Exploratory Data Analysis

Understanding the Dataset

The dataset above was obtained through the TMDB API. The movies available in this dataset are in correspondence with the movies that are listed in the **MovieLens Latest Full Dataset** comprising of 26 million ratings on 45,000 movies from 27,000 users. Let us have a look at the features that are available to us.

Features

- **adult:** Indicates if the movie is X-Rated or Adult.
- **belongs_to_collection:** A stringified dictionary that gives information on the movie series the particular film belongs to.
- **budget:** The budget of the movie in dollars.
- **genres:** A stringified list of dictionaries that list out all the genres associated with the movie.
- **homepage:** The Official Homepage of the move.
- **id:** The ID of the move.
- **imdb_id:** The IMDB ID of the movie.
- **original_language:** The language in which the movie was originally shot in.
- **original_title:** The original title of the movie.
- **overview:** A brief blurb of the movie.
- **popularity:** The Popularity Score assigned by TMDB.
- **poster_path:** The URL of the poster image.
- **production_companies:** A stringified list of production companies involved with the making of the movie.
- **production_countries:** A stringified list of countries where the movie was shot/produced in.
- **release_date:** Theatrical Release Date of the movie.
- **revenue:** The total revenue of the movie in dollars.
- **runtime:** The runtime of the movie in minutes.
- **spoken_languages:** A stringified list of spoken languages in the film.
- **status:** The status of the movie (Released, To Be Released, Announced, etc.)
- **tagline:** The tagline of the movie.
- **title:** The Official Title of the movie.
- **video:** Indicates if there is a video present of the movie with TMDB.
- **vote_average:** The average rating of the movie.
- **vote_count:** The number of votes by users, as counted by TMDB.

There are a total of **45,466 movies** with **24 features**. Most of the features have very few NaN values (apart from **homepage** and **tagline**). We will attempt at cleaning this dataset to a form suitable for analysis in the next section.

Data Wrangling

The data that was originally obtained was in the form of a JSON File. This was converted manually into a CSV file to arrive at an input that could be loaded into a Pandas DataFrame effortlessly. In other words, the dataset we have in our hands is already relatively clean. We will however attempt at learning more about our features and performing appropriate wrangling steps to arrive at a form that is more suitable for analysis.

Let us start by removing the features that are not useful to us.

The original title refers to the title of the movie in the native language in which the movie was shot. As such, I will prefer using the translated, Anglicized name in this analysis and hence, will drop the original titles altogether. We will be able to deduce if the movie is a foreign language film by looking at the ***original_language*** feature so no tangible information is lost in doing so.

We see that the majority of the movies have a recorded revenue of 0. This indicates that we do not have information about the total revenue for these movies. Although this forms the majority of the movies available to us, we will still use revenue as an extremely important feature going forward from the remaining 7000 moves.

The ***budget feature*** has some unclean values that makes Pandas assign it as a generic object. We proceed to convert this into a numeric variable and replace all the non-numeric values with NaN. Finally, as with budget, we will convert all the values of 0 with NaN to indicate the absence of information regarding budget.

As we move forward trying to answer certain questions, we will have to construct several features suitable for that particular query. For now, we will construct two very important features:

- **year:** The year in which the movie was released.
- **return:** The ratio of revenue to budget.

The **return** feature is extremely insightful as it will give us a more accurate picture of the financial success of a movie. Presently, our data will not be able to judge if a <latex iconid='0'>100 million did better than a <latex iconid='1'>200,000. This feature will be able to capture that information.

A return value > 1 would indicate profit whereas a return value < 1 would indicate a loss.

We have close to **5000 movies** for which we have data on revenue and budget ratio. This is close to **10% of the entire dataset**. Although this may seem small, this is enough to perform very useful analysis and discover interesting insights about the world of movies.

There are close to **0 adult movies** in this dataset. The **adult** feature therefore is not of much use to us and can be safely dropped.

Exploratory Data Analysis

Title and Overview Wordclouds

Are there certain words that figure more often in Movie Titles and Movie Blurbs? I suspect there are some words which are considered more potent and considered more worthy of a title. Let us find out!

The word **Love** is the most commonly used word in movie titles. **Girl**, **Day** and **Man** are also among the most commonly occuring words. I think this encapsulates the idea of the ubiquitious presence of romance in movies pretty well.

Life is the most commonly used word in Movie titles. **One** and **Find** are also popular in Movie Blurbs. Together with **Love**, **Man** and **Girl**, these wordclouds give us a pretty good idea of the most popular themes present in movies.

Production Countries

The Full MovieLens Dataset consists of movies that are overwhelmingly in the English language (more than 31000). However, these movies may have shot in various locations around the world. It would be interesting to see which countries serve as the most popular destinations for shooting movies by filmmakers, especially those in the United States of America and the United Kingdom.

	num_movies	country
0	21153	United States of America
1	4094	United Kingdom
2	3940	France
3	2254	Germany
4	2169	Italy
5	1765	Canada
6	1648	Japan
7	964	Spain
8	912	Russia
9	828	India

Unsurprisingly, the **United States** is the most popular destination of production for movies given that our dataset largely consists of English movies. **Europe** is also an extremely popular location with the UK, France, Germany and Italy in the top 5. **Japan** and **India** are the most popular Asian countries when it comes to movie production.

Franchise Movies

Let us now have a brief look at Franchise movies. I was curious to discover the longest running and the most successful franchises among many other things. Let us wrangle our data to find out!

Highest Grossing Movie Franchises

The **Harry Potter** Franchise is the most successful movie franchise raking in more than 7.707 billion dollars from 8 movies. The **Star Wars** Movies come in a close second with a 7.403 billion dollars from 8 movies too. **James Bond** is third but the franchise has significantly more movies compared to the others in the list and therefore, a much smaller average gross.

Most Successful Movie Franchises (by Average Gross)

We will use the average gross per movie to gauge the success of a movie franchise. However, this is not a very potent metric as the revenues in this dataset have not been adjusted for inflation. Therefore, revenue statistics will tend to strongly favor franchises in the recent times.

Longest Running Franchises

Finally, in this subsection, let us take a look at the franchises which have stood the test of time and have managed to deliver the largest number of movies under a single banner. This metric is potent in the way that it isn't affected by inflation. However, this does not imply that successful movie franchises tend to have more movies. Some franchises, such as Harry Potter, have a predefined storyline and it wouldn't make sense to produce more movies despite its enormous success.

The **James Bond** Movies is the largest franchise ever with over 26 movies released under the banner. **Friday the 13th** and **Pokemon** come in at a distant second and third with 12 and 11 movies respectively.

Production Companies

Highest Earning Production Companies

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

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 **Paramaount Pictures** are the second and the third highest earning companies with 55 billion dollars and 48 billion dollars in revenue respectively.

Most Succesful Production Companies

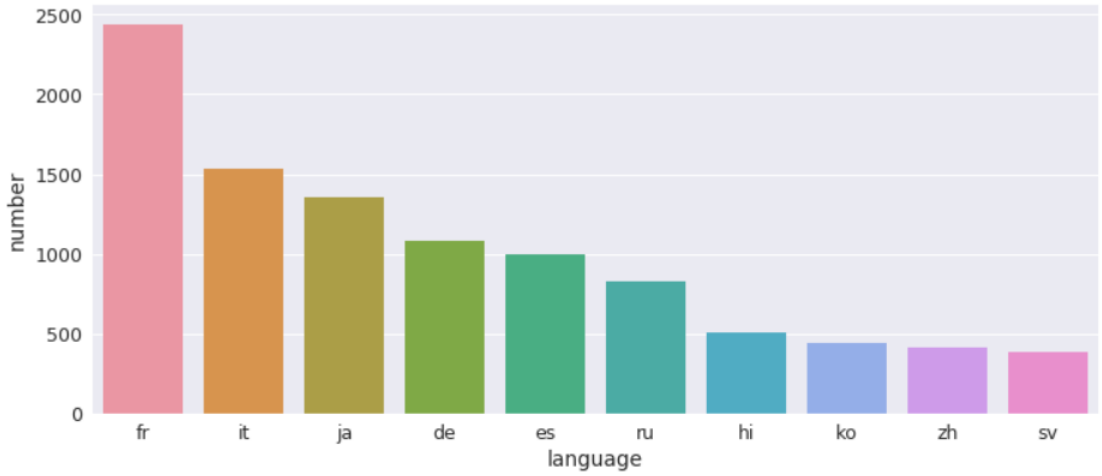
Which production companies produce the most succesful movies on average? Let us find out. We will only consider those companies that have made at least 15 movies.

Pixar Animation Studios has produced the most successful movies, on average. This is not surprising considering the amazing array of movies that it has produced in the last few decades: Up, Finding Nemo, Inside Out, Wall-E, Ratatouille, the Toy Story Franchise, Cars Franchise, etc. **Marvel Studios** with an average gross of 615 million dollars comes in second with movies such as Iron Man and The Avengers under its banner.

Original Language

In this section, let us look at the languages of the movies in our dataset. From the production countries, we have already deduced that the majority of the movies in the dataset are English. Let us see what the other major languages represented are.

There are over 93 languages represented in our dataset. As we had expected, English language films form the overwhelmingly majority. French and Italian movies come at a very distant second and third respectively. Let us represent the most popular languages (apart from English) in the form of a bar plot.



As mentioned earlier, **French** and **Italian** are the most commonly occurring languages after English. **Japanese** and **Hindi** form the majority as far as Asian Languages are concerned.

Popularity, Vote Average and Vote Count

In this section, we will work with metrics provided to us by TMDB users. We will try to gain a deeper understanding of the popularity, vote average and vote count features and try and deduce any relationships between them as well as other numeric features such as budget and revenue.

Let us examine the summary statistics and the distribution of each feature one by one.

The Popularity score seems to be an extremely skewed quentity with a mean of only **2.9** but maximum values reaching as high as 547, which is almost 1800% greater than the mean. However, as can be seen from the distribution plot, almost all movies have a popularity score less than 10 (the 75th percentile is at 3.678902).

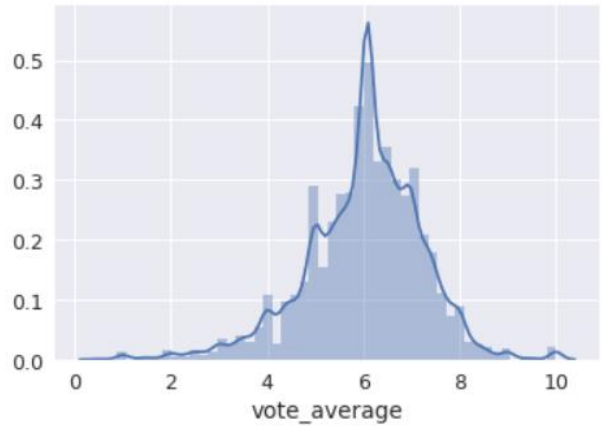
Most Popular Movies by Popularity Score

Minions is the most popular movie by the TMDB Popularity Score. **Wonder Woman** and **Beauty and the Beast**, two extremely successful woman centric movies come in second and third respectively.

As with popularity scores, the distribution of vote counts is extremely skewed with the median vote count standing at a paltry 10 votes. The most votes a single movie has got stands at 14,075. TMDB Votes, therefore, are not as potent and suggestive as its IMDB Counterpart. Nevertheless, let us check which the most voted on movies on the website are.

Most Voted on Movies

Inception and **The Dark Knight**, two critically acclaimed and commercially successful Christopher Nolan movies figure at the top of our chart.

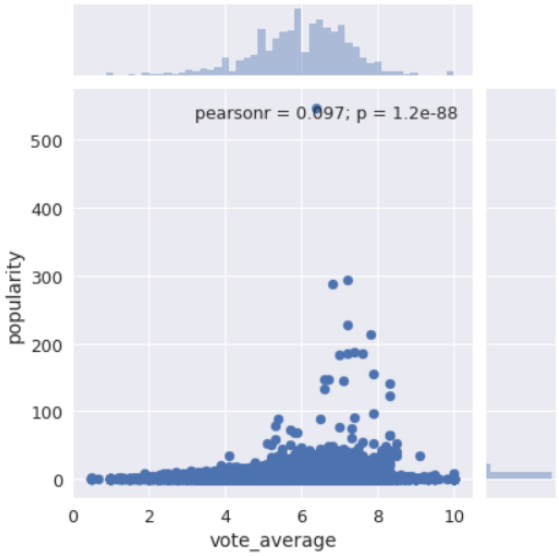


It appears that TMDB Users are extremely strict in their ratings. The mean rating is only a **5.6** on a scale of 10. Half the movies have a rating of less than or equal to 6. Let us check what the most critically acclaimed movies as per TMDB are. We will only consider those movies that have more than 2000 votes (similar to IMDB's criteria of 5000 votes in selecting its top 250).

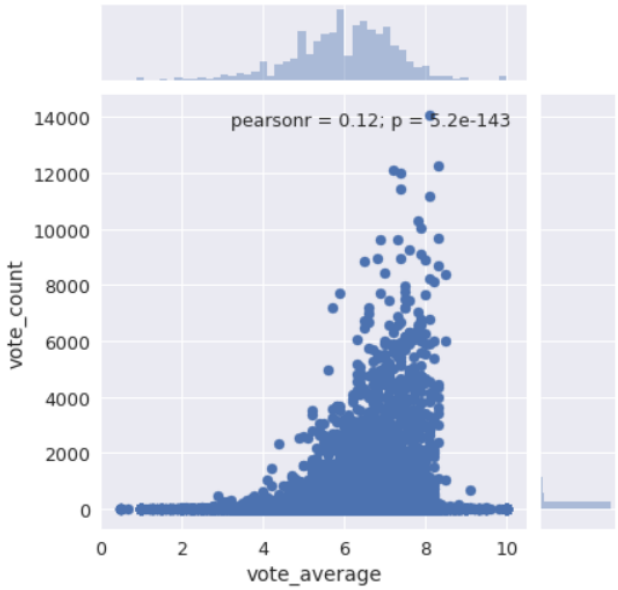
Most Critically Acclaimed Movies

The Shawshank Redemption and **The Godfather** are the two most critically acclaimed movies in the TMDB Database. Interestingly, they are the top 2 movies in IMDB's Top 250 Movies list too. They have a rating of over 9 on IMDB as compared to their 8.5 TMDB Scores.

Do popularity and vote average share a tangible relationship? In other words, is there a strong positive correlation between these two quanitties? Let us visualise their relationship in the form of a scatterplot.



Surprisingly, the Pearson Coefficient of the two aforementioned quantities is a measly **0.097** which suggests that **there is no tangible correlation**. In other words, popularity and vote average and independent quantities. It would be interesting to discover how TMDB assigns numerical popularity scores to its movies.



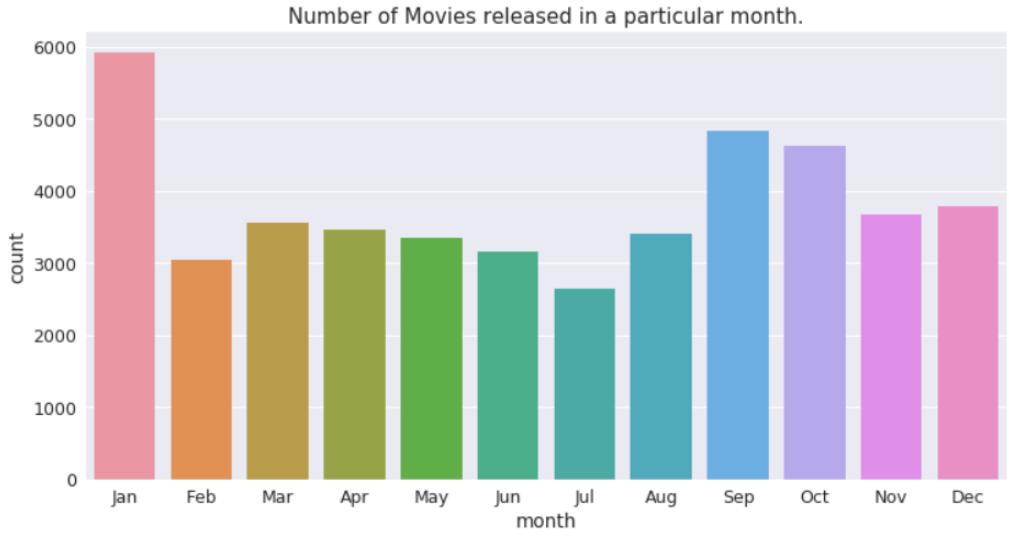
There is a very small correlation between Vote Count and Vote Average. A large number of votes on a particular movie does not necessarily imply that the movie is good.

Movie Release Dates

Release Dates can often play a very important role in determining the success and the revenue generated by a particular movie. In this section, we will try and gain insights about release dates in terms of years, months and days of the week.

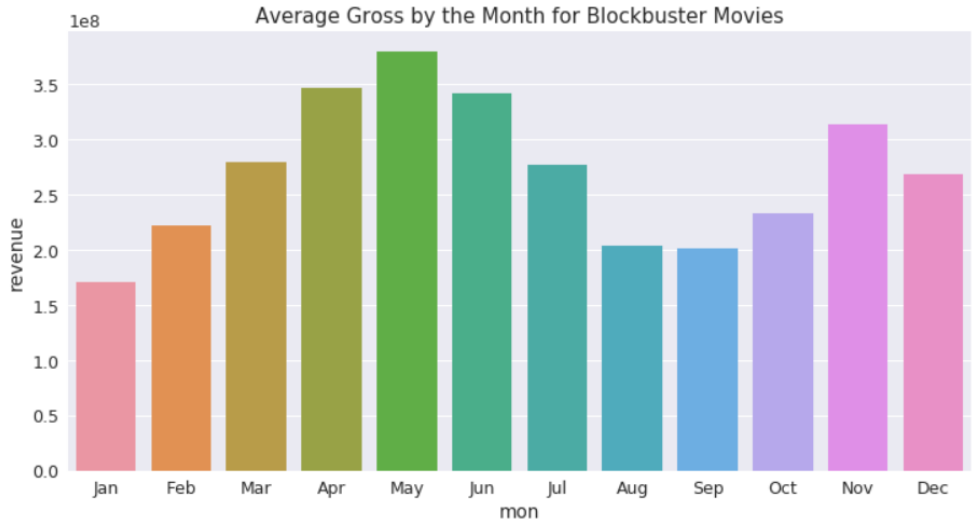
We have already constructed the **year** feature in our preliminary data wrangling step. Let us now extract the month and day too for each movie with a release date.

With these features in hand, let us now check the most popular and most successful months and days.

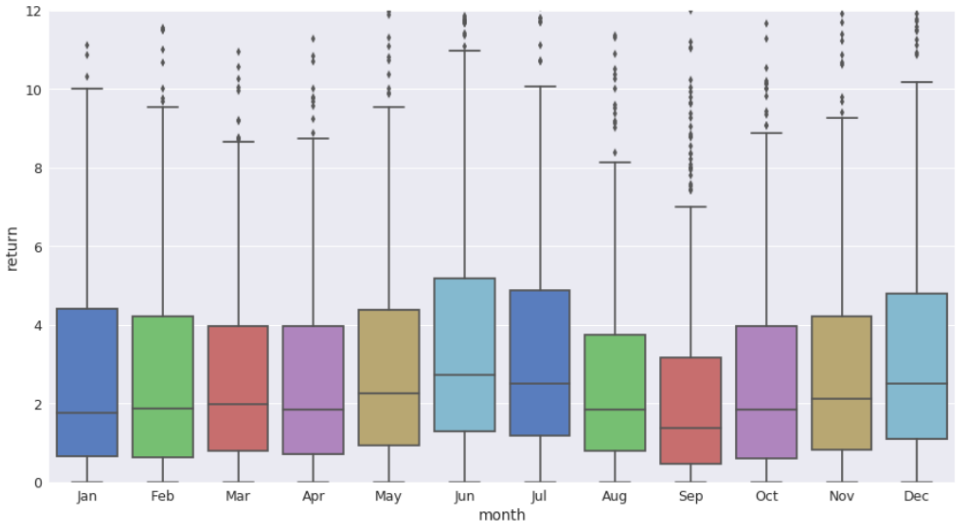


It appears that **January** is the most popular month when it comes to movie releases. In Hollywood circles, this is also known as the *the dump month* when sub par movies are released by the dozen.

In which months do bockbuster movies tend to release? To answer this question, we will consider all movies that have made in excess of 100 million dollars and calculate the average gross for each month.

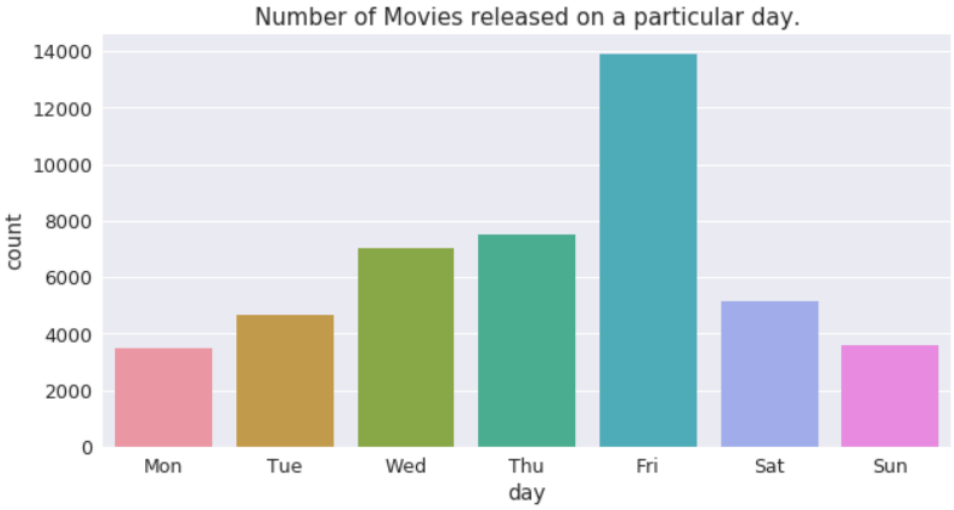


We see that the months of **April**, **May** and **June** have the highest average gross among high grossing movies. This can be attributed to the fact that blockbuster movies are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment, Do some months tend to be more successful than others? Let us visualise the boxplot between the return and the months.



The months of **June** and **July** tend to yield the highest median returns. **September** is the least successful months on the aforementioned metrics. Again, the success of June and July movies can be attributed to them being summer months and times of vacation. September usually denotes the beginning of the school/college semester and hence a slight reduction in the consumption of movies.

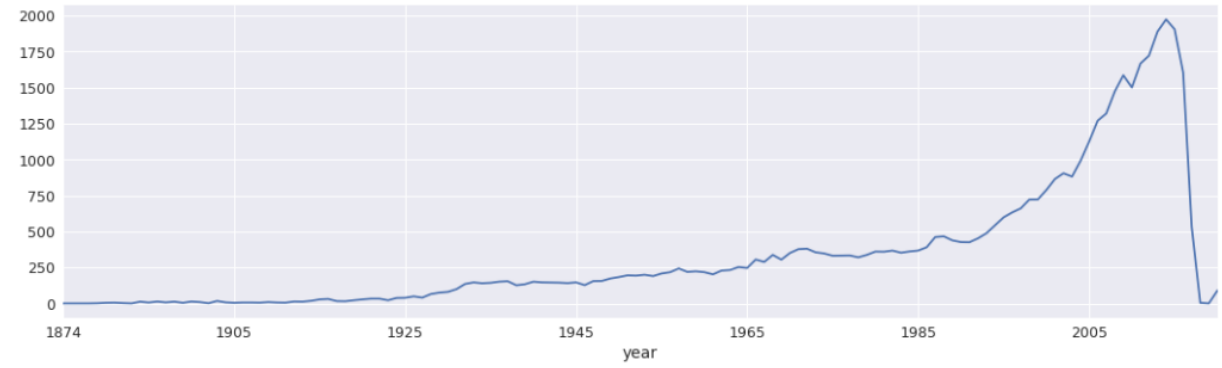
Let us now have a look at the most popular days as we did for months.



Friday is clearly the most popular day for movie releases. This is understandable considering the fact that it usually denotes the beginning of the weekend. **Sunday** and **Monday** are the least popular days and this can be attributed to the same aforementioned reason.

Number of Movies by the year

The Dataset of 45,000 movies available to us does not represent the entire corpus of movies released since the inception of cinema. However, it is reasomnable to assume that it does include almost every major film released in Hollywood as well as other major film industries across the world (such as Bollywood in India). With this assumption in mind, let us take a look at the number of movies produced by the year.



We notice that there is a sharp rise in the number of movies **starting the 1990s decade.** However, we will not look too much into this as it is entirely possible that recent movies were oversampled for the purposes of this dataset.

Next, let us take a look at the earliest movies represented in the dataset.

Earliest Movies Represented

The oldest movie, **Passage of Venus**, wss a series of photographs of the transit of the planet Venus across the Sun in 1874. They were taken in Japan by the French astronomer Pierre Janssen using his 'photographic revolver'. This is also the oldest movie on both IMDB and TMDB.

Finally, in this section, let us construct a heatmap to indicate movie releases by month and year for all movies released in this century. This will give us a good idea of the *hot* and *cold* months for movie buffs.

Movie Status

Although not entirely relevant to our analysis of movies, gathering information on the various kinds of movies based on their status of release can provide us interesitng insight on the nature of the movies present in our dataset. My preliminary hunch was that almost every movie has the **Released** status. Let's find out, Almost every movie is indeed released. However, it is interesting to see that MovieLens has user ratings for movies that are still in the planning, production and post production stage. We might take this information into account while building our collaborative filtering recommendation engine.

Spoken Languages

Does the number of spoken languages influence the success of a movie? To do this, we will convert our **spoken_languages** feature to a numeric feature denoting the number of languages spoken in that film.

Most movies have just one language spoken in the entire duration of the film. **19** is the higest number of languages spoken in a film. Let us take a look at all the films with more than 10 spoken languages.

The movie with the most number of languages, **Visions of Europe** is actually a collection of 25 short films by 25 different European directors. This explains the sheer diversity of the movie in terms of language.

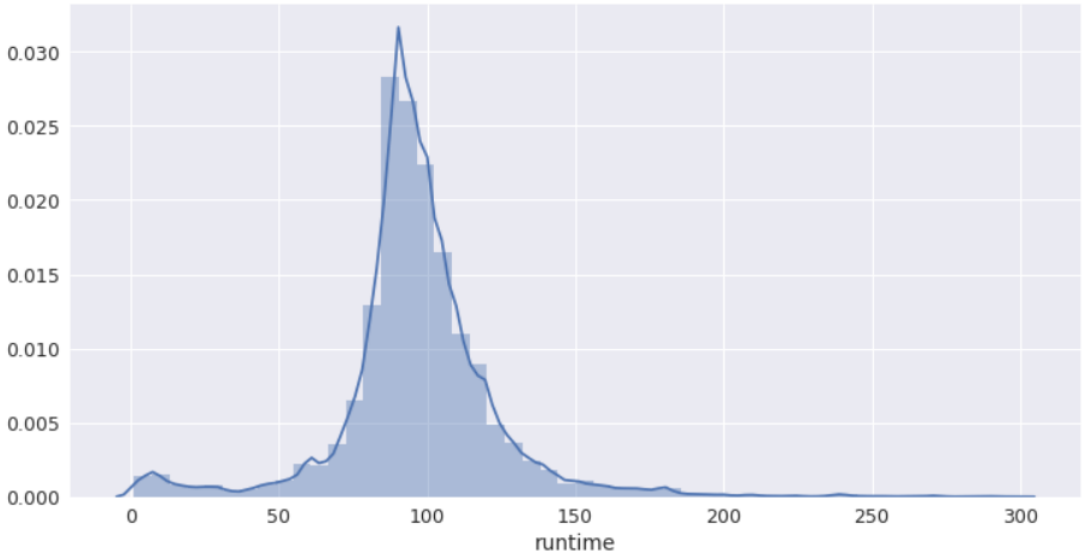
The **Spearman Coefficient** is 0.018 indicating no correlation between the two quantities.

Runtime

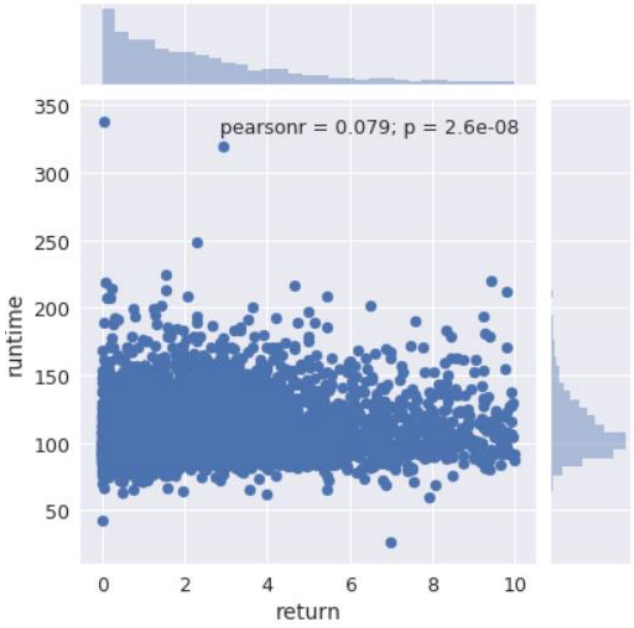
From its humble beginnings of 1 minute slient, black & white clips to epic 3 hour visual delights, movies have a come a long way in terms of runtime. In this section, let us try and gain some additional insights about the nature of movie lengths and their evolution over time.

The average length of a movie is about 1 hour and 30 minutes. The longest movie on record in this dataset is a **staggering 1256 minutes (or 20 hours) long.**

We are aware that most movies are less than 5 hours (or 300 minutes) long. Let us plot a distribution of these mainstream movies.



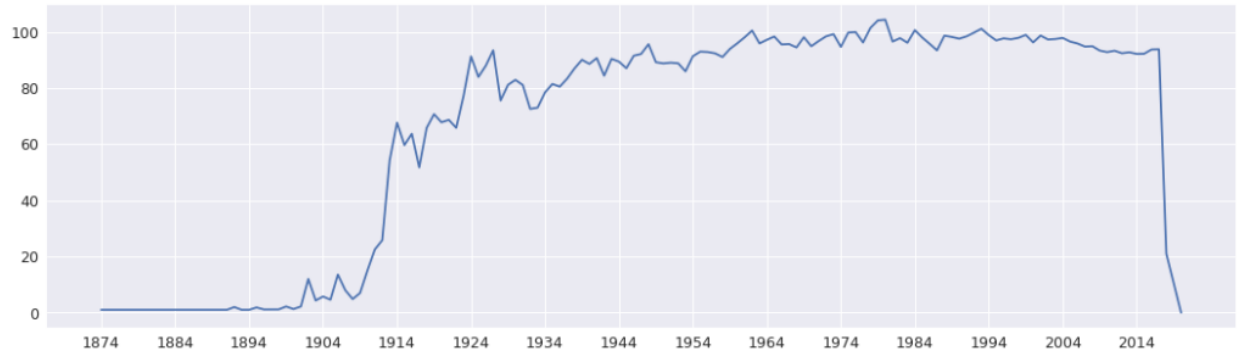
Is there any meaningful relationship between runtime and return? Let us find out!



There seems to be relationship between the two quantities. **The duration of a movie is independent of its success.** However, I have a feeling this might not be the case with duration and budget. A longer movie should entail a higher budget. Let us find out if this is really the case.

The two quantities have a much weaker correlation than I had expected. In retrospect, the genre of the movie tends to have a much greater impact on budget. A 3 hour art film will cost significantly lesser than a 90 minute Sci-Fi movie.

Next, I'd like to see the average lengths of movies through time, right from the 1890s to the 2017s. It would be interesting to see the trends in what filmmakers adjudged would be the appropriate length of a movie at that time.



We notice that films started hitting the **60 minute mark as early as 1914**. Starting **1924**, films started having the traiditonal 90 minute duration and has remained more or less constant ever since.

Finally in this section, let us see the longest and the shortest movies of all time (with respect to the movies in the dataset).

Shortest Movies

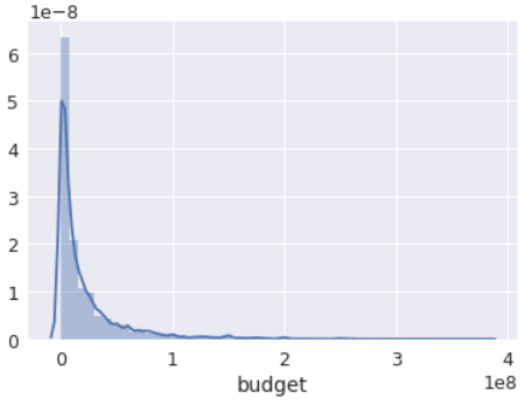
We see that every movie in this list except **A Gathering of Cats** were filmed in the late 1890s and the beginning of the 20th century. All these movies were one minute long.

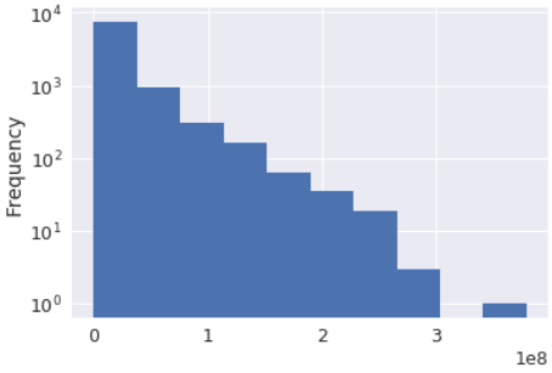
Longest Movies

We notice that almost all the entries in the above chart are actually miniseries and hence, do not count as feature length films. We cannot gather too much insight from this list of longest movies as there is no way of distinguishing feature length films from TV Mini Series from our dataset (except, of course, by doing it manually).

Budget

Let us now turn our attention to budget. We expect budgets to be a skewed quantity and also heavily influenced by inflation. Nevertheless, it would be interesting to gather as much insights as possible from this quantity as budget is often a critical feature in predicting movie revenue and success. As a start, let us gather the summary statistics for our budget, The mean budget of a film 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.



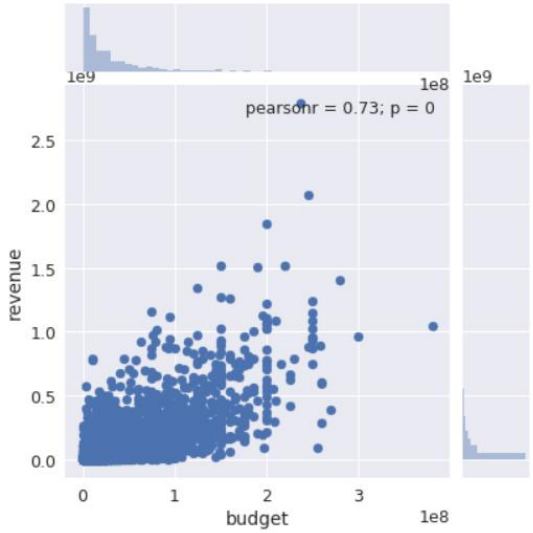


The distribution of movie budgets shows an exponential decay. More than 75% of the movies have a budget smaller than 25 million dollars. Next, let us take a look at the most expensive movies of all time and the revenue & returns that they generated.

Most Expensive Movies of all Time

Two **Pirates of the Carribean** films occupy the top spots in this list with a staggering budget of over **300 million dollars**. All the top 10 most expensive films 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.

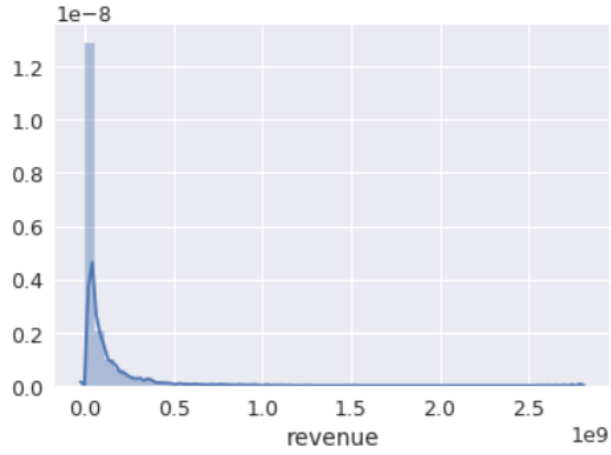


The pearson r value of **0.73** between the two quantities indicates a very strong correlation.

Revenue

The final numeric feature we will explore is the revenue. The revenue is probably the most important numeric quantity associated with a movie. We will try to predict the revenue for movies given a set of features in a later section. The treatment of revenue will be very similar to that of budget and we will once again begin by studying the summary statistics.

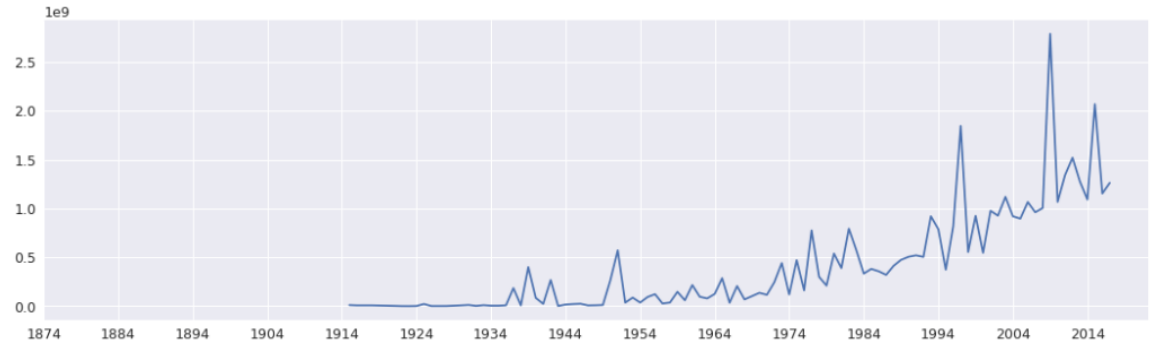
The mean gross of a movie is **68.7 million dollars** whereas the median gross is much lower at **16.8 million dollars**, suggesting the skewed nature of revenue. The lowest revenue generated by a movie is **just 1 dollar** whereas the highest grossing movie of all time has raked in an astonishing *2.78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



The distribution of revenue undergoes exponential decay just like budget. We also found that the two quantities were strongly correlated. Let us now take a look at the highest and least grossing movies of all time.

Highest Grossing Films of All Time

These figures have not been adjusted for inflation. Therefore, we see a disproportionate number of movies from very recent times in the top 10 list. To get an understanding of the revenue garnered by movies, let us plot the maximum revenue through the years.



As can be seen from the figure, the maximum gross has steadily risen over the years. 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**. Both these movies were directed by James Cameron.

Returns

We will not look too much into returns for the time being. Let us just 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.

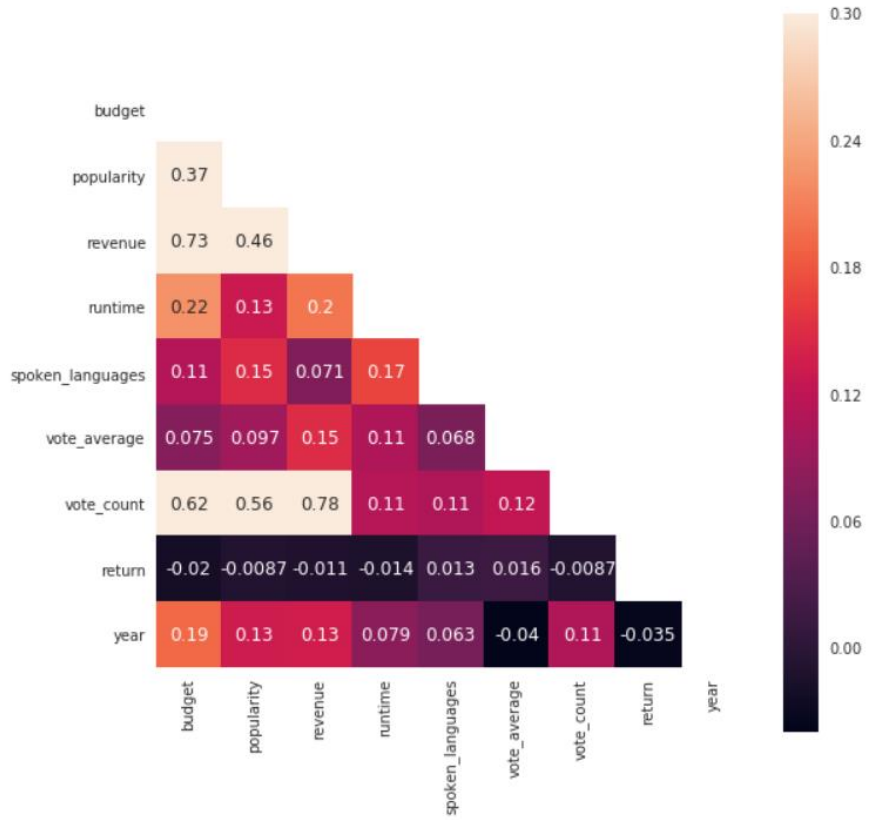
Most Successful Movies

title	budget	revenue	return	year
E.T. the Extra-Terrestrial	10500000.0	792965326.0	75.520507	1982
Star Wars	11000000.0	775398007.0	70.490728	1977
Jaws	7000000.0	470654000.0	67.236286	1975
The Exorcist	8000000.0	441306145.0	55.163268	1973
Four Weddings and a Funeral	6000000.0	254700832.0	42.450139	1994
The Godfather	6000000.0	245066411.0	40.844402	1972
Look Who's Talking	7500000.0	296000000.0	39.466667	1989
Annabelle	6500000.0	255273813.0	39.272894	2014
Dirty Dancing	6000000.0	213954274.0	35.659046	1987
The Sound of Music	8200000.0	286214286.0	34.904181	1965

Worst Box Office Disasters

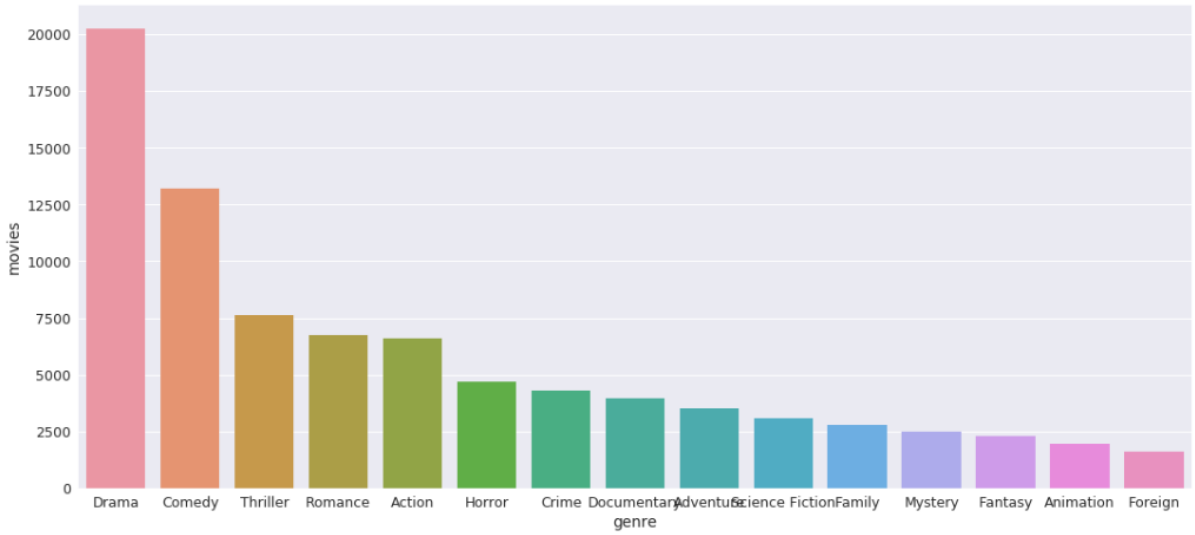
title	budget	revenue	return	year
Chaos	20000000.0	10289.0	0.000514	2005
5 Days of War	20000000.0	17479.0	0.000874	2011
Special Forces	10000000.0	10759.0	0.001076	2011
Foodfight!	65000000.0	73706.0	0.001134	2012
Term Life	16500000.0	21256.0	0.001288	2016
Laurence Anyways	9500000.0	12250.0	0.001289	2012
The Good Night	15000000.0	20380.0	0.001359	2007
Cherry 2000	10000000.0	14000.0	0.001400	1987
Twice Born	13000000.0	18295.0	0.001407	2012
All The Queen's Men	15000000.0	23000.0	0.001533	2001

With these analyses in place, we are in a good position to construct our correlation matrix.



Genres

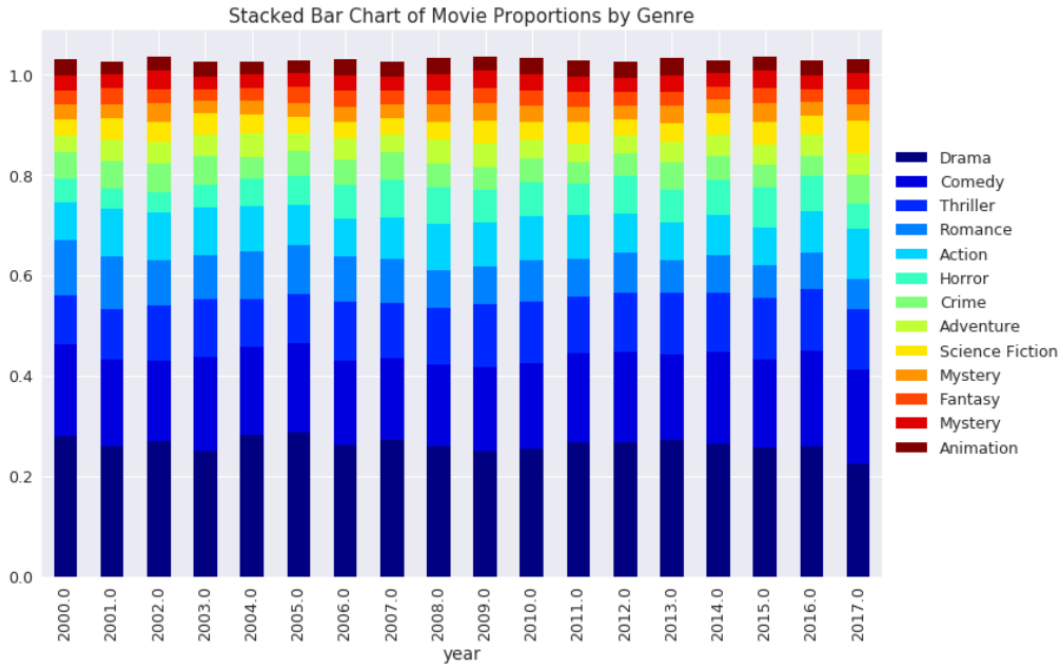
TMDB defines 32 different genres for our set of 45,000 movies. Let us now have a look at the most commonly occuring genres in movies.

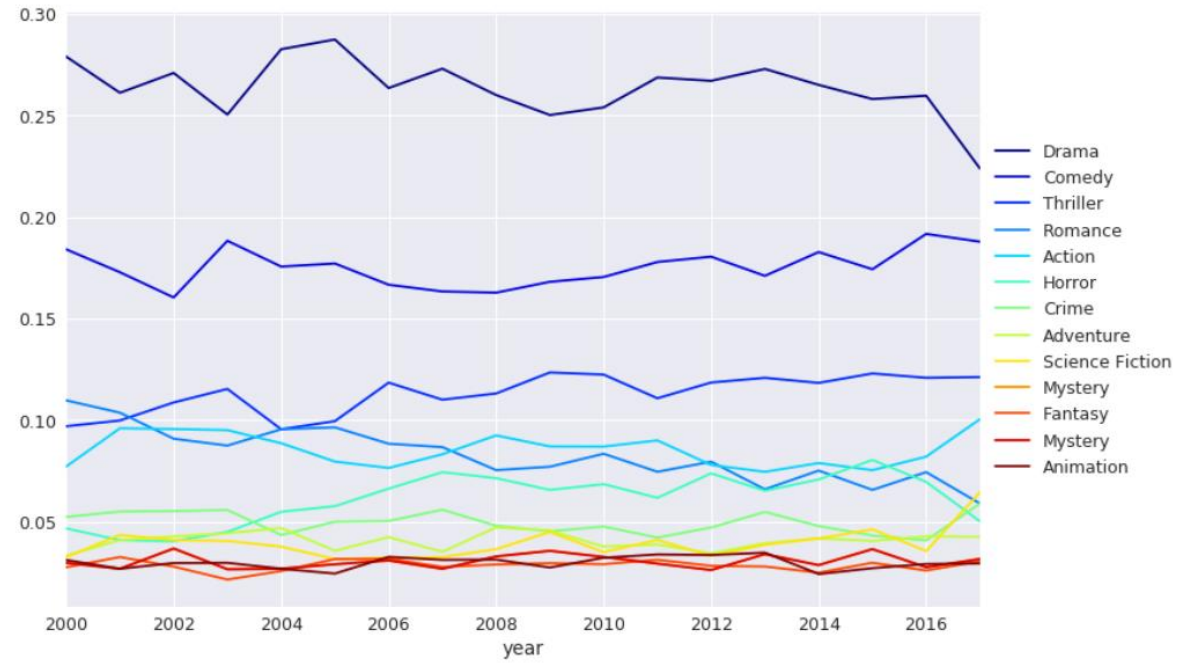


Drama is the most commonly occurring genre with almost half the movies identifying itself as a drama film. **Comedy** comes in at a distant second with 25% of the movies having adequate doses of humor. Other major genres represented in the top 10 are Action, Horror, Crime, Mystery, Science Fiction, Animation and Fantasy.

The next question I want to answer is the trends in the share of genres of movies across the world. Has the demand for Science Fiction movies increased? Do certain years have a disproportionate share of Animation Movies? Let's find out!

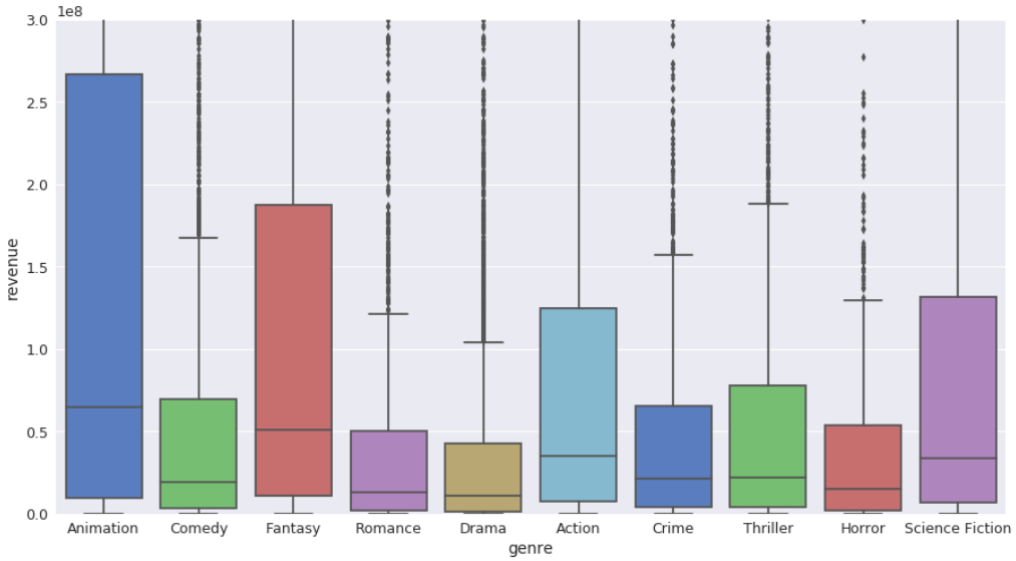
We will only be looking at trends starting 2000. We will consider only those themes that appear in the top 15 most popular genres. We will exclude Documentaries, Family and Foreign Movies from our analysis.



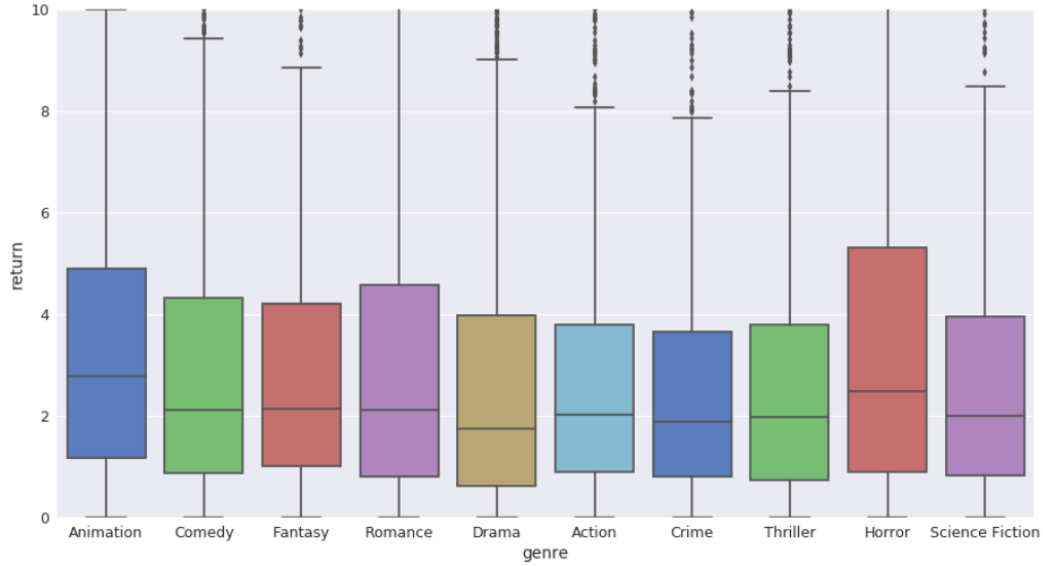


The proportion of movies of each genre has remained fairly constant since the beginning of this century except for **Drama**. The proportion of drama films has fallen by over 5%. **Thriller** movies have enjoyed a slight increase in their share.

One question that I have always had is that if some genres are particularly more successful than others. For example, we should expect Science Fiction and Fantasy Movies to bring in more revenue than other genres but when normalized with their budget, do they prove to be as successful? We will visualize two violin plots to answer this question. One will be genres versus the revenue while the other will be versus returns.



Animation movies has the largest 25-75 range as well as the median revenue among all the genres plotted. **Fantasy** and **Science Fiction** have the second and third highest median revenue respectively.



From the boxplot, it seems like **Animation** Movies tend to yield the highest returns on average. **Horror** Movies also tend to be a good bet. This is partially due to the nature of Horror movies being low budget compared to Fantasy Movies but being capable of generating very high revenues relative to its budget.

Cast and Crew

Let us now take a look at the cast and crew of our movies. We do not have these details with us in our main dataset. However, we have a separate file consisting of the full cast and crew credits of all the Movielens Movies. Let us take a look at this credits data.

	cast	crew	id
0	[{'cast_id': 14, 'character': 'Woody (voice)',...	[{'credit_id': '52fe4284c3a36847f8024f49', 'de...	862
1	[{'cast_id': 1, 'character': 'Alan Parrish', '...	[{'credit_id': '52fe44bfc3a36847f80a7cd1', 'de...	8844
2	[{'cast_id': 2, 'character': 'Max Goldman', 'c...	[{'credit_id': '52fe466a9251416c75077a89', 'de...	15602
3	[{'cast_id': 1, 'character': "Savannah 'Vannah...	[{'credit_id': '52fe44779251416c91011acb', 'de...	31357
4	[{'cast_id': 1, 'character': 'George Banks', '...	[{'credit_id': '52fe44959251416c75039ed7', 'de...	11862

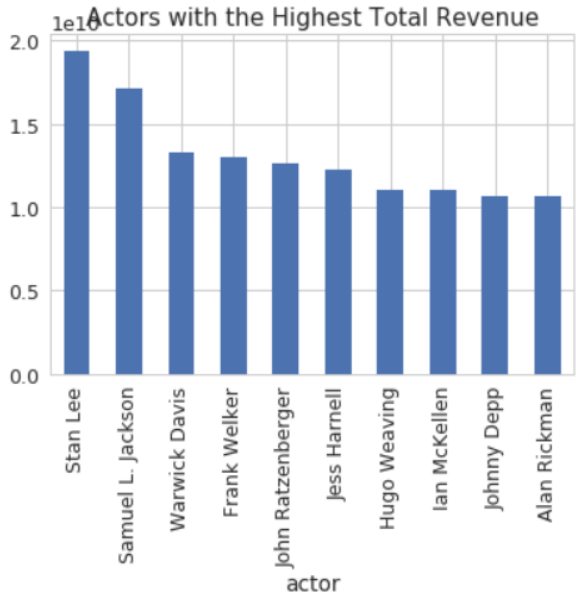
Credits Dataset :

- **cast:** A stringified list of dictionaries consisting of cast names and the corresponding characters they played.
- **crew:** A stringified list of dictionaries consisting of crew names and the function they performed.
- **id:** The TMDB ID of movie.

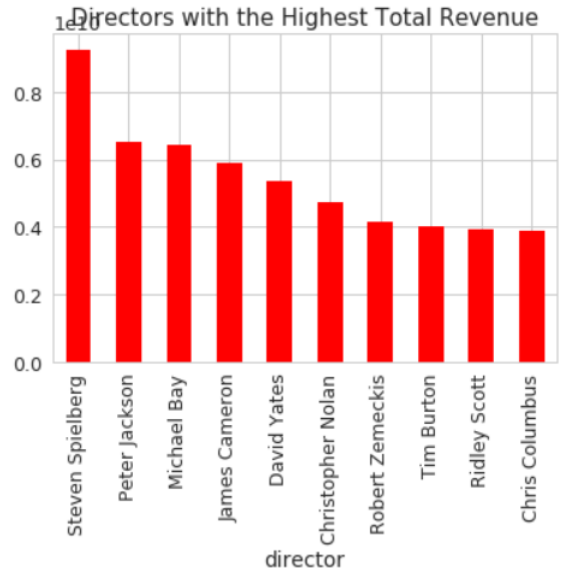
We need to perform a left join of our original movies metadata dataframe with the credits dataframe on the TMDB Movie ID. Before we are able to perform this join, we need to make sure that the ID column of our main dataframe is clean and of type integer. To do this, let us try to perform an integer conversion of our IDs and if an exception is raised,we will replace the ID with NaN. We will then proceed to drop these rows from our dataframe.

Let us now take a look at the actors and the directors who have raked in the most amount of money with their movies.

Actors with the Highest Total Revenue

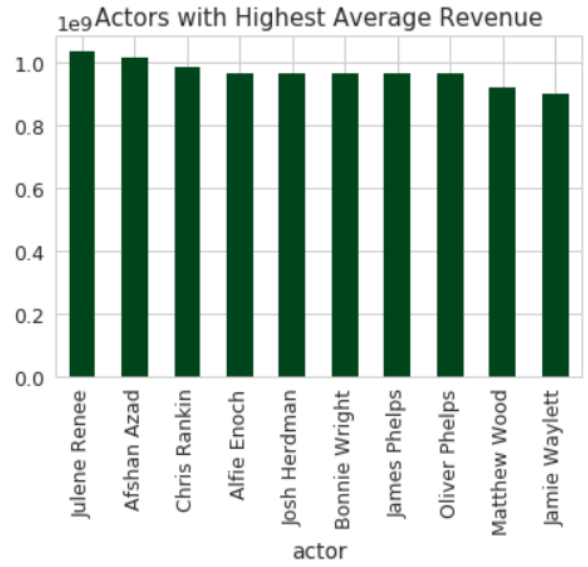


Directors with the Highest Total Revenue

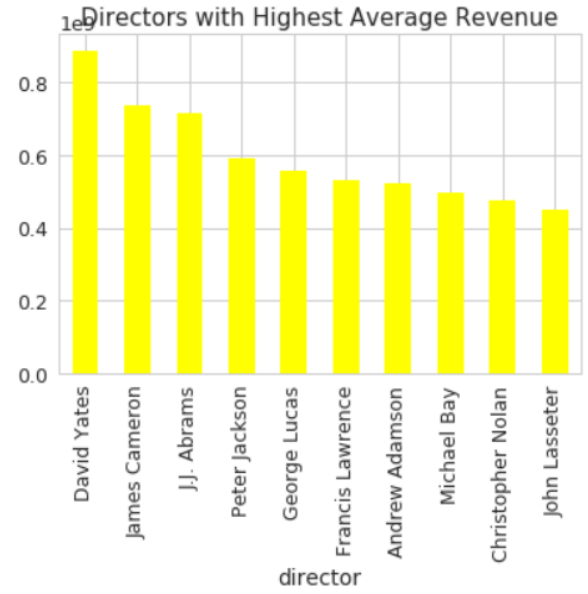


For average revenues, we will consider only actors and directors who have acted and directed in at least 5 movies respectively.

Actors with Highest Average Revenue



Directors with Highest Average Revenue



Which actors and directors are the safest bet? For this, we will consider the average return brought in by a particular director or actor. We will only consider those movies that have raked in at least 10 million dollars. Also, we will only consider actors and directors that have worked in at least 5 films.

Most Successful Actors

actor	return
Jami Gertz	3.099099e+06
Donna Mitchell	2.479289e+06
Andrew McCarthy	2.479280e+06
Nicholas Pryor	1.770914e+06
James Spader	1.549551e+06
Michael Bowen	1.033038e+06
Robert Downey Jr.	5.165189e+05
Martin Kove	3.395458e+05
William Zabka	3.395441e+05
Brad Pitt	2.817390e+05

Most Successful Directors

director	return
John G. Avildsen	169822.750468
George A. Romero	60.656935
George Lucas	54.328446
Martha Coolidge	49.553131
Davis Guggenheim	46.243000
George Miller	39.076975
James Wan	27.659835
Guy Hamilton	25.779057
John Badham	25.433532
John Carpenter	25.075058

John G. Avildsen has an unnaturally high return. All the other directors in the list are nowhere close to the level of return that he has gained. Let us have a look at his movies.

	title	budget	revenue	return	year
1853	Rocky	1000000.0	117235147.0	1.172351e+02	1976.0
2307	Rocky V	42000000.0	119946358.0	2.855866e+00	1990.0
2315	The Karate Kid	8000000.0	90815558.0	1.135194e+01	1984.0
2316	The Karate Kid, Part II	113.0	115103979.0	1.018619e+06	1986.0
5226	Joe	106000.0	19319254.0	1.822571e+02	1970.0
5658	Neighbors	8500000.0	29916207.0	3.519554e+00	1981.0

The Karate Kid, Part II has a budget of only 113 dollars. This seems like an anomaly since official figures state that the movie cost **13 million dollars**. So, although he has directed amazing movies, he does not belong to this list.

We will end our Exploratory Data Analysis over here. Let us use some of the insights we gained in this section and build some useful predictive models.