Movie Ratings Analysis Project:

Names: Jezlea Ortega and Ishmam Fardin

The objective of this project is to explore a movie ratings dataset from MovieLens to uncover intriguing patterns and insights, such as the average rating of movies, the most popular movies, and the connections between movie genres and ratings. The dataset had multiple csv files and provided user ratings for movies, and information on what genre each movie is.

Data Collection:

Using Pandas library we read in the movies.csv and ratings.csv files which were downloaded from https://grouplens.org/datasets/movielens/latest/

Data Preprocessing:

Using the merge method, we merged the ratings.csv and movies.csv, and removed any rows with empty cells. Using the drop method, we removed the movieId, timestamp, and userId columns to shorten runtimes for future uses.

Data Exploration:

There are a total of 58,020 movies in the movies.csv file which was found by fixating on the ‘title’ column, dropping duplicates, and then finding the length. There are a total of 19 unique genres and it was found by isolating each individual genre in the ‘genres’ string through a split function. We utilized the ‘|’ symbol separator between each genre

to split the string into multiple substrings. Furthermore, we next created a new dataframe with each genre maintaining their own respective rows. Afterwards, creating a new set with each individualistic genre name listed out separately as substrings. Due to the ‘no genres listed’ substring being ignored in this particular project, we located the index location for this unique genre, and simply deleted it from our set. The final step in this particular problem was to print out the length of our set, which would equate to the amount of unique genres. The next portion of this subsection asked us to find the average number of ratings per movie, which we calculated to be 515.700. Utilization of the groupby function made this portion relatively simple, allowing us to group each movie with their individual ratings. After grouping the title column with the rating column we simply counted the amount of ratings per movie up, using count, and found the average using mean.

Data Analysis

In order to determine the average rating per movie we once again utilized the groupby function. Again, grouping the title column with their individual ratings and deriving the average using the mean function. To calculate the top 10 highest-rated (average rating) movies with a minimum number of ratings to be considered, we created a function that takes in a dataframe (such as merge) and a number (which would be the minimum amount of ratings). The first step is to initialize a copy of the dataframe from the parameter to a variable called new_df, so that we don’t update the parameter itself. We decided to add a column called count with each row being 1 to new_df. This helps us when we do groupby and aggregate finding the average and finding the sum of count for a specific title at the same time. When we execute this , it gives us the necessary

information to find what we're looking for since we have an average rating and the total number of ratings for each movie, we just have to filter out some things. We then filter out movie titles that don't meet the minimum requirements for total number of user ratings, and after receiving the necessary results, sort_values(ascending = False) organizes the list in descending order to get the highest rated at the top. Finally, we use slicing to only look at the first index to the 9th index. Through a combination of groupby, agg, count, and sort_value methods, identification of the top ten most rated movies was possible. Progressing forward, our next problem statement asked us to analyze the distribution of ratings across different genres. We started this problem by first copying our separated genres dataframe into a new dataframe, preserving the original data for future use. Then, we fixated on the genres and ratings columns once again, using a groupby function to group the data and taking the mean to determine the distribution. As we execute this line of code, we also ignore any information associated with 'no genres listed' as per the directions. Further, analyzing the distribution we produce a horizontal bar graph with genres equally to the array, otherwise known as the x-axis, and ratings equating to the index values, aka the y axis. In execution of this visualization, it is apparent Film-Noir had the most average ratings, and Horror maintained the least.

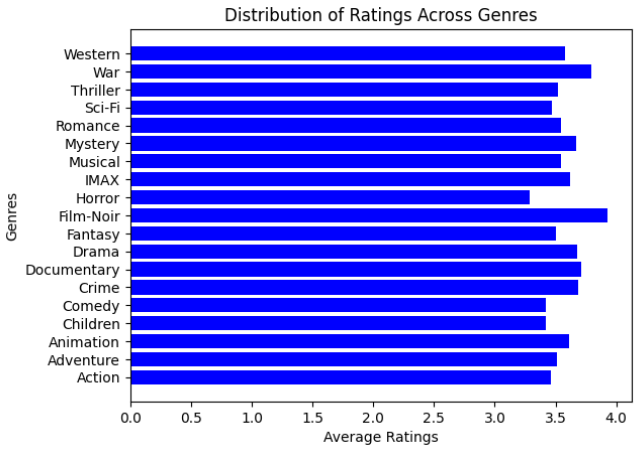
Data Analysis: Investigating the Relationship Between Release year and Average Per Genre

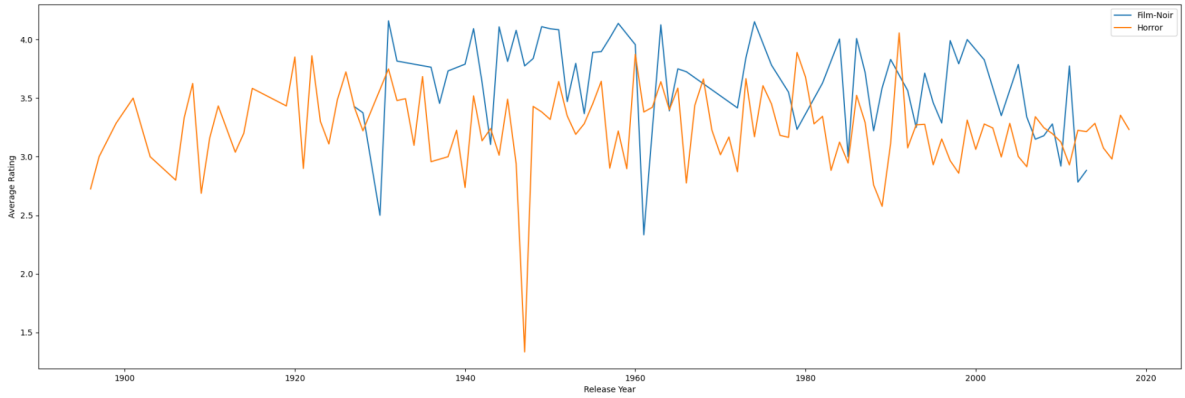
First and foremost, to begin our analysis we needed to separate the release year from its respective movie titles, giving it its own individualistic column. In order to do so, we initialize another copy of our merged data set, and use str.extract to take out solely the release year from each movie title into a variable called 'test_random'. As a result,

producing a new merged data set with ‘release year’ listed as an individual column. When we used the unique() method to check if the release years were of the correct format (only 4 numbers), we noticed ‘nan’ in the list. We used the method dropna() with parameters for the release year column and inplace as true to update the release year column to remove nan/null values. Similar to earlier, we used the split and explode methods to separate the genre’s from each row and have each row with one genre. We then created a function that takes in a genre name and uses it to find average rating in relation to release year. We create a copy of the dataframe ‘test_random’ into a variable called ‘allGenres.’ We first filter out the genres we don’t need and then drop the genre’s columns since it’s no longer relevant (since all of them would be the same genres). We then group by release year and find average rating using groupby and mean methods. Then we created a dictionary called ‘avg_rating_genre_list’ , and then we use a for loop to go through every genre in ‘unique_genres’ list (from earlier to find the number of unique genres) and set the key of the dictionary to the name of the genre, and the value of the key to a variable called ‘data’ which will hold the data we got from using the function. Finally we create all 19 graphs by creating a grid of 20 subplots and then using a for loop with zip method to go through our dictionary and the 20 subplots at the same time. In each iteration, we plot data from each genre and label axis/title on one subplot.

Film Noir vs Horror

When looking at the distribution of average ratings across all genres, we noticed a disparity between the two. So we decided to look at the average ratings per year to get a better idea of how each compares over the years. It also gives us a good idea of what information we have. In the average rating vs release year graph, we see that there are no values for Film Noir before 1920 while Horror stretches from before the 1900s all the way to 2020. Based on that, we inferred that it's possible we have a lot more user ratings / data for Horror compared to Film Noir. So we checked how many ratings we got for each genre, and found that Film Noir had 272742 user ratings in the dataset while Horror had a whopping 2070791 user ratings, which is close to 10 times more than Film-Noir. Based on that graph , we can also see that the ratings for Horror tanked a lot between 1940 and 1960 , which negatively impacted its overall average rating. The

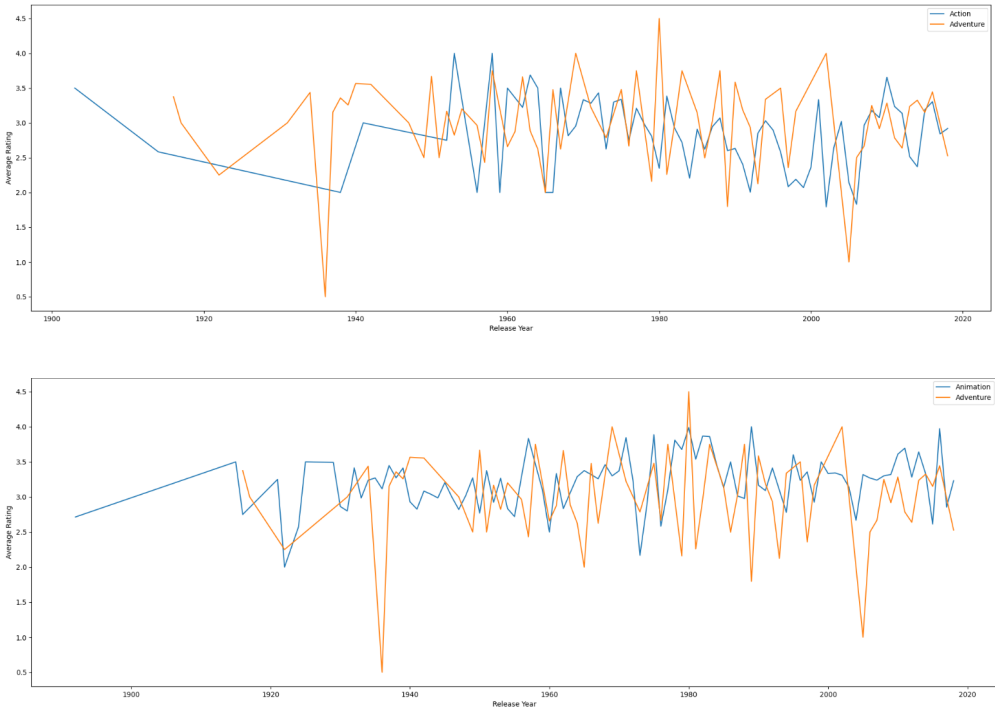


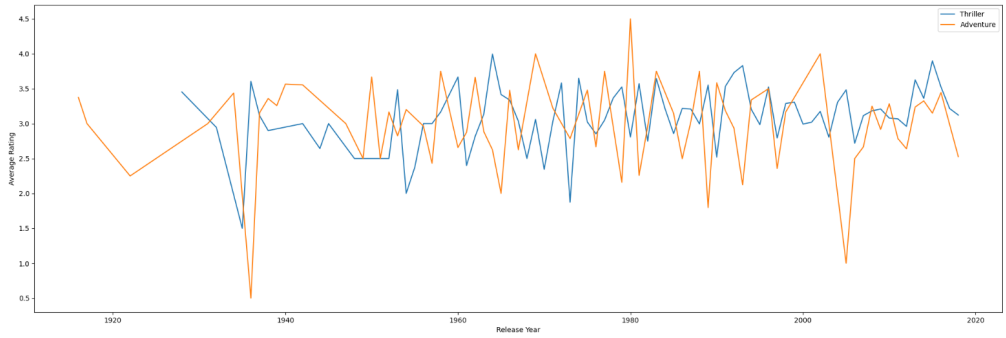


graph tells us that Film-Noir came out later than Horror and as we approached 2020, Film-noir ratings seem to be on the decline and the number of user ratings seem to diminish as well (telling us that it's popularity declined). Overall, despite Film-Noir's decline near the end, the data we have for Film Noir tended to be higher than horror because Horror's ratings were more diverse and watched by many kinds of people (people who enjoy horror on a regular basis, people who are checking it out based on hype, etc), so the ratings were lower. Film-noir is probably more niched given how little user ratings it has , so it had a fanbase that enjoyed it more and was catered towards a specific audience which resulted in higher ratings.

Adventure Vs. Action Vs. Animation Vs. Thriller

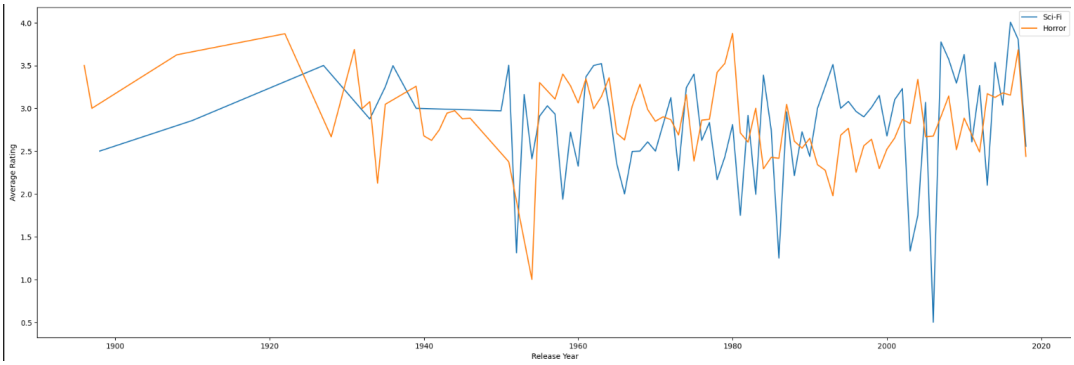
Through visual analysis of the distribution of ratings per year across a multitude of genres, a similarity in the line graphs of adventure, action, animation, thriller were remarkably apparent. The striking similarity between these genres ratings throughout the years, demonstrates a correlation between the four. We can infer that when one of these genres proves to be popular, the other respective genres also maintain an increased popularity. Vice Versa, as one genre decreases in popularity it is highly likely the others will face a decline as well. This may be attributed to the similarity in each genre's features. Likewise, a direct correlation can be attributed to a large quantity of movies typically being a part of all four genres. In each genre's cases, 1980 proves to be their most successful year, reaching peak ratings. Alternatively, around 1930 is when these genres reach the least popularity.





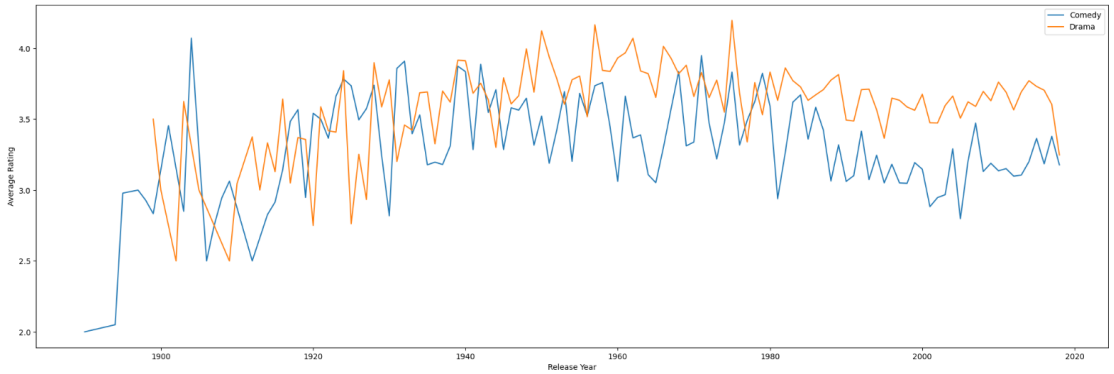
Sci-Fi Vs. Horror

Similarly to previously described relations between genres, we can determine there is a direct relationship between the Sci-Fi and Horror genres due to their near identical nature. Given the high similarities between both visual representations of the genre, we can determine throughout the years when one genre's ratings increases so does the other. Again, these can be attributed to the related features in each genre and the high likelihood a large variety of movies were categorized as both genres. In the case of Horror, peak ratings appeared to be around release year 1920, whilst it's popularity reached lowested around between 1940-1960. The Sci-Fi genre appears most successful around 2000-2020, whilst it was least successful also appeared between 2000-2020.



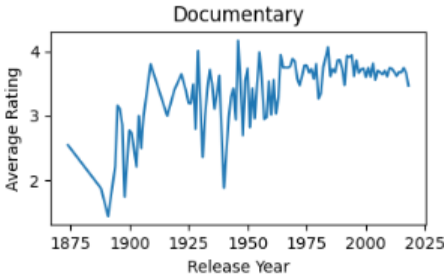
Comedy Vs. Drama

The last distinguishable relationship we can identify proves to be between the Comedy and Drama genres. These genres share a direct relationship and also have a high likelihood of sharing many movie titles and attributes. Typically, as seen in the visualization, as comedy reaches high success, popularity, and ratings, drama significantly increases alongside it. Vice versa, in the case of one's success plummeting. Particularly for the comedy genre, the highest ratings received belonged to the release year between 1900-1920. Alternatively, from 1878-1900 the comedy genre reached its lowest popularity. The drama genre achieved peak popularity between 1960-1980. Whilst alternatively, the genre receives its lowest average ratings from 1900-1920.



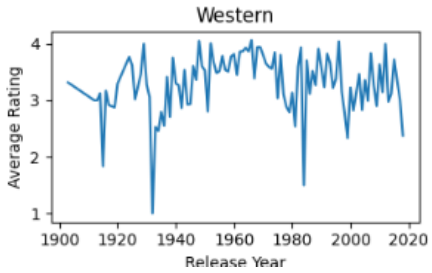
Documentary

The distribution of the average ratings per release year for the documentary genre appeared to share no common



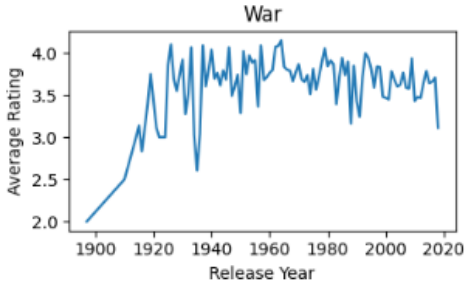
correlation with other genres. Furthermore, between the release years 1865 and 1900, the documentary genre seemingly lacked in popularity, scoring their lowest average rating. Vice versa, in the release years 1925-1950 a remarkably high new record of ratings demonstrates audience members liked documentary genres at an increased rate.

Western



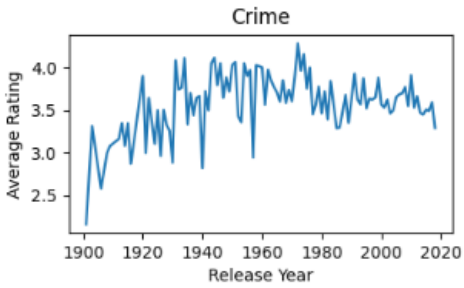
Similar to the documentary genre, the western category shared no striking similarity between other respective Genres. Unfortunately, between 1920-1940 viewers seemed to lack interest in all movies pertaining to the western genres. Besides lack of popularity, a lack of movies releases between this time period could be the source of such a decline. However, the western genre succeedingly highly with a multitude of peak rating periods, three to be exact: 1920-1940, 1940-1960, and 1960-1980. Audience appeared to be highly interested in movies pertaining this genre during such peak periods.

War



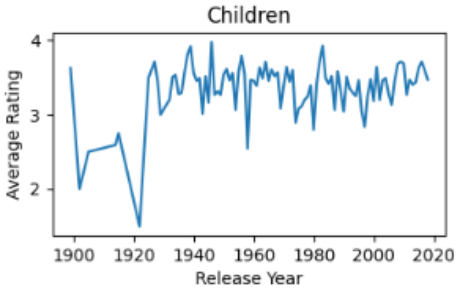
Throughout the release years, the war genre maintains relatively consistent high ratings. An almost persistent series of high ratings makes logical sense due to the war genre containing one of the highest average ratings with a 3.797. Despite a relatively consistent high success, unfortunately, around 1900 the war genre suffers it's lowest overall average ratings. Between 1920 and 1960, about 5 respective peak ratings can be visualized on the graph.

Crime



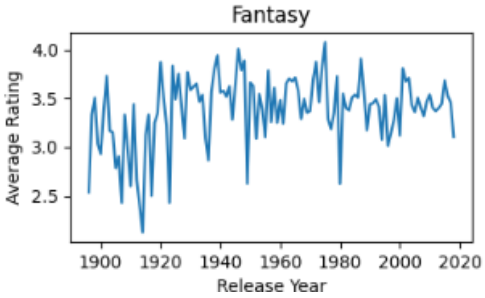
The visual representation of the crime genres average ratings against release years maintains a consistently fluctuating series of averages. In 1900 the crime genres reaches its lowest average rating in history, demonstrating a lack of interest and popularity for this category during this particular time. On the opposite spectrum, between 1960-1980 the genre achieves its highest success, proving to be significantly popular within such a time frame.

Children



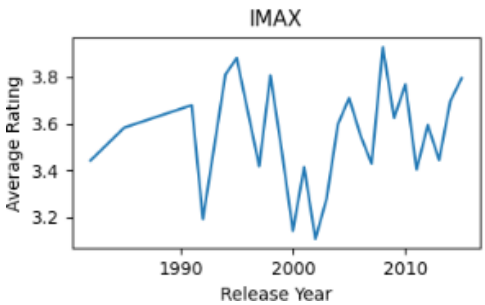
Based of the coded line graph representation of the average ratings for the children genre, we can infer that in 1920 it extremely lacked popularity. However, the children genre faces a multitude of peak popularity periods which attribute to their high average ratings ever: 1930-1990.

Fantasy



In the preceding years, a fluctuation of average ratings can be seen throughout the fantasy genre. In the release years 1900-1920, the genre faces it's lowest recorded average ratings, demonstrating it's least successful/popular time period. Vice versa, in 1960-1980 the highest recorded overall average rating is achieved, a perfect 4.0.

IMAX



Differently from other respective genres, the IMAX

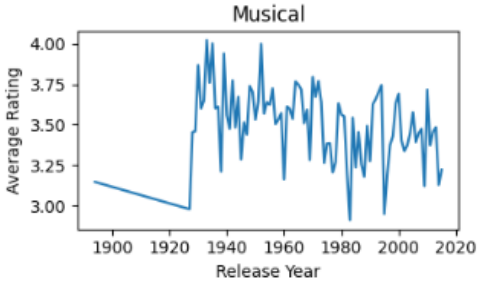
genre maintains significantly less fluctuations in

their average ratings throughout the release years.

The lowest point of popularity/ratings, is depicted in 2000-2010. Whilst the highest

average ratings for the IMAX genre can be seen between 2000-2010.

Musical



Throughout the release years the musical genre

faces a number of significantly increasing and

decreasing average ratings. However, in 1980-2000

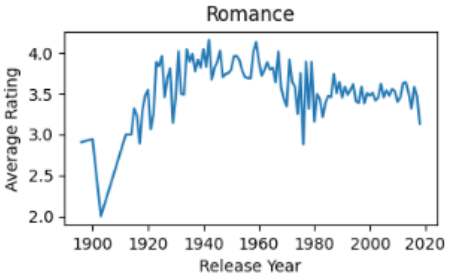
the genre faces it's largest sharp decline, reaching below a 3.0 rating. As a result,

demonstrating 1980-2000 to be it's least popular time period. Alternatively, in

1920-1940 the genre reaches a record high average rating of 4.0, becoming it's most

successful period of time.

Romance



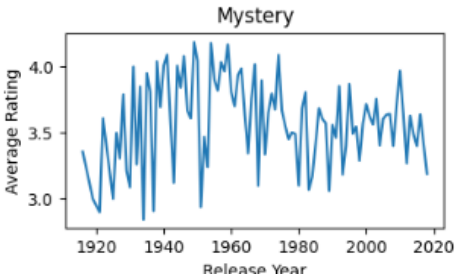
Based on the line graph, we can infer that throughout the

release years of movies categorized as romance, 1900-

1920 encompasses the lowest average rating for this

genre. As such romance was widely unpopular within this respective period. However, in 1940 the romance genre was increasingly popular, having an average rating of 4.0 compared to it's lower ratings throughout the years.

Mystery



The final line graph for our respective genres average ratings compared against release years is for mystery. In the mystery line graph we can conclude that based on the graphs lowest recorded average rating, 1920-1940 had the sharpest decline in popularity. Whilst from 1940-1960, the mystery genre experiences it's high period of acclaim throughout the release years.

Results and Conclusion

In conclusion, basic mathematical calculations regarding our dataframes determined a totality of 58,020 movies, 19 unique genres, and an average quantity of 515.700 ratings per movie. Through further analyzation of movie ratings we were able to determine the top ten highest rated movies in descending sequence: *Planet Earth II , Planet Earth,*

Shawshank Redemption, Band of Brothers, Godfather, Usual Suspects, Godfather: Part II, Schindler's List, and *Seven Samurai.* An investigation of the relationship between release year and average rating per genre gives us a plethora of information regarding our data. In analyzing the distribution of ratings Film-Noir maintains the highest versus the lowest being Horror, which may be attributed to the large quantity of user ratings for Horror. The increased quantity of negative ratings belonging to Horror, led to an overall high decline in 1940-1960, resulting in their lowest distribution. Whilst, the increased tendency of high ratings given in the Film-Noirs later release, substantially creates a rise in average ratings. In terms of other genres, a direct correlation can be found with a multitude, indicating as one genre's popularity increases the other respective ones also rise, and vice versa: Correlation One: adventure, action, animation, and thriller, Correlation Two: Sci-Fi and Horror.

One extensive limitation to our project is the analysis speed of a large quantity of our code. Due to the expansive size of our ratings and movies dataframes, oftentimes applying our code to each fraction of data takes an extensive period of time. As such, making for inefficient coding in the long run. A potential resolution to this issue may be to decrease the data size in order to process information at a more efficient rate. Another solution however that would allow us to maintain the expansive data set would be to process information in portions rather than totality. Depending on the problem statement, we could utilize only necessary portions of the dataframe to determine an answer. In turn, processing significantly less information at a time and increasing analysis speed.