IDS 572: Extra Assignment

Recommender system

Archan Patel - 661105271 - apate381@uic.edu

Jasbir Singh - 651837003 - jasingh3@uic.edu

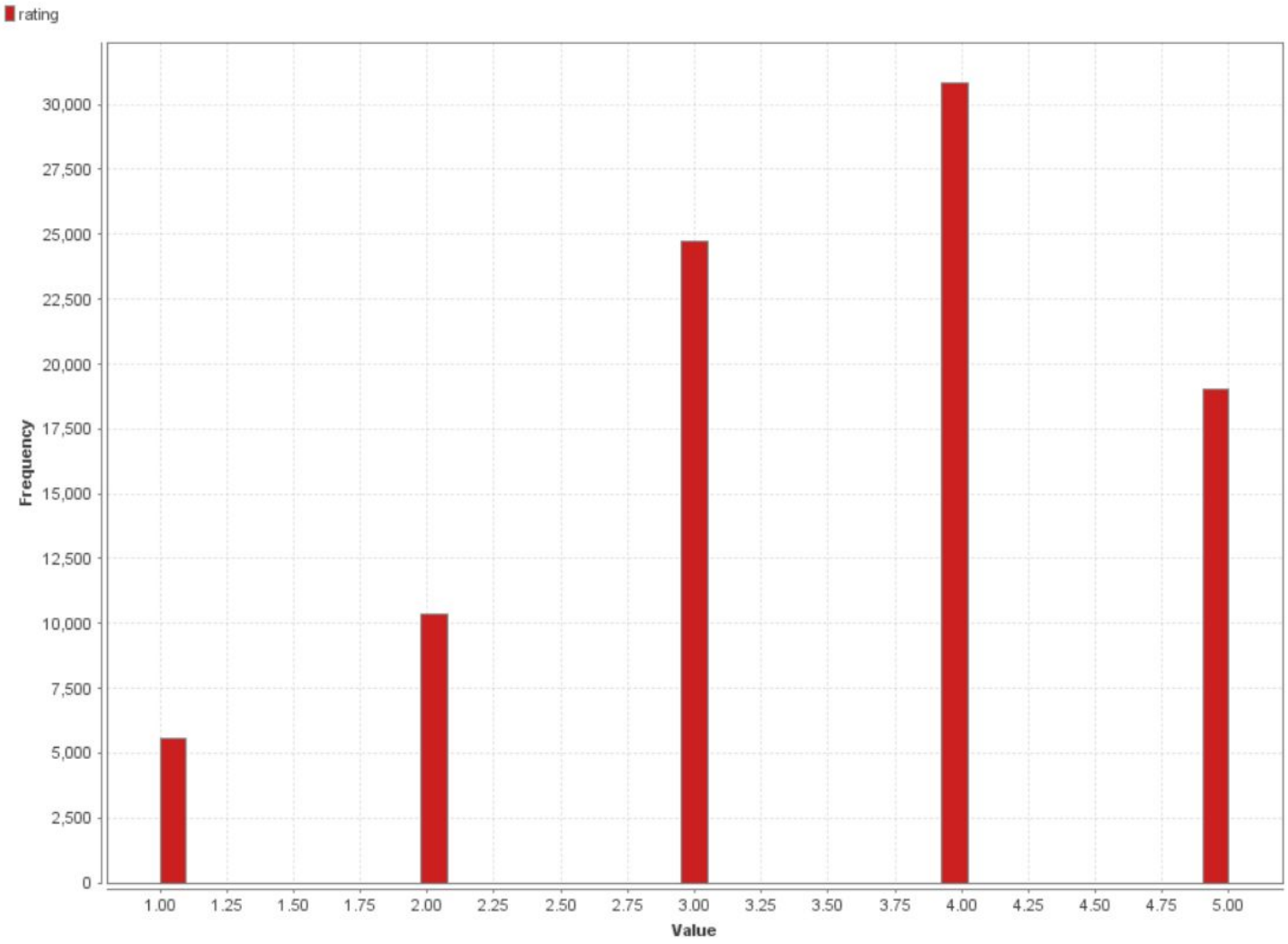
Jeetyog Rangnekar - 656052696 - jrangn2@uic.edu

Assignment Question 1

(a)

What is the overall distribution of ratings

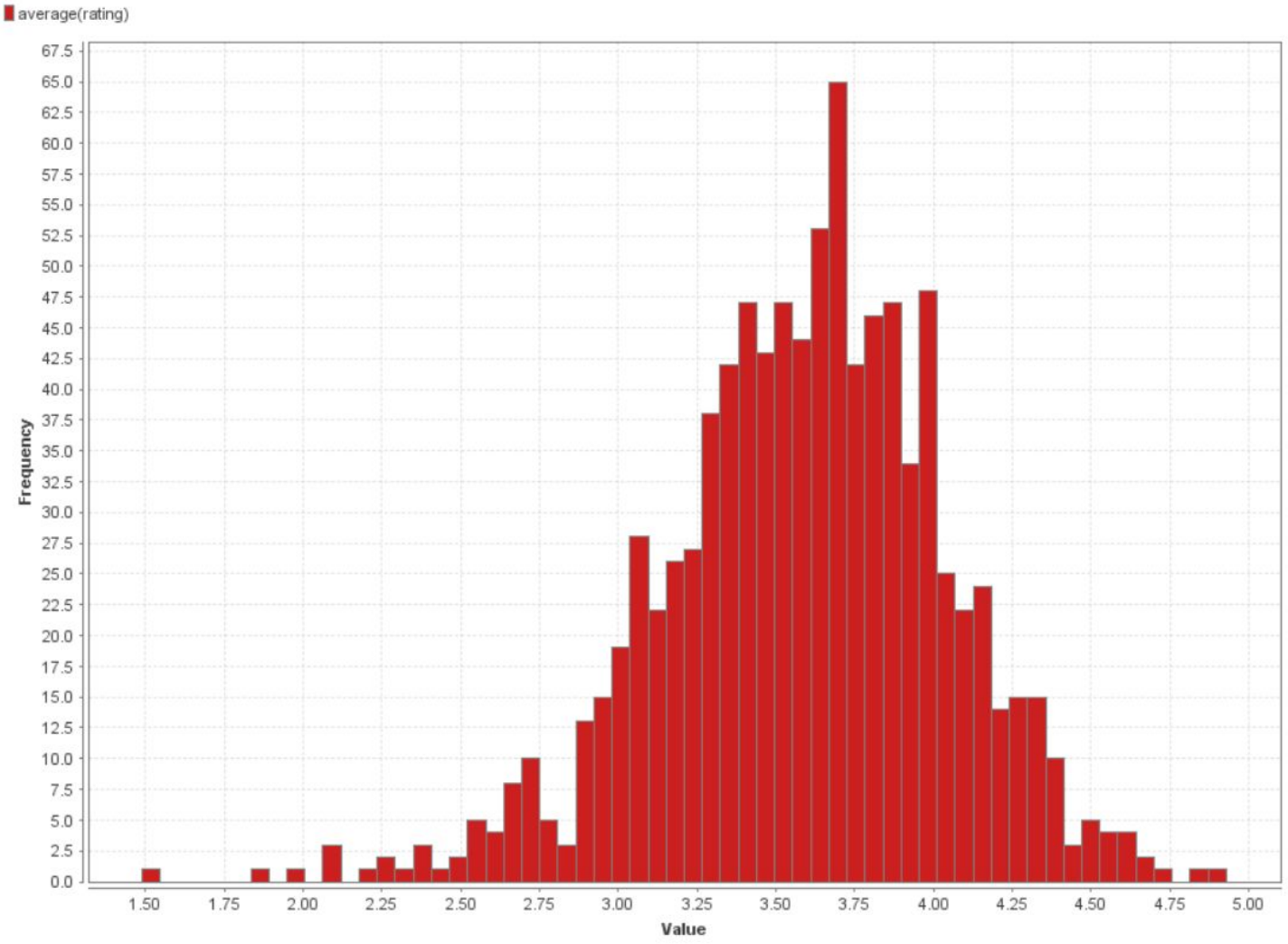
- Below figure displays the overall distribution of ratings.



- It can be seen that most of the users have given 3 or higher ratings to movies. More than 30,000 records have rating 4, which is near to ⅓ of all the records. While rating 1 has been given for the lowest amount of time - around 5500 which is 6% of all the records.
- The average rating is 3.524.

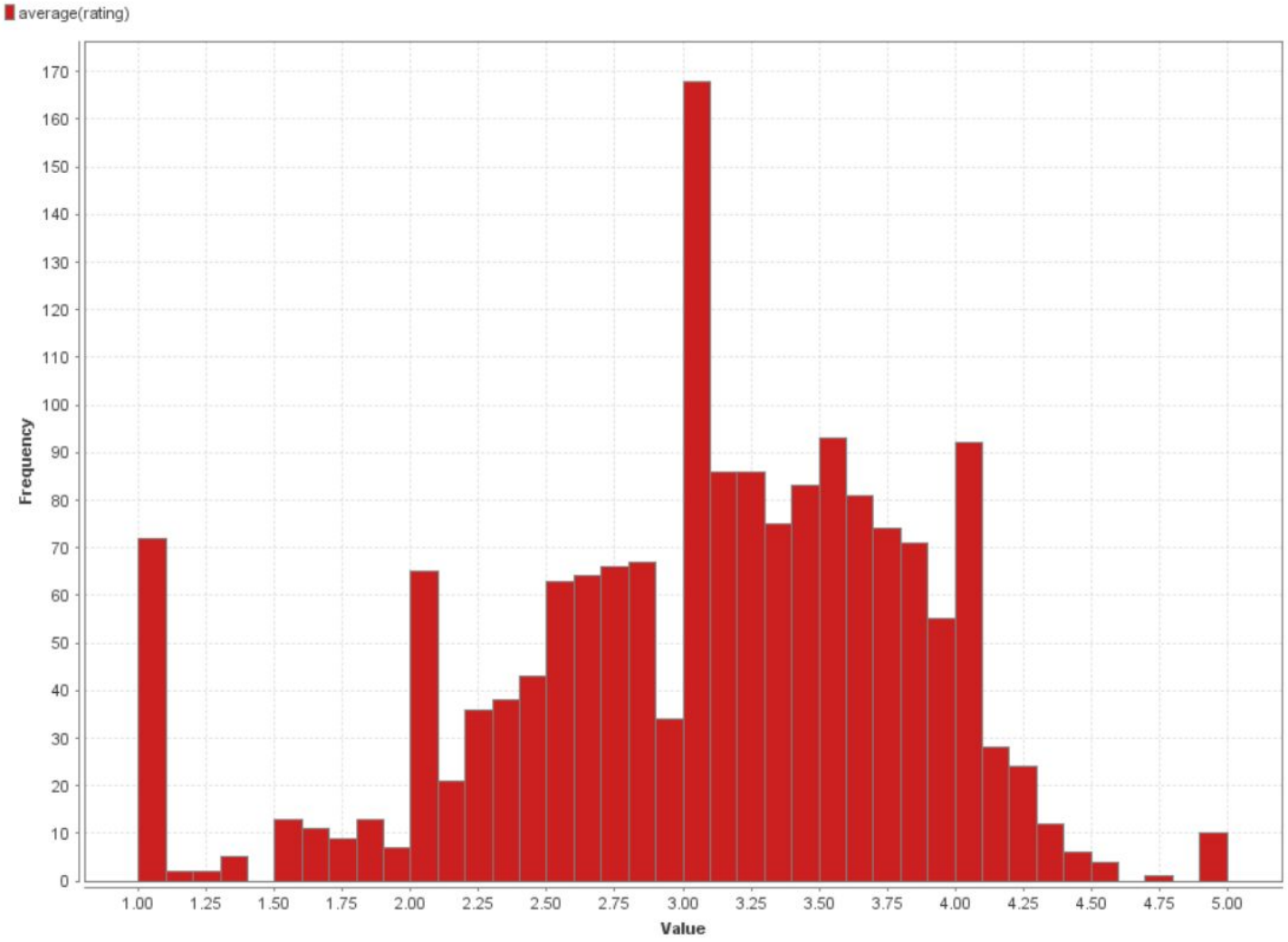
On average, how do users rate movies; what ratings do movies have on average?

- Below figure displays the distribution of average rating of 943 users.



  - It can be seen that the distribution has slightly left skewed bell shaped curve. Most of the user's average ratings lies within the range of 3 to 4. Very few users rate maximum & minimum ratings.

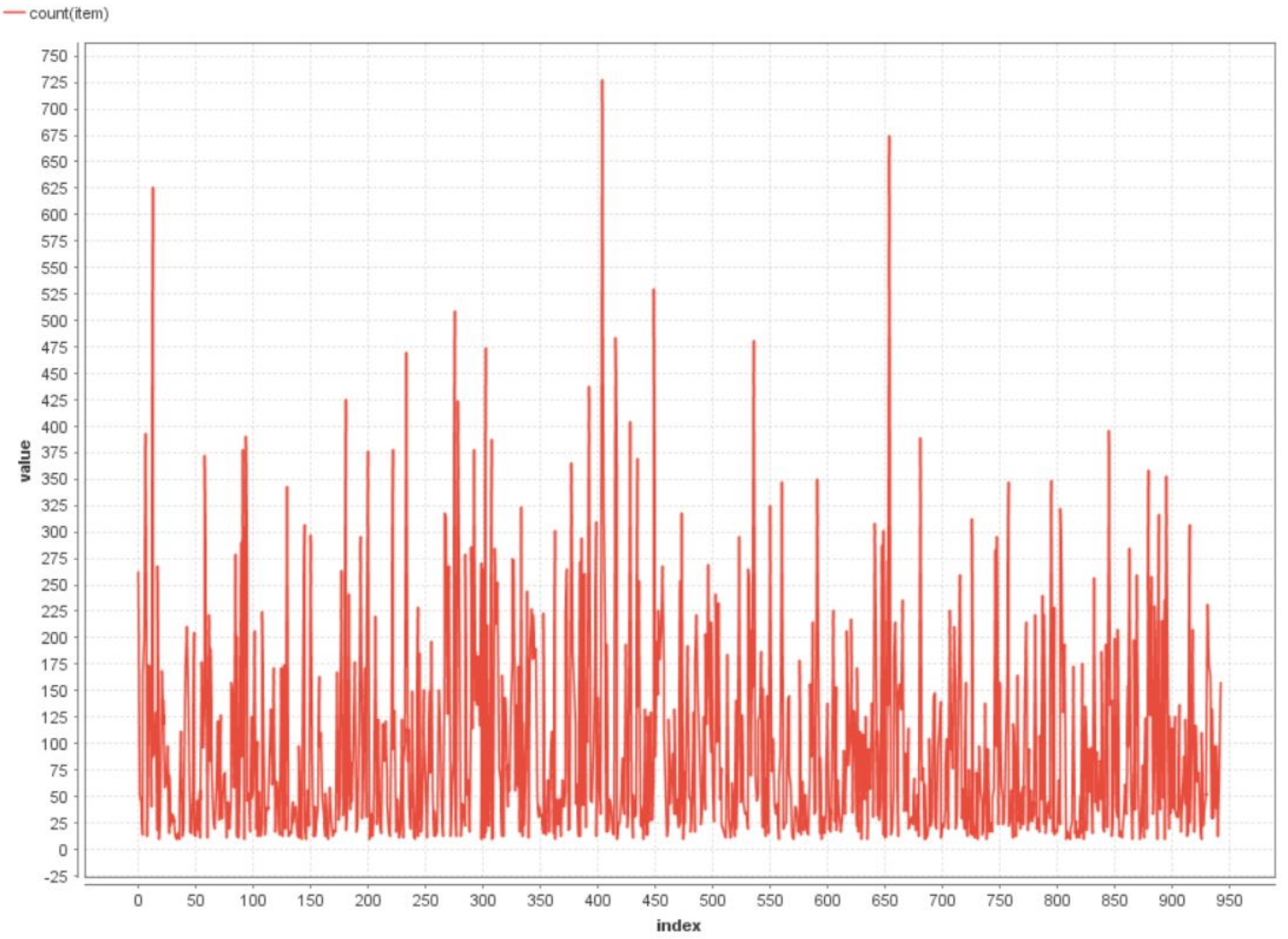
- Below figure displays the distribution of 1682 movies as per their average rating.



  - It can be seen that most movies have ratings between 2.5 to 4.0. Except 2 instances, where more than 50 movies have average rating of 1.0 & 2.0
  - There are 10 movies who has the average rating of 4.9 - which is the highest. While, on the other hand there are around 72 movies with the lowest average rating of 1.0.
  - Around 169 movies has the average rating of 3.0

How many movies do users rate, and how many ratings do movies get?

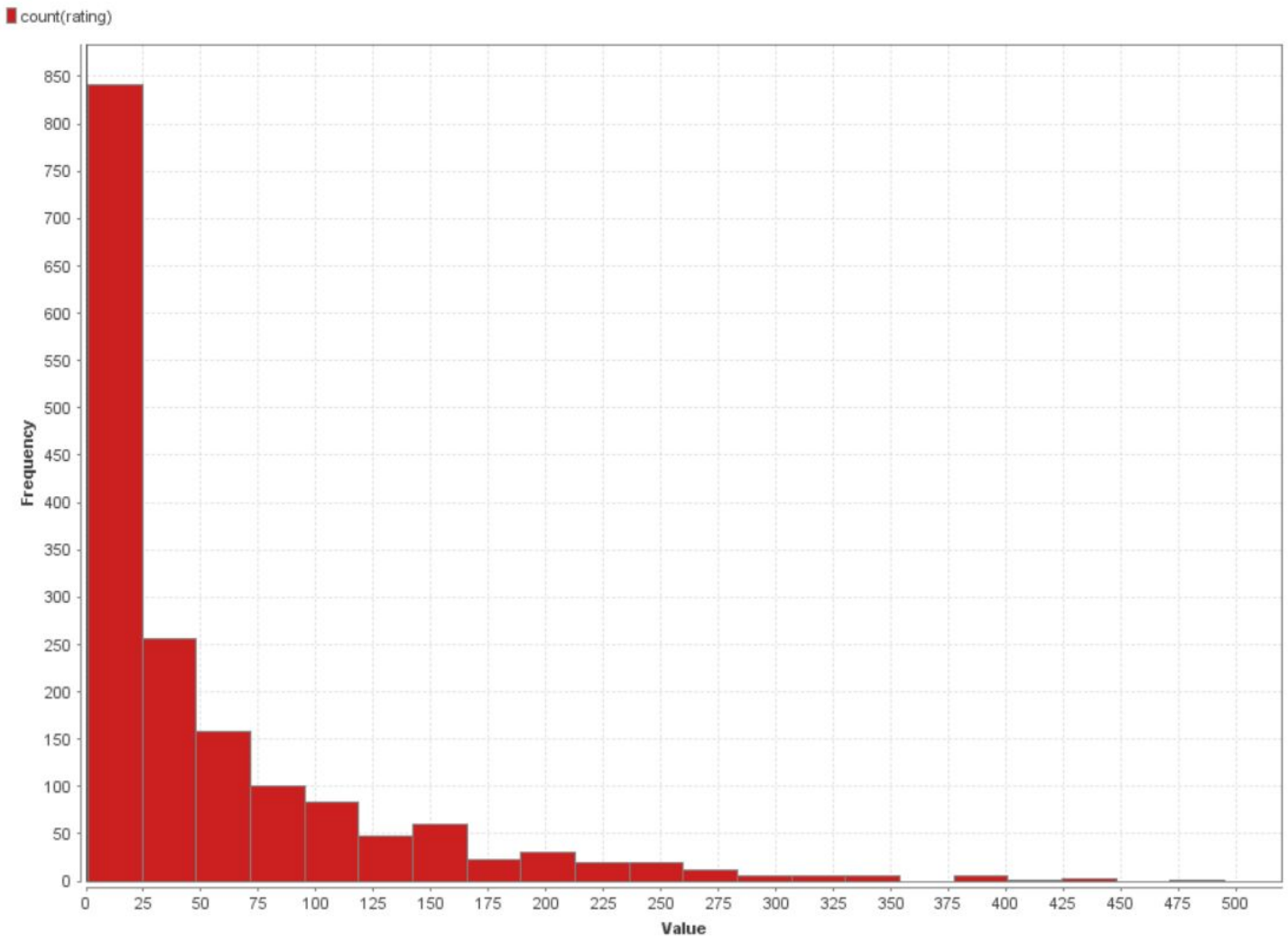
- Below images display the distribution of number of movies rated by users.



count(item)	Integer	0	Min 10	Max 727	Average 96.045

  - On average, a user rates 96 movies. One of the user has rated 727 movies which is the highest count of rated movies by a user, while the lowest count is 10.

- Below figure displays the distribution of count of ratings by movies.



  - It can be seen that the distribution is right skewed. Most of the movies get around 1-25 number of reviews. Very few movies get large number of reviews.

			Min	Max	Average
count(rating)	Integer	0	1	495	53.911

  - On average, a movie gets around 54 reviews. Highest number of reviews received for a movie is 495, while the lowest is 1.

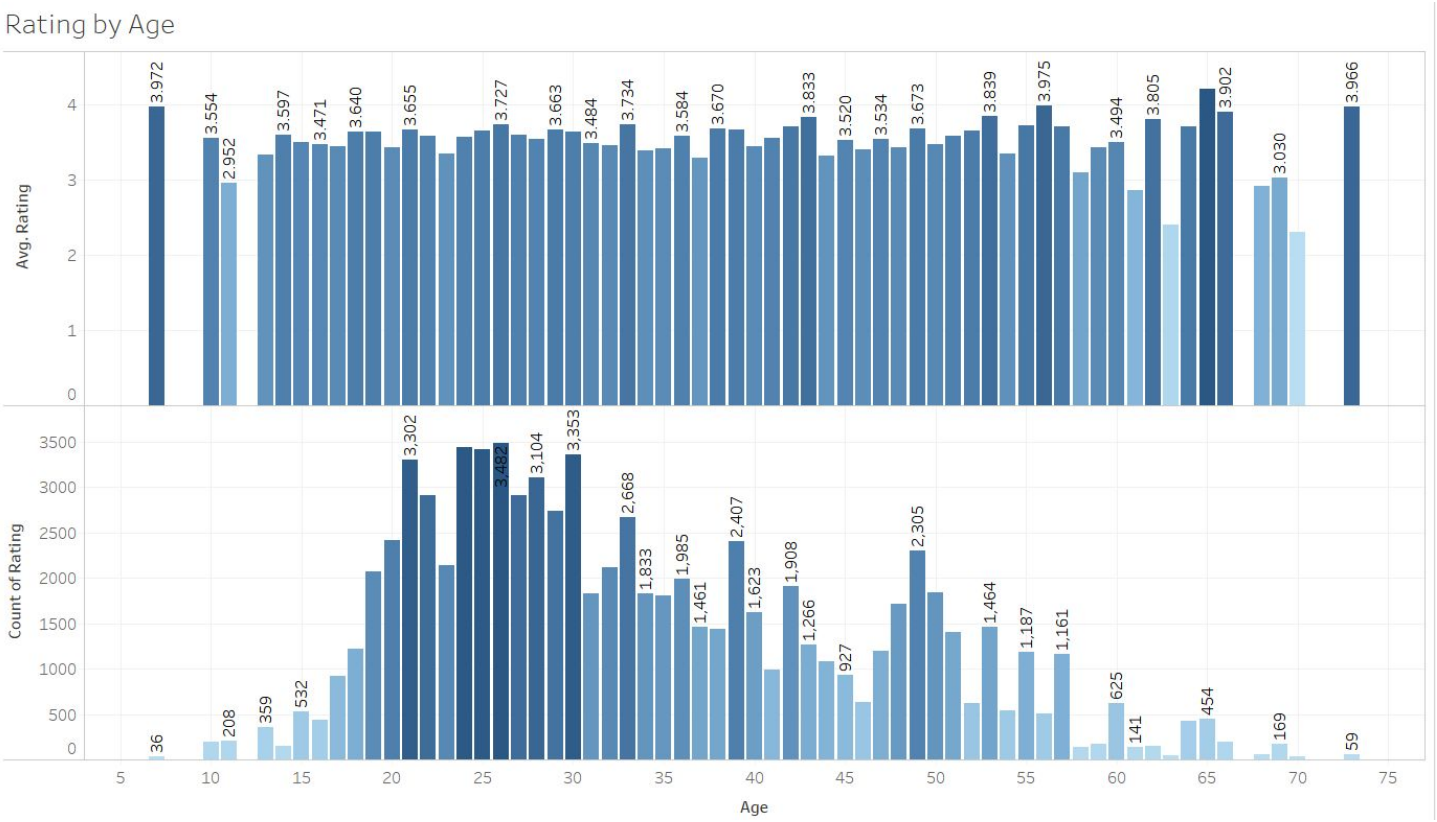
How are rating levels distributed, do many people have high/low ratings?

- The distribution of the ratings is plotted earlier in the report and it was observed that, most of the users give rating of 3 or 4. Very few users gives the maximum & minimum ratings.
- Around 90% of the users have given a minimum rating of 3, while around 15.5% of the users have given a minimum rating of 4.

(b)

(Graphs below have been created in Tableau)

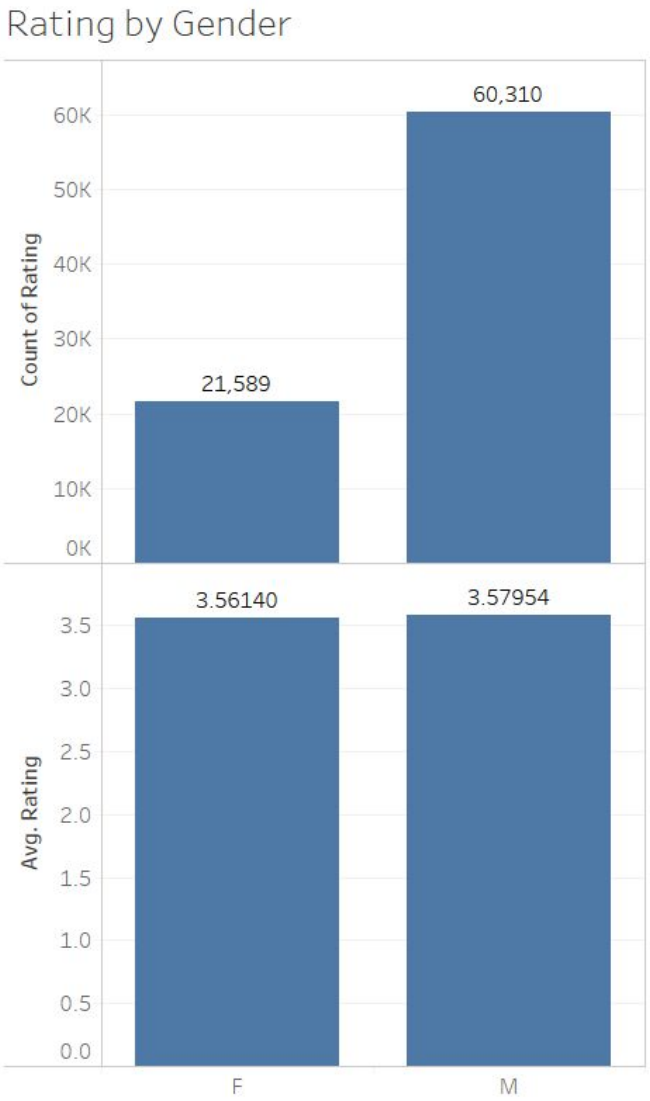
- **Ratings by Age**
    - Below image displays the distribution of ratings as per the age.



    - We can see that users between Age 20 to 30 tends to give ratings to more number of movies. The graph is more or less right skewed which means, elder people tends to give ratings to less number of movies.
    - Regarding the average ratings as per age, there is no significant difference in the average ratings among different age groups. Lowest average rating has been observed among the users of Age 70, which is 2.303

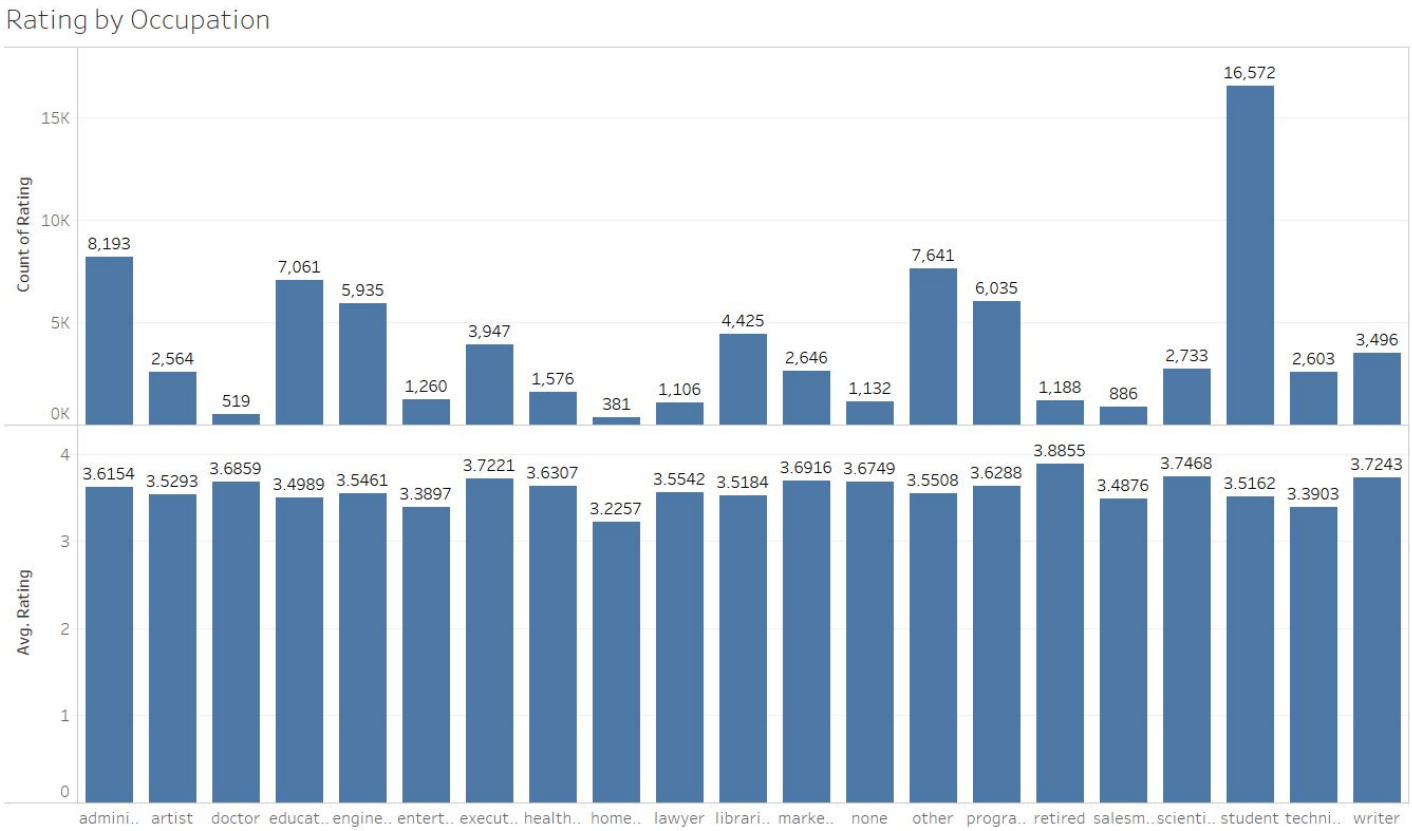
      While, the highest has been observed among the users of age 65, which is 4.5. But both of these users have lower number of rating count.
    - Users with age between 20-30, who gave the ratings to highest number of movies, have an average rating of around 3.5.

- **Ratings by Gender**



  - We can see that there is not much difference in terms of average rating between Males & Females.
  - However, when it comes to rating the movie, Males seems to rate more number of movies than Females - the count is almost triple for Males.

- **Ratings by Occupation**



  - We can see that average rating is consistent among different occupations. The highest average rating is observed for retired & lowest for homemakers.
  - However, when it comes to rating a movie, students seem to rate highest number of movies. 20% of all ratings are provided by students. While, homemakers tend to rate lower number of movies.
  - Homemakers also have the least average rating among all occupations as stated above - which means that they rate low number of movies and tends to give low ratings to movies.

- **Ratings by Genre**
  - Since, we do not have a separate Genre dataset and one movie has multiple Genre assigned to it - we found the average of rating and count of rating for each genre. Based on that, here are the findings.
    - Highest number of ratings is for genre Drama, while the lowest is for Documentary. Drama also has high average rating.
    - Highest average rating is for genre Film-Noir, while the lowest is for Fantasy.

Assignment Question 2

- For accessing the performance, we can take Error rate into consideration. Because the aim here is to predict the rating which is closest to the actual rating. So error rate becomes critical as we don't want to have a predicted rating too far from the actual one.
    - For example, consider a true movie rating to be 4.0 and predicted ratings to be 3.9, 4.1, 4.3 etc.
    - By accessing the performance using accuracy, none of this prediction is correct. But, error rate will say that 4.1 is closer to the actual rating than the other ones. The problem statement here deals with predicting the rating closest to the actual one - not whether a user will provide rating to a movie or not. Hence, error rate seems to be a reasonable criteria here for accessing the performance.

(a)

- Optimize parameters operator with Global average method gave the same results for different value of Min Rating & Range. Here is the summary of the performance of the Global average method. (Min Rating =2, Range = 5)

	RMSE	**MAE**	**NMAE**
Training	1.126	0.945	0.189
Testing	1.122	0.945	0.189

    - We can see that RMSE value is slightly different for training & testing dataset - while MAE & NMAE remains same for both.
    - It is interesting to notice that RMSE for test data is less than RMSE for training data.

- Using optimize parameter operator with User Item Baseline, below is the summary of the best performing model.
  Parameters: num_iter = 10 (didn't affect the performance), reg u = 1, reg i = 2

	RMSE	**MAE**	**NMAE**
Training	0.913	0.722	0.181
Testing	0.962	0.758	0.190

- By comparing this performance with the Global average, we can see that User Item Baseline is giving better performance.
- File '2(a).xlsx' attached with this report contains the performance summary of different parameters for User Item Baseline. Based on the performance summary, we can conclude the following about the effect of parameter changes on the performance.
  (num iter has also been involved below although not available in the file since it was one of the parameter considered)

		Training			Testing		
Parameter	**Change**	**RMSE**	**MAE**	**NMAE**	**RMSE**	**MAE**	**NMAE**
num iter	**Increase**	No Change	No Change	No Change	No Change	No Change	No Change
num iter	**Decrease**	No Change	No Change	No Change	No Change	No Change	No Change
reg u	**Increase**	Increase	Increase	Increase	Increase	Increase	Increase
reg u	**Decrease**	Decrease	Decrease	Decrease	Decrease	Decrease	Decrease
reg i	**Increase**	Increase	Increase	Increase	Increase	Increase	Increase
reg i	**Decrease**	Decrease	Decrease	Decrease	Decrease	Decrease	Decrease

(b)

- Using Optimize Parameter operator, we found out the performance with different parameters like Number of factors, Learning Rate & Regularization.
- File '2-(b).xlsx' attached with this report contains the performance summary of different parameters. By looking at the performance, we can say that
    - As we increase the number of factors, the training error is getting decreased. However, the testing error keeps getting increased. So, we can say that model approaches overfit when we increase the number of factors.
      As we decrease the number of factors, the difference between training & testing error starts getting decreased.
    - Similarly, as we increase the learning rate, the training error keeps getting decreased, which means the model approaches overfit. However, with lower learning rate, we found the training & testing error to be high. So, we can conclude that Learning rate has to be optimal for a better model.
    - Regularization parameter was used to reduce the overfit of the model. If the parameter value is increased, error rate is increasing, but with decrease in parameter value, the model approaches overfit.
- So by comparing all these parameter values & their performance, below is the summary of the model we chose as the best mode.
  Parameters: Num Factors = 3, Learning Rate = 0.015, Regularization = 0.015

	RMSE	**MAE**	**NMAE**
Training	0.845	0.663	0.166
Testing	0.943	0.742	0.186

(c)

- **User k-NN**
  - Using Optimize Parameter operator, we found out the performance of User k-NN with different values of k & Correlation mode as Pearson & Cosine Similarity.
  - File 'User-knn-2(c).xlsx' attached with this report contains the performance summary of different parameters for User k-NN. By looking at the performance we can say that
    - With increasing value of k, the training error keeps on increasing.
    - Error rate is lesser with Pearson measure compared to cosine similarity measure. However, difference between training error & test error is more in Pearson than cosine similarity. Hence, we can say that with Pearson measure, the model approaches overfit.
    - reg_u, reg_i & shrinkage does not have any impact on the performance *(Since, they did not have any impact on the performance, we removed as a parameter from optimized parameter operator to save the computing time)*
  - So by comparing all these parameter values & their performance, below is the summary of the model we chose as the best mode.
    Parameters: k = 50, Correlation mode = cosine, reg_u = 1, reg_i = 1, shrinkage = 5

	RMSE	**MAE**	**NMAE**
Training	0.920	0.720	0.180
Testing	0.957	0.750	0.188

- **Item k-NN**
  - Using Optimize Parameter operator, we found out the performance of Item k-NN with different values of k & Correlation mode as Pearson & Cosine Similarity.
  - File 'Item-knn-2(c).xlsx' attached with this report contains the performance summary of different parameters. By looking at the performance, we found it similar to the User k-NN.
    - With increasing value of k, the training error keeps on increasing.
    - Error rate is lesser with Pearson measure compared to cosine similarity measure. However, difference between training error & test error is more in Pearson than cosine similarity. Hence, pearson measure introduces overfit - we can say.
    - reg_u, reg_i & shrinkage does not have any impact on the performance. *(Since, they did not have any impact on the performance, we removed as a parameter from optimized parameter operator to save the computing time)*
  - So by comparing all these parameter values & their performance, below is the summary of the model we chose as the best mode.
    Parameters: k = 80, Correlation mode = cosine, reg_u = 1, reg_i = 1, shrinkage = 5

	RMSE	**MAE**	**NMAE**
Training	0.9	0.707	0.177
Testing	0.948	0.744	0.186

- **Model Comparison**
  - Below is the summary of all different method performances.

	Training			Testing		
Method	**RMSE**	**MAE**	**NMAE**	**RMSE**	**MAE**	**NMAE**
Global Average	1.126	0.945	0.189	1.122	0.945	0.189
User Item Baseline	0.913	0.722	0.181	0.962	0.758	0.190
Matrix Factorization	0.845	0.663	0.166	0.943	0.742	0.186
User k-NN	0.920	0.720	0.180	0.957	0.750	0.188
Item k-NN	0.9	0.707	0.177	0.948	0.744	0.186

  - By looking at the performance, we can say Matrix Factorization & Item k-NN are best models among all, since the testing error is least in them.
  - But, if we have to choose one, we would go ahead with Item k-NN, since the difference between training & testing error is lesser compared to Matrix Factorization. *(Although, the testing error is lesser in Matrix Factorization, but the difference is only 0.005 which is negligible)*