Sistemas de Recomendacion

RIWS 2021/2022

Ricardo Sánchez Arias - ricardo.sanchez1@udc.es

1. Realizar una exploración inicial indicando el número de valor vacíos (NA), el número de muestras duplicadas, y el número de usuarios, productos y puntuaciones que hay en el DataFrame creado

Resultados despues de exploración

```
Total number elements: 403344
Number of empty values (NA): 0
Number of duplicated samples: 0
Number of users: 610
Number of movies: 9724
Number of ratings: 10
Number of timestamps: 85043
```

Tras la exploración de los valores del archivo facilitado, no se encontraron valores vacíos (NA), ni valores duplicados.

Existen un total de 610 usuarios.

Existen un total de 9724 productos.

Existen 10 ratings. Esto es de 0 a 5 en incrementos de 0.5.

Existen un total de 85043 timestamps.

2. Si eliminamos del dataset los productos con menos de 5 puntuaciones, ¿cuántos productos quedan? A continuación, sobre el dataset obtenido, si eliminamos los usuarios con menos de 10 puntuaciones, ¿cuántos usuarios quedan? ¿y cuántos productos? ¿de qué tamaño es ahora la matriz?

```
Removing movies with less than 5 ratings...
Number of movies: 3650

Removing users with less than 10 ratings...
Number of users: 610

Size of matrix: (90274, 4)
```

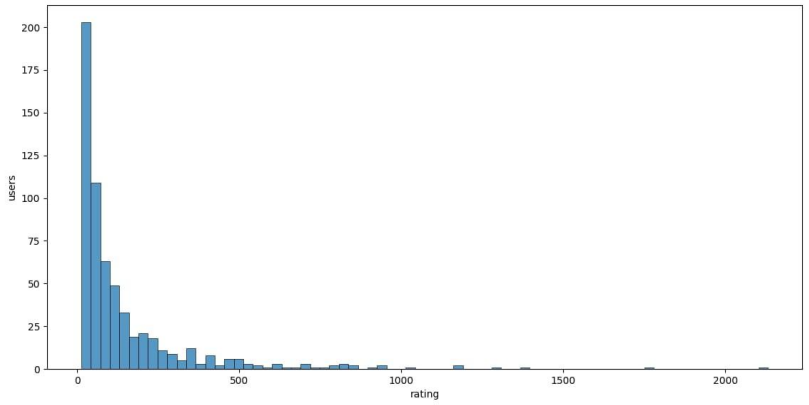
Después de eliminar los productos con menos de 5 ratings, quedan 3650 productos.

A este resultado le eliminamos los usuarios con menos de 10 ratings, quedando los mismos usuarios, 610.

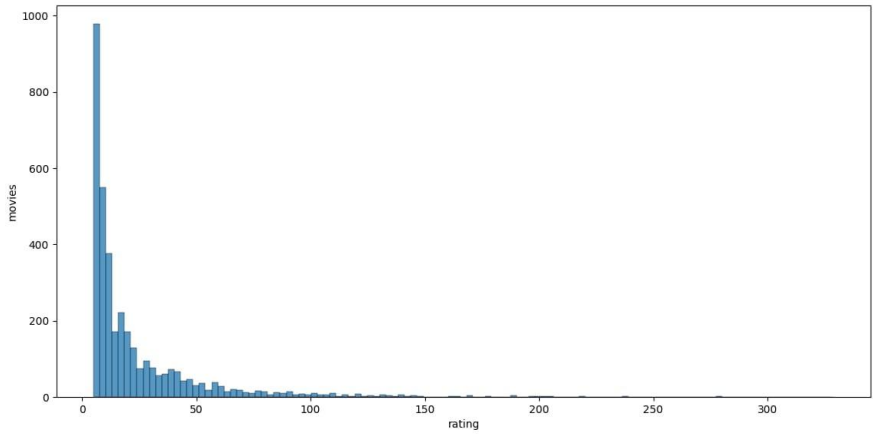
En total, después de este filtrado, la matrix tiene un tamaño de 90274 x 4

3. Representar en un histograma el número de puntuaciones por usuario. Hacer lo mismo con el número de puntuaciones por producto

Número de puntuaciones por usuario



Número de puntuaciones por producto

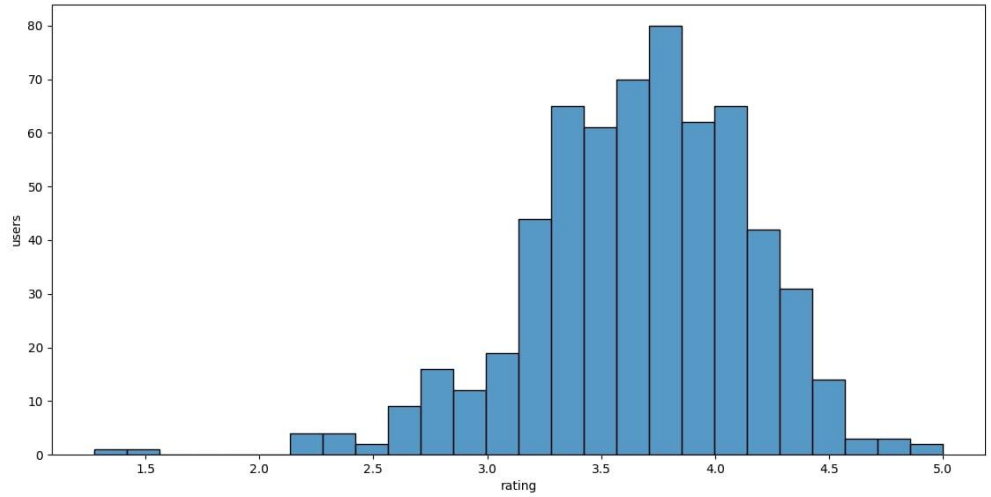


En estas graficas se puede observar cuantos usuarios/productos tienen un numero x de valoraciones, e.g., aproximadamente 200 usuarios tienen sobre 50 valoraciones.

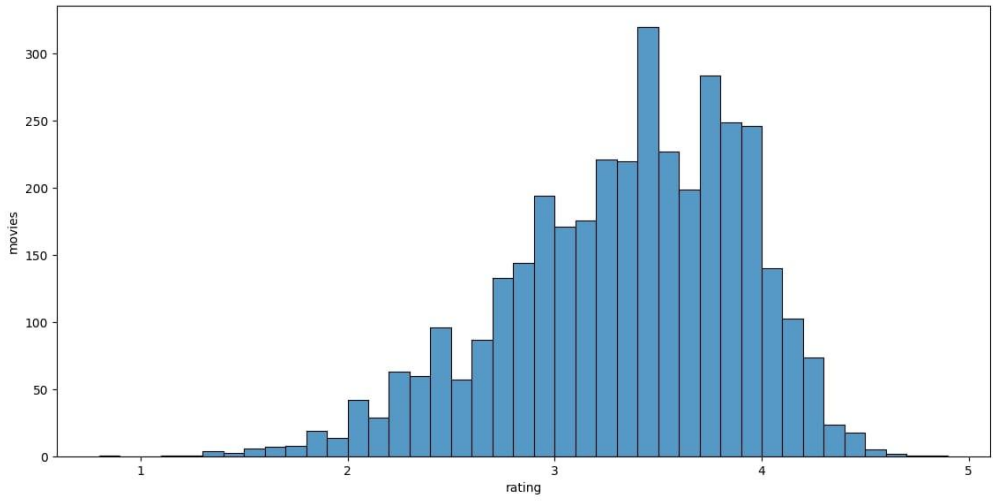
Se aplicaria lo mismo para la gráfica de productos.

4. Representar en un histograma la media de puntuaciones por usuario. Hacer lo mismo con la media de puntuaciones por producto.

Media de puntuaciones por usuario



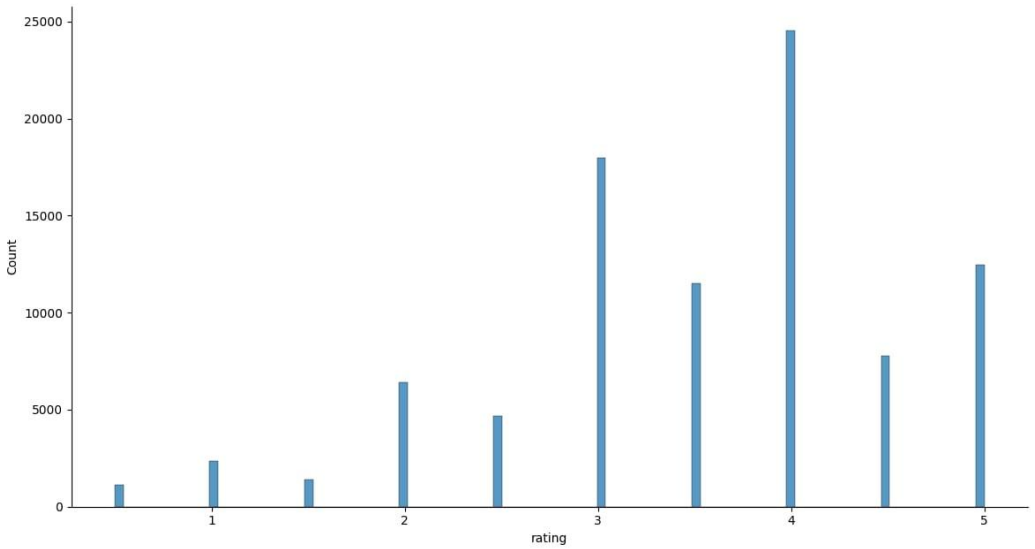
Media de puntuaciones por producto



En estas graficas se reflejan de los distintas medias depuntuaciones, cuantos usuarios/productos als tienen.

Con esto podemos saber que notas son las mas usadas para valorar productos.

5. Representar en un diagrama de barras la distribución de las puntuaciones



En esta gráfica se puede observar que la mayoría de las puntuaciones están en torno al 3.5 y 4.

6. Crear un objeto de tipo surprise.Dataset a partir del DataFrame empleado previamente. Considerando la tarea de predicción y las métricas RMSE y MAE, aplicar cross_validate (con k = 5). ¿Cuál es el algoritmo que mejor se comporta y por qué? Comparar el funcionamiento de los algoritmos. Justificar la respuesta

SVD Algorithm

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8591  0.8598  0.8652  0.8506  0.8535  0.8576  0.0051
MAE (testset)     0.6581  0.6568  0.6616  0.6544  0.6580  0.6578  0.0023
Fit time          1.05    1.04    1.03    0.96    1.00    1.01    0.03
Test time         0.22    0.20    0.19    0.19    0.18    0.20    0.01
```

NormalPredictor Algorithm

```
* NormalPredictor Algorithm
```

```
Evaluating RMSE, MAE of algorithm NormalPredictor on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.4037  1.3986  1.3967  1.4048  1.4027  1.4013  0.0031
MAE (testset)     1.1222  1.1166  1.1178  1.1191  1.1211  1.1194  0.0020
Fit time          0.09    0.11    0.11    0.13    0.12    0.11    0.01
Test time         0.17    0.09    0.17    0.10    0.18    0.14    0.04
```

BaselineOnly Algorithm

```
* BaselineOnly Algorithm

Evaluating RMSE, MAE of algorithm BaselineOnly on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8594  0.8646  0.8514  0.8614  0.8559  0.8585  0.0045
MAE (testset)     0.6641  0.6619  0.6571  0.6623  0.6582  0.6607  0.0027
Fit time          0.25    0.25    0.24    0.24    0.24    0.24    0.01
Test time         0.10    0.09    0.08    0.14    0.11    0.10    0.02
```

KNNWithZScore Algorithm msd

```
* KNNWithZScore Algorithm msd

Evaluating RMSE, MAE of algorithm KNNWithZScore on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8649  0.8586  0.8637  0.8663  0.8557  0.8618  0.0040
MAE (testset)     0.6588  0.6515  0.6551  0.6610  0.6499  0.6552  0.0042
Fit time          0.34    0.36    0.36    0.37    0.46    0.38    0.04
Test time         1.70    1.65    1.70    1.69    1.90    1.73    0.09
```

KNNWithZScore Algorithm cosine

```
* KNNWithZScore Algorithm cosine

Evaluating RMSE, MAE of algorithm KNNWithZScore on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8680  0.8547  0.8735  0.8687  0.8634  0.8657  0.0064
```

```
MAE (testset)     0.6594  0.6525  0.6635  0.6620  0.6567  0.6588  0.0039
Fit time          0.68    0.68    0.79    0.66    0.66    0.69    0.05
Test time         1.76    1.76    1.91    1.74    1.63    1.76    0.09
```

KNNWithZScore Algorithm pearson

```
* KNNWithZScore Algorithm pearson

Evaluating RMSE, MAE of algorithm KNNWithZScore on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8597  0.8581  0.8501  0.8721  0.8676  0.8615  0.0077
MAE (testset)     0.6495  0.6503  0.6483  0.6560  0.6559  0.6520  0.0033
Fit time          0.71    0.74    0.74    0.90    0.78    0.77    0.07
Test time         1.62    1.59    1.68    2.01    1.69    1.72    0.15
```

El algoritmo que mejor se comporta es el KNNWithZScore Algorithm pearson ya que cuenta con el menor valor encuanto a la RMSE y MAE.

La RMSE baja indica que el error cuadrático medio entre las predicciones del modelo y los valores reales es la mas baja entre los modelos.

La MAE indica que el error absoluto medio entre las predicciones del modelo y los valores reales son las mas bajas.