|  |
| --- |
| Analyzing Data from Movie Lens |
| DS 501 – CASE II |
| Data extracted from Movie Lens were analyzed, which produced some information and conjectures were produced in order to help movie industry professional for decision making process. |
|  |

**Bir Kafle**

**Mohammed Ayub**

**Suman Lama**



**Introduction:**

This report walks through all the stages taken before coming into some conclusions in order to aid movie professional for decision making on the types of movies they should be making and investing on. The computing tool along with various statistical and mathematical knowledge were used to analyze raw data. Then, the information obtained from the processed data was used to suggest some advices for movie business professionals.

The computation tool and mathematical knowledge were not only sufficient to understand and analyze data. Some rational assumptions were made which were aided by outer research and related data. During this process some data were excluded, to better suit the need and help the process. Or, some data were neglected as suggested by mathematical analysis (for example: exclusion of outliers to balance the bias-variance tradeoff instead of making model more flexible).

**Data Collection:**

Data from Movie Lens were extracted from the site. The required data were present in three different tables; users (fields : 'user\_id', 'gender', 'age', 'occupation', 'zip'), ratings (fields: 'user\_id', 'movie\_id', 'rating', 'timestamp'), and movie (fields: 'movie\_id', 'title', 'genres'). The format of the initially available file was changed to more suitable file format (.HDF5: Hierarchical Data File) to ease the computing. The raw data from all the above three tables were merged to produce a single data frame which was then processed and analyzed.

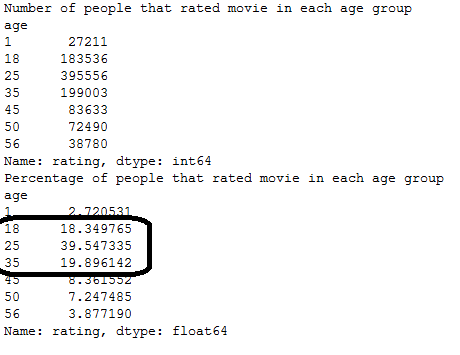
**Tools Used:**

Python was the language of choice used for writing codes to process the data, because of various available packages required to process the above data. On the other hand, python has the features of scripting, functional and object oriented language which is very handy, and also has all the packages required for creating graphs/drawings to produce visual reports. Design and presentation becomes more conspicuous and sophisticated cause of all these features available within the same language.

Statistical methods have been used to analyze and understand the data; hence can be considered as one of the tools used. Statistical concepts played vital role in visualizing data and during interpretation of data. Initially, statistical methods were used to choose data, and in the end they were used to reach a conclusion and also in some sense back the findings that were obtained.

**Results and Findings:**

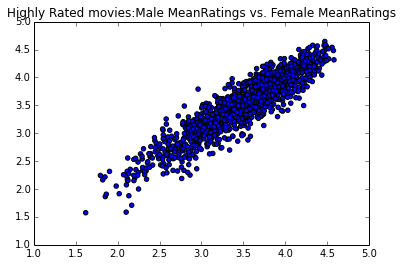
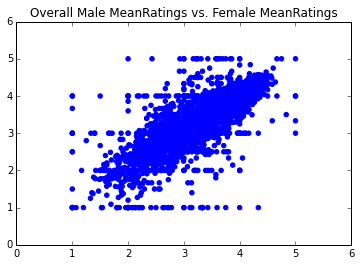
Various mathematical concepts, especially statistical, were used to analyze and understand the data. Several assumptions were made while analyzing and processing the data. For example; age group between 18 and 45 were used to produce conjectures and to draw conclusion as the population of US for this age category is more (36.5% of US population is between 18 and 44; US Census Data 2010). Also, this age category had the highest count than any other age category, and it would make sense to focus the group which watched more movies. On the other hand, this age group is more outgoing and economically stable- which means their view directly impacts how well a movie does. Socially, this age group is also indulges more in entertainment.



*Fig 1: Count and Percentage of people who rated movies from different age groups.*

Statistical concepts such as mean, median, variance, standard deviation, and correlation were used while analyzing the data. These concepts help to understand the data and summarize the data. The volume of the data required us to use mathematical analysis to understand and extract information from raw data.

The data was classified (men vs women, different age groups, genres) and processed. The computed value was compared pre and post classification which helped to understand the correlation between various groups. When the outliers were removed (this was done by selecting movies with more than 200 ratings present) the movie ratings from male and female had a strong correlation.

Fig 2 a) Overall Men Vs Women Mean Ratings b) Men/ Women Mean Ratings w/o Outlier

From the above two figures we can deduce that the figure on the right shows that male ratings are correlated with female ratings. But, the figure on the left has so many outlier points which makes difficult to interpret and define the relationship. If we had tried to include all those outlying data points we had to come up with more flexible model, and this would cloud the final conclusion. The best approach for these data points was to exclude those outliers and analyze the data.

The ratings distribution for various age groups helped us to deduce that people of all age group were more likely to rate in between 3 and 5, as the percentage of people rating between 3 and 5 constituted nearly or more than 80%. Also, it can be said that age group above 55 are more lenient in ratings as the percentage of people in age group above 55 who gave ratings of 4 (36.8%) and ratings of 5 (27.7%) are the highest than any other age group for the same ratings. Figure 3 illustrates provides a visual information about the Age Group and Ratings distribution and helps to justify above mentioned points.

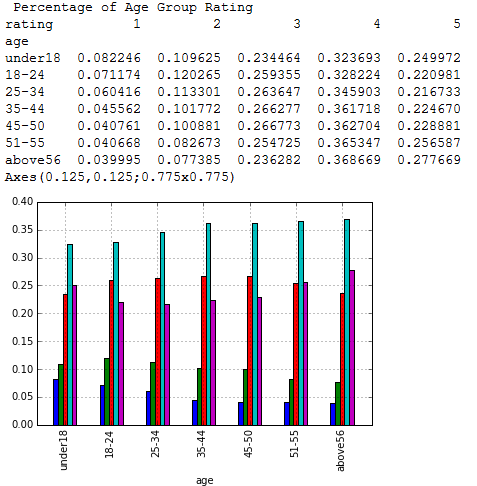


Fig 3: Ratings Distribution for various age groups.

**Validation of Results:**

According to the information obtained from the data that was processed it was found that age group between 18 and 44, are the primary consumers of movie. The data obtained from Census Bureau of US 2010 shows that age group between 18 and 44 comprises of 36.5% population. The percentage of people who rated the movie compromised mostly from age group between 18 and 44, and this validates the result and assumptions made.

Also, the processed data revealed that movie of comedy and drama performed well than any other movie. The results were compared with the financially successful movies at that time. Out of top 10 grossing movies of that time, 3 were comedy and 3 were drama. This validates the result obtained and proposes that meaningful information was obtained after processing of raw data.

**Business Intelligence Propositions:**

As mentioned above, comedy and drama were the most popular movies. The result was obtained by using separate different data subsets. Initially, popular genre among the age group between 18 and 44, excluding other age groups was computed. The top five most popular genres for this age group were Comedy, Drama, Comedy/Drama, Comedy/Romance and Drama/Romance. Second popular genres among the age group below 18 and above 45 were computed. The top five most popular genres for these age groups were Drama, Comedy, Comedy/Romance, Comedy/Drama and Drama/Romance.

So, the proposition is movies should be one of these top five genres. Since, these genres best represents the majority of the audience (age group between 18 and 44), given the fact that population of this particular age group is higher. Also, the income distribution of this age group is good compared with other age group, which makes this group to be the prime consumers of the movie industry.

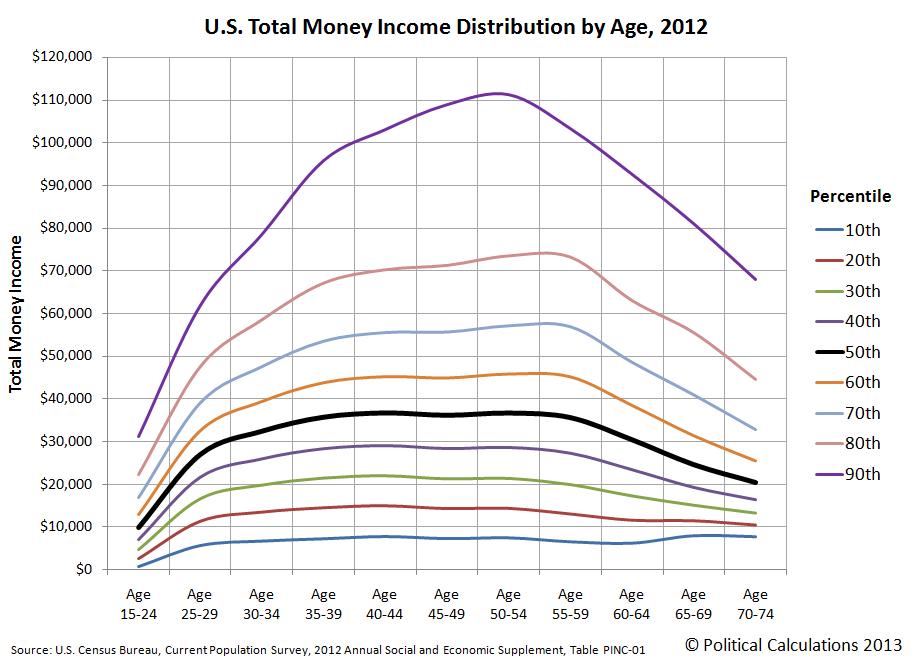


Fig 4: US Income Distribution for various age groups.

**Conclusion**:

The raw data available from Movie Lens were difficult to comprehend initially, but with the help of computing tool and mathematical knowledge, information were generated from raw data. The informative data were then used to draw the figures and tables so that data could be easily grasped by any user.

Through this case study it helped to implement mathematical concepts in processing/selection of data and also refining and finally aided to back the results that were obtained. Overall, the project was helpful in understanding the importance of computing tools and mathematical skills and concepts to process the data.

**Appendix/Citations**:

1. Code Attached
2. Table Attached
3. DS -501 Slides link:

<http://users.wpi.edu/~xkong/course/ds501/f14e/index.html>

1. Python for Data Analysis: Wes McKinney, Edition: 2012
2. Python Documentation

<https://docs.python.org/2/>

1. Pandas

[www.pandas.pydata.org](http://www.pandas.pydata.org)

1. Matplotlib

[www.matplotlib.org](http://www.matplotlib.org)