Data Intake Report

Name: G2M insights for Cab Investment firm
Report date: 13/03/2023
Internship Batch: LISUM19
Version:<1.0>
Data intake by: Salman Ali Sayyed
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/SalmanSD07/DataGlacier_G2M.git

Tabular data details:

Total number of observations	359392
Total number of files	4
Total number of features	14
Base format of the file	.csv
Size of the data	41.1+ MB

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

Proposed Approach:

- Mention approach of dedup validation (identification)
  For analysing the dataset I had started by reading the csv files and drop the duplicates based on their primary key provided. Then I had visualized the trends on yellow and pink cab across years based on the amount of profit they make per ride, percentage of users using yellow and pink cab services in different cities. Then I had merged the given files and and created a dataframe master_data which is further used for visualization.
  Trends of yellow and Pink Cab across different gender had also been visualized with the help of stacked bar chart.
- Mention your assumptions (if you assume any other thing for data quality analysis)
  For the descriptive statistics using the chisqaure test I had assumed that Pink Cab is more popular among user in different cities which was later rejected based on the hypothesis result.s

Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.
Please do not forget to remove this section while converting the file into pdf.