The Data

For my initial capstone project (“A Guide to build a movie recommendation engine”), the primary data source is MovieLens dataset which includes 3 .dat file for movies, users and ratings. I loaded each dataset to different dataframe and did these steps:

1- Searching null value and missing value with querying data frames with command like

“.isnull().values.any()” and didn’t find any missing value.

2- Trying to find outliers and check the values if they are in the scope or not? For example, “Ratings” data frame, has one column ‘Rating’ which must have a value between 1 to 5, and apply command “.groupby(‘Rating’ )” to dataframe and check this column has any other value other than 1 to 5 or not? If yes, replace with appropriate value.

3- Searching for duplicate rows in data frames and if find any, investigate more to figure out what strategy must take(remove or keep). In the “User” data frame, I found duplicated rows which has different “UserId” but the other columns(‘Gender’,’Occupation’,’Zip Code’) are the same and it sounds to me, they are the same users, so for more investigation I merged “Users” data frame with “Ratings” data frame to correlate the users with their ratings and then perform grouping to see how many rows exist for each of these users (which they seem the same users) and figured out each of these user has a lot of ratings, so I had to decide to keep these rows as they are or delete duplicated rows from users and just keep one of them and merged all related rows in Ratings dataframe and assigned them to that one existent user. After thinking about the final result of this project, I came up with the solution to keep them as they are like different users, because the existence of these duplicated users don’t have any side effects on any final reports and analysis.

Now I have 3 data frames(Users, Movies and Ratings) which are cleaned and ready for further exploration.