Tidy Data with R

`separate()` and `unite()`

- `spread()` and `gather()` help you reshape the layout of your data to place variables in columns and observations in rows.

- `separate()` and `unite()` allow you split and combine cells to place a single, complete value in each cell.

Tidy Data with R

`separate()`

- `separate()` turns a single character column into multiple columns by splitting the values of the column wherever a separator character appears.
- So, for example, we can use `separate()` to tidy table3, which combines values of cases and population in the same column.

Tidy Data with R

(BEFORE)

```
# Data set three
table3

## Source: local data frame [6 x 3]
##
##        country year              rate
## 1 Afghanistan 1999      745/19987071
## 2 Afghanistan 2000     2666/20595360
## 3      Brazil 1999   37737/172006362
## 4      Brazil 2000   80488/174504898
## 5       China 1999 212258/1272915272
## 6       China 2000 213766/1280428583
```

Tidy Data with R

```
separate(table3, rate,
    into = c("cases", "population"))

## Source: local data frame [6 x 4]
##
##       country year  cases population
## 1 Afghanistan 1999    745   19987071
## 2 Afghanistan 2000   2666   20595360
## 3      Brazil 1999  37737  172006362
## 4      Brazil 2000  80488  174504898
## 5       China 1999 212258 1272915272
## 6       China 2000 213766 1280428583
```

Tidy Data with R

- To use `separate()` pass separate the name of a data frame to reshape and the name of a column to separate.
- Also give `separate()` an into argument, which should be a vector of character strings to use as new column names.
- `separate()` will return a copy of the data frame with the column removed.
- The previous values of the column will be split across several columns, one for each name in into.

Where to Separate?

- By default, `separate()` will split values wherever a non-alphanumeric character appears.
- Non-alphanumeric characters are characters that are neither a number nor a letter.
- For example, in the code above, `separate()` split the values of rate at the forward slash characters.

Tidy Data with R

Specifying a Character

If you wish to use a specific character to separate a column, you can pass the character to the sep argument of `separate()`.

```
separate(table3, rate,
    into = c("cases", "population"),
    sep = "/")
```

Multiple Separation

- You can also pass an integer or vector of integers to `sep`. `separate()` will interpret the integers as positions to split at.
- Positive values start at 1 at the far-left of the strings;
- negative value start at -1 at the far-right of the strings.
- The length of `sep` should be one less than the number of names in into.

Tidy Data with R

- **Example:** You can use this arrangement to separate the last two digits of each year.

Tidy Data with R

(Mid Columns : year into century and year)

```
separate(table3, year,
   into = c("century", "year"), sep = 2)
```

```
## Source: local data frame [6 x 4]
##
##       country century year              rate
## 1 Afghanistan      19   99      745/19987071
## 2 Afghanistan      20   00     2666/20595360
## 3      Brazil      19   99   37737/172006362
## 4      Brazil      20   00   80488/174504898
## 5       China      19   99 212258/1272915272
## 6       China      20   00 213766/1280428583
```