In [5]:
```
import pandas as pd
```

In [6]:
```
tmdb=pd.read_excel("/content/python work.xlsx")
```

In [7]:
```
tmdb
```

Out[7]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	
0	135397.0	tt0369610	32.985763	150000000.0	1.513529e+09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.
1	76341.0	tt1392190	28.419936	150000000.0	3.784364e+08	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.m
2	262500.0	tt2908446	13.112507	110000000.0	2.952382e+08	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries
3	140607.0	tt2488496	11.173104	200000000.0	2.068178e+09	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.co
4	168259.0	tt2820852	9.335014	190000000.0	1.506249e+09	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://v
...	...	...	...	...	...	...	...	
10864	21449.0	tt0061177	0.064317	0.0	0.000000e+00	What's Up, Tiger Lily?	Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...	

	id	imdb_id	popularity	budget	revenue	original_title	cast
10865	22293.0	tt0060666	0.035919	19000.0	0.000000e+00	Manos: The Hands of Fate	Harold P. Warren|Tom Neyman|John Reynolds|Dian...
10866	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10867	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10868	NaN	NaN	10866.000000	NaN	NaN	NaN	NaN

10869 rows × 21 columns

In [8]:
```
#datacleaning
#removing empty rows
```

In [9]:
```
tmdb.isnull().sum()
```

Out[9]:

	0
id	3
imdb_id	13
popularity	2
budget	3
revenue	3
original_title	3
cast	79
homepage	7933
director	47
tagline	2827
keywords	1496
overview	7
runtime	3
genres	26
production_companies	1033
release_date	3
vote_count	3
vote_average	3
release_year	3
budget_adj	3
revenue_adj	3

dtype: int64

In [10]:
```
tmdb=tmdb.dropna(how="all").reset_index(drop=True)
```

In [11]:
```
tmdb.duplicated().sum()
```

Out[11]:
```
np.int64(1)
```

In [12]:
```
# Drop exact duplicate rows
tmdb= tmdb.drop_duplicates().reset_index(drop=True)
```

In [13]:
```
tmdb.duplicated().sum()
```

Out[13]:
```
np.int64(0)
```

In [14]:
```
#checking data types
tmdb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   id                    10865 non-null  float64
 1   imdb_id               10855 non-null  object 
 2   popularity            10866 non-null  float64
 3   budget                10865 non-null  float64
 4   revenue               10865 non-null  float64
 5   original_title        10865 non-null  object 
 6   cast                  10789 non-null  object 
 7   homepage              2936 non-null   object 
 8   director              10821 non-null  object 
 9   tagline               8041 non-null   object 
 10  keywords              9372 non-null   object 
 11  overview              10861 non-null  object 
 12  runtime               10865 non-null  float64
 13  genres                10842 non-null  object 
 14  production_companies  9835 non-null   object 
 15  release_date          10865 non-null  object 
 16  vote_count            10865 non-null  float64
 17  vote_average          10865 non-null  float64
 18  release_year          10865 non-null  float64
 19  budget_adj            10865 non-null  float64
 20  revenue_adj           10865 non-null  float64
dtypes: float64(10), object(11)
memory usage: 1.7+ MB
```

In [15]:
```
# Convert 'id' column to Int64 (nullable integer type)
tmdb['id'] = tmdb['id'].astype('Int64')
```

In [16]:
```
#convert revenue and budget collumn to int64
tmdb['revenue']=tmdb['revenue'].astype('Int64')
tmdb['budget']=tmdb['budget'].astype('Int64')
```

In [17]:
```
#convert release date to datetime
tmdb['release_date']=pd.to_datetime(tmdb['release_date'])
```

In [18]:
```
#convert budget_adj and revenue_adj to Int64 after rounding
tmdb['budget_adj']=tmdb['budget_adj'].round().astype('Int64')
```

```
tmdb['revenue_adj']=tmdb['revenue_adj'].round().astype('Int64')
```

In [19]:
```
#convert release year to int
tmdb['release_year']=tmdb['release_year'].astype('Int64')
```

In [20]:
```
#convert vote_count to Int64
tmdb['vote_count']=tmdb['vote_count'].astype('Int64')
```

In [21]:
```
tmdb.head()
```

Out[21]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	home
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insu
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-v epi
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.

5 rows × 21 columns

In [22]:
```
#replace zeros in thes columns "budget", "revenue", "budget_adj", "revenue_adj", "runtime" as nan
tmdb['budget']=tmdb['budget'].replace(0,pd.NA)
tmdb['revenue']=tmdb['revenue'].replace(0,pd.NA)
```

```
tmdb['budget_adj']=tmdb['budget_adj'].replace(0,pd.NA)
tmdb['revenue_adj']=tmdb['revenue_adj'].replace(0,pd.NA)
tmdb['runtime']=tmdb['runtime'].replace(0,pd.NA)
```

In [23]:
```
tmdb.columns
```

Out[23]:
```
Index(['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_title',
       'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview',
       'runtime', 'genres', 'production_companies', 'release_date',
       'vote_count', 'vote_average', 'release_year', 'budget_adj',
       'revenue_adj'],
      dtype='object')
```

In [32]:
```
# Create the 'profit' column
tmdb['profit'] = tmdb['revenue'] - tmdb['budget']
```

In [33]:
```
tmdb.head()
```

Out[33]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	home
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insu
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-v epi
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.

5 rows × 22 columns

ANALYSIS

In [66]:
```
#total number of movies
tmdb['id'].count()
```

Out[66]:
```
np.int64(10865)
```

In [67]:
```
#total number of directors
tmdb['director'].nunique()
```

Out[67]:
```
5067
```

In [68]:
```
#total number of production companies
tmdb['production_companies'].nunique()
```

Out[68]:
```
7445
```

In [69]:
```
#years used for analysis
tmdb['release_year'].unique()
```

Out[69]:
```
<IntegerArray>
[2015, 2014, 1977, 2009, 2010, 1999, 2001, 2008, 2011, 2002, 1994, 2012, 2003,
 1997, 2013, 1985, 2005, 2006, 2004, 1972, 1980, 2007, 1979, 1984, 1983, 1995,
 1992, 1981, 1996, 2000, 1982, 1998, 1989, 1991, 1988, 1987, 1968, 1974, 1975,
 1962, 1964, 1971, 1990, 1961, 1960, 1976, 1993, 1967, 1963, 1986, 1973, 1970,
 1965, 1969, 1978, 1966, <NA>]
Length: 57, dtype: Int64
```

In [36]:
```
#TOP 10 MOVIES BY PROFIT AND VOTE COUNT

# Clean the 'original_title' column to ensure it's purely string for robust grouping
tmdb['original_title'] = tmdb['original_title'].fillna('Unknown Title').astype(str)

tmdb.groupby("original_title")[['profit','vote_count']].sum().sort_values(by="profit",ascending=False).head(10)
```

Out[36]:

original_title	profit	vote_count
Avatar	2544505847	8458
Star Wars: The Force Awakens	1868178225	5292
Titanic	1645034188	4675
Jurassic World	1363528810	5562
Furious 7	1316249360	2947
The Avengers	1288080742	9024
Harry Potter and the Deathly Hallows: Part 2	1202817822	3750
Avengers: Age of Ultron	1125035767	4304
Frozen	1124219009	3691
The Net	1084279658	201

In [81]:
```
 tmdb.groupby('release_year').agg(
    total_movies=('id', 'count'),
    total_profit=('profit', 'sum')
).sort_values(by='total_movies', ascending=False)

#yearly analysis by total number of movies and profit
```

Out[81]:

release_year	total_movies	total_profit
2014	700	16676201357
2013	659	15782743325
2015	629	19032145273
2012	588	16596845507
2011	540	14966694704
2009	533	14069305071
2008	496	11843373159
2010	489	13341222037
2007	438	11686103667
2006	408	8726299999
2005	364	9259452998
2004	307	9822506642
2003	281	9202155213
2002	266	9004589344
2001	242	8022044875
2000	227	6110774823
1999	224	6433209130
1998	210	5274145395
1996	204	4736626727
1997	192	6380241248
1994	184	4704384415

release_year	total_movies	total_profit
1993	178	4575828325
1995	175	5608098320
1988	145	2313076703
1989	137	3603940908
1991	133	3048121485
1992	133	3873046070
1990	132	3454904971
1987	125	2148546726
1986	121	1665354869
1985	109	1666899046
1984	105	1833438131
1981	82	1320251458
1982	81	1811756359
1983	80	1505073842
1980	78	1002065664
1978	65	1005100089
1977	57	1743308485
1979	57	950330150
1971	55	336284710
1973	55	1057686099
1974	47	650582027

release_year	total_movies	total_profit
1976	47	629426536
1966	46	52468689
1975	44	808207270
1964	42	294678387
1970	41	454070023
1967	40	636334637
1972	40	457235917
1968	39	168271762
1965	35	399281854
1963	34	115411882
1962	32	166879846
1960	32	108198052
1969	31	206862989
1961	31	299083188

In [30]:
```
#DIRECTORS WITH MOST MOVIES
tmdb.groupby("director")['id'].count().sort_values(ascending=False).head(10)
```

Out[30]:

director	id
Woody Allen	45
Clint Eastwood	34
Martin Scorsese	29
Steven Spielberg	29
Ridley Scott	23
Ron Howard	22
Steven Soderbergh	22
Joel Schumacher	21
Brian De Palma	20
Barry Levinson	19

dtype: Int64

In [48]:
```
#GENRE WITH THE HIGHEST AVERAGE BUDGET
tmdb.groupby("genres")['budget'].mean().sort_values(ascending=False).head(10)
```

Out[48]:

genres	budget
Adventure|Fantasy|Action|Western|Thriller	425000000.0
Thriller|Action|Adventure|Science Fiction	209000000.0
Action|Adventure|Science Fiction|Fantasy	200000000.0
Family|Fantasy|Adventure	200000000.0
Action|Drama|Horror|Science Fiction|Thriller	200000000.0
Adventure|Action|Fantasy	198000000.0
Action|Family|Fantasy	195000000.0
Action|Family|Science Fiction|Adventure|Mystery	190000000.0
Animation|Adventure|Comedy|Family|Action	185000000.0
Fantasy|Adventure|Action|Family|Romance	180000000.0

dtype: Float64

In [50]:
```
#GENRE WITH THE HISHEST PROFIT
tmdb.groupby("genres")['profit'].sum().sort_values(ascending=False).head(10)
```

Out[50]:

genres	profit
Comedy	12183078642
Drama	9050102799
Comedy|Romance	7822616677
Adventure|Fantasy|Action	5820583556
Action|Adventure|Science Fiction	4832602017
Adventure|Fantasy|Family	4799516484
Animation|Family	4780137559
Comedy|Drama|Romance	4525934209
Adventure|Action|Science Fiction	4332926988
Adventure|Action|Thriller	4007872468

dtype: Int64

In [84]:
```
#top 10 highest rated movies an their profit

tmdb.groupby("original_title")['vote_average'].mean().sort_values(ascending=False).head(10)
```

Out[84]:

original_title	vote_average
The Story of Film: An Odyssey	9.2
The Mask You Live In	8.9
Life Cycles	8.8
Black Mirror: White Christmas	8.8
Pink Floyd: Pulse	8.7
Opeth: In Live Concert At The Royal Albert Hall	8.6
The Art of Flight	8.5
Queen - Rock Montreal	8.5
Doctor Who: The Time of the Doctor	8.5
John Mayer: Where the Light Is Live in Los Angeles	8.5

dtype: float64

In [86]:
```
tmdb.groupby('production_companies').agg(
    total_movies=('id', 'count'),
    total_revenue=('revenue', 'sum')
).sort_values(by='total_movies', ascending=False).head(10)
#top 10 production companies by total movies and total revenue
```

Out[86]:

production_companies	total_movies	total_revenue
Paramount Pictures	156	8097344275
Universal Pictures	133	3792411919
Warner Bros.	84	2537558369
Walt Disney Pictures	76	3077561290
Columbia Pictures	72	4925479137
Metro-Goldwyn-Mayer (MGM)	72	1130564548
New Line Cinema	61	2386329840
Touchstone Pictures	51	2096772630
20th Century Fox	50	2061377892
Twentieth Century Fox Film Corporation	49	2515343511

In [ ]: