Data Intake Report

Name: Market Insights for XYZ

Report date: 20/09/2022

Internship Batch: LISUM13:30

Version: 1.0

Data intake by: Oluwayomi Omodemi

Data storage location: https://github.com/YOMIUS/Cab-Data-Analysis

**Tabular data details:**

Cab­\_Data.csv

|  |  |
| --- | --- |
| **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** | 19MB |

City.csv

|  |  |
| --- | --- |
| **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** | 608 Bytes |

Customer\_ID.csv

|  |  |
| --- | --- |
| **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** | 1.5MB |

Transaction\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 440908 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 10MB |

**Master Dataset**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 1 |
| **Total number of features** | 14 |
| **Base format of the file** | .csv |
| **Size of the data** | 40MB |

**Proposed Approach:**

The provided datasets were combined into a main dataset so exploratory data analysis could be carried out on it. Before joining the datasets, to ensure the datatypes of the features, empty spaces in the columns for population and users are removed and the datatype is converted from object to an integer.

The datasets are loaded using pandas and converted to dataframes so we can explore the data contained within. After performing a check for null values and seeing as there are none, the datasets are merged on corresponding values in each other.

To create the master dataset:

* Cab\_data is merged to City on the City field
* The emergent result is merged as with transaction data on the transaction id field
* Next result dataframe is then merged with the customer data on the customer id field

The result is an expansion on the initial cab data provided. The dataframe contains 14 columns properly ordered and corresponding to the values in cab data.

The features contained in the data are as follows:

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Data type |
| Transaction\_ID: | The transaction id of the rides users made from 2016 to 2018 | Integer (int64) |
| Date\_of\_Travel: | The date the ride took place (The data is from 01/01/2016 to 31/12/2018 ). | Datetime  (datetime64[ns]) |
| Company: | The company the cab ride was for (Either Pink Cab or Yellow Cab) | String (object) |
| City: | The city the trip took place in | String (object) |
| KM\_Travelled: | The distance the recorded ride covered for the trip | Float (float64) |
| Price\_Charged: | The amount the user paid for the trip | Float (float64) |
| Cost\_of\_Trip: | The actual cost to the company for the trip | Float (float64) |
| Population: | The total population of the city the trip took place in | Integer (int64) |
| Users: | The total number of users for the respective cab companies per city | Integer (int64) |
| Customer\_ID: | The customer id of the user | Integer (int64) |
| Payment\_Mode: | The mode of payment used by the user to pay for the trip | String (object) |
| Gender: | The gender of the user (Male or Female) | String (object) |
| Age: | The age of the user | Integer (int64) |
| Income\_(USD/Month): | The monthly income of the user | Integer (int64) |