Data Intake Report – Cab Services Analysis

Report Date: April 15th, 2025

Prepared by: Fangning Luo

## 1. Data Sources Overview

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Description** | **Source** | **Format** | **Size** | **Date Range** |
| Cab\_Data.csv | Trip details incl. transaction ID, date or travel, company, city, distance travelled, price charged, and cost of trip | Provided CSV | CSV | 359,392 rows | 2016/1/2 – 2018/12/31 |
| City.csv | City population and user base | Provided CSV | CSV | 20 rows | - |
| Customer\_ID.csv | Customer demographic info | Provided CSV | CSV | 49,171 rows | - |
| Transaction\_ID.csv | Link between trip and customer | Provided CSV | CSV | 440,098 rows | 2016/1/2 – 2018/12/31 |

## 2. Schema Summary

### a. Cab

| **Column** | **Data Type** | **Notes** |
| --- | --- | --- |
| Transaction ID | String | Primary key |
| Date of Travel | Int | Original trip date (serial date format since 1900/1/1) |
| Company | String | 'Pink Cab' or 'Yellow Cab' |
| City | String | Matches city dataset |
| KM Travelled | Float | Distance in kilometers |
| Price Charged | Float | Revenue from customer |
| Cost of Trip | Float | Operational cost |
| Year, Month, Day | Integer | Extracted for temporal analysis |

### b. City

| **Column** | **Data Type** | **Notes** |
| --- | --- | --- |
| City | String | Matches city dataset |
| Population | Float | Number of Population in the city |
| Users | Float | Number of Users in the city |

### c. Customer

| **Column** | **Data Type** | **Notes** |
| --- | --- | --- |
| Customer ID | String | Customer ID |
| Gender | String | Gender (Female / Male) |
| Age | Integer | Age of customer |
| Income (USD/Month) | Integer | Income per month |

### d. Transaction

| **Column** | **Data Type** | **Notes** |
| --- | --- | --- |
| Transaction ID | String | Primary key |
| Customer ID | String | Customer ID |
| Payment Mode | String | Cash / Credit Card |

## 3. Data Volume

* Total trips: 359,392
* Time period: Jan 2nd, 2016 – Dec 31st, 2018
* Number of cities: 20
* Number of unique customers: 49171

## 4. Initial Data Quality Summary

| **Check** | **Status** | **Notes** |
| --- | --- | --- |
| Missing values | ✅ None | No missing value found |
| Duplicate entries | ✅ None | No duplicates in primary keys |
| Date format consistency | ✅ Clean | All dates parsed successfully |
| Mismatched keys (joins) | ✅ None | All transaction have matching customer ID |

## 5. Preprocessing Summary

* Converted serial data format to Year-Month-Day format in `cab` dataset
* Merged `cab` + `transaction` using `Transaction ID` to get `cab\_customer` dataset
* Added derived columns: profit, profit margin, year-month key

## 6. Risks and Assumptions

* Population in `city` dataset is assumed static over years
* No time variation in `customer` demographics (e.g. age remains constant)
* Users in `city` dataset assumed to represent unique customer base, though overlap between cities is unclear