CS 6630: Visualizing Movies Metadata

Ram Kashyap S and Mohammed Musaddiq

Nov 10, 2017

1 Basic Info

Project Title: Visualizing Movies Metadata

Team member 1: Ram Kashyap S
Email: u1082810@utah.edu
uID: u1082810

Team member 2: Mohammed Musaddiq
Email: mohammed.musaddiq@utah.edu
uID: u1068996

Project repository:
`https://github.com/ramkashyap-s/dataviscourse-pr-movies-viz`

2 Overview and Motivation

The reason we chose this project is the relevance and appeal it would have to a public audience considering the millions of movie fans across the world. Furthermore, being movie buffs ourselves, we are also motivated by personal interest in visualizing different aspects of movies and sharing the same fellow fans.

3 Related Work

We explored many example visualizations online, assessed them according to our project scope and arrived at our final design. We are particularly inspired by this project `https://cips1.engineering.asu.edu/movie_analysis/`

The table and line charts that we implemented in our milestone are similar to the ones that we have seen in class.

4 Questions

The primary questions we are trying to answer with our visualization are:

- For a given actor/director, view or compare how the following parameters vary over time:
  - Rating of movies they have acted in/directed
  - Gross earnings of movies they have acted in/directed
  - Budget of movies they have acted in/directed
  - Number of movies they have acted in/directed per genre
  - Proportion comparison of movies they have acted in/directed across genres

Benefits: The above visualization would allow users to get valuable insights into how an actor's/director's movies have fared over time based on various aspects they are interested about and wish to compare. It would also tell them about what kind of movies the actor/director is usually involved in.

- For a set of filters such as movie rating, year(s) and genre, provide a list of movies matching that criteria. Then, for a particular movie in this list, provide the ability to visualize other meta-data such as Director, Actor, Language etc

Benefits: This visualization would help users in researching and finding all movies based on the genre(s) they like, how recent or old the movie is or how good/bad the movie's ratings are. Further, they can also view other potentially information about the movie they are interested in or are curious about.

5 Data

We have obtained the metadata for 5000+ movies spanning across 100 years in 66 countries from here:

`https://github.com/sundeepblue/movie_rating_prediction/blob/master/movie_metadata.csv`

The data contains 28 variables and close to 5000 movie records. There are 2399 unique director names and thousands of actors/actresses.

5.1 Data Processing

Our data source is from a csv file which has empty values in some columns for few records

We decided to handle data cleaning on the fly. We tried to delete the movie records which had empty/zero values in columns. But, this strategy proved to be unhelpful as there are non-empty columns which we could use in our visualizations.

- For table, we are checking for empty/missing string values. We are checking for null/empty 'NaN' values in plot.
- In the movie titles column, we have an extraneous character Â which is due to a different encoding. But, it isn't affecting the visualizations.

6 Exploratory Data Analysis

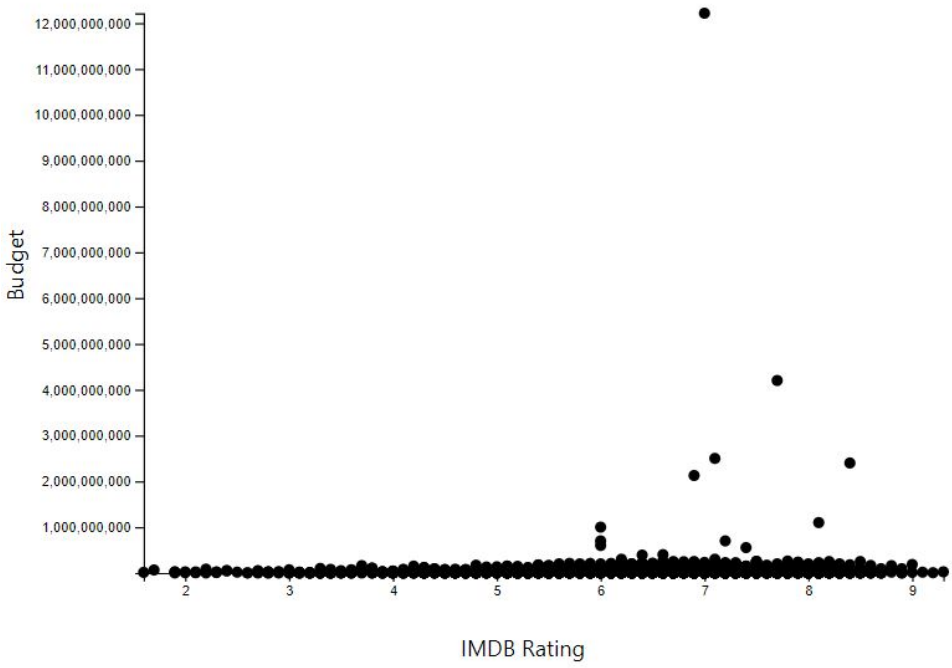

Figure 1: Initial plot - rating vs budget

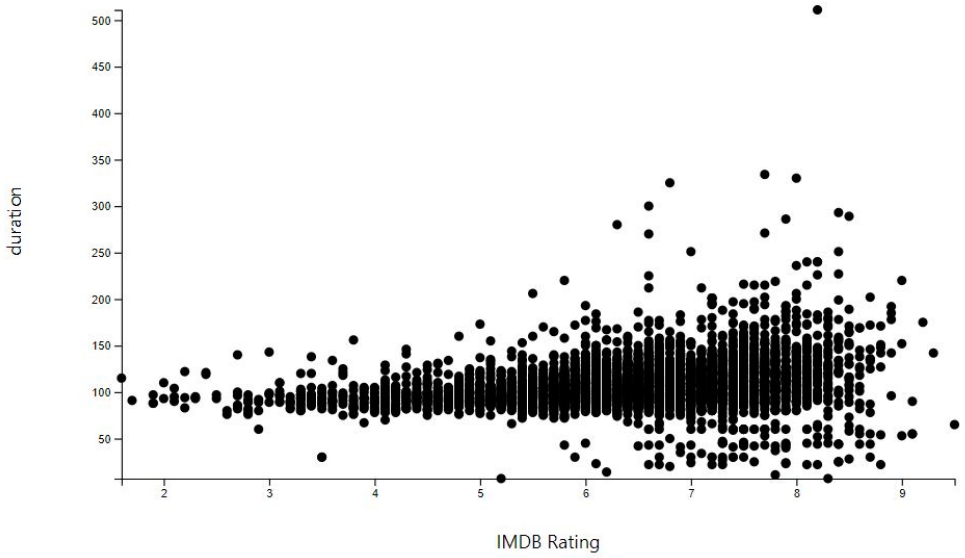

Figure 2: Initial plot - rating vs duration

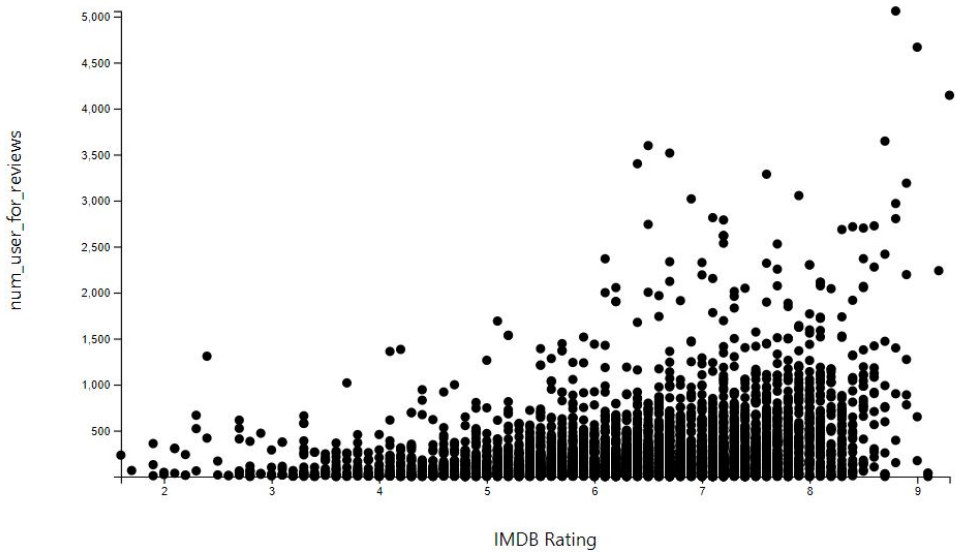

Figure 3: Initial plot - rating vs number of user reviews

We have created few scatter plots like the one shown above. We understood that rating has linear correlation with factors such as number of reviews, gross, number of movie facebook likes for movies released after 2008, etc.

From the plot Rating vs Budget as shown in Figure 1, we realized that we need to adjust our y-axis to eliminate the outliers from the plots to get meaningful insights. In the upcoming weeks, we will try to remove the outliers from the plots that we have identified to get meaningful insights.

7 Design Evolution

Initial Design

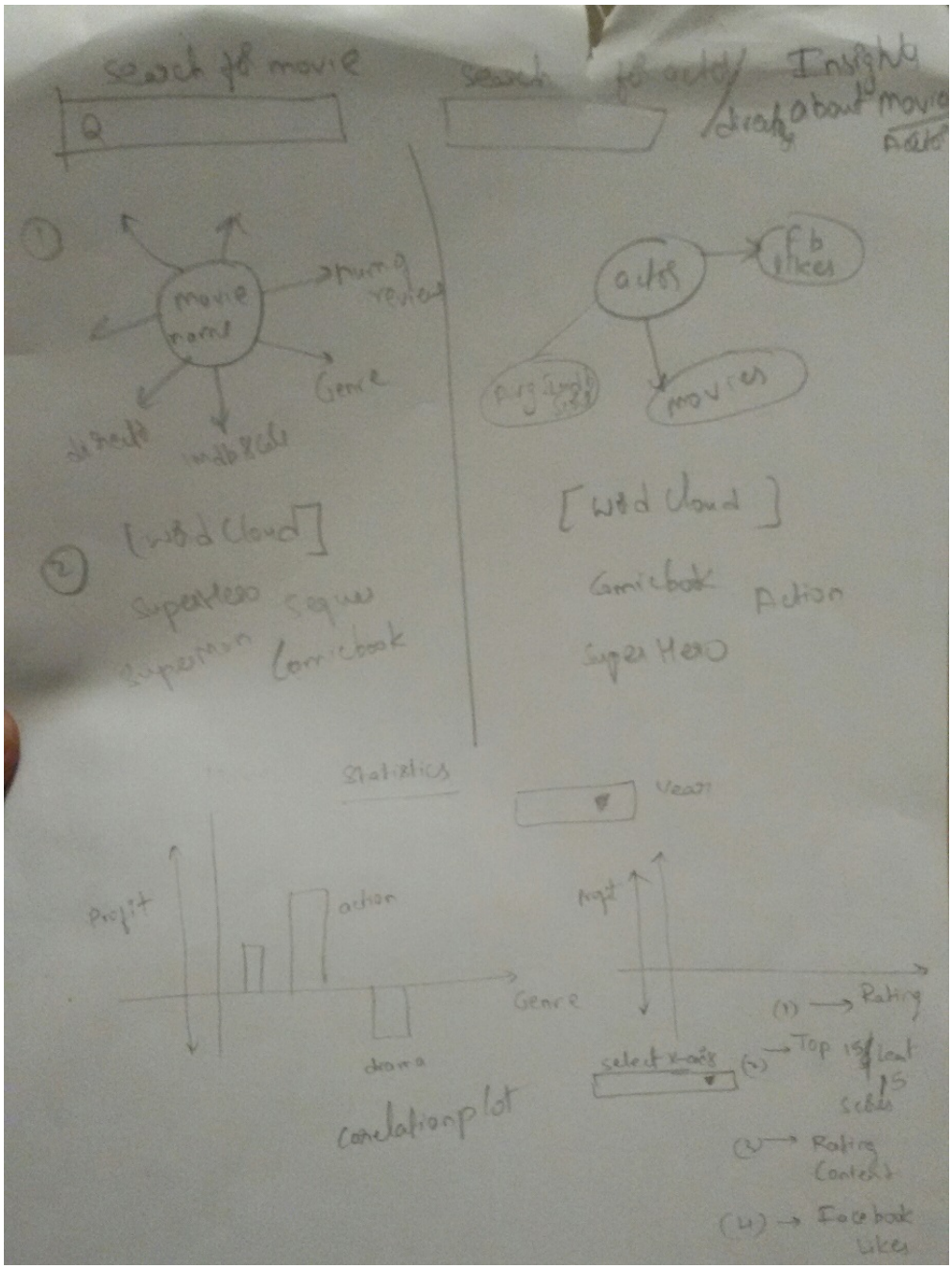

Figure 4: Initial Design

Implement a search filter that would allow the user to select a movie and an actor/director. Based on the selection, we will visualize that movie and actor/director as a graph with the movie and actor/director as central nodes and their associated metadata as child nodes attached to them.

We feel there might be a need to allow users to find/explore movies in other ways instead of restricting them to using a search filter that allows them to filter only by title.

We will also visualize the movie's plot keywords using a word cloud that would give the user some idea about what the movie is about.

Next, we will allow the user to select various movie attributes such as genre, rating using a drop-down list and visualize their correlation with movie profit using bar-chart.

Initial Design 2

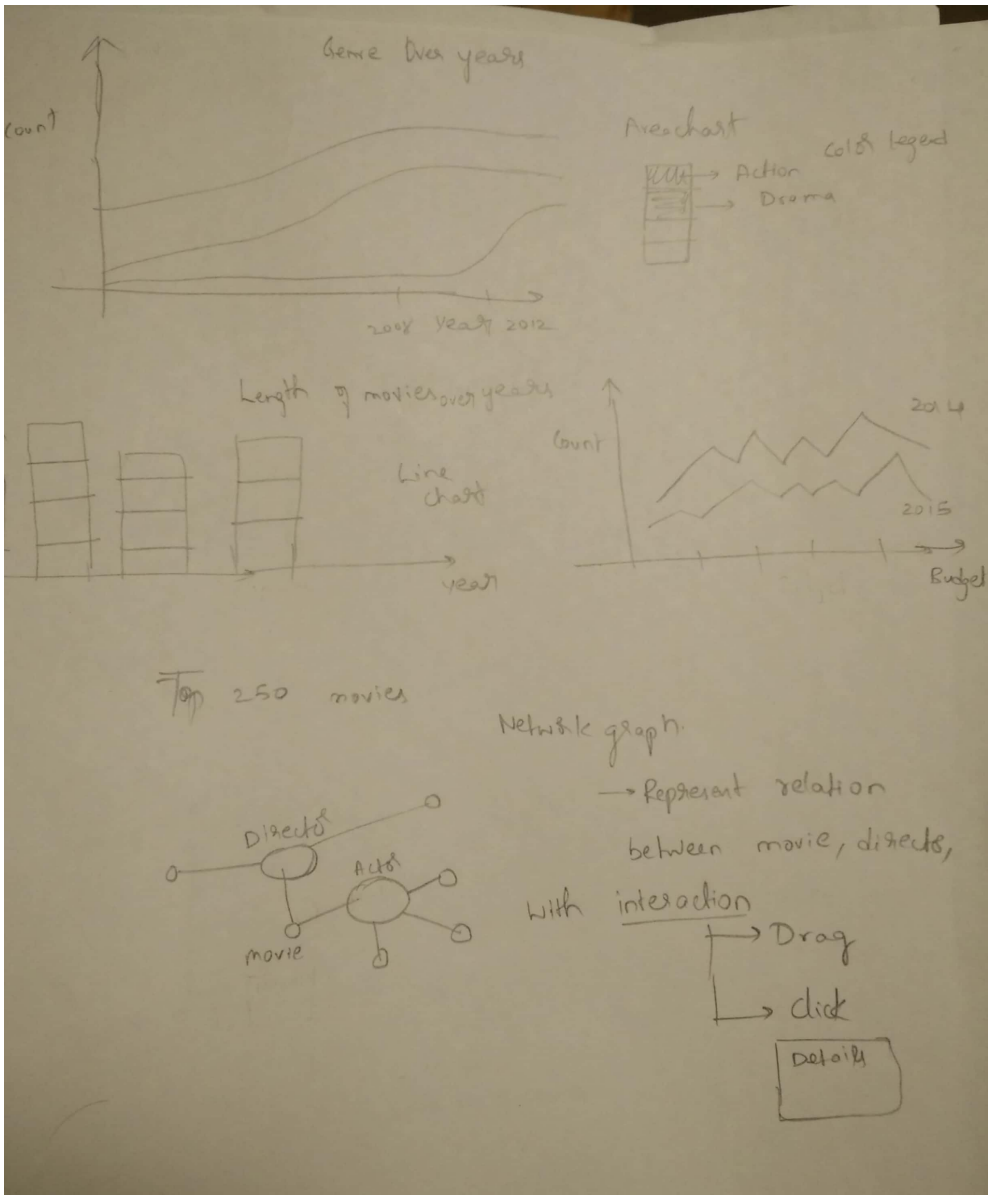

Figure 5: Initial Design 2

For top 250 movies implement a network graph which will represent relations between movie, director and actor. The size of the node for actor and director would be according to the number of connections(degree). The graph would be interactive with drag for visualizing the connections and click for visualizing the

details of node.

We feel that users might not get much information by looking at the connections in the graph.

Exploring data:

Implement area chart for genre trend over the years. Hue is used to encode different genres.

Implement a stacked bar chart for duration of the movies over the years.

Implement a line chart for budget over the years

Initial Design 3

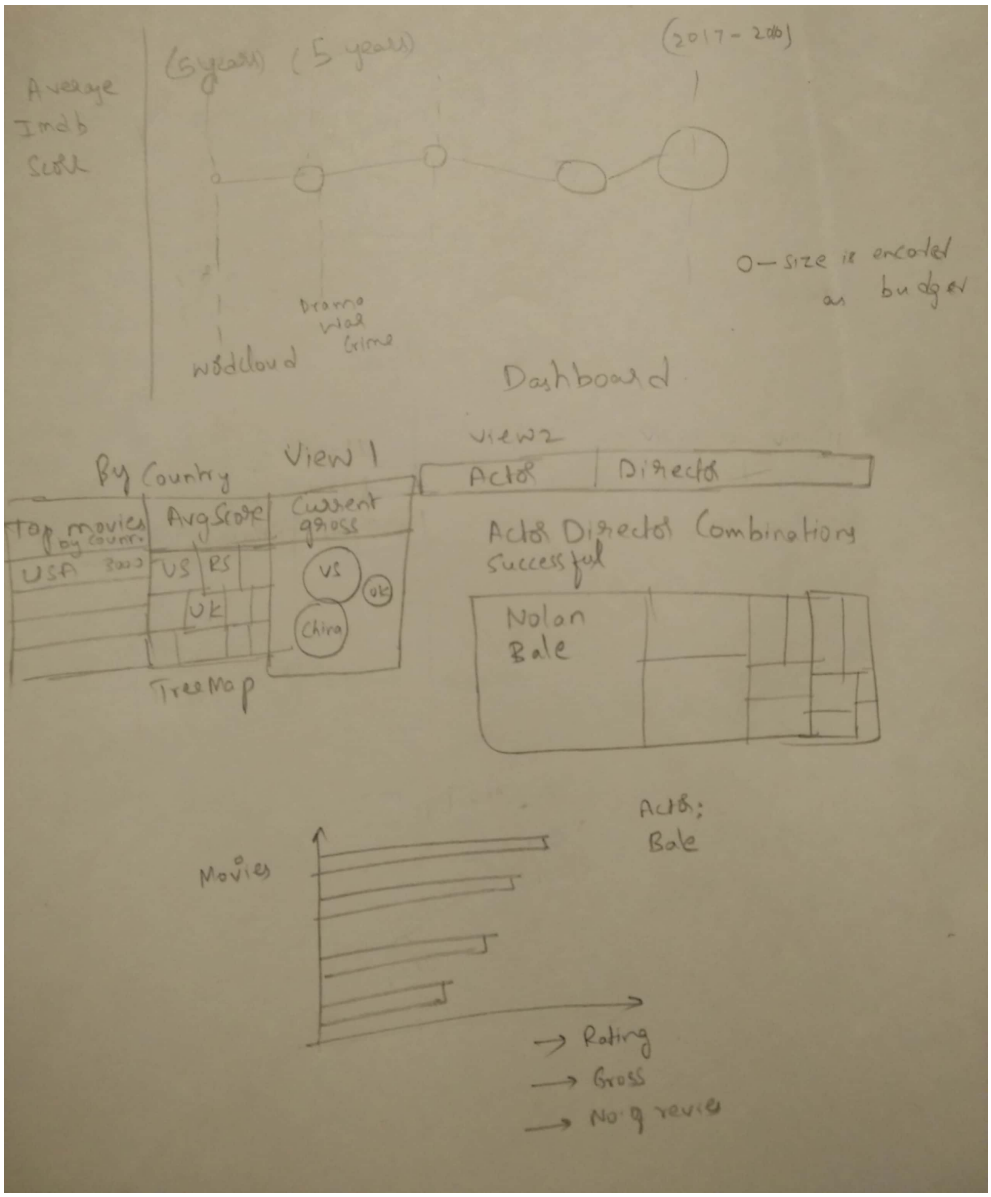

Figure 6: Initial Design 3

Implement a dashboard with different views
View by country: By each country display top twenty movies contributed
Implement a tree map with average scores
Implement a heat map of countries by selected year gross

View by actor and director:

Implement a tree map with successful actor and director combinations

Implement bar chart for top movies by rating, grossing, number of reviews, etc. for an actor or a director

Finalized Design

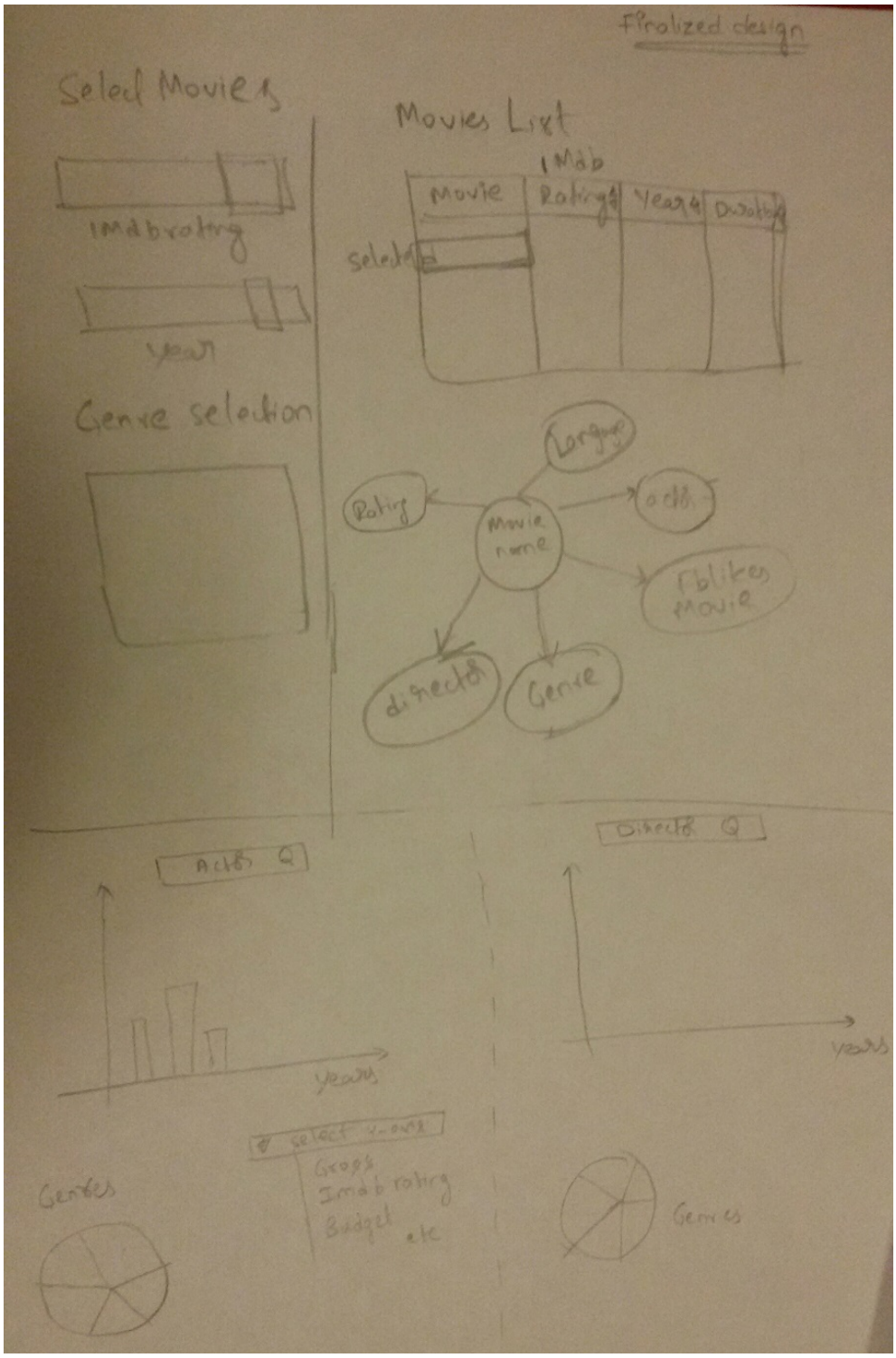

Figure 7: Finalized Design

We plan to implement a Movie Rating chart with a brush that would enable selection of a range of ratings that a user can specify as well as a brush-enabled Year chart that would allow the user to specify a range of years. We will also provide single/multiple genre selection using check-boxes.

We feel this is better than our initial design since earlier, the users did not have the ability to apply certain obvious filters to their search in case they did not know a movie name to begin with and just wanted to explore.

Then, we would visualize all movies matching the above criteria specified by the user using a dynamic table.

Finally, the user can select a movie from the above list he wants to know more about. We will visualize that movie as a graph with the movie as a central node and its associated metadata as child nodes attached to it.

Next, for the Actor/Director visualization, we plan to implement a search filter that would allow the user to search for and select an Actor/Director of interest. Based on the selection, we will provide a bar-chart based visualization of how different attributes of their movies have changed over the years. These attributes can be chosen using a drop-down list.

7.1 Changes from project proposal

Based on the discussions with professor, we decided not to include pie chart for genres as a movie can have more than one genre and it would be complicated to inference from such a chart. Instead, we will try to implement a bar chart for genres.

Also, we have realized that scatter plot is much suited for some plots, than a bar chart. So, we are implementing scatter plots, line charts or bar charts according to the data that we want to represent.

8 Implementation

8.1 Filters

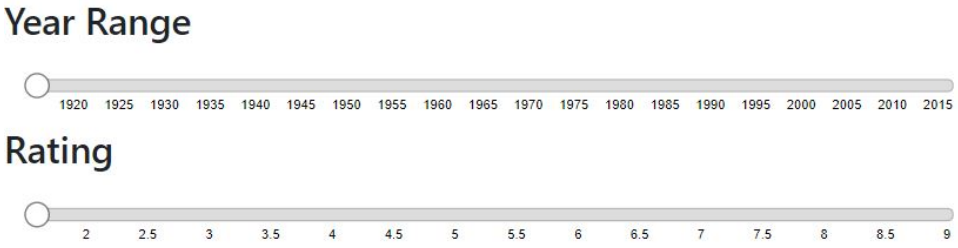

Figure 8: Selectors

For this milestone, we have implemented selectors for year and rating. Slider would give the values selected from the right up to the handler position.

Also, we have populated the check boxes required for selecting the genres.

Figure 9: Selectors

The filters are not connected with table or other views. The filter selection would later be used to populate the movie data in the other views.

8.2 Table

The intent behind the table is to list all the movies (along with attributes such as IMDB rating, Budget and Gross) that match the filter criteria specified by the user. As for the interactive functionality, we have enabled sorting on each column that would allow the user to easily sort the entries in ascending/descending order. At the moment, we have not actually linked the table to the filter criteria but plan to do so in the next phase.

Movies matching the specified criteria:

Movie Title	IMDB Score	Budget	Gross
Avatar	7.9	237000000	760505847
Pirates of the Caribbean: At World's End	7.1	300000000	309404152
Spectre	6.8	245000000	200074175
The Dark Knight Rises	8.5	250000000	448130642
John Carter	6.6	263700000	73058679
Spider-Man 3	6.2	258000000	336530303
Tangled	7.8	260000000	200807262
Avengers: Age of Ultron	7.5	250000000	458991599
Harry Potter and the Half-Blood Prince	7.5	250000000	301956980
Batman v Superman: Dawn of Justice	6.9	250000000	330249062

Figure 10: Table

8.3 Line chart

The intent behind the line chart is to analyze how a particular actor's or director's movies have performed over the years using their respective IMDB ratings. As for the interactive functionality, we plan to implement a search filter or a drop down that would allow selection of actor and link the same to the link chart to update dynamically. Also, we would have a tool-tip to show few details about the movie.

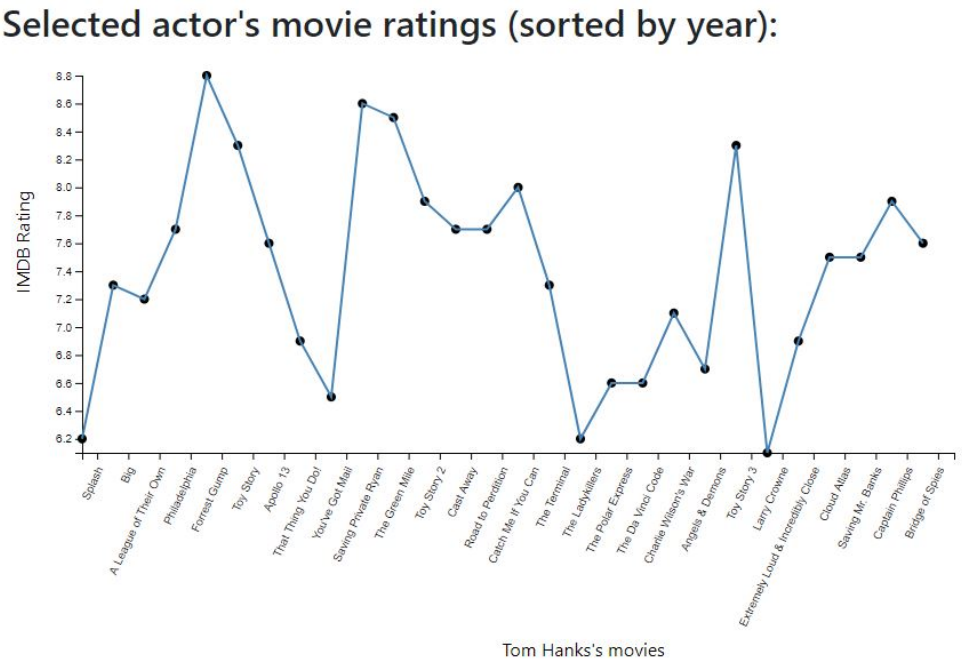

Figure 11: Actor/Director plot

9 Evaluation

9.1 Sliders

We have realized that we would need second slider handle for selecting movies in a range. So, we would implement a second handler in the upcoming weeks.

9.2 Table

We have thought about the scenario where the number of movies matching the user's filter criteria could be too high to visualize efficiently. The gross is colored red if the gross is less than budget or blue otherwise. So we plan to further improve this table by implementing a fixed-header scroll feature that would allow the user to scroll through any number of resultant entries without negatively impacting the screen space.

9.3 Line chart

This feature could be further enhanced by including selection of other attributes such as budget/gross to be analyzed for the selected actor.