Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 08.03.2021
Internship Batch: LISP01
Version: <1.0>
Data intake by: Pelinsu ÇELEBİ
Data intake reviewer: **<**intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets

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	20.1 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	4 KB

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 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	8.58 MB

Proposed Approach:

- The datasets are merged into one based on common keys in each set by inner joining them
- The Date of Travel is filtered to capture dates between 31.01.201 and 31.12.2018 as required
- No null values found that needs to be handled
- Outliers detected in Price Charged column and rows containing outliers are discarded from the dataset
- Two companies were compared based different properties obtained by grouping, aggregating, counting the different columns of the dataset and inspecting the visualizations
- K-means clustering performed on Customer data to obtain the different segments of customers to inspect