Data Intake Report

Name: XYZ Cab Industry analysis

Report date: October 14, 2024

Internship Batch: LISUM38

Version: 1.0

Data Intake by: Kirtoria Ward

Data Intake Reviewer: < >

Data Storage Location: Google Drive

Tabular Data Details:

File Name	Total Number of Observations	Total Number of Features	Base Format	Size
Cab_Data.csv	359,392	7	.csv	20.7 MB
Customer_ID.csv	49,171	4	.csv	1.0 MB
Transaction_ID.csv	440,098	3	.csv	8.8 MB
City.csv	20	3	.csv	1 KB

Introduction:

This report provides an overview of the data intake process, data cleaning steps, and key insights gained from the initial exploration of the datasets provided for the XYZ Cab Industry analysis. The goal is to help XYZ make an informed investment decision by identifying key trends in cab usage, fares, and customer preferences.

Overview of Data:

Four datasets were provided for analysis, covering cab transactions, customer demographics, payment modes, and city information from 2016 to 2018:

1. Cab_Data.csv: Contains transaction details such as date of travel, company name (Yellow or Pink Cab), city, kilometers traveled, price charged, and cost of the trip.
2. Customer_ID.csv: A mapping table that links customer demographics using a unique Customer ID.
3. Transaction_ID.csv: A mapping table linking transactions to customers, including the payment method used.
4. City.csv: Lists U.S. cities, their population, and the number of cab users.

Data Cleaning Steps:

Several steps were taken to ensure the datasets were ready for analysis:

1. Removing Duplicates: Duplicates were removed from cab_data, customer_data, and transaction_data to ensure clean non-repetitive data.
2. Handling Missing Values:
   - In cab_data, any missing numeric values (such as price charged) were filled with 0.
   - In customer_data, any missing categorical values were filled with 'Unknown' to maintain consistency.

No additional issues with data quality, such as inconsistent data types, were identified.

Key insights from Initial Exploration:

Several important insights were identified during the initial exploratory data analysis:

1. Cab Company Usage: Yellow Cab had significantly more rides than Pink Cab, suggesting it holds a larger market share.
2. Average Fares by City: Certain cities, such as New York, exhibited higher average fares compared to smaller cities like Nashville, indicating regional price variations.

3. Payment Preferences: The majority of customers preferred using card payments over cash, which suggests a stronger reliance on digital payment infrastructure.

Challenges and Assumptions:

1. Challenges:

- One of the main challenges was dealing with missing information in the data. For example, some numbers were missing, and in those cases, they were replaced with 0 to avoid leaving gaps

2. Assumptions:

- It was assumed that when the price of a ride was missing in the data, it meant that the information wasn't recorded properly. Setting these prices to 0 helped to ensure that the analysis wasn't affected too much.

- The number of cab users in each city was taken from the City.csv file, and it was assumed that this data accurately represents the number of users in each location.

Conclusion:

The data has been successfully cleaned and prepared for analysis, with key insights already emerging from the initial exploration. A presentation will involve deeper analysis to identify which cab company is a better investment for XYZ based on cab usage trends, regional performance, and customer preferences.