Práctica 2: Limpieza y análisis de datos

Las 1.000 Películas Mejor Valoradas en FilmAffinity (2013–2023)

Enero 2024

Índice de contenidos

Integrantes del Grupo

- Juan Antonio Tora Cánovas
- Tim Thorp

1 Descripción del dataset

El dataset original se puede consultar en los siguientes links:

- En nuestro repositorio GitHub, bajo el directorio «/data»: **GitHub Repository**
- A través del siguiente link de ZENODO: **Enlace ZENODO**

Este dataset constaba de 12 columnas, de las que solo usamos 6. Son las siguientes:

- Duración
- Género
- País
- Puntuación Media
- Director
- Reparto

¿Por qué podría ser importante este dataset y qué problema resuelve?

Este estudio se enfoca en un conjunto de datos de películas, buscando entender qué factores contribuyen significativamente a las puntuaciones de las mismas. Mediante el análisis de una variedad de características como género, director, reparto y duración, investigaremos cómo diferentes elementos pueden afectar la recepción de las películas. Por ejemplo, ¿tienen las películas de animación mejores puntuaciones en promedio? ¿Influye la duración de la película en su puntuación? El objetivo es descubrir patrones ocultos y proporcionar una comprensión más profunda de los factores que contribuyen al éxito de una película.

Este análisis es importante porque ayuda a identificar las tendencias en las preferencias de la audiencia y puede guiar a los creadores de contenido en la toma de decisiones. Al comprender mejor qué hace que una película sea bien recibida, los productores y directores pueden ajustar sus enfoques para satisfacer mejor las expectativas del público.

---

2 Integración y selección

Para comenzar, cargamos el archivo CSV y examinamos su estructura.

```
df <- read.csv('../data/dataset_movie_info.csv', dec=',')
str(df)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ Título                : chr  "Coco" "Joker" "Parásitos" "Spider-Man: Cruzando el Multiverso" ...
##  $ Título.Original       : chr  "Coco" "Joker" "Gisaengchung" "Spider-Man: Across the Spider-Verse"
##  $ Año                   : int  2017 2019 2019 2023 2014 2014 2018 2014 2015 2016 ...
##  $ Duración              : int  109 121 132 140 169 74 126 103 94 106 ...
##  $ Género                : chr  "Animación, Fantástico, Comedia, Drama" "Thriller, Drama" "Intriga, C
##  $ País                  : chr  "Estados Unidos" "Estados Unidos" "Corea del Sur" "Estados Unidos" .
##  $ Puntuación.Media      : num  8 8 8 7.9 7.9 7.9 7.8 7.8 7.8 7.8 ...
##  $ Número.de.Puntuaciones: chr  "53.934" "70.870" "61.739" "10.674" ...
##  $ Director              : chr  "Lee Unkrich, Adrián Molina" "Todd Phillips" "Bong Joon-ho" "Joaquim
##  $ Reparto               : chr  "Anthony Gonzalez, Gael García Bernal, Benjamin Bratt, Alanna Ubach,
##  $ Sinopsis              : chr  "Miguel es un joven con el sueño de convertirse en leyenda de la músi
##  $ Enlace                : chr  "https://www.filmaffinity.com/es/film893369.html" "https://www.filma
```

Observamos que cada película puede incluir varios géneros en una misma fila, separados por comas. Para facilitar su uso en nuestro análisis, es necesario crear una variable *dummy* para cada género individual.

```
library(tidyr)
library(dplyr)

# Convertimos la columna de géneros en un formato largo,
# donde cada fila es una película con un único género
df_long <- df %>%
  separate_rows(Género, sep = ",") %>%  # Separamos los géneros por comas
  mutate(Género = trimws(Género))       # Eliminamos espacios en blanco

# Ahora, creamos una lista de géneros únicos
unique_genres <- unique(df_long$Género)

# Creamos variables dummy para cada género único
# Cada película tendrá un 1 si pertenece a ese género, y un 0 si no pertenece
for(genre in unique_genres) {
  df[[genre]] <- ifelse(grepl(genre, df$Género), 1, 0)
}
```

Hacemos lo mismo para los directores. Para evitar la creación de cientos de dimensiones, limitamos nuestra selección a los directores que han dirigido al menos 4 películas.

```
df_long <- df %>%
  separate_rows(Director, sep = ",") %>%
  mutate(Director = trimws(Director))

# Contamos el número de películas que ha dirigido (o co-dirigido) cada director
director_counts <- table(df_long$Director)

# Identificamos los directores que han dirigido al menos 4 películas
frequent_directors <- which(director_counts >= 4)

# Extraemos los nombres de estos directores
frequent_directors_list <- names(frequent_directors)

for(director in frequent_directors_list) {
  df[[director]] <- ifelse(grepl(director, df$Director), 1, 0)
}
```

Análogamente, creamos las variables dummy para los actores. Limitamos nuestra selección a los actores que han aparecido en al menos 8 películas.

```
df_long <- df %>%
  separate_rows(Reparto, sep = ",") %>%
  mutate(Reparto = trimws(Reparto))

actor_counts <- table(df_long$Reparto)
frequent_actors <- which(actor_counts >= 8)
frequent_actors_list <- names(frequent_actors)

for(actor in frequent_actors_list) {
```

```
  df[[actor]] <- ifelse(grepl(actor, df$Reparto), 1, 0)
}
```

Eliminamos las columnas `Año`, `Título`, `Título.Original`, `Número.de.Puntuaciones`, `Sinopsis` y `Enlace` de nuestro conjunto de datos, ya que no las utilizaremos en nuestro análisis.

También descartamos las columnas `Género`, `Director` y `Reparto`, ya que las hemos reemplazado con las correspondientes variables *dummy*.

```
df <- select(df, -Año, -Género, -Director, -Reparto, -Título, -Título.Original,
             -Número.de.Puntuaciones, -Sinopsis, -Enlace)
```

Mostramos la estructura del dataframe actualizado.

```
str(df)
```

```
## 'data.frame':    1000 obs. of  59 variables:
##  $ Duración            : int  109 121 132 140 169 74 126 103 94 106 ...
##  $ País                : chr  "Estados Unidos" "Estados Unidos" "Corea del Sur" "Estados Unidos" ...
##  $ Puntuación.Media    : num  8 8 8 7.9 7.9 7.9 7.8 7.8 7.8 7.8 ...
##  $ Animación           : num  1 0 0 1 0 0 0 0 1 1 ...
##  $ Fantástico          : num  1 0 0 0 0 0 0 0 1 1 ...
##  $ Comedia             : num  1 0 1 0 0 0 0 0 1 0 ...
##  $ Drama               : num  1 1 1 0 1 1 1 1 0 1 ...
##  $ Thriller            : num  0 1 1 0 0 0 0 0 0 0 ...
##  $ Intriga             : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ Acción              : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ Aventuras           : num  0 0 0 1 1 0 0 0 1 0 ...
##  $ Ciencia ficción     : num  0 0 0 1 1 1 0 0 0 0 ...
##  $ Romance             : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ Bélico              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Infantil            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Musical             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Western             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cine negro          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Terror              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Anthony Russo       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Asghar Farhadi      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Chad Stahelski      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Clint Eastwood      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Damien Chazelle     : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ Denis Villeneuve    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ François Ozon       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Guy Ritchie         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hirokazu Koreeda    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hong Sang-soo       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ James Gunn          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Joe Russo           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Martin Scorsese     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Pablo Larraín       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Paolo Sorrentino    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Rodrigo Sorogoyen   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Steven Spielberg    : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ Taika Waititi       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Woody Allen         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ V48                 : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Adam Driver         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bárbara Lennie      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Benedict Cumberbatch: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bradley Cooper      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Carey Mulligan      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cate Blanchett      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Domhnall Gleeson    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Jake Gyllenhaal     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Jon Bernthal        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Manolo Solo         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Matthew McConaughey : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ Michael Stuhlbarg   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Olivia Colman       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Oscar Isaac         : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ Paul Dano           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Rooney Mara         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Scarlett Johansson  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Timothée Chalamet   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Tom Hanks           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Willem Dafoe        : num  0 0 0 0 0 0 0 0 0 0 ...
```

---

3 Limpieza de los datos

3.1 Gestión de valores perdidos

En la estructura actualizada, observamos que hay una variable dummy `V48` con valor positivo (1) para todas las películas. Esta variable corresponde a una cadena vacía, así que la eliminamos.

```
df <- select(df, -V48)
```

No hay valores `NA` presentes en el conjunto de datos:

```
sum(is.na(df))
```

```
## [1] 0
```

Tampoco hay cadenas vacías:

```
sum(df=="")
```

```
## [1] 0
```

Solo hay ceros en las columnas que corresponden a las variables *dummy*:

```r
colSums(df==0)
```

```
##             Duración                 País     Puntuación.Media
##                    0                    0                    0
##            Animación           Fantástico              Comedia
##                  896                  903                  810
##                Drama             Thriller              Intriga
##                  213                  814                  942
##              Acción            Aventuras      Ciencia ficción
##                  900                  915                  942
##             Romance              Bélico             Infantil
##                  897                  972                  987
##             Musical              Western           Cine negro
##                  980                  980                  996
##              Terror       Anthony Russo       Asghar Farhadi
##                  963                  996                  996
##       Chad Stahelski       Clint Eastwood      Damien Chazelle
##                  996                  996                  996
##     Denis Villeneuve        François Ozon          Guy Ritchie
##                  994                  996                  996
##     Hirokazu Koreeda        Hong Sang-soo           James Gunn
##                  993                  996                  996
##            Joe Russo      Martin Scorsese        Pablo Larraín
##                  996                  996                  996
##     Paolo Sorrentino    Rodrigo Sorogoyen     Steven Spielberg
##                  996                  996                  995
##        Taika Waititi          Woody Allen          Adam Driver
##                  996                  995                  991
##       Bárbara Lennie Benedict Cumberbatch       Bradley Cooper
##                  991                  988                  991
##       Carey Mulligan       Cate Blanchett     Domhnall Gleeson
##                  992                  992                  992
##      Jake Gyllenhaal         Jon Bernthal          Manolo Solo
##                  990                  992                  992
##  Matthew McConaughey    Michael Stuhlbarg        Olivia Colman
##                  992                  991                  992
##          Oscar Isaac            Paul Dano          Rooney Mara
##                  992                  992                  992
##   Scarlett Johansson    Timothée Chalamet            Tom Hanks
##                  991                  991                  991
##         Willem Dafoe
##                  989
```

3.2 Gestión de valores extremos

Primero, examinamos visualmente las variables continuas `Duración` y `Puntuación.Media`.

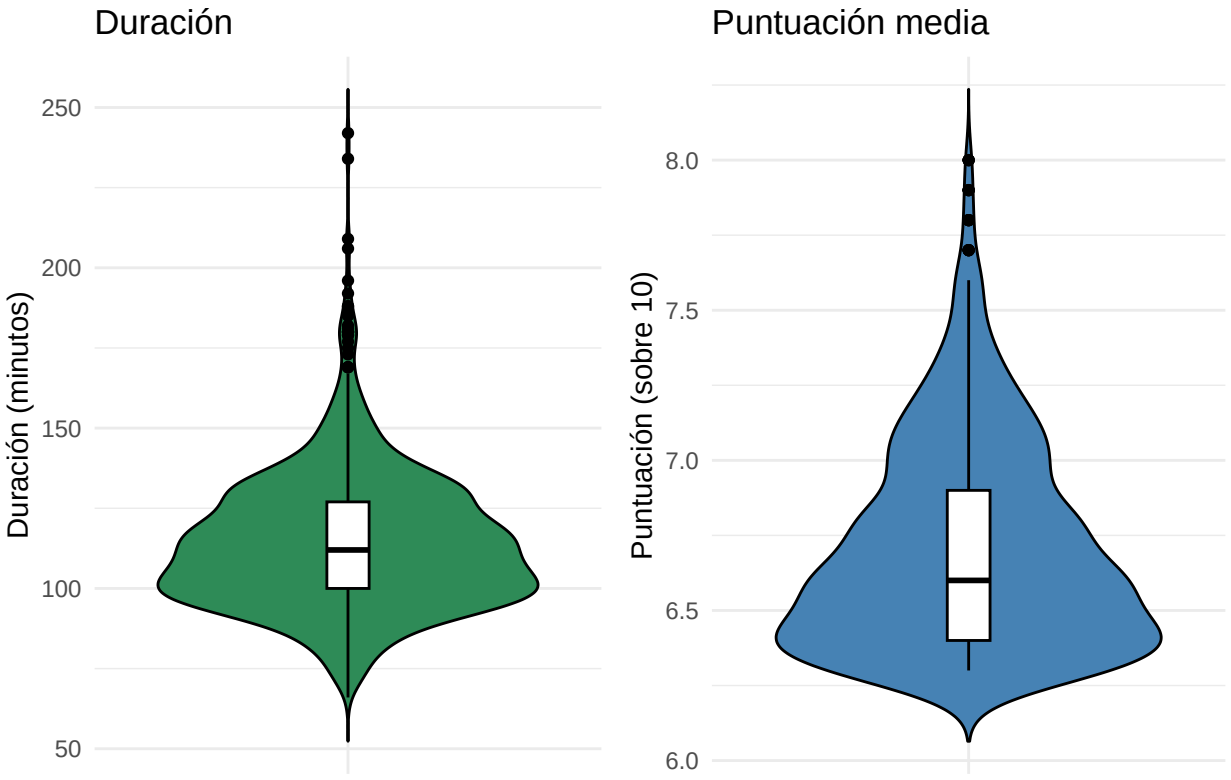
```r
library(ggplot2)
library(grid)
library(gridExtra)

duration_plot <- ggplot(df, aes(x = "", y = Duración)) +
```

```
  geom_violin(fill = "seagreen", color = "black", trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white", color = "black") +
  labs(x = "",
       y = "Duración (minutos)",
       title = "Duración") +
  theme_minimal()

rating_plot <- ggplot(df, aes(x = "", y = Puntuación.Media)) +
  geom_violin(fill = "steelblue", color = "black", trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white", color = "black") +
  labs(x = "",
       y = "Puntuación (sobre 10)",
       title = "Puntuación media") +
  theme_minimal()

grid.arrange(duration_plot, rating_plot, ncol = 2)
```



La distribución de `Duración` parece ser aproximadamente normal, aunque presenta una cola hacia la derecha (hacia arriba en este caso) y valores atípicos (*outliers*) en la parte superior.

La distribución de `Puntuación.Media` también presenta asimetría, con valores atípicos en el extremo superior. Pero esto es esperable: en la primera parte del proyecto cuando realizamos el *scraping* para el dataset, nos centramos en las 1.000 películas con las mejores puntuaciones, efectivamente cogiendo la parte superior de la distribución.

```r
sort(boxplot.stats(df$Duración)$out)
```

```
##  [1] 169 169 169 173 175 177 179 179 180 180 180 180 181 182 185 188 188 192 196
## [20] 206 209 234 242
```

El boxplot de `Duración` sugiere que las películas con una duración de 169 minutos o más se catalogan como valores atípicos. Sin embargo, estos valores son válidos — las 3 películas más largas del dataset son:

1. La Liga de la Justicia de Zack Snyder (242 minutos)
2. An Elephant Sitting Still (234 minutos)
3. El irlandés (209 minutos)

```r
sort(boxplot.stats(df$Puntuación.Media)$out)
```

```
##  [1] 7.7 7.7 7.7 7.7 7.8 7.8 7.8 7.8 7.9 7.9 7.9 8.0 8.0 8.0
```

Por otro lado, las películas con una puntuación igual o superior a 7,7 se han identificado como valores atípicos, pero como antes, estos también son válidos. Las 3 películas mejor puntuadas del dataset son:

1. Coco (8,0)
2. Parásitos (8,0)
3. Joker (8,0)

Por lo tanto, no realizamos ningún cambio.

---

4 Análisis de los datos

4.1 Selección de los grupos de datos

Aunque nos habría gustado hacer un ANOVA comparando las medias de todos los géneros, una de las asunciones de ANOVA es la independencia de las observaciones. Dado que cada película puede tener más de un solo género, no podemos cumplir con esta asunción.

Por lo tanto, para el primer análisis, vamos a realizar una comparación de medias con respecto a un género concreto.

Los grupos son los siguientes:

1. Puntuaciones medias de las películas que pertenezcan al género «Animación».
2. Puntuaciones medias de las películas que NO pertenezcan al género «Animación».

A continuación, dividimos el dataset en dos grupos.

```r
animation_ratings <- df[df$Animación==1, "Puntuación.Media"]
non_animation_ratings <- df[df$Animación==0, "Puntuación.Media"]
```

4.2 Comprobación de la normalidad

Para comprobar si los datos siguen una distribución normal, empleamos el test de Shapiro.

```r
shapiro.test(df$Puntuación.Media)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Puntuación.Media
## W = 0.90939, p-value < 2.2e-16
```

```r
shapiro.test(df$Duración)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Duración
## W = 0.94301, p-value < 2.2e-16
```

```r
shapiro.test(animation_ratings)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  animation_ratings
## W = 0.95052, p-value = 0.0006833
```

```r
shapiro.test(non_animation_ratings)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  non_animation_ratings
## W = 0.90638, p-value < 2.2e-16
```

El p-valor en cada caso es muy inferior a 0.05. Por lo tanto, se rechaza la hipótesis nula a favor de la hipótesis alternativa: los datos no siguen una distribución normal.

4.3 Comprobación de la homogeneidad de la varianza

Para comprobar la homogeneidad de la varianza, empleamos el test de Fligner. Se trata de un test no paramétrico que compara las varianzas basándose en la mediana. Es una alternativa cuando no se cumple la condición de normalidad en las muestras.

```r
library(stats)
fligner.test(list(animation_ratings, non_animation_ratings))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(animation_ratings, non_animation_ratings)
## Fligner-Killeen:med chi-squared = 17.348, df = 1, p-value = 3.112e-05
```

Al ser p-valor < 0.05, se rechaza la hipótesis nula a favor de la hipótesis alternativa: los datos presentan heterocedasticidad (es decir, que tienen varianzas distintas).

4.4 Análisis 1: Correlación entre duración y puntuación media

Dado que las variables `Duración` y `Puntuación.Media` no siguen una distribución normal, no es posible aplicar la correlación de Pearson de manera apropiada. En su lugar, hemos optado por el método de Kendall. Este también es un test no paramétrico que no asume la normalidad de los datos.

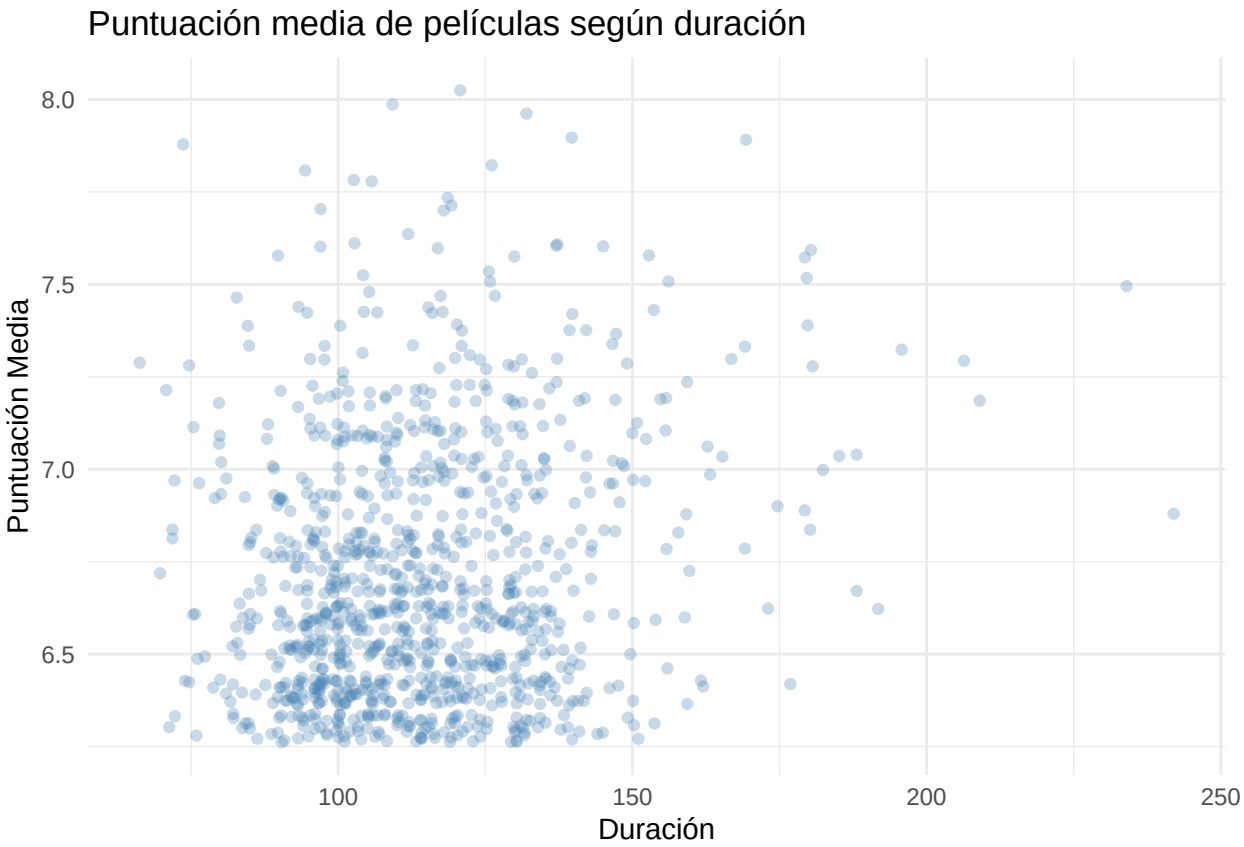
```
cor.test(df$Duración, df$Puntuación.Media, method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$Duración and df$Puntuación.Media
## z = 4.0961, p-value = 4.203e-05
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.0912933
```

El p-valor obtenido es menor a 0,001 ($4,203 \times 10^{-5}$), lo que nos permite descartar la hipótesis nula de que no existe correlación con un nivel de confianza del 99,9 %. El coeficiente de correlación de Kendall (τ) resultó ser 0,09, lo que sugiere una correlación significativa y positiva, aunque débil.

Por lo tanto, podemos concluir que las películas más largas tienden a tener puntuaciones medias ligeramente más altas. Esta correlación (débil) puede visualizarse mediante un gráfico de dispersión.

```
ggplot(data=df, aes(Duración, Puntuación.Media)) +
  geom_jitter(color="steelblue", alpha=0.3) +
  theme_minimal() +
  labs(title = "Puntuación media de películas según duración",
       x = "Duración",
       y = "Puntuación Media")
```



4.5 Análisis 2: Comparación de Puntuación Media (Animación vs. No Animación)

A pesar de que los datos no presentan una distribución normal, contamos con más de 100 observaciones (películas) en cada categoría:

```
table(factor(df$Animación, levels=c(0,1), labels=c("No animación", "Animación")))
```

```
##
## No animación    Animación
##          896          104
```

Con un número de muestras mayor a 30 en cada grupo, podemos apoyarnos en el teorema del límite central y suponer que las medias de ambos grupos se distribuyen de manera normal.

Procedemos a comparar las puntuaciones medias entre los grupos "Animación" y "No Animación" mediante una prueba t de Student. Nuestra hipótesis es que las películas animadas obtienen mejores puntuaciones, por lo que aplicamos un test unilateral (`alternative = "greater"`).

Debido a la heterocedasticidad en nuestros datos, incluimos el argumento `var.equal = FALSE`.

```
t.test(animation_ratings, non_animation_ratings, alternative = "greater", var.equal = FALSE)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  animation_ratings and non_animation_ratings
## t = 4.9069, df = 119.31, p-value = 1.48e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1356618       Inf
## sample estimates:
## mean of x mean of y
##  6.893269  6.688393
```

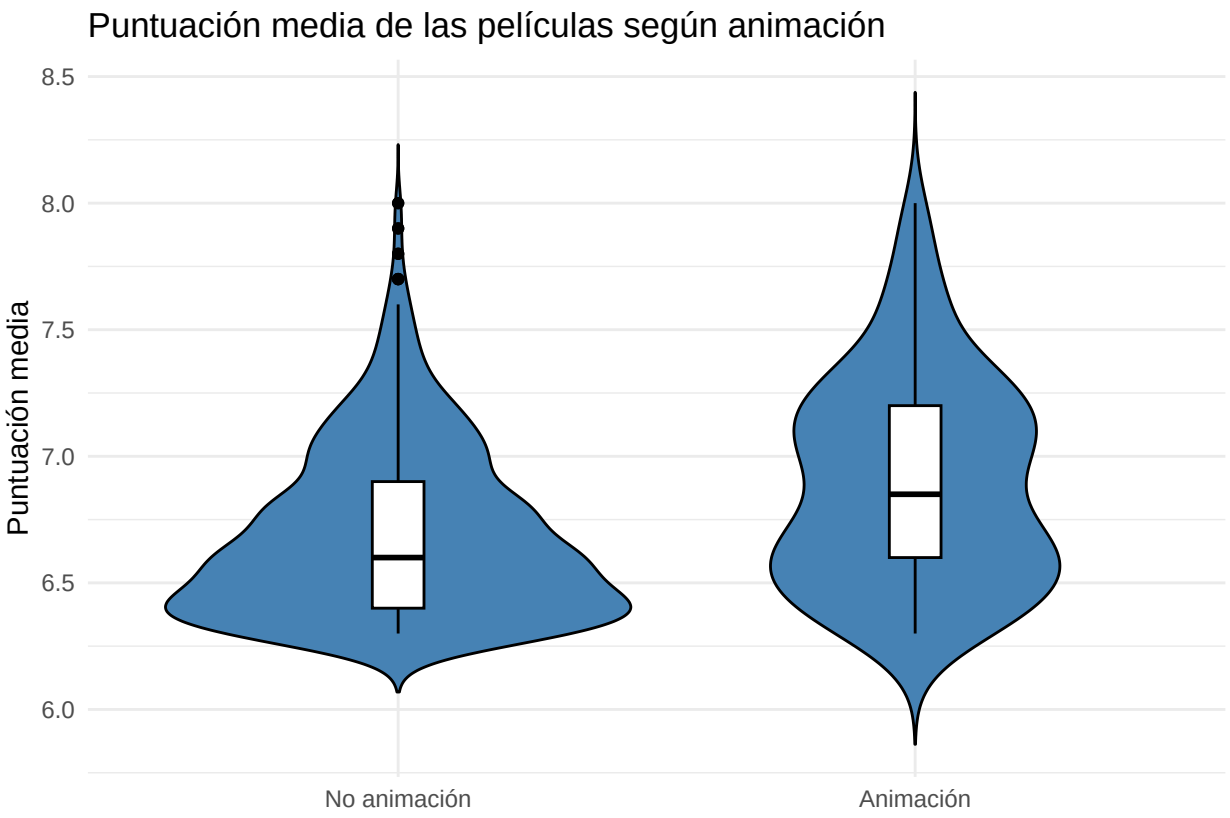
Los resultados de la prueba t-Student revelan un p-valor de $1,48 \times 10^{-6}$.

Esto nos lleva a rechazar la hipótesis nula (al ser menor que 0,05), aceptando así la hipótesis alternativa: la puntuación media de las películas animadas es superior a la de las no animadas.

El intervalo de confianza se sitúa en [0,136, ∞]. Con un 95 % de confianza, afirmamos que la puntuación media de las películas animadas es al menos 0,136 puntos superior a la de las no animadas.

Este resultado se ilustra en la siguiente gráfica.

```
ggplot(df,
       aes(x = factor(Animación, levels=c(0,1), labels=c("No animación", "Animación")),
           y = Puntuación.Media)) +
    geom_violin(fill = "steelblue", color = "black", trim = FALSE) +
    geom_boxplot(width = 0.1, fill = "white", color = "black") +
    labs(x = "",
         y = "Puntuación media",
         title = "Puntuación media de las películas según animación") +
    theme_minimal()
```



En el gráfico de violín, es evidente la distinta distribución de ambos grupos.

Las puntuaciones de las películas no animadas se concentran mayormente en la parte inferior, presentando una distribución asimétrica con una cola larga hacia la derecha (hacia arriba en este caso). Las películas con puntuaciones de 7,7 o más son consideradas *outliers* en este grupo.

En cambio, las películas animadas muestran una distribución aproximadamente bimodal, con un valor mediano notablemente más alto.

4.6 Análisis 3: Modelo *Random Forest*

En este análisis, desarrollaremos un modelo *Random Forest* (bosque aleatorio) con el objetivo de identificar las variables que ejercen mayor influencia en la puntuación media de las películas.

4.6.1 Preparación de los datos para el modelo

Los modelos *Random Forest* son eficientes en identificar automáticamente las características más relevantes durante el entrenamiento. No obstante, un exceso de variables y/o categorías puede incrementar el tiempo de procesamiento.

```
length(unique(df$País))
```

```
## [1] 61
```

Vemos que hay 61 países únicos en el dataset. Para reducir el número de categorías, decidimos agrupar los países con menos de 10 películas bajo la categoría «Otro».

```r
# Contamos el número de películas por país
country_counts <- table(df$País)

# Identificamos los países que aparecen menos de 10 veces
infrequent_countries <- which(country_counts < 10)

# Extraemos los nombres de estos países
infrequent_countries_list <- names(infrequent_countries)

# Reemplazamos los nombres de estos países en el dataframe por "Otro"
df$País[df$País %in% infrequent_countries_list] <- "Otro"

# Convertimos la columna País en un factor
df$País <- factor(df$País)

# Volvemos a comprobar el número de países únicos
length(unique(df$País))
```

```
## [1] 17
```

Con este ajuste, el número de países se ha reducido de 61 a 17, incluyendo la categoría «Otro».

4.6.2 Creación del modelo

Procedemos a ajustar el modelo *Random Forest* a nuestros datos (este proceso puede tardar varios minutos al ejecutar el código).

```r
library(caret)

train_control <- trainControl(
  method = "cv",              # para la validación cruzada
#  verboseIter = TRUE,        # para un log detallado (comentar antes de knit to PDF)
)

# Establecemos un seed para asegurar reproducibilidad
set.seed(123)
rf_model <- train(
  Puntuación.Media ~ .,       # la formula para el modelo (puntuación media como variable dependiente)
  data = df,
  method = "rf",              # el tipo de modelo (Random Forest)
  trControl = train_control, # el objeto control que hemos creado en el paso anterior
  metric = "Rsquared",        # la métrica que queremos optimizar en el modelo
)

# Mostramos el resumen del modelo por pantalla
rf_model
```

```
## Random Forest
##
## 1000 samples
##   57 predictor
##
```

```
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared    MAE
##    2    0.3363008  0.07865795  0.2740522
##   37    0.3461414  0.07372519  0.2757341
##   72    0.3511703  0.07084242  0.2787110
##
## Rsquared was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

Observamos que el coeficiente de determinación (R^2) es bastante bajo (0,079), lo que indica que nuestro modelo solo explica un 7,9 % de la varianza en las puntuaciones de las películas. Es muy probable que haya otros factores con mayor influencia en la puntuación de la película, como el guion o el presupuesto, para dar unos ejemplos.

Sin embargo, sería interesante analizar cuáles son los factores que más han contribuido al modelo.

4.6.3 Análisis de los resultados del modelo

A continuación, analizamos las 10 variables más importantes en el modelo.

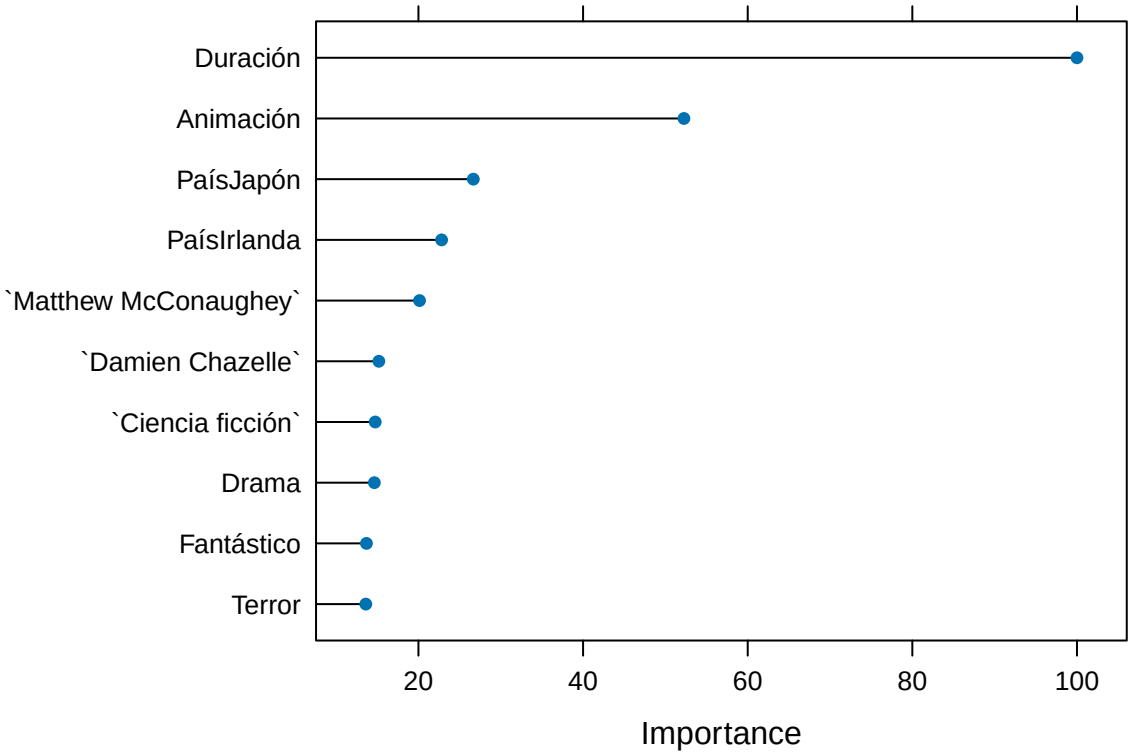
```r
# Escalamos las importancias de las variables, estableciendo el máximo en 100
importance <- varImp(rf_model, scale = TRUE)

# Convertimos la importancia de las variables en un dataframe para poder manipularlas mejor
importance_df <- as.data.frame(importance$importance)

# Ordenamos el dataframe por la importancia de las variables en orden descendente
importance_sorted <- importance_df[order(-importance_df$Overall),]

# Extraemos los nombres de las 10 variables más importantes
top_features_names <- rownames(head(importance_sorted, 10))

# Creamos y mostramos un gráfico de las 10 variables más importantes en el modelo
plot(importance, top = 10)
```



Observamos que `Duración` ha sido la variable más influyente, seguido por `Animación`, lo cuál es consistente con los hallazgos de la prueba t del análisis anterior.

La mayoría de las variables de mayor importancia son los géneros, pero también observamos algunos países (Japón, Irlanda), un actor (Matthew McConaughey) y un director (Damien Chazelle).

4.7 Análisis 4: Regresión lineal

Una de las limitaciones del modelo *Random Forest* es que no podemos deducir fácilmente la dirección de cada variable. Es decir, si la variable en cuestión tiende a aumentar o disminuir la puntuación de la película.

Sin embargo, podemos complementar el modelo *Random Forest* con un modelo lineal y así deducir la dirección de influencia. Es importante destacar que no podamos incluir todos los géneros a la vez: cada película puede tener más de un género, lo que crea problemas de colinealidad.

```
# Creamos variables dummy para los países Japón y Irlanda
df$Japón <- as.numeric(df$País == "Japón")
df$Irlanda <- as.numeric(df$País == "Irlanda")

summary(lm(Puntuación.Media ~ Duración + Animación + Japón + Irlanda +
             `Matthew McConaughey` + `Damien Chazelle`, data=df))
```

```
##
## Call:
## lm(formula = Puntuación.Media ~ Duración + Animación + Japón +
##     Irlanda + `Matthew McConaughey` + `Damien Chazelle`, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78478 -0.24505 -0.06027  0.20654  1.36957
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.2671949  0.0594460 105.427  < 2e-16 ***
## Duración              0.0035572  0.0005055   7.037 3.66e-12 ***
## Animación             0.2236663  0.0375192   5.961 3.47e-09 ***
## Japón                 0.1106989  0.0560178   1.976 0.048416 *
## Irlanda               0.4071359  0.1049389   3.880 0.000111 ***
## `Matthew McConaughey` 0.3043255  0.1176762   2.586 0.009848 **
## `Damien Chazelle`     0.4160173  0.1651792   2.519 0.011939 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3288 on 993 degrees of freedom
## Multiple R-squared:  0.1108, Adjusted R-squared:  0.1054
## F-statistic: 20.62 on 6 and 993 DF,  p-value: < 2.2e-16
```

El R^2 ajustado, que toma en cuenta el número de variables en el modelo, es de 0,105. Esto indica que el modelo lineal ha logrado explicar un 10,5 % de la varianza en las puntuaciones con estas 6 variables.

Observamos que todos los coeficientes son positivos y, además, significativos con un nivel de confianza del 95 %, lo que respalda los resultados del modelo *Random Forest*.

El coeficiente de `Duración` sugiere que un aumento de un minuto en la duración corresponde a un aumento de 0,0036 en la puntuación de la película. Por ejemplo, comparando una película de 2 horas con otra de 3 horas, la más larga tendría, en promedio, una puntuación 0,22 más alta $(0,0036 \times 60)$.

De manera similar, las películas del género `Animación` tienen, en promedio, una puntuación 0,22 más alta en comparación con las que no pertenecen a este género.

Las películas de Japón e Irlanda muestran un aumento en la puntuación de 0,11 y 0,41, respectivamente.

Contrario a lo esperado, las películas en las que aparece Matthew McConaughey tienen una puntuación 0,30 más alta. *(Perdón, no podía resistir.)*

Finalmente, las películas dirigidas por Damien Chazelle también presentan una puntuación 0,42 más alta.

A continuación, comprobamos los coeficientes de los otros géneros por separado.

```
coefficients(lm(Puntuación.Media ~ `Ciencia ficción`, data=df))
```

```
##       (Intercept) `Ciencia ficción`
##        6.70414013        0.09585987
```

```
coefficients(lm(Puntuación.Media ~ Drama, data=df))
```

```
## (Intercept)       Drama
##  6.66572770  0.05587332
```

```
coefficients(lm(Puntuación.Media ~ Fantástico, data=df))
```

```
## (Intercept)  Fantástico
##  6.70099668  0.08972497
```

```
coefficients(lm(Puntuación.Media ~ Terror, data=df))
```

```
## (Intercept)      Terror
##   6.7165109  -0.1840785
```

Observamos que las películas de `Ciencia ficción` (+0,10), `Drama` (+0,06) y `Fantástico` (+0,09) corresponden a un ligero aumento de la puntuación, mientras que las películas de `Terror` suelen estar asociadas con una puntuación notablemente peor (-0,18). Aunque no se ha incluido en la memoria para ahorrar espacio, todos estos coeficientes son significativos con un nivel de confianza de 95 %.

Los coeficientes del modelo lineal proporcionan información valiosa sobre el impacto general de una variable, pero es importante destacar que los modelos *Random Forest* pueden identificar interacciones y que los efectos de una variable pueden no ser lineales.

Por ejemplo, a pesar de que normalmente las películas más largas estén asociadas con puntuaciones más altas, esta tendencia puede variar según el género u otros factores. Al emplear nuestro modelo *Random Forest* para realizar predicciones, una mayor duración podría aumentar la puntuación estimada en ciertos géneros, mientras que en otros podría reducirla.

---

5 Representación de los resultados

Este apartado ya ha quedado resuelto con diferentes gráficas y/o tablas, incluidas durante el desarrollo de todos los ejercicios anteriores.

---

6 Resolución del problema

Hemos llegado a varias conclusiones que responden a nuestro problema inicial: entender qué factores influyen significativamente en las puntuaciones de las películas.

1. **Correlación entre Duración y Puntuación Media**: Hemos encontrado una correlación débil pero significativa entre la duración de las películas y sus puntuaciones medias. En general, las películas más largas tienden a tener puntuaciones ligeramente más altas.
2. **Comparación de Puntuación Media entre Géneros (Animación vs. No Animación)**: Las películas de animación, en promedio, tienen puntuaciones más altas en comparación con aquellas que no son de animación.
3. **Influencia de Factores Específicos (Modelo *Random Forest*)**: El modelo *Random Forest* ha identificado varios factores como significativamente influyentes en las puntuaciones de las películas: la duración, ciertos géneros (como Animación, Ciencia ficción, Drama, Fantástico y Terror), países específicos (Japón e Irlanda), un actor (Matthew McConaughey) y un director (Damien Chazelle). A pesar de estos resultados, el coeficiente de determinación relativamente bajo sugiere que hay otros factores con una mayor influencia en las puntuaciones.

4. **Análisis de Regresión Lineal**: Mediante el análisis de regresión lineal, hemos podido estimar la dirección y la magnitud del impacto de estas variables. Por ejemplo, una larga duración muestra un efecto positivo en las puntuaciones, lo cual es consistente con los resultados de la correlación inicial. Además, se ha observado que ciertos géneros, como animación, ciencia ficción, drama y fantástico, tienden a incrementar las puntuaciones, mientras que el terror tiende a reducirlas.

Estos hallazgos nos ofrecen una mayor comprensión de los factores que contribuyen al éxito de una película desde la perspectiva de su puntuación. No obstante, es importante reconocer las limitaciones de nuestro estudio. En la primera fase del proyecto, nos centramos únicamente en las 1.000 películas mejor valoradas, lo cual introdujo un sesgo en los datos. Un enfoque más amplio, que incluya películas con puntuaciones inferiores, podría revelar más factores que afectan negativamente a las valoraciones.

Aunque no fue objeto de nuestro estudio, también sería relevante considerar otras variables que podrían influir en la puntuación, como el presupuesto, si la película es una secuela o si está basada en un libro, entre otros aspectos.

Dicho esto, los resultados siguen siendo valiosos para productores, directores y otros profesionales de la industria cinematográfica, ya que ofrecen información sobre varios aspectos a tener en cuenta para potencialmente mejorar la recepción de sus películas.

---

7 Código

El código R utilizado para la limpieza, análisis y representación de los datos está incluido en esta memoria, así como en el siguiente repositorio privado:

- **GitHub Repository**

Los archivos relevantes se encuentran **bajo el directorio «/code»**:

- **Memoria.Rmd**
- **Memoria.R**

Para crear el archivo `Memoria.R`, primero se desarrolló `Memoria.Rmd` y después con el siguiente código se creó el de extensión «.R»:

```
library(knitr)
purl("Memoria.rmd")
```

```
## [1] "Memoria.R"
```

A continuación, exportamos un fichero CSV con los datos finales analizados.

```
write.csv(df, '../data/data_analyzed.csv', row.names = FALSE)
```

---

8 Vídeo

Para acceder al vídeo de presentación de la práctica, clic aquí: **Google Drive**

Para acceder al repositorio GitHub de la práctica, clic aquí: **GitHub Repository**

Tabla de Contribuciones

Las iniciales representan la confirmación por parte del grupo de que el integrante ha participado en dicho apartado.

Contribuciones	Firma
Investigación previa	JATC, TT
Redacción de las respuestas	JATC, TT
Desarrollo del código	JATC, TT
Participación en el vídeo	JATC, TT

Bibliografía Utilizada

1. Amat Rodrigo, J. (2016). Análisis de la homogeneidad de varianza (homocedasticidad). Recuperado el 2 de enero de 2024, de https://rpubs.com/Joaquin_AR/218466
2. Calvo, M., Subirats, L., & Pérez, D. (2019). *Introducción a la limpieza y análisis de los datos*. Editorial UOC.
3. Dalgaard, P. (2008). *Introductory statistics with R*. Springer Science & Business Media.
4. Farouni, R. (2016). Random Forest Regression using Caret. Recuperado el 3 de enero de 2024, de https://gist.github.com/rfarouni/9be5c651af60d5d7cc6c9b529e821b47.
5. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.
6. McKinney, W. (2012). *Python for Data Analysis*. O'Reilly Media, Inc.
7. Osborne, J. W. (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. *Newborn and Infant Nursing Reviews*, 10(1), 1527-3369.
8. Squire, M. (2015). *Clean Data*. Packt Publishing Ltd.