Data Intake Report

Name: <G2M Insight for Cab Investment Firm – Week 2>
Report date: <August 14th 2022>
Internship Batch:<LISUM12>
Version:<1.0>
Data intake by:<Richard Coltenback>
Data intake reviewer:<None>
Data storage location: https://github.com/RColtenback/Cab-Investment-Week-2

Tabular data details:

Cab_Data.csv	
Total number of observations	<359393>
Total number of files	<1 file>
Total number of features	<7>
Base format of the file	<.csv >
Size of the data	<20.1 MB>
City.csv	
Total number of observations	<21>
Total number of files	<1 file>
Total number of features	<3>
Base format of the file	<.csv >
Size of the data	<4.00 KB>
Customer_ID.csv	
Total number of observations	<49172>
Total number of files	<1 file>
Total number of features	<4>
Base format of the file	<.csv >
Size of the data	<1.00 MB>
Transaction_ID.csv	
Total number of observations	<440098>
Total number of files	<1 file>
Total number of features	<3>
Base format of the file	<.csv >
Size of the data	<8.58 MB>

Proposed Approach:

- When combining the datasets in Jupyter Notebook, the duplicates were combined. I checked by going into excel and making sure each data entry had only one count for each entry.
- There are some transactions that are specifically for gender and some that are specifically for distance and cost. These don’t change the way we look at the data, just something of note.