**Data Imputation**

As you may know data these has lots of missing values. They may be a result of the way data is primarily being collected. For example, from surveys or questionnaire where a person chooses not to answer a particular set of questions. There is not much a surveyor can do when handling such data. There also may be a number of other reasons that the data is missing. Schematically, several cases are possible: (i) the value is missing because if was forgotten or lost, (ii) the value is missing because it was not applicable to the instance, (iii) the value is missing because it is of no interest to the instance [cite… <https://www.researchgate.net/publication/308007055_Missing_Data>]

Missing values lead to incomplete data which can have problems further when working with it. These create a number of potential challenges when doing statistical analysis. A more obvious observation being missing data can relatively shrink the sample size of data. Hence, Missing values can increase the chances of making Type I and Type II errors, reduce statistical power, and limit the reliability of confidence intervals. [cite… <https://www.researchgate.net/publication/320928605_Data_Imputation>].

1. **Type of Missingness**

A number of mechanisms for missingness can be defined depending upon the relation between the observed values and the unobserved or the missing value. They are classified into there major categories:

a) Missing completely at random (MCAR): When we make this assumption about data, we assume that the missing data is completely unrelated to the rest of the information being provided in the data i.e. it states that the missingness of data is unrelated to the observed as well as the missing data in the data. Such missingness is said to be ignorable.

b) Missing at random (MAR)-a weaker assumption than MCAR-: This type of missingness assumes that the probability that a value is missing is related to the observed data i.e. the observed data is statistically related to the missing variables and it is possible to estimate the missing value from the observed data.

[cite… <https://www.researchgate.net/publication/308007055_Missing_Data>]

c) Not missing at random (NMAR): Here, the missingness of the data is dependent both on the observed as well as the missing values and hence determining the missingness is usually unattainable. For such data, any standard imputation methods give wrong answers and hence can’t be used.

The missing data namely MCAR and MAR can be dealt by imputing them. Data imputation simply means to replace the missing values in dataset with alternate values derived from computations or otherwise.

Our dataset consisted mostly of ‘unknown’ rather than Null values. They were labeled as ‘unknown’ as depicted in Figure 1.1

A screenshot of a computer

Description automatically generated

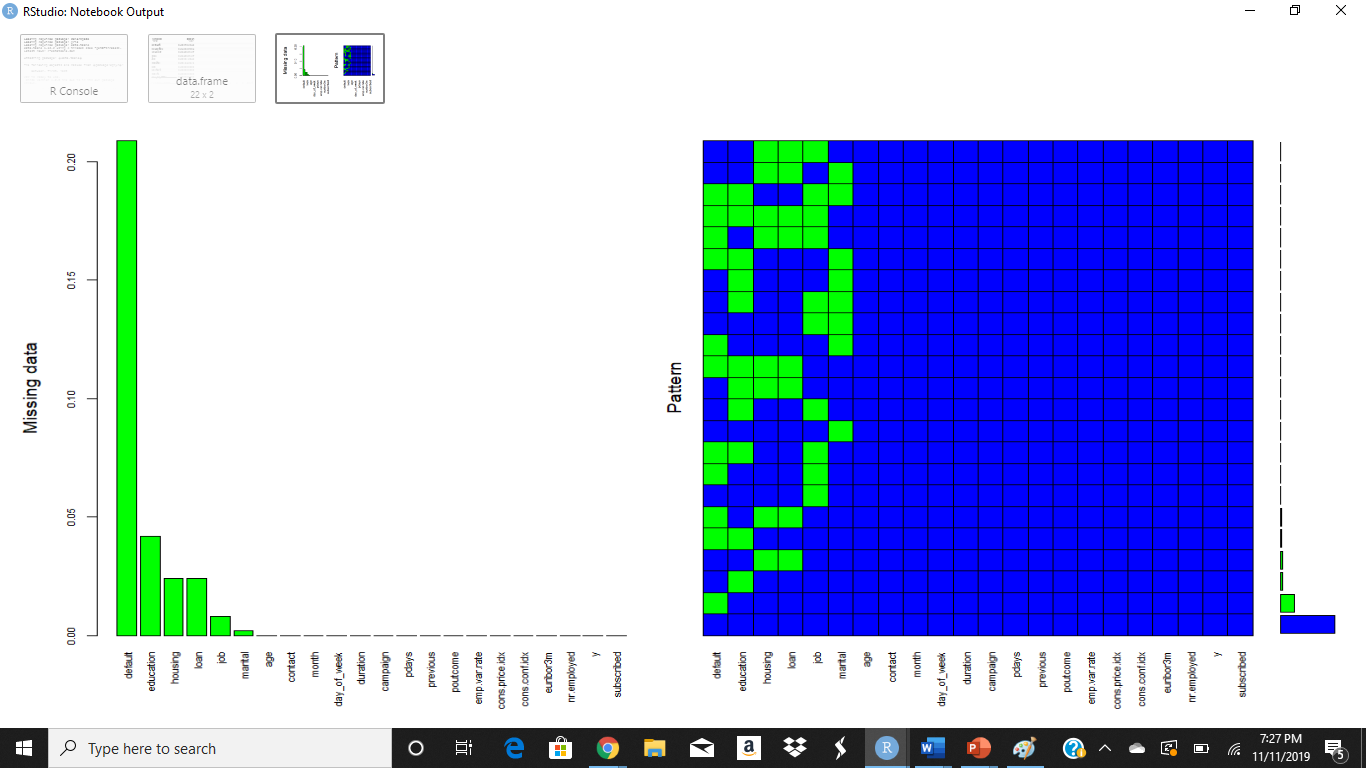
**Figure 1.1**

1. **Dealing with the unknowns**

As the number of unknows were high and were interfering with the results, we had to make a choice on how to deal with them. As we had no null values or blanks in the dataset but only unknowns, we replaced all the unknowns in the datasets by ‘NA’. The following code was used for the same.

data[data == 'unknown'] <- NA

Figure 1.2 gives us tradeoff between the proportion of missing data for each variable and the second plot gives a gist of how the variables had missing data in combination.



**Figure 1.2**

The Missing data plot tells you the percent of NA across each variable. As can be seen the variable default which means whether the person had credit in default had the highest percent of NA namely 23% followed by education at 4% and then housing. The next plot i.e. the Pattern plot tells us how the NA have occurred in combination. For example, in the fourth row from bottom as can be seen housing and loan are both green stating that both housing and loan have maximum number of rows having NA occurring simultaneously.

After replacing unknown with NA, the next step was data imputation. Rather than choosing mean or median for data imputation or completely getting rid of the variables having unknown, we used the MICE package for imputation.

1. **Multivariate Imputation by Chained Equations (MICE)**

MICE or Multivariate Imputation by Chained Equations deals with methods that impute data giving plausible results. As the name suggests, it creates multiple imputations for the same set of missing data. As this process involves assigning values to the missing data multiple times, it creates a set of complete datasets which have multiple values for the same set of missing observations. The advantage of using MICE over single imputation methods such as mean is that it handles the statistical uncertainty which arises in missing values when using methods of single imputation.

MICE function automatically detects values that are missing.  The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data [cite… <https://www.rdocumentation.org/packages/mice/versions/3.6.0/topics/mice>].

MICE makes the use of method based on FCS or  Fully Conditional Specification which imputes missing values variable-by-variable by defining an imputation model for each variable with missing data, conditioned on the remaining variable. [cite… <https://cs.brown.edu/research/pubs/theses/ugrad/2019/farley.edwin.pdf>].

MICE assumes that the data are MAR or Missing at Random and hence uses only the observed values to predict them. The command we used for imputing missing data using mice package is as follows:

idata <- idata <- mice(data, m=5, maxit=5, meth ='pmm', seed=500)

Here, the arguments for the above command are as follows:

data : The data frame used. Here its data

m : Number of imputations. By default, it is set to 5.

maxit : A scalar value giving maximum number of iterations. By default, it is set to 5

meth : The method used for carrying out imputation . Here ‘PMM’ is used.

The output of running the above command is as follows:

A screenshot of a cell phone

Description automatically generated

**Figure 1.3**

As you can see, Figure 1.3 tells you which all variables used PMM method for imputation. PMM stands for Predictive Mean Matching. This method forms a candidate set of donors of possibly 3 to 10 members from the pool of donors for each missing entry. These donors selected have predicted a value for the missing entry. This donor predicted value is one of the closest values to that of the predicted value for the missing entry. Of these candidate donors, one donor is randomly chosen, and the observed value of that donor is selected as the replacement for the missing entry.

Other methods for imputation in MICE include using **Logreg** for logistic regression imputation, **polyreg** for polytomous regression imputation and **polr** for proportional odds model