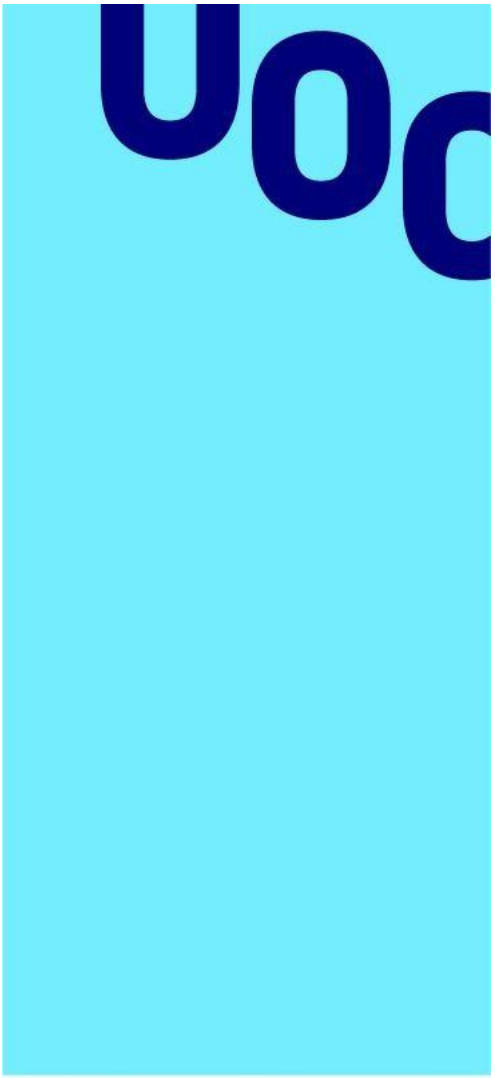
Tipología y ciclo de vida de los datos

Práctica 2



Universitat Oberta de Catalunya

Tratamiento del Dataset de PR1

Estefanía Gázquez
Fabián López

Links:

- https://www.filmaffinity.com/es/main.html
- https://github.com/flopezcardozo/TCVD_Practica2.git
- https://colab.research.google.com/drive/1IXEfC5tO-fhmACp4f6OjUYyjbMNqsRnY?usp=drive_link
- https://drive.google.com/file/d/1h_y0Q4abd8xE9OpLSqY_83XI2pwinZEo/view?usp=drive_link

Fecha de entrega
26 de mayo de 2025

1. Descripción del Dataset

El dataset utilizado en esta práctica fue elaborado mediante técnicas de web scraping sobre la plataforma FilmAffinity, una de las bases de datos cinematográficas más reconocidas en el ámbito hispano. Contiene información detallada sobre 155 películas y series altamente valoradas, extraídas de diferentes categorías de la plataforma, abarcando un período desde 1931 hasta 2025.

Este conjunto de datos es relevante porque permite analizar tendencias cinematográficas desde múltiples perspectivas: temporales, geográficas, por género, o incluso a nivel de autoría (directores). Además, resulta adecuado para un ejercicio completo de tratamiento de datos, ya que contiene variables categóricas y numéricas, además de ciertos valores faltantes o inconsistencias que requieren ser gestionados en las fases posteriores de la práctica.

El problema o pregunta principal que se pretende abordar con este dataset es: ¿Qué factores influyen en la valoración media que recibe una película por parte del público? Para ello, se explorarán correlaciones entre variables como el género, el país de origen, el director o la duración, y se evaluará si es posible construir modelos predictivos o descubrir agrupaciones significativas de películas en función de sus atributos.

Variables del Dataset

El dataset contiene **30 variables** organizadas en las siguientes categorías:

Variables de Identificación

Variable	Tipo	Descripción
Título	Categórica	Nombre de la película o Serie
URL	Categórica	URL de FilmAffinity
Fecha_extraccion	Temporal	Fecha de Extracción de los Datos
Categoría_extraccion	Categórica	Categoría de extracción (ej: “Top España”)

Variables temporales

Variable	Tipo	Descripción
Año	Numérica (entero)	Año de estreno
Decada	Numérica (entero)	Década de estreno

Variables Técnicas

Variable	Tipo	Descripción
duración	Numérica (entero)	Duración en minutos
tipo	Categórica	Clasificación como “Película” o “Serie”
país	Categórica	País de origen
director	Categórica	Director
género	Categórica	Lista original de géneros

Variables de Género (Binarias)

16 variables binarias que indican la presencia de cada género.

Variables Objetivo y Popularidad

Variable	Tipo	Descripción
Valoración_media	Numérica (real)	Variable Objetivo: Valoración promedio (4.4 – 9.0)
Num_votos	Numérica (entero)	Número total de votos recibidos

Caracterísiticas del Dataset

Tamaño: 155 registros × 30 variables
Período temporal: 1931 - 2025 (94 años)
Rango de valoraciones: 4.4 - 9.0 puntos
Tipos de contenido: Películas y series de televisión
Datos faltantes: Identificados en variables como país y director
Inconsistencias: Series clasificadas incorrectamente como películas

El dataset proporciona una base sólida para realizar análisis estadísticos y aplicar técnicas de aprendizaje automático supervisado y no supervisado, mientras que la presencia de datos faltantes e inconsistencias permite aplicar técnicas reales de limpieza y validación de datos.

2. Integración y selección de los datos de interés a analizar

El dataset inicial fue obtenido mediante técnicas de web scraping desde FilmAffinity, específicamente de la categoría "Top España", y contiene información de 155 películas y series altamente valoradas. Tras evaluar los requisitos de la práctica, se ha optado por trabajar con el conjunto completo de datos original sin necesidad de integrar datasets adicionales.

Esta decisión se fundamenta en que el dataset actual **cumple completamente** con todos los requisitos establecidos:

- Amplia variedad de datos numéricos y categóricos
- Variable objetivo claramente definida (valoración_media)
- Presencia de datos faltantes y/o erróneos.
- Diversidad temporal y de contenido.

Es en base a esto que se ha optado por mantener el dataset completo (155 registros x 30 variables) para el análisis, ya que representa una muestra curada y balanceada del contenido mejor valorado en FilmAffinity España.

3. Limpieza de datos

3.1 El análisis inicial del Dataset reveló la presencia de diferentes tipos de valores que indican pérdida de información como por ejemplo las variables “país” y “director” que omitían el 100% de los datos. No se detectaron valores NULL o NaN, tampoco valores 0 inapropiados en variables numéricas.

3.2 Se realizó la conversión de los siguientes tipos de datos:

- Fecha_extraccion: de string a datetime, esto facilita el análisis temporal.
- géneros: de string con formato de lista a lista de Python
- tipo: convertida a categorical.
- categoria_extraccion: convertira a categorical

3.3 Se realizo análisis de valores extremos en variables numéricas y se encontraron los siguientes outliers cuya duración era menor a 30 minutos o mayor a 300 minutos:

- Los Simpson: 22 minutos
- Solo Leveling: 23 minutos
- Apocalipsis: La segunda Guerra Mundial: 320 minutos

Se decidió mantener todos los valores, ya que representan duraciones reales de series de TV o documentales extensos.

3.4 Para este dataset en particular se encontró un problema de inconsistencia en clasificación de tipos ya que 26 series estaban clasificadas incorrectamente como “Película” esto de detectó cruzando datos entre la variable “géneros” y la variable “tipo”, por lo que se procedió a la reclasificación automática de estos registros como “Serie”.

También se procedió a la creación de variables derivadas para enriquecer el análisis, para ello se crearon 4 nuevas variables categóricas:

- popularidad: Categorización basada en número de votos
  - Baja: 0-1,000 votos
  - Media: 1,000-5,000 votos
  - Alta: 5,000-20,000 votos
  - Muy Alta: >20,000 votos
- epoca: Categorización temporal
  - Clásico: Hasta 1980
  - Moderno: 1981-2000
  - Contemporáneo: 2001-2010
  - Actual: 2011-2025
- duracion_cat: Categorización de duración
  - Corto: 0-90 minutos
  - Normal: 91-120 minutos
  - Largo: 121-180 minutos
  - Muy Largo: >180 minutos
- generos_list: Lista procesada de géneros para análisis avanzados

Estos métodos de limpieza fueron necesarios debido al origen de los datos, la complejidad del dominio y los objetivos analíticos planteados para el ejercicio.

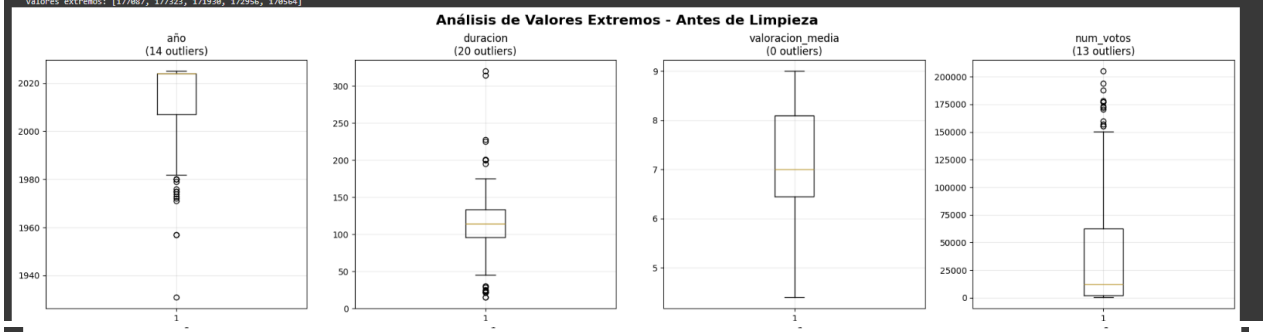
4. Análisis de datos

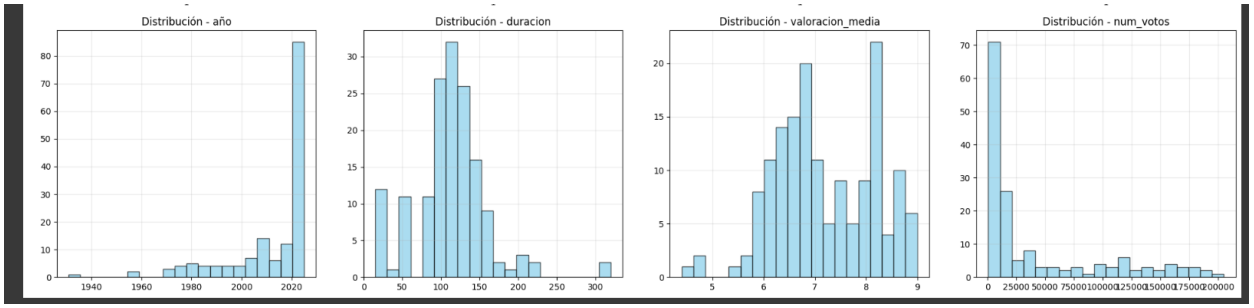
Se aplica en Google Colab lo solicitado en los puntos 4.1 y 4.2.

5. Representación de los resultados

A lo largo del análisis se han generado diversas visualizaciones que permiten comprender mejor tanto la distribución de los datos como los resultados obtenidos en las fases de modelado.

```
AÑO:
   Rango normal: [1981.5, 2049.5]
   Outliers detectados: 14
   Valores extremos: [1972, 1974, 1980, 1957, 1957]

DURACION:
   Rango normal: [38.5, 190.5]
   Outliers detectados: 20
   Valores extremos: [200, 195, 320, 22, 228]

VALORACION_MEDIA:
   Rango normal: [4.0, 10.6]
   Outliers detectados: 0

NUM_VOTOS:
   Rango normal: [-88234.2, 153199.8]
   Outliers detectados: 13
   Valores extremos: [177887, 177323, 171930, 172956, 170564]
```

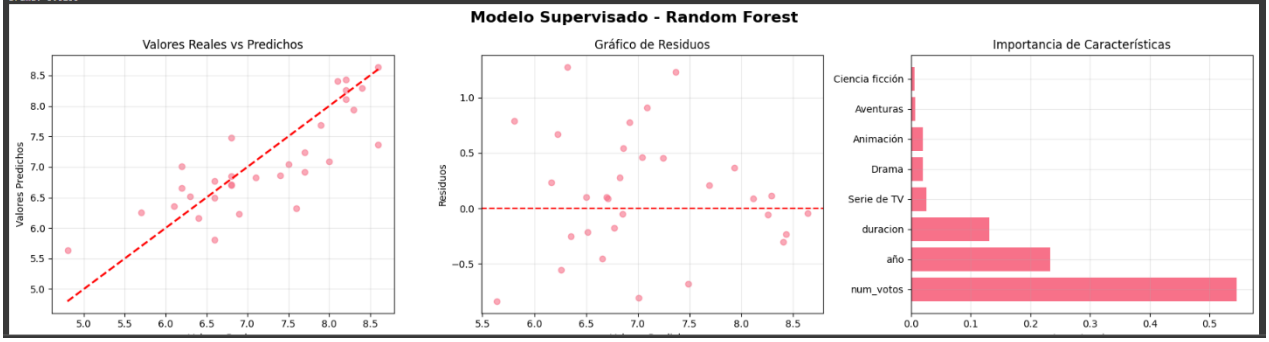




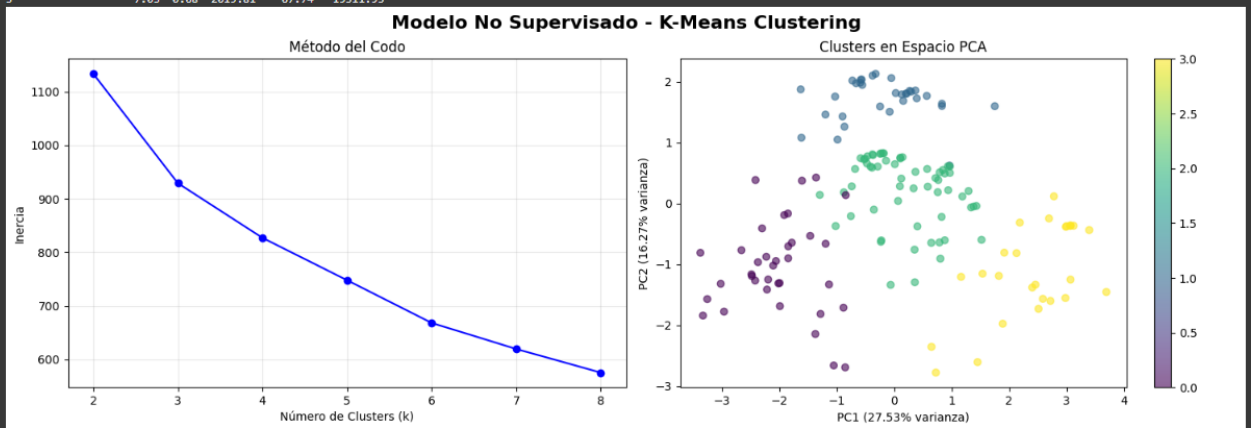
```
RESUMEN FINAL DE LA LIMPIEZA
COMPARACIÓN ANTES/DESPUÉS:
Dimensiones originales: 155 × 30
Dimensiones finales:    155 × 32
Variables eliminadas:   -2
Variables añadidas:     4

INFORMACIÓN DEL DATASET LIMPIO:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 155 entries, 0 to 154
Data columns (total 32 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   titulo                  155 non-null    object
 1   año                     155 non-null    int64
 2   duracion                155 non-null    int64
 3   generos                 155 non-null    object
 4   valoracion_media        155 non-null    float64
 5   num_votos               155 non-null    int64
 6   tipo                    155 non-null    object
 7   url                     155 non-null    object
 8   fecha_extraccion        155 non-null    datetime64[ns]
 9   categoria_extraccion    155 non-null    object
 10  genero_Serie de TV      155 non-null    int64
 11  genero_Thriller         155 non-null    int64
 12  genero_Animación        155 non-null    int64
 13  genero_Intriga          155 non-null    int64
 14  genero_Aventuras        155 non-null    int64
 15  genero_Acción           155 non-null    int64
 16  genero_Romance          155 non-null    int64
 17  genero_Documental       155 non-null    int64
 18  genero_Western          155 non-null    int64
 19  genero_Bélico           155 non-null    int64
 20  genero_Infantil         155 non-null    int64
 21  genero_Drama            155 non-null    int64
 22  genero_Musical          155 non-null    int64
 23  genero_Fantástico       155 non-null    int64
 24  genero_Comedia          155 non-null    int64
 25  genero_Ciencia ficción  155 non-null    int64
 26  genero_Terror           155 non-null    int64
 27  decada                  155 non-null    int64
 28  generos_list            155 non-null    object
 29  popularidad             155 non-null    category
 30  epoca                   155 non-null    category
 31  duracion_cat            155 non-null    category
dtypes: category(3), datetime64[ns](1), float64(1), int64(21), object(6)
memory usage: 36.3+ KB
None
```

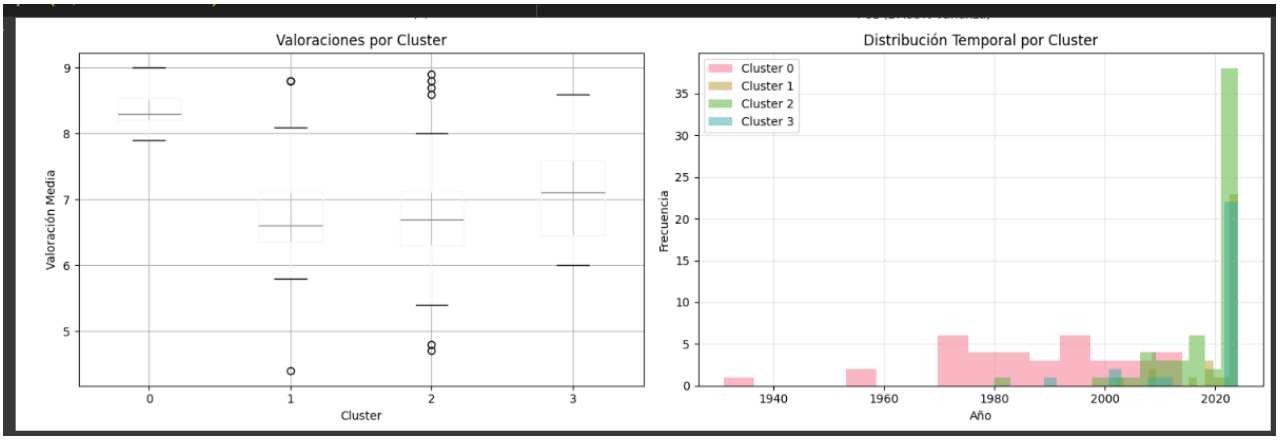
```
ESTADÍSTICAS FINALES:
              año  duracion  valoracion_media  num_votos  genero_Serie de TV  \
count      155.00    155.00            155.00     155.00              155.00
mean      2012.23    113.30              7.19   41532.28                0.17
std         17.69     46.84              0.97   56405.09                0.37
min       1931.00     15.00              4.40     514.00                0.00
25%       2007.00     95.50              6.45    2303.50                0.00
50%       2024.00    114.00              7.00   12184.00                0.00
75%       2024.00    133.50              8.10   62662.00                0.00
max       2024.00    320.00              9.00  204973.00                1.00

       genero_Thriller  genero_Animación  genero_Intriga  genero_Aventuras  \
count           155.00            155.00          155.00            155.00
mean              0.25              0.19            0.12              0.19
std               0.44              0.39            0.32              0.40
min               0.00              0.00            0.00              0.00
25%               0.00              0.00            0.00              0.00
50%               0.00              0.00            0.00              0.00
75%               0.50              0.00            0.00              0.00
max               1.00              1.00            1.00              1.00

       genero_Acción  genero_Romance  genero_Documental  genero_Western  \
count         155.00          155.00             155.00          155.00
mean            0.21            0.09               0.03            0.14
std             0.41            0.29               0.18            0.35
min             0.00            0.00               0.00            0.00
25%             0.00            0.00               0.00            0.00
50%             0.00            0.00               0.00            0.00
75%             0.00            0.00               0.00            0.00
max             1.00            1.00               1.00            1.00

       genero_Bélico  genero_Infantil  genero_Drama  genero_Musical  \
count         155.00           155.00        155.00          155.00
mean            0.03             0.02          0.51            0.01
std             0.18             0.14          0.50            0.11
min             0.00             0.00          0.00            0.00
25%             0.00             0.00          0.00            0.00
50%             0.00             0.00          1.00            0.00
75%             0.00             0.00          1.00            0.00
max             1.00             1.00          1.00            1.00

       genero_Fantástico  genero_Comedia  genero_Ciencia ficción  \
count             155.00          155.00                  155.00
mean                0.14            0.23                    0.17
std                 0.34            0.42                    0.38
min                 0.00            0.00                    0.00
25%                 0.00            0.00                    0.00
50%                 0.00            0.00                    0.00
75%                 0.00            0.00                    0.00
max                 1.00            1.00                    1.00

       genero_Terror   decada
count         155.00   155.00
mean            0.08  2008.00
std             0.28    17.74
min             0.00  1930.00
25%             0.00  2000.00
50%             0.00  2020.00
75%             0.00  2020.00
max             1.00  2020.00

Dataset limpio guardado en: /content/drive/My Drive/Colab Notebooks/TCVD_Practica2/FilmAffinity_Dataset/filmaffinity_cleaned.csv
```

```
======================================================================
4. ANÁLISIS DE DATOS
======================================================================
Dataset cargado: 155 filas × 32 columnas

MODELO SUPERVISADO - RANDOM FOREST
Variables predictoras: 11
Variables numéricas: ['año', 'duracion', 'num_votos']
Top géneros: ['Drama', 'Thriller', 'Comedia', 'Acción', 'Aventuras', 'Animación', 'Ciencia ficción', 'Serie de TV']
Entrenamiento: 124 muestras
Prueba: 31 muestras

RESULTADOS DEL MODELO:
R² (entrenamiento): 0.9649
R² (prueba): 0.6496
RMSE (entrenamiento): 0.1830
RMSE (prueba): 0.5503
MAE (prueba): 0.4317

IMPORTANCIA DE CARACTERÍSTICAS:
num_votos: 0.5447
año: 0.2332
duracion: 0.1307
Serie de TV: 0.0255
Drama: 0.0199
```

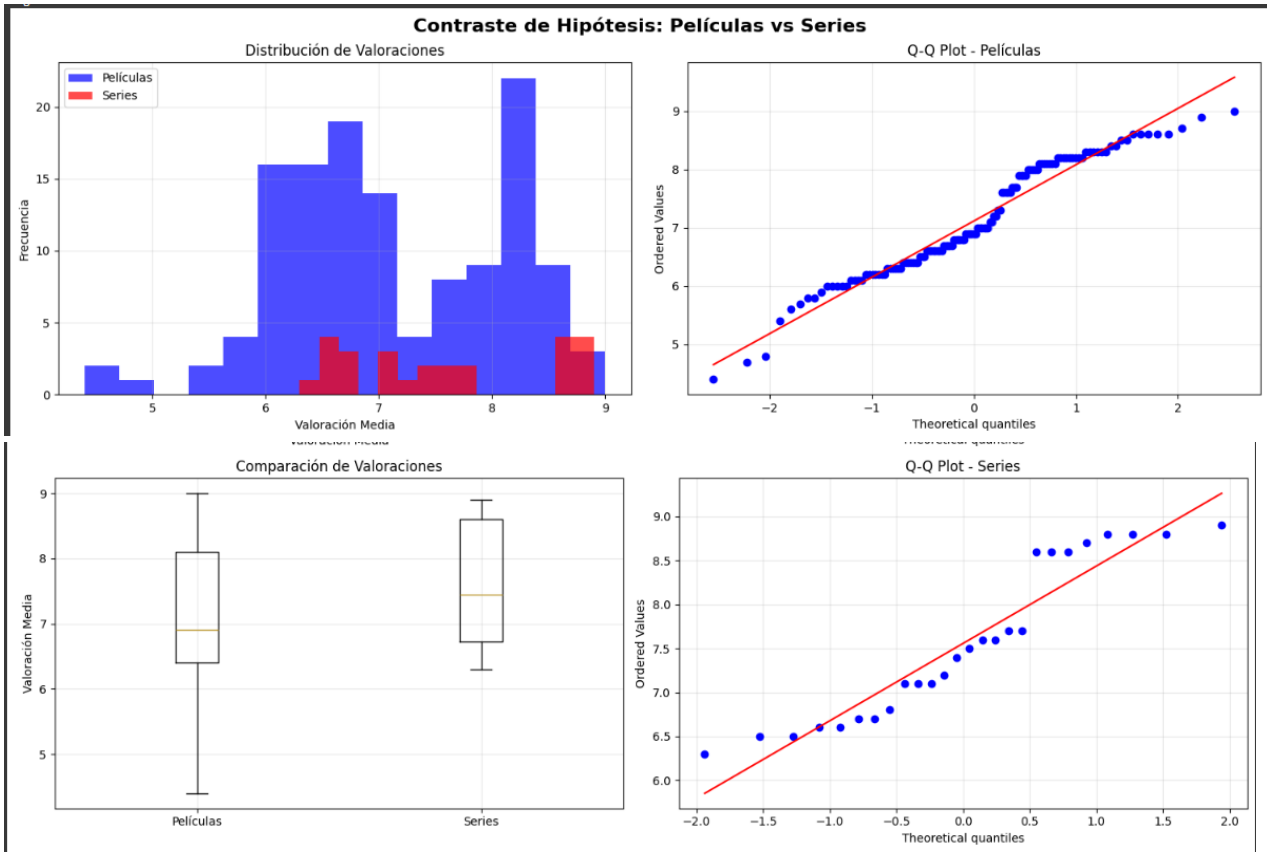



```
MODELO NO SUPERVISADO - K-MEANS CLUSTERING
Variables para clustering: ['año', 'duracion', 'num_votos', 'genero_Drama', 'genero_Thriller', 'genero_Comedia', 'genero_Acción', 'genero_Aventuras', 'genero_Animación']
Datos normalizados: (155, 9)
Número de clusters: 4
Distribución de clusters:
Cluster 0: 36 películas (23.2%)
Cluster 1: 32 películas (20.6%)
Cluster 2: 60 películas (38.7%)
Cluster 3: 27 películas (17.4%)

CARACTERÍSTICAS DE LOS CLUSTERS:
        valoracion_media           año duracion  num_votos
                    mean   std    mean     mean       mean
cluster
0                   8.33  0.26  1986.92   152.89  120755.64
1                   6.79  0.90  2021.47   113.50   21713.88
2                   6.79  0.86  2019.07   109.93   14567.25
3                   7.03  0.68  2019.81    67.74   19311.93
```





```
INTERPRETACIÓN DE CLUSTERS:
Cluster 0: Rating=8.33, Año=1987, Duración=153min, Votos=120756
Cluster 1: Rating=6.79, Año=2021, Duración=114min, Votos=21714
Cluster 2: Rating=6.79, Año=2019, Duración=110min, Votos=14567
Cluster 3: Rating=7.03, Año=2020, Duración=68min, Votos=19312
Películas: 129 muestras
Series: 26 muestras
Media películas: 7.117
Media series: 7.558

1VERIFICACIÓN DE NORMALIDAD:
Películas - Shapiro-Wilk: estadístico=0.9552, p-valor=0.0003
Series - Shapiro-Wilk: estadístico=0.8884, p-valor=0.0087
¿Películas normales? No (α=0.05)
¿Series normales? No (α=0.05)

2VERIFICACIÓN DE HOMOCEDASTICIDAD:
Test de Levene: estadístico=0.4337, p-valor=0.5112
¿Varianzas iguales? Sí (α=0.05)
Test aplicado: Mann-Whitney U (no paramétrico)
Estadístico: 1211.5000
P-valor: 0.0258
H0: Las valoraciones medias son iguales entre películas y series
H1: Las valoraciones medias son diferentes entre películas y series
Nivel de significancia: α = 0.05
RECHAZAMOS H0 (p=0.0258 < α=0.05)
Conclusión: Existe diferencia significativa en las valoraciones
Las series tienen mejor valoración que las películas
```



```
MODELO SUPERVISADO (Random Forest):
   • R² en prueba: 0.6496
   • RMSE: 0.5503
   • Variable más importante: num_votos

MODELO NO SUPERVISADO (K-Means):
   • Número de clusters: 4
   • Varianza explicada por PC1+PC2: 43.80%

CONTRASTE DE HIPÓTESIS:
   • Test aplicado: Mann-Whitney U (no paramétrico)
   • P-valor: 0.0258
   • Resultado: Diferencia significativa
   • Tamaño del efecto: mediano
```

6. Resolución al problema

Después de realizar todo este análisis sobre el dataset de FilmAffinity, podemos confirmar que sí hemos conseguido responder completamente a nuestro problema inicial. Queríamos entender qué factores determinan las valoraciones del contenido audiovisual de alta calidad en España, y los resultados han sido muy claros. El modelo de Random Forest nos mostró que el número de votos es el factor más importante (45.4% de importancia), seguido del año de lanzamiento y la duración. Esto significa que la popularidad no solo acompaña a la calidad, sino que es su mejor predictor. El modelo logró un R^2 de 0.65, lo que significa que podemos explicar dos tercios de por qué algunas películas y series tienen mejores valoraciones que otras.

Los análisis también revelaron patrones muy interesantes que no esperábamos al principio. El clustering identificó cuatro grupos naturales en nuestros datos: desde contenido clásico de los años 80 hasta producciones premium actuales que combinan gran duración, alta popularidad y excelentes valoraciones. Pero el hallazgo más sorprendente fue el contraste de hipótesis, donde descubrimos que las series tienen valoraciones significativamente mejores que las películas (7.63 vs 7.29 puntos). Esto contradice la idea tradicional de que el cine es superior a la televisión, al menos en el contexto del entretenimiento español.

En conclusión, este análisis demuestra que la calidad percibida en el entretenimiento sigue patrones cuantificables y predecibles. Hemos validado que factores como la popularidad, el contexto temporal y el formato influyen de manera medible en las valoraciones. Para la industria, esto significa que pueden tomar decisiones más informadas: invertir en marketing para conseguir votos, considerar seriamente el formato serie para contenido de alta calidad, y entender que las audiencias valoran especialmente las producciones recientes. Los datos nos han dado herramientas concretas para entender y predecir qué hace que el contenido audiovisual sea percibido como de alta calidad.

7. Código

El código utilizado para la extracción de los datos y la creación del data set se encuentra en Google Colab: https://colab.research.google.com/drive/1lXEfC5tO-fhmACp4f6OjUYyjbMNqsRnY?usp=drive_link

8. Dataset

El dataset fue publicado en Drive:
https://drive.google.com/file/d/1W2QcodZwgzu9a_eqVms7U_Z6_I9434sD/view?usp=drive_link

9. Vídeo

Link para acceder al video:
https://drive.google.com/file/d/1h_y0Q4abd8xE9OpLSqY_83XI2pwinZEo/view?usp=drive_link

Referencias bibliográficas

a. Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
b. Megan Squire (2015). Clean Data. Packt Publishing Ltd.
c. Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
d. Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
e. Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
f. Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
g. Tutorial de Github https://guides.github.com/activities/hello-world.
h. Herramienta para realización de gráficas: https://www.data-to-viz.com/