

Day 4 Cheatsheet

Data Summarization

Functions

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

Library/Package	Piece of code	Example of usage	What it does
dplyr	summarize()	df <- df %>% summarize(mean_x = mean(x))	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()	df %>% summarize(across(c('col_a', 'col_b'), ~ sum(.x)))	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()	df %>% count(factor_name)	Count the number of groups in a factor variable of a data frame or tibble
dplyr	group_by()	df %>% count(factor_name)	Groups data into rows that contain the same specified value(s)
dplyr	ungroup()	df %>% count(factor_name)	Undo a grouping that was done by group_by()
Base R	plot()	plot(x, y)	Creates a scatterplot of x and y vector data
Base R	boxplot()	boxplot(x, y)	Creates a boxplot of y against levels of x
Base R	hist()	hist(x)	Creates a histogram of x
Base R	density()	plot(density(x))	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	<code>factor(x)</code> or <code>as.factor(x)</code>	Factor	Coerces object <code>x</code> into a factor (which is used to represent categorical data). This function can be used to coerce object <code>x</code> into other data types, i.e., <code>as.character</code> , <code>as.numeric</code> , <code>as.data.frame</code> , <code>as.matrix</code> , <code>as.Date</code> etc.
Base R	<code>levels(x)</code>	<code>levels(factor_obj)</code>	Returns or sets the value of the levels in an object <code>x</code> .
Base R	<code>rep()</code>	<code>rep(1:3)</code>	Replicates the values in <code>x</code> to make a vector.
Base R	<code>seq()</code>	<code>seq(from = 0, to = 1, by = 0.2)</code>	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.