Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 11/12/2022
Internship Batch: LISUM15
Version: <1.0>
Data intake by: Neenu Santhosh
Data intake reviewer: <intern who reviewed the report>
Data storage location: <location URL e.g.: GitHub, cloud>

Tabular data details:

Transaction Id Data

Total number of observations	440098
Total number of files	<Number of files received>
Total number of features	3
Base format of the file	csv
Size of the data	8,788 KB

Customer Id Data

Total number of observations	49171
Total number of files	<Number of files received>
Total number of features	4
Base format of the file	csv
Size of the data	1,027 KB

City Data

Total number of observations	20
Total number of files	<Number of files received>
Total number of features	3
Base format of the file	csv
Size of the data	1 KB

Cab Data

Total number of observations	359392
Total number of files	<Number of files received>
Total number of features	7
Base format of the file	csv
Size of the data	20,663 KB

Proposed Approach:

1. I Imported all the data sets using pandas.
2. I did sanity checks like null values check, duplicate value check, and data type check.
3. There were no null values and duplicates. However, the datatype had some issues, hence changed them using the 'as type and date time' function.
4. I used the Left join and merge function to create a final data set
5. Once the data was clean, I did the univariate and bivariate analysis using the boxplot, histogram, and correlation heatmap
6. I developed my hypothesis and proved it true/false based on my analysis

Assumption

The price charged is the amount charged to the customer
The cost of the trip is the total expense of the trip for the company