W.P.I. D.S.501; Introduction to Datascience

To: Dr R. Paffenroth, Dr X. Kong

From: Matt Curcio

Due Date: December, 2014

Subject: Case Study #2

Using a publicly available movie database, MovieLens, The will be described by way of descriptive statistics. After the description and then extrapolate and expand upon this information to provide movie genres which could be geared toward particular audiences.

This case study uses a publicly available movie database, MovieLens; University of Michigan, to find trends among the participants of the database and then expand upon the findings to make a case for the development of specific movie types or genres. The database was complied in 2000. It is a database with movie ratings from 6,040 users containing information on approximately 3,900 movies producing a total of over a million ratings.

The majority of the analysis will be using basic descriptive statistics. Through out this analysis I intend to point out what users watch and what types of movies they prefer.

The first analysis was to investigate the ages and what proportions tend to watch movies and possibly their habits. Figure #1 show the percentage of Americans that consider themselves movie-goers. By far the largest single proportion is the group of Americans between the age of 25-34. However the second and third largest groups are near to this major age group. If the three regions between 18 and 44 are combined into one super group, this would represent over 75% (77.7%)of the movie going population. I will hence forth call this super group the 'Young Americans.'

Another grouping that I would like to point out is the older set of Americans in the age range of 45-49, 50-55, and 56+. Combined I would like to refer to this group as the 'Baby Boomers.' I would like you to keep this grouping in mind for further work done later. This Baby Boomers consist of nearly one out of every 5 Americans, (19.5%). This may become important since this group generally has a larger percentage of disposable income.

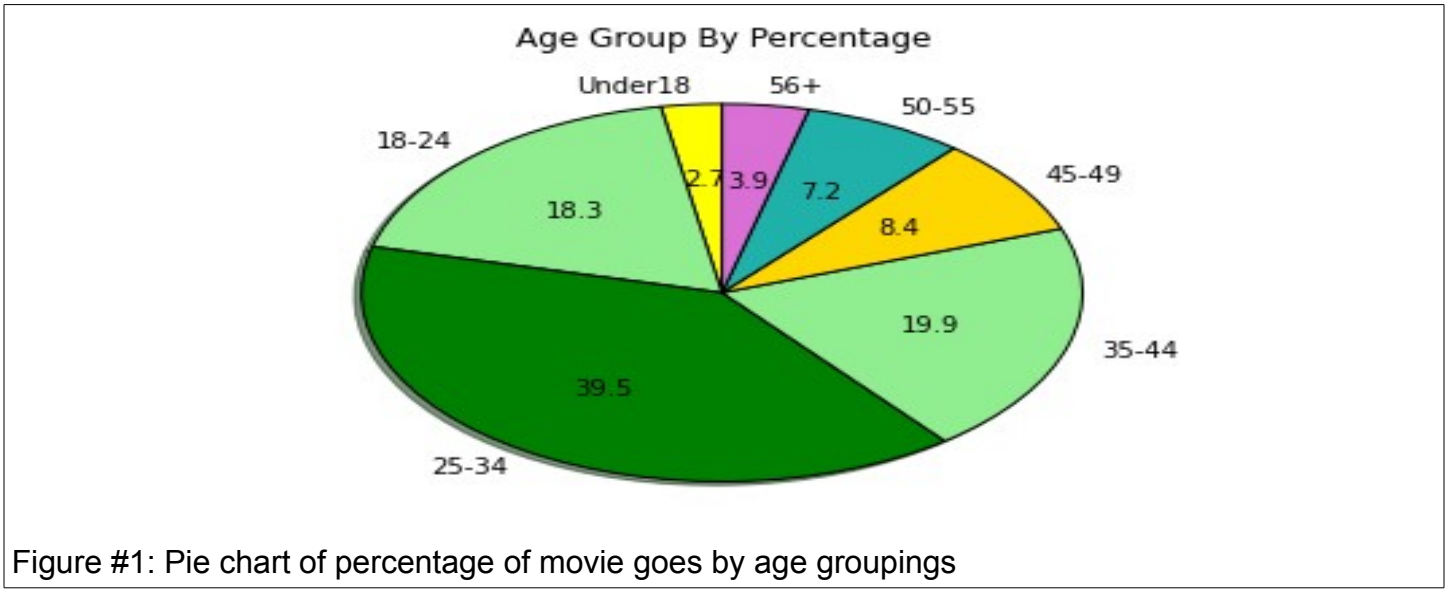

Figure #1: Pie chart of percentage of movie goes by age groupings

The next question that was asked was, 'How does each of these age ranges rate the movies they see?' It is interesting to note that the lowest ratings were given by the 18-24 age group, a part of the Young Americans group. Conversely as the movie going population ages they tend to rate movies more favorably. So much so that the three highest ratings were the averages produced by the Baby Boomers. The Baby Boomers overall rate the movies they see with an overall 3.71 star rating. More investigation would be needed to determine why this occurs but it does seem that this older group rates movies 6% higher than the 18-24 age range. See figure #2.

Percent Change of Baby Boomer rating to 18-24 age ratings = $\frac{3.722 - 3.5}{3.5} x\,100\% = 6.3\%$

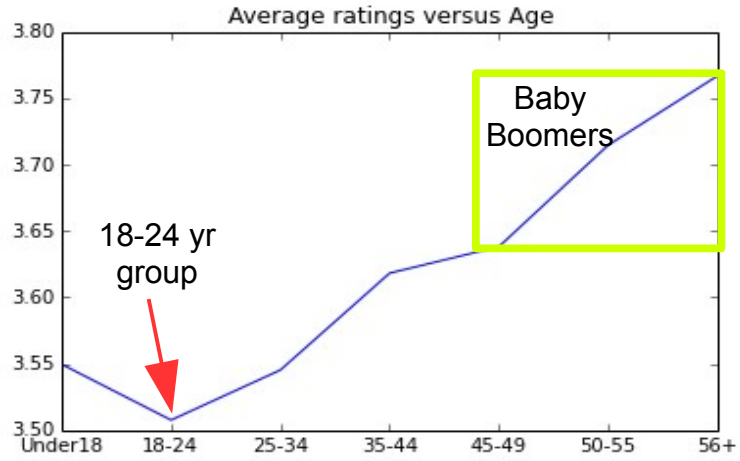

Figure #2: Average ratings given to movies by age groupings

The idea that Baby Boomers tend to rate the movies they see more highly is also born out by the boxplots of the population occupations. In the survey, there were 20 categories of occupations. Among the 20 occupations is a 'retired' category. Although any person in the survey population could have answered that they were retired this category is predominately used by Americans between the ages of 56 plus. Here again it points to the idea that older movie goers are more apt to rate movies higher. See figure #3.

In contrast with the boxplots of almost all of the other categories were had the inter-quartile range (IQR) much lower with the means being closer to 3 with the top of the third quartile reaching less than 4.0.

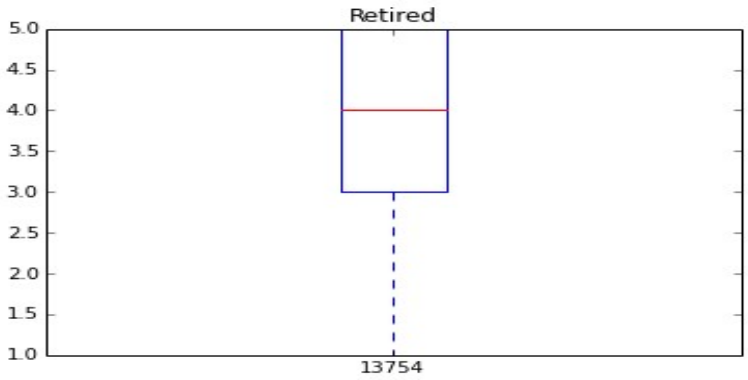

Figure #3: Boxplot of movie ratings by occupation, retired.

There was one oddity that was observed when investigating the occupations of Americans and their overall ratings. The boxplot of one occupation stood out similarly to the retired occupational category and it was writers. See figure #4 . It appears writers are more apt to rate the movies they see more favorably than the general population. More research would be needed to pin down the reasoning for this but I will speculate. I would speculate that writers are more to give higher ratings because they may appreciate the time and effort that went into writing involved in the movie. Another explanation could be that we have a very small number of writers in our pool of users and thus movie rates overall.

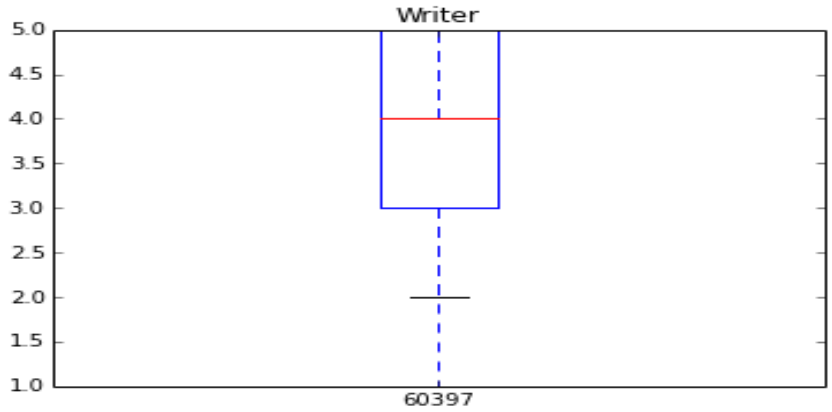

Figure #4: Boxplot of movie ratings by occupation, writer.

Now I will discuss movie attributes that may be most interesting to viewers. There was an analysis of the terms which the largest population grouping call their 'likes.' Since it was determined that the largest movie going group was the 'Young Americans,' we wanted to find out what movie genres that appeal to them. The top four movie genre were comedy (17%), drama (16%), action (13%), and thriller (9%). See figure #5.

What is interesting is that many of the terms were found to overlap with each other. In other words, movies cannot be classified into one genre. Movies have become sophisticated and mix many features together. It is now common to hear of movies described as 'dramodies', meaning the combination of a drama and a comedy.

I would propose that any new movie would be an amalgam of these terms also. This analysis suggests that finding movies with descriptors in multiple genre would be the most profitable. Pairing interesting leading men and women together in a comedy format that requires some action. One move that came up during this research was 'Romancing The Stone.' An action film that had as its main characters fall in love in the midst of the thrills that they undergo, the travails of being chased and escaping unfriendly antagonists. This movie in particular had its comic moments as well. While being chased through out the perilous jungle they interject comic relief.

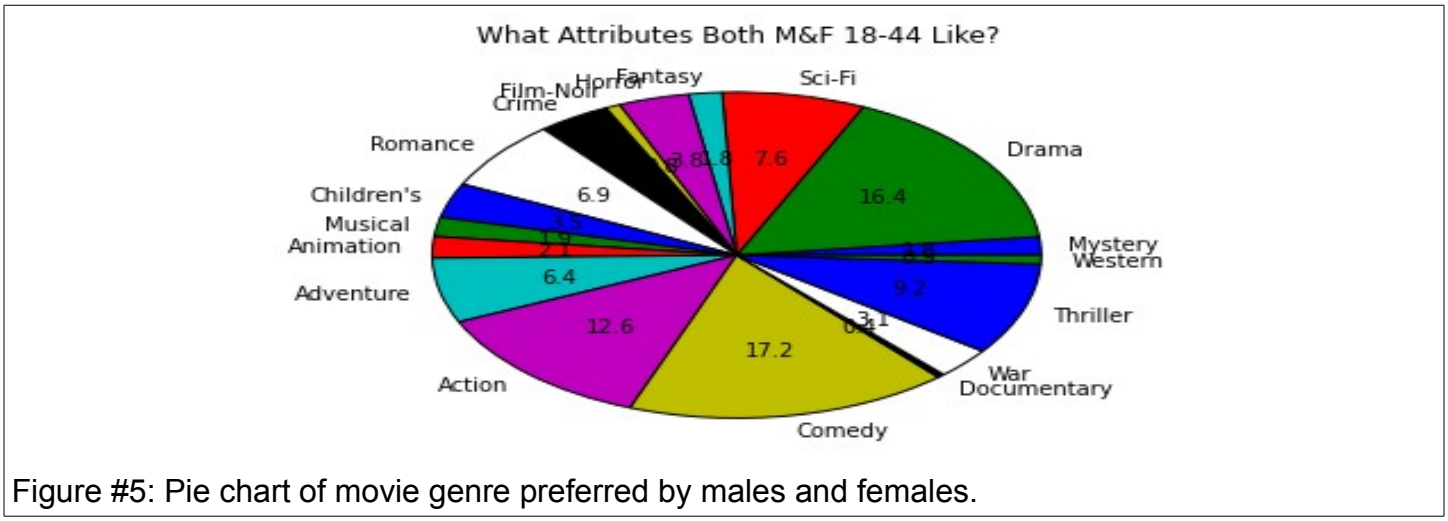

Figure #5: Pie chart of movie genre preferred by males and females.

Two additional plots were made to further investigate occupation versus ratings as well as occupation versus standard deviations of the ratings provided. The plot of occupation versus occupation was slightly different than the boxplots shown earlier because the boxplots show more information about the spread of the data for each occupation. Information from the boxplot data is more akin to looking at the standard deviations of the occupations. The Standard deviation data provides less overall information than the boxplots which show not only the inter-quartile ranges but also the outliers.

Occupation By Category
1. "other" or not specified 2. "academic/educator" 3. "artist" 4. "clerical/admin" 5. "college/grad student" 6. "customer service" 7. "doctor/health care" 8. "executive/managerial" 9. "farmer" 10. "homemaker" 11. "K-12 student" 12. "lawyer" 13. "programmer" 14. "retired" 15. "sales/marketing" 16. "scientist" 17. "self-employed" 18. "technician/engineer" 19. "tradesman/craftsman" 20. "unemployed" 21. "writer"

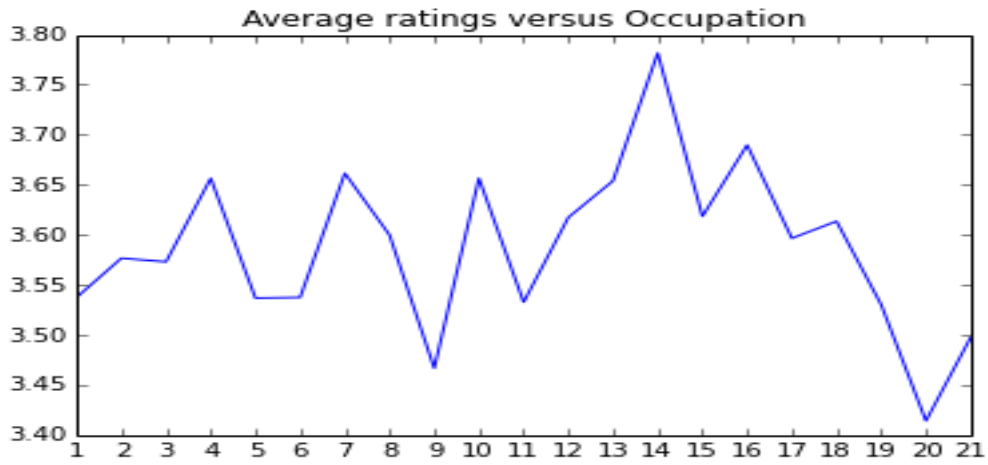

Figure #6: Mean of movie ratings given by occupation, 21 categories.

There are several very unusual points of interest in this plot of job categories. The two lowest points of this graph are ratings from 'homemaker' and 'unemployed.' These two job descriptions unfortunately not unknown for having copious amounts of disposable income, although this may be debatable with the 'homemaker.' Therefore it might provide a rationalization for making movies which do not cater to these groups. As I noted earlier that this might be considered an unusual find because it was commonly thought that movies produced in the depths of the U.S. Depression (ca. 1930's) were frequented by the masses of unemployed for enjoyment and escapism. The reasons for this divergence is not with in the scope of this report because it is multifaceted.

It is good to see corroborating evidence from our previous boxplots of the retired giving hi marks to movies in seen again in the plot. As mentioned earlier I will use this fact later to promote movie ideas for this group.

The last movie analysis undertaken here is looking at the standard deviation of the ratings for each of the twenty occupations surveyed. See figure #6. Looking carefully at the highest values in the figure #6 we see that groups 11 and 20 have the largest standard deviation for their ratings. These groups are k-12 students and the unemployed. In other words, this report is suggesting that these groups are hard to please. The groups are not monolithic and may reflect the general increase in diversity of the youth in America. Due to the fact that young students do not consistently rate highly (or debatable like) all the same movie attributes designing movies for the 'Young Americans' would be difficult. As mentioned above it would be best to attempt a multi-pronged approach and formulate a movie with as many of the top attributes as possible. One suggestion would be to write a thriller in space with drama, attractive male and female roles which interact with each on a comic level).

Alternately, we can see that our retired group has one of the smallest standard deviations. See figure #7. To this writer it implies that the retired group are more loyal to their ratings. If that groups finds something they like the retired will act more as a solid voting block and all join in together. This may also have its roots in American social history and be related to the homogeneity of that population. I have called the older aged group the 'Baby Boomers' because this strata of the population was born under good times after World War II and in the midst of the prosperous 1950's and 1960's.

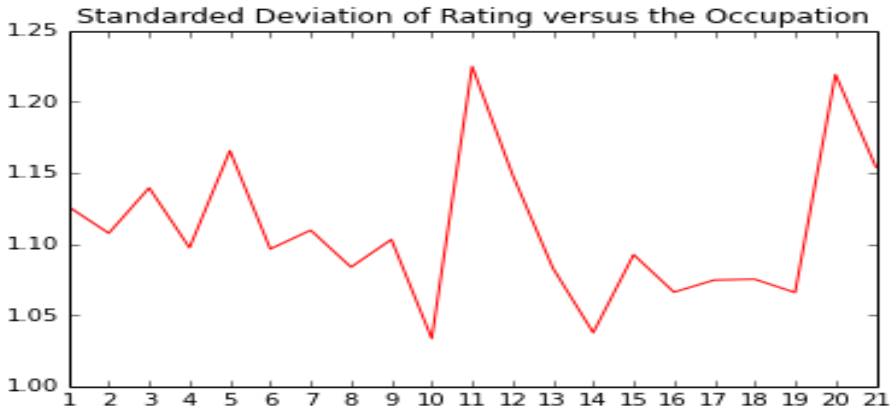

Figure #7: Standard deviation of movie ratings given by occupation, 21 categories.

In conclusion, predicting movie trends among changing demographics is a difficult undertaking. The MovieLens data which was used for this project was limited but basic trends could be discerned from it. The two major ideas that were found in this report was that the majority of movie-goers are between 18 and 44, known as the Young Americans. Though this group is very large and presumably very diverse however there is one pattern that sticks out. It is the trend for movies that 'hit' all the most liked genres. An amalgamation of genres would be in the best interest for a movie company to invest in. Movies that include 3, 4 or more genres would be best received by the Young Americans. The top genres include ranked in order.

Top Rated Movie Genres
1. Comedy 2. Drama 3. Action 4. Sci-fi 5. Thriller

The second movie trend that was gleaned from this data was that there is a population of older Americans, known as the Baby Boomers that are, seemingly, more homogenous in their movie viewing habits. This group is more apt to rate a movie more highly and has a relatively larger percent of disposable income to draw from. They have grown up with the movies as destination and maybe more loyal because of those miscellaneous facts.

Therefore this movie reporter would highly recommend a multi-pronged approach to making movies. Make movies for the Baby Boomers which are devoted and rate movies generally highly and conversely make movies for the vast number of youth that have diverse tastes and want a movie that hits all the favorite highlights.