# COMP40610 Visual Exploration Tool Design Document

Student Name: Sanad Masannat

Student Number: 24217734

Title:

Average IMDb rating & Engagement based on Chosen Factors

Screenshot:  
A close-up of a graph

Description automatically generated

Dataset overview:

The dataset I have chosen is an adaptation of the Tidy Tuesday dataset, specifically the [13th of July](https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-13) dataset. The data here is all about Scooby Doo episodes and contains many features which include but not limited to: Series Name, network, season, IMDb score, engagement, air data, run time, guest characters, each of the characters voice actors, if they caught or unmasked the monster, features about the monster( type, subtype, gender, motive, if it was real, gender and if it had caught any of the characters) and how many times catchphrases were said. As we can see, there is a lot of qualitative variables here. While count of catchphrases is numerical, I figured it would be better to focus on IMDb and engagement scores and see how a select few qualitative variables affect it. Thus, the factors I decided on were air date, who caught the monster and who unmasked the monster, the network it aired on and type of monster. To use those variables, I performed some preprocessing on the dataset to clean data, handle NULL by either replacing it with False or No Monster. Then I created 2 new columns, one to hold an array on how caught the monster(s) and the other on who unmasked them. Finally, I handled any spelling errors on the dataset and used interpolation to get missing IMDb scores. The new dataset created is found [here](https://raw.githubusercontent.com/SanMas99/University_Work/main/Information%20Visualisation/ScoobyDooCleaned.csv).

Design Considerations:

Overall goal: The goal of my tool was to see how specific factors and combinations of factors can affect IMDb scores. The available factors are: who unmasked and who caught the monster, years, the network and the monster type.

Scatter Graph: This is a scatter graph showing all episodes with IMDb score being plotted against the engagement score. Each network has a different colour, and the shape denotes the season. As one may have noticed from the screenshot, I have used a logarithmic y axis to show the engagement score. The reason I chose to do this was because of the large disparity between the maximum and minimum. The lack of log made majority of the data appear on the x-axis, thus difficult to visualise. As this will be the main graph in which factor selection will b displayed, we need to be able to visualize the data. There was the option of filtering data above certain values but that meant I would have to lose a lot of data before it became remotely visible.

Bar Chat 1: This is a bar graph which plots each of the networks against their average IMDb scores. I chose to sort the networks in descending order of average IMDB scores as opposed to sorting alphabetically or chronologically as the latter is very difficult due to concurrent times while the former would make showing how the network affected the IMDb score lower. There were not many other options of graph types I could use here as we needed a graph that supports plots of nominal vs quantitative variables, and a bar chart would be perfect in this regard to showcase what I aimed to here. The colour palette used was that of the previous graph

Line Graph: This is a line graph which plots change in average IMDb scores over time in years. There were 2 routes I could have taken here; the first one was to show multiple lines for multiple networks or one lie for all the years. I decided on the latter because I wanted to see the years influence on IMDb score here and because filtering by networks via interactivity would change the graph more. One major disadvantage here is how the shape ended up turning out towards the later years due to, again, multiple TV networks running episodes concurrently. However, had I went with the multiple line graph, it would make it difficult to see how ratings would have changed over the years.

Bar Graph 2: This is a grouped bar graph which plots each of the characters (if they caught or unmasked the monster as separate bars) against their average IMDb scores. Grouping the bars by character allows for a direct comparison between the characters and how it affected score. This method highlights each character’s overall effectiveness in solving mysteries and their effect on IMDb and engagement. An alternative approach could have been using to merge the two categories within a single bar, however that could change the data in a way that would’ve been factually untrue (i.e. one character having a higher total over the other when they may have caught less or unmasked less). One issue here is that the data was inflated due to me flattening the array leading a much higher count than in reality, but the order still holds.

Bar Graph 3: This is a bar graph which plots each of the monster types against their average IMDb scores. By plotting each type separately, it provides a clear view of how the audience rated episodes featuring specific monsters. The bar format makes it easy to identify top-performing and underperforming monster types at a glance due to sorting via average IMDb score. However, this approach does not convey additional details such as episode writing which is important for ratings. A dot plot here is possible but makes the visualizing of data much more difficult.

Interactivity:

The interaction I chose to use here is cross-filtering combined with selecting filters. The user can click on the monster to filter data based on the monster selection or select who caught the monster, who unmasked it or network. They can even combine any of the filtering to see specific data which matches all the filters. Using this, we can select one specific feature to focus on. When combined they can they see which combinations gave the highest IMDb score or better engagement.