Looking at Data

Raphael Carvalho

02/06/2019

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.

Looking at Data

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"
```

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   12
```

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

```
nrow(plants)
```

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

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

```
ncol(plants)
```

```
## [1] 12
```

If you are curious as to how much space the dataset is occupying in memory, you can use object.size(plants).

```
object.size(plants)
```

```
## 1058704 bytes
```

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] "Accepted.Symbol"          "Synonym.Symbol"
##  [3] "Scientific.Name"          "Duration"
##  [5] "Active.Growth.Period"     "Foliage.Color"
##  [7] "pH..Minimum."             "pH..Maximum."
##  [9] "Precipitation..Minimum."  "Precipitation..Maximum."
## [11] "Shade.Tolerance"          "Temperature..Minimum...F."
```

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.

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

```
head(plants)
```

```
##   Accepted.Symbol Synonym.Symbol              Scientific.Name
## 1           ABELM             NA                   Abelmoschus
## 2            ABES             NA       Abelmoschus esculentus
## 3           ABIES             NA                        Abies
## 4            ABBA             NA               Abies balsamea
## 5           ABBAB             NA Abies balsamea var. balsamea
## 6           ABUTI             NA                      Abutilon
##            Duration Active.Growth.Period Foliage.Color pH..Minimum.
## 1              <NA>                <NA>          <NA>           NA
## 2 Annual, Perennial                <NA>          <NA>           NA
## 3              <NA>                <NA>          <NA>           NA
## 4         Perennial    Spring and Summer         Green            4
## 5         Perennial                <NA>          <NA>           NA
## 6              <NA>                <NA>          <NA>           NA
##   pH..Maximum. Precipitation..Minimum. Precipitation..Maximum.
## 1           NA                     NA                      NA
## 2           NA                     NA                      NA
## 3           NA                     NA                      NA
## 4            6                     13                      60
## 5           NA                     NA                      NA
## 6           NA                     NA                      NA
##   Shade.Tolerance Temperature..Minimum...F.
## 1            <NA>                       NA
## 2            <NA>                       NA
## 3            <NA>                       NA
## 4        Tolerant                      -43
## 5            <NA>                       NA
## 6            <NA>                       NA
```

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)
```

```
##    Accepted.Symbol Synonym.Symbol                  Scientific.Name
## 1            ABELM             NA                       Abelmoschus
## 2             ABES             NA           Abelmoschus esculentus
## 3            ABIES             NA                            Abies
## 4             ABBA             NA                   Abies balsamea
## 5            ABBAB             NA      Abies balsamea var. balsamea
## 6            ABUTI             NA                          Abutilon
## 7             ABTH             NA             Abutilon theophrasti
## 8            ACACI             NA                           Acacia
## 9            ACCO2             NA                 Acacia constricta
## 10           ACCOC             NA Acacia constricta var. constricta
##             Duration Active.Growth.Period Foliage.Color pH..Minimum.
## 1              <NA>                <NA>          <NA>           NA
## 2  Annual, Perennial                <NA>          <NA>           NA
## 3              <NA>                <NA>          <NA>           NA
## 4         Perennial    Spring and Summer         Green            4
## 5         Perennial                <NA>          <NA>           NA
## 6              <NA>                <NA>          <NA>           NA
```

```
## 7             Annual                 <NA>          <NA>           NA
## 8               <NA>                 <NA>          <NA>           NA
## 9          Perennial    Spring and Summer         Green            7
## 10         Perennial                 <NA>          <NA>           NA
##    pH..Maximum. Precipitation..Minimum. Precipitation..Maximum.
## 1            NA                      NA                      NA
## 2            NA                      NA                      NA
## 3            NA                      NA                      NA
## 4           6.0                      13                      60
## 5            NA                      NA                      NA
## 6            NA                      NA                      NA
## 7            NA                      NA                      NA
## 8            NA                      NA                      NA
## 9           8.5                       4                      20
## 10           NA                      NA                      NA
##    Shade.Tolerance Temperature..Minimum...F.
## 1             <NA>                       NA
## 2             <NA>                       NA
## 3             <NA>                       NA
## 4         Tolerant                      -43
## 5             <NA>                       NA
## 6             <NA>                       NA
## 7             <NA>                       NA
## 8             <NA>                       NA
## 9       Intolerant                      -13
## 10            <NA>                       NA
```

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

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

```
##      Accepted.Symbol Synonym.Symbol                  Scientific.Name
## 5152           ZIZAN             NA                           Zizania
## 5153            ZIAQ             NA                  Zizania aquatica
## 5154          ZIAQA2             NA   Zizania aquatica var. aquatica
## 5155           ZIPA3             NA                Zizania palustris
## 5156           ZIPAP             NA Zizania palustris var. palustris
## 5157          ZIZAN2             NA                       Zizaniopsis
## 5158            ZIMI             NA             Zizaniopsis miliacea
## 5159           ZIZIA             NA                             Zizia
## 5160            ZIAP             NA                     Zizia aptera
## 5161            ZIAU             NA                      Zizia aurea
## 5162            ZITR             NA                  Zizia trifoliata
## 5163           ZOSTE             NA                           Zostera
## 5164            ZOMA             NA                    Zostera marina
## 5165           ZOYSI             NA                           Zoysia
## 5166            ZOJA             NA                   Zoysia japonica
##       Duration Active.Growth.Period Foliage.Color pH..Minimum.
## 5152      <NA>                 <NA>          <NA>           NA
## 5153    Annual               Spring         Green          6.4
## 5154    Annual                 <NA>          <NA>           NA
## 5155    Annual                 <NA>          <NA>           NA
## 5156    Annual                 <NA>          <NA>           NA
## 5157      <NA>                 <NA>          <NA>           NA
## 5158 Perennial    Spring and Summer         Green          4.3
```

```
## 5159      <NA>                 <NA>          <NA>           NA
## 5160 Perennial                 <NA>          <NA>           NA
## 5161 Perennial                 <NA>          <NA>           NA
## 5162 Perennial                 <NA>          <NA>           NA
## 5163      <NA>                 <NA>          <NA>           NA
## 5164 Perennial                 <NA>          <NA>           NA
## 5165      <NA>                 <NA>          <NA>           NA
## 5166 Perennial                 <NA>          <NA>           NA
##      pH..Maximum. Precipitation..Minimum. Precipitation..Maximum.
## 5152           NA                      NA                       NA
## 5153          7.4                      30                       50
## 5154           NA                      NA                       NA
## 5155           NA                      NA                       NA
## 5156           NA                      NA                       NA
## 5157           NA                      NA                       NA
## 5158          9.0                      35                       70
## 5159           NA                      NA                       NA
## 5160           NA                      NA                       NA
## 5161           NA                      NA                       NA
## 5162           NA                      NA                       NA
## 5163           NA                      NA                       NA
## 5164           NA                      NA                       NA
## 5165           NA                      NA                       NA
## 5166           NA                      NA                       NA
##      Shade.Tolerance Temperature..Minimum...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
```

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)
```

```
##  Accepted.Symbol Synonym.Symbol                    Scientific.Name
##  ABBA   :   1    Mode:logical   Abelmoschus                  :   1
##  ABBAB  :   1    NA's:5166      Abelmoschus esculentus       :   1
##  ABELM  :   1                   Abies                        :   1
##  ABES   :   1                   Abies balsamea               :   1
##  ABIES  :   1                   Abies balsamea var. balsamea:   1
##  ABTH   :   1                   Abutilon                     :   1
##  (Other):5160                   (Other)                      :5160
```

```
##               Duration              Active.Growth.Period
##  Perennial        :3031   Spring and Summer   : 447
##  Annual           : 682   Spring              : 144
##  Annual, Perennial: 179   Spring, Summer, Fall:  95
##  Annual, Biennial :  95   Summer              :  92
##  Biennial         :  57   Summer and Fall     :  24
##  (Other)          :  92   (Other)             :  30
##  NA's             :1030   NA's                :4334
##       Foliage.Color   pH..Minimum.    pH..Maximum.
##  Dark Green  :  82   Min.   :3.000   Min.   : 5.100
##  Gray-Green  :  25   1st Qu.:4.500   1st Qu.: 7.000
##  Green       : 692   Median :5.000   Median : 7.300
##  Red         :   4   Mean   :4.997   Mean   : 7.344
##  White-Gray  :   9   3rd Qu.:5.500   3rd Qu.: 7.800
##  Yellow-Green:  20   Max.   :7.000   Max.   :10.000
##  NA's        :4334   NA's   :4327    NA's   :4327
##  Precipitation..Minimum. Precipitation..Maximum.     Shade.Tolerance
##  Min.   : 4.00           Min.   : 16.00          Intermediate: 242
##  1st Qu.:16.75           1st Qu.: 55.00          Intolerant  : 349
##  Median :28.00           Median : 60.00          Tolerant    : 246
##  Mean   :25.57           Mean   : 58.73          NA's        :4329
##  3rd Qu.:32.00           3rd Qu.: 60.00
##  Max.   :60.00           Max.   :200.00
##  NA's   :4338            NA's   :4338
##  Temperature..Minimum...F.
##  Min.   :-79.00
##  1st Qu.:-38.00
##  Median :-33.00
##  Mean   :-22.53
##  3rd Qu.:-18.00
##  Max.   : 52.00
##  NA's   :4328
```

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)
```

```
## < table of extent 0 >
```

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  12 variables:
##  $ Accepted.Symbol          : Factor w/ 5166 levels "ABBA","ABBAB",..: 3 4 5 1 2 7 6 8 15 16 ...
##  $ Synonym.Symbol           : logi  NA NA NA NA NA NA ...
##  $ 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
##  $ Active.Growth.Period     : Factor w/ 8 levels "Fall, Winter and Spring",..: NA NA NA 4 NA NA NA NA
##  $ Foliage.Color            : Factor w/ 6 levels "Dark Green","Gray-Green",..: NA NA NA 3 NA NA NA NA
##  $ pH..Minimum.             : num  NA NA NA 4 NA NA NA NA 7 NA ...
##  $ pH..Maximum.             : num  NA NA NA 6 NA NA NA NA 8.5 NA ...
##  $ Precipitation..Minimum.  : int  NA NA NA 13 NA NA NA NA 4 NA ...
```

```
##  $ Precipitation..Maximum.  : 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 ...
##  $ Temperature..Minimum...F.: int  NA NA NA -43 NA NA NA NA -13 NA ...
```