Capstone Project #2 –
Recommendation System for movies

Problem:
It has been estimated that there are approximately 500,000 movies currently in existence.There are currently over 135,000 cinema screens worldwide on which approximately 8,000 movies are released internationally each year. In the evolutionary view, this situation creates `selection pressure’ on individual movies, as not all movies are equally popular; industry-supply and audience-demand for specific movies (in fact, for specific movie stories) are asymmetrical(or at least seems so), as most movies lose money. The audience will not be spending their time watching every movie available to them, they will pick randomly something to watch. If a person picks a movie and does not enjoy it there will be no positive word-of-mouth. This doesn’t necessarily mean the movie was bad, it might mean it was not interesting to that individual.

This is where the recommendation system is helpful. Recommendation system helps the user find items of their interest and Helps the item provider to deliver their items to the right user. It increases revenues for business through increased consumption.Movie Recommendation systems are becoming increasingly important in today’s extremely busy world as it helps audiences to make the right choices, without having to expend their cognitive resources.

In this project goal is to Build Content Based and Collaborative Filtering Based Recommendation Engines for movies.

Clients:

Producers and distributors of movies.

Data: These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

https://www.kaggle.com/rounakbanik/the-movies-dataset

Acknowledgements: This dataset is an ensemble of data collected from TMDB and GroupLens.

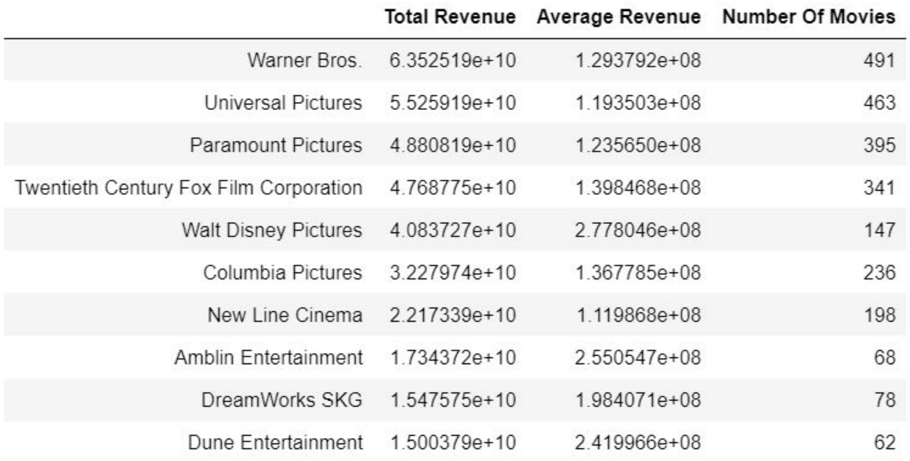
The Movie Details, Credits and Keywords have been collected from the TMDB Open API. This product uses the TMDb API but is not endorsed or certified by TMDb. Their API also provides access to data on many additional movies, actors and actresses, crew members, and TV shows.

The dataset has a record of 45466 movies with 24 columns(features). After looking closer to the elements, I observed more than 50% of its data with null values. As revenue is the feature that I am interested in, I checked for Nan values in that column. 38052 records of the movies have recorded revenue of 0, indicating that we do not have valuable information about the total revenue for these movies. Although this forms most of the movies available to us, we will still use revenue as a vital feature to advance from the remaining 7414 movies. The budget feature had some unclean values that make Pandas assign it as a generic object. I converted this into a numeric variable and replaced all the non-numeric values with NaN. Extracted feature "Release Year" from "Release Date." "Release Year" is the year in which the movie was released. I calculated the Net Profit/Loss using features "Revenue" and "budget." This feature is incredibly insightful as it will give us a more accurate picture of a movie's financial success. Presently, our data will not judge if a 200 million budget movie that earned 100 million did better than a 50,000 budget movie taking in 200,000. This feature will be able to capture that information. A value > 1 would indicate profit, whereas a return value < 1 would indicate a loss. A few features like adult, id, original_title, poster_path, video does not provide useful information. I dropped these features from the data frame.

By exploring clean Movie metadata we would try to answer below questions.

1. Which Production companies make most money in movie business?
2. Which movies are more popular?
3. Which movies have been most voted by TMDB voters?
4. Which movies are most Critically Acclaimed?
5. Does Release Month play a significant role in determining the success and the revenue generated by a particular movie?
6. Which are the most expensive movies of all time?
7. How strong a correlation does the budget hold with the revenue?
8. Which are the Highest Grossing Films of All Time?
9. Which are the least and the most successful movies of all time?

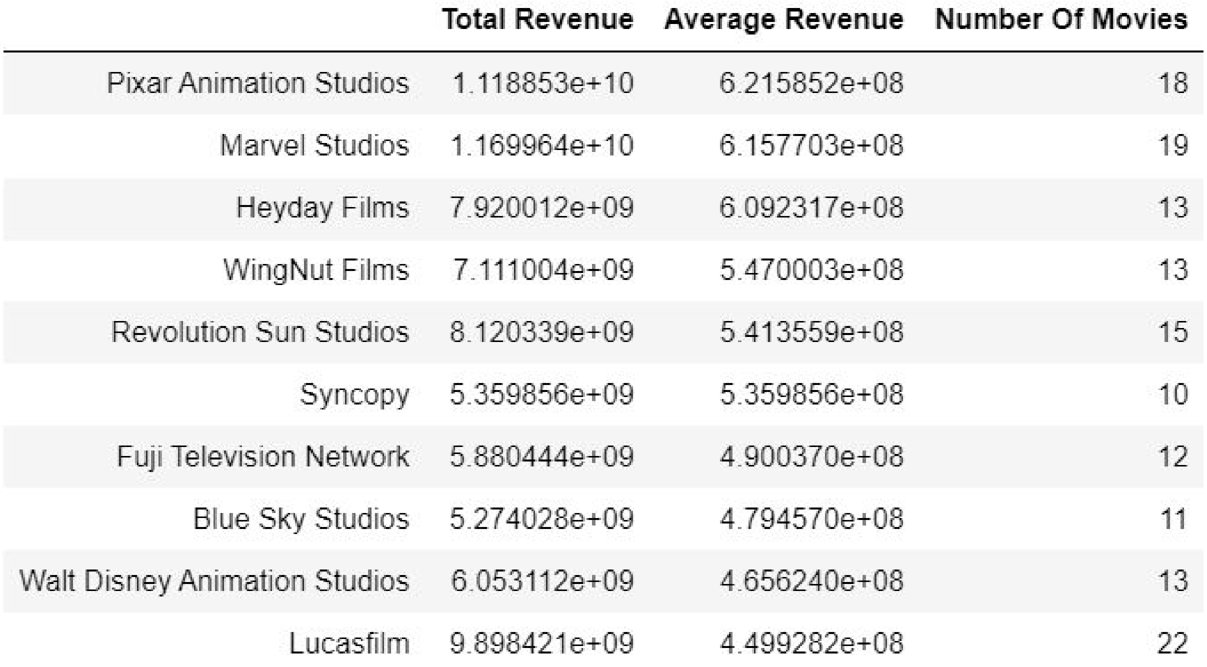
Let us find out which production companies have earned the most money from the movie making business.



	Total Revenue	Average Revenue	Number Of Movies
Warner Bros.	6.352519e+10	1.293792e+08	491
Universal Pictures	5.525919e+10	1.193503e+08	463
Paramount Pictures	4.880819e+10	1.235650e+08	395
Twentieth Century Fox Film Corporation	4.768775e+10	1.398468e+08	341
Walt Disney Pictures	4.083727e+10	2.778046e+08	147
Columbia Pictures	3.227974e+10	1.367785e+08	236
New Line Cinema	2.217339e+10	1.119868e+08	198
Amblin Entertainment	1.734372e+10	2.550547e+08	68
DreamWorks SKG	1.547575e+10	1.984071e+08	78
Dune Entertainment	1.500379e+10	2.419966e+08	62

Warner Bros is the highest-earning production company of all time, earning a staggering 63.5 billion dollars from close to 500 movies. Universal Pictures and Paramount Pictures are the second and the third highest-earning companies with 55 billion dollars and 48 billion dollars in revenue.

As we are aware, Warner Bros and Universal Pictures are bigger studios compared to others on the list. Thus it would be more appropriate to look at the average revenue of studios. We will consider studios that have produced at least ten movies.

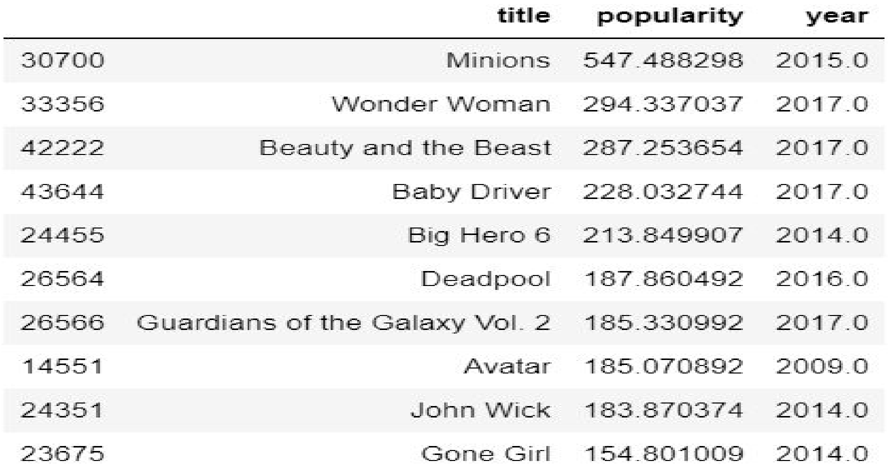


	Total Revenue	Average Revenue	Number Of Movies
Pixar Animation Studios	1.118853e+10	6.215852e+08	18
Marvel Studios	1.169964e+10	6.157703e+08	19
Heyday Films	7.920012e+09	6.092317e+08	13
WingNut Films	7.111004e+09	5.470003e+08	13
Revolution Sun Studios	8.120339e+09	5.413559e+08	15
Syncopy	5.359856e+09	5.359856e+08	10
Fuji Television Network	5.880444e+09	4.900370e+08	12
Blue Sky Studios	5.274028e+09	4.794570e+08	11
Walt Disney Animation Studios	6.053112e+09	4.656240e+08	13
Lucasfilm	9.898421e+09	4.499282e+08	22

Pixar Animation Studios has produced the most successful movies, on average. This is no surprise, though Pixar has made just 18 movies. It includes the Toy Story Franchise, Up, Finding Nemo, Inside Out, Wall-E, Ratatouille, Cars Franchise, Incredibles, etc., which the audience has received well across the world as well as critically acclaimed. Marvel Studios, with an average gross of 615 million dollars, comes in second.

To answer questions like Which movies are more popular? And Which moves have been most voted by TMDB voters? We need to look at features popularity, vote_count, and vote_average. As these features are of type object, I converted them to float type.

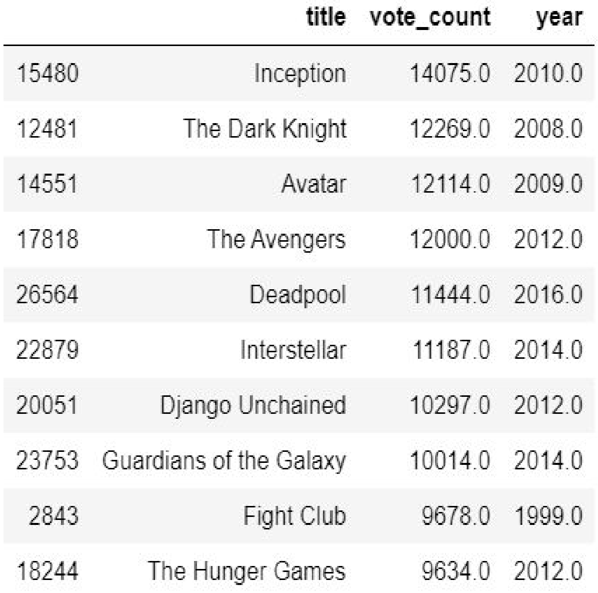
First, let us find the answer to Which movies are the most popular?



	title	popularity	year
30700	Minions	547.488298	2015.0
33356	Wonder Woman	294.337037	2017.0
42222	Beauty and the Beast	287.253654	2017.0
43644	Baby Driver	228.032744	2017.0
24455	Big Hero 6	213.849907	2014.0
26564	Deadpool	187.860492	2016.0
26566	Guardians of the Galaxy Vol. 2	185.330992	2017.0
14551	Avatar	185.070892	2009.0
24351	John Wick	183.870374	2014.0
23675	Gone Girl	154.801009	2014.0

"Minions" is the most popular movie by the TMDB Popularity Score. I guess no arguing about liking cute minions. It is also interesting to note that Minions' title characters Talk less than a few words in the movie and yet most popular. Wonder Woman and Beauty and the Beast come in second and third respectively, both of which are women-centric stories almost tieing up for the second place.

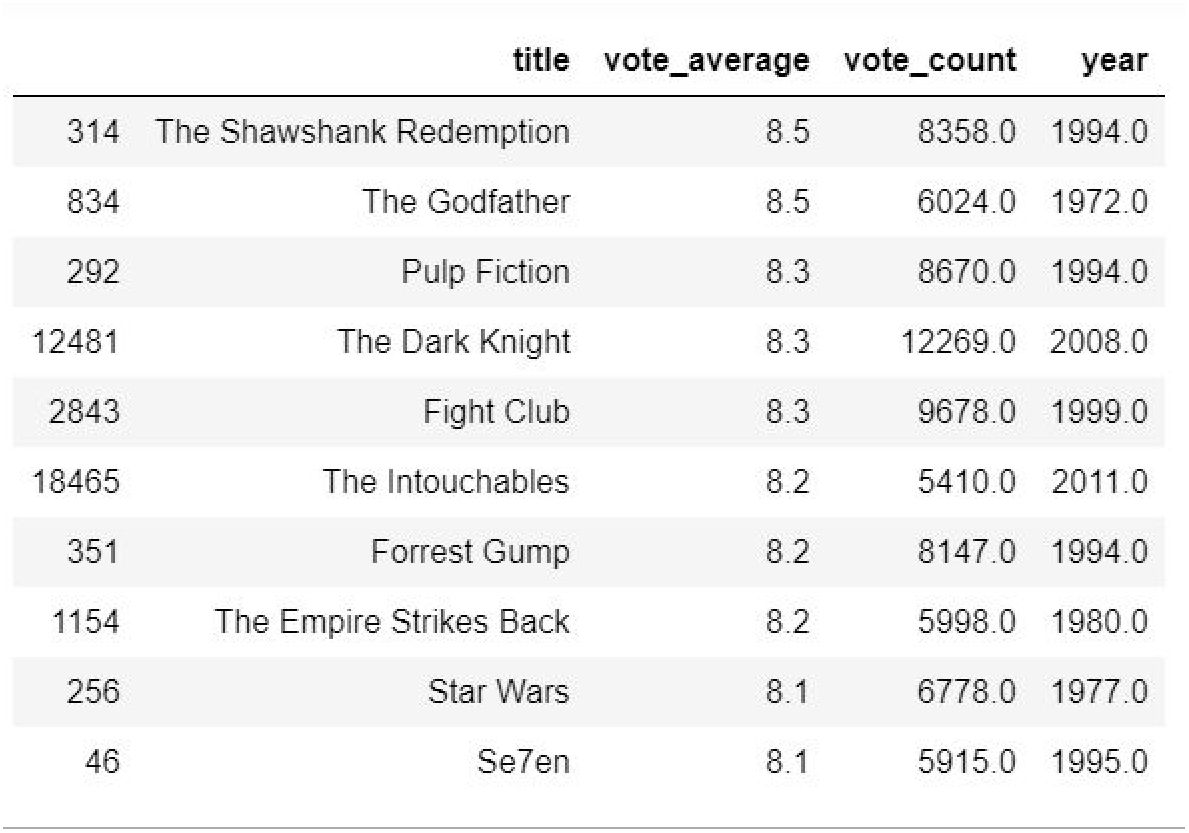
Now that we know which movies are most popular, let us see which movies people most voted on in TMDB.



	title	vote_count	year
15480	Inception	14075.0	2010.0
12481	The Dark Knight	12269.0	2008.0
14551	Avatar	12114.0	2009.0
17818	The Avengers	12000.0	2012.0
26564	Deadpool	11444.0	2016.0
22879	Interstellar	11187.0	2014.0
20051	Django Unchained	10297.0	2012.0
23753	Guardians of the Galaxy	10014.0	2014.0
2843	Fight Club	9678.0	1999.0
18244	The Hunger Games	9634.0	2012.0

Inception and The Dark Knight, two critically acclaimed movies, are at the top of our chart. It is interesting to note that Christopher Nolan directed both of these.

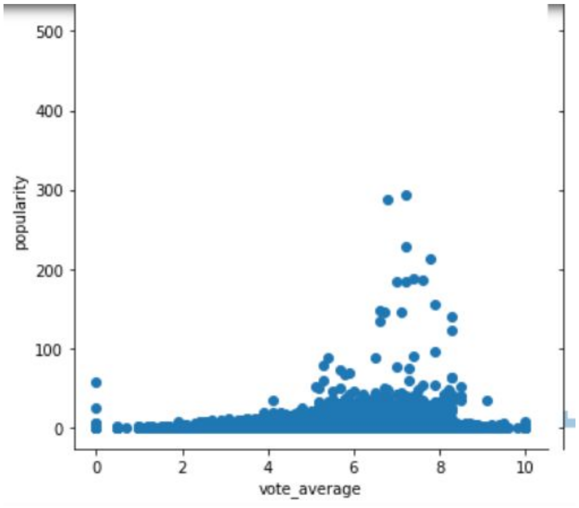
Let us check what the most critically acclaimed movies as per TMDB are. We will only consider those movies with more than 5000 votes (similar to IMDB's criteria of 5000 options in selecting its top 250).



	title	vote_average	vote_count	year
314	The Shawshank Redemption	8.5	8358.0	1994.0
834	The Godfather	8.5	6024.0	1972.0
292	Pulp Fiction	8.3	8670.0	1994.0
12481	The Dark Knight	8.3	12269.0	2008.0
2843	Fight Club	8.3	9678.0	1999.0
18465	The Intouchables	8.2	5410.0	2011.0
351	Forrest Gump	8.2	8147.0	1994.0
1154	The Empire Strikes Back	8.2	5998.0	1980.0
256	Star Wars	8.1	6778.0	1977.0
46	Se7en	8.1	5915.0	1995.0

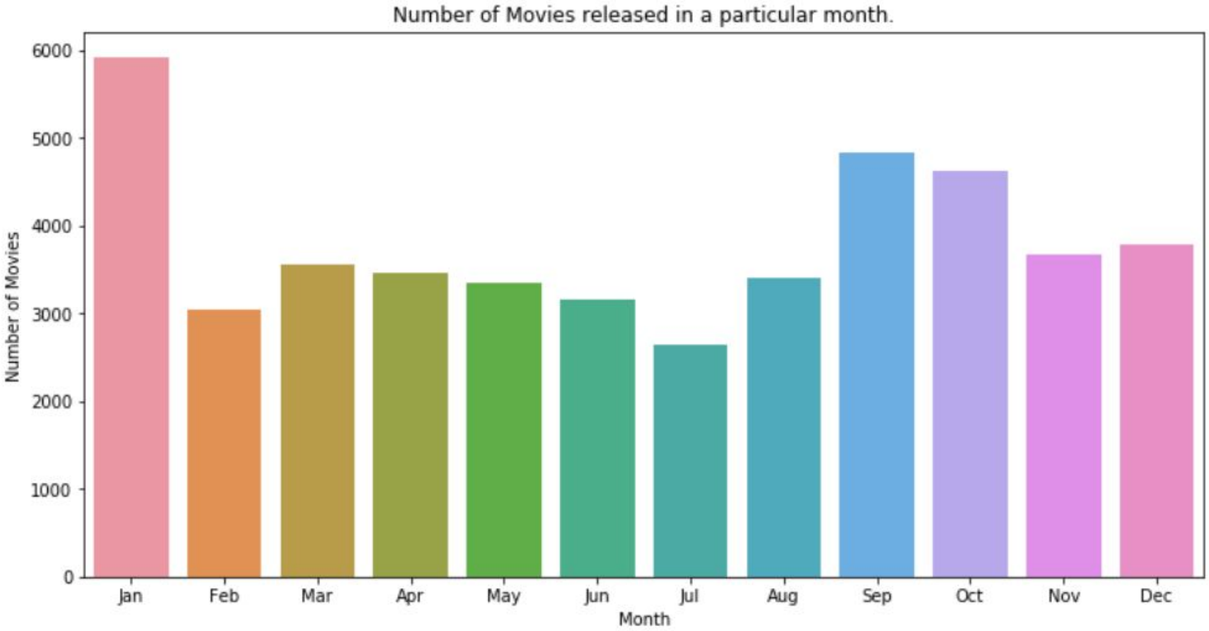
The Shawshank Redemption and The Godfather are the two most critically acclaimed movies in the TMDB Database. The Academy Awards is going to agree with me on this.

Do popularity and vote average share a tangible relationship? In other words, is there a strong positive correlation between these two quantities?

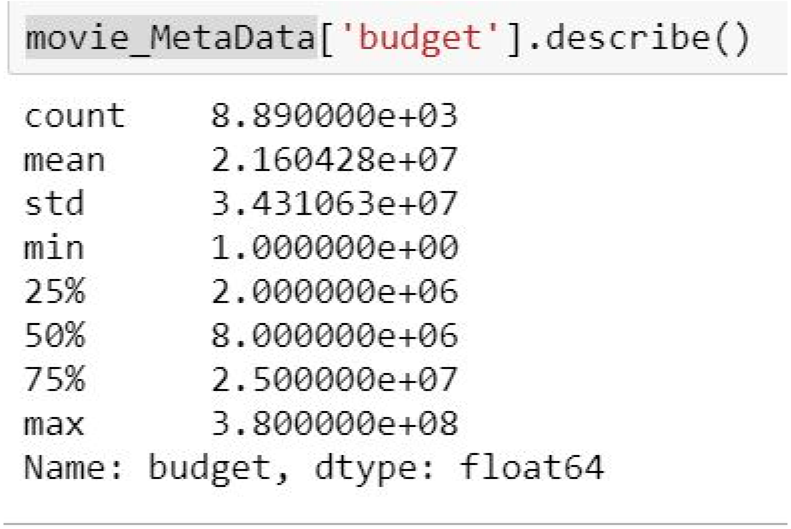


Surprisingly, the Pearson Coefficient of the two quantities, is 0.154, suggesting no definite correlation. In other words, popularity and vote average and independent quantities.

Release Dates can often play a significant role in determining the success and the revenue generated by a particular movie. This section will try and gain insights about release dates in terms of months. We have already constructed the year feature in our preliminary data wrangling step. Let us now extract the month for each movie with a release date.



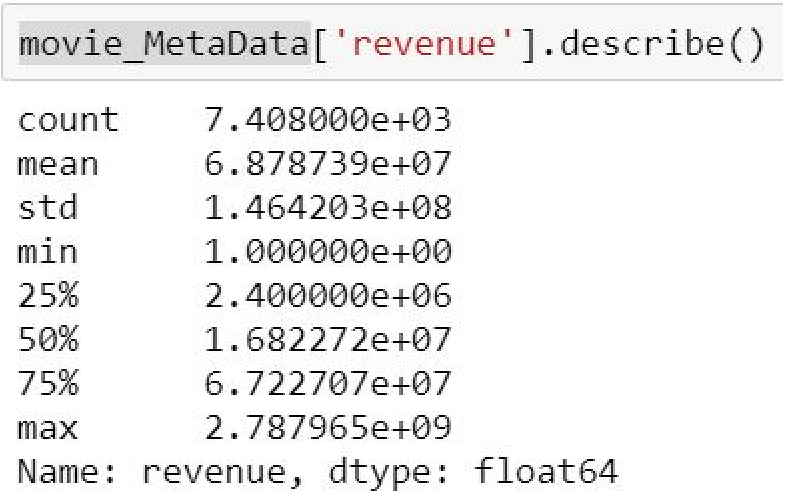
It appears that January is the most popular month when it comes to movie releases. This is also known as the dump month in Hollywood circles when the dozen release subpar movies.



```
movie_MetaData['budget'].describe()
```

```
count    8.890000e+03
mean     2.160428e+07
std      3.431063e+07
min      1.000000e+00
25%      2.000000e+06
50%      8.000000e+06
75%      2.500000e+07
max      3.800000e+08
Name: budget, dtype: float64
```

A film's mean budget is 21.6 million dollars, whereas the median budget is far smaller at 8 million dollars. This strongly suggests the mean being influenced by outliers.

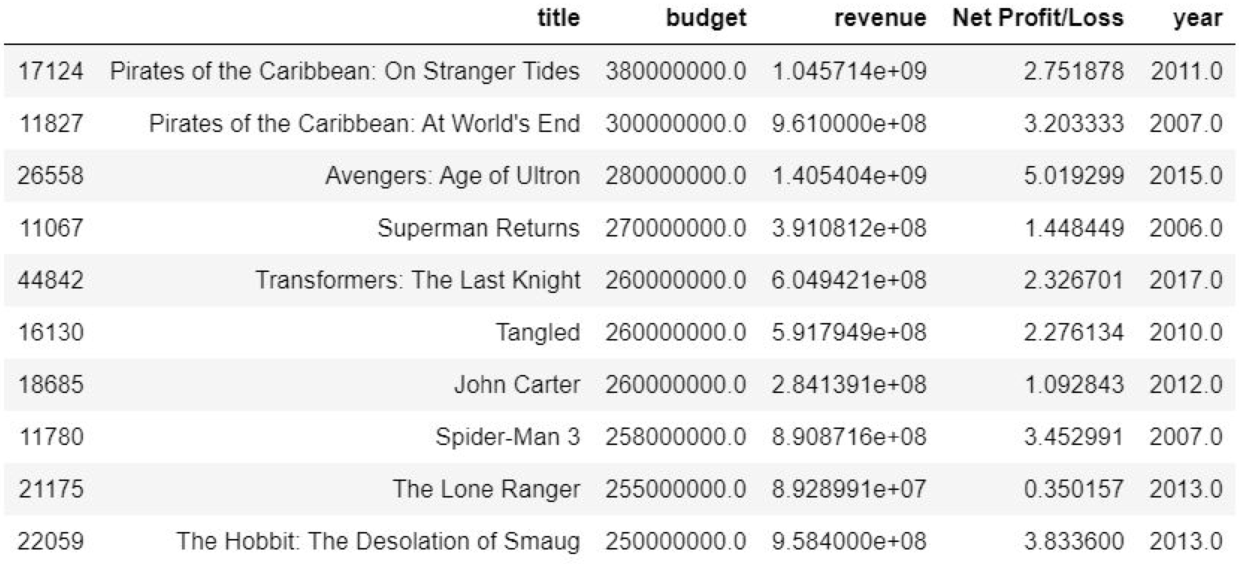


```
movie_MetaData['revenue'].describe()
```

```
count    7.408000e+03
mean     6.878739e+07
std      1.464203e+08
min      1.000000e+00
25%      2.400000e+06
50%      1.682272e+07
75%      6.722707e+07
max      2.787965e+09
Name: revenue, dtype: float64
```

The mean gross of a movie is 68.7 million dollars, whereas the median gross is much lower at 16.8 million dollars, suggesting the skewed nature of revenue. The most insufficient revenue generated by a movie is just 1 dollar, whereas the highest-grossing film of all time has raked in an astonishing *2.78 billion dollars.

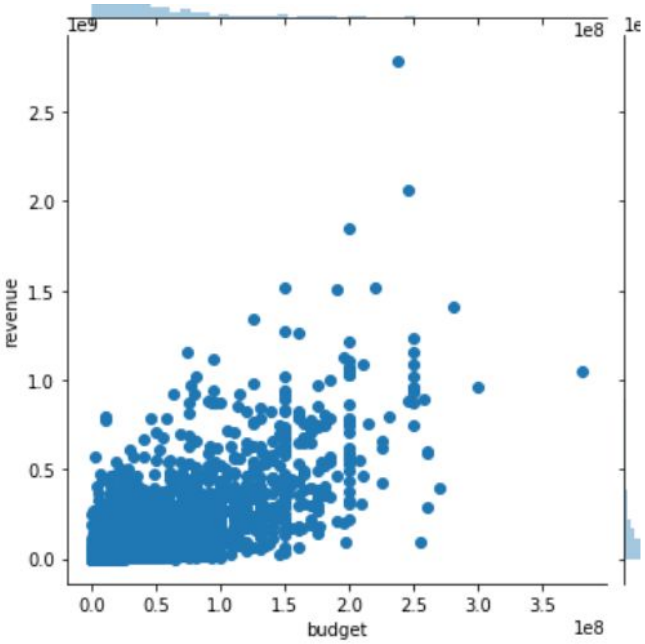
Let us take a look at the most expensive movies of all time and the revenue & returns.



	title	budget	revenue	Net Profit/Loss	year
17124	Pirates of the Caribbean: On Stranger Tides	380000000.0	1.045714e+09	2.751878	2011.0
11827	Pirates of the Caribbean: At World's End	300000000.0	9.610000e+08	3.203333	2007.0
26558	Avengers: Age of Ultron	280000000.0	1.405404e+09	5.019299	2015.0
11067	Superman Returns	270000000.0	3.910812e+08	1.448449	2006.0
44842	Transformers: The Last Knight	260000000.0	6.049421e+08	2.326701	2017.0
16130	Tangled	260000000.0	5.917949e+08	2.276134	2010.0
18685	John Carter	260000000.0	2.841391e+08	1.092843	2012.0
11780	Spider-Man 3	258000000.0	8.908716e+08	3.452991	2007.0
21175	The Lone Ranger	255000000.0	8.928991e+07	0.350157	2013.0
22059	The Hobbit: The Desolation of Smaug	250000000.0	9.584000e+08	3.833600	2013.0

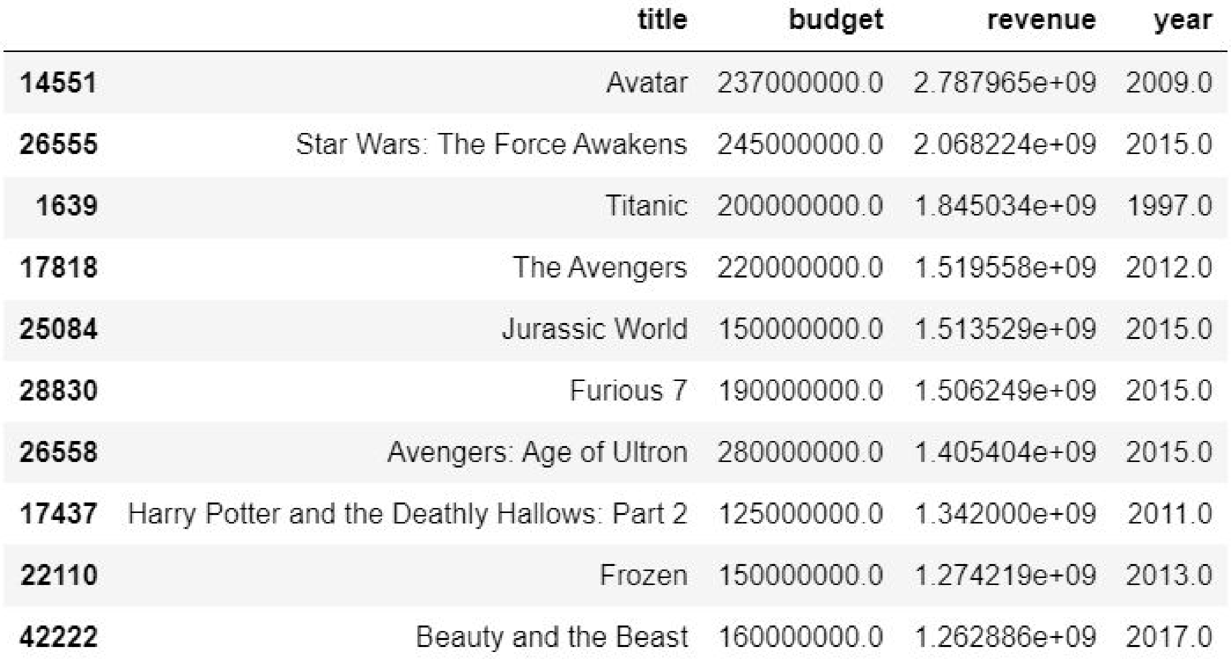
Two Pirates of the Caribbean films occupy the top spots in this list with a staggering budget of over 300 million dollars. All the top 10 most expensive movies made a profit on their investment except for The Lone Ranger, which managed to recoup less than 35% of its investment, taking in a paltry 90 million dollars on a 255 million dollar budget.

How strong a correlation does the budget hold with the revenue? A stronger correlation would directly imply more accurate forecasts.



The scatterplot above shows a positive correlation between budget and revenue.

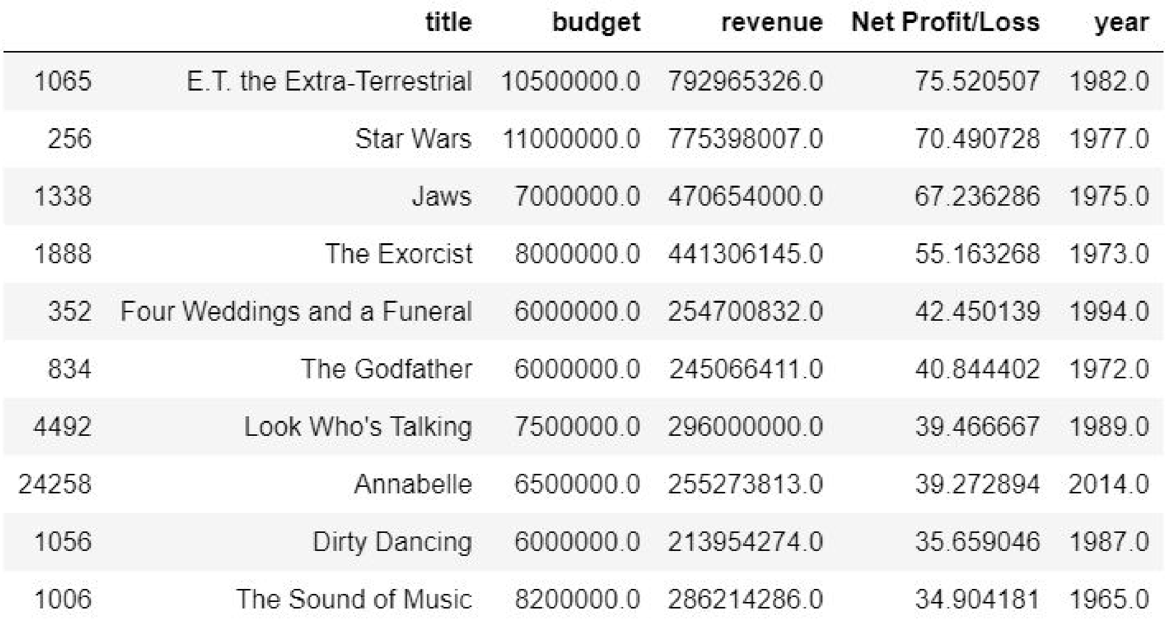
Let us see which are Highest Grossing Films of All Time?



	title	budget	revenue	year
14551	Avatar	237000000.0	2.787965e+09	2009.0
26555	Star Wars: The Force Awakens	245000000.0	2.068224e+09	2015.0
1639	Titanic	200000000.0	1.845034e+09	1997.0
17818	The Avengers	220000000.0	1.519558e+09	2012.0
25084	Jurassic World	150000000.0	1.513529e+09	2015.0
28830	Furious 7	190000000.0	1.506249e+09	2015.0
26558	Avengers: Age of Ultron	280000000.0	1.405404e+09	2015.0
17437	Harry Potter and the Deathly Hallows: Part 2	125000000.0	1.342000e+09	2011.0
22110	Frozen	150000000.0	1.274219e+09	2013.0
42222	Beauty and the Beast	160000000.0	1.262886e+09	2017.0

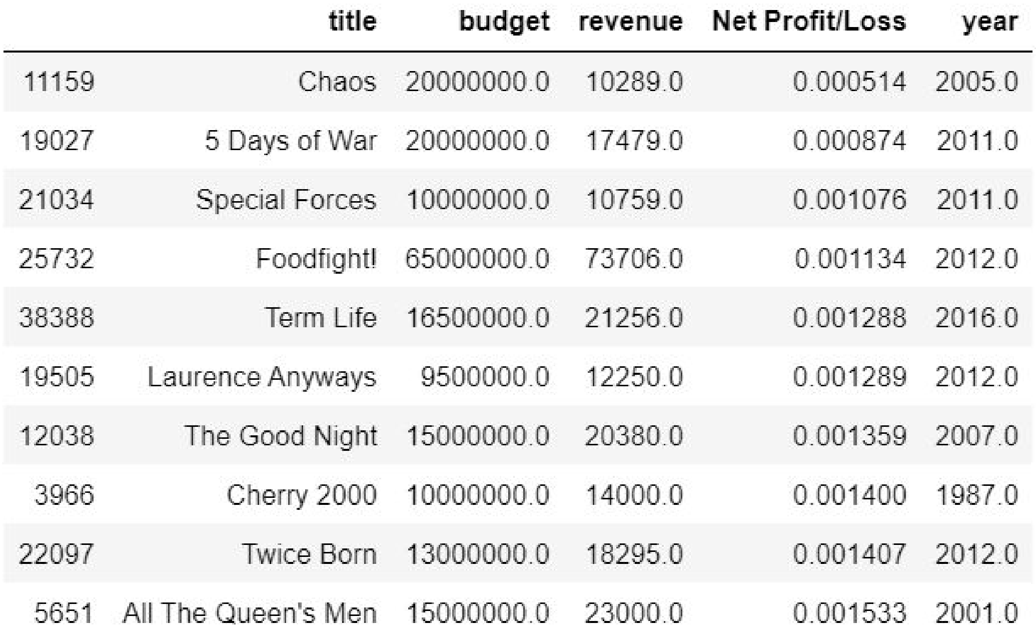
The world of movies broke the 1 billion dollar mark in 1997 with the release of Titanic. It took another 12 years to break the 2 billion dollar mark with Avatar. James Cameron directed both these movies.

The highest-grossing movie does not necessarily mean the movie made the highest profit of all. Let us check the least and the most successful movies of all time. To do this, we will only consider those movies which have a budget greater than 5 million dollars.



	title	budget	revenue	Net Profit/Loss	year
1065	E.T. the Extra-Terrestrial	10500000.0	792965326.0	75.520507	1982.0
256	Star Wars	11000000.0	775398007.0	70.490728	1977.0
1338	Jaws	7000000.0	470654000.0	67.236286	1975.0
1888	The Exorcist	8000000.0	441306145.0	55.163268	1973.0
352	Four Weddings and a Funeral	6000000.0	254700832.0	42.450139	1994.0
834	The Godfather	6000000.0	245066411.0	40.844402	1972.0
4492	Look Who's Talking	7500000.0	296000000.0	39.466667	1989.0
24258	Annabelle	6500000.0	255273813.0	39.272894	2014.0
1056	Dirty Dancing	6000000.0	213954274.0	35.659046	1987.0
1006	The Sound of Music	8200000.0	286214286.0	34.904181	1965.0

E.T. the Extra-Terrestrial is the most successful movie! It is interesting to note that most of the successful movies in the top 10 list are released between 1965 - 1989.



	title	budget	revenue	Net Profit/Loss	year
11159	Chaos	20000000.0	10289.0	0.000514	2005.0
19027	5 Days of War	20000000.0	17479.0	0.000874	2011.0
21034	Special Forces	10000000.0	10759.0	0.001076	2011.0
25732	Foodfight!	65000000.0	73706.0	0.001134	2012.0
38388	Term Life	16500000.0	21256.0	0.001288	2016.0
19505	Laurence Anyways	9500000.0	12250.0	0.001289	2012.0
12038	The Good Night	15000000.0	20380.0	0.001359	2007.0
3966	Cherry 2000	10000000.0	14000.0	0.001400	1987.0
22097	Twice Born	13000000.0	18295.0	0.001407	2012.0
5651	All The Queen's Men	15000000.0	23000.0	0.001533	2001.0

Chaos is the least successful movie. We can observe here that most of the movies listed in top 10 are released between 2000-2012! So is it that older movies were more successful than newer ones? We cannot certainly say so as these figures have not been adjusted for inflation.

Below is the link to the Github repository of jupyter notebook files with Data wrangling and statistical analysis code.

https://github.com/lasyabheemendra/Sprigboard-DatascienceProjects/blob/master/Capstone2_Recommendation%20System%20for%20movies/Data%20Analysis%20of%20Movie%20Data.ipynb

PowerPoint presentation of the above report can be found in the link below.

https://github.com/lasyabheemendra/Sprigboard-DatascienceProjects/blob/master/Capstone2_Recommendation%20System%20for%20movies/Recommendation%20system.pptx