¿CÚAL PONEMOS HOY?

Las 1000 mejores películas de la historia del cine, según *IMDB*

Diana Celine Pérez Rojas, Ariadna Romero Montero
Tipología y ciclo de vida de los datos
Universitat Oberta de Catalunya
Junio 2023

Índice

1. Descripción del conjunto de datos

El conjunto de datos motivo de este análisis contiene las mil películas con mejor puntuación en el sitio web IMDB (`https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows`), que se especializa en el recopilado de información sobre contenido audiovisual (televisión, cine, plataformas de *streaming*). Estas puntuaciones fueron las otorgadas por los usuarios de esta plataforma.

El conjunto de datos se compone de 1000 filas y 16 columnas:

Nombre de la variable	**Tipología**	**Descripción**
Poster Link	Cadena de caracteres	Enlace/url al cartel que acompaña la ficha de cada película en IMDB
Series_Title	Cadena de caracteres	Nombre
Released_Year	Cadena de caracteres	Año de estreno
Certificate	Cadena de caracteres	Clasificación/Código de contenido
Runtime	Cadena de caracteres	Duración total
Genre	Cadena de caracteres	Género
IMDB_Rating	Numérica	Puntuación
Overview	Cadena de caracteres	Sinopsis breve
Meta_score	Numérica	Puntuaciones del sitio web Metacritic.com
Director	Cadena de caracteres	Director/es
Star1-Star4	Cadena de caracteres	Actores y actrices principales
No_of_votes	Numérica	Número de votos obtenidos por la película
Gross	Numérica	Recaudación total

De acuerdo al planteamiento de la Práctica 1, es de nuestro interés encontrar algún patrón o características que distinguen a las películas que gozan de cierto nivel de reconocimiento entre la crítica no especializada. De esta forma, se podría prematuramente determinar si una película estrenada en este año podría encontrarse después de unos meses dentro de esta lista.

```
> dim(data)
[1] 1000   16
> str(data)
'data.frame':   1000 obs. of  16 variables:
 $ Poster_Link  : chr  "https://m.media-amazon.com/images/M/MV5BMDFkYTc0MGEtZmNhMC00ZDIzLWI
images/M/MV5BM2MyNjYxNmUtYTAwNi00MTYxLWJmNWYtYzZlODY3ZTk3OTFlXkEyXkFqcGdeQXVyNzkwMjQ"| __tr
k2Mw@@._V1_UX67_CR0,0,67,98_AL_.jpg" "https://m.media-amazon.com/images/M/MV5BMWMwMGQzZTIt
 $ Series_Title : chr  "The Shawshank Redemption" "The Godfather" "The Dark Knight" "The Go
 $ Released_Year: chr  "1994" "1972" "2008" "1974" ...
 $ Certificate  : chr  "A" "A" "UA" "A" ...
 $ Runtime      : chr  "142 min" "175 min" "152 min" "202 min" ...
 $ Genre        : chr  "Drama" "Crime, Drama" "Action, Crime, Drama" "Crime, Drama" ...
 $ IMDB_Rating  : num  9.3 9.2 9 9 9 8.9 8.9 8.9 8.8 8.8 ...
 $ Overview     : chr  "Two imprisoned men bond over a number of years, finding solace and
patriarch transfers control of his clandestine empire to his reluctant son." "When the mena
one of th"| __truncated__ "The early life and career of Vito Corleone in 1920s New York Cit
 $ Meta_score   : int  80 100 84 90 96 94 94 94 74 66 ...
 $ Director     : chr  "Frank Darabont" "Francis Ford Coppola" "Christopher Nolan" "Francis
 $ Star1        : chr  "Tim Robbins" "Marlon Brando" "Christian Bale" "Al Pacino" ...
 $ Star2        : chr  "Morgan Freeman" "Al Pacino" "Heath Ledger" "Robert De Niro" ...
 $ Star3        : chr  "Bob Gunton" "James Caan" "Aaron Eckhart" "Robert Duvall" ...
 $ Star4        : chr  "William Sadler" "Diane Keaton" "Michael Caine" "Diane Keaton" ...
 $ No_of_Votes  : int  2343110 1620367 2303232 1129952 689845 1642758 1826188 1213505 2067
 $ Gross        : chr  "28,341,469" "134,966,411" "534,858,444" "57,300,000" ...
```

Figura 1: Primer visionado al conjunto de datos

2. Integración y selección

En primer lugar, se añadirá una columna que contenga un identificador único para cada película.

Seguidamente, eliminamos las columnas que no nos aportan información relevante para el análisis: *Poster_Link* y *Overview*.

En adición, se renombrará la columna que contiene el nombre de la película de 'Series_Title' a 'Movies_Title'.

```
# Adición de una columna con un identificador único para cada película
data <- data %>% mutate(ID=row_number())
# Eliminación de las columnas innecesarias
data <- data %>% select(-c('Poster_Link', 'Overview'))
# Renombrar variable "Series_Title" a "Movies_Title"
data <- data %>% rename(Movies_Title = Series_Title)
```

Figura 2: Integración y selección del conjunto de datos

3. Limpieza de los datos

Integración y corrección

En primer lugar, procedemos a transformar la columna *Runtime*, ya que junto al número de minutos de duración de la película, aparece la cadena de caracteres 'min'. Para reemplazar todas las ocurrencias de esta cadena en la columna *Runtime*, se hará uso de `gsub`.

Seguidamente, se reducirán los valores de la columna *Genre* a uno: la variable almacenará únicamente el primer valor antes de la coma. Para ello, se usará una combinación entre `sapply`, que permite aplicar una función al vector indicado, `strsplit`, que divide una cadena de caracteres por el caracter indicado - en este caso una coma -, y, finalmente, `trimws` - para eliminar cualquier espacio en blanco de la cadena de caracteres resultante -.

A continuación, se transformará a las variables *Released_Year*, *Runtime*, *Meta_score* y *IMDB_Rating* a numéricas. Al ejecutar este función, R lanza una advertencia o *Warning*, donde indica que valores nulos han sido introducidos por coerción. Esto se debe a que, como se verá a continuación, la variable *Meta_score* contiene originalmente valores nulos.

Finalmente, se procederá a discretizar las variables de clasificaciones/puntuaciones (*IMDB_Rating* y *Meta_score*). Para no perder información valiosa respecto a posibles relaciones entre las puntuaciones y otras variables numéricas, como podría ser la recaudación, se optará por que las variables discretizadas conformen una nueva variable (su nombre se verá acompañado de un '_mod' al final.

IMDB_Rating Esta clasificación toma valores de 0 a 10, siendo el mínimo 7.6 y el máximo 9.3. En este caso, se opta por una discretización basada en el usuarios, donde se han definido 3 categorías: '7.5-8' (678 películas), '8-8.5' (289 películas) y '8.5-10' (33 películas).

Meta_score Esta clasificación toma valores de 0 a 100, siendo el mínimo valor 29 y el máximo 100. Para esta discretización, se opta por *equal-width binning* de 5 'bins', que llevarán asociadas una etiqueta.

Valores nulos

```
> colSums(is.na(data))
   Movies_Title   Released_Year     Certificate         Runtime           Genre
              0               1               0               0               0
    IMDB_Rating    Meta_score        Director           Star1           Star2
              0             157               0               0               0
          Star3           Star4    No_of_Votes           Gross              ID
              0               0               0             169               0
IMDB_Rating_mod  Meta_score_mod
              0             157
```

Figura 3: Valores nulos

En el conjunto de datos hay tres columnas con valores nulos: *Released_Year* con 1 valor, *Meta_score* con 157 valores, *Gross* con 169 valores y, finalmente, *Meta_score_mod* con 157

valores.

En el caso de *Released_Year*, la película en cuestión es 'Apollo 13'. Al tratarse de una única fila, se opta por imputar directamente el valor que corresponde al año de lanzamiento de la película - 1995 de acuerdo a la ficha de esta en IMDB (`https://www.imdb.com/title/tt0112384/`) -.

Finalmente, para los valores nulos en *Meta_score* y *Gross*, inicialmente se hizo un análisis de correlaciones para determinar si existía una relación lineal entre las variables; en el caso de *Meta_score*, muestra una ligera relación negativa con la variable *Released_Year* (-0.31), y para *Gross*, muestra una relación positiva con *No_of_votes* (0.56). Finalmente, se decide hacer la imputación mediante un *MissForest*, ya que es un método robusto y permite trabajar con un conjunto de datos mixtos. Los valores de correlación, no obstante, no parecen conducir a una relación fuerte que permita realizar predicciones sobre los valores nulos. En adición, el número de valores únicos en lo que se refiere a las variables categóricas, permite crear el supuesto de que será difícil para alguno de los algoritmos de imputación de encontrar un patrón que permite inferir un valor a imputar.

Si observamos el ajuste de la media y de la mediana respectivas de estas variables, como es de esperar, la mediana se muestra más robusta a los valores atípicos dentro de la distribución. Por ello, se ha optado por hacer una imputación directa en este caso.

Valores en blanco

```
> data["Certificate"][data["Certificate"] == '']
  [1] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
 [27] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
 [53] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
 [79] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
```

Figura 4: Valores en blanco en la variable Certificate

En cuanto a la columna *Certificate*, lo que se encuentran son valores en blanco. Para poder corregir este hecho, inicialmente se sustituirán los valores en blanco por valores nulos.

Los valores nulos en estas categorías se corresponden con aquellas películas que obtuvieron una clasificación 'Not rated' en su ficha de IMDB. En consecuencia, se imputará este valor para estos valores nulos.

4. Análisis de los datos

Análisis descriptivo de los datos

Previo a acometer pruebas estadísticas al conjunto de datos, se realizará un análisis exploratorio inicial de las variables:

En cuanto a las variables numéricas, los estadísticos de resumen indican que:

Released_Year : La mayoría de las películas fueron estrenadas antes de los años 2000.

Runtime : De media, las películas incluidas en este ranking tienen una duración de 119 minutos (cerca de dos horas).

IMDB_Rating : En cuanto al rating que obtienen por parte de los usuarios de IMDB, de mediana reciben una puntuación de 7.9 sobre 10.

Meta_score : En lo que respecta al rating obtenido por Metacritic, los resultados son más diversos: la nota mínima es de 28 sobre 100 y la mediana es de 79.

No_of_votes En cuanto a los votos obtenidos, presentan una mayor dispersión

Gross: Es la variable numérica que presenta mayor dispersión, de acuerdo a la varianza calculada: su valor mínimo es de 1305, mientras que el valor máximo es de 936.662.225.

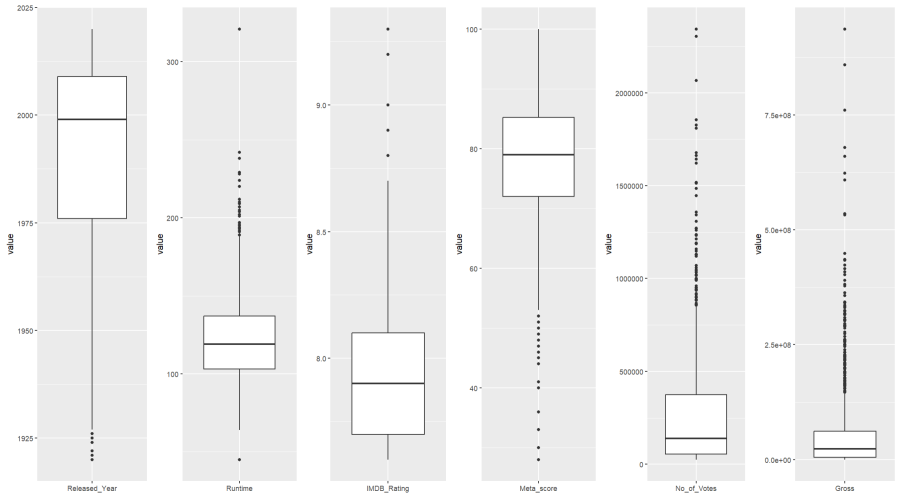

Figura 5: Diagramas de caja para las variables numéricas

En cuanto a las variables categóricas [1]:

Certificate La mayoría de las películas incluidas en el listado obtuvieron la calificación 'U', 'A', o 'UA'. Sorprende que una categoría restrictiva (R) esté dentro del top.

Genre Sobresalen los títulos dramáticos y, en menor medida, los de acción.

IMDB_Rating_mod La mayoría de las películas dentro del ranking tenían una puntuación entre 7-5 y 8.

Meta_score_mod En cuanto a la puntuación obtenido por los críticos, las películas acostumbran a recibir una calificación de 'B', un equivalente a un notable.

Director El top 10 está conformado por directores conocidos como Steven Spielberg, Alfred Hitchcock y Hayao Miyazaki, con más de 10 películas dentro del ranking.

[1] Para las variables *Director*, *Star1*, *Star2*, *Star3* y *Star4*, únicamente se hace un análisis sobre los 10 principales valores, por aparición, dentro del conjunto de datos.

Star1-Star4 En cuanto a los actores y actrices principales de las películas, destacan Tom Hanks, Robert de Niro, Clint Eastwood o Al Pacino (con 10 o más apariciones).

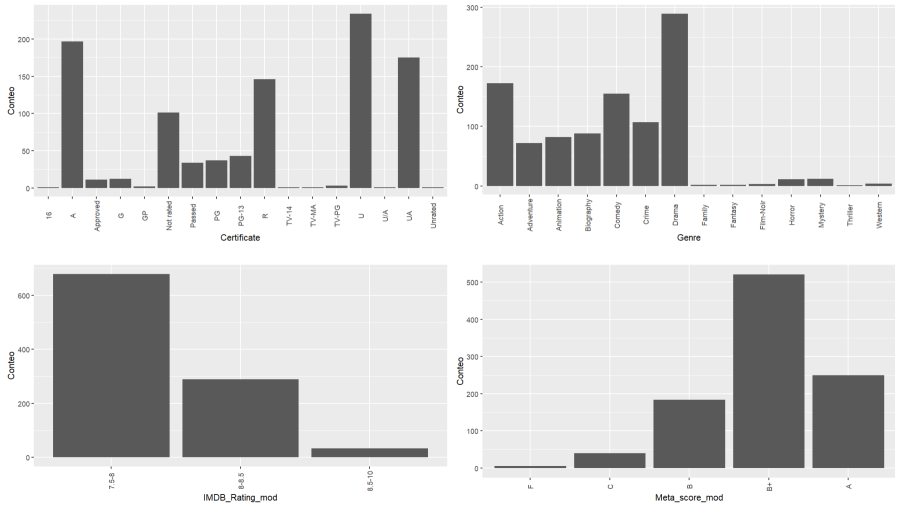

Figura 6: Variables categóricas

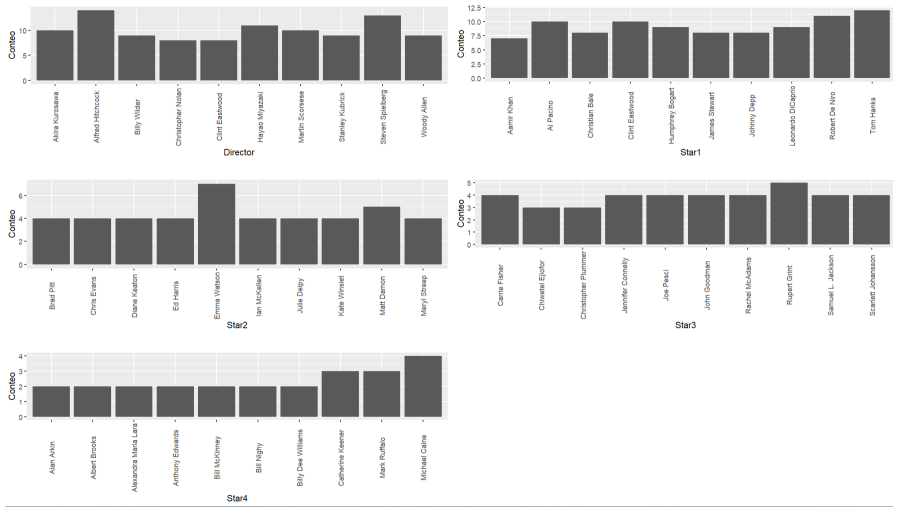

Figura 7: Variables categóricas (2)

Análisis

Una vez preparados y realizada la limpieza de datos, comenzamos con el análisis.

En primer lugar, se realiza un bucle que recorre cada variable en el conjunto de datos. Para cada variable numérica, se realiza una prueba de normalidad (Shapiro-Wilk) para evaluar si sigue una distribución normal. Además, se realiza una prueba de homogeneidad de varianzas (Levene) entre la variable actual y la variable *IMDB_Rating_mod*. Los resultados de estas pruebas nos permitirán determinar si las variables numéricas cumplen con los supuestos necesarios para posteriores análisis estadísticos. A continuación, se llevan a cabo pruebas de correlación de Spearman entre la variable *IMDB_Rating* y *Runtime*, *No_of_Votes*, *Gross* y *Released_Year*. Estas pruebas evalúan la relación entre las variables y la calificación en IMDB. Los resultados nos proporcionarán el valor de rho y el valor p correspondientes a cada correlación. A través del valor de p, podremos determinar si existe correlación (valor de p ¡0.05) y, dependiendo de si el valor de rho es positivo o negativo, podremos saber si se trata de una correlación positiva o negativa.

```
Variable: Released_Year
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.8998864
Valor p: 5.031752e-25

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 8.020892
Valor p: 0.0003501855

Variable: Runtime
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.9378209
Valor p: 5.129456e-20

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 8.603535
Valor p: 0.0001974284

Variable: IMDB_Rating
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.9193035
Valor p: 1.104189e-22

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 5.625945
Valor p: 0.003718499

Variable: Meta_score
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.9648652
Valor p: 8.429763e-15

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 7.122143
Valor p: 0.0008487512

Variable: No_of_Votes
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.7329853
Valor p: 3.150821e-37

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 71.81318
Valor p: 7.303783e-30

Variable: Gross
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.5985256
Valor p: 4.738402e-43

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 6.040706
Valor p: 0.002467872

Variable: ID
Prueba de normalidad (Shapiro-Wilk):
Estadístico: 0.954809
Valor p: 5.388677e-17

Prueba de homogeneidad de varianzas (Levene):
Estadístico: 176.2733
Valor p: 2.824147e-66
```

Figura 8: Resultados de los tests Shapiro-Wilk y Levene

En el segundo análisis, se realiza una prueba de Kruskal-Wallis para evaluar si hay diferencias significativas en la calificación de IMDB entre diferentes categorías de variables. En este caso, se realizan pruebas para las mismas variables que en el análisis anterior, pero esta vez también se añaden variables categóricas como 'Genre' y 'Director'. Al igual que en el análisis anterior, los valores de p obtenidos nos permitirán determinar si las variables analizadas son factores determinantes en la calificación de IMDB.

En tercer lugar, se realiza una regresión lineal para examinar la relación entre las variables independientes (*Runtime*, *Gross*, *Released_Year*, *No_of_Votes* y *Genre*) y la variable dependiente (*IMDB_Rating*). Esta vez, se excluye la variable 'Director' debido a que presenta demasiados valores únicos y sería complicado realizar el análisis. Con este análisis, podremos ver cuánta variabilidad explica nuestro modelo (R-cuadrado), qué variables muestran valores de p más bajos (lo que significa una mayor influencia en la calificación de IMDB) y si esta influencia es positiva o negativa, lo cual se visualiza a través del signo de puntuación en la columna 'Estimate' de manera similar a la prueba de Spearman.

```
> # Realizar una regresión lineal 1
> model <- lm(IMDB_Rating ~ Runtime + Gross + Released_Year + No_of_Votes + as.factor(Genre), data = data)
>
> # Obtener los resultados de la regresión
> summary(model)

Call:
lm(formula = IMDB_Rating ~ Runtime + Gross + Released_Year +
    No_of_Votes + as.factor(Genre), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55511 -0.15105 -0.01681  0.13357  0.78086

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.426e+01  6.158e-01  23.165   <2e-16 ***
Runtime                     2.394e-03  2.614e-04   9.159   <2e-16 ***
Gross                      -8.076e-10  8.623e-11  -9.366   <2e-16 ***
Released_Year              -3.394e-03  3.113e-04 -10.902   <2e-16 ***
No_of_Votes                 5.980e-07  2.593e-08  23.060   <2e-16 ***
as.factor(Genre)Adventure -3.127e-02  2.986e-02  -1.047    0.295
as.factor(Genre)Animation  1.506e-01  2.950e-02   5.104   4e-07 ***
as.factor(Genre)Biography  1.325e-02  2.805e-02   0.472    0.637
as.factor(Genre)Comedy     2.900e-02  2.443e-02   1.187    0.235
as.factor(Genre)Crime      4.073e-02  2.668e-02   1.526    0.127
as.factor(Genre)Drama      4.657e-02  2.116e-02   2.201    0.028 *
as.factor(Genre)Family    -1.278e-03  1.501e-01  -0.009    0.993
as.factor(Genre)Fantasy    2.703e-02  1.514e-01   0.179    0.858
as.factor(Genre)Film-Noir -1.406e-02  1.235e-01  -0.114    0.909
as.factor(Genre)Horror    -3.498e-02  6.592e-02  -0.531    0.596
as.factor(Genre)Mystery   -1.997e-02  6.320e-02  -0.316    0.752
as.factor(Genre)Thriller  -4.961e-02  2.111e-01  -0.235    0.814
as.factor(Genre)Western    2.313e-01  1.071e-01   2.160    0.031 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2101 on 982 degrees of freedom
Multiple R-squared:  0.4281,    Adjusted R-squared:  0.4182
F-statistic: 43.23 on 17 and 982 DF,  p-value: < 2.2e-16
```

Figura 9: Resultados de la primera regresión

Finalmente, se introduce en la regresión la variable *Meta_score*, que es significativa de acuerdo a su *p-value*, pero mejora el ajuste del modelo ligeramente, pero no así el R-cuadrado ajustado.

En resumen, mediante la aplicación de pruebas estadísticas y modelos de regresión, realizamos un análisis detallado de los datos para explorar las relaciones entre las variables y la calificación en IMDB.

```
> # Realizar una regresión lineal 2
> model1 <- lm(IMDB_Rating ~ Runtime + Gross + Released_Year + No_of_Votes + as.factor(Genre) + Meta_score, data = data)
>
> # Obtener los resultados de la regresión
> summary(model1)

Call:
lm(formula = IMDB_Rating ~ Runtime + Gross + Released_Year +
    No_of_Votes + as.factor(Genre) + Meta_score, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54437 -0.14827 -0.02248  0.13511  0.74584

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.246e+01  6.422e-01  19.409  < 2e-16 ***
Runtime                    2.270e-03  2.544e-04   8.923  < 2e-16 ***
Gross                     -7.994e-10  8.378e-11  -9.542  < 2e-16 ***
Released_Year             -2.657e-03  3.172e-04  -8.377  < 2e-16 ***
No_of_Votes                5.847e-07  2.526e-08  23.153  < 2e-16 ***
as.factor(Genre)Adventure -4.335e-02  2.905e-02  -1.492 0.135902
as.factor(Genre)Animation  1.105e-01  2.913e-02   3.793 0.000158 ***
as.factor(Genre)Biography  1.579e-03  2.730e-02   0.058 0.953871
as.factor(Genre)Comedy     1.150e-02  2.384e-02   0.482 0.629763
as.factor(Genre)Crime      3.092e-02  2.596e-02   1.191 0.233869
as.factor(Genre)Drama      2.547e-02  2.074e-02   1.228 0.219673
as.factor(Genre)Family    -1.326e-02  1.458e-01  -0.091 0.927578
as.factor(Genre)Fantasy    5.207e-02  1.471e-01   0.354 0.723422
as.factor(Genre)Film-Noir -8.119e-02  1.203e-01  -0.675 0.500007
as.factor(Genre)Horror    -5.119e-02  6.408e-02  -0.799 0.424592
as.factor(Genre)Mystery   -3.547e-02  6.143e-02  -0.577 0.563789
as.factor(Genre)Thriller  -6.555e-02  2.051e-01  -0.320 0.749314
as.factor(Genre)Western    2.359e-01  1.041e-01   2.267 0.023596 *
Meta_score                 4.686e-03  6.083e-04   7.704 3.22e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2042 on 981 degrees of freedom
Multiple R-squared:  0.4607,    Adjusted R-squared:  0.4508
F-statistic: 46.55 on 18 and 981 DF,  p-value: < 2.2e-16
```

Figura 10: Resultados de la segunda regresión

5. Resolución del problema

Después de realizar un análisis exhaustivo del conjunto de datos IMDb Top 1000 y aplicar modelos analíticos, se llegaron a las siguientes conclusiones:

Se encontraron correlaciones significativas entre la calificación de IMDb y algunas variables específicas. La duración de las películas (*Runtime*) y el número de votos (*No_of_Votes*) mostraron una correlación positiva moderada con la calificación de IMDb. Esto significa que las películas más largas y con un mayor número de votos tienden a tener una calificación más alta en IMDb. Por otro lado, no se encontró una correlación significativa entre la recaudación bruta (*Gross*) y la calificación de IMDb, lo que indica que el éxito financiero de una película no está necesariamente relacionado con su calificación en IMDb. Además, el año de lanzamiento (*Released_Year*) presentó una correlación negativa débil con la calificación de IMDb, lo que sugiere que las películas más recientes tienden a tener una calificación ligeramente más baja.

Se realizaron pruebas de Kruskal-Wallis para evaluar si había diferencias significativas en la calificación de IMDb en función de diferentes variables. Se encontraron diferencias significativas en la calificación de IMDb entre las diferentes categorías del año de lanzamiento (*Released_Year*) y el número de votos (*No_of_Votes*). Esto indica que tanto el año de lanzamiento como el número de votos pueden influir en la calificación de IMDb. Sin embargo, no se encontraron diferencias significativas en la calificación de IMDb entre los géneros de las películas (*Genre*) ni entre los directores.

El modelo de regresión lineal mostró que la duración de las películas, las ganancias brutas, el año de lanzamiento y el número de votos son variables significativas para predecir la calificación de IMDb. Sin embargo, el género de las películas no mostró un efecto significativo en la calificación. El modelo tuvo un R-cuadrado de 0.4155, lo que indica que aproximadamente el 41.55 % de la variabilidad en la calificación de IMDb se explica por las variables incluidas en el modelo.

En resumen, se puede concluir que la duración de las películas, el número de votos y el año de lanzamiento son factores importantes que influyen en la calificación de IMDb. Por otro lado, las ganancias brutas y el género de las películas no parecen ser factores determinantes en la calificación.

6. Contribuciones

Contribuciones	**Firma**
Investigación Previa	Diana Celine Pérez Rojas, Ariadna Romero Montero
Redacción de las respuestas	Diana Celine Pérez Rojas, Ariadna Romero Montero
Desarrollo del código	Diana Celine Pérez Rojas, Ariadna Romero Montero
Participación en el vídeo	Diana Celine Pérez Rojas, Ariadna Romero Montero