

# Data Intake Report

Name: Decoding the US Cab Industry: Data-Driven Investment Intelligence

Report date: 11-09-2023

Internship Batch: LISUM 25

Version: 1.0

Data intake by: Hande Gul Atasagun

Data intake reviewer: -

Data storage location: [https://github.com/hgatasagun/DataGlacier\\_HandeGulAtasagun](https://github.com/hgatasagun/DataGlacier_HandeGulAtasagun)

## Tabular data details: Cab\_Data

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	21.2MB

Transaction ID	int64
Date of Travel	int64
Company	object
City	object
KM Travelled	float64
Price Charged	float64
Cost of Trip	float64

## Tabular data details: Customer\_ID

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1.1MB

Customer ID	int64
Gender	object
Age	int64
Income (USD/Month)	int64

## Tabular data details: Transaction\_ID

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	9MB

Transaction ID	int64
Customer ID	int64
Payment_Mode	object

## Tabular data details: City

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	759 bytes

City	object
Population	float64
Users	float64

#### Tabular data details: city\_coordinates

Total number of observations	19
Total number of files	1
Total number of features	2
Base format of the file	csv
Size of the data	666 bytes

City	object
Coordinates	object

#### Proposed Approach:

- A **'data frame inspection function'** has been written to examine the data type, shape, quantiles, and the first 5 rows of the columns in the datasets.
- To determine the presence of duplicates in the data, initially, the uniqueness of 'Transaction ID' in the "Cab\_Data" dataset was examined to check whether each row had unique values. Considering that the same customer could make multiple transactions, the count of unique values based on the 'Customer ID' was not checked. The uniqueness of 'Customer ID' in the "Customer\_ID" dataset and 'Transaction ID' in the "Transaction\_ID" dataset was also reviewed separately. Additionally, a **'duplicate detection function'** was implemented to examine whether there were any duplicate rows. The analysis results showed that there were no repeated data entries in the datasets.
- The erroneous data types in the datasets have been corrected. Additionally, a **'dataframe column types detection function'** has been implemented to facilitate data manipulation. This function accurately identifies columns that were treated as numerical but are categorical and columns that, despite being categorical, exhibit cardinality characteristics.
- It has been determined that there are no missing values in the datasets. Furthermore, outliers in the datasets have been identified using various functions, and necessary corrections have been made.