Day 5 Cheatsheet

Data Cleaning

Major concepts

- **Most important rule of data handling - Always be looking at your data!**
- `NA` - general missing data
- `NaN` - stands for “Not a Number”, happens when you do 0/0.
- `Inf` and `-Inf` - Infinity, happens when you take a positive number (or negative number) by 0.

Functions

Library/Package	Piece of code	Example of usage	What it does
Base `R`	`is.na(x)`	`is.na(x)`	checks if `x` is `NA`.
Base `R`	`is.nan(x)`	`is.nan(x)`	checks if `x` is `NaN`.
Base `R`	`is.infinite(x)`	`is.infinite(x)`	checks if `x` is `Inf` or `-Inf`.
`naniar`	`pct_complete(x)`	`pct_complete(x)`	Reports the percentage of data that is complete in `x`.
`naniar`	`gg_miss_var(x)`	`gg_miss_var(x)`	Reports as a plot the percentage of data that is complete in `x`.
`tidyr`	`drop_na(df)`	`drop_na(df)`	Drops rows of `NA` from a given data frame/tibble
`dplyr`	`case_when()`	`df <- arrange(df, mpg)`	This function allows you to vectorise multiple `if_else()` statements. If no cases match, NA is returned.
`dplyr`	`mutate()`	`df <- mutate(df, newcol = wt/2.2)`	Adds a new column that is a function of existing columns
`dplyr`	`separate()`	`df %>% separate(x, c("A", "B"))`	Separate a character column into multiple columns with a regular expression or numeric locations
`dplyr`	`unite()`	`df %>% unite("z", x:y, remove = FALSE)`	Unite multiple columns together into one column
`stringr`	`str_detect`	`df %>% filter(str_detect(col_name, "string_pattern"))`	Returns logical vector to indicate if string pattern was detected
`stringr`	`str_replace`	`str_replace(vector), "replace_me","with_me")`	Replaces all instances of one specified string with another specified string

* This format was adapted from the cheatsheet format from AlexsLemonade.