FIT5147
VISUALISATION PROJECT

IMDb Analysis

Abstract

This project is an extension of the Exploration project done on IMDb (Internet Movie Database) – online portal for checking out different media elements such as Movies, TV series, Videos, Video Games etc., along with fans and critic reviews and ratings. The aim of this visualization project is creating an interactive webpage using R-Shiny that helps the intended audience to explore the evolution of the media industry over the last century. To help with this, three sections (with subsections) were included in the webpages with narrative text. The approach followed was deep driven approach where the contents are drilled down as the user moves through the sections.

Submitted by

Raguram Ramakrishnasamy Dhandapani
rram0013@student.monash.edu
30151325

Table of Contents

List of Tables & Figures

1. INTRODUCTION

The project focuses on the **Internet Movie Database (IMDb)**, a repository of all media in the entertainment industry such as movies, tv series, videos, video games etc. The analysis carried out in the exploration project answered the below key questions (to list a few)

- At the overall level, which genre tops the entertainment industry across the years and across certain media types?
- How ratings for media are influenced by the Release year, Run time (selected media types)?
- How connected are the cast and crew for the movies?

Now in the Visualisation project, these explored questions are visually reflected in a webpage with interactive visualisations and concrete information. The media which are considered for this project are Movies, TV series and Video Games. Some of the key findings to be communicated to the audience are as below:

- The trend of different genre of media released across the last 60 years
- The top-rated Movies/TV series/Video Games across the different years
- Looking at the cast and crew of the Movies and inter connections between directors

The audience considered for this project are the movie buffs and media critics. The reason for this kind of audience is that, the webpages show the different genre of media across years with respective top-rated ones. Every movie buff would be interested to look at the media industry for good content and top performers. At the same time the media critics (mainly Movie) would like to see the trend along with the network of directors for successful combinations.

2. DESIGN METHODOLOGY

2.1 Five Sheet Design

The five-design sheet methodology was used to create the designs of the visualizations and narrative feature of the webpages. From the sheet 1 to sheet 5, the process was followed streamlined and the final visual elements and the narration that need to be shown on the webpage were finalised. See APPENDIX for the FIVE sheets.

In this first sheet, different ideas were put forth based on the exploration project results. From all the ideas brainstormed, 5-6 of the key ideas were filtered and then combined into 3 sections.

In the second, third and fourth sheets, different designs of the final webpages were put forth with at least half of the visualisation techniques varying between them. Different kinds of charts from the initial design such as bar charts, line charts with the simple ones to start with to complex ones such as network plots, word clouds were included in the design sheets,

In the Final sheet, the design of the webpage was finalised with 3 sections complying with the Drill down narrative approach. The below table shows the different sections in the final sheet design and the proposed visual implementation of them.

Sections	Content	Design Plan
Section - 1	Trend of Genres across years	A line chart for the trend with colour differentiating the genres
	Number of releases per year and Genres in them	A bar chart with height showing the total and word cloud for Genres
Section 2	List of top-rated media based on filters	A data table to show the list
	Relation between the ratings and votes for selected media	A heatmap to show the relation between the parameters

Section 3	Network plot between the director and actors of selected movie	A network plot with dynamic interactions
	To show the average rating of director, actor and crew	Showing the ratings for Cast and crew with a bar chart
	IMDb webpage of the movie	Embedded Webpage for the related cast/crew/movie

Table 01 – Five Sheet design - 5th Sheet

2.2 Design Elements

The webpages are designed with the idea that the audience will understand the media industry that has evolved over the last century. Below is the design method and narration implemented:

- Presentation
  - Cascaded tabs were used for implementation
  - Tabs can be selected from the navigation bar

- Narrative approach
  - The approach used in the design was **Drill Down approach (FUNNEL type)**
  - The data was presented at the overall level and then drilled down to the inner elements

3. IMPLEMENTATION

After the design was finalised in the 5-sheet design, the next step is the implementation of those pages. But implementation had its own challenges since everything designed cannot implement. There will be cases where the implementation medium does not support for certain items and some of the items may be redundant with the existing items.

3.1 Design Implementation and modifications

The webpage is implemented using R and R-Shiny. Data Wrangling and cleaning was done using R and cleaned datasets were used in R-Shiny for the webpage development. Below are the set of libraries used for implementation.

Library	Purpose
DPLYR	Data manipulation (modifying data, joining datasets)
TIDYR	Data cleaning actions
GGPLOT2	Creating variety of plots
WORDCLOUD	Creating the word clouds in the report
RCOLORBREWER	To Add colours to word cloud
SHINY	To create interactive webpage
TM	Text manipulations
STRINGR	String Manipulation
PLOTLY	Interactive Charts
NETWORKD3	Implementing Network diagrams
DT	To create interactive Data tables

Table 02 – Libraries used in R for webpage implementation

Compared to initial proposed plan, some modifications were done in the implemented version. These changes were made in the design during implementation due to design constraints in the tool and redundancy/complicated contents in them for the audience. All these modification comments are provided in the below table against the initial design.

Sections	Content	Initial Design	Implementation Modifications
Section - 1	Trend of Genres across years	A line chart for the trend with colours to identify the genres	None
	Number of releases per year and Genres in them	A bar chart with height showing the total and word cloud for Genres	None
Section 2	List of top-rated media based on filters	A data table to show the list	A BAR chart was used to view the 20 top-rated items and a data table was used as a subset for the chart
	Relation between the ratings and votes for selected media	A heatmap to show the relation between the parameters - ***Not implemented***	*This was discarded as this does not add much value for the audience*
Section 3	Network plot between the director and actors of selected movie	A network plot with dynamic interactions	Instead of relation in a movie, relation between two directors were analysed
	A bar chart to show the average rating of director, actor and crew	Showing the ratings for Cast - ***Not implemented***	*This was discarded as this does not add much value for the audience*
	IMDb webpage of the movie	Embedded Webpage for the related information - ***Implemented in a different way***	Instead of an embedded webpage, a link to the webpage was provided based on click interaction in network plot
	Movies of the two directors	***New entity***	A bar chart of the movies from the two directors
	Word cloud Data Table	***New entity***	From the bar chart showing the movies of the two directors, word clouds (showing genres of movie for each director) and cast of the movie were included

Table 03 – Implementation design with modifications

3.2 Implementation decisions for the Visualisation

As per the design and modifications mentioned above, the visualisation for the webpages were created. The approach followed for the webpage creation was **Drill Down** (Funnel approach) where the overall information was first conveyed and then a certain part of it was drilled down and analysed (as shown in Fig 01).

In the design,

- First page shows the trends of the different genres for different media across years
  (has interactive elements to look at certain media type, select genres and options to look genres for any year)

- Second page is the drill down to look at the top-rated media in the industry for each media
  (has multiple filters to choose based on genres, media, release year etc and an interactive bar chart that refreshes a data table to look at items with same rating in the filtered list)

- Third page is used to deep dive into the Movie industry and look at the relation between the Directors, Movies and Cast
  (interactive network plot and an interactive bar chart to look at the cast of movies by the directors selected)

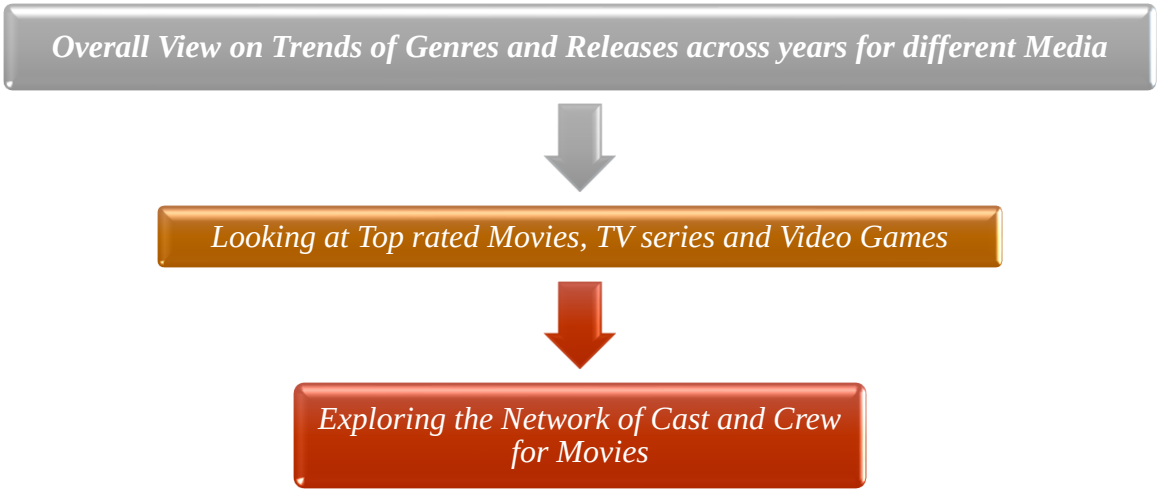


Figure 01 – The Funnel Approach used in the narration of Visualisations

The reason for choosing this method is because of the audience the webpage is designed for. Since the audience are Movie buffs and media critics, giving them an overall view in the first page and allowing them to interact with the deep driven data in the subsequent pages will be the appropriate approach to go with.

3.3 Choosing Visual representations

The next main task is to choose the right set of charts/graphs and visual techniques that help the viewers understand the content effectively. Changes made during implementation were done on basis of this. Different charts/graphs and visual elements chosen for the webpages present the content in the best possible way that makes it easier for the user to interpret the message.

Basic charts (Bar charts, pie charts) were chosen to represent the initial data and messages, so that as the user progresses through the webpages, they will be able to understand the contents presented with complex charts (Network plots, word clouds). This is also important for creating interactions between the charts in the same pages as the user will understand the changes happening in the page.

For example, in section-03 to show the list of movies based on directors, which involved sorting and colouring a **BAR chart,** where length of the bar was the Rating and colour of the bar was Director, was chosen instead of a *sorted table* or *a word cloud with title the size of rating*, as this was visually clearer to the user and indicates the message as implied.

In the similar way, a lot of charts and interactions were chosen so that the user gets a clear message with ease of interactions in the webpage.

3.4 Implementation Challenges

While implementing the above design and pages, there were quite some challenges faced. Below are the list of challenges and the steps to overcome those

- **The Dataset sizes**
  For this project, 4 datasets were considered same as the exploration project and the total dataset size was around 2GB in total. Handling them in total would cripple the webpages and to sort this out, the Data wrangling/Checking done in the exploration project was used and then the datasets were split into multiple small files and were used in each individual section of the webpages as per the need

- **Building Interactions**
  This task was the most crucial part of the webpages. Apart from the filters and select options included, interactions among the charts and providing information in between the charts was one of the challenging tasks carried out. To implement this GGPLOTLY (part of the Plotly) package was used and chart constructed in GGPLOT were added interactions based on selection in the charts using GGPLOTLY. On-

click data based on clicks was used to subset the data for the other charts through codes (instead of build in plotly functionalities)

- **Building Network Plot**
  The network plot in section 03 of the webpage (showing the relation between cast/crew) was built using the NETWORKD3 package which has a very clean and interactive output but needs a lot of prework that need to be done. Data for these charts were prepared dynamically in the code based on the user selection and then the chart was updated accordingly. Along with building this plot, interactions were added on top of it using JavaScript

4. USER GUIDE

This section of the report is used to explain all the charts/graphs and visual elements presented in the webpages. The focus in each of the table is on the *Visual Element*, its *Purpose*, *image as in webpage*, *Interactions* present in it and *a small explanation* of what the element does. Each section below has the information of the elements contained in them.

4.1 Section 01 – Genre Analysis

The **First** section introduces the audience to the webpages and has different visualisations that show the trend of the Genres in the industry. Below is the explanation of different visual elements included in this section along with their purpose.

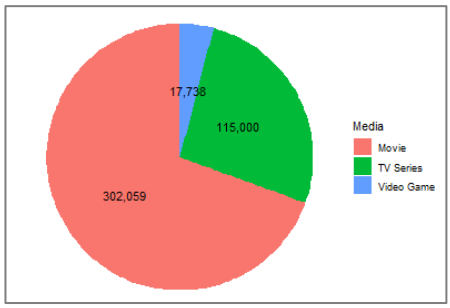
Visual Element	**PIE Chart**
Purpose	To show the split of total releases across different media during past 60 years
Visualisation	
Interactivity	This is a static chart with no interaction. This is part of the introduction text
Filters	No Filters
Interpretation	This chart shows that Movies have dominated the media followed by the TV series and then by Video Games

Table 04 – Section 01 Visual element 01

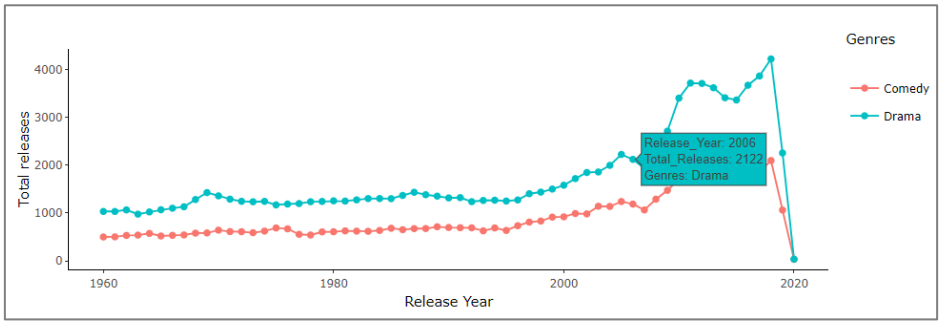
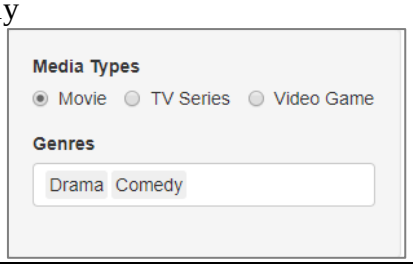
Visual Element	**LINE Chart**
Purpose	To show the trend of Genres in the past 60 years with the number of releases
Visualisation	
Interactivity	1. 'On hover' information that shows the Release Year, number of Releases and the genre 2. Updates itself based on Filter selection
Filters	The side panel on the left contains option for changing the Media Type or Genres and chart will update accordingly 
Interpretation	The chart helps us understand the trend of the genres for the three media types and we would also be able to compare the different trends across years

Table 05 – Section 01 Visual element 02

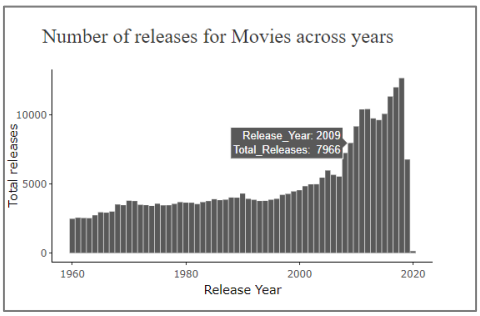
Visual Element	**COLUMN Chart**
Purpose	To show the total releases per year in the past 60 years
Visualisation	
Interactivity	1. 'On hover' information that shows the Release Year, number of Releases 2. Updates itself based on Filter selection 3. 'On-click' action updates the two graphs to the right of it
Filters	The side panel on the left contains option for changing the Media Type and chart will update accordingly
Interpretation	The chart helps us understand the number of releases that has happened over the years in the selected media

Table 06 – Section 01 Visual element 03

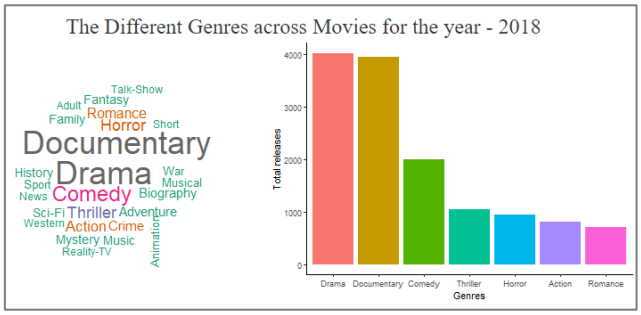
Visual Element	**WORD CLOUD and COLUMN chart**
Purpose	To show the genres across a selected year
Visualisation	
Interactivity	Updates itself based on ‘On-click’ action of the chart to its left
Filters	The chart reflecting the number of releases in each media across years
Interpretation	This gives an idea on the peak genre during any year for the selected media

Table 07 – Section 01 Visual element 04

4.2 Section 02 – Top Rated

The Second section focuses on the Top-rated media in the industry. There are a series of filters that allows the user to streamline the search according to their interest. Below is the explanation of different visual elements included in this section along with their purpose.

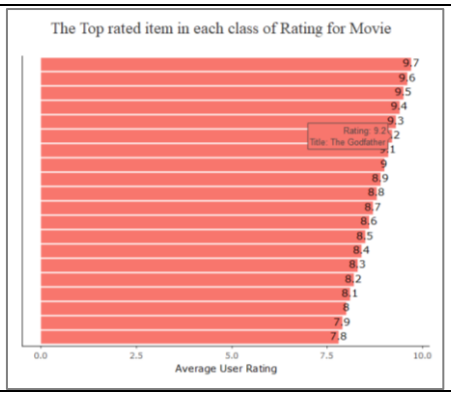
Visual Element	**BAR chart**
Purpose	To show Top 20 Ratings in the list of media selected with the top ones in each of them
Visualisation	
Interactivity	1. ‘On hover’ information that shows the Rating of that column and the name of media 2. Updates itself based on Filter selection 3. ‘On-click’ action updates the table to the right of it
Filters	Updated based on the multiple filters at the top of the Page
Interpretation	Each column is the top item based on number of votes for that rating in the data based on filters and shows top 20 ratings available

Table 08 – Section 02 Visual element 01

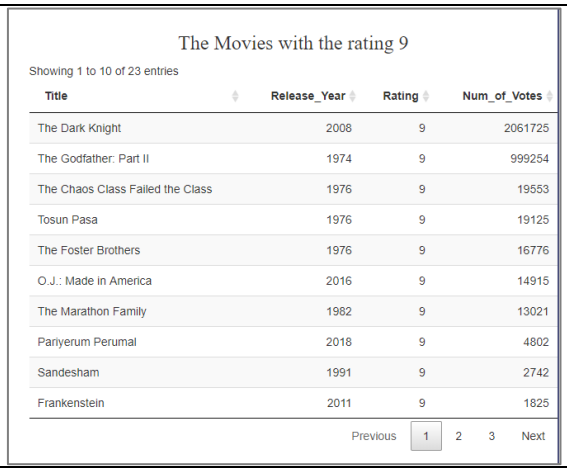
Visual Element	**DATA TABLE**
Purpose	Shows all items with the selected Rating in the dataset
Visualisation	
Interactivity	1. Column sort and page selectors to navigate in the table
Filters	Updates itself based on the selection on the chart to its right and Filter selection
Interpretation	Contains all the items of the selected media with the Rating as per the selection in the chart to its right sorted based on Number of votes You can navigate across the table and sort it based on any of the columns

Table 09 – Section 02 Visual element 02

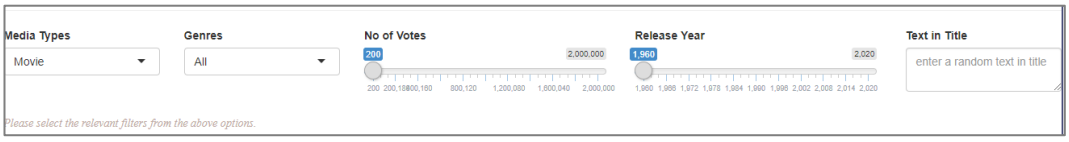
Filter Element	**Multiple Filters**
Purpose	Allows the user to filter the media content for viewing
Image	
Interactivity	Drop down list, Slider input and Random text input for user selection
Interpretation	Based on the selection made in the above filters, the data for the charts and tables in the page gets altered

Table 10 – Section 02 Filter element 01

4.3 Section 03 – Exploring Cast

The Third section focused mainly on the Movie industry. Here we have the options to look at the network between two directors and understand their genres and crew. Below is the explanation of different visual elements included in this section along with their purpose.

Visual Element	**NETWORK DIAGRAM**
Purpose	Shows the network of 2 Directors, their Movies and its Cast with links in between them

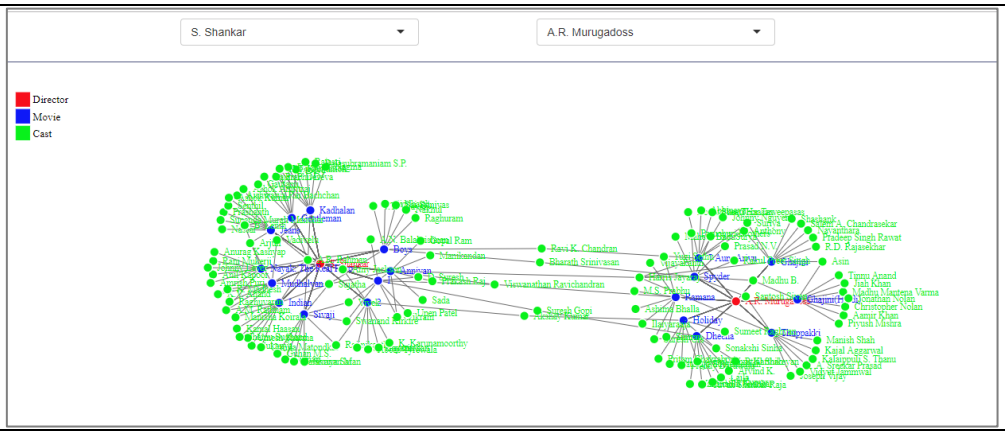
Visualisation	
Interactivity	1. Click and drag the nodes to get a clear view of the network 2. 'On-click' interaction in the nodes that will take you directly to the IMDb webpage of the concerned node 3. 'On-Hover' functionality on each item to highlight that node and its connection
Filters	Updates itself based on the selection of the two directors on Top of it
Interpretation	The network is drawn between two directors where the Directors are at the centre connected to their movies and finally to the cast of the movies If the directors, movies have any cast in common they will be shown once and linked to corresponding factors

Table 11 – Section 03 Visual element 01

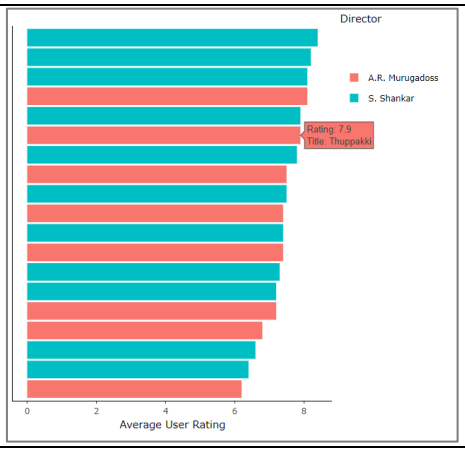
Visual Element	**BAR CHART**
Purpose	Lists the movies directed by the 2 directors in order of Rating
Visualisation	
Interactivity	1.'On-click' interaction in the bars that will display the cast and crew of the movie with their details 2. 'On-Hover' functionality on each row that shows the movie rating and name
Filters	Selection of the two directors on Top of page
Interpretation	The chart shows the movies of the two directors sorted by Rating and coloured according to the director. Clicking on any movie will display the cast and crew of the movie

Table 12 – Section 03 Visual element 02

Visual Element	**WORD CLOUD**
Purpose	Variety of genres handled by the directors
Visualisation	
Interactivity	1. No interaction in the word clouds 2. Updates itself based on the selection of the two directors on Top of page
Filters	Selection of the two directors on Top of page
Interpretation	The word cloud indicates the variety of genres in the movies by that director

Table 13 – Section 03 Visual element 03

Visual Element	**DATA TABLE**
Purpose	To show the cast and crew of the selected movie
Visualisation	Cast of the movie **Thuppakki** directed by **A.R. Murugadoss** released on 2012 Name | Category | Current_Age A. Sreekar Prasad | editor | NA Joseph Vijay | actor | 45 years Kajal Aggarwal | actress | 34 years Vidyut Jammwal | actor | 39 years Manish Shah | producer | NA Kalaippuli S. Thanu | producer | NA Harris Jayaraj | composer | 44 years Santosh Sivan | cinematographer | NA
Interactivity	1. No interaction in the Table 2. Updates itself based on the selection of the movie from the chart to its left 3. The title of the table also updates itself based on selection
Filters	Selection of the movie from the bar chart to its left
Interpretation	This table helps us understand the different crew members of the team for the selected movie from Bar chart

Table 14 – Section 03 Visual element 04

5. CONCLUSION

The overall intent for the project was to design webpages that will help the users get a feel of exploring the media industry and understand the trends in evolution of each media. As part of it, there were three sections that were included in the webpages that followed the Drill down approach.

In the design of the pages, a variety of visualisation elements were included to convey the intended message. The method used in the webpage design and the visual elements included will provide an easier, impactful and engaging ambience for the user to interpret the information provided.

6. REFLECTION

The project was completed using R (for Data wrangling and manipulation) and R-Shiny (for creating the webpages and implementing visual elements). During the implementation of this project, there were a lot of hurdles faced and learning that I have gained. To list a few,

- **Choosing the correct libraries that will help you deliver what you intended -** I initially chose GGPLOT, IGRAPH for creating graphs, interactions and network visualisation. But after spending some time with them and going over the materials, I realised there were much effective and easier to implement packages such as PLOTLY & NETWORKD3
- **Building a story to convey in your design –** The initial five sheet design had a few problems while implementing as the story was not proper in it and some elements were difficult to implement in the given tool. So, some changes were made to the initial design in terms of storytelling and implementation

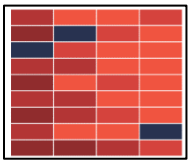
7. REFERENCES

1. http://www.imdb.com (Information courtesy of IMDb)
2. https://en.wikipedia.org/wiki/IMDb
3. https://shiny.rstudio.com/articles/tag-glossary.html
4. https://www.color-hex.com/
5. https://rdrr.io/cran/networkD3/man/networkD3-shiny.html

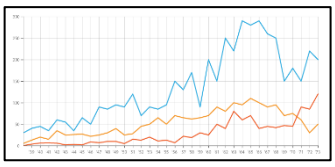
8. APPENDIX – 5 sheet design

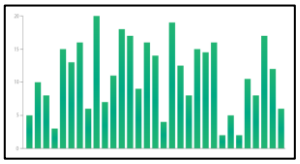
SHEET – 1 – Ideas sheet

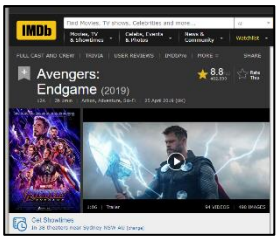
1. IDEAS:



- Plotting the genres across years in a line chart with trends
- Deep diving into specific media types with a word cloud/bar chart and trend of genres
- Showing the trend of genres with respect to each media type
- Showing the scale of ratings, runtime, num of votes for the selected media types through a heatmap plot
- Showing top-rated movies, TV series, Video Games based on ratings and votes (table format) in specific genres
- Displaying leading crew members of the selected media
- Viewing the related webpage in IMDB website for selected movies that will show in the same window
- Setting up a network plot to visualise the cast and crew
- Plotting the age of the crew for selected media







2. FILTER:

Considering the initial ideas, below ideas are filtered to next stage

- Visualising the trends of top genres across selected media type through bar/line charts and word cloud
- Showing the scale of ratings, runtime, num of votes for selected media type in violin plot
- Displaying leading crew members of the selected media
- Displaying IMDB webpage of media
- Table view for the top-rated movies/TV series with respect to num of votes
- Showing the network plot of the cast and crew for selected crew member
- Showing the age factor of the leading cast and crew of the selected media

3. CATEGORIZE:

Grouping the filtered ideas together

- Visualising the trends of top genres across selected media type through bar/line charts and word cloud
- Showing the scale of ratings, runtime, num of votes for selected media type in violin plot
- Table view for the top-rated movies/TV series with respect to num of votes
- Displaying leading crew members of the selected media
- Displaying IMDB webpage of media
- Showing the network plot of the cast and crew for selected crew member
- Showing the age factor of the leading cast and crew of the selected media

4. COMBINE AND REFINE:

Based on the grouping created above, below are the combined and refined ideas

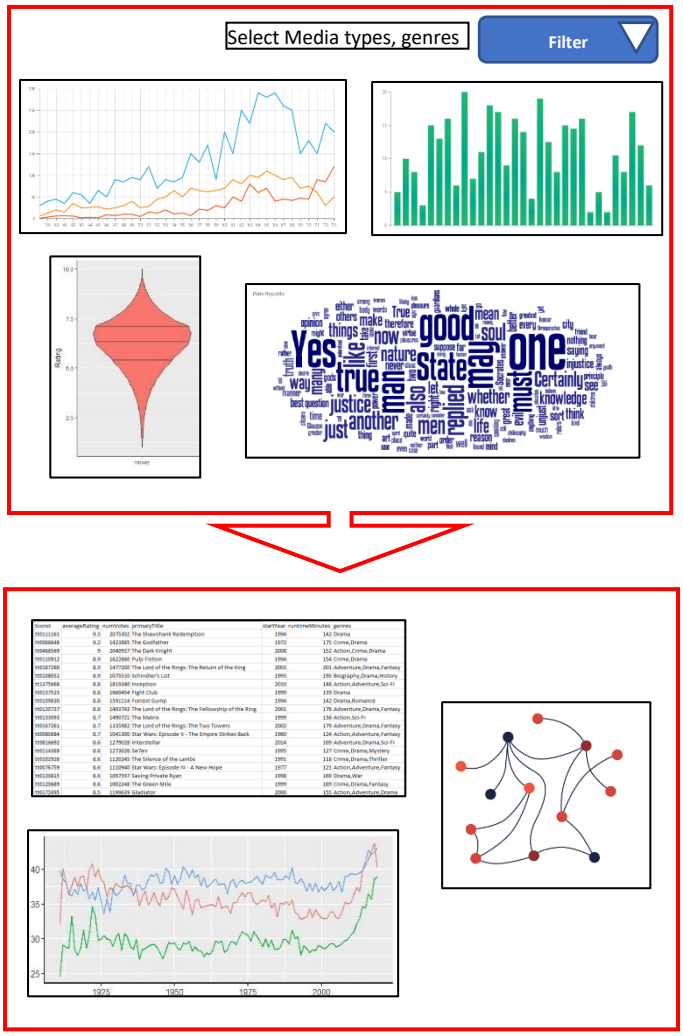
- Visualising the trends of top genres across selected media type through bar/line charts and using a violin plot to see the distribution of the data for ratings, runtime, num of votes for the same media
- Using the table view for the top-rated movies/TV series with respect to num of votes and adding in features such as genre, release year along with the filters across votes, ratings and few other factors
- Displaying leading crew members of the selected media from previous table with age and a network plot of the directors other top-rated movies
- Displaying IMDB webpage of media

5. QUESTIONS

- The solutions provided above will give a clean picture on the data and contents to be narrated along with deep dive analysis for certain cases
- There are 3 different designs that will provide a funnel approach to help the user to understand evolution of the Media industry across years
- The information provided in the solution has its accuracy based on the data from the IMDB website

SHEET – 2

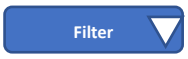
LAYOUT



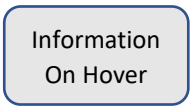
INFORMATION

Title	IMDb Analysis
Authors	Raguram (rram0013-30151325)
Sheet number	**Sheet-2**
Task	
Date	30-May-2019

OPERATIONS



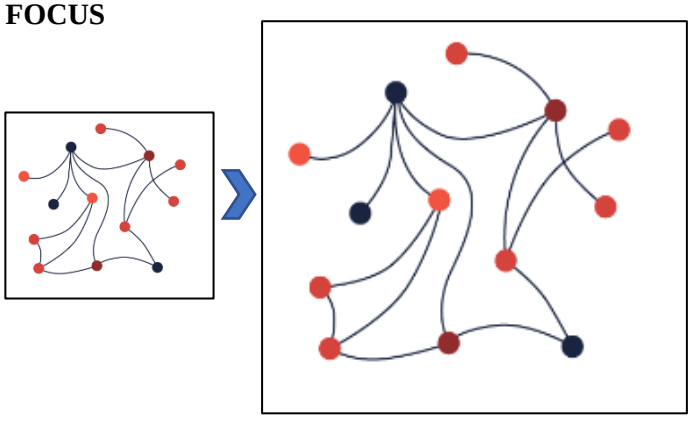
Filter option
This filter operation will be used to select the required media type

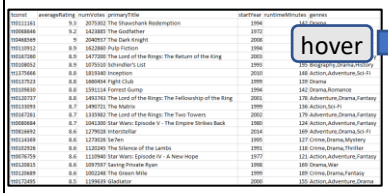


Hover Option
This shows some data on the cast when hovered over the table

Click option
When clicked over the director in the Table shows the network of director and his movies

FOCUS





On hover:
Leading cast, director, their ages

DISCUSSION

Page-1 detail:
Trend of Genres across years,
number of releases per year and word cloud of genres of selected media,
ratings in a ribbon plot for selected genre and media

Page-2 detail:
The list of top-rated movies in form of table
network plot between the director and actors of selected movie
Average age of actors, actress and directors across years

Advantages:

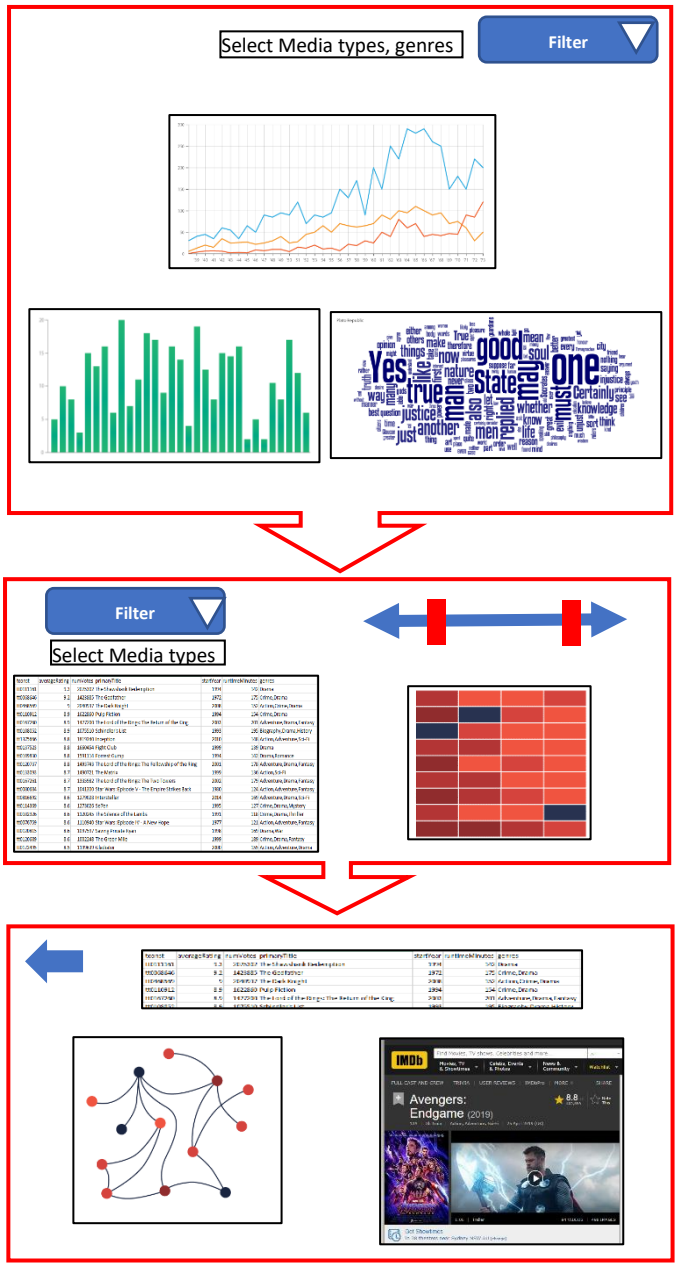
- Clear picture of trends and released across years
- Ratings analysis clearly
- Top rated movies and visual link between cast of movie and other crew members

Disadvantages:

- Too many contents in a single page
- Limited interactions for the user

SHEET – 3

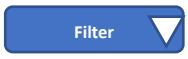
LAYOUT



INFORMATION

Title	IMDb Analysis
Authors	Raguram (rram0013-30151325)
Sheet number	**Sheet-3**
Task	
Date	30-May-2019

OPERATIONS



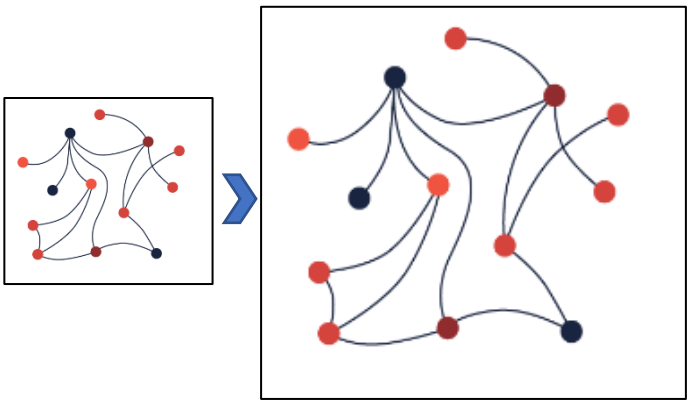
Filter option
This filter operation will be used to select the required media type

Slider Option
This is used to select the time frame for which the tables will be refreshed

Click option
When clicked over the movie or tv series in the Table shows the network of director with his movies/tv series and it's the IMDB webpage

Page Navigation
Used to navigate between the two last 2 pages where we will be seeing specific data

FOCUS



DISCUSSION

Page-1 detail:
Trend of Genres across years,
number of releases per year and word cloud of genres of selected media

Page-2 detail:
The list of top-rated media in form of table
Heatmap between the ratings and votes received in the media

Page-3 detail:
Network plot between the director and actors of selected media
IMDB Webpage of the movie selected in Realtime

Advantages:

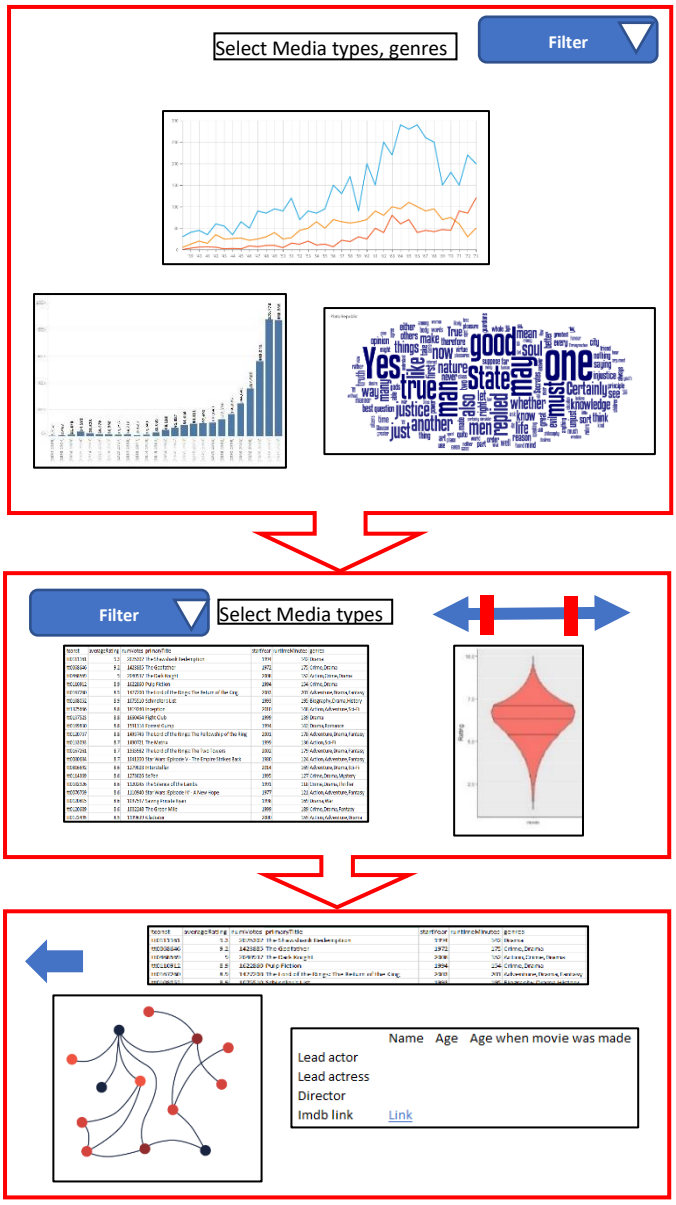
- Clear picture of trends and releases across years
- Individual movie analysis

Disadvantages:

- Contents to be covered are to less across individual pages (if combined are too much in a single page)

SHEET – 4

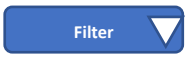
LAYOUT



INFORMATION

Title	IMDb Analysis
Authors	Raguram (rram0013-30151325)
Sheet number	**Sheet-4**
Task	
Date	30-May-2019

OPERATIONS



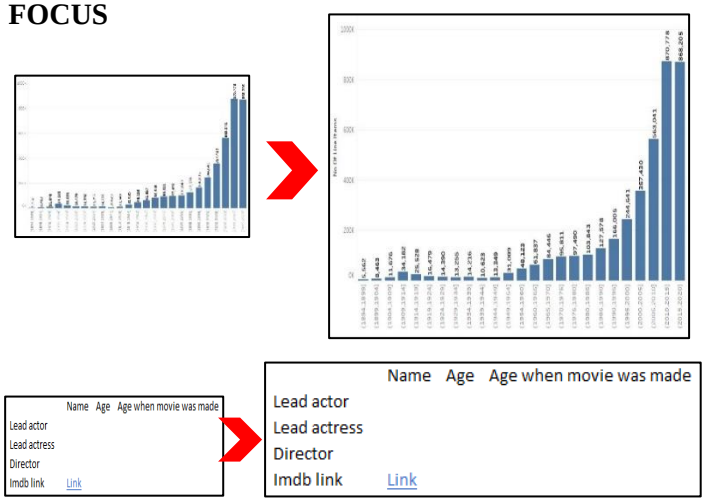
Filter option
This filter operation will be used to select the required media type

Click option
When clicked over the movie or tv series in the Table shows the network of director with his movies/tv series and crew details of that movie

Page Navigation
Used to navigate between the two last 2 pages where we will be seeing specific data

Slider Option
This is used to select the time frame for which the tables will be refreshed

FOCUS



DISCUSSION

Page-1 detail:
Trend of Genres across years,
number of releases per year and word cloud of genres of selected media

Page-2 detail:
The list of top-rated media in form of table
Ribbon plot of Ratings of the selected media

Page-3 detail:
Network plot between the director and actors of selected media
A table showing the details of the media/movie

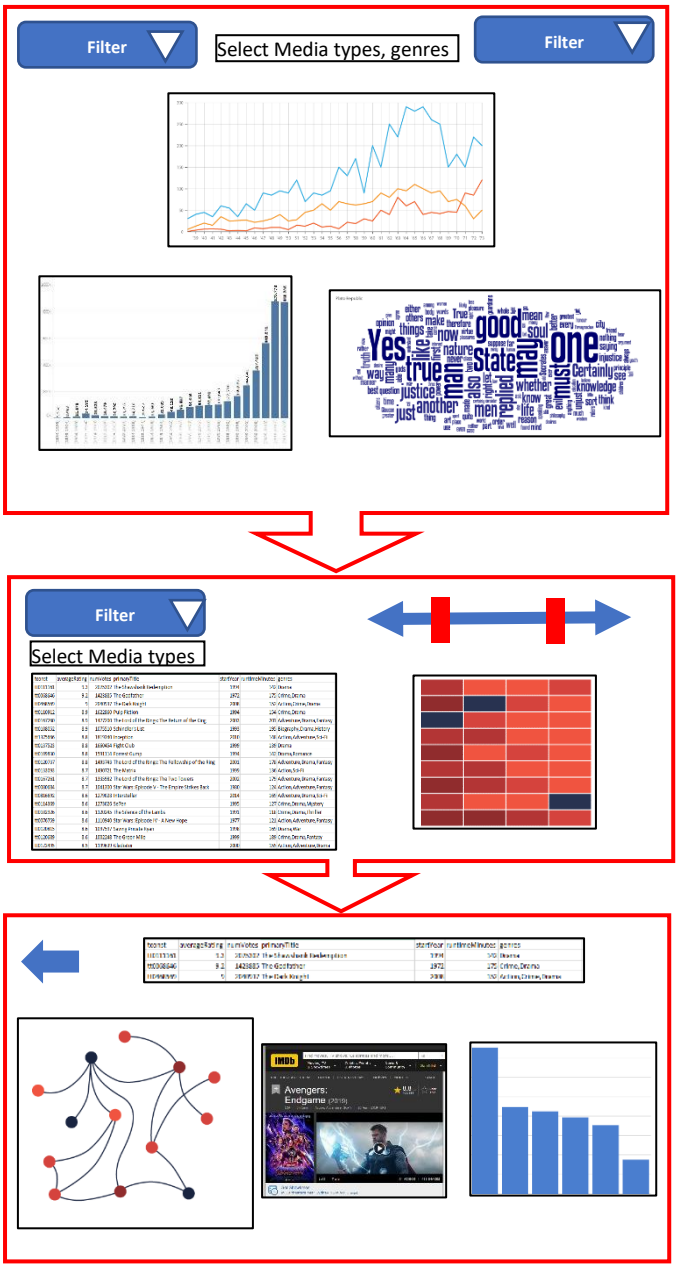
Advantages:

- Clear picture of trends and releases across years
- Individual movie analysis

Disadvantages:

- Even though the content is higher than Sheet 3 design, some of the pages are underfilled with data/graphs/interactions

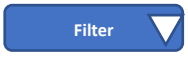
LAYOUT



INFORMATION

Title	IMDb Analysis
Authors	Raguram (rram0013-30151325)
Sheet number	**Sheet-5**
Task	
Date	30-May-2019

OPERATIONS

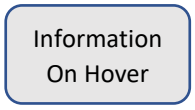


Filter option
This filter operation will be used to select the required media type and respective genres

Click option
When clicked over the movie or tv series in the Table shows detailed view of that movie with the IMDB webpage and directors other top-rated movies

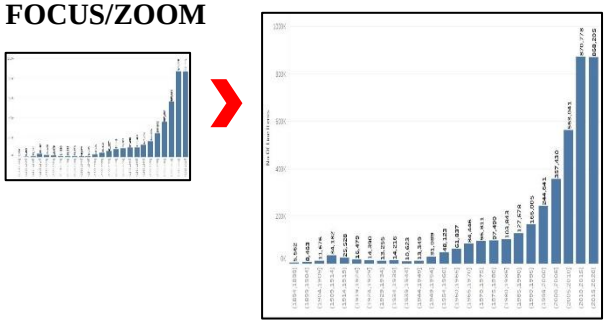
Page Navigation
Used to navigate between the two last 2 pages where we will be seeing specific data

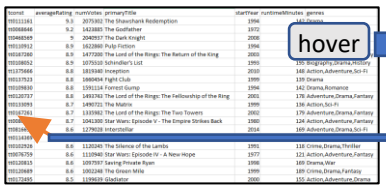
Slider Option
This is used to select the time frame for which the tables will be refreshed



Hover Option
This shows some data on the cast when hovered over the table

FOCUS/ZOOM





Leading cast, director, their ages

Clicking will take the user to the next page for movie specific and director details

DETAILS

- The visualisation is implemented used R Shiny and the data manipulations for this is done using R
- For data manipulation, R requires libraries such as dplyr, tidyverse and in along with the Shiny library other libraries are used for visualisation

Page-1 detail:
Trend of Genres across years,
Number of releases per year and word cloud of genres of selected media

Page-2 detail:
The list of top-rated media in form of table
Heat-map between the ratings and num of votes for selected media

Page-3 detail:
Network plot between the director and actors of selected movie
A bar chart to show the average rating of director, actor and crew
IMDb webpage of the movie