In [2]:
```
import pandas as pd                                    #Suyash Pratap S
ingh
import numpy as np                                     #181B226
```

In [3]:
```
df = pd.read_csv(r'C:\Users\Admin\Desktop\movie_metadata.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_faceb
0	Color	James Cameron	723.0	178.0	0.0	
1	Color	Gore Verbinski	302.0	169.0	563.0	
2	Color	Sam Mendes	602.0	148.0	0.0	
3	Color	Christopher Nolan	813.0	164.0	22000.0	
4	NaN	Doug Walker	NaN	NaN	131.0	

5 rows × 28 columns

In [5]:
```
df.shape
```

Out[5]:
```
(5043, 28)
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5043 entries, 0 to 5042
Data columns (total 28 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   color                      5024 non-null   object 
 1   director_name              4939 non-null   object 
 2   num_critic_for_reviews     4993 non-null   float64
 3   duration                   5028 non-null   float64
 4   director_facebook_likes    4939 non-null   float64
 5   actor_3_facebook_likes     5020 non-null   float64
 6   actor_2_name               5030 non-null   object 
 7   actor_1_facebook_likes     5036 non-null   float64
 8   gross                      4159 non-null   float64
 9   genres                     5043 non-null   object 
 10  actor_1_name               5036 non-null   object 
 11  movie_title                5043 non-null   object 
 12  num_voted_users            5043 non-null   int64  
 13  cast_total_facebook_likes  5043 non-null   int64  
 14  actor_3_name               5020 non-null   object 
 15  facenumber_in_poster       5030 non-null   float64
 16  plot_keywords              4890 non-null   object 
 17  movie_imdb_link            5043 non-null   object 
 18  num_user_for_reviews       5022 non-null   float64
 19  language                   5031 non-null   object 
 20  country                    5038 non-null   object 
 21  content_rating             4740 non-null   object 
 22  budget                     4551 non-null   float64
 23  title_year                 4935 non-null   float64
 24  actor_2_facebook_likes     5030 non-null   float64
 25  imdb_score                 5043 non-null   float64
 26  aspect_ratio               4714 non-null   float64
 27  movie_facebook_likes       5043 non-null   int64  
dtypes: float64(13), int64(3), object(12)
memory usage: 1.1+ MB
```

In [7]:
```
#count total rows in each column which contain null values
df.isna().sum()
```

Out[7]:
```
color                        19
director_name               104
num_critic_for_reviews       50
duration                     15
director_facebook_likes     104
actor_3_facebook_likes       23
actor_2_name                 13
actor_1_facebook_likes        7
gross                       884
genres                        0
actor_1_name                  7
movie_title                   0
num_voted_users               0
cast_total_facebook_likes     0
actor_3_name                 23
facenumber_in_poster         13
plot_keywords               153
movie_imdb_link               0
num_user_for_reviews         21
language                     12
country                       5
content_rating              303
budget                      492
title_year                  108
actor_2_facebook_likes       13
imdb_score                    0
aspect_ratio                329
movie_facebook_likes          0
dtype: int64
```

In [8]:
```
df.isnull()
```

Out[8]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_fac
0	False	False	False	False	False	
1	False	False	False	False	False	
2	False	False	False	False	False	
3	False	False	False	False	False	
4	True	False	True	True	False	
...	...	...	...	...	...	
5038	False	False	False	False	False	
5039	False	True	False	False	True	
5040	False	False	False	False	False	
5041	False	False	False	False	False	
5042	False	False	False	False	False	

5043 rows × 28 columns

In [9]:
```
df.describe()
```

Out[9]:

	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	acto
count	4993.000000	5028.000000	4939.000000	5020.000000	
mean	140.194272	107.201074	686.509212	645.009761	
std	121.601675	25.197441	2813.328607	1665.041728	
min	1.000000	7.000000	0.000000	0.000000	
25%	50.000000	93.000000	7.000000	133.000000	
50%	110.000000	103.000000	49.000000	371.500000	
75%	195.000000	118.000000	194.500000	636.000000	
max	813.000000	511.000000	23000.000000	23000.000000	

In [10]:
```
print("Total number of null values = ",df.isnull().sum().sum())
```

```
Total number of null values =  2698
```

In [11]:
```
clean_data= df.dropna()
clean_data.head()
```

Out[11]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_faceb
0	Color	James Cameron	723.0	178.0	0.0	
1	Color	Gore Verbinski	302.0	169.0	563.0	
2	Color	Sam Mendes	602.0	148.0	0.0	
3	Color	Christopher Nolan	813.0	164.0	22000.0	
5	Color	Andrew Stanton	462.0	132.0	475.0	

5 rows × 28 columns

In [18]:
```
print(df.shape)
print(clean_data.shape)
print("The data we lost %s during cleaning"%(df.shape[0]-clean_data.shape[0]))
```

```
(5043, 28)
(3756, 28)
The data we lost 1287 during cleaning
```

In [26]:
```
df['language']=df['language'].fillna(value='No info')
df
```

Out[26]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_fac
0	Color	James Cameron	723.0	178.0	0.0	
1	Color	Gore Verbinski	302.0	169.0	563.0	
2	Color	Sam Mendes	602.0	148.0	0.0	
3	Color	Christopher Nolan	813.0	164.0	22000.0	
4	NaN	Doug Walker	NaN	NaN	131.0	
...	...	...	...	...	...	
5038	Color	Scott Smith	1.0	87.0	2.0	
5039	Color	NaN	43.0	43.0	NaN	
5040	Color	Benjamin Roberds	13.0	76.0	0.0	
5041	Color	Daniel Hsia	14.0	100.0	0.0	
5042	Color	Jon Gunn	43.0	90.0	16.0	

5043 rows × 28 columns

In [28]:
```
mean_bud = clean_data['budget'].mean()
df['budget']=df['budget'].fillna(value=mean_bud)
print(mean_bud)
df
```

```
46236849.637912676
```

Out[28]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_fac
0	Color	James Cameron	723.0	178.0	0.0	
1	Color	Gore Verbinski	302.0	169.0	563.0	
2	Color	Sam Mendes	602.0	148.0	0.0	
3	Color	Christopher Nolan	813.0	164.0	22000.0	
4	NaN	Doug Walker	NaN	NaN	131.0	
...	...	...	...	...	...	
5038	Color	Scott Smith	1.0	87.0	2.0	
5039	Color	NaN	43.0	43.0	NaN	
5040	Color	Benjamin Roberds	13.0	76.0	0.0	
5041	Color	Daniel Hsia	14.0	100.0	0.0	
5042	Color	Jon Gunn	43.0	90.0	16.0	

5043 rows × 28 columns

In [30]:
```
clean_data.isnull()
```

Out[30]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_fac
0	False	False	False	False	False	
1	False	False	False	False	False	
2	False	False	False	False	False	
3	False	False	False	False	False	
5	False	False	False	False	False	
...	...	...	...	...	...	
5026	False	False	False	False	False	
5027	False	False	False	False	False	
5033	False	False	False	False	False	
5035	False	False	False	False	False	
5042	False	False	False	False	False	

3756 rows × 28 columns

In [27]:
```
df['title_year'].fillna(method='ffill',limit = 3)
df
```

Out[27]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_fac
0	Color	James Cameron	723.0	178.0	0.0	
1	Color	Gore Verbinski	302.0	169.0	563.0	
2	Color	Sam Mendes	602.0	148.0	0.0	
3	Color	Christopher Nolan	813.0	164.0	22000.0	
4	NaN	Doug Walker	NaN	NaN	131.0	
...	...	...	...	...	...	
5038	Color	Scott Smith	1.0	87.0	2.0	
5039	Color	NaN	43.0	43.0	NaN	
5040	Color	Benjamin Roberds	13.0	76.0	0.0	
5041	Color	Daniel Hsia	14.0	100.0	0.0	
5042	Color	Jon Gunn	43.0	90.0	16.0	

5043 rows × 28 columns

In [ ]: