Data Intake Report

Name: G2M: Insight for Cab Investment Firm
Report date: 07/14/2023
Internship Batch: LIMSUM23: 30
Version: 1.0
Data intake by: Susan Zhang
Data intake reviewer: N/A
Data storage location: https://github.com/20szha/VC/blob/new-branch/Week%202/G2M_Insight_for_Cab_Investment_Firm.ipynb
Tabular data details:

Customer_ID.csv

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1,027 KB

Cab_Data.csv

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,663 KB

City_Data.csv

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

Transaction_ID.csv

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8,788 KB

After merging into Master_Data.csv

Total number of observations	359,392
Total number of files	1 (merged from the top 4 files)

Total number of features	14
Base format of the file	.csv
Size of the data	47,439 KB

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

Proposed Approach:

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

The following steps were taken to explore the data:

1. Read in all given .csv files
2. Cleaned up data if necessary
    a. Converted excel time serial number into dates
3. Merge data into single master dataframe using the .merge() function
    a. Merged on “City”, “Transaction ID”, and “Customer ID” to avoid duplicate records
    b. Filled NA’s with ‘’
        i. Original dataset did not have any anyways and those with no City, Customer ID, or Transaction ID matches should’ve been discarded
4. Explore the data and evaluate for better company to invest in
    a. For example: sns.pairplot() function and .corr() function to visualize the relationship between each feature and the .groupby() function to compare the various features against one another

Conclusion: It appears that the Yellow Cab would be a better investment.