Conociendo los datos

Carlos Iván Espinosa

7 de octubre de 2016

Contents

Pueden descargar este documento en pdf haciendo clic aquí

Una de las primeras cosas que se debe hacer cuando trabajamos con un conjunto de datos es conocer las características de estos datos. ¿Qué tipo de variables son? ¿Cómo están almacenadas? ¿Cuál es la distribución y sus dimensiones? ¿Existen datos erróneos? ¿Hay datos que faltan?

Usaremos `R` para contestar algunas de estas preguntas, veremos que funciones podemos utilizar y como tenemos que usarlas. Al final de esta lección espero que hayan logrado comprender la naturaleza de los datos. Seguiremos usando los datos de amebiasis de Loja.

Características del objeto

Lo primero que vamos hacer es conocer cuáles son las características del objeto que hemos generado. Lo primero cargar los datos. Vuelva a descargar los datos desde aquí

```
ameLoja<-read.table("AMEBIASIS_LOJA.csv", header=TRUE, sep=';')
```

> ¿Tuvo algún error?

No olvide que debe poner los datos en la carpeta del nuevo proyecto, vuélvalo a intentar. si prefiere no mover los datos a la carpeta de este proyecto no hay problema, copie la dirección completa donde tiene los datos (será algo como esto`C:/Users/UPTL/Documents/GitHub/cargarDatos/AMEBIASIS_LOJA.csv`) y reemplácelo por el nombre del archivo. ¿Funcionó?, recuerde que necesitamos usar el `/` y no `\` que sale por defecto cuando hago copio desde la ventana de windows.

Una vez que tenemos los datos subidos vamos a ver algunas de las características, en la anterior lección ya vimos algunas funciones.

```
class(ameLoja)
```

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

Con la función `class` podemos saber qué tipo de objeto tenemos. En este caso tenemos un data.frame. Los data.frame son matrices que tienen tanto variables cualitativas como cuantitativas, y es el objeto que por defecto se genera cuando utilizo una función como read.table o read_excel.

Ahora que sabemos que estamos trabajando con un data.frame (o trama de datos en español) sabemos que tiene dos dimensiones; las filas y las columnas. Pero ¿cuántas filas y cuántas columnas tienen mi objeto? Podemos utilizar un par de funciones para saberlo.

```
dim(ameLoja)
```

```
## [1] 3019    9
```

```
ncol(ameLoja)
```

```
## [1] 9
```

```
nrow(ameLoja)
```

```
## [1] 3019
```

La función `dim` nos da cuantas filas y columnas tiene el objeto y con las funciones `ncol`y `nrow` podemos ver columnas y filas por separado. Aunque parezca que esto es poco práctico, esta información puede ayudarnos a generar algunos de los análisis.

Ahora que sabemos forma y tamaño del conjunto de datos, vamos a hacernos una idea de lo que hay dentro. Usaremos la función `names` que nos permite conocer el nombre de las diferentes variables.

```
names(ameLoja)
```

```
## [1] "Cantón"           "Distrito"         "Dis.Distribucion"
## [4] "Sexo"             "Edad.en.años"     "N.X"
## [7] "N.Y"              "Consultas"        "Parroquia"
```

Nuestra data.frame tiene unos nombres de variables bastante descriptivos, aunque hay alguno que no está muy claro, como N.Y, efectivamente no quiere decir la distancia a Nueva York y N.X la distancia a Xalapa, estas dos variables corresponden a las coordenadas geográficas latitud x y longitud y. Ahora, necesitamos echar un vistazo a los datos reales. Sin embargo, nuestra base de datos contiene 3019 observaciones (filas), así que es poco práctico ver toda la tabla a la vez.

La función `head` permite hacer una vista previa de la parte superior del conjunto de datos y la función `tail` la parte inferior de estos datos. Adicionalmente, yo podría cambiar la cantidad de observaciones que me presentan estas funciones agregando un valor en estas funciones. Veamos.

```
head(ameLoja, 5)
```

```
##   Cantón Distrito Dis.Distribucion   Sexo Edad.en.años                 N.X
## 1   LOJA    11D01             LOJA Hombre            1 683.887.999.999.509
## 2   LOJA    11D01             LOJA Hombre           13 683.887.999.999.509
## 3   LOJA    11D01             LOJA Hombre           14 683.887.999.999.509
## 4   LOJA    11D01             LOJA Hombre            2 683.887.999.999.509
## 5   LOJA    11D01             LOJA Hombre            2  68.989.299.999.942
##                   N.Y Consultas               Parroquia
```

```
## 1 957.489.600.000.001         1           CHUQUIRIBAMBA
## 2 957.489.600.000.001         2           CHUQUIRIBAMBA
## 3 957.489.600.000.001         1           CHUQUIRIBAMBA
## 4 957.489.600.000.001         1           CHUQUIRIBAMBA
## 5 956.987.200.000.001         1 TAQUIL (MIGUEL RIOFRÍO)
```

```
tail(ameLoja, 6)
```

```
##         Cantón Distrito          Dis.Distribucion  Sexo Edad.en.años
## 3014 GONZANAMÁ    11D06 CALVAS,GONZANAMA,QUILANGA Mujer           49
## 3015 GONZANAMÁ    11D06 CALVAS,GONZANAMA,QUILANGA Mujer            5
## 3016 GONZANAMÁ    11D06 CALVAS,GONZANAMA,QUILANGA Mujer           52
## 3017 GONZANAMÁ    11D06 CALVAS,GONZANAMA,QUILANGA Mujer           55
## 3018 GONZANAMÁ    11D06 CALVAS,GONZANAMA,QUILANGA Mujer            7
## 3019 GONZANAMÁ    11D06 CALVAS,GONZANAMA,QUILANGA Mujer           82
##                       N.X                N.Y Consultas Parroquia
## 3014 673.123.999.999.634 953.243.300.000.001         1 GONZANAMÁ
## 3015 673.123.999.999.634 953.243.300.000.001         1 GONZANAMÁ
## 3016 673.123.999.999.634 953.243.300.000.001         1 GONZANAMÁ
## 3017 673.123.999.999.634 953.243.300.000.001         1 GONZANAMÁ
## 3018 673.123.999.999.634 953.243.300.000.001         1 GONZANAMÁ
## 3019 673.123.999.999.634 953.243.300.000.001         1 GONZANAMÁ
```

Como vemos este código me mostró las cinco primeras y seis ultimas observaciones del objeto, usted puede probar otros valores.

Ahora vamos a utilizar la función `str` para ver qué tipo de variables tenemos en este objeto.

```
str(ameLoja)
```

```
## 'data.frame':    3019 obs. of  9 variables:
##  $ Cantón          : Factor w/ 16 levels "CALVAS","CATAMAYO",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ Distrito        : Factor w/ 9 levels "11D01","11D02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Dis.Distribucion: Factor w/ 9 levels "CALVAS,GONZANAMA,QUILANGA",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Sexo            : Factor w/ 2 levels "Hombre","Mujer": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Edad.en.años    : int  1 13 14 2 2 22 3 30 36 4 ...
##  $ N.X             : Factor w/ 94 levels "560.125.999.999.995",..: 56 56 56 56 50 56 56 56 50 56 ...
##  $ N.Y             : Factor w/ 94 levels "948.835.300.000.001",..: 73 73 73 73 62 73 73 73 62 73 ...
##  $ Consultas       : int  1 2 1 1 1 1 2 1 1 1 ...
##  $ Parroquia       : Factor w/ 62 levels "12 DE DICIEMBRE (CAB.EN ACHIOTES)",..: 15 15 15 15 56 15 15
```

Revise los datos, ¿ve algo raro? Fíjese bien en las variables N.X y N.Y, lo ve, tenemos un problema los datos no son números, los está tomando como factores. Mire atentamente los datos, puede verlo mejor en el resultado anterior de head o tail. Claro la separaciones de miles ha sido utilizado un punto. Bueno por ahora, sabemos que hay un problema que hay que resolver, pero lo dejaremos para más adelante. La función str además nos da información parecida a la de dim ya que nos da los datos de cuantas variables (columnas) y cuantas observaciones (filas).

Características de las variables

Ahora que ya sabemos cómo está nuestra tabla de datos podemos fijarnos en nuestras variables. Una función que se utilizará mucho en R es la función `summary`, esta función nos da información de resumen. Cuando la función sumary es utilizada en un objeto, nos mostrará algunos datos descriptivos de las variables.

```
summary(ameLoja)
```

```
##        Cantón        Distrito                          Dis.Distribucion
##  LOJA     :1511   11D01  :1511   LOJA                          :1511
##  ESPÍNDOLA: 435   11D05  : 435   ESPINDOLA                     : 435
##  PALTAS   : 223   11D03  : 223   PALTAS                        : 223
##  SARAGURO : 208   11D02  : 217   CATAMAYO,CHAGUARPAMBA,OLMEDO: 217
##  CATAMAYO : 197   11D08  : 208   SARAGURO                      : 208
##  CALVAS   : 144   11D06  : 189   CALVAS,GONZANAMA,QUILANGA     : 189
##  (Other)  : 301   (Other): 236   (Other)                       : 236
##      Sexo       Edad.en.años                      N.X
##  Hombre:1179   Min.   :  0.00   699.479.999.999.254: 391
##  Mujer :1840   1st Qu.: 10.00   700.035.999.999.237: 252
##                Median : 22.00   674.264.999.999.636: 210
##                Mean   : 27.76   699.197.706.600.898: 203
##                3rd Qu.: 41.00   649.964.999.999.821: 178
##                Max.   :600.00   660.316.999.999.752: 133
##                                 (Other)            :1652
##                    N.Y         Consultas                    Parroquia
##  955.951.800.000.001: 391   Min.   : 1.000   LOJA              :1372
##  955.719.700.000.001: 252   1st Qu.: 1.000   AMALUZA           : 308
##  949.251.800.000.001: 210   Median : 1.000   CATACOCHA         : 179
##  955.929.493.221.639: 203   Mean   : 1.858   CARIAMANGA        : 133
##  9552096            : 178   3rd Qu.: 2.000   CATAMAYO (LA TOMA): 133
##  9521662            : 133   Max.   :41.000   MACARÁ            : 105
##  (Other)            :1652                    (Other)           : 789
```

Como vemos la salida es diferente para cada variable, esta salida es dependiente de su clase. Para los datos numéricos como *edad.en.años* summary muestra los siguientes datos; el mínimo, primer cuartil, la mediana, la media, el tercer cuartil, y el máximo. Los valores obtnidos 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') aparece en los datos. Por ejemplo, la variable sexo aparece hombres 1179 y mujeres 1840.

Se puede ver que R limita los resultados de las variables categóricas en 6 niveles incluyendo una nueva categoría denominada 'Other', esto lo hace con el fin de estandarizar la salida. Dado que es una variable categórica / Factor, podemos ver cuántas veces ocurre cada valor realmente en los datos con table(ameLoja$Cantón).

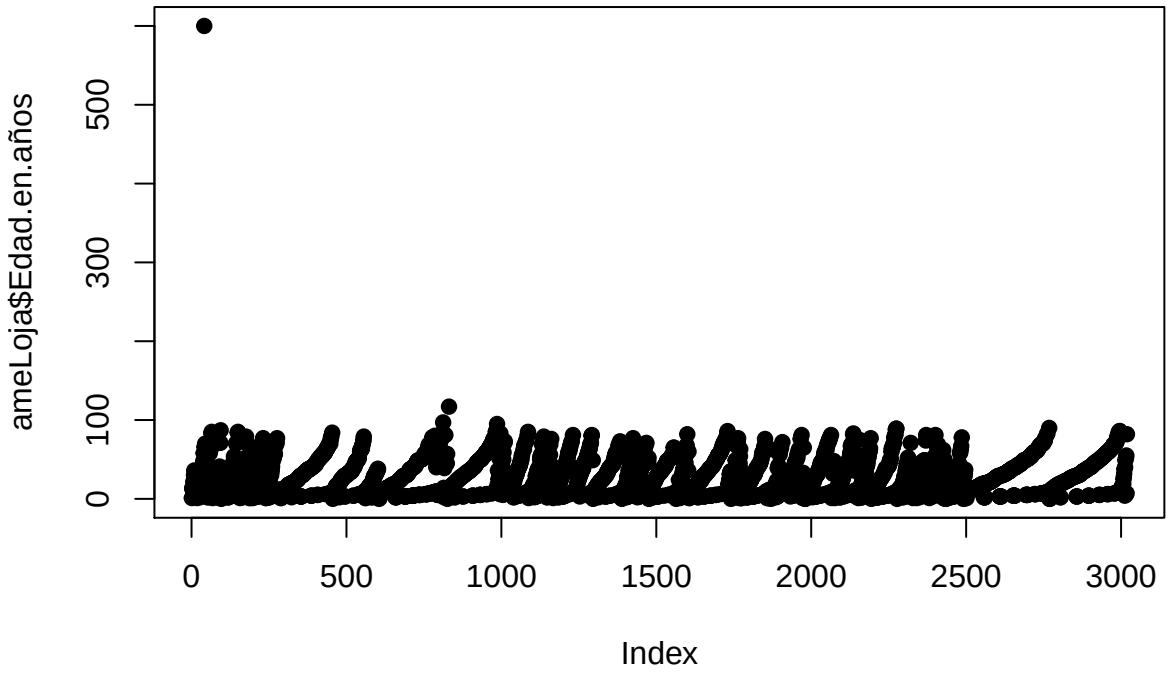
```
table(ameLoja$Cantón)
```

```
##
##       CALVAS     CATAMAYO       CELICA CHAGUARPAMBA    ESPÍNDOLA
##          144          197           23           12          435
##    GONZANAMÁ        LOJA       MACARÁ       OLMEDO       PALTAS
##           40         1511          105            8          223
##       PINDAL      PUYANGO     QUILANGA     SARAGURO    SOZORANGA
##           21           59            5          208            7
##    ZAPOTILLO
##           21
```

Ahora podemos ver todas las categorías de la variable Cantón, cada uno con su frecuencia.

En los datos de amebiasis tenemos una variable numérica edad años, vamos a verificar si los datos de esta variable son correctos. Utilizaremos un gráfico para ver como se distribuye esta variable.

```
plot(ameLoja$Edad.en.años, pch=19)
```



¿Qué es lo que ven?

Hay un valor que se encuentra fuera de lo normal, supuestamente hay una persona con 600 años. Esto es poco probable así que deberiamos corregir este dato antes de continuar.

Ejercicio 2

Descargue los datos de plantas que los puede encontrar aquí súbalos a la consola y descríbalos.

1. ¿Cuáles son las dimensiones de este objeto? (cuantas variables y cuantas observaciones)
2. Utilice head y tail para ver las características internas de los datos ¿qué ven?
3. Describa cuantas variables categóricas y numéricas tiene este set de datos.

4. Describa al cuatro variables, dos numéricas y dos categóricas. Realice un gráfico descriptivo de cada una de estas variables.

5. Revisar si alguna de las variables númericas tienen valores erróneos, ubicarlos y responder ¿por qué considera que son erróneos?.