Data Intake Report

Name: <G2M Case Study>
Report date: <07/03/2021>
Internship Batch:<LISP01>
Version:<1.0>
Data intake by:<Nonhlanhla Luphade>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/shelovescode000/DataSets>

Tabular data details:

Total number of observations	<359393>
Total number of files	<5>
Total number of features	<22>
Base format of the file	<.csv>
Size of the data	<56.156 MB>

Proposed Approach:

- Five data sets (Cab_data.csv, USHoliday.csv, Customer_ID.csv, Transaction_ID.csv and City.csv) were combined to one dataset.
- The full data set consists of 22 variables (7 derived features)
- The data was collected from 31/01/2016 to 31/12/2018

Assumptions:

- The difference between the profit_charged and cost_of_trip can be used to calculate profit.
- The customer IDs are the same for both companies.
- The number of cab users is an approximate of the total number of all cab users in the city.

Manipulations:

- Created an age group column for age groups (16-24, 25-34, 35-44 and 55+) using the age column.
- Created a holiday column which checks whether it is a holiday or a normal day.
- Created an income class column which uses the income column values to check if a customer is in low-income class or middle-income or high-income class.
    - Low-income class = [income < lower quartile]
    - Middle-income class = [lower quartile < income < mean]
    - High-income class = [income > mean]
- Created distance travelled column which places the distances in km to 3 categories (short [0-10 km], normal [10-30 km] and long [>30 km]).