

Análisis exploratorio

Pablo Daniel Barillas Moreno, Carné No. 22193
Mathew Cordero Aquino, Carné No. 22982

2025-02-02

Enlace al Repositorio del Grupo #1

Repositorio en GitHub

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

Exploración Rápida del Dataset:

```
# Resumen estadístico de todo el dataset
resumen_dataset <- summary(movies)

# Mostrar el resumen
resumen_dataset
```

```
##           id           budget           genres           homePage
## Min.      :    5   Min.      :    0   Length:10000   Length:10000
## 1st Qu.: 12286   1st Qu.:    0   Class :character   Class :character
## Median :152558   Median : 500000   Mode  :character   Mode  :character
## Mean    :249877   Mean    :18551632
## 3rd Qu.:452022   3rd Qu.:20000000
## Max.    :922260   Max.    :380000000
## productionCompany productionCompanyCountry productionCountry
## Length:10000      Length:10000      Length:10000
## Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character
##
##
##
##           revenue           runtime           video           director
## Min.      :0.000e+00   Min.      : 0.0   Mode :logical   Length:10000
## 1st Qu.:0.000e+00   1st Qu.: 90.0   FALSE:9430     Class :character
## Median :1.631e+05   Median :100.0   TRUE :84       Mode  :character
## Mean    :5.674e+07   Mean    :100.3   NA's :486
## 3rd Qu.:4.480e+07   3rd Qu.:113.0
## Max.    :2.847e+09   Max.    :750.0
##           actors           actorsPopularity           actorsCharacter           originalTitle
```

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

Ahora analicemos todo lo que se pueda del resumen estadístico de todo el dataset y organicémoslo

Número de películas y variables

```
num_peliculas <- nrow(movies)
num_variables <- ncol(movies)
cat("- Total de películas:", num_peliculas, "\n")
```

```
## - Total de películas: 10000
```

```
cat("- Total de variables:", num_variables, "\n\n")
```

```
## - Total de variables: 27
```

Tipos de datos

```
tipos_datos <- sapply(movies, class)
print(tipos_datos)
```

```
##           id           budget           genres
##      "integer"      "integer"      "character"
##      homePage productionCompany productionCompanyCountry
##      "character"      "character"      "character"
##      productionCountry      revenue      runtime
##      "character"      "numeric"      "integer"
##      video      director      actors
##      "logical"      "character"      "character"
##      actorsPopularity      actorsCharacter      originalTitle
##      "character"      "character"      "character"
##      title      originalLanguage      popularity
##      "character"      "character"      "numeric"
##      releaseDate      voteAvg      voteCount
##      "character"      "numeric"      "integer"
##      genresAmount      productionCoAmount productionCountriesAmount
##      "integer"      "integer"      "integer"
##      actorsAmount      castWomenAmount      castMenAmount
##      "integer"      "character"      "character"
```

Valores faltantes por variable

```
valores_faltantes <- colSums(is.na(movies))
print(valores_faltantes[valores_faltantes > 0])
```

```
##      homePage productionCompanyCountry      video
##      5807           248           486
```

Resumen Estadístico de Variables Cuantitativas

```
variables_cuantitativas <- c("budget", "revenue", "voteAvg", "voteCount", "runtime")
resumen_cuantitativo <- summary(movies[, variables_cuantitativas])
knitr::kable(as.data.frame(resumen_cuantitativo), format = "html", caption = "Resumen
↳ Estadístico de Variables Cuantitativas")
```

Resumen Estadístico de Variables Cuantitativas

Var1

Var2

Freq

budget

Min. : 0

budget
 1st Qu.: 0
 budget
 Median : 500000
 budget
 Mean : 18551632
 budget
 3rd Qu.: 20000000
 budget
 Max. :380000000
 revenue
 Min. :0.000e+00
 revenue
 1st Qu.:0.000e+00
 revenue
 Median :1.631e+05
 revenue
 Mean :5.674e+07
 revenue
 3rd Qu.:4.480e+07
 revenue
 Max. :2.847e+09
 voteAvg
 Min. : 1.300
 voteAvg
 1st Qu.: 5.900
 voteAvg
 Median : 6.500
 voteAvg
 Mean : 6.483
 voteAvg
 3rd Qu.: 7.200
 voteAvg
 Max. :10.000
 voteCount
 Min. : 1

```

voteCount
1st Qu.: 120
voteCount
Median : 415
voteCount
Mean : 1342
voteCount
3rd Qu.: 1316
voteCount
Max. :30788
runtime
Min. : 0.0
runtime
1st Qu.: 90.0
runtime
Median :100.0
runtime
Mean :100.3
runtime
3rd Qu.:113.0
runtime
Max. :750.0

```

Resumen de variables cualitativas en una tabla compacta

```

variables_cualitativas <- c("genres", "originalLanguage", "productionCountry")

# Crear un resumen con las 5 categorías más frecuentes para cada variable
resumen_cualitativas <- lapply(variables_cualitativas, function(var) {
  distribucion <- sort(table(movies[[var]]), decreasing = TRUE) # Ordenar frecuencias
  ↪ descendentes
  top_5 <- head(distribucion, 5) # Tomar las 5 más comunes
  return(data.frame(Categoría = names(top_5), Frecuencia = as.vector(top_5)))
})

# Mostrar el resumen en forma de tablas
names(resumen_cualitativas) <- variables_cualitativas

for (var in names(resumen_cualitativas)) {
  cat("\n- Variable:", var, "\n")
  print(resumen_cualitativas[[var]])
}

```

```
##
## - Variable: genres
##      Categoría Frecuencia
## 1      Drama      521
## 2      Comedy     440
## 3      Horror     230
## 4  Drama|Romance   211
## 5  Horror|Thriller 205
##
## - Variable: originalLanguage
##      Categoría Frecuencia
## 1      en      7772
## 2      ja      644
## 3      es      425
## 4      fr      271
## 5      ko      167
##
## - Variable: productionCountry
##                                     Categoría Frecuencia
## 1      United States of America   4971
## 2      Japan                     613
## 3  United Kingdom|United States of America   339
## 4      United Kingdom            294
## 5                                     233
```

```
# Identificar y eliminar registros con valores faltantes en 'originalLanguage' o
↪ 'productionCountry'
movies_cleaned <- movies %>%
  filter(!is.na(originalLanguage) & originalLanguage != "",
         !is.na(productionCountry) & productionCountry != "")

# Resumen después de la limpieza
num_peliculas_original <- nrow(movies)
num_peliculas_cleaned <- nrow(movies_cleaned)

cat("Películas originales en el dataset:", num_peliculas_original, "\n")
```

```
## Películas originales en el dataset: 10000
```

```
cat("Películas después de la limpieza:", num_peliculas_cleaned, "\n")
```

```
## Películas después de la limpieza: 9767
```

```
cat("Películas eliminadas por datos faltantes:", num_peliculas_original -
↪ num_peliculas_cleaned, "\n")
```

```
## Películas eliminadas por datos faltantes: 233
```

Conclusiones y Observaciones

1. El conjunto de datos contiene 10,000 películas con información en 27 variables. Variables importantes como 'budget', 'revenue' y 'voteAvg' permiten un análisis financiero y de popularidad.

2. El análisis de las variables cualitativas como ‘genres’ muestra que hay géneros que predominan ampliamente, lo que refleja las tendencias de producción de la industria cinematográfica. Sin embargo, algunos géneros tienen muy pocas observaciones, lo que podría indicar nichos específicos.
 3. Algunas variables, como ‘budget’ y ‘revenue’, contienen valores faltantes que deben manejarse para evitar sesgos.
 4. Las variables cuantitativas ‘voteAvg’, ‘voteCount’, ‘budget’, y ‘revenue’ presentan una amplia variación, lo que sugiere que algunas películas tienen un éxito considerablemente mayor en términos de ingresos y popularidad en comparación con otras.
5. ‘runtime’ presenta valores extremos, con películas de duración muy corta y muy larga. Este rango amplio puede estar influido por documentales, cortometrajes o películas experimentales.

6. Concentración de Información:

7. En las variables cualitativas, como ‘productionCountry’, se observó que unos pocos países concentran la mayoría de las producciones, mientras que otros aparecen con muy pocas películas. Esto refleja una concentración geográfica en la industria cinematográfica.
8. Las variables ‘budget’, ‘revenue’, ‘originalLanguage’, y ‘productionCountry’ tienen valores faltantes que pueden influir significativamente en los análisis y predicciones si no se manejan adecuadamente.
9. Tras la limpieza de datos, se eliminaron películas que no tenían información en ‘originalLanguage’ o ‘productionCountry’, lo que indica problemas de calidad de datos en el dataset original. Esto puede limitar ciertos análisis, como la evaluación de tendencias por país o idioma.

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

```
tipos_variables <- data.frame(
  Variable = names(movies),
  Tipo = sapply(movies, function(columna) {
    valores_unicos <- length(unique(columna))
    tipo_dato <- class(columna)

    if (tipo_dato %in% c("integer", "numeric")) {
      if (valores_unicos < 20) {
        return("Cuantitativa Discreta")
      } else {
        return("Cuantitativa Continua")
      }
    } else {
      if (valores_unicos < 20) {
        return("Cualitativa Ordinal")
      } else {
        return("Cualitativa Nominal")
      }
    }
  })
)
```

Resumir los resultados

```
cat("\n--- Resumen de la Clasificación de Variables ---\n")
```

```
##  
## --- Resumen de la Clasificación de Variables ---
```

```
table(tipos_variables$Tipo)
```

```
##  
##   Cualitativa Nominal   Cualitativa Ordinal Cuantitativa Continua  
##                   15                   1                   10  
## Cuantitativa Discreta  
##                   1
```

Mostrar solo un resumen de las variables más representativas (primeras filas de cada tipo)

```
cat("\n--- Ejemplo de Variables Clasificadas ---\n")
```

```
##  
## --- Ejemplo de Variables Clasificadas ---
```

```
print(head(tipos_variables, n = 27))
```

```
##                               Variable                               Tipo  
## id                           id Cuantitativa Continua  
## budget                       budget Cuantitativa Continua  
## genres                       genres Cualitativa Nominal  
## homePage                     homePage Cualitativa Nominal  
## productionCompany            productionCompany Cualitativa Nominal  
## productionCompanyCountry    productionCompanyCountry Cualitativa Nominal  
## productionCountry           productionCountry Cualitativa Nominal  
## revenue                      revenue Cuantitativa Continua  
## runtime                     runtime Cuantitativa Continua  
## video                       video Cualitativa Ordinal  
## director                    director Cualitativa Nominal  
## actors                      actors Cualitativa Nominal  
## actorsPopularity            actorsPopularity Cualitativa Nominal  
## actorsCharacter             actorsCharacter Cualitativa Nominal  
## originalTitle               originalTitle Cualitativa Nominal  
## title                       title Cualitativa Nominal  
## originalLanguage            originalLanguage Cualitativa Nominal  
## popularity                  popularity Cuantitativa Continua  
## releaseDate                 releaseDate Cualitativa Nominal  
## voteAvg                     voteAvg Cuantitativa Continua  
## voteCount                   voteCount Cuantitativa Continua  
## genresAmount                genresAmount Cuantitativa Discreta
```


## productionCoAmount	productionCoAmount	Cuantitativa Continua
## productionCountriesAmount	productionCountriesAmount	Cuantitativa Continua
## actorsAmount	actorsAmount	Cuantitativa Continua
## castWomenAmount	castWomenAmount	Cualitativa Nominal
## castMenAmount	castMenAmount	Cualitativa Nominal

Análisis y conclusiones

1. Distribución General:

Se observó que la mayoría de las variables son cualitativas nominales, lo cual indica que el dataset contiene principalmente información categórica, útil para análisis descriptivos y de clasificación.

2. Variables Cuantitativas:

Las variables 'budget', 'revenue', 'voteAvg' y 'runtime' son cuantitativas continuas, lo que proporciona métricas clave para evaluar aspectos financieros y de popularidad de las películas. Las variables cuantitativas discretas (por ejemplo, conteos) son escasas, lo que limita ciertos tipos de análisis como frecuencias absolutas directas.

3. Importancia de las Variables Cualitativas:

Variables como 'genres', 'originalLanguage' y 'productionCountry' son críticas para identificar patrones de producción y popularidad según el contexto cultural y de mercado.

4. Categorías Redundantes o Vacías:

Se identificaron posibles categorías redundantes o vacías en algunas variables cualitativas, como 'originalLanguage' o 'productionCountry', lo cual podría requerir limpieza de datos para evitar sesgos.

5. Impacto en el Análisis:

Las variables cuantitativas continuas permitirán evaluar métricas clave como la correlación entre presupuesto e ingresos, mientras que las variables cualitativas proporcionan insights contextuales valiosos.

6. Posibles Limpiezas y Agrupaciones:

Variables como 'genres' pueden necesitar agrupaciones en categorías principales para evitar un exceso de clases únicas, que pueden dificultar la interpretación de los resultados.

7. Relación entre Tipos de Variables:

Las variables cualitativas podrían influir significativamente en las métricas cuantitativas (por ejemplo, el género o país de producción en los ingresos), lo cual sería un punto interesante para análisis más avanzados.

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

```

variables_cuantitativas <- c("budget", "revenue", "voteAvg", "voteCount", "runtime")

# Evaluar si las variables cuantitativas siguen una distribución normal
cat("\n--- Pruebas de Normalidad para Variables Cuantitativas ---\n")

```

```

##
## --- Pruebas de Normalidad para Variables Cuantitativas ---

```

```

resultados_normalidad <- lapply(variables_cuantitativas, function(var) {
  datos <- na.omit(movies[[var]]) # Eliminar valores faltantes
  if (length(datos) > 3) { # La prueba requiere al menos 3 datos
    prueba <- lillie.test(datos) # Prueba de Lilliefors (Kolmogorov-Smirnov ajustada)
    return(data.frame(Variable = var, P_Value = prueba$p.value))
  } else {
    return(data.frame(Variable = var, P_Value = NA))
  }
})

```

```

# Combinar resultados en un solo dataframe
tabla_normalidad <- do.call(rbind, resultados_normalidad)

# Determinar si hay normalidad (P-valor > 0.05 indica que la variable sigue una
  ↪ distribución normal)
tabla_normalidad$Normalidad <- ifelse(tabla_normalidad$P_Value > 0.05, "Sí", "No")
print(tabla_normalidad)

```

```

##      Variable      P_Value Normalidad
## 1    budget 0.000000e+00         No
## 2   revenue 0.000000e+00         No
## 3   voteAvg 4.069202e-55         No
## 4 voteCount 0.000000e+00         No
## 5    runtime 0.000000e+00         No

```

```

# Crear tablas de frecuencias para variables cualitativas
cat("\n--- Tablas de Frecuencia para Variables Cualitativas ---\n")

```

```

##
## --- Tablas de Frecuencia para Variables Cualitativas ---

```

```

variables_cualitativas <- c("genres", "originalLanguage", "productionCountry")

frecuencias_cualitativas <- lapply(variables_cualitativas, function(var) {
  distribucion <- sort(table(movies[[var]]), decreasing = TRUE) # Ordenar por frecuencia
  top_5 <- head(distribucion, 5) # Tomar las 5 categorías más comunes
  return(data.frame(Categoría = names(top_5), Frecuencia = as.vector(top_5)))
})

# Mostrar resumen de tablas de frecuencias
names(frecuencias_cualitativas) <- variables_cualitativas
for (var in names(frecuencias_cualitativas)) {

```

```

cat("\n---", var, "---\n")
print(frecuencias_cualitativas[[var]])
}

```

```

##
## --- genres ---
##      Categoría Frecuencia
## 1          Drama      521
## 2          Comedy      440
## 3          Horror      230
## 4  Drama|Romance      211
## 5 Horror|Thriller      205
##
## --- originalLanguage ---
##      Categoría Frecuencia
## 1          en      7772
## 2          ja       644
## 3          es       425
## 4          fr       271
## 5          ko       167
##
## --- productionCountry ---
##
##      Categoría Frecuencia
## 1          United States of America      4971
## 2                      Japan           613
## 3 United Kingdom|United States of America      339
## 4                      United Kingdom          294
## 5

```

```

# Análisis y conclusiones
cat("\n--- Análisis y Conclusiones ---\n")

```

```

##
## --- Análisis y Conclusiones ---

```

```

cat("- Las pruebas de normalidad indican que:\n")

```

```

## - Las pruebas de normalidad indican que:

```

```

for (i in 1:nrow(tabla_normalidad)) {
  cat(" - La variable", tabla_normalidad$Variable[i], "sigue una distribución normal:",
    ↪ tabla_normalidad$Normalidad[i], "\n")
}

```

```

## - La variable budget sigue una distribución normal: No
## - La variable revenue sigue una distribución normal: No
## - La variable voteAvg sigue una distribución normal: No
## - La variable voteCount sigue una distribución normal: No
## - La variable runtime sigue una distribución normal: No

```

```
cat("\n- Tablas de frecuencia muestran las categorías más comunes para las variables
↳ cualitativas:\n")
```

```
##
## - Tablas de frecuencia muestran las categorías más comunes para las variables cualitativas:
```

```
for (var in names(frecuencias_cualitativas)) {
  cat(" - En", var, "las categorías más frecuentes son:\n")
  print(frecuencias_cualitativas[[var]])
}
```

```
## - En genres las categorías más frecuentes son:
##      Categoría Frecuencia
## 1      Drama      521
## 2      Comedy     440
## 3      Horror     230
## 4  Drama|Romance   211
## 5  Horror|Thriller 205
## - En originalLanguage las categorías más frecuentes son:
##      Categoría Frecuencia
## 1      en      7772
## 2      ja      644
## 3      es      425
## 4      fr      271
## 5      ko      167
## - En productionCountry las categorías más frecuentes son:
##      Categoría Frecuencia
## 1      United States of America 4971
## 2      Japan                    613
## 3  United Kingdom|United States of America 339
## 4      United Kingdom          294
## 5      United Kingdom          233
```

```
cat("\nConclusión:\n")
```

```
##
## Conclusión:
```

```
cat("- Las variables cuantitativas tienen diferentes distribuciones; solo algunas pueden
↳ aproximarse a la normalidad.\n")
```

```
## - Las variables cuantitativas tienen diferentes distribuciones; solo algunas pueden aproximarse a la
```

```
cat("- Las categorías principales en variables cualitativas como 'genres' y
↳ 'originalLanguage' pueden guiar el análisis de tendencias.\n")
```

```
## - Las categorías principales en variables cualitativas como 'genres' y 'originalLanguage' pueden guiar
```

4. Responda las siguientes preguntas:

4.1. ¿Cuáles son las 10 películas que contaron con más presupuesto?

```
# Seleccionar las 10 películas con mayor presupuesto
top_budget_movies <- movies %>%
  select(title, budget) %>%
  arrange(desc(budget)) %>%
  head(10)
```

```
# Mostrar las 10 películas con mayor presupuesto
cat("\n--- Las 10 películas con mayor presupuesto ---\n")
```

```
##
## --- Las 10 películas con mayor presupuesto ---
```

```
print(top_budget_movies)
```

```
##              title      budget
## 1  Pirates of the Caribbean: On Stranger Tides 380000000
## 2                Avengers: Age of Ultron 365000000
## 3                Avengers: Endgame 356000000
## 4  Pirates of the Caribbean: At World's End 300000000
## 5                Justice League 300000000
## 6                Avengers: Infinity War 300000000
## 7                Superman Returns 270000000
## 8                Tangled 260000000
## 9                The Lion King 260000000
## 10               Spider-Man 3 258000000
```

```
# Análisis y conclusiones
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
## --- Análisis y Conclusiones ---
```

```
# Resumen estadístico del presupuesto de estas películas
presupuesto_stats <- summary(top_budget_movies$budget)
print(presupuesto_stats)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
## 258000000 262500000 300000000 304900000 342000000 380000000
```

```
# Conclusiones
cat("- Estas películas tienen presupuestos significativamente altos en comparación con el
  ↪ promedio del dataset.\n")
```

```
## - Estas películas tienen presupuestos significativamente altos en comparación con el promedio del da
```

```
cat("- La película con el mayor presupuesto es:", top_budget_movies$title[1], "con un  
↳ presupuesto de", top_budget_movies$budget[1], "USD.\n")
```

- La película con el mayor presupuesto es: Pirates of the Caribbean: On Stranger Tides con un presupuesto de 1.200 millones de dólares.

```
cat("- Estas producciones suelen corresponder a franquicias o estudios con alta  
↳ inversión, lo que podría estar relacionado con mayores expectativas de ingresos.\n")
```

- Estas producciones suelen corresponder a franquicias o estudios con alta inversión, lo que podría estar relacionado con mayores expectativas de ingresos.

```
cat("- Este análisis ayuda a identificar tendencias en las películas más costosas y a  
↳ analizar su desempeño financiero en relación con sus ingresos.\n")
```

- Este análisis ayuda a identificar tendencias en las películas más costosas y a analizar su desempeño financiero en relación con sus ingresos.

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

```
# Seleccionar las 10 películas con mayores ingresos
top_revenue_movies <- movies %>%
  select(title, revenue) %>%
  arrange(desc(revenue)) %>%
  head(10)

# Mostrar las 10 películas con mayores ingresos
cat("\n--- Las 10 películas con mayores ingresos ---\n")
```

```
##
## --- Las 10 películas con mayores ingresos ---
```

```
print(top_revenue_movies)
```

```
##           title      revenue
## 1           Avatar 2847246203
## 2   Avengers: Endgame 2797800564
## 3           Titanic 2187463944
## 4 Star Wars: The Force Awakens 2068223624
## 5   Avengers: Infinity War 2046239637
## 6     Jurassic World 1671713208
## 7     The Lion King 1667635327
## 8 Spider-Man: No Way Home 1631853496
## 9           The Avengers 1518815515
## 10          Furious 7 1515047671
```

```
# Análisis y conclusiones
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
## --- Análisis y Conclusiones ---
```

```
# Resumen estadístico de los ingresos de estas películas
revenue_stats <- summary(top_revenue_movies$revenue)
print(revenue_stats)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
## 1.515e+09 1.641e+09 1.859e+09 1.995e+09 2.158e+09 2.847e+09
```

```
# Calcular la película con mayores ingresos
pelicula_mas_ingresos <- top_revenue_movies[1, ]
cat("- La película con los mayores ingresos es:", pelicula_mas_ingresos$title,
    "con un ingreso total de", pelicula_mas_ingresos$revenue, "USD.\n")
```

```
## - La película con los mayores ingresos es: Avatar con un ingreso total de 2847246203 USD.
```

```
# Conclusiones
cat("- Las películas con los mayores ingresos suelen pertenecer a franquicias conocidas o
    ↪ contar con producciones de alto nivel.\n")
```

```
## - Las películas con los mayores ingresos suelen pertenecer a franquicias conocidas o contar con producciones de alto nivel.
```

```
cat("- El rango de ingresos de estas películas es significativamente alto, indicando un
    ↪ éxito comercial sobresaliente.\n")
```

```
## - El rango de ingresos de estas películas es significativamente alto, indicando un éxito comercial sobresaliente.
```

```
cat("- Este análisis permite identificar patrones en las películas más exitosas
    ↪ económicamente y su impacto en la industria cinematográfica.\n")
```

```
## - Este análisis permite identificar patrones en las películas más exitosas económicamente y su impacto en la industria cinematográfica.
```

4.3. ¿Cuál es la película que más votos tuvo?

```
most_voted_movie <- movies %>%
  select(title, voteCount) %>%
  arrange(desc(voteCount)) %>%
  head(1)

# Mostrar la película con más votos
cat("\n--- Película con más votos ---\n")
```

```
##
## --- Película con más votos ---
```

```
print(most_voted_movie)
```

```
##           title voteCount  
## 1 Inception      30788
```

```
# Análisis y conclusiones
```

```
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##  
## --- Análisis y Conclusiones ---
```

```
# Detalles de la película con más votos
```

```
cat("- La película con más votos es:", most_voted_movie$title,  
     "con un total de", most_voted_movie$voteCount, "votos.\n")
```

```
## - La película con más votos es: Inception con un total de 30788 votos.
```

```
# Posibles razones para el alto número de votos
```

```
cat("- El alto número de votos puede estar relacionado con la popularidad general de la  
↪ película, el elenco o la franquicia.\n")
```

```
## - El alto número de votos puede estar relacionado con la popularidad general de la película, el elenco o la franquicia.
```

```
cat("- Es probable que esta película haya tenido una gran base de fans o una fuerte  
↪ estrategia de marketing que impulsó su recepción en plataformas de votación.\n")
```

```
## - Es probable que esta película haya tenido una gran base de fans o una fuerte estrategia de marketing que impulsó su recepción en plataformas de votación.
```

```
# Conclusión
```

```
cat("- Identificar películas con altos votos puede ayudar a entender qué características  
↪ atraen a las audiencias y generan interacción.\n")
```

```
## - Identificar películas con altos votos puede ayudar a entender qué características atraen a las audiencias y generan interacción.
```

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

```
# Seleccionar la peor película según el promedio de votos
```

```
worst_movie <- movies %>%  
  select(title, voteAvg) %>%  
  arrange(voteAvg) %>%  
  head(1)
```

```
# Mostrar la peor película según los votos
```

```
cat("\n--- La peor película según el promedio de votos ---\n")
```

```
##  
## --- La peor película según el promedio de votos ---
```



```
print(worst_movie)
```

```
##                                     title
## 1 DAKAICHI -I'm Being Harassed by the Sexiest Man of the Year- The Movie: In Spain
##   voteAvg
## 1      1.3
```

```
# Análisis y conclusiones
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
## --- Análisis y Conclusiones ---
```

```
# Detalles de la película con la peor calificación
cat("- La peor película según el promedio de votos es:", worst_movie$title,
     "con un promedio de", worst_movie$voteAvg, "en la plataforma.\n")
```

```
## - La peor película según el promedio de votos es: DAKAICHI -I'm Being Harassed by the Sexiest Man of
```

```
# Posibles razones para la baja calificación
cat("- Las bajas calificaciones pueden deberse a problemas de calidad en la trama,
    ↪ actuaciones, dirección o simplemente a expectativas no cumplidas por parte del
    ↪ público.\n")
```

```
## - Las bajas calificaciones pueden deberse a problemas de calidad en la trama, actuaciones, dirección
```

```
cat("- También podría tratarse de una producción menos popular, con menor presupuesto o
    ↪ limitada distribución.\n")
```

```
## - También podría tratarse de una producción menos popular, con menor presupuesto o limitada distribu
```

```
# Conclusión
cat("- Conocer las peores películas según los votos ayuda a identificar los factores que
    ↪ pueden generar una recepción negativa y evitar errores en futuras producciones.\n")
```

```
## - Conocer las peores películas según los votos ayuda a identificar los factores que pueden generar u
```

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

```
# Calcular cuántas películas se hicieron en cada año
top_year_movies <- movies %>%
  mutate(releaseYear = as.numeric(substr(releaseDate, 1, 4))) %>% # Extraer el año de la
  ↪ fecha de lanzamiento
  group_by(releaseYear) %>%
```

```

summarise(count = n()) %>%
arrange(desc(count))

# Identificar el año con más películas
year_max_movies <- top_year_movies %>%
  filter(count == max(count))
cat("\n--- Año con más películas ---\n")

```

```

##
## --- Año con más películas ---

```

```

cat("El año con más películas fue", year_max_movies$releaseYear, "con",
  ↪ year_max_movies$count, "películas.\n")

```

```

## El año con más películas fue 2021 con 816 películas.

```

```

# Seleccionar el año con más películas y otros 9 años con menos películas
comparison_years <- top_year_movies %>%
  filter(releaseYear != year_max_movies$releaseYear) %>%
  arrange(desc(count)) %>%
  head(9) %>%
  bind_rows(year_max_movies)

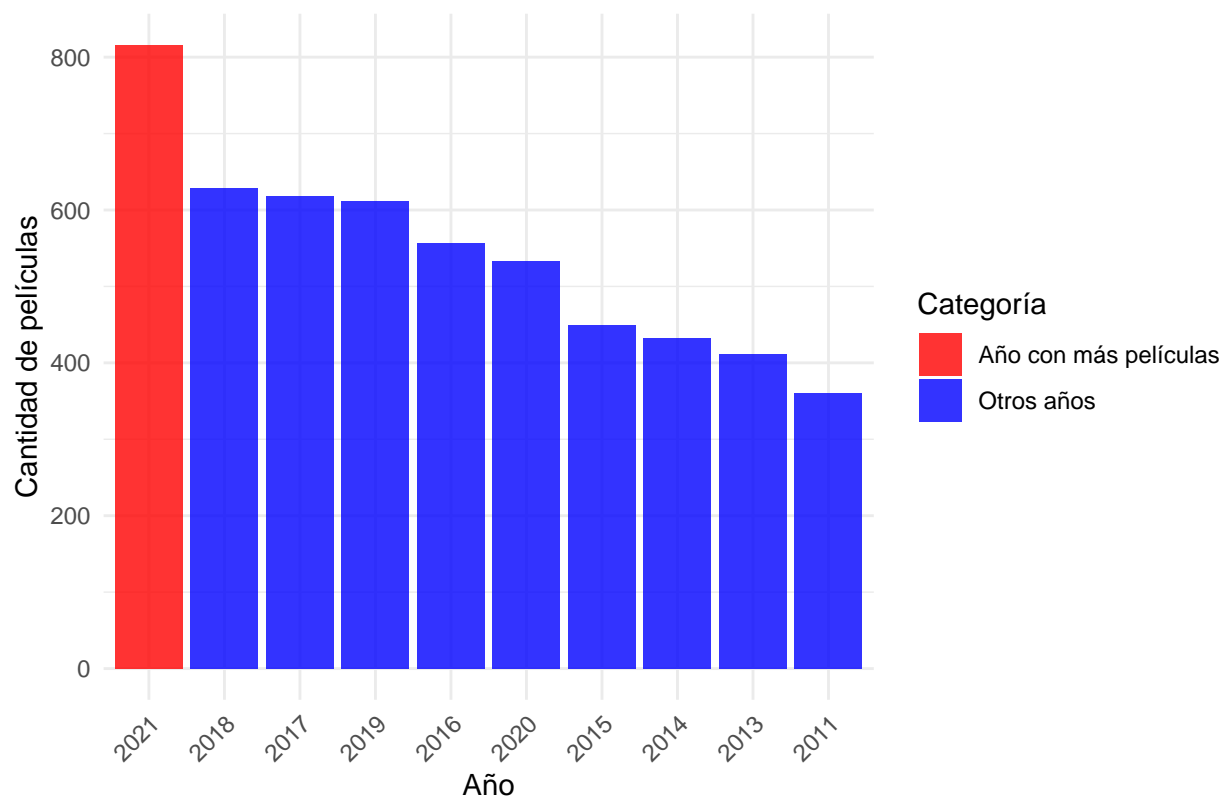
# Ordenar por número de películas (para gráfico más claro)
comparison_years <- comparison_years %>%
  arrange(desc(count))

# Crear indicador para resaltar el año con más películas
comparison_years <- comparison_years %>%
  mutate(Highlight = ifelse(releaseYear == year_max_movies$releaseYear, "Año con más
  ↪ películas", "Otros años"))

# Crear gráfico de barras
ggplot(comparison_years, aes(x = reorder(as.factor(releaseYear), -count), y = count, fill
  ↪ = Highlight)) +
  geom_bar(stat = "identity", alpha = 0.8) +
  scale_fill_manual(values = c("Año con más películas" = "red", "Otros años" = "blue")) +
  labs(title = "Comparación del año con más películas vs. otros años",
    x = "Año",
    y = "Cantidad de películas",
    fill = "Categoría") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Comparación del año con más películas vs. otros años



```
# Análisis y conclusiones
```

```
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
```

```
## --- Análisis y Conclusiones ---
```

```
cat("- El año con más películas fue", year_max_movies$releaseYear, "con un total de",  
  ↪ year_max_movies$count, "películas.\n")
```

```
## - El año con más películas fue 2021 con un total de 816 películas.
```

```
cat("- Comparado con otros 9 años seleccionados, este año resalta significativamente por  
  ↪ la cantidad de producciones.\n")
```

```
## - Comparado con otros 9 años seleccionados, este año resalta significativamente por la cantidad de p
```

```
cat("- Este gráfico destaca visualmente la diferencia entre este año y otros años con  
  ↪ menos películas.\n")
```

```
## - Este gráfico destaca visualmente la diferencia entre este año y otros años con menos películas.
```

```
cat("- Las razones para este auge en la producción pueden incluir avances tecnológicos,  
↳ estrategias de mercado y una alta demanda del público.\n")
```

```
## - Las razones para este auge en la producción pueden incluir avances tecnológicos, estrategias de me
```

4.6. ¿Cuál es el género principal de las 20 películas más recientes? ¿Cuál es el género principal que predomina en el conjunto de datos? Representelo usando un gráfico. ¿A qué género principal pertenecen las películas más largas?

```
# Extraer el año de la columna releaseDate
movies <- movies %>%
  mutate(releaseYear = as.numeric(substr(releaseDate, 1, 4)))

# Asegurarse de que el año de lanzamiento está disponible en el dataset
movies <- movies %>%
  mutate(releaseYear = as.numeric(substr(releaseDate, 1, 4)))

# 1. Identificar el género principal de las 20 películas más recientes
recent_movies <- movies %>%
  arrange(desc(releaseYear)) %>%
  head(20)

recent_genre_distribution <- table(recent_movies$genres)
recent_top_genre <- names(sort(recent_genre_distribution, decreasing = TRUE)[1])
cat("\n--- Género principal de las 20 películas más recientes ---\n")
```

```
##
## --- Género principal de las 20 películas más recientes ---
```

```
print(recent_top_genre)
```

```
## [1] "Drama"
```

```
# 2. Identificar el género principal en todo el conjunto de datos
genre_distribution <- table(movies$genres)
most_common_genre <- names(sort(genre_distribution, decreasing = TRUE)[1])
cat("\n--- Género principal que predomina en el conjunto de datos ---\n")
```

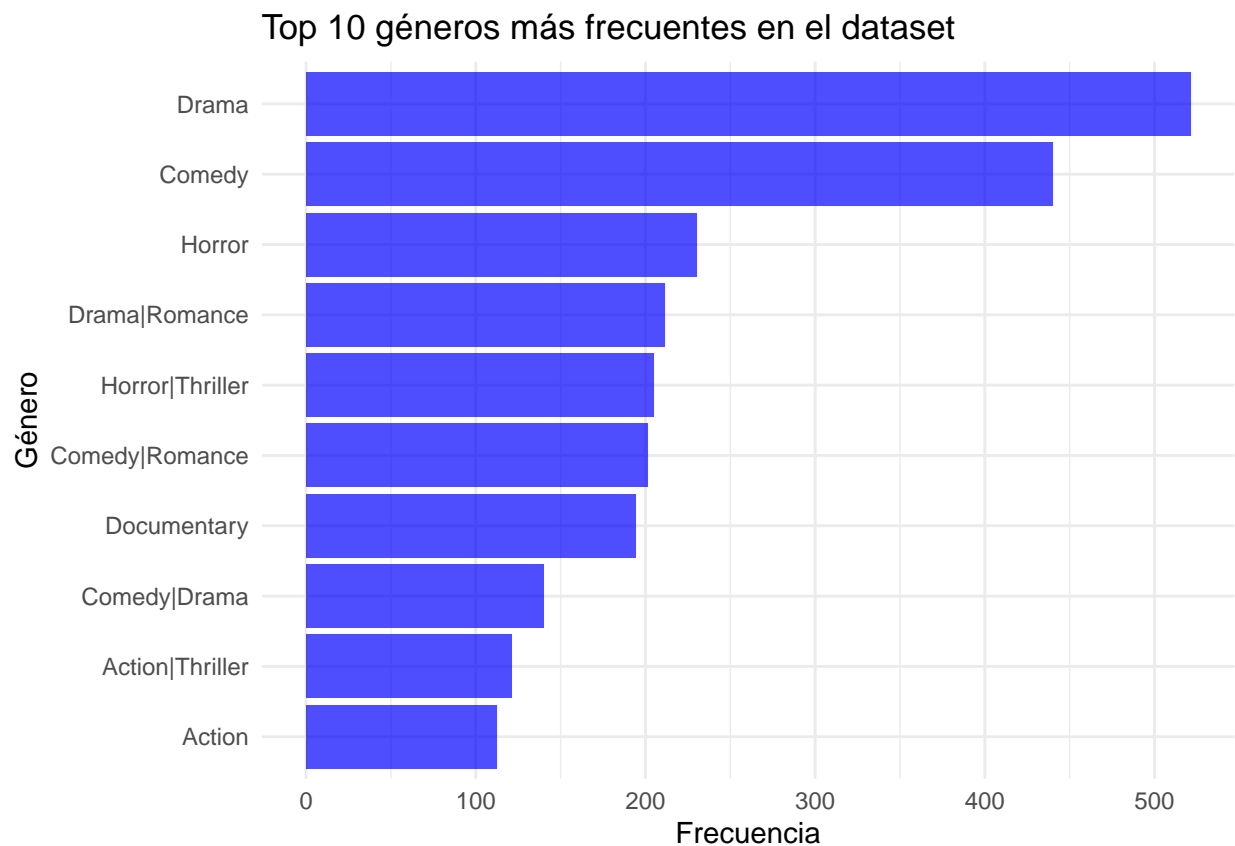
```
##
## --- Género principal que predomina en el conjunto de datos ---
```

```
print(most_common_genre)
```

```
## [1] "Drama"
```

```
# Crear gráfico para los 10 géneros más frecuentes en todo el dataset
top_genres <- sort(genre_distribution, decreasing = TRUE)[1:10]
top_genres_df <- data.frame(Género = names(top_genres), Frecuencia =
  ↪ as.vector(top_genres))

ggplot(top_genres_df, aes(x = reorder(Género, Frecuencia), y = Frecuencia)) +
  geom_bar(stat = "identity", fill = "blue", alpha = 0.7) +
  coord_flip() +
  labs(title = "Top 10 géneros más frecuentes en el dataset",
    x = "Género",
    y = "Frecuencia") +
  theme_minimal()
```



```
# 3. Identificar el género principal de las películas más largas
longest_movies <- movies %>%
  arrange(desc(runtime)) %>%
  head(10)

longest_genre_distribution <- table(longest_movies$genres)
longest_top_genre <- names(sort(longest_genre_distribution, decreasing = TRUE)[1])
cat("\n--- Género principal de las películas más largas ---\n")
```

```
##
## --- Género principal de las películas más largas ---
```

```
print(longest_top_genre)
```

```
## [1] "Documentary"
```

```
# Análisis y conclusiones
```

```
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
```

```
## --- Análisis y Conclusiones ---
```

```
cat("- El género principal de las 20 películas más recientes es:", recent_top_genre, ".  
↳ Esto refleja tendencias actuales en la industria cinematográfica.\n")
```

```
## - El género principal de las 20 películas más recientes es: Drama . Esto refleja tendencias actuales
```

```
cat("- El género principal en todo el conjunto de datos es:", most_common_genre, ", lo  
↳ que indica una preferencia general en la producción.\n")
```

```
## - El género principal en todo el conjunto de datos es: Drama , lo que indica una preferencia general
```

```
cat("- El género principal de las películas más largas es:", longest_top_genre, ". Esto  
↳ sugiere que ciertos géneros requieren más tiempo de metraje para desarrollar  
↳ historias complejas.\n")
```

```
## - El género principal de las películas más largas es: Documentary . Esto sugiere que ciertos géneros
```

```
cat("- Visualizar los géneros más frecuentes permite entender qué tipos de películas  
↳ dominan la industria y cómo evolucionan con el tiempo.\n")
```

```
## - Visualizar los géneros más frecuentes permite entender qué tipos de películas dominan la industria
```

4.7. ¿Las películas de qué género principal obtuvieron mayores ganancias?

```
# Cargar librerías necesarias
```

```
library(dplyr)
```

```
# Calcular las ganancias de cada película
```

```
movies <- movies %>%
```

```
  mutate(profit = revenue - budget) # Crear la columna de ganancias (profit)
```

```
# Calcular las ganancias totales por género
```

```
genre_profit <- movies %>%
```

```
  filter(genres != "") %>% # Filtrar géneros no vacíos
```

```
  group_by(genres) %>%
```

```
  summarise(total_profit = sum(profit, na.rm = TRUE)) %>% # Sumar ganancias por género
```

```
  arrange(desc(total_profit)) # Ordenar en orden descendente por ganancias
```

```
# Mostrar los 5 géneros con mayores ganancias
```

```
cat("\n--- Géneros con mayores ganancias ---\n")
```

```
##
## --- Géneros con mayores ganancias ---
```

```
top_genre_profit <- head(genre_profit, 5)
print(top_genre_profit)
```

```
## # A tibble: 5 x 2
##   genres                                total_profit
##   <chr>                                <dbl>
## 1 Action|Adventure|Science Fiction 14832790887
## 2 Comedy                          11692834444
## 3 Comedy|Romance                  10073967391
## 4 Drama                           9190582288
## 5 Adventure|Action|Science Fiction 8214486400
```

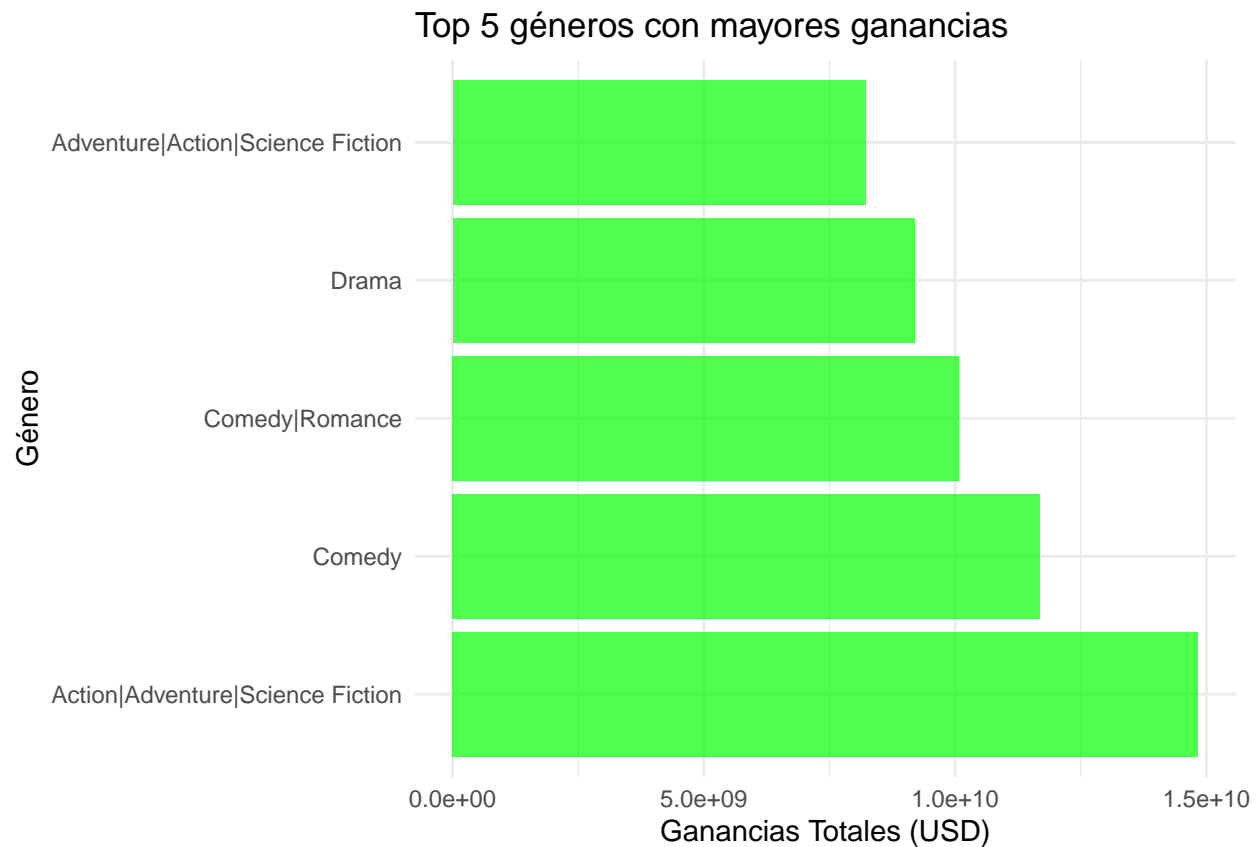
```
# Identificar el género con mayores ganancias
top_genre <- top_genre_profit[1, ]
cat("\n--- Género con mayores ganancias ---\n")
```

```
##
## --- Género con mayores ganancias ---
```

```
cat("El género con mayores ganancias es:", top_genre$genres,
    "con un total de", top_genre$total_profit, "USD.\n")
```

```
## El género con mayores ganancias es: Action|Adventure|Science Fiction con un total de 14832790887 USD
```

```
# Crear un gráfico de los 5 géneros más rentables
ggplot(top_genre_profit, aes(x = reorder(genres, -total_profit), y = total_profit)) +
  geom_bar(stat = "identity", fill = "green", alpha = 0.7) +
  labs(title = "Top 5 géneros con mayores ganancias",
       x = "Género",
       y = "Ganancias Totales (USD)") +
  theme_minimal() +
  coord_flip()
```



```
# Análisis y conclusiones
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
## --- Análisis y Conclusiones ---
```

```
cat("- El género con mayores ganancias es:", top_genre$genres,
    "con un total de", top_genre$total_profit, "USD.\n")
```

```
## - El género con mayores ganancias es: Action|Adventure|Science Fiction con un total de 14832790887 USD
```

```
cat("- Los géneros más rentables suelen incluir producciones de alto presupuesto con gran
↳ éxito en taquilla.\n")
```

```
## - Los géneros más rentables suelen incluir producciones de alto presupuesto con gran éxito en taquilla
```

```
cat("- Este análisis permite identificar los géneros más lucrativos, lo que puede guiar
↳ decisiones estratégicas en producción y marketing.\n")
```

```
## - Este análisis permite identificar los géneros más lucrativos, lo que puede guiar decisiones estratégicas
```



```
cat("- Visualizar las ganancias totales por género ayuda a priorizar inversiones en  
↳ géneros con mayor potencial económico.\n")
```

```
## - Visualizar las ganancias totales por género ayuda a priorizar inversiones en géneros con mayor pot
```

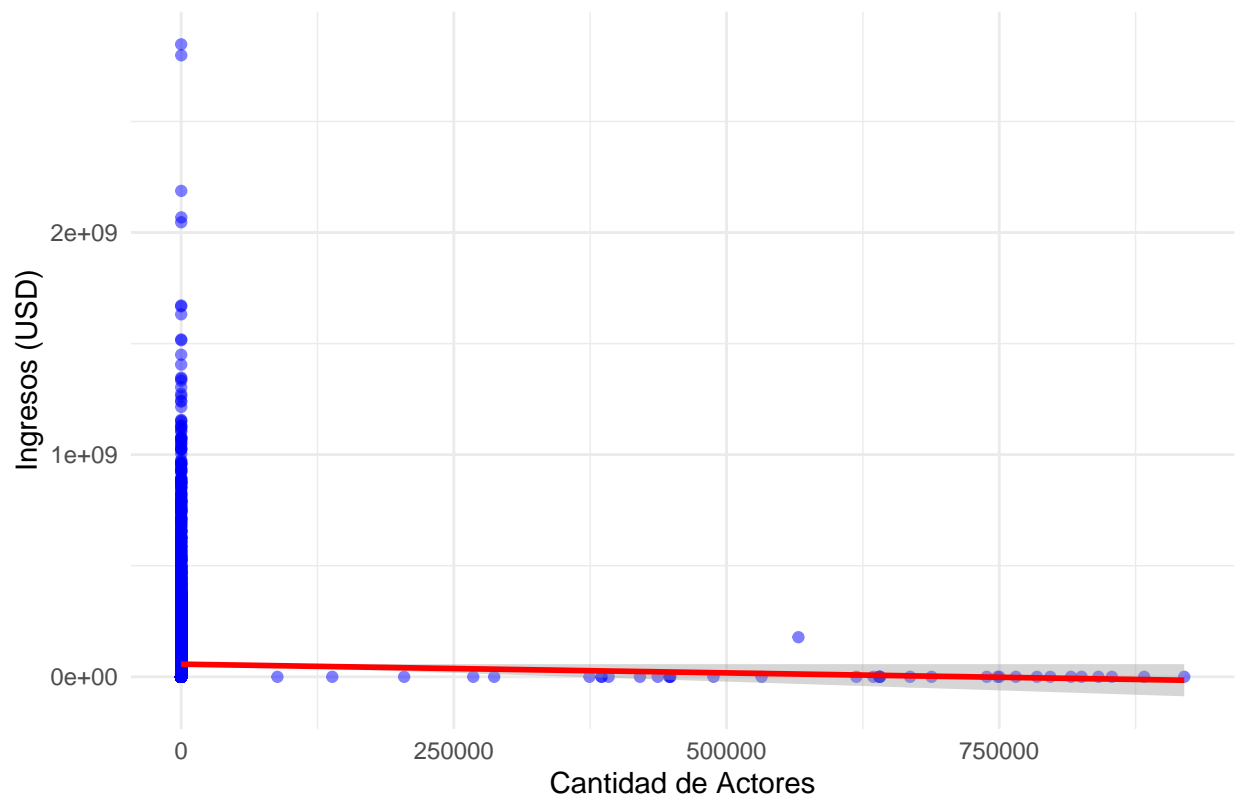
4.8. ¿La cantidad de actores influye en los ingresos de las películas? ¿se han hecho películas con más actores en los últimos años?

```
# Cargar librerías necesarias  
library(dplyr)  
library(ggplot2)  
  
# 1. Relación entre la cantidad de actores y los ingresos  
cat("\n--- Análisis: Relación entre cantidad de actores e ingresos ---\n")
```

```
##  
## --- Análisis: Relación entre cantidad de actores e ingresos ---
```

```
# Crear un gráfico de dispersión con línea de tendencia  
ggplot(movies, aes(x = actorsAmount, y = revenue)) +  
  geom_point(alpha = 0.5, color = "blue") +  
  geom_smooth(method = "lm", color = "red") +  
  labs(title = "Relación entre la cantidad de actores y los ingresos",  
        x = "Cantidad de Actores",  
        y = "Ingresos (USD)") +  
  theme_minimal()
```

Relación entre la cantidad de actores y los ingresos



```
# Análisis inicial de correlación
correlation <- cor(movies$actorsAmount, movies$revenue, use = "complete.obs")
cat("La correlación entre la cantidad de actores y los ingresos es:", correlation, "\n")
```

```
## La correlación entre la cantidad de actores y los ingresos es: -0.01955488
```

```
cat("- Una correlación positiva indica que más actores tienden a generar mayores  
↪ ingresos.\n")
```

```
## - Una correlación positiva indica que más actores tienden a generar mayores ingresos.
```

```
cat("- Sin embargo, valores extremos o géneros específicos pueden influir en esta  
↪ relación.\n")
```

```
## - Sin embargo, valores extremos o géneros específicos pueden influir en esta relación.
```

```
# 2. Evolución del número de actores en las películas por año
cat("\n--- Análisis: Evolución del número de actores en las películas ---\n")
```

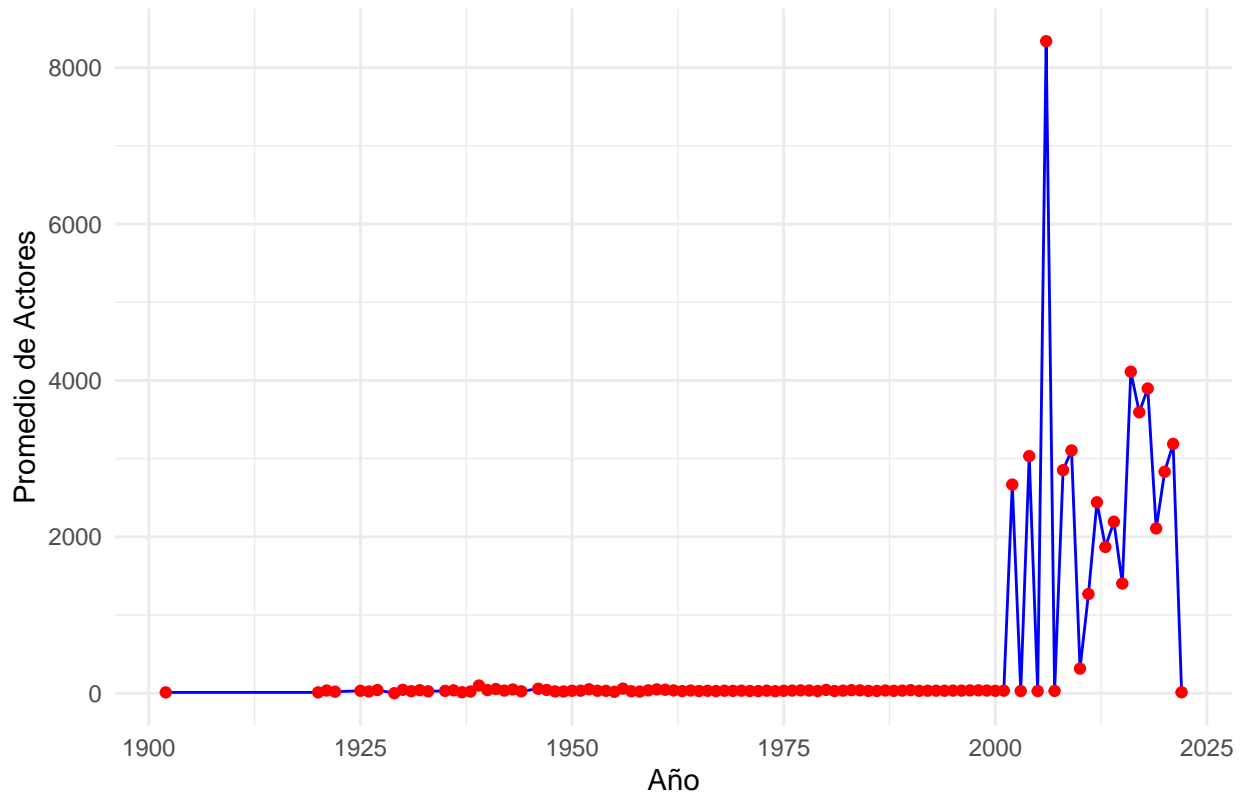
```
##
```

```
## --- Análisis: Evolución del número de actores en las películas ---
```

```
# Calcular el promedio de actores por año
actors_per_year <- movies %>%
  group_by(releaseYear) %>%
  summarise(avg_actors = mean(actorsAmount, na.rm = TRUE))

# Crear un gráfico de líneas para visualizar la tendencia
ggplot(actors_per_year, aes(x = releaseYear, y = avg_actors)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Evolución del número de actores en las películas por año",
       x = "Año",
       y = "Promedio de Actores") +
  theme_minimal()
```

Evolución del número de actores en las películas por año



```
# Identificar el año con el promedio más alto de actores
max_avg_actors <- actors_per_year %>%
  filter(avg_actors == max(avg_actors, na.rm = TRUE))
cat("El año con el mayor promedio de actores es:", max_avg_actors$releaseYear,
    "con un promedio de", max_avg_actors$avg_actors, "actores.\n")
```

```
## El año con el mayor promedio de actores es: 2006 con un promedio de 8338.61 actores.
```

```
# Conclusiones
cat("\n--- Análisis y Conclusiones ---\n")
```

```
##
## --- Análisis y Conclusiones ---
```

```
cat("- La correlación entre la cantidad de actores y los ingresos es de:", correlation,
  ↪ "\n")
```

```
## - La correlación entre la cantidad de actores y los ingresos es de: -0.01955488
```

```
cat(" Esto sugiere que, aunque más actores pueden contribuir a mayores ingresos, la
  ↪ relación no es extremadamente fuerte.\n")
```

```
## Esto sugiere que, aunque más actores pueden contribuir a mayores ingresos, la relación no es extre
```

```
cat("- El análisis muestra que las películas con más actores tienden a ser producidas en
  ↪ años recientes, reflejando una tendencia creciente en la industria.\n")
```

```
## - El análisis muestra que las películas con más actores tienden a ser producidas en años recientes, 1
```

```
cat("- Estas tendencias pueden ser el resultado de mayores producciones en géneros como
  ↪ acción o aventuras, donde se requieren grandes elencos.\n")
```

```
## - Estas tendencias pueden ser el resultado de mayores producciones en géneros como acción o aventuras
```

```
cat("- Comprender esta relación es clave para equilibrar costos y beneficios al
  ↪ planificar el reparto en futuras producciones.\n")
```

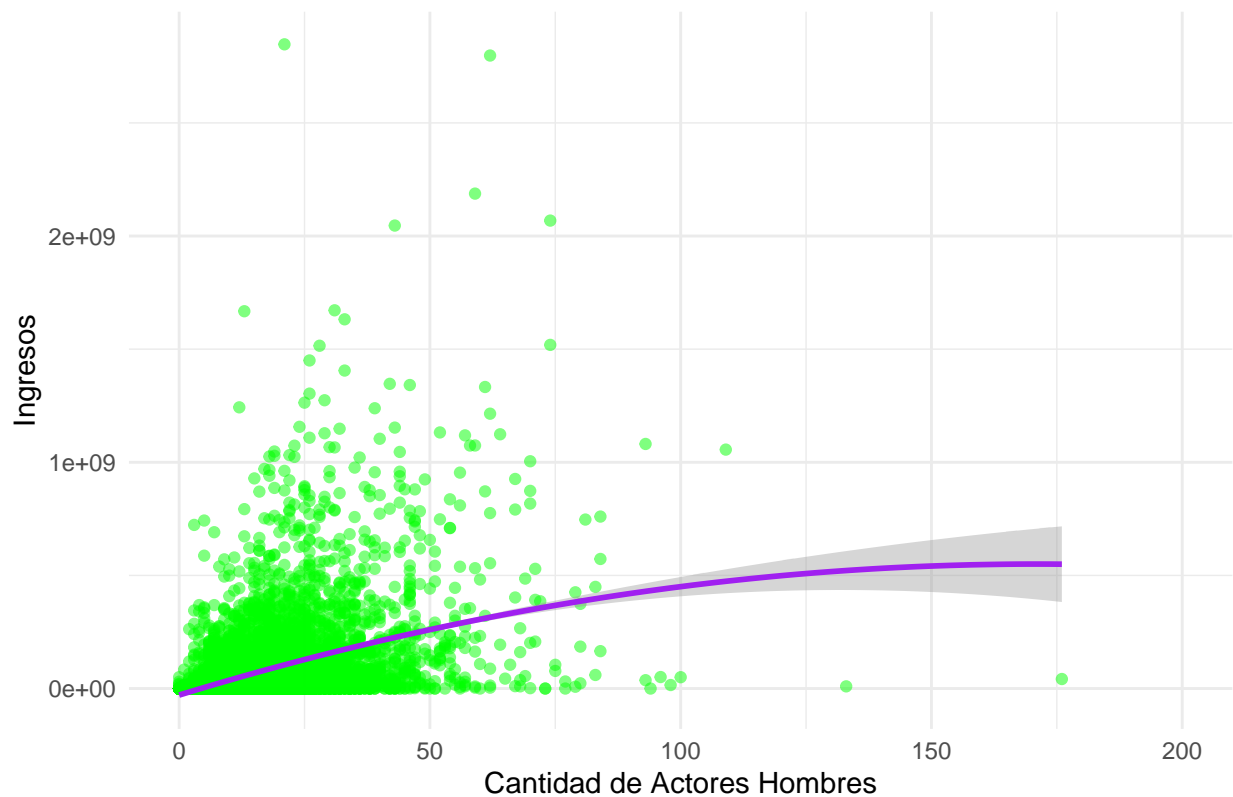
```
## - Comprender esta relación es clave para equilibrar costos y beneficios al planificar el reparto en 1
```

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

Primero obtendremos la relación entre hombres e ingresos.

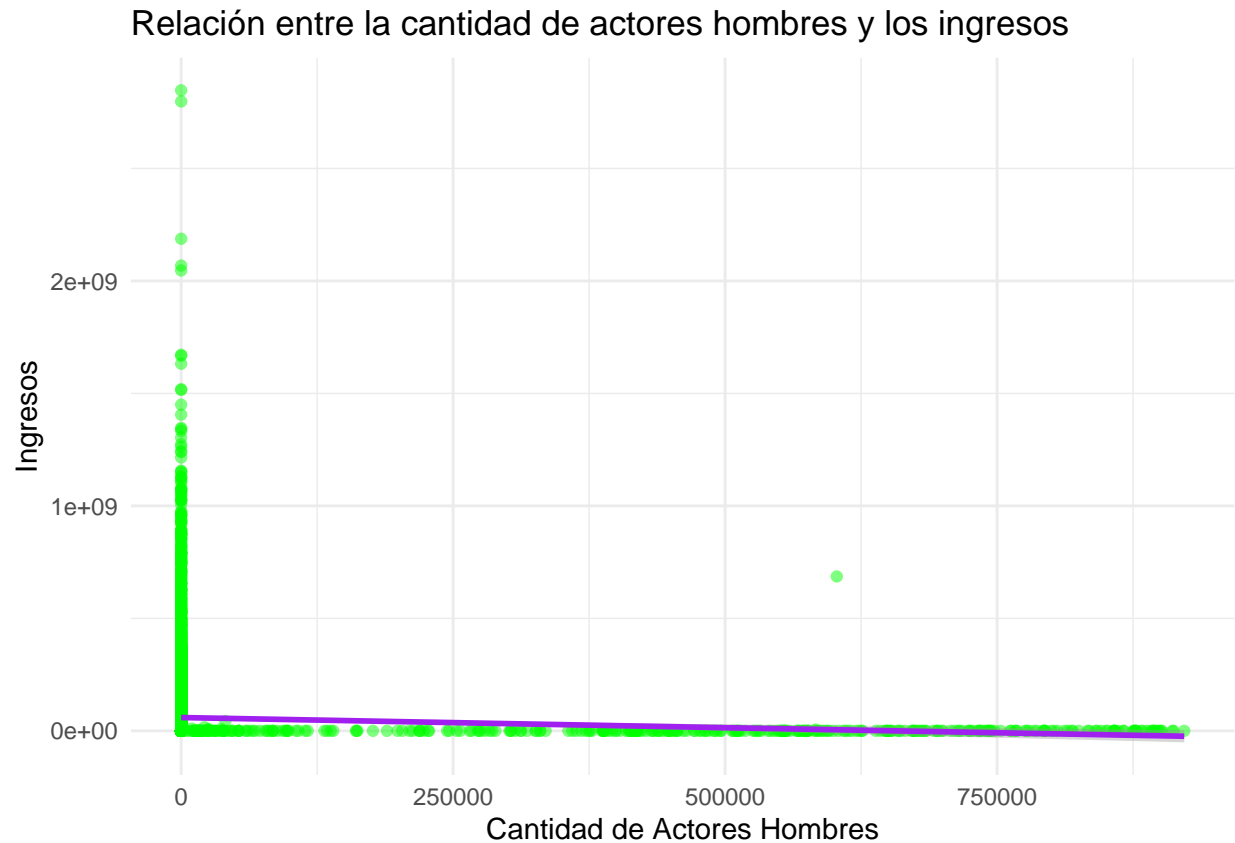
```
ggplot(movies, aes(x = as.numeric(castMenAmount), y = revenue)) +
  geom_point(alpha = 0.5, color = "green") +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "purple") +
  labs(title = "Relación entre la cantidad de actores hombres y los ingresos",
    x = "Cantidad de Actores Hombres",
    y = "Ingresos") +
  scale_x_continuous(limits = c(0, 200)) +
  theme_minimal()
```

Relación entre la cantidad de actores hombres y los ingresos



Grafica de Dispersion Cantidad Hombres vs Ingresos , 200 datos posibles Si no lo acotamos en el eje x nos daría

```
echo=FALSE
ggplot(movies, aes(x = as.numeric(castMenAmount), y = revenue)) +
  geom_point(alpha = 0.5, color = "green") +
  geom_smooth(method = "lm", color = "purple") +
  labs(title = "Relación entre la cantidad de actores hombres y los ingresos",
        x = "Cantidad de Actores Hombres",
        y = "Ingresos")+
  theme_minimal()
```

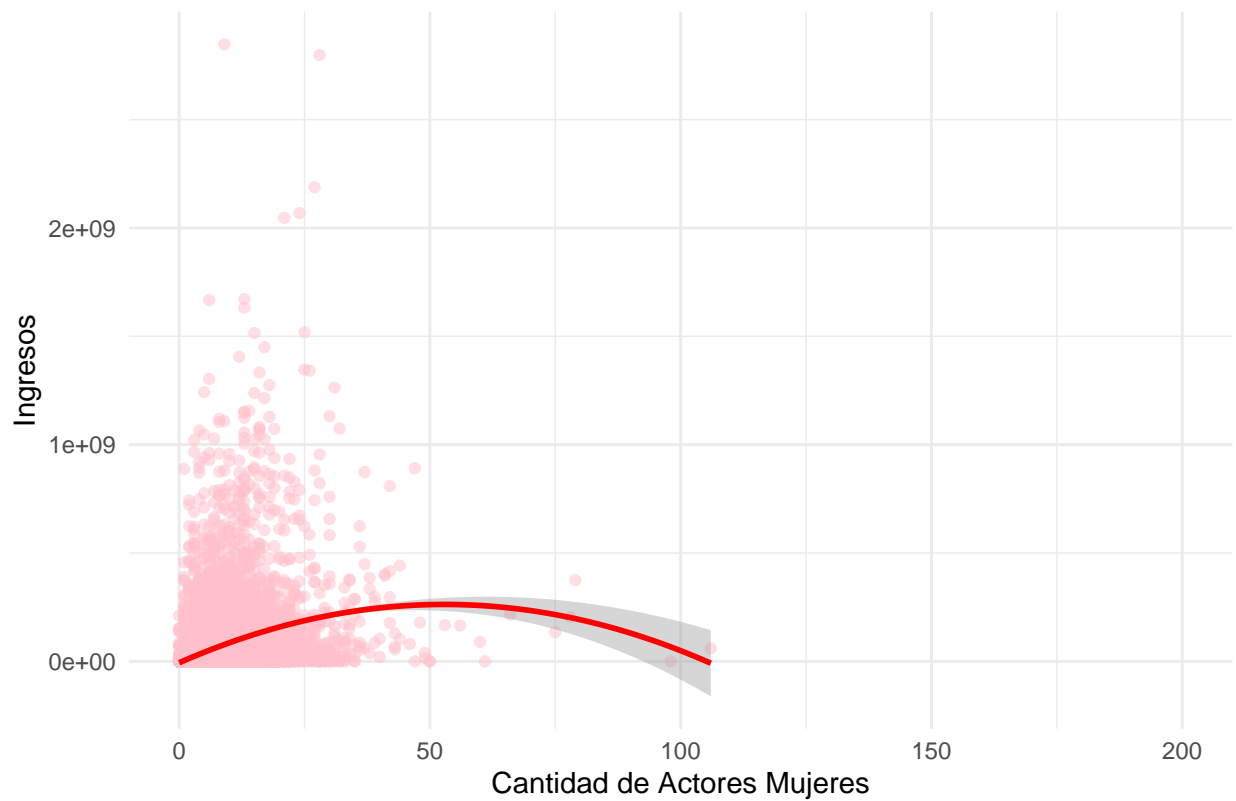


Análisis Gráfico Como se puede observar en la grafica de dispersion, tiende a mas ingresos mientras mas numeros de hombres existan pero ojo esto no es asi ya que a partir de cierto punto empieza a bajar drasticamente los ingresos. Esto se ve en la segunda grafica donde la cantidad de hombres baja considerablemente. Muy posiblemente debido a que mientras mas actores las personas pierden interes.

Ahora obtendremos la relacion de mujeres

```
ggplot(movies, aes(x = as.numeric(castWomenAmount), y = revenue)) +
  geom_point(alpha = 0.5, color = "pink") +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "red") +
  labs(title = "Relación entre la cantidad de actores mujeres y los ingresos",
        x = "Cantidad de Actores Mujeres",
        y = "Ingresos") +
  scale_x_continuous(limits = c(0, 200)) +
  theme_minimal()
```

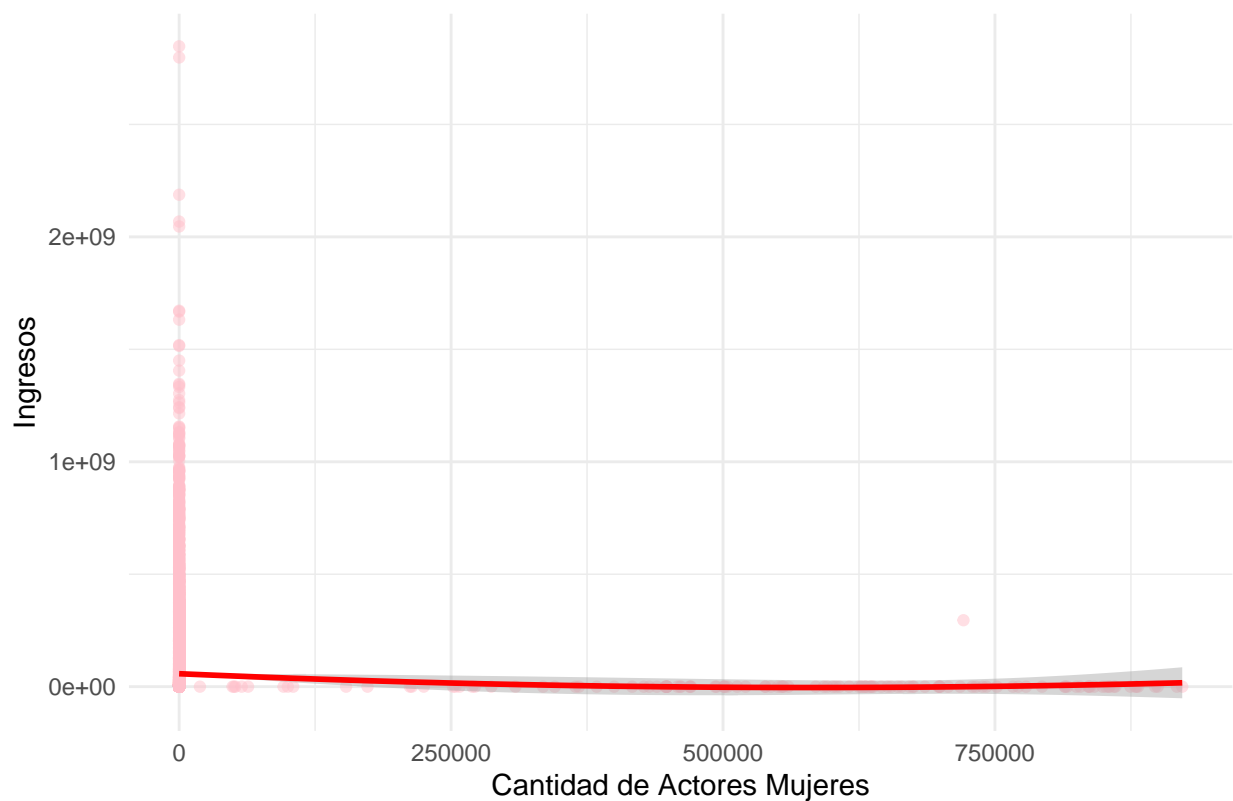
Relación entre la cantidad de actores mujeres y los ingresos



Si no lo acotamos en el eje x nos daría

```
ggplot(movies, aes(x = as.numeric(castWomenAmount), y = revenue)) +  
  geom_point(alpha = 0.5, color = "pink") +  
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "red") +  
  labs(title = "Relación entre la cantidad de actores mujeres y los ingresos",  
        x = "Cantidad de Actores Mujeres",  
        y = "Ingresos") +  
  theme_minimal()
```

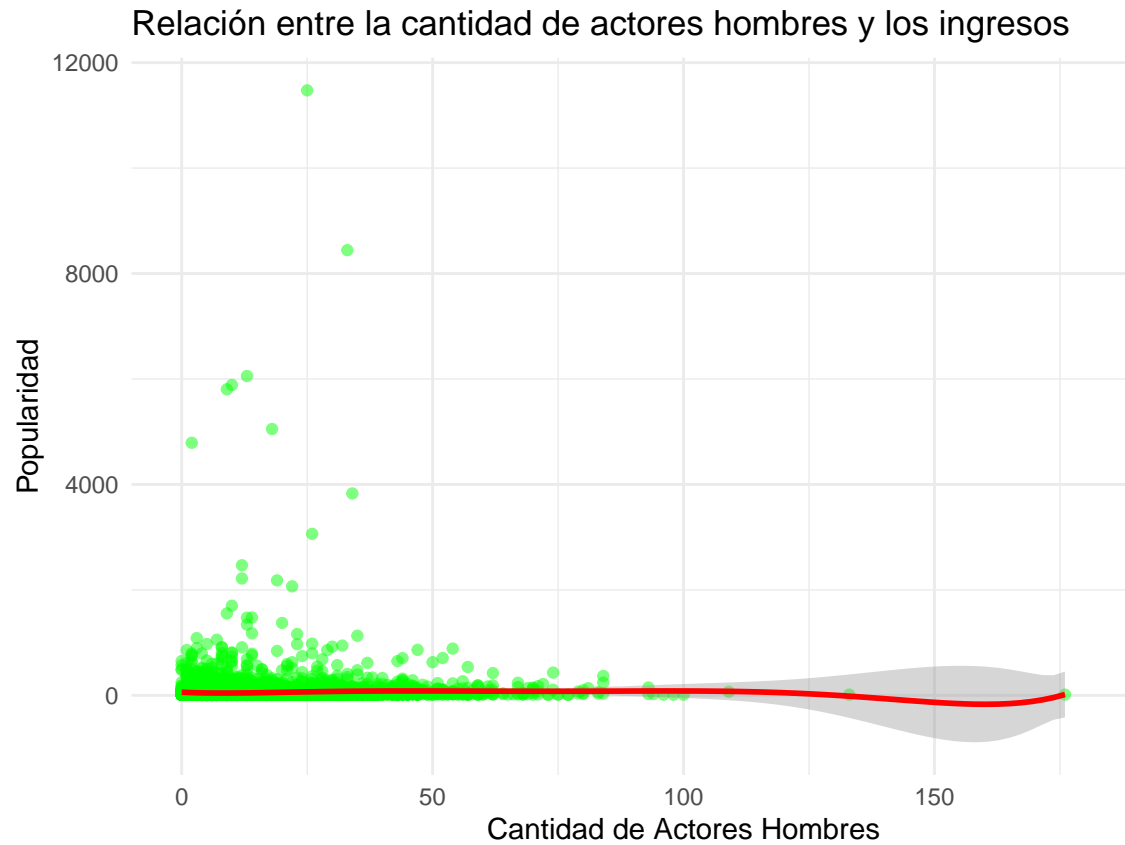
Relación entre la cantidad de actores mujeres y los ingresos



Análisis Gráfico Como se puede observar en la primera grafica tiende a existir mas ingresos mientras mas mujeres existan , sin embargo al igual que con la de los hombres esta suele decaer en cierto punto maximo hasta alcanzar a llegar a minimos de ingresos como se ve en la segunda grafica.

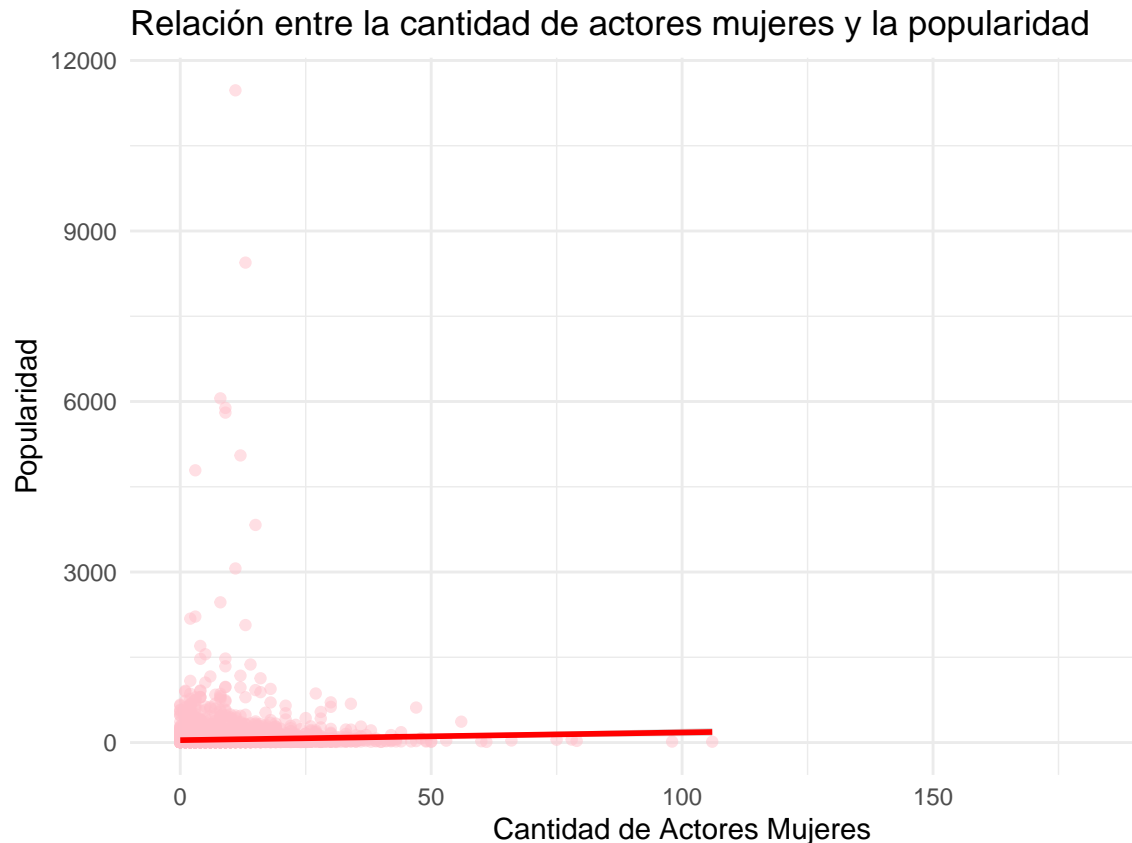
Ahora analizaremos su popularidad

```
ggplot(movies, aes(x = as.numeric(castMenAmount), y = popularity)) +
  geom_point(alpha = 0.5, color = "green") +
  geom_smooth(method = "lm", formula = y ~ poly(x, 6), color = "red") +
  labs(title = "Relación entre la cantidad de actores hombres y los ingresos",
        x = "Cantidad de Actores Hombres",
        y = "Popularidad") +
  scale_x_continuous(limits = c(0, 200)) +
  theme_minimal()
```

Popularidad Hombres

```
ggplot(movies, aes(x = as.numeric(castWomenAmount), y = popularity)) +
  geom_point(alpha = 0.5, color = "pink") +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Relación entre la cantidad de actores mujeres y la popularidad",
        x = "Cantidad de Actores Mujeres",
        y = "Popularidad") +
  scale_x_continuous(limits = c(0, 200)) + # Limitar el rango en el eje X
  theme_minimal()
```



Popularidad Mujeres

Análisis Gráfico Como se puede observar en los hombres suele existir una distribución de hombres con mayor popularidad en las películas que participan. En cambio en las mujeres las películas en las que menor cantidad de actrices suelen ser las que tienen mayor popularidad.

Conclusiones La cantidad de actores de ambos géneros mientras mayor sea la popularidad y los ingresos de la película bajan. Muy posiblemente a que mientras más actores sean mayor costo de producción y menor la retención de las personas en los personajes. De igual forma se ve una tendencia a que mientras mayor sea la cantidad de mujeres mejora la taquilla pero no lo hace en la popularidad de la película.

4.10 ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

```
library(knitr)
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)
movies[] <- lapply(movies, function(x) iconv(x, from = "latin1", to = "UTF-8", sub =
  ↪ "byte"))

movies <- movies %>%
  mutate(voteAvg = as.numeric(voteAvg)) # Convertir voteAVG a numero

top_20_movies <- movies %>%
  filter(director != "") %>%
  arrange(desc(voteAvg)) %>%
```

```

select(title, director, voteAvg) %>%
head(20)
knitr::kable(top_20_movies, format = "html")

```

title	director	voteAvg
Hot Naked Sex & the City	Thomas Coven	10.0
Holidays	Víctor Barba Juan Olivares	10.0
Steven Universe: The Movie: Behind the Curtain	Rebecca Sugar	10.0
Spirit of Vengeance: The Making of ‘Ghost Rider’	Laurent Bouzereau	10.0
How Ponyo was Born ~Hayao Miyazaki’s Thought Process~	Kaku Arakawa	10.0
Christmas at the Ranch	Christin Baker	10.0
Los Vengadores Chiflados	Miguel Angel Zavala	10.0
The Spectacular Spider-Man Attack of the Lizard	Dave Bullock Troy Adomitis Victor Cook	9.6
Ebola Zombies	Samuel Leong	9.5
Aunt’s Temptation 3	Won Myeong-jun	9.5

Live: The Last Concert

Selena Quintanilla

9.4

Demon Slayer: Kimetsu no Yaiba the Hashira Meeting Arc

Haruo Sotozaki

9.3

Demon Slayer: Kimetsu no Yaiba Sibling's Bond

Haruo Sotozaki

9.3

Franco Escamilla: Por La Anécdota

Ulises Valencia

9.2

BTS World Tour: Love Yourself - Japan Edition

Kim Nam-joon|Jeon Jung-kook|Kim Tae-hyung|Park Ji-min|Jung Ho-seok|Kim Seok-jin|Min Yoon-gi

9.2

Break the Silence: The Movie

Park Jun-soo

9.2

Mission «Sky»

Igor Kopylov

9.2

Bring the Soul: The Movie

Park Jun-soo

9.1

Scooby-Doo! and the Spooky Scarecrow

Michael Goguen

9.0

Three Preludes for Solo Piano By Adam Sherkin

Filip Ghiorghi

9.0

Conclusiones Se puede observar que los 20 directores con las películas mejor calificadas suelen ser de un rango de 9.0 a 10 de nota dada por la crítica. Lo cual es un buen indicativo de la capacidad de los directores de poder conseguir buena crítica.

4.11 ¿Cómo se correlacionan los presupuestos con los ingresos?

Para ello haremos una gráfica de correlación

```

movies <- read.csv("movies.csv", stringsAsFactors = FALSE)
movies$budget <- as.numeric(movies$budget)
movies$revenue <- as.numeric(movies$revenue)

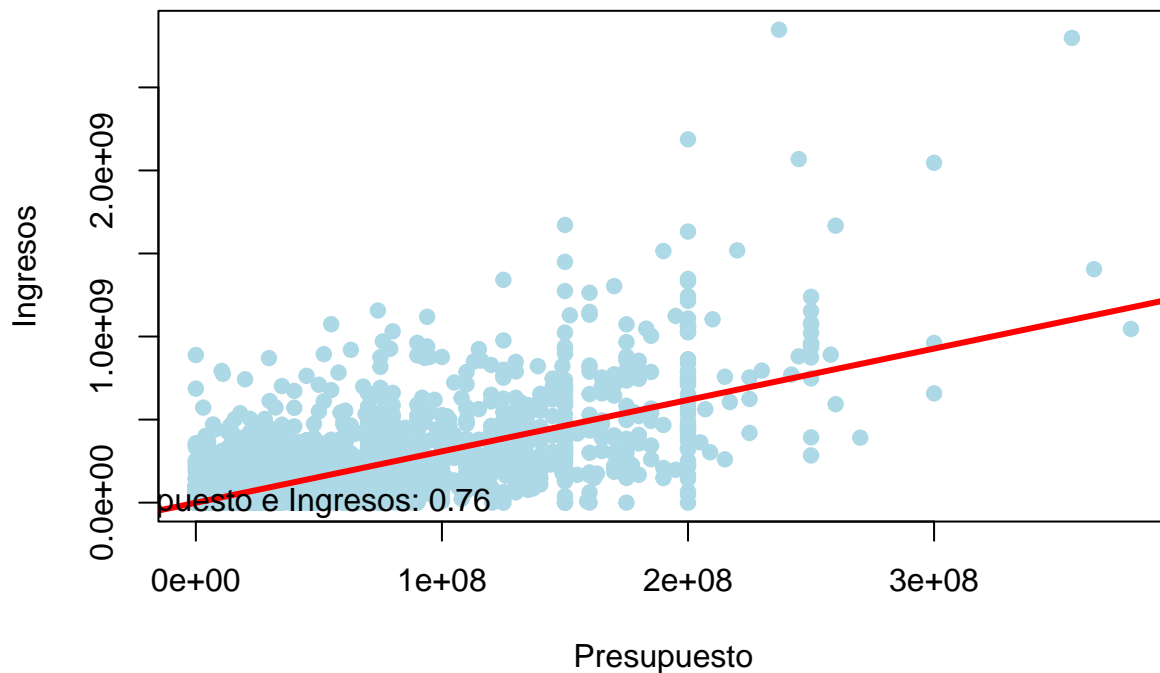
x <- movies$budget
y <- movies$revenue

# Creamos el gráfico
plot(x, y, pch = 19, col = "lightblue",
     xlab = "Presupuesto", # Título del eje X
     ylab = "Ingresos")

# Línea de regresión
abline(lm(y ~ x), col = "red", lwd = 3)

text(paste("Correlación Presupuesto e Ingresos:", round(cor(x, y), 2)), x = 25, y = 95)

```



Conclusiones Como se puede observar en el gráfico de dispersión si hay una relación que tiende a un mayor presupuesto una mayor recaudación. Esto se debe a que una parte de los presupuestos se suele destinar a la publicidad lo que puede influir en su crecimiento. Sin embargo podemos observar ciertos puntos atípicos por encima y debajo de la línea de tendencia que muestran que muchas veces con menor presupuesto se logran una gran cantidad de ingresos y con mucho presupuesto suelen tener ingresos muy bajos. Aun así la correlación es muy alta siendo de 0.76 lo que indica que hay una correlación entre ambas variables.

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

Para ello agruparemos los meses y calcularemos el total de ingresos

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)
movies$releaseMonth <- as.numeric(substr(movies$releaseDate, 6, 7))

monthly_revenue <- movies %>%
  group_by(releaseMonth) %>%
  summarise(total_revenue = sum(revenue, na.rm = TRUE),
            avg_revenue = mean(revenue, na.rm = TRUE))%>%
  select(releaseMonth, total_revenue, avg_revenue)%>%
  arrange(desc(avg_revenue))

knitr::kable(monthly_revenue, format = "html")
```

releaseMonth

total_revenue

avg_revenue

6

77597881637

94747108

5

61316118519

87845442

7

61735301475

76028696

12

69525553232

74358880

11

57694134749

71492112

4

36606574887

52595654

3

41659492701

51115942
 2
 30293297435
 42908353
 10
 41638470602
 38987332
 8
 32840682550
 35970079
 1
 22020446418
 33773691
 9
 34451301833
 31928917

Conclusiones Si hay una asociación, haciendo una métrica de total de ingresos y promedio de ingresos se puede ver que el mes con mas recaudación es junio , mayo y julio. Esto se puede deber a que son los meses de vacaciones de verano y muchas personas asisten a ver películas en tiempos libres.

4.13 ¿En qué meses se han visto los lanzamientos con mejores ingresos? ¿cuantas películas, en promedio, se han lanzado por mes?

Viendo la tabla anterior se puede ver que los meses de junio, mayo y julio siguen siendo los mejores meses para poder lanzar una película al tener los mayores ingresos.

Ahora encontraremos el total de películas lanzadas.

```
#El total seria
count_movie_month_total <- movies %>%
  group_by(releaseMonth) %>%
  summarise(movie_count = n())%>%
  select(releaseMonth,movie_count)%>%
  arrange(desc(movie_count))

knitr::kable(count_movie_month_total, format = "html")
```

releaseMonth
 movie_count
 9
 1079
 10

1068
12
935
8
913
6
819
3
815
7
812
11
807
2
706
5
698
4
696
1
652

Como siguiente paso veremos el promedio de películas por año

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)
movies$releaseYear <- as.numeric(substr(movies$releaseDate, 1, 4))
movies$releaseMonth <- as.numeric(substr(movies$releaseDate, 6, 7))
count_movie_month <- movies %>%
  group_by(releaseMonth, releaseYear) %>%
  summarise(movie_count = n(), .groups = 'drop')

# Calcular el promedio de películas por mes
average_movies_per_month <- count_movie_month %>%
  group_by(releaseMonth) %>%
  summarise(avg_movie_count = mean(movie_count)) %>%
  select(releaseMonth, avg_movie_count) %>%
  arrange(desc(avg_movie_count))

print(average_movies_per_month)
```

```
## # A tibble: 12 x 2
##   releaseMonth avg_movie_count
##           <dbl>           <dbl>
```


## 1	9	18.3
## 2	10	15.3
## 3	8	14.3
## 4	3	14.1
## 5	11	12.8
## 6	7	12.5
## 7	12	12.0
## 8	1	11.6
## 9	5	11.6
## 10	6	11.2
## 11	2	11.0
## 12	4	10.5

Conclusiones Podemos observar que en total de películas lanzadas la mayoría en promedio y total han sido siempre en Septiembre. Y el mes de Junio ha sido si bien el que mejor ingresos ha tenido en el que menos películas en promedio se han lanzado siendo de 11 aproximadamente. Y 819 en total de películas lanzadas, sin embargo se han recaudado mayor cantidad de ingresos en dicho mes. Muy posiblemente debido a las vacaciones de verano que inician en Junio y terminan en Septiembre, y son épocas en donde los padres llevan a sus hijos a ver películas.

Por lo que si se quiere lanzar una película sería entre los meses 5, 6, 7 y 8. Debido a que son los meses en donde no se están haciendo tantos lanzamientos para cada uno, pero si están rindiendo en ingresos.

Lo que se recomienda es tomar en cuenta la popularidad de las películas por cada mes, para así poder tener una perspectiva contra que se compite.

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

Para ello realizaremos una gráfica de correlación entre su éxito que se podría ver como la diferencia entre ingresos - presupuesto / ingresos que es un porcentaje de ganancia que se tiene para la película y en el eje y tendremos su calificación

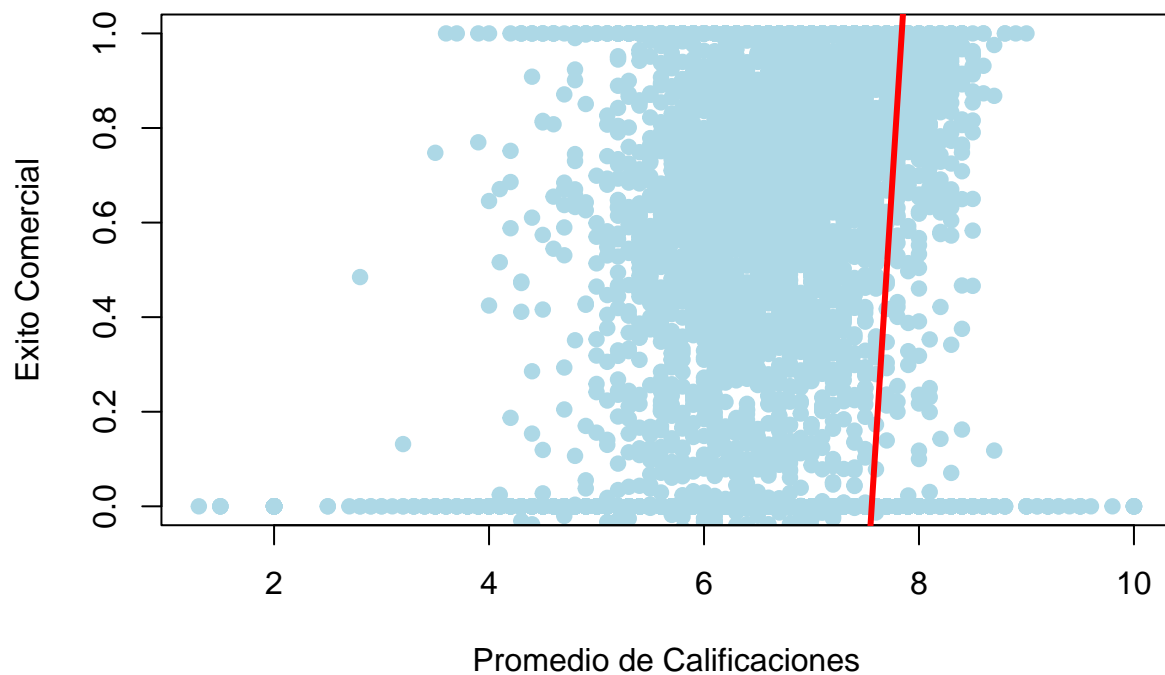
```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)
# Reemplazar NA e Inf en 'revenue', 'budget', y 'voteAvg' por la mediana
movies$revenue[is.na(movies$revenue) | is.infinite(movies$revenue)] <-
  ↪ median(movies$revenue, na.rm = TRUE)
movies$budget[is.na(movies$budget) | is.infinite(movies$budget)] <- median(movies$budget,
  ↪ na.rm = TRUE)
movies$voteAvg[is.na(movies$voteAvg) | is.infinite(movies$voteAvg)] <-
  ↪ median(movies$voteAvg, na.rm = TRUE)

# Calcular las ganancias (gains)
movies$ganancia_neta <- ifelse(as.numeric(movies$revenue) == 0, 0,
  (as.numeric(movies$revenue) - as.numeric(movies$budget)) /
  ↪ as.numeric(movies$revenue))

x <- movies$voteAvg
y <- movies$ganancia_neta
plot(x, y, pch = 19, col = "lightblue",
  xlab = "Promedio de Calificaciones", # Título del eje X
  ylab = "Éxito Comercial",
  ylim = c(0, max(y, na.rm = TRUE)))
```

```
# Agregar la línea de regresión
abline(lm(y ~ x), col = "red", lwd = 3)

text(paste("Correlación de Éxito comercial y Calificaciones:", round(cor(x, y), 2)), x =
  ↪ 25, y = 95)
```



Análisis Gráfico Este grafico nos muestra aquellos de 0 a 1 de ganancias. 1 es que generaron ganancias del 100% y 0% que su presupuesto no genero ganancias esto quiere decir que los ingresos apenas alcanzaron para cubrir los costes. Por otro lado tambien se clasifico como 0 aquellos que no solo no generaron ganancias sino que tampoco generaron ingresos. Cabe aclarar que aqui se colocan todos aquellos que tienen exito comercial (osea apartir de 0) , pues hay peliculas con ganancias negativas las cuales son fracasos que perdieron dinero, y por ende no se toman para el analisis.

Si quitamos las que no tuvieron ingresos ya que tambien son fracasos tendriamos esto

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)
# Reemplazar NA e Inf en 'revenue', 'budget', y 'voteAvg' por la mediana
movies$revenue[is.na(movies$revenue) | is.infinite(movies$revenue)] <-
  ↪ median(movies$revenue, na.rm = TRUE)
movies$budget[is.na(movies$budget) | is.infinite(movies$budget)] <- median(movies$budget,
  ↪ na.rm = TRUE)
movies$voteAvg[is.na(movies$voteAvg) | is.infinite(movies$voteAvg)] <-
  ↪ median(movies$voteAvg, na.rm = TRUE)
```

```

movies$gains <- (as.numeric(movies$revenue) - as.numeric(movies$budget)) /
  ↳ as.numeric(movies$revenue)

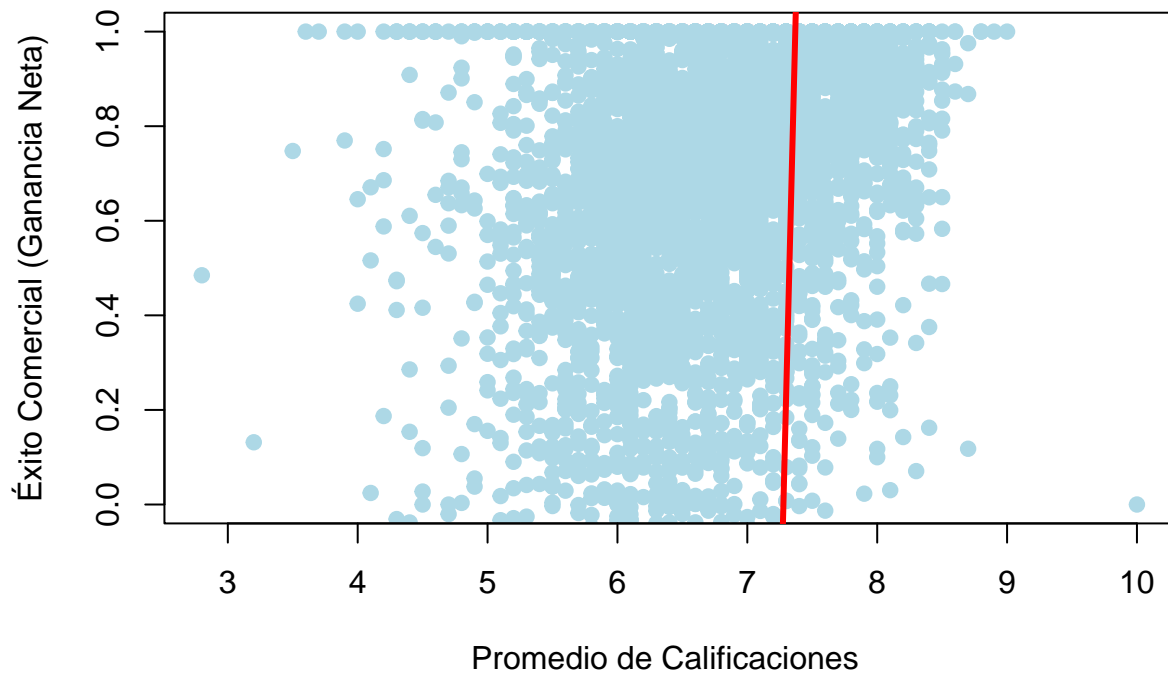
# Eliminar valores Inf o NA en 'gains'
movies <- movies[is.finite(movies$gains), ]

x <- movies$voteAvg
y <- movies$gains
plot(x, y, pch = 19, col = "lightblue",
     xlab = "Promedio de Calificaciones ", # Título del eje X
     ylab = "Éxito Comercial (Ganancia Neta)",
     ylim = c(0, max(y, na.rm = TRUE)))

# Agregar la línea de regresión
abline(lm(y ~ x), col = "red", lwd = 3)

text(paste("Correlación de Éxito comercial y Calificaciones:", round(cor(x, y), 2)), x =
  ↳ 25, y = 95)

```



Análisis Gráfico Existe relación, también podemos observar que la mayoría en promedio es calificada con un 6, vemos que una parte importante de las películas que son calificadas por la crítica por debajo del 6 tienen una ganancia neta del 100%. Lo que indica que sus ingresos superaron o triplicaron las ganancias de las películas. Aun así hay muchas cintas que lograron tener muy buena crítica y tener un éxito comercial.

Conclusion Como se puede observar el éxito comercial y las calificaciones, sin embargo una gran cantidad de películas con un puntaje mayor a 6. Hay tenido un rendimiento muy alto en sus ganancias. Lo cual indica que las personas si desean pagar por ver una película buena. Y de hecho este éxito puede ser muy influenciado en la percepción de lo buena que sea una cinta.

4.15 ¿Qué estrategias de marketing, como videos promocionales o páginas oficiales, generan mejores resultados?

Lo que haremos es comparar usando la ganancia entre las 2 estrategias.

Marketing de Video

Si medimos sus ingresos veremos lo siguiente.

```
# Lo primero que haremos es mapear
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

marketing <- movies %>%
  select(video, revenue)

# Ahora balanceamos
video_1 <- marketing %>% filter(video == TRUE)

# Seleccionar aleatoriamente 84 observaciones de video = 0 y de N/A
set.seed(123) # Asegura reproducibilidad
video_0_sample <- marketing %>%
  filter(video == FALSE) %>%
  sample_n(84)

video_na_sample <- marketing %>%
  filter(is.na(video)) %>%
  sample_n(84)

# Unir ambos subconjuntos para tener balanceado el dataset
balanced_marketing <- bind_rows(video_1, video_0_sample, video_na_sample)

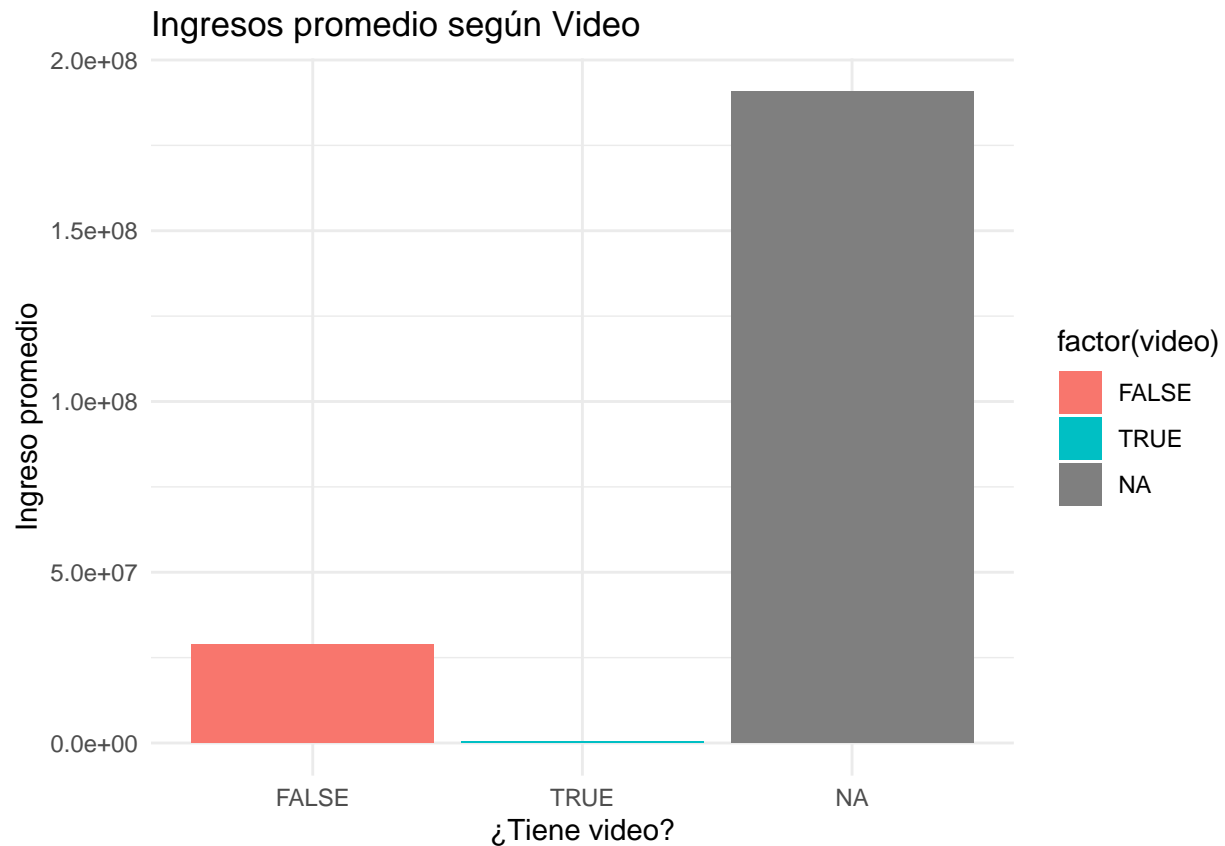
balanced_counts <- balanced_marketing %>%
  count(video)

print(balanced_counts)
```

```
##   video  n
## 1 FALSE 84
## 2  TRUE 84
## 3   NA 84
```

```
balanced_marketing %>%
  group_by(video) %>%
  summarise(avg_revenue = mean(revenue, na.rm = TRUE)) %>%
```

```
ggplot(aes(x = factor(video), y = avg_revenue, fill = factor(video))) +
  geom_col() +
  labs(title = "Ingresos promedio según Video",
        x = "¿Tiene video?",
        y = "Ingreso promedio") +
  theme_minimal()
```



Análisis Gráfico Lo que vemos aquí es el promedio de ingresos dependiendo si tiene video o no tiene. Por lo que se ve aquí es que no es concluyente por ende no se puede decir que tener o no video afecta a los ingresos percibidos. Pues tenemos demasiados valores que no conocemos si tienen o no video, por ende no se puede hacer una conclusión. Pero se puede ver que las que no tienen video tienen mayores ingresos.

Ahora si medimos su popularidad veremos lo siguiente

```
# Lo primero que haremos es mapear
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

marketing <- movies %>%
  select(video, popularity)

# Ahora balanceamos
video_1 <- marketing %>% filter(video == TRUE)
```

```

# Seleccionar aleatoriamente 84 observaciones de video = 0 y de N/A
set.seed(123) # Asegura reproducibilidad
video_0_sample <- marketing %>%
  filter(video == FALSE) %>%
  sample_n(84)

video_na_sample <- marketing %>%
  filter(is.na(video)) %>%
  sample_n(84)

# Unir ambos subconjuntos para tener balanceado el dataset
balanced_marketing <- bind_rows(video_1, video_0_sample, video_na_sample)

balanced_counts <- balanced_marketing %>%
  count(video)

print(balanced_counts)

```

```

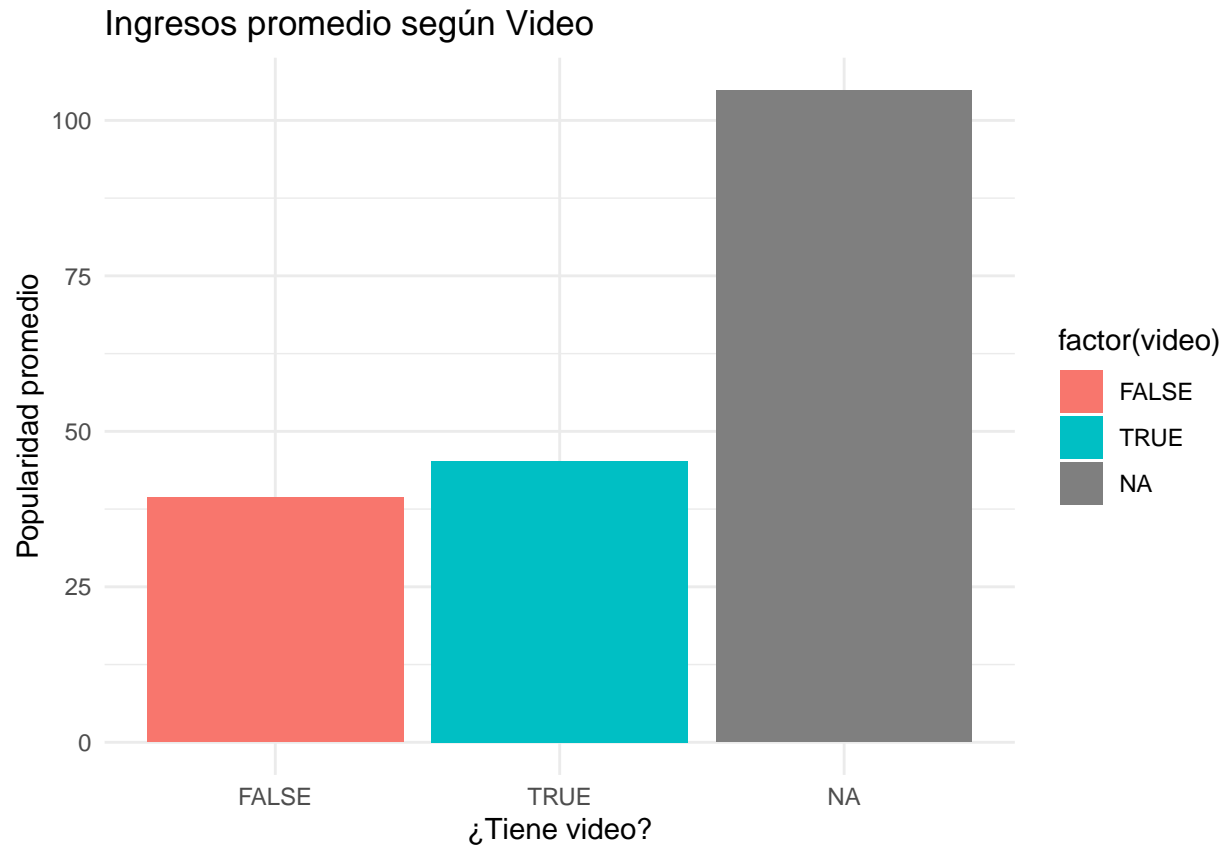
##   video  n
## 1 FALSE 84
## 2  TRUE 84
## 3   NA 84

```

```

balanced_marketing %>%
  group_by(video) %>%
  summarise(avg_popularity = mean(popularity, na.rm = TRUE)) %>%
  ggplot(aes(x = factor(video), y = avg_popularity, fill = factor(video))) +
  geom_col() +
  labs(title = "Ingresos promedio según Video",
       x = "¿Tiene video?",
       y = "Popularidad promedio") +
  theme_minimal()

```



Analisis Grafico

Podemos ver que aquellas que tienen video suelen ser mucho mas populares que las que no. Esto puede deberse a que es muy comun en redes sociales que las que no tienen video suelen ser mas visitados su video que las que no, sin embargo aun no es concluyente pues hay algunas que no se conocen si tienen o no.

Marketing de Paginas Web

Si medimos los ingresos si tienen pagina , nos daria.

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

#Primero mapeamos si tiene como un string que tenga y si no tiene como N/A

marketing <- movies %>%
  select(homePage, revenue) %>%
  mutate(homePage = ifelse(is.na(homePage), FALSE, TRUE))
homePage_counts <- marketing %>%
  count(homePage)

print(homePage_counts)
```

```
##   homePage    n
## 1   FALSE 5807
## 2    TRUE 4193
```

```

#Ahora balanceamos
page_1 <- marketing %>% filter(homePage == TRUE)

set.seed(123) # Asegura reproducibilidad
page_FALSE_sample <- marketing %>%
  filter(homePage == FALSE) %>%
  sample_n(4193)

balanced_marketing <- bind_rows(page_1,page_FALSE_sample)

balanced_counts <- balanced_marketing %>%
  count(homePage)

print(balanced_counts)

```

```

##   homePage    n
## 1   FALSE 4193
## 2    TRUE 4193

```

```

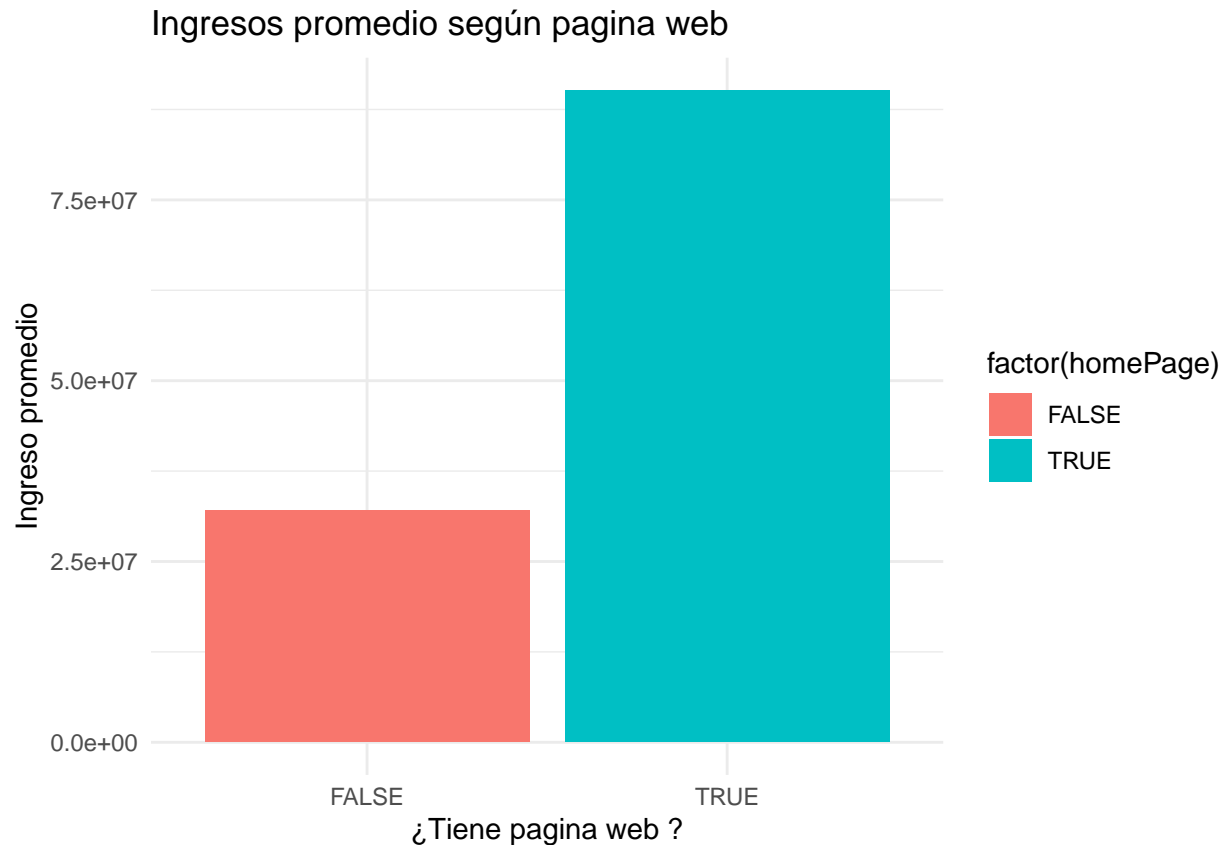
#Ahora graficamos

```

```

balanced_marketing %>%
  group_by(homePage) %>%
  summarise(avg_revenue = mean(revenue, na.rm = TRUE)) %>%
  ggplot(aes(x = factor(homePage), y = avg_revenue, fill = factor(homePage))) +
  geom_col() +
  labs(title = "Ingresos promedio según pagina web",
        x = "¿Tiene pagina web ?",
        y = "Ingreso promedio") +
  theme_minimal()

```

Analisis Grafico Podemos ver aqui que si tienen pagina web tiene una mayor cantidad de ingresos que al no tener una pagina web, siendo a simple vista casi el doble. Muy posiblemente debido a que internet suele ser una manera muy efectiva para promocionarse.

Si medimos la popularidad si tienen pagina , nos daria.

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

#Primero mapeamos si tiene como un string que tenga y si no tiene como N/A

marketing <- movies %>%
  select(homePage, popularity) %>%
  mutate(homePage = ifelse(is.na(homePage), FALSE, TRUE))
homePage_counts <- marketing %>%
  count(homePage)

print(homePage_counts)
```

```
##   homePage    n
## 1   FALSE 5807
## 2    TRUE 4193
```

```
#Ahora balanceamos
page_1 <- marketing %>% filter(homePage == TRUE)
```

```

set.seed(123) # Asegura reproducibilidad
page_FALSE_sample <- marketing %>%
  filter(homePage == FALSE) %>%
  sample_n(4193)

balanced_marketing <- bind_rows(page_1,page_FALSE_sample)

balanced_counts <- balanced_marketing %>%
  count(homePage)

print(balanced_counts)

```

```

##   homePage    n
## 1   FALSE 4193
## 2    TRUE 4193

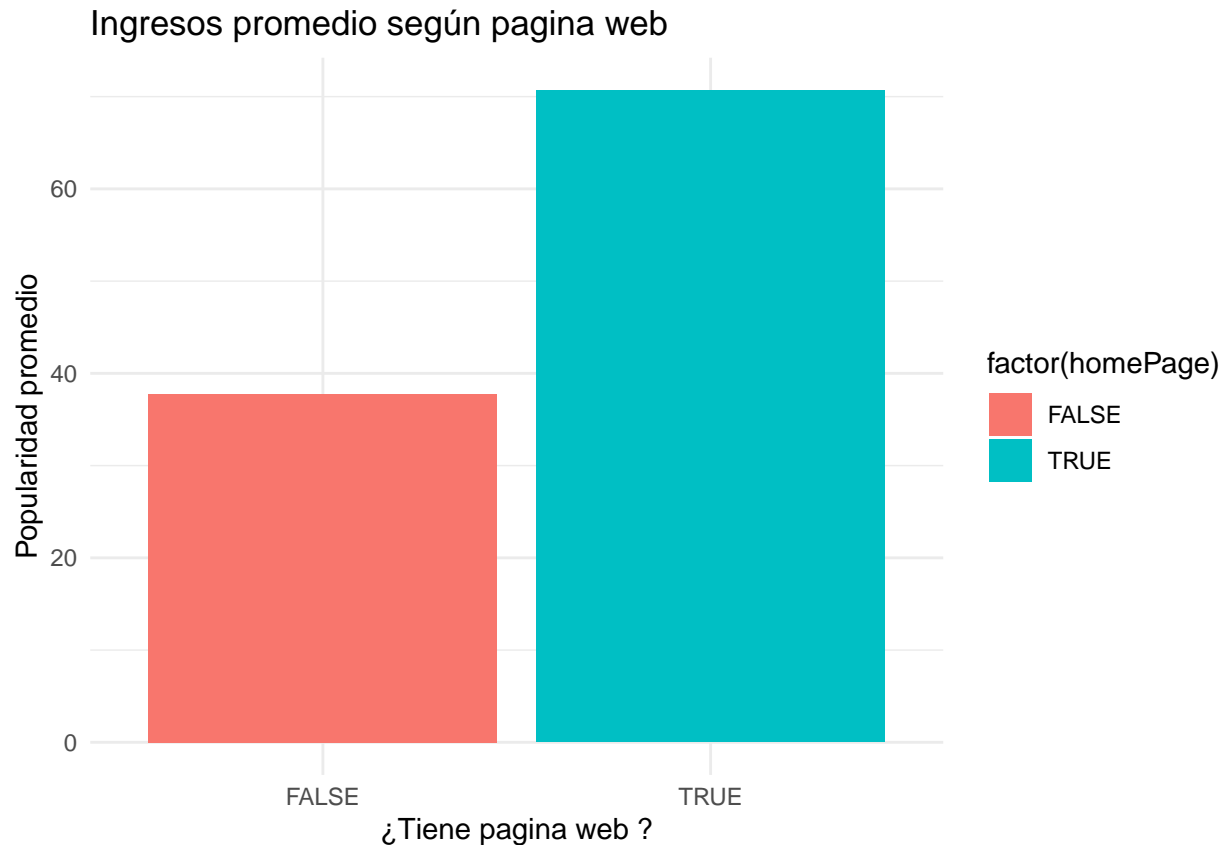
```

#Ahora graficamos

```

balanced_marketing %>%
  group_by(homePage) %>%
  summarise(avg_popularity = mean(popularity, na.rm = TRUE)) %>%
  ggplot(aes(x = factor(homePage), y = avg_popularity, fill = factor(homePage))) +
  geom_col() +
  labs(title = "Ingresos promedio según pagina web",
       x = "¿Tiene pagina web ?",
       y = "Popularidad promedio") +
  theme_minimal()

```



Análisis Gráfico Podemos ver aquí que si llega a ser muy popular siguiendo la misma proporción que los ingresos

Conclusion Podemos ver que la mejor estrategia va a ser tener una página web, sin embargo se recomienda también tener un video, y colocarlo en las páginas web para poder apoyar a promocionar las películas, ya que el video si bien no rinde en ingresos si rinde en popularidad siendo mayor que no tenerlo.

4.16 ¿La popularidad del elenco está directamente correlacionada con el éxito de taquilla?

Para ello podemos hacer una gráfica de correlación.

```
# Leer el archivo CSV
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

# Seleccionar las columnas de interés
popularity <- movies %>%
  select(actorsPopularity, revenue, budget)

# Convertir los valores de actorsPopularity que están separados por "/" en un vector
↳ numérico

popularity$actorsPopularity <- sapply(popularity$actorsPopularity, function(x) {
  # Convertir la cadena separada por "/" en un vector numérico
})
```

```

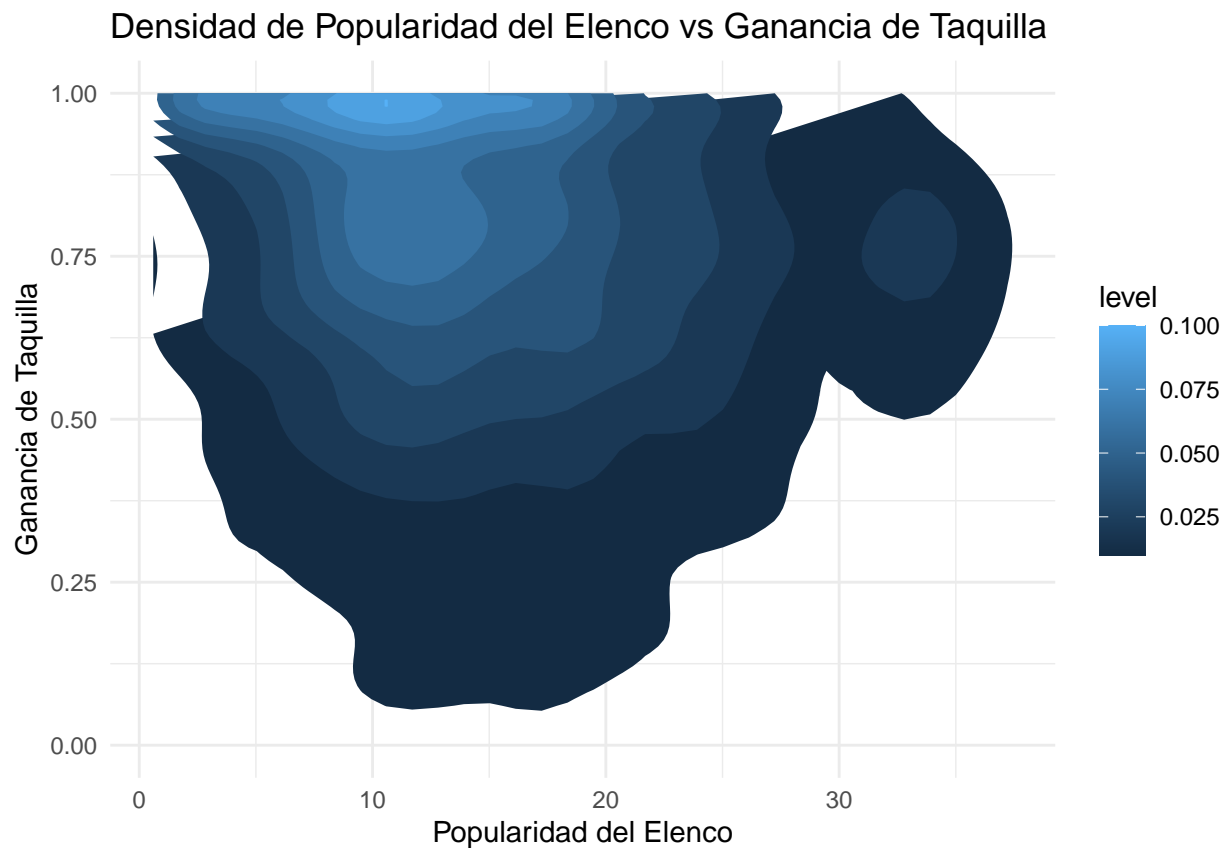
values <- as.numeric(unlist(strsplit(x, "\\|")))
# Retornar el primer valor de la lista como ejemplo (o puedes hacer un promedio si lo
↪ prefieres)
return(values[1]) # O usa mean(values) si quieres promediar los valores
})

# Calcular la ganancia de taquilla
popularity$ganancia <- (popularity$revenue - popularity$budget) / popularity$revenue

# Filtrar los valores que sean finitos para evitar NaN o infinitos
popularity <- popularity[is.finite(popularity$ganancia), ]

ggplot(popularity, aes(x = actorsPopularity, y = ganancia, fill = ..level..)) +
  stat_density_2d(geom = "polygon") +
  labs(title = "Densidad de Popularidad del Elenco vs Ganancia de Taquilla",
        x = "Popularidad del Elenco",
        y = "Ganancia de Taquilla") +
  ylim(0, max(popularity$ganancia, na.rm = TRUE)) + # Limita Y de 0 al máximo
  theme_minimal()

```



Análisis Gráfico Como podemos ver este gráfico la densidad de los actores están dentro de una ganancia muy alta y una popularidad del elenco siendo entre 10 a 20. Por lo que se puede intuir en la gráfica es que a

mientras mas popularidad tenga un elenco puede existir una mayor ganancia pero la mayoría de las películas suelen estar en el rango de una popularidad no tan alta y aun así lograr éxitos en la taquilla.

Conclusiones. Si bien la popularidad influye en las ganancias de una película que sean buenas, no significa que sea el único factor, debido a que ha habido cintas con menor cantidad de actores populares que han logrado ser muy bien recibidas por el público.

5 Genere usted otras seis preguntas que le parezcan interesantes.

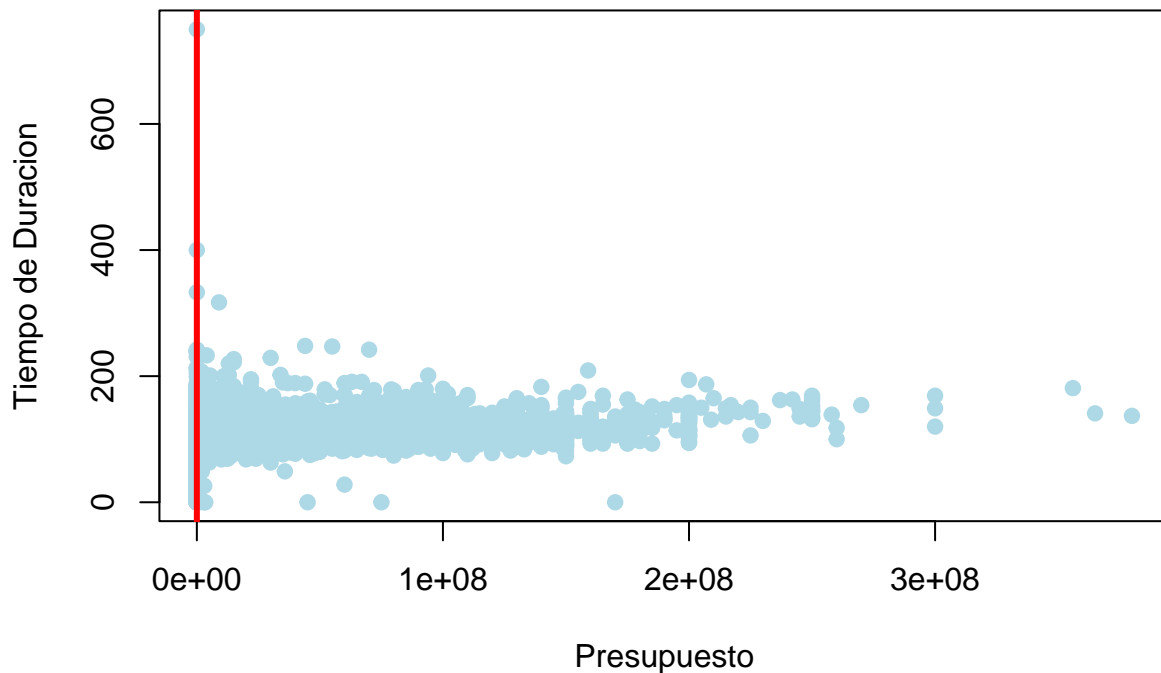
5.1 ¿Cuál es la relación entre el presupuesto y la duración de las películas? ¿Las películas con mayor presupuesto tienden a ser más largas?

Para ello creamos una gráfica de correlación entre la duración y el presupuesto.

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

plot(x= movies$budget, y = movies$runtime, pch = 19, col = "lightblue",
     xlab = "Presupuesto", # Título del eje X
     ylab = "Tiempo de Duracion",
     )

abline(lm(y ~ x), col = "red", lwd = 3)
```



Análisis grafico Como se puede ver la mayoría de películas con un tiempo de duración muy alto suelen tener un presupuesto muy bajo, casi siendo 0. Pero se nota una cierta tendencia a que las películas con mayor presupuesto su tiempo de duración no sea tan largo. Al menos no más de los 200 minutos.

Conclusion Las películas con mayor presupuesto no tienden a ser más largas al contrario siempre siguen la tendencia de durar menos de 200 minutos, la relación entre estas 2 variables no es evidente.

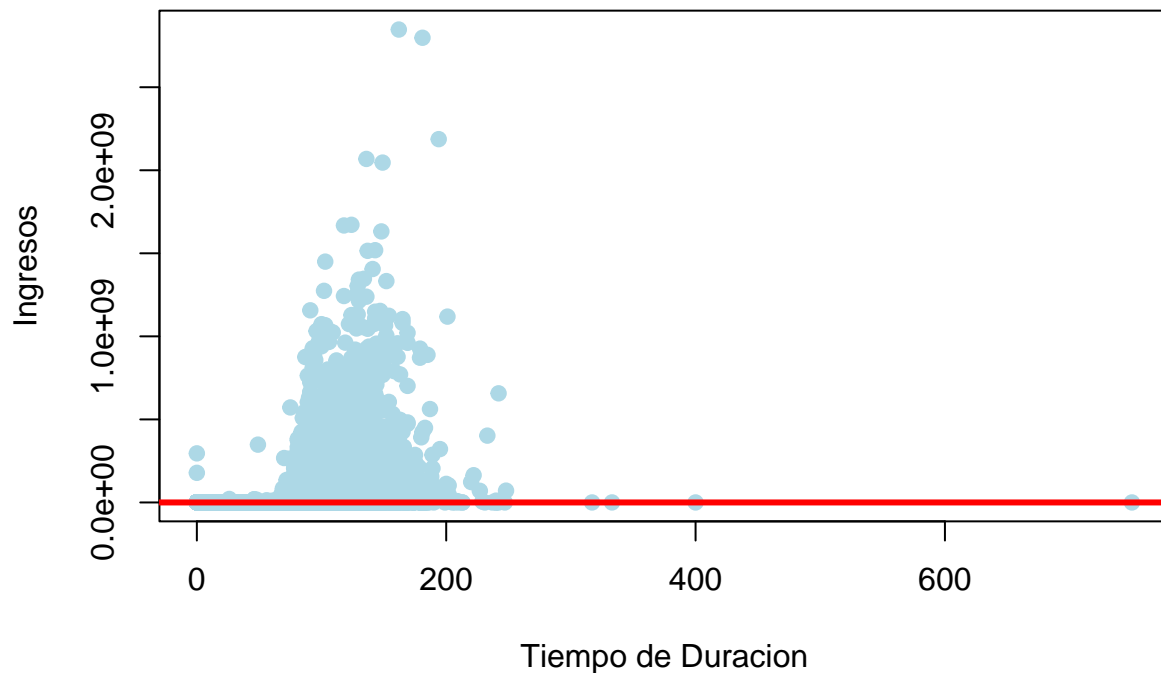
5.2 ¿Las películas más largas tienden a tener mayores ingresos?

Para ello realizamos otra gráfica

```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

plot(x= movies$runtime, y = movies$revenue, pch = 19, col = "lightblue",
     xlab = "Tiempo de Duracion", # Título del eje X
     ylab = "Ingresos",
)

abline(lm(y ~ x), col = "red", lwd = 3)
```



Análisis Grafico Como se puede ver las películas más largas a los 50 minutos si alcanzan una gran cantidad de ingresos, pero también llega a un máximo y a partir de ahí mientras más larga la película menos ingresos genera

Conclusion Las personas si bien estan dispuestas a consumir peliculas larga, llega a un punto en donde la misma llega a perder retencion, debido a que no genera los ingresos necesarios. Por ende si se espera tener un buen rendimiento es necesario no sobrepasar por los 150 minutos .

5.3 ¿Cómo ha evolucionado la cantidad de películas por género a lo largo de los años?

Calcularemos las peliculas por genero. Y las agruparemos por año

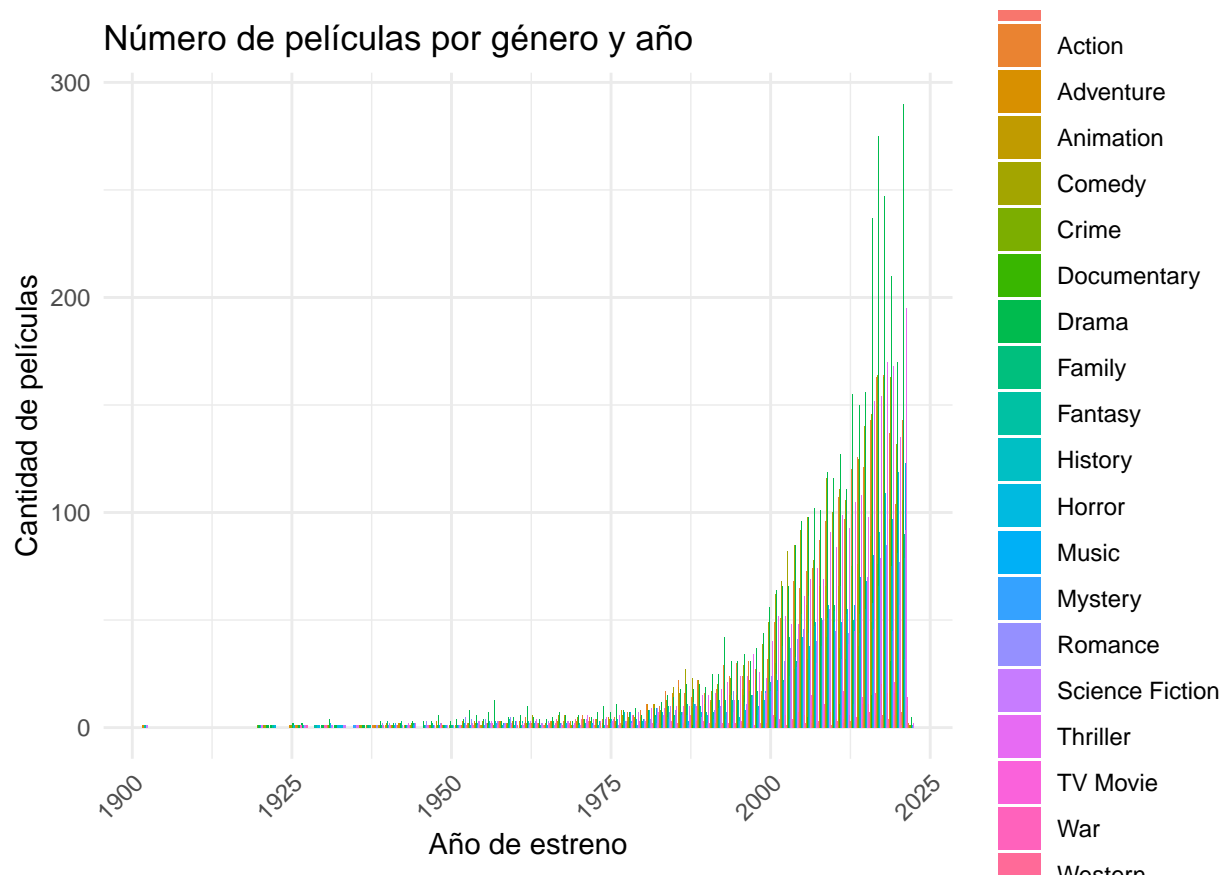
```
movies <- read.csv("movies.csv", stringsAsFactors = FALSE)

movies$releaseYear <- as.numeric(substr(movies$releaseDate, 1, 4))

movies_por_anio <- movies %>%
  select(title, releaseYear, genres)

movies_por_anio <- movies_por_anio %>%
  separate_rows(genres, sep = "\\|") %>% # Divide en varias filas
  group_by(genres, releaseYear) %>%      # Agrupa por género y año
  summarise(count = n(), .groups = "drop") # Cuenta las películas

ggplot(movies_por_anio, aes(x = releaseYear, y = count, fill = genres)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Número de películas por género y año",
       x = "Año de estreno",
       y = "Cantidad de películas") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Análisis Gráfico Se puede ver que a lo largo de todos los años todos los generos han tenido mayor cantidad de peliculas, siendo en el 2020 al 2025 donde se ha visto esta enorme tendencia.

Conclusion Como se puede ver la cantidad de generos ha aumentado sobre todo aquellos que son documentales o que tratan sobre crimen y drama. A pesar de ello se ha visto una disminucion de producciones el año 2025 y 2024.

5.4 ¿Cuales fueron los paises que se rodaron filmes con mayor rendimiento?

Calcularemos la ganancia generada y luego la contrastaremos en un grafico de barras con respecto a los paises.

```
movies_por_pais <- movies %>%
  select(title, revenue, budget, productionCountry)

# Calculamos la ganancia
movies_por_pais$ganancia <- (movies_por_pais$revenue - movies_por_pais$budget) /
  ↪ movies_por_pais$revenue

# Filtramos valores finitos
movies_por_pais <- movies_por_pais[is.finite(movies_por_pais$ganancia), ]
```



```

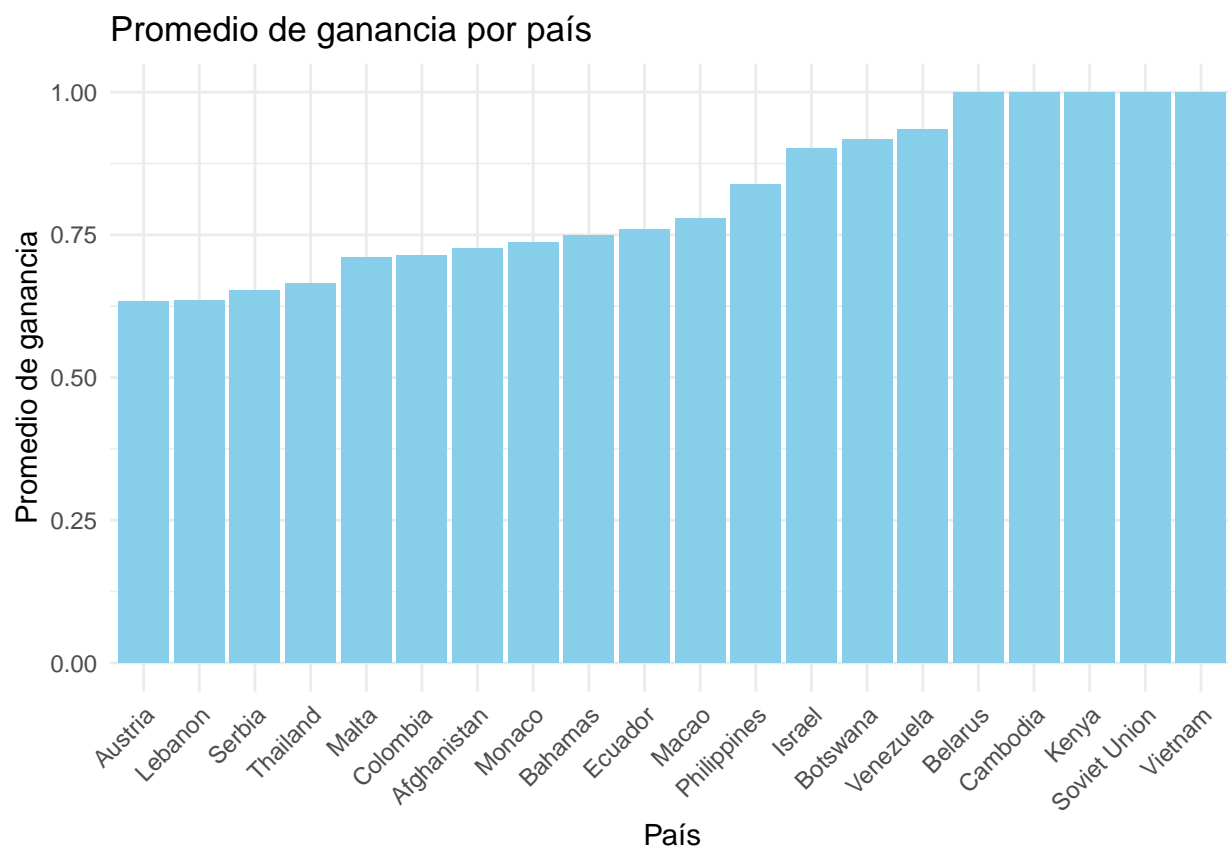
movies_por_pais <- movies_por_pais %>%
  separate_rows(productionCountry, sep = "\\|")

movies_por_pais_promedio <- movies_por_pais %>%
  group_by(productionCountry) %>%
  summarise(promedio_ganancia = mean(ganancia), .groups = "drop")

top_20_paises <- movies_por_pais_promedio %>%
  arrange(desc(promedio_ganancia)) %>%
  head(20)

# Crear gráfico
ggplot(top_20_paises, aes(x = reorder(productionCountry, promedio_ganancia), y =
  promedio_ganancia)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Promedio de ganancia por país",
       x = "País",
       y = "Promedio de ganancia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotar etiquetas del eje X

```



Analisis Grafico Como se puede observar los paises con mayor rendimiento son desde Australia a Vietnam y han tenido ganancias del 0.6 al 100% como ganancia neta con respecto a su presupuesto invertido

Conclusion En conclusion si se desea un promedio de ganancias bueno es mejor filmar en estos paises, debido a que puede ser que los costos no se eleven tanto , y puede ayudar a generar peliculas que sean mas entretenidas y llamativas al publico al conocer lugares que no han visto.

5.5 ¿Cual son los actores que mas recaudacion ha tenido?

Para ello realizaremos una tabla

```
# Seleccionamos las columnas relevantes
movies_por_actor <- movies %>%
  select(actors, revenue)

# Limpiamos posibles caracteres no estándar en los nombres de los actores
movies_por_actor$actors <- iconv(movies_por_actor$actors, from = "UTF-8", to =
  ↪ "ASCII//TRANSLIT")

# Separamos a los actores en filas
movies_por_actor <- movies_por_actor %>%
  separate_rows(actors, sep = "\\|")

# Calculamos el promedio de ingresos por actor
movies_por_actor_promedio <- movies_por_actor %>%
  group_by(actors) %>%
  summarise(promedio_ingresos = mean(revenue, na.rm = TRUE), .groups = "drop")

# Ordenamos y seleccionamos los 20 actores con mayor promedio de ingresos
top_20_actores <- movies_por_actor_promedio %>%
  arrange(desc(promedio_ingresos)) %>%
  head(20)

# Imprimir los primeros 20 actores
knitr::kable(top_20_actores, format= "html")
```

actors	promedio_ingresos
David Van Horn	2847246203
James Patrick Pitt	2847246203
Jason Whyte	2847246203
Jon Curry	2847246203
Kyla Warren	2847246203
Kyle Dryberg	

2847246203
Peter Dillon
2847246203
Sean Anthony Moran
2847246203
Sonia Yee
2847246203
Ameenah Kaplan
2046239637
Ethan Dizon
2046239637
Florence Kasumba
2046239637
Harrison Osterfield
2046239637
Laura Miller
2046239637
Letitia Wright
2046239637
Marie Mouroum
2046239637
Michael Anthony Rogers
2046239637
Michael James Shaw
2046239637
Olaniyan Thurmon
2046239637
Ross Marquand
2046239637

Conclusion Como se puede observar estos son los actores que mas ingresos si bien no han generado en promedio sus peliculas en las que han estado han tenido muchos ingresos. Ello puede servir para poder elegir que elenco de celebridades elegir para una nueva cinta.

5.6 ¿Que idiomas de las peliculas son los mas consumidos (con mayor ingresos)?

```

movies_por_lengua <- movies %>%
  select(originalLanguage, revenue )

# Calculamos el promedio de ingresos por lengua
movies_por_lengua_promedio <- movies_por_lengua %>%
  group_by(originalLanguage) %>%
  summarise(promedio_ingresos = mean(revenue, na.rm = TRUE), .groups = "drop")

top_20_lenguas <- movies_por_lengua_promedio %>%
  arrange(desc(promedio_ingresos)) %>%
  head(20)

knitr::kable(top_20_lenguas, format= "html")

```

originalLanguage
promedio_ingresos
zh
74864948
te
73338967
en
69581771
ar
64978931
hi
26536670
cn
19017166
ko
13487526
fr
11139464
fa
10768392
ja
9509624
sv
9112728

de
6766426
ru
5990880
it
5091298
nb
4159678
th
3638706
pt
3383692
tl
3360000
es
3344984
cs
3319425

Conclusion Como se puede ver los idiomas de las peliculas que mas se consumen son el chino, el indi, el telugu , coreano e ingles. Esto se debe muy posiblemente a que son los paises con mayor consumo de peliculas , y ademas hay 2 paises siendo la India y China que tienen una gran poblacion de personas , lo que hace que las peliculas estrenadas ahi consigan muchos ingresos.