Third: Evaluate Data Quality Issues in the Data Provided

Using the programming language of your choice (SQL, Python, R, Bash, etc...) identify at least one data quality issue. We are not expecting a full blown review of all the data provided, but instead want to know how you explore and evaluate data of questionable provenance.

Commit your code and findings to the git repository along with the rest of the exercise.

Data quality issues in Users data:

-> Users data out 495 entries there are 283 duplicate rows

-> User_id which is the primary key of the table, which is supposed to be unique has only 42.8% distinct values. Duplicates in user id would cause trouble in performing analysis.

-> State field has 11.3% (56) missing values. State field should be set to NOT NULL constraint because it helps us to perform "geographic analysis" or "spatial analysis" i.e., interpreting data in relation to geographic or spatial characteristics.

Data quality issues in receipts data:

-> PurchasedItemCount doesn't match the item count in the rewardsReceiptItemList.

-> there are some userid's in receipts which are not present in user's table. Need to discuss with the data providers regarding this.

Data quality issues in brands data:

->brand code in the brands data has 76.9% distinct values. I suggest brand data should have details of each brand and should be organized properly. I see that is a lot of redundancy while modelling a Database one should normalize this table to avoid redundancy problems.

-> There are 54 records where bar code and brand code have same values.

-> Need to discuss with the data provider for better understanding of brand code and brands data.

Note: Please refer to the code file and data report's for your reference