COMPSCI5099:
Information Visualisation

Group Project Report: Multiview Visualisation

Submitted by:

Group # 18	
Student Name	**GUID**
Ahnaf Ismat Tasin	2739690T
Kirtika	2726518K
Mansi Jain	2762889J
Mohammad Shahmidul Islam	2753415I
Subhomoy Haldar	2739302H

Demo Video: https://www.youtube.com/watch?v=hbDbljEdGxU

This is a link to a YouTube video that demonstrates all three systems and explains the basics of their design and implementation.

Table of Contents

A. Design and Implementation

1. The Data

The **IMDb Top 100 Movies** is a list of the top one hundred movies of all time compiled by IMDb. All one-hundred movies in the list date from 1972 to 2015 and have 9 different columns for each row-entry in the list.

Link to the source data: IMDb Top 100 Movies | Kaggle

In the dataset[1], the rows contain the movies and each movie is an *item*. Each of the nine columns represent the *attributes* of each movie. Below is how our dataset can be categorized for each movie.

1. *Genre*: This is a categorical feature since genres are qualitative data and cannot be arranged in a hierarchical order, hence it is unordered. Examples of genres are horror, crime, drama etc
2. *Category*: This is the age group in which a film is rated. It is diverging ordinal ordered data because certain categories serve a larger audience than others. For example, PG-13 will have a larger audience because the viewing age is over 13, whereas R begins at the age of 18, drawing a smaller audience.
3. *Year of release*: It is a quantitative discrete variable, as the dataset is sorted with imdb rating so the years in which movies are released are a diverging quantitative ordered attribute.
4. *Run time*: The length of the movies is measured in minutes which is a continuous variable and it is diverging ordered.
5. *Votes*: It is the number of voting each movie received. It is diverging continuous data.

6. *Imdb rating*: This is a score for a movie calculated by taking the average of all votes. This is a sequential continuous attribute.
7. *Gross total*:This is how much money a movie earned. These values are discrete because they are rounded up to their nearest ten thousands.

2. The Tasks

The primary objective ws to provide users with elegant data visualizations for exploratory analysis and decision-making purposes. Users desire the ability to explore how data is spread across dimensions and discover the interrelations between the attributes. Therefore, we have provided users with a graphical user interface (GUI) platform that has a dynamic interface using which they can examine the full dataset.

Users are able to quickly and easily refine the information being presented, as well as choose a subset of the data based on their particular concerns or requirements. Users have the ability to do a faceted search by using the tools found in the sidebar to filter data. For instance, users may narrow down their search by selecting just "crime" and "drama" in the genre category from the drop-down menu, and then using the slider menu to set the range of the runtime of the movies they are interested in seeing. This would compile a list of movies that fulfill all of the requirements and bring the visualization interface up to date.

Users are also able to interact with the systems by being able to investigate correlations or links that exist between each target object, such as a movie, by selecting a portion of a data from the interface, and these selections will be highlighted in the adjacent interface, allowing users to see the correlation. The brushes selectively highlight on both plots.They can *locate* an object of interest, or a *target* object, by using a mouse-hover gesture to bring up tooltips that provide further information about the selected movie.

Users will also be able to *query* the selected data. When the user has selected the data that is of interest to them, all of their selections will be *summarized* and displayed in a data frame. From this data frame, the user will be able to *identify*, or obtain information about all the aforementioned attributes of each movie, and *compare* each movie.

3. The Core Systems

Three different visualization systems have been implemented and each system is a clearly separate integrated system that can be used to perform all the tasks outlined in Section A.2 (The Task).

Each system utilizes *multiview composition* where **two** coordinated visualizations are displayed simultaneously to provide users with different perspectives or complementary information about the dataset.

We took the approach linked views supporting *brushing and linking*, where interaction with one view (e.g., selecting a data point or applying a filter) affects the other view accordingly, providing a coordinated and seamless exploration experience and helps users identify relationships that may not be easily discernible from a single visualization.

The visualizations in our systems allow rectangular selection and the bar chart allows interval selection. The selection area can be moved around by dragging. The size can be changed by scrolling on the selected area. Click outside the selection area to deselect and reset the plot.

The implementation of each system is outlined below.

System A: Year, Runtime and distribution of Genre.

We present the variation in movie runtimes across different years, the number of movies released in each year from the selected list, and offer users the ability to filter movies based on genre. Every individual data point in the scatter plot represents a movie, and the color intensity of each point corresponds to the IMDb ranking of that movie, where darker colors represent higher ranks.

System B: Income, Rating and Category.

We demonstrate the relationship between a movie's IMDb rating and the income it generated (and mean income of all the movies in the list is also denoted on the visualization), along with the quantity of movies from the curated list released per year. Additionally, users have the option to filter movies by categories. Each distinct point within the scatter plot symbolizes a movie, and the color intensity of each point corresponds to the inverse ranking of that movie, where darker colors represent lower ranks.

System C: Movie Name, Category and Votes count.

We depict the influence of a movie's category on the total number of votes it accumulates, illustrating the impact of genre on a film's popularity. Users are also provided with the option to filter movies by release year. As with previous visualizations, each individual point in the scatter plot represents a movie, and the color intensity of each point corresponds to the inverse ranking of that movie, where darker colors represent lower ranks.

4. Generalized Selection

We were offered the choice of using Python through Altair[2] or JavaScript through Vega. Our team members had prior experience with Python, so we went with the former. Unfortunately, there is no documentation or example showing augmented interaction with tooltips i.e using the right mouse button. For this, we would have to augment the code with JavaScript and extend the project.

We managed to implement partial Generalized Selection[3] through brushing and linking, which would highlight a subset of data across the views. We allow for categorical and range filtering through the controls on the left sidebar. This is evident in our tasks in the following ways:

1. System A allows for filtering by genre. A film can have multiple genres and keeping one genre active would highlight all movies that have the specified genre in its list.
2. System B allows for filtering by the movie category with a multi-selection widget as well. In his case, only one category is assigned per movie.
3. System C allows for filtering by the year of release range as well the votes.

5. Demo Videos

The top of this document includes a link to a YouTube video that demonstrates all three systems and explains the basics of their design and implementation.

6. Design Comparison

Design Decision # 1

Choosing the appropriate method for the second visualisation for each system.

Choices across our three systems:

We chose the Stacked bar chart for System A, Line Chart for System B; and Pie Chart for System C.

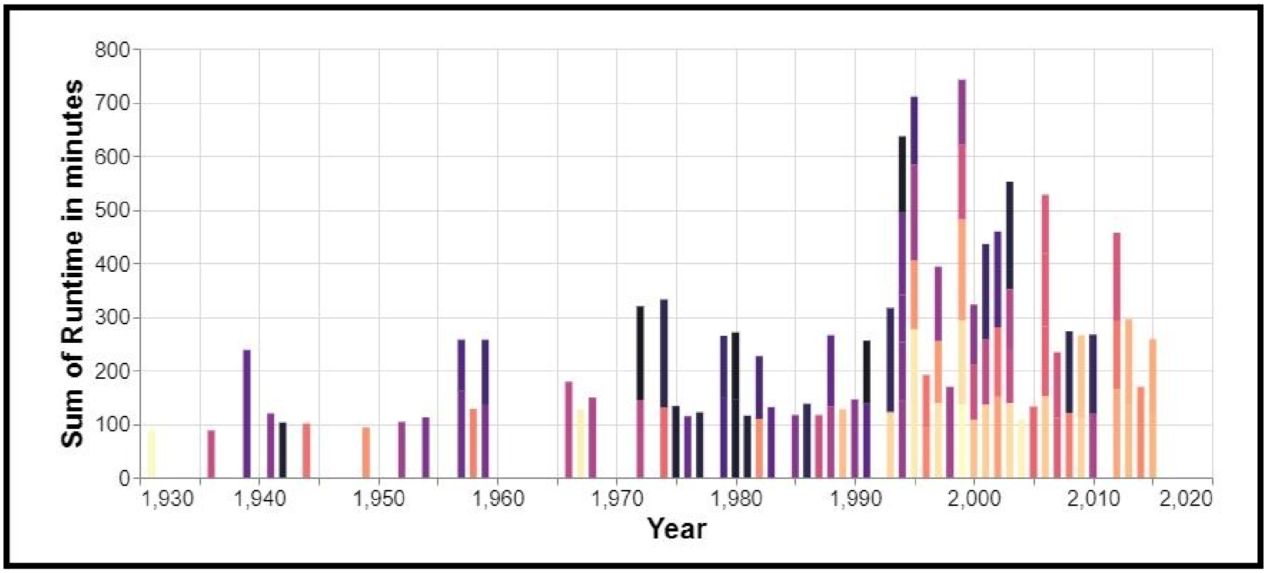


Fig 6.1.a System A: Stacked Bar Chart

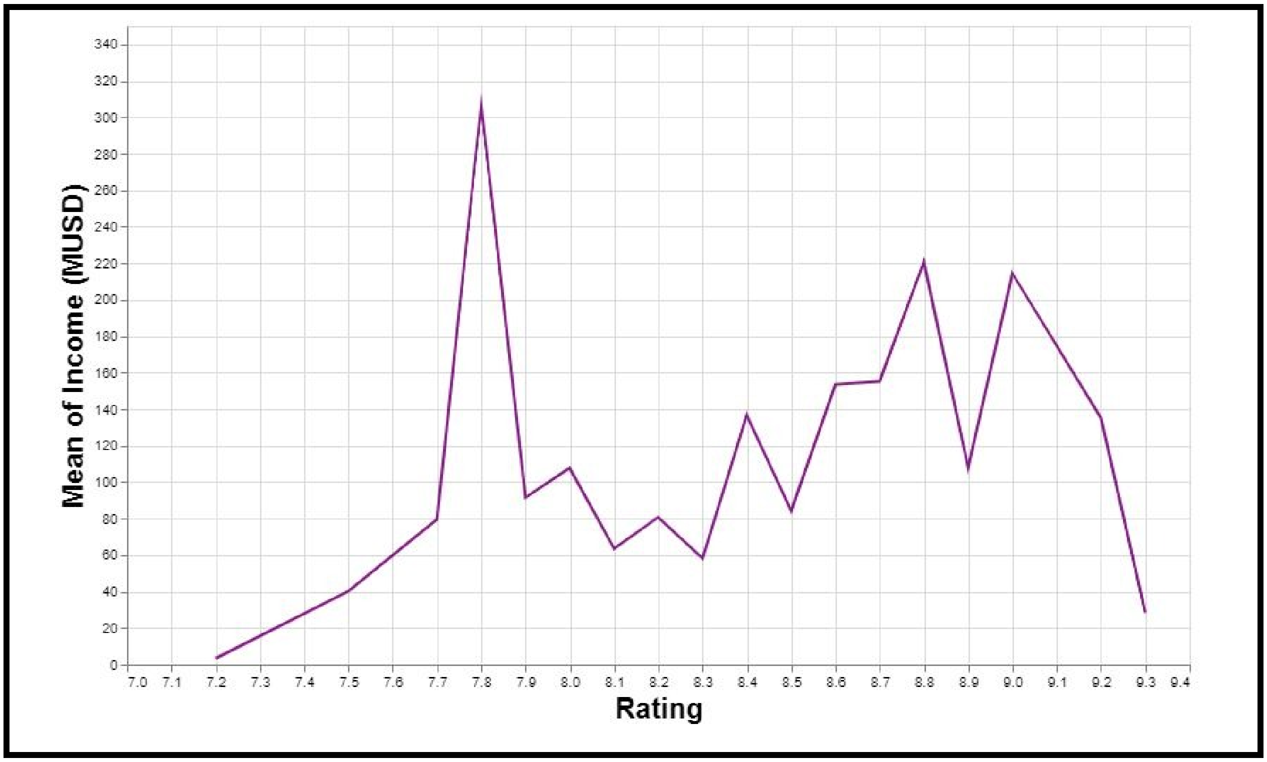


Fig 6.1.b System B: Line Chart

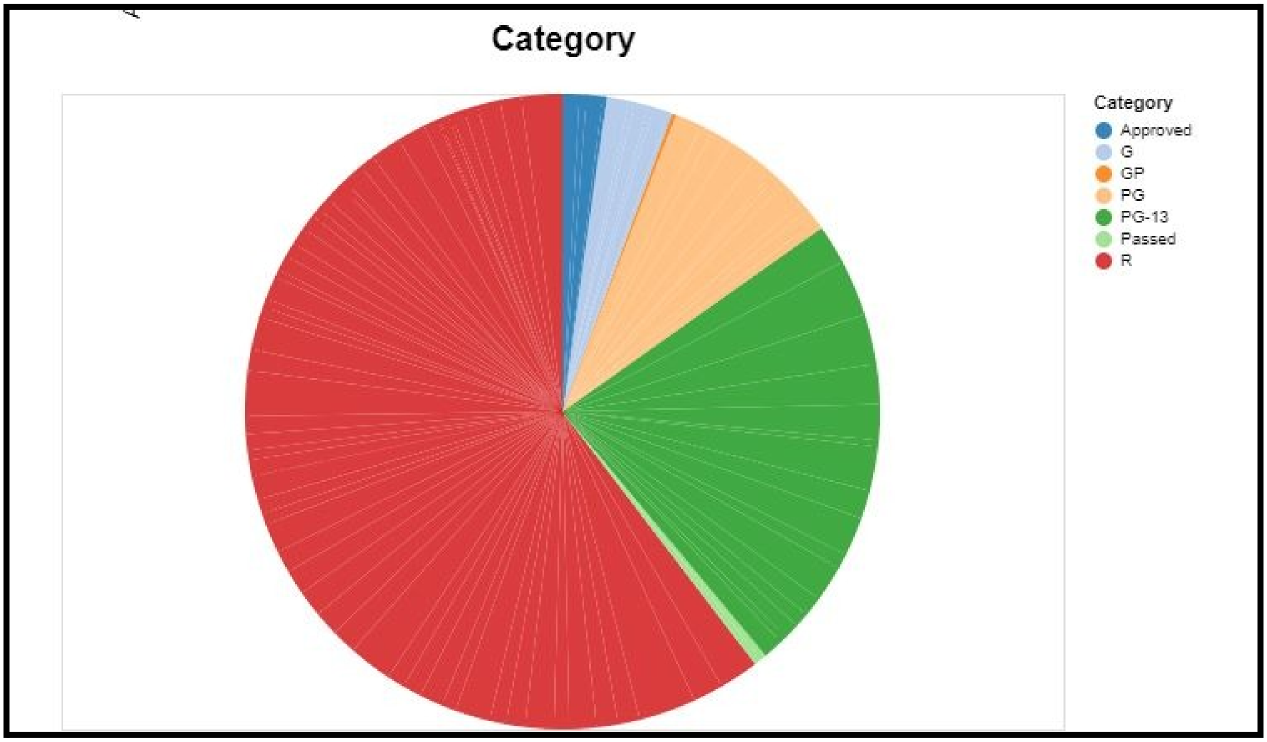


Fig 6.1.c System C: Pie Chart

Alternatives and Justification of our Choices:

For System A, we could have used a pie chart, but it would result in too many segments, making it difficult to determine the relative sizes of the slices.

For System B, an alternative to a line chart would be a radar chart. However, radar charts are primarily used to visualize multivariate data with three or more variables, making them less suitable for displaying two numerical values, ‘Rating’ and ‘Income’.

For System C, we could have chosen a stacked bar chart but that would have resulted the bar for R-rated movies to be too tall and for the ‘GP’, ‘Passed’ or ‘Approved’ categories, the bars would have had almost no substantial height, making the visualization very sparse and not the optimal choice.

Design Decision # 2

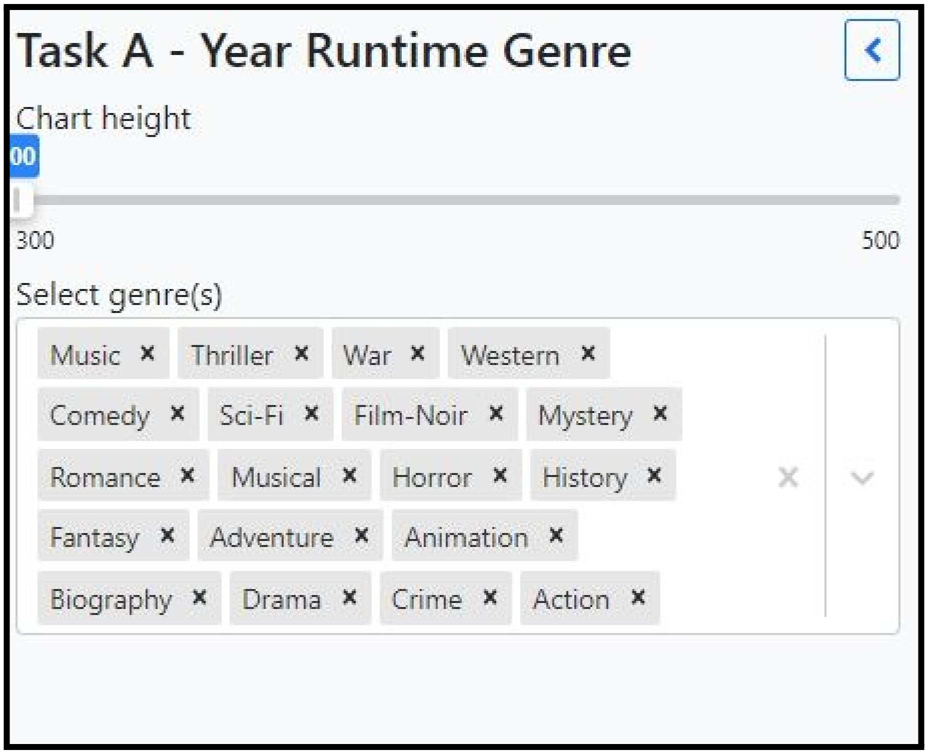
Incorporating dynamic filtering and zooming components in the User Interface

Choices across our three systems:

In System A, we implemented a ‘drop-down multi-select’ tool for filtering the Genre. The ’chart height’ slider tool is used for zooming into the visualization. The zoom magnitude is scaled appropriately for the spread in the visualization (arbitrarily, 300 to 500 units, so a magnitude range of 200 units).

In System B, implemented a ‘drop-down multi-select’ tool for filtering the Category. The ’chart height’ slider tool used here has wider range due to the greater dispersion of data points, providing more opportunities for zooming and exploration of the visualized data.

In System C, we used a slider for specifying the range of ‘Votes’ and the ‘Year of Release’ as both these are numerical values. The ‘chart height’ tool works similarly as in the other systems.



Task A - Year Runtime Genre

Chart height

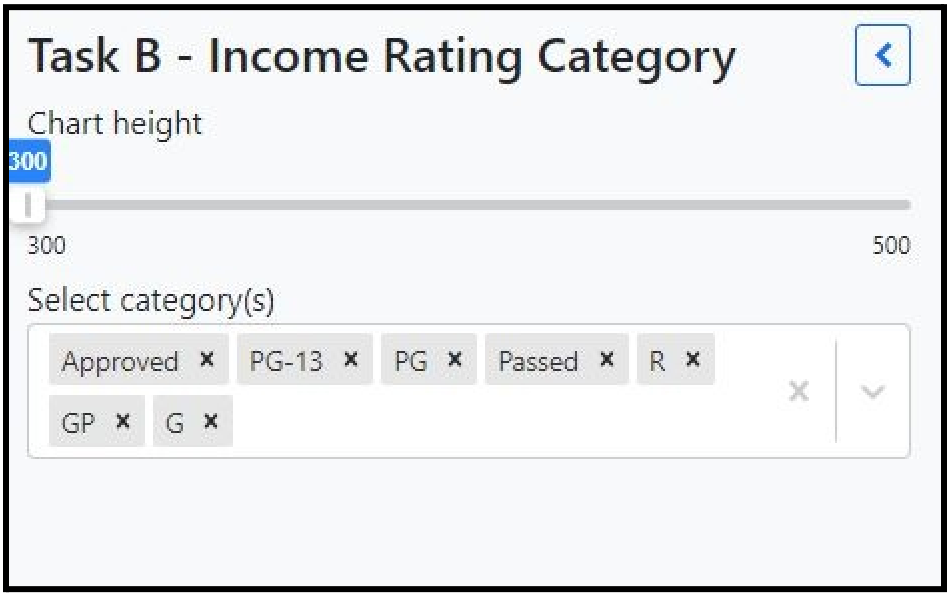
00

300 500

Select genre(s)

Music × Thriller × War × Western ×
Comedy × Sci-Fi × Film-Noir × Mystery ×
Romance × Musical × Horror × History ×
Fantasy × Adventure × Animation ×
Biography × Drama × Crime × Action ×

Fig 6.2.a System A: Drop-down Multi-select and Chart Height



Task B - Income Rating Category

Chart height

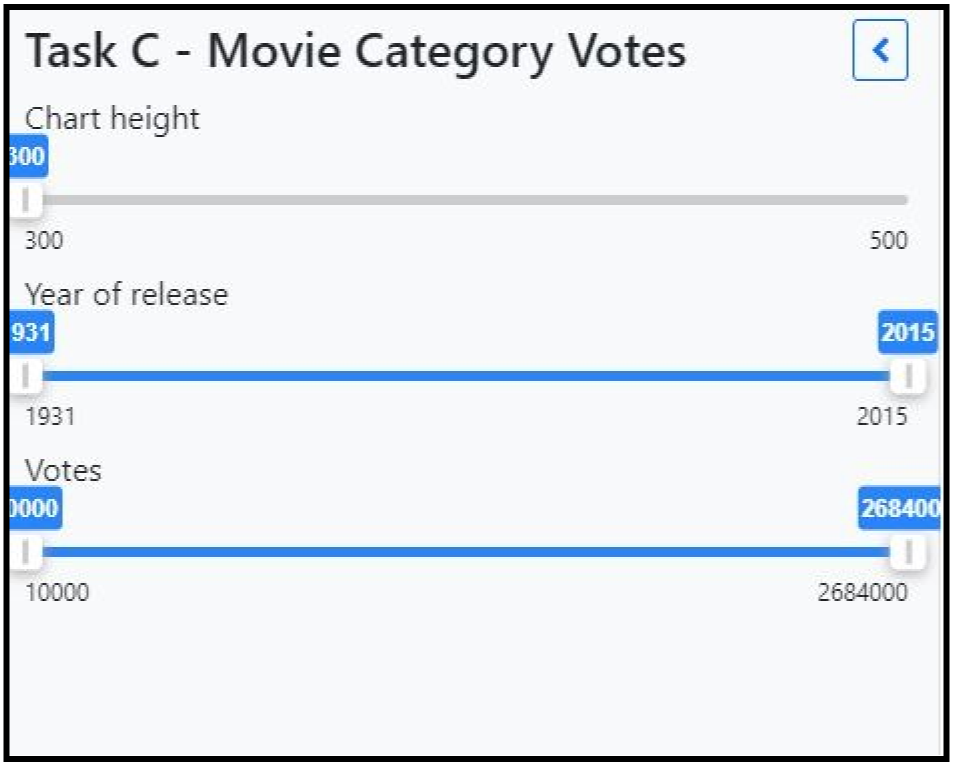
300

300 500

Select category(s)

Approved × PG-13 × PG × Passed × R ×
GP × G ×

Fig 6.2.b System B: Drop-down Multi-select and Chart Height



Task C - Movie Category Votes

Chart height
300

300 500

Year of release
931 2015

1931 2015

Votes
0000 268400

10000 2684000

Fig 6.2.c System C: Slider and Chart Height

Alternatives and Justification of our Choices:

For System A and B, we could have used a series of checkboxes, which would have cluttered the interface but we went with our selection to save space.

For System C, we could have used two separate input fields for the lower and upper limits of the range. Although this method offers precise value input, we chose slider because it is faster to use and more intuitive.

Design Decision # 3

Selecting varying color schemes and with opposite meanings of intensity in each system

Choice across our three systems:

For System A, we chose the ‘Magma’ color theme and the color intensity of each point corresponds to the IMDb ranking of that movie, where darker colors represent higher ranks.

For System B, we chose the Yellow Green Blue color theme and the color intensity of each point corresponds to the inverse ranking of that movie, where darker colors represent lower ranks.

For System C, we chose the ‘Light Teal Blue’ color theme and the color intensity of each point corresponds to the inverse ranking of that movie, where darker colors represent lower ranks.

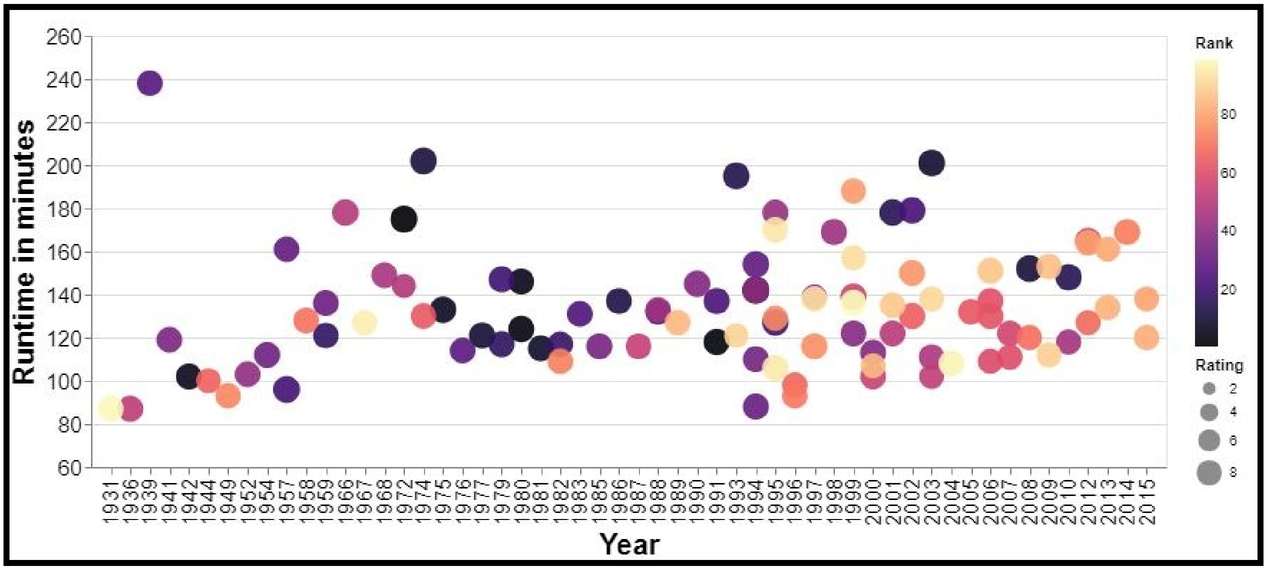


Fig 6.3.a The Scatter Plot in System A using the Magma color theme

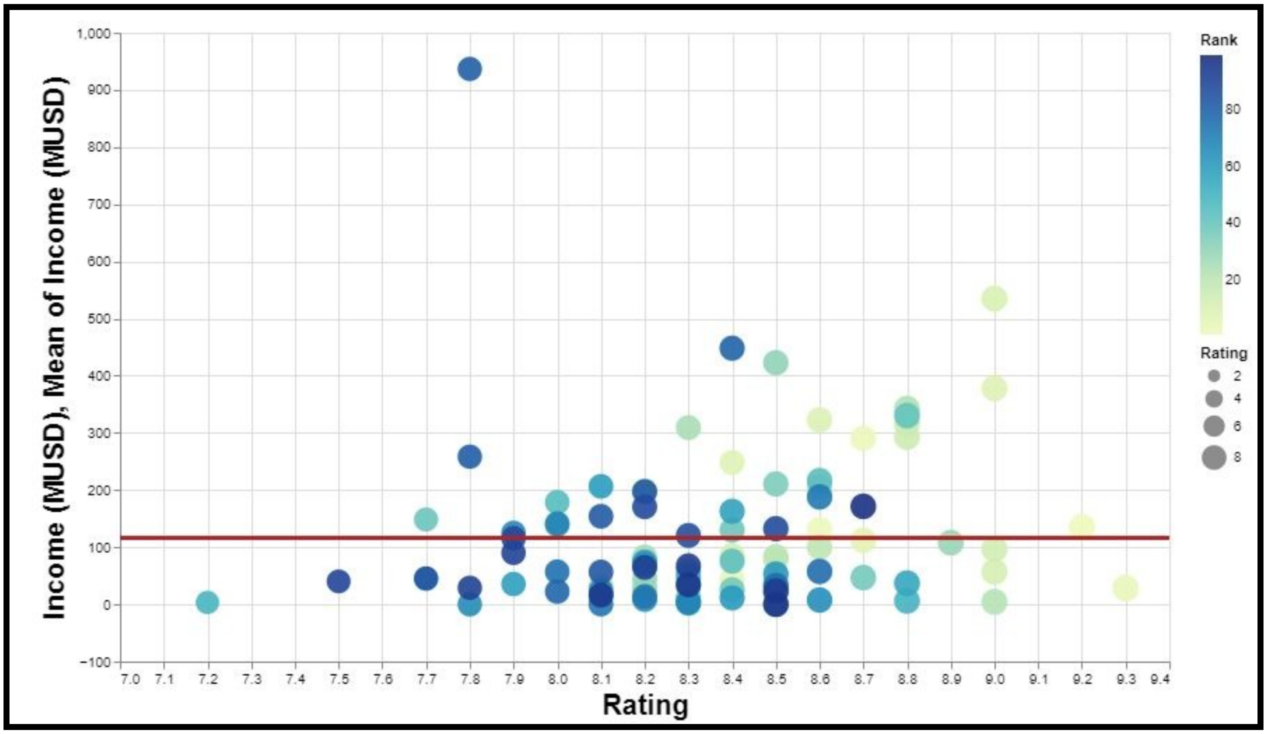


Fig 6.3.b The Scatter Plot in System B using the Yellow-Green-Blue color theme

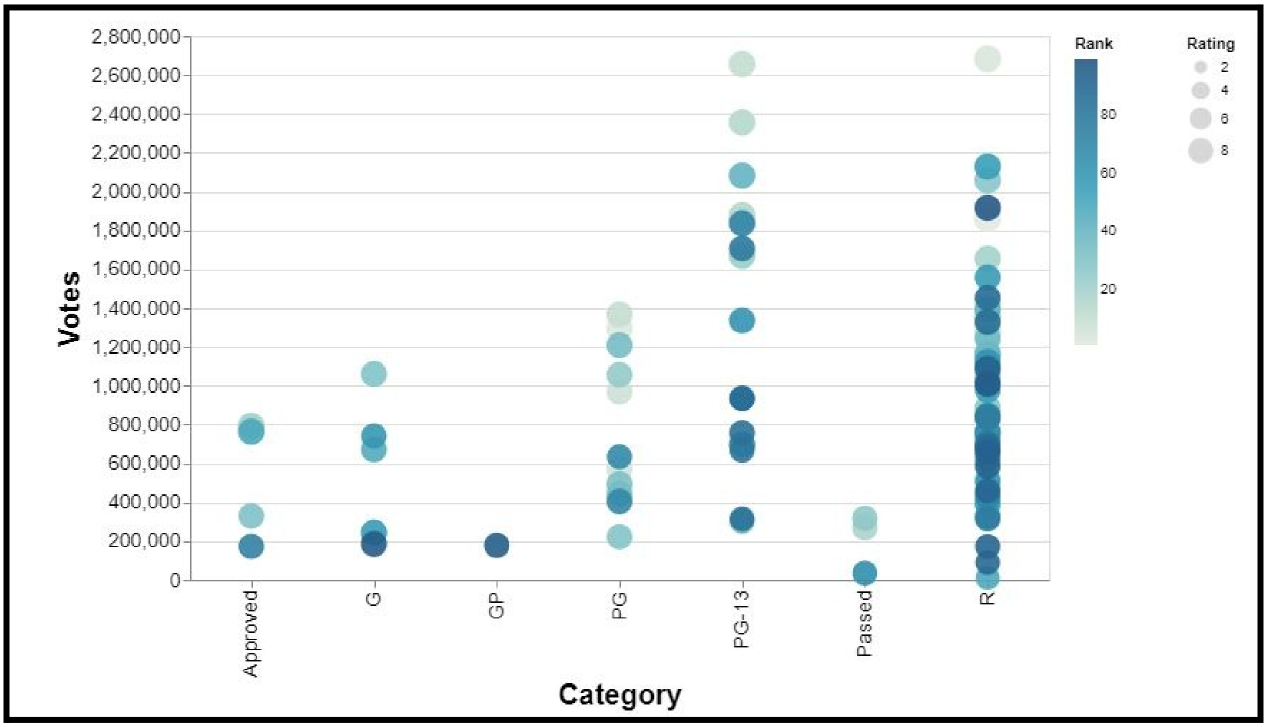


Fig 6.3.b The Scatter Plot in System C using the Light Teal Blue color theme

Alternative and Justification of our Choice:

We could have used a Black Grey White color theme. But, in doing so, it can make it difficult to distinguish between data points and can result in a lack of contrast, making it challenging to identify patterns and trends in the data. Additionally, this color scheme can be perceived as monotonous and unengaging, potentially leading to a loss of interest from the audience.

Design Decision # 4

Varying the chart width for the scatter plot visualization for each system

Choice across our three systems:

In System A and B, the chart width is wider because there is a lot of data on the X axis.

In System C,the chart width is more compact

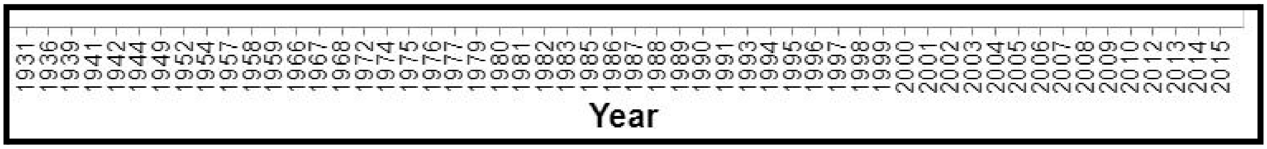


Fig 6.4.a System A Chart Width

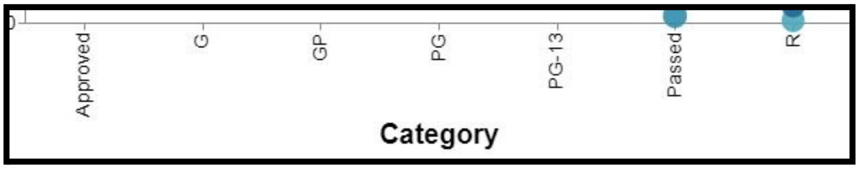


Fig 6.4.b System C Chart Width

Alternative and Justification of our Choice:

An alternative would be to keep the same chart width across all visualizations. But in doing so, we would either be wasting a lot of space on the horizontal axis or cramming the axis with too many points. We varied the chart width in our visualizations in order to optimize the use of available space and improve the visual appeal of the visualization.

Design Decision # 5

Incorporating different font sizes across our systems

Choice across our three systems:

In reference to the scatter plot visualization of each system,

In System A, the font size is kept considerably bigger than the label text font

In System B, the font size is kept significantly smaller than the label text font

In System C, the font size is kept relatively the same size as the label text.

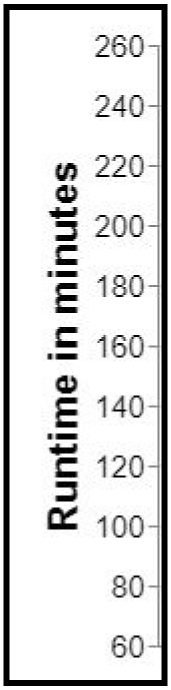


Fig 6.5.a System A: Font Size compared to Label text font size

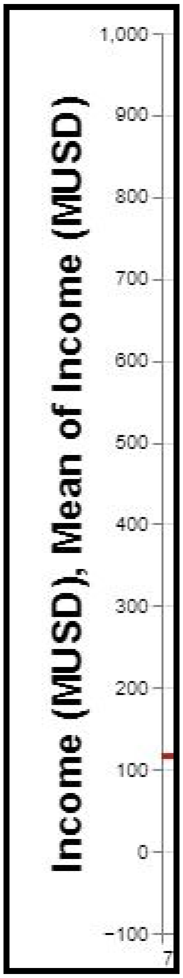


Fig 6.5.b System B: Font Size compared to Label text font size

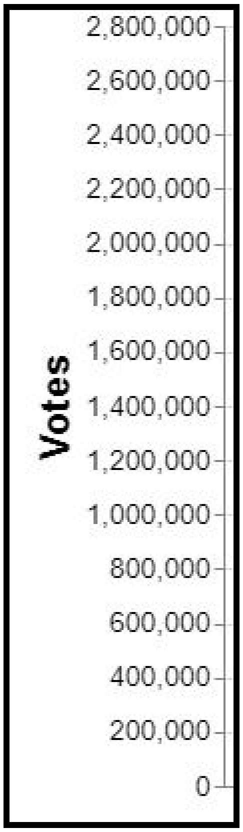


Fig 6.5.c System C: Font Size compared to Label text font size

Alternative and Justification:

For each of the systems, we could have kept a a uniform font size but in doing so, we would be unable to emphasize specific information or create a visual hierarchy. Hence, we chose to vary the typography to impact the legibility, readability, and overall effectiveness of the visualizations.

Design Decision # 6

Incorporating more interactions on a specific visualization

Choice across our three systems:

In System B, we added an additional layer of visualization by including a 'floating line' - a red horizontal line in the figure below - to represent the average income of the selected movies. This line dynamically adjusts as new groups of movies are selected, recalculating the mean income of the selected movies and providing a clear point of comparison against individual movie data points. By incorporating this feature, we aimed to provide deeper insights into the distribution and relative performance of the selected movies compared to System A and System C.

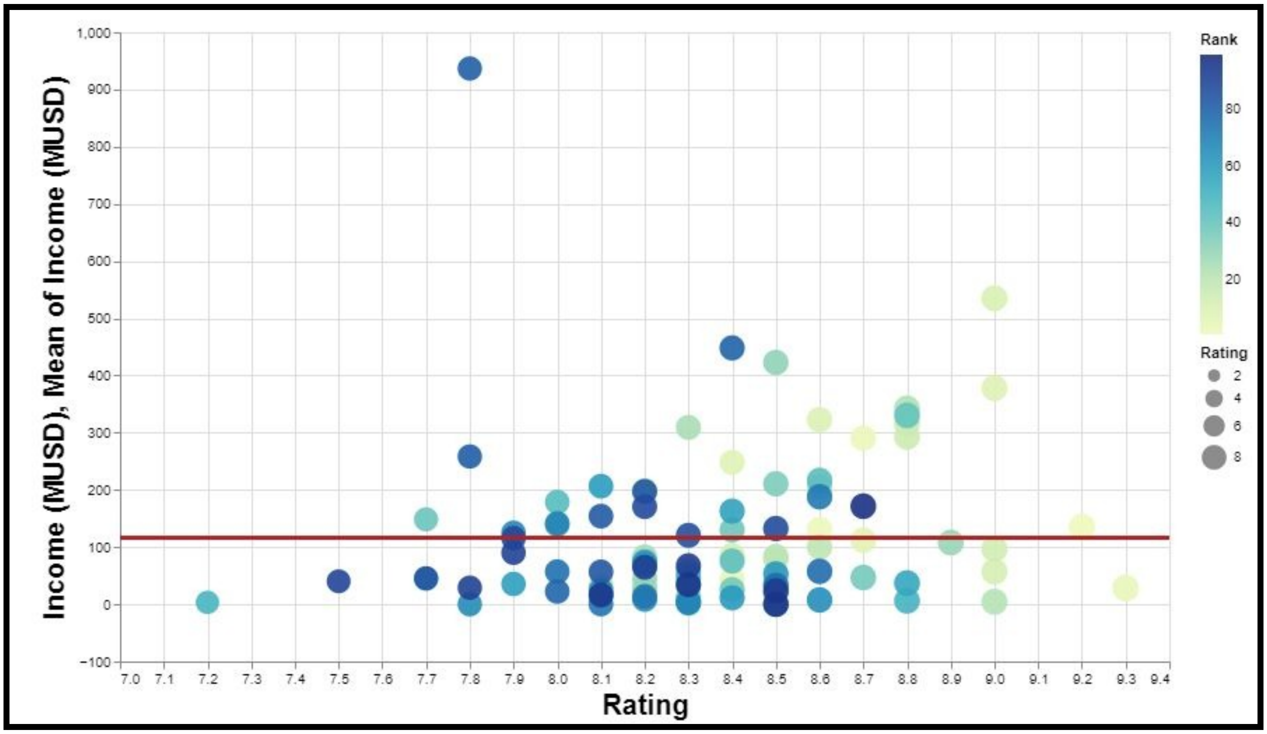


Fig 6.5.b System B: Font Size compared to Label text font size

Alternative and Justification:

Opting to maintain uniformity across all visualizations, without exploring opportunities for creativity, presents a potential alternative. However, by doing so, we would forgo the chance to innovate and potentially enhance the effectiveness of the overall impact of the visualizations on their intended audience.

B. Evaluation

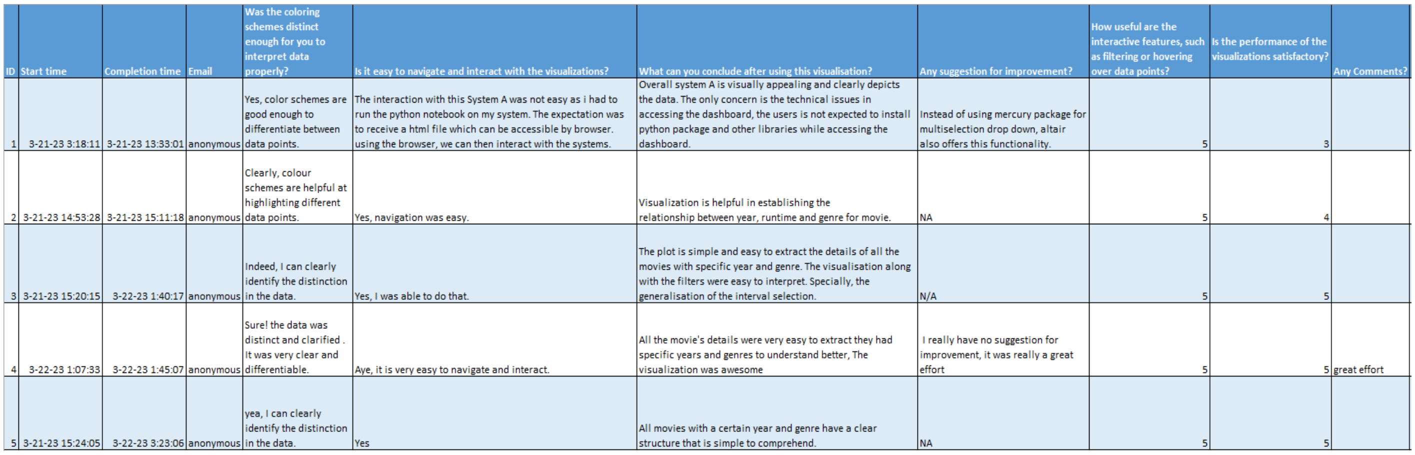
7. User Evaluation Comparison

We chose the questionnaire method for our user evaluation due to its simplicity and speed. For our review, we considered five PhD students at the University of Glasgow. All subjects are current students at the engineering department, with two of them from the school of computing science and three of them from varied departments across the other schools under the engineering department.

Our primary focus was on assessing performance, which meant evaluating how effectively users are able to understand and interact with the visualizations presented to them. To achieve this, we looked at two key factors: response time and accuracy. Response time refers to the amount of time it takes for users to complete a task, while accuracy assesses how well users can comprehend the information provided. We also considered the process through which users engage with data visualizations, and the steps they take to draw conclusions or insights.

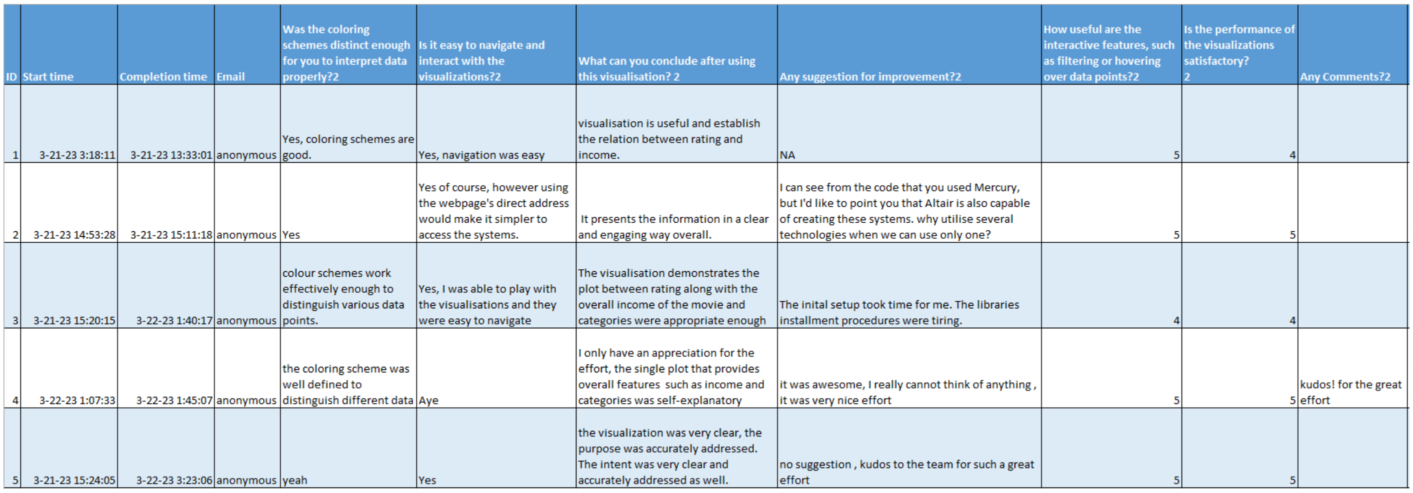
In addition, we considered the user experience of the visualizations. This involved assessing the usefulness and appeal of the visualizations to the users. User feedback gave us great information into how the visualizations were utilized and helped us discover areas that required development. Ultimately, our approach to data collecting and analysis enabled us to get an in-depth comprehension of both user performance and experience. We were able to discover areas for improvement and make educated judgments about how to improve the user experience of our visualizations.

Appendix:



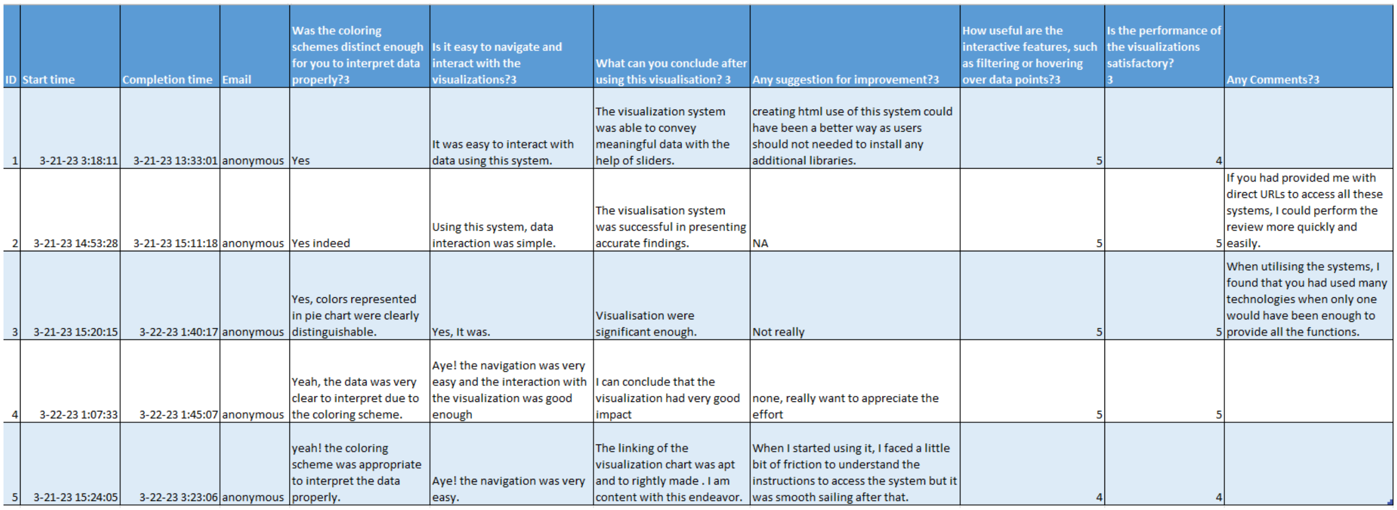
ID	Start time	Completion time	Email	Was the coloring schemes distinct enough for you to interpret data properly?	Is it easy to navigate and interact with the visualizations?	What can you conclude after using this visualisation?	Any suggestion for improvement?	How useful are the interactive features, such as filtering or hovering over data points?	Is the performance of the visualizations satisfactory?	Any Comments?
1	3-21-23 3:18:11	3-21-23 13:33:01	anonymous	Yes, color schemes are good enough to differentiate between data points.	The interaction with this System A was not easy as i had to run the python notebook on my system. The expectation was to receive a html file which can be accessible by browser. using the browser, we can then interact with the systems.	Overall system A is visually appealing and clearly depicts the data. The only concern is the technical issues in accessing the dashboard, the users is not expected to install python package and other libraries while accessing the dashboard.	Instead of using mercury package for multiselection drop down, altair also offers this functionality.	5	3	
2	3-21-23 14:53:28	3-21-23 15:11:18	anonymous	Clearly, colour schemes are helpful at highlighting different data points.	Yes, navigation was easy.	Visualization is helpful in establishing the relationship between year, runtime and genre for movie.	NA	5	4	
3	3-21-23 15:20:15	3-22-23 1:40:17	anonymous	Indeed, I can clearly identify the distinction in the data.	Yes, I was able to do that.	The plot is simple and easy to extract the details of all the movies with specific year and genre. The visualisation along with the filters were easy to interpret. Specially, the generalisation of the interval selection.	N/A	5	5	
4	3-22-23 1:07:33	3-22-23 1:45:07	anonymous	Sure! the data was distinct and clarified . It was very clear and differentiable.	Aye, it is very easy to navigate and interact.	All the movie's details were very easy to extract they had specific years and genres to understand better, The visualization was awesome	I really have no suggestion for improvement, it was really a great effort	5	5	great effort
5	3-21-23 15:24:05	3-22-23 3:23:06	anonymous	yea, I can clearly identify the distinction in the data.	Yes	All movies with a certain year and genre have a clear structure that is simple to comprehend.	NA	5	5	

Fig 7.1 Feedback for System A



ID	Start time	Completion time	Email	Was the coloring schemes distinct enough for you to interpret data properly?2	Is it easy to navigate and interact with the visualizations?2	What can you conclude after using this visualisation? 2	Any suggestion for improvement?2	How useful are the interactive features, such as filtering or hovering over data points?2	Is the performance of the visualizations satisfactory? 2	Any Comments?2
1	3-21-23 3:18:11	3-21-23 13:33:01	anonymous	Yes, coloring schemes are good.	Yes, navigation was easy	visualisation is useful and establish the relation between rating and income.	NA	5	4	
2	3-21-23 14:53:28	3-21-23 15:11:18	anonymous	Yes	Yes of course, however using the webpage's direct address would make it simpler to access the systems.	It presents the information in a clear and engaging way overall.	I can see from the code that you used Mercury, but I'd like to point you that Altair is also capable of creating these systems. why utilise several technologies when we can use only one?	5	5	
3	3-21-23 15:20:15	3-22-23 1:40:17	anonymous	colour schemes work effectively enough to distinguish various data points.	Yes, I was able to play with the visualisations and they were easy to navigate	The visualisation demonstrates the plot between rating along with the overall income of the movie and categories were appropriate enough	The inital setup took time for me. The libraries installment procedures were tiring.	4	4	
4	3-22-23 1:07:33	3-22-23 1:45:07	anonymous	the coloring scheme was well defined to distinguish different data	Aye	I only have an appreciation for the effort, the single plot that provides overall features such as income and categories was self-explanatory	it was awesome, I really cannot think of anything , it was very nice effort	5	5	kudos! for the great effort
5	3-21-23 15:24:05	3-22-23 3:23:06	anonymous	yeah	Yes	the visualization was very clear, the purpose was accurately addressed. The intent was very clear and accurately addressed as well.	no suggestion , kudos to the team for such a great effort	5	5	

Fig 7.2 Feedback for System B



ID	Start time	Completion time	Email	Was the coloring schemes distinct enough for you to interpret data properly?3	Is it easy to navigate and interact with the visualizations?3	What can you conclude after using this visualisation? 3	Any suggestion for improvement?3	How useful are the interactive features, such as filtering or hovering over data points?3	Is the performance of the visualizations satisfactory? 3	Any Comments?3
1	3-21-23 3:18:11	3-21-23 13:33:01	anonymous	Yes	It was easy to interact with data using this system.	The visualization system was able to convey meaningful data with the help of sliders.	creating html use of this system could have been a better way as users should not needed to install any additional libraries.	5	4	
2	3-21-23 14:53:28	3-21-23 15:11:18	anonymous	Yes indeed	Using this system, data interaction was simple.	The visualisation system was successful in presenting accurate findings.	NA	5	5	If you had provided me with direct URLs to access all these systems, I could perform the review more quickly and easily.
3	3-21-23 15:20:15	3-22-23 1:40:17	anonymous	Yes, colors represented in pie chart were clearly distinguishable.	Yes, It was.	Visualisation were significant enough.	Not really	5	5	When utilising the systems, I found that you had used many technologies when only one would have been enough to provide all the functions.
4	3-22-23 1:07:33	3-22-23 1:45:07	anonymous	Yeah, the data was very clear to interpret due to the coloring scheme.	Aye! the navigation was very easy and the interaction with the visualization was good enough	I can conclude that the visualization had very good impact	none, really want to appreciate the effort	5	5	
5	3-21-23 15:24:05	3-22-23 3:23:06	anonymous	yeah! the coloring scheme was appropriate to interpret the data properly.	Aye! the navigation was very easy.	The linking of the visualization chart was apt and to rightly made . I am content with this endeavor.	When I started using it, I faced a little bit of friction to understand the instructions to access the system but it was smooth sailing after that.	4	4	

Fig 7.3 Feedback for System C

8. Future Work

After a thorough examination of the collected data, we can determine that for what the systems offer, they are working at a decently efficient level. We can make this claim as users reported good user experience with only a couple of suggestions of improvement with the systems.

Users said that they will find it much simpler to access the systems if they have a direct URL link to pages where they can complete their analysis quickly and effectively. By implementing the planned Altair system modifications and subsequently generating URLs to webpages, users will eventually have a more satisfying experience and systems will become more accessible. It is also advised by one of our subjects that more study be conducted on the Altair system because it appears to be capable of building web applications in Jupyter Notebook, a task for which we employed the Mercury framework[4].

Based on the results of our tests, the key change we would do to better our systems would be to host them on a website and distribute the URL for people to access from any computer. This would simplify the user experience and allow User Evaluation to be conducted on a larger scale, given that anybody may use our systems. This would make our systems accessible to the constructive criticism of a larger, more diversified user base. As a secondary update, we would examine Altair in greater detail and use it to its full potential in order to simplify our code and make it scalable in the future by preventing dependancy issues in our technology stack.

References

[1] Mrityunjay Pathak. 2023. IMDb top 100 movies. (January 2023). Retrieved March 27, 2023 from https://www.kaggle.com/datasets/themrityunjaypathak/imdb-top-100-movies

[2] Vega-Altair: Declarative visualization in python. Retrieved March 27, 2023 from https://altair-viz.github.io/

[3] Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. 2008. Generalized selection via interactive query relaxation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008). DOI:http://dx.doi.org/10.1145/1357054.1357203

[4] Build data web apps in Jupyter Notebook. Retrieved March 27, 2023 from https://runmercury.com/