Case Study #3
Data Wranglers

Ardavasd Ardhaldjian, Joshua Levy, River Yan, & Anson Zheng

Motivations and Why The Topic is Interesting.

- The film industry is gigantic and worth over $235 billion dollars.
  - Recently Warner Brothers signed a deal to have it have its film management system controlled by AI.
- Its an industry that touches everyone, we all have a favorite movie or tv show.
- Personally I am always interested about which movie is going to be the next big summer blockbuster.

Motivations and Interests Cont.

- The dataset that we obtained from MovieLens gives us insight into the consumers who watch these movies. It can tell us:
  - Who watches these movies
  - What their favorite movies are
  - And how managers can market movies to various consumer groups to remain competitive in this growing and shifting industry
  - And more!

2. Results

“ Number of Movies Rated Above 4.5 Overall

29

Number of Movies Rated Above 4.5 Among Male and Female

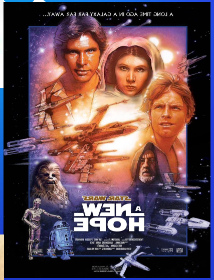
29

70

“ Number of Movies have Median Ratings above 4.5 Among Male and Female Over Age 30

105

187

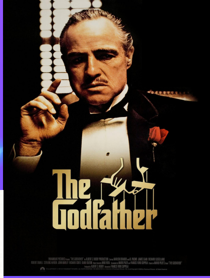


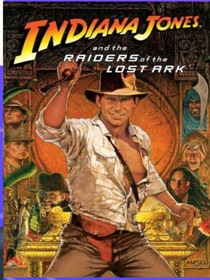
10 Most Popular Movies

Definition of popular:

- Best rated
- Most rated

- The Usual Suspects (1995)
- Star Wars: Episode IV - A New Hope (1977)
- The Shawshank Redemption (1994)
- Schindler's List (1993)
- The Silence of the Lambs (1991)
- The Godfather (1972)
- One Flew Over the Cuckoo's Nest (1975)
- Raiders of the Lost Ark (1981)
- Saving Private Ryan (1998)
- The Sixth Sense (1999)





Gender Conjecture

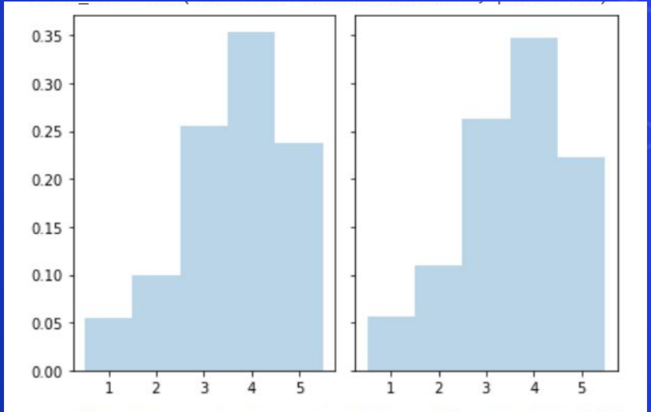
“

Hypothesis Testing:

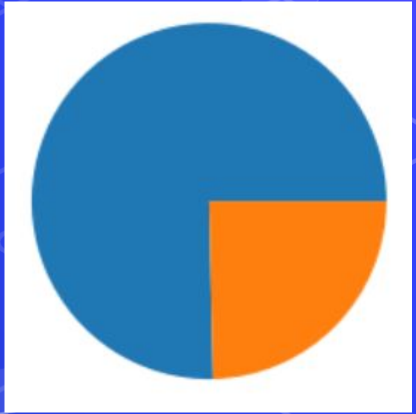
$H_0: \mu_{male} <= \mu_{female}$

$H_a: \mu_{male} > \mu_{female}$

Distribution of ratings by male(left) and female(right)



Distribution of male(blue) and female(orange) in data



Age Conjecture

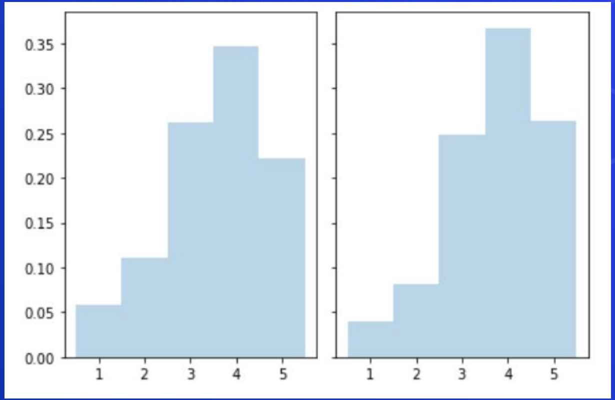
❝

Hypothesis Testing:

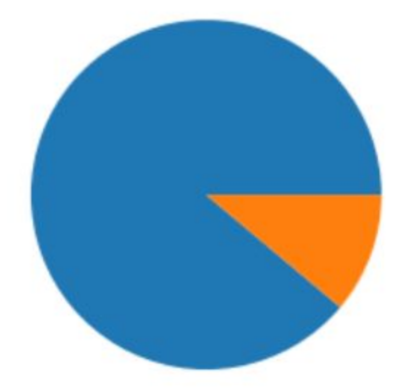
H_0: $\mu_{Below50}$ <= $\mu_{Above50}$

H_a: $\mu_{Below50}$ > $\mu_{Above50}$

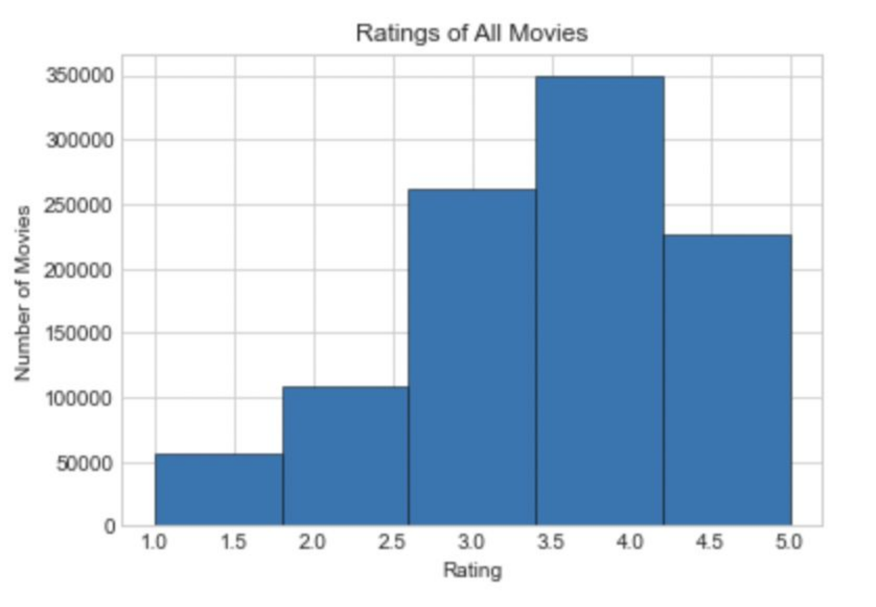
Distribution of ratings by users below 50(left) and ratings by users above 50(right).

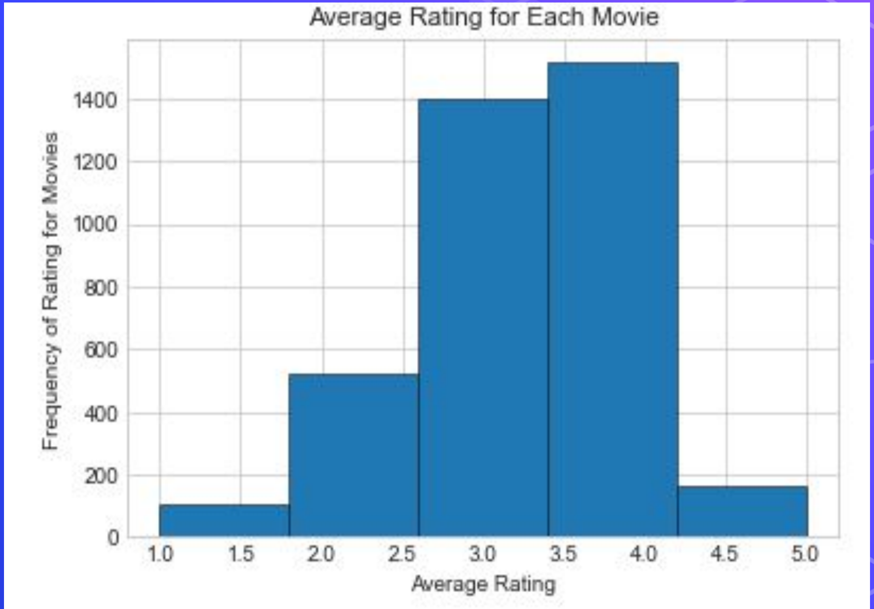


Distribution of users below 50(blue) and users above 50(orange) in data.



Part 2 Results

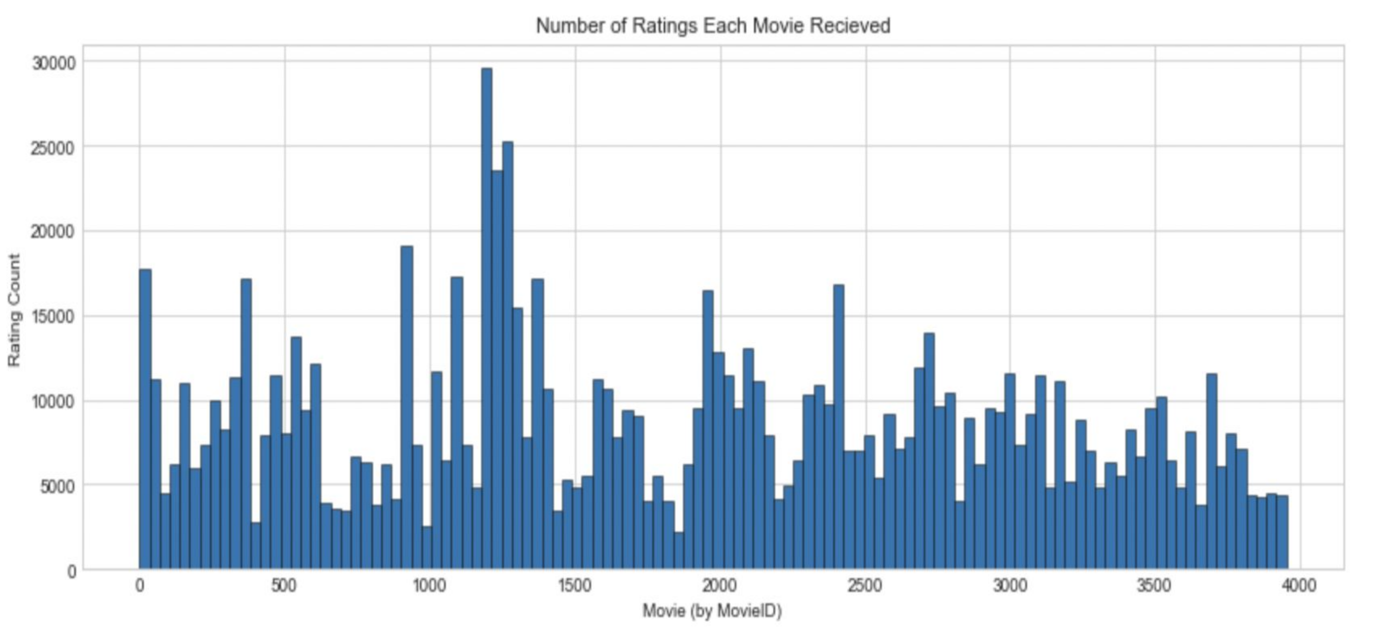




Most movies received a rating of about 4 and the majority of movies had a rating above 3.

Majority of movies received a rating between 3 and 4.

Part 2 Results



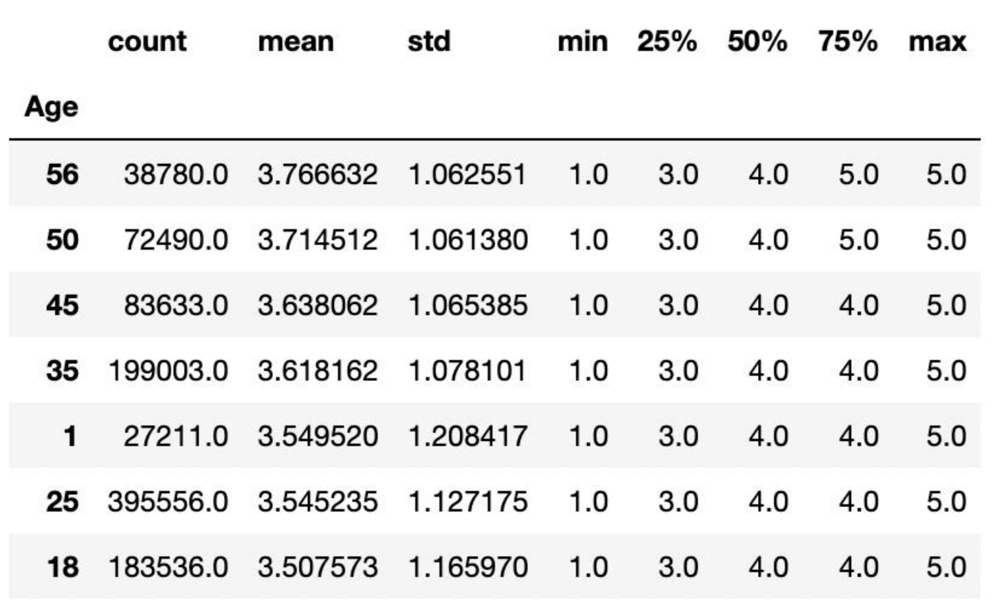
The most highly rated set of movies had close to 30000 ratings. Most movies had about 10000 or less ratings but a few had over 15000 ratings.

Age Conjecture

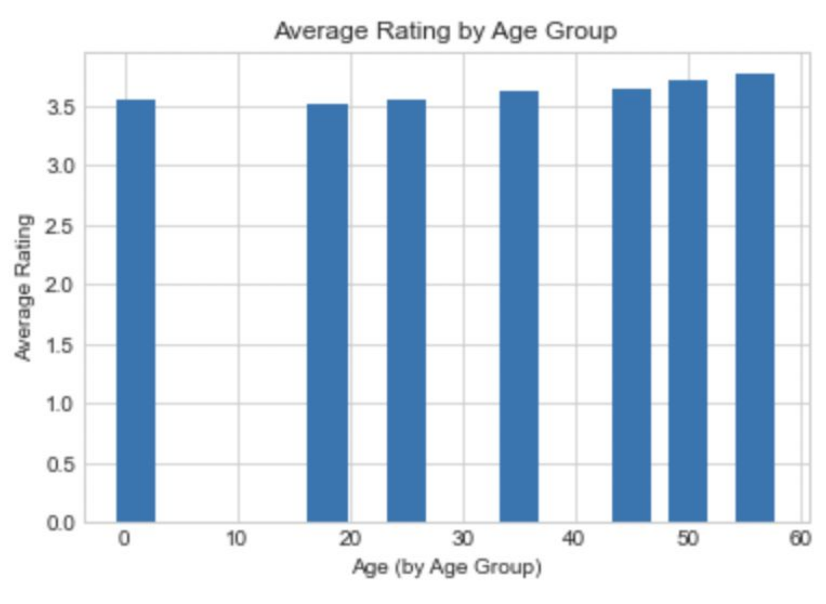
Hypothesis Testing:

$H_0: \mu_{AvgGenderRatingi} = \mu_{AvgGenderRatingj}$

$H_a: \mu_{AvgGenderRatingi} \neq \mu_{AvgGenderRatingj}$



Age	count	mean	std	min	25%	50%	75%	max
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

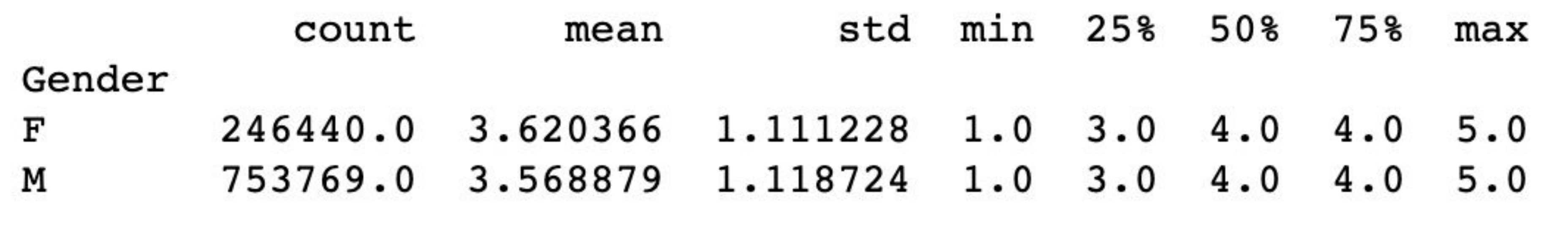


Gender Conjecture

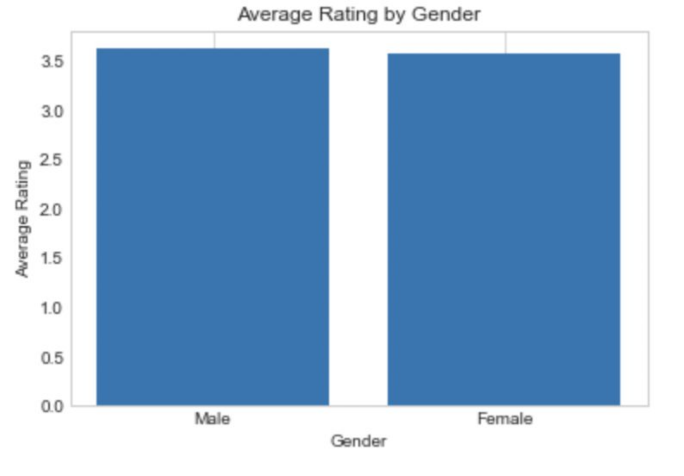
Hypothesis Testing:

$H_0: \mu_{AvgMaleRating} = \mu_{AvgFemaleRating}$

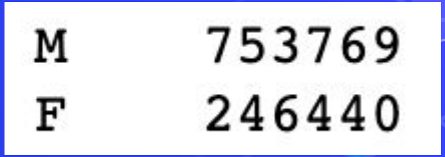
$H_a: \mu_{AvgMaleRating} \neq \mu_{AvgFemaleRating}$



```
           count      mean       std  min  25%  50%  75%  max
Gender
F       246440.0  3.620366  1.111228  1.0  3.0  4.0  4.0  5.0
M       753769.0  3.568879  1.118724  1.0  3.0  4.0  4.0  5.0
```

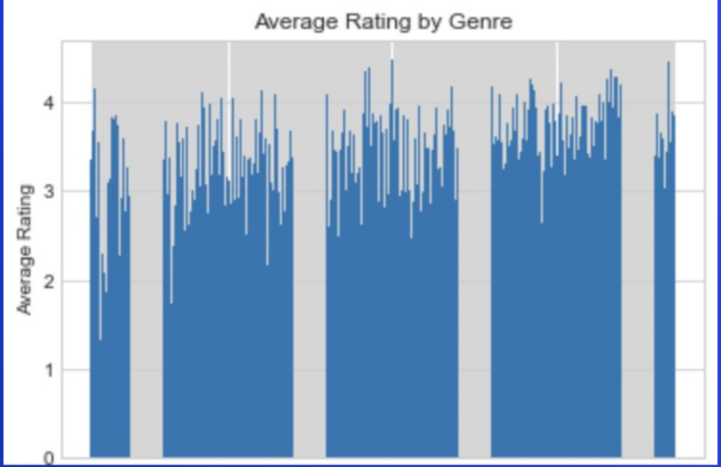


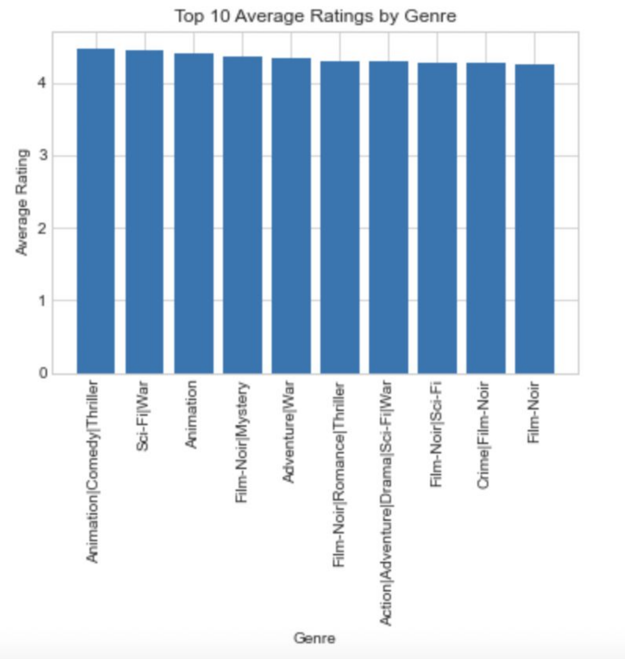
Average Rating by Gender

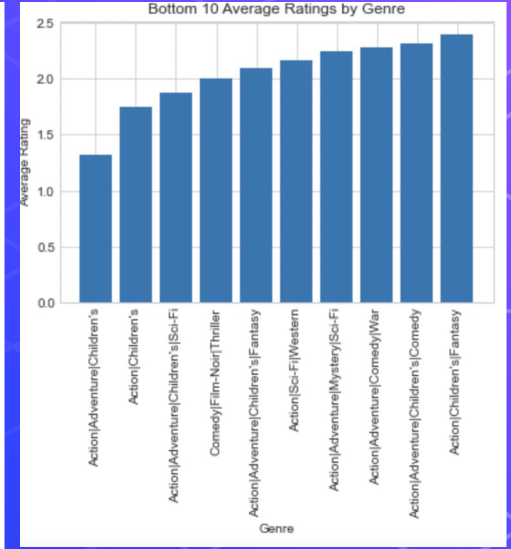


```
M      753769
F      246440
```

Movie Genre Conjecture







Hypothesis Testing:

H_0: $\mu_{AvgGenreRating.i} = \mu_{AvgGenreRating.j}$, while $i \neq j$

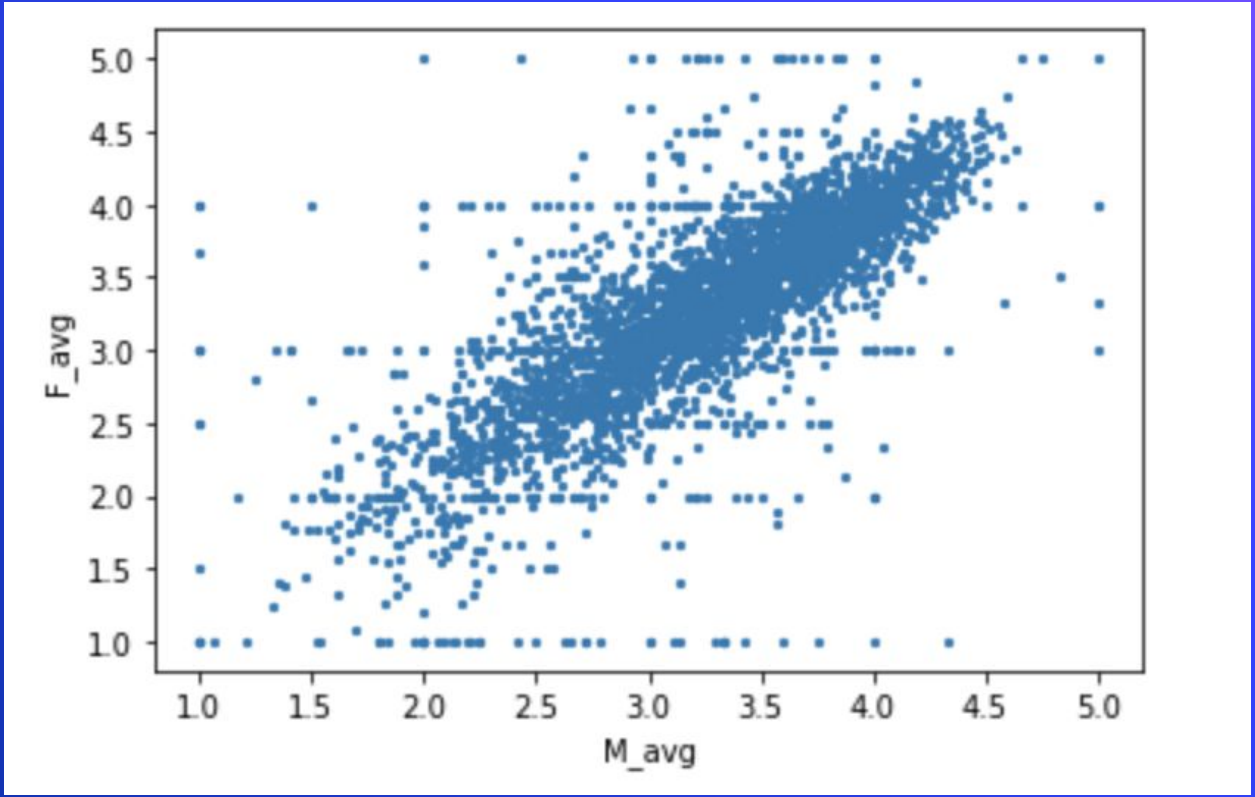
H_a: $\mu_{AvgGenreRating.i} \neq \mu_{AvgGenreRating.j}$, while $i \neq j$

Problem 3: Results (Tools used)

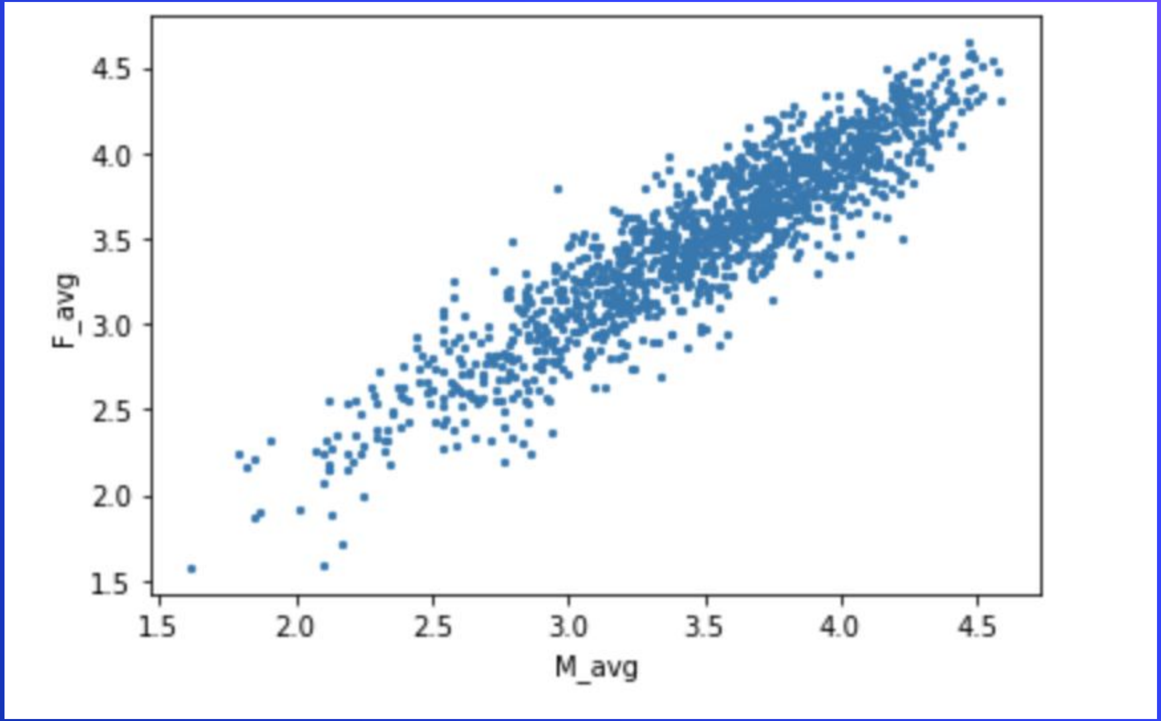
Tools Used:

- Pandasql (Joining tables and and recreating them)
- Pandas (.head(), .unique(), .shape()

Problem 3: Results part 1



Problem 3: Results part 2

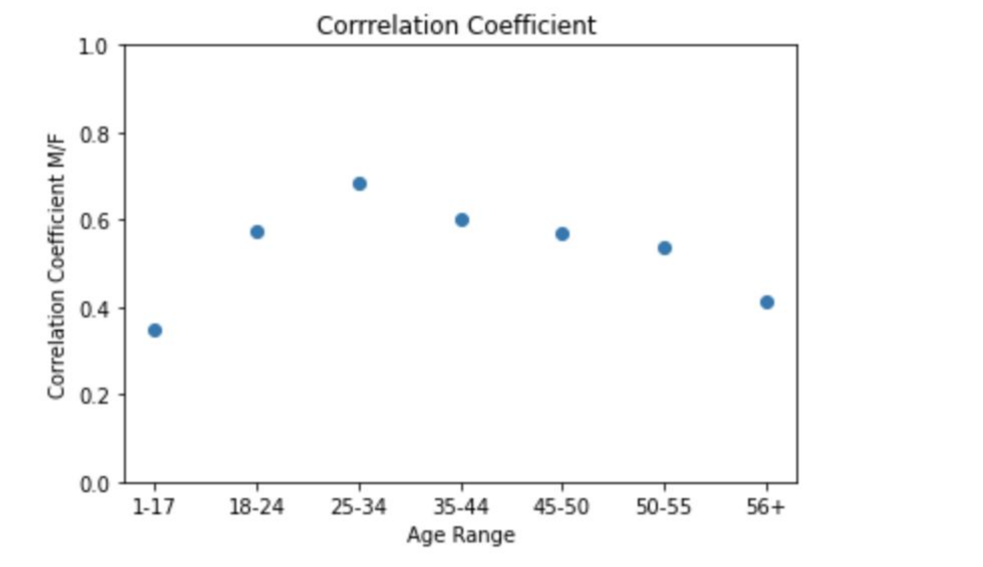


Problem 4: Results part 3

Correlations Coefficients

All Age groups: 0.76

1 - 17: 0.34
18 - 24: 0.57
25 - 34: 0.68
35 - 44: 0.59
45 - 50: 0.56
50 - 55: 0.53
56+: 0.41

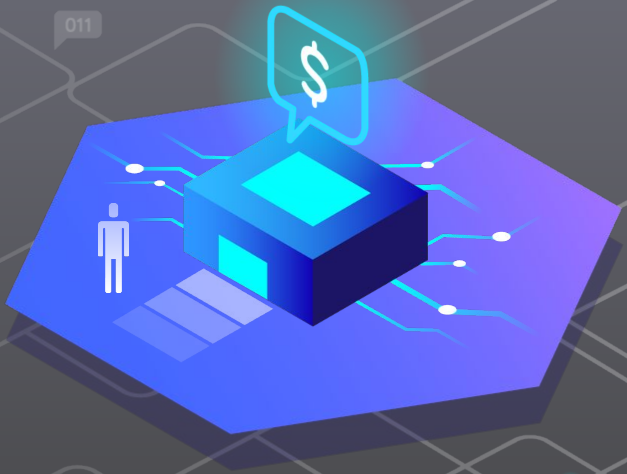


Hypothesis Testing:

H_0: $\mu_{AvgMovieRatingByMaleAgeGroup} = \mu_{AvgMovieRatingByFemaleAgeGroup}$

H_a: $\mu_{AvgMovieRatingByMaleAgeGroup} \neq \mu_{AvgMovieRatingByFemaleAgeGroup}$

3. Storytelling



How does everything fit together?

What our results tell us:

- Older Audiences consistently rate movies higher
- Females on average rate movies higher
- Correlation between age groups (Male and Female) between 25 - 34 tend to rate movies higher than children or middle aged adults

How does everything fit together?

Based on these results if you were a film executive looking for the next big hit:

- Marketing Movies to older audiences is a successful strategy
- AVOID childrens movies. They were consistently rated the lowest.
- Cater movies towards female audiences

The End