Data Intake Report

Name: Cab Industry Investment Analyst
Report date: 17th June 2024
Internship Batch: LISUM34
Version: 1.0
Data intake by: Fenil Mavani
Data intake reviewer:
Data storage location:

Tabular data details:

Cab_Data.csv

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

City.csv

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	4KB

Customer_ID.csv

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

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

Proposed Approach:

- Identify duplicates: We will use the duplicated() method in pandas to identify any duplicate rows in each dataset based on unique identifiers such as 'Transaction ID', 'Customer ID', and other relevant columns.
- Remove duplicates: Once identified, duplicates will be removed using the drop_duplicates() method to ensure data quality and accuracy.

Assumptions:

- Unique Identifiers: It is assumed that 'Transaction ID' in the Transaction_ID.csv and Cab_Data.csv files are unique identifiers for each transaction.
- Date Formatting: Dates in the Cab_Data.csv are in a consistent format and correctly represent the period from 31/01/2016 to 31/12/2018.
- No Missing Critical Data: It is assumed that there are no critical missing data in the key columns such as 'Customer ID', 'Transaction ID', and 'City'.
- Consistent Data Types: All numeric fields such as 'Price Charged', 'Cost of Trip', 'Income (USD/Month)', etc., are in a consistent format and do not contain any non-numeric values.