3. Movielens dataset analysis

3.1 Data Extraction

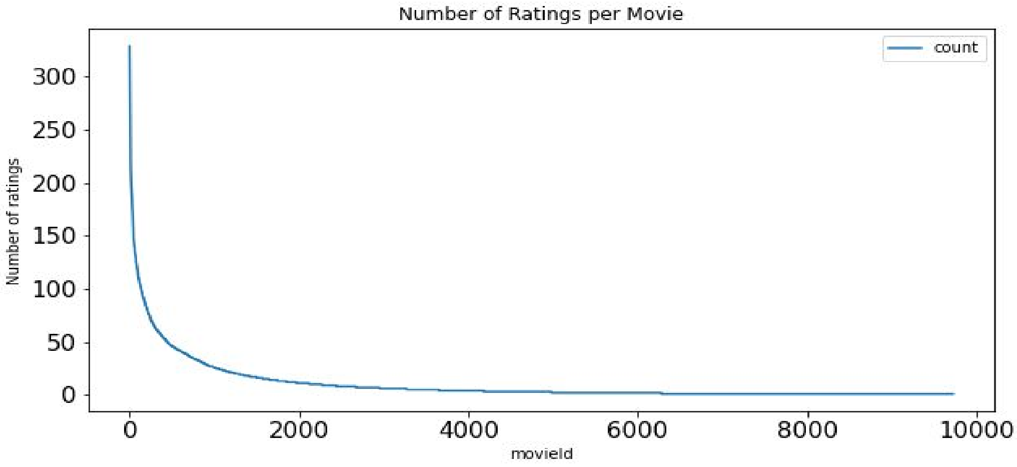
The “ml-latest-small” dataset used for this project was taken from GroupLens Research at https://grouplens.org/datasets/movielens.The dataset “ml-lastest-small “ includes 100836 ratings across 9742 movies with each user being identified by an anonymous id. In comparison the “ml-latest-small “ dataset has less data on individual users compared to “ml-1m Dataset” as it does not include additional user details such as age group and occupation. Despite this, our group decided to use the “ml-latest-small” dataset because the data is much more up to date. This is important for analyzing modern trends in genres, and allows us to more accurately assess our recommendation system.

3.2 Machine Learning

Machine Learning was an interesting concept that I wanted to further analyze using this project and test different methods learned in class to see which method yields the best accuracy in predicting similar movies.

3.2.1 Exploratory Data Analysis

Looking at the frequency of movie ratings per movie, it is visible that only a small faction of movies are rated frequently meaning the majority of movies in this dataset are rarely rated. Ultimately this behaviour results in a highly skewed distribution of ratings for the popularly rated movies compared to the less known movies.



3.2.2 Machine learning models

For this section, I decided to use the scikit-learn machine learning models we learned in class of SVC, K-nearest neighbors, Gaussian Naive Bayes, and Multinomial Naive Bayes. Each method was able to run in reasonable time except for SVC, which has a complexity between $O(n_{features} \times n^2_{samples})$ to $O(n_{features} \times n^3_{samples})$. As a result, I commented the SVC model out of the submitted code but recorded the accuracy score from when I allowed it to finish executing.

To prepare the data, I used scikit-learn's train_test_split on X = ['movieId','userId'] and Y = normalization of user ratings. Although all ratings were done from a scale from 0-5, a user's standard of average may differ from another user's standard of average. My approach for calculating the normalized rating for each rating is as follows:

1. Calculate average rating for each user
2. Subtract the average value from the actual rating for movies rated by the user.

Prediction Results:

SVC prediction score: 0.76076
MultinomialNB prediction score: 0.04485
KNN prediction score: 0.72986
GaussianNB prediction score: 0.76394

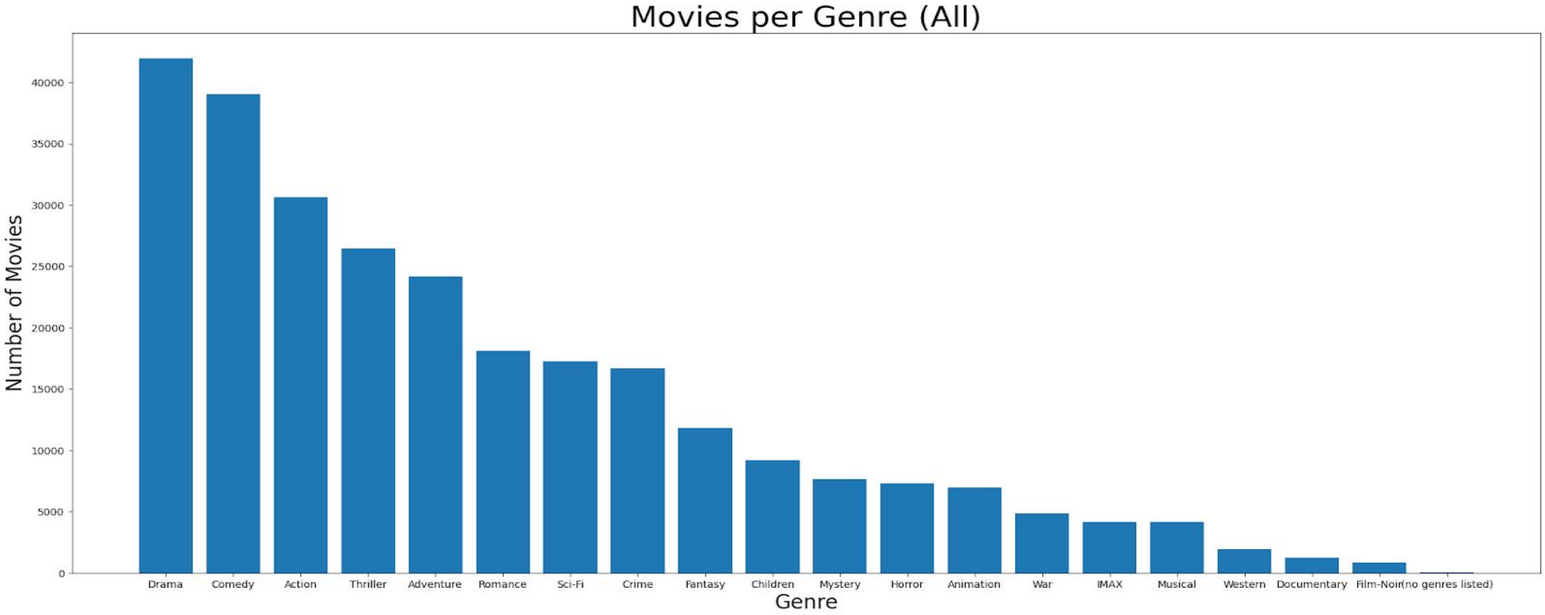
3.3 Movie Genre Analysis - Does genre affect rating?

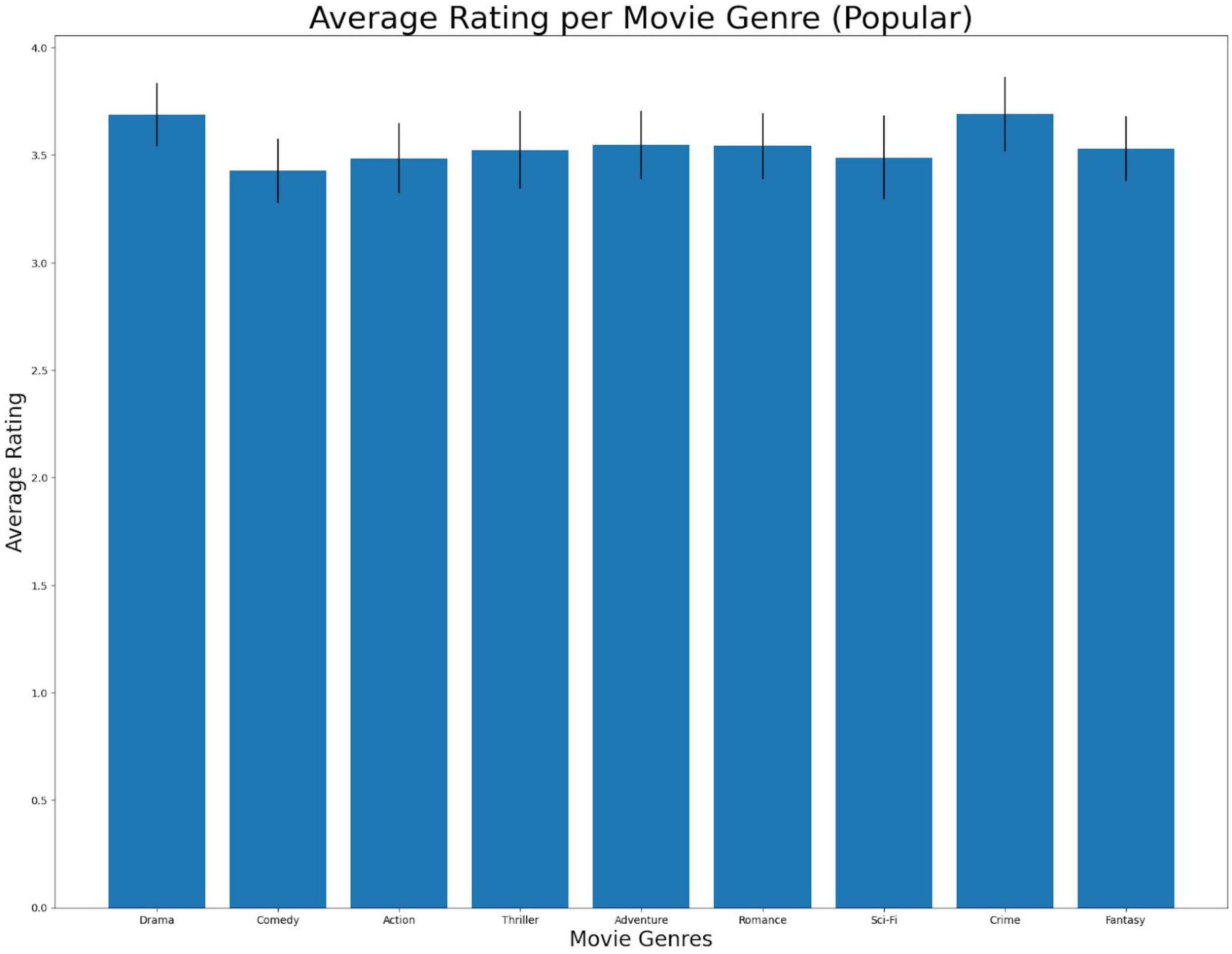
3.3.1 Cleaning the data

For this section, I wanted to analyze how the movie genre affected a movie's rating. To begin cleaning the data, I split the genres labelled for each movie into a list. Furthermore, I used pandas.DataFrame.explode() to divide each movie with multiple genres into a new row with only one genre. To clarify, given a row [movieId , 'action,thriller'], the movie will be divided into a row of [movieId, action], and [movieId, thriller].

3.3.2 Average rating of popular movie genres

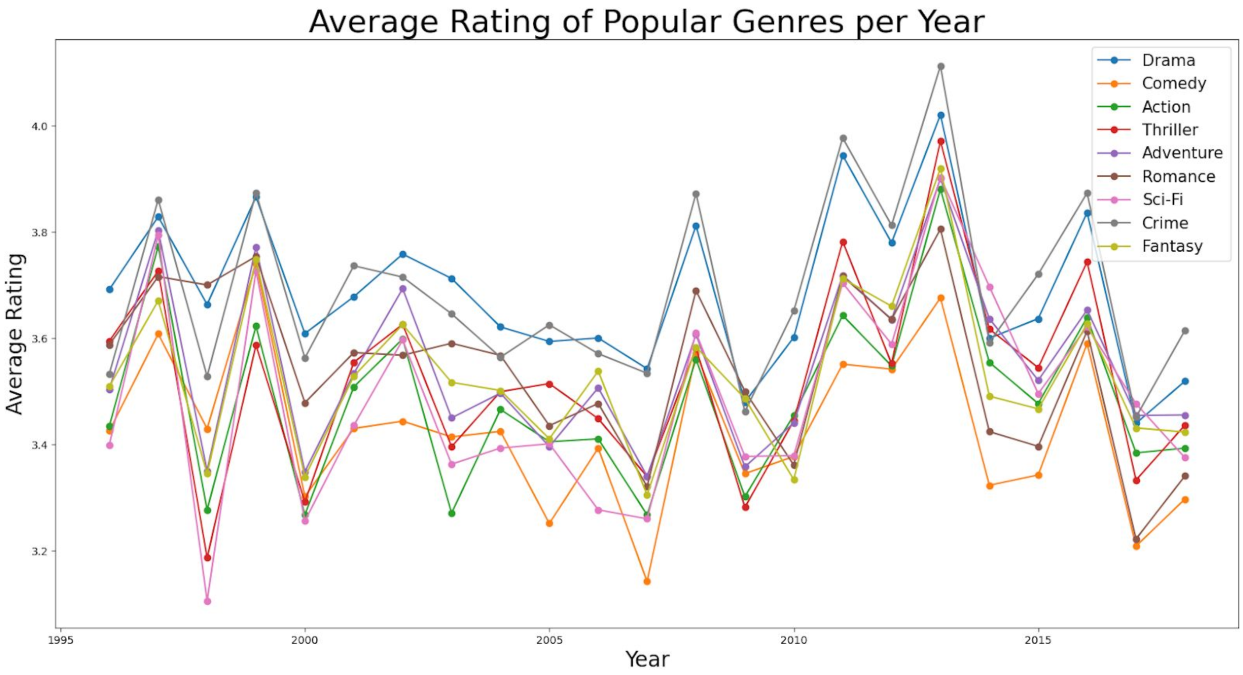
Looking at the count of movies per genre, it is evident that the most popular genres of a movie are Drama and Comedy. Other popular movies with a movie count greater than 15000 are Action, Thriller, Adventure, Romance, Sci-fi, Crime and Fantasy. To further analyze the results, I plotted the average rating for each popular genre. I used genres with a high movie count in order to have a higher sample size of rating. Genres with lower movie count may be skewed higher or lower due to less samples. Lastly, I ran an ANOVA test to determine if there is a significant difference in average ratings per genre. With a ANOVA p-value of 2.39816e-05
It can be concluded that there is a difference in rating based on genre.


Movies per Genre (All)


Average Rating per Movie Genre (Popular)

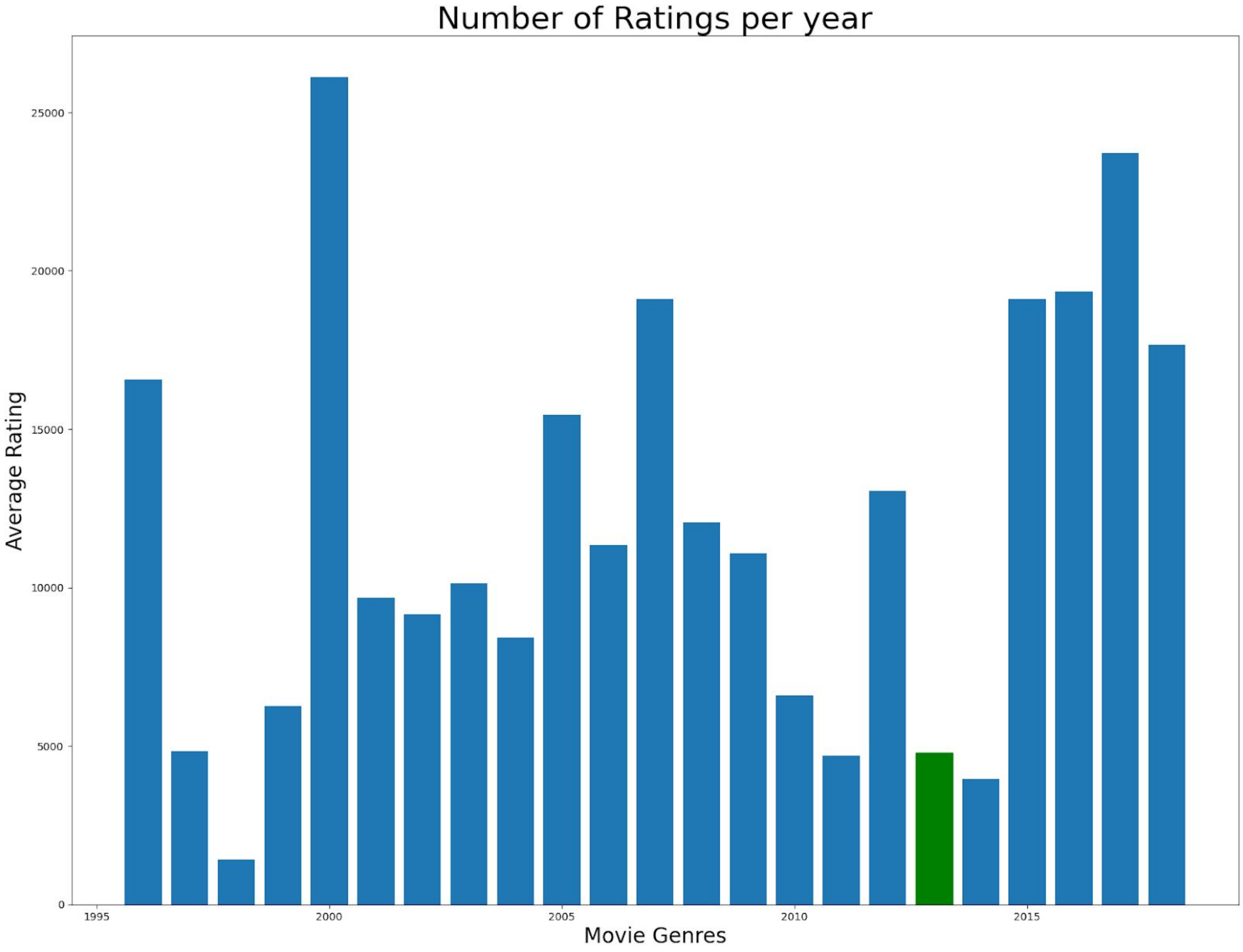
3.3.3 Average Rating of Movie Genre vs Time

Under observation, the average rating of the Drama genre compared to other popular genres is almost one standard deviation higher. In 3.3.2, I plotted the average genre rating for the popular movies as a whole, but I wanted to determine whether the average ratings of movies have changed over time. Has society always enjoyed Drama movies as a whole or is it a new trend for modern movies? In order to answer this I plotted the average rating for the popular genre for each year between 1995 - 2018.



From the plot above, it is visible that Drama movies are highly rated every year jumping between the highest and second highest rated with the crime movie genre. One unusual observation from this plot is the sudden positive spike in ratings for many movie genres. This effect can occur in two possibilities. Either movies in 2013 were simply really good, or the number of ratings in 2013 were simply lower compared to other years causing ratings to be inflated.

To determine whether or not the ratings are inflated, I plotted the number of ratings per genre per year. Labelled by the green bar, it is visible that 2013 had a below average rating count per genre compared to other years.



Lastly, for curiosity reasons I filtered the data to find the most highly rated movies per popular genre released in 2013..

Crime: Kick-Ass 2 , Side Effects , Gangster Squad
Thriller: Prisoners , Side Effects, Captain Phillips
Fantasy: Hansel & Gretel: Witch Hunters, Man of Steel, Jack the Giant Slayer
Sci-Fi: Gravity, Star Trek Into Darkness, Dark Skies
Action: Gravity, This Is the End , Kick-Ass 2
Drama: Prisoners, Side Effects, Captain Phillips

Conclusion

Overall, I can conclude that there is an effect on movie rating based on genre. Looking at the average rating of popular genres, it is visible that historically, movie genres such as comedy and romance have a lower average rating compared to the drama and crime genres. Additionally, the two most highly rated movies genres have been similar over the years with drama and crime.

Limitations

The biggest limitation with the “ml-latest-small” dataset is that it no longer includes in-depth data about the users rating the movies. In the past, movielens datasets had additional user data such as occupation and age group, but that dataset has no longer been updated since 2003. Although the data would have been useful to analyze additional questions such as the effect of age, gender or occupation on genre or rating, the data may be too old to be relevant for today’s society. There could be external social norms or an increase in movie standards in our present society and technological improvements that differ from the past data.

Accomplishment Statement

QiZhong (Francis) Wan

- Formed ideas and questions for the general direction of the project
- Created the machine learning using various scikit-learn machine learning models learned in class
- Used Pandas to clean and transform data for machine learning and genre analysis
- I used various statistical analysis tool learned in class and visualized the results using matplotlib
- Created the recommender system