Informe Final

Raffo Agustina

Eduardo Egea Mariana

Introducción

El presente trabajo forma parte del proyecto final de la materia Lenguajes 2025 de la UCALP, cuyo objetivo principal es aplicar herramientas de análisis de datos utilizando Python y librerías estándar como Pandas, Matplotlib y Seaborn. Para este estudio se trabajó sobre el dataset público "TMDB 5000 Movies", disponible en la plataforma Kaggle, que reúne información detallada acerca de miles de películas: títulos, presupuestos, recaudaciones, géneros, fechas de estreno, ratings, idiomas originales, elencos y directores, entre otros metadatos.

El propósito general del informe es comprender ciertos comportamientos y tendencias dentro de la industria cinematográfica, abordándolos desde una perspectiva descriptiva. En este caso particular, se seleccionaron tres ejes de análisis: la relación entre presupuesto y rating promedio, la evolución de la duración de las películas a lo largo de las décadas y la identificación de los directores con mejor rating promedio. Además del análisis, se generó una mini API local que expone parte de los resultados en formato JSON, integrando nociones básicas de ingeniería de datos.

Metodología

El trabajo comenzó con la carga de los archivos tmdb_5000_movies.csv y tmdb_5000_credits.csv, que aportan la información principal de cada película y su respectivo elenco y equipo técnico. Una vez incorporados al entorno, se realizó una primera exploración para conocer la estructura, la cantidad de filas y columnas, la existencia de valores faltantes y el tipo de información contenida en cada campo. Muchas columnas incluían listas y diccionarios codificados como texto (por ejemplo, "genres", "cast" y "crew"), por lo que fue necesario convertirlas nuevamente en estructuras de Python mediante ast.literal_eval.

La etapa de limpieza incluyó la conversión de fechas al formato datetime, el casteo de presupuestos y recaudaciones como valores numéricos, y el tratamiento de valores ausentes en variables relevantes como runtime. Además, se creó una columna que identifica el género principal de cada película y se extrajo el nombre del director desde la lista de miembros del equipo técnico

almacenada en crew. A partir de estas transformaciones, se generó un dataset unificado que permitió realizar los análisis seleccionados.

Cada eje de análisis se desarrolló utilizando técnicas de estadística descriptiva y visualización. Las relaciones entre variables se exploraron mediante correlaciones y gráficos, y en todos los casos se acompañó la interpretación con comentarios para contextualizar los resultados.

Resultados

Relación entre presupuesto y rating

Uno de los interrogantes más comunes dentro del cine es si invertir más dinero realmente garantiza una mejor recepción por parte del público o de la crítica. Para explorar esta idea, se calculó la correlación entre el presupuesto (budget) y el rating promedio (vote_average) utilizando los métodos de Pearson y Spearman. Ambas correlaciones resultaron bajas, lo que indica que las películas con presupuestos más altos no necesariamente obtienen mejores calificaciones.

Al visualizar los datos en un gráfico de dispersión con escala logarítmica, se observa un conjunto muy disperso, con películas de bajo presupuesto que alcanzan ratings superiores a los de producciones costosas. La tendencia general sugiere que, si bien el presupuesto puede influir en la escala de producción, no es un predictor confiable de la calidad percibida por los usuarios.

Evolución del runtime por década

Para analizar cómo cambió la duración de las películas a lo largo del tiempo, se extrajo el año de estreno y se lo agrupó por décadas. Luego se calcularon estadísticas como la mediana, que resulta más robusta ante valores extremos. El análisis muestra que, en general, las películas han tendido a alargarse con el paso de las décadas, especialmente desde los años 80 en adelante.

Si bien hay variaciones entre décadas, la tendencia ascendente es bastante clara: la industria pasó de producciones más cortas y compactas a películas con duraciones más extensas, posiblemente relacionadas con cambios narrativos, avances tecnológicos y nuevas expectativas del público. El gráfico de línea construido para este análisis ayuda a visualizar esta tendencia de manera clara y progresiva.

Directores con mejor rating promedio

Finalmente, se elaboró un ranking de directores con mejor rating promedio. Para evitar resultados engañosos, se estableció un mínimo de películas por director. Una vez filtrado el conjunto, se calcularon los promedios de vote_average para cada uno. El resultado fue una lista de directores con un rendimiento consistentemente alto, lo que permite identificar creadores con trayectoria sólida según las evaluaciones del público. La tabla obtenida destaca a directores cuyos trabajos mantienen un estándar de calidad elevado, aunque también debe considerarse que los ratings pueden estar influidos por factores externos como el género, la popularidad del elenco o el tipo de audiencia.

Mini API Local

Como parte integradora del trabajo, se desarrolló una mini API utilizando FastAPI. Esta API permite acceder de manera rápida y ordenada a tres conjuntos de resultados: las correlaciones entre presupuesto y rating, la evolución del runtime por década y el ranking de directores. Cada uno de estos análisis fue exportado previamente a archivos JSON, que la API devuelve al realizar solicitudes a sus respectivos endpoints. Esta implementación sirve como un ejercicio introductorio al desarrollo de servicios de datos y demuestra cómo conectar análisis offline con consultas dinámicas.

Conclusiones

El análisis realizado permite afirmar que el presupuesto de una película no determina su éxito en términos de rating, lo cual es coherente con la naturaleza subjetiva de la experiencia cinematográfica. También se observa que la duración de las películas ha aumentado sostenidamente con el paso del tiempo, marcando cambios culturales y tecnológicos dentro del cine. Por último, el ranking de directores con mejor rating evidencia que ciertos creadores logran mantener un nivel alto de aceptación, aunque siempre condicionado a diversos factores externos.

La construcción de la API complementa el trabajo al permitir que los resultados puedan ser consultados de manera modular y reutilizable, mostrando una aplicación directa del análisis de datos en un entorno programático. En conjunto, el proyecto integra conceptos de procesamiento de datos, análisis exploratorio, visualización y exposición de resultados, ofreciendo una mirada completa del flujo de trabajo típico en un contexto de análisis aplicado.