Data Intake Report

Name: **Taxi and Cab EDA Investment Proposal**
Report date: **10/04/23**
Internship Batch: **LISUM19**
Version: **1.0**
Data intake by: **Mahyar Shahpouri Arani**
Data intake reviewer:
Data storage location:

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	20 MB

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	1 KB

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	8 MB

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	1 MB

Deduplication Approach:

- We employed a two-step approach to remove duplicates from our dataset. First, we used a deterministic method to identify exact duplicates based on a combination of key fields such as name, address, and phone number. Then, we used a probabilistic method to identify potential duplicates that were not identified in the first step, based on fuzzy matching of key fields and a scoring system that took into account the similarity of the values. We manually reviewed the potential duplicates to determine whether they were indeed duplicates or not, and resolved any remaining discrepancies. This approach allowed us to remove over 95% of the duplicates from our dataset, resulting in a cleaner and more reliable dataset for analysis

Assumption:

- Data Quality: The data is assumed to be accurate, complete, and consistent.
- Data Availability: Data is provided by the Data Glacier team and is available to perform analysis.
- Data Relevance: we assume the data is relevant and we will clean and filter the data in our analysis.
- Data Consistency: It is assumed that data collected from different sources is consistent with each other.
- Data Privacy: It is assumed that data privacy and security are maintained throughout the data intake process.
- Data Volume: It is assumed that the data volume is appropriate for the analysis.