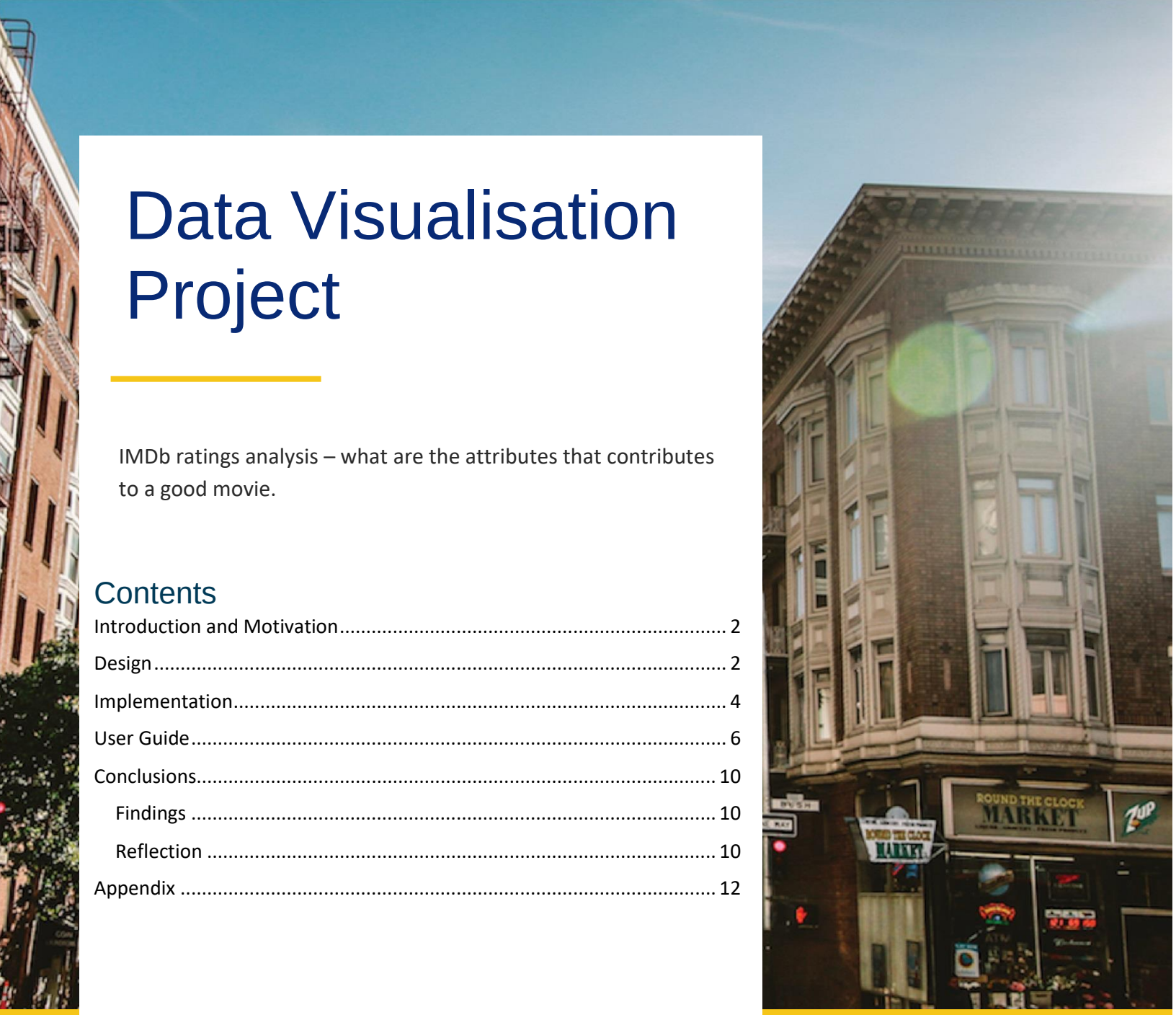


Data Visualisation Project

IMDb ratings analysis – what are the attributes that contributes to a good movie.

Contents

NOVEMBER 1

By: Xiaoyu Tian 28540964
Tutor: Angel Das and Mohit Gupta

Introduction and Motivation

This report is to introduce and present the narrative visualizations based on the exploration of the IMDb dataset. As I am a film lover, I am always interested in exploring every aspect of a good movie. Recently, I have browsed the IMDb top rated 250 movies. Although it's not surprising that those famous movies are on that list, I am still wondering what kind of topic or attributes of a movie that the viewers mostly like. Also, I'm very curious that if there are any correlations between those attributes. Therefore, the aim of the narrative visualizations is to convey information about the relationship between different attributes of a movie with the IMDb rating, as well as how those attributes correlate with each other, and let the users explore the movies freely. By using the five design sheets method, I have come up with different ideas and designs to explore these questions. The final design is realized by using the D3 version 4 and it is based on the five design sheets which will be presented in the appendix for the reference, and the details of the design and implementation of the visualizations will be discussed in this report. Therefore, I hereby present my visualizations to the target users who are also movie lovers and reviewers who wish to explore more about the movies.

Design

The five design sheet will be first briefly described, and they can be found in the appendix of this report.

Sheet 1:
This sheet displays the brainstorm process. After filtering out some ideas due to their irrelevance or incompatible, for example stacked area plot is not good for comparison and the shape of the areas are misleading, the rest ideas are categorized into 5 groups, bar chart, proportional, spatial, distribution and line graph. Useful ideas are further combined and refined into our following different designs.

Sheet 2:
The first design contains two stacked bar charts with lines showing the trend, and bubble chart showing different genres and their corresponding average rating. The first bar chart will display the number of moves over year and there is a slider bar below it to select the year range. The second bar chart shows the number of movies over different IMDb scores. Also, the line graph is designed to be shown within the bar chart, representing the trend for each genre. However, the line graph of this design will be very messy and distract the users focus on the bar chart.

Sheet 3:
This design mainly focuses on the scatterplot in the center of the page, which brings more focus to the users. On the left there is a filter section which contains different filtering tools that help the users to manipulate the data which are shown in the plot. The main reason for these filter items is to reduce

the complexity of the data. Also, it is natural to visually encode data that changes over time, hence when the users open the page, by default they will see the scatterplot of IMDb votes over a time series to get a basic idea how the movies are distributed throughout the history. Furthermore, the selection bar below the scatterplot enables the users to manipulate the variables display in the scatterplot freely, for the users to see different relationships between each variable. Also, there is a button above the plot. By ticking the button, the scatterplot will only show the movies with awards. The pie chart will give the users an overall view on how the movie genres are composite of, which somehow represents the preference of people's movie type.

Sheet 4:
The third design combine the tabular dataset and spatial dataset, where it includes a choropleth map showing the world movies distribution, a boxplot showing statistical distribution of ratings for each genre and a dot chart where the size representing the number of votes. The including of a world choropleth map is the main part of this design, as when human brain processes visual information, the upper sections is mainly concerned with the spatial location. Hence, giving the interactive choropleth map will give the users a very clear overview of the distributions of the movies in the world. As another half of the brain is devoted to analysing the visual features, such as colour, size and shape (Knauff, 2013). By combining the statistical part under the map with boxplot and dot chart, it gives the users information in different dimensions. However, the main issue with this design is with the narrative part. The users could not easily link how the geometry might affect the IMDb ratings, as well as how each attribute correlates with each other. Also, the three different charts increase the complexity of the design, which might distract the focus of the users.

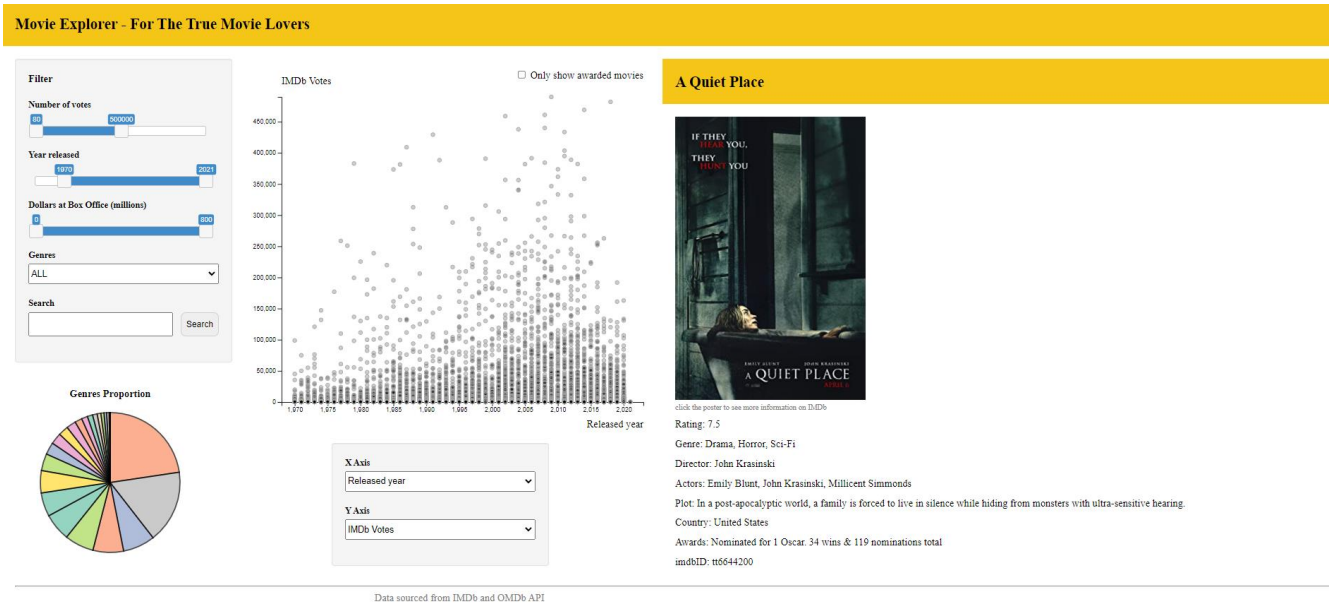
Sheet 5:
According to the previous discussion, sheet 3 is selected as the final realisation. One major reason to choose this design is that it gives the users a significant freedom to interact with the map. The users can apply filter items and select any variables that they wish to see the relationships, and it encourages the users to explore, which is the main aim of this project. The pie chart might not be the best chart for human eye consumption, but it works well here as the point of it is to help the users the gain a very brief ideas how each genre contributes to all the movies and let the users focus on the main part which is the scatterplot.

focus+context views in which data in the view is shown in more detail around the focal point on the view and in less detail elsewhere in the view, so as to provide context. Probably the best known focus+context view is the fisheye lens, in which the view is distorted so as to achieve an effect similar to the fisheye lens in photography. Other types of focus+context views remove detail for objects away from the focus or provide detail on a separate layer, for instance using the metaphor of a magnifying lens.

Implementation

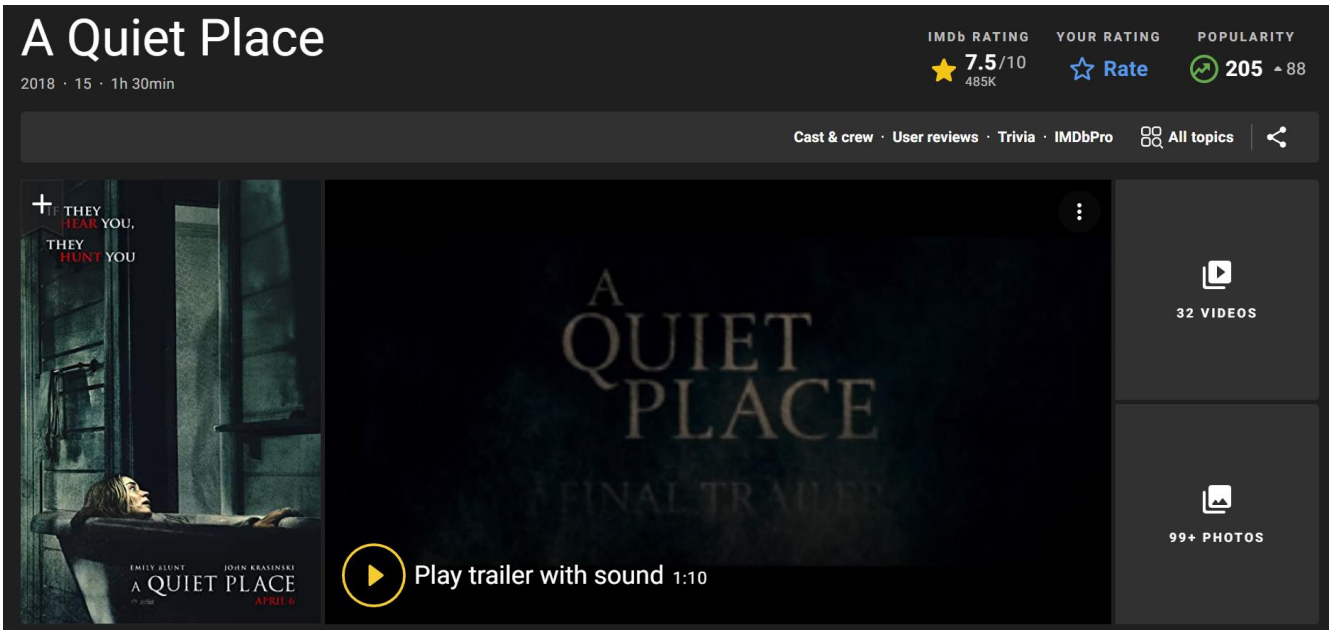
The final design implementation is based on D3 version 4. The idea of this design is inspired by the work of “mine-cetinkaya-rundel”. His visualization is based on Shiny and the the github repository link is as follows: https://github.com/rstudio/shiny-examples/tree/master/051-movie-explorer. Only the ideas are used as a reference here and no template code is used for my project as its under D3 instead of Shiny.

The first challenging part is to obtain the data. Although the IMDb provides free download for the dataset, it only contains some basic attributes of a movie, such as title name, number of votes and ratings. The box office data, movie award information as well as other information is not provided. Although “mine-cetinkaya-rundel” provided data for the reproducible work, the IMDb data provided are too old to use, but luckily the source to the data is provided which is the OMDb API. However, unlike normal website downloads, the OMDb API must obtain a key and use this key in javascript code directly connect to and download from the API. As new to javascript, I have to learn how to download the data first and combine the data with IMDb datasets. I learnt from “Coding Shiksha”’s youtube channel, and the link is provided as follows: https://www.youtube.com/watch?v=8iuPNq553U0. Furthermore, I’m also inspired by this tutorial of the ideas including a movie poster to design, which will be detailly discuss below. Another challenge is the extensive data wrangling due to complexity and large observations. There were more 550 thousand of observations and the data is messy, and I have to change the format, clear the data and filter out unwanted data, which eventually returns 66732 observations. The wrangling process is crucial, as the page will take too long to load if using such a large dataset. As previously mentioned, the joining process in javascript between a csv file from IMDb with the OMDb API data is complex. The file for the wrangling process is provided in the zip during submission.



The final completed version is as above. Most functions in the design are implemented here. One thing to mention is that due to the dynamic properties of the filter bar is required in the design, the jQuery is used for the slider bar in the filter section.

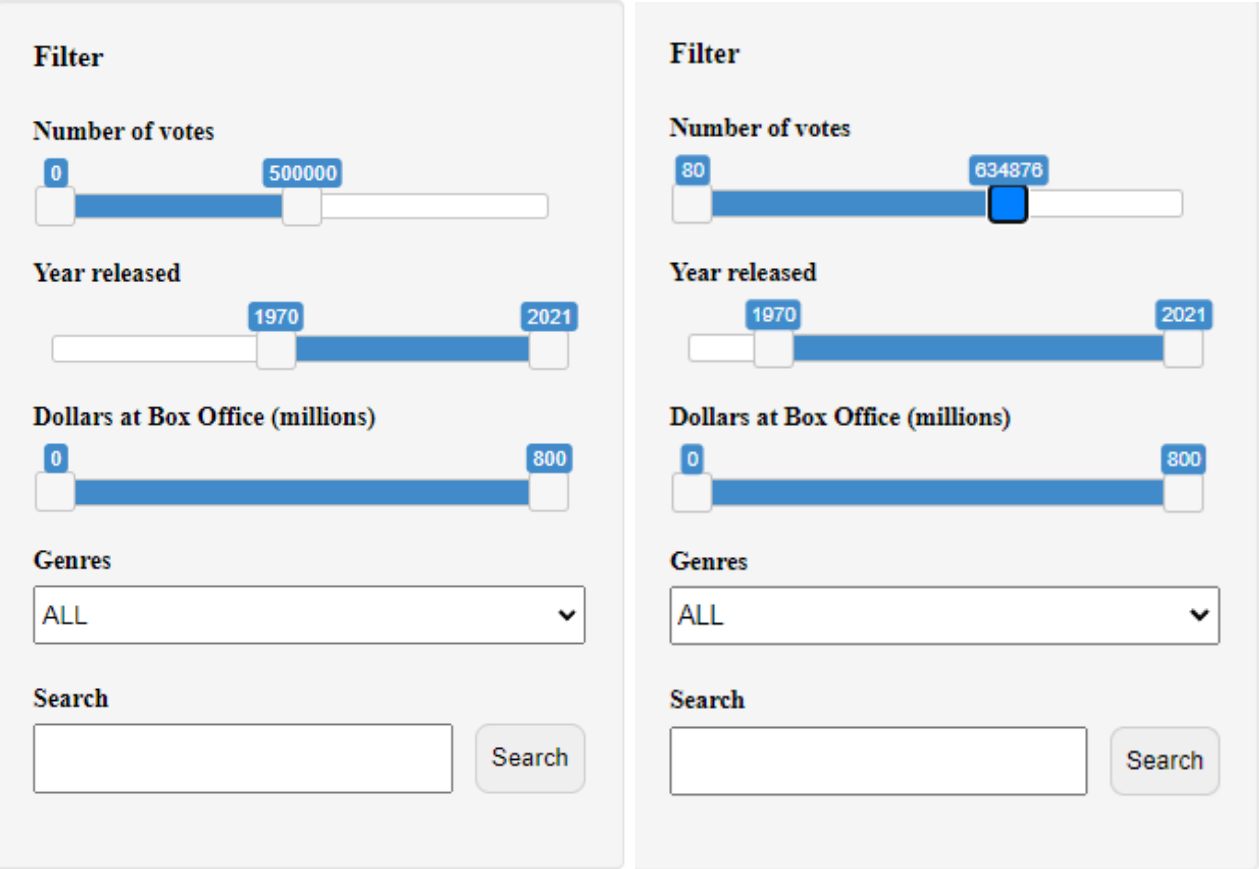
Also, there are several differences in the layout with the five-design sheet. First, we can see a highlighted title is added “Movie Explorer – For the true movie lovers”, this should immediately tell what this page is used for and who are the target users. The highlighting color used is the IMDb color, which could also remind the users what this page is about. As previously mentioned, inspired by “Coding Shiksha”’s youtube tutorial, I have added another section to the right of the scatterplot, which contains all the details of a selected movie. This is better than original design where all those information are put into the tooltip when hover over the point, and the poster is nicely shown in this section. In this way, the users can easily obtain all the information of the selected movie. By seeing the poster, it also gives the users a direct image of the movie. Also, it encourages the users to explore the movies by clicking on the scatterplot, because it’s fun to see different movie posters and their detail information. Furthermore, the links to the IMDb page are provided as well, by simply clicking on the poster, this will lead the users to the IMDb page for them to explore more about the movie that arouse their interest.



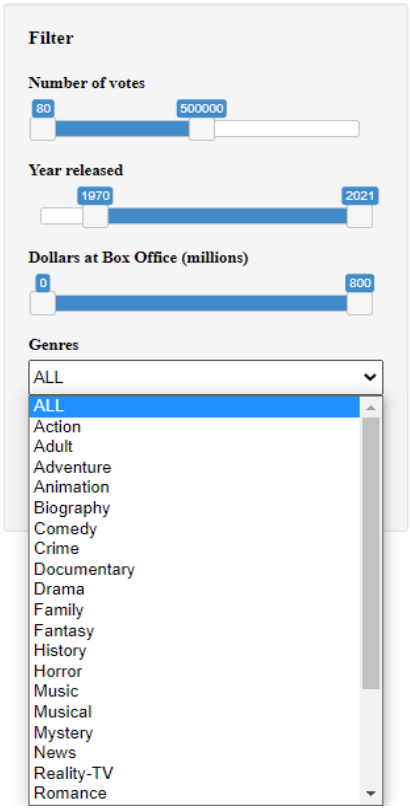
According to the design, when select a certain genre of movies, the pie chart should highlight the selected genre and its proportion of the total data. Also, there should be a button above the scatterplot for the users to perform logarithms transformation to the axis. However, due time limits, these functions are not implemented in this project.

User Guide

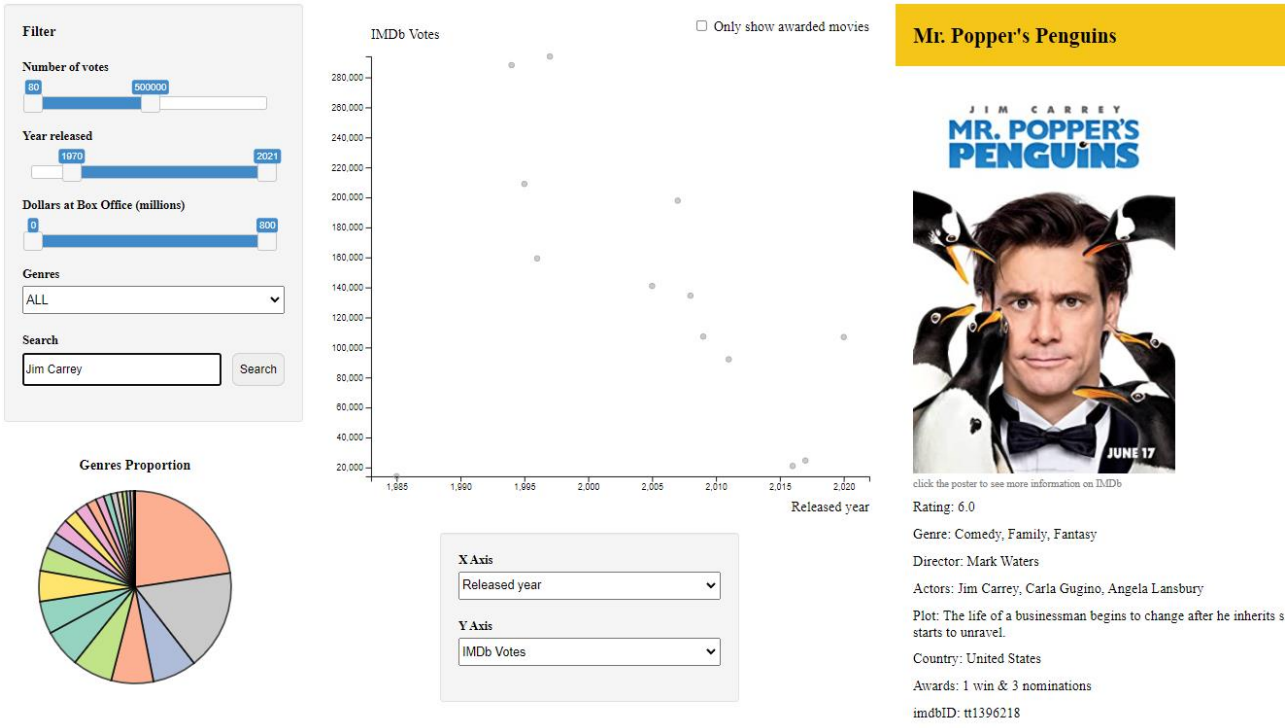
This section will explain all the interactive features of the narrative visualizations. First, the data is sourced from IMDb and OMDb API. However, some movie data are not provided for free download, hence not all the movies in history is able to be reflected in this visualization. In our visualizations, 66732 observations are included. Please note that due to the large dataset, it would take around 10 seconds to load the page. Please wait patiently while loading the visualizations. Details of the interactive features are as follows:



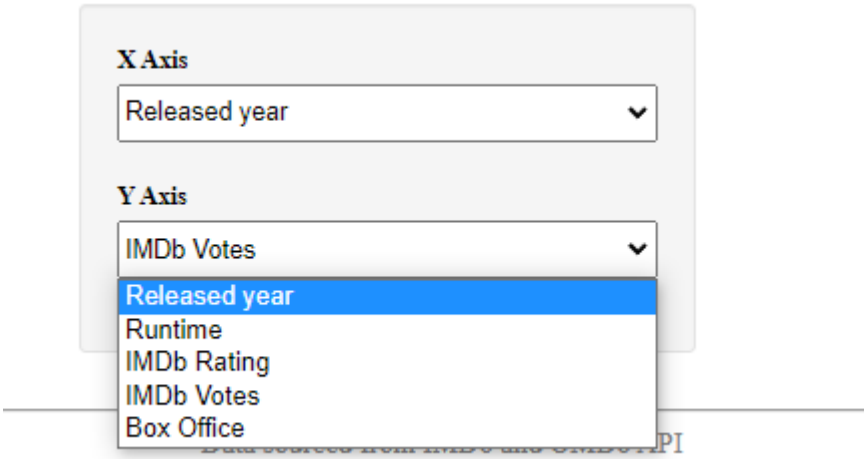
The filter bar allows the users to play around and make the selection of the data to be presented freely, and the selection data will be presented in the scatterplot autonomously. The users can click and slide the slider bars to make the selection of the range among variables which include the number of votes, the year released and the box office revenue for the movies. When click and hold the slider bar button, the button will be highlighted in blue and that tells the users they are making the movement of the slider.



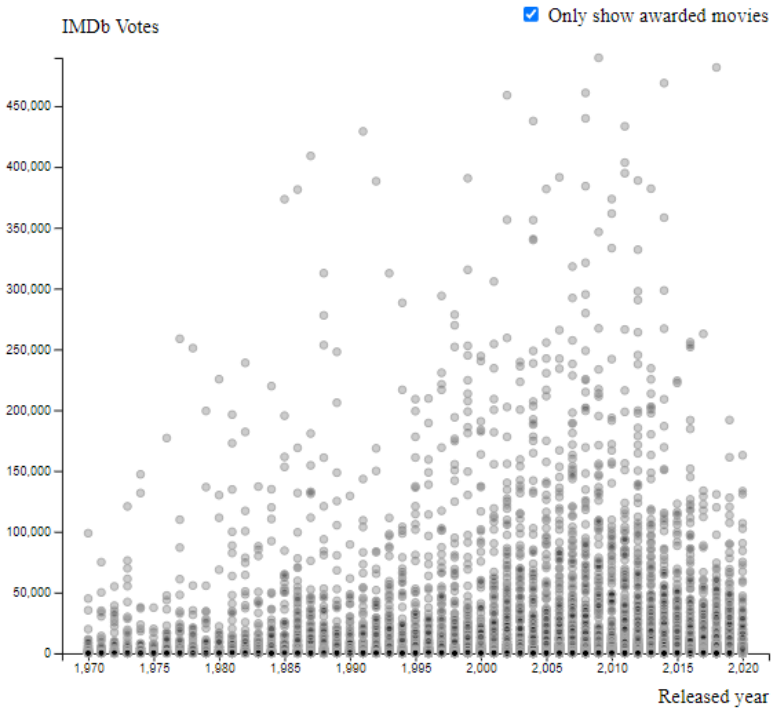
Below the slider bar, there is a selection bar, where the users can make the selection of a certain genres, for example, drama. Then the scatterplot will only show the movies that falls within that genre.



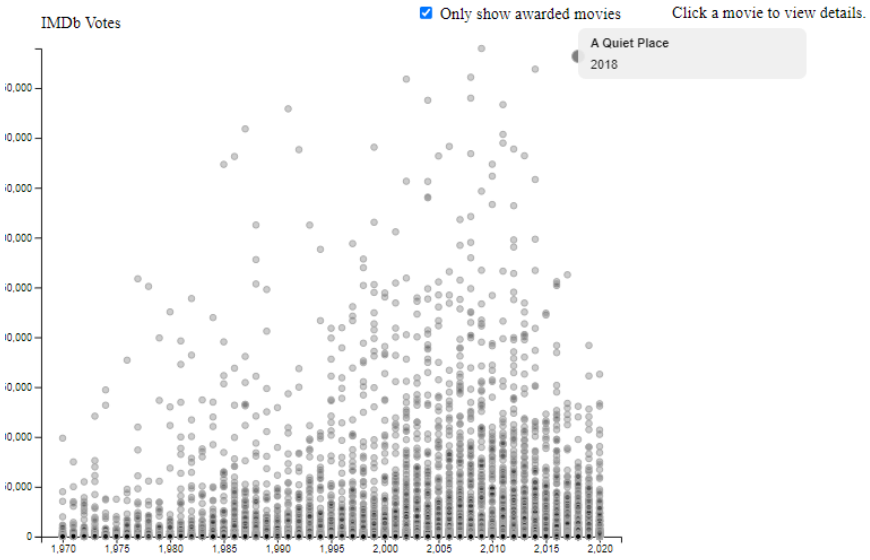
Following by the genre selection bar, a search bar is provided for the users to search for movies based on the movie title names, main actor/actress names and director names. For example, if we put in the actor's name "Jim Carrey", the movies he involved will be presented in the scatterplot as above. Again, as previously discussed, the data do not include all the movies in history, and hence here the movies in the scatterplot does not represent all the movies that the certain actors or directors are involved with.



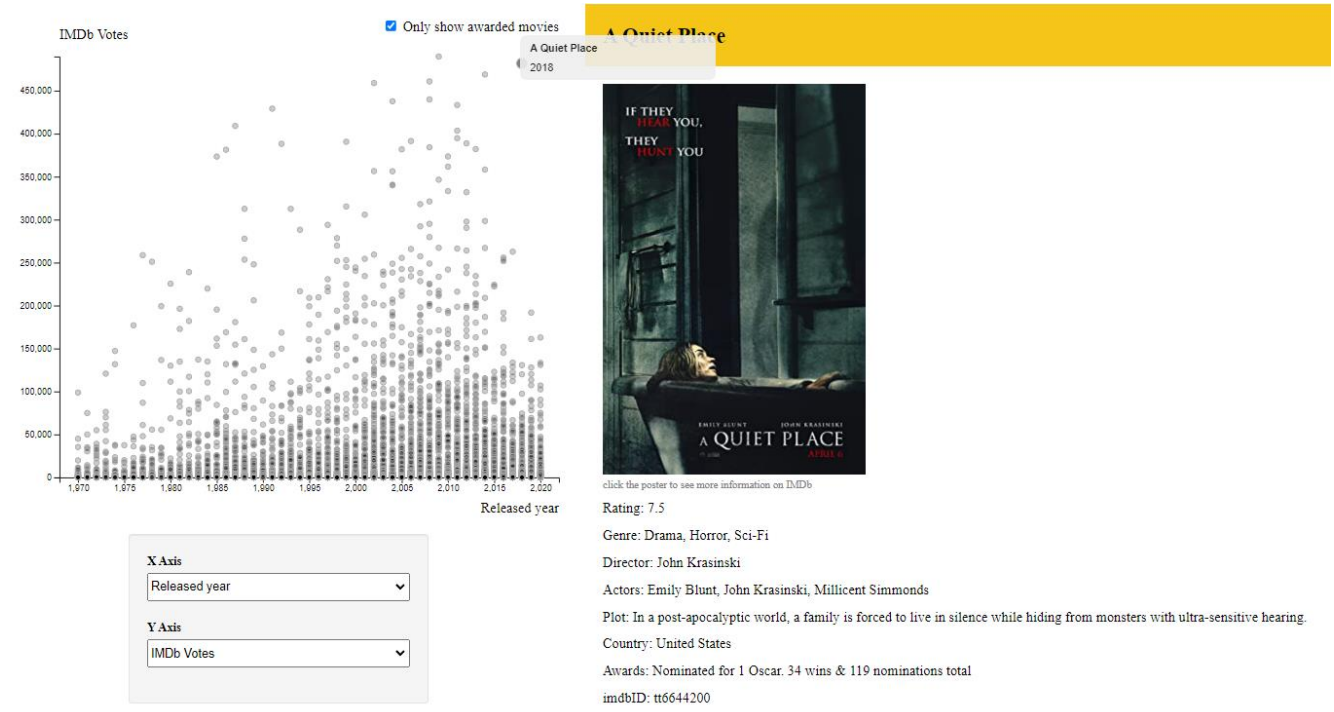
Under the scatterplot, the users need to select the variables that they wish to see the relationships of for the x and y axis among the selections of released year, runtime, rating, number of votes and box office data.



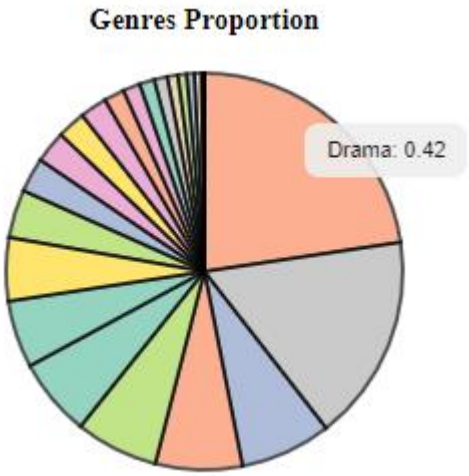
There is a button above the plot which is only showing the movies with awards. The awards include any kind of awards such as Oscar or nominations.



When the mouse hover over a point in the scatter plot, tooltip is shown as above which includes the movie title and the year the movie is released. When the html page is opened, we can see that nothing appears next to the plot. There is a message “Click a movie to view details”. This guides the users to click on the point in the plot to see more details.



After clicking the point in the plot, the poster of the movie as well as all the movie details are shown to the right of the plot. Also, we can see a message now appears below the poster that leads the users to click on the poster, and this will lead the users to the IMDb page of that movie.



When the mouse hover over the pie chart, the proportion of a genre will be displayed as above.

Conclusions

Findings

In general, the average rating of IMDb movies is not increasing throughout the years, but we see the variation to the average rating become larger in the more recent years. No significant trend to see the rating is associated with runtime. However, it is interesting to see that the number of votes has a similar distribution with box office over the IMDb rating. If we put y axis as IMDb rating, while putting number of votes or Box office data in x axis, we can see that the data is heteroscedastic and has barriers. The first barrier appears around rating 9-10, while another barrier in high-voting-low-rating area. One possible explanation is, the more popular the movies are (more voting), the more likely the movie to achieve a reasonably high rating score due to their popularity. Furthermore, it is interesting to see that there is positive association between IMDb votes and Box office. The plots of the findings will be displayed in the appendix.

Again, there are limitations in the dataset, where not all the movies are included in this project. But still, this narrative visualization encourages the users to explore the relationship between attributes, so as to explore the movie itself freely.

Reflection

The data obtaining process and wrangling process was cumbersome, but I'm very appreciated that this project allows me to learn more about the Javascript and D3. Also, the experience in producing interactive visualizations is beneficial. It is unfortunate that due to time limits, some functions are not able to be implemented. Still some improvement can be done in the future, for example, the use of colour palletes can be improved.

Bibliography

Pretorius, A. Johannes, Helen C. Purchase, and John T. Stasko. Tasks for multivariate network analysis. In *Multivariate Network Visualization*, pp. 77-95. Springer International Publishing, 2014.

Ward, Matthew O., Georges Grinstein, and Daniel Keim. Chapters 11-13 of *Interactive data visualization: foundations, techniques, and applications (2nd Ed)*. CRC Press, 2015.

Munzner, Tamara. Chapters 11-14 of *Visualization Analysis and Design*. CRC Press, 2014.

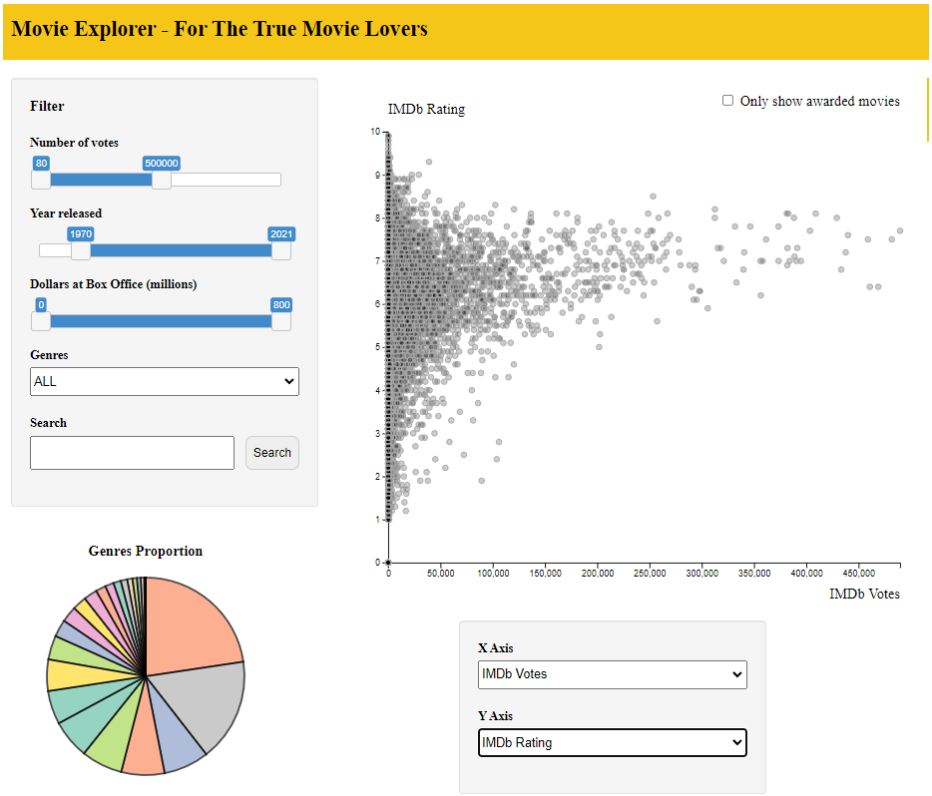
Segel, Edward, and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16, no. 6 : 1139-1148, 2010.

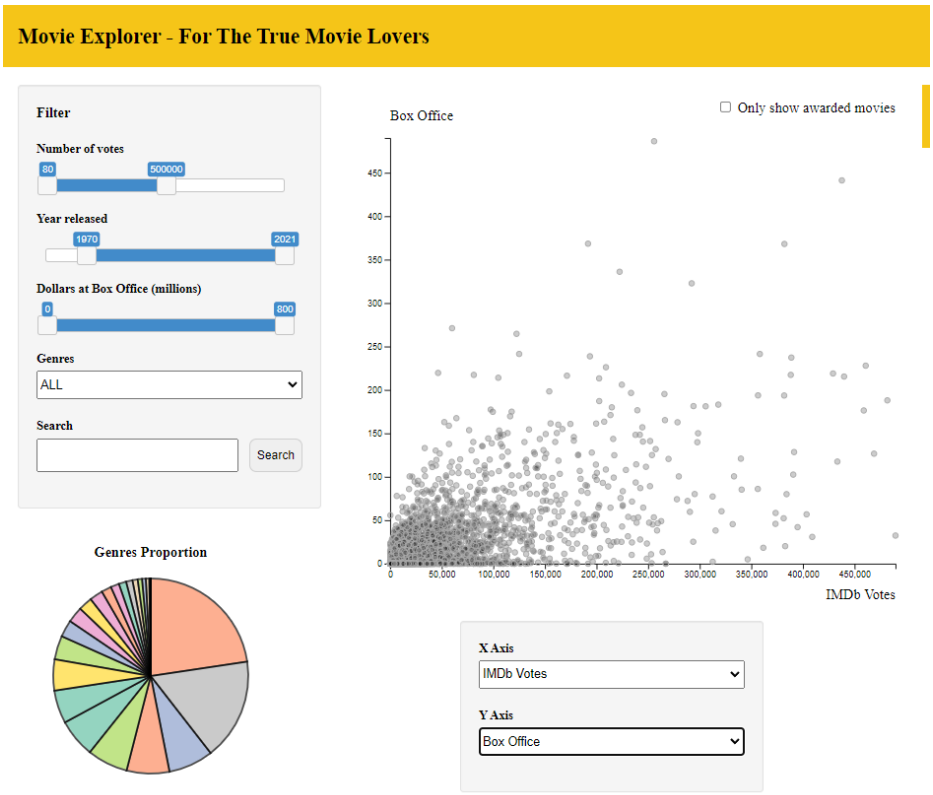
Dataset

IMDb. 2021. IMDb Movies. [online] Available at: https://www.imdb.com/interfaces/ [Accessed 10 August 2021]

OMDb API [online] Available at: http://www.omdbapi.com/ [Accessed 30 October 2021]

Appendix

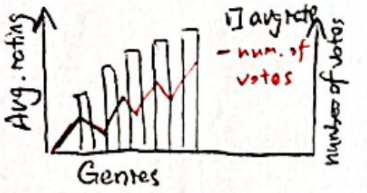

Movie Explorer - For The True Movie Lovers
Filter
Number of votes
80
500000
Year released
1970
2021
Dollars at Box Office (millions)
0
800
Genres
ALL
Search
Search
Genres Proportion
IMDb Rating
Only show awarded movies
IMDb Votes
X Axis
IMDb Votes
Y Axis
IMDb Rating


Movie Explorer - For The True Movie Lovers
Filter
Number of votes
80
500000
Year released
1970
2021
Dollars at Box Office (millions)
0
800
Genres
ALL
Search
Search
Genres Proportion
Box Office
Only show awarded movies
IMDb Votes
X Axis
IMDb Votes
Y Axis
Box Office

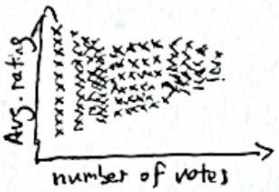
IDEAS

SHEET 1

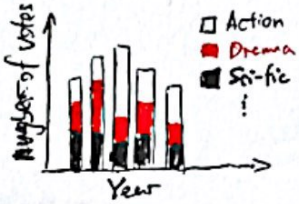
① Bar Chart



② Scatter Plot



③ Stacked Bar Chart

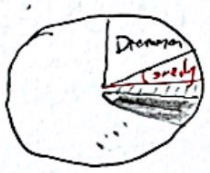


④ Symbol map

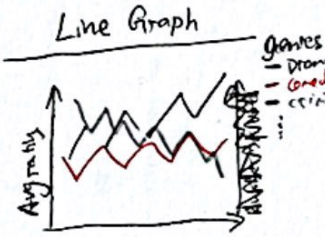
⑤ Choropleth map

⑥ Bubble chart

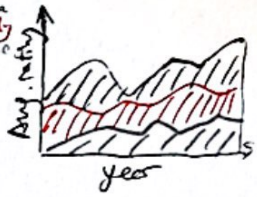
⑦ Pie chart



⑧ Line Graph



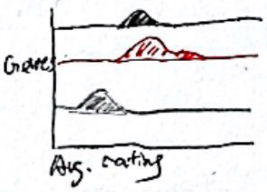
⑨ Stacked area



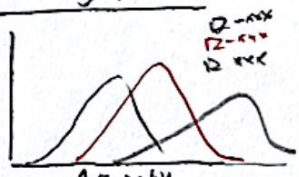
⑩ Radar Chart

⑪ Boxplot

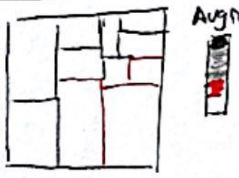
⑫ Ridgeline Graph



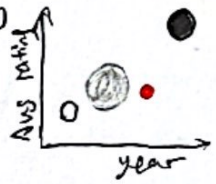
⑬ Density plot



⑭ Tree Map



⑮ Dot chart



FILTER

⑨ Stacked area

- not good for comparison
- The slopes and the shape of the coloured area can be misleading

⑩ RadarChart

- not relevent to this topic, it would be hard to fit 20+genres in one chart, and the visual is meaningless.

④ Symbol map

- not usually appearing
- most movies are from US. if the size of symbol represent the count of movies, the symbol in US will be extremely large.

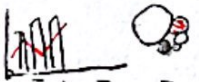
CATEGORIZE

Bar	Proportional	Spatial	Distribution	Line
① ③	⑥ ⑦ ⑭ ⑮	④	⑨ ⑪ ⑫ ⑬	⑧

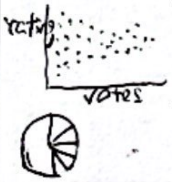
COBINE REFINE

~~① ⑬ can be combined into one, with each bubble has different size/colour based on movie genres and their count~~

① ⑥ ⑧ can be put in one app. bar chart and Line graph can be put into one with same x-axis, and the buble chart can be used to refine the selection of the genres.



② ⑦ can be combined
- Scatterplot shows relationship between 2 variables
- pie chart used to refine the choice of different genres.

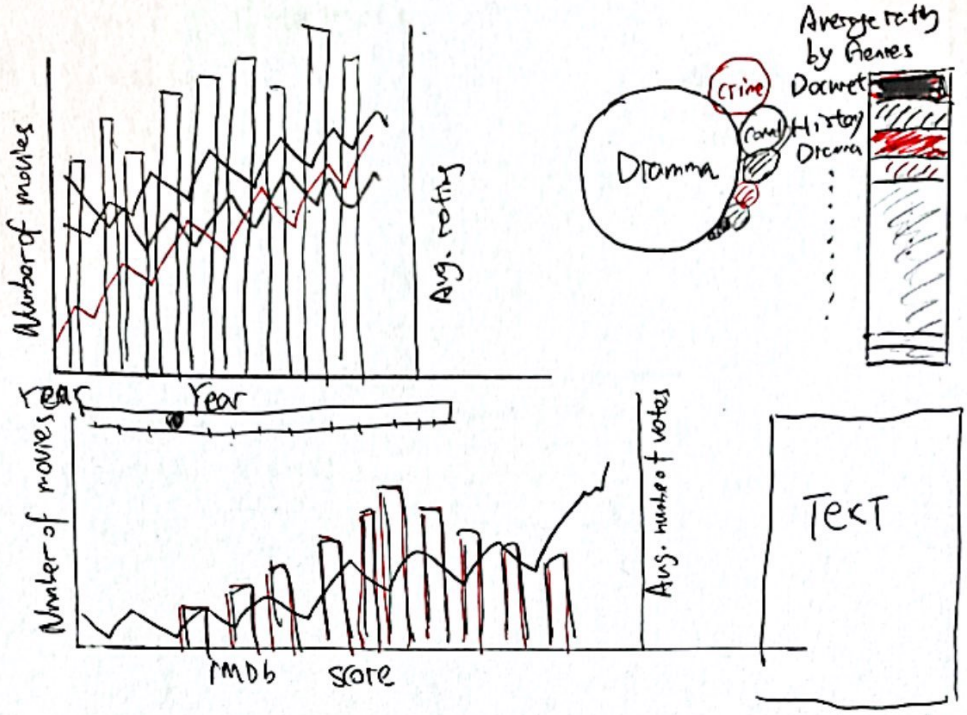


⑤ ⑪ can be combined together
- the map shows the distribution of film
- select different countries to compare their boxplot of ratings.

Questions:

- Is showing location geographically necessary to deliver the key message?
- Can users filter accordig to date period, location or genres?
- Can users view overall graph after filtering are applied?

LAYOUT



Information

Title: FIT5147 Data visualisation project

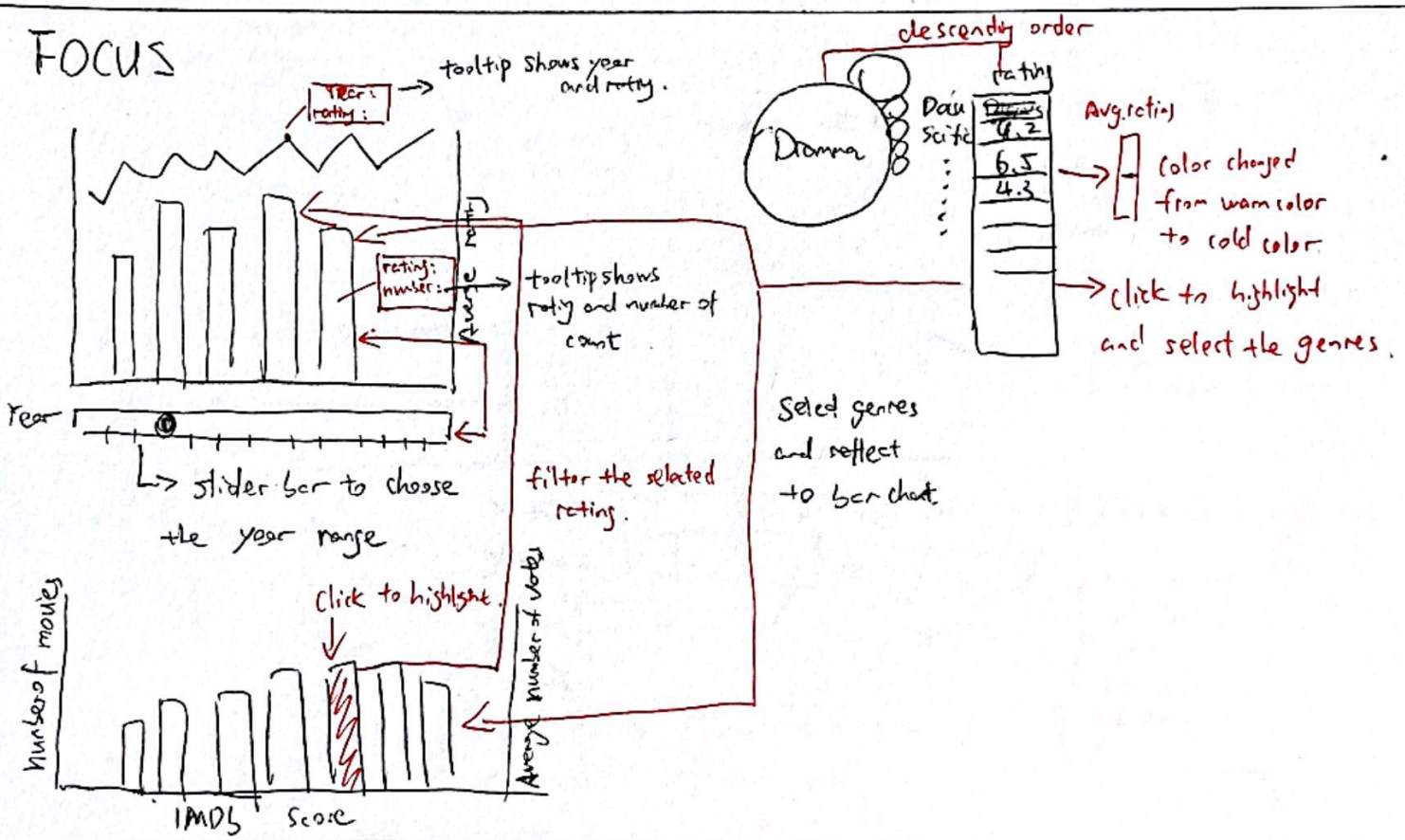
Author: Xiaoyu Tian

Date: 11/10/2021

Sheet: 2

Task: IMDb rating project.

FOCUS



OPERATIONS

- Toolips can be shown when the mouse hover over the chart
- Slider bar can be used to select the time range and filter the bar chart
- the second bar chart can be clicked on and reflect in the first bar chart with selected ratings.
- Bubble chart and the list of genres with their ratings is in descendig order

Click on the generes to ~~select~~ filter the selected genres and shows the data in the bar chart.

Discussion:

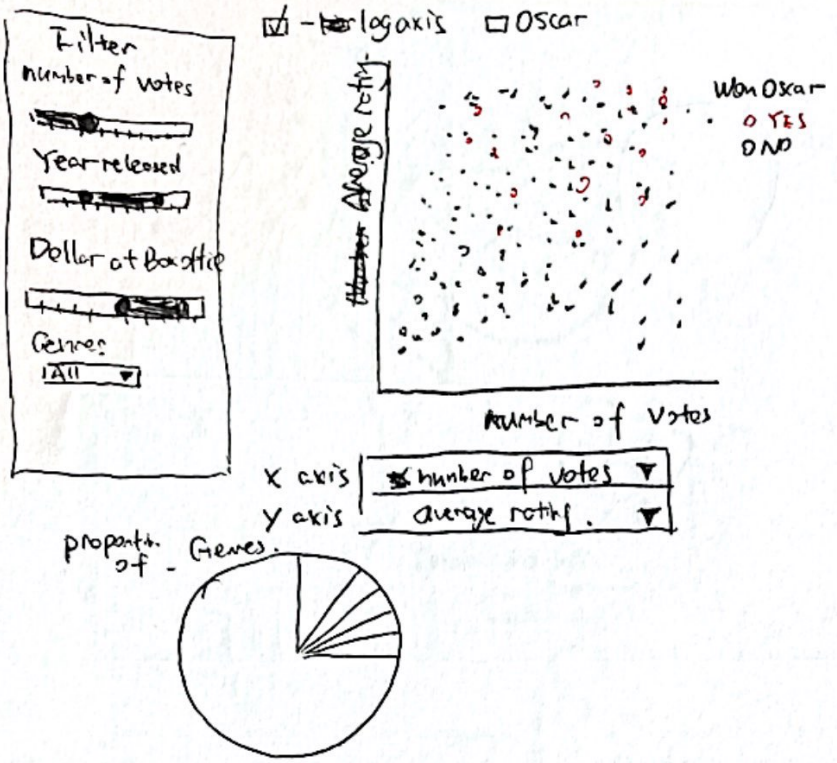
pons:

- present overall summary and trend for the average rating
- users can filter data to view the selected data

Cons:

- How genres contributes to the average rating is not obvious
- Since there are a lot of genres, the line graph would be messy if shown all.

LAYOUT



INFO

Title: FIT5147 Data visualisation project

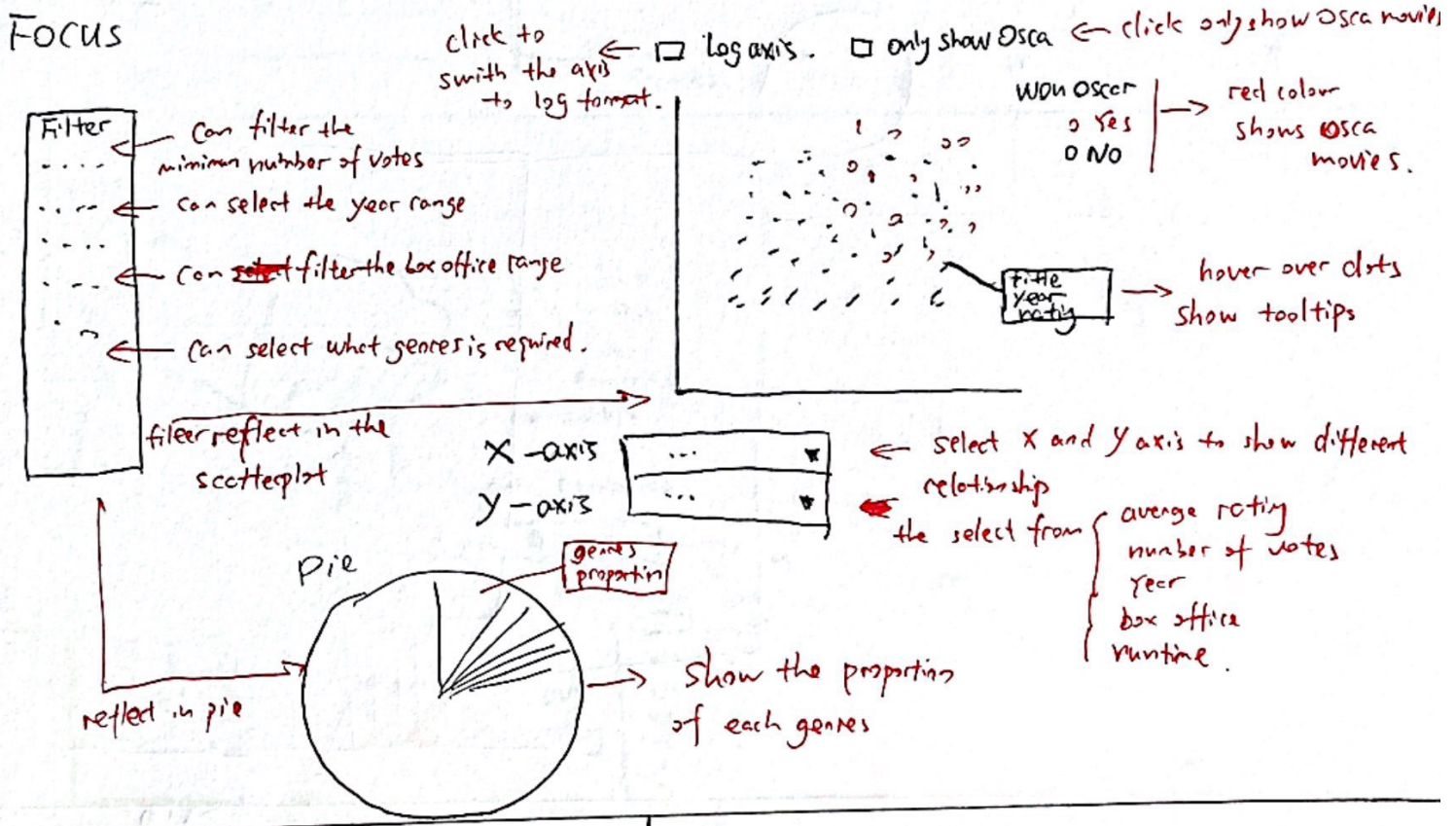
Author: Xiaoyu Tian

Date: 11/10/2021

Sheet 3

Task: IMDb rotty project.

FOCUS



OPERATIONS

- The filter bar leave a lot of variable to filter include: Vote cont, year, dollar data, and genres.
- The X and y axis can be changed with the selection of rotty, number of votes, year, box office and runtime.
- The filtered data will present in the scatterplot and pie chart.
- There are buttons to click on to select log transformation and ~~locate~~ shown Osca movies

Discussions:

Pros:

- All the elements in the dataset can be shown relationship between
- The users can filter the data as their requested

Cons:

- Does not show clearly trends between variables.

LAYOUT

Info

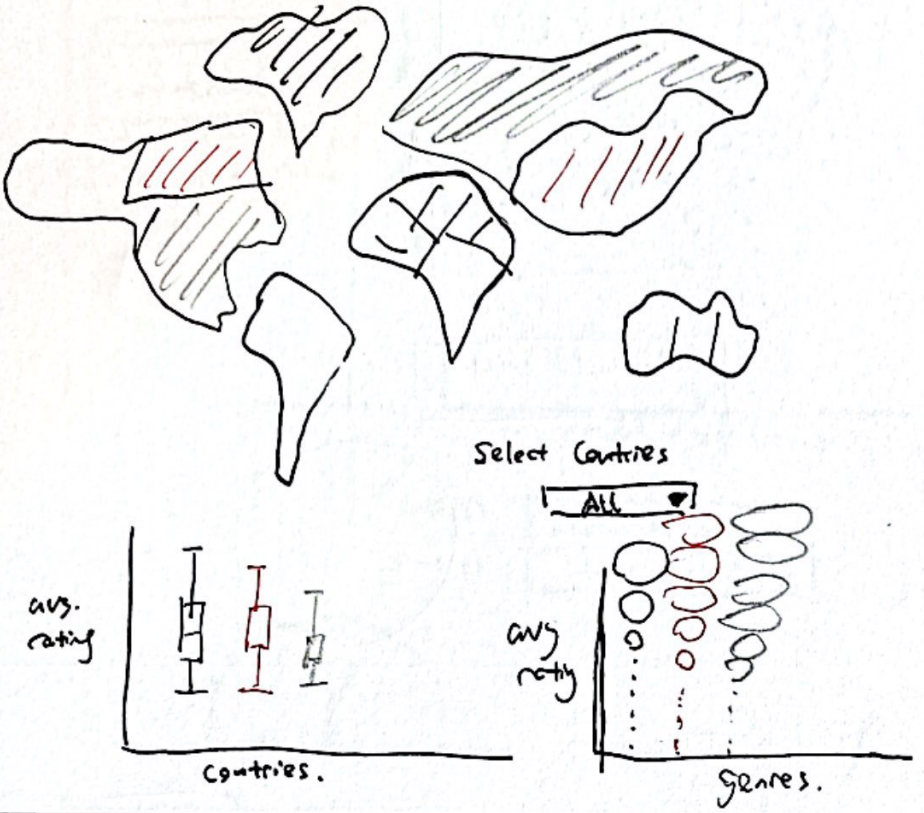
Title : FIT5147 Data visualisation project

Author : Xiaoyu Tian

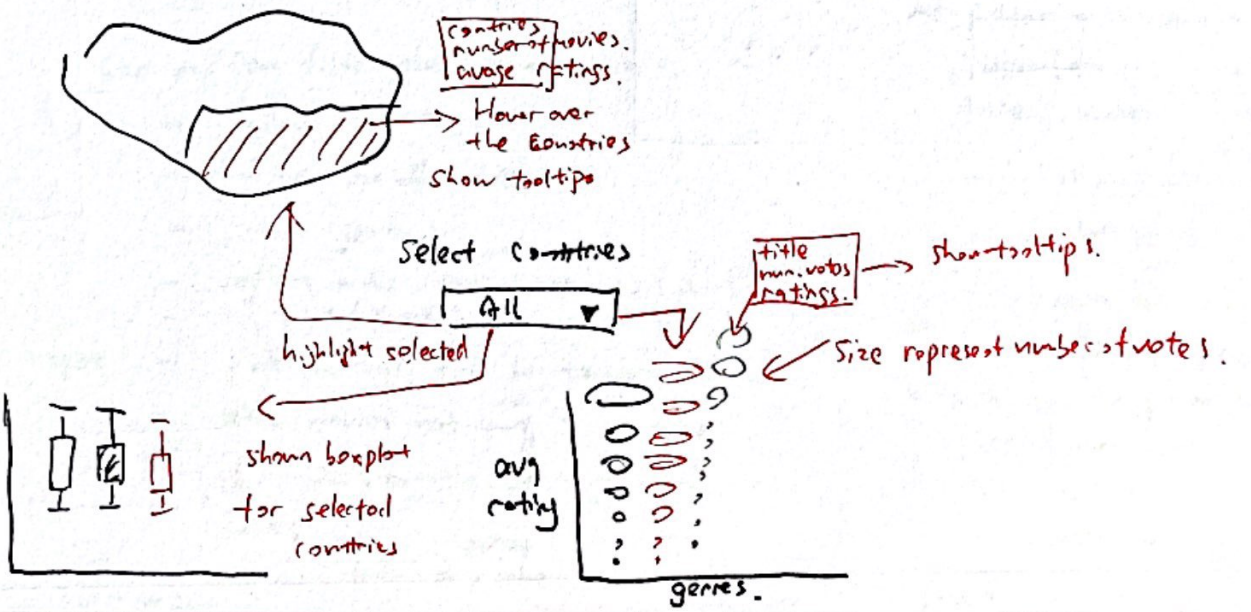
Date : 11/10/2021

Sheet 4

Task : IMDB rating project



Focus



OPERATIONS :

- The map shows the geometry distribution of the movies, Mouse hover around shows the tooltip
- select bar to select the countries to highlight and it will affect the boxplot shown in the boxplot graph
- the size of the dot graph represents the number of votes for each genres
- when hover over the dot, tooltips shows the movie details

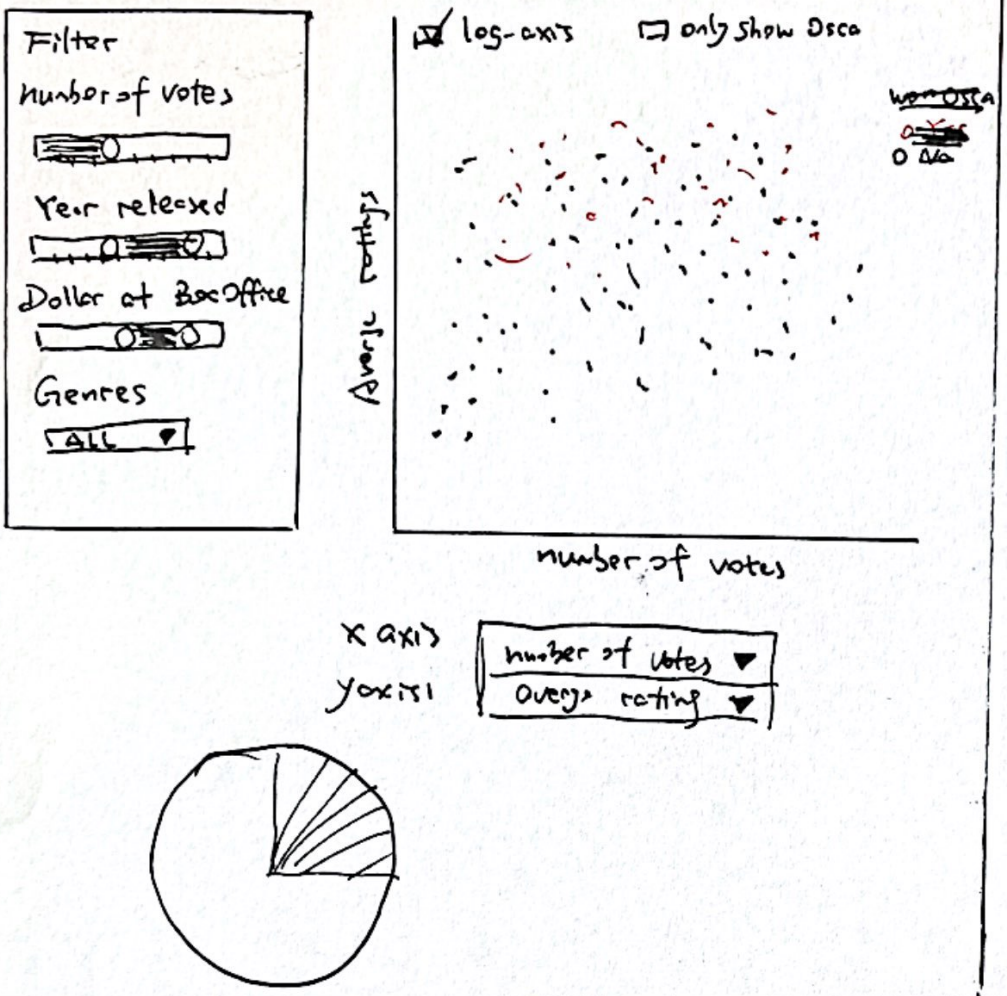
Discussion :

Pros : Clearly shown the geometry distribution of the movies.

- present the statistical data to see the mean, median and the outliners.

Cons : the relationship between the genres and rating not very obvious, also the number of the votes.

LAYOUT



INFO

Title: FIT5147 Data Visualisation project

Author: Xiaoyu Tian

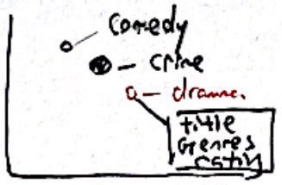
Date: 11/10/2021

Sheet 5

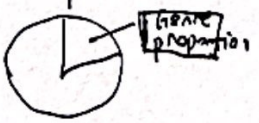
Task: IMDb rating project

Focus

- Filter bar — Can filter the minimum number of votes.
  - select year range
  - filter Box office data
  - select genres
  - reflect in the scatterplot and pie chart
  - select x/y axis's
- Scatterplot — click buttons for log transformation and hide movies not won Oscar prize
  - Color for the dot represent different genres



  - tooltip will be shown when hover over
- Pie chart represents the proportion of each genres
  - tooltip shown when hover



OPERATION

- Filter menu allows users to filter between, year, box office genres, number of votes
- highlight the selected data in the ~~dataset~~ scatterplot
- only shown the selected genres in the scatterplot
- highlight the proportion in the pie chart when select genres.
- change x/y axis would reflect on scatterplot

Details

Dataset

→ A combined dataset from IMDb and Omdb and web scraping. the dataset contains detail information for each movies around the world.

Dependencies

→ R Shiny

Estimate of time:

- 2 weeks.