Data Intake Report

Name: Data Glacier Final Project
Report date: 05.06.2022
Internship Batch:LISUM-09
Version:<1.0>
Data intake by:Abdullah GÖK
Data intake reviewer:
Data storage location: https://github.com/abdullahgk/Data-Glacier-Internship-Assignments/tree/main/Data%20Glacier%20Final%20Project

Tabular data details:

Total number of observations	13647309
Total number of files	2
Total number of features	48
Base format of the file	csv
Size of the data	2,2 GB

DATA CLEANING WALKTHROUGH

1. **Our data had many problems at first , having many nan values , ambiguous column names , string typo errors , datetime mismatch and mixed datatypes in a single column**
2. **I translated and updated column names with English to make it more clear.**
3. **'Last_Date_Prim_Cust', 'Spouse_Index' columns were having so many missing rows and I want to deleted them not to cause any error on my EDA.**
4. **“Gross_Income” were having nan values and I filled them with mean gross income .**
5. **I filled “Segmentation” with the categorical object “not provided”.**
6. **'Cust_Type_Begin_Month' I filled nan values with “0” and corrected values to meaningful ones like 4.0 to 4 etc.**
7. **The rest of the columns I dropped missing rows because their percentage was so low to effect the general analysis.**
8. **I separated date column with “month” and created a new column for EDA.**
9. **In several rows(for example Loans,Funds..) I replaced 0 with “No” and 1 “Yes” in order to understand data better.**
10. **'Customer_Seniority' column was having object typo errors. I cleaned the rows using str.strip() method and change the type with integer.**
11. **As there was only one type of “Adress Type” column so I dropped it.**