Investigate_a_Dataset

September 17, 2020

1 Explore IMDb movies data

1.1 Table of Contents

Introduction
Data Wrangling
Exploratory Data Analysis
Conclusions
Introduction
The data set that we are dealing with is collected from The Movie Database (TMDb) and contains a lot of important information about more than 4000 movie. the data explains a lot of movies features and how it affect the success or failure of each one to observe the features that make a movie succeed and how can we determine its success.

```
In [10]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
```

Data Wrangling
In this section, we are going to load our data, explore it and clean it to make it ready for making our analysis and driving our observations from it.

1.1.1 Loading and explore data :

```
In [12]: df = pd.read_csv('tmdb_5000_movies.csv')

         print(df.shape)
         df.head()
```

```
(4803, 20)
```

```
Out[12]:       budget                                             genres  \
         0  237000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
         1  300000000  [{"id": 12, "name": "Adventure"}, {"id": 14, "...
         2  245000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
         3  250000000  [{"id": 28, "name": "Action"}, {"id": 80, "nam...
         4  260000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
```

```
                                       homepage      id  \
0                   http://www.avatarmovie.com/   19995
1  http://disney.go.com/disneypictures/pirates/     285
2   http://www.sonypictures.com/movies/spectre/  206647
3             http://www.thedarkknightrises.com/   49026
4           http://movies.disney.com/john-carter   49529

                                            keywords original_language  \
0  [{"id": 1463, "name": "culture clash"}, {"id":...                 en
1  [{"id": 270, "name": "ocean"}, {"id": 726, "na...                 en
2  [{"id": 470, "name": "spy"}, {"id": 818, "name...                 en
3  [{"id": 849, "name": "dc comics"}, {"id": 853,...                 en
4  [{"id": 818, "name": "based on novel"}, {"id":...                 en

                             original_title  \
0                                    Avatar
1  Pirates of the Caribbean: At World's End
2                                   Spectre
3                     The Dark Knight Rises
4                                John Carter

                                            overview  popularity  \
0  In the 22nd century, a paraplegic Marine is di...  150.437577
1  Captain Barbossa, long believed to be dead, ha...  139.082615
2  A cryptic message from Bonds past sends him o...  107.376788
3  Following the death of District Attorney Harve...  112.312950
4  John Carter is a war-weary, former military ca...   43.926995

                                production_companies  \
0  [{"name": "Ingenious Film Partners", "id": 289...
1  [{"name": "Walt Disney Pictures", "id": 2}, {"...
2  [{"name": "Columbia Pictures", "id": 5}, {"nam...
3  [{"name": "Legendary Pictures", "id": 923}, {"...
4        [{"name": "Walt Disney Pictures", "id": 2}]

                                production_countries release_date     revenue  \
0  [{"iso_3166_1": "US", "name": "United States o...   2009-12-10  2787965087
1  [{"iso_3166_1": "US", "name": "United States o...   2007-05-19   961000000
2  [{"iso_3166_1": "GB", "name": "United Kingdom"...   2015-10-26   880674609
3  [{"iso_3166_1": "US", "name": "United States o...   2012-07-16  1084939099
4  [{"iso_3166_1": "US", "name": "United States o...   2012-03-07   284139100

   runtime                                     spoken_languages    status  \
0    162.0  [{"iso_639_1": "en", "name": "English"}, {"iso...  Released
1    169.0           [{"iso_639_1": "en", "name": "English"}]  Released
2    148.0  [{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...  Released
3    165.0           [{"iso_639_1": "en", "name": "English"}]  Released
```

```
4    132.0           [{"iso_639_1": "en", "name": "English"}]  Released

                                                tagline  \
0                      Enter the World of Pandora.
1  At the end of the world, the adventure begins.
2                             A Plan No One Escapes
3                                  The Legend Ends
4            Lost in our world, found in another.

                                     title  vote_average  vote_count
0                                   Avatar           7.2       11800
1  Pirates of the Caribbean: At World's End           6.9        4500
2                                  Spectre           6.3        4466
3                    The Dark Knight Rises           7.6        9106
4                              John Carter           6.1        2124
```

In [14]:
```
# Knowing the data type of each column :
df.dtypes
```

Out[14]:
```
budget                    int64
genres                   object
homepage                 object
id                        int64
keywords                 object
original_language        object
original_title           object
overview                 object
popularity              float64
production_companies     object
production_countries     object
release_date             object
revenue                   int64
runtime                 float64
spoken_languages         object
status                   object
tagline                  object
title                    object
vote_average            float64
vote_count                int64
dtype: object
```

In [15]:
```
# Knowing if any column contains null values :
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 20 columns):
budget                  4803 non-null int64
genres                  4803 non-null object
```

```
homepage                1712 non-null object
id                      4803 non-null int64
keywords                4803 non-null object
original_language       4803 non-null object
original_title          4803 non-null object
overview                4800 non-null object
popularity              4803 non-null float64
production_companies    4803 non-null object
production_countries    4803 non-null object
release_date            4802 non-null object
revenue                 4803 non-null int64
runtime                 4801 non-null float64
spoken_languages        4803 non-null object
status                  4803 non-null object
tagline                 3959 non-null object
title                   4803 non-null object
vote_average            4803 non-null float64
vote_count              4803 non-null int64
dtypes: float64(3), int64(4), object(13)
memory usage: 750.5+ KB
```

```
In [17]: # Let's see the difference between original_title and title :
         sum(df['original_title'] == df['title'])
```

```
Out[17]: 4542
```

As shown above, 4542 movie has the same 'title' and 'original_title'. So, lets reduce redundancy and remove one of them. But now let's see some of the rows when title and original_title are difference :

```
In [18]: df[df['original_title'] != df['title']]
```

```
Out[18]:          budget                                                genres  \
         97     15000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
         215   130000000  [{"id": 12, "name": "Adventure"}, {"id": 14, "...
         235    97250400  [{"id": 14, "name": "Fantasy"}, {"id": 12, "na...
         317    94000000  [{"id": 18, "name": "Drama"}, {"id": 36, "name...
         474           0  [{"id": 9648, "name": "Mystery"}, {"id": 18, "...
         488    86000000  [{"id": 12, "name": "Adventure"}, {"id": 14, "...
         492     8000000  [{"id": 35, "name": "Comedy"}, {"id": 16, "nam...
         561    74500000  [{"id": 12, "name": "Adventure"}, {"id": 18, "...
         678    65000000  [{"id": 28, "name": "Action"}, {"id": 18, "nam...
         719    60000000  [{"id": 10402, "name": "Music"}, {"id": 99, "n...
         786    68490000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
         861    47000000                       [{"id": 18, "name": "Drama"}]
         985    60000000  [{"id": 12, "name": "Adventure"}, {"id": 18, "...
         1023          0                       [{"id": 18, "name": "Drama"}]
         1028          0  [{"id": 18, "name": "Drama"}, {"id": 878, "nam...
```

```
1095        110  [{"id": 28, "name": "Action"}, {"id": 18, "nam...
1136   31000000  [{"id": 18, "name": "Drama"}, {"id": 12, "name...
1140   33000000  [{"id": 28, "name": "Action"}, {"id": 35, "nam...
1142          0  [{"id": 12, "name": "Adventure"}, {"id": 35, "...
1255   42000000  [{"id": 53, "name": "Thriller"}, {"id": 18, "n...
1260   10000000  [{"id": 35, "name": "Comedy"}, {"id": 10749, "...
1284   40000000  [{"id": 28, "name": "Action"}, {"id": 18, "nam...
1285   25000000  [{"id": 28, "name": "Action"}, {"id": 36, "nam...
1287   25000000  [{"id": 12, "name": "Adventure"}, {"id": 16, "...
1298   80341000  [{"id": 12, "name": "Adventure"}, {"id": 18, "...
1304    3860000  [{"id": 28, "name": "Action"}, {"id": 18, "nam...
1341   36500000  [{"id": 28, "name": "Action"}, {"id": 14, "nam...
1357   36000000  [{"id": 28, "name": "Action"}, {"id": 18, "nam...
1460   35000000  [{"id": 10752, "name": "War"}, {"id": 12, "nam...
1471   41677699  [{"id": 16, "name": "Animation"}, {"id": 10751...
...         ...                                               ...
4434     852510                      [{"id": 18, "name": "Drama"}]
4444          0                      [{"id": 18, "name": "Drama"}]
4452     800000                      [{"id": 18, "name": "Drama"}]
4457          0  [{"id": 18, "name": "Drama"}, {"id": 53, "name...
4459          0  [{"id": 18, "name": "Drama"}, {"id": 10769, "n...
4464     800000  [{"id": 27, "name": "Horror"}, {"id": 35, "nam...
4474     750000                      [{"id": 18, "name": "Drama"}]
4482     700000                      [{"id": 18, "name": "Drama"}]
4500          0  [{"id": 53, "name": "Thriller"}, {"id": 9648, ...
4505          0  [{"id": 18, "name": "Drama"}, {"id": 35, "name...
4510          0  [{"id": 35, "name": "Comedy"}, {"id": 18, "nam...
4535    2000000  [{"id": 28, "name": "Action"}, {"id": 18, "nam...
4536          0               [{"id": 99, "name": "Documentary"}]
4546          0  [{"id": 28, "name": "Action"}, {"id": 35, "nam...
4573          0  [{"id": 10749, "name": "Romance"}, {"id": 18, ...
4585          0                     [{"id": 27, "name": "Horror"}]
4591          0                      [{"id": 18, "name": "Drama"}]
4605          0                      [{"id": 18, "name": "Drama"}]
4609          0                     [{"id": 35, "name": "Comedy"}]
4615     300000  [{"id": 18, "name": "Drama"}, {"id": 53, "name...
4672     200000                   [{"id": 37, "name": "Western"}]
4677          0  [{"id": 10749, "name": "Romance"}, {"id": 18, ...
4684          0                     [{"id": 27, "name": "Horror"}]
4685          0               [{"id": 99, "name": "Documentary"}]
4695     180000  [{"id": 18, "name": "Drama"}, {"id": 35, "name...
4699          0                      [{"id": 18, "name": "Drama"}]
4719     120000  [{"id": 18, "name": "Drama"}, {"id": 10749, "n...
4751          0  [{"id": 18, "name": "Drama"}, {"id": 10749, "n...
4790          0  [{"id": 18, "name": "Drama"}, {"id": 10769, "n...
4792      20000  [{"id": 80, "name": "Crime"}, {"id": 27, "name...

                                                 homepage      id  \
```

97	NaN	315011
215	NaN	1979
235	http://www.asterixauxjeuxolympiques.com/index.php	2395
317	http://www.theflowersofwarmovie.com/	76758
474	NaN	330770
488	NaN	9992
492	NaN	293644
561	NaN	1997
678	NaN	300168
719	http://www.thisisit-movie.com	13576
786	http://themonkeykingmovie.com	381902
861	NaN	2841
985	NaN	10047
1023	NaN	7504
1028	NaN	593
1095	http://ent.sina.com.cn/hjj/	1494
1136	NaN	79
1140	NaN	27936
1142	http://pourquoijaipasmangemonpere.com/	280391
1255	NaN	80278
1260	http://www.die-fabelhafte-welt-der-amelie.de	194
1284	NaN	14392
1285	NaN	19495
1287	NaN	77459
1298	http://www.redclifffilm.com	12289
1304	NaN	44865
1341	http://oostrov.ru	16911
1357	NaN	365222
1460	NaN	11876
1471	NaN	12429
...	...	...
4434	NaN	2009
4444	NaN	146882
4452	NaN	60243
4457	NaN	905
4459	NaN	78705
4464	http://www.deadsnow.com	14451
4474	NaN	8416
4482	http://tartanvideo.com/film.asp?ProjectID={C66...	18734
4500	NaN	181940
4505	NaN	375950
4510	NaN	385636
4535	NaN	346
4536	http://theotherdreamteam.com/	84318
4546	NaN	45380
4573	http://www.wie-im-himmel-derfilm.de/start.html	464
4585	NaN	19204
4591	NaN	10238

```
4605                              http://www.dogtooth.gr/   38810
4609                                                  NaN   11956
4615                                                  NaN    2786
4672                                                  NaN     391
4677                                                  NaN   53256
4684                                                  NaN  402515
4685                                                  NaN  126141
4695   http://www.miramax.com/movie/children-of-heaven/   21334
4699                                                  NaN  344466
4719                                                  NaN   40652
4751                                                  NaN   42109
4790                                                  NaN   13898
4792                                                  NaN   36095

                                                  keywords original_language  \
97    [{"id": 1299, "name": "monster"}, {"id": 7671,...                ja
215   [{"id": 657, "name": "fire"}, {"id": 720, "nam...                en
235   [{"id": 271, "name": "competition"}, {"id": 12...                fr
317   [{"id": 173251, "name": "forced prostitution"}...                zh
474   [{"id": 428, "name": "nurse"}, {"id": 658, "na...                fr
488   [{"id": 1158, "name": "grandfather grandson re...                en
492                      [{"id": 209714, "name": "3d"}]                es
561   [{"id": 380, "name": "brother brother relation...                en
678                                                  []                zh
719   [{"id": 3490, "name": "pop star"}, {"id": 6027...                en
786             [{"id": 207411, "name": "monkey king"}]                zh
861   [{"id": 90, "name": "paris"}, {"id": 549, "nam...                fr
985   [{"id": 222, "name": "schizophrenia"}, {"id": ...                fr
1023  [{"id": 818, "name": "based on novel"}, {"id":...                hi
1028  [{"id": 1228, "name": "1970s"}, {"id": 1565, "...                ru
1095  [{"id": 351, "name": "poison"}, {"id": 478, "n...                zh
1136  [{"id": 548, "name": "countryside"}, {"id": 10...                zh
1140  [{"id": 156052, "name": "unemployment"}, {"id"...                fr
1142  [{"id": 10336, "name": "animation"}, {"id": 34...                fr
1255  [{"id": 3434, "name": "thailand"}, {"id": 6941...                en
1260  [{"id": 90, "name": "paris"}, {"id": 128, "nam...                fr
1284  [{"id": 782, "name": "assassin"}, {"id": 1402,...                zh
1285                                                 []                en
1287  [{"id": 1299, "name": "monster"}, {"id": 10987...                fr
1298  [{"id": 158145, "name": "flaming arrow"}, {"id...                zh
1304  [{"id": 779, "name": "martial arts"}, {"id": 7...                zh
1341  [{"id": 818, "name": "based on novel"}, {"id":...                ru
1357               [{"id": 5565, "name": "biography"}]                cn
1460  [{"id": 924, "name": "italian"}, {"id": 1605, ...                en
1471  [{"id": 456, "name": "mother"}, {"id": 1357, "...                ja
...                                                ...               ...
4434  [{"id": 570, "name": "rape"}, {"id": 922, "nam...                ro
4444  [{"id": 254, "name": "france"}, {"id": 2071, "...                fr
```

4452	[{"id": 1625, "name": "emigration"}, {"id": 60...	fa
4457	[{"id": 212, "name": "london england"}, {"id":...	de
4459	[]	en
4464	[{"id": 1191, "name": "norway"}, {"id": 10327,...	no
4474	[{"id": 90, "name": "paris"}, {"id": 131, "nam...	it
4482	[{"id": 10183, "name": "independent film"}]	en
4500	[]	en
4505	[{"id": 11612, "name": "hospital"}, {"id": 130...	fr
4510	[{"id": 445, "name": "pornography"}, {"id": 57...	sl
4535	[{"id": 233, "name": "japan"}, {"id": 1462, "n...	ja
4536	[{"id": 2070, "name": "olympic games"}, {"id":...	en
4546	[]	zh
4573	[{"id": 30, "name": "individual"}, {"id": 240,...	sv
4585	[{"id": 612, "name": "hotel"}, {"id": 1706, "n...	it
4591	[{"id": 1156, "name": "sister sister relations...	sv
4605	[{"id": 255, "name": "male nudity"}, {"id": 29...	el
4609	[{"id": 90, "name": "paris"}, {"id": 977, "nam...	fr
4615	[{"id": 90, "name": "paris"}, {"id": 5918, "na...	fr
4672	[{"id": 2987, "name": "gang war"}, {"id": 4942...	it
4677	[{"id": 572, "name": "sex"}, {"id": 154937, "n...	de
4684	[{"id": 321, "name": "terror"}, {"id": 8087, "...	pt
4685	[]	en
4695	[{"id": 1155, "name": "brother sister relation...	fa
4699	[]	ro
4719	[{"id": 1965, "name": "sandstorm"}, {"id": 151...	fr
4751	[]	pt
4790	[]	fa
4792	[{"id": 233, "name": "japan"}, {"id": 549, "na...	ja

	original_title \
97	
215	4: Rise of the Silver Surfer
235	Astérix aux Jeux Olympiques
317	
474	Évolution
488	Arthur et les Minimoys
492	Don Gato: El inicio de la pandilla
561	Deux frères
678	
719	Michael Jackson's This Is It
786	
861	Un long dimanche de fiançailles
985	Joan of Arc
1023	1947: Earth
1028	
1095	
1136	
1140	Micmacs à tire-larigot

```
1142                  Pourquoi j'ai pas mangé mon père
1255                                    Lo imposible
1260               Le fabuleux destin d'Amélie Poulain
1284
1285                                          Nomad
1287                                 Un monstre à Paris
1298                                       Chi bi
1304
1341                                  Obitaemyy Ostrov
1357                                               3
1460                              Le Hussard sur le toit
1471
...                                                ...
4434                      4 luni, 3 sptmîni i 2 zile
4444                                Le bonheur d'Elza
4452
4457                              Die Büchse der Pandora
4459                                       Cama adentro
4464                                            Død snø
4474                                    Il conformista
4482                     L.I.E. Long Island Expressway
4500                                       Arnolds Park
4505                                 Médecin de campagne
4510                                Julija in Alfa Romeo
4535
4536                                Kita svajoni komanda
4546
4573                                   Så som i himmelen
4585               ...E tu vivrai nel terrore! L'aldilà
4591                                 Viskningar och rop
4605
4609                              Chacun Cherche Son Chat
4615                                       Pierrot le fou
4672                              Per un pugno di dollari
4677                                           Drei
4684                                         Solitude
4685                                      Chats perchés
4695
4699                                      Lumea e a mea
4719  Une femme mariée: Suite de fragments d'un film...
4751                              Gabriela, Cravo e Canela
4790
4792
```

```
                                                overview  popularity  \
97    From the mind behind Evangelion comes a hit la...    9.476999
215   The Fantastic Four return to the big screen as...   60.810723
235   Astérix and Obélix have to win the Olympic Gam...   20.344364
```

317	A Westerner finds refuge with a group of women...	12.516546
474	11-year-old Nicolas lives with his mother in a...	3.300061
488	Arthur is a spirited ten-year old whose parent...	27.097932
492	Top Cat has arrived to charm his way into your...	0.719996
561	Two tigers are separated as cubs and taken int...	8.884318
678	Huo An, the commander of the Protection Squad ...	9.568884
719	A compilation of interviews, rehearsals and ba...	15.798622
786	Taking place 500 years after the Havoc in Heav...	4.726290
861	In 1919, Mathilde was 19 years old. Two years ...	23.054510
985	In 1429 a teenage girl from a remote French vi...	21.084542
1023	It's 1947 and the borderlines between India an...	1.246883
1028	Ground control has been receiving strange tran...	24.132271
1095	During China's Tang dynasty the emperor has ta...	9.950505
1136	One man defeated three assassins who sought to...	23.607392
1140	A man and his friends come up with an intricat...	7.663515
1142	Based on the novel 'Evolution Man' by Roy Lewi...	5.794466
1255	In December 2004, close-knit family Maria, Hen...	47.559928
1260	At a tiny Parisian café, the adorable yet pain...	73.720244
1284	A heroic tale of three blood brothers and thei...	6.884467
1285	The Nomad is a historical epic set in 18th-cen...	8.212018
1287	Paris,1910. Emile, a shy movie projectionist, ...	17.375562
1298	In the early third century, the land of Wu is ...	14.837500
1304	Ip Man's peaceful life in Foshan changes after...	19.947265
1341	On the threshold of 22nd century, furrowing th...	2.785832
1357	When a band of brutal gangsters led by a crook...	19.167377
1460	In a time of war and disease, a young officer ...	2.877488
1471	The son of a sailor, 5-year old Sosuke lives a...	39.586760
...	...	...
4434	Gabita is pregnant, abortion is strictly forbi...	9.270133
4444	A young Parisian woman of Caribbean descent re...	0.007254
4452	A married couple are faced with a difficult de...	12.049373
4457	The rise and inevitable fall of an amoral but ...	1.824184
4459	Buenos Aires is in a deep recession. As the mo...	0.038706
4464	Eight medical students on a ski trip to Norway...	11.205726
4474	A weak-willed Italian man becomes a fascist fl...	8.429295
4482	In this biting and disturbing coming-of-age ta...	7.731064
4500	When strangers Frank Delano and his Uncle Bobb...	0.006069
4505	All the people in this countryside area, can c...	2.651304
4510	Tilen (18), an attractive high school student,...	0.061248
4535	A samurai answers a village's request for prot...	39.756748
4536	The incredible story of the 1992 Lithuanian ba...	0.388401
4546	Three thieves try to steal a valuable jade tha...	1.685020
4573	A musical romantic tragedy about a famous comp...	4.321249
4585	A young woman inherits an old hotel in Louisia...	8.022122
4591	When a woman dying of cancer in turn-of-the ce...	11.347855
4605	Three teenagers are confined to an isolated co...	28.858238
4609	When Chloe (Garance Clavel), a young Parisian,...	1.410387
4615	Pierrot escapes his boring society and travels...	7.791898

```
4672  The Man With No Name enters the Mexican villag...   38.771062
4677  Hanna and Simon are in a 20 year marriage with...    5.937602
4684  After finding an old storage locker filled wit...    0.018716
4685  Paris 2002. Yellow cats appear on the walls. C...    0.092562
4695  Zohre's shoes are gone; her older brother Ali ...    7.072118
4699  Larisa is 16 and lives in a city by the sea. I...    0.327622
4719  Charlotte is young and modern, not a hair out ...    1.112792
4751  In 1925, Gabriela becomes cook, mistress, and ...    0.557602
4790  Various women struggle to function in the oppr...    1.193779
4792  A wave of gruesome murders is sweeping Tokyo. ...    0.212443

                                      production_companies  \
97    [{"name": "Cine Bazar", "id": 5896}, {"name": ...
215   [{"name": "Ingenious Film Partners", "id": 289...
235   [{"name": "Constantin Film", "id": 47}, {"name...
317   [{"name": "Beijing New Picture Film Co. Ltd.",...
474   [{"name": "Ex Nihilo", "id": 3307}, {"name": "...
488   [{"name": "Canal Plus", "id": 104}, {"name": "...
492   [{"name": "Anima Estudios", "id": 9965}, {"nam...
561   [{"name": "Path\u00e9 Renn Productions", "id":...
678   [{"name": "Shanghai Film Group", "id": 3407}, ...
719            [{"name": "Columbia Pictures", "id": 5}]
786           [{"name": "Filmko Pictures", "id": 9175}]
861   [{"name": "Gerber Pictures", "id": 975}, {"nam...
985   [{"name": "Columbia Pictures", "id": 5}, {"nam...
1023  [{"name": "Cracking the Earth Films", "id": 22...
1028  [{"name": "Mosfilm", "id": 5120}, {"name": "Cr...
1095  [{"name": "Beijing New Picture Film Co. Ltd.",...
1136  [{"name": "Beijing New Picture Film Co. Ltd.",...
1140  [{"name": "France 2 Cin\u00e9ma", "id": 83}, {...
1142  [{"name": "Path\u00e9 Films", "id": 4959}, {"n...
1255  [{"name": "Summit Entertainment", "id": 491}, ...
1260  [{"name": "France 3 Cin\u00e9ma", "id": 591}, ...
1284  [{"name": "Applause Pictures", "id": 5346}, {"...
1285  [{"name": "Wild Bunch", "id": 856}, {"name": "...
1287  [{"name": "Europa Corp", "id": 1075}, {"name":...
1298  [{"name": "Metropolitan Filmexport", "id": 656...
1304  [{"name": "The Weinstein Company", "id": 308},...
1341  [{"name": "Art Pictures Studio", "id": 3451}, ...
1357  [{"name": "Mandarin Films Distribution Co.", "...
1460  [{"name": "France 2 Cin\u00e9ma", "id": 83}, {...
1471  [{"name": "Studio Ghibli", "id": 10342}, {"nam...
...                                                 ...
4434  [{"name": "Saga Film", "id": 859}, {"name": "M...
4444  [{"name": "France T\u00e9l\u00e9vision", "id":...
4452  [{"name": "Asghar Farhadi Productions", "id": ...
4457              [{"name": "Nero Films", "id": 4903}]
4459  [{"name": "Libido Cine", "id": 28810}, {"name"...
```

```
4464  [{"name": "Euforia Film", "id": 3553}, {"name"...
4474  [{"name": "Paramount Pictures", "id": 4}, {"na...
4482                                                 []
4500  [{"name": "The Picture Factory", "id": 637}, {...
4505  [{"name": "France 2 Cin\u00e9ma", "id": 83}, {...
4510  [{"name": "Perfo Production", "id": 34679}, {"...
4535             [{"name": "Toho Company", "id": 882}]
4536                                                 []
4546                                                 []
4573  [{"name": "GF Studios AB", "id": 242}, {"name"...
4585            [{"name": "Fulvia Film", "id": 13682}]
4591  [{"name": "Cinematograph AB", "id": 7445}, {"n...
4605  [{"name": "Greek Film Center", "id": 7254}, {"...
4609  [{"name": "Vertigo Productions", "id": 2756}, ...
4615  [{"name": "Dino de Laurentiis Cinematografica"...
4672  [{"name": "United Artists", "id": 60}, {"name"...
4677  [{"name": "X-Filme Creative Pool", "id": 1972}...
4684       [{"name": "Gravitas Ventures", "id": 44632}]
4685      [{"name": "Les Films du Jeudi", "id": 54259}]
4695  [{"name": "The Institute for the Intellectual ...
4699                                                 []
4719  [{"name": "Orsay Films", "id": 2325}, {"name":...
4751  [{"name": "United Artists", "id": 60}, {"name"...
4790  [{"name": "Jafar Panahi Film Productions", "id...
4792            [{"name": "Daiei Studios", "id": 881}]

                                    production_countries release_date  \
97               [{"iso_3166_1": "JP", "name": "Japan"}]   2016-07-29
215   [{"iso_3166_1": "DE", "name": "Germany"}, {"is...   2007-06-13
235   [{"iso_3166_1": "BE", "name": "Belgium"}, {"is...   2008-01-13
317   [{"iso_3166_1": "CN", "name": "China"}, {"iso_...   2011-12-15
474   [{"iso_3166_1": "BE", "name": "Belgium"}, {"is...   2015-09-14
488            [{"iso_3166_1": "FR", "name": "France"}]   2006-12-13
492   [{"iso_3166_1": "IN", "name": "India"}, {"iso_...   2015-10-30
561   [{"iso_3166_1": "FR", "name": "France"}, {"iso...   2004-04-07
678   [{"iso_3166_1": "HK", "name": "Hong Kong"}, {"...   2015-02-19
719   [{"iso_3166_1": "US", "name": "United States o...   2009-10-28
786   [{"iso_3166_1": "HK", "name": "Hong Kong"}, {"...   2016-01-22
861            [{"iso_3166_1": "FR", "name": "France"}]   2004-10-26
985            [{"iso_3166_1": "FR", "name": "France"}]   1999-10-18
1023  [{"iso_3166_1": "CA", "name": "Canada"}, {"iso...   1998-09-16
1028           [{"iso_3166_1": "RU", "name": "Russia"}]   1972-03-20
1095  [{"iso_3166_1": "CN", "name": "China"}, {"iso_...   2006-12-21
1136            [{"iso_3166_1": "CN", "name": "China"}]   2002-12-19
1140           [{"iso_3166_1": "FR", "name": "France"}]   2009-10-28
1142           [{"iso_3166_1": "FR", "name": "France"}]   2015-04-08
1255  [{"iso_3166_1": "US", "name": "United States o...   2012-09-09
1260  [{"iso_3166_1": "FR", "name": "France"}, {"iso...   2001-04-25
```

```
1284  [{"iso_3166_1": "CN", "name": "China"}, {"iso_...   2007-12-12
1285  [{"iso_3166_1": "FR", "name": "France"}, {"iso...   2005-07-17
1287           [{"iso_3166_1": "FR", "name": "France"}]   2011-10-12
1298            [{"iso_3166_1": "CN", "name": "China"}]   2008-07-10
1304  [{"iso_3166_1": "CN", "name": "China"}, {"iso_...   2013-01-08
1341           [{"iso_3166_1": "RU", "name": "Russia"}]   2008-12-18
1357        [{"iso_3166_1": "HK", "name": "Hong Kong"}]   2015-12-19
1460           [{"iso_3166_1": "FR", "name": "France"}]   1995-09-20
1471            [{"iso_3166_1": "JP", "name": "Japan"}]   2008-07-19
...                                                 ...          ...
4434  [{"iso_3166_1": "BE", "name": "Belgium"}, {"is...   2007-08-24
4444  [{"iso_3166_1": "GP", "name": "Guadaloupe"}, {...   2011-04-11
4452  [{"iso_3166_1": "IR", "name": "Iran"}, {"iso_3...   2011-03-15
4457          [{"iso_3166_1": "DE", "name": "Germany"}]   1929-01-30
4459        [{"iso_3166_1": "AR", "name": "Argentina"}]   2004-09-21
4464           [{"iso_3166_1": "NO", "name": "Norway"}]   2009-01-09
4474  [{"iso_3166_1": "IT", "name": "Italy"}, {"iso_...   1970-10-21
4482  [{"iso_3166_1": "US", "name": "United States o...   2001-01-20
4500  [{"iso_3166_1": "US", "name": "United States o...   2007-07-20
4505           [{"iso_3166_1": "FR", "name": "France"}]   2016-03-23
4510         [{"iso_3166_1": "SI", "name": "Slovenia"}]   2015-09-17
4535            [{"iso_3166_1": "JP", "name": "Japan"}]   1954-04-26
4536                                                []   2012-09-28
4546  [{"iso_3166_1": "CN", "name": "China"}, {"iso_...   2006-01-01
4573           [{"iso_3166_1": "SE", "name": "Sweden"}]   2004-09-03
4585            [{"iso_3166_1": "IT", "name": "Italy"}]   1981-04-22
4591           [{"iso_3166_1": "SE", "name": "Sweden"}]   1972-12-21
4605           [{"iso_3166_1": "GR", "name": "Greece"}]   2009-06-01
4609           [{"iso_3166_1": "FR", "name": "France"}]   1996-04-03
4615  [{"iso_3166_1": "FR", "name": "France"}, {"iso...   1965-08-29
4672  [{"iso_3166_1": "IT", "name": "Italy"}, {"iso_...   1964-09-12
4677          [{"iso_3166_1": "DE", "name": "Germany"}]   2010-12-23
4684  [{"iso_3166_1": "US", "name": "United States o...   2014-10-21
4685           [{"iso_3166_1": "FR", "name": "France"}]   2004-12-05
4695             [{"iso_3166_1": "IR", "name": "Iran"}]   1997-08-01
4699          [{"iso_3166_1": "RO", "name": "Romania"}]   2015-06-05
4719           [{"iso_3166_1": "FR", "name": "France"}]   1964-12-04
4751           [{"iso_3166_1": "BR", "name": "Brazil"}]   1983-03-24
4790             [{"iso_3166_1": "IR", "name": "Iran"}]   2000-09-08
4792            [{"iso_3166_1": "JP", "name": "Japan"}]   1997-11-06

        revenue  runtime                                   spoken_languages  \
97     77000000    120.0  [{"iso_639_1": "it", "name": "Italiano"}, {"is...
215   289047763     92.0  [{"iso_639_1": "en", "name": "English"}, {"iso...
235   132900000    116.0  [{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...
317    95311434    145.0  [{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
474           0     81.0     [{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
488   107944236     94.0           [{"iso_639_1": "en", "name": "English"}]
```

492	0	89.0	[]
561	62172050	109.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
678	121545703	127.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
719	0	111.0	[{"iso_639_1": "en", "name": "English"}]
786	193677158	120.0	[{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
861	0	133.0	[{"iso_639_1": "co", "name": ""}, {"iso_639_1"...
985	66976317	148.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
1023	528972	101.0	[{"iso_639_1": "hi", "name": "\u0939\u093f\u09...
1028	0	167.0	[{"iso_639_1": "ru", "name": "P\u0443\u0441\u0...
1095	0	114.0	[{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
1136	177394432	99.0	[{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
1140	14000000	100.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
1142	0	100.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
1255	180274123	113.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
1260	173921954	122.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...
1284	0	126.0	[{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
1285	0	112.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
1287	0	90.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
1298	127814609	150.0	[{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
1304	64076736	130.0	[{"iso_639_1": "cn", "name": "\u5e7f\u5dde\u8b...
1341	21834845	115.0	[{"iso_639_1": "ru", "name": "P\u0443\u0441\u0...
1357	156844753	105.0	[{"iso_639_1": "cn", "name": "\u5e7f\u5dde\u8b...
1460	15000000	135.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...
1471	187479518	100.0	[{"iso_639_1": "ja", "name": "\u65e5\u672c\u8a...
...	...	...	...
4434	1185783	113.0	[{"iso_639_1": "ro", "name": "Rom\u00e2n\u0103"}]
4444	0	78.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
4452	0	123.0	[{"iso_639_1": "fa", "name": "\u0641\u0627\u06...
4457	0	109.0	[{"iso_639_1": "de", "name": "Deutsch"}]
4459	0	83.0	[{"iso_639_1": "es", "name": "Espa\u00f1ol"}]
4464	1984662	91.0	[{"iso_639_1": "no", "name": "Norsk"}]
4474	0	107.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
4482	1667192	97.0	[{"iso_639_1": "en", "name": "English"}]
4500	0	103.0	[{"iso_639_1": "en", "name": "English"}]
4505	0	102.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
4510	0	83.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
4535	271841	207.0	[{"iso_639_1": "ja", "name": "\u65e5\u672c\u8a...
4536	0	89.0	[{"iso_639_1": "en", "name": "English"}]
4546	0	98.0	[{"iso_639_1": "zh", "name": "\u666e\u901a\u8b...
4573	0	132.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
4585	0	87.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
4591	0	91.0	[{"iso_639_1": "sv", "name": "svenska"}]
4605	110197	94.0	[{"iso_639_1": "el", "name": "\u03b5\u03bb\u03...
4609	0	91.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
4615	0	110.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
4672	14500000	99.0	[{"iso_639_1": "it", "name": "Italiano"}, {"is...
4677	2611555	119.0	[{"iso_639_1": "en", "name": "English"}, {"iso...
4684	0	89.0	[{"iso_639_1": "en", "name": "English"}]

```
4685          0     59.0     [{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
4695     900000     89.0  [{"iso_639_1": "fa", "name": "\u0641\u0627\u06...
4699          0    104.0  [{"iso_639_1": "ro", "name": "Rom\u00e2n\u0103"}]
4719          0     95.0     [{"iso_639_1": "fr", "name": "Fran\u00e7ais"}]
4751          0     99.0    [{"iso_639_1": "pt", "name": "Portugu\u00eas"}]
4790          0     90.0  [{"iso_639_1": "fa", "name": "\u0641\u0627\u06...
4792      99000    111.0  [{"iso_639_1": "ja", "name": "\u65e5\u672c\u8a...

        status                                            tagline  \
97    Released                  A god incarnate. A city doomed.
215   Released                Discover the secret of the Surfer.
235   Released                                               NaN
317   Released                                               NaN
474   Released                                               NaN
488   Released            Adventure awaits in your own backyard.
492   Released                                               NaN
561   Released  Two infant tiger cubs, separated from their pa...
678   Released                   When the Eagle meets the Dragon
719   Released                Like You've Never Seen Him Before
786   Released                                               NaN
861   Released                                     Never let go
985   Released                                               NaN
1023  Released                                               NaN
1028  Released                                               NaN
1095  Released  Unspeakable secrets are hidden within the Forb...
1136  Released          One man's strength will unite an empire.
1140  Released                                  Non Stop Madness
1142  Released                  These feet are made for walkin'
1255  Released    Nothing is more powerful than the human spirit.
1260  Released          One person can change your life forever.
1284  Released                                               NaN
1285  Released                             Courage know no limit
1287  Released                                               NaN
1298  Released                       The future will be decided.
1304  Released  In Martial Arts there is no right or wrong, on...
1341  Released                                               NaN
1357  Released                                               NaN
1460  Released                                               NaN
1471  Released     Welcome To A World Where Anything Is Possible.
...        ...                                               ...
4434  Released               At what moment do we begin to live?
4444  Released                                               NaN
4452  Released                           Ugly truth, sweet lies.
4457  Released                                               NaN
4459  Released                                               NaN
4464  Released                                  Eins, Zwei, Die!
4474  Released    Bertolucci's Masterpiece about Sex and Politics
4482  Released  On the Long Island Expressway there are lanes ...
```

4500	Released	NaN
4505	Released	NaN
4510	Released	NaN
4535	Released	The Mighty Warriors Who Became the Seven Natio...
4536	Released	NaN
4546	Released	NaN
4573	Released	NaN
4585	Released	The seven dreaded gateways to Hell are conceal...
4591	Released	A haunting and shattering film experience.
4605	Released	The cat is the most feared animal there is!
4609	Released	No man. No job. No cat.
4615	Released	NaN
4672	Released	In his own way he is perhaps, the most dangero...
4677	Released	Imagine the possibilities.
4684	Released	NaN
4685	Released	NaN
4695	Released	A Little Secret...Their Biggest Adventure!
4699	Released	NaN
4719	Released	She Loves Two Men... She is Married to One!
4751	Released	NaN
4790	Released	NaN
4792	Released	Madness. Terror. Murder.

	title	vote_average	vote_count
97	Shin Godzilla	6.5	143
215	Fantastic 4: Rise of the Silver Surfer	5.4	2589
235	Asterix at the Olympic Games	5.0	471
317	The Flowers of War	7.1	187
474	Evolution	6.4	47
488	Arthur and the Invisibles	6.0	639
492	Top Cat Begins	5.3	9
561	Two Brothers	6.9	180
678	Dragon Blade	5.9	145
719	This Is It	6.7	247
786	The Monkey King 2	6.0	24
861	A Very Long Engagement	7.1	346
985	The Messenger: The Story of Joan of Arc	6.2	367
1023	Earth	6.6	9
1028	Solaris	7.7	357
1095	Curse of the Golden Flower	6.6	203
1136	Hero	7.2	635
1140	Micmacs	6.8	148
1142	Why I Did (Not) Eat My Father	5.3	125
1255	The Impossible	7.0	2025
1260	Amélie	7.8	3310
1284	The Warlords	6.3	83
1285	Nomad: The Warrior	4.3	17
1287	A Monster in Paris	6.5	313

1298	Red Cliff	7.1	205
1304	The Grandmaster	6.3	273
1341	The Inhabited Island	5.3	23
1357	Ip Man 3	6.5	379
1460	The Horseman on the Roof	6.7	25
1471	Ponyo	7.5	926
...	...	...	...
4434	4 Months, 3 Weeks and 2 Days	7.3	156
4444	Elza	0.0	0
4452	A Separation	7.7	469
4457	Pandora's Box	7.6	45
4459	Live-In Maid	7.8	3
4464	Dead Snow	6.1	311
4474	The Conformist	7.6	127
4482	L.I.E.	6.7	31
4500	Carousel of Revenge	0.0	0
4505	The Country Doctor	6.0	63
4510	Juliet and Alfa Romeo	6.0	1
4535	Seven Samurai	8.2	878
4536	The Other Dream Team	6.9	7
4546	Crazy Stone	6.9	22
4573	As It Is in Heaven	6.9	42
4585	The Beyond	6.6	117
4591	Cries and Whispers	7.8	115
4605	Dogtooth	6.9	332
4609	When the Cat's Away	6.1	16
4615	Pierrot le Fou	7.6	127
4672	A Fistful of Dollars	7.6	883
4677	Three	6.3	31
4684	American Beast	0.0	0
4685	The Case of the Grinning Cat	7.7	3
4695	Children of Heaven	7.8	112
4699	The World Is Mine	0.0	0
4719	The Married Woman	7.1	20
4751	Gabriela	6.0	2
4790	The Circle	6.6	17
4792	Cure	7.4	63

[261 rows x 20 columns]

As shown above, original_title column contains non English words that is not necessary for us. So, we are going to remove that column and keep 'title' column only.

1.1.2 Data Cleaning :

Removing unnecessary columns :

```
In [21]: # Removing 'original_title' column :
```

```
#df.drop('original_title' , axis=1 , inplace = True)
df.shape
```

```
Out[21]: (4803, 19)
```

As we can see from the dataframe info, the column 'homepage' contains only 1712 values out of 4803 and the rest is null values. This is completely not enough that we are going to remove that column too.

```
In [22]: # Remove homepage column :
         df.drop('homepage' , axis=1 , inplace = True)
         df.shape
```

```
Out[22]: (4803, 18)
```

Filling null values :

```
In [23]: # Replace null values in Overview column with 'Not exist'
         df['overview'].fillna('Not exist' , inplace=True)
```

Release_date column contains only one null value. So, lets see the whole row of that null value :

```
In [25]: df[df['release_date'].isnull()]
```

```
Out[25]:       budget genres      id keywords original_language  \
         4553       0     []  380097       []                en

                                                 overview  popularity  \
         4553  1971 post civil rights San Francisco seemed li...         0.0

              production_companies production_countries release_date  revenue  runtime  \
         4553                   []                   []          NaN        0      0.0

              spoken_languages    status tagline                       title  \
         4553               []  Released     NaN  America Is Still the Place

               vote_average  vote_count
         4553           0.0           0
```

As shown above, most of the values are missing with 'America Is Still the place' movie (that does not have realse date value), so it is better to remove the whole row

```
In [26]: # Removing the row that contains null release date value :
         df = df[df['release_date'].notnull()]
```

```
In [27]: df.shape
```

```
Out[27]: (4802, 18)
```

```
In [28]: # Replace null values in runtime column with its average :
         runtime_avg = df['runtime'].notna().mean()
         df['runtime'].fillna(runtime_avg , inplace=True)
```

```
/opt/conda/lib/python3.6/site-packages/pandas/core/generic.py:5434: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
  self._update_inplace(new_data)
```

```
In [29]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4802 entries, 0 to 4802
Data columns (total 18 columns):
budget                  4802 non-null int64
genres                  4802 non-null object
id                      4802 non-null int64
keywords                4802 non-null object
original_language       4802 non-null object
overview                4802 non-null object
popularity              4802 non-null float64
production_companies    4802 non-null object
production_countries    4802 non-null object
release_date            4802 non-null object
revenue                 4802 non-null int64
runtime                 4802 non-null float64
spoken_languages        4802 non-null object
status                  4802 non-null object
tagline                 3959 non-null object
title                   4802 non-null object
vote_average            4802 non-null float64
vote_count              4802 non-null int64
dtypes: float64(3), int64(4), object(11)
memory usage: 712.8+ KB
```

```
In [30]: # Replace null values in tagline column with 'Not exist'
         df['tagline'].fillna('Not exist' , inplace=True)
```

```
/opt/conda/lib/python3.6/site-packages/pandas/core/generic.py:5434: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
  self._update_inplace(new_data)
```

```
In [31]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4802 entries, 0 to 4802
Data columns (total 18 columns):
budget                  4802 non-null int64
genres                  4802 non-null object
id                      4802 non-null int64
keywords                4802 non-null object
original_language       4802 non-null object
overview                4802 non-null object
popularity              4802 non-null float64
production_companies    4802 non-null object
production_countries    4802 non-null object
release_date            4802 non-null object
revenue                 4802 non-null int64
runtime                 4802 non-null float64
spoken_languages        4802 non-null object
status                  4802 non-null object
tagline                 4802 non-null object
title                   4802 non-null object
vote_average            4802 non-null float64
vote_count              4802 non-null int64
dtypes: float64(3), int64(4), object(11)
memory usage: 712.8+ KB
```

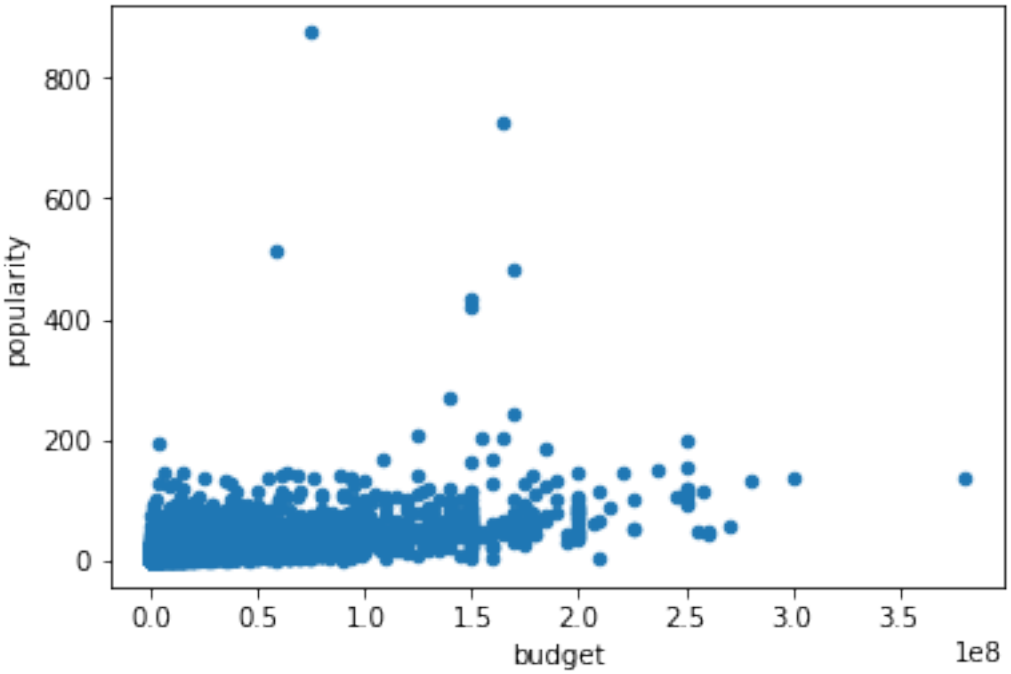
Exploratory Data Analysis

After cleaning our data, we're ready to move on to exploration. We are going to compute statistics and create visualizations so that we can analyse and answer some questions about the data.

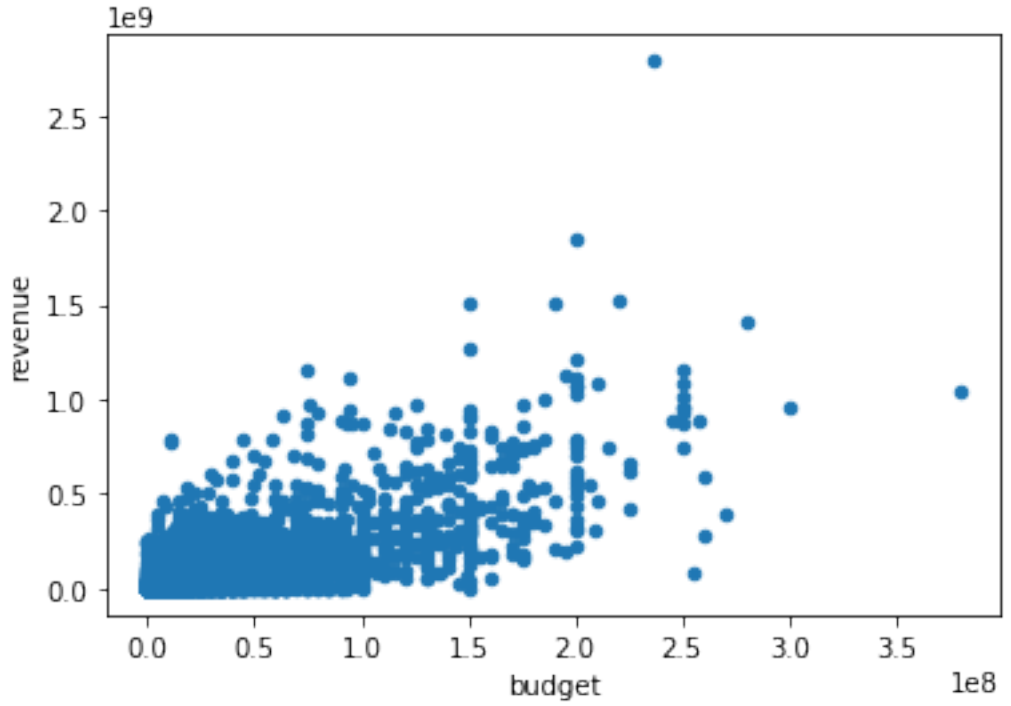
1.1.3 How does the movie budget affect its success?

Let's plot the movies budget with the dependent features that determine the success of the movie (popularity, revenue and vote_avrage) and see if there is a clear relationship between the movie budget and any of that features :

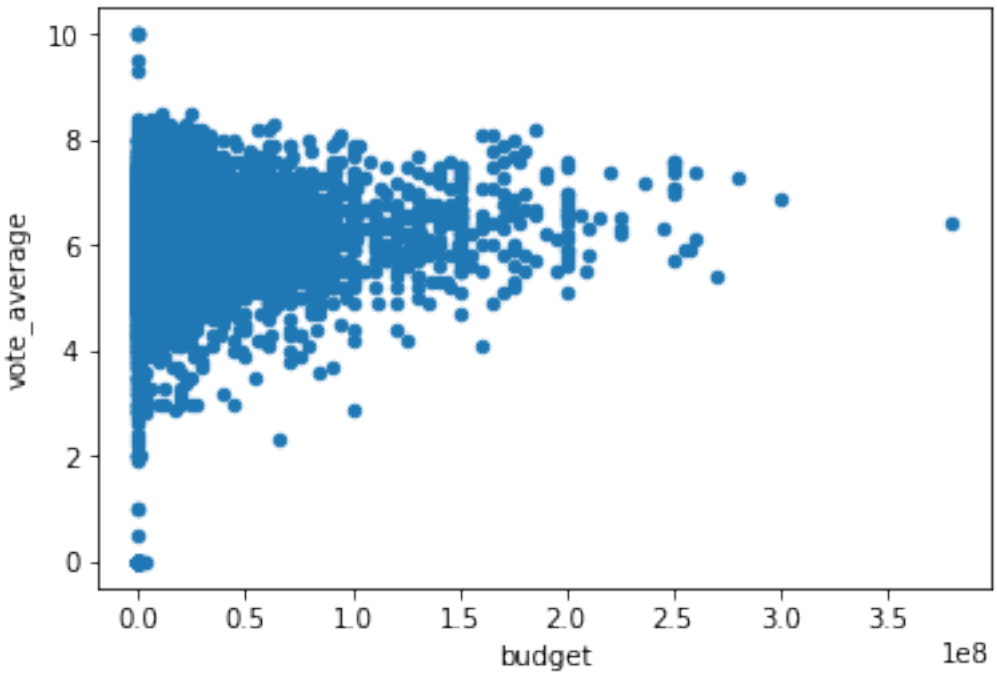
```
In [32]: # scatter plot between budget and popularity :
         df.plot(x='budget' , y='popularity' , kind='scatter');
```



In [33]:
```
# scatter plot between budget and popularity :
df.plot(x='budget' , y='revenue' , kind='scatter');
```



```
In [34]: # scatter plot between budget and popularity :
         df.plot(x='budget' , y='vote_average' , kind='scatter');
```



Observation : As seen from the above three plots, the most feature that has a releationship with budget is the revenue. It is very clear that revenue is directly propotional to the movie budget.

While the other two plots show that there isn't a clear relationship between the movie budget and the popularity of the movie or its vote average. Some of the low budget movies got very high popularity that high budget movies did not get it. Also the vote average has a common range that is between 4 and 8 despite of how much the movie budget was.

1.1.4 How do movie budget and revenue change overtime?

```
In [35]: df['release_date'].describe()

Out[35]: count           4802
         unique          3280
         top       2006-01-01
         freq              10
         Name: release_date, dtype: object
```

Now, we want to see on average how much the money spent and gained from movies changed between the 20th and 21st centuries.

So, let's Order the movies by their release date and get the first movie came in the 21st century so that we can split the movies into before and after 2000 :

```
In [37]: # getting the first movie came in the 21st century

         df['release_date'] = pd.to_datetime(df['release_date'])
         release_date_sorted = pd.DataFrame(df['release_date'].sort_values())
         release_date_sorted.reset_index(inplace=True)
         release_date_sorted.drop('index',axis=1,inplace=True)
         release_date_sorted.iloc[1308,0]
```

```
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:3: 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: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
  This is separate from the ipykernel package so we can avoid doing imports until
```

```
Out[37]: Timestamp('2000-01-01 00:00:00')
```

Getting all the movies that came in the 20th century and obtaining their budget and revenue averages :

```
In [38]: before_2000 = df.sort_values('release_date').iloc[0:1308,:];
         before_2000.shape
```

```
Out[38]: (1308, 18)
```

```
In [39]: print("average budget before 2000 : " , before_2000['budget'].mean())
         print("average revenue before 2000 : " , before_2000['revenue'].mean())
```

```
average budget before 2000 :  20843081.8677
average revenue before 2000 :  67828967.5673
```

Getting all the movies that came in the 21th century and obtaining their budget and revenue averages :

```
In [40]: after_2000 = df.sort_values('release_date').iloc[1308:,:];
         after_2000.shape
```

```
Out[40]: (3494, 18)
```

```
In [42]: print("average budget after 2000 : " , after_2000['budget'].mean())
         print("average revenue after 2000 : " , after_2000['revenue'].mean())
```

```
average budget after 2000 :  32123805.2198
average revenue after 2000 :  87686765.2736
```

Observation : We can see that the movies average budget and revenue in the 21st century increased by about 20 million over the movies in the 20th century and this is how the release date of the movie affects its budget and revenue.

1.1.5 Does the movies industry affected by the movie Original language ?

Let's start by seeing how many original languages exist in the data set and the count of each movie category :

```
In [43]: df['original_language'].value_counts()
```

```
Out[43]: en    4504
         fr      70
         es      32
         zh      27
         de      27
         hi      19
         ja      16
         it      14
         cn      12
         ru      11
         ko      11
         pt       9
         da       7
         sv       5
         nl       4
         fa       4
         he       3
         th       3
         cs       2
         ro       2
         ar       2
         ta       2
         id       2
         vi       1
         pl       1
         hu       1
         tr       1
         xx       1
         te       1
         sl       1
         ps       1
         ky       1
         no       1
         af       1
         nb       1
         el       1
         is       1
         Name: original_language, dtype: int64
```

Most of the movies has English language. So, when we are driving observations about the movies budget and revenue according to their original language, we should look for the average, as the sum would definitly be at its highest value with English movies (which are 4504 movie).

```
In [45]: df.groupby('original_language').vote_count.mean().sort_values(ascending=False)

Out[45]: original_language
         en    719.183393
         ja    715.750000
         id    632.500000
         ko    588.818182
         nb    583.000000
         da    450.428571
         it    403.928571
         el    332.000000
         pt    321.444444
         no    311.000000
         es    301.875000
         pl    260.000000
         de    242.444444
         fr    228.828571
         cn    220.916667
         fa    155.250000
         zh    146.592593
         te    135.000000
         he    112.666667
         xx    109.000000
         th    108.333333
         ru     95.181818
         af     94.000000
         nl     92.750000
         sv     87.000000
         ro     78.000000
         ar     53.500000
         hi     44.000000
         ps     32.000000
         hu     28.000000
         is     26.000000
         cs     12.000000
         tr      7.000000
         ta      4.000000
         sl      1.000000
         vi      1.000000
         ky      0.000000
         Name: vote_count, dtype: float64

In [46]: df.groupby('original_language').vote_average.mean().sort_values(ascending=False)

Out[46]: original_language
         te    7.500000
         id    7.400000
         he    7.400000
```

```
fa    7.375000
ar    7.300000
nl    7.175000
da    7.128571
pl    7.100000
xx    7.100000
sv    7.060000
ja    7.050000
it    7.028571
af    6.900000
el    6.900000
is    6.900000
nb    6.700000
ko    6.672727
es    6.659375
hu    6.500000
cn    6.500000
fr    6.430000
pt    6.388889
ru    6.354545
de    6.325926
zh    6.300000
ps    6.300000
no    6.100000
en    6.067029
hi    6.010526
sl    6.000000
th    5.966667
ta    5.850000
cs    5.650000
vi    5.000000
tr    4.300000
ro    3.650000
ky    0.000000
Name: vote_average, dtype: float64
```

```
In [47]: df.groupby('original_language').budget.mean().sort_values(ascending=False)
```

```
Out[47]: original_language
te    4.000000e+07
en    3.040111e+07
zh    2.202560e+07
ja    1.429361e+07
ko    1.429091e+07
ru    1.397273e+07
xx    1.200000e+07
cn    1.072641e+07
tr    1.000000e+07
```

```
da    9.742857e+06
de    8.616354e+06
fr    8.480997e+06
nl    7.375000e+06
es    5.847683e+06
sv    5.000000e+06
th    4.833333e+06
nb    3.500000e+06
af    3.000000e+06
it    2.967859e+06
pl    2.159280e+06
hi    1.715789e+06
vi    1.300000e+06
pt    1.133333e+06
id    1.050000e+06
no    8.000000e+05
he    6.666667e+05
ro    4.262550e+05
fa    2.450000e+05
ps    4.600000e+04
is    1.000000e+01
el    0.000000e+00
hu    0.000000e+00
ta    0.000000e+00
cs    0.000000e+00
sl    0.000000e+00
ar    0.000000e+00
ky    0.000000e+00
Name: budget, dtype: float64
```

```
In [48]: df.groupby('original_language').revenue.mean().sort_values(ascending=False)
```

```
Out[48]: original_language
te    1.000000e+08
en    8.649805e+07
ja    6.602892e+07
xx    5.526056e+07
zh    4.173498e+07
cn    3.374016e+07
da    2.989889e+07
ko    2.535645e+07
es    1.865218e+07
de    1.396191e+07
fr    1.246151e+07
th    1.203408e+07
pl    1.070000e+07
af    9.879971e+06
ru    9.510074e+06
```

```
hi    7.447231e+06
nl    6.680779e+06
nb    4.159678e+06
pt    4.026498e+06
he    3.708616e+06
it    3.029494e+06
id    2.274881e+06
no    1.984662e+06
vi    6.390000e+05
ro    5.928915e+05
fa    2.250000e+05
el    1.101970e+05
is    1.100000e+01
hu    0.000000e+00
ta    0.000000e+00
sv    0.000000e+00
ky    0.000000e+00
ps    0.000000e+00
tr    0.000000e+00
cs    0.000000e+00
ar    0.000000e+00
sl    0.000000e+00
Name: revenue, dtype: float64
```

Observation : We can observe that there are movies with some specific languages (English and Japanese for example) that have highest budget, revenue and they are the most common movies to have the highest vote_count.

Also we can see that these movies are not the ones with the highest vote average because of that high vote counts.

Number of votes extremly affected by the original language of the movie. We can see that English movies have the largest number of votes (3239202 votes for all English movies), while the summation of total votes for movies with some languages are just less than 10 votes!

Conclusions

Finally, we can conclude that each movie success can be determined according to some variables. These variables depend on a lot of components that we actually can use to make indecations about whether ther movie would succeed or not before it is even published.

```
In [50]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[50]: 0
```

```
In [ ]:
```