Data Intake Report

Name: G2M Cab Case Study
Report date: 8/14/2022
Internship Batch: LISUM12
Version:1.0
Data intake by: Dima Musa
Data intake reviewer:
Data storage location: https://github.com/dowoma/DataSets

Tabular data details (Cab_Data):

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	20663 KB

Tabular data details (City):

Total number of observations	19
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	1 KB

Tabular data details (Customer_ID):

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1027 KB

Tabular data details (Transaction_ID):

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	8788 KB

Proposed Approach:

- Mention approach of dedup validation (identification)
- Mention your assumptions (if you assume any other thing for data quality analysis)

- Use .duplicated to check for duplicates in each .csv and in master data set.
- Checked for null values.
- Merged on Transaction_ID and Customer_ID and used .groupby and .apply to merge changing columns and keep same 'Customer ID', 'Gender', 'Age', 'Income', 'Age Group', 'Income Group'

- Checked distribution of continuous values
- Calculated profit as Price_Charged – Cost_of_Trip; profit has some outliers.
- Assumed Silicon Valley refers to combined cities of Campbell, Cupertino, Gilroy, Los Altos, Los Gatos, Milpitas, Morgan Hill, Mountain View, Palo Alto, San Jose, Santa Clara, Saratoga and Sunnyvale of CA
- Assumed Orange County refers to combined cities of Anaheim, Santa Ana, Irvine, and other smaller cities of CA.
- Assumed users refers to cab users of city including both companies