Investigate_a_Dataset

November 10, 2019

1 Project: Investigate a Dataset (The Movie Database 'TMDb')

1.1 Name : Mohamed Ahmed Makki

1.2 E-mail: ugtta@yahoo.com

1.3 Table of Contents

Introduction

my project in "Investigate a Dataset" will be in (The Movie Database 'TMDb'). this database available in a very famous website {The Movie Database (TMDb) | https://www.themoviedb.org} & {IMDb: Ratings and Reviews for New Movies and TV Shows | https://www.imdb.com}. containing all information and data about moviesreleased across many of years till nowadays. my job will be - importing the necessary packages for coding (numpy, pandas, matplotlib . . etc) - loading the proper/chosen dataset that will be investigated here. - data wrangling and cleaning - exploratory data analysis - then posing some quetions about this data. - collecting all information to answer the posed questions - draw necessary graphs that will ilustrate and answer our questions

In [2]:
```
#   import necessary packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Data Wrangling

starting with loading the data, check for cleanliness, and then trim and clean dataset for analysis. Make sure that

```
In [3]: # loading data (imdb-movies)
        df = pd.read_csv('https://d17h27t6h515a5.cloudfront.net/topher/2017/October/59dd1c4c_tmd
        df.head()
```

```
Out[3]:        id    imdb_id  popularity     budget     revenue  \
        0  135397  tt0369610   32.985763  150000000  1513528810
        1   76341  tt1392190   28.419936  150000000   378436354
        2  262500  tt2908446   13.112507  110000000   295238201
        3  140607  tt2488496   11.173104  200000000  2068178225
        4  168259  tt2820852    9.335014  190000000  1506249360

                     original_title  \
        0                Jurassic World
        1            Mad Max: Fury Road
        2                     Insurgent
        3  Star Wars: The Force Awakens
        4                     Furious 7

                                                        cast  \
        0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
        1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
        2  Shailene Woodley|Theo James|Kate Winslet|Ansel...
        3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
        4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                                homepage          director  \
        0                  http://www.jurassicworld.com/   Colin Trevorrow
        1                    http://www.madmaxmovie.com/     George Miller
        2     http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
        3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
        4                       http://www.furious7.com/         James Wan

                                tagline      ...        \
        0              The park is open.      ...
        1             What a Lovely Day.      ...
        2     One Choice Can Destroy You      ...
        3  Every generation has a story.      ...
        4            Vengeance Hits Home      ...

                                                overview runtime  \
        0  Twenty-two years after the events of Jurassic ...     124
        1  An apocalyptic story set in the furthest reach...     120
        2  Beatrice Prior must confront her inner demons ...     119
        3  Thirty years after defeating the Galactic Empi...     136
        4  Deckard Shaw seeks revenge against Dominic Tor...     137

                                             genres  \
```

```
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller
2         Adventure|Science Fiction|Thriller
3   Action|Adventure|Science Fiction|Fantasy
4                      Action|Crime|Thriller

                                  production_companies release_date vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15       5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15       6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15       2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15       5292
4  Universal Pictures|Original Film|Media Rights ...       4/1/15       2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
4           7.3          2015  1.747999e+08  1.385749e+09

[5 rows x 21 columns]
```

In [4]:
```
# size of table (rows, columns)
df.shape
```

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

In [5]:
```
df.duplicated().sum()
```

Out[5]:
```
1
```

In [6]:
```
df.drop_duplicates(inplace=True)
```

In [7]:
```
df.shape
```

Out[7]:
```
(10865, 21)
```

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10865 entries, 0 to 10865
Data columns (total 21 columns):
id                      10865 non-null int64
imdb_id                 10855 non-null object
popularity              10865 non-null float64
budget                  10865 non-null int64
revenue                 10865 non-null int64
original_title          10865 non-null object
cast                    10789 non-null object
```

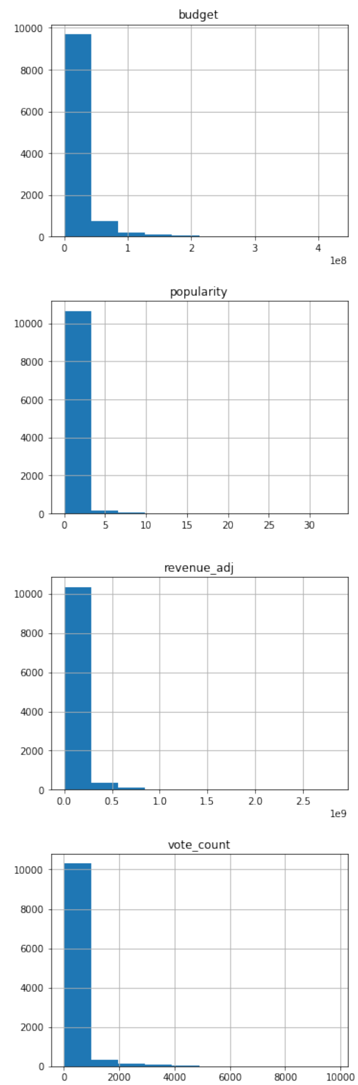
```
homepage                2936 non-null object
director                10821 non-null object
tagline                 8041 non-null object
keywords                9372 non-null object
overview                10861 non-null object
runtime                 10865 non-null int64
genres                  10842 non-null object
production_companies    9835 non-null object
release_date            10865 non-null object
vote_count              10865 non-null int64
vote_average            10865 non-null float64
release_year            10865 non-null int64
budget_adj              10865 non-null float64
revenue_adj             10865 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.8+ MB
```

1.3.1 Data Cleaning : by removing duplicated rows, removing some unwanted columns (in this invetigation), finding rows with missing Values. . and more as follows:

In [9]:
```
# After discussing the structure of the data and any problems that need to be
#   cleaned, perform those cleaning steps in the second part of this section.
df.describe()
```

Out[9]:
```
                  id    popularity        budget       revenue       runtime  \
count   10865.000000  10865.000000  1.086500e+04  1.086500e+04  10865.000000
mean    66066.374413      0.646446  1.462429e+07  3.982690e+07    102.071790
std     92134.091971      1.000231  3.091428e+07  1.170083e+08     31.382701
min         5.000000      0.000065  0.000000e+00  0.000000e+00      0.000000
25%     10596.000000      0.207575  0.000000e+00  0.000000e+00     90.000000
50%     20662.000000      0.383831  0.000000e+00  0.000000e+00     99.000000
75%     75612.000000      0.713857  1.500000e+07  2.400000e+07    111.000000
max    417859.000000     32.985763  4.250000e+08  2.781506e+09    900.000000

         vote_count  vote_average  release_year    budget_adj   revenue_adj
count  10865.000000  10865.000000  10865.000000  1.086500e+04  1.086500e+04
mean     217.399632      5.975012   2001.321859  1.754989e+07  5.136900e+07
std      575.644627      0.935138     12.813260  3.430753e+07  1.446383e+08
min       10.000000      1.500000   1960.000000  0.000000e+00  0.000000e+00
25%       17.000000      5.400000   1995.000000  0.000000e+00  0.000000e+00
50%       38.000000      6.000000   2006.000000  0.000000e+00  0.000000e+00
75%      146.000000      6.600000   2011.000000  2.085325e+07  3.370173e+07
max     9767.000000      9.200000   2015.000000  4.250000e+08  2.827124e+09
```

In [10]:
```
df.hist(figsize=(20,20));
```




```
In [11]: # lets find columns with missing values
         df.columns[df.isnull().any()]
```

```
Out[11]: Index(['imdb_id', 'cast', 'homepage', 'director', 'tagline', 'keywords',
                'overview', 'genres', 'production_companies'],
               dtype='object')
```

```
In [12]: # counting the missing values in each columns
         df.isnull().sum()
```

```
Out[12]: id                          0
         imdb_id                    10
         popularity                  0
         budget                      0
```

```
revenue                    0
original_title             0
cast                      76
homepage                7929
director                  44
tagline                 2824
keywords                1493
overview                   4
runtime                    0
genres                    23
production_companies    1030
release_date               0
vote_count                 0
vote_average               0
release_year               0
budget_adj                 0
revenue_adj                0
dtype: int64
```

In [13]:
```
# we have to drop some columns that we are not going to us becuase they are out of our
df.drop(['imdb_id', 'cast', 'homepage', 'tagline', 'keywords',
       'overview', 'production_companies', 'budget', 'revenue'], axis =1, inplace=True)
```

In [14]:
```
df.shape
```

Out[14]:
```
(10865, 12)
```

In [15]:
```
# let's drop the these rows with missing data (NaN)
df.dropna(inplace=True)
```

In [16]:
```
# counting again the missing values in each columns after cleaning some columns
df.isnull().sum()
```

Out[16]:
```
id                0
popularity        0
original_title    0
director          0
runtime           0
genres            0
release_date      0
vote_count        0
vote_average      0
release_year      0
budget_adj        0
revenue_adj       0
dtype: int64
```

In [17]:
```
# our final data after cleaning some columns and rows
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10800 entries, 0 to 10865
Data columns (total 12 columns):
id                10800 non-null int64
popularity        10800 non-null float64
original_title    10800 non-null object
director          10800 non-null object
runtime           10800 non-null int64
genres            10800 non-null object
release_date      10800 non-null object
vote_count        10800 non-null int64
vote_average      10800 non-null float64
release_year      10800 non-null int64
budget_adj        10800 non-null float64
revenue_adj       10800 non-null float64
dtypes: float64(4), int64(4), object(4)
memory usage: 1.1+ MB
```

In [18]:
```
# we get rid of rows with missing data, but let's chech if some columns have zero value
df.loc[df.budget_adj ==0]
```

Out[18]:

	id	popularity	original_title \
30	280996	3.927333	Mr. Holmes
36	339527	3.358321	Solace
72	284289	2.272044	Beyond the Reach
74	347096	2.165433	Mythica: The Darkspore
75	308369	2.141506	Me and Earl and the Dying Girl
88	301875	1.959765	Equals
92	370687	1.876037	Mythica: The Necromancer
95	258509	1.841779	Alvin and the Chipmunks: The Road Chip
100	326359	1.724712	Frozen Fever
101	254302	1.661789	High-Rise
103	292040	1.646664	Spooks: The Greater Good
116	297291	1.380320	The Scorpion King: The Lost Throne
119	86828	1.360827	Absolutely Anything
122	277355	1.342839	Everly
125	223485	1.329702	Slow West
128	309245	1.293140	Mistress America
130	245706	1.284541	True Story
132	263109	1.253580	Shaun the Sheep Movie
134	321751	1.245224	A Perfect Day
139	193687	1.161812	Z for Zachariah
140	300803	1.144808	Dragonheart 3: The Sorcerer's Curse
143	378373	1.128081	Brothers of the Wind
146	241257	1.065888	Regression
147	245698	1.063055	Pawn Sacrifice
148	353326	1.046518	The Man Who Knew Infinity

151	290637	1.036825	Pay the Ghost
152	244458	1.027620	The Voices
153	308504	1.021441	Last Knights
158	290762	0.953647	Miss You Already
161	324807	0.938432	A Bigger Splash
...	...	...	...
10830	4772	0.380321	Cul-de-sac
10831	1888	0.529721	The Fortune Cookie
10833	3001	0.737730	How to Steal a Million
10834	12639	0.310688	Return of the Seven
10836	38720	0.239435	Walk Don't Run
10837	19728	0.291704	The Blue Max
10838	22383	0.151845	The Professionals
10839	13353	0.276133	It's the Great Pumpkin, Charlie Brown
10840	34388	0.102530	Funeral in Berlin
10842	36540	0.253437	Winnie the Pooh and the Honey Tree
10843	29710	0.252399	Khartoum
10844	23728	0.236098	Our Man Flint
10845	5065	0.230873	Carry On Cowboy
10846	17102	0.212716	Dracula: Prince of Darkness
10847	28763	0.034555	Island of Terror
10849	28270	0.206537	Gambit
10850	26268	0.202473	Harper
10851	15347	0.342791	Born Free
10852	37301	0.227220	A Big Hand for the Little Lady
10853	15598	0.163592	Alfie
10854	31602	0.146402	The Chase
10856	20277	0.140934	The Ugly Dachshund
10857	5921	0.131378	Nevada Smith
10858	31918	0.317824	The Russians Are Coming, The Russians Are Coming
10859	20620	0.089072	Seconds
10860	5060	0.087034	Carry On Screaming!
10861	21	0.080598	The Endless Summer
10862	20379	0.065543	Grand Prix
10863	39768	0.065141	Beregis Avtomobilya
10864	21449	0.064317	What's Up, Tiger Lily?

	director	runtime	\
30	Bill Condon	103	
36	Afonso Poyart	101	
72	Jean-Baptiste LÃĽonetti	95	
74	Anne K. Black	108	
75	Alfonso Gomez-Rejon	105	
88	Drake Doremus	101	
92	A. Todd Smith	0	
95	Walt Becker	92	
100	Chris Buck|Jennifer Lee	8	
101	Ben Wheatley	119	

103	Bharat Nalluri	104
116	Mike Elliott	105
119	Terry Jones	85
122	Joe Lynch	90
125	John Maclean	84
128	Noah Baumbach	84
130	Rupert Goold	100
132	Mark Burton|Richard Starzack	85
134	Fernando LeÃşn de Aranoa	106
139	Craig Zobel	97
140	Colin Teague	97
143	Gerado Olivares|Otmar Penker	98
146	Alejandro AmenÃąbar	106
147	Edward Zwick	114
148	Matt Brown	108
151	Uli Edel	94
152	Marjane Satrapi	101
153	Kazuaki Kiriya	115
158	Catherine Hardwicke	112
161	Luca Guadagnino	120
...	...	...
10830	Roman Polanski	113
10831	Billy Wilder	125
10833	William Wyler	123
10834	Burt Kennedy	95
10836	Charles Walters	114
10837	John Guillermin	156
10838	Richard Brooks	117
10839	Bill Melendez	25
10840	Guy Hamilton	102
10842	Wolfgang Reitherman	25
10843	Basil Dearden|Eliot Elisofon	134
10844	Daniel Mann	108
10845	Gerald Thomas	93
10846	Terence Fisher	90
10847	Terence Fisher	89
10849	Ronald Neame	109
10850	Jack Smight	121
10851	James Hill	95
10852	Fielder Cook	95
10853	Lewis Gilbert	114
10854	Arthur Penn	135
10856	Norman Tokar	93
10857	Henry Hathaway	128
10858	Norman Jewison	126
10859	John Frankenheimer	100
10860	Gerald Thomas	87
10861	Bruce Brown	95

```
10862             John Frankenheimer      176
10863                 Eldar Ryazanov       94
10864                    Woody Allen       80

                                        genres release_date  vote_count  \
30                               Mystery|Drama       6/19/15         425
36                         Crime|Drama|Mystery        9/3/15         474
72                                    Thriller       4/17/15          81
74                    Action|Adventure|Fantasy       6/24/15          27
75                                Comedy|Drama       6/12/15         569
88              Drama|Romance|Science Fiction         9/4/15         135
92                    Fantasy|Action|Adventure      12/19/15          11
95            Adventure|Animation|Comedy|Family     12/17/15         278
100                 Adventure|Animation|Family        3/9/15         475
101               Action|Drama|Science Fiction       9/26/15         161
103                            Thriller|Action       4/11/15         114
116                   Action|Fantasy|Adventure        1/9/15          22
119                     Comedy|Science Fiction       6/26/15         199
122                            Thriller|Action       1/23/15         169
125                   Romance|Thriller|Western       4/16/15         229
128                                     Comedy       8/14/15         132
130                        Crime|Drama|Mystery       4/17/15         354
132          Family|Animation|Comedy|Adventure        2/5/15         340
134                               Comedy|Drama       8/28/15         102
139             Drama|Science Fiction|Thriller       8/13/15         181
140                   Action|Adventure|Fantasy       2/24/15          59
143                     Adventure|Drama|Family      12/24/15          11
146                    Horror|Mystery|Thriller       10/1/15         310
147                                      Drama       9/16/15         148
148                                      Drama       9/17/15         104
151                            Horror|Thriller       9/16/15         114
152               Horror|Thriller|Comedy|Crime        2/6/15         371
153                           Action|Adventure        4/3/15         237
158                       Comedy|Drama|Romance       9/12/15         139
161               Crime|Drama|Mystery|Thriller      11/26/15          69
...                                        ...           ...         ...
10830           Comedy|Drama|Foreign|Thriller         2/1/66          18
10831                           Romance|Comedy      10/19/66          17
10833                     Comedy|Crime|Romance       7/13/66          67
10834                           Action|Western      10/19/66          14
10836                           Comedy|Romance        1/1/66          11
10837               War|Action|Adventure|Drama       6/21/66          12
10838                 Action|Adventure|Western       11/1/66          21
10839                         Family|Animation      10/27/66          49
10840                                 Thriller      12/22/66          13
10842                         Animation|Family        1/1/66          12
10843       Adventure|Drama|War|History|Action        6/9/66          12
10844  Adventure|Comedy|Fantasy|Science Fiction       1/16/66          13
```

10845	Comedy|Western	3/1/66	15
10846	Horror	1/9/66	16
10847	Science Fiction|Horror	6/20/66	13
10849	Action|Comedy|Crime	12/16/66	14
10850	Action|Drama|Thriller|Crime|Mystery	2/23/66	14
10851	Adventure|Drama|Action|Family|Foreign	6/22/66	15
10852	Western	5/31/66	11
10853	Comedy|Drama|Romance	3/29/66	26
10854	Thriller|Drama|Crime	2/17/66	17
10856	Comedy|Drama|Family	2/16/66	14
10857	Action|Western	6/10/66	10
10858	Comedy|War	5/25/66	11
10859	Mystery|Science Fiction|Thriller|Drama	10/5/66	22
10860	Comedy	5/20/66	13
10861	Documentary	6/15/66	11
10862	Action|Adventure|Drama	12/21/66	20
10863	Mystery|Comedy	1/1/66	11
10864	Action|Comedy	11/2/66	22

	vote_average	release_year	budget_adj	revenue_adj
30	6.4	2015	0.0	2.700677e+07
36	6.2	2015	0.0	2.056620e+07
72	5.5	2015	0.0	4.222338e+04
74	5.1	2015	0.0	0.000000e+00
75	7.7	2015	0.0	0.000000e+00
88	5.6	2015	0.0	1.839999e+06
92	5.4	2015	0.0	0.000000e+00
95	5.7	2015	0.0	2.150550e+08
100	7.0	2015	0.0	0.000000e+00
101	5.4	2015	0.0	0.000000e+00
103	5.6	2015	0.0	0.000000e+00
116	4.5	2015	0.0	0.000000e+00
119	5.8	2015	0.0	4.774472e+06
122	5.1	2015	0.0	0.000000e+00
125	6.6	2015	0.0	2.107664e+05
128	6.4	2015	0.0	2.300396e+06
130	6.0	2015	0.0	4.342117e+06
132	6.9	2015	0.0	5.492398e+07
134	6.3	2015	0.0	1.566238e+06
139	5.5	2015	0.0	1.090043e+05
140	4.5	2015	0.0	0.000000e+00
143	7.5	2015	0.0	0.000000e+00
146	5.2	2015	0.0	1.625741e+07
147	6.6	2015	0.0	0.000000e+00
148	7.1	2015	0.0	1.055465e+07
151	5.3	2015	0.0	0.000000e+00
152	6.0	2015	0.0	0.000000e+00
153	6.3	2015	0.0	3.352102e+06

```
158          7.2      2015     0.0  0.000000e+00
161          5.8      2015     0.0  1.781601e+06
...          ...       ...     ...           ...
10830        6.7      1966     0.0  0.000000e+00
10831        6.4      1966     0.0  0.000000e+00
10833        7.3      1966     0.0  0.000000e+00
10834        5.1      1966     0.0  0.000000e+00
10836        5.8      1966     0.0  0.000000e+00
10837        5.5      1966     0.0  0.000000e+00
10838        6.0      1966     0.0  0.000000e+00
10839        7.2      1966     0.0  0.000000e+00
10840        5.7      1966     0.0  0.000000e+00
10842        7.9      1966     0.0  0.000000e+00
10843        5.8      1966     0.0  0.000000e+00
10844        5.6      1966     0.0  0.000000e+00
10845        5.9      1966     0.0  0.000000e+00
10846        5.7      1966     0.0  0.000000e+00
10847        5.3      1966     0.0  0.000000e+00
10849        6.1      1966     0.0  0.000000e+00
10850        6.0      1966     0.0  0.000000e+00
10851        6.6      1966     0.0  0.000000e+00
10852        6.0      1966     0.0  0.000000e+00
10853        6.2      1966     0.0  0.000000e+00
10854        6.0      1966     0.0  0.000000e+00
10856        5.7      1966     0.0  0.000000e+00
10857        5.9      1966     0.0  0.000000e+00
10858        5.5      1966     0.0  0.000000e+00
10859        6.6      1966     0.0  0.000000e+00
10860        7.0      1966     0.0  0.000000e+00
10861        7.4      1966     0.0  0.000000e+00
10862        5.7      1966     0.0  0.000000e+00
10863        6.5      1966     0.0  0.000000e+00
10864        5.4      1966     0.0  0.000000e+00

[5636 rows x 12 columns]
```

In [19]:
```
# same for revenue_adj and we'll find that about half of these two columns have zero va
df.loc[df.revenue_adj ==0]
```

Out[19]:
```
           id  popularity                      original_title  \
48     265208    2.932340                           Wild Card
67     334074    2.331636                            Survivor
74     347096    2.165433              Mythica: The Darkspore
75     308369    2.141506      Me and Earl and the Dying Girl
92     370687    1.876037            Mythica: The Necromancer
93     307663    1.872696                                Vice
100    326359    1.724712                        Frozen Fever
101    254302    1.661789                           High-Rise
```

103	292040	1.646664	Spooks: The Greater Good
116	297291	1.380320	The Scorpion King: The Lost Throne
122	277355	1.342839	Everly
133	157827	1.251681	Louder Than Bombs
140	300803	1.144808	Dragonheart 3: The Sorcerer's Curse
143	378373	1.128081	Brothers of the Wind
145	294963	1.073349	Bone Tomahawk
147	245698	1.063055	Pawn Sacrifice
149	346808	1.041922	Momentum
151	290637	1.036825	Pay the Ghost
152	244458	1.027620	The Voices
154	314405	1.008474	Il racconto dei racconti
156	157843	0.973316	Queen of the Desert
158	290762	0.953647	Miss You Already
159	251516	0.953046	Kung Fury
164	228968	0.917040	Kidnapping Mr. Heineken
165	347969	0.913085	The Ridiculous 6
166	237756	0.906860	Kill Me Three Times
169	311291	0.894477	45 Years
174	342474	0.861179	Jenny's Wedding
175	277217	0.848748	Descendants
176	207936	0.843174	Tumbledown
...	...	...	...
10834	12639	0.310688	Return of the Seven
10836	38720	0.239435	Walk Don't Run
10837	19728	0.291704	The Blue Max
10838	22383	0.151845	The Professionals
10839	13353	0.276133	It's the Great Pumpkin, Charlie Brown
10840	34388	0.102530	Funeral in Berlin
10841	42701	0.264925	The Shooting
10842	36540	0.253437	Winnie the Pooh and the Honey Tree
10843	29710	0.252399	Khartoum
10844	23728	0.236098	Our Man Flint
10845	5065	0.230873	Carry On Cowboy
10846	17102	0.212716	Dracula: Prince of Darkness
10847	28763	0.034555	Island of Terror
10849	28270	0.206537	Gambit
10850	26268	0.202473	Harper
10851	15347	0.342791	Born Free
10852	37301	0.227220	A Big Hand for the Little Lady
10853	15598	0.163592	Alfie
10854	31602	0.146402	The Chase
10855	13343	0.141026	The Ghost & Mr. Chicken
10856	20277	0.140934	The Ugly Dachshund
10857	5921	0.131378	Nevada Smith
10858	31918	0.317824	The Russians Are Coming, The Russians Are Coming
10859	20620	0.089072	Seconds
10860	5060	0.087034	Carry On Screaming!

10861	21	0.080598	The Endless Summer
10862	20379	0.065543	Grand Prix
10863	39768	0.065141	Beregis Avtomobilya
10864	21449	0.064317	What's Up, Tiger Lily?
10865	22293	0.035919	Manos: The Hands of Fate

	director	runtime	\
48	Simon West	92	
67	James McTeigue	96	
74	Anne K. Black	108	
75	Alfonso Gomez-Rejon	105	
92	A. Todd Smith	0	
93	Brian A Miller	96	
100	Chris Buck|Jennifer Lee	8	
101	Ben Wheatley	119	
103	Bharat Nalluri	104	
116	Mike Elliott	105	
122	Joe Lynch	90	
133	Joachim Trier	109	
140	Colin Teague	97	
143	Gerado Olivares|Otmar Penker	98	
145	S. Craig Zahler	132	
147	Edward Zwick	114	
149	Stephen S. Campanelli	96	
151	Uli Edel	94	
152	Marjane Satrapi	101	
154	Matteo Garrone	125	
156	Werner Herzog	128	
158	Catherine Hardwicke	112	
159	David Sandberg	31	
164	Daniel Alfredson	95	
165	Frank Coraci	119	
166	Kriv Stenders	90	
169	Andrew Haigh	95	
174	Mary Agnes Donoghue	94	
175	Kenny Ortega	112	
176	Sean Mewshaw	105	
...	...	...	
10834	Burt Kennedy	95	
10836	Charles Walters	114	
10837	John Guillermin	156	
10838	Richard Brooks	117	
10839	Bill Melendez	25	
10840	Guy Hamilton	102	
10841	Monte Hellman	82	
10842	Wolfgang Reitherman	25	
10843	Basil Dearden|Eliot Elisofon	134	
10844	Daniel Mann	108	

10845	Gerald Thomas	93
10846	Terence Fisher	90
10847	Terence Fisher	89
10849	Ronald Neame	109
10850	Jack Smight	121
10851	James Hill	95
10852	Fielder Cook	95
10853	Lewis Gilbert	114
10854	Arthur Penn	135
10855	Alan Rafkin	90
10856	Norman Tokar	93
10857	Henry Hathaway	128
10858	Norman Jewison	126
10859	John Frankenheimer	100
10860	Gerald Thomas	87
10861	Bruce Brown	95
10862	John Frankenheimer	176
10863	Eldar Ryazanov	94
10864	Woody Allen	80
10865	Harold P. Warren	74

	genres	release_date	vote_count \
48	Thriller|Crime|Drama	1/14/15	481
67	Crime|Thriller|Action	5/21/15	280
74	Action|Adventure|Fantasy	6/24/15	27
75	Comedy|Drama	6/12/15	569
92	Fantasy|Action|Adventure	12/19/15	11
93	Thriller|Science Fiction|Action|Adventure	1/16/15	181
100	Adventure|Animation|Family	3/9/15	475
101	Action|Drama|Science Fiction	9/26/15	161
103	Thriller|Action	4/11/15	114
116	Action|Fantasy|Adventure	1/9/15	22
122	Thriller|Action	1/23/15	169
133	Drama	5/18/15	43
140	Action|Adventure|Fantasy	2/24/15	59
143	Adventure|Drama|Family	12/24/15	11
145	Horror|Western|Adventure|Drama	10/23/15	220
147	Drama	9/16/15	148
149	Thriller|Action	8/1/15	100
151	Horror|Thriller	9/16/15	114
152	Horror|Thriller|Comedy|Crime	2/6/15	371
154	Romance|Fantasy|Horror	5/14/15	211
156	Drama|History	9/3/15	30
158	Comedy|Drama|Romance	9/12/15	139
159	Action|Comedy|Science Fiction|Fantasy	5/28/15	487
164	Drama|Action|Crime|Thriller	3/12/15	131
165	Comedy|Western	12/11/15	252
166	Comedy|Thriller	4/10/15	96

169	Drama	8/28/15	167
174	Comedy|Drama	7/31/15	92
175	Music|Action|Adventure|Comedy|Family	7/31/15	262
176	Music|Romance|Comedy	4/18/15	22
...	...	...	...
10834	Action|Western	10/19/66	14
10836	Comedy|Romance	1/1/66	11
10837	War|Action|Adventure|Drama	6/21/66	12
10838	Action|Adventure|Western	11/1/66	21
10839	Family|Animation	10/27/66	49
10840	Thriller	12/22/66	13
10841	Western	10/23/66	12
10842	Animation|Family	1/1/66	12
10843	Adventure|Drama|War|History|Action	6/9/66	12
10844	Adventure|Comedy|Fantasy|Science Fiction	1/16/66	13
10845	Comedy|Western	3/1/66	15
10846	Horror	1/9/66	16
10847	Science Fiction|Horror	6/20/66	13
10849	Action|Comedy|Crime	12/16/66	14
10850	Action|Drama|Thriller|Crime|Mystery	2/23/66	14
10851	Adventure|Drama|Action|Family|Foreign	6/22/66	15
10852	Western	5/31/66	11
10853	Comedy|Drama|Romance	3/29/66	26
10854	Thriller|Drama|Crime	2/17/66	17
10855	Comedy|Family|Mystery|Romance	1/20/66	14
10856	Comedy|Drama|Family	2/16/66	14
10857	Action|Western	6/10/66	10
10858	Comedy|War	5/25/66	11
10859	Mystery|Science Fiction|Thriller|Drama	10/5/66	22
10860	Comedy	5/20/66	13
10861	Documentary	6/15/66	11
10862	Action|Adventure|Drama	12/21/66	20
10863	Mystery|Comedy	1/1/66	11
10864	Action|Comedy	11/2/66	22
10865	Horror	11/15/66	15

	vote_average	release_year	budget_adj	revenue_adj
48	5.3	2015	2.759999e+07	0.0
67	5.4	2015	1.839999e+07	0.0
74	5.1	2015	0.000000e+00	0.0
75	7.7	2015	0.000000e+00	0.0
92	5.4	2015	0.000000e+00	0.0
93	4.1	2015	9.199996e+06	0.0
100	7.0	2015	0.000000e+00	0.0
101	5.4	2015	0.000000e+00	0.0
103	5.6	2015	0.000000e+00	0.0
116	4.5	2015	0.000000e+00	0.0
122	5.1	2015	0.000000e+00	0.0

133	6.3	2015	1.012000e+07	0.0
140	4.5	2015	0.000000e+00	0.0
143	7.5	2015	0.000000e+00	0.0
145	6.3	2015	1.655999e+06	0.0
147	6.6	2015	0.000000e+00	0.0
149	5.8	2015	1.839999e+07	0.0
151	5.3	2015	0.000000e+00	0.0
152	6.0	2015	0.000000e+00	0.0
154	5.7	2015	1.104000e+07	0.0
156	6.0	2015	1.379999e+07	0.0
158	7.2	2015	0.000000e+00	0.0
159	7.7	2015	5.796172e+05	0.0
164	5.8	2015	0.000000e+00	0.0
165	4.8	2015	5.519998e+07	0.0
166	5.1	2015	0.000000e+00	0.0
169	6.0	2015	0.000000e+00	0.0
174	5.2	2015	0.000000e+00	0.0
175	6.7	2015	0.000000e+00	0.0
176	6.6	2015	0.000000e+00	0.0
...	...	...	...	...
10834	5.1	1966	0.000000e+00	0.0
10836	5.8	1966	0.000000e+00	0.0
10837	5.5	1966	0.000000e+00	0.0
10838	6.0	1966	0.000000e+00	0.0
10839	7.2	1966	0.000000e+00	0.0
10840	5.7	1966	0.000000e+00	0.0
10841	5.5	1966	5.038511e+05	0.0
10842	7.9	1966	0.000000e+00	0.0
10843	5.8	1966	0.000000e+00	0.0
10844	5.6	1966	0.000000e+00	0.0
10845	5.9	1966	0.000000e+00	0.0
10846	5.7	1966	0.000000e+00	0.0
10847	5.3	1966	0.000000e+00	0.0
10849	6.1	1966	0.000000e+00	0.0
10850	6.0	1966	0.000000e+00	0.0
10851	6.6	1966	0.000000e+00	0.0
10852	6.0	1966	0.000000e+00	0.0
10853	6.2	1966	0.000000e+00	0.0
10854	6.0	1966	0.000000e+00	0.0
10855	6.1	1966	4.702610e+06	0.0
10856	5.7	1966	0.000000e+00	0.0
10857	5.9	1966	0.000000e+00	0.0
10858	5.5	1966	0.000000e+00	0.0
10859	6.6	1966	0.000000e+00	0.0
10860	7.0	1966	0.000000e+00	0.0
10861	7.4	1966	0.000000e+00	0.0
10862	5.7	1966	0.000000e+00	0.0
10863	6.5	1966	0.000000e+00	0.0

```
10864            5.4          1966  0.000000e+00          0.0
10865            1.5          1966  1.276423e+05          0.0

[5952 rows x 12 columns]
```

In [20]:
```
# we can't delete all these big rows with zero values
# we can't also replace the zero values by mean method as the big numbers of zero data
# the best method in our case is to use the non_zeros mean (mean of columns for values

nonzero_budget_adj_mean = df[df.budget_adj !=0].mean()

nonzero_budget_adj_mean
```

Out[20]:
```
id              4.545144e+04
popularity      9.931836e-01
runtime         1.071017e+02
vote_count      4.090292e+02
vote_average    6.032552e+00
release_year    2.001250e+03
budget_adj      3.692239e+07
revenue_adj     1.022921e+08
dtype: float64
```

In [21]:
```
nonzero_revenue_adj_mean = df[df.revenue_adj !=0].mean()

nonzero_revenue_adj_mean
```

Out[21]:
```
id              4.458150e+04
popularity      1.045387e+00
runtime         1.079587e+02
vote_count      4.363709e+02
vote_average    6.149072e+00
release_year    2.000918e+03
budget_adj      3.516846e+07
revenue_adj     1.151223e+08
dtype: float64
```

In [22]:
```
# we find now the non_zero_ mean for both budget_adj and revenue_adj
# nonzero_budget_adj_mean = 3.692239e+07
# nonzero_revenue_adj_mean = 1.151223e+08
# now we may fill the the data with zero values with above mentioned calculated mean
df.budget_adj.replace((0, 3.692239e+07), inplace=True)
df.revenue_adj.replace((0, 1.151223e+08), inplace=True)
```

In [23]:
```
# now lets check again if we have zeros values in these two columns
df.loc[df.budget_adj ==0]
df.loc[df.revenue_adj ==0]
# now we'll find no zero values in our data frame
```

```
Out[23]: Empty DataFrame
         Columns: [id, popularity, original_title, director, runtime, genres, release_date, vote
         Index: []
```

```
In [ ]:
```

Exploratory Data Analysis

starting posing some questions and try to get the proper answers from available dataset and by drawing the necessary graphs

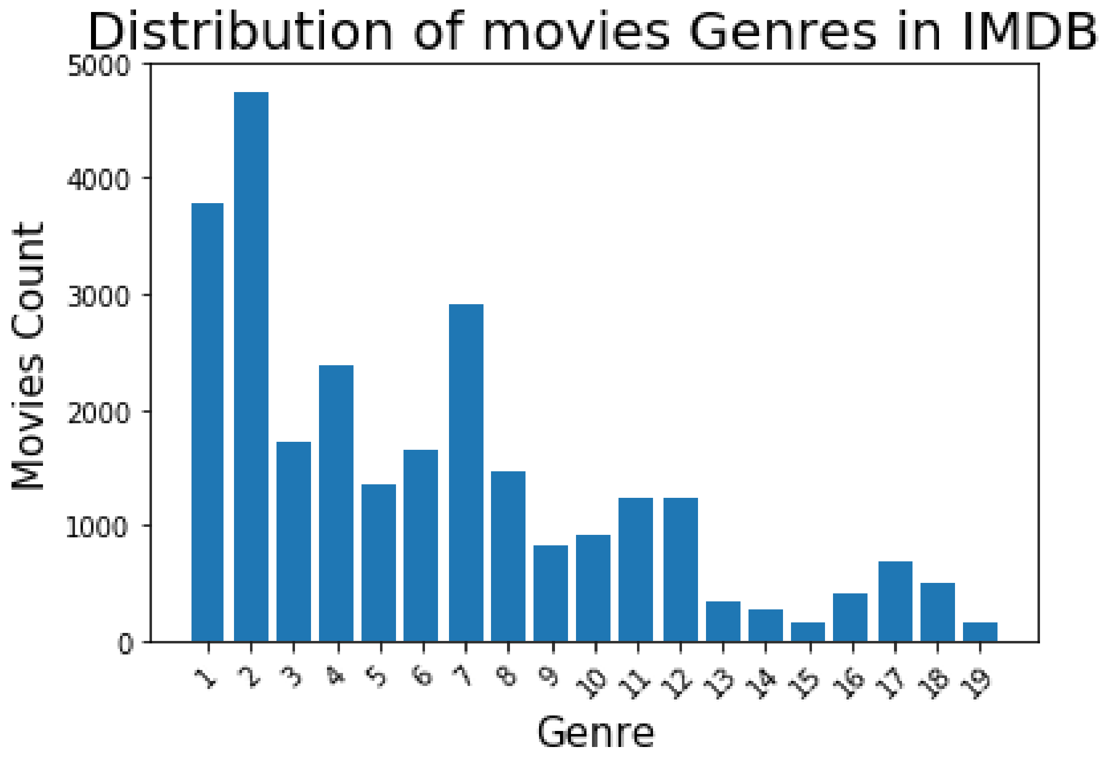
1.3.2 Research Question 1 (Distribution of movies genres in IMDB)

```
In [24]: # as the genres column have data contains more than one string, so we have to separate
         # using (str.contains) allows seperating all these strings
         # then this will let me count of iteration of each or values_count of each genres
         comedy_films =df[df['genres'].str.contains('Comedy')]
         drama_films =df[df['genres'].str.contains('Drama')]
         romance_films =df[df['genres'].str.contains('Romance')]
         action_films =df[df['genres'].str.contains('Action')]
         crime_films =df[df['genres'].str.contains('Crime')]
         horror_films =df[df['genres'].str.contains('Horror')]
         thriller_films =df[df['genres'].str.contains('Thriller')]
         adventure_films =df[df['genres'].str.contains('Adventure')]
         mystery_films =df[df['genres'].str.contains('Mystery')]
         fantasy_films =df[df['genres'].str.contains('Fantasy')]
         family_films =df[df['genres'].str.contains('Family')]
         sci_fi_films =df[df['genres'].str.contains('Science Fiction')]
         history_films =df[df['genres'].str.contains('History')]
         war_films =df[df['genres'].str.contains('War')]
         western_films =df[df['genres'].str.contains('Western')]
         music_films =df[df['genres'].str.contains('Music')]
         animation_films =df[df['genres'].str.contains('Animation')]
         documentary_films =df[df['genres'].str.contains('Documentary')]
         tv_films =df[df['genres'].str.contains('TV')]

         comedy_films.shape[0], drama_films.shape[0], romance_films.shape[0], action_films.shape
```

```
Out[24]: (3782,
          4754,
          1708,
          2378,
          1353,
          1636,
          2904,
          1466,
          809,
          912,
          1223,
```

```
 1223,
 332,
 270,
 164,
 402,
 692,
 509,
 162)
```

```
In [25]: location = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]
         height = [comedy_films.shape[0], drama_films.shape[0], romance_films.shape[0], action_f


         label = ['comedy', 'drama', 'romance', 'action', 'crime, horror', 'thriller', 'adventur


         index = np.arange(len(location))
         plt.bar(index, height, tick_label='label')
         plt.xlabel('Genre', fontsize=15)
         plt.ylabel('Movies Count', fontsize=15)
         plt.xticks(index, location, fontsize=10, rotation=45)
         plt.title('Distribution of movies Genres in IMDB', fontsize=20);
```



1.4 from previous graph will find that Drama films are most numbers of films available in IMDB, then comedy films, thriller and finally the action films, in other hand the Tv-films and western films are the less numbers of films available in IMDB.

1.4.1 Research Question 2 (relationship between Revenues_adj, Budget_adj anf Popularity, are movies with high revenues or high budget have a higher popularity and viceversa)

In [26]:
```python
# computing the mean of Revenue_adj
df.revenue_adj.mean()
```

Out[26]: 73262288.600239351

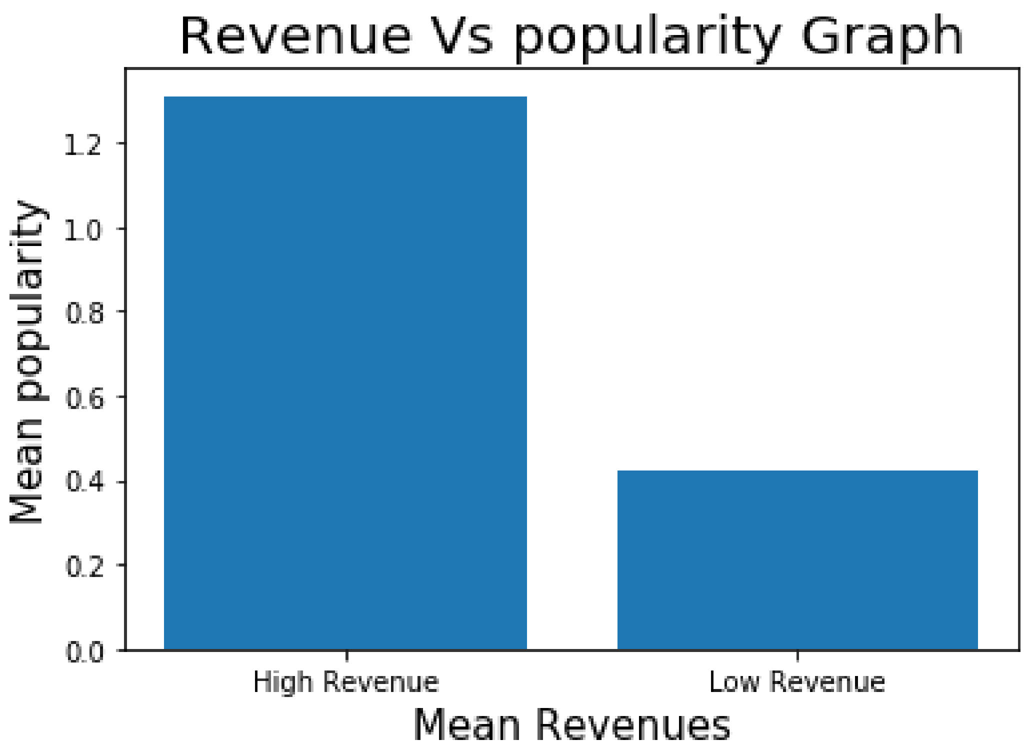
In [27]:
```python
high_revenue= df.query('revenue_adj>73262288.6')
low_revenue= df.query('revenue_adj<=73262288.6')
```

In [28]:
```python
high_popularity = high_revenue['popularity'].mean()
low_popularity = low_revenue['popularity'].mean()

high_popularity, low_popularity
```

Out[28]: (1.3098792565765764, 0.42108383052959497)

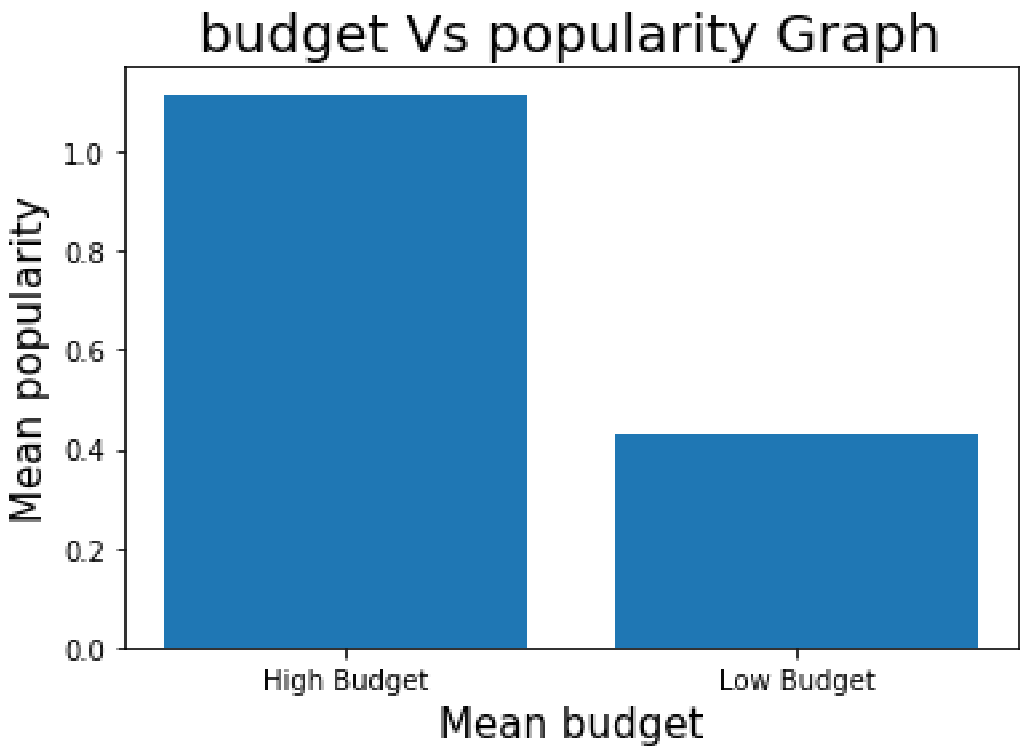
In [29]:
```python
locations = [1, 2]
heights =[high_popularity, low_popularity]
labels = ['High Revenue', 'Low Revenue']

plt.bar(locations, heights, tick_label=labels)
plt.title('Revenue Vs popularity Graph', fontsize=20)
plt.xlabel('Mean Revenues', fontsize=15)
plt.ylabel('Mean popularity', fontsize=15)
```

Out[29]: Text(0,0.5,'Mean popularity')

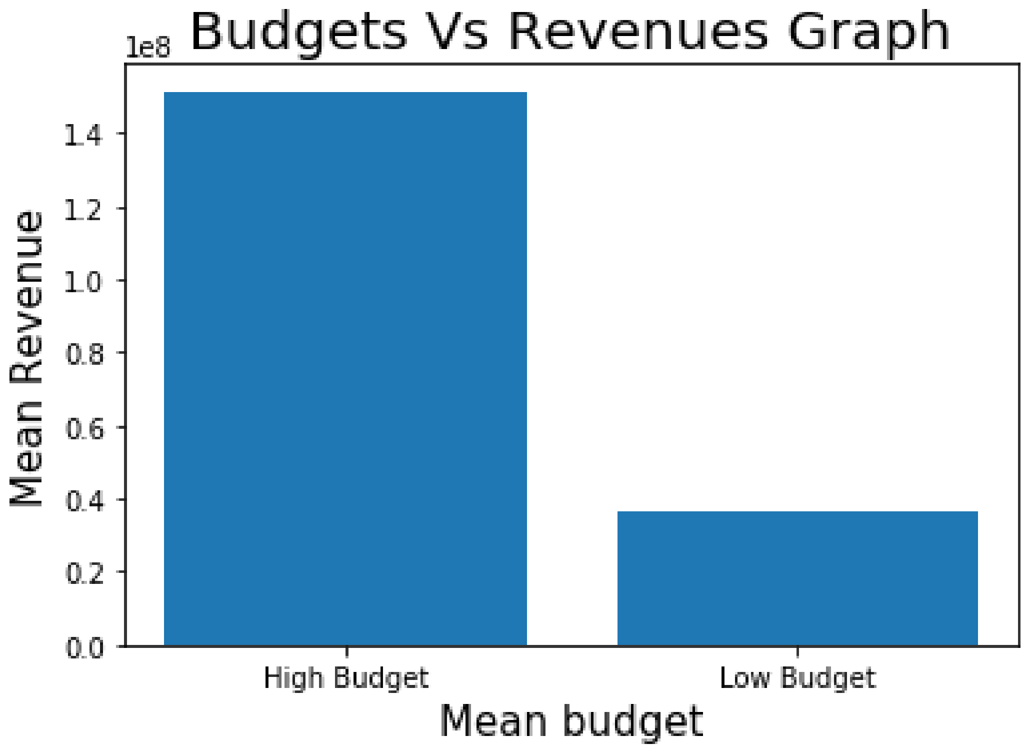



```
In [30]: # same here for Budget: computing the mean of Budget_adj

         df.budget_adj.mean()

Out[30]: 27016937.528131425

In [31]: high_budget= df.query('budget_adj>27016937.5')
         low_budget= df.query('budget_adj<=27016937.5')

In [32]: high_popularity = high_budget['popularity'].mean()
         low_popularity = low_budget['popularity'].mean()

         high_popularity, low_popularity

Out[32]: (1.1160031563225057, 0.43064931909684434)

In [33]: locations = [1, 2]
         heights =[high_popularity, low_popularity]
         labels = ['High Budget', 'Low Budget']

         plt.bar(locations, heights, tick_label=labels)
         plt.title('budget Vs popularity Graph', fontsize=20)
         plt.xlabel('Mean budget', fontsize=15)
         plt.ylabel('Mean popularity', fontsize=15)
```

```
Out[33]: Text(0,0.5,'Mean popularity')
```




```
In [34]: # Now we have all information to know if movies with higher budget will get a higher re
         high_revenue_adj = high_budget['revenue_adj'].mean()
         low_revenue_adj = low_budget['revenue_adj'].mean()

         high_revenue_adj, low_revenue_adj

Out[34]: (151517560.65381107, 36561502.686105058)

In [35]: locations = [1, 2]
         heights =[high_revenue_adj, low_revenue_adj]
         labels = ['High Budget', 'Low Budget']

         plt.bar(locations, heights, tick_label=labels)
         plt.title('Budgets Vs Revenues Graph', fontsize=20)
         plt.xlabel('Mean budget', fontsize=15)
         plt.ylabel('Mean Revenue', fontsize=15)

Out[35]: Text(0,0.5,'Mean Revenue')
```



1.4.2 Research Question 3 (which genres of movies are most popular over time, then from year to year)

```
In [36]: df.groupby('genres')['popularity'].sum()
```

```
Out[36]: genres
         Action                                                37.269991
         Action|Adventure                                      10.360652
         Action|Adventure|Animation                             1.818651
         Action|Adventure|Animation|Comedy|Drama                0.370019
         Action|Adventure|Animation|Comedy|Family               0.063246
         Action|Adventure|Animation|Drama|Family                0.132458
         Action|Adventure|Animation|Family                      1.616152
         Action|Adventure|Animation|Family|Fantasy              1.603381
         Action|Adventure|Animation|Family|Mystery              0.201030
         Action|Adventure|Animation|Family|Science Fiction      4.262132
         Action|Adventure|Animation|Fantasy                     0.070257
         Action|Adventure|Animation|Fantasy|Horror              0.155075
         Action|Adventure|Animation|Fantasy|Science Fiction     0.401188
         Action|Adventure|Animation|Science Fiction             1.710748
         Action|Adventure|Animation|Science Fiction|Crime       0.559451
         Action|Adventure|Animation|Science Fiction|Thriller    2.846465
         Action|Adventure|Comedy                               20.086228
```

```
Action|Adventure|Comedy|Crime                          2.116193
Action|Adventure|Comedy|Crime|Drama                    2.221544
Action|Adventure|Comedy|Crime|Foreign                  0.021222
Action|Adventure|Comedy|Crime|Romance                  0.146110
Action|Adventure|Comedy|Crime|Thriller                 5.241053
Action|Adventure|Comedy|Drama                          2.070309
Action|Adventure|Comedy|Drama|Family                   0.675253
Action|Adventure|Comedy|Drama|Mystery                  0.571693
Action|Adventure|Comedy|Drama|Romance                  0.133281
Action|Adventure|Comedy|Drama|Science Fiction          1.853547
Action|Adventure|Comedy|Drama|Thriller                 0.222379
Action|Adventure|Comedy|Drama|War                      0.492877
Action|Adventure|Comedy|Drama|Western                  2.267704
                                                         ...
War|Drama|Action                                       6.415818
War|Drama|Action|Adventure|History                     0.757082
War|Drama|Foreign|History                              0.267577
War|Drama|History                                      2.537700
War|Drama|History|Action                               1.779861
War|Drama|History|Action|Romance                       0.294611
War|Drama|History|Thriller                             0.523770
War|Drama|Mystery|Romance                              0.756105
War|Drama|Romance                                      0.522053
War|History                                            0.137661
War|History|Action|Adventure|Drama                     1.319068
Western                                                8.227824
Western|Action                                         0.301410
Western|Action|Adventure                               0.386204
Western|Action|Adventure|Drama                         0.526108
Western|Action|Comedy                                  0.363695
Western|Action|Drama|Science Fiction                   0.510296
Western|Adventure                                      1.272227
Western|Animation|Adventure|Comedy|Family              1.040588
Western|Animation|Family|Comedy|Music                  0.837906
Western|Comedy                                         0.262123
Western|Comedy|Drama|Music                             0.360746
Western|Drama                                          3.301304
Western|Drama|Adventure|Thriller                       9.110700
Western|Drama|Comedy|Romance                           0.293473
Western|Drama|Crime|Romance                            0.393664
Western|History                                        0.128234
Western|History|War                                    0.948560
Western|Horror|Thriller                                0.354484
Western|Thriller                                       0.387592
Name: popularity, Length: 2031, dtype: float64
```

In [37]:
```
# which genres are most popular.
# to answer this question we have first to consider that the data in genres column as i
```

```
# this is clearly shown in the previous code, we'' find the column 'genres' have more t
# the best method and the simple one is to consider the first string in each row as the
# to split the first string
genres_new = df['genres'].str.split("|", n = 0, expand = True)
genres_new
```

Out[37]:

```
                     0               1               2               3  \
0               Action       Adventure  Science Fiction        Thriller
1               Action       Adventure  Science Fiction        Thriller
2            Adventure  Science Fiction        Thriller            None
3               Action       Adventure  Science Fiction         Fantasy
4               Action           Crime        Thriller            None
5              Western           Drama       Adventure        Thriller
6      Science Fiction          Action        Thriller       Adventure
7                Drama       Adventure  Science Fiction            None
8               Family       Animation       Adventure          Comedy
9               Comedy       Animation          Family            None
10              Action       Adventure           Crime            None
11     Science Fiction         Fantasy          Action       Adventure
12               Drama  Science Fiction            None            None
13              Action          Comedy  Science Fiction            None
14              Action       Adventure  Science Fiction            None
15               Crime           Drama         Mystery         Western
16               Crime          Action        Thriller            None
17     Science Fiction          Action       Adventure            None
18             Romance         Fantasy          Family           Drama
19                 War       Adventure  Science Fiction            None
20              Action          Family  Science Fiction       Adventure
21              Action           Drama            None            None
22              Action           Drama        Thriller            None
23               Drama         Romance            None            None
24              Comedy           Drama            None            None
25              Action            None            None            None
26              Comedy            None            None            None
27               Crime          Comedy          Action       Adventure
28               Drama        Thriller         History            None
29              Action  Science Fiction        Thriller            None
...                ...             ...             ...             ...
10836           Comedy         Romance            None            None
10837              War          Action       Adventure           Drama
10838           Action       Adventure         Western            None
10839           Family       Animation            None            None
10840         Thriller            None            None            None
10841          Western            None            None            None
10842        Animation          Family            None            None
10843        Adventure           Drama             War         History
10844        Adventure          Comedy         Fantasy  Science Fiction
10845           Comedy         Western            None            None
```

10846	Horror	None	None	None
10847	Science Fiction	Horror	None	None
10848	Adventure	Science Fiction	None	None
10849	Action	Comedy	Crime	None
10850	Action	Drama	Thriller	Crime
10851	Adventure	Drama	Action	Family
10852	Western	None	None	None
10853	Comedy	Drama	Romance	None
10854	Thriller	Drama	Crime	None
10855	Comedy	Family	Mystery	Romance
10856	Comedy	Drama	Family	None
10857	Action	Western	None	None
10858	Comedy	War	None	None
10859	Mystery	Science Fiction	Thriller	Drama
10860	Comedy	None	None	None
10861	Documentary	None	None	None
10862	Action	Adventure	Drama	None
10863	Mystery	Comedy	None	None
10864	Action	Comedy	None	None
10865	Horror	None	None	None

	4
0	None
1	None
2	None
3	None
4	None
5	None
6	None
7	None
8	None
9	None
10	None
11	None
12	None
13	None
14	None
15	None
16	None
17	None
18	None
19	None
20	Mystery
21	None
22	None
23	None
24	None
25	None

```
26         None
27         None
28         None
29         None
...         ...
10836      None
10837      None
10838      None
10839      None
10840      None
10841      None
10842      None
10843    Action
10844      None
10845      None
10846      None
10847      None
10848      None
10849      None
10850   Mystery
10851   Foreign
10852      None
10853      None
10854      None
10855      None
10856      None
10857      None
10858      None
10859      None
10860      None
10861      None
10862      None
10863      None
10864      None
10865      None

[10800 rows x 5 columns]
```

In [38]:
```
# now we'll consider only the first column from multiple columns created from splitting
genres_adj = genres_new[0]
genres_adj
```

Out[38]:
```
0                 Action
1                 Action
2              Adventure
3                 Action
4                 Action
5                Western
```

```
6        Science Fiction
7                  Drama
8                 Family
9                 Comedy
10                Action
11       Science Fiction
12                 Drama
13                Action
14                Action
15                 Crime
16                 Crime
17       Science Fiction
18               Romance
19                   War
20                Action
21                Action
22                Action
23                 Drama
24                Comedy
25                Action
26                Comedy
27                 Crime
28                 Drama
29                Action
              ...
10836             Comedy
10837                War
10838             Action
10839             Family
10840           Thriller
10841            Western
10842          Animation
10843          Adventure
10844          Adventure
10845             Comedy
10846             Horror
10847    Science Fiction
10848          Adventure
10849             Action
10850             Action
10851          Adventure
10852            Western
10853             Comedy
10854           Thriller
10855             Comedy
10856             Comedy
10857             Action
10858             Comedy
```

```
10859           Mystery
10860            Comedy
10861       Documentary
10862            Action
10863           Mystery
10864            Action
10865            Horror
Name: 0, Length: 10800, dtype: object
```

In [39]:
```
# adding the new column (genres_adj) to the dataset, this column has only on string eac
df['genres_adj'] = genres_adj
```

In [40]:
```
# deleting the old column (genres) with multiple strings in each row.
df.drop(['genres'], axis=1)
```

Out[40]:
```
           id  popularity                             original_title  \
0      135397   32.985763                             Jurassic World
1       76341   28.419936                         Mad Max: Fury Road
2      262500   13.112507                                  Insurgent
3      140607   11.173104               Star Wars: The Force Awakens
4      168259    9.335014                                  Furious 7
5      281957    9.110700                               The Revenant
6       87101    8.654359                         Terminator Genisys
7      286217    7.667400                                The Martian
8      211672    7.404165                                    Minions
9      150540    6.326804                                 Inside Out
10     206647    6.200282                                    Spectre
11      76757    6.189369                           Jupiter Ascending
12     264660    6.118847                                 Ex Machina
13     257344    5.984995                                     Pixels
14      99861    5.944927                    Avengers: Age of Ultron
15     273248    5.898400                          The Hateful Eight
16     260346    5.749758                                    Taken 3
17     102899    5.573184                                    Ant-Man
18     150689    5.556818                                 Cinderella
19     131634    5.476958   The Hunger Games: Mockingjay - Part 2
20     158852    5.462138                               Tomorrowland
21     307081    5.337064                                   Southpaw
22     254128    4.907832                                San Andreas
23     216015    4.710402                         Fifty Shades of Grey
24     318846    4.648046                              The Big Short
25     177677    4.566713        Mission: Impossible - Rogue Nation
26     214756    4.564549                                      Ted 2
27     207703    4.503789               Kingsman: The Secret Service
28     314365    4.062293                                  Spotlight
29     294254    3.968891             Maze Runner: The Scorch Trials
...       ...         ...                                        ...
10836   38720    0.239435                              Walk Don't Run
```

10837	19728	0.291704	The Blue Max
10838	22383	0.151845	The Professionals
10839	13353	0.276133	It's the Great Pumpkin, Charlie Brown
10840	34388	0.102530	Funeral in Berlin
10841	42701	0.264925	The Shooting
10842	36540	0.253437	Winnie the Pooh and the Honey Tree
10843	29710	0.252399	Khartoum
10844	23728	0.236098	Our Man Flint
10845	5065	0.230873	Carry On Cowboy
10846	17102	0.212716	Dracula: Prince of Darkness
10847	28763	0.034555	Island of Terror
10848	2161	0.207257	Fantastic Voyage
10849	28270	0.206537	Gambit
10850	26268	0.202473	Harper
10851	15347	0.342791	Born Free
10852	37301	0.227220	A Big Hand for the Little Lady
10853	15598	0.163592	Alfie
10854	31602	0.146402	The Chase
10855	13343	0.141026	The Ghost & Mr. Chicken
10856	20277	0.140934	The Ugly Dachshund
10857	5921	0.131378	Nevada Smith
10858	31918	0.317824	The Russians Are Coming, The Russians Are Coming
10859	20620	0.089072	Seconds
10860	5060	0.087034	Carry On Screaming!
10861	21	0.080598	The Endless Summer
10862	20379	0.065543	Grand Prix
10863	39768	0.065141	Beregis Avtomobilya
10864	21449	0.064317	What's Up, Tiger Lily?
10865	22293	0.035919	Manos: The Hands of Fate

	director	runtime	release_date	vote_count	\
0	Colin Trevorrow	124	6/9/15	5562	
1	George Miller	120	5/13/15	6185	
2	Robert Schwentke	119	3/18/15	2480	
3	J.J. Abrams	136	12/15/15	5292	
4	James Wan	137	4/1/15	2947	
5	Alejandro GonzÃąlez IÃśÃąrritu	156	12/25/15	3929	
6	Alan Taylor	125	6/23/15	2598	
7	Ridley Scott	141	9/30/15	4572	
8	Kyle Balda|Pierre Coffin	91	6/17/15	2893	
9	Pete Docter	94	6/9/15	3935	
10	Sam Mendes	148	10/26/15	3254	
11	Lana Wachowski|Lilly Wachowski	124	2/4/15	1937	
12	Alex Garland	108	1/21/15	2854	
13	Chris Columbus	105	7/16/15	1575	
14	Joss Whedon	141	4/22/15	4304	
15	Quentin Tarantino	167	12/25/15	2389	
16	Olivier Megaton	109	1/1/15	1578	

17	Peyton Reed	115	7/14/15	3779
18	Kenneth Branagh	112	3/12/15	1495
19	Francis Lawrence	136	11/18/15	2380
20	Brad Bird	130	5/19/15	1899
21	Antoine Fuqua	123	6/15/15	1386
22	Brad Peyton	114	5/27/15	2060
23	Sam Taylor-Johnson	125	2/11/15	1865
24	Adam McKay	130	12/11/15	1545
25	Christopher McQuarrie	131	7/23/15	2349
26	Seth MacFarlane	115	6/25/15	1666
27	Matthew Vaughn	130	1/24/15	3833
28	Tom McCarthy	128	11/6/15	1559
29	Wes Ball	132	9/9/15	1849
...	...	...	...	...
10836	Charles Walters	114	1/1/66	11
10837	John Guillermin	156	6/21/66	12
10838	Richard Brooks	117	11/1/66	21
10839	Bill Melendez	25	10/27/66	49
10840	Guy Hamilton	102	12/22/66	13
10841	Monte Hellman	82	10/23/66	12
10842	Wolfgang Reitherman	25	1/1/66	12
10843	Basil Dearden|Eliot Elisofon	134	6/9/66	12
10844	Daniel Mann	108	1/16/66	13
10845	Gerald Thomas	93	3/1/66	15
10846	Terence Fisher	90	1/9/66	16
10847	Terence Fisher	89	6/20/66	13
10848	Richard Fleischer	100	8/24/66	42
10849	Ronald Neame	109	12/16/66	14
10850	Jack Smight	121	2/23/66	14
10851	James Hill	95	6/22/66	15
10852	Fielder Cook	95	5/31/66	11
10853	Lewis Gilbert	114	3/29/66	26
10854	Arthur Penn	135	2/17/66	17
10855	Alan Rafkin	90	1/20/66	14
10856	Norman Tokar	93	2/16/66	14
10857	Henry Hathaway	128	6/10/66	10
10858	Norman Jewison	126	5/25/66	11
10859	John Frankenheimer	100	10/5/66	22
10860	Gerald Thomas	87	5/20/66	13
10861	Bruce Brown	95	6/15/66	11
10862	John Frankenheimer	176	12/21/66	20
10863	Eldar Ryazanov	94	1/1/66	11
10864	Woody Allen	80	11/2/66	22
10865	Harold P. Warren	74	11/15/66	15

	vote_average	release_year	budget_adj	revenue_adj	genres_adj
0	6.5	2015	1.379999e+08	1.392446e+09	Action
1	7.1	2015	1.379999e+08	3.481613e+08	Action

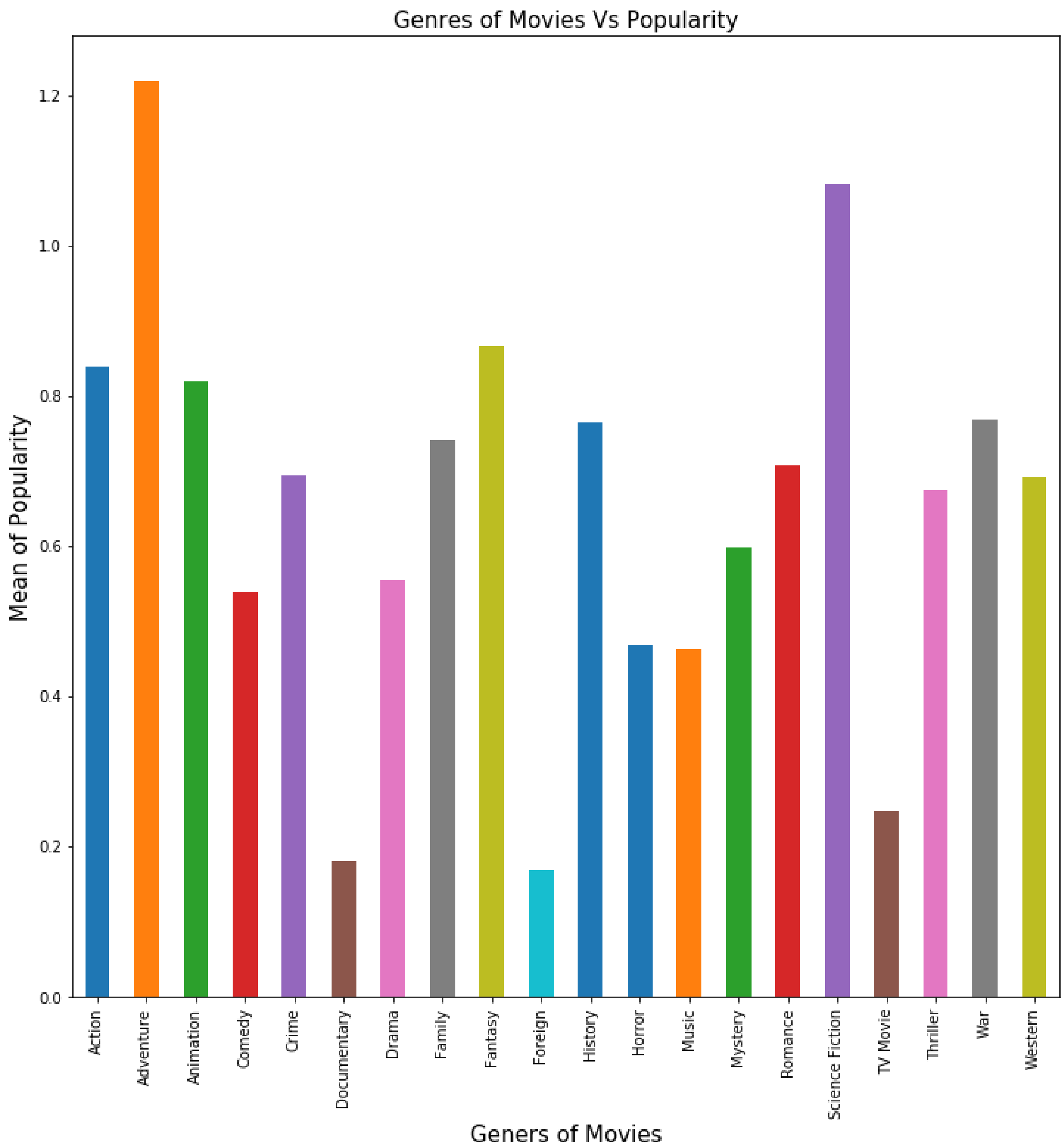
2	6.3	2015	1.012000e+08	2.716190e+08	Adventure
3	7.5	2015	1.839999e+08	1.902723e+09	Action
4	7.3	2015	1.747999e+08	1.385749e+09	Action
5	7.2	2015	1.241999e+08	4.903142e+08	Western
6	5.8	2015	1.425999e+08	4.053551e+08	Science Fiction
7	7.6	2015	9.935996e+07	5.477497e+08	Drama
8	6.5	2015	6.807997e+07	1.064192e+09	Family
9	8.0	2015	1.609999e+08	7.854116e+08	Comedy
10	6.2	2015	2.253999e+08	8.102203e+08	Action
11	5.2	2015	1.619199e+08	1.692686e+08	Science Fiction
12	7.6	2015	1.379999e+07	3.391985e+07	Drama
13	5.8	2015	8.095996e+07	2.241460e+08	Action
14	7.4	2015	2.575999e+08	1.292632e+09	Action
15	7.4	2015	4.047998e+07	1.432992e+08	Crime
16	6.1	2015	4.415998e+07	2.997096e+08	Crime
17	7.0	2015	1.195999e+08	4.771138e+08	Science Fiction
18	6.8	2015	8.739996e+07	4.989630e+08	Romance
19	6.5	2015	1.471999e+08	5.984813e+08	War
20	6.2	2015	1.747999e+08	1.923127e+08	Action
21	7.3	2015	2.759999e+07	8.437300e+07	Action
22	6.1	2015	1.012000e+08	4.328514e+08	Action
23	5.3	2015	3.679998e+07	5.240791e+08	Drama
24	7.3	2015	2.575999e+07	1.226787e+08	Comedy
25	7.1	2015	1.379999e+08	6.277435e+08	Action
26	6.3	2015	6.255997e+07	1.985944e+08	Comedy
27	7.6	2015	7.451997e+07	3.714978e+08	Crime
28	7.8	2015	1.839999e+07	8.127872e+07	Drama
29	6.4	2015	5.611998e+07	2.863562e+08	Action
...	...	...	...	...	...
10836	5.8	1966	8.061618e+07	1.343603e+08	Comedy
10837	5.5	1966	8.061618e+07	1.343603e+08	War
10838	6.0	1966	8.061618e+07	1.343603e+08	Action
10839	7.2	1966	8.061618e+07	1.343603e+08	Family
10840	5.7	1966	8.061618e+07	1.343603e+08	Thriller
10841	5.5	1966	5.038511e+05	1.343603e+08	Western
10842	7.9	1966	5.038511e+05	1.343603e+08	Animation
10843	5.8	1966	5.038511e+05	1.343603e+08	Adventure
10844	5.6	1966	5.038511e+05	1.343603e+08	Adventure
10845	5.9	1966	5.038511e+05	1.343603e+08	Comedy
10846	5.7	1966	5.038511e+05	1.343603e+08	Horror
10847	5.3	1966	5.038511e+05	1.343603e+08	Science Fiction
10848	6.7	1966	3.436265e+07	8.061618e+07	Adventure
10849	6.1	1966	3.436265e+07	8.061618e+07	Action
10850	6.0	1966	3.436265e+07	8.061618e+07	Action
10851	6.6	1966	3.436265e+07	8.061618e+07	Adventure
10852	6.0	1966	3.436265e+07	8.061618e+07	Western
10853	6.2	1966	3.436265e+07	8.061618e+07	Comedy
10854	6.0	1966	3.436265e+07	8.061618e+07	Thriller

```
10855           6.1          1966  4.702610e+06  8.061618e+07          Comedy
10856           5.7          1966  4.702610e+06  8.061618e+07          Comedy
10857           5.9          1966  4.702610e+06  8.061618e+07          Action
10858           5.5          1966  4.702610e+06  8.061618e+07          Comedy
10859           6.6          1966  4.702610e+06  8.061618e+07         Mystery
10860           7.0          1966  4.702610e+06  8.061618e+07          Comedy
10861           7.4          1966  4.702610e+06  8.061618e+07     Documentary
10862           5.7          1966  4.702610e+06  8.061618e+07          Action
10863           6.5          1966  4.702610e+06  8.061618e+07         Mystery
10864           5.4          1966  4.702610e+06  8.061618e+07          Action
10865           1.5          1966  1.276423e+05  8.061618e+07          Horror

[10800 rows x 12 columns]
```

In [52]:
```python
# Now we can use groupby to see which genres are most popular over entire time
df.groupby('genres_adj')['popularity'].mean()
```

Out[52]:
```
genres_adj
Action             0.837782
Adventure          1.217868
Animation          0.817977
Comedy             0.538260
Crime              0.694063
Documentary        0.179317
Drama              0.553444
Family             0.739779
Fantasy            0.865390
Foreign            0.167124
History            0.764636
Horror             0.468638
Music              0.462125
Mystery            0.596896
Romance            0.707231
Science Fiction    1.082355
TV Movie           0.245873
Thriller           0.673381
War                0.767041
Western            0.690646
Name: popularity, dtype: float64
```

In [118]:
```python
# ploting the last result in graph
df1 = df.groupby('genres_adj')['popularity'].mean().plot(kind='bar', figsize=(12,12));
plt.title('Genres of Movies Vs Popularity', fontsize=15)
plt.xlabel('Geners of Movies', fontsize=15)
plt.ylabel('Mean of Popularity', fontsize=15);
```



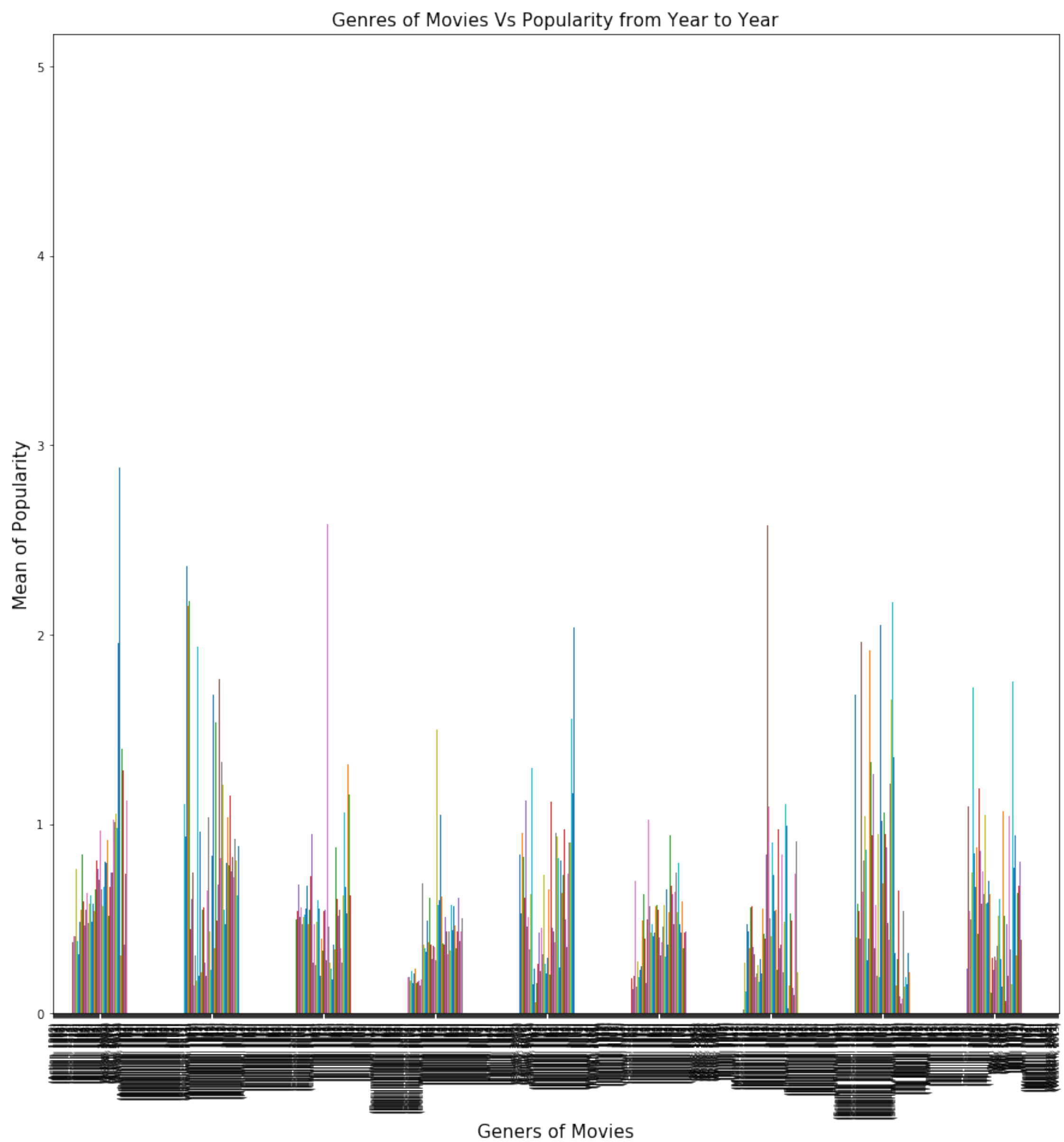
In [74]:
```
# using groupby again to see which genres are most popular from year to year
df.groupby(['genres_adj', 'release_year'])['popularity'].mean()
```

Out[74]:
```
genres_adj  release_year
Action      1960            0.590724
            1961            0.540904
            1962            0.299207
            1963            1.008599
            1964            0.254216
            1965            0.268987
            1966            0.254542
```

	1967	0.530274
	1968	0.368664
	1969	0.420294
	1970	0.227680
	1971	0.508694
	1972	0.343920
	1973	0.455597
	1974	0.331369
	1975	0.271900
	1976	0.374327
	1977	0.407406
	1978	0.409209
	1979	0.763054
	1980	0.381385
	1981	0.314431
	1982	0.483564
	1983	0.546067
	1984	0.840223
	1985	0.589887
	1986	0.462538
	1987	0.547405
	1988	0.636180
	1989	0.476612
		...
War	2014	1.273797
	2015	2.131503
Western	1961	0.210021
	1962	0.516593
	1964	0.127679
	1966	0.246072
	1967	0.139647
	1968	0.621202
	1970	0.568645
	1971	0.285940
	1972	0.476664
	1973	0.592252
	1975	0.162767
	1977	0.241629
	1979	0.262123
	1980	0.223935
	1982	0.360746
	1990	0.457183
	1992	0.841580
	1993	0.293473
	1994	0.363695
	1999	0.354484
	2002	1.040588
	2003	0.680803

```
                2004            0.780069
                2006            0.463068
                2007            1.150389
                2013            0.390628
                2014            0.760452
                2015            4.929120
Name: popularity, Length: 828, dtype: float64
```

In [127]:
```python
# plotting the previous result
df.groupby(['genres_adj', 'release_year'])['popularity'].mean().plot(kind='bar', figsi
plt.title('Genres of Movies Vs Popularity from Year to Year', fontsize=15)
plt.xlabel('Geners of Movies', fontsize=15)
plt.ylabel('Mean of Popularity', fontsize=15)
plt.xticks(fontsize=7, rotation=90);
```



1.4.3 Research Question 4 (what kind of movies that get the higher revenues)

In [91]:
```
# calculating the maximum revenue by any movie
df.revenue_adj.max()
```

Out[91]:
```
2827123750.41189
```

In [92]:
```
# choosing movies with revenues more tha 1000,000000,
max_revenue_movies = df.query('revenue_adj>1000000000')
max_revenue_movies
```

Out[92]:
```
          id  popularity                                 original_title  \
0     135397   32.985763                                 Jurassic World
3     140607   11.173104                   Star Wars: The Force Awakens
4     168259    9.335014                                      Furious 7
8     211672    7.404165                                        Minions
14     99861    5.944927                        Avengers: Age of Ultron
1329      11   12.037933                                      Star Wars
1334     840    1.104816            Close Encounters of the Third Kind
1386   19995    9.432768                                         Avatar
1921   12155    5.572950                            Alice in Wonderland
1930   10193    2.711136                                    Toy Story 3
2412    1893    3.526029        Star Wars: Episode I - The Phantom Menace
2633     120    8.575419  The Lord of the Rings: The Fellowship of the Ring
2634     671    8.021423           Harry Potter and the Philosopher's Stone
2875     155    8.466668                                The Dark Knight
3374   12445    5.711315     Harry Potter and the Deathly Hallows: Part 2
3522   38356    0.760503                   Transformers: Dark of the Moon
3911     121    8.095275              The Lord of the Rings: The Two Towers
3912     672    6.012584        Harry Potter and the Chamber of Secrets
4180    8587    4.782688                                  The Lion King
4361   24428    7.637767                                   The Avengers
4363   49026    6.591277                          The Dark Knight Rises
4365   37724    5.603587                                        Skyfall
4949     122    7.122455      The Lord of the Rings: The Return of the King
4955      12    3.440519                                   Finding Nemo
5231     597    4.355219                                        Titanic
5422  109445    6.112766                                         Frozen
5425   68721    4.946136                                     Iron Man 3
6190     674    5.939927            Harry Potter and the Goblet of Fire
6555      58    4.205992         Pirates of the Caribbean: Dead Man's Chest
6977     809    2.191033                                       Shrek 2
7269     238    5.738034                                  The Godfather
7309    1891    5.488441                        The Empire Strikes Back
7387     285    4.965391       Pirates of the Caribbean: At World's End
7987    1892    4.828854                             Return of the Jedi
```

8094	1642	1.136610	The Net
8095	532	1.115152	A Close Shave
8457	602	4.480733	Independence Day
8889	601	2.900556	E.T. the Extra-Terrestrial
9806	578	2.563191	Jaws
10110	12230	2.631987	One Hundred and One Dalmatians
10223	329	2.204926	Jurassic Park
10398	9325	2.550704	The Jungle Book
10594	9552	2.010733	The Exorcist
10690	15121	1.313676	The Sound of Music
10758	1924	1.210324	Superman

	director	runtime \
0	Colin Trevorrow	124
3	J.J. Abrams	136
4	James Wan	137
8	Kyle Balda|Pierre Coffin	91
14	Joss Whedon	141
1329	George Lucas	121
1334	Steven Spielberg	135
1386	James Cameron	162
1921	Tim Burton	108
1930	Lee Unkrich	103
2412	George Lucas	136
2633	Peter Jackson	178
2634	Chris Columbus	152
2875	Christopher Nolan	152
3374	David Yates	130
3522	Michael Bay	154
3911	Peter Jackson	179
3912	Chris Columbus	161
4180	Roger Allers|Rob Minkoff	89
4361	Joss Whedon	143
4363	Christopher Nolan	165
4365	Sam Mendes	143
4949	Peter Jackson	201
4955	Andrew Stanton|Lee Unkrich	100
5231	James Cameron	194
5422	Chris Buck|Jennifer Lee	102
5425	Shane Black	130
6190	Mike Newell	157
6555	Gore Verbinski	151
6977	Andrew Adamson|Kelly Asbury|Conrad Vernon	93
7269	Francis Ford Coppola	175
7309	Irvin Kershner	124
7387	Gore Verbinski	169
7987	Richard Marquand	135
8094	Irwin Winkler	114

8095	Nick Park	30
8457	Roland Emmerich	145
8889	Steven Spielberg	115
9806	Steven Spielberg	124
10110	Clyde Geronimi|Hamilton Luske|Wolfgang Reitherman	79
10223	Steven Spielberg	127
10398	Wolfgang Reitherman	78
10594	William Friedkin	122
10690	Robert Wise	174
10758	Richard Donner	143

	genres	release_date	vote_count	\
0	Action|Adventure|Science Fiction|Thriller	6/9/15	5562	
3	Action|Adventure|Science Fiction|Fantasy	12/15/15	5292	
4	Action|Crime|Thriller	4/1/15	2947	
8	Family|Animation|Adventure|Comedy	6/17/15	2893	
14	Action|Adventure|Science Fiction	4/22/15	4304	
1329	Adventure|Action|Science Fiction	3/20/77	4428	
1334	Science Fiction|Drama	11/16/77	600	
1386	Action|Adventure|Fantasy|Science Fiction	12/10/09	8458	
1921	Family|Fantasy|Adventure	3/3/10	2853	
1930	Animation|Family|Comedy	6/16/10	2924	
2412	Adventure|Action|Science Fiction	5/19/99	2823	
2633	Adventure|Fantasy|Action	12/18/01	6079	
2634	Adventure|Fantasy|Family	11/16/01	4265	
2875	Drama|Action|Crime|Thriller	7/16/08	8432	
3374	Adventure|Family|Fantasy	7/7/11	3750	
3522	Action|Science Fiction|Adventure	6/28/11	2456	
3911	Adventure|Fantasy|Action	12/18/02	5114	
3912	Adventure|Fantasy|Family	11/13/02	3458	
4180	Family|Animation|Drama	6/23/94	3489	
4361	Science Fiction|Action|Adventure	4/25/12	8903	
4363	Action|Crime|Drama|Thriller	7/16/12	6723	
4365	Action|Adventure|Thriller	10/25/12	6137	
4949	Adventure|Fantasy|Action	12/1/03	5636	
4955	Animation|Family	5/30/03	3692	
5231	Drama|Romance|Thriller	11/18/97	4654	
5422	Animation|Adventure|Family	11/27/13	3369	
5425	Action|Adventure|Science Fiction	4/18/13	6882	
6190	Adventure|Fantasy|Family	11/5/05	3406	
6555	Adventure|Fantasy|Action	6/20/06	3181	
6977	Adventure|Animation|Comedy|Family|Fantasy	5/19/04	1676	
7269	Drama|Crime	3/15/72	3970	
7309	Adventure|Action|Science Fiction	1/1/80	3954	
7387	Adventure|Fantasy|Action	5/19/07	2626	
7987	Adventure|Action|Science Fiction	5/23/83	3101	
8094	Crime|Drama|Mystery|Thriller|Action	7/28/95	201	
8095	Family|Animation|Comedy	12/24/95	115	

8457	Action|Adventure|Science Fiction	6/25/96	2000
8889	Science Fiction|Adventure|Family|Fantasy	4/3/82	1830
9806	Horror|Thriller|Adventure	6/18/75	1415
10110	Adventure|Animation|Comedy|Family	1/25/61	913
10223	Adventure|Science Fiction	6/11/93	3169
10398	Family|Animation|Adventure	10/18/67	928
10594	Drama|Horror|Thriller	12/26/73	1113
10690	Drama|Family|Music|Romance	3/2/65	620
10758	Adventure|Fantasy|Action|Science Fiction	12/14/78	518

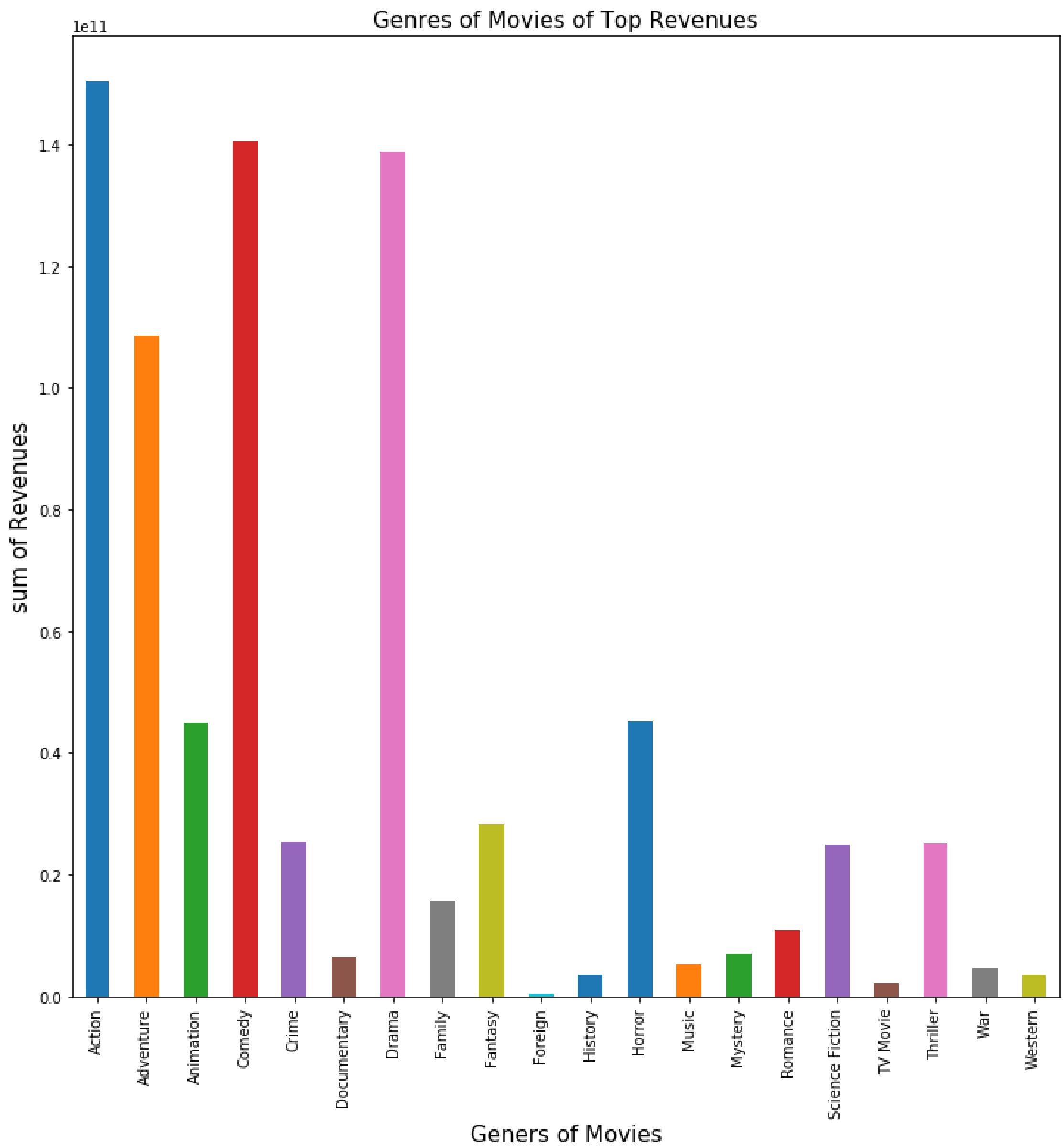
	vote_average	release_year	budget_adj	revenue_adj	genres_adj
0	6.5	2015	1.379999e+08	1.392446e+09	Action
3	7.5	2015	1.839999e+08	1.902723e+09	Action
4	7.3	2015	1.747999e+08	1.385749e+09	Action
8	6.5	2015	6.807997e+07	1.064192e+09	Family
14	7.4	2015	2.575999e+08	1.292632e+09	Action
1329	7.9	1977	3.957559e+07	2.789712e+09	Adventure
1334	7.0	1977	7.195562e+07	1.092965e+09	Science Fiction
1386	7.1	2009	2.408869e+08	2.827124e+09	Action
1921	6.3	2010	2.000000e+08	1.025467e+09	Family
1930	7.5	2010	2.000000e+08	1.063172e+09	Animation
2412	6.3	1999	1.505411e+08	1.209981e+09	Adventure
2633	7.8	2001	1.145284e+08	1.073080e+09	Adventure
2634	7.2	2001	1.539360e+08	1.202518e+09	Adventure
2875	8.1	2008	1.873655e+08	1.014733e+09	Drama
3374	7.7	2011	1.211748e+08	1.287184e+09	Adventure
3522	6.1	2011	1.890326e+08	1.089358e+09	Action
3911	7.8	2002	9.576865e+07	1.122902e+09	Adventure
3912	7.2	2002	1.212261e+08	1.062776e+09	Adventure
4180	7.7	1994	6.620002e+07	1.159592e+09	Family
4361	7.3	2012	2.089437e+08	1.443191e+09	Science Fiction
4363	7.5	2012	2.374361e+08	1.026713e+09	Action
4365	6.8	2012	1.899489e+08	1.052849e+09	Action
4949	7.9	2003	1.114231e+08	1.326278e+09	Adventure
4955	7.4	2003	1.114231e+08	1.024887e+09	Animation
5231	7.3	1997	2.716921e+08	2.506406e+09	Drama
5422	7.5	2013	1.404050e+08	1.192711e+09	Animation
5425	6.9	2013	1.872067e+08	1.137692e+09	Action
6190	7.3	2005	1.674845e+08	1.000353e+09	Adventure
6555	6.8	2006	2.163338e+08	1.152691e+09	Adventure
6977	6.5	2004	1.731668e+08	1.061904e+09	Adventure
7269	8.3	1972	3.128737e+07	1.277914e+09	Drama
7309	8.0	1980	4.762866e+07	1.424626e+09	Adventure
7387	6.8	2007	3.155006e+08	1.010654e+09	Adventure
7987	7.8	1983	7.082424e+07	1.253819e+09	Adventure
8094	5.6	1995	3.148127e+07	1.583050e+09	Crime
8095	7.4	1995	3.148127e+07	1.583050e+09	Family
8457	6.6	1996	1.042663e+08	1.135764e+09	Action

```
8889            7.2          1982  2.372625e+07  1.791694e+09  Science Fiction
9806            7.3          1975  2.836275e+07  1.907006e+09           Horror
10110           6.6          1961  2.917944e+07  1.574815e+09        Adventure
10223           7.4          1993  9.509661e+07  1.388863e+09        Adventure
10398           7.0          1967  2.614705e+07  1.345551e+09           Family
10594           7.2          1973  3.928928e+07  2.167325e+09            Drama
10690           7.2          1965  5.674862e+07  1.129535e+09            Drama
10758           6.7          1978  1.838485e+08  1.003539e+09        Adventure
```

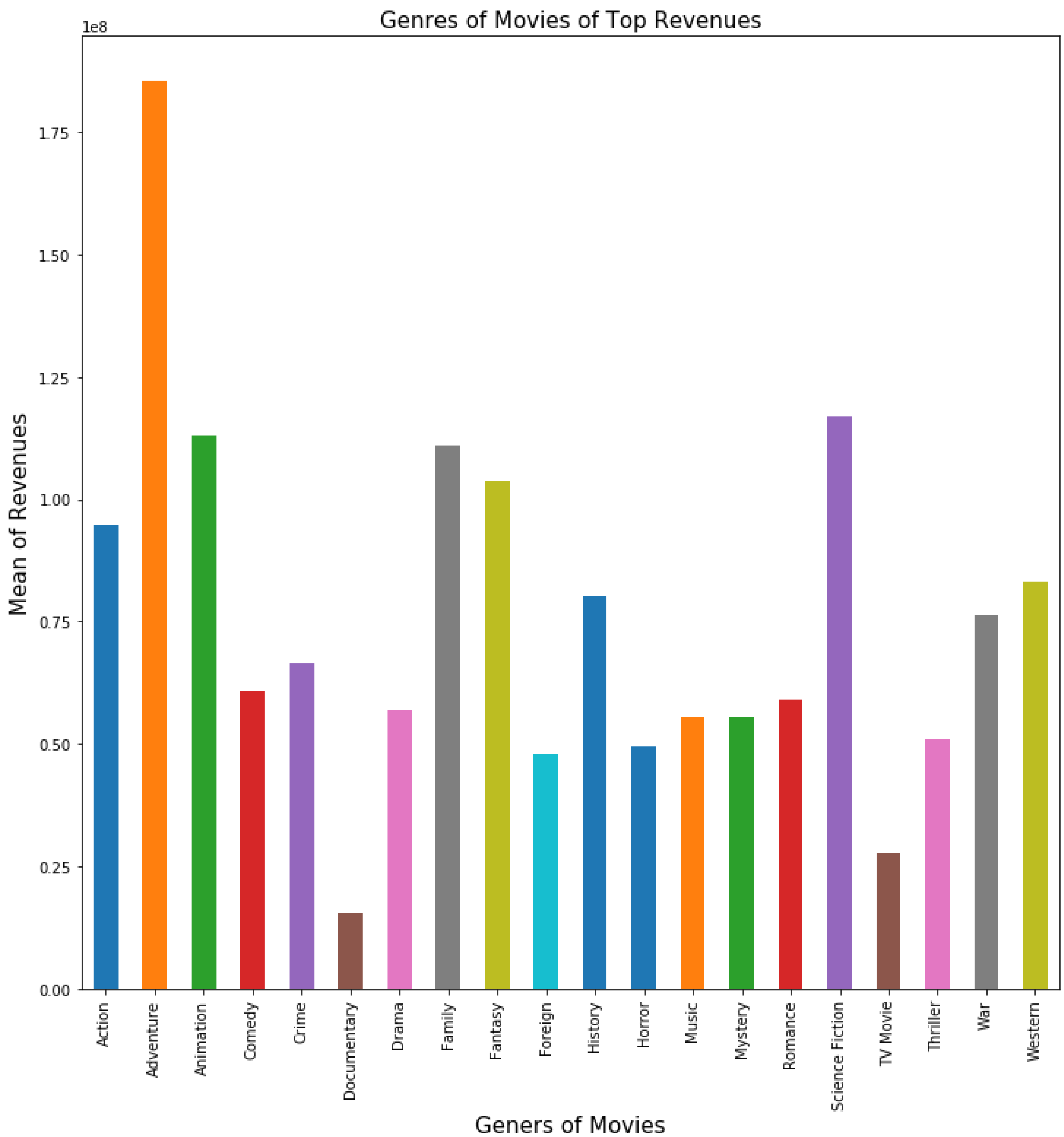
In [94]:
```python
df.groupby('genres_adj')['revenue_adj'].sum()
```

Out[94]:
```
genres_adj
Action             1.503273e+11
Adventure          1.086293e+11
Animation          4.493714e+10
Comedy             1.406284e+11
Crime              2.523616e+10
Documentary        6.538793e+09
Drama              1.388816e+11
Family             1.573888e+10
Fantasy            2.820308e+10
Foreign            4.316645e+08
History            3.526445e+09
Horror             4.520415e+10
Music              5.307794e+09
Mystery            6.914789e+09
Romance            1.092444e+10
Science Fiction    2.477224e+10
TV Movie           2.042366e+09
Thriller           2.499944e+10
War                4.501118e+09
Western            3.487604e+09
Name: revenue_adj, dtype: float64
```

In [116]:
```python
# Define Genres of movies with top sum of revenues
df.groupby('genres_adj')['revenue_adj'].sum().plot(kind='bar', figsize=(12,12))
plt.title('Genres of Movies of Top Revenues', fontsize=15)
plt.xlabel('Geners of Movies', fontsize=15)
plt.ylabel('Sum of Revenues', fontsize=15);
```




```
In [128]: # Define Genres of movies with top mean of revenues
          df.groupby('genres_adj')['revenue_adj'].mean().plot(kind='bar', figsize=(12,12));
          plt.title('Genres of Movies of Top Revenues', fontsize=15)
          plt.xlabel('Geners of Movies', fontsize=15)
          plt.ylabel('Mean of Revenues', fontsize=15);
```



Conclusions

1.5 from this investigation we can finally conclude the answers of the questions that were posed at the beginning of this investigation:

1- the populirty of movies are connected with both the budget of movie and of course the revenue of this movie. accordingly the movies with high budget or/and high revenue will be of higher popularity and viceversa.

2- the number of most movies genres (count) in IMDB are Drama films then Comedy film, Thriller and finally Action films, in the other hand the less number of movies genres available in IMDB are the Tv films and Western Films.

3- the most genres of movies with high popularity are adventure films in the foreground, then science fiction, fantasy, action and animation . . and these films with less popularity are foreign films, ducumentaries and tv-films.

4- what kind of movies that win higher revenues: in case we consider the higher sum of revenue, Action films are in the foreground then the comedy films, drama then adventure films, but in case we consider it with the higher mean avenues. then the adventure films wil be the first.

1.6 Submitting your Project

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```
In [2]: from subprocess import call
        call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

Out[2]: 0

In [ ]:
```