| Please choose a lesson, or type 0 to return to course menu.

1: Basic Building Blocks 2: Workspace and Files 3: Sequences of Numbers

4: Vectors 5: Missing Values 6: Subsetting Vectors

7: Matrices and Data Frames 8: Logic 9: Functions

10: lapply and sapply 11: vapply and tapply 12: Looking at Data

13: Simulation 14: Dates and Times 15: Base Graphics

Selection: 12

| | 0%

| Whenever you're working with a new dataset, the first thing you should do is look at it! What is

| the format of the data? What are the dimensions? What are the variable names? How are the variables

| stored? Are there missing data? Are there any flaws in the data?

...

|==== | 4%

| This lesson will teach you how to answer these questions and more using R's built-in functions.

| We'll be using a dataset constructed from the United States Department of Agriculture's PLANTS

| Database (http://plants.usda.gov/adv\_search.html).

...

|======== | 8%

| I've stored the data for you in a variable called plants. Type ls() to list the variables in your

| workspace, among which should be plants.

> ls()

[1] "flags" "my\_seq" "ok" "old.dir" "plants" "viewinfo" "x"

| Excellent work!

|============ | 12%

| Let's begin by checking the class of the plants variable with class(plants). This will give us a

| clue as to the overall structure of the data.

> class(plants)

[1] "data.frame"

| Your dedication is inspiring!

|=============== | 17%

| It's very common for data to be stored in a data frame. It is the default class for data read into

| R using functions like read.csv() and read.table(), which you'll learn about in another lesson.

...

|=================== | 21%

| Since the dataset is stored in a data frame, we know it is rectangular. In other words, it has two

| dimensions (rows and columns) and fits neatly into a table or spreadsheet. Use dim(plants) to see

| exactly how many rows and columns we're dealing with.

> dim(plants)

[1] 5166 10

| Keep up the great work!

|======================= | 25%

| The first number you see (5166) is the number of rows (observations) and the second number (10) is

| the number of columns (variables).

...

|=========================== | 29%

| You can also use nrow(plants) to see only the number of rows. Try it out.

> nrow(plants)

[1] 5166

| You are doing so well!

|=============================== | 33%

| ... And ncol(plants) to see only the number of columns.

> ncol(plants)

[1] 10

| Perseverance, that's the answer.

|================================== | 38%

| If you are curious as to how much space the dataset is occupying in memory, you can use

| object.size(plants).

> object.size(plants)

644232 bytes

| Excellent job!

|====================================== | 42%

| Now that we have a sense of the shape and size of the dataset, let's get a feel for what's inside.

| names(plants) will return a character vector of column (i.e. variable) names. Give it a shot.

> names(plants)

[1] "Scientific\_Name" "Duration" "Active\_Growth\_Period" "Foliage\_Color"

[5] "pH\_Min" "pH\_Max" "Precip\_Min" "Precip\_Max"

[9] "Shade\_Tolerance" "Temp\_Min\_F"

| You got it right!

|========================================== | 46%

| We've applied fairly descriptive variable names to this dataset, but that won't always be the case.

| A logical next step is to peek at the actual data. However, our dataset contains over 5000

| observations (rows), so it's impractical to view the whole thing all at once.

...

|============================================== | 50%

| The head() function allows you to preview the top of the dataset. Give it a try with only one

| argument.

> head(plants)

Scientific\_Name Duration Active\_Growth\_Period Foliage\_Color pH\_Min pH\_Max

1 Abelmoschus <NA> <NA> <NA> NA NA

2 Abelmoschus esculentus Annual, Perennial <NA> <NA> NA NA

3 Abies <NA> <NA> <NA> NA NA

4 Abies balsamea Perennial Spring and Summer Green 4 6

5 Abies balsamea var. balsamea Perennial <NA> <NA> NA NA

6 Abutilon <NA> <NA> <NA> NA NA

Precip\_Min Precip\_Max Shade\_Tolerance Temp\_Min\_F

1 NA NA <NA> NA

2 NA NA <NA> NA

3 NA NA <NA> NA

4 13 60 Tolerant -43

5 NA NA <NA> NA

6 NA NA <NA> NA

| That's a job well done!

|================================================== | 54%

| Take a minute to look through and understand the output above. Each row is labeled with the

| observation number and each column with the variable name. Your screen is probably not wide enough

| to view all 10 columns side-by-side, in which case R displays as many columns as it can on each

| line before continuing on the next.

...

|====================================================== | 58%

| By default, head() shows you the first six rows of the data. You can alter this behavior by passing

| as a second argument the number of rows you'd like to view. Use head() to preview the first 10 rows

| of plants.

> head(plants, 10)

Scientific\_Name Duration Active\_Growth\_Period Foliage\_Color pH\_Min

1 Abelmoschus <NA> <NA> <NA> NA

2 Abelmoschus esculentus Annual, Perennial <NA> <NA> NA

3 Abies <NA> <NA> <NA> NA

4 Abies balsamea Perennial Spring and Summer Green 4

5 Abies balsamea var. balsamea Perennial <NA> <NA> NA

6 Abutilon <NA> <NA> <NA> NA

7 Abutilon theophrasti Annual <NA> <NA> NA

8 Acacia <NA> <NA> <NA> NA

9 Acacia constricta Perennial Spring and Summer Green 7

10 Acacia constricta var. constricta Perennial <NA> <NA> NA

pH\_Max Precip\_Min Precip\_Max Shade\_Tolerance Temp\_Min\_F

1 NA NA NA <NA> NA

2 NA NA NA <NA> NA

3 NA NA NA <NA> NA

4 6.0 13 60 Tolerant -43

5 NA NA NA <NA> NA

6 NA NA NA <NA> NA

7 NA NA NA <NA> NA

8 NA NA NA <NA> NA

9 8.5 4 20 Intolerant -13

10 NA NA NA <NA> NA

| Perseverance, that's the answer.

|========================================================== | 62%

| The same applies for using tail() to preview the end of the dataset. Use tail() to view the last 15

| rows.

> tail(plants, 15)

Scientific\_Name Duration Active\_Growth\_Period Foliage\_Color pH\_Min pH\_Max

5152 Zizania <NA> <NA> <NA> NA NA

5153 Zizania aquatica Annual Spring Green 6.4 7.4

5154 Zizania aquatica var. aquatica Annual <NA> <NA> NA NA

5155 Zizania palustris Annual <NA> <NA> NA NA

5156 Zizania palustris var. palustris Annual <NA> <NA> NA NA

5157 Zizaniopsis <NA> <NA> <NA> NA NA

5158 Zizaniopsis miliacea Perennial Spring and Summer Green 4.3 9.0

5159 Zizia <NA> <NA> <NA> NA NA

5160 Zizia aptera Perennial <NA> <NA> NA NA

5161 Zizia aurea Perennial <NA> <NA> NA NA

5162 Zizia trifoliata Perennial <NA> <NA> NA NA

5163 Zostera <NA> <NA> <NA> NA NA

5164 Zostera marina Perennial <NA> <NA> NA NA

5165 Zoysia <NA> <NA> <NA> NA NA

5166 Zoysia japonica Perennial <NA> <NA> NA NA

Precip\_Min Precip\_Max Shade\_Tolerance Temp\_Min\_F

5152 NA NA <NA> NA

5153 30 50 Intolerant 32

5154 NA NA <NA> NA

5155 NA NA <NA> NA

5156 NA NA <NA> NA

5157 NA NA <NA> NA

5158 35 70 Intolerant 12

5159 NA NA <NA> NA

5160 NA NA <NA> NA

5161 NA NA <NA> NA

5162 NA NA <NA> NA

5163 NA NA <NA> NA

5164 NA NA <NA> NA

5165 NA NA <NA> NA

5166 NA NA <NA> NA

| Your dedication is inspiring!

|============================================================= | 67%

| After previewing the top and bottom of the data, you probably noticed lots of NAs, which are R's

| placeholders for missing values. Use summary(plants) to get a better feel for how each variable is

| distributed and how much of the dataset is missing.

> summary(plants)

Scientific\_Name Duration Active\_Growth\_Period

Abelmoschus : 1 Perennial :3031 Spring and Summer : 447

Abelmoschus esculentus : 1 Annual : 682 Spring : 144

Abies : 1 Annual, Perennial: 179 Spring, Summer, Fall: 95

Abies balsamea : 1 Annual, Biennial : 95 Summer : 92

Abies balsamea var. balsamea: 1 Biennial : 57 Summer and Fall : 24

Abutilon : 1 (Other) : 92 (Other) : 30

(Other) :5160 NA's :1030 NA's :4334

Foliage\_Color pH\_Min pH\_Max Precip\_Min Precip\_Max

Dark Green : 82 Min. :3.000 Min. : 5.100 Min. : 4.00 Min. : 16.00

Gray-Green : 25 1st Qu.:4.500 1st Qu.: 7.000 1st Qu.:16.75 1st Qu.: 55.00

Green : 692 Median :5.000 Median : 7.300 Median :28.00 Median : 60.00

Red : 4 Mean :4.997 Mean : 7.344 Mean :25.57 Mean : 58.73

White-Gray : 9 3rd Qu.:5.500 3rd Qu.: 7.800 3rd Qu.:32.00 3rd Qu.: 60.00

Yellow-Green: 20 Max. :7.000 Max. :10.000 Max. :60.00 Max. :200.00

NA's :4334 NA's :4327 NA's :4327 NA's :4338 NA's :4338

Shade\_Tolerance Temp\_Min\_F

Intermediate: 242 Min. :-79.00

Intolerant : 349 1st Qu.:-38.00

Tolerant : 246 Median :-33.00

NA's :4329 Mean :-22.53

3rd Qu.:-18.00

Max. : 52.00

NA's :4328

| Perseverance, that's the answer.

|================================================================= | 71%

| summary() provides different output for each variable, depending on its class. For numeric data

| such as Precip\_Min, summary() displays the minimum, 1st quartile, median, mean, 3rd quartile, and

| maximum. These values help us understand how the data are distributed.

...

|===================================================================== | 75%

| For categorical variables (called 'factor' variables in R), summary() displays the number of times

| each value (or 'level') occurs in the data. For example, each value of Scientific\_Name only appears

| once, since it is unique to a specific plant. In contrast, the summary for Duration (also a factor

| variable) tells us that our dataset contains 3031 Perennial plants, 682 Annual plants, etc.

...

|========================================================================= | 79%

| You can see that R truncated the summary for Active\_Growth\_Period by including a catch-all category

| called 'Other'. Since it is a categorical/factor variable, we can see how many times each value

| actually occurs in the data with table(plants$Active\_Growth\_Period).

> table(plants$Active\_Growth\_Period)

Fall, Winter and Spring Spring Spring and Fall Spring and Summer

15 144 10 447

Spring, Summer, Fall Summer Summer and Fall Year Round

95 92 24 5

| You got it!

|============================================================================= | 83%

| Each of the functions we've introduced so far has its place in helping you to better understand the

| structure of your data. However, we've left the best for last....

...

|================================================================================ | 88%

| Perhaps the most useful and concise function for understanding the \*str\*ucture of your data is str().

| Give it a try now.

> str(plants)

'data.frame': 5166 obs. of 10 variables:

$ Scientific\_Name : Factor w/ 5166 levels "Abelmoschus",..: 1 2 3 4 5 6 7 8 9 10 ...

$ Duration : Factor w/ 8 levels "Annual","Annual, Biennial",..: NA 4 NA 7 7 NA 1 NA 7 7 ...

$ Active\_Growth\_Period: Factor w/ 8 levels "Fall, Winter and Spring",..: NA NA NA 4 NA NA NA NA 4 NA ...

$ Foliage\_Color : Factor w/ 6 levels "Dark Green","Gray-Green",..: NA NA NA 3 NA NA NA NA 3 NA ...

$ pH\_Min : num NA NA NA 4 NA NA NA NA 7 NA ...

$ pH\_Max : num NA NA NA 6 NA NA NA NA 8.5 NA ...

$ Precip\_Min : int NA NA NA 13 NA NA NA NA 4 NA ...

$ Precip\_Max : int NA NA NA 60 NA NA NA NA 20 NA ...

$ Shade\_Tolerance : Factor w/ 3 levels "Intermediate",..: NA NA NA 3 NA NA NA NA 2 NA ...

$ Temp\_Min\_F : int NA NA NA -43 NA NA NA NA -13 NA ...

| You got it right!

|==================================================================================== | 92%

| The beauty of str() is that it combines many of the features of the other functions you've already

| seen, all in a concise and readable format. At the very top, it tells us that the class of plants is

| 'data.frame' and that it has 5166 observations and 10 variables. It then gives us the name and class

| of each variable, as well as a preview of its contents.

...

|======================================================================================== | 96%

| str() is actually a very general function that you can use on most objects in R. Any time you want to

| understand the structure of something (a dataset, function, etc.), str() is a good place to start.

...

|============================================================================================| 100%

| In this lesson, you learned how to get a feel for the structure and contents of a new dataset using a

| collection of simple and useful functions. Taking the time to do this upfront can save you time and

| frustration later on in your analysis.

...