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).

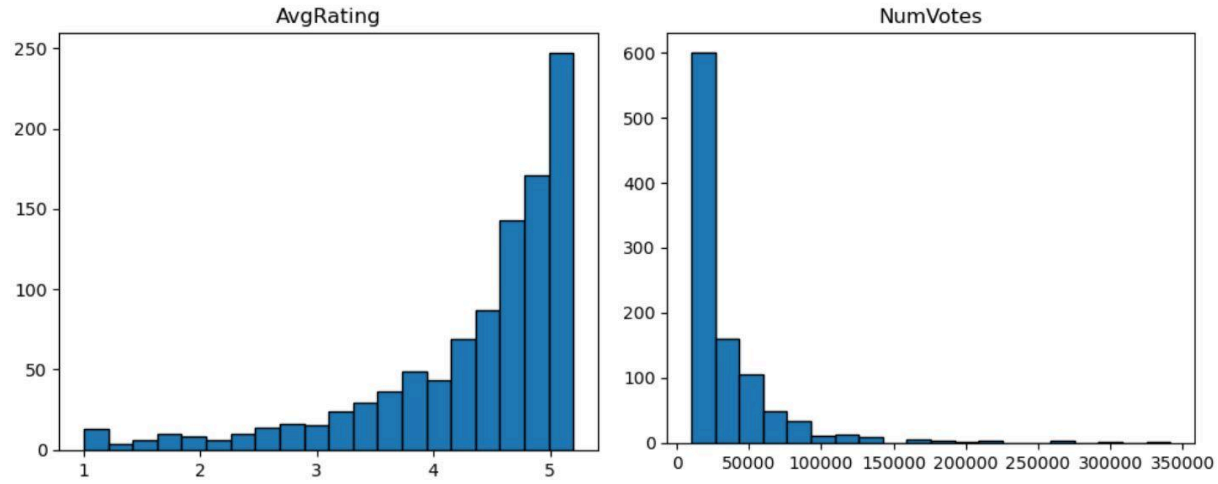
In order to get a better understanding of the data a function was created which outputted histograms of each of the three numerical variables: AvgRating, NumVotes, and ReleaseYear. The histogram for AvgRating, which can be seen in Figure 1, showed that very few films are able to secure a rating of three 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

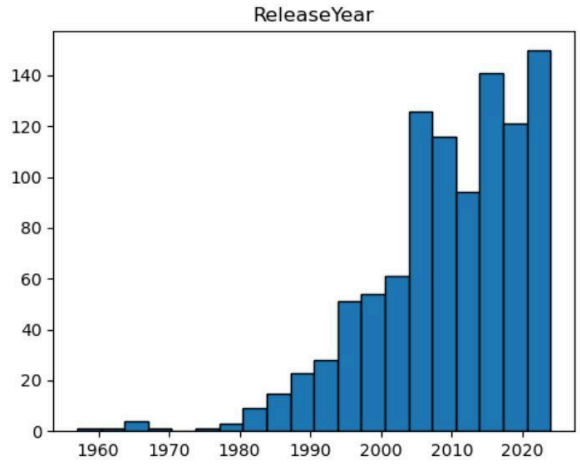


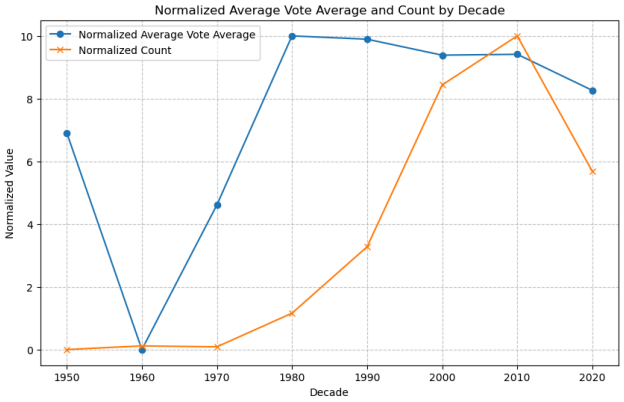
Finally, the histogram for ReleaseYear, shown in figure 3, showed that there was definitely a bias toward newer movies within the dataset. Extremely few movies within the dataset were released before 1980, with the earliest being released in 1957. A potential explanation for this is that truly bad films 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





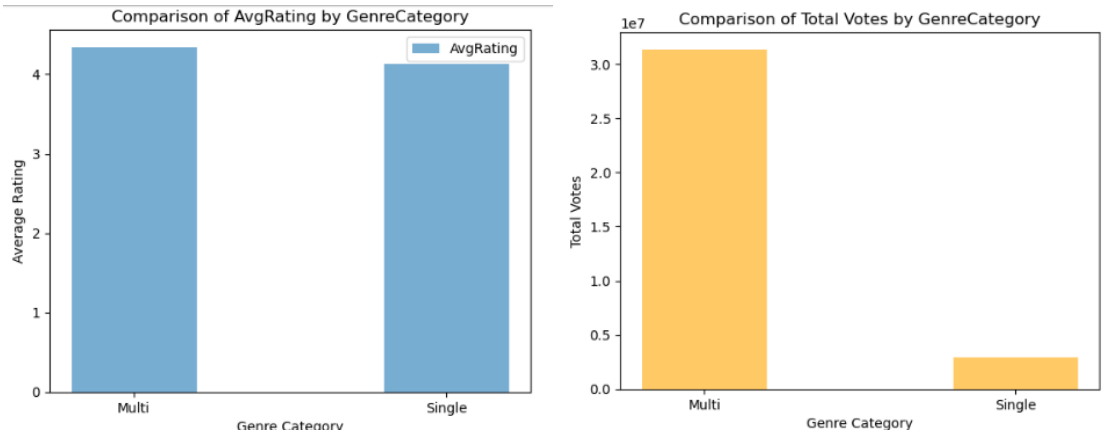
The data used in this project was mostly clean, though some slight modifications had to be made to prepare it for modeling. The first thing done was to drop the columns id and title from the dataset, as they had no predicting power. At this point null values were checked for, though thankfully none were 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

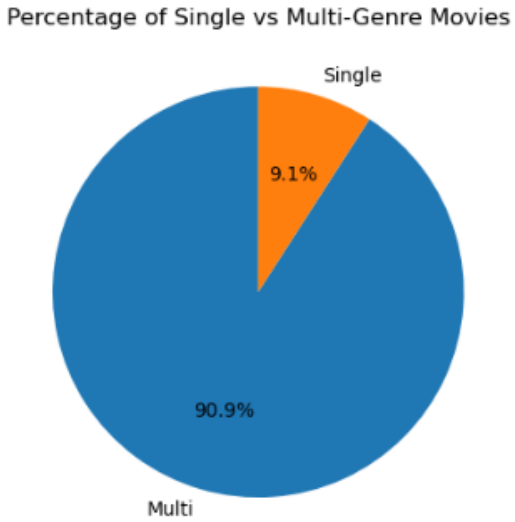
To analyze how a movie having multiple genres may impact its quality we split them into two categories, Single and Multi. A bar chart was then made of the average score for each Single and Multi, so they could be easily compared. This can be seen in figure 5. The results of these graphs showed that the

Figures 5

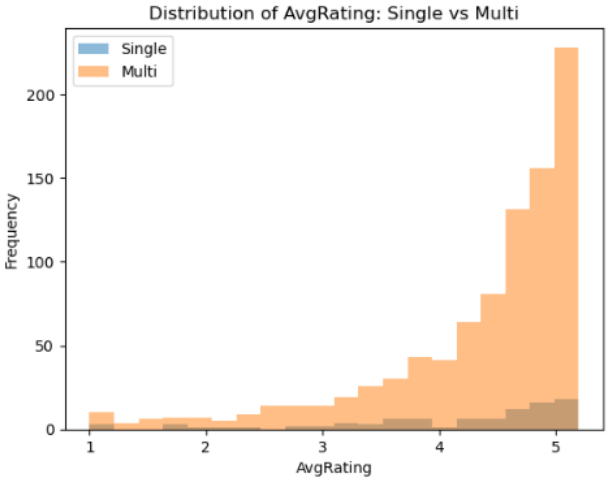
Bar Charts for Single and Multi



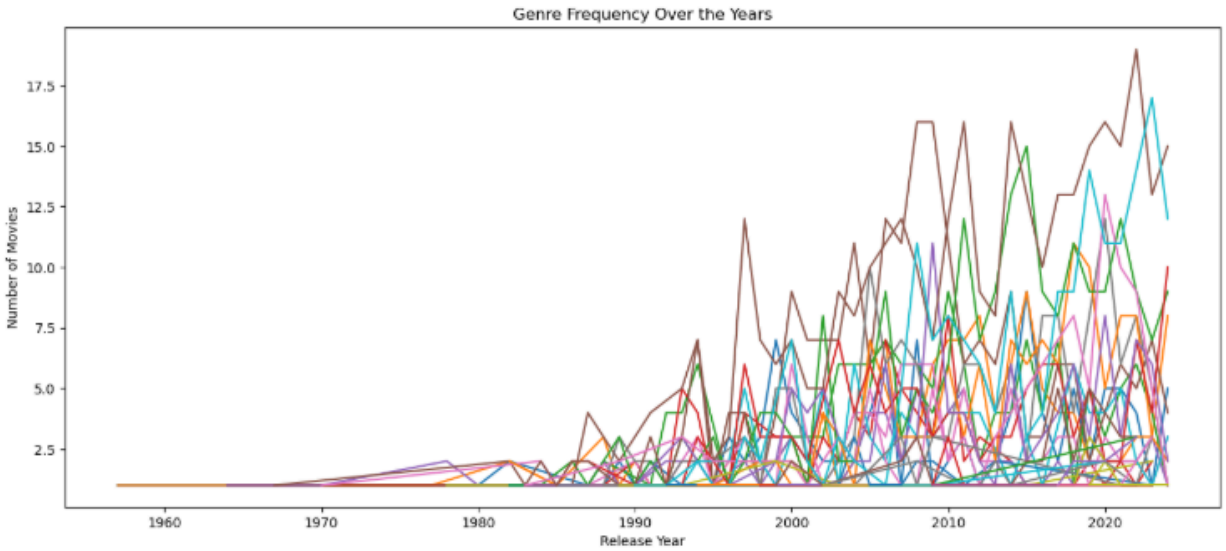
However, this revealed that there was no major difference between single and multi-genre movies when it comes to average rating however there was a massive discrepancy in the number of votes that single and multi-received. The reason for this is because 90.9% of movies that are



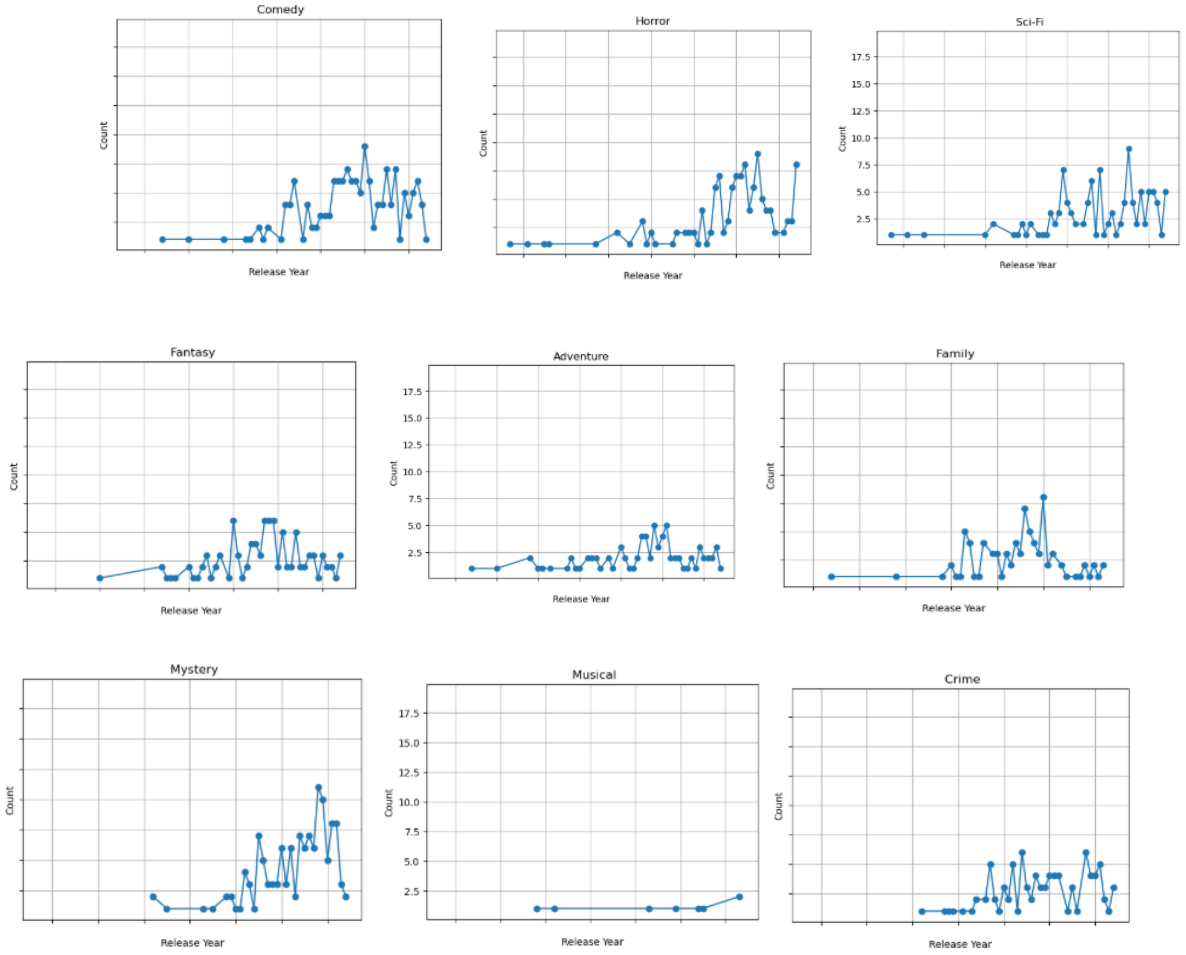
rated low are multi genre versus 9.1%.

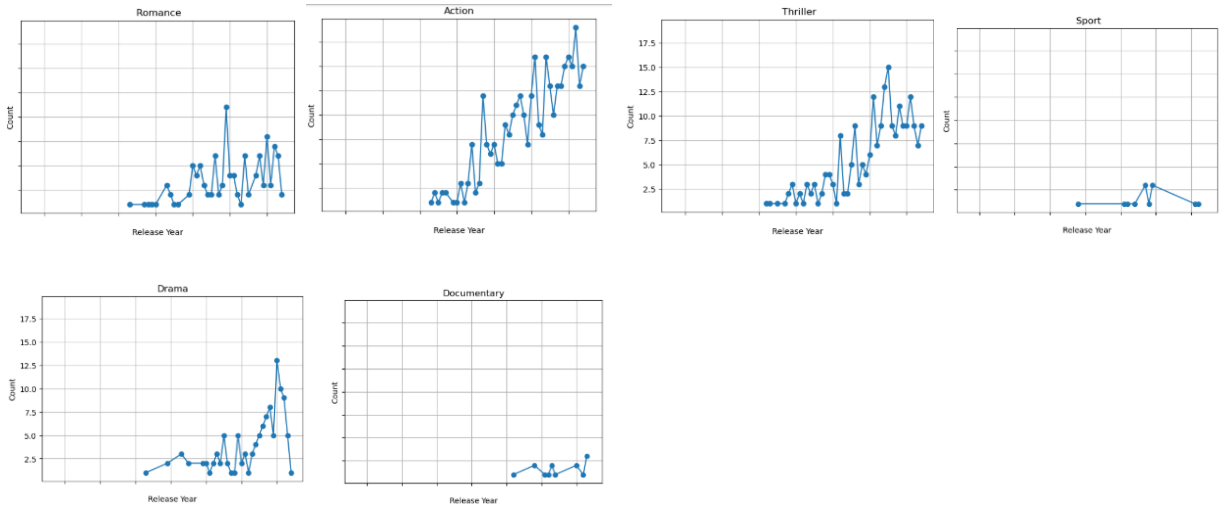


Next to show the popularity of these games over time we generated a graph that would show genre frequency over the years.

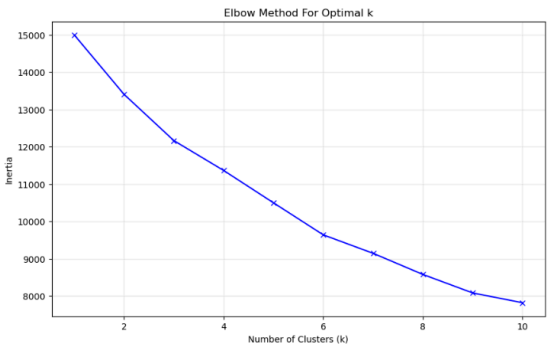


This was pretty difficult to see so we split the genres up onto their own charts to better be Able to see the trends of each genre over time.

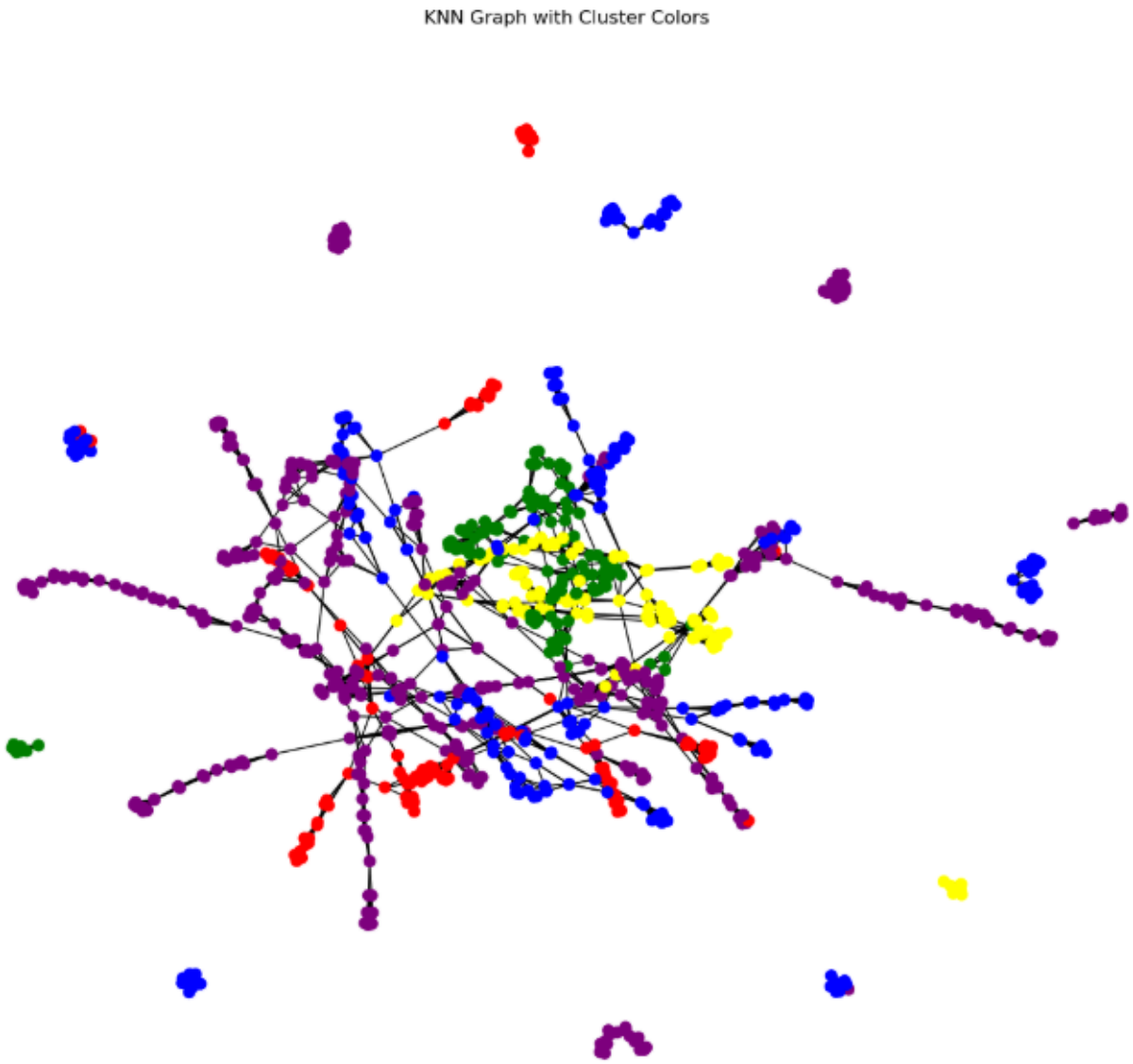




We saw here that action movies had the most amount of low rated entries. followed by mystery fantasy sci-fi horror and comedy. Romance and comedy could be rated low because they are often paired together, and comedy is very subjective. So even if the movie is well produced and acted, if the comedy Falls flat for a particular person then they will rate it lower. Musicals and documentaries made have less low ratings because there are less musicals that are released and it requires a lot more of a production and investment to create the movies that are based on musicals or musical movies so you're more likely to get a higher quality product versus a single war movie or another we can just pick up your phone and start recording. documentaries may have less movies that are rated low along with sports because you are just telling what happened so lower the documentaries just rely on are you conveying the information poorly rather than pure artistic Merit. Next an elbow

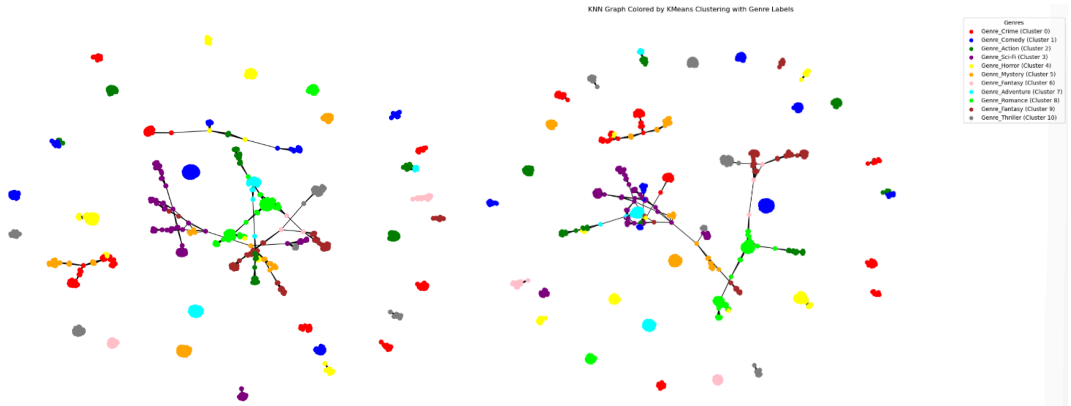


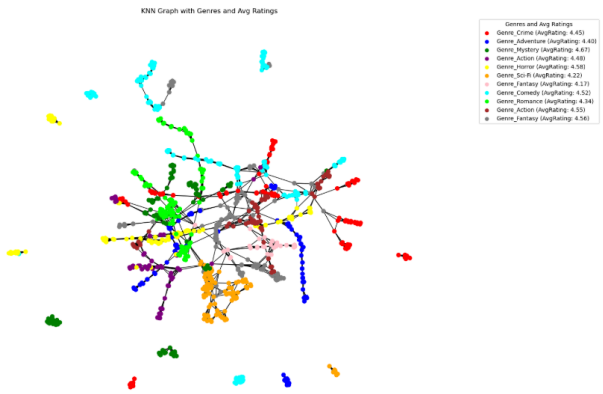
Because of the size of the data and the fact that there is no massive drop in the slope we decided to settle on five for the cluster count. next we built the cluster.



Due to the amount of overlap present in the graft like we do to only measure similarities between different groups it was a bit hard to read but there wasn't an overarching message that was clear and that was there isn't that much of a difference between the different types of movies as far as the Rings go. There is a lot of overlap so there is no singular calendar that works on roses that are overwhelmingly negative in their rates. We made a few more clustering models,

altering some settings to see if we could parse out and notice some more trends. We made a graph that clustered based on the number of genres rather than choosing five for example and colored them based on the genre, then also labeled them all based on the genre there and tried that. Due to the small Size of the data set and not having that many categories or variables it did not offer too much variance in the results of the cluster model, but it is still giving us information like the spread of certain clusters based on their scores.





Storytelling & Conclusion

The insights that we gain from this project where that people watch more action romance and comedy movies but also rate them lower. So, they are not being turned away by repeated failure to the genre I am assuming because it is one of the more popular genres, so people do not

really get tired of it they just give more movies ratings. What we learned was that a lot more movies were released that score lower in genomes like romance and comedies overall. There may be lower rated movies today, but they are not scoring worse than movies from the past. There are just more movies being released into the lower barrier of Entry, so movies are not necessarily getting worse, that's just more bad movies. The outliers do not have a strong genre correlation as they are very and feature all genres. We obtained our initial goals of understanding if movies were getting worse or not or if genres were oversaturated. I think what we could do better next time is rather than create a customer model for a data set that is this limited variable to predict movies of certain genres would fail or how likely they were to score lower. a project on what movie genre would be rated the lowest of all time or how low would it rate with the next release.

Impact

The potential impact of our project is That those were more risk averse this card for making romance comedy or action films. They make sure that they are more likely to fail or that the movies will be lower rated. There is also the worry that they may be contributing to an oversaturated genre. feel like they don't want to create for that genre. there may also be continued conversations from customers/users about oversaturation in genres that have more films that score low so people may look at genres like action or romance and think that they do not need to see movies like that or they are not worth watching because there's already so many.

GitHub