Data Quality Assessment

Data Quality Issues

During the examination of the provided data, several instances of incomplete data were observed across multiple tables. Incomplete data can significantly impact data analysis, decision-making processes, and overall data integrity. This document highlights the identification of incomplete data issues and provides a sample SQL query to detect such issues.

SQL Queries

User Table:

```
SELECT *
FROM User
WHERE last_login IS NULL OR state IS NULL OR Signupsource IS NULL;
```

Brand Table:

```
SELECT *
FROM Brand
WHERE brand_code IS NULL;
```

```
SELECT *
FROM Brand
WHERE top_brand IS NULL;
```

Category Table:

```
SELECT *
FROM Category
WHERE category_code IS NULL;
```

```
SELECT *
FROM Category
WHERE category_name IS NULL;
```

```
SELECT *
```

```
FROM Category
WHERE category_code IS NULL OR category_name IS NULL;
```

Receipt Table:

```
SELECT *
FROM Receipt
WHERE bonuspoints_earned IS NULL
   OR bonuspoints_earnedreason IS NULL
   OR finished_date IS NULL
   OR pointsawarded_date IS NULL
   OR total_spent IS NULL
   OR purchaseditem_count IS NULL
   OR points_earned IS NULL
   OR purchase_date IS NULL;
```

These queries will help you identify rows with missing values or inconsistencies in the data, enabling you to take corrective actions to improve the overall data quality and integrity.

Explanation of SQL Queries

The SQL queries provided are designed to identify instances of missing or incomplete data within each table of the provided dataset. By executing these queries, data analysts can pinpoint specific records that require attention and remediation to enhance data quality and integrity.

For instance, queries targeting the User table isolate rows where last_login, state, or Signupsource information is absent, facilitating subsequent data cleansing efforts. Similarly, queries applied to the Brand, Category, and Receipt tables enable the identification of records with missing values in critical fields, aiding in the detection and resolution of data quality issues across different dimensions of the dataset.

These queries serve as valuable tools for data quality assessment and form part of a broader strategy to ensure the reliability, completeness, and consistency of organizational data assets. Through systematic identification and resolution of data quality issues, organizations can enhance decision-making processes and derive greater value from their data resources.

Identifying Data Quality Issues

Data quality issues can arise from various sources such as data entry errors, inconsistencies, missing values, outliers, or discrepancies between different data sources. Below are some SQL queries to help identify potential data quality issues in the Fetch Rewards database:

Detecting Duplicates:

```
SELECT receipt_id, COUNT(*)
FROM Receipt
GROUP BY receipt_id
HAVING COUNT(*) > 1;
```

This query identifies duplicate entries in the Receipt table based on the receipt_id field.

Checking for Inconsistent Data Types:

```
SELECT *
FROM Receipt
WHERE NOT ISDATE(purchase_date);
```

This query checks if the purchase_date field contains inconsistent data types, such as non-date values.

Identifying Outliers in Numeric Fields:

```
SELECT *
FROM Receipt
WHERE total_spent < 0 OR purchaseditem_count < 0 OR points_earned < 0;
```

This query identifies receipts with negative values in numeric fields like total_spent, purchaseditem_count, or points_earned.

Exploring and Evaluating Data of Questionable Provenance

When dealing with data of questionable provenance, it's crucial to take a systematic approach to explore and evaluate its quality and reliability. Here's how you can proceed:

Data Profiling:
Perform data profiling to gain insights into the data distribution, patterns, and anomalies. Use summary statistics, histograms, and frequency distributions to identify outliers, missing values, and data inconsistencies.

Data Quality Assessment:
Assess the completeness, accuracy, consistency, and integrity of the data. Verify data against predefined business rules, domain constraints, and validation criteria. Flag records with anomalies or inconsistencies for further investigation.

Data Cleaning and Standardization:
Implement data cleaning and standardization techniques to address missing values, duplicates, outliers, and inconsistencies. Use techniques like imputation, normalization, and data transformation to improve data quality and consistency.

Data Integration and Validation:
Integrate data from multiple sources and validate it against reference data or authoritative sources. Resolve discrepancies and conflicts between different data sets by establishing data reconciliation processes and resolving data conflicts through consensus or arbitration.

Data Documentation and Metadata Management:

Document data lineage, sources, transformations, and quality assessments to establish transparency and traceability. Maintain metadata repositories and data dictionaries to provide context and insights into the data's origin, structure, and semantics.

Collaborative Review and Feedback:
Engage stakeholders, subject matter experts, and data custodians in collaborative reviews and feedback sessions to validate data assumptions, interpretations, and decisions. Leverage domain knowledge and expertise to validate data quality and relevance in the context of business objectives and requirements.

By following these practices, you can systematically explore and evaluate data of questionable provenance, improve its quality and reliability, and make informed decisions based on trustworthy and actionable insights.