Utilizing User Reviews and Demographics to Identify Successful Movies and Ideal Audiences

Analyst: Aaron Godfrey

Date: November 20th, 2022

Table of Contents

ABSTRACT

The goal of this project was to take a 1,000,000 observation movie rating data set, provided by the University of Minnesota's MovieLens project, and attempt to create a prediction algorithm to determine a hypothetical movie's rating. The motivation of this project was to take an introductory dive into the recommendation engine system that many companies such as Netflix and Hulu utilize to provide their users a unique and enjoyable experience. While some factors were unfortunately discarded from the analysis, the bulk of the testing was done with data relating to each reviews score, the genre(s) of the film being reviewed, the reviewer's age, gender, and their occupation. The data contained a wide variety of ratings, each with multiple different genres but there were some clear concentrations of observations when it came to these factors. After multiple logistic-based regression algorithms, we determined that there was very little significance between the factors observed and the ratings of the films. Many reasons as to why this was the case were considered but ultimately, it was decided that perhaps the purpose of the project combined with the methodology chosen was the culprit. In the future, this project would require a new direction to test the data in and perhaps a different data set with more variables to consider.

PROJECT PLAN

Organization Details and Description:

Headquarters — Minneapolis, Mn

IPO — Private

Research Team Count — 11-50 people

MovieLens is part of a larger research team called GroupLens, which is based at the University of Minnesota. The main drive behind their research is in the business of creating recommendation systems for various mediums, such as movies and books, and they operate in attempt to further the quality of personalization. They primarily work with both undergraduate and graduate students to give hands-on experience with large datasets, building impressive systems to build up their repertoire. Additionally, they offer many of their datasets to the public for free use, with some being for the specific intent of practicing skills learned within the data science track of courses at a university. These include uses for machine learning, sentiment analysis, and more.

Analysis Opportunity:

Using their dataset of 1,000,209 anonymous ratings of 3,900 movies made by 6,040 users, with details such as 21 different occupations of users, 18 different movie genres, and seven different age brackets, this project aims to ponder at three different research questions to create some predictive models. The models will serve as

pseudo-guidelines for a potential movie recommendation algorithm in attempt to try and further curate a user experience in the world of film.

Research Questions:

In the modern age of entertainment and streaming services, many media consumers look to find movies that suit their tastes and provide an enjoyable experience. Using the dataset from MovieLens, some of the factors that can help curate that experience can be explored. This project will aim to research the following questions with the help of MovieLens.

Question 1: Does age/gender affect the ratings of certain genres?

There is a common assumption that certain generations of adults prefer certain genres of movies over others. This question aims to prey on that assumption and see if there is any significant correlation between the factors. With reviews being more impactful then ever, good movie production companies will tend to cater their film towards specific demographics in order to maximize box office revenue. The recent trend into streaming service releases only further exacerbate this need to have good reviews circulating and these services are as catered as you can get when marketing to consumers.

Question 2: What genres tend to rate higher than others?

With the current sandbox of superhero movies racking in the box office money, there does stand the chance that certain genres of movies may rate higher than others

and this question aims to determine the significance of this phenomena. Knowing if certain themes and motifs resonate with more viewers is a powerful ability when it comes to selling tickets. Some genres are more applicable to a broader audience, such as action or adventure, whereas others may have a more niche audience, like horror or thrillers. But knowing more about each of these genres and their scope is important, as the less popular a genre is, the more risk a studio takes attempting to develop a film in that category.

Question 3: Is the timestamp of the review indicative of any user demographic trait or rating?

The timestamp at which the review was taken is an enigma in the dataset as it could potentially contain heavy implications of quality or be a complete nonfactor. Rather than using this value to develop an insight on the film itself, this variable will be analyzed in attempt to tell us more about the audience watching and how their own life can directly correlate to the success or failure of a film. Perhaps people that review later into the movie end up awarding low scores or maybe females are more patient with their reviews than males are; these are some examples of correlations that this variable can help us identify. Perhaps developing studio productions requires more thought about the viewers than we originally thought.

Hypotheses:

Hypothesis 1: Gender will absolutely have an effect in genre ratings, but age will only influence the review score.

There are common stereotypes about certain genders liking certain genres, so much so that there are genres made to appeal to these observances such as “chick flicks”. However, with the launch of social media, the newer generations have the ability to view more and more content, allowing for more opinions to be made. I intend to dive deep into this hypothesis.

Hypothesis 2: Comedy and action/adventure will fluctuate greatly in score whereas other “niche” genres will be more consistent but lower.

There are the obvious standout films over the past few decades that continually have a presence in pop culture and the majority of them are comedy or action/adventure. With that being said, because of the popularity of these genres, many companies will try to replicate their success and will fail more often than not, leading to a large number of movies in the genre but they are rated much worse compared to the standard. While this still happens in genres such as horror too, it is not on the same scale, meaning the average score should not be as variable.

Hypothesis 3: Users with professions that require higher education will be more likely to have longer timestamps. Additionally, the shorter the timestamp, the worse the review score.

The first statement is not largely based on any facts nor experience per say but generally speaking, if the user endured the educational challenges for longer than others, they probably have more patience. The second prediction however is based on the idea that if the viewer truly does not enjoy the film, they are more likely to not even

finish the film compared to those that do enjoy it. This would mean they write a lower score.

Data:

The data from MovieLens comes from their own users on their service and includes structured variables for many different attributes. However, there are three main divisions of data points: Users, Movies, and Ratings. These sections connect with each other through shared ID values and help to paint the larger picture at hand.

Users

In this particular dataset, the user's demographic data was collected while their identities remained anonymous. These users voluntarily submitted their data pertaining to their gender, age, occupation, and zip code. Their age is bracketed into 7 groups starting at under 18 and ending at 56 and over and their occupations are divided into 21 categories, ranging from farmer to programmer, including some special cases such as student and self-employed. Each user also has a unique user ID.

Movies

The movies portion of the data have a much simpler layout. Each movie has a title, a year of release, a genre, and a movie ID value. There are 18 different genres, ranging from Action to Western. The movie ID is used to connect user reviews to a title.

Ratings

Finally, the ratings portion is mainly the connector piece between the other two portions. A rating has a user ID, a movie ID, a rating based on 5 stars (users can only review with a whole star), and a timestamp as to when the review was made since the beginning of the movie. Additionally, each user is required to have at least 20 ratings in order to be considered in this dataset.

Proposed Methodology:

Unlike other reviews and recommendation-based datasets, this project does not contain any written reviews that would require processes such as sentimental analysis. While that may mean that the procedures might be easier, since we are dealing with purely numeric values, that means that we should perform a more robust analysis.

Given that our data is primarily structured data, we will undergo exploratory data analysis on most of the fields in order to attempt some efficient filtering of weak variables. Our end goal for this project will be to have some primitive guidelines for a recommendation algorithm so we need to develop some way to classify key components for good reviews.

Regarding our first research question, this will most likely contain our greatest factors. We will probably lead the search with some simple regression tests of ratings, evaluating males and females separately while also testing the age brackets as well.

Regarding our second research question, this will widely rely on how we approach the situation. The main concept that we need to address the issue with is variance among reviews and how to properly scale results against one another. Certain genres will have significantly more entries than others, making some answers more concrete than others. Outlier analysis will have to be heavily implemented on this portion as we will need to find ways to treat each section of the data fairly. Plotting the genres individually and analyzing their distribution patterns is going to be the main approach, paying particular attention to mean square error and conducting subsequent hypothesis testing when appropriate.

Finally, our third research question is when we can possibly start looking into creating a starting recommendation engine. After we determine some key demographics to analyze through the exploratory data analysis, we can then investigate adding in some classification models, like decision trees or neural networks. Afterwards, we can possibly look into some light machine learning by training and testing data for predictions.

Output Summary:

Question 1: Does age/gender affect the ratings of certain genres?

When it comes to responding to the first research question, we can summarize the tests we conduct in tables plotting out the genders and age brackets versus key numbers such as adjusted R^2 values, standard deviation, F and t statistics, and more. Similar to a previous lab of mine (see below) we can layout a simple summary of stats

with each gender and age bracket represented in their own table or combined tables if we determine that two of the variables significantly correlate with one another (NOTE: the example tables would not dignify the use of adjusted R^2 as there is only one independent variable).

Carrier Delay Graphs/Models

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4533313	4533313	1560.77	<.0001
Error	1929	5602864	2904.54311		
Corrected Total	1930	10136177			

Root MSE	53.89381	R-Square	0.4472
Dependent Mean	53.86950	Adj R-Sq	0.4470
Coeff Var	100.04514		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	40.22367	1.27416	31.57	<.0001
Int_carrier_delay		1	0.86538	0.02190	39.51	<.0001

2 Variables:	ARR_DELAY Int_carrier_delay

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ARR_DELAY	4164	28.52906	54.75757	118795	1.00000	1399	ARR_DELAY
Int_carrier_delay	1931	15.76851	56.00415	30449	0	1399	

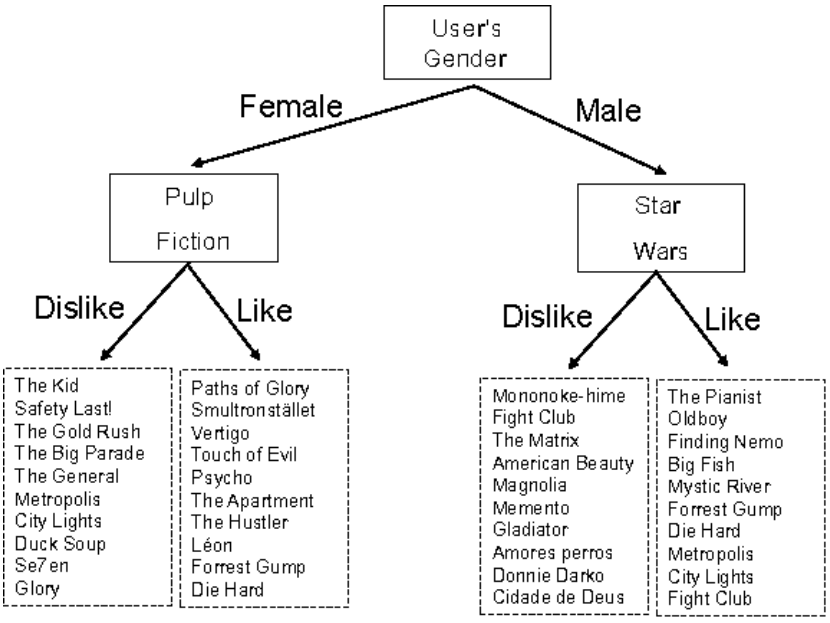
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	ARR_DELAY	Int_carrier_delay
ARR_DELAY	1.00000	0.66876
ARR_DELAY		<.0001
	4164	1931
Int_carrier_delay	0.66876	1.00000
	<.0001	
	1931	1931

Question 2: What genres tend to rate higher than others?

Regarding the second research question, as said before the output format will heavily depend on how the situation is eventually approached but we can bet on the fact that the results will be distribution-plot related. Perhaps by overlapping or displaying multiple plots in a grid, the graphs will be able to be comparable to one another as we will find some way to “fairly” compare the difference in review density per genre. Additionally, if hypothesis testing does occur, summary tables will be provided.

Question 3: Is the timestamp of the review indicative of any user demographic trait or rating?

As aforementioned with this question, we intend to summarize our findings by providing some light recommendation systems, such as a decision tree (see below). We can then add in modifiers as well such as confidence interval values per choice so that the reported models are both informative and interactive. Ideally, we have high confidence in most decisions so that the models can include the demographics of everyone, allowing anyone to interact with the model, but in the event that we can only predict one facet (i.e. if you reviewed the movie faster, you are male but our results do not necessarily prove that slower reviews are from women) then we will create either placeholder labels or disclaimers in the model.



Hopefully after researching these questions, we can have a more developed direction in which to aim our predictions towards and then layout a simple recommendation algorithm to summarize the project results.

LITERATURE REVIEW

As public perception continues to dominate the success of businesses worldwide, companies relentlessly demand answers from researchers on how to improve the reach of their influence. The simple answer to this dilemma is marketing. Advertisements all throughout their evolution from simple word-of-mouth to digital pauses in media consumption have stood the test of time by allowing the brand name and services of a company to nest within the mind of the consumer. But as is the case with most other commodities, nothing is free; so, to capitalize on advertising opportunities, businesses must know their consumer.

Generally speaking, people are more willing to act upon urges that relate to them in some kind of emotional or behavioral way. For example, when people decide on a movie to watch, they tend to reflect on their current mood and decide based on if they want to perpetuate that feeling or shift to a different tone. This is the same tactic that promoters aim to manipulate because if you can get the customer to care, you can get the sale. Countless studies have been done on men and women to evaluate how each of them handle their emotions similarly or differently. One such example comes from researchers Robin W. Simon from Florida State University and Leda E. Nath from

the University of Wisconsin, where in this study they evaluated the frequencies of both positive and negative emotions. Their findings deduced that negative emotions are reported by women significantly more often than men, whereas positive emotions are immutable based on gender alone but instead are affected by surrounding demographics such as income, marital status, and financial stability (Simon 2004). They also concluded that their evidence does support established social theory, such as higher status people in society tend to be more positive in their emotions and vice versa for those of lower status. While these concepts have already been rather well cemented in the logic of the common person, it is studies like these that ended up proving to companies that there are definitive trends to prey upon when designing marketing schemes.

Know that we have a better insight on how basic human emotions tend to pan out, we can begin to evaluate bias involved in emotions, or alternatively, how human emotions affect other humans' emotions. If the goal is to extend influence, then replicating positive experiences amongst the population should be the goal so we then look to studies in those fields. One in particular stood out rather keenly as they tested various aggregation techniques in attempt to mitigate popularity biases in group recommendations. Computer engineering researchers Emre Yalcin and Alper Bilge from Turkey conducted tests on groups ranking preferences for various situations and then, after employing many different clustering techniques, deduced that the two methods that reduced the most popularity bias and consequently improved group recommendation quality were those focused on reducing misery, specifically the

“average without misery” and “least misery” methods (Yalcin 2021). From this study, we can then infer that the key to advertising your services to others is to group customers based off of their emotional responses to products, specifically not invoking negative responses.

They go on to mention that utilizing methods like these can allow “platforms, such as Netflix or steam… [to] provide more visibility of the niche items in their produced recommendation lists and boost such items’ sales”. This is what we aim to further explore in this project. The specific area that we will conduct the research in is within the film industry. Like many entertainment moguls that exist today, the key to streamlining movies and tv series to the masses is through recommendation engines, which need some kind of basis to run off of. By using a movie review database, we can potentially see how the emotions of reviewers connect with other reviewers and see how that can determine the consensus-deemed success of the movie, while also looking for trends amongst the audience as well.

In addition to all of the sources referenced thus far, there is one source that conducted a very similar project at the University of Stirling. Researcher Karen Boyle took to the Internet Movie Database (IMDb) to analyze gender-targeted comedy films. While she ended up framing her discussion around testing if the service was predominantly male instead of female, this project will use her insight on gender imbalance when analyzing the success of certain films over others. Boyle’s work provides a great steppingstone for this study, giving a baseline on how to compare reviews from movies and identifying the demographic differences in reviewers, and

her work will be cited to allow for explorations into her processes. This project aims to not only pick up where she left off, but then to explore further into more demographic possibilities.

EXPLORATORY DATA ANALYSIS

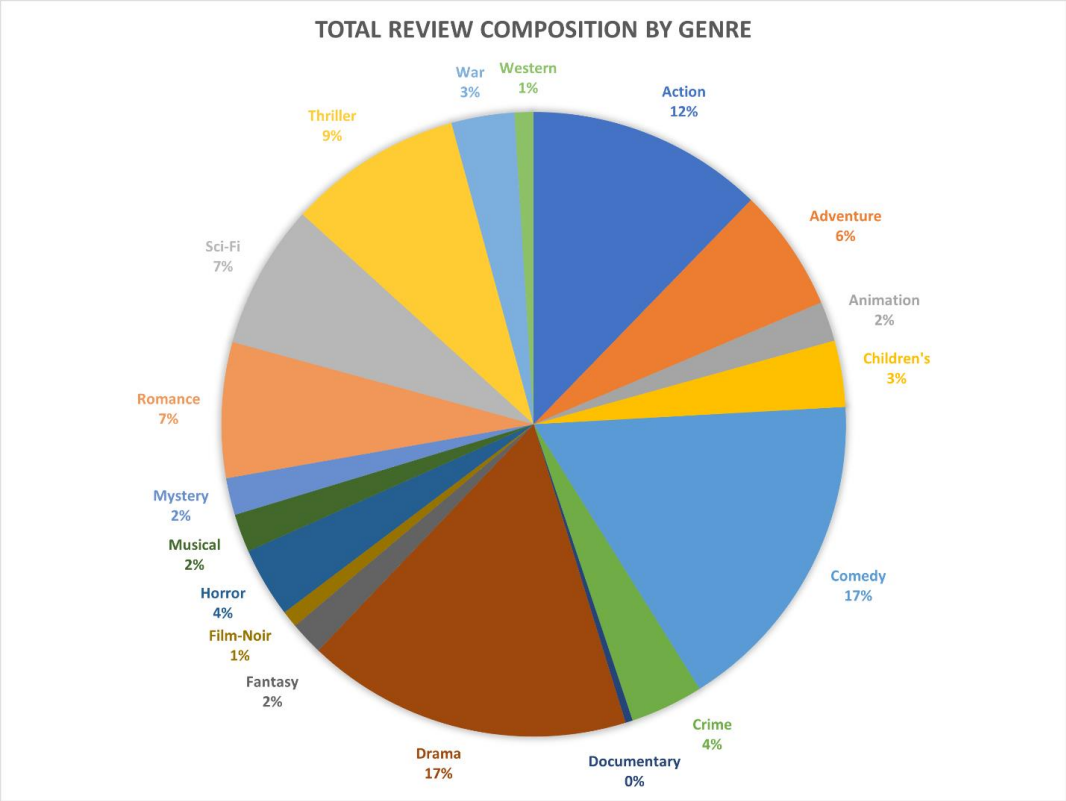
The data from this project comes from GroupLens.org, specifically the MovieLens division, and all of the data came structured. While this is review data, there was no text reviews, only a 5-star based system. The raw data was dissected, sorted, and graphed in Microsoft Excel. Each observation consisted of a user (indexed with an ID instead of a name), rating score, movie title, and time stamp at which the review was made. Due to the nature of this project, all variables were kept for analysis, however the data was consolidated into a single file to simplify the exploration process.

Amongst the ~1,000,000 ratings, there were 3,900 movies and 6,040 users, each with at least 10 reviews. While originally there were 10 different variables, the analyzed variables were as follows:

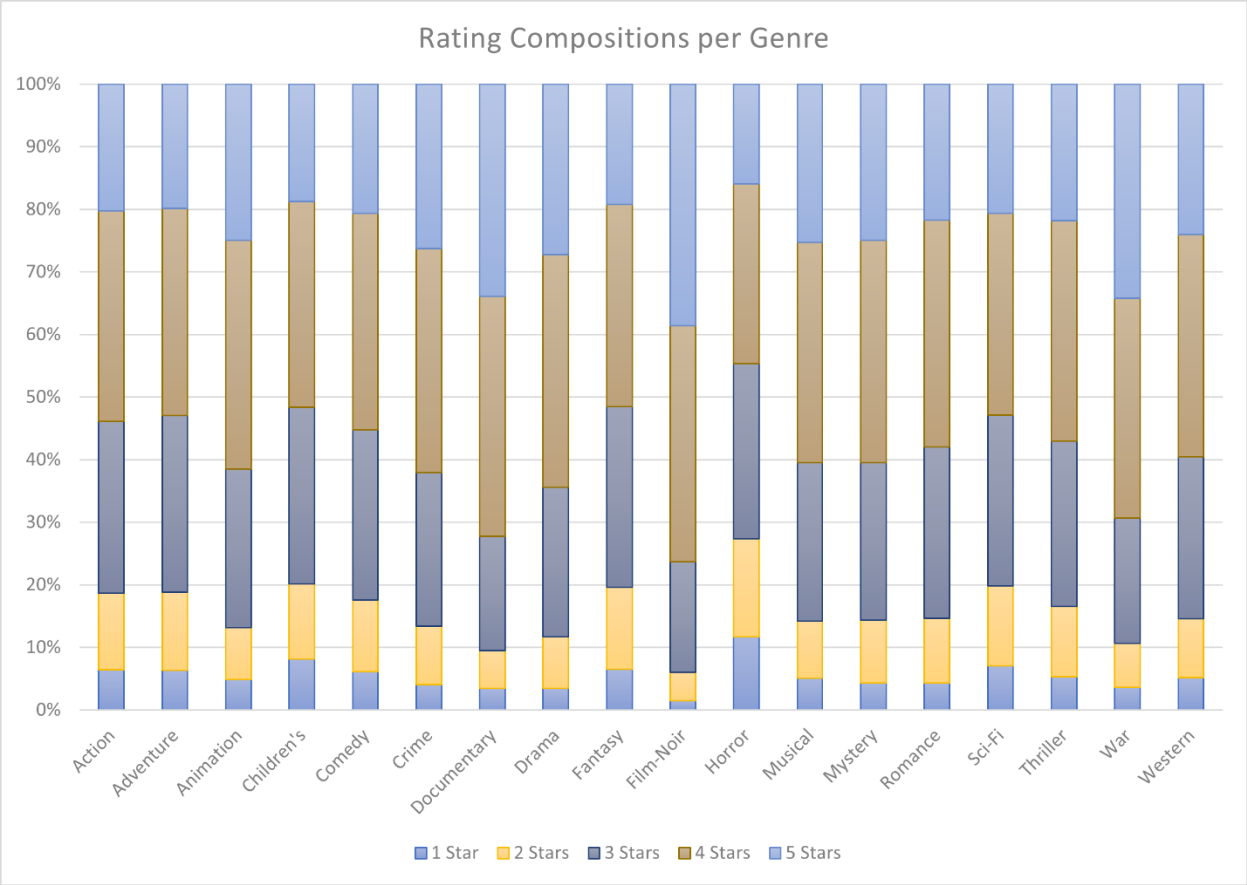
- Rating: integer value 1 through 5, with 5 being the best score
- Gender: Single character denoting the reviewer as male or female
- Age-Bracket: Age of the reviewer, within the ranges, under 18, 56+, or the various decades in-between

- Occupation: Employment status of the reviewer, one of 20 options including unemployed, student, etc.
- Genre: The Genre(s) of the film, multiple are listed when appropriate but only 18 different genres are selectable

Given that this dataset is based around movie reviews, I wanted to start exploring a variable that would be universal to all films, as going through each film individually not only has a broad scale of factors affecting its success but also trying to render each ratings value into a usable statistic would require a high amount of processing power. The answer to this query was to investigate the genre of the films. Each movie entry in the data set had at least one genre from a list of 18 attached to it. For the sake of contextualizing the diversity of movies in the data set, I started out by plotting the relative frequency of each genre in the data set (Note: Films with multiple genres contributed to each genre, e.g., Jaws was counted as both an Action movie and a Horror movie).



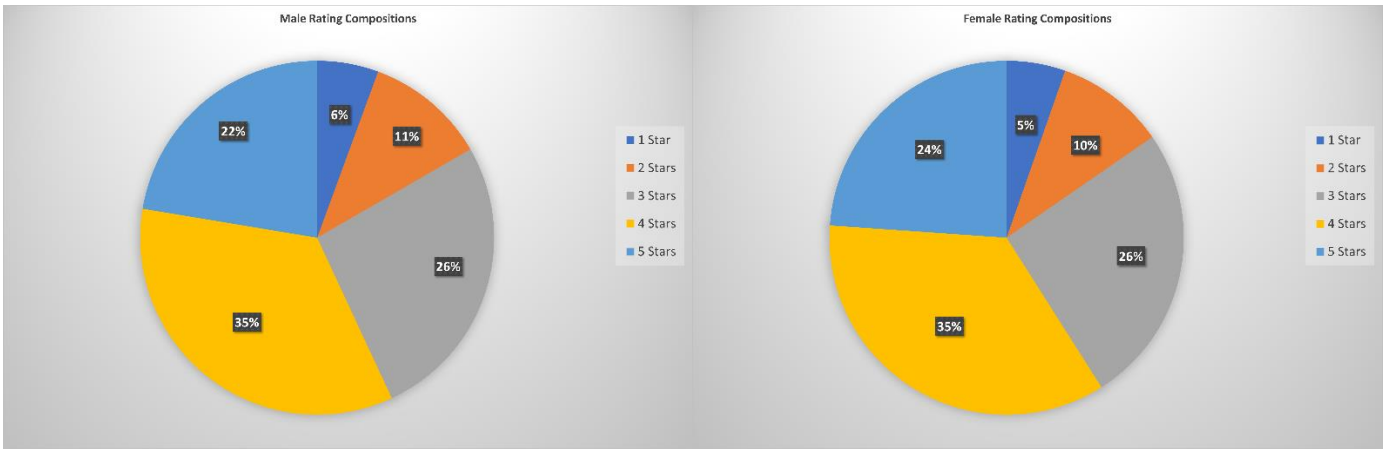
This pie chart helps us get a feel for what the landscape of the MovieLens reviews tended to focus on. As some would expect, there are a high concentration of comedies, dramas, and action films with other categories such as documentaries and film-noirs being barely represented in the full spectrum. This beginning perspective will help us to understand the biases in reviews as we delve deeper into the data analysis however to give the genres more of an even playing field, below is a distribution of review scores per genre. Remember that these reviews operate on a 5-star basis.



By looking at the previous two charts in-tandem, we can begin to formulate possible conjectures related to the ratings. One of these that we can start off with is the horror genre. According to the stacked bar-plot, the horror genre has the most relative 1-star reviews, which to horror fans is probably not that surprising. Horror is a more niche genre as most people do not like to tempt their fears. Generally speaking, many people that watch horror films do so because they enjoy being scared but consequently, this often means that they will be more critical of their horror films as some concepts scare them more than others. Other genres can have their reviews explained in a similar fashion. Comedy, for example, is commonly referred to as

subjective and therefore, while more people enjoy laughing than being frightened, people will still tend to be critical of the comedy they perceive from entertainment.

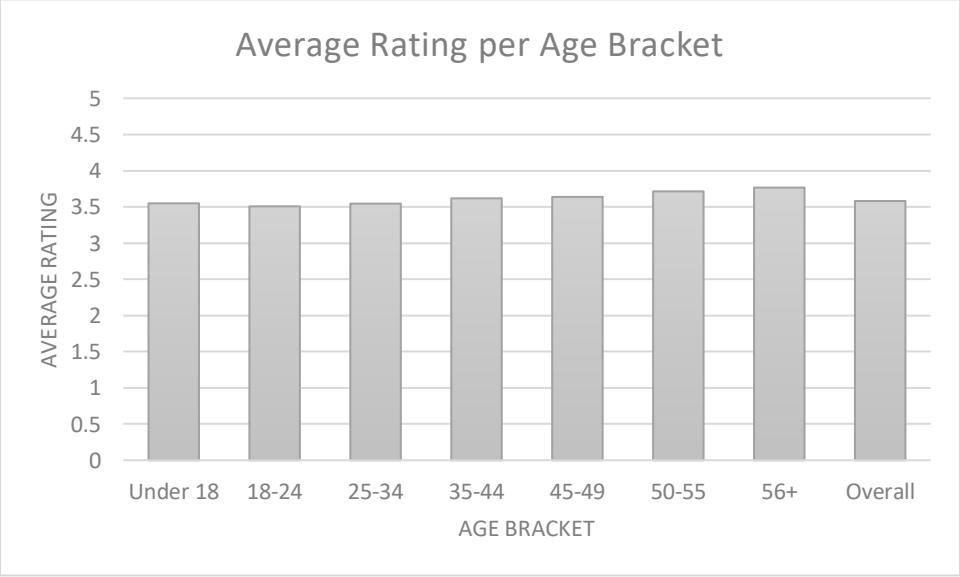
Moving into a deeper exploration, I decided to take other variables and test them alongside the ratings. The next pivotal variable I examined was gender. This category has so much potential to provide very valuable feedback as this is one of the main factors that movie recommendation engines will use to function. Starting out simple, I looked at the composition of the ratings between both males and females and the results are shown below.



As shown in later graphs, there were significantly more male than female reviewers so I figured it would be better to start out with a more balanced look at such a fundamental variable. Between both male and female users, the most frequent review is 4 stars, followed by 3 stars, and both ratings have equal relative frequency amongst the reviews. To add more complexity, we broke down those rating frequencies by genre as well, appendix C and D, and immediately more clear trends are seen. Between both genders, 4 star ratings are virtually always more plentiful in every

genre, however the concentration of reviews are different in certain genres. For example, men appear to have a much higher frequency of action movie reviews than women, but women have a higher romance review frequency.

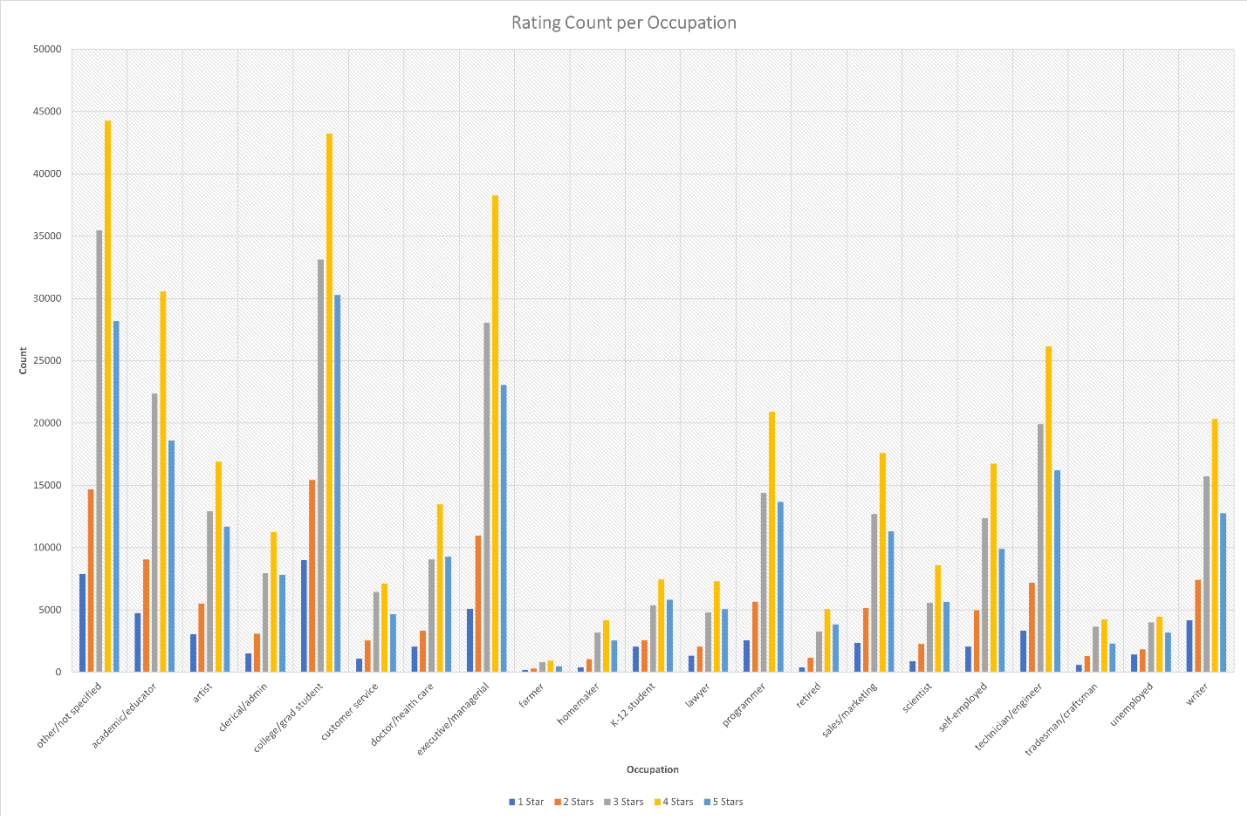
Graphs in appendixes G through K begin to consider the age of the users. Each plot shows the distributions for each star rating, and we consistently see the middle age groups being the most frequent reviewers. Specifically, users within the ages of 25-34 write the most reviews for every genre and the star rating does not appear to affect this. Additionally, we begin to see that as the ratings go up, the frequency of comedy and action ratings starts to decline however dramas start to dramatically increase their presence. While there may be something going on with those genres, we should also take a look at the mean scores of each age group to see if there is any major indication that age could be significant.



This graph that covers that relationship does not seem promising as all age brackets have an average rating of just above 3.5 stars which appears to be inline with the

overall average rating. If the brackets were more different from each other, it would have implied a heavy bias towards age which would imply that age is a significant factor in review scores. Regardless we will continue to test the variable to see if there still is some effect on the model.

Finally, we took a quick look at occupation to see if there were any significant correlations, we could spot but as you can see in the graph below, it does not look to be too promising.



The data appears to continue to follow the same trend of 4 stars being very popular, but it does provide a different insight on the users making reviews. Immediately, it is noticeable that farmers do not appear to be leaving many reviews, along with other

manual labor-based positions such as homemaker and craftsman. Conversely, more mentally challenging positions, like college student, executives, and engineers seem to have significantly more reviews. Naturally, there are also quite a few participants with occupations not listed in the ~20 options that still submitted reviews.

From this early look at the data, we can make some early conjectures on what will be major factors in determining what makes a critically-acclaimed film. Popular genres will tend to expand the audience the film will reach but high-quality niche genres will garner higher praise from reviewers. While males tend to watch more movies, women still have a big presence, especially in genres like romance and drama. Middle-aged viewers will also tend to express their ratings more often. When we move on to develop our prediction models and attempt to create a recommendation system, these factors and more should be heavily considered in the process if we wish to have positive feedback rates from our system.

METHODOLOGY

Research Question 1: Does age/gender affect the ratings of certain genres?

As we proceeded with our exploratory data analysis, we realized that while we were going to save the occupation of the raters to compare with the timestamp variable, I instead opted to include them along with the zip code in this analysis as well. Now that we are testing age, gender, occupation, and location in this analysis,

there must be some general guidelines. First, all the categories will be evaluated amongst their own values i.e., age data will be compared exclusively to other age data at first. Since all the variables being dealt with are qualitative, we will choose to undergo logistic regression using maximum likelihood estimation, having each of the variables represented in their own dummy variables. These dummy variables will have to be created for the data set, but they will still be attached to each individual rating.

After each individual plot, we will then generate main effect models and subsequent interaction models using average ratings as our performance checker to note if any of the plots would imply there is a relationship between the sets of variables. With whatever models this process generates, we will complete multiple 10-fold cross-validations and average the returning mean square error values to determine how effective the models are.

Research Question 2: What genres tend to rate higher than others?

While this question can be answered, and to an extent was answered, within simple EDA procedures, I instead took it upon myself to evolve the question slightly. While our EDA did give a good representation of each genre individually, that is not necessarily the end all be all of the question because, quite frankly, very few movies in the data set had only one genre, with some movies having up to six different genres. Naturally, I want to modify this question to account for that and in the end, the resulting model will be also more applicable to other movie environments. So, the

question we will be asking instead is: How well do certain genres rate and do certain combinations affect the ratings significantly?

As we plan to do for the first research question, we will run logistic regression models as our main method, due to the nature of every variable being categorical. However, to alleviate some computing power and time, we will apply a variable screening method. As of now, the current plan is to scan the data set for the frequency of the genres and additionally how often they appear with other genres. Furthermore, we will investigate popular genre pairings as well as eliminate non-existent pairings i.e. I'm sure there is no movie that is both Children's and Film-Noir. This should hopefully reduce the number of models that will be generated and, consequently, reduce the number of main effects models to evaluate as well.

Depending on the number of models our filtering leaves us with, we will run multiple 5-folds cross-validations (if the number is high) or 10-folds cross-validations (if the number is low. Once all is said and done, to experiment with other methods of creating our recommendation engine, we will create a classification-based decision tree and using test data, create confusion matrices to produce accuracy scores.

Finally, since the data set was released in 2003, we will provide our recommendation engines with some more contemporary films that are not in the data set to see how much the public's view has changed when rating movies.

Techniques Used:

- Logistic Regression (using maximum likelihood estimate):
    - Calculates the estimated effect that each variable (given the others are held constant) would have on the classification of a movie's rating
- K-fold Cross-Validation
    - Divides a data set into k subsets, turns one of the subsets into a test set, and evaluates the effectiveness of the model produced from the other k-1 folds
- Main Effects Model
    - Used to visualize how multiple variables affect a mean response when treated as independently
- Interaction Model
    - Accompanies a main effect model. Used to visualize how multiple variables affect a mean response when the variables interact to affect the mean response
- Decision Tree (Classification)
    - Predict the classification of a variable based on other independent variables on a tree-like binary basis (ex. Is the movie a comedy? Yes or No), using entropy to determine the binary splits

DATA VISUALIZATIONS

At this point in the project, I had realized that I was very ambitious with my desires but with the methods that I was able to learn and execute, I ended up severely limiting the proposed methodology, as well as completely removing the second question, partially due to time constraints after finishing the first question and partially because the EDA did indeed explore the topic enough to find some key correlations. More discourse about the reasoning behind these decisions will be located in the challenges and next steps portion of the report but as of now, the data visualizations and analysis will focus solely on the first research question: Does age/gender affect the ratings of certain genres?

Given that all my data is categorical, the methodology was very limited in options so to best answer the research question, we used exclusively multinomial logistic regression. Trying to see the effect of gender, age, and occupation with genres on the ratings of movies was a daunting task however results were still able to be collected, after a saga of different methods to reduce computing load. First, all the non-genre related variables had to get dummy variables associated with them but afterwards we were able to jump straight into creating the regression model. Every regression model was created with the rating 1 as the reference value, meaning the later obtained log odds equations would be each rating in relation to the 1-star ratings.

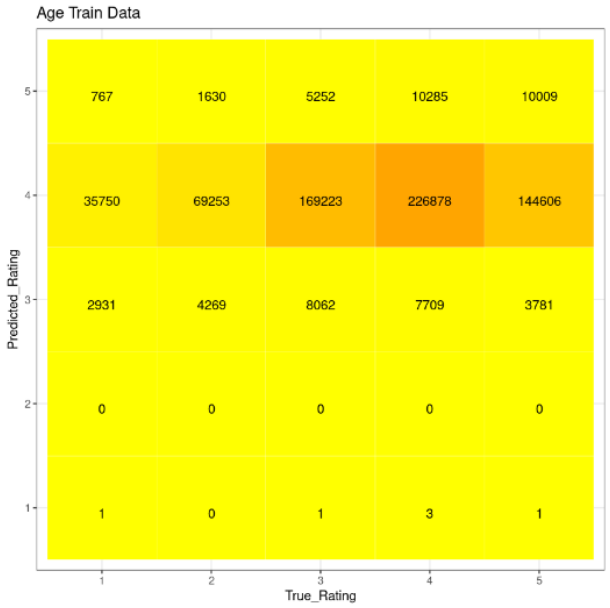
Starting with the gender set of data, this was the only data segment to converge on an error value during the regression process, meaning that after many iterations, it

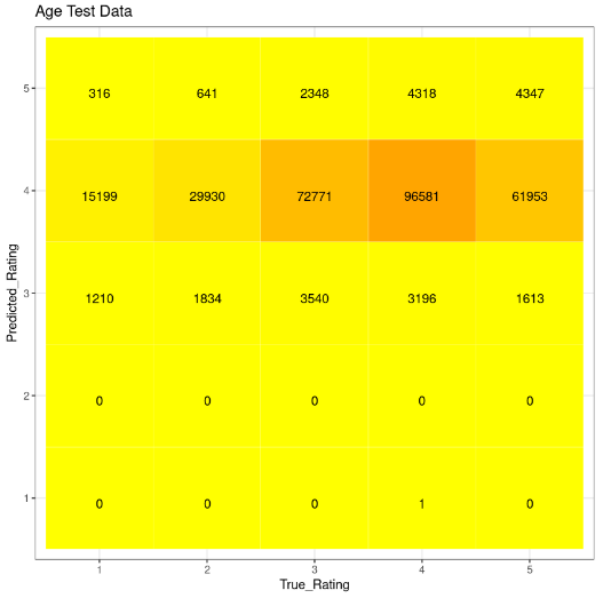
finally determined that the model could not be any more optimized with the given parameters (See Appendix M). But once the model was complete, we obtained an accuracy score of about 35% (this will be a common theme throughout). If we want the data to be significant on some regard, we will look for an accuracy value of at very least 70% but ideally in the 80-90's. This is because with a 70% accuracy, the model is more correct in its guesses than not, however having a 35% accuracy is not much better than what a random guess would be. The accuracy score was obtained through the resulting confusion matrices, which allows us to have a better look at where the model gained its inaccuracy.

As you can see, the model did not predict a single rating of 1-star or 2-stars and instead concentrated most of its predictions on 3-stars and up. This is most likely due to the composition of the ratings being very imbalanced. However, we can notice that the true 4-stars amount consistently high between the train and test matrices, which correlates to the high concentration of 4-star ratings that we noted in the exploratory

data analysis. After sweeping through the corresponding p-values for the regression model, we also are 95% confident that every variable in the model is statistically significant.

Looking at the age data set, we came to similar conclusions as the gender data set. Unlike the other segment, this set did not converge however the error stayed around the same amount (See Appendix N). After the model was generated, we also followed it up with some confusion matrices to evaluate the performance and the accuracy score was also about 35%.





These matrices also highlight that the model did not have much to consider when regarding lower ratings, but this time there were a few that squeaked by. This time around though there seems to be slightly more overall 5-star predictions and slightly less overall 3-star, but the main concentration still hovers around the true 4-star

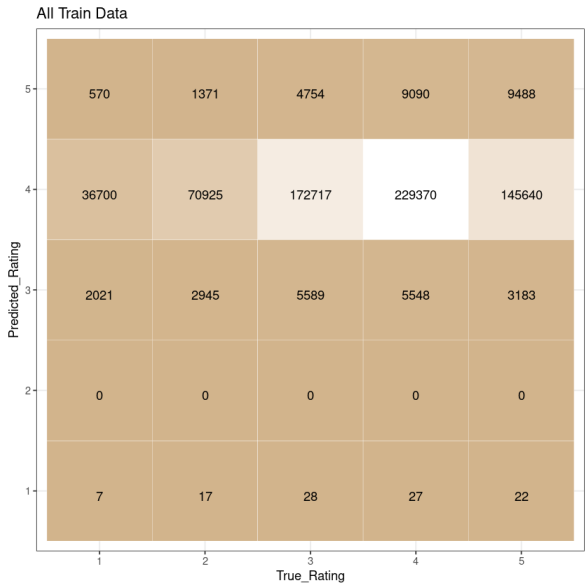
area. With this model we are also 95% confident that every variable in the model is statistically significant.

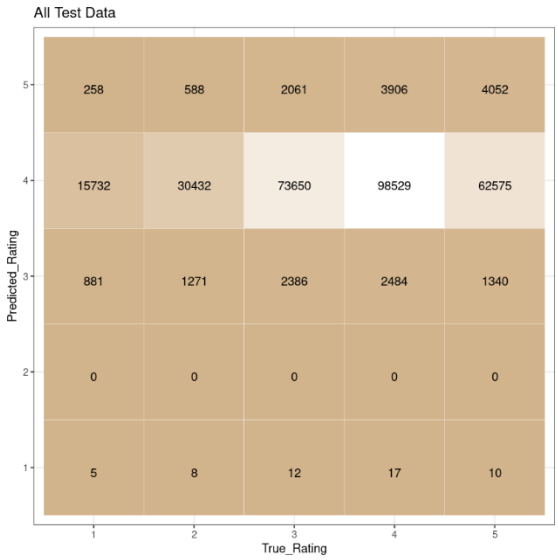
Finally, we looked at the occupation data set which also did not converge in the regression process. In a similar fashion to the age data set, the error value approached the general error for the gender data set, so chances are, given a few more iterations, the model would nearly have converged (See Appendix 0). The resulting confusion matrices from the model once again provided a 35% accuracy score.

The same trends are once again followed, with there being substantially less 1 and 2-star ratings but most of them being centered around true 4-star.

Given that the data thus far does not seem like any are that significant, I decided to do one final regression with all variables included to see if the combination of them

would lead to a more accurate model. The confusion matrix of which is seen below but unfortunately, it still seems to be that the same trend repeats itself.





Ultimately, we also determined the log odds equations for each data set, which is comprised of each variable and their own unique weight value representing the importance and effect of that variable in the rating. Snippets of each will be shown in chart form followed by the full chart for proportion perspective in the appendix. The log odds values are beneficial as they help us determine how much the model prefers any given rating over a rating of 1. Therefore, the log odds of each rating above a two were positive and in favor of the higher rating.

ANALYSIS

Overall, from the data collected, the models that I have created seem to show that there is not a strong relationship between any of the demographic values nor the genres and the rating of a movie. Given that about 35% of all ratings were 4-star ratings and our models consistently can only have a 35% accuracy score, even the model seems to understand that predicting a rating of 4 is simply the most optimal you can get. There could be many reasons as to why this could have happened this way so let's reflect on the process as a whole so we can see what happened.

First of all, to look at why this happened we need to understand the severity of the results. One common theme through all of the models is the fact that relatively no predictions of one or two stars were made. This is the unfortunate side effect of not being able to normalize the pool of data values. In other words, because the observations were so heavily skewed toward higher ratings in terms of frequency, the models were especially unsure on how to handle what makes a movie rate poorly, and the lack of extra variables such as budgets, directors, etc. only exaggerated the issue. This proposal is only exaggerated by looking at the most frequent correct guess, which is a rating of 4 stars unanimously across each model. Coincidently, looking at the many graphs from the EDA will show that in virtually every single way to slice the data, there was always the highest frequency of 4-star ratings. This allowed the regression model to have way more experience determining what could make a movie a 4-star film and therefore it would look for those patterns in the data.

Consequently, this could also indicate that neither gender nor age nor occupation can truly determine the score of a film based off the genres of it because there seem to be other factors that make them good or bad. Due to the nature of logistic regression, these models saved countless hours of testing each individual grouping of gender, age, and occupation but if the model still says the resulting regression is not accurate, then these factors clearly do not have that much influence with the given data. To go back to a previous statement, lets take the idea of a "chick flick" back into consideration. Many of these films would be considered dramas but that doesn't mean that all women who watch a drama automatically think its iconic. Each film could cover a different tone or theme that goes even further than a genre. Coming-of-age, family struggles, love pains, all of these themes are common in chick flicks, but they aren't necessarily genres. These are some of the other reasons why people may choose to like a film—because they can resonate with the meaning and purpose of the film. This is where a variable such as a synopsis would add more possibly factors. A sentiment analysis score of that brief description would probably end up being a better correlator.

Aside from pure sensical factors, there are some numerical and computational factors that need to be considered in the analysis as well. For example, one variable potential cause is the split amount of the test and train data. Similar to the reasoning for wanting at least a 70% accuracy score, I decided to do a 70-30 split of test and train data to ensure that I was giving the machine ample amounts of information to learn from and following it up with a brief, yet complex, set of data to test its effectiveness.

Overall, this factor is one that I would brush off as not a significant point of error as the split I elect to use if rather common in studies and accuracy scores that low would not be significantly affected by tame split data.

Another computational factor that probably ended up causing a much bigger effect on the data was the regression technique itself. This is partially due to my inexperience to multinomial regression in coding. While I am somewhat confident that the method I employed is nearly optimized for this kind of situation, there is the chance that I chose an incorrect path or library to use when dealing with a dataset of this caliber. Similarly, I tried to adjust the maximum number of iterations the model would undergo to reduce the error value but I was unable to understand the reference material of the package so the iterations were kept at the default 100. While this may not have been that big of an issue due to the fact that the gender model converged (See Appendix M) and the occupation model was seemingly near convergence (See Appendix O), this could still have been a significant factor.

Overall, I do not believe that any minor computational adjustment would have saved the accuracy of any model and I think the only saving grace would simply be testing different parameters or using a different regression technique. Even at surface level, determining that you cannot predict the success of a film based off of only your age, gender, job, and the genre of the film is not the craziest concept out there. More likely than not, we would need some additional information or to repurpose our analysis to find a different connection.

ETHICAL RECOMMENDATIONS

While the effects of my specific results may not have major ethical applications, broadening the topic to movie ratings and recommendation in general can allow for a better discussion to be made. Rating movies is effectively exploring the subjectivity of art and because of that, the ethics in question revolve around expression of feelings. While constructive criticism in art is not immoral or negative, the argument could be made that ignoring art based only off of others input can be disrespectful. Similarly, judging a movie based purely off of a predicted rating is also not fair to the creator of the film and the results from this experiment, if they were significant, could have threatened the appreciation of a given piece.

Despite this discussion, there are clearly better recommendation systems that have been implemented in the popular streaming platforms today. Netflix, HBO, Hulu, and many more have their own algorithms set in place as an attempt to keep their users interested in exploring the library of films they offer. Does this mean that they are ruining certain films if they are not recommended by the engine? This also leads to the dilemma that is the saturation of the film industry. The sheer number of films available to the public through whatever medium is entirely too much content for one person to consume so the recommendation engines produced by these companies exist not necessarily to shun the “bad” movies but more help the user to find films that they would enjoy watching without forcing them to undergo trial and error. This purpose is even reflected in most open forum reviews as when they write their

reviews, they do so in respect to a potential viewer, communicating the importance of the movie and if the film was worth their time. The ethical issue that would spawn from these recommendation machines (which some forum websites already propagate) is if they allow for the review of an unreleased film. There the issue becomes more apparent as potential viewers are discouraged from exploring a film purely off of the speculation of others. Rumors rarely lead to good reputation and even if they end up false, they can harshly tarnish the work of many people without any justified reason, which is unethical.

With these situations in mind, it gives a new outlook on the potential ramifications of this project. Given that the process attempted provided a prediction of the rating rather than predicting if a user would like the film, we are running the risk of being ethically unsound in the creation of this project. The current prediction output could possibly allow people to discount a film based purely off of simply genre, age group, gender, and occupation, which is clearly not fair to the film. Instead, the results of this project should be used more or less as a foundation to creating a more robust recommendation system whose intent is to offer suggestions of movies rather than predicting the success of one.

CHALLENGES

With the confident and hopeful approach to the problem from the early planning stages to the data visualizations, there was clearly many changes, some unavoidable,

sometime sensitive, but all of them ended up shaping the project into what it became, which while fell short of initial hopes still came to some kind of conclusion.

The first set of challenges came with the exploratory data analysis. Unlike some of my peers, my dataset was virtually the same dataset that I downloaded from the website i.e., there was no filtering that needed to occur to eliminate irrelevant observations. While I did not employ all of the fields provided in the dataset, the total entry count stayed at 1,000,000. This was already a heavy burden on my computing capabilities and in-turn, my runtimes per execution were taxing on my schedule. As I was rusty in my R and Python data munging skills, I decided to resort to Microsoft Excel for the EDA, which was a more comfortable environment at the time, despite it crashing my laptop whenever I tried to enter a reference to a cell. While Excel was able to provide what I needed and generate the graphs I wanted, the whole data munging and exploration process can easily be mitigated in other attempts by investing more time in the R and Python environments or other data visualization software.

The next set of challenges involved the actual implementation of the methodology: the bread and butter of the project. These challenges were mainly due to my novice nature in multinomial logistic regression. I previously had some idea of multinomial logistic regression but the more I learned, the more difficult it was to apply what I learned to a project of this scale and proportion. I suddenly had to investigate runtime complexities, different R and Python libraries, and eventually I landed upon a method that worked and allowed me to do some kind of analysis on a

project of my side. But that analysis then turned into the not-very-significant outcome that was shown in my data visualizations. At that point, it was also simply defeating just knowing that there was some *thing* just not clicking in the data and finally that became the conclusion. If I were to reattempt this analysis, I would probably resort to possibly looking into performing a neural network instead of multinomial logistic regression. While the topic is also somewhat new to me, it seems like it is more averse to larger datasets and may offer more flexibility in execution.

RECOMMENDATIONS/NEXT STEPS

Overall, this analysis came to inconclusive results and inaccurate data models, but that is not necessarily a bad thing. This project pushed my coding skills and understanding of data way out of their comfort zones and thankfully the project idea was in my interests in exploring. In the end, I learned many lessons on the process of data analysis and also realized that not all data sets will work for a given regression model.

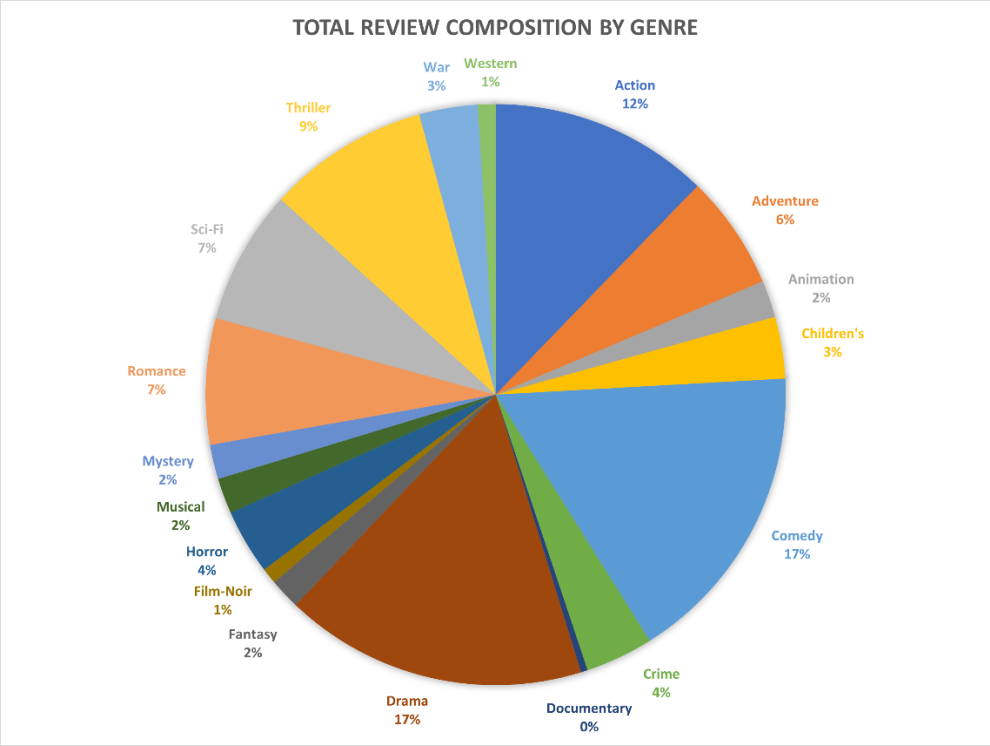
I would say the main aspect that I would have changed for this entire process would have been to use invest more time in considering the direction and end goal of the project. As I mentioned in the ethical recommendations, perhaps treating the data with an interpolative mindset, like correlating the movie scores with each other, instead of an extrapolative one, like predicting the rating of a hypothetical film, would have yielded stronger conclusions and more insight than what was determined here. Additionally, spending more time to dive deeper into regression methods and example of their implementation would have significantly reduced the workload and hours spent trying to understand how to implement and analyze the models.

As alluded to in the challenges section, if I were to recomplete this project, I would have traded multinomial logistic regression for a neural network, as the size and complexity of the data would probably not be as intense in a neural network. Additionally, neural networks thrive on a large amount of input data, and I believe one million observations would fit that niche. However, if I did stay with the logistic regression approach, I would try to reframe the questions or try to investigate the data in such a way that some simple logistic regression could be involved. Not only would this significantly reduce the variability of the answers we would achieve but it would also, consequently, probably result in a better accuracy score. Perhaps looking at each star rating individually and trying to classify a film in that regard. This would demand much more time and investment but there would probably be some way to automate through the many different graphs and models needed to cover the dataset.

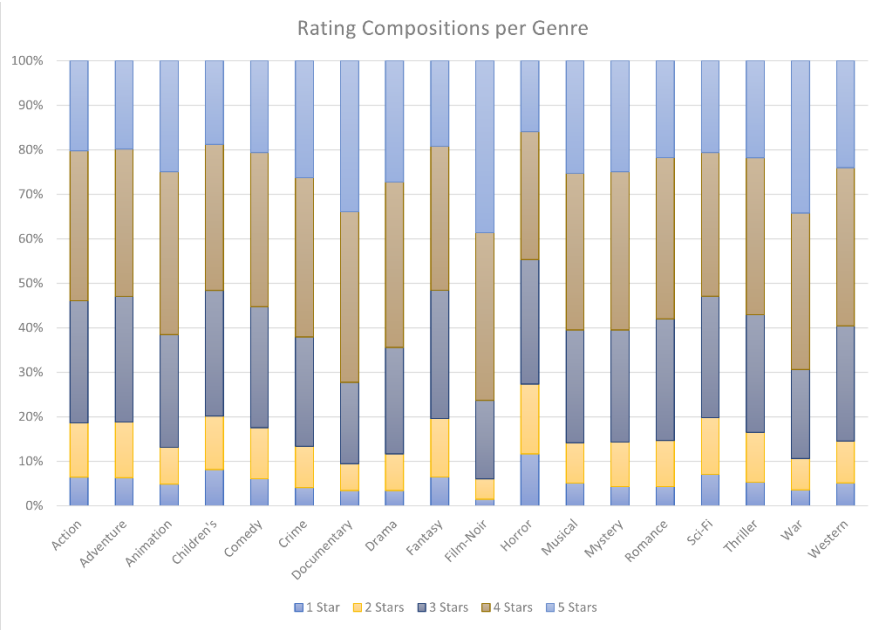
Finally, I think in general, while I did enjoy looking at the MovieLens dataset and appreciated its simplicity, perhaps a different data set would have been better, especially one with quantitative data. Limiting my options to exclusively categorical analysis was very taxing and I feel it limited my potential to explore the data appropriately.

APPENDIX

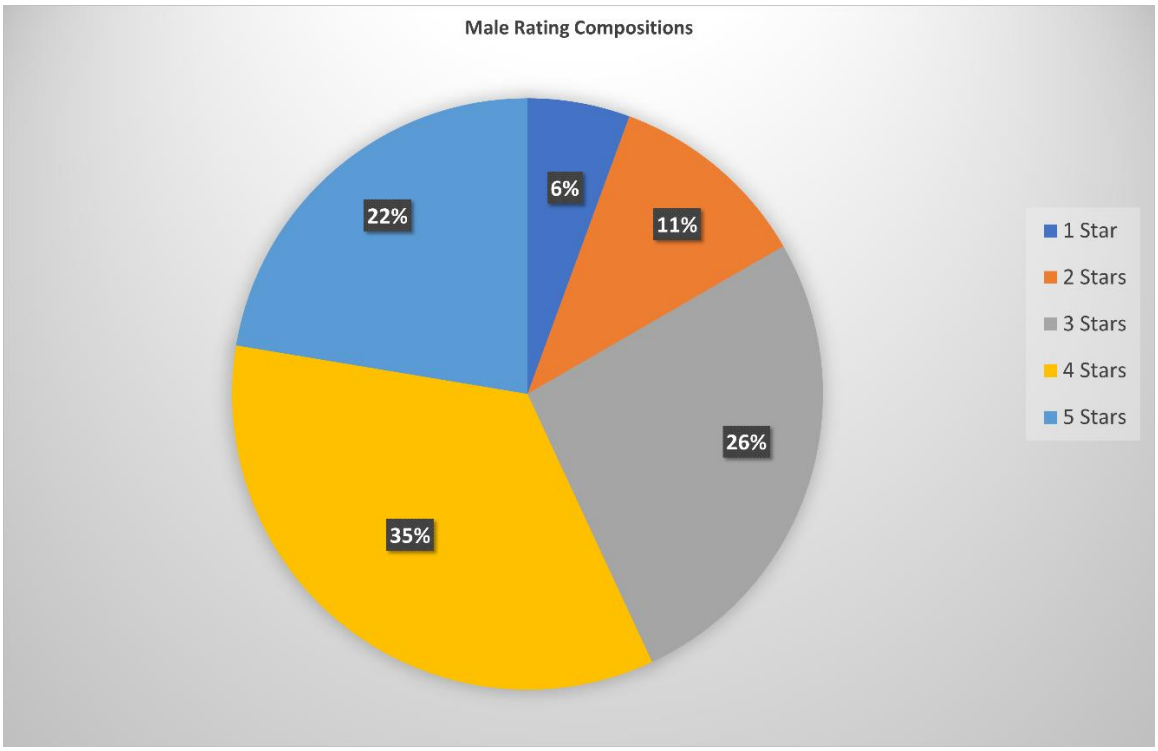
A. "Total Review Composition by Genre"



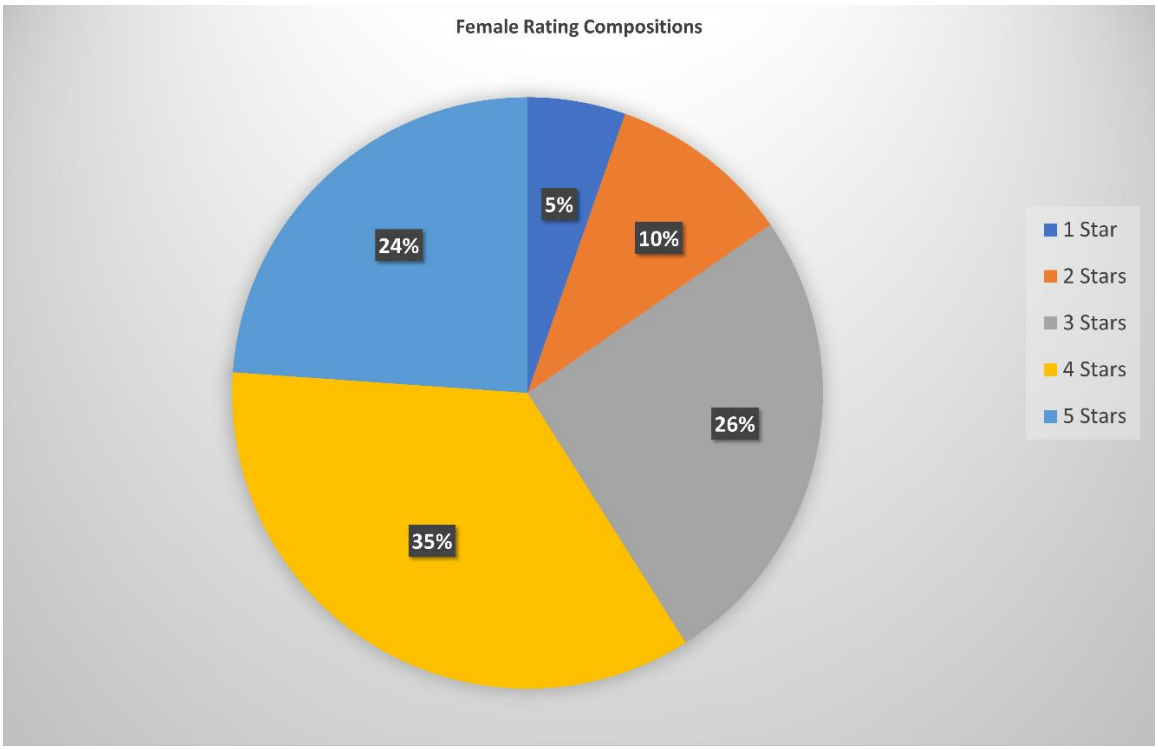
B. "Rating Compositions per Genre"



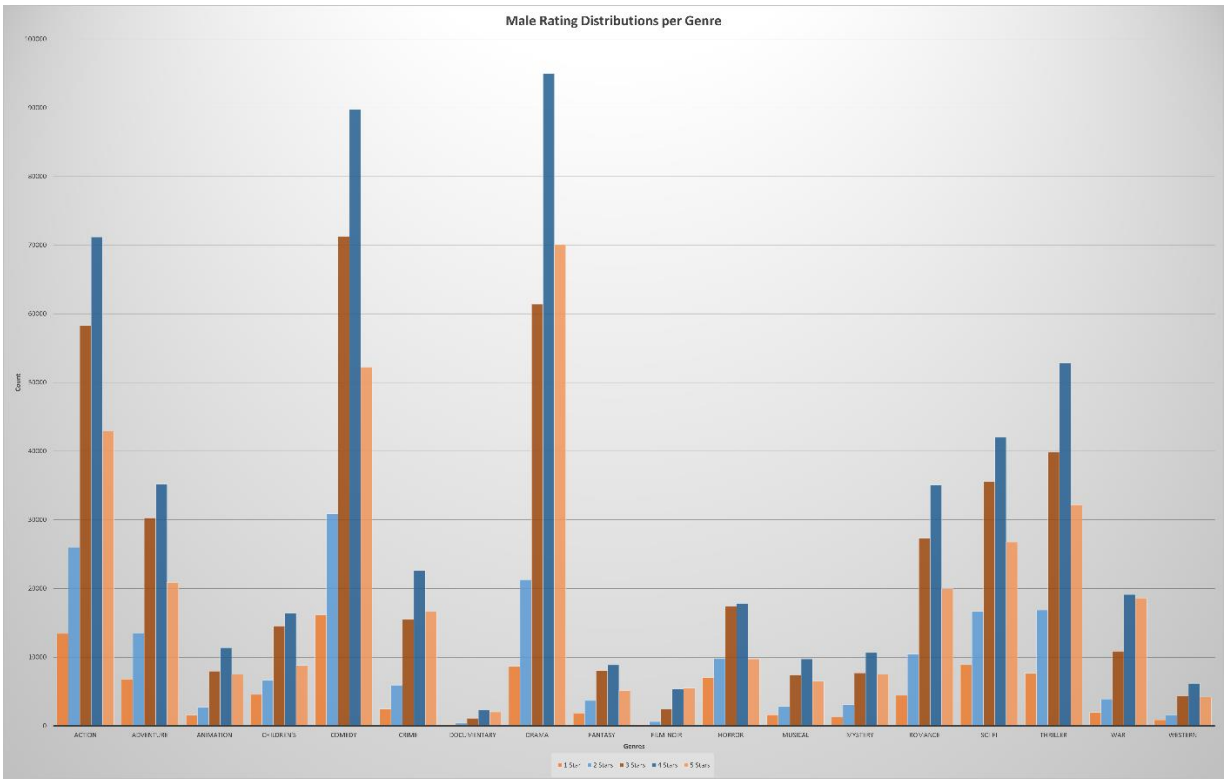
C. “Male Rating Compositions”



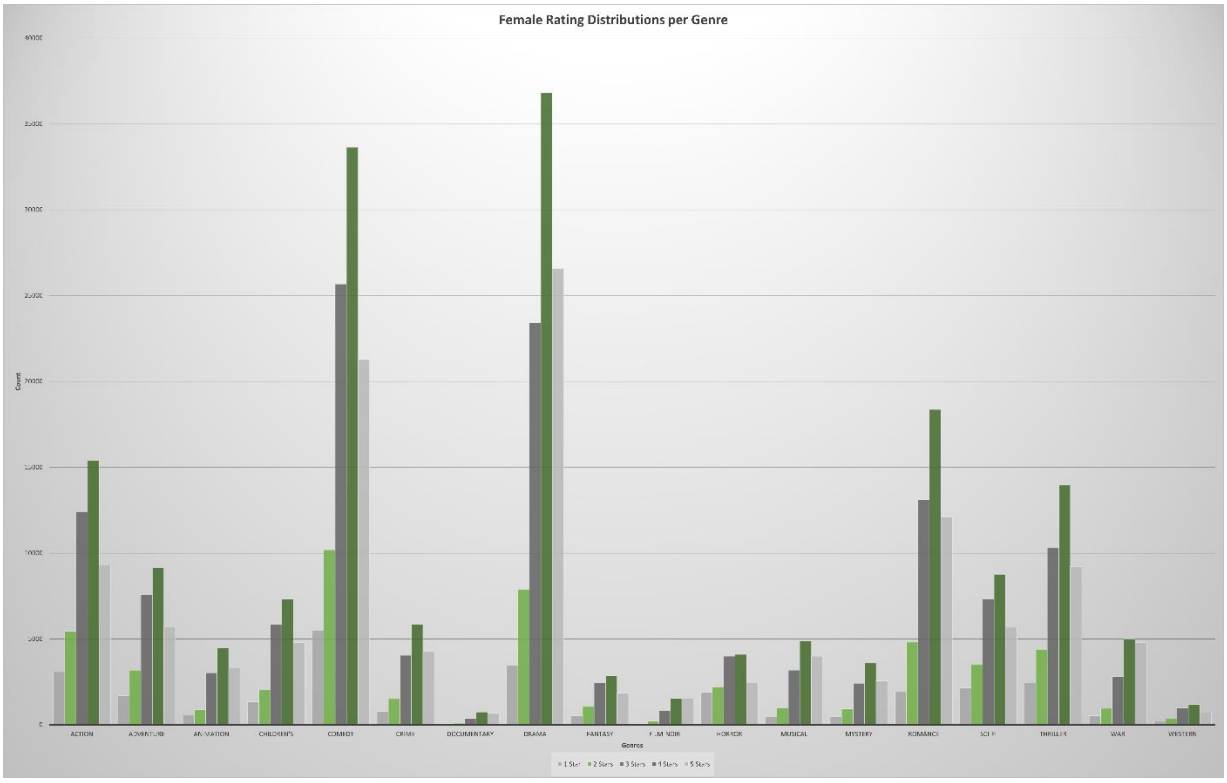
D. “Female Rating Compositions”



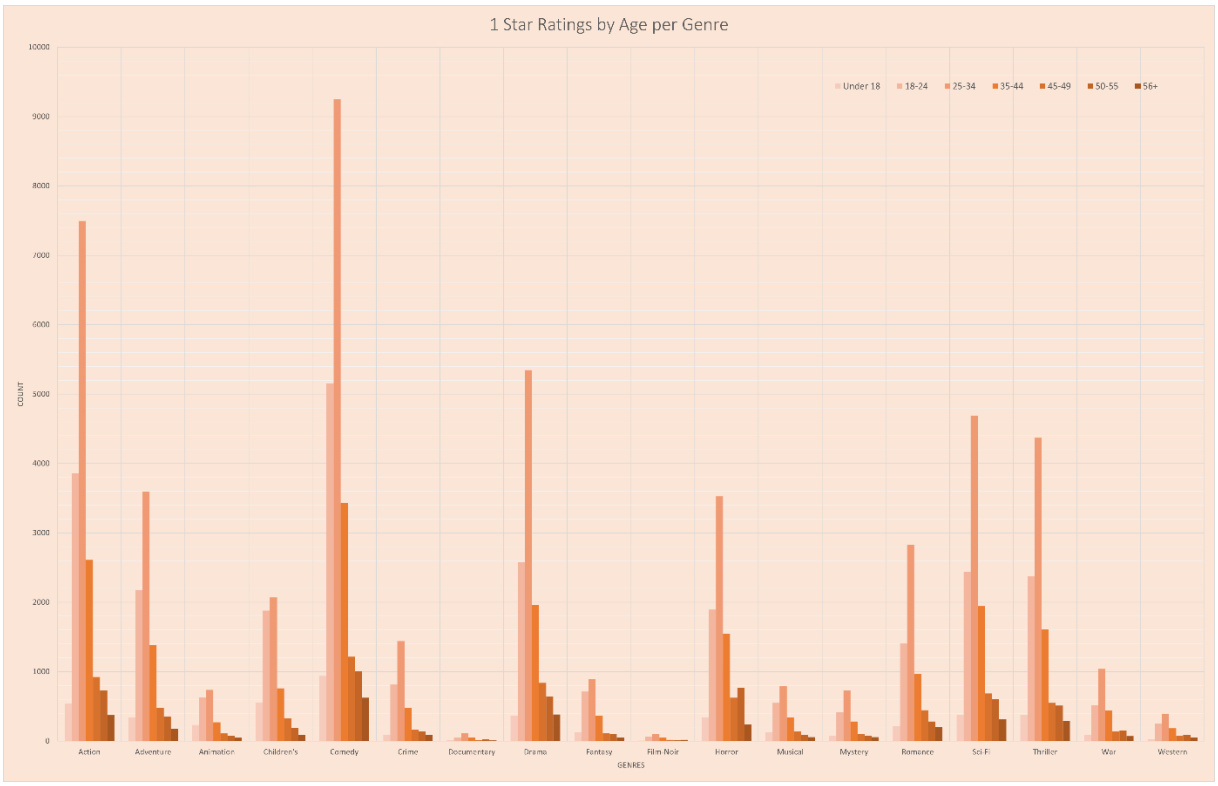
E. “Male Rating Distributions per Genre”



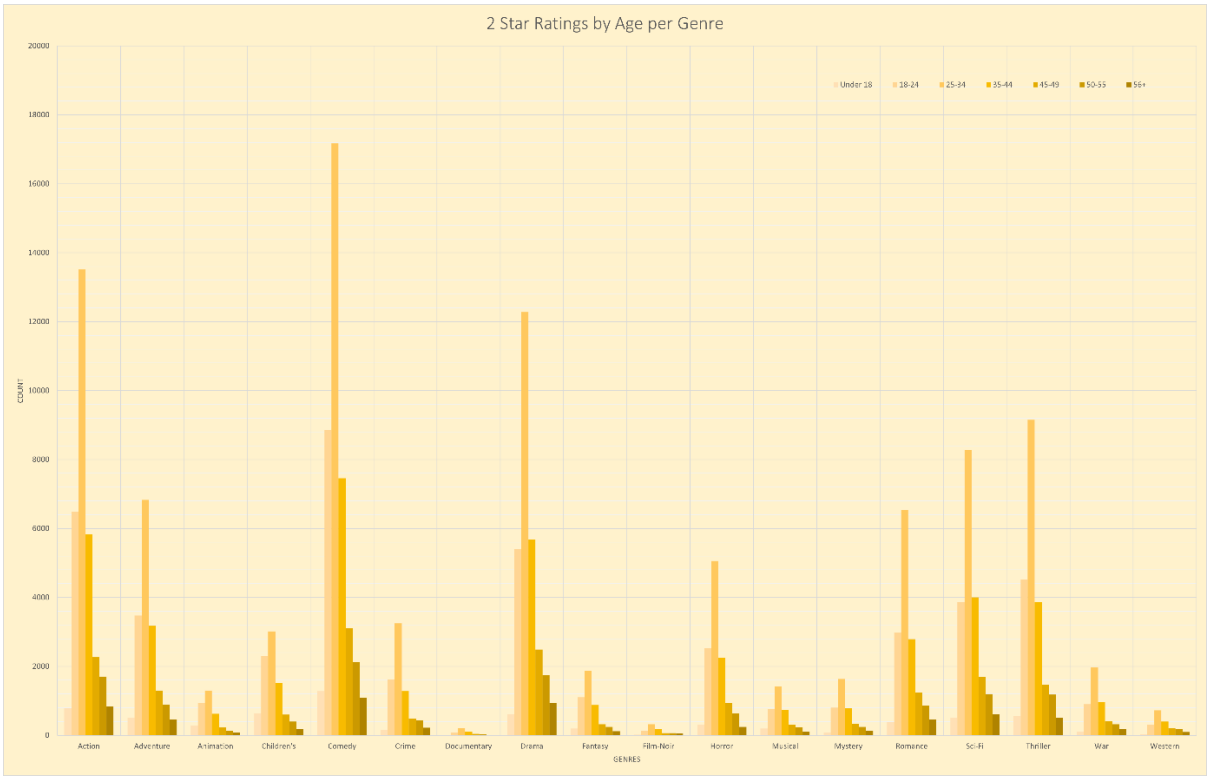
F. “Female Rating Distributions per Genre”



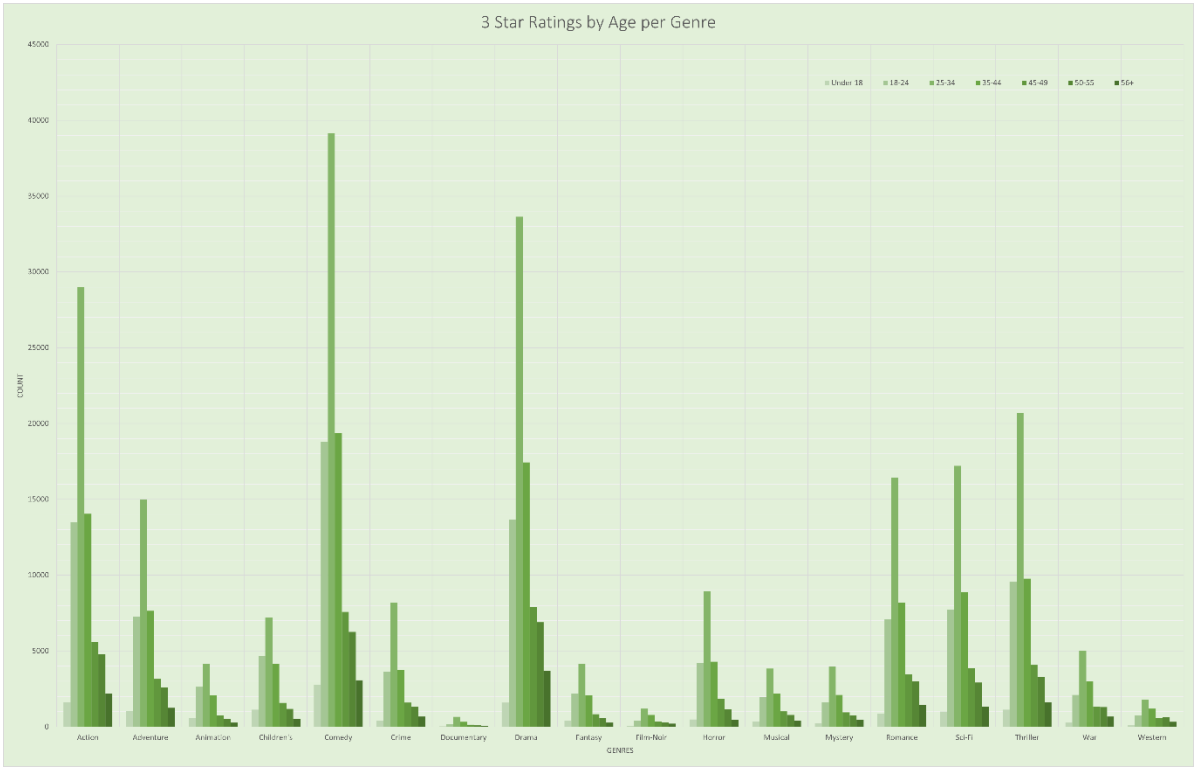
G. “1 Star Ratings by Age per Genre”



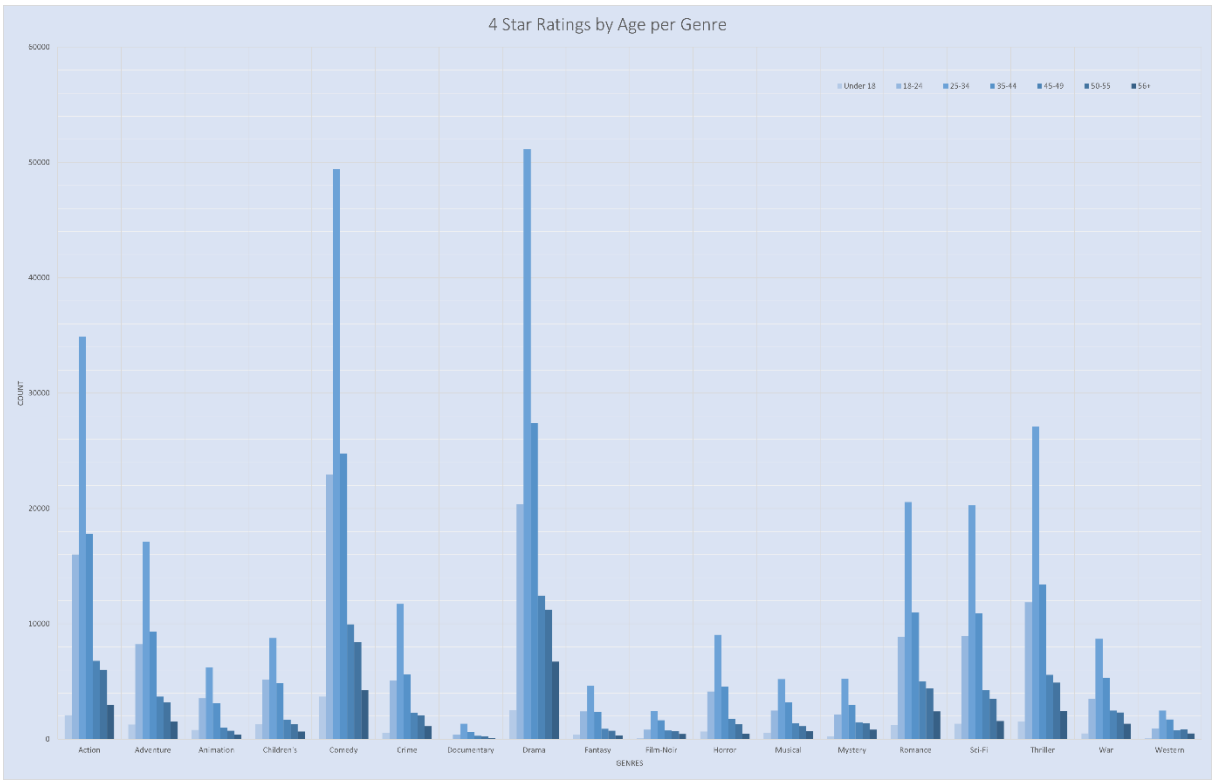
H. “2 Star Ratings by Age per Genre”



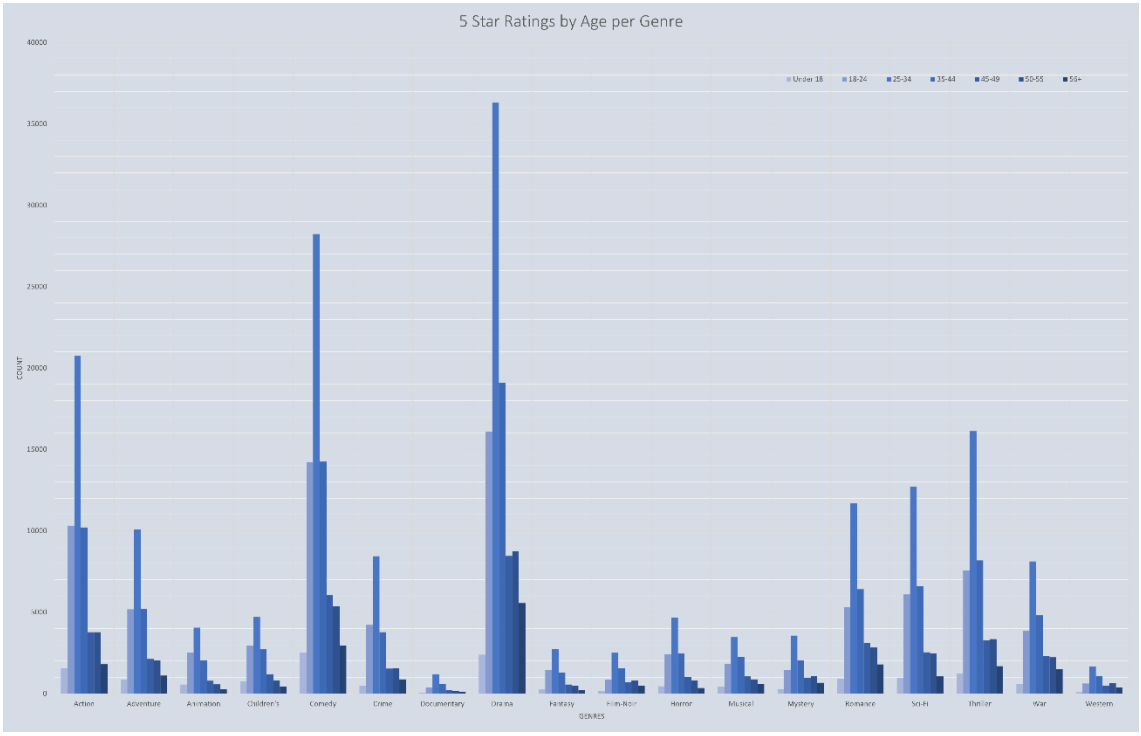
I. “3 Star Ratings by Age per Genre”



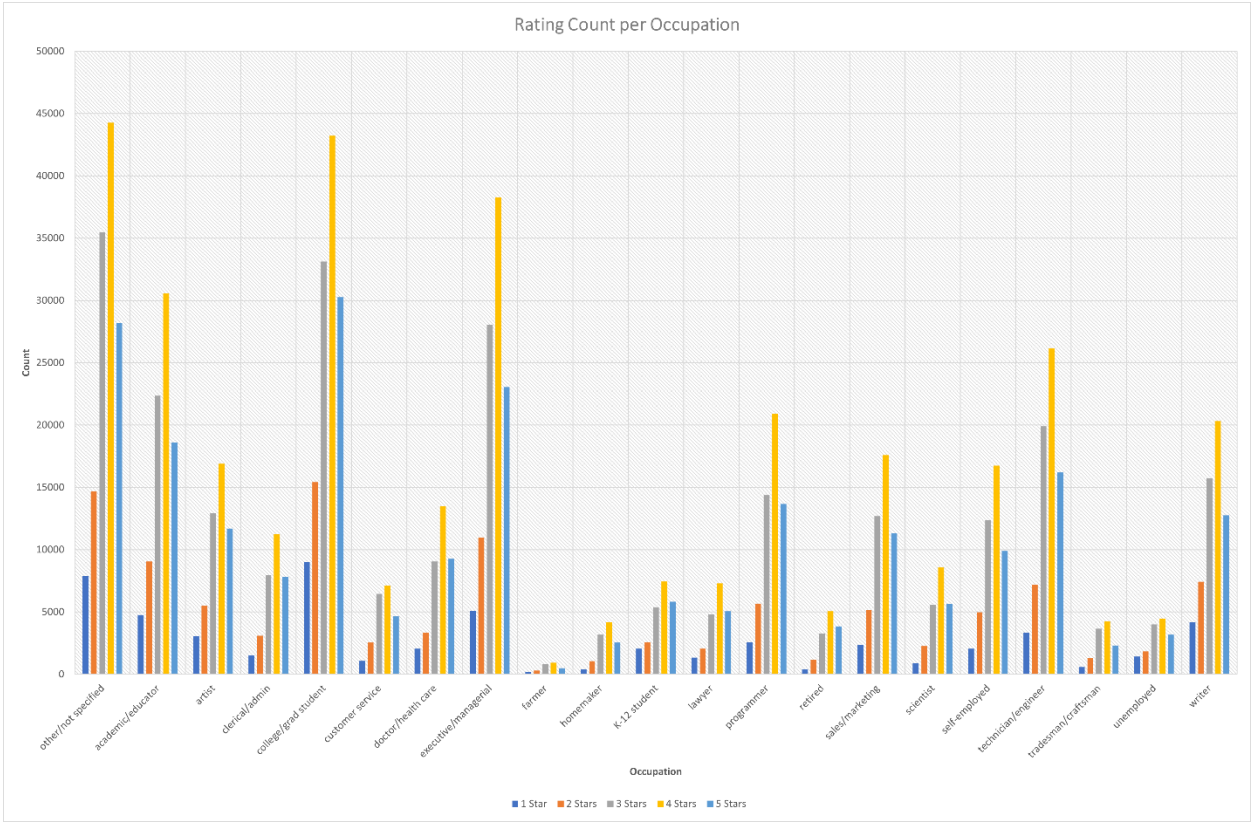
J. “4 Star Ratings by Age per Genre”



K. “5 Star Ratings by Age per Genre”



I. “Rating Count per Occupation”



M. Gender model error reduction iterations

```
# weights:  105 (80 variable)
initial  value 1126933.254611
iter  10 value 1026724.837500
iter  20 value 1025337.181886
iter  30 value 1024905.941414
iter  40 value 1022930.937124
iter  50 value 1018538.098390
iter  60 value 1009251.921507
iter  70 value 1007749.838305
iter  80 value 1006467.005190
iter  90 value 1005108.441337
iter  90 value 1005108.440390
iter  90 value 1005108.440386
final  value 1005108.440386
converged
```

N. Age model error reduction iterations

```
# weights:  130 (100 variable)
initial  value 1127268.017696
iter  10 value 1033983.665595
iter  20 value 1032815.567743
iter  30 value 1032349.911803
iter  40 value 1028732.527724
iter  50 value 1022442.109856
iter  60 value 1013182.887205
iter  70 value 1010758.046289
iter  80 value 1009341.651985
iter  90 value 1007648.393605
iter 100 value 1005610.362760
final  value 1005610.362760
stopped after 100 iterations
```

O. Occupation model error reduction iterations

```
# weights:  200 (156 variable)
initial  value 1125475.103862
iter  10 value 1037159.685389
iter  20 value 1035765.174489
iter  30 value 1035220.819119
iter  40 value 1030993.430783
iter  50 value 1021329.208082
iter  60 value 1014098.912676
iter  70 value 1012157.845547
iter  80 value 1010328.001540
iter  90 value 1009850.854165
iter 100 value 1006713.699229
final  value 1006713.699229
stopped after 100 iterations
```

P. Gender data log odds chart:

Rating	(Intercept)	GenderM	Action1	Adventure1	Animation1
2	0.462494	0.084563	-0.05923	0.110316	0.243987
3	1.337904	0.046072	-0.11993	0.149033	0.625199
4	1.618934	0.014554	-0.2108	0.115433	0.900895
5	1.121509	-0.03688	-0.32262	0.1148	1.14481

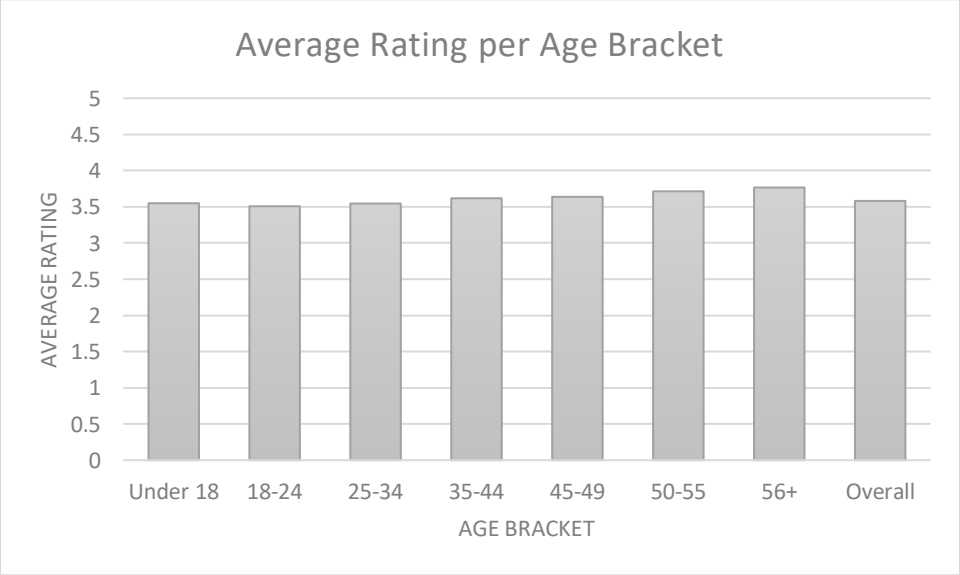
Q. Age data log odds chart:

Rating	(Intercept)	Age.Bracket25-34	Age.Bracket35-44	Age.Bracket45-49	Age.Bracket50-55
2	0.480887	0.005627	0.244564	0.251569	0.200356
3	1.092714	0.162356	0.49419	0.49158	0.690375
4	1.403607	0.099951	0.454888	0.437762	0.658936
5	0.855311	0.097222	0.434961	0.446315	0.752018

R. Occupation data log odds chart:

Rating	(Intercept)	Occupationartist	Occupationclerical/admin	Occupationcollege/grad_student	Occupationcustomer_service
2	0.067291	0.350143	0.816733	0.328488	0.557928
3	1.064466	0.258613	0.682004	0.222825	0.779408
4	1.354128	0.198331	0.642736	0.231381	0.712595
5	0.646953	0.35341	0.82538	0.302037	0.732779

S. Average Rating vs Age Bracket



T. All model error reduction iterations

```
# weights:  235 (184 variable)
initial  value 1126625.851969
iter  10 value 1023638.557923
iter  20 value 1021571.805842
iter  30 value 1021101.964435
iter  40 value 1020668.449481
iter  50 value 1019005.250069
iter  60 value 1017719.157799
iter  70 value 1013606.598660
iter  80 value 1010316.738468
iter  90 value 1006931.227671
iter 100 value 1006186.556999
final  value 1006186.556999
stopped after 100 iterations
```

U.

Rating	(Intercept)	GenderM	Age.Bracket25-34	Age.Bracket35-44	Age.Bracket45-49
2	0	0.00E+00	0.952239352	1.51E-11	0
3	0	8.72E-13	0.001144843	2.22E-16	0
4	0	0.00E+00	0.020666724	0.00E+00	0
5	0	2.30E-03	0.147132334	0.00E+00	0

REFERENCES

About MovieLens. Movielens. (n.d.). Retrieved from https://movielens.org/info/about

Boyle, K. (2015, November 12). Gender, comedy and reviewing culture on the Internet Movie Database. Stirling Online Research Repository. Retrieved from https://dspace.stir.ac.uk/handle/1893/22497#.Yyfq6nbMKbg

Gershman, A., Meisels, A., Lüke, K.-H., Rokach, L., Schclar, A., & Sturm, A. (1970, January 1). *[PDF] a decision tree based recommender system: Semantic scholar*. Semantic Scholar. Retrieved from https://www.semanticscholar.org/paper/A-Decision-Tree-Based-Recommender-System-Gershman-Meisels/adb6e8f0fc75985d63e2bb34a31e49b1545d81b4

Grouplens Research. Crunchbase. (n.d.). Retrieved from https://www.crunchbase.com/organization/grouplens-research

Simon, R. W., & Nath, L. E. (2004, March 1). *Gender and emotion in the United States: Do men and women differ in self-reports of feelings and expressive behavior?* American Journal of Sociology. Retrieved from https://www.journals.uchicago.edu/doi/10.1086/382111

Yalcin, E., & Bilge, A. (2021, April 26). *Investigating and counteracting popularity bias in group recommendations*. Science Direct. Retrieved from https://www.sciencedirect.com/science/article/pii/S0306457321001047?casa_token=XFBQkvkX6CwAAAAA%3ArnVajs1jk9EbRlLYdB1B_vdGE6oP8GSqMf8c7GQn8VtVAgO6nC4rNbJsn7BB_48sbNzZm5fqk3k

Dataset Reference

Apparently sometime during the formation of the report, the original website containing the dataset has been taken down or restructured to where my original links are no longer active. To my knowledge, the citation below is still the same data set, just hosted on Kaggle instead of MovieLens themselves.

Golden, O. (2021, January 23). *Movielens 1M Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/odedgolden/movielens-1m-dataset