Familiarizandose con los datos

David H. Duncan

January 20, 2016

Cada vez que usted está trabajando con un nuevo conjunto de datos, lo primero que debe hacer es mirarlo! ¿Cuál es el formato de los datos? ¿Cuáles son las dimensiones? ¿Cuáles son los nombres de las variables? ¿Cómo se almacenan las variables? ¿Existen datos que faltan? ¿Hay errores en los datos?

Esta lección le enseñará cómo responder a estas preguntas y más sobre uso funciones incorporadas de R. Vamos a estar usando un conjunto de datos construido a partir de la base de datos del Departamento de Plantas Agrícolas de los Estados Unidos (http://plants.usda.gov/adv_search.html).

Copie y pegue el siguiente en la consola:

```
.datapath <- file.path(path.package('swirl'), 'Courses',
                       'ConoceR', 'Familiarizandose_con_los_datos',
                       'plant-data.txt')
# Leer los datos
plants <- read.csv(.datapath, strip.white=TRUE, na.strings="")
```

Ha almacenado los datos en una variable llamada plants. Escriba ls() para listar las variables en el espacio de trabajo, entre los cuales debe aparecer plants.

```
ls()
```

```
## [1] "plants"
```

A continuación, copie y pegue este paso que realice unos cambios para hacer el conjunto más amigable

```
# Remove annoying columns
.cols2rm <- c('Accepted.Symbol', 'Synonym.Symbol')
plants <- plants[, !(names(plants) %in% .cols2rm)]

# Make names pretty
names(plants) <- c('Nombre_científico', 'Duración', 'Periodo_crecimiento',
                   'Color_follaje', 'pH_Mín', 'pH_Max',
                   'Precip_Mín', 'Precip_Max',
                   'Tolerancia_de_sombra', 'Temp_Mín_F')
```

Empecemos por por verificar la clase de la variable plants con class(plants). Esto nos dará una idea de la estructura general de los datos.

```
class(plants)
```

```
## [1] "data.frame"
```

Es muy común que los datos se almacenen en una trama de datos (data.frame). Esta es la clase predeterminada para los datos leídos en R utilizando funciones como read.csv() y read.table(), sobre las cuales aprenderemos en otra lección.

Dado que el conjunto de datos se almacena en una trama de datos, sabemos que debe ser es rectangular. En otras palabras, tiene dos dimensiones (filas y columnas) y encaja perfectamente en una tabla u hoja de cálculo. Utilice dim(plants) para ver exactamente con cuántas filas y columnas estamos tratando.

```
dim(plants)
```

```
## [1] 5166   10
```

El primer número que se ve (5166) es el número de filas (observaciones) y el segundo número (10) es el número de columnas (variables).

También puede utilizar nrow(plants) para ver sólo el número de filas. El comando Nrow quiere decir n de número y row = fila. Inténtalo.

```
nrow(plants)
```

```
## [1] 5166
```

. . . Y ncol(plants) para ver sólo el número de columnas.

```
ncol(plants)
```

```
## [1] 10
```

Ahora que sabemos forma y tamaño del conjunto de datos, vamos a tener una idea de lo que hay dentro. names(plants) devolverá un vector de caracteres con los nombres de las columnas (es decir, variables). Intentalo.

```
names(plants)
```

```
##  [1] "Nombre_científico"    "Duración"             "Periodo_crecimiento"
##  [4] "Color_follaje"        "pH_Mín"               "pH_Max"
##  [7] "Precip_Mín"           "Precip_Max"           "Tolerancia_de_sombra"
## [10] "Temp_Mín_F"
```

Ahora, ya que esté usted trabajando en RStudio ahora mismo, le comento que también se puede ver muchos de estos detalles en la ventanilla de ENVIRONMENT (entorno). Allí, justo a la izquierda del objeto de datos plants verá usted una flechita en un cículo azul. Púlselo para que revele su detalle.

¿Ve usted las mismas dimensiones de esta trama de datos plants, y los nombres y las clases de variables que contiene? Este ejemplo es un poco feo porque tiene muchos valores perdidos representados por 'NA'. Parece a las letras de una canción de pop, ¿no?

Hemos aplicado unos nombres de variables bastante descriptivos para este conjunto de datos, pero no siempre será así. Un siguiente paso lógico es dar un vistazo a los datos reales. Sin embargo, nuestra base de datos contiene más de 5.000 observaciones (filas), así que es poco práctico ver toda la tabla a la vez.

La función head() le permite hacer una vista previa de la parte superior del conjunto de datos. Dese la oportunidad con un solo argumento.

```
head(plants)
```

```
##              Nombre_científico          Duración Periodo_crecimiento
## 1                   Abelmoschus              <NA>                <NA>
## 2       Abelmoschus esculentus Annual, Perennial                <NA>
## 3                         Abies              <NA>                <NA>
## 4                Abies balsamea         Perennial   Spring and Summer
## 5 Abies balsamea var. balsamea         Perennial                <NA>
## 6                      Abutilon              <NA>                <NA>
##   Color_follaje pH_Mín pH_Max Precip_Mín Precip_Max Tolerancia_de_sombra
## 1          <NA>     NA     NA         NA         NA                 <NA>
## 2          <NA>     NA     NA         NA         NA                 <NA>
## 3          <NA>     NA     NA         NA         NA                 <NA>
## 4         Green      4      6         13         60             Tolerant
## 5          <NA>     NA     NA         NA         NA                 <NA>
## 6          <NA>     NA     NA         NA         NA                 <NA>
##   Temp_Mín_F
## 1         NA
## 2         NA
## 3         NA
## 4        -43
## 5         NA
## 6         NA
```

Tome un minuto para mirar el resultado y entender la salida anterior. Cada fila se etiqueta con el número de observación y cada columna con el nombre de la variable. Es probable que su pantalla no sea lo suficientemente amplia como para ver las 10 columnas de lado a lado, en cuyo caso R muestra tantas columnas como pueda en cada línea antes de continuar al siguiente.

Por defecto, la función head() muestra las primeras seis filas de los datos. Puede modificar este comportamiento al pasar como segundo argumento el número de filas que desea ver. Use la head() para obtener una vista previa de las primeras 10 filas de plantas.

```
head(plants, 10)
```

```
##                    Nombre_científico          Duración Periodo_crecimiento
## 1                         Abelmoschus              <NA>                <NA>
## 2             Abelmoschus esculentus Annual, Perennial                <NA>
## 3                               Abies              <NA>                <NA>
## 4                      Abies balsamea         Perennial   Spring and Summer
## 5       Abies balsamea var. balsamea         Perennial                <NA>
## 6                            Abutilon              <NA>                <NA>
## 7                Abutilon theophrasti            Annual                <NA>
## 8                             Acacia              <NA>                <NA>
## 9                   Acacia constricta         Perennial   Spring and Summer
## 10 Acacia constricta var. constricta         Perennial                <NA>
##    Color_follaje pH_Mín pH_Max Precip_Mín Precip_Max Tolerancia_de_sombra
## 1           <NA>     NA     NA         NA         NA                 <NA>
## 2           <NA>     NA     NA         NA         NA                 <NA>
## 3           <NA>     NA     NA         NA         NA                 <NA>
## 4          Green      4    6.0         13         60             Tolerant
## 5           <NA>     NA     NA         NA         NA                 <NA>
## 6           <NA>     NA     NA         NA         NA                 <NA>
```

```
## 7           <NA>     NA     NA         NA         NA                 <NA>
## 8           <NA>     NA     NA         NA         NA                 <NA>
## 9          Green      7    8.5          4         20            Intolerant
## 10          <NA>     NA     NA         NA         NA                 <NA>
##    Temp_Mín_F
## 1          NA
## 2          NA
## 3          NA
## 4         -43
## 5          NA
## 6          NA
## 7          NA
## 8          NA
## 9         -13
## 10         NA
```

Lo mismo se aplica en el uso de la función tail() para obtener una vista previa del final del conjunto de datos. Utilice tail() para ver las últimas 15 filas.

```
tail(plants, 15)
```

```
##                     Nombre_científico  Duración Periodo_crecimiento
## 5152                          Zizania      <NA>                 <NA>
## 5153                 Zizania aquatica    Annual               Spring
## 5154   Zizania aquatica var. aquatica    Annual                 <NA>
## 5155                Zizania palustris    Annual                 <NA>
## 5156 Zizania palustris var. palustris    Annual                 <NA>
## 5157                       Zizaniopsis      <NA>                 <NA>
## 5158            Zizaniopsis miliacea Perennial   Spring and Summer
## 5159                           Zizia      <NA>                 <NA>
## 5160                     Zizia aptera Perennial                 <NA>
## 5161                      Zizia aurea Perennial                 <NA>
## 5162                 Zizia trifoliata Perennial                 <NA>
## 5163                          Zostera      <NA>                 <NA>
## 5164                  Zostera marina Perennial                 <NA>
## 5165                          Zoysia      <NA>                 <NA>
## 5166                  Zoysia japonica Perennial                 <NA>
##      Color_follaje pH_Mín pH_Max Precip_Mín Precip_Max
## 5152          <NA>     NA     NA         NA         NA
## 5153         Green    6.4    7.4         30         50
## 5154          <NA>     NA     NA         NA         NA
## 5155          <NA>     NA     NA         NA         NA
## 5156          <NA>     NA     NA         NA         NA
## 5157          <NA>     NA     NA         NA         NA
## 5158         Green    4.3    9.0         35         70
## 5159          <NA>     NA     NA         NA         NA
## 5160          <NA>     NA     NA         NA         NA
## 5161          <NA>     NA     NA         NA         NA
## 5162          <NA>     NA     NA         NA         NA
## 5163          <NA>     NA     NA         NA         NA
## 5164          <NA>     NA     NA         NA         NA
## 5165          <NA>     NA     NA         NA         NA
## 5166          <NA>     NA     NA         NA         NA
```

```
##      Tolerancia_de_sombra Temp_Mín_F
## 5152                 <NA>         NA
## 5153           Intolerant         32
## 5154                 <NA>         NA
## 5155                 <NA>         NA
## 5156                 <NA>         NA
## 5157                 <NA>         NA
## 5158           Intolerant         12
## 5159                 <NA>         NA
## 5160                 <NA>         NA
## 5161                 <NA>         NA
## 5162                 <NA>         NA
## 5163                 <NA>         NA
## 5164                 <NA>         NA
## 5165                 <NA>         NA
## 5166                 <NA>         NA
```

Después de la vista previa de la parte superior e inferior de los datos, usted probablemente ha notado un montón de NAs, que son los marcadores de posición de R para los valores perdidos. Use summary(plants) para obtener una mejor idea de cómo se distribuye cada variable y que tanto de la base de datos no se encuentra.

```
summary(plants)
```

```
##                     Nombre_científico              Duración
##  Abelmoschus                 :   1    Perennial        :3031
##  Abelmoschus esculentus      :   1    Annual           : 682
##  Abies                       :   1    Annual, Perennial: 179
##  Abies balsamea              :   1    Annual, Biennial :  95
##  Abies balsamea var. balsamea:   1    Biennial         :  57
##  Abutilon                    :   1    (Other)          :  92
##  (Other)                     :5160    NA's             :1030
##           Periodo_crecimiento      Color_follaje      pH_Mín
##  Spring and Summer   : 447     Dark Green  :  82   Min.   :3.000
##  Spring              : 144     Gray-Green  :  25   1st Qu.:4.500
##  Spring, Summer, Fall:  95     Green       : 692   Median :5.000
##  Summer              :  92     Red         :   4   Mean   :4.997
##  Summer and Fall     :  24     White-Gray  :   9   3rd Qu.:5.500
##  (Other)             :  30     Yellow-Green:  20   Max.   :7.000
##  NA's                :4334     NA's        :4334   NA's   :4327
##      pH_Max         Precip_Mín      Precip_Max       Tolerancia_de_sombra
##  Min.   : 5.100   Min.   : 4.00   Min.   : 16.00   Intermediate: 242
##  1st Qu.: 7.000   1st Qu.:16.75   1st Qu.: 55.00   Intolerant  : 349
##  Median : 7.300   Median :28.00   Median : 60.00   Tolerant    : 246
##  Mean   : 7.344   Mean   :25.57   Mean   : 58.73   NA's        :4329
##  3rd Qu.: 7.800   3rd Qu.:32.00   3rd Qu.: 60.00
##  Max.   :10.000   Max.   :60.00   Max.   :200.00
##  NA's   :4327     NA's   :4338    NA's   :4338
##    Temp_Mín_F
##  Min.   :-79.00
##  1st Qu.:-38.00
##  Median :-33.00
##  Mean   :-22.53
```

```
##  3rd Qu.:-18.00
##  Max.   : 52.00
##  NA's   :4328
```

summary() proporciona una salida diferente para cada variable, dependiendo de su clase. Para los datos numéricos como Precip_Mín, summary() muestra el mínimo, primer cuartil, la mediana, la media, el tercer cuartil, y el máximo. Estos valores nos ayudan a entender cómo se distribuyen los datos.

Para las variables categóricas (llamadas variables 'factor' en R), summary() muestra el número de veces que cada valor (o 'nivel') se produce en los datos. Por ejemplo, cada valor de Nombre_cientifico sólo aparece una vez, ya que es única para una planta específica. Por el contrario, el resumen de Duration (también una variable de factor) nos dice que nuestro conjunto de datos contiene 3031 plantas perennes, 682 plantas anuales, etc.

Se puede ver que R trunca el resumen para Periodo_crecimiento incluyendo una nueva categoría denominada 'Other'. Dado que es una variable categórica / Factor, podemos ver cuántas veces cada valor realmente ocurre en los datos con table(plants$Periodo_crecimiento).

```
table(plants$Periodo_crecimiento)
```

```
##
## Fall, Winter and Spring                  Spring         Spring and Fall
##                      15                     144                      10
##       Spring and Summer    Spring, Summer, Fall                  Summer
##                     447                      95                      92
##         Summer and Fall             Year Round
##                      24                       5
```

Cada una de las funciones que hemos introducido hasta el momento tiene su utilidad para ayudar a entender mejor la estructura de los datos. Sin embargo, hemos dejado lo mejor para lo último ...

Quizás la función más útil y concisa para la comprensión de la estructura de sus datos es str() de la palabra E*str*uctura. Dese una oportunidad ahora.

```
str(plants)
```

```
## 'data.frame':    5166 obs. of  10 variables:
##  $ Nombre_científico   : Factor w/ 5166 levels "Abelmoschus",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Duración            : Factor w/ 8 levels "Annual","Annual, Biennial",..: NA 4 NA 7 7 NA 1 NA 7 7
##  $ Periodo_crecimiento : Factor w/ 8 levels "Fall, Winter and Spring",..: NA NA NA 4 NA NA NA NA 4 NA
##  $ Color_follaje       : Factor w/ 6 levels "Dark Green","Gray-Green",..: NA NA NA 3 NA NA NA NA 3 NA
##  $ pH_Mín              : 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_Mín          : int  NA NA NA 13 NA NA NA NA 4 NA ...
##  $ Precip_Max          : int  NA NA NA 60 NA NA NA NA 20 NA ...
##  $ Tolerancia_de_sombra: Factor w/ 3 levels "Intermediate",..: NA NA NA 3 NA NA NA NA 2 NA ...
##  $ Temp_Mín_F          : int  NA NA NA -43 NA NA NA NA -13 NA ...
```

La belleza de str() es que combina muchas de las características de las otras funciones que ya has visto, todo ello en un formato conciso y fácil de leer. En la parte superior, se nos dice que la clase de las plantas es 'data.frame' y que cuenta con 5166 observaciones y 10 variables. A continuación, nos da el nombre y la categoría de cada variable, así como una vista previa de su contenido.

str() es en realidad una función muy general que se puede utilizar en la mayoría de los objetos en R. Cada vez que usted quiere entender la estructura de algo (un conjunto de datos, función, etc.), str() es un buen modo para comenzar.

En esta lección, ha aprendido a tener una idea de la estructura y contenido de un nuevo conjunto de datos utilizando una colección de funciones simples y útiles. Tomarse el tiempo para hacer esto por adelantado puede ahorrarle tiempo y frustración más tarde durante su análisis.

La lección le hubiera enseñado a usted algunas funciones de gran utilidad. Entres todas ellas, dos sobresalten porque muestran muchos detalles importantes de un tramo de datos en una sola función. Antes de seguir, entonces, ¿cuales son?

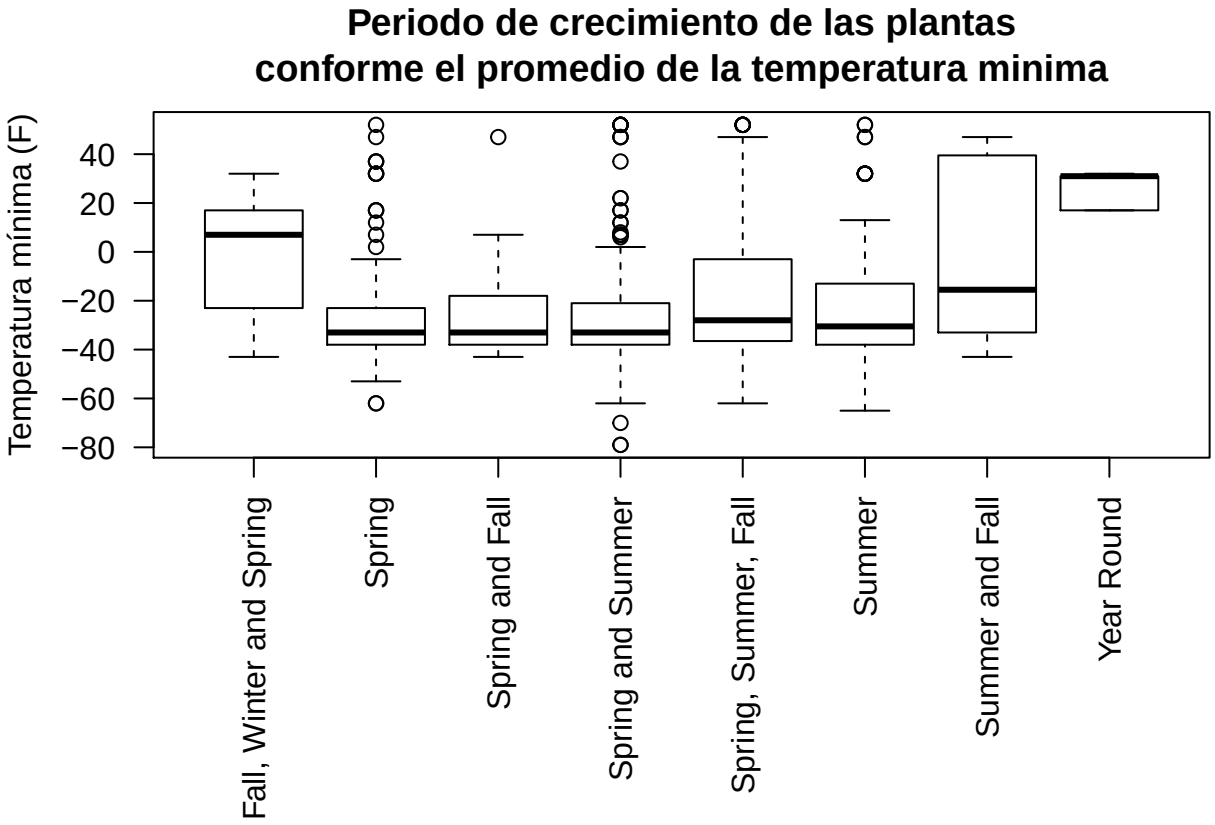
- str() y summary()
- nrow() y tail()
- names() y head()
- tail() y summary()

Con variables cuantitativas la función summary() calcula, como condición base, los límites (inferior y superior), los cuartiles 1 y 3, la mediana y la media. Antes de seguir, entonces, ¿qué calcula para factores (variables cualitativas)? Pista: Usted vió ejemplos hace unos pasos y se los puede revisar por deslizar hacia arriba en la consola.

- Frecuencia de casos en cada nivel del factor
- Lo mismo que calcula para variables cuantitativas
- La significancia de cada factor

Por último, le dejo con una gráfica sencilla para visualizar este conjunto de datos. Debiera haber aparecido en la ventanilla de PLOTs ya. No es nada sensacional, sino un ejemplo de una presentación para echarse un vistazo a la distribución de valores.

```
par(mar=c(10,4,3.5,1))
boxplot(plants$Temp_Mín_F~plants$Periodo_crecimiento,
        las=2, ylab = "Temperatura mínima (F)",
        main="Periodo de crecimiento de las plantas\nconforme el promedio de la temperatura minima")
```



En esta grafica se ve la distribucion de temperaturas minimas para las 5166 especie de plantas, categorizadas por su periodo de crecimiento anual. ¿Que detalle se llama a su atencion de esta grafica? Puede ser de la realizacion de la grafica, o del contenido. Lo que sea, comentelo en nuestras paginas del curso, o tuítearlo con #MOOCUTPLAnalisisDatos!

Felicitaciones, ha terminado usted otra lección. ¡No se olvide de guardar pedazos de código que le parece útil en su álbum de recortes! Hasta la próxima oportunidad, le espero mucho éxito en sus labores.