Data Intake Report

Name: <EDA G2M>
Report date: <>
Internship Batch:<LISUM19>
Version:<1.0>
Data intake by:<Connor Walker>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <github>

Tabular data details:

Total number of observations	<440098>
Total number of files	<4>
Total number of features	<14>
Base format of the file	<.csv>
Size of the data	<20.2MB>

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

Proposed Approach:

- To validate the deduplication process, we will compare the unique identifiers across all four data files and identify any discrepancies or inconsistencies.
- Our assumptions for the data quality analysis include assuming that missing values are null values, and that any extreme outliers outside of the 99th percentile are valid data points unless there is evidence to suggest otherwise.