Data Intake Report

Name**:** G2M insight for Cab Investment firm
Report date: 04/09/2021
Internship Batch: LISUM02
Version: 1.0
Data intake by: Memudu Alimatou Sadia
Data intake reviewer: N/A
Data storage location: https://github.com/memudualimatou/CAB-INVESTMENT-EDA

Data Preparation:

Tabular data details: Total number of observations	359392
Total number of files	4
Total number of features	19
Base format of the file	.csv
Data name	Global_cab_data
Size of the data	46.1 MB

Proposed Approach:

- Global data formation

A global dataset was built by combining 3 different datasets which are Cab_Data.csv, Transaction_ID.csv and Customer_ID.csv and performing inner join with the column they have in common. The city.csv was not included due to its values mostly related to the cities and uncommon data related to others datasets.
The Transaction_ID.csv and Cab_data.csv was inner join through the Transaction ID column present in both datasets was formed the Trans_cab.csv dataset. Then, this new dataset (Trans_cab.csv) was inner-joined with the Customer_ID.csv through a common Customer ID column to form our global_cab_data.csv dataset which details are listed in the table above.

- Assumptions for data quality improvement

1. 5th January 2018 is the day with the highest number of trips from both companies due to unknown reason.
2. The price charged is related to the KM travelled
3. The cost of price depend on the price of fuel
4. The yellow cab is user's preferred company for undiscovered reasons, nothing justify this likeness from the datasets.