Data Exploration and Storyboard on the Dataset - IMDB 5000 Dataset

Group 16 - Akshat Tyagi (at3761), Harshit Srivastava (hs3500), Hamza Mirza (online, hm1800)

Data Set

The dataset contains information about movies that are listed on IMDB, aggregated at the movie level. So, it contains information in the following manner:
Director - Name of the director (categorical)
Actor - Names of 3 actors (categorical)
Facebook likes (0-1M) - director, actor and total cast (quantitative)
Genre - movie genre (categorical)
No. of user reviews (0-5000) - no. of reviews submitted by users (quantitative)
No. of critic reviews (0-800) - no. of reviews submitted by critics (quantitative)
IMDB rating (0-10) - IMDB rating considering both user reviews and critics' reviews (quantitative, ordinal)
Duration (0-500) - Duration of the movie in mins (quantitative, ordinal)
Title_year - Year when the movie was released. (quantitative, temporal)
Language - Language in which movie was released. (categorical)
Budget - Amount of money spent on making a movie. (quantitative)
Gross - Amount earned by the money during its run. (quantitative)
Movie_title - The name of the movie. (Categorical)
Plot_keywords - Some words used to describe a movie (Categorical).

Dataset Link - (https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset)

Analytical Questions and Proxy Tasks

We will explore the IMDB movie reviews dataset to answer some pertinent questions regarding how movies are rated based on a plethora of factors. We would like to answer the following questions -

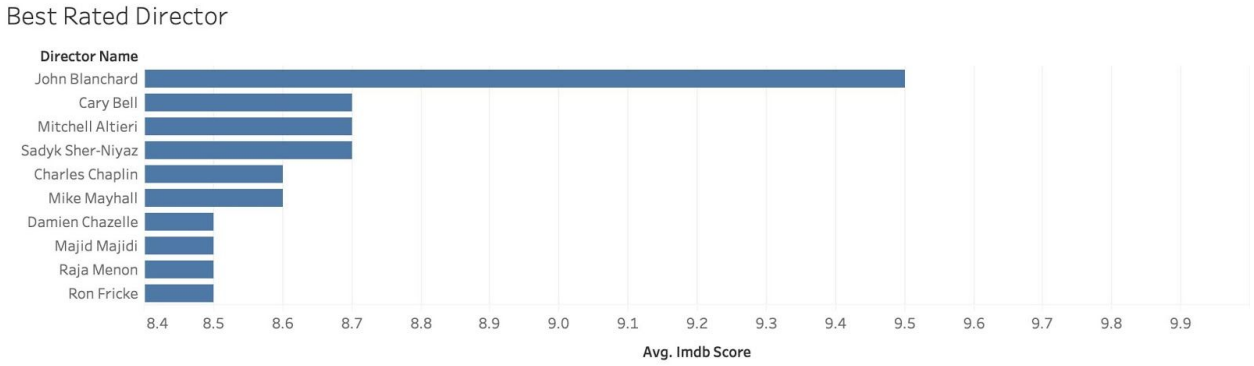
Q1. Who is the best-rated director?
Proxy Task: Analyzing distribution of movie ratings for different directors
Proxy Value: best-rated -> high IMDB rating
Q2. Which are the best-rated movies?
Proxy Task: Visualizing features of best-rated movies
Proxy Value: best-rated -> high IMDB rating
Q3. Does the movie rating change with time?
Proxy Task: Does the IMDB rating change with title_year?
Proxy Value: movie rating -> IMDB Rating, time -> title_year
Q4. Are movies in some languages better rated than movies in others?
Proxy Task: Visualizing rating distribution for movies of different languages
Proxy Value: better rated -> high IMDB rating, language -> language
Q5. A movie is a hit if it earns more money than it costs. What are some of the biggest hits? Do some directors have more hits than others?
Proxy Task: Do movie_titles from some directors have a higher difference between budget and gross than others?
Proxy Value: movie -> movie_title, director -> director, earns -> gross, costs -> budget.
Q6. Does the plot keyword of movies change with time?
Proxy Task: Do aggregated plot_titles for the movie_title change with title_year?
Proxy Value: plot keyword -> plot_titles, movie -> movie_title, time -> title_year
Q7. Does the movie genre have any effect on the movie's rating?
Proxy Task: What is the average movie rating for a given genre? Do any specific genre has a higher average rating compared to other genres?
Proxy Value: higher average rating -> avg(IMDB Rating), genre -> genres
Q8. Does the budget of the movie relate to the IMDB rating?
Proxy Task: What is the average IMDB rating for a given budget range?
Proxy Value: Average IMDB rating -> avg(IMDB rating), budget range -> budget

In order to answer these questions, we would like to create interactive visualizations where we can change the visualized data based on the attributes. The website would consist of a dashboard with multiple visualizations answering the above questions. We would like our website to be dynamic such that it can process the data for the visualizations on the backend in less time. For this, we would query our data with cross filter, which would be used for the visualizations. With the interactive nature of the visualizations, not only will we answer some important questions about the dataset, but we would also like the user to interact with the visualizations and form their own questions.

Story Design

Data Analysis

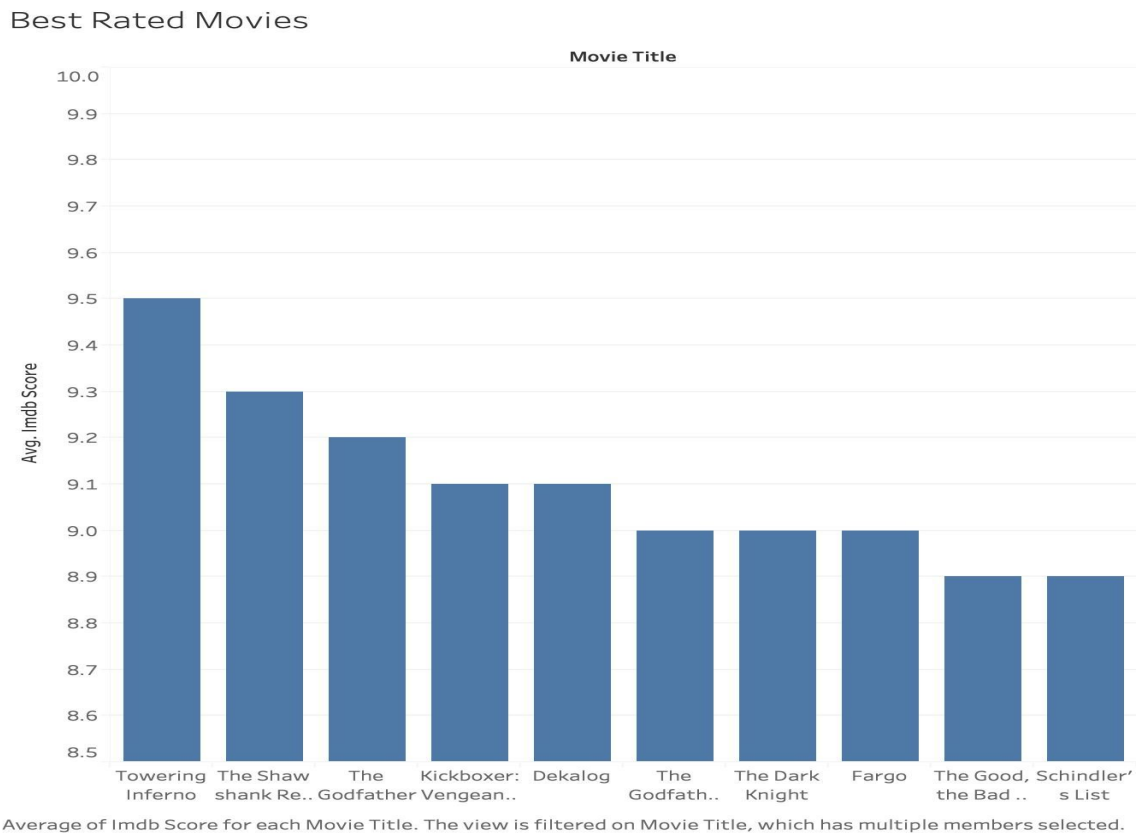
Q1. Who is the best-rated director?



Average of Imdb Score for each Director Name. The view is filtered on Director Name, which has multiple members selected.

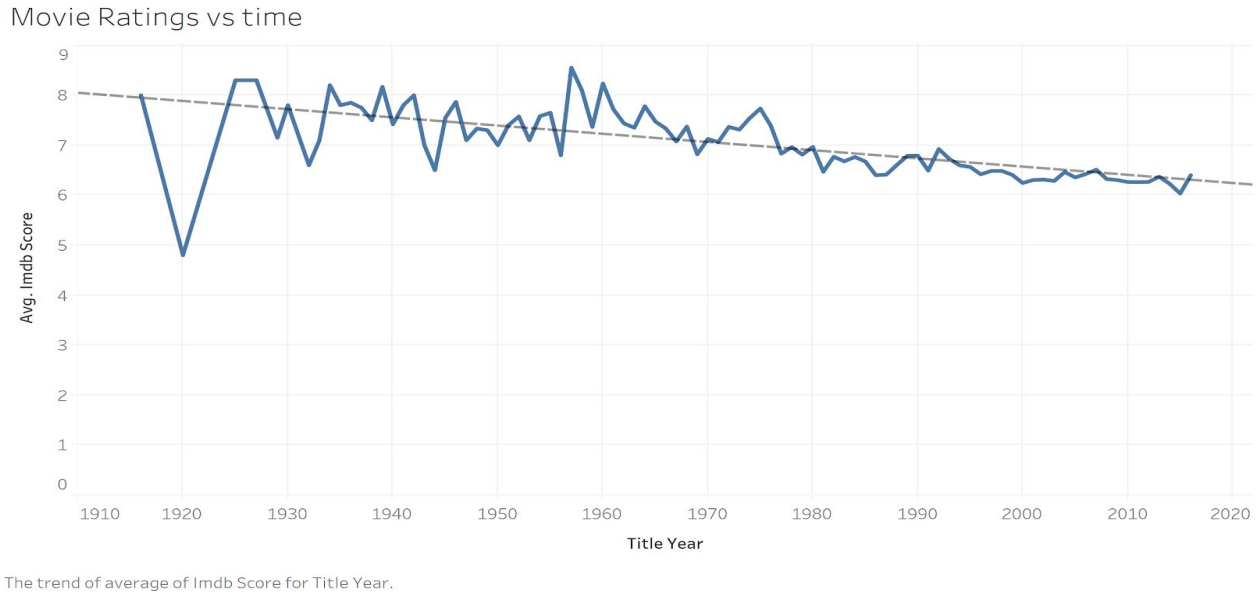
It was observed that John Blanchard , Cary Bell, and Mitchell Altieri are the best directors because their movies have the highest ratings.

Q2. Which are the best-rated movies?



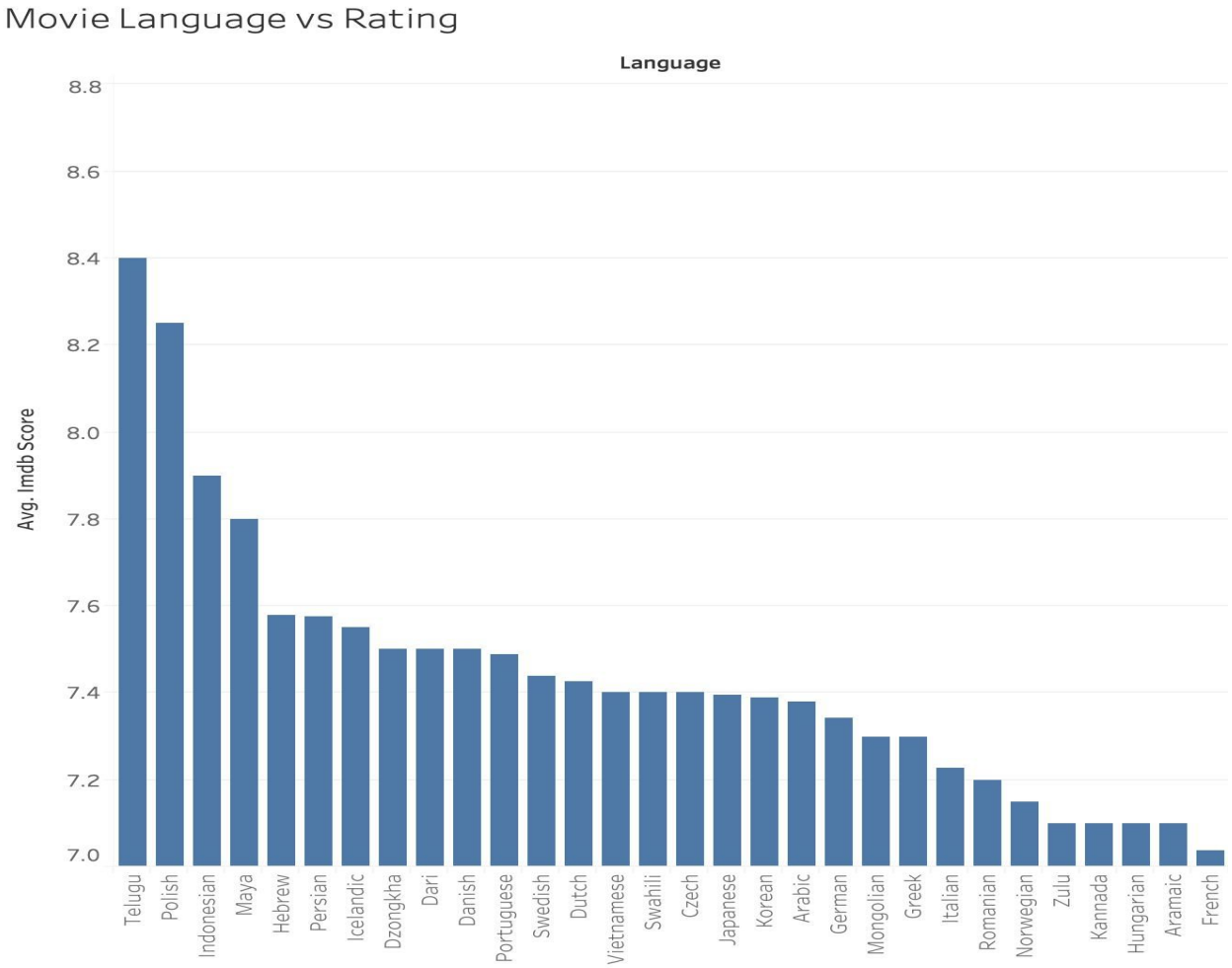
Towering Inferno, The Shawshank Redemption, and The Godfather are the highest rated movies.

Q3. Does the movie rating change with time?



Yes, it was observed that time time, the average IMDB rating of a movie decreased, as shown above.

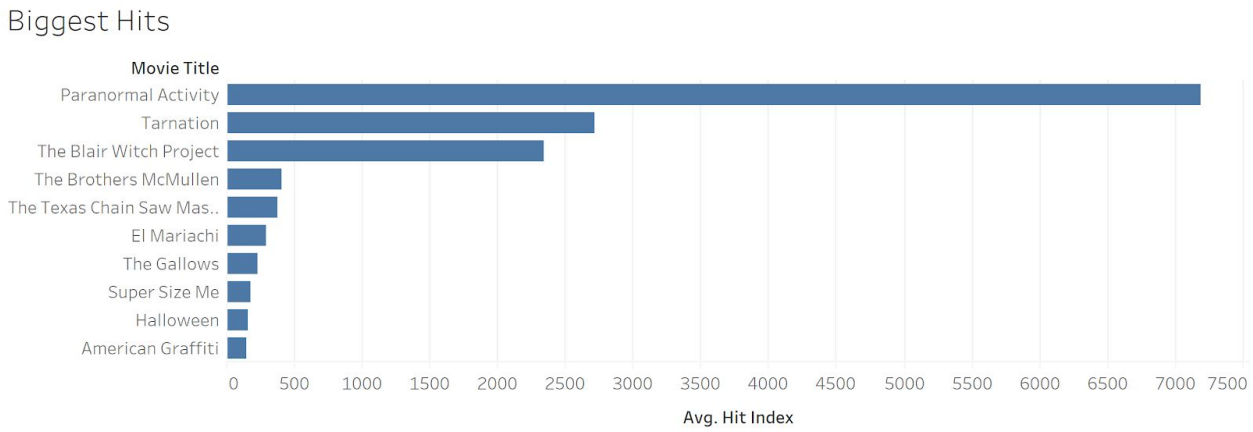
Q4. Are movies in some languages better rated than movies in others?



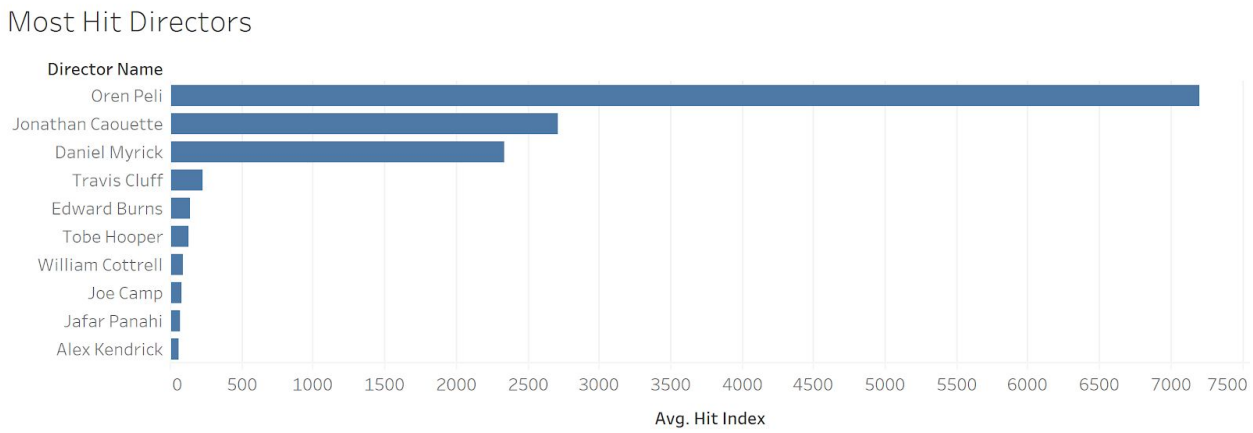
Average of Imdb Score for each Language. The view is filtered on Language, which keeps 30 of 48 members.

The above graph shows top 30 highly-rated languages. As shown above, it's clear that while movies in Telugu, Polish, and Indonesian have high ratings, movies in Chinese, Tamil, and Bosnian have low ratings. Thus, movies in some languages are rated better than others.

Q5. A movie is a hit if it earns more money than it costs. What are some of the biggest hits? Do some directors have more hits than others?



Average of Hit Index for each Movie Title. The view is filtered on Movie Title, which keeps 10 of 4,916 members.



Average of Hit Index for each Director Name. The view is filtered on Director Name, which keeps 10 of 2,399 members.

Based on the analysis, Paranormal Activity is the biggest hit, followed by Tarnation and The Blair Witch Project. Oren Peli, Jonathan Caouette, and Daniel Myrick are the highest rated directors.

Q6. Does the plot keyword of movies change with time?

There was no change observed.

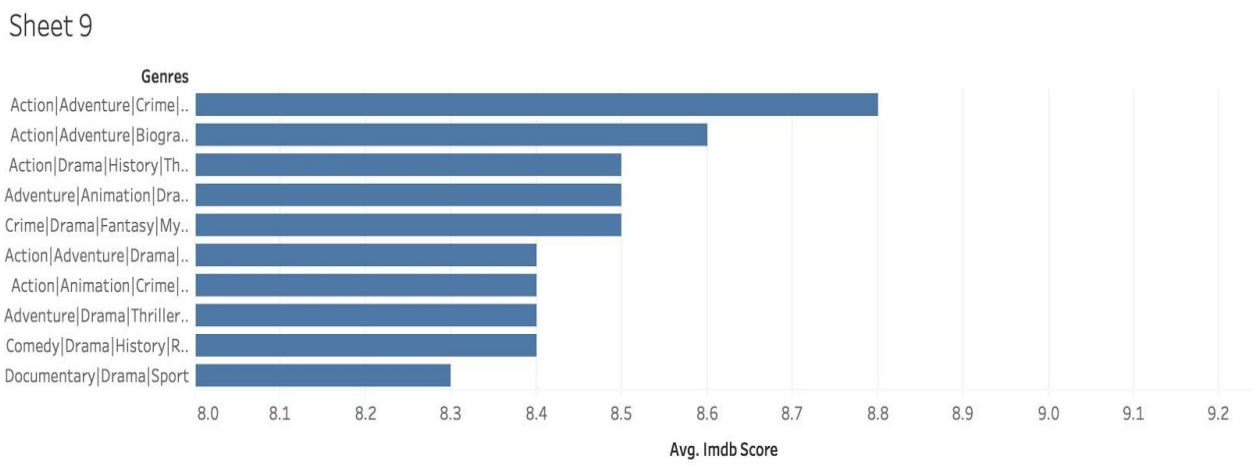
We have decided to omit this question as its not pertinent to our storytelling and we were rightly advised by the TAs to not consider this question.

Plot Keywords. Color shows details about Plot Keywords. Size shows count of Plot Keywords.

Q7. Does the movie genre have any effect on the movie's rating?

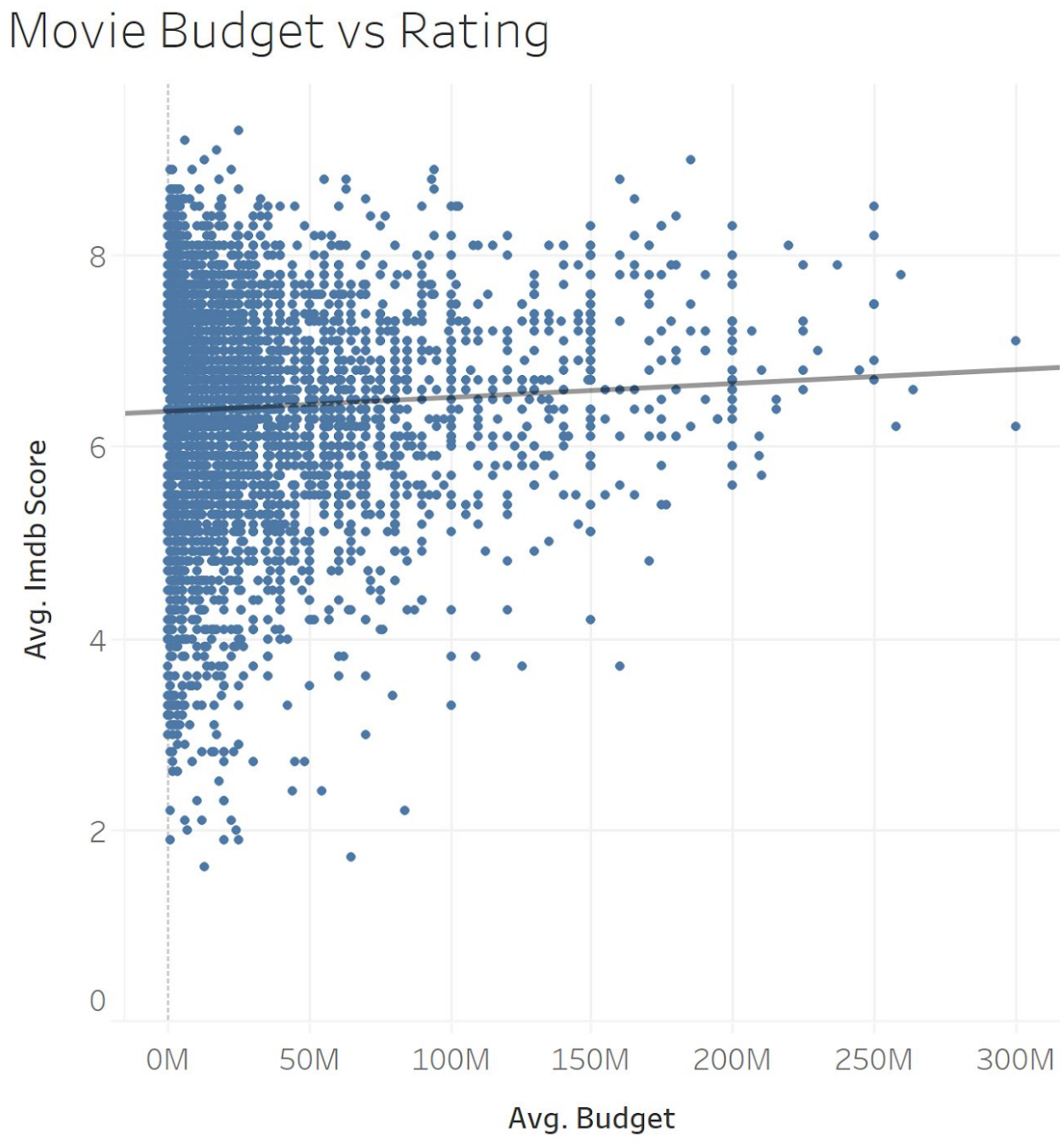
Movies with tags such as Action, Adventure, Crime etc have better ratings than ones with the tags - Documentary, Drama, Sport.

We have decided to omit this question because of multi-class genres present in the data and there's no good way to aggregate it, on top of it, it doesn't really help with the story-telling.



Average of Imdb Score for each Genres. The view is filtered on Genres, which keeps 10 of 914 members.

Q8. Does the budget of the movie relate to the IMDB rating?

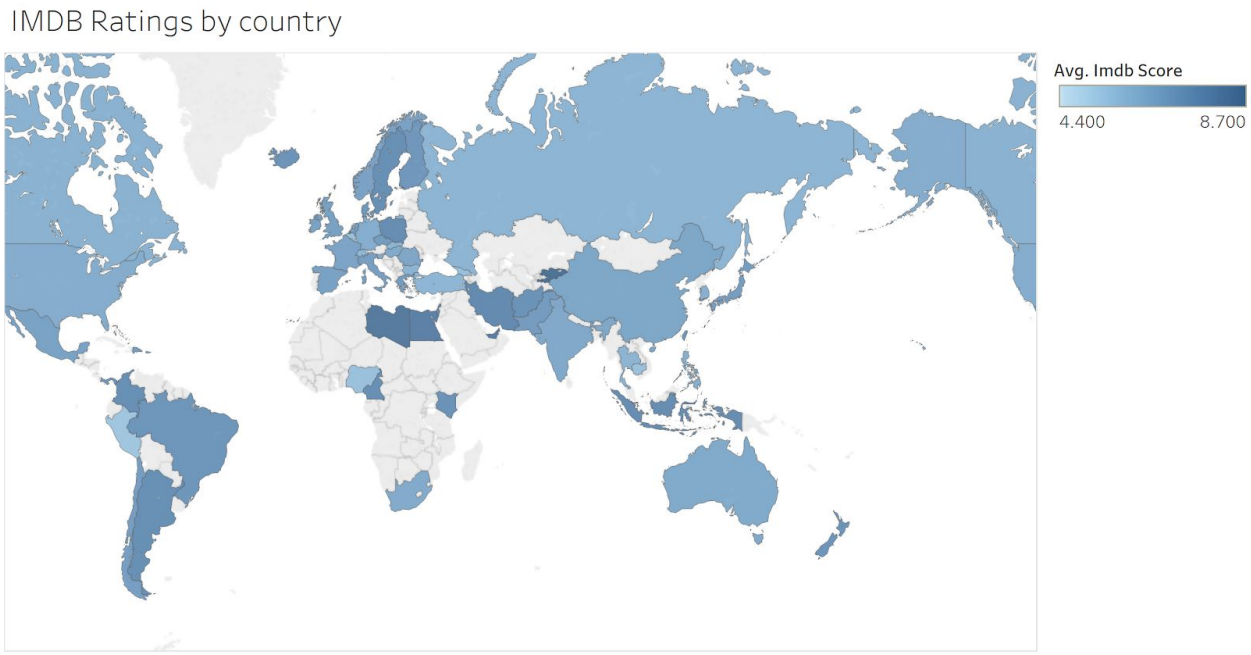


Average of Budget vs. average of Imdb Score. Details are shown for Movie Title. The view is filtered on average of Budget, which ranges from 218 to 323,253,073.

Yes, as shown above, the average imdb score and average movie budget have a positive correlation.

Additional Questions Asked after observing the dataset -

Q9. Does movie rating vary by country?



Yes. Countries like Iran, Egypt, and Libya have better rated movies than Russia, Nigeria etc.

Q11. Is the money a movie makes relating to its rating?

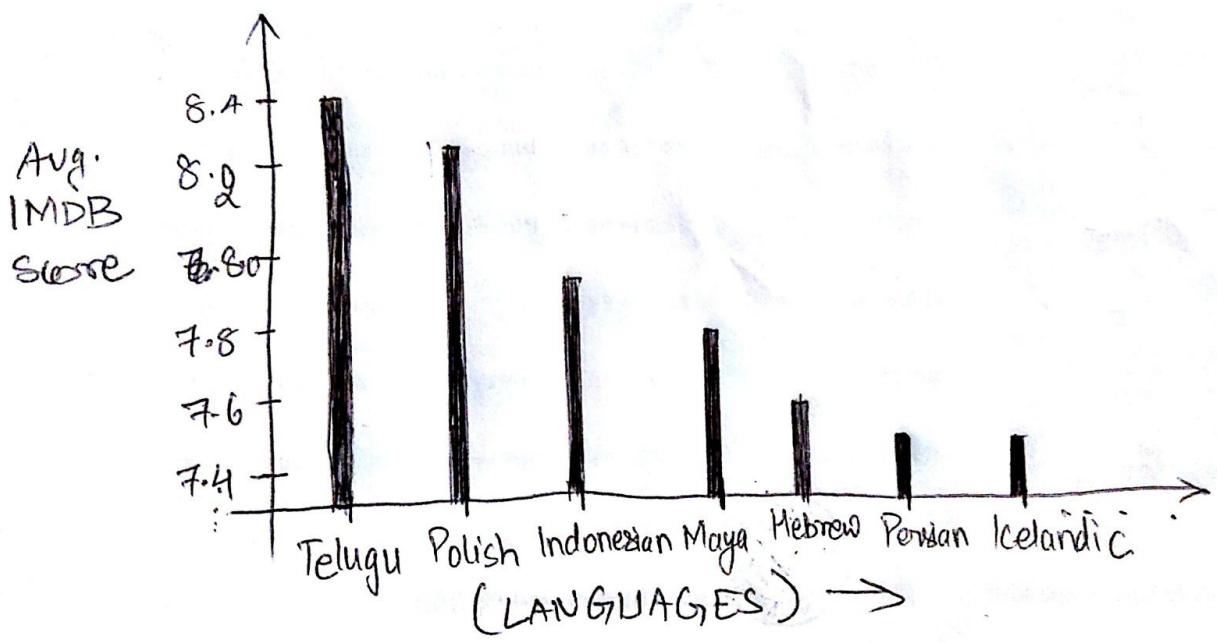
Average of Gross vs. average of Imdb Score. Details are shown for Movie Title.

Yes. Movies which have a good IMDB rating tend to earn more, as shown above.

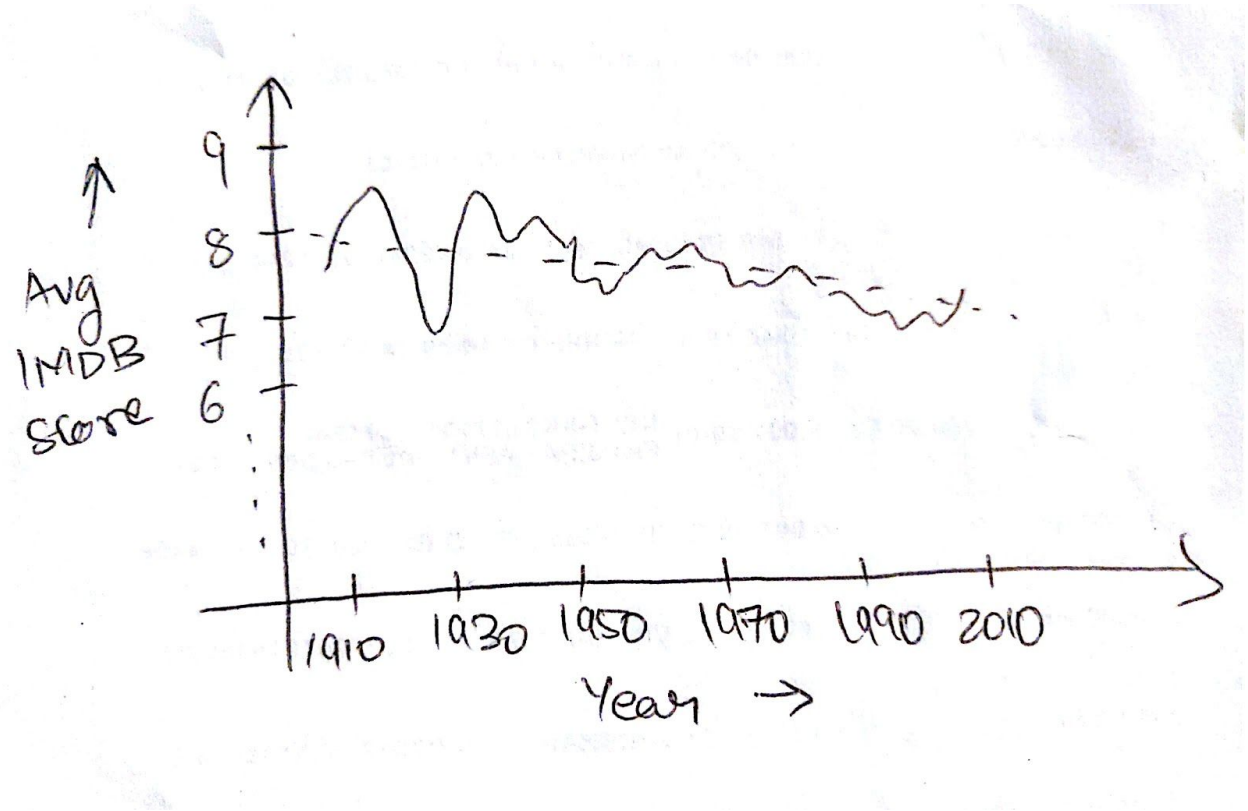
Storyboard

The IMDB dataset is fascinating because it is allowing us to observe some general trends about movies. A movie's performance depends on a plethora of factors, and we will observe a few of them today.

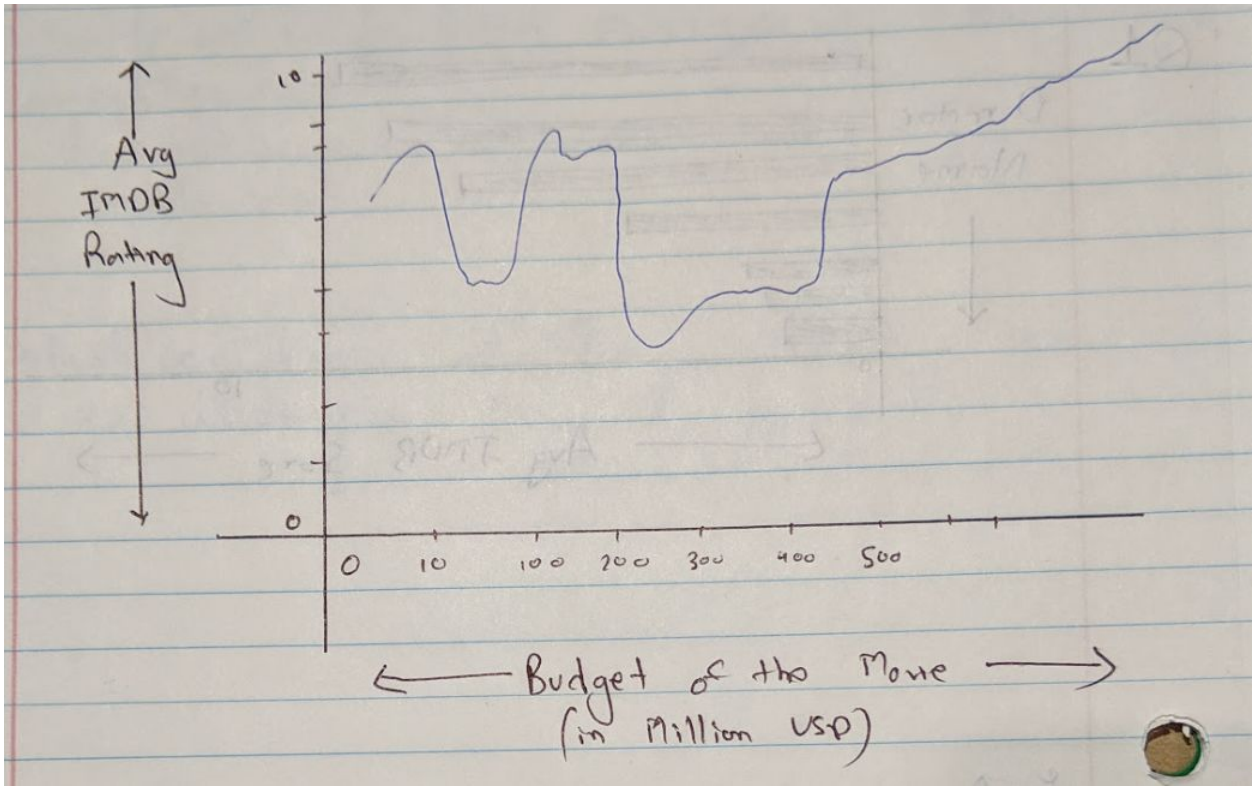
First, we observed that it's not just the content of the movie, but also its origins that affect its rating. As shown below, we can see that movies from some parts of the world and some languages do better than others.



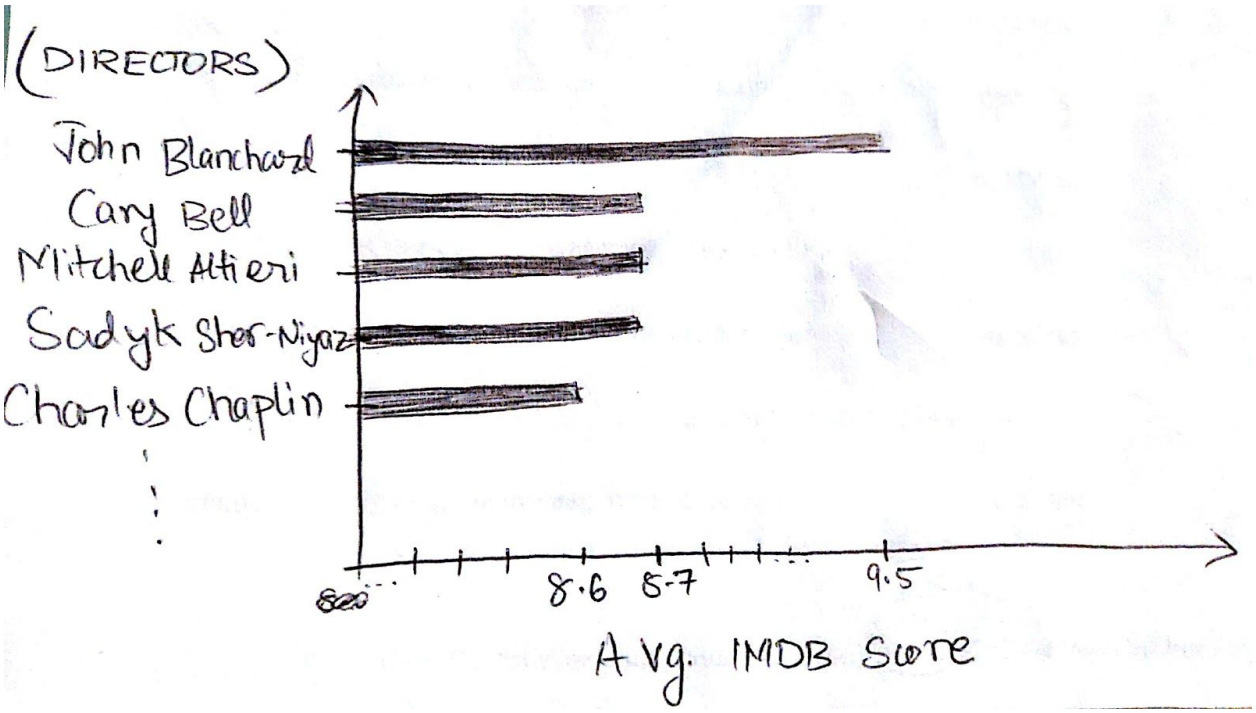
People mostly say that movies made in the olden days were better than the ones made today. We observed this hypothesis and concluded that the hypothesis holds true, based on the ratings of movies based on their release date, as shown below.



Now that we have established that a movie's origins plays a crucial role in its rating, we would also like to see if spending more money means that the movie is good. As shown below, the amount of money spent on a movie affects its rating.

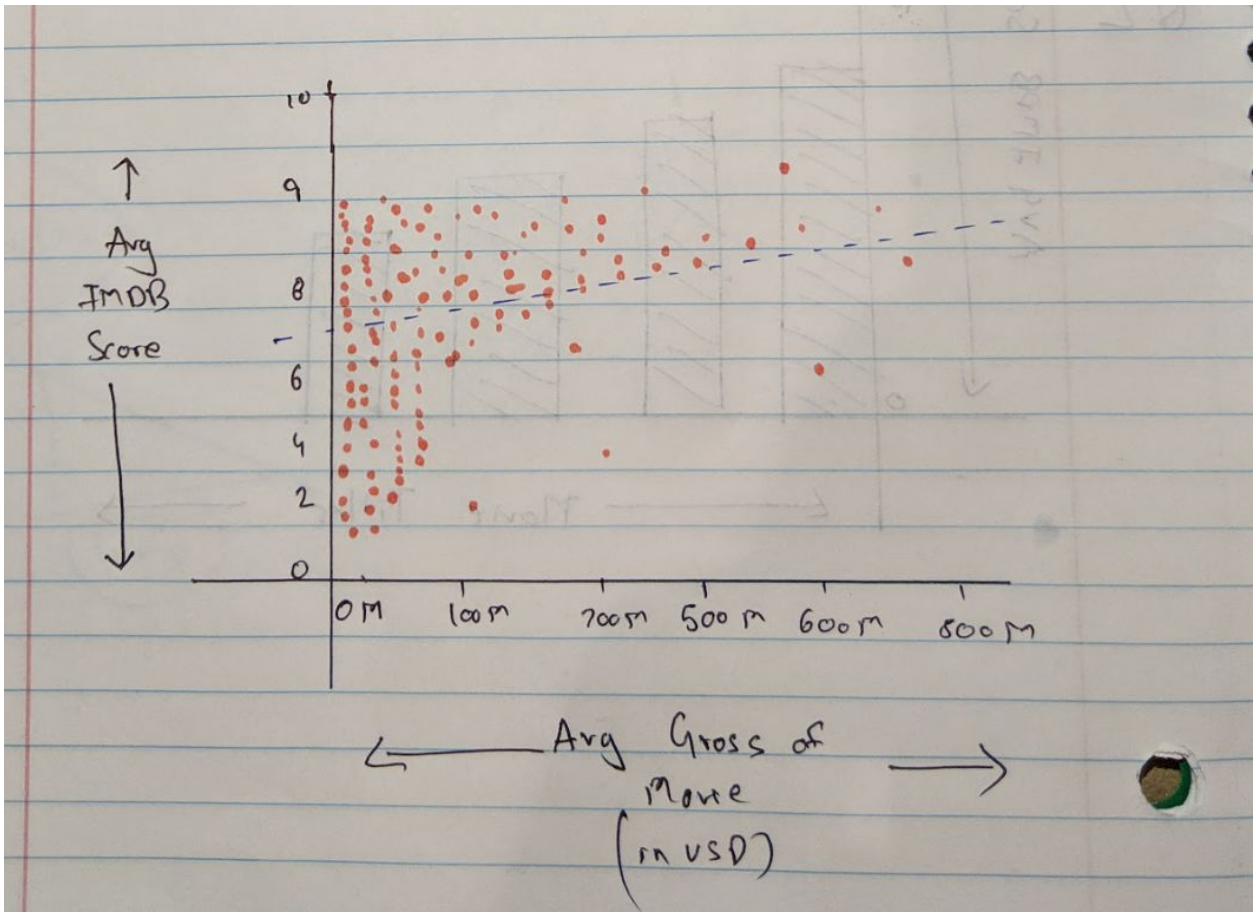


Since how well a movie is made depends a lot on directors, let's look at some of the best directors out there, based on how hit their movies were -

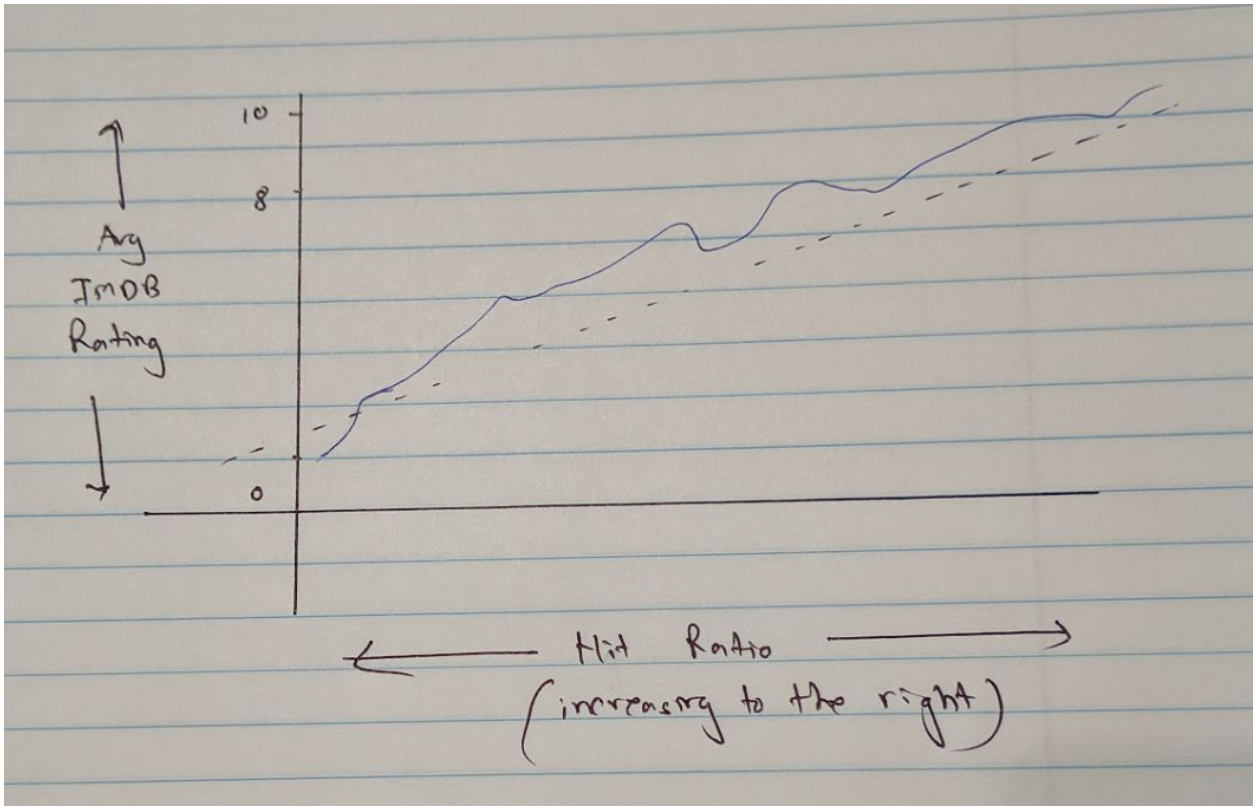


So far, we have observed indicators that could potentially predict the performance of a movie. Now let's look at some indicators which quantify a how good or bad a movie is.

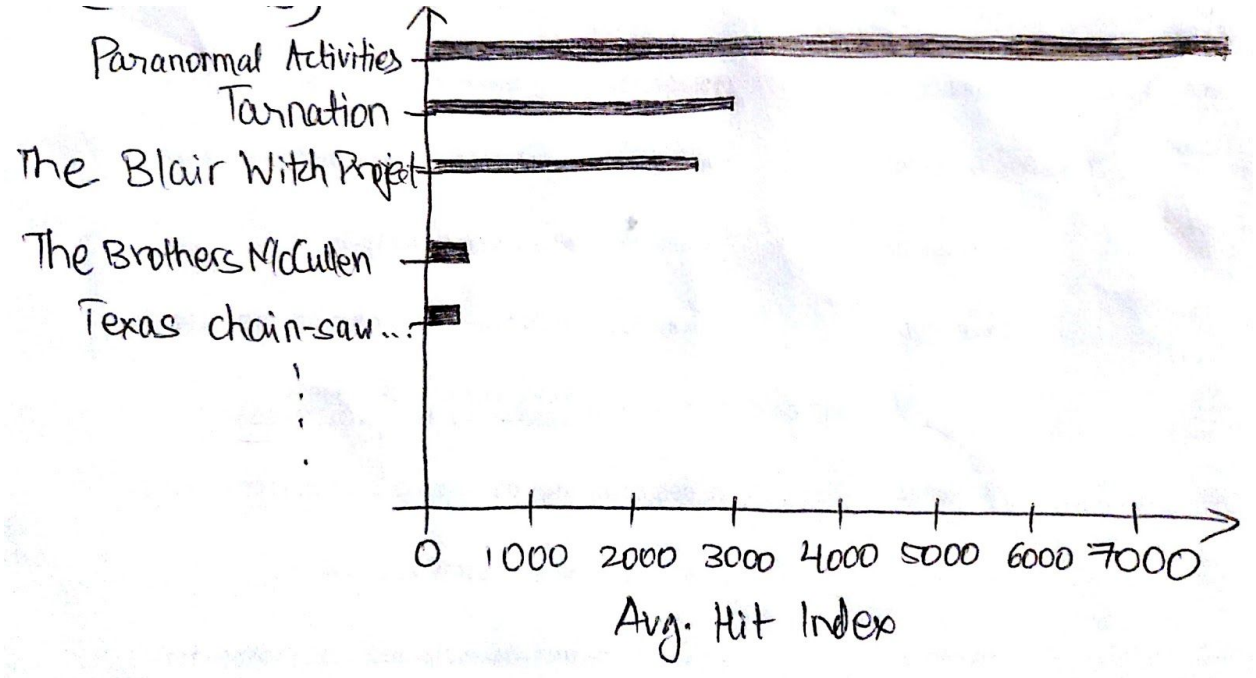
By observing the visualization below, its clear that the money a movie makes is proportional to its IMDB rating. This means that there is a general trend where audience appreciates well made movies.



Now lets see if movies that are hit, are hits because they're well made or not. We can observe from the graph below that well rated movies turn out to be hits. Some, like Paranormal activity make a lot of money for a small budget.



To recapitulate, let's look at the highest rated movies of all times -



We can conclude that a movie's success is heavily dependant on its rating. But it's also dependant on a plethora of factors, including the language its made in, time when it was released, and the region it comes from.

Changelog

We answered two additional questions which we felt were germane to the analysis. Those questions are -

Q1. Does the average movie rating vary by country?
Q2. Does the amount of money a movie earn depend on its rating?

Phase 1 Update

By now, we have a working web page implemented. We have added the text from our stories and are currently working on adding the visualizations. We have added three new choropleth maps, with the possibility of adding a fourth, in addition to the original proposal. The new maps are as follows - Average movie budget for each country, average number of facebook likes per country, average number of votes per country. We are deliberating on adding a fourth choropleth map to visualize the average budget per country.

LIST OF IMPLEMENTED ITEMS -

1. Choropleth Map to visualize the average IMDB score grouped by the country of origin.
2. Choropleth Map to visualize the average movie budget grouped by the country of origin.
3. Choropleth Map to visualize the average number of facebook likes for the movie grouped by the country of origin.
4. Choropleth Map to visualize the average number of votes for the movie grouped by the country of origin.
5. Helper functions to load specific data, group them according to some feature, and scale them up.
6. Helper functions to create choropleth maps and assign specific properties to them.
7. Helper functions to create bar charts.

8. Bar chart for language vs imdb_score.

LIST OF ITEMS TO BE IMPLEMENTED

1. Helper functions to create scatter plots.
2. Creation of remaining bar charts and scatter plots using the helper functions for other visualizations in the project.
3. Making the visualizations interactive by adding tooltips, animations, filters etc.

CURRENT VERSION

https://github.com/NYU-VIS-FALL2018/storytelling-group-16