Data Intake Report

Name: <Cab EDA Notebook>
Report date: <11/13/2022>
Internship Batch:<LISUM215>
Version:<1.0>
Data intake by:<Abilesh Ayyanar>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/abileshayyanar/CabEDA>

Tabular data details:

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	<20663 KB>

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	<1 KB>

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	<8788 KB>

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	<1027 KB>

Dedup validation: Transaction ID, Customer ID, and City information are all found on multiple files, so one instance of each was removed.

Assumptions: Using a boxplot, I saw that there were not major outliers in any of the data sets, so I assumed that it would be accurate to directly compare the average values for things like total cost and distance traveled.