📊 Análisis de Visionados: Exploración y Modelado de Datos

Este proyecto tiene como objetivo explorar el catálogo de contenidos de Netflix y sus tendencias de consumo mediante el uso de Python para la preparación de datos y Power BI para la visualización interactiva.
A través del análisis se busca comprender patrones de producción, tipos de contenido, géneros predominantes y comportamiento del usuario en la plataforma para realizar predicciones basadas en datos usando técnicas de modelado estadístico.

🗂️ Estructura del Proyecto Pfinal

```
├── data/ # Datos crudos y procesados
├── notebooks/ # Notebooks de Jupyter con el análisis y Scripts de procesamiento
├── results/ # Gráficos y archivos de resultados
├── README.md # Descripción del proyecto e informe
```

🛠️ Instalación y Requisitos

Este proyecto usa Python 3.13.5 y requiere las siguientes bibliotecas:
- pandas
- numpy
- matplotlib
- seaborn

📊 Resultados y Conclusiones

Se realizó un proceso de limpieza y transformación de datos que incluyó:

-Estandarización de los nombres de las columnas (que empiecen por mayúscula)

-Unificación de los ID para más adelante hacer merge

-Se trabajó con dos conjuntos de datos: **df1**, centrado en la información del contenido (películas y series), y **df2**, con el registro de actividad de los usuarios.

-Se analizaron los valores nulos y se aplicaron estrategias de imputación específicas:
-`genre_secondary`: se reemplazó con 'None' al ser información faltante no esencial.
-`imdb_rating`: se imputó con la media del campo.

-`production_budget` y `box_office_revenue`: en algunos casos se imputaron con 0, ya que representaban producciones sin datos financieros o no aplicables.

-En **df2**, las variables `watch_duration_minutes` y `progress_percentage` se completaron con la mediana agrupada por la acción de visualización (`action`), mientras que `user_rating` se completó con la media global.

Conclusiones

- El análisis del dataset **df1** permitió explorar la composición del catálogo de contenido. Se observó una proporción significativa de películas frente a series, y una alta concentración de géneros como drama, comedia y acción.
- -En términos económicos, se calculó el ROI (retorno de inversión) comparando `box_office_revenue` con `production_budget`, detectando una amplia variabilidad: algunas producciones superan ampliamente su presupuesto, mientras que otras no recuperan la inversión. Esto se ve en el gráfico**:** la mayoría de los puntos están **muy concentrados en la esquina inferior izquierda**, lo que indica que:
  -La mayoría de las películas/series tienen **presupuestos bajos o moderados** y también generan **ingresos bajos**.
  -Es típico: el mercado audiovisual tiene muchas producciones pequeñas y pocas superproducciones.

  -Se ven **unos pocos puntos muy altos (outliers)**, con ingresos de cientos de millones o más.
  -Esas son las **películas taquilleras** o "blockbusters". Suelen tener presupuestos grandes, pero también un ROI alto.

  -Aunque parece que a mayor presupuesto **tienden** a crecer los ingresos, **la correlación no es muy fuerte**. Gastar más **no garantiza** ganar más.
  -En el gráfico hay películas de bajo presupuesto que logran buenos ingresos y algunas caras que no recuperan su inversión.

- -En cuanto a si el tipo de dispositivo afecta a la acción de visualización se aplicó una prueba de Chi-cuadrado.

  El resultado ($\chi^2 = 6.65$, $p = 0.88$) indicó que **no hay evidencia estadísticamente significativa** de que el dispositivo afecte el tipo de acción de visualización, es decir, los patrones de interacción son similares sin importar el dispositivo empleado.

- El Dashboard Muestra tres KPI (Progreso medio de visualización: mide el grado promedio de completitud de los contenidos. - Valoración media del usuario: refleja la satisfacción general y du2ración media de visualización:** permite estimar el tiempo promedio invertido por sesión.

  La segunda fila responde a qué géneros predominan en el catálogo, producción por país de origen que son **Estados Unidos (USA)**, seguido por **India** y **Japón**. La distribución por género **Adventure**, **Animation**, **Comedy** son los más vistos y Recuento títulos por contenido Con películas y series dominando el mercado.

  La tercera fila responde a dispositivos más utilizados para ver contenido (**Desktop y Laptop**), evolución de lanzamientos por año, que hay un **Crecimiento Exponencial:** A partir del año **2000** aproximadamente, se observa un aumento drástico y sostenido en el número de títulos añadidos a la plataforma e **Irregularidad:** La línea muestra mucha irregularidad (subidas y bajadas agudas) después del 2000 lo que sugiere

una estrategia de adquisición y producción muy dinámica y no lineal año tras año, aunque la tendencia general es de fuerte crecimiento.

Y por último el gráfico de Netflix Original indica que la mayor parte del catálogo todavía está compuesta por contenido licenciado o de terceros (False).

No obstante, el contenido etiquetado como **Netflix Original (True)** ocupa una porción significativa, reflejando la considerable inversión de la plataforma en producción propia.

🔄 Próximos Pasos

- Explorar el impacto de factores externos como campañas de marketing de las producciones.

-Realizar otro tipo de análisis predictivo como por ejemplo hacia que tipo de producciones se debería centrar la compañia.

-Incrementar la inversión en géneros con alta tasa de completitud y retorno financiero positivo y potenciar las producciones originales ya que tienen mejor rendimiento.

✒️ Autores

- Maria Cristina Martinez Gutierrez

- [@Mariacris155](https://github.com/Mariacris155)