Familiarize Yourself with the Dataset

In the lab sessions, we will work with the "All Beauty" category of the Amazon Review Data, and we will use the 5-core subset. You can download the dataset and find information about it here: https://nijianmo.github.io/amazon/index.html

Exercise 1

Download and import the 5-core dataset.

Exercise 2

Clean the dataset from missing ratings and duplicates (cases where the same user has rated the same item multiple times) if any. How many observations does the cleaned dataset have?

```
Obervations in the cleaned dataset: 4092
```

Exercise 3

Create a test set by extracting the latest (in time) positively rated item (rating ≥ 4) by each user. Remove users that do not appear in the training set. How many observations does the training and test set have?

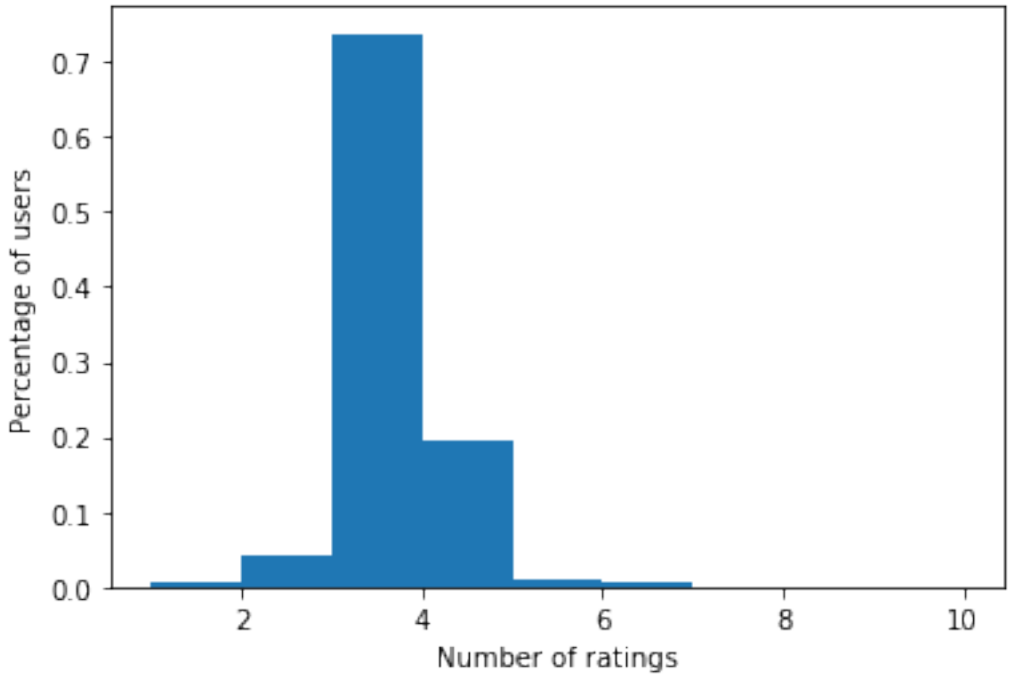
```
Observations in training set: 3133
Observations in test set: 949
```

Exercise 4

4.1

Compute the number of ratings per user in the training set. What is the summary statistics of the number of ratings, and how does a histogram look like?
Reflect on how a collaborative filtering and a content-based recommender system, respectively, will perform for users with few ratings.

```
Summary statistics:
 count  981.000000
  mean    3.193680
   std    0.610454
   min    1.000000
   25%    3.000000
   50%    3.000000
   75%    3.000000
   max    9.000000
Histogram:
```

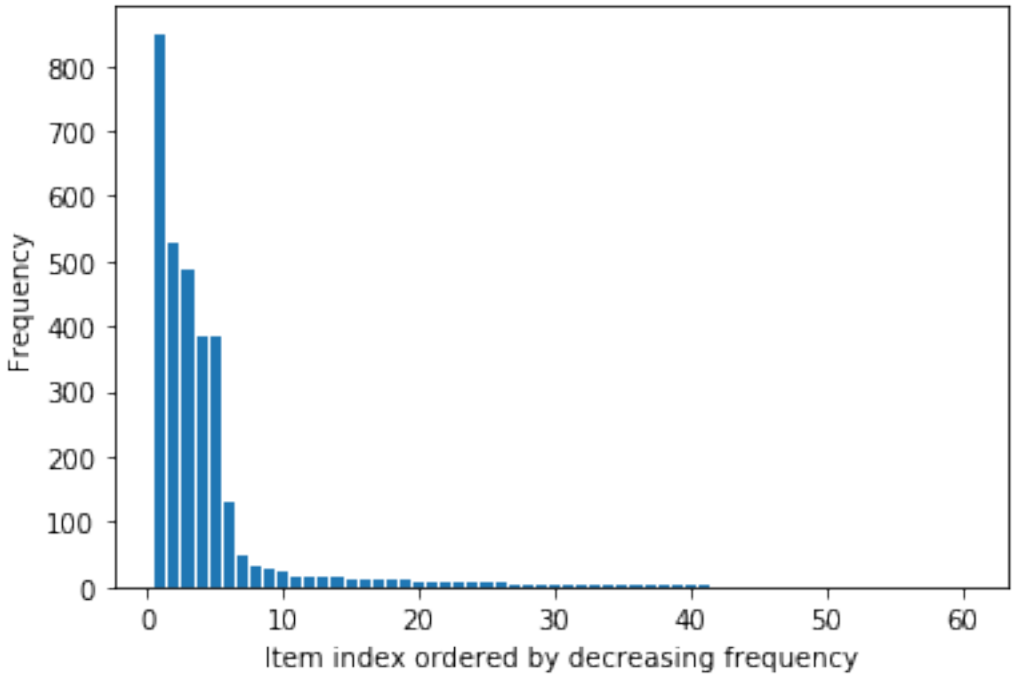


4.2

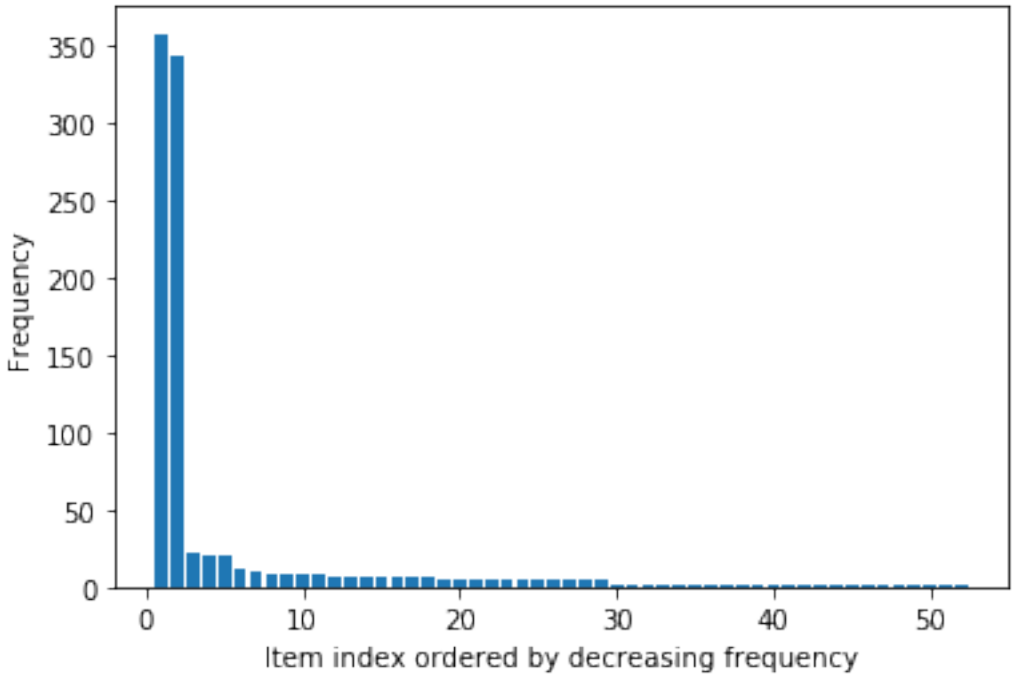
Compute the number of ratings per item in the training set. How does a barplot of the number of ratings ordered by decreasing frequency look like?
Reflect on how it will affect the prediction process of a recommender system if only a small fraction of the items are rated frequently.

Repeat this exercise on the test set and reflect on how the evaluation of a recommender system can be affected by popular items.

`Training set:`


Frequency
800
700
600
500
400
300
200
100
0
0
10
20
30
40
50
60
Item index ordered by decreasing frequency

Test set:

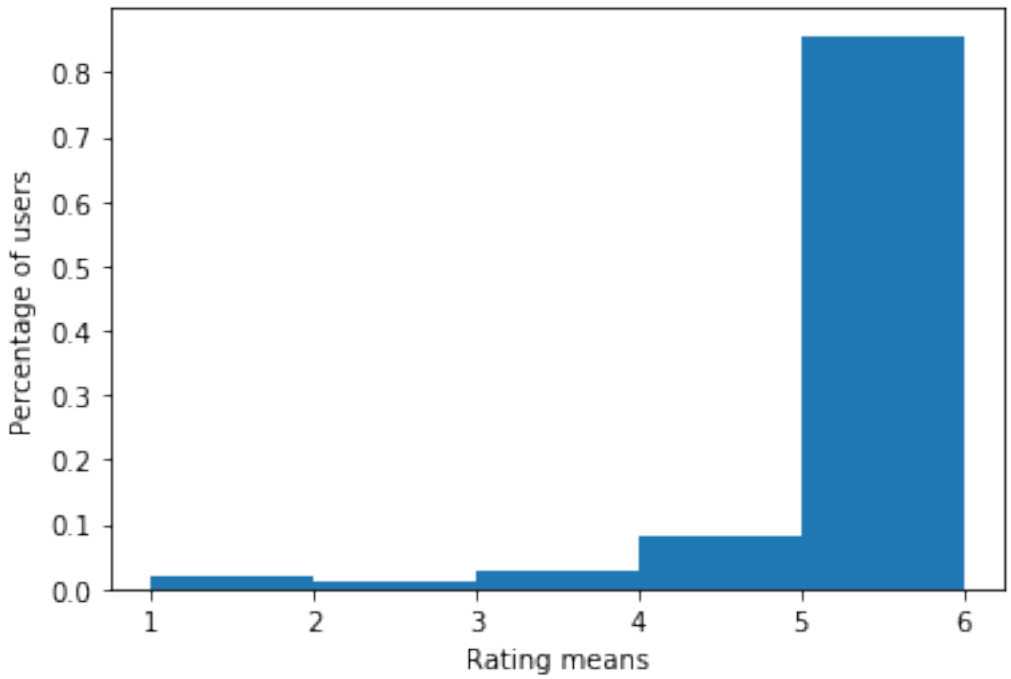

Frequency
350
300
250
200
150
100
50
0
0
10
20
30
40
50
Item index ordered by decreasing frequency

4.3

Compute the mean rating per user in the training set. What is the summary statistics of the rating means, and how does a histogram look like?
Reflect on how a recommender system can take into account if different users rate on different "scales" (e.i. a rating of 3 may be high for one user while low for another).

Repeat this exercise with mean rating per item.

```
Mean Rating per User
Summary statistics:
 count  981.000000
  mean    4.770014
   std    0.718303
   min    1.000000
   25%    5.000000
   50%    5.000000
   75%    5.000000
   max    5.000000
Histogram:
```




```
Mean Rating per Item
Summary statistics:
 count  60.000000
  mean   4.022345
   std   0.920234
```

```
   min    1.000000
   25%    3.464286
   50%    4.154762
   75%    4.747879
   max    5.000000
Histogram:
```