CMPT 353 – Computational Data Science

GROUP PROJECT: ANALYSIS AND COMPUTATION ON MOVIES

BY:

AARISH KAPILA | 301269929

ALI NANJI | 301361228

QIZHONG (FRANCIS) WAN | 301351323

GitLab Repo: https://csil-git1.cs.surrey.sfu.ca/francisw/cmpt353-group-project.git

August 08, 2020

1. Recommender System

1.1 Intro

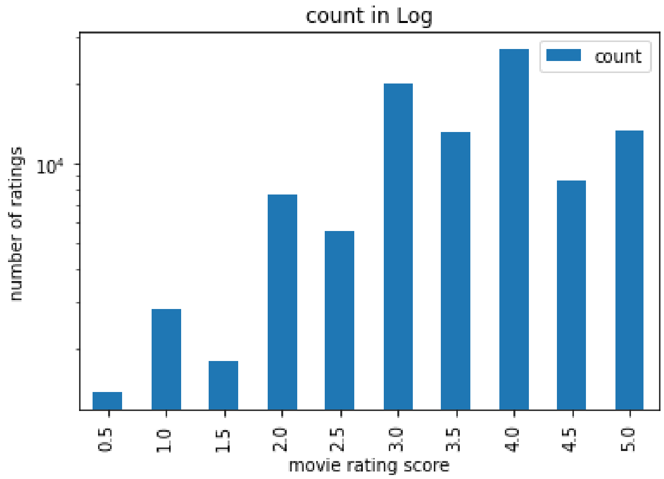
As a team, we always wondered how Netflix, YouTube or other online media platforms recommend items. We found through our research that with the applications of data science and machine learning, a recommender system can be built to support such businesses. Hence, we decided to build a simple item based recommender system. An item based recommender system is simply recommending items based upon a given item (which is assumed to be an item liked by the user).

We also came across two major collaborative filtering techniques used to implement such systems. First one is called Memory based and relies on taking a preferences matrix for items by users, using this matrix to predict missing preferences and suggest items. The second one is Model based, where methods are designed using machine learning algorithms such as k-nearest neighbors, random forest, etc. to predict unrated user ratings.

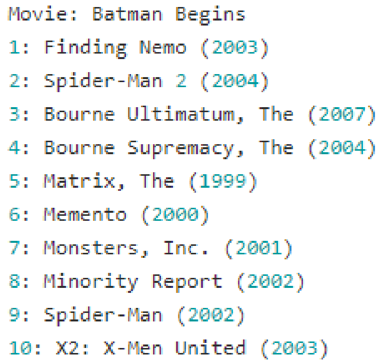
After using various classifiers, k-nearest neighbors seemed to have the highest accuracy score. So, we decided to use that to create a model and fit the data. Also, the algorithm itself does not make any assumptions about the underlying distribution of the data but relies on the similarity of the item features. It will measure the distances between the target movie and other movies in the database and will then rank them and return the top k-nearest neighbors as the most similar movie recommendations.

1.2 Data Cleaning and Analysis

We chose the “`ml-latest-small`” dataset from the grouplens website. The dataset contains two major CSV files, the first one with movies and their respective ids (`movies.csv`) and the second one with movie ratings and other corresponding features. The first step was to create a userId - movieId 2d matrix, which was accomplished using the `pivot()` function. Then we log normalized the counts to standardize the ratings. Many users had just 1-2 movies rated which seemed irrelevant to include in our analysis. So, we chose a threshold of at least 55 ratings per user to be considered in our data. The same goes for the movies. Movies which are rated at least 55 times, will be considered in our analysis. Below is a plot for the movie ratings with the counts in log.



Now we were finally able to apply the knn algorithm. We observed that, as we increased the value of k, the boundary was getting smoother. So, in general any value between 4-10 seemed adequate. We then took a movie as input and suggested films based on the KNN inferences. We used a modified version of `get_close_matches()` to retrieve the index of the movie in our database. At last, our function `recommendation()` prints the recommended movies based upon the given target. Below are ten movie recommendations for the movie 'Batman Begins'.



```
Movie: Batman Begins
1: Finding Nemo (2003)
2: Spider-Man 2 (2004)
3: Bourne Ultimatum, The (2007)
4: Bourne Supremacy, The (2004)
5: Matrix, The (1999)
6: Memento (2000)
7: Monsters, Inc. (2001)
8: Minority Report (2002)
9: Spider-Man (2002)
10: X2: X-Men United (2003)
```

With our system we observed that the films that are generally projected as recommendations are the ones with the highest number of ratings. That means our model is heavily driven by highly-rated films. So, if a movie, which is not highly rated, is supposed to be the best match, will be completely ignored by our model. Also, when a new movie is added to the database, it won't be recommended much at the start. There is also an issue with wasting space with the sparse matrix of movieId-userId as lots of values are zeros.

1.3 Conclusion

We were able to implement a recommender system but with some issues. While researching, we found out that the sparse matrix problem could be solved by doing a matrix factorization. The level of difficulty can depend upon the data. But we would like to conclude that we experienced a good insight of various data science and machine learning techniques while working on the system.

2. What factors influence the average rating of a movie?

2.1 Intro

The success of a movie can be measured in many ways. Some of these include awards and nominations, return on investment, or reaped critical acclaim. However, when a common man is trying to decide what movie to watch, mostly they simply look up the ratings of a movie and decide whether the movie is worth watching or not. If a movie has a good rating, it is likely to attract the attention of many viewers.

With our interest in machine learning, the natural thing to come to our mind was whether we could train a model to predict the rating of a movie (on a scale of 1-10) based on certain variables. We start by analyzing whether the average rating of movies has changed overtime, followed by analyses to see if there is a relationship between the average rating of a movie and three variables: the genre, the length of a movie, and whether a movie is an adult movie or not. Following this analysis, we attempt to train a model to predict the ratings of a movie.

2.2 Initial Data Cleaning

The data we work on for this part of the project was taken from IMDB. The dataset included over 30 variables for various titles in multiple files. Initially the data set was too big to be read into one notebook as some of the files included over 7 million lines of data.

After further inspection, we noticed that not all titles were movies, some were tv series, tv episodes, videos, etc. Using the `2_initial_data_cleaning.py` program, we created a data frame with only IDs (“`tconst`”) of titles that were movies, and this data frame was then used to filter all the other data files to only include rows on titles that were movies and create new GZIP files of the data.

As we looked through the data further, we also noticed that not all movies had ratings. These movie titles were then excluded form the dataset using the GZIP files created form the previous program and the `2_clean_data_for_ratings_analysis.py` program. This file then created new GZIP files for the needed IMDB files. This data cleaning process reduced the data size to 200,000+ rows, which was manageable in Jupyter Notebooks. For each of the analysis below, the data is cleaned or transformed as needed.

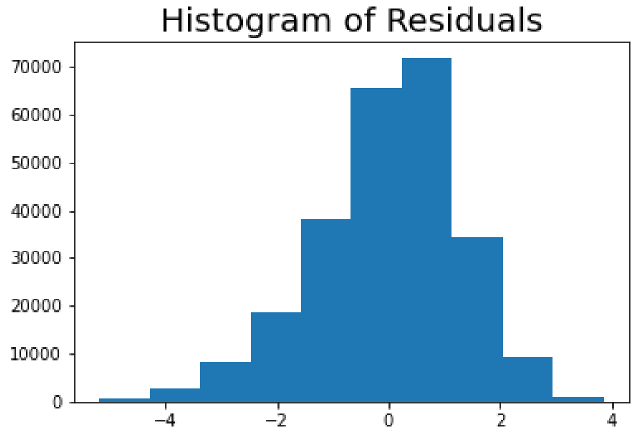
For all the analyses we set $\alpha = 0.05$. Additionally, for every tests, our data satisfies the following basic assumptions:

- The samples are representative of the population
- The samples are independent and identically-distributed

2.3 Variables Analysis

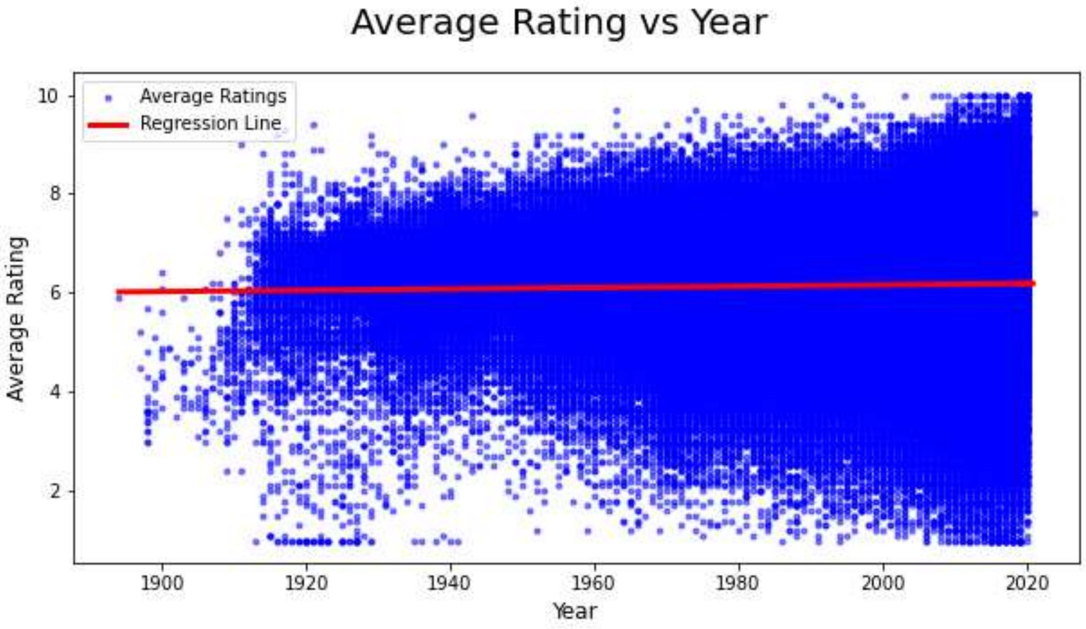
2.3.1 Year vs Rating

We were interested in testing if the average ratings have changed over the years. For this, we did not need to do any data cleaning. We simply read the columns required from different files and merged the data frames into a single data frame named “merged” which is then used for the rest of the analysis.



Before we do the analysis, we need to ensure the assumptions of the Ordinary Least Squares (OLS) assumptions are satisfied. The basic assumptions are fulfilled, but the only one we need to look at is the normal distribution of the residuals. As seen in the plot above, we can see that the data looks close to normal, and `n > 40` so we assume that it is normal.

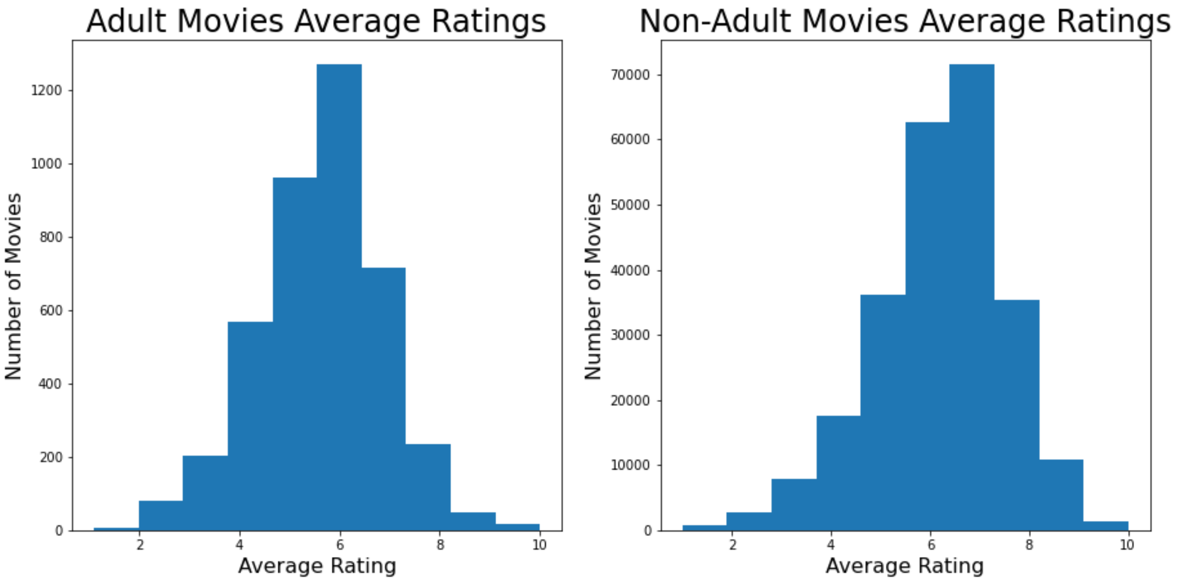
Although the p-value less than alpha, which means that the slope of the Regression line is not 0, the slope is ~`0.00135`, which is pretty much negligible. Additionally, we can see in the graph below that regression line is pretty much a straight line and the slope is not noticeable by the naked eye. Furthermore, as per the explained variance value, only `0.068%` of the variance in the ratings is explained by the change in years, which is a very small proportion. Therefore, we can conclude that the year of a movie’s release does not help predict the rating a movie.



2.3.2 Adult Movies vs Rating

Another variable we thought that would possibly impact the rating of a movie was whether a movie was an adult movie or not. To complete this analysis, we started by simply reading the columns needed from different files and merged the data frames into a single data frame named “merged.” We then separated the data into 2 data frames, one for adult movies and one for non-adult movies.

We planned on doing a T-test on these two data frames to test if the means of the average ratings differ. In order to do this test, among other assumptions we need to make sure that the data is normally distributed and have the same variance. We did a normal test on both datasets and they both failed. We then decided to create a histogram of both datasets (shown below), and we can see that both datasets look normally distributed. Because we have fairly large datasets, by the Central Limit Theorem, we can assume normality and do the test.

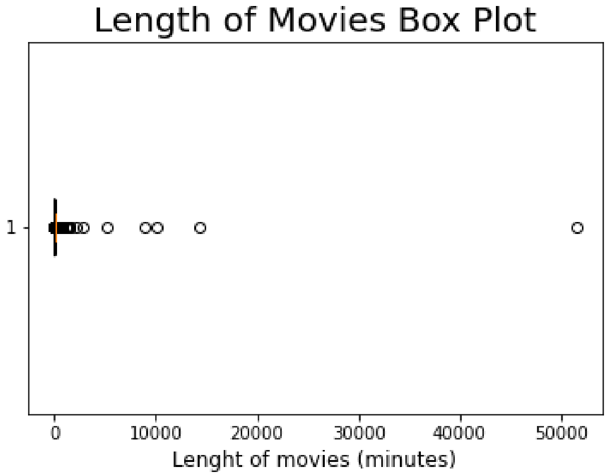


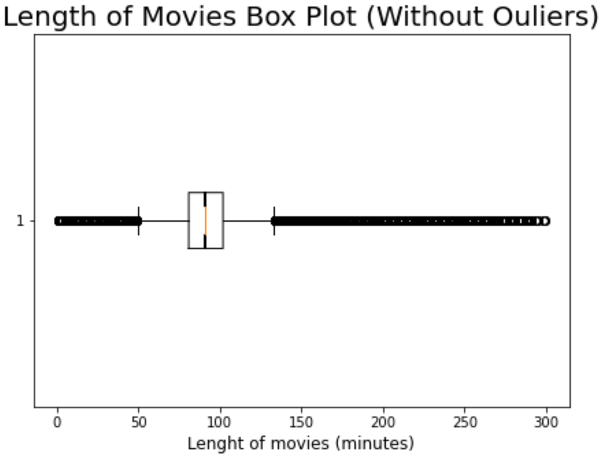
Unfortunately, the datasets have a very low p-value on the Levene Test, meaning the variances of the data sets are not similar and therefore we can’t perform the normal T-test.

However, we can do a variation of the T-test that does not assume equal variances. On this variation of the T-test, we get a p-value of `5.956e-145`, which is much lower than the α, which suggests that the two data sets have different means. In order to do further analysis with a clear conscious, we also did a Mann-Whitney U-test on our datasets. This test does not have any other assumptions but two: the dataset observations are independent, and the values are ordinal. Since our datasets satisfy these assumptions, we can perform a Mann-Whitney U-test on our datasets. With this test, we get a p-value of `8.608e-170` which confirms that the data sets do not have similar values when sorted. Therefore, we can conclude that whether a movie is an adult movie or not does affect the rating of the movie.

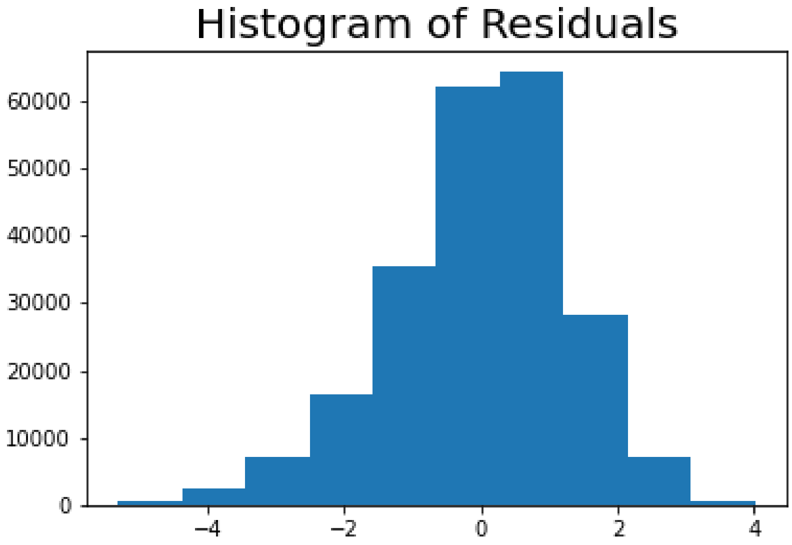
2.3.3 Length of a Movie vs Rating

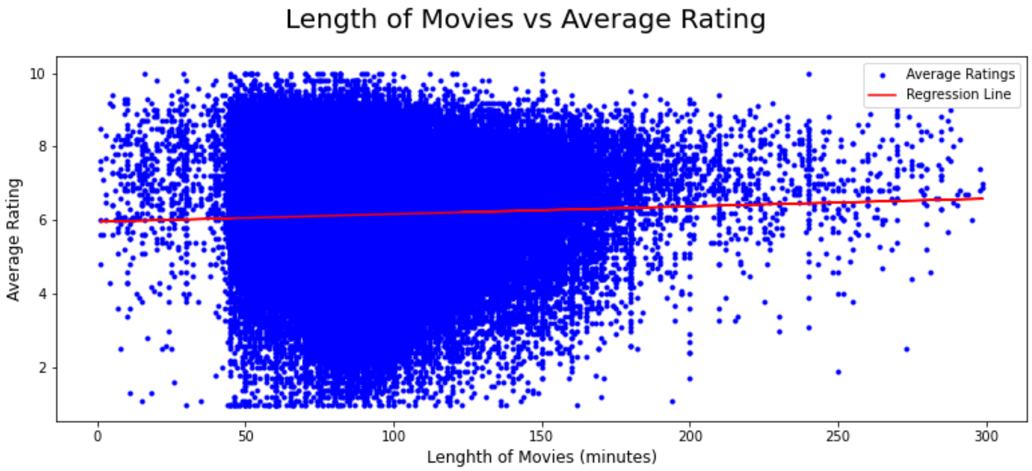
Another variable that could potentially affect the rating of a movie is its length. Does the length of a movie affect its rating? This is the question we try to answer in this part of the analysis. We start of by collecting the required columns into one data frame. Something interesting we noticed in our data were the outliers. As seen in the Box Plot below on the left, quite a few of the movies were far from the mean than other data points. To make the data more realistic, we filtered out all movies that were over 5 hours long. The updated Box Plot is displayed below on the right.
*Interesting Fact: The longest movie in our dataset is an experimental art movie that is over 35 days long, precisely 857 hours long.





Like one of the previous analysis, in order to be able to do an OLS test, we need to make sure the residuals are normally distributed to satisfy the assumption of the test. As seen in the plot below, the histogram of residuals looks normal, and since it has over 40 data points, by the Central Limit Theorem we can claim that the residuals are normally distributed and complete the test.



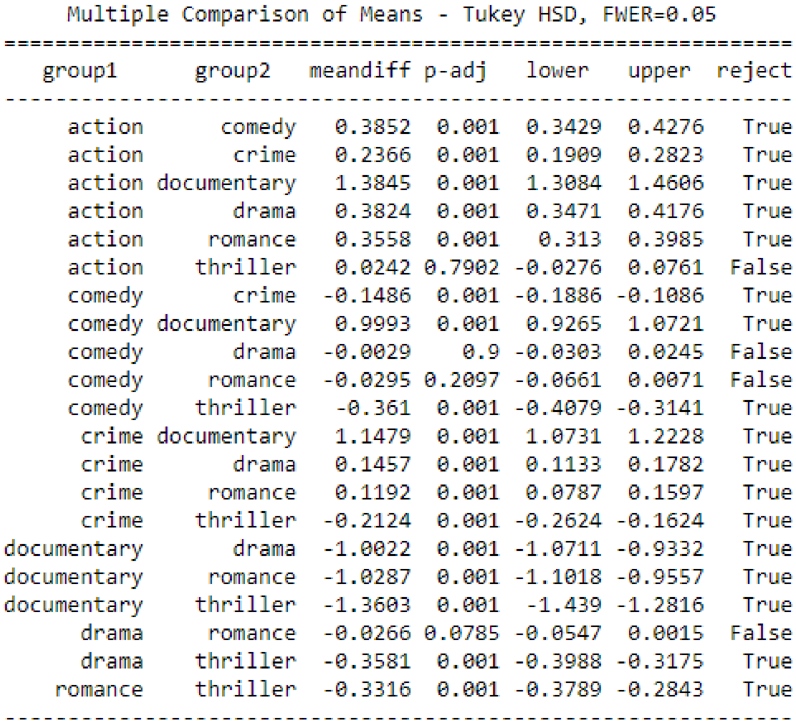


After completing the OLS test, we get a p-value of `1.984e-70`. Our results are significant with results $p<\alpha$. As we can see from the plot above, the regression line does slightly increase and the length of the movie increases. There is a possibility that the slope increases because there are fewer movies that are over 200 minutes long, but we can not ignore all of those movies because they still are valid data entries. Therefore, we conclude that although we have a very small slope, the average ratings do depend linearly on the length of the movies.

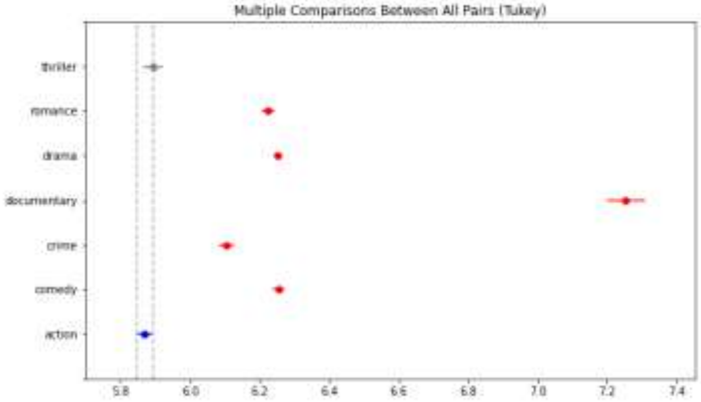
2.3.3 Genres of a Movie vs Rating

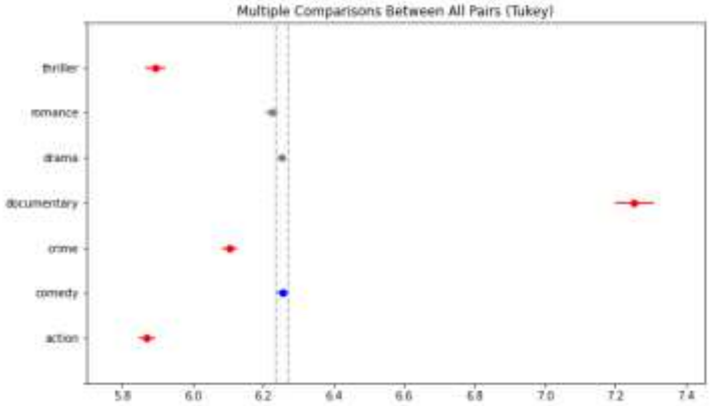
The last variable we were interested in testing is the relationship between the genres of a movie and its average rating. Each movie had between 1 and 3 genres saved in the data frame as a comma separated string. In order to do the analysis, the strings were split into an array and then exploded so that each genre of each movie is in its own row.

We then checked to see how many unique genres there were and found out there were 27 unique genres. Doing an ANOVA (Analysis of Variance) test on all of these would not be reasonable. We then decided to do an ANOVA test for only those genres that had at least 20,000 movies. This brought it down to 7 genres. After doing an ANOVA test on these genres, we see get a p-value of 0, which means there is a difference between the means of the groups.



```
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
==============================================================
   group1      group2   meandiff p-adj   lower   upper  reject
--------------------------------------------------------------
     action      comedy   0.3852  0.001  0.3429  0.4276   True
     action       crime   0.2366  0.001  0.1909  0.2823   True
     action documentary   1.3845  0.001  1.3084  1.4606   True
     action       drama   0.3824  0.001  0.3471  0.4176   True
     action     romance   0.3558  0.001   0.313  0.3985   True
     action    thriller   0.0242 0.7902 -0.0276  0.0761  False
     comedy       crime  -0.1486  0.001 -0.1886 -0.1086   True
     comedy documentary   0.9993  0.001  0.9265  1.0721   True
     comedy       drama  -0.0029    0.9 -0.0303  0.0245  False
     comedy     romance  -0.0295 0.2097 -0.0661  0.0071  False
     comedy    thriller   -0.361  0.001 -0.4079 -0.3141   True
      crime documentary   1.1479  0.001  1.0731  1.2228   True
      crime       drama   0.1457  0.001  0.1133  0.1782   True
      crime     romance   0.1192  0.001  0.0787  0.1597   True
      crime    thriller  -0.2124  0.001 -0.2624 -0.1624   True
documentary       drama  -1.0022  0.001 -1.0711 -0.9332   True
documentary     romance  -1.0287  0.001 -1.1018 -0.9557   True
documentary    thriller  -1.3603  0.001  -1.439 -1.2816   True
      drama     romance  -0.0266 0.0785 -0.0547  0.0015  False
      drama    thriller  -0.3581  0.001 -0.3988 -0.3175   True
    romance    thriller  -0.3316  0.001 -0.3789 -0.2843   True
--------------------------------------------------------------
```





Since we had significance in the ANOVA test, we can do a post hoc test. We decided to use the Tukey's HSD (Honest Significance Difference) test. We organize the data as required using the pd.melt( ) method and perform a pairwise Tukey's HSD test. The results of the test show that all the genres have different means, except for 4 pairs. Romance, drama, and comedy movies have similar means, and thriller and action movies have similar means. After looking further into those pairs, we noticed that there were over 3000 movies that had both genres, action and thriller. Also, there were over 3500 movies that had the genres romance, comedy, and drama. This may explain the similarity in the means of these pairs. Since most of the pairs had different means, we can conclude that the genre of a movie does influence its average rating.

We will further analyze the relationship between genres and ratings of a movie in the next part of our analysis with a different data set.

2.4 Prediction Model

After analyzing the variables, we decided to train a model to predict the rating of a movie with the genres of the movie, the length of a movie, and whether it is an adult movie or not.

We had to first organize the genres into indicator variables for each movie and concatenate it along with the other variables. The data was then split into a training and testing set. The different models we tested are as follows:

- Bayesian Classifier
- K-nearest neighbors classifier
    - `KNeighborsClassifier(n_neighbors = 10)`
- Random Forest Classifier
    - `RandomForestClassifier(n_estimators = 100, max_depth=20, min_samples_leaf=10)`
- Perceptrons
    - `MLPClassifier(solver = 'sgd', hidden_layer_sizes=(150,100,50), activation='logistic')`

We later realized that it may be harder for the model to predict the exact rating of a movie to the decimal value, but it might be easier to predict if the ratings were split into bins. We then rounded the ratings and completed the tests. We were not able to get any significant results. The Random Forest Classifier gave the best results with 0.38 on the training data and 0.37 on the testing data.

Then, we split the runtime minutes into 4 bins based on the quartiles. The 4 quartiles were as follows in order:

- 1-81 minutes
- 82-91 minutes
- 92-102 minutes
- 103-300 minutes

This transformation did not yield any better results either. The highest score we got was 0.375 on the training data and 0.372 on the testing data.

To avoid p-hacking, we decided not to fiddle with the data any further. All the models that we used did not yield any satisfactory results. From the results we got, we can conclude that these variables are not enough to predict the rating of a movie.

3. Movielens dataset analysis

3.1 Data Extraction

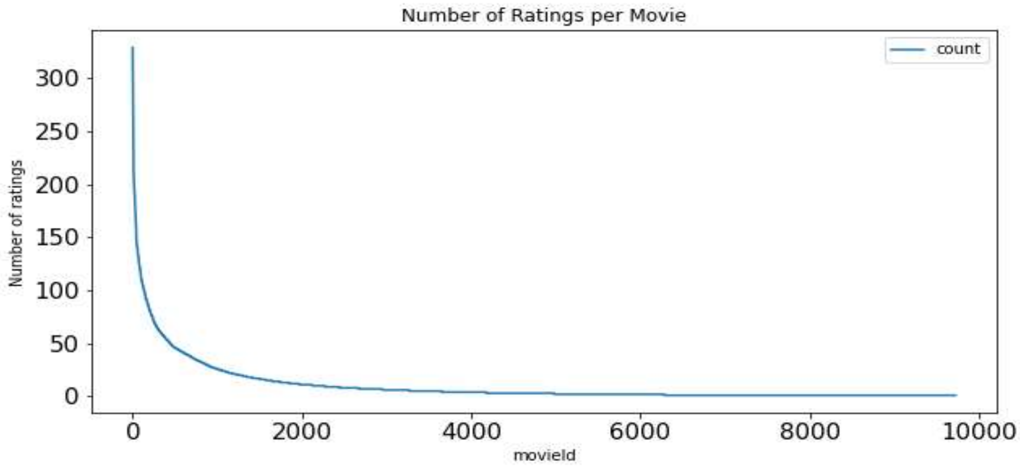
The “`ml-latest-small`” dataset used for this project was taken from GroupLens Research. The dataset “ml-lastest-small “ includes 100,836 ratings across 9742 movies with each user being identified by an anonymous id. In comparison the “`ml-latest-small`” dataset has less data on individual users compared 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 we 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 fraction of the movies are rated frequently meaning the majority of movies in this dataset are rarely rated. Ultimately this behavior 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, we decided to use the scikit-learn machine learning models we learned in class, namely 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, we commented the SVC model out of the submitted code but recorded the accuracy score from when we allowed it to finish executing.

To prepare the data, we 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. Our 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

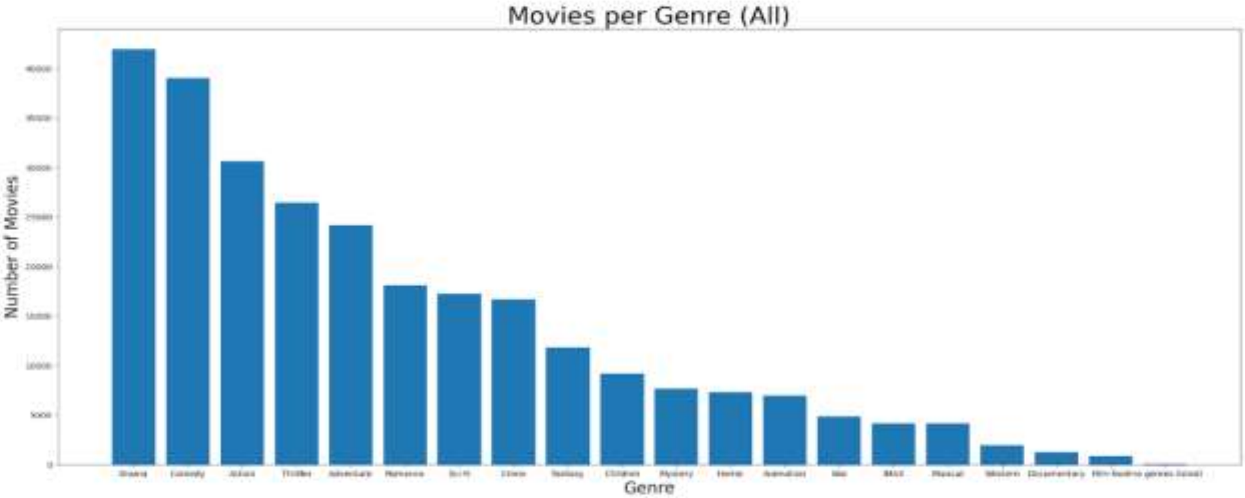
3.3 Movie Genre Analysis - Does genre affect rating?

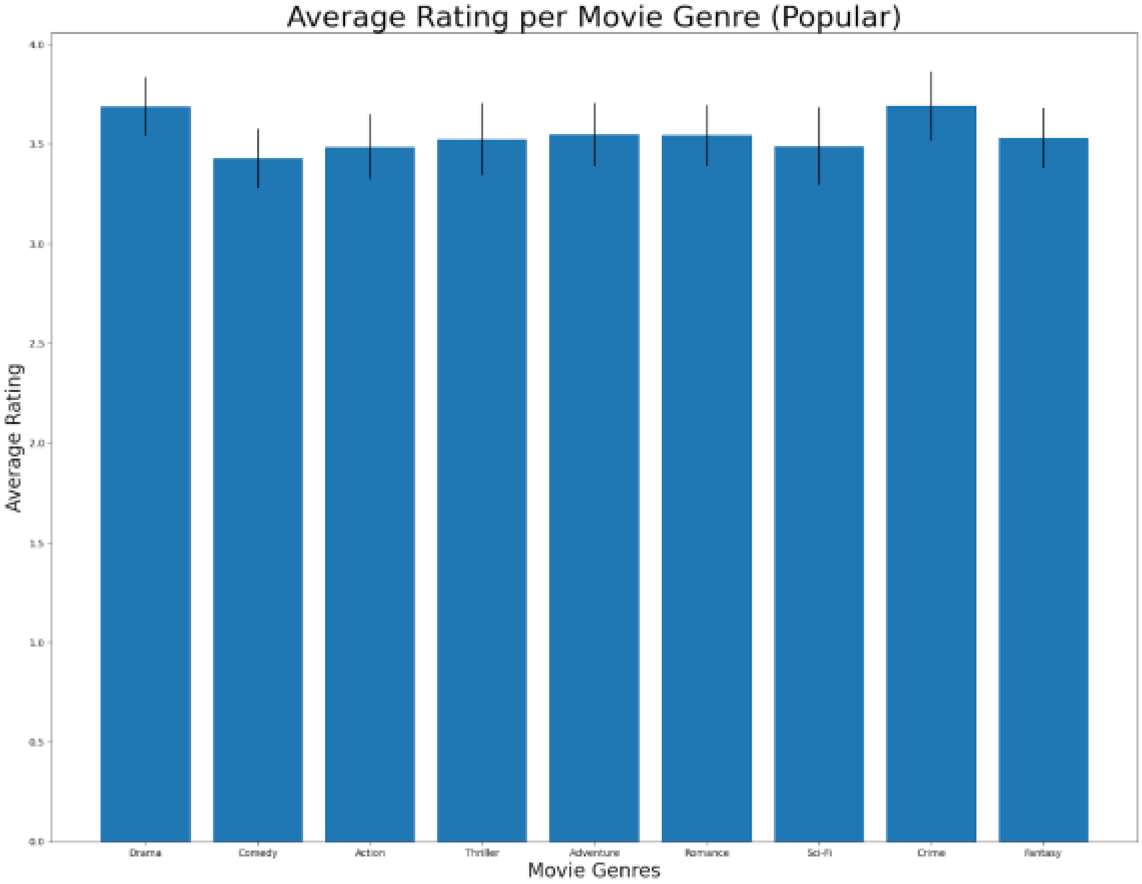
3.3.1 Cleaning the data

For this section, we wanted to analyze how the movie genre affected a movie's rating. To begin cleaning the data, we split the genres labelled for each movie into a list. Furthermore, we 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

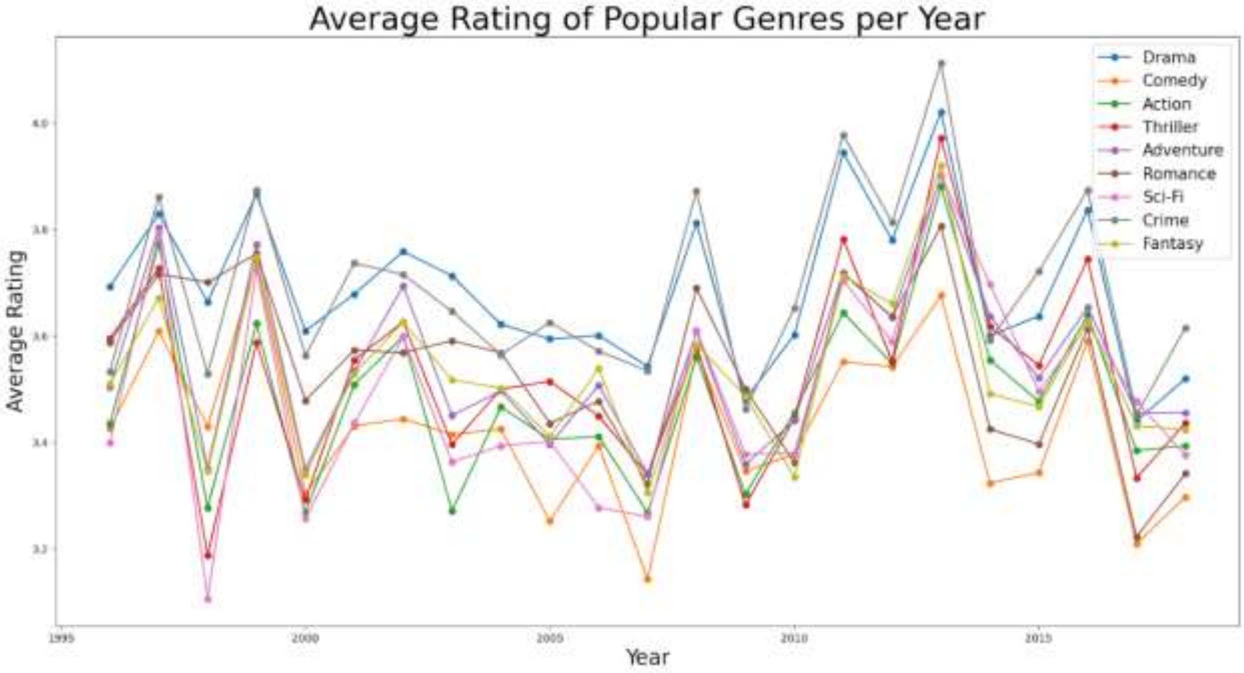
Looking at the count of movies per genre, it is evident that the most popular genres of a movie are Drama and Comedy. Other popular movies with a movie count greater than 15,000 are Action, Thriller, Adventure, Romance, Sci-fi, Crime and Fantasy. To further analyze the results, we plotted the average rating for each popular genre. We 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, we ran an ANOVA test to determine if there is a significant difference in average ratings per genre. With an ANOVA p-value of `2.39816e-05` it can be concluded that there is a difference in rating based on genre.





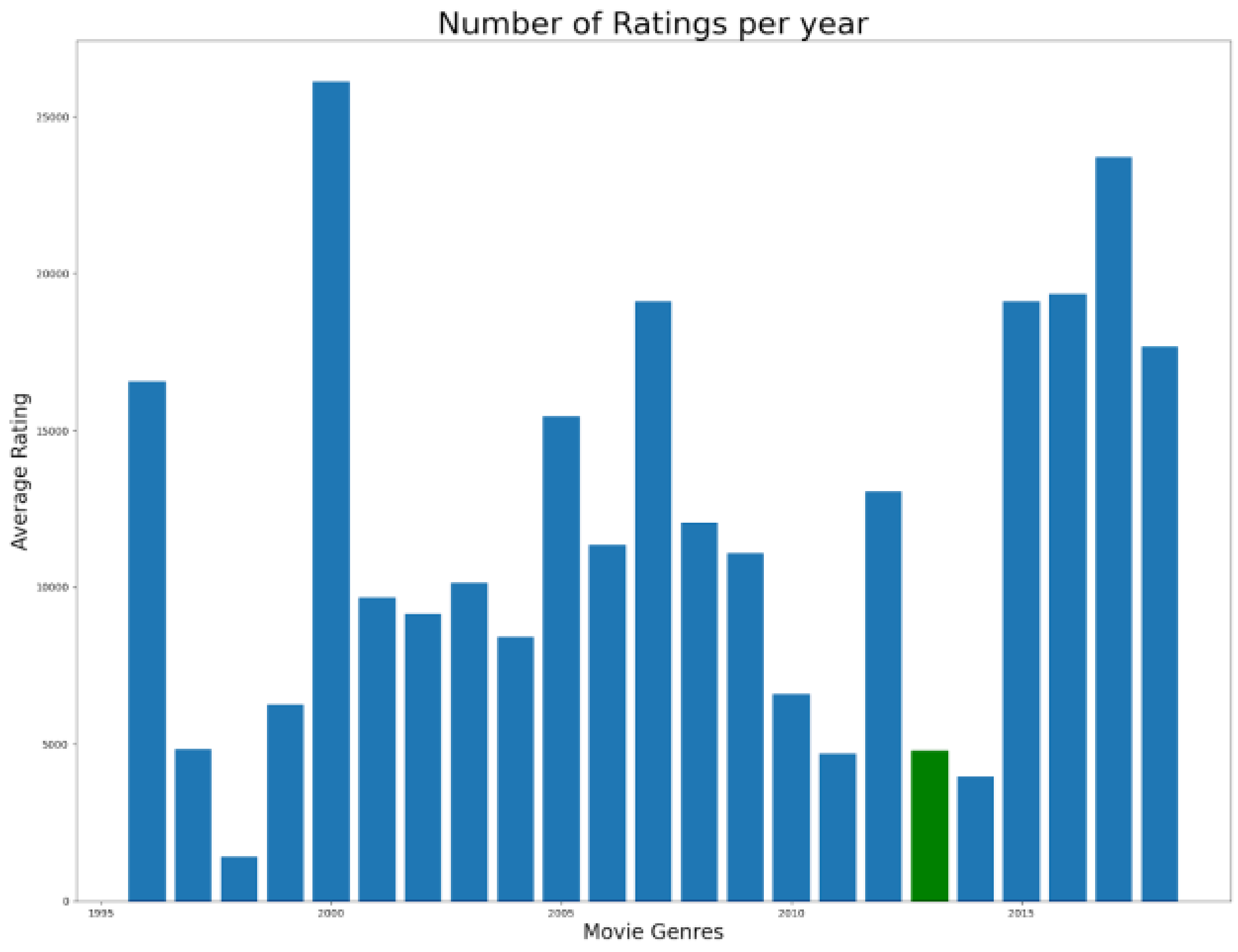
3.3.3 Average Rating of Movie Genre vs Time

Under observation, the average rating of the Drama genre compared to other popular genres is almost one standard deviation higher. In 3.3.2, we plotted the average genre rating for the popular movies as a whole, but we 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, we 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, we 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 we 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

3.4 Conclusion

Overall, we 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 or problems we faced

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. The lack of data in the IMDB dataset also limited our ability to do further analysis. If the dataset included data such as the revenue or the budget of a movie, it would have helped predict the ratings better. If we had more time, we would create a web scrapper to get more data and do better analysis.

References

Data Sources

https://grouplens.org/datasets/movielens
https://www.imdb.com/interfaces/

Accomplishment Statements

Ali Nanji

CMPT 353: Computational Data Science May – Aug 2020

- Collaborated with a group of 3 students to brainstorm questions we are interested in answering using data science tools and techniques.
- Utilized libraries such as Matplotlib, Pandas, Scikit-Learn, and Numpy in python to clean and analyze a dataset on movies from IMDB.
- Developed and trained a model to predict the rating of a movie based on the various features of a movie such as genre, length, and others.
- Completed various statistical tests on various features to determine which of the features would assist in predicting the rating of a movie.
- Maintained constant communication between team members through text and voice channels to ensure tasks are completed on time.

QiZhong (Francis) Wan

CMPT 353: Computational Data Science May – Aug 2020

- 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
- Helped with creating the recommender system before giving the responsibility to Aarish

AARISH KAPILA

CMPT 353: Computational Data Science May – Aug 2020

- Retrieved the data to work on for our project.
- Cleaned and organized the data for the recommender system.
- Using scikit-learn, experimented with various models to be used in our recommender system.
- Helped with various questions by suggesting which statistical test would be appropriate for particular questions.
- Mainly worked and finished the movie recommender system.