Ardavasd Ardhaldjian, Joshua Levy, River Yan, & Anson Zheng
Data Wranglers
DS501 Case Study #3
10/28/21

# Final Report

## Table of Contents

## 1. Data Analysis Methodologies (Ardavasd )

Before performing any of the data analysis our team started by transforming the .dat files into individual pandas dataframes (using double colin as value separators). At this point we used many of the built-in functions within pandas to learn more about the data. Most notably '.head' was used to peak into the first few rows; '.shape' was used to find the dimensions of our sets; and '.unique' was used to find number of duplicate rows;

```
[4]: print(movie.head())
     print(rating.head())
     print(user.head())
```

```
   MovieID                               Title                        Genres
0        1                    Toy Story (1995)   Animation|Children's|Comedy
1        2                      Jumanji (1995)  Adventure|Children's|Fantasy
2        3             Grumpier Old Men (1995)                Comedy|Romance
3        4            Waiting to Exhale (1995)                  Comedy|Drama
4        5  Father of the Bride Part II (1995)                        Comedy
   UserID  MovieID  Rating  Timestamp
0       1     1193       5  978300760
1       1      661       3  978302109
2       1      914       3  978301968
3       1     3408       4  978300275
4       1     2355       5  978824291
   UserID Gender  Age  Occupation Zip-code
```

**Figure 1**

```
# Explore Dimension of all 3 data sets
print(movie.shape) # ncol = 3, nrow = 3883
print(rating.shape) # ncol = 4, nrow = 1000209
print(user.shape) # ncol = 5, nrow = 6040
```

```
# Explore unique values of MovieID's in data
print(len(movie['MovieID'].unique()))
print(len(rating['MovieID'].unique()))
# There are 3883 unique values of MovieID in
# Thus, not all movies are rated
```

```
3883
3706
```

**Figure 2**

After the initial data frame analysis we combined all three of the dataframes into one master data frame using MovieID and UserID as primary keys.

```
Data = pd.merge(rating, movie, on = "MovieID")
Data = pd.merge(Data, user, on = "UserID")
```

```
Data.head(5)
```

| | Unnamed: 0_x | UserID | MovieID | Rating | Timestamp | Unnamed: 0_y | Title | Genres | Counts | Score | Unnamed: 0 | Gender | Age | Occupation | Zip-code |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 1193 | 5 | 978300760 | 1176 | One Flew Over the Cuckoo's Nest (1975) | ['Drama'] | 129 | 3.94 | 0 | F | 1 | 10 | 48067 |
| **1** | 1 | 1 | 661 | 3 | 978302109 | 655 | James and the Giant Peach (1996) | ['Animation', "Children's", 'Musical'] | 1 | 1.00 | 0 | F | 1 | 10 | 48067 |
| **2** | 2 | 1 | 914 | 3 | 978301968 | 902 | My Fair Lady (1964) | ['Musical', 'Romance'] | 676 | 3.91 | 0 | F | 1 | 10 | 48067 |
| **3** | 3 | 1 | 3408 | 4 | 978300275 | 3339 | Erin Brockovich (2000) | ['Drama'] | 43 | 3.74 | 0 | F | 1 | 10 | 48067 |
| **4** | 4 | 1 | 2355 | 5 | 978824291 | 2286 | Bug's Life, A (1998) | ['Animation', "Children's", 'Comedy'] | 43 | 2.51 | 0 | F | 1 | 10 | 48067 |

**Figure 3**

Additional data analysis was done using various python libraries and visualization packages such as Pandas, Numpy and MatPlotLib. The MatplotLib library was used to create bar charts, histograms and scatter plots. These visualizations were helpful at finding distributions in average ratings for age ranges and gender.

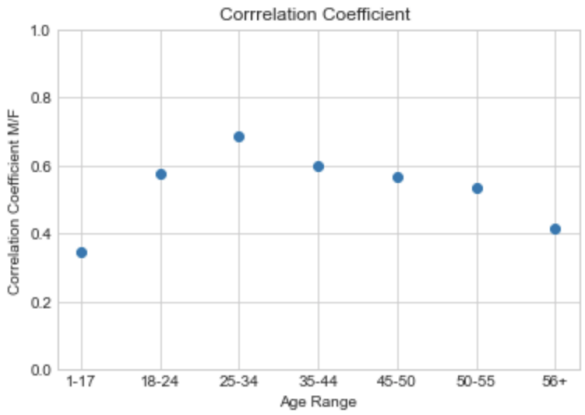


**Figure 4**

## 2. Findings (River)

To begin, there were several findings in the data after performing calculations in Problem 1. In the first part of Problem 1 we found that 29 movies had an average rating higher than 4.5. In addition, we found that there are also 29 movies that had an average rating higher than 4.5 that were rated by men. There were also 70 movies that had an average rating higher than 4.5 from the ratings of women. Another finding in the data was that 105 movies had a median rating over 4.5 among men over age 30, while 187 movies had a median rating over 4.5 among women over age 30. Another interesting finding from Problem 1 was that the ten most popular movies were Usual Suspects, Star Wars: Episode IV - A New Hope (1977), Shawshank Redemption, Schindler's List (1993), Silence of the Lambs, Godfather, One Flew Over the Cuckoo's Nest (1975), Raiders of the Lost Ark (1981), Saving Private Ryan (1998), and Sixth Sense. Furthermore, we found that the data distribution of gender was 888939 ratings for Male and 111270 ratings for Female.

We also found three interesting findings in the data for Problem 2. The first interesting finding was when we analyzed the average ratings of the movies and created bar plots. In the first graph of Part 2, we found that based on all the movies, the most ratings were between 3.5 and 4.0. As shown in Figure 4 below, the histogram shows the number of movies that received a rating on a 5-point scale. Most movies received a rating of about 4 and most movies had a rating above 3.

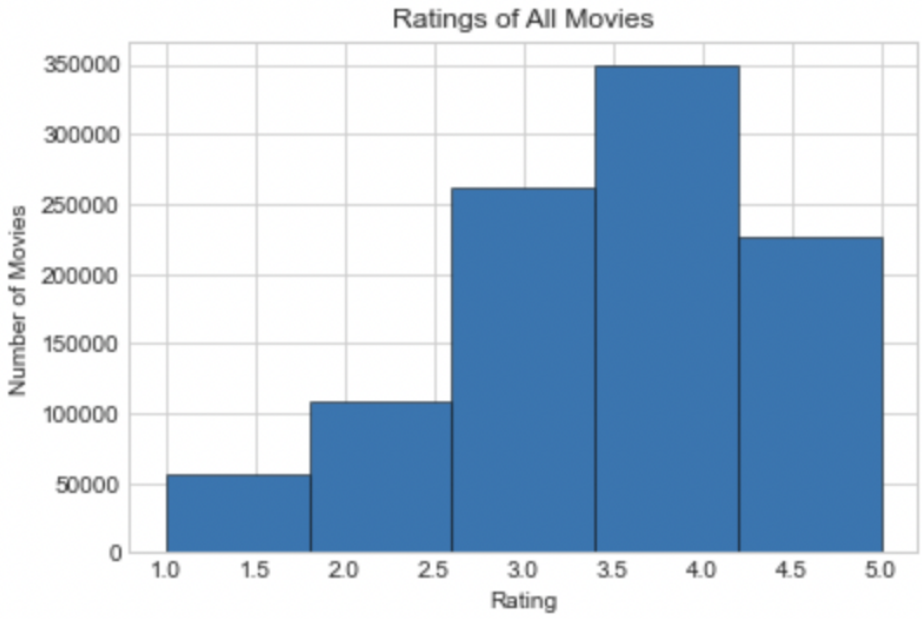


**Figure 5**

In the second graph of Part 2, we saw that depending on the movie, the number of ratings varied quite a bit between each other depending on what the movie was. As shown in Figure 5, the histogram gives the number of ratings each movie received. The most highly rated movie had close to 30000 ratings.

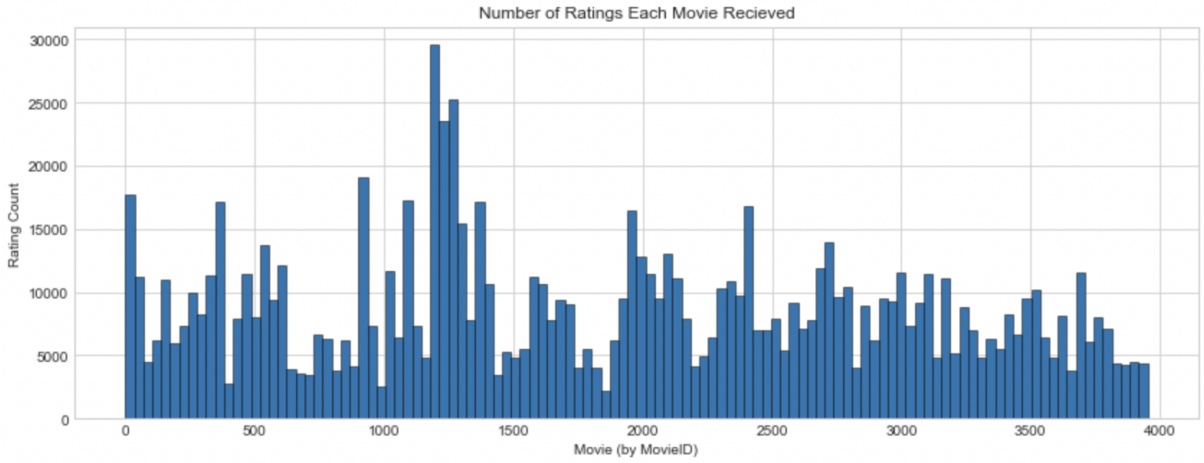


**Figure 6**

In the third graph of Part 2, we particularly noticed that the averaging rating for each movie had a similar shape to a normal distribution. The most popular range was also 3.5 - 4.0. As displayed in Figure 6 below, the histogram shows the average rating each movie received (on a 5 point scale). Majority of movies received a rating between 3 and 4.

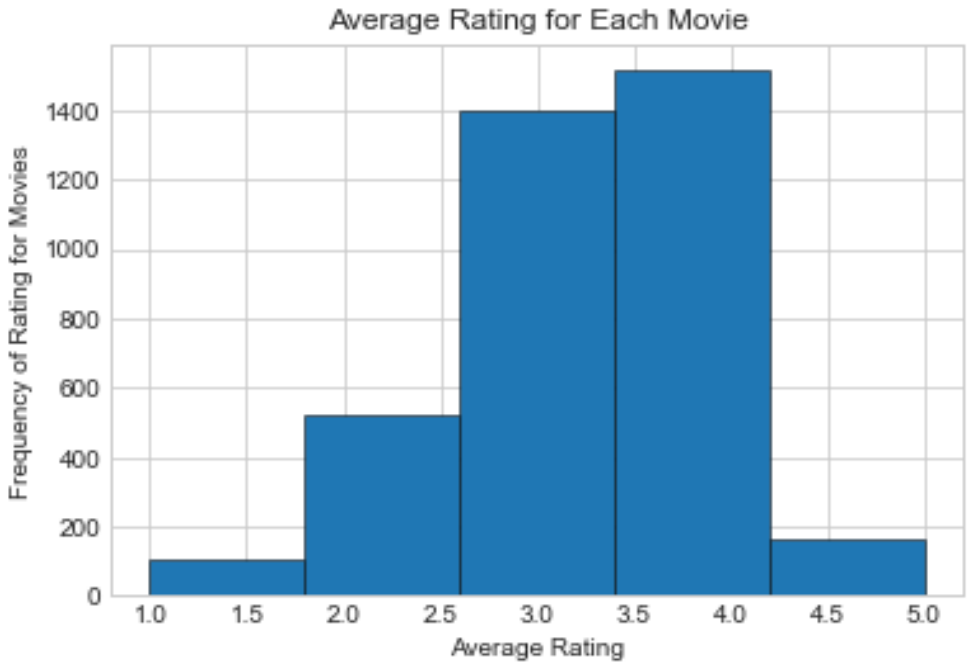


**Figure 7**

In the last graph of Part 2, we found that the amount of the data lowered because there was a filter added where the movie had to be rated for at least 100 times (Depicted in Figure 7). This is helpful because it reduces outliers and gets rid of movies with a lower number of ratings, which could affect the statistics of the data such as mean, maximum, and minimum.

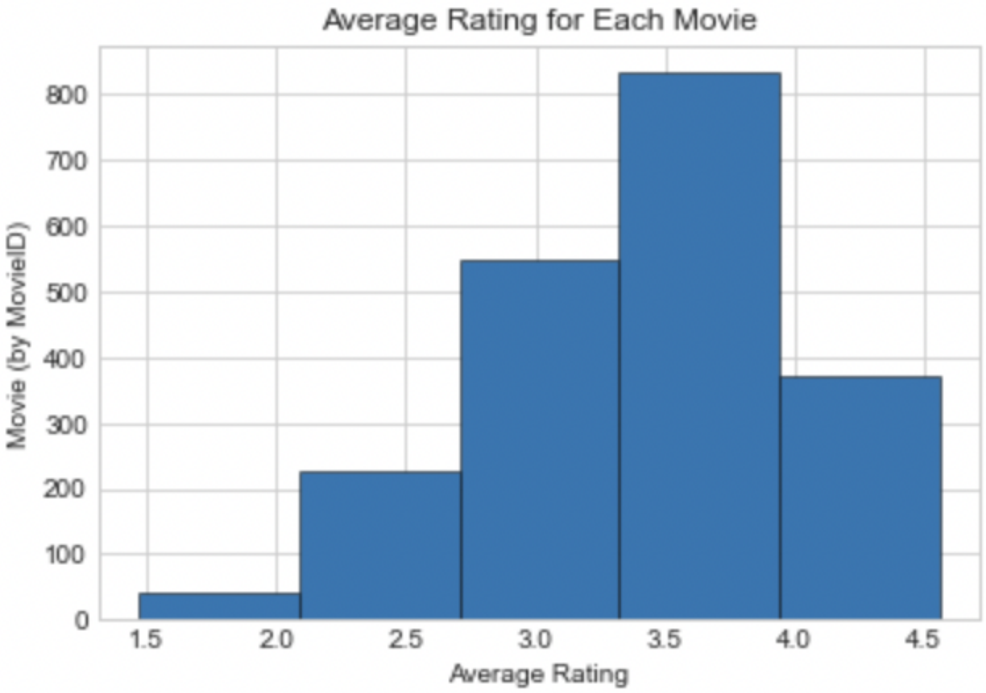


**Figure 8**

From the last question of Part 2, I was able to derive three major conjectures about age, gender, and genre. First, a conjecture about the gender distribution is that they are more likely to give a higher rating because of their average ratings. The null hypothesis is that the average ratings are the same for each gender, while the alternative hypothesis is that the average ratings were different for each gender. For females on average they give a rating of 3.62, while males on average give a rating of 3.57. As shown in Figure 8 below, the mean rating for the females was higher than males. I acquired this data by using Pandas' describe function and then sorting the means from highest to lowest.

```
Gender
F    3.620366
M    3.568879
```

**Figure 9**

However, one thing to keep in mind is that the overall dataset of 1 million entries is skewed a lot more towards males, where about 75% of the data are males and only 25% are females. This number will affect the average ratings. As shown in Figure 9 below, it displays the number of males and females in the entire dataset. This was derived from the value_counts function.

```
M    753769
F    246440
```

**Figure 10**

In addition, as shown in Figure 10 below, the male and female average ratings are very similar, where the males rate movies minutely higher.

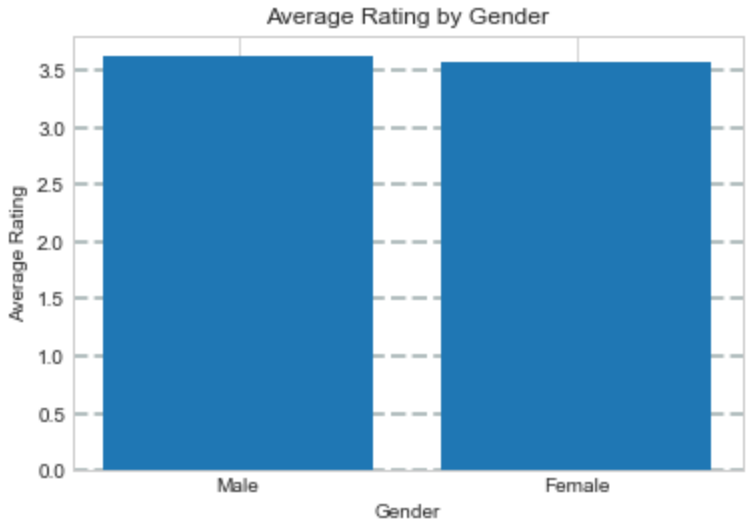


**Figure 11**

Second, a conjecture about the age is that the oldest people rate the movies the highest. Therefore, they are more likely to rate a movie a 5. The null hypothesis is that the average ratings are the same for each age, while the alternative hypothesis is that the average ratings were different for each age. While teenage to middle aged adults are the least likely to rate a movie a 5 because their average was the lowest. Lastly, little kids have the medium average where they are more likely to give a higher rating than teenagers and middle aged adults, but less likely to give a higher rating than the oldest people. As shown in Figure 11 below, the mean rating for the ages of 56 is the highest mean rating out of all the ages while the age of 18 had the lowest mean rating. I acquired this data by using Pandas' describe function and then sorting the means from highest to lowest.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Age** | | | | | | | | |
| **56** | 38780.0 | 3.766632 | 1.062551 | 1.0 | 3.0 | 4.0 | 5.0 | 5.0 |
| **50** | 72490.0 | 3.714512 | 1.061380 | 1.0 | 3.0 | 4.0 | 5.0 | 5.0 |
| **45** | 83633.0 | 3.638062 | 1.065385 | 1.0 | 3.0 | 4.0 | 4.0 | 5.0 |
| **35** | 199003.0 | 3.618162 | 1.078101 | 1.0 | 3.0 | 4.0 | 4.0 | 5.0 |
| **1** | 27211.0 | 3.549520 | 1.208417 | 1.0 | 3.0 | 4.0 | 4.0 | 5.0 |
| **25** | 395556.0 | 3.545235 | 1.127175 | 1.0 | 3.0 | 4.0 | 4.0 | 5.0 |
| **18** | 183536.0 | 3.507573 | 1.165970 | 1.0 | 3.0 | 4.0 | 4.0 | 5.0 |

**Figure 12**

On the top of the table above, the graph displayed in Figure 12 visually shows that the age of 56 rates movies slightly higher on average. Conversely, the age of 18 rated movies slightly lower on average compared to the other ages.

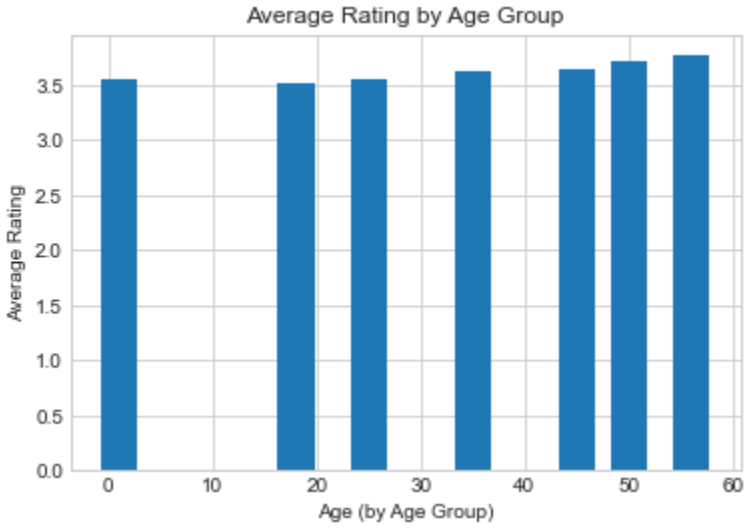


**Figure 13**

Third, the highest average rated movie genre was animation, comedy, and thriller movies, while the lowest average rated genre was the action, adventure, and children genre. The null hypothesis is that the average ratings are the same for each movie genre, while the alternative hypothesis is that the average ratings were different for each movie genre. This means that the animation, comedy, and thriller genre is more likely to receive a higher rating, while action, adventure, and children is less likely to receive a lower rating. As shown in Figure 13 below, the mean for the "Animation|Comedy|Thriller" is the highest mean out of all the genres. I acquired this data by using Pandas' describe function and then sorting the means from highest to lowest.

| Genres | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Animation\|Comedy\|Thriller** | 688.0 | 4.473837 | 0.739339 | 1.0 | 4.0 | 5.0 | 5.00 | 5.0 |
| **Sci-Fi\|War** | 1367.0 | 4.449890 | 0.805507 | 1.0 | 4.0 | 5.0 | 5.00 | 5.0 |
| **Animation** | 459.0 | 4.394336 | 0.819555 | 1.0 | 4.0 | 5.0 | 5.00 | 5.0 |
| **Film-Noir\|Mystery** | 1584.0 | 4.367424 | 0.776372 | 1.0 | 4.0 | 5.0 | 5.00 | 5.0 |
| **Adventure\|War** | 1644.0 | 4.346107 | 0.794099 | 1.0 | 4.0 | 5.0 | 5.00 | 5.0 |
| **Film-Noir\|Romance\|Thriller** | 445.0 | 4.294382 | 0.751029 | 1.0 | 4.0 | 4.0 | 5.00 | 5.0 |
| **Action\|Adventure\|Drama\|Sci-Fi\|War** | 2990.0 | 4.292977 | 0.844432 | 1.0 | 4.0 | 4.0 | 5.00 | 5.0 |
| **Film-Noir\|Sci-Fi** | 1800.0 | 4.273333 | 0.875051 | 1.0 | 4.0 | 4.0 | 5.00 | 5.0 |
| **Crime\|Film-Noir** | 867.0 | 4.264129 | 0.811660 | 1.0 | 4.0 | 4.0 | 5.00 | 5.0 |

**Figure 14**

In addition, Figure 14 below generally shows the average ratings for each movie. This is a chaotic-looking graph, but we wanted to see the general distribution of average ratings based on movie genres. This led us to focus on the next two graphs below to examine the average ratings of movie genres with the top ten highest rated movie genres and the top ten lowest rated movie genres.

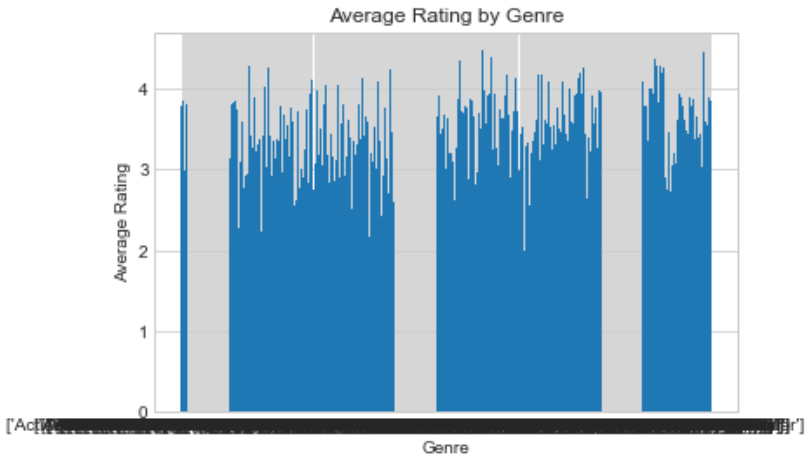


**Figure 15**

To supplement the conjecture above, the graph below depicted through Figure 15, the "Animation|Comedy|Thriller" genre has the highest bar of the bar plot with the highest average rating.

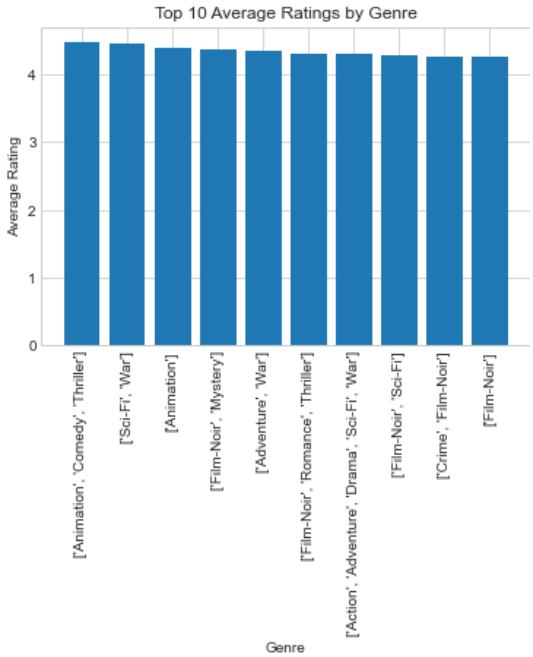


**Figure 16**

To also supplement the conjecture stated above, the “Action|Adventure|Children’s” genre has the lowest average rating for a genre seen in Figure 16 below.

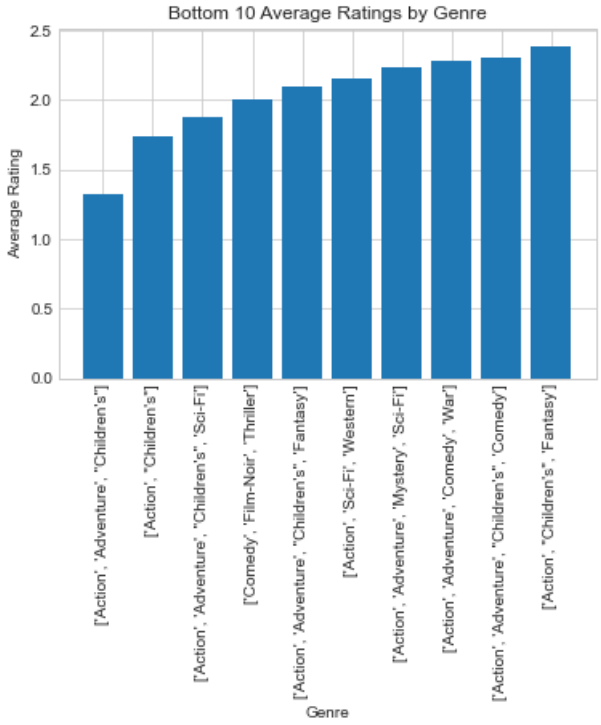


**Figure 17**

In Problem 3 we explored the relationship between mean rating of movies of men and women. At first we graphed a scatter plot using all of the movies and found that there were some differences of average rating for certain movies (as shown in the graph below).

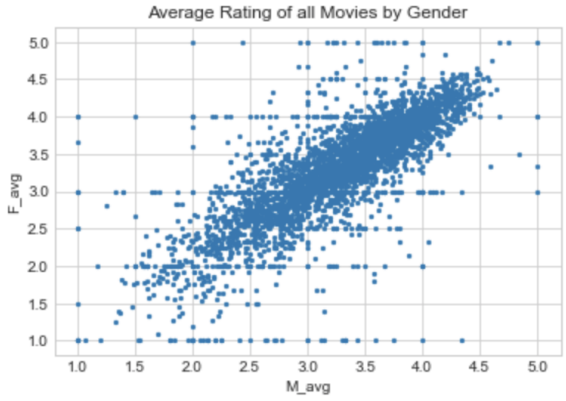


**Figure 18**

After further exploration of the data and analysis of the graph it is clear that many of the differences in average rating are generated from movies that don’t have enough ratings to be

statistically significant. A simple filter to remove movies with less than 200 ratings can be used to gain a clearer picture into these relationships.

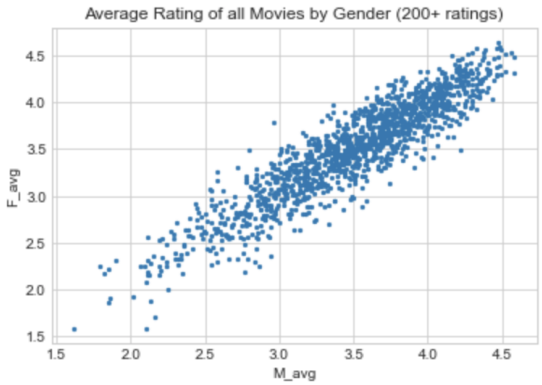


**Figure 19**

From the image above we can see that the function for the line of best fit is about y = x and the variation among data points is much less compared to the previous image.

## 3. Conjectures ( Anson )

Multiple conjectures were made, notably for Gender and Age. For Gender, all ratings are separated by Gender, male and female. Hypothesis testing was applied to explore if the mean of ratings by male is different from the mean of ratings by female. The null hypothesis was that the mean of rating by male is lower than mean of rating by female, which implies that males are harder to please since on average they give lower ratings. The alternate hypothesis was that the mean of rating by male is higher than mean of ratings by female, which implies that males are easier to please since on average they give higher ratings. The variance of both was examined and determined to be roughly the same, so a two-sample t-test with equal variance was used. After observing the result of the t-test, it was concluded that the p-value is very

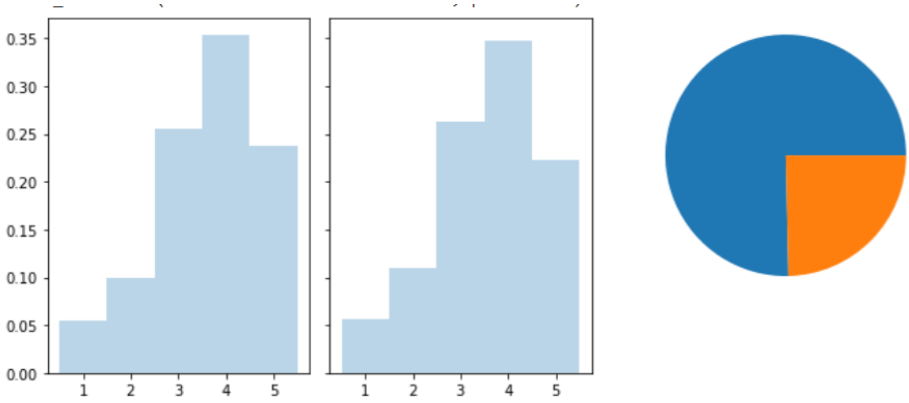


**Figure 20**

high, so the null hypothesis was not rejected, so this does not imply that male are easier to please. Afterwards, two histograms were made to see if the distribution of the ratings looks different. Turns out that the distribution of rating by male(left) is very similar to the distribution of rating by female(right). Then, a pie chart was made to observe the number of ratings made by male, and the number of ratings made by females. As demonstrated on the pie chart, the rating made by males(blue) is three times as many as ratings made by females(orange). Therefore, this conjecture is not supported by the data, and there is no sufficient evidence showing that males are easier to please. A similar conjecture was made about age, ratings are separated by Age, users aged below 50 and users aged above 50. The null hypothesis was that the mean of rating by users aged below 50 is lower than mean of rating by users aged below 50, which implies that users aged below 50 are harder to please since on average they give lower ratings. The alternate hypothesis was that the mean of rating by users aged below 50 is higher than mean of ratings by users aged above 50 , which implies that users aged below 50 are easier to please since on average they give higher ratings. The variance of rating by users aged below 50 and rating by users aged above 50 was examined and determined to be roughly the same, so applied a two-sample t-test with equal variance. After observing the result of the t-test, it was concluded that the p-value is high, which means the null hypothesis was not rejected, so this does not translate to users below the age of 50 being easier to please.  Two histograms were also made to see if the distribution of the ratings looks different for users age above 50 and below 50. From observing the histograms, the distribution of rating by users aged below 50(left) does not look very different to the distribution of rating by users aged above 50(right). Then, a pie chart was made to observe the number of ratings made by users aged

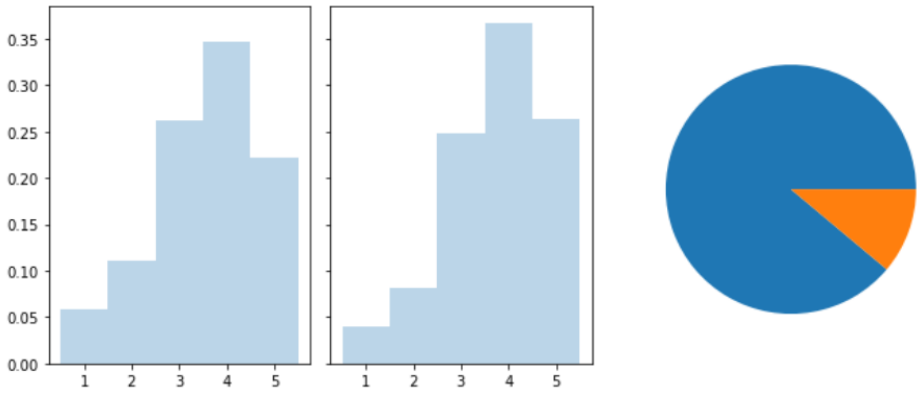


**Figure 21**

below 50, and the number of ratings made by users aged above 50. As demonstrated on the pie chart, the rating made by users aged below 50(blue) is eight times as many as ratings made by users aged above 50(orange). Therefore, this conjecture is not supported by the data, and there is

no sufficient evidence showing that Therefore, this conjecture is not supported by the data, and there is no sufficient evidence showing that males are easier to please.

## 4. Surprises in the Data (River)

Throughout the dataset of a million movies, there were a few surprises that came along the way. The first surprise that the group saw was that the “Children|Action|Adventure” genre was rated the lowest. We feel that a movie with that type of genre should be rated much higher based on past movies we have seen. However, this does make sense because the ages that gave the highest average ratings were much older in age. Therefore, they may not rate children's movies as high.

Similar to what we found that was surprising above, another surprise for us was the fact that animation and animation-related movies were rated so high. The genre with the highest average rating was the “Animation|Comedy|Thriller|” genre and the third highest genre were animation movies. This is surprising because we thought the animation movies were more tailored to younger kids rather than older adults.

## 5. Business Decision (Joshua)

The dataset contains a list of ratings from individual movie goers after watching a movie. The ratings were taken over the course of the year 2000. This dataset lists individuals that can be sorted by gender, age group, and rating they gave for each movie. By using these factors, data scientists can help inform managers and executives about decisions on what kind of movies to make and market. The dataset would be useful for movie marketing executives looking to market movies to a certain age group or gender. For example from the dataset we can presume that movies which have the genre of children's/action/adventure are unpopular because they have the lowest average rating. From this an executive can make a decision as to whether a movie in this category is worth making as it is rather unpopular towards most audiences. Inversely a executive might decide from this data to greenlight more movies that are animated/comedy/thrillers as they are extremely popular among audiences.

## 6. Member Contributions

Josh: The parts within Questions 1 through 3 were split up evenly among group members. I (Josh) worked on the first 3 sub questions in question 2. I had trouble with the second sub question. The numbers and labels were not matching up and after hours of trying to figure it out I talked it over with another group member and realized the problem was with the data frame I had created and was using to display the histogram (didn’t actually happen but, if we need page fluff add it). By using the original data frame and the same parameters I was able to get the histogram to display the data properly. Additionally, I helped other group members with their data visualizations. Later, I returned the favor and helped the same group member with his questions from part 1.

Anson: Members of the group decided to take on different parts of the project. I started the project by pre-processing the datasets from MovieLens before the first team meeting. Through research on the internet, I imported the dat file as tables, and assigned each attribute their corresponding name according to the text file README, then I exported those tables to csv files, which makes it easier to use later on. After the first meeting I started to tackle all the problems in Question 1, I was able to complete my problems with some support from a teammate. I also provided a teammate some guidance, because I remembered the graph shown in class is similar to what we are asked to make in part 3. I also worked on explaining the conjectures in part 4, and the meaning behind the results of those hypothesis testing and plots.

River: I was responsible for the last two parts of Problem 2. For part D of Problem 2, I wrote the code for the graph that was creating a histogram of the average rating for movies which are rated more than 100 times. In addition, I developed the conjectures for the last part of Problem 2 and performed hypothesis testing. I coded the statistics summary for the three conjectures I created for Age, Gender, and Movie Genres. I also aided with my teammates on brainstorming ideas for the report. I wrote the section on the findings of the data where I summarize my findings of the data and everyone else's findings as well. Also, I wrote the section for what the surprises we found interesting were.

Ardavasd: I was responsible for all of Problem 3. I wrote the code for Problem 3 in the Jupyter Notebook and included the additional information in the slideshow. Within Problem 3, I used pandasql which allowed me to recreate and restructure some of our tables for ease of use. For The conjecture portion in problem 3 I created a scatterplot of the distribution of correlation coefficients for average movies ratings between Men and Women among different age groups. I also contributed to creating and presenting the presentation.