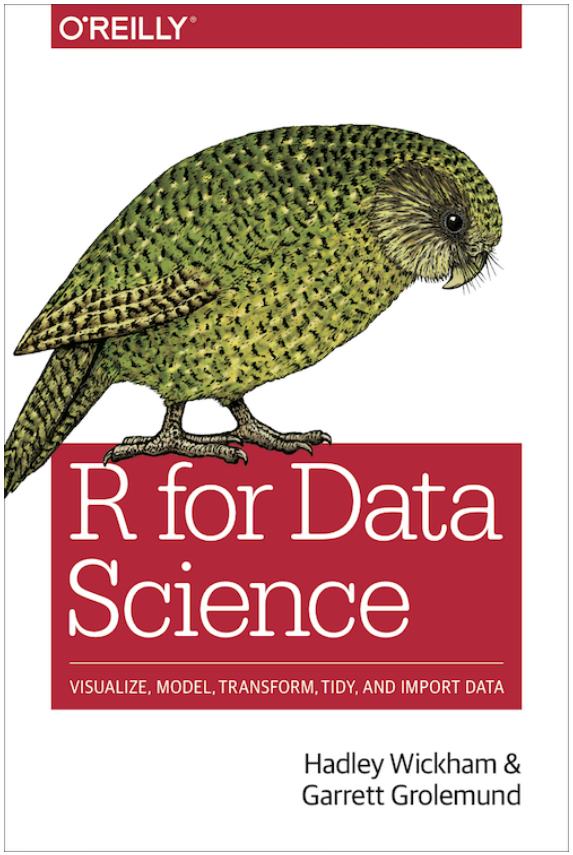
Socio-Informatics 348

Data Transformation
Missing Values

Dr Lisa Martin

Department of Information Science
Stellenbosch University

Today's Reading



R for Data Science, Chapter 18

Explicit Missing Values

```
treatment <- tribble(
  ~person,           ~treatment, ~response,
  "Derrick Whitmore", 1,          7,
  NA,                 2,          10,
  NA,                 3,          NA,
  "Katherine Burke",  1,          4
)
```

Explicit Missing Values

Last observation carried forward

- When missing values represent repeated measurements

```
treatment |>
  fill(everything())
#> # A tibble: 4 × 3
#>   person           treatment response
#>   <chr>                <dbl>    <dbl>
#> 1 Derrick Whitmore         1        7
#> 2 Derrick Whitmore         2       10
#> 3 Derrick Whitmore         3       10
#> 4 Katherine Burke          1        4
```

- can enter individual variables
- `fill(direction = "down")` or `fill(direction = "up")`...

Explicit Missing Values

Fixed values

- Sometimes missing values represent a specific value, like 0
- Item might be skipped in a survey instead of answered with zero, e.g. number of children
- Software/survey might also have a default missing value, e.g. -99

```
x <- c(1, 4, 5, 7, NA)
coalesce(x, 0)
#> [1] 1 4 5 7 0
```

```
x <- c(1, 4, 5, 7, -99)
na_if(x, -99)
#> [1]  1  4  5  7 NA
```

NA vs NaN

A special type of missing value:

- **N**ot **a** **N**umber
- Generally behaves just like NA
- Use `is.nan()` to test for NaN

```
x <- c(NA, NaN)
x * 10
#> [1]  NA NaN
x == 1
#> [1] NA NA
is.na(x)
#> [1] TRUE TRUE
```

NA vs NaN

- Generally encountered when performing a mathematical operation that has an indeterminate result

```
0 / 0
#> [1] NaN
0 * Inf
#> [1] NaN
Inf - Inf
#> [1] NaN
sqrt(-1)
#> Warning in sqrt(-1): NaNs produced
#> [1] NaN
```

Implicit Missing Values

An explicit missing value is the presence of an absence.
An implicit missing value is the absence of a presence.

```
stocks <- tibble(
  year  = c(2020, 2020, 2020, 2020, 2021, 2021, 2021),
  qtr   = c(   1,    2,    3,    4,    2,    3,    4),
  price = c(1.88, 0.59, 0.35,   NA, 0.92, 0.17, 2.66)
)
```

- Sometimes, we want the missing values to be explicit, e.g. for analysis or visualisation
- How do we move from implicit to explicit missing values?

Implicit Missing Values

Pivoting

```
stocks |>
  pivot_wider(
    names_from = qtr,
    values_from = price
  )
#> # A tibble: 2 × 5
#>    year   `1`   `2`   `3`   `4`
#>   <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1  2020  1.88  0.59  0.35 NA
#> 2  2021 NA     0.92  0.17  2.66
```

- Making data wider can make implicit missing values explicit because every combination of the rows and new columns must have some value.

Implicit Missing Values

Complete

```
stocks |>
  complete(year, qtr)
#> # A tibble: 8 × 3
#>    year   qtr price
#>   <dbl> <dbl> <dbl>
#> 1  2020     1  1.88
#> 2  2020     2  0.59
#> 3  2020     3  0.35
#> 4  2020     4 NA
#> 5  2021     1 NA
#> 6  2021     2  0.92
#> # ℹ 2 more rows
```

- Providing combinations of variables that should exist.

Implicit Missing Values

Complete

```
stocks |>
  complete(year = 2019:2021, qtr)
#> # A tibble: 12 × 3
#>    year   qtr price
#>   <dbl> <dbl> <dbl>
#> 1  2019     1 NA
#> 2  2019     2 NA
#> 3  2019     3 NA
#> 4  2019     4 NA
#> 5  2020     1  1.88
#> 6  2020     2  0.59
#> # ℹ 6 more rows
```

- Able to provide additional/custom values of variables

Implicit Missing Values

Joins / Anti-joins

- Anti-join between x and y shows rows in x that do not have a match in y
- Which destination airports in the flights data do not have a matching airport in the airports data?

```
flights |>
  distinct(faa = dest) |>
  anti_join(airports)
#> Joining with `by = join_by(faa)`
#> # A tibble: 4 × 1
#>   faa
#>   <chr>
#> 1 BQN
#> 2 SJU
#> 3 STT
#> 4 PSE
```

Implicit Missing Values

Joins / Anti-joins

- Which tail number in the flights data do not have a matching plane in the planes data?

```
flights |>
  distinct(tailnum) |>
  anti_join(planes)
#> Joining with `by = join_by(tailnum)`
#> # A tibble: 722 × 1
#>   tailnum
#>   <chr>
#> 1 N3ALAA
#> 2 N3DUAA
#> 3 N542MQ
#> 4 N730MQ
#> 5 N9EAMQ
#> 6 N532UA
#> # ℹ 716 more rows
```

Implicit Missing Values

Joins / Anti-joins

- We were not aware of these implicit missing values until we performed the anti-join
- These were simply missing rows in the data, not explicit NA values

Factors and Empty Groups

- A final type of missingness: An empty group
- A factor level that has no observations

```
health <- tibble(
  name   = c("Ikaia", "Oletta", "Leriah", "Dashay", "Tresaun"),
  smoker = factor(c("no", "no", "no", "no", "no"), levels = c("yes", "no")),
  age    = c(34, 88, 75, 47, 56),
)
```

Factors and Empty Groups

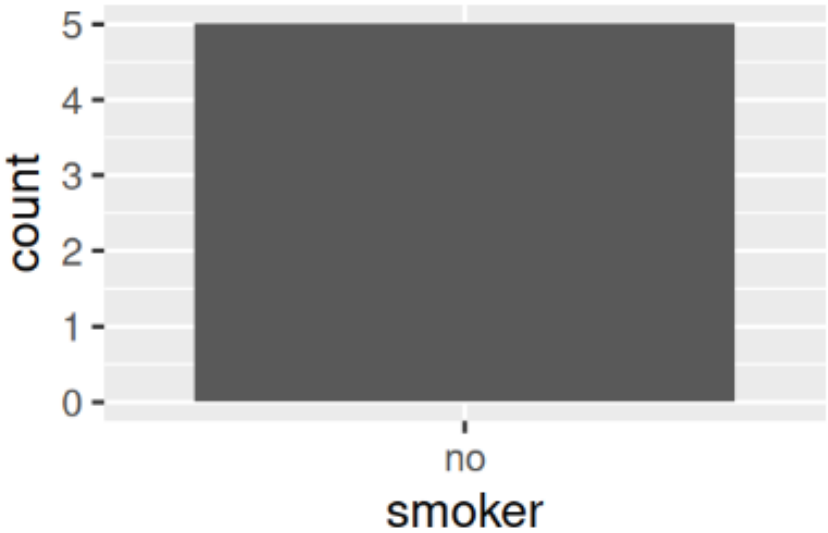
```
health |> count(smoker)
#> # A tibble: 1 × 2
#>   smoker     n
#>   <fct>  <int>
#> 1 no         5
```

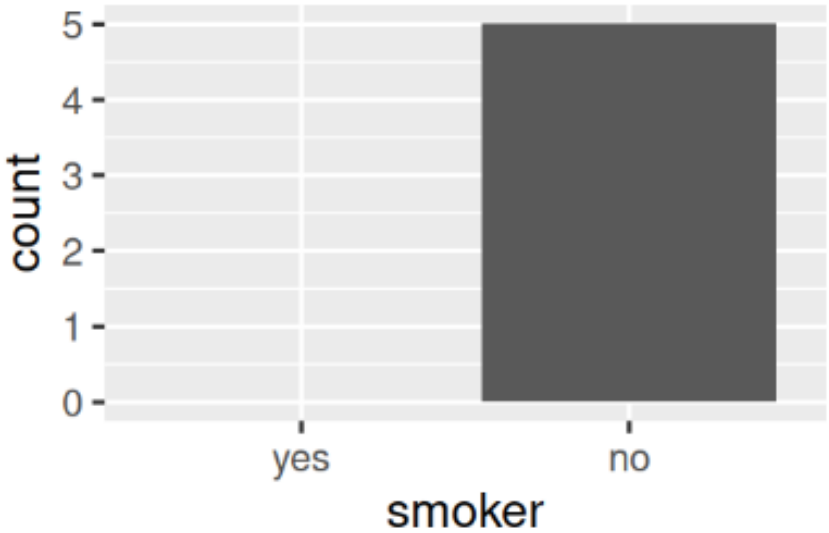
```
health |> count(smoker, .drop = FALSE)
#> # A tibble: 2 × 2
#>   smoker     n
#>   <fct>  <int>
#> 1 yes        0
#> 2 no         5
```

Factors and Empty Groups

```r
ggplot(health, aes(x = smoker)) +
  geom_bar() +
  scale_x_discrete()

ggplot(health, aes(x = smoker)) +
  geom_bar() +
  scale_x_discrete(drop = FALSE)
```





Factors and Empty Groups

```
health |>
  group_by(smoker, .drop = FALSE) |>
  summarize(
    n = n(),
    mean_age = mean(age),
    min_age = min(age),
    max_age = max(age),
    sd_age = sd(age)
  )
#> # A tibble: 2 × 6
#>   smoker     n mean_age min_age max_age sd_age
#>   <fct>  <int>    <dbl>   <dbl>   <dbl>  <dbl>
#> 1 yes        0      NaN     Inf    -Inf   NA
#> 2 no         5       60      34      88   21.6
```

Factors and Empty Groups

- Get 'cleaner' NAs by using `complete` instead of `.drop = FALSE`

```
health |>
  group_by(smoker) |>
  summarize(
    n = n(),
    mean_age = mean(age),
    min_age = min(age),
    max_age = max(age),
    sd_age = sd(age)
  ) |>
  complete(smoker)
#> # A tibble: 2 × 6
#>   smoker     n mean_age min_age max_age sd_age
#>   <fct>  <int>    <dbl>   <dbl>   <dbl>  <dbl>
#> 1 yes       NA       NA      NA      NA   NA
#> 2 no         5       60      34      88   21.6
```