# Analyzing data from Movie-Lens

DS-501 Case Study 2 Team #6 Praneeth Nooli & Rahul Ghadge

# **Introduction**

One day, Rahul suggested Praneeth “Hey buddy, Fast and Furious is great movie series you should watch it!” and a week later Praneeth suggested Rahul “Dude, Sinister is awesome movie, do watch.” But a month later when they discussed about it again none watched the movie suggested by other. Because both liked different genre of movies. Rahul liked action movies whereas Praneeth liked horror flicks.

We know what movies that we ourselves enjoy, but how are we supposed to know what movies another person may prefer? We don’t know what movies they like, or even what movies they’ve seen for that matter. The question of recommending a movie is more daunting than its face value. We certainly don’t want to disappoint our friends, nor lead them into a two-hour wasted investment of their precious free time. Yet we must attempt to recommend a movie based solely on their personal background like- age, gender, geographic location, and occupation for example.

This question extends far beyond maintaining personal social relationships. Movie streaming businesses, such as Netflix®, acquire new customers every day. Each time a new member logs-in for the first time, Netflix® must recommend a movie to that new user based solely on login demographic information. Ensuring a positive first experience is vital in maintaining and growing a satisfied subscriber-base, making this question all the more important.

How do we recommend movies to others based solely on their personal information? This question seems both incredibly difficult and complicated to answer, yet surprisingly relevant both socially for individuals, and economically for direct-to-consumer movie services. A question that is both difficult yet relevant is precisely one that is interesting. Moreover it’s a question worth asking and, as we will do in this report, worth considering how to answer (or, at the very least, answer up to some reasonable level of certainty). We will begin to explore how to answer this question through the following four objectives:

Objective 1: Collect data and analyze basic details.

Objective 2: Extending our investigations to histograms.

Objective 3: Correlation-Men vs. Women.

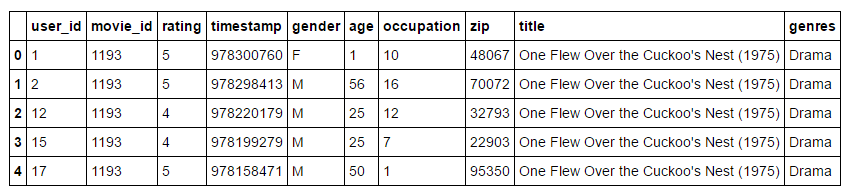
Objective 4: Business Intelligence.

# **Collect data and analyze basic details.**

**Data**: This report uses the *MovieLens 1 Million Dataset*, from <http://grouplens.org/datasets/movielens/>. The biggest advantage of this dataset is that it has already been cleaned and organized, and is ready for analysis. The dataset is comprised of three data files, *users*, *movies*, and *ratings.* *Users* contains the profile information for approx. 6000 users who submitted ratings – including a user ID, as well as gender, age, occupation and zip code. *Movies* contains basic information of the around 4000 movies that were rated– including a movie ID, title and genres. Finally *ratings* contains data about each of the 1 Million movie ratings submitted– including the user and movie IDs along with the rating and a timestamp.

***Import, merge, and store data*.**

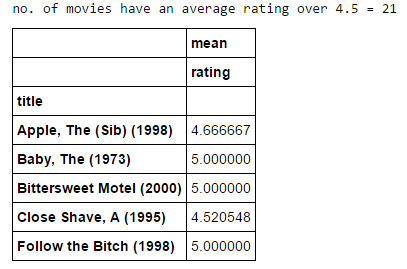
Each of the three data files described above were downloaded directly from website link provided. Each individual file was read into Python as a Pandas Data Frame. Thanks to the relational nature of these files, the data was merged into a single Data Frame containing pertinent information for every movie rating. This main Data Frame was then stored in HDF5 file. An example of the first 5 rows of this data frame appears below.



***Basic statistics and analysis of data collected.***

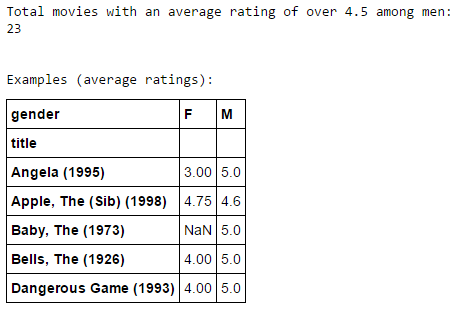
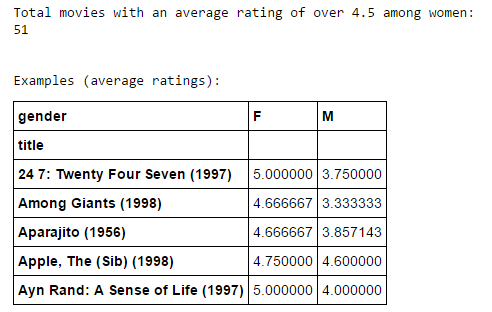
*How many movies have an average rating over 4.5 overall?*

First, we considered the number of movies with an average rating above 4.5 overall (out of a best-possible 5). Using a Pandas pivot table, we determined there were 21 movies with an average rating of at least 4.5. Five examples including movie name and overall rating are displayed below.



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

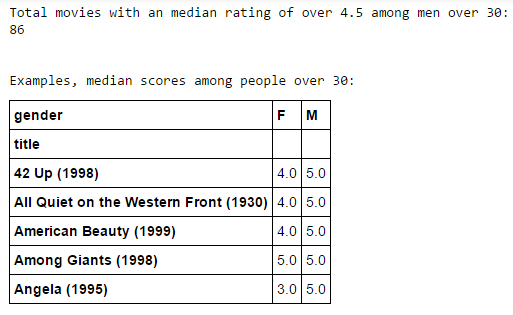
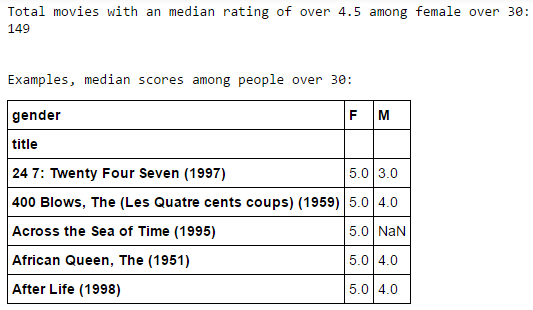
These rating were submitted by both men and women who, likely, have differing tastes in movies. Thus while only 29 movies had an average rating of 4.5 overall, many movies may receive high average ratings amongst a specific gender. We counted the number of movies with an average rating over 4.5 among men and women (again with a pivot table). There were **23** movies satisfying a men’s average rating above 4.5, however among women’s ratings, there were **51** movies with an average rating over 4.5. Five examples of each, including the average rating amongst the other gender, are shown below.

These results do not indicate that women are “easy to please” whereas men are more difficult. Among the users submitting ratings, 4,331 are males while only 1,709 are females. Accordingly, over 750 thousand ratings were submitted by men, compared to only 250 thousand by women. Thus while there are 70 movies with high average ratings by women (more than twice the number for men, at 29), these ratings are being masked in the overall average ratings by the dominant number of male ratings. This is why the number of movies with an overall average of 4.5, and an average of 4.5 among men, are so similar.

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

We used a pivot table to collect all movies (by title) with a median rating above 4.5 among both men and women whose age is over 30. There were **86** movies with a median above 4.5 among men over 30, and **149** qualified movies among women’s rating. Five examples of each (including the median among the opposite gender) are shown below.

Observe that many of these “popular” movies (in terms of median) have a median rating of 5.0. In fact, this means that these movies have *all* ratings of 5.0. This doesn’t indicate overwhelming success of the movie, as these calculations do not factor in the *number* of times each film was rated. This will be explored further in the second objective.

*What are the ten most popular movies in the data set?*

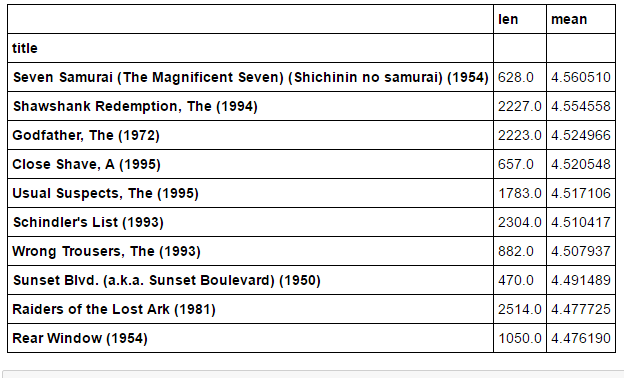
We are considering a movie is “Popular” if:

1 – It has above-average ratings amongst both men and women (i.e. an average rating amongst

Men which is higher than the average of *all* ratings given by men)

2 – The total number of ratings is higher than the average number of ratings per movie

Our interest is the most popular movie *in this data set.* Therefore the “society” or “group” in consideration is specifically the individuals who submitted ratings (i.e. samples). This is small subset of people over the globe (i.e. population). Due to sampling bias, these ratings are certainly not reflective of the world population. Therefore these movies are “Popular” in the sense that they are the most popular of the 3,900 movies rated, *among the individuals who submitted ratings.* Once we determined the “Popular” movies given this definition, the 10 *most* popular were chosen based on highest average overall rating.



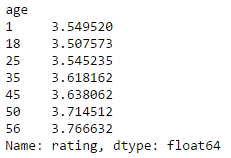
*Conjectures about how easy various groups are to please.*

Before we even recommend a movie, our success or failure could depend on the inherent *happiness* level in various groups of people: how easy they are to please.

***Conjecture 1:* *The older a person gets, the more difficult they are to please.***

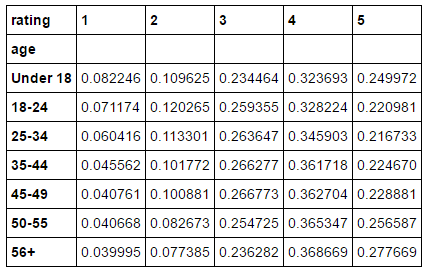
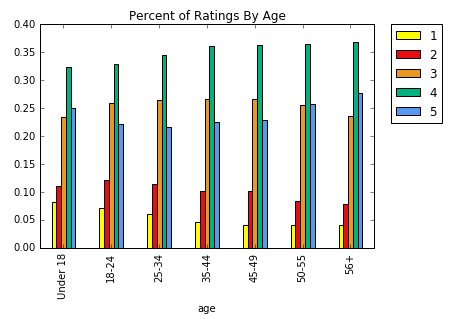
As a person grows in years and life experience, it becomes more and more difficult to impress them, or show them something “new.” Therefore we postulate that as a person ages, they are more difficult to please. Within movie data, this will be reflected by lower ratings from higher age groups.

First, we computed the average among all ratings provided by each age group:



As we can see, there is actually an *increase* in average rating in higher age groups. While this difference of 0.25 in the average is not enough to support or refute our hypothesis, it suggests that there may be a difference in overall ratings among age groups, and thus we investigate further.

Beyond average ratings, a better measure of how “easy” or “difficult” it is to please a group of people is *percent* of high and low ratings given. That is, if a group is difficult to please, the percentage of *low* ratings out of *all* ratings submitted by that group will be higher than a group that is easier to please. Similarly, a difficult group will submit a lower total fraction of *high* ratings. The following table and plot displays the percentage of each rating (1 – 5) submitted by each age group.

These results confirm the general trend observed in the average ratings. Of those ratings submitted by individuals over the age of 56, 4% of those ratings were 1 whereas 27.8% of them were 5. On the other hand, among 18-24 year olds, 7.1% of ratings were 1 whereas only 22.1% were 5. That is, the oldest age groups submitted 5% more ratings of **5**, and 3% fewer ratings of **1** than did the 18-24 age group. The plot demonstrates that this, in fact, appears to be a trend: in general as age increases, the percentage of both 4 and 5 ratings increases whereas the percentage of both 1 and 2 ratings decreases.

*Conclusion:* Given this movie-rating data, this conjecture appears to be false. In fact, our data supports a trend that is just the opposite: an older person is more likely to rate a movie as 5 and less likely to rate a movie as 1 than is a younger person.

# **Extending our investigations to histograms**

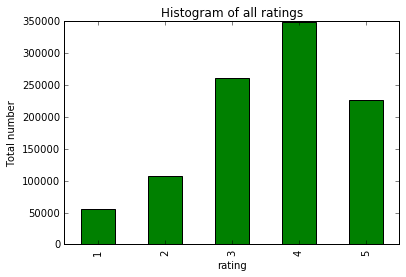
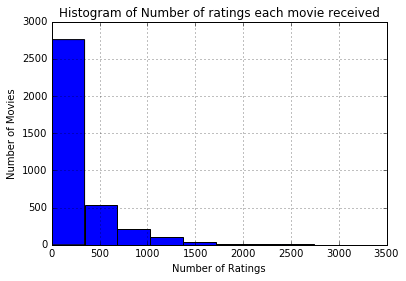
The inferences drawn in the previous objective were largely based on averages and percentages. One shortcoming of such inferences is the ignorance of the *number* of times each movie was rated. Regardless of rating numbers, a movie with 1,000 ratings is surely considered more “popular” than a movie only viewed once. Therefore we expand our analysis to consider how many times various movies were rated, relative to various predictors in our data set. An excellent way to visualize such cumulative data is using histograms.

*Plot a histogram of the ratings of all movies.*

First, we tabulate the *total* number of each type of rating (1-5) submitted in the data set. This can be summarized using a histogram on all movie ratings.

*Plot a histogram of the ratings each movie received.*

On the other hand, we consider the per-movie *quantity* of ratings: how many ratings each movie received in the data set. This is akin to understanding how many times each movie has been *viewed* within the data set. In this case, the horizontal axis represents the number of ratings received by a movie, and the vertical axis represents the total number of movies with that number of ratings.

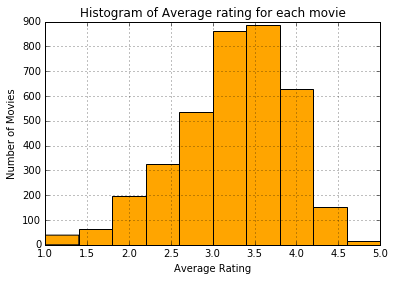
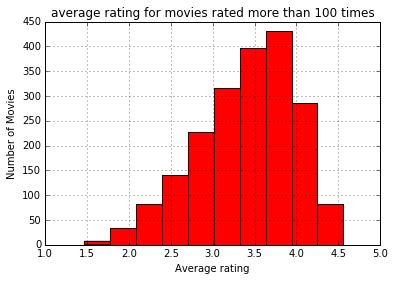
 

*Plot a Histogram of the average rating for each movie.*

In addition to knowing the *quantity* of ratings for each movie, we are also interested in the *quality* of those ratings. Therefore we calculate the average overall rating for each movie, displayed as a histogram as shown below. The horizontal axis now represents the average rating, while the vertical axis represents the number of movies with that average rating.

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

Next, we combine the two previous measures, and consider the *quality* of ratings in terms of *quantity*. In particular, we apply a fixed quantification limit− in this case 100 ratings−eliminating movies with few numbers of ratings from analysis, and then once again consider the average rating of each movie. A histogram summarizing this data can be found below on right side. The reasoning for this quantity cut-off is explained by the following two questions.

** **

*What do you observe about the tails of the histograms?*

The tails of the histograms are noticeably different. When considering all movies, there were a number of titles with average ratings between 1 and 1.5, as well as between 4.5 and 5. When we restrict to movies with at least 100 ratings, all such extreme averages disappear−all averages range between 1.5 and 4.5. This makes sense: if a movie is only rated once, and given a 1, then it has an average rating of 1. On the other hand, when a movie has been viewed many times, in order to achieve an average rating of 1 it needs to receive a very large percent of 1 ratings. For this reason, when we restrict to movies viewed more than 100 times, extreme average ratings disappear.

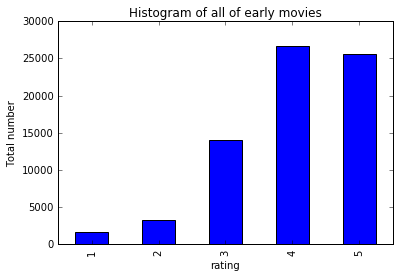
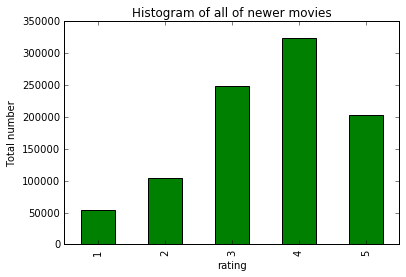
*Which highly rated movies do you trust are actually good?*

We would ultimately like to use the *quality* of ratings to determine the *quality* of a movie. If a movie has been viewed a single time, with a rating of 5, this movie now has an average rating of 5. This is an extremely high “quality” average rating. However the lack of *quantity* in ratings makes this average deceiving: it really only indicates that exactly one person enjoyed this movie. On the other hand, if a movie has over 100 views and an average rating of 4.5, in order to achieve a high average *many* people rated this movie highly. Therefore it is much more likely that this is a good movie. Thus a high quality rating, in the presence of a high *quantity* of ratings, is more indicative of a high quality *movie*. That is, among highly rated movies, we trust that those rated more than 100 times are actually good movies. Herein lies the motivation for restricting our data to at least 100 ratings.

***Conjecture 1: People prefer to watch newer movies than older ones.***

Newer movies are likely being watched by audiences looking to be entertained. They watch these films with a critical eye, analyzing their level of enjoyment. This should lead to a distribution that appears normal. Audiences watching much older movies (i.e. movies made before 1960) likely have some nostalgia tied up in these older films. This will lead to unbalanced distributions.

**Data:** Although the release year of each movie was not a predictor in the MovieLens data set, it was included within the title for each film. We extracted the year from each title, and classified the movies by year of release. The average release year among these films was 1960. Therefore we separated the movies into “older” and “newer” films, based on pre- or post- 1960 release. We then calculated the total number of ratings of each type (1-5) for these two sets. This is summarized in the following two histograms.

Once again, we frame this question in terms of *distributions* as there is an inherent gap in quantity of ratings: newer movies have been rated much more frequently. Nevertheless, we can see there is a distinct difference between these two distributions. Among older movies, the total number of 4 and 5 ratings is nearly equal. This total dominates the low (1 and 2) ratings, and is approximately twice the number of 3 ratings. This leads to an unbalanced distribution accumulating towards the high-end. The distribution of total ratings among newer movies, however, is much closer to a normal one.

*Conclusion:* People prefer to watch newer movies more but at the same time they watch good old movies as well. This has been supported by our data, as seen in the histograms above.

# **Correlation-Men vs Women**

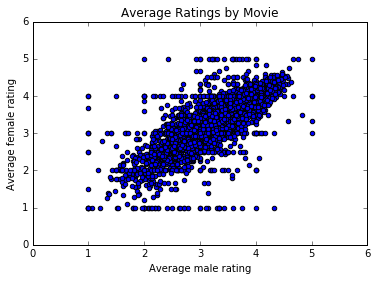
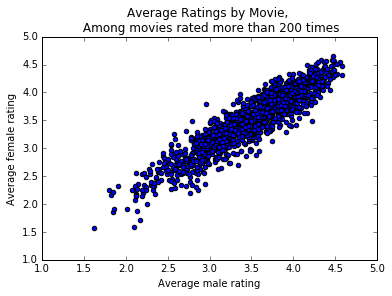
After exploring many aspects of the data concerning movies and their ratings, we will analyze trends among those *submitting* the ratings - users. We are interested in determining when the ratings among one group of users may be indicative of those from another group. In particular, we will compare the ratings of men versus the ratings of women.

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

First, we compute (via pivot table) the average rating among men, and among women, for each movie title in the data set.

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

As seen below, mean ratings of movies with few total ratings can be extremely misleading, and are not an accurate measure of the perceived quality of a movie. Therefore we restrict our scatter plot to only those movies rated at least 200 times.

*Compute the correlation coefficient between the ratings of men and women.*

As seen in above plots, there seems to be an overall linear dependence between the average ratings of men, and the average ratings of women, particularly in those movies with more than 200 views. Given this perceived linear relationship, we therefore calculate the correlation coefficient between the mean ratings of men and the mean ratings of women.



*What do you observe?*

These correlation coefficients are both relatively high. A correlation coefficient closer to one indicates a stronger linear relationship between the data being tested. Therefore, there is a reasonable linear relationship between the average male rating and average female rating among all movies, and a rather strong linear relationship among movies with at least 200 ratings.

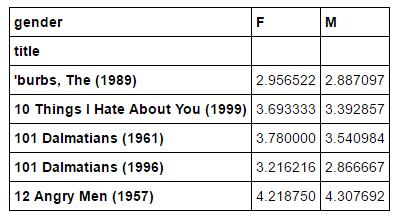
*Are the ratings similar or not?*

This data presented above is somewhat misleading. First, a high (0.92) correlation indicates that average men’s rating is predictable in a linear fashion from the average women’s rating. A high correlation indicates linear predictability, but it is the nature of the trend that indicates-this linear relationship is likely equality. More-over, based on the high correlation values, one may predict that ratings between men and women are similar. Ho-wever, this correlation is for the mean rating per title between men and women. That is, on average men and women rate movies in a manner that is linearly related. This doesn't indicate that the ratings themselves are similar For example, there could be a movie for which both men and women have an average rating of 3, where women always rate as 1 or 5 and all men rate it as 3.

***Conjecture 1: When they are tired Male and female rate more similarly***

Male and female tend to give similar rating to a movie when they are tired. Considering the fact that at the end of the day before sleep or during wee hour’s human mind is tired. So there should be collinearity between ratings given by male and female during this time.

**Data:** We filtered out the data on timestamp basis which was submitted during the time 10PM to 5AM. In this data we determined the average ratings per movie title by gender. Then to identify similarity in ratings we computed the correlation coefficient.





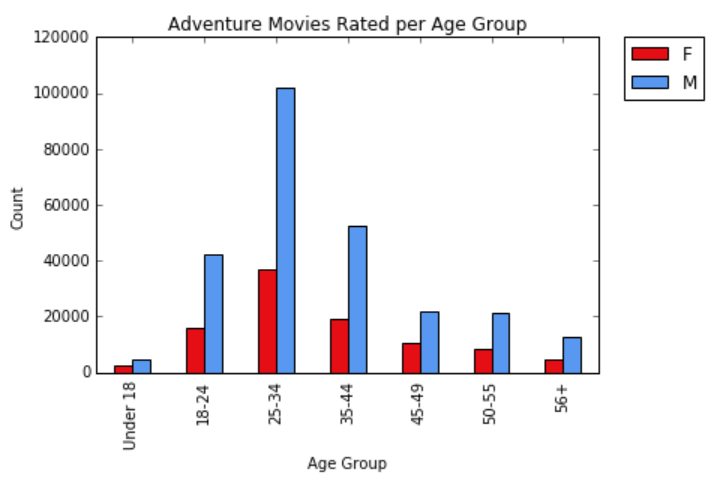
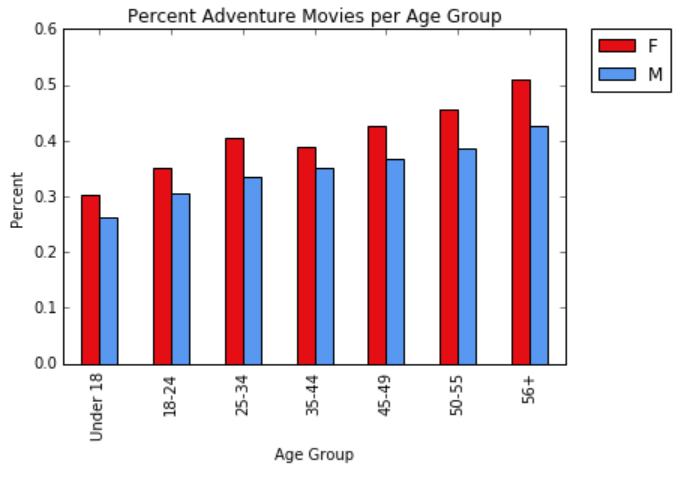
*Conclusion:* We can see there is a very high correlation coefficient close to 0.87. Hence, it is significant that male and female tend to give similar rating in between 10AM and 5PM.

# **Business Intelligence**

So as discussed in the introduction section, we tried to identify a solution to a million dollar question as to What genre of Movie should Netflix recommend to a first-time user based on their log-in information? All the data analysis done so far is pretty much helpful and gives us some idea about identifying a solution. We focus our analysis on the two most “active” genres - Comedy and Adventure and propose two specific Business Questions.

**Data:** We filter out movies with genre as Comedy or Adventure. And then plot a histogram of ages of both male and female users and their ratings given to Adventure movies.

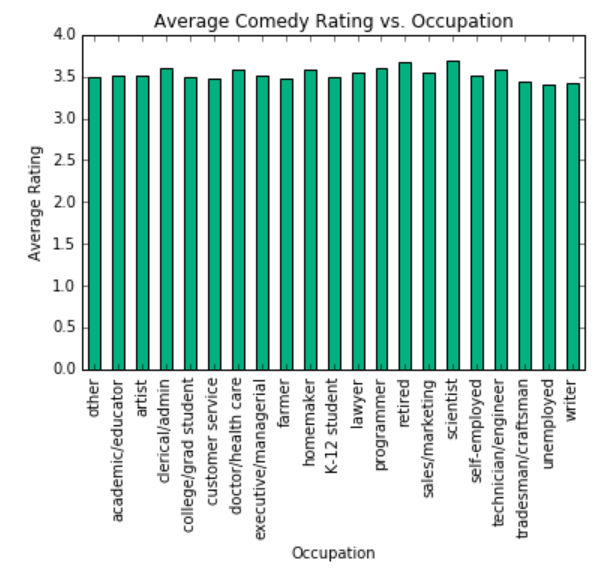
**Which gender and age group watches the most adventure movies?**

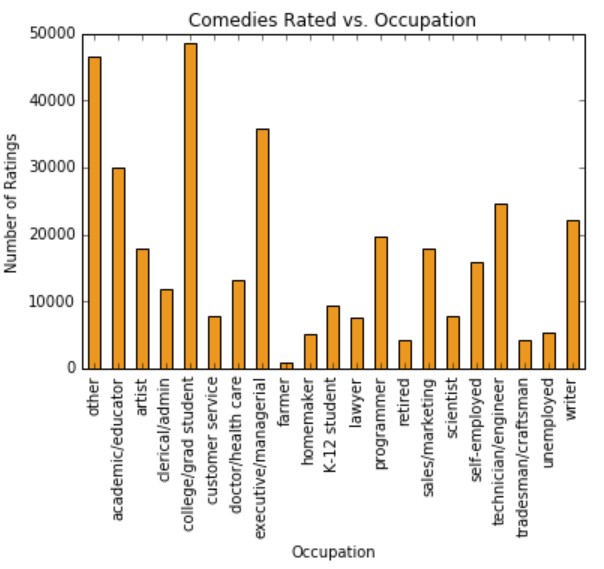
 

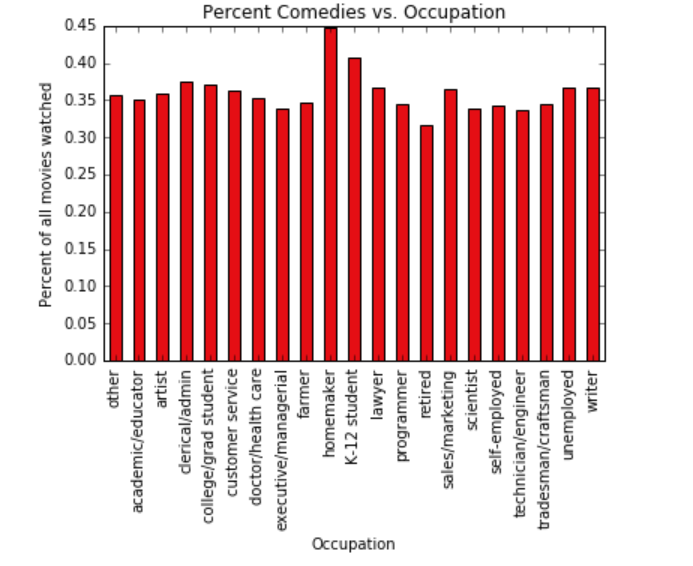
Conclusion:-

1. Women show a particularly strong interest in Adventure movies
2. Nearly half of movies watched by older women were adventure movies
3. Older people are more interested in adventure movies than younger individuals

**Which occupation is most likely to enjoy a comedy?**







Conclusion: College students, homemakers, and children all show high interest in comedies compared to other occupations

# **Data Limitations**

While the *MovieLens 1M Dataset* was sufficient to begin answering the questions presented in this report, we recognize that this data has its limitations.

Lack of latest data - All movies rated in this data set were released *before* the year 2,000. Since that time, hundreds of movies generating billions in revenue have been produced. A more up-to-date data set would be advantageous.

Biased Data - The *MovieLens 1M* data set provided a wide breadth of films (ranging from 1920-2000) and, moreover, included a complete set of demographic information for each user. Although the quantity of the data collected was appropriate for this study, we recognize that the rating-responses carry significant bias. Most of this bias arises due to under-representation, as users resided within the United States, and only about 6,000 users’ submitted ratings. Additional sampling bias arises from unbalanced gender-sampling, with a 3:1 male to female ratio among the users.

# **Conclusions**

In this paper, we began to explore the difficult-yet-relevant problem of making individual movie recommendations based solely on background information. First we retrieved the *MovieLens 1M Ratings* Data set, and performed some basic analysis. After examining “popular” movies based on submitted ratings, we found that our data suggested that older people are easier to please than younger folks: the oldest sample users gave the *highest* percentage of 5 ratings and *lowest* percentage of 1 ratings among all ratings they submitted. Next we extended our analysis to histograms to consider cumulative data. After generating summary-based histograms, we found that people prefer to watch newer movies more but at the same time they watch good old movies as well. Hence *older* movies receive disproportionately more high-end ratings than do newer movies, which we postulate is due to a nostalgia factor. Next we considered the ratings of men versus those of women, and particularly when the two were predictable. Using a combination of correlation and linear regression, we determined that ratings between genders could be predicted among highly-viewed movies made recently, and comedies.

Lastly, we extended our recommendation problem to business intelligence. In particular, among “active genres,” which genre of movie should Netflix recommend to a first-time user? By researching on these active genres (Comedy, and Adventure), we concluded –

1. Women show a particularly strong interest in Adventure movies
2. Nearly half of movies watched by older women were adventure movies
3. Older people are more interested in adventure movies than younger individuals
4. College students, homemakers, and children all show high interest in comedies compared to other occupations

Thus we have successfully completed case study on Movie Lens data by analyzing 1M datasets.