Data Intake Report

Name: <G2M insight for Cab Investment firm>
Report date: <10/11/2022>
Internship Batch:<LISUM14>
Version:<1.0>
Data intake by:<Zhuoming Yang>
Data intake reviewer:<>
Data storage location: <GitHub >

Tabular data details:

Total number of observations	<358,392 of rows>
Total number of files	<5>
Total number of features	<60>
Base format of the file	<.csv >
Size of the data	<102.7MB>

Proposed Approach:

- **Data: Merged_data.csv, Cab_Data.csv, City.csv, Customer_ID.csv, Transaction_ID.csv, StormEvent_2016, StormEvent_2017, StormEvent_2018:**
- **Cab_Data.csv** – this file includes details of transaction for 2 cab companies.
- **Customer_ID.csv** – this is a mapping table that contains a unique identifier which links the customer's demographic details.
- **Transaction_ID.csv** – this is a mapping table that contains transaction to customer mapping and payment mode.
- **City.csv** – this file contains list of US cities, their population and number of cab users.
- **StormEvent_2016_2017_2019** – this file contains records of yearly storm information US
    - Customer_data['customer_id'] = transaction_data['customer_id']
    - Transaction_data['transactoin_id']= cab_data['transaction_id']
    - Cab_data['city'] = city_data['city']
    - Join every data together by their foreign key -> Merged_data
    - **Important Feature:**
        - **Location: State, City, Population**
        - **Customer information: Income(usd/month), gender**
        - **Trip information: KM_travelled, Price_Charged, Profit, Loss**
- Assumptions:
    - Null Values
    - Duplicates
    - Date Format need to be converted into a better format for analyzing
    - Data Types need to be change