lecture5_note_data_wrangling

Shukyee Chan

2023-10-05

1. … to *reshape* a table (long <-> wide) with `pivot_longer` and `pivot_wider`
2. … to *stack* tables by row or by column with `bind_rows` and `bind_cols` (or, alternatively, `cbind` and `rbind`)
3. … to *merge* two tables with `inner_join`, `full_join`, `left_join`, `right_join`, `semi_join`, and `anti_join`

Outline of In-Class Demo

For this in-class demonstration, we will continue working on the external parts of the V-Dem data from 1984 to 2022. The data are located here: `_DataPublic_/vdem/1984_2022/vdem_1984_2022_external`

1. Reshape the V-Dem dataset
    1. `pivot_longer`: Make it a long table where each variable gets its own row. That is, a row in the new dataset is a *country-year-observation*.
    2. `pivot_wider`: Widen the above long table so that each *Year* has its own column.
2. Stack multiple subsets of the V-Dem datasets by row and by columns
    1. `bind_cols`: Merge the following two subsets of the V-Dem data: `_DataPublic_/vdem/1984_2022/vdem_1984_2022` and `_DataPublic_/vdem/1984_2022/vdem_1984_2022_index`
    2. `bind_rows`: Merge the following two subsets of the V-Dem data: `_DataPublic_/vdem/1984_2022/vdem_1984_2022` and `_DataPublic_/vdem/1945_1983/vdem_1945_1983_external`
3. Join multiple regional subsets of the V-Dem datasets
    1. Make a new data frame that contains the following variables: `country_name`, `year`, `e_regionpol_6C`, `e_fh_status`, `e_gdppc`, and `e_gdp`
    2. Create two separate subsets of the above data frames. Each subset include a subset of countries/regions that are within the *region* (defined by `e_regiongeo` and `e_regionpol_6C` respectively) where *China* is located.
    3. Explore the behavior of `inner_join`, `full_join`, `left_join`, `right_join`, `semi_join`, and `anti_join` with the two data frames.
4. Validate V-dem's GDP data with World Bank data

```r
library(tidyverse)
```

```r
d <- read_csv("_DataPublic_/vdem/1984_2022/vdem_1984_2022_external.csv")
```

1. Reshape the V-Dem dataset

```
# Want: Each row contain country-year-variable
# want to make a row contain only one variable of the country

# take a look at the names of the variable
d |> select(country_name) |> distinct()
```

```
## # A tibble: 181 x 1
##    country_name
##    <chr>
##  1 Mexico
##  2 Suriname
##  3 Sweden
##  4 Switzerland
##  5 Ghana
##  6 South Africa
##  7 Japan
##  8 Burma/Myanmar
##  9 Russia
## 10 Albania
## # ... with 171 more rows
```

```
d_subset <- d |>
  select(country_name, year,starts_with("e"))

d_subset_long <- d_subset |>
  pivot_longer(cols = starts_with("e"))

# transform: make a wide dataset with a lot of columns into each columns become has its own row
```

```
d_subset_wide_year <- d_subset_long |>
  pivot_wider(names_from = year, values_from = value)
```

2. Stack multiple subsets of the V-Dem datasets

```
d_VdemIndex <- read_csv("_DataPublic_/vdem/1984_2022/vdem_1984_2022_index.csv")
```

```
d_stack <- bind_cols(d, d_VdemIndex)
```

```
# Want: Stack two tables by rows
d_1945_1983 <- read_csv("_DataPublic_/vdem/1945_1983/vdem_1945_1983_external.csv")
```

```
d_1945_2022 <- bind_rows(d, d_1945_1983)

d_1945_2022 |>
  select(year) |>
  distinct() |>
  arrange(year)
```

```
## # A tibble: 78 x 1
##     year
##    <dbl>
##  1  1945
##  2  1946
##  3  1947
##  4  1948
##  5  1949
##  6  1950
##  7  1951
##  8  1952
##  9  1953
## 10  1954
## # ... with 68 more rows
```