Data Wrangling

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`, `summarize`, `ungroup`

Case Study

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, take a subset of indicators of *economic development* (along with country-year identifiers).
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. Make a new dataset contains the following indicators:
    1. Average development level from 1984 to 2022.
    2. Magnitude of economic growth from 1984 to 2022.

In-class Exercise

Further reading

- R for Data Science (2e) Chapters 4, 5, 8: https://r4ds.hadley.nz/
- V-Dem documentation: https://v-dem.net/

Load the `tidyverse` Packages

This section loads the packages we need in this lecture.

```
library(tidyverse)
```

Read and Write 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.
```

Check Basic Information of the Dataset

```
dim(d)
```

```
## [1] 6789  211
```

```
names(d)
```

```
##   [1] "country_name"              "country_text_id"
##   [3] "country_id"                "year"
##   [5] "historical_date"           "project"
##   [7] "historical"                "histname"
##   [9] "codingstart"               "codingend"
##  [11] "codingstart_contemp"       "codingend_contemp"
##  [13] "codingstart_hist"          "codingend_hist"
##  [15] "gapstart1"                 "gapstart2"
##  [17] "gapstart3"                 "gapend1"
##  [19] "gapend2"                   "gapend3"
##  [21] "gap_index"                 "COWcode"
##  [23] "e_v2x_api_3C"              "e_v2x_api_4C"
##  [25] "e_v2x_api_5C"              "e_v2x_civlib_3C"
##  [27] "e_v2x_civlib_4C"           "e_v2x_civlib_5C"
##  [29] "e_v2x_clphy_3C"            "e_v2x_clphy_4C"
##  [31] "e_v2x_clphy_5C"            "e_v2x_clpol_3C"
##  [33] "e_v2x_clpol_4C"            "e_v2x_clpol_5C"
##  [35] "e_v2x_clpriv_3C"           "e_v2x_clpriv_4C"
##  [37] "e_v2x_clpriv_5C"           "e_v2x_corr_3C"
##  [39] "e_v2x_corr_4C"             "e_v2x_corr_5C"
##  [41] "e_v2x_cspart_3C"           "e_v2x_cspart_4C"
##  [43] "e_v2x_cspart_5C"           "e_v2x_delibdem_3C"
##  [45] "e_v2x_delibdem_4C"         "e_v2x_delibdem_5C"
##  [47] "e_v2x_EDcomp_thick_3C"     "e_v2x_EDcomp_thick_4C"
##  [49] "e_v2x_EDcomp_thick_5C"     "e_v2x_egal_3C"
##  [51] "e_v2x_egal_4C"             "e_v2x_egal_5C"
##  [53] "e_v2x_egaldem_3C"          "e_v2x_egaldem_4C"
##  [55] "e_v2x_egaldem_5C"          "e_v2x_elecoff_3C"
##  [57] "e_v2x_elecoff_4C"          "e_v2x_elecoff_5C"
##  [59] "e_v2x_execorr_3C"          "e_v2x_execorr_4C"
##  [61] "e_v2x_execorr_5C"          "e_v2x_feduni_3C"
##  [63] "e_v2x_feduni_4C"           "e_v2x_feduni_5C"
##  [65] "e_v2x_frassoc_thick_3C"    "e_v2x_frassoc_thick_4C"
##  [67] "e_v2x_frassoc_thick_5C"    "e_v2x_freexp_3C"
##  [69] "e_v2x_freexp_4C"           "e_v2x_freexp_5C"
##  [71] "e_v2x_freexp_altinf_3C"    "e_v2x_freexp_altinf_4C"
##  [73] "e_v2x_freexp_altinf_5C"    "e_v2x_gencl_3C"
##  [75] "e_v2x_gencl_4C"            "e_v2x_gencl_5C"
##  [77] "e_v2x_gencs_3C"            "e_v2x_gencs_4C"
##  [79] "e_v2x_gencs_5C"            "e_v2x_gender_3C"
##  [81] "e_v2x_gender_4C"           "e_v2x_gender_5C"
##  [83] "e_v2x_genpp_3C"            "e_v2x_genpp_4C"
##  [85] "e_v2x_genpp_5C"            "e_v2x_jucon_3C"
##  [87] "e_v2x_jucon_4C"            "e_v2x_jucon_5C"
##  [89] "e_v2x_libdem_3C"           "e_v2x_libdem_4C"
##  [91] "e_v2x_libdem_5C"           "e_v2x_liberal_3C"
##  [93] "e_v2x_liberal_4C"          "e_v2x_liberal_5C"
##  [95] "e_v2x_mpi_3C"              "e_v2x_mpi_4C"
##  [97] "e_v2x_mpi_5C"              "e_v2x_partip_3C"
##  [99] "e_v2x_partip_4C"           "e_v2x_partip_5C"
## [101] "e_v2x_partipdem_3C"        "e_v2x_partipdem_4C"
## [103] "e_v2x_partipdem_5C"        "e_v2x_polyarchy_3C"
## [105] "e_v2x_polyarchy_4C"        "e_v2x_polyarchy_5C"
## [107] "e_v2x_pubcorr_3C"          "e_v2x_pubcorr_4C"
```

```
## [109] "e_v2x_pubcorr_5C"           "e_v2x_suffr_3C"
## [111] "e_v2x_suffr_4C"             "e_v2x_suffr_5C"
## [113] "e_v2xcl_rol_3C"             "e_v2xcl_rol_4C"
## [115] "e_v2xcl_rol_5C"             "e_v2xcs_ccsi_3C"
## [117] "e_v2xcs_ccsi_4C"            "e_v2xcs_ccsi_5C"
## [119] "e_v2xdd_dd_3C"              "e_v2xdd_dd_4C"
## [121] "e_v2xdd_dd_5C"              "e_v2xdl_delib_3C"
## [123] "e_v2xdl_delib_4C"           "e_v2xdl_delib_5C"
## [125] "e_v2xeg_eqdr_3C"            "e_v2xeg_eqdr_4C"
## [127] "e_v2xeg_eqdr_5C"            "e_v2xeg_eqprotec_3C"
## [129] "e_v2xeg_eqprotec_4C"        "e_v2xeg_eqprotec_5C"
## [131] "e_v2xel_frefair_3C"         "e_v2xel_frefair_4C"
## [133] "e_v2xel_frefair_5C"         "e_v2xel_locelec_3C"
## [135] "e_v2xel_locelec_4C"         "e_v2xel_locelec_5C"
## [137] "e_v2xel_regelec_3C"         "e_v2xel_regelec_4C"
## [139] "e_v2xel_regelec_5C"         "e_v2xlg_legcon_3C"
## [141] "e_v2xlg_legcon_4C"          "e_v2xlg_legcon_5C"
## [143] "e_v2xme_altinf_3C"          "e_v2xme_altinf_4C"
## [145] "e_v2xme_altinf_5C"          "e_v2xps_party_3C"
## [147] "e_v2xps_party_4C"           "e_v2xps_party_5C"
## [149] "e_boix_regime"              "e_democracy_breakdowns"
## [151] "e_democracy_omitteddata"    "e_democracy_trans"
## [153] "e_fh_cl"                    "e_fh_pr"
## [155] "e_fh_rol"                   "e_fh_status"
## [157] "e_wbgi_cce"                 "e_wbgi_gee"
## [159] "e_wbgi_pve"                 "e_wbgi_rle"
## [161] "e_wbgi_rqe"                 "e_wbgi_vae"
## [163] "e_lexical_index"            "e_uds_median"
## [165] "e_uds_mean"                 "e_uds_pct025"
## [167] "e_uds_pct975"               "e_coups"
## [169] "e_legparty"                 "e_autoc"
## [171] "e_democ"                    "e_p_polity"
## [173] "e_polcomp"                  "e_polity2"
## [175] "e_bnr_dem"                  "e_chga_demo"
## [177] "e_ti_cpi"                   "e_vanhanen"
## [179] "e_peaveduc"                 "e_peedgini"
## [181] "e_area"                     "e_regiongeo"
## [183] "e_regionpol"                "e_regionpol_6C"
## [185] "e_cow_exports"              "e_cow_imports"
## [187] "e_gdp"                      "e_gdp_sd"
## [189] "e_gdppc"                    "e_gdppc_sd"
## [191] "e_miinflat"                 "e_pop"
## [193] "e_pop_sd"                   "e_total_fuel_income_pc"
## [195] "e_total_oil_income_pc"      "e_total_resources_income_pc"
## [197] "e_radio_n"                  "e_miferrat"
## [199] "e_mipopula"                 "e_miurbani"
## [201] "e_miurbpop"                 "e_pefeliex"
## [203] "e_peinfmor"                 "e_pelifeex"
## [205] "e_pematmor"                 "e_wb_pop"
## [207] "e_civil_war"                "e_miinteco"
## [209] "e_miinterc"                 "e_pt_coup"
## [211] "e_pt_coup_attempts"
```

Select Variables (Columns) of Interest

```
d_s <- d |>
  select(country_name, country_id, year, e_fh_cl, e_gdp, e_gdppc)

d_s
```

```
## # A tibble: 6,789 x 6
##    country_name country_id  year e_fh_cl   e_gdp e_gdppc
##    <chr>             <dbl> <dbl>   <dbl>   <dbl>   <dbl>
##  1 Mexico                3  1984       4  93563.    11.7
##  2 Mexico                3  1985       4  94259.    11.5
##  3 Mexico                3  1986       4  92750.    11.1
##  4 Mexico                3  1987       4  93220.    10.9
##  5 Mexico                3  1988       4  94687.    10.8
##  6 Mexico                3  1989       3  98145.    11.0
##  7 Mexico                3  1990       4 103254.    11.4
##  8 Mexico                3  1991       4 107374.    11.6
##  9 Mexico                3  1992       3 111533.    11.9
## 10 Mexico                3  1993       4 114611.    12.0
## # ... with 6,779 more rows
```

Rename Variables of Interest

```
d_s <- d_s |>
  rename("FH Civil Liberty" = "e_fh_cl",
         "GDP" = "e_gdp",
         "GDP per capita" = "e_gdppc")

# Check out functions that can allow you to rename variables in batch
```

Filter Observations (Rows) of Interest