Day 4 Cheatsheet

Data Summarization

Functions

Library/Package	Piece of code	Example of usage	What it does
Base R	min(x)	min(x)	Returns the minimum value of all values in an object x.
Base R	sum(x)	sum(x)	Returns the sum of all values (values must be integer, numeric, or logical) in object x.
Base R	mean(x)	mean(x)	Returns the arithmetic mean of all values (values must be integer or numeric) in object x or logical vector x.
Base R	log(x)	log(x)	Gives the natural logarithm of object x. log2(x) can be used to give the logarithm of the object in base 2. Or the base can be specified as an argument.
Base R	range(x)	range(x)	Gives the min and max for object x.
Base R	sd(x)	sd(x)	Gives the standard deviation for object x.
Base R	sqrt(x)	sqrt(x)	Gives the square root for object x.
Base R	quantile(x)	<pre>quantile(x, probs = .5)</pre>	Produces sample quantiles corresponding to the given probabilities x.
Base R	<pre>summary(x)</pre>	summary(x)	Returns a summary of the values in object x.
Base R	rowSums()	rowSums(df)	Calculates sums for each row
Base R	colSums()	colSums(df)	Calculates sums for each column
Base R	rowMeans()	rowMeans(df)	Calculates means for each row
Base R	colMeans()	colMeans(df)	Calculates means for each column

Library/Package	Piece of code	Example of usage	What it does
dplyr	summarize()	<pre>df <- df %>% summarize(mean_x = mean(x))</pre>	Summarizes multiple values in an object into a single value. This function can be used with other functions to retrieve a single output value for the grouped values. summarize and summarise are synonyms in this package. However, note that this function does not work in the same manner as the base R summary function.
dplyr	across()	<pre>df %>% summarize(across(c('col_a', 'col_b'), ~ sum(.x)))</pre>	Use the across function with summarize to summarize across multiple columns of your data.
Base R	unique()	unique(df)	Returns a vector, data frame or array like x but with duplicate elements/rows removed.
Base R	table()	table(x)	Builds a contingency table of the counts at each combination of factor levels.
dplyr	count()	<pre>df %>% count(factor_name)</pre>	Count the number of groups in a factor variable of a data frame or tibble
dplyr	group_by()	<pre>df %>% count(factor_name)</pre>	Groups data into rows that contain the same specified value(s)
dplyr	ungroup()	<pre>df %>% count(factor_name)</pre>	Undo a grouping that was done by
Base R	plot()	plot(x, y)	dplyr::group_by() Creates a scatterplot of x and y vector data
Base R	<pre>boxplot()</pre>	<pre>boxplot(x, y)</pre>	Creates a boxplot of y against levels of x
Base R	hist()	hist(x)	Creates a histogram of x
Base R	density()	<pre>plot(density(x))</pre>	Creates a kernel density plot of x when used with plot()

Data Classes

Major concepts

- Character strings or individual characters, quoted
- Numeric any real number(s)
- Double a special subset of numeric that contains fractional values.
- Integer any integer(s)/whole numbers
- Factor categorical/qualitative variables
- Logical variables composed of TRUE or FALSE
- Date/POSIXct represents calendar dates and times
- matrix Two-dimensional class of data where all rows and columns consist of the same data type.
- data frame Two-dimensional class of data where all columns can be of different data types.
- list Can be of varying dimensions and can hold any kind of data type. Can hold vectors, strings, matrices, models, list of other lists.

Functions

Library/Package	Piece of code	Example of usage	What it does
Base R	<pre>factor(x) or as.factor(x)</pre>	Factor	Coerces object x into a factor (which is used to represent categorical data). This function can be used to coerce object x into other data types, i.e., as.character, as.numeric, as.data.frame, as.matrix, as.Date etc.
Base R	levels(x)	<pre>levels(factor_obj)</pre>	Returns or sets the value of the levels in an object x.
Base R	rep()	rep(1:3)	Replicates the values in x to make a vector.
Base R	seq()	<pre>seq(from = 0, to = 1, by = 0.2)</pre>	Creates a vector of a sequence of numbers based on the specified arguments.

[•] lubridate is a powerful, widely used R package from "tidyverse" family to work with Date / POSIXct class objects

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