Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 08 July 2021
Internship Batch: LISUM01
Version: 1.0
Data intake by: Vatsal Vinesh Mandalia
Data intake reviewer: None
Data storage location: https://github.com/Vatsal-2414/Vatsal-Mandalia-Data-Glacier

Tabular data details:

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

File name: Cab_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	759 B

File name: City.csv

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	1.1 MB

File name: Customer_ID.csv

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	9 MB

File name: Transaction_ID.csv

Total number of observations	342
Total number of files	1
Total number of features	6
Base format of the file	.csv
Size of the data	16 KB

File name: US Holiday Dates (2004-2021).csv

Proposed Approach:

- Mention approach of dedup validation (identification)
  - The values in the ‘Date of Travel’ column in the Cab data were in a 5-digit Excel Serial format. Converted this into datetime format (%Y-%m-%d) and stored in a new column named ‘Travel_Date’.
  - No duplicate rows found using features ‘Transaction ID’ and ‘Travel_Date’.
  - The Cab_data and the city data were merged on ‘City’ column through inner join. There was no cab ride data for the city of ‘SAN FRANCISCO CA’ in the merged dataframe. ‘Transaction ID’ used to search for duplicate rows.
  - Performed inner join on ‘Customer ID’ between Transaction and Customer data. No null values were present in the merged dataframe.
  - The merged dataframes from the above two steps were inner joined on ‘Transaction ID’. 80,706 observations with only details on the transaction and customer demography were lost.
  - Acquired a third-party dataset – US Holiday Dates (2004 - 2021) from Kaggle. Merged this at the end. Using the start and end dates of the cab ride data, the US Holiday dataset was subsetted.
- Mention your assumptions (if you assume any other thing for data quality analysis)