Data Intake Report – Week 2 (Cab Investment Case Study)

Prepared by: Naga Pavithra Jajala

Date: 06.04.2025

Purpose:

To understand the quality, structure, and integration approach of the datasets provided for XYZ's Cab Industry investment analysis.

Dataset Summary

Dataset Name	Description
Cab_Data.csv	Contains ride-level details for two companies, including travel date, kilometers travelled, price charged, and cost of trip.
Customer_ID.csv	Contains demographic information of customers (Gender, Age, Income).
Transaction_ID.csv	Maps transaction IDs to customer IDs and their payment mode.
City.csv	Contains details about cities, including population and number of cab users.

Data Shape (Rows & Columns)

Dataset	Rows	Columns
Cab_Data	359,392	7
Customer_ID	49,171	4
Transaction_ID	440,098	3
City	20	3

(Values based on .shape method)

Missing Value Analysis

Each dataset was checked for missing (null) values using .isnull().sum().
Result: No missing values were found in any of the datasets. No imputation was necessary.

Duplicate Records

All datasets were checked using .duplicated().sum().

Dataset	Duplicate Rows Found
Cab_Data	0
Customer_ID	0
Transaction_ID	0
City	0

No duplicate rows were present. No action required.

Data Merge & Master Table Creation

The following merges were performed to create a single unified dataset (merged_df) for analysis:

1. **Cab_Data** merged with **Transaction_ID** using Transaction ID
    - **Join type:** inner
    - **Reason:** To retain only valid transactions from both tables.
2. The result was merged with **Customer_ID** using Customer ID
    - **Join type:** inner
    - **Reason:** To enrich transaction data with customer demographic info.
3. Final merge with **City** using City
    - **Join type:** left
    - **Reason:** To retain city information even if some cab records lacked a match.

Final Dataset Summary

- Final merged dataset: **merged_df**
- Shape after merging: **[Insert rows] rows × [Insert columns] columns** (use merged_df.shape)
- Ready for EDA and visualization