Data Intake Report

Name: G2M Insight for cab Investment Firm
Report date: 14/03/2024
Internship Batch: LISUM31
Version: 1.0
Data intake by: Aditya Mukherjee
Data intake reviewer
Data storage location: Github

Tabular data details:

Cab Dataset:

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	20663 kb

Transaction ID Dataset:

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	8788 kb

Customer ID Dataset:

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1027 kb

City Dataset:

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	1 kb

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

The deduplication process is crucial for ensuring the integrity and reliability of the dataset before conducting any analysis. Our approach to deduplication validation includes the following steps:

- Initial Dataset Inspection: Begin with a preliminary inspection of each dataset to identify obvious duplicates or inconsistencies. This involves checking for rows where all or a subset of key columns are identical.

- Unique Identifier Verification: Ensure that supposed unique identifiers, such as Transaction ID and Customer ID, do not have duplicates within their respective datasets. This can be achieved using aggregation methods to count occurrences of each identifier and flagging any counts greater than one.

- Combining Datasets: When merging datasets, we check for inconsistencies that could introduce duplicates, such as differing formats or multiple entries for a single identifier across tables. For instance, a single Transaction ID should not be linked to multiple Customer IDs unless the business logic explicitly allows for this scenario.

- After-Merge Check: Post-merging, examine the resulting dataset for duplicates that could have arisen from the merging process. This could include duplicated transactions or customer entries that were not evident before the datasets were combined.

- Row-wise Deduplication: For the final, merged dataset, apply row-wise deduplication, considering all relevant columns. This step involves removing duplicate rows that do not add unique information to the dataset.

Assumptions for Data Quality Analysis

- During our data quality analysis, we made several assumptions to streamline the process and ensure the analysis was focused and effective:

- Data Completeness: We assumed that the datasets provided were complete and represented the entire scope of transactions, customers, and operational data required for analysis. Missing data, unless obviously problematic, was presumed to be minimal or non-impactful.

- Correctness of Unique Identifiers: We assumed that Transaction ID and Customer ID fields were correctly assigned and unique where expected. Any anomaly in these identifiers was treated as a data quality issue.

- Standardization of Formats: Where data from different datasets were merged (e.g., dates, city names), we assumed that the formats were standardized or could be standardized without loss of information. Any deviation from standard formats required correction before analysis.

- Accuracy of Transactional Data: It was assumed that the transactional data (e.g., KM Travelled, Price Charged, Cost of Trip) were accurately recorded at the source. The analysis did not incorporate a mechanism to verify the accuracy of these records against external sources.

- Temporal Consistency: For the analysis of trends and seasonality, we assumed that the Date of Travel data points were consistently recorded and reflective of actual travel dates. This is crucial for accurately assessing patterns and seasonal effects.

- These assumptions were necessary to maintain focus on actionable insights and recommendations, acknowledging that perfect data quality is often unattainable in real-world scenarios. However, any significant deviations from these assumptions would need to be addressed through data cleaning, transformation, or further validation steps.