CHAPTER IV

# Exploring and Managing Data

## Introduction to Data Exploration

*What is data exploration?[[1]](#footnote-1)*

Data exploration is usually the first step after you have sourced data with the right format. Users conduct data exploration to understand the information they gathered and to begin analyzing the data. Data in real life often comes in a non-rigid manner that is not user-friendly. For analysis, you will need to narrow down the large bulks of raw data that you collected and prepare the data for moving forward to data analysis, narration, and visualization.

Often, you will find data from different sources in different formats. By exploring the data, you will form a general idea about the basic attributes of the dataset that include: what the variables are (how many observations there are), if the dataset contains missing values, if the data make sense in real life, etc.

*What questions do we have in mind when exploring data?*

The key to exploring data is to prepare the data for further analysis. You will first need to look at the license and owner of data and make sure that you will be able to use it and share your findings. Second, you should look at the format of data and see if you can work with the specific data format. When proceeding to the third step, you can delve deeper into the content of data and see if the dataset includes the information you want, and if it is relevant to the question you have in mind. A more detailed process of data exploration is included below.

## Exploring and Managing Data in Excel

Excel is a spreadsheet program that is used for storing, organizing, and manipulating data. Many open data sources provide data in Excel format, which is easily downloadable and allows for others to analyze the data without having to convert the data to a different format. Excel can be used to sort, explore, manage, and analyze data. Again, keep in mind that Excel is not an open-source software.

*How is Excel organized?*

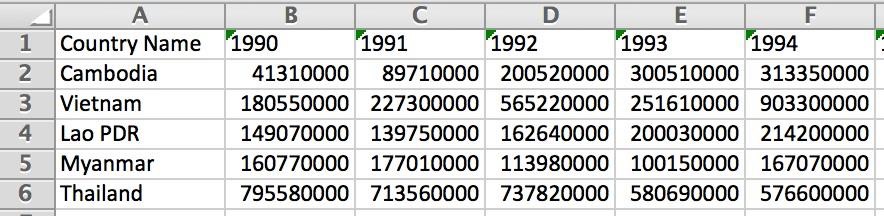
Each spreadsheet is made up of a grid of cells formed by columns (identified by letters) and rows (identified by numbers). Each cell has a unique reference made using the column letter and row number, such as B6 or AA10. Cells can contain data as text, numbers, or formulas. Formulas make calculations based on information contained in other cells, or aggregate data across a range of cells to calculate statistics such as the average, median, or sum. An Excel workbook can be comprised of multiple spreadsheets.

*Exploring data in Excel*

When first opening data in Excel, take a moment to become familiar with the contents:

* *What variables are presented?* Are these the variables you expected or need?
* *Are there missing values?*
* *Is the data formatted in a crosstab?* A crosstab will have one variable in the columns and another in the rows, such as countries listed in the rows, years listed in the columns, and each cell containing the value of official development assistance (ODA) received by that country in that year (see example of *Incorrect format* below).
* *Do variables have consistent coding?* For example, the variable *respondents sex* may have text values such as female or male in some cells, while binary variables (0 or 1) in others. Another example would be multiple spellings of a country name: Democratic Republic of the Congo, DRC, Dem. Rep. Congo.
* *Are there multiple spreadsheets in the workbook?* Check along the bottom of the screen to see if there are multiple tabs (each representing a different spreadsheet). If so, click through to see what data is presented in each spreadsheet.

Once familiar with the data presented in the workbook, it is time to decide if the data needs to be formatted or cleaned in any way. The standard way to format data for analysis is to list the variables across the top of the spreadsheet in Row 1. Each observation is then listed in each row. This means that data presented in a crosstab format will need to be reformatted for analysis.



*Incorrect format: Country name in row, year in column*

*Correct format: Variable names in Row 1*

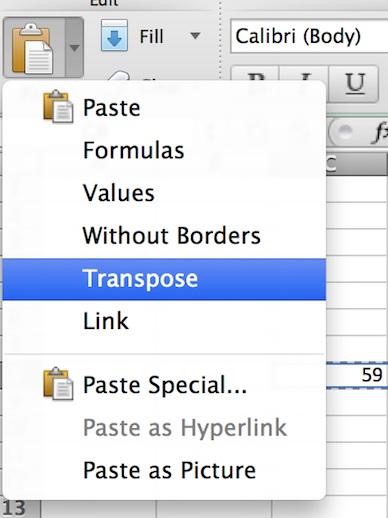
*Managing Data in Excel*

*Import data*

To import data that is not currently saved in Excel format, open up Excel and navigate to “File” and choose “Import”. From there, select the file format of the data source you would like to open in Excel. If it is a text file and each data value is separated by a common character (this is called a delimiter: it could be a comma, colon, period, etc.), you can define the delimiter when importing the data and Excel will separate the data values into individual cells.

Similarly, if you open data in Excel and there are multiple data values within a single cell that are separated by a delimiter, go to the “Data” tab and choose “Text to columns”. This command can distribute the contents of a single cell across multiple cells.

*Format data*

If the data is not formatted correctly (if it does not have the variable names in Row 1 and only one observation per row), it is necessary to reformat the data before proceeding with analysis. Helpful tips for formatting data:

* Start a new sheet with variable names in Row 1. Data will be copy and pasted to this new sheet.
* To repeat information in multiple cells (such as a country name), type the information in a cell. Hover over the cell until the small black square appears in the bottom right corner. Drag the square down to cover the cells where the information is to be repeated.
* If reformatting data in a crosstab format, the paste command “transpose” will transfer the data from horizontal to vertical, or vice versa.

*Putting data into tables*

You will have some additional controls over data if you put it into a table (such as sorting and filtering as described below). To put data into a table, select the area of the data and click on the “Table” tab and select “New”. You will see that your variable names are now treated as table headers.

*Sort data*

Select a column of alphanumeric data in a range of cells, or make sure that the active cell is in a table column containing alphanumeric data. On the Data tab, in the Sort & Filter group, do one of the following:

1) To sort in ascending alphanumeric order, click Sort A to Z.

2) To sort in descending alphanumeric order, click Sort Z to A.

*Creating filters*

When you [put your data in a table](https://support.office.com/en-us/article/put-your-data-in-a-table-e81aa349-b006-4f8a-9806-5af9df0ac664), filtering controls are added to the table headers automatically. For quick filtering:

1)Click the arrow Filter drop-down arrow

Filter drop-down arrow in the table header of the column you want to filter.

2) In the list of text or numbers, uncheck the **(Select All)** box at the top of the list, and then check the boxes of the items you want to show in your table.

3) Click **OK**.

*Pivot tables in Excel*

Creating pivot tables in Excel is a great way to analyze and summarize large data sets. To create a pivot table, highlight the data to be included in the pivot table and go to the “Insert” tab. Select “Pivot table”. This will bring up a dialogue box - click OK. The new pivot table will appear on a new worksheet. On the right hand side of the worksheet will be a window to control what variables appear in the pivot table. When desired variables are checked, they will appear in the pivot table and Excel will automatically sum the totals in each category. In order to change from sum to some over measure (such as average), right click on the column header “Sum of [variable name]” and select “Field Settings.” Options will be given to display the sum, count, average, standard deviation, product, and variance.[[2]](#footnote-2)

*Formulas in Excel*

A formula performs calculations or other actions on the data in your worksheet. A formula always starts with an equal sign (=), which can be followed by numbers, math operators (such as a plus or minus sign), and functions, which can really expand the power of a formula.

## Exploring and Managing Data with R

*What is R and RStudio?*

R is a language and environment for statistical computing and graphics. R can be considered a different implementation of S, a statistical programming language for organizing, visualizing, and analyzing data.

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

*When do we use R?*

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and more) and graphical techniques, and is highly extensible (users and developers can expand R and add to its capabilities). The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

For detailed instructions on how to explore data with R, please refer to Appendix A.

1. Explanation from Techopedia, <http://www.techopedia.com/definition/28789/data-exploration> [↑](#footnote-ref-1)
2. More resources on creating pivot tables can be found at [Office Support](https://support.office.com/en-US/article/Overview-of-PivotTable-and-PivotChart-reports-527c8fa3-02c0-445a-a2db-7794676bce96)  [↑](#footnote-ref-2)