

Subject: Analysis and Data Discrepancies in Recent Dataset

Greetings,

I hope this email finds you well. Below is an analysis based on the three datasets we received. Please note that some discrepancies were identified in the data, and several assumptions were made to overcome these challenges. I've outlined the key issues and assumptions, as well as the findings that resulted from them.

Data Discrepancies:

1. **Missing Birth Dates:** Approximately 3.7% of users have missing birth dates in the Users table, which may affect analyses that require age-related segmentation.
2. **Unclear Category Definitions:** The definitions for Categories 1-4 in the Products table are vague. Based on assumptions, the following definitions were applied:
 - **Category 1:** Department
 - **Category 2:** Class
 - **Category 3:** Subclass
 - **Category 4:** Item Type
3. **Missing Brand Information:** 27% of products in the Products table lack an associated brand. This is particularly important for any brand-related analysis and could skew results.
4. **Data Type Inconsistencies:** Several issues with data formats were noted:
 - Dates are recorded in different formats across the datasets.
 - Flagging columns are inconsistently stored as integers and text.
5. **The Limited Date Range** (mid-June 2024 to early September): Restricts the ability to analyze annual performance trends.

Findings Based on Assumptions and Data Discrepancies:

1. **Leading Brands by Receipts (Users Aged 21+):**
 - NERDS CANDY: 6
 - DOVE: 6
 - TRIDENT: 4
 - GREAT VALUE: 4
 - SUAVE: 2
2. **Leading Brands by Sales (Users with Accounts > 6 Months):**
 - CVS: \$72.00
 - TRIDENT: \$46.72
 - DOVE: \$42.88
 - COORS LIGHT: \$34.96
 - QUAKER: \$16.60

3. **Leading Brands for Dips and Salsa:**

- By **Sales**: TOSTITOS (\$239.54)
- By **Receipt Count**: TOSTITOS (66)

Recommendations for Further Analysis:

A more comprehensive analysis could be conducted if the following gaps are addressed:

- **Missing Data**: Resolving missing birth dates and brand associations will improve the reliability of results.
- **Expanded Transaction Table**: Adding more users to the transaction table will offer a better representation of the overall customer base.
- **Wider Date Range**: Including prior year data and expanding the date range for transactions will allow for a more complete view of yearly performance trends, particularly for brands, departments, and other key metrics.

Please let me know if you have any questions or need further clarifications on the findings or recommendations.

Best regards,
Ankitha Husom