Data Wrangling (1)

Haohan Chen

Last update: September 28, 2023

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?

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

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

2. **Subset by columns**

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

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

3. **Subset by rows**

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

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

4. **Summarize the data**

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

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

        1. Average level of education quality from 1984 to 2022
        2. Change of education quality from 1984 to 2022

    3. Examine the data and *briefly* discuss: Which countries perform the best and the worst in terms of education quality in the past four decades?

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

0. Load the `tidyverse` Packages

This section loads the packages we need in this lecture.

```
library(tidyverse)
```

1. Import and Export the V-Dem Data

This section loads the VDEM dataset and describe its basic information

```
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.
```