Analysis of Movie Data over Time

Grecia Plasencia Acosta, Lionel Dsilva, Joseph Griffin

Week 2/3 Group Project

***Abstract*— This document displays findings from an aggregate of movie datasets. Multiple subsets of data were merged to analyze movies by genre and ratings over time from just after 1900 all the way to 2018. Comedy and drama have been the most popular genres, number of movies produced was found to increase over time, number of ratings was found to increase at first but decrease in recent years, average rating was found to decrease over time, and positive correlation was found between multiple genres as movies can be labeled as more than one genre. These relationships were visualized using different graphing methods and further explored using the Tableau software.**

***Keywords*— Tableau, Visualization, Movies, Ratings, Decades, Graphs**

I. Description of the Datasets

This document puts forth the relationships found when analyzing movie datasets. The dataset we accessed from GroupLens [2] was compiled by randomly selected movies released through 2018. Additionally, the number of reviews provided for each movie was also random. Each movie has at least one review, but the number varies per movie. Our analysis operates under the assumption that the population size is representative of the entire population. Due to the nature of the dataset, it may be possible that this data is not representative of the population as a whole. However, there is a large sample size here that would suggest it is indeed.

The dataset used had four different subsets, each containing different information about the movies. These subsets were general movie information, user ratings, tag information, and link information. Of these, two subsets of the dataset were used, namely the ratings and movies subset. The first subset contains information about each movie present. The second subset contains movie ratings left by each user. Subset contents are represented in the tables below.

Table 1: MOVIE TABLE

Variable	**Type**	**Note**
movieId	Integer	Unique key across all subsets
title	String	Movie name with release date
genre	Sting	Movie genre; each movie has different multiple genres; separated by ‘|’ symbols.

Table 2: RATING TABLE

Variable	**Type**	**Note**
userId	Integer	Unique key across rating and tag

		subsets.
movieId	Integer	Unique key across all subsets.
rating	Integer (Float)	Rating for each movie, if multiple ratings are left they appear as multiple values in a single column.
timestamp	Integer (Time)	Time at which rating was logged.

Table 3: TRANSFORMED DATASET

Variable	**Type**	**Note**
movieId	Integer	Unique key across all subsets
title	String	Movie name
year_released	Integer (Float)	Year of release; extracted from title.
(no genre listed)	Boolean (T/F)	Dummy variable for movies with no genres listed.
Action	Boolean (T/F)	Dummy variable for Action genre.
Adventure	Boolean (T/F)	Dummy variable for Adventure genre.
Animation	Boolean (T/F)	Dummy variable for Animation genre.
Children	Boolean (T/F)	Dummy variable for Children genre.
Comedy	Boolean (T/F)	Dummy variable for Comedy genre.
Crime	Boolean (T/F)	Dummy variable for Crime genre.
Documentary	Boolean (T/F)	Dummy variable for Documentary genre.
Drama	Boolean (T/F)	Dummy variable for Drama.

Fantasy	Boolean (T/F)	Dummy variable for Fantasy genre.
Film-Noir	Boolean (T/F)	Dummy variable for Film-Noir genre.
IMAX	Boolean (T/F)	Dummy variable for IMAX genre.
Musical	Boolean (T/F)	Dummy variable for Musical genre.
Mystery	Boolean (T/F)	Dummy variable for Mystery genre.
Romance	Boolean (T/F)	Dummy variable for Romance genre.
Sci-Fi	Boolean (T/F)	Dummy variable for Sci-Fi genre.
Thriller	Boolean (T/F)	Dummy variable for Thriller genre.
War	Boolean (T/F)	Dummy variable for War genre.
Western	Boolean (T/F)	Dummy variable for Western genre.
rating	Integer (Float)	Rating for a given movie, calculated as the mean of all ratings for a given movie.
num_ratings	Integer	Cumulative sum of all ratings received by a movie, derived from users.

II. Data Processing

The software used for this assignment was Tableau [1] and all the data was sourced from GroupLens [2]. The data was cleaned and joined with Python, and secondary visualizations of the exploratory nature were made as such. The two subsets were then joined and transformed into a singular dataset to visualize its results using Python aggregations equivalent to SQL queries of the similar sort. All visualizations are based on the relationship between movie releases, their average rating, and the number of users that rated a movie. Each is defined by noticeable trends that are given by each visualization.

The subsets used were cleaned and transformed into a singular dataset that can be used for an exploratory analysis. The two secondary subsets, ‘links’ and ‘tags’ were not used as they contain user comments and links to database pages, which is unusable for visualization purposes. The subsets were joined using the Movie ID variable as a unique key. The release year variable was extracted from the movie titles, and ratings were derived from the mean aggregation of all ratings received by a movie. Further, the number of ratings for each movie was derived from a cumulative sum aggregation of each user who was logged as giving a particular movie a rating.

III. Results

A. Genres

One of the first things that was visualized in this work was the number of movies that were in the dataset and the decade in which they were released. As previously stated the data contained years but for this work it became easier to group by decade. As evidenced in Figure 1, the amount of movies released as the decades went by grew significantly. Two of the genres that most noticeably increased were comedy and drama.

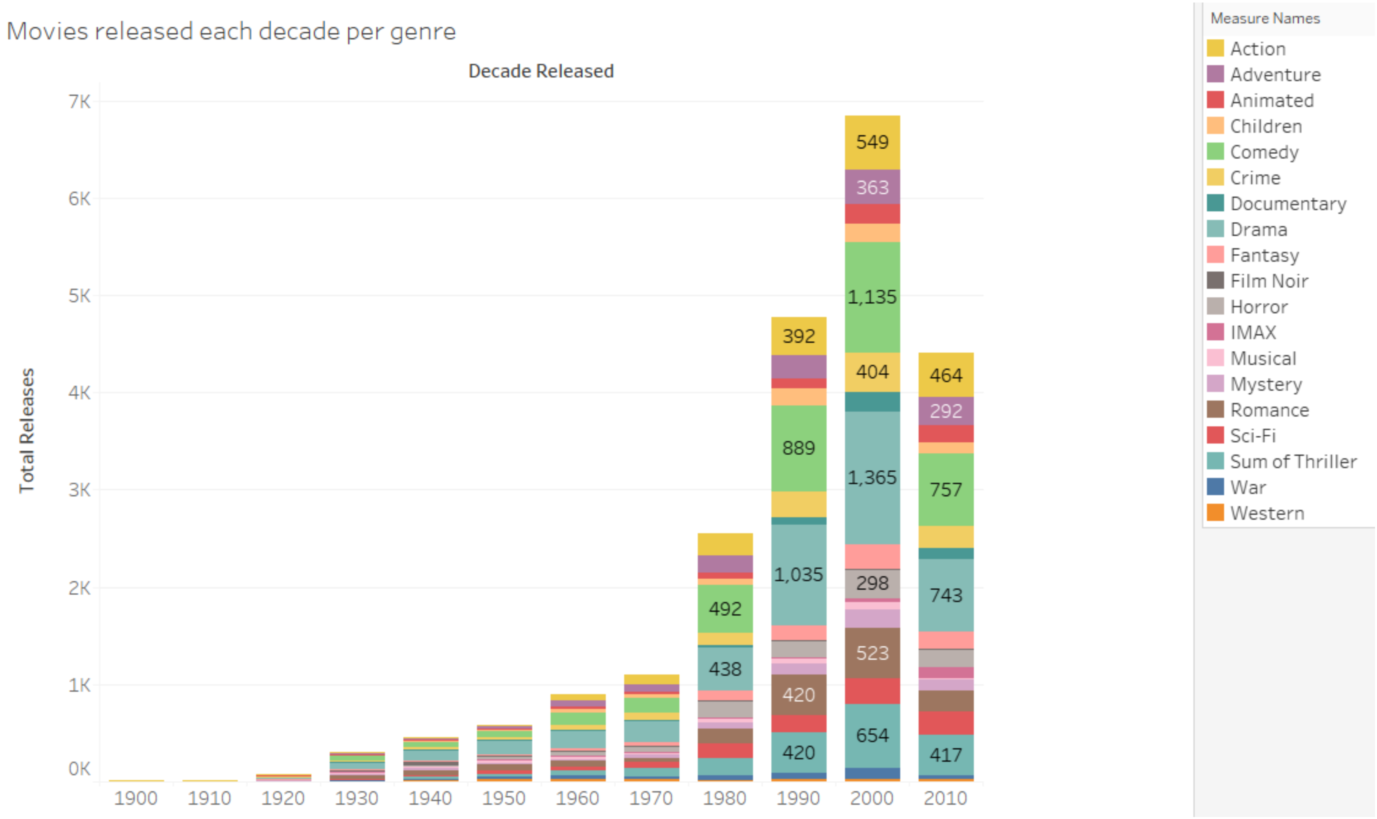

Figure 1. Movies Released Each Decade per Genre

This trend is further evidenced in Figure 2. When compared to the rest of the genres, comedy and drama movies are abundant. Plenty of movies in those genres have been consistently released throughout the decades analyzed. Figure 2 also displays which genres have decreased in production. Throughout all of the decades present in the dataset, there were only 85 total film noir movies. Film noir became emblematic of the 1940s and 1950s [3] and the data visualized supports that given that there were so few movies of that genre.

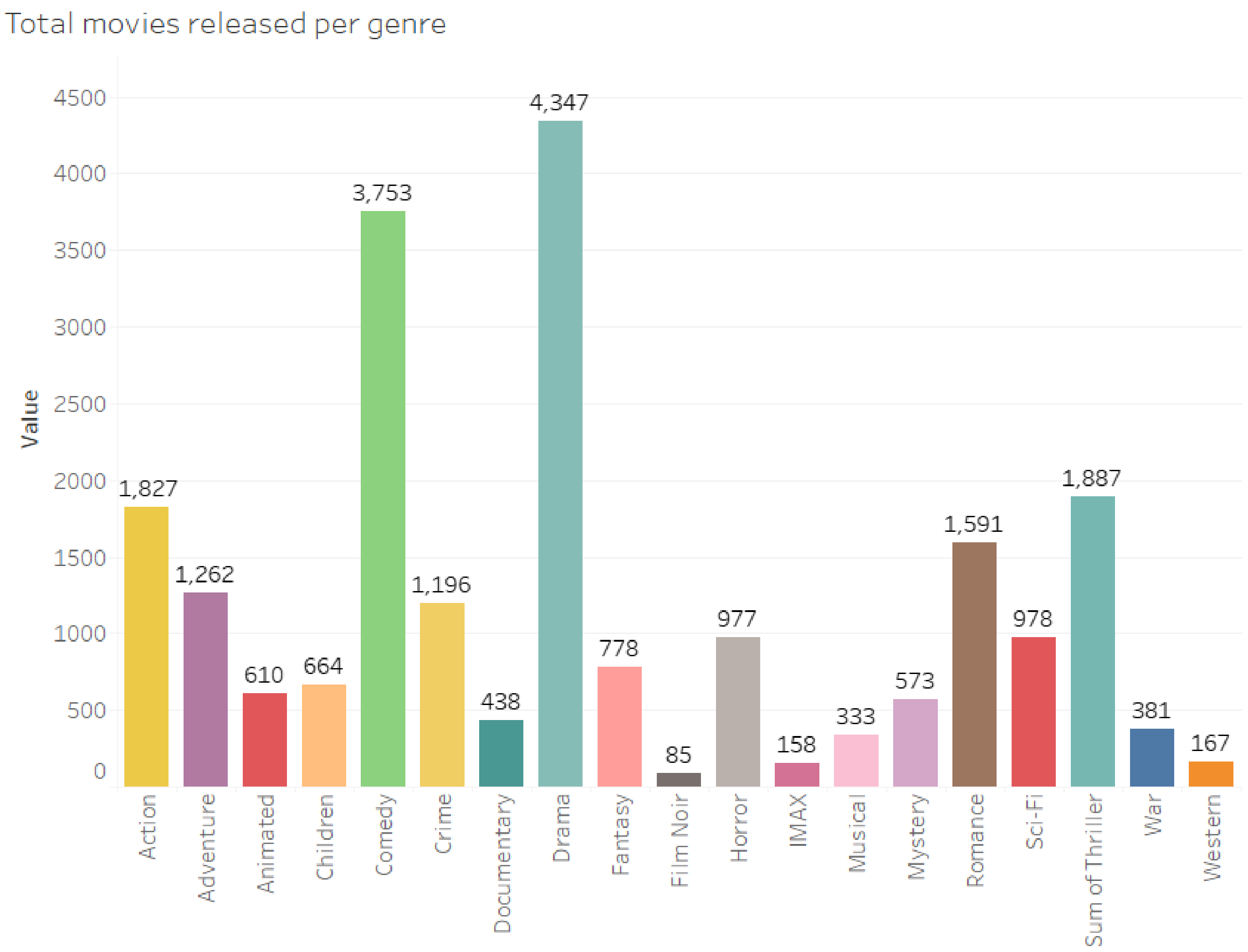

Figure 2. Total Movies Released per Genre

B. Ratings

Part of this work focused on the importance of ratings on films across the decades. Given the few movies that were released in the 1900s and 1910s there are very few reviews. It follows that the few reviews would have a higher weight. Figure 3 shows the average ratings per decade across all genres, but again keep in mind the small number of reviews for the first two decades of the 20th century. Looking beyond the first two decades in the visualization, the average rating has a definite downward trend over time. This is especially interesting given the investments in movie production costs growing massively in recent years.

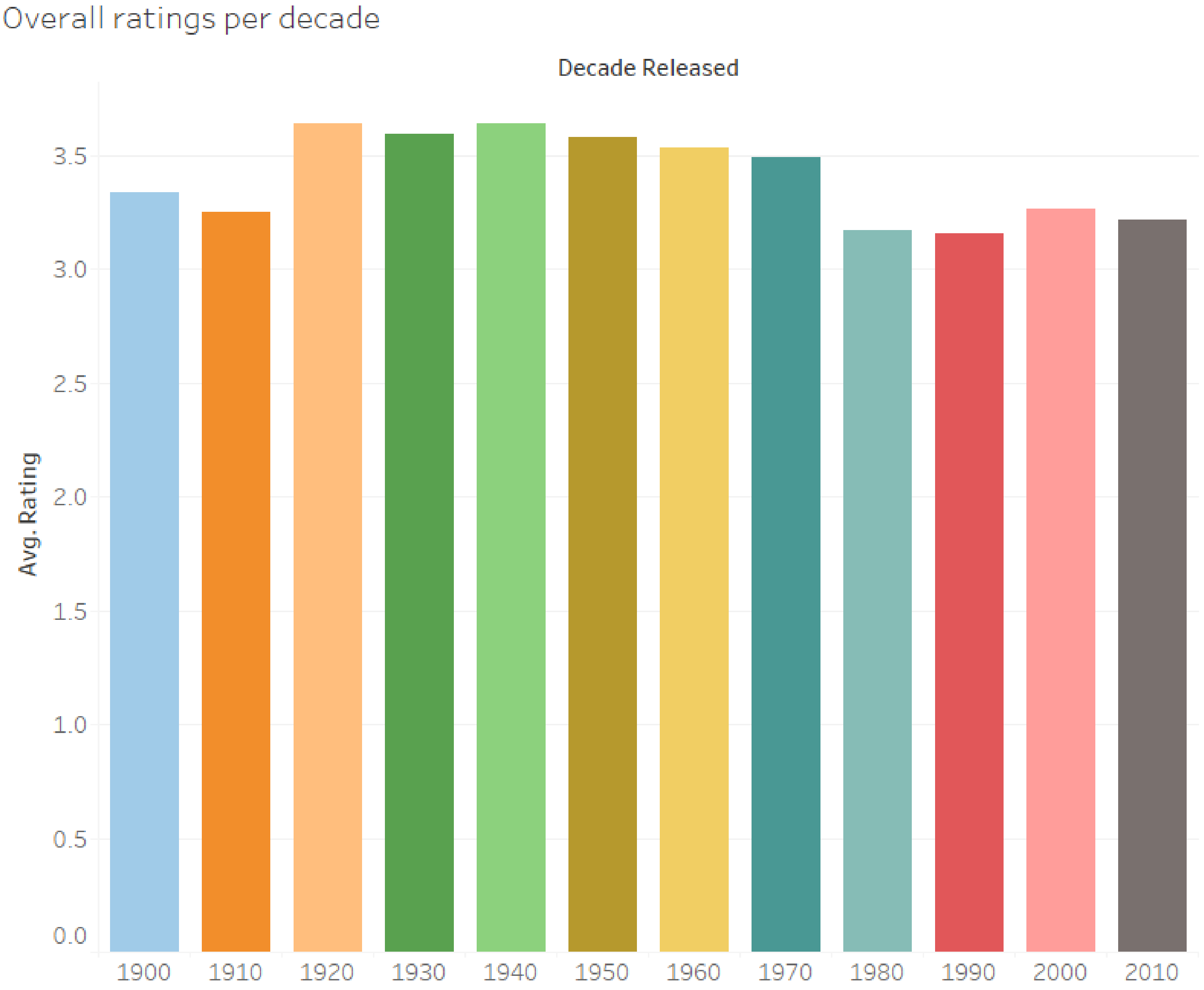

Figure 3. Average Ratings per Decade, All Genres Included

Figure 4 displays the distribution of ratings collected in the dataset. The number of ratings available for each decade is shown in this figure. There is a glaring discrepancy between the number of ratings available in the first decades of the dataset and in recent years. The major differences in the rating quantity available in the 1900s, 1910s, and 1920s becomes more apparent here. This number peaks in the 1980s and 1990s, as represented by the density in the colored bar for those decades. However, it is with noting that from 2000 there appears to be a decrease in the quantity of ratings available. The early increase likely coincides with the growth of the movie industry and the increased number of films. The drop after the 1990s likely coincides with increased internet accessibility. Perhaps as mass quantities of information about movies became more and more available via the internet, people began to feel less of a need to share their thoughts on the movie.

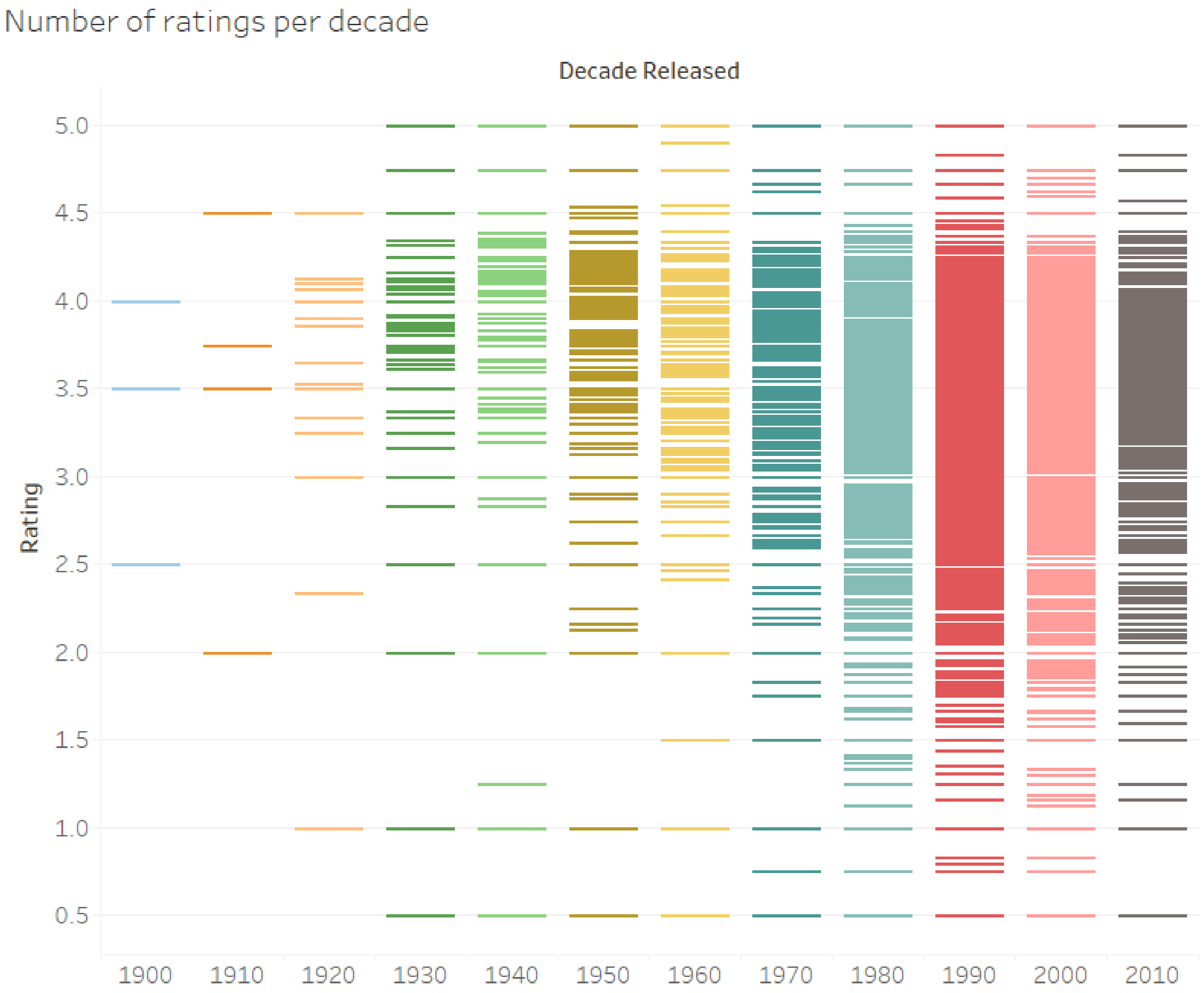

Figure 4. Number of Ratings per Decade

A combined depiction of the average ratings and number of ratings over time is given in Figure 5. Once again, the inconsistency of ratings due to small sample size in the early decades is made evident. After this, a much more consistent downward trend in average rating can be seen. There is a slight uptick when approaching the present day, but this also coincides with a decrease in reviews and a decrease in movies in the dataset, particularly in the last decades. This will be further evidenced in Figure 6. The more dramatic curve can be seen showing the number of ratings. This number continues to increase until reaching a distinct peak in the mid 1990s. Interestingly, there is no real change to the average rating slope around this time. Following the peak, there is a large drop in the number of ratings, and this continues through the end of the dataset in 2018.

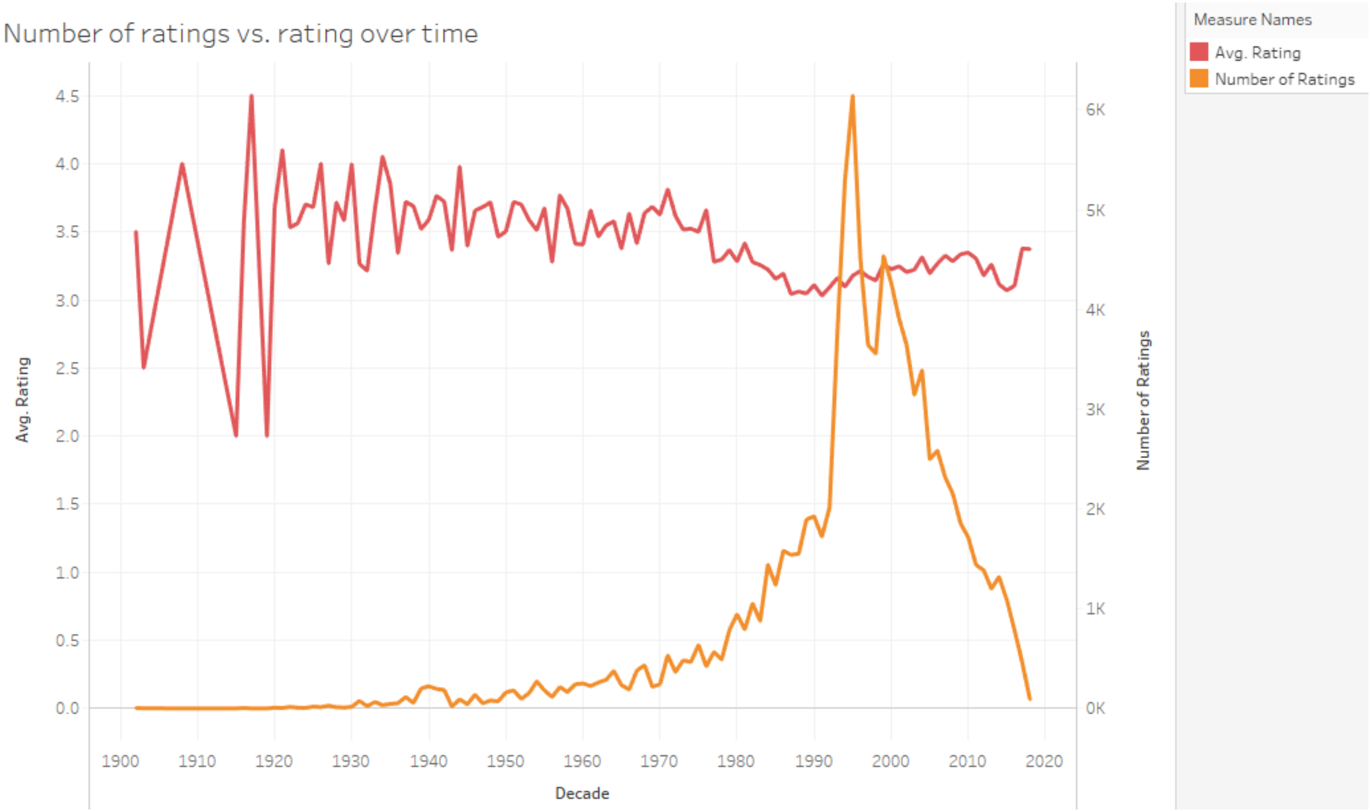

Figure 5. Number of Ratings vs Rating over Time

C. Production

Another part of this analysis focused on the production of films. It is established in Figure 6 that the production of films continued to increase into the 1980s. This aligns with the data in Figure 1, which also showed a great increase in movies in the same decade. In the years to follow, the number of movies remains high. There is a noticeable drop in the last few years of the dataset, but again we attribute this to simply being the edge of the dataset and not being representative of the entire population. This drop was also seen in Figure 1, however it is displayed in more detail in this figure.

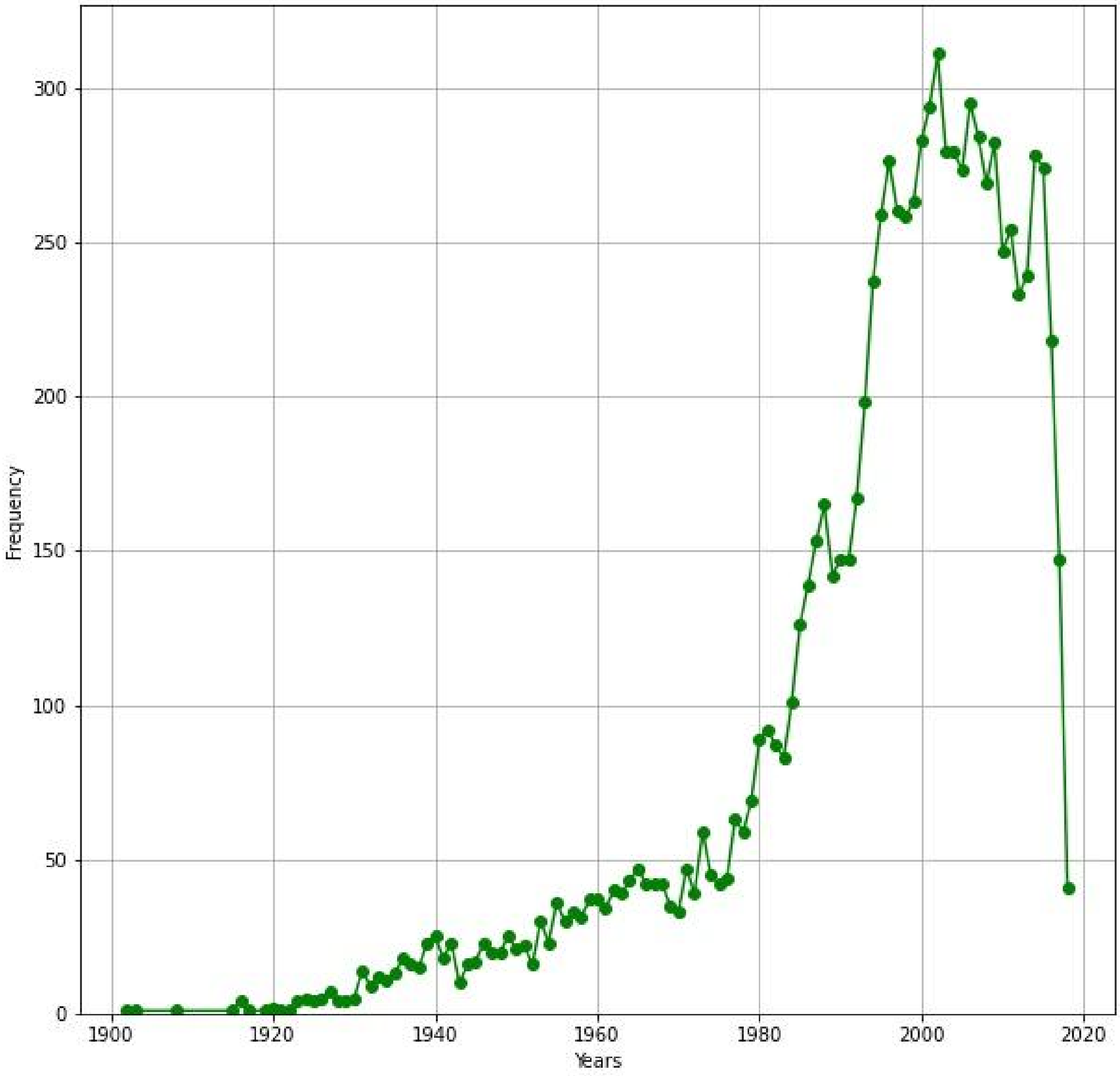

Figure 6. Total Movies over Time

Changes in film production have a direct relationship with audience demands. Shown in Figure 7 is the correlation between genres in movies. Films rarely fit into one particular genre; most fall into multiple categories. We are able to see the rates at which the different genres overlap in Figure 7. One relationship that stands out is the high correlation between children's films and animated films. On the other side of the spectrum, the low correlation between dramas and comedies is also quite noticeable.

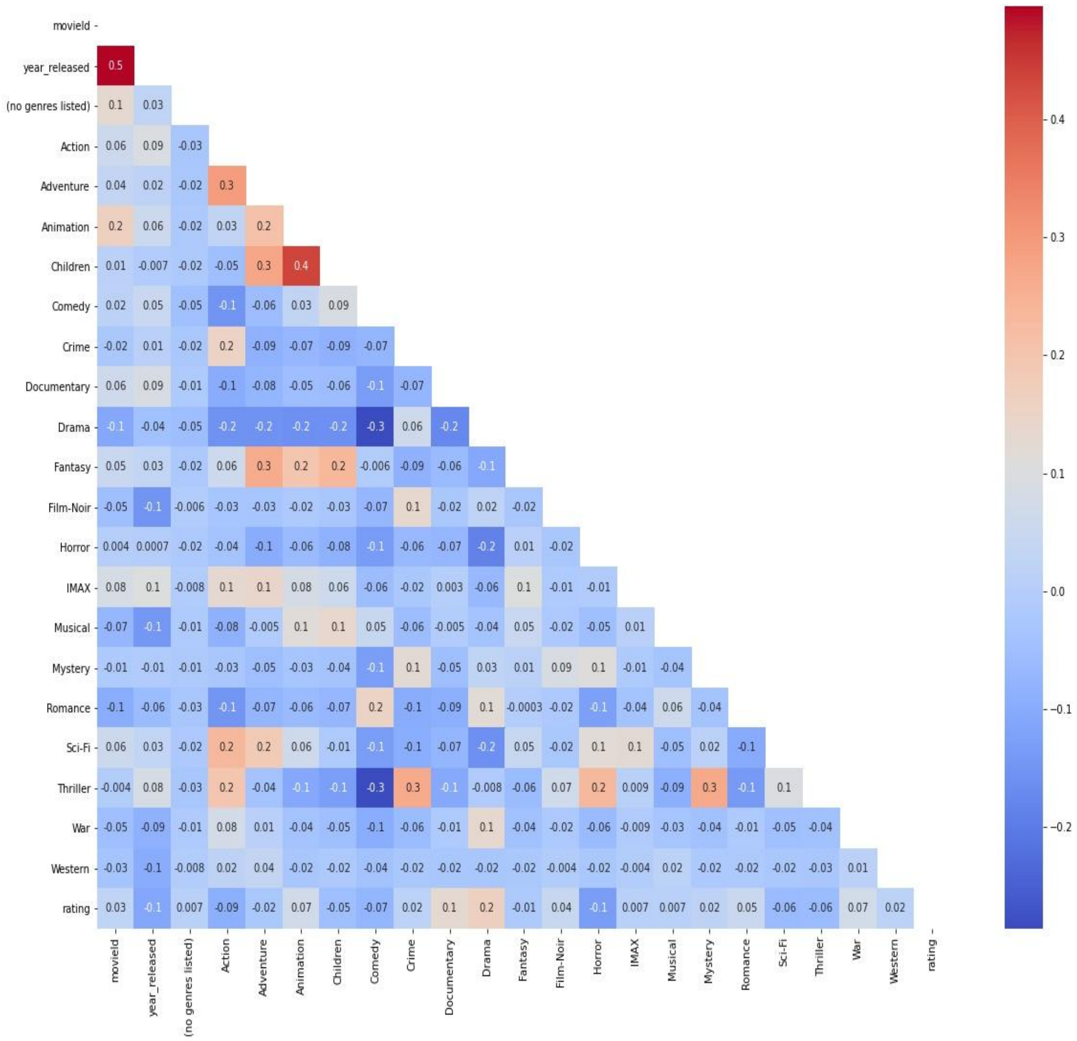

Figure 7. Genre Correlation Matrix

Lastly, complementing Figure 1, a distribution of all movies by genre is displayed over time in Figure 8. Similar trends in popular movie genres can be seen here as well.

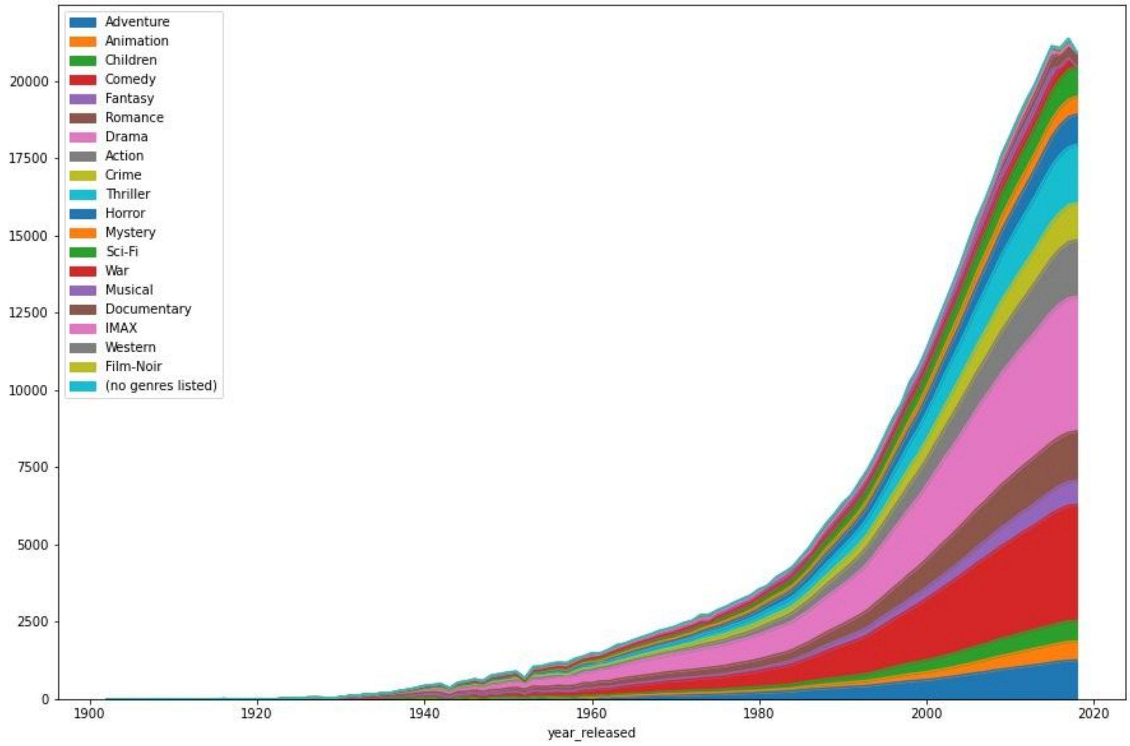

Figure 8. Area Plot Depicting Genres Produced over Time

IV. Conclusions / Hypothesis

Consecutive joins with the ratings subset allowed for a multitude of observations to be noted. We had hypothesized that total movie releases, number of ratings and average ratings would all increase over time. The hypothesis turned out to be partially supported. Other interesting findings were noted as well. We saw that comedy and drama have long been the most popular movie genres, number of movies increased over time, number of ratings increased early on but decreased in recent years, and average rating consistently decreased. Movie releases by genre can be seen in Figures 1, 3, and 8. Not only do these two genres have by far the most total releases over time, they have held their positions as the most popular genres over the last century as well.

Number of movie releases can be most clearly seen in Figure 6. There is a clear and obvious increase in movies released from the early 1900s into the 2000s. There is a drop in the most recent 10 years of the dataset, but given this small year range being on the end of the dataset, we attribute this simply that: the edge of the dataset and not data truly reflective of the entire population.

Perhaps our most interesting findings were the decrease in the number of ratings in recent years as well as decrease in the average rating over time (Figures 3, 4, 5). It is logical that number of ratings increased for essentially the entire 20th century as the number of movies released continued to increase (Figure 6), but the curious part is the severe drop just before the 21st century. Movie releases continued to increase, but ratings did not. We attribute this to the increase in internet access. As plenty of movie information and reviews became instantly available on traditional platforms and on newer platforms such as YouTube, perhaps the majority of people felt it was less necessary to share their thoughts.

The consistent decrease in movie ratings over time is also curious (Figures 3, 5). Perhaps as more and more content has been released (including more and more high quality content), people have become more and more picky and hold new movies to higher standards.

Lastly, we also found positive correlation between multiple genres in Figure 7. This finding may seem logical, but our correlation matrix is definitely a very valuable tool. We see a number of genres that are highly correlated: children and animation, adventure and action, children and adventure, fantasy and adventure, thriller and crime, thriller and mystery. This data has countless applications in movie recommendation algorithms, where specific genres can be grouped together to recommend similar movie types with the same tags. Netflix and other online streaming platforms continue to increase in use, and user-customized movie and show suggestions are a major component of these platforms as they try to please the customer with content they enjoy.

V. Theories and Practices Applied to Story

Throughout the work there were different types of visualization techniques used. The graphs used throughout were chosen to best represent the data being displayed. In Figure 1 a bar graph was chosen to display not only the increase in the amount of films released per year but also to provide the amount of films released per genre during each decade. It was the best choice given the ordering and ranking that were shown, the principle of closure was displayed. Figure 2 was a labeled bar graph to easily display totals, the nature of the data in this graph was quantitative thus proper labeling was exhibited. The similarity of the data can be easily seen in this figure. Average ratings also came across easier in a bar graph, there was less detail necessary for this graph so labels were excluded as a simplification. Figure 4 excluded labels and detail, focusing on relative size instead. The proximity of data points was the point of the figure. Simplicity was also considered for Figure 5 given that there were only two variables being displayed. Hue was arbitrary but continuity was important.

Figure 6 is a simple trend line and only requires the reader to understand how many movies were produced from initial collection dates to the last updated dates - so simplicity was a driving factor in creating this visualization. Figure 8 on the other hand is slightly complex in comparison. This was created as an alternative to Figure 1 and Figure 2 - while retaining visual information, but sacrificing the precision that Figure 2 displays with the number of movies per genre. Like the aforementioned figures, colour hues and the area under the curve (AUC) serves as a visual indicator for observations. Figure 7 carries values that show us the correlation between different variables within the dataset itself, and is organized into a heatmap to make positive and negative correlations easier to discern - heatmaps, like other visualization techniques use contrasting colour hues to note observations and in this case, a dark red square shows a positive correlation and a dark blue square shows a negative correlation - lighter squares with either grey or white squares show a negligible correlation between variables as such.

References

[1] Tableau, Seattle, Washington, USA, Tableau Desktop Software, version 2021.1.1 64-bit, May 16 2021. [Online]. Available:https://www.tableau.com/products/trial Accessed on: May 22, 2021.

[2] GroupLens Research, “MovieLens 100K Dataset,” GroupLens, Dec-2019. [Online]. Available: https://grouplens.org/datasets/movielens/100k/. [Accessed: 19-May-2021].

[3] R. Schwartz, “Neo-Noir: The New Film Noir Style from Psycho to Collateral,” in *Neo-noir: the new film noir style from Psycho to Collateral*, Lanham, MD: Scarecrow, 2005, pp. ix-ix.