Investigate_a_Dataset

February 3, 2021

1 Project: Investigate a Dataset (Replace this with something more specific!)

1.1 Table of Contents

Introduction

1.2 Questions:

```
<li>What is the perfect time of a movie ?</li>
<li>What genres are most popular?</li>
<li>Which actors appeared the most?</li>
<li>Charcteristics Associated with Successful Movies</li>
<li>Charcteristics Associated with unsuccessful Movies</li>
```

```
In [1]: # importing libraries
        import pandas as pd
        import numpy as np
        %matplotlib inline
        import matplotlib.pyplot as plt
```

Data Wrangling

```
In [2]: # loading data
        df = pd.read_csv('tmdb-movies.csv')
        df.head()
```

```
Out[2]:        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 [3]: # some 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
```

> **Tip**: You should *not* perform too many operations in each cell. Create cells freely to explore your data. One option that you can take with this project is to do a lot of explorations in an initial notebook. These don't have to be organized, but make sure you use enough comments to understand the purpose of each code cell. Then, after you're done with your analysis, create a duplicate notebook where you will trim the excess and organize your steps so that you have a flowing, cohesive report.

Tip: Make sure that you keep your reader informed on the steps that you are taking in your investigation. Follow every code cell, or every set of related code cells, with a markdown cell to describe to the reader what was found in the preceding cell(s). Try to make it so that the reader can then understand what they will be seeing in the following cell(s).

1.2.1 Data Cleaning (Replace this with more specific notes!)

```
In [4]: df.isna().head()
```

```
Out[4]:       id  imdb_id  popularity  budget  revenue  original_title   cast  \
        0  False    False       False   False    False           False  False
        1  False    False       False   False    False           False  False
        2  False    False       False   False    False           False  False
        3  False    False       False   False    False           False  False
        4  False    False       False   False    False           False  False

           homepage  director  tagline    ...      overview  runtime  genres  \
        0     False     False    False    ...         False    False   False
        1     False     False    False    ...         False    False   False
        2     False     False    False    ...         False    False   False
        3     False     False    False    ...         False    False   False
        4     False     False    False    ...         False    False   False

           production_companies  release_date  vote_count  vote_average  release_year  \
        0                 False         False       False         False         False
        1                 False         False       False         False         False
        2                 False         False       False         False         False
        3                 False         False       False         False         False
        4                 False         False       False         False         False

           budget_adj  revenue_adj
        0       False        False
        1       False        False
        2       False        False
        3       False        False
        4       False        False

        [5 rows x 21 columns]
```

```
In [5]: # removing nulls
        # df.dropna(inplace=True)
        # removing null values from the data set will affect the data and the results of the ana
        # In addition, nearly all null values are in the homepage column so those movies might d
        # home page.
```

removing null values from the data set will affect the data and the results of the analysis. In addition, nearly all null values are in the homepage column so those movies might did not have home page.

```
In [6]: # dulicated values
        df[df.duplicated() == True]
```

```
Out[6]:          id    imdb_id  popularity    budget  revenue original_title  \
        2090  42194  tt0411951     0.59643  30000000   967000         TEKKEN

                                                           cast homepage  \
        2090  Jon Foo|Kelly Overton|Cary-Hiroyuki Tagawa|Ian...      NaN

                      director               tagline     ...       \
        2090  Dwight H. Little  Survival is no game     ...

                                                       overview runtime  \
        2090  In the year of 2039, after World Wars destroy ...      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      3/20/10        110           5.0          2010  30000000.0

              revenue_adj
        2090     967000.0

        [1 rows x 21 columns]
```

```
In [8]: # there is one duplicated row
        # removing duplicates
        df.drop_duplicates(inplace=True)
```

```
In [9]: # add new column with profit
        df["profit"] = df["revenue"] - df["budget"]
        df.head(1)
```

```
Out[9]:        id    imdb_id  popularity     budget     revenue  original_title  \
        0  135397  tt0369610   32.985763  150000000  1513528810  Jurassic World

                                                         cast  \
        0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...

                                homepage          director            tagline  \
        0  http://www.jurassicworld.com/  Colin Trevorrow  The park is open.

               ...     runtime                                     genres  \
        0     ...         124  Action|Adventure|Science Fiction|Thriller

                                          production_companies release_date vote_count  \
        0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15       5562
```

```
  vote_average  release_year    budget_adj   revenue_adj      profit
0          6.5          2015  1.379999e+08  1.392446e+09  1363528810

[1 rows x 22 columns]
```

```
In [10]: # just take a look on the dtypes
         df.dtypes
```

```
Out[10]: 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             object
         vote_count                int64
         vote_average            float64
         release_year              int64
         budget_adj              float64
         revenue_adj             float64
         profit                    int64
         dtype: object
```

```
In [11]: # converting the data type of release_date column to datetime
         df["release_date"] = df["release_date"].astype("datetime64")
```

```
In [12]: # let's see that it goes well
         df.dtypes
```

```
Out[12]: 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
profit                               int64
dtype: object
```

NOTE: There are some movies with negative profits

Exploratory Data Analysis

2 Q1: What is the perfect time of a movie ?

```
In [13]: # getting over-average profits
         high_profit = df[df["profit"] > df["profit"].mean()]
```

```
In [14]: df["runtime"].describe()
```

```
Out[14]: count    10865.000000
         mean       102.071790
         std         31.382701
         min          0.000000
         25%         90.000000
         50%         99.000000
         75%        111.000000
         max        900.000000
         Name: runtime, dtype: float64
```

```
In [15]: # getting the average runtime of top movies
         average_perfect_time = high_profit["runtime"].mean()
         average_perfect_time
```

```
Out[15]: 112.22333000997008
```

```
In [48]: #plotting runtime and profit to detect the relqtionship
         df.plot.scatter(x="runtime", y="profit")
         df.plot.xlabel = "runtime"
         df.plot.ylabel = "profit"
         df.plot.title = "The relatirela
```

```
  File "<ipython-input-48-56bf835f1fe2>", line 5
df.plot.title = "The relatirela
```

```
                                       ^
    SyntaxError: EOL while scanning string literal
```

```
In [138]: # getting popular movies
          high_pop = df[df["popularity"] > df["popularity"].mean()]
          high_pop.head()
```

```
Out[138]:        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     ...      runtime  \
          0              The park is open.     ...          124
          1             What a Lovely Day.     ...          120
          2     One Choice Can Destroy You     ...          119
          3  Every generation has a story.     ...          136
          4            Vengeance Hits Home     ...          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...   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      profit
0           6.5          2015  1.379999e+08  1.392446e+09  1363528810
1           7.1          2015  1.379999e+08  3.481613e+08   228436354
2           6.3          2015  1.012000e+08  2.716190e+08   185238201
3           7.5          2015  1.839999e+08  1.902723e+09  1868178225
4           7.3          2015  1.747999e+08  1.385749e+09  1316249360

[5 rows x 22 columns]
```

```
In [139]: high_pop["runtime"].mean()
```

```
Out[139]: 107.80359477124183
```

3 Conclusion1

So the average runtime of the most successful movies is between 107 and 113 minutes although the best movie (in profits) is avatar whose runtime is 162 but it is not standard. Personally, I get bored with movies that are longer than 2 hours. It seems that there is not a clear relationship between runtime and profit but we can see that most runtimes are less than 200 minutes.

3.1 Q2: What genres are most popular?

```
In [25]: # get movies with high profits
         high_profit = df[df["profit"] > df["profit"].mean()]
```

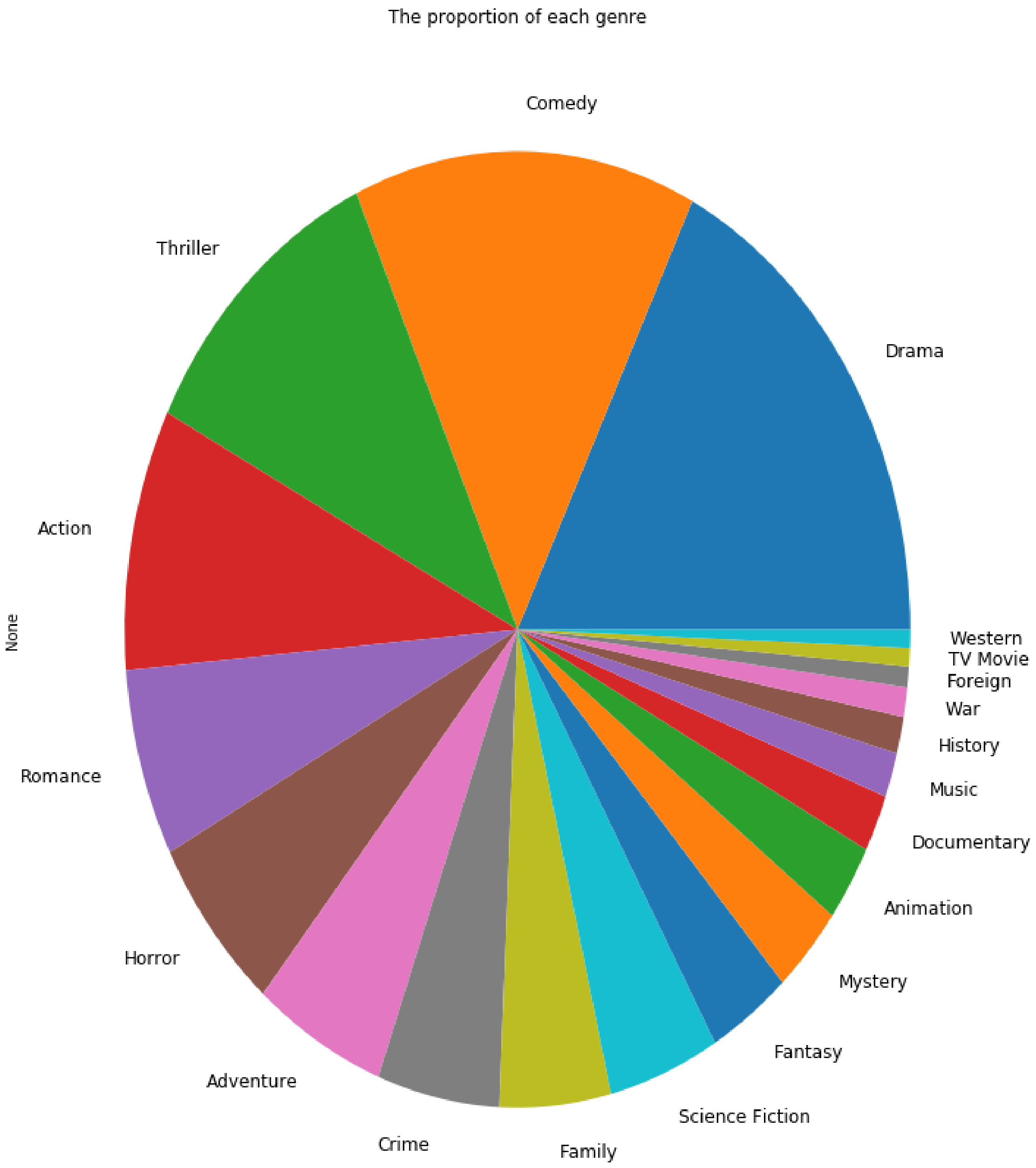
```
In [26]: def extractCats(df, col):
             """This function extracts substrings seperated by a specidic character.
                df = dataFrame
                var = column name that you want to extract data from
             """
             all_cats = df[col].str.cat(sep="|")
             all_cats = pd.Series(all_cats.split("|"))
             count = all_cats.value_counts()
             return count
```

```
In [27]: count = extractCats(df, "genres")
         count
```

```
Out[27]: Drama              4760
         Comedy             3793
         Thriller           2907
         Action             2384
         Romance            1712
         Horror             1637
         Adventure          1471
         Crime              1354
         Family             1231
         Science Fiction    1229
         Fantasy             916
         Mystery             810
         Animation           699
         Documentary         520
         Music               408
         History             334
         War                 270
         Foreign             188
         TV Movie            167
         Western             165
         dtype: int64
```

```
In [45]: count.plot(kind="pie", fontsize = 12, figsize=(12, 15))
         plt.title("The proportion of each genre")
```

```
Out[45]: Text(0.5,1,'The proportion of each genre')
```



So, We can see that the most common genre is Drama then Comedy while western was the least one

```
In [29]: # getting high profit movies
         high_profit = df[df["profit"] > df["profit"].mean()]
         high_profit.head()
```

```
Out[29]:        id    imdb_id  popularity     budget     revenue  \
         0  135397  tt0369610   32.985763  150000000  1513528810
```

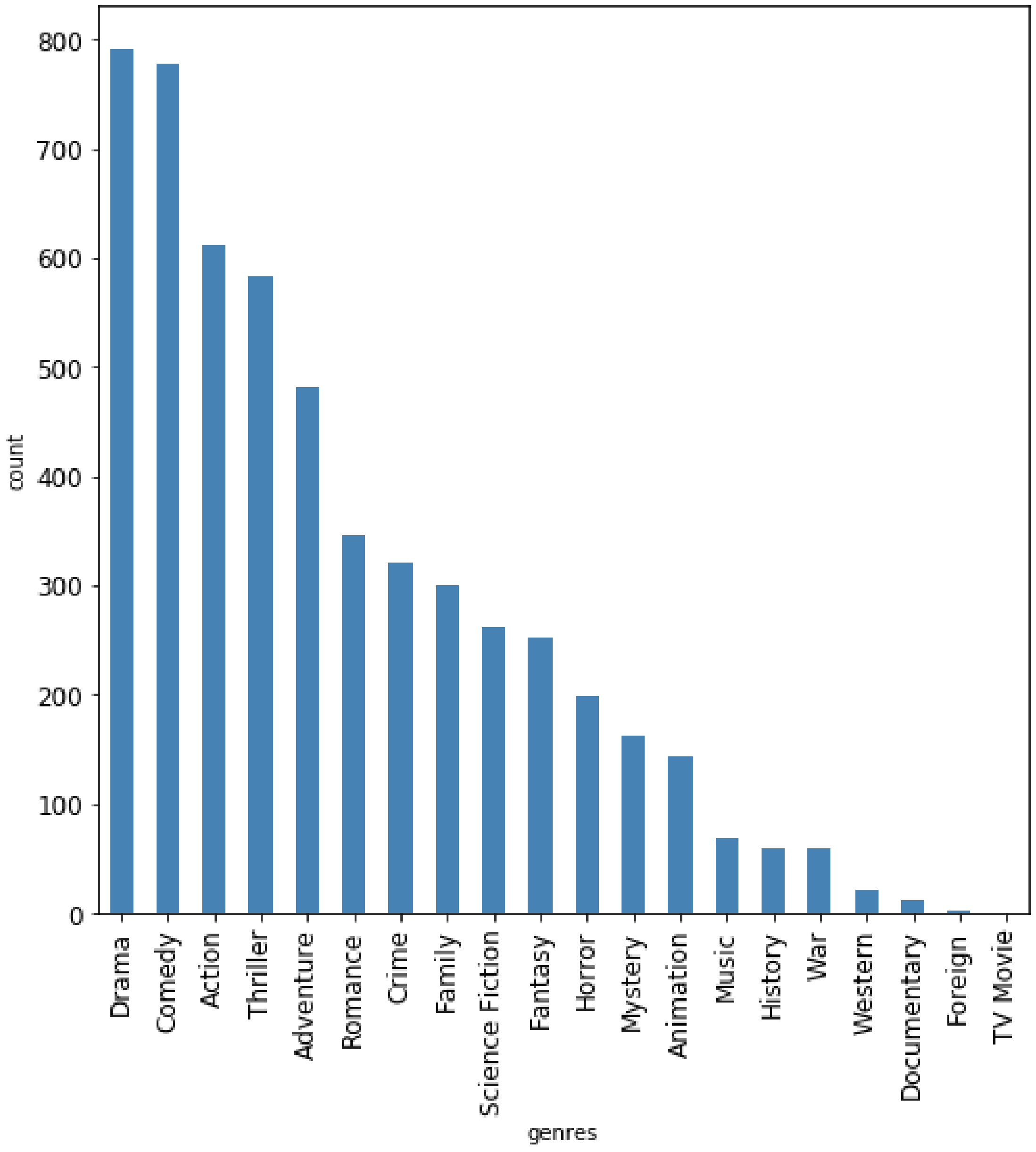
```
1   76341  tt1392190   28.419936  150000000   378436354
2  262500  tt2908446   13.112507  110000000   295238201
3  140607  tt2488496   11.173104  200000000  2068178225
4  168259  tt2820852    9.335014  190000000  1506249360

                 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     ...     runtime  \
0              The park is open.     ...         124
1             What a Lovely Day.     ...         120
2     One Choice Can Destroy You     ...         119
3  Every generation has a story.     ...         136
4            Vengeance Hits Home     ...         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...   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      profit
```

```
0           6.5          2015  1.379999e+08  1.392446e+09  1363528810
1           7.1          2015  1.379999e+08  3.481613e+08   228436354
2           6.3          2015  1.012000e+08  2.716190e+08   185238201
3           7.5          2015  1.839999e+08  1.902723e+09  1868178225
4           7.3          2015  1.747999e+08  1.385749e+09  1316249360

[5 rows x 22 columns]
```

```
In [30]: # Extracting genres by calling extractCts
         high_profit_count = extractCats(high_profit, "genres")
         high_profit_count
```

```
Out[30]: Drama              791
         Comedy             777
         Action             611
         Thriller           582
         Adventure          482
         Romance            345
         Crime              320
         Family             300
         Science Fiction    261
         Fantasy            253
         Horror             198
         Mystery            162
         Animation          144
         Music               68
         History             60
         War                 59
         Western             22
         Documentary         11
         Foreign              2
         TV Movie             1
         dtype: int64
```

```
In [44]: # bar chart that shows each genre with its count ( genres of top movies of n profit)
         high_profit_count.plot.bar(color="steelblue", fontsize = 12, figsize=(8, 8))
         plt.xlabel("genres")
         plt.ylabel("count")
         plt.show()
```



```
In [32]: lowest_profit = df[df["profit"] <= 0]
         lowest_profit.head()
```

```
Out[32]:        id    imdb_id  popularity     budget   revenue  \
         48  265208  tt2231253    2.932340   30000000         0
         57  210860  tt3045616    2.575711   60000000  30418560
         59  201088  tt2717822    2.550747   70000000  17752940
         66  205775  tt1390411    2.345821  100000000  93820758
         67  334074  tt3247714    2.331636   20000000         0
```

```
            original_title  \
48               Wild Card   
57               Mortdecai   
59                Blackhat   
66  In the Heart of the Sea   
67                Survivor   

                                                 cast  \
48  Jason Statham|Michael Angarano|Milo Ventimigli...   
57  Johnny Depp|Gwyneth Paltrow|Ewan McGregor|Paul...   
59  Chris Hemsworth|Leehom Wang|Tang Wei|Viola Dav...   
66  Chris Hemsworth|Benjamin Walker|Cillian Murphy...   
67  Pierce Brosnan|Milla Jovovich|Dylan McDermott|...   

                                   homepage        director  \
48                                      NaN      Simon West   
57            http://mortdecaithemovie.com/     David Koepp   
59  http://www.legendary.com/film/blackhat/    Michael Mann   
66  http://www.intheheartoftheseamovie.com/      Ron Howard   
67                http://survivormovie.com/  James McTeigue   

                                              tagline    ...    runtime  \
48        Never bet against a man with a killer hand.    ...         92   
57                         Sophistication Has a Name.    ...        106   
59                       We are no longer in control.    ...        133   
66  Based on the incredible true story that inspir...    ...        122   
67                 His Next Target is Now Hunting Him    ...         96   

                                     genres  \
48                     Thriller|Crime|Drama   
57                         Comedy|Adventure   
59      Mystery|Crime|Action|Thriller|Drama   
66  Thriller|Drama|Adventure|Action|History   
67                    Crime|Thriller|Action   

                                 production_companies release_date vote_count  \
48  Current Entertainment|Lionsgate|Sierra / Affin...   2015-01-14        481   
57  Lionsgate|Mad Chance|OddLot Entertainment|Huay...   2015-01-21        696   
59  Universal Pictures|Forward Pass|Legendary Pict...   2015-01-13        584   
66  Imagine Entertainment|Spring Creek Productions...   2015-11-20        805   
67  Nu Image Films|Winkler Films|Millennium Films|...   2015-05-21        280   

   vote_average  release_year    budget_adj   revenue_adj    profit
48          5.3          2015  2.759999e+07  0.000000e+00 -30000000
57          5.3          2015  5.519998e+07  2.798506e+07 -29581440
59          5.0          2015  6.439997e+07  1.633270e+07 -52247060
66          6.4          2015  9.199996e+07  8.631506e+07  -6179242
67          5.4          2015  1.839999e+07  0.000000e+00 -20000000
```

```
         [5 rows x 22 columns]
```

```
In [148]: lowest_profit_genres = extractCats(lowest_profit, "genres")
          lowest_profit_genres
```

```
Out[148]: Drama              3070
          Comedy             2370
          Thriller           1849
          Action             1421
          Horror             1196
          Romance            1028
          Family              803
          Adventure           802
          Science Fiction     780
          Crime               767
          Fantasy             555
          Mystery             522
          Animation           501
          Documentary         427
          Music               264
          History             216
          TV Movie            166
          Foreign             166
          War                 163
          Western             119
          dtype: int64
```

```
In [41]: popular = df[df["popularity"] > df["popularity"].mean()]
         popular.head()
```

```
Out[41]:        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...
```

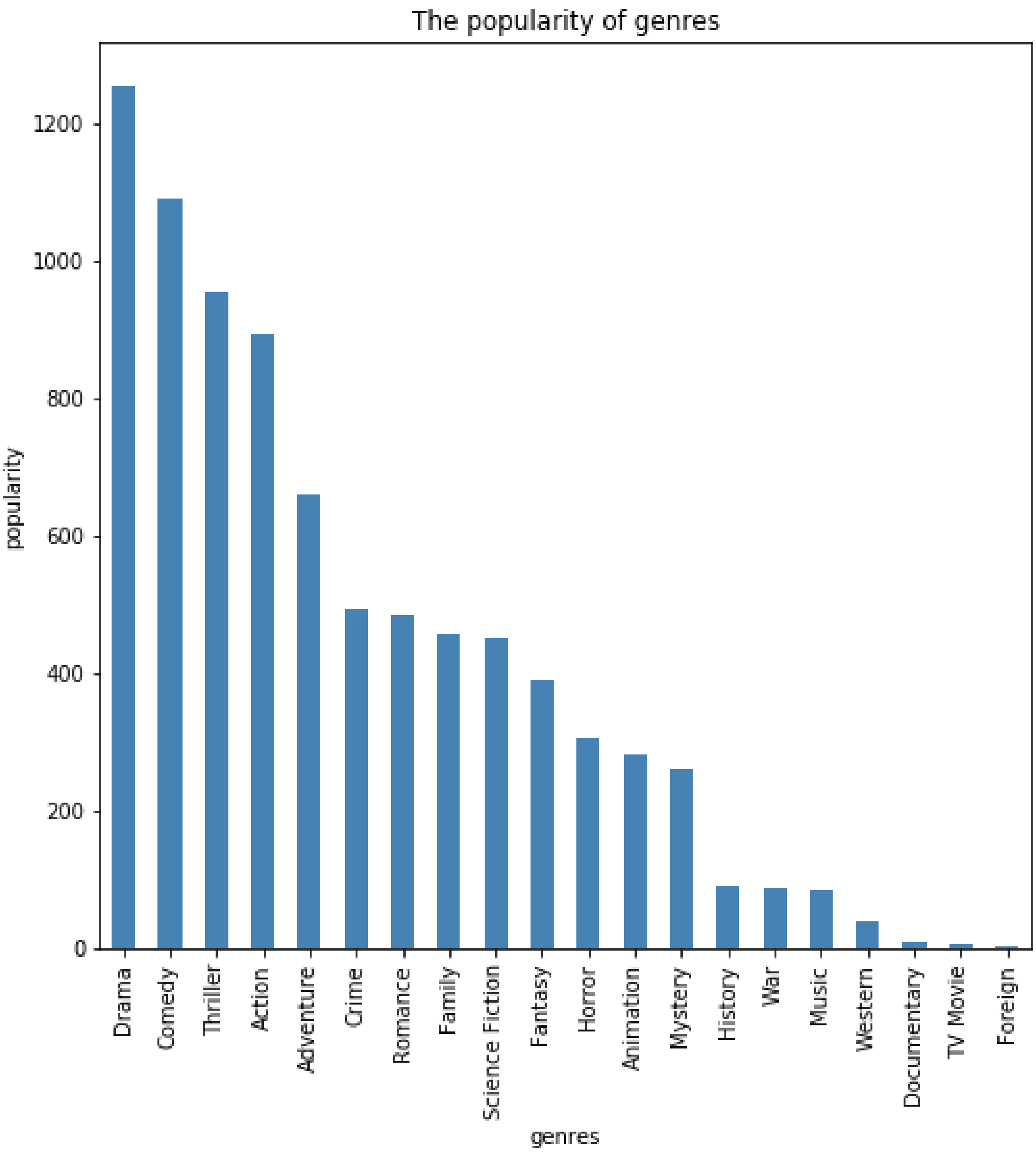
```
3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                            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     ...     runtime  \
0              The park is open.     ...         124
1             What a Lovely Day.     ...         120
2     One Choice Can Destroy You     ...         119
3  Every generation has a story.     ...         136
4            Vengeance Hits Home     ...         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...   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      profit
0          6.5          2015  1.379999e+08  1.392446e+09  1363528810
1          7.1          2015  1.379999e+08  3.481613e+08   228436354
2          6.3          2015  1.012000e+08  2.716190e+08   185238201
3          7.5          2015  1.839999e+08  1.902723e+09  1868178225
4          7.3          2015  1.747999e+08  1.385749e+09  1316249360

[5 rows x 22 columns]
```

```
In [42]: pop_count = extractCats(popular, "genres")
         pop_count
```

```
Out[42]: Drama              1255
         Comedy             1091
         Thriller            952
         Action              892
         Adventure           661
```

```
Crime               494
Romance             485
Family              456
Science Fiction     452
Fantasy             390
Horror              306
Animation           282
Mystery             260
History              90
War                  88
Music                84
Western              39
Documentary           9
TV Movie              6
Foreign               1
dtype: int64
```

```
In [49]: pop_count.plot.bar(figsize=(8, 8), color="steelblue")
         plt.xlabel("genres")
         plt.ylabel("popularity")
         plt.title("The popularity of genres")
         plt.show()
```



4 Conclusion2:

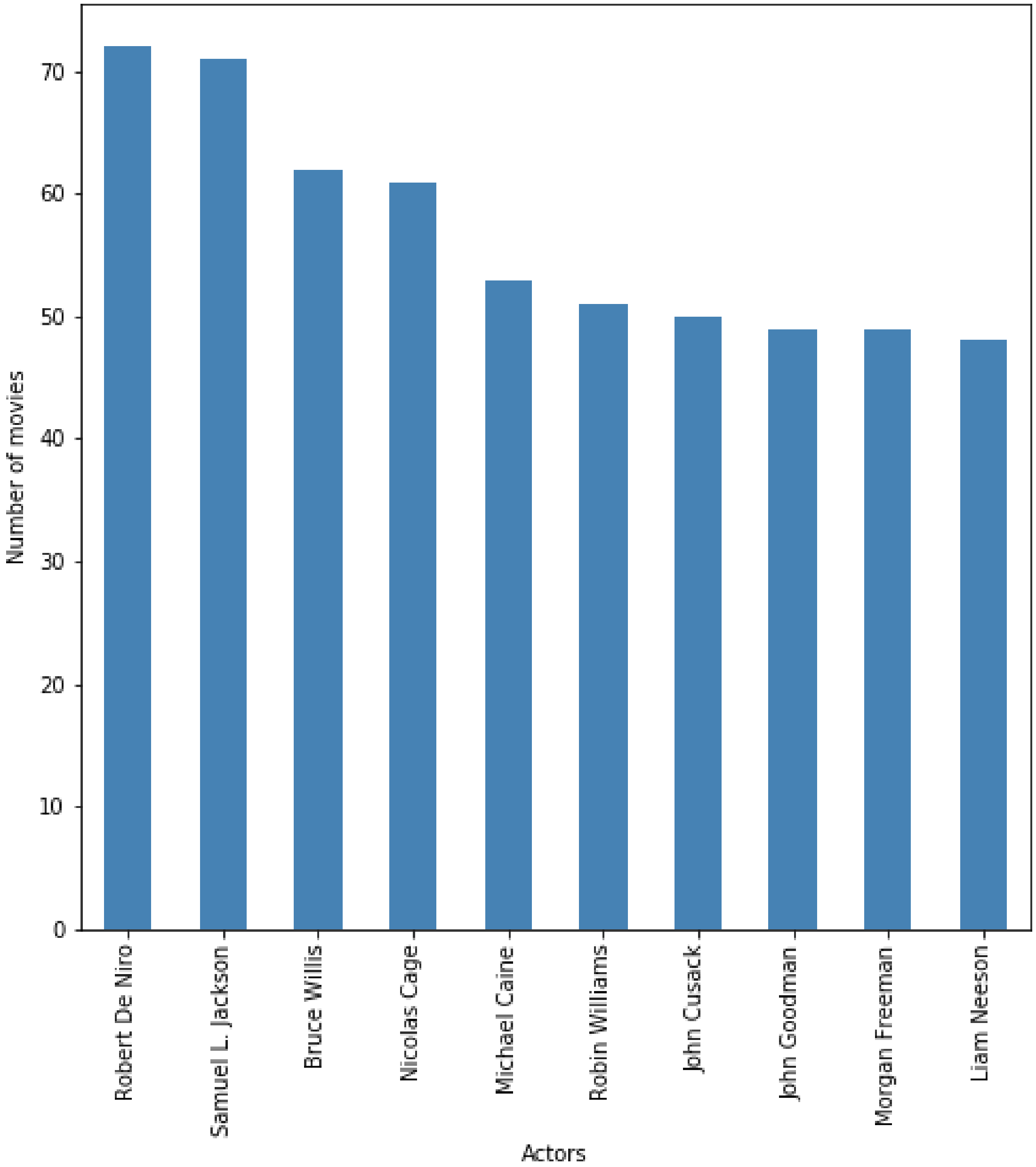
So, The most common genre in general is Drama 4760 while the least one is western and the most genre the made profits is alse Drama while the least one is also Drama. This contrast maybe because the number of Drama movies is high so the possiblity to success and fail is also high. Drama is the best genre in everything.

5 Q3: Which actors appeared the most?

```
In [34]: # Getting actors
         actors_count = extractCats(df, "cast")
         actors_count.head(20)
```

```
Out[34]: Robert De Niro        72
         Samuel L. Jackson     71
         Bruce Willis          62
         Nicolas Cage          61
         Michael Caine         53
         Robin Williams        51
         John Cusack           50
         John Goodman          49
         Morgan Freeman        49
         Liam Neeson           48
         Susan Sarandon        48
         Julianne Moore        47
         Alec Baldwin          47
         Gene Hackman          46
         Johnny Depp           46
         Tom Hanks             46
         Christopher Walken    46
         Sylvester Stallone    45
         Dennis Quaid          45
         Willem Dafoe          45
         dtype: int64
```

```
In [37]: # bar chart between the actor on x axis and the number of their movies on y axis
         actors_count.head(10).plot.bar(color="steelblue", figsize=(8, 8))
         plt.xlabel("Actors")
         plt.ylabel("Number of movies")
         plt.show()
```



6 Conclusion3:

Robert De Niro appeared the most with 72 movies then Samuel L.Jackson with 71 movies

7 Charcteristics Associated with Successful Movies

```
In [152]: positive_profit = df.query("profit > 0")
          positive_profit.head()
```

```
Out[152]:        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     ...     runtime  \
          0            The park is open.     ...         124   
          1           What a Lovely Day.     ...         120   
          2   One Choice Can Destroy You     ...         119   
          3  Every generation has a story.     ...         136   
          4          Vengeance Hits Home     ...         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...   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      profit
0          6.5          2015  1.379999e+08  1.392446e+09  1363528810
1          7.1          2015  1.379999e+08  3.481613e+08   228436354
2          6.3          2015  1.012000e+08  2.716190e+08   185238201
3          7.5          2015  1.839999e+08  1.902723e+09  1868178225
4          7.3          2015  1.747999e+08  1.385749e+09  1316249360

[5 rows x 22 columns]
```

```
In [153]: # Sorting the dataFarme of positive profits
          positive_profit.sort_values(by="profit", ascending=False).head()
```

```
Out[153]:           id    imdb_id  popularity     budget     revenue  \
          1386   19995  tt0499549    9.432768  237000000  2781505847
          3     140607  tt2488496   11.173104  200000000  2068178225
          5231     597  tt0120338    4.355219  200000000  1845034188
          0     135397  tt0369610   32.985763  150000000  1513528810
          4     168259  tt2820852    9.335014  190000000  1506249360

                              original_title  \
          1386                        Avatar
          3     Star Wars: The Force Awakens
          5231                       Titanic
          0                   Jurassic World
          4                        Furious 7

                                                             cast  \
          1386  Sam Worthington|Zoe Saldana|Sigourney Weaver|S...
          3     Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
          5231  Kate Winslet|Leonardo DiCaprio|Frances Fisher|...
          0     Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
          4     Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                                         homepage         director  \
          1386                        http://www.avatarmovie.com/    James Cameron
          3     http://www.starwars.com/films/star-wars-episod...      J.J. Abrams
          5231              http://www.titanicmovie.com/menu.html    James Cameron
          0                         http://www.jurassicworld.com/  Colin Trevorrow
          4                              http://www.furious7.com/        James Wan

                                                  tagline     ...     runtime  \
          1386                Enter the World of Pandora.     ...         162
```

```
3                  Every generation has a story.     ...          136
5231  Nothing on Earth could come between them.     ...          194
0                              The park is open.     ...          124
4                            Vengeance Hits Home     ...          137

                                         genres  \
1386   Action|Adventure|Fantasy|Science Fiction
3      Action|Adventure|Science Fiction|Fantasy
5231                     Drama|Romance|Thriller
0     Action|Adventure|Science Fiction|Thriller
4                         Action|Crime|Thriller

                                   production_companies release_date  \
1386  Ingenious Film Partners|Twentieth Century Fox ...   2009-12-10
3             Lucasfilm|Truenorth Productions|Bad Robot   2015-12-15
5231  Paramount Pictures|Twentieth Century Fox Film ...   1997-11-18
0     Universal Studios|Amblin Entertainment|Legenda...   2015-06-09
4     Universal Pictures|Original Film|Media Rights ...   2015-04-01

     vote_count vote_average  release_year    budget_adj   revenue_adj  \
1386       8458          7.1          2009  2.408869e+08  2.827124e+09
3          5292          7.5          2015  1.839999e+08  1.902723e+09
5231       4654          7.3          1997  2.716921e+08  2.506406e+09
0          5562          6.5          2015  1.379999e+08  1.392446e+09
4          2947          7.3          2015  1.747999e+08  1.385749e+09

          profit
1386  2544505847
3     1868178225
5231  1645034188
0     1363528810
4     1316249360

[5 rows x 22 columns]
```

```
In [167]: # Getting top 50 movies in profit
          top_50 = positive_profit.head(50)
          top_50.describe()
```

```
Out[167]:                  id  popularity        budget       revenue     runtime  \
          count      50.000000   50.000000  5.000000e+01  5.000000e+01   50.000000
          mean   224370.260000    6.110215  8.689600e+07  3.952077e+08  120.460000
          std     73140.041735    5.550223  7.060099e+07  4.517118e+08   16.563064
          min     76341.000000    2.883233  0.000000e+00  9.064511e+06   91.000000
          25%    167369.500000    3.343932  2.825000e+07  8.918731e+07  109.500000
          50%    254224.000000    4.607380  7.100000e+07  2.297503e+08  118.500000
          75%    280051.000000    6.085384  1.462500e+08  5.293634e+08  130.000000
          max    339527.000000   32.985763  2.800000e+08  2.068178e+09  167.000000
```

```
        vote_count  vote_average  release_year    budget_adj   revenue_adj  \
count    50.000000     50.000000          50.0  5.000000e+01  5.000000e+01
mean   2120.480000      6.758000        2015.0  7.994428e+07  3.635909e+08
std    1402.367724      0.683655           0.0  6.495288e+07  4.155746e+08
min     396.000000      5.200000        2015.0  0.000000e+00  8.339346e+06
25%    1120.500000      6.225000        2015.0  2.598999e+07  8.205229e+07
50%    1652.000000      6.800000        2015.0  6.531997e+07  2.113702e+08
75%    2790.000000      7.300000        2015.0  1.345499e+08  4.870141e+08
max    6185.000000      8.000000        2015.0  2.575999e+08  1.902723e+09

             profit
count  5.000000e+01
mean   3.083117e+08
std    4.046298e+08
min    4.333790e+06
25%    3.745799e+07
50%    1.562816e+08
75%    3.956134e+08
max    1.868178e+09
```

```
In [155]: positive_profit["release_year"].value_counts()
```

```
Out[155]: 2013    194
          2014    184
          2011    181
          2015    179
          2012    169
          2010    160
          2008    157
          2006    150
          2007    145
          2009    139
          2005    135
          2004    116
          2003    107
          2002     97
          2001     92
          1993     85
          1999     81
          2000     80
          1997     80
          1996     76
          1998     76
          1995     74
          1989     71
          1992     69
          1988     66
```

```
1990    65
1994    65
1987    65
1986    60
1985    54
1991    53
1984    46
1983    45
1982    40
1980    37
1981    36
1979    25
1977    23
1978    22
1973    17
1974    17
1976    16
1975    15
1971    14
1967    12
1968    11
1972    10
1970    10
1961     9
1962     9
1964     8
1960     7
1963     6
1966     5
1969     4
1965     4
Name: release_year, dtype: int64
```

8 Conclusion 4

1- The average profit of the top 50 movies in profit is "308311700" (assuming that USD is the defaul currency). 2- The average runtime is 120.46 minutes. 3- The average Budget is 868960000. 4- The average revenue is 395207700. 5- most of them were released in 2013. 6- The average profit is 308311700.

9 Characteristics Associated with unseccessful movies

```
In [156]: # Getting movies with negative profits
          negative_profit = df[df["profit"] <= 0]
          negative_profit.head()
```

```
Out[156]:         id    imdb_id  popularity     budget   revenue  \
48  265208  tt2231253    2.932340   30000000         0
57  210860  tt3045616    2.575711   60000000  30418560
59  201088  tt2717822    2.550747   70000000  17752940
66  205775  tt1390411    2.345821  100000000  93820758
67  334074  tt3247714    2.331636   20000000         0

             original_title  \
48                Wild Card
57                Mortdecai
59                 Blackhat
66  In the Heart of the Sea
67                 Survivor

                                                 cast  \
48  Jason Statham|Michael Angarano|Milo Ventimigli...
57  Johnny Depp|Gwyneth Paltrow|Ewan McGregor|Paul...
59  Chris Hemsworth|Leehom Wang|Tang Wei|Viola Dav...
66  Chris Hemsworth|Benjamin Walker|Cillian Murphy...
67  Pierce Brosnan|Milla Jovovich|Dylan McDermott|...

                                   homepage        director  \
48                                      NaN      Simon West
57            http://mortdecaithemovie.com/     David Koepp
59  http://www.legendary.com/film/blackhat/    Michael Mann
66  http://www.intheheartoftheseamovie.com/      Ron Howard
67                http://survivormovie.com/  James McTeigue

                                              tagline    ...    runtime  \
48        Never bet against a man with a killer hand.    ...         92
57                         Sophistication Has a Name.    ...        106
59                       We are no longer in control.    ...        133
66  Based on the incredible true story that inspir...    ...        122
67                 His Next Target is Now Hunting Him    ...         96

                                     genres  \
48                     Thriller|Crime|Drama
57                         Comedy|Adventure
59      Mystery|Crime|Action|Thriller|Drama
66  Thriller|Drama|Adventure|Action|History
67                    Crime|Thriller|Action

                                 production_companies release_date vote_count  \
48  Current Entertainment|Lionsgate|Sierra / Affin...   2015-01-14        481
57  Lionsgate|Mad Chance|OddLot Entertainment|Huay...   2015-01-21        696
59  Universal Pictures|Forward Pass|Legendary Pict...   2015-01-13        584
66  Imagine Entertainment|Spring Creek Productions...   2015-11-20        805
67  Nu Image Films|Winkler Films|Millennium Films|...   2015-05-21        280
```

```
   vote_average  release_year   budget_adj   revenue_adj    profit
48          5.3          2015  2.759999e+07  0.000000e+00 -30000000
57          5.3          2015  5.519998e+07  2.798506e+07 -29581440
59          5.0          2015  6.439997e+07  1.633270e+07 -52247060
66          6.4          2015  9.199996e+07  8.631506e+07  -6179242
67          5.4          2015  1.839999e+07  0.000000e+00 -20000000

[5 rows x 22 columns]
```

```
In [157]: negative_profit.describe()
```

```
Out[157]:                  id   popularity        budget       revenue       runtime  \
          count    7092.000000  7092.000000  7.092000e+03  7.092000e+03  7092.000000
          mean    77337.273830     0.369477  6.628385e+06  2.302138e+06    98.751128
          std     98638.368109     0.346509  1.729492e+07  9.535256e+06    35.048770
          min        17.000000     0.000065  0.000000e+00  0.000000e+00     0.000000
          25%     13306.000000     0.162038  0.000000e+00  0.000000e+00    89.000000
          50%     27247.000000     0.287430  0.000000e+00  0.000000e+00    96.000000
          75%     97453.500000     0.471296  4.000000e+06  0.000000e+00   107.000000
          max    414419.000000     8.411577  4.250000e+08  1.730000e+08   900.000000

                  vote_count  vote_average  release_year    budget_adj   revenue_adj  \
          count  7092.000000   7092.000000   7092.000000  7.092000e+03  7.092000e+03
          mean     55.688804      5.836309   2001.856317  8.236126e+06  2.797028e+06
          std     102.107416      0.979386     13.125779  2.075159e+07  1.119590e+07
          min      10.000000      1.500000   1960.000000  0.000000e+00  0.000000e+00
          25%      14.000000      5.200000   1996.000000  0.000000e+00  0.000000e+00
          50%      23.000000      5.900000   2007.000000  0.000000e+00  0.000000e+00
          75%      52.000000      6.500000   2012.000000  4.748721e+06  0.000000e+00
          max    1777.000000      9.200000   2015.000000  4.250000e+08  1.819387e+08

                       profit
          count  7.092000e+03
          mean  -4.326247e+06
          std    1.201443e+07
          min   -4.139124e+08
          25%   -2.704157e+06
          50%    0.000000e+00
          75%    0.000000e+00
          max    0.000000e+00
```

```
In [158]: negative_profit["release_year"].value_counts()
```

```
Out[158]: 2014    516
          2013    465
          2015    450
          2012    419
          2009    394
          2011    359
```

```
2008    339
2010    329
2007    293
2006    258
2005    229
2004    191
2003    174
2002    169
2001    150
2000    147
1999    143
1998    134
1996    128
1994    119
1997    112
1995    101
1993     93
1991     80
1988     79
1990     67
1989     66
1992     64
1986     61
1987     60
1984     59
1985     55
1981     46
1978     43
1966     41
1971     41
1980     41
1982     41
1973     38
1983     35
1977     34
1964     34
1979     32
1965     31
1976     31
1970     31
1972     30
1974     30
1975     29
1968     28
1963     28
1967     28
1969     27
1960     25
```

```
1962     23
1961     22
Name: release_year, dtype: int64
```

10 Conclusion 5:

1- The average profit of the lowest 50 movies in profit is "308311700" (assuming that USD is the defaul currency). 2- The average runtime is 98.75 minutes. 3- The average Budget is 662838500. 4- The average revenue is 230213800. 5- most of them were released in 2014. 6- The average loss is 432624700.

11 General Conclusions

1-The dataset has samples and this enough to investigate the dataset 2-Removing Nulls would affect the data hence the results of analysis and most of them are in hompage column which I did not used to investigate the dataset. 3-There is only one duplicated row and I removed it. 4-I added an extra column that holds th profit. 5-The relationship between runtime and profits check Q1 up. 6- To see which genres are most Popular or most common or whuch of them made the top profits check Q2 7-To see The actors that appered the most check Q3 8- to see the characteristics of successful movies check 'The characteristics associated with successful movies' 9- to see the characteristics of unsuccessful films check ' The characteristics associated with unsuccessful movies'. 10- The massive amount of values is useful as it make our analysis more accurate.

11.1 OPTIONAL: Question for the reviewer

If you have any question about the starter code or your own implementation, please add it in the cell below.

For example, if you want to know why a piece of code is written the way it is, or its function, or alternative ways of implementing the same functionality, or if you want to get feedback on a specific part of your code or get feedback on things you tried but did not work.

Please keep your questions succinct and clear to help the reviewer answer them satisfactorily.

Your question

```
In [159]: from subprocess import call
          call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

Out[159]: 0

In [ ]:
```