Data Intake Report

Project: Cab Company Investment Analysis for XYZ

Data Source: The data was provided by the client, XYZ, and consists of four datasets containing information on two cab companies operating in the US. The data covers a period from January 31, 2016, to December 31, 2018.

Data Description:
The four datasets include:
- Cab_Data.csv: Transaction details for two cab companies.
- Customer_ID.csv: A mapping table containing unique identifiers linking customer demographic details.
- Transaction_ID.csv: A mapping table containing transaction to customer mapping and payment mode.
- City.csv: A list of US cities, their population, and the number of cab users.

Data Quality Assessment:
An initial assessment of data quality revealed that some missing values and duplicate records need to be addressed. Data preprocessing steps, such as imputation or dropping rows/ columns, will be performed accordingly.

Data Structure:
The data is structured in four separate tables, which need to be merged or joined to create a comprehensive dataset for analysis. Relationships between tables are based on transaction IDs, customer IDs, and city names.

Variable Descriptions:
After merged the four datasets and preprocessed the data, we extracted the following features:
1. Transaction ID: A unique identifier for each transaction.
2. Date of Travel: The date on which the cab ride took place.
3. Company: The cab company's name (either Pink Cab or Yellow Cab).
4. City: The city where the cab ride took place.
5. KM Travelled: The distance traveled in kilometers during the cab ride.
6. Price Charged: The total fare charged to the customer.
7. Cost of Trip: The cost incurred by the cab company for the trip.
8. Year: The year of the transaction.
9. Margin: The profit margin of the trip (Price Charged - Cost of Trip).
10. Customer ID: A unique identifier for each customer.
11. Payment_Mode: The payment method used by the customer (Card or Cash).
12. Gender: The gender of the customer.
13. Age: The age of the customer.
14. Income (USD/Month): The customer's monthly income in USD.
15. Month: The month of the transaction.
16. Income Class: The income class of the customer (e.g., Low, Medium, or High).
17. Age Group: The age group of the customer (e.g., 20-29, 30-39).
18. Users
19. Population

Initial Findings:
A preliminary examination of the data suggests differences in pricing, customer preferences, and profit margins between the two cab companies. Further analysis will be required to identify patterns, trends, and relationships **that may influence XYZ's investment decision.**

Limitations and Assumptions:

The data is limited to a three-year period, which may not capture long-term trends or recent changes in the cab industry.

The data only covers two cab companies, and the results may not be generalizable to the entire cab industry.

It is assumed that the data accurately reflects real-world transactions and customer behavior.

External factors, such as weather, holidays, or local events, are not considered in the datasets but may impact cab usage and preferences.