# **Missing Values**

## **Introduction**

Missing data, or missing values, occur when no data / no value is stored for certain observations within a variable.

Incomplete data is an unavoidable problem in dealing with most data sources. Missing data is a common occurrence in both data science competitions and business datasets, and may have a significant impact on the conclusions that can be derived from the data.

## **Why is data missing?**

The source of missing data can be very different. These are just a few examples:

* A value is missing because it was forgotten, lost or not stored properly
* For a certain observation, the value of the variable does not exist
* The value can't be known or identified

In many organisations, information is collected manually into a form by a person talking with a client on the phone, or alternatively, by customers filling forms online. Often, the person entering the data does not complete all the fields in the form. Many of the fields are not compulsory, and therefore, those values will be missing. The reasons for omitting the information in those fields can vary: perhaps the client does not want to disclose some information, for example income, or they do not know the answer, or the answer is not applicable for a certain circumstance, or on the contrary, the person in the organisation wants to spare the client some time, and therefore omits asking questions they think are not so relevant.

There are other cases where the value for a certain variable does not exist. For example, in the variable 'total debt as a percentage of total income' (very common in financial data), if the person has no income, then the total percentage of 0 does not exist, and therefore it will be a missing value.

It is important to understand how the missing data are introduced in the dataset, or in other words, the mechanisms by which missing information is introduced in a dataset. Depending on the mechanism, we may choose to process the missing values differently. In addition, by knowing the source of missing data, we may choose to take action to control that source, and decrease the number of missing data looking forward during data collection.

## **Missing Data Mechanism**

There are 3 mechanisms that lead to missing data, 2 of them involve missing data randomly or almost-randomly, and the third one involves a systematic loss of data.

**Missing Completely at Random, MCAR:**

A variable is missing completely at random (MCAR) if the probability of being missing is the same for all the observations. When data is MCAR, there is absolutely no relationship between the data missing and any other values, observed or missing, within the dataset. In other words, those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than other. If values for observations are missing completely at random, then disregarding those cases would not bias the inferences made.

**Missing at Random, MAR:**

MAR occurs when there is a relationship between the propensity of missing values and the observed data. In other words, the probability of an observation being missing depends on available information (i.e., other variables in the dataset). For example, if men are more likely to disclose their weight than women, weight is MAR. The weight information will be missing at random for those men and women who do not disclose their weight, but as men are more prone to disclose it, there will be more missing values for women than for men.

In a situation like the above, if we decide to proceed with the variable with missing values (in this case weight), we might benefit from including gender to control the bias in weight for the missing observations.

**Missing Not at Random, MNAR:**

Missing data is not at random (MNAR) when there is a mechanism or a reason why missing values are introduced in the dataset. For example, MNAR would occur if people failed to fill in a depression survey because of their level of depression. Here, the missing of data is related to the outcome, depression. Similarly, when a financial company asks for bank and identity documents from customers in order to prevent identity fraud, typically, fraudsters impersonating someone else will not upload documents, because they don't have them, because they are fraudsters. Therefore, there is a systematic relationship between the missing documents and the target we want to predict: fraud.

Understanding the mechanism by which data is missing is important to decide which methods to use to impute the missing values.