

# Online Dating

The Use of the Internet to Meet Partners



**Alekya Kumar**  
**Trupti Jadhav**

# Agenda

- **01 Introduction**
- **02 Information of Dataset**
- **03 Methodologies/Analysis**
- **04 Findings**
- **05 Conclusion**

# Dataset Overview

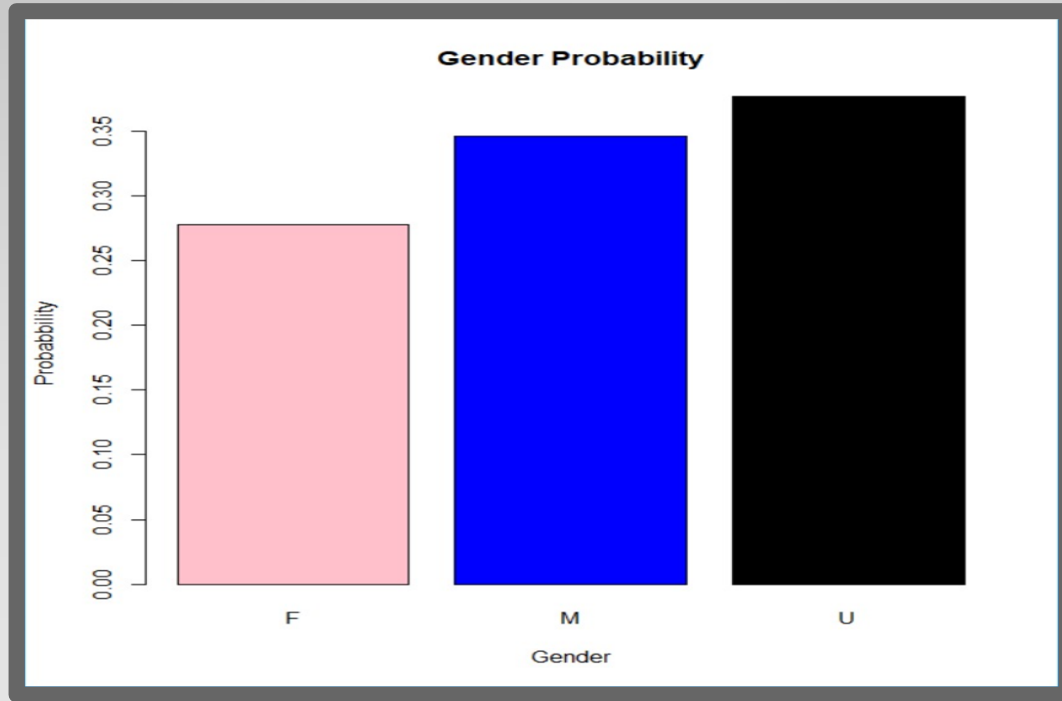
- ★ Data contains 17,359,346 anonymous ratings of 168,791 profiles made by 135,359 users.
- ★ UserID - user who provided rating.
- ★ ProfileID - user who has been rated.
- ★ UserIDs range between 1 and 135,359.
- ★ ProfileIDs range between 1 and 220,970 (not every profile has been rated).
- ★ Ratings are on a 1-10 scale where 10 is the best (integer ratings only).
- ★ Only users who provided at least 20 ratings were included.
- ★ Users who provided constant ratings were excluded.



# Dataset

	UserID	Profile	Gender_user	Gender_profile	Ratings
1	1	3346	F	M	9
2	1	41755	F	M	2
3	1	128	F	M	1
4	1	22552	F	F	10
5	1	12124	F	U	10
6	1	127227	F	F	7
7	1	34396	F	M	2
8	1	133879	F	M	2
9	1	74829	F	M	2
10	1	115087	F	M	2
11	1	127060	F	M	1
12	1	91997	F	M	9
13	1	90280	F	F	4
14	1	109792	F	M	4
15	1	120599	F	M	9
16	1	108446	F	M	6
17	1	110866	F	M	8
18	1	41869	F	M	1
19	1	130355	F	M	9
20	1	50790	F	M	6

# Data Analysis



Gender	freq
F	61365
M	76441
U	83164

# Data Analysis

## mean\_ratings\_user

full\$Gender_user	full\$Ratings
F	5.356985
M	6.706718
U	4.928286

## mean\_ratings\_profile

full\$Gender_profile	full\$Ratings
F	6.358190
M	5.031327
U	5.860144

## male\_male\_mean

male_male\$Gender_profile	male_male\$Ratings
M	4.461443

## female\_female\_mean

female_female\$Gender_profile	female_female\$Ratings
F	5.13514

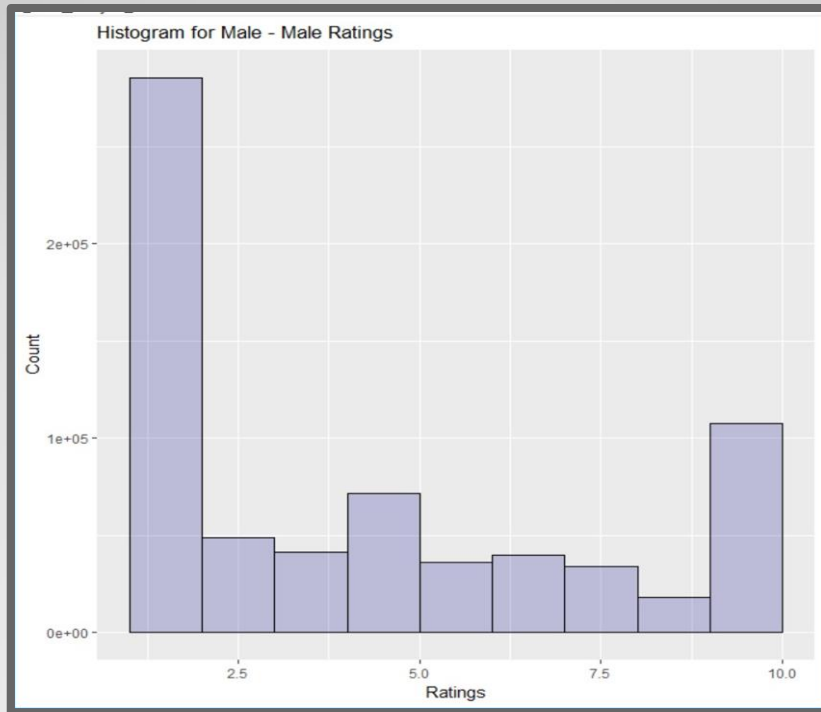
## male\_female\_mean

male_female\$Gender_profile	male_female\$Ratings
F	6.922843

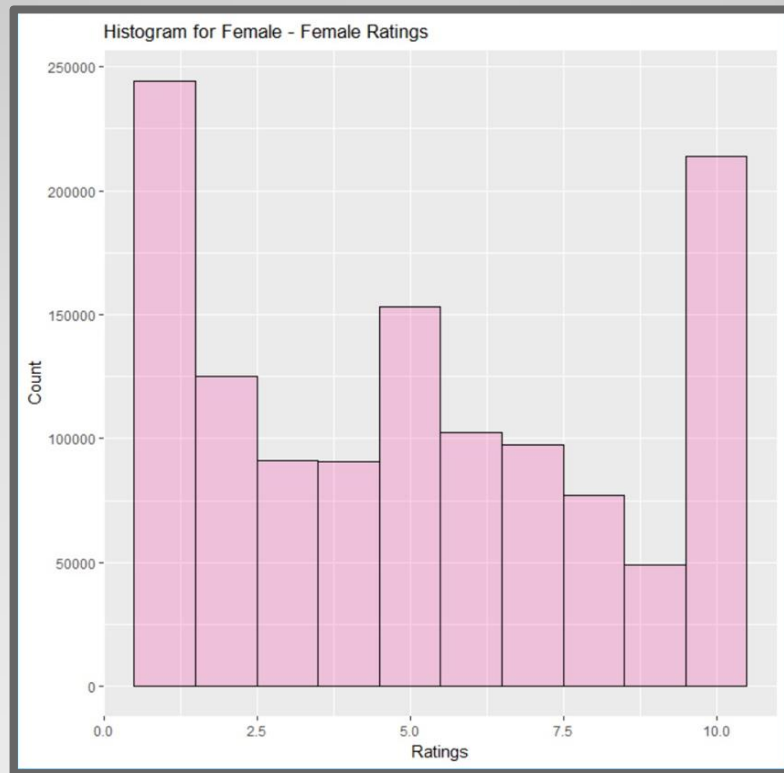
## female\_male\_mean

female_male\$Gender_profile	female_male\$Ratings
M	5.483944

# Male-Male Ratings

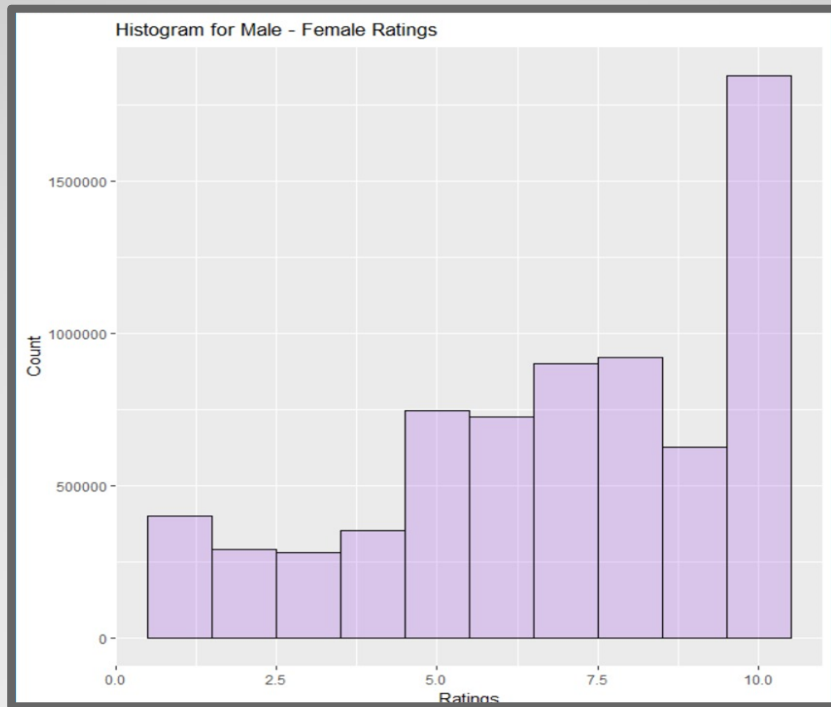


# Female-Female Ratings

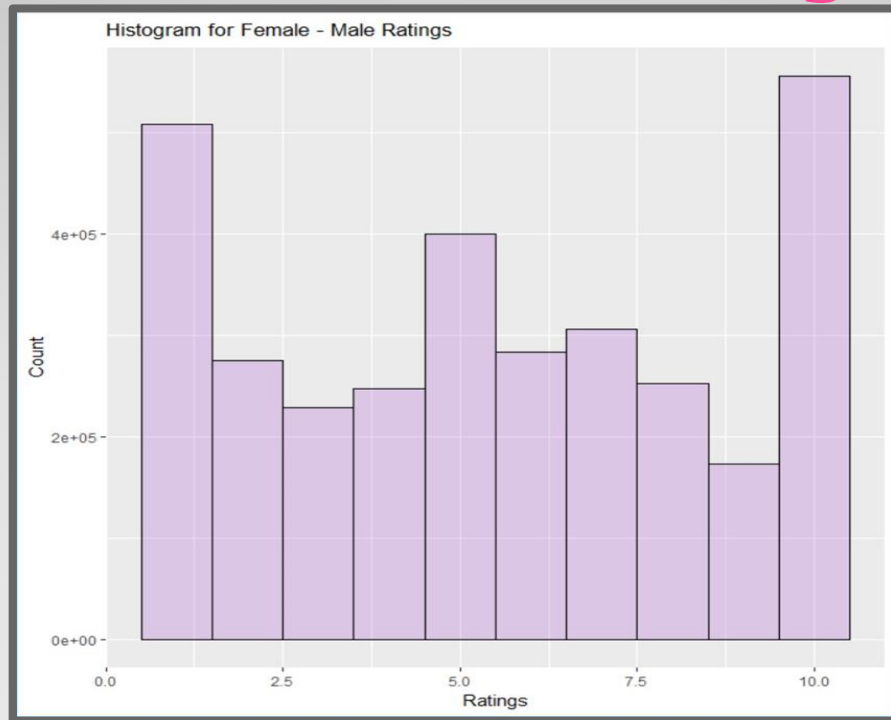




# Male-Female Ratings



# Female-Male Ratings



# Why Collaborative Filtering ?

- ❑ **We don't have the set of features about a profile to come up with recommendations, instead we have ratings. So Content based recommender system was ruled out.**
- ❑ **Based on similarities/likings of the users and taking the likings which are repeated the most, we have come up with certain recommendations.**

# Building Recommender System


**A total of  $3 \times 3 = 9$  models were constructed and evaluated.**

**Similarity(Cosine similarity)**

**→ Normalization**

- ☐ **Non-normalized**
- ☐ **centered**
- ☐ **z-score**

# Cosine Similarity Errors



	RMSE	MSE	MAE
UBCF_c_non_normalized	6.062238	36.75073	5.239639
UBCF_c_center_normalized	2.910479	8.47089	2.416934
UBCF_c_zscore_normalized	2.913338	8.48754	2.419143

Centering based Normalization outperformed Z-score normalization and both of those normalization outperformed the model constructed using non-normalized data.

# Building Recommender System

**Similarity(Pearson Correlation as similarity)**

→ **Normalization**

- ☐ **Non-normalized**
- ☐ **Centered**
- ☐ **Z-score**

# Pearson Correlation Errors

	RMSE	MSE	MAE
UBCF_p_non_normalized	6.346035	40.27216	5.595615
UBCF_p_center_normalized	2.620766	6.868416	2.139462
UBCF_p_zscore_normalized	2.812717	7.911377	2.322087



Centering based Normalization outperformed Z-score normalization and both of those normalization outperformed the model constructed using non-normalized data.

# Building Recommender System


**Similarity(Euclidean Distance as similarity)**

→ **Normalization**

- ☐ **Non-normalized**
- ☐ **centered**
- ☐ **z-score**



# Euclidean Distance Errors



	RMSE	MSE	MAE
UBCF_e_non_normalized	6.05472	36.65964	5.235555
UBCF_e_center_normalized	2.909454	8.46492	2.41519
UBCF_e_zscore_normalized	2.91632	8.50494	2.4208

Centering based Normalization outperformed Z-score normalization and both of those normalization outperformed the model constructed using non-normalized data.

# Next Part...

- ❑ **After building and evaluating various models, the best one was chosen, which is Pearson correlation with centered normalization model and predictions were also made for the same**
- ❑ **Build a model based on Popularity metric.**
- ❑ **Compared the results for both the recommender lists using the same set of users.**

## Top 10 recommendation for 5 users(Pearson-centered norm.)

	1	2	3	4	5	6	7	8	9	10
User 10	113157	71570	14349	97992	22319	30166	104785	77036	74257	125439
User 11	54929	156148	10148	65602	143706	22319	14258	26084	63063	117775
User 12	74829	100855	139552	42490	36793	205930	194912	65218	93891	21497
User 13	71570	33216	81470	14349	58395	7242	74257	86018	77036	61278
User 14	71636	22319	30400	81470	33216	52391	93681	133232	71570	81498

# Top 10 recommendation for 5 users(Popularity)

	1	2	3	4	5	6	7	8	9	10
User 10	156148	22319	71636	121859	33216	71570	117981	113157	65602	10148
User 11	156148	22319	71636	31116	121859	33216	71570	117981	113157	65602
User 12	22319	71636	31116	33216	71570	117981	113157	65602	10148	98678
User 13	156148	71636	121859	33216	71570	117981	65602	10148	98678	14258
User 14	156148	22319	71636	31116	121859	33216	71570	113157	65602	10148

# Conclusion

- ❑ As the tables show, the 3 UBCF model employing raw, non-normalized data were the poorest performing UBCF models assessed herein.
- ❑ Thus normalization of the data appears to improve the UBCF model regardless of which similarity metrics is employed.
- ❑ The errors for Pearson Correlation approach is lesser than the other two approaches.
- ❑ As such, the Pearson Correlation should be preferred over Cosine similarity and Euclidean Distance metrics when developing a user based collaborative filtering recommender for our dataset.
- ❑ From the popularity method, we can see that few sets of profiles are constant being picked many times. This is because those set of profiles have higher popularity within the users.



Questions/Comments?

**THANKYOU!!!**