Investigate_a_Dataset

January 8, 2023

> **Tip**: Welcome to the Investigate a Dataset project! You will find tips in quoted sections like this to help organize your approach to your investigation. Once you complete this project, remove these **Tip** sections from your report before submission. First things first, you might want to double-click this Markdown cell and change the title so that it reflects your dataset and investigation.

1 Project: Investigate a Dataset - TMDb movie data

1.1 Table of Contents

1.1.1 Dataset Description

This data set contains information about 10,000 movies collected from The Movie Database (TMDb),including user ratings and revenue.

1.1.2 Question(s) for Analysis

Q1:Which movies have high and least profit?

Q2:What factors are associated with Profit Earned?

Q3:Which movies with longest and shortest runtime values?

Q4: Which movies with largest and lowest budgets?

```
In [3]: import pandas as pd
        import numpy as np
        import csv
        from datetime import datetime
        import matplotlib.pyplot as plt
        import seaborn as sns
        % matplotlib inline
```

Data Wrangling

After observing the dataset, we will be keeping only relevent data for our data anaylsis process to answer the suggested questions.

In [4]:
```
# 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('tmdb-movies.csv')
```

In [5]:
```
#view first rows of the dataset
df.head()
```

Out[5]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	...	overview	runtime
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Colin Trevorrow	The park is open.	...	Twenty-two years after the events of Jurassic ...	124
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.com/	George Miller	What a Lovely Day.	...		
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insurgent	Robert Schwentke	One Choice Can Destroy You	...		
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-wars-episod...	J.J. Abrams	Every generation has a story.	...		
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.com/	James Wan	Vengeance Hits Home	...		

```
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 [6]:
```
#dataset summary
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
```

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

```
Out[7]: (10866, 21)
```

```
In [8]: #dataset info
        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
```

```
In [9]: #duplicated rows
        df.duplicated().sum()
```

```
Out[9]: 1
```

```
In [10]: #missing date in the dataset
         df.isnull().sum()
```

```
Out[10]: id                         0
         imdb_id                   10
```

```
popularity                0
budget                    0
revenue                   0
original_title            0
cast                     76
homepage               7930
director                 44
tagline                2824
keywords               1493
overview                  4
runtime                   0
genres                   23
production_companies    1030
release_date              0
vote_count                0
vote_average              0
release_year              0
budget_adj                0
revenue_adj               0
dtype: int64
```

```
In [11]: df.columns[df.isnull().any()]
```

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

Obsevations from the data set:

No unit of currency is mentioned in the dataset. So for my analysis I will take it as dollar as it is the most used international currency.

1.1.3 Data Cleaning

1.Removing unused column such as: id, imdb_id, popularity, budget_adj, revenue_adj, homepage, keywords, overview, production_companies and vote_average.

2.Removing the duplicacy in the rows.

3.Some movies in the database have zero budget or zero revenue, that is there value has not been recorded so we will be discarding such entries

4.Changing format of budget and revenue column.

5.Changing release date column into date format.

```
In [12]: #droping info that wont be used
         df.drop(['id', 'imdb_id', 'popularity', 'budget_adj', 'revenue_adj', 'homepage', 'keywo
         df.head()
```

Out[12]:

	budget	revenue	original_title	cast	director
0	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow
1	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller
2	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke
3	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams
4	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan

	tagline	runtime	genres	release_date	vote_count	release_year
0	The park is open.	124	Action|Adventure|Science Fiction|Thriller	6/9/15	5562	2015
1	What a Lovely Day.	120	Action|Adventure|Science Fiction|Thriller	5/13/15	6185	2015
2	One Choice Can Destroy You	119	Adventure|Science Fiction|Thriller	3/18/15	2480	2015
3	Every generation has a story.	136	Action|Adventure|Science Fiction|Fantasy	12/15/15	5292	2015
4	Vengeance Hits Home	137	Action|Crime|Thriller	4/1/15	2947	2015

In [13]:
```
#remove duplicates
df.drop_duplicates(keep ='first', inplace=True)
df.shape
```

Out[13]: (10865, 11)

In [14]:
```
# remove movies with zero budget & revenue
temp_list=['budget', 'revenue']
df[temp_list] = df[temp_list].replace(0, np.NAN)
df.dropna(subset = temp_list, inplace = True)
rows, col = df.shape
df.shape
```

Out[14]: (3854, 11)

```
In [15]: #checking data types
         df.dtypes

Out[15]: budget            float64
         revenue           float64
         original_title     object
         cast               object
         director           object
         tagline            object
         runtime             int64
         genres             object
         release_date       object
         vote_count          int64
         release_year        int64
         dtype: object

In [16]: #changing data type for budget and revenue
         change_type=['budget', 'revenue']
         df[change_type]=df[change_type].applymap(np.int64)
         df.dtypes

Out[16]: budget             int64
         revenue            int64
         original_title    object
         cast              object
         director          object
         tagline           object
         runtime            int64
         genres            object
         release_date      object
         vote_count         int64
         release_year       int64
         dtype: object

In [17]: #change release date into datetime format
         df['release_date'] = pd.to_datetime(df['release_date'])
         df['release_date'].head()

Out[17]: 0   2015-06-09
         1   2015-05-13
         2   2015-03-18
         3   2015-12-15
         4   2015-04-01
         Name: release_date, dtype: datetime64[ns]
```

Exploratory Data Analysis

1.1.4 Research Question 1: Which movies have high and least profit?

In [18]:
```
#calculate profit
df.insert(2,'profit_earned',df['revenue']-df['budget'])
df.head()
```

Out[18]:

	budget	revenue	profit_earned	original_title	cast	director	tagline	runtime	genres	release_date	vote_count	release_year
0	150000000	1513528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	124	Action|Adventure|Science Fiction|Thriller	2015-06-09	5562	2015
1	150000000	378436354	228436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	120	Action|Adventure|Science Fiction|Thriller	2015-05-13	6185	2015
2	110000000	295238201	185238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	119	Adventure|Science Fiction|Thriller	2015-03-18	2480	2015
3	200000000	2068178225	1868178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Every generation has a story.	136	Action|Adventure|Science Fiction|Fantasy	2015-12-15	5292	2015
4	190000000	1506249360	1316249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	Vengeance Hits Home	137	Action|Crime|Thriller	2015-04-01	2947	2015

In [19]:
```
def calculate(column):
    high= df[column].idxmax()
    high_details=pd.DataFrame(df.loc[high])
    low= df[column].idxmin()
    low_details=pd.DataFrame(df.loc[low])
    info=pd.concat([high_details, low_details], axis=1)

    return info
```

```
        calculate('profit_earned')
```

Out[19]:

	1386	\
budget	237000000	
revenue	2781505847	
profit_earned	2544505847	
original_title	Avatar	
cast	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...	
director	James Cameron	
tagline	Enter the World of Pandora.	
runtime	162	
genres	Action|Adventure|Fantasy|Science Fiction	
release_date	2009-12-10 00:00:00	
vote_count	8458	
release_year	2009	

	2244
budget	425000000
revenue	11087569
profit_earned	-413912431
original_title	The Warrior's Way
cast	Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
director	Sngmoo Lee
tagline	Assassin. Hero. Legend.
runtime	100
genres	Adventure|Fantasy|Action|Western|Thriller
release_date	2010-12-02 00:00:00
vote_count	74
release_year	2010

Column with id 1386 (Avatar) shows the highest earned profit with 2.54451e+09 .

Whereas the column with id 2244 (The Warrior's Way) shows the lowest earned profit with -4.13912e+08 (which means it didn't make profit, on the contrary it lost alot of money).

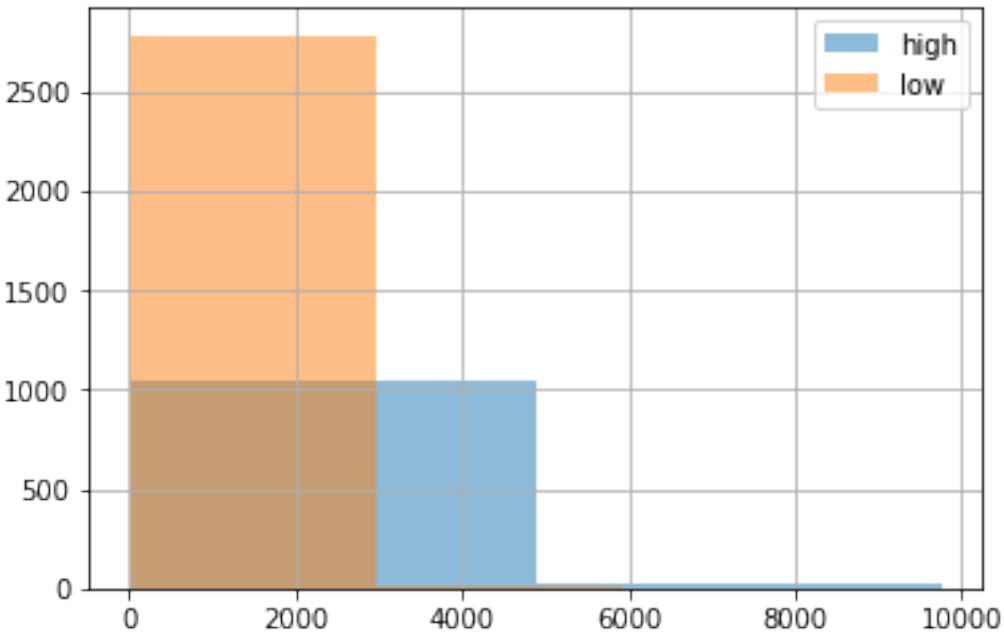
In [ ]:

1.1.5 Research Question 2 : What factors are associated with Profit Earned?

```
In [20]: df.profit_earned.mean()
         high = ( df.profit_earned > df.profit_earned.mean())
         low =  ( df.profit_earned < df.profit_earned.mean())

In [21]: # vote count vs profit earned
         df.vote_count[high].mean()

Out[21]: 1255.7467166979361

In [22]: df.vote_count[low].mean()
```

```
Out[22]: 249.35724533715924
```

```
In [23]: df.vote_count[high].hist(alpha= 0.5, bins=2, label='high');
         df.vote_count[low].hist(alpha= 0.5, bins=2, label='low');
         plt.legend();
```

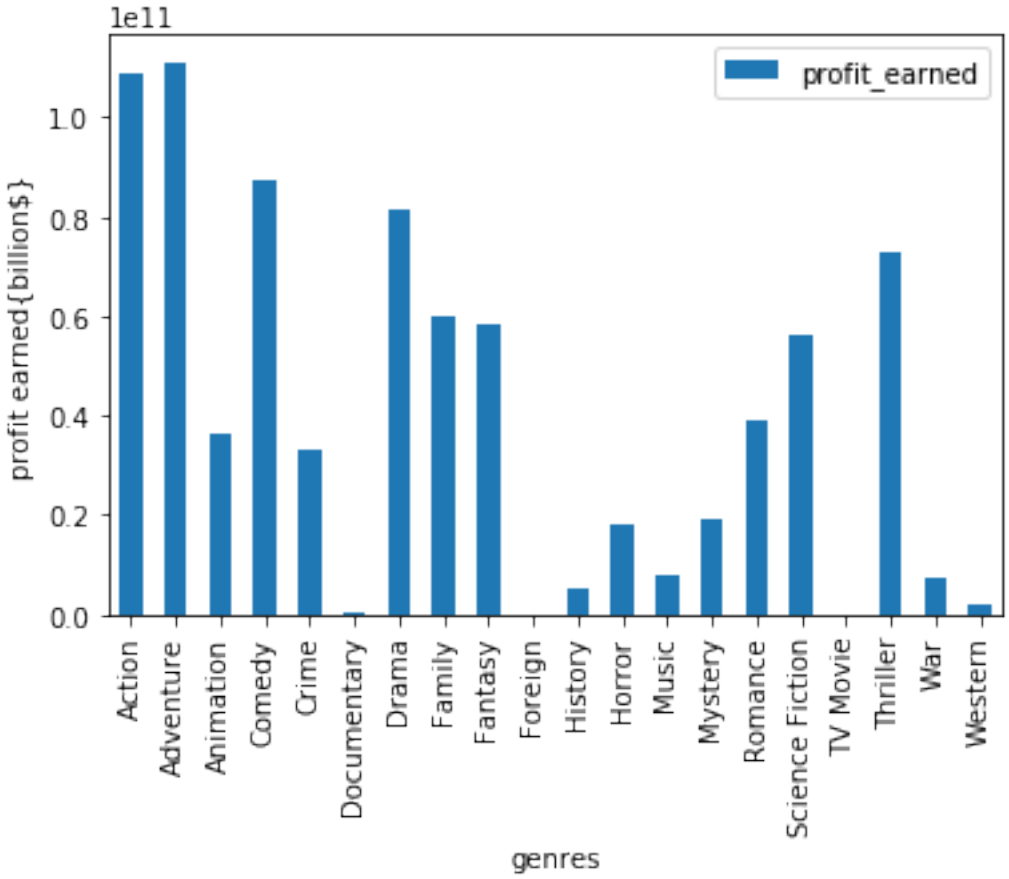


this histogram presents the there is a colrellation between the profit earned and the the votes, as the profit earned for movies are higher with repect to the vote count.

```
In [24]: # genres vs profit earned
         genres = (df.genres.str.split('|', expand=True).stack().to_frame(name='genre'))
         genres.index = genres.index.droplevel(1)
```

```
In [29]: (genres.join(df['profit_earned']).groupby('genre').sum().plot(kind='bar'))
         plt.xlabel("genres")
         plt.ylabel("profit earned{billion$}")
```

```
Out[29]: Text(0,0.5,'profit earned{billion$}')
```



this bar chart represents that the most profit earned is in adventure movies followed by action (which doesn't have this much differnce)

1.1.6 Research Question 3 : which movies with longest and shortest runtime values?

```
In [26]: def calculate(column):
             high= df[column].idxmax()
             high_details=pd.DataFrame(df.loc[high])
             low= df[column].idxmin()
             low_details=pd.DataFrame(df.loc[low])
             info=pd.concat([high_details, low_details], axis=1)

             return info

         calculate('runtime')
```

```
Out[26]:                                                    2107   \
         budget                                          18000000
         revenue                                           871279
         profit_earned                                  -17128721
```

```
original_title                                                Carlos
cast            Edgar RamÃŋrez|Alexander Scheer|Fadi Abi Samra...
director                                             Olivier Assayas
tagline                            The man who hijacked the world
runtime                                                         338
genres                               Crime|Drama|Thriller|History
release_date                                   2010-05-19 00:00:00
vote_count                                                       35
release_year                                                   2010

                                                               5162
budget                                                           10
revenue                                                           5
profit_earned                                                    -5
original_title                                          Kid's Story
cast            Clayton Watson|Keanu Reeves|Carrie-Anne Moss|K...
director                                       Shinichiro Watanabe
tagline                                                        NaN
runtime                                                         15
genres                                  Science Fiction|Animation
release_date                                   2003-06-02 00:00:00
vote_count                                                       16
release_year                                                   2003
```

Column with id 2107 (Carlos) shows the highest runtime wiht 338 minutes.
Whereas the column with id 5162 (Kid's Story) shows the lowest runtime with 15 minutes.

1.1.7 Research Question 4 : Which movies with largest and lowest budgets?

```
In [28]: def calculate(column):
             high= df[column].idxmax()
             high_details=pd.DataFrame(df.loc[high])
             low= df[column].idxmin()
             low_details=pd.DataFrame(df.loc[low])
             info=pd.concat([high_details, low_details], axis=1)

             return info

         calculate('budget')
```

```
Out[28]:                                                          2244  \
         budget                                                425000000
         revenue                                                11087569
         profit_earned                                       -413912431
         original_title                                 The Warrior's Way
         cast            Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
         director                                              Sngmoo Lee
         tagline                                 Assassin. Hero. Legend.
```

```
runtime                                                       100
genres                 Adventure|Fantasy|Action|Western|Thriller
release_date                                  2010-12-02 00:00:00
vote_count                                                     74
release_year                                                 2010

                                                             2618
budget                                                          1
revenue                                                       100
profit_earned                                                  99
original_title                                        Lost & Found
cast           David Spade|Sophie Marceau|Ever Carradine|Step...
director                                               Jeff Pollack
tagline        A comedy about a guy who would do anything to ...
runtime                                                        95
genres                                              Comedy|Romance
release_date                                  1999-04-23 00:00:00
vote_count                                                     14
release_year                                                 1999
```

Column with id 2244 (The Warrior's Way) shows the highest budget with $425000000
Whereas the column with id 2618 (Lost & Found) shows the lowest budget with $1
Conclusions

Avatar is the highest earned profit with $2.54451e+09

The Warrior's Way shows the lowest earned profit with $-4.13912e+08 (which means it didn't make profit, on the contrary it lost alot of money).

Carlos shows the highest runtime wiht 338 minutes.

Kid's Story shows the lowest runtime with 15 minutes.

The Warrior's Way) shows the highest budget with $425000000

(Lost & Found) shows the lowest budget with $1

1.2 Submitting your Project

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

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

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

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