Trabajo Final Lenguajes 2025

Alumnos:

Jauregui Lorda, María Lina

Lara, Gonzalo Francisco

Fecha:

14/12/2025

Objetivos

Objetivo general

Aplicar técnicas de análisis de datos y programación para extraer conocimiento significativo a partir de datasets reales, integrando conceptos vistos a lo largo de la materia.

Objetivos específicos

- Aplicar técnicas de **preprocesamiento y limpieza de datos**.
- Utilizar **estadística descriptiva y visualizaciones** para responder preguntas de interés.
- Evaluar **relaciones entre variables clave** como género, rating, presupuesto, revenue y duración.
- Elaborar un **informe académico con estructura formal**.
- Publicar los resultados principales mediante una **mini-API local**, integrando nociones básicas de ingeniería de datos.

Ejes de análisis

El análisis se estructuró en torno a cuatro preguntas principales:

1. ¿Cuál es la **rentabilidad (ROI)** de las películas y cómo varía según el género?
2. ¿Existe relación entre **budget, revenue, popularidad y puntuación promedio**?
3. ¿Cómo fue la **evolución de la duración de las películas** en los últimos 50 años?
4. ¿Cuáles son los **directores con mejor rating promedio**?

Metodología

Exploración inicial de los datasets

Se analizaron ambos datasets para comprender:

- Tipo de información disponible.
- Tamaño y estructura.
- Variables relevantes.
- Posibles relaciones entre tablas.

El dataset principal fue movies_df, por su mayor volumen y riqueza de variables.

Limpieza y preprocesamiento de datos

Las principales etapas del preprocesamiento fueron:

- Eliminación de valores nulos y atípicos en las variables **budget** y **revenue**.
- Selección de columnas relevantes para mejorar la legibilidad del análisis.
- Filtrado de películas con estado **"Released"**.
- Procesamiento de la columna **genres**, que presentaba una estructura tipo diccionario:
    - Eliminación de la columna original.
    - Creación de variables binarias (1/0) para cada género.

El resultado de este proceso fue la creación de un **dataset enriquecido (nivel gold)** listo para análisis.

Procesamiento del dataset de créditos

Del dataset credits_df se conservaron únicamente:

- El identificador de la película.
- El director.

Se implementó una función para identificar al director dentro de la columna crew y se realizó la limpieza correspondiente.

Integración de datos

Finalmente, ambos datasets fueron integrados mediante un **INNER JOIN**, obteniendo un dataset final sobre el cual se desarrollaron todas las preguntas de negocio.

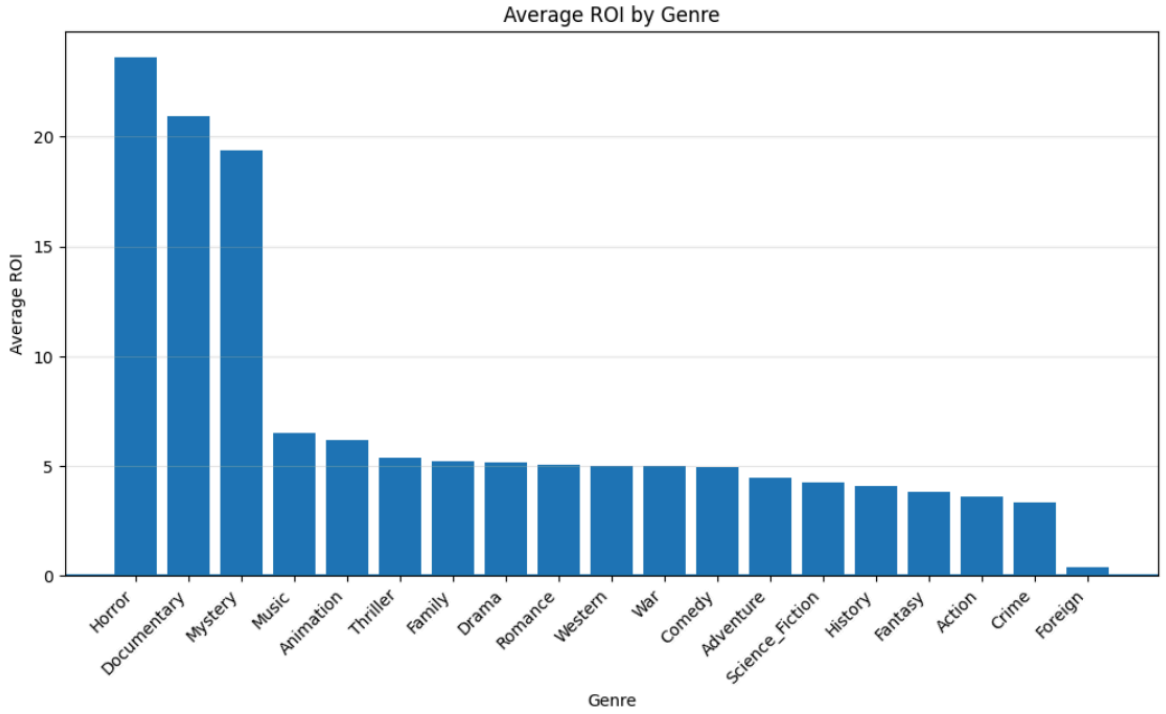
Resultados y discusión

1- Rentabilidad (ROI) y ROI por género

El **ROI** fue definido como:

ROI = revenue / budget

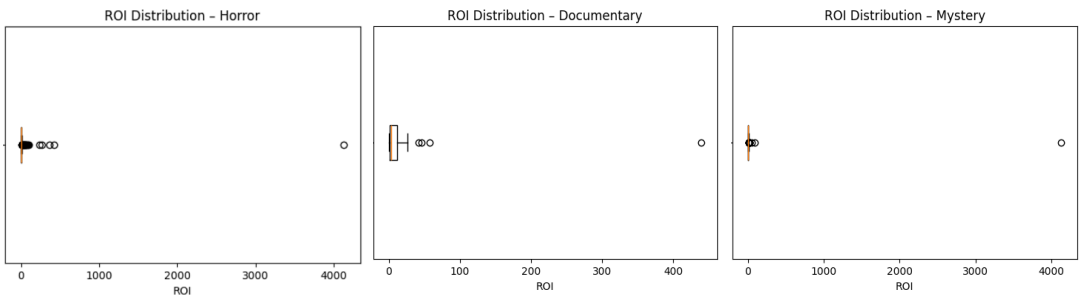
Primero se calculó el ROI para cada película y luego se agruparon los resultados por género.



El análisis inicial mostró valores extremadamente altos en tres géneros:

- Horror
- Documentary
- Mystery

Un análisis de outliers reveló que dos películas distorsionaban significativamente los resultados:



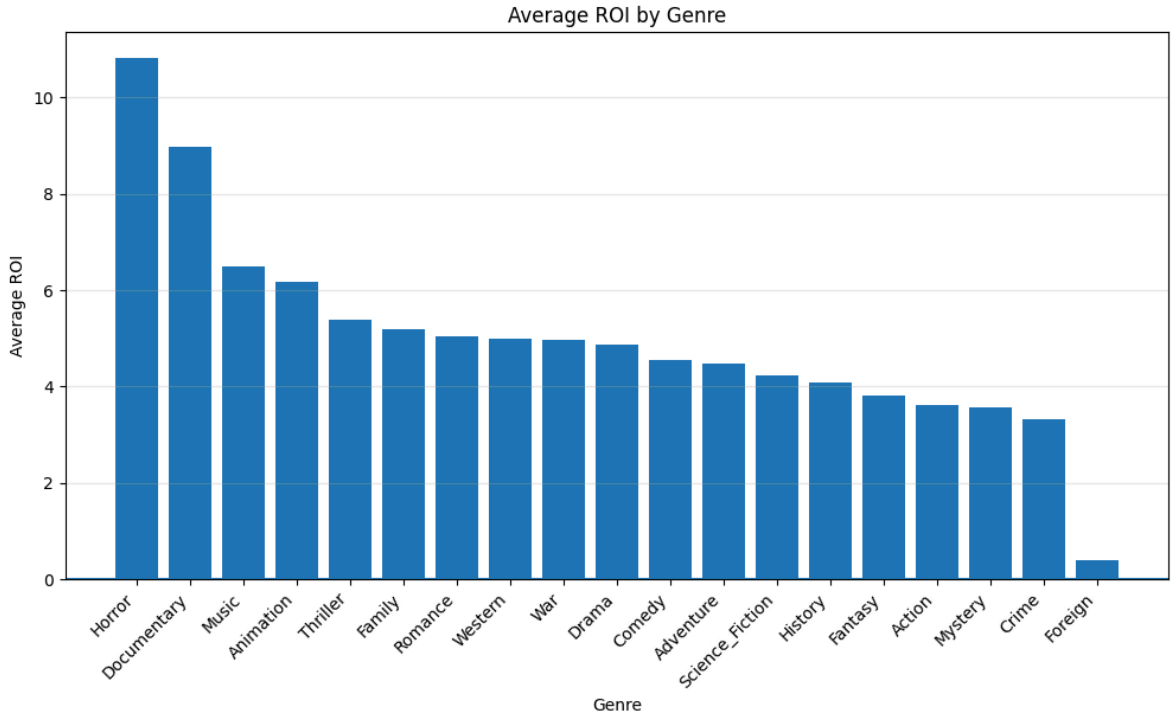
- *The Blair Witch Project*

	movie_id	title	ROI
3116	2667	The Blair Witch Project	4133.333333
3056	176	Saw	86.593058
3170	473	Pi	53.685867

- *Super Size Me*

	movie_id	title	ROI
3171	9372	Super Size Me	439.616585
3117	9459	Woodstock	57.508517
3083	1781	An Inconvenient Truth	46.243000

Al excluir estos títulos, los resultados se volvieron más representativos. Aun así, el género **Horror** continuó destacándose por su alto ROI promedio, lo que puede explicarse por sus bajos presupuestos y altos retornos relativos.



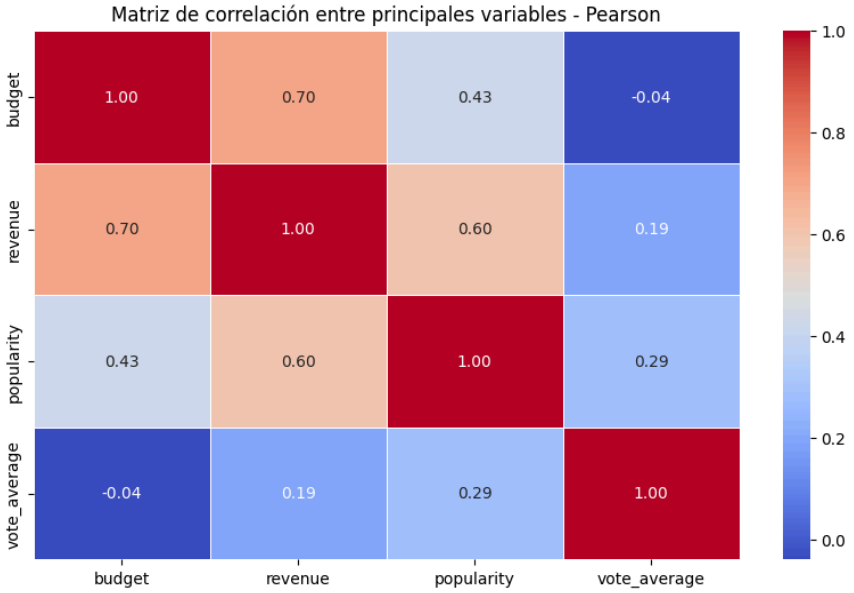
2- Relación entre budget, revenue, popularidad y rating

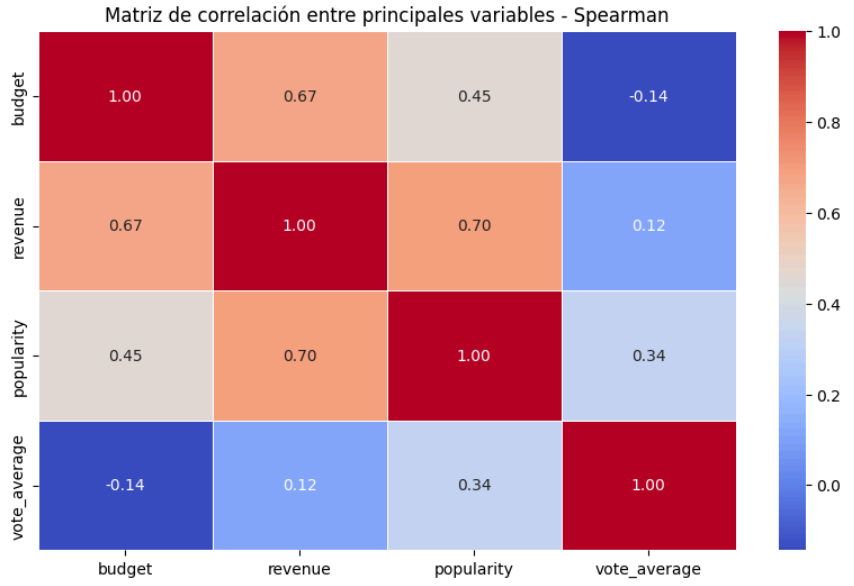
Se construyó una **matriz de correlación**, utilizando tanto el coeficiente de **Pearson** como el de **Spearman**.

Los principales hallazgos fueron:

- Existe una **correlación positiva fuerte** entre **budget y revenue**.
- El **revenue** también presenta una correlación positiva con la **popularidad**.
- No se encontró una relación significativa entre:
    - Budget y puntuación promedio.
    - Revenue y puntuación promedio.

Esto sugiere que el éxito económico no necesariamente está asociado a una mejor valoración crítica.

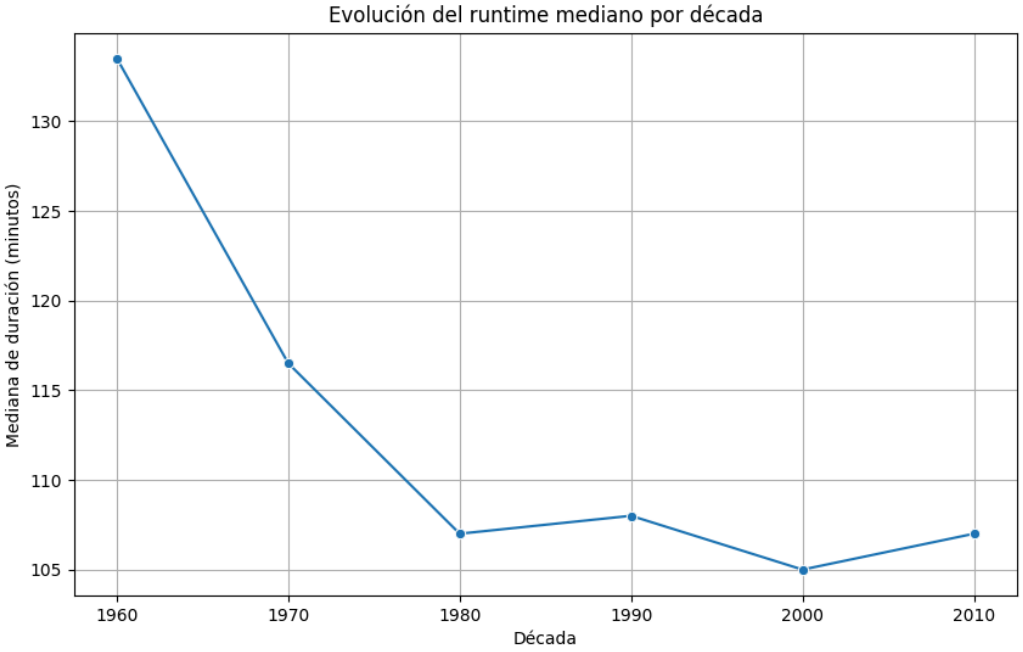




3- Evolución de la duración de las películas

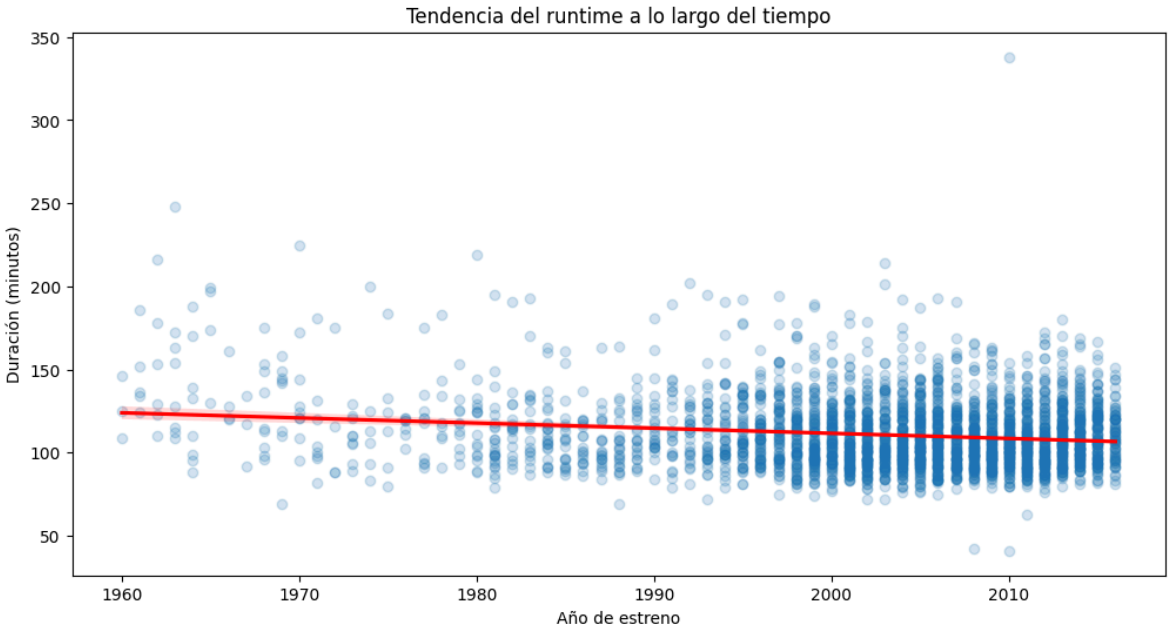
Se validó previamente el formato de fechas y luego se analizaron las duraciones promedio por década.

Un primer gráfico de líneas mostró ruido en décadas tempranas (especialmente los años 60), debido a una **baja cantidad de observaciones**.



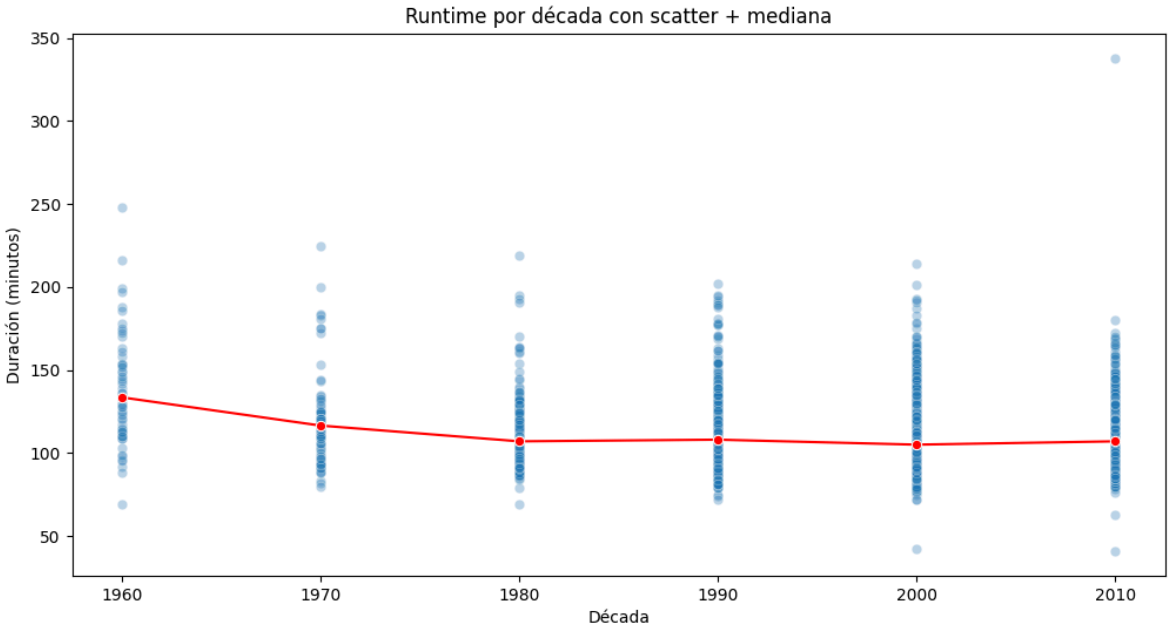
Para mejorar el análisis se utilizaron:

- Gráficos de dispersión.
- Regresión lineal.



- Análisis de la **mediana** de duración por período.

Estos enfoques permitieron identificar la tendencia real sin que los valores extremos distorsionaran los resultados.

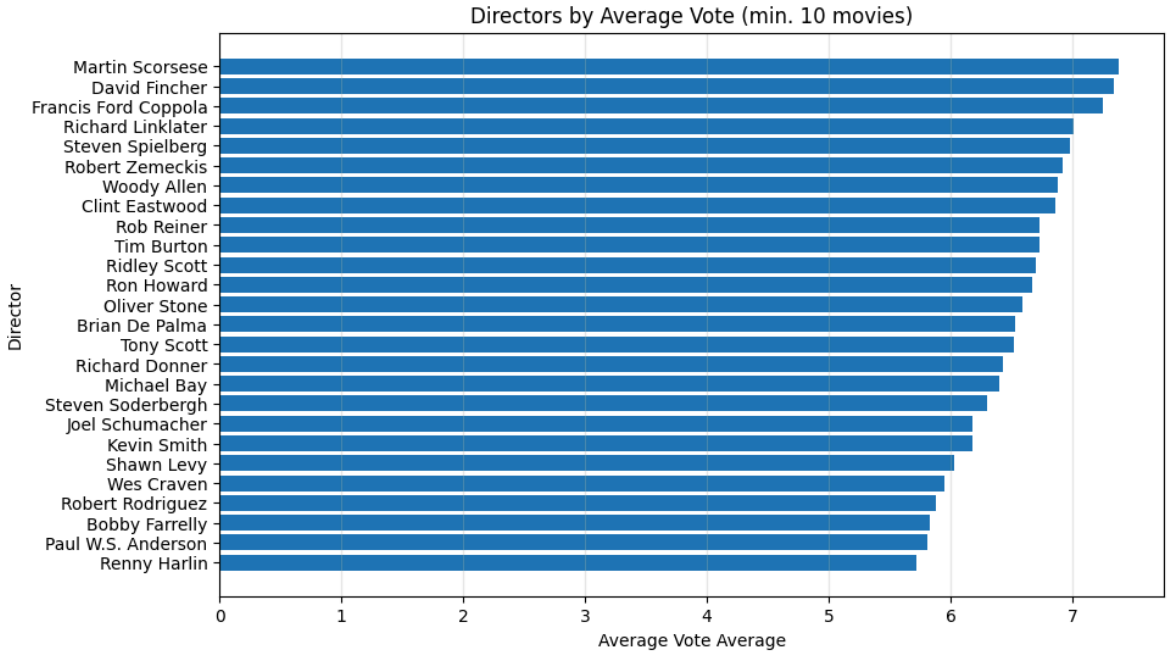


4- Directores con mejor rating promedio

Se agruparon los directores según la cantidad de películas dirigidas y se estableció un umbral mínimo de **10 películas lanzadas**.

director	
Steven Spielberg	27
Clint Eastwood	19
Ridley Scott	16
Robert Rodriguez	16
Martin Scorsese	16
Renny Harlin	14
Steven Soderbergh	14
Tim Burton	14
Robert Zemeckis	13
Oliver Stone	13
Michael Bay	12

Sobre este subconjunto se calculó el **rating promedio**, que luego fue visualizado gráficamente, permitiendo identificar a los directores más consistentes en términos de valoración.



Conclusiones

A partir del análisis realizado, se pueden extraer las siguientes conclusiones:

- El **ROI es altamente dependiente del género**, destacándose el horror como uno de los más rentables, incluso luego de eliminar outliers.
- Un **mayor presupuesto aumenta las probabilidades de mayor revenue**, pero no garantiza mejores valoraciones por parte del público o la crítica.
- La **popularidad** se relaciona más con el desempeño económico que con el rating promedio.
- La duración de las películas no presenta una variación drástica en los últimos 50 años, aunque el análisis debe realizarse con cuidado para evitar sesgos por baja cantidad de datos.
- Analizar directores con un volumen mínimo de obras permite obtener métricas más estables y comparables.

En conjunto, este trabajo demuestra cómo el uso de técnicas de EDA, visualización y modelado básico permite transformar datos crudos en información valiosa, integrando de manera práctica conceptos de programación, estadística y análisis de datos.

Bibliografía

- https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson
- https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Spearman
- https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata