Data Intake Report

Project Title: Cab Industry Analysis for Investment Decision

Prepared by: Ira Mishra

Date: 03/16/2025

1. Objective

The purpose of this report is to document the data intake process for the exploratory data analysis (EDA) phase of the cab industry analysis. The analysis aims to provide insights into customer demographics, cab company performance, and potential investment opportunities.

---

2. Data Sources

The following datasets have been provided and used for analysis:

1. **Cab Data**: Contains details about rides, pricing, and company information.
2. **Customer Data**: Includes customer demographic details such as age, gender, and income.
3. **Transaction Data**: Captures transaction details including payment method and amount.
4. **City Data**: Provides information on city-wise population and number of users.

All datasets are stored in CSV format and sourced from [Google Drive/GitHub].

---

3. Data Dictionary

Column Name	Dataset	Description
Transaction ID	Transaction Data	Unique ID for each transaction
Customer ID	Customer Data	Unique identifier for customers

Payment Mode	Transaction Data	Mode of payment (Cash/Card)
Gender	Customer Data	Gender of the customer
Age	Customer Data	Age of the customer
Income (USD/Month)	Customer Data	Monthly income of the customer
Date of Travel	Cab Data	Date when the trip occurred
Company	Cab Data	Name of the cab company (Pink Cab/Yellow Cab)
City	City Data	City where the ride occurred
KM Travelled	Cab Data	Distance traveled in kilometers
Price Charged	Cab Data	Amount charged to the customer
Cost of Trip	Cab Data	Cost incurred by the cab company
Population	City Data	Population of the respective city
Users	City Data	Number of users in that city

---

4. Data Quality Assessment

Check	Status
Missing Values	Identified & to be handled
Duplicate Records	No major duplicates found
Data Consistency	Some inconsistencies in city names
Outliers Detection	Outliers detected in Price Charged
Format Issues	Date column needs formatting

---

5. Assumptions & Considerations

- Data is assumed to be representative of the overall cab industry trends.
- Missing values in demographic fields will be handled using median/mode imputation.
- Revenue will be calculated based on 'Price Charged' column.

- City-wise user data will be used to understand market penetration.

---

6. Next Steps

- **Data Cleaning**: Address missing values, format inconsistencies, and duplicate records.
- **EDA Execution**: Generate summary statistics, visualizations, and initial insights.
- **Hypothesis Formation**: Identify key business questions and validate using data.

Deliverables for Week 2:

- EDA Notebook
- Data Intake Report (this document)
- EDA Recommendations & Hypothesis Results
