Investigate_a_Dataset_Report2_me

April 14, 2020

1 Project: TMDb movie data

1.1 Table of Contents

Introduction

In this project we will be analyzing a data set (cleaned from original data on Kaggle) associated with information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

In particular , we'll be interested in finding trends among the popular movies whith high revenues and ratings, and how they differed from other movies with lower ratings and revenues .

Questions to answer during this analyses :

Research Question 1 (Which genres are most popular from year to year ?)

Research Question 2 : What were the most and the least popular movies in this dataset ?

Research Question 3 : What are the 10 most popular Production companies and the respective success year ?

Research Question 4 : What kinds of properties are associated with movies that have high revenues ?

Research Question 5 : Who are the 5 most successful directors ?

Research Question 6 : How is the evolution of movies' budget through years ?

Research Question 7 : Is there a relationship between budget ,revenue with their respective adjustements ?

```
In [1]: #importing necessary libraries
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import datetime
        %matplotlib inline
```

Data Wrangling

1.1.1 General Properties

```
In [2]: # Loading data from "tmdb-movies.csv" file
        df=pd.read_csv('tmdb-movies.csv',header=0)
        df.head()
```

Out[2]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	...	\
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Colin Trevorrow	The park is open.	...	
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.com/	George Miller	What a Lovely Day.	...	
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insurgent	Robert Schwentke	One Choice Can Destroy You	...	
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-wars-episod...	J.J. Abrams	Every generation has a story.	...	
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.com/	James Wan		...	

```
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]
```

After executing some primordial observation codes , The first vision of the data gives us some information : It is data set with 10866 lines and 21 columns . These rows are movies and each one of them is associated with two unique identifiers (id and Imdb id) In the next columns we have : the popularity score , budget , revenue , original title , cast , home page , release date , revenue , etc ..

We probably won't analyze movies Id and imdb_Id (because they are very specific to the movie) , original title , homepage , tagline and overview won't be really useful for inspecting the trends between movies .

```
In [3]: df.shape
```

```
Out[3]: (10866, 21)
```

It is data set with 10866 lines and 21 columns .

```
In [4]: #convert release_date to timestamp
        df['release_date']=pd.to_datetime(df['release_date'])
```

```
In [5]: #cheking the conversion
        df.dtypes

Out[5]: id                             int64
        imdb_id                       object
        popularity                   float64
        budget                         int64
        revenue                        int64
        original_title                object
        cast                          object
        homepage                      object
        director                      object
        tagline                       object
        keywords                      object
        overview                      object
        runtime                        int64
        genres                        object
        production_companies          object
        release_date          datetime64[ns]
        vote_count                     int64
        vote_average                 float64
        release_year                   int64
        budget_adj                   float64
        revenue_adj                  float64
        dtype: object

In [6]: df.tail()

Out[6]:           id    imdb_id  popularity  budget  revenue  \
        10861     21  tt0060371    0.080598       0        0
        10862  20379  tt0060472    0.065543       0        0
        10863  39768  tt0060161    0.065141       0        0
        10864  21449  tt0061177    0.064317       0        0
        10865  22293  tt0060666    0.035919   19000        0

                        original_title  \
        10861        The Endless Summer
        10862                Grand Prix
        10863       Beregis Avtomobilya
        10864    What's Up, Tiger Lily?
        10865  Manos: The Hands of Fate

                                                        cast homepage  \
        10861  Michael Hynson|Robert August|Lord 'Tally Ho' B...      NaN
        10862  James Garner|Eva Marie Saint|Yves Montand|Tosh...      NaN
        10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...      NaN
        10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...      NaN
        10865  Harold P. Warren|Tom Neyman|John Reynolds|Dian...      NaN
```

```
                director                                                tagline  \
10861         Bruce Brown                                                   NaN
10862  John Frankenheimer  Cinerama sweeps YOU into a drama of speed and ...
10863      Eldar Ryazanov                                                   NaN
10864         Woody Allen                                WOODY ALLEN STRIKES BACK!
10865    Harold P. Warren      It's Shocking! It's Beyond Your Imagination!

          ...                                                  overview runtime  \
10861     ...       The Endless Summer, by Bruce Brown, is one of ...      95
10862     ...       Grand Prix driver Pete Aron is fired by his te...     176
10863     ...       An insurance agent who moonlights as a carthie...      94
10864     ...       In comic Woody Allen's film debut, he took the...      80
10865     ...       A family gets lost on the road and stumbles up...      74

                      genres  \
10861             Documentary
10862  Action|Adventure|Drama
10863          Mystery|Comedy
10864           Action|Comedy
10865                  Horror

                                  production_companies release_date  \
10861                                   Bruce Brown Films   2066-06-15
10862  Cherokee Productions|Joel Productions|Douglas ...   2066-12-21
10863                                          Mosfilm   2066-01-01
10864                           Benedict Pictures Corp.   2066-11-02
10865                                       Norm-Iris   2066-11-15

      vote_count  vote_average  release_year     budget_adj  revenue_adj
10861         11           7.4          1966       0.000000          0.0
10862         20           5.7          1966       0.000000          0.0
10863         11           6.5          1966       0.000000          0.0
10864         22           5.4          1966       0.000000          0.0
10865         15           1.5          1966  127642.279154          0.0

[5 rows x 21 columns]
```

Clearly we have a problem in the datetime format , python is proceeding with number of the year between 17 and 68 as they are in the next century (example : year 1966 becomes 2066 after conversion) . That's why we're going to integrate a function to correct this :

```
In [73]: def fix_date(x):

             if x.year -100 > 1960:
                 year = x.year-100
```

```
        else:

            year = x.year

        return datetime.date(year,x.month,x.day)


    df['release_date'] = df['release_date'].apply(fix_date)
```

In [74]:
```
df.tail()
```

Out[74]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	...	overview	runtime	genres
10861	21	tt0060371	0.080598	0	0	The Endless Summer	Michael Hynson|Robert August|Lord 'Tally Ho' B...	NaN	Bruce Brown	NaN	...	The Endless Summer, by Bruce Brown, is one of ...	95	Documentary
10862	20379	tt0060472	0.065543	0	0	Grand Prix	James Garner|Eva Marie Saint|Yves Montand|Tosh...	NaN	John Frankenheimer	Cinerama sweeps YOU into a drama of speed and ...	...	Grand Prix driver Pete Aron is fired by his te...	176	
10863	39768	tt0060161	0.065141	0	0	Beregis Avtomobilya	Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...	NaN	Eldar Ryazanov	NaN	...	An insurance agent who moonlights as a carthie...	94	
10864	21449	tt0061177	0.064317	0	0	What's Up, Tiger Lily?	Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...	NaN	Woody Allen	WOODY ALLEN STRIKES BACK!	...	In comic Woody Allen's film debut, he took the...	80	
10865	22293	tt0060666	0.035919	19000	0	Manos: The Hands of Fate	Harold P. Warren|Tom Neyman|John Reynolds|Dian...	NaN	Harold P. Warren	It's Shocking! It's Beyond Your Imagination!	...	A family gets lost on the road and stumbles up...	74	

```
10862  Action|Adventure|Drama
10863           Mystery|Comedy
10864            Action|Comedy
10865                  Horror

                                    production_companies release_date  \
10861                                  Bruce Brown Films   1966-06-15
10862  Cherokee Productions|Joel Productions|Douglas ...   1966-12-21
10863                                           Mosfilm   1966-01-01
10864                           Benedict Pictures Corp.   1966-11-02
10865                                       Norm-Iris   1966-11-15

      vote_count  vote_average  release_year     budget_adj  revenue_adj
10861         11           7.4          1966       0.000000          0.0
10862         20           5.7          1966       0.000000          0.0
10863         11           6.5          1966       0.000000          0.0
10864         22           5.4          1966       0.000000          0.0
10865         15           1.5          1966  127642.279154          0.0

[5 rows x 21 columns]
```

In [75]:
```
#some summary statistics
df.describe()
```

Out[75]:

	id	popularity	budget	revenue	runtime
count	10866.000000	10866.000000	1.086600e+04	1.086600e+04	10866.000000
mean	66064.177434	0.646441	1.462570e+07	3.982332e+07	102.070863
std	92130.136561	1.000185	3.091321e+07	1.170035e+08	31.381405
min	5.000000	0.000065	0.000000e+00	0.000000e+00	0.000000
25%	10596.250000	0.207583	0.000000e+00	0.000000e+00	90.000000
50%	20669.000000	0.383856	0.000000e+00	0.000000e+00	99.000000
75%	75610.000000	0.713817	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	10866.000000	10866.000000	10866.000000	1.086600e+04	1.086600e+04
mean	217.389748	5.974922	2001.322658	1.755104e+07	5.136436e+07
std	575.619058	0.935142	12.812941	3.430616e+07	1.446325e+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%	145.750000	6.600000	2011.000000	2.085325e+07	3.369710e+07
max	9767.000000	9.200000	2015.000000	4.250000e+08	2.827124e+09

--> After executing some summary statistics , observations :

--> The least popular movie got ~0 score and the most popular one got ~33 popularity score

--> the minimum budget,revenue and runtime are null which is weird (we should investigate about it)

--> minimum vote count is very far from the max vote count (10 and 9767)

--> the majority of movies runtime was between 90 and 111 minutes / the majority of movies were rated between 5 and 6

--> The oldest movie in this data set is released in 1960 and the newest one was released in 2015 .

--> The majority of movies in this dataset were released between 1995 and 2011

--> These movies had a mean adjusted budget of ~150 million $

In [76]: df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    10866 non-null  int64
 1   imdb_id               10856 non-null  object
 2   popularity            10866 non-null  float64
 3   budget                10866 non-null  int64
 4   revenue               10866 non-null  int64
 5   original_title        10866 non-null  object
 6   cast                  10790 non-null  object
 7   homepage              2936 non-null   object
 8   director              10822 non-null  object
 9   tagline               8042 non-null   object
 10  keywords              9373 non-null   object
 11  overview              10862 non-null  object
 12  runtime               10866 non-null  int64
 13  genres                10843 non-null  object
 14  production_companies  9836 non-null   object
 15  release_date          10866 non-null  object
 16  vote_count            10866 non-null  int64
 17  vote_average          10866 non-null  float64
 18  release_year          10866 non-null  int64
 19  budget_adj            10866 non-null  float64
 20  revenue_adj           10866 non-null  float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

1.1.2 Data Cleaning

Columns with missing values : imdb_id /cast/homepage/homepage/director/tagline/keywords/overvie

We are going to drop the columns we are not going to use :

id , imdb_id , cast , homepage , tagline ,keywords, overview

```
In [77]: df.drop(['id', 'imdb_id', 'cast', 'homepage', 'tagline','keywords', 'overview'],axis=1,

In [78]: df.head()
```

Out[78]:

	popularity	budget	revenue	original_title	director	runtime	genres	production_companies	release_date	vote_count	vote_average	release_year	budget_adj	revenue_adj
0	32.985763	150000000	1513528810	Jurassic World	Colin Trevorrow	124	Action|Adventure|Science Fiction|Thriller	Universal Studios|Amblin Entertainment|Legenda...	2015-06-09	5562	6.5	2015	1.379999e+08	1.392446e+09
1	28.419936	150000000	378436354	Mad Max: Fury Road	George Miller	120	Action|Adventure|Science Fiction|Thriller	Village Roadshow Pictures|Kennedy Miller Produ...	2015-05-13	6185	7.1	2015	1.379999e+08	3.481613e+08
2	13.112507	110000000	295238201	Insurgent	Robert Schwentke	119	Adventure|Science Fiction|Thriller	Summit Entertainment|Mandeville Films|Red Wago...	2015-03-18	2480	6.3	2015	1.012000e+08	2.716190e+08
3	11.173104	200000000	2068178225	Star Wars: The Force Awakens	J.J. Abrams	136	Action|Adventure|Science Fiction|Fantasy	Lucasfilm|Truenorth Productions|Bad Robot	2015-12-15	5292	7.5	2015	1.839999e+08	1.902723e+09
4	9.335014	190000000	1506249360	Furious 7	James Wan	137	Action|Crime|Thriller	Universal Pictures|Original Film|Media Rights ...	2015-04-01	2947	7.3	2015	1.747999e+08	1.385749e+09

For a better and clearer analyzing we can make the 'original_title' as the first column .

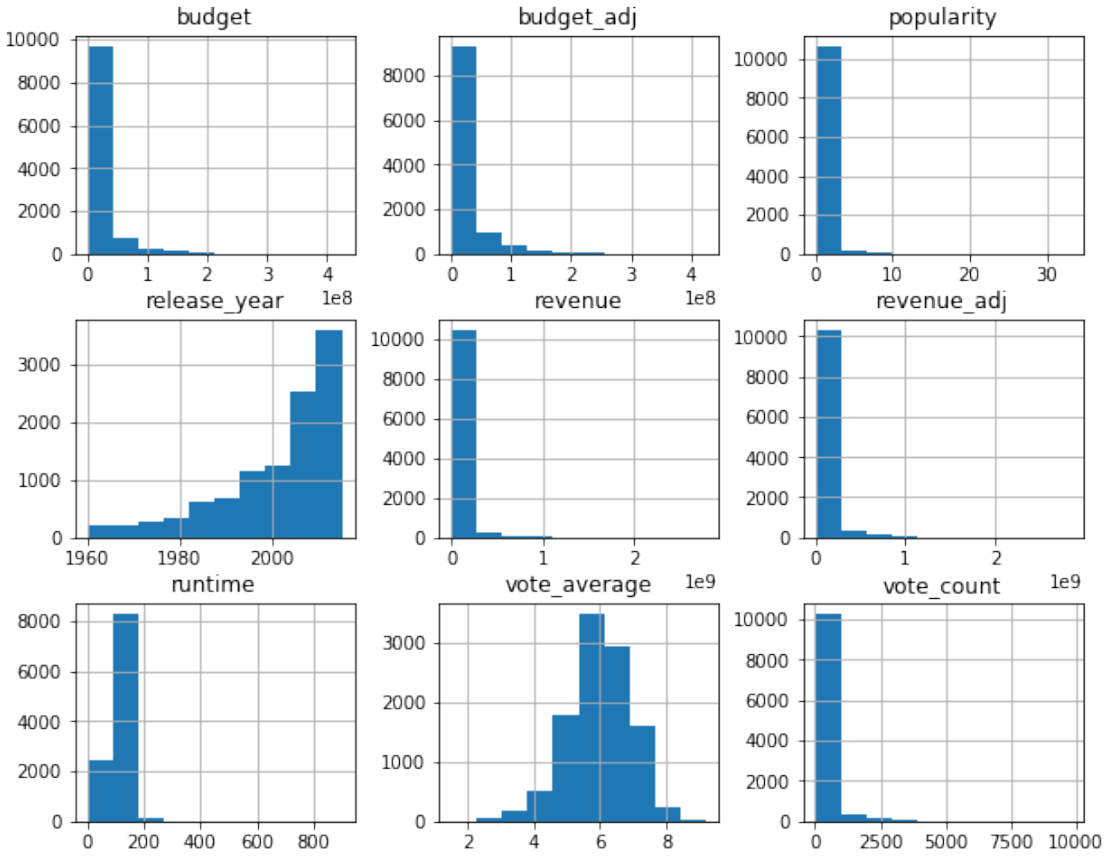
```
In [79]: #change  'original_title' position as the first column
         swap=df['original_title']
         df.drop(['original_title'],axis=1,inplace=True)
         df.insert(0,'original_title',swap)

In [80]: df.head()
```

Out[80]:

	original_title	popularity	budget	revenue
0	Jurassic World	32.985763	150000000	1513528810
1	Mad Max: Fury Road	28.419936	150000000	378436354
2	Insurgent	13.112507	110000000	295238201

```
3  Star Wars: The Force Awakens   11.173104  200000000  2068178225
4                     Furious 7    9.335014  190000000  1506249360

           director  runtime                                     genres  \
0   Colin Trevorrow      124  Action|Adventure|Science Fiction|Thriller
1     George Miller      120  Action|Adventure|Science Fiction|Thriller
2  Robert Schwentke      119         Adventure|Science Fiction|Thriller
3       J.J. Abrams      136   Action|Adventure|Science Fiction|Fantasy
4         James Wan      137                      Action|Crime|Thriller

                               production_companies release_date  vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...   2015-06-09        5562
1  Village Roadshow Pictures|Kennedy Miller Produ...   2015-05-13        6185
2  Summit Entertainment|Mandeville Films|Red Wago...   2015-03-18        2480
3          Lucasfilm|Truenorth Productions|Bad Robot   2015-12-15        5292
4  Universal Pictures|Original Film|Media Rights ...   2015-04-01        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
```

In [81]:
```
df.hist(figsize=(10,8));
```



--> These plots agree with summary statistics we saw before

We notice that the plots of revenue , revenue_adj and vote count are skewed to the right

--> The majority of movies had a budget aound 20 million$
-->Vote average was between 5 and 7 --> The majority had a runtime of ~150minutes

```
In [82]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   original_title        10866 non-null  object
 1   popularity            10866 non-null  float64
 2   budget                10866 non-null  int64
 3   revenue               10866 non-null  int64
 4   director              10822 non-null  object
 5   runtime               10866 non-null  int64
```

```
 6   genres                10843 non-null  object
 7   production_companies  9836 non-null   object
 8   release_date          10866 non-null  object
 9   vote_count            10866 non-null  int64
 10  vote_average          10866 non-null  float64
 11  release_year          10866 non-null  int64
 12  budget_adj            10866 non-null  float64
 13  revenue_adj           10866 non-null  float64
dtypes: float64(4), int64(5), object(5)
memory usage: 1.2+ MB
```

we have missing data from 3 data columns : director ,genres , production_companies

```
In [83]: #checking the null rows of production_companies
         df[df.production_companies.isnull()]
```

Out[83]:

	original_title	popularity	budget	revenue	director	runtime	genres	production_companies	release_date	vote_count	vote_average
228	Racing Extinction	0.584363	0	0	Louie Psihoyos	90	Adventure|Documentary	NaN	2015-01-24	36	7.8
259	Crown for Christmas	0.476341	0	0	Alex Zamm	84	TV Movie	NaN	2015-11-27	10	7.6
295	12 Gifts of Christmas	0.417191	0	0	Peter Sullivan	84	Family|TV Movie	NaN	2015-11-26	12	6.3
298	The Girl in the Photographs	0.370258	0	0	Nick Simon	95	Crime|Horror|Thriller	NaN	2015-09-14	10	4.7
328	Advantageous	0.367617	0	0	Jennifer Phang	92	Science Fiction|Drama|Family	NaN			
...	...	...	...	...	...	...	...				
10804	Interiors	0.149259	0	0	Woody Allen	93	Drama	NaN			
10806	Gates of Heaven	0.138635	0	0	Errol Morris	85	Documentary	NaN			
10816	The Rutles: All You Need Is Cash	0.064602	0	0	Eric Idle|Gary Weis	76	Comedy	NaN			
10842	Winnie the Pooh and the Honey Tree	0.253437	0	0	Wolfgang Reitherman	25	Animation|Family	NaN			
10853	Alfie	0.163592	0	0	Lewis Gilbert	114	Comedy|Drama|Romance	NaN			

```
328                    NaN   2015-06-23          29           6.4
...                    ...          ...         ...           ...
10804                  NaN   1978-08-02          35           6.3
10806                  NaN   1978-10-01          12           5.9
10816                  NaN   1978-03-22          14           6.0
10842                  NaN   1966-01-01          12           7.9
10853                  NaN   1966-03-29          26           6.2

       release_year  budget_adj  revenue_adj
228            2015         0.0          0.0
259            2015         0.0          0.0
295            2015         0.0          0.0
298            2015         0.0          0.0
328            2015         0.0          0.0
...             ...         ...          ...
10804          1978         0.0          0.0
10806          1978         0.0          0.0
10816          1978         0.0          0.0
10842          1966         0.0          0.0
10853          1966         0.0          0.0

[1030 rows x 14 columns]
```

In [84]:
```
#checking the null rows of production_companies
df[df.director.isnull()]
```

Out[84]:
```
                                          original_title  popularity   budget  \
532                      Iliza Shlesinger: Freezing Hot    0.126594        0
548                          Sense8: Creating the World    0.108072        0
556                                        With This Ring    0.100910        0
1032             Marvel Studios: Assembling a Universe    0.291253        0
1054                                  Unlocking Sherlock    0.269468        0
1203                                       Free to Play    0.119891   150000
1241                                          Dance-Off    0.135376        0
1288                 Top Gear: The Perfect Road Trip 2    0.038364        0
1852                           The Diary of Anne Frank    0.256703        0
1872                                               Paa    0.091395  3250000
1895                  Doctor Who: The Waters of Mars    0.056777        0
2221                  Scott Pilgrim vs. the Animation    0.281852        0
2286                  Bo Burnham: Words, Words, Words    0.207234        0
2290      Across the Line: The Exodus of Charlie Wright    0.203502        0
2315                           Listen to Your Heart    0.171615        0
2318                        Barbie in A Mermaid Tale    0.170408        0
2376                   Doctor Who: A Christmas Carol    0.068411        0
2397                  The Making of The Walking Dead    0.033048        0
2401    Opeth: In Live Concert At The Royal Albert Hall    0.067753        0
2403                    Yu-Gi-Oh! 3D: Bonds Beyond Time    0.067620        0
3171   The Assassination of Jesse James: Death Of An ...    0.019819        0
```

3224	John Mayer: Where the Light Is Live in Los Ang...	0.224721	0
3276	Kismat Konnection	0.147657	4180000
3285	Tropic Thunder: Rain of Madness	0.136883	0
3357	Bill Burr: Why Do I Do This?	0.042517	0
3365	Foo Fighters: Live at Wembley Stadium	0.002475	0
3369	Wizards On Deck With Hannah Montana	0.001682	0
3910	Steve Jobs: One Last Thing	0.002006	0
4679	Barbie in A Mermaid Tale 2	0.421746	0
4797	Doctor Who: The Snowmen	0.167501	0
4872	Party Bercy	0.090552	0
4939	The Men Who Built America	0.003183	0
5413	The Brave Little Toaster to the Rescue	0.324270	0
5866	Jinxed	0.211825	5000000
5915	Phineas and Ferb: Mission Marvel	0.168210	0
5972	Top Gear: The Perfect Road Trip	0.165605	0
6033	Russell Brand: Messiah Complex	0.048587	0
6181	North and South, Book I	0.000065	0
7579	La hora frÃŋa	0.443952	0
7767	Doctor Who: The Runaway Bride	0.126603	0
7814	Transformers: Beginnings	0.040311	0
9593	Peter Pan	0.001662	0
10386	The Making of 'The Nightmare Before Christmas'	0.118854	0
10426	Magical Mystery Tour	0.114034	0

	revenue	director	runtime	\
532	0	NaN	71	
548	0	NaN	25	
556	0	NaN	105	
1032	0	NaN	43	
1054	0	NaN	60	
1203	0	NaN	75	
1241	0	NaN	0	
1288	0	NaN	94	
1852	0	NaN	100	
1872	0	NaN	133	
1895	0	NaN	62	
2221	0	NaN	4	
2286	0	NaN	63	
2290	0	NaN	94	
2315	0	NaN	0	
2318	0	NaN	75	
2376	0	NaN	62	
2397	0	NaN	30	
2401	0	NaN	163	
2403	0	NaN	60	
3171	0	NaN	32	
3224	0	NaN	164	
3276	11000000	NaN	153	

```
3285          0      NaN       30
3357          0      NaN       55
3365          0      NaN      120
3369          0      NaN       68
3910          0      NaN       60
4679          0      NaN       74
4797          0      NaN       60
4872          0      NaN      120
4939          0      NaN      360
5413          0      NaN       74
5866          0      NaN       75
5915          0      NaN       44
5972          0      NaN       84
6033          0      NaN       99
6181          0      NaN      561
7579          0      NaN       92
7767          0      NaN       60
7814          0      NaN       22
9593          0      NaN       52
10386         0      NaN       25
10426         0      NaN       55

                                          genres  \
532                                       Comedy
548                 Documentary|Science Fiction
556                             Comedy|Romance
1032                   TV Movie|Documentary
1054                   TV Movie|Documentary
1203                            Documentary
1241                 Romance|Music|Comedy
1288                            Documentary
1852                                  Drama
1872                 Drama|Family|Foreign
1895          Science Fiction|Adventure|Family
2221              TV Movie|Animation|Adventure
2286                                  Comedy
2290           Drama|Action|Thriller|Crime
2315                   Drama|Music|Romance
2318              Animation|Family|Fantasy
2376                                   NaN
2397                            Documentary
2401                                  Music
2403                        Animation|Fantasy
3171                            Documentary
3224                                  Music
3276           Drama|Comedy|Romance|Foreign
3285                          Action|Comedy
3357                                  Comedy
```

```
3365                                                Music
3369                                               Family
3910                                          Documentary
4679                                    Animation|Family
4797                                                 NaN
4872                                               Comedy
4939                                 Documentary|History
5413                             Animation|Comedy|Family
5866                               Family|Fantasy|Comedy
5915                  Action|Adventure|Science Fiction
5972              TV Movie|Action|Adventure|Documentary
6033                                               Comedy
6181                               Drama|History|Western
7579   Horror|Thriller|Science Fiction|Mystery|Foreign
7767                            Science Fiction|TV Movie
7814         Animation|Action|Thriller|Science Fiction
9593         Action|Adventure|Animation|Family|Fantasy
10386                                         Documentary
10426                                               Music

                                    production_companies release_date  \
532                                New Wave Entertainment   2015-01-23
548                                               Netflix   2015-08-10
556         Lifetime Television|Sony Pictures Television   2015-01-24
1032                           Marvel Studios|ABC Studios   2014-03-18
1054                                                 NaN   2014-01-19
1203                                               Valve   2014-03-19
1241                                                 NaN   2014-01-01
1288                                    2 Entertain Video   2014-11-17
1852   Darlow Smithson Productions|British Broadcasti...   2009-01-09
1872                                           A B Corp   2009-12-04
1895                                          BBC Wales   2009-12-19
2221                                           Titmouse   2010-08-10
2286                                                 NaN   2010-10-19
2290                                                 NaN   2010-12-28
2315                                                 NaN   2010-08-14
2318                                                 NaN   2010-01-26
2376                                                 NaN   2010-12-25
2397                                                 NaN   2010-07-31
2401                                                 NaN   2010-09-21
2403                                                 NaN   2010-01-23
3171                               New Wave Entertainment   2008-01-01
3224                                                 NaN   2008-07-01
3276                                     Tips Industries   2008-07-18
3285                                                 NaN   2008-08-26
3357                                                 NaN   2008-09-23
3365                                                 NaN   2008-08-25
3369                                                 NaN   2008-02-11
```

3910	NaN	2011-11-02
4679	Mattel	2012-02-23
4797	BBC Television UK	2012-12-25
4872	TF1 VidÃľo	2012-09-23
4939	NaN	2012-10-16
5413	NaN	1997-11-01
5866	NaN	2013-11-29
5915	Disney Television Animation	2013-08-16
5972	BBC	2013-11-17
6033	NaN	2013-11-25
6181	NaN	1985-11-03
7579	NaN	2007-09-14
7767	NaN	2007-07-06
7814	DreamWorks Home Entertainment	2007-10-16
9593	Burbank Films Australia	1988-01-01
10386	Buena Vista Home Entertainment	1993-10-03
10426	MPI Home Video	1967-12-25

	vote_count	vote_average	release_year	budget_adj	revenue_adj
532	14	6.6	2015	0.000000e+00	0.000000e+00
548	12	7.5	2015	0.000000e+00	0.000000e+00
556	14	6.5	2015	0.000000e+00	0.000000e+00
1032	32	6.3	2014	0.000000e+00	0.000000e+00
1054	11	7.2	2014	0.000000e+00	0.000000e+00
1203	40	7.0	2014	1.381637e+05	0.000000e+00
1241	18	5.7	2014	0.000000e+00	0.000000e+00
1288	12	6.8	2014	0.000000e+00	0.000000e+00
1852	19	7.5	2009	0.000000e+00	0.000000e+00
1872	11	6.1	2009	3.303301e+06	0.000000e+00
1895	25	8.0	2009	0.000000e+00	0.000000e+00
2221	19	7.7	2010	0.000000e+00	0.000000e+00
2286	13	6.5	2010	0.000000e+00	0.000000e+00
2290	11	6.0	2010	0.000000e+00	0.000000e+00
2315	29	7.3	2010	0.000000e+00	0.000000e+00
2318	35	6.3	2010	0.000000e+00	0.000000e+00
2376	11	7.7	2010	0.000000e+00	0.000000e+00
2397	42	8.4	2010	0.000000e+00	0.000000e+00
2401	10	8.6	2010	0.000000e+00	0.000000e+00
2403	10	6.0	2010	0.000000e+00	0.000000e+00
3171	20	6.8	2008	0.000000e+00	0.000000e+00
3224	16	8.5	2008	0.000000e+00	0.000000e+00
3276	11	5.8	2008	4.233448e+06	1.114065e+07
3285	12	6.9	2008	0.000000e+00	0.000000e+00
3357	10	8.0	2008	0.000000e+00	0.000000e+00
3365	10	8.4	2008	0.000000e+00	0.000000e+00
3369	14	6.1	2008	0.000000e+00	0.000000e+00
3910	11	6.6	2011	0.000000e+00	0.000000e+00
4679	22	6.0	2012	0.000000e+00	0.000000e+00

```
4797           10           7.8          2012  0.000000e+00  0.000000e+00
4872           15           6.4          2012  0.000000e+00  0.000000e+00
4939           11           5.3          2012  0.000000e+00  0.000000e+00
5413           11           7.0          1997  0.000000e+00  0.000000e+00
5866           19           7.0          2013  4.680167e+06  0.000000e+00
5915           13           5.9          2013  0.000000e+00  0.000000e+00
5972           35           7.6          2013  0.000000e+00  0.000000e+00
6033           12           6.8          2013  0.000000e+00  0.000000e+00
6181           17           6.0          1985  0.000000e+00  0.000000e+00
7579           10           4.9          2007  0.000000e+00  0.000000e+00
7767           18           7.6          2007  0.000000e+00  0.000000e+00
7814           34           5.8          2007  0.000000e+00  0.000000e+00
9593           28           6.6          1988  0.000000e+00  0.000000e+00
10386          15           7.5          1993  0.000000e+00  0.000000e+00
10426          15           5.8          1967  0.000000e+00  0.000000e+00
```

In [85]:
```
#checking the null rows of production_companies
df[df.genres.isnull()]
```

Out[85]:

	original_title	popularity	budget	revenue	director	runtime
424	Belli di papÃ	0.244648	0	0	Guido Chiesa	100
620	All Hallows' Eve 2	0.129696	0	0	Antonio Padovan|Bryan Norton|Marc Roussel|Ryan...	90
997	Star Wars Rebels: Spark of Rebellion	0.330431	0	0	Steward Lee|Steven G. Lee	44
1712	Prayers for Bobby	0.302095	0	0		
1897	Jonas Brothers: The Concert Experience	0.020701	0	0		
2370	Freshman Father	0.081892	0	0		
2376	Doctor Who: A Christmas Carol	0.068411	0	0		
2853	Vizontele	0.130018	0	0		
3279	ìêÿřì ë	0.145331	0			
4547	London 2012 Olympic Opening Ceremony: Isles of...	0.520520	0	0		
4732	The Scapegoat	0.235911	0	0		
4797	Doctor Who: The Snowmen	0.167501	0	0		
4890	Cousin Ben Troop Screening	0.083202	0	0		
5830	Doctor Who: The Time of the Doctor	0.248944	0	0		
5934	Prada: Candy	0.067433	0	0		
6043	Bombay Talkies	0.039080	0	0		
6530	Saw Rebirth	0.092724	0	0		
8234	Viaggi di nozze	0.028874	0	0		
8614	T2 3-D: Battle Across Time	0.273934	0	0		
8878	Mom's Got a Date With a Vampire	0.038045	0	0		
9307	Goldeneye	0.094652	0	0		
9799	The Amputee	0.175008	0	0		
10659	The Party at Kitty and Stud's	0.344172	5000	0		

```
1712         0                                    Russell Mulcahy       88
1897         0                                    Bruce Hendricks       76
2370         0                                      Michael Scott        0
2376         0                                                NaN       62
2853         0                                   YÄślmaz ErdoÄan      110
3279         0                                      Kim Jin-Yeong       96
4547         0                                        Danny Boyle      220
4732         0                                    Charles Sturridge      100
4797         0                                                NaN       60
4890         0                                       Wes Anderson        2
5830         0                                        James Payne       60
5934         0                          Wes Anderson|Roman Coppola        3
6043         0  Anurag Kashyap|Dibakar Banerjee|Zoya Akhtar|Ka...      127
6530         0                          Jeff Shuter|Daniel Viney        6
8234         0                                      Carlo Verdone      103
8614         0                                      James Cameron       12
8878         0                                        Steve Boyum       85
9307         0                                           Don Boyd      105
9799         0                                        David Lynch        5
10659        0                                      Morton Lewis       71

      genres                production_companies release_date  vote_count  \
424      NaN                                 NaN   2015-10-29          21
620      NaN  Ruthless Pictures|Hollywood Shorts   2015-10-06          13
997      NaN                                 NaN   2014-10-03          13
1712     NaN         Daniel Sladek Entertainment   2009-02-27          57
1897     NaN                                 NaN   2009-02-27          11
2370     NaN                                 NaN   2010-06-05          12
2376     NaN                                 NaN   2010-12-25          11
2853     NaN                                 NaN   2001-02-02          12
3279     NaN                                 NaN   2008-08-13          11
4547     NaN                                 BBC   2012-07-27          12
4732     NaN                    Island Pictures   2012-09-09          12
4797     NaN                  BBC Television UK   2012-12-25          10
4890     NaN                                 NaN   2012-01-01          14
5830     NaN                                 NaN   2013-12-25          26
5934     NaN                                 NaN   2013-03-25          27
6043     NaN           Viacom 18 Motion Pictures   2013-05-03          12
6530     NaN                                 NaN   2005-10-24          24
8234     NaN                                 NaN   1995-12-15          44
8614     NaN                                 NaN   1996-01-01          14
8878     NaN                Walt Disney Pictures   2000-10-13          16
9307     NaN                   Anglia Television   1989-08-26          10
9799     NaN                                 NaN   1974-01-01          11
10659    NaN             Stallion Releasing Inc.   1970-02-10          10

       vote_average  release_year   budget_adj  revenue_adj
424             6.1          2015      0.00000          0.0
```

```
          620             5.0          2015      0.00000          0.0
          997             6.8          2014      0.00000          0.0
          1712            7.4          2009      0.00000          0.0
          1897            7.0          2009      0.00000          0.0
          2370            5.8          2010      0.00000          0.0
          2376            7.7          2010      0.00000          0.0
          2853            7.2          2001      0.00000          0.0
          3279            6.1          2008      0.00000          0.0
          4547            8.3          2012      0.00000          0.0
          4732            6.2          2012      0.00000          0.0
          4797            7.8          2012      0.00000          0.0
          4890            7.0          2012      0.00000          0.0
          5830            8.5          2013      0.00000          0.0
          5934            6.9          2013      0.00000          0.0
          6043            5.9          2013      0.00000          0.0
          6530            5.9          2005      0.00000          0.0
          8234            6.7          1995      0.00000          0.0
          8614            6.7          1996      0.00000          0.0
          8878            5.4          2000      0.00000          0.0
          9307            5.3          1989      0.00000          0.0
          9799            5.0          1974      0.00000          0.0
          10659           3.0          1970  28081.84172          0.0
```

In [86]:
```
#we are going to drop these null rows
df.dropna(inplace=True)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9807 entries, 0 to 10865
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   original_title        9807 non-null   object
 1   popularity            9807 non-null   float64
 2   budget                9807 non-null   int64
 3   revenue               9807 non-null   int64
 4   director              9807 non-null   object
 5   runtime               9807 non-null   int64
 6   genres                9807 non-null   object
 7   production_companies  9807 non-null   object
 8   release_date          9807 non-null   object
 9   vote_count            9807 non-null   int64
 10  vote_average          9807 non-null   float64
 11  release_year          9807 non-null   int64
 12  budget_adj            9807 non-null   float64
 13  revenue_adj           9807 non-null   float64
dtypes: float64(4), int64(5), object(5)
```

```
memory usage: 1.1+ MB
```

```
In [87]: #cheking duplicates
         df.duplicated().sum()
```

```
Out[87]: 1
```

```
In [88]: df[df.duplicated()]
```

```
Out[88]:      original_title  popularity    budget  revenue          director  runtime  \
         2090         TEKKEN     0.59643  30000000   967000  Dwight H. Little       92

                                                 genres    production_companies  \
         2090  Crime|Drama|Action|Thriller|Science Fiction  Namco|Light Song Films

              release_date  vote_count  vote_average  release_year  budget_adj  \
         2090   2010-03-20         110           5.0          2010  30000000.0

               revenue_adj
         2090     967000.0
```

We have one duplicated row (TEKKEN movie) so we have to drop it

```
In [89]: df.drop_duplicates(inplace=True)
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9806 entries, 0 to 10865
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   original_title        9806 non-null   object
 1   popularity            9806 non-null   float64
 2   budget                9806 non-null   int64
 3   revenue               9806 non-null   int64
 4   director              9806 non-null   object
 5   runtime               9806 non-null   int64
 6   genres                9806 non-null   object
 7   production_companies  9806 non-null   object
 8   release_date          9806 non-null   object
 9   vote_count            9806 non-null   int64
 10  vote_average          9806 non-null   float64
 11  release_year          9806 non-null   int64
 12  budget_adj            9806 non-null   float64
 13  revenue_adj           9806 non-null   float64
dtypes: float64(4), int64(5), object(5)
memory usage: 1.1+ MB
```

```
In [90]: df.tail()
```

Out[90]:

	original_title	popularity	budget	revenue	\
10861	The Endless Summer	0.080598	0	0	
10862	Grand Prix	0.065543	0	0	
10863	Beregis Avtomobilya	0.065141	0	0	
10864	What's Up, Tiger Lily?	0.064317	0	0	
10865	Manos: The Hands of Fate	0.035919	19000	0	

	director	runtime	genres	\
10861	Bruce Brown	95	Documentary	
10862	John Frankenheimer	176	Action|Adventure|Drama	
10863	Eldar Ryazanov	94	Mystery|Comedy	
10864	Woody Allen	80	Action|Comedy	
10865	Harold P. Warren	74	Horror	

	production_companies	release_date	\
10861	Bruce Brown Films	1966-06-15	
10862	Cherokee Productions|Joel Productions|Douglas ...	1966-12-21	
10863	Mosfilm	1966-01-01	
10864	Benedict Pictures Corp.	1966-11-02	
10865	Norm-Iris	1966-11-15	

	vote_count	vote_average	release_year	budget_adj	revenue_adj
10861	11	7.4	1966	0.000000	0.0
10862	20	5.7	1966	0.000000	0.0
10863	11	6.5	1966	0.000000	0.0
10864	22	5.4	1966	0.000000	0.0
10865	15	1.5	1966	127642.279154	0.0

We see that there are some null values in budget , revenu , budget_adj , revenue_adj :

--> we're going to replace these null values with the mean

```
In [91]: df.budget_adj.isnull().sum()
```

Out[91]: 0

```
In [92]: #Filling null values of budget with the mean
         mean=df['budget'].mean()
         df['budget'].replace(0,mean,inplace=True)
```

```
In [93]: df.tail()
```

Out[93]:

	original_title	popularity	budget	revenue	\
10861	The Endless Summer	0.080598	1.612525e+07	0	
10862	Grand Prix	0.065543	1.612525e+07	0	
10863	Beregis Avtomobilya	0.065141	1.612525e+07	0	
10864	What's Up, Tiger Lily?	0.064317	1.612525e+07	0	

```
10865  Manos: The Hands of Fate    0.035919  1.900000e+04        0

                director  runtime                  genres  \
10861        Bruce Brown       95             Documentary
10862  John Frankenheimer      176  Action|Adventure|Drama
10863     Eldar Ryazanov        94          Mystery|Comedy
10864        Woody Allen        80           Action|Comedy
10865    Harold P. Warren       74                  Horror

                                    production_companies release_date  \
10861                                  Bruce Brown Films   1966-06-15
10862  Cherokee Productions|Joel Productions|Douglas ...   1966-12-21
10863                                          Mosfilm     1966-01-01
10864                           Benedict Pictures Corp.    1966-11-02
10865                                       Norm-Iris     1966-11-15

       vote_count  vote_average  release_year     budget_adj  revenue_adj
10861          11           7.4          1966       0.000000          0.0
10862          20           5.7          1966       0.000000          0.0
10863          11           6.5          1966       0.000000          0.0
10864          22           5.4          1966       0.000000          0.0
10865          15           1.5          1966  127642.279154          0.0
```

In [94]:
```
#Filling null values of revenue with the mean
mean=df['revenue'].mean()
df['revenue'].replace(0,mean,inplace=True)
```

In [95]:
```
#Filling null values of revenue adjusts with the mean
mean=df['revenue_adj'].mean()
df['revenue_adj'].replace(0,mean,inplace=True)
```

In [96]:
```
#Filling null values of budget adjusts with the mean
mean=df['budget_adj'].mean()
df['budget_adj'].replace(0,mean,inplace=True)
```

In [97]:
```
##Filling null values of runtime with the mean

mean=df['runtime'].mean()
df['runtime'].replace(0,mean,inplace=True)
```

In [98]:
```
df.tail()
```

Out[98]:
```
                 original_title  popularity        budget       revenue  \
10861        The Endless Summer    0.080598  1.612525e+07  4.407785e+07
10862                Grand Prix    0.065543  1.612525e+07  4.407785e+07
10863       Beregis Avtomobilya    0.065141  1.612525e+07  4.407785e+07
10864    What's Up, Tiger Lily?    0.064317  1.612525e+07  4.407785e+07
10865  Manos: The Hands of Fate    0.035919  1.900000e+04  4.407785e+07
```

```
                director  runtime                 genres  \
10861        Bruce Brown     95.0            Documentary
10862  John Frankenheimer    176.0  Action|Adventure|Drama
10863     Eldar Ryazanov     94.0          Mystery|Comedy
10864        Woody Allen     80.0           Action|Comedy
10865    Harold P. Warren     74.0                 Horror

                                  production_companies release_date  \
10861                                  Bruce Brown Films   1966-06-15
10862  Cherokee Productions|Joel Productions|Douglas ...   1966-12-21
10863                                           Mosfilm   1966-01-01
10864                           Benedict Pictures Corp.   1966-11-02
10865                                       Norm-Iris   1966-11-15

       vote_count  vote_average  release_year    budget_adj   revenue_adj
10861          11           7.4          1966  1.935070e+07  5.685528e+07
10862          20           5.7          1966  1.935070e+07  5.685528e+07
10863          11           6.5          1966  1.935070e+07  5.685528e+07
10864          22           5.4          1966  1.935070e+07  5.685528e+07
10865          15           1.5          1966  1.276423e+05  5.685528e+07
```

```
In [99]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9806 entries, 0 to 10865
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   original_title        9806 non-null   object
 1   popularity            9806 non-null   float64
 2   budget                9806 non-null   float64
 3   revenue               9806 non-null   float64
 4   director              9806 non-null   object
 5   runtime               9806 non-null   float64
 6   genres                9806 non-null   object
 7   production_companies  9806 non-null   object
 8   release_date          9806 non-null   object
 9   vote_count            9806 non-null   int64
 10  vote_average          9806 non-null   float64
 11  release_year          9806 non-null   int64
 12  budget_adj            9806 non-null   float64
 13  revenue_adj           9806 non-null   float64
dtypes: float64(7), int64(2), object(5)
memory usage: 1.1+ MB
```

--> After cleaning , now we have 9806 rows instead of 10866

Exploratory Data Analysis

As a beginning , I've chosen popularity as my dependent variable and release_year , genres , producton_companies and budget as independent ones .

1.1.3 Research Question 1 (Which genres are most popular from year to year ?)

Cleaned dataset :

```
In [100]: df.tail()
```

Out[100]:

	original_title	popularity	budget	revenue	director	runtime	genres	production_companies	release_date	vote_count	vote_average	release_year	budget_adj	revenue_adj
10861	The Endless Summer	0.080598	1.612525e+07	4.407785e+07	Bruce Brown	95.0	Documentary	Bruce Brown Films	1966-06-15	11	7.4	1966	1.935070e+07	5.685528e+07
10862	Grand Prix	0.065543	1.612525e+07	4.407785e+07	John Frankenheimer	176.0	Action|Adventure|Drama	Cherokee Productions|Joel Productions|Douglas ...	1966-12-21	20	5.7	1966	1.935070e+07	5.685528e+07
10863	Beregis Avtomobilya	0.065141	1.612525e+07	4.407785e+07	Eldar Ryazanov	94.0	Mystery|Comedy	Mosfilm	1966-01-01	11	6.5	1966	1.935070e+07	5.685528e+07
10864	What's Up, Tiger Lily?	0.064317	1.612525e+07	4.407785e+07	Woody Allen	80.0	Action|Comedy	Benedict Pictures Corp.	1966-11-02	22	5.4	1966	1.935070e+07	5.685528e+07
10865	Manos: The Hands of Fate	0.035919	1.900000e+04	4.407785e+07	Harold P. Warren	74.0	Horror	Norm-Iris	1966-11-15	15	1.5	1966	1.276423e+05	5.685528e+07

```
In [101]: dfw=df.groupby(['genres','popularity' ])[['release_year']].mean()
          dfw.sort_values(by='release_year')
```

Out[101]:

genres	popularity	release_year
Action|Adventure|Western	1.872132	1960.0
Comedy|Romance	0.875173	1960.0
Horror	0.065808	1960.0
Comedy|Family	0.114188	1960.0
Drama	0.138777	1960.0
...		...
	0.547827	2015.0

```
                                0.520250          2015.0
                                0.474788          2015.0
                                0.661471          2015.0
Documentary|Music               0.227580          2015.0

[9803 rows x 1 columns]
```

1.1.4 Research Question 2 : What were the most and the least popular movies in this dataset ?

```
In [102]: #Most popular
          df[df['popularity']==df['popularity'].max()]
```

```
Out[102]:   original_title  popularity       budget       revenue         director  \
          0  Jurassic World   32.985763  150000000.0  1.513529e+09  Colin Trevorrow

             runtime                                     genres  \
          0    124.0  Action|Adventure|Science Fiction|Thriller

                                       production_companies release_date  vote_count  \
          0  Universal Studios|Amblin Entertainment|Legenda...   2015-06-09        5562

             vote_average  release_year    budget_adj   revenue_adj
          0           6.5          2015  1.379999e+08  1.392446e+09
```

```
In [103]: #Least popular
          df[df['popularity']==df['popularity'].min()]
```

```
Out[103]:      original_title  popularity        budget       revenue       director  \
          9977   The Hospital    0.000188  1.612525e+07  4.407785e+07  Arthur Hiller

                runtime                genres production_companies release_date  \
          9977    103.0  Mystery|Comedy|Drama   Simcha Productions   1971-12-14

                vote_count  vote_average  release_year    budget_adj   revenue_adj
          9977          10           6.4          1971  1.935070e+07  5.685528e+07
```

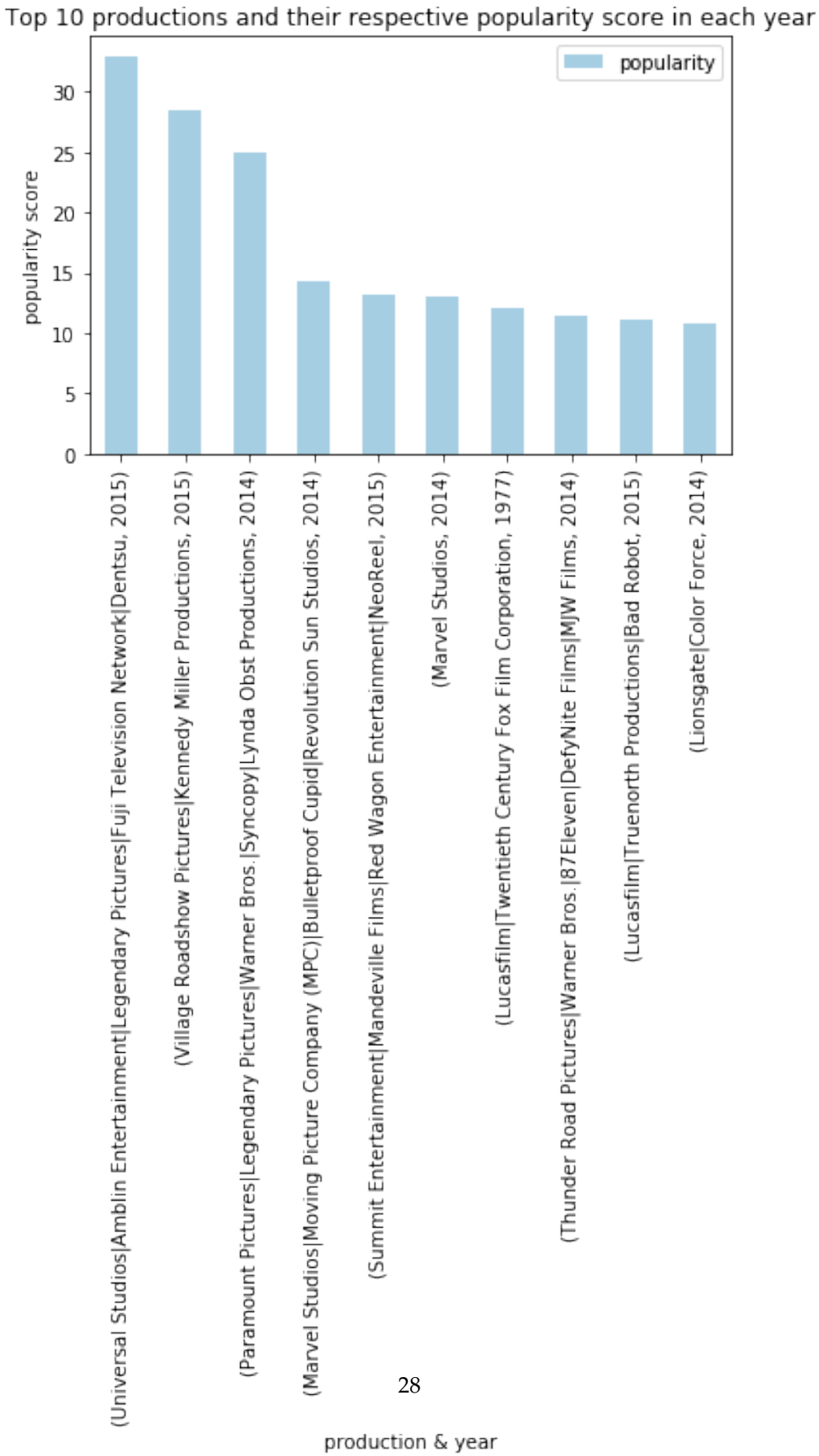
1.1.5 Research Question 3 : What are the 10 most popular Production companies and the respective success year ?

```
In [104]: df_00=df.groupby(['production_companies','release_year' ])[['popularity']].max()
          df_00.sort_values(by='popularity',ascending=False).head(10)
```

```
Out[104]:                                                                 popularity
          production_companies                                release_year
          Universal Studios|Amblin Entertainment|Legendar... 2015          32.985763
          Village Roadshow Pictures|Kennedy Miller Produc... 2015          28.419936
          Paramount Pictures|Legendary Pictures|Warner Br... 2014          24.949134
          Marvel Studios|Moving Picture Company (MPC)|Bul... 2014          14.311205
          Summit Entertainment|Mandeville Films|Red Wagon... 2015          13.112507
```

```
Marvel Studios                                          2014          12.971027
Lucasfilm|Twentieth Century Fox Film Corporation        1977          12.037933
Thunder Road Pictures|Warner Bros.|87Eleven|Def...      2014          11.422751
Lucasfilm|Truenorth Productions|Bad Robot               2015          11.173104
Lionsgate|Color Force                                   2014          10.739009
```

In [105]:
```
#Plotting
ax=df_00.sort_values(by='popularity',ascending=False).head(10).plot(kind='bar' , title
ax.set_xlabel("production & year")
ax.set_ylabel("popularity score")
plt.show()
```


Top 10 productions and their respective popularity score in each year
popularity
popularity score
0
5
10
15
20
25
30
(Universal Studios|Amblin Entertainment|Legendary Pictures|Fuji Television Network|Dentsu, 2015)
(Village Roadshow Pictures|Kennedy Miller Productions, 2015)
(Paramount Pictures|Legendary Pictures|Warner Bros.|Syncopy|Lynda Obst Productions, 2014)
(Marvel Studios|Moving Picture Company (MPC)|Bulletproof Cupid|Revolution Sun Studios, 2014)
(Summit Entertainment|Mandeville Films|Red Wagon Entertainment|NeoReel, 2015)
(Marvel Studios, 2014)
(Lucasfilm|Twentieth Century Fox Film Corporation, 1977)
(Thunder Road Pictures|Warner Bros.|87Eleven|DefyNite Films|MJW Films, 2014)
(Lucasfilm|Truenorth Productions|Bad Robot, 2015)
(Lionsgate|Color Force, 2014)
production & year

--> We see that the most popular properties got their positions with their latest productions (2015/2014 . .)

1.1.6 Research Question 4 : What kinds of properties are associated with movies that have high revenues ?

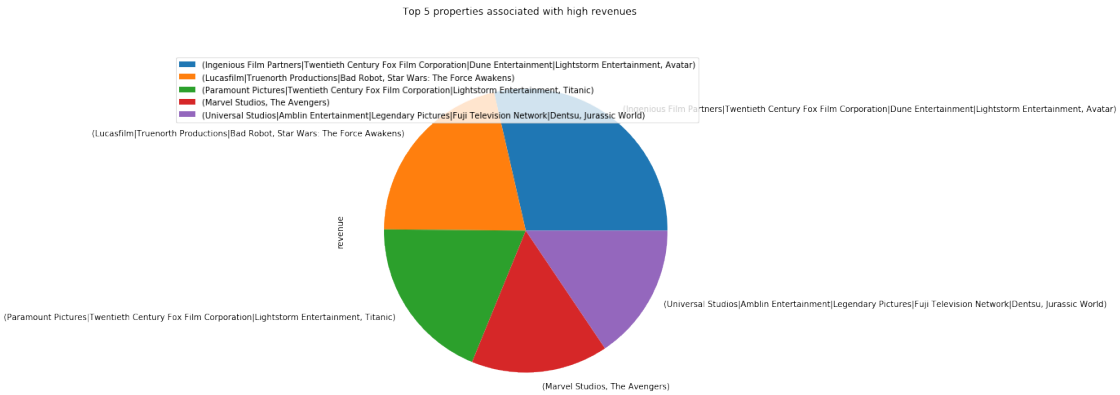
```
In [106]: df_01=df.groupby(['production_companies','original_title' ])[['revenue']].max()
          df_01.sort_values(by='revenue',ascending=False).head()
```

Out[106]:

production_companies	original_title	
Ingenious Film Partners|Twentieth Century Fox F...	Avatar	2.781
Lucasfilm|Truenorth Productions|Bad Robot	Star Wars: The Force Awakens	2.068
Paramount Pictures|Twentieth Century Fox Film C...	Titanic	1.845
Marvel Studios	The Avengers	1.519
Universal Studios|Amblin Entertainment|Legendar...	Jurassic World	1.513

We'll generate a plot to see how the cinema's "CAKE" of production companies is devided :

```
In [107]: #Plotting
          df_01.sort_values(by='revenue',ascending=False).head().plot(kind='pie',title='Top 5 pr
          plt.show()
```



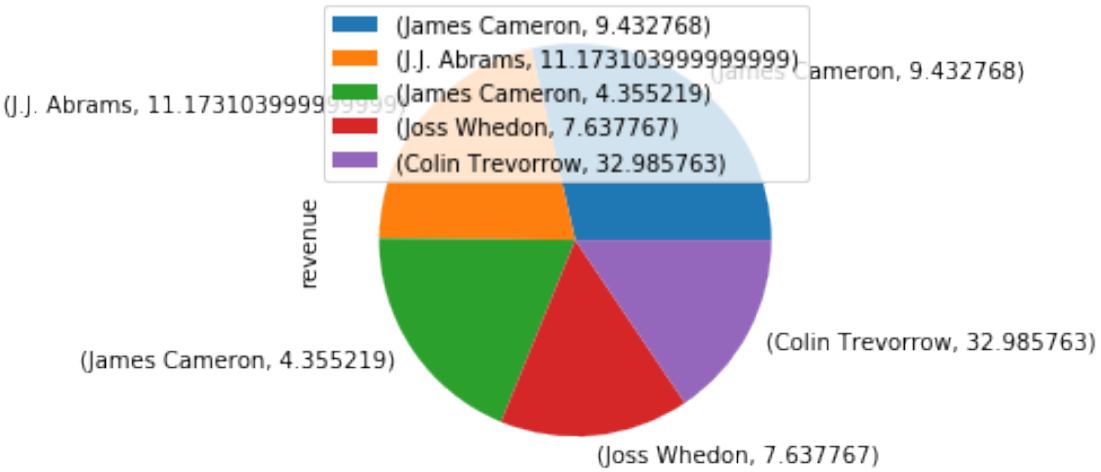
1.1.7 Research Question 5 : Who are the 5 most successful directors ?

```
In [108]: df_02=df.groupby(['director','popularity' ])[['revenue']].max()
          df_02.sort_values(by='revenue',ascending=False).head()
```

Out[108]:

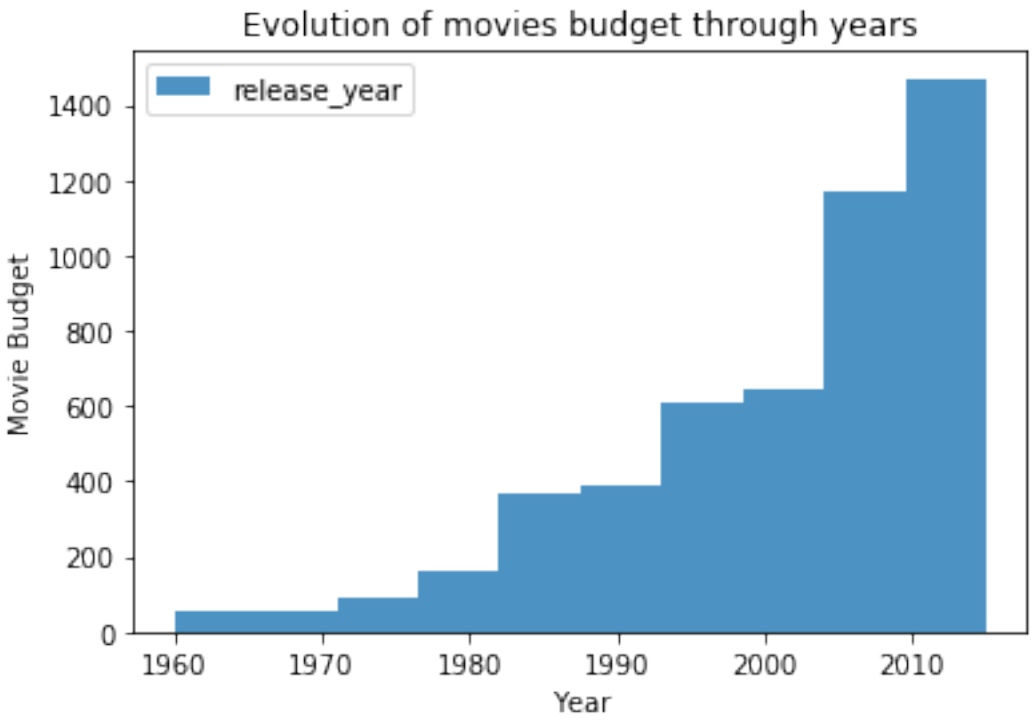
director	popularity	revenue
James Cameron	9.432768	2.781506e+09
J.J. Abrams	11.173104	2.068178e+09
James Cameron	4.355219	1.845034e+09
Joss Whedon	7.637767	1.519558e+09
Colin Trevorrow	32.985763	1.513529e+09

In [109]:
```
#Plotting
df_02.sort_values(by='revenue',ascending=False).head().plot(kind='pie',subplots=True);
plt.show()
```



1.1.8 Research Question 6 : How is the evolution of movies' budget through years ?

In [110]:
```
dfr=df.groupby(['budget','revenue' ])[['release_year']].max()
dfr.sort_values(by='release_year');
ay=dfr.plot(kind='hist',title='Evolution of movies budget through years', alpha=0.8);
ay.set_xlabel("Year")
ay.set_ylabel("Movie Budget")
plt.show()
```

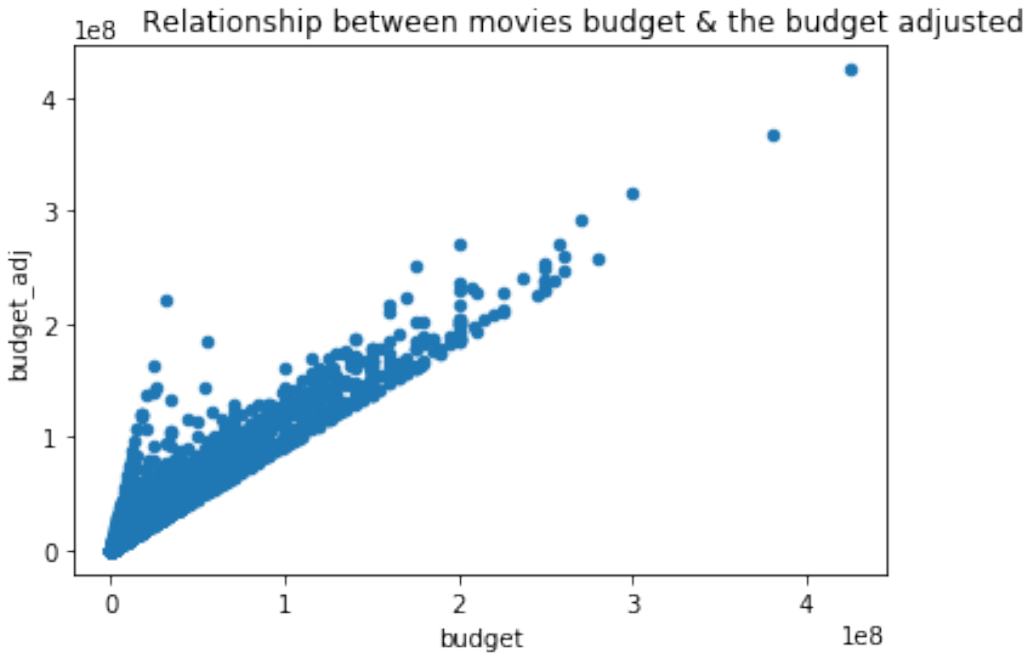


This plot is skewed to the left

Movies budget has increased heavily since the 80's

1.1.9 Research Question 7 : Is there a relationship between budget ,revenue with their respective adjustements ?

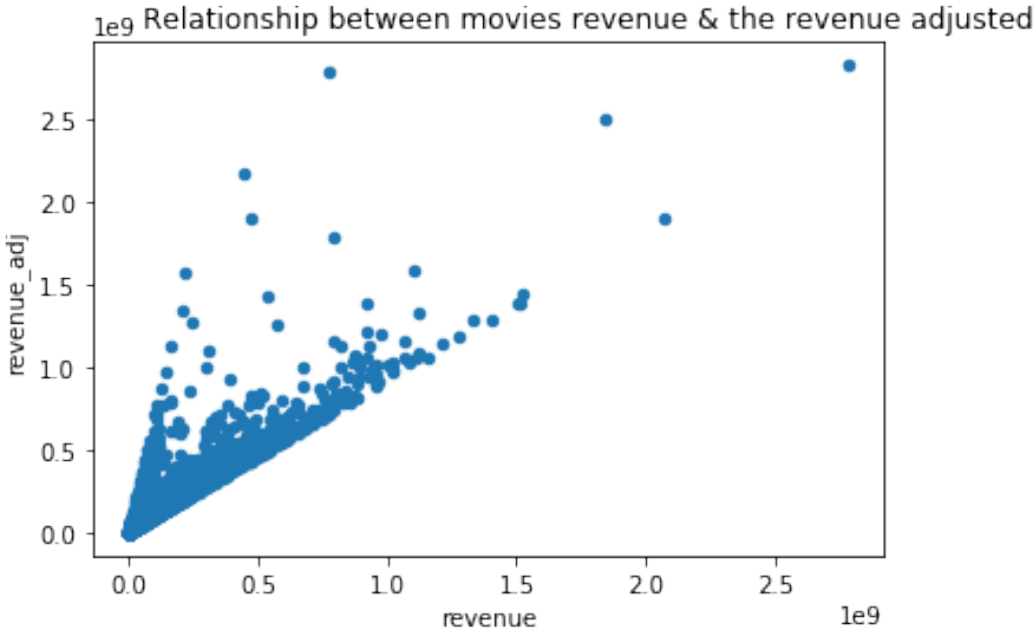
```
In [111]: df.plot(x='budget', y='budget_adj', kind='scatter' , title='                    Rela

Out[111]: <matplotlib.axes._subplots.AxesSubplot at 0x1b945e65f48>
```



--> The budget needed and the budget adjusted are highly correlated

```
In [112]: df.plot(x='revenue', y='revenue_adj', kind='scatter' , title='                    Re
```

```
Out[112]: <matplotlib.axes._subplots.AxesSubplot at 0x1b945eae808>
```



--> The revenue needed accomplished and the revenue adjusted are highly correlated

==> This correlation is due to that numerical values of the budget are very close to their adjusted ones thanks to the approximative statistic studies accomplished to assign the budget . (Same case with the revenue)

Limitations

During data Wrangling : "release_date" column was not in a datetime format . After conversion , the years before 1969 have increased 100 year (1966 became 2066) that's why i tried to create a function to fix this error and apply it in my code .

The release dates were not sorted , that's why some operations needed that the dates get sorted .

Many production companies were not mentioned (1030 companys with Nan) .

Some rows had irrelevant null values .

Conclusions

---> To get clear analyses to this dataset , I had to execute some cleaning operations by dropping unecessary columns (Id , Imdb_id , cast', 'homepage', 'tagline','keywords', 'overview') . I also had to fill columns having null values with their respective mean ('revenue, 'revenue_adj' , 'budget' , 'budget_adj' , 'runtime') . Furthermore , I found a duplicated row so i dropped it .

During my analyses , I have chosen popularity score as my dependent variable to inspect and investigate its relationships and trends with the other variables .

This dataset after analyzing it can give us information about movie industry .

We analyzed the distribution of poplarity between movies , production companies and directors .

We observed the evolution of financial amount through the years .

1.1.10 Resources i refered to :

https://stackoverflow.com/questions/11927715/how-to-give-a-pandas-matplotlib-bar-graph-custom-colors https://cmdlinetips.com/2018/02/how-to-sort-pandas-dataframe-by-columns-and-row/ https://www.geeksforgeeks.org/different-ways-to-iterate-over-rows-in-pandas-dataframe/ https://www.pythonprogramming.in/find-minimum-and-maximum-value-of-all-columns-from-pandas-dataframe.html https://stackoverflow.com/questions/15741759/find-maximum-value-of-a-column-and-return-the-corresponding-row-values-using-pan https://stackoverflow.com/questions/12169170/find-the-max-of-two-or-more-columns-with-pandas https://stackoverflow.com/questions/29583312/pandas-sum-of-duplicate-attributes https://stackoverflow.com/questions/51914255/pandas-how-to-sum-values-in-a-column-for-duplicate-rows https://www.geeksforgeeks.org/python-working-with-date-and-time-using-pandas/ https://riptutorial.com/python/example/2789/iterate-over-dates https://knowledge.udacity.com/questions/116305

https://knowledge.udacity.com/questions/117107 https://stackoverflow.com/questions/21487329/add-x-and-y-labels-to-a-pandas-plot https://seaborn.pydata.org/tutorial.html https://asq.org/quality-resources/scatter-diagram#Use

In [114]:
```
from subprocess import call
call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

Out[114]:
```
4294967295
```

In [ ]:

In [ ]: