Hi,

There are several observations I found in the data. First and most importantly, barcode is not having a consistent type in the receipts dataset, and the brands datasets. Thus, when we try to link these two tables, we only get 89 valid data. Within the 89 valid data, there are only 16 unique datasets, which implies that we only identified 16 brands among all items. This hugely weakens our power to analyze -- we have more than 1000 brands, and almost 7000 items, but we can analyze 89 of them. The current dataset is unable to draw any valid conclusion for us because it is not representative enough. In order to resolve this problem, I will further need the information about different naming rules between the receipts dataset and brands datasets.

Then, most rows only have partial columns filled in, that is to say, most dataset does not have full information. For example, there are 3851 items does not have a barcode information, and this would a problem when joining tables. I will not be able to locate brands for these items, for example.

Lastly, there are duplicated rows in the users, which should not be hard to resolve. I would simply delete the duplicated rows. But I am still curious about why the duplicated information exists, and how we can avoid them.