Data Intake Report

Name: G2M Insight for Cab Investment firm

Report date: 10 June 2022

Internship Batch: LISUM10: 30

Version: 1.0

Data intake by: Ray Ng

Data intake reviewer:

Data storage location: https://github.com/DataGlacier/DataSets

**Tabular data details:**

Cab\_Data.csv

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** |  |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 20.2 MB |

City.csv

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** |  |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 B |

Customer\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** |  |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1 MB |

Transaction\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** |  |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 20.2 MB |

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

**Proposed Approach:**

* Load all data as Pandas dataframe
* This data assumes there are no duplicates in Transaction ID (Cab\_Data and Transaction\_ID), and in Customer ID (Customer\_Data), City (City), so index these columns when loading each dataframe
* Cab\_ID and Transaction\_ID merged into one Pandas dataframe using the join function
* Added additional columns to Cab\_Data: Profit Generated, Trips, Yellow Cab Trips, Pink Cab Trips, Share of Yellow Cab Trips, Share of Pink Cab Trips – to make summaries easier to access
* Added additional columns to Customer\_ID: Age Group, Income Bracket – for making a breakdown of users by demographic characteristic
* Converted some columns to readable/processable data
* Used the groupby function to show summaries of data, sometimes merging with other frames, such as breakdown of users by city, and the City dataframe
* Used Pandas to make several plots including scatterplot, stacked column, and time series

**Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.**

**Please do not forget to remove this section while converting the file into pdf.**