ASSIGNMENT 2

COMP5048-VISUAL ANALYTICS

Group 10
Semester 2, 2020

THE UNIVERSITY OF SYDNEY

Assignment Coversheet – GROUP ASSIGNMENT

Please fill in your details below. Use one form for each group assignment.

Personal Details of Students

Group Name/Number					
Family Name	Given Name (s)	Student Number (SID)	Unikey	Contribution + percentage	Signature
Bele	**Utkarsh**	**490590751**	**Ubel0548**	Task 7 + Meeting Minutes for Week 8 (15%)	Utkarsh
Lobo	**Roland**	**490548682**	**Lrol0801**	Task 6 + Meeting Minutes for Week 9 (15%)	Roland
Lobo	**Wilson**	**490548707**	**Wlob3430**	Task 1 + Meeting Minutes for Week 10 (15%)	Wilson
Saif	**Mohammed**	**490278701**	**Moha6885**	Task 4 +Task 5 + Conclusion + Meeting Minutes for Week 6 + IMDB data integration (20%)	M Saif
Sajnani	**Ishan**	**490198281**	**Isaj7533**	Task 3 + Introduction + Aims and Contribution + Meeting Minutes for Week 7+ Data pre-processing (20%)	Ishan
Tonape	**Shwetali**	**490174344**	**Ston2220**	Task 2 + Meeting Minutes for Week 11 (15%)	Shwetali

Assignment Details:

Assignment Title		Netflix Content Analysis from 2008 to 2020			
Assignment number		**Assignment 2**			
Unit of Study Tutor		**Seokhee Hong**			
Group or Tutorial ID		**Assignment 2 Group- 10**			
Due Date	**19/11/2020**	Submission Date	**18/11/2020**	Word Count	**11101**

Declaration:

1. I understand that all forms of plagiarism and unauthorised collusion are regarded as academic dishonesty by the university, resulting in penalties including failure of the unit of study and possible disciplinary action.
2. I have completed the **Academic Honesty Education Module** on Canvas.
3. I understand that failure to comply with the Academic Dishonesty and Plagiarism in Coursework Policy can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the ***University of Sydney By-Law 1999*** (as amended).
4. This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.
5. The assessment has not been submitted previously for assessment in this or any other unit, or another institution.
6. I acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment may:
    a. Reproduce this assignment and provide a copy to another member of the school; and/or
    b. Use similarity detection software (which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking).
7. I have retained a duplicate copy of the assignment.

Please type in your group number here to acknowledge this declaration:	**Assignment 2 Group- 10**

Table of Contents

1. INTRODUCTION

1.1 Dataset and Task

The dataset that we have chosen for our assignment work is “Netflix Movies and TV Shows”. The dataset provides information about the content available on Netflix, at a country level from 2008 to 2020. This information about the content available ranges from its casting to the maturity rating of the same.

Pre-processing of the dataset was done in order to remove outliers and bring the data into the right shape and at the right level with an aim to get better visualization and get efficient results. The general data pre-processing included:

- Splitting the “date_added” column in the dataset into “year_added” and “month_added”, using python (the code for the same has been attached in the **source code zip folder**).
- In the country column we had TV Shows and Movies being released in multiple countries and to illustrate the same in the dataset they were separated by comma in their respective rows. Using Power Query, we were able to split the comma delimited values of countries in different rows.

All the other carried-out data pre-processing exclusive to the respective task are listed below with the task description. With an aim to perform exploratory analysis on the Netflix dataset, a series of tasks have been brainstormed and designed, ranging from trend analysis on the Netflix content to the duration analysis of TV shows and Movies over the years . These tasks have been segregated to represent a clear and comprehensive visualization. These are:

- **Task 1:** Creating a map visualization to explore the trend of Netflix content.
- **Task 2:** Displaying top content producing countries.

Pre-Processing for Task 2
We split the shows and movies released in different countries into separate rows to be able to easily count the content release in each country. We also used the data from 2000 for our visualization to avoid any skewness in the visualization

- **Task 3:** Creating a visualization to illustrate the type of content available on Netflix and analyze monthly addition of content on Netflix.

Pre-Processing for Task 3
Aggregating the original dataset based on the respective years to get the count of TV Shows and Movies using Python.

- **Task 4:** Creating a visualization representing the maturity rating analysis of content available on Netflix
- **Task 5:** Creating a visualization to show find out the highest earning genres and production companies by integrating the Netflix dataset with IMDB dataset

Pre-Processing for Task 5
We performed a join operation on the Netflix dataset with IMDB movies dataset. To successfully integrate the same, we performed Inner join on the Titles column using Tableau prep.

- **Task 6**: Creating a visualization to carry out genre analysis on the content available on Netflix
- **Task 7:** Creating a visualization to represent the duration analysis of Tv shows and Movies available on Netflix.

1.2 Aims and Contribution

Aim : The dataset provides information pertaining to different TV Shows and Movies added on Netflix over the years from 2008 to 2020. The aim of our study is to make use of this data to find out patterns in the content added on Netflix globally, at country level. This may vary from analyzing the type of content uploaded on Netflix, the genre of content, average duration of the TV Shows and Movies.

Contribution: With the use of all the parameters in the dataset and creating visualization, very useful features were extracted from the same. Our team members tried visualizing the data and its underlying tasks using different analytics tools like Tableau, Tableau Prep, Power Query, Power BI and using programming languages like Python (Seaborn, Matplotlib). Evaluating all the visualizations created, the best ones which suited their task the most was selected. For the evaluation process, each team member followed either Think aloud or Cognitive walkthrough.

With the successful completion of this project, our team has gained a significant amount of knowledge on using different tools for visualization of different tasks. Also, most importantly the team has learnt more on extracting the underlying patterns, as we aimed to prove the expected and unearth the unexpected. All the team members showed great team spirit and commendable team contribution during the project.

2. TASK 1

2.1 Design

2.1.1 Analysis

The objective of this visualization is to understand the global adoption of Netflix and the major trend change in Netflix's offerings. The pre-processing required for this visualization was the splitting of the Country column (containing multiple countries separated by comma) into multiple rows with a single country entry. The analysis of the visualization is as follows:

- Netflix originated in United states and then was introduced to neighbouring countries.
  it was available in only 8 countries in year 2013, which grew to 22 in year 2015, then to 56 in 2016 and is currently present in 110 as of year 2020, as per the dataset.
- Netflix initially focused on streaming movie contents and TV shows only gained popularity after 2015. The rate at which movies are added over the years is significantly more than TV shows with and average distribution 70-30% for movies and tv shows respectively.
- Contents added after 2013, spread across various genres, this signifies Netflix attempt to cater to a wide range of audience. Currently Netflix offers 42 genres options. TV shows offer more variety of genres than movies with not much difference.

2.1.2 Visualization

The visualization for Task 1 is a dashboard consists of 5 graphs that are combined to analyse the global trend of Netflix adoption.

Graph 1: A map with countries coloured based on the content available,
Graph 2: A diverging bar graph that shows the no. of Movie(s) and TV show(s) added in a particular year.
Graph 3: A pie chart that shows the distribution of the content added based on Movie or TV show
Graph 4: A line chart that show the number of genres offer across all years
Graph 5: Table view chart that shows five columns namely total content, no. of Countries the content is available in, genre options available, no. of Movie(s) and no. of TV show(s). Selecting a country on "map" will update the other graph based on the country selected. Selecting the year in graph 2 and 4, will update the "map", "pie" chart and "table" chart to reflect data for that" year". Selecting the "type" in diverging pie graph, will update the other charts to reflect data for that "type". Dashboard also offers filters for Country, Year and Type to look at data pertaining to required fields.

2.2 Implementation

Load the pre-processed excel file as Data source in Tableau.

Graph 1 in Figure 1: Select "Longitude" and "Latitude" from "Measures" section and place it in the "columns" and "rows" section respectively. Select "Country" from the "Dimension" section and place it in "Detail" in "Marks" section. Next select "Show Id" from "Dimension" and place it in "Color" in "Marks" section then right click on it and select measure "Count (Distinct)". Click on "Colors" then "Edit Colors" and select "Red-Gold", click "Apply" and "OK". Next select "Date Added" from "Dimension" and place it in "Pages" section and make sure it takes "YEAR (Date Added)" else right click and select "Year". Now right click on "YEAR (Date Added)" and filter out the null values.

Graph 2 in Figure 1 : Select "Date Added" from "Dimension" and place it in "Column" section and make sure it takes "YEAR (Date Added)" else right click and select "Year". Select "Show Id" from "Dimension" and place it in "Rows" section then right click and select measure "Count(Distinct)" Hold ctrl key and drag "CNTD(Show Id)" and drop it on "Color" in "Marks" section. From marks section select "Bar" Chart.

Graph 3 in Figure 1: Select "Pie" chart from "Marks" section. Select "Show Id" from "Dimension" and place it in "Angle" in "Marks" section then right click and select measure "Count (Distinct)". Select "Type" from "Dimension" and place it in "Colors" in Marks section. Add "Type" and "CNTD (Show Id)" to "Label" in "Marks" section, right click on "CNTD (Show Id) and select "Quick Table Calculation" and pick "Percent of Total".

Graph 4 in Figure 1: Select "Date Added" from "Dimension" and place it in "Column" section and make sure it takes "YEAR (Date Added)" else right click and select "Year". Select "Listed In" from "Dimension" and place it in "Rows" section then right click on it and select measure "Count(Distinct)" Hold ctrl key and drag "CNTD(Listed In)" from rows and drop it on "Color" in "Marks" section. From marks section select "Line" Chart.

Graph 5 in Figure 1: Select "Show Id" from "Dimension" and place it in "Colum" section then right click on it and select measure "Count (Distinct)". Select "Country" from "Dimension" and place it in "Columns", also select measure "Count (Distinct)". Select "Listed In" from "Dimension" and place it in "Columns", also select measure "Count (Distinct)". Create a calculated field "Movie Count" with code "IIF([Type]= "Movie", [Show Id], NULL)" then duplicate these field edit to count TV Show instead. Right click on "Show Me" in the top right and select “Text Tables”. Move the “Measure Names” to “Column section”. Customize appearance and edit alias for preferred names.

Make some cosmetic changes for better appearance for every graph.

2.3 Evaluation

2.3.1. Results

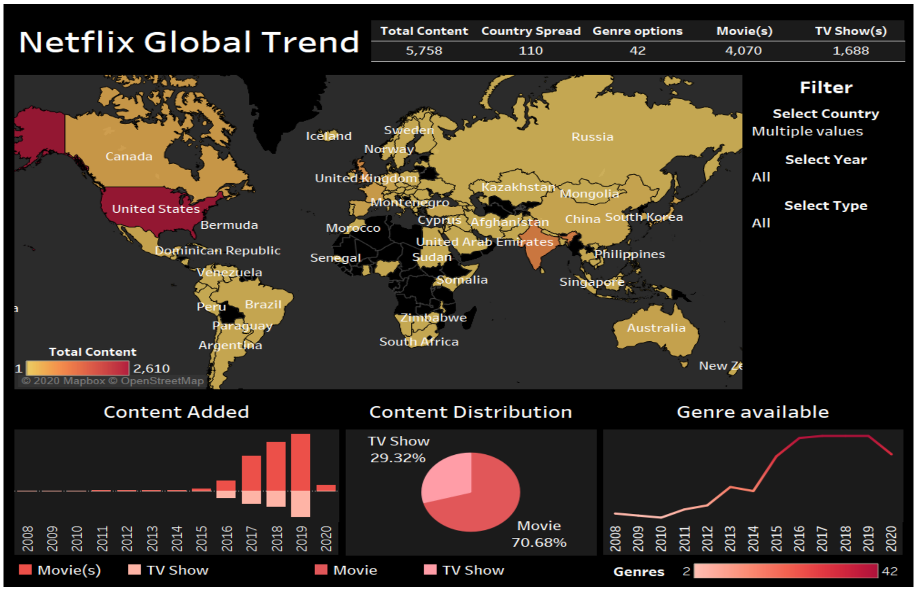


Figure 1 : Netflix Global Trend VA system

2.3.2 Discussions

For this visualization we used Think Aloud evaluation technique. This direct observation method was performed on two users which involved asking them to think out loud as they were performing a task. The user actions on this visualization pertaining to what they are looking at, what they are thinking, doing and feeling were noted down. This helped us in determining the user expectations and in identifying the aspects of the visualization that were unclear or perplexing.

The Dashboard was showcased to 2 participants addressing the following tasks:

- **Task 1 :** Tell us about the understanding of the different segments on the Dashboard
- **Task 2 :** Netflix was present in how many countries in the year of 2016
- **Task 3 :** Looking at the Map tells us the country with the highest TV Show content availability in year 2018.
- **Task 4 :** What can you say about the content available in the country of Canada in 2019.

For all the tasks as listed above, the thoughts and actions of the participants were noted down. The single ease question score for these tasks was asked to every participant and noted down. This included just one question on - How difficult or easy did participants find the task? The range of the score was set from 1 – 5, with 1 being the toughest and 5 being the easiest. Below mentioned is the observation for the same.

Single Ease Question	Participant 1	Participant 2
Task 1	5	5
Task 2	5	5
Task 3	5	4
Task 4	4	5

3. TASK 2

3.1 Design

3.1.1 Analysis

The visualization for Task 2 is a moving/ animated bar chart that shows the top 10 countries on Netflix from the year 2000 to 2020 with the highest content. To design the visualization, data was pre-processed. The countries column was split and segregated into new rows (i.e. A country column with value – India, United States, United Kingdom; got split into three different rows with values India, United States and United Kingdom respectively). This was a necessary step to have the exact count of the content in each country. The analysis of the visualization is as follows:

- United States is consistently the top content producing country (with a huge margin between the second content producing country). United States is followed by India and United Kingdom with a close competition between the two.
- Content from France and Canada was in demand from 2014, as we can see a peak in their content count. There is also huge peak in the overall number of contents from 2014, stating the gaining popularity of Netflix.
- There is a drop in the overall content in 2019 and 2020, which could be due to Covid-19 and worldwide lockdown.
- The highest content of 436 was produced in 2018 by the United States.

3.1.2 Visualization

The visualization is a moving bar chart. The still bar chart for the same are shown in Figure 1 (a), (b), (c). The bars in the bar chart represent the countries, each with a different color. The number of content (i.e. movies and tv shows) produced by each country is displayed next to each bar and the year is displayed in the center- right corner in bold.

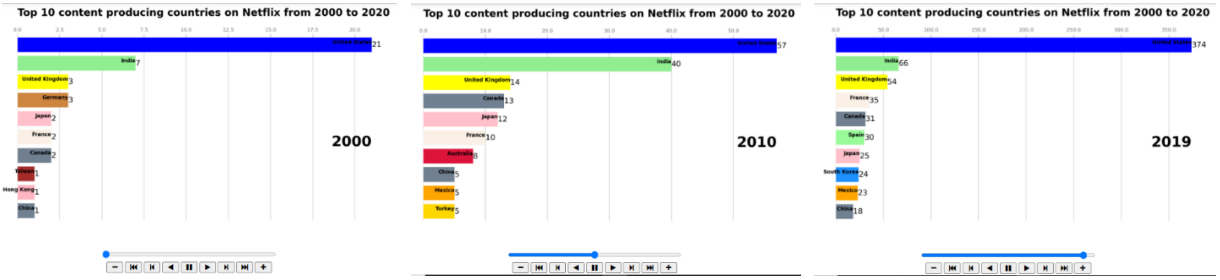
3.2 Implementation

The data visualization tool/ platform used for the visualization is python and the libraries used are MATLAB like plotting framework 'matplotlib.pyplot', tick locating and formatting configuration 'matplotlib.ticker', live animations 'matplotlib.animation', data manipulation and analysis library 'pandas' and for working with arrays 'NumPy'.

The pre-processed dataset (.csv file) was read and only selected columns (show id, country, release year) are used. The data before 2000 displayed only 1-3 countries with a few content counts. For better visualization and since the visualization displays the top 10 countries, we visualized data from 2000. The code has been added in the appendix.

3.3 Evaluation

3.3.1. Results



(a) (b) (c)

Figure 2: Still images of moving bar chart for the year 2000 (a), 2010 (b) and 2019 (c)

3.3.2 Discussions

We have applied Cognitive Walkthrough as the method for evaluation. The questions that were asked for evaluation were:

Question 1 : Whether the user could easily understand the visualization

Question 2 : Whether the visualization seemed complex.

Based on the evaluation from the three experts to whom we showed the visualization commented that the same was readable, could be easily interpreted and wasn't too complex.

4. TASK 3

4.1 Design

4.1.1 Analysis

For this task we have divided our analysis into two parts. These are :

Part 1 : For the first part of the analysis, we are aiming to analyze the trends in the addition of content available on Netflix. With this we aim to understand, whether Netflix has been focusing on TV shows or Movies over the years from 2008 to 2020. We have analyzed the same using a line chart **(Figure 3)** overviewing the trends in addition of content on Netflix over the years from 2008 to 2020. Also, we are designing two funnel graphs**(Figure 4, Figure 5)** , depicting the growth of Movies and TV-shows on Netflix, over the years since inception. Furthermore, we dive deep into the maturity rating of the content that Netflix has been focusing on, in two of the most content producing countries. The same has been represented using an area chart (**Figure 6).** In the same we have tried to contrast the growth of content of different ratings in top content producing countries.

The key takeaways extracted from the descriptive analysis performed are:

- Netflix has added most of its content during the period of 2016-2019, as shown in Figure 3. The same can be accounted for and supported with an increase in the demand for the same, in both developing and developed countries.

- In the period of 2016-2019, a total of 4013 movies and 1874 TV shows were released. In its period of inception, from 2008-2012, Netflix only added 21 movies and 4 Tv Shows. The same has been analyzed and represented in Figure 4 and Figure 5.
- In the US, we have more TV Show and movies belonging to TV-MA rating whereas in India, we have more content belonging to TV-14 rating. Further analysis on the overall maturity ratings in done in Task 4. The same can be accounted for different cultures and values in respective countries.

Part 2: With the second part of the analysis, we aim to analyze the trends in the monthly cycle for addition of content on Netflix. Holding onto the same, we intent to recommend the months, wherein the content producers can release their TV-shows and Movies to gain more viewership and reach out to more audiences. From the observations made in the first part of the analysis, we have only considered time from 2016 to 2019 as Netflix has added majority of its content during this time. The key takeaways extracted from the descriptive analysis performed are:

- In the graph, shown below in Figure 7, we can observe a steep spike in the content addition during the months of October-December consistently during the period of 2016-2019.
- As an analysis summary since, most the content is added within these three months. My advice and data-based recommendation to any new content producer would be to add content in month of January, to get a good range of viewing audience.

4.1.2 Visualization

Part 1 Visualization : The visualizations implemented for the first part of our analysis include a line chart (Figure 3) ; with a hue of content type, two funnel graphs (Figure 4, Figure 5); depicting the growth of movies and TV-shows separately over the years. As discovered in the previous task that United States and India are the highest content producing countries. With the same, we aim to understand the type of content, in terms of rating that has been added on Netflix, available in these two countries. The same has been represented using a set of area chart, being framed in a dashboard (Figure 6), allowing comparison between the two.

Part 2 Visualization : The visualization implemented for the second part of our analysis is in form of a stacked area chart (Figure 7) , wherein we hue the graph based out on the content type.

4.2 Implementation

To explain and illustrate the implementation we have taken the Figure wise approach, as we go along explaining the tools used to represent the same.

4.2.1 Tableau

Data Pre-Processing : To begin with the analysis, we did some pre-processing on the data using Power Query and Python, as we have multiple countries in one row separated by comma, so using Power Query we segregated that into different rows. Also, using Python, we added two more columns deriving from the “date_added” column in the dataset. These new columns signified “month_added” and “year_added”. The source code for the same, has been attached in the **source code zip folder.**

Figure 3: In order to implement the same, firstly, the pre-processed dataset was imported as (.csv). Once the data source is imported, we selected the year added in the columns tab and the count of show id in the rows tab. For adding the hue, type was selected as color and the line graph was selected.

Figure 6: With the analysis findings from our previous task, we have the names of top content producing countries and we aim to analyze the type of content they are focusing on these two countries. In order to achieve the same, we tried representing the same using two area charts, segregated based out on the content type. Initially the same was done separately for US and India, and later were combined in form of a dashboard.

Figure 7: With an aim to analyze the monthly trend in addition of content, we implemented a stacked area chart using Tableau. To implement the same, we choose the Year added and month added in the columns and count of distinct show ids in that respective months. Since, we are focusing on a period of years (i.e. 2016-2019), we choose year added as filter selecting those respective years.

4.2.1 Power BI

Data Pre-Processing : In order to carry out the analysis on TV shows and Movies separately, I have used Python for aggregating the data, by grouping the data based on Content type and aggregating the count of respective shows and movies in respective years. Exporting the data frame, I have used Power BI to visualize the same. The code for the data pre-processing has been attached in the **source code zip folder.**

Figure 4 and Figure 5 are funnel graphs representing the overall growth of TV shows and movies over the years. To implement the same, firstly I loaded the data using .csv file and then on two different sheets. Once the data was successfully imported, I had to change the data summarization for years, as we don't aim to summarize years. To create the funnel graph, I selected the respective chart type and added year in the group section and added respective values as aggregated count of Movies and TV shows. After successful data representation, I worked on the aesthetics of the visualization as I changed the background color to black and changed data labels to white color and furthermore increased their size for better visibility.

4.3 Evaluation

4.3.1. Results

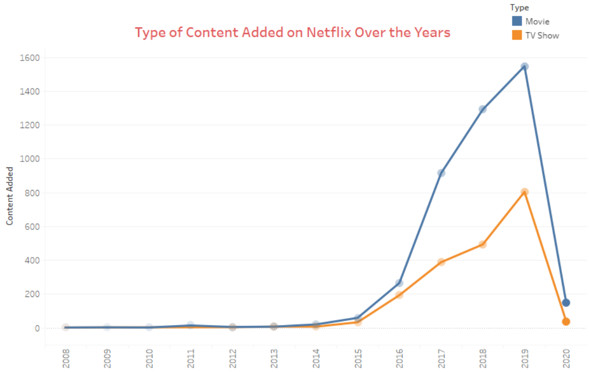


Figure 3 : Overall Growth in Content

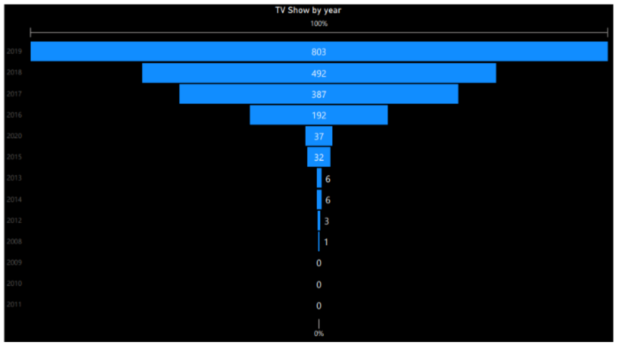
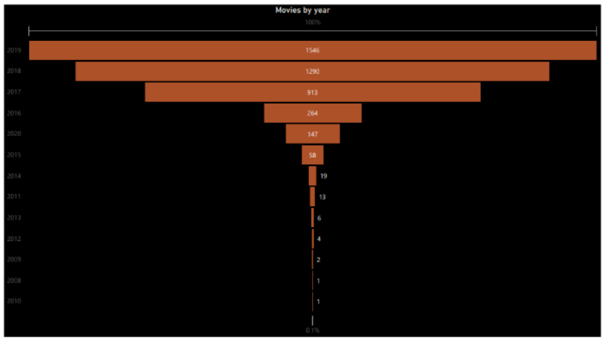


Figure 4 : Growth in TV show content



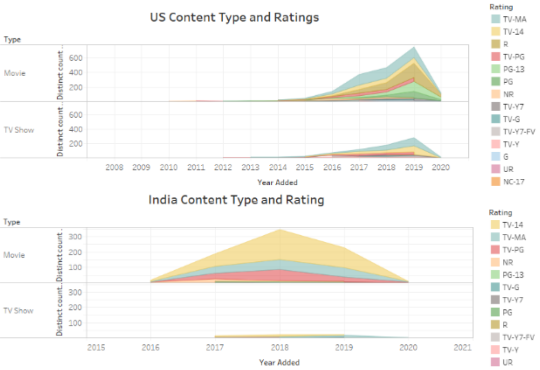


Figure 5 : Growth in Movies content

Figure 6: Content Preference Analysis

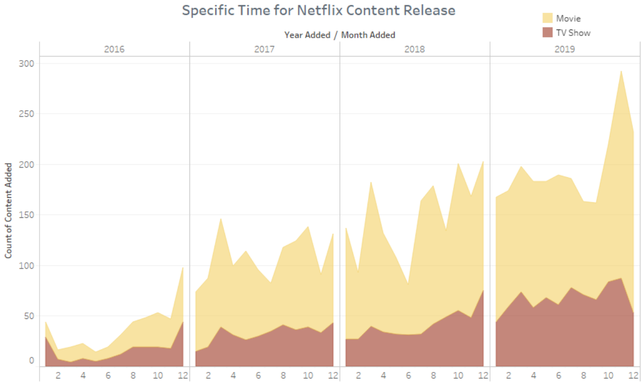


Figure 7: Monthly Trends in Content Addition

4.3.2 Discussions

For this set of tasks and visualization we used Think Aloud evaluation technique. This direct observation method was performed on two users which involved asking them to think out loud as they were performing a task. The user actions on this visualization pertaining to what they are looking at, what they are thinking, doing and feeling were noted down. This helped us in determining the user expectations and in identifying the aspects of the visualization that were unclear or perplexing.

The line charts, funnel graphs, area chart dashboard and the stacked area chart (Figure 3, Figure 4 , Figure 5, Figure 6 and Figure 7) were showcased to 2 participants addressing the following tasks:

- **Task 1:** Tell us what this visualization represents
- **Task 2:** Understand the comparison and contrast between the type of content added on Netflix.
- **Task 3 :** This task was only for Figure 7, wherein the user was asked to identify the spike in the growth of content and subsequently to identify the corresponding month as well.

For all the tasks as listed above, the thoughts and actions of the participants were noted down. The single ease question score for these tasks was asked to every participant and noted down. This included just one question on - How difficult or easy did participants find the task? The range of the score was set from 1 – 5, with 1 being the toughest and 5 being the easiest. Below mentioned is the observation for the same.

Single Ease Question	Participant 1	Participant 2
Task 1	5	5
Task 2	5	5
Task 3	5	4

5. TASK 4

5.1 Design

5.1.1 Analysis

We wanted to find out the distribution of Netflix content in terms of maturity ratings and availability across the movie and Tv Show categories. We designed the visual by taking the count of ratings as Y-axis and type of rating as X-axis. The bar graph clearly shows the type of content that is available and its quantity. Separation on type further expresses the presence of different categories of content. The separation on type was important because it is evident that both categories have dissimilar availabilities of mature content.

Maturity rating visualization indicates that most of the content – whether it's movies or TV shows, belongs to Mature category. TV-14 rating is assigned to media which needs parental guidance for viewers below 14 years of age and this type of content is 2nd most common (across both Movies and Tv shows). It is also evident that there are far higher R rated movies than R rated Tv shows. Therefore, it can be said that majority of the content available on Netflix is not suitable for children. Among the 2 categories of content – Movies and Tv Shows, it appears that there is more rating data for Movies. After analyzing the dataset, we were able to conclude that the dataset indeed has more movies and fewer TV shows, thereby validating our conclusion.

5.1.2 Visualization

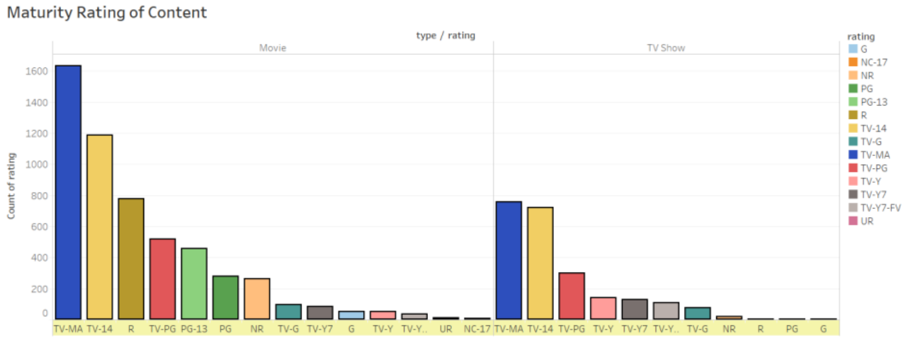
We had to determine the count of the maturity rating rows and divide them according the different ratings. Bar graph was the best suited visual for this task, because it clearly shows the frequency distribution, relative proportions between multiple categories and can be arranged in ascending order.

5.2 Implementation

For task 4, we had to summarize the type of content available on Netflix according to their maturity rating. The visualization displays the different maturity rating content in a bar graph. The graph was created on Tableau. The graph uses the following columns – Rating, Count(Rating) and Type (indicates whether Movie or Tv Show). We also filtered out null values – Rows where rating wasn't present.

5.3 Evaluation

5.3.1. Results



Count of rating for each rating broken down by type. Color shows details about rating. The view is filtered on rating, which excludes Null.

Figure 8 : Maturity Rating Analysis of Content on Netflix

5.3.2 Discussions

Netflix offers the option to set profiles that block mature content for children. According to our analysis, vast majority of content on Netflix is not suitable for children. We believe that Netflix realized the fact that they have mature content and therefore decided to add this setting. The internet isn't governed by the same censorship laws as broadcasted Television. This makes publishing R rated content easier and explains the presence of R rated movies on Netflix. During the genre analysis, we learned that in general, "Drama" content is widely available on Netflix. However, after joining IMDB movie dataset we realized that "Action" movies tend to bring in more money for the production companies.

For this set of tasks and visualization we used Think Aloud evaluation technique. This direct observation method was performed on two users which involved asking them to think out loud as they were performing a task. The user actions on this visualization pertaining to what they are looking at, what they are thinking, doing and feeling were noted down. This helped us in determining the user expectations and in identifying the aspects of the visualization that were unclear or perplexing.

The Bar graph was showcased to two participants addressing the following tasks:

- **Task 1:** Tell us what this visualization represents
- **Task 2:** Looking at the graph, identify the content rating which is the most available.
- **Task 3 :** Spot the difference, if any, between the 2 types of content available.

For all the tasks as listed above, the thoughts and actions of the participants were noted down. The single ease question score for these tasks was asked to every participant and noted down. This included just one question on - How difficult or easy did participants find the task? The range of the score was set from 1 – 5, with 1 being the toughest and 5 being the easiest. Below mentioned is the observation for the same.

Single Ease Question	Participant 1	Participant 2
Task 1	5	5
Task 2	5	5
Task 3	5	4

6. TASK 5

6.1 Design

6.1.1 Analysis

For task 5, we wanted to find out the highest earning genres and production companies. In order to carry out this task, we had to refer to multiple columns of data. We also had to join another dataset, because Netflix dataset didn't have the worldwide income and gross earning columns. This visualization only considers movies because the income data for TV shows was not available in the IMDB dataset. Tree maps was chosen to visualize multiple columns of data. The size of the tiles represents the popularity of Genres while the shade of the tiles represents the revenue, they bring in. The combination of shape and shades simplify the message of the visualization. Darker shades represent higher earnings by the production company and larger size represents higher earnings from the genres. After visualizing three columns- Worldwide gross income, Genre and Production Companies, we were able to understand the type of movies that are popular and their relationship with profitability. We were also able to link the profitability with the production companies.

From the visual – "Movie genres and production houses with the highest profit", we were able to identify the Genres and production companies that perform well at the box office. Action movies tend to perform well at the box office and Marvel studios has produced the highest earning movies in that genre. Marvel studios have so far earned $1,697,820,296 from movies that belong to "Action, Adventure and Sci-Fi" in gross revenue. Columbia Pictures is the 2nd highest earning production house with movies belonging to the same genre. 2nd highest earning genre is "Crime, Drama, Thriller" and has made Warner Bros $1,074,251,311.

6.1.2 Visualization

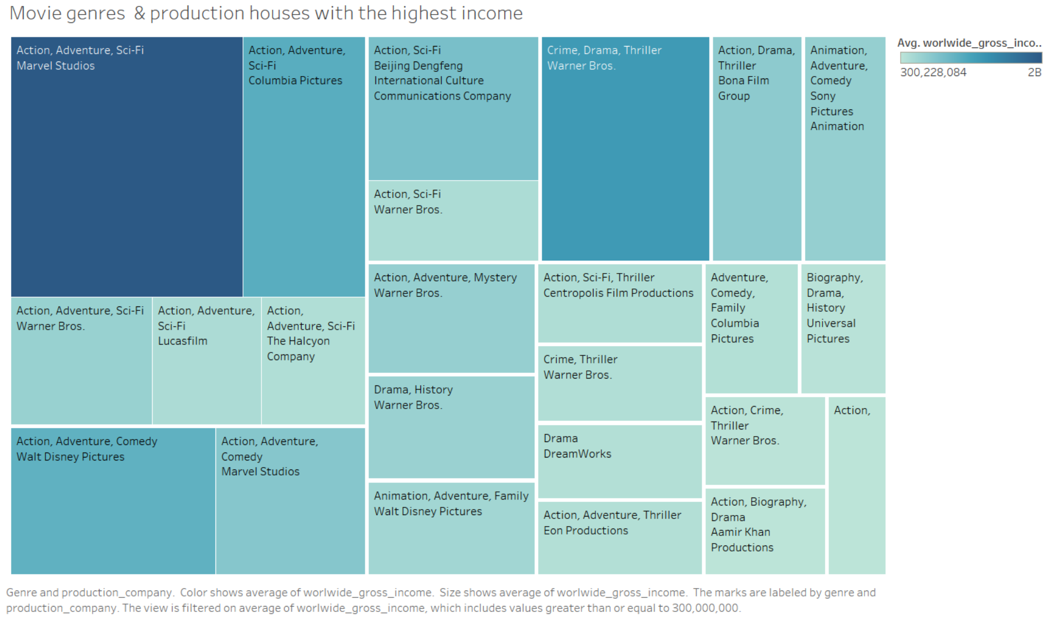
We had to visualize multiple columns to find out the highest earning genres and production companies. Tree maps was the visual of choice to represent multiple column magnitudes in a single frame. The size and shade express the magnitude of values from separate columns.

6.2 Implementation

To implement task 5, we performed the following pre-processing steps- Joined Netflix dataset with IMDB movies dataset. Inner join was done on Titles column using tableau prep. We chose Worldwide gross income, Genre and Production Companies columns for the visualization on Tableau. The graph of choice was a treemap. Lastly, we filtered out the data where gross worldwide earnings were lower than 300,000,000 (300 million). This was done to improve the visual.

6.3 Evaluation

6.3.1. Results



Movie genres & production houses with the highest income

Genre and production_company. Color shows average of worlwide_gross_income. Size shows average of worlwide_gross_income. The marks are labeled by genre and production_company. The view is filtered on average of worlwide_gross_income, which includes values greater than or equal to 300,000,000.

Figure 9 : Movie Genres and Production Houses with Highest Incomes

6.3.2 Discussions

Marvel has been extremely successful at the box office in the "Action, Adventure and Sci-Fi" genre and this can be attributed to the success of their Avengers franchise. Columbia films (Owned by Sony corporation of America) have earned millions of dollars from the same genre and some of this success can be explained by the success of Spiderman movies. Warner bros are 3rd when it comes to earnings in the same genre indicating that their DC movies division isn't as popular as marvel's cinematic universe. However, they have the highest worldwide income in the "Crime, Drama, Thriller" category. Apart from Action, the other moneymaking genres are Drama and Comedy. The revenue from these movies on average is the same among the competing production companies.

For task 5, we used cognitive walk through technique for evaluation. We asked 2 participants to evaluate the visualization against a set criterion and asked them questions relating to their experience with the visualization. The questions asked were:

- **Question 1:** Is the user able to understand the visualization?
- **Question 2:** Is the visualization complex or easy to understand?

After deciding the evaluation questions, we brought in the participants to view the visualization. We asked them if it was easy to read the text that was written in the tiles. We asked if the color contrast was balanced or was hindering the readability. The participants agreed that the font size was optimal, and the colors were balanced.

We also asked the participants if the intended message of the visualization was clear. We asked them whether they were able to decipher the intention behind using different shades and sizes of tiles. The participants confirmed that the message from the visual was clear and could tell that the shade represented the magnitude of a column and size represented the magnitude of another column.

7. TASK 6

7.1 Design

7.1.1 Analysis

For the given task, an analysis of movie and tv show genres available on Netflix is performed. For creating the visualizations, we have depicted the data by using pie chart, tree map and bubble chart to find the top genres available in both the movie and tv show category. A further analysis of finding the top genres in United States of America and the amount of movie genre content released each year was done. Our aim was to present an aesthetically pleasing and easily understandable graphs and charts for the data. To start with the analysis, we pre-processed the initial dataset by using Power query in Excel. Since, a movie or tv show can belong to different genres, the column had to be separated such that there was a unique row for each genre content. The above-mentioned graphs were presented in the form of a dashboard for quick comparison.

Following information was incurred from the visualizations created:

- Going against the revenue generation analysis (carried out in Task 5) , Drama is the most widely available genre in both movies and tv shows category which is then followed by comedy.
- Along with Drama and Comedy, Netflix also focusses on providing movies in the Action & Adventure, Documentaries and Thriller categories as well.
- Romantic and Reality TV shows also together form a major part of the tv show genre content available on Netflix.
- Movies belonging to the Drama Genre had seen a rise in availability in years 2016 to 2018. Other top categories also showed a similar trend.
- In USA, Movie Drama is the most widely available category.
- In USA which is the most content producing country, a wide availability of documentaries and comedy genre can be seen.

7.1.2 Visualization

The visualization for Top 5 movie genres on Netflix is presented using a pie chart. Furthermore, the visualizations for Top 5 tv show genres are shown by means of a tree map. This helps to find the genre which has the most availability by means of containers. A line chart was used to present the amount of top genre content released each year from 2000 to 2020. Finally, a bubble chart was used to depict the top genres available in USA on Netflix. The availability of the content can be seen from the size of the bubble. To consolidate all the graphs, a dashboard was used for quick comparison and analysis.

7.2 Implementation

Following steps were performed for the implementation:

- The Netflix dataset was pre-processed in Excel using Power Query. The ‘listed in’ columns consisted of different kind of genres and hence was used to segregated to include individual rows.

- The Pie-chart depicting the ‘Top 5 movie genres’ on Netflix was implemented by using the Tableau as a visualization tool. Here the count of ‘listed in’ column was used. Pie-chart was selected from the Show-me. Enter the ‘listed in’ in filter and select Top 5 for this column to just show the top 5 genres.
- Similarly, A tree map was selected as the necessary chart for presenting the ‘Top 5 TV Show genre’. The filters were used to select only the needed categories.
- A line chart was used for presenting the amount of movie genre content yearly. The filter was applied on ‘release year’ to include only years 2000-2020.
- Finally, a bubble chart was used to show the top genres in USA. Here the filter of countries consisted of a query to select only USA as the country.
- All these charts and graphs were consolidated in the form of a dashboard to get single view of all.

7.3 Evaluation

7.3.1. Results

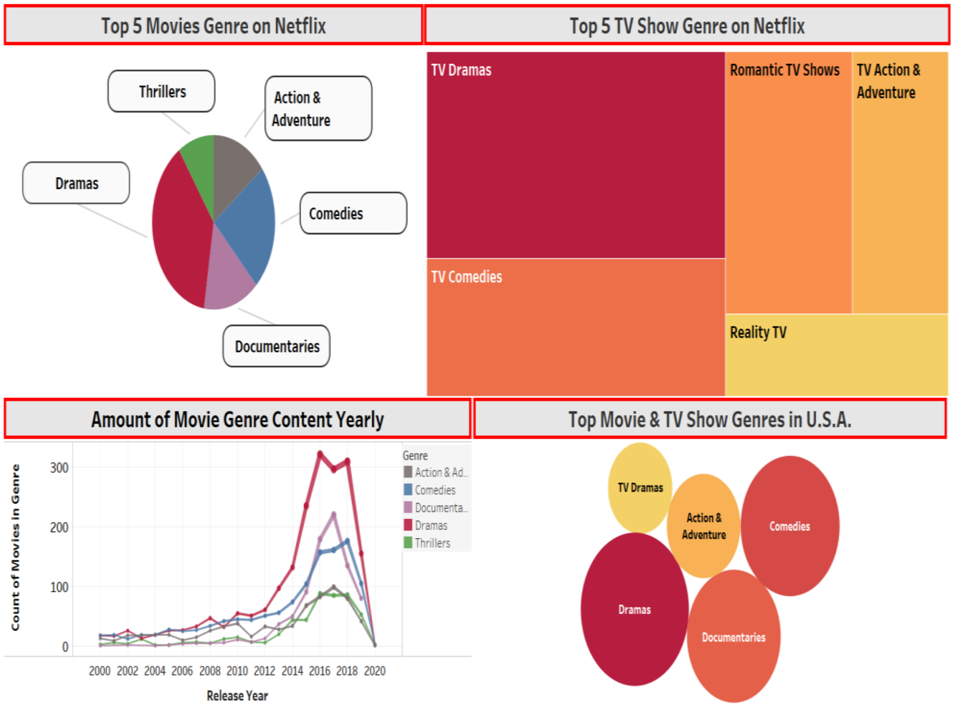


Figure 10: Dashboard of Genre Analysis

7.3.2 Discussions

For this group of visualizations, we used the think aloud evaluation technique. This observation was performed on two users which consisted of users orally walking themselves through the presented visualizations like they were performing a task. The user actions on these visualizations pertaining to what they are looking at, what they are thinking, feeling and doing were logged. This helped us to determine the user’s expectations and identifying the details of the visualizations that were uncertain or vague.

The dashboard consisting of pie chart, tree map, line chart and bubble chart were showcased to 2 participants addressing the following task:

- **Task 1:** Tell us what this visualization represents.
- **Task 2:** Looking at the graphs determine the most popular of the shown genre in both the Movie and TV Show category.
- **Task 3:** Find the period during which the top genres were widely added by Netflix.

For all the tasks mentioned above, the thoughts and actions of the participants were noted down. The single ease question score for these tasks was asked to every participant and logged. This included just one question on – How difficult or easy did the participants find the task? The range of the score was set from 1 – 5, with 1 being the toughest and 5 being the easiest. The observation for the same is tabulated below.

Single Ease Questions	Participant 1	Participant 2
Task 1	5	5
Task 2	4	4
Task 3	5	4

8. TASK 7

8.1 Design

8.1.1 Analysis

In this visualisation we have tried to make a connection between the average duration of movies and TV shows with regard to the years passed by and what is the trend that we can extrapolated from it. Below mentioned is the summary of the analysis carried out:

- The Movie Graph show us that the average duration of the total run time has been decreasing with each year. The highest use to be in 1964 with 228 mins which has now dropped to just 104 mins in 2020. This is significant decrease in movie duration.
- The TV shows graph also shows a decrease in the average runtime of the TV shows with each year. The highest used to be in 1992 with 11 seasons which has now decreased to just 1 season in 2020.

These are the results that were generated from the visualisation about the movie and TV average duartion in the newtflix dataset.

8.1.2 Visualization

For this graph we designed it using an area chart in Power BI and put time intervals to highlight a relation between the movie and TV show average duration with repect to years passing to the present on the netflix dataset. We have mainly used Power BI and Excel to make this visualization and our main aim was to gain insights from the data wheteher that the duration has increased or has stayed consistant or even decreased and we gained some interesting insights that were not expected.

8.2 Implementation

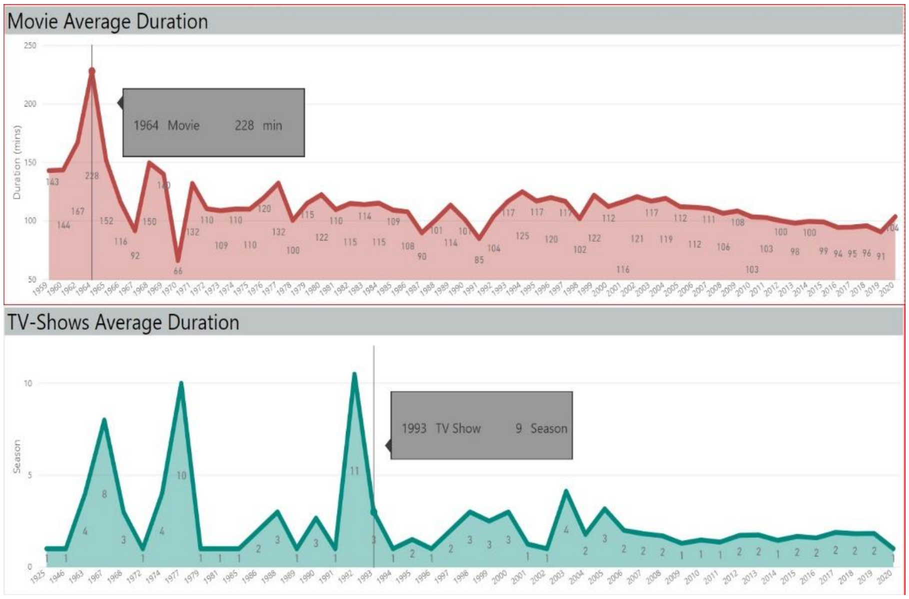
We mainly used MS Excel and Power BI for this visualization and the following steps show how the visualization was created.

- Import Netflix titles in power bi desktop.

- cleaning data: split duration column to two columns, "duration" and "duration unit".
- used area chart to show Movie average duration over the release years.
- X-axis = release year
- Y-axis = Average of duration
- enabled data labels for clear understanding of the trend.
- and filtered type to "Movie"
- Used same process for showing TV shows average duration over the release year.
- X-axis = release year
- Y-axis = Average of duration (seasons)
- enabled data labels for clear understanding of the trend.
- and filtered type to "TV-Shows"
- created a tooltip page for better visual of the details by creating table with columns; year, type, duration, unit.

8.3 Evaluation

8.3.1. Results



<u>Figure 11 :</u> Content Duration Analysis

8.3.2 Discussions

There is general trend decrease in the average duration of both TV shows and Movies which can be due to various factors like people's short attention span and lives getting busier etc. And for these set of tasks the Cognitive walktrough was used as the method for evaluation of this visualisation and three cognitive walktrough experts were asked these two question

Question 1: Will the user understand the visualization?

Question 2: Does the visualization seem complex to the user?

The experts were very easily able to understand what the visualization was about and they were able to understand the insights that the visualization had generated.

9. CONCLUSION

Netflix is a huge media streaming platform and achieved its current size due to the massive growth that it has experienced over the years. It originated in the US in 2008 but has since expanded to more than 100 countries worldwide. Their growth was fueled by the addition of content, most of which was added between 2016 and 2019. This content was added because of the demand in both developed and developing countries. Netflix prefers adding content during the holidays and our analysis shows that months of October, November and December observe a rise in content addition. As per our observation, more movies are added compared to TV shows.

USA has led the content contribution race and produces the highest number of Movies and Tv shows available on the platform. India has closed the gap with UK and secured 2nd place in recent years, owing to the production of indie content.

Drama is the most widely available genre on Netflix, followed by comedy. Romantic and Reality TV genres are highly popular when it comes to Tv shows. However, it should be noted that there are more movies available on the platform compared to Tv shows. Movies make up 70% of the content on Netflix. However, the average duration of TV shows and movies has been on a steady decline. Movies used to last for over 200 mins in 1960s but only run for 1.5 hours now. TV shows used to have multiple seasons before (8+), but now mostly end in 1-3 seasons. Most of this content is for mature audiences. Viewers need to be above 17 years of age in order to view most of the media on the platform.

On joining IMDB dataset, we found out that even though drama is widely available on Netflix, it's the action genre that brings in the most amount of money. Marvel, Walt Disney and Warner bros have produced some of the highest earning movies in Hollywood.

10. REFERENCES

- IMDB dataset source https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

11. APPENDIX

11.1 Weekly Group Meeting Minutes

11.1.1. Week 6

Meeting Name	**Assignment Group 10: Meeting 1**		
Date of the Meeting	5th October 2020	Time	07:00-08:00pm
Meetings prepared by	Mohammed Saif	Location	Zoom
Attendees	Mohammed Saif, Ishan Sajnani, Wilson Lobo, Roland Lobo, Utkarsh Bele and Shwetali Tonape		
Meeting Objectives			
• **To brainstorm around the given assignment datasets and finalize the one that the team will pursue for the course of Assignment 2**			
Discussions			
1) Pros and Cons of all the dataset were discussed 2) Dataset Decided (Netflix movie and tv shows dataset)			

3) Collaboration tool decided – Dropbox
4) Sub-topics for Individual visualizations assigned
5) For the next meeting, everyone agreed to research around the dataset and finalize the questions in the next meeting.

Agenda for next meeting

Exploratory Analysis Questions and Individual topics assigned

11.1.2. Week 7

Meeting Name	**Assignment Group 10: Meeting 2**		
Date of the Meeting	11th October 2020	Time	09:00-10:00pm
Meetings prepared by	Ishan Sajnani	Location	Zoom
Attendees	Mohammed Saif, Ishan Sajnani, Wilson Lobo, Roland Lobo, Utkarsh Bele and Shwetali Tonape		

Meeting Objectives

- **To brainstorm around the selected Netflix dataset and write down the exploratory analysis questions pertaining to the same.**
- **Divide the questions among the team members.**

Discussions

1) Team brainstormed around the Netflix dataset and took a column by column approach to discuss the questions.
2) Discussion in considering an additional dataset for IMDB ratings.
3) Checking the usability index of the data individually and post merging the same with the original dataset.
4) Note was taken for the following questions:
    a. Understanding what content/genre is available in different countries
    b. Change in movie/tv trends from 2000- 2020? genre/ content? (not enough data)
    c. Is Netflix increasingly focusing on TV rather than movies in recent years? (insights on TV/movies content)
    d. Network analysis of Actors / Directors and find interesting insights (more data)
    e. Combine IMDB data to find out highly rated content and it's cast
    f. Avg. duration of movie/tv shows. Segregate year by year
    g. What maturity rating content does well?
    h. Which country produces the most content? or the top ten countries with the most content? (Using a bar chart or pie chart, tree maps can also be used in this case)
    i. Clusters of movies based on sequels prequels or the different seasons

Individual Visualization Accountability	
Name	Exploratory Questions
Mohammed Saif	What maturity rating content does well?
Ishan Sajnani	Is Netflix increasingly focusing on TV rather than movies in recent years? (insights on TV/movies content)
Wilson Lobo	Change in movie/tv trends from 2000- 2020.
Roland Lobo	Understanding what content/genre is available in different countries
Shwetali Tonape	Which country produces the most content? or the top ten countries with the most content?
Utkarsh Bele	Avg. duration of movie/tv shows (Segregated year on year)
Agenda for next meeting	
Review the individual progress on visualizations	

11.1.3. Week 8

Meeting Name	**Assignment Group 10: Meeting 3**		
Date of the Meeting	22nd October 2020	Time	09:30-10:30pm
Meetings prepared by	Utkarsh Bele	Location	Zoom
Attendees	Mohammed Saif, Ishan Sajnani, Wilson Lobo, Roland Lobo, Utkarsh Bele and Shwetali Tonape		
Meeting Objectives			
• **To review and analyse the teams progress on the visualisations assigned to individual team members and discuss about the presentation details.**			
Discussions			
1) Individual team members presented their visualisations to the team for their feedback. 2) The team discussed what will be the main visualization in focus around which the other visualisations will revolve. 3) The team discussed that coming Sunday will be when the entire team will finalise all the visualisations. 4) The presentation will be completed and finalised by Tuesday which will give the necessary amount of time to the team for any error corrections.			
Agenda for next meeting			
Finalizing Individual assigned Visualizations			

11.1.4. Week 9

<table>
<tr><td>Meeting Name</td><td colspan="3"><b>Assignment Group 10: Meeting 4</b></td></tr>
<tr><td>Date of the Meeting</td><td>25th October 2020</td><td>Time</td><td>07:00-9:00pm</td></tr>
<tr><td>Meetings prepared by</td><td>Roland Lobo</td><td>Location</td><td>Zoom</td></tr>
<tr><td>Attendees</td><td colspan="3">Mohammed Saif, Ishan Sajnani, Wilson Lobo, Roland Lobo, Utkarsh Bele and Shwetali Tonape</td></tr>
<tr><td colspan="4">Meeting Objectives</td></tr>
<tr><td colspan="4">• <b>Reviewing each team members visualisations, compilation into presentation slides and deciding the flow of the upcoming presentation.</b></td></tr>
<tr><td colspan="4">Discussions</td></tr>
<tr><td colspan="4">1) Individual team members presented their finalised visualisations to the team.
2) The team discussed the flow of the presentation.
3) The team decided the time period allotted to each member for the final presentation based on the part taken.
4) Future discussions and graphs to be included for the final report.</td></tr>
<tr><td colspan="4">Agenda for next meeting</td></tr>
<tr><td colspan="4"><b>Prepared scripts for the recording of the presentation and submission.</b></td></tr>
</table>

11.1.5. Week 10

<table>
<tr><td>Meeting Name</td><td colspan="3"><b>Assignment Group 10: Meeting 5</b></td></tr>
<tr><td>Date of the Meeting</td><td>5th November 2020</td><td>Time</td><td>06:00-6.30pm</td></tr>
<tr><td>Meetings prepared by</td><td>Wilson Lobo</td><td>Location</td><td>Zoom</td></tr>
<tr><td>Attendees</td><td colspan="3">Mohammed Saif, Ishan Sajnani, Wilson Lobo, Roland Lobo, Utkarsh Bele and Shwetali Tonape</td></tr>
<tr><td colspan="4">Meeting Objectives</td></tr>
<tr><td colspan="4">• <b>Decide every team member part in preparing the final report.</b>
• <b>Decide the structure of the report</b></td></tr>
<tr><td colspan="4">Discussions</td></tr>
<tr><td colspan="4">1) Ishan, will work on the introduction part, write a brief about the datasets and the task.
2) By next meeting, every team member agreed to complete the Design and implementation part of their visualization
3) Saif, will be writing the Conclusion as well
4) Everyone needs to evaluate and justify their final visualization and write a visual analysis or storytelling not exceeding 1 page
5) For personal reflection, everyone will have to write 0.5 page per week. More details to be discussed in next week's meeting.</td></tr>
</table>

Agenda for next meeting
Prepared scripts for the recording of the presentation and submission.

11.1.6. Week 11

Meeting Name	**Assignment Group 10: Meeting 6**		
Date of the Meeting	12th November 2020	Time	6:00- 7:00 pm
Meetings prepared by	Shwetali Tonape	Location	Zoom
Attendees	Mohammed Saif, Ishan Sajnani, Wilson Lobo, Roland Lobo, Utkarsh Bele and Shwetali Tonape		
Meeting Objectives			
• **Reviewing each team members progress on the report and assigning individual tasks for the final report.**			
Discussions			
1) Team members discussed their report progress 2) Compile the individual work on Sunday (15th November) and final report edit on Wednesday (18th November) 3) Team members discussed on evaluation methods and personal reflection			
Agenda for next meeting			
Finalize and compile the report			

11.2 Weekly Team's Personal Reflection

11.2.1 Week 6

1. **Ishan Sajnani :** In Week 6, our team organized our first team meeting, wherein we aimed to brainstorm around the assignment datasets and discuss the usability and data shape of each dataset. It was quite interesting to discuss the usability index of the datasets. A prior notice about the agenda was communicated to all the team members, so that everyone comes prepared to the meeting with their notes on using different datasets. The meeting kicked off by discussing the datasets one by one, the real challenging part was narrowing down from top 3 to the selected one. For the same, we referred to the usability index and the data quotient option on Kaggle, as a metrics along with the shape of the data. The shape of the data was considered because it would determine the amount and the level of the pre-processing that the data needs. My opinion was with two datasets, one was the Netflix dataset as everyone has prior knowledge about the dataset. According to me, having the subject knowledge about the dataset is important to understand the past and carry out exploratory and descriptive analysis. Having said that, my second choice of the dataset was NBA dataset.
2. **Mohammed Saif:** Looked at different datasets and found Netflix dataset the most intriguing. Consulted with group and started brainstorming on different things that we could do with the dataset. In order to decide the tasks, I first looked at the dataset. Examined what columns are available, checked null values and looked at aspects that would need preprocessing.

3. **Shwetali Tonape:** In the first week, I spend some time to understand the requirements of the assignment, the marking rubrics and the chosen Netflix dataset. After understanding the Netflix dataset, I made a list of all questions that can be considered to visualize the dataset.
4. **Roland Lobo:** After finalizing the dataset for our group assignments, during the group meeting scheduled I participated in first discussing the pros and cons of Netflix dataset selected. The put forward my views of different visualization that could fit with the dataset. Finally, before the next meeting we had to do a research on the same.
5. **Utkarsh Bele:** In this week I mainly focused on understanding the datasets available. After discussion with the team it was decided that we will work on the Netflix dataset. I looked at the Netflix dataset and tried to understand it.
6. **Wilson Lobo:** Went through the assignment document and develop a understanding of what is expected. Found NBA dataset, University data and Netflix dataset as potential dataset. Shared my views and understanding regarding the assignment and dataset on the call to the group. Shared my opinion on the visualization that would complement the dataset.

11.2.2 Week 7

1. **Ishan Sajnani :** In Week 7, as planned everyone took time to study the dataset and define descriptive and exploratory task that could be carried out as analysis. We had our second meeting during Week 7, wherein everyone was prepared with their set of questions and tasks. One of the challenges that I faced during this week, was finding the right dataset to integrate in-order to explore more tasks. In the meeting, we discussed all the questions and tasks that could be covered. I was assigned the task to analyze, that if Netflix is increasingly focusing on what type of content and analyze monthly patterns in the addition of content.
2. **Mohammed Saif:** I started thinking about what columns I will use to visualize and analyze my task. Apart from the columns I also had to decide which application to use for visualization. Checked out different visualization applications and chose Tableau for its ease of use and wide selection of graphs.
3. **Shwetali Tonape:** After deciding on my task in the zoom meeting, I researched and studied various visualization tool to understand the best fit for my visualization. I worked on the various visualization tool, and finally decided to work on python.
4. **Roland Lobo:** Brainstorming around the Netflix dataset helped to come up with various questions that could be addressed and represented by means of a visualization. Discussion to work with other related dataset like IMDB to help us expand our analysis. Finally, at the end of the meeting I undertook the task of performing Genre Analysis on the contents in the Netflix dataset.
5. **Utkarsh Bele:** In this week I looked at the datasets and its working. Tried to understand the what the dataset is about and all the columns and the entire data. Since the questions were decided I looked at the task assigned.
6. **Wilson Lobo:** Started my research on the Netflix data, analyzed the different columns and identified crucial insights that can be derived. Also looked for additional dataset that would support the findings. Imported the dataset in various tools and looked by unique visualization. Tableau offered good visualization considering the large Netflix data other tools showed few issues.

11.2.3 Week 8

1. **Ishan Sajnani :** In Week 8, now everyone is working on their assigned task. For my task, firstly I carried out some pre-processing using Python and Power Query. Furthermore, I imported the data into tableau and was able to prepare the line chart and gradient bar chart from the same. For the presentation purposes, I was able to add animation to the line chart, as I was able to visualize the growth in the TV show and movies content over the years. During our team meeting in the Week 8, everyone discussed their progress on assigned task, and we had discussions pertaining to the Assignment presentation. For the same, the team brainstormed and discussed which visualization or which task would hold the center stage in our analysis and which visualizations will revolve around the same. During the meeting, the team established the deadline of weekend to meet and collaborate all the findings and draft a presentation. Also, set the deadlines for the final presentation to be completed by Week 9, just before the assignment presentation due date.
2. **Mohammed Saif:** Began working on Tableau and visualized the required columns in a bar graph. The message was visible but after spending more time on it realized that segregating on type will make the visual more appealing. Segregation will also increase the visibility on the similarities and differences in content rating between Movies and Tv Shows.
3. **Shwetali Tonape:** Worked around the Netflix dataset and created various visualizations on python and tableau. Researched more on different available visualizations on GitHub including different datasets and took inspiration for my visualization.
4. **Roland Lobo**: The required pre-processing was done on the dataset to make it suitable for visualizing. A draft of the visualization was presented in the meeting to get some feedback before I could finalize my visualization and present it as a dashboard.
5. **Utkarsh Bele:** In this week I tried to understand the task that I was given about making the visualization. I also tried to view the dataset and what else I will need to make any changes to the dataset.I looked at the dataset and if any changes or make any pre-processing on the dataset.
6. **Wilson Lobo:** Finalized Tableau as my tool to be used. Implemented different visualizations (learned from YouTube, GitHub and tableau learning forum) and demonstrated to group and take a note of their views. Focused on keeping the graph simple and picked the task to identify the global trend of Netflix.

11.2.4 Week 9

1. **Ishan Sajnani :** As planned by Week 9, we will be finishing our presentation. In order to do the same, we aim to combine all the visualizations and decide a good analysis storyline to communicate the same. In week 9, I used Power BI to visualize the growth of TV shows and movies by representing them in a funnel graph. In order to do the same, firstly I aggregated the data using Python. Starting with grouping by year and aggregating the count of shows based on content type. I did it separately for different content types in different data frames and then at the end, I merged the two and exported the final .csv file. Subsequently, I imported the data from that csv file into Power BI and selected funnel graph as the desired data graph. In the same, I grouped the data based on year and computed the aggregate of movies and Tv-Shows separately. Below mentioned is the result of the same. In Week 9, I also worked on the second part of the analysis which aimed at analyzing monthly trends in the addition of content on Netflix. Holding onto the same, we intent to recommend the months, wherein the content producers can release their TV-shows and Movies to gain more viewership and reach out to more audiences. From the observations made in the first part of the analysis, we have only considered time from 2016 to

2019 as Netflix has added majority of its content during this time. As we aim to recommend the months, wherein a new content producer can launch his/her show or movie on Netflix. From our analysis, as per my analysis, my advice and data-based recommendation to any new content producer would be to add content in month of January, to get a good range of viewing audience.

2. **Mohammed Saif:** The first visual looked great but wanted to see if more insights could be uncovered. Searched for datasets and found IMDB dataset. The dataset only contained movies but had considerable amount of titles that were present in the Netflix. Moreover, the dataset had columns for production companies and gross income and could enhance on the insights we had.
3. **Shwetali Tonape:** After finalizing the visualization, I worked on the slides for the presentation. Analyzed and Reviewed the visualization and found the pros and cons. Finally, compiled all that into the final slide and worked on the script for my presentation.
4. **Roland Lobo:** I presented my final set of visualizations to be included in the presentation. Preparation for the presentation was done taking in consideration the time constraint allotted to each member. Final report discussion to add or improve visualization was done.
5. **Utkarsh Bele:** I mainly focused on making the necessary pre-processing on the dataset. I made a rough visualisation and showed it to the team. The team decided that there were some changes needed in the visualisation. I made the necessary changes in the visualisation. I then made the visualisation look visually appealing and some design changes. I then stated working on the presentation and report.
6. **Wilson Lobo:** My main visualization was finalized, and I started looking for graphs that could complement my main visualization, I felt this would make it easier for users to understand more details in a concise way. Also focused on incorporating the changes/suggestions highlighted by team members. Prepared for the presentation with finalized script and practiced are transition.

11.2.5 Week 10

1. **Ishan Sajnani :** With our presentation and peer review done in Week 9. This week we had a meeting, highlighting the reviews of everyone from the presentation. The aim of the same was to have a retrospective view of things and record the lessons learned. By recording the lessons learned we aim to improve our visualizations for the final report. In Week 10, I also worked on the visualization aiming to analyze the maturity rating of the content that Netflix has been focusing on, in two of the most content producing countries. The same has been represented using an area chart, as represented below. In the same we have tried to contrast the growth of content of different ratings in top content producing countries. Along with the same, we divided the work among ourselves for different sections of the final report. In the same, I was assigned the work of writing the introduction section of the report along with my report on my share of task.
2. **Mohammed Saif:** In order to join the 2 datasets, I decided to use Tableau prep. Previous experience working with Tableau prep had already equipped me with skills needed to combine and analyze the 2 datasets. Performed a join and managed to create the second visualization – Highest earning genres and production companies.
3. **Shwetali Tonape:** Read some research papers, journals and books to better understand the flow for the report. Read up on various story- telling of visualizations. Finally, started working on the report and combined all the knowledge that I gained in the past few weeks into the report.

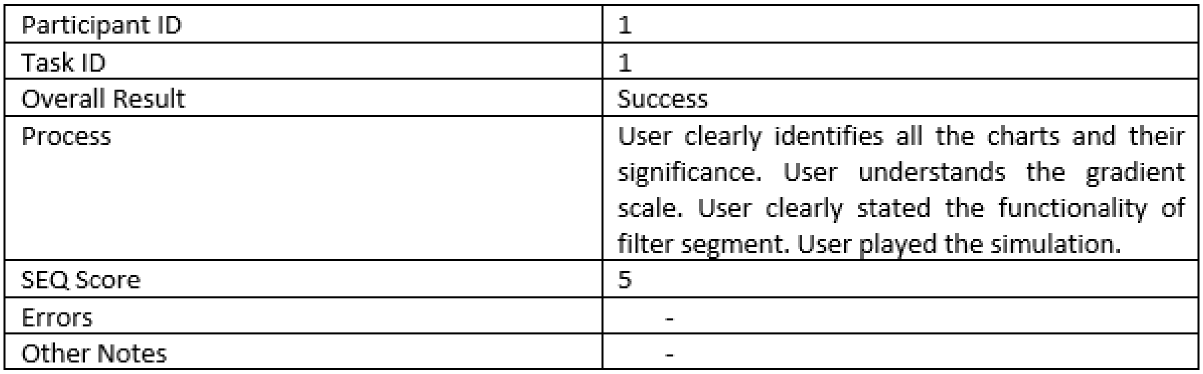
4. **Roland Lobo**: I added another visualization related to the genre content over the years. This was presented using a line graph. Further, I started my work on the design and implementation of the final report for my part.
5. **Utkarsh Bele:** The team decided on the structure of the report. I started working on the final report in this week. I started working on the design and implementation part on my individual part.
6. **Wilson Lobo:** I started working on the report and researched on how to evaluate and how to perform visual analysis. Also, research on technique of think-out-aloud and drafted task for the candidate to evaluate the visualization.

11.2.5 Week 11

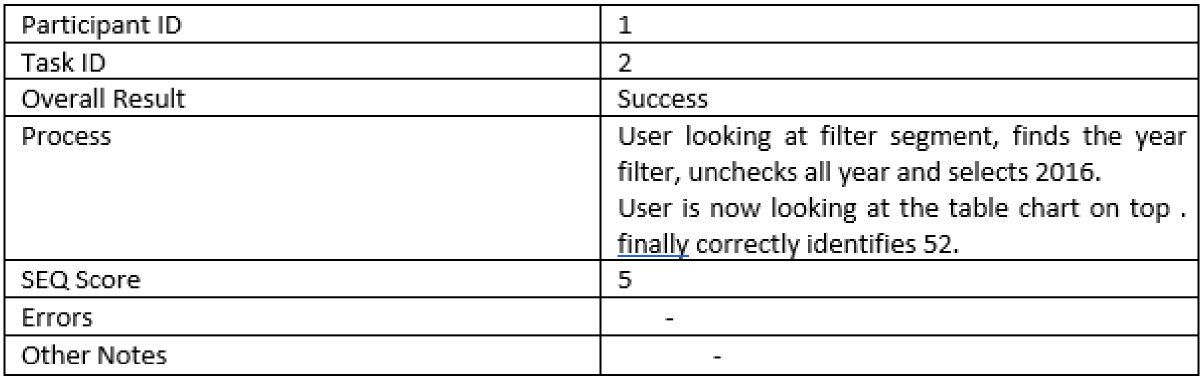
1. **Ishan Sajnani :** During the final week, we plan on compiling everyone's content . As everyone will present the Design, Implementation and Evaluation part of their visualization. In the evaluation we are using the Cognitive walkthrough and the think aloud techniques. In our meeting, we discussed the format and the parameters like Tasks and number of participants to be included in the think aloud and number of experts in the cognitive walkthrough.
2. **Mohammed Saif:** Discussed and decided on the structure of the report. Noted down all the preprocessing and implementation steps. Analyzed the visuals and used the evaluation techniques to critique the visualization. Peer reviewed the report and learned about group mates' perspectives through their task reports.
3. **Shwetali Tonape:** Reviewed my report and after suggestions from team members, updated the report accordingly. Read on different evaluation techniques used and applied it on the visualization. After, a final review worked on the minor editing.
4. **Roland Lobo:** Finalize the report and discuss with the group if any modifications are needed on my part.
5. **Utkarsh Bele:** In this week I almost completed all my part of the report. I asked the team members about any changes need in my part. I made the necessary changes in my report. I started working on my evaluation methods used for my task and the personal reflection part as well.
6. **Wilson Lobo:** Finalized the report structure with team members. Asked my brother and a friend to be a candidate for think-out-aloud evaluation. Conducted the evaluation and prepared the template to be added in appendix. Coordinated with team members and reviewed the report. Created a Demo of the Netflix global trend to be submitted along with report.

11.3 Evaluation Methods

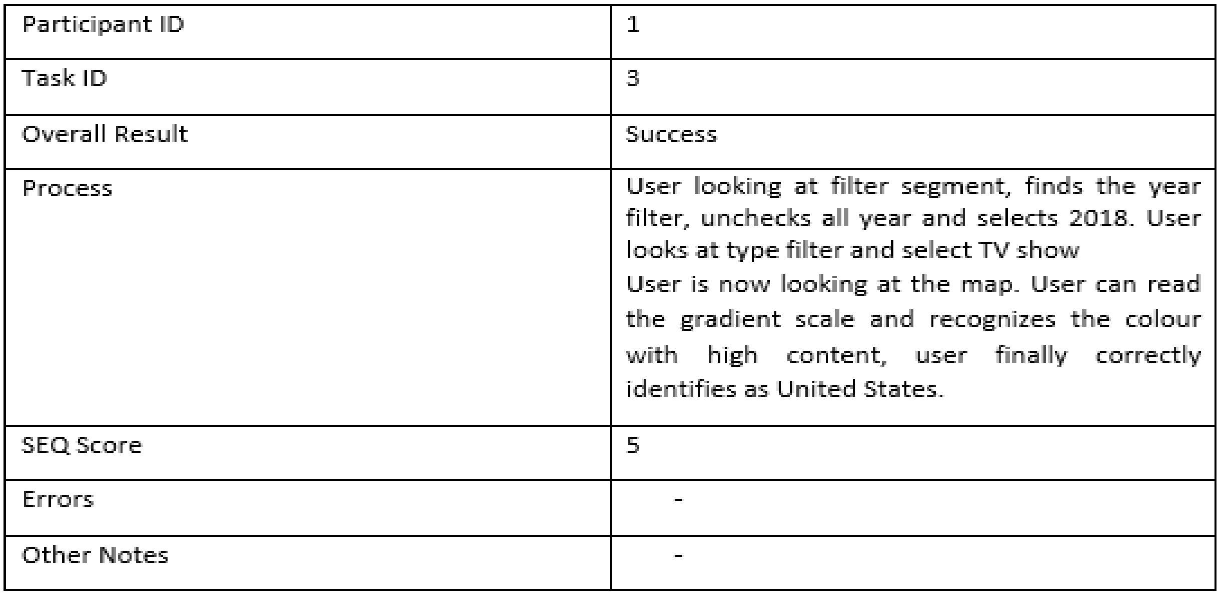
11.3.1 Task 1 – Think Aloud Technique



Participant ID	1
Task ID	1
Overall Result	Success
Process	User clearly identifies all the charts and their significance. User understands the gradient scale. User clearly stated the functionality of filter segment. User played the simulation.
SEQ Score	5
Errors	-
Other Notes	-



Participant ID	1
Task ID	2
Overall Result	Success
Process	User looking at filter segment, finds the year filter, unchecks all year and selects 2016. User is now looking at the table chart on top . finally correctly identifies 52.
SEQ Score	5
Errors	-
Other Notes	-



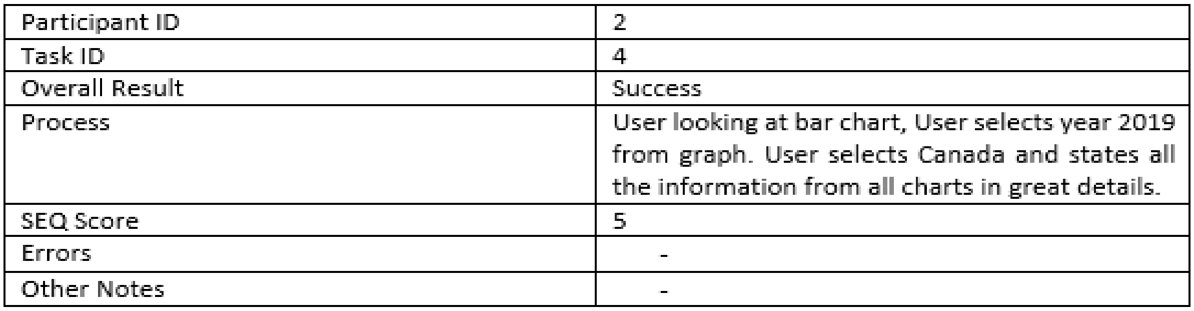
Participant ID	1
Task ID	3
Overall Result	Success
Process	User looking at filter segment, finds the year filter, unchecks all year and selects 2018. User looks at type filter and select TV show User is now looking at the map. User can read the gradient scale and recognizes the colour with high content, user finally correctly identifies as United States.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	1
Task ID	4
Overall Result	Success
Process	User now clicks on Canada and starts stating the information from all charts and realizes he didn't select 2019 from the filter. Finally, identifies all information correctly.
SEQ Score	4
Errors	-
Other Notes	Missed details from content distribution chart.

Participant ID	2
Task ID	1
Overall Result	Success
Process	User clearly identifies all the charts and their significance. User clearly stated the functionality of filter segment. User is aware of the actions of each graph. User played the simulation.
SEQ Score	5
Errors	-
Other Notes	-

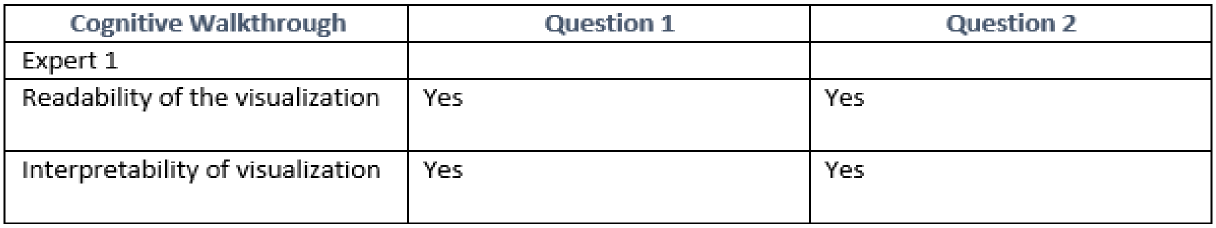
Participant ID	2
Task ID	2
Overall Result	Success
Process	User looking at bar chart, User selects year 2016 from graph. Acknowledges that all graphs refreshed. User is now looking at the table chart on top . Finally correctly identifies 52
SEQ Score	5
Errors	-
Other Notes	

Participant ID	2
Task ID	3
Overall Result	Success
Process	User looking at bar chart, User selects year 2018 from graph. User selects the TV show on pie graph. User is now looking at the map and hovers over US and India. User finally correctly identifies as United States.
SEQ Score	4
Errors	-
Other Notes	User was confused with gradient for India

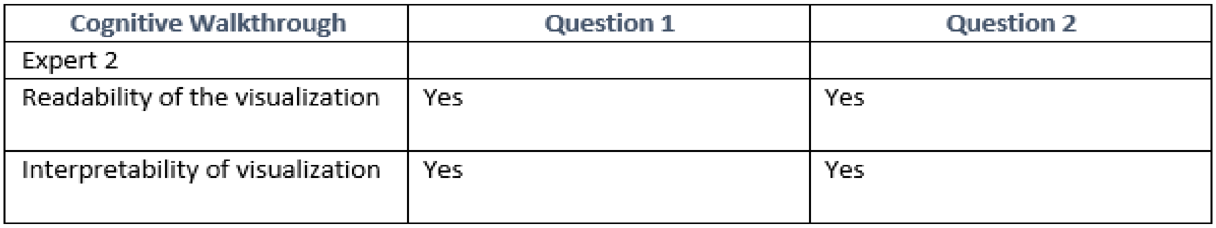


Participant ID	2
Task ID	4
Overall Result	Success
Process	User looking at bar chart, User selects year 2019 from graph. User selects Canada and states all the information from all charts in great details.
SEQ Score	5
Errors	-
Other Notes	-

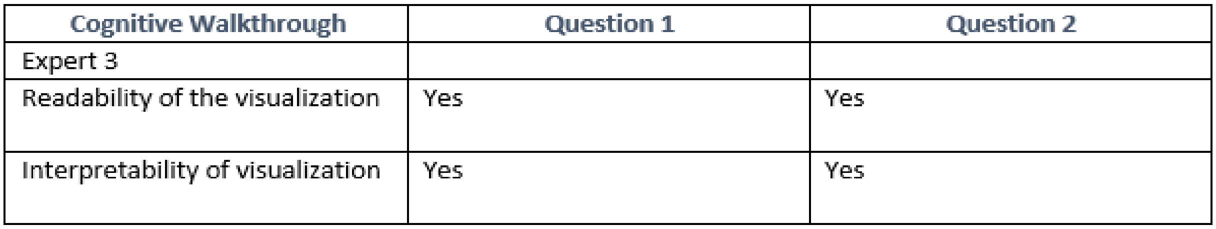
11.3.2 Task 2 – Cognitive Walkthrough



Cognitive Walkthrough	Question 1	Question 2
Expert 1		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes



Cognitive Walkthrough	Question 1	Question 2
Expert 2		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes



Cognitive Walkthrough	Question 1	Question 2
Expert 3		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes

11.3.3 Task 3 – Think Aloud Technique

Participant ID	1
Task ID	1
Overall Result	Success
Process	The user was easily able to understand what the visualization represents and understood what we intent to communicate with the same.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	1
Task ID	2
Overall Result	Success
Process	The user was easily able to understand the comparison between the content added in the US and India and user was successfully able to point out the reason for the same.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	1
Task ID	3
Overall Result	Success
Process	User easily identified the spike in the addition of content during the months of October-December.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	2
Task ID	1
Overall Result	Success
Process	The user was easily able to understand what the visualization represents and understood what we intent to communicate with the same.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	2
Task ID	2
Overall Result	Success
Process	The user was easily able to understand the comparison between the content added in the US and India. With the same the user was able to make out which rating is preferred in which country.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	2
Task ID	3
Overall Result	Success
Process	The user looked at the stacked area chart and read the legend. The User is easily able to identify and analyze the spike signifying the growth of content in years of 2018 and 2019 successfully.
SEQ Score	4
Errors	-
Other Notes	-

11.3.4 Task 4 – Think Aloud Technique

Participant ID	1
Task ID	1
Overall Result	Success
Process	User looked at the bar graph User has identified the intent of the graph.
SEQ Score	5
Errors	-
Other Notes	-
Participant ID	1
Task ID	2
Overall Result	Success
Process	User looked at the graph. User was able to easily identify the most available maturity rated content.
SEQ Score	5
Errors	-
Other Notes	-

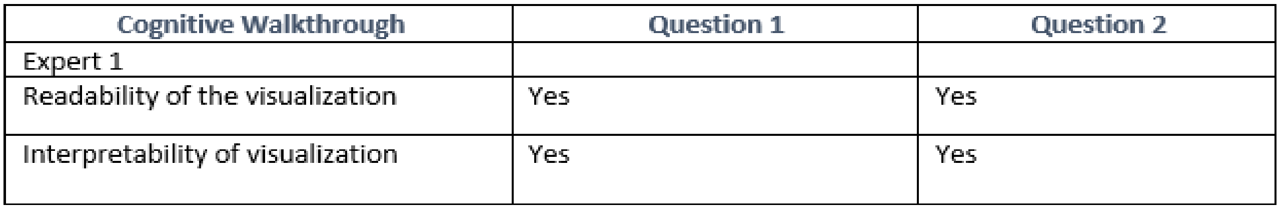
Participant ID	1
Task ID	3
Overall Result	Success
Process	User looking at the stacked area chart. User identifies that there is a higher amount of movie rated R content.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	2
Task ID	1
Overall Result	Success
Process	User looked at the bar graph User has identified that the graph is about maturity ratings.
SEQ Score	5
Errors	-
Other Notes	-

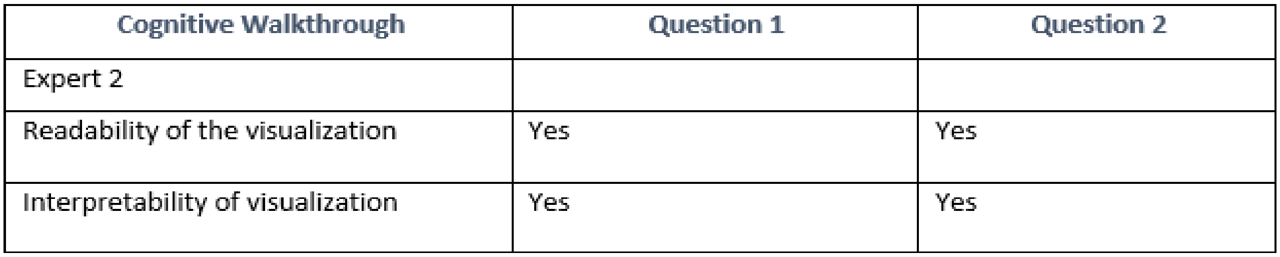
Participant ID	2
Task ID	2
Overall Result	Success
Process	User looked at the graph. User was able to easily identify the most available maturity rated content.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	2
Task ID	3
Overall Result	Success
Process	User looking at the stacked area chart. User was able to spot the difference between the 2 categories- fewer R rated TV Shows.
SEQ Score	4
Errors	-
Other Notes	-

11.3.5 Task 5 – Cognitive Walkthrough

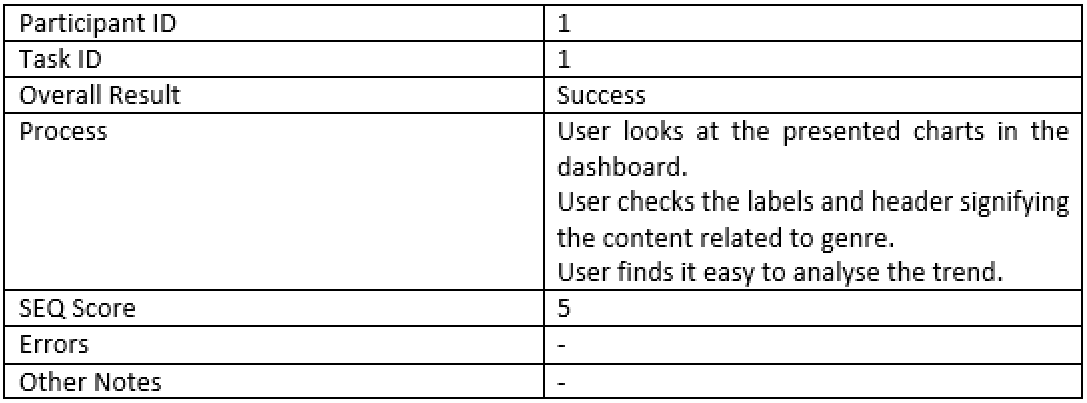


Cognitive Walkthrough	Question 1	Question 2
Expert 1		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes



Cognitive Walkthrough	Question 1	Question 2
Expert 2		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes

11.3.6 Task 6- Think Aloud Technique



Participant ID	1
Task ID	1
Overall Result	Success
Process	User looks at the presented charts in the dashboard. User checks the labels and header signifying the content related to genre. User finds it easy to analyse the trend.
SEQ Score	5
Errors	-
Other Notes	-

Participant ID	1
Task ID	2
Overall Result	Success
Process	User looks at the pie chart, tree map and bubble chart in the dashboard. User identifies the distributed regions in the graphs and charts. User finds the popular genre from the visualisation. User finds it easy to infer the needed information
SEQ Score	4
Errors	-
Other Notes	-

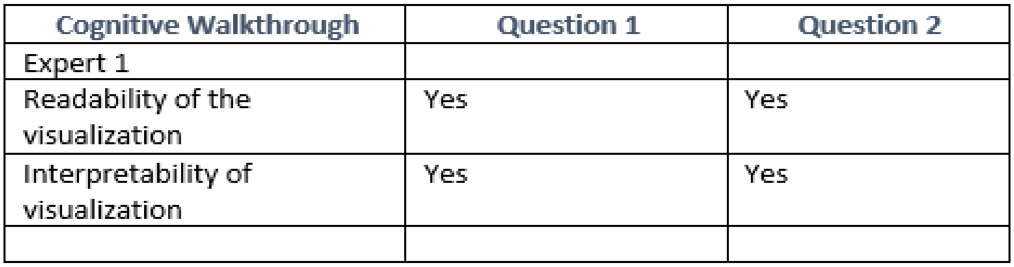
Participant ID	1
Task ID	3
Overall Result	Success
Process	User looks at the line graph in the dashboard. User observes the spike in the trend for the genres in the year 2016 to 2018. User finds it easy to analyse the trend.
SEQ Score	4
Errors	-
Other Notes	-

Participant ID	2
Task ID	1
Overall Result	Success
Process	User looks at the presented charts in the dashboard. User checks the labels and header signifying the content related to genre. User observes the distinguishable colour scheme portrayed in the visualisation for each genre. User finds it easy to analyse the data presented.
SEQ Score	5
Errors	-
Other Notes	-

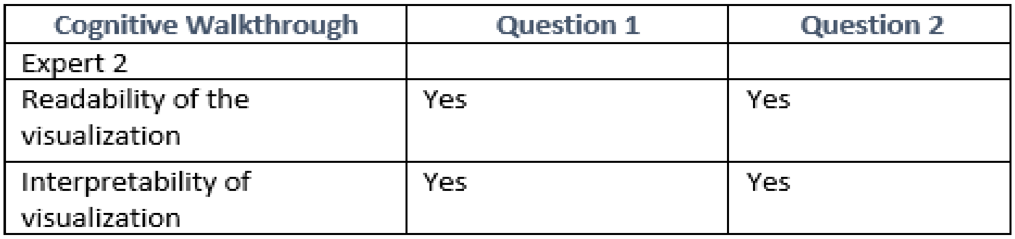
Participant ID	2
Task ID	2
Overall Result	Success
Process	User looks at the pie chart, tree map and bubble chart in the dashboard. User identifies the distributed regions in the graphs and charts. User finds the popular genre from the visualisation. User finds it easy to conclude his thoughts.
SEQ Score	4
Errors	-
Other Notes	-

Participant ID	2
Task ID	3
Overall Result	Success
Process	User looks at the line graph in the dashboard. User observes the change in trend in the later years signifying a rise in the genre availability in the period from 2016 to 2018. User finds it easy to analyse the trend.
SEQ Score	4
Errors	-
Other Notes	-

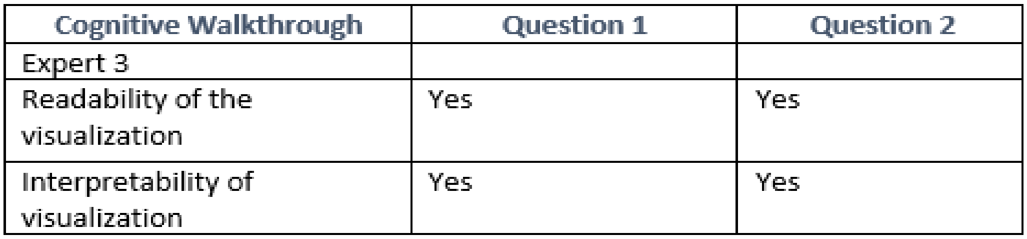
11.3.7 Task 7 – Cognitive Walkthrough



Cognitive Walkthrough	Question 1	Question 2
Expert 1		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes



Cognitive Walkthrough	Question 1	Question 2
Expert 2		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes



Cognitive Walkthrough	Question 1	Question 2
Expert 3		
Readability of the visualization	Yes	Yes
Interpretability of visualization	Yes	Yes