Data Intake Report

Name: <Pue Leu Nae Park>
Report date: <08.MAR.2021>
Internship Batch:<LIAP01>
Version:<1.0>
Data intake by:<Pue Leu Nae Park>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/reneeparkkr/DataGlacier.git>

Tabular data details:

- Cab_Data

Total number of observations	<359392>
Total number of files	<4>
Total number of features	<7>
Base format of the file	<.csv>
Size of the data	<20.2MB>

- City

Total number of observations	<20>
Total number of files	<4>
Total number of features	<3>
Base format of the file	<.csv>
Size of the data	<759B>

- 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	<1MB>

- 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	<8.6MB>

- WeatherEvents_Jan2016-Dec2020

Total number of observations	<6274506 >
Total number of files	<1>
Total number of features	<13>
Base format of the file	<.csv>
Size of the data	<762MB>

-advisorsmith_cost_of_living_index

Total number of observations	<510 >
Total number of files	<1>
Total number of features	<3>
Base format of the file	<.csv>
Size of the data	<10KB>

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Users in ‘City.csv’ data assumed as users for XYZ company (only used by pink and yellow taxi)
- Profits are calculated by (Price_Charged - Cost_of_Trip) for each trip
- Income is divided in 3 classes which are “high”, “middle”and “low”.
- Precipitation” mostly consists of rain, snow, hail so the number of Precipitations will be distributed to rain, snow, hail vectors by their ratio.