DATA AT WORK

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

JORGE CAMÕES

FREE SAMPLE CHAPTER
SHARE WITH OTHERS
DATA AT WORK

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

xiv

## 1 The Building Blocks of Data Visualization

1

- Spatial Organization of Stimuli ...................................................... 4
- Seeing Abstract Concepts ............................................................. 6
  - Charts .................................................................................. 7
  - Networks .............................................................................. 9
  - Maps ................................................................................. 10
  - Volume: Figurative Visualizations ................................................. 11
  - Visualization in Excel ............................................................... 12
- Retinal Variables ....................................................................... 12
- From Concepts to Charts ............................................................. 16
- The Proto-Chart ........................................................................ 17
- Chart Effectiveness .................................................................... 18
- Takeaways ............................................................................... 23

## 2 Visual Perception

24

- Perception and Cognition ............................................................ 25
  - Cognitive Offloading ................................................................ 26
  - A False Dichotomy .................................................................. 27
  - Charts and Tables ................................................................... 27
- Eye Physiology .......................................................................... 29
  - The Retina ............................................................................ 29
  - Cones ................................................................................. 30
  - The Arc of Visual Acuity ............................................................. 31
  - Saccades ............................................................................. 32
  - Impact of Eye Physiology on Visualization ............................... 34
- Pre-Attentive Processing ............................................................. 36
  - Salience .............................................................................. 36
  - Impact of Pre-Attentive Processing and Salience on Visualization .............................................................................. 37
- Working Memory ...................................................................... 40
  - Impact of Working Memory on Visualization .......................... 41
- Gestalt Laws ............................................................................. 43
  - Law of Proximity .................................................................... 47
  - Law of Similarity .................................................................... 47
Law of Segregation ................................................................. 48
Law of Connectivity .............................................................. 48
Law of Common Fate ............................................................. 49
Law of Closure ................................................................ 50
Law of Figure/Ground .......................................................... 50
Law of Continuity ................................................................ 51
Impact of Gestalt Laws on Visualization ................................. 52
The Limits of Perception .......................................................... 53
Why We Need Grid Lines and Reference Lines: Weber’s Law ..... 55
Being Aware of Distortions: Stevens’ Power Law ..................... 56
Context and Optical Illusions .................................................. 58
Impact of the Limits of Perception on Visualization .................. 59
Takeaways ........................................................................ 60

3 Beyond Visual Perception ......................................................... 62
Social Prägnanz .................................................................... 63
Breaking the Rules ................................................................ 64
The Tragedy of the Commons .................................................. 65
Color Symbolism ................................................................ 68
Representing Time .................................................................. 69
Axis Folding ........................................................................ 69
Don’t Make Me Think! ............................................................ 70
Literacy and Experience .......................................................... 71
Graphic Literacy .................................................................... 71
Familiarity with the Subject ..................................................... 74
Information Asymmetry .......................................................... 75
Organizational Contexts ........................................................... 75
Wrong Messages from the Top ................................................... 76
Impression Management ......................................................... 77
Takeaways ........................................................................ 78

4 Data Preparation ................................................................ 79
Problems with the Data ............................................................ 80
Structure without Content ....................................................... 80
Content without Structure ....................................................... 81
What Does “Well-Structured Data” Mean, Anyway? ................. 83
A Helping Hand: Pivot Tables ..................................................... 84
Extracting the Data ................................................................ 86
The PDF Plague .................................................................... 88
“Can It Export to Excel?” ........................................................ 89
Cleansing Data ........................................................... 90
Transforming Data ....................................................... 90
Loading the Data Table ............................................... 91
Data Management in Excel .......................................... 91
  Organizing the Workbook ........................................... 93
  Links Outside of Excel ............................................... 93
  Formulas ..................................................................... 93
  Cycles of Production and Analysis ............................... 94
Takeaways ................................................................. 95

5 Data Visualization ................................................... 96
  From Patterns to Points ............................................... 97
    Shape Visualization .................................................. 99
    Point Visualization .................................................. 103
    Outlier Visualization ............................................... 104
    Data Visualization Tasks .......................................... 106
  The Construction of Knowledge ................................... 106
    Data .......................................................................... 107
    Information .............................................................. 108
    Knowledge .............................................................. 108
    Wisdom ..................................................................... 109
  Defining Data Visualization ......................................... 110
  Languages, Stories, and Landscapes ......................... 111
  Graphical Literacy ..................................................... 112
  Graphical Landscapes ................................................. 113
    Profiling ................................................................. 113
    Dashboards ............................................................. 114
    Infographics ............................................................ 116
  A Crossroad of Knowledge .......................................... 120
    Statistics ................................................................. 120
    Design ....................................................................... 120
    Applications ............................................................ 120
    Content and Context ................................................ 121
  Data Visualization in Excel ........................................ 121
    The Good ................................................................. 122
    The Bad ..................................................................... 122
    The Ugly ................................................................. 124
    Beyond the Excel Chart Library ....................... 125
    Don’t Make Excel Charts ........................................... 128
Takeaways ................................................................. 131
6 Data Discovery, Analysis, and Communication

Where to Start? ................................................................. 133
   The Visual Information-Seeking Mantra .................................................. 134
   Focus plus Context ........................................................................ 137
Asking Questions ........................................................................ 138
   A Classification of Questions .......................................................... 139
Selecting and Collecting the Data .................................................. 140
Searching for Patterns ....................................................................... 142
Setting Priorities ........................................................................ 147
Reporting Results ................................................................. 148
   Clarification ........................................................................ 148
   The Human Dimension .............................................................. 149
   The Design .......................................................................... 150
Project: Monthly Births ............................................................... 151
   Defining the Problem ................................................................ 151
   Collecting the Data ................................................................ 152
   Assessing Data Availability ......................................................... 152
   Assessing Data Quality ............................................................ 154
   Adjusting the Data ................................................................ 154
   Exploring the Data ................................................................ 155
   Embracing Seasonality ............................................................. 156
   Communicating Our Findings ..................................................... 161
Takeaways .............................................................................. 162

7 How to Choose a Chart .................................................................. 163
Task-Based Chart Classification ................................................... 166
Audience Profile ........................................................................ 170
Sharing Visualizations ............................................................... 173
   Screens and Projectors ................................................................ 173
   Smartphones and Vertical Displays .............................................. 174
   PDF Files ........................................................................ 174
   Excel Files ...................................................................... 175
   Sharing Online ...................................................................... 175
Takeaways .............................................................................. 176
8 A Sense of Order

The Bar Chart ................................................................. 180
  Vertical and Horizontal Bars .................................................. 181
  Color Coding ..................................................................... 182
  Ordering ........................................................................... 182
  Chart Size ........................................................................ 185
  Breaks in the Scale ............................................................. 187
  Changing Metrics to Avoid Breaks in the Scale ......................... 188
  Evolution and Change .......................................................... 190
  A Special Bar Chart: The Population Pyramid ......................... 190

Dot Plots .......................................................................... 192
Slope Charts ....................................................................... 194
Strip Plots ........................................................................... 195
Speedometers ........................................................................ 196
Bullet Charts ....................................................................... 197
Alerts .................................................................................. 198
Takeaways ........................................................................... 199

9 Parts of a Whole: Composition Charts ................................. 200
What Is Composition? ............................................................ 202
  Composition or Comparison? .................................................. 202
  Pie Charts ........................................................................... 205
    Critique ........................................................................... 205
    Damage Control ................................................................. 206
  Donut Charts ....................................................................... 210
    Donuts as Multi-Level Pies ................................................. 212
  Actual Hierarchical Charts: Sunburst Charts and Treemaps ........ 213
  Stacked Bar Chart ............................................................... 217
  Pareto Chart ....................................................................... 218
Takeaways ........................................................................... 221

10 Scattered Data ................................................................ 222
The Data ............................................................................. 225
Distribution .......................................................................... 227
  Showing Everything: Transparencies and Jittering ..................... 227
  Quantifying Impressions ....................................................... 228
  Mean and Standard Deviation ................................................ 229
  The Median and the Interquartile Range ..................................... 229
  Outliers ............................................................................. 230
  Box-and-Whisker Plots .......................................................... 232
    Z-Scores ......................................................................... 233
### CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Pareto Chart Revisited</td>
<td>235</td>
</tr>
<tr>
<td></td>
<td>Excel Maps</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>Histograms</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>Bin Number and Width</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>Histograms and Bar Charts</td>
<td>245</td>
</tr>
<tr>
<td></td>
<td>Cumulative Frequency Distribution</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>Takeaways</td>
<td>248</td>
</tr>
<tr>
<td>11</td>
<td>Change Over Time</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>Focus on the Flow: The Line Chart</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Scales and Aspect Ratios</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td>Focus on the Relationships: Connected Scatter Plots</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>Sudden Changes: The Step Chart</td>
<td>259</td>
</tr>
<tr>
<td></td>
<td>Seasonality: The Cycle Plot</td>
<td>261</td>
</tr>
<tr>
<td></td>
<td>Sparklines</td>
<td>263</td>
</tr>
<tr>
<td></td>
<td>Animation</td>
<td>266</td>
</tr>
<tr>
<td></td>
<td>Takeaways</td>
<td>270</td>
</tr>
<tr>
<td>12</td>
<td>Relationships</td>
<td>271</td>
</tr>
<tr>
<td></td>
<td>Understanding Relationships</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td>Curve Fitting</td>
<td>274</td>
</tr>
<tr>
<td></td>
<td>The Scatter Plot</td>
<td>276</td>
</tr>
<tr>
<td></td>
<td>Scatter Plot Design</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Clusters and Groupings</td>
<td>281</td>
</tr>
<tr>
<td></td>
<td>Multiple Series and Subsets</td>
<td>282</td>
</tr>
<tr>
<td></td>
<td>Profiles</td>
<td>284</td>
</tr>
<tr>
<td></td>
<td>Bubble Charts</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>Takeaways</td>
<td>291</td>
</tr>
<tr>
<td>13</td>
<td>Profiling</td>
<td>292</td>
</tr>
<tr>
<td></td>
<td>The Need to Solve</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>Panel Charts</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>Bar Charts with Multiple Series</td>
<td>298</td>
</tr>
<tr>
<td></td>
<td>Horizon Chart</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>Reorderable Matrix</td>
<td>304</td>
</tr>
<tr>
<td></td>
<td>Small Multiples</td>
<td>307</td>
</tr>
<tr>
<td></td>
<td>Profiling in Excel</td>
<td>310</td>
</tr>
<tr>
<td></td>
<td>Takeaways</td>
<td>311</td>
</tr>
</tbody>
</table>
14 Designing for Effectiveness .............................................. 312

The Aesthetic Dimension ........................................................... 315
A Wrong Model ........................................................................ 316
The Design Continuum .............................................................. 318

Tools Are Not Neutral: Defaults .............................................. 320

Reason and Emotion ................................................................ 321
A.I.D.A. ................................................................................ 321
Does Reason Follow Emotion? .................................................. 326
Emotion and Effectiveness ........................................................ 328
Occam’s Razor ...................................................................... 329

Designing Chart Components ................................................... 332
Pseudo-3D ............................................................................. 333
Textures .............................................................................. 337
Titles ................................................................................ 338
Fonts ................................................................................ 339
Annotations ........................................................................ 339
Grid Lines ........................................................................... 342
Clip Art .............................................................................. 343
The Secondary Axis ................................................................ 344
Legends ............................................................................. 346
Backgrounds ....................................................................... 347

Ordering the Data ..................................................................... 347
Number of Series ...................................................................... 351
Chart Type ........................................................................... 352
Grouping ............................................................................ 352
Residual Category .................................................................. 353
Context ................................................................................ 353
Small Multiples ..................................................................... 354

Lying and Deceiving with Charts ............................................... 355
Data, Perception, and Cognition .................................................. 356
Exaggerating Differences ............................................................ 356
Distorting Time Series ............................................................... 357
Aspect Ratio ......................................................................... 357
Omitting Points ....................................................................... 358
Mistaking Variation for Evolution ................................................ 358
Double Axes ......................................................................... 359
Pseudo 3D ............................................................................ 359
Context .............................................................................. 360
When Everything Goes ............................................................ 362

Takeaways ............................................................................... 364
# 15 Color: Beyond Aesthetics

Quantifying Color ........................................................................ 367
  The RGB Model ..................................................................... 368
  The HSL Model ..................................................................... 368
Stimuli Intensity ........................................................................ 370
The Functional Tasks of Color ..................................................... 372
  Categorize ........................................................................... 373
  Group ............................................................................... 376
  Emphasize .......................................................................... 378
  Sequence ............................................................................ 378
  Diverge .............................................................................. 382
  Alert ................................................................................. 386
Color Symbolism ........................................................................ 386
The Role of Gray ...................................................................... 387
Color Staging .......................................................................... 389
Color Harmony ........................................................................ 392
  General Principles ................................................................. 392
  The Classical Rules ................................................................. 393
  Complementary Colors ............................................................ 394
  Split Complementary Colors ...................................................... 394
  Triadic Harmony ................................................................... 396
  Analogous Colors ................................................................... 396
  Rectangle ........................................................................... 396
  Warm Colors and Cool Colors ..................................................... 398
Sources for Color Palettes .......................................................... 399
  Excel ................................................................................ 399
  Beyond Excel ....................................................................... 402
  Color Blindness .................................................................... 403
Takeaways ............................................................................. 405

# 16 Conclusion

It’s All About Pragmatism, Not Aesthetics ........................................ 407
Say Goodbye to the Old Ways ..................................................... 407
Find Your Own Data Visualization Model ....................................... 408
In Business Visualization, Hard Work Is Not Always the Best Work .......... 408
Organizational Literacy ........................................................... 409
Reason and Emotion ................................................................ 409
Play with Constraints ................................................................ 410
The Tools .............................................................................. 411

Index ................................................................................... 412
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 study¹ 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.

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.

### Content without Structure

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

**Figure 4.1** A table with names and addresses.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Surname</th>
<th>Address</th>
<th>City</th>
<th>Zip Code</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>John</td>
<td>Doe</td>
<td>S Main St</td>
<td>Torrington</td>
<td>06790</td>
<td>Connecticut</td>
</tr>
<tr>
<td>1001</td>
<td>Mary</td>
<td>Poppins</td>
<td>SW 11th St</td>
<td>Lowton</td>
<td>73501</td>
<td>Oklahoma</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
<th>Marital Status</th>
<th>Children</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Mary T. Poppins</td>
<td>Female</td>
<td>34</td>
<td>5.38</td>
<td>182</td>
<td>Married</td>
<td>4</td>
<td>Librarian</td>
</tr>
<tr>
<td>1000</td>
<td>John F. Doe</td>
<td>Male</td>
<td>82</td>
<td>6.17</td>
<td>138</td>
<td>Widower</td>
<td>2</td>
<td>Retired</td>
</tr>
</tbody>
</table>

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.

---

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.

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.

---

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 article³ 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.

---

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.

![Figure 4.4](image)

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 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Item</th>
<th>Quintile</th>
<th>Year</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Lowest 20</td>
<td>2012</td>
<td>39</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Lowest 20</td>
<td>2013</td>
<td>40</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Second 20</td>
<td>2012</td>
<td>47</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Second 20</td>
<td>2013</td>
<td>52</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Third 20</td>
<td>2012</td>
<td>49</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Third 20</td>
<td>2013</td>
<td>56</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Fourth 20</td>
<td>2012</td>
<td>59</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Fourth 20</td>
<td>2013</td>
<td>59</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Highest 20</td>
<td>2012</td>
<td>71</td>
</tr>
<tr>
<td>Food Total</td>
<td>Eggs</td>
<td>Highest 20</td>
<td>2013</td>
<td>76</td>
</tr>
</tbody>
</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.
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

<table>
<thead>
<tr>
<th>Series ID</th>
<th>Year</th>
<th>Period</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M01</td>
<td>11.7</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M02</td>
<td>11.6</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M03</td>
<td>11.3</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M04</td>
<td>10.8</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M05</td>
<td>10.4</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M06</td>
<td>10.1</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M07</td>
<td>10.0</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M08</td>
<td>9.9</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M09</td>
<td>10.0</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M10</td>
<td>10.1</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M11</td>
<td>10.2</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M12</td>
<td>10.3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Series ID</th>
<th>Year</th>
<th>Period</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M01</td>
<td>11.7</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M02</td>
<td>11.6</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M03</td>
<td>11.3</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M04</td>
<td>10.8</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M05</td>
<td>10.4</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M06</td>
<td>10.1</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M07</td>
<td>10.0</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M08</td>
<td>9.9</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M09</td>
<td>10.0</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M10</td>
<td>10.1</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M11</td>
<td>10.2</td>
</tr>
<tr>
<td>LASST010000000000003</td>
<td>2010</td>
<td>M12</td>
<td>10.3</td>
</tr>
</tbody>
</table>

**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.
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.

“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.

- **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.
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 worldwide⁴ 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

---

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. 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.

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

2D plane, 9, 11, 23
3D charts. See also pseudo-3D effects
  grid lines in, 334
  maps, 238–240
  pie charts, 19, 21, 65, 77, 207, 337
3D effect, 22
3D Maps, 238–240
3D pie charts, 19, 21, 65, 77, 207, 337
3D spaces, 11

A
abstract concepts, 6–12
Adobe Color CC color palette, 403
aesthetics, 315–319
  considerations, 313, 314
  described, 23
  design continuum, 318–319
  evaluation criteria, 316
  importance of, 313, 314
  overvaluing, 408
  vs. pragmatism, 407
aggregation, 91
A.I.D.A. (Attention, Interest, Desire, Action), 321–326
alerts, 198–199, 386
analogous colors, 396, 397
anchor points
  considerations, 75, 232, 248
  importance of, 136
  pie charts and, 203, 204
animations, 266–269
  charts, 103
  Keynote, 266
  pattern detection with, 267
  PowerPoint, 266
  vs. small multiples, 311
  time periods, 266–269
annotations, 167, 339–341
Anscombe, Francis, 224
Anscombe’s quartet, 223–224
applications, 121
arc of visual acuity, 31–32, 60
area, 6
art, clip, 22, 343
aspect ratios, 173, 254–256, 357–358
attention, 27, 33, 65–66
Attention, Interest, Desire, Action (A.I.D.A.), 321–326
audience
  attracting, 65–66
  considerations, 78
  literacy, 176
  messages, 71, 313
  “tragedy of the commons” and, 65–67
audience profile, 170–172
availability, 152–153
axes
  double, 359
  plotting data along, 16–17
  secondary, 344–346
  slope charts, 194
axis folding, 69–70

B
backgrounds, 22, 326, 347
bamboo chart, 72–73, 167
bandlines, 266
bar charts, 180–192. See also bars
  bad defaults for, 321
  breaks in scale, 187–189, 199
  chart size, 185–187
  color coding, 182, 183
  combining with strip plot, 72–73
  compression, 185
  described, 18
  evolution/change, 190
  grouped, 298–299
  vs. histograms, 245, 248
horizontal vs. vertical, 181–182
labels, 181
multiple series, 181–182, 298–299, 310
ordering values, 182–185
overview, 180
population pyramids, 190–192, 267–269, 352
stacked, 18, 202, 203, 217–218
bar height, 56
bars
comparing, 26, 56, 254, 298, 334
considerations, 352
distortions and, 56–58
vs. dots, 192
error, 127
in histograms, 245, 248
horizontal, 180, 181–182
length of, 180
non-aligned, 56, 203
omitting points, 258
pseudo 3D and, 359
Stevens’ power law, 56–58
vertical, 180, 181–182, 219
Weber’s law, 56
Bertin, Jacques, 13, 111, 142, 304–306
Better Life Index, 285
Beveridge curve, 170–171
BI (business intelligence) systems, 89–90
bin number/width, 238, 241, 242–244
bins, 241, 242–244
births, monthly (project), 151–162
black-and-white charts, 387–389
blind spot, 30, 32
“box,” 232
box-and-whisker plots, 232–234, 241
bubble charts
considerations, 286, 291
distortion and, 56, 57–58
example, 286–290
lollipop charts, 127, 192–193
overview, 286–287
relationships, 286–290
bullet charts, 172, 197–198, 199
business intelligence (BI) systems, 89–90
business visualization, 112, 131
C
Cairo, Alberto, 116–117, 313
categories
charts, 164
color, 367, 373–375, 376
grouping, 48, 298–299
grouping/ordering data, 349, 350
residual, 353
categorizing questions, 138–140
chart types
3D. See 3D charts
bamboo charts, 72–73, 167
bar. See bar charts
bubble. See bubble charts
bullet charts, 172, 197–198, 199
combo charts, 166
composition. See composition charts
considerations, 22
data reduction charts, 167–168, 169, 351
described, 23, 163
donut charts, 18, 210–213
fan charts, 208–209
Gapminder, 287
helium charts, 290, 291
hierarchical charts, 212, 213–216
horizon charts, 70, 299–303
line. See line charts
lollipop charts, 127, 192–193
maps. See maps
overview, 352
panel charts, 217, 295–297
Pareto charts, 218–220, 221, 235–237
pie charts. See pie charts
point comparison, 167–168, 351
proportion charts, 166, 212–213, 218, 221
proto-charts, 17–18, 21, 23
slope charts, 194–195
step charts, 259–261
strip plots, 18, 72, 73, 195–196, 232
sunburst charts, 213, 216
task-based classification, 166–169, 176
transformations, 17–18
charts
aesthetics of, 23
animating, 103
answering questions with, 138–140, 147
aspect ratio, 357–358
audience profile, 170–172
“bad,” 22
black-and-white, 387–389
categories, 164
choosing, 19–21, 163–176
combo, 166
complex, 131
components, 332–347
concepts, 16–17
considerations, 6, 7, 21, 22
consistency, 22
defined, 7
dynamic, 85, 86
effectiveness of. See effectiveness
evaluation criteria, 316–317
exaggerating differences, 356
false dichotomy, 27–28
“flat,” 147–148
fonts. See fonts
“Graphenstein,” 19
vs. graphs, 7
grayscale, 387–389
“high-impact,” 21
“hooks,” 150
legends. See legends
line, 8, 18, 20
low-density, 294
lying/deceiving with, 355–363
memorable, 66–67
multiple series. See multiple series
charts
number of series in, 351–355
overview, 7–9
profile. See profiling
reducing use of, 94
reusing/recycling, 94
seasonality and. See seasonality
simplification, 44, 52, 329–332
size, 34
“spaghetti,” 38–39, 138, 353
subjectivity in, 15
vs. tables, 27–28, 76
titles. See titles
transformations, 21
classes. See bins
classifications, 166–169
cleansing data, 90
Cleveland study, 53–56, 58, 179
Cleveland, William, 53–56, 60, 254, 357
clinical trials, 27
clip art, 22, 343
closure, law of, 50
clusters, 274, 281–282
Coase, Ronald, 142
cognition, 25–28, 356
cognitive offloading, 26
cognitive style, 121
color, 365–405
aesthetic quality of, 366
alerts, 386
analogous, 396, 397
bar charts, 182, 183
categorizing by, 367, 373–375, 376
classical rules, 393–394
complementary, 394, 395
considerations, 22, 42, 367
consistency, 22
cool, 396
diverging scales, 382–385
emphasizing items via, 37, 38, 378
fill, 21
functional qualities of, 366
functional tasks of, 372–386
grouping data by, 376–377
horizon charts, 299–303
HSL, 368–370
vs. hue, 369
for identification, 376–377, 388
message tone, 392
overview, 365–366
pie charts, 48
preferences, 367
pure, 371
quantifying, 367–370
rectangle rule, 396, 397
removing, 387–389
RGB, 368
role of gray, 387–389
sequences, 378–382
split complementary, 394, 395
standard pale/tte, 392
stimuli intensity, 366, 370–372
suitability to task, 366
triadic, 396
verbalizing, 367
warm, 396
color blindness, 388, 403–404
color codes, 42
color coding, 182, 183, 198, 374–375
color differentiation, 22
color harmony, 371, 392–399
color palettes, 399–404
Adobe Color CC, 403
ColorBrewer, 402
considerations, 371
Excel, 131, 399–401
LCD-friendly, 173
color ramps, 378–382
color staging, 389–391
color symbolism, 68, 366, 386–387, 392
color wheels, 393–394, 396, 398
ColorBrewer color palette, 402
combo charts, 166
common fate, law of, 49
commons, tragedy of, 65–67
communications, 63, 89
comparisons
absolute vs. relative, 360–362
bars, 26, 56, 254, 298, 334
bubble charts, 286
categories, 349
charts/tables, 27
vs. composition, 202–204
considerations, 53, 134, 178–179
data points, 167–168, 199, 351
dates, 195
distributions, 233–234
donut charts, 210
fan charts, 208
overview, 177–179
Pareto charts, 235–237
perception and, 27
sparklines, 264
structure without content, 81
complementary colors, 394, 395
composition, 139, 140, 201, 202–204, 221
composition charts, 200–221
composition vs. comparison, 202–204
considerations, 201–202, 221
donut charts, 18, 210–213
overview, 200–202
Pareto charts, 218–220, 221, 235–237
pie charts. See pie charts
stacked bar charts, 18, 217–218
sunburst charts, 213, 216	
treemaps, 21, 213, 214–215, 216
compression, 185
concepts
abstract, 6–12
charts, 16–17
cones, 30–31
connectivity, law of, 48–49, 52
constellation naming, 44
constraints, 115, 362, 410
Consumer Expenditure Survey, 84
content
considerations, 74, 80, 121
content without structure, 81–83
importance, 36–40, 60
structure without content, 80–81
context, 353–354
considerations, 121
focus and, 137–138
graphic lies, 360–362
optical illusions and, 58–59
organizational, 75–78
overview, 353–354
small multiples, 354–355
context entity, 137
continuity, law of, 51–52
cool colors, 396
coordinate pairs, 7, 16, 23
coordinate system, 4–5
coordinates, 4, 7, 8, 9, 13, 16, 23
Cotgreave, Andy, 192
covariation, 272–273, 276
Crystal Xcelsius, 125
cumulative effects, 219, 248
cumulative frequency distribution, 246–247
cumulative values, 219, 221, 235–237
curve fitting, 274–275
cycle plots, 261–263
cyclic patterns, 155, 156

D
D3 language, 411
dashboards
car, 196–197
Excel, 115–116
executive, 196
graphical landscapes, 114–116
SAP BusinessObjects Dashboards, 125
speedometers, 196–197
Stephen Few on, 114
data. See also information
adjusting, 154
aggregating, 91
analyzing, 142–147, 155–156
availability, 152–153
categories. See categories
cherry-picking, 358
cleansing, 90
cognition and, 356
collecting, 140–141, 152
communicating findings, 161–162
comparing. See comparisons
complexity of, 132–133
considerations, 225
corrections to, 330
derived, 91
dirty, 90
discovery, 132–140
diagnostic judgment, 147–148
emotion and, 326
encoding, 91
extracting, 86–90
grouping. See grouping data
hiding, 353
inconsistent, 154
interrogating, 138–140
missing, 50, 141, 154
non-alert levels, 197
ordering, 347–351
overview, 107
perception and, 356
primary, 140–141
problems with, 80–83
qualitative, 14
quality, 154, 160
quantitative, 14
relevancy, 147–148
removing, 91
reporting results, 148–150
scattered. See scattered data
seasonality. See seasonality
secondary, 140–141
selecting, 140–141
standardized, 91
transformations, 17–18
transforming, 90–91
variation vs. evolution, 358–359
well-structured, 83–86
data discovery, 132–140
data integration, 131
data integrity, 83, 95, 347
data journalism, 111
data management, Excel, 91–94
data points, 177–199
anchor. See anchor points
cloud of, 8
comparing, 167–168, 199, 351
connections between, 9, 59
considerations, 6, 7, 352
differences among, 9, 59
in networks, 9
omitting, 358
overlapping, 228, 248
overview, 107
patterns and, 311
plotting. See plots/plotting
profiles and. See profiling
reading, 97–106
working with, 103–104
data preparation, 79–95
data preprocessing system, 28
data reduction charts, 167–168, 169, 351
data sensing, 3
data stories, 111–112
data tables, 16, 91
data types, 14, 15
data validation rules, 80
data visualization, 96–131
  aesthetic dimension, 315–319
  asking questions, 138–140, 147
  building blocks of, 1–23
  business visualizations, 112, 131
  comparisons. See comparisons considerations, 2, 407
  construction of knowledge, 106–109
  defining, 97, 110–111
  effectiveness of. See effectiveness evaluation criteria, 316–317
  in Excel, 12, 122–130
  familiarity with subject, 74–75
  figurative, 6, 11–12
  graphical landscapes, 111, 112, 113–119, 185
  graphical literacy. See graphical literacy
  impact of eye physiology on, 34–35
  impact of working memory on, 41–42, 60
  interactive, 173, 175
  knowledge/skills, 120–121
  languages, 111–112
  limits of, 59
  lying/deceiving with, 355–363
  outliers. See outliers overview, 1–2, 96–97
  patterns, 97–106
  points. See data points
  reason vs. emotion, 409–410
  shapes, 97, 98, 99–103
  sharing, 173–175
  stories, 111–112
tasks, 106
tools for, 411
data visualization model, 408
data-driven annotations, 167
dates. See also time entries
  comparing, 195
  considerations, 22
  meaningful, 41
  representing, 69
decoration stage, 318, 319
demographic indicators, 141
demographics, 141, 151, 208, 338
dependency ratios, 144–147
depth, 23
derived data, 91
design
  annotations, 339
  backgrounds, 347
  chart components, 332–347
  clip art, 343
  for effectiveness, 312–314
  exploration stages, 150
  fonts, 339
  grid lines, 342–343
  inconsistencies, 333
  legends, 346
  number of series, 351–355
  Occam’s razor, 329–332
  ordering data, 347–351
  pseudo-3D elements, 333–336
  reason/emotion, 321–332
  removing clutter, 330
  scatter plots, 279–281
  secondary axis, 344–346
textures, 337
titles, 338–339
design continuum, 318–319
design skills, 120
diagrams, 10
dichotomy, false, 27
differences, exaggerating, 356
DIKW Pyramid, 107–109
dimensions, 6, 393
display alerts, 198–199
distances, 8
distortion, 56–58, 60, 318
distributions, 227–231
  comparing, 233–234, 248
  considerations, 227
  described, 227
  jittering, 228
  properties, 223
  questions about, 139, 140
  scattered data, 227–231
  shapes, 223
  studying, 227
  transparencies, 227–228
  z-scores, 233–234
diverging scales, 382–385
The Dollar Street project, 149
donut charts, 18, 210–213
dot plots, 192–193
double axes, 359
dual-axis charts, 344–346, 359
dynamic charts, 85, 86
equivalence, 346
error bars, 127, 192
Escher bonus, 334
Escher, M.C., 333
ETL (Extract, Transform, and Load)
  process, 80, 86–91
evolution
  vs. change, 190
  line charts, 251
  questions about, 139, 140
  scatter plots, 256
  step charts, 259
  vs. variation, 358–359
Excel
  advantages, 122–123
  alerts, 198
  color palettes, 131, 399–401
  considerations, 122, 411
  data extraction, 86–87
  data management in, 91–94
  data structure, 92
  data visualization in, 122–130
  defaults, 320–321
  disadvantages, 123–125
  exporting to, 89–90
  formulas, 93–94
  links outside of, 80, 93–94, 95
  networks in, 12
  profiling in, 124, 310
  visualization in, 12
  workbooks, 80, 93–94, 95
Excel 2003, 320
Excel 2007, 122
Excel 2010, 122
Excel 2016, 122–125
Excel chart library, 7, 123–127, 164–165, 216
Excel charts
  considerations, 7, 92, 294, 411
  default, 92
  using, 128–130
Excel dashboards, 115–116
Excel files, 80, 110, 175, 407
Excel histograms, 245
Excel library, 295
Excel maps, 12, 238–240, 310
Excel online, 175
Excel visualizations, 411

E
editorial dimension, 39, 113, 212–213, 355
editorial judgment, 147–148
effectiveness
  charts, 18–23, 408
  considerations, 164, 313
  data visualization, 18–23, 408
  designing for, 312–314
  emotion and, 328–329
  information producer/consumer, 326
  scope, 312–313
Einstein, Albert, 206
electrocardiogram, 74
elements. See also entities; objects
  adding, 330
  categorizing by color, 367, 373–375, 378
  emphasizing by color, 378
  missing, 51–52
  pseudo-3D, 333–336
emotion, 149, 150, 321–332, 409–410
encoding data, 91
encoding stage, 318, 319
entities, 137, 304, 307. See also elements; objects
exceptions, 139
exploded slices, 21, 22, 207
exporting to Excel, 89–90
Extract, Transform, and Load (ETL) process, 80, 86–91
extracting data, 86–90
eye movements, 32–34
eye physiology, 29–35
eye–brain system, 2, 3, 25, 44, 61

F
Fairfield, Hannah, 174
false dichotomy, 27
fan charts, 208–309
Few, Stephen
  bandlines, 266
  on dashboards, 114
  on data visualization, 110, 407
  on Excel 2007, 122
  on pie charts, 204, 205
  simplicity and, 23
figurative visualizations, 6, 11–12
figure/ground, law of, 50–51
files
  Excel, 80, 110, 175, 407
  PDF, 88–89, 174–175
  sharing, 173–175
fill color, 21
focus, 137–138
focus entity, 137
focus-plus-context approach, 137–138
fonts, 22, 339
formatting
  conditional, 143, 198
  considerations, 160, 178
  titles, 339
forms
  law of closure and, 50
  missing data, 50
  simplification of, 44, 52
formulas, 85, 92, 93–94
frames, 22, 55
frameworks, 204
Freedman–Diaconis’s rule, 242
functional stage, 318, 319

G
Gapminder chart, 287
Gapminder Foundation, 149
garbage in, garbage out (GIGO), 83
geometric primitives, 6, 16, 17, 23
geometric shapes, 6
Gestalt laws, 43–53, 60
GIGO (garbage in, garbage out), 83
“gore” fonts, 339
graph theory, 7
“Graphenstein” charts, 19
graphicacy, 112–113
graphical illiteracy, 77, 321
graphical landscapes, 111, 112, 113–119, 185
graphical literacy
  considerations, 112, 295
  described, 112
  emotional components, 321
  “epiphanies,” 112–113
  low, 77, 314, 321, 409
  overview, 71–74
graphical tables, 384–385
graphics
  clip art, 22, 343
  infographics, 313
  simplification, 44, 52
  sparklines, 263–266
graphs, 7, 8, 315, 316
  gray, 387–389
  grayscale charts, 387–389
  grid lines, 55–56
  in 3D charts, 334
  overview, 342–343
  grouping data
    by category, 349–350
    by color, 376–377
    considerations, 352
    scatter plots, 281–282
  grouping items. See also ordering items
    bar charts, 298–299
    by category, 298–299
    Gestalt laws, 43–53, 60
    “meaningful groups,” 282–283
    by theme, 147
  groupings, 281–282
H 
hearing, 3
Heer, Jeffrey, 300
helium charts, 290, 291
hierarchical charts, 212, 213–216
“high-impact charts,” 21
histograms, 240–245, 246, 248
horizon charts, 70, 299–303
HSL color model, 368–370
hues, 369, 372, 373
human perspective, 149

I 
ilusions, optical, 58–59
illustrations, 12
images. See also graphics
clip art, 22, 343
infographics, 313
simplification, 44, 52
impressions
management of, 77–78
quantifying, 228–229
validating, 228–229
inconsistencies, 333
infographics
vs. business charts, 94
considerations, 116–117
graphical landscapes, 111, 112, 113–119, 185
increasing audience with, 313
Napoleon’s troops, 117–119
information. See also data
asymmetry, 75, 170
displaying most important, 114–115
loss of, 222
new, 17
noise, 222
overview, 108
sharing, 173–175, 328
useful vs. useless, 44, 222
working memory and, 40–42, 60
information units, 41
interactive visualizations, 173, 175
interface design, 70, 71
interquartile range, 229–230
IT roles, 92

J 
journalism, data, 107, 111
journalists, 75, 314
JPEG format, 175

K 
Kanizsa’s Triangle, 58
key performance indicators (KPIs), 197–198, 199
Keynote animations, 266
knowledge, 106–109, 120–121
Kosslyn, Stephen, 30, 31
KPIs (key performance indicators), 197–198, 199
Krug, Steve, 70, 71

L 
labels/labeling, 41, 60, 181
Laffer curve, 273
landscapes
drawing with coordinate system, 4–5
graphical, 111, 112, 113–119, 185
languages, 111–112
law of closure, 50
law of common fate, 49
law of connectivity, 48–49, 52
law of continuity, 51–52
law of figure/ground, 50–51
law of proximity, 47, 52
law of segregation, 48, 52
law of similarity, 47–48
legends
borders, 346
c onsiderations, 22, 34, 35, 41
design, 346
elimination of, 42
overview, 346
pie charts and, 22, 207
replacing with labels, 60
unnecessary, 68
lies/lying, 355–363
line charts
  aspect ratios, 254–256
  described, 18
  example of, 8
  markers, 251, 270
  overview, 250–254
  scales, 254–256
  sparklines, 263–266
  time periods and, 250–256
  variables in, 256
linear scale, 248
lines
  bandlines, 266
  breaks in, 51–52
  considerations, 6, 326, 352
  grid lines. See grid lines
  in networks, 9
  reference, 55–56, 57, 279, 291
literacy, 71–74, 409. See also graphical literacy
loading data tables, 91
log scale, 240, 247
lollipop charts, 127, 192–193
low-density charts, 294
luminance, 369, 372, 393
Lumira, 411
M
Mackinlay, Jock D., 15
macula lutea, 29
Magritte, René, 2
makeup stage, 318, 319
Malofiej Awards, 117
management, wrong messages from, 74–75
maps
  3D Maps, 238–240
  considerations, 225
  in Excel, 12, 238–240, 310
  overview, 6, 10–11
  treemaps, 21, 213, 214–215, 216
matrix, reorderable, 304–306
McGill, Robert, 53
mean, 229
median, 229–230
memorization, 40–41
memory
  cognitive offloading, 26
  maximum number of objects stored, 351–355
  working, 40–42, 60
message tone, 392
metrics, 188–189
Microsoft Power BI, 122, 124, 126, 310
Minard, Charles Joseph, 117–119
mnemonics, 41
Monthly Births project, 151–162
movies. See animations
multiple series charts
  bar charts, 181–182, 298–299, 310
donut charts, 210
Excel maps, 310
patterns and, 58
scatter plots, 282–283
N
Napoleon’s troops infographic, 117–119
narratives, 111–112
National Snow and Ice Data Center (NSIDC), 295–297
network diagram, 10
networks, 6, 9–10, 12
New York Times, 170, 174
noise, 44, 248
nominal data type, 15
nominal variables, 14, 15, 181, 351
NSIDC (National Snow and Ice Data Center), 295–297
O
objects. See also elements; entities
  common fate of, 49
  connections, 48–49, 52
  features of, 60
  figure vs. ground, 50–51
  grouping, 44, 45, 48, 52–53
  maximum number stored, 351–355
  pre-attentive processing, 36–40, 60
  prominence of, 36–40, 60
proximity, 47, 52
roles of, 390
segregated, 48, 52
similarity, 47–48
simplification, 44, 52, 329–332
types of, 338
Occam’s razor, 329–332
OECD (Organization for Economic Cooperation and Development), 285
Office palettes, 399
online sharing, 175
optical illusions, 58–59
ordering items
alphabetically, 182–185
keys, 199, 347–349, 351
questions about, 139, 140
values, 182–185
ordinal data type, 15
ordinal variables, 14, 15, 372
Organization for Economic Cooperation and Development (OECD), 285
organizational contexts, 75–78
organizational literacy, 409
Ortega y Gasset, José, 63
outlier detection chart, 167
outlier visualization, 97, 98, 104–105
outliers, 230–231
   considerations, 228, 274
defining, 231
dot plots and, 192, 193
   identifying, 231
   lollipop charts and, 193
overview, 230
reading, 97–98
working with, 104–105
overstimulation, 78, 192, 394

P
panel charts, 217, 295–297
Pareto charts, 218–220, 221, 235–237
Pareto principle, 219
parsimony, 329
patterns
   considerations, 58–59
data points and, 311
hiding, 311
searching for, 142–147
time, 267
visualizing data, 97–106
PDF files, 88–89, 174–175
Penrose triangle, 333
perception, 25–28. See also visual perception
charts vs. tables, 27–28
cognition and, 25–28, 356
comparisons and, 27
considerations, 53, 62
data and, 356
false dichotomy, 27
limits of, 53–59
overview, 25
perspective, 346
photoreceptor cells, 29–30, 367, 368
pie charts, 205–209
   3D, 19, 21, 77, 207, 337
   anchor points, 203, 204
   color, 48
   considerations, 21, 201, 202, 205–206
   correcting/improving, 206–207
   critiques, 19–22, 23, 204, 205–206
   described, 18, 205
donut charts, 18, 210–213
   examples, 19–21, 22, 66, 201–206
   exploded slices, 21, 22, 207
   fan charts, 208–309
   legends and, 22, 207
   multi-level, 212–213
   number of slices, 22
   popularity of, 201
   pseudo-3D and, 21, 22, 336, 337
   reading, 202, 203
   sectograms, 201
   slices, 22, 205–207, 212
   sunburst charts, 213, 216
   transformations, 21
   Tufted on, 205
pivot tables, 84–86, 95
plane, 6
Planisphaerium celeste, 43
Playfair, William, 128–129, 253
plots/plotting
  along axes, 16–17
  box-and-whisker plots, 232–234, 241
cycle plots, 261–263
dot plots, 192–193
overlapping points, 228, 248
scatter plots. See scatter plots
strip plots, 18, 72, 73, 195–196, 232
PNG format, 175
point comparison charts, 167–168, 351
point visualization, 97, 98, 103–104
points. See data points
political issues, 89
population density charts, 184–185,
  192–193
population projections, 144–146
population pyramids, 190–192, 267–269,
  352
population statistics, 133–136, 144–147, 267
Power BI, 123, 124, 126, 310, 411
PowerPoint presentations, 77, 266, 320
pragmatism vs. aesthetics, 407
prägnanz, 44, 63–64
pre-attentive processing, 36–40, 60
presentations, 42, 115, 267, 407
primary data, 140–141
primus inter pares, 38
principle of least effort, 24
priorities, setting, 147–148
profiling, 292–311
  bar charts/multiple series, 298–299,
    310
described, 294
Excel, 124, 310
graphical landscapes, 113–114
overview, 292–295
panel charts, 295–297
questions about, 139, 140
reorderable matrix, 304–306
scatter plots, 284–285
small multiples, 114, 307–310, 354–355
projectors, 173–174
proportion, 201
proportion charts, 166, 212–213, 218, 221
proto-charts, 17–18, 21, 23
proximity, law of, 47, 52
pseudo-3D effects, 333–336
  considerations, 22
data distortion and, 359
  grid lines and, 334
overview, 333–334
pie charts and, 21, 22, 336, 337
use of, 336
pyramids
  DIKW, 107–109
  population, 190–192, 267–269, 352
Python language, 411
Q
QlikView, 126, 411
qualitative variables, 14
quality, data, 154, 160
quantifying color, 367–370
quantifying impressions, 228–229
quantitative data, 14
quantitative data type, 15
quantitative variables, 14
quartiles, 229–230
questions, asking, 138–140, 147
R
R language, 411
ratios
  aspect, 173, 254–256, 357–358
  dependency, 144–147
  overview, 253–254
reason, 150, 321–322, 409–410
rectangles, color, 396, 397
redundancy, 346
reference lines, 55–56, 57, 279, 291
reference points, 311
reification, 58
relationships, 271–291
  analyzing, 273–275
  bubble charts, 286–290
  clusters/groupings, 281–282
  considerations, 113
correlation, 273, 275, 277–279, 291
direction, 273, 275
inverted-u shape, 273, 275
linear, 273, 275
overview, 271–273
positive vs. negative, 273, 277
questions about, 139, 140
scatter plots. See scatter plots
shape, 273, 275
strength, 273, 275
time periods and, 253, 256–259
visualization, 274, 275
reorderable matrix, 304–306
reporting results, 148–150
representations, 2
residual category, 353
retina, 29–30, 31
retinal variables, 12–15, 53–55
RGB color model, 368
Rice rule, 242
river, metaphorical, 249–250
Rosling, Hans, 49, 266
rules, 63, 64–71, 78, 332

S
saccade movements, 32–34
Sagan, Carl, 75
salience, 36–40, 60
SAP BusinessObjects Dashboards, 125
SAP Lumira, 411
Saramago, José, 109
saturation, 369, 393
scales
breaks in, 187–189, 199
common, 54, 264
log, 103, 240, 247
vertical, 185, 203
working with, 254–256
scatter plots, 276–285
clusters/groupings, 281–282
connected, 256–259
curve fitting, 274–275
design, 279–281
multiple series, 282–283
overview, 276–279
profiling, 284–285
subsets, 282–283
time periods and, 256–259
variables, 256–259
scattered data, 222–248
box-and-whisker plots, 232–234, 241
considerations, 225
cumulative frequency distribution,
246–247
curve fitting, 274–275
distribution, 227–231
Excel maps, 238–240
histograms, 240–245, 246, 248
overview, 222–225
Pareto charts, 235–237, 248
Schwartz, Barry, 164
screens, 173–174
seasonality, 156–160
considerations, 157, 297
cycle plots, 261–263
cyclic patterns, 155, 156
geography and, 157
time periods and, 261–263, 296–297
working with, 156–160
secondary axis, 344–346
secondary data, 140–141
sectograms, 201
segregation, law of, 48, 52
senses, 27
sequences, 378–382
shape visualization, 97, 98, 99–103
shapes
considerations, 99
reading, 97–98
working with, 99–103
sharing visualizations, 173–175
Shneiderman, Ben, 136
signal, 44
similarity, law of, 47–48
simplification, 44, 47–48
slices, pie, 22, 205–207, 212. See also pie
charts
slope, 254
slope charts, 194–195
small multiples, 114, 307–310, 354–355
smartphones, 174
smell, 3
social conventions, 64, 78
social networks, 355
social prégnanz, 63–64
social relationships, 63–65
software applications, 121
spaghetti chart, 38–39, 138, 353
sparklines, 185, 186, 263–266
speedometers, 196–197
Spence, Ian, 206
split complementary colors, 394, 395
spreadsheet errors, 93–94
spreadsheets, 93–94, 122, 142, 411
stacked bar charts, 18, 202, 203, 217–218
standard deviation, 229
standardization, 91
statistics
  considerations, 120
  human perspective, 149
  knowledge of, 120
step charts, 259–261
Stevens’ power law, 56–58
stimuli, 3–5
stimuli intensity, 366, 370–372
stories, 111–112
strip plots, 18, 72, 73, 195–196, 232
structure
  content without structure, 81–83
  Excel, 92
  structure without content, 80–81
subtitles, 338
sunburst charts, 213, 216
symbolism, color, 68, 366, 386–387, 392
symbols, 343

table values, 23
Tableau, 126, 175, 411
tables
  vs. charts, 27–28, 76
data, 16, 91
  false dichotomy, 27–28
  graphical, 384–385
  loading, 91
  pivot, 84–86, 95
  value differences, 3
task-based chart classification, 166–169,
  176
taste, 3
technical issues, 89–90
textures, 337
time patterns, 267
time periods, 249–270
  animations, 266–269
  considerations, 22, 249–250
  cycle plots and, 261–263
  direction, 22
  flow of, 250–256
  line charts and, 250–256
  relationships and, 253, 256–259
  representing, 69
  scatter plots and, 256–259
  seasonality, 261–263, 296–297
  small multiples, 267–269, 270
  sparklines, 263–266
  step charts and, 259–261
  sudden changes, 259–261
time series, 357
titles, 22, 236, 338–339
Tobler, Waldo, 47
tools, 25–26, 320–321, 411
touch, 3
“tragedy of the commons,” 65–67
transformations, 17–18, 21, 23
transforming data, 90–91
transparencies, 227–228
treemaps, 21, 213, 214–215, 216
Trendalyzer, 266
triadic harmony, 396
Triangulum constellation, 43
Tufte, Edward
  on pie charts, 205
  on PowerPoint presentations, 320
small multiples, 307–309
sparklines, 185, 186, 263–266
on statistics, 65
umbō, 29
units of measurement, 291
user interface, 70, 71
user roles, 92

values
  cumulative, 219, 221, 235–237
  ordering, 182–185
  table, 23
variables
  covariation between, 272–273, 276
  line charts, 256
  multiple, 283
  nominal, 14, 15, 181, 351
  ordinal, 14, 15, 372
  qualitative, 14
  quantitative, 14
  retinal, 12–15
  scatter plots, 256–259
variation vs. evolution, 358–359
vertical displays, 174
vision, 3
visual acuity, 31–32, 60
Visual Information-Seeking Mantra, 136
visual perception, 24–78. See also
  perception
  axis folding, 69–70
  breaking the rules, 64–71
  color symbolism, 68, 366, 386–387, 392
  considerations, 60
  familiarity with subject, 74–75
  impression management, 77–78
  information asymmetry, 75, 170
  organizational contexts, 75–78
  resources, 59–60
  social prägnanz, 63–64
  “tragedy of the commons,” 65–67
  wrong messages from management, 76–77
visual rhetoric, 15
visual stimuli, 3–5
visualization. See data visualization
Visualoop, 117
volume, 6, 11–12

Walmart growth chart, 305–309
Ware, Colin, 36, 59
warm colors, 396
Weber’s law, 55–56, 57
“whiskers,” 232
Wilkinson, Leland, 111
wisdom, 109
workbooks, 80, 93–94, 95
working memory, 40–42, 60

x coordinates, 7

y coordinates, 7
Yarbus, Alfred, 32, 33

z-scores, 233–234