Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: <>

Internship Batch: LISUM15

Version: 1.0

Data intake by:

Data intake reviewer:

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

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 848 681 |
| **Total number of files** | 4 |
| **Total number of features** | 17 |
| **Base format of the file** | .csv |
| **Size of the data** | 32.1 MB |

**Cab\_Data**

|  |  |
| --- | --- |
| **Total number of observations** | 359 392 |
| **Total number of files** | 1 |
| **Total number of features** | Transaction ID (int), no missing values  Date of Travel (int), no missing values  Company (str), no missing values  City (str), no missing values  KM Travelled (float), no missing values  Price Charged (float), no missing values  Cost of Trip (float), no missing values  **Total: 7 features** |
| **Base format of the file** | .csv |
| **Size of the data** | 21,2 MB |

1. “Transaction ID” feature is primary key (unique for table). Assuming 1 row means 1 travel (no travels have 2 different Transaction IDs)
2. “Date of Travel” feature have format “days since 1899-12-30” as the latest date is 2018-12-31 and the maximum feature’s value is 43465. The minimum value of this feature in this case will correspond to 2016-01-02
3. “Company” feature – cab company name in text format
4. “City” feature – city name and (in some cases) state code in text format
5. “KM travelled” feature – float value of travel distance rounded to 2 digits after decimal point
6. “Price Charged” feature – the amount in US dollars received for a trip from customer, float with 2 digits after decimal point
7. “Cost of Trip” feature – float with 4 digits after decimal point
8. Assume “Cost of Trip” feature is total cost summing up waiting time fee, cost of trip distance by counter and other costs applied (i.e., this feature is displaying real direct costs of each cab company)
9. Assume all values are correct. Observed differences are due to companies’ price policies and marketing campaigns. No other factors.
10. We have no info about number of passengers in each trip. Assume that each trip was taken by 1 person with corresponding customer ID

**City**

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | 1 |
| **Total number of features** | City (str), no missing values  Population (int), no missing values  Users (int), no missing values  **Total 3 features** |
| **Base format of the file** | .csv |
| **Size of the data** | 0.8 KB |

1. “City” feature is primary key (unique for table). Text, contains city name and (in some cases) state code. Similar to “City” feature in Cab\_Data table
2. “Population” feature is integer with group separator ‘,’ (comma). IMPORTANT: Data contained in table not suite the official US Census Bureau data (see <https://www.moderncities.com/article/2017-jun-top-100-us-cities-ranked-by-2016-population>). For example, population of Miami, FL in 2016 was 453,579 but the value of 1,339,155 provided in file. In reverse, for Boston with official population 673,184 provided value is 248,968). What exactly does “Population” feature mean?
3. “SAN FRANCISCO CA” NOT mentioned in Cab\_Data table. Assume that data in “City” table contains reference information, but not fully corresponds with “City” feature in Cab\_Data table
4. “Users” feature is integer with group separator ‘,’ (comma). Assume this is potential customers of cab companies (estimated number of people that use any cab company services) got from outer source (marketing department, agencies, etc.)

**Customer\_ID**

|  |  |
| --- | --- |
| **Total number of observations** | 49 171 |
| **Total number of files** | 1 |
| **Total number of features** | Customer ID (int), no missing values  Gender (str), no missing values  Age (int), no missing values  Income (USD/month) (int), no missing values  **Total: 4 features** |
| **Base format of the file** | .csv |
| **Size of the data** | 1.1 MB |

1. “Customer ID” feature is primary key (unique for table). Assuming all records means different people (1 person’s info is contained only in 1 record). Therefore, the client bases of the cab companies are not crossing (totally different) (??? Check)
2. “Gender” feature – binary feature “Male”/”Female”
3. “Age” feature – 18-65 years. Assume all customers are valid credit card holders (check income vs age)
4. “Income (USD/month)” feature – compare with data about income in USA

**Transaction\_ID**

|  |  |
| --- | --- |
| **Total number of observations** | 440 098 |
| **Total number of files** | 1 |
| **Total number of features** | Transaction ID (int), no missing data  Customer ID (int), no missing data  Payment\_Mode (str), no missing data  **Total: 3 features** |
| **Base format of the file** | .csv |
| **Size of the data** | 9.0 MB |

1. The table is mapping table between Cab\_Data table and Customer\_ID. Check if it’s fully correspond with records in these tables.
2. Number of observations exceeds number of records in Cab\_Data. So not all the trips info is available
3. “Transaction ID” feature corresponds with “Transaction ID” feature in Cab\_Data table
4. “Customer ID” feature corresponds with “Customer ID” feature in Customer\_ID table
5. “Payment\_Mode” feature – binary feature “Card”/”Cash”

**Proposed Approach:**

* Merge tables with corresponding ID features into new table to make cross-table analysis and checks
* Mention your assumptions (if you assume any other thing for data quality analysis)

**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.**