Tidying and organizing data: extras

Packages

```
library(tidyverse)
```

The pig feed data again

```
my_url <- "http://ritsokiguess.site/STAC32/pigs1.txt"
pigs <- read_table(my_url)
pigs
```

pig	feed1	feed2	feed3	feed4
1	60.8	68.7	92.6	87.9
2	57.0	67.7	92.1	84.2
3	65.0	74.0	90.2	83.1
4	58.6	66.3	96.5	85.7
5	61.7	69.8	99.1	90.3

Make longer (as before)

```
pigs
```

pig	feed1	feed2	feed3	feed4
1	60.8	68.7	92.6	87.9
2	57.0	67.7	92.1	84.2
3	65.0	74.0	90.2	83.1
4	58.6	66.3	96.5	85.7
5	61.7	69.8	99.1	90.3

```
pigs %>% pivot_longer(-pig, names_to="feed",
                      values_to="weight") -> pigs_longer
```

Make wider two ways 1/2

`pivot_wider` is inverse of `pivot_longer`:

```
pigs_longer %>%
  pivot_wider(names_from=feed, values_from=weight)
```

pig	feed1	feed2	feed3	feed4
1	60.8	68.7	92.6	87.9
2	57.0	67.7	92.1	84.2
3	65.0	74.0	90.2	83.1
4	58.6	66.3	96.5	85.7
5	61.7	69.8	99.1	90.3

we are back where we started.

Make wider 2/2

Or

```
pigs_longer %>%
  pivot_wider(names_from=pig, values_from=weight)
```

feed	1	2	3	4	5
feed1	60.8	57.0	65.0	58.6	61.7
feed2	68.7	67.7	74.0	66.3	69.8
feed3	92.6	92.1	90.2	96.5	99.1
feed4	87.9	84.2	83.1	85.7	90.3

Disease presence and absence at two locations

Frequencies of plants observed with and without disease at two locations:

```
Species     Disease present          Disease absent
        Location X Location Y  Location X Location Y
A               44         12          38         10
B               28         22          20         18
```

This has two rows of headers, so I rewrote the data file:

```
Species  present_x present_y    absent_x  absent_y
A               44         12          38         10
B               28         22          20         18
```

Read into data frame called `prevalence`.

Species	present_x	present_y	absent_x	absent_y
A	44	12	38	10
B	28	22	20	18

Lengthen and separate

```
prevalence %>%
  pivot_longer(-Species, names_to = "column",
               values_to = "freq") %>%
  separate(column, into = c("disease", "location"))
```

Species	disease	location	freq
A	present	x	44
A	present	y	12
A	absent	x	38
A	absent	y	10
B	present	x	28
B	present	y	22
B	absent	x	20
B	absent	y	18

Making longer, the better way

```
prevalence %>%
  pivot_longer(-Species, names_to=c("disease", "location"),
               names_sep="_", values_to="frequency")%>%
  arrange(Species, location, disease) -> prevalence_longer
prevalence_longer
```

Species	disease	location	frequency
A	absent	x	38
A	present	x	44
A	absent	y	10
A	present	y	12
B	absent	x	20
B	present	x	28
B	absent	y	18
B	present	y	22

Making wider, different ways

```
prevalence_longer %>%
  pivot_wider(names_from=c(Species, location), values_from=fre
```

disease	A_x	A_y	B_x	B_y
absent	38	10	20	18
present	44	12	28	22

```
prevalence_longer %>%
  pivot_wider(names_from=location, values_from=frequency)
```

Species	disease	x	y
A	absent	38	10
A	present	44	12
B	absent	20	18
B	present	28	22

Interlude

```
pigs_longer
```

pig	feed	weight
1	feed1	60.8
1	feed2	68.7
1	feed3	92.6
1	feed4	87.9
2	feed1	57.0
2	feed2	67.7
2	feed3	92.1
2	feed4	84.2
3	feed1	65.0
3	feed2	74.0
3	feed3	90.2
3	feed4	83.1
4	feed1	58.6

What if summary is more than one number?

eg. quartiles:

```
pigs_longer %>%
  group_by(feed) %>%
  summarize(r=quantile(weight, c(0.25, 0.75)))
```

feed	r
feed1	58.6
feed1	61.7
feed2	67.7
feed2	69.8
feed3	92.1
feed3	96.5
feed4	84.2
feed4	87.9

this also works

```
pigs_longer %>%
  group_by(feed) %>%
  summarize(r=list(quantile(weight, c(0.25, 0.75)))) %>%
  unnest(r)
```

feed	r
feed1	58.6
feed1	61.7
feed2	67.7
feed2	69.8
feed3	92.1
feed3	96.5
feed4	84.2
feed4	87.9

or, even better, use enframe:

```
quantile(pigs_longer$weight, c(0.25, 0.75))
```

```
##     25%     75%
## 65.975 90.225
```

```
enframe(quantile(pigs_longer$weight, c(0.25, 0.75)))
```

name	value
25%	65.975
75%	90.225

A nice look

```
pigs_longer %>%
  group_by(feed) %>%
  summarize(r=list(enframe(quantile(weight, c(0.25, 0.75))))) %>%
  unnest(r) %>%
  pivot_wider(names_from=name, values_from=value)
```

feed	25%	75%
feed1	58.6	61.7
feed2	67.7	69.8
feed3	92.1	96.5
feed4	84.2	87.9

A hairy one

18 people receive one of three treatments. At 3 different times (pre, post, followup) two variables y and z are measured on each person:

id	treatment	pre_y	post_y	fu_y	pre_z	post_z	fu_z
A.1	A	3	13	9	0	0	9
A.2	A	0	14	10	6	6	3
A.3	A	4	6	17	8	2	6
A.4	A	7	7	13	7	6	4
A.5	A	3	12	11	6	12	6
A.6	A	10	14	8	13	3	8
B.1	B	9	11	17	8	11	27
B.2	B	4	16	13	9	3	26
B.3	B	8	10	9	12	0	18
B.4	B	5	9	13	3	0	14
B.5	B	0	15	11	3	0	25
B.6	B	4	11	14	4	2	9

Attempt 1

```
repmes %>% pivot_longer(contains("_"),
                        names_to=c("time", "var"),
                        names_sep="_"
                         )
```

id	treatment	time	var	value
A.1	A	pre	y	3
A.1	A	post	y	13
A.1	A	fu	y	9
A.1	A	pre	z	0
A.1	A	post	z	0
A.1	A	fu	z	9
A.2	A	pre	y	0
A.2	A	post	y	14
A.2	A	fu	y	10
A.2	A	pre	z	6

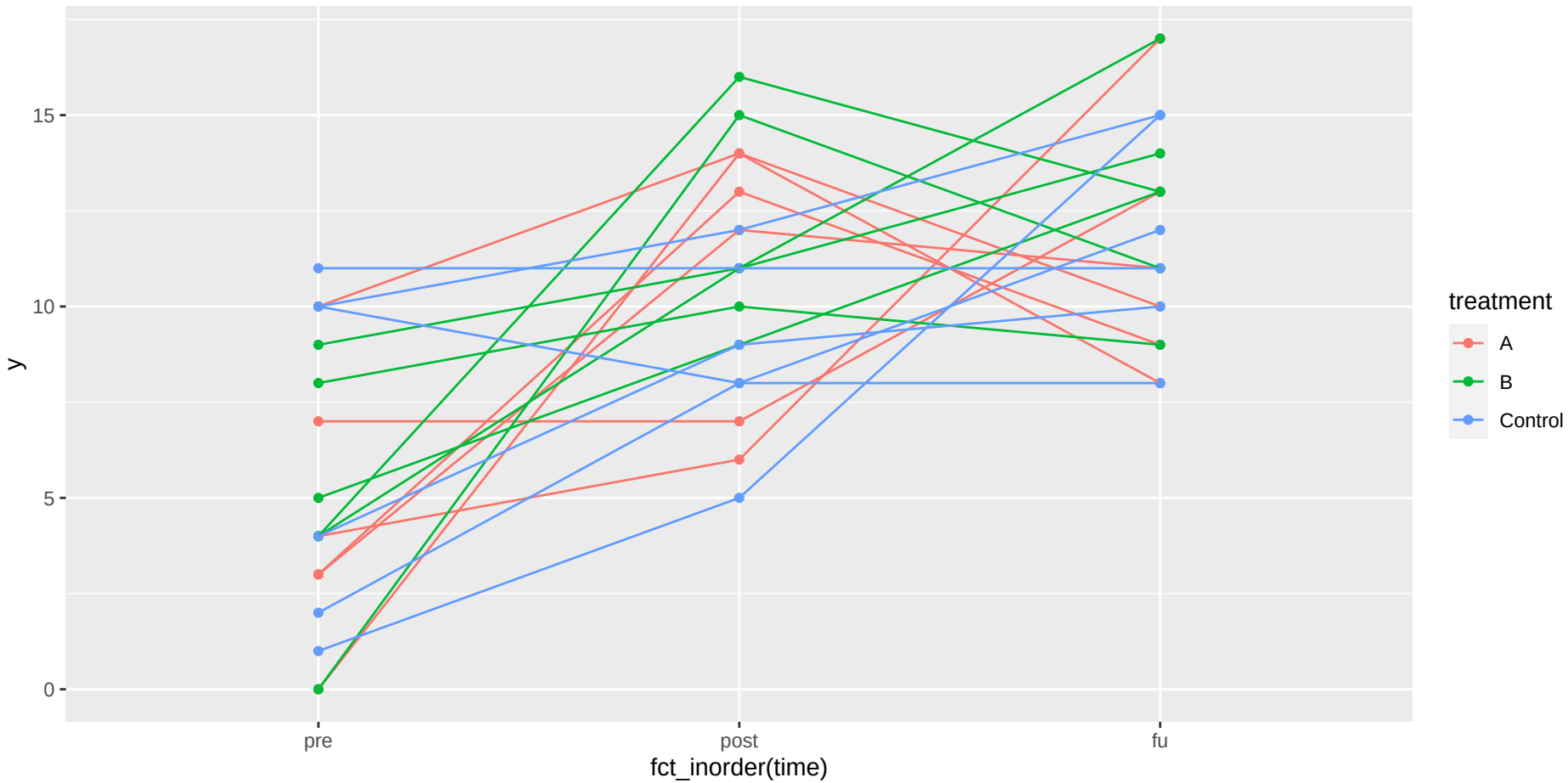
Attempt 2

```
repmes %>% pivot_longer(contains("_"),
                        names_to=c("time", ".value"),
                        names_sep="_"
                        ) -> repmes3
repmes3
```

id	treatment	time	y	z
A.1	A	pre	3	0
A.1	A	post	13	0
A.1	A	fu	9	9
A.2	A	pre	0	6
A.2	A	post	14	6
A.2	A	fu	10	3
A.3	A	pre	4	8
A.3	A	post	6	2
A.3	A	fu	17	6

make a graph

```
ggplot(repmes3, aes(x=fct_inorder(time), y=y,
                    colour=treatment, group=id)) +
  geom_point() + geom_line()
```



or do the plot with means

```
repmes3 %>% group_by(treatment, ftime=fct_inorder(time)) %>%
  summarize(mean_y=mean(y)) %>%
  ggplot(aes(x=ftime, y=mean_y, colour=treatment,
             group=treatment)) +
    geom_point() + geom_line() -> g
```

the plot

```
g
```