CASE STUDY 2

Jinal Jain jjjain@wpi.edu
Manasee Godsay mrgodsay@wpi.edu
Mihir Sawant msawant@wpi.edu
Rushikesh Naidu ranaidu@wpi.edu
Umesh Nair uunair@wpi.edu

1. INTRODUCTION

Many movie companies, like Netflix, nowadays solely don't make a film or show based on the plot or the thought that "subscribers might like it". The decisions are usually based on a number of factors and almost entirely on data. It is essential to analyze the demographics of the population and the ratings given to every movie across different genres for a company to decide what type of movies to produce and make.

In Case Study 2 we are looking at the MovieLens 1M Data Set. The goal of this case study is to understand and analyze some basic details of the data and make conjectures. Begins with downloading the 1M Movie Lens dataset from http://grouplens.org/datasets/movielens/ and performing analysis in Python. The downloaded data file has three separate data files: "Ratings", "Movies" and "Users" which are merged into a single Pandas DataFrame. In order to understand how HDF5 format is used to store data, the entire dataset was stored into an HDF5 file. Then the single data file undergoes analysis where several questions are answered. In the first part number of movies that have an average rating over 4.5 overall, among men and among women, number of movies that have a median rating over 4.5 among men and women over age 30, ten most popular movies and conjectures about how easy are various groups are to please are reported. In the second part visualization by plotting histograms for ratings of all movies, number of ratings and average rating for each movie, average rating for movies that are rated more than hundred times and conjectures about the distribution of the ratings were reported. In the third part scatter plots for men versus women and their mean ratings for every movie and movies rated more than 200 times, correlation coefficients between the ratings of men and women and Conjectures about under what circumstances the rating given by one gender can be used to predict the rating given by the other gender. And finally, business question is where the conjectures in the first three parts of the analysis are used to come up with insights that a movie company might be interested in.

2. ANALYSIS & OBSERVATIONS

After downloading the MovieLens 1M Data Set, it was combined together into a single Pandas DataFrame. The following analyses were performed and observations were obtained:

I. BASIC ANALYSIS

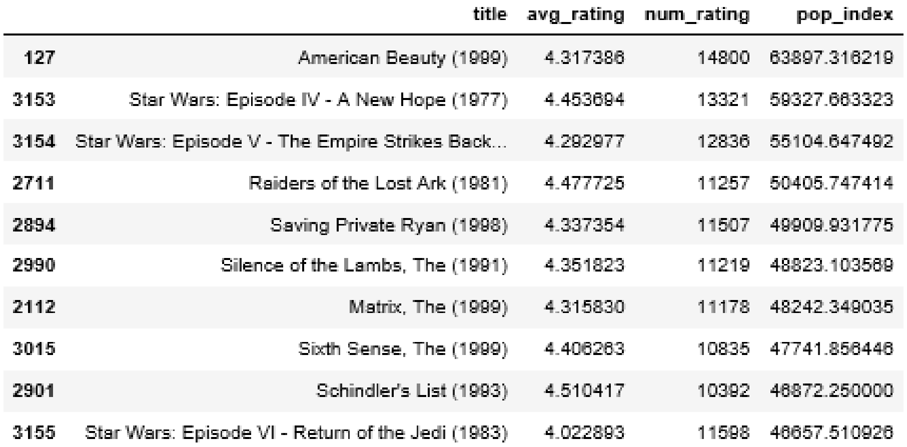
1. Number of movies having an average rating over 4.5 overall:
There are 29 movies in the dataset with an average rating over 4.5 overall.

2. Number of movies having an average rating over 4.5
(i) Among men: 23 movies
(ii) Among women: 51 movies

3. Number of movies having a median rating over 4.5
(i) Among men over age 30: 86 movies
(ii) Among women over age 30: 149 movies

4. Defining Popularity: The popularity is calculated by considering the average of ratings for a movie multiplied by the number of ratings. A movie with higher sum of ratings is considered to be more popular than the one with lower sum of ratings.

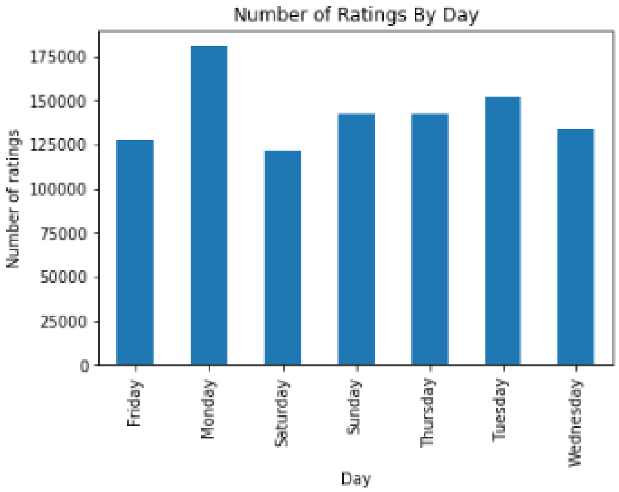
Based on the above definition the 10 most popular movies are:



	title	avg_rating	num_rating	pop_index
127	American Beauty (1999)	4.317386	14800	63897.316219
3153	Star Wars: Episode IV - A New Hope (1977)	4.453694	13321	59327.663323
3154	Star Wars: Episode V - The Empire Strikes Back...	4.292977	12836	55104.647492
2711	Raiders of the Lost Ark (1981)	4.477725	11257	50405.747414
2894	Saving Private Ryan (1998)	4.337354	11507	49909.931775
2990	Silence of the Lambs, The (1991)	4.351823	11219	48823.103569
2112	Matrix, The (1999)	4.315830	11178	48242.349035
3015	Sixth Sense, The (1999)	4.406263	10835	47741.856446
2901	Schindler's List (1993)	4.510417	10392	46872.250000
3155	Star Wars: Episode VI - Return of the Jedi (1983)	4.022893	11598	46657.510926

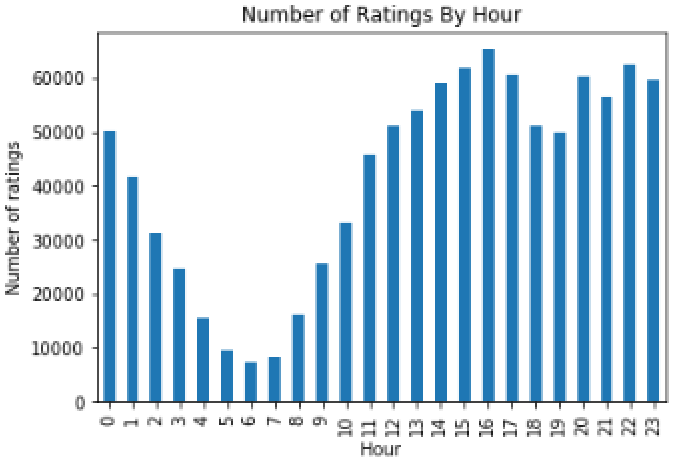
Based on our logic and definition, “American Beauty” is more popular than “Star Wars: Episode IV – A New Hope” because even though the average rating for the latter is more but the number of ratings for American Beauty is higher and hence, it lies above in the list.

5. Conjecture 1: Saturday and Sunday has the largest user traffic
To test the conjecture, plotted a histogram of Number of Ratings by Day, with day on the horizontal axis and number of ratings on the vertical axis.



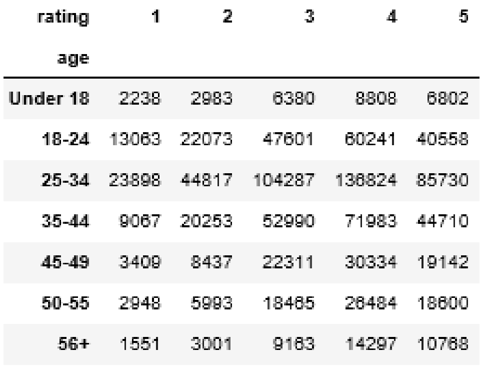
FALSE: From the above histogram, Monday and Tuesday gets the highest number of ratings as compared to Saturday and Sunday. This indicates that the user traffic is more on weekdays than on weekends.

Conjecture 2: Most users view movies at night.
To test the conjecture, plotted a histogram of Number of Ratings by Hour, with day on the horizontal axis and number of ratings on the vertical axis.



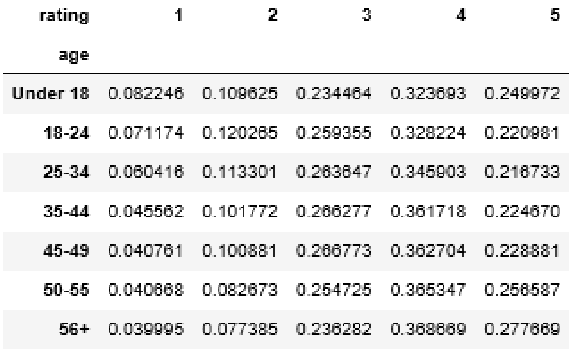
FALSE: We assume that users rate movies immediately after watching them. Based on that we can see that users have rated maximum between 3pm & 5pm and 10pm & 11pm. This indicates that users watch movies between lunch and dinner hours.

Conjecture 3: Users over the age of 50 are easier to please
To test the conjecture, we first calculated the frequency of the ratings across each age group.



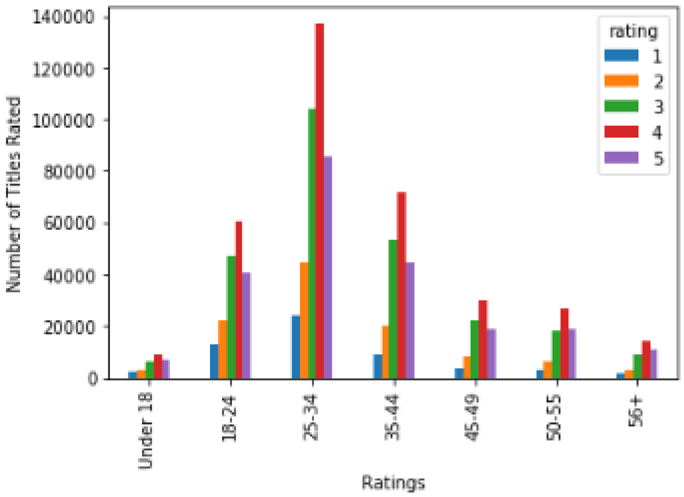
rating	1	2	3	4	5
age					
Under 18	2238	2983	6380	8808	6802
18-24	13063	22073	47601	60241	40558
25-34	23898	44817	104287	136824	85730
35-44	9067	20253	52990	71983	44710
45-49	3409	8437	22311	30334	19142
50-55	2948	5993	18465	26484	18600
56+	1551	3001	9163	14297	10768

Then normalized those values



rating	1	2	3	4	5
age					
Under 18	0.082246	0.109625	0.234464	0.323693	0.249972
18-24	0.071174	0.120265	0.259355	0.328224	0.220981
25-34	0.060416	0.113301	0.263647	0.345903	0.216733
35-44	0.045562	0.101772	0.266277	0.361718	0.224670
45-49	0.040761	0.100881	0.266773	0.362704	0.228881
50-55	0.040668	0.082673	0.254725	0.365347	0.256587
56+	0.039995	0.077385	0.236282	0.368669	0.277669

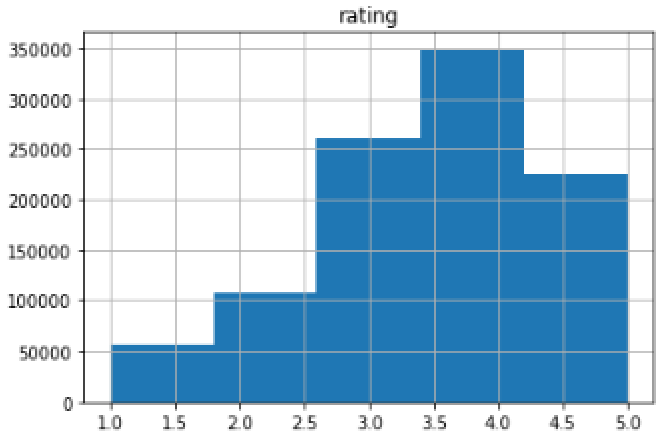
Plotted a histogram to represent the distribution of normalized ratings across each age group



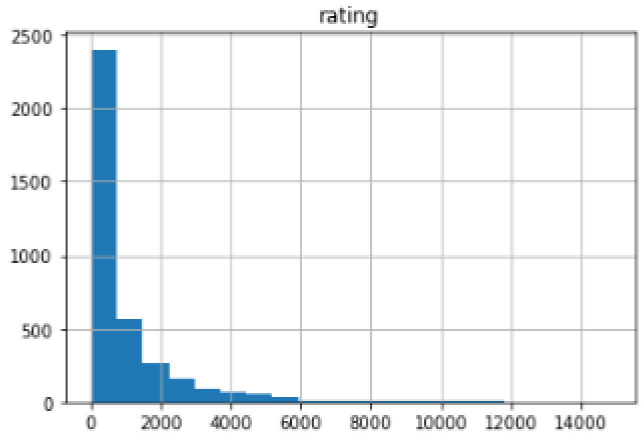
TRUE: The percentage of ratings of 4 and 5 given by the users older than the age of 50 is higher than any other age group.

II. EXPANDING BASIC ANALYSIS TO HISTOGRAMS

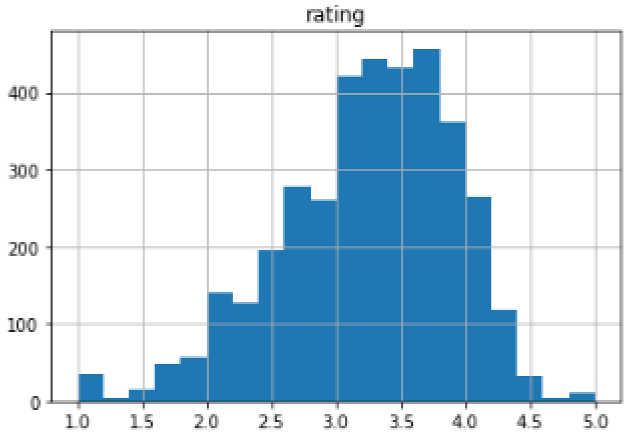
1. Plotting a histogram of the ratings of all movies



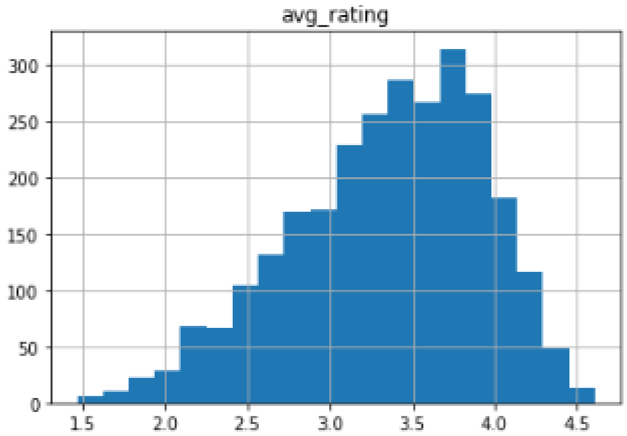
2. Plotting a histogram of the number of ratings each movie received



3. Plotting a histogram of the average rating for each movie



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



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?

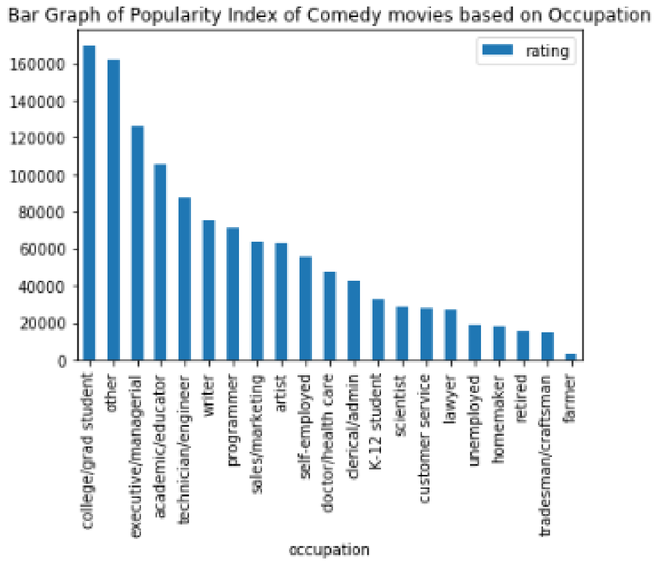
The tails of the histogram with all the movies are more populated than the histograms with movie which got 100+ ratings. This is because of sample bias and the ratings tend to spread out with a larger sample size.

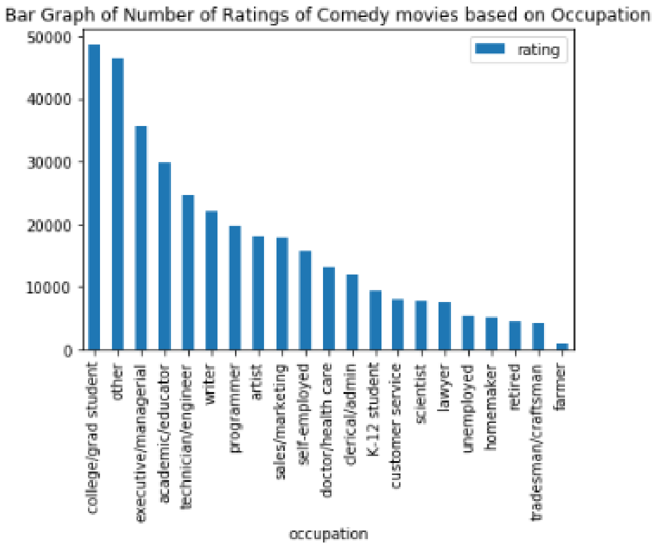
Which highly rated movies would you trust are actually good? Those rated more than 100 times or those rated less than 100 times?

The ones rated more than 100 times are most trustworthy. This is because the bigger the sample size the better the representation

5. Conjecture 1: Scientists like Comedy Movies

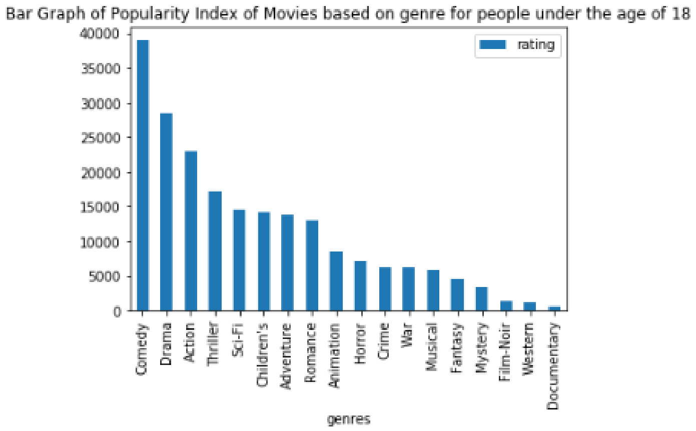
To test this, plotted a bar graph of popularity index and Number of Ratings of Comedy Movies based on Occupation

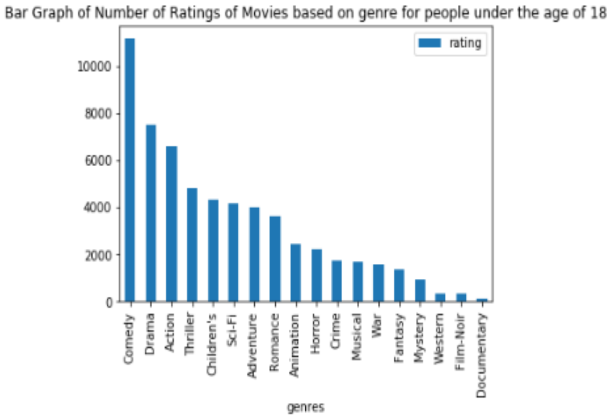

Bar Graph of Popularity Index of Comedy movies based on Occupation


Bar Graph of Number of Ratings of Comedy movies based on Occupation

TRUE: Even though total number of ratings are less for scientists but the average rating is highest.

Conjecture 2: Users under the age of 18 prefer "children's", "animation" and "sci-fi" movies
To test this, plotted a bar graph of popularity index and Number of Ratings of Comedy Movies based on Occupation

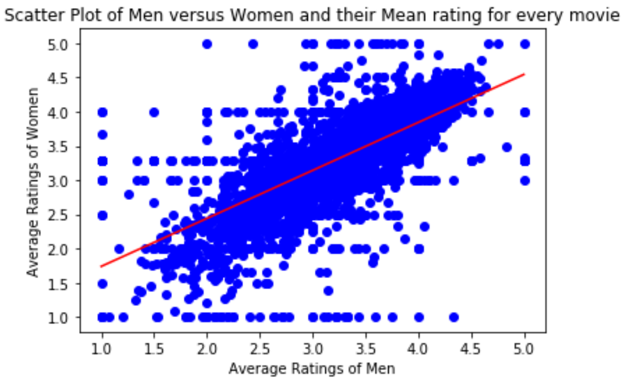

Bar Graph of Popularity Index of Movies based on genre for people under the age of 18


Bar Graph of Number of Ratings of Movies based on genre for people under the age of 18

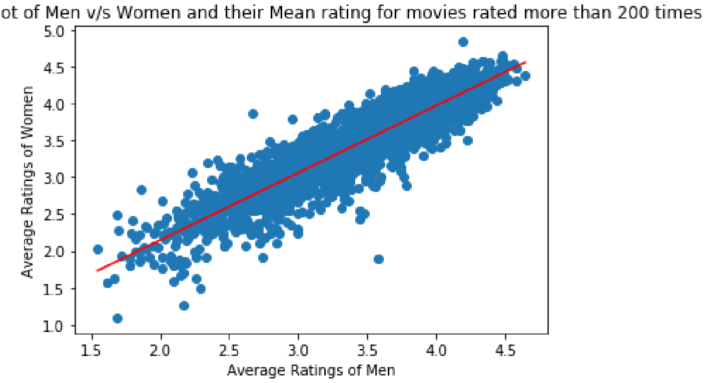
TRUE: Because even though total number of ratings are more for children's movies but the average rating is more for sci-fi movies

III. CORRELATION: MEN VERSUS WOMEN

1. Men versus Women and their mean rating for every movie


Scatter Plot of Men versus Women and their Mean rating for every movie

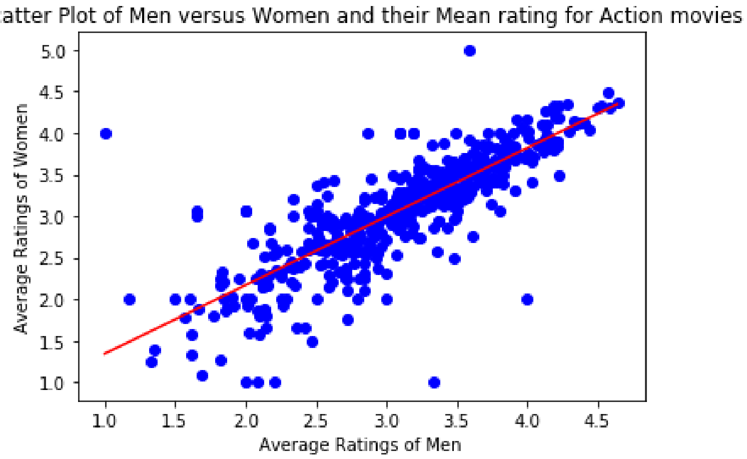
2. Men versus Women and their mean rating for movies rated more than 200 times



3. Correlation between the ratings of men and women

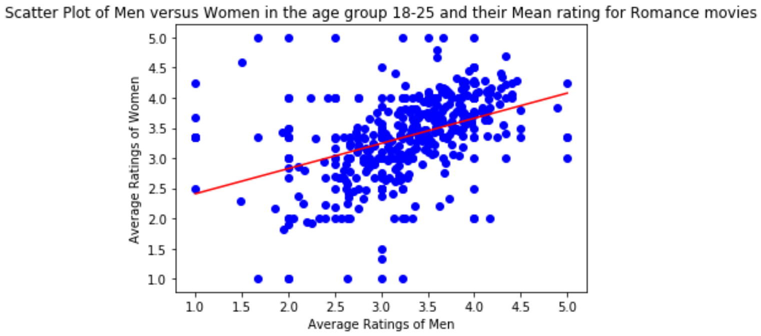
We observe that the correlation coefficient between the ratings of men and women is 0.691705 which means that there is a 69% correlation between them, which is pretty significant.

4. Conjecture 1: We can predict the ratings of Action movies given by women looking at the ratings given by men



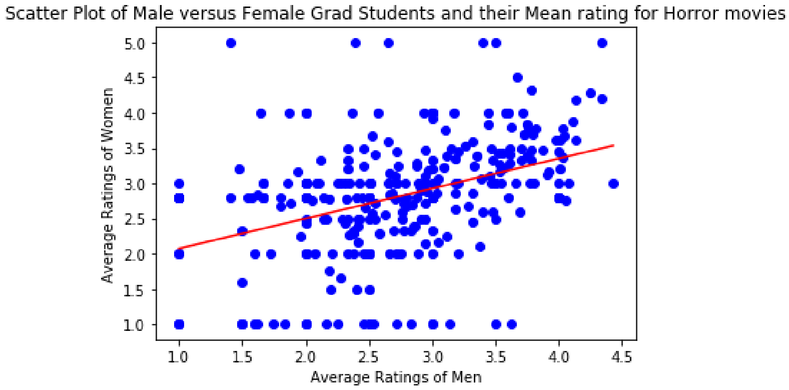
TRUE: The correlation coefficient is 0.82 and so we can predict the ratings of Action movies given by women looking at the ratings given by men. Looking at the ratings given by men, the test error in predicting the ratings for women is 16.3%. Men and Women have pretty similar preferences when it comes to Action movies.

Conjecture 2: We can predict ratings given by women in the age group 18-25 for Romance movie looking at the men's ratings



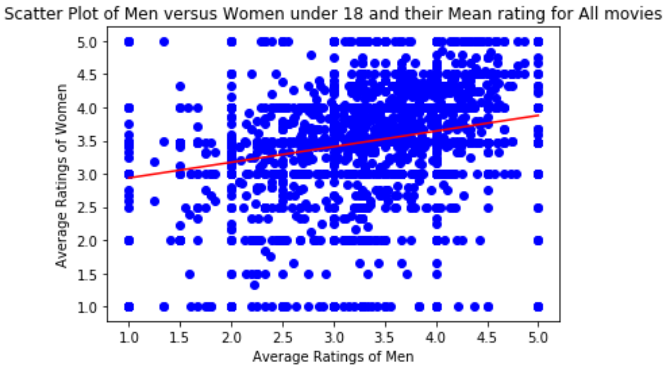
FALSE: The correlation coefficient is 0.414689 for the ratings of Romance movies given by women in age group 18-25 and the ratings given by men in the same age group. This means that if we predict women’s ratings from men’s rating we have a 42% chance of getting the prediction wrong since we have observed a 42% test error from the data.

Conjecture 3: We can predict a female grad student rating for a horror movie based on a male grad student



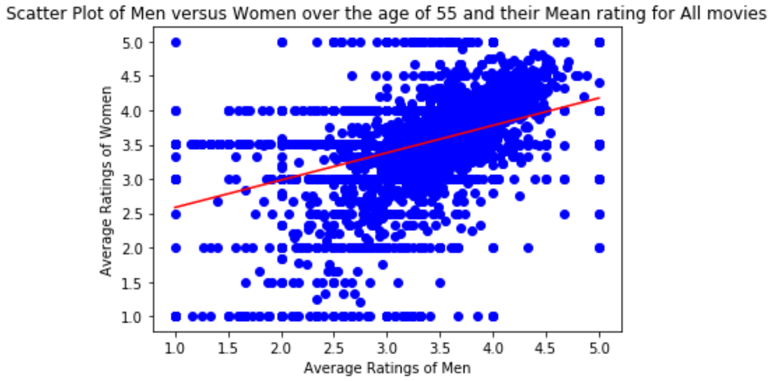
FALSE: The correlation coefficient is 0.40747 for the mean ratings of Horror movies given by female grad student and the mean ratings given by male grad students. This means that if we predict female college/grad student’s ratings from male college/grad student’s rating we have a 50.5% chance of getting the prediction wrong since we have observed a 50.5% test error from the data.

Conjecture 4: Men and women have similar preferences when they're younger



FALSE: The correlation coefficient is 0.2454 for the mean ratings of movies given by women under 18 and the mean ratings given by men under 18. The correlation coefficient is too low and we cannot conclude that men and women have similar preferences when they’re younger.

Conjecture 5: Men and women have similar preferences when they're older



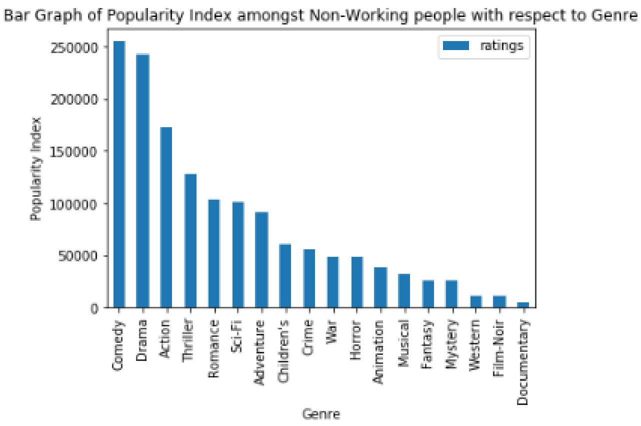
FALSE: The correlation coefficient is 0.4287 for the mean ratings of movies given by older women and the mean ratings given by older men. The correlation coefficient is average and we cannot conclude that men and women have similar preferences when they're older.

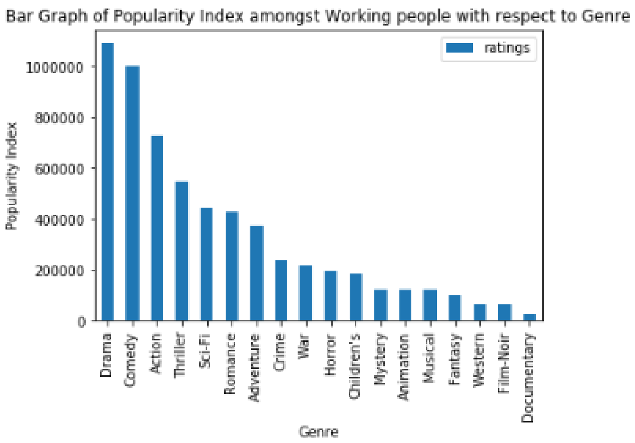
3. BUSINESS INTELLIGENCE

For online movie companies, like Amazon Video who acquire new customers every day, requires a recommender system solely based on their registration and without any information about previous ratings or browsing history. There are many facets to build such a system but we particularly want to focus on: "What genre of movies should Amazon Video suggest to a new user?" But this is a broad question and to make it specific we divided the question amongst "Working" and "Non-working" users. Non-working means that the users who have zero working hours and the remaining are classified as Working. In our data we considered "homemaker", "college/gradstudent", "retired", "unemployed" and "K-12 student" as Non-working and the remaining as Working. Based on this we asked the following specific questions:

1. What kind of movies is watched by Working and Non-working users?

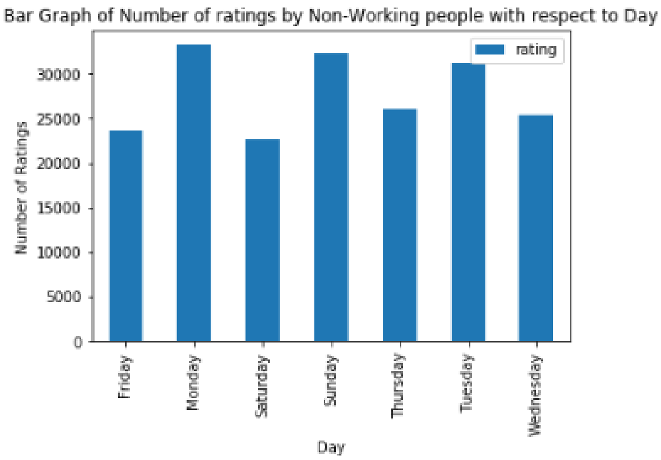
To test this, plotted a bar graph of popularity index amongst Non-working and Working based on genre

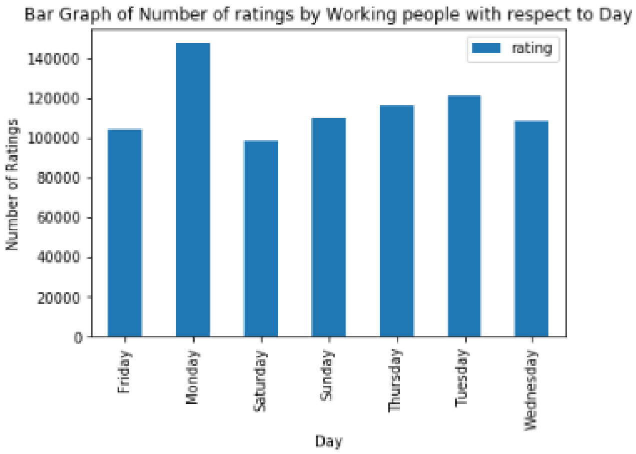




Based on the Popularity Index graphs for both we can see that Comedy and Drama are watched by Working and Non-working users. More Drama and Comedy movies should be suggested to them.

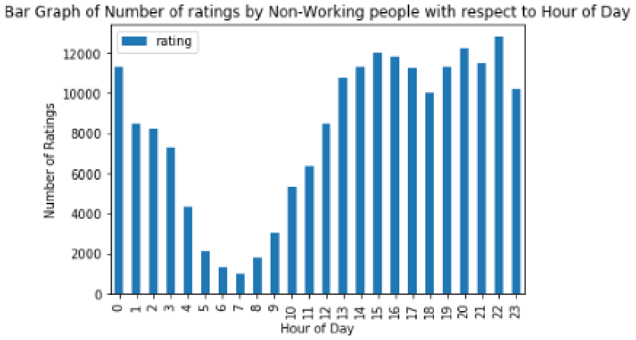
2. What is the best day of the week to recommend movies to Working and Non-Working users?

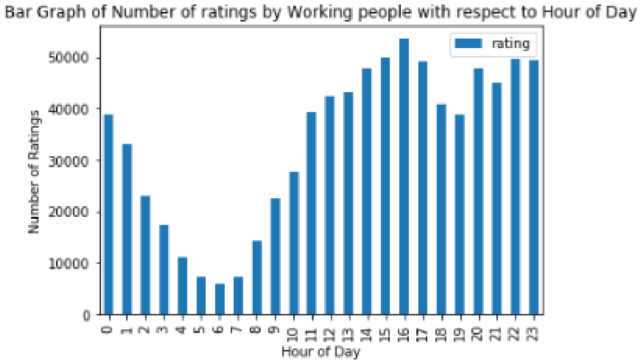




Based on the number of ratings per day we can see that for Non-working users the distribution is almost uniform with Sunday, Monday and Tuesday as the modal bars and hence, we can recommend them movies more on Sunday, Monday and Tuesday. But for Working users there is one modal bar, which is for Monday and hence, Monday is the best day of the week to recommend movies to them.

3. What is the best time of the day to recommend movies to Working and Non-Working users?





Based on the number of ratings for every Hour of the Day we can see that for non-working users 8pm to 10pm is the best time to recommend movies. For working users, 3pm to 5pm is the best time to recommend movies.