**Elaborado por: Andersson Tellez Beltran**

**Paso 1: Alcance del proyecto y captura de datos**

* Identificar y recopilar los datos que usaras para tu proyecto.

Los datos que voy a usar para este proyecto lo extraje del siguiente link:

<https://www.kaggle.com/datasets/floatingcoder/tmdb-20000-movies-dataset>

* Explicar para qué casos de uso final deseas preparar los datos, por ejemplo: tabla de análisis,

aplicación de fondo, base de datos de fuentes de verdad, etc.)

Los datos los usare para generar una tabla de análisis, realizar una estadística básica a los datos.

**Paso 2: Explorar y evaluar los datos, el EDA.**

1. Cargamos los datos a la herramienta

text = pd.read\_csv('Documents\Movies.csv', index\_col=1)

Tabla

Descripción generada automáticamente

1. Indicamos índice en el DF

text.set\_index('C#', drop=True, append=False, inplace=True, verify\_integrity=False)

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

1. Se reemplazan los valores nulos

text = text.fillna(0)

1. Se ordenan

text.sort\_values(by='id',ascending=True, inplace=True)

Tabla

Descripción generada automáticamente con confianza media

1. Se convierte float a entero

text.popularity=(text.popularity\*1000)

Tabla

Descripción generada automáticamente Tabla

Descripción generada automáticamente

1. Se eliminan duplicados

text.drop\_duplicates(subset=['id'],inplace=True)

Tabla

Descripción generada automáticamente

1. Volvemos a ordenar

text.sort\_values(by='C#',ascending=True, inplace=True)

Tabla

Descripción generada automáticamente con confianza media

1. Exploramos la data

Texto

Descripción generada automáticamente

1. Descripción numerica

desc=text.describe()

Tabla

Descripción generada automáticamente

1. Conversión de serie para poder extraer el año

text['release\_date'] = pd.to\_datetime(text['release\_date'], errors='coerce')

text['Año'] = text['release\_date'].dt.year

Tabla

Descripción generada automáticamente

1. Studio variable Target

text.original\_language.value\_counts()

target1 = text.groupby(by = ['Año','original\_language']).adult.count() (target1.tail(11).unstack(level=0).plot.bar()) plt.show()

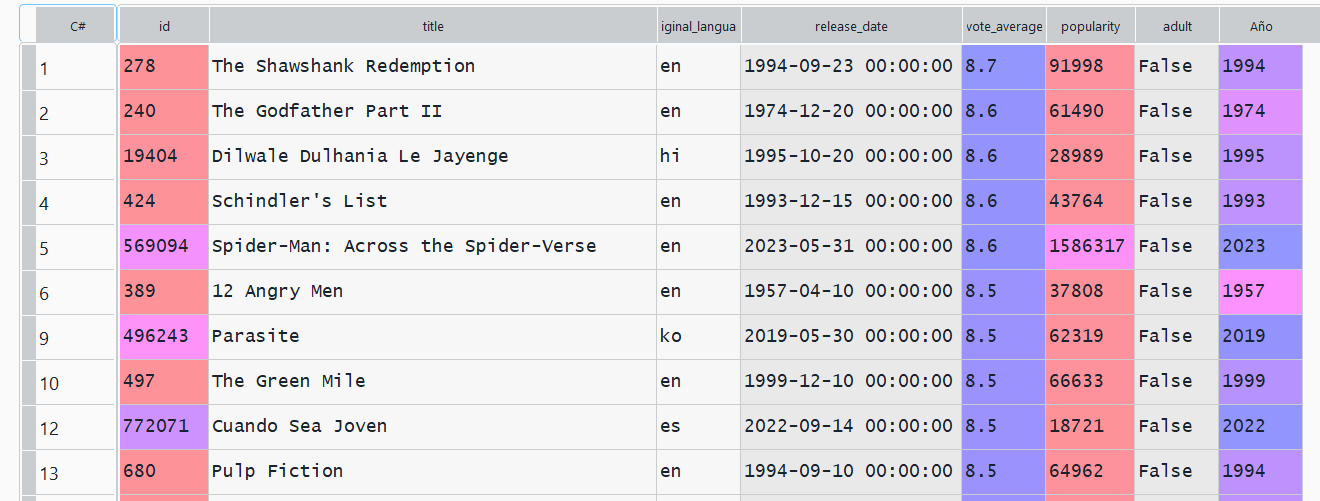
Gráfico

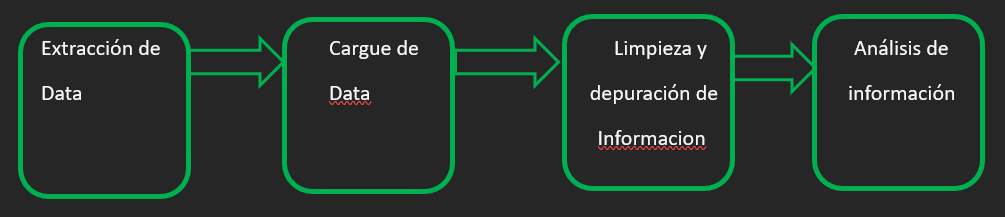
Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamente

**Paso 3: Definir el modelo de datos**

El modelo de datos que elegí para trabajar es un modelo plano ya que es bidimensional y todas las filas están relacionadas entre si.





Escogí trabajar con Spyder en Python por que tengo conocimiento especifico en este lenguaje, lo he usado para crear programas de escritorio y para automatizar procesos, en cuanto a la visualización de los datos y análisis prefiero trabajar con power BI, no lo use en este proyecto ya que al no ser tan robusto se puede trabajar directamente con Python y sus métodos.

La frecuencia para actualizar datos seria semanal ya que aumentaría la cantidad de espectadores en las peliculas recientes y es mejor tener un umbral mas largo para consolidar la información.

**Paso 5: Completar la redacción del Proyecto**

• ¿Cuál es el objetivo del proyecto?

El objetivo del proyecto que elegí es netamente informativo con el fin de obtener un ranking del top de los resultados finales.

• ¿Qué preguntas quieres hacer?

¿El desarrollo del cine en el pasar del tiempo?

¿Preferencia de los espectadores de acuerdo con su idioma?

¿Ranking mayor y menor de las películas?

• ¿Por qué eligió el modelo que eligió?

Fue el modelo que mas se adapta al tipo de información que encontré.

• Incluya una descripción de cómo abordaría el problema de manera diferente en los siguientes

escenarios:

o Si los datos se incrementaran en 100x.

Buscar e implementar herramientas robustas que soporten la cantidad de datos.

o Si las tuberías se ejecutaran diariamente en una ventana de tiempo especifica.

En lo posible que se ejecutaran en un umbral de tiempo no pico para que no generara conflicto con procesos externos.

o Si la base de datos necesitara ser accedido por más de 100 usuarios funcionales.

Pues trataría de desglosar el uso final del usuario y a si validar si necesita acceso a toda la data de la tabla o mas bien segmentada y estructurar la finalidad en redes diferentes.

o Si se requiere hacer analítica

Implementaría una herramienta para tal fin como power BI.