Detailed Analysis of EDA Results

Receipts Data Analysis

1. Total Rows: 1043

- The dataset consists of 1043 rows, indicating a moderate amount of data for analysis.

2. Percentage of Duplicate Records: 0.00%

- The absence of duplicate records suggests good data quality and uniqueness in the dataset, specifically in the _id field which serves as the primary key.

3. Data Types:

- _id: object

- rewardsReceiptItemList: object

- The data types indicate that the _id field is used for unique identification, while rewardsReceiptItemList contains nested JSON objects that represent detailed receipt items.

4. Normalized rewardsReceiptItemList DataFrame:

- Contains detailed item information with the following distinct counts in categorical columns:

  - item_name: 104 distinct values

  - brand: 23 distinct values

- Analysis:

  - The diversity in item_name (104 distinct items) indicates a wide range of products captured in the receipts.

  - The brand column with 23 distinct values.

  - The normalization process helps in analyzing item-level data separately, making it easier to perform detailed analyses such as frequency counts, price analysis, and item categorization.

Users Data Analysis

1. Total Rows: 9000

- The dataset is relatively large, with 9000 rows, which is substantial for drawing meaningful insights and trends.

2. Percentage of Duplicate Records: 0.00%

- Similar to the receipts data, the absence of duplicate records in the users data signifies good data quality and unique user entries.

3. Data Types:

- _id: object
- creation_date: object
- user_settings: object
- The data types indicate that the _id field is the unique identifier, creation_date records the date of user creation, and user_settings contains nested JSON objects related to user preferences.

4. Categorical Columns and Distinct Counts:

- user_settings.language: 3 distinct values
- user_settings.timezone: 24 distinct values
- Analysis:
  - The language column shows that there are only 3 distinct languages
  - The timezone column with 24 distinct values

Brands Data Analysis

1. Total Rows: 125

- The dataset is relatively small, with 125 rows.

2. Percentage of Duplicate Records: 0.00%

- The lack of duplicate records implies that each brand entry is unique and reliable for analysis.

3. Data Types:

- _id: object
- category: object
- categoryCode: object
- topBrand: object
- brandCode: object
- These data types indicate that the _id field uniquely identifies each brand, while other fields categorize and provide additional information about the brands.

4. Distinct Values in Categorical Columns:

- category: 6 distinct values
- categoryCode: 6 distinct values
- topBrand: 2 distinct values
- brandCode: 5 distinct values
- Analysis:
  - The category and categoryCode columns both having 6 distinct values suggest a straightforward mapping between categories and their codes, making it easier to categorize and analyze brands.
  - The topBrand column indicates whether a brand is a top brand or not, with only 2 distinct values, suggesting a binary classification.
  - The brandCode column having 5 distinct values indicates a few unique brand codes, which could be used to group and analyze brand performance.
  - This data can be used to understand brand distribution across categories, identify top-performing brands, and explore brand code usage.