Data Intake Report

Name: EDA Cab Industry

Report date: July 2nd, 2024

Internship Batch: LISUM 34

Version:<1.0>

Data intake by: Ha My Pham

Data intake reviewer: Data Glacier

Data storage location: <https://github.com/DataGlacier/DataSets/tree/main>

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 4 |
| **Total number of features** | 12 |
| **Base format of the file** | csv |
| **Size of the data** | ~32MB |

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

**Proposed Approach:**

* Mention approach of dedup validation (identification)
* Mention your assumptions (if you assume any other thing for data quality analysis)

Merging dataset

Check for NA

Drop columns

Encode data

Run a random forest to classify users’ choice of pink or yellow cab.

Run feature importance to find the most important variables in deciding which company to use 🡪 Price is the most important 🡪 Lean over company with cheaper price.