Data Intake Report

Name: G2M Insight for Cab Investment Firm

Report date: Wednesday February 17, 2021

Internship Batch:<Enter your batch code from Canvas course>

Version:<1.0>

Data intake by: Laith Adi

Data intake reviewer:<intern who reviewed the report>

Data storage location: <location URL eg: github, cloud>

**Tabular data details**

Note: The information on the tables are on the original datasets. Before any data processing (joining of tables, filtering of attributes/features, etc.) was done.

Table: **Cab\_Data.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 359,392 |
| **Total number of files** | 1 |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 19.2+ MB |

Table: **City.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** | 608.0+ bytes |

Table: **Customer\_ID.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 49,171 |
| **Total number of files** | 1 |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.5+ MB |

Table: **Transaction\_ID.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 440,098 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 10.1+ MB |

Table: **Income.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 39,030 |
| **Total number of files** | 1 |
| **Total number of features** | 80 |
| **Base format of the file** | .csv |
| **Size of the data** | 23.8+ MB |

**Attributes for Each Table**

|  |  |
| --- | --- |
| **Table** | **Attributes** |
| Cab\_Data.csv  (cab\_data) | Transaction ID  Date of Travel  Company  City  KM Travelled  Price Charged  Cost of Trip |
| City.csv  (city) | City  Population  Users |
| Customer\_ID.csv  (customer) | Customer ID  Gender  Age  Income (USD/Month) |
| Transaction\_ID.csv  (transaction) | Transaction ID  Customer ID  Payment\_Mode |
| Income.csv  (income) | State  state\_ab  city  family\_mean  \*there much more but we do not care about the others\* |

**Final Data**

Table: **final\_data**

|  |  |
| --- | --- |
| **Total number of observations** | 359,392 |
| **Total number of files** | -- |
| **Total number of features** | 16 |
| **Base format of the file** | Stored as Pandas Dataframes |
| **Size of the data** | -- |

**Important features**

**Analysis:**

1. Data exploration
   1. Identify the different tables and their attributes
   2. Assumptions on the data
      1. The Income.csv is from 2012-2016 so we will assume that the numbers do not differ as much carrying into 2017 – 2018