Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 20-10-2024

Internship Batch: LISUM38

Version:1.0

Data intake by: Vuppu Sree Manasa

Data intake reviewer: N/A

Data storage location: Github

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | **359392** |
| **Total number of files** | 4 |
| **Total number of features** | **17 features and 35 derived features** |
| **Base format of the file** | .csv |
| **Size of the data** |  |

**Proposed Approach:**

* Mention approach of dedup validation (identification)

**Exploratory, Predictive modelling**

* Mention your assumptions (if you assume any other thing for data quality analysis)

***\*\*Assumptions:\*\****

• Since an MLR regression model is attempted to be built, assumptions around No Multicollinearity, Homoscadascity, No pattern in error and Error normally distributed around mean 0 are made.

* Outliers are present in the Price\_Charged feature; however, further study is to be conducted to understand if these are valid outliers.

• Profit from rides is calculated by holding other factors constant, using only the Price\_Charged and Cost\_of\_Trip features for profit calculations.

• The "Users" feature in the city dataset is interpreted as the number of cab users in that city. We assume this includes users of other cab services, such as Yellow and Pink cabs. The following is used to calculate the proportion of users using Users and Population considering other factors as constant.

Based on the availability of cabs, considering Yellow Cab Company is older than the Pink Cab Company (in reality, one must study their start date and analyse their growth graph to truly gauge their numbers and spread.)