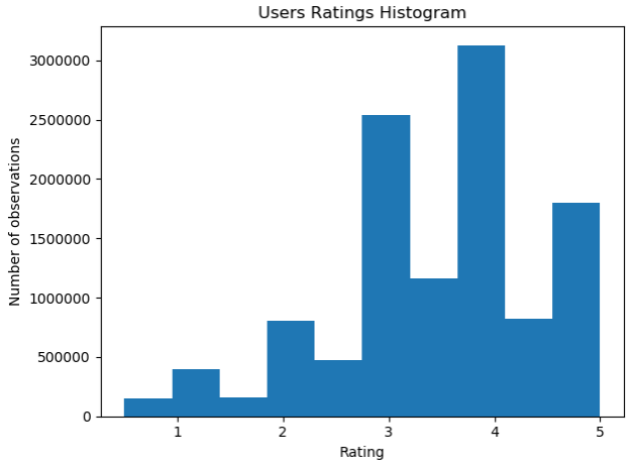
# Moran Sorka

# Data Scientist – Home Test

## Data Exploration:

The dataset contains 7,565 movies rated by 265,917 users.

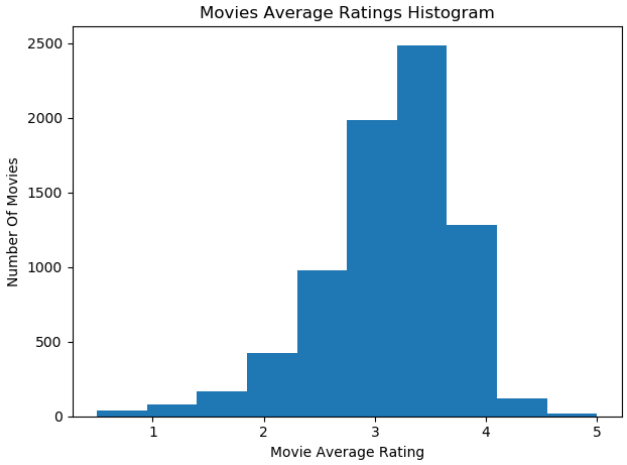
1. Histogram of users ratings:
    - Possible rating values: {0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5}
    - The average rating is 3.53, the mode is 4 (27.4%)
    - 75% of user's ratings it integers (1, 2, 3, 4, 5)
    - 15% of user's ratings is 5




```
Name: rating, dtype: float64
count    1.143764e+07
mean     3.532719e+00
std      1.066919e+00
min      5.000000e-01
25%      3.000000e+00
50%      4.000000e+00
75%      4.000000e+00
max      5.000000e+00
```

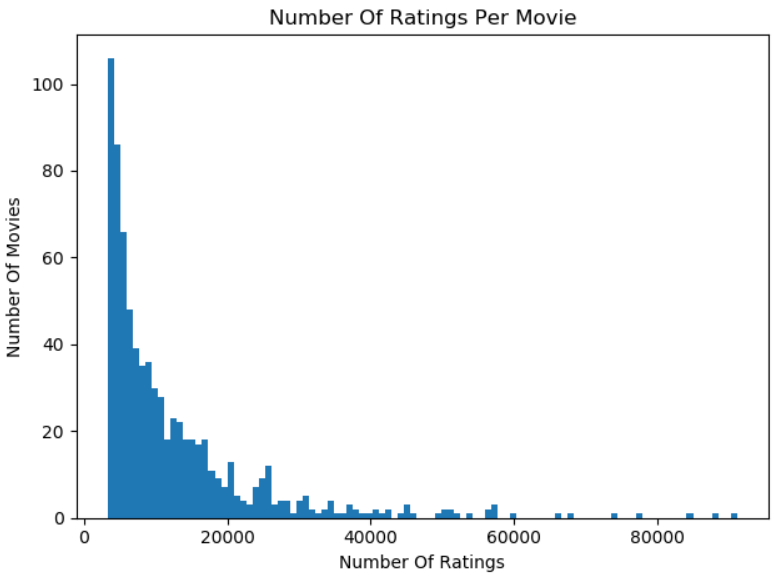
```
Name: rating, dtype: float64
4.0    0.273382
3.0    0.221849
5.0    0.157231
3.5    0.101828
4.5    0.071988
2.0    0.070589
2.5    0.041721
1.0    0.034574
1.5    0.013670
0.5    0.013168
```

2. Histogram of movies average ratings:
    - The average rating per movie is 3.12 with std of 0.62
    - Only 5% of the movies have average rating greater than 4



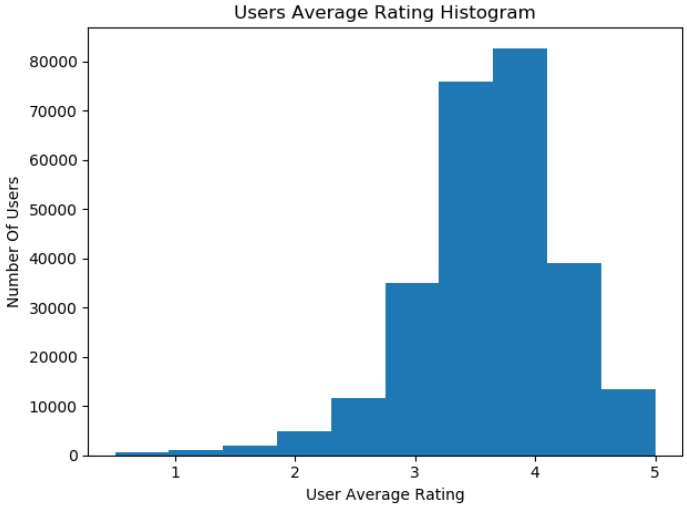

```
count    7565.000000
mean        3.127278
std         0.620652
min         0.500000
25%         2.800000
50%         3.225922
75%         3.552916
max         5.000000
```

3. Histogram of number of ratings per movie:
    - Right skewed distribution
    - The median is 45 rating users per movie
    - 75% of the movies were ranked by less than 469 users



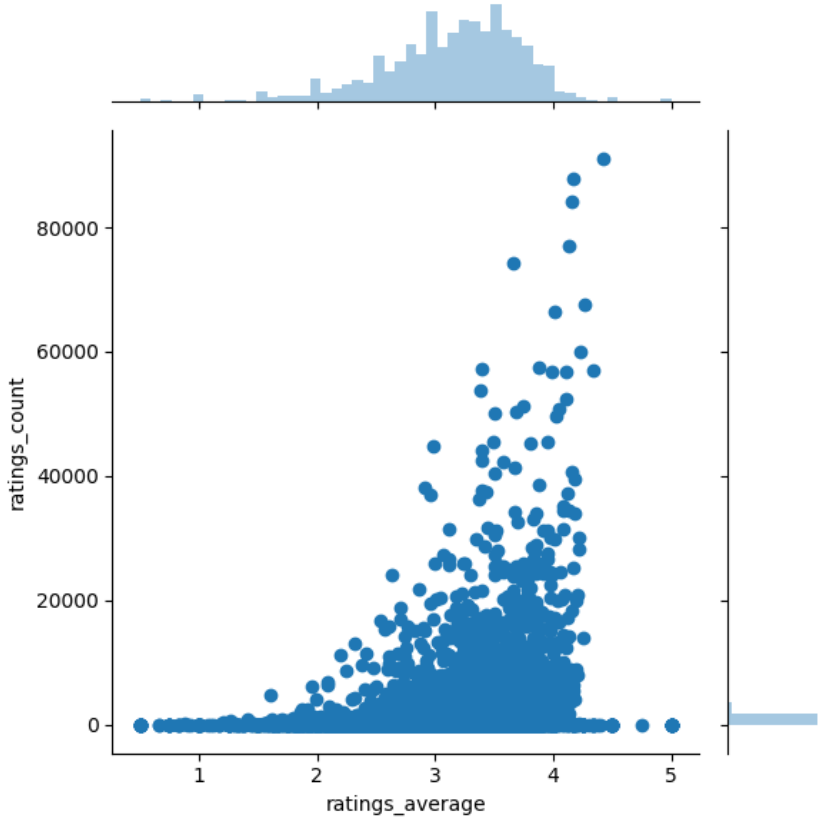

```
Name: ratings_average, dtype: float64
count     7565.000000
mean      1511.773695
std       5305.245799
min          1.000000
25%          7.000000
50%         45.000000
75%        469.000000
max      91082.000000
```

4. Plotting histogram of users average rating:
    - The average rating per user is: 3.6 with std of 0.63
    - Only 5% of the movies have average rating greater than 4




```
Name: rating, dtype: float64
count    265917.000000
mean          3.617960
std           0.628718
min           0.500000
25%           3.285714
50%           3.666667
75%           4.000000
max           5.000000
Name: rating, dtype: float64
```

5. Relationship between ratings average and number of rating users:



## Q.1:

I choose to use IMDB's weighted rating because this formula provides a Bayesian estimation which considers the number of ratings each movie has received, minimum ratings required and the mean ratings for all movies. Thus, this metric can use as an appropriate benchmark for this task.

$$\text{Weighted Rating (WR)} = \left(\frac{v}{v+m}.R\right) + \left(\frac{m}{v+m}.C\right)$$

R = average rating for the movie

v = number of ratings for the movie

m = minimum ratings required to be listed in the Top Rated list

C = the mean ratings across the whole report

Top 10 movies:

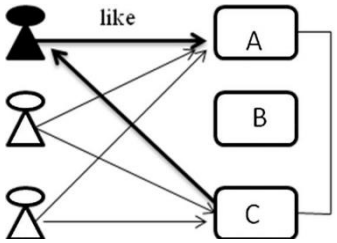
| | title | ratings_count | ratings_average | wr |
|---|---|---|---|---|
| 1560715 | The Million Dollar Hotel | 91082 | 4.42901 | 4.39563 |
| 71479 | Sleepless in Seattle | 57070 | 4.33981 | 4.29256 |
| 921500 | Once Were Warriors | 67662 | 4.26653 | 4.22978 |
| 230555 | License to Wed | 60024 | 4.23072 | 4.19142 |
| 1333567 | Terminator 3: Rise of the Machines | 87901 | 4.16998 | 4.14491 |
| 2011099 | The Thomas Crown Affair | 30043 | 4.21439 | 4.14144 |
| 1458620 | Murder She Said | 28280 | 4.21303 | 4.13615 |
| 342641 | Confession of a Child of the Century | 39600 | 4.18207 | 4.12787 |
| 3128961 | Solaris | 84078 | 4.15225 | 4.12676 |
| 4177393 | The Talented Mr. Ripley | 33987 | 4.17829 | 4.11627 |

## Q.2:

Splitting data into test and train sets

## Q.3:

I choose to implement Item-based collaborative filtering models (Model 1). Those models identify similar movies based on users' previous ratings and predict the user's rating for unrated movies.




I used grid search for tuning the models parameters.

- KNN-With-Means – clustering based algorithm
- SVD - Matrix Factorization based algorithm

.

- I compared the models using RMSE
- KNN is unsupervised learning model (pre-computed), we can limit the number of neighbors and make it more scalable. Many parameters to tune, sensitivity to noise.
- SVD is better dealing with scalability and sparsity data.
- In real life problems the recommendations must be "up to date" and response time must be very fast (trade-off). Evaluating the recommendation model prior to launch can be out-of-date.
- CTR, clicks for BPR, retention rate

| | userId | movieId | title | rating | pred_rating |
|---|---|---|---|---|---|
| 128548 | 270896 | 858 | Sleepless in Seattle | 5 | 4.88273 |
| 2584884 | 270896 | 2324 | Local Color | 4 | 4.68356 |
| 1486899 | 270896 | 750 | Murder She Said | 5 | 4.67674 |
| 1421467 | 270896 | 296 | Terminator 3: Rise of the Machines | 5 | 4.66335 |
| 382240 | 270896 | 58559 | Confession of a Child of the Century | 5 | 4.65642 |
| 1708623 | 270896 | 4993 | 5 Card Stud | 5 | 4.61439 |
| 3404490 | 270896 | 1089 | Point Break | 5 | 4.59511 |
| 4090172 | 270896 | 778 | Monsieur  Hulot's Holiday | 5 | 4.5277 |
| 1194042 | 270896 | 2692 | The Red Elvis | 4.5 | 4.51276 |
| 5286536 | 270896 | 924 | Dawn of the Dead | 4.5 | 4.49816 |
| 1651795 | 270895 | 318 | The Million Dollar Hotel | 5 | 5 |
| 989161 | 270895 | 527 | Once Were Warriors | 5 | 5 |
| 66510 | 270895 | 110 | Three Colors: Red | 4 | 4.81531 |
| 2973529 | 270895 | 457 | Sissi | 4 | 4.77675 |
| 5579123 | 270895 | 161 | Ocean's Eleven | 4 | 4.54363 |
| 7390992 | 270895 | 539 | Psycho | 4 | 4.30331 |
| 577416 | 270895 | 339 | Night on Earth | 3 | 4.29808 |
| 6371711 | 270895 | 454 | Romeo + Juliet | 5 | 4.29355 |
| 5486128 | 270895 | 597 | Titanic | 5 | 4.22635 |
| 921498 | 270895 | 500 | Reservoir Dogs | 4 | 4.18779 |
| 989160 | 270894 | 527 | Once Were Warriors | 2 | 3.01235 |
| 382239 | 270894 | 58559 | Confession of a Child of the Century | 3 | 2.92704 |
| 4211378 | 270894 | 1213 | The Talented Mr. Ripley | 3 | 2.92062 |
| 331283 | 270894 | 4226 | Shriek If You Know What I Did Last Friday the Thirteenth | 3 | 2.89734 |
| 3213037 | 270894 | 593 | Solaris | 5 | 2.89693 |
| 1458619 | 270894 | 541 | The Man with the Golden Arm | 4 | 2.86831 |
| 4774038 | 270894 | 745 | The Sixth Sense | 0.5 | 2.85499 |
| 1560713 | 270894 | 293 | A River Runs Through It | 2 | 2.82644 |
| 7925887 | 270894 | 5995 | Miffo | 1.5 | 2.78824 |
| 230553 | 270894 | 2762 | Young and Innocent | 1.5 | 2.76378 |