Data Intake Report

Name: <Exploratory Data Analysis>
Report date: <19/11/22>
Internship Batch:<LISUM15>
Version:<1.0>
Data intake by:<Xiaoyao Yin>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/DataGlacier/DataSets>

Tabular data details:

<Cab_data>

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	<21.2 MB >

<City>

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	<4 KB >

<Customer_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	<1.1 MB >

<Transaction_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	<9 MB >

Proposed Approach:

- **Dedupe is a validation technique used in data science to ensure that data is accurate and consistent. It is often used to clean up data sets before further analysis is conducted. Dedupe can be used to identify and correct errors in data, standardize data formats, or merge data from multiple sources.**
- Assumptions: The data is free of bias. The data is accurate.