Data Intake Report

Name: Cab Industry Analysis
Report date: 05/12/2024
Internship Batch: LISUM33
Version: 1.0
Data intake by: Manhui Zhu
Data Intake reviewer:
Data Storage location: https://github.com/DataGlacier/DataSets.git

Tabular data details:

Name of data	Transaction_ID.csv
Total number of observations	440098
Total number of features	3
Base format of the file	.csv
Size of the data	10.1+ MB

Name of data	Customer_ID.csv
Total number of observations	49171
Total number of features	4
Base format of the file	.csv
Size of the data	1.5 + MB

Name of data	City.csv
Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	608.0 + bytes

Name of data	Cab_Data.csv
Total number of observations	359392
Total number of features	7
Base format of the file	.csv
Size of the data	19.2 + MB

Name of data	full.csv
Total number of observations	359392
Total number of features	21
Base format of the file	.csv
Size of the data	50.0 + MB

Proposed Approach:

- For Transaction_ID data, I check number of unique values in `Transaction ID` column, it is the same as total number of observations, which means each row with different Transaction ID value is the unique row.
- For Customer _ID data, I check number of unique values in `Customer ID` column, it is the same as total number of observations, which means each row with different Customer ID value is a unique row.
- For City data, I check number of unique values in `City` column, it is the same as total number of observations, which means each row with different City is a unique row.
- For Cab data, I check number of unique values in `Transaction ID ` column, it is the same as total number of observations, which means each row with different Transaction ID value is a unique row.
- The full dataset combines the above four datasets but only keeps the transaction that included in Cab_data.csv.
- The full dataset is consisting of 21 features, including 7 derived features. They are: `Profit_of_trip`, `Price_per_KM`, `Cost_per_km`, `Year`, `Month`, `Age_group`, `income_groups`.

Assignment:

XYZ is a private equity firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry. Our objective is to do exploratory data analysis and provide actionable insights to help XYZ firm understand the cab industry, assisting decision – making of choose more appropriate company to make investment.