Data Intake Report

Name**:** G2M insight for Cab Investment firm
Report date: 20 June 2021
Internship Batch: LISUM01
Version:<1.0>
Data intake by: Payal Upadhyay
Data intake reviewer:<Payal Upadhyay>
Data storage location: < https://github.com/payal-upadhyay/EDA-Notebook/upload/JupyterNotebook>

Tabular data details:

Cab_Data:

Total number of observations	1066
Total number of files	1
Total number of features	8
Base format of the file	.csv
Size of the data	75.1+ KB

City:

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	608.0+ bytes

Customer_ID:

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.5+MB

Transaction_ID:

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	10.1+MB

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

Proposed Approach:

- We are analyzing data sets using graphs to see patterns, trends and anomalies as well as testing hypothesis there we are conducting an Exploratory Data Analysis

Before importing on Jupyter:

- The given files contained data outside the time frame of 31-01-2016 to 31-12-2018 so the data was filtered to only contain the data of the given time frame
- We added an extra column in Cab_Data.csv called Profit which is Price Charged – Cost of trip
- We changed the date format to dd-mm-yy in Cab_Data.csv
- The data intake report was taken after the above changes

Mention approach of dedup validation (identification)

- All the datasets were joined before the data was visualized
- After joining the data we made sure that there are no duplicate values in the table
- as well as no null values by using a function in JupyterNotebook.
- No null values or duplicate values were returned

Mention your assumptions (if you assume any other thing for data quality analysis)

- We encountered a few negative values for profit which were kept assuming passengers earned discounts on trips.

- In order to conduct hypothesis testing using t-test we assumed equal sample variances