Data Intake Report

Name: XYZ’s Cab Company Investment Case Study
Report date: MAR 2025
Data intake by: Raj Pawar
Data storage location: Local (Original file source - GitHub)

Tabular data details:

1. **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	20.1 MB

2. **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	759 bytes

3. **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 MB

4. **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	8.58 MB

Proposed Approach:

1. Deduplication Validation:

- Checked for full-row duplicates in each file using duplicated() method
- Verified uniqueness of key identifiers: Transaction ID, Customer ID
- No major duplicate issues found in base files
- After merging, validated again to ensure no duplication carried over

2. Assumptions for Data Quality:

- Date of Travel column in Excel-style number format, converted to datetime
- Population and Users columns in City.csv were stored as strings with commas and then cleaned and converted to integers
- Income data assumed to be monthly in USD
- Data assumed to be consistent in time range (Jan 2016 – Dec 2018)
- All missing value checks returned zero/null, so assumed no major NA issues
- Basic outlier and negative value checks done (price, distance, cost)