Data Intake Report

Name: **G2M insight for Cab Investment firm**
Report date: **7 March 2025**
Internship Batch: **LISUM43**
Version: **1.0**
Data intake by: **Asha K C**
Data intake reviewer: **-None-**
Data storage location: **https://github.com/Asha-KC-07/Cabs-DataSets**

Tabular data details:

Cab_Data.csv

Total number of observations	# of rows: 359392
Total number of files	# of files: 1
Total number of features	# of columns: 7
Base format of the file	Comma separated (.csv)
Size of the data	20.1 MB

City.csv

Total number of observations	# of rows: 20
Total number of files	# of files: 1
Total number of features	# of columns: 3
Base format of the file	Comma separated (.csv)
Size of the data	1 KB

Customer_ID.csv

Total number of observations	# of rows: 49171
Total number of files	# of files: 1
Total number of features	# of columns: 4
Base format of the file	Comma separated (.csv)
Size of the data	1 MB

Transaction_ID.csv

Total number of observations	# of rows: 440098
Total number of files	# of files: 1
Total number of features	# of columns: 3
Base format of the file	Comma separated (.csv)
Size of the data	8.58 MB

Proposed Approach:

- **Step 1:** Initial data Insights - Load all the data sets, identify available columns along with their data types.
- **Step 2**: Identify potential analysis that can be performed - Profit analysis, customer age vs gender vs cab rides, City based cab demands, Pink cab vs Yellow cab profit comparison.
- **Step 3**: Clean data sets – remove duplicates if any, convert travel date in Cab_Data.csv to proper format, convert 'Population' and 'Users' in City.csv to numeric values, join datasets to have master data.
- **Step 4**: Identify the additional data sets required: US holiday data (considering major holidays New Year, Independence Day, Thanksgiving, Christmas), US weather data, other cab industry data
- **Step 5**: Perform analysis and come up with final recommendations – plot the analysis from merged data. Provide insights from derived graphs/comparison reports.

Reports analyzed are:

1. **Seasonal trends in Cab usage –** Line plot showing monthly cab usage
2. **Cab demand around major US holidays –** Bar plot showing number of rides on US holidays
3. **Monthly cab usage vs Monthly revenue trend –** Line graphs showing comparison between the number of rides per month and corresponding monthly revenue of the month
4. **Pink cab & Yellow cab Seasonal trends vs Profitability –** Line plot of number of rides per company and a line plot showing monthly revenue per company
5. **Profit margin per company –** Bar plot showing average profit margin for Pink cab and yellow cab
6. **Profit by age, Ride count by income –** a line plot showing average profit per ride vs customer's age and a scatter plot showing ride count of customer based on their income
7. **City-wise profit –** bar plot showing profit of cabs per city

Python libraries used for visualization: **Pandas, Matplotlib, Seaborn**