The Checklist for Exploring Data

Congratulations! You were able to open your file in a spreadsheet editor such as Excel. Now have a look at the following points.

Check your metadata

- ☐ Do you know who released the data?
- ☐ Do you understand how you can use the data under its licence?
- ☐ Do you know the release date and/or time span of the data?
- ☐ Have you counted the number of rows?
- ☐ Have you counted the number of columns?
- ☐ Does the title of the file correspond to the content?
- ☐ Do you have a data dictionary, i.e. an explanation of the columns and potentially abbreviations?
- ☐ Do you know the unit of analysis, i.e. what's a ‘record’, what’s in a row?
- ☐ Do you have a complicated structure such as a longitudinal dataset, e.g. a file with countries over time?
- ☐ Are there any notes, caveats or comments regarding you dataset?

Check your data formats

- ☐ Have you looked at the first 5 rows?
- ☐ Have you looked at the last 5 rows?
- ☐ Have you looked at 5 random rows?
- ☐ If there are **numbers**, are they recognised as a number format?
- ☐ Are the units consistent?
- ☐ If there are **dates**, are they recognised as a date format?
- ☐ Do you understand the date format (e.g. 13/12 vs 12/13)?
- ☐ If there are **categorical** variables (e.g. months), are they stored as strings (e.g. “Dec”) or as factors/other (e.g. “12”)?
- ☐ Do you know the complete list of allowed answers for categorical variables? E.g. for quarter: spring, summer, autumn, winter and Q1-Q4.
- ☐ If there are **string** variables (= text), do you know if it was entered manually? (If so, take extra care!)

Check your missing data

- ☐ Do you know what symbol stands for a missing data or "NA"?
- ☐ Are there any other codes such as "unknown" or "-99" that might also be missing data?
- ☐ Did you count the number of missing data?

- ☐ Have you counted the number of non-missing entries for each variable?
- ☐ Are there any rows that are empty or almost empty?
- ☐ Are there any columns that are empty or almost empty (less than 5% of valid data)?
- ☐ Are there more missing data than you expected?
- ☐ Do you suspect a systematic process that explains the missing data?

Check your textual data

- ☐ Is your string variable consistent? E.g. if you have postcodes do they all have a space between the outcode and incode and upper case? (“EC2A 4JE”)
- ☐ Is the length of variables consistent, e.g. does *year* always have four digits?
- ☐ Do you know the minimum length of your string variables?
- ☐ And do you know the maximum length of your string variables?
- ☐ Are there any unusual symbols or abbreviations?
- ☐ Are you aware of the steps you have to do for cleaning the data?
- ☐ Have you thought of the context?

Check your categorical data

- ☐ Did you create a table of counts (frequency table) for categorical variables?
- ☐ If you have groups, have you compared them?

Check your numerical data

- ☐ Have you calculated any averages, e.g. mean or median?
- ☐ Have you calculated a measure of spread, such as the standard deviation?
- ☐ Do you know the extreme values, small and large?
- ☐ Do you know why there are extreme values?
- ☐ Have you wondered if some of the data is censored and/or truncated?
- ☐ Did you inspect them visually, e.g. with a histogram?
- ☐ If the distribution is skewed and/or extreme, have you looked at the "non-extreme" range?