Data Intake Report

Project overview

XYZ, a private investment firm in the US, is evaluating the cab industry to identify the most profitable company for investment. This analysis uses data provided on two cab companies to generate actionable insights and support XYZ's decision-making process.

The deliverables include:

1. An Exploratory Data Analysis (EDA) Notebook.
2. A Data Intake Report (current document).
3. Insights and recommendations on the cab companies.

Data Overview

The project uses the following four datasets:

- Cab_Data.csv: Contains transaction details for two cab companies, including travel date, distance, price, and cost.
- City.csv: Provides population and number of cab users for various US cities.
- Customer_ID.csv: Contains demographic information for customers (age, gender, income).
- Transaction_ID.csv: Maps transactions to customer IDs and includes the payment mode.

Data Cleaning Steps

The following steps were performed to ensure clean, usable data for analysis:

1. Handling Missing Values:
    - Checked all datasets for missing values.
    - Removed or imputed rows where necessary.
2. Data Type Conversion:
    - Date of Travel in Cab_Data.csv was converted to datetime format.
    - Population, Price Charged, and Income columns were converted to numeric types.
3. Standardization:
    - Standardized city names across all datasets (removed spaces, unified case).
    - Cleaned numeric columns by removing commas and invalid values.
4. Duplicate Removal:
    - Identified and removed duplicate rows in all datasets.
5. Feature Engineering:
    - Added a Margin column as the difference between Price Charged and Cost of Trip.
    - Extracted Month from the Date of Travel to analyze seasonal trends.

Data Transformations

The following transformations were applied to prepare the final master dataset:

1. Merging Datasets:
    - Cab_Data was merged with Transaction_ID using Transaction ID.
    - Added customer demographics by merging Customer_ID on Customer ID.
    - Added city population and cab users by merging City.csv on City.
2. Final dataset includes the following fields:
    - Transaction ID, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip, Margin, Customer Demographics (Age, Gender, Income), City Population, Users.
3. Validation:
    - Ensured that all merged columns align correctly and contain valid data.

Challenges and Solutions

Challenges	Solutions Implemented
Missing or invalid values in numeric fields	Converted data to numeric, handled missing values using imputation or removal.
Data type inconsistencies	Converted date columns and numeric fields to appropriate formats.
Duplicate rows across datasets	Removed duplicates to ensure clean data.
Merging datasets with non-uniform keys	Standardized city names and unified column naming conventions.

Final Prepared Dataset

Column Name	Description
`Transaction ID`	Unique ID for each cab transaction.
`Date of Travel`	Travel date for the cab ride.
`Company`	Cab company name (e.g., Pink Cab).
`City`	City where the cab ride occurred.
`KM Travelled`	Distance travelled during the ride.
`Price Charged`	Revenue collected for the ride.
`Cost of Trip`	Operational cost of the trip.
`Margin`	Difference between Price Charged and Cost of Trip.
`Customer ID`	Unique identifier for the customer.
`Gender`	Gender of the customer.
`Age`	Age of the customer.
`Income (USD/Month)`	Monthly income of the customer.
`Population`	Population of the city.
`Users`	Number of cab users in the city.