

**Data Science Intern at Data Glazier**

**Week 8: Report on Group Project**

**Topic: Bank Marketing (Campaign)**

**Group Name: Campaign Catalysts**

**Specialization:** Data Science

**Batch Code:** LISUM19

**Date:** 2nd May 2023

**Submitted to:** Data Glacier

**Handling missing data for Bank marketing dataset**

Data cleaning is an important step in data pre-processing due to its ability to help improve the

quality of the dataset for a more reliable output. The presence of impurities in real-world data

application has brought about the development of several methods to eradicate this problem

to help improve the accuracy and usability of existing data (Müller and Freytag, 2005). [3] The data

cleaning process involves the detection or removal of outliers, smoothing noisy data, filling in

missing values and resolving inconsistency within a dataset (Han, Pei and Kamber, 2011). [4]

There is exactly no one way of dealing with missing data. There are different solutions for data imputation depending on the kind of problem and it always difficult to provide a general solution, and care should be taken when it comes to removing missing values in any given data set since doing so will introduce biasness in the model.

**Imputing or Deleting missing values of the Data:**

Before we decide to remove, replace or impute the data, we have to understand and establish the reason why data is missing.

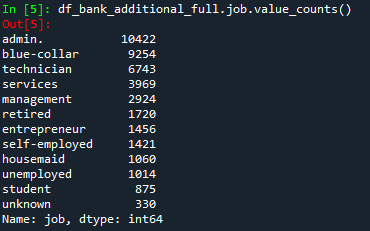
* Missing at Random: This means that the tendency for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
* Missing Completely at Random: The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
* Missing not at Random: Possible reasons are that the missing value depends on the hypothetical value or missing value is dependent on some other variable’s value.

**Data cleaning of categorical features in the data set:**

We have the following unknown values for some of the features in the data set:

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Unknown Values** | **Minimum value** | **Maximum value** |
| Job | 990 | unknown | admin |
| Marital | 80 | unknown | married |
| Education | 1731 | illiterate | university degree |
| Default | 8597 | yes | no |
| Housing | 990 | unknown | yes |
| Loan | 990 | unknown | no |
| contact | 0 | nil | nil |
| month | 0 | nil | nil |
| day\_of\_week | 0 | nil | nil |
| poutcome | 35563 | success | nonexistence |
| y | 0 | yes | no |

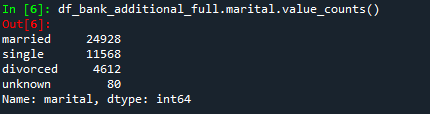
**Categorical Feature: Job**



From the above entries in the data set, the admin job-type records and makes up about 25% of the entire data set column as compared to the unknown entries making about 0.8% of the entire column followed by the student entries with 2.1%.

So for the unknown values for the categorical feature job will be imputed with the admin value type since that’s the most prevalent jog type in the job column.

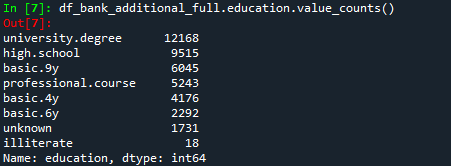
**Categorical Feature: Marital**



From the above entries in the data set, the married value-type records and makes up about 60.8% of the entire data set column as compared to the unknown entries making about 0.19% of the entire column followed by the divorced entries with 11.1%.

So for the unknown values for the categorical feature marital will be imputed with the married value type since that’s the most prevalent jog type in the job column.

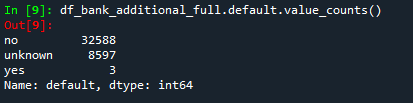
**Categorical Feature: Education**



From the above entries in the data set, the university. degree type records and makes up about 29.5% of the entire data set column as compared to the unknown entries making about 4.2% of the entire column followed by the illiterate entries with 0.04%.

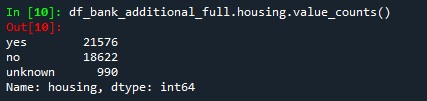
So for the unknown values for the categorical feature education will be imputed with the university. degree value type since that’s the most prevalent jog type in the job column.

**Categorical Feature: Default**



From the above entries in the data set, the no-type records and makes up about 79.1% of the entire data set column as compared to the unknown entries making about 20.8% of the entire column followed by the yes entries with 0.007%, making this a very skewed distribution and more biased towards the majority entries meaning there is no enough information concerning this feature, hence could be discarded or remove from the data set to eliminate any further biasness.

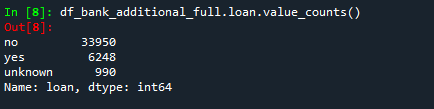
**Categorical Feature: Housing**



From the above entries in the data set, the yes-type records and makes up about 52.4% of the entire data set column as compared to the no entries making about 45.2% of the entire column followed by the unknown entries with 2.4%.

Ideally the difference in the numbers of both entries are not that much so for the unknown values for the categorical feature housing will be imputed with using the random choice mode of selection for this purpose, meaning the 2.4% entries for the unknown entries will be randomly shared across the two.

**Categorical Feature: Loan**



From the above entries in the data set, the no value-type records and makes up about 82.4% of the entire data set column as compared to the unknown entries making about 2.4% of the entire column followed by the yes entries with 15.1%.

So for the unknown values for the categorical feature loan will be imputed with the no type since that’s the most prevalent jog type in the job column.

**References:**

[3]: Müller, H. & Freytag, J.C. (2005) Problems, methods, and challenges in comprehensive data

cleansing. Professoren des Inst. Für Informatik.

[4]: Han, J., Kamber, M. & Pei, J. (2011) Data mining: concepts and techniques. Elsevier