In [1]:

```
# importing the library
import numpy as np
import pandas as pd
```

In [3]:

```
#Load the data
titles = pd.read_csv("credits.csv")
credits  = pd.read_csv("titles.csv")

# merge the data
df_combined = pd.merge(titles,credits,on= "id",how="left")

#check result
df_combined.shape
df_combined.head()
```

Out[3]:

	person_id	id	name	character	role	title	type	description	release_y
0	3748	tm84618	Robert De Niro	Travis Bickle	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
1	14658	tm84618	Jodie Foster	Iris Steensma	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
2	7064	tm84618	Albert Brooks	Tom	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
3	3739	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
4	48933	tm84618	Cybill Shepherd	Betsy	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19

In [9]:

```
df_combined.tail()
```

Out[9]:

	person_id	id	name	character	role	title	type	description
77796	736339	tm1059008	Adelaida Buscato	María Paz	ACTOR	Lokillo	MOVIE	A controversial TV host and comedian who has b...
77797	399499	tm1059008	Luz Stella Luengas	Karen Bayona	ACTOR	Lokillo	MOVIE	A controversial TV host and comedian who has b...
77798	373198	tm1059008	Inés Prieto	Fanny	ACTOR	Lokillo	MOVIE	A controversial TV host and comedian who has b...
77799	378132	tm1059008	Isabel Gaona	Cacica	ACTOR	Lokillo	MOVIE	A controversial TV host and comedian who has b...
77800	1950416	tm1059008	Julian Gaviria	NaN	DIRECTOR	Lokillo	MOVIE	A controversial TV host and comedian who has b...

In [16]:

```
# drop duplicates
df_combined.drop_duplicates(inplace =True)
```

In [17]:

```
df_combined.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 77801 entries, 0 to 77800
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   person_id             77801 non-null  int64  
 1   id                    77801 non-null  object 
 2   name                  77801 non-null  object 
 3   character             68029 non-null  object 
 4   role                  77801 non-null  object 
 5   title                 77800 non-null  object 
 6   type                  77801 non-null  object 
 7   description           77763 non-null  object 
 8   release_year          77801 non-null  int64  
 9   age_certification     46658 non-null  object 
 10  runtime               77801 non-null  int64  
 11  genres                77801 non-null  object 
 12  production_countries  77801 non-null  object 
 13  seasons               14710 non-null  float64
 14  imdb_id               74302 non-null  object 
 15  imdb_score            73851 non-null  float64
 16  imdb_votes            73764 non-null  float64
 17  tmdb_popularity       77790 non-null  float64
 18  tmdb_score            76664 non-null  float64
dtypes: float64(5), int64(3), object(11)
memory usage: 11.3+ MB
```

In [18]:

```
df_combined.head()
```

Out[18]:

	person_id	id	name	character	role	title	type	description	release_y
0	3748	tm84618	Robert De Niro	Travis Bickle	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
1	14658	tm84618	Jodie Foster	Iris Steensma	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
2	7064	tm84618	Albert Brooks	Tom	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
3	3739	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19
4	48933	tm84618	Cybill Shepherd	Betsy	ACTOR	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	19

In [19]:

```python
# handling missing values
df_combined['imdb_score'].fillna(df_combined['imdb_score'].mean(),inplace =True)
```

```
C:\Users\prana\AppData\Local\Temp\ipykernel_30908\1817718760.py:2: FutureWarning: A va
lue is trying to be set on a copy of a DataFrame or Series through chained assignment
using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because th
e intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({c
ol: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the
operation inplace on the original object.


  df_combined['imdb_score'].fillna(df_combined['imdb_score'].mean(),inplace =True)
```

In [20]:

```python
df_combined.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 77801 entries, 0 to 77800
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   person_id             77801 non-null  int64  
 1   id                    77801 non-null  object 
 2   name                  77801 non-null  object 
 3   character             68029 non-null  object 
 4   role                  77801 non-null  object 
 5   title                 77800 non-null  object 
 6   type                  77801 non-null  object 
 7   description           77763 non-null  object 
 8   release_year          77801 non-null  int64  
 9   age_certification     46658 non-null  object 
 10  runtime               77801 non-null  int64  
 11  genres                77801 non-null  object 
 12  production_countries  77801 non-null  object 
 13  seasons               14710 non-null  float64
 14  imdb_id               74302 non-null  object 
 15  imdb_score            77801 non-null  float64
 16  imdb_votes            73764 non-null  float64
 17  tmdb_popularity       77790 non-null  float64
 18  tmdb_score            76664 non-null  float64
dtypes: float64(5), int64(3), object(11)
memory usage: 11.3+ MB
```

In [21]:

```
# convert text columns
text_cols = ['id','name','character','role','title','type','description','age_certifi
df_combined[text_cols] =df_combined[text_cols].astype('string')
```

In [22]:

```
# convert categorical columns
df_combined['type'] =df_combined['type'].astype('category')
df_combined['role']=df_combined['role'].astype('category')
df_combined['age_certification']=df_combined['age_certification'].astype('category')
```

In [24]:

```
# fix the  numeric
df_combined['seasons']=df_combined['seasons'].astype('Int64')
```

In [25]:

```
#imdb moves -> interger (nullble)
df_combined['imdb_votes']=df_combined['imdb_votes'].astype('Int64')

# float columns -> rouding form
flo_columns = ['imdb_score', 'tmdb_score','tmdb_popularity']
df_combined[flo_columns] =df_combined[flo_columns].astype('float64')


# interger columns
df_combined['person_id'] =df_combined['person_id'].astype('int64')
df_combined['release_year']=df_combined['release_year'].astype('int64')
df_combined['runtime']=df_combined['runtime'].astype('int64')
```

In [27]:

```
df_combined.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 77801 entries, 0 to 77800
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype   
---  ------                --------------  -----   
 0   person_id             77801 non-null  int64   
 1   id                    77801 non-null  string  
 2   name                  77801 non-null  string  
 3   character             68029 non-null  string  
 4   role                  77801 non-null  category
 5   title                 77800 non-null  string  
 6   type                  77801 non-null  category
 7   description           77763 non-null  string  
 8   release_year          77801 non-null  int64   
 9   age_certification     46658 non-null  category
 10  runtime               77801 non-null  int64   
 11  genres                77801 non-null  string  
 12  production_countries  77801 non-null  string  
 13  seasons               14710 non-null  Int64   
 14  imdb_id               74302 non-null  string  
 15  imdb_score            77801 non-null  float64 
 16  imdb_votes            73764 non-null  Int64   
 17  tmdb_popularity       77790 non-null  float64 
 18  tmdb_score            76664 non-null  float64 
dtypes: Int64(2), category(3), float64(3), int64(3), string(8)
memory usage: 9.9 MB
```

In [28]:
```
df_combined.isnull()
```

Out[28]:

	person_id	id	name	character	role	title	type	description	release_year	a
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
77796	False	False	False	False	False	False	False	False	False	
77797	False	False	False	False	False	False	False	False	False	
77798	False	False	False	False	False	False	False	False	False	
77799	False	False	False	False	False	False	False	False	False	
77800	False	False	False	True	False	False	False	False	False	

77801 rows × 19 columns

In [30]:
```
df_combined.isnull().sum()
```

Out[30]:

```
person_id                   0
id                          0
name                        0
character                9772
role                        0
title                       1
type                        0
description                38
release_year                0
age_certification       31143
runtime                     0
genres                      0
production_countries        0
seasons                 63091
imdb_id                  3499
imdb_score                  0
imdb_votes               4037
tmdb_popularity            11
tmdb_score               1137
dtype: int64
```

In [41]:

```
# checking -> missing the character
df_combined['charater']=df_combined['character'].fillna('Unknown')
```

In [48]:

```
# removing the missing values from age_certification
df_combined['age_certification'] =(
    df_combined['age_certification'].cat.add_categories('Not Rated').fillna('Not Rate
)
```

In [49]:

```
# removing the null values form seasons
df_combined['seasons']=df_combined['seasons'].fillna(0)
```

In [50]:

```
# Imdb_votes remove null values
df_combined['imdb_votes']=df_combined['imdb_votes'].fillna(0)
```

In [51]:

```
#tmdb score -> fill with mean(recommeded)
df_combined['tmdb_score']=df_combined['tmdb_score'].fillna(df_combined['tmdb_score'].
```

In [52]:

```
df_combined.isnull().sum()
```

Out[52]:

```
person_id               0
id                      0
name                    0
character            9772
role                    0
title                   1
type                    0
description            38
release_year            0
age_certification       0
runtime                 0
genres                  0
production_countries    0
seasons                 0
imdb_id              3499
imdb_score              0
imdb_votes              0
tmdb_popularity        11
tmdb_score              0
charater                0
dtype: int64
```

In [53]:

```
# character -> keep but fill logically
df_combined['character'] =df_combined['character'].fillna('Unknown')
```

In [54]:

```
# title
df_combined=df_combined.dropna(subset=['title'])
```

In [55]:

```
# description
df_combined['description'] =df_combined['description'].fillna('No description availab
```

```
C:\Users\prana\AppData\Local\Temp\ipykernel_30908\555764813.py:2: SettingWithCopyWarni
ng:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df_combined['description'] =df_combined['description'].fillna('No description availa
ble')
```

In [56]:

```
df_combined['imdb_id']=df_combined['imdb_id'].fillna('Not Available ')
```

```
C:\Users\prana\AppData\Local\Temp\ipykernel_30908\1104850663.py:1: SettingWithCopyWarn
ing:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df_combined['imdb_id']=df_combined['imdb_id'].fillna('Not Available ')
```

In [57]:

```
# tmdb_popularity
df_combined['tmdb_popularity'] =df_combined['tmdb_popularity'].fillna(
    df_combined['tmdb_popularity'].mean()
)
```

```
C:\Users\prana\AppData\Local\Temp\ipykernel_30908\3465524233.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_combined['tmdb_popularity'] =df_combined['tmdb_popularity'].fillna(
```

In [60]:

```
df_combined.isnull().sum()
```

Out[60]:

```
person_id               0
id                      0
name                    0
character               0
role                    0
title                   0
type                    0
description             0
release_year            0
age_certification       0
runtime                 0
genres                  0
production_countries    0
seasons                 0
imdb_id                 0
imdb_score              0
imdb_votes              0
tmdb_popularity         0
tmdb_score              0
charater                0
dtype: int64
```

In [62]:

```
df_combined.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 77800 entries, 0 to 77800
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype   
---  ------                --------------  -----   
 0   person_id             77800 non-null  int64   
 1   id                    77800 non-null  string  
 2   name                  77800 non-null  string  
 3   character             77800 non-null  string  
 4   role                  77800 non-null  category
 5   title                 77800 non-null  string  
 6   type                  77800 non-null  category
 7   description           77800 non-null  string  
 8   release_year          77800 non-null  int64   
 9   age_certification     77800 non-null  category
 10  runtime               77800 non-null  int64   
 11  genres                77800 non-null  string  
 12  production_countries  77800 non-null  string  
 13  seasons               77800 non-null  Int64   
 14  imdb_id               77800 non-null  string  
 15  imdb_score            77800 non-null  float64 
 16  imdb_votes            77800 non-null  Int64   
 17  tmdb_popularity       77800 non-null  float64 
 18  tmdb_score            77800 non-null  object  
 19  charater              77800 non-null  string  
dtypes: Int64(2), category(3), float64(2), int64(3), object(1), string(9)
memory usage: 11.1+ MB
```

```
In [ ]:
```