## Overview

Data cleaning is the process of organizing and transforming raw data into a dataset that can be easily accessed and analysed. A data cleaning plan is a written proposal outlining how you plan to transform your raw data into clean and usable data.

While this is not a comprehensive list of every possible data-cleaning step, these are common steps that typically apply to the tools or technologies, data-cleaning steps, and considerations or potential issues that may arise during the clean-up process.

1. Understand the data source, objective and goals of the procedure to be performed.
2. Identify the data tool to be considered
3. Identify the applicable library/dependencies to be utilised on the data tool.
4. Read raw data into the Python data tool, Jupyter Notebook.
   * Extract data from a web portal folder, un-zip data, and download it to a CSV data file.
   * Extract data via the GET request from an API
   * Create an API key where applicable.
   * Ensure that all of the data cleanings for these files were done in the Jupyter Notebook, never touching the raw data
5. Review the data
   * Rows, columns, values, missing data
   * Check your columns against your data dictionary
6. Understand data
   * Valid
   * Accurate
   * Complete
   * Consistent
   * Uniform
7. Perform the required data procedure and drop any unimportant columns not required.
8. Describe the data to gain more understanding of the data procedure.
9. Visualisation is important, utilise data visual libraries to gain more understanding of the data procedure.
10. Rename columns to match your data dictionary where applicable
11. Check variable classes/types and update as needed
12. Transform/Normalize variables
    * Remove symbols from columns such as % or $
    * Remove white space from strings
13. Standardize
    * Make sure all of your variables are measured the same way within your dataset and the same way across time
    * Rescale variables
14. Validate data
    * You can also create reports/markdowns that report issues in the dataset
    * Things to check for in your validation process
      + Values out of range
      + This is a chance to check that an ID wasn’t entered incorrectly
15. Merge and/or append data as needed
    * Across time, forms, data sources
    * Again do validation checks, do you have the correct number of rows and columns after the merge?
16. Export clean data to a specified format (.csv, .txt, .xlsx)
    * Again, review recommendations for best formats for [long-term data storage](https://dmptool.org/general_guidance#file-formats)
17. Data transform strategies to clean and harmonise data in the database should be considered on data filtered for
    * Active Company Status because we have “Active” and “Active - Proposal to Strike off”.
    * Accounts Category identified as ‘SMALL’ because we have other details as “TOTAL EXEMPTION SMALL ”.
    * Data obtained from the GET request should have an iteration element to avoid manual pull requests of the data set
    * Frequent data accuracy and completeness procedure should be considered.