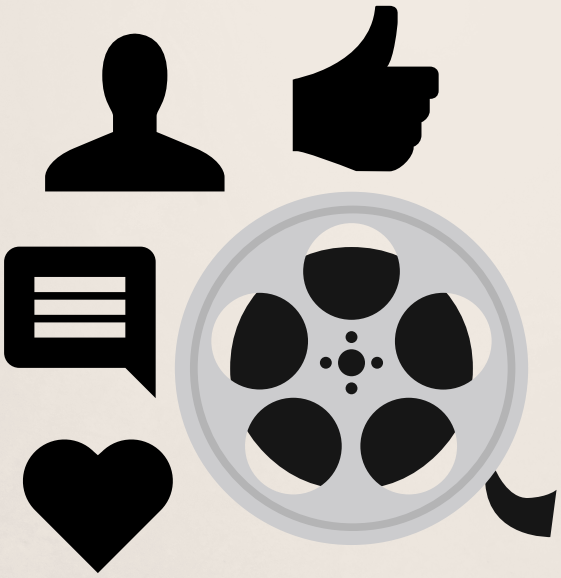


Movie Recommender system

IDS 572 Assignment 5

Group Details

NagaShrikanth Ammanabrolu	**676837954**
Suresh Sappa	**667192596**
Sagar Kanchi	**669850639**

Table of Contents

Introduction

The MovieLens dataset of 100,000 ratings arising from 943 users and 1682 movies is our target dataset, which has been divided into training and testing segments. The primary goal of this project is to identify the segments of users and how they rate movies. A recommender system has to be built using the available rating information to score the incoming data and predict the likelihood of the user liking a particular movie. The movies with a higher probability of being liked by the new incoming user are recommended. We begin the analysis by carrying out data exploration and then perform recommender system analysis using different techniques, whose results will be summarized below.

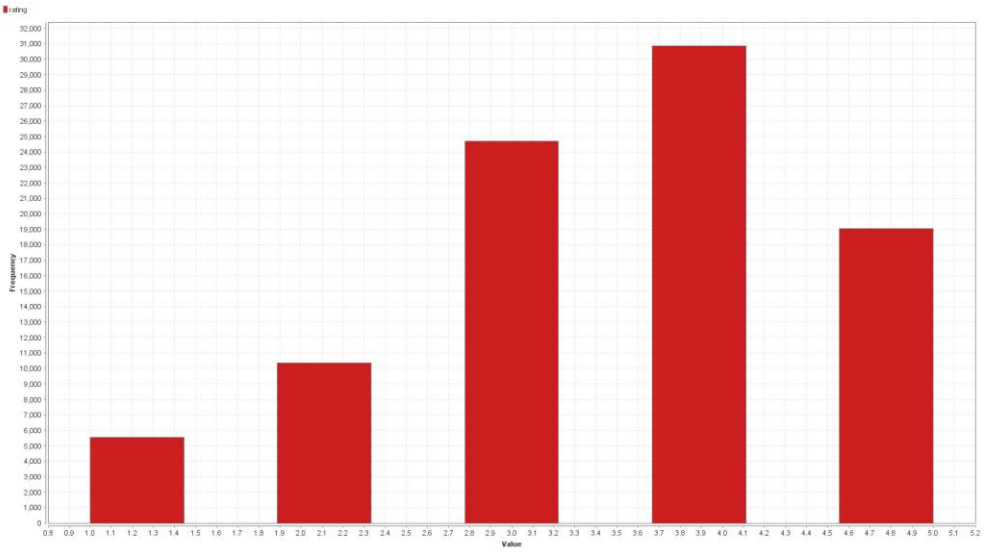
Question No. 1) (a)
Solution.

Data Exploration: Overall distribution of ratings

The first step in the analysis is Data Exploration, where an understanding of the users, movies and ratings ought to be established. The ratings for the movies range on a scale of 1-5 and the overall distribution of these ratings in the dataset is as described below.

Overall Distribution of Ratings		
Rating	**Number of instances**	**Overall Fraction of Ratings**
Rating 1-star	5,568 Ratings	6.14%
Rating 2-star	10,375 Ratings	11.50%
Rating 3-star	24,721 Ratings	27.30%
Rating 4-star	30,858 Ratings	34.01%
Rating 5-star	19,048 Ratings	21.03%

A graphical representation of these overall distribution of ratings is obtained as shown below



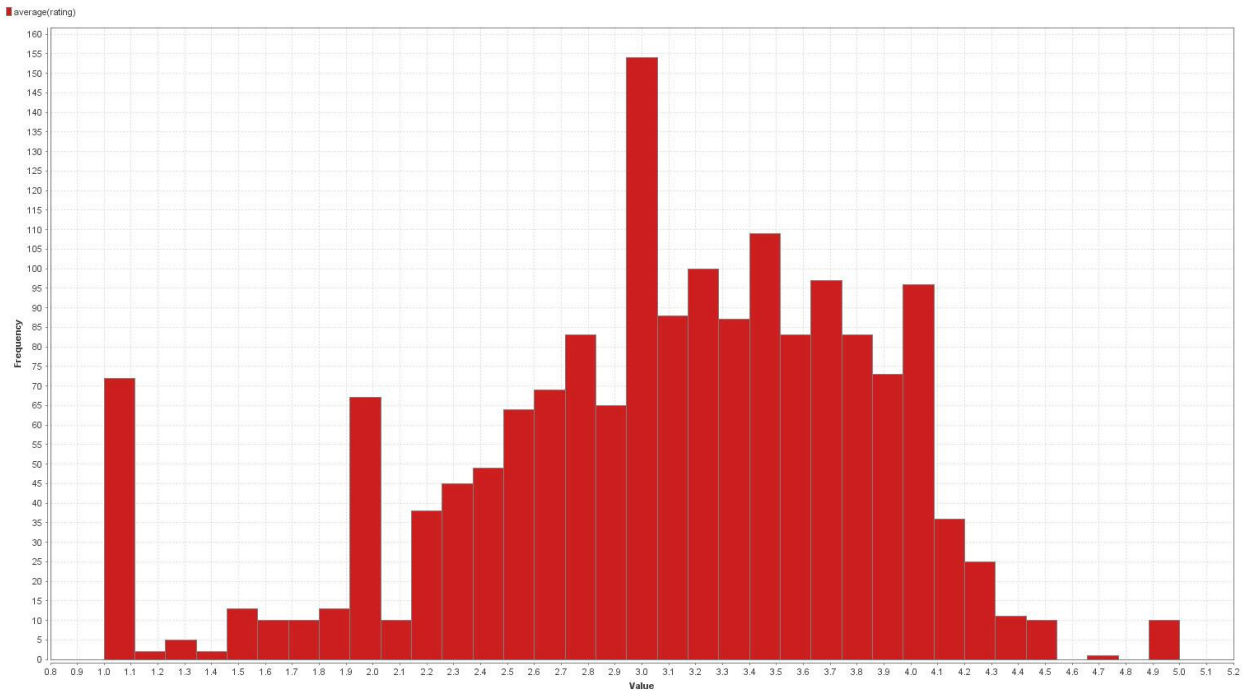
From this information, we can summarize that most of the users who rated movies, have a higher tendency to rate the movie 4 stars, with 34.01% doing so. The next most popular rating, at 27.30% is a rating of 3 stars. 5 stars rated movies share for 21.03% in the dataset and is the third most popular choice among movie raters.

The least popular rating, which users tend to reserve for a select few is a 1-star rating.

Data Exploration: Distribution of ratings by Users and Movies

Now that we have an understanding of the overall distribution of movie ratings, we now obtain the distribution of ratings by users and movies. The data exploration is performed on the training dataset again and the resulting distribution of ratings are as summarized below.

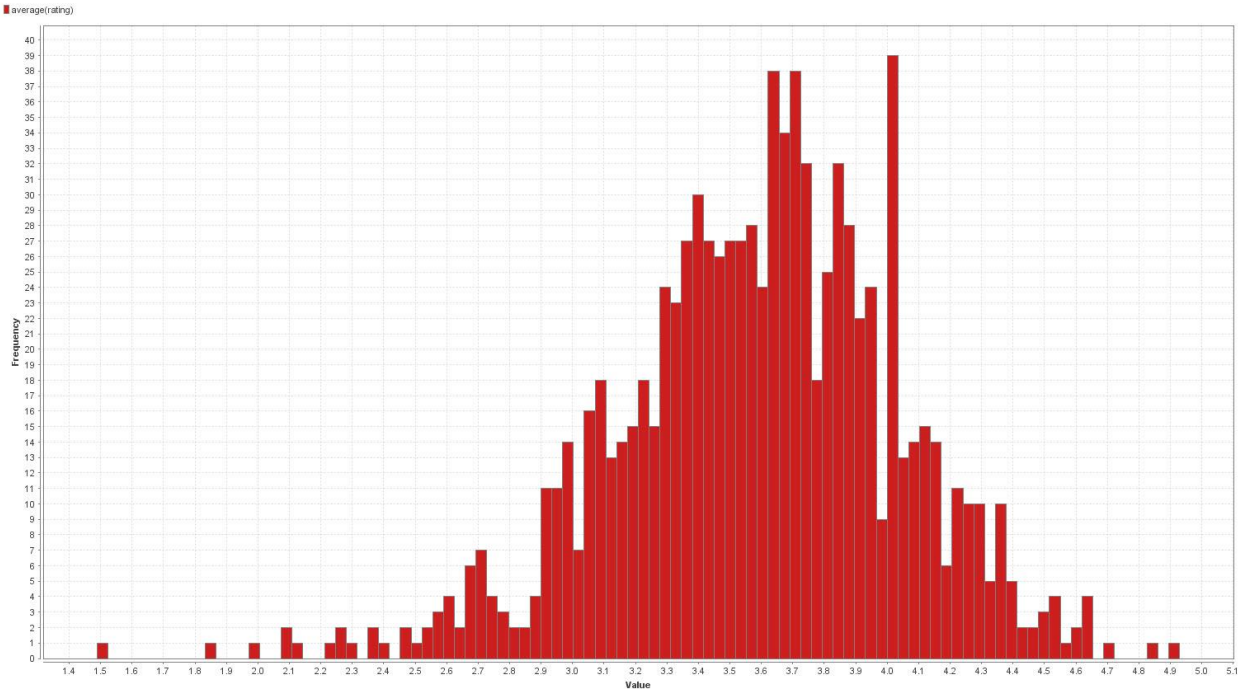
(A) Average Distribution of ratings by Movies



The average distribution of ratings by movies is as shown above. We find that 1049 items out of the total 1680 movie items in the training dataset have a rating of 3 stars or above. We obtain that 62% of the movies on an average have a rating of 3-stars or above.

The average rating by movies in the training dataset is found out to be 3.067, with a standard deviation of 0.8.

(B) Average Distribution of ratings by Users



The average distribution of ratings by users is as shown above. We find that 861 users out of the total 943 users in the training dataset gave the movies a rating of 3-stars or above. We obtain that 91.3% of the users on an average gave a rating of 3-stars or above.

The average rating by movies in the training dataset is found out to be 3.588, with a standard deviation of 0.453.

Data Exploration: Distribution of Rating counts by Users and Movies

Overall Distribution of Rating Counts			
Rating Counts by Movies		**Rating Counts by Users**	
Movie ID's	**Rating Count**	**User ID's**	**Rating Count**
Movie ID 50	495 users	**User ID 405**	727 movies
Movie ID 100	443 users	**User ID 655**	675 movies
Movie ID 181	439 users	**User ID 13**	626 movies
Movie ID 258	412 users	**User ID 450**	530 movies
Movie ID 286	400 users	**User ID 276**	508 movies
Movie ID 294	398 users	**User ID 416**	483 movies
Movie ID 1	392 users	**User ID 537**	480 movies
Movie ID 288	386 users	**User ID 303**	474 movies
Movie ID 121	384 users	**User ID 234**	470 movies
Movie ID 174	379 users	**User ID 393**	438 movies

From the above summarization of the top 10 movies and users with the most rating counts, we find that Movie ID 50 and User ID 405 are the most popular items in the dataset.

Movie ID 50 has been rated by 495 users out of the total 943 users in the training dataset, which sums up to approximately 52.5% of the entire users. User ID 405 is the most popular user on the training dataset, having rated 727 movies out of the total 1682 movies, which sums up to approximately 43% of the entire movie database.

Data Exploration: Distribution of rating levels

The ratings of the movies are distributed across 5 different categories and could be scaled into one of the four sections as described below. These rating levels include the said rating and anything higher than the rating. For example, a rating level of 'Higher than 4-star' includes all the users who have given an average rating of 4-star or above.

Distribution of Rating Levels		
Rating Level	**Number of users**	**Overall Fraction of Ratings**
Higher than 4-star	173 users	18.34%
Higher than 3-star	861 users	91.30%
Higher than 2-star	941 users	99.78%
Higher than 1-star	943 users	100%

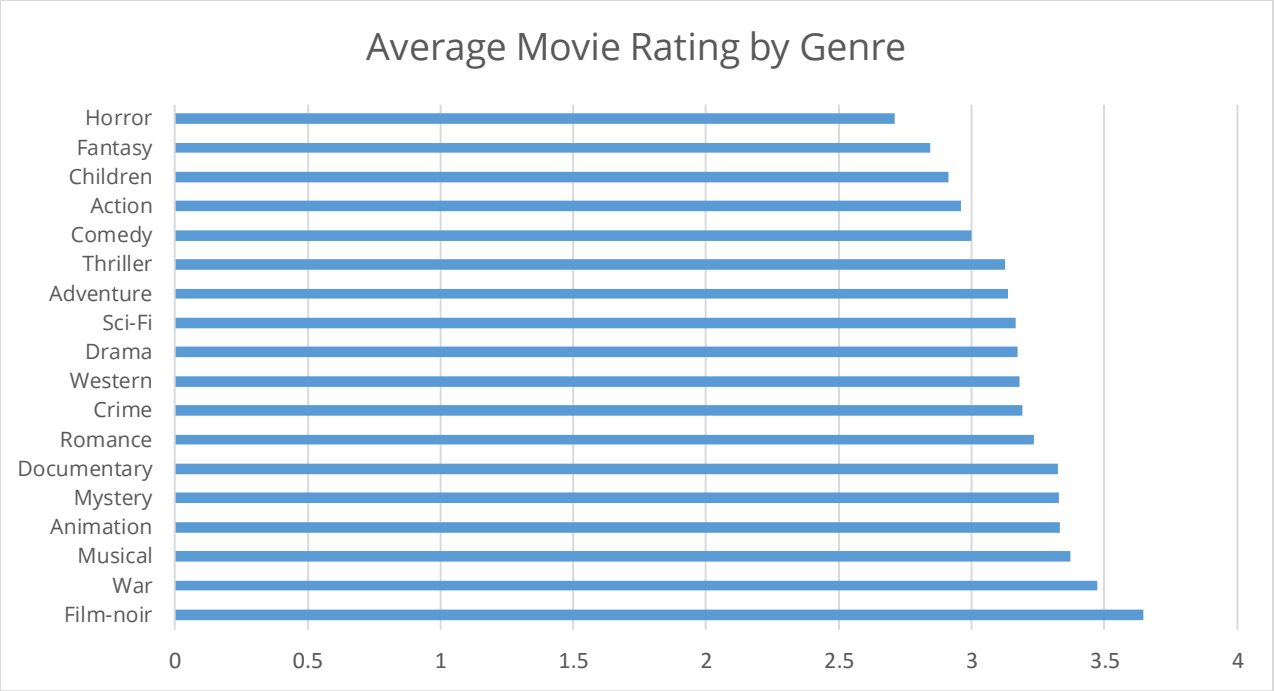
Question No. 1) (b)
Solution.

Data Exploration: Ratings by genre

The next step in the process of Data Exploration is to understand how demographic variations affect the movie ratings. Here, we will be analyzing the ratings of movies by genres. The primary purpose of such an analysis is to see if there are any specific genres of movies that are performing particularly well in comparison to others.

An information of the Top 5 movie genres and their average movie ratings are summarized below.

Distribution of Ratings by Genre		
Movie Genre	**Average Movie Rating**	**Number of movies**
Film-Noir	3.647	23 movies
War	3.473	71 movies
Musical	3.372	56 movies
Animation	3.3304	42 movies
Mystery	3.33	61 movies



A summarization of the average movie ratings by all of the genres is as shown above.

After grouping the dataset by items aggregating across users and ratings, we perform an Inner join with the 'u_item.csv' data. Later using aggregate operator for different genres, we gather each of the movie/genre pairs and filter for only the cases when a movie belongs to a genre, and then find the average rating.

(*A summarization of the Ratings by Genre for the Top-5 Genres is given in the Appendix*)

Data Exploration: Ratings by user age groups

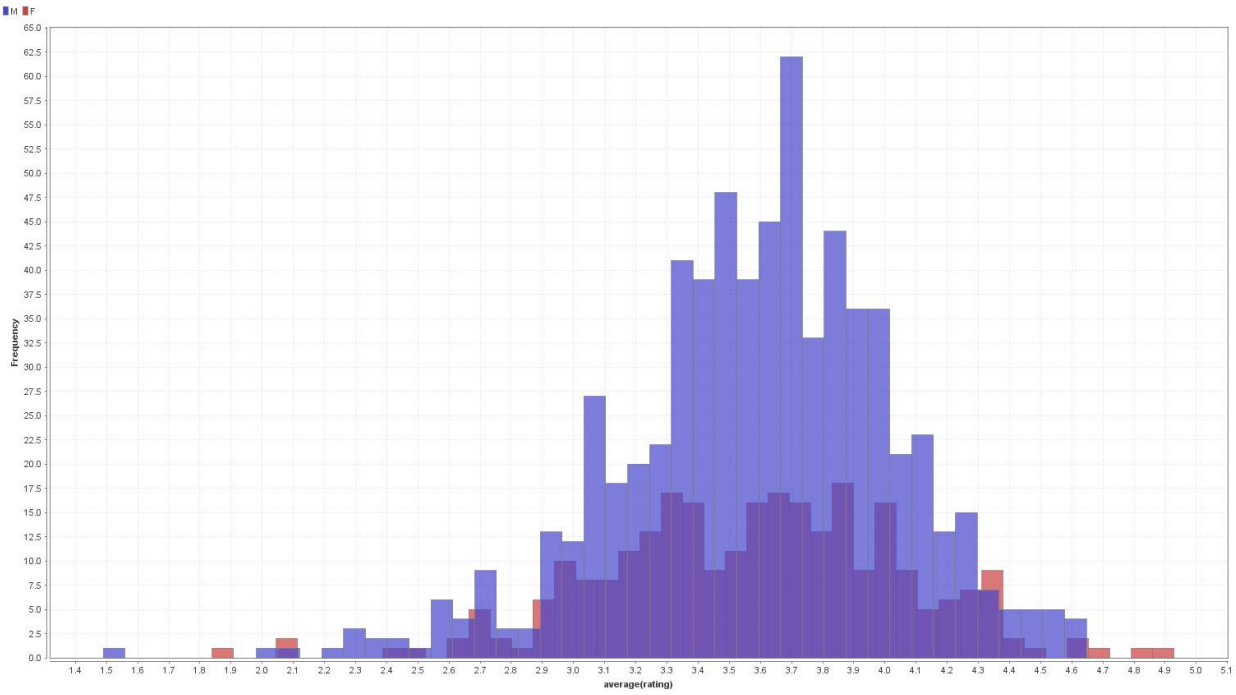
For the purpose of summarizing the rating of movies by age groups, after grouping the dataset by users and aggregating across ratings and items, we perform an Inner join with the 'u_user.csv' data. Using the 'Generate attribute, we create six new attributes for the user age groups. Later, using aggregate operator for different age groups, we gather each of the user/age group pairs and filter for only the cases when a user belongs to a particular age group, and then find the average rating.

Distribution of Ratings by Age groups		
Age groups of users	**Average Movie Rating**	**Number of users**
Under 20 years of age	3.480	77 users
20 to 30 years of age	3.568	332 users
30 to 40 years of age	3.587	241 users
40 to 50 years of age	3.603	168 users
50 to 60 years of age	3.706	94 users
60 years of age and above	3.633	31 users

From the above table, we can see that the average movie rating tends to remain somewhat same across all of the age groups. Age groups of 20 to 30 years of age have the highest number of users who have rated movies in the given training dataset. Hence, we can safely conclude that age of users does not play a major role in the way users rate movies.

Data Exploration: Ratings by gender

After aggregating the user demographics information, we can plot a histogram of the users' gender versus their average movie rating as shown below.

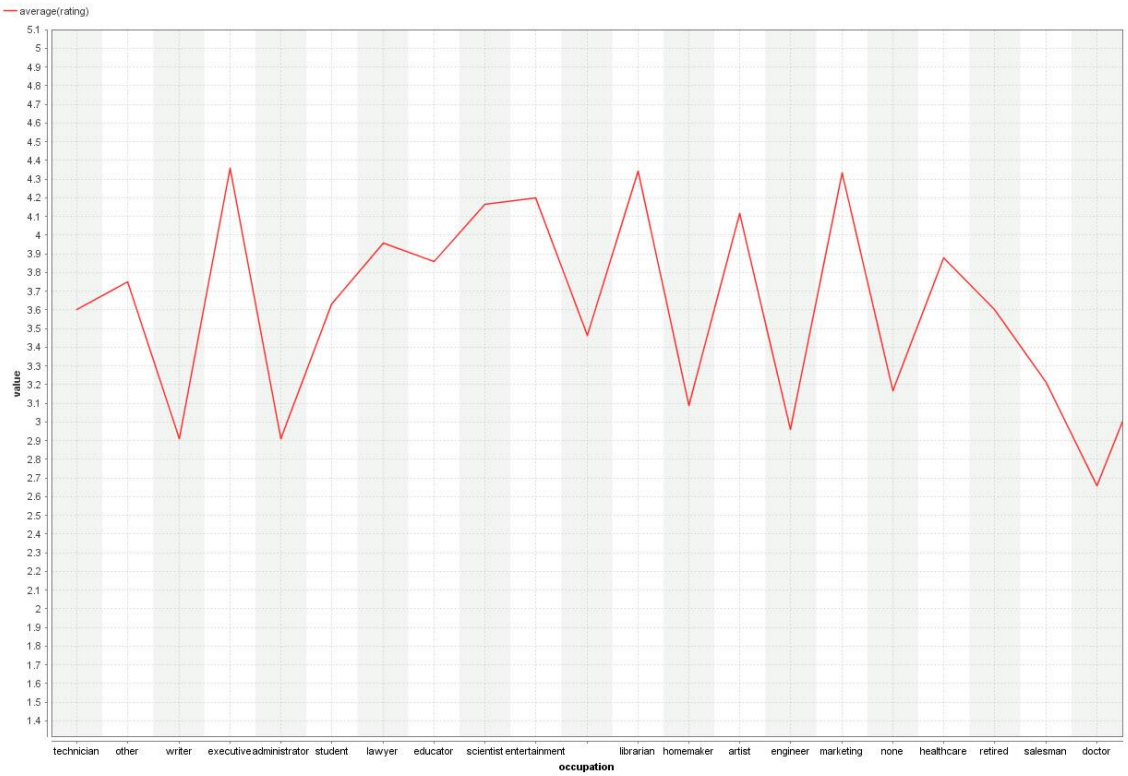


We can conclude from the graph that there are significantly more male users as compared to female users who rate movies. It could also be seen that only females tend to rate movies higher than 4.6 on an average.

Data Exploration: Ratings by occupation

From the output of the 'Inner join' we obtain for the users' information, we find that the occupation with label 'student' has the most number of users (See Appendix). Later, using a series plot of Average movie ratings against Occupations, we find the plot as shown below.

From the plot, we can see that users with the occupations 'Executive', 'Librarian', and 'Marketing' tend to rate the movies the highest. Users with the occupation described as 'Doctor' are the ones who tend to rate the movies the lowest on an average. However, we also observe that with the exception of 'Doctors', users tend to rate movies in the range of 3-4 stars, irrespective of their occupation.



Question No. 2)
Solution.

Collaborative filtering based rating prediction

We make use of the different collaborative filtering approaches found in the Collaborative Filtering based Item Recommendation (CFRP) operators. Different approaches such as Global average, User Item baseline, User k-NN, Item k-NN, and Matrix factorization operators would be used for predicting ratings.

Different measures that are often used to evaluate performance of collaborative filtering are Mean absolute error (MAE), Normalized MAE (NMAE) and Root mean square error (RMSE). Our primary focus would be on RMSE, since it implies a penalty factor for outliers and other variables trudging too far away from the mean. Higher error values will have a higher penalty factor and would also help us aggregate the errors.

Question No. 2) (a)
Solution.

Global Average method and User-Item Baseline methods

(A) Global Average Method

We first use the Global Average Method operator to predict movie ratings. The model is

applied to both the training and the test datasets and the errors mentioned earlier are taken into consideration. The summarization of the different error measures for this model are described in the Appendix. We find that there is only a variation of 0.004 in the RMSE between the Training and Test dataset. MAE and NMAE remain the same across both the training and the test dataset.

(B) User-Item Baseline Method

We now use User-Item Baseline model operator to predict movie ratings. As mentioned earlier, reducing RMSE is our primary concern and there were observations in changed RMSE by changing the parameters. We have changed for different values of number of iterations, regularization parameter for user biases and item biases. Changing the number of iterations didn't make a difference in the RMSE values. Decreasing the regularization parameters reduced the RMSE and increasing them increases the RMSE.

Thus, we find the optimal parameters which reduces the RMSE and the summarization of different error measures are described in the Appendix.

By comparing the error measurements across both the models, we find that User-Item Baseline method is better since it yields a lower error rate across RMSE, MAE and NMAE.

Question No. 2) (b)
Solution.

Matrix Factorization

(A) Varying Number of Factors

We perform exploratory analysis of the effects of varying the number of factors on the performance measurement. This analysis is performed by varying the Number of factors and keeping the rest of the parameters as constant. We find that increasing the Number of factors decreases error in the Training dataset, but increases the error in the Testing dataset.

The table summarizing the effects of varying number of factors on the performance measures is described in the Appendix.

(B) Varying Learning Rate

Next, we perform the same procedure as mentioned above with the Learning Rate attribute. This analysis is performed by varying the Learning Rate and keeping the rest of the parameters as constant. We find that increasing the Learning Rate decreases error in the Training dataset, but increases the error in the Testing dataset. The table summarizing the effects of varying number of factors on the performance measures is described in the Appendix.

(C) Optimal value of Parameters

To compensate for the overfitting of the model, we also tried varying the Regularization parameter. After performing the said analysis, the table given below represents the best parameters which yield us the best results, with the remaining parameters set to their defaults.

Matrix Factorization Parameters	Associated Label or Value
Number of Factors	5
Learning Rate	0.01
Regularization	0.03

The final error report for Matrix Performance operator is described in the Appendix.

Question No. 2) (c)
Solution.

User k-NN and Item k-NN

(A) User k-NN

Next we do the analysis using the User k-NN model operator. Here, we have measured the performance for different values of 'k' across Cosine similarity measure and Pearson measure. The values of 'k' are varied, by keeping all of the remaining parameters as constant to their defaults. We find that increasing the value of 'k' reduces the error rate in test dataset only up to an extent of 100. Any further increase does not yield us significant improvement in performance measures. We also find that Pearson measure yields us a better performance measure as compared to the Cosine similarity measure.

The performance measure report for different values of 'k' and distance measuring methods are summarized in the Appendix.

(B) Item k-NN

Similarly, we do the analysis using the Item k-NN model operator. We have measured the performance for different values of 'k' across Cosine similarity measure and Pearson measure. The values of 'k' are varied, by keeping all of the remaining parameters as constant to their defaults. We find that increasing the value of 'k' reduces the error rate in test dataset only up to an extent of 100. Any further increase does not yield us significant improvement in performance measures. We also find that Pearson measure yields us a better performance measure as compared to the Cosine similarity measure.

The performance measure report for different values of 'k' and distance measuring methods are summarized in the Appendix.

Evaluative comparison of the different Best models

The best models from each of the five collaborative filtering models are being summarized below for an overall picture of how these models perform against each other.

Evaluative comparison of the different models						
Model	**RMSE**		**MAE**		**NMAE**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
Global Average	1.126	1.122	0.945	0.945	0.236	0.236
User Item Baseline	0.913	0.958	0.722	0.757	0.180	0.189
Matrix Factorization	0.822	0.958	0.647	0.755	0.162	0.189
User k-NN	0.770	0.949	0.598	0.746	0.149	0.187
Item k-NN	0.696	0.938	0.542	0.735	0.135	0.184

From the above table, we could conclude that Item k-NN is the Collaborative Filtering model which yields us a better model as compared to the other models. Since Item k-NN collaborative filtering model yields us the lowest error rate across all possible performance measurement metrics, it is our best model.

Question No. 3)
Solution.

Determining the 'Best Model'

Now, our primary goal here is to recommend movies to users with a 'high' value of prediction. For this purpose, we need to determine a 'Best model', based on its performance. We make use of the overall accuracy, true class precision and the true class recall to make this informed-decision, whether to recommend a particular movie to a user or not. Our goal is to increase the number of movies that we recommend (recall) and to ensure that these movies will match the taste of the user (precision). The true class recall is our primary area of concern here, but care must also be taken so that true class precision is not compromised to a large extent.

Measuring recall requires knowing whether each item is relevant; for a movie recommender, this would involve asking many users to view all movies to measure how successfully we recommend each one to each user. Similarly, this approximation to precision suffers from the same biases as the recall approximation. [1] In the "recommend some good items" task it is likely that we will prefer a system with a high precision, while in the "recommend all good items" task, a higher recall rate is more important than precision. Typically, we can expect a trade-off between these quantities—while allowing longer recommendation lists typically improves recall, it is also likely to reduce the precision. [2]

To determine our 'Best model', we have we run all of our best models to obtain individual performance metrics across a single value of cut-off first of 4-star rating or higher. Below is the summarization.

Comparing Performance of the different models			
Model	**Overall Accuracy (%)**	**True Class Recall (%)**	**True Class Precision (%)**
Global Average	42%	0%	0%
User Item Baseline	59.08%	36.86%	83.24%
Matrix Factorization	59.83%	37.92%	84.07%
User k-NN	62.28%	43.74%	83.29%
Item k-NN	62.40%	42.82%	84.82%

To further confirm our best model, we have considered User k-NN and Item k-NN (Both, which yield better true class recall accuracy) and tried for different values of 'cut-off'. This summarization table is provided in the Appendix and we find that User k-NN yields us a better model for every cut-off value in terms of True class recall accuracy.

Although User k-NN and Item k-NN yield us almost similar results, we choose 'User k-NN' as our best model. True Class recall (sensitivity) is the fraction of relevant instances that are retrieved. As the probability that a relevant document is retrieved by the query is important here, we choose to go with the True class recall for selecting our 'Best Model'. Moreover, the User k-NN model does not exhibit significant decrease in the True class precision value.

Determining the Optimal 'Cut-off'

Now, we have to consider the optimal value of 'Cut-off', above which we will be recommending the movie to a user. The performance results that we obtained earlier are for a fixed 'cut-off' of a 4-star or higher rating prediction. To further fine-tune our model, we have to find the optimal value of 'cut-off' for which we obtain the best performance.

Comparing Performance of the different Cut-off values			
Cut-Off Value	**Overall Accuracy (%)**	**True Class Recall (%)**	**True Class Precision (%)**
4	62.28%	43.74%	83.29%
3.95	63.83%	47.72%	82.54%
3.9	64.95%	51.16%	81.53%
3.85	65.95%	54.51%	80.48%
3.8	66.85%	57.80%	79.44%
3.75	67.71%	61.16%	78.41%
3.7	68.63%	64.64%	77.54%
3.5	70.36%	75.81%	73.80%
3.415	70.57%	80.12%	72.19%

Hence, we choose a cut-off of 3.415 average movie rating. A cut-off of 3.415 will yield us an approximate true recall accuracy of 80.12% and a relatively decent precision level of 72.19%.

Distribution of Errors across Users and Movies

(A) Errors across Users

Next, we have to consider the distribution of errors across Users. For this, we obtain newly generated attributes such as the 'Error' which is a difference between the prediction and the rating, the 'Squared Error' and the 'Absolute Error'. All of these errors are aggregated across each individual user to get an overall distribution of errors across each user.

Key Findings: The error distribution across users is a normally distributed graph. From the graph, it is safe to say that the average of errors is distributed normally across users. The squared error distribution is right-skewed across users and the absolute error distribution is slightly right-skewed.

(B) Errors across Movies

Similarly, we have to consider the distribution of errors across Movies. For this, we obtain newly generated attributes such as the 'Error' which is a difference between the prediction and the rating, the 'Squared Error' and the 'Absolute Error'. All of these errors are aggregated across each individual movie listing to get an overall distribution of errors across each user.

Key Findings: The error distribution across movies is a normally distributed graph. From the graph, it is safe to say that the average of errors is distributed normally across movies. The squared error distribution is heavily right-skewed across movies and the absolute error distribution is slightly right-skewed.

By observing the error distribution graphs, it could be said that errors are equally distributed across users and movies.

The relevant graphs of all the error distributions are shown in the Appendix.

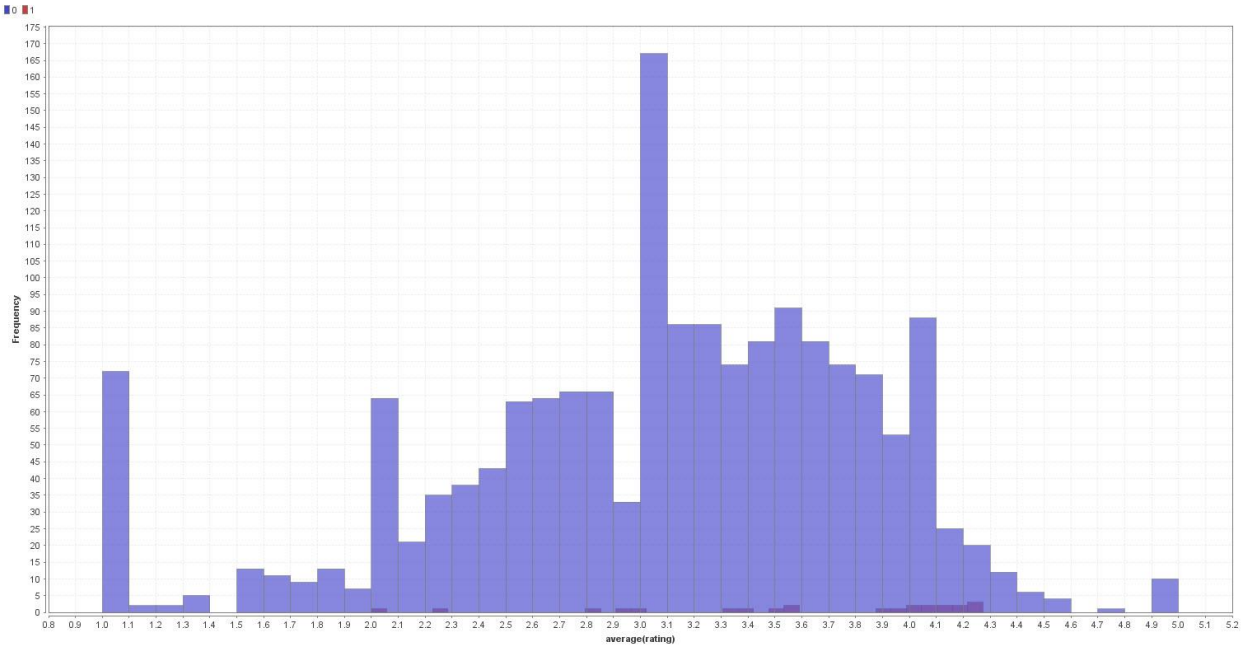
APPENDIX

Question No. 1 (b) (i)

Ratings by Genre for the Top-5 Genres

Row No.	Film-Noir	count(Film-N...	average(ave...
1	0	1657	3.059
2	1	23	3.647

Average Movie Rating for Film-Noir Genre



Distribution of Movie ratings for Film-Noir genre

Row No.	War	count(War)	average(ave...
1	0	1609	3.049
2	1	71	3.473

Row No.	Musical	count(Music...	average(ave...
1	0	1624	3.056
2	1	56	3.372

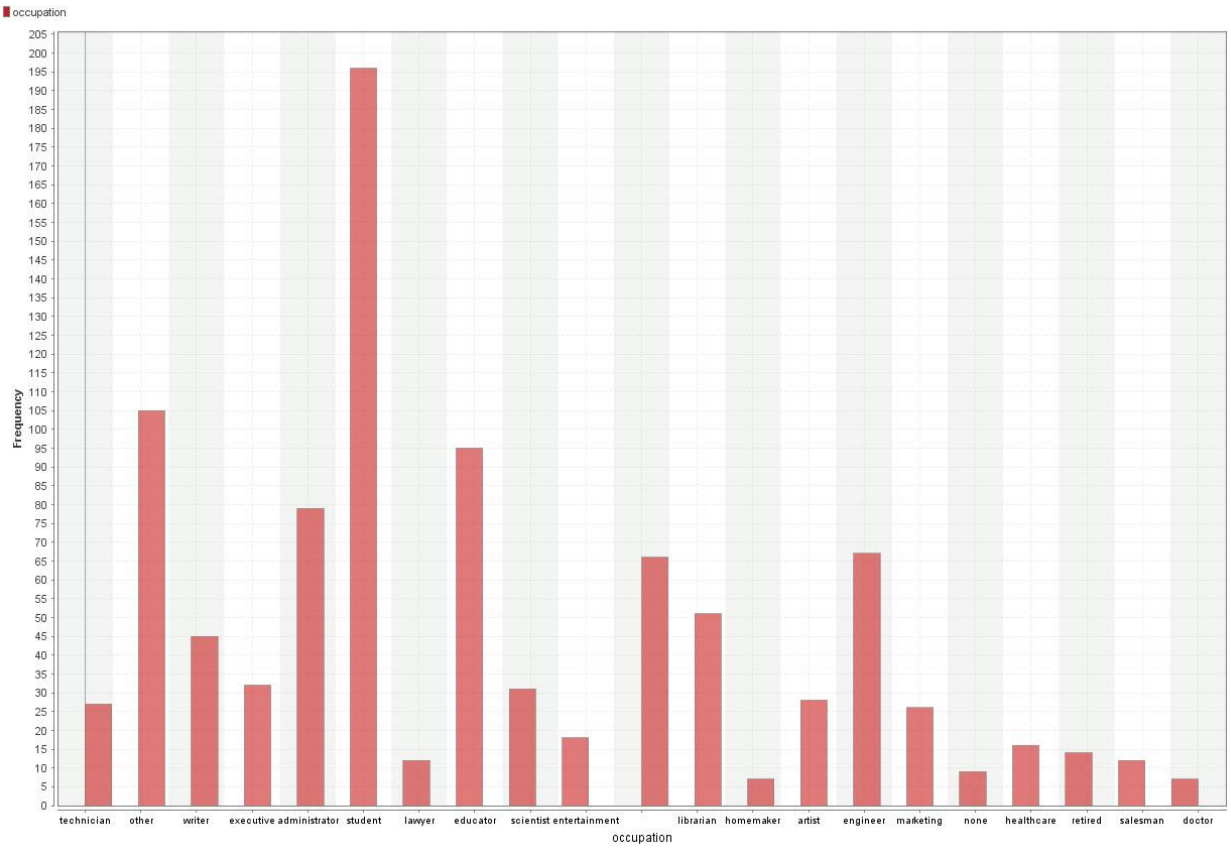
Row No.	Animation	count(Anim...	average(ave...
1	0	1638	3.061
2	1	42	3.304

Row No.	Mystery	count(Myste...	average(ave...
1	0	1619	3.057
2	1	61	3.330

Distribution of Movie ratings for the next top Movie Genres

Question No. 1 (b) (ii)

Ratings by Occupation



Distribution of Movie Ratings across different Occupations

Question No. 2 (a)

Global Average Method Error Report		
Error Measurement	**Training Dataset**	**Test Dataset**
RMSE	1.126	1.122
MAE	0.945	0.945
NMAE	0.236	0.236

User-Item Baseline Parameters	Associated Label or Value
Number of Iterations	5
Regularization parameter for User biases	3.0
Regularization parameter for Item biases	5.0

User-Item Baseline Method Error Report		
Error Measurement	**Training Dataset**	**Test Dataset**
RMSE	0.913	0.958
MAE	0.722	0.757
NMAE	0.180	0.189

Question No. 2 (b)

Matrix Factorization – Varying the Number of factors						
Number of Factors	**RMSE**		**MAE**		**NMAE**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
5	0.814	0.954	0.639	0.754	0.160	0.188
10	0.743	0.972	0.582	0.764	0.146	0.191
15	0.685	0.996	0.537	0.776	0.134	0.194
18	0.657	1.012	0.514	0.791	0.129	0.198
20	0.635	1.000	0.497	0.784	0.124	0.196

Matrix Factorization – Varying the Learning Rate						
Learning Rate	**RMSE**		**MAE**		**NMAE**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
0.01	0.816	0.957	0.640	0.754	0.160	0.189
0.02	0.805	0.973	0.631	0.763	0.158	0.191
0.03	0.798	0.968	0.624	0.756	0.156	0.189

Matrix Factorization Final Error Report		
Error Measurement	**Training Dataset**	**Test Dataset**
RMSE	0.822	0.958
MAE	0.647	0.755
NMAE	0.162	0.189

Question No. 2 (c)

User k-NN Error Report (Cosine Correlation Measure)						
Value of K	RMSE		MAE		NMAE	
	Training	Testing	Training	Testing	Training	Testing
25	0.924	0.959	0.725	0.756	0.181	0.189
50	0.922	0.957	0.724	0.754	0.181	0.189
80	0.923	0.957	0.725	0.754	0.181	0.189
100	0.924	0.957	0.726	0.755	0.182	0.189
120	0.924	0.958	0.727	0.755	0.182	0.189

User k-NN Error Report (Pearson Measure)						
Value of K	RMSE		MAE		NMAE	
	Training	Testing	Training	Testing	Training	Testing
25	0.770	0.949	0.598	0.746	0.149	0.187
50	0.790	0.949	0.615	0.745	0.154	0.186
80	0.804	0.949	0.627	0.746	0.157	0.186
100	0.809	0.949	0.632	0.746	0.158	0.187
120	0.813	0.950	0.635	0.746	0.159	0.187

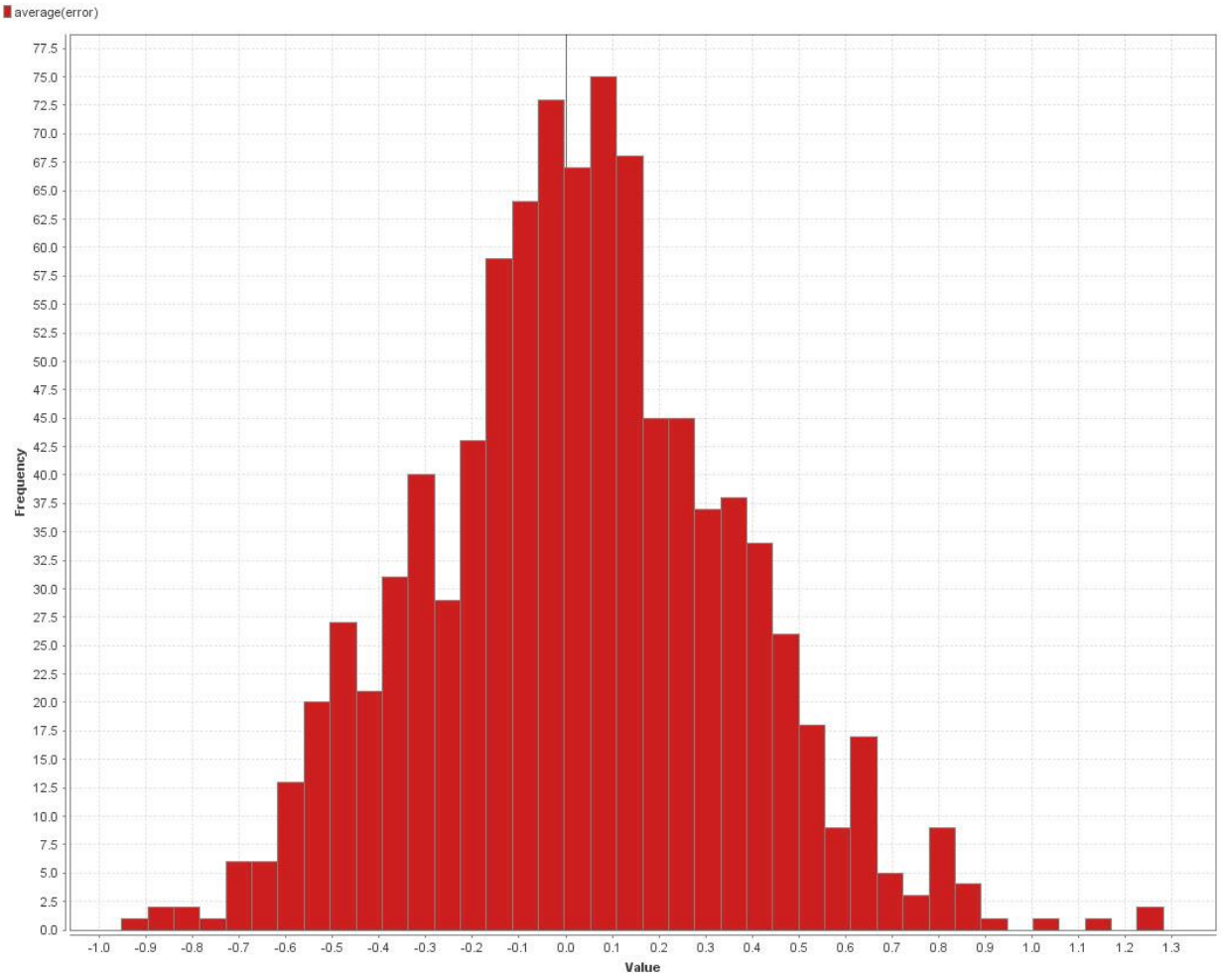
Item k-NN Error Report (Cosine Correlation Measure)						
Value of K	RMSE		MAE		NMAE	
	Training	Testing	Training	Testing	Training	Testing
25	0.894	0.940	0.700	0.738	0.175	0.184
50	0.898	0.943	0.704	0.741	0.176	0.185
80	0.902	0.945	0.709	0.743	0.177	0.186
100	0.904	0.947	0.711	0.745	0.178	0.186
120	0.905	0.947	0.712	0.746	0.178	0.186

Item k-NN Error Report (Pearson Measure)						
Value of K	RMSE		MAE		NMAE	
	Training	Testing	Training	Testing	Training	Testing
25	0.696	0.938	0.542	0.735	0.135	0.184
50	0.725	0.937	0.566	0.736	0.141	0.184
80	0.745	0.938	0.583	0.736	0.146	0.184
100	0.754	0.939	0.591	0.737	0.148	0.184
120	0.760	0.939	0.596	0.737	0.149	0.184

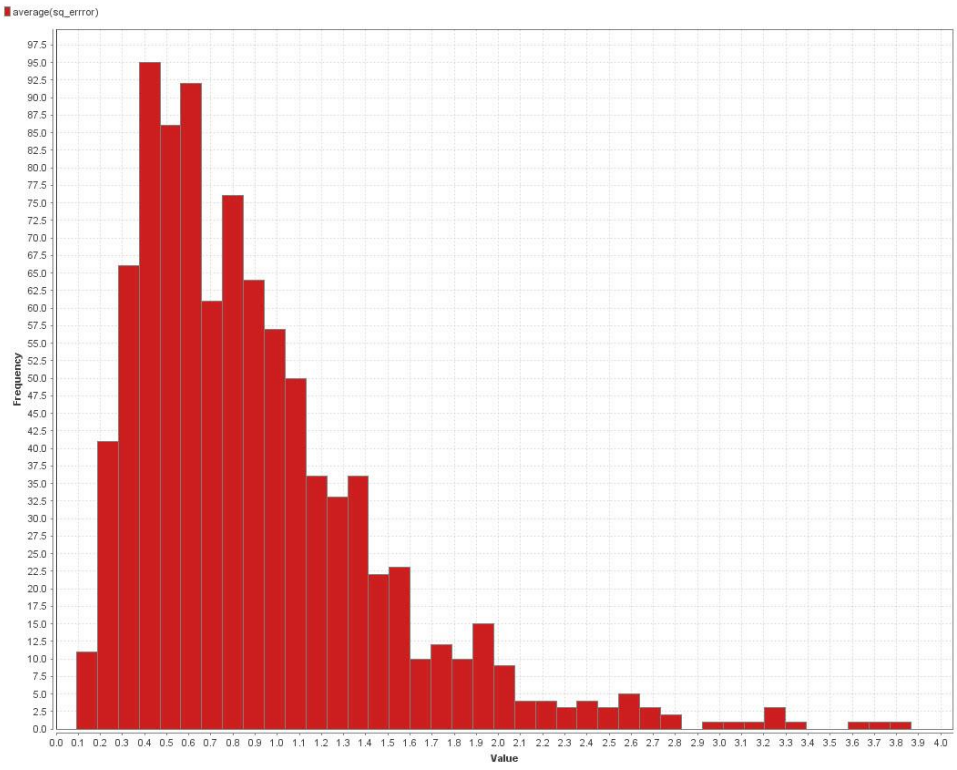
Question No. 3)

Comparing Performance of User and Item k-NN						
Cut-Off Values	**Overall Accuracy**		**True Class Recall**		**True Class Precision**	
	User k-NN	**Item k-NN**	**User k-NN**	**Item k-NN**	**User k-NN**	**Item k-NN**
4	62.28%	62.40%	43.74%	42.82%	83.29%	84.82%
3.95	63.83%	63.71%	47.72%	46.63%	82.54%	83.52%
3.9	64.95%	64.97%	51.16%	50.06%	81.53%	82.72%
3.85	65.95%	66.21%	54.51%	53.65%	80.48%	81.84%

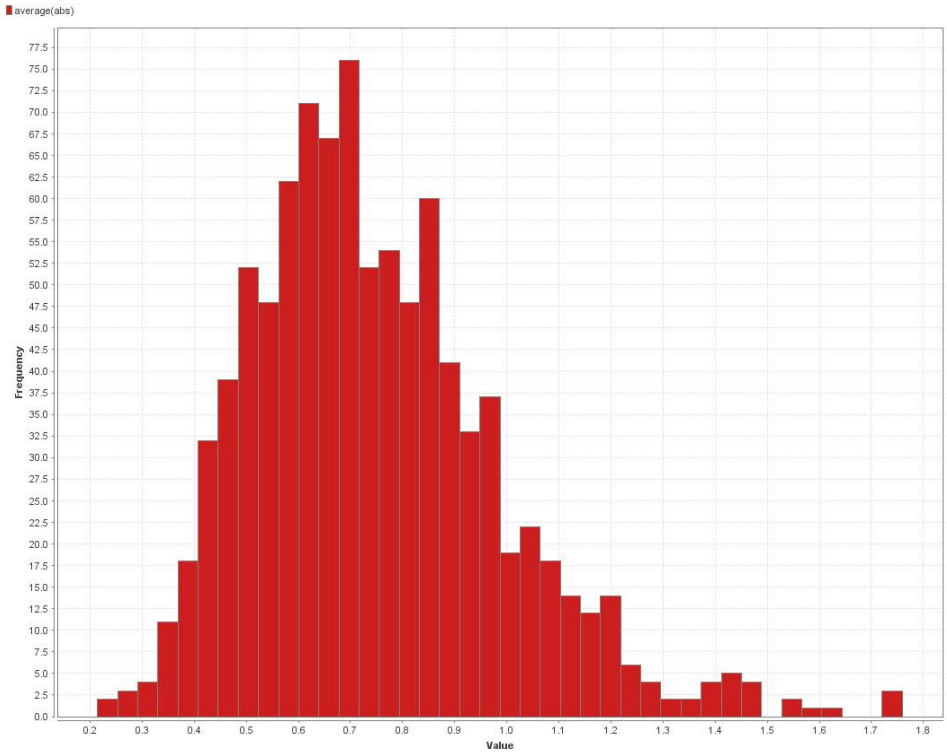
Average Error Distribution by Users



Average Error Rate

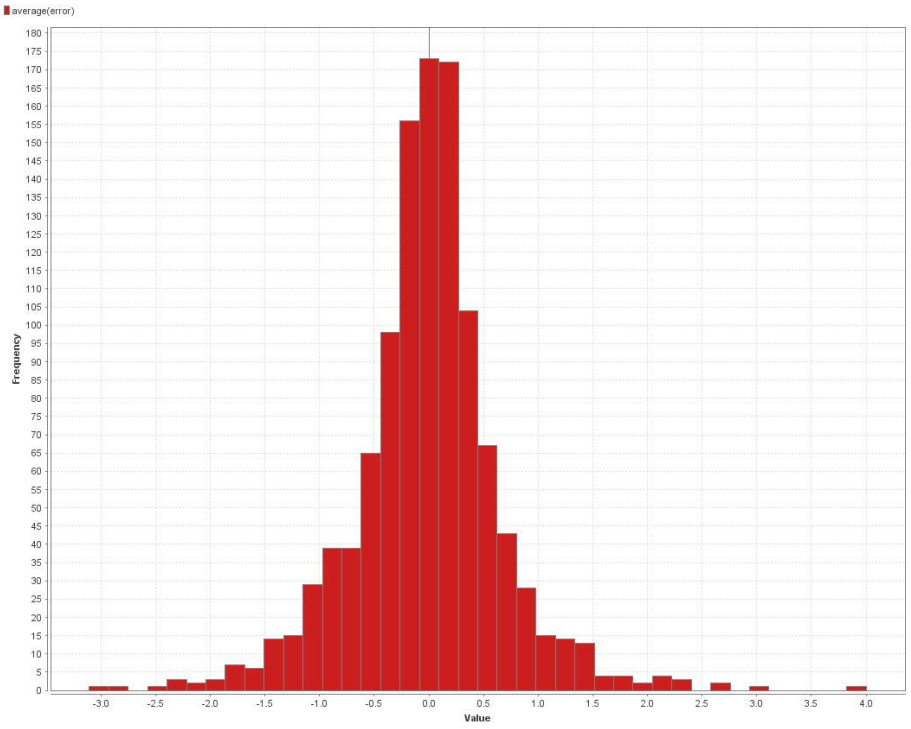


Average Squared Error

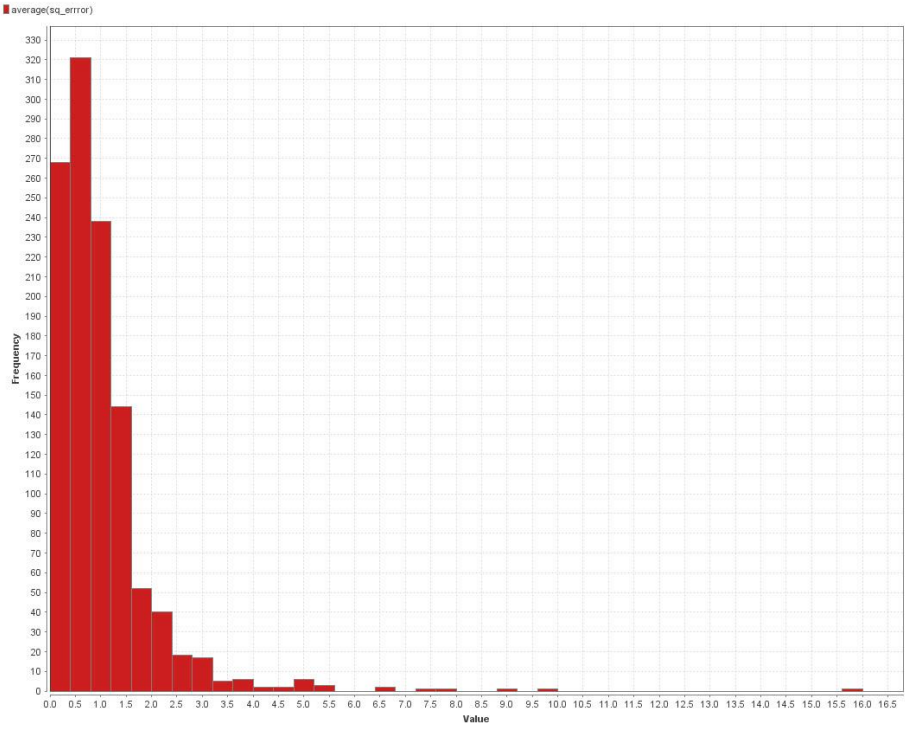


Average Absolute Error

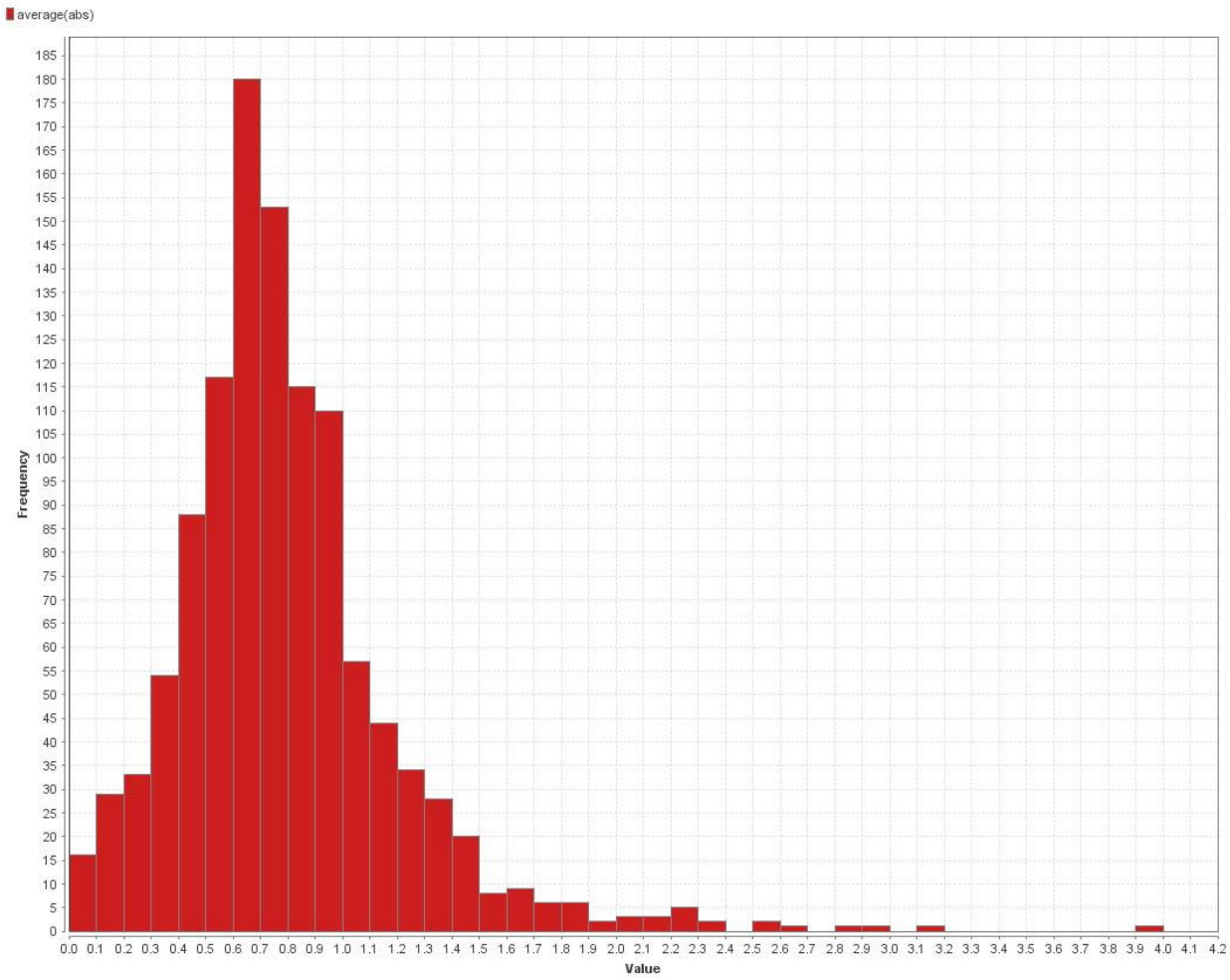
Average Error Distribution by Movies



Average Error Rate



Average Squared Error Rate



Average Absolute Error Rate