test

## Data

The supplied dataset contains 2500 measurements of 20 numerical variables.

The analysis team and report writer were not supplied with any additional information about the data, e.g.:

* what has been measured
* variable importance, interactions
* measurement scale, units, reliability

Therefore, variables are treated equally and without bias in the analysis and modelling. Decisions are made on statistical grounds and stated assumptions.

|  |
| --- |
| Tip |
| Additional contextual information about the data could result in alternative modelling decisions and outcomes and should be considered. |

### Missing data

Before proceeding with [Exploratory Data Analysis](#sec-exploratory-data-analysis), it is important to explore the extent of missing data and whether there are patterns to the missingness. Some classification models assume no missing data, and depending on the amount, prevalence and patterning of missing data, different assumptions and techniques can be applied.

We investigate missingness to see whether the assumption that missing data is ‘missing at random’ (MAR) holds - that is, is the probability of missingness only dependent on **observed** variables? If so and there is a significant amount of missing data (>5%), multiple imputation approaches may be explored to *impute* missing values by estimation through statistical inference.

If the missing data is not MAR - that is, it appears to depend on **unobserved** variables, there may be other, more appropriate imputation methods such as maximum likelihood imputation. Unobserved influences may include the measurement processes or chemical compositions of the samples.

*Imputation* can introduce bias into the analysis if the assumptions of the imputation method are violated, or the imputed values differ significantly from true missing values.

|  |
| --- |
| Caution |
| The client should consider missingness in their data:   * Are there any concerns regarding missing data? * What are the reasons for missing data? Can they be mitigated? * Can data be collected where it has been identified as missing? * Is there likely to be (more, less, similar amount of) missing data in the future? |