ETL CLASS PROJECT – PHASE 1

Autor: Alejandro Arteaga Jaramillo

Código: 22500232

Problema y contexto:

La industria del cine ha experimentado numerosos cambios a lo largo de los años, impulsados por la evolución de las tecnologías de animación, la llegada de las plataformas de streaming y los cambios culturales que han obligado a la industria a adaptarse. Como resultado, las películas más taquilleras del pasado difieren significativamente de las actuales, aunque comparten ciertas características.

Por ello, el análisis de datos en esta industria es fundamental para comprender la evolución de las tendencias cinematográficas, identificar patrones de éxito y optimizar la toma de decisiones en producción, distribución y recomendación de contenido.

Un análisis exhaustivo de los datos puede ayudar a responder las siguientes preguntas clave:

1. ¿Qué factores influyen en el éxito de una película?

Es posible analizar cómo variables como el género, el presupuesto, la calificación del público, la duración e incluso la fecha de estreno impactan en el desempeño de una película en el mercado.

2. ¿Cómo han cambiado las preferencias del público?

La cultura del público ha evolucionado en la última década, lo que ha influido en el interés por ciertos géneros cinematográficos. Además, la llegada del streaming ha transformado los patrones de consumo, lo que permite identificar nuevas tendencias en la manera en que se disfrutan las películas.

3. ¿Cómo afectan el presupuesto y la estrategia de lanzamiento al éxito comercial?

En la actualidad, una estrategia de lanzamiento bien planificada puede ser determinante para el éxito de una película o serie. Además, el análisis de la relación entre la fecha de estreno y los ingresos generados permite identificar los períodos más propicios para lanzamientos exitosos.

4. ¿Cómo han evolucionado las críticas y la percepción del público?

Desde sus inicios, el cine ha sido objeto de críticas, tanto positivas como negativas. Anteriormente, solo un grupo selecto de expertos era considerado "crítico de cine", pero hoy en día cualquier persona con acceso a internet puede influir en la percepción de una película a través de plataformas como Rotten Tomatoes o IMDb. Analizar la evolución de estos estándares de calidad permite entender cómo ha cambiado la recepción del público a lo largo del tiempo.

Importancia del análisis:

Este tipo de análisis no solo beneficia a la industria cinematográfica, sino también a plataformas de streaming, productoras y analistas de datos que buscan optimizar la oferta de contenido y mejorar la experiencia del usuario mediante la implementación de modelos predictivos.

Descripción del dataset:

Para este análisis se utilizó el dataset "Full IMDb Movies Data", disponible en la plataforma Kaggle. Este dataset contiene información detallada sobre películas desde 1990 hasta la actualidad, lo que permite realizar un estudio exhaustivo de la industria cinematográfica en las últimas décadas.

El dataset consta de 21 columnas y 903,263 registros, con las siguientes variables principales:

1. id: A unique identifier for each movie.
2. title: The name of the movie.
3. vote_average: The average rating the movie has received from users (on a scale, typically from 0 to 10).
4. vote_count: The total number of votes or ratings submitted for the movie.
5. status: The current state of the movie (e.g., "Released," "Post-Production").
6. release_date: The date when the movie was officially released.
7. revenue: The total earnings the movie made (usually in USD).
8. runtime: The duration of the movie in minutes.
9. adult: Indicates whether the movie is classified as adult content (e.g., "True" or "False").
10. budget: The total cost of producing the movie (usually in USD).
11. imdb_id: The unique identifier for the movie on IMDb (Internet Movie Database).
12. original_language: The language in which the movie was originally produced (e.g., "en" for English).
13. original_title: The original title of the movie in its native language.
14. overview: A brief summary or description of the movie's plot.
15. popularity: A metric indicating how popular the movie is (typically based on views, searches, or ratings).
16. tagline: A short phrase or slogan associated with the movie.
17. genres: The categories or genres the movie belongs to (e.g., Action, Comedy, Drama).
18. production_companies: The names of the companies involved in producing the movie.
19. production_countries: The countries where the movie was produced.
20. spoken_languages: The languages spoken in the movie.
21. keywords: Important terms or phrases associated with the movie, often used for categorization or search.

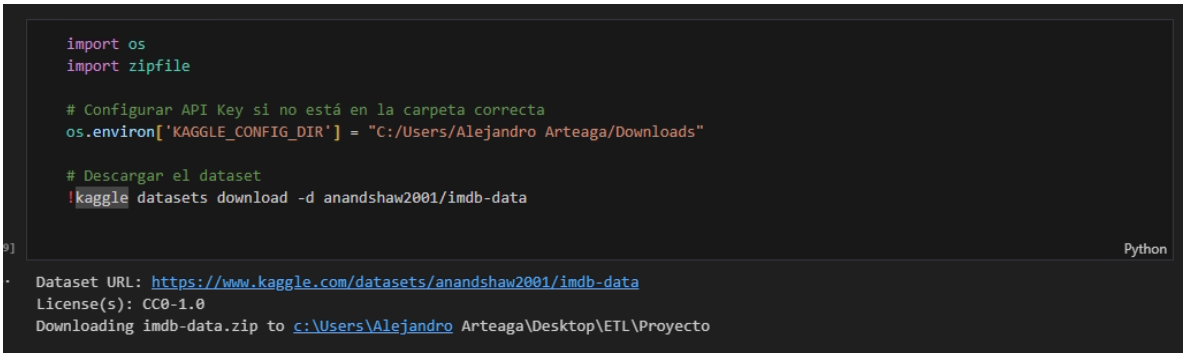
A partir de las categorías anteriores, es posible realizar diversos análisis aplicados, tales como:

- Identificar tendencias en la industria cinematográfica (géneros más populares, películas más taquilleras).
- Desarrollar sistemas de recomendación basados en géneros o valoraciones.
- Analizar la relación entre presupuesto e ingresos para evaluar la rentabilidad de las películas.
- Determinar los factores clave para el éxito de una película.
- Visualizar tendencias a lo largo del tiempo (estrenos de películas por año).

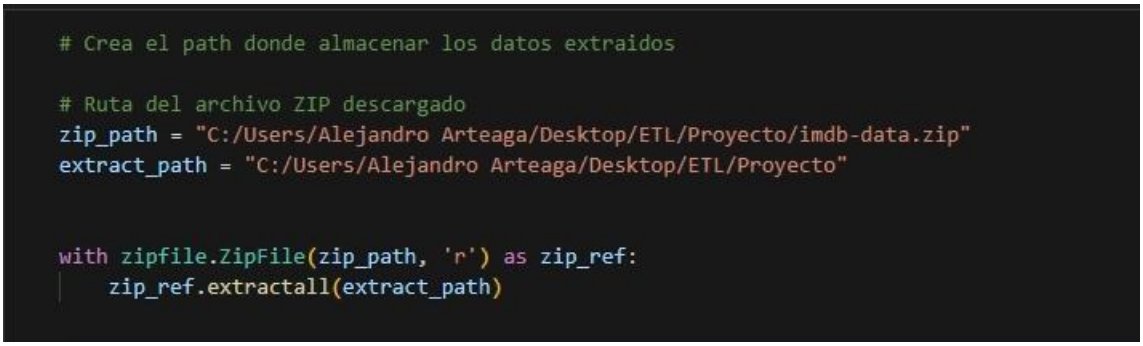
Un análisis de este tipo no solo es útil para estudios de cine, sino también para plataformas de streaming, inversionistas, investigadores y expertos en marketing. Dependiendo del enfoque, puede contribuir a la mejora en la producción de contenido, optimización de estrategias comerciales y una mejor comprensión de la evolución de la industria cinematográfica, proporcionando información clave sobre su futuro.

Process and evidence of data extraction:

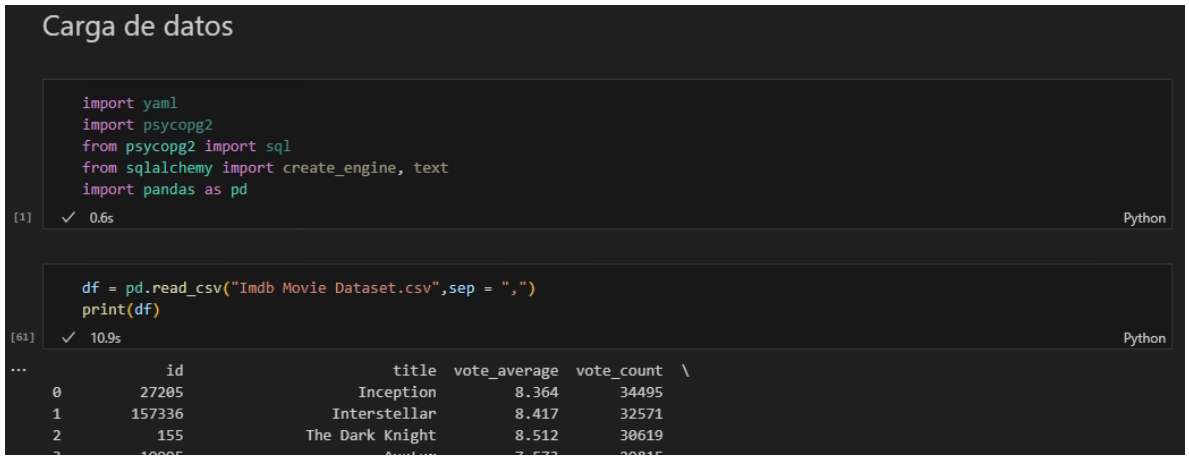
- Se configura la API key descargada desde el perfil de kaggle



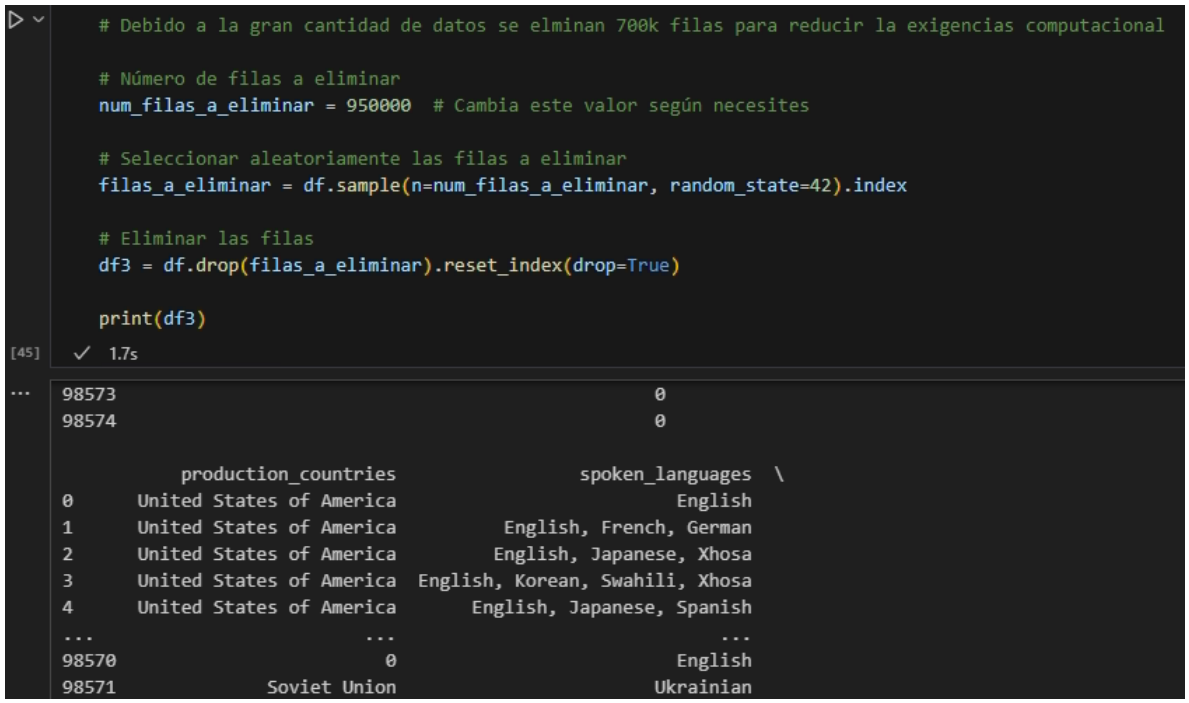
- Se descomprimen los datos descargados desde kaggle en la carpeta especificada.



- Se crea el dataframe (df) a partir del archivo .csv descomprimido



- El dataset cuenta con más de 1 millon de filas, para temas de análisis y reducción de carga computacional se toma la decisión de eliminar aleatoriamente una cantidad considerable de filas, dejando para el análisis un total de 50.000



- Se crea la database y la tabla donde se realizara la carga de datos.

```
    db_name = "etl_proyecto"
    try:
        with conn.cursor() as cur:
            cur.execute(sql.SQL("CREATE DATABASE {}").format(sql.Identifier(db_name)))
            print(f"Base de datos '{db_name}' creada exitosamente.")
    except psycopg2.errors.DuplicateDatabase:
        print(f"La base de datos '{db_name}' ya existe.")
```

19] ✓ 0.0s

```
La base de datos 'etl_proyecto' ya existe.
```

```
engine = create_engine(f"postgresql://{db_user}:{db_password}@{db_host}:{db_port}/{db_name}")

with engine.connect() as conn:
    conn.execute(text("""
        CREATE TABLE IF NOT EXISTS tabla_proyecto_ETL2 (

            id BIGINT,
            title VARCHAR(1000),
            vote_average FLOAT,
            vote_count BIGINT,
            status VARCHAR(255),
            release_date VARCHAR(1000),
            revenue BIGINT,
            runtime BIGINT,
            adult BOOL,
            budget BIGINT,
            imdb_id VARCHAR(1000),
            original_language VARCHAR(1000),
            original_title VARCHAR(1000),
            overview VARCHAR(1000),
            popularity FLOAT,
            tagline VARCHAR(1000),
            genres VARCHAR(1000),
            production_companies VARCHAR(1000),
            production_countries VARCHAR(1000),
            spoken_languages VARCHAR(1000),
            keywords VARCHAR(1000)

        );
    """))

    conn.commit()  # Asegúrate de confirmar los cambios
    print("Tabla 'tabla_proyecto_ETL2' creada exitosamente en PostgreSQL.")
```

- Debido a la longitud de unas columnas (Keywords) se restringen a un límite de caracteres, para evitar inconvenientes con la carga de datos, posteriormente se cargan los datos del dataframe a la tabla creada en la base de datos.

```
df3['keywords'] = df3['keywords'].astype(str).str[:1000]
```

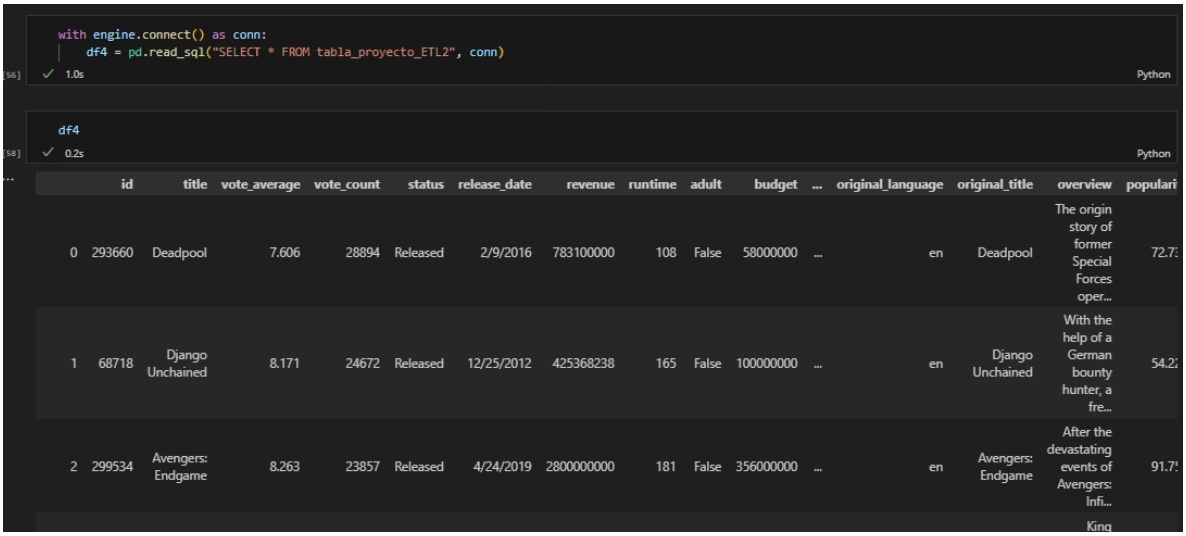
✓ 0.0s Python

```
engine = create_engine(f"postgresql://{db_user}:{db_password}@{db_host}:{db_port}/{db_name}")

with engine.connect() as conn:
        stmt = text("""
        INSERT INTO tabla_proyecto_ETL2 (id, title, vote_average, vote_count, status, release_date, revenue, runtime, adult, budget, imdb_id, original_language,
        VALUES (:id, :title, :vote_average, :vote_count, :status, :release_date, :revenue, :runtime, :adult, :budget, :imdb_id, :original_language, :original_tit
    """)
        conn.execute(stmt, df3.to_dict(orient="records"))
        conn. commit()


print("Los datos se cargaron exitosamente")
```

✓ 23.5s Python

```
Los datos se cargaron exitosamente
```

Por último se crea el dataframe4 (df4) el cual lee de la tabla creada los datos y los imprime.



```
with engine.connect() as conn:
    df4 = pd.read_sql("SELECT * FROM tabla_proyecto_ETL2", conn)
```

✓ 1.0s Python

```
df4
```

✓ 0.2s Python

	id	title	vote_average	vote_count	status	release_date	revenue	runtime	adult	budget	...	original_language	original_title	overview	populari
0	293660	Deadpool	7.606	28894	Released	2/9/2016	783100000	108	False	58000000	...	en	Deadpool	The origin story of former Special Forces oper...	72.7
1	68718	Django Unchained	8.171	24672	Released	12/25/2012	425368238	165	False	100000000	...	en	Django Unchained	With the help of a German bounty hunter, a fre...	54.2
2	299534	Avengers: Endgame	8.263	23857	Released	4/24/2019	2800000000	181	False	356000000	...	en	Avengers: Endgame	After the devastating events of Avengers: Infi...	91.7
														King	

De igual forma se rectifica esta info utilizando pgadmin4, donde se evidencia la creación de la base de datos, la tabla y se corrobora las dimensiones de la misma.