**Summary Report on Data Quality and Exploratory Findings**

**1. Data Overview**

The transactions.csv dataset contains transactional data with the following columns:

* trans\_id: Transaction ID (integer)
* user\_id: User ID (integer)
* product: Product name (string)
* amount: Transaction amount (string, should be numeric)
* trans\_date: Transaction date (string, should be datetime)

**2. Data Quality Issues**

During the data exploration phase, I have identified several data quality issues that need to be addressed:

1. **Incorrect Data Types**:
   * The amount column is stored as a string, but it should be a numeric type to allow for arithmetic operations.
   * The trans\_date column is stored as a string, but it should be converted to a datetime type for proper date manipulation.
2. **Missing Values**:
   * There is 1 missing value in the trans\_date column. This missing value needs to be handled appropriately.
3. **Non-Numeric Values in amount**:
   * Two rows contain non-numeric values in the amount column which should be corrected.
4. **Non-date Values in trans\_date**:
   * Two rows contain non-date values in the trans\_date column which should be corrected.
5. **Misplaced Data**:
   * One row have date values in the amount column that are supposedly in trans\_date column.

**3. Exploratory Data Analysis**

**Structure and Data Types**:

* The dataset has 84 records.
* The columns are trans\_id (int64), user\_id (int64), product (object), amount (object), and trans\_date (object).

**Summary Statistics**:

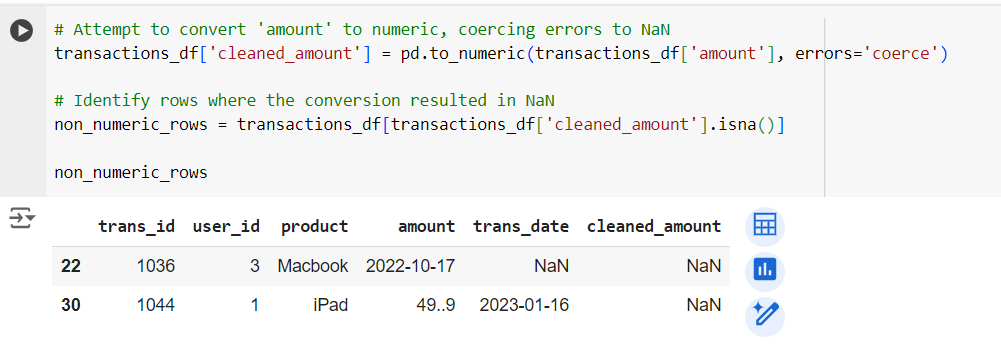
* The trans\_id ranges from 1014 to 1097.
* The user\_id in transaction-1.csv ranges from 1 to 9, but the user\_id in users-1.csv ranges from 1 to 5 only.
* The amount column contains numerical values stored as strings.
* The trans\_date column contains date information stored as strings.

**Missing Values**:

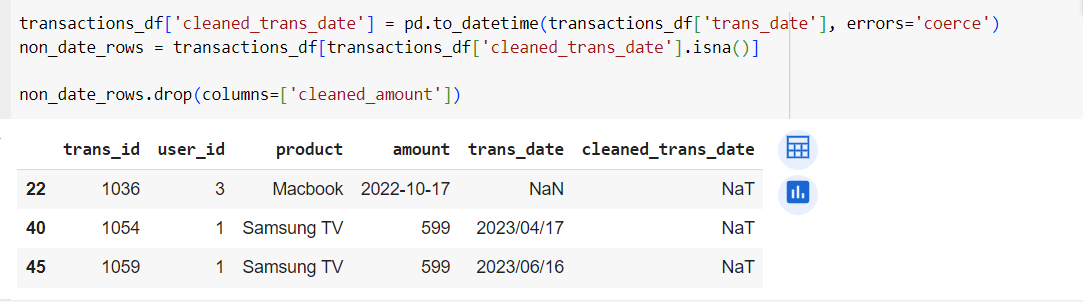
* The trans\_date column has 1 missing value. Other columns have no missing values.

**Identified Non-Numeric Values in amount**:

* Two identified rows with non-numeric values in the amount column:



* Three identified rows with non-date values in the trans\_date column:



**Categorical Columns**:

* The product column contains 5 unique products: "Samsung TV", "iPad", "Macbook", "iPhone", and "AirPods".
* Frequency counts for each product:
  + Samsung TV: 18 occurrences
  + iPad: 18 occurrences
  + Macbook: 18 occurrences
  + iPhone: 18 occurrences
  + AirPods: 12 occurrences

**4. Recommended Data Cleaning and Transformation Steps**

To address the identified data quality issues, the following steps are recommended:

1. **Convert Data Types**:
   * Convert the amount column to a numeric type.
   * Convert the trans\_date column to a datetime type.
2. **Handle Missing Values**:
   * Drop rows with missing values in the trans\_date column for simplicity.
3. **Standardize Product Names**:
   * Ensure consistency in product names by converting them to uppercase.
4. **Calculate Total Spent by Each User**:
   * Group transactions by user\_id and calculate the total amount spent by each user.
5. **Correct Non-Numeric Values in amount**:
   * Identify and correct non-numeric values in the amount column.
6. **Correct Misplaced Data**:
   * Swap values between amount and trans\_date columns where necessary.