Data Wrangling (1)

Haohan Chen

Last update: October 05, 2023

```
library(tidyverse)
```

Objectives of this Lecture

This lecture introduces data wrangling with R. Using V-Dem data as an example, we will learn how to use the wrangle data with a set of `tidyverse` functionality. Specifically, we will focus on functions...

1. to import and export data: `read_csv` , `write_csv` (with a brief introduction to other data import/ export functions from `readr`).
2. to take a subset of *columns* in the existing data: `select`
3. to rename columns: `rename`
4. to take a subset of *rows* by some simple conditions: `slice_`
5. to take a subset of *rows* by some more complicated conditions: `filter`
6. to sort the rows based on the value of one or multiple columns: `arrange`
7. to perform (4) (5) (6) group by group: `group_by`, `ungroup`
8. to create new columns in the data: `group_by`, `mutate`, `ungroup`
9. to summarize the data: `group_by`, `summarise`, `ungroup`

Outline of In-Class Demo

To demonstrate the above functionality, we will use real-world political data from V-Dem. Specifically, we will use the above function to explore the state of global economic development from 1984 to 2022. Our effort will take the following step (with one-on-one mappings with the above tools).

1. Read a part of pre-processed V-Dem data into R: 1984-2022 "external" data in the V-Dem dataset.
2. Consulting the dataset's codebook and take a **subset** of indicators of *economic development* (along with country-year identifiers).
    - *See a list of country-yer identifiers on p. 5 of the codebook (under "1.7 Identifier Variables in the V-Dem Datasets").*
    - *See a list of development indicators on p. 23 of the codebook (under "9. Background Factors").*
3. Rename the column to name their names informative to readers.

4. Find the country-year with the *highest* and *lowest* level of economic development. In addition, create a dataset containing a random sample of country-year in the dataset.
5. Create a dataset focusing on the economic development of Asian countries and regions; Create a dataset that contains only countries/ regions whose development level pass certain threshold.
6. Create a dataset whose rows are sorted by the development level of country-year.
7. Create a dataset that contains the year of the higest development level for each country/ region respectively.
8. Add the following economic indicators to the data:
    1. Country-year development level with reference to that of 1984.
    2. Year-on-year economic growth.
9. Perform a data availability/ integrity check. Then aggregate the data into a new country-level dataset which contains the following indicators:
    1. Average development level from 1984 to 2022.
    2. Magnitude of growth from 1984 to 2022.

In-Class Exercise

The quality of education has a decisive effect on a country's future development. Applying the data wrangling tools we introduce in this lecture, perform the following task:

1. **Coodbook lookup**. Look up the codebook, answer the following questions:
    1. What indicators regarding the quality of education are available in the V-Dem datasets? 9 Background Factors (E)

    9.1.1 Education 15+ (E) (e_peaveduc) What is the average years of education among citizens older than 15?

    9.1.2 Educational inequality, Gini (E) (e_peedgini) How unequal is the level of education achieved by the population aged 15 years and older?

    2. What are the data's coverage (i.e., for which countries and years do we have data?)

```
setwd('C:/Users/warre/OneDrive/Documents/GitHub/HKU_POLI3148_23Fall/')
set.seed(52)
d <- read_csv("_DataPublic_/vdem/1984_2022/vdem_1984_2022_external.csv")
```

```
## Rows: 6789 Columns: 211
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (3): country_name, country_text_id, histname
## dbl  (207): country_id, year, project, historical, codingstart, codingend, c...
## date   (1): historical_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
d |> select(country_name, country_id, year) |> distinct()
```

```
## # A tibble: 6,789 x 3
##    country_name country_id  year
##    <chr>              <dbl> <dbl>
##  1 Mexico                 3  1984
##  2 Mexico                 3  1985
##  3 Mexico                 3  1986
##  4 Mexico                 3  1987
##  5 Mexico                 3  1988
##  6 Mexico                 3  1989
##  7 Mexico                 3  1990
##  8 Mexico                 3  1991
##  9 Mexico                 3  1992
## 10 Mexico                 3  1993
## # i 6,779 more rows
```

```
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
## # i 171 more rows
```

```
d |> select(year) |> distinct()
```

```
## # A tibble: 39 x 1
##     year
##    <dbl>
##  1  1984
##  2  1985
##  3  1986
##  4  1987
##  5  1988
##  6  1989
##  7  1990
##  8  1991
##  9  1992
## 10  1993
## # i 29 more rows
```

3. What are their sources? Provide the link to least 1 source.

e_peedgini: Source(s): Clio Infra (clio-infra.eu), drawing on Mitchell (1998a, 1998b, 1998c), United States Census Bureau (2021), UNESCO, Földvári and van Leeuwen (2010a), Leeuwen, van Leeuwen-Li, Földvári (2011), Leeuwen, van Leeuwen-Li, Földvári (2012a), Leeuwen, van Leeuwen-Li, Földvári (2012b), Didenko, Foldvari, van Leeuwen (2012).

2. **Subset by columns**

    1. Create a dataset containing only the country-year identifiers and indicators of education quality.

```
education <- d |> select(country_name,  year, e_peaveduc, e_peedgini)
```

2. Rename the columns of education quality to make them informative.

```
education_renamed <- education |> rename( "average_years_of_postsecondary_education" = "e_peaveduc" , "
```

3. **Subset by rows**

    1. List 5 countries-years that have the highest education level among its population.

```
education_renamed |> slice_max(average_years_of_postsecondary_education, n=5)
```

```
## # A tibble: 13 x 4
##    country_name    year average_years_of_postsecondary_~1 postsecondary_gini_i~2
##    <chr>          <dbl>                             <dbl>                  <dbl>
##  1 United Kingdom  2010                              13.3                   6.07
##  2 United Kingdom  2011                              13.3                  NA
##  3 United Kingdom  2012                              13.3                  NA
##  4 United Kingdom  2013                              13.3                  NA
##  5 United Kingdom  2014                              13.3                  NA
##  6 United Kingdom  2015                              13.3                  NA
##  7 United Kingdom  2016                              13.3                  NA
##  8 United Kingdom  2017                              13.3                  NA
##  9 United Kingdom  2018                              13.3                  NA
## 10 United Kingdom  2019                              13.3                  NA
## 11 United Kingdom  2020                              13.3                  NA
## 12 United Kingdom  2021                              13.3                  NA
## 13 United Kingdom  2022                              13.3                  NA
## # i abbreviated names: 1: average_years_of_postsecondary_education,
## #   2: postsecondary_gini_inequality_index
```

2. List 5 countries-years that suffer from the most severe inequality in education.

```
education_renamed |> slice_max(postsecondary_gini_inequality_index, n=5)
```

```
## # A tibble: 5 x 4
##   country_name  year average_years_of_postsecondary_edu~1 postsecondary_gini_i~2
##   <chr>        <dbl>                              <dbl>                  <dbl>
## 1 Burkina Faso  1984                              0.301                   97.0
## 2 Burkina Faso  1985                              0.322                   96.9
## 3 Burkina Faso  1986                              0.343                   96.7
## 4 Burkina Faso  1987                              0.364                   96.4
## 5 Burkina Faso  1988                              0.385                   96.1
## # i abbreviated names: 1: average_years_of_postsecondary_education,
## #   2: postsecondary_gini_inequality_index
```

4. **Summarize the data**

1. Check data availability: For which countries and years are the indicators of education quality available?

```
cleaned_data <- education_renamed |> filter_at(vars(c(postsecondary_gini_inequality_index, average_years

education_renamed |> mutate(gini_missing = is.na(postsecondary_gini_inequality_index)) |> group_by(coun
```

```
## # A tibble: 181 x 2
##    country_name number_missing_gini
##    <chr>                      <int>
##  1 Afghanistan                   12
##  2 Albania                       39
##  3 Algeria                       12
##  4 Angola                        12
##  5 Argentina                     12
##  6 Armenia                       12
##  7 Australia                     12
##  8 Austria                       12
##  9 Azerbaijan                    12
## 10 Bahrain                       39
## # i 171 more rows
```

```
education_renamed |> mutate(yrs_missing = is.na(average_years_of_postsecondary_education)) |> group_by(
```

```
## # A tibble: 181 x 2
##    country_name number_missing_years_of_education
##    <chr>                                    <int>
##  1 Afghanistan                                  0
##  2 Albania                                     39
##  3 Algeria                                      0
##  4 Angola                                       0
##  5 Argentina                                    0
##  6 Armenia                                      0
##  7 Australia                                    0
##  8 Austria                                      0
##  9 Azerbaijan                                   0
## 10 Bahrain                                     39
## # i 171 more rows
```

```
summary(cleaned_data)
```

```
##  country_name            year      average_years_of_postsecondary_education
##  Length:5015        Min.   :1984   Min.   : 0.301
##  Class :character   1st Qu.:1994   1st Qu.: 4.840
##  Mode  :character   Median :2003   Median : 7.489
##                     Mean   :2003   Mean   : 7.360
##                     3rd Qu.:2013   3rd Qu.:10.118
##                     Max.   :2022   Max.   :13.300
##
##  postsecondary_gini_inequality_index
##  Min.   : 3.771
##  1st Qu.:18.726
```

```
##  Median :27.937
##  Mean   :34.298
##  3rd Qu.:46.602
##  Max.   :96.983
##  NA's   :1637
```

2. Create two types of country-level indicators of education quality

    1. Average level of education quality from 1984 to 2022

```
education_renamed |> group_by(country_name) |> summarise(avg_gini_index = mean(postsecondary_gini_inequa
```

```
## # A tibble: 181 x 2
##    country_name   avg_gini_index
##    <chr>                   <dbl>
##  1 Austria                  6.35
##  2 Barbados                 6.98
##  3 Denmark                  8.17
##  4 Switzerland              8.28
##  5 United Kingdom           8.38
##  6 Japan                    9.33
##  7 Norway                   9.58
##  8 Australia                9.60
##  9 Tajikistan              10.8
## 10 Hungary                 11.2
## # i 171 more rows
```

```
education_renamed |> group_by(country_name) |> summarise(avg_gini_index = mean(postsecondary_gini_inequa
```

```
## # A tibble: 181 x 2
##    country_name avg_gini_index
##    <chr>                 <dbl>
##  1 Burkina Faso           91.3
##  2 Mali                   87.9
##  3 Niger                  85.3
##  4 Somalia                84.7
##  5 Afghanistan            77.8
##  6 Benin                  76.9
##  7 The Gambia             76.7
##  8 Guinea                 73.4
##  9 Burundi                73.0
## 10 Nepal                  69.8
## # i 171 more rows
```

```
education_renamed |> group_by(country_name) |> summarise(average_years_of_education = mean(average_years
```

```
## # A tibble: 181 x 2
##    country_name average_years_of_education
##    <chr>                             <dbl>
##  1 Burkina Faso                      0.982
##  2 Niger                             1.06
```

```
##  3 Mali                              1.25
##  4 Somalia                           1.29
##  5 Burundi                           1.86
##  6 Mozambique                        2.36
##  7 Benin                             2.39
##  8 Angola                            2.46
##  9 Senegal                           2.54
## 10 Guinea                            2.62
## # i 171 more rows
```

```
education_renamed |> group_by(country_name) |> summarise(average_years_of_education = mean(average_years
```

```
## # A tibble: 181 x 2
##    country_name   average_years_of_education
##    <chr>                               <dbl>
##  1 Germany                              12.9
##  2 Australia                            12.9
##  3 United Kingdom                       12.9
##  4 Canada                               12.7
##  5 Switzerland                          12.7
##  6 Japan                                12.6
##  7 Norway                               12.4
##  8 France                               12.0
##  9 South Korea                          12.0
## 10 New Zealand                          11.9
## # i 171 more rows
```

2. Change of education quality from 1984 to 2022

```
education_renamed |> group_by(country_name) |> arrange(year, by.group=TRUE) |> mutate(change_in_years_o
```

```
## # A tibble: 179 x 3
## # Groups:   country_name [179]
##    country_name  year change_in_years_of_education
##    <chr>        <dbl>                        <dbl>
##  1 Botswana      2022                         5.17
##  2 Singapore     2022                         4.52
##  3 Libya         2022                         4.07
##  4 Cuba          2022                         3.84
##  5 Chad          2022                         3.82
##  6 Egypt         2022                         3.82
##  7 Jordan        2022                         3.82
##  8 South Korea   2022                         3.54
##  9 Saudi Arabia  2022                         3.49
## 10 Algeria       2022                         3.35
## # i 169 more rows
```

```
education_renamed |> group_by(country_name) |> arrange(year , by.group=TRUE) |> mutate(change_in_years_
```

```
## # A tibble: 179 x 3
## # Groups:   country_name [179]
```

```
##    country_name  year change_in_years_of_education
##    <chr>        <dbl>                         <dbl>
##  1 Tajikistan    2022                        -0.252
##  2 North Korea   2022                         0
##  3 Russia        2022                         0.230
##  4 Azerbaijan    2022                         0.252
##  5 Uzbekistan    2022                         0.272
##  6 Kyrgyzstan    2022                         0.301
##  7 Switzerland   2022                         0.328
##  8 Armenia       2022                         0.336
##  9 Germany       2022                         0.350
## 10 Georgia       2022                         0.387
## # i 169 more rows
```

```
education_renamed   |> group_by(country_name) |> arrange(year) |> mutate(change= last(na.omit(postsecon
```

```
## # A tibble: 179 x 3
## # Groups:   country_name [179]
##    country_name         year change
##    <chr>               <dbl>  <dbl>
##  1 Costa Rica           2022  4.12
##  2 New Zealand          2022  3.16
##  3 Spain                2022  2.30
##  4 Trinidad and Tobago  2022  2.30
##  5 Switzerland          2022  1.72
##  6 Lebanon              2022  0.718
##  7 Seychelles           2022  0.696
##  8 France               2022 -0.287
##  9 Venezuela            2022 -0.395
## 10 Jamaica              2022 -0.597
## # i 169 more rows
```

```
education_renamed   |> group_by(country_name) |> arrange(year) |> mutate(change= last(na.omit(postsecon
```

```
## # A tibble: 179 x 3
## # Groups:   country_name [179]
##    country_name  year change
##    <chr>        <dbl>  <dbl>
##  1 Nepal         2022  -39.8
##  2 Botswana      2022  -34.0
##  3 Haiti         2022  -31.5
##  4 Egypt         2022  -30.8
##  5 Iran          2022  -30.3
##  6 Angola        2022  -29.5
##  7 India         2022  -29.0
##  8 Nigeria       2022  -27.5
##  9 Malawi        2022  -27.2
## 10 Uganda        2022  -26.8
## # i 169 more rows
```

3. Examine the data and *briefly* discuss: Which countries perform the best and the worst in terms of

African Nations generally do not see much education of individuals past the age of 15, where as Western developed nations sees higher education levels for individuals older than 15. Those same African nations also see a higher inequality, with a few very educated individuals and a lot of uneducated individuals. We see that Singapore and other developing Asian nations have seen higher growth of education level. However, in previous soviet states, the education years have gone down. This may be due to the following: while the soviets valued education, the value of education has become less relevant in the post soviet world as many states have become petro-states.

Submission requirement: You will submit your outputs through Moodle. In your submission:

1. Attach a PDF document rendered by Rmarkdown
2. In the text field of your submission, include the link to the corresponding Rmarkdown file in your *DaSPPA portfolio* GitHub repo.

Due: October 4, 2023

Note: Please only use the functions we cover in this lecture for this exercise. There is absolutely no need to perform any data visualization for this exercise... We will get there in later lectures.

Further reading

- R for Data Science (2e) Chapters 4, 5, 8: https://r4ds.hadley.nz/
- `readr` documentation (note: read the "cheatsheet"): https://readr.tidyverse.org/
- `dplyr` documentation (note: read the "cheatsheet"): https://dplyr.tidyverse.org/
- V-Dem documentation: https://v-dem.net/

Demo