

Exploratory Data Analysis (EDA) Report

Project Name: G2M Insight for Cab Investment Firm

Report Date: 03/14/25

Internship Batch: LISUM43

Version: 1.0

Data Intake By: Manil Shangle

Data Intake Reviewer:

Data Storage Location:

1. Overview of Provided Datasets

The dataset consists of four files covering transactions, customer demographics, city details, and payment modes. The time range of the data is **January 31, 2016 - December 31, 2018**.

Summary of Datasets:

Dataset Name	Observations	Features	File Type	Size
Cab_Data.csv	359,392	7	.csv	20.6 MB
Customer_ID.csv	49,171	4	.csv	1.0 MB
Transaction_ID.csv	440,098	3	.csv	8.7 MB
City.csv	20	3	.csv	1 KB

2. Data Cleaning & Preprocessing

Steps Taken:

- ✓ **Checked for missing values** and handled them appropriately.
- ✓ **Removed duplicate rows** from each dataset.
- ✓ **Standardized column names** (converted to lowercase and replaced spaces with underscores).
- ✓ **Converted date columns** to proper datetime format.
- ✓ **Merged datasets** using key identifiers:
 - Cab_Data merged with Transaction_ID (on Transaction ID)
 - Transaction_ID merged with Customer_ID (on Customer ID)

- Cab_Data merged with City (on City)
 - Profit = Price Charged - Cost of Trip
 - Price per KM = Price Charged / KM Travelled
 - Profit per KM = Profit / KM Travelled
 - Number of Rides per Customer
-

3. Hypothesis Testing & Key Findings

Hypothesis 1: Is there seasonality in cab usage?



Finding: The data shows **seasonal trends** in cab usage, with higher demand during holiday periods and summer months.

Hypothesis 2: Which company has more users?



Finding: Company X has **more unique users**, but Company Y shows **higher retention rates**.

Hypothesis 3: Does more customers mean more profit?



Finding: More customers do not always mean higher profits. Instead, profit depends on trip distance and pricing strategy.

Hypothesis 4: What payment mode is most popular?



Finding: Credit/Debit card payments dominate, with cash usage being significantly lower.

Hypothesis 5: Does population affect cab usage?



Finding: There is a **strong correlation** between city population and cab usage, with larger cities generating higher ride demand.

4. Business Insights & Recommendations



Key Insights:

- Company X has a **wider customer base**, while Company Y has a **better profit margin per ride**.
- Seasonal trends suggest **higher revenue opportunities in summer and holiday seasons**.

- The shift to **cashless payments** suggests investments in **digital payment systems** will benefit cab companies.
- **Cities with larger populations** show higher demand, making urban areas better for investment.

Recommendation:

Based on profitability, retention, and market trends, **Company Y is the better investment option for XYZ**, as it ensures **sustained profit growth despite fewer users**.

5. Conclusion

The analysis reveals key factors affecting the cab market, including seasonality, customer behavior, and city demographics. The findings provide a **data-driven approach** for XYZ's investment decision in the cab industry.