Fetch Rewards Data Quality Evaluation

1. Overview

The goal of this document is to identify and evaluate data quality issues within the Fetch Rewards dataset using Python. We will focus on key quality aspects such as completeness, consistency, accuracy, and validity. The analysis will cover the `Users`, `Receipts`, and `Brands` datasets.

---

2. Python Code for Data Quality Analysis

```
import pandas as pd
import json
import numpy as np

# Load the data
users_file = 'users.json'
receipts_file = 'receipts.json'
brands_file = 'brands.json'

def load_json(file):
    with open(file, 'r') as f:
        return pd.json_normalize(json.load(f))

users_df = load_json(users_file)
receipts_df = load_json(receipts_file)
brands_df = load_json(brands_file)

# Function to check missing values
def check_missing_values(df, name):
    print(f"\nMissing Values in {name}:")
    print(df.isnull().sum())

# Function to check duplicate entries
def check_duplicates(df, key_column, name):
    duplicates = df[df.duplicated(subset=[key_column])]
    print(f"\nDuplicate Records in {name} based on {key_column}: {len(duplicates)}")
    if not duplicates.empty:
        print(duplicates.head())

# Function to check data types and inconsistencies
```

```
def check_data_types(df, name):
    print(f"\nData Types and Inconsistencies in {name}:")
    print(df.dtypes)
    for col in df.columns:
        if df[col].dtype == 'object':
            print(f"Unique values in {col}: {df[col].unique()[:5]}")

# Perform data quality checks
check_missing_values(users_df, 'Users')
check_duplicates(users_df, '_id.$oid', 'Users')
check_data_types(users_df, 'Users')

check_missing_values(receipts_df, 'Receipts')
check_duplicates(receipts_df, '_id.$oid', 'Receipts')
check_data_types(receipts_df, 'Receipts')

check_missing_values(brands_df, 'Brands')
check_duplicates(brands_df, '_id.$oid', 'Brands')
check_data_types(brands_df, 'Brands')
```

---

3. Findings and Observations

Users Table:

- **Missing Values:**
  - Some `state` and `signUpSource` fields are missing.
- **Duplicates:**
  - Found duplicate `user_id` values that need to be investigated.
- **Data Type Issues:**
  - `createdDate` and `lastLogin` fields should be converted to timestamps.

Receipts Table:

- **Missing Values:**
  - Some receipts have missing `totalSpent` values.
  - `purchaseDate` is missing in some rows, making it hard to track sales trends.
- **Duplicates:**
  - No duplicate receipt IDs detected.
- **Data Type Issues:**

- `totalSpent` values are sometimes stored as strings instead of floats.

Brands Table:

- **Missing Values:**
  - Some `brandCode` values are missing, which can impact reporting.
- **Duplicates:**
  - No duplicate brand IDs detected.
- **Data Type Issues:**
  - `topBrand` should be a boolean, but some values are stored as strings.

---

4. Recommendations

1. **Handle Missing Values:**
   - Impute missing values where possible (e.g., use default values for `state`).
   - Investigate missing purchase dates and ensure data completeness.
2. **Fix Data Types:**
   - Convert timestamps to `datetime` format for consistency.
   - Ensure numeric fields such as `totalSpent` are stored as floats.
3. **Deduplication:**
   - Investigate and remove duplicate records in the Users table.
4. **Validation Rules:**
   - Implement validation checks during data ingestion to prevent inconsistencies.

---

5. Conclusion

This analysis highlights several data quality issues in the Fetch Rewards dataset that should be addressed to ensure accurate reporting and data integrity. Implementing data validation, transformation, and regular quality checks will help maintain a reliable data warehouse.