



7316 - INTRODUCTION TO R

Module 2: Data manipulation in R - Assignment

Teacher: Mickaël Buffart (mickael.buffart@hhs.se)

In this assignment you will work on some tables from David Card and Alan Krueger's seminal 1994 paper on the minimum wage introduction in New Jersey. The paper is available on the course webpage.

1 BASIC SETUP: TIDYVERSE AND RIO

1. Install the packages "Tidyverse" and "Rio", if you have not yet installed them.
2. Create a project for this assignment. Create a data folder in it.
3. Download the dataset `card_krueger_public.dta` (in `module_2_assignment.zip`) from the course website. Copy it into your project data folder. Load the data.

2 TAKING A FIRST LOOK

4. Describe your dataset: what is its structure: how many variables, types of variables, etc.

You will notice that the data does not contain any variable names. We therefore refer to the codebook (on the course website) to find the necessary variables. I have prepared a csv file with the variable names and labels called `variable_names.csv`.

5. Load this list and assign each column the appropriate variable name.

The dataset is still a bit large, given that we only want to replicate a two tables.

6. Drop all variables except `SHEET`, `CHAIN`, `STATE`, `EMPFT`, `EMPPT`, `NMGRS`, `EMPFT2`, `EMPPT2`, `NMGRS2`, `STATUS2`

3 SUMMARIZING THE DATA

7. Check if `EMPFT` contains missing values.

We now want to get a feeling for the kind of observations we are dealing with. Card and Krueger sample restaurants of different US fast food chains (Burger King, KFC, Roy Rogers, Wendy's). We would like to know the distribution of the different chains across New Jersey and Pennsylvania (table 2 in the paper)

8. Create a separate dummy variable for each chain that equals 1 (or `TRUE`) if the store belongs to this chain and 0 (`FALSE`) otherwise
9. Tabulate the mean of each of these 4 variables by State
10. Save the tabulated values into a `matrix`

11. Remove the stata dummy and transpose the matrix, rename the columns such that it corresponds to the *Distributon of Store Types* section of Table 2 and turn it into a tibble.
12. Print the table

4 TYDING UP THE DATASET

If you look at the data, you will realize that the values for a single store are spread across several columns. The number of full-time employees is recorded in the variable `EMPFT` for the first year and `EMPFT2` for the second year. This violates the tidy principle that each observation has its own row. To make the tidying easier, we first reduce the number of variables by aggregating full-time employment, part-time employment and managers into one variable for full-time equivalents (`FTE`).

13. Aggregate the employment for each store and period into two new variables called `FTE1` and `FTE2`. Follow the paper and use the formula $FTE = EMPFT + 0.5 * EMPPT + NMGRS$
14. Order the data in `FTE1` ascending order and `FTE2` descending order.
15. Gather the data, such that for each store you have two observations of `FTE`, one for each year. Save this new dataset as `data.tidy`.