Data Intake Report

Name:	Data Analytics for Investment Decision- Assignment 2
Report date:	02/03/2021
Internship Batch:	LISP01
Version:	1.0
Data intake by:	Ajaegbu Ebuka Emmanuel
Data intake reviewer:	<intern who reviewed the report>
Data storage location:	https://github.com/EEAjaegbu/Investment-Analytisc--Assignmet-2

DATA

Customer ID Data File

Total number of observations	49171
Total number of files	1
Total number of features	5
Base format of the file	.txt
Size of the data	1304 KB

Cabdata Data File

Total number of observations	359392
Total number of files	1
Total number of features	8
Base format of the file	.csv
Size of the data	24,437 KB

City Data File

Total number of observations	20
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1 KB

Transaction_ID Data File

Total number of observations	440098
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	11,688 KB

Proposed Approach:

Dedup Validation (Identification):

The danger of duplicates in the data used for analysis is that it bias the result of the analysis. To identify duplicates entry or cases in the data, duplicated () method in pandas will be used to find duplicated rows while the drop_duplicates () method in pandas will be used in removing duplicates entry in the data.

Assumption:

The two sample T test was used to test statistically the hypothesis of no difference between the mean of two unknown population and he Pearson test was used to test for significant of correlation between tow variables. We assume that the data set are randomly drawn from two unknown IID (independently and identically distributed) population.