Homework 1: Data (1)

Due on January 19th 2022 at 11 PM

Your final output should consist of raw code and a pdf file with answers. Output should be uploaded on Github before the due date/hours.

The goal of this assignment is to introduce basic data manipulation techniques. The recommended software is R. Packages from the tidyverse universe are especially recommended for this assignment. Document your code as much as possible to allow easy grading.

The data

We use data from the French SRCV (Statistics and Resources on Living Condition) datasets from year 2004 to 2019. Each year, two surveys are conducted:

- dathh: is a survey of households in France. The data is longitudinal, with household identifier given by *idmen* and time identifier given by *year*. Additional variables include year of the last migration *myear* (available until year 2014), year of moving into the dwelling *datent*, a dummy variable indicating marriage status *mstatus*, and *location*, which is a categorical variable indicating the geographical location of the household defined as the following:

  - Paris: household locates in Paris.
  - Rural: household locates in rural area.
  - Urban X to Y: household locates in a city with number of inhabitants from X to Y (thousands).

  And one more categorical variable *move* defined after 2014 as the following:

  - the household lives at the same address as in the previous survey
  - the household has moved since last survey

- datind: is a longitudinal data of individuals with individual identifier given by *idind*. Individuals may belong to a household with identifier *idmen*, and time identifier given by *year*. The individual dataset includes basic information on individual's gender, age, and wage. Additional variables include employment status of the individual *empstat*, a dummy variable indicating whether or not the survey is responded by the individual *respondent*, and a categorical variable *profession*, where each code indicates a different profession.

Exercise 1   Basic Statistics

Open the corresponding dataset, and report the following statistics:

- Number of households surveyed in 2007.
- Number of households with marital status “Couple with kids” in 2005.
- Number of individuals surveyed in 2008.
- Number of individuals aged between 25 and 35 in 2016.
- Cross-table gender/profession in 2009.
- Distribution of wages in 2005 and 2019. Report the mean, the standard deviation, the inter-decile ratio D9/D1 and the Gini coefficient.
- Distribution of age in 2010. Plot an histogram. Is there any difference between men and women?
- Number of individuals in Paris in 2011.

Exercise 2   Merge Datasets

In the first part of this exercise, we will learn how to merge datasets.

- Read all individual datasets from 2004 to 2019. Append all these datasets.
- Read all household datasets from 2004 to 2019. Append all these datasets.
- List the variables that are simultaneously present in the individual and household datasets.
- Merge the appended individual and household datasets.

In the second part, we use the newly created dataset from the previous to answer the following questions:

- Number of households in which there are more than four family members
- Number of households in which at least one member is unemployed
- Number of households in which at least two members are of the same profession
- Number of individuals in the panel that are from household-Couple with kids
- Number of individuals in the panel that are from Paris.
- Find the household with the most number of family members. Report its idmen.
- Number of households present in 2010 and 2011.

Exercise 3 Migration

- Find out the year each household enters and exit the panel. Report the distribution of the time spent in the survey for each household.
- Based on *datent*, identify whether or not a household moved into its current dwelling at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.
- Based on *myear* and *move*, identify whether or not household migrated at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.
- Mix the two plots you created above in one graph, clearly label the graph. Do you prefer one method over the other? Justify.
- For households who migrate, find out how many households had at least one family member changed his/her profession or employment status.

Exercise 4 Attrition

Compute the attrition across each year, where attrition is defined as the reduction in the number of individuals staying in the data panel. Report your final result as a table in proportions.

Hint: Construct a year of entry and exit for each individual.