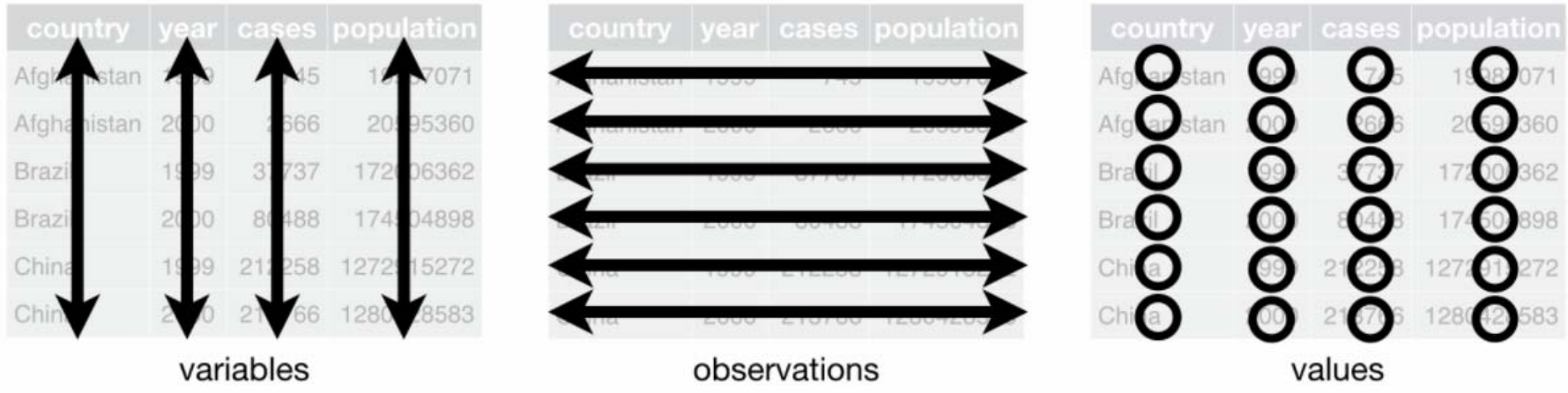
Illustrating tidyr::gather()



Source: Data Science with R by Garrett Grolemund

ME447 Visualizing Data
Fall 2017–18

Richard Layton

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

Data encoded in the column names

Not tidy.

Use tidyr::gather()

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

gather()

```
VA_wide %>%
  gather(location_sex
         , death_rate
         , Rural.Male:Urban.Female
         ) -> VA_long1
```

VA_long1

age_group	location_sex	death_rate
50-54	Rural.Male	11.7
55-59	Rural.Male	18.1
60-64	Rural.Male	26.9
65-69	Rural.Male	41.0
70-74	Rural.Male	66.0
50-54	Rural.Female	8.7
55-59	Rural.Female	11.7
60-64	Rural.Female	20.3
65-69	Rural.Female	30.9
70-74	Rural.Female	54.3
50-54	Urban.Male	15.4
55-59	Urban.Male	24.3
60-64	Urban.Male	37.0
65-69	Urban.Male	54.6
70-74	Urban.Male	71.1
50-54	Urban.Female	8.4
55-59	Urban.Female	13.6
60-64	Urban.Female	19.3
65-69	Urban.Female	35.1
70-74	Urban.Female	50.0

gather() has 3 primary arguments.

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

gather()

```
VA_wide %>%
  gather(location_sex
         , death_rate
         , Rural.Male:Urban.Female
         ) -> VA_long1
```

VA_long1

age_group	location_sex	death_rate
50-54	Rural.Male	11.7
55-59	Rural.Male	18.1
60-64	Rural.Male	26.9
65-69	Rural.Male	41.0
70-74	Rural.Male	66.0
50-54	Rural.Female	8.7
55-59	Rural.Female	11.7
60-64	Rural.Female	20.3
65-69	Rural.Female	30.9
70-74	Rural.Female	54.3
50-54	Urban.Male	15.4
55-59	Urban.Male	24.3
60-64	Urban.Male	37.0
65-69	Urban.Male	54.6
70-74	Urban.Male	71.1
50-54	Urban.Female	8.4
55-59	Urban.Female	13.6
60-64	Urban.Female	19.3
65-69	Urban.Female	35.1
70-74	Urban.Female	50.0

The 1st argument:
new variable name for gathering original column names

gather() creates new column location_sex
writes the old column names as data values in the new column

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

gather()

```
VA_wide %>%
  gather(location_sex
         , death_rate
         , Rural.Male:Urban.Female
         ) -> VA_long1
```

VA_long1 data frame

age_group	location_sex	death_rate
50-54	Rural.Male	11.7
55-59	Rural.Male	18.1
60-64	Rural.Male	26.9
65-69	Rural.Male	41.0
70-74	Rural.Male	66.0
50-54	Rural.Female	8.7
55-59	Rural.Female	11.7
60-64	Rural.Female	20.3
65-69	Rural.Female	30.9
70-74	Rural.Female	54.3
50-54	Urban.Male	15.4
55-59	Urban.Male	24.3
60-64	Urban.Male	37.0
65-69	Urban.Male	54.6
70-74	Urban.Male	71.1
50-54	Urban.Female	8.4
55-59	Urban.Female	13.6
60-64	Urban.Female	19.3
65-69	Urban.Female	35.1
70-74	Urban.Female	50.0

The 2nd argument:
new variable name for gathering original column values

gather() creates new column death_rate
writes the old column values as data values in the new column

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

gather()

```
VA_wide %>%
  gather(location_sex
         , death_rate
         , Rural.Male:Urban.Female
         ) -> VA_long1
```

VA_long

age_group	location_sex	death_rate
50-54	Rural.Male	11.7
55-59	Rural.Male	18.1
60-64	Rural.Male	26.9
65-69	Rural.Male	41.0
70-74	Rural.Male	66.0
50-54	Rural.Female	8.7
55-59	Rural.Female	11.7
60-64	Rural.Female	20.3
65-69	Rural.Female	30.9
70-74	Rural.Female	54.3
50-54	Urban.Male	15.4
55-59	Urban.Male	24.3
60-64	Urban.Male	37.0
65-69	Urban.Male	54.6
70-74	Urban.Male	71.1
50-54	Urban.Female	8.4
55-59	Urban.Female	13.6
60-64	Urban.Female	19.3
65-69	Urban.Female	35.1
70-74	Urban.Female	50.0

The 3rd argument:
names of original columns being gathered

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

gather()

```
VA_wide %>%
  gather(location_sex
         , death_rate
         , Rural.Male:Urban.Female
         ) -> VA_long1
```

VA_long

age_group	location_sex	death_rate
50-54	Rural.Male	11.7
55-59	Rural.Male	18.1
60-64	Rural.Male	26.9
65-69	Rural.Male	41.0
70-74	Rural.Male	66.0
50-54	Rural.Female	8.7
55-59	Rural.Female	11.7
60-64	Rural.Female	20.3
65-69	Rural.Female	30.9
70-74	Rural.Female	54.3
50-54	Urban.Male	15.4
55-59	Urban.Male	24.3
60-64	Urban.Male	37.0
65-69	Urban.Male	54.6
70-74	Urban.Male	71.1
50-54	Urban.Female	8.4
55-59	Urban.Female	13.6
60-64	Urban.Female	19.3
65-69	Urban.Female	35.1
70-74	Urban.Female	50.0

All other columns are copied as many time as needed.

VA_wide

age_group	Rural.Male	Rural.Female	Urban.Male	Urban.Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

gather()

```
VA_wide %>%
  gather(location_sex
         , death_rate
         , Rural.Male:Urban.Female
         ) -> VA_long1
```

VA_long1 data frame

age_group	location_sex	death_rate
50-54	Rural.Male	11.7
55-59	Rural.Male	18.1
60-64	Rural.Male	26.9
65-69	Rural.Male	41.0
70-74	Rural.Male	66.0
50-54	Rural.Female	8.7
55-59	Rural.Female	11.7
60-64	Rural.Female	20.3
65-69	Rural.Female	30.9
70-74	Rural.Female	54.3
50-54	Urban.Male	15.4
55-59	Urban.Male	24.3
60-64	Urban.Male	37.0
65-69	Urban.Male	54.6
70-74	Urban.Male	71.1
50-54	Urban.Female	8.4
55-59	Urban.Female	13.6
60-64	Urban.Female	19.3
65-69	Urban.Female	35.1
70-74	Urban.Female	50.0

The data frame is now in long form

Data that was in the column names is now a variable

What is now still not tidy?