Data Intake Report for XYZ Cab Industry Analysis

Nishi Gandhi

October 5, 2024

1 Introduction

This report provides an overview of the datasets used for the analysis of the cab industry, focusing on usage patterns, fare structures, and customer demographics. The objective is to derive actionable insights that can guide strategic decisions for investments in the cab sector.

2 Datasets Overview

2.1 Cab Data

The Cab Data dataset contains transaction details for two cab companies. The key columns are shown in Table 1.

Table 1: Cab Data

Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged
10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95
10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52
10000013	42371	Pink Cab	ATLANTA GA	9.04	125.20
10000014	42376	Pink Cab	ATLANTA GA	33.17	377.40
10000015	42372	Pink Cab	ATLANTA GA	8.73	114.62

2.2 City Data

The City Data dataset provides information about US cities, including their population and number of cab users, as shown in Table 2.

Table 2: City Data

City	Population	Users
NEW YORK NY	8,405,837	302,149
CHICAGO IL	1,955,130	164,468
LOS ANGELES CA	1,595,037	144,132
MIAMI FL	1,339,155	17,675
SILICON VALLEY	1,177,609	27,247

2.3 Customer ID Data

The Customer ID dataset links customer demographic details to unique identifiers. Table 3 summarizes its key columns.

Table 3: Customer ID Data

Customer ID	Gender	Age	Income (USD/Month)
29290	Male	28	10,813
27703	Male	27	9,237
28712	Male	53	11,242
28020	Male	23	23,327
27182	Male	33	8,536

2.4 Transaction ID Data

The Transaction ID dataset links transactions to customers and payment modes. Table 4 summarizes its key columns.

Table 4: Transaction ID Data

Transaction ID	Customer ID	Payment Mode
10000011	29290	Card
10000012	27703	Card
10000013	28712	Cash
10000014	28020	Cash
10000015	27182	Card

3 Data Exploration and Cleaning

During the exploratory data analysis (EDA), we checked for missing values, data type discrepancies, and duplicates across all datasets. Data types were adjusted as necessary to ensure consistency and facilitate analysis.

4 Key Insights

The following insights were derived from the EDA:

- Trends in customer usage indicate peak periods during weekends and holidays.
- Average fares differ significantly between the two cab companies.
- Customer demographics, particularly age and income, influence cab usage patterns.

5 Hypotheses

The following hypotheses were generated during the analysis:

- There is a seasonality in the number of customers using the cab service.
- Company A has a higher average fare per ride compared to Company B.
- Customers using Company A are more likely to pay with credit cards than those using Company B.
- The number of rides increases during weekends compared to weekdays.

- Larger cities have more cab users than smaller cities.
- Customer demographics (like age or income) influence the choice of cab company.

6 Conclusion

The analysis highlights critical patterns in cab usage and fare structures, suggesting strategic opportunities for investments in the cab industry. Understanding customer demographics and their impact on usage will be essential for tailoring services and maximizing revenue.

7 References

Used Latex for creating report and .ipynb file is pushed to github.
Company A= Yellow Cab
Company B = Pink Cab