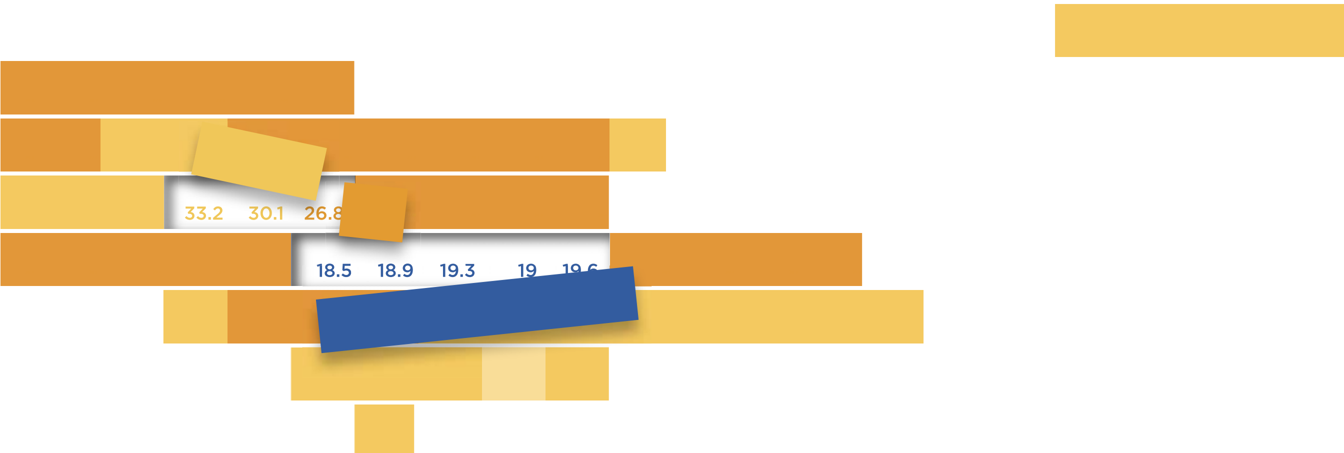
DATA AT WORK

Best practices for creating effective charts and information graphics in Microsoft® Excel®

JORGE CAMÕES

FREE SAMPLE CHAPTER



DATA AT WORK

Best practices for creating effective charts and information graphics in Microsoft® Excel®

JORGE CAMÕES

New Riders
VOICES THAT MATTER™
www.newriders.com

DATA AT WORK
Best practices for creating effective charts and information graphics in Microsoft® Excel®

Jorge Camões

New Riders
Find us on the Web at www.newriders.com
New Riders is an imprint of Peachpit, a division of Pearson Education.
To report errors, please send a note to errata@peachpit.com

Acquisitions Editor: Nikki Echler McDonald
Production Editor: Kim Wimpsett
Development Editor: Dan Foster
Copy Editor: Jan Seymour
Proofreader: Scout Festa
Compositor: WolfsonDesign
Indexer: Karin Arrigoni
Cover and Interior Designer: Mimi Heft

ISBN 13: 9780134268637
ISBN 10: 0134268636

9 8 7 6 5 4 3 2 1

Printed and bound in the United States of America

To my family

Acknowledgments

I'd like to first thank Alberto Cairo. In non-English speaking countries, there are a few oases when it comes to publishing original data visualization books, but the landscape is basically a barren desert. I wanted to help change that, so I wrote the first manuscript of this book in my mother tongue, Portuguese. I then asked Alberto if he would read it.

Not only can Alberto read Portuguese, but we also share a similar view of what we think data visualization is all about, in spite of working in different areas. To make a long story short, he liked the book and introduced me to his acquisitions editor, Nikki McDonald, and so my data visualization journey took a turn. With the help of Nikki, my development editor Dan Foster, copy editor Jan Seymour, and production editor Kim Wimpsett, my poor manuscript became a real book. Alberto read several chapters of the English version and provided invaluable feedback.

Stephen Few also read a few chapters and saved me from myself once or twice, for which I am very appreciative.

If I know how to make a few charts that you can't find in the Excel charts library, that's because I learned it from, or was inspired by, Jon Peltier, the true Excel charts master. I'm deeply grateful to Jon for all the knowledge and generosity he shared with the community for well over 10 years.

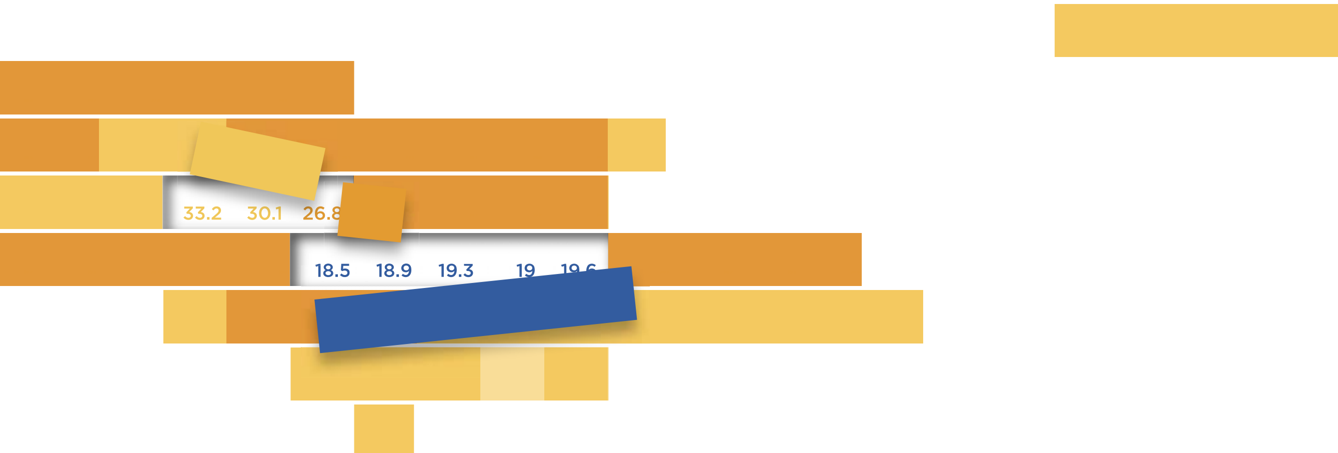
Andreas Lipphardt's untimely death was the saddest moment along my data visualization journey. I wrote a few posts for his company's blog, and we talked often about working together in the future. I still wonder what would have happened if we had.

Finally, I thank my family. When I was more interested in writing this book than actually make a living, Teresa and the kids were very patient and supportive.

About the Author

Jorge Camões studied statistics and information management and has been consulting businesses on how to effectively use information visualizations since 2010, with clients in the top 25 pharma companies and major retailers. Prior to starting his consulting business, Camões worked for 10 years in the business intelligence department of the Portuguese subsidiary of Merck & Co. Camões runs the popular data visualization blog Excelcharts.com. He works from his home in Lisbon, Portugal.

Contents



INTRODUCTION

No data point is an island,
Entire of itself,
Every data point is a piece of the continent,
A part of the pattern.

The venerable poet John Donne must be turning in his grave with this paraphrase of his beautiful meditation “No man is an island,” but I couldn’t find a better way to express the nature of data, which have a context and a web of relationships. The path to knowledge lies in discovering and making these relationships visible.

Social change and technological progress have made the world a more uncertain place. As another poet, Luís de Camões (not related), said, “Change doesn’t change like it used to.” In an effort to cope with uncertainty, we put technology at the service of mass data production and retrieval. This has been called by many names over the years. Today we call it “Big Data.”

Acquiring and storing data has become the goal; the more data, the better. But are we missing the point? We no longer need *more* data if it's not accompanied by the right skills that turn it into truly *better* data. We need to consider how those who need the data will use it, and for what purpose. Otherwise, it's pointless to continue accumulating useless data, collecting digital dust in a forgotten folder on a hard disk. Waiting. Or, worse yet, making pie charts.

A Quantitative Change

Suppose that the data you work with is now updated daily rather than monthly, multiplying its total volume by 30. As Arthur C. Clark told us, **a quantitative change of this magnitude forces a qualitative change** in organizational culture, our attitude toward data, and data's role in decision making. Just imagine if the data allowed you to react to whatever is happening (rather than merely acknowledging what happened weeks ago) so that you become aware of its impact on all levels of the organization, beginning with how each person interprets their roles and tasks.

Only a planetary catastrophe would prevent the ever escalating volume of data. In the past, much of human experience was absent from our data monitoring systems, but it's now beginning to be quantified. In a few years, we'll reminisce affectionately over the complaints about information overload that we have today.

This is where data visualization begins. But beware. Data visualization is marketed today as the miracle cure that will open the doors to success, whatever its shape. We have enough experience to realize that in reality it's not always easy to distinguish between real usefulness and zealous marketing. After the initial excitement over the prospects of data visualization comes disillusionment, and after that the possibility of a balanced assessment. The key is to get to this point quickly, without disappointments and at a lower cost. This book is designed to help get you there.

A Language for Multiple Users

Data visualization helps us manage information. To make the most of this information, we must first accept the fact that "data visualization" does not exist *as a single entity.* Instead, think of it as a blanket term: It exists differently for each group of people who use it.

Visualization is like a language. Paraphrasing the Portuguese writer José Saramago, “There is no English; there are languages in English.” For example, although people from the United States, Wales, and South Africa all speak English, they’d likely have some difficulty communicating because their versions of English are all so different, having changed from their common core over the years based on their geographical and social contexts.

Data visualization is a graphical language, used differently depending on the “speaker.” A graphic designer, a statistician, or a manager starts from the same foundations of data visualization, but each has different goals, skills, and contexts, which are reflected in their different visualization choices.

A Wrong Model

Imagine that we all wish to write poetry. For the unfortunate not blessed with the gift of rhyming, the word processor offers some models that help with writing reports in the form of folk poetry. Seems absurd? Well, this is what happens with data visualization, too, when we look to spreadsheet chart templates to help overcome our weaknesses.

Graphic designers have made visualization the fashion phenomenon it is today—their poetry meant to be seen by large audiences and evidenced in data journalism, books, blogs, and social networks. Results vary between the brilliance of many visualizations in the *New York Times*, for instance, and the mediocrity of many infographics created by marketing departments as clickbait.

Meanwhile, millions of charts made with spreadsheets remain hidden within business organizations. The obscure, everyday users of office tools, unaware of better visualization models adapted to their contexts, mistakenly see the designers’ work as a reference to imitate, often with catastrophic results. Peer pressure, the *this-is-what-the-client-wants*, vendor sales tactics, and a lack of training feed the illusion that there is beauty in bad poetry.

There is not. **The purpose of data visualization in organizations is not to make beautiful charts; it is to make** effective **charts.** And, as we shall see, if your charts are effective, they’re also likely to be beautiful, even in aspects with strong associations to aesthetics, such as the use of color.

A Better Model

Visualizations crafted by graphic designers are often appealing, but in a business context we can't use the same model. At a time when graphic literacy in organizations is still low, we must evaluate this model's usefulness, beginning with four simple concepts:

- **Process.** Visual displays of information in business organizations and in the media have different goals and different production and consumption processes, which should not be mixed up.
- **Asymmetry.** Information asymmetry—whereby one party has more or better information than the other—is generally less evident within an organization than, say, between journalists and their readers. Graphical representations must adapt to this difference, adding detail in the former and finding the core message in the latter.
- **Model.** If you hire a data visualization expert, make sure she is aligned with your organization's specific interests or focus, because her data visualization model may prove incompatible with the organizational culture, daily work processes, available tools, and skill sets. It's almost impossible, for example, to convince an Excel user to learn a few lines of code, so this cannot be an expectation.
- **Technology.** Almost everything you need to understand about data visualization can be learned and practiced in a spreadsheet, which is an everyday tool people are familiar with.

Today, business organizations are encouraged to become more efficient and effective. Improving the return on investment (ROI) of their data should be a top priority. This is achieved by adhering to data visualization principles and best practices, and especially through a change of perspective, which has negligible costs, both in absolute financial terms and when compared to the results of past practices.

In fact, **many data visualization best practices are no different from the rules of etiquette**. A set of rules that is merely a ritualization of common sense is easy to understand, but must be internalized and practiced.

In short, data visualization in an organizational context has unique characteristics that must be identified and respected. The display of business data is not art, nor is it an image to attract attention in a newspaper, or a moment of leisure between more serious tasks. Business visualization is first and foremost an effective way to discover and communicate complex information, taking advantage of the noblest of our senses, sight, to support the organization's mission and goals.

Data Visualization for the Masses

I write a blog about data visualization (excelcharts.com), and over the years I have often been tempted to move away from the worksheet and devote myself to true visualization tools. This would be the normal path. But the spreadsheet is the only tool that the vast majority of us have access to in an organizational context, and getting data visualization to the average person must start from this contingency if we want to encourage learning and increase graphical literacy. Then, at some later point, people and organizations will assess whether the tool adequately satisfies their needs and can then make a natural and demanding transition to other applications. Or not.

This is therefore a book about data visualization for the masses—that is, for those who, with the support of a spreadsheet, use visual representations of data as an analytical and communications tool: students in their academic work, sellers in their sales analysis, product managers in planning their budgets, and managers in their performance assessments.

The Labor Market

Taking into account the economic circumstances of today, is it justified to invest in statistics, data analysis, and data visualization skills? As I mentioned, with the exception of a scenario of global catastrophe, it's difficult to imagine a future that does not involve an increase in the volume of data and the need to use it. In fact, these skills are becoming central to the vast universe of what we call "knowledge workers." Compared to other skills, these skills cut across more areas of activity, ensuring some competitive advantages in the labor market within the expected social, economic, and technological trends.

A study[1] by consultants McKinsey & Company on "Big Data" estimates that in 2018, in the United States alone, there will be a shortage of up to 190,000 people with high analytical skills, and a shortage of about 1.5 million managers and analysts with analytical skills to use data in the process of decision-making.

1 McKinsey & Company. Big data: The next frontier for innovation, competition, and productivity. 2011.

It's wise to read these reports with some skepticism, of course, considering their unknown agendas. Nevertheless, this study indicates the need for qualified human resources in this area, of which data visualization is an essential part.

My View of Data Visualization

I have on my desk a report that includes hundreds of charts, all of which are inefficient, ugly, and useless. There isn't a single chart *I* am proud of. And, yes, it was I who made them, many years ago, as one of my first professional tasks. Even more embarrassing is that I remember the report's commercial success.

I had not yet realized it, but working with data would become as normal for me as breathing. I didn't pay much attention to it at the time, until one day I stumbled upon a book: *The Visual Display of Quantitative Information*, by a certain Edward Tufte. For me, this was the Book of Revelation. In it, I discovered data visualization as a concept and as a field of study, and it was love at first sight.

Over the years, I realized that there are no universal rules and goals in this field. Subjectivity, personal aesthetic sensibilities, the task at hand, the profile of skills and interests, the audience—these all conspire to minimize things that we take for granted, such as the importance of effectiveness in the transmission of the message.

Within this relativism, the easy answer is to accept that anything goes. Throughout this book, you'll see examples of dead ends where this path sometimes takes us. But if we accept that there is no one-size-fits-all perspective, and that there are no universal rules, we still must seek a coherent theory for each group of practitioners and consumers.

My view of **data visualization is an exercise in everyday normality**: Simply give the eyes what they *need* to see, so that the visualization goals are met at minimal cost, in the same natural way we use vision to check whether we can cross a roadway.

To take advantage of vision, we must understand that there is no difference in nature between the physical landscape around us and the graphical landscape we create on a screen or on a sheet of paper.

Organization of the Book

This book follows a narrow path between theory that's too abstract to be useful for everyday tasks and practice that's too focused on a concrete task to help us understand the general rules. I tried to follow this path in every chapter, showing how theory applies in each example and how the specific task always has a theoretical framework that explains, justifies, and generalizes it. It's important to understand *why*, not just how.

To begin to understand data visualization, the first part of this book describes the context in which the action takes place: the characteristics of the human senses, the objects we use when making charts, the role of perception, how knowledge is acquired, and the many ways of defining data visualization.

In the second part of the book, we'll recognize that a chart is a visual argument, an answer to a question, and that the quality of this answer begins with the chart type you choose. Then, we'll format the chart. You'll see that the best chart formatting serves the content and is not distinguished from it, praising its qualities and reducing its flaws.

Throughout the book, we'll analyze data visualization in an organizational context, including good practices in data management, the Excel chart library, how to avoid bad software defaults, and how to use application flexibility to go beyond what the Excel library seems to offer.

The Limits of This Book

I wrote this book with a particular reader profile in mind: those who are not paid professionally for their aesthetic talents and artistic skills.

You might find this problematic, because designing a chart seems to require these skills. But I totally reject that. You need not be artistically talented to create effective charts.

I believe in increasing graphical literacy, and for that to happen we can help build a safety net of basic criteria for producing effective visual representations. I believe this will be useful at the professional level and will also contribute (marginally) to a more critical citizenship.

This book focuses on identifying the basic principles of data visualization for an organizational environment, as performed by individuals who have certain skills and who use a very specific tool: the spreadsheet. The intersection of these factors defines the main limits of this book:

- **Major visualization types.** In the first chapter, you'll see data visualization classified into three major groups: charts (we define "charts" in the first chapter), networks, and maps. Although they have some common principles, networks and maps are excluded from this book because they have a specific vocabulary that must be addressed in the proper context.

- **The chart.** A chart is just one part of the information communication within an organization, just like a single paragraph of a story. Since this is an introductory book, there will be a balance between this concept of the "graphical landscape" and the idea of a chart as the minimum unit of data visualization.
- **Excel.** The spreadsheet software I use now is Excel 2016, with which I made all the charts for this book. When it was necessary to refer to application features and capabilities, I tried to be as generic as possible in order to include other versions of Excel and even other spreadsheet programs.
- **Chart types.** Due to its flexibility, Excel allows us to go beyond its library. Throughout this book, you'll find many examples of this flexibility. But there are hard limits (charts that Excel just can't do) and soft limits (charts that would be so difficult to create and with such a low cost-benefit ratio that in practice we should not attempt to use them regularly). For Excel, networks and maps represent such exceptions.
- **Not a manual.** Although written with Excel users in mind, this book is not a manual of techniques, tips, and tricks.
- **No retouching.** It's important for me to ensure that the charts you'll find in this book are true to the original made in Excel, so they have not been retouched by additional software, even in the management of text elements, in which Excel is especially limited. However, for inclusion in the book with the highest possible quality, the charts were exported to PDF, which led to some minor changes that I have tried to minimize.

There's also a practical limitation regarding the data. I wanted to use real data, not some fake business indicators, but this poses problems of confidentiality and limited interest. To circumvent that, I used official statistics as a proxy for business data. Except for a few specialized contexts, we can use the same methodology and chart types. Both are in deep need of a more effective approach.

Break the Rules!

Data visualization is not a science; it is a crossroads at which certain scientific knowledge is used to justify and frame subjective choices. This doesn't mean that rules don't count. Rules exist and are effective when applied within the context for which they were designed.

You'll find many rules in this book—so many rules that the temptation to break them (intelligently) may be overwhelming. If this is your case, congratulations, that's the spirit. I myself could not resist and tried to test the limits and possible alternatives. I invite you to do the same.

Companion Website

As I said, this book is not a manual. It will not teach you how to make a chart in Excel. You won't find even a single formula.

That's why we set up a comprehensive companion website for the book:

- **dataatworkbook.com**

On the website, you'll find:

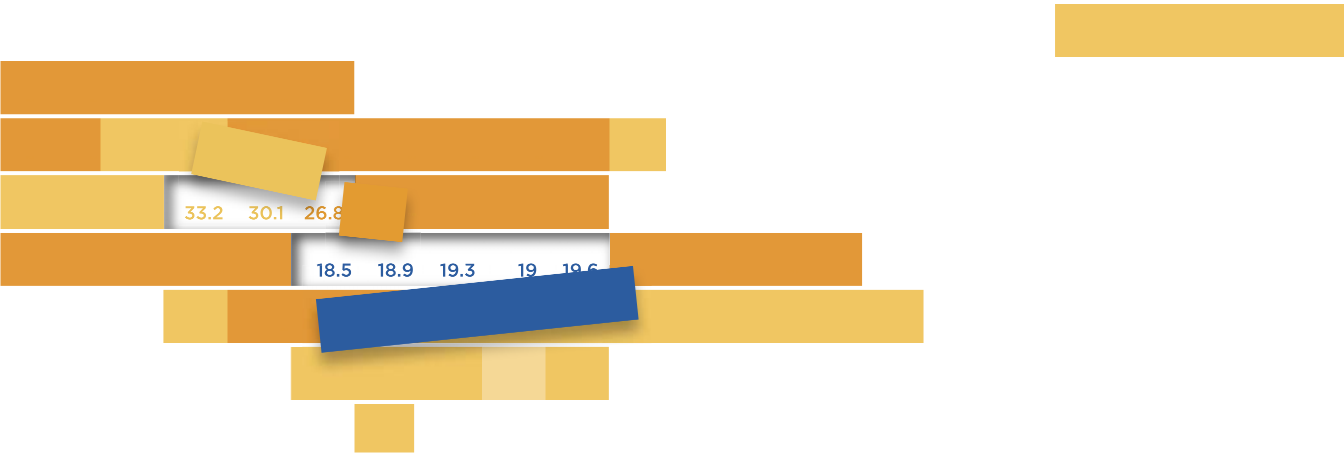
- All the relevant original charts in Excel files that you can download and play with. I've also included brief comments for each chart to help you learn how to make them. When you see the icon, it means that the chart is available to download.
- Links to the original data sources and, when possible, a dynamic bookmark to the most recent data.
- Links to other content referenced in the book. You'll find icons sprinkled throughout the book that invite you to read a relevant paper, watch a video, go to a web page, and so on. When you see this icon, it means that you'll find a link on the companion website.

I welcome your comments, suggestions, and change requests. I ask you to add them liberally on the website for the benefit of all.

I'll try to be aware of comments and suggestions made on social media and consumer reviews on major online book retailers and address them on the book's website, if needed.

Over time, I'll add original charts not published in the book as well as additional resources, so be sure to check in often.

You can find me on most social media, but I confess that Twitter is the only service I use regularly. I will tweet about new content, so if you follow me (@camoesjo) you won't miss it!



4

DATA PREPARATION

Jacques Bertin defines his semiology of graphics as a "visual transcription of a data table." In a perfect world, this table materializes in front of us when we need it, ready to use. In everyday reality, however, things involve more sweat and less magic. People coined the expression "data janitor" for a reason.

In a data visualization project, data extraction costs and data preparation are often overlooked, either by management that doesn't understand the level of detail required or by data analysts making overly optimistic assumptions. This translates into many hours of data cleansing that most people don't see. If not taken into account, these labor-intensive tasks can consume several times the resources available for a project, whether it's a simple chart for an upcoming meeting or an organization-wide project.

Brilliant visualizations cannot redeem bad data, either in content or in structure. Many spreadsheet users are not familiar with well-structured data, and that's one more reason to discuss data preparation.

We can summarize all preparation work on the data table, regarding both structure and content, by the acronym ETL, for *Extract, Transform, and Load*. ETL is just as applicable to your Excel files as it is to large, formal systems.

This chapter is not strictly about data visualization. If the tables you need actually materialize in front of you, ready to use, if you know how to structure the tables to take advantage of pivot tables, and if you organize sheets in your workbook by content type, it's probably safe to skip this chapter. In a more sophisticated organization, most of the issues discussed here are not relevant, and most of the data comes from internal systems. However, many people still struggle with these basic issues, so if you're in this category, read on.

Problems with the Data

Let's split data problems into two broad categories: 1) **structure without content**, and 2) **content without structure**. The first category affects our data in particular; the second is common in data we get from other sources.

Structure without Content

Even if you've never seen a table for which multiple users can enter data (such as a table for telemarketing operators), you can imagine how much garbage data is collected: incomplete ZIP codes, multiple abbreviations for the same entity, misspellings, logical inconsistencies...you name it.

It's challenging to define good data validation rules without forcing exclusions: What happens when a few ZIP codes are missing from a lookup table? Suppose, though, that you can maintain a table with a minimum number of errors. **Figure 4.1** represents an example of such a table. To make things more interesting, try linking this table to a second table containing other personal data (**Figure 4.2**). First, you'll have to split the field Name into Name and Surname, to be able to join both tables. Now, is John Doe in the first table the same person referred to as John F. Doe in the second table? The solution in these cases is to have common fields in both tables that are not subject to different interpretations (social security or driver's license numbers are good candidates). If there are no safe common fields, you'll need to allocate additional resources to determine whether it's the same person. Multiply this process by thousands of records and you have a problem on your hands that, if not anticipated, would generate serious time and resource management issues.

ID	Name	Surname	Address	City	Zip Code	State
1000	John	Doe	S Main St	Torrington	CT 06790	Connecticut
1001	Mary	Poppins	SW 11th St	Lowton	OK 73501	Oklahoma

Figure 4.1 A table with names and addresses.

ID	Name	Gender	Age	Height	Weight	Marital Status	Children	Occupation
1001	Mary T. Poppins	Female	34	5.38	182	Married	4	Librarian
1000	John F. Doe	Male	82	6.17	138	Widower	2	Retired

Figure 4.2 A table with socio-demographic characteristics. To get a better feel for structure without content, imagine that there are many more rows (records) and many entry errors in them.

A few other special cases also belong to the category of structure without content. One of the most common is a break in a time series, whereby you still get the same measure (an unemployment rate, for example), but changes in methodologies, concepts, technologies, or regional administrative boundaries make comparisons meaningless. Or, at least, comparisons must be carried out with extra care—the same care you should use when comparing countries that use different ways of measuring the same reality. For example, infant mortality rate depends on how a country defines "live birth." Because the definition is not the same across countries, this may affect country rankings in international comparisons.[1]

Content without Structure

Suppose you're a data provider, perhaps at the U.S. Census Bureau or at a small public relations company. The moment you release the data, you cease controlling it. You don't know how people will read and *reuse* the data. They may want to cross-check it if they suspect that the data is not telling the whole story. Or they will misunderstand the concepts. Whatever they do, first they must have access to the data in a format they can use.

Providers often make it hard to use the data beyond the format in which they released it; they're often unaware of this issue or focus on the end user and forget the data professional, who probably needs a more specific format.

1 MacDorman, Marian F. and Matthews, T.J. "Behind International Rankings of Infant Mortality: How the United States Compares with Europe." *NCHS Data Brief*, No. 23, November 2009.

Go to the web page

Data providers should then ask themselves two simple questions: How many data reuse issues are we causing by releasing the data in this format? Is this reuse friction level acceptable for our data dissemination goals?[2] Typical answers are, respectively, "a lot" and "no." The end result is that data reuse friction levels can range from none (rare), to mildly annoying, to a source of a barrage of unprintable curses.

Go to the web page

Let me give you an unfair example. Suppose you want to know the military budget as a percentage of GDP in each country. There are several sources, but you could start with the CIA's website publication *The World Factbook*. Country profiles in the *Factbook* contain several sections and subsections.

Figure 4.3 displays the Military section for the United Kingdom. You can manually open this section and copy the data you need for each country, or you could use a scraping tool that automates the process. If you're unable to automate the process, you'll have a few long and boring days ahead of you. Because the data are not displayed the way you need it, time and resource costs will increase since you'll have to structure it first.

Figure 4.3 UK Military data in the *The World Factbook* from the CIA.

Transportation :: UNITED KINGDOM

Military :: UNITED KINGDOM

Military branches:

Army, Royal Navy (includes Royal Marines), Royal Air Force (2013)

Military service age and obligation:

16-33 years of age (officers 17-28) for voluntary military service (with parental consent under 18); no conscription; women serve in military services, but are excluded from ground combat positions and some naval postings; must be citizen of the UK, Commonwealth, or Republic of Ireland; reservists serve a minimum of 3 years, to age 45 or 55; 17 years 6 months of age for voluntary military service by Nepalese citizens in the Brigade of Gurkhas; 16-34 years of age for voluntary military service by Papua New Guinean citizens (2012)

Manpower available for military service:

males age 16-49: 14,856,917

females age 16-49: 14,307,316 (2010 est.)

Manpower fit for military service:

males age 16-49: 12,255,452

females age 16-49: 11,779,679 (2010 est.)

Manpower reaching militarily significant age annually:

male: 383,989

female: 365,491 (2010 est.)

Military expenditures:

2.49% of GDP (2012)

2.48% of GDP (2011)

2.49% of GDP (2010)

country comparison to the world: 28

Transnational Issues :: UNITED KINGDOM

2 I'm not implying they do it on purpose; they may not be able to reduce friction due to technological reasons.

I said this is an unfair example because the *Factbook* actually allows us to jump between the country profile level and the list level. At the bottom of the page on the website, you'll see "country comparison to the world: 28." If you click the number 28, you'll get a list of all countries sorted by military expenditures as a percentage of GDP. Then you can choose a country from that list and return to the profile view. This nice feature is still quite rare, unfortunately.

These two broad categories of structure without content and content without structure try to make sense of the variety of issues when using data presented in an unfriendly format. Hadley Wickham brilliantly captured the difference between well-structured and poorly structured data in an excellent article[3] in which he quotes the first paragraph of Leo Tolstoy's *Anna Karenina*: "Happy families are all alike; every unhappy family is unhappy in its own way." The "happy family" dataset is structured according to some rules that make it similar to other "happy families," while **there is a virtually infinite number of ways to create an unhappy dataset**.

What Does "Well-Structured Data" Mean, Anyway?

The acronym GIGO (*garbage in, garbage out*) summarizes the issues we deal with every day: Results and insights depend on data quality. We can handle data critically (being aware of the "garbage" and factoring it in to the evaluation of results) or uncritically ("if the data has been subject to extensive processing by the computer, it can't be wrong").

Data integrity becomes essential when the volume of data increases and we need to update, filter, and aggregate it, and use data as a basis for derivative calculations. A clean, consistent, and well-structured table means lower update and maintenance costs and more flexibility to multiply the perspectives from which we can analyze the data.

This may not be good news for the user accustomed to the loose spreadsheet environment, where storage, presentation, intermediate calculations, and parameters often share the same sheet. Let's start untangling this mess with a concrete example.

3 Wickham, Hadley. "Tidy Data." *Journal of Statistical Software*, Vol. 59, No. 10, August 2014.

The first step toward improving data structures is understanding that storing data and presenting data are two very different things. You should never use storage and presentation features together in a single worksheet. Share your source table if requested, of course, but otherwise bury it deep down in a data-only sheet. If you have a well-structured table, you'll never have to touch it again, except when using a client like a pivot table or when adding a variable. **In Excel, tables are for storing data, and pivot tables are for analyzing and presenting data.**

A Helping Hand: Pivot Tables

Ah, pivot tables! Pivot tables are great at many levels. They can even serve as a litmus test for checking how well a table is structured. If every single cross-tabulation is done easily and you don't have to change the pivot table following an update, you can be reasonably sure that you have a well-structured table.

Figure 4.4 shows a sample of one of the output formats for the Consumer Expenditure Survey. Assuming we know the meaning of the Series ID, this is the typical manner of *presenting* the data, with time periods in columns and entities in rows.

Consumer Expenditure Survey
Years: 1984 to 2013

Series ID	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
CXU080110LB0101M	35	30	30	28	28	33	30	31	28	30
CXU080110LB0102M	26	24	24	23	23	24	23	26	23	24
CXU080110LB0103M	36	29	28	29	28	32	30	34	27	33
CXU080110LB0104M	37	32	34	28	28	32	29	29	29	31
CXU080110LB0105M	38	33	34	30	30	36	35	35	29	32
CXU080110LB0106M	43	35	32	33	30	37	33	34	35	33
CXU080110LB01A1M	36	31	30	28	28	32	30	32	28	31
CXU080110LB01A2M	34	29	27	28	27	35	29	28	26	29
CXU190904LB0101M	30	29	31	31	30	33	35	42	43	46

Figure 4.4 Sample output from the Consumer Expenditure Survey (Bureau of Labor Statistics).

Go to the web page

Think of the table as a cross tabulation (Series ID × Year) that must be uncrossed so that we can use it. Unlike other output formats from the Bureau of Labor Statistics, you can get all the data you need in a single table, and it's very easy to reverse it to the right format, resulting in the table you see in **Figure 4.5**.

Series ID	Year	Value
CXU080110LB0101M	1984	35
CXU080110LB0101M	1985	30
CXU080110LB0101M	1986	30
CXU080110LB0101M	1987	28
CXU080110LB0101M	1988	28
CXU080110LB0101M	1989	33
CXU080110LB0101M	1990	30
CXU080110LB0101M	1991	31
CXU080110LB0101M	1992	28
CXU080110LB0101M	1993	30

Figure 4.5 Un-pivoting the data table.

Series ID contains multiple variables, so we must parse it and look for the descriptive text for each code. **Figure 4.6** shows how the final table will look.

Category	Item	Quintile	Year	Value
Food Total	Eggs	Lowest 20	2012	39
Food Total	Eggs	Lowest 20	2013	40
Food Total	Eggs	Second 20	2012	47
Food Total	Eggs	Second 20	2013	52
Food Total	Eggs	Third 20	2012	49
Food Total	Eggs	Third 20	2013	56
Food Total	Eggs	Fourth 20	2012	59
Food Total	Eggs	Fourth 20	2013	59
Food Total	Eggs	Highest 20	2012	71
Food Total	Eggs	Highest 20	2013	76

Figure 4.6 A few rows of the final data table.

Creating dynamic charts in Excel requires knowledge of advanced formulas, but often we only need them because the data table is not properly structured. **Figure 4.7** shows a simple dynamic chart (not a pivot chart) that you can create without a single formula. It displays the proportion of food expenditure away from home, over the years, for the selected income quintile. Select a different quintile and the chart will update.

From Figure 4.6 we can see that **a well-structured table is essentially a list of observations and their characteristics** (category and item, income quintile, and time) **and the associated measure** (expenditure). In a pivot table, measures are usually placed in the Values area, while characteristics go into the Rows, Columns, or Filters areas.

Download the original chart

Category	Food Total
Quintile	Highest 20

Percentage	Items		
Year	Food	Food at home	Food away from home
1984	100%	53%	47%
1985	100%	52%	48%
1986	100%	51%	49%
1987	100%	50%	50%
1988	100%	51%	49%
1989	100%	49%	51%
1990	100%	49%	51%
1991	100%	55%	45%
1992	100%	55%	45%

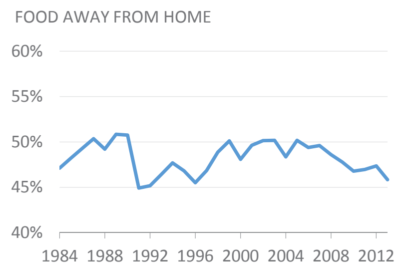


Figure 4.7 A dynamic chart using a pivot table.

In a well-structured table that can be easily used as a pivot table source, the content of each column must be understood as a group (years, quintiles), and the values in each measure should be comparable (expenditure in dollars in a column and expenditure units in a second column).

Reality can get more complicated, and so will the structure. Suppose you get expenditure by gender. Ideally, you'd add a new column ("Gender") with two values (Male, Female). But if they are averages instead of totals, you can't aggregate them, and, in this case, you have to add them as measures.

Extracting the Data

You successfully complete the first stage in the ETL process when you access a file that you can edit and manipulate. When you get a text file, you may need to open it in a text editor (such as the free Notepad++ for Windows) to solve multiple small issues with Search and Replace. Do your computer's regional settings and the text share the same symbols for decimal places and thousands separators? (Some may use periods while others use commas.) Are there any strange characters? Can they be removed?

Extraction can be a very long and rocky journey, so let's start with a smooth example first, again from the Bureau of Labor Statistics. I'm looking for the monthly unemployment rate, at the state level, for a period of several years. **Figure 4.8** shows a sample of the output. There are several output options, including an Excel file, but for now we'll work with a tab-delimited text file. I'm getting the data for each state, which means that I'll have to consolidate them into a single table, removing all unwanted text.

THE WAY YOU PASTE DATA CHANGES THE OUTPUT

Scenario 1: Direct paste from web page to Excel

Series Id: LASST010000000000003

Seasonally Adjusted
Area: Alabama
Area Type: Statewide
Measure: unemployment rate
State/Region/Division: Alabama

Series ID	Year	Period	Value
LASST010000000000003	2010	M01	11.7
LASST010000000000003	2010	M02	11.6
LASST010000000000003	2010	M03	11.3
LASST010000000000003	2010	M04	10.8
LASST010000000000003	2010	M05	10.4
LASST010000000000003	2010	M06	10.1
LASST010000000000003	2010	M07	10.0
LASST010000000000003	2010	M08	9.9
LASST010000000000003	2010	M09	10.0
LASST010000000000003	2010	M10	10.1
LASST010000000000003	2010	M11	10.2
LASST010000000000003	2010	M12	10.3

Scenario 2: From web page to Notepad+ and from Notepad+ to Excel

Series Id: LASST010000000000003

Seasonally Adjusted
Area: Alabama
Area Type: Statewide
Measure: unemployment rate
State/Region/Division: Alabama

Series ID	Year	Period	Value
LASST010000000000003	2010	M01	11.7
LASST010000000000003	2010	M02	11.6
LASST010000000000003	2010	M03	11.3
LASST010000000000003	2010	M04	10.8
LASST010000000000003	2010	M05	10.4
LASST010000000000003	2010	M06	10.1
LASST010000000000003	2010	M07	10
LASST010000000000003	2010	M08	9.9
LASST010000000000003	2010	M09	10
LASST010000000000003	2010	M10	10.1
LASST010000000000003	2010	M11	10.2
LASST010000000000003	2010	M12	10.3

Figure 4.8 Pasting data into Excel, from a web page and from a text editor.

Figure 4.8 explains why you should have a text editor between a web page and the spreadsheet. Scenario 1, on the left, shows the result of a direct paste from the web page, while scenario 2 shows what happens when you paste to Notepad++ first: Excel recognizes the tab character and automatically parses the text.

As with the example on expenditure, we'll have to find what the Series ID codes mean. You may want to split the Series ID codes into multiple columns using the Text to Column function in Excel. Also, create a real date from the Year and Period columns.

When extracting data from other public sources, you may run into some limits imposed by the organization. The United Nations Population Division doesn't allow you to select more than five variables or countries in each query (**Figure 4.9**). Other organizations impose limitations at the cell level. The Eurostat limits each query to 750,000 cells. Depending on how high the limit is or how detailed are the data you need, you may have to run multiple queries to get all the data and then merge the results into a single file.

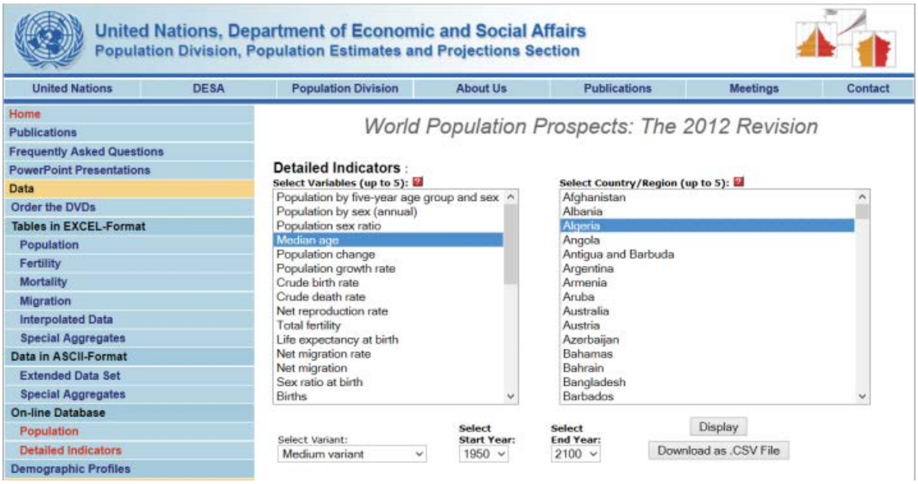


Figure 4.9 Extracting data from the UN Population Division.

The PDF Plague

With more or less pain, the chance of getting a text file from official statistical offices is high. Other data providers, such as professional associations, may have other, more restrictive policies regarding data dissemination.

Many years ago, I needed to get data on the various types of electricity consumption (high voltage, low voltage, domestic, industrial, public roads, and so on) at a very detailed regional level. The data were available only in large sheets of paper, where someone had elegantly *handwritten* all these thousands of values. It was an admirable job, almost worthy of a Charles Dickens novel. It also had an unanticipated cost, because my organization had to purchase a copy of all those sheets and hire someone to enter the data manually.

Today, no sane organization would share its data in this format. With all the technology we have in our hands, that would be ridiculous, right? Well, not so fast. Let's abstract for a moment from the technology and focus on the goal: getting a few thousand values into an editable table. Now tell me: What difference does it make if we have handwritten numbers on a sheet of paper or a PDF file with such a twisted formatting that the cost of extracting the data is higher than entering them by hand? Actually, there is a difference: I found those handwritten sheets only once, while I keep stumbling upon data tables in PDF files, to my despair and exasperation.

If you're a data provider, you have a degree of control over your data when you share them in a PDF. You might persuade some people not to use the data in a way different than you intend. This is not wrong if you have a strong reason to do it, but it will anger your users, even if that's not your plan. Again, make sure that the way you share your data is aligned with your goals. In addition to presenting your data the way you want people to see it by default, provide a link to the raw data. That way everyone is happy.

If you're a user of internal data, you might assume that you'll never have to extract data from PDF files. But, sooner or later, you will. And there will not be a quick fix. You may be able to open simple and well-behaved PDFs in Word 2013 or 2016, so there's no harm if you try that first. If that doesn't work, try copying the data from the PDF and pasting it into the text editor (such as Notepad++), and then from the text editor into Excel. Then you can try an additional application, such as the free tool Tabula, to extract the data into CSV or XLS files. None of the solutions will be entirely satisfactory, but the cost of editing the table should be lower than manual data entry.

Go to the web page

"Can It Export to Excel?"

Internal business intelligence (BI) systems should allow you full control over the content you want to extract and how you want to extract it. Unfortunately, that's not always the case. Let me paint a grim and somewhat exaggerated picture here.

First, you have to solve a *communications* problem. You, the business user, and the IT people apparently don't speak the same language: They don't understand why a market share above 100 percent is not possible, and you don't understand that they must have a rule for each of your beloved exceptions. So when you get the data from IT, crosscheck it to make sure you've got the right data.

Second, there is a *political* problem. The data you want and the way you want it may not fit into the current formal corporate policies regarding access privileges, data security, or data dissemination. You can also be caught in a power struggle between IT and other areas, and they may start dragging their feet to avoid granting you access to the data.

Finally, there may be *technical* issues. The eternal question "Can it export to Excel?" forced BI vendors to make this option available. After so many years, I think they still hate it, judging from the output files I have to deal with. If the application can export data to CSV or Excel, there's hardly a reason to create unfriendly table structures that force the user to take additional steps to clean the data. This means extra work for you, but if in every update the format is wrong but consistent, you might use a macro to correct it and solve the problem.

Cleansing Data

I'll assume that you survived the previous stage of the ETL process and you're now the proud owner of a nice-looking table. But the smile will vanish from your face if you now find a record of a *123-year-old* new mom living in a city called *Cincinatti, TX.*

The second stage of ETL, transformation, deals with data manipulation, but the first transformation, data cleansing, is so important and specific that it deserves to be promoted to its own step. Data cleansing suggests, of course, that the data is dirty. Data is dirty because it contains typos or inconsistencies or fails in some way to meet a standard.

All this "dirt" must be cleansed before any serious analysis can take place, and again a pivot table can be very handy for this purpose. If you count every category in a field, you'd soon find only one reference to Cincinatti, TX, while there are many references to Cincinnati, OH. So, you'll probably need to change that record because the city name is misspelled and associated with the wrong state. And what about the 123-year old new mom? Check the age range. She's probably only 23. Please note the word "probably"; **just because a value seems strange, that doesn't mean it's not real**. Be sure to cross-check against a lookup table and against other fields for logical inconsistencies, and don't forget to have a log that includes all your edits.

Transforming Data

One of the benefits of making data cleansing an autonomous step is that now transformation can focus on adapting the dataset to the goals of the analysis. If you're using a spreadsheet, you're now moving from the cell level to the column level where you add, remove, or change variables. Here are a few examples of possible data transformations:

- **Encoding:** If a column includes answers to an open question (where there are no predefined answers), you must add one or more columns to categorize those answers. For example, if you asked people to name three of their preferred movie actors, you'd have to parse the answer and code every one of the names.
- **Aggregation:** The level of detail may be excessive for the purposes of analysis, and we'll need to aggregate the data at a higher level. Our 23-year-old new mom can belong to a larger category (for example, ages 20–24), or data at the daily level can hide a pattern that can only be spotted at the week level.

Read the blog post

- **Derived data:** If we're studying obesity and have weight and height data, we can calculate Body Mass Index (BMI) and add it as a new variable.
- **Removal:** Changes in project scope may make some of the observations irrelevant, or some variables may only be needed to calculate derived data (like BMI above). Keep in the dataset only the data you need.
- **Standardization:** If we need to link our new table to other tables in our system, some standardization may be needed, including changes in table structure and in labeling (for example, M/F instead of Male/Female).

Loading the Data Table

The last stage of the ETL process occurs when the data becomes usable. This can take many forms, such as uploading the file to a system such as a new table, appending the file to an existing table such as an update, or, in Excel, simply changing the data format from a range to a table. In recent Excel versions, you can also add the file to the data model.

Data Management in Excel

It's hard to find a tool that, like Excel, combines power, flexibility, and ease of use for some basic tasks when compared to other similar tools. The problem is that Excel training often focuses too much on the tool and leaves out task-specific aspects.

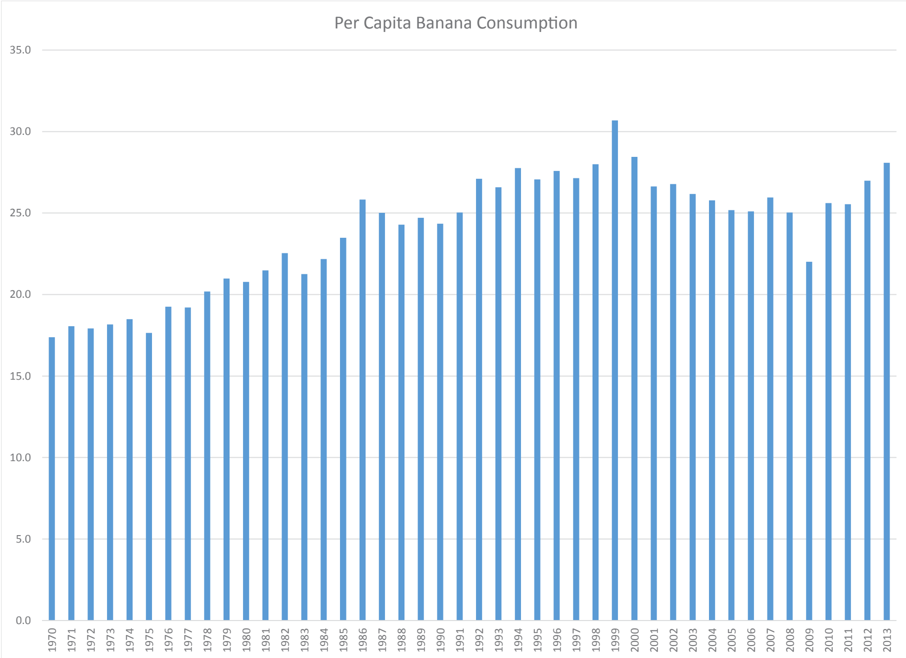


Figure 4.10 A default chart when pressing F11.

For example, take chart making. Knowing how to "make charts in Excel" and knowing how to "make charts" are two different creatures. Give a monkey a banana every time it presses F11, and you get a (very low-paid) Excel chart maker (**Figure 4.10**).

The same happens with the data. Unlike database applications, Excel does not impose any kind of structure, and because users lack the right training, they believe that this is the natural way to manage their data. Sure, people in IT make data structure a top priority, but they don't really understand business needs, do they?

Many organizations can gain much if there's a greater mutual understanding of IT and user roles. Users must obtain a minimum level of literacy with data structures. They must see how structuring the loose spreadsheet environment maximizes the power of functions and formulas that take advantage of that environment (pivot tables and lookup formulas, for example). This simplifies chart making, adds interaction, and reduces updating and maintenance costs. IT personnel and data users may sometimes have a conflicting relationship, but a greater proximity and understanding may help them all realize that users are not always a danger to system security, and IT is not always unaware of business needs.

Organizing the Workbook

The number of worksheets in an Excel file is virtually unlimited, and, surprisingly, we can use all we want without incurring extra costs. Hence, an Excel file that has some level of complexity must be organized in a way that clearly separates the results (charts, tables), intermediate calculations, parameters, and data tables in different, specialized sheets.

Links Outside of Excel

An IT-managed BI system in an Excel-centric organization risks becoming a dual BI system in which users get the data from the formal system, but all the actual analysis is done in Excel. This can quickly get out of control, with isolated file archipelagos in each computer, and impossible-to-reconcile data.

You can't eradicate Excel as a BI tool unless you uninstall it. The organization should have a better understanding of why users keep using Excel. If the formal BI model can't address those needs, it should provide direct access to data in a safe and controlled manner, which again requires a closer relationship between users and IT.

The ideal scenario is to create one or more tables that closely match the user's needs, connected to her workbook and from where she can refresh data.

Formulas

When one of the papers that shaped recent economic policy worldwide[4] draws conclusions based on faulty Excel formulas, and when news of millions of dollars being lost due to spreadsheet errors is common, the least we can do is to assume that a formula is a potential threat. With all other things equal, using fewer formulas makes a spreadsheet simpler to maintain, improves performance, and produces fewer errors.

Calculations with a database query are faster and errors are often easier to spot (you get to the needle-in-a-haystack frustration level much faster in Excel than when using database queries). You can connect your workbook to a query in an external database that performs all the calculations before feeding the data into the spreadsheet. And there are many other ways to avoid formulas, such as

---

4 Reinhart, Carmen M. and Kenneth S. Rogoff. "Growth in a Time of Debt." *American Economic Review: Papers and Proceedings, Vol. 100, No. 2: 573-578,* 2010.

Go to the web page

using pivot tables instead of aggregate formulas or using a data model instead of lookups. Array formulas and calculations in tables are also safer and faster. Finally, named ranges are your friends; use them extensively.

So, as a mantra, you should think, "Avoid Excel formulas." This seems to contradict the very nature of the application, but when you avoid formulas, your workbook becomes safer and more solid. Note that the point is not to turn your workbook into a formula-free zone (that's almost impossible) but to think about better alternatives. Also, you should infer from the techniques suggested above that "avoid formulas" doesn't equal "hardcode data" (entering a value instead of a formula).

Cycles of Production and Analysis

There is a major difference between business visualization and media infographics.[5] Unlike most infographics, which aren't updated after they're published, business visualizations usually include a set of representations that remain useful from cycle to cycle and cut across the organization. Charts on market share and growth are updated for each cycle. They are seen at various levels of regional detail and are common to the multiple markets in which the organization operates.

Think of business charts as the three Rs of ecology:

- They should be *reused* across multiple markets.
- They should be *recycled* by updating the data.
- Their number should be *reduced*, making business visualization more cost-effective at multiple levels.

This does not cover all the data visualization needs in an organization, and you may use many charts only once, but try to evaluate whether a chart has the potential to be used more than once. If the answer is "yes," you should evaluate whether it makes sense to spend extra resources to prepare it for repurposing (by adding interaction or creating a database query, for example).

This is just a small part of the many things that relates to data management in Excel. If it were possible to synthesize this management in a single word, that word would be "structure." Recent Excel versions have introduced new features that suggest more investment in the data structure (including tables, data models, Power Pivot, slicers, PowerBI, and so on). This, in turn, allows you to manage a growing volume of data more effectively.

Go to the web page

5 Check the work of one of my preferred designers, Adolfo Arranz, at Visualoop to make the concept of differences at several levels crystal clear.

Takeaways

- Data preparation is possibly the least thankful part of any data visualization process because it is slow, invisible, and undervalued. If you don't have access to a properly formatted table, assume that you'll spend much more time than anticipated preparing it.
- Pivot tables can help you structure your data tables.
- Although you can paste a few numbers to make a quick chart, the data source for more permanent charts should reside outside of Excel, and preferably be connected to a database query.
- Bring data into Excel as close as possible to its final format to avoid manipulating data inside Excel.
- Assume that formulas are a thread to data integrity, and avoid them whenever possible.
- Structure your workbook so that each sheet has a single purpose.

INDEX

C

D

E

H

I

J

K

L

P

Q

R

S

T