CS5044:
Practical 02
Joint Report

150015294

200020111

Word Count: 968

The visualisation created for this project explores a large data set of movies available on various streaming platforms. This report aims to provide the reader with a background knowledge of how to use the visualisation and justify and describe design decisions. Files relating to the ideation process, such as sketches, forming the question, and data set analysis can be found in the files under the "Ideation Process" folder of the project. Further, a README.md file is provided to give an overview of all files used in the visualisation to offer an easy way to understand the code decomposition.

1. DATA & QUESTIONS

1.1 Describe the dataset and Original data sources

The data set used, provided on Kaggle by Ruchi Bhatia, contains various attributes about 16,000+ movies that are available on Hulu, Disney+, Prime Video, or Netflix [1]. A description of all attributes can be seen in *Table 1* below. Some data required cleaning before it could be passed. For example, certain movies were adjusted as their runtimes were checked and found to be incorrect.

Figure 1. Attributes Available in the Dataset

Attribute Name	Description	Notes
ID	Unique ID for all the movies in the data set	
Title	The Title of the Movie	16,744 Unique Values
Year	The year in which the movie was produced	Ranges: 1902 - 2020
Age	The Target Age Group	5 Categories: all, 7+, 13+, 16+, 18+, unknown values were replaced with "Unknown"
IMDb	IMDb Rating of the Movie	Possible Range: 0 - 10, but the data set has a max of 9.3
Rotten Tomatoes	Rotten Tomatoes Rating in Percent	Possible Range: 0 - 100, but 69% of the data was null, so we did not use this attribute in the final vis
Netflix	If the movie is on Netflix	0 if no, 1 if yes
Hulu	If the movie is on Hulu	0 if no, 1 if yes
Prime Video	If the movie is on Prime Video	0 if no, 1 if yes
Disney+	If the movie is on Disney+	0 if no, 1 if yes
Type	Movie or series	all values were movies, so all values were 0
Directors	All the Directors listed in any movie	This was not used in the final version as this attribute is less helpful in answering our main question
Genres	Genres of the Movie, stored as comma separated lists	Theres ~27 Genres Total. Null values made 3% of the visualisation and are not included if any filters are applied.
Country	Where the movie was produced	Although we wanted to consider this attribute, we felt that Language was a more important visualisation to do as its more helpful in deciding what platform to get.
Language	The languages the movie is available in, stored as a comma separated list	4% missing was replaced with "Unknown". There also exists a "None" category, which we did not want to remove in case they refer to silent movies / music videos / short films.
Runtime	The length of the movie.	4% missing was removed from runtime visualisation, but not anywhere else. There were a few high outliers (400 minutes +), so these values were checked. We found that a few of them had incorrect movie lengths, so we fixed them to match the IMDb listing when parsing the data. One TV show made it on, so this was removed as it was 15 episodes. Low values (0-5 minutes) were not removed as they could be shorts or music videos.

1.2 The Question

The question we are aiming to answer is:

Which platform should I subscribe to? (Between Netflix, Hulu, Disney+ and Prime)

We particularly liked this question as it lends itself well to other sub-questions that a user may want to explore, including (but not limited to):

- Which platform has the most movies in my native language?
- Do any platforms share a significant portion of their libraries? (Useful if a user wants to subscribe to more than 1 platform)
- Who has the most movies of my favourite genre?

We found that Netflix, Hulu, Disney and Prime have movies with significant differences in rating, genres, ages and so on. Therefore, we wanted to make a visualization that could allow users to choose a streaming platform best suited to their taste.

2. DESCRIPTION OF VISUALISATION

2.1 Design Description

There were multiple other visualisations considered for this project, which have been compiled into the “Brainstorming” PDF in the Ideation Process Folder. However, to answer this question posed, we decided on the following visual analysis:

- ***Shared Titles***: Chord Diagram
- ***Genres***: Circle-Packed Donut Chart
- ***Target Age***: Bar Charts
- ***Language***: Lollipop Chart
- ***Rating and Runtime***: Side-by-Side Boxplots

Figure 2 shows the visual encoding table of all the attributes and justifications for how these variables were visualised. Evaluation of expressiveness and justification are largely based on Munzner’s explanation of expressiveness as well as Illinsky and Steele’s chapter on choosing appropriate visual encoding [2] [3].

Figure 2. Visual Encoding Table

Attribute	Views (If Seen in More Than 1)	Attribute Type	Visual Variable	Variable Details	Expressive	Justification
All Platforms (Netflix, Hulu, Disney and Prime)	All Views (Except Age Bar Charts)	Categorical	Colour	The colour of all visualisations are set by the platform it is representing.	Yes	As these attributes are shown in every view, we utilised the same colour scheme as much as possible to keep the visualisations consistent. Colour is considered a good attribute for categorical data.
	Overview		Colour, Size	The arc colour represents the platform, and the chord colour represents the other platform they share the proportion of titles with.	Yes	In this case, the gradient utilises colours of the shared platform under the source platofrms arc to visualise how much of their collection they share with other platforms. This is particularly useful as it creates an easy visualisation to quickly understand how titles are shared with one or more other platforms.
	Ages		Position / Placement	The location on the x axis (stacked and grouped) and y axis (stacked only) indicate the platform the bar chart belongs to.	Yes	Position and Placement are considered good attributes for categorical data. Therefore, grouping in 2 different views allows for the benefits of both graphs to improve the quality of the visualisation.
Count of Occurences for Platform (derived attribute)	All Views except Language	Quantitative	Size / Area	The area / size of the chord diagram, donut chart, bar chart, and box plots	Yes	By counting the number of times a platform occurs in a given data set, we are able to derive a quantitative value. Area and size (for example, the area of a donut chart) are considered good attributes for quantitative data.
	Languages View		Position / Placement	The position on the y axis represents the count of movies in the language view	Yes	Position and placement are considered good attributes for quantiative data. By utilising this lollipop chart, we are able to show the categorical data in a quantiative way.
Genre	Genre (And Filter)	Categorical	Size / Area, Angle	The size of the donut chart represents the count of that genre. The angle of the pie chart represents how many movies exist by platform for that genre	Maybe	This attribute is classed as possibly effective as size is not considered a traditional effective method for showcasing categorical data. However, as this visualisation focuses on the breakdown by genre rather than genres in comparison to each other, the size acts more as a way to visualise an extra attribute. It further assists the visualising by limiting the amount of information in a single view, allowing the viewer to zoom in to see less-popular genres.
Age	Age (And Filter)	Ordinal / Quantitative	Size / Length	The size / height of the bar chart represents the total number of movies for that age group.	Yes	Age can be considered both an Ordinal and Quantitiative measurement. This is because there is an implied order (with all representing 0+, and the remaining categories being numerical). However, there also exists "Unknown" values, and therefore the data cannot be considered only quantitiative. As such, a bar char chart was chosen as it represents a good comparison in two views to easily see the breakdown of age compared to other ages and compared to its value in other platforms.
IMDb Rating	IMDb Rating (And Filter)	Quantitative	Position (Y) and Size	Visualised in side by side box plots agains the y axis. The size of a point represents the count of outliers with that rating.	Yes	We decided that IMDb rating was a better alternative than using the rotten tomatoes rating. This is because 69% of the data for Rotten Tomatoes are null values. With IMDb being a quantitative value, with a range between 0-10, position on the y axis is a fitting encoding method for visualising this data.
Runtime	Runtime (And Filter)					Like IMDb rating, position on the y axis is a fitting visualisation method for this quantiative variable. Moreover, the size of the outliers allows the visualisation to not be overhwleming with too much data. This follows the overview -> zoom -> more details method, as described by Schneiderman [4].
Language	Language (And Filter)	Categorical	Position (Y and X Axis)	The position on the y axis represents which language the data is showing, the position on the x axis represents the count of that language	Yes	Since this categorical data has so many categories, we decided to go with a lollipop diagram. This is because it is a visualisation that allows for easy paging without misleading or obscuring the view.
Year	N/A (Filter)	Quantitative	NA	NA	Maybe	Although year could be an interesting visualisation, we believed we should allow people to filter by year without its own visualisaiton. By filtering by year, users can still get the same experience, learning more about movies within a specific year range rather than comparing years against each other.

2.2 Introduction to web layout

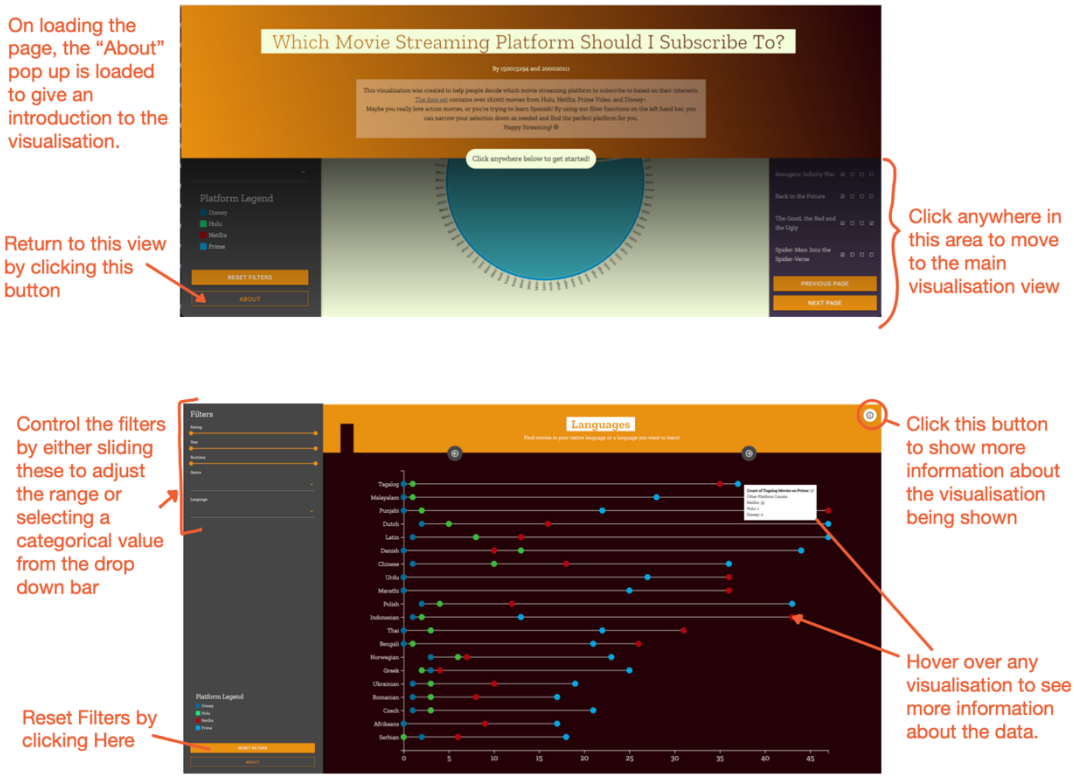
We wanted to make this visualisation look more like an engaging, interactive info-graphic rather than a traditional visualisation dashboard. We thought this approach would better fit the data set and the question, as our main goal was to assist users who were trying to decide between platforms. As such, we approached the design as a typical website.

2.3 Interaction zones

On the left side of the web page is an interactive zone where users can manually configure their movie preferences. Filtering is applied to all charts. On the right is the display zone, consisting of 5 “pages”, which show a total of 6 visualisations. Users can immediately get new, updated visualisations by interacting with the filter.

There are two kinds of interactive methods: multi-selector boxes (categorical data) and slider bars (quantitative data). *Figure 3* visualises the basic information on how to interact with the visualisation.

Figure 3. Introduction to Interaction Instructions



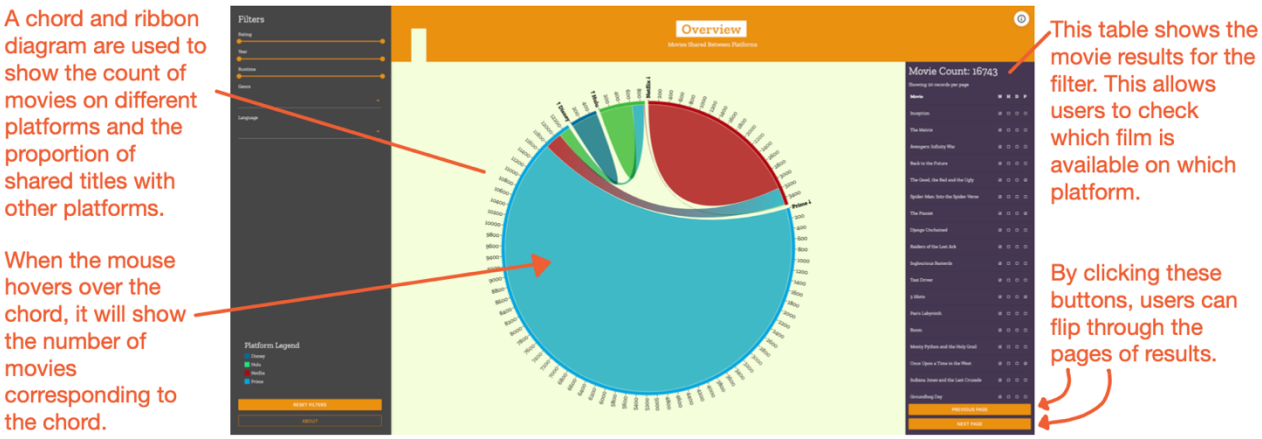
2.4 Display Zones

Figure 3. Introduction to Interaction Instructions

display zone more fit to their preferences. This section describes the different views and how to interact with them. Descriptions of each visualisation are provided in Figures *4* – *8*.

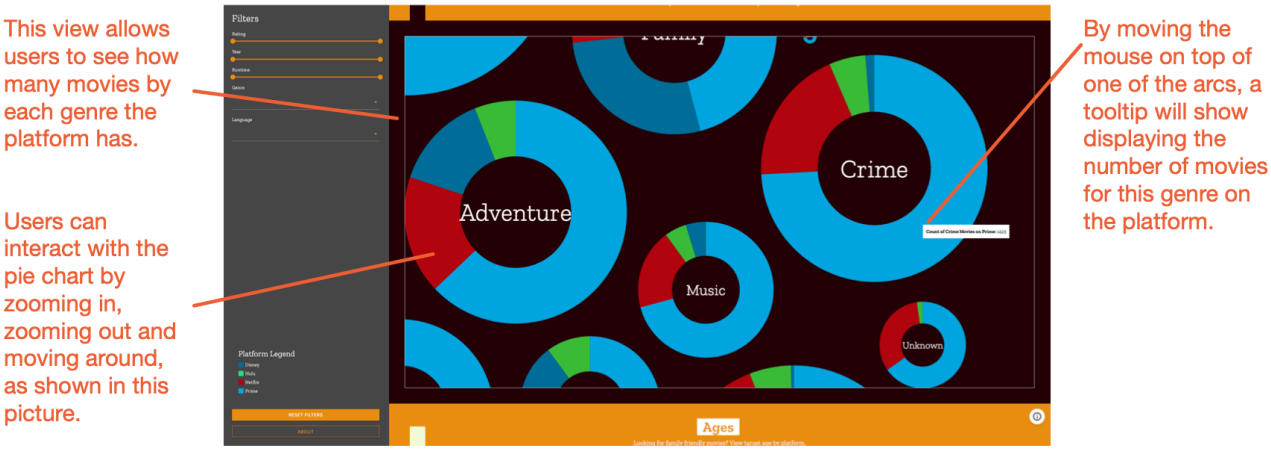
2.4.1 Overview: Shared Titled Chord Diagram

Figure 4. Overview



2.4.2 Genres Analysis: Circle-Pack Donut (Pie) Chart Pie

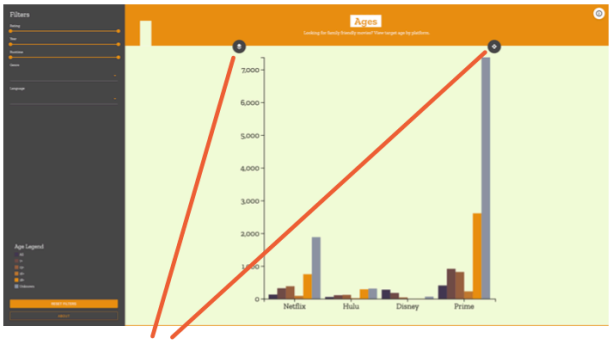
Figure 5. Genres

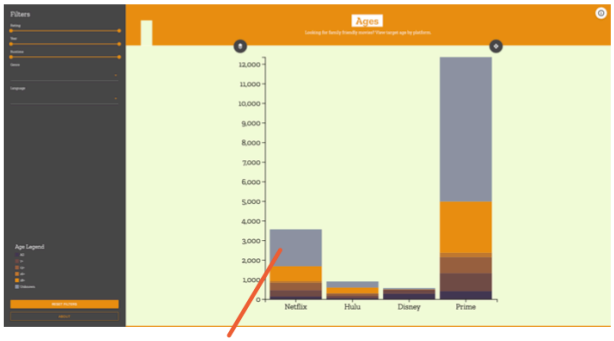


2.4.3 Target Age Analysis: Bar chart

Figure 6. Age

These bar charts show the breakdown of target-age for all the movies on each platform.





These two buttons at the top of the chart toggle the different views, between stacked and grouped bar charts.

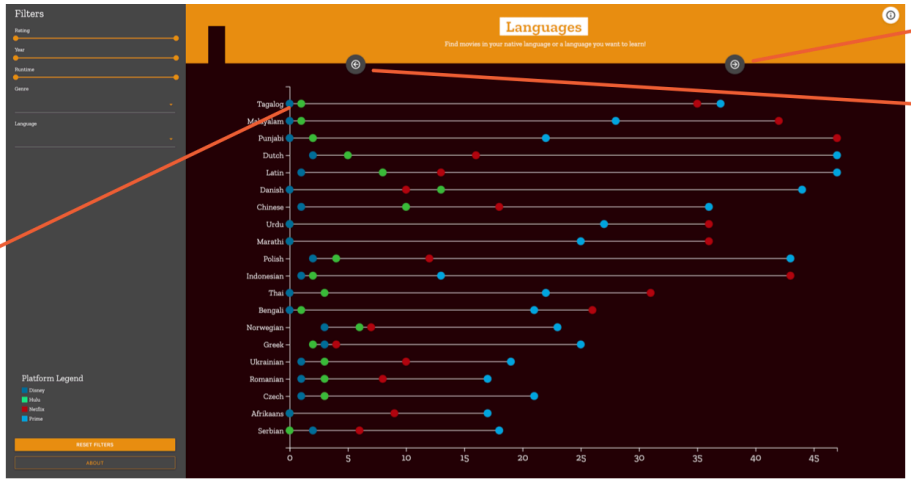
By hovering over any element, more specific information is displayed.

2.4.4 Language Analysis: Lollipop chart

Figure 7. Language

View the lollipop chart to see the distribution of languages by platform.

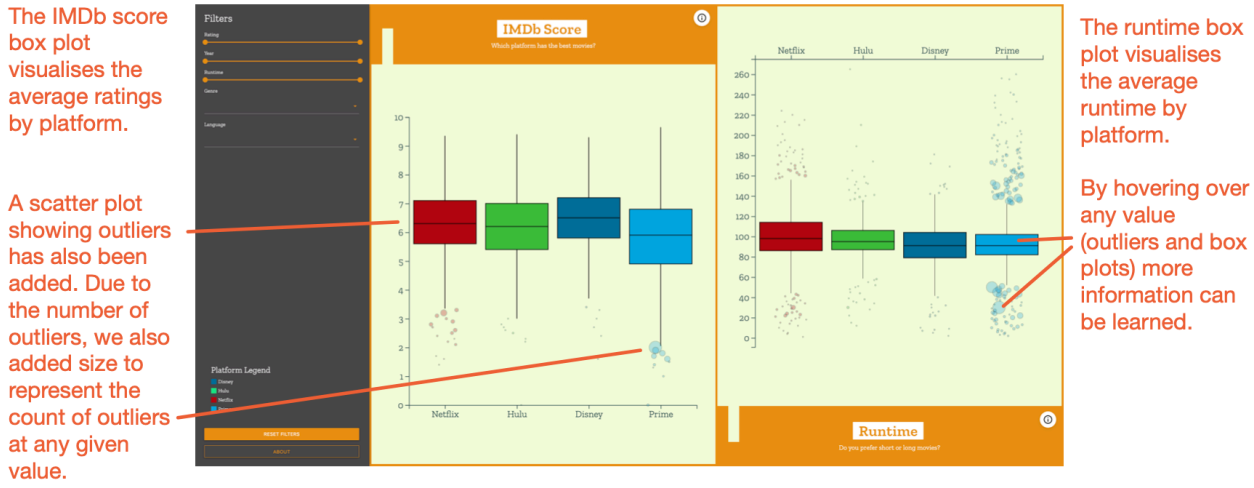
Hover over any of these elements for more information.



These two buttons at the top of the chart allow users to flip through the pages of languages. These were added as there are over 180 + languages in the data set and that was too much information to show on one page.

2.4.5 Running and Rating Analysis: Boxplot

Figure 8. Rating and Runtime



3. IMPLEMENTATION & TOOLS USED

Various examples helped us to code the visualisation, and the links are provided in the code wherever it was inspired from another source. However, we also utilised various libraries to improve the visualisation and functions of website. *Figure 9* lists libraries imported for the project.

Figure 9. Libraries, Frameworks, and Tools Used

Library	Details	Usage	Source
D3.js	d3 version 6	Used to create all visualisations	[5]
noUISlider Library	minified version from Materialize	Utilised to create multi-sliders for filtering qualitative data	[6]
noUI Materialize CSS	Materialize CSS	Used to make the noUI slider nicer	[7]
jQuery	minified jQuery 3.6.0	Used for selecting the results of the multi-select drop down option boxes	[8]
Materialize Framework	minified CSS and JS	Used to create the multi-selct drop down options	[9]
Gradient CSS Generator	online tool	Used to create gradient for the "About" Section	[10]
Chroma.js Color Palette Helper	online tool	Used to create a color-blind friendly palette for sequential data in Age visualisation	[11]
Google Fonts	Zilla Slab (Highlight, 400, 600, and 700)	Used for fonts throughout all the pages	[12]
Material Design Icons	Online Tool	Used for SVG icons for buttons	[13]

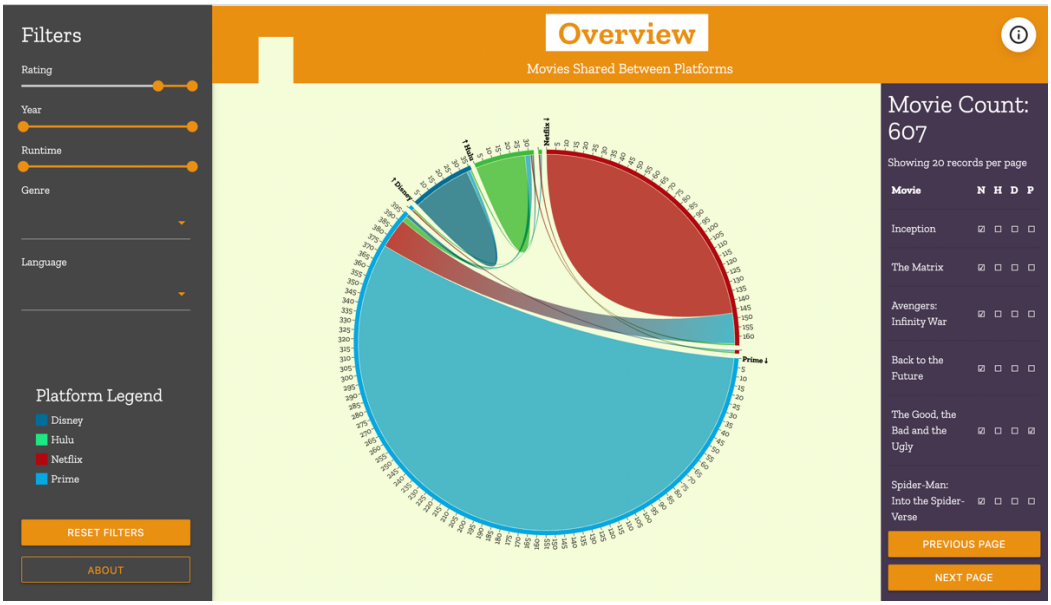
Each visualisation is done in its own JavaScript file, under the scripts/ folder. Further, the README.md file in the project folder describes all the files, code, functions, and variables.

4. INSIGHTS & CRITICAL DISCUSSION

4.1 Briefly describe the insights people can gather from your visualisation.

This visualisation aimed to answer questions that people may ask themselves when deciding which streaming platform to subscribe to. This was achieved through the various visualisations created for the views. *Figures 10 – 15* visualise examples of questions that users may ask themselves and showcases how they may arrive at these insights.

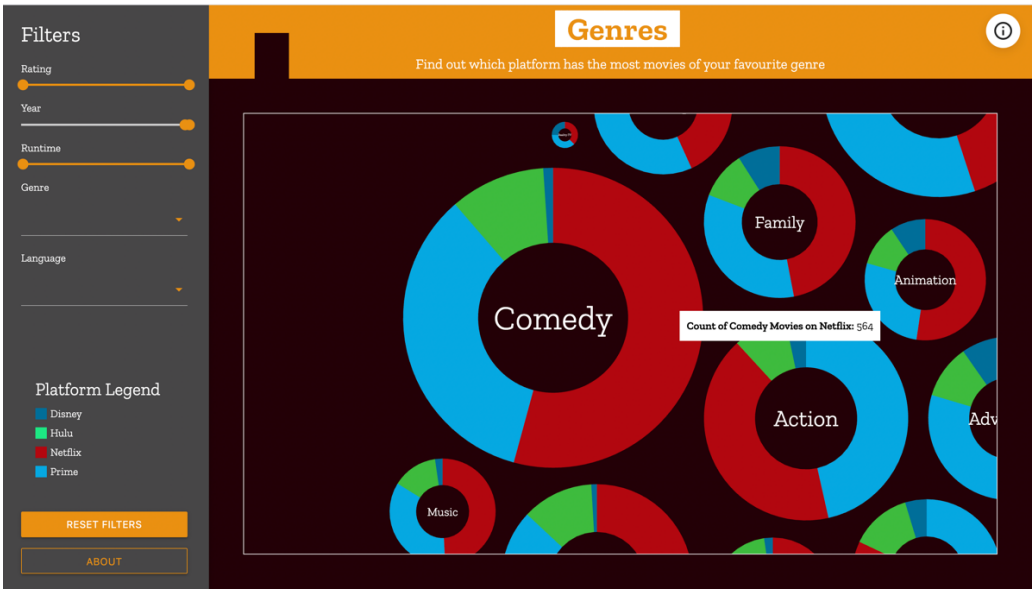
Figure 10. Overview Question Example



Q: Who has the most movies with an IMDb Score 8+, and how many of these titles are shared?

A: Prime has the most movies, with just over 370 movies being Prime exclusive. Approximately 20 movies are shared between Netflix and Prime. Notably one movie is shared between Hulu, Netflix and Prime.

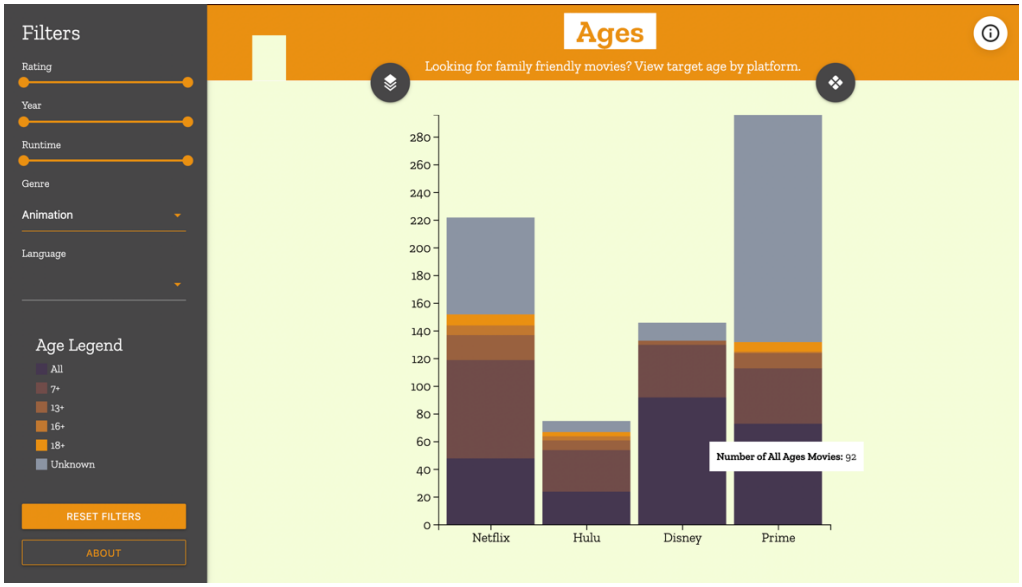
Figure 11. Genre Question Example



Q: My favourite genre is comedy, but I only like more recent films. Who has the most comedy movies from 2017 onwards?

A: Netflix has the most comedy movies, with 564 titles compared to 11 Disney comedy movies.

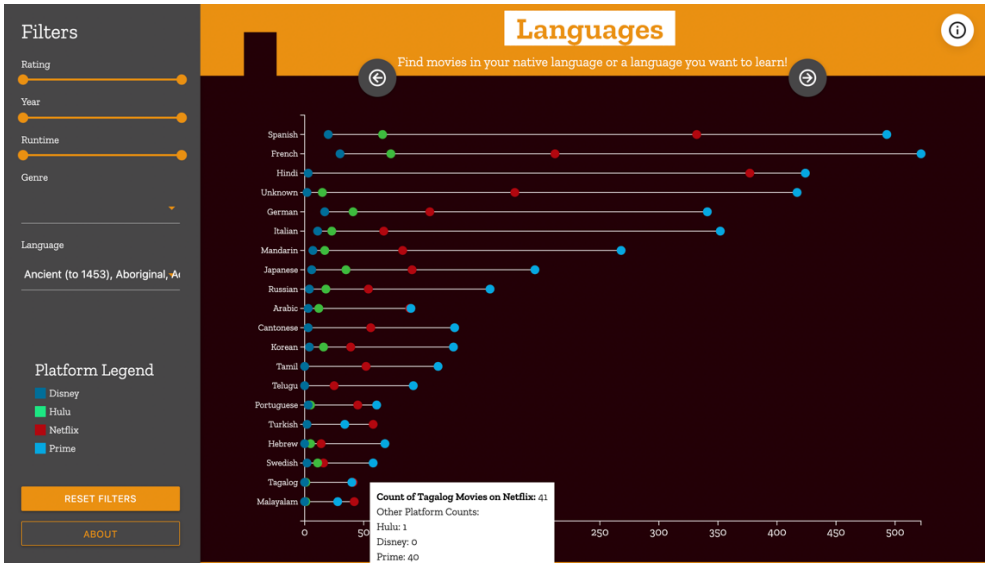
Figure 12. Ages Question Example



Q: I will be sharing my Netflix with my family and children, so who has the most "All-Age" animated movies?

A: Disney has the most All Age Titles, with 92 compared to Hulu's 24.

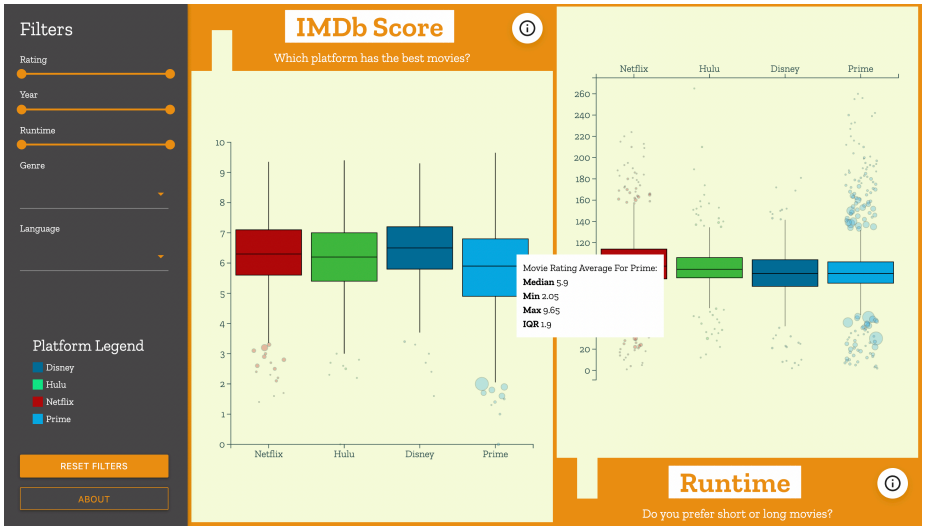
Figure 13. Language Question Example



Q: I will be sharing my subscription with my Lola, whose native language is Tagalog. Which Platform has the most movies in Tagalog?

A: Netflix has the most, with 41 movies in Tagalog. However, Prime is very close as well, with 40 movies in Tagalog.

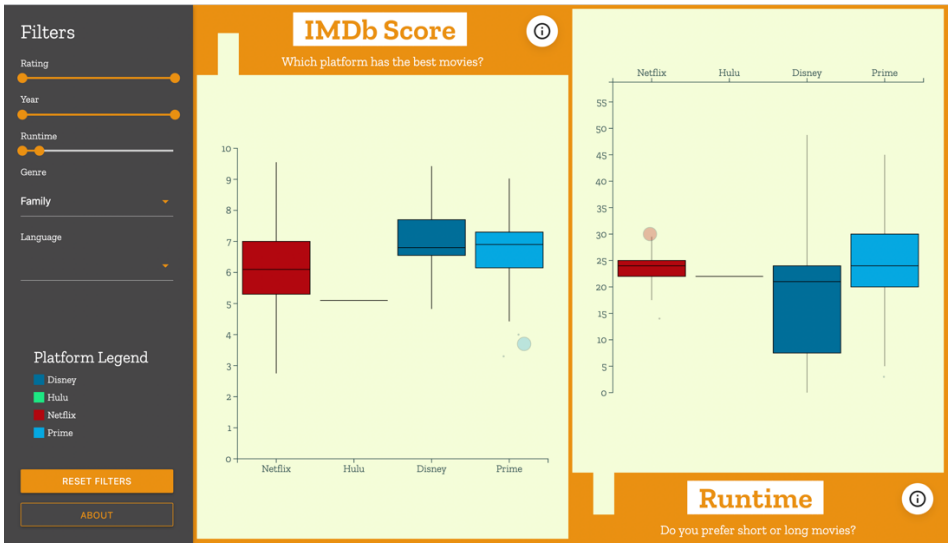
Figure 14. IMDb Score Question Example



Q: Wow, Prime has a lot of titles. But do they have the best movies?

A: Actually, no. Prime has the lowest average, with a median IMDb rating of 5.9 compared to Disney's median of 6.5. Also, Prime has the most low-rated movies, as shown by the larger outliers.

Figure 15. Runtime Question Example



Q: My kids can only watch 30 minutes of tv at a time, so I need family friendly films shorter than this. Which platform has the largest range under this time?

A: Disney+ offers the most titles in this category, compared to 1 Hulu title.

4.2 Discuss the limitations of the visualisation

Although the design was built to be responsive to changes in window size, this could still be improved. We were able to resolve a large proportion of issues in Safari however some features still contained issues due to browser specific bugs. For example, at certain heights / on mobile there are a few issues, as shown in *Figures 16 and 17.*

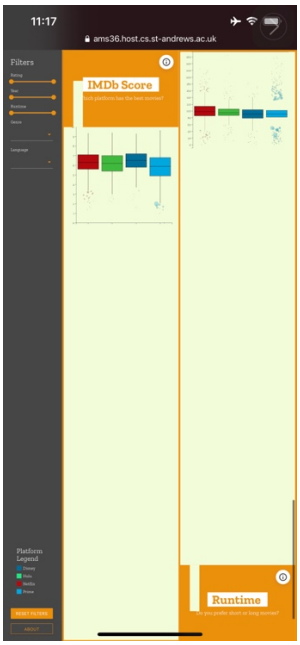


Figure 16. These do not center on Safari iOS

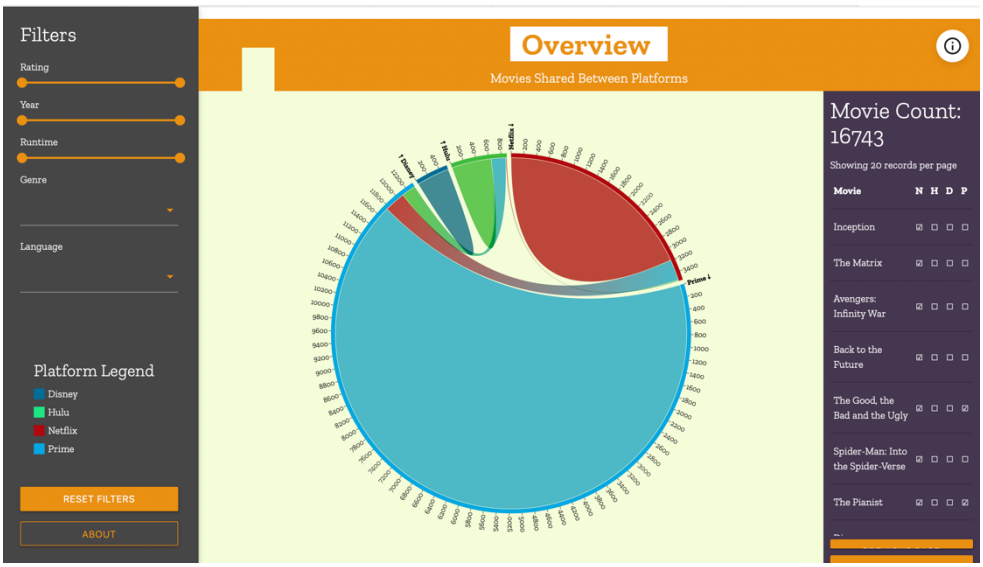


Figure 17. Paging Buttons get cut off and extra white space is added on Safari

Further, this visualisation could also be improved by connecting a separate data set that would allow the user to learn more about the movie titles specifically. If this was implemented, the visualisation could have served a dual purpose, also assisting people to choose which movie to watch based on their preferences and subscription platforms.

Finally, this visualisation could be improved by applying an ordering function and search bar to the movie list on the overview page. With the amount of data here, it would take ~800 clicks to get to the end of the list. As such, ordering functions and search bars could prove incredibly helpful here.

Links to Hosted Version:

150015294: https://ams36.host.cs.st-andrews.ac.uk/practical2/index.html

200020111: https://ym49.host.cs.st-andrews.ac.uk/D3/CS5044_P2/index.html

Works Cited

[1] R. Bhatia, "Movies on Netflix, Prime Video, Hulu and DIsney+," 2020. [Online]. Available: https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney. [Accessed 30 March 2021].

[2] T. Munzner, Visualization Analysis & Design, Boca Raton: CRC Press / Taylor & Francis Group, 2015.

[3] N. Iliinksy and J. Steele, "Chapter 4. Choose Appropriate Visual Encodings," in *Designing Data Visualizations*, ISBN: 9781449312282, O'Reilly Media, Inc., 2011.

[4] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proceedings 1996 IEEE Symposium on Visual Languages,* pp. 336-343, 1996. doi: 10.1109/VL.1996.545307.

[5] M. Bostock, "D3: Data-Driven Documents," [Online]. Available: https://d3js.org. [Accessed April 2021].

[6] L. Gersen, "noUISlider," [Online]. Available: https://github.com/leongersen/noUiSlider. [Accessed April 2021].

[7] Materialize, "Range," [Online]. Available: https://materializecss.com/range.html. [Accessed April 2021].

[8] OpenJS Foundation, "jQuery: Write Less, Do More," [Online]. Available: https://jquery.com. [Accessed April 2021].

[9] Materialize, "Select," [Online]. Available: https://materializecss.com/select.html. [Accessed April 2021].

[10] "CSS Gradient," [Online]. Available: https://cssgradient.io. [Accessed April 2021].

[11] G. Aisch, "Chroma.js Color Palette Helper," [Online]. Available: https://gka.github.io/palettes/#/9|s|00429d,96ffea,ffffe0|ffffe0,ff005e,93003a|1|1. [Accessed April 2021].

[12] "Google Fonts," Google, [Online]. Available: https://fonts.google.com. [Accessed April 2021].

[13] A. Andrews, "Material Design Icons," [Online]. Available: https://materialdesignicons.com. [Accessed April 2021].