Data Intake Report

Overview of Data Sources

This project includes four primary datasets provided for analysis. Each dataset is critical for answering business questions

around customer demographics, cab usage, and financial performance across two cab companies:

1. **Cab_Data.csv**: This dataset contains detailed records of cab transactions, including trip distance, price charged, and cost of trip.

It is essential for analyzing the financial performance and operational details of each trip.

2. **Transaction_ID.csv**: This file maps transactions to specific customers and records the payment mode used.

Linking this file to Cab_Data provides a connection to customer data, enabling analysis of customer behavior.

3. **Customer_ID.csv**: This dataset includes demographic information for each customer, such as age, gender, and monthly income.

This information will help us explore customer segments and understand demographic influences on cab usage.

4. **City.csv**: Provides city-level data, including population and the number of cab users. This data enriches the analysis by

adding geographical context and enabling city-wise comparisons of cab demand and usage.

Data Quality Assessment

Each dataset was reviewed for data quality, focusing on missing values and duplicates. Here are the

findings:

- **Missing Values**: None of the datasets contained missing values, which suggests that the data collection process was consistent.

- **Duplicates**: No duplicate rows were found in any of the datasets, indicating that each record is unique and accurately represents
  individual transactions or customer entries.

- **Outliers**: Outlier detection identified some high values in the 'Price Charged' field. These will be examined further during analysis
  to understand if they represent premium services, long trips, or potential errors.

Overall, the data is clean and ready for integration, with only minor outliers to consider.

Integration Strategy

The integration process aims to combine all datasets into a comprehensive master dataset for analysis.
Here is the approach to achieve this:

- **Step 1**: Merge Cab_Data with Transaction_ID using the 'Transaction ID' field. This links transaction details with customer IDs and payment modes.
- **Step 2**: Merge the resulting data with Customer_ID on 'Customer ID' to add customer demographics.
- **Step 3**: Merge with City data on 'City' to provide geographical insights into cab usage patterns and market demand.

This master dataset will support in-depth analysis by consolidating financial, demographic, and geographical data for each transaction.

Preliminary Insights

During the initial exploration, several insights emerged that will shape the focus of the analysis:

- **Customer Demographics**: Both cab companies have a diverse age and income demographic, with slight variations that may influence
  customer preferences and spending habits.

- **Revenue and Profit**: Preliminary aggregation shows that one company achieves significantly higher revenue and profit, suggesting
  differences in pricing, distance traveled, or market positioning.

- **Geographical Demand**: Trip volumes and revenues vary considerably by city, with certain cities showing much higher demand.
  This points to potential geographic markets that may be more lucrative or underserved.

These findings provide a foundation for formulating hypotheses and investigating patterns, such as seasonality, demographic influences,
and city-specific trends.