Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: Apr 26th, 2022

Internship Batch: LISUM08

Version: 1.0

Data intake by: Leonardo Pinheiro de Queiroz

Data intake reviewer: Leonardo Pinheiro de Queiroz

Data storage location: <https://github.com/leopqrz/Data_Glacier_Internship/tree/main/week_02>

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 5 |
| **Total number of features** | 20 |
| **Base format of the file** | .csv |
| **Size of the data** | 31.2 MB |

**Proposed Approach:**

* **cab\_data** dataset: Convert the excel serial date of Date of Travel column into datetime object and add the columns: WeekDay, Month, Day and Year
* **city\_data** dataset: Convert strings to integers on Population and Users columns
* Merge the datasets: **cab\_data**, **customer\_ID**, **transaction\_ID** and **city\_data** into **all\_data** dataset, based on the commom columns Transaction ID, Customer ID and City excluding transactions that are not in all datasets.
* Create the new attribute: **Profit**, as the difference between Price Charged and Cost of Trip attributes.
* Merge the **holidays** dataset into the previous merged dataset based on the Date of Travel and Date columns, excluding eastern holidays and the column Date.
* On holidays column, fill the cells NaN as Not a holiday.
* On the final merged all\_data dataset there's no duplicate rows neither missing data
* Exploratory Data Analysis (EDA):
  + Decriptive Analysis (Univariate Analysis):
    - Numerical Attributes
    - Categorical Attributes
  + Correlational Analysis
    - Categorical vs Categorical (Contingency Table and Chi-Square Test)
    - Numerical vs Numerical (Heatmap, Scatter Plot)
  + Contextual Analysis
    - Time-based Analysis
    - Agent-based Analysis