Lab 1

Mario Fernando Rocha 23501, Luis Pedro Lira 23669,Juan Francisco Martínez 23617

2026-01-16

Análisis exploratorio de Datos

Carros

```r
movies <- read.csv("movies_2026.csv")
```

Pregunta 1: Exploración rápida de los Datos

#Dimensiones

Cambia SOLO el encabezado del chunk:

```
``` r
nrow(movies)
ncol(movies)
dim(movies)
names(movies)
str(movies)
summary(movies)
```

#Datos Faltantes y Duplicados

```r
# Cantidad de NA por columna
na_conteo <- colSums(is.na(movies))

# Porcentaje de NA por columna
na_porcentaje <- round(colMeans(is.na(movies)) * 100, 2)

# Tabla resumen de faltantes
faltantes <- data.frame(
  variable = names(movies),
  na = as.vector(na_conteo),
  na_pct = as.vector(na_porcentaje)
)
```
```

```r
head(faltantes, 15)  # Top 15 columnas con más NA

# Duplicados (filas repetidas)
sum(duplicated(movies))
```

Dimension del conjunto de datos

El conjunto de datos está compuesto por 19,883 observaciones y 28 variables, lo que representa un volumen considerable de información. Este tamaño permite realizar un análisis exploratorio robusto y obtener conclusiones generales representativas del comportamiento de la industria cinematográfica.

Estructura del data set

El dataset presenta una combinación de variables cuantitativas y cualitativas, lo que permite analizar tanto aspectos financieros y de popularidad como características narrativas, de producción y distribución de las películas.

Presupuesto

El dataset presenta una combinación de variables cuantitativas y cualitativas, lo que permite analizar tanto aspectos financieros y de popularidad como características narrativas, de producción y distribución de las películas.

#Resumen de variables Numéricas

```r
# Detectar columnas numéricas
num_cols <- sapply(movies, is.numeric)
datos_num <- movies[, num_cols]

# Resumen estadístico simple por variable numérica
res_num <- data.frame(
  variable = names(datos_num),
  media = sapply(datos_num, mean, na.rm = TRUE),
  mediana = sapply(datos_num, median, na.rm = TRUE),
  sd = sapply(datos_num, sd, na.rm = TRUE),
  min = sapply(datos_num, min, na.rm = TRUE),
  max = sapply(datos_num, max, na.rm = TRUE)
)

head(res_num)
```

#variables Categóricas

```r
cat_cols <- sapply(movies, is.character) | sapply(movies, is.factor)
names(movies[, cat_cols])
```

Pregunta número 2

```r
variable_type <- data.frame(
  Variable = names(movies),
  Tipo = c(
    "Cuantitativa discreta", "Cuantitativa continua", "Cuantitativa continua",
    "Cuantitativa continua", "Cualitativa nominal", "Cualitativa nominal",
    "Cualitativa nominal", "Cualitativa nominal", "Cualitativa nominal",
    "Cualitativa nominal", "Cuantitativa continua", "Cualitativa nominal",
    "Cuantitativa discreta", "Cualitativa nominal", "Cuantitativa discreta",
    "Cualitativa nominal", "Cualitativa nominal", "Cuantitativa discreta",
    "Cualitativa nominal", "Cuantitativa discreta", "Cuantitativa continua",
    "Cualitativa nominal", "Cuantitativa continua", "Cualitativa nominal",
    "Cuantitativa discreta", "Cuantitativa discreta", "Cuantitativa discreta",
    "Cuantitativa discreta"
  )
)

variable_type
```

Interpretación

Las variables del conjunto de datos se clasifican en cuantitativas y cualitativas. Las variables cuantitativas incluyen tanto variables continuas, relacionadas con aspectos financieros y de popularidad, como variables discretas asociadas a conteos y características de producción. Por otro lado, las variables cualitativas son mayoritariamente de tipo nominal y describen atributos narrativos, de idioma, elenco y producción de las películas. Esta diversidad de tipos de variables permite realizar distintos análisis descriptivos y comparativos a lo largo del estudio.

###Pregunta número 3

Separar las variables

```r
vars_cuant <- variable_type$Variable[grepl("Cuantitativa", variable_type$Tipo)]
vars_cual  <- variable_type$Variable[grepl("Cualitativa", variable_type$Tipo)]

vars_cuant
vars_cual
```

##normalizar

```r
set.seed(123)

normalidad_list <- lapply(vars_cuant, function(v){
  x <- movies[[v]]
  x <- x[!is.na(x)]
  x <- x[is.finite(x)]   # por si hay Inf/NaN

  n <- length(x)

  if(n < 3){
    return(data.frame(
      Variable = v,
```

```
      n = n,
      muestra = n,
      p_value = NA,
      decision = "No aplica (n<3)"
    ))
  }

  m <- min(n, 5000)
  xs <- sample(x, m)

  p <- shapiro.test(xs)$p.value
  decision <- ifelse(p < 0.05, "Rechaza normalidad", "No rechaza normalidad")

  data.frame(
    Variable = v,
    n = n,
    muestra = m,
    p_value = p,
    decision = decision
  )
})

normalidad <- bind_rows(normalidad_list) %>%
  arrange(p_value)

normalidad
```

Normalidad de las variables cuantitativas

Los resultados de la prueba de Shapiro–Wilk muestran que todas las variables cuantitativas analizadas presentan valores p inferiores a 0.05, por lo que se rechaza la hipótesis de normalidad en todos los casos. Este resultado indica que las variables no siguen una distribución normal, lo cual es consistente con la exploración inicial del conjunto de datos, donde se observaron distribuciones altamente asimétricas y la presencia de valores extremos. En consecuencia, para los análisis posteriores se emplean métodos descriptivos y no paramétricos.

##Pregunta 4

###Top 10 películas con más presupuesto

```
top10_budget <- movies %>%
  filter(!is.na(budget)) %>%
  arrange(desc(budget)) %>%
  select(title, budget, revenue, releaseYear) %>%
  head(10)

top10_budget
```

Interpretación Las diez películas con mayor presupuesto corresponden principalmente a franquicias reconocidas y producciones de gran escala. Esto refleja una estrategia de alta inversión por parte de los estudios, enfocada en películas con alto potencial comercial y amplio alcance internacional.

###Top 10 películas con más ingresos

```r
top10_revenue <- movies %>%
  filter(!is.na(revenue)) %>%
  arrange(desc(revenue)) %>%
  select(title, revenue, budget, releaseYear) %>%
  head(10)

top10_revenue
```

Interpretación Las películas con mayores ingresos corresponden a producciones altamente populares y con distribución global. Si bien existe coincidencia entre películas con alto presupuesto y altos ingresos, no todas las producciones costosas garantizan un éxito financiero equivalente, lo que sugiere la influencia de otros factores como recepción del público y estrategia de marketing.

Película con más votos

```r
pelicula_mas_votos <- movies %>%
  filter(!is.na(voteCount)) %>%
  arrange(desc(voteCount)) %>%
  select(title, voteCount, voteAvg, revenue, budget, releaseYear) %>%
  head(1)

pelicula_mas_votos
```

Interpretación La película con mayor número de votos es “Inception”, lo que indica una alta participación del público y un interés sostenido a lo largo del tiempo. Su elevada calificación promedio sugiere que, además de ser ampliamente vista, fue bien recibida por los usuarios.

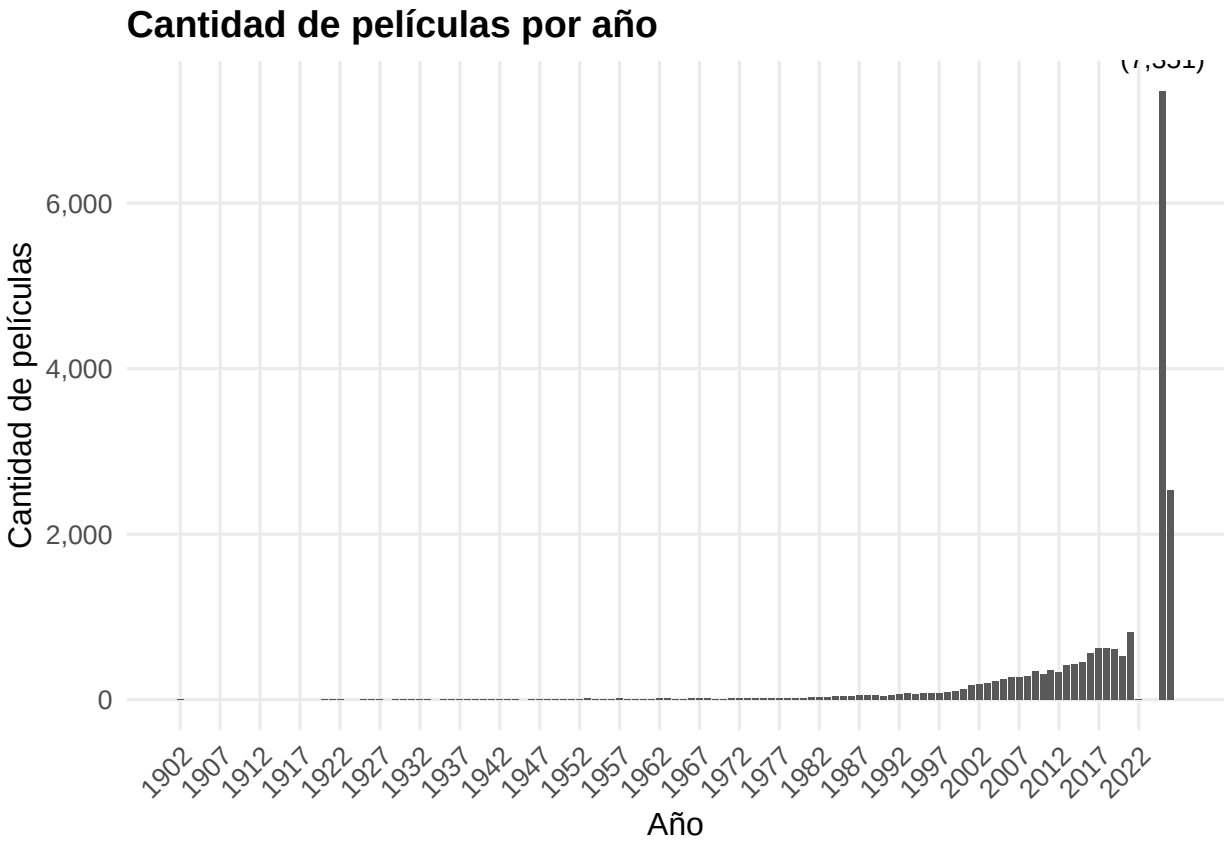
Película con menos votos

```r
peliculas_0_votos <- movies %>%
  filter(!is.na(voteCount), voteCount == 0) %>%
  select(title, voteCount, voteAvg, releaseYear)

peliculas_0_votos

min_votos_pos <- movies %>%
  filter(!is.na(voteCount), voteCount > 0) %>%
  summarise(min_voteCount = min(voteCount)) %>%
  pull(min_voteCount)

peliculas_menos_votos_con_voto <- movies %>%
  filter(!is.na(voteCount), voteCount == min_votos_pos) %>%
  select(title, voteCount, voteAvg, releaseYear)

min_votos_pos
peliculas_menos_votos_con_voto
```

Interpretación Las películas con menor número de votos no permiten identificar de manera confiable la “peor película”, ya que una sola votación no es representativa de la percepción general del público. Por esta razón, las calificaciones con muy pocos votos deben interpretarse con cautela y no se consideran concluyentes.

###Películas por año

```
library(dplyr)
library(ggplot2)
library(scales)

peliculas_por_anio <- movies %>%
  filter(!is.na(releaseYear)) %>%
  mutate(releaseYear = as.integer(releaseYear)) %>%
  group_by(releaseYear) %>%
  summarise(cantidad = n()) %>%
  arrange(releaseYear)

anio_mas <- peliculas_por_anio %>%
  filter(cantidad == max(cantidad))
ggplot(peliculas_por_anio, aes(x = releaseYear, y = cantidad)) +
  geom_col(width = 0.85) +
  geom_text(
    data = anio_mas,
    aes(label = paste0("Máximo: ", releaseYear, "\n(", comma(cantidad), ")")),
    vjust = -0.4,
    size = 3.5
  ) +
  scale_x_continuous(
    breaks = seq(
      min(peliculas_por_anio$releaseYear),
      max(peliculas_por_anio$releaseYear),
      by = 5
    )
  ) +
  scale_y_continuous(labels = comma) +
  labs(
    title = "Cantidad de películas por año",
    x = "Año",
    y = "Cantidad de películas"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

El gráfico de barras muestra la cantidad de películas producidas por año. Se observa un crecimiento muy marcado en la producción cinematográfica a partir de la década de 1990, con un incremento aún más pronunciado después del año 2000. El año con mayor cantidad de películas producidas corresponde a uno de los años más recientes, lo que refleja tanto el crecimiento de la industria audiovisual como una mayor disponibilidad de información para producciones modernas. En contraste, los años más antiguos presentan una cantidad significativamente menor de películas registradas.

###Género de películas

```
movies2 <- movies %>%
  mutate(
    main_genre = ifelse(is.na(genres) | genres == "", "Sin género", sub("\\|.*", "", genres)),

    release_dt = suppressWarnings(ymd(releaseDate)),

    release_dt = ifelse(is.na(release_dt) & !is.na(releaseYear),
                        as.Date(paste0(releaseYear, "-12-31")),
                        release_dt) %>% as.Date()
  )
```

####top 20 películas más recientes

```
top20_recientes <- movies2 %>%
  filter(!is.na(release_dt)) %>%
  arrange(desc(release_dt)) %>%
```

```
  select(title, releaseYear, releaseDate, main_genre, runtime) %>%
  head(20)

top20_recientes
```

Interpretación

El análisis de las 20 películas más recientes muestra una diversidad considerable de géneros, lo que indica que las producciones actuales no se concentran en un solo tipo narrativo. Sin embargo, al analizar el conjunto de datos completo, se observa que el género Drama es el que predomina, seguido por otros géneros como Comedy, Action y Animation. Esto sugiere que, aunque existe variedad en los lanzamientos recientes, el drama continúa siendo un género central dentro de la industria cinematográfica.

####Género predominante

```
freq_genero <- movies2 %>%
  group_by(main_genre) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  mutate(porcentaje = round(100 * n / sum(n), 2))

freq_genero
```

Interpretación

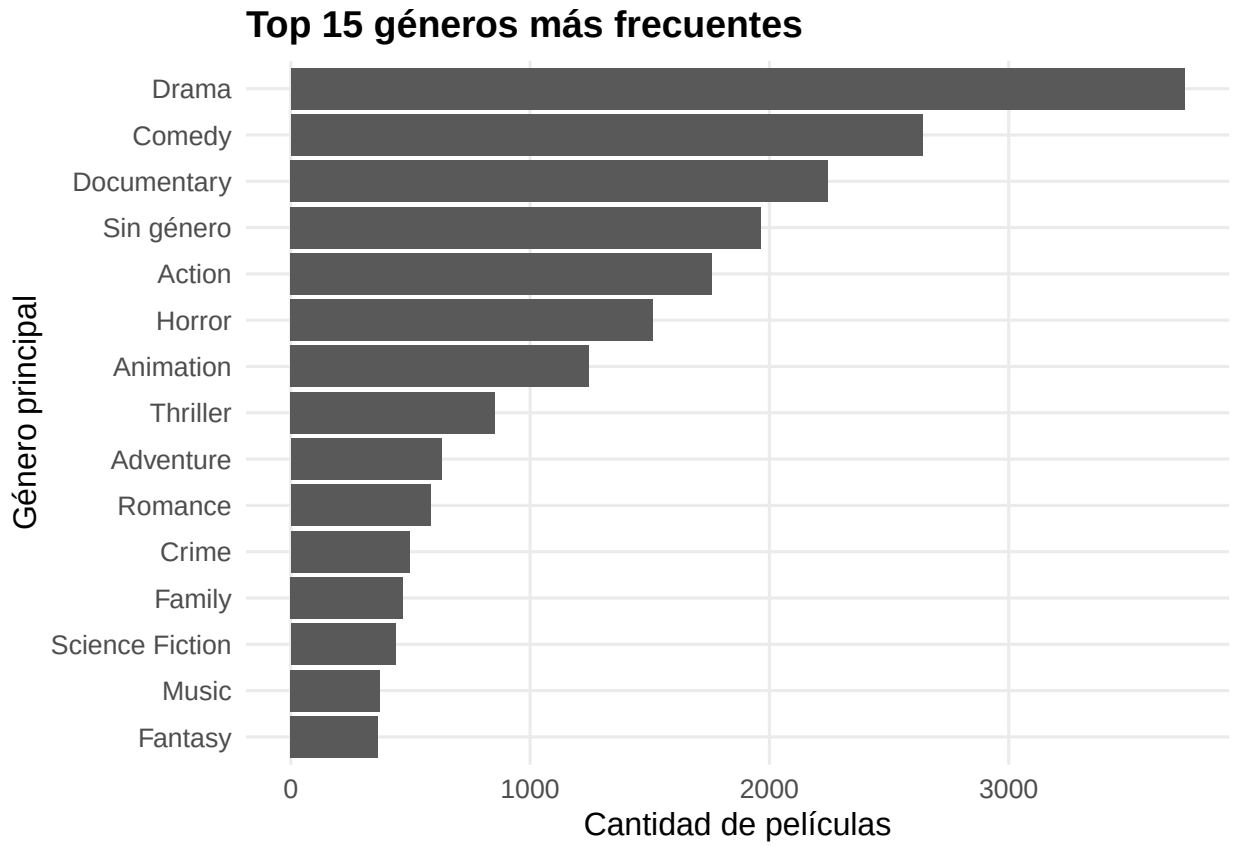
El análisis de frecuencias muestra que el género Drama es el más predominante en el conjunto de datos, representando aproximadamente el 18.8% de las películas. Este resultado sugiere que las producciones cinematográficas tienden a concentrarse en narrativas dramáticas, seguidas por géneros como Comedy y Documentary. La presencia de la categoría "Sin género" se debe a registros con información incompleta y no altera la tendencia general observada.

####Género predominante

```
genero_predominante <- freq_genero %>% head(1)
genero_predominante
```

####Gráfico de top 15 géneros

```
freq_top15 <- freq_genero %>% head(15)

ggplot(freq_top15, aes(x = reorder(main_genre, n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Top 15 géneros más frecuentes",
    x = "Género principal",
    y = "Cantidad de películas"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

El gráfico de los 15 géneros más frecuentes refuerza los resultados de la tabla de frecuencias, evidenciando que el género Drama lidera claramente el conjunto de datos, seguido por Comedy y Documentary. A partir de estos géneros se observa una disminución gradual en la cantidad de películas por categoría.

####Películas más largas

```
top20_largas <- movies2 %>%
  filter(!is.na(runtime), runtime > 0) %>%
  arrange(desc(runtime)) %>%
  select(title, runtime, releaseYear, main_genre) %>%
  head(20)

top20_largas
```

Interpretación

El análisis de las películas con mayor duración muestra la presencia de valores extremos, con duraciones que superan ampliamente el promedio habitual de una película. Estas producciones corresponden principalmente a contenidos especiales, como documentales extensos, registros históricos o producciones no convencionales. Este comportamiento explica la alta asimetría observada en la variable runtime durante la exploración inicial de los datos.

####Género de las películas más largas

```r
# Crear variable de género principal (primer género)
movies2 <- movies2 %>%
  mutate(
    main_genre = ifelse(
      is.na(genres) | genres == "",
      "Sin género",
      sapply(strsplit(as.character(genres), "\\|"), `[`, 1)
    )
  )

# Crear top 50 películas más largas
top50_largas <- movies2 %>%
  filter(!is.na(runtime), runtime > 0) %>%
  arrange(desc(runtime)) %>%
  head(50)
```

```r
freq_genero_largas <- top50_largas %>%
  group_by(main_genre) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  mutate(porcentaje = round(100 * n / sum(n), 2))

freq_genero_largas
```

Género de las películas más largas ####Género predominante en las más largas

```r
genero_top_largas <- freq_genero_largas %>% head(1)
genero_top_largas
```

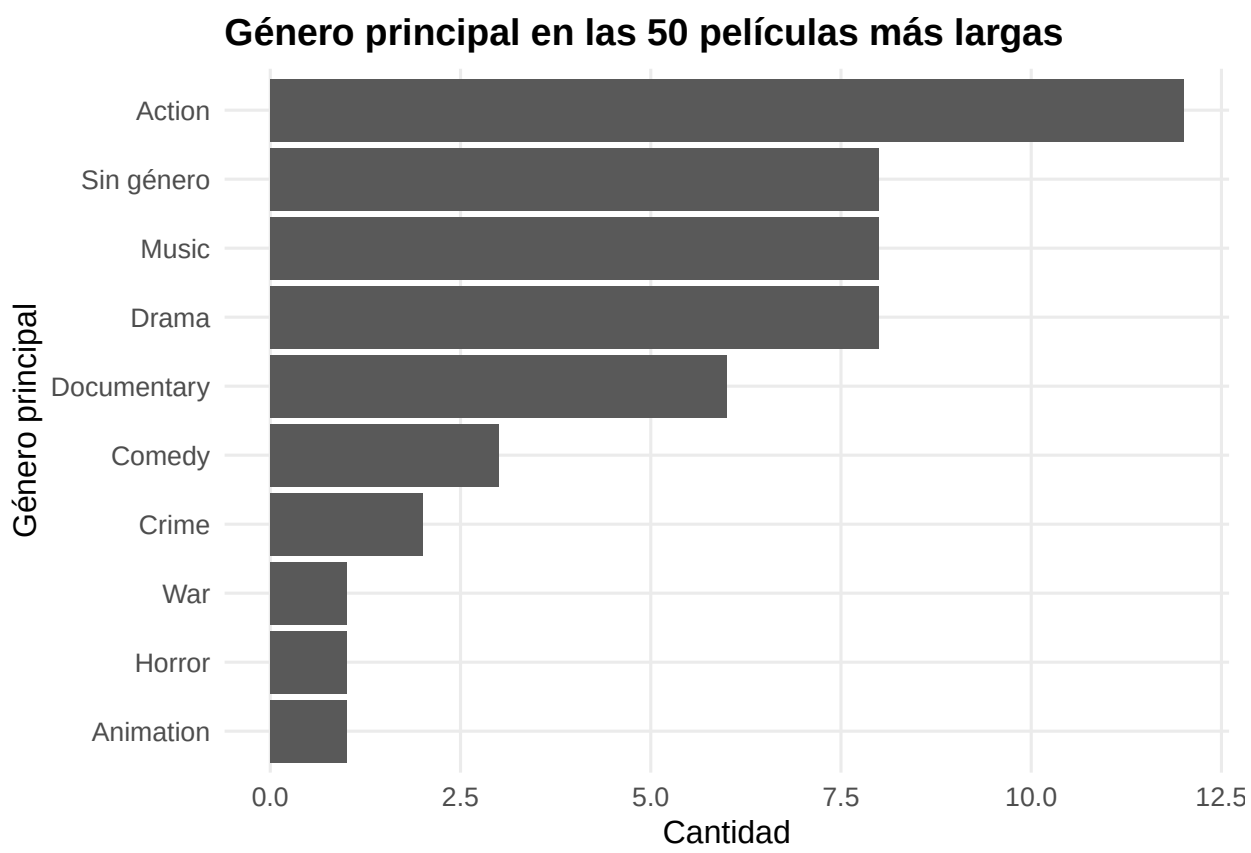
Interpretación

El análisis de las 50 películas con mayor duración muestra que estas se distribuyen entre varios géneros principales. El género Action presenta la mayor frecuencia, representando el 24% de las películas más largas, seguido por Drama y Music, cada uno con el 16%. Asimismo, una proporción relevante de películas aparece clasificada como "Sin género", lo que se asocia a información incompleta en el conjunto de datos. Estos resultados indican que las películas de mayor duración no se concentran en un solo género, sino que corresponden a producciones diversas y, en muchos casos, contenidos especiales o no convencionales.

####Gráfico de los géneros predominantes en las películas más largas

```r
ggplot(freq_genero_largas, aes(x = reorder(main_genre, n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Género principal en las 50 películas más largas",
    x = "Género principal",
    y = "Cantidad"
  ) +
  theme_minimal(base_size = 12) +
  theme(
```

```
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

El análisis de la ganancia promedio por género muestra que el género Animation es el que presenta el mayor beneficio promedio por película. Esto sugiere que, aunque el número de producciones animadas es menor en comparación con otros géneros, estas tienden a ser altamente rentables de manera individual.

####Género con más ganancias

```
movies_gan <- movies %>%
  mutate(
    main_genre = ifelse(is.na(genres) | genres == "", "Sin género", sub("\\|.*", "", genres)),
    ganancia = revenue - budget
  ) %>%
  filter(!is.na(budget), !is.na(revenue), budget > 0, revenue > 0)

# Resumen por género
gan_por_genero <- movies_gan %>%
  group_by(main_genre) %>%
  summarise(
    peliculas = n(),
    ganancia_total = sum(ganancia, na.rm = TRUE),
    ganancia_promedio = mean(ganancia, na.rm = TRUE),
    ganancia_mediana = median(ganancia, na.rm = TRUE)
```

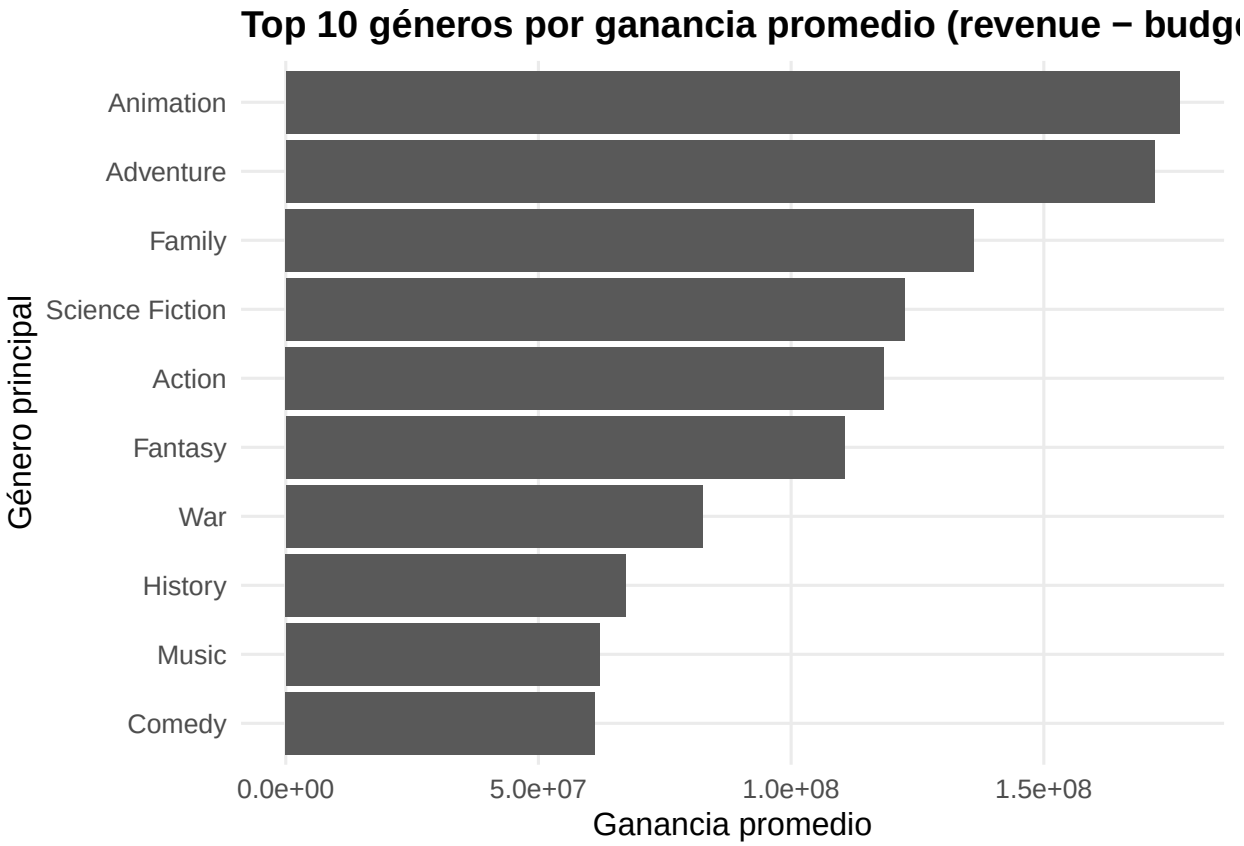
```
  ) %>%
  arrange(desc(ganancia_promedio))

gan_por_genero
```

El análisis de la ganancia total indica que el género Action concentra la mayor cantidad de beneficios acumulados. Esto se explica principalmente por el alto número de películas producidas dentro de este género, lo que incrementa su ganancia total, aunque su rentabilidad promedio sea menor que la de otros géneros.

```
genero_mayor_promedio <- gan_por_genero %>% head(1)
genero_mayor_promedio
```

```
genero_mayor_total <- gan_por_genero %>%
  arrange(desc(ganancia_total)) %>%
  head(1)

genero_mayor_total
```

```
library(ggplot2)

top10_gan <- gan_por_genero %>% head(10)

ggplot(top10_gan, aes(x = reorder(main_genre, ganancia_promedio), y = ganancia_promedio)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Top 10 géneros por ganancia promedio (revenue - budget)",
    x = "Género principal",
    y = "Ganancia promedio"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```

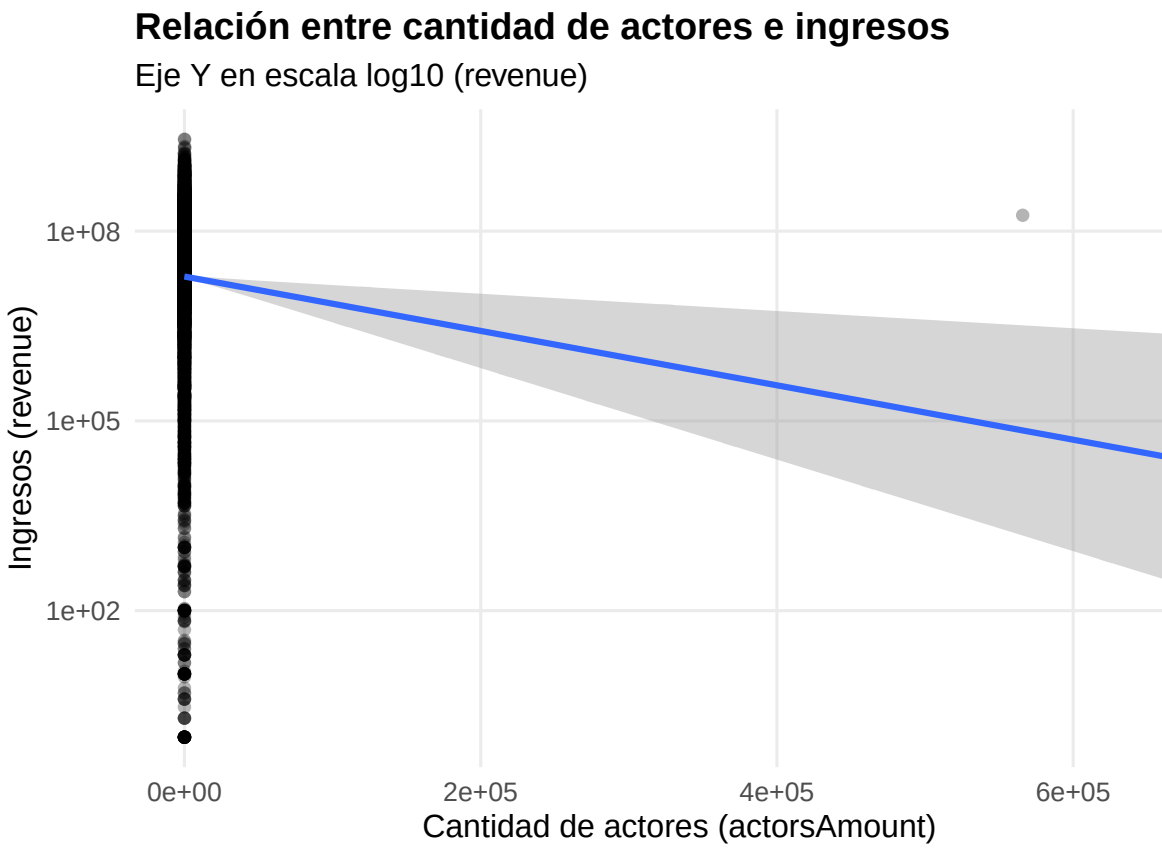


El gráfico de los 10 géneros con mayor ganancia promedio refuerza que Animation, Adventure y Family lideran en términos de rentabilidad promedio, mientras que géneros como Action y Comedy, aunque populares, presentan una ganancia promedio más moderada.

```
##limpiar datos
df_actores_ing <- movies %>%
  filter(!is.na(actorsAmount), !is.na(revenue)) %>%
  filter(actorsAmount > 0, revenue > 0)

summary(df_actores_ing$actorsAmount)
summary(df_actores_ing$revenue)
```

```
##gráfico de relación entre cantidad de actores e ingresos
ggplot(df_actores_ing, aes(x = actorsAmount, y = revenue)) +
  geom_point(alpha = 0.3) +
  scale_y_log10() +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Relación entre cantidad de actores e ingresos",
    subtitle = "Eje Y en escala log10 (revenue)",
    x = "Cantidad de actores (actorsAmount)",
    y = "Ingresos (revenue)"
```

```
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Análisis de actores

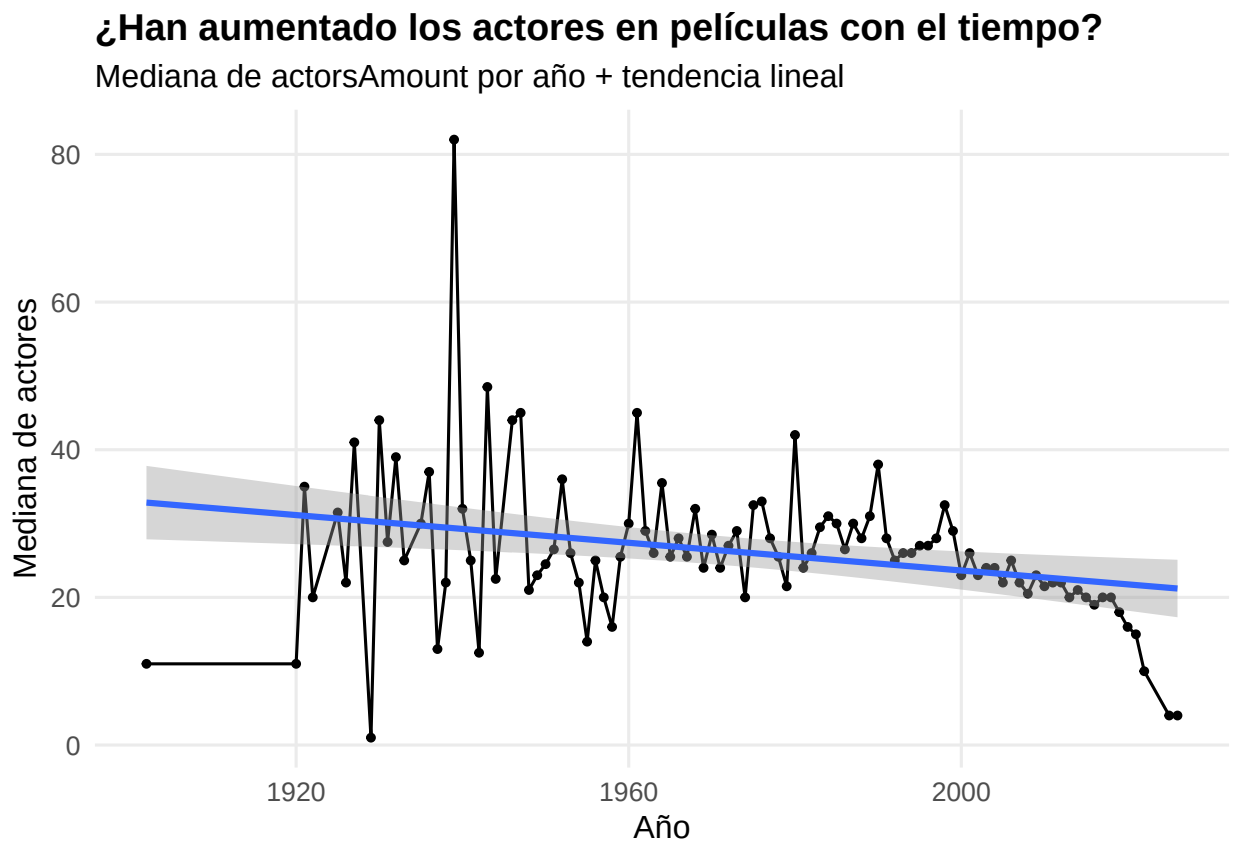
Interpretación

El gráfico de dispersión muestra una alta variabilidad en los ingresos de las películas para cualquier cantidad de actores. Aunque se observa una ligera tendencia negativa en la línea de ajuste, la dispersión de los datos es elevada, lo que indica que la cantidad de actores no tiene una relación clara ni directa con los ingresos de las películas. Por lo tanto, no se encuentra evidencia suficiente para afirmar que un mayor número de actores implique mayores ingresos.

```
##Actores por año
df_actores_anio <- movies %>%
  filter(!is.na(releaseYear), !is.na(actorsAmount)) %>%
  mutate(releaseYear = as.integer(releaseYear)) %>%
  filter(actorsAmount > 0)

actores_por_anio <- df_actores_anio %>%
  group_by(releaseYear) %>%
  summarise(
    peliculas = n(),
    promedio_actores = mean(actorsAmount, na.rm = TRUE),
    mediana_actores = median(actorsAmount, na.rm = TRUE)
  ) %>%
```

```
  arrange(releaseYear)

actores_por_anio
```

```
##Gráfico de tendencia
ggplot(actores_por_anio, aes(x = releaseYear, y = mediana_actores)) +
  geom_line() +
  geom_point(size = 1) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "¿Han aumentado los actores en películas con el tiempo?",
    subtitle = "Mediana de actorsAmount por año + tendencia lineal",
    x = "Año",
    y = "Mediana de actores"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El análisis de la cantidad de actores a lo largo del tiempo, utilizando la mediana por año, no muestra un incremento sostenido en el número de actores en las películas. Por el contrario, la tendencia lineal sugiere una ligera disminución en la mediana de actores en producciones más recientes. Esto indica que, a lo largo del tiempo, no se han producido películas con elencos significativamente más grandes de manera sistemática.

```r
##Comparación de los últimos años con años recientes
max_anio <- max(df_actores_anio$releaseYear, na.rm = TRUE)
corte <- max_anio - 5

comparacion <- df_actores_anio %>%
  mutate(periodo = ifelse(releaseYear >= corte, "Últimos 5 años", "Años anteriores")) %>%
  group_by(periodo) %>%
  summarise(
    peliculas = n(),
    promedio_actores = mean(actorsAmount, na.rm = TRUE),
    mediana_actores = median(actorsAmount, na.rm = TRUE)
  )

corte
comparacion
```

###Cantidad de hombres y mujeres en las películas

```r
##preparar datos
df_cast <- movies %>%
  filter(!is.na(castWomenAmount), !is.na(castMenAmount)) %>%
  mutate(
    cast_total = castWomenAmount + castMenAmount,
    pct_women = ifelse(cast_total > 0, castWomenAmount / cast_total, NA)
  ) %>%
  filter(!is.na(pct_women), cast_total > 0)

summary(df_cast$castWomenAmount)
summary(df_cast$castMenAmount)
summary(df_cast$pct_women)
```

Interpretación

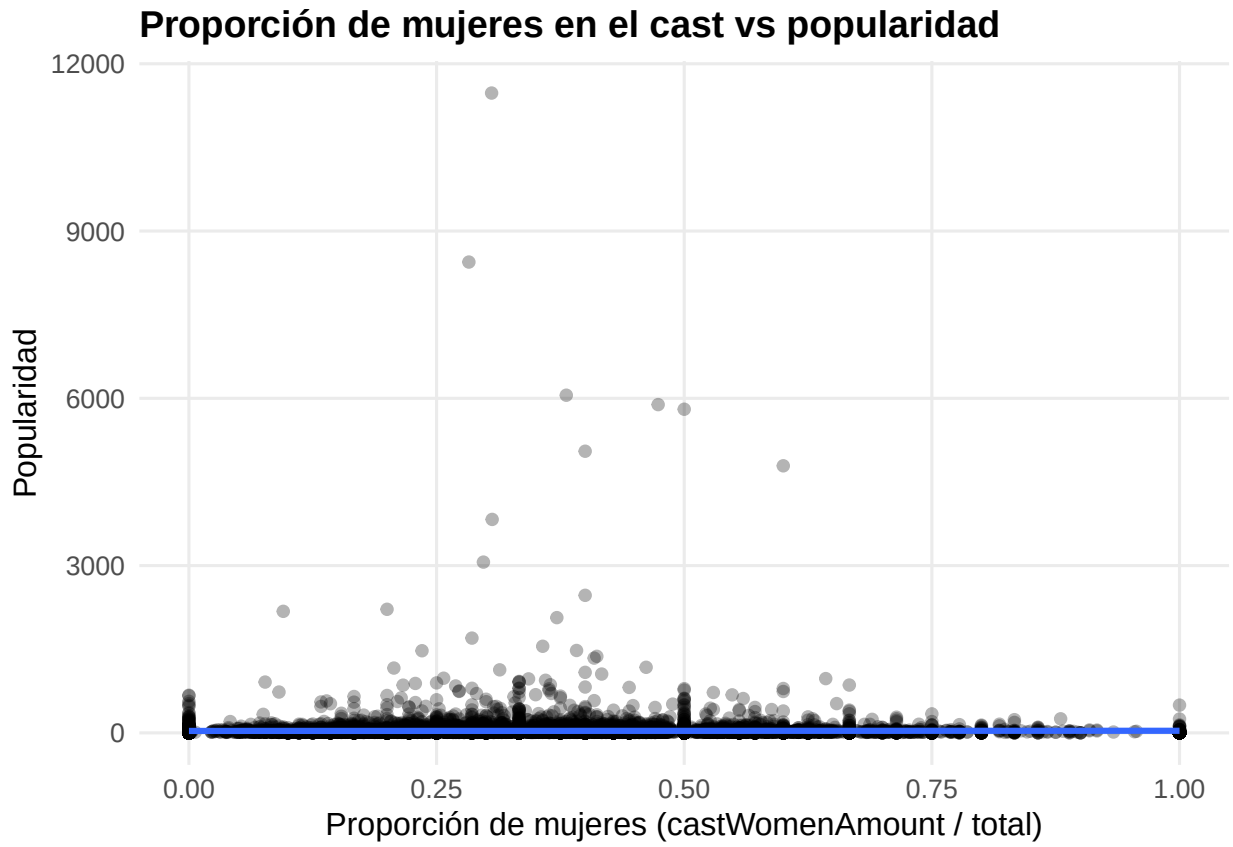
La correlación de Spearman entre la cantidad de mujeres en el elenco y la popularidad de las películas es positiva y estadísticamente significativa (= 0.57, p < 0.05). Esto indica que las películas con un mayor número de actrices tienden a presentar mayores niveles de popularidad. Sin embargo, este resultado debe interpretarse con cautela, ya que la cantidad de mujeres en el reparto puede estar asociada al tamaño total del elenco y a producciones de mayor escala.

```r
# Popularidad vs cantidad de mujeres/hombres y proporción
cor_pop_women <- cor.test(df_cast$castWomenAmount, df_cast$popularity, method = "spearman")
cor_pop_men   <- cor.test(df_cast$castMenAmount,   df_cast$popularity, method = "spearman")
cor_pop_pctw  <- cor.test(df_cast$pct_women,       df_cast$popularity, method = "spearman")

cor_pop_women
cor_pop_men
cor_pop_pctw
```

Interpretación

La correlación entre la cantidad de hombres en el elenco y la popularidad de las películas es positiva y estadísticamente significativa (= 0.61, p < 0.05). Este resultado sugiere que películas con elencos más numerosos, independientemente del género de los actores, tienden a ser más populares.

```r
ggplot(df_cast, aes(x = pct_women, y = popularity)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Proporción de mujeres en el cast vs popularidad",
    x = "Proporción de mujeres (castWomenAmount / total)",
    y = "Popularidad"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El gráfico de dispersión entre la proporción de mujeres en el elenco y la popularidad muestra una alta dispersión de los datos y una línea de tendencia prácticamente horizontal. Esto indica que no existe una relación clara entre el balance de género en el reparto y la popularidad de las películas. En consecuencia, la proporción de mujeres en el elenco no parece ser un factor determinante en la recepción del público.

```r
df_cast_rev <- df_cast %>%
  filter(!is.na(revenue), revenue > 0)

summary(df_cast_rev$revenue)
```

Interpretación

Las correlaciones de Spearman entre la cantidad de mujeres y hombres en el elenco y los ingresos de las

películas son positivas y estadísticamente significativas. Esto sugiere que las películas con elencos más numerosos tienden a generar mayores ingresos. Sin embargo, estas relaciones deben interpretarse con cautela, ya que la cantidad de actores y actrices está estrechamente asociada al tamaño general de la producción y no necesariamente al género de los integrantes del reparto.

```
cor_rev_women <- cor.test(df_cast_rev$castWomenAmount, df_cast_rev$revenue, method = "spearman")
cor_rev_men   <- cor.test(df_cast_rev$castMenAmount,   df_cast_rev$revenue, method = "spearman")
cor_rev_pctw  <- cor.test(df_cast_rev$pct_women,       df_cast_rev$revenue, method = "spearman")

cor_rev_women
cor_rev_men
cor_rev_pctw
```

Interpretación

Al analizar la proporción de mujeres en el elenco en relación con los ingresos, la correlación obtenida es muy débil y cercana a cero, aunque estadísticamente significativa debido al gran tamaño del conjunto de datos. Este resultado indica que el balance de género en el reparto no tiene una relación relevante con el éxito comercial de las películas, una vez que se considera el tamaño total del elenco.

####20 mejores actores según las películas

```
top20_mejor_calif <- movies %>%
  filter(!is.na(voteAvg), !is.na(voteCount)) %>%
  filter(voteCount >= 100) %>%   # puedes cambiar 100 por 50 o 200
  arrange(desc(voteAvg), desc(voteCount)) %>%
  select(title, director, voteAvg, voteCount, releaseYear) %>%
  head(20)

top20_mejor_calif
```

Interpretación

El listado de las 20 películas mejor calificadas muestra que las producciones con mayor calificación promedio suelen contar con un número considerable de votos, lo que refuerza la confiabilidad de la evaluación del público. En este grupo aparecen tanto películas clásicas ampliamente reconocidas como producciones más recientes, lo que sugiere que las altas calificaciones no dependen únicamente del año de estreno, sino de la recepción general de la audiencia.

```
directores_top20 <- top20_mejor_calif %>%
  distinct(director)

directores_top20
```

Interpretación

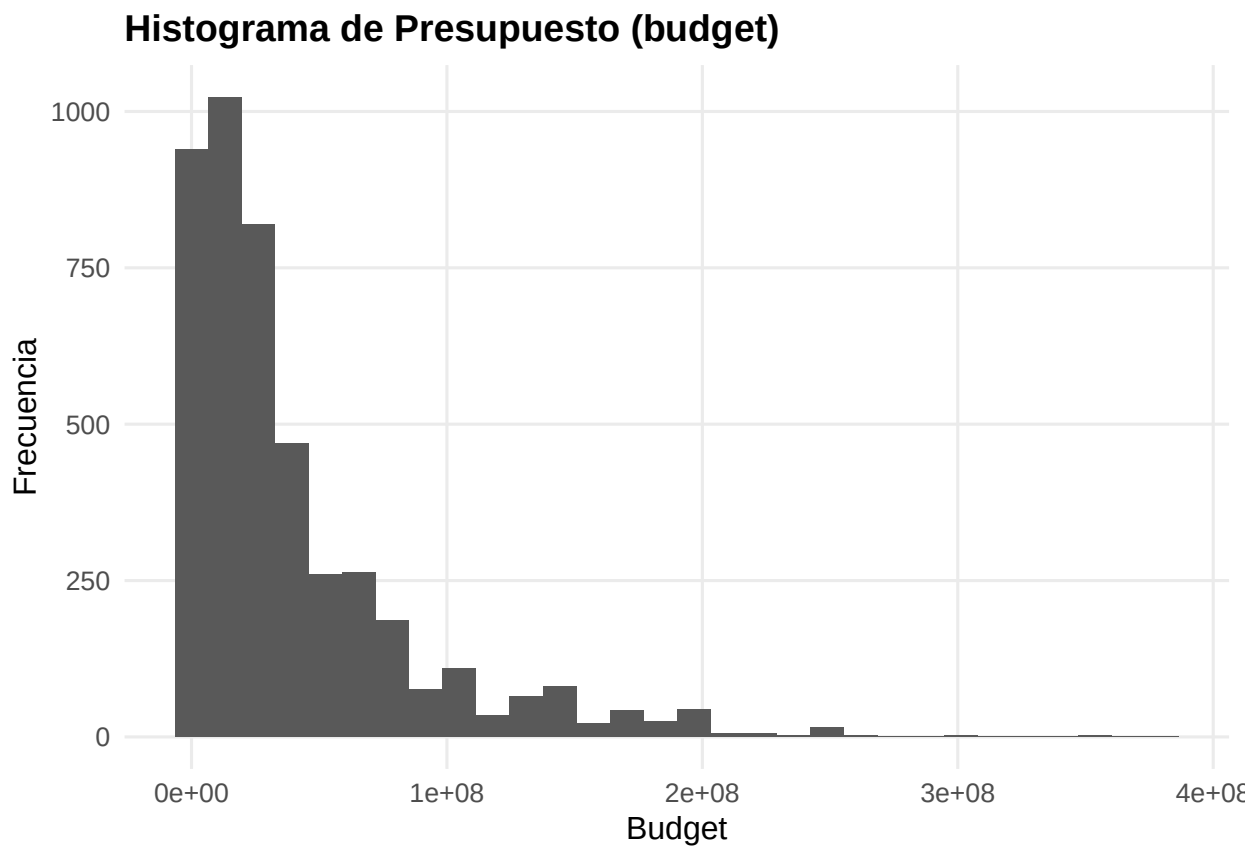
El conjunto de directores asociados a las 20 películas mejor calificadas es diverso, con varios directores que aparecen una sola vez. Esto indica que no existe un único director dominante en este grupo, sino que la alta calificación puede alcanzarse desde distintos estilos de dirección y contextos cinematográficos.

####Correlación de presupuestos con ingresos

```
##limpiar datos
df_br <- movies %>%
  filter(!is.na(budget), !is.na(revenue)) %>%
```

```r
  filter(budget > 0, revenue > 0)

summary(df_br$budget)
summary(df_br$revenue)
```

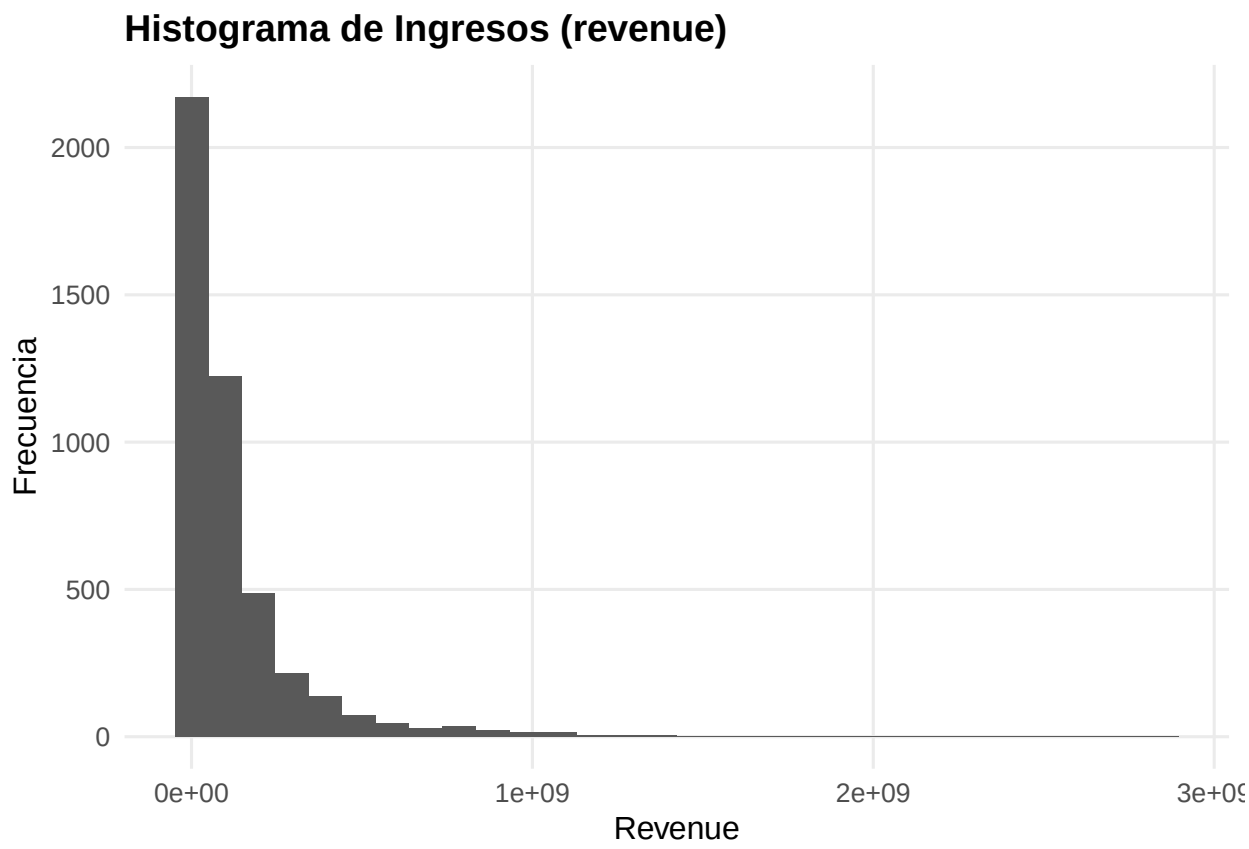
```r
##histograma
ggplot(df_br, aes(x = budget)) +
  geom_histogram(bins = 30) +
  labs(title = "Histograma de Presupuesto (budget)", x = "Budget", y = "Frecuencia") +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El histograma del presupuesto muestra una distribución fuertemente asimétrica hacia la derecha, con la mayoría de las películas concentradas en presupuestos bajos y un número reducido de producciones con presupuestos extremadamente altos. Este patrón es característico de la industria cinematográfica, donde pocas películas concentran grandes inversiones.

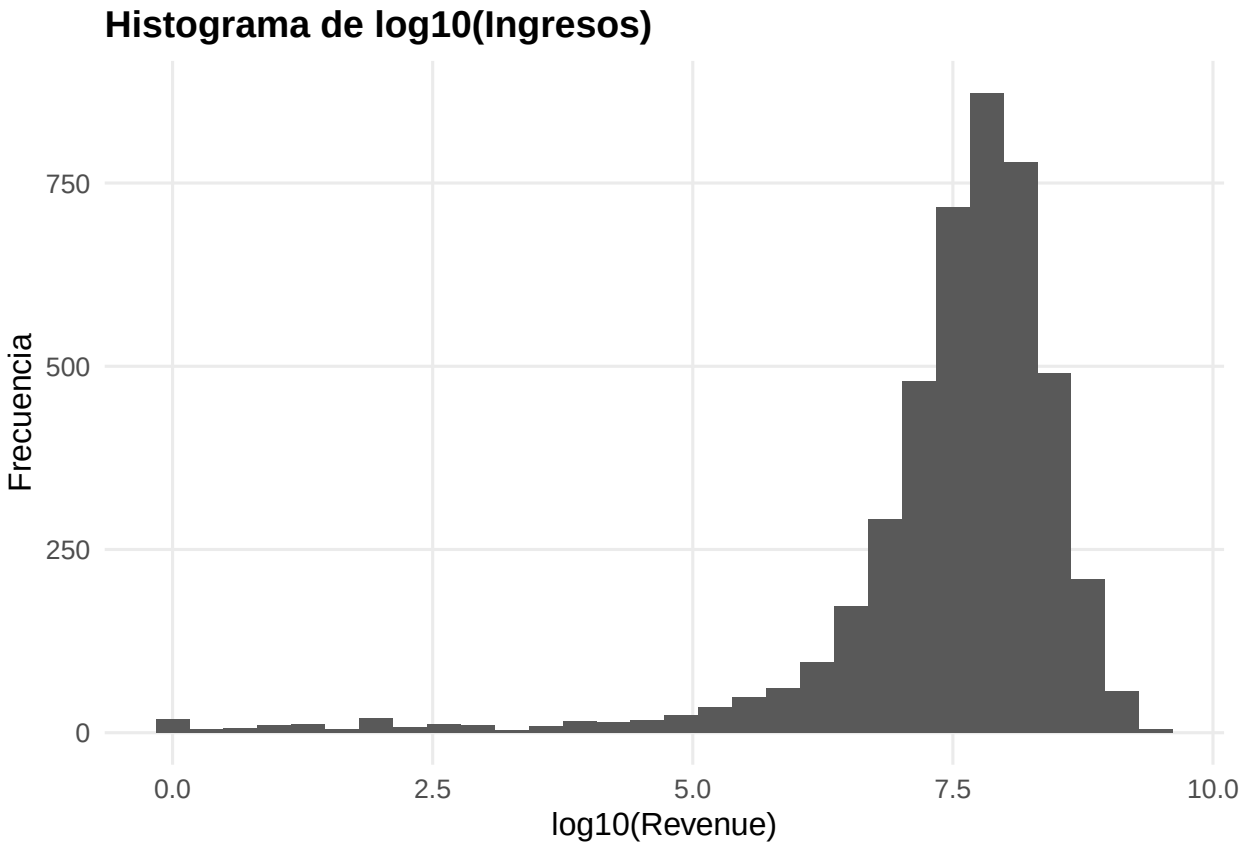
```r
ggplot(df_br, aes(x = revenue)) +
  geom_histogram(bins = 30) +
  labs(title = "Histograma de Ingresos (revenue)", x = "Revenue", y = "Frecuencia") +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El histograma de los ingresos presenta una asimetría aún más marcada que la del presupuesto, evidenciando que la mayoría de las películas genera ingresos moderados o bajos, mientras que un pequeño grupo de producciones alcanza ingresos muy elevados. Esto refleja la naturaleza altamente concentrada del éxito comercial en la industria.

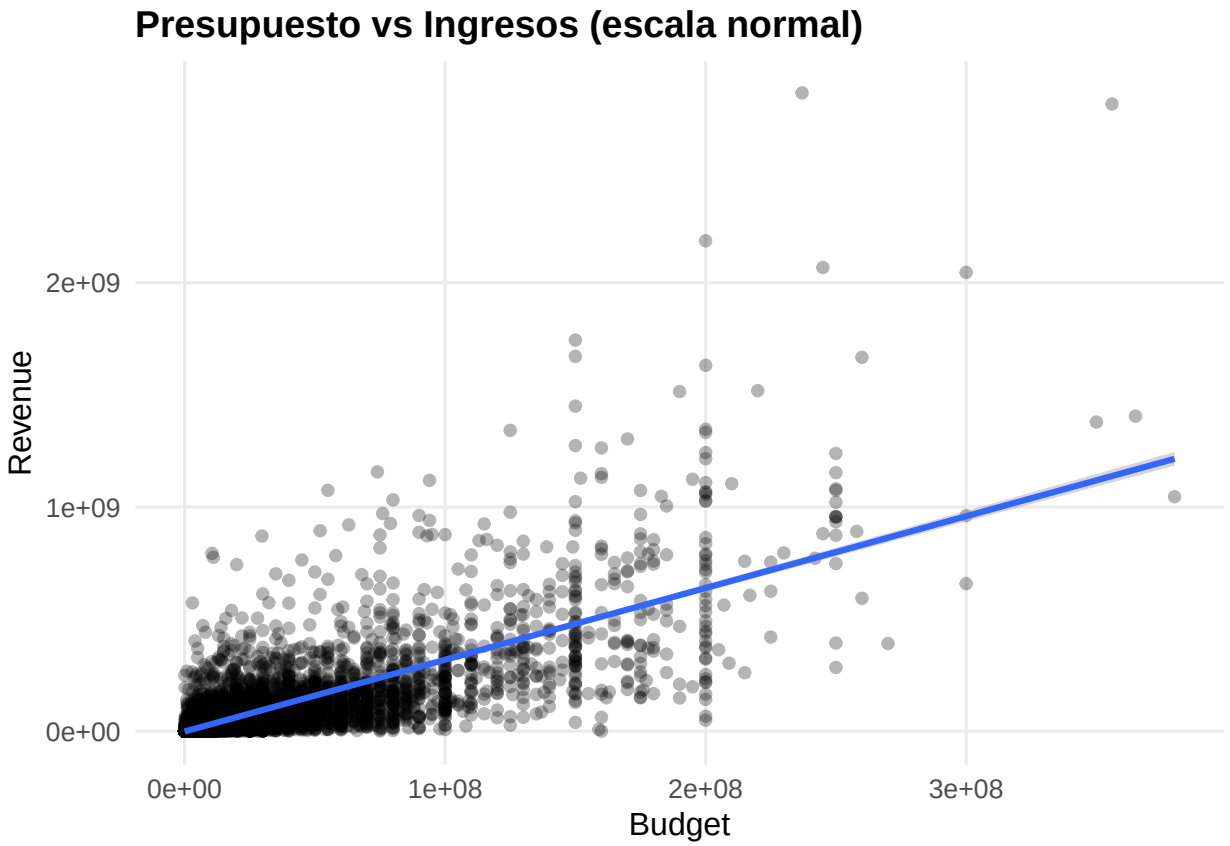
```
ggplot(df_br, aes(x = log10(revenue))) +
  geom_histogram(bins = 30) +
  labs(title = "Histograma de log10(Ingresos)", x = "log10(Revenue)", y = "Frecuencia") +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

Al aplicar una transformación logarítmica a los ingresos, la distribución se vuelve más cercana a una forma simétrica. Esto facilita la visualización y el análisis de los datos, y confirma que la asimetría observada en la escala original se debe a la presencia de valores extremos.

```
ggplot(df_br, aes(x = budget, y = revenue)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Presupuesto vs Ingresos (escala normal)",
    x = "Budget",
    y = "Revenue"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El gráfico de dispersión entre presupuesto e ingresos muestra una relación positiva: en general, las películas con mayor presupuesto tienden a generar mayores ingresos. Sin embargo, la dispersión de los puntos indica que un alto presupuesto no garantiza necesariamente un alto ingreso, lo que sugiere la influencia de otros factores como el género, la recepción del público o las estrategias de marketing.

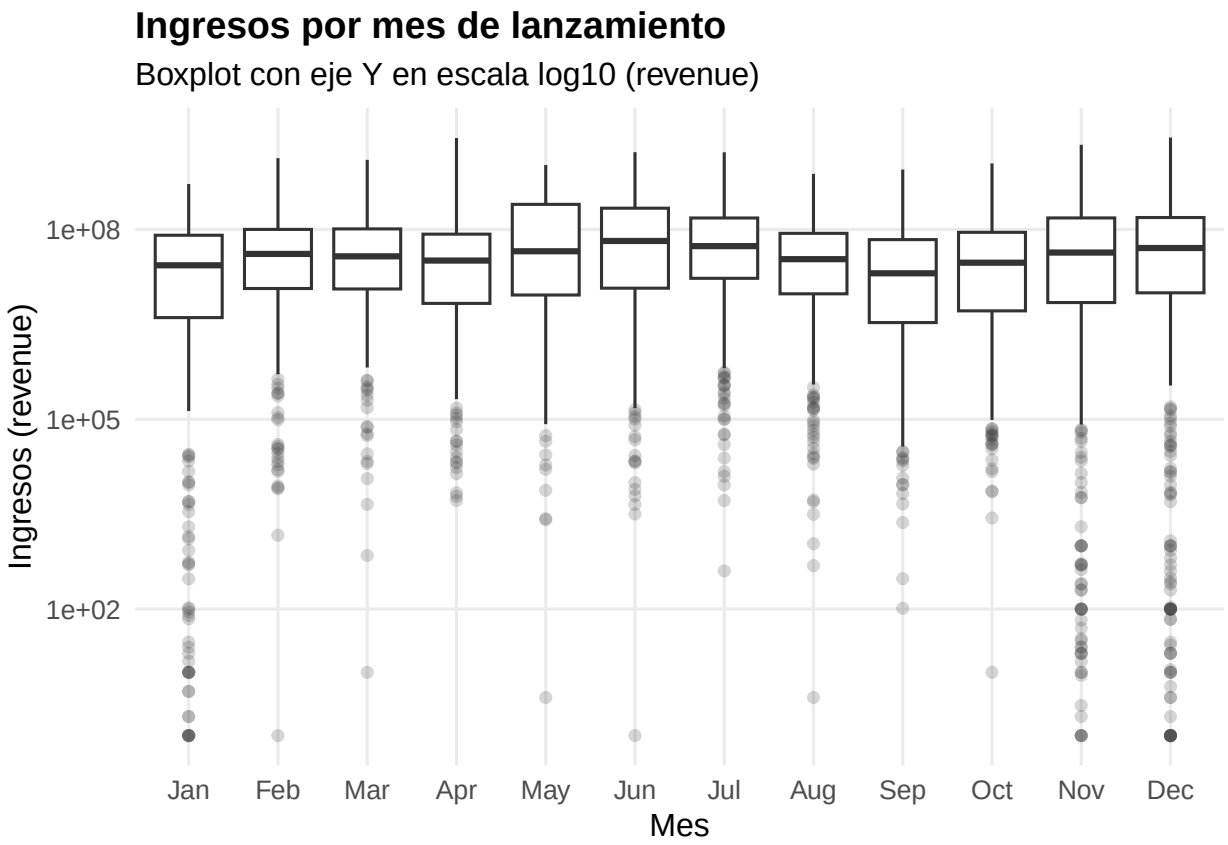
####Meses de lanzamiento con ingresos

```
# Preparar datos con mes
df_mes <- movies %>%
  mutate(release_dt = suppressWarnings(ymd(releaseDate))) %>%
  filter(!is.na(release_dt), !is.na(revenue), revenue > 0) %>%
  mutate(
    mes_num = month(release_dt),
    mes = factor(month(release_dt, label = TRUE, abbr = TRUE),
                 levels = month(1:12, label = TRUE, abbr = TRUE))
  )

table(df_mes$mes)
```

```
##resumen de mes
resumen_mes <- df_mes %>%
  group_by(mes) %>%
  summarise(
    peliculas = n(),
    revenue_mediana = median(revenue, na.rm = TRUE),
    revenue_promedio = mean(revenue, na.rm = TRUE)
```

```
  ) %>%
  arrange(desc(revenue_mediana))

resumen_mes
```

```
ggplot(df_mes, aes(x = mes, y = revenue)) +
  geom_boxplot(outlier.alpha = 0.2) +
  scale_y_log10() +
  labs(
    title = "Ingresos por mes de lanzamiento",
    subtitle = "Boxplot con eje Y en escala log10 (revenue)",
    x = "Mes",
    y = "Ingresos (revenue)"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

El análisis de los ingresos según el mes de lanzamiento muestra diferencias claras entre los distintos meses del año. A partir del boxplot y del resumen estadístico, se observa que los meses de junio, julio y diciembre presentan las mayores medianas de ingresos, destacando junio como el mes con la mediana más alta.

Los meses de verano (junio y julio) y el cierre de año (diciembre) concentran tanto un mayor número de

películas con ingresos como valores más elevados en términos medianos, lo que sugiere que estos periodos están asociados a un mejor desempeño comercial. En contraste, meses como septiembre y octubre presentan medianas de ingresos más bajas.

La prueba de Kruskal–Wallis confirma que las diferencias en los ingresos entre los meses de lanzamiento son estadísticamente significativas ($p < 0.05$), lo que indica que el mes de estreno sí está asociado al desempeño comercial de las películas.

```
mes_mejor <- resumen_mes %>% head(1)
mes_mejor
```

```
kruskal.test(revenue ~ mes, data = df_mes)
```

####Mes con lanzamientos y mejores ingresos

```
##limpiar datos
movies2 <- movies %>%
  mutate(
    release_dt = suppressWarnings(ymd(releaseDate)),
    release_dt = ifelse(is.na(release_dt) & !is.na(releaseYear),
                        as.Date(paste0(releaseYear, "-12-31")),
                        release_dt) %>% as.Date(),
    mes = factor(month(release_dt, label = TRUE, abbr = TRUE),
                 levels = month(1:12, label = TRUE, abbr = TRUE))
  )
```

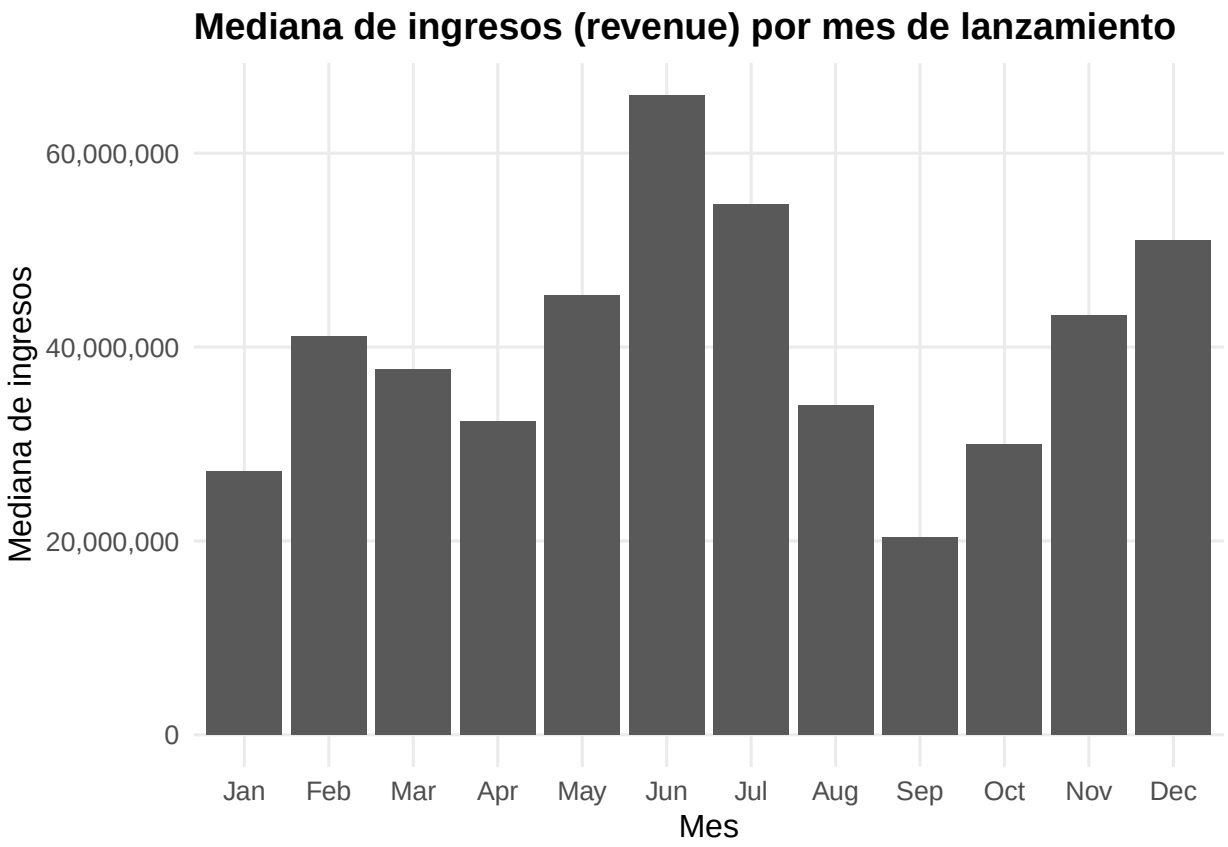
```
##tabla de ingresos por mes
df_rev_mes <- movies2 %>%
  filter(!is.na(release_dt), !is.na(revenue), revenue > 0)

ingresos_por_mes <- df_rev_mes %>%
  group_by(mes) %>%
  summarise(
    peliculas_con_revenue = n(),
    revenue_mediana = median(revenue, na.rm = TRUE),
    revenue_promedio = mean(revenue, na.rm = TRUE)
  ) %>%
  arrange(desc(revenue_mediana))

ingresos_por_mes
```

Los resultados muestran que junio, julio y diciembre son los meses con mayores ingresos medianos, siendo junio el mes con mejor desempeño comercial.

```
##top 3 mejores meses
top3_meses <- ingresos_por_mes %>% head(3)
top3_meses
```

```
##Gráfico de ingresos por mes
ggplot(ingresos_por_mes, aes(x = mes, y = revenue_mediana)) +
  geom_col() +
  scale_y_continuous(labels = comma) +
  labs(
```

```
    title = "Mediana de ingresos (revenue) por mes de lanzamiento",
    x = "Mes",
    y = "Mediana de ingresos"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

El gráfico de la mediana de ingresos por mes de lanzamiento muestra que los meses de junio y julio presentan las mayores medianas de ingresos, seguidos por diciembre y noviembre. Esto sugiere que los lanzamientos realizados durante el verano y el cierre de año tienden a obtener un mejor desempeño comercial. En contraste, meses como septiembre y enero presentan medianas de ingresos más bajas, indicando un menor rendimiento relativo en esos periodos.

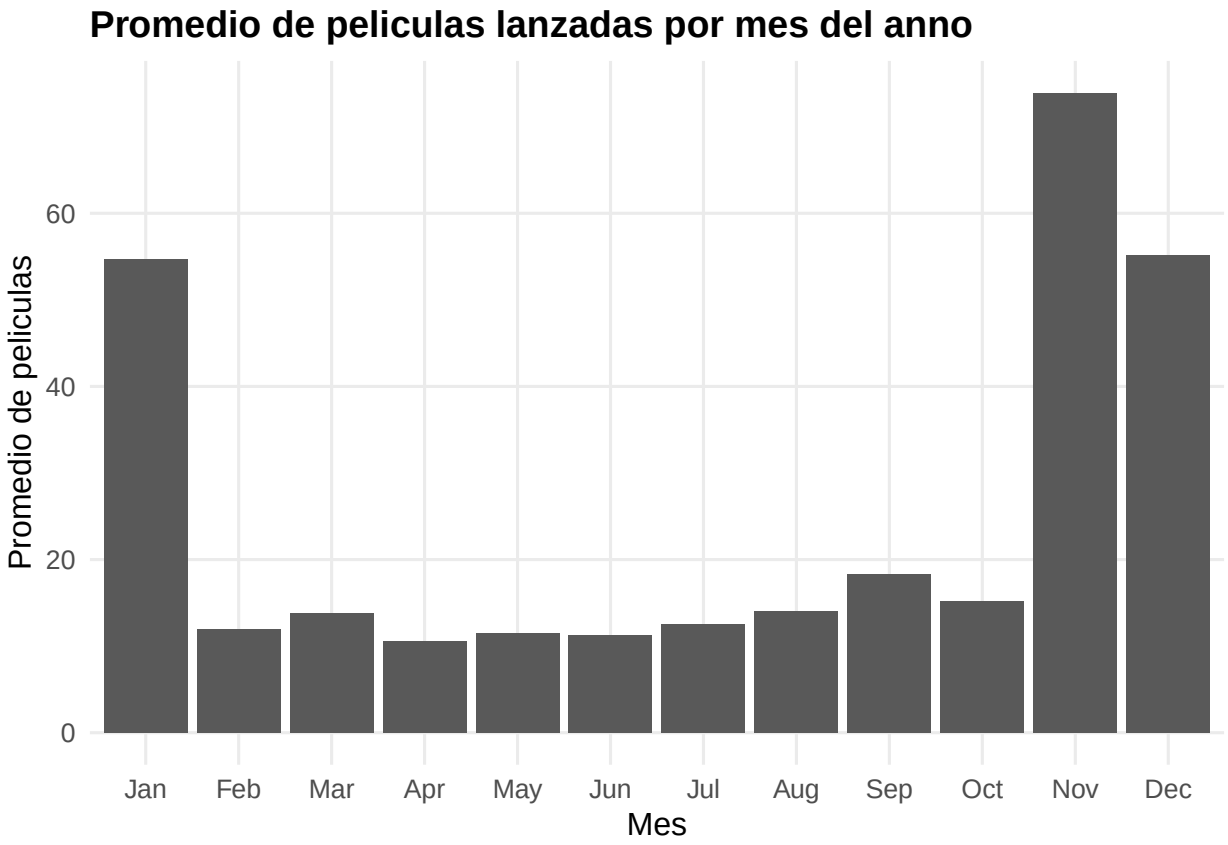
```
## promedio de películas por mes
lanz_por_anio_mes <- movies2 %>%
  filter(!is.na(release_dt)) %>%
  mutate(anio = year(release_dt)) %>%
  group_by(anio, mes) %>%
  summarise(peliculas = n()) %>%
  ungroup()
promedio_por_mes_del_anio <- lanz_por_anio_mes %>%
  group_by(mes) %>%
```

```
  summarise(
    promedio_peliculas = mean(peliculas),
    mediana_peliculas  = median(peliculas)
  ) %>%
  arrange(desc(promedio_peliculas))

promedio_por_mes_del_anio
```

Interpretación

El promedio de películas lanzadas por mes muestra que noviembre es el mes con mayor cantidad de estrenos en promedio a lo largo de los años, seguido por diciembre y enero. Esto sugiere que la industria cinematográfica concentra una mayor cantidad de lanzamientos hacia el cierre y el inicio del año. La diferencia observada entre el promedio y la mediana indica que existen algunos años con picos elevados de estrenos, particularmente en estos meses, mientras que en otros meses la cantidad de lanzamientos se mantiene más estable.

```
##gráfico
ggplot(promedio_por_mes_del_anio, aes(x = mes, y = promedio_peliculas)) +
  geom_col() +
  labs(
    title = "Promedio de peliculas lanzadas por mes del anno",
    x = "Mes",
    y = "Promedio de peliculas"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

El gráfico del promedio de películas lanzadas por mes muestra que noviembre es el mes con mayor cantidad de estrenos en promedio, seguido por diciembre y enero. Esto indica que la industria cinematográfica concentra una mayor actividad de lanzamientos hacia el cierre y el inicio del año. En contraste, meses como abril, mayo y febrero presentan promedios más bajos, lo que sugiere una menor intensidad de estrenos durante esos periodos.

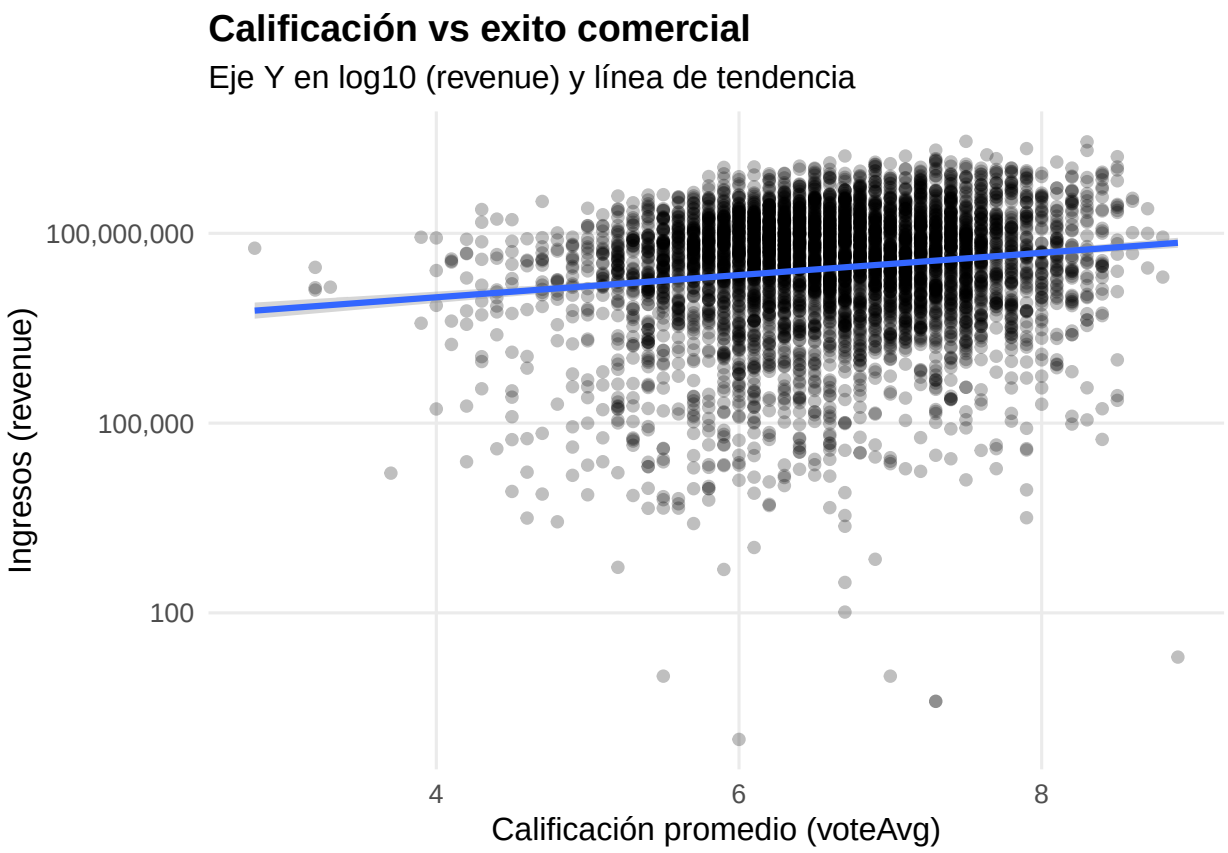
####4.14

```
##preparar datos
df_calif <- movies %>%
  filter(!is.na(voteAvg), !is.na(voteCount), !is.na(revenue)) %>%
  filter(voteAvg > 0, revenue > 0) %>%
  filter(voteCount >= 100)

summary(df_calif$voteAvg)
summary(df_calif$revenue)
summary(df_calif$voteCount)
```

```
##correlacion
cor_spear <- cor.test(df_calif$voteAvg, df_calif$revenue, method = "spearman")
cor_spear
```

```
##grafico
ggplot(df_calif, aes(x = voteAvg, y = revenue)) +
  geom_point(alpha = 0.25) +
```

```
  scale_y_log10(labels = comma) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Calificación vs exito comercial",
    subtitle = "Eje Y en log10 (revenue) y línea de tendencia",
    x = "Calificación promedio (voteAvg)",
    y = "Ingresos (revenue)"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold")
  )
```



Interpretación

La correlación de Spearman entre la calificación promedio de los usuarios y los ingresos de las películas es positiva pero débil (0.11), aunque estadísticamente significativa. El gráfico de dispersión muestra una ligera tendencia ascendente, lo que indica que películas mejor calificadas tienden a generar mayores ingresos en promedio. Sin embargo, la alta dispersión de los datos sugiere que la calificación por sí sola no es un factor determinante del éxito comercial.

Las calificaciones muestran una relación positiva débil con los ingresos, lo que sugiere que una buena recepción del público ayuda al desempeño comercial, pero no lo garantiza.

```
## exito como gancia
df_gan <- movies %>%
```

```
  filter(!is.na(voteAvg), !is.na(voteCount), !is.na(revenue), !is.na(budget)) %>%
  filter(voteAvg > 0, revenue > 0, budget > 0) %>%
  filter(voteCount >= 100) %>%
  mutate(ganancia = revenue - budget)

cor_gan <- cor.test(df_gan$voteAvg, df_gan$ganancia, method = "spearman")
cor_gan
```

###4.15

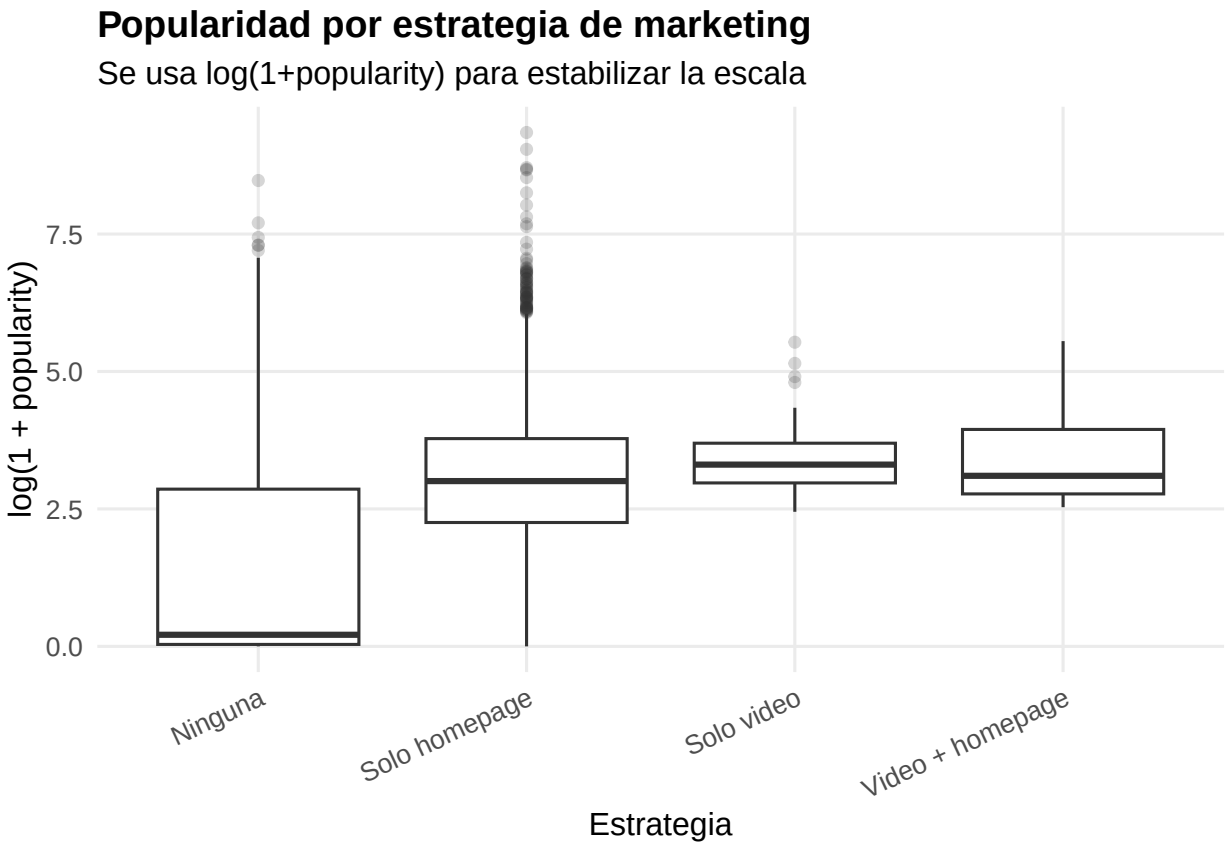
```
##limpieza de datos
movies_mkt <- movies %>%
  mutate(
    tiene_video = ifelse(is.na(video), FALSE, video),
    tiene_homepage = ifelse(is.na(homePage) | homePage == "", FALSE, TRUE),
    estrategia = case_when(
      tiene_video == FALSE & tiene_homepage == FALSE ~ "Ninguna",
      tiene_video == TRUE  & tiene_homepage == FALSE ~ "Solo video",
      tiene_video == FALSE & tiene_homepage == TRUE  ~ "Solo homepage",
      tiene_video == TRUE  & tiene_homepage == TRUE  ~ "Video + homepage"
    )
  )
table(movies_mkt$estrategia)
```

```
##popularidad de las estrategias
res_pop <- movies_mkt %>%
  filter(!is.na(popularity)) %>%
  group_by(estrategia) %>%
  summarise(
    peliculas = n(),
    pop_mediana = median(popularity, na.rm = TRUE),
    pop_promedio = mean(popularity, na.rm = TRUE)
  ) %>%
  arrange(desc(pop_mediana))

res_pop
```

```
res_rev <- movies_mkt %>%
  filter(!is.na(revenue), revenue > 0) %>%
  group_by(estrategia) %>%
  summarise(
    peliculas = n(),
    rev_mediana = median(revenue, na.rm = TRUE),
    rev_promedio = mean(revenue, na.rm = TRUE)
  ) %>%
  arrange(desc(rev_mediana))

res_rev
```

```
##grafico de popularidad por estrategia
ggplot(movies_mkt, aes(x = estrategia, y = log1p(popularity))) +
  geom_boxplot(outlier.alpha = 0.2) +
```

```
  labs(
    title = "Popularidad por estrategia de marketing",
    subtitle = "Se usa log(1+popularity) para estabilizar la escala",
    x = "Estrategia",
    y = "log(1 + popularity)"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        axis.text.x = element_text(angle = 25, hjust = 1),
        plot.title = element_text(face = "bold"))
```

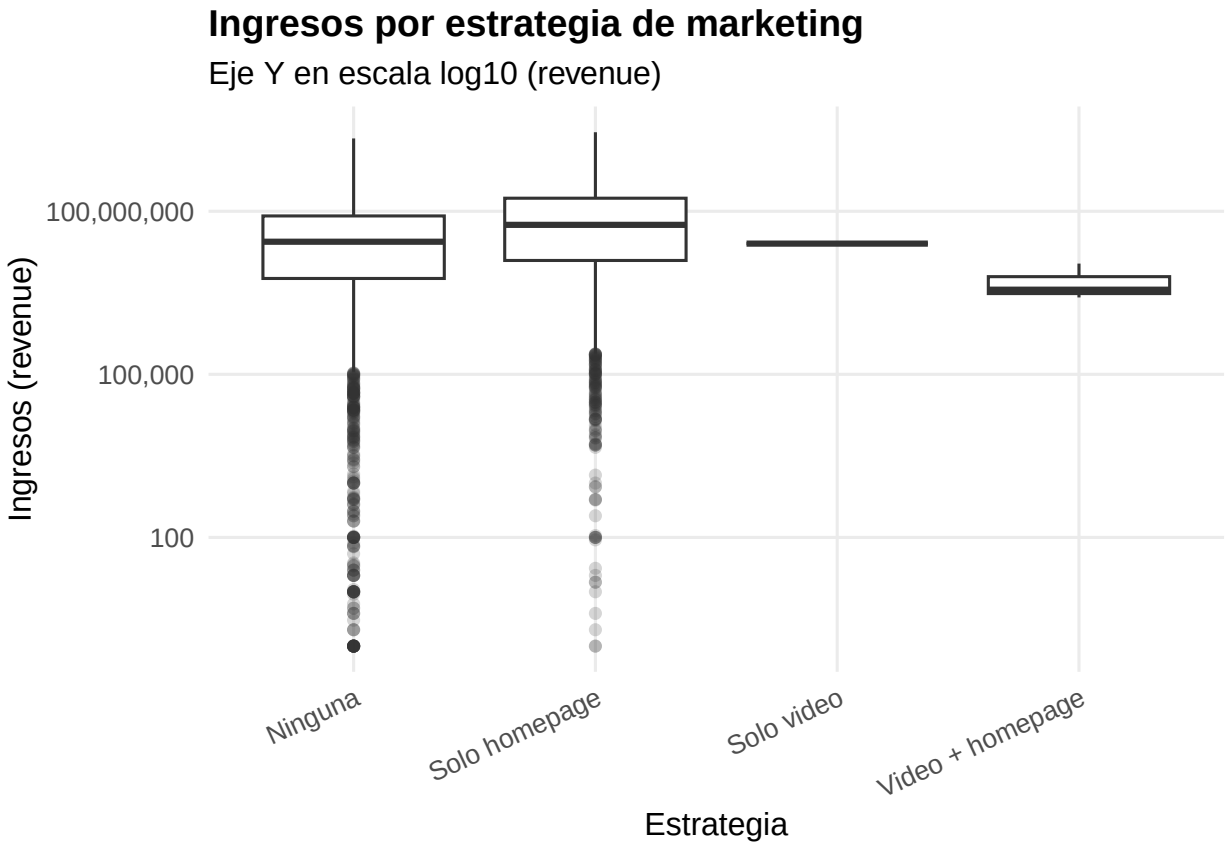


Interpretación

El boxplot de popularidad por estrategia de marketing muestra diferencias claras entre las categorías analizadas. Las películas que cuentan con algún tipo de estrategia de marketing digital (homepage, video o ambos) presentan niveles de popularidad considerablemente mayores que aquellas que no utilizan ninguna estrategia. En particular, las películas con homepage y las que combinan video con homepage muestran las medianas de popularidad más altas. Esto sugiere que la presencia de recursos promocionales en línea está asociada con una mayor visibilidad y atención del público.

```
##grafico de ingresos por estrategia
df_rev_plot <- movies_mkt %>%
  filter(!is.na(revenue), revenue > 0)

ggplot(df_rev_plot, aes(x = estrategia, y = revenue)) +
  geom_boxplot(outlier.alpha = 0.2) +
  scale_y_log10(labels = comma) +
```

```
  labs(
    title = "Ingresos por estrategia de marketing",
    subtitle = "Eje Y en escala log10 (revenue)",
    x = "Estrategia",
    y = "Ingresos (revenue)"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        axis.text.x = element_text(angle = 25, hjust = 1),
        plot.title = element_text(face = "bold"))
```



Interpretación

El boxplot de ingresos por estrategia de marketing indica que las películas que utilizan una homepage presentan una mediana de ingresos más alta en comparación con aquellas que no emplean ninguna estrategia de marketing. Si bien las categorías "Solo video" y "Video + homepage" cuentan con muy pocas observaciones, se observa que, en general, la presencia de una estrategia de marketing se asocia con mayores ingresos. La alta dispersión refleja que el marketing no garantiza el éxito comercial, pero sí parece contribuir positivamente al desempeño económico.

```
kruskal.test(popularity ~ estrategia, data = movies_mkt)
```

```
kruskal.test(revenue ~ estrategia, data = df_rev_plot)
```

Interpretación

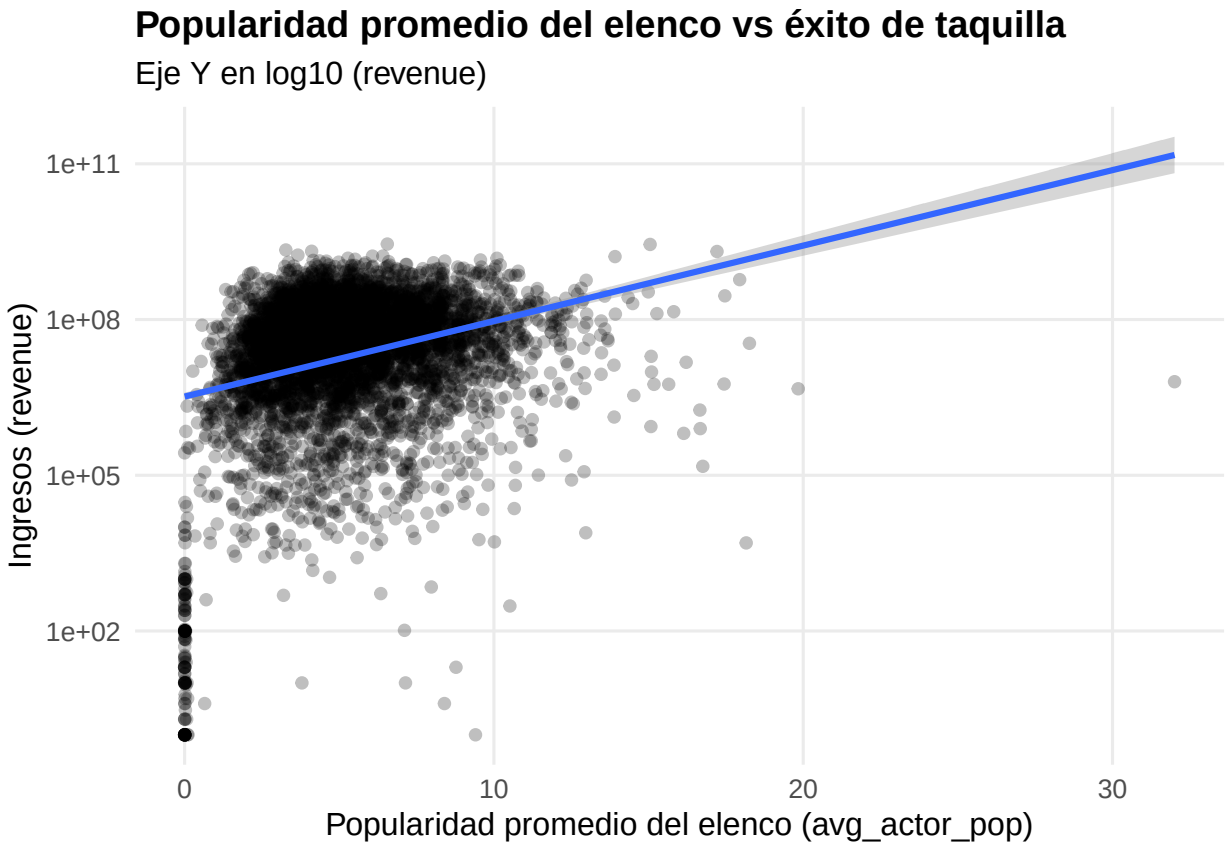
Las pruebas de Kruskal–Wallis confirman que existen diferencias estadísticamente significativas tanto en la popularidad como en los ingresos entre las distintas estrategias de marketing ($p < 0.05$). Esto indica que el tipo de estrategia utilizada está asociado con resultados distintos en términos de visibilidad y desempeño comercial. No obstante, estas diferencias deben interpretarse con cautela, ya que algunas categorías presentan un número reducido de observaciones.

####4.16

```
##limpieza de datos
extraer_numeros <- function(s){
  nums <- str_extract_all(s, "[0-9]+\\.?[0-9]*")[[1]]
  as.numeric(nums)
}

movies_pop <- movies %>%
  mutate(
    actor_pop_nums = lapply(actorsPopularity, extraer_numeros),
    avg_actor_pop = sapply(actor_pop_nums, function(x) if(length(x)==0) NA else mean(x, na.rm=TRUE)),
    max_actor_pop = sapply(actor_pop_nums, function(x) if(length(x)==0) NA else max(x, na.rm=TRUE))
  )

summary(movies_pop$avg_actor_pop)
summary(movies_pop$max_actor_pop)
```

```
## correlacion
df_pop_rev <- movies_pop %>%
  filter(!is.na(revenue), revenue > 0) %>%
  filter(!is.na(avg_actor_pop))


cor_avg <- cor.test(df_pop_rev$avg_actor_pop, df_pop_rev$revenue, method = "spearman")
cor_max <- cor.test(df_pop_rev$max_actor_pop, df_pop_rev$revenue, method = "spearman")

cor_avg
cor_max
```

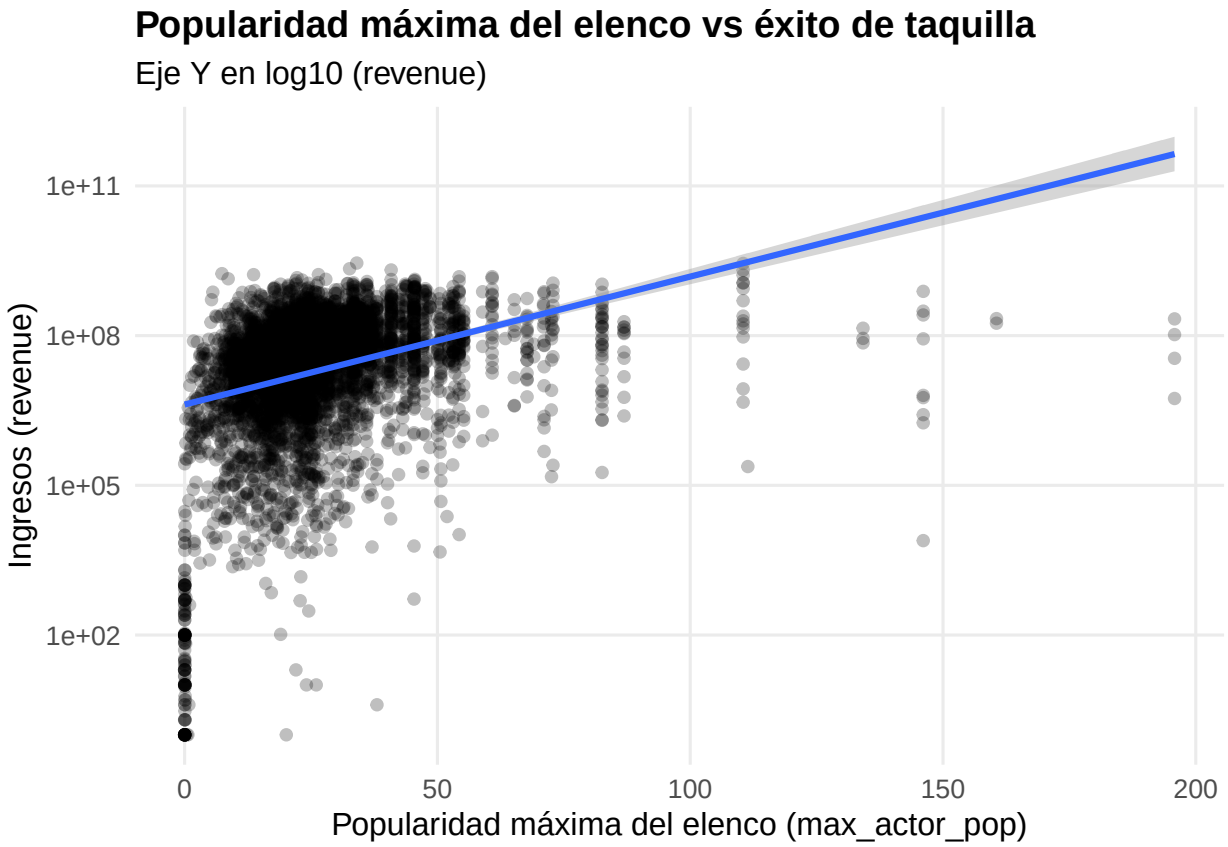
```
##grafico con el promedio popularidad del elenco
ggplot(df_pop_rev, aes(x = avg_actor_pop, y = revenue)) +
  geom_point(alpha = 0.25) +
  scale_y_log10() +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Popularidad promedio del elenco vs éxito de taquilla",
    subtitle = "Eje Y en log10 (revenue)",
    x = "Popularidad promedio del elenco (avg_actor_pop)",
    y = "Ingresos (revenue)"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El gráfico de dispersión entre la popularidad promedio del elenco y los ingresos de las películas muestra una relación positiva clara. La línea de tendencia indica que, en promedio, las películas con elencos más populares tienden a generar mayores ingresos. Aunque existe una alta dispersión de los datos, el patrón general sugiere que la popularidad promedio del elenco está asociada con un mejor desempeño en taquilla.

```
##grafico con el máximo del elenco
ggplot(df_pop_rev, aes(x = max_actor_pop, y = revenue)) +
  geom_point(alpha = 0.25) +
  scale_y_log10() +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Popularidad máxima del elenco vs éxito de taquilla",
    subtitle = "Eje Y en log10 (revenue)",
    x = "Popularidad máxima del elenco (max_actor_pop)",
    y = "Ingresos (revenue)"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(face = "bold"))
```



Interpretación

El gráfico que relaciona la popularidad máxima del elenco con los ingresos de las películas muestra una tendencia positiva aún más marcada. Esto sugiere que contar con al menos un actor o actriz de alta popularidad dentro del elenco puede estar asociado con un mayor éxito comercial. Sin embargo, la dispersión observada indica que la presencia de una estrella no garantiza ingresos elevados, sino que actúa como un factor complementario dentro del desempeño general de la película.

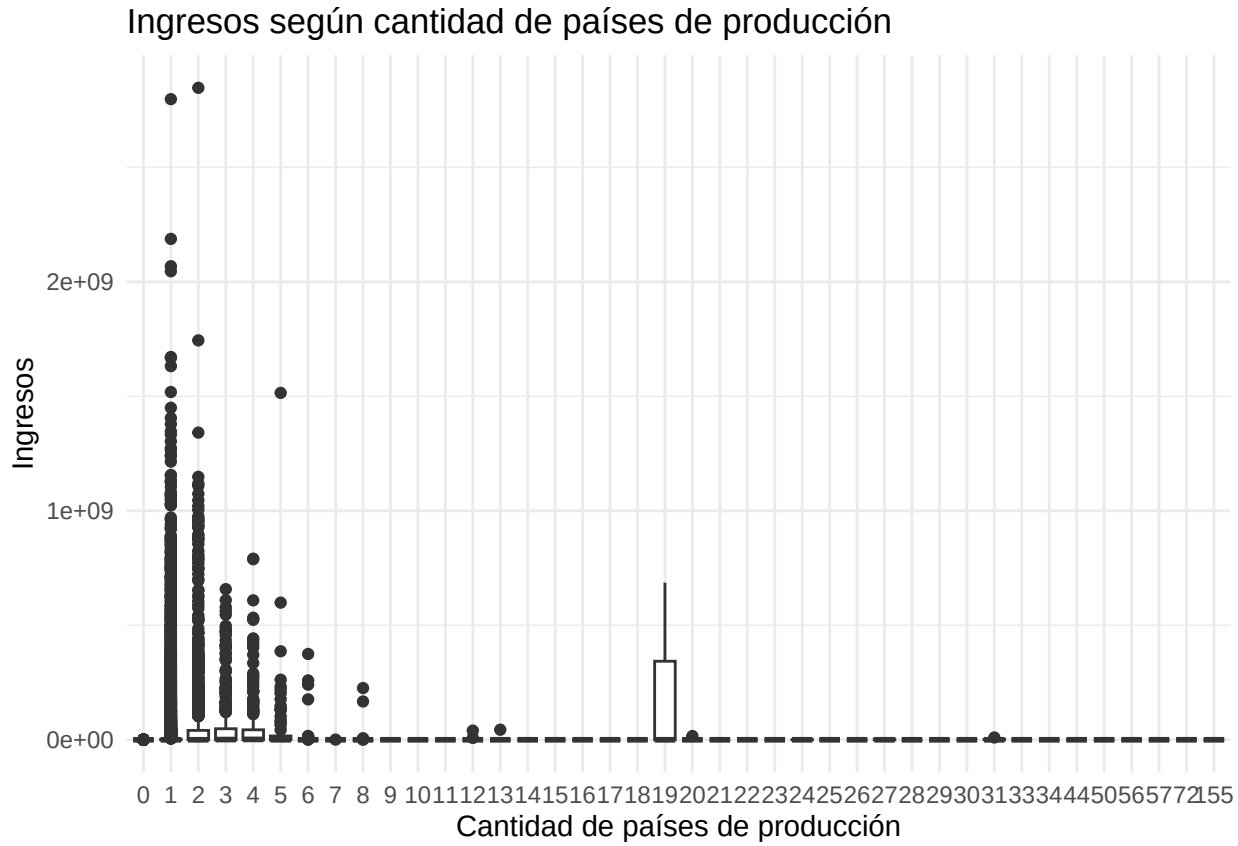
5. Análisis Exploratorio Adicional

En esta sección se plantean seis preguntas adicionales con el objetivo de explorar patrones complementarios en el conjunto de datos, sin repetir los análisis realizados en los incisos anteriores.

```
movies_clean <- movies %>%
  filter(
    !is.na(revenue),
    !is.na(popularity),
    !is.na(runtime),
    !is.na(productionCountriesAmount),
    !is.na(genresAmount),
    !is.na(productionCoAmount),
    !is.na(castWomenAmount),
    !is.na(castMenAmount),
    !is.na(originalLanguage),
    !is.na(releaseYear)
  )
```

5.1 ¿Existe diferencia en los ingresos según la cantidad de países de producción?

```
ggplot(movies_clean, aes(x = factor(productionCountriesAmount), y = revenue)) +
  geom_boxplot() +
  labs(
    x = "Cantidad de países de producción",
    y = "Ingresos",
    title = "Ingresos según cantidad de países de producción"
  ) +
  theme_minimal()
```



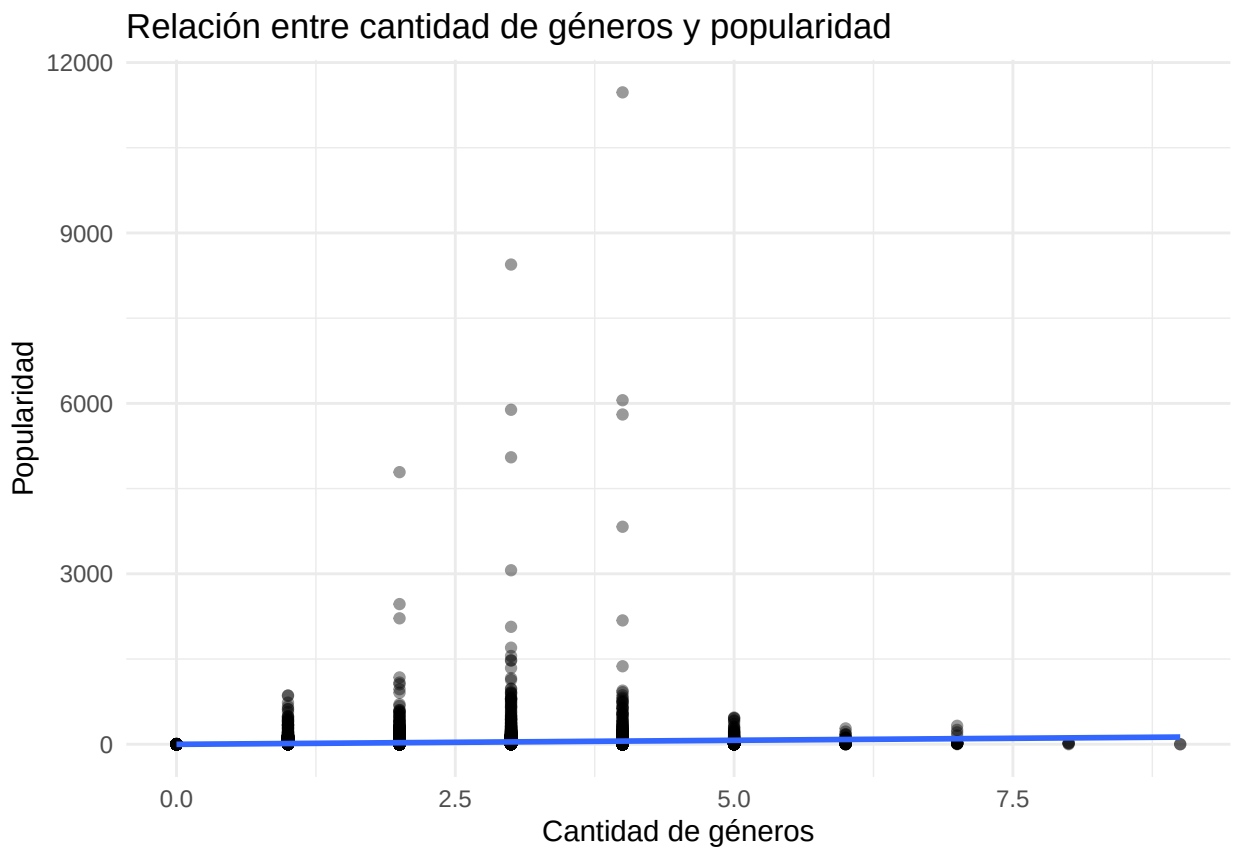
Interpretación

El boxplot muestra que las películas producidas en uno o pocos países concentran la mayor cantidad de ingresos y presentan una alta variabilidad. No se observa una tendencia clara de que un mayor número de países de producción implique mayores ingresos; por el contrario, las coproducciones con muchos países son menos frecuentes y no destacan consistentemente por mayores ingresos.

5.2 ¿La cantidad de géneros influye en la popularidad y calificación de las películas?

```
ggplot(movies_clean, aes(x = genresAmount, y = popularity)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
```

```
    x = "Cantidad de géneros",
    y = "Popularidad",
    title = "Relación entre cantidad de géneros y popularidad"
  ) +
  theme_minimal()
```



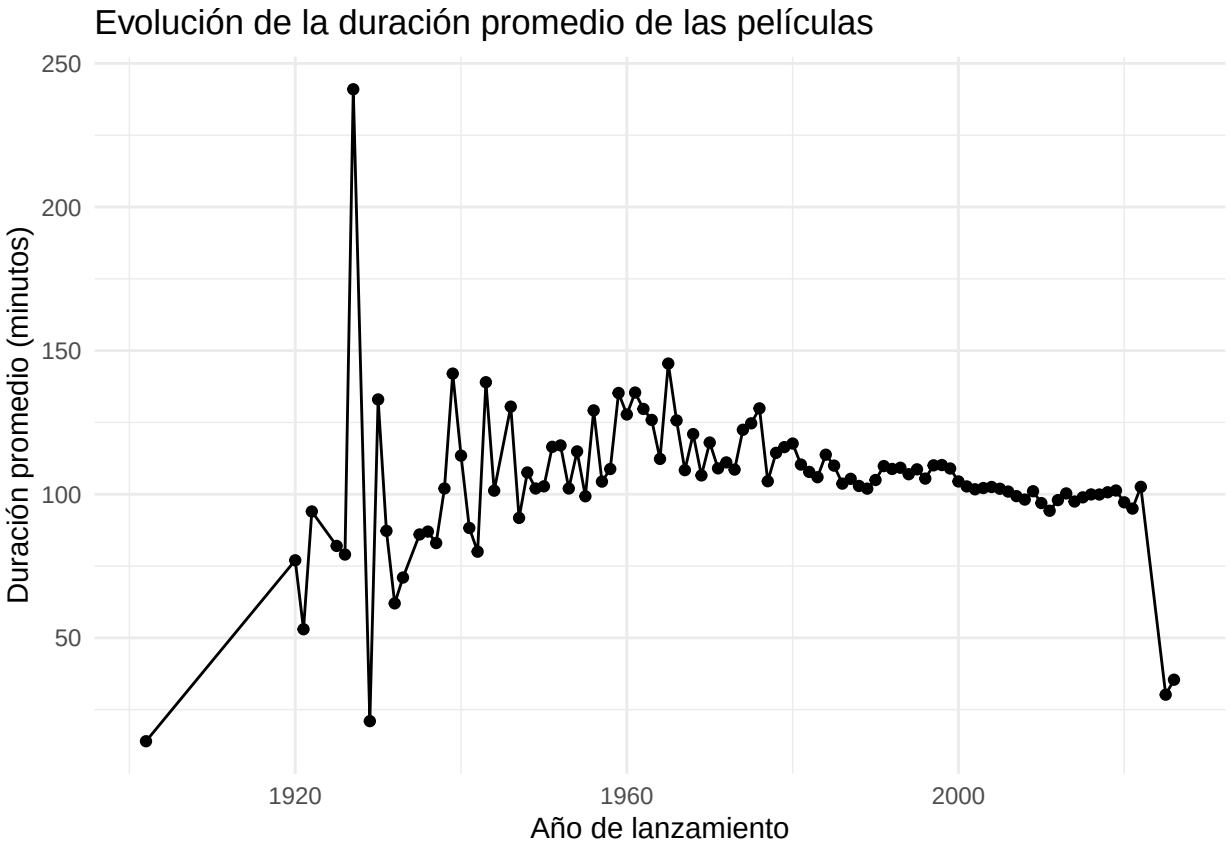
Interpretación

El gráfico de dispersión entre la cantidad de géneros y la popularidad muestra una alta dispersión de los datos y una tendencia prácticamente horizontal. Esto indica que incluir más géneros en una película no se asocia de manera clara con un aumento en su popularidad. La mayoría de películas exitosas se concentra en uno o pocos géneros.

5.3 ¿Cómo ha cambiado la duración promedio de las películas a lo largo del tiempo?

```
runtime_year <- movies_clean %>%
  group_by(releaseYear) %>%
  summarise(runtime_promedio = mean(runtime))

ggplot(runtime_year, aes(x = releaseYear, y = runtime_promedio)) +
  geom_line() +
  geom_point() +
  labs(
    x = "Año de lanzamiento",
    y = "Duración promedio (minutos)",
```

```
    title = "Evolución de la duración promedio de las películas"
  ) +
  theme_minimal()
```

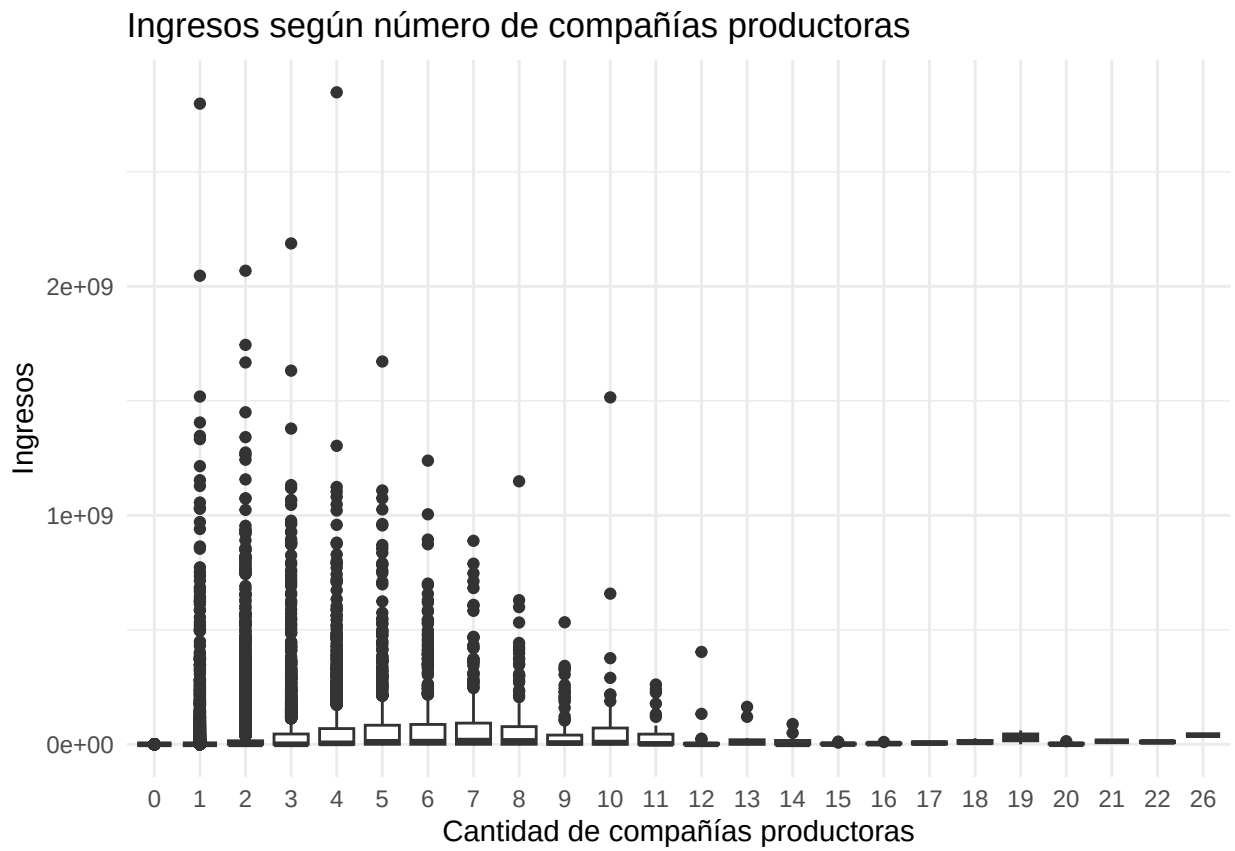


Interpretación

La evolución de la duración promedio de las películas a lo largo del tiempo muestra variaciones importantes en los primeros años, seguidas de una estabilización en décadas más recientes alrededor de una duración cercana a los 100 minutos. Esto sugiere que, con el tiempo, la industria ha convergido hacia duraciones más estándar para los largometrajes.

5.4 ¿Las películas con más compañías productoras generan mayores ingresos?

```
ggplot(movies_clean, aes(x = factor(productionCoAmount), y = revenue)) +
  geom_boxplot() +
  labs(
    x = "Cantidad de compañías productoras",
    y = "Ingresos",
    title = "Ingresos según número de compañías productoras"
  ) +
  theme_minimal()
```



Interpretación

El boxplot indica que las películas con pocas compañías productoras concentran los mayores ingresos y presentan una gran variabilidad. No se observa una relación directa entre un mayor número de compañías productoras y mayores ingresos, lo que sugiere que la colaboración entre muchas compañías no garantiza un mejor desempeño comercial.

5.5 ¿Existen diferencias en la popularidad de las películas según su idioma original?

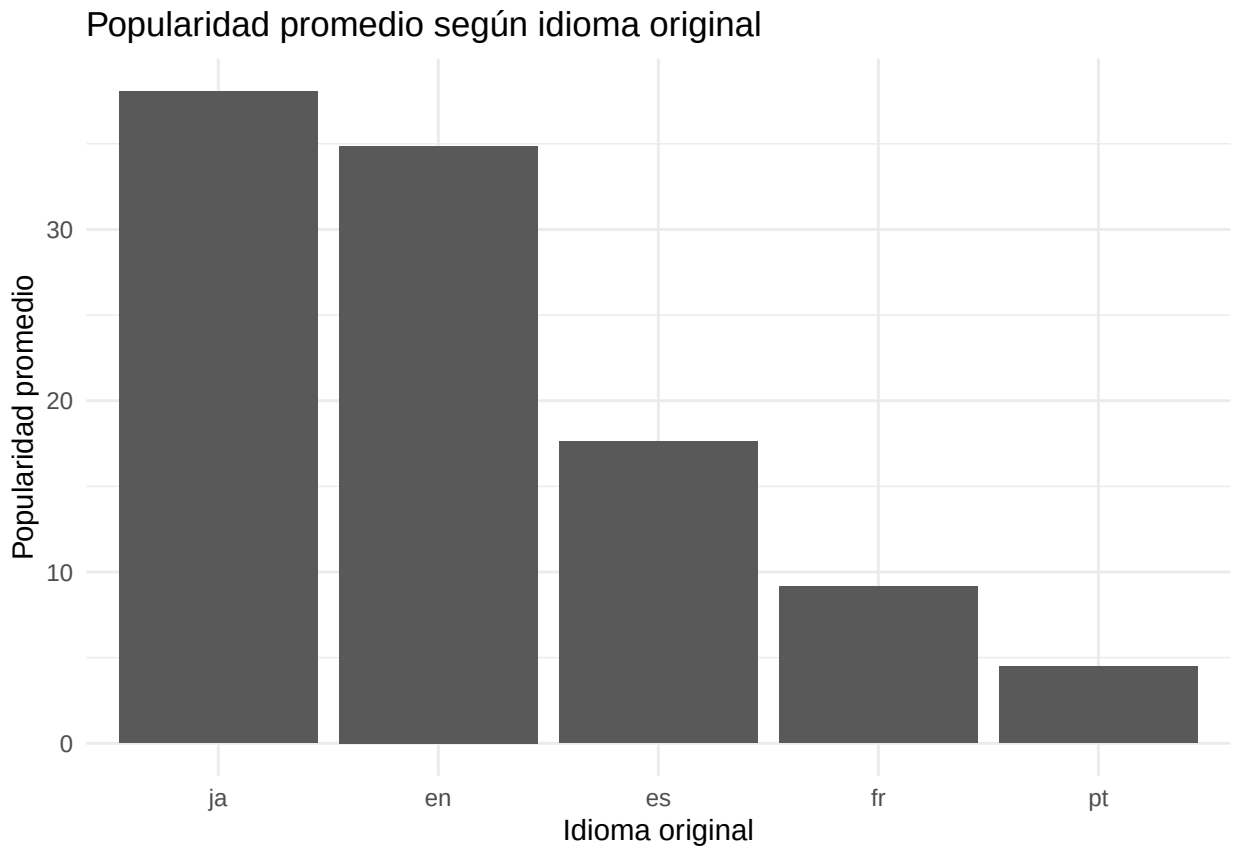
En este análisis se consideran únicamente los idiomas más frecuentes en el conjunto de datos, con el fin de evitar sesgos causados por idiomas con muy pocas observaciones.

```
# Seleccionar los 5 idiomas más frecuentes
top_languages <- movies_clean %>%
  group_by(originalLanguage) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  slice_head(n = 5)
```

```
# Filtrar solo películas en esos idiomas
movies_lang <- movies_clean %>%
  filter(originalLanguage %in% top_languages$originalLanguage)
```

```
# Calcular popularidad promedio por idioma
pop_lang <- movies_lang %>%
  group_by(originalLanguage) %>%
  summarise(popularidad_promedio = mean(popularity))
```

```r
# Gráfico de barras
ggplot(pop_lang, aes(
  x = reorder(originalLanguage, -popularidad_promedio),
  y = popularidad_promedio
)) +
  geom_col() +
  labs(
    x = "Idioma original",
    y = "Popularidad promedio",
    title = "Popularidad promedio según idioma original"
  ) +
  theme_minimal()
```



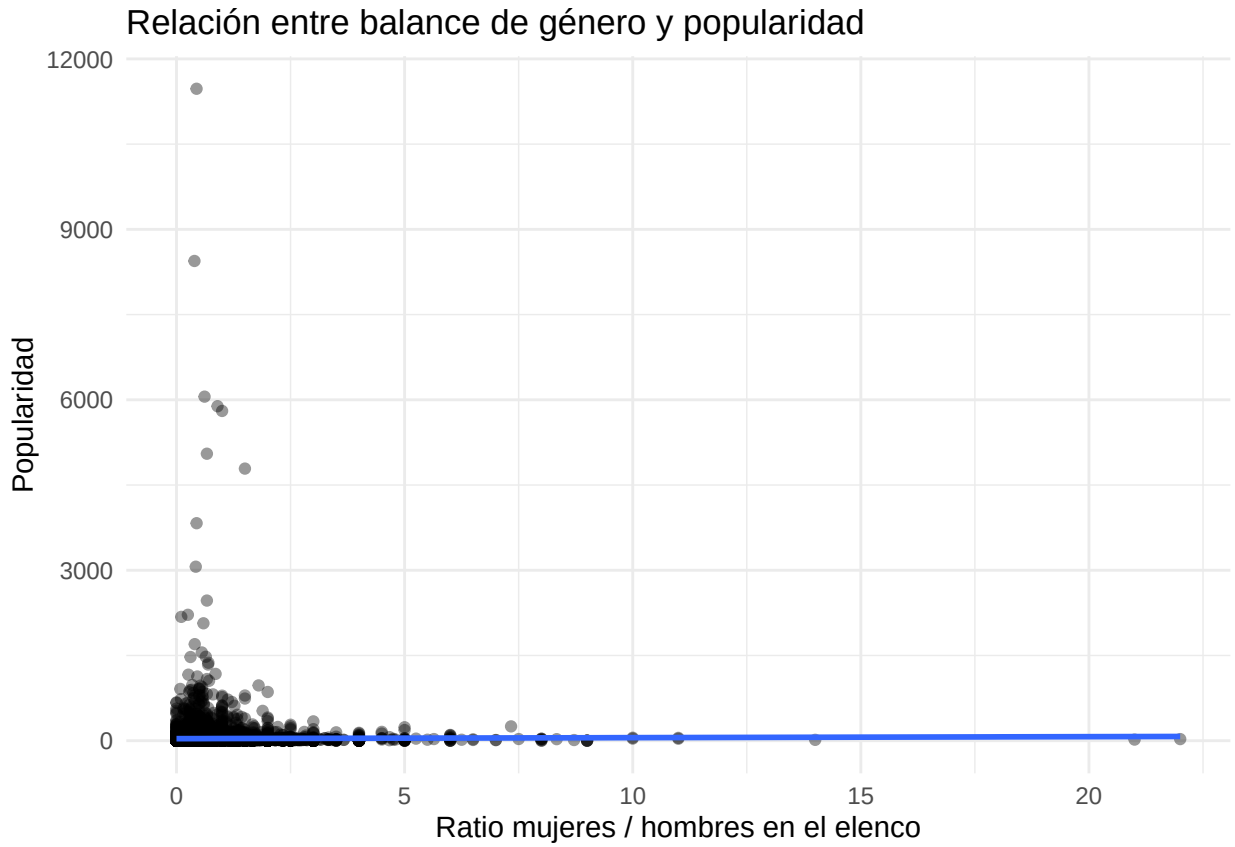
Interpretación

El gráfico de barras muestra diferencias claras en la popularidad promedio según el idioma original de las películas. Algunos idiomas presentan valores promedio más altos de popularidad, lo que sugiere una mayor visibilidad o alcance internacional. Sin embargo, estas diferencias pueden estar influenciadas por el número de películas y el tamaño del mercado asociado a cada idioma.

5.6 ¿El balance de género en el elenco influye en la popularidad de las películas?

```r
movies_gender <- movies_clean %>%
  mutate(ratio_genero = castWomenAmount / castMenAmount) %>%
  filter(is.finite(ratio_genero))
```

```r
ggplot(movies_gender, aes(x = ratio_genero, y = popularity)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "Ratio mujeres / hombres en el elenco",
    y = "Popularidad",
    title = "Relación entre balance de género y popularidad"
  ) +
  theme_minimal()
```



Interpretación

El gráfico que relaciona el ratio mujeres/hombres en el elenco con la popularidad muestra una alta concentración de películas con ratios bajos y una línea de tendencia prácticamente horizontal. Esto indica que el balance de género en el reparto no tiene una relación clara con la popularidad de las películas, reforzando que este factor no es determinante en la recepción del público.

LINK REPOSITORIO GITHUB git@github.com:Cisco890/Lab1MineriaDatos.git