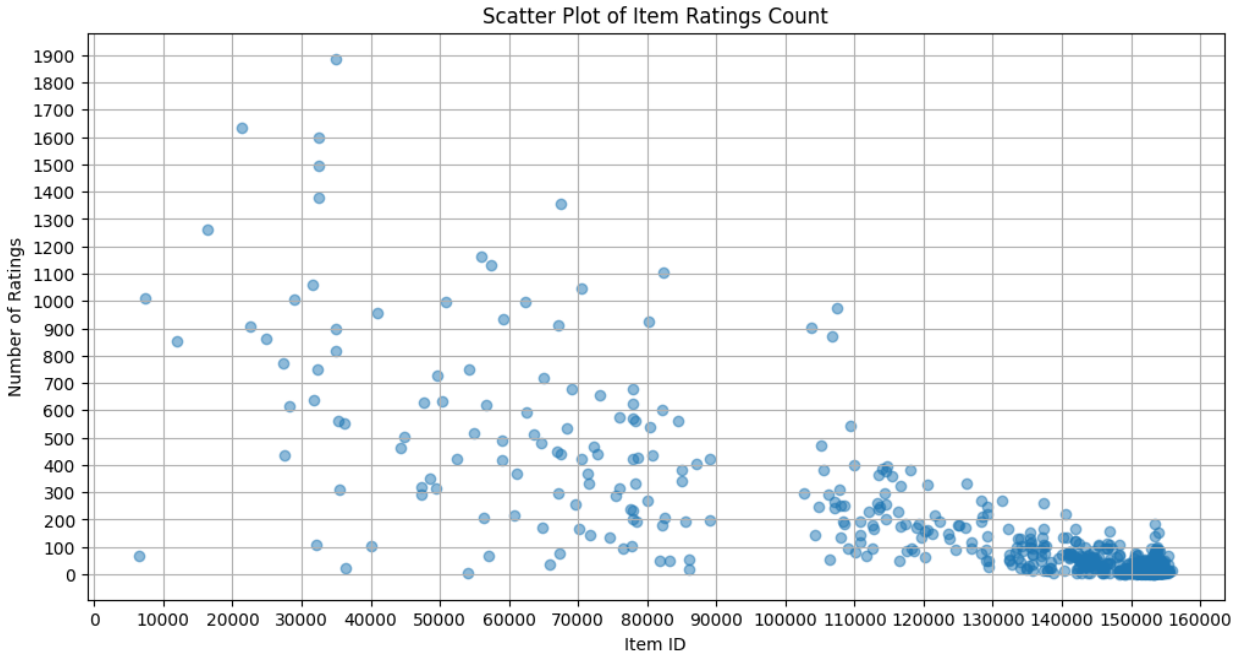
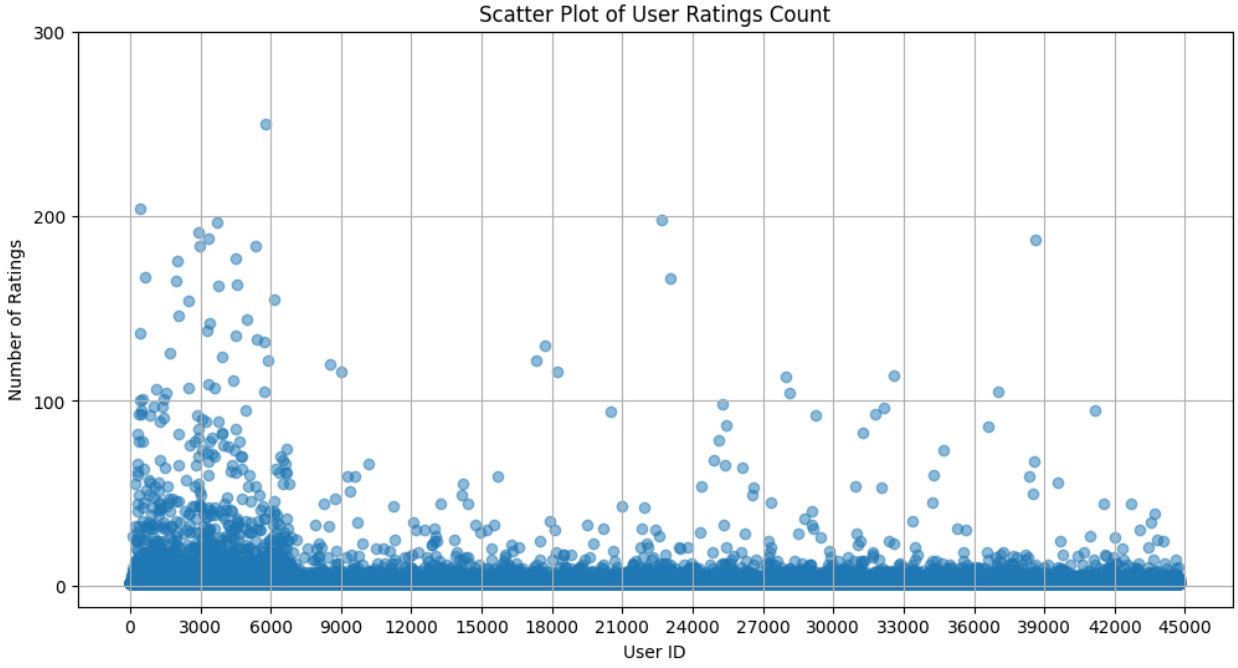
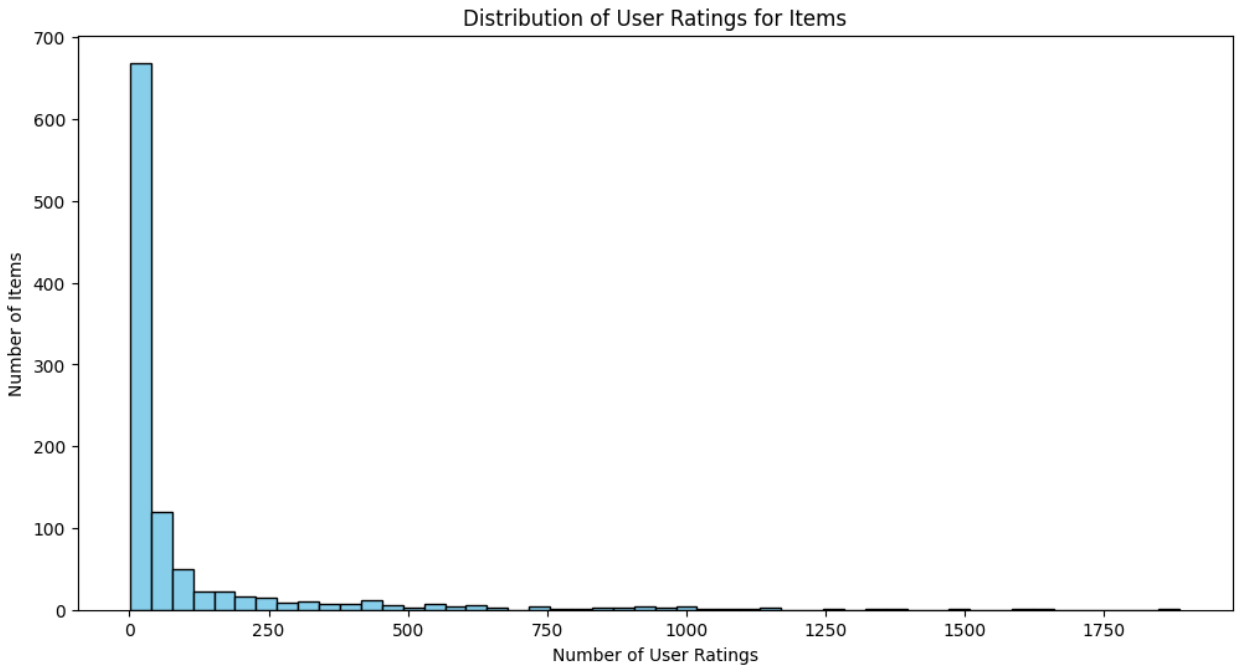
DMG Assignment 2
Tarushi Gandhi 2020579
Shreya Bhatia 2020542







Number of user ratings for individual items is very skewed, data is sparse, item-item similarity will work better. This can also be observed through the above scatter plots.
We used pearson correlation as our similarity measure
We first calculate k nearest neighbors to the given item and then we compute the rating using the formula

$$r_{xi} = b_{xi} + \frac{\sum_{j \in N(i;x)} s_{ij} \cdot (r_{xj} - b_{xj})}{\sum_{j \in N(i;x)} s_{ij}}$$

baseline estimate for r_{xi}

$$b_{xi} = \mu + b_x + b_i$$

- μ = overall mean movie rating
- b_x = rating deviation of user x
  = (*avg. rating of user* x) $- \mu$
- b_i = rating deviation of movie i

Results
RMSE for validation sets
Fold 0:
RMSE 1.2268427820716816
Fold 1:
RMSE 1.2119851907275256
Fold 2:
RMSE 1.210863465852417

Fold 3:
RMSE 1.2062653053982821
Fold 4:
RMSE 1.2117064604506345
Mean RMSE 1.2135326409001082

RMSE for test set
RMSE = 1.2017464266278848