

Managing Data in Spreadsheets

The previous chapter started our discussion of file-based data storage with a presentation of several file formats used for tabular data. In this and the following chapters of the book, I present different software tools that are commonly used for file-based data management and processing. While our focus in this book is on the R statistical toolkit, we start with one of the most widely used type of data management software: spreadsheets. A spreadsheet is a big table where users enter data into cells and perform calculations with them. The most common spreadsheet software that probably all readers are familiar with is Microsoft's *Excel* (part of the Microsoft Office suite). Apple's *Numbers* is a similar package, available only for the macOS platform. There are also different free spreadsheet systems such as the *Calc* software that is available as part of OpenOffice or LibreOffice.

Spreadsheets are not especially designed for managing research data, but they can be used for this purpose. This, however, comes with significant limitations. Most of these limitations are due to the fact that spreadsheets were developed to facilitate data management and processing *by humans*, and not by computers. In other words, they support a workflow where humans enter information primarily for other humans to look at. This is why there are almost no constraints to ensure the consistency of a table, but also why spreadsheets have so many features to format data for visual consumption.

Therefore, as we will see below, spreadsheets do not help much to ensure that a table is correctly formatted, and that the content of the cells in a table conforms with the type of a variable. Also, working with spreadsheets means that you manage data through pointing and clicking

with your mouse and through manual editing, which makes it difficult, if not impossible, to replicate your revisions later. It is therefore not a surprise that things can go wrong when you rely on spreadsheet software for data management; the *Economist* even covered some of the most popular cases in a report on Excel errors in science (The Economist, 2016). I nevertheless include spreadsheets in this book, because they – perhaps unfortunately – still constitute one of the main ways in which scholars manage their data. It is my hope that this chapter can prevent you from committing some of the fundamental mistakes that can occur when using Excel or similar tools; we discuss some of them at the end of the chapter.

For the illustration below, we rely on Microsoft Excel Version 2019, the most frequently used spreadsheet software. This is the most recent version of the software for Mac users. While the basic functions we cover here are also available in earlier versions of Excel and in Excel for Windows, they are sometimes accessed under different names and with different menu entries. For the most important functions, I mention these differences in the text. Still, the screenshots presented below are based on Excel for Mac.

5.1 APPLICATION: SPATIAL INEQUALITY

Inequality is traditionally defined as an unequal distribution of income, wealth, or some other quantity across the individuals in a society. However, there are other types of inequality, for example, between different regions in a country. This “spatial” inequality is what we cover in the example for this chapter, by computing a national-level estimate of economic inequality across a country’s different locations. More precisely, we are interested in the extent to which regions in a country differ with respect to their economic performance.

Most economic indicators are provided at the level of individual countries – think of the gross domestic product (GDP) or national growth estimates. While important for comparisons between countries, these national indicators cannot capture variation *within* countries. This is why economists have started to systematically collect data on economic variables at the subnational level for large samples of countries. These data allows us to examine regional variation in economic outcomes, but at the same time compare these patterns across countries.

One of the first global datasets of this type is the G-Econ dataset (Nordhaus, 2006; Chen and Nordhaus, 2011). G-Econ divides up the

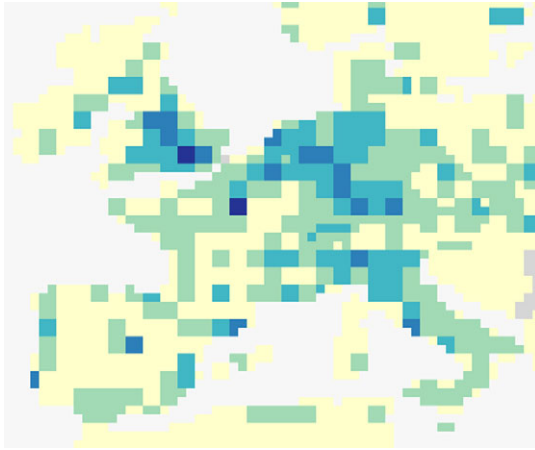


FIGURE 5.1. G-Econ data for Europe in 2005. Visualization created with the PRIO-Grid mapping tool at <https://grid.prio.org/>, see Tollefsen et al. (2012).

different countries into equal-sized, quadratic cells and reports economic indicators for each of them. The most important cell-based indicator is the Gross Cell Product (GCP), which is the equivalent of the GDP, but measured at the level of cells. It contains data for four different years (1990, 1995, 2000, and 2005). The map in Figure 5.1 shows a visualization of the data for Western Europe. We can clearly see the different cells, each colored according to its GCP value with economically wealthy areas displayed in darker shading. The plot shows the strong economic differences within Europe, with London and Paris being among the major economic hubs. Also, we see considerable variation within countries, for example, in Italy (the north vs. the south).

In this chapter, we use G-Econ to show how to perform simple data operations with spreadsheet software, in order to produce an estimate of spatial inequality for the countries contained in G-Econ. We use Version 4.0 of the data, released in May 2011. The online repository for this book contains a copy of the data, see the file `g-econ.xls`. The data comes in a single Excel file with two sheets (tables), shown as different tabs in Excel (see Figure 5.2).

One tab is called “definitions”, and it contains a list with the variables in the dataset and their descriptions. The other is called “GEcon40” and contains the actual data, in a large table with 27,445 rows. The first row in the table contains the column names, each corresponding to one entry in the documentation. Take some time to familiarize yourself with the data.

27442	5101.825	Zimbabwe	390.3988	390.3988	390.3
27443	71.39693	Zimbabwe	707.9332	707.9332	707.9
27444	3474.651	Zimbabwe	665.1368	665.1368	665.1
27445	1844.421	Zimbabwe	619.0456	619.0456	619.0
27446	11.89949	Zimbabwe	536.7775	536.7775	536.7
27447					
27448					

FIGURE 5.2. The different sheets in G-Econ's Excel file.

In particular, see if you can make sense of the variables and the values contained in the dataset. There are a few things to note:

- Some cells do not seem to belong to actual countries. One of the entities in the data is “Antarctica,” which consists of many cells without economic activity (hence the missing values, coded as #N/A). Other small entities are also listed in the dataset (e.g., Svalbard or other small islands). This is not a problem in itself, but we need to keep this in mind when we assign cells to countries.
- For some variables in the data, the measurement units do not seem to be correct. For example, the column DIS_OCEAN is supposed to contain distances to the nearest (ice-free) ocean in kilometers and has many values in excess of 100,000. This is not really plausible, so this is likely a scaling issue and the values are probably given in meters rather than kilometers.
- Variables D1 and D2 are supposed to contain the same information (distance to the coast) according to the codebook, but in the actual data there are some rows where the values for these variables actually differ. Thus, there must be different ways in which the distance to the coast was computed.
- Did you spot how the most important variables in the data, the gross cell products, are encoded in the data? They are stored in the MER* and PPP* columns. Recall that this table design is not optimal, as we discussed in Chapter 3. Data structure and content are not independent, since we need to add additional columns to the table if we want to add more years to our database.

How do we work with spreadsheets in practice? In the following, we go through some basic steps to familiarize ourselves with spreadsheets, and to compute national-level estimates of spatial inequality. Since most readers will be familiar with spreadsheet software, much of this should be easy and straightforward.

5.2 SPREADSHEET TABLES AND (THE LACK OF) STRUCTURE

Spreadsheets do not care about data structure: Columns do not have clearly defined types (such as text or numeric columns). Rather, you can insert any value into any cell of a spreadsheet. This is why there is no real data definition you have to do before you can work with a table; if you need a new table, all you have to do is open a new Excel workbook (a collection of tables), or add a new empty sheet to an existing workbook (using the + sign next to the different tabs, see Figure 5.2).

In a spreadsheet, each table always has the same rectangular structure, where columns are labeled with capital letters and rows with numbers. This is problematic, since for a social science data table we usually want to name variables (i.e., columns) ourselves. Therefore, in a spreadsheet, we usually define columns by inserting their names in the first row, as is done in the G-Econ dataset. However, this is a *convention* rather than a requirement of the software – for example, nothing prevents you from adding new data *above* the row with the column names, which would break the structure. You may have noticed that in the G-Econ table, the first row appears to be fixed – it does not move when you scroll up and down in the table. While this does not mean that Excel treats the column names in any other way, it is a simple display adjustment to always keep the header names visible. This is called “freezing” the display of a part of the dataset. You can enable this by selecting the entire line below the row you want to freeze, and then clicking on **Window** >> **(Un)freeze Panes**.

Also, Excel does not really care about the type of data that goes into particular columns. All you can do is change the formatting of the cells. Simply select the entire column by clicking on the column header (e.g., column G for the DIS_LAKE variable in G-Econ) and click on **Format** >> **Cells**. This brings up the dialogue box in Figure 5.3.

In the list, you can choose different formats for the cells – since you selected an entire column, any settings you change here apply to the entire column. Now select “Number” and tick the box to use the 1000 separator (a comma). This will change the display of the different values in the column, but it does not define a fixed type for the column such that it stores, for example, only text or only numbers. Rather, we can still mix numbers and text in that column, as we do, for example, for the variable name in row 1 (text) and the data values in the remaining columns (numbers). Defining a format for a set of cells does not change their internal values, it only affects how they are displayed on the screen. In the above example, Excel still keeps simple numbers in the cells, but

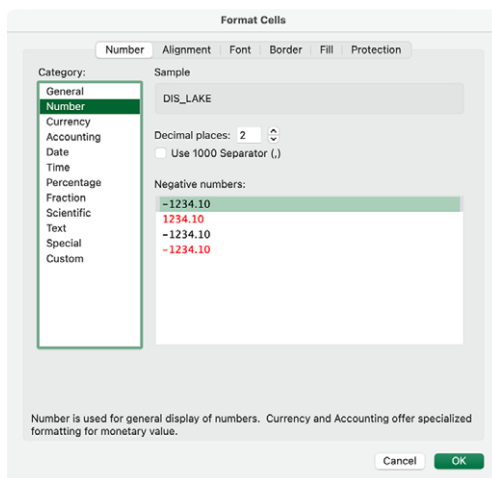


FIGURE 5.3. Excel's cell formatting dialogue box.

adjusts their display such that, for instance, the value 156602 is shown as 156,602.00. This may sometimes be confusing, since *what we see is not exactly what Excel stores internally*. When you select a cell, you can always view its true, unformatted content in the formula bar at the top, below the toolbars.

In Excel, each workbook consists of one or more sheets (tables) and it is stored in a single file. These files use the extension .xlsx, or .xls for the older legacy Excel file format (which is what the G-Econ database does). Any modifications you make to the data or the appearance of the table must be saved to the file. This is the same for all the common spreadsheet tools, where the data content and the presentation of the table are kept in the same file. None of these tools allow us to define strict types for the columns in a table, which would avoid coding mistakes and erroneous input.

5.3 RETRIEVING DATA FROM A TABLE

In many cases, you need to retrieve subsets of your table: For example, you may be interested only in selected variables from the G-Econ dataset, which are required for an analysis you want to conduct. As many data operations in spreadsheets, this is a manual operation, where you select the columns you need by clicking on their header (the capital letters), and then copy-paste them to a new sheet. You can select multiple columns at

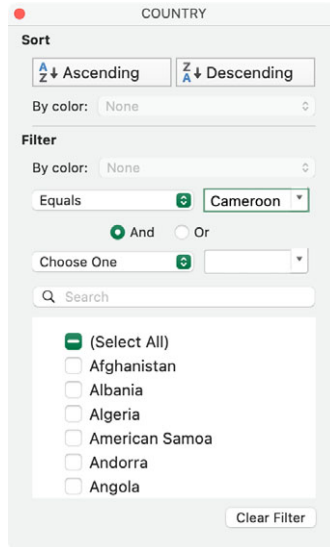




FIGURE 5.4. Excel's sort/filter dialogue box.

the same time – if they are not adjacent, you can hold down **command** (Windows: **Ctrl**) while clicking to select them. The retrieval of particular rows works in a similar way, by clicking (while pressing the Command/Ctrl keys if necessary) on the row numbers on the left. If you want to extract a set of rows with a particular value, this can be done by using one of Excel's filtering features. For example, let us extract all the cell values for Cameroon from G-Econ: Select the COUNTRY column, go to the **Data** tab in the menu, and select **Filter**. This displays a small arrow symbol next to the column name. Click on it, and you will see the dialogue box shown in Figure 5.4.

This allows you to sort and filter your data. For example, to extract the entries for Cameroon in G-Econ, choose the operator **Equals** in the drop-down menu in the Filter section, and make sure that only **Cameroon** is selected in the input field. This will display all rows for Cameroon from the data – the original row numbers remain the same, but are shown in blue to indicate that some rows are not displayed. As you can see, you can specify other filtering conditions, such as values that **Begin with** a particular sequence of characters. If you apply this filtering mechanism to a column with numbers, you will be given different selection operators, such as **Greater than** or **Less than**. You can also link different conditions to each other using logical operators such as **AND** or **OR**. Once you have

filtered the rows that you are interested in, you can select and copy-paste them to a different sheet for further processing.

Excel also allows you to sort the data in your spreadsheet, using again the dialogue box in Figure 5.4. If you select the COUNTRY column, bring up the Sort/Filter dialogue ( ) and click on Ascending (Windows: A to Z), Excel will first ask you about the extent of the data that should be sorted. Here, you need to select **Expand the selection**, or Excel will *only sort values in the COUNTRY column*, which breaks the entire logic of a table because the values in the sorted column are assigned to different rows. While filtering is temporary in the sense that it displays a subset of the table but leaves the underlying data unchanged, sorting permanently changes the order of the rows in the table. Also, it is useful to note that Excel implicitly assumes that the first line of the table (the one containing the variable names) is static and therefore does not include it in the sorting. Importantly, again, this is an *assumption* the software makes, since there is no mechanism in Excel that lets you assign fixed column names that the software then works with.

5.4 CHANGING TABLE STRUCTURE AND CONTENT

Due to the absence of a fixed data structure that is maintained by the software, changes to the logical design of the table and its content are easy: You can access new columns simply by using some of the empty ones in your spreadsheet, or insert them by right-clicking on the header of a column. While there is an upper limit to the number of columns in any sheet (16,384), this is unlikely to matter in practice since tables of this size are impossible to navigate by a user. For each new column you insert, you can follow the convention to specify the variable name at the top, although, again, this is something that the software does not require. As for all the others, there is no preset type for the new column, but you can change the display of the values by formatting cells as we have shown above.

Changing the actual data in a spreadsheet is also done in a similar way, simply by manually editing the content of the cells in your table. Alternatively, similar to a word processor, you can use the standard search and replace feature to do this for multiple values. For example, take a look at the PRECAVNEW80_08 column in the G-Econ table. There are many erroneous entries in this column with the value #DIV/0!. If you want to work with the dataset for your analysis, these values can cause problems

down the road: Obviously, we do not really know what to do with these values, and the occurrence of text in a column that is supposed to be numeric can interfere with mathematical operations you perform with it. Select the `PRECAVNEW80_08` column by clicking on its header, and go **Edit** **Find** **Replace** (Windows: **Home** **Find & Select** **Replace**). In the dialogue box that shows up, enter the value we are looking for (`#DIV/0!`) and the value we want to replace it with. In this case, it is probably safest to simply remove the strange values, which we can do by leaving the “Replace with” field blank. If you then click “Replace all,” the erroneous values in the column will disappear.

The above operations for changing your data structure and updating the data values in your table show the data workflow that is common to all spreadsheet tools. Most of the data work consists of manual operations: navigating to a particular cell of a table and typing in a certain value. This workflow may sometimes be convenient, for example, when we create a new dataset that requires human coding and manual input. But even for these tasks, the lenient approach of spreadsheets when it comes to data types and structure can cause problems, since the software does not automatically recognize mistakes (e.g., when entering text for a variable with only numeric values). Thus, the software lets you do almost anything with your data – this freedom is likely something you pay for with inaccuracies in your data and inefficiencies in your workflow. However, before we turn to the question of how you get your data out of a spreadsheet, let us first complete our description of the two remaining data operations and how they can be done in Excel.

5.5 AGGREGATING DATA FROM A TABLE

Excel has different features that allow you to aggregate your data. For our purpose of computing an indicator of spatial inequality, we present only one here: The Pivot table. A Pivot table is a tool to summarize data in a flexible way, for example, by allowing users to introduce different grouping dimensions over which the values in the data can be aggregated. This is what we need for our application. Let us start with a simple example that counts the number of cells per country in G-Econ. Before we proceed, you need to select the entire G-Econ table to make sure that all the data is included in the Pivot table. You can do this via the menu (**Edit** **Select All**) or by clicking the square box at the top left of the table. Now, you can access the **Summarize with PivotTable** feature via the **Data** menu (Windows:

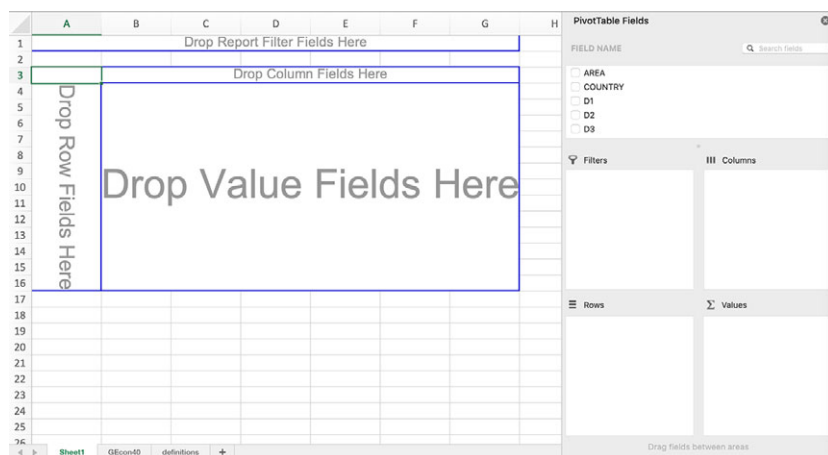


FIGURE 5.5. Setting up a Pivot table.

Insert > **Tables** > **PivotTable**). It starts with a simple dialogue box where you need to specify the subset of the data you want use as the basis for your Pivot table. In our case, this defaults to the part of the sheet that is filled with data, so there is nothing to change here. Also, we choose to have the new Pivot table placed in a new sheet as part of your Excel workbook.

This creates a new sheet with an empty roster for a Pivot table. The structure of such a table is simple: it consists of column and row fields, which are the main levels for grouping. The value field at the center of the table is the one that displays the summarized data. Figure 5.5 shows the basic configuration of a Pivot table.

The part on the right is where you set up the field(s) you want to use as grouping level(s) for your table. You can simply drag and drop field names from the box at the top into the empty boxes below. First, drag and drop the **COUNTRY** field into the “Rows” box. This adds all the different values for this field (the individual countries) as rows in your Pivot table. So far, however, we do not have any summary values we are computing for each of these countries. To set this up, drag the **PPP2005_40** field (the gross cell product computed using Purchasing Power Parity for the year 2005) from the list at the top and drop it into the “Values” box at the bottom right. This way, Excel computes aggregate (summary) statistics over the gross cell product for each of the grouping levels (currently, the countries on the left). The default summary statistics Excel computes is to *count* the number of observations for each country, which is why the entry in the “Values” box reads “Count of PPP2005_40.” For Afghanistan, for

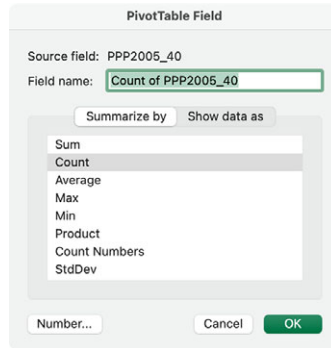


FIGURE 5.6. Changing the aggregation function for a Pivot table.

example, there are 92 cells in the dataset, while Albania is much smaller with only 9 cells.

Now, let us compute the *standard deviation* of all gross cell products for each country, as a measure of the spread of the gross cell product values and hence the spatial inequality. For this, we need to change the aggregation function; rather than counting the number of observations, we want to compute the standard deviation of all gross cell product values for 2005 for each country. You can change the aggregation function by clicking on the small “i” symbol next to the entry in the “Values” box (in Excel for Windows, this is a little arrow that brings up a drop-down list, where you can select “Field Settings”). This brings up a dialogue box, shown in Figure 5.6.

The count function (highlighted) is what we currently use for the Pivot table. If you change this to StdDev, the values in the Pivot table will reflect the standard deviation of the gross cell product values for each country. This turns out to give us many invalid values, such as the familiar #DIV/0! values (which indicate a division by zero), but also #N/A values, which are present in the original data sheet. Here, we see some of the problems resulting from Excel’s (and other spreadsheets’) sloppy use of data types. In the main data, we have many values in the (numeric) gross cell product variable that are not numbers but strings (text). Since we cannot compute a mathematical sum over text values, Excel simply outputs these values directly. We will later work with tools that require us to specify a fixed type for each column of a table and that allow for a consistent coding of missing values to properly exclude them from computations.

5.6 EXPORTING SPREADSHEET DATA

We now have the values of spatial inequality and could continue to analyze them in Excel. However, in order to show you a complete workflow and a nice R plot at the end of this chapter, we export the data to a text file (CSV) for further analysis in R. To create a CSV file from our Pivot table, we simply save the worksheet with the table in this format: **File** » **Save As**. There are different CSV file formats available; I recommend that you choose the one with UTF-8 support. During the export process, Excel warns you that only the active sheet will be saved to CSV, since this file format cannot deal with multiple tables in one file – each CSV file is supposed to contain exactly one table. You can acknowledge this warning and proceed with “OK.” The second warning refers to the fact that many Excel features such as colored cells or formatted text are lost when saving the file as a CSV. This is why you should not keep the entire workbook in this format (the CSV file will be saved anyway).

5.7 RESULTS: SPATIAL INEQUALITY

If you take a look at the values we have computed in our Pivot table above, there are some that reflect our basic intuition of spatial inequality. You see that some of the inequality estimates are very large, for example, for South Korea (86.04) or the United Kingdom (67.63). These are countries that are very strong economically, but where economic activity is disproportionately concentrated in single economic centers such as Seoul or London. Although comparable in terms of overall economic performance, other countries such as Germany (41.93) have lower values of spatial inequality, because there are several economic hubs in the country.

A plot of the overall distribution of spatial inequality using the exported CSV file (see Figure 5.7) shows that a large number of countries have very low values of spatial inequality, or in other words, a relatively even geographic distribution of economic activity. This could be due to a number of reasons. For example, the size of the country could affect its spatial inequality; if a small country only consists of three cells, the spread of economic activity within that country will likely be limited. Another reason could be the level of economic development. Many countries have very low levels of economic activity throughout, which will also affect the scores we have computed. All this suggests that the way in which we compute spatial inequality may not be entirely satisfactory, and that other measures may be preferable.

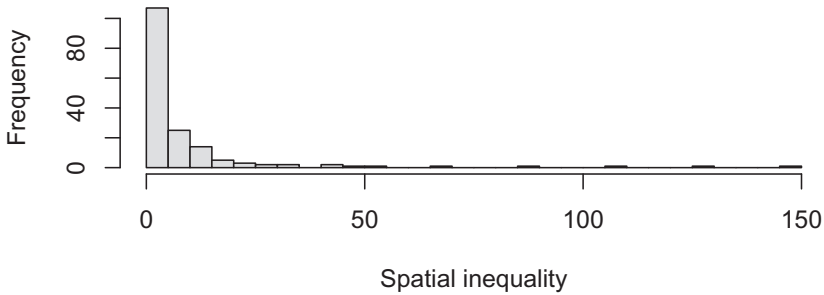


FIGURE 5.7. Histogram of spatial inequality scores.

5.8 SUMMARY AND OUTLOOK

Although we have only scratched the surface in terms of Excel's functionality, you should now have a good idea of how data processing with spreadsheets works. Spreadsheets leave you considerable freedom when it comes to the structure of your table(s), and the data you store in them. They also let you adjust the visual display of your data. For example, you can increase the font size for a particular cell, or change its background color. This is because spreadsheets are designed for a mix of different tasks: for data *storage* in a (loosely defined) tabular structure, but also for the visual *presentation* of your data in different fonts or colors. The interaction with spreadsheets works mostly by means of manual editing; for example, you navigate the spreadsheet using your mouse, and edit content, delete rows or columns, or execute other tasks such as sorting your data.

The features that spreadsheets offer may be suitable for certain tasks in the data collection process. For example, if you manually put together a human-coded dataset, spreadsheets can serve as useful and intuitive tools for coders to enter the coded information. However, for many (if not most) other tasks in social science data management and processing, spreadsheets are not a good choice: The lack of a pre-defined structure of your table makes it difficult to spot and fix errors in your data, such as wrong labels, inconsistently coded missing values, or text in numeric columns. Also, the fact that you work with spreadsheets through manual interactions makes it difficult to replicate the data processing steps you completed. There is no record of what you did; if something went wrong, you likely have to repeat most steps yourself. If others want to replicate your work, it would be hardly possible.

While in general I explicitly recommend against the use of spreadsheet software in data management, there are certain applications for which it can be useful, for example, for the manual creation of a new dataset. Also, several datasets in the social sciences are still distributed in Excel format, which is why it is difficult to avoid this software completely. If you need to manage and process your data with Excel, here is some advice for avoiding major problems down the road. The suggestions below apply in particular to a data processing workflow where the data is eventually exported from Excel to be analyzed in a statistical toolkit, such as R or Stata. If you use Excel solely to prepare your data for humans to look at, most of them do not apply. Much of what I mention here overlaps with recommendations from other scholars, see, for example, Data Carpentry (2017).

- *Use one table per sheet:* In our example above, we worked with different sheets that are all part of the same Excel file (recall that you can switch between them using the tabs at the bottom). I strongly recommend that you store at most one table per sheet, not multiple ones. This makes it easier to maintain a consistent type for a given column. Also, exporting the data becomes much easier, since you can select the entire sheet – and not just a subset of it – and save it to a separate file.
- *Stick to the rectangular table format:* Within any given sheet, strictly keep the rectangular table structure intact. That is, do not let sub-headers or any other layout elements interrupt the table structure. Also, do not use the “merge cells” feature, which allows you to combine adjacent cells into bigger ones, as this also breaks the structure of the table and creates major problems when exporting your data. The same applies to comments you place somewhere outside the main table. If you want to record and preserve comments in a spreadsheet, create a separate column and place your annotations there. It also makes your life easier if you do *not* leave a margin of empty cells around your main table, as is often done for visual purposes. The top-left cell (A1) is where your main data should start.
- *Use proper variable names:* A correctly formatted table requires that columns be named. In Excel, you do this by placing the variable names in the first row. You can use any text as variable names, but most other software packages are more restrictive here. This is why you should not use white spaces, special characters, or mathematical symbols in variable names. Also, numeric characters at the beginning of variable

names are usually not a good idea (and not permitted in some tools, for example in R). If you require variable names with multiple words, you can separate them with an underscore (as in `new_variable`). Since you may later be using these variable names in statistical code, it is useful to keep them relatively short.

- *Make sure that data values are valid:* Since spreadsheet software such as Excel lets you enter almost anything into the cells of a table, it is up to you to make sure that the data in a particular column complies with the type of that column. For example, a column recording the time of an observation (the year) should only have values such as 1816 or 1946, but not 1990s or 2001 (and 2002). The former are purely numeric values, the latter are not. Non-numeric values can be recognized in Excel due to the fact that they are left-aligned in the cell, while true numbers are right-aligned. Alternatively, you can use one of Excel's functions such as `ISNUMBER()` to test whether a content of a cell is a number, and similar functions also exist for other types. It is also important to use consistent coding for missing values; I recommend to leave these cells empty.
- *Do not use formatting elements to store information:* A frequent mistake when using Excel is to use formatting elements such as cell coloring, font styles, or others for storing information. For example, if you create a country-level dataset with information about whether a country is democratic or autocratic, it is *not* a good idea to color democracies in green and autocracies in red. This information can be used by humans only, and it is lost when you export the data in any format other than Excel. Therefore, it is best not to care about font styles, cell backgrounds, etc., but stick to the defaults set by the spreadsheet software.