**Case Study 2 report**

***Importing the MovieLens data set and merging it into a single Pandas DataFrame:***

Initially the data was converted to csv format for convenience sake. Using different transformations, it was combined to one file. After combining, certain label names were changed for the sake of convenience. Also, a few other label values were changed. The age attribute was discretized to provide more information and for better analysis. The timestamp attribute was also converted into date and time. The dates generated were used to extract the month and year of the same for analysis purposes. Thus, we extracted extra information which can be used for analysis. Also, another attempt was made to extract information from the zipcode attribute in the table using the ‘geocoder’ library in Python. This produced a lot of errors, missing values and took a lot of root time and hence could not be considered.

The data was then converted to a single Pandas data frame and different analysis was performed.

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

21 (Over 4.5. =4.5 is not considered)



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

It was found that men have rated 23 movies an average rating of 4.5 and above and women have rated 51 movies.

***How many movies have a median rating over 4.5 among men over age 30? How about women over age 30?***

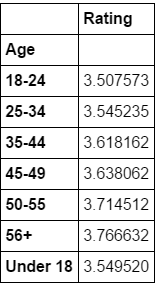
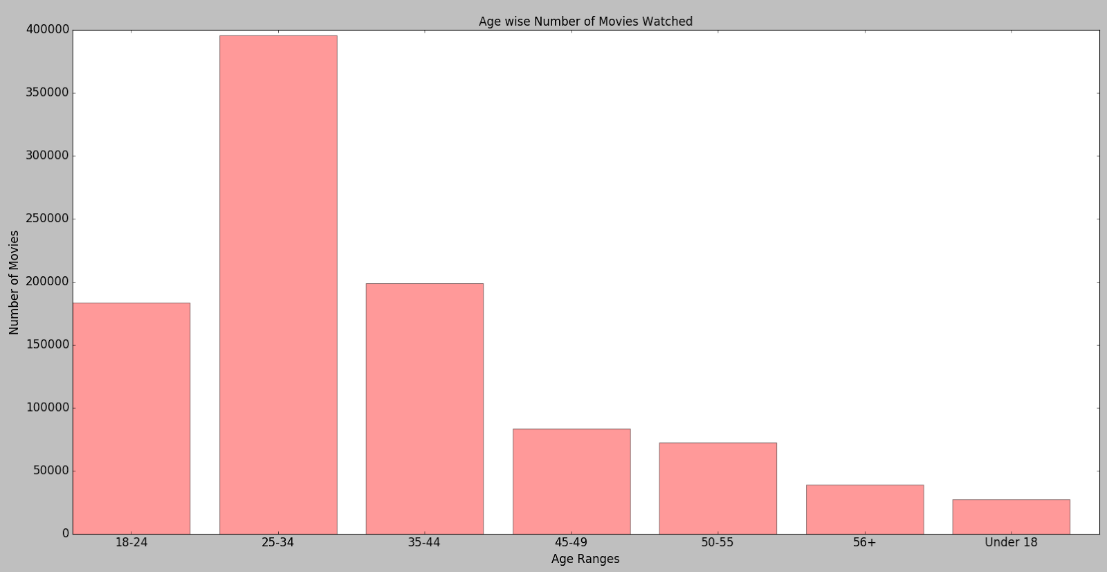
Considering men and women both, around 381 movies for men and 381 for women have an average rating of 4.5 and above.

***What are the ten most popular movies?***

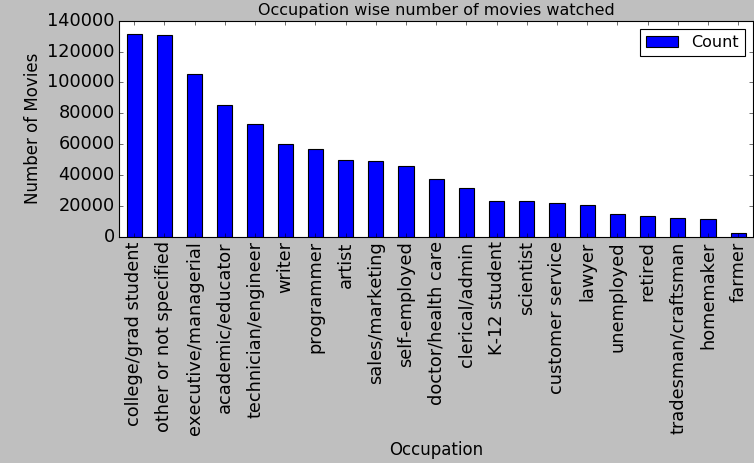
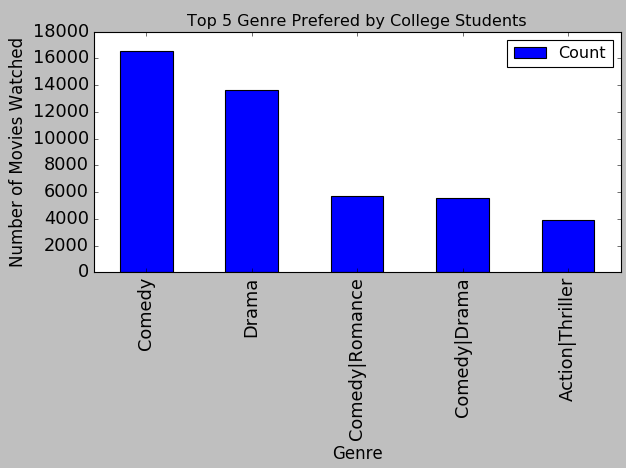


We’ve considered the number of ratings as a measure of popularity. We believe a movie can achieve a high rating but with low number of ratings. Thus, just the average rating cannot be considered as a measure for popularity. Thus, a measure of popularity can be the maximum number of ratings a movie received because it can be considered to be popular since a lot of are talking about it and a lot of people are rating it.

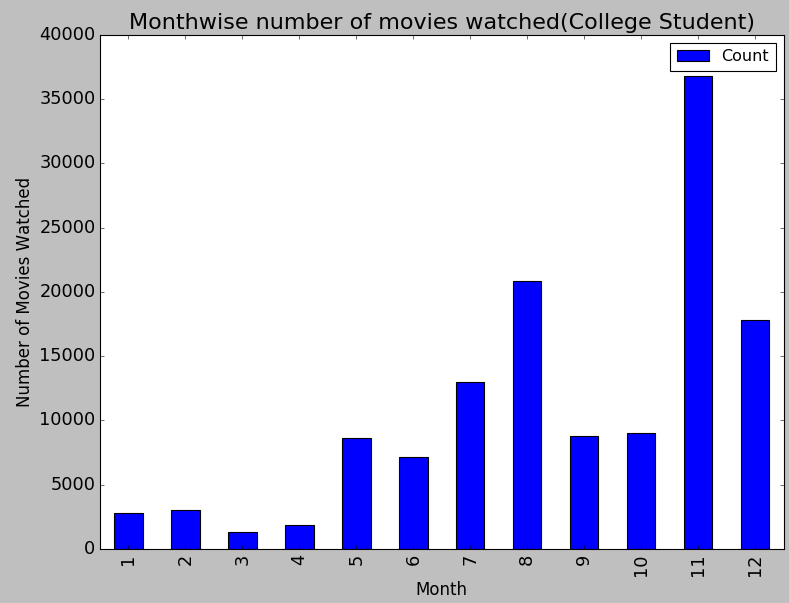
***Make some conjectures about how easy various groups are to please? Support your answers with data!***

The age group 25-34 seems to have contributed through their ratings the highest. This implies two things. Firstly, it shows that the younger working generation is active on social networking websites and it can be implied that they watch a lot of movies in one form another. This information is critical. Also, we see that age groups 18-24 & 35-44 come after the 25-34. Hence, these age groups can be effectively targeted to improve sales. For example, we know that the age groups ’25-34’ & ’35-44’ are the working class and data shows they watch a lot of movies. Companies like Netflix can offer executive discounts to this lot of population since they’re interested in watching movies and a discount can drive them towards improving sales. A decent number of people from the population visit retail stores like Walmart regularly. Walmart can tie up with companies like Netflix or theatres and offer discounts to regular or loyal customers, thus improving sales on both sides. Whereas the age group ’18-24’ represents a lot of students. These companies can promote or let students avail special packages through college events and other activities. Also, looking at their average ratings, it shows they’re not very critical and provide open minded reviews. Thus, this class of population is a good target.

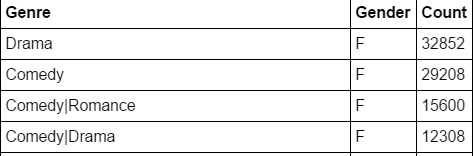
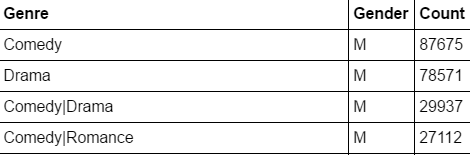
***Occupation wise number of movies watched***

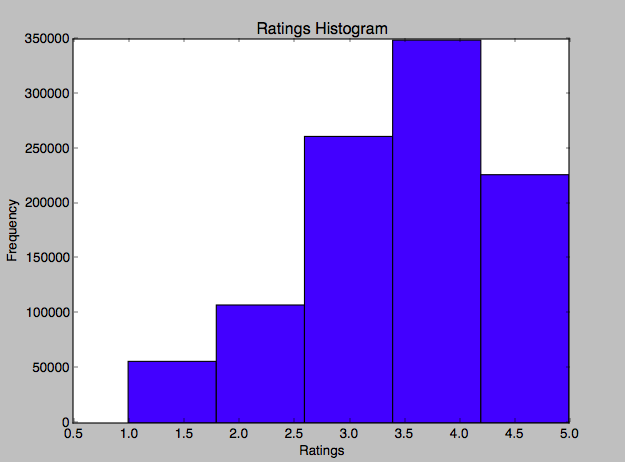
The graph shows that students tend to watch a lot of movies. Naturally, this habit of students is not surprising since a lot of students’ love watching movies and some of them view this as a social activity to enjoy with your friends. Also, further analysis proves that students love watching Comedy and Drama genres. This gives direction for strategical decision making for companies in the film industry. As stated above, they can offer exclusive discounts to students to elevate their sales.

Icing on the cake, the graph shows that college students tend to watch a lot of movies in the month of November. November indicates Thanksgiving break. Thus, targeting audience during family holidays especially during the month of November will benefit these companies.

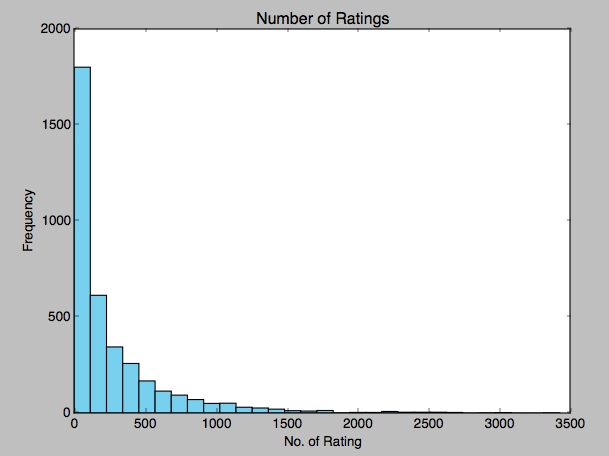
Below, we’ve compared the genres & the number of ratings and feedback that the audience prefers to give. It seems that the audience prefers to comment on movies of the genre ‘Comedy’ & ’Drama’ or a combination of both in a movie. Thus, it can be concluded that these genres are the movies of interest for the population. Sales of such movies can be increased by effective marketing, more publicity, availability of discounts and cheap rates.



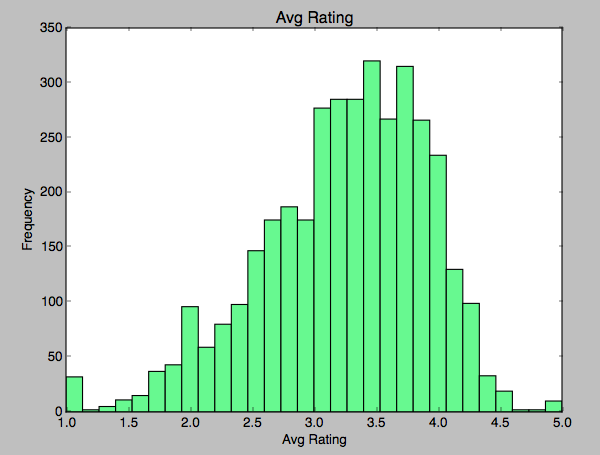
***Problem 2***

a.

The histogram shows the general distribution of the ratings for all movies. The histogram shows that the audience isn’t really critical. A very low population of people have contributed with ratings as low as 0-2.5. Most of the ratings lie between 2.5-5 which indicates the audience is generous. Maximum ratings are in the range 3.5-4. Movies with such ratings can be used to analyze upcoming movies of similar taste and to predict the crowd response on these movies.

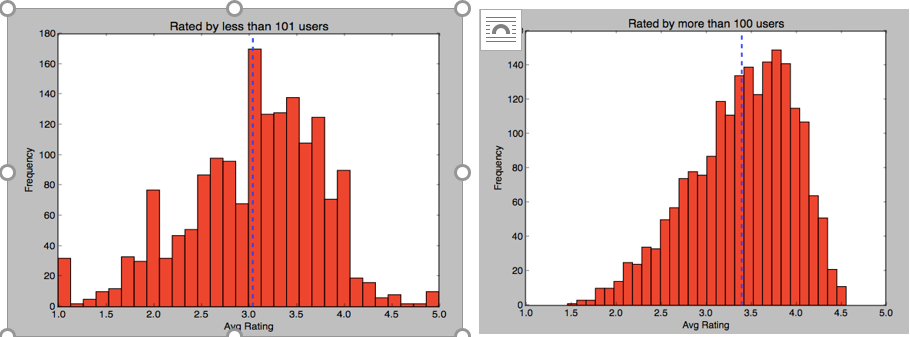
b.

The histogram shows that a lot of movies have very few ratings. Thus, they have not been adequately rated and thus can cause some consistency while analyzing data. There are a few movies who have considerable amount of ratings too which can be considered for analysis.

c.

The figure provides a histogram by finding the mean of the ratings for each movie. It gives a Gaussian distribution. The rations between 3.0 to 4.0 give the maximum amount of ratings. Plus, the range of rating illustrates that a lot of generous ratings have been provided.

d.



In both scenarios, histograms are skewed to the left which conveys that the tail is on the left but when all

movies are considered skewness is less as compared to when only those movies are considered which were rated more than 100 times.

Above two graphs indicate the histogram of average ratings when rated more than 100 times and when rated

less than 101 times. In both graphs, there is a vertical dashed line that indicates the average value of these ratings.

So when movies rated more than 100 times it has 3.40 average rating over all movies but when movies rated less than 101 times it has 3.04 average rating over all movies.

These average values show that if more users rate the movie then on an average movie is good as compared to those movies that are rated by few users.

***Conjecture 1:***

Table 1 represents top 5 genre that were rated by maximum users and Table 2 represents top 5 Genre having

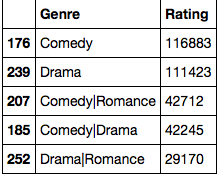
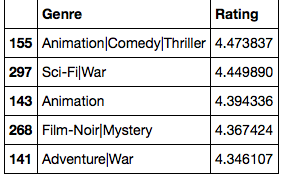
on an average highest ratings:

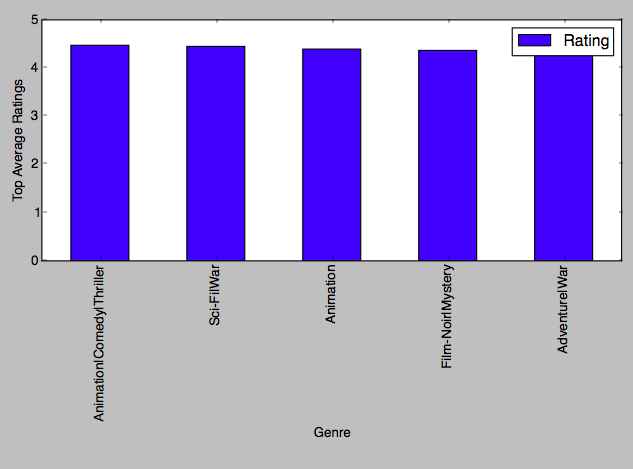
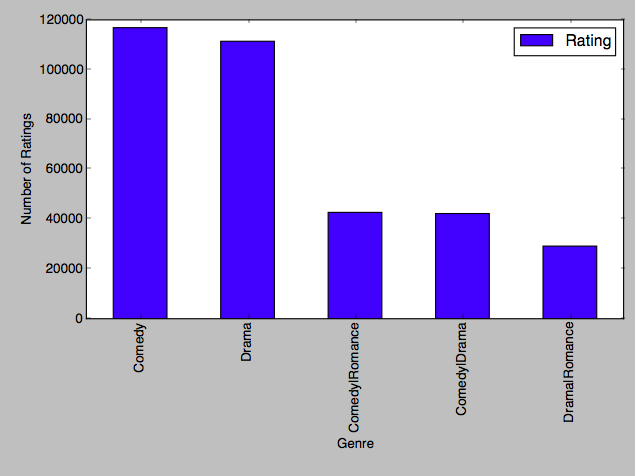
Genre that were rated by maximum users may not be the true representation of movie ratings as ratings can be given by

users and bots. This represents high bias in the data.

On the other hand, Average rating in table 2 may have sampling biases which means it was rated by few users who rated movies high and ignore ones who rated movies low and that leads to high rating.

Table 1 Table 2

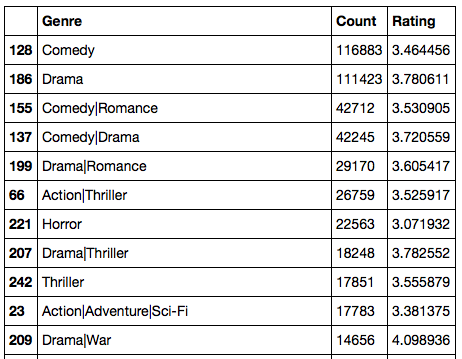
 



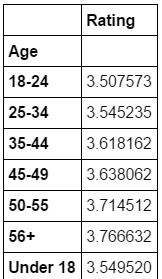
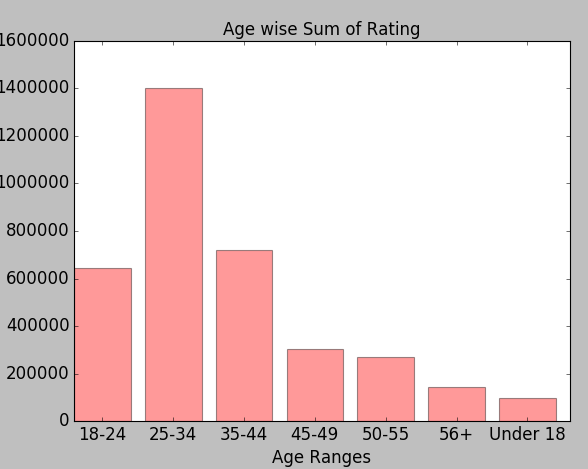
To overcome above biased ratings one should seek to look for those Genre that show the true representation of

ratings by considering legitimate users and by considering enough users or samples.

Count + Rating:



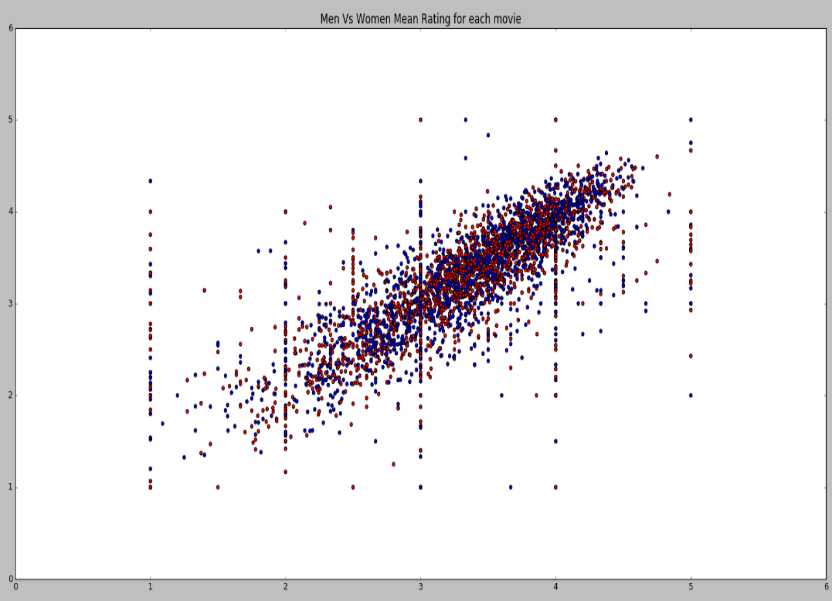
*Conjecture 2:*

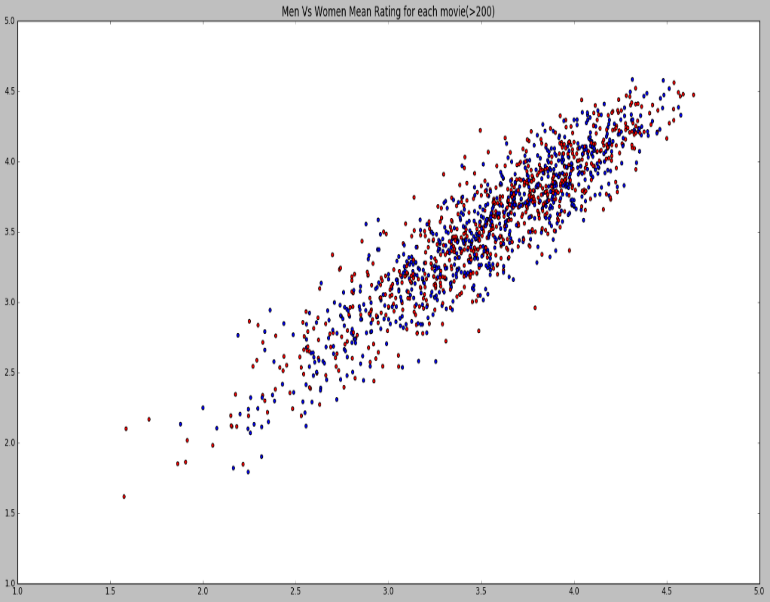


We can see that sum of rating for age group 25-34 is the highest of all. But, average rating for each group is same and we can find out from the figure on the right hand side. Hence we cannot predict about Ratings and Ages with this method.

***Problem 3***

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



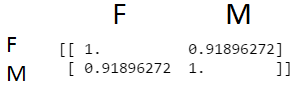


**Left Figure:** The above scatter plot shows that the average rating of me and women show a linearly increasing trend. It says that excluding a few movies and a few ratings, men and women tend to think alike.

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

The above scatter plot was produced by segregating only those movie ratings who have been rated more than 200 times. The average of these ratings for me versus women was plotted. It shows a similar linear increasing trend as in the scatter plot where ‘number of ratings > 200’ was not considered. Thus, people are like minded and they like what everyone likes to watch.

***Compute the \*correlation coefficent\* between the ratings of men and women***



*What do you observe?*

The correlation coefficient shows that there is very high correlation between the ratings of men and women. Thus, indicating that men and women think alike when it comes to movies. A correlation coefficient of 0.92 is very high and shows high relevance. The above scatter plots also illustrate the same. Hence, it can be said that if a man happens to like a particular, we can expect his other half to also like the movie. But if we plan to develop a model that predicts according to the above analysis, it will not be that efficient since this will produce high collinearity because of the high correlation. A solution to this problem is to use an attribute individually or in a combination with another attribute.

***Are the ratings similar or not? Support your answer with data*!**

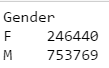
As we can see from the above scatter plot, ratings are almost similar as both Males and Females follow the linear trend.

Average Rating overall for men and women:

https://lh6.googleusercontent.com/mqKWVMdcidQo0sOzyeCg-SrIew8UBm_abmKS8RZWf-xZAjeOKlcQcP1NLcwxExjx02CN2OuFR5CLW768nroL5ZHQV68utx4h9us_j-APOBkzof1_tkWY7-IW0sIMRanamN0c_TTN

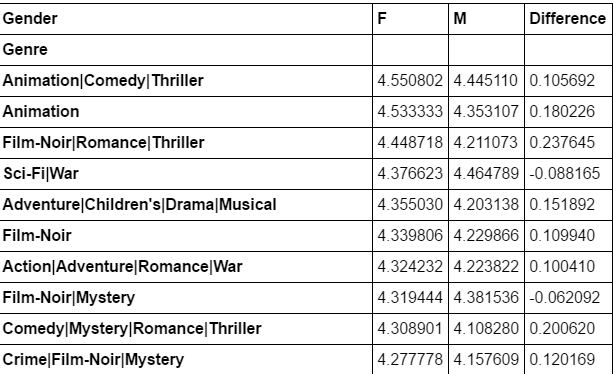
You can say that average ratings are almost similar. But there may be some discrepancy in above results because as you can see from below results, number of movies rated for men is much higher than women

***Number of Movies Rated*:**

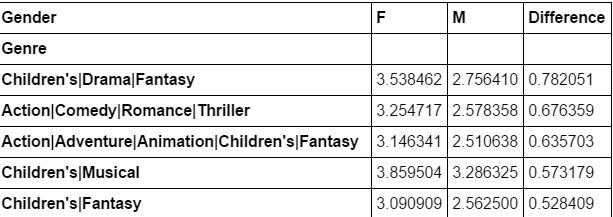


Though number of average ratings are similar, count of number of movies largely differ. Hence, we cannot accurately predict just on the basis of this analysis. More filtering is required.

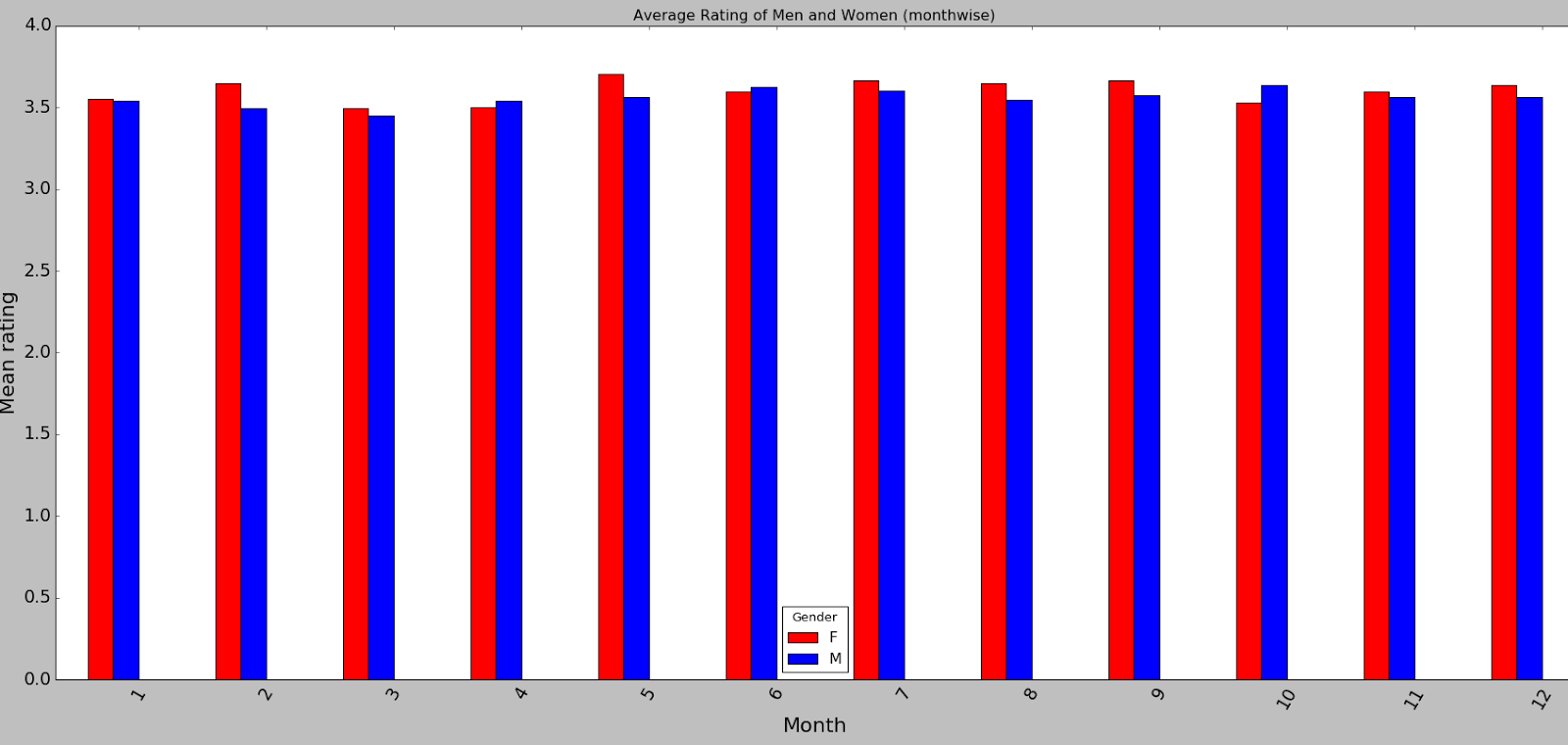
***Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.***



These genres are highly rated by men and women both and on observing, you can see a very slight difference in the ratings. This implies that they are similar and they prove the analysis explained by the scatter plots.

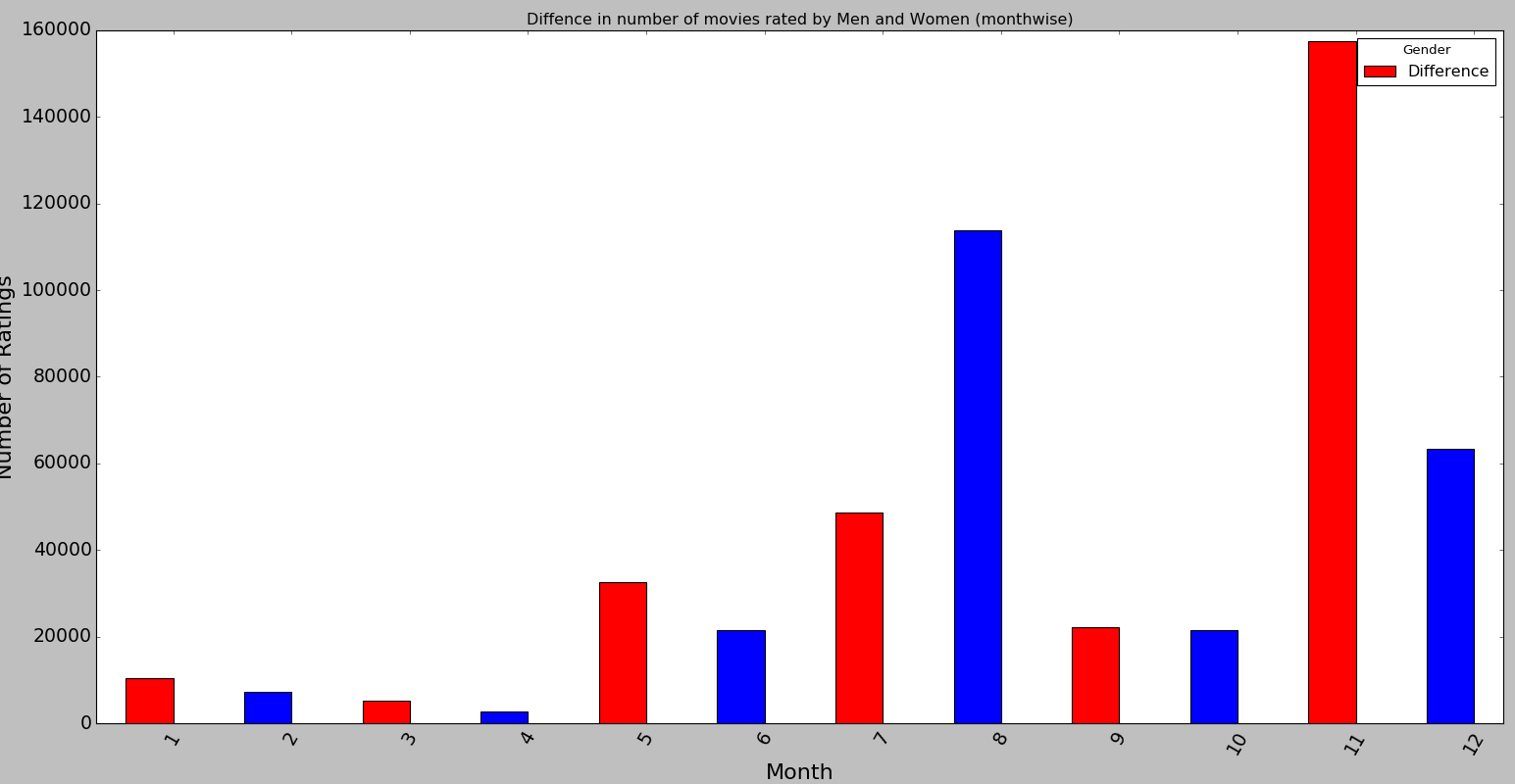
These are some of the special cases where difference in Rating of genre is greater than 0.5. This value is not large enough though. Hence we can use to predict a general trend that if a male viewer likes a certain genre then what is possibility of a female liking it.

Month wise average rating for men and women:

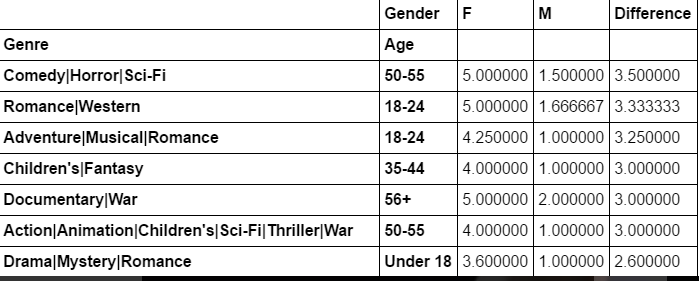


Although, there is high similarity in the ratings provided by men and women, there isn’t a large difference in monthly average rating. But, compared to other months, January tends to hold the most similar ratings between the two sexes compared to February or May.

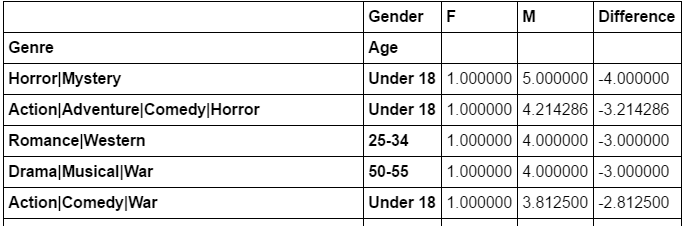
Moreover, below plot depicts that the difference in number of movies rated in the first 4 months is less compared to the months that will follow. This strengthens our conjectures that, in January the difference in number of movies rated is less and also the rating average is similar.



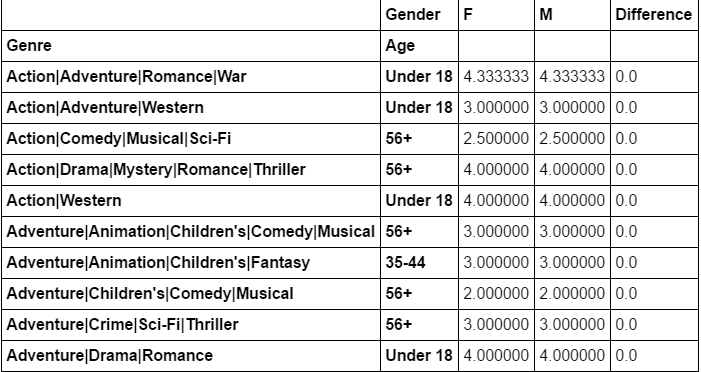
Average Rating group by Genre and Age:



These genres depict disagreement between the two sexes. The table shows that elders tend to like Comedy|Horror|Sci-Fi movies which is interesting as we’d think the opposite about them.



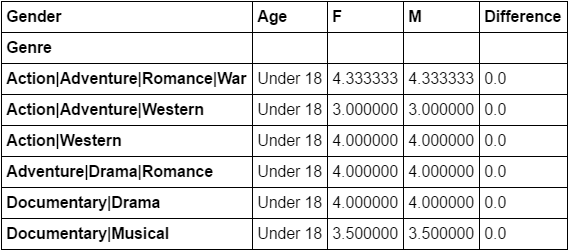
However, these are few due to similarity compared to similar liking which is shown below:



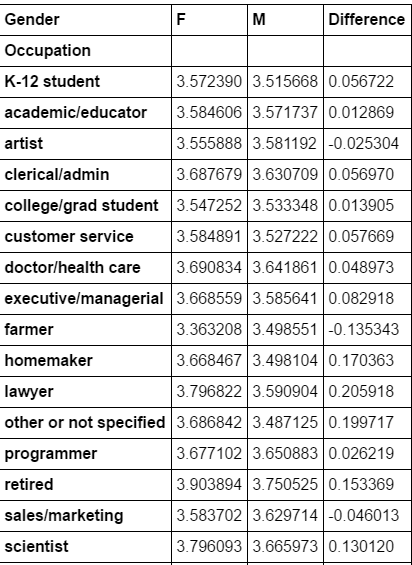
You can say that, for various genres with given age, we can predict that if a male tends to like that genre, then females will also like them.

Below table shows genres which are liked by both, males and females.

These are the Genres where Male and Female have similar views if their age is Under 18. This table shows that a lot of energetic youth are into action and adventure movies. A lot of teenagers who are interested in Theatre & Drama are providing positive feedback for genres like Drama, Musical, Western and Romance. A lot of teenagers are drawn towards Documentaries which help them learn.



Occupation wise Average Rating Trend of Male and Female:



The above table relates occupation, sex and the average rating. It can be used to link like-minded people. For example, movies that doctors will love can be used to predict the probability that a female likes the movie, given that a male likes it.

***Problem 4***

Initial analysis for the number of movies showed that men rated more number of movies than women. Also, further analysis showed that the number of ratings above 4.5 was more for women than men. Thus, it can be implied than women are more easy to please and they are generous while giving out feedback.

It was also observed that a lot of users in the age group ’25-34’ were contributing heavily to ratings. These age groups consist of that part of the population which belong to the working class i.e. they are company executives and managers. Thus, we concluded that companies in the movie industry must concentrate their marketing in such areas. Such marketing strategies can be implemented during the month of November, when we have found that most of the population goes to watch movies. Also, ‘Comedy’ & ‘Drama’ genres are very popular as we’ve found.

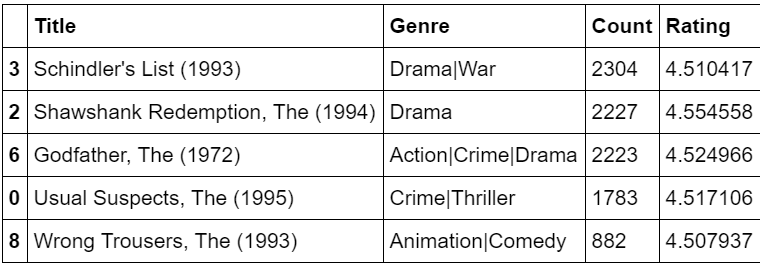
*Business Question:*

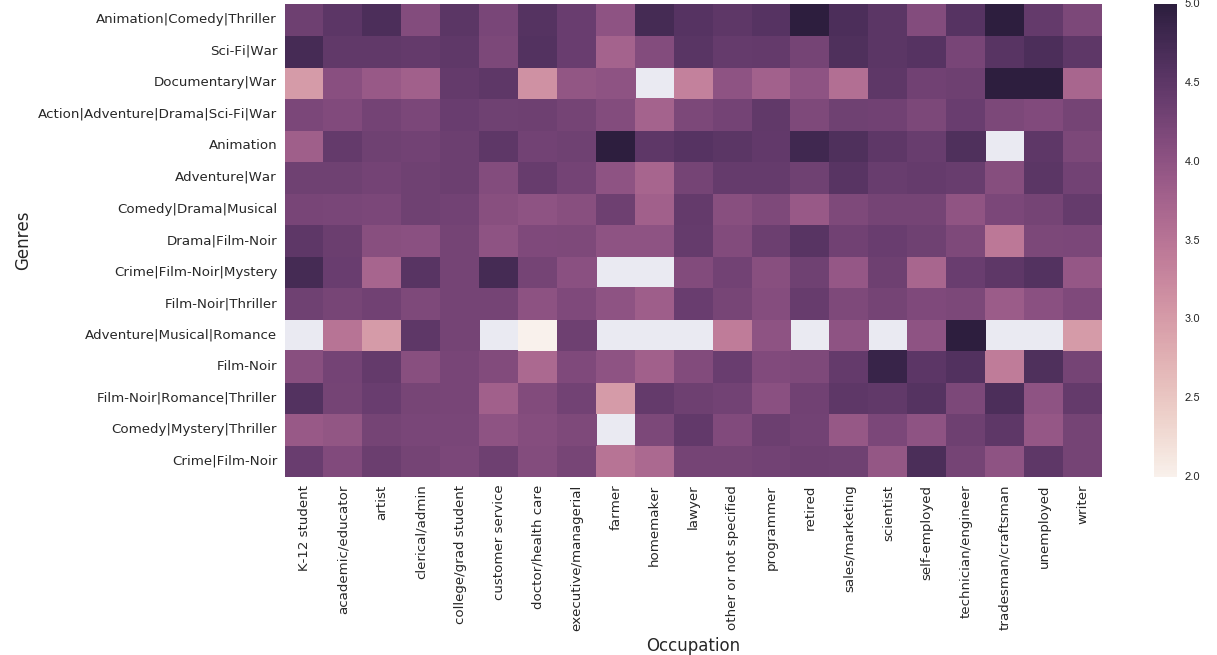
**Market-wise Customer Segmentation by Smart selection of the Target Audience. Analysis is to be done to find out the best Package to Target User using Data Science which can increase the loyalty of customer towards the company or increase the profit.**

*We will do this by finding out analysis about Target Audience. For Example: Which genres are preferred by males and females? Which genres are extremely disagreed upon by both the sexes? Also, categorizing by age, which age groups must be concentrated upon for marketing and increasing sales*? And so on.

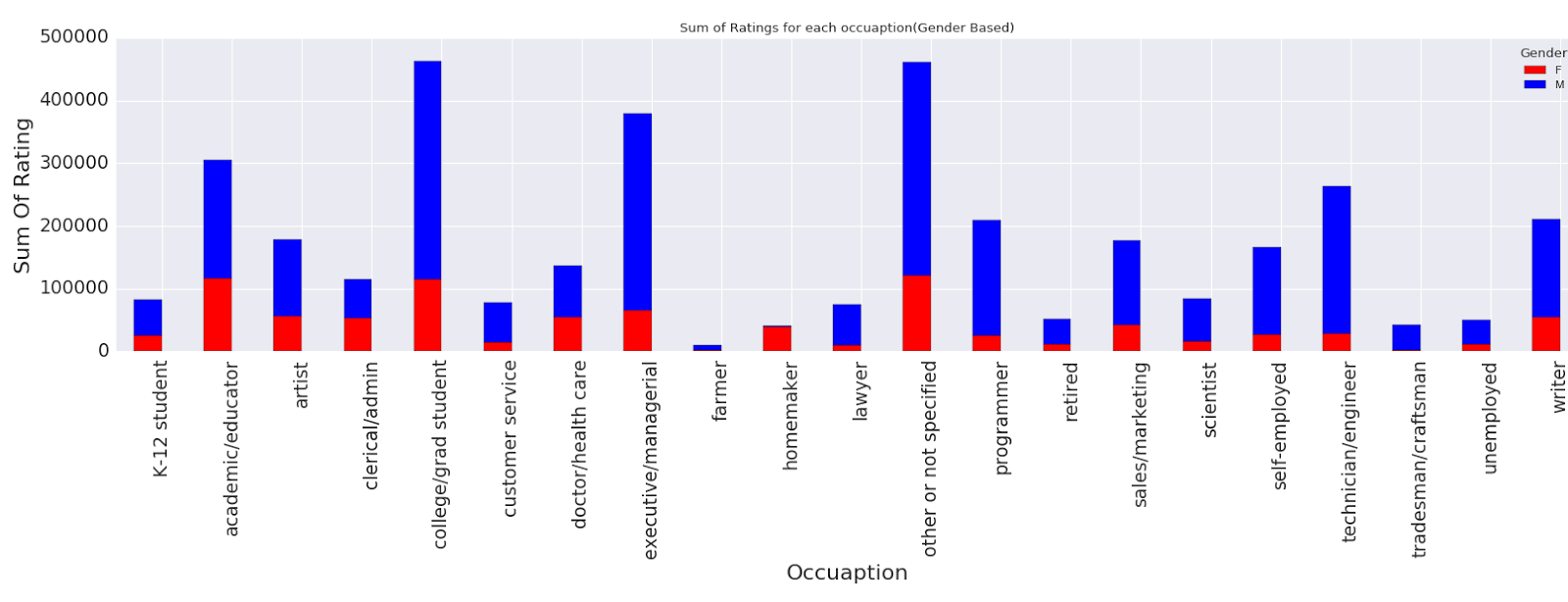
We have already done in detail analysis in Conjectures about Target Audience that needs to be considered.

There are few movies which has overall rating over 4.5 with the fact that they have been rated by many number of users. These can be considered as all-time favorite movies and can be recommended to all users. Hence, this can be recommended or should be part of package that we will suggest to target audience.



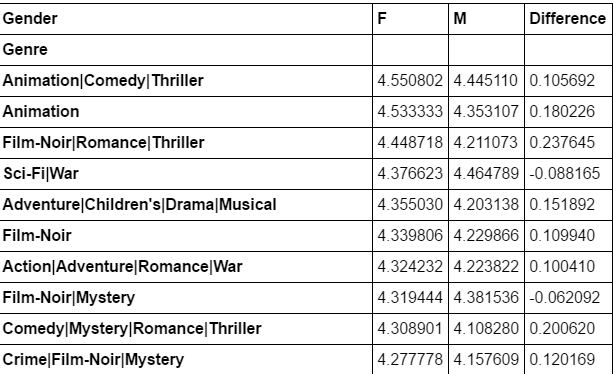


From the Above Graph, we can state the relationship between Occupation and Genres of Movies that an individual prefer. For Example: Farmer do not prefer to watch Comedy|Mistery|Thriller and College Student Prefer Animation|Comedy|Thriller.



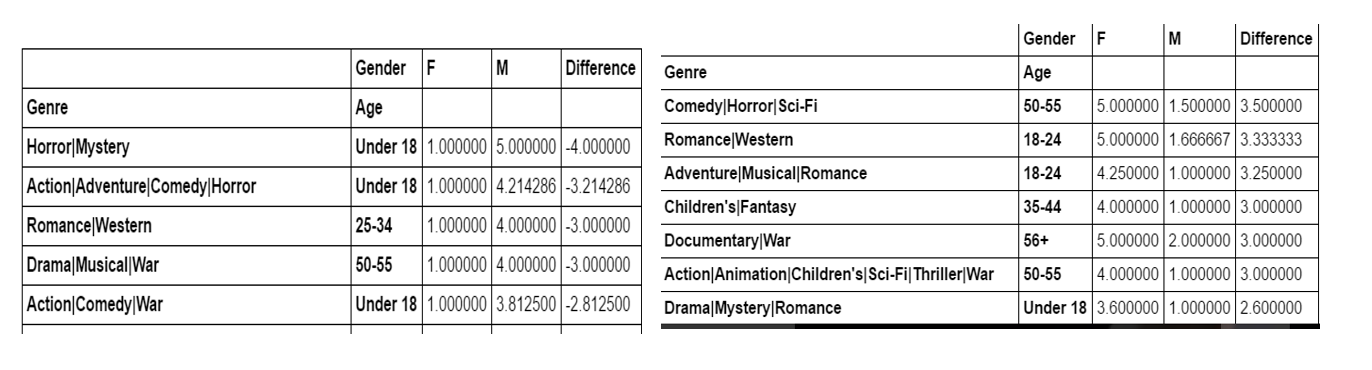
We can find out from the above graph the Target Audience that the company should consider. For Example: College Student tends to rate more movies than any other groups. Moreover, company can find out about the gender Biasness from the above graph. For Example: there are no female farmers who rates the movies

As we’ve found out, the genres Animation/Comedy/Thriller are very popular among both the sexes. They’ve provided very similar ratings and feedback on them. We have recorded those genres that have a difference of less than 0.5 in ratings. Thus, we know that in order improve sales, these genre of movies must be promoted, more discounts and corporate coupons must be provided to improve sales.



Below table shows the genres which are most disagreed upon and the age groups that disagree the most on these lines.

As we can see there is great disagreement between teenagers on topics of genre of movies.



The above table shows that the age group 50-55 also show huge disagreement in the genre of movies they prefer. Thus, majorly, it’s the elderly and the teenagers who contribute in a disagreement.

Thus, it makes a lot of sense to concentrate efforts on the working class rather than convincing the elderly and the youth. Also, collaborations between organizations and companies can lead to better marketing on both sides since the working class watches lots of movies. Unlike, the others this range of population agrees in their liking of movies. There is hardly any disagreement faced.