Q1. Which labels/features/records can be used to represent user preference (i.e., how much a user likes the item)?
A: Rating records. The records include Book-Rating score from users.

Also, what is the total range of the labels/ features/records?
A: After refering Book-Crossing Dataset-Format, it metioned that "Book-Rating" expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.
So we could find out the total range is 0-10.

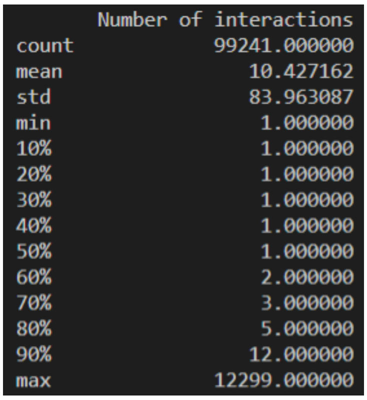
Does the higher the value is, the better a user prefers an item, or conversely the lower the better?
A: Yes, but only correct to 1-10 except 0. As metioned like above, 0 expressed implicit meaning, so we couldn't figure out 0 is higher better or lower better.

Q2. What did you do to clean the data? Have you done any transformation, integration or deletion? Briefly explain why you did all these cleaning actions.
A:

1. **I used "ISO-8859-1" encoding format to open csv file (use chardet to detect the encoding format), because it seemed like error occured when I used "utf-8" encoding format.**
2. **I tried to use a separator with regex format to find the semicolon which not in column value but between each column.**
3. **I found out some columns have double quotes and tab symbols that not required. So I replaced and removed them to let the data look more clearly.**
4. **I deleted the rating records whose ISBN format is not correct and not in books records.**
5. **I replaced some decode char like '&amp;', enter line symbol and double quotes to display data correctly.**

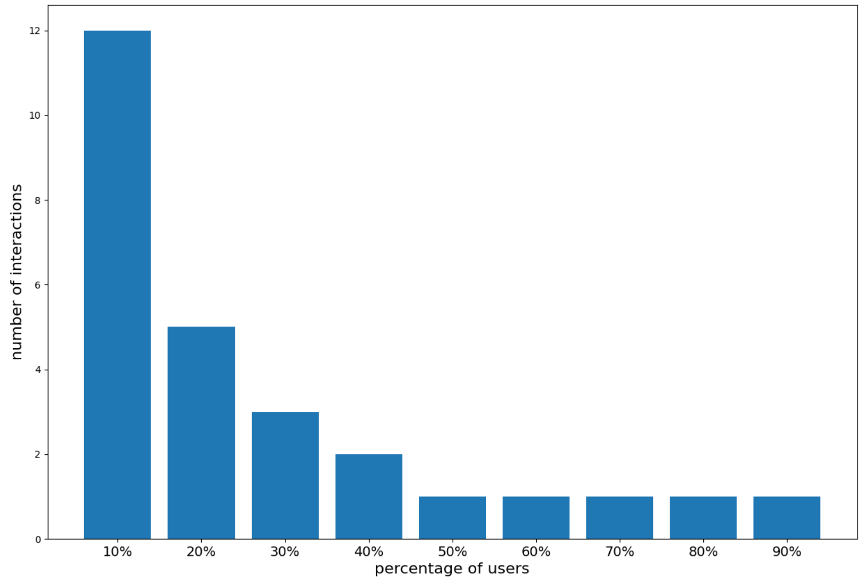
Q3. Sort the users by the number of interactions, and observe what is the minimum number of interactions (Y) generated by the top-X% users? Draw some appropriate figures and briefly explain what your insight is.

A: We can get the percentiles information as below.



```
       Number of interactions
count            99241.000000
mean                10.427162
std                 83.963087
min                  1.000000
10%                  1.000000
20%                  1.000000
30%                  1.000000
40%                  1.000000
50%                  1.000000
60%                  2.000000
70%                  3.000000
80%                  5.000000
90%                 12.000000
max              12299.000000
```

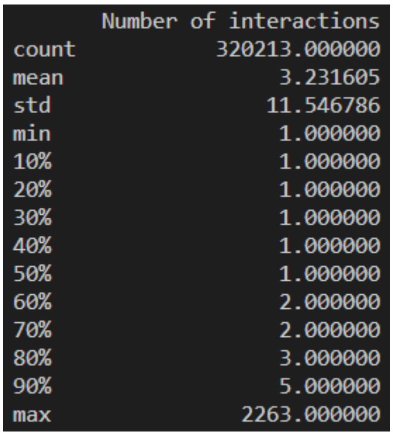
And according to the wiki, the explaination of percentiles said: Pk表示至少有k%的資料小於或等於這個數，而同時也有(100-k)%的資料大於或等於這個數。
Then, we can get the reverse result from above, and get below figure.



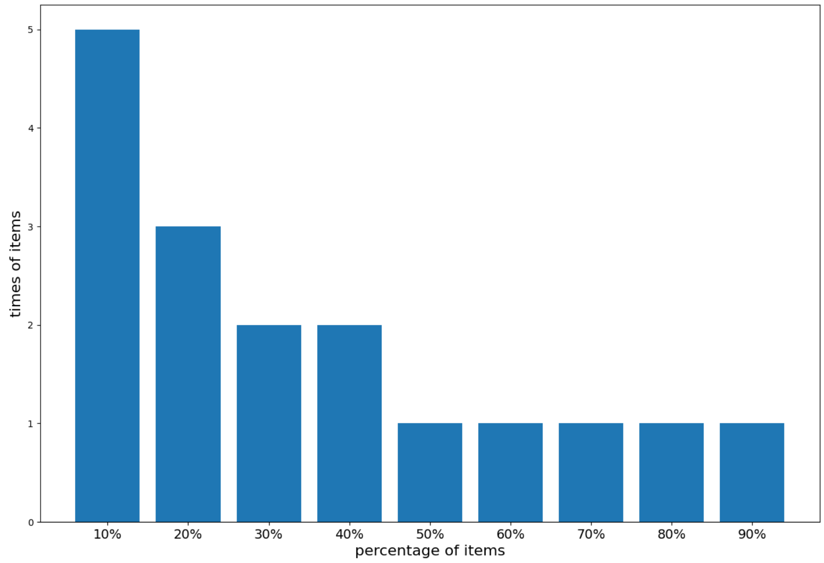
As we can see at the above figure, we can figure out:
Top 10% users minimum number of interactions (Y) = 12, and top 90% users only give one interaction.

Q4. Sort the items by the number of interactions, and observe what is the minimum number of times (Y) that the top-X% items have been interacted with? Draw some appropriate figures and briefly explain what your insight is.

A: We can get the percentiles information as below.



```
        Number of interactions
count            320213.000000
mean                  3.231605
std                  11.546786
min                   1.000000
10%                   1.000000
20%                   1.000000
30%                   1.000000
40%                   1.000000
50%                   1.000000
60%                   2.000000
70%                   2.000000
80%                   3.000000
90%                   5.000000
max                2263.000000
```



According to the wiki, the explaination of percentiles said: Pk表示至少有k%的資料小於或等於這個數，而同時也有(100-k)%的資料大於或等於這個數。

As we can see at the above figure, we can figure out:

Top 10% items minimum number of interactions (Y) = 5, and top 90% items minimum number of interactions were only one.

Q5. Did you use other datasets or resources to get more information regarding the users and items?

A: No.

And if you don't, what features do you think would be beneficial and where to find them?

A: Books-Ratings records map with user id and ISBN. So I think it could find out how much like the book through the rating score given by users. And even can recommend books by other users who rated similar books highly.

Q6. Freely study your dataset and come up with a question/idea to analyze.

a. The motivation of the question needs to be meaningful and valuable. Why do you want to study this question? What do you expect to bring out from the analysis?

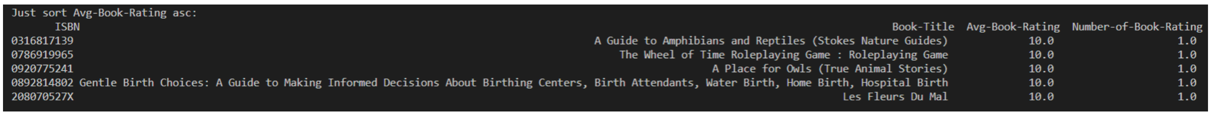
A: I want to know what the top 5 most rated books are. And it’s very suitable for recommending the most rated books to anyone.

b. Define and formulate your question clearly. It is suggested that you provide an example for easy understanding.

A: What are the top 5 most highly rated books? And these books' rating counts have to be more than mean counts.

c. At least 1 figure to show your analysis results and several sentences for explanations. Make sure that the figure is representative enough for the question.

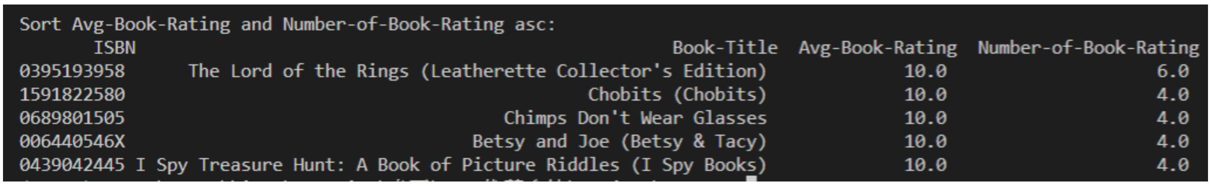
A: First, I get the top 5 most highly rated books without filter rating counts.



```
Just sort Avg-Book-Rating asc:
      ISBN                                                                                                                                   Book-Title  Avg-Book-Rating  Number-of-Book-Rating
0316817139                                                                                    A Guide to Amphibians and Reptiles (Stokes Nature Guides)             10.0                    1.0
0786919965                                                                                        The Wheel of Time Roleplaying Game : Roleplaying Game             10.0                    1.0
0920775241                                                                                                       A Place for Owls (True Animal Stories)             10.0                    1.0
0892814802 Gentle Birth Choices: A Guide to Making Informed Decisions About Birthing Centers, Birth Attendants, Water Birth, Home Birth, Hospital Birth             10.0                    1.0
208070527X                                                                                                                            Les Fleurs Du Mal             10.0                    1.0
```

But the result shows these books can reference only one person’s rating, it’s not objective enough.

So, add the filter of condition on rating counts has to be more than mean counts.



```
Sort Avg-Book-Rating and Number-of-Book-Rating asc:
      ISBN                                                   Book-Title  Avg-Book-Rating  Number-of-Book-Rating
0395193958      The Lord of the Rings (Leatherette Collector's Edition)             10.0                    6.0
1591822580                                            Chobits (Chobits)             10.0                    4.0
0689801505                                    Chimps Don't Wear Glasses             10.0                    4.0
006440546X                                 Betsy and Joe (Betsy & Tacy)             10.0                    4.0
0439042445 I Spy Treasure Hunt: A Book of Picture Riddles (I Spy Books)             10.0                    4.0
```

Then, we can get these top 5 highly rated books for more objective enough.

d. Did the results turn out to be the same as your expectation? Why or why not?

A: Yes. Because the result can reference the rating records more than one person’s opinion.