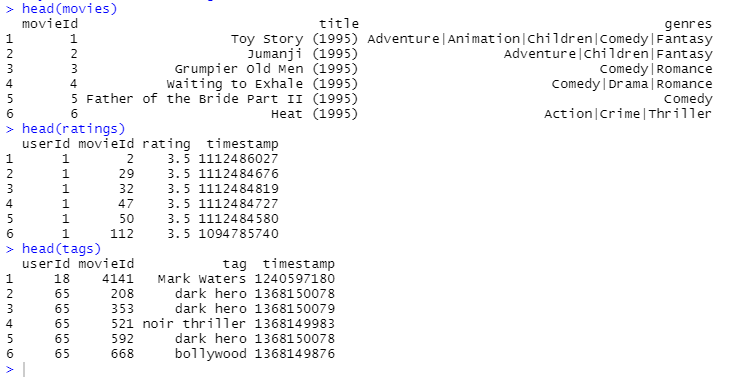
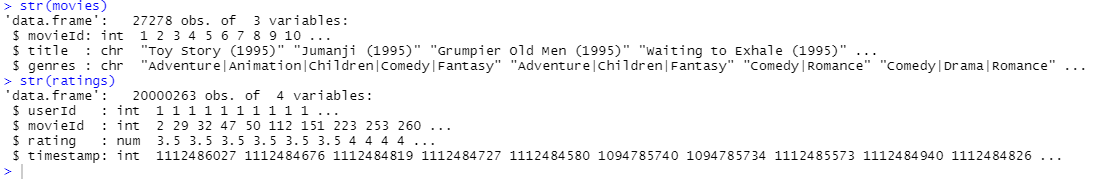
Movie recommender system for movie ratings. In this report we will create a prediction model that predicts movie ratings given by users. We will be using the 20M version of Movielens generated by GroupLens research Lab.

Data Overview





From the overview above, we can see that UserId is a unique ID for the user and movieId unique ID for the movie. The rating column is the prediction that we want to make and the rest of the columns can be used as predictors.

We can see that the movies dataset has 27278 rows and 3 columns. Ratings dataset has 20000263 rows and 4 columns.

* userId : Unique Id for user
* movieId: unique Id for movie, consistent between ratings,  tags ,movies and links. movieId refers to the same movies across these four data files.
* rating: Rating between 0 and 5 for the movie
* title: The movie title, titles are entered manually and imported from <https://www.themoviedb.org/>.
* timestamp: Date and time the rating was given
* genre: genres  list separated by pipeline |  includes the following : Action, Adventure, Animation, Children’s, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War , Western.

Scatter chart

Description automatically generated with medium confidence

Do a left join between ratings and movies using movieID. We will use this to get better correlation and visualization.

**Data Analysis**

Before we build the prediction model , lets analyze and visualize each predictor to gain insight on the distribution.

Graphical user interface, text

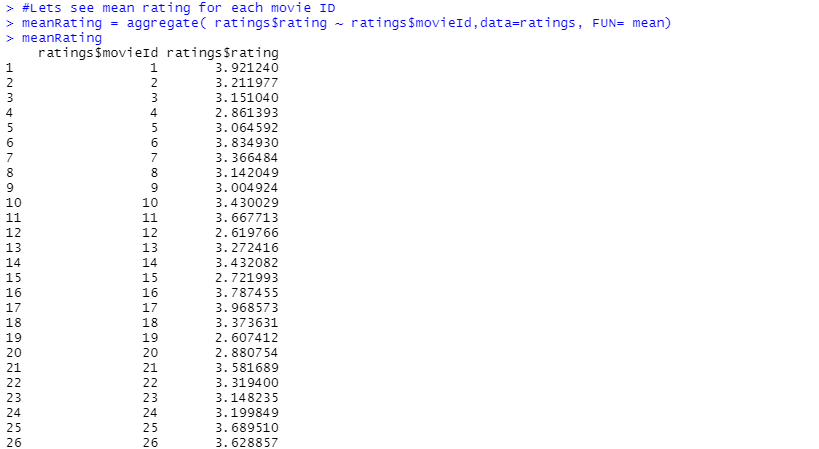
Description automatically generated

Chart, histogram

Description automatically generated

This histogram actually shows that the ratings are normally distributed . There are certain movies that aren’t rated much at all while some are rated numerous times by users.

Lets see the mean rating for each movie ID



With 500 observations, it is keen to note that the median seems to lie around 3 ~ 4.

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

The histogram shows some users are more active in rating the movies. The ratings are approximately normally distributed but skewed to the left.

Graphical user interface, text, application

Description automatically generated

Chart, histogram

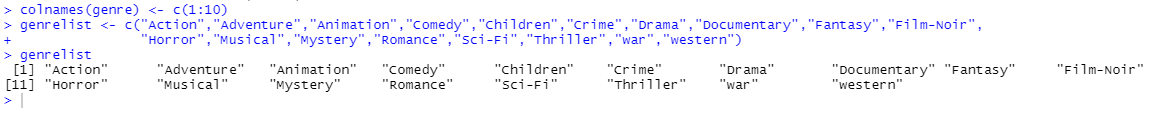
Description automatically generated

The ratings are approximately negatively skewed.

The genre is a pipeline of list separated by ‘|’. Lets get all the genres into a data frame for preprocessing.

Text

Description automatically generated



We can create a one-hot encoding for a genre matrix of each film. We might want to create a search matrix to search for ratings specified by genre.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated with low confidence

Lets take a look at our target rating. We won’t be using it for predictions but is worth to analysis.

Chart, line chart

Description automatically generated

The graph shows that users have a higher tendency to give round numbers for the rating. We can also see that users tend to give higher ratings.