Data Intake Report

Name: Cab Usage Analysis
Report date: 07.14.2024
Internship Batch: LISUM35
Version: 1.0
Data intake by: Alua Birgebayeva
Data storage location: https://github.com/alua222/eda_notebook

Tabular data details: Cab_Data

Total number of observations	359392
Total number of features	7
Base format of the file	csv
Size of the data	19.2MB

Tabular data details: Customer_Data

Total number of observations	49171
Total number of features	4
Base format of the file	csv
Size of the data	1.51MB

Tabular data details: Transaction_Data

Total number of observations	440098
Total number of features	3
Base format of the file	csv
Size of the data	10.08MB

Tabular data details: City_Data

Total number of observations	20
Total number of features	3
Base format of the file	csv
Size of the data	0.01MB

Tabular data details: Holiday_Data

Total number of observations	9
Total number of features	2
Base format of the file	tibble
Size of the data	none

Proposed Approach:

Deduplication Validation (Identification):

1. **Identification of Duplicates**:
    - Use functions to identify and remove duplicate rows in each dataset. For example:
        - For `cab_data`, `customer_data`, `transaction_data`, and `city_data`, ensure each entry is distinct.
2. **Validation of Key Fields**:
    - Ensure that key fields such as `Transaction ID` and `Customer ID` are unique.
    - Check for duplicates in these key fields to maintain data integrity.

Assumptions for Data Quality Analysis:

1. **Date Formatting**:
    - Assume that dates provided in integer format (e.g., Excel serial date) need conversion to a standard date format.
    - Convert these dates to a readable format to ensure consistency.
2. **Handling Missing Values**:
    - Assume that NA values in key columns (e.g., `Company`, `Holiday`) need to be handled appropriately.
    - Fill NA values with placeholders or remove rows with missing key information to avoid analysis errors.
3. **Consistency of Categorical Data**:
    - Assume that all categorical data (e.g., `Company` names, `City` names) should be consistent across files.
    - Normalize the data if necessary to ensure consistency across datasets.

By following this proposed approach, I ensured that the datasets are clean, consistent, and ready for further analysis..