Data Intake Report

Name: <G2M insights for Cab investment>
Report date: <4/18/2022>
Internship Batch:<LISMU 08>
Version:<1.0>
Data intake by:<Uchenna Ilodigwe>
Data intake reviewer:None
Data storage location: https://github.com/uchennailodigwe/VCWeek2
Tabular data details:

	Cab_Data.csv	Transaction_ID.csv	City.csv	Customer_ID
Observations	359392	440098	19	49171
Total number of files	1	1	1	1
Total number of features	7	3	3	4
Base format of the file	.csv	.csv	.csv	.csv
Size of the data	20663kb	8788kb	1kb	1027kb

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

Proposed Approach:

- Mention approach of dedup validation (identification)
  - Merge Files to get the necessary features for the analysis.
  - 17 Features(including 5 derived features)
  - Timeframe of the data: 2016-01-31 to 2018-12-31
  - Total data points : 359392

- Mention your assumptions (if you assume any other thing for data quality analysis)
  - Outliers are present in Price_Charged feature but due to unavailability of trip duration details ,we are not treating this as outlier.
  - Profit/loss of rides are calculated keeping other factors constant and only Price_Charged and Cost_of_Trip features used to calculate profit/loss margin.
  - Users feature of city dataset is treated as number of cab users in the city.
  - we have assumed that this can be other cab users as well(including Yellow and
  - Pink cab)