Data Intake Report

Name: **Cab Industry Analysis**

Report date: **15/07/2022**

Internship Batch: **LISUM 11:30**

Version: **1.0**

Data intake by: **Anastasia Apanasiuk**

Data intake reviewer: **-//-**

Data storage location: **https://github.com/apaanastasia93/data_glacier.git**

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	21.2 MB

Tabular data details: 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	759 B

Tabular data details: 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	1.05 MB

Tabular data details: Transaction_ID.csv

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	9 MB

Proposed Approach:

- Dedup validation:
  All four datasets contain different data, therefore duplicating data can be excluded.
  The datasets contain collected data from 01 January 2016 till 31 December 2018. There are two common columns 'Transaction ID' and 'Customer ID' which can be used for merging datasets.

- Assumptions:
  'Cost of Trip' contains data of what the cost of a trip was, while 'Price Charged' shows what price had been payed by a customer.
  New features have been created:

1. Income feature is representing a flat company's income after subtracting costs of trip:
   Income *= Price Charged - Cost of Trip*
2. Price Per KM shows what price each company charges per km of a trip:
   Price_Per_KM *= Cost of Trip / KM Travelled*