Data Intake Report

Name: Cab Industry Investment Analysis

Report date: Mar 12 2023

Internship Batch: Feb to May 2023

Version: 1

Data intake by: Sai Teja Reddy Jonnala

Data intake reviewer:

Data storage location:

**Tabular data details:**

Cab\_Data

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | <Number of files received> |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 21.2 MB |

City

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | <Number of files received> |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 B |

Customer\_ID

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | <Number of files received> |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.05 MB |

Transaction\_ID

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** | <Number of files received> |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 19 B |

**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)

When it comes to data quality analysis, identifying and handling duplicate data is an important step to ensure that the results are accurate and reliable. The proposed approach for deduplication validation can involve the following steps :

1. Identify the columns in the dataset that can be used to identify duplicates. In the case of these datasets, unique identifiers such as Transaction ID and Customer ID can be used to identify duplicate transactions and customers respectively.
2. Check for exact duplicates by comparing all the columns in the dataset. Any rows that have the same values in all columns can be considered duplicates and should be removed from the dataset.
3. Check for partial duplicates by comparing a subset of columns in the dataset. This can be useful in cases where the datasets have additional columns that may contain different values for the same record. For example, in the case of the City\_Data dataset, two cities with the same population and user count can be considered duplicates even if they have different names.
4. Use data cleaning techniques such as normalization to standardize the data before performing the deduplication validation. This can help to identify duplicates that may not be immediately obvious due to formatting or data entry errors.

**Assumptions :**

* We assume that the unique identifiers (Transaction ID and Customer ID) are truly unique and have not been duplicated or misused in any way.
* We assume that the datasets have been properly formatted and do not contain any data entry errors or anomalies.
* We assume that there are no additional datasets that need to be merged or integrated with these datasets, which may lead to additional duplicates or discrepancies that need to be resolved.

**You should fully investigate and understand each data set. Review the Source Documentation Understand the field names and data types Identify relationships across the files Field/feature transformations Determine which files should be joined versus which ones should be appended Create master data and explain the relationship Identify and remove duplicates Perform other analysis like NA value and outlier detection**

**Cab\_Data:**

* Field names: Transaction ID, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip
* Data types: integer, integer, string, string, float, float, float
* Relationships: The Transaction ID in this dataset can be used to join with the Transaction\_Data dataset.
* Transformations: The Date of Travel field can be transformed into a datetime format for easier analysis and visualization.
* Other analysis: Outlier detection can be performed on the KM Travelled, Price Charged, and Cost of Trip fields.

**City\_Data:**

* Field names: City, Population, Users
* Data types: string, string, string
* Relationships: The City field in this dataset can be used to join with the Cab\_Data dataset.
* Transformations: The Population and Users fields can be transformed into integer format for easier analysis.
* Other analysis: None needed for this dataset.

**Customer\_Data:**

* Field names: Customer ID, Gender, Age, Income (USD/Month)
* Data types: integer, string, integer, integer
* Relationships: The Customer ID in this dataset can be used to join with the Transaction\_Data dataset.
* Transformations: None needed for this dataset.
* Other analysis: Outlier detection can be performed on the Age and Income fields.

**Transaction\_Data:**

* Field names: Transaction ID, Customer ID, Payment\_Mode
* Data types: integer, integer, string
* Relationships: The Transaction ID and Customer ID in this dataset can be used to join with the Cab\_Data and Customer\_Data datasets.
* Transformations: None needed for this dataset.
* Other analysis: None needed for this dataset.