

Cal State LA University

IMDb Movie Dataset Analysis

By: Bita Sadat Faraji Tabrizi

1. **Data set URL’s**

<https://www.kaggle.com/tmdb/tmdb-movie-metadata>

<https://www.kaggle.com/PromptCloudHQ/imdb-data>

This first dataset contains information about the IMDB movies. This dataset has movies from 1916 to 2017. The second dataset contains IMDB data from 2006 to 2016. Due to the fact that the “rating” variable is missing from the first dataset, the second dataset was used to provide rating for the common movies. Overall, in this project IMDB movies from years 2006 to 2016 is analyzed.

The first dataset was cleaned as they were a lot of duplicate values. The overall combined dataset contains 1000 rows and 20 columns. The 20 columns are as follows:

|  |  |  |
| --- | --- | --- |
| Budget | Overview | Spoken Languages |
| Genres | Popularity | Status |
| Homepage | Production Companies | Tagline |
| ID | Production Countries | Runtime |
| Keywords | Release date | Vote Average |
| Original language | Revenue | Vote Count |
| Original Title | Language | Rating |

1. **Data Cleaning**
2. **Illegal values:**

**Before:**



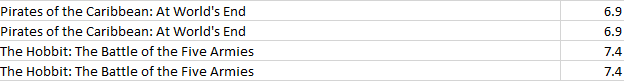
**After:**



The unwanted “symbols” and “id” are deleted and only the last genre type is saved to avoid large number of duplications for each movie. This is done using “=TRIM () function. In general, each movie can be categorized by several genre but to avoid the repetition of genres, only the last genre was considered.

1. **Duplicates:**

**Before:**



**After:**



Highlighted all the duplicate cells using duplicate function. Then, Filtered all the highlighted values and eliminated the entire row which contained duplicates.

1. **Misspelling**

**Before:**



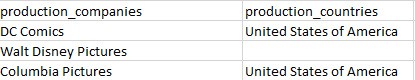
**After:**



The keywords contained junk values as shown. The junk values are removed to get the keywords “hostage” and “texas”.

1. **Missing value:**

**Before:**



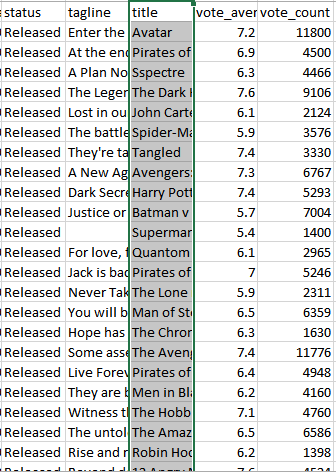
**After:**



The production country was missing on Walt Disney Pictures Company, so it was filled with the appropriate country. This was filled based on a search done on the Internet.

1. **Remove Unwanted observations**

**Before:**



**After:**



In this dataset there are irrelevant observations that don’t actually fit the specific problems that is being solved.

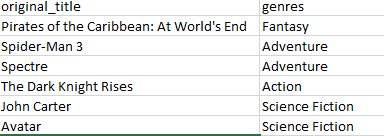
1. **Sorting:**

In this dataset the genres were sorted by adding custom filter to all the attributes rows. This sorting makes it easier to read the excel file and observe which movies are under the same genres.

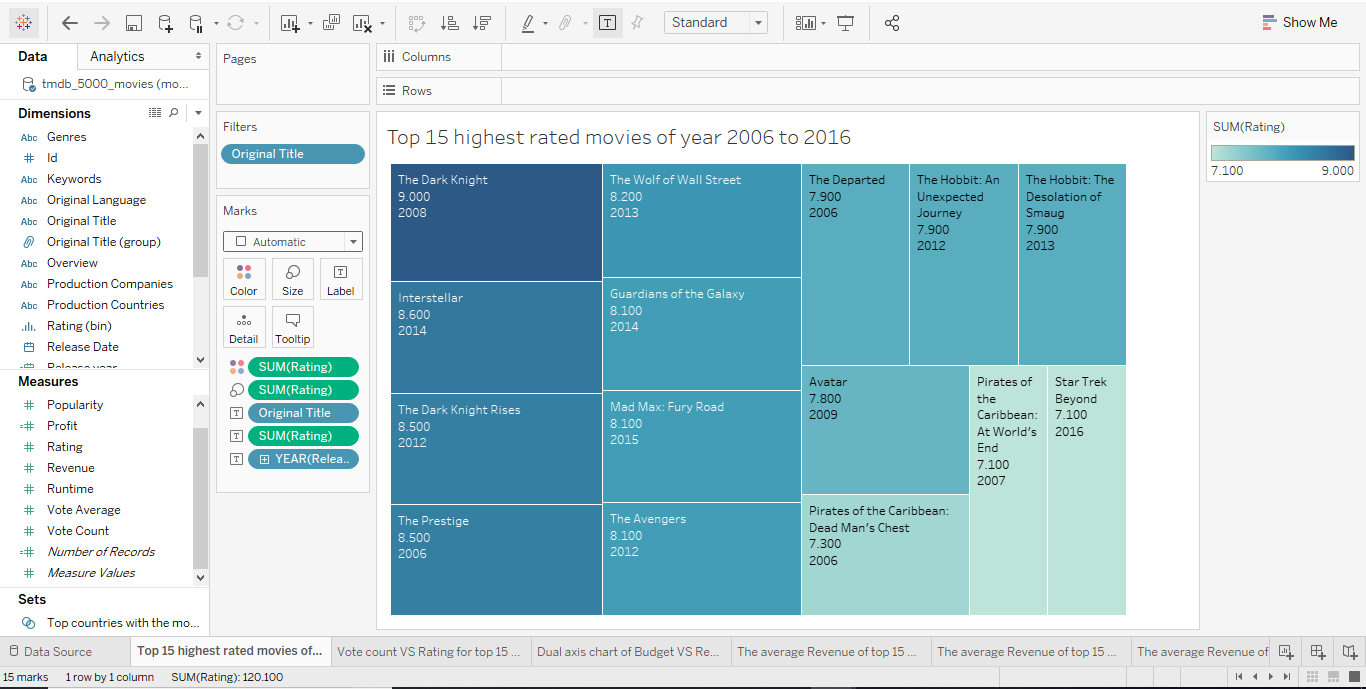
**Before:**



**After:**

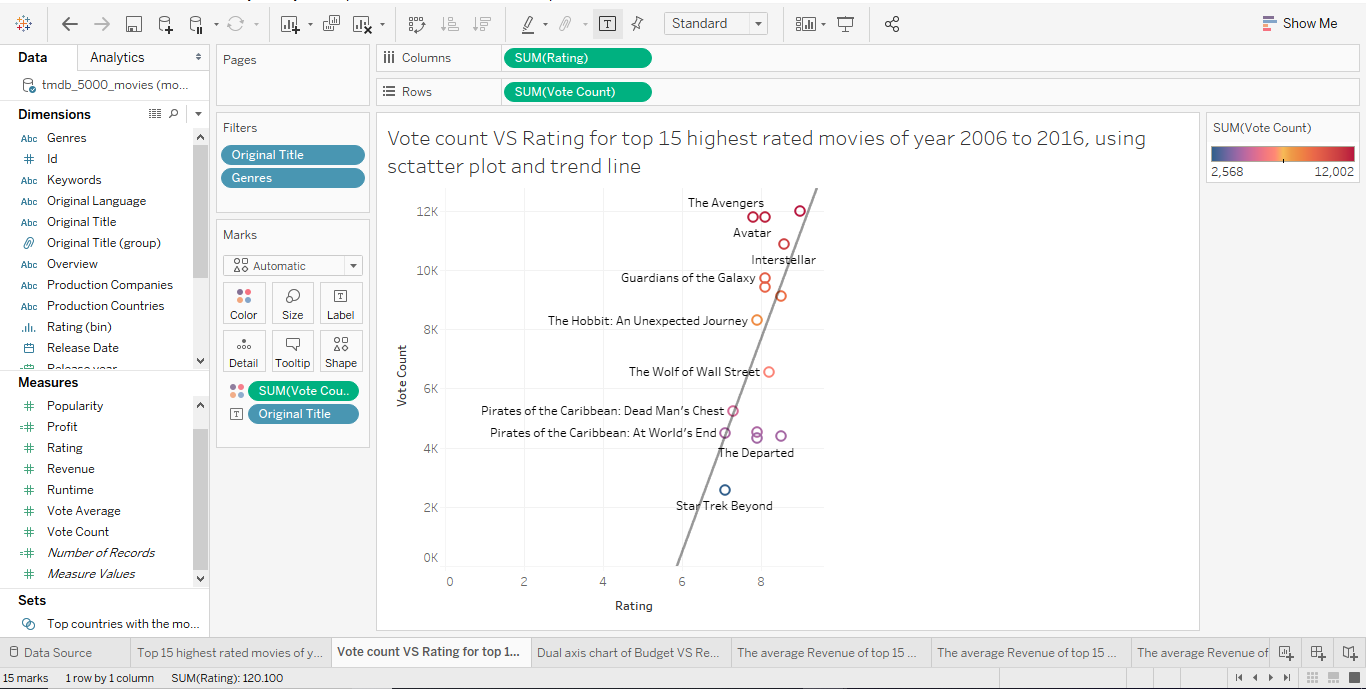


1. **Data Visualizations**
2. **Top 15 highest rated movies of year 2006 to 2016**



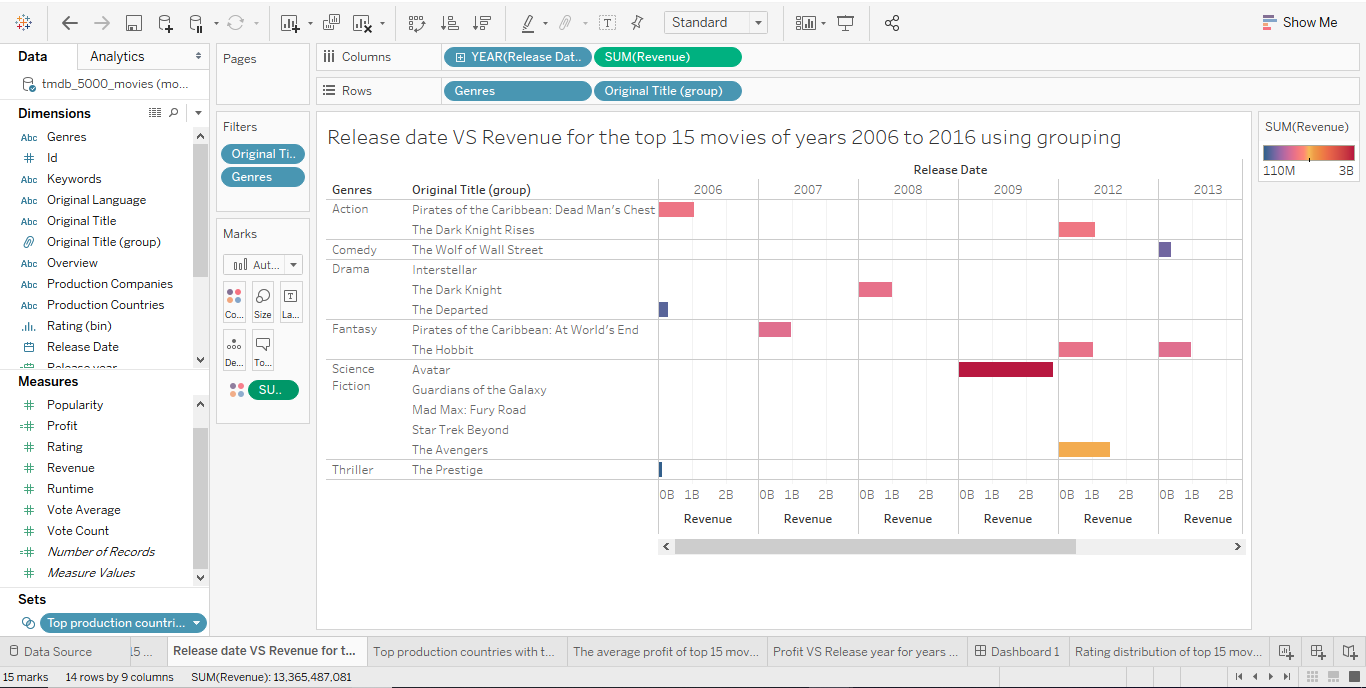
A Tree Mapwas used to show the top 15 movies of years 2006 to 2016. As it can be seen above The Dark Knight has the highest rate, which is 9.0. The IMDB ratings ranges from 1 to 10, with 1 being the lowest and 10 being the highest. This visual is appropriate for this representation as there is a scale of color, with the darkest color showing the highest rated movie.

1. **Vote count VS Rating for top 15 highest rated movies of year 2006 to 2016**



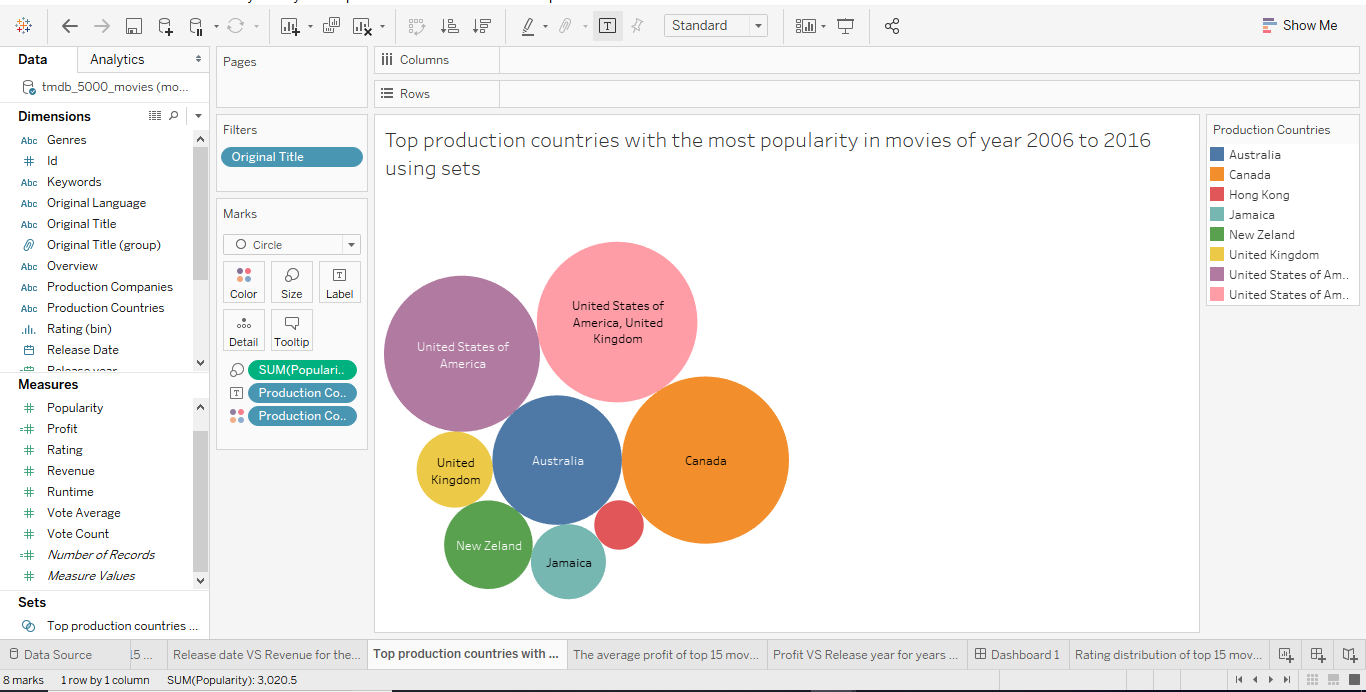
**Scatter Plot** was used to show the correlation between vote count and rating. As shown above as the rating increases the vote count increases as well. The points in the graph are connected using Trend line**.** This line shows the linear relationship between vote count and rating. The Dark Knight again is at the top, as both its vote count and rating are the highest.

1. **Release date VS Revenue for the top 15 movies of years 2006 to 2016**

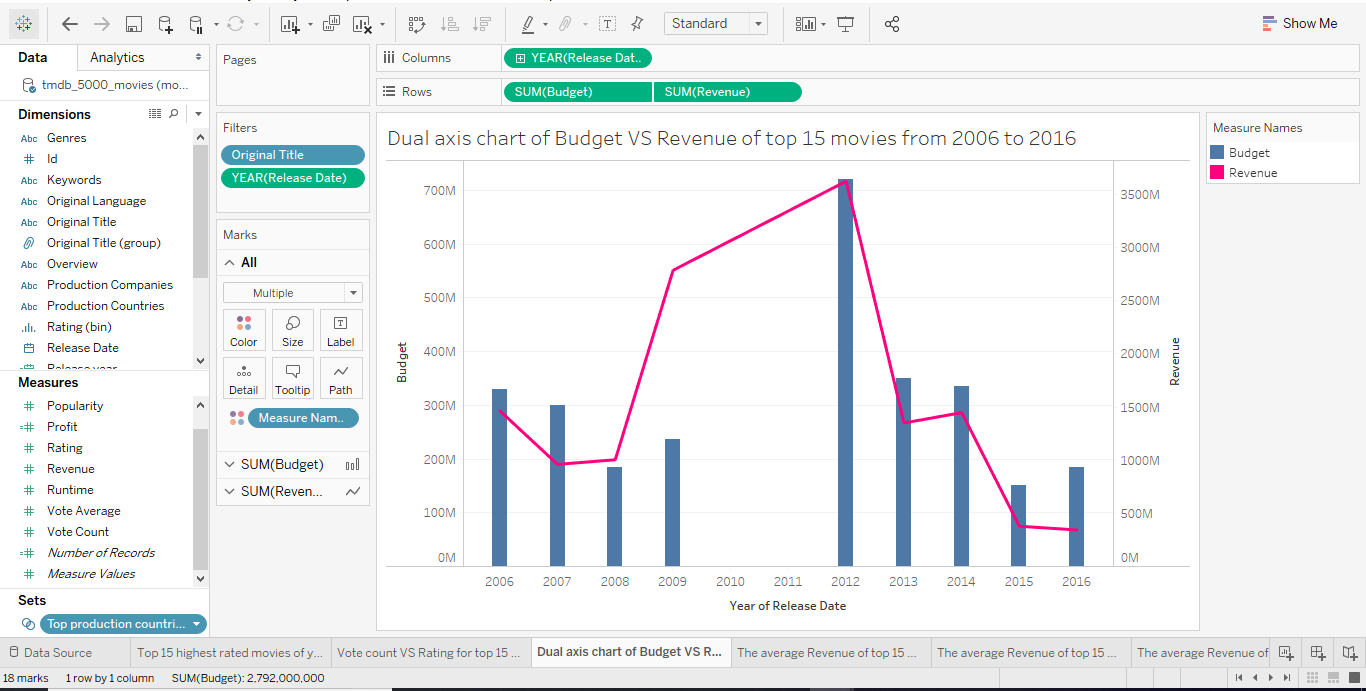


A Bar Chart is used in this visualization to show the Release date against the Revenue generated. In addition, in this visualization **Grouping** method is used to group the two hobbit movies as one. This bar chart shows that the highest revenue is in year 2009 for the movie Avatar, which is 2,787,965,000. And the lowest revenue in year 2006 belongs to the movie The Departed, which is 289,847,354. The bar chart was a good way to represent this visualization as the highest and the lowest revenue can be clearly seen from the visual, with highest revenue being the darkest color.

1. **Top production countries with the most popularity in movies of years 2006 to 2016**

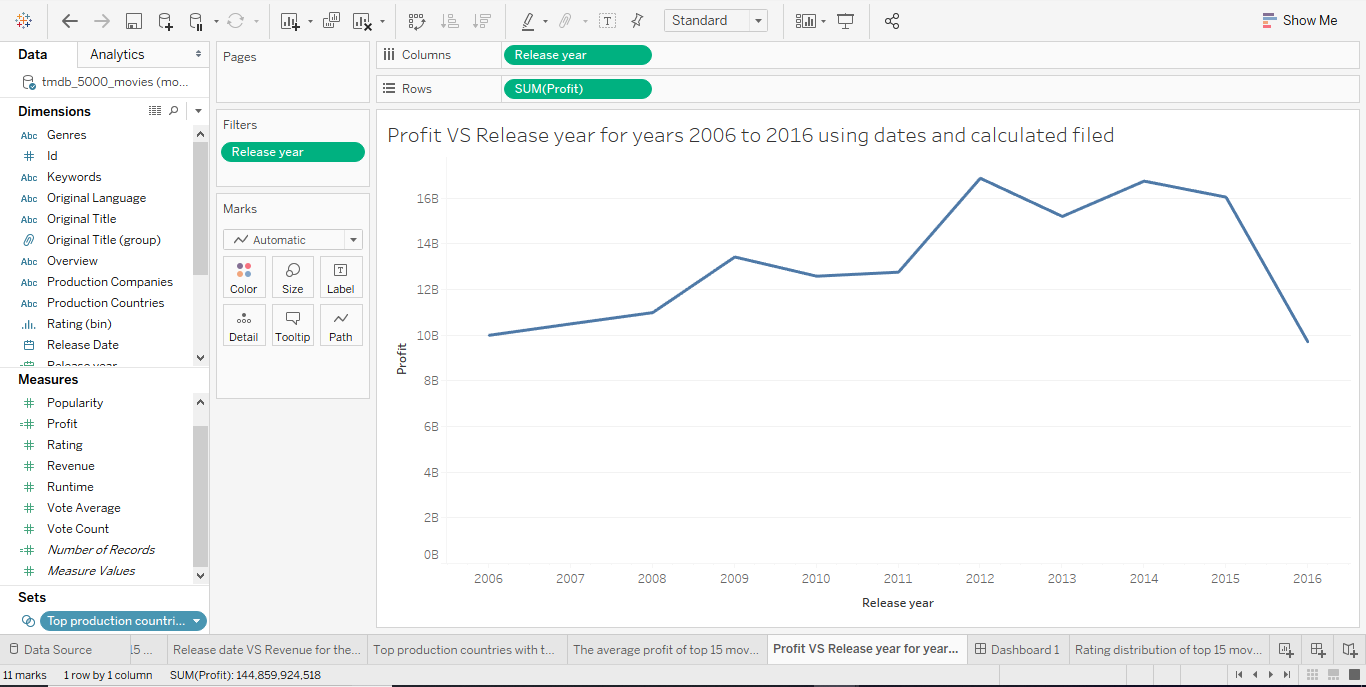
To show the top production countries with the most popularity, packed bubble is used. This visualization provides a clear view, using different sizes of bubbles. This was done by using **Sets.** This shows Canada has the highest popularity with 742.2. In the sets the top 15 are chosen based on popularity.

1. **Dual axis chart of Budget VS Revenue of top 15 movies from 2006 to 2016**



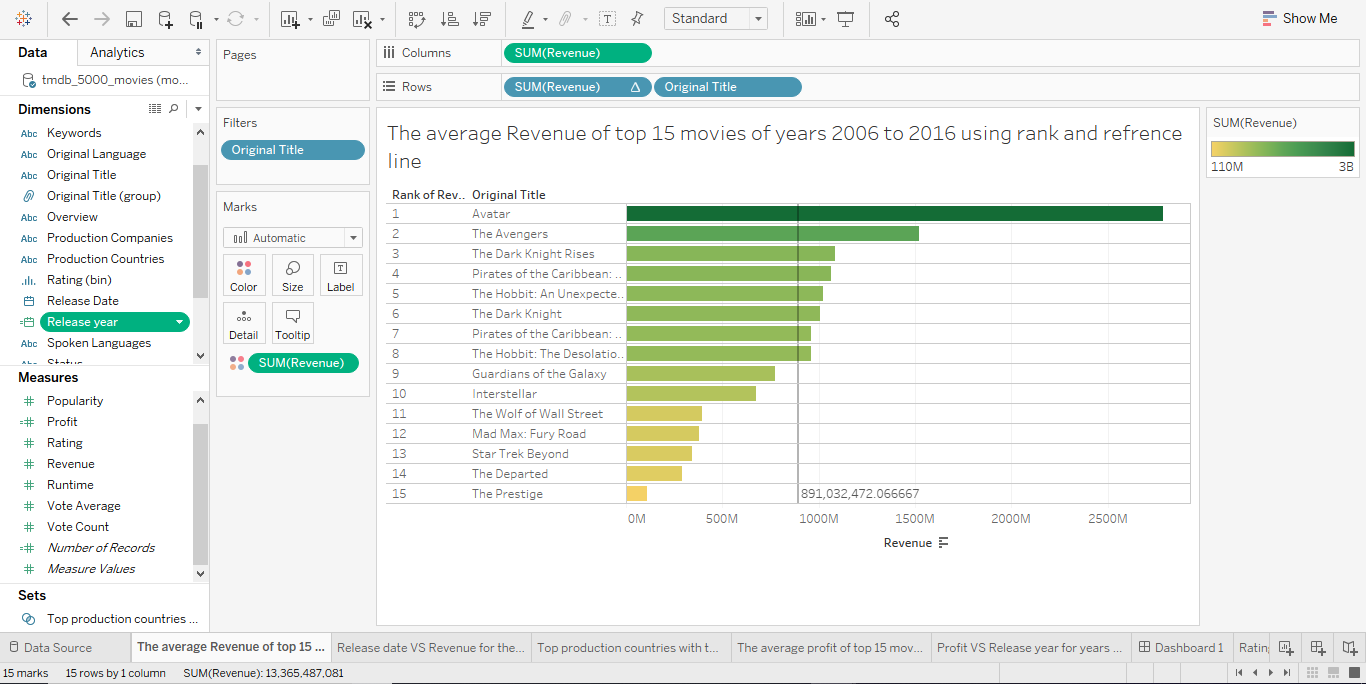
In this visualization **Dual Axis Chart** is used to show the correlation between budget and revenue for years 2006 to 2016. The bars represent the budget, while the revenue is represented using a line. As shown above, the highest budget along with the highest revenue is present in year 2012. This visualization provides a color for each axes, to make it easier to understand.

1. **Profit VS Release year for years 2006 to 2016**



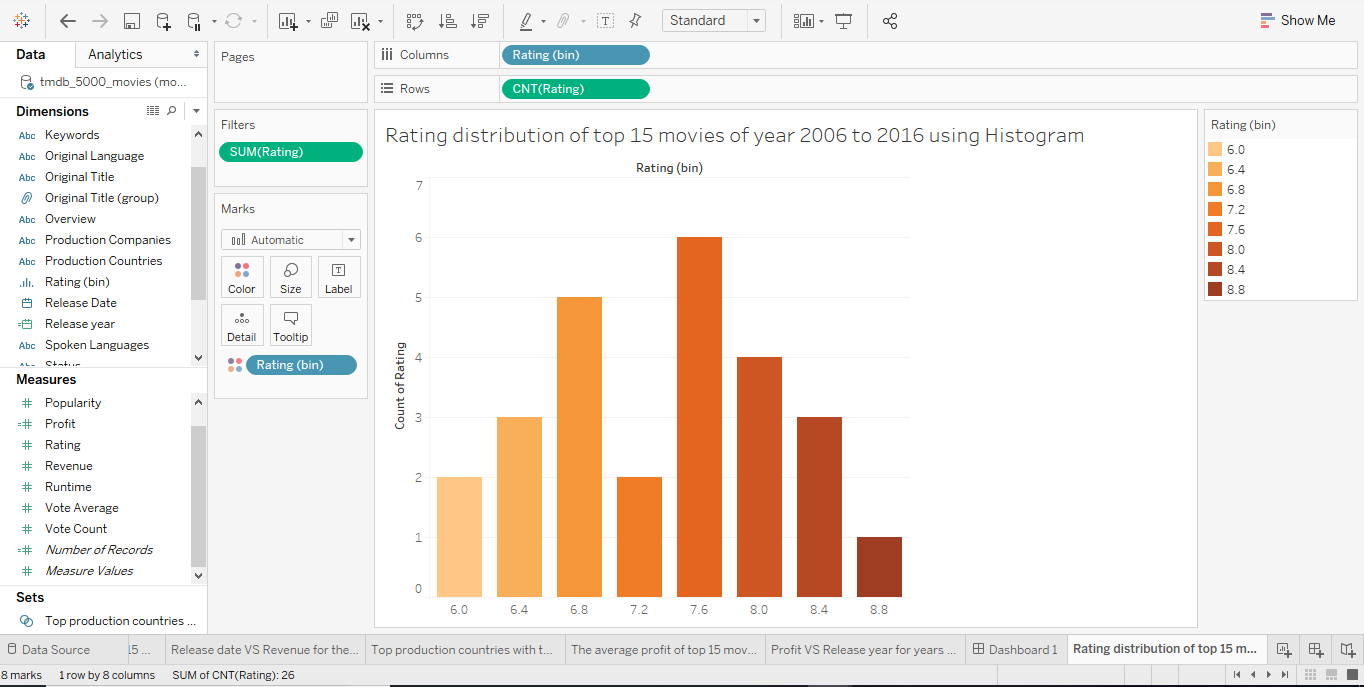
This visualization shows the relationship between Profit and Release year. Release year is obtained using **dates**,from the release date variable. In this visualization the **calculated field** Profit is used. This is calculated using the formula Profit= Revenue – Budget. Moreover, the visualization shows that the highest profit was at year 2012 which was 16,877,927,592. Overall the line shows an increase in profit over the years till year 2015, but a sudden drop in 2016.

1. **The average Revenue of top 15 movies of years 2006 to 2016**



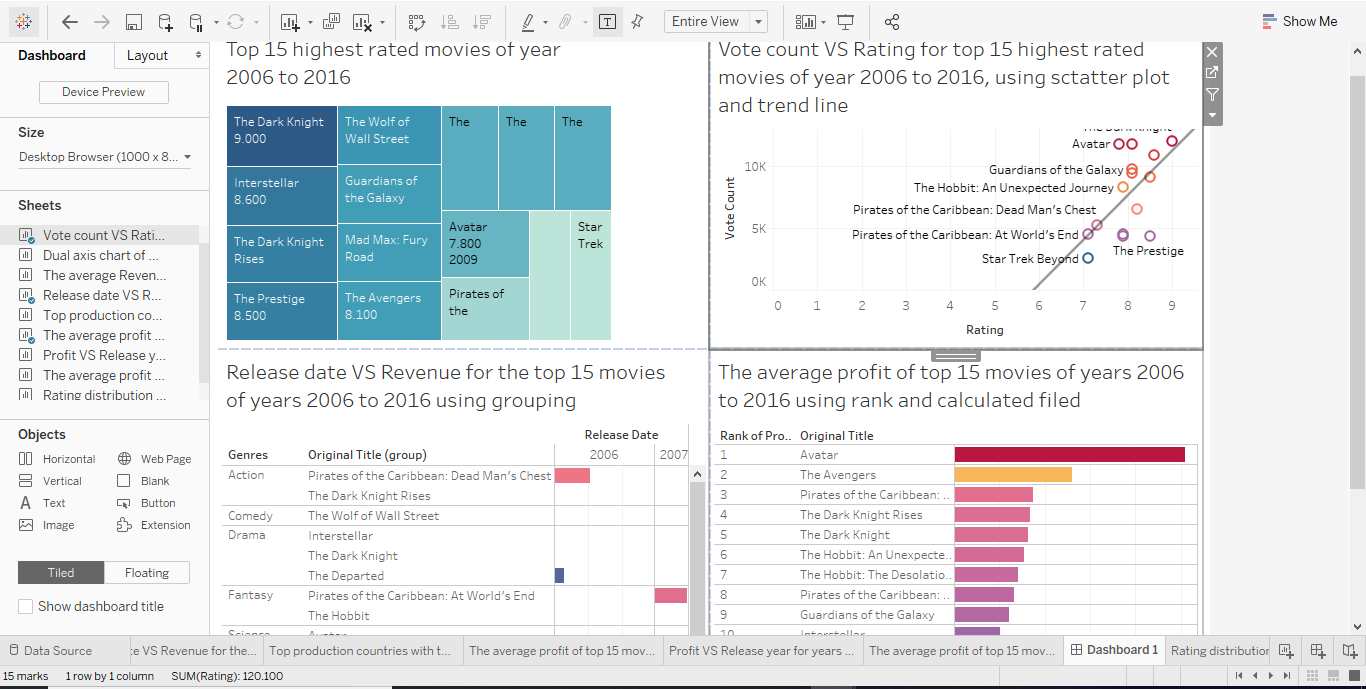
In this visualization a Horizontal Bar chart is used to show the average revenue of top 15 movies of year 2006 to 2016. The average revenue is 891,032,472.066667 which was calculated using the **Reference line**. Also, each movie is **Ranked** based on the average revenue. As it can be seen above, Avatar has the highest average revenue.

1. **Rating distribution of top 15 movies of year 2006 to 2016**

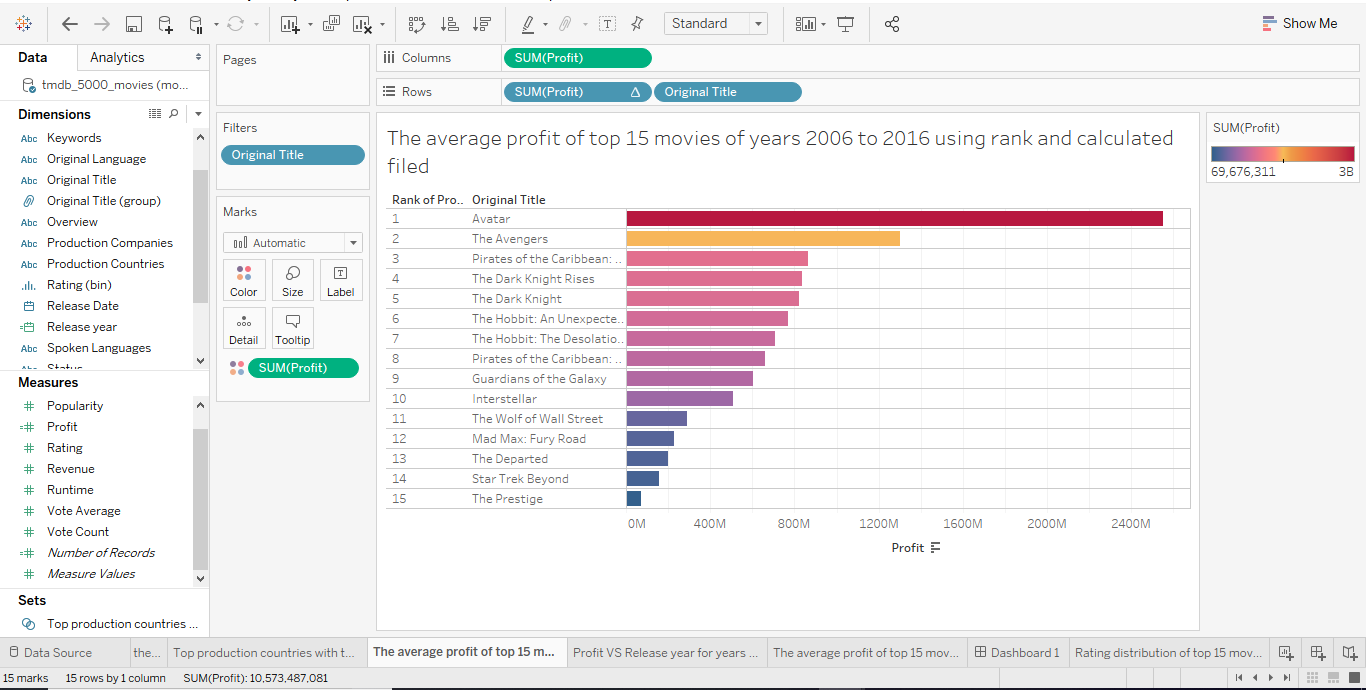
In this visualization **Histogram** was used to show the rating distribution of top 15 movies. A histogram looks like a bar chart but groups values for a continuous measure into ranges, or bins.

The bin size that was used is 0.4. As it can be seen 7.4 has the highest rating, with the count of rating being 6.

1. **Dashboard**



In this Dashboard 4 different visualizations are presented. **Tree Map** shows top 15 highest rated movies, **Scatter Plot** that shows Vote count VS Rating for top 15 highest rated movies, **Bar Chart** that compares Release date and Revenue of top 15 highest ranked movies, and **Horizontal Bar** that shows the average profit of top 15 highest rated movies.

 The additional visualization that was used in Dashboard is the average profit of top 15 movies of years 2006 to 2016. In this visualization the **calculated field** Profit is used. As mentioned in question six, the formula is Profit= Revenue – Budget. Also the titles are **ranked** based on the profit, using the quick table calculation in profit. The visualizations shows that Avatar has the highest average profit, making it the first ranked with the value of 2,550,965,087.

1. **Story telling:**

**How it all started?** A recorded sequence of film / video images displayed on a screen with sufficient rapidity as to create the illusion of motion and continuity, called as Movie. Film as information which is experience good. The feature of experience good is that consumers must experience it before they know what it is? It can be experienced by analyzing rating, Therefore the ratings of the movies by critics as well as people determine how popular the movie can be. A high IMDb rating makes an insignificant contribution towards a movie's success. A film's genre, voting, releasing date, budget and revenue at the box office all matter for the popularity and success of the movies in a particular period. IMDb (Internet Movie Database) is an [online database](https://en.wikipedia.org/wiki/Online_database) of information related to films, television programs, home videos and video games, and internet streams, including cast, production crew and personnel biographies, plot summaries, trivia, and fan reviews and ratings. IMDb registered users can cast a vote (from 1 to 10) on every released title in the database. Individual votes are then aggregated and summarized as a single IMDb rating, visible on the title’s main page. By “released title” we mean that the movie (or TV show) must have been shown publicly at least once (including festival screening). To make sure all its users have access to quality, IMDb uses different metrics to rank movies. Though IMDb does not reveal the algorithm it uses, it is evident that it ranks movies based on;

1. Number of votes
2. Movie ratings
3. User reviews
4. How helpful is the review to other users?

While IMDb uses the above metrics, it is evident that the number of votes a movie project determines its ratings and how high it ranks.

**Does Rating matter?**

IMDb ratings are critical. They not only help IMDb to rank films but also help users make decisions on whether to watch your movie or not. Once a new title is released, IMDb makes it available to users. Users have the power to cast votes and review all new releases in regards to their experience.

To keep the system fair to all each user is only allowed to rate one movie at a time. This means that one user cannot rate the same movie twice.

However, all IMDb users, have the privilege to update their rating often as they wish. Once a user updates the rating, the new rating overwrites the previous one. This is one of the strategies IMDb uses to reduce ballot stuffing and maintain quality and fairness.

To makes sure users have access to high-quality movies, IMDb further allows users to rate other user reviews on how helpful they are. They achieve this by prompting users to answer with a yes or no. A user review is termed as useful if it helps other users make decisions. For instance, if a movie is of poor quality and a user leaves a poor review, it helps other users to save time and money, and in this case, it is regarded as helpful.

Rating represents the relationship between rating and box office success**.** Correlation between the box office success and positive reviews and also the relationship between the box office success and movie performance.

**5 chosen questions form part C**

1. Top 15 highest rated movies of year 2006 to 2016
2. Dual axis chart of Budget VS Revenue of top 15 movies from 2006 to 2016
3. Release date VS Revenue for the top 15 movies of years 2006 to 2016
4. Vote count VS Rating
5. The average Revenue of top 15 movies of years 2006 to 2016

**Analysis:**

Analysis has different aspects which needs to be considered. The top rated movies are determined using different measures. This dataset contains movies from year 2006 to 2016. It contains 1000 data entries with the following list: Budget, Id, Keywords, Original language, Original title, Overview, Popularity, Production Company, Production Country, Release date, Revenue, Runtime, Spoken languages, Status, Tagline Title, Vote average, Vote count, their ratings, and genres. (Distribution of votes for each rating). The different types of genres are (action, animation, comedy, fantasy, drama, and Science Fiction).

**Top 15 highest rated movies of year 2006 to 2016**

This story point determines the top 15 movies of year 2006 to 2016 and it uses rating as the measure. This help us to determine that which movie is popular based on the count of rating. According to this analysis, we can determine the most top movie on basis of rating count. In year, 2014 the highest rating is 8.6 for the movie Interstellar directed by Christopher Nolan.

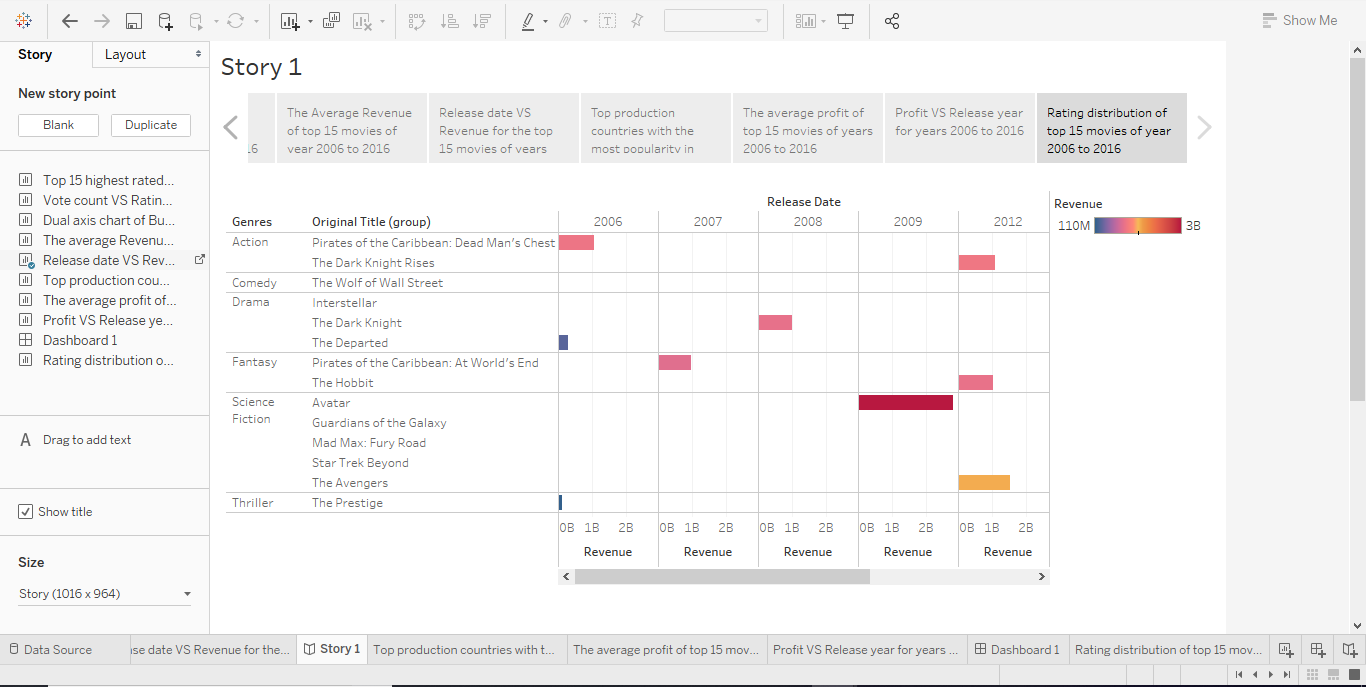
**Dual axis chart of Budget VS Revenue of top 15 movies from 2006 to 2016**

This story point determines the budget and revenue of the top 10 movies between 2006- 2016 as the measure. This helps us to determine the highest budget and revenue movie year 2012 where, budget is 720,000,000 and revenue is 3,625,600,577.

**Release date Vs Revenue of Top 15 Movies from 2006-2016**

This story point determines the revenue and releasing date of top 15 movies of year 2006 to 2016. Here The Hobbit: an unexpected journey and the desolation of smaug are grouped.

On base of the count of rating we can predict the highest revenue movie in as 2,787,965,087 in year 2009 and lowest in the year 2006 as 289,847,354.

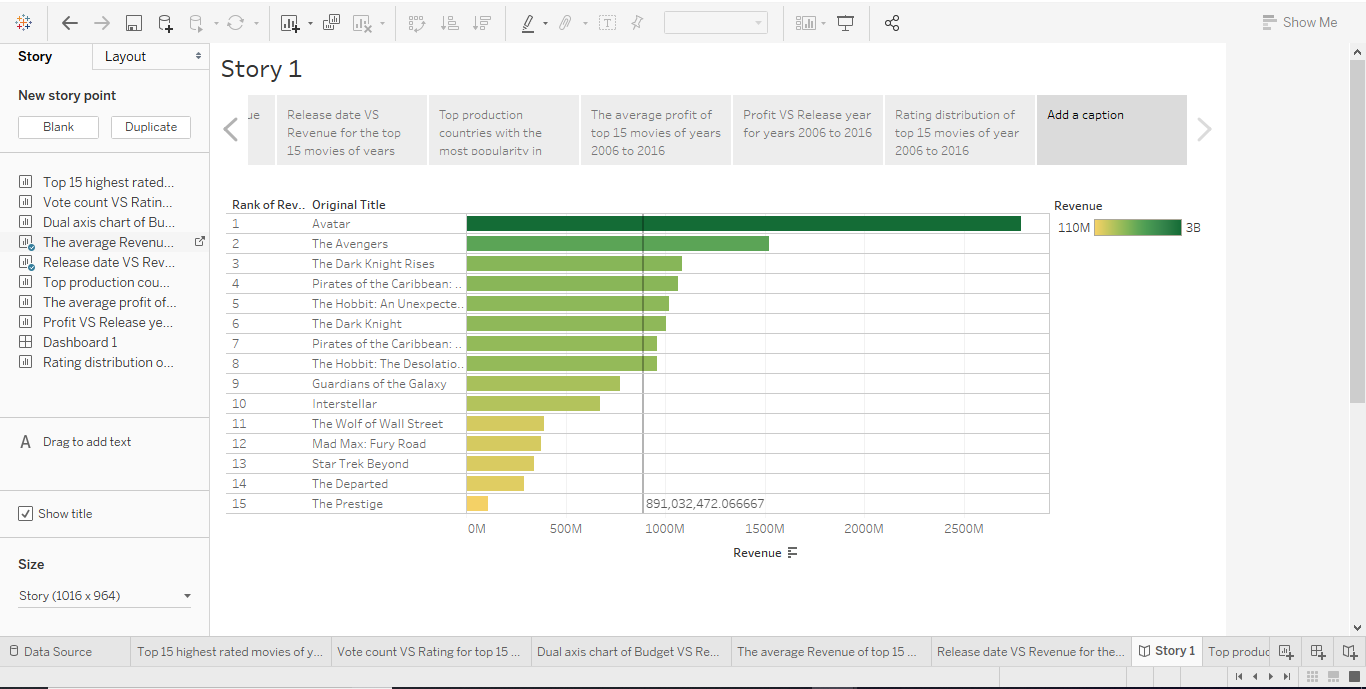


**Vote count VS Rating:**

This story point determines the Correlation between the rating and vote count of top 15 movies of year 2006 to 2016. The Dark Knight has the highest rating of 9.000 and vote count of 12,002.

**The average Revenue of top 15 movies of year 2006 to 2016**

This story point determines the top 15 movies of year 2006 to 2016. This analysis help us to determine that ranked movie is popular based on average revenue. The average revenue is 891,032,472.066667 which was calculated using the Reference line. This analysis help us to determine that ranked movie is popular based on its revenue.



**Reference:**

1. Wikipedia. IMDb. <https://en.wikipedia.org/wiki/IMDb>. accessed April 5, 2019.
2. IMDb Help Center. FAQ for IMDb Ratings. [https://help.imdb.com/article/imdb/track-movies-tv/faq-for-imdb-ratings/G67Y87TFYYP6TWAV#](https://help.imdb.com/article/imdb/track-movies-tv/faq-for-imdb-ratings/G67Y87TFYYP6TWAV) .accessed by April 5, 2019.
3. Posting Bros. How IMDB Rating Algorithm Works. <https://clflaggingexpert.net/how-imdb-rating-algorithm-works/>. accessed by April 5, 2019.