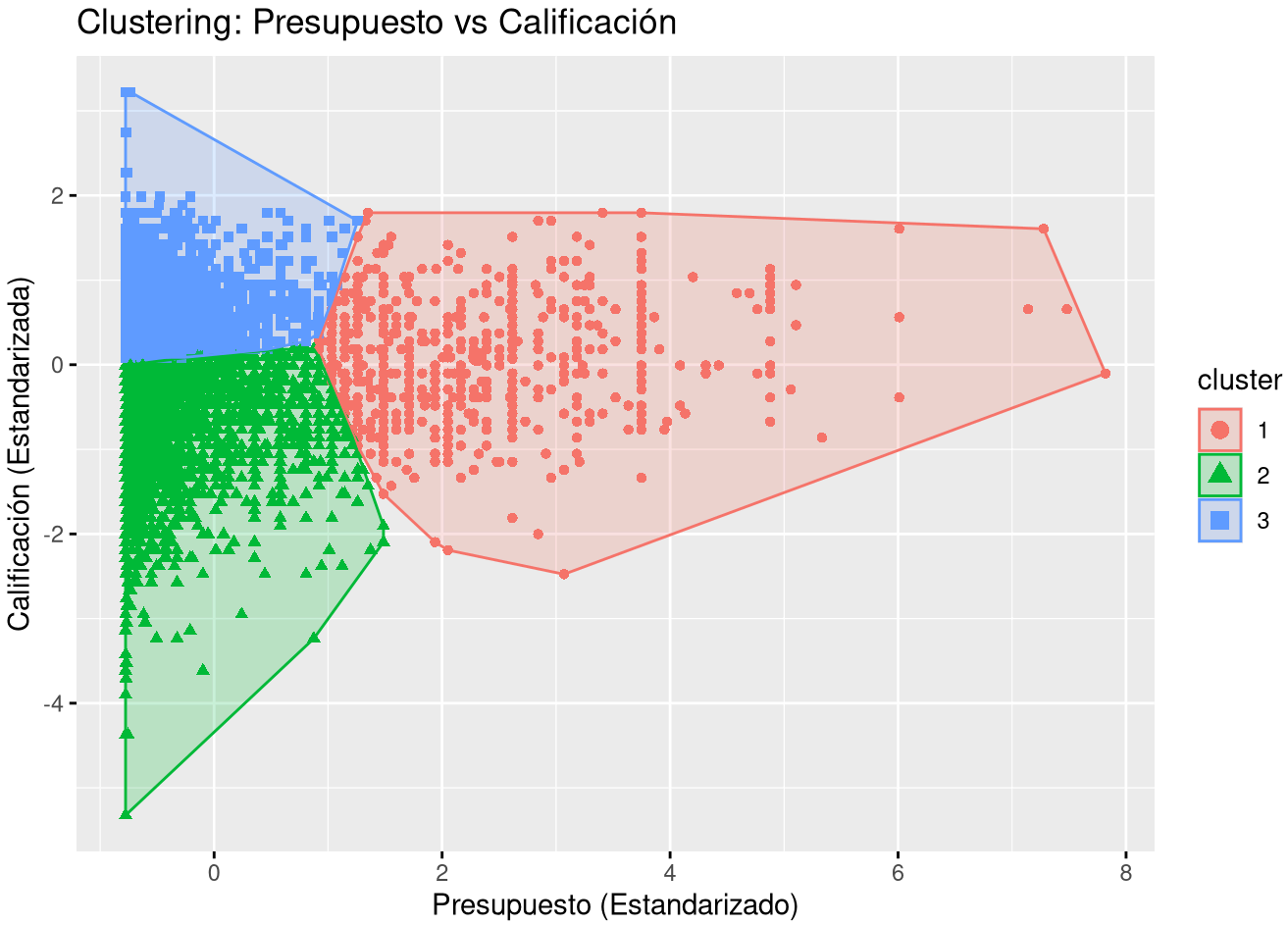
Al comparar coproducciones (más de un país) contra producciones de un solo país, se puede evaluar si hay ventaja en ingresos promedio. Este análisis es útil porque las coproducciones pueden ampliar mercado y distribución. El resultado depende mucho de cómo esté distribuido el dataset y de si hay outliers grandes. Por eso, además del promedio, vale la pena revisar mediana o escala log.

5.5 ¿Existe una duración (runtime) “ideal” que



Al agrupar por rangos de duración, las películas de 120–150 minutos tienen la calificación promedio más alta (~6.17). Las muy cortas (<90) tienen una calificación promedio mucho menor (~2.70), lo que sugiere menor aceptación. Esto no significa que la duración cause la calificación, pero sí marca un patrón interesante. Podría explorarse si el género o el año explican parte de esta diferencia.

5.6 ¿Se pueden identificar grupos según presupuesto




```
##   cluster    budget  voteAvg
## 1       1 136429551 6.705811
## 2       2  25081252 5.873428
## 3       3  16939485 7.510606
```

El clustering de presupuesto vs calificación separa tres perfiles, incluyendo uno de alto presupuesto con buena calificación promedio (~6.71) y otro de presupuesto menor con calificación aún más alta (~7.51). Esto sugiere que gastar mucho no garantiza la mejor crítica, aunque puede ayudar. La estandarización permite comparar patrones sin que el presupuesto domine por escala. Los promedios por cluster ayudan a interpretar qué tipo de películas caen en cada grupo.