How can we deal with missing data in R?

2017-02-16

What are NAs?

Basic vocabulary:

- `NA` stands for *not available*
- `NaN` stands for *not a number*
- `Inf` stands for *infinite*
- `NULL` stands for a *null object*

But really… what are NAs?

- NAs are **placeholders**

Some R quirks

Most important thing to remember: NAs are **contagious**!

```
NA > 1
```

```
## [1] NA
```

```
NA/3
```

```
## [1] NA
```

```
NA == NA # This last one is important!
```

```
## [1] NA
```

Some R quirks (2)

Which items are missing? Which are not?

```
vec <- c("R", "ladies", NA, "Paris")

# Won't work
vec == NA
vec != NA
```

```
## [1] NA NA NA NA
## [1] NA NA NA NA
```

```
# Yey!
is.na(vec)
!is.na(vec)
```

```
## [1] FALSE FALSE  TRUE FALSE
## [1]  TRUE  TRUE FALSE  TRUE
```

Some R quirks (3)

```
vect <- c(2, 2, 2, NA)

sum(vect)
sum(vect, na.rm = TRUE)
```

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

How can we deal with missing values?

Ignore the missing values and work only with complete cases

- Lose key information, bias your analysis
- Values may be missing for a reason!

Impute the missing values

- Lots of methods!
- But other shortcomings

Before treatment… exploration

- How many are there?
- Where are the missing values?
- Are they related?
- Can I make assumptions to help with the imputation?

Simple example with `airquality`

```
# Libraries

library(dplyr)
library(zoo)   # locf imputation
library(VIM)   # visualization

# Data
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Some info on the data

```
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

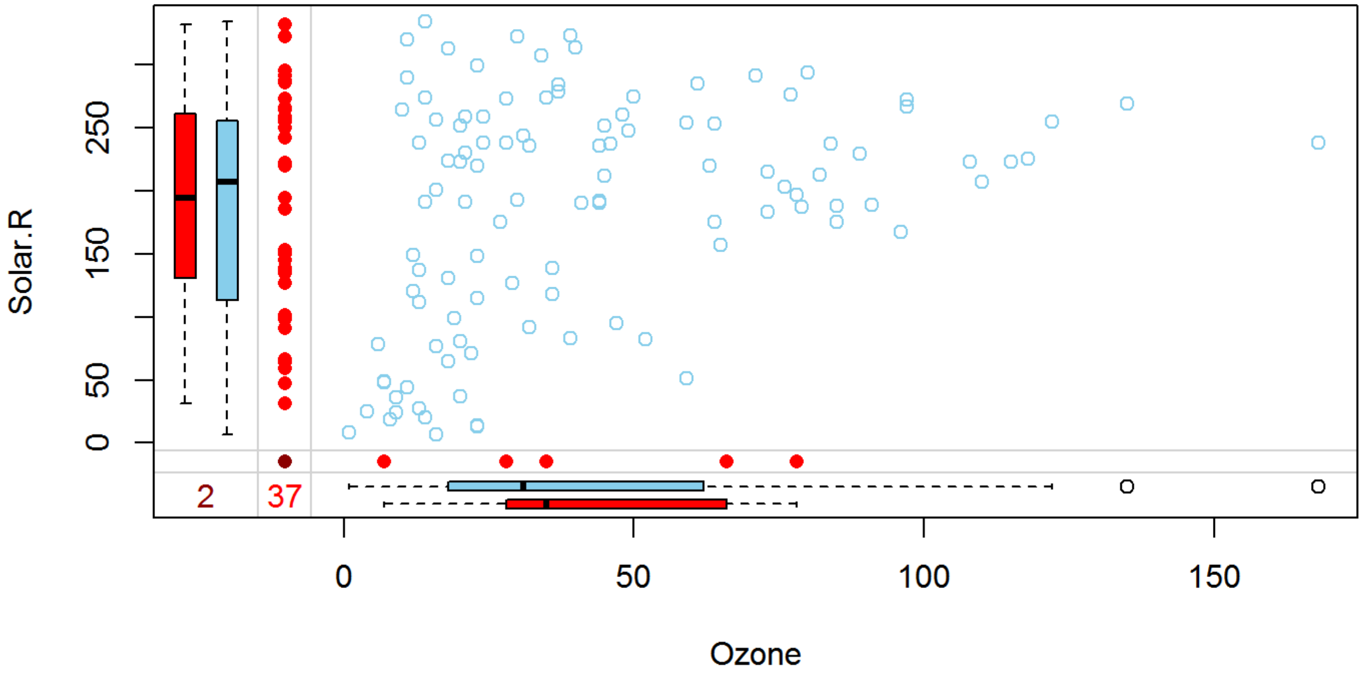
Exploration

```
airquality %>%
  group_by(Month) %>%
  summarise(
    miss_ozone = sum(is.na(Ozone)),
    miss_solar = sum(is.na(Solar.R)),
    miss_both  = sum(is.na(Ozone) & is.na(Solar.R)),
    n_month    = n()
    )
```

```
## # A tibble: 5 × 5
##   Month miss_ozone miss_solar miss_both n_month
##   <int>      <int>      <int>     <int>   <int>
## 1     5          5          4         2      31
## 2     6         21          0         0      30
## 3     7          5          0         0      31
## 4     8          5          3         0      31
## 5     9          1          0         0      30
```

Exploration (2)

```
airquality %>%
  select(Ozone, Solar.R) %>%
  marginplot()
```



Imputation

```
# mean imputation with dplyr
airquality %>%
  mutate_at(
    .cols = vars(Solar.R, Ozone),
    .funs = funs(ifelse(is.na(.), mean(., na.rm = T), .))
    ) %>%
  head()
```

```
##      Ozone  Solar.R Wind Temp Month Day
## 1 41.00000 190.0000  7.4   67     5   1
## 2 36.00000 118.0000  8.0   72     5   2
## 3 12.00000 149.0000 12.6   74     5   3
## 4 18.00000 313.0000 11.5   62     5   4
## 5 42.12931 185.9315 14.3   56     5   5
## 6 28.00000 185.9315 14.9   66     5   6
```

Imputation (2)

```
# mean imputation with zoo
airquality %>%
  na.aggregate() %>%
  head()
```

```
##      Ozone  Solar.R Wind Temp Month Day
## 1 41.00000 190.0000  7.4   67     5   1
## 2 36.00000 118.0000  8.0   72     5   2
## 3 12.00000 149.0000 12.6   74     5   3
## 4 18.00000 313.0000 11.5   62     5   4
## 5 42.12931 185.9315 14.3   56     5   5
## 6 28.00000 185.9315 14.9   66     5   6
```

Imputation (3)

```
# locf imputation with zoo
airquality %>%
  na.locf() %>%
  head()
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    18     313 14.3   56     5   5
## 6    28     313 14.9   66     5   6
```

Links & packages

More on visualization with VIM: https://cran.r-project.org/web/packages/VIMGUI/vignettes/VIM-Imputation.pdf)

"Tagged" missing values (importing from STATA and SPSS): http://haven.tidyverse.org/reference/tagged_na.html

Summary of different R packages for imputation https://www.rstudio.com/rviews/2016/11/30/missing-values-data-science-and-r/)

More on imputation methods (in French & with some math): http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-idm.pdf