

```
In [1]: import pandas as pd
import numpy as np
import sklearn
```

```
In [2]: data = pd.read_csv('C:/Users/kommula_sravani/Desktop/tmdb-movies.csv')
```

```
In [3]: data.head()
```

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt Bryce Dallas Howard Irrfan Khan Vi...	http://www.jurassicworld.com/	Colin Trevorrow
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy Charlize Theron Hugh Keays-Byrne Nic...	http://www.madmaxmovie.com/	George Miller
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley Theo James Kate Winslet Ansel...	http://www.thedivergentseries.movie/#insurgent	Robert Schwentke
3	140607	tt2488496	11.713104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford Mark Hamill Carrie Fisher Adam D...	http://www.starwars.com/films/star-wars-episod...	J.J. Abrams
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel Paul Walker Jason Statham Michelle ...	http://www.furious7.com/	James Wan

5 rows x 21 columns

```
In [4]: data.shape
```

```
Out[4]: (10866, 21)
```

```
In [5]: data.describe()
```

	id	popularity	budget	revenue	runtime	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	10866.000000	10866.000000	1.086600e+04	1.086600e+04	10866.000000	10866.000000	10866.000000	10866.000000	1.086600e+04	1.0866
mean	66064.177434	0.646441	1.462570e+07	3.982332e+07	102.070863	217.389748	5.974922	2001.322658	1.755104e+07	5.1364
std	92130.136561	1.000185	3.091321e+07	1.170035e+08	31.381405	575.619058	0.935142	12.812941	3.430616e+07	1.4463
min	5.000000	0.000065	0.000000e+00	0.000000e+00	0.000000	10.000000	1.500000	1960.000000	0.000000e+00	0.0000
25%	10596.250000	0.207583	0.000000e+00	0.000000e+00	90.000000	17.000000	5.400000	1995.000000	0.000000e+00	0.0000
50%	20669.000000	0.383856	0.000000e+00	0.000000e+00	99.000000	38.000000	6.000000	2006.000000	0.000000e+00	0.0000
75%	75610.000000	0.713817	1.500000e+07	2.400000e+07	111.000000	145.750000	6.600000	2011.000000	2.085325e+07	3.3697
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	900.000000	9767.000000	9.200000	2015.000000	4.250000e+08	2.8271

```
In [6]: data.nunique()
```

id	10865
imdb_id	10855
popularity	10814
budget	557
revenue	4702
original_title	10571
cast	10719
homepage	2896
director	5067
tagline	7997
keywords	8804
overview	10847
runtime	247
genres	2039
production_companies	7445
release_date	5909
vote_count	1289
vote_average	72
release_year	56
budget_adj	2614
revenue_adj	4840
dtype:	int64

```
In [7]: movies = data.drop(['imdb_id', 'popularity', 'homepage', 'director', 'tagline', 'keywords', 'overview', 'production_companies'])
```

```
In [8]: movies.head()
```

	id	budget	revenue	original_title	cast	runtime	genres	release_year
0	135397	150000000	1513528810	Jurassic World	Chris Pratt Bryce Dallas Howard Irrfan Khan Vi...	124	Action Adventure Science Fiction Thriller	2015
1	76341	150000000	378436354	Mad Max: Fury Road	Tom Hardy Charlize Theron Hugh Keays-Byrne Nic...	120	Action Adventure Science Fiction Thriller	2015
2	262500	110000000	295238201	Insurgent	Shailene Woodley Theo James Kate Winslet Ansel...	119	Adventure Science Fiction Thriller	2015
3	140607	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford Mark Hamill Carrie Fisher Adam D...	136	Action Adventure Science Fiction Fantasy	2015
4	168259	190000000	1506249360	Furious 7	Vin Diesel Paul Walker Jason Statham Michelle ...	137	Action Crime Thriller	2015

```
In [9]: sum(movies.duplicated())
```

```
Out[9]: 1
```

```
In [10]: movies.drop_duplicates(inplace=True)
print(sum(movies.duplicated()))
print(movies.shape)
```

```
0
(10865, 8)
```

```
In [11]: movies.isnull().sum()
```

id	0
budget	0
revenue	0
original_title	0
cast	76
runtime	0
genres	23
release_year	0
dtype:	int64

```
In [12]: movies = movies.dropna()
```

```
In [13]: movies.isnull().sum()
```

id	0
budget	0
revenue	0
original_title	0
cast	0
runtime	0
genres	0
release_year	0
dtype:	int64

```
In [14]: movies
```

	id	budget	revenue	original_title	cast	runtime	genres	release_year
0	135397	150000000	1513528810	Jurassic World	Chris Pratt Bryce Dallas Howard Irrfan Khan Vi...	124	Action Adventure Science Fiction Thriller	2015
1	76341	150000000	378436354	Mad Max: Fury Road	Tom Hardy Charlize Theron Hugh Keays-Byrne Nic...	120	Action Adventure Science Fiction Thriller	2015
2	262500	110000000	295238201	Insurgent	Shailene Woodley Theo James Kate Winslet Ansel...	119	Adventure Science Fiction Thriller	2015
3	140607	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford Mark Hamill Carrie Fisher Adam D...	136	Action Adventure Science Fiction Fantasy	2015
4	168259	190000000	1506249360	Furious 7	Vin Diesel Paul Walker Jason Statham Michelle ...	137	Action Crime Thriller	2015
...
10861	21	0	0	The Endless Summer	Michael Hynson Robert August Lord 'Tally Ho' B...	95	Documentary	1966
10862	20379	0	0	Grand Prix	James Garner Eva Marie Saint Yves Montand Tosh...	176	Action Adventure Drama	1966
10863	39768	0	0	Beregis Avtomobilya	Innokentiy Smoktunovskiy Oleg Efremov Georgi Z...	94	Mystery Comedy	1966
10864	21449	0	0	What's Up, Tiger Lily?	Tatsuya Mihashi Akiko Wakabayashi Mie Hama Joh...	80	Action Comedy	1966
10865	22293	19000	0	Manos: The Hands of Fate	Harold P. Warren Tom Neyman John Reynolds Dian...	74	Horror	1966

10767 rows x 8 columns

```
In [15]: for x in movies.index:
if movies.loc[x, "budget"] < 10000:
    movies.drop(x, inplace = True)
```

```
In [16]: movies
```

	id	budget	revenue	original_title	cast	runtime	genres	release_year
0	135397	150000000	1513528810	Jurassic World	Chris Pratt Bryce Dallas Howard Irrfan Khan Vi...	124	Action Adventure Science Fiction Thriller	2015
1	76341	150000000	378436354	Mad Max: Fury Road	Tom Hardy Charlize Theron Hugh Keays-Byrne Nic...	120	Action Adventure Science Fiction Thriller	2015
2	262500	110000000	295238201	Insurgent	Shailene Woodley Theo James Kate Winslet Ansel...	119	Adventure Science Fiction Thriller	2015
3	140607	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford Mark Hamill Carrie Fisher Adam D...	136	Action Adventure Science Fiction Fantasy	2015
4	168259	190000000	1506249360	Furious 7	Vin Diesel Paul Walker Jason Statham Michelle ...	137	Action Crime Thriller	2015
...
10835	5923	12000000	20000000	The Sand Pebbles	Steve McQueen Richard Attenborough Richard Cre...	182	Action Adventure Drama War Romance	1966
10841	42701	75000	0	The Shooting	Will Hutchins Millie Perkins Jack Nicholson Wa...	82	Western	1966
10848	2161	5115000	12000000	Fantastic Voyage	Stephen Boyd Raqueel Welch Edmond O'Brien Donal...	100	Adventure Science Fiction	1966
10855	13343	700000	0	The Ghost & Mr. Chicken	Don Knotts Joan Staley Liam Redmond Dick Sarge...	90	Comedy Family Mystery Romance	1966
10865	22293	19000	0	Manos: The Hands of Fate	Harold P. Warren Tom Neyman John Reynolds Dian...	74	Horror	1966

5093 rows x 8 columns

```
In [17]: for x in movies.index:
if movies.loc[x, 'Revenue'] < 5000 :
    movies.drop(x,inplace = True)
```

```
In [18]: movies.drop(movies[movies['runtime']== 0].index, inplace = True)
```

```
Out[19]: movies
```

	id	budget	revenue	original_title	cast	runtime	genres	release_year
0	135397	150000000	1513528810	Jurassic World	Chris Pratt Bryce Dallas Howard Irrfan Khan Vi...	124	Action Adventure Science Fiction Thriller	2015
1	76341	150000000	378436354	Mad Max: Fury Road	Tom Hardy Charlize Theron Hugh Keays-Byrne Nic...	120	Action Adventure Science Fiction Thriller	2015
2	262500	110000000	295238201	Insurgent	Shailene Woodley Theo James Kate Winslet Ansel...	119	Adventure Science Fiction Thriller	2015
3	140607	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford Mark Hamill Carrie Fisher Adam D...	136	Action Adventure Science Fiction Fantasy	2015
4	168259	190000000	1506249360	Furious 7	Vin Diesel Paul Walker Jason Statham Michelle ...	137	Action Crime Thriller	2015
...
10822	396	7500000	33736689	Who's Afraid of Virginia Woolf?	Elizabeth Taylor Richard Burton George Segal S...	131	Drama	1966
10828	5780	3000000	13000000	Tom Curtain	Paul Newman Julie Andrews Lila Kedrova HansJf...	128	Mystery Thriller	1966
10829	6644	4653000	6000000	El Dorado	John Wayne Robert Mitchum James Caan Charlene ...	120	Action Western	1966
10835	5923	12000000	20000000	The Sand Pebbles	Steve McQueen Richard Attenborough Richard Cre...	182	Action Adventure Drama War Romance	1966
10848	2161	5115000	12000000	Fantastic Voyage	Stephen Boyd Raqueel Welch Edmond O'Brien Donal...	100	Adventure Science Fiction	1966

3807 rows x 8 columns

Which are the movies with the third lowest and third highest budget?

```
In [29]: budget_movies = movies.sort_values(by = 'budget', ascending=False).head(3)
budget_movies
```

	id	budget	revenue	original_title	cast	runtime	genres	release_year	year
2244	46528	425000000	11087569	The Warrior's Way	Kate Bosworth Jang Dong-gun Geoffrey Rush Dann...	100	Adventure Fantasy Action Western Thriller	2010	1970
3375	1865	380000000	1021683000	Pirates of the Caribbean: On Stranger Tides	Johnny Depp PenÃ©lope Cruz Geoffrey Rush Ian M...	136	Adventure Action Fantasy	2011	1970
7387	285	300000000	961000000	Pirates of the Caribbean: At World's End	Orlando Depp Johnny Lee Miller Keira Knightley Geof...	169	Adventure Fantasy Action	2007	1970

```
In [21]: budget_movies = movies.sort_values(by = 'budget', ascending=False).tail(3)
budget_movies
```

	id	budget	revenue	original_title	cast	runtime	genres	release_year
7447	23827	15000	193355800	Paranormal Activity	Katie Featherston Micah Sloate Mark Fredrichs A...	86	Horror Mystery	2007
7277	692	12000	6000000	Pink Flamingos	Divine David Lochary Mary Vivian Pearce Mink S...	93	Comedy	1972
1354	985	10000	7000000	Eraserhead	Jack Nance Charlotte Stewart Allen Joseph Jean...	89	Science Fiction Fantasy Horror Drama	1977

Which are the movies with most and least earned revenue?

```
In [22]: high = movies['revenue'].max()
movies[movies['revenue'] == high]['original_title']
```

```
Out[22]: 1386 Avatar
Name: original_title, dtype: object
```

```
In [23]: low = movies['revenue'].min()
movies[movies['revenue'] == low]['original_title']
```

```
Out[23]: 3752 The Good Doctor
Name: original_title, dtype: object
```

What is the average runtime of movies in the year 2006?

```
In [24]: movies.groupby('release_year').mean()['runtime']
```

1979	115.437500
1980	113.304348
1981	105.133333
1982	113.640000
1983	107.677419
1984	108.619048
1985	104.400000
1986	104.957447
1987	103.304348
1988	104.642857
1989	106.705882
1990	108.000000
1991	110.163265
1992	112.830189
1993	111.794118
1994	109.274194
1995	112.454545
1996	109.905882
1997	114.722222
1998	111.271739
1999	108.850877
2000	109.169811
2001	109.008264
2002	106.984000
2003	108.982906
2004	107.068966
2005	109.717791
2006	108.255952
2007	107.876543
2008	108.560241
2009	106.872832
2010	106.813559
2011	106.463918
2012	107.638158
2013	109.269663
2014	109.296970
2015	110.868750
Name:	runtime, dtype: float64

```
In [25]: movies[movies['release_year'] == 2006]['runtime'].mean()

Out[25]: 108.25595238095238
```

What is the average number of words in movie titles between the year 2000-2005?

```
In [26]: movies['year'] = pd.DatetimeIndex(movies['release_year']).year
```

```
In [25]: movies[movies['release_year'] == 2006]['runtime'].mean()
```

```
Out[25]: 108.25595238095238
```

What is the average number of words in movie titles between the year 2000-2005?

```
In [26]: movies['year'] = pd.datetimeIndex(movies['release_year']).year
movies['original_title'].str.split().str.len()
```

0	2
1	4
2	4
3	5
4	2
...	...
10822	5
10828	2
10829	2
10833	3
10848	2
Name:	original_title, Length: 3807, dtype: int64

```
In [27]: movies[movies['release_year'].between(2000,2005)]['original_title'].str.split().str.len().mean()
```

```
Out[27]: 2.694980694980695
```

What is the most common Genre for Vin Diesel & Emma Watson movies?

```
In [28]: def tit_for_tat(datal, column2, data1_column):
dataset_temp = movies.copy()
dataset_temp[data1_column] = dataset_temp[data1_column].str.contains(datal)
dataset_temp.drop(dataset_temp[dataset_temp[data1_column] == False].index, inplace = True)
dataset_temp.drop(['budget', "revenue", "original_title", "genres", "cast"])
dataset_temp.insert(0, "New_ID", range(0, len(dataset_temp)))

set_genre = set()
for temp_genre in dataset_temp[column2].str.split('|').values:
    if isinstance(temp_genre, float): continue
    set_genre = set_genre.union(temp_genre)

genre_count = dict()
for s in set_genre: genre_count[s] = 0
for temp_genre in dataset_temp[column2].str.split('|'):
    if temp_genre == float and pd.isnull(temp_genre): continue
    for s in [s for s in temp_genre if s in set_genre]:
        genre_count[s] += 1
genre_occurences = {}
for k,v in genre_count.items():
    genre_occurences.append([k,v])
genre_occurences.sort(key = lambda x:x[1], reverse = True)
return genre_occurences, genre_count

count, count1 = tit_for_tat("Vin Diesel", 'genres', 'cast')
print(str("The most common genre of movies performed by Vin Diesel is :- \n" + str(count[:1]) + "\n"))
count, count1 = tit_for_tat("Emma Watson", 'genres', 'cast')
print(str("The most common genre of movies performed by Emma Watson is :- \n" + str(count[:1])))

The most common genre of movies performed by Vin Diesel is :-
[['Action', 14]]

The
```