

# 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>

<b>Total number of observations</b>	<359392>
<b>Total number of files</b>	<1>
<b>Total number of features</b>	<7>
<b>Base format of the file</b>	<csv>
<b>Size of the data</b>	<21.2 MB >

### <City>

<b>Total number of observations</b>	<20>
<b>Total number of files</b>	<1>
<b>Total number of features</b>	<3>
<b>Base format of the file</b>	<csv>
<b>Size of the data</b>	<4 KB >

### <Customer\_ID>

<b>Total number of observations</b>	<440098>
<b>Total number of files</b>	<1>
<b>Total number of features</b>	<3>
<b>Base format of the file</b>	<csv>
<b>Size of the data</b>	<1.1 MB >

### <Transaction\_ID>

<b>Total number of observations</b>	<49171>
<b>Total number of files</b>	<1>

<b>Total number of features</b>	<4>
<b>Base format of the file</b>	<csv>
<b>Size of the data</b>	<9 MB >

**Proposed Approach:**

- **Dedupe technique is used to ensure that data is accurate and consistent. It is used to clean up data sets before further analysis is conducted.**
- Assumptions: The data is free of bias. The data is accurate.