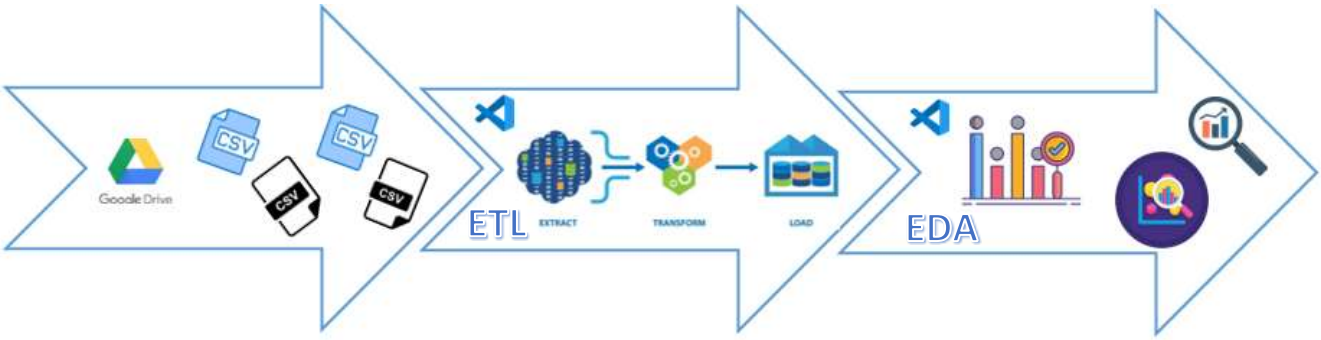
Informe de EDA



El Análisis Exploratorio de Datos (EDA) consiste en examinar y explorar un conjunto de datos con el objetivo de descubrir patrones, tendencias, relaciones y características importantes de los datos.

Algunas de las cosas que se pueden realizar, entre otras son:

-Descripción de datos: Se examina la estructura y composición de los datos, incluyendo la cantidad de observaciones, variables, tipos de datos y distribuciones de valores.

-Visualización de datos: Se utilizan gráficos y diagramas para representar visualmente los datos.

-Tratamiento de datos faltantes o erróneos: Se detectan y manejan los valores faltantes o inconsistentes en el conjunto de datos.

-Análisis de correlación: Se examina la relación entre las diferentes variables del conjunto de datos.

-Identificación de outliers (valores atípicos): Se buscan observaciones que se desvíen significativamente de la tendencia general de los datos.

En este informe, se detallará:

- Patrones ocultos, variables relevantes, problemas en los datos y la toma de decisiones informadas sobre el enfoque analítico adecuado.

Lenguaje utilizado:

- Python

Librerías utilizadas:

- Pandas - Numpy - Seaborn - Matplotlib(pyplot)

Giselle Pereira Nuñez

Se abre “archivo.csv”, donde se encuentran los datos de plataformas y reseñas de las mismas.

Tratado como un Dataframe, se le realizan:

1)Visualización de tipo de dato de cada campo:

- Id, type, title, director, cast, country, date_added, rating, listed_in, description, duration_type, timestamp, movieId, plataform: Estas columnas de tipo "object".
- release_year: Es una columna de tipo "int64".
- duration_int: Es una columna de tipo "int64".
- userId: Es una columna de tipo "int64".
- score: Es una columna de tipo "float64".

```
id              object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
listed_in       object
description     object
duration_int     int64
duration_type   object
userId           int64
score          float64
timestamp       object
movieId         object
plataform       object
dtype: object
```

2)Cantidad de valores faltantes (se describirán solo los campos que tengan faltantes):

- director: Hay 3,960,615 valores faltantes en esta columna, siendo un 35.9% sobre la totalidad de valores de este campo.
- cast: Hay 2,550,859 valores faltantes en esta columna, siendo un 23.1% sobre la totalidad de valores de este campo.
- country: Hay 5,510,080 valores faltantes en esta columna, siendo un 50% sobre la totalidad de valores de este campo.
- date_added: Hay 4,577,546 valores faltantes en esta columna, siendo un 41.5% sobre la totalidad de valores de este campo.

Giselle Pereira Nuñez

- description: Hay 1,815 valores faltantes en esta columna, siendo un 0.01% sobre la totalidad de valores de este campo.
- duration_type: Hay 231,117 valores faltantes en esta columna, siendo un 2.1% sobre la totalidad de valores de este campo.

Finalmente, el porcentaje de los datos faltantes en su totalidad es de 8.5%.

```
id                0
type              0
title             0
director    3960615
cast        2550859
country     5510080
date_added  4577546
release_year      0
rating            0
listed_in         0
description    1815
duration_int      0
duration_type  231117
userId            0
score             0
timestamp         0
movieId           0
plataform         0
dtype: int64
```

```
id              0.0
type            0.0
title           0.0
director       35.9
cast           23.1
country        50.0
date_added     41.5
release_year    0.0
rating          0.0
listed_in       0.0
description     0.0
duration_int    0.0
duration_type   2.1
userId          0.0
score           0.0
timestamp       0.0
movieId         0.0
plataform       0.0
dtype: float64
```

3)Relleno del campo “country” con ‘no declarado’.

4)Estadísticas descriptivas (se realizan sobre los campos numericos):

- El conteo de valores sin los valores faltantes.
- La media Indica el valor típico de la columna, útil para tener una idea general de la distribución de los datos.
- La desviación estándar mide la variabilidad de los datos alrededor de la media. Una desviación estándar grande indica que los datos están más dispersos alrededor de la media, mientras que una desviación estándar pequeña indica que los datos están más agrupados cerca de la media.
- El valor mínimo y el valor máximo indican los límites del rango de valores observados.
- Los cuartiles dividen los datos en cuatro partes iguales. El primer cuartil indica el valor que es mayor que el 25% de los valores en la columna. El segundo cuartil indica el valor que es mayor que el 50% de los valores en la columna. El tercer cuartil indica el valor que es mayor que el 75% de los valores en la columna.

	release_year	duration_int	userId	score
count	1.102429e+07	1.102429e+07	1.102429e+07	1.102429e+07
mean	2.010819e+03	6.569939e+01	8.997251e+04	3.533455e+00
std	1.538663e+01	5.175968e+01	8.686601e+04	1.059692e+00
min	1.920000e+03	0.000000e+00	1.000000e+00	5.000000e-01
25%	2.010000e+03	3.000000e+00	2.855800e+04	3.000000e+00
50%	2.016000e+03	8.300000e+01	5.684300e+04	3.500000e+00
75%	2.019000e+03	1.010000e+02	1.168670e+05	4.000000e+00
max	2.021000e+03	6.010000e+02	2.708960e+05	5.000000e+00

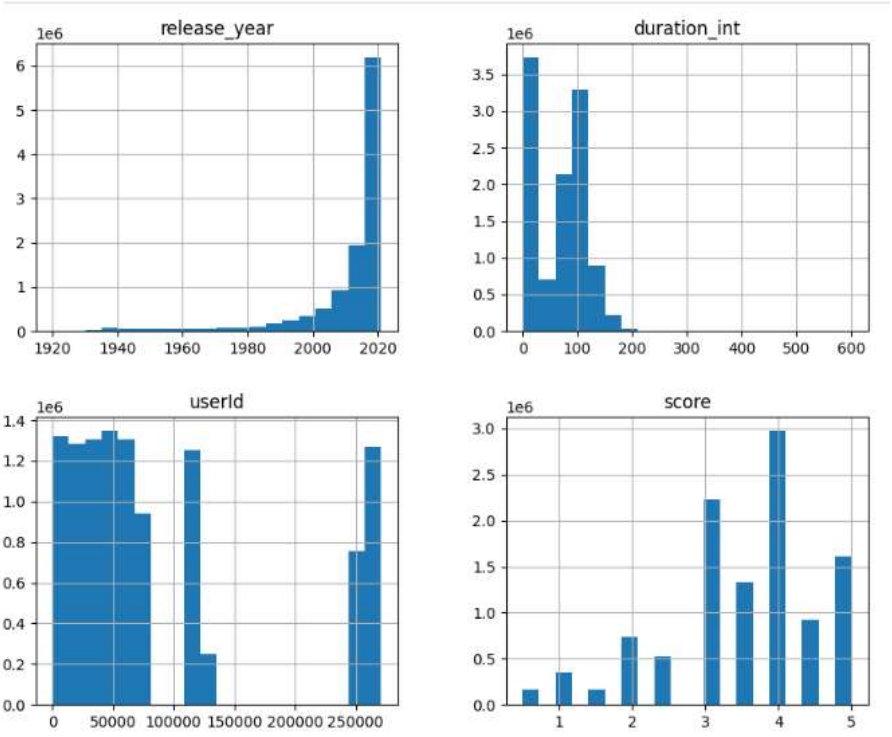
5)Reviso Correlaciones (también se aplican en los campos numéricos)

- Los valores de la matriz de correlación varían entre -1 y 1. Un valor de 1 indica una correlación perfecta positiva (dos variables que aumentan juntas), mientras que un valor de -1 indica una correlación perfecta negativa (dos variables que disminuyen juntas). Un valor de 0 indica que no hay correlación entre las variables.

	release_year	duration_int	userId	score
release_year	1.000000	-0.134076	-0.000367	0.000517
duration_int	-0.134076	1.000000	-0.000043	-0.000325
userId	-0.000367	-0.000043	1.000000	0.004608
score	0.000517	-0.000325	0.004608	1.000000

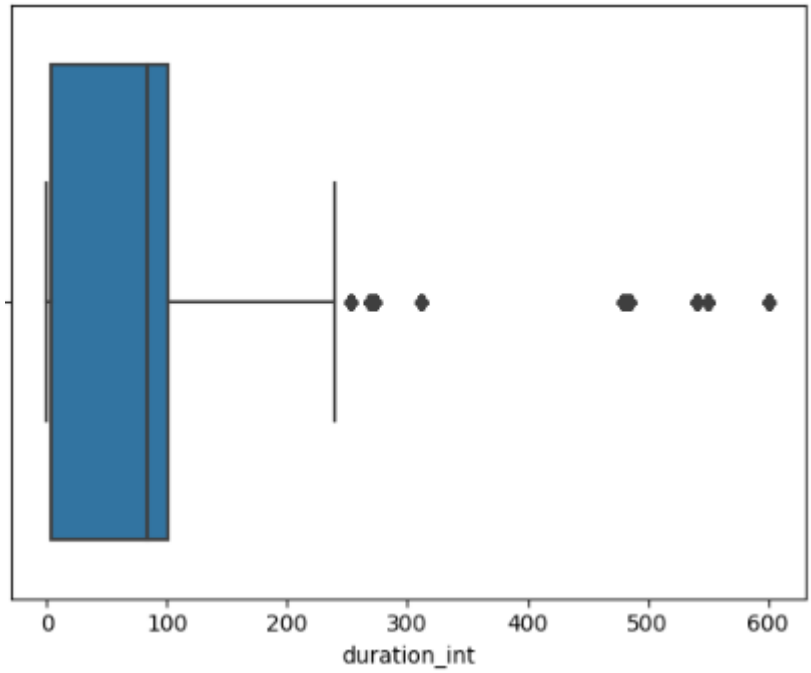
6)Visualización:

- El campo “release_year”, tiene rangos de años aproximados desde 1930 a 2020, siendo este último año donde predominan las cantidades de películas y series.
- El campo “duration_int”, tiene un rango aproximado desde 0 a 200 del tiempo de duración de películas y series. Siendo predominantes los rangos de 0 a 20 (apróx) y de 95 a 110 (apróx).
- El campo “userId” es el que predomina en la totalidad de los datos numéricos.
- El campo “score”, con rango desde 0 a 5. Predomina el puntaje 4 para películas y series.

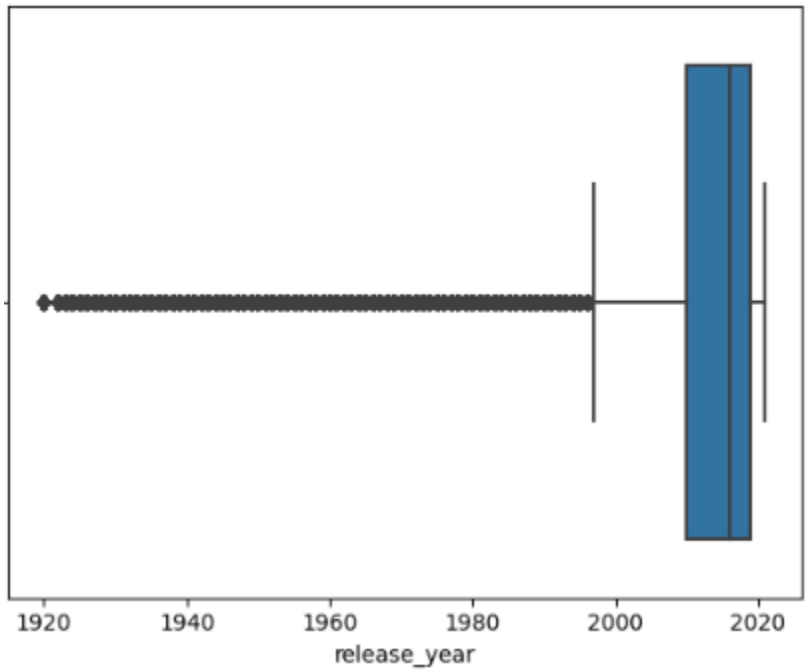


7)Valores atípicos de cada campo numérico:

- Campo “duration_int”, se observan valores atípicos de por ejemplo 600 minutos. Luego de verificar se concluye que es real, ya que hay películas que duran más de 10hs.



- Campo “release_year”, se observan valores atípicos también reales, ya que se refieren al año de lanzamiento de la película.



El Análisis Exploratorio de Datos es importante porque proporciona una comprensión profunda del conjunto de datos. Permite descubrir patrones ocultos, seleccionar variables relevantes, identificar problemas en los datos y tomar decisiones informadas sobre el enfoque analítico adecuado. En última instancia, el EDA ayuda a construir modelos más precisos y a obtener conocimientos valiosos de los datos.