PROYECTO FINAL DE LENGUAJES

Título del trabajo: Análisis Exploratorio de Datos de Películas
Materia: Lenguajes 2025
Integrantes: Román Raffo, Joaquín Del Percio.
Fecha: *7/12/2025*
**Universidad Católica de La Plata (UCALP)*

1. Introducción y objetivos

El objetivo de este trabajo es aplicar técnicas de análisis exploratorio de datos (EDA) sobre un dataset de películas obtenido desde Kaggle. A partir de este dataset realizamos distintos análisis estadísticos descriptivos y visualizaciones con Python, utilizando librerías como pandas, numpy, matplotlib y seaborn.

Además de las visualizaciones y estadísticas, se generaron resúmenes en archivos .csv que luego se exponen mediante una mini-API local desarrollada con Flask, cumpliendo con los requisitos de la materia.

Los ejes principales analizados fueron:

- Rentabilidad (ROI) por género.
- Relación entre presupuesto y rating.
- Evolución de la duración de las películas en los últimos 50 años.
- Distribución del rating por idioma.

El objetivo general es comprender mejor las características del dataset y extraer conclusiones que permitan interpretar patrones en la industria cinematográfica.

2. Metodología

2.1. Lectura y limpieza del dataset

El dataset original contenía variables como: título, género, país, presupuesto, recaudación, duración, idioma original, director, actores, rating, fechas, entre otras.

Se realizaron los siguientes pasos:

- Conversión de tipos: fechas a datetime, variables numéricas a float o int.
- Estandarización: nombres de columnas en minúsculas y con guiones bajos.
- Tratamiento de nulos:
    - En presupuesto y revenue, se filtraron valores cero o nulos para evitar ROIs inválidos.
    - En idioma y género, se reemplazaron nulos por “Unknown”.
- Creación de columnas derivadas:
    - roi = revenue / budget
    - decade para análisis por décadas (ej. 1990, 2000, etc.)
- Filtrado temporal: Para los análisis de evolución, se consideraron solo películas desde los años 70 en adelante.

3. Resultados y discusión

3.1. Rentabilidad (ROI) por género

Se calculó ROI para cada película y luego se analizaron los promedios por género.

Hallazgos principales:

- Algunos géneros como horror y documental suelen tener alto ROI, probablemente porque requieren presupuestos bajos.

- Los géneros como acción y aventura tienen ROI más moderado debido a presupuestos mucho más altos.
- Hay géneros muy irregulares (como ciencia ficción) donde hay casos extremadamente rentables, pero también pérdidas grandes.

Interpretación:
La rentabilidad no está exclusivamente asociada a la popularidad del género, sino al equilibrio entre presupuesto y retorno. Los géneros "baratos" tienden a ganar en promedio.

3.2. Relación entre presupuesto y rating

Se calcularon correlaciones Pearson y Spearman y se generó un gráfico de dispersión.

Resultados:

- La correlación entre presupuesto y rating es muy baja, cercana a cero.
- Visualmente, hay películas de bajo presupuesto con ratings muy altos y películas enormes con ratings mediocres.

Interpretación:
Un presupuesto alto puede mejorar la producción, pero no garantiza calidad percibida por el público. La relación entre dinero invertido y calidad percibida es prácticamente nula.

3.3. Evolución de la duración de películas (últimos 50 años)

Se analizó el runtime promedio por década.

Hallazgos:

- Desde 1970 hasta 2000 la duración promedio fue subiendo de manera constante.

- Entre 2010 y 2020 se observa estabilidad, con una leve tendencia a películas un poco más largas.
- Los outliers pertenecen principalmente a géneros épicos o documentales extensos.

Interpretación:
Las películas actuales tienden a ser ligeramente más largas que las de décadas anteriores, aunque la diferencia no es tan marcada como suele creerse.

3.4. Distribución del rating por idioma

Se agruparon películas por idioma original y se calculó la distribución del rating.

Resultados:

- El inglés domina numéricamente, pero otros idiomas como japonés, coreano y francés muestran promedios relativamente altos.
- Idiomas minoritarios tienen mucha variabilidad porque existen pocas películas en el dataset.

Interpretación:
Los idiomas con menor representación muestran más variabilidad, pero no necesariamente peores resultados. Idiomas con industrias fuertes (como Japón o Corea) poseen promedios competitivos incluso comparados al inglés.

4. Mini-API local

Se desarrolló una API con Flask que expone resultados en formato JSON.
Los endpoints principales son:

- /roi → ROI promedio por género

- /correlaciones → correlaciones entre presupuesto y rating
- /runtime → duración promedio por década
- /rating_idioma → distribución del rating por idioma

Los JSON se generan desde el notebook y se guardan en la carpeta salidas/.

5. Conclusiones

- La rentabilidad depende más del presupuesto que del género en sí.
- La relación entre presupuesto y rating es prácticamente inexistente.
- La duración de las películas ha aumentado levemente en las últimas décadas, pero no de manera explosiva.
- Los idiomas no dominantes presentan distribuciones variadas, pero no necesariamente peores; al contrario, algunos promedian muy bien.

En general, el análisis permitió comprender patrones importantes de la industria cinematográfica y aplicar herramientas fundamentales de análisis de datos, visualización y desarrollo de APIs.

6. Bibliografía y fuentes

- Dataset de Kaggle (Movies Dataset / TMDB / similar).
- Documentación oficial de:
    - pandas
    - matplotlib
    - seaborn
    - Flask
- Apuntes de cátedra y presentaciones de la materia.