**Final Project**

Jaws: The Revenge, Superbabies: Baby Geniuses 2, and The Human Centipede are all incredibly different movies. Yet they all have one thing in common, they are considered some of the worst movies of all time. Whether they be terrible children’s movies, entries into horror franchises that should have ended long before, or just movies which are genuinely unpleasant to watch, some movies blur the line between subjective taste and being objectively bad. This begs the question, what makes a movie truly terrible? In particular, is there any genre of movie which is more likely than the others to create the worst movies of all time?

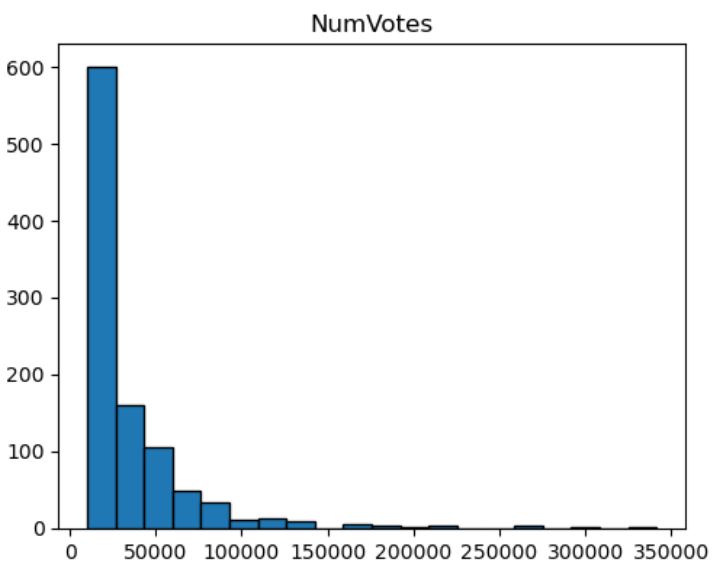
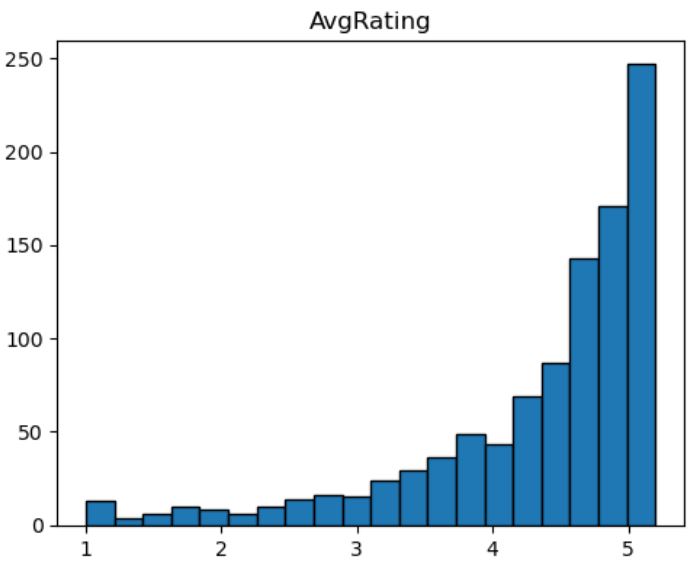
The data for this project was found on Kaggle, and it was composed of 1,000 rows and 6 columns (https://www.kaggle.com/datasets/octopusteam/imdb-top-1000-worst-rated-titles). The data showed the 1,000 lowest rated films on the film and tv-show rating website IMDB. The columns were id (the film’s IMDB id), title, genres (a comma separated list of each film’s genre), averageRating (the film’s rating on IMDB, ranging from 1 to 10), numVotes (the number of IMDB votes affecting the film’s IMDB rating), and releaseYear (the film’s release year).

In order to get a better understanding of the data a function was created which outputted histograms of each of the three numerical variables: AvgRating, NumVotes, and ReleaseYear. The histogram for AvgRating, which can be seen in Figure 1, showed that very few films are able to secure a rating of three or less. This makes sense, in order to maintain a rating that low a movie will have to be practically universally hated. Similarly, ratings got closer to 5.2, the highest rated movie in the dataset, they began to occur more.

The histogram for NumVotes, which is shown in Figure 2, was the opposite of AvgRating’s. This time the histogram was incredibly right skewed, with the majority of films having less than 5,000 ratings and no film having less than 10,000 ratings, which seemed to be the minimum amount of reviews required to be included in the dataset. In fact, nearly all films had between 10,000 and 50,000 ratings.

**Figures 1 and 2**

*Histogram for AvgRating and NumVotes*

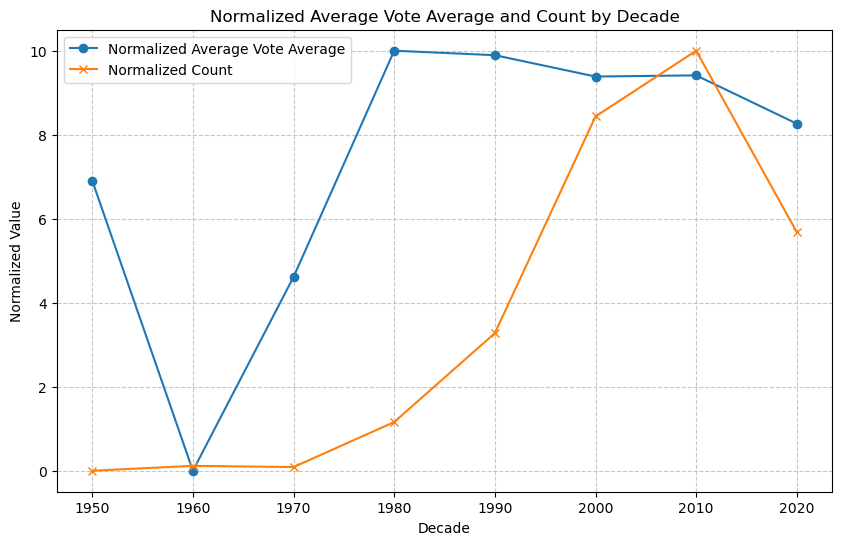
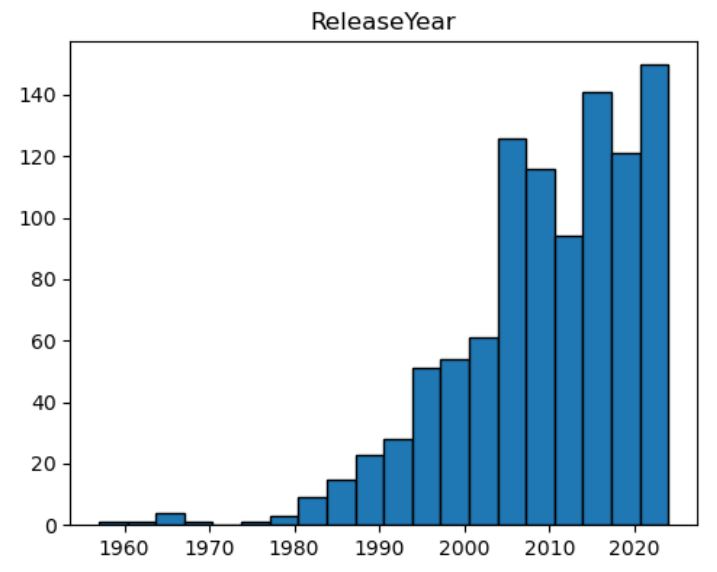


Finally, the histogram for ReleaseYear, shown in figure 3, showed that there was definitely a bias toward newer movies within the dataset. Extremely few movies within the dataset were released before 1980, with the earliest being released in 1957. A potential explanation for this is that truly bad films from the 50s and earlier may not be watched enough today to reach the minimum rating threshold of 10,000 reviews. Regardless, this is a bias of the dataset which is worth noting.

This trend was further explored in Figure 4, which shows the average rating of each decade, along with how many films from that decade are present within the dataset. This graphic showed that while movies from the 1980s and later made up the vast majority of the dataset, the average film from earlier years had a quality much lower than in later years. This may suggest that only films who are notorious for their low quality are the ones able to reach the 10,000 rating minimum.

**Figures 3 and Figure 4**

*Histogram and Line Graph for ReleaseYear*



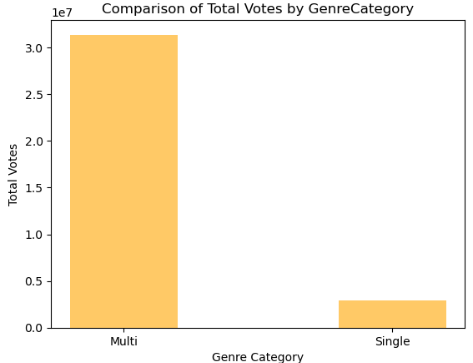
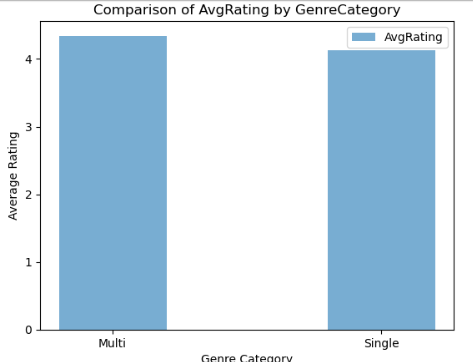
The data used in this project was mostly clean, though some slight modifications had to be made to prepare it for modeling. The first thing done was to drop the columns id and title from the dataset, as they had no predicting power. At this point null values were checked for, though thankfully none were present within the data. The larger change made was to split and one hot-encode the genres column, so that it was changed to be a boolean type, where each genre was its own column. Any of the newly created genre columns with less than 100 positive instances were then dropped from the dataset. After this step was taken a total of fifteen columns were present in the dataset.

**Methods**

To analyze how a movie having multiple genres may impact its quality we split them into two categories, Single and Multi. Bar charts were then made for the average score and total votes for each Single and Multi, so they could be easily compared. The values for total votes were scaled so that the units were in thousands. This can be seen in figures 5 and 6.

**Figures 5 and 6**

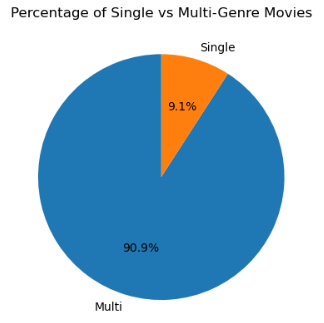
*Bar Charts for Single and Multi Genre Movies*



The results of these graphs showed that there was very little difference between the average scores of Multi and Single. However, there was a clear divide between the popularity of movies which have one genre, as opposed to multiple, with single-genres movies being, on average, much less reviewed. A further, and similar, disparity is shown in Figure 7. This figure shows that in addition to single genre films being unpopular, they are also relatively rare, making up only 9.1% of the dataset.

**Figure 7**

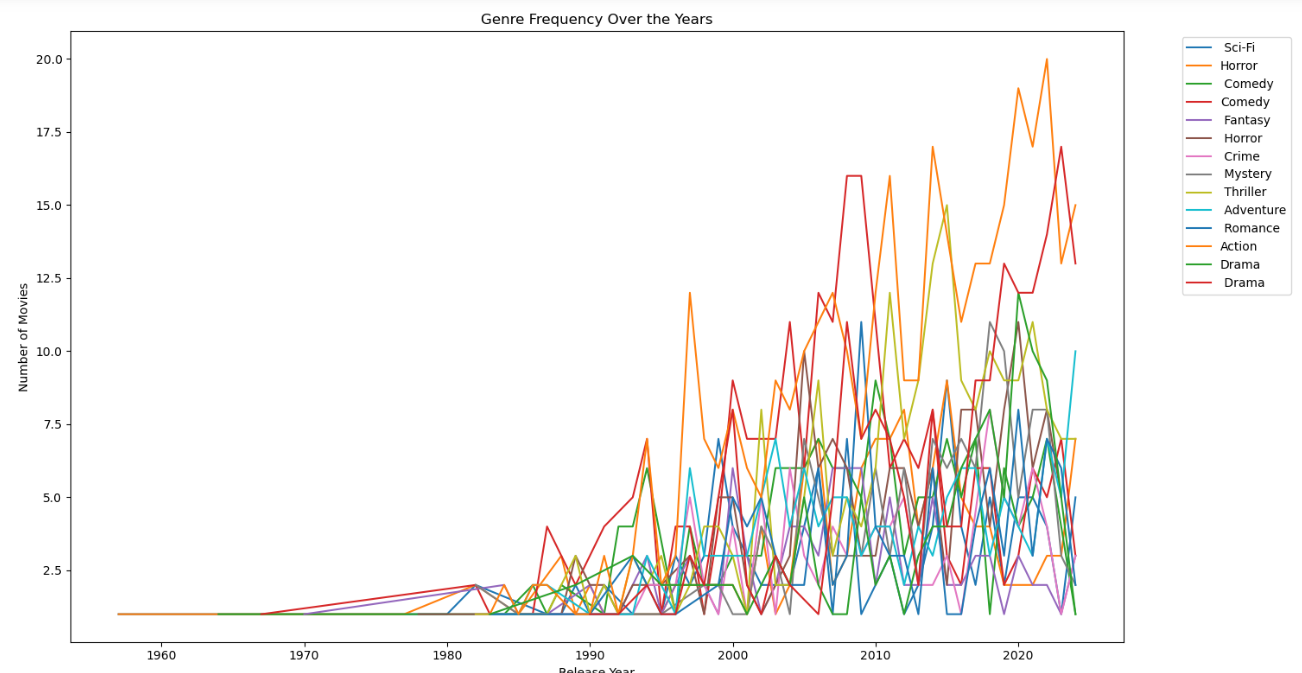
*Pie Chart for Single and Multi Genre Movies*



Another important factor which needed to be investigated was the popularity of genres over time. In order to do this a line graph was created which showed the amount of films of each genre which are present within the dataset, per year. This can be seen in Figure 8.

**Figure 8**

*Line Graph for Genres over Time*

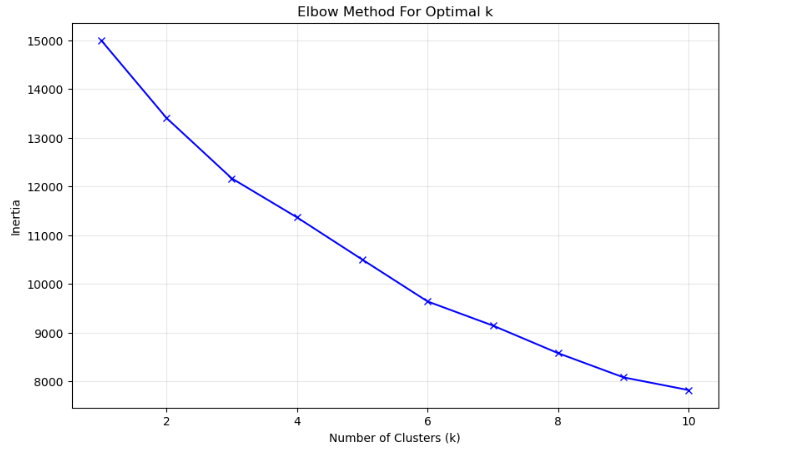


This graph shows that action movies are the most appearing movie type, especially recently. Also common were mysteries, fantasies, sci-fis movies, horror movies, and comedies. One potential reason that comedy films may be rated low is because comedy is very subjective. So even if the movie is well produced and acted, if the comedy falls flat for a particular person then they will rate it lower. Musicals and documentaries may appear less often in the dataset because there are less musicals created and the ones which do release generally require much more production and investment to be made. Therefore, it makes sense that the average quality of a high budget production, like a musical, would be greater than genres such as comedies or war films, which can be created with next to no budget. Additionally, there may be less documentaries or sports movies within the dataset because they are more so telling a story which has already happened, as opposed to having to create a new one.

To determine how many centroids should be used within the clustering algorithm an elbow graph was created. This graphic can be seen as Figure 9. Because of the size of the dataset and the fact that there is no clear drop in the elbow curve we decided to settle on five for the cluster count.

**Figure 9**

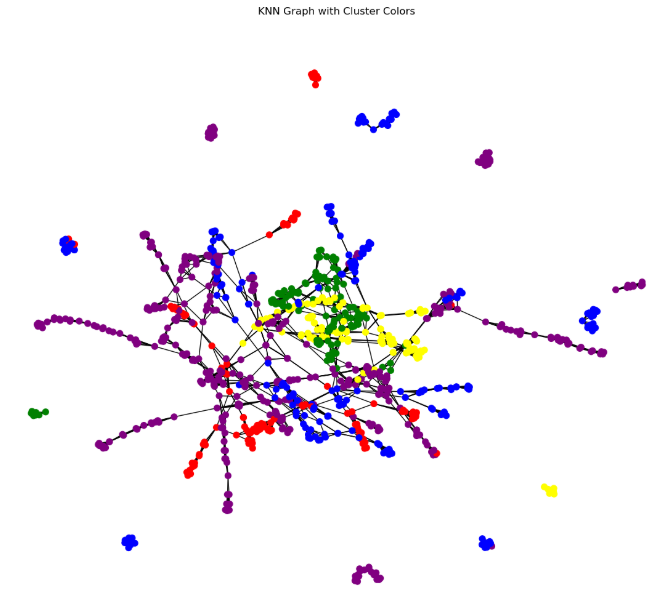
*Elbow Graph for Optimal k-value*



At this point the initial cluster model was created. This model outputs a physical representation of the cluster, which can be seen in Figure 10.

**Figure 10**

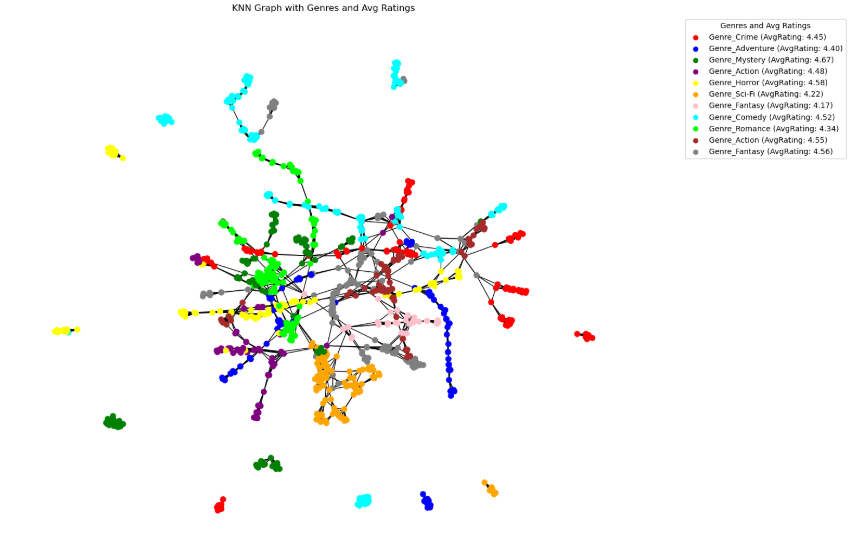
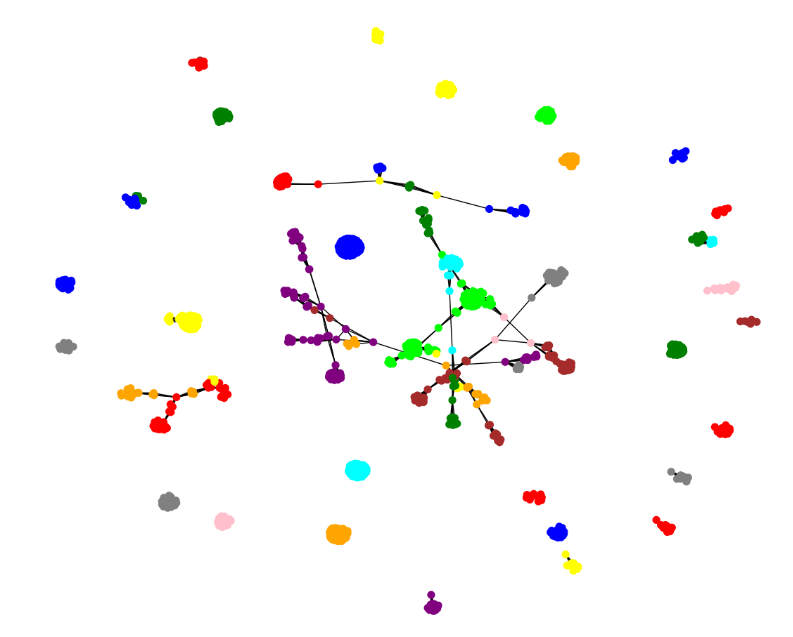
*Cluster Graph 1*



This initial clustering output had some strong clusters, notably around the edges, however, it still had a fair amount of overlap between the different clusters. In the hopes of creating a more accurate model with tighter clusters a few more clustering models were created, with slightly altered settings. These initial attempts can be seen in Figures 11 and 12.

**Figures 11 and 12**

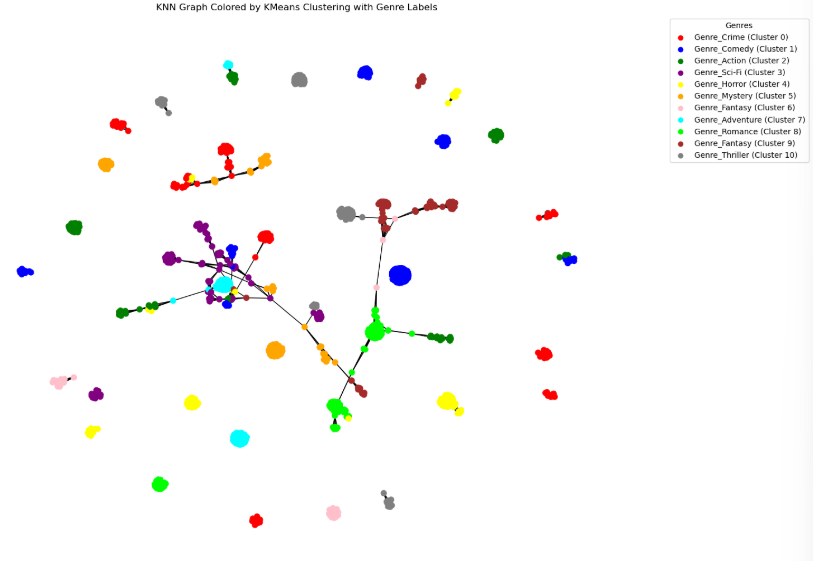
*Cluster Graphs 2 and 3*



These newer clusters, especially cluster graph 3, tended to be tighter, likely because the amount of centroids was increased. One final attempt to create the optimal clustering model was to cluster each datapoint by genre intentionally, to test how important genre is. This final clustering output can be seen in Figure 13.

**Figure 13**

*Cluster Graphs 4*



This final clustering output gave some interesting information, notably showing which genres tended to have connections with each other. Ultimately, however, this graph seemed to have been too separated, with too many clusters to give data which is too meaningful.

**Storytelling & Conclusion**

The insights that we gained from this project include that people tend to watch more action, romance, and comedy movies while still rating them poorly. So, despite their bad experiences with the genres they are not being turned away from them. An additional thing which was learned was that a lot more movies are released within the genres which tend to score lower, such as romance and comedies. So, despite these types of films not always being well received, people are watching them and the studios are making them. Additionally, while there may be many more lower rated movies today, the few older films within the dataset tend to score oddly poorly. One important note is that movies likely have not gotten worse overall, however, it is simply that the barrier of entry to create a film has never been lower. In the end, the stated goal of understanding if movies were getting worse or not or if genres were oversaturated was completed, as it was found that genre is in fact an important factor. One way that this study could be improved in the future would be to integrate specific user’s scores into the model, so that their individual tastes can be taken into account. This would allow for film recommenders to potentially be created at an individual level.

**Impact**

One potential impact of our project is that it can potentially help the studios and companies which create movies avoid the seemingly common movie-making pitfalls. For instance, there is a clear abundance of poorly made comedies and action films. So, if studios wish to remain off of the IMDB bottom 1,000 list they should ensure that any action or comedy films they create are made with intentionality, and that they do not simply become another failed film. Another takeaway is that users should not reward studios who continue to make uninspired, repetitive films. This study shows that action films in particular are made over and over, and frequently disappoint. Therefore, if customers really want to see more inventive films be created they should show studios their preference by no longer watching the inspired action films when they are released. Over time, this will send a clear message to studios that the previously acceptable quality of film is no longer good enough.

**Link to Code:**

https://github.com/bigbadraj/Final-Group-Project

**Sources:**

https://www.kaggle.com/datasets/octopusteam/imdb-top-1000-worst-rated-titles