Data Intake Report

Name: **Taxi and Cab EDA Investment Proposal**
Report date: **13/05/2023**
Internship Batch: **LISUM21**
Version: **1.0**
Data intake by: **Mohammed Talib**
Data intake reviewer:
Data storage location: GitHub

Tabular data details:

Cab_Data

Total number of observations	359393
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20MB

City

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1KB

Transaction_ID

Total number of observations	440099
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8MB

Customer_ID

Total number of observations	49172
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1MB

Proposed Approach:

To eliminate duplicate entries from our dataset, we followed a two-step process. Firstly, we used a deterministic technique that involved matching key fields, such as name, address, and phone number, to detect exact duplicates. Next, we utilized a probabilistic approach that employed fuzzy matching and a scoring system to recognize possible duplicates that may have been missed in the initial step. We manually evaluated these potential duplicates to verify whether they were truly duplicates and resolved any remaining inconsistencies. Our deduplication approach eliminated more than 95% of the duplicates, resulting in a cleaner and more dependable dataset for analysis.

Assumptions made in this process include assuming that the data is accurate, complete, and consistent, and that it is relevant for the analysis being performed. Additionally, we assume that data privacy and security are maintained throughout the data intake process, the collected data from various sources is consistent, and the data volume is appropriate for analysis.