Tipología y ciclo de vida de los datos

Práctica 1

¿Cómo podemos capturar los datos de la web?
25% nota final

Fecha de entrega
12 de noviembre de 2024

Realizada por:

Jorge Moreno Fuentes

Jorge Martín Rota

Repositorio de datos:

https://github.com/jmf3192/UOC_PRACT1

1. Contexto:

Explicar en qué contexto específico se han recolectado los datos y argumentar por qué el sitio web seleccionado es una fuente pertinente y fiable de esa información. Indicar la dirección del sitio web.

Se ha elegido IMDb como fuente de datos para crear un conjunto de datos basado en las 250 mejores películas según su clasificación en el sitio web. IMDb (Internet Movie Database) es una plataforma reconocida y confiable en la industria del cine, que ofrece información detallada sobre películas, series de televisión, actores, directores, y otros elementos relacionados con la industria cinematográfica. La página utilizada para este proyecto es IMDb Top 250. Esta lista define las 250 películas mejor valoradas según los usuarios de IMDb.

IMDb es una buena fuente para este tipo de práctica de scraping porque ofrece datos públicos bien organizados y de fácil acceso. Se ha elegido IMDb por la fiabilidad de su información, por su popularidad en la industria y la gran cantidad de datos que proporciona. Esta fuente de datos permite explorar aspectos variados de las películas como su popularidad, los directores, los géneros y el reparto. Además, su estructura clara facilita el uso de técnicas de scraping sin requerir autenticación o el uso de APIs.

2. Título:

Definir un título conciso y que sea descriptivo para el dataset.

Análisis de las 250 mejores películas de IMDb: Extracción y exploración de datos del cine.

3. Descripción del dataset:

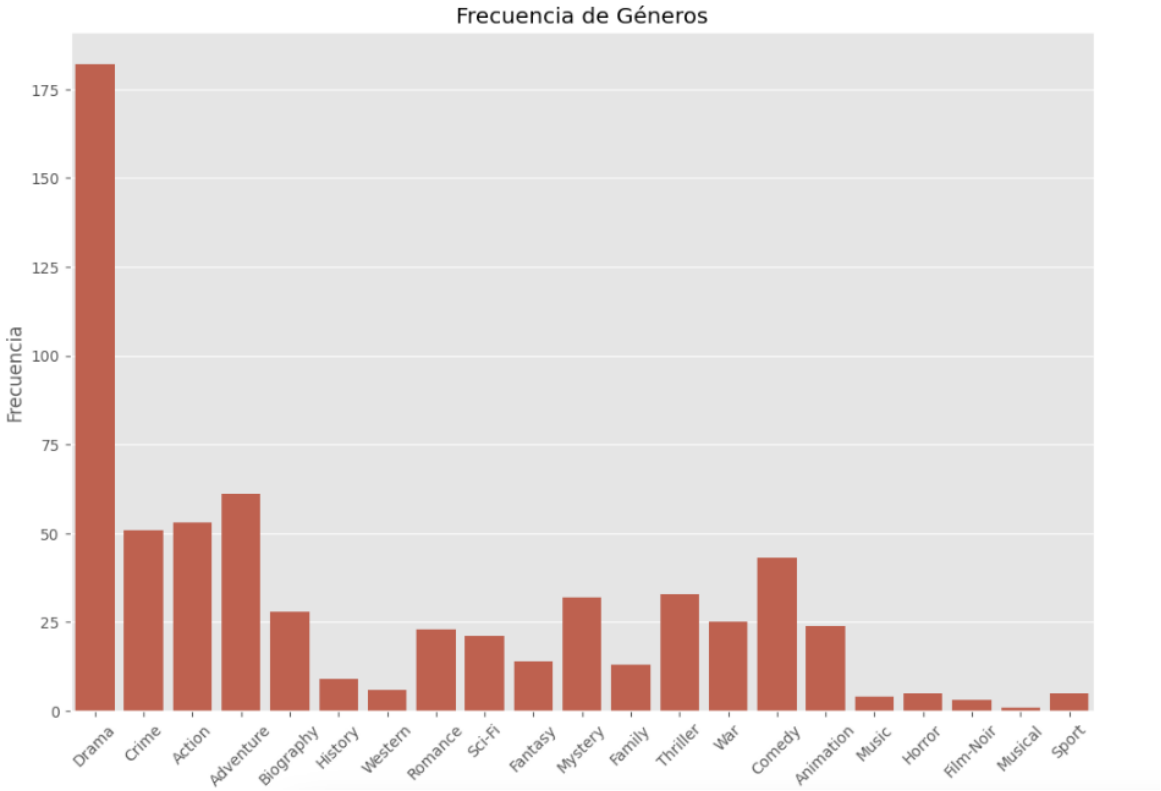
Desarrollar una breve descripción del conjunto de datos que se ha extraído. Es necesario que esta descripción sea coherente con el título elegido.

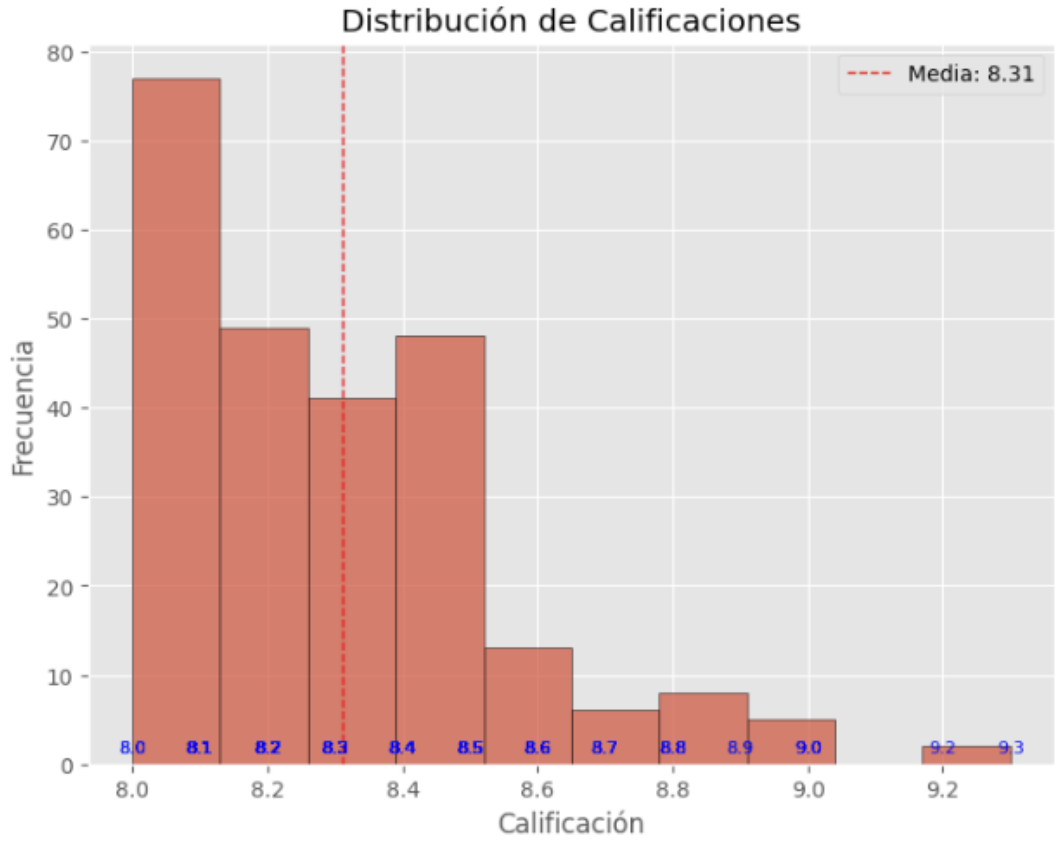
El dataset extraído contiene información detallada sobre las 250 películas mejor valoradas en IMDb. El objetivo es analizar diversos aspectos del cine y su impacto en la audiencia. Para cada película, se han recopilado datos como el título, la calificación, el género, la fecha de lanzamiento, el director, los escritores y los actores principales.

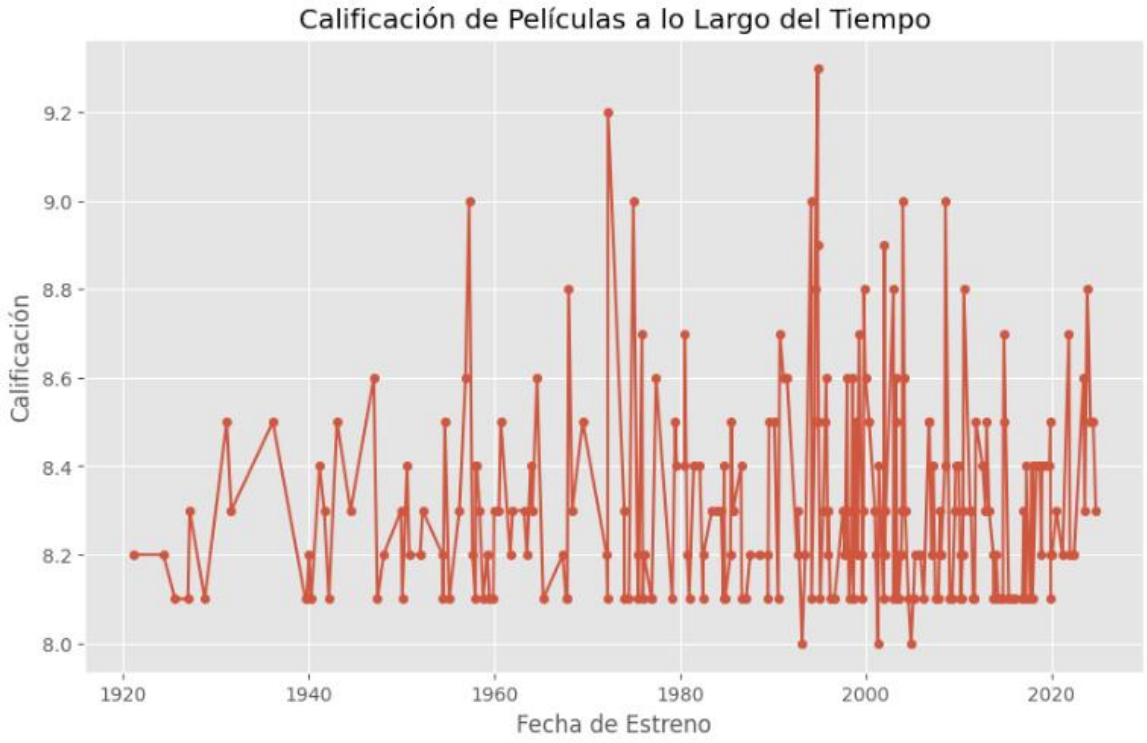
Estos datos permiten explorar patrones de popularidad, analizar el perfil de los directores y actores más recurrentes en el cine de alta valoración y entender las características que comparten las películas mejor clasificadas. Además, el dataset incluye enlaces directos a las páginas individuales de IMDb de cada película, lo que permite realizar futuras exploraciones o análisis adicionales. Este conjunto de datos proporciona una visión completa y organizada de las películas más influyentes según la comunidad de IMDb.

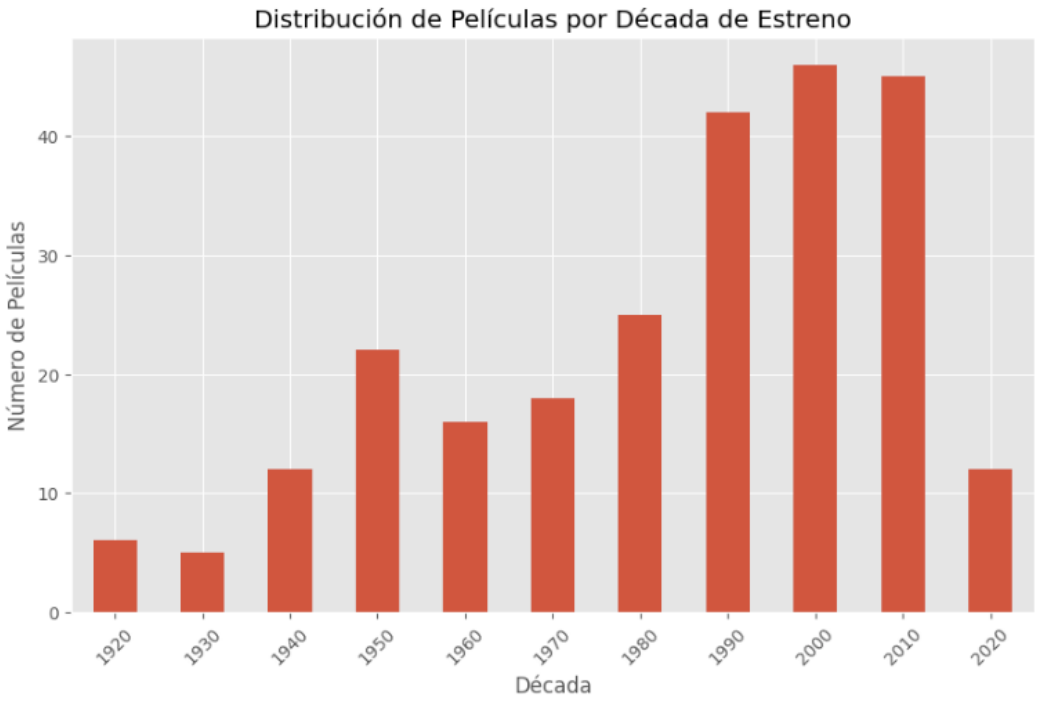
4. Representación gráfica:

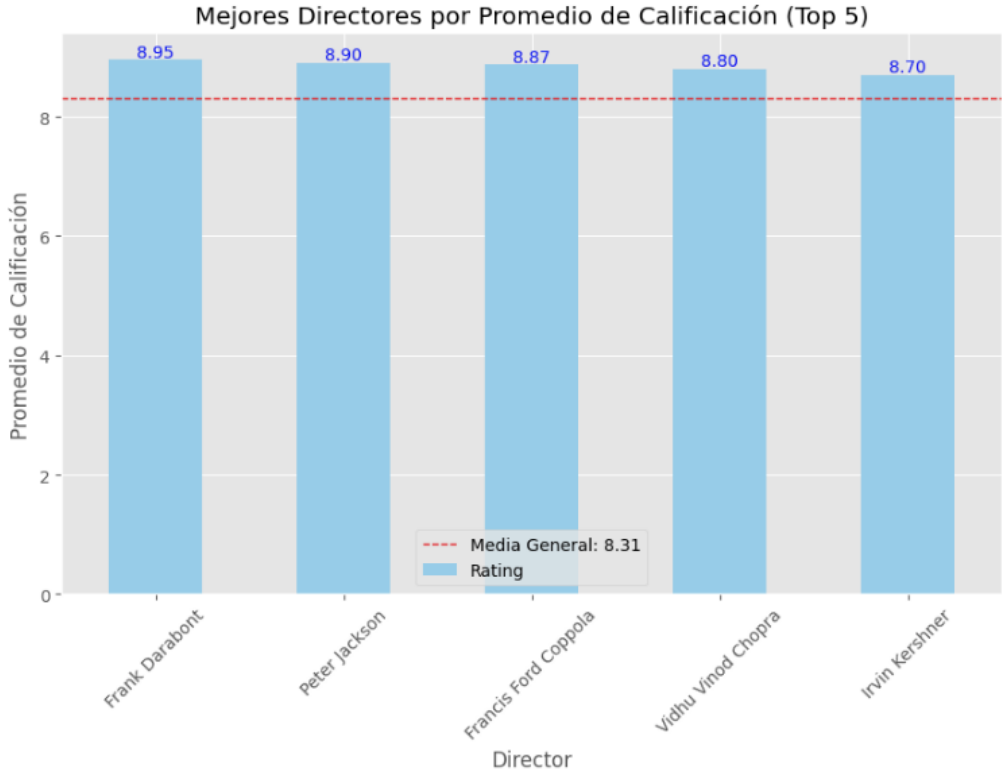
Dibujar un esquema o diagrama que refleje visualmente el dataset y el proyecto elegido

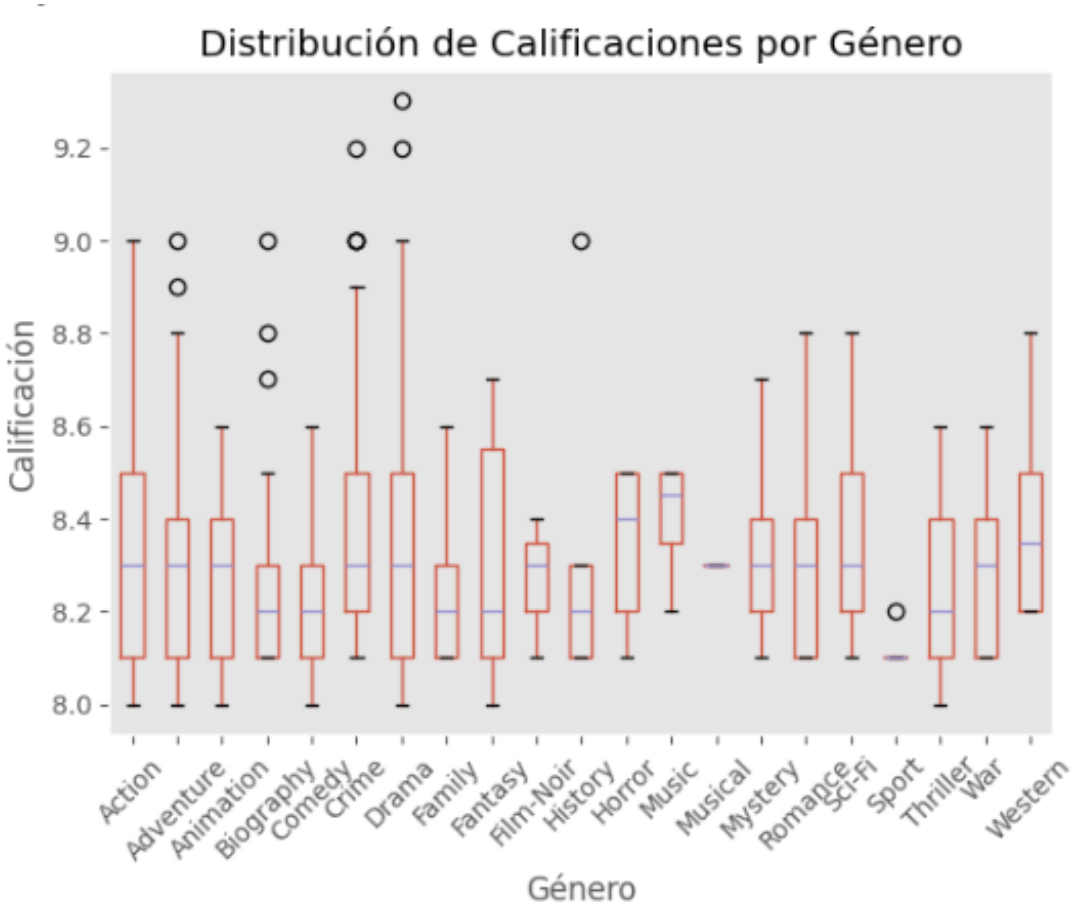


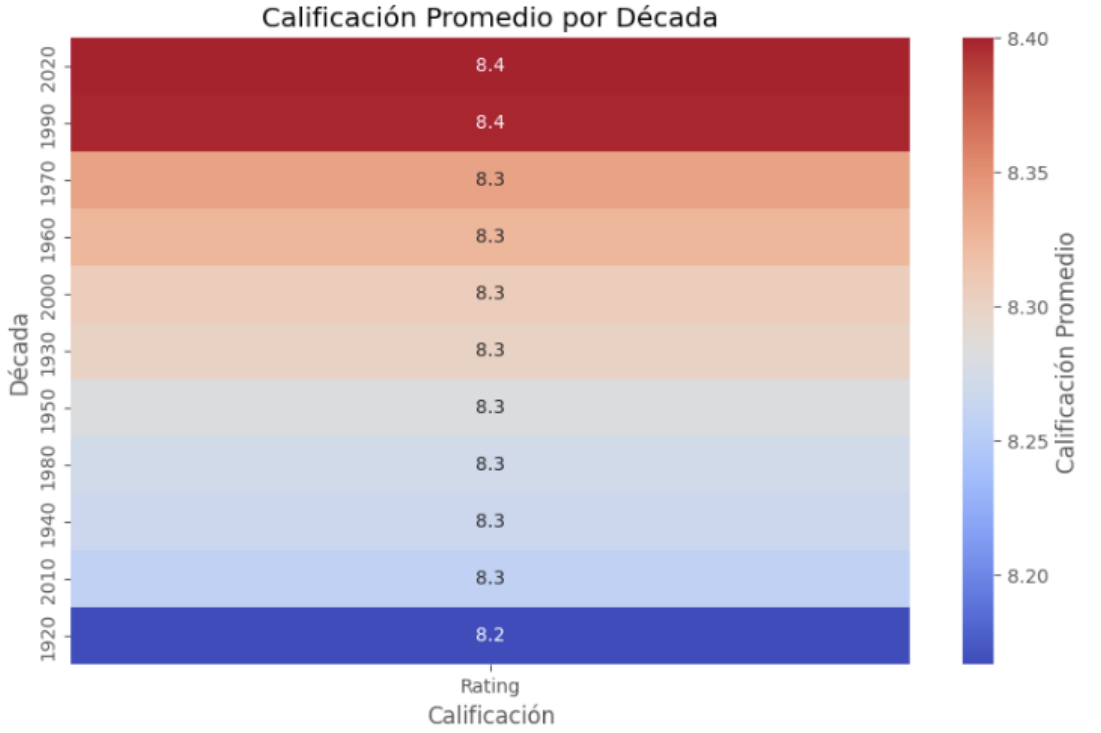

Distribución de Calificaciones
Media: 8.31
Frecuencia
Calificación
80
70
60
50
40
30
20
10
0
8.0
8.2
8.4
8.6
8.8
9.0
9.2
8.0
8.1
8.2
8.3
8.4
8.5
8.6
8.7
8.8
8.9
9.0
9.2
9.3

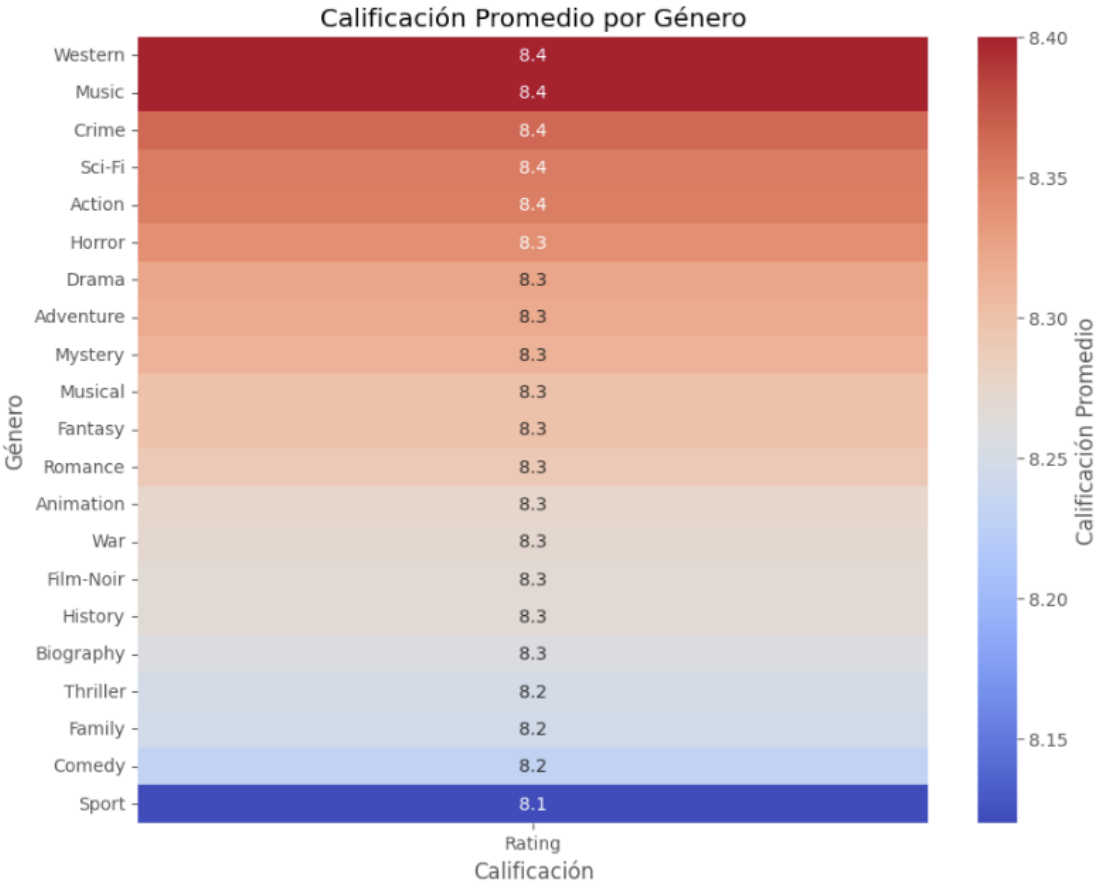

Calificación de Películas a lo Largo del Tiempo
Calificación
Fecha de Estreno
9.2
9.0
8.8
8.6
8.4
8.2
8.0
1920
1940
1960
1980
2000
2020


Distribución de Películas por Década de Estreno
Número de Películas
0
10
20
30
40
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
Década


Mejores Directores por Promedio de Calificación (Top 5)
Promedio de Calificación
0
2
4
6
8
8.95
8.90
8.87
8.80
8.70
Media General: 8.31
Rating
Frank Darabont
Peter Jackson
Francis Ford Coppola
Vidhu Vinod Chopra
Irvin Kershner
Director


Distribución de Calificaciones por Género
Calificación
9.2
9.0
8.8
8.6
8.4
8.2
8.0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Film-Noir
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western
Género


Calificación Promedio por Década
Década
2020
1990
1970
1960
2000
1930
1950
1980
1940
2010
1920
8.4
8.4
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.2
Rating
Calificación
8.40
8.35
8.30
8.25
8.20
Calificación Promedio


Calificación Promedio por Género
Género
Western
Music
Crime
Sci-Fi
Action
Horror
Drama
Adventure
Mystery
Musical
Fantasy
Romance
Animation
War
Film-Noir
History
Biography
Thriller
Family
Comedy
Sport
8.4
8.4
8.4
8.4
8.4
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.3
8.2
8.2
8.2
8.1
Rating
Calificación
8.40
8.35
8.30
8.25
8.20
8.15
Calificación Promedio

5. Contenido:

Explicar los campos que se incluyen en el dataset y el período de tiempo al que pertenecen los datos.

Este dataset contiene información sobre las 250 películas mejor valoradas en IMDb, proporcionando una visión completa de los elementos que hacen que estas películas sean tan valoradas por la audiencia. La popularidad de estas películas, medida a través de las calificaciones y reseñas de millones de usuarios, sugiere que han dejado una impresión duradera en el público. Este conjunto de datos ofrece una plataforma para analizar diversos factores, desde los aspectos narrativos hasta los elementos de producción que pueden haber contribuido al éxito de estas películas.

Al desglosar la información contenida en cada columna, se pueden investigar preguntas fundamentales sobre las características de las películas mejor valoradas:

- **Title (Título):** Cada película es identificada por su título único, lo cual es fundamental para poder investigar de manera individual o grupal los atributos de cada obra. Esto permite tanto el análisis a nivel de película como en comparaciones entre ellas.

- **Rating (Calificación):** La calificación promedio otorgada por los usuarios de IMDb es una medida significativa del impacto de la película. Analizar esta columna puede revelar qué aspectos generan una alta valoración en el público y cómo estos factores cambian a lo largo del tiempo o varían según otros factores, como el género o el director. La calificación es, en esencia, una medida de resonancia emocional y cultural entre el público y la obra.

- **Genre (Género):** Los géneros asignados a cada película ofrecen una visión sobre qué tipo de historias capturan la atención y el interés del público de forma consistente. Este atributo permite investigar si ciertos géneros, como el drama, la comedia o la acción, están asociados con una mayor popularidad. Analizar los géneros también permite ver si el atractivo de los temas cinematográficos es temporal, dependiendo de las preferencias sociales y culturales de cada época, o si hay géneros que mantienen un atractivo universal.

- **Web (Página de IMDb):** Un enlace directo a la página de IMDb de la película puede proporcionar detalles adicionales sobre cada obra, como su recaudación, duración y otros datos que pueden enriquecer el análisis. Esto es útil especialmente para estudios que deseen profundizar en las características específicas de una película en particular o acceder a datos actualizados.

- **Date (Fecha de estreno)**: La fecha de estreno de cada película permite explorar cómo han cambiado las preferencias de la audiencia a lo largo del tiempo. Esto también ofrece la oportunidad de analizar cuáles fueron las épocas doradas del cine, en las que surgieron más películas con calificaciones altas. Es una columna crucial para observar cómo ciertas décadas han producido películas icónicas y cuáles características comunes podrían haber influido en su éxito.

- **Director:** El director o los directores de cada película representan el estilo y la visión que hicieron posible la realización de la obra. Analizar a los directores permite identificar quiénes han logrado un impacto duradero en el cine y cómo su estilo puede contribuir al éxito de una película. También puede ayudar a descubrir patrones en las carreras de los directores y ver cómo algunos nombres recurrentes han marcado la historia del cine.

- **Writers (Guionistas):** Los guionistas que han creado las historias de cada película permiten investigar los temas, enfoques y narrativas que han cautivado a la audiencia. Al estudiar los guionistas, se puede ver qué tipo de historias y estilos de escritura han sido especialmente influyentes y cómo estos temas han evolucionado con el tiempo. También permite identificar si ciertos guionistas recurrentes tienden a participar en películas bien valoradas.

- **Actors (Actores):** Los actores principales de cada película son fundamentales para el atractivo de la obra, ya que pueden dar vida a los personajes de manera que resuene en la audiencia. Esta columna es útil para explorar qué actores han aparecido con más frecuencia en las películas mejor valoradas y cómo su presencia influye en la popularidad de una película. Los actores no solo contribuyen con su interpretación, sino que también pueden aportar un elemento de familiaridad o de prestigio que eleva el perfil de la obra.

Este dataset proporciona un vistazo a las películas favoritas del público y también permite estudiar patrones: **¿hay géneros más comunes en las películas mejor valoradas? ¿Algunos directores o actores aparecen más de una vez? ¿Qué tendencias han cambiado con el tiempo?** Son algunas de las preguntas que el dataset puede resolver y que se pueden estudiar en futuras prácticas.

6. Propietario:

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en su defecto, justificar esta búsqueda con análisis similares. Indicar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto elegido.

El propietario de los datos originales es IMDb (Internet Movie Database). Es una plataforma ampliamente utilizada y reconocida en la industria cinematográfica. Es propiedad de Amazon. IMDb recopila y organiza datos sobre películas, series, actores, directores y otros aspectos del cine y la televisión. Toda la información presentada en IMDb es accesible públicamente. Esto permite su uso para análisis y estudios siempre que se respeten las políticas de uso de la plataforma.

Para esta práctica de scraping, hemos seguido un enfoque respetuoso y ético. En primer lugar, el acceso a los datos se realiza mediante técnicas de web scraping. Se incluyen medidas para evitar saturar los servidores de IMDb como incluir pausas aleatorias entre las solicitudes. Además, se ha utilizado un *User-Agent* válido para identificar la naturaleza de la solicitud y cumplir con los términos de uso del sitio.

Este proyecto se inspira en estudios y análisis similares de datos de IMDb que han sido utilizados en el pasado para explorar tendencias cinematográficas, identificar patrones de popularidad o evaluar la influencia de ciertos actores y directores en la industria. La naturaleza educativa y analítica de este trabajo está en línea con el uso responsable y permitido de los datos de IMDb, ya que no se involucra ningún uso comercial ni modificación de la información original. Además, no se emplearon APIs externas como fuente de los datos. Esto sigue las indicaciones de la práctica y asegura que la extracción se haga directamente desde el sitio de IMDb de manera controlada y respetuosa.

7. Inspiración:

Explicar por qué puede ser interesante este conjunto de datos y qué preguntas se pretenden responder con ellos. Es necesario comparar con los análisis anteriores o análisis similares presentados en el apartado 6.

Este conjunto de datos resulta altamente relevante porque permite comprender qué hace que ciertas películas se conviertan en íconos dentro del universo cinematográfico, según la comunidad global de IMDb, un sitio con millones de usuarios activos. Al analizar aspectos como calificación, género, director, guionistas y elenco, podemos identificar los atributos comunes de las películas mejor valoradas y observar cómo cambian las preferencias del público a lo largo de las décadas. Esta información no solo es útil para la industria cinematográfica, sino que también contribuye a estudios culturales y sociológicos sobre el impacto del cine en diferentes generaciones y contextos históricos.

Preguntas de investigación detalladas:

1. ¿Qué géneros predominan entre las películas mejor valoradas?
    - Analizar los géneros en el contexto de las calificaciones más altas nos permite observar si ciertos temas o estilos logran una conexión duradera con el público o si su popularidad varía en función de los periodos históricos y las tendencias sociales. Por ejemplo, en las visualizaciones de géneros con altas calificaciones, se observa que el drama y el crimen tienen un lugar destacado, lo que sugiere que el público valora temas profundos y moralmente complejos. Al distribuir estos géneros a lo largo de las décadas, se pueden detectar patrones de cambio: tal vez en ciertas décadas se prefieren películas de aventuras o ciencia ficción, mientras que en otras, el drama y el thriller ganan relevancia.
2. ¿Existen directores o actores recurrentes en las películas de alta valoración?

- Identificar patrones de directores y actores recurrentes en el éxito cinematográfico nos ayuda a entender qué estilo y qué enfoque artístico resuena en la audiencia. Directores como Christopher Nolan, Quentin Tarantino y Martin Scorsese aparecen frecuentemente en las películas mejor valoradas. Esto sugiere que, más allá de la historia en sí, el estilo narrativo y visual particular de estos directores juega un papel clave en cómo se perciben sus películas. De forma similar, la presencia de actores recurrentes en estos filmes, como Robert De Niro o Morgan Freeman, indica que el público puede verlos como un sello de calidad o intensidad emocional, elementos que enriquecen la experiencia cinematográfica.

3. ¿Cómo han evolucionado las tendencias de popularidad a lo largo de las décadas?

    - Al estudiar la evolución de las calificaciones de las películas a lo largo del tiempo, agrupadas por década, podemos ver cómo cambian las preferencias del público en función de factores sociales, políticos y culturales. Por ejemplo, las películas con temas de guerra o espionaje eran populares durante y después de los períodos de conflicto, como la Guerra Fría o la Segunda Guerra Mundial

4. ¿Qué impacto tiene el reparto o el equipo creativo en la valoración de una película?

    - El análisis del reparto y el equipo creativo revela el peso que tienen ciertos nombres en el éxito de una película. No solo la presencia de directores famosos influye, sino que guionistas y actores con estilo o popularidad particulares pueden añadir un valor adicional. Los actores conocidos, que a menudo son considerados como talentosos en roles dramáticos complejos, podrían contribuir a calificaciones altas en películas de drama. La estructura narrativa creada por guionistas reconocidos también puede tener una influencia profunda en la aceptación de las películas, lo cual es evidente al analizar películas que reciben altas calificaciones debido a sus tramas complejas y bien desarrolladas.

El análisis de datos de IMDb no es nuevo en la industria ni en los estudios académicos, y se han realizado investigaciones previas que exploran estos factores desde una perspectiva más general o enfocada en un solo aspecto, como el género o el director. Sin embargo, este enfoque es innovador porque utiliza un análisis multifactorial que considera varias dimensiones simultáneamente: género, directores, escritores y actores. Esto permite no solo ver cómo cada factor individual impacta la calificación, sino también entender las interacciones entre estos elementos y cómo contribuyen colectivamente al éxito de una película.

Al sumar estos resultados a los estudios previos, el análisis se convierte en una herramienta poderosa para entender los patrones de éxito en el cine desde múltiples ángulos. Permite identificar no solo qué géneros o directores tienen más éxito, sino también cómo la combinación de varios factores puede crear un producto que resuena más profundamente en la audiencia. Este enfoque también ofrece una perspectiva única sobre la evolución de la industria, mostrando cómo ciertos nombres, géneros y estilos han mantenido su popularidad o cómo han surgido nuevas preferencias en la audiencia de IMDb.

8. Licencia:

Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under
- Database Contents License.
- Otra (especificar cuál)

Para este proyecto, la **licencia CC BY-NC-SA 4.0 (Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International)** es una opción adecuada para el dataset resultante. Esta licencia permite que otros usuarios compartan, adopten y utilicen el dataset siempre que:

- Den el crédito adecuado al autor.
- No lo usen con fines comerciales.
- Compartan el resultado bajo los mismos términos de esta licencia.

La elección se basa en los siguientes puntos:

1. Reconocimiento del Propietario Original: El dataset se basa en datos públicos de IMDb, propiedad de Amazon. Al aplicar esta licencia, permitimos su uso y reutilización mientras se reconoce a IMDb como fuente original de los datos. Con esta licencia se cumplen con los principios éticos de atribución.
2. No Comercial: La licencia restringe el uso comercial del dataset, este punto es importante ya que los datos originales de IMDb no fueron creados para fines

comerciales sin autorización. Esto asegura que el dataset mantenga un uso educativo y de investigación sin fines de lucro.

3. Compatibilidad con el Propósito Educativo: Al requerir que cualquier trabajo derivado utilice la misma licencia se fomenta un uso responsable y colaborativo del dataset en la comunidad educativa. Con este punto se asegura que futuras versiones del conjunto de datos sigan siendo accesibles para fines no comerciales.

9. Código:

Código implementado para la obtención del dataset, preferiblemente en Python o, alternativamente, en R.

```
# Comenzamos importando las librerias necesarias

import requests

import json

import pandas as pd

from bs4 import BeautifulSoup

import time

import random


# Comenzamos creando una función inicializadora que nos ayudará a identificar la

# página escecífica donde queremos hacer el scrapping.

class IMDbScraper:

    def __init__(self):

        self.url = "https://www.imdb.com/chart/top/?ref_=nv_mv_250"

        self.data = []

        self.headers = {
```

```
            "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/92.0.4515.159 Safari/537.36"
        }
# La siguiente función será específica para verificar el hearder y el user-agent
    def check_user_agent(self):
        print(f"User-Agent utilizado: {self.headers['User-Agent']}")

# Continuaremos creando otra funciòn que nos permita descargar la web objetivo en html
    def download_html(self):
        response = requests.get(self.url, headers=self.headers)
        if response.status_code == 200:
            return response.text
        else:
            print(f"Error al descargar la página. Estado: {response.status_code}")
            return None

# Ahora configuramos función donde podamos obtener los datos básicos de las 250 películas
    def scrape_data(self, html):
        start = html.find('<script type="application/ld+json">')
        end = html.find('</script>', start)
        json_data = html[start + len('<script type="application/ld+json">'):end].strip()

        data = json.loads(json_data)
# Para ejecutar esta función, hacemos una iteración para cada uno de los datos que queremos obtener.
```

```
        for item in data['itemListElement']:
            movie = item['item']
            title = movie['name']
            rating = movie['aggregateRating']['ratingValue']
            genre = movie.get('genre', 'N/A')
            web = movie['url']
            self.data.append([title, rating, genre, web])

# Para los datos más complejos, creamos una función a parte con BeautifulSoup.
    def scrape_movie_details(self, url):
        response = requests.get(url, headers=self.headers)
        if response.status_code == 200:
            soup = BeautifulSoup(response.text, 'html.parser')

            json_data = soup.find("script", type="application/ld+json").string
            movie_data = json.loads(json_data)

            date = movie_data.get("datePublished", "N/A")
            director = ", ".join([d["name"] for d in movie_data.get("director", [])])
            writers = ", ".join([w["name"] for w in movie_data.get("creator", []) if w["@type"]
== "Person"])
            actors = ", ".join([a["name"] for a in movie_data.get("actor", [])])

            time.sleep(random.uniform(1, 2))
```

```
            return date, director, writers, actors
        else:
            print(f"Error al acceder a la URL: {url}")
            return "N/A", "N/A", "N/A", "N/A"


# Ahora tenemos una función que genera el dataframe con toda la información obtenida previamente.
    def save_to_dataframe(self):
        df = pd.DataFrame(self.data, columns=["Title", "Rating", "Genre", "Web"])
        df[['Date', 'Director', 'Writers', 'Actors']] = df['Web'].apply(lambda url: pd.Series(self.scrape_movie_details(url)))


        return df


# Por último,creamos una función que ejecuta toddo el código
    def run(self):
        print("Scraping las 250 mejores películas de IMBb")


        self.check_user_agent()


        html = self.download_html()
        if html:
            self.scrape_data(html)
            df = self.save_to_dataframe()
            display(df)
            df.to_csv('films.csv', index=False)
```

```
        else:
            print("No se pudo obtener HTML.")


scraper = IMDbScraper()
scraper.run()
```

10. Dataset:

Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción del mismo.

El dataset se puede encontrar en:

https://github.com/jmf3192/UOC_PRACT1/tree/main/dataset

11. Vídeo:

Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

El enlace es el siguiente:

https://drive.google.com/file/d/1l_v0zIJjMLqeij2nO6K_-mciAg09DM-g/view?usp=drive_link

CONTRIBUCIONES	FIRMA
Investigación previa	Jorge Moreno Fuentes, Jorge Martín Rota
Redacción de las respuestas	Jorge Moreno Fuentes, Jorge Martín Rota
Desarrollo del código	Jorge Moreno Fuentes, Jorge Martín Rota
Participación en el vídeo	Jorge Moreno Fuentes, Jorge Martín Rota