**Q1.** (5%) Which labels/features/records can be used to represent user preference (i.e., how much a user likes the item)? Also, what is the total range of the labels/ features/records? Does the higher the value is, the better a user prefers an item, or conversely the lower the better?

A1. We can thru the Book-Rate to represent the user preference.

Total range of is: 0 to 10, and it’s higher the better for Book-Rate for items.

**Q2.** (25%) 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.

A2. First drop the less related columns (eg, "Image-URL-S", "Image-URL-M", "Image-URL-L") then merge the three dataset and do the following steps to cleaning date.

**Step1.** Check if have any data are null or messed in merged data (eg, Str <-> Int)

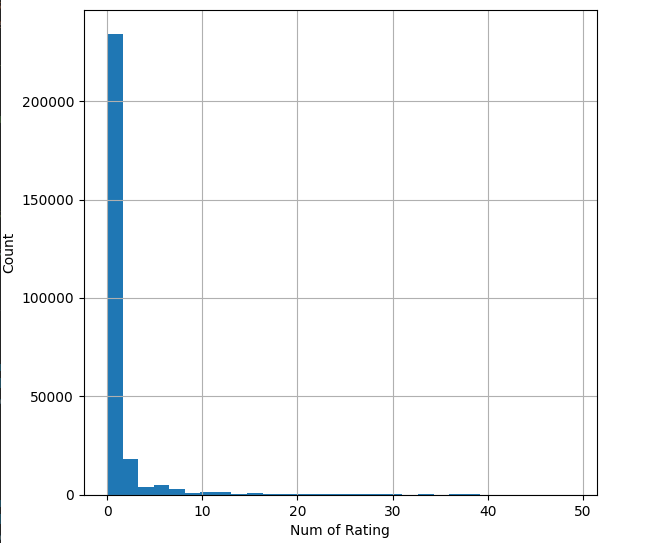
**Step2.** Find out there’ve **few strings** in the ‘Year-Of-Public’, then replace the right number into the columns, and find there’ve **many NA** in columns, the way I chose is the fill the mean in the NA columns.

**Step3.** Find out there’ve many unnormal in the ‘Age’, the way I chose is drop the age which is equal to 0 or above 105 years old, also filled the mean in NA columns.

**Step4.** Find out location have included Country/State/City so need to sperate it.

**Step5.** Since there’ve many 0 points in Books rating, I think the users who will give the 0 points is more like a prank or system algo have an error, some of them maybe is true, but not be so much like this case, so if use the rating data 0 score can be removed.

**Q3.** (15%) 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.

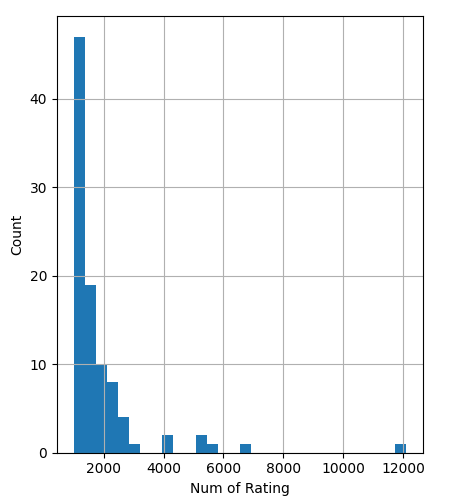
A3. There’ve total 277,286 users, refer figure1 and can know there do have many users are not giving any scores, so I think it ‘outlier’ must need removed.

After removed the 0 rating of users, total users only have 98,002 leave.

一張含有 文字, 橙色 的圖片

自動產生的描述

Then I start to see the distribution of top 25 users, can see except top 1 have over 10 thousand, other users are usually around 1000 to 2000, and after calculated we can know there’ve 96 of 98,002 are > rating of 1000 **(0.1%)**, so I use the cutoff point 1000 to draw a histogram **(the min interaction generated by top 0.1% users are 1000)**

一張含有 桌 的圖片

自動產生的描述x

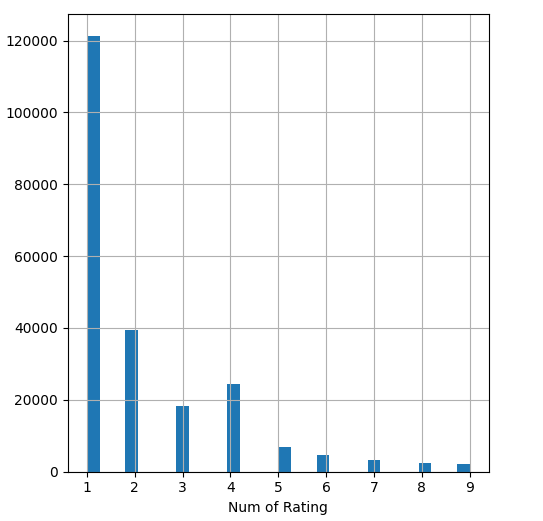
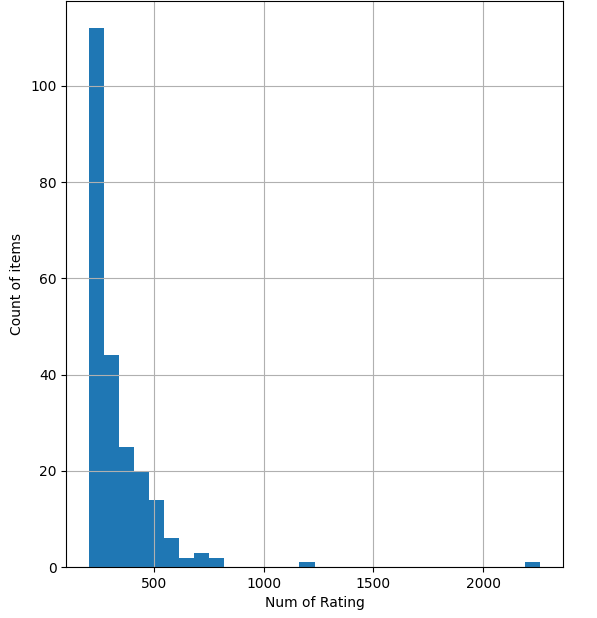
**Q4.** (15%) Sort the items by the number of interactions, and observe what is the minimum number of times (*Y*) that the top-*X*% items have been interacted with? Draw some appropriate figures and briefly explain what your insight is.

A4. There’ve total 238,289 books, and the 0 scores just have 14,093, this seems not like an outlier, so I won’t removed it, I choose filled the mean values into it

一張含有 文字 的圖片

自動產生的描述

Also refer the figure, all the counts are all intensive **under 10**, and after calculated we can know there’ve 230 of 238289 > rating of 200 **(0.1%)**, so the cutoff point I use is the 200 **(the min interaction generated by top 0.1% items are 200)**



**Q5.** (20%) Did you use other datasets or resources to get more information regarding the users and items? If you do, what did you use and why did you choose them? Where did you apply these extra features to help the analysis? And if you don't, what features do you think would be beneficial and where to find them?

A5. I didn’t have find the extra dataset, but I think **user education level and job** are going to have strong correlation with rating of books, like I said there have many users are not giving any scores, so maybe we can search out some specific job or education of users don’t like to give scores

And for books I think can add the **genres of books**, this might help a lot for recommend to specific user who like to read what kind of books

**Q6.** (20%) 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?

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

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.

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

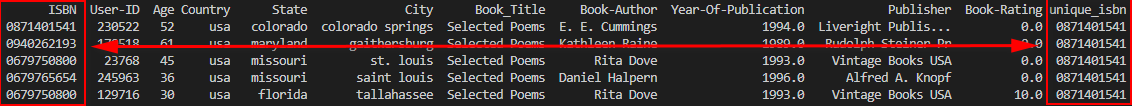
A6. Find out each book will depend on difference publication or publisher will provided the different ISBN, based on our target this might have some problem to trace the user preference (not sure yet), since our goal is to recommend the books for user, not specific publisher of books.

And through the figure we can know around 218,881(90%) the books ISBN is good, so we just need to deal with this 10%,

一張含有 桌 的圖片

自動產生的描述一張含有 文字, 裝置, 儀錶 的圖片

自動產生的描述

My thought is to create the new columns of unique ISBN, every same books which have >2 ISBN going to use the first ISBN to replace all. so can through the figure before cleaning the same book have different ISBN, after I choose the first ISBN to replace all, now the book “Selected Poems” have same ISBN.

I haven’t done my result yet, but will try observe the precision of recommendation have better after unique the ISBN or not.