# Data Intake Report

Name: Cab Industry EDA – Investment Assessment for XYZ
Report date: 14-09-2025
Internship Batch: LISUM49
Version: 1.0
Data intake by: Magdalena Ivanova
Data intake reviewer: -
Data storage location: https://github.com/magdalena-ivanova-ds/Data-Glacier-Virtual-Internship/tree/main/week2

**File: Cab_Data.csv**

| | |
|---|---|
| **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** | 20 663 KB |

**File: Customer_ID.csv**

| | |
|---|---|
| **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** | 1 027 KB |

**File: Transaction_ID.csv**

| | |
|---|---|
| **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** | 8 788 KB |

**File: City.csv**

| | |
|---|---|
| **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 |

**Proposed Approach:**

• Deduplication validation:

1. Duplicate IDs are checked in each source table (Cab_Data, Transaction_ID, Customer_ID) before merging but not explicitly removed from the master dataset.
2. 2. Transaction ID and Customer ID are used as join keys across datasets to build the master table.
3. 3. Duplicate counts are displayed for each source table; merging uses left joins to preserve all cab records.

• Assumptions:

1. The Users column in City.csv is interpreted as the number of cab users in each city; Population is the total city population.
2. Monetary values (Price Charged, Cost of Trip) are used for revenue/cost/profit calculations; currency is assumed to be USD.
3. Dates are parsed from Excel serial format, not ambiguous string dates.
4. Missing values are left as-is in non-critical fields; profit and margin are calculated only when both Price Charged and Cost of Trip are available.