# Exploratory Visual System for IMDB Movies


Fig. 1. Central Dashboard View

**Abstract**— Among all the methods humans use to explore the data, visual exploration is proven to be the most effective method. We have created an interactive dashboard for IMDB movies dataset in order to allow users to explore the same visually. We have implemented this system using various packages offered in R platform. This work is inspired by various IEEE papers and articles. This dashboard attempts to demonstrate some aspects of data exploration which are uniquely suitable for visual interactions with movies data. This is best exemplified by network graph among other interactions explained by certain scenarios in this report. Users can explore movies based on their choices with visual interaction. This dashboard translates these interactions into filters at data level and updates the visualizations accordingly. For example, users can explore movies based on IMDB rating, Genre, Movie duration, country etc. This dashboard can also be used for analysis of relations between movies, actors, budget, gross, ratings,return on investment etc. to gain some interesting insights on the roles played by these attributes.

**Index Terms**— Visual Analytics, Interactive Dashboard, Visual exploration of movies, Movies Dashboard, Actors network, ROI of Movies, Movies Watchlist, R Implementation.

✦

## 1 Introduction

The film industry or motion picture industry comprises the technological and commercial institutions of filmmaking. Currently, the largest markets by box office are United States, China, United Kingdom, Japan, South Korea and India. The cinema of the United States, often generally referred to as Hollywood, has had a profound effect on cinema across the world since the early 20th century. The United States cinema (Hollywood) is the oldest film industry in the world which originated more than 121 years ago and the largest film industry in terms of revenue. Between 2009-2015, Hollywood consistently grossed $10 billion (or more) annually. The Internet Movie Database (abbreviated IMDb) is an online database of information related to films, television programs and video games, including cast, production crew, fictional characters, biographies, plot summaries, trivia and reviews, operated by IMDb.com, Inc.

The data used for this dashboard is obtained from Kaggle which was founded in 2010 as a platform for predictive modelling and
analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. The data was scraped for 5000+ movies from IMDB website using a Python library called "scrapy". This dataset contains attributes related to actors, review ratings, budget, box-office gross collection, year of release etc.

The design for this exploratory dashboard is inspired by various IEEE papers and articles. Our main idea is to demonstrate how visual exploration of data can provide unique insights not easily feasible by traditional methods.

After analysing various JavaScript and R libraries, we have found R shiny with DCR package most suitable for our design. It provides a layer of abstraction above raw data and dynamically adjusts the working set based on the user selection of various graph areas, thereby minimizing lines of code necessary for developing such dashboard. In addition to offering rapid development, the code is highly maintainable.

## 2 System Design

The Primary goal of this system design is to assist users to provide a watch list of movies based on their preferences using this highly interactive dashboard and views.

For instance, a user who prefers to watch short and slick Hollywood Thrillers spanning around 100 min, this easy-to-use interactive dashboard can be helpful. The user can can seamlessly switch between various visualizations to get the list of movies that match his/her criteria. This process would have been very complicated for an average movie-goer to go online and search all one by one.

Our Secondary goal is to assist users who are keen to watch movies of their favourite artists sharing screen space. Navigating through Actors Network Visualization dashboard, such users can find it extremely easy to check out the networks of their favourite actors and movies associated among them. Additionally we can explore the actors network and actors who are working in closed networks.

In addition, few movie-lovers are particularly interested in Economics and profitability of movie making. ROI dashboard provide such insights using interactive bubble charts with comparison of budget vs box-office collection.

The visual design of these dashboards are developed on R platform with various libraries such as dcr, shiny, plotly, vis-network etc.

### 2.1 Introduction to dataset

This dataset contains 5000+ movies from IMDB website. It has 28 variables for 5043 movies and 4906 posters (998MB), spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses.

Below are the 28 variables:

| Variable Name | Variable Description |
|---|---|
| movie_title | Name of the movie |
| color | If picture was black/white or in color |
| num_critic_for_reviews | Total number of critics |
| movie_facebook_likes | Total number of likes the movie got on Facebook |
| duration | Length of the movie in minutes |
| director_name | Name of the director |
| director_facebook_likes | Number of likes the director has on Facebook |
| actor_3_name | The third actor's name |
| actor_3_facebook_likes | Number of likes 3rd actor has on Facebook |
| actor_2_name | The seconds actor's name |
| actor_2_facebook_likes | Number of likes 2nd actor has on Facebook |
| actor_1_name | The first actor's name |
| actor_1_facebook_likes | Number of likes the 1st actor has on Facebook |
| gross | Gross/ profit w/o tax removed |
| genres | Type of movie (horror, action) |
| num_voted_users | Total number of user that voted |
| cast_total_facebook_likes | Total number of likes the cast got on Facebook |
| facenumber_in_poster | Total number of actor faces on the poster |
| plot_keywords | Keywords in the movie |
| movie_imdb_link | Link to movie on IMDB |
| num_user_for_reviews | Number of comments |
| language | Language movie is in |
| country | Country movie was made |
| content_rating | Content rating( PG, G) |
| budget | Budget used to make movie |
| title_year | Year movie was made/released |
| imdb_score | Average IMDB Rating |
| aspect_ratio | Aspect ratio/ ratio of the width to the height of an image/screen |

Table1: Data Variables

Although dataset contains many variables, certain variables such as aspect_ratio, facenumber_in_poster etc., were not involved in building the dashboards. We added some extra-calculated variables to this dataset such as duration range of the movies whenever it was necessary for us to incorporate into our dashboard.

### 2.2 Central Interactive Dashboard:

Central dashboard consists of below depicted visualizations, which are self-explanatory. All the visualizations listed below relate to working set binding. Each visualization can be used to narrow/expand the selection of the data which is reflected in other visualizations accordingly. This dashboard can assist the user in shortlisting a watch list of the movies that satisfy his/her preferences.

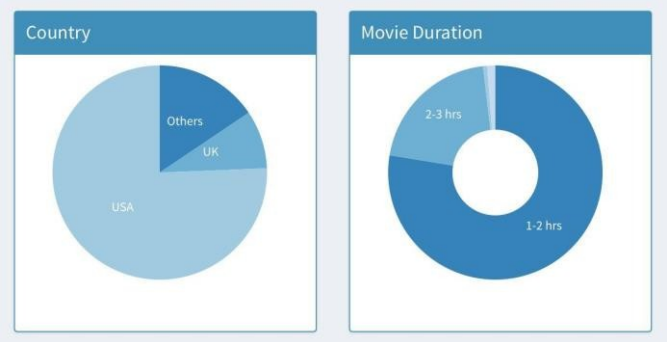


Fig. 2 Pie chart for Countries and Donut chart for Movie Duration

The above dashboard consists of interactive Pie and donut Chart which lets the user to click on a portion of chart and displays the corresponding results in remaining charts. Pie Chart has all the movies categorised into the countries of production like USA, UK and other countries. Donut chart has movies categorised based on ranges of movies duration.

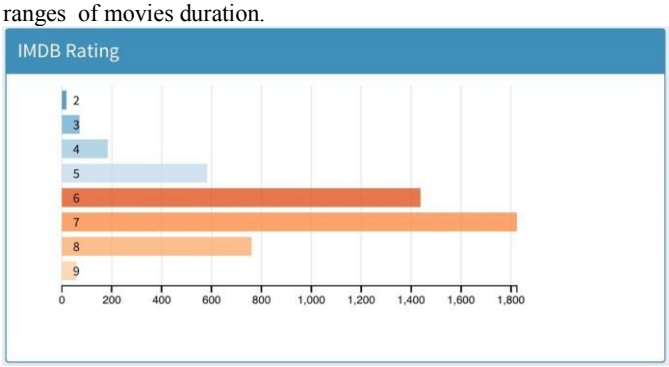


Fig. 3. IMDB Ratings horizontal bar chart

The above section on the central dashboard helps in drilling down the movie list with a certain movie rating.

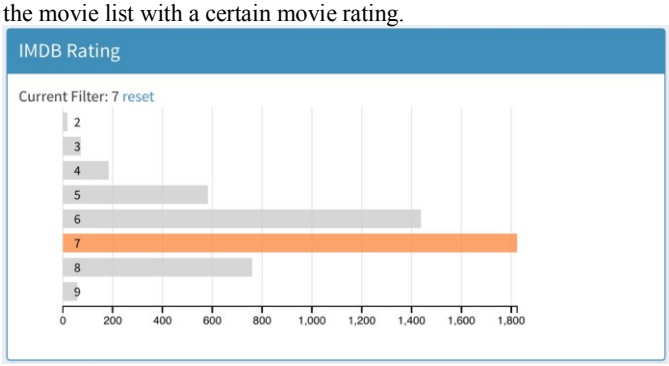


Fig. 4.1 IMDB Ratings horizontal bar chart with filter on IMDB rating =7

After selecting a particular movie rating , user is presented with various genres with associated number of movies within that rating criterion.[Please refer image below]

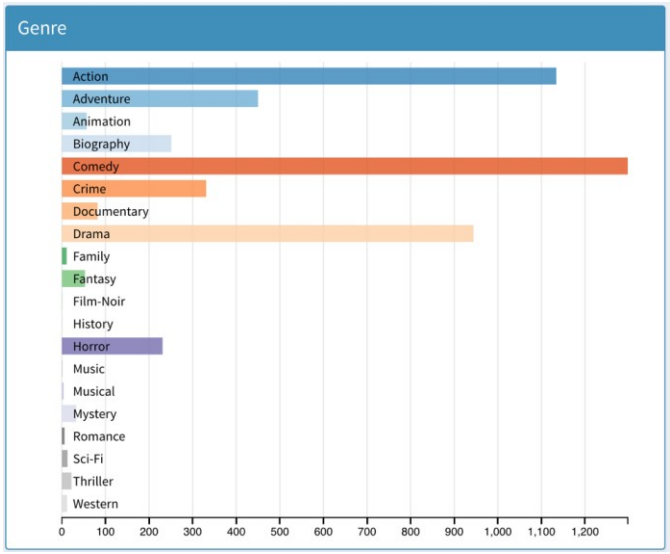


Fig. 4.2 Movie Genre with associated count- row chart

Using this dashboard users can perform wider range of explorative tasks as huge amount of combinations of drill down filters are possible. For example, users can find a movie genre with highest number of movies for rating less than 7.0. Users can explore the same movie genre and associated count of movies for ratings greater than or equal to 7.0. If this count decreases, it can be concluded that audience were more critical than embracing it. If this count increases, then it can be concluded that the genre received more positive response than negative. (Such analysis can actually be extended across US / UK/Other countries)

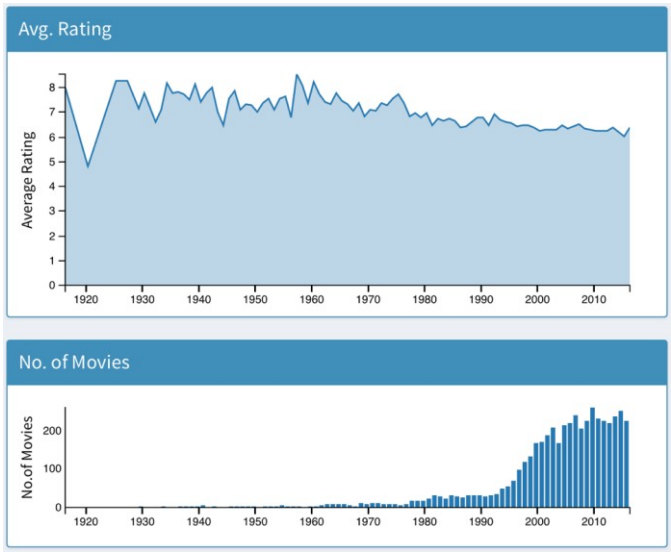


Fig. 5. Average movie rating and number of movies across the years

Above dashboard section displays the number of movies made across different years through an interactive bar chart. As clearly seen, a major chunk of movies is made in the 21st century. With a drag and mouse feature embedded within this visualization, user can set the range of years he wishes to see movies made in and accordingly the other visuals are interactively modified.

Further users can also see current shortlisted movies using data table below which contains drop-down option to set the number movies displayed in the same page. Table contains data such as country of make, genre, Duration etc., There is also search option
included within this to aid the user to reach a certain movie in his/her mind and get the details accordingly. The user can also sort the list as per his/her preferences by clicking on the variable name.


Fig. 6: List of movies meeting the user preferences

## 2.3 Additional Views:

The Visualization system also offers additional views not connected to the central dashboard but in separate tabs. This separated view helps users process information more efficiently without getting overwhelmed.

These visuals can be accessed using hidden sidebar:

Side bar has three sections which are separate views serving different kind of purpose for the use. One is the main dashboard, second is the Network Visualization and other is bubble chart consisting of ROI component of movies.

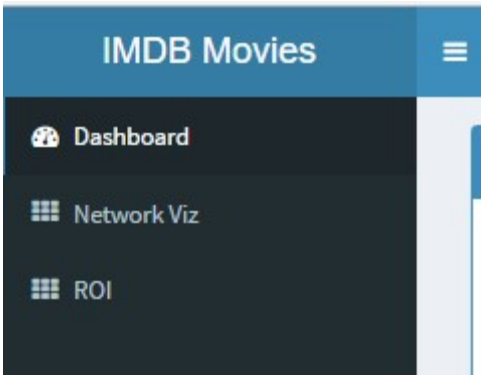


Fig. 7: Sidebar for Navigation

### 2.3.1 Actors Network Visualization:

The Network Visualization (Network Viz) offers elegant display of associations between various actors. This visualization is highly processor-intensive and scans the dataset multiple times to find "Between-ness" among various actors by analyzing the common movies among them. Such exploration can easily be done using visual dashboard and it is very hard for user to understand such relations using traditional methods of data exploration.

Network Visualization might be extremely helpful for people who are looking for a movie starring two famous actors in the same movie. User can get information like which all the other movies a certain actor has starred in. We can know if the actors have closed network and also check if the success of an actor is related to his network.

From a network graph perspective, it is challenging to decide a center of the graph based on actors with their linking to films they played in. Here, we are using a graph metric of 'betweenness centrality'. Betweenness is a centrality measure that considers the importance of connecting disparate groups. It gives a high "betweenness score" to films or actors which are on multiple short paths between pairs of other films or actors. Seen on this way, a film in the center is "often" on the shortest path from pairs of random other films or actors. Films and actors are the nodes of the network which is called a bipartite graph (an edge always links a film with an actor).

The network graph represents the center of all the rated films (With chosen category 'Good', 'Average' or 'Bad') in the IMDB database. It plots the the 50 most central nodes with their direct connections.

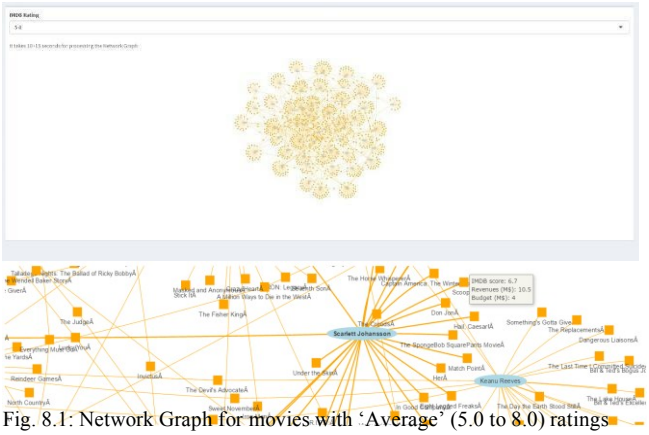

Fig. 8.1: Network Graph for movies with 'Average' (5.0 to 8.0) ratings

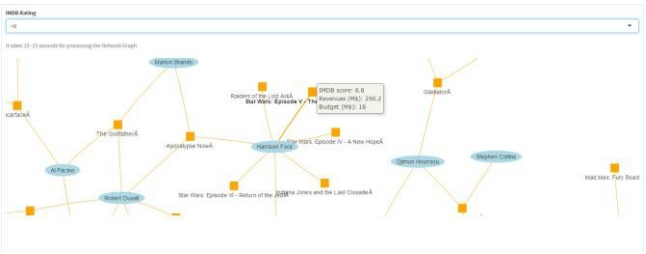

Fig. 8.2: Network Graph for movies with 'Good'' (Greater than 8.0) ratings

In addition to the powerful explorative powers this visualization offers, users also have an option to explore such connections between actors for "Good" movies (Having IMDB ration >8), "Average Movies" (Having IMDB rating between 5 and 8), and "Bad Movies" (Having IMDB rating below 5).

This is achieved by using a drop-down selection option embedded within this network visualization.

Fig. 9: Network Graph: Drop down range selection

### 2.3.2 Exploring ROI of Movies using bubble chart

With the help of bubble chart between Gross and budget which was done using "plotly" in R, users can explore Return on Investment (ROI) for various movies by using highly flexible and interactive bubble chart.

This Chart consists of a drop-down element which has the options to view the bubble chart for all movies and a specific genre by selection.

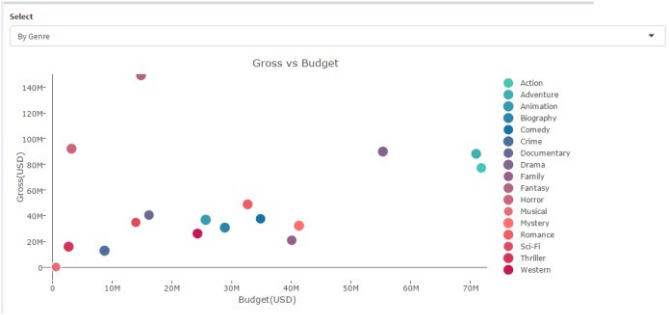


Fig. 10: Bubble chart showing results by movie genre

This chart also shows genre wise movies with an assorted color for each genre. Users have option to drill down using interactive options:


Fig. 11: Bubble chart : Zoomed in view

Color distribution can be seen on the bubble chart for different kind of genres across. Size of the bubble is indicating the IMDB rating. As obvious as it looks, movies with genres such as action, adventure, animation tend to have more budget and at the same yielding high gross as well. This is opposed to Thriller and horror movies which have less budget and less gross too.

So, which class of movies have least budget but yield high gross on an average? These kinds of questions can be answered by narrowing down the bubble chart using the various drill down options provided at the bottom of the bubble chart. So, the assessment measure like slope of the line joining the origin and the point would give a good indication of Return of Investment (R.O.I) measure. A movie with higher slope would yield high gross with relatively less budget.

Based on different view operations in this visualisation, it can be easily interpreted that on an average, Horror movies have relatively low budget and a high gross compared to Action movies which have high budget but moderate returns with respect to the budget used.

Mouse-over Operation: When the user hovers over his/her mouse pointer on any bubble point, then the information like Movie name, Genre, IMDB rating and exact figures of Budget and Gross for that movie can be seen.

## 3 Similar Work

### Live plasma: a music and movie discovery engine[3]

Live plasma is a discovery search engine, with colored lines and nodes representing the connections between artists or movies. Home page of this search engine contains a single search box for users to enter the name of either a musician, a movie or an actor. It results in giving a web of relationships based on data pulled from an API.

### Relevance to our Visual System

Relevance to our Visual System: Impressed by the lines and node design in this search engine, we decided to create a similar kind of visual to represent how different actors are connected to the films they played in.

## 4 Team Member Contributions

Bharadwaj worked on POC (Proof of Concept) for R platform with Shiny and DCR library as our technology to work with. He explored and checked feasibility of various other JavaScript libraries such as D3, DC etc. These libraries offer more flexibility and control over the graphs and the workings set. However, the code is hard to debug and maintain. R platform with Shiny and DCR library offers a tradeoff between control and flexibility over high level of abstraction and code maintainability.

Anvesh and Varun explored the dataset and performed some initial analysis necessary form brainstorming an abstract design of the dashboard. Vasanth contributed to some important ideas about visualizations and how they should interact.

Bharadwaj and Anvesh worked on translating the ideas from the research papers and articles into system design, Setting up the development platform, translation of detailed design into code and necessary documentation.

## 5 Conclusion

"A picture is worth a thousand words" is a famous English idiom which refers to the notion that complex ideas can be conveyed with a

single still image. But an interactive image as in our present cases might convey more than that with more nuances of information while offering control to the user.

Many of the exploratory options incorporated in these visual dashboards and views can offer the user a control and flexibility through seamless visual representations:

Central Dashboard helps user to explore and shortlist movies using user friendly interactive visualisations such as filtering on pie chart containing country of movie making, donut chart consisting of movies based on duration of preference.

Two separate bar charts which show the genre wise count and rating based count of movies would change according to the user filtering in the earlier charts.

Based on the user's preferences, these interactive views within the same dashboard would ease the job of a user to explore the preferences and arrive at a shortlisted movie set without having to scroll through multiple pages of information.

Network visualisation tab conveys information about the actor network frequently sharing the screen space. It helps in exploring associations between two actors, relation between actors with closed network and their success because of the network.

With the help of bubble chart visualisation present in a different view tab, users can explore the various movies by genre and access the information about budget, gross and ROI of various kinds of movies within the same chart.

Visualisations like these can let the user draw some notable insights from movie data which would not have been possible with traditional search and find methods of exploring information.

## REFERENCES


[1] S. Chuan, IMDB 5000 Movie dataset, kaggle, https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset. Accessed on May 2017.

[2] R. Gupta, N Garg, A. Das, A novel method to measure the reliability of the 111 bollywooed movie rating system, Feb 2013

[3] J. Harris, Liveplasma music and movie discovery web app – review, http://jasonharris.ca/2011/10/liveplasma-music-and-movie-discoveryweb-app-review . Accessed on May 2017

[4] Academy Award's best actor and actress winners and nominees from 2000-2004 downloaded from http://www.imdb.com/Sections/Awards. Accessed on May 2017.

[5] Davidson, G.S., Wylie, B.N. and Boyack, K.W. Cluster stability and the use of noise in interpretation of clustering. Proc. IEEE Information Visualization 2001.23-30.

[6] Internet Movie Database (IMDb) network provided for GD'05 at http://www.ul.ie/gd2005. Accessed on May 2017.

*[7]* Nooy, W.d., Mrvar, A. and Batagelj, V. *Exploratory* Social Network Analysis with Pajek. Cambridge University Press, 2005.

[8] Paley, W. Bradford. Illuminated Diagrams: Using Light and Print to Comparative Advantage. *InfoVis Conference* 2002.

[9] S. Maddineni, Visualizing Hollywood BoxOffice Revenue. NYC Data Science Conference 2016.

[10] J. Dauenhauer, J. Hockett, J. Mammarelli and M. Yarem, Information Analysis on Movie Genres. Published March 2014.