Group Name: Profit
Name: Jinghao Shen
Email: jshen30@ucsc.edu
Country: United States
College/Company: University of California, Santa Cruz
Specialization: Data Science
Date: 10/26/2023

Github Repo link:
https://github.com/jshen1s1/BankMarketing

Data Variable Information:

1. age: int
2. job: string - 17 categories
3. marital: string - 3 categories
4. education: string - 4 categories
5. default: string - binary
6. balance: int
7. housing: string - binary
8. loan: string - binary
9. contact: string - 3 categories
10. day: int
11. month: string - 12 categories
12. duration: int
13. campaign: int
14. pdays: int
15. previous: int
16. poutcome: string - 4 categories
17. y: string - binary

- #6 annually, count in euros
- #10 + #11 = last contact date and month
- #12 count in seconds
- #13 number of contacts performed during this campaign and for this client
- #14 number of days that passed by after the client was last contacted from a previous campaign
- #14 -1 as not previously contacted
- #15 number of contacts performed before this campaign and for this client
- #17 outcome (target)

Possible Data Problems:

1. Missing attribute value: None
2. Duplicate: None
3. Categorical columns: #2, #3, #4, #5, #7, #8, #9, #11, #16, #17

4. Long-tail distribution: #9, #14, #15, and #16

Possible Approaches:

- Normalization: Replace string variables with a 4-byte integer. This will save us storage space and be more appropriate for modeling.
    - Create tables for each column and populate them with unique values
    - Use LabelEncoder from sklearn.preprocessing for label encoding
    - Use get_dummies for one-hot encoding
- Outliers, Skewed:
    - Remove suspicious data
    - Grouping categories: Reduce the number of categories. Combine tailed data.
- Correlation: Get an idea of how numerical attributes relate to each other.
    - Use tools from ibmdbpy to read the result as a graph
    - Combine #10 and #11 into one column