**Key R and RStudio concepts for beginners to understand**

1. package
2. repository
3. library (system and user)
4. script
5. IDE
6. An update to Rstudio does not update R (nor visa versa)
7. An update to R does not update extra packages
8. The need to update regularly
9. library() and require() functions
10. When R changes a second-level version number, the user library needs recreating (e.g. 3.5.3 – 3.6.0)

**Data science**

From: Garrett Grolemund & Hadley Wickham (2017)



**Workflow when reading, cleaning, and exploring data prior to analysis**

E.g. see: <https://sebastiansauer.github.io/data-cleansing/>

1. Set-up a well-designed directory structure
2. Use best-practice for file naming
3. Read the data - was it imported as expected?
4. Capture any metadata describing the data
5. Identify missing values - keep, impute, or exclude?
6. Check distribution for each variable (graphs)
7. Identify outliers (graphs, tests)
8. Check for overall plausibility, errors (e.g., typos), and expected balance (graphs, tables)
9. Identify highly-correlated variables
10. Identify variables with (nearly) no variance
11. Identify variables with strange names or values - adjust if required
12. Rename some variables and/or values, if required
13. Check variable classes (e.g., characters versus factors) - adjust if required
14. Remove/transform some variables (maybe your intended model does not like categorial variables)
15. Check some overall pattern (statistical/numerical summaries and graphs)
16. Centre and/or scale variables (if required for analysis)
17. Save copies of the cleaned data
18. Archive raw data, cleaned data, and cleaning script file