Data Intake Report

Name: Cab Industry Analysis – Investment Readiness Assessment
Report date: 2025-06-13
Internship Batch: LISUM46: 30 May - 30 Aug 25
Version: 1.0
Data intake by: Ruslan Kurmashev
Data intake reviewer: N/A
Data storage location: https://github.com/Ruslan-Kurmashev/Data-Glacier-Internship-W2---cab-investment-case-

Tabular data details:

Total number of observations	359392
Total number of files	4
Total number of features	7
Base format of the file	.csv
Size of the data	29.7 MB

Proposed Approach:
- Deduplication validation was performed using `Transaction ID` and `Customer ID`. Duplicates were checked using `.duplicated()` on primary identifiers. No duplicate trip records were found.
- Assumptions:
  • Each row represents a unique trip.
  • `Profit` was computed as `Price Charged - Cost of Trip`.
  • Dates were converted from Excel serial format.
  • Joins between tables were assumed to be one-to-one via `Customer ID` and `Transaction ID`.
  • Outliers in price and profit were not removed unless they were data errors, as they reflect real business behavior.

Additional Data Notes

Missing Values
None detected across key columns after merging. No imputation was required. All relevant fields are complete and ready for analysis.
Feature Engineering
The following derived features were created to support business and behavioral analysis:
- Profit = Price Charged - Cost of Trip
- Profit_per_KM
- Trip_Category (Short, Medium, Long)
- Income_Class (Low, Medium, High)
- Year, Month, Weekday
- Day_Type (Weekend/Weekday)
- Distance_Bin (binned distance variable)

Data Join Validation
All joins between tables were performed using primary identifiers (Transaction ID, Customer ID, and City). The number of rows in the final merged dataset (359,392) matches the original trip-level table, confirming that no records were lost or duplicated.