

Lab 04: CS631

Working with Tidy Data

Alison Hill





Let's review

Data wrangling to date!

From `dplyr`:

- `filter`
- `arrange`
- `mutate`
- `group_by`
- `summarize`
- `glimpse`
- `distinct`
- `count`
- `tally`
- `pull`
- `top_n`

Let's add from `dplyr`:

- `select`
- `rename`
- `recode`
- `case_when`

From `tidyr`:

- `gather`
- `separate`
- `spread`
- `unite`

Plus 1 other package:

- `skimr::skim`

Un-tidy cakes

```
# A tibble: 2 x 4          # A tibble: 2 x 4
  series challenge    cake pie_tart  series challenge    cake pie_tart
<fct>   <chr>      <dbl>   <dbl>  <fct>   <chr>      <dbl>   <dbl>
1 1      showstopper    5       5 1 3      showstopper   12      17
2 1      signature     12      4 2 3      signature     24      12

# A tibble: 2 x 4          # A tibble: 2 x 4
  series challenge    cake pie_tart  series challenge    cake pie_tart
<fct>   <chr>      <dbl>   <dbl>  <fct>   <chr>      <dbl>   <dbl>
1 2      showstopper    8      17 1 4      showstopper   27       9
2 2      signature     21      7 2 4      signature     11      15
```

Still un-tidy cakes

```
cakes_untidy %>%  
  bind_rows()
```

```
# A tibble: 16 x 4  
  series challenge    cake pie_tart  
  <fct>   <chr>      <dbl>   <dbl>  
1 1      showstopper    5       5  
2 1      signature    12       4  
3 2      showstopper    8      17  
4 2      signature    21       7  
5 3      showstopper    12      17  
6 3      signature    24      12  
7 4      showstopper    27       9  
8 4      signature    11      15  
9 5      showstopper    20       6  
10 5     signature     4       7  
11 6     showstopper    12       0  
12 6     signature    20      17  
13 7     showstopper    19       3  
14 7     signature    11      10  
15 8     showstopper    26      12  
16 8     signature    21       8
```

Finally tidy cakes

```
cakes_tidy <- cakes_untidy %>%  
  gather(bake_type, num_bakes, cake:pie_tart,  
         factor_key = TRUE) %>%  
  arrange(series)  
cakes_tidy
```

```
# A tibble: 32 x 4  
  series challenge  bake_type num_bakes  
  <fct>   <chr>      <fct>      <dbl>  
1 1      showstopper cake          5  
2 1      signature  cake         12  
3 1      showstopper pie_tart       5  
4 1      signature  pie_tart       4  
5 2      showstopper cake          8  
6 2      signature  cake         21  
7 2      showstopper pie_tart      17  
8 2      signature  pie_tart       7  
9 3      showstopper cake         12  
10 3     signature  cake         24  
# ... with 22 more rows
```

Know Your Tidy Data

```
glimpse(cakes_tidy)
```

```
Observations: 32
```

```
Variables: 4
```

```
$ series      <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5,  
$ challenge   <chr> "showstopper", "signature", "showstopper", "signature", "  
$ bake_type   <fct> cake, cake, pie_tart, pie_tart, cake, cake, pie_tart, pie  
$ num_bakes   <dbl> 5, 12, 5, 4, 8, 21, 17, 7, 12, 24, 17, 12, 27, 11, 9, 15,
```



```
library(skimr)
skim(cakes_tidy)
```

Table: Data summary

Name cakes_tidy Number of rows 32
Number of columns 4

Column type frequency:
character 1
factor 2
numeric 1

Group variables None

Variable type: character

skim_variable n_missing complete_rate min max empty n_unique whitespace

challenge 0 1 9 11 0 2 0

```
skim(cakes_tidy) %>%  
  summary()
```

Table: Data summary

Name cakes_tidy Number of rows 32
Number of columns 4

Column type frequency:
character 1
factor 2
numeric 1

Group variables None

Benefits of Tidy Data

```
cakes_tidy %>%  
  count(challenge, bake_type, wt = num_bakes, sort = TRUE)
```

```
# A tibble: 4 x 3  
  challenge    bake_type      n  
  <chr>        <fct>    <dbl>  
1 showstopper cake        129  
2 signature   cake        124  
3 signature   pie_tart     80  
4 showstopper pie_tart     69
```

```
cakes_tidy %>%  
  count(series, bake_type, wt = num_bakes)
```

```
# A tibble: 16 x 3  
  series bake_type      n  
  <fct>   <fct>     <dbl>  
1 1      cake      17  
2 1      pie_tart    9  
3 2      cake      29  
4 2      pie_tart   24  
5 3      cake      36  
6 3      pie_tart   29  
7 4      cake      38  
8 4      pie_tart   24  
9 5      cake      24  
10 5     pie_tart   13  
11 6      cake      32  
12 6     pie_tart   17  
13 7      cake      30  
14 7     pie_tart   13  
15 8      cake      47  
16 8     pie_tart   20
```

```
library(skimr)
cakes_tidy %>%
  group_by(bake_type) %>%
  select_if(is.numeric) %>%
  skim()
```

Table: Data summary

Name Piped data Number of rows 32
Number of columns 2

Column type frequency:
numeric 1

Group variables bake_type

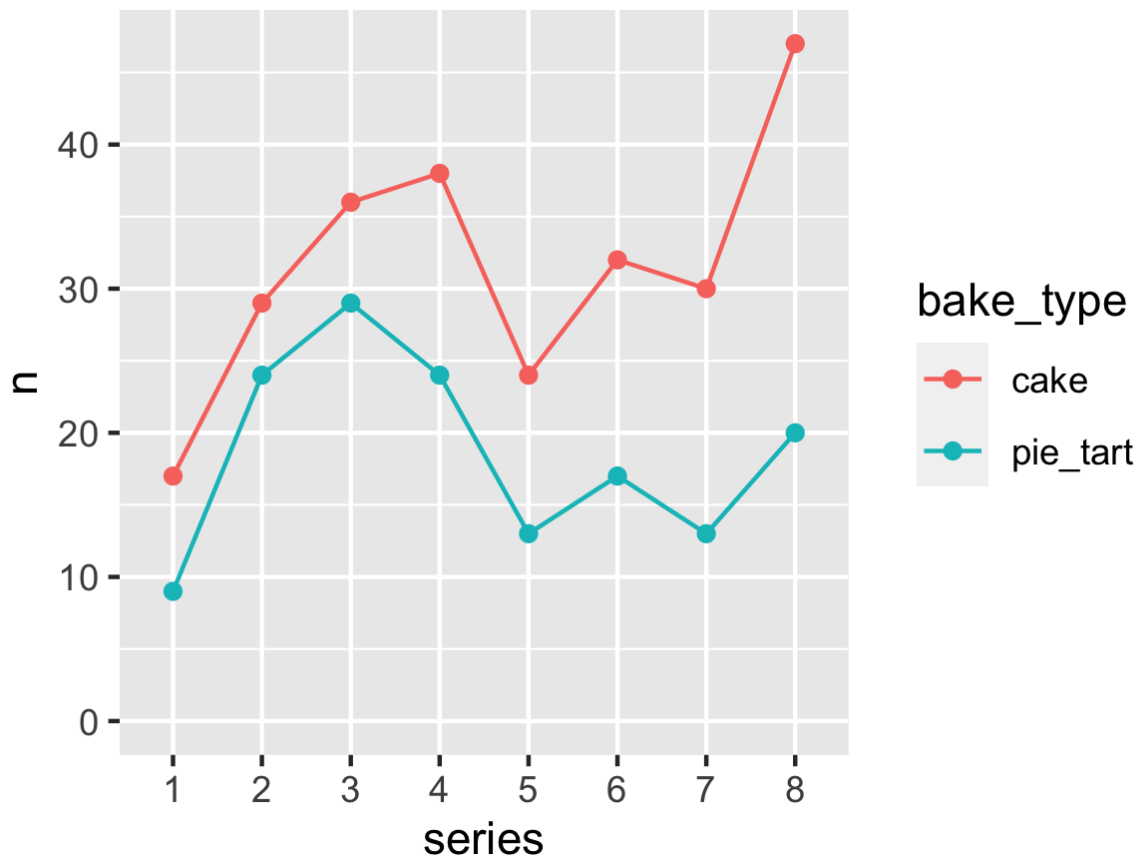
Variable type: numeric

skim_variable bake_type n_missing complete_rate mean sd p0 p25 p50 p75
p100 hist

```
cakes_by_series <- cakes_tidy %>%  
  count(series, bake_type, wt = num_bakes)  
cakes_by_series
```

```
# A tibble: 16 x 3  
  series bake_type      n  
  <fct>  <fct>    <dbl>  
1 1      cake      17  
2 1      pie_tart    9  
3 2      cake      29  
4 2      pie_tart   24  
5 3      cake      36  
6 3      pie_tart   29  
7 4      cake      38  
8 4      pie_tart   24  
9 5      cake      24  
10 5      pie_tart   13  
11 6      cake      32  
12 6      pie_tart   17  
13 7      cake      30  
14 7      pie_tart   13  
15 8      cake      47  
16 8      pie_tart   20
```

```
ggplot(cakes_by_series, aes(x = series, y = n,  
                             color = bake_type,  
                             group = bake_type)) +  
  
  geom_point() +  
  geom_line() +  
  expand_limits(y = 0)
```



You have 2 challenges today!

Described here Reference lab here



Tidy Data:

<http://r4ds.had.co.nz/tidy-data.html>

<http://moderndive.com/4-tidy.html>

<http://vita.had.co.nz/papers/tidy-data.html>

<https://github.com/jennybc/lotr-tidy#readme>