Data Intake Report

Name: G2M insight for Cab Investment
Report date: 14-11-2023
Internship Batch: LISUM27
Version:1.0
Data intake by: Ayo-John Oluwaseun
Data intake reviewer:
Data storage location: https://github.com/Sarjzman/VC

Tabular data details:

Total number of observations	282193
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	90.6 MB

Proposed Approach:

Dedup validation (identification) ; To maintain transparency and facilitate reproducibility, the deduplication approach and methods were thoroughly documented. The documentation outlines the columns used for deduplication, the Python functions applied, and the rationale behind choosing specific deduplication strategies.

The rigorous deduplication process undertaken in Python ensures that the dataset used for subsequent analysis is free from redundancies, contributing to the reliability and accuracy of the findings. By adopting appropriate methods and validating the results, the integrity of the data is preserved, instilling confidence in the outcomes of the analysis.

Assumptions and Data Quality Decisions:

Transaction ID and Customer ID as Unique Identifiers:

For the purpose of analysis, each unique transaction ID is considered to represent a distinct ride, providing a basis for tracking individual transactions. Similarly, each unique customer ID is treated as a unique customer, allowing for customer-centric analysis.
Handling Outliers:

The 'Date of Travel' column, 'Users,' and 'Population' were identified as outliers due to inconsistencies. As a result, these data points were considered for removal to ensure the integrity and accuracy of the analysis. Outliers, in this context, refer to values that deviate significantly from the expected or typical range, potentially impacting the reliability of insights derived from the dataset.

These decisions are essential for maintaining data quality, ensuring that the dataset is well-structured, and aligning with the analytical goals. Removing outliers helps mitigate the impact of inconsistent or erroneous data, contributing to more reliable and meaningful analyses.