In [20]:

```python
# Import the numpy and pandas packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Task 1: Reading and Inspection

- ### Subtask 1.1: Import and read

Import and read the movie database. Store it in a variable called `movies`.

In [64]:

```python
#write your code here
movies = pd.read_csv('IMDB_Movies.csv')
OrgData = movies
movies
```

Out[64]:

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross	genres	...	num_user_for_reviews	language	country	content_rating	budget	title_year	actor_2_facebook_likes
0	Color	James Cameron	723.0	178.0	0.0	855.0	Joel David Moore	1000.0	760505847.0	Action|Adventure|Fantasy|Sci-Fi	...	3054	English	USA	PG-13	237000000.0	2009.0	936.0
1	Color	Gore Verbinski	302.0	169.0	563.0	1000.0	Orlando Bloom	40000.0	309404152.0	Action|Adventure|Fantasy	...	1238	English	USA	PG-13	300000000.0	2007.0	5000.0
2	Color	Sam Mendes	602.0	148.0	0.0	161.0	Rory Kinnear	11000.0	200074175.0	Action|Adventure|Thriller	...	994	English	UK	PG-13	245000000.0	2015.0	393.0
3	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0	Christian Bale	27000.0	448130642.0	Action|Thriller	...	2701	English	USA	PG-13	250000000.0	2012.0	23000.0
4	NaN	Doug Walker	NaN	NaN	131.0	NaN	Rob Walker	131.0	NaN	Documentary	...		NaN	NaN	NaN	NaN	NaN	12.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5038	Color	Scott Smith	1.0	87.0	2.0	318.0	Daphne Zuniga	637.0	NaN	Comedy|Drama	...	6	English	Canada	NaN	NaN	2013.0	470.0
5039	Color	NaN	43.0	43.0	NaN	319.0	Valorie Curry	841.0	NaN	Crime|Drama|Mystery|Thriller	...	359	English	USA	TV-14	NaN	NaN	593.0
5040	Color	Benjamin Roberds	13.0	76.0	0.0	0.0	Maxwell Moody	0.0	NaN	Drama|Horror|Thriller	...	3	English	USA	NaN	1400.0	2013.0	0.0
5041	Color	Daniel Hsia	14.0	100.0	0.0	489.0	Daniel Henney	946.0	10443.0	Comedy|Drama|Romance	...	9	English	USA	PG-13	NaN	2012.0	719.0
5042	Color	Jon Gunn	43.0	90.0	16.0	16.0	Brian Herzlinger	86.0	85222.0	Documentary	...	84	English	USA	PG	1100.0	2004.0	23.0

5043 rows × 28 columns

- ### Subtask 1.2: Inspect the dataframe

Inspect the dataframe's columns, shapes, variable types etc.

In [65]:

```python
#write your code here
movies.shape
```

Out[65]:

```
(5043, 28)
```

In [66]:

```python
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5043 entries, 0 to 5042
Data columns (total 28 columns):
color                        5024 non-null object
director_name                4939 non-null object
num_critic_for_reviews       4993 non-null float64
duration                     5028 non-null float64
director_facebook_likes      4939 non-null float64
actor_3_facebook_likes       5020 non-null float64
actor_2_name                 5030 non-null object
actor_1_facebook_likes       5036 non-null float64
gross                        4159 non-null float64
genres                       5043 non-null object
actor_1_name                 5036 non-null object
movie_title                  5043 non-null object
num_voted_users              5043 non-null int64
cast_total_facebook_likes    5043 non-null int64
actor_3_name                 5020 non-null object
facenumber_in_poster         5030 non-null float64
plot_keywords                4890 non-null object
movie_imdb_link              5043 non-null object
num_user_for_reviews         5023 non-null object
language                     5031 non-null object
country                      5038 non-null object
content_rating               4740 non-null object
budget                       4551 non-null float64
title_year                   4935 non-null float64
actor_2_facebook_likes       5030 non-null float64
imdb_score                   5043 non-null float64
aspect_ratio                 4714 non-null float64
movie_facebook_likes         5043 non-null int64
dtypes: float64(12), int64(3), object(13)
memory usage: 1.1+ MB
```

Task 2: Cleaning the Data

- ### Subtask 2.1: Inspect Null values

Find out the number of Null values in all the columns and rows. Also, find the percentage of Null values in each column. Round off the percentages upto two decimal places.

In [67]:

```python
# Write your code for column-wise null count here
movies.isnull().sum(axis=0).sort_values(ascending=False)
```

Out[67]:

```
gross                        884
budget                       492
aspect_ratio                 329
content_rating               303
plot_keywords                153
title_year                   108
director_name                104
director_facebook_likes      104
num_critic_for_reviews        50
actor_3_name                  23
actor_3_facebook_likes        23
num_user_for_reviews          20
color                         19
duration                      15
facenumber_in_poster          13
actor_2_name                  13
actor_2_facebook_likes        13
language                      12
actor_1_name                   7
actor_1_facebook_likes         7
country                        5
movie_facebook_likes           0
genres                         0
movie_title                    0
num_voted_users                0
movie_imdb_link                0
imdb_score                     0
cast_total_facebook_likes      0
dtype: int64
```

- ### Subtask 2.2: Drop unecessary columns

For this assignment, you will mostly be analyzing the movies with respect to the ratings, gross collection, popularity of movies, etc. So many of the columns in this dataframe are not required. So it is advised to drop the following columns.

- color
- director_facebook_likes
- actor_1_facebook_likes
- actor_2_facebook_likes
- actor_3_facebook_likes
- actor_2_name
- cast_total_facebook_likes
- actor_3_name
- duration
- facenumber_in_poster
- content_rating
- country
- movie_imdb_link
- aspect_ratio
- plot_keywords

In [70]:

```python
# Write your code for dropping the columns here. It is advised to keep inspecting the dataframe after each set of opera
movies = movies.drop([

    'color',
    'director_facebook_likes',
    'actor_1_facebook_likes',
    'actor_2_facebook_likes',
    'actor_3_facebook_likes',
    'actor_2_name',
    'cast_total_facebook_likes',
    'actor_3_name',
    'duration',
    'facenumber_in_poster',
    'content_rating',
    'country',
    'movie_imdb_link',
    'aspect_ratio',
    'plot_keywords'],axis=1)
```

In [71]:

```python
movies
```

Out[71]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes
0	James Cameron	723.0	760505847.0	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237000000.0	2009.0	7.9	33000
1	Gore Verbinski	302.0	309404152.0	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	300000000.0	2007.0	7.1	0
2	Sam Mendes	602.0	200074175.0	Action|Adventure|Thriller	Christoph Waltz	Spectre	275868	994	English	245000000.0	2015.0	6.8	85000
3	Christopher Nolan	813.0	448130642.0	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	250000000.0	2012.0	8.5	164000
4	Doug Walker	NaN	NaN	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens ...	8		NaN	NaN	NaN	7.1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5038	Scott Smith	1.0	NaN	Comedy|Drama	Eric Mabius	Signed Sealed Delivered	629	6	English	NaN	2013.0	7.7	84
5039	NaN	43.0	NaN	Crime|Drama|Mystery|Thriller	Natalie Zea	The Following	73839	359	English	NaN	NaN	7.5	32000
5040	Benjamin Roberds	13.0	NaN	Drama|Horror|Thriller	Eva Boehnke	A Plague So Pleasant	38	3	English	1400.0	2013.0	6.3	16
5041	Daniel Hsia	14.0	10443.0	Comedy|Drama|Romance	Alan Ruck	Shanghai Calling	1255	9	English	NaN	2012.0	6.3	660
5042	Jon Gunn	43.0	85222.0	Documentary	John August	My Date with Drew	4285	84	English	1100.0	2004.0	6.6	456

5043 rows × 13 columns

- ### Subtask 2.3: Drop unecessary rows using columns with high Null percentages

Now, on inspection you might notice that some columns have large percentage (greater than 5%) of Null values. Drop all the rows which have Null values for such columns.

In [72]:

```python
# Write your code for dropping the rows here
## round with 2 for 2 decimal points
round(movies.isnull().sum().sort_values(ascending=False)/len(movies)*100,2)
```

Out[72]:

```
gross                     17.53
budget                     9.76
title_year                 2.14
director_name              2.06
num_critic_for_reviews     0.99
num_user_for_reviews       0.40
language                   0.24
actor_1_name               0.14
movie_facebook_likes       0.00
imdb_score                 0.00
num_voted_users            0.00
movie_title                0.00
genres                     0.00
dtype: float64
```

In [73]:

```python
movies = movies[movies['gross'].notnull()]
movies = movies[movies['budget'].notnull()]
```

In [74]:

```python
round(movies.isnull().sum().sort_values(ascending=False)/len(movies)*100,2)
```

Out[74]:

```
language                  0.08
actor_1_name              0.08
num_critic_for_reviews    0.03
movie_facebook_likes      0.00
imdb_score                0.00
title_year                0.00
budget                    0.00
num_user_for_reviews      0.00
num_voted_users           0.00
movie_title               0.00
genres                    0.00
gross                     0.00
director_name             0.00
dtype: float64
```

- ### Subtask 2.4: Drop unecessary rows

Some of the rows might have greater than five NaN values. Such rows aren't of much use for the analysis and hence, should be removed.

In [75]:

```python
# Write your code for dropping the rows here
(movies.isnull().sum(axis=1).sort_values(ascending=False) > 5).sum()
```

Out[75]:

```
0
```

In [76]:

```python
movies = movies[movies.isnull().sum(axis=1).sort_values(ascending=False) <= 5]
movies
```

```
c:\users\karan\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  """Entry point for launching an IPython kernel.
```

Out[76]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes
0	James Cameron	723.0	760505847.0	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237000000.0	2009.0	7.9	33000
1	Gore Verbinski	302.0	309404152.0	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	300000000.0	2007.0	7.1	0
2	Sam Mendes	602.0	200074175.0	Action|Adventure|Thriller	Christoph Waltz	Spectre	275868	994	English	245000000.0	2015.0	6.8	85000
3	Christopher Nolan	813.0	448130642.0	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	250000000.0	2012.0	8.5	164000
5	Andrew Stanton	462.0	73058679.0	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter	212204	738	English	263700000.0	2012.0	6.6	24000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5033	Shane Carruth	143.0	424760.0	Drama|Sci-Fi|Thriller	Shane Carruth	Primer	72639	371	English	7000.0	2004.0	7.0	19000
5034	Neill Dela Llana	35.0	70071.0	Thriller	Ian Gamazon	Cavite	589	35	English	7000.0	2005.0	6.3	74
5035	Robert Rodriguez	56.0	2040920.0	Action|Crime|Drama|Romance|Thriller	Carlos Gallardo	El Mariachi	52055	130	Spanish	7000.0	1992.0	6.9	0
5037	Edward Burns	14.0	4584.0	Comedy|Drama	Kerry Bishé	Newlyweds	1338	14	English	9000.0	2011.0	6.4	413
5042	Jon Gunn	43.0	85222.0	Documentary	John August	My Date with Drew	4285	84	English	1100.0	2004.0	6.6	456

3891 rows × 13 columns

- ### Subtask 2.5: Fill NaN values

You might notice that the `language` column has some NaN values. Here, on inspection, you will see that it is safe to replace all the missing values with `'English'`.

In [77]:

```python
# Write your code here
round(movies.isnull().sum().sort_values(ascending=False)/len(movies)*100,2)
```

Out[77]:

```
language                  0.08
actor_1_name              0.08
num_critic_for_reviews    0.03
movie_facebook_likes      0.00
imdb_score                0.00
title_year                0.00
budget                    0.00
num_user_for_reviews      0.00
num_voted_users           0.00
movie_title               0.00
genres                    0.00
gross                     0.00
director_name             0.00
dtype: float64
```

In [78]:

```python
movies.groupby('language').language.count().sort_values(ascending=False)
```

Out[78]:

```
language
English       3707
French          37
Spanish         26
Mandarin        15
German          13
Japanese        12
Hindi           10
Cantonese        8
Italian          7
Korean           5
Portuguese       5
Norwegian        4
Hebrew           3
Persian          3
Dutch            3
Danish           3
Thai             3
Dari             2
Indonesian       2
Aboriginal       2
Icelandic        1
Hungarian        1
Arabic           1
Aramaic          1
Bosnian          1
Telugu           1
Czech            1
Swedish          1
Russian          1
Romanian         1
Dzongkha         1
None             1
Filipino         1
Mongolian        1
Maya             1
Kazakh           1
Vietnamese       1
Zulu             1
Name: language, dtype: int64
```

In [79]:

```python
movies.language.describe()
```

Out[79]:

```
count        3888
unique         38
top       English
freq         3707
Name: language, dtype: object
```

In [80]:

```python
movies.language = movies.language.fillna('English')
```

In [81]:

```python
round(movies.isnull().sum().sort_values(ascending=False)/len(movies)*100,2)
```

Out[81]:

```
actor_1_name              0.08
num_critic_for_reviews    0.03
movie_facebook_likes      0.00
imdb_score                0.00
title_year                0.00
budget                    0.00
language                  0.00
num_user_for_reviews      0.00
num_voted_users           0.00
movie_title               0.00
genres                    0.00
gross                     0.00
director_name             0.00
dtype: float64
```

- ### Subtask 2.6: Check the number of retained rows

You might notice that two of the columns viz. `num_critic_for_reviews` and `actor_1_name` have small percentages of NaN values left. You can let these columns as it is for now. Check the number and percentage of the rows retained after completing all the tasks above.

In [84]:

```python
# Write your code for checking number of retained rows here
len(movies)/len(OrgData) * 100
```

Out[84]:

```
77.15645449137418
```

Checkpoint 1: You might have noticed that we still have around `77%` of the rows!

Task 3: Data Analysis

- ### Subtask 3.1: Change the unit of columns Convert the unit of the `budget` and `gross` columns from `$` to `million $`.

In [87]:

```python
# Write your code for unit conversion here
movies['budget'] = movies['budget']/1000000
movies['gross'] = movies['gross']/1000000
```

In [88]:

```python
movies
```

Out[88]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes
0	James Cameron	723.0	760.505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237.0000	2009.0	7.9	33000
1	Gore Verbinski	302.0	309.404152	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	300.0000	2007.0	7.1	0
2	Sam Mendes	602.0	200.074175	Action|Adventure|Thriller	Christoph Waltz	Spectre	275868	994	English	245.0000	2015.0	6.8	85000
3	Christopher Nolan	813.0	448.130642	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	250.0000	2012.0	8.5	164000
5	Andrew Stanton	462.0	73.058679	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter	212204	738	English	263.7000	2012.0	6.6	24000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5033	Shane Carruth	143.0	0.424760	Drama|Sci-Fi|Thriller	Shane Carruth	Primer	72639	371	English	0.0070	2004.0	7.0	19000
5034	Neill Dela Llana	35.0	0.070071	Thriller	Ian Gamazon	Cavite	589	35	English	0.0070	2005.0	6.3	74
5035	Robert Rodriguez	56.0	2.040920	Action|Crime|Drama|Romance|Thriller	Carlos Gallardo	El Mariachi	52055	130	Spanish	0.0070	1992.0	6.9	0
5037	Edward Burns	14.0	0.004584	Comedy|Drama	Kerry Bishé	Newlyweds	1338	14	English	0.0090	2011.0	6.4	413
5042	Jon Gunn	43.0	0.085222	Documentary	John August	My Date with Drew	4285	84	English	0.0011	2004.0	6.6	456

3891 rows × 13 columns

- ### Subtask 3.2: Find the movies with highest profit

1. Create a new column called `profit` which contains the difference of the two columns: `gross` and `budget`.
2. Sort the dataframe using the `profit` column as reference.
3. Extract the top ten profiting movies in descending order and store them in a new dataframe - `top10`

In [89]:

```python
# Write your code for creating the profit column here
movies['profit'] = movies['gross'] - movies['budget']
movies
```

Out[89]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit
0	James Cameron	723.0	760.505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237.0000	2009.0	7.9	33000	523.505847
1	Gore Verbinski	302.0	309.404152	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	300.0000	2007.0	7.1	0	9.404152
2	Sam Mendes	602.0	200.074175	Action|Adventure|Thriller	Christoph Waltz	Spectre	275868	994	English	245.0000	2015.0	6.8	85000	-44.925825
3	Christopher Nolan	813.0	448.130642	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	250.0000	2012.0	8.5	164000	198.130642
5	Andrew Stanton	462.0	73.058679	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter	212204	738	English	263.7000	2012.0	6.6	24000	-190.641321
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5033	Shane Carruth	143.0	0.424760	Drama|Sci-Fi|Thriller	Shane Carruth	Primer	72639	371	English	0.0070	2004.0	7.0	19000	0.417760
5034	Neill Dela Llana	35.0	0.070071	Thriller	Ian Gamazon	Cavite	589	35	English	0.0070	2005.0	6.3	74	0.063071
5035	Robert Rodriguez	56.0	2.040920	Action|Crime|Drama|Romance|Thriller	Carlos Gallardo	El Mariachi	52055	130	Spanish	0.0070	1992.0	6.9	0	2.033920
5037	Edward Burns	14.0	0.004584	Comedy|Drama	Kerry Bishé	Newlyweds	1338	14	English	0.0090	2011.0	6.4	413	-0.004416
5042	Jon Gunn	43.0	0.085222	Documentary	John August	My Date with Drew	4285	84	English	0.0011	2004.0	6.6	456	0.084122

3891 rows × 14 columns

In [90]:

```python
# Write your code for sorting the dataframe here
movies.sort_values(by='profit',ascending=False)
```

Out[90]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit
0	James Cameron	723.0	760.505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237.000000	2009.0	7.9	33000	523.505847
29	Colin Trevorrow	644.0	652.177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	150.000000	2015.0	7.0	150000	502.177271
26	James Cameron	315.0	658.672302	Drama|Romance	Leonardo DiCaprio	Titanic	793059	2528	English	200.000000	1997.0	7.7	26000	458.672302
3024	George Lucas	282.0	460.935665	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	1470	English	11.000000	1977.0	8.7	33000	449.935665
3080	Steven Spielberg	215.0	434.949459	Family|Sci-Fi	Henry Thomas	E.T. the Extra-Terrestrial	281842	515	English	10.500000	1982.0	7.9	34000	424.449459
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2334	Katsuhiro Ôtomo	105.0	0.410388	Action|Adventure|Animation|Family|Sci-Fi|Thriller	William Hootkins	Steamboy	13727	79	Japanese	2127.519898	2004.0	6.9	973	-2127.109510
2323	Hayao Miyazaki	174.0	2.298191	Adventure|Animation|Fantasy	Minnie Driver	Princess Mononoke	221552	570	Japanese	2400.000000	1997.0	8.4	11000	-2397.701809
3005	Lajos Koltai	73.0	0.195888	Drama|Romance|War	Marcell Nagy	Fateless	5603	45	Hungarian	2500.000000	2005.0	7.1	607	-2499.804112
3859	Chan-wook Park	202.0	0.211667	Crime|Drama	Min-sik Choi	Lady Vengeance	53508	131	Korean	4200.000000	2005.0	7.7	4000	-4199.788333
2988	Joon-ho Bong	363.0	2.201412	Comedy|Drama|Horror|Sci-Fi	Doona Bae	The Host	68883	279	Korean	12215.500000	2006.0	7.0	7000	-12213.298588

3891 rows × 14 columns

In [91]:

```python
# Write your code for top10 movies
top10 = movies.sort_values(by='profit',ascending=False).head(10)
top10
```

Out[91]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit
0	James Cameron	723.0	760.505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237.0	2009.0	7.9	33000	523.505847
29	Colin Trevorrow	644.0	652.177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	150.0	2015.0	7.0	150000	502.177271
26	James Cameron	315.0	658.672302	Drama|Romance	Leonardo DiCaprio	Titanic	793059	2528	English	200.0	1997.0	7.7	26000	458.672302
3024	George Lucas	282.0	460.935665	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	1470	English	11.0	1977.0	8.7	33000	449.935665
3080	Steven Spielberg	215.0	434.949459	Family|Sci-Fi	Henry Thomas	E.T. the Extra-Terrestrial	281842	515	English	10.5	1982.0	7.9	34000	424.449459
794	Joss Whedon	703.0	623.279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	220.0	2012.0	8.1	123000	403.279547
17	Joss Whedon	703.0	623.279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	220.0	2012.0	8.1	123000	403.279547
509	Roger Allers	186.0	422.783777	Adventure|Animation|Drama|Family|Musical	Matthew Broderick	The Lion King	644348	656	English	45.0	1994.0	8.5	17000	377.783777
240	George Lucas	320.0	474.544677	Action|Adventure|Fantasy|Sci-Fi	Natalie Portman	Star Wars: Episode I - The Phantom Menace	534658	3597	English	115.0	1999.0	6.5	13000	359.544677
66	Christopher Nolan	645.0	533.316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark Knight	1676169	4667	English	185.0	2008.0	9.0	37000	348.316061

- ### Subtask 3.3: Drop duplicate values

After you found out the top 10 profiting movies, you might have notice a duplicate value. So, it seems like the dataframe has duplicate values as well. Drop the duplicate values from the dataframe and repeat `Subtask 3.2`.

In [92]:

```python
# Write your code for dropping duplicate values here
movies.drop_duplicates(keep='first',inplace=True)
```

In [93]:

```python
# Write code for repeating subtask 2 here
movies
```

Out[93]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit
0	James Cameron	723.0	760.505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237.0000	2009.0	7.9	33000	523.505847
1	Gore Verbinski	302.0	309.404152	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	300.0000	2007.0	7.1	0	9.404152
2	Sam Mendes	602.0	200.074175	Action|Adventure|Thriller	Christoph Waltz	Spectre	275868	994	English	245.0000	2015.0	6.8	85000	-44.925825
3	Christopher Nolan	813.0	448.130642	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	250.0000	2012.0	8.5	164000	198.130642
5	Andrew Stanton	462.0	73.058679	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter	212204	738	English	263.7000	2012.0	6.6	24000	-190.641321
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5033	Shane Carruth	143.0	0.424760	Drama|Sci-Fi|Thriller	Shane Carruth	Primer	72639	371	English	0.0070	2004.0	7.0	19000	0.417760
5034	Neill Dela Llana	35.0	0.070071	Thriller	Ian Gamazon	Cavite	589	35	English	0.0070	2005.0	6.3	74	0.063071
5035	Robert Rodriguez	56.0	2.040920	Action|Crime|Drama|Romance|Thriller	Carlos Gallardo	El Mariachi	52055	130	Spanish	0.0070	1992.0	6.9	0	2.033920
5037	Edward Burns	14.0	0.004584	Comedy|Drama	Kerry Bishé	Newlyweds	1338	14	English	0.0090	2011.0	6.4	413	-0.004416
5042	Jon Gunn	43.0	0.085222	Documentary	John August	My Date with Drew	4285	84	English	0.0011	2004.0	6.6	456	0.084122

3856 rows × 14 columns

In [94]:

```python
top10 = movies.sort_values(by='profit',ascending=False).head(10)
top10
```

Out[94]:

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit
0	James Cameron	723.0	760.505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	English	237.0	2009.0	7.9	33000	523.505847
29	Colin Trevorrow	644.0	652.177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	150.0	2015.0	7.0	150000	502.177271
26	James Cameron	315.0	658.672302	Drama|Romance	Leonardo DiCaprio	Titanic	793059	2528	English	200.0	1997.0	7.7	26000	458.672302
3024	George Lucas	282.0	460.935665	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	1470	English	11.0	1977.0	8.7	33000	449.935665
3080	Steven Spielberg	215.0	434.949459	Family|Sci-Fi	Henry Thomas	E.T. the Extra-Terrestrial	281842	515	English	10.5	1982.0	7.9	34000	424.449459
17	Joss Whedon	703.0	623.279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	220.0	2012.0	8.1	123000	403.279547
509	Roger Allers	186.0	422.783777	Adventure|Animation|Drama|Family|Musical	Matthew Broderick	The Lion King	644348	656	English	45.0	1994.0	8.5	17000	377.783777
240	George Lucas	320.0	474.544677	Action|Adventure|Fantasy|Sci-Fi	Natalie Portman	Star Wars: Episode I - The Phantom Menace	534658	3597	English	115.0	1999.0	6.5	13000	359.544677
66	Christopher Nolan	645.0	533.316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark Knight	1676169	4667	English	185.0	2008.0	9.0	37000	348.316061
439	Gary Ross	673.0	407.999255	Adventure|Drama|Sci-Fi|Thriller	Jennifer Lawrence	The Hunger Games	701607	1959	English	78.0	2012.0	7.3	140000	329.999255

In [ ]:

Checkpoint 2: You might spot two movies directed by `James Cameron` in the list.

- ### Subtask 3.4: Find IMDb Top 250

1. Create a new dataframe `IMDb_Top_250` and store the top 250 movies with the highest IMDb Rating (corresponding to the column: `imdb_score`). Also make sure that for all of these movies, the `num_voted_users` is greater than 25,000. Also add a `Rank` column containing the values 1 to 250 indicating the ranks of the corresponding films.
2. Extract all the movies in the `IMDb_Top_250` dataframe which are not in the English language and store them in a new dataframe named `Top_Foreign_Lang_Film`.

In [101…

```python
# Write your code for extracting the top 250 movies as per the IMDb score here. Make sure that you store it in a new dataframe
# and name that dataframe as 'IMDb_Top_250'
IMDb_Top_250 = movies[movies['num_voted_users'] > 25000].sort_values(by='imdb_score',ascending=False).head(250)
IMDb_Top_250
```

Out[101…

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit
1937	Frank Darabont	199.0	28.341469	Crime|Drama	Morgan Freeman	The Shawshank Redemption	1689764	4144	English	25.00	1994.0	9.3	108000	3.341469
3466	Francis Ford Coppola	208.0	134.821952	Crime|Drama	Al Pacino	The Godfather	1155770	2238	English	6.00	1972.0	9.2	43000	128.821952
2837	Francis Ford Coppola	149.0	57.300000	Crime|Drama	Robert De Niro	The Godfather: Part II	790926	650	English	13.00	1974.0	9.0	14000	44.300000
66	Christopher Nolan	645.0	533.316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark Knight	1676169	4667	English	185.00	2008.0	9.0	37000	348.316061
4498	Sergio Leone	181.0	6.100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	780	Italian	1.20	1966.0	8.9	20000	4.900000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4931	John Carney	232.0	9.437933	Drama|Music|Romance	Glen Hansard	Once	90827	329	English	0.18	2007.0	7.9	26000	9.257933
2605	Ang Lee	287.0	128.067808	Action|Drama|Romance	Chen Chang	Crouching Tiger, Hidden Dragon	217740	1641	Mandarin	15.00	2000.0	7.9	0	113.067808
3029	David O. Russell	410.0	93.571803	Biography|Drama|Sport	Christian Bale	The Fighter	275869	389	English	25.00	2010.0	7.9	36000	68.571803
2177	Tim Burton	111.0	56.362352	Fantasy|Romance	Johnny Depp	Edward Scissorhands	357581	588	English	20.00	1990.0	7.9	16000	36.362352
2487	George Cukor	82.0	72.000000	Drama|Family|Musical|Romance	Jeremy Brett	My Fair Lady	66959	258	English	17.00	1964.0	7.9	0	55.000000

250 rows × 14 columns

In [102…

```python
IMDb_Top_250['Rank'] = IMDb_Top_250['imdb_score'].rank(method='first',ascending=False)
IMDb_Top_250
```

Out[102…

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit	Rank
1937	Frank Darabont	199.0	28.341469	Crime|Drama	Morgan Freeman	The Shawshank Redemption	1689764	4144	English	25.00	1994.0	9.3	108000	3.341469	1.0
3466	Francis Ford Coppola	208.0	134.821952	Crime|Drama	Al Pacino	The Godfather	1155770	2238	English	6.00	1972.0	9.2	43000	128.821952	2.0
2837	Francis Ford Coppola	149.0	57.300000	Crime|Drama	Robert De Niro	The Godfather: Part II	790926	650	English	13.00	1974.0	9.0	14000	44.300000	3.0
66	Christopher Nolan	645.0	533.316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark Knight	1676169	4667	English	185.00	2008.0	9.0	37000	348.316061	4.0
4498	Sergio Leone	181.0	6.100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	780	Italian	1.20	1966.0	8.9	20000	4.900000	5.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4931	John Carney	232.0	9.437933	Drama|Music|Romance	Glen Hansard	Once	90827	329	English	0.18	2007.0	7.9	26000	9.257933	246.0
2605	Ang Lee	287.0	128.067808	Action|Drama|Romance	Chen Chang	Crouching Tiger, Hidden Dragon	217740	1641	Mandarin	15.00	2000.0	7.9	0	113.067808	247.0
3029	David O. Russell	410.0	93.571803	Biography|Drama|Sport	Christian Bale	The Fighter	275869	389	English	25.00	2010.0	7.9	36000	68.571803	248.0
2177	Tim Burton	111.0	56.362352	Fantasy|Romance	Johnny Depp	Edward Scissorhands	357581	588	English	20.00	1990.0	7.9	16000	36.362352	249.0
2487	George Cukor	82.0	72.000000	Drama|Family|Musical|Romance	Jeremy Brett	My Fair Lady	66959	258	English	17.00	1964.0	7.9	0	55.000000	250.0

250 rows × 15 columns

In [106…

```python
IMDb_Top_250[IMDb_Top_250['language']!='English']
```

Out[106…

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	profit	Rank
4498	Sergio Leone	181.0	6.100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	780	Italian	1.200000	1966.0	8.9	20000	4.900000	5.0
4747	Akira Kurosawa	153.0	0.269061	Action|Adventure|Drama	Takashi Shimura	Seven Samurai	229012	596	Japanese	2.000000	1954.0	8.7	11000	-1.730939	17.0
4029	Fernando Meirelles	214.0	7.563397	Crime|Drama	Alice Braga	City of God	533200	749	Portuguese	3.300000	2002.0	8.7	28000	4.263397	20.0
2373	Hayao Miyazaki	246.0	10.049886	Adventure|Animation|Family|Fantasy	Bunta Sugawara	Spirited Away	417971	902	Japanese	19.000000	2001.0	8.6	28000	-8.950114	23.0
4259	Florian Henckel von Donnersmarck	215.0	11.284657	Drama|Thriller	Sebastian Koch	The Lives of Others	259379	407	German	2.000000	2006.0	8.5	39000	9.284657	35.0
4921	Majid Majidi	46.0	0.925402	Drama|Family	Bahare Seddiqi	Children of Heaven	27882	130	Persian	0.180000	1997.0	8.5	0	0.745402	39.0
2323	Hayao Miyazaki	174.0	2.298191	Adventure|Animation|Fantasy	Minnie Driver	Princess Mononoke	221552	570	Japanese	2400.000000	1997.0	8.4	11000	-2397.701809	47.0
2970	Wolfgang Petersen	96.0	11.433134	Adventure|Drama|Thriller|War	Jürgen Prochnow	Das Boot	168203	426	German	14.000000	1981.0	8.4	11000	-2.566866	49.0
4105	Chan-wook Park	305.0	2.181290	Drama|Mystery|Thriller	Min-sik Choi	Oldboy	356181	809	Korean	3.000000	2003.0	8.4	43000	-0.818710	57.0
4659	Asghar Farhadi	354.0	7.098492	Drama|Mystery	Shahab Hosseini	A Separation	151812	264	Persian	0.500000	2011.0	8.4	48000	6.598492	58.0
1329	S.S. Rajamouli	44.0	6.498000	Action|Adventure|Drama|Fantasy|War	Tamannaah Bhatia	Baahubali: The Beginning	62756	410	Telugu	18.026148	2015.0	8.4	21000	-11.528148	59.0
1298	Jean-Pierre Jeunet	242.0	33.201661	Comedy|Romance	Mathieu Kassovitz	Amélie	534262	1314	French	77.000000	2001.0	8.4	39000	-43.798339	61.0
2734	Fritz Lang	260.0	0.026435	Drama|Sci-Fi	Brigitte Helm	Metropolis	111841	413	German	6.000000	1927.0	8.3	12000	-5.973565	65.0
4033	Thomas Vinterberg	349.0	0.610968	Drama	Thomas Bo Larsen	The Hunt	170155	249	Danish	3.800000	2012.0	8.3	60000	-3.189032	73.0
2829	Oliver Hirschbiegel	192.0	5.501940	Biography|Drama|History|War	Thomas Kretschmann	Downfall	248354	564	German	13.500000	2004.0	8.3	14000	-7.998060	81.0
2551	Guillermo del Toro	406.0	37.623143	Drama|Fantasy|War	Ivana Baquero	Pan's Labyrinth	467234	1083	Spanish	13.500000	2006.0	8.2	27000	24.123143	86.0
4000	Juan José Campanella	262.0	20.167424	Drama|Mystery|Thriller	Ricardo Darín	The Secret in Their Eyes	131831	231	Spanish	2.000000	2009.0	8.2	33000	18.167424	97.0
3550	Denis Villeneuve	226.0	6.857096	Drama|Mystery|War	Lubna Azabal	Incendies	80429	156	French	6.800000	2010.0	8.2	37000	0.057096	98.0
2047	Hayao Miyazaki	212.0	4.710455	Adventure|Animation|Family|Fantasy	Christian Bale	Howl's Moving Castle	214091	330	Japanese	24.000000	2004.0	8.2	13000	-19.289545	100.0
2830	Alejandro Amenábar	157.0	2.086345	Biography|Drama|Romance	Belén Rueda	The Sea Inside	64556	140	Spanish	10.000000	2004.0	8.1	0	-7.913655	110.0
2914	Je-kyu Kang	86.0	1.110186	Action|Drama|War	Min-sik Choi	Tae Guk Gi: The Brotherhood of War	31943	224	Korean	12.800000	2004.0	8.1	0	-11.689814	119.0
4461	Thomas Vinterberg	98.0	1.647780	Drama	Ulrich Thomsen	The Celebration	65951	258	Danish	1.300000	1998.0	8.1	5000	0.347780	127.0
3553	José Padilha	142.0	0.008060	Action|Crime|Drama|Thriller	Wagner Moura	Elite Squad	81644	107	Portuguese	4.000000	2007.0	8.1	11000	-3.991940	143.0
3423	Katsuhiro Ôtomo	150.0	0.439162	Action|Animation|Sci-Fi	Mitsuo Iwata	Akira	106160	430	Japanese	1100.000000	1988.0	8.1	0	-1099.560838	147.0
4267	Alejandro G. Iñárritu	157.0	5.383834	Drama|Thriller	Adriana Barraza	Amores Perros	173551	361	Spanish	2.000000	2000.0	8.1	11000	3.383834	152.0
3456	Vincent Paronnaud	242.0	4.443403	Animation|Biography|Drama|War	Catherine Deneuve	Persepolis	70194	158	French	7.300000	2007.0	8.0	14000	-2.856597	166.0
3344	Karan Johar	210.0	4.018695	Adventure|Drama|Thriller	Shah Rukh Khan	My Name Is Khan	69759	235	Hindi	12.000000	2010.0	8.0	27000	-7.981305	175.0
4144	Walter Salles	71.0	5.595428	Drama	Fernanda Montenegro	Central Station	28951	257	Portuguese	2.900000	1998.0	8.0	0	2.695428	177.0
4284	Ari Folman	231.0	2.283276	Animation|Biography|Documentary|Drama|History|War	Ari Folman	Waltz with Bashir	46107	156	Hebrew	1.500000	2008.0	8.0	0	0.783276	181.0
4897	Sergio Leone	122.0	3.500000	Action|Drama|Western	Clint Eastwood	A Fistful of Dollars	147566	235	Italian	0.200000	1964.0	8.0	0	3.300000	183.0
1171	Yimou Zhang	283.0	0.084961	Action|Adventure|History	Jet Li	Hero	149414	841	Mandarin	31.000000	2002.0	7.9	0	-30.915039	212.0
2863	Clint Eastwood	251.0	13.753931	Drama|History|War	Yuki Matsuzaki	Letters from Iwo Jima	132149	316	Japanese	19.000000	2006.0	7.9	5000	-5.246069	226.0
3264	Michael Haneke	447.0	0.225377	Drama|Romance	Isabelle Huppert	Amour	70382	190	French	8.900000	2012.0	7.9	33000	-8.674623	229.0
3510	Yash Chopra	29.0	2.921738	Drama|Musical|Romance	Shah Rukh Khan	Veer-Zaara	34449	119	Hindi	7.000000	2004.0	7.9	2000	-4.078262	233.0
3677	Christophe Barratier	112.0	3.629758	Drama|Music	Jean-Baptiste Maunier	The Chorus	44151	110	French	5.500000	2004.0	7.9	0	-1.870242	237.0
4415	Fabián Bielinsky	94.0	1.221261	Crime|Drama|Thriller	Ricardo Darín	Nine Queens	38215	125	Spanish	1.500000	2000.0	7.9	0	-0.278739	243.0
4640	Cristian Mungiu	233.0	1.185783	Drama	Anamaria Marinca	4 Months, 3 Weeks and 2 Days	44763	172	Romanian	0.590000	2007.0	7.9	14000	0.595783	244.0
2605	Ang Lee	287.0	128.067808	Action|Drama|Romance	Chen Chang	Crouching Tiger, Hidden Dragon	217740	1641	Mandarin	15.000000	2000.0	7.9	0	113.067808	247.0

Checkpoint 3: Can you spot `Veer-Zaara` in the dataframe?

- ### Subtask 3.5: Find the best directors

1. Group the dataframe using the `director_name` column.
2. Find out the top 10 directors for whom the mean of `imdb_score` is the highest and store them in a new dataframe `top10director`.

In [109…

```python
# Write your code for extracting the top 10 directors here
top10director = movies.groupby('director_name').imdb_score.mean().sort_values(ascending=False).head(10)
top10director
```

Out[109…

```
director_name
Charles Chaplin      8.600000
Tony Kaye            8.600000
Ron Fricke           8.500000
Damien Chazelle      8.500000
Majid Majidi         8.500000
Alfred Hitchcock     8.500000
Sergio Leone         8.433333
Christopher Nolan    8.425000
Asghar Farhadi       8.400000
Richard Marquand     8.400000
Name: imdb_score, dtype: float64
```

Checkpoint 4: No surprises that `Damien Chazelle` (director of Whiplash and La La Land) is in this list.