Part II

Chapter One

Habits of a Good Data Analyst

By working with different data sets each week, Makeover Monday is a unique opportunity for participants to develop analytical skills that can be used in real life. Working with constantly changing topics, data sizes, and data complexity forces you to be considerate and deliberate about your analysis and helps you develop the skills to make you better at your job. This chapter covers the common characteristics of the best data analysis we have seen in Makeover Monday.

Approaching Unfamiliar Data

Over the course of 52 weeks, you will find 52 completely different data sets. It is almost impossible to have the knowledge to understand all 52 topics immediately. When faced with this situation, you need to have tools in your arsenal to help you overcome this deficit quickly.

While Makeover Monday is a great opportunity to practice data visualization and data analysis, it is also a chance to simulate real-life business scenarios. When specific requirements or guidelines are provided, considering those requirements for your makeover is a good idea. Think about it this way: if you are given specific requirements at work or specific questions that need to be answered as part of your analysis, will you ignore them? If you do, what will happen? I bet the outcome will not be in your favor. So why not treat Makeover Monday the same way and practice real-life business scenarios with a different data set every week?

In the corporate world, when you kick off any kind of dashboard project, you should engage with stakeholders, meet with the people who will be using it, the people who own and/or govern the data, and others to get the full picture before you start analyzing and visualizing the data.

When it comes to producing something of value for your (internal or external) customer, ask yourself:

- What does the customer need?
- What do they want to know?
- Are there specific guidelines when it comes to style?
- How can I give them more than they asked for? Are there additional insights in the data of which they may not be aware?
- How do I communicate with my customer? Do I need to confirm any additional information before presenting my work?

Occasionally we provide specific requirements for a data set to give our community an opportunity to practice the above approach. When we do, consider treating it just like a work project. You are allowed to explore the data and find your own insights, while considering the requirements provided.

A good data analyst can identify and understand the requirements. A great data analyst will ask questions when they do not clearly understand. It can be difficult to ask questions in a virtual

environment like Makeover Monday, but the advantage is that Makeover Monday provides a safe space to ask lots of questions. If you take what you have learned, you will feel more comfortable asking questions when working with stakeholders in your company and it will become much easier.

Identify the Challenges

A good place to start when working with unfamiliar data is to identify the challenges you are facing. When we say identify, we literally mean writing the challenges down. Nearly every chart you see will be embedded within an article or story of some sort. Read the article. It is likely to have a lot of information that will give you enough context to get started.

Once you have read the article, write down the keywords, write down the definitions, and identify any acronyms and what they mean. This list can go on and on. The more challenges you identify, the less of a struggle it will be to work with unfamiliar data.

For example, we used a data set for the NCAA basketball tournament. Eva did not know what the scores meant or how the tournament was structured, so she researched it. Eva looked at regional clusters, different seeds, rounds, and trends, and she created time series charts, bar charts, scatter plots, heat maps, highlight tables, box and whisker plots, and bump charts. You name it, she tried them all. She looked at winning margins, upsets, and number of wins by team, and continued her analysis until she found something that (i) made sense to her, (ii) she could explain to someone else who does not know basketball, and (iii) was simple. The result was the simple timeline visualization in Figure 1.1.

Gain Insights from Metadata

When creating a data visualization, we are often tempted to begin exploring the data immediately rather than focusing on doing analysis first. It is beneficial to first get familiar with the data itself. Use the following questions as a guide to becoming familiar with the data.

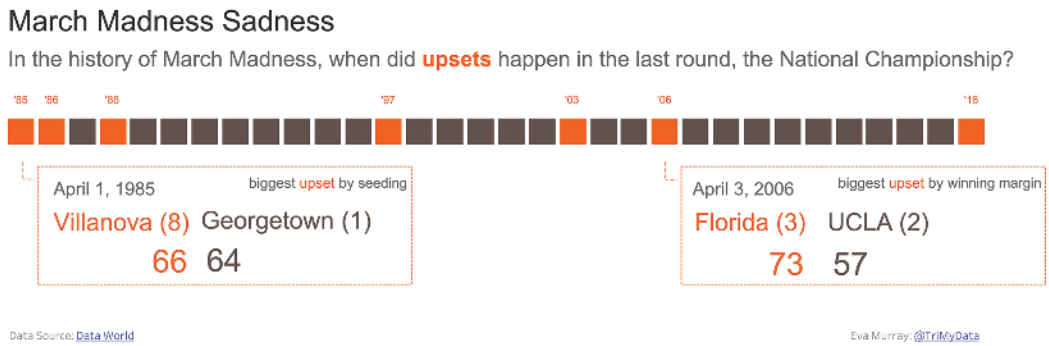

FIGURE 1.1 Upsets in the NCAA basketball tournament championship game.

What Data Types Are in the Data set?

Data types include numeric values, date fields, text fields, Boolean fields, spatial objects, and so on. Knowing the data types before starting your analysis helps you during the analytical process. If a field is translated as text that should be numeric, now is the time to change it. If there are specific default aggregations (for example, one of your metrics should always be calculated as an average), set these up as default. Give fields more understandable names, so they are clearer and easier to work with.

What Range of Values Do the Fields Contain?

A text field containing product categories may hold less than a dozen different values, while a field of individual product names may include hundreds or thousands of different values. Understanding the possible values or range for each field helps you get a better feel for the data. Become familiar with the data by sampling the first few rows of the data set.

Is the Data Complete?

Identifying missing or incomplete data is critical for accurate analysis. For example, when working with dates, first check the range of dates, then look for any missing dates. Comparing years, quarters, or months is challenging and could be inaccurate if some dates are missing.

With geographical data, check for missing locations. For example, if you have state-level data for the United States, check to see if all states are accounted for in the data. If there are some missing, verify whether that is to be expected or if there is a problem. Decide what to do if there is indeed a problem. Make certain you understand the impact of proceeding with incomplete data. If the data set should have data for every state and you do not notice it, then you could make assumptions about aggregations at the region or country level that are incorrect.

Explore the Data

Once you have some context, begin exploring the data. What do the field names mean? How much variety do the fields have? Is there a data hierarchy? What happens when I compare data across fields? Does one field affect another?

These are all very simple questions to ask and even simpler to answer. Explore the data by building lots and lots of charts. Remember, you are ultimately a data analyst. A good data analyst needs to be able to explore the data to find a story or an insight.

In Figure 1.2, Daniel Caroli chose to compare one field to another. He saw distributions in the data and thought to compare ages across years. Daniel used *all* of the data and included a box plot as a summary to help give context. He made it easier to understand by including some simple text. This is a great example of making the complex simple when working with unfamiliar data.

Once you have a grasp on the fields, focus on the metrics. How varied are they? Is there a wide or narrow distribution? What do the metrics mean? What is the proper way to aggregate the data? Can the data even be aggregated? Are there relationships between the metrics?

One trap people fell into with data about the top 500 YouTube gaming channels was to confuse YouTube *views* with YouTube *videos*. Views and videos are *not* the same thing. One video can be viewed many times, yet some people used views and videos interchangeably. This is a clear sign that the data analyst did not take the time

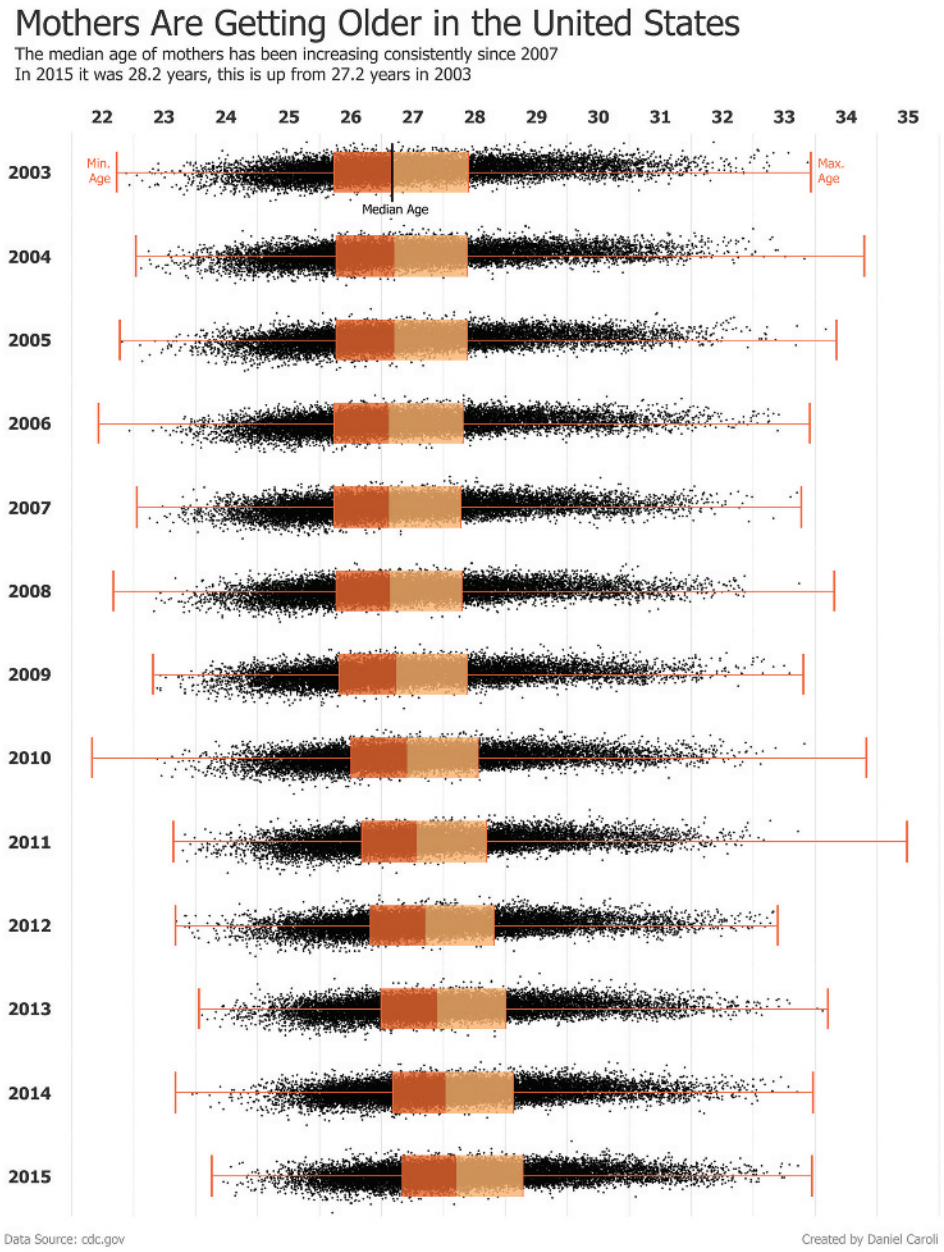

FIGURE 1.2 Median age of mothers in the United States when giving birth.

to *understand* the data. Understanding the data is fundamental to ensuring your analysis is correct.

In his visualization (Figure 1.3), Marc Soares demonstrated how to compare metrics effectively, how to research a topic to provide sufficient context, and how to explain the topic to his audience.

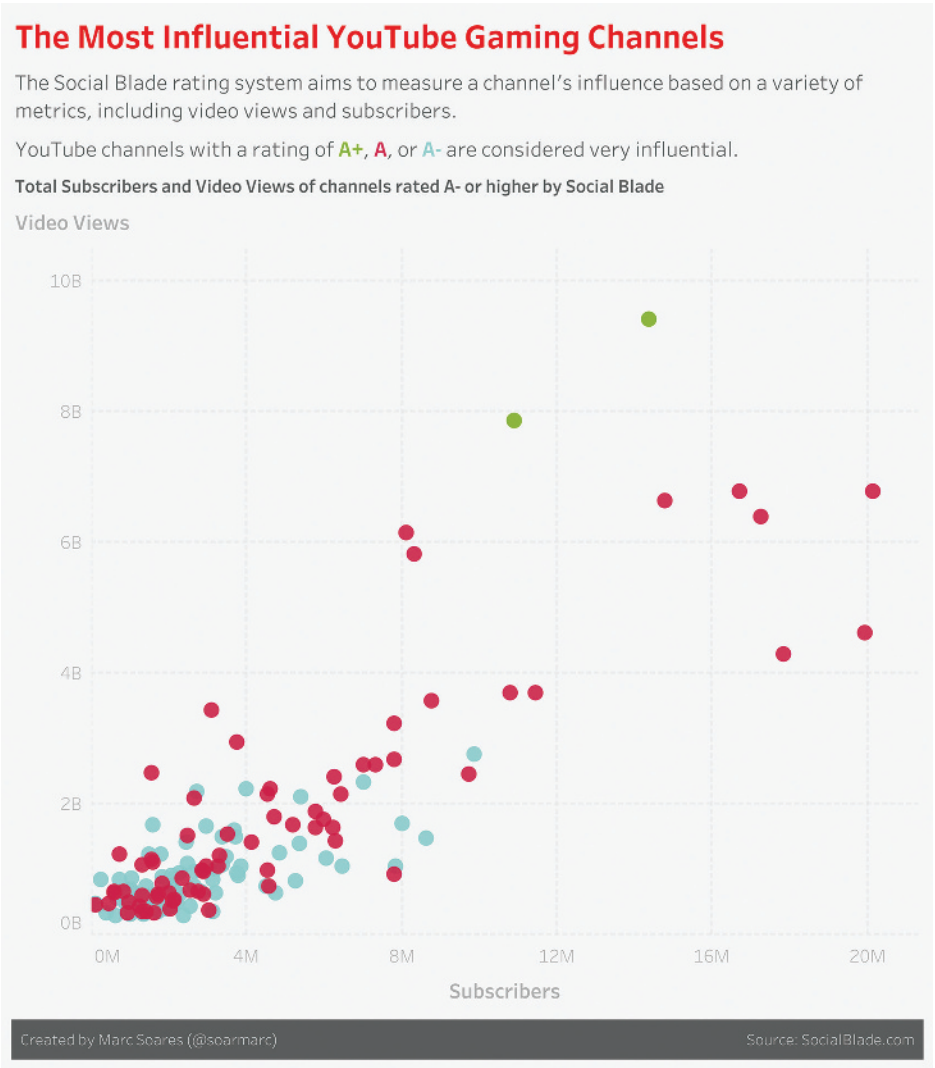

FIGURE 1.3 Scatterplot of YouTube views vs. channel subscribers.

If you want to be great at data analysis, you have to practice, you have to be able to explore and understand the data to find insights, and you have to communicate your findings well.

Charlie Hutcheson has created over 100 Makeover Monday visualizations and through this practice has developed excellent analytical skills. He demonstrated this particularly well when visualizing life expectancy data (Figure 1.4). He explored the data, found some

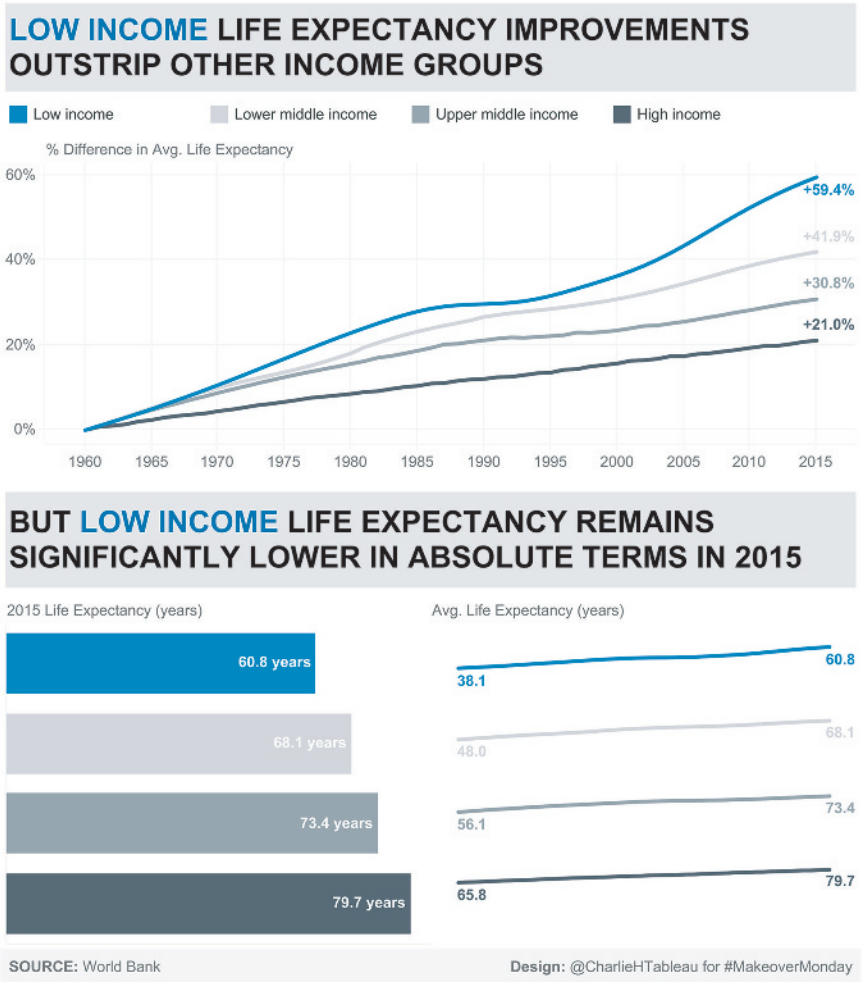

FIGURE 1.4 Analysis of life expectancy for different income levels.

insights, then used his data visualization and storytelling skills to communicate his findings succinctly, clearly, and simply.

Now that you have taken the time to read the article for context and explored the fields and metrics to gain more understanding, it is time to think about how to most effectively simplify the data.

Remove Unnecessary Fields

Consider a wide data set, that is, a data set that has a lot of fields and metrics. After you have taken the time to understand which data is important, remove all of the fields you do not need.

The data set about Chicago taxi trips was both wide (19 columns) and tall (105 million rows). It is very unlikely that you will need all 19 columns and all 105M rows. Pooja Gandhi limited the data set to only three fields and four metrics, yet she was able to create a stunning visualization that shows the when and where of Chicago taxi trips, as can be seen in Figure 1.5.

Focus on a Subset of the Data

Closely related to eliminating the fields you do not need is reducing the complexity of the data through filtering or limiting the data set. For example, in week 16 of 2017, we challenged everyone to visualize 784 million records of UK medical prescriptions. There is no good way to visualize this much data in a single chart, so to create an effective design you are almost forced to focus on a subset of the data. In Figure 1.6, Adam Crahen focused his analysis on a single drug compared to other drugs, with the number of prescriptions aggregated up to the monthly level.

Through the use of highlighting, Adam intentionally guides his audience to the data the analysis is about. He uses color in his title to clarify what the line represents, reducing clutter by eliminating the need for a legend. Ultimately, the highlighting helps provide focus.

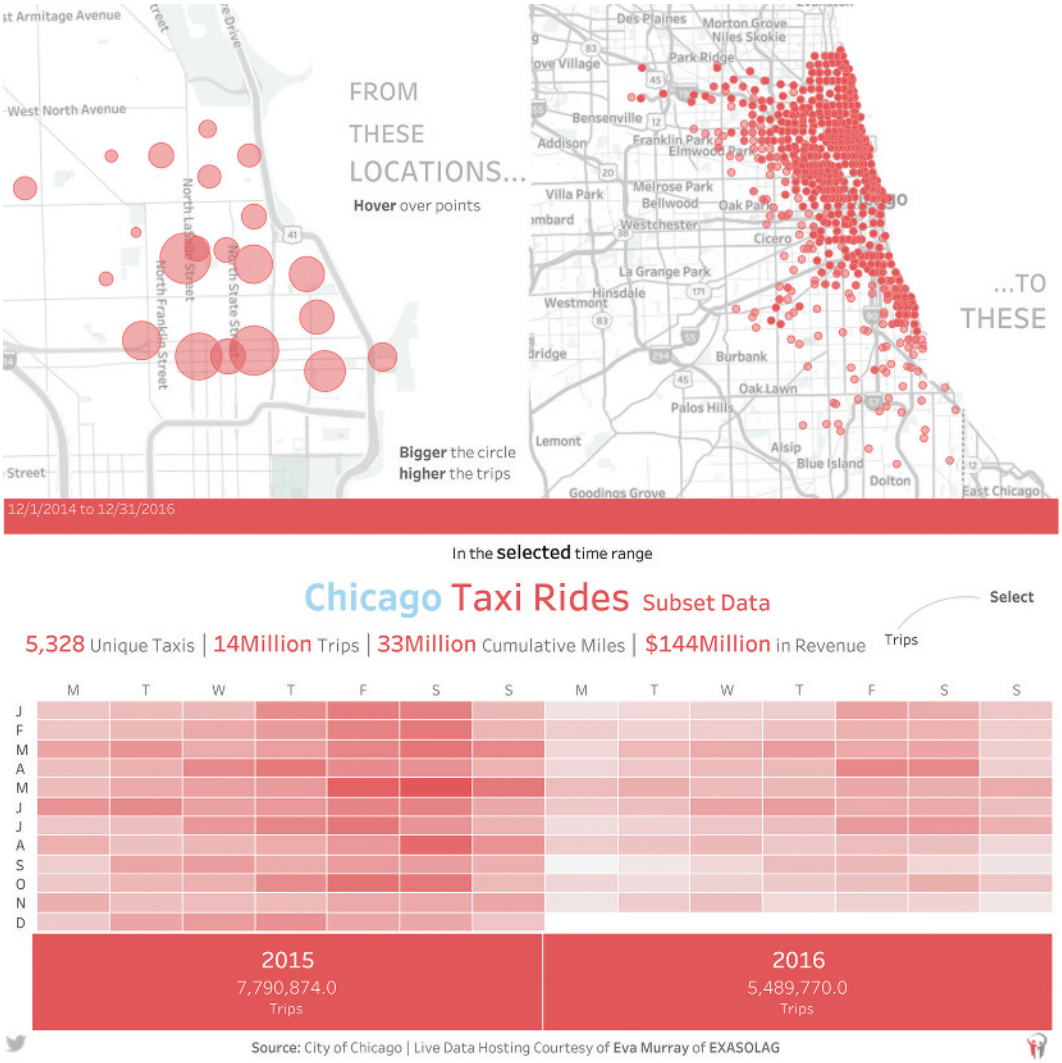

FIGURE 1.5 Visualization of Chicago taxi trips by Pooja Gandhi.

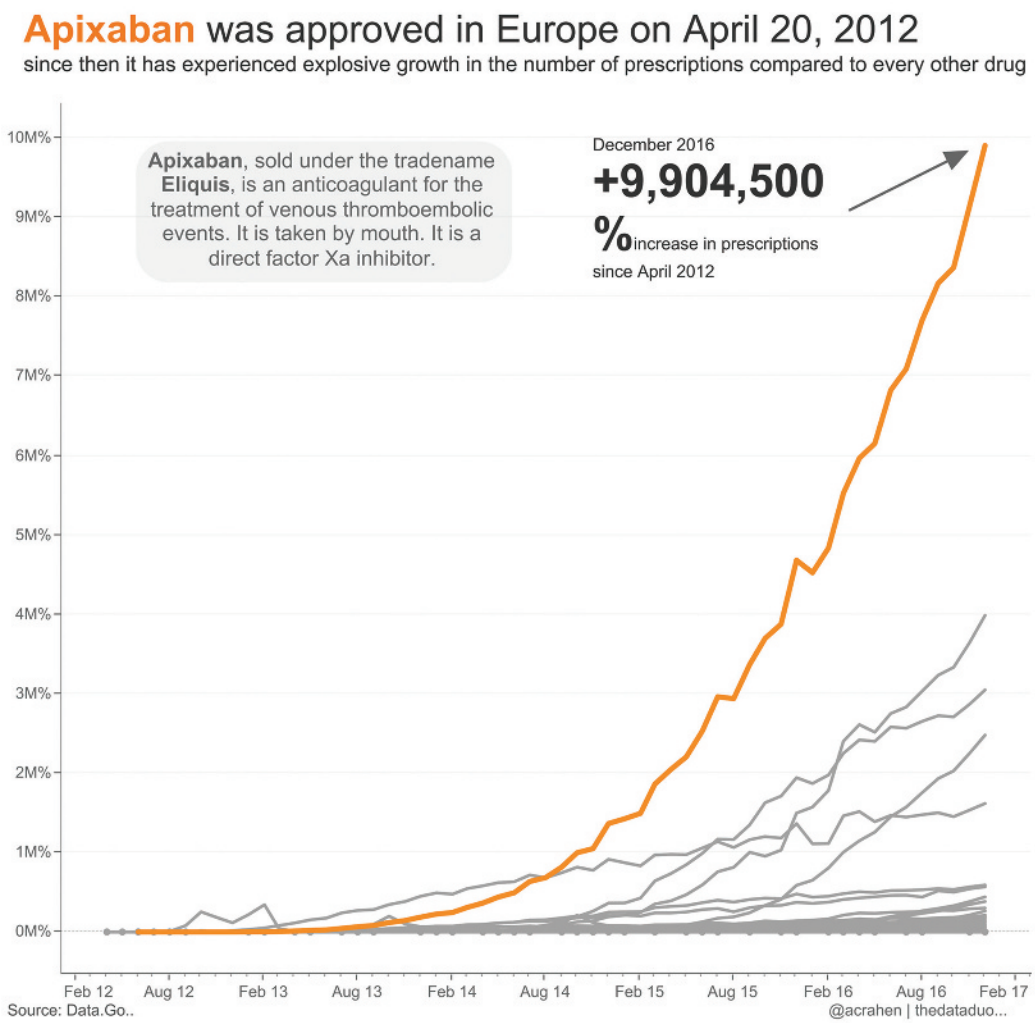

FIGURE 1.6 The growth of Apixaban prescriptions in the United Kingdom.

The idea with highlighting is to move the focus to one particular field while keeping all others in scope but moving them to the background. Sean Miller used highlighting to effectively allow his audience to pick one state's income distribution and compare that state to all others, as seen in Figure 1.7. Essentially Sean is viewing a subset of the data to create simplicity.

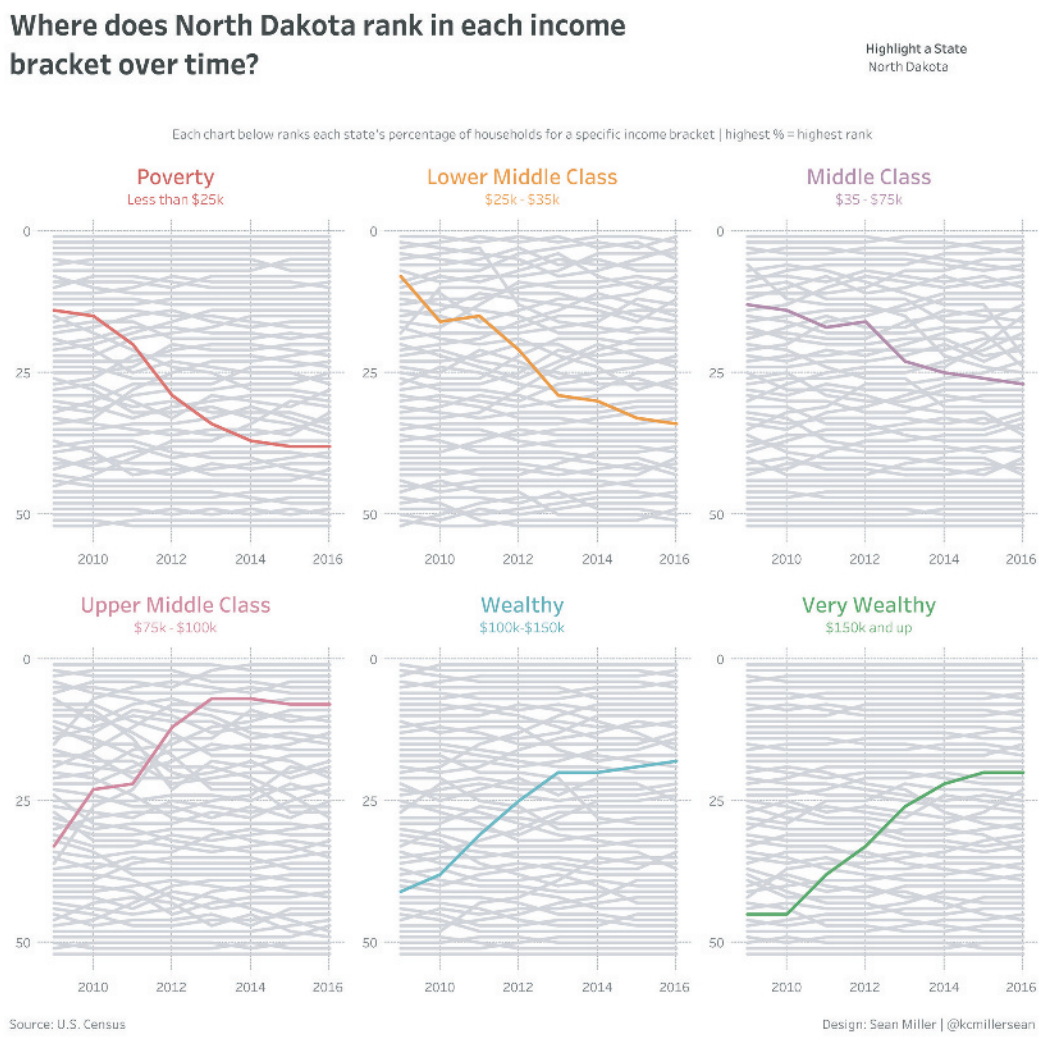

FIGURE 1.7 Example of combining highlighting and context.

Analysis versus Visualization

Visualizing every data point in a data set will not necessarily add to our understanding of the data. We have to be able to differentiate between **analysis** and **visualization**.

Yes, we need to make use of all the available data to identify outliers and trends, to build more accurate analytical and predictive models, and to increase the certainty of our assumptions and conclusions. Visualizations, however, should represent our findings in a way the audience can comprehend easily and quickly. For example, the audience does not need to see every data point as dots on a map or every product in the product portfolio in a scatter plot. Often, aggregated data or a representative sample of the data is sufficient to support the analysis.

Design the visualizations so that they provide *enough* information to focus the audience on the insights, such as certain trends or outliers. Combining aggregate level analysis with detailed information should be done with care.

In Figure 1.8 about air quality in the United States, Pooja Gandhi combines aggregated analysis via the tile map and heatmap, while providing detailed data at the county level in the dot plot in the lower half of her visualization.

Pooja provides a state filter to reduce the complexity. Within each year (i.e., the vertical panes dividing the data into bands), a single dot represents a county. The horizontal distribution of the data points shows increasing levels of ozone measurements and the growth in the number of readings and measurement stations over time, from left to right.

Pooja very effectively combined the two concepts of analysis and visualization. The top half of her dashboard has an analytical and exploratory focus, while the bottom half provides an impactful visual representation of the data to show changes over time.

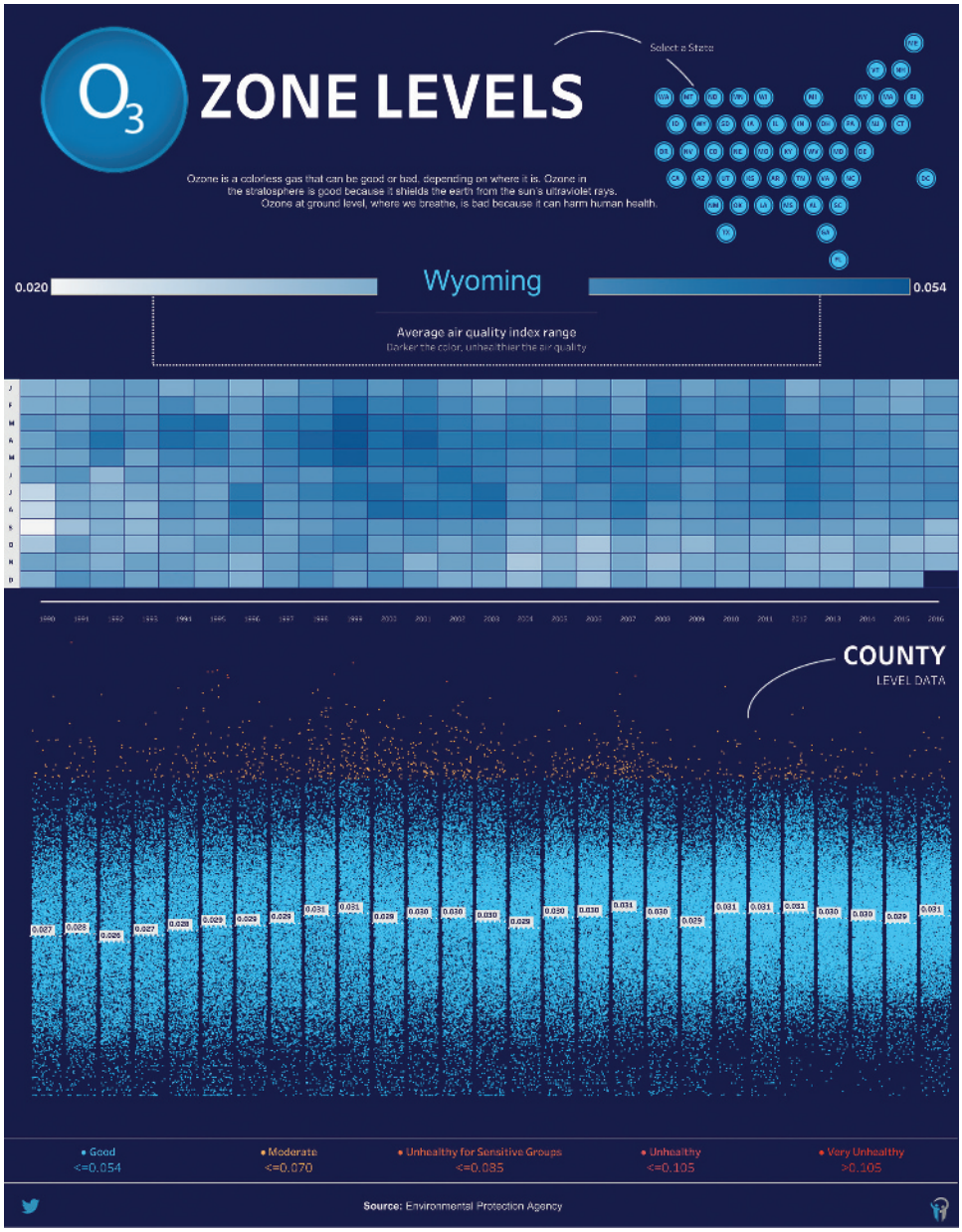

FIGURE 1.8 Combine multiple charts to allow aggregate and detailed analysis.

Take Your Time

While we encourage people not to spend more than an hour on Makeover Monday each week, we do not want you to think that you should rush through your analysis. By taking your time, slowing down, and thinking through the analysis, you will create better work.

Complicated data sets take time to understand, especially those that are wide and tall. Some data sets are more difficult to understand than others. In those cases, take your time to really understand the data. Analysis takes time. Make sure your calculations are accurate. Verify your numbers. Slow down.

Not taking the necessary amount of time to work through your analysis thoroughly, step by step, can easily lead to incorrect assumptions. One of our challenges involved a very simple data set about German car production, including production and export numbers per month for passenger cars and trucks.

Does that data tell you anything about sales? Not really. It does not tell you how many cars were sold and neither does it tell you when people buy cars in Germany. Some people saw that production decreased during August and December compared to the other months. Yet too many people did not bother to try to understand *why* production drops during those months. Germans probably take holidays during August, but maybe it is also a conscious choice by the manufacturer to reduce production output during that month. Those are possible explanations, but not definite conclusions.

Be careful with the statements you add into your visualizations. You could be right, but you could also be way off. Either way, confirm.

Build Context Through Additional Research

Many data sets warrant additional research to get a clearer picture of the data. For Makeover Monday, we provide participants with an article that gives the story behind the data. This can be a short news story, a longer opinion piece, or even an academic research paper. We encourage everyone to use that information to help them understand the context of the data, and to learn about the data collection process and the intention behind the original visualization that is being used as the basis for the makeover.

While a number of people dive straight into data visualization as soon as the data is published, we find that those who read the available article often find the analysis process easier and are able to show the insights they find more clearly. They use the context, the data definitions, and the available commentary to enhance and strengthen their story.

If you want to build your skills as an analyst, we strongly recommend setting time aside for research to support your data analysis and visualization. Here are some steps you can take to approach a data set and gain a deeper understanding of its topic and background.

Read the Available Information

Most charts, dashboards, and data visualizations do not exist in isolation. Rather, they are part of a larger data story, news article, blog, forum discussion, white paper, academic research paper, or other publication. Take the time to read and understand the information supporting the chart and take note of key information including:

Purpose of the Study: Investigation or Analysis

- Why was the data collected in the first place and by whom?
- Does the purpose give you ideas for building your own data story?
- Is there an interesting insight in the origins of the research that you can use to guide your analysis?

Definitions of Key Metrics and Dimensions

- What do the different field names mean and what does this tell you about the values they contain?
- Are there any assumptions stated with regard to specific values?
- Are there comments on data quality and completeness?

Data Collection Process

- How, when, and where was data collected?
- How reliable are the data collection methods that were used? For example, is data being collected by highly reliable sensors that automatically send data to a database, or is the data collected through questionnaires filled in manually by study participants?
- How was the data processed and treated following the collection and storage in the database?

Insights Shared in the Article

- What are the key messages and conclusions shared by the original authors?
- Are there any claims that are unsubstantiated and that you can target in your analysis?
- Are there data points the original authors were unable to explain or that they disregarded in their work?

These questions are intended to guide you and not to prescribe a strict process to follow. They should encourage you to dig a little further in your analysis and to ensure that your conclusions are based on sufficient research given the available information.

Seek Additional Information

Aside from the initial research, it can be helpful to seek out information from secondary sources, such as those used by the original authors or commentary included with the original analysis. Doing

this research gives you an idea of the type of information that influenced the original research and lets you understand the critique that others may have already written about the data you are about to analyze.

For data sets that provide lots of options for analysis and visualization, spend a little extra time to get a more comprehensive understanding of the topic, the data implications, and the work already done by others. This will save you time in the long run as you build your research and analytical skills while creating stronger arguments to support your analysis.

Find Insights

Every week, the Makeover Monday data is ripe for analysis. Consider applying analytical thinking to your work. What is the data *really* telling you? What can you add to the conversation? What are the unknown unknowns? Thinking like this will help you move from being a good data analyst to a great data analyst.

Sean Hughes, in analyzing the salaries of the Obama and Trump administrations (Figure 1.9), went through 15 iterations, exploring the fields and grouping them in all sorts of ways before settling on comparing salaries by gender in the two administrations. Sean took the salaries and binned them into $30 000 increments. Lastly Sean used colors associated with the political parties, gave the chart a clear title and subtitle, and included a reference line for context.

By looking at the data from multiple perspectives, Sean was able to identify insights that were not immediately obvious. Sean has organized the chart in a way that clearly shows the differences between the administrations. He uses colors that are associated with the political parties and a subtitle that explains what he found.

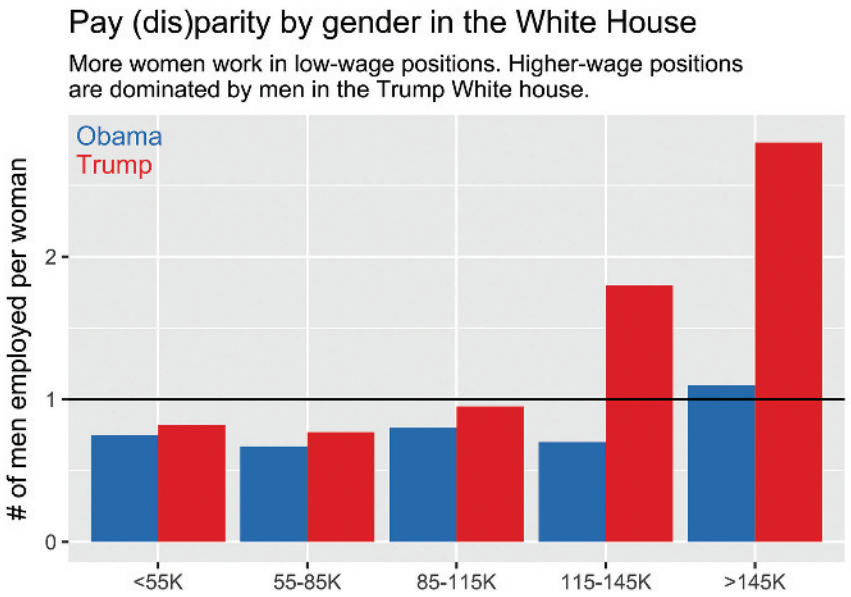

FIGURE 1.9 Comparison of White House salaries by Sean Hughes.

Educating Your Audience

When we gave our community data about solar eclipses, this provided a great opportunity for finding insights and teaching the audience about a topic that fascinates many people. This data set was great for exploring and made it possible to use visual analysis as a way to find a story.

It was a chance to explore how to:

- Put different metrics and fields into the view and manipulate them for unique visualizations so that fascinating patterns could emerge
- Swap axes and understand how this changes the visuals and makes a story more impactful or changes its focus
- Size and color fields to see what happens

Many participants were able to identify patterns and find insights in the data that resulted in informative and stunning visualizations. In Figure 1.10, Marc Reid taught his audience about solar eclipses through his interactive visualization. He showed patterns in the data and used those patterns as the focus of his work.

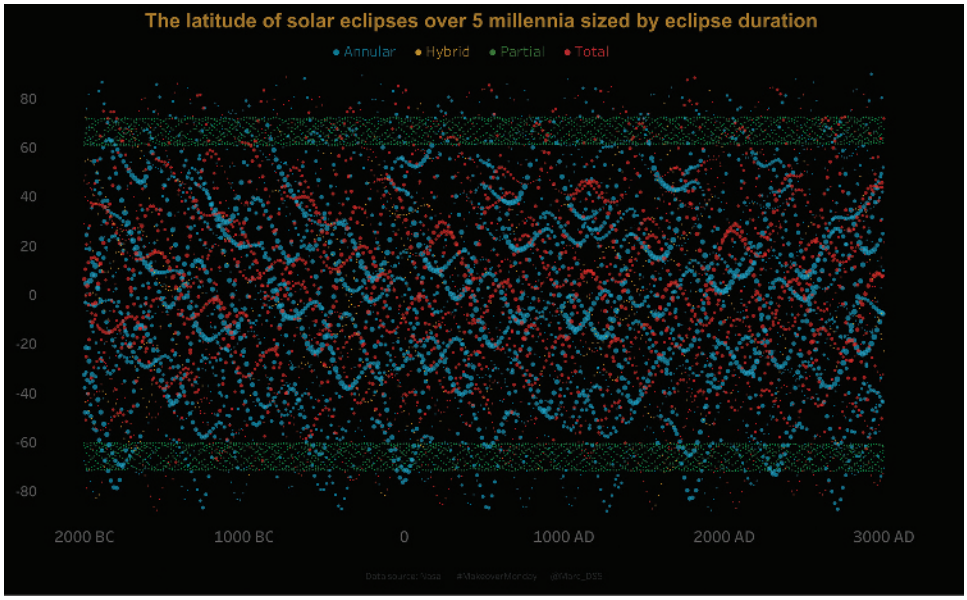

FIGURE 1.10 Plot of solar eclipses by Marc Reid.

Sebastián's visualization in Figure 1.11 revealed Saros cycles in the data, which made for a really great visualization of the pattern throughout the series and over time. He also added a number of explanations to help his audience better understand the life cycle of a solar eclipse, turning his dashboard into an interactive astronomy lesson.

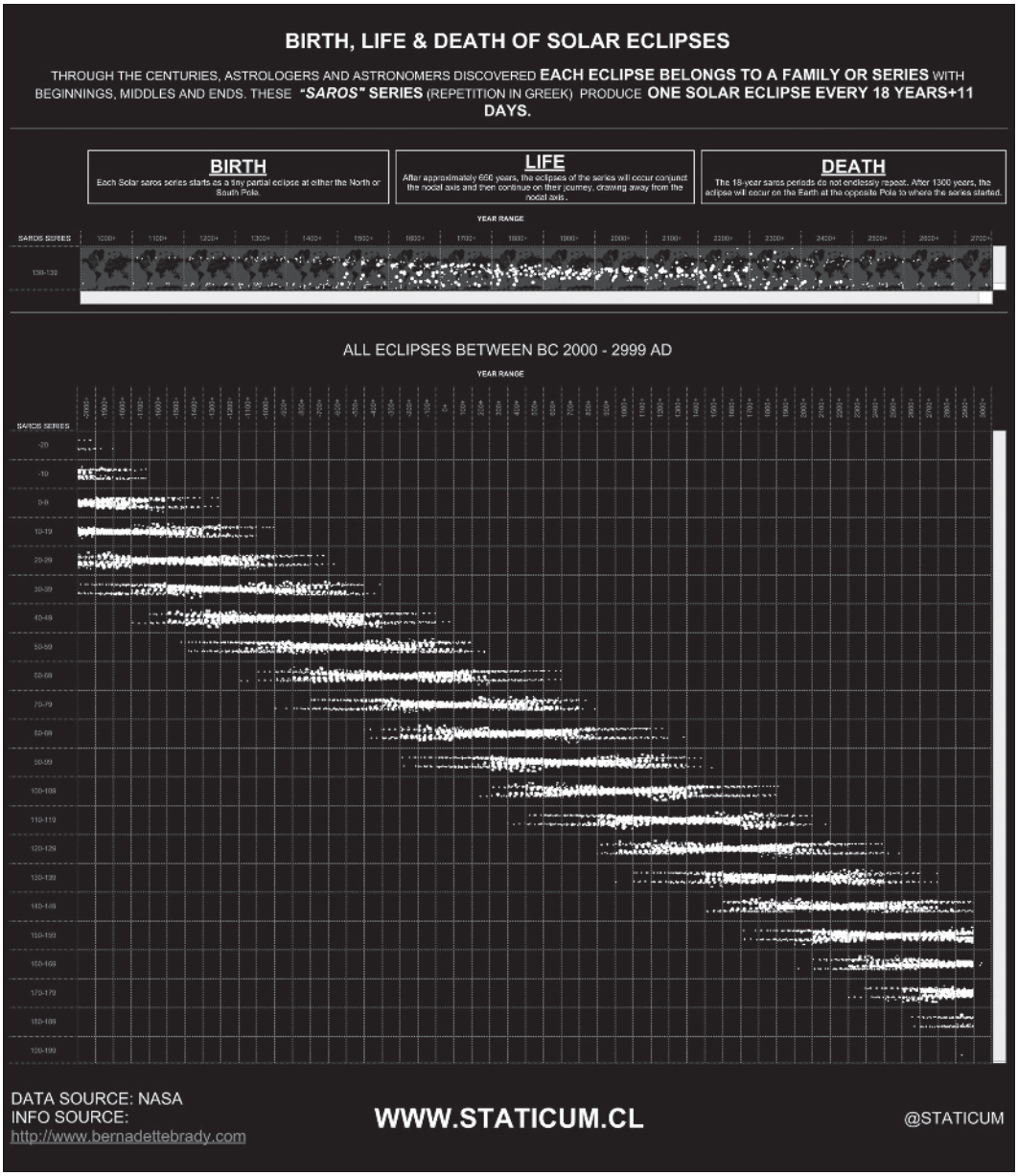

FIGURE 1.11 Design can be used to educate your audience.

Communicate Clearly

Great data analysts are exceptional communicators. They make the complex simple by taking their insights and displaying them clearly and effectively to a broad audience. Rarely are people born as great communicators; even more rare is someone who is innately great at communicating data. Storytelling is a skill we can all learn. Books like Cole Nussbaumer Knaflic's *Storytelling with Data* (Wiley, 2015) help teach the basics of communicating data effectively.

Combining theoretical learning with books and practical learning through projects like Makeover Monday creates a platform to practice clear and effective communication of information through data visualization.

This can be a daunting challenge when the subject is very "niche" and your audience might not be familiar with it. Take week four of 2018 as an example, when Eva posted a data set about turkey vultures.

- What are turkey vultures?
- Why should we care about their migration patterns?
- What do the patterns even mean?

When you communicate your insights clearly, it can pull your audience in. The best compliment from those viewing your data visualizations is if they start caring about a topic previously unfamiliar to them and possibly even take actions to make a difference.

Lindsay Betzendahl created the stunning visualization in Figure 1.12 to explain turkey vultures and their migration patterns simply.

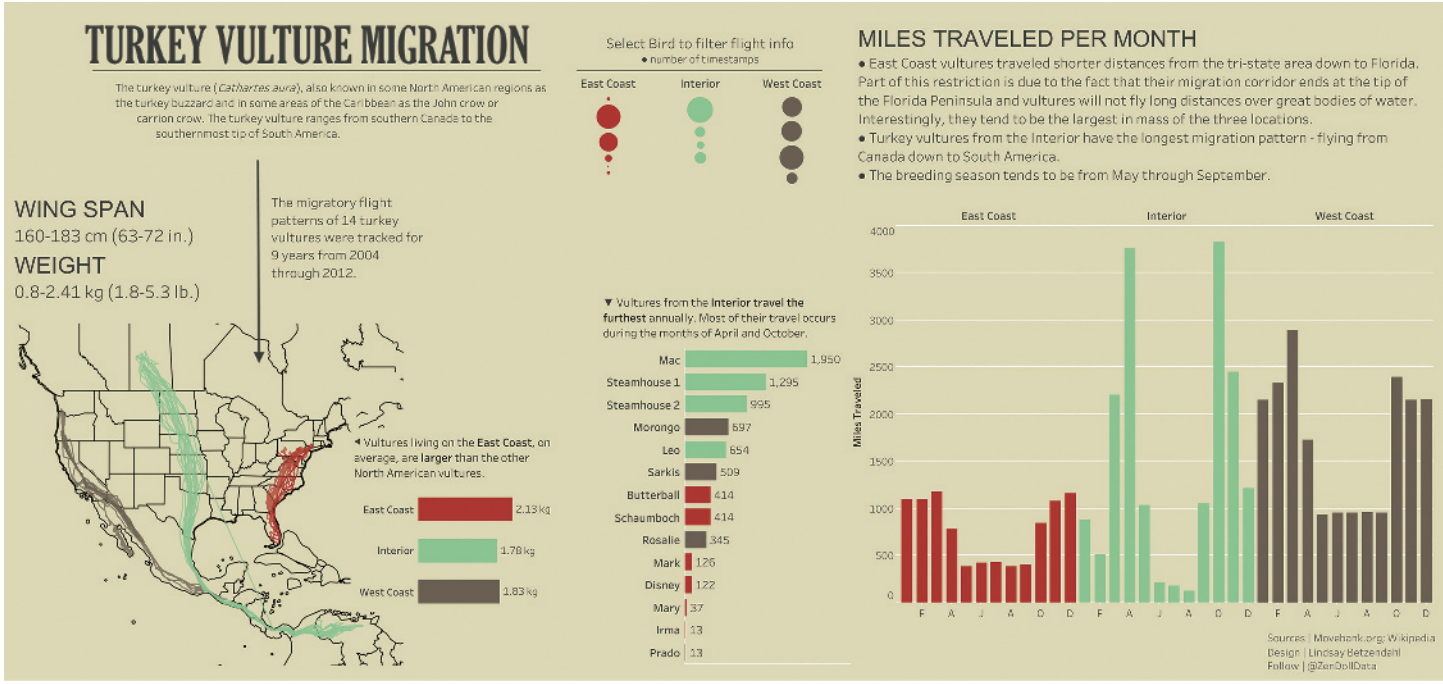

FIGURE 1.12 Turkey vulture migration patterns by Lindsay Betzendahl.

Lindsey communicated clearly by:

- Using the subtitle to explain the subject
- Including basic characteristics of the birds
- Using annotations to highlight specific insights
- Including additional information as context through a series of bullet points
- Using consistent, simple, soft colors that help the audience connect the analytical components

In the same week, Klaus Schulte taught us about one of the three nonmigratory birds in the data set (see Figure 1.13).

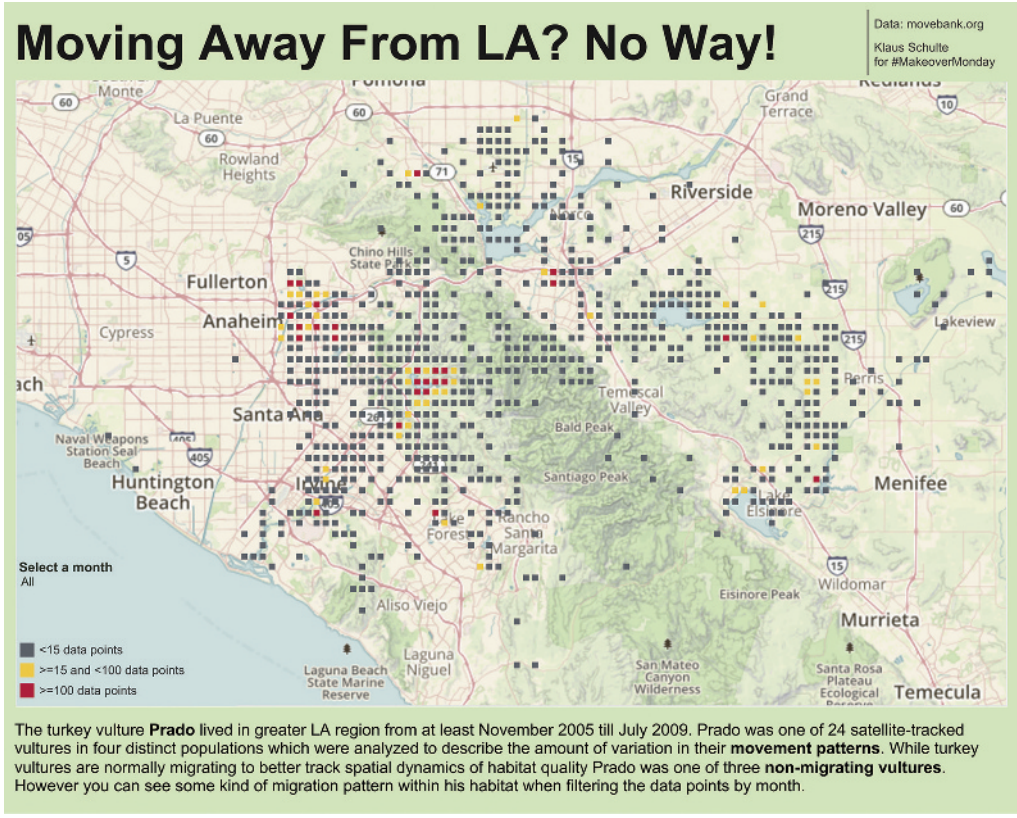

FIGURE 1.13 Simplify the data to communicate the analysis more effectively.

Klaus communicated clearly by:

- Summarizing his analysis in the footer
- Choosing colors that are easy to distinguish from each other
- Telling his audience how to interact
- Using simple shapes that frame the visualization

Ask Questions

The final characteristic common to those highly effective data analysts in the Makeover Monday community is having the confidence to ask questions. Through our day-to-day work and our engagement with the data visualization community we often come across people who clearly do not understand a particular topic as well as they would like to, yet they are afraid to ask questions. Typically, we are told this is because they do not want to look unknowledgeable, but, for us, asking questions shows curiosity and it is this curiosity that separates great data analysts from everyone else. They want to know the "why" and they will continue to ask questions and explore the data until they feel like they know why.

When we gave the Makeover Monday community a data set about the ethnicity of players in Major League Baseball (MLB), Mike Cisneros wanted to provide additional contextual information. How does MLB compare with other major sports leagues in America? He wanted to know the "how" and the "why" of baseball's demographics. What he found, by researching and including additional data, was that Major League Soccer, where more than 28% of players are African-American or Black international athletes, was the most ethnically balanced league in North American professional sports. The design of his visualization in Figure 1.14 effectively supports and highlights his findings.

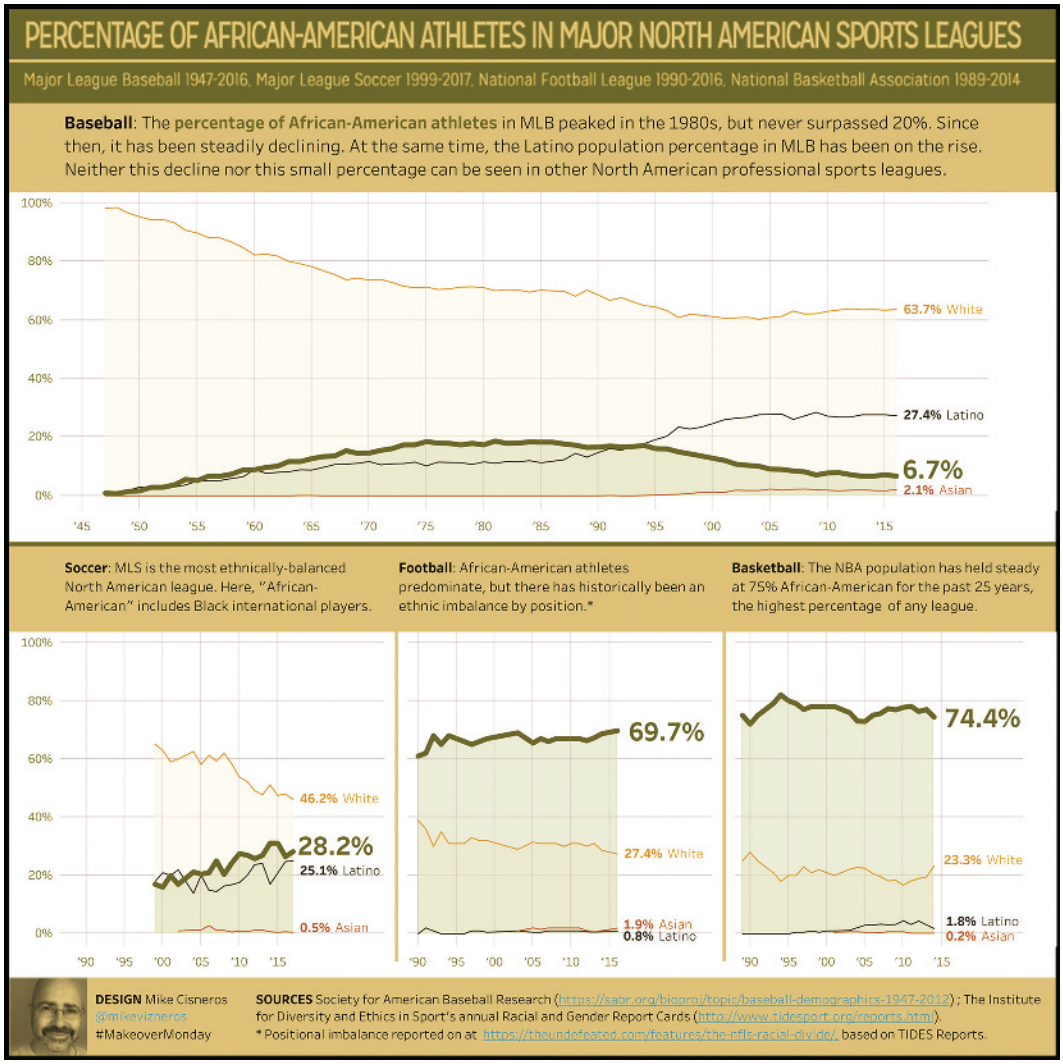

FIGURE 1.14 Comparison of ethnicities across major sports leagues in North America.

Great data analysts do not settle for reporting a single number. They will dig deeper to provide context. Paweł Wróblewski, when visualizing data about personality types (Figure 1.15), could have simply reported the results of the Myers–Briggs personality test. However, he wanted the reader to be able to compare their own personality type with the rest of the United States. By doing so,

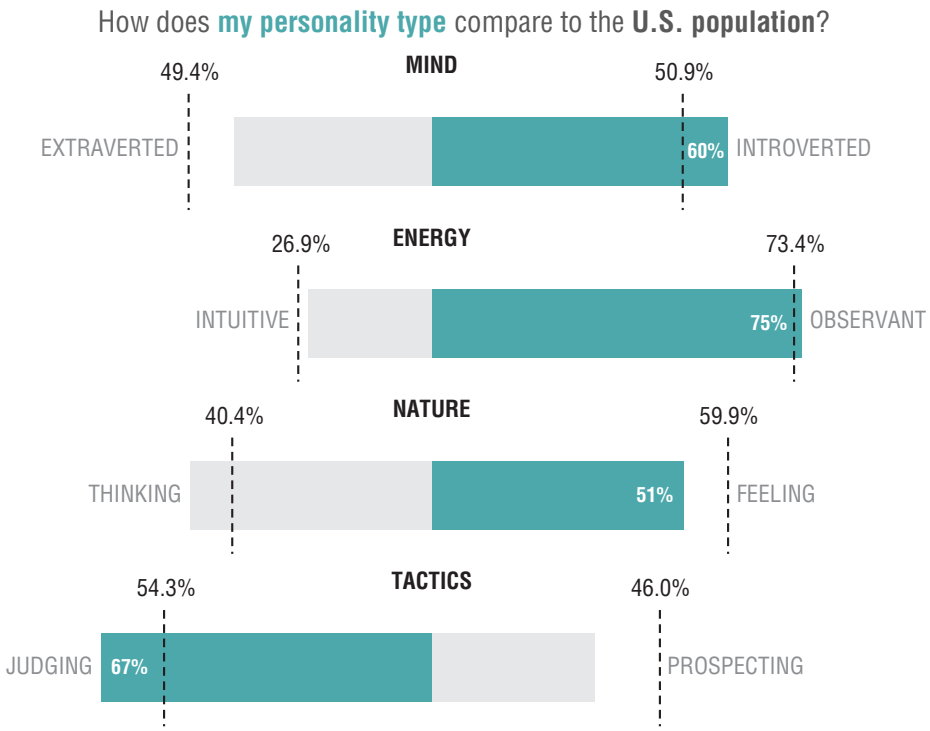

FIGURE 1.15 Allowing the user to answer their own question draw them into the analysis.

Paweł gave the reader more context. He answered the question "compared to what?"

The Makeover Monday community has struggled at times with asking questions. For one week we provided data about Andy's marathon and Eva's triathlon and a list of questions we were curious to know the answers to, yet very few people bothered to ask us any questions. It was clear early into the week that many people did not understand the data and saw our list of questions as requirements and actually opted out of tackling the data altogether or focused on something "safe" instead.

If you are unsure about any aspect of a data set or the context provided, you have to ask question to get clarification. This applies in your professional environment at work and also for projects like

Makeover Monday where we do not expect people to understand everything every single week. If your boss does not appreciate you asking questions, then you probably need a new boss. Curiosity should be celebrated, especially for analysts. Without curiosity, analysts will simply regurgitate the numbers in the data.

It is obvious from his visualization (Figure 1.16) and its storytelling that Joel Gluck was curious about the topic of the water footprint of the foods we eat. Joel clearly wanted to know *why*. He did additional research and called out the main contributors to water usage as well as other harmful impacts on the planet caused by our food choices. By going further into the topic and making it his personal goal to share his insights in a way that will make his audience think and act, Joel showed that as an analyst he can take the given data far beyond reporting numbers.

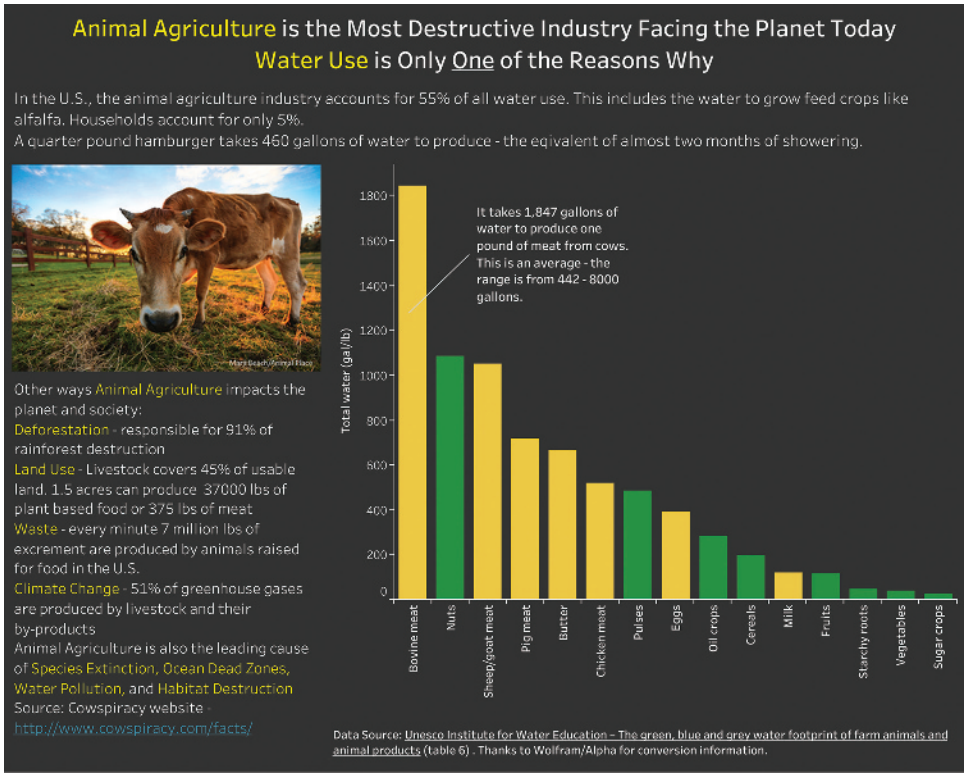

FIGURE 1.16 Including additional data is an effective method for explaining the why.

Summary

Becoming a good data analyst takes practice, diligence, and dedication. It will not come easy and it will take a long time. One of the most satisfying experiences is when you can look back on your previous work, see a progression in your analytical skills, and appreciate how far you have come. This chapter outlined five key habits to help you become a good data analyst:

1. *Grasp unfamiliar data*. Create your own systematic approach to exploring the data in order to understand what it is trying to tell you.
2. *Take your time*. Speeding through the analysis will lead to mistakes and incorrect conclusions. Give yourself enough time to thoroughly understand the data.
3. *Find insights*. Once you understand the data, you will be in a position to develop insights. Take those insights and use them to educate your audience.
4. *Communicate clearly*. If you have done great analysis and you communicate it poorly, the analysis will not have the desired impact. It is critical to develop exceptional communication skills.
5. *Ask questions*. If you have questions or you do not understand, you have to get over your insecurities and ask. The more questions you ask, the better your questions will become. The better your questions become, the better your analysis will be.

By developing these habits, you will improve by leaps and bounds. The more consistent your habits become, the more effective you will be as a data analyst.