Name: Shatabdi Pal
Email Id: shatabdi@pdx.edu
Country: USA

Data Intake Report

Name: Bank Marketing (Campaign)
Report date: 07/03/2024
Internship Batch: LISUM34
Version: 1.0
Data intake by: Shatabdi Pal
Data intake reviewer:
Data storage location:

Tabular data details: bank-full.csv

Total number of observations	452111
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	4503 KB

Tabular data details: bank.csv

Total number of observations	4521
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	451 KB

Tabular data details: bank_additional_full.csv

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	5669 KB

Tabular data details: bank_additional.csv

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	571 KB

Proposed Approach of Deduplication Validation:

1. Data Selection: Selected bank-additional.csv over bank.csv due to common columns and additional data in bank-additional.csv. The extra data in 'bank-additional.csv' provides more comprehensive information for our analysis, hence the selection.
2. Data Integration: Merged bank-additional.csv and bank-additional-full.csv to incorporate all relevant data.
3. Duplicate Detection: Identified duplicate records based on criteria such as identical values across specified columns.
4. Deduplication Process: Remove duplicate records to ensure each entry in the final dataset is unique and represents distinct data instances.

Assumptions for Data Quality Analysis:

1. **Consistency:** Assumed that the dataset follows consistent data formats and conventions across all files (bank.csv, bank-full.csv, bank-additional.csv, bank-additional-full.csv).
2. **Completeness:** Based on the given dataset information, assume all files (bank-additional.csv, bank-additional-full.csv) are complete and do not contain null values.
3. **Data Integrity:** Expected that there are no discrepancies or anomalies that could compromise data integrity during the merge and deduplication processes.
4. **Common Key**: Relied on a shared key or set of keys to accurately identify and remove duplicate records across the merged dataset (bank-additional.csv and bank-additionalfull.csv).
By following this approach and assuming these conditions, you can ensure the final data set is cleaned of duplicates and ready for further analysis.