Data Intake Report

Name: Cab Transactions Data
Report date: 28 December 2021
Internship Batch: LISUM05
Version: 1.0
Data intake by: Diego Martínez
Data intake reviewer:
Data storage location: Dataset at my repo / Dataset at DG's repo (clones)
Tabular data details:

Filename	Cab_Data
Total number of observations	359392
Total number of features	7
Base format of the file	.csv
Size of the data	21 MB

Filename	City
Total number of observations	19
Total number of features	3
Base format of the file	.csv
Size of the data	4 KB

Filename	Customer_ID
Total number of observations	49170
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 MB

Filename	Transaction_ID
Total number of observations	440097
Total number of features	3
Base format of the file	.csv
Size of the data	8,6 MB

Approach:

- I will consider Cab_Data.csv as the “main” dataframe, since each rows represents a transaction and contains the “Price Charged” and “Cost of Trip” fields, which are extremely relevant.
- To ease plotting and data wrangling, I will merge/join the four dataframes into a single “master data”, with which I will be working on in the EDA. This is inefficient memory-wise, but since the size of the data is manageable the convenience outweights the (memory) efficiency.

Assumptions:

- The field "Cost of Trip" includes every possible expense, in such a way that the profit of a trip can be defined as the difference between its "Price Charged" and "Cost of Trip".
- The field "Date of Travel", whose values are integers, represents days; two adjacent integers represent two adjacent days and the lowest integer in the record corresponds to the starting date indicated in the assignment (31/01/2016).