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	<pre>is.infinite(x)</pre>	<pre>is.infinite(x)</pre>	checks if x is Inf or -Inf.
naniar	<pre>pct_complete(x)</pre>	<pre>pct_complete(x)</pre>	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	<pre>drop_na(df)</pre>	<pre>drop_na(df)</pre>	Drops rows of NA from a given data frame/tibble
dplyr	case_when()	<pre>df <- arrange(df, mpg)</pre>	This function allows you to vectorise multiple if_else() statements. If no cases match, NA is returned.
dplyr	mutate()	<pre>df <- mutate(df, newcol = wt/2.2)</pre>	Adds a new column that is a function of existing columns
dplyr	separate()	<pre>df %>% separate(x, c("A", "B"))</pre>	Separate a character column into multiple columns with a regular expression or numeric locations
dplyr	unite()	<pre>df %>% unite("z", x:y, remove = FALSE)</pre>	Unite multiple columns together into one column
stringr	str_detect	<pre>df %>% filter(str_detect(col_na "string_pattern"))</pre>	Returns logical vector mego indicate if string pattern was detected
stringr	str_replace	<pre>str_replace(vector), "replace_me","with_me")</pre>	Replaces all instances of one specified string with another specified string

^{*} This format was adapted from the cheatsheet format from AlexsLemonade.