Chapter Six: Data Manipulation

Dr. Simona Simona

4/12/2020

This chapter will address the following topics:

- Reading in data
- Recoding data
- Cleaning data

Most quantitative social science research is based on secondary data as we have already noted in the previous chapter. Secondary data however, is usually messy and disorganised, it does not come to us in a 'ready to be coooked' format. This is because in most cases, these data are not collected for purposes of research. They may be collected for research but certainly not for our research. If we have to use to answer our research questions, we have to do the work of repurposing and that involves transforming and reorganising it so that it can be suitable to answer our research questions.

It is estimated that 80% of the data analyst's time is spent on data manipulation. That is a large chunk of time, but it only illustrates the importance of data manipulation in the data analysis workflow. There are many other terms that refer to data manipulation including data management, data munging, data transformation, janitor work and more recently data wrangling. Data manipulation is the process of getting data from various sources, cleaning and transforming it for visualisation and modelling purposes. In fact, data can only be useful if the data analyst is able to prepare and clean it for substantive modelling work. This process is even much more so in today's world where data availability has expanded exponentially. In this chapter, I introduce facilities available in R programming that help turn messy and unrefined raw data into clear and actionable bits of information. We shall look at some of the most common data manipulation procedures like reading, recoding, subsetting, selecting, merging and saving data. A the end of the chapter we will do a case study of collecting and manipulating the the worldbank data that we are using in this book. We will go through the process of collecting, saving the csv files, reading them in R and working on it until it is in the format we are using it in this book.

Reading in the data

As we can imagine, for us to be analyse any type of data, we first have to bring it into R. When you are a new R user, this process may not be that straightfoward. There are different mathods of reading in data in the R environment and these depend on the file format the original data is saved in and the R package that you are using to read the data in. The most common data file formats that you will encounter would be excel files (.xlsx), comma separated files (.csv), text files (.txt), table files (.tab). You will also reading data saved in other programmes such as SPSS (.sav), Stata

(.dta), SAS (.sas) among others. There are many R packages that can be used to read data into R. We will be considering an appropriate package to use for each data file.

We can read data in using RStudio by clicking on the `Import Dataset` tab in the **Environment** window. But this may not be the best method because you can not assign it to an object of your choice. The data may come in with long and complicated names that may make it difficult to deal with. We could also us `file.choose()` function to interactively read in the data as we did in **Chapter Three**. For large datasets though, this may be complicated to do. We shall instead use the **working dictory** in this chapter.

The working directory

The working directory is a folder on your computer that