Investigate_a_Dataset

March 13, 2023

1 PROJECT - TMDb Movie Data

1.1 Table of content

```
<li><a href="#intro">Introduction</a></li>
<li><a href="#wrangling">Data Wrangling</a></li>
<li><a href="#eda">Exporatory Data Analysis</a></li>
<li><a href="#conclusions">Conclusions</a></li>
```

Introduction

1.1.1 Dataset Description

In this project I will be analysing a data set containing information about 10,000 movies collected from The Movie Databasedata associated with movie genres in correlation with their rating and in particular I will be interested in finding trends amongst the movie generes that are highly rated

1.1.2 Question(s) for Analysis

Research Question 1
Which genres are highly rated
Research Question 2
Does movie budget influence movie rating or revenue?
Research Question 3
WWhat kinds of properties are associated with movies that have high rating?

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        % matplotlib inline
        # imported seaborn to make my visualization look nice.

        df = pd.read_csv('tmdb-movies.csv')
        df.head()
```

```
Out[1]:        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 [2]: # Upgrade pandas to use dataframe.explode() function.
        !pip install --upgrade pandas==0.25.0
```

```
Requirement already up-to-date: pandas==0.25.0 in /opt/conda/lib/python3.6/site-packages (0.25.0
Requirement already satisfied, skipping upgrade: pytz>=2017.2 in /opt/conda/lib/python3.6/site-p
Requirement already satisfied, skipping upgrade: numpy>=1.13.3 in /opt/conda/lib/python3.6/site-
Requirement already satisfied, skipping upgrade: python-dateutil>=2.6.1 in /opt/conda/lib/python
Requirement already satisfied, skipping upgrade: six>=1.5 in /opt/conda/lib/python3.6/site-packa
```

Data Wrangling

1.1.3 General Properties

```
In [3]: # Load your data and print out a few lines. Perform operations to inspect data
        #   types and look for instances of missing or possibly errant data.
```

```
In [4]: # data exploration
        df.shape
```

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

```
In [5]: # data exploration
        df.describe()
```

```
Out[5]:                  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
```

```
In [6]: # data exploration
        df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                2936 non-null object
director                10822 non-null object
tagline                 8042 non-null object
keywords                9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies    9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

1.1.4 Data Cleaning

The final two columns ending with "_adj" show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time, therefore the budget and revenue columns are irrelevant to the analysis, hence will be dropped

```
In [7]: df = pd.read_csv('tmdb-movies.csv')
        df.drop(['budget', 'revenue'], axis = 1, inplace = True)
        df.head()
```

Out[7]:

	id	imdb_id	popularity	original_title	cast	homepage	director	tagline	keywords	overview	runtime	genres
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Colin Trevorrow	The park is open.	monster|dna|tyrannosaurus rex|velociraptor|island	Twenty-two years after the events of Jurassic ...	124	Action|Adventure|Science Fiction|Thriller
1	76341	tt1392190	28.419936	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.com/	George Miller	What a Lovely Day.	future|chase|post-apocalyptic|dystopia|australia	An apocalyptic story set in the furthest reach...	120	
2	262500	tt2908446	13.112507	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insurgent	Robert Schwentke	One Choice Can Destroy You	based on novel|revolution|dystopia|sequel|dyst...	Beatrice Prior must confront her inner demons ...	119	
3	140607	tt2488496	11.173104	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.	android|spaceship|jedi|space opera|3d	Thirty years after defeating the Galactic Empi...	136	
4	168259	tt2820852	9.335014	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.com/	James Wan	Vengeance Hits Home	car race|speed|revenge|suspense|car	Deckard Shaw seeks revenge against Dominic Tor...	137	

```
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
```

There are some columns with missing data that are not particularly relevant to the analysis, hence i will drop the following columns:
1) homepage
2) tagline
3) keywords
4) production_companies

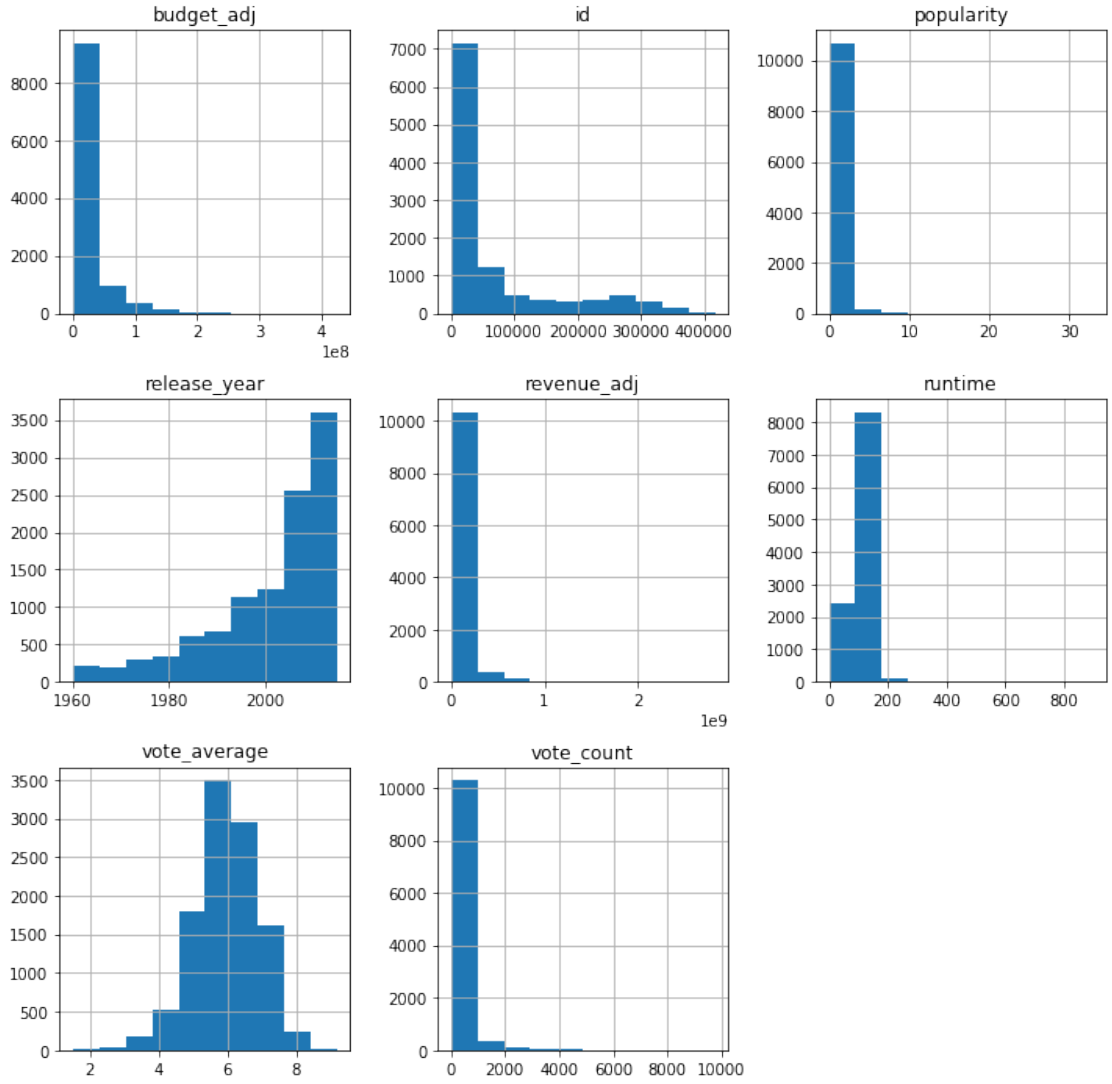
```
In [8]: df.drop(['homepage', 'tagline', 'keywords', 'production_companies'], axis = 1, inplace =
        df.head()

Out[8]:       id    imdb_id  popularity               original_title  \
0  135397  tt0369610   32.985763               Jurassic World
1   76341  tt1392190   28.419936           Mad Max: Fury Road
2  262500  tt2908446   13.112507                     Insurgent
3  140607  tt2488496   11.173104  Star Wars: The Force Awakens
4  168259  tt2820852    9.335014                     Furious 7

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

                                            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 release_date  vote_count  \
0  Action|Adventure|Science Fiction|Thriller       6/9/15        5562
1  Action|Adventure|Science Fiction|Thriller      5/13/15        6185
2         Adventure|Science Fiction|Thriller      3/18/15        2480
3   Action|Adventure|Science Fiction|Fantasy     12/15/15        5292
4                     Action|Crime|Thriller       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
```

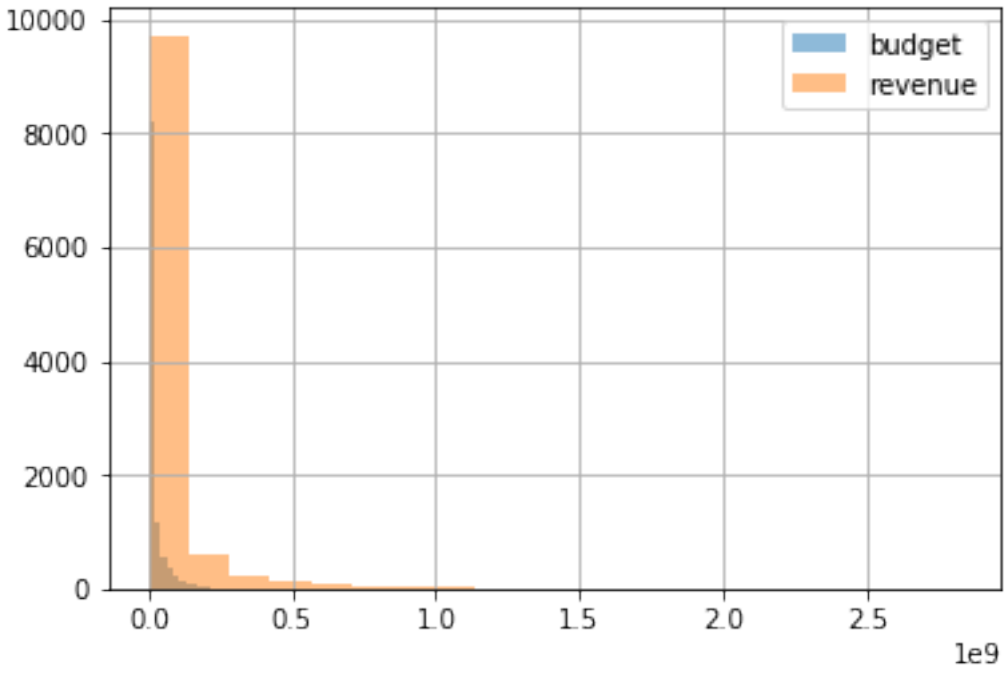
In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 15 columns):
id                10866 non-null int64
imdb_id           10856 non-null object
popularity        10866 non-null float64
original_title    10866 non-null object
cast              10790 non-null object
director          10822 non-null object
overview          10862 non-null object
runtime           10866 non-null int64
genres            10843 non-null object
release_date      10866 non-null object
vote_count        10866 non-null int64
vote_average      10866 non-null float64
release_year      10866 non-null int64
budget_adj        10866 non-null float64
revenue_adj       10866 non-null float64
dtypes: float64(4), int64(4), object(7)
memory usage: 1.2+ MB
```

In [10]:
```
#looked up the histogram of the whole data frame
df.hist(figsize=(12,12));
```




```
In [11]: df.budget_adj.hist(alpha=0.5, bins=20, label='budget')
         df.revenue_adj.hist(alpha=0.5, bins=20, label='revenue')
         plt.legend();
```



Exploratory Data Analysis

1.1.5 Research Question 1 - Which genres are highly rated?

```
In [12]: df.head()
```

Out[12]:

	id	imdb_id	popularity	original_title	cast	director	overview	runtime
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	Twenty-two years after the events of Jurassic ...	124
1	76341	tt1392190	28.419936	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	An apocalyptic story set in the furthest reach...	120
2	262500	tt2908446	13.112507	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	Beatrice Prior must confront her inner demons ...	119
3	140607	tt2488496	11.173104	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Thirty years after defeating the Galactic Empi...	136
4	168259	tt2820852	9.335014	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	Deckard Shaw seeks revenge against Dominic Tor...	137

```
                                              genres release_date  vote_count  \
0  Action|Adventure|Science Fiction|Thriller       6/9/15        5562
1  Action|Adventure|Science Fiction|Thriller      5/13/15        6185
2         Adventure|Science Fiction|Thriller      3/18/15        2480
3   Action|Adventure|Science Fiction|Fantasy     12/15/15        5292
4                       Action|Crime|Thriller       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
```

To Analyze based on movie genre, first step is to split the genre into singular genres for each movie.

```
In [13]: splitted_genre = df.assign(genres=df['genres'].str.split('|')).explode('genres')
```

Then sort the counts in ascending order

```
In [14]: splitted_genre.groupby('genres').count().id.sort_values()
```

```
Out[14]: genres
         Western             165
         TV Movie            167
         Foreign             188
         War                 270
         History             334
         Music               408
         Documentary         520
         Animation           699
         Mystery             810
         Fantasy             916
         Science Fiction    1230
         Family             1231
         Crime              1355
         Adventure          1471
         Horror             1637
         Romance            1712
         Action             2385
         Thriller           2908
         Comedy             3793
         Drama              4761
         Name: id, dtype: int64
```

This shows that Drama, followed by commedy is the highest type of genre although not exclusive.

```
In [15]: splitted_genre.groupby('genres').mean().vote_average.sort_values()

Out[15]: genres
         Horror             5.337447
         Science Fiction    5.665041
         Thriller           5.750413
         Action             5.787421
         TV Movie           5.788024
         Fantasy            5.863537
         Comedy             5.905167
         Adventure          5.940585
         Mystery            5.946790
         Foreign            5.981383
         Family             5.997563
         Romance            6.042874
         Western            6.083030
         Crime              6.124059
         Drama              6.165301
         War                6.297778
         Animation          6.403147
         History            6.410479
         Music              6.480392
         Documentary        6.908462
         Name: vote_average, dtype: float64
```

To answer the question what genre are highly rated I found the mean count of the vote_average collumn, and it returned the data above.

According to the data, the following are the top rated genres

A. Documentary
B. Music
C. History
D. Animation
E. War

1.1.6 Research Question 2 - Does movie budget influence movie rating or revenue?

Does budget spent on movie production influence the revenue?

```
In [16]: # to have an overview of how many rows have zero value in the budget column
         df.query('budget_adj==0')

Out[16]:            id    imdb_id  popularity                  original_title  \
         30     280996  tt3168230    3.927333                      Mr. Holmes
         36     339527  tt1291570    3.358321                          Solace
         72     284289  tt2911668    2.272044                Beyond the Reach
         74     347096  tt3478232    2.165433          Mythica: The Darkspore
         75     308369  tt2582496    2.141506  Me and Earl and the Dying Girl
         ...       ...        ...         ...                             ...
         10860    5060  tt0060214    0.087034             Carry On Screaming!
```

```
10861      21  tt0060371    0.080598              The Endless Summer
10862   20379  tt0060472    0.065543                      Grand Prix
10863   39768  tt0060161    0.065141            Beregis Avtomobilya
10864   21449  tt0061177    0.064317         What's Up, Tiger Lily?

                                                      cast  \
30     Ian McKellen|Milo Parker|Laura Linney|Hattie M...
36     Abbie Cornish|Jeffrey Dean Morgan|Colin Farrel...
72     Michael Douglas|Jeremy Irvine|Hanna Mangan Law...
74     Melanie Stone|Kevin Sorbo|Adam Johnson|Jake St...
75     Thomas Mann|RJ Cyler|Olivia Cooke|Connie Britt...
...                                                  ...
10860  Kenneth Williams|Jim Dale|Harry H. Corbett|Joa...
10861  Michael Hynson|Robert August|Lord 'Tally Ho' B...
10862  James Garner|Eva Marie Saint|Yves Montand|Tosh...
10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...
10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...

                    director  \
30                Bill Condon
36               Afonso Poyart
72     Jean-Baptiste LÃľonetti
74               Anne K. Black
75           Alfonso Gomez-Rejon
...                          ...
10860            Gerald Thomas
10861              Bruce Brown
10862       John Frankenheimer
10863           Eldar Ryazanov
10864              Woody Allen

                                                 overview  runtime  \
30     The story is set in 1947, following a long-ret...      103
36     A psychic doctor, John Clancy, works with an F...      101
72     A high-rolling corporate shark and his impover...       95
74     When Teelaâs sister is murdered and a powerf...      108
75     Greg is coasting through senior year of high s...      105
...                                                  ...      ...
10860  The sinister Dr Watt has an evil scheme going...       87
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

                       genres release_date  vote_count  vote_average  \
30               Mystery|Drama      6/19/15         425           6.4
36         Crime|Drama|Mystery       9/3/15         474           6.2
72                    Thriller      4/17/15          81           5.5
```

```
74      Action|Adventure|Fantasy      6/24/15          27           5.1
75                  Comedy|Drama      6/12/15         569           7.7
...                           ...          ...         ...           ...
10860                     Comedy      5/20/66          13           7.0
10861                Documentary      6/15/66          11           7.4
10862     Action|Adventure|Drama     12/21/66          20           5.7
10863             Mystery|Comedy       1/1/66          11           6.5
10864              Action|Comedy      11/2/66          22           5.4

       release_year  budget_adj   revenue_adj
30             2015         0.0  2.700677e+07
36             2015         0.0  2.056620e+07
72             2015         0.0  4.222338e+04
74             2015         0.0  0.000000e+00
75             2015         0.0  0.000000e+00
...             ...         ...           ...
10860          1966         0.0  0.000000e+00
10861          1966         0.0  0.000000e+00
10862          1966         0.0  0.000000e+00
10863          1966         0.0  0.000000e+00
10864          1966         0.0  0.000000e+00

[5696 rows x 15 columns]
```

In [17]:
```
# to have an overview of how many rows have zero value in the budget column
df.query('revenue_adj==0')
```

Out[17]:
```
            id    imdb_id  popularity                 original_title  \
48      265208  tt2231253    2.932340                      Wild Card
67      334074  tt3247714    2.331636                       Survivor
74      347096  tt3478232    2.165433        Mythica: The Darkspore
75      308369  tt2582496    2.141506  Me and Earl and the Dying Girl
92      370687  tt3608646    1.876037      Mythica: The Necromancer
...        ...        ...         ...                           ...
10861       21  tt0060371    0.080598            The Endless Summer
10862    20379  tt0060472    0.065543                     Grand Prix
10863    39768  tt0060161    0.065141          Beregis Avtomobilya
10864    21449  tt0061177    0.064317        What's Up, Tiger Lily?
10865    22293  tt0060666    0.035919      Manos: The Hands of Fate

                                                    cast             director  \
48     Jason Statham|Michael Angarano|Milo Ventimigli...          Simon West
67     Pierce Brosnan|Milla Jovovich|Dylan McDermott|...       James McTeigue
74     Melanie Stone|Kevin Sorbo|Adam Johnson|Jake St...        Anne K. Black
75     Thomas Mann|RJ Cyler|Olivia Cooke|Connie Britt...  Alfonso Gomez-Rejon
92     Melanie Stone|Adam Johnson|Kevin Sorbo|Nicola ...       A. Todd Smith
...                                                  ...                 ...
10861  Michael Hynson|Robert August|Lord 'Tally Ho' B...         Bruce Brown
```

```
10862  James Garner|Eva Marie Saint|Yves Montand|Tosh...   John Frankenheimer
10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...       Eldar Ryazanov
10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...          Woody Allen
10865  Harold P. Warren|Tom Neyman|John Reynolds|Dian...     Harold P. Warren

                                                 overview  runtime  \
48     When a Las Vegas bodyguard with lethal skills ...        92
67     A Foreign Service Officer in London tries to p...        96
74     When Teelaâs sister is murdered and a powerf...       108
75     Greg is coasting through senior year of high s...       105
92     Mallister takes Thane prisoner and forces Mare...         0
...                                                  ...       ...
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 release_date  vote_count  vote_average  \
48         Thriller|Crime|Drama      1/14/15         481           5.3
67        Crime|Thriller|Action      5/21/15         280           5.4
74     Action|Adventure|Fantasy      6/24/15          27           5.1
75                 Comedy|Drama      6/12/15         569           7.7
92     Fantasy|Action|Adventure     12/19/15          11           5.4
...                         ...          ...         ...           ...
10861              Documentary      6/15/66          11           7.4
10862   Action|Adventure|Drama     12/21/66          20           5.7
10863            Mystery|Comedy       1/1/66          11           6.5
10864             Action|Comedy      11/2/66          22           5.4
10865                    Horror     11/15/66          15           1.5

       release_year    budget_adj  revenue_adj
48             2015  2.759999e+07          0.0
67             2015  1.839999e+07          0.0
74             2015  0.000000e+00          0.0
75             2015  0.000000e+00          0.0
92             2015  0.000000e+00          0.0
...             ...           ...          ...
10861          1966  0.000000e+00          0.0
10862          1966  0.000000e+00          0.0
10863          1966  0.000000e+00          0.0
10864          1966  0.000000e+00          0.0
10865          1966  1.276423e+05          0.0

[6016 rows x 15 columns]
```

The data above returns 5696 rows with zero value in budget, this is an area to investigate further as to why and how a movie budget is zero. But for the sake of this analysis i will continue

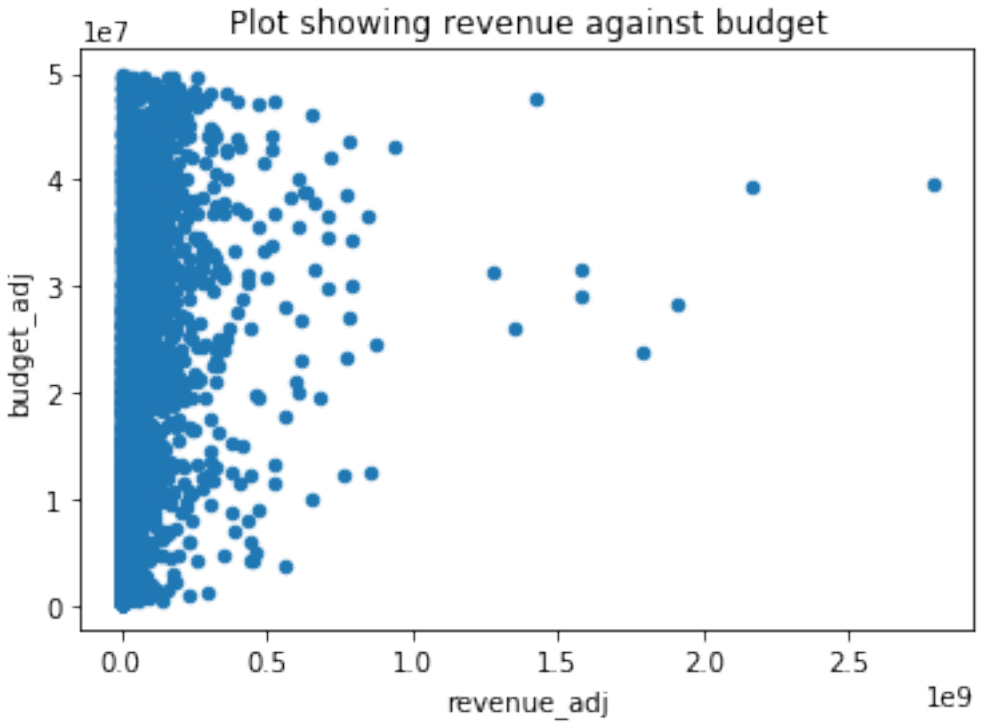
my analysis without these columns.

In order to plot the relationship between movie budget and revenue, I excluded the rows with budget below 50,000 to avoid outliers and to create a mederate sample size.

```
In [18]: # Created a function to plot the result of queries, to avoid repetitive code.

         def query_plot(query,x,y):
             df.query(query).plot.scatter(x = x , y = y)
```

```
In [19]: query_plot('budget_adj > 20_0000 & budget_adj < 50_000_000', x = 'revenue_adj', y = 'bu
         plt.title('Plot showing revenue against budget')
```

```
Out[19]: Text(0.5,1,'Plot showing revenue against budget')
```



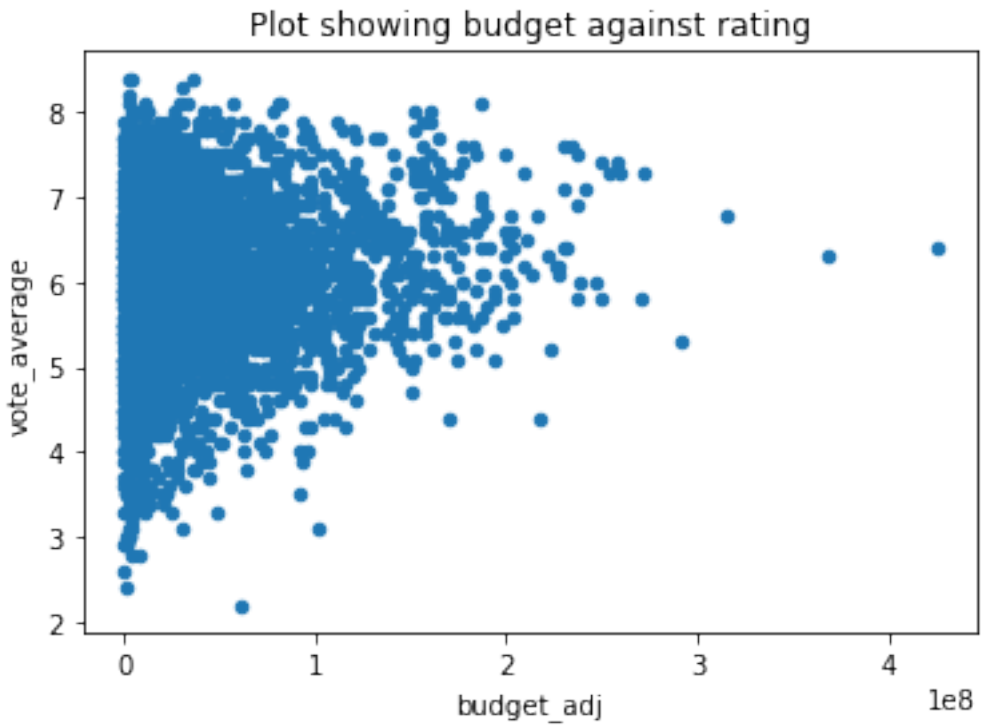
I observed there is no strong correlation between revenue and budget according to the scatter-plot.

QUESTION:

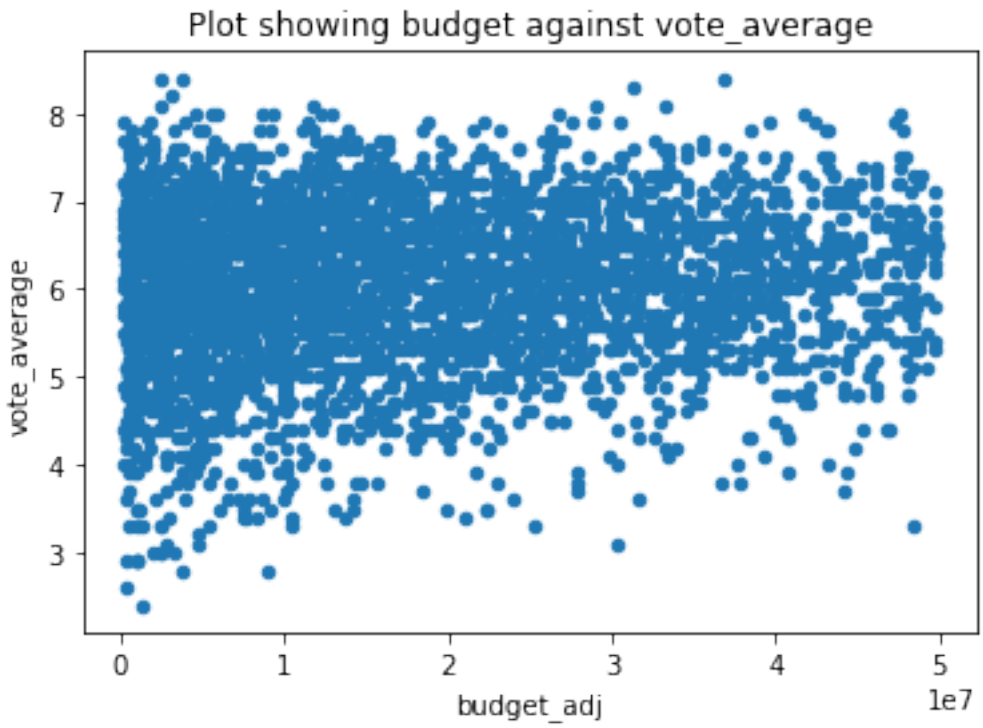
Does budget spent on movie production influence the rating.

```
In [20]: query_plot('budget_adj > 200000', x = 'budget_adj', y = 'vote_average')
         plt.title('Plot showing budget against rating')
```

```
Out[20]: Text(0.5,1,'Plot showing budget against rating')
```

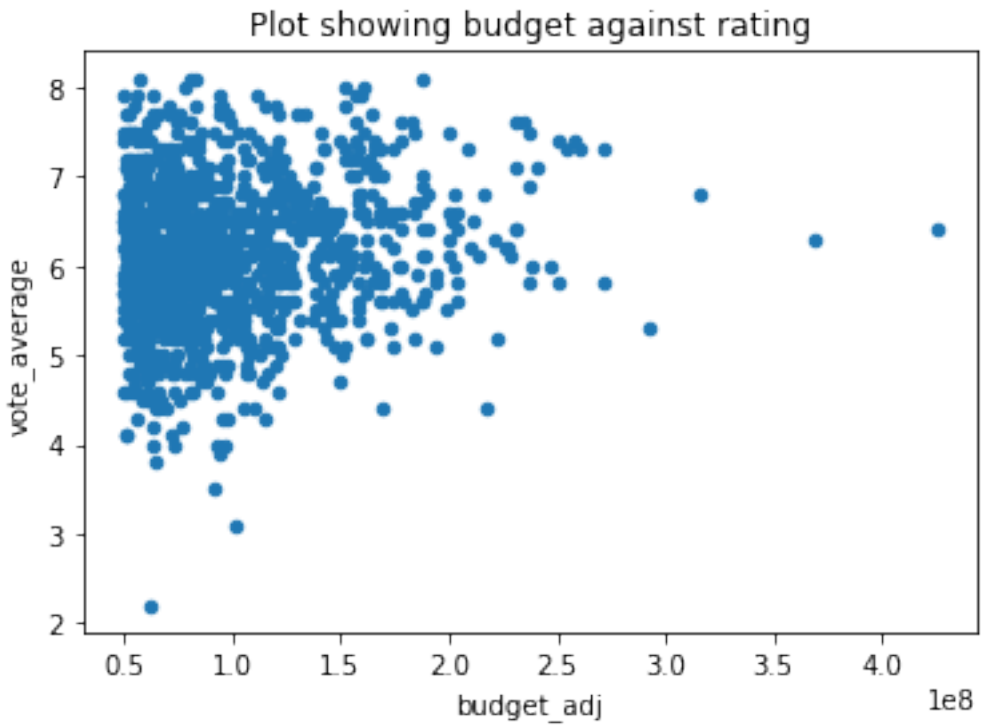



```
In [21]: query_plot('budget_adj > 20_0000 & budget_adj < 50_000_000', x = 'budget_adj', y = 'vot
         plt.title('Plot showing budget against vote_average')

Out[21]: Text(0.5,1,'Plot showing budget against vote_average')
```




```
In [22]: query_plot('budget_adj > 50_000_000', x = 'budget_adj', y = 'vote_average')
         plt.title('Plot showing budget against rating')
```

```
Out[22]: Text(0.5,1,'Plot showing budget against rating')
```



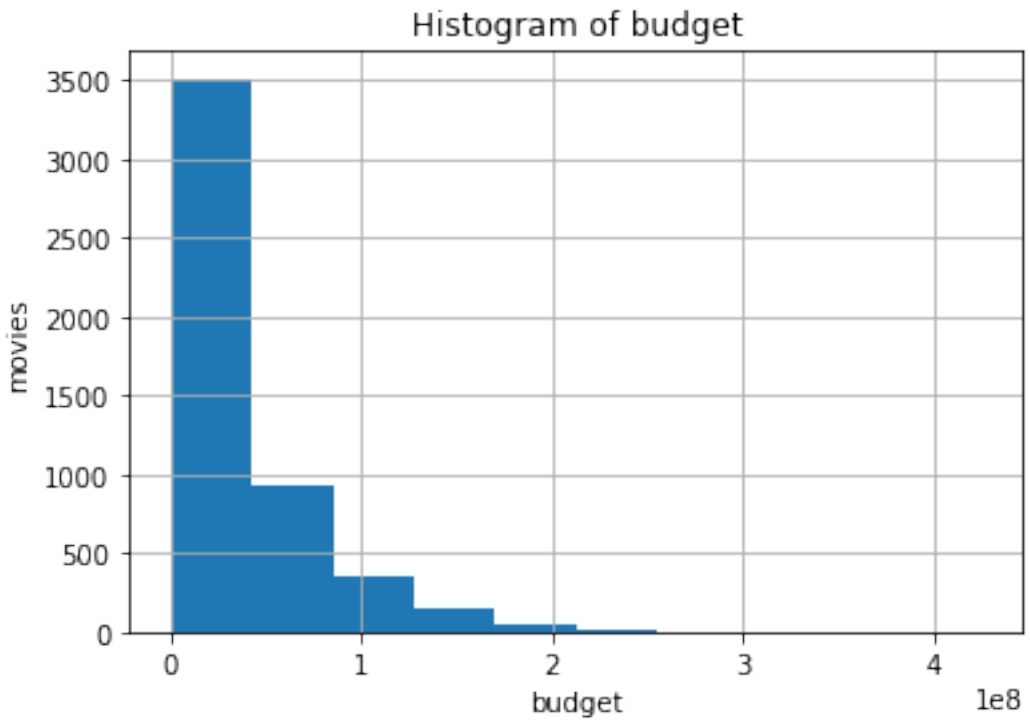

```
In [23]: df.query('budget_adj > 200000').budget_adj.describe()
```

```
Out[23]: count    5.034000e+03
         mean     3.788305e+07
         std      4.207485e+07
         min      2.025573e+05
         25%      9.150250e+06
         50%      2.374361e+07
         75%      5.078480e+07
         max      4.250000e+08
         Name: budget_adj, dtype: float64
```

According to the summary statistics, 50% of the movie have budget below 50Million.

```
In [24]: df.query('budget_adj > 200000').budget_adj.hist()
         plt.title('Histogram of budget')
         plt.xlabel("budget")
         plt.ylabel("movies")
```

```
Out[24]: Text(0,0.5,'movies')
```



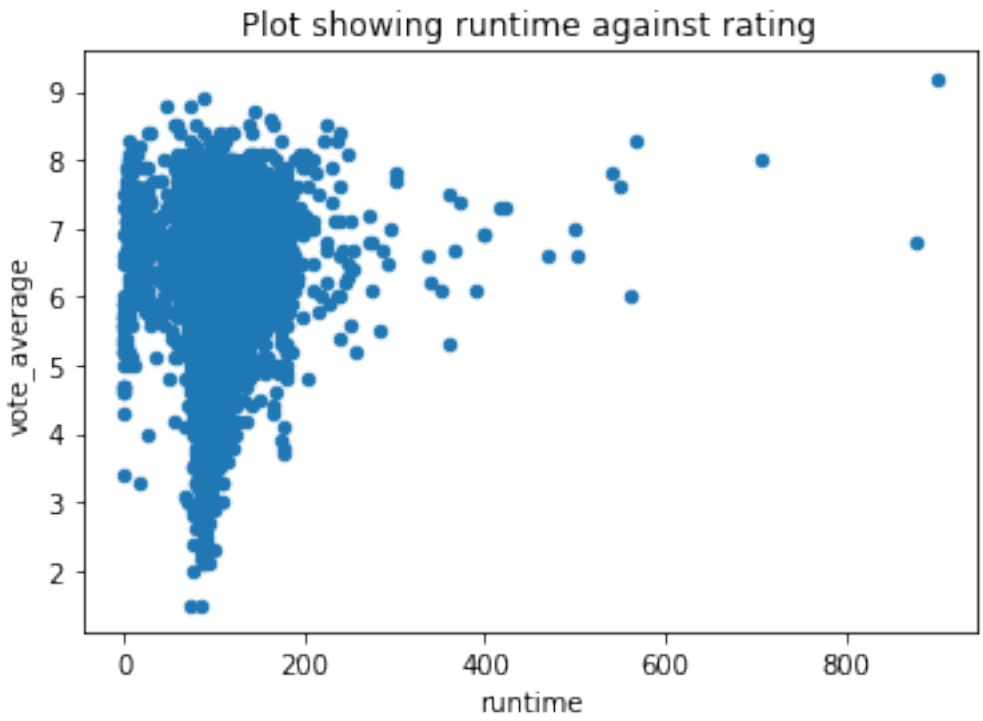
It is logical to assume that the higer the budget spent, the higher the movie rating, but the scatterplot doesnt support that assumption, instead it shows that there is no strong correlation between the movie budget and the movie rating.

1.1.7 Research Question 3 - What kinds of properties are associated with movies that have high rating?

A scatterplot to compare runtime and rating.

```
In [25]: df.plot.scatter(x = 'runtime', y = 'vote_average')
         plt.title('Plot showing runtime against rating')
```

```
Out[25]: Text(0.5,1,'Plot showing runtime against rating')
```




```
In [26]: df[df.runtime==df.runtime.max()]
```

```
Out[26]:          id    imdb_id  popularity                original_title  \
3894  125336  tt2044056    0.006925  The Story of Film: An Odyssey

                                                    cast      director  \
3894  Mark Cousins|Jean-Michel Frodon|Cari Beauchamp...  Mark Cousins

                                               overview  runtime       genres  \
3894  The Story of Film: An Odyssey, written and dir...      900  Documentary

     release_date  vote_count  vote_average  release_year  budget_adj  \
3894       9/3/11          14           9.2          2011         0.0

      revenue_adj
3894          0.0
```

1) I noticed that the movie with the highest runtime has the highest rating - (The Story of Film: An Odyssey)

2) But looking at the result of the scattered plot holistically, it's obvious that the run time does'nt necesarily influence the movie rating, so it might be safe to say that the movie(The Story of Film: An Odyssey) is am outlier.

```
In [27]: df.head()
```

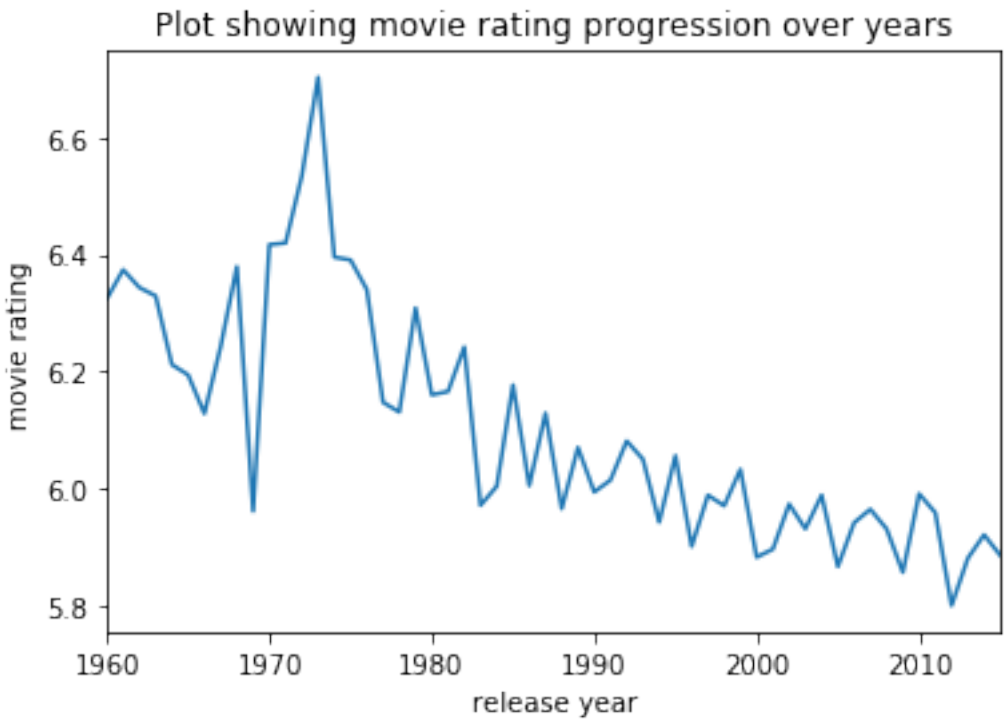
Out[27]:

	id	imdb_id	popularity	original_title	cast	director	overview	runtime	genres	release_date	vote_count	vote_average	release_year	budget_adj	revenue_adj
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	Twenty-two years after the events of Jurassic ...	124	Action|Adventure|Science Fiction|Thriller	6/9/15	5562	6.5	2015	1.379999e+08	1.392446e+09
1	76341	tt1392190	28.419936	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	An apocalyptic story set in the furthest reach...	120	Action|Adventure|Science Fiction|Thriller	5/13/15	6185	7.1	2015	1.379999e+08	3.481613e+08
2	262500	tt2908446	13.112507	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	Beatrice Prior must confront her inner demons ...	119	Adventure|Science Fiction|Thriller	3/18/15	2480	6.3	2015	1.012000e+08	2.716190e+08
3	140607	tt2488496	11.173104	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Thirty years after defeating the Galactic Empi...	136	Action|Adventure|Science Fiction|Fantasy	12/15/15	5292	7.5	2015	1.839999e+08	1.902723e+09
4	168259	tt2820852	9.335014	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	Deckard Shaw seeks revenge against Dominic Tor...	137	Action|Crime|Thriller	4/1/15	2947	7.3	2015	1.747999e+08	1.385749e+09

In [28]:
```
Year_average = df.groupby('release_year').mean()
```

In [29]:
```
Year_average['vote_average'].plot()
plt.title('Plot showing movie rating progression over years')
plt.xlabel("release year")
plt.ylabel("movie rating")
```

Out[29]: Text(0,0.5,'movie rating')



I noticed that the movie rating reduces over the years.

Conclusions

--- Research Question 1

Which genres are the highest rated?

It was observed that the following five movie genre had the highest rating in these order A. Documentary B. Music C. History D. Animation E. War

--- Research Question 2

Does movie budget influence movie rating or revenue?

It may be logical to assumed that the higher the budget spent, the higher the movie rating, but the analysis does'nt support such assumption, instead it exposes that there is no strong correlation between the movie budget and the movie rating.

Also there the analysis exposes that the movie budget doesnt influence the revenue.

--- Research Question 3

WWhat kinds of properties are associated with movies that have high rating?

A. The genre of the movie is a major property that influeces the movie rating.

B. The analysis exposed that the run time does not influence the movie rating.

1.2 Project Limitation

After querying the data, it was noticed that 5696 rows with zero value in the budget was returned and 6016 rows with zero value in revenue was returned, this is an area to investigate further as to why and how a movie budget/revenue is recorded as zero.

```
In [30]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[30]: 0
```