Apuntes Data Cleaning

Pilar Amat Rodrigo

7/12/2017

Contents

Data_document: weather in https://assets.datacamp.com/production/repositories/34/datasets/b3c1036d9a60a9dfe0f99051d2474a54f76055ea/weather.rds

Packages: library(dplyr) // library(tidyr)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

1. Exploring raw data:

1. Check the kind of the data: **class(x)**
2. View its dimension: **dim(x)**
3. Read the column names: **names(x)**
4. Read the structure: **str(x)**, we can also use **glimpse(x)** from the dplyr package. Summary could be a nice approach too **summary(x)**
5. Look at the data *perse*: **head(x,n=)**, **tail(x,n=)**. Only recommend **print(x)** when it's a small set.
6. Visualizing the data: view a histogram **hist(x&a)**, or plotting data **plot(x&a,x&b)**.

Principles of tidy data

Observations as rows, variables as columns & one type of observational unit per table

2. Tidyr

Gathering columns into key-value pairs

The most important function in tidyr is gather(). It should be used when you have columns that are not variables and you want to collapse them into key-value pairs.

The easiest way to visualize the effect of gather() is that it makes wide datasets long:

```
gather(data, key, values, ...)
    data: The name of the data object
    key: The name of the new column whose values will be what are now the column headers
    value: The name of the new column whose values will be the actual data object measurements
    ... The columns to gather (or, in this case, the column to exclude from gathering, use -)
```

```
bmi <- read.csv("bmi_clean.txt")
bmi_long<- gather(bmi, Year, Values, - Country)

#Dimensiones de la tabla original
dim(bmi)
```

```
## [1] 199  30
```

```
#Dimensiones de la tabla tras el gather
dim(bmi_long)
```

```
## [1] 5771    3
```

The opposite of gather() is spread()

which takes key-values pairs and spreads them across multiple columns. This is useful when values in a column should actually be column names (i.e. variables). It can also make data more compact and easier to read.

The easiest way to visualize the effect of spread() is that it makes long datasets wide:

```
spread(data, key, values)
    data: is a data frame
    key: The name of the column whose values will become column headers for the wide dataset
    value: The name of the column containing the actual BMI measurements.
```

```
bmi_wide <- spread(bmi_long, Year, Values) # my_key es el nombre de la columna que contiene las etiquet


#Volvemos a la posición inicial con la función spread
dim(bmi_wide)
```

```
## [1] 199  30
```

Separate Columns

The **separate()** function allows you to separate one column into multiple columns. Unless you tell it otherwise, it will attempt to separate on any character that is not a letter or number. You can also specify a specific separator using the sep argument.

```
separate(data, col, into)
    data: is a data frame
    col: the name of the column to be split
    into: a character vector of the new column names -> c("year","month")
    sep: separator "-"
```

Unite Columns

The opposite of separate() is **unite()**, which takes multiple columns and pastes them together. By default, the contents of the columns will be separated by underscores in the new column, but this behavior can be altered via the sep argument.

```
  unite(data, col,...)
      data: is the data frame
```

col: is the name of the new column.
...: followed by the names of the columns that are being merged.

3. Type Conversions

as.character()

as.numeric()

as.integer()

as.factor()

as.logical()

3.1 Lubridate Package

Dates

A package called lubridade very useful over dates. Insert the dates as it's specified in the method. library(lubridate)

ymd() year moth day

hms ()

ymd_hms ()

3.2 stingr package

Manipulating strings

str_trim () deleting white spaces

str_pad() add zeros or completing with zeros

```
library(stringr)

#Completa hasta el campo 10 con ceros por la izquierda.

str_pad("459", width=10, side= "left", pad="0")
```

```
## [1] "0000000459"
```

str_detect() Detect a pattern

str_replace(element_container, element_to_replace, replacement_element) Find and replace a pattern.

tolower() lowercase

toupper() uppercase

3.3 Missing Values/ Special (inf / NaN)

finding Na

is.na() to find NAs

any(is.na())

sum(is.na()) Conseguimos contar los TRUE como 1

complete.cases(df) TRUE is complete, FALSE is there is anydata missing.

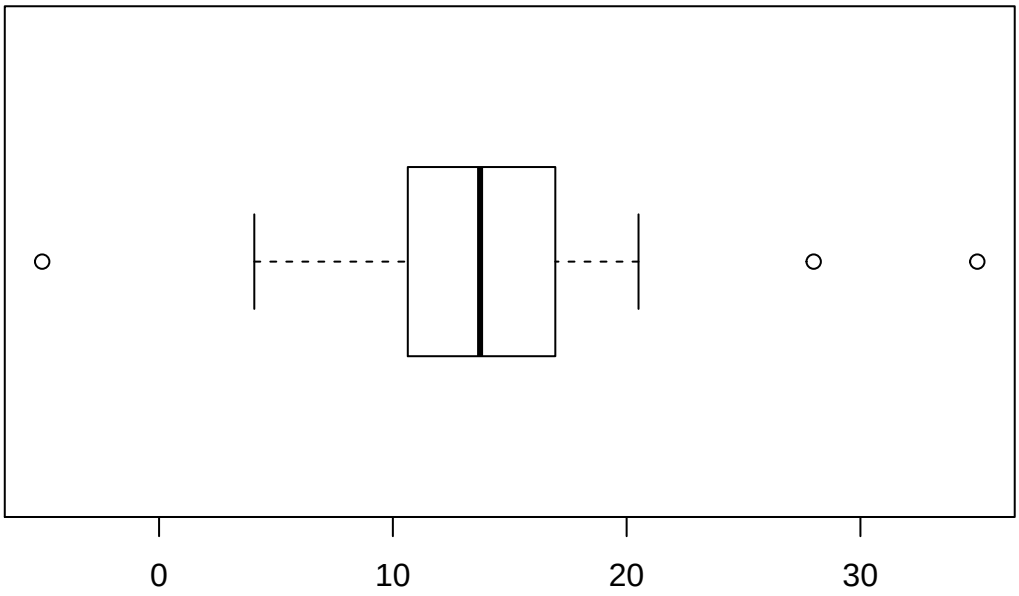
na.omit()

3.4 Outliers

Better use plots

```
#simulate some data to see outlier
set.seed(10)
x<- c(rnorm(30, mean=15, sd=5), -5, 28, 35)

#view a boxplot
boxplot(x, horizontal = TRUE)
```



4. Exercise

Clean a real dataset and leave it ready for analysing. Dataset: weather.rds

4.1 Read and explore data:

Take a look**

```
#import dataset
weather <- readRDS("weather.rds")
```

```r
# Verify that weather is a data.frame
class(weather)
```

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

```r
# Check the dimensions
dim(weather)
```

```
## [1] 286  35
```

```r
# View the column names
colnames(weather)
```

```
##  [1] "X"       "year"    "month"   "measure" "X1"      "X2"      "X3"
##  [8] "X4"      "X5"      "X6"      "X7"      "X8"      "X9"      "X10"
## [15] "X11"     "X12"     "X13"     "X14"     "X15"     "X16"     "X17"
## [22] "X18"     "X19"     "X20"     "X21"     "X22"     "X23"     "X24"
## [29] "X25"     "X26"     "X27"     "X28"     "X29"     "X30"     "X31"
```

```r
# View the structure of the data
str(weather)
```

```
## 'data.frame':    286 obs. of  35 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ year    : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
##  $ month   : int  12 12 12 12 12 12 12 12 12 12 ...
##  $ measure: chr  "Max.TemperatureF" "Mean.TemperatureF" "Min.TemperatureF" "Max.Dew.PointF" ...
##  $ X1      : chr  "64" "52" "39" "46" ...
##  $ X2      : chr  "42" "38" "33" "40" ...
##  $ X3      : chr  "51" "44" "37" "49" ...
##  $ X4      : chr  "43" "37" "30" "24" ...
##  $ X5      : chr  "42" "34" "26" "37" ...
##  $ X6      : chr  "45" "42" "38" "45" ...
##  $ X7      : chr  "38" "30" "21" "36" ...
##  $ X8      : chr  "29" "24" "18" "28" ...
##  $ X9      : chr  "49" "39" "29" "49" ...
##  $ X10     : chr  "48" "43" "38" "45" ...
##  $ X11     : chr  "39" "36" "32" "37" ...
##  $ X12     : chr  "39" "35" "31" "28" ...
##  $ X13     : chr  "42" "37" "32" "28" ...
##  $ X14     : chr  "45" "39" "33" "29" ...
##  $ X15     : chr  "42" "37" "32" "33" ...
##  $ X16     : chr  "44" "40" "35" "42" ...
##  $ X17     : chr  "49" "45" "41" "46" ...
##  $ X18     : chr  "44" "40" "36" "34" ...
##  $ X19     : chr  "37" "33" "29" "25" ...
##  $ X20     : chr  "36" "32" "27" "30" ...
##  $ X21     : chr  "36" "33" "30" "30" ...
##  $ X22     : chr  "44" "39" "33" "39" ...
##  $ X23     : chr  "47" "45" "42" "45" ...
##  $ X24     : chr  "46" "44" "41" "46" ...
##  $ X25     : chr  "59" "52" "44" "58" ...
##  $ X26     : chr  "50" "44" "37" "31" ...
##  $ X27     : chr  "52" "45" "38" "34" ...
##  $ X28     : chr  "52" "46" "40" "42" ...
##  $ X29     : chr  "41" "36" "30" "26" ...
##  $ X30     : chr  "30" "26" "22" "10" ...
```

```
##  $ X31    : chr  "30" "25" "20" "8" ...
```

```
# Load dplyr package
library(dplyr)

# Look at the structure using dplyr's glimpse()
glimpse(weather)
```

```
## Observations: 286
## Variables: 35
## $ X       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...
## $ year    <int> 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, ...
## $ month   <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12...
## $ measure <chr> "Max.TemperatureF", "Mean.TemperatureF", "Min.Temperat...
## $ X1      <chr> "64", "52", "39", "46", "40", "26", "74", "63", "52", ...
## $ X2      <chr> "42", "38", "33", "40", "27", "17", "92", "72", "51", ...
## $ X3      <chr> "51", "44", "37", "49", "42", "24", "100", "79", "57",...
## $ X4      <chr> "43", "37", "30", "24", "21", "13", "69", "54", "39", ...
## $ X5      <chr> "42", "34", "26", "37", "25", "12", "85", "66", "47", ...
## $ X6      <chr> "45", "42", "38", "45", "40", "36", "100", "93", "85",...
## $ X7      <chr> "38", "30", "21", "36", "20", "-3", "92", "61", "29", ...
## $ X8      <chr> "29", "24", "18", "28", "16", "3", "92", "70", "47", "...
## $ X9      <chr> "49", "39", "29", "49", "41", "28", "100", "93", "86",...
## $ X10     <chr> "48", "43", "38", "45", "39", "37", "100", "95", "89",...
## $ X11     <chr> "39", "36", "32", "37", "31", "27", "92", "87", "82", ...
## $ X12     <chr> "39", "35", "31", "28", "27", "25", "85", "75", "64", ...
## $ X13     <chr> "42", "37", "32", "28", "26", "24", "75", "65", "55", ...
## $ X14     <chr> "45", "39", "33", "29", "27", "25", "82", "68", "53", ...
## $ X15     <chr> "42", "37", "32", "33", "29", "27", "89", "75", "60", ...
## $ X16     <chr> "44", "40", "35", "42", "36", "30", "96", "85", "73", ...
## $ X17     <chr> "49", "45", "41", "46", "41", "32", "100", "85", "70",...
## $ X18     <chr> "44", "40", "36", "34", "30", "26", "89", "73", "57", ...
## $ X19     <chr> "37", "33", "29", "25", "22", "20", "69", "63", "56", ...
## $ X20     <chr> "36", "32", "27", "30", "24", "20", "89", "79", "69", ...
## $ X21     <chr> "36", "33", "30", "30", "27", "25", "85", "77", "69", ...
## $ X22     <chr> "44", "39", "33", "39", "34", "25", "89", "79", "69", ...
## $ X23     <chr> "47", "45", "42", "45", "42", "37", "100", "91", "82",...
## $ X24     <chr> "46", "44", "41", "46", "44", "41", "100", "98", "96",...
## $ X25     <chr> "59", "52", "44", "58", "43", "29", "100", "75", "49",...
## $ X26     <chr> "50", "44", "37", "31", "29", "28", "70", "60", "49", ...
## $ X27     <chr> "52", "45", "38", "34", "31", "29", "70", "60", "50", ...
## $ X28     <chr> "52", "46", "40", "42", "35", "27", "76", "65", "53", ...
## $ X29     <chr> "41", "36", "30", "26", "20", "10", "64", "51", "37", ...
## $ X30     <chr> "30", "26", "22", "10", "4", "-6", "50", "38", "26", "...
## $ X31     <chr> "30", "25", "20", "8", "5", "1", "57", "44", "31", "30...
```

```
# View a summary of the data
summary(weather)
```

```
##        X              year          month          measure
##  Min.   :  1.00   Min.   :2014   Min.   : 1.000   Length:286
##  1st Qu.: 72.25   1st Qu.:2015   1st Qu.: 4.000   Class :character
##  Median :143.50   Median :2015   Median : 7.000   Mode  :character
##  Mean   :143.50   Mean   :2015   Mean   : 6.923
##  3rd Qu.:214.75   3rd Qu.:2015   3rd Qu.:10.000
##  Max.   :286.00   Max.   :2015   Max.   :12.000
```

```
##       X1                  X2                  X3
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##       X4                  X5                  X6
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##       X7                  X8                  X9
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X10                 X11                 X12
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X13                 X14                 X15
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X16                 X17                 X18
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X19                 X20                 X21
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X22                 X23                 X24
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##      X25                X26                X27
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X28                X29                X30
##  Length:286         Length:286         Length:286
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##      X31
##  Length:286
##  Class :character
##  Mode  :character
##
##
##
```

```
# View first 6 rows
head(weather, n=6)
```

```
##   X year month           measure X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12
## 1 1 2014    12  Max.TemperatureF 64 42 51 43 42 45 38 29 49  48  39  39
## 2 2 2014    12 Mean.TemperatureF 52 38 44 37 34 42 30 24 39  43  36  35
## 3 3 2014    12  Min.TemperatureF 39 33 37 30 26 38 21 18 29  38  32  31
## 4 4 2014    12    Max.Dew.PointF 46 40 49 24 37 45 36 28 49  45  37  28
## 5 5 2014    12    MeanDew.PointF 40 27 42 21 25 40 20 16 41  39  31  27
## 6 6 2014    12     Min.DewpointF 26 17 24 13 12 36 -3  3 28  37  27  25
##   X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30
## 1  42  45  42  44  49  44  37  36  36  44  47  46  59  50  52  52  41  30
## 2  37  39  37  40  45  40  33  32  33  39  45  44  52  44  45  46  36  26
## 3  32  33  32  35  41  36  29  27  30  33  42  41  44  37  38  40  30  22
## 4  28  29  33  42  46  34  25  30  30  39  45  46  58  31  34  42  26  10
## 5  26  27  29  36  41  30  22  24  27  34  42  44  43  29  31  35  20   4
## 6  24  25  27  30  32  26  20  20  25  25  37  41  29  28  29  27  10  -6
##   X31
## 1  30
## 2  25
## 3  20
## 4   8
## 5   5
## 6   1
```

```
# View the last 6 rows
tail(weather, n=6)
```

```
##       X year month            measure   X1   X2   X3   X4   X5   X6   X7
## 281 281 2015    12 Mean.Wind.SpeedMPH    6 <NA> <NA> <NA> <NA> <NA> <NA>
## 282 282 2015    12  Max.Gust.SpeedMPH   17 <NA> <NA> <NA> <NA> <NA> <NA>
## 283 283 2015    12    PrecipitationIn 0.14 <NA> <NA> <NA> <NA> <NA> <NA>
```

```
## 284 284 2015    12         CloudCover    7 <NA> <NA> <NA> <NA> <NA> <NA>
## 285 285 2015    12             Events Rain <NA> <NA> <NA> <NA> <NA> <NA>
## 286 286 2015    12     WindDirDegrees  109 <NA> <NA> <NA> <NA> <NA> <NA>
##       X8   X9  X10  X11  X12  X13  X14  X15  X16  X17  X18  X19  X20  X21
## 281 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 282 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 283 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 284 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 285 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 286 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##      X22  X23  X24  X25  X26  X27  X28  X29  X30  X31
## 281 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 282 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 283 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 284 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 285 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 286 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
```

4.2 Tidy data:

Column names are values**

The weather dataset suffers from one of the five most common symptoms of messy data: column names are values. In particular, the column names X1-X31 represent days of the month, which should really be values of a new variable called day.

```
# Load the tidyr package
library(tidyr)

# Gather the columns
weather2 <- gather(weather, day, value, c(X1:X31), na.rm = TRUE)

# View the head
head(weather2)
```

```
##   X year month           measure day value
## 1 1 2014    12  Max.TemperatureF  X1    64
## 2 2 2014    12 Mean.TemperatureF  X1    52
## 3 3 2014    12  Min.TemperatureF  X1    39
## 4 4 2014    12    Max.Dew.PointF  X1    46
## 5 5 2014    12    MeanDew.PointF  X1    40
## 6 6 2014    12     Min.DewpointF  X1    26
```

Values are variable names

Our data suffer from a second common symptom of messy data: values are variable names. Specifically, values in the measure column should be variables (i.e. column names) in our dataset.

```
# First remove column of row names
weather2 <- weather2[, -1]

# Spread the data
weather3 <- spread(weather2, measure, value)

# View the head
head(weather3)
```

```
##   year month day CloudCover    Events Max.Dew.PointF Max.Gust.SpeedMPH
## 1 2014    12  X1          6      Rain             46                29
## 2 2014    12 X10          8      Rain             45                29
## 3 2014    12 X11          8 Rain-Snow             37                28
## 4 2014    12 X12          7      Snow             28                21
## 5 2014    12 X13          5                       28                23
## 6 2014    12 X14          4                       29                20
##   Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 1           74                    30.45               64
## 2          100                    29.58               48
## 3           92                    29.81               39
## 4           85                    29.88               39
## 5           75                    29.86               42
## 6           82                    29.91               45
##   Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF Mean.Humidity
## 1                  10                22             40            63
## 2                  10                23             39            95
## 3                  10                21             31            87
## 4                  10                16             27            75
## 5                  10                17             26            65
## 6                  10                15             27            68
##   Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 1                     30.13               52                   10
## 2                      29.5               43                    3
## 3                     29.61               36                    7
## 4                     29.85               35                   10
## 5                     29.82               37                   10
## 6                     29.83               39                   10
##   Mean.Wind.SpeedMPH Min.DewpointF Min.Humidity Min.Sea.Level.PressureIn
## 1                 13            26           52                    30.01
## 2                 13            37           89                    29.43
## 3                 13            27           82                    29.44
## 4                 11            25           64                    29.81
## 5                 12            24           55                    29.78
## 6                 10            25           53                    29.78
##   Min.TemperatureF Min.VisibilityMiles PrecipitationIn WindDirDegrees
## 1               39                  10            0.01            268
## 2               38                   1            0.28            357
## 3               32                   1            0.02            230
## 4               31                   7               T            286
## 5               32                  10               T            298
## 6               33                  10            0.00            306
```

4.3 Clean data:
Clean up dates**

Now that the weather dataset adheres to tidy data principles, the next step is to prepare it for analysis. We'll start by combining the year, month, and day columns and recoding the resulting character column as a date. We can use a combination of base R, stringr, and lubridate to accomplish this task.

```
# Load the stringr and lubridate packages
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
# Remove X's from day column
weather3$day <- str_replace(weather3$day,"X","")

# Unite the year, month, and day columns
weather4 <- unite(weather3, date, year, month, day, sep = "-")

# Convert date column to proper date format using lubridates's ymd()
weather4$date <- ymd(weather4$date)

# Rearrange columns using dplyr's select()
weather5 <- select(weather4, date, Events, CloudCover:WindDirDegrees)

# View the head of weather5
head(weather5)
```

```
##         date    Events CloudCover Max.Dew.PointF Max.Gust.SpeedMPH
## 1 2014-12-01      Rain          6             46                29
## 2 2014-12-10      Rain          8             45                29
## 3 2014-12-11 Rain-Snow          8             37                28
## 4 2014-12-12      Snow          7             28                21
## 5 2014-12-13                    5             28                23
## 6 2014-12-14                    4             29                20
##   Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 1           74                    30.45              64
## 2          100                    29.58              48
## 3           92                    29.81              39
## 4           85                    29.88              39
## 5           75                    29.86              42
## 6           82                    29.91              45
##   Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF Mean.Humidity
## 1                  10                22             40            63
## 2                  10                23             39            95
## 3                  10                21             31            87
## 4                  10                16             27            75
## 5                  10                17             26            65
## 6                  10                15             27            68
##   Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 1                     30.13               52                   10
## 2                      29.5               43                    3
## 3                     29.61               36                    7
## 4                     29.85               35                   10
## 5                     29.82               37                   10
## 6                     29.83               39                   10
##   Mean.Wind.SpeedMPH Min.DewpointF Min.Humidity Min.Sea.Level.PressureIn
## 1                 13            26           52                    30.01
## 2                 13            37           89                    29.43
## 3                 13            27           82                    29.44
## 4                 11            25           64                    29.81
```

```
## 5                 12           24          55                   29.78
## 6                 10           25          53                   29.78
##   Min.TemperatureF Min.VisibilityMiles PrecipitationIn WindDirDegrees
## 1               39                  10            0.01            268
## 2               38                   1            0.28            357
## 3               32                   1            0.02            230
## 4               31                   7               T            286
## 5               32                  10               T            298
## 6               33                  10            0.00            306
```

A closer look at column types

It's important for analysis that variables are coded appropriately. This is not yet the case with our weather data. Recall that functions such as as.numeric() and as.character() can be used to coerce variables into different types.

```
# View the structure of weather5
str(weather5)
```

```
## 'data.frame':    366 obs. of  23 variables:
##  $ date                     : Date, format: "2014-12-01" "2014-12-10" ...
##  $ Events                   : chr  "Rain" "Rain" "Rain-Snow" "Snow" ...
##  $ CloudCover               : chr  "6" "8" "8" "7" ...
##  $ Max.Dew.PointF           : chr  "46" "45" "37" "28" ...
##  $ Max.Gust.SpeedMPH        : chr  "29" "29" "28" "21" ...
##  $ Max.Humidity             : chr  "74" "100" "92" "85" ...
##  $ Max.Sea.Level.PressureIn : chr  "30.45" "29.58" "29.81" "29.88" ...
##  $ Max.TemperatureF         : chr  "64" "48" "39" "39" ...
##  $ Max.VisibilityMiles      : chr  "10" "10" "10" "10" ...
##  $ Max.Wind.SpeedMPH        : chr  "22" "23" "21" "16" ...
##  $ MeanDew.PointF           : chr  "40" "39" "31" "27" ...
##  $ Mean.Humidity            : chr  "63" "95" "87" "75" ...
##  $ Mean.Sea.Level.PressureIn: chr  "30.13" "29.5" "29.61" "29.85" ...
##  $ Mean.TemperatureF        : chr  "52" "43" "36" "35" ...
##  $ Mean.VisibilityMiles     : chr  "10" "3" "7" "10" ...
##  $ Mean.Wind.SpeedMPH       : chr  "13" "13" "13" "11" ...
##  $ Min.DewpointF            : chr  "26" "37" "27" "25" ...
##  $ Min.Humidity             : chr  "52" "89" "82" "64" ...
##  $ Min.Sea.Level.PressureIn : chr  "30.01" "29.43" "29.44" "29.81" ...
##  $ Min.TemperatureF         : chr  "39" "38" "32" "31" ...
##  $ Min.VisibilityMiles      : chr  "10" "1" "1" "7" ...
##  $ PrecipitationIn          : chr  "0.01" "0.28" "0.02" "T" ...
##  $ WindDirDegrees           : chr  "268" "357" "230" "286" ...
```

```
# Replace "T" with "0" (T = trace)
weather5$PrecipitationIn <- str_replace(weather5$PrecipitationIn,"T","0")

# Convert characters to numerics
weather6 <- mutate_at(weather5, vars(CloudCover:WindDirDegrees), funs(as.numeric))

# Look at result
str(weather6)
```

```
## 'data.frame':    366 obs. of  23 variables:
##  $ date                     : Date, format: "2014-12-01" "2014-12-10" ...
##  $ Events                   : chr  "Rain" "Rain" "Rain-Snow" "Snow" ...
##  $ CloudCover               : num  6 8 8 7 5 4 2 8 8 7 ...
```

```
##  $ Max.Dew.PointF           : num  46 45 37 28 28 29 33 42 46 34 ...
##  $ Max.Gust.SpeedMPH        : num  29 29 28 21 23 20 21 10 26 30 ...
##  $ Max.Humidity             : num  74 100 92 85 75 82 89 96 100 89 ...
##  $ Max.Sea.Level.PressureIn : num  30.4 29.6 29.8 29.9 29.9 ...
##  $ Max.TemperatureF         : num  64 48 39 39 42 45 42 44 49 44 ...
##  $ Max.VisibilityMiles      : num  10 10 10 10 10 10 10 10 10 10 ...
##  $ Max.Wind.SpeedMPH        : num  22 23 21 16 17 15 15 8 20 23 ...
##  $ MeanDew.PointF           : num  40 39 31 27 26 27 29 36 41 30 ...
##  $ Mean.Humidity            : num  63 95 87 75 65 68 75 85 85 73 ...
##  $ Mean.Sea.Level.PressureIn: num  30.1 29.5 29.6 29.9 29.8 ...
##  $ Mean.TemperatureF        : num  52 43 36 35 37 39 37 40 45 40 ...
##  $ Mean.VisibilityMiles     : num  10 3 7 10 10 10 10 9 6 10 ...
##  $ Mean.Wind.SpeedMPH       : num  13 13 13 11 12 10 6 4 11 14 ...
##  $ Min.DewpointF            : num  26 37 27 25 24 25 27 30 32 26 ...
##  $ Min.Humidity             : num  52 89 82 64 55 53 60 73 70 57 ...
##  $ Min.Sea.Level.PressureIn : num  30 29.4 29.4 29.8 29.8 ...
##  $ Min.TemperatureF         : num  39 38 32 31 32 33 32 35 41 36 ...
##  $ Min.VisibilityMiles      : num  10 1 1 7 10 10 10 5 1 10 ...
##  $ PrecipitationIn          : num  0.01 0.28 0.02 0 0 0 0 0 0.43 0.01 ...
##  $ WindDirDegrees           : num  268 357 230 286 298 306 324 79 311 281 ...
```

Extreme or missing data

Let's see where are the NA values

```
# Count missing values
sum(is.na(weather6))
```

```
## [1] 6
```

```
# Find missing values
summary(weather6)
```

```
##       date                Events            CloudCover    Max.Dew.PointF
##  Min.   :2014-12-01   Length:366         Min.   :0.000   Min.   :-6.00
##  1st Qu.:2015-03-02   Class :character   1st Qu.:3.000   1st Qu.:32.00
##  Median :2015-06-01   Mode  :character   Median :5.000   Median :47.50
##  Mean   :2015-06-01                      Mean   :4.708   Mean   :45.48
##  3rd Qu.:2015-08-31                      3rd Qu.:7.000   3rd Qu.:61.00
##  Max.   :2015-12-01                      Max.   :8.000   Max.   :75.00
##
##  Max.Gust.SpeedMPH  Max.Humidity     Max.Sea.Level.PressureIn
##  Min.   : 0.00     Min.   :  39.00   Min.   :29.58
##  1st Qu.:21.00     1st Qu.:  73.25   1st Qu.:30.00
##  Median :25.50     Median :  86.00   Median :30.14
##  Mean   :26.99     Mean   :  85.69   Mean   :30.16
##  3rd Qu.:31.25     3rd Qu.:  93.00   3rd Qu.:30.31
##  Max.   :94.00     Max.   :1000.00   Max.   :30.88
##  NA's   :6
##  Max.TemperatureF Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF
##  Min.   :18.00    Min.   : 2.000      Min.   : 8.00     Min.   :-11.00
##  1st Qu.:42.00    1st Qu.:10.000      1st Qu.:16.00     1st Qu.: 24.00
##  Median :60.00    Median :10.000      Median :20.00     Median : 41.00
##  Mean   :58.93    Mean   : 9.907      Mean   :20.62     Mean   : 38.96
##  3rd Qu.:76.00    3rd Qu.:10.000      3rd Qu.:24.00     3rd Qu.: 56.00
##  Max.   :96.00    Max.   :10.000      Max.   :38.00     Max.   : 71.00
##
```

```
##  Mean.Humidity   Mean.Sea.Level.PressureIn Mean.TemperatureF
##  Min.   :28.00   Min.   :29.49             Min.   : 8.00
##  1st Qu.:56.00   1st Qu.:29.87             1st Qu.:36.25
##  Median :66.00   Median :30.03             Median :53.50
##  Mean   :66.02   Mean   :30.04             Mean   :51.40
##  3rd Qu.:76.75   3rd Qu.:30.19             3rd Qu.:68.00
##  Max.   :98.00   Max.   :30.77             Max.   :84.00
##
##  Mean.VisibilityMiles Mean.Wind.SpeedMPH Min.DewpointF     Min.Humidity
##  Min.   :-1.000       Min.   : 4.00      Min.   :-18.00   Min.   :16.00
##  1st Qu.: 8.000       1st Qu.: 8.00      1st Qu.: 16.25   1st Qu.:35.00
##  Median :10.000       Median :10.00      Median : 35.00   Median :46.00
##  Mean   : 8.861       Mean   :10.68      Mean   : 32.25   Mean   :48.31
##  3rd Qu.:10.000       3rd Qu.:13.00      3rd Qu.: 51.00   3rd Qu.:60.00
##  Max.   :10.000       Max.   :22.00      Max.   : 68.00   Max.   :96.00
##
##  Min.Sea.Level.PressureIn Min.TemperatureF Min.VisibilityMiles
##  Min.   :29.16            Min.   :-3.00    Min.   : 0.000
##  1st Qu.:29.76            1st Qu.:30.00    1st Qu.: 2.000
##  Median :29.94            Median :46.00    Median :10.000
##  Mean   :29.93            Mean   :43.33    Mean   : 6.716
##  3rd Qu.:30.09            3rd Qu.:60.00    3rd Qu.:10.000
##  Max.   :30.64            Max.   :74.00    Max.   :10.000
##
##  PrecipitationIn  WindDirDegrees
##  Min.   :0.0000   Min.   :  1.0
##  1st Qu.:0.0000   1st Qu.:113.0
##  Median :0.0000   Median :222.0
##  Mean   :0.1016   Mean   :200.1
##  3rd Qu.:0.0400   3rd Qu.:275.0
##  Max.   :2.9000   Max.   :360.0
##
```

```
# Find indices of NAs in Max.Gust.SpeedMPH
ind <- which(is.na(weather6$Max.Gust.SpeedMPH))

# Look at the full rows for records missing Max.Gust.SpeedMPH
weather6[ind, ]
```

```
##           date Events CloudCover Max.Dew.PointF Max.Gust.SpeedMPH
## 161 2015-05-18    Fog          6             52                NA
## 205 2015-06-03                 7             48                NA
## 273 2015-08-08                 4             61                NA
## 275 2015-09-01                 1             63                NA
## 308 2015-10-12                 0             56                NA
## 358 2015-11-03                 1             44                NA
##     Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 161          100                    30.30               58
## 205           93                    30.31               56
## 273           87                    30.02               76
## 275           78                    30.06               79
## 308           89                    29.86               76
## 358           82                    30.25               73
##     Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF Mean.Humidity
## 161                  10                16             48            79
```

```
## 205                  10               14             45             82
## 273                  10               14             57             68
## 275                  10               15             62             65
## 308                  10               15             51             65
## 358                  10               16             42             57
##     Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 161                     30.23               54                    8
## 205                     30.24               52                   10
## 273                     29.99               69                   10
## 275                     30.02               74                   10
## 308                     29.80               64                   10
## 358                     30.13               60                   10
##     Mean.Wind.SpeedMPH Min.DewpointF Min.Humidity Min.Sea.Level.PressureIn
## 161                 10           43           57                    30.12
## 205                  7           43           71                    30.19
## 273                  6           54           49                    29.95
## 275                  9           59           52                    29.96
## 308                  8           48           41                    29.74
## 358                  8           40           31                    30.06
##     Min.TemperatureF Min.VisibilityMiles PrecipitationIn WindDirDegrees
## 161               49                   0               0             72
## 205               47                  10               0             90
## 273               61                  10               0             45
## 275               69                  10               0             54
## 308               51                  10               0            199
## 358               47                  10               0            281
```

Obvious error

There is a humidity =1000 that needs to be replaced to 100

```
# Review distributions for all variables
summary(weather6)
```

```
##       date                Events            CloudCover    Max.Dew.PointF
##  Min.   :2014-12-01   Length:366         Min.   :0.000   Min.   :-6.00
##  1st Qu.:2015-03-02   Class :character   1st Qu.:3.000   1st Qu.:32.00
##  Median :2015-06-01   Mode  :character   Median :5.000   Median :47.50
##  Mean   :2015-06-01                      Mean   :4.708   Mean   :45.48
##  3rd Qu.:2015-08-31                      3rd Qu.:7.000   3rd Qu.:61.00
##  Max.   :2015-12-01                      Max.   :8.000   Max.   :75.00
##
##  Max.Gust.SpeedMPH  Max.Humidity     Max.Sea.Level.PressureIn
##  Min.   : 0.00     Min.   :  39.00   Min.   :29.58
##  1st Qu.:21.00     1st Qu.:  73.25   1st Qu.:30.00
##  Median :25.50     Median :  86.00   Median :30.14
##  Mean   :26.99     Mean   :  85.69   Mean   :30.16
##  3rd Qu.:31.25     3rd Qu.:  93.00   3rd Qu.:30.31
##  Max.   :94.00     Max.   :1000.00   Max.   :30.88
##  NA's   :6
##  Max.TemperatureF Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF
##  Min.   :18.00    Min.   : 2.000      Min.   : 8.00     Min.   :-11.00
##  1st Qu.:42.00    1st Qu.:10.000      1st Qu.:16.00     1st Qu.: 24.00
##  Median :60.00    Median :10.000      Median :20.00     Median : 41.00
##  Mean   :58.93    Mean   : 9.907      Mean   :20.62     Mean   : 38.96
##  3rd Qu.:76.00    3rd Qu.:10.000      3rd Qu.:24.00     3rd Qu.: 56.00
```

```
##  Max.   :96.00    Max.   :10.000      Max.   :38.00     Max.   : 71.00
##
##  Mean.Humidity   Mean.Sea.Level.PressureIn Mean.TemperatureF
##  Min.   :28.00   Min.   :29.49             Min.   : 8.00
##  1st Qu.:56.00   1st Qu.:29.87             1st Qu.:36.25
##  Median :66.00   Median :30.03             Median :53.50
##  Mean   :66.02   Mean   :30.04             Mean   :51.40
##  3rd Qu.:76.75   3rd Qu.:30.19             3rd Qu.:68.00
##  Max.   :98.00   Max.   :30.77             Max.   :84.00
##
##  Mean.VisibilityMiles Mean.Wind.SpeedMPH Min.DewpointF     Min.Humidity
##  Min.   :-1.000       Min.   : 4.00      Min.   :-18.00   Min.   :16.00
##  1st Qu.: 8.000       1st Qu.: 8.00      1st Qu.: 16.25   1st Qu.:35.00
##  Median :10.000       Median :10.00      Median : 35.00   Median :46.00
##  Mean   : 8.861       Mean   :10.68      Mean   : 32.25   Mean   :48.31
##  3rd Qu.:10.000       3rd Qu.:13.00      3rd Qu.: 51.00   3rd Qu.:60.00
##  Max.   :10.000       Max.   :22.00      Max.   : 68.00   Max.   :96.00
##
##  Min.Sea.Level.PressureIn Min.TemperatureF Min.VisibilityMiles
##  Min.   :29.16            Min.   :-3.00    Min.   : 0.000
##  1st Qu.:29.76            1st Qu.:30.00    1st Qu.: 2.000
##  Median :29.94            Median :46.00    Median :10.000
##  Mean   :29.93            Mean   :43.33    Mean   : 6.716
##  3rd Qu.:30.09            3rd Qu.:60.00    3rd Qu.:10.000
##  Max.   :30.64            Max.   :74.00    Max.   :10.000
##
##  PrecipitationIn  WindDirDegrees
##  Min.   :0.0000   Min.   :  1.0
##  1st Qu.:0.0000   1st Qu.:113.0
##  Median :0.0000   Median :222.0
##  Mean   :0.1016   Mean   :200.1
##  3rd Qu.:0.0400   3rd Qu.:275.0
##  Max.   :2.9000   Max.   :360.0
##
```

```
# Find row with Max.Humidity of 1000
ind <- which(weather6$Max.Humidity==1000)

# Look at the data for that day
weather6[ind, ]
```

```
##           date                Events CloudCover Max.Dew.PointF
## 135 2015-04-21 Fog-Rain-Thunderstorm          6             57
##     Max.Gust.SpeedMPH Max.Humidity Max.Sea.Level.PressureIn
## 135                94         1000                   29.75
##     Max.TemperatureF Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF
## 135               65                  10                20             49
##     Mean.Humidity Mean.Sea.Level.PressureIn Mean.TemperatureF
## 135            71                     29.6                56
##     Mean.VisibilityMiles Mean.Wind.SpeedMPH Min.DewpointF Min.Humidity
## 135                    5                 10            36           42
##     Min.Sea.Level.PressureIn Min.TemperatureF Min.VisibilityMiles
## 135                    29.53               46                   0
##     PrecipitationIn WindDirDegrees
## 135            0.54            184
```

```
# Change 1000 to 100
weather6$Max.Humidity[ind] <- 100
```

For Visibility happens to be a -1 also, that needs to be replaced by 10

```
# Look at summary of Mean.VisibilityMiles
summary(weather6$Mean.VisibilityMiles)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.000   8.000  10.000   8.861  10.000  10.000
```

```
# Get index of row with -1 value
ind <- which(weather6$Mean.VisibilityMiles==-1)

# Look at full row
weather6[ind,]
```

```
##           date Events CloudCover Max.Dew.PointF Max.Gust.SpeedMPH
## 192 2015-06-18                 5             54                23
##     Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 192           72                    30.14               76
##     Max.VisibilityMiles Max.Wind.SpeedMPH MeanDew.PointF Mean.Humidity
## 192                  10                17             49            59
##     Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 192                     30.04                67                   -1
##     Mean.Wind.SpeedMPH Min.DewpointF Min.Humidity Min.Sea.Level.PressureIn
## 192                 10            45           46                    29.93
##     Min.TemperatureF Min.VisibilityMiles PrecipitationIn WindDirDegrees
## 192               57                  10               0            189
```

```
# Set Mean.VisibilityMiles to the appropriate value
weather6$Mean.VisibilityMiles[ind] <- 10
```

Finishing touches

Before officially calling our weather data clean, we want to put a couple of finishing touches on the data. These are a bit more subjective and may not be necessary for analysis, but they will make the data easier for others to interpret, which is generally a good thing.

There are a number of stylistic conventions in the R language. Depending on who you ask, these conventions may vary. Because the period (.) has special meaning in certain situations, we generally recommend using underscores (_) to separate words in variable names. We also prefer all lowercase letters so that no one has to remember which letters are uppercase or lowercase.

Finally, the events column (renamed to be all lowercase in the first instruction) contains an empty string ("") for any day on which there was no significant weather event such as rain, fog, a thunderstorm, etc. However, if it's the first time you're seeing these data, it may not be obvious that this is the case, so it's best for us to be explicit and replace the empty strings with something more meaningful.

```
# Clean up column names
names(weather6) <- tolower(colnames(weather6))

# Print the first 6 rows of weather6
head(weather6,n=6)
```

```
##         date    events cloudcover max.dew.pointf max.gust.speedmph
## 1 2014-12-01      Rain          6             46                29
## 2 2014-12-10      Rain          8             45                29
## 3 2014-12-11 Rain-Snow          8             37                28
```

```
## 4 2014-12-12      Snow          7             28               21
## 5 2014-12-13                    5             28               23
## 6 2014-12-14                    4             29               20
##   max.humidity max.sea.level.pressurein max.temperaturef
## 1           74                    30.45              64
## 2          100                    29.58              48
## 3           92                    29.81              39
## 4           85                    29.88              39
## 5           75                    29.86              42
## 6           82                    29.91              45
##   max.visibilitymiles max.wind.speedmph meandew.pointf mean.humidity
## 1                  10                22            40            63
## 2                  10                23            39            95
## 3                  10                21            31            87
## 4                  10                16            27            75
## 5                  10                17            26            65
## 6                  10                15            27            68
##   mean.sea.level.pressurein mean.temperaturef mean.visibilitymiles
## 1                     30.13               52                   10
## 2                     29.50               43                    3
## 3                     29.61               36                    7
## 4                     29.85               35                   10
## 5                     29.82               37                   10
## 6                     29.83               39                   10
##   mean.wind.speedmph min.dewpointf min.humidity min.sea.level.pressurein
## 1                 13           26           52                    30.01
## 2                 13           37           89                    29.43
## 3                 13           27           82                    29.44
## 4                 11           25           64                    29.81
## 5                 12           24           55                    29.78
## 6                 10           25           53                    29.78
##   min.temperaturef min.visibilitymiles precipitationin winddirdegrees
## 1               39                  10            0.01            268
## 2               38                   1            0.28            357
## 3               32                   1            0.02            230
## 4               31                   7            0.00            286
## 5               32                  10            0.00            298
## 6               33                  10            0.00            306
```