Investigate_a_Dataset

December 14, 2021

1 Project: Investigate a Dataset - [TMDb Movies]

1.1 Table of Contents

1.1.1 Dataset Description

Tip: In this section of the report, provide a brief introduction to the dataset you've selected/downloaded for analysis. Read through the description available on the homepage-links present here. List all column names in each table, and their significance. In case of multiple tables, describe the relationship between tables.

```
In [1]: # Use this cell to set up import statements for all of the packages that you
        #   plan to use.
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        %matplotlib inline
        import seaborn as sns
        import csv
        import datetime as datetime
```

Data Wrangling

```
In [2]: # 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.
        df=pd.read_csv('Database_TMDb_movie_data/tmdb-movies.csv')
        df.shape
```

```
Out[2]: (10866, 21)
```

```
In [3]: df.head(2) #checking columns and first 2 raws

Out[3]:       id    imdb_id  popularity     budget     revenue      original_title  \
        0  135397  tt0369610   32.985763  150000000  1513528810      Jurassic World
        1   76341  tt1392190   28.419936  150000000   378436354  Mad Max: Fury Road

                                                cast  \
        0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
        1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...

                            homepage         director             tagline  \
        0  http://www.jurassicworld.com/  Colin Trevorrow   The park is open.
        1    http://www.madmaxmovie.com/    George Miller  What a Lovely Day.

              ...                                                overview runtime  \
        0     ...       Twenty-two years after the events of Jurassic ...     124
        1     ...       An apocalyptic story set in the furthest reach...     120

                                          genres  \
        0  Action|Adventure|Science Fiction|Thriller
        1  Action|Adventure|Science Fiction|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

           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 rows x 21 columns]

In [4]: df.info() #checking data types

<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
```

```
In [5]: df.nunique() #checking unique values
```

```
Out[5]: id                    10865
        imdb_id                10855
        popularity             10814
        budget                   557
        revenue                 4702
        original_title         10571
        cast                   10719
        homepage                2896
        director                5067
        tagline                 7997
        keywords                8804
        overview               10847
        runtime                  247
        genres                  2039
        production_companies    7445
        release_date            5909
        vote_count              1289
        vote_average              72
        release_year              56
        budget_adj              2614
        revenue_adj             4840
        dtype: int64
```

```
In [6]: df.describe()
```

```
Out[6]:                   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
```

Questions

Which 5 movies had the highest and lowest profit?

Whats the Highest and Lowest (Budget , Revenue? Relationship between Revenue , Budget and Profit Which movie had the greatest and least runtime? RelationShip between Profit and Vote Average RelationShip between Vote Count and Vote Average Relationship between RunTime and Vote Count Relationship between RunTime and Profit Which genres are most popular ? <a href="#8">Whichre Top 5 Production Studios?

2 Data Cleaning

let's check the Duplicates

```
In [7]: df.duplicated().sum() #checking Duplicates
```

```
Out[7]: 1
```

```
In [8]: df.drop_duplicates(inplace=True) #Droping Duplicates
```

```
In [9]: df.duplicated().sum() #checking duplicates after drop
```

```
Out[9]: 0
```

Let's check the Columns and Drop the unused

```
In [10]: df.columns #checking Columns names
```

```
Out[10]: Index(['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_title',
                'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview',
                'runtime', 'genres', 'production_companies', 'release_date',
                'vote_count', 'vote_average', 'release_year', 'budget_adj',
                'revenue_adj'],
               dtype='object')
```

```
In [11]: #Removing unwanted Columns
         df=df.drop(['imdb_id','homepage','tagline','overview','budget_adj','revenue_adj','keywo

         print("Afetr Removing Unused Columns (Rows,Columns) : ",df.shape)

Afetr Removing Unused Columns (Rows,Columns) :  (10865, 14)


In [12]: df.info() #checking columns Types

<class 'pandas.core.frame.DataFrame'>
Int64Index: 10865 entries, 0 to 10865
Data columns (total 14 columns):
id                      10865 non-null int64
popularity              10865 non-null float64
budget                  10865 non-null int64
revenue                 10865 non-null int64
original_title          10865 non-null object
cast                    10789 non-null object
director                10821 non-null object
runtime                 10865 non-null int64
genres                  10842 non-null object
production_companies    9835 non-null object
release_date            10865 non-null object
vote_count              10865 non-null int64
vote_average            10865 non-null float64
release_year            10865 non-null int64
dtypes: float64(2), int64(6), object(6)
memory usage: 1.2+ MB


In [13]: df['release_date']=pd.to_datetime(df['release_date']) #Converting realease date Column

In [14]: df['net_profit']=df['revenue']-df['budget'] #Creating New column (profit)

In [15]: df.head(2)

Out[15]:        id  popularity     budget     revenue      original_title  \
         0  135397   32.985763  150000000  1513528810      Jurassic World
         1   76341   28.419936  150000000   378436354  Mad Max: Fury Road

                                                  cast         director  \
         0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...  Colin Trevorrow
         1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...    George Miller

            runtime                                     genres  \
         0      124  Action|Adventure|Science Fiction|Thriller
         1      120  Action|Adventure|Science Fiction|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

   vote_average  release_year  net_profit
0           6.5          2015  1363528810
1           7.1          2015   228436354
```

Let's check the Na Values

```
In [16]: df.isnull().sum()
```

```
Out[16]: id                       0
         popularity               0
         budget                   0
         revenue                  0
         original_title           0
         cast                    76
         director                44
         runtime                  0
         genres                  23
         production_companies  1030
         release_date             0
         vote_count               0
         vote_average             0
         release_year             0
         net_profit               0
         dtype: int64
```

```
In [17]: # Columns we need to check for na
         columns = ['budget', 'revenue']
         # Replace 0 with NAN
         df[columns] = df[columns].replace(0, np.NaN)
         # Drop rows which contains NAN
         df.dropna(subset = columns, inplace = True)
         print("Afetr Droping rows contains NAN: ",df.shape)
```

```
Afetr Droping rows contains NAN:  (3854, 15)
```

```
In [18]: df.describe() #checking data
```

```
Out[18]:                 id   popularity        budget       revenue      runtime  \
         count    3854.000000  3854.000000  3.854000e+03  3.854000e+03  3854.000000
         mean    39888.185262     1.191554  3.720370e+07  1.076866e+08   109.220291
         std     67222.527399     1.475162  4.220822e+07  1.765393e+08    19.922820
         min         5.000000     0.001117  1.000000e+00  2.000000e+00    15.000000
         25%      6073.500000     0.462368  1.000000e+07  1.360003e+07    95.000000
         50%     11321.500000     0.797511  2.400000e+07  4.480000e+07   106.000000
```

```
75%     38573.250000     1.368324  5.000000e+07  1.242125e+08   119.000000
max    417859.000000    32.985763  4.250000e+08  2.781506e+09   338.000000

        vote_count  vote_average  release_year    net_profit
count  3854.000000   3854.000000   3854.000000  3.854000e+03
mean    527.720291      6.168163   2001.261028  7.048292e+07
std     879.956821      0.794920     11.282575  1.506195e+08
min      10.000000      2.200000   1960.000000 -4.139124e+08
25%      71.000000      5.700000   1995.000000 -1.321535e+06
50%     204.000000      6.200000   2004.000000  2.002019e+07
75%     580.000000      6.700000   2010.000000  8.170331e+07
max    9767.000000      8.400000   2015.000000  2.544506e+09
```

Exploratory Data Analysis

Tip: Now that you've trimmed and cleaned your data, you're ready to move on to exploration. **Compute statistics** and **create visualizations** with the goal of addressing the research questions that you posed in the Introduction section. You should compute the relevant statistics throughout the analysis when an inference is made about the data. Note that at least two or more kinds of plots should be created as part of the exploration, and you must compare and show trends in the varied visualizations.

Tip: - Investigate the stated question(s) from multiple angles. It is recommended that you be systematic with your approach. Look at one variable at a time, and then follow it up by looking at relationships between variables. You should explore at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent variable of interest. Lastly, you should perform both single-variable (1d) and multiple-variable (2d) explorations.

Research Question 1 (Which 5 movies had the highest and lowest profit??)

In [19]:
```
#top 5 movies
top_5=df.nlargest(5,'net_profit')
top_5
```

Out[19]:
```
          id  popularity       budget       revenue  \
1386   19995    9.432768  237000000.0  2.781506e+09
3     140607   11.173104  200000000.0  2.068178e+09
5231     597    4.355219  200000000.0  1.845034e+09
0     135397   32.985763  150000000.0  1.513529e+09
4     168259    9.335014  190000000.0  1.506249e+09

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

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

      runtime                                      genres  \
1386      162   Action|Adventure|Fantasy|Science Fiction
3         136   Action|Adventure|Science Fiction|Fantasy
5231      194                     Drama|Romance|Thriller
0         124  Action|Adventure|Science Fiction|Thriller
4         137                      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  net_profit
1386        8458           7.1          2009  2544505847
3           5292           7.5          2015  1868178225
5231        4654           7.3          1997  1645034188
0           5562           6.5          2015  1363528810
4           2947           7.3          2015  1316249360
```

As We can see Top 5 movies Made a Profit is : 1. Avatar (With 2.544 Billion Dollars) 2. Star Wars: The Force Awakens (With 1.868 Billion Dollar) 3. Titanic (With 1.645 Billion Dollar) 4. Jurassic World (With 1.363 Billion Dollar) 5. Furious 7 (With 1.316 Billion Dollar)

```
In [20]: lowest_5=df.nsmallest(5,'net_profit') #lowest 5 movies
         lowest_5
```

```
Out[20]:          id  popularity       budget     revenue     original_title  \
         2244  46528    0.250540  425000000.0  11087569.0  The Warrior's Way
         5508  57201    1.214510  255000000.0  89289910.0    The Lone Ranger
         7031  10733    0.948560  145000000.0  25819961.0          The Alamo
         3484  50321    0.921653  150000000.0  38992758.0    Mars Needs Moms
         4970  10009    1.653031  100000000.0       250.0       Brother Bear

                                                      cast  \
         2244  Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
         5508  Johnny Depp|Armie Hammer|William Fichtner|Hele...
         7031  Dennis Quaid|Billy Bob Thornton|Jason Patric|P...
```

```
3484  Seth Green|Joan Cusack|Dan Fogler|Breckin Meye...
4970  Joaquin Phoenix|Jeremy Suarez|Rick Moranis|Joa...

                        director  runtime  \
2244                  Sngmoo Lee      100
5508              Gore Verbinski      149
7031           John Lee Hancock      137
3484                 Simon Wells       88
4970  Aaron Blaise|Robert Walker       85

                                       genres  \
2244  Adventure|Fantasy|Action|Western|Thriller
5508                  Action|Adventure|Western
7031                      Western|History|War
3484               Adventure|Animation|Family
4970        Animation|Adventure|Family|Fantasy

                                  production_companies release_date  \
2244                          Boram Entertainment Inc.   2010-12-02
5508  Walt Disney Pictures|Jerry Bruckheimer Films|I...   2013-07-03
7031         Imagine Entertainment|Touchstone Pictures   2004-04-07
3484                     Walt Disney Animation Studios   2011-03-09
4970  Walt Disney Pictures|Walt Disney Feature Anima...   2003-10-20

      vote_count  vote_average  release_year  net_profit
2244          74           6.4          2010  -413912431
5508        1607           6.0          2013  -165710090
7031          60           5.9          2004  -119180039
3484         129           5.5          2011  -111007242
4970         753           6.8          2003   -99999750
```

Now with the Lowest 5 Movies : 1. The Warrior's Way (with -413 Million Dollars) 2. The Lone Ranger (with -165 Million Dollars) 3. The Alamo (With -119 Million Dollars) 4. Mars Needs Moms (With -111 Million Dollars) 5. Brother Bear (with -99 Million Dollar)

Research Question 2 (What's the Highest and Lowest (Budget , Revenue?)

```
In [21]: def minmax(x):
             # function 'idmin' to find the lowest profit movie.
             min_index = df[x].idxmin()
             # function 'idmax' to find Highest profit movie.
             high_index = df[x].idxmax()
             high = pd.DataFrame(df.loc[high_index,:])
             low = pd.DataFrame(df.loc[min_index,:])
              # print the movie with high and low Budget and Revenue
             print("Movie With the Highest "+ x + " : ",df['original_title'][high_index])
             print("Movie With the Lowest "+ x + "  : ",df['original_title'][min_index])
```

```
        return pd.concat([high,low],axis = 1)

    # minmax function.
    minmax('budget')
```

```
Movie With the Highest budget :  The Warrior's Way
Movie With the Lowest budget  :  Lost & Found
```

Out[21]:

	2244	\
id	46528	
popularity	0.25054	
budget	4.25e+08	
revenue	1.10876e+07	
original_title	The Warrior's Way	
cast	Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...	
director	Sngmoo Lee	
runtime	100	
genres	Adventure|Fantasy|Action|Western|Thriller	
production_companies	Boram Entertainment Inc.	
release_date	2010-12-02 00:00:00	
vote_count	74	
vote_average	6.4	
release_year	2010	
net_profit	-413912431	
	2618	
id	39964	
popularity	0.090186	
budget	1	
revenue	100	
original_title	Lost & Found	
cast	David Spade|Sophie Marceau|Ever Carradine|Step...	
director	Jeff Pollack	
runtime	95	
genres	Comedy|Romance	
production_companies	Alcon Entertainment|Dinamo Entertainment	
release_date	1999-04-23 00:00:00	
vote_count	14	
vote_average	4.8	
release_year	1999	
net_profit	99	

In [22]:
```
minmax('revenue')
```

```
Movie With the Highest revenue :  Avatar
Movie With the Lowest revenue  :  Shattered Glass
```

Out[22]:

		1386 \
id		19995
popularity		9.43277
budget		2.37e+08
revenue		2.78151e+09
original_title		Avatar
cast	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...	
director		James Cameron
runtime		162
genres	Action|Adventure|Fantasy|Science Fiction	
production_companies	Ingenious Film Partners|Twentieth Century Fox ...	
release_date	2009-12-10 00:00:00	
vote_count		8458
vote_average		7.1
release_year		2009
net_profit		2544505847
		5067
id		13537
popularity		0.462609
budget		6e+06
revenue		2
original_title		Shattered Glass
cast	Hayden Christensen|Peter Sarsgaard|ChloÃń Sevi...	
director		Billy Ray
runtime		94
genres		Drama|History
production_companies	Lions Gate Films|Cruise/Wagner Productions|Bau...	
release_date	2003-11-14 00:00:00	
vote_count		46
vote_average		6.4
release_year		2003
net_profit		-5999998

In [23]:
```
minmax('net_profit')
```

```
Movie With the Highest net_profit :  Avatar
Movie With the Lowest net_profit  :  The Warrior's Way
```

Out[23]:

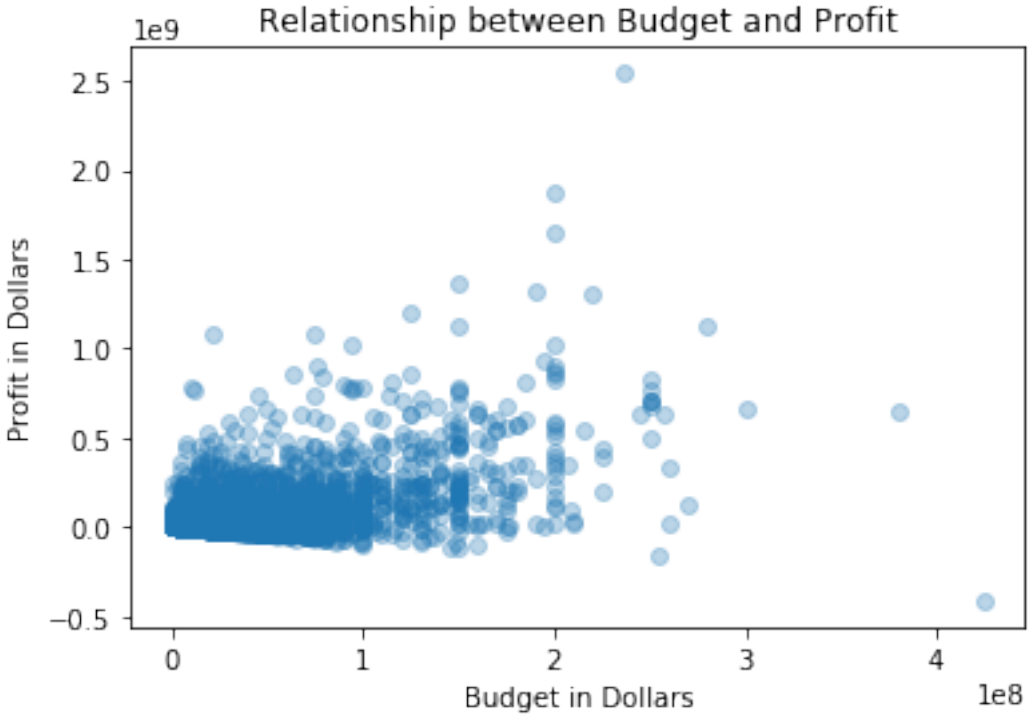
		1386 \
id		19995
popularity		9.43277
budget		2.37e+08
revenue		2.78151e+09
original_title		Avatar
cast	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...	
director		James Cameron

```
runtime                                                       162
genres                          Action|Adventure|Fantasy|Science Fiction
production_companies  Ingenious Film Partners|Twentieth Century Fox ...
release_date                                    2009-12-10 00:00:00
vote_count                                                   8458
vote_average                                                  7.1
release_year                                                 2009
net_profit                                                2544505847

                                                             2244
id                                                          46528
popularity                                                0.25054
budget                                                     4.25e+08
revenue                                                1.10876e+07
original_title                                    The Warrior's Way
cast                  Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
director                                                 Sngmoo Lee
runtime                                                         100
genres                        Adventure|Fantasy|Action|Western|Thriller
production_companies                          Boram Entertainment Inc.
release_date                                    2010-12-02 00:00:00
vote_count                                                      74
vote_average                                                   6.4
release_year                                                  2010
net_profit                                                -413912431
```

Research Question 3 (Relationship between Revenue , Budget and Profit)

In [24]:
```
# x-label and y-label
plt.xlabel('Revenue in Dollars')
plt.ylabel('Profit in Dollars')
# title
plt.title('Relationship between Revenue and Profit')
plt.scatter(df['revenue'], df['net_profit'],alpha=0.3)
plt.show()
# x-label and y-label
plt.xlabel('Budget in Dollars')
plt.ylabel('Profit in Dollars')
# title
plt.title('Relationship between Budget and Profit')
plt.scatter(df['budget'], df['net_profit'],alpha=0.3)
plt.show()
```

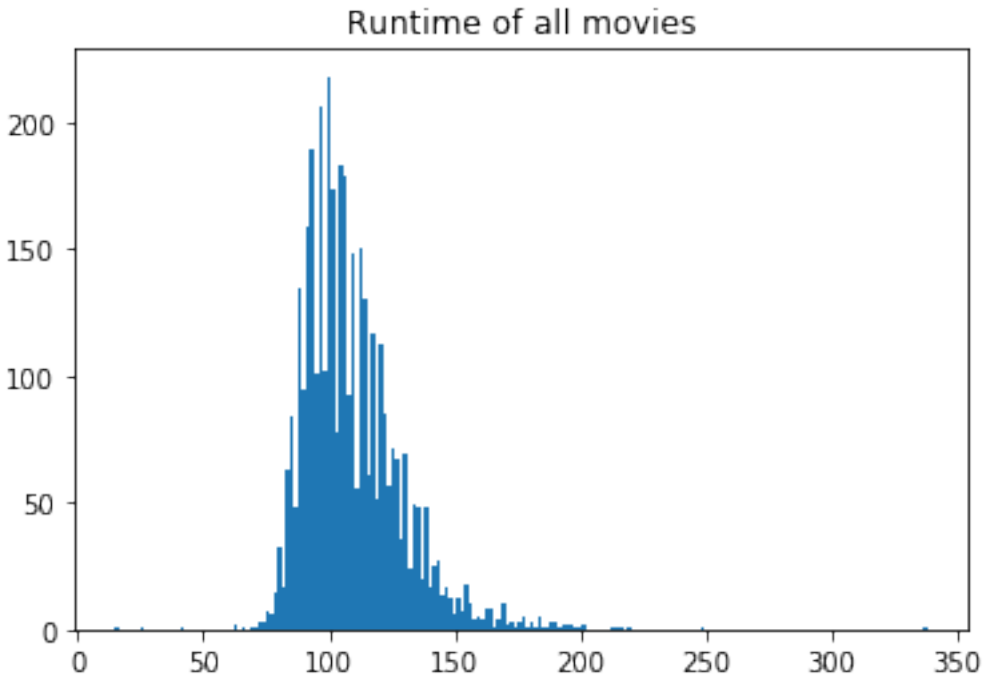

Relationship between Revenue and Profit
1e9
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Profit in Dollars
0.0
0.5
1.0
1.5
2.0
2.5
Revenue in Dollars
1e9


Relationship between Budget and Profit
1e9
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Profit in Dollars
0
1
2
3
4
Budget in Dollars
1e8

We can see that there's a strong relationship between profit and revenue, higher the profit, higher the revenue. ALSO We can see that there no as such relationship between budget and profits, But yes there are very less flims which didnt make profit when the budget was greater then 20M Dollar.

Research Question 4 (Which movie had the greatest and least runtime?)

In [25]:
```
# first we can check the distribution of Runtime of all Movies with Histogram
plt.title('Runtime of all movies')
plt.hist(df['runtime'], bins = 200);
plt.show()
```



In [26]:
```
# Runtime Average
df['runtime'].mean()
```

Out[26]:
```
109.22029060716139
```

As we can see the Average runtime for All movies around 110 Minute

In [27]:
```
df.nlargest(1,'runtime')
```

Out[27]:
```
          id  popularity      budget    revenue original_title  \
2107   43434    0.534192  18000000.0  871279.0         Carlos

                                                    cast         director  \
2107  Edgar RamÃŋrez|Alexander Scheer|Fadi Abi Samra...  Olivier Assayas
```

```
      runtime                          genres  \
2107      338  Crime|Drama|Thriller|History

                                    production_companies release_date  \
2107  Egoli Tossell Film AG|Canal+|Arte France|Films...   2010-05-19

      vote_count  vote_average  release_year  net_profit
2107          35           6.2          2010   -17128721
```

Movie with greatest runtime : Carlos with 338 minutes record

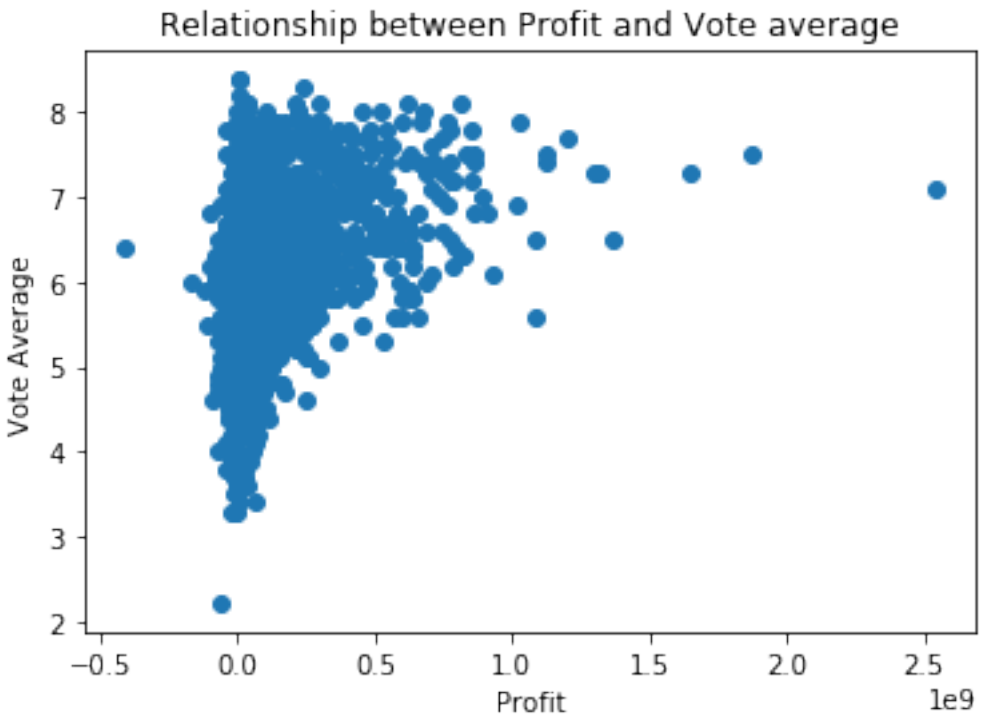
```
In [28]: df.nsmallest(1,'runtime')
```

```
Out[28]:          id  popularity  budget  revenue original_title  \
         5162  24914    0.208637    10.0      5.0    Kid's Story

                                                    cast             director  \
         5162  Clayton Watson|Keanu Reeves|Carrie-Anne Moss|K...  Shinichiro Watanabe

               runtime                     genres production_companies release_date  \
         5162       15  Science Fiction|Animation          Studio 4ÂřC   2003-06-02

               vote_count  vote_average  release_year  net_profit
         5162          16           6.8          2003          -5
```

Movie with least runtime : Kid's Story As a 15 Min Record

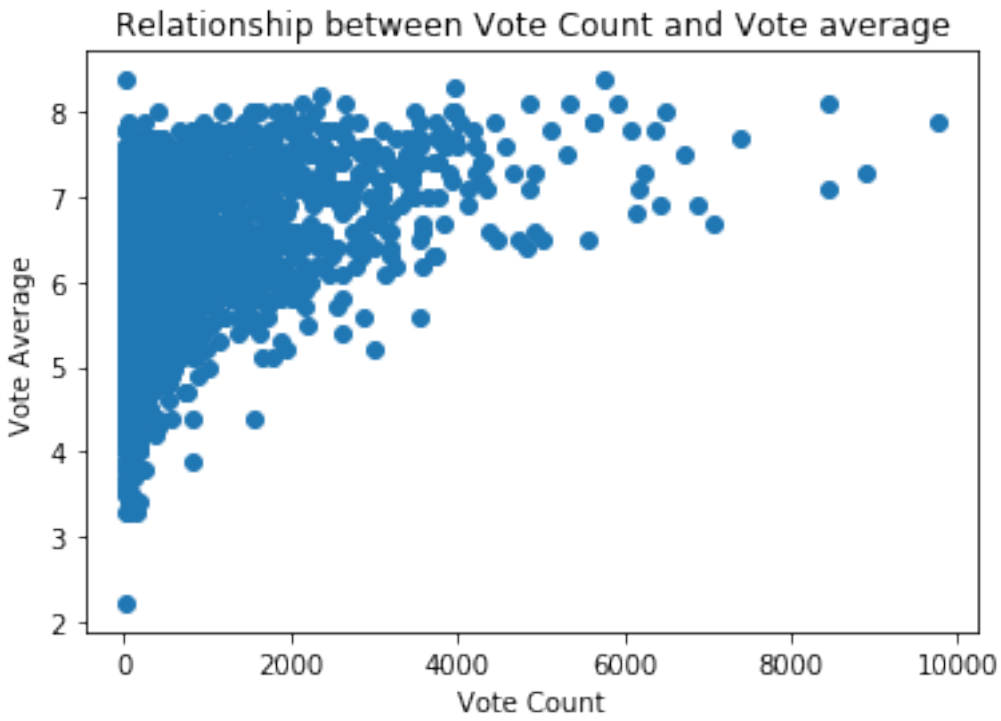
Relationship Between Profit and Vote Average

```
In [29]: # x-label and y-label
         plt.xlabel('Profit')
         plt.ylabel('Vote Average')
         # title
         plt.title('Relationship between Profit and Vote average')
         plt.scatter(df['net_profit'], df['vote_average'])
         plt.show()
```



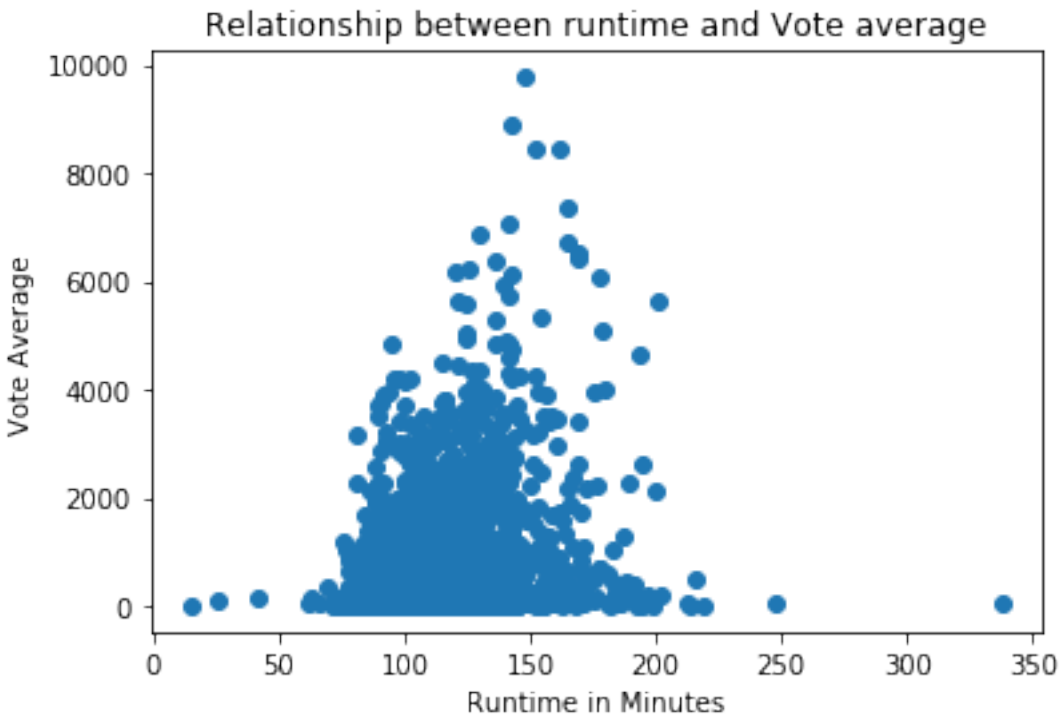
RelationShip between Vote count and Vote Average

```
In [30]: # x-label and y-label
         plt.xlabel('Vote Count')
         plt.ylabel('Vote Average')
         # title
         plt.title('Relationship between Vote Count and Vote average')
         plt.scatter(df['vote_count'], df['vote_average'])
         plt.show()
```



Research Question 5 (Relationship between RunTime and Vote Count)

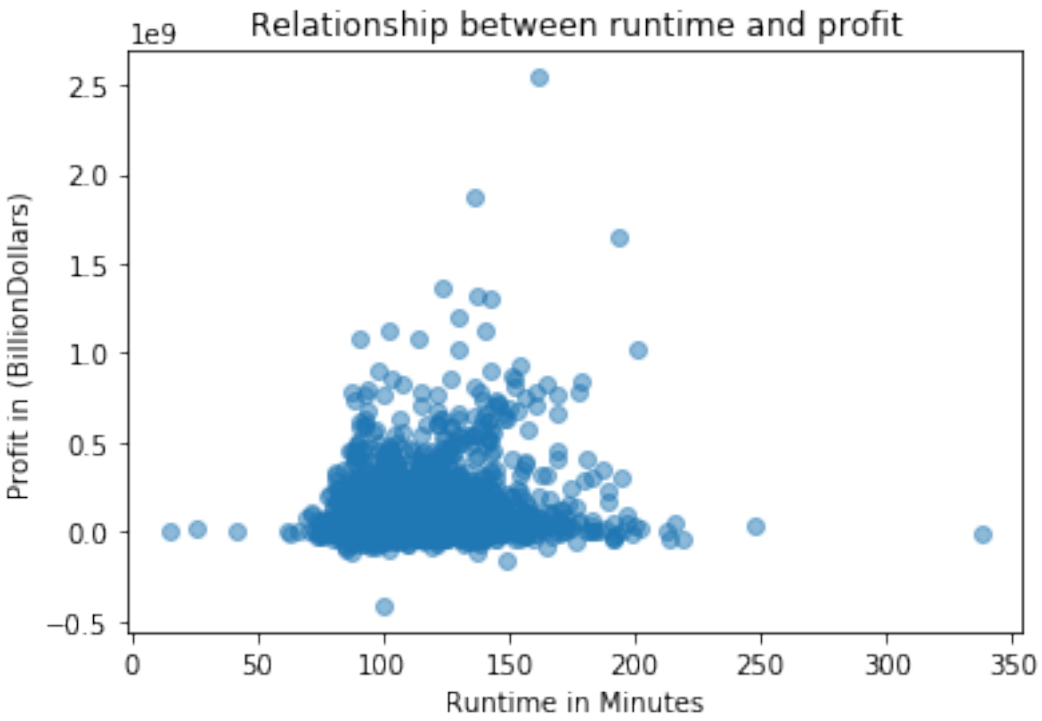
```
In [31]: # x-label and y-label
         plt.xlabel('Runtime in Minutes')
         plt.ylabel('Vote Average')
         # title
         plt.title('Relationship between runtime and Vote average')
         plt.scatter(df['runtime'], df['vote_count'])
         plt.show()
```



As we can see that most votes goes to the Runtime average for all the movies that's around 110 Minutes

Research Question 6 (Relationship between RunTime and Profit)

```
In [32]: # x-label and y-label
         plt.xlabel('Runtime in Minutes')
         plt.ylabel('Profit in (BillionDollars)')
         # title
         plt.title('Relationship between runtime and profit')
         plt.scatter(df['runtime'], df['net_profit'],alpha=0.5)
         plt.show()
```



Most of the movies have runtime in range of 85 to 130 Minutes

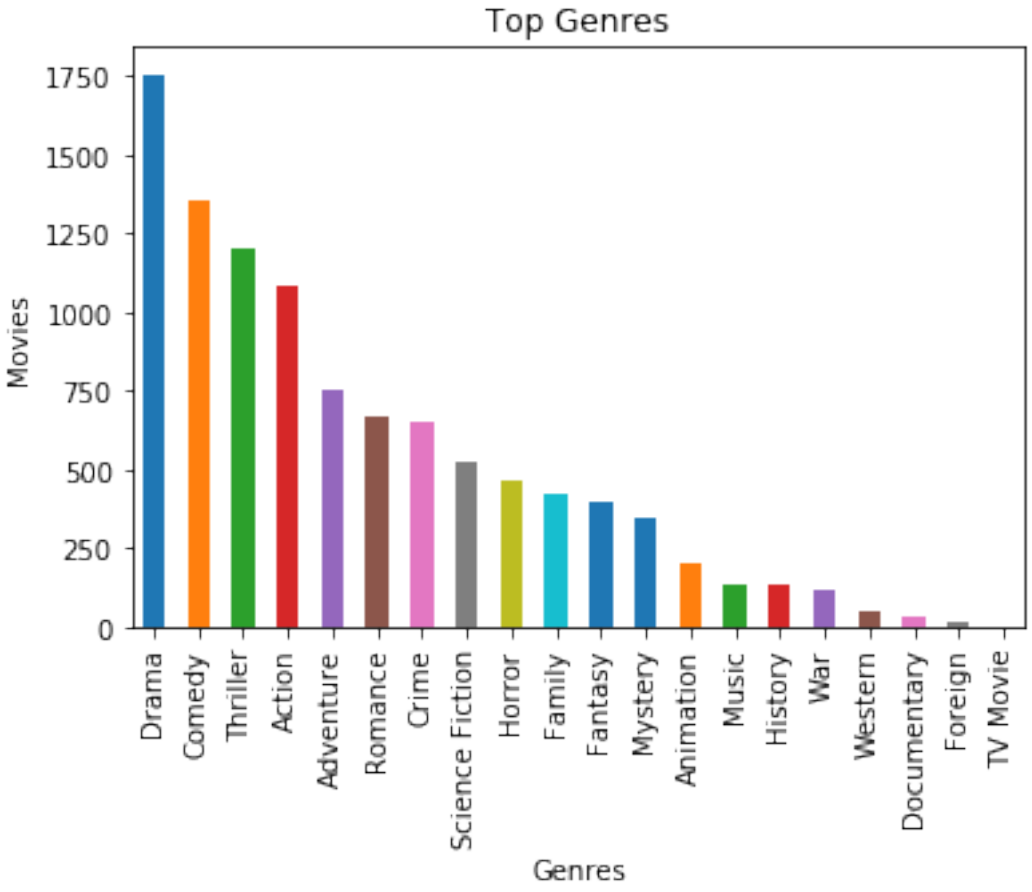
Research Question 7 (Which genres are most popular ?)

```
In [33]: genres_count = pd.Series(df['genres'].str.cat(sep = '|').split('|')).value_counts(ascen
         genres_count

Out[33]: Drama              1756
         Comedy             1358
         Thriller           1204
         Action             1085
         Adventure           749
         Romance             667
         Crime               651
         Science Fiction     519
         Horror              463
         Family              425
         Fantasy             396
         Mystery             344
         Animation           201
         Music               136
         History             129
         War                 119
         Western              52
```

```
Documentary            35
Foreign                13
TV Movie                1
dtype: int64
```

So the Top 10 Genres are Drama, Comedy, Action, Thriller, Adventure, Romance, Crime, Family, Scince Fiction, Fantasy Lets visualize this with a plot

```
In [34]: # we can review the answer by diagram graph
         diagram = genres_count.plot.bar()
         # x-label and y-label
         diagram.set_xlabel('Genres')
         diagram.set_ylabel('Movies')
         # title
         diagram.set(title = 'Top Genres')
         plt.show()
```



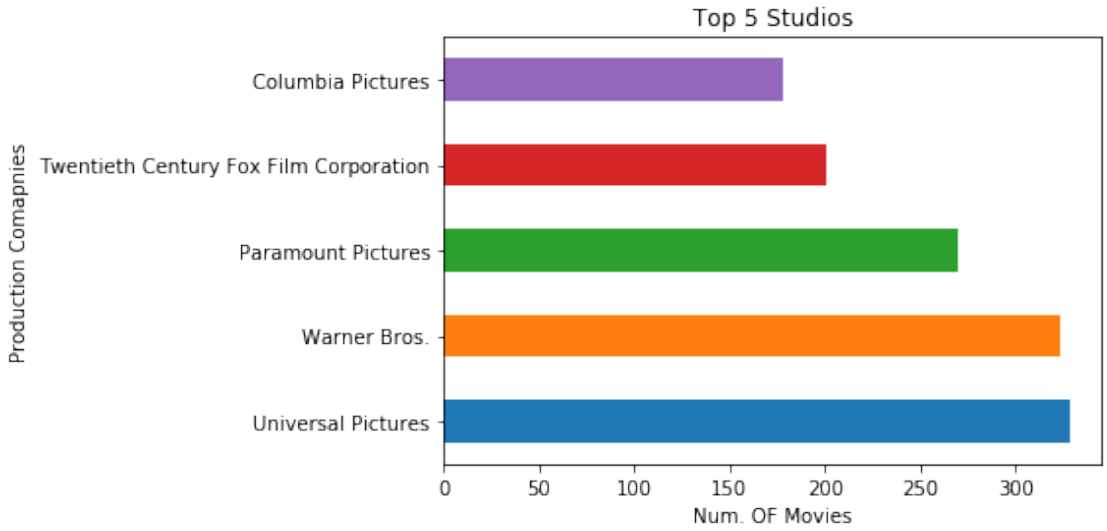
Research Question 8 (Which're Top 5 Production Studios?)

```
In [35]: prod_company_count = pd.Series(df['production_companies'].str.cat(sep = '|').split('|')
         top_5_studios=prod_company_count.head()
         top_5_studios
```

```
Out[35]: Universal Pictures                          329
         Warner Bros.                                324
         Paramount Pictures                          270
         Twentieth Century Fox Film Corporation      201
         Columbia Pictures                           178
         dtype: int64
```

As we can see the Top 5 Studios is : Universal Pictures With 329 Movie Warner Bros. With 324 Movie Paramount Pictures With 270 Movie Twentieth Century Fox Film Corporation With 201 Movie Columbia Pictures With 178 Movie ##### and here's a Plot to visualize the result

```
In [36]: # we can review the answer by diagram graph
         diagram = top_5_studios.plot.barh()
         # x-label and y-label
         diagram.set_xlabel('Num. OF Movies')
         diagram.set_ylabel('Production Comapnies')
         # title
         diagram.set(title = 'Top 5 Studios')
         plt.show()
```



As we can see the Top production Company is Universal pictures

Conclusions

After investigating the Tmdb Movies data set we Figured out that

1. Top 5 Movies made a profit is :
2. Avatar (With 2.544 Billion Dollars)
3. Star Wars: The Force Awakens (With 1.868 Billion Dollar)

4. Titanic (With 1.645 Billion Dollar)
5. Jurassic World (With 1.363 Billion Dollar)
6. Furious 7 (With 1.316 Billion Dollar)
7. Lowest 5 Profitable Movies :
8. The Warrior's Way (with -413 Million Dollars)
9. The Lone Ranger (with -165 Million Dollars)
10. The Alamo (With -119 Million Dollars)
11. Mars Needs Moms (With -111 Million Dollars)
12. Brother Bear (with -99 Million Dollar)
13. Movie With the Highest budget : The Warrior's Way
14. Movie With the Lowest budget : Lost & Found
15. Average runtime for All movies around 110 Minute
16. Top 5 Geners for all the time is :

    1. Drama
    2. Comedy
    3. Thriller
    4. Action
    5. Adventure

17. Top 5 Production Companies

18. Universal Pictures

19. Warner Bros.

20. Paramount Pictures

21. Twentieth Century Fox Film Corporation

22. Columbia Pictures

Data Limitation

The conclusion is not full proof that given the above requirement the movie will be a big hit but it can be.

Also, we also lost some of the data in the data cleaning steps where we dont know the revenue and budget of the movie, which has affected our analysis.

This conclusion is not error proof.

```
In [37]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[37]: 0
```