Data Intake Report

Name: <Bank Marketing Campaign>
Report date: <18/03/2024>
Internship Batch:<LISUM31>
Version:<1.0>
Data intake by:<Glacier Analysis Group>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <GitHub>

Tabular data details, Bank_Additional Data:

Total number of observations	4119
Total number of files	1
Total number of features	21
Base format of the file	csv
Size of the data	0.032 MB

Tabular data details, Bank_Additional_Full Data:

Total number of observations	41,188
Total number of files	1
Total number of features	21
Base format of the file	csv
Size of the data	6.6 MB

Tabular data details, Bank Data:

Total number of observations	4,521
Total number of files	1
Total number of features	17
Base format of the file	csv
Size of the data	0.6006 MB

Tabular data details, Bank_Full Data:

Total number of observations	45,211
Total number of files	1
Total number of features	17
Base format of the file	csv
Size of the data	5.9 MB

Proposed Approach:

Approach of dedup validation (identification):

- We loaded the datasets using the pandas function pd.read_csv() and specified the delimiter as “;” to read the datasets.
- Utilizing pandas functions like info(), we identified the number of observations, features and size of each Dataframe.

Assumptions

- We assumed that the data follows a consistent format across all columns, including date format, numerical format, and categorical format.
- The data entry process was assumed to be accurate, minimizing errors and inconsistencies in the dataset.
- The dataset was assumed to be complete, with no missing values that could impact the analysis.
- The source of the data was assumed to be reliable and trustworthy, reducing the risk of data inaccuracies or biases.