**Building user-based recommendation model for Amazon**

## **Problem Statement:**

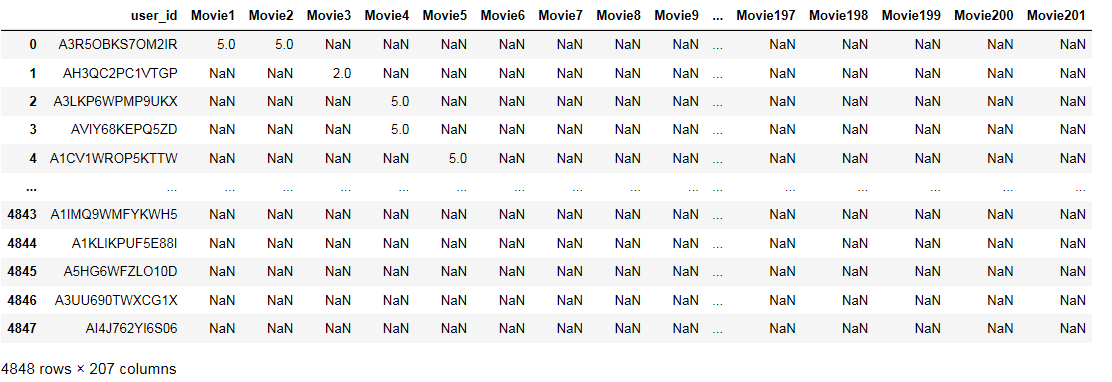
DESCRIPTION

The dataset provided contains movie reviews given by Amazon customers. Reviews were given between May 1996 and July 2014.

**Data Dictionary**  
UserID – 4848 customers who provided a rating for each movie  
Movie 1 to Movie 206 – 206 movies for which ratings are provided by 4848 distinct users

**Data Considerations**  
- All the users have not watched all the movies and therefore, all movies are not rated. These missing values are represented by NA.  
- Ratings are on a scale of -1 to 10 where -1 is the least rating and 10 is the best.

**Dataset:-**



* Dataset have 4848 rows and 207 columns, 4848 rows contains number of user id and ratings & 207 columns have Movie1 to Movie 206.

**Analysis Task**

- Exploratory Data Analysis:

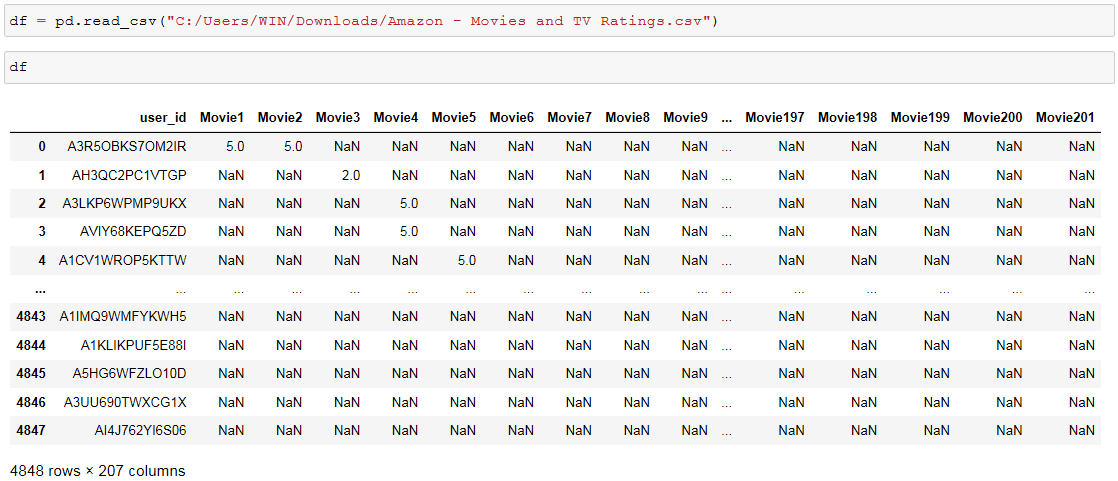
* Which movies have maximum views/ratings?
* What is the average rating for each movie? Define the top 5 movies with the maximum ratings.
* Define the top 5 movies with the least audience.
  + Recommendation Model: Some of the movies hadn’t been watched and therefore, are not rated by the users. Netflix would like to take this as an opportunity and build a machine learning recommendation algorithm which provides the ratings for each of the users.
* Divide the data into training and test data
* Build a recommendation model on training data
* Make predictions on the test data

# **Solution**:

We start by importing all libraries to perform analysis



**Import data into Python environment.**

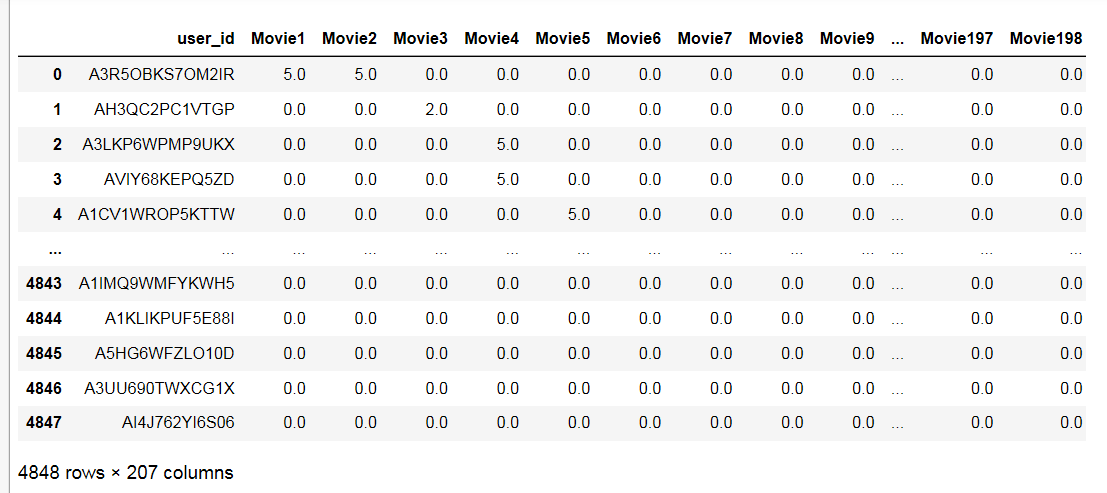


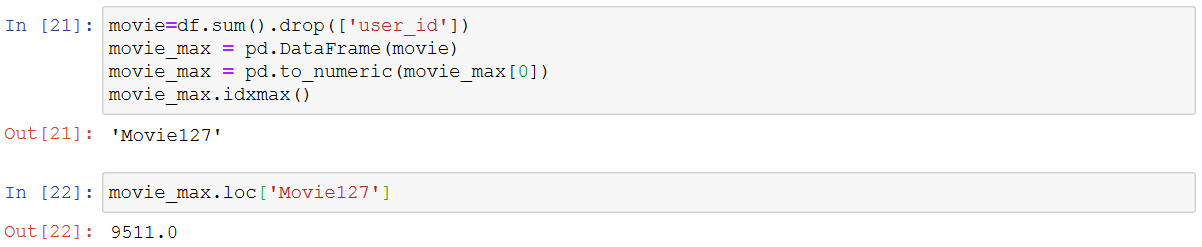
**The Tasks are :**

1. **Which movies have maximum views/ratings?**
   * We are going to fill NA value with 0 to perform this task



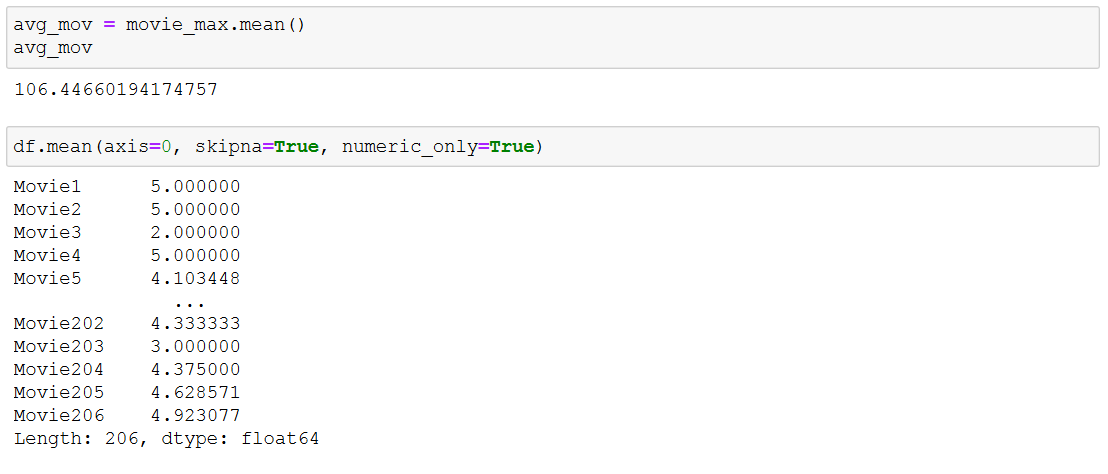
Output :-





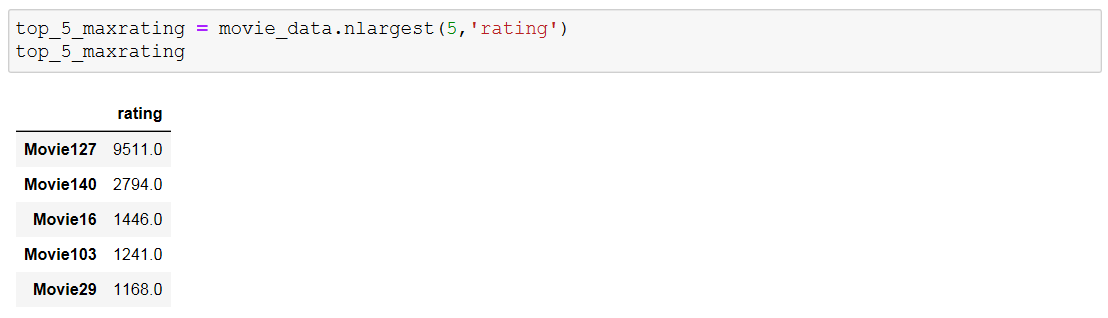
* + Output is clearly shows that ‘Movie127’ has maximum ratings 9511.

1. **What is the average rating for each movie?**

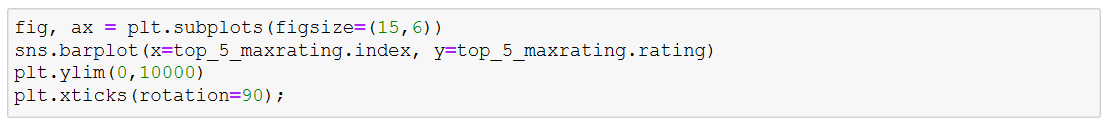


* + Output is showing 106.44 is average rating of all movies and below output shows average rating for each movies.

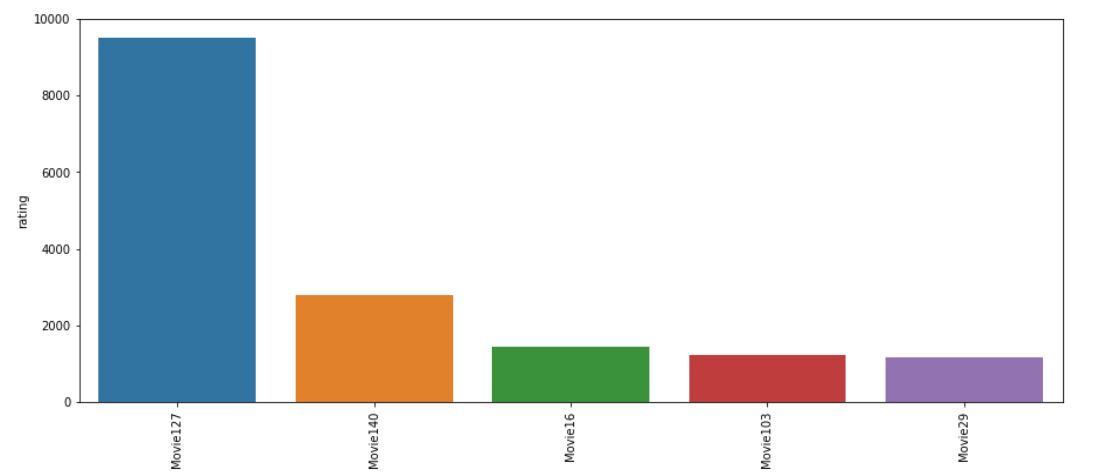
**3. Define the top 5 movies with the maximum ratings.**



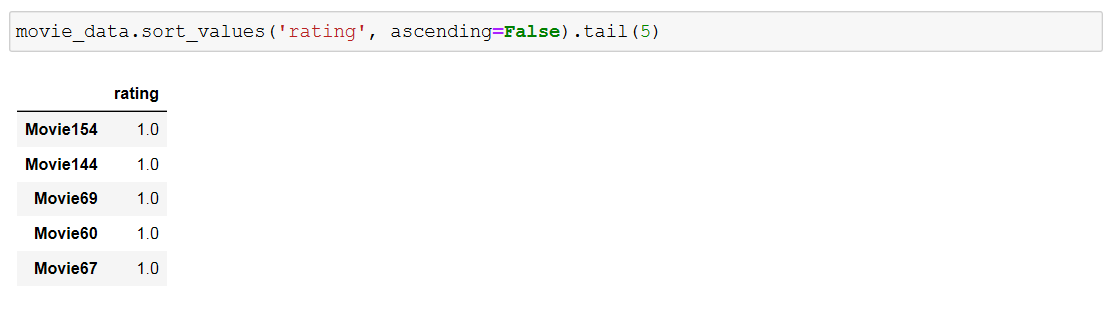
* + As we can see in outputMovie 127, 140, 16, 103, 29 is highly watched and highly rated movies.
  + Visualization of highest rated movies shown below,



Output:-

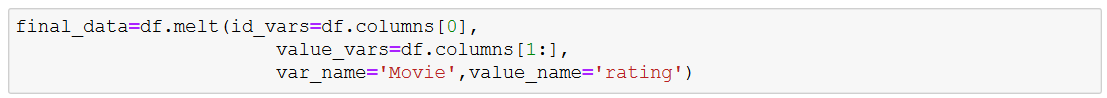


**4. Define the top 5 movies with the least audience.**

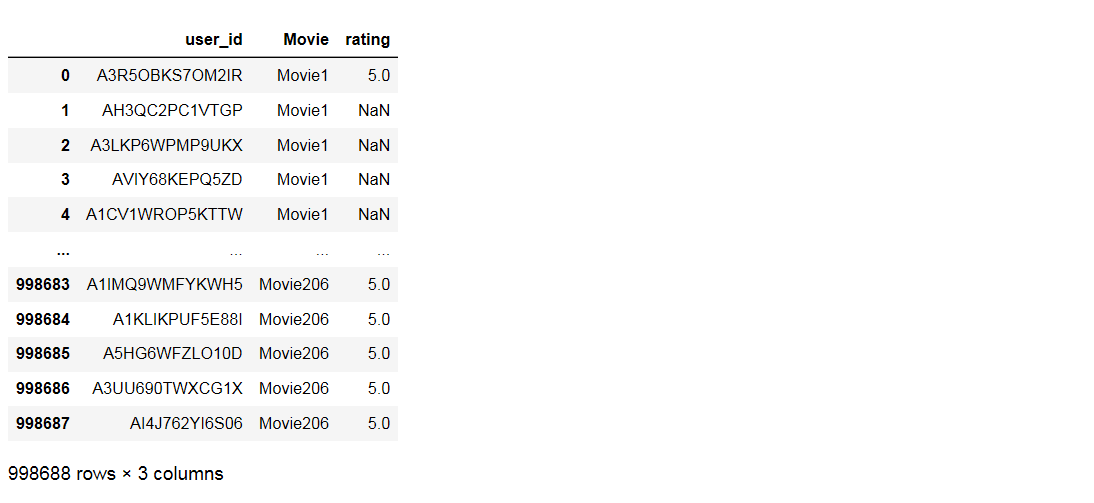


**Recommendation Model**

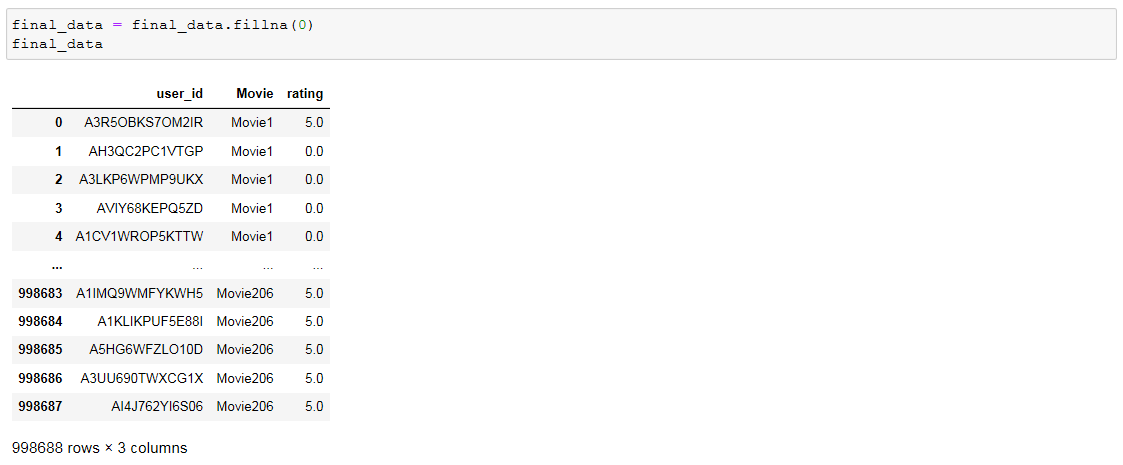
**Data preparation :-**



Output:-



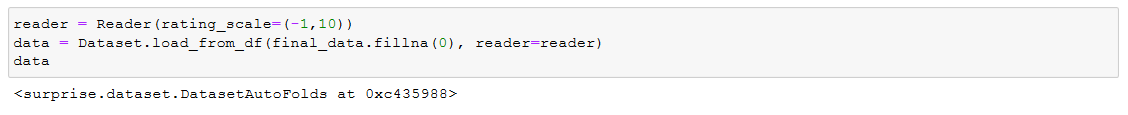
* + Unpivot a DataFrame from wide to long format by using melt function now data frame have 998688 rows and 3 columns.



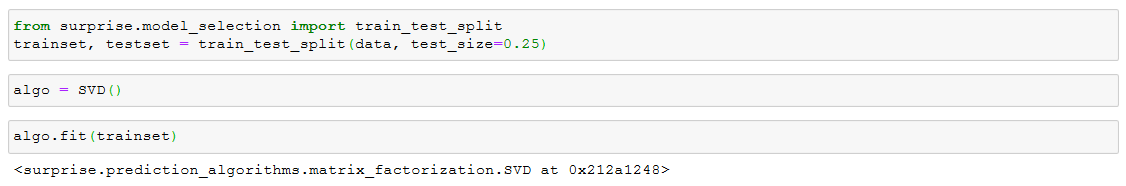
* + Now we prepare final data replace NA with zero
* Importing all libraries to make recommendation model



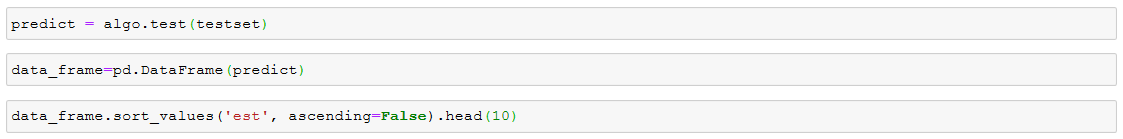
* Ratings are on a scale of -1 to 10 where -1 is the least rating and 10 is the best.



* Dividing data in to 70 % trainset and 25% testset.
* Importing SVD model and fit model to train set data.



* Now predicting on test set data



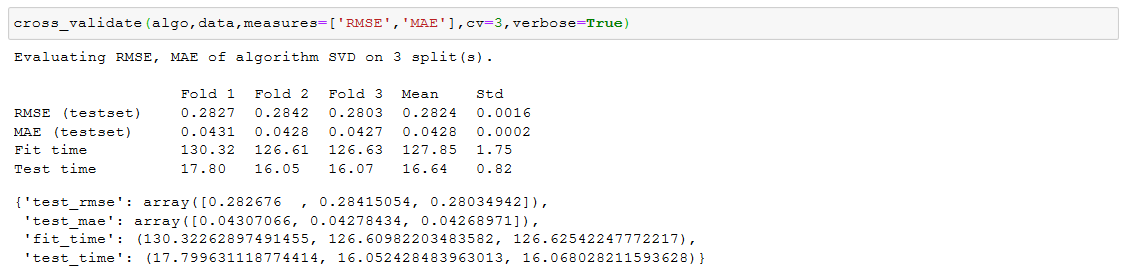
Output:-



* To predict accuracy of the model calculate Root Mean square Error



* + We can see RMSE of model is 0.2742 which is low indicate this is good model.
* We can also check RMSE with cross validation techniques.



* + As we can see from Cross-validation techniques having 3 folds of train and test set after predicting on test data we calculate RMSE for each folds and mean of RMSE for these folds is 0.2824 which is low also mean of Mean Absolute Error of these fold is 0.0428 which is also low indicate that model perform good on test data.