Case Study 2 - Analyzing data from MovieLens

DS501 - Introduction to Data Science

Worcester Polytechnic Institute

1. Report: please prepare a report based on what you found in the data.

- What data you collected?
  - This report collects a variety of data about movie-goers over a period of many years, such as what gender they are, what movies they see, and how they rate those movies. A variety of insightful trends using this data was gleaned in the charts below.
- Why this topic is interesting or important to you? (Motivations)
  - Almost everybody enjoys movies as a form of entertainment in some form or another, and I enjoy learning about how people perceive different movies. The charts I produced from the data below gave me valuable insights into how people think so that when I meet new people I can be more informed about their preferences.
- How did you analyze the data?
  - The original data was read into a CSV document, and then the R ggplot 2 framework was used to analyze various segments of the data and display the resulting information in various charts and graphs to summarize the conclusions made.
- What did you find in the data? (please include figures or tables in the report)
  - The charts and insights described in this report show important trends in how individuals who watch movies tend to perceive those movies, and what types of movies they prefer. Some of the insights were more counterintuitive than others, but all of the insights are supported by the data.

2. R Code with RMarkdown

How to submit: Upload on Course Webpage and Send an email to ndingari@wpi.edu with the subject: "DS501 Case study 2".

Introduction

Desired outcome of the case study. In this case study we will look at the movies data set from MovieLens. It contains data about users and how they rate movies. The idea is to analyze the data set, make conjectures, support or refute those conjectures with data, and tell a story about the data!

Problem 1: Importing the MovieLens data set and merging it into a single data frame

Report some basic details of the data you collected. For example:

- How many movies have an average rating over 4.5 overall?

Table 1: Movies over 4.5 stars on average (11 total)

Title	Mean Rating
Aiqing wansui (1994)	5.000
Entertaining Angels: The Dorothy Day Story (1996)	5.000
Great Day in Harlem, A (1994)	5.000
Marlene Dietrich: Shadow and Light (1996)	5.000
Prefontaine (1997)	5.000
Saint of Fort Washington, The (1993)	5.000
Santa with Muscles (1996)	5.000
Someone Else's America (1995)	5.000
Star Kid (1997)	5.000
They Made Me a Criminal (1939)	5.000
Pather Panchali (1955)	4.625

- How many movies have an average rating over 4.5 among men? How about women?

Table 2: Movies over 4.5 stars on average for males (18 total)

Title	Mean Rating
Aiqing wansui (1994)	5.000000
Delta of Venus (1994)	5.000000
Entertaining Angels: The Dorothy Day Story (1996)	5.000000
Great Day in Harlem, A (1994)	5.000000
Leading Man, The (1996)	5.000000
Letter From Death Row, A (1998)	5.000000
Little City (1998)	5.000000
Love Serenade (1996)	5.000000
Marlene Dietrich: Shadow and Light (1996)	5.000000
Prefontaine (1997)	5.000000
Quiet Room, The (1996)	5.000000
Saint of Fort Washington, The (1993)	5.000000
Santa with Muscles (1996)	5.000000
Star Kid (1997)	5.000000
They Made Me a Criminal (1939)	5.000000
Little Princess, The (1939)	4.666667
Two or Three Things I Know About Her (1966)	4.666667
Pather Panchali (1955)	4.625000

Table 3: Movies over 4.5 stars on average for females (16 total)

Title	Mean Rating
Everest (1998)	5.000000
Faster Pussycat! Kill! Kill! (1965)	5.000000
Foreign Correspondent (1940)	5.000000
Maya Lin: A Strong Clear Vision (1994)	5.000000
Mina Tannenbaum (1994)	5.000000
Prefontaine (1997)	5.000000
Someone Else's America (1995)	5.000000
Stripes (1981)	5.000000

Title	Mean Rating
Telling Lies in America (1997)	5.000000
Visitors, The (Visiteurs, Les) (1993)	5.000000
Year of the Horse (1997)	5.000000
Schindler's List (1993)	4.632911
Close Shave, A (1995)	4.631579
Shawshank Redemption, The (1994)	4.562500
Wallace & Gromit: The Best of Aardman Animation (1996)	4.533333
Shall We Dance? (1996)	4.529412

- How many movies have an median rating over 4.5 among men over age 30? How about women over age 30?

Table 4: Movies over 4.5 stars on average for males over 30 (27 total)

Title	Mean Rating
Aiqing wansui (1994)	5.000000
Anna (1996)	5.000000
Aparajito (1956)	5.000000
Delta of Venus (1994)	5.000000
Entertaining Angels: The Dorothy Day Story (1996)	5.000000
Faithful (1996)	5.000000
Faust (1994)	5.000000
For the Moment (1994)	5.000000
Great Day in Harlem, A (1994)	5.000000
Leading Man, The (1996)	5.000000
Little City (1998)	5.000000
Love Serenade (1996)	5.000000
Marlene Dietrich: Shadow and Light (1996)	5.000000
Mondo (1996)	5.000000
Prefontaine (1997)	5.000000
Rendezvous in Paris (Rendez-vous de Paris, Les) (1995)	5.000000
Romper Stomper (1992)	5.000000
Santa with Muscles (1996)	5.000000
Sliding Doors (1998)	5.000000
Star Kid (1997)	5.000000
The Deadly Cure (1996)	5.000000
They Made Me a Criminal (1939)	5.000000
Tough and Deadly (1995)	5.000000
Two or Three Things I Know About Her (1966)	5.000000
World of Apu, The (Apur Sansar) (1959)	5.000000
Whole Wide World, The (1996)	4.666667
Pather Panchali (1955)	4.571429

Table 5: Movies over 4.5 stars on average for females over 30 (32 total)

Title	Mean Rating
Angel Baby (1995)	5.000000

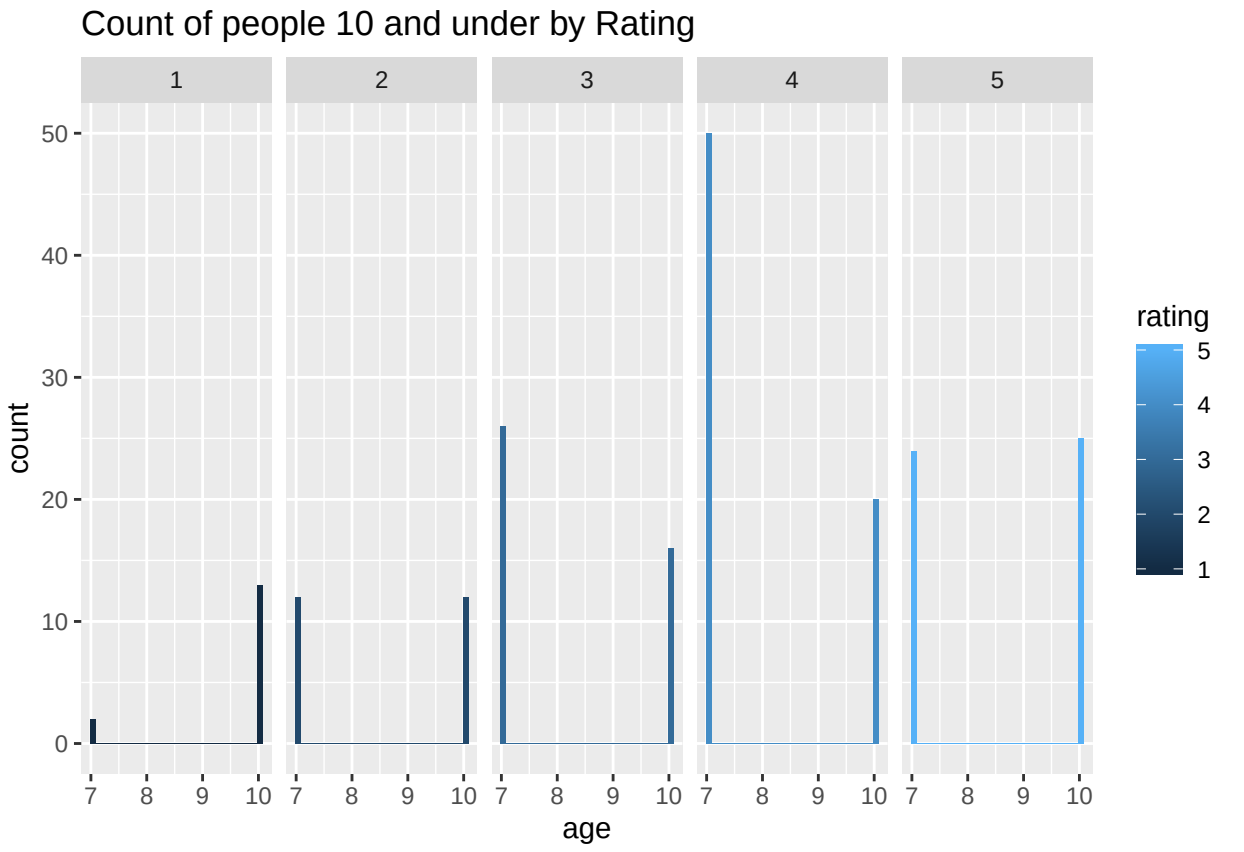
Title	Mean Rating
Bent (1997)	5.000000
Best Men (1997)	5.000000
Chairman of the Board (1998)	5.000000
Evil Dead II (1987)	5.000000
Farinelli: il castrato (1994)	5.000000
Fear of a Black Hat (1993)	5.000000
Foreign Correspondent (1940)	5.000000
Great Expectations (1998)	5.000000
Higher Learning (1995)	5.000000
In the Bleak Midwinter (1995)	5.000000
Last Dance (1996)	5.000000
Letter From Death Row, A (1998)	5.000000
Ma vie en rose (My Life in Pink) (1997)	5.000000
Mina Tannenbaum (1994)	5.000000
Night Flier (1997)	5.000000
Safe (1995)	5.000000
Stripes (1981)	5.000000
SubUrbia (1997)	5.000000
Swept from the Sea (1997)	5.000000
Turbulence (1997)	5.000000
Visitors, The (Visiteurs, Les) (1993)	5.000000
Women, The (1939)	5.000000
Wrong Trousers, The (1993)	5.000000
Grand Day Out, A (1992)	4.800000
Once Were Warriors (1994)	4.800000
Heavy Metal (1981)	4.666667
Schindler's List (1993)	4.666667
It's a Wonderful Life (1946)	4.576923
Close Shave, A (1995)	4.555556
Mr. Smith Goes to Washington (1939)	4.538462
Christmas Carol, A (1938)	4.533333

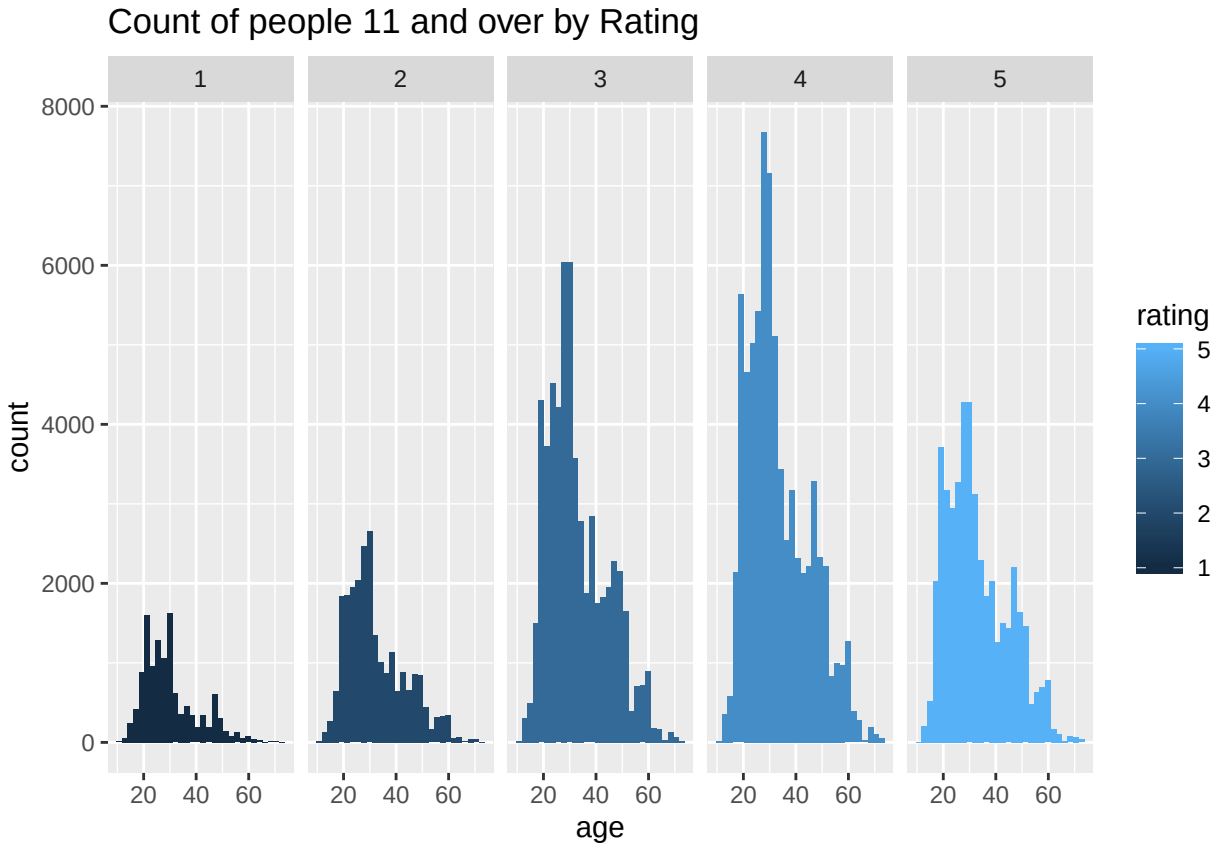
- What are the ten most popular movies?
  - In this report, the most popular movies are defined as the ones that have been seen/rated the greatest number of times

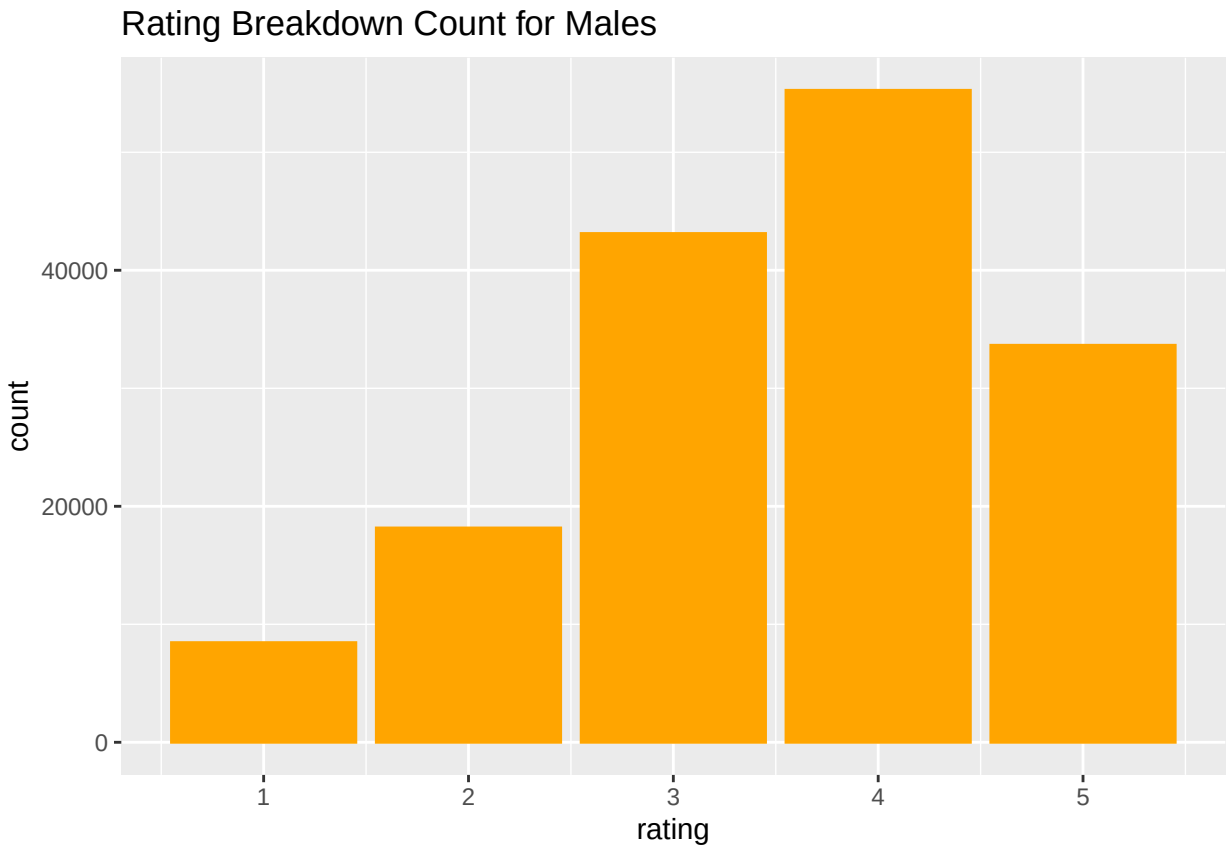
Table 6: Top 10 Popular Movies

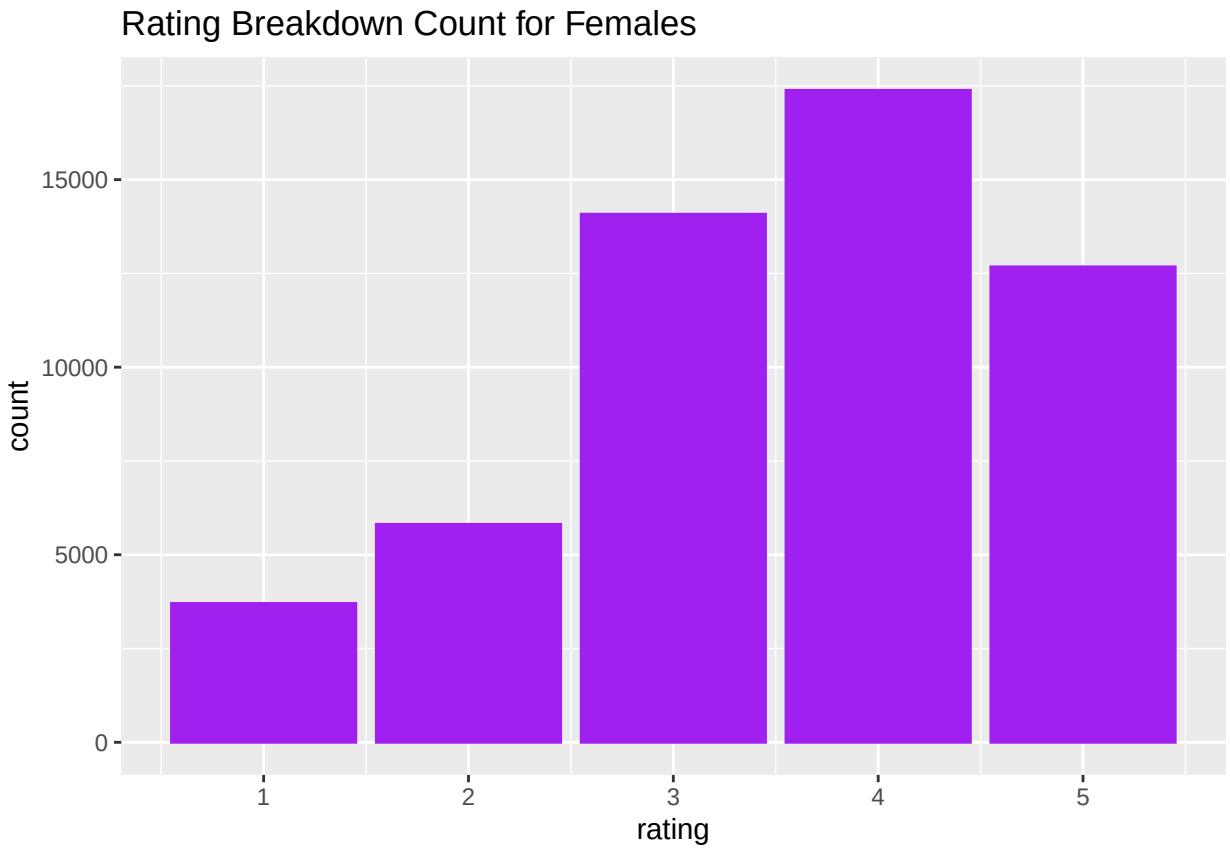
Movie	View/Rating Count
Star Wars (1977)	2915
Return of the Jedi (1983)	2535
Empire Strikes Back, The (1980)	2202
Fargo (1996)	1524
English Patient, The (1996)	1443
Toy Story (1995)	1356
Princess Bride, The (1987)	1296
Independence Day (ID4) (1996)	1287
Godfather, The (1972)	1239
Men in Black (1997)	1212

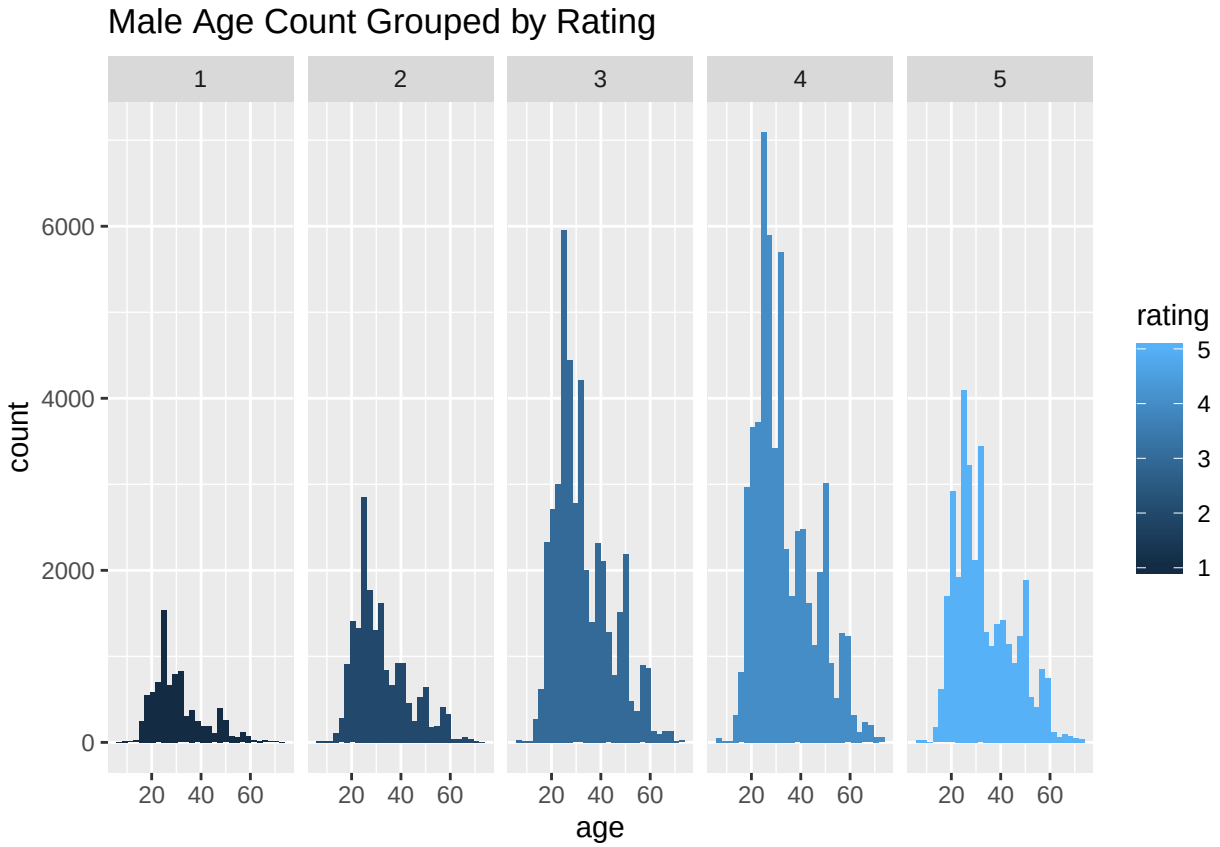
- Make some conjectures about how easy various groups are to please? Support your answers with data!
  - For example, one might conjecture that people between the ages of 1 and 10 are the easiest to please since they are all young children. This conjecture may or may not be true, but how would you support or disprove either conclusion with with data?
    * If people between the ages of 1-10 were easiest to please, then a significantly higher percentage of them should be giving high ratings to movies than people of other ages do.
    * Looking at the "Count of people 10 and under by Rating" chart and the "Count of people 11 and over by Rating" charts below, both charts seem to have a similar distribution of most people giving a lot of 4 star ratings followed by slightly fewer giving 3 and 5 star ratings. This trend is not significantly different with the 10 and under group than in the 11 and over group, therefore 1-10 year olds do not appear to be significantly easier to please.
  - Be sure to come up with your own conjectures and support them with data!
    * Another conjecture the data supports is that males and females have roughly similar patterns in how they distribute their movie ratings. For both males and females, based on the "Rating Breakdown Count for Males" and "Rating Breakdown Count for Females" charts below, both charts show that people of each gender tend to give more 4 star ratings than any other ratings. Also both genders tend to give slightly fewer 3 and 5 ratings, and even fewer 1 and 2 star ratings. This trend is the same with both genders.
    * A fascinating trend the data shows is that a large percentage of females who rate movies 1 star are 20 years of age or younger. For males who rate movies 1 star, the age distribution is much more even, although slightly more are also younger. The phenomenon of younger people rating movies 1 star is far more prevalent among females than males. See the charts below with "Age Count Grouped by Rating" in the title.

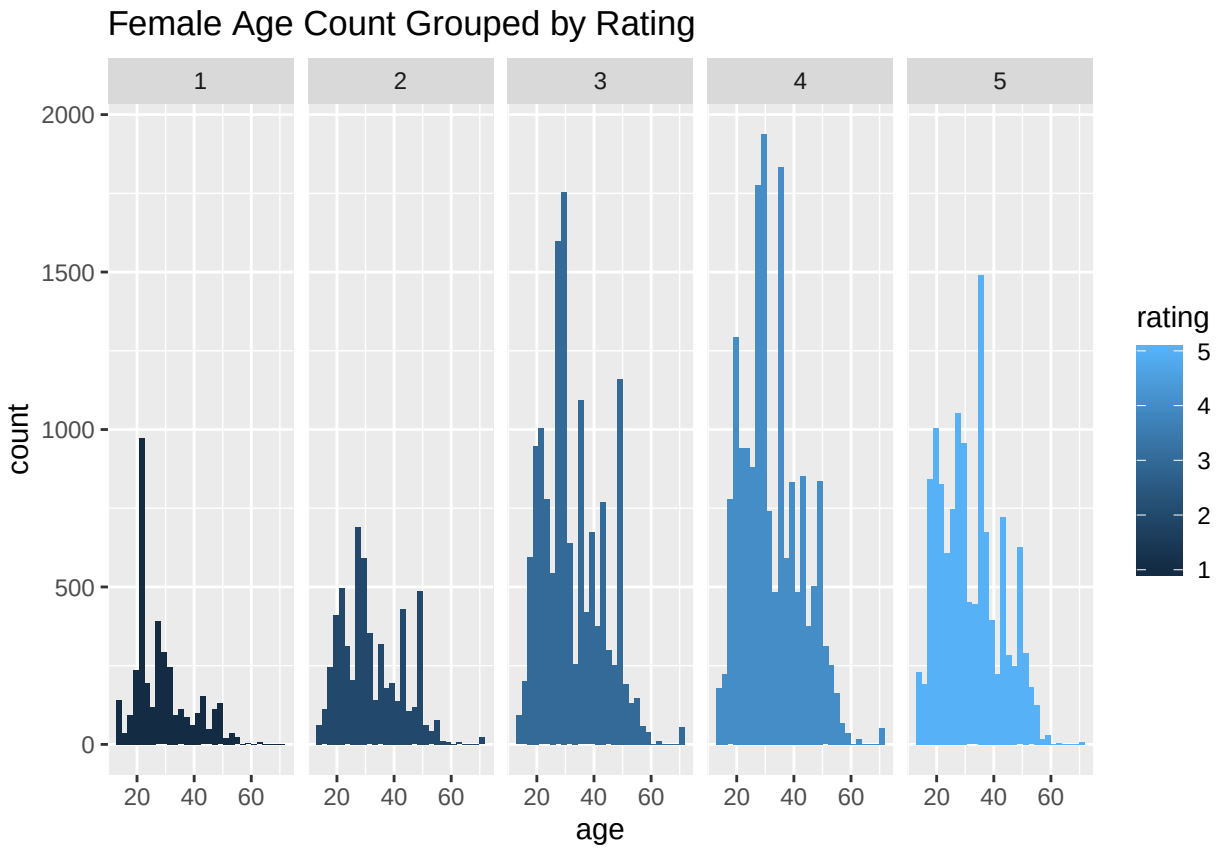



Count of people 11 and over by Rating
1
2
3
4
5
8000
6000
4000
2000
0
count
20
40
60
age
rating
5
4
3
2
1


Rating Breakdown Count for Males
count
40000
20000
0
1
2
3
4
5
rating


Rating Breakdown Count for Females
15000
10000
5000
0
count
1
2
3
4
5
rating

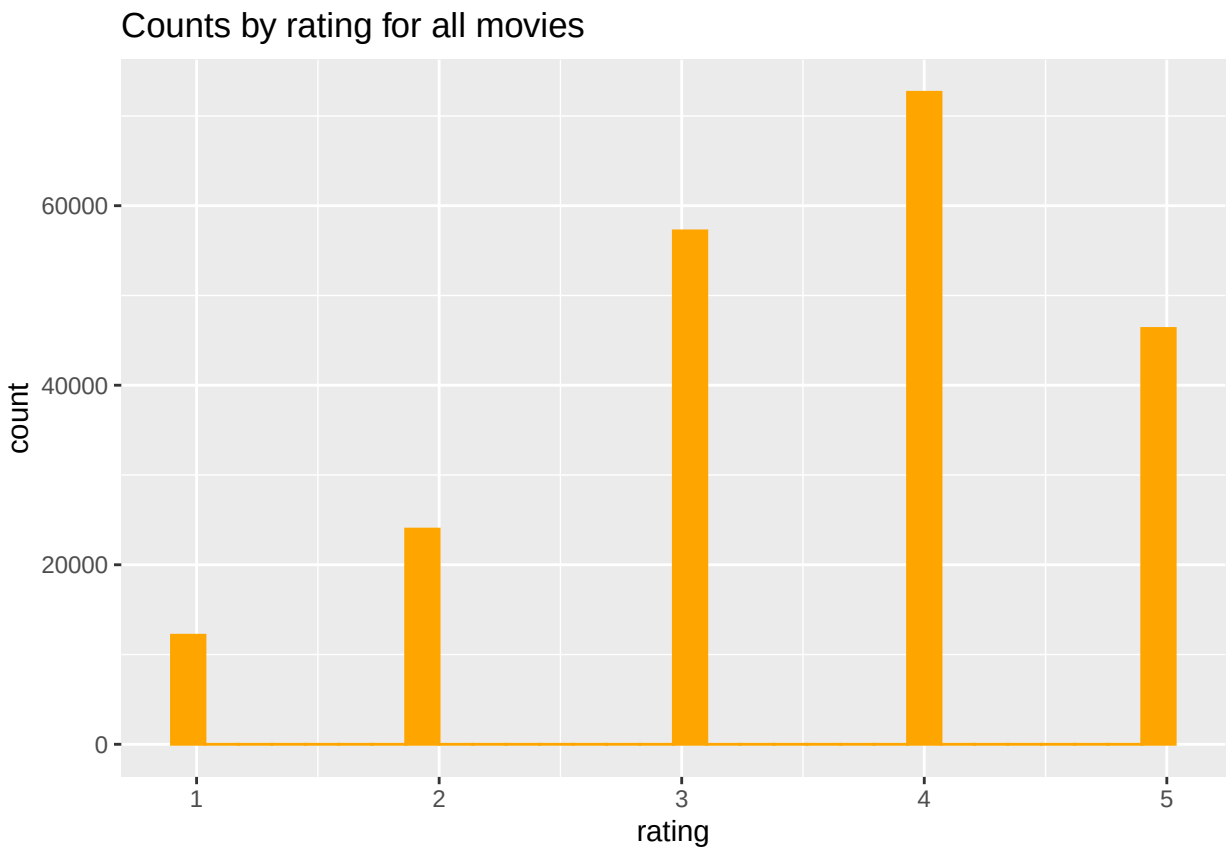

Male Age Count Grouped by Rating
1
2
3
4
5
count
6000
4000
2000
0
20
40
60
age
rating
5
4
3
2
1



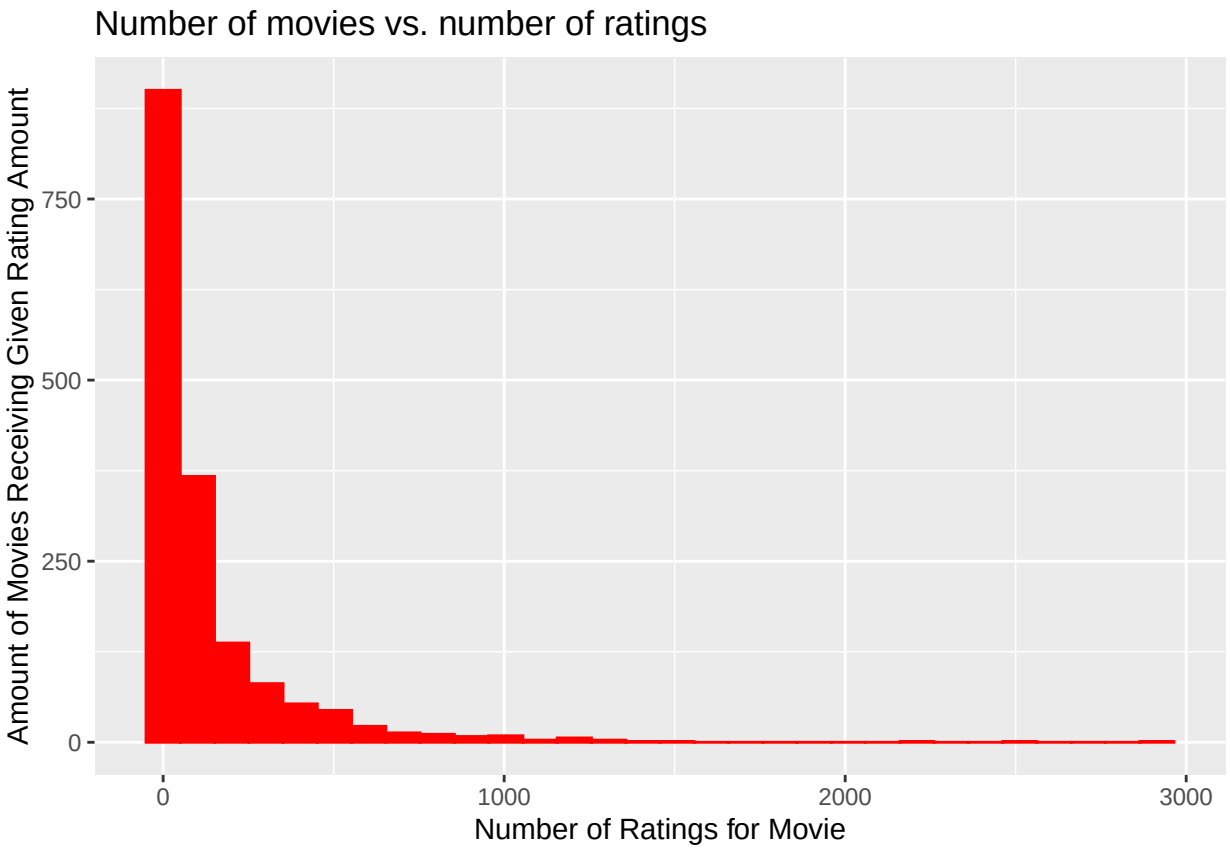
Problem 2: Expand our investigation to histograms

An obvious issue with any inferences drawn from Problem 1 is that we did not consider how many times a movie was rated.

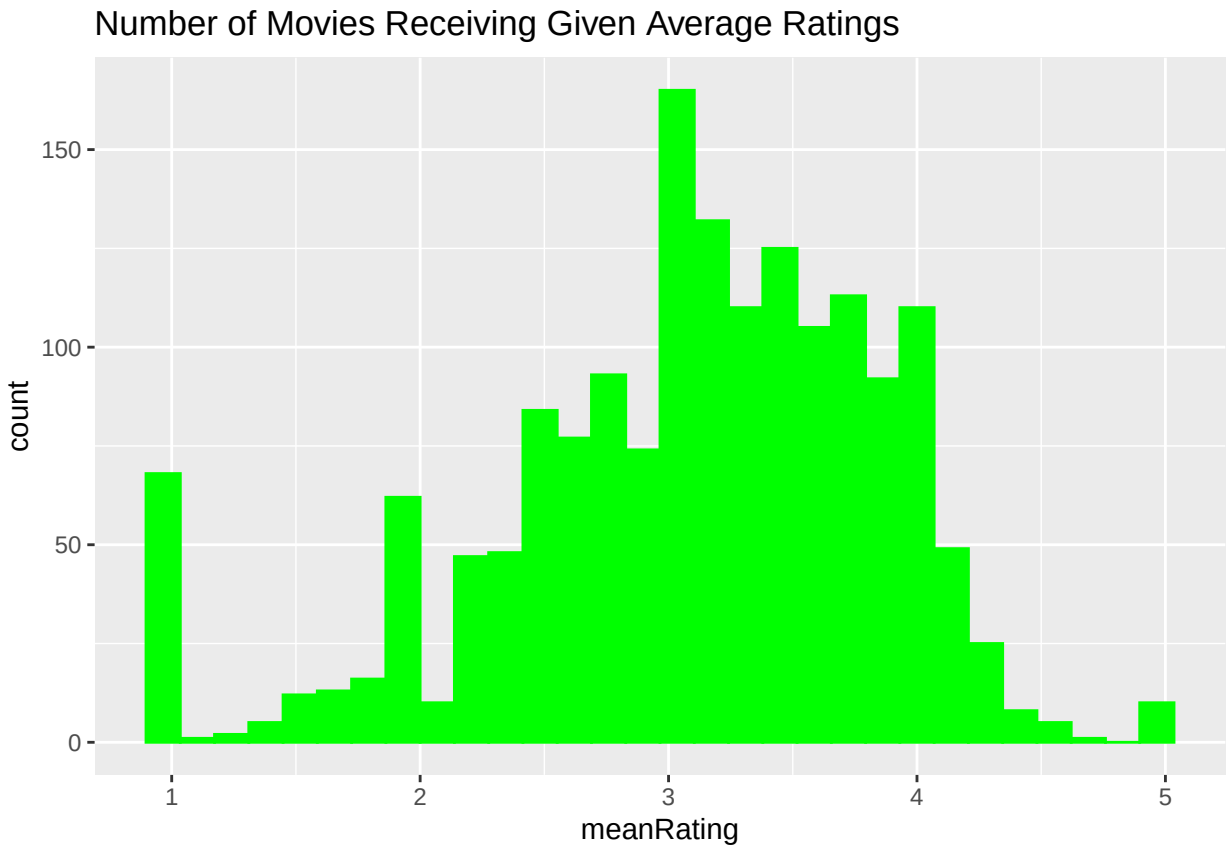
- Plot a histogram of the ratings of all movies.



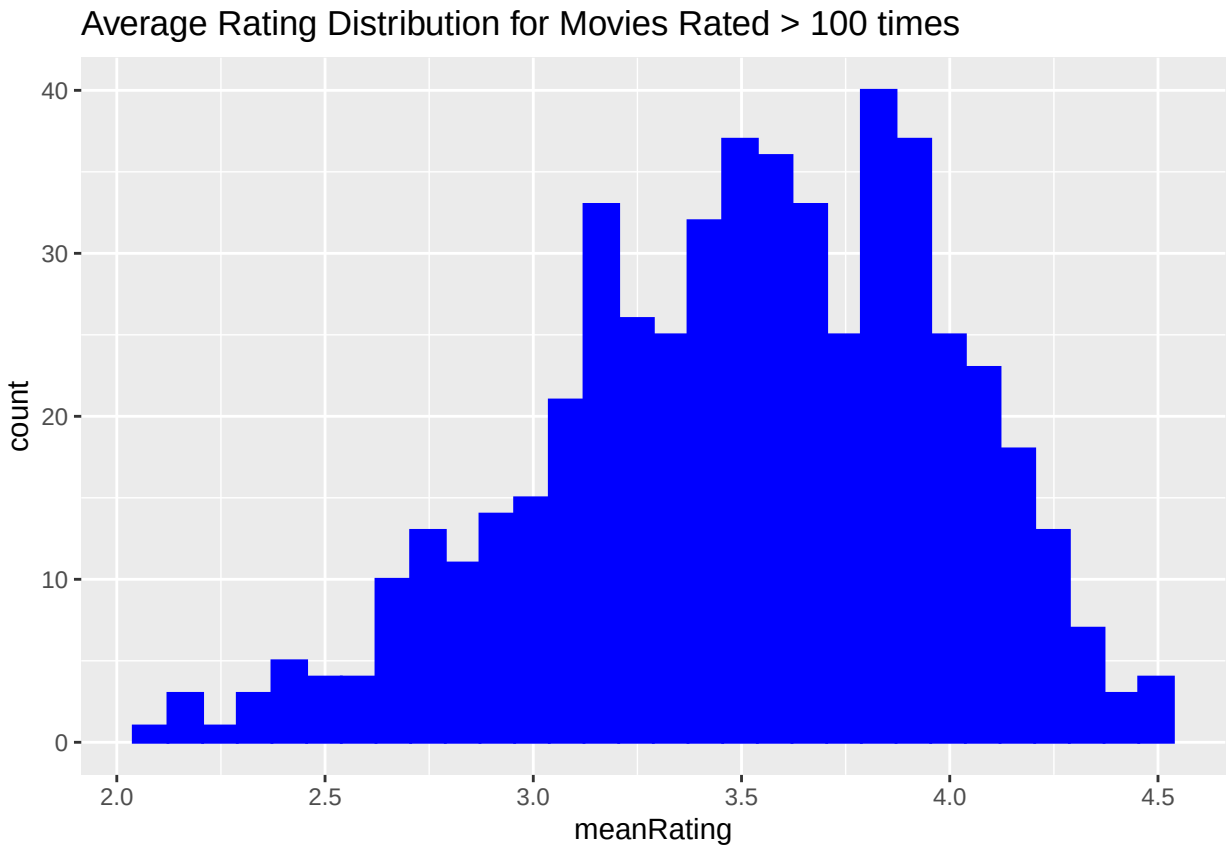
- Plot a histogram of the number of ratings each movie received.



- Plot a histogram of the average rating for each movie.

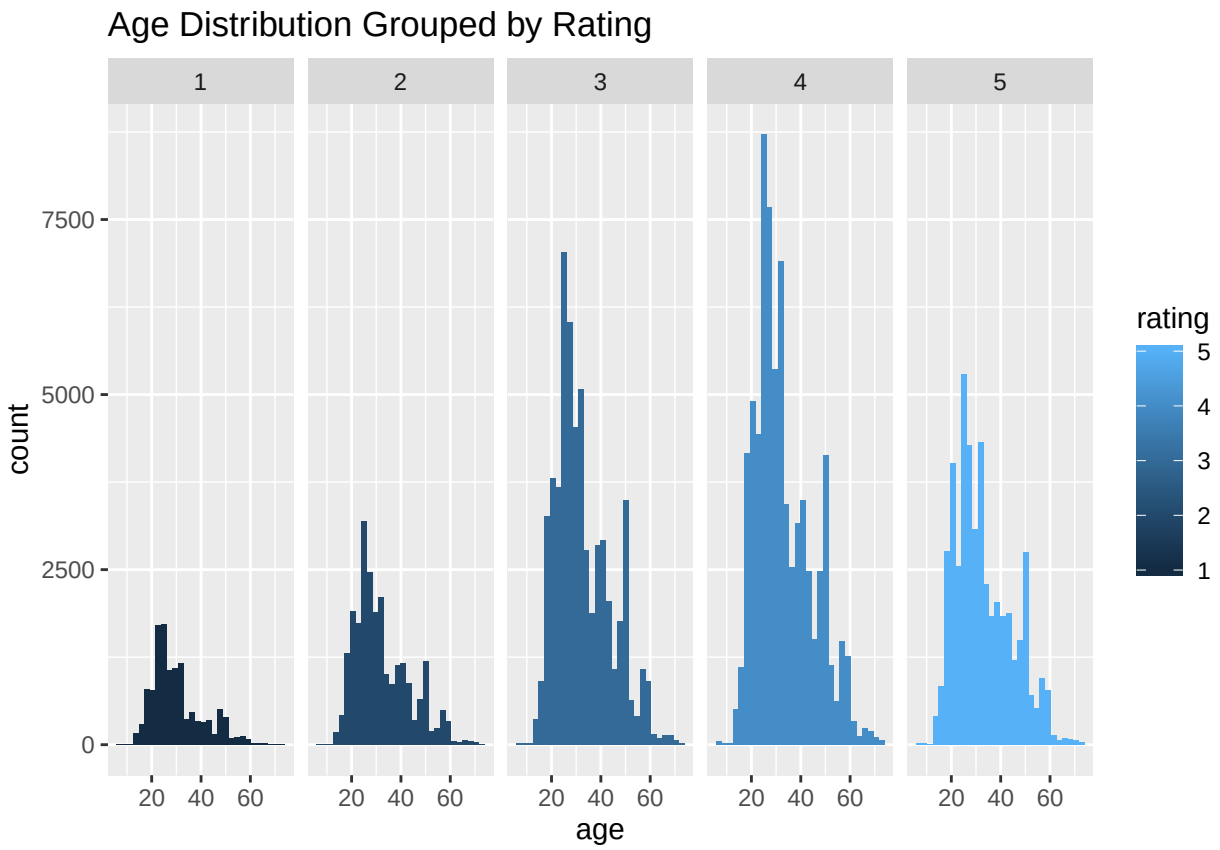


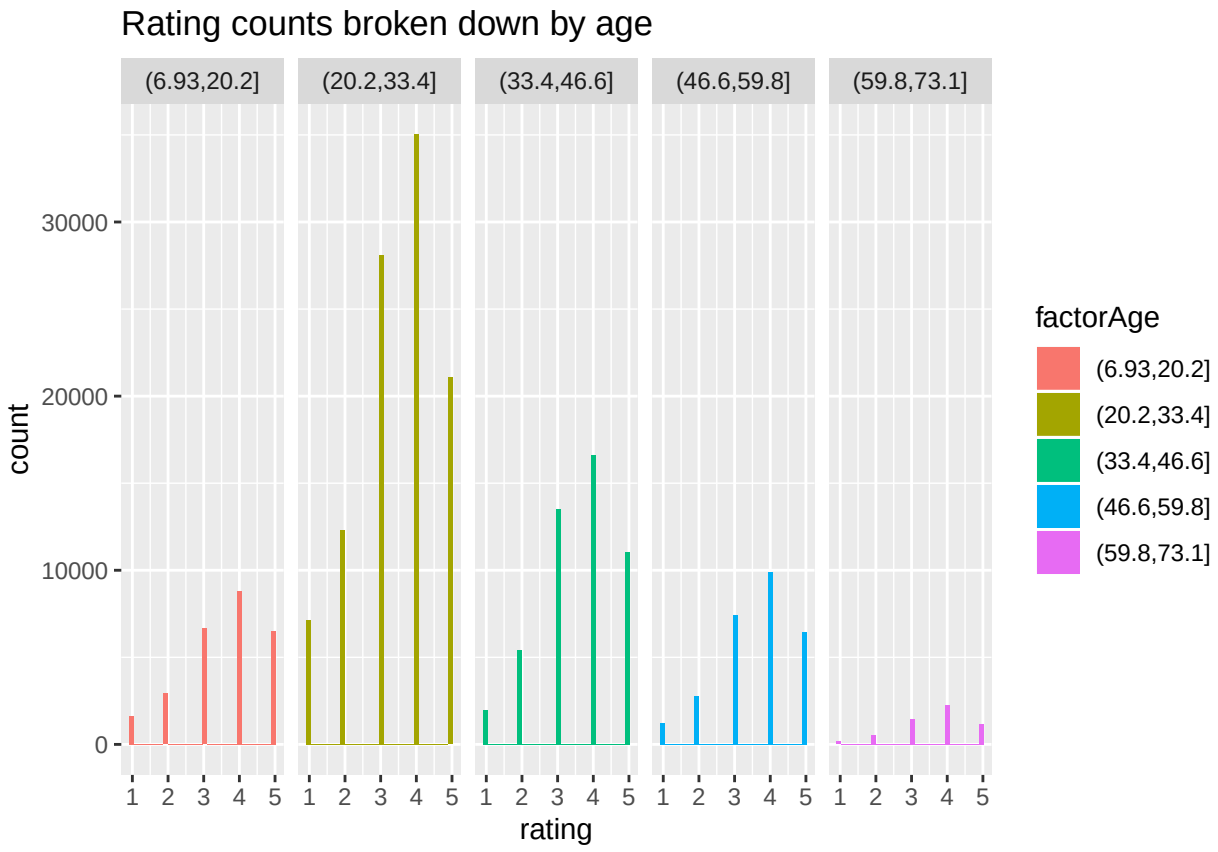
- Plot a histogram of the average rating for movies which are rated more than 100 times.



- Results
  - What do you observe about the tails of the histogram where you use all the movies versus the one where you only use movies rated more than 100 times?
    - For the histogram using all the movies, there are a significantly higher percentage of movies where the mean rating is low than for just the movies with over 100 ratings. For the histogram with just the movies over 100 ratings, the outliers are less pronounced and the distribution is smoother.
  - Which highly rated movies would you trust are actually good? Those rated more than 100 times or those rated less than 100 times?
    - The highly rated movies that are actually good are more likely the ones with over 100 ratings because more people have had the opportunity to rate the movies, and the outliers of fewer data points skew the data less.
- Make some conjectures about the distribution of ratings? Support your answers with data!
  - For example, what age range do you think has more extreme ratings? Do you think children are more or less likely to rate a movie 1 or 5?
  - Be sure to come up with your own conjectures and support them with data!
    - The "Rating Counts by Age" chart below groups rating counts into 5 equal age ranges. For example it has data about the rating count of movie goers between 7 and 20 years old, then it has more information about moviegoers between the ages of 20 and 33 etc. Based on this chart, every age range seems to have roughly the same distribution of ratings - with most individuals giving 4 stars and slightly fewer giving 3 or 5 stars. There appears to be no particular age range that is significantly more likely to give movies extreme ratings.
    - The "Age Distribution Grouped by Rating" chart shows within each rating class, how many people of each age group gave movies that rating. Like the chart mentioned above, each

rating grouping here has roughly the same distribution of the number people of each age giving the movies that rating. Each group has a high number of 20-40 year olds followed by lower numbers of people of younger and older ages. In general, the data indicates that a high number of 20-40 year olds rated movies and a lower number of people of different ages rated movies. No particular rating class appears to have a significant difference in which age group gives that rating more often.

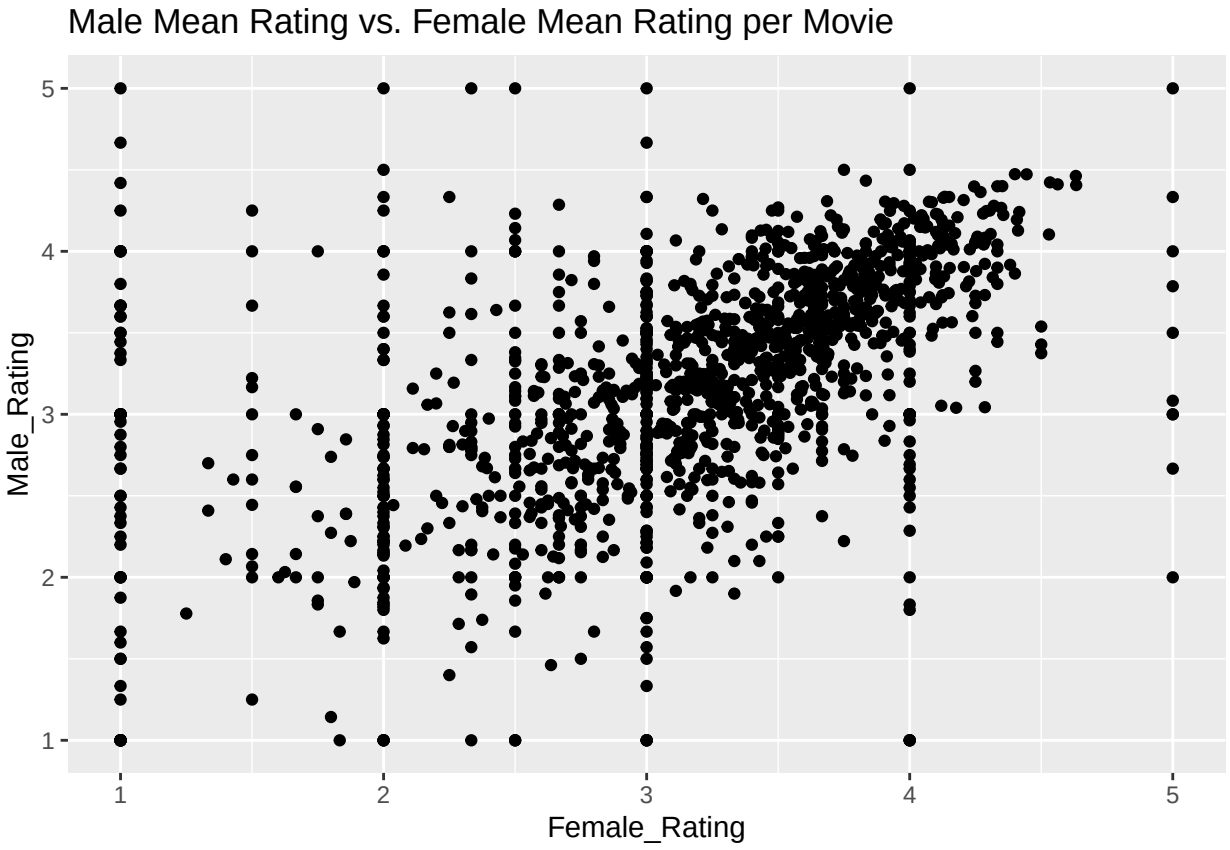




Problem 3: Correlation: Men versus women

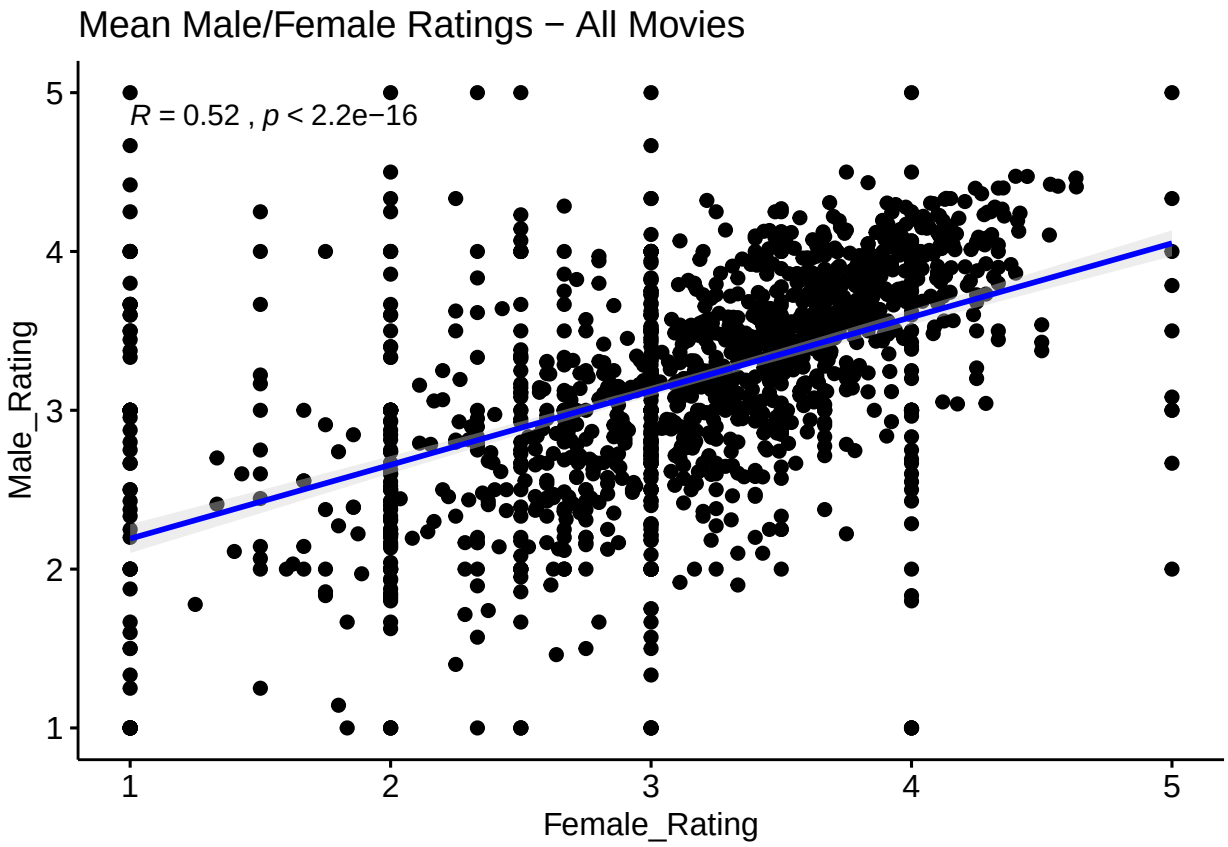
Let us look more closely at the relationship between the pieces of data we have.

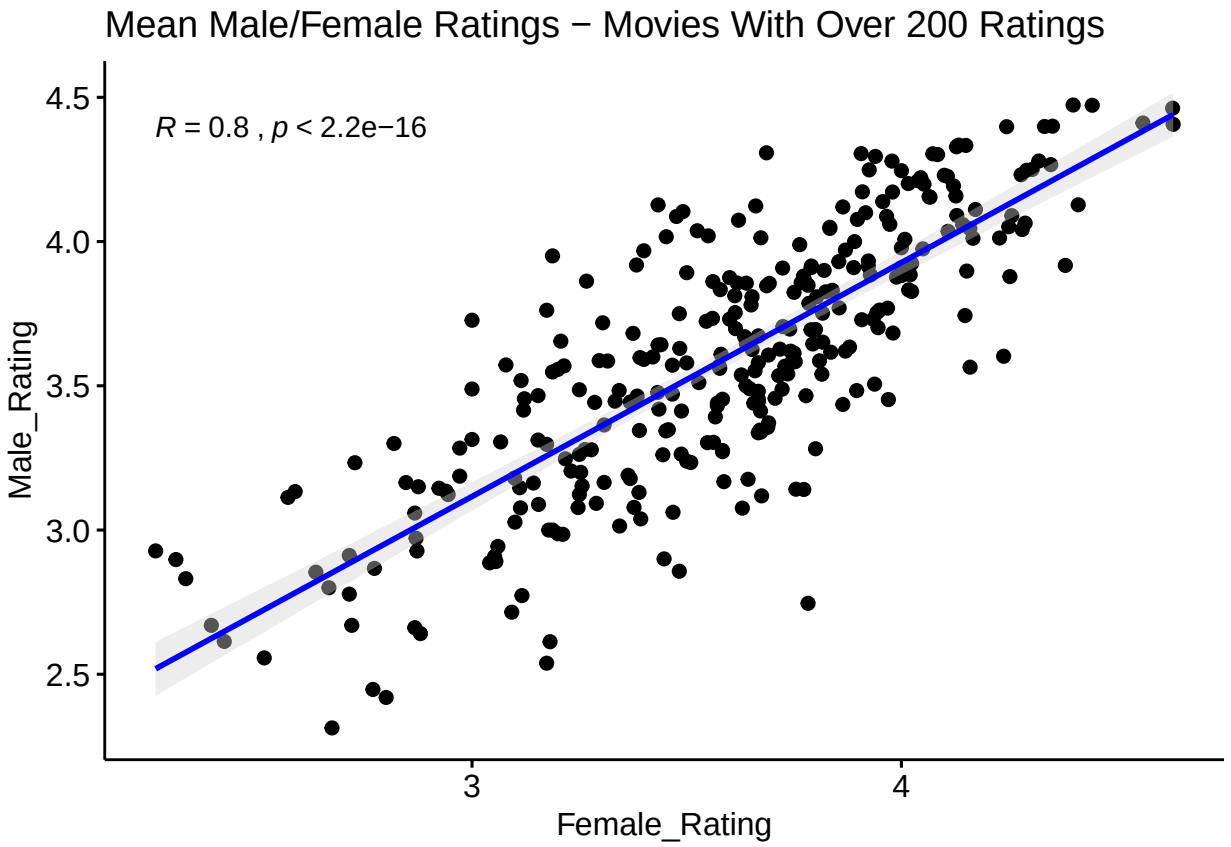
- Make a scatter plot of men versus women and their mean rating for every movie.



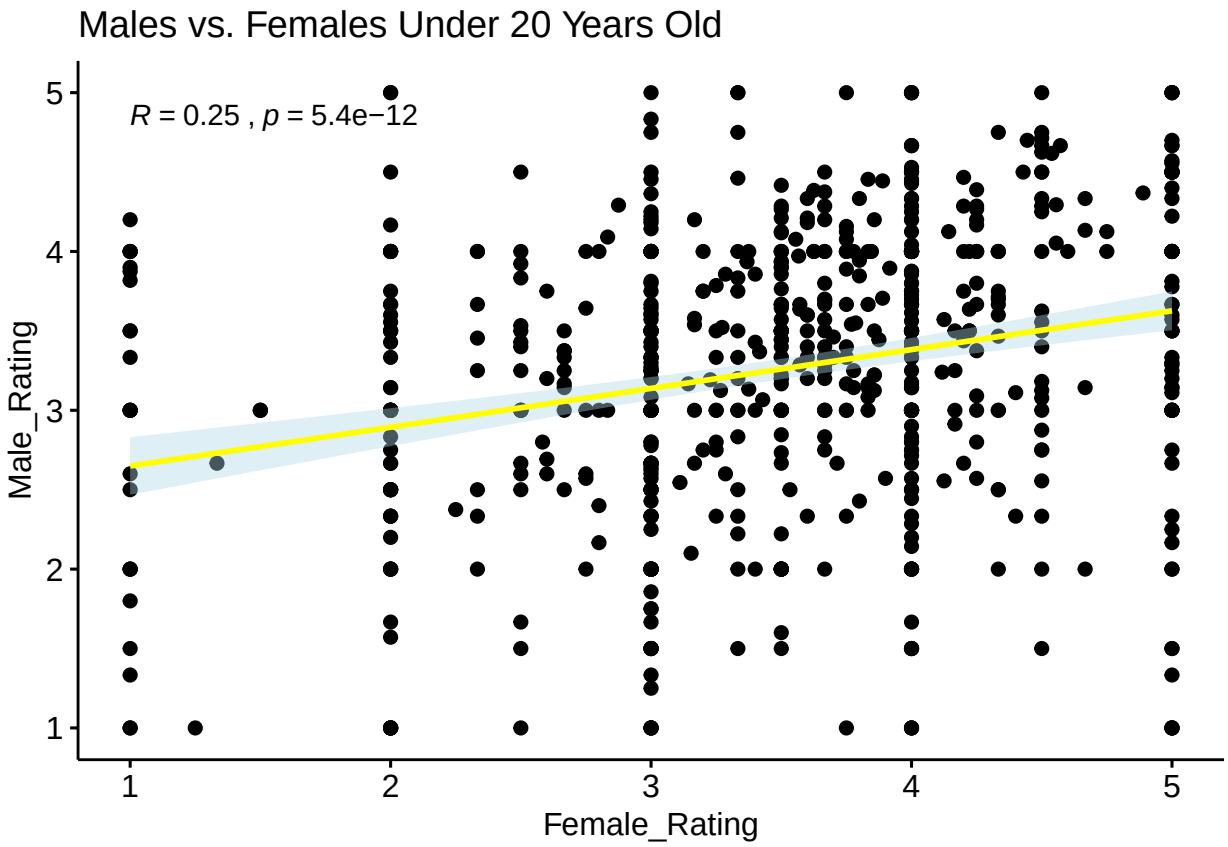
- Make a scatter plot of men versus women and their mean rating for movies rated more than 200 times.

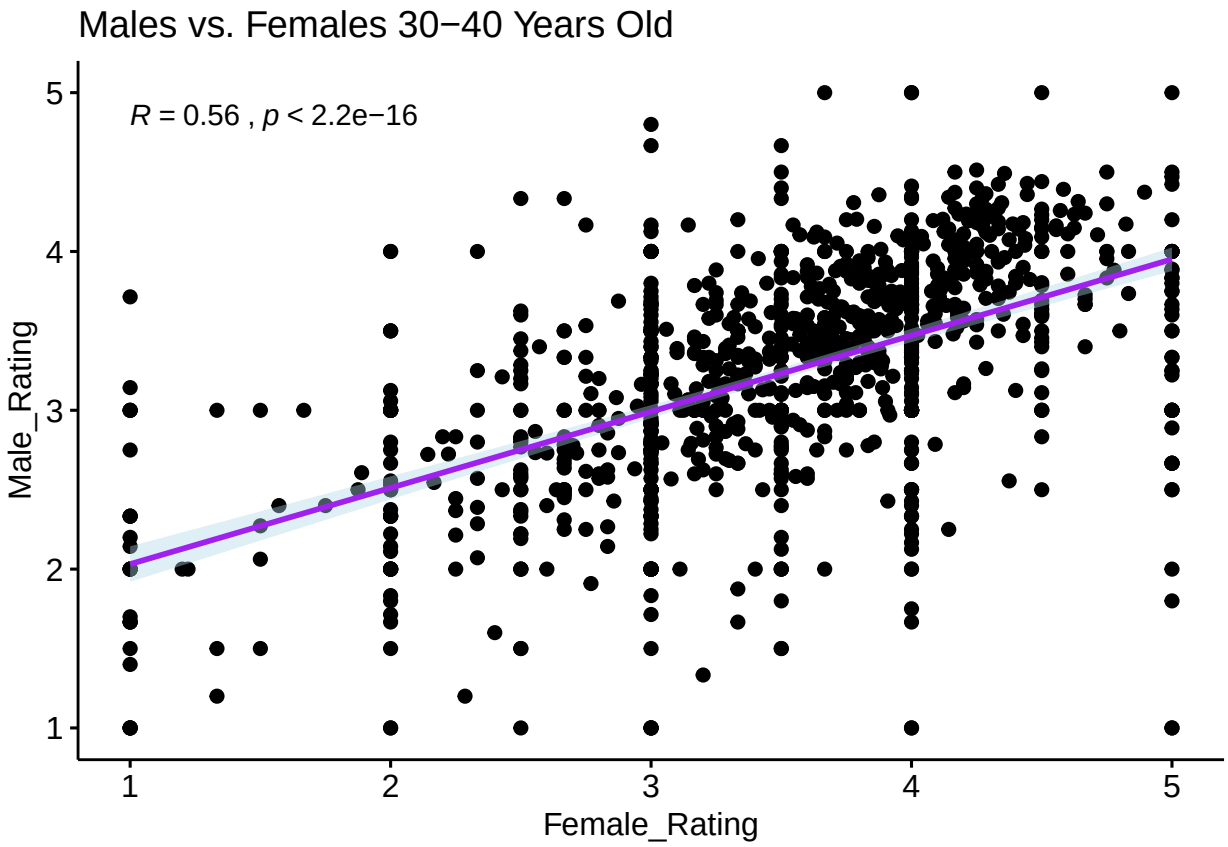
- Compute the correlation coefficent between the ratings of men and women.

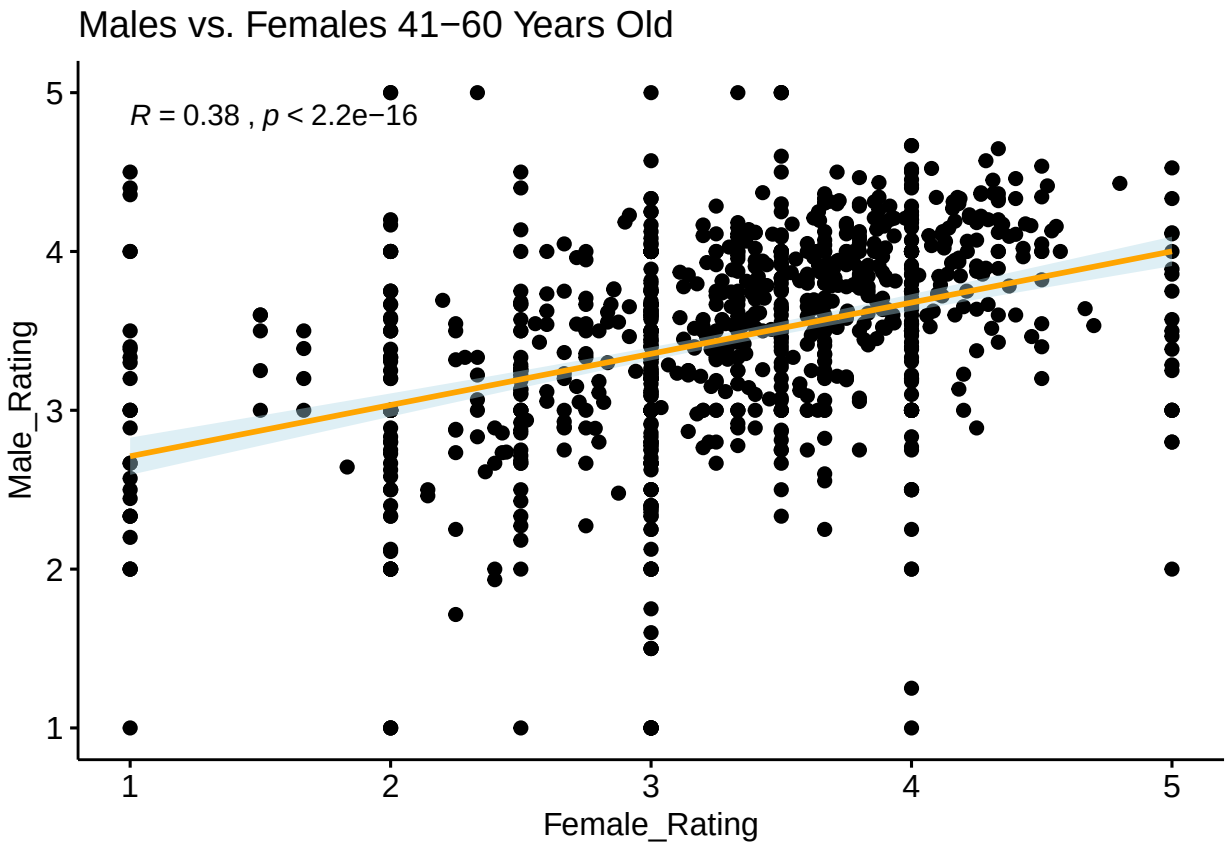

Mean Male/Female Ratings − All Movies
R = 0.52 , p < 2.2e−16
Male_Rating
Female_Rating
1
2
3
4
5



- Results
  - What do you observe?
    - Based on the above plots, the correlation coefficient for the mean female/male ratings for all the movies was about 0.52.
    - The correlation coefficient for the female/male ratings of the movies receiving over 200 ratings is around 0.8.
  - Are the ratings similiar or not? Support your answer with data!
    - As described above, the ratings for the movies that received over 200 ratings were significantly more positively correlated than the plot for all the movies.
- Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.
  - For example, are men and women more similar when they are younger or older?
  - Be sure to come up with your own conjectures and support them with data! -Based on the below plots, the correlation between males and females when they are younger than 20 years old is very low - approximately 0.25.
    - Between the ages of 30 and 40, the genders have a much higher correlation of around 0.56, and then between the ages of 41 and 60 years old the correlation drops again slightly to 0.38.
    - This data supports the earlier conjecture that more females younger than 20 years of age tend to give movies extremely low ratings than males younger than 20 years of age.


Males vs. Females Under 20 Years Old
R = 0.25 , p = 5.4e−12
Male_Rating
Female_Rating
1
2
3
4
5


Males vs. Females 30−40 Years Old
R = 0.56 , p < 2.2e−16
Male_Rating
Female_Rating
1
2
3
4
5



Problem 4: Open Ended Question: Business Intelligence

- Do any of your conjectures in Problems 1, 2, and 3 provide insights that a movie company might be interested in?
  - For problem 1, a fascinating insight is that generally speaking, people of different age groups and genders have similar patterns in how they rate movies. Both genders individually as well many of the distinct age groups all tended to most commonly give 4 star ratings followed by 3 and 5 star ratings. Then the remaining ratings were least common.
  - Also found in problem 1, younger females disproportionately tend to give low ratings compared to other demographics. This suggests that movie companies might want to put slightly more effort in catering toward younger females.
  - There are a significantly greater number of movies that received ratings from fewer than 500 individuals than movies that were rated by more people, which indicates that most movies are not overwhelmingly popular. However, there are stronger correlations between genders once the number of ratings for a movie is higher. This indicates that when more people rate a movie, the data tends to converge between different groups.
  - Movie viewers of more extreme ages (very young or very old) tend to vary more in how they view movies then middle aged people (e.g age 30-40). This indicates that the movie industry might need to put slightly more effort into appealing to certain younger or older individuals.
- Propose a business question that you think this data can answer.
  - Does the movie industry need to expend a lot of money and energy creating that have a distinct appeal to different genders?
  - Answer: For most movies that are rated by a large number of individuals, based on the data they tend to appeal roughly equally to males and females based on the positive correlation charts

between the genders. For certain niche movies rated by fewer individuals, there do tend to be movies that one gender prefers strongly over the other. However, since the movies rated by more people are the money-makers, in general the movie industry does not have to strongly consider the difference between genders. They should just make a small percentage of movies that cater to each gender then spend most of their energy on creating movies with universal appeal.

Done

All set!