Data Intake Report

Name: Cab Investment

Report date: September 8, 2022

Internship Batch: LISUM13

Version: 1.0

Data intake by: Nurein Umeya

Data intake reviewer:

Data storage location: https://github.com/nureinumeya1999/Cab-Investment-

**Cab\_Data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 1 |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 20663 KB |

**City details:**

|  |  |
| --- | --- |
| **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** | 1 KB |

**Customer\_ID details:**

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | 1 |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1027 KB |

**Transaction\_ID details:**

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 8788 KB |

**Proposed Approach:**

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

**Data Deduplication**

Data deduplication was achieved on each .csv data file independently. For the ‘Cab\_Data’ file, it was assumed the ‘Transaction ID’ label uniquely identifies an observation and assessing duplication on this feature did not reveal duplicates. The ‘City’ file was small enough to assess no duplication by eye. For the ‘Customer\_ID’ file, it was assumed the ‘Customer ID’ label uniquely identifies an observation and assessing duplication on this feature did not reveal duplicates. For the ‘Transaction\_ID’ file, it was assumed the ‘Transaction ID’ label uniquely identifies an observation and assessing duplication on this feature did not reveal duplicates. Granting the aforementioned labels the assumption of uniquely identifying an observation was done so due to the fact the labels were in the format of integers and given the name with postfix “ID”, making it safe to assume each was unique to the observation.

**NA Removal**

The purpose of this data intake will be to assist the company in making their decision to invest in one of the given cab companies, so observations with a ‘Transaction ID’ in the ‘Transaction ID.csv’ file that is not present in any observations in the ‘Cab\_Data.csv’ file will be discarded in the final dataset, as these observations will not have any associated cab data, notably the ‘Company’ feature associated with them. Thus, the final dataset will only contain transactions that had the associated cab data collected. In the final dataset, these to-be-discarded observations will show up as observations with ‘NaN’ labels corresponding to features originating from the ‘Cab\_Data.csv’ file.