Functions by chapter

Mark Huber

Chapter 1: Introduction to Data Science

Name	Package	Input	Output	Description
head	utils	Various objects	Various	Returns the first part of the object.
summary	base	Various objects	Statistics	Generic base R function that is used to produce a summary of various types of data. When applied to numerical vectors, tries to find minimum, first quantile, median, mean, third quantile, and maximum of every numerical variable, and for categorical data, returns counts of the different levels.
plot	graphics	Two numerical vectors	Graphical Object	Scatterplot formed from input.

Chapter 2: R and R Studio

Name	Package	Input	Output	Description
write_lines	readr	vector of strings	text file	Writes out a vector of strings to a text file.

Chapter 3: The Tidyverse

The first set of functions in this chapter are *transforming* functions. They take as input a tibble (or data frame) and return a tibble.

Name	Package	Input	Output	Description
slice	dplyr	tibble	tibble	Keeps a fixed set of rows (observations) from a tibble.
select	dplyr	tibble	tibble	Keeps a subset of columns (variables) from a tibble.
mutate	dplyr	tibble	tibble	Adds or replaces a column with data formed as a function of other columns.
summarize	dplyr	tibble	tibble	Creates a summary tibble where entries are found via vector functions applied to columns of data.

The next set of functions are *statistics*, functions that take a vector of arbitrary length of numbers, and run a function on those numbers. Note that packages base, stats, and utils are built in to R, you never have to load those manually.

Name	Package	Input	Output	Description
mean	base	numerical vector	number	Finds the sample mean.
median	stats	numerical vector	number	Finds the sample median.
sum	base	numerical vector	number	Adds the numbers together.
max	base	numerical vector	number	Finds the largest value.
min	base	numerical vector	number	Finds the smallest value.

Finally in this chapter is the tibble function from the package of the same name that takes a set of vectors of equal length, and returns a tibble whose columns come from the input vectors.

Name	Package	Input	Output	Description
tibble	tibble	one or more vectors	tibble	Creates a tibble using columns from the input vectors.

Chapter 4: Importing Data

The first group of functions read in a file to a tibble, or write out a tibble to a file.

Name	Package	Input	Output	Description
read_csv	readr	file	tibble	Reads in a comma separated value file.
read_xlsx	readxl	file	tibble	Reads in an Excel file.
read_dta	haven	file	tibble	Reads in a Stata file.
read_rds	readr	file	tibble	Reads in an R file.
write_rds	readr	tibble	file	Writes a tibble to an R file.

Next comes a helpful function that *concatenates* two strings, that is, it places the second string at the end of the first string to make a new string.

Name	Package	Input	Output	Description
str_c	stringr	two strings	one string	Places second string at the end of the first string to make a new string.

Finally comes a function for downloading a file from a website.

Name	Package	Input	Output	Description
download.file	utils	URL	file	Downloads a file from the web.

Chapter 5: Transforming Data

First comes two more functions for transforming tibbles.

Name	Package	Input	Output	Description
filter	dplyr	tibble	tibble	Keeps rows that meet a condition.
arrange	dplyr	tibble	tibble	Sorts the rows in ascending order by one or more columns.

Next is a function for dealing with floating point arithmetic when checking if numbers are equal.

Name	Package	Input	Output	Description
near	dplyr	two numbers	one boolean	Checks if the two numbers are the same within machine epsilon.

Finally, a *helper function* for arrange.

Name	Package	Input	Output	Description
desc	dplyr	a vector	a vector	Transforms a vector so that it will be sorted in descending order.

Chapter 6: Graphical Grammars

The grammar of graphics system used by the package ggplot2 creates a graphical object built out of *layers*. These layers are connected together by a + symbol.

The heart of a ggplot is created by the *geometry* functions, that creates layers based on the data.

Name	Package	Input	Output	Description
ggplot	ggplot2	data and mapping	layer	Creates initial layer.
geom_point	ggplot2	data and mapping	layer	Creates scatterplot layer.
geom_bar	ggplot2	data and mapping	layer	Creates bar plot layer (with stat either ‘count’ or ‘identity’)
geom_smooth	ggplot2	data and mapping	layer	Creates line or curve estimate of point cloud layer.

Name	Package	Input	Output	Description
geom_boxplot	ggplot2	data and mapping	layer	Creates boxplot layer.

The function aes is used to create a mapping needed for a geometry.

Name	Package	Input	Output	Description
aes	ggplot2	vectors	mapping	Creates a mapping of the vectors for a geometry.

Coordinate transforms change the way coordinates are displayed on the graphic. The function aes is used to create a mapping needed for a geometry.

Name	Package	Input	Output	Description
coord_flip	ggplot2	none	layer	Swaps horizontal and vertical axis.
coord_quickmap	ggplot2	none	layer	Turns latitude and longitude into Cartesian coordinates.

Themes change the overall look of the graphic.

Name	Package	Input	Output	Description
theme_minimal	ggplot2	none	layer	Removes background grey squares.

Facet functions split the single graphics into multiple smaller graphics based on a variable.

Name	Package	Input	Output	Description
facet_wrap	ggplot2	model	layer	Breaks graph into multiple graphs based on predictor variable.

The last function does not create a graphical layer. Instead, it reorders a vector of data based on another vector. This is useful in making bar plots whose bars go from longest to shortest. Facet functions split the single graphics into multiple smaller graphics based on a variable.

Name	Package	Input	Output	Description
reorder	stats	categorical vector and sortable vector	categorical vector	Reorders the categorical vector based on the values of the second, sortable vector.

Chapter 7: More Graphical Grammars

Begin with two more geometries, a coordinate transform, and two more themes.

Name	Package	Input	Output	Description

Name	Package	Input	Output	Description
geom_histogram	ggplot2	data and mapping	layer	Creates a histogram layer.
geom_area	ggplot2	data and mapping	layer	Creates an area graph layer.
coord_polar	ggplot2	none	layer	Transforms from Cartesian to polar coordinates.
theme_void	ggplot2	none	layer	Removes all axes and lables.
theme_classic	ggplot2	none	layer	Delivers a classic look.

Next come four functions for making the graph look prettier.

Name	Package	Input	Output	Description
labs	ggplot2	labels to change	layer	Changes various labels in the graph.
xlab	ggplot2	x axis label	layer	Changes the x axis label.
ylab	ggplot2	y axis label	layer	Changes the y axis label.
scale_fill_manual	ggplot2	color mapping	layer	Changes how values get turned into fill colors.

The next function is specifically for creating correlogram graphs from a correlation matrix.

Name	Package	Input	Output	Description
cor	stats	numerical vectors	correlation matrix	Estimates correlations for multiple vectors of data.
ggcorrplot	ggcorrplot	correlation matrix	graphical object	Creates a correlogram from a correlation matrix.

The final two functions are useful for manipulating tibbles with mutate. The first pulls out the row names of the tibble as a vector. The second takes a boolean vector, and returns the yes value for each TRUE entry, and the no value for each FALSE entry.

Name	Package	Input	Output	Description
rownames	tibble	tibble	vector	Pulls out the row names of tibble as a vector of strings.

Name	Package	Input	Output	Description
ifelse	base	boolean vector, yes value, no value	vector	Returns the yes value for the TRUE entries, and no value for the FALSE entries in the boolean vector.

Chapter 8: Grouping observations

The main function of this chapter is group_by, which breaks the observations into groups, where functions like summarize, filter, mutate, and slice can operate on each group.

Name	Package	Input	Output	Description
group_by	dplyr	tibble	tibble with groups	Breaks observations into groups so that other functions such as summarize, filter, mutate, and slice operate on each group separately.

The rank is a useful function that tells the rank of components of a sortable vector.

Name	Package	Input	Output	Description
rank	base	sortable vector	rank vector	Returns the ranks of all the items in the sortable vector.

Chapter 9: Combining datasets as sets

There are three functions that take as input two tibbles that must have the same column names in the same order. The output is a tibble that contains observations from one or the other or both input tibbles depending on the function.

Name	Package	Input	Output	Description
union	dplyr	two tibbles	tibble	Returns all observations in at least one of the tables.
intersect	dplyr	two tibbles	tibble	Returns all observations in both tables
setdiff	dplyr	two tibbles	tibble	Returns observations in the first table but not the second