Entrega 2

Gabriela Aldana, Valeria Arce, Juan Esteban Torres
Adriana Pineda
Probabilidad y Estadistica 2
24/05/2025

Marco metodológico y objetivo:

Full IMDb Movies DatLa:

Para el análisis se utilizó el conjunto de datos Full IMDb Movies Data, (Internet Movie Database). La información es recolectada a partir de contribuciones de usuarios, estudios de cine y fuentes oficiales, asegurando una base de datos amplia y actualizada. El conjunto de datos cuenta con 1048541 registros y 21 variables, incluyendo tanto categóricas (género, idioma, clasificación por edad, entre otras) como numéricas (presupuesto, recaudación, duración, calificación, etc.) Se realizó una selección de variables relevantes enfocada en aquellas con mayor impacto potencial en el éxito financiero y popularidad de las películas.

El objetivo de este análisis es identificar los factores clave que influyen en la popularidad y el éxito financiero de las películas en IMDb, considerando variables como el género, la valoración del público, el presupuesto y la recaudación, con el fin de comprender tendencias y patrones relevantes en la industria cinematográfica.

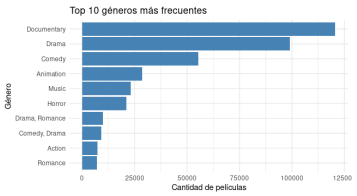
El análisis se estructuró en varias etapas. Primero, se realizó un análisis descriptivo univariado y bivariado para examinar el comportamiento de variables individuales (categóricas y numéricas) y detectar patrones generales. Se utilizaron gráficos estadísticos (barras, treemaps, histogramas, boxplots, dispersión) para visualizar relaciones y tendencias relevantes, especialmente en torno a género, idioma, clasificación por edad, presupuesto, recaudación y calificación promedio. En la fase inferencial, se aplicaron pruebas de hipótesis como la t de una muestra (comparando la media de calificación con un valor hipotético), proporciones (para evaluar la distribución de películas para adultos) y t de dos muestras (para comparar ingresos entre películas para adultos y no adultas). También se evaluó la normalidad de variables clave (calificación, ingresos, presupuesto, popularidad) mediante la prueba de Kolmogorov-Smirnov, asimetría y gráficos de densidad; en casos de alta asimetría, se aplicaron transformaciones logarítmicas para mejorar la validez del análisis. Finalmente, se exploraron relaciones entre variables cuantitativas y categóricas, como la recaudación media por idioma, presupuesto por género y calificación según clasificación por edad.

Análisis descriptivo de variables categóricas:

- Análisis descriptivo de variables categóricas:
  - Univariado:

	genres	popularidad_media	votos_media	revenue_media	presupuesto_media
1	Action	6.378005	5.822976	11077456	4161306.4
2	Adventure	7.194411	6.100763	20016521	6626223.3
3	Comedy	3.189137	5.890226	3285491	1170168.7
4	Drama	3.107950	5.986937	2053171	871872.3
5	Horror	4.259944	5.380656	1921775	740914.2

Al analizar las gráficas propuestas, se pueden identificar algunas tendencias claras en los usuarios. Por ejemplo, los géneros más comunes son Documentary, Drama y Comedy, destacándose claramente sobre otros géneros menos frecuentes.

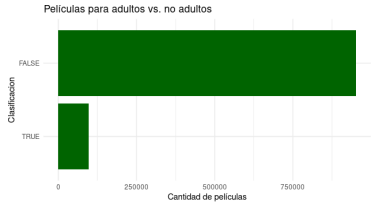


	genres	n
1	Documentary	120668
2	Drama	98981
3	Comedy	55356
4	Animation	28577
5	Music	23134
6	Horror	21114
7	Drama, Romance	9971
8	Comedy, Drama	9030
9	Action	7439
10	Romance	7062

(gráfica 1, tabla 1)

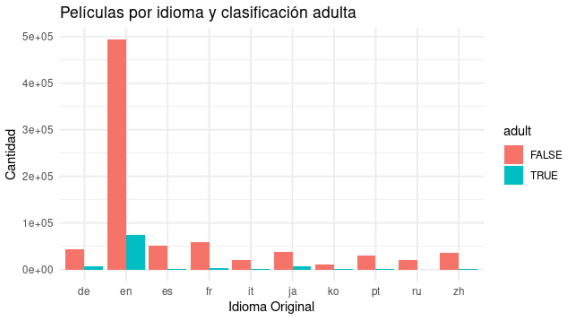
En cuanto al idioma original, el inglés domina en todo el sentido, mientras que idiomas como el francés, español y alemán quedan no tanto.

Por otro lado, se puede observar que la mayoría de las películas no están etiquetadas como contenido para adultos, lo cual probablemente influye en que puedan llegar a una audiencia más amplia.



- <u>Bivariado:</u>

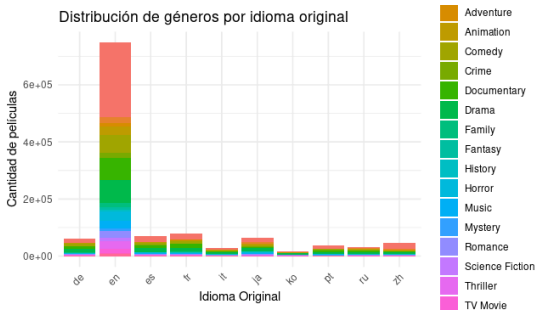
Al analizar las gráficas relacionadas con los datos se identifican patrones interesantes sobre cómo se distribuyen y clasifican las películas. Por ejemplo, en el gráfico de barras agrupadas se observa que la mayoría de los idiomas tienen un predominio de películas no clasificadas como contenido adulto. Esto ocurre especialmente en inglés, aunque este idioma también concentra el mayor número de películas para adultos.

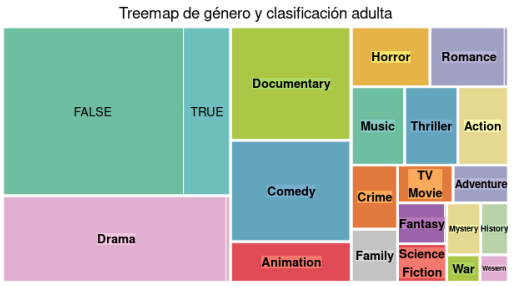


(gráfica 2)

En el gráfico de barras apiladas se muestran los géneros como Drama y Documentary se encuentran en una variedad de idiomas, mientras que otros como TV Movie o Western aparecen casi exclusivamente en inglés. Esto podría reflejar diferencias culturales o enfoques de mercado más específicos para los estudios.

En el treemap podemos notar algo diferente donde se muestra que géneros como Animation, Comedy y Family están mayoritariamente asociados con contenido apto para todo público. En cambio, Horror, Crime y Thriller tienden a incluir una mayor proporción de películas clasificadas como

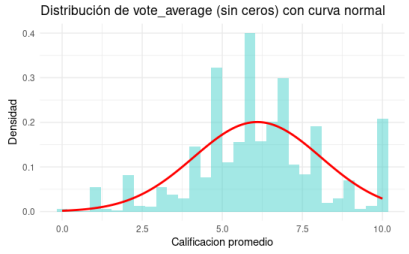




adultas, ayudando a visualizar qué género llega a ser más 'apto' para el público.

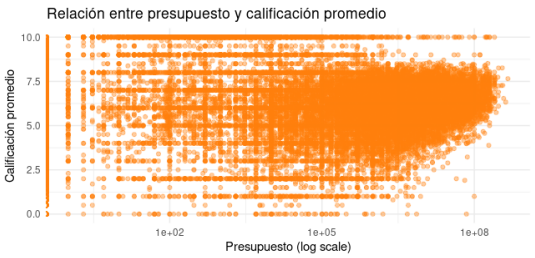
- Análisis descriptivo de variables cuantitativas (univariado y bivariado):

El histograma de calificación promedio, sin contar los ceros, muestra una concentración entre 5 y 7. Aunque se incluye una curva normal, la forma real no es simétrica, apoyando la conclusión de que los datos no siguen una distribución normal.



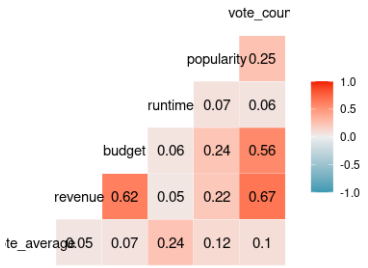
Por otro lado, el gráfico de dispersión entre la calificación promedio y la recaudación muestra una nube de puntos concentrada en calificaciones medias entre 4 y 7, con tendencia a tomar una forma ascendente.

Al observar la relación entre el presupuesto y la calificación promedio, usando una escala logarítmica, vemos que las películas con presupuestos más altos tienden a alcanzar



calificaciones de 6 a 8. Sin embargo, también hay varias producciones con presupuestos elevados que obtienen calificaciones bajas.

Finalmente, la matriz de correlación nos confirma que la recaudación está moderadamente correlacionada con el presupuesto (0.62) y con la popularidad (0.67). En contraste, la calificación promedio tiene una baja correlación con las demás variables numéricas, indicando que el éxito financiero de una película no necesariamente está relacionado con cómo es percibida por el público.

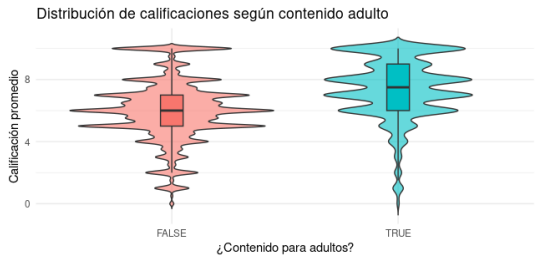


En cuanto a la distribución de las variables, los resultados sugieren que ninguna de las variables sigue una distribución normal en su totalidad.

- Análisis descriptivo bivariado entre variables categóricas y cuantitativas:

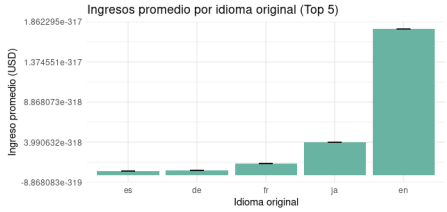
Se presentan tres gráficos que permiten analizar variables numéricas segmentadas por

categorías. El primero es un boxplot del presupuesto según los cinco géneros más comunes, donde se observa una gran variabilidad, con algunos géneros concentrando presupuestos bajos y otros

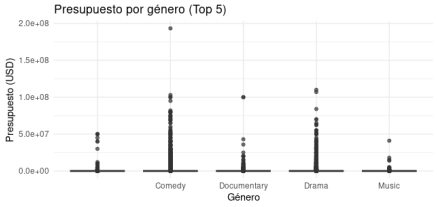


mostrando valores extremos.

El segundo gráfico muestra la recaudación por idioma original, destacando una clara concentración en el idioma inglés, aunque también se observan casos aislados en otros idiomas con ingresos significativos.



El tercero representa la calificación promedio según si el contenido es para adultos o no. La mayoría de películas no son para adultos y presentan calificaciones moderadas, mientras que las de contenido adulto tienen mayor dispersión.



- <u>Pruebas de hipótesis:</u>

Se realizaron dos pruebas de hipótesis con el objetivo de contrastar supuestos sobre las características generales del conjunto de datos.

La primera fue una prueba t para una muestra, aplicada a la variable vote_average.

- Hipótesis nula (H_0): la media de las calificaciones es igual a 7.
- Hipótesis alternativa (H_1): la media de las calificaciones es distinta de 7.

```
        One Sample t-test

data:  datos$vote_average
t = -1639.3, df = 1048574, p-value < 2.2e-16
alternative hypothesis: true mean is less than 7
95 percent confidence interval:
     -Inf 2.046535
sample estimates:
mean of x
  2.04156
```

Esta hipótesis se propuso considerando que una puntuación de 7 suele reflejar una valoración positiva. Sin embargo, la media observada fue de 2.04 y el p-valor fue menor a 0.001, por lo que se rechaza la hipótesis nula, indicando una diferencia significativa respecto al valor esperado. Esto sugiere que, en promedio, las películas tienen calificaciones notablemente más bajas.

La segunda fue una prueba de proporciones, enfocada en la variable is_adult.

- Hipótesis nula (H_0): la proporción de películas para adultos es igual o menor al 5%.
- Hipótesis alternativa (H_1): la proporción de películas para adultos es mayor al 5%.

```
        1-sample proportions test with continuity correction

data:  n_adultos out of n_total, null probability 0.05
X-squared = 39921, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is greater than 0.05
95 percent confidence interval:
 0.09206069 1.00000000
sample estimates:
         p
0.09252557
```

Este análisis buscaba evaluar si el contenido adulto era marginal. Los resultados mostraron una proporción observada de 9.25% y un p-valor menor a 0.001, por lo que se rechaza la hipótesis nula, concluyendo que la proporción de películas para adultos es significativamente mayor al 5%.

- **Análisis de normalidad:**

Para el análisis de normalidad utilizamos las variables continuas: vote_average, revenue popularity y budget que fueron consideradas a partir del objetivo principal. Este proceso incluyó la aplicación de la prueba de Kolmogorov-Smirnov, el examen de los estadísticos de asimetría, y la visualización mediante gráficos de densidad para verificar su conformidad con una distribución normal.

Estos resultados indican que aparentemente la variable vote_average no sigue una distribución normal, ya que se rechaza la hipótesis nula con un p-value extremadamente pequeño.

- **Revenue**

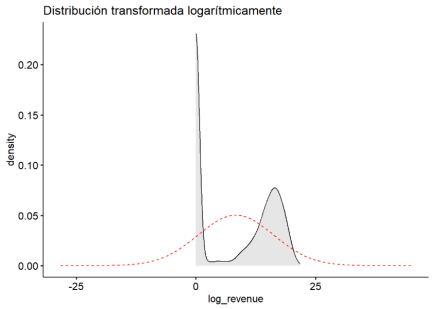
Tras aplicar la prueba de Kolmogorov-Smirnov a la variable revenue, se obtuvieron los siguientes resultados:

```
  Asymptotic one-sample Kolmogorov-Smirnov test

data:  datos_limpios$revenue
D = 0.39331, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Estos resultados indican que aparentemente la variable revenue no sigue una distribución normal, ya que se rechaza la hipótesis nula con un p-value extremadamente pequeño.

Adicionalmente, el coeficiente de asimetría (skewness) arrojó un valor de **8.382533.** Esto indica una alta inclinación de la distribución hacia la cola derecha indicando una una alta asimetría notable. Debido a esta alta asimetría, se optó por hacer una transformación logarítmica. Tras la transformación, el nuevo valor de asimetría fue de **0.03466874**, implicando a una distribución más simétrica que se puede observar en la gráfica de densidad.



En el gráfico de densidad de la variable ya transformada presenta una distribución bimodal. Esta forma sugiere la existencia de dos subgrupos distintos de películas en términos de ingresos, uno con recaudaciones muy bajas y otro con recaudaciones considerablemente mayores. A pesar de aplicar una transformación logarítmica para corregir la asimetría inicial, la distribución resultante tampoco se ajusta totalmente a una distribución normal.

- **Budget**

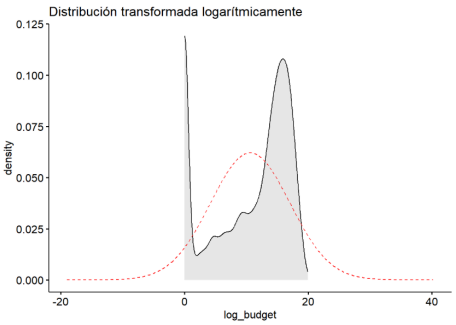
Tras aplicar la prueba de Kolmogorov-Smirnov a la variable budget, se obtuvieron los siguientes resultados:

```
Asymptotic one-sample Kolmogorov-Smirnov test

data:  datos_limpios$budget
D = 0.34659, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Estos resultados indican que aparentemente la variable popularity no sigue una distribución normal, ya que se rechaza la hipótesis nula con un p-value extremadamente pequeño.

Adicionalmente, el coeficiente de asimetría (skewness) arrojó un valor de **4.665613**, indicando una asimetría positiva considerable. Debido a esto, se optó por aplicar una transformación logarítmica. Como resultado, el nuevo coeficiente de asimetría dio como resultado **-0.6100768**. Si bien este valor representa una reducción significativa de la asimetría original, ahora sugiere una ligera asimetría negativa, lo cual es observable en el gráfico de densidad.

Tras 

la transformación logarítmica de la variable, la gráfica nos muestra una bimodalidad. Se identifica una alta concentración de valores próximos a cero, lo cual nos da a entender a una posible tendencia a presupuestos mínimos o nulos, por otro lado existe un segundo pico entre los valores 18 y 20. Esto nos sugiere dos grupos de películas por nivel de presupuesto. A pesar de hacer la transformación, la variable transformada no se ajusta totalmente a la distribución normal.

- **<u>Prueba de hipótesis (use α = 0.05)</u>**

<u>Revenue y Adults:</u>

Para la prueba de hipótesis, se investiga si la clasificación de contenido de una película se relaciona con sus ingresos. Con esto en mente se seleccionaron la variable escalar revenue (transformada logarítmicamente) y la categórica binaria adult que destina si una película es destinada a un público adulto o no ('TRUE' para adultos y 'FALSE' para no adultos). El objetivo es determinar si existe una diferencia estadísticamente significativa en la media de los ingresos entre estos dos grupos de películas.

Se realizaron las siguientes pruebas de hipótesis con α = 0.05:

- **Hipótesis Nula (H0)**: La media de revenue es igual para el grupo de películas clasificadas como **adult = TRUE** y el grupo de películas clasificadas como **adult = FALSE.**

  H0: µadult = TRUE = µadult = FALSE

- **Hipótesis Alternativa (H1)**: La media de revenue es diferente entre el grupo de películas clasificadas como **adult = TRUE** y el grupo de películas clasificadas como **adult = FALSE**.

H1: µadult = TRUE != µadult = FALSE

Antes de realizar la t de Welch para la comparación de medias, se procedió a verificar los supuestos que garantizan la validez y fiabilidad de sus resultados.

<u>Independencia de las Muestras:</u>

Se asume que los dos grupos que son comparados, películas clasificadas para adultos(adult = TRUE) y las que se clasifican para no adultos(adult = FALSE) son independientes, esto ya que la misma naturaleza de la variable adult lo señala. Una película se clasifica en una única categoría, excluyendo su pertenencia simultánea a ambas.

Normalidad:

El supuesto de normalidad establece que la variable escalar, en nuestro caso revenue, debe seguir una distribución aproximadamente normal dentro de cada uno de los grupos definidos por la variable categórica adult. Para evaluar este supuesto, se examinó visualmente la distribución de la variable escalar para cada categoría mediante gráficos de densidad.

El gráfico muestra la distribución de revenue para películas adultas y no adultas. El grupo no adulto (adult = FALSE) presenta una distribución más unimodal, mientras que el grupo adulto (adult = TRUE) muestra una forma más irregular y dispersa. Ambas distribuciones difieren visualmente de la normalidad, sin embargo se sigue optando por t de Welch ya que el tamaño de la muestra es mayor a 30.

Igualdad de Varianzas:

Para validar la igualdad de varianzas se requiere de que las varianzas de la variable escalar sean iguales o muy similares entre los dos grupos, adult = TRUE y adult = FALSE. Para verificar este supuesto, se aplicó la prueba correspondiente

```
 F test to compare two variances

data:  log_revenue by adult
F = 2.5095, num df = 69355, denom df = 214, p-value = 2.22e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.05605 3.00776
sample estimates:
ratio of variances
          2.509496
```

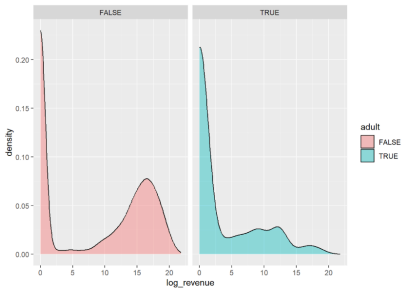
Al hacer la prueba F para comparar varianzas indica un estadístico F = 2.51 con un valor p extremadamente pequeño ($p < 0.001$) nos lleva a rechazar la hipótesis nula de igualdad de varianzas entre los grupos. Esto sugiere que las varianzas de revenue son significativamente diferentes entre ambos grupos,por lo cual optamos por usar la prueba t de Welch.

Después de analizar los supuestos y optar por t de Welch, se realizó la prueba con estos resultados:

```
 Welch Two Sample t-test

data:  log_revenue by adult
t = 15.744, df = 217.34, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 4.709977 6.058008
sample estimates:
mean in group FALSE  mean in group TRUE
           8.356081            2.972089
```

Con los resultados de la prueba podemos decir que existe evidencia estadística suficiente para concluir que la media de revenue difiere significativamente entre las películas



clasificadas para adultos y las que no lo son. Las películas no adultas (adult = FALSE), con una media de revenue de 8.356, presentan ingresos promedio mayores que las películas para adultos (adult = TRUE), cuya media es de 2.972. Rechazamos H0

Pruebas ANOVAS:

En esta parte queremos explorar las relaciones entre variables numéricas y conjuntos de variables categóricas binarizadas mediante el Análisis de Varianza (ANOVA). El objetivo es

determinar si la pertenencia a ciertas categorías como géneros o idiomas tiene un efecto estadísticamente significativo en la media de una variable de interés como popularidad o presupuesto.

Popularidad y géneros:

Se busca evaluar si existe una diferencia estadísticamente significativa en la media de la variable popularity de las películas en función de su pertenencia a uno o más de los siguientes géneros: Drama, Comedia, Thriller, Acción y Horror.

Para el modelo ANOVA se tienen las siguientes hipótesis:

- **Hipótesis Nula (H0):**

Ninguno de los géneros considerados tiene un efecto significativo sobre la media de la popularidad.

H0: βDrama = βComedy = βThriller = βAction = βHorror =0

- **Hipótesis Alternativa (H1):**

Para el modelo general: Al menos uno de los géneros considerados tiene un efecto significativo sobre la media de la popularidad.

H1: Al menos un βj != 0

Para realizar este análisis, la variable categórica genres, que puede contener múltiples valores por película, se transformó en un conjunto de variables binarias. Cada nueva variable representa uno de los géneros que nos interesa. Drama, Comedia, Thriller, Acción u Horror, tomando el valor 1 si la película pertenece a ese género y 0 si no. Esto permite evaluar el efecto individual de la pertenencia a cada uno de estos géneros.

La prueba ANOVA fue realizada en R y como

```
              Df    Sum Sq Mean Sq F value   Pr(>F)
Drama          1    156269  156269  83.426  < 2e-16 ***
Comedy         1    113007  113007  60.330 8.13e-15 ***
Thriller       1      1594    1594   0.851    0.356
Action         1    138167  138167  73.762  < 2e-16 ***
Horror         1     36769   36769  19.630 9.41e-06 ***
Residuals  69566 130307181    1873
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

resultado nos arrojó lo siguiente:

El análisis ANOVA para la variable popularity y los géneros seleccionados mostró que tanto Drama, Comedy, Action y Horror influyen significativamente en la popularidad (todos p < 0.001). Por otro lado, Thriller no presentó un efecto significativo (p = 0.356). Estos resultados sugieren que ciertos géneros están asociados con diferentes niveles de popularidad.

Presupuesto e idioma

Se busca evaluar si existe una diferencia estadísticamente significativa en la media de la variable budget de las películas en función de su idioma original, considerando los siguientes idiomas: inglés (en), español (es), francés (fr), japonés (ja) y ruso (ru).

Para el modelo ANOVA se tienen las siguientes hipótesis:

- **Hipótesis Nula (H0):**

Para el modelo general: Ninguno de los idiomas considerados tiene un efecto significativo sobre la media del presupuesto.

H0: βen = βes = βfr = βja = βru = 0

- **Hipótesis Alternativa (H1):**

Para el modelo general: Al menos uno de los idiomas considerados tiene un efecto significativo sobre la media del presupuesto.

H1: Al menos un βj != 0

De forma similar al análisis de géneros, la

variable original_language se transformó en un conjunto de variables binarias. Cada nueva variable representa uno de los idiomas de interés, tomando el valor 1 si la película tiene ese idioma original y 0 si no.

A continuación, se presenta la tabla resumen

```
              Df    Sum Sq   Mean Sq  F value   Pr(>F)
en             1 2.090e+18 2.090e+18 2617.080  < 2e-16 ***
es             1 1.577e+16 1.577e+16   19.750 8.84e-06 ***
fr             1 2.258e+16 2.258e+16   28.271 1.06e-07 ***
ja             1 1.023e+14 1.023e+14    0.128    0.720
ru             1 3.879e+13 3.879e+13    0.049    0.826
Residuals  69566 5.556e+19 7.987e+14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

del modelo ANOVA obtenida de R

El análisis ANOVA para la variable budget y los idiomas seleccionados indicó que tanto el inglés, español y francés influyen significativamente en el presupuesto. Por otro lado, japonés y ruso no presentaron un efecto significativo. Estos resultados sugieren que ciertos idiomas, notablemente el inglés, español y francés, están asociados con diferentes niveles de presupuesto cinematográfico.

- **Análisis de bondad de ajuste**

Se realizaron pruebas de bondad de ajuste Chi-cuadrado con el objetivo de poder evaluar si las distribuciones de frecuencias observadas para ciertas variables se ajustan a ciertas distribuciones.

Se realizó la prueba a las variables original_language y adult, esto para poder identificar si se distribuyen de manera uniforme, lo que implicaría que todas sus respectivas categorías tienen la misma frecuencia esperada.

- **Original_language:**

```
        Chi-squared test for given probabilities

data:  tabla_idioma
X-squared = 3387160, df = 105, p-value < 2.2e-16
```

Para la variable seleccionada se hizo la prueba para evaluar si podemos o no rechazar H0: Las frecuencias de los diferentes idiomas originales de las películas son uniformes.

La prueba Chi-cuadrado para la bondad de ajuste a la variable nos arrojó un resultado que permite rechazar la hipótesis nula debido a p-value casi nulo, por lo tanto, se concluye que las frecuencias de los idiomas originales son desiguales.

- **Adult:**

Para la variable seleccionada se hizo la prueba para evaluar si podemos o no rechazar H0: Las frecuencias de las distribuciones entre adulto y no adulto son uniformes.

```
        Chi-squared test for given probabilities

data:  tabla_adultos
X-squared = 68715, df = 1, p-value < 2.2e-16
```

La prueba Chi-cuadrado para la bondad de ajuste a la variable nos arrojó un resultado que permite rechazar la hipótesis nula debido a p-value casi nulo, por lo tanto, se concluye que existe un desequilibrio significativo en la cantidad de películas para adultos versus no adultos.

- **Géneros (Poisson):**

Adicionalmente, se investigó si la variable num_generos se ajusta a una distribución de Poisson. Esta variable fue creada para representar el conteo de cuántos de los cinco géneros cinematográficos (Drama, Comedia, Thriller, Acción y Horror) están asociados a cada película, a partir de la variable original genres.

Para llevar a cabo esta evaluación, se estimó

primero el parámetro λ de la distribución de Poisson mediante la media muestral de num_generos. Posteriormente, se calcularon las frecuencias esperadas bajo este modelo de Poisson y se compararon con las frecuencias

```
        Chi-squared test for given probabilities

data:  as.numeric(tabla_generos)
X-squared = 13650, df = 1, p-value < 2.2e-16
```

observadas utilizando la prueba Chi-cuadrado.

La prueba Chi-cuadrado para num_generos respecto a una distribución de Poisson arrojó un resultado que nos indica poder rechazar H0, indicando que la distribución observada de num_generos difiere de la que se esperaría bajo un modelo de Poisson.

- **<u>Análisis de independencia</u>**

Para el análisis de independencia en las variables categóricas se decidió que las variables genres, original_language y adult se agruparan, esto resultando en tres nuevas variables a analizar en pares, genres_agrupado, original_language_agrupado y adult_agrupado.

Independencia entre idioma y género

En esta parte decidimos evaluar si la clasificación de una película según su género (genres_agrupado) es independiente de su idioma original (original_language_agrupado), por lo tanto se plantearon las siguiente hipótesis:

- **Hipótesis Nula (H0):** El género agrupado de una película es independiente de su idioma original agrupado.
- **Hipótesis Alternativa (H1):** El género agrupado de una película no es independiente de su idioma original agrupado

Con esto en mente se desarrolló la prueba chi cuadrado con estas dos variables transformadas en sus agrupadas, esto dándonos el siguiente resultado:

```
        Pearson's Chi-squared test with simulated p-value (based on 10000
        replicates)

data:  tabla_contingencia_agrupada
X-squared = 7486.3, df = NA, p-value = 9.999e-05
```

Viendo los resultados llegamos a la conclusión que existe una fuerte evidencia estadística para afirmar que el género de una película (agrupado) y su idioma original (agrupado) no son independientes, por lo tanto se rechaza H0. Dado este resultado se decidió optar por hacer el análisis post-hoc de residuos.

Al hacer el análisis se decidió elegir los géneros más importantes como Drama, Comedia, Thriller, Acción y Horror, esto se decidió debido a la cantidad de datos resultantes.

- **Drama:** Este género está fuertemente sobrerrepresentado en **francés (fr)** (residual ≈ 11.0) y **español (es)** (residual ≈ 8.2). En notable contraste, muestra una significativa subrepresentación en **inglés (en)** (residual ≈ -20.5).
- **Comedia:** Presenta una sobrerrepresentación significativa en **francés (fr)** (residual ≈ 8.7) y **español (es)** (residual ≈ 6.4). Por otro lado, está subrepresentado en **inglés (en)** (residual ≈ -5.2) y de forma muy marcada en **japonés (ja)** (residual ≈ -10.2).
- **Acción:** Las películas de Acción están sobrerrepresentadas en varios idiomas asiáticos, destacando **japonés (ja)** (residual ≈ 10.4) y **chino (cn)** (residual ≈ 15.1). En contraparte, este género

está significativamente subrepresentado en **español (es)** (residual ≈ -7.9) y **francés (fr)** (residual ≈ -6.8).

- **Horror:** Este género evidencia una sobrerrepresentación extremadamente alta en **inglés (en)** (residual ≈ 22.1). Inversamente, está subrepresentado en **francés (fr)** (residual ≈ -9.5), **árabe (ar)** (residual ≈ -5.2) y **japonés (ja)** (residual ≈ -3.9).
- **Thriller:** Muestra una fuerte sobrerrepresentación en **inglés (en)** (residual ≈ 11.3). Sin embargo, está notablemente subrepresentado en **japonés (ja)** (residual ≈ -8.3).

Independencia entre adulto y género

En esta parte decidimos evaluar se investigó si la clasificación de una película según su género (genres_agrupado) es independiente de si su contenido es para adultos (adult_agrupado).

- **Hipótesis Nula (H0):** El género agrupado de una película es independiente de su clasificación de contenido adulto agrupado.
- **Hipótesis Alternativa (H1):** El género agrupado de una película no es independiente de su clasificación de contenido adulto agrupado.

Con esto en mente se desarrolló la prueba chi cuadrado con estas dos variables transformadas en sus agrupadas, esto dándonos el siguiente resultado:

```
Pearson's Chi-squared test with simulated p-value (based on 10000
replicates)

data:  tabla_contingencia
X-squared = 251.82, df = NA, p-value = 9.999e-05
```

Viendo los resultados llegamos a la conclusión que existe una fuerte evidencia estadística para afirmar que el género de una película (agrupado) y si su contenido es para adultos o no (agrupado) no son independientes, por lo tanto se rechaza H0. Dado este resultado se decidió optar por hacer el análisis post-hoc de residuos. Haciendo el análisis correspondiente:

- Los géneros **Family** y **Thriller** tienden a estar sobrerrepresentados en contenido no adulto (FALSE) (residuales ≈ 2.7 y ≈ 3.0, respectivamente) y, por lo tanto, subrepresentados en contenido adulto.
- Los géneros **Action**, **Adventure**, y **Music** también muestran una tendencia a estar más asociados con contenido no adulto (FALSE) (residuales ≈ 2.0, ≈ 1.2, y ≈ 1.9 respectivamente para FALSE), aunque con menor intensidad que Family o Thriller.
- Los géneros **Crime** y **Fantasy** muestran una leve tendencia a estar más asociados con contenido adulto (TRUE) (residuales ≈ 1.5 y ≈ 1.6 respectivamente para TRUE).
- Otros géneros muestran que están muy sobrerrepresentada en contenido adulto (TRUE) (residual ≈ 14.0) y, consecuentemente, muy subrepresentada en contenido no adulto (FALSE) (residual ≈ -14.0).

<u>Conclusión:</u>

El presente estudio se propuso identificar los factores clave que influyen en la popularidad y el éxito financiero de las películas, utilizando un amplio conjunto de datos de IMDb. A través de un análisis descriptivo y diversas pruebas estadísticas inferenciales, se han obtenido hallazgos significativos que permiten comprender mejor las dinámicas de la industria cinematográfica.

Inicialmente, el análisis exploratorio reveló una alta variabilidad y asimetría en variables financieras cruciales como el presupuesto (budget) y los ingresos (revenue), así como en la popularidad (popularity). Las pruebas de normalidad confirmaron que estas variables, junto con la calificación promedio (vote_average), no siguen una distribución normal en su forma original. Si bien las transformaciones logarítmicas aplicadas a budget, revenue y popularity ayudaron a mitigar la asimetría, la normalidad estricta no se alcanzó en todos los casos, evidenciando la complejidad inherente a estas métricas. La calificación promedio (vote_average), por ejemplo, mostró una distribución particular que sugiere patrones de calificación específicos por parte de los usuarios.

Las pruebas de hipótesis arrojaron luz sobre relaciones específicas. Se encontró una diferencia estadísticamente significativa en los ingresos (logarítmicos) promedio entre las películas clasificadas como contenido para adultos y aquellas que no lo son, sugiriendo que la clasificación de contenido tiene una implicación económica. Además, los análisis de varianza (ANOVA) exploratorios indicaron que tanto el género cinematográfico como el idioma original parecen influir de manera significativa en la popularidad y el presupuesto de las películas, respectivamente, con ciertos géneros e idiomas destacando en estas métricas.

Los análisis de bondad de ajuste revelaron que las distribuciones de variables categóricas como el idioma original, el género y la clasificación de contenido adulto no son uniformes; ciertos idiomas y géneros son mucho más prevalentes que otros, y la proporción de películas para adultos difiere significativamente de una distribución equitativa. Además, se encontró que el número de géneros principales asociados a una película no sigue una distribución de Poisson, lo que sugiere que la combinación de géneros no es un proceso aleatorio simple.

Finalmente, las pruebas de independencia Chi-cuadrado confirmaron la existencia de asociaciones significativas entre los pares de variables categóricas analizadas: género e idioma original, género y clasificación de contenido adulto, e idioma original y clasificación de contenido adulto. El análisis post-hoc de los residuales estandarizados detalló estas interdependencias, mostrando, por ejemplo, fuertes afinidades entre ciertos géneros e idiomas (como Animación y japonés, o Drama y francés/español), y entre géneros y la clasificación de contenido (como la mayor propensión de géneros de "Otros" y "Romance" hacía contenido adulto).

En conjunto, este estudio subraya que el éxito y la popularidad de una película no dependen de un único factor, sino de una compleja interacción de variables. El género, el idioma original y la clasificación de contenido adulto demuestran ser elementos interconectados que se asocian de manera diferencial con el presupuesto, los ingresos y la popularidad. Si bien el idioma inglés y ciertos géneros como el Horror o la Ciencia Ficción muestran una fuerte presencia en indicadores clave, la diversidad de la producción cinematográfica global se refleja en nichos específicos donde otros idiomas y géneros prosperan. Estos hallazgos pueden ofrecer perspectivas valiosas para la toma de decisiones en la industria cinematográfica, aunque se reconoce la

necesidad de considerar modelos más complejos y variables adicionales para una comprensión más profunda.