Data.table Lab

David Gerard

2019-07-01

Learning Objectives

- Manipulating data.tables.
- Tidying data.tables.

College Scorecard

For this lab, **use data.table and not the tidyverse**.

The data in "college_score.csv" contains a subset of the variables found in the 2016 to 2017 College Scorecard database. These data contain information on colleges in the United States. The variables included are:

- `UNITID` and `OPEID`: Identifiers for the colleges.
- `INSTNM`: Institution name
- `ADM_RATE`: The Admission Rate.
- `SAT_AVE`: Average SAT equivalent score of students admitted.
- `UGDS`: Enrollment of undergraduate certificate/degree-seeking students
- `COSTT4_A`: Average cost of attendance (academic year institutions)
- `AVGFACSAL`: Average faculty salary
- `GRAD_DEBT_MDN`: The median debt for students who have completed
- `AGE_ENTRY`: Average age of entry
- `ICLEVEL`: Level of institution (1 = 4-year, 2 = 2-year, 3 = less than 2-year).
- `MN_EARN_WNE_P6`: Mean earnings of students working and not enrolled 6 years after entry (so students who graduated in the 2009 to 2010 academic year).

1. Use `fread()` and relative paths to load in the `data.table` from college_score.csv.
2. Use data.table to calculate the average median debt for each level of institution.
3. Filter out any rows with an `NA` in `SAT_AVG`.
4. Add the variable `SAT_RANK` to the data frame which contains the rank of the institution based on `SAT_AVG`. That is, the school with the highest `SAT_AVG` has a rank of 1, the next highest has a rank of 2, etc. Add this variable by reference.
5. Recode `ICLEVEL` to more human-readable levels. Do this by reference.

World Bank

For this lab, **use data.table and not the tidyverse**.

The World Bank is an international organization that provides loans to countries with the goal of reducing poverty. The data frames in the data folder were all taken from the public data repositories of the World Bank.

- fertility.csv: Contains the fertility rate information for each country for each year. For the variables `1960` to `2017`, the values in the cells represent the fertility rate in total births per woman for the that year. Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
- life_exp.csv: Contains the life expectancy information for each country for each year. For the variables `1960` to `2017`, the values in the cells represent life expectancy at birth in years for the given year. Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

1. Use relative paths and `fread()` to read these data into R.
2. These data are messy. The observational units in fertility.csv and life_exp.csv are locations in space-time (e.g. Aruba in 2017). Recall that tidy data should have one observational unit per row. Make these data tidy now. Make sure that the new year variable is a numeric.
3. Join the tidied data.tables together.
4. For 1960, plot fertility rate by life expectancy. Add a loess curve. You can use ggplot2 for this, or you can try using base R's plotting functions.