**Data Analyst Exercise**

**Introduction**

Latest MovieTweetings snapshot files include three data sets: movies, ratings, and users. Each data set contains several variables. After these data sets were imported to R, data validation processes were applied. Through the data validation process, Genre, Rating, and Date information were mainly both focused and studied on. As a result, three non-trivial observations were found and stated in the following sections.

**Data Preparation**

Firstly, data sets were imported to R. Separators in data files are stated as “::”. Since “movies” data set includes movie names that may contain “:”, its conversion to convenient data table is different compared to other data sets. Briefly, the function to split the columns was used. Then, general information about the data were extracted.

As to “movies” data set, it includes movies’ IDs, names, genres, and release years. There are 31,698 different movies having release year from 1878 to 2018. 222 of movies do not have any genre.

For “ratings” data set, it includes users’ ID, movies’ ID, ratings done by corresponding user and rating time stamp. Rating time stamp is the length of time from 01.01.1970 00:00:00 to the time that user rated the movie. This length of time is represented by seconds. There are 727,362 ratings in total. Additionally, all movies have been rated. Average rating of all the rated movies is 7.301803.

As to “users” data set, it consists of two variables: users’ IDs and Users’ Twitter IDs. There are 54,702 users. All users rated at least one movie.

**Data Validation**

Data validation is a very important process for the analysis. Data must have meaning as a whole and in parts. At this point, rating information, which is a pretty important factor in data analysis, has been studied on. Two questions about rating information were answered. First one is whether all ratings lie between 0 and 10 as it should be. The other one is whether anyone rated a movie before its release year.

As to first question, there is no rating out the range from 0 and 10. For the second one, rating year was derived from the rating time stamp. It was found that there are three ratings submitted before the release years of movies. These three ratings were removed from the “ratings” data set.

**Observations**

As mentioned before, three main factors, which are genre, rating, and date information, regarding movies were mainly both focused and studied on.

As to first observation, the number of ratings and average rating per movie were calculated and added to “ratings” data table as variables. Then, movies were set in decreasing order of number of ratings per movie. The most rated 30 movies’ average ratings were printed, and it was detected that they all are around or more than average rating of all movies, 7.30. Then, can we say that “good” movies are rated more than “bad” ones? Before answering this question, the definition of “good” and “bad” must be made. Histogram of movie mean ratings was plotted and stated below.

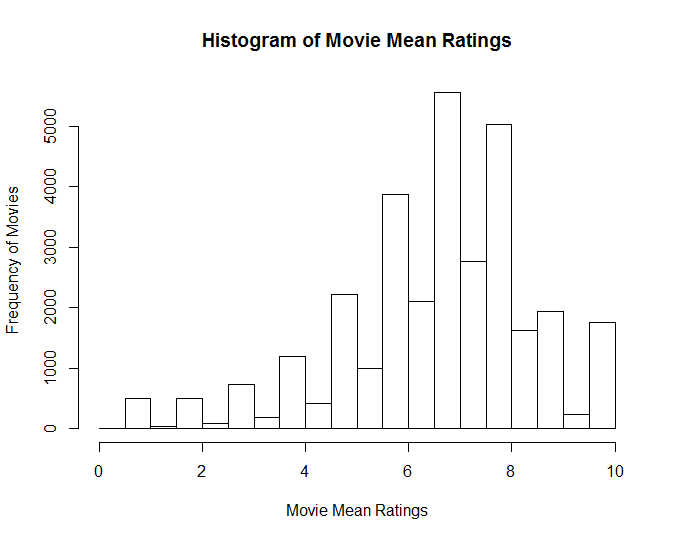


Figure 1

Movie mean ratings are accumulated around 7. Then, “good” movies were grouped as the ones having average rating more than 8. Movies having average rating less than 5 were grouped as “bad” movies. Movies other than “good” and “bad” ones were removed for this observation. A variable called “isGood” was created and it represents “good” movies as 1 and “bad” ones as 0. In order to remove movies with low number of ratings, box-plot of rating numbers was constructed. Movies having rating counts which are in lower whisker or below this point were removed. The relation between the variables “isGood” and rating counts was examined by one-way ANOVA test. P-value of the test is 0.0000000000000002 which means there is a significant relationship between “isGood” and rating counts. Then, we fail to reject that “good” movies are rated more than “bad” ones.

The next observation was found using genre information. Most-rated 1000 movies were extracted from data and genre information regarding these movies was represented by data table. Same process was applied to the highest-rated 1000 movies. However, for these movies, some of them have pretty low number of ratings; therefore, a constraint was added which chooses movies with rating number more than average number of ratings per a movie. Movies which were included in both data tables were removed from the two data tables. There are 659 movies left in most-rated data table and highest-rated data table as well.

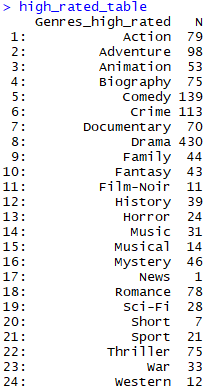
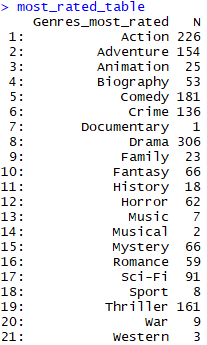


Figure 2: The total numbers shows how many times a genre used by   
most- and highest-rated 659 movies

The percentage of genres used by these two data tables (highest- and most-rated 1000 movies) were calculated. The percentages, corresponding to genres that are included in just other data table, were set to zero. The percentages of genres in two data tables were subtracted from each other and the deviation of these values were examined and stated in plot shown below.



Figure 3

Differences in percentages of genres used by most-rated movies and highest-rated movies deviate around zero which means it may not be said that most-rated movies have similar genres with highest-rated movies.

As to third observation, in order to analyze the effect of time on user ratings, one of most rated movies, whose rating score is slightly more than average, were selected. The reason to choose one of most rated movies is that more data results in more accuracy in an observation. This movie is “Gravity”. Its release year is 2013, which is the earliest rating time in the data set. This eliminates the condition that the number of ratings per particular length of time decreases to some degree over years after the release date of a movie.

Firstly, a graph showing ratings through time was plotted stated in Figure 4. However, it does not show the number of ratings (the density). Therefore, monthly mean ratings were calculated and stated through time in Figure 5.



Figure 4

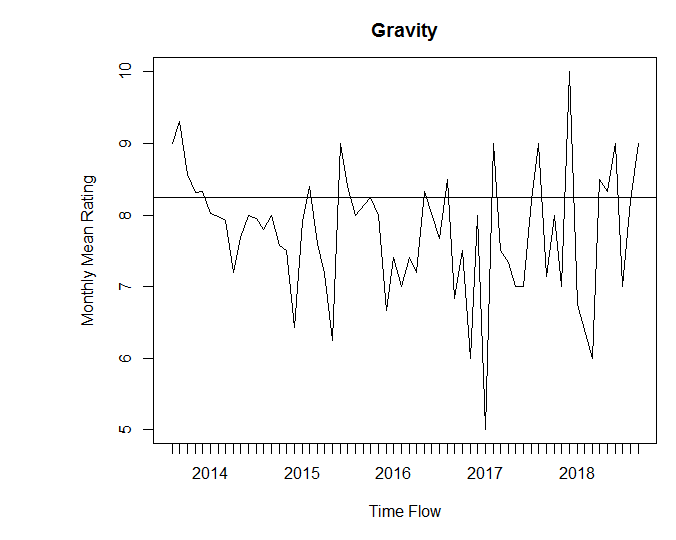


Figure 5

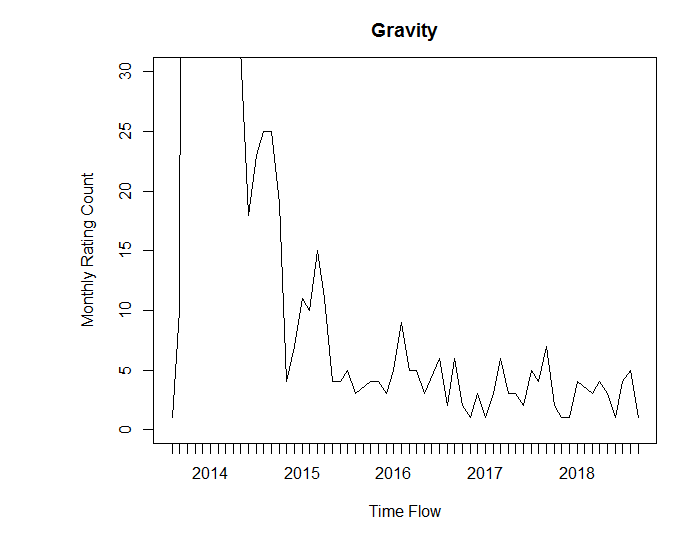


Figure 6

Figure 5 shows that monthly mean rating generally fluctuates below average rating of the movie especially a while after the movie was released (after the beginning of 2015). Rating count per movie per month was plotted and shown in Figure 6 as well. It can be seen that fluctuations are similar for both graphs stated in Figure 5 and Figure 6. For example, there was downward trend in monthly mean rating graph at the beginning of 2015. Then, rating increased with upward trend for a while. Same fluctuation can be seen in monthly rating count graph as well. When rating count increased, mean rating increased as well. Same processes can be observed multiple times through time. Additionally, correlation value was calculated as 0.16 which shows weak correlation in fluctuations between these two graphs. Therefore; it can be said that people who rated higher than average rating are more than people who rated lower than average rating. That is the reason why average rating line is above the region in which monthly mean rating fluctuates. If a movie was rated relatively high early on, that affects people and most of them rated high as can be seen in Figure 4 and 5.