Data Intake Report

Name: G2M Strategy for Investment in Cab Company

Report date: 21-11-2023

Internship Batch: LISUM 27

Version:<1.0>

Data intake by: Siddharth Dudugu

Data intake reviewer: -

Data storage location: GitHub:

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 359391 |
| **Total number of files** | 1 |
| **Total number of features** | 19 |
| **Base format of the file** | ipynb |
| **Size of the data** | - |

**Proposed Approach:**

Deduplication Validation Approach:

1. Identification of Duplicates:
   * Utilized unique identifiers such as 'Transaction ID' and 'Customer ID' for identifying potential duplicate records.
   * Employed pandas functions like duplicated() to identify and flag rows with identical values in key fields.
   * Considered additional fields such as 'Date of Travel,' 'City,' etc., to enhance the accuracy of duplicate identification.
2. Handling Duplicates:
   * Created a new deduplicated dataset using the drop\_duplicates() function in pandas.
   * Selected appropriate criteria for keeping the first occurrence or last occurrence based on the context of the data.
3. Verification and Validation:
   * Reviewed and compared the deduplicated dataset with the original to ensure essential information retention.
   * Validated the impact of deduplication on key metrics, including profit, demand, and customer loyalty.
   * Assessed the impact on analyses and visualizations to ensure consistency and reliability.

Assumptions for Data Quality Analysis:

1. Consistency of Unique Identifiers:
   * Assumed that 'Transaction ID' and 'Customer ID' are consistent and accurately represent unique records.
   * Verified that these identifiers are not duplicated or misassigned.
2. Data Integrity:
   * Assumed that data integrity is maintained, ensuring accurate and consistent values in critical fields such as 'Date of Travel' and 'City.'
3. Assumption of Accuracy:
   * Assumed that the provided data accurately reflects the transactions, customer details, and financial information of the cab companies.
   * Conducted verification of accuracy through cross-referencing with external sources and random sampling checks.
4. Handling Missing Values:
   * Assumed that missing values were appropriately handled during data preprocessing.
   * Identified and addressed any remaining missing values during the analysis.
5. Assumption of Timeliness:
   * Assumed that the data reflects the specified time period (from 31/01/2016 to 31/12/2018) with no outliers or discrepancies.
6. Currency of External Data:
   * If third-party data (e.g., weather data) was utilized, assumed that it is current and relevant to the time period under analysis.

These assumptions were made in the context of our data quality analysis, and the entire deduplication process has been completed, ensuring the reliability and accuracy of our findings.