Chapter 5: Using basic operations to shape your story

Imagine that you have the following data set in Table 5.1. This raw data is very messy and does not tell a clear story. To help readers understand it, you would probably consider putting it in a line graph to show how energy usage changes over time.

Table 5.1: Average daily energy usage (in kWh) of traditional and energy efficient classrooms in three different U.S. cities

Location	Classroom Type	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Detroit	Traditional	13.4	16.2	36.7	34	38.5	36.4	29.7	15.4	14.2
Detroit	Efficient	9.8	11.3	15.9	17.8	26.2	25.6	20.1	17.1	9.1
Baltimore	Traditional	12.1	14.6	33.0	26.6	34.7	32.8	26.7	13.9	12.8
Baltimore	Efficient	8.0	9.3	20.5	21.8	22.5	21.0	16.5	14.0	7.5
Austin	Traditional	14	13.2	15.9	17.8	19.1	18.7	16.2	12.8	13.8
Austin	Efficient	12.2	10.4	11.1	16.6	16.1	15.4	13.5	13.4	11.9

Your first attempt at a line graph might look something like Figure 5.1:

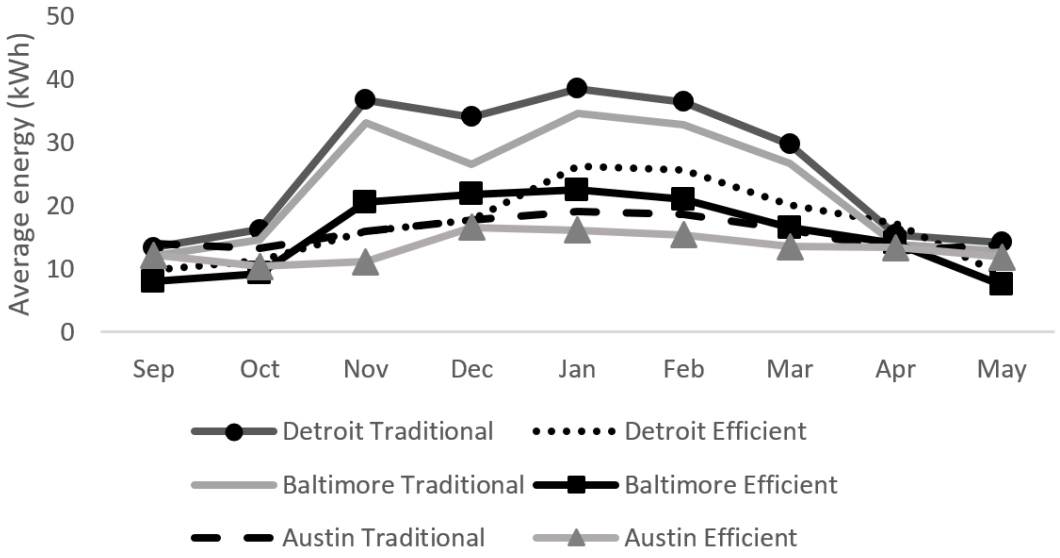


Figure 5.1: Average daily energy usage (in kWh) of energy efficient and traditional classrooms in three different cities

Figure 5.1 contains too much information for us to understand its story. We might next consider breaking the data up into multiple line graphs as in Figure 5.2:

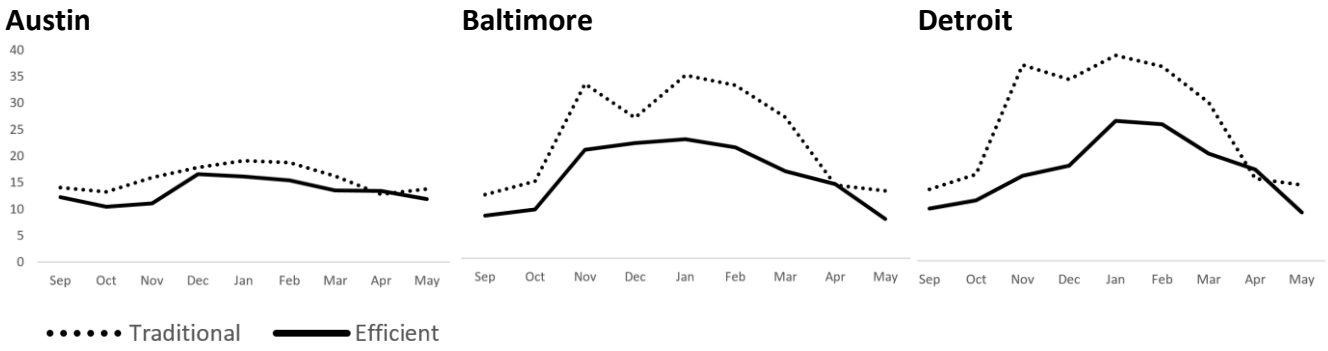


Figure 5.2: Energy data grouped into multiple line charts

Figure 5.2 is certainly much easier to understand than Figure 5.1—and we can see that the energy efficient classrooms use less energy than the traditional, particularly in Detroit and Baltimore. However, even these graphs do not contain a clear-cut answer to the question many readers will have, which is "how much energy overall do the efficient classrooms save?"

Summarizing Data

This is where **summarizing** the data comes in. When we summarize data, we try to communicate the largest amount of information as simply as possible. We use simple mathematical calculations—such as sums, percentages, or averages—to combine rows or columns of our raw data, enabling us to see broad patterns.

Thus, we might transform table 5.1 from the monthly breakdown into an average for the school year as illustrated in Figure 5.3

Location	Classroom Type	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Detroit	Traditional	13.4	16.2	36.7	34	38.5	36.4	29.7	15.4	14.2
Detroit	Efficient	9.8	11.3	15.9	17.8	26.2	25.6	20.1	17.1	9.1
Baltimore	Traditional	12.1	14.6	33.0	26.6	34.7	32.8	26.7	13.9	12.8
Baltimore	Efficient	8.0	9.3	20.5	21.8	22.5	21.0	16.5	14.0	7.5
Austin	Traditional	14	13.2	15.9	17.8	19.1	18.7	16.2	12.8	13.8
Austin	Efficient	12.2	10.4	11.1	16.6	16.1	15.4	13.5	13.4	11.9

Location	Classroom Type	Average daily usage Sep-May
Detroit	Traditional	26.1
Detroit	Efficient	17.0
Baltimore	Traditional	23.0
Baltimore	Efficient	15.7
Austin	Traditional	15.7
Austin	Efficient	13.4

Figure 5.3: Simplifying the data story by averaging for the entire year

Once the data has been condensed in this way, we can now report it as a bar chart that allows readers to grasp more easily the relative differences between efficient and traditional buildings in each city.

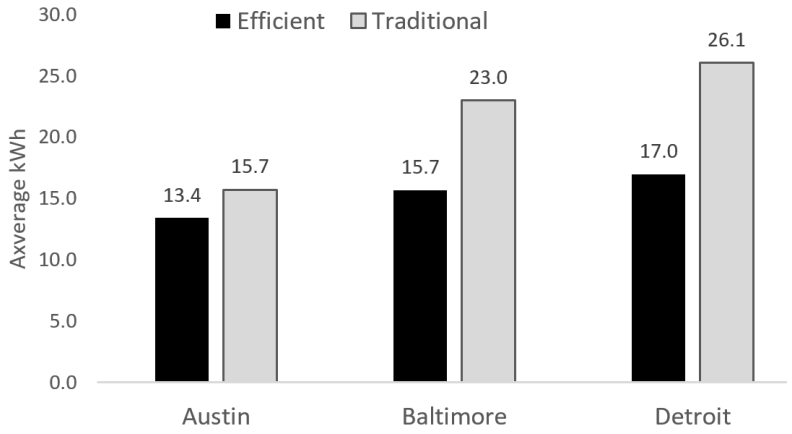
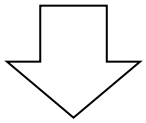


Figure 5.4: Average daily energy usage (in kWh) of traditional and energy efficient classrooms in three different U.S. cities

However, the data can be further condensed to just report one average for traditional buildings and one average for energy efficient buildings in all three cities as is illustrated in Figure 5.5.

Location	Classroom Type	Average daily usage Sep-May
Detroit	Traditional	26.1
Detroit	Efficient	17.0
Baltimore	Traditional	23.0
Baltimore	Efficient	15.7
Austin	Traditional	15.7
Austin	Efficient	13.4



Classroom Type	Average daily usage Sep-May
Traditional	21.59
Efficient	15.35

Figure 5.3: Simplifying the data story by averaging data for all traditional and all efficient buildings

This summarization now allows us to produce Figure 5.4

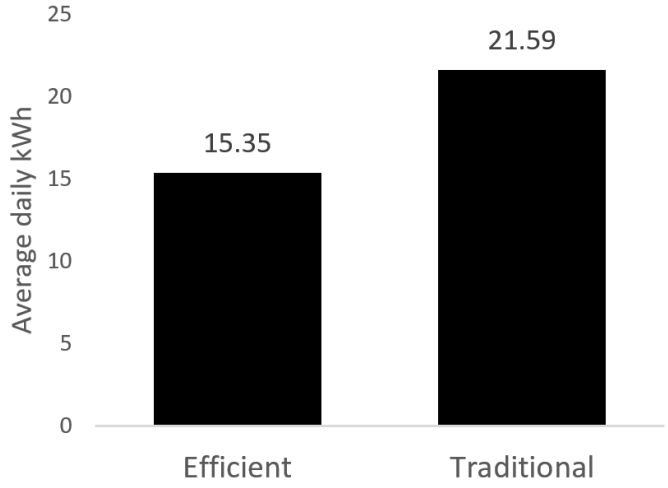


Figure 5.4: Average daily energy usage (in kWh) of energy efficient and traditional classrooms in three different U.S. cities during the school year

Figure 5.4 helps readers quickly estimate that the energy efficient classrooms average approximately 30% less energy over the school year than the traditional classrooms. We lose some nuance and detail when we summarize the data this way, but we also gain the clarity of a simple and persuasive story.

A case study in engineering visualizations: Three ways to summarize data

In the last section we used averages to summarize our data on energy usage. But you actually have additional choices for the types of calculations you use to summarize your data. The most common choices for reporting summarized data include

- Sums (i.e., adding up raw counts)
- Percentages
- Averages

Less common choices include

- Ratios
- Differences (i.e., calculating the differences between two averages or sums)
- Medians
- Rankings

To illustrate the choices you have in what calculations to report, let's look at another dataset. Table 5.2 tallies the number of visualizations found in five industry and five academic engineering reports, summarizing them by visualization type.

Table 5.2: Data visualizations in five industry and five academic engineering reports

Visualization Type	*Industry/Govt (331 pages)*	*Academic (90 pages)*
Tables	76	22
Illustrations & Diagrams	46	25
Line graphs	26	40
Bar graphs	25	2
Other	1	11
Pie graphs	1	0

This data is not nearly as complicated as the energy data beginning this chapter and you can probably pick out a story with a little effort. However, we can make it easier for readers to find the story in this data and also improve its credibility by experimenting with different ways to report and summarize the data.

Readers are probably interested in learning which visualizations are the most common and if there are any differences in the type of visualizations found in industry versus academic reports. One way present this information might be to use a stacked bar graph of the **raw counts** as in Figure 5.5 below.

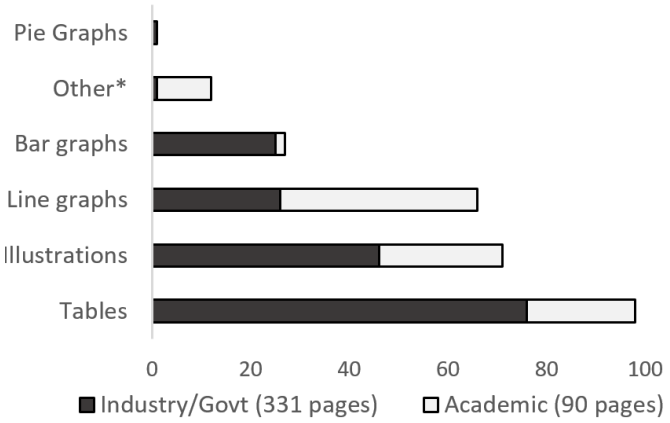


Figure 5.5: Total visualizations found in engineering reports (raw counts)

Figure 5.5 helps us quickly grasp that tables are the most common visualizations in these engineering reports while pie graphs are rarely found. We can also see that bar graphs are rare in academic reports while line graphs appear more common. However, some readers will question the credibility of Figure 5.5 saying that it over-represents industry reports since these reports were over three times as long as academic reports.

One way to address this issue might be to report the **percentage** of visualizations in each type of report. In other words, we could calculate the percent of industry visualizations that are tables, bar graphs, line graphs, etc. And then we could do the same for academic visualizations. Using this calculation method, we might produce something like the cluster bar graph in Figure 5.

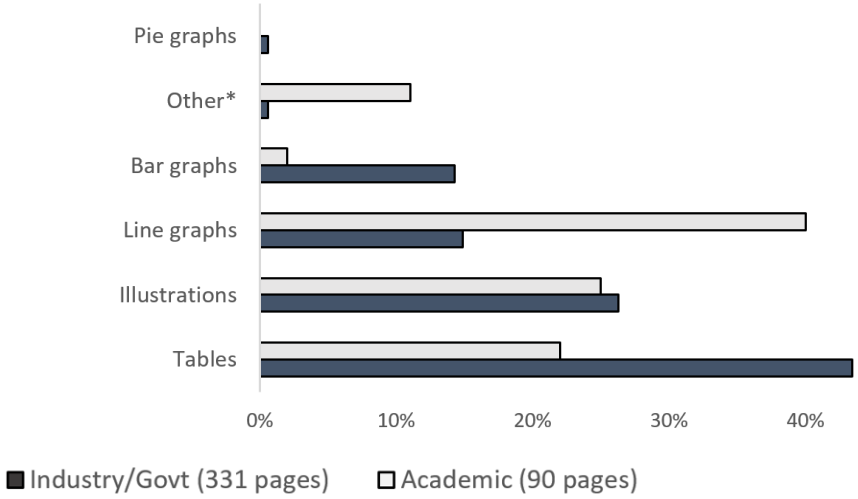


Figure 5.6: Percentage of visualization types found in engineering reports

Figure 5.6 changes our story quite a bit. When we look at our data as percentages, we see that line graphs in academic reports rival tables in industry reports as the most common visualization. We also see that the percentage of tables in academic visualizations is approximately half of that found in industry documents.

Yet another way to display the data is to report **the average** number of visualizations per report page as in Figure 5.7. How does this switch affect the story?

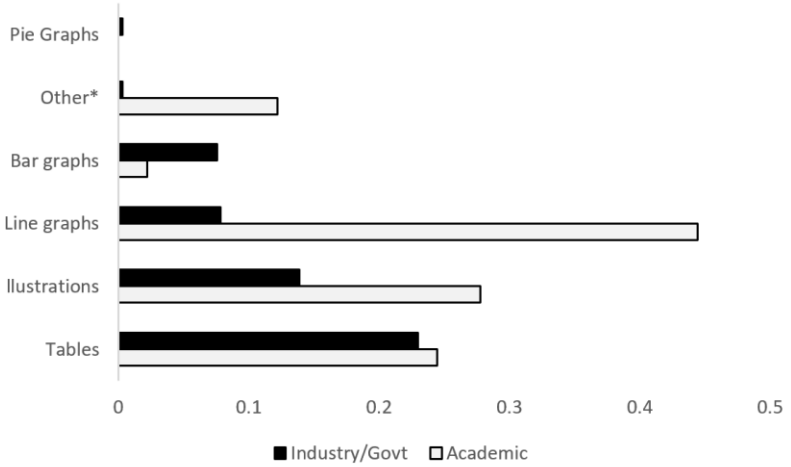


Figure 5.7: Average number of visualizations per page found in engineering reports

In our previous figures, industry visualizations seem to outnumber visualizations found in academic reports. However, when we average by the number of pages analyzed as in Figure 5.7, suddenly academic visualizations appear more common than industry ones. Moreover, line graphs in academic reports jump out as the most common visualization.

So which visualization is best?

All three visualization (and our table of raw data) are ethical: all three give readers the information they need to interpret the data. Therefore, any decisions about which is best depends upon our purpose and the expectations of our audience. If our purpose is to simply show which visualizations are most and least common in engineering, then Figure 5.5 (or perhaps a non-stacked version of it) is probably a good choice. If, however, our purpose is to look at how data reporting differs in industry and academic documents then Figures 5.6 or 5.7 provide a clearer picture of how the two differ with 5.7 providing comparatively accurate information about which type of publication has the most visualizations.

The point of this section is that your choices for data reporting also include the choice of operations for summarizing your data. You can report raw numbers, percentages, or averages. Your choice of these operations influences both the story you communicate and readers' perceptions of your credibility.

Exercise 5.1

Imagine we have conducted a study of students in a class, recording the number of pens and pencils they are carrying on four different dates (yes, this is a silly study). Using the data below, create **three different visualizations** to illustrate three different stories you might tell about this data. Your audience is fellow students and instructors who want to learn more about students' habits and practices. Experiment with summarizing and averaging the data in order to make your story as clear as possible. Be prepared to defend your design choices.

In addition to creating the three visualizations, write up approximately 1-4 sentences for each visualization describing its main and (if relevant) secondary stories.

Be sure that each of your visualizations has a complete caption and that axes are labeled. Follow the design principles from Chapter 4.

You can copy and paste the data into Excel or you can download a spreadsheet at

_____

Table A: Number of pens and pencils carried by students on four different dates

Student	Gender	Major	Year	14-Oct	21-Oct	28-Oct	4-Nov
Arjun	M	Science	Junior	1	1	2	1
Carlos	M	Business	Junior	2	2	3	2
Cerice	F	Science	Soph	2	2	3	2
David	M	Humanities	Soph	3	2	3	2
Felipe	M	Business	Junior	2	2	2	2
Henri	M	Science	Junior	1	2	2	2
Jenna	F	Business	Soph	5	4	4	3
Kaitlyn	F	Humanities	Soph	3	5	6	5
Landon	M	Business	Junior	1	1	2	0
Leo	M	Humanities	Fresh	2	3	4	4
Marcel	M	Science	Junior	1	0	2	1
Maria	F	Humanities	Fresh	6	8	9	4
Matt	M	Business	Junior	2	3	2	2
Mustafa	M	Business	Junior	3	1	3	2
Nisha	F	Humanities	Fresh	4	4	8	5
Ryan	M	Science	Junior	1	1	2	1
Tiffany	F	Science	Senior	2	1	3	2
Zach	M	Science	Senior	1	1	1	1

Creating new variables

In Chapter two we saw how a simple choice of whether to focus on gold medals or total medals altered our story about who "won" the Olympics.

In fact, how to report the Olympics medals is controversial and many systems have been proposed.[1] For instance, the British newspaper, *The Guardian*, has suggested assigning a point value for each type of medal. In this system, a gold medal is worth four points, a silver two, and a bronze one. Such a calculation produces the following table:

Table 5.3: 2008 Weighted Olympic Medals

Country	*Gold (4 pts)*	*Silver (2 pts)*	*Bronze (1 pt.)*	*Weighted Total*
China	51	21	28	**274**
USA	36	38	36	**256**
Russia	23	21	29	**163**
Britain	19	13	15	**117**
Australia	14	15	17	**103**
Germany	16	10	15	**99**
France	7	16	18	**78**

Many people believe that the weighted total in Table 5.3 represents a fairer version of the Olympics winnings. And, of course, other weightings could be proposed, such as awarding three or five points for a gold medal instead of four.

The point here is that our choices for displaying data go beyond simply summarizing and reporting the data we have collected or have been given. We can also experiment with new ways to calculate the data, such as creating a new variable. You and your classmates may have already experimented with creating new variables when working on the browser data from the last chapter. Such creative calculations need to be balanced against how they will affect our credibility.

In addition to creating a new variable such as a weighted ranking, you can also combine your data with other sources to create a more nuanced story of what the data says. For instance, many Olympics observers note that because China and the US are two of the three most populous countries in the world, it is little surprise that they dominate contests such as the Olympics. In this sense, the accomplishments of Australia—a far smaller country which ranks 52nd in world population—seem much more remarkable than those of China, the US, or Russia, which rank 1, 3 and 9 in world population respectively.

Thus, we could revise our story about the Olympic medals by finding reliable data on the population of each country and averaging our data by those numbers. Table 5.4 shows how the story changes if we display the data by medals per capita (i.e., per person):

[1] It should be noted that the official position of the Olympic committee is that the Olympics is a contest among individuals and not nations.

Table 5.4: 2008 Gold Olympic Medals per Capita[2]

Country	*Gold Medals*	*Population*	*Gold Medals per million residents*
Jamaica	6	2,705,827	**2.22**
Bahrain	1	1,234,571	**0.81**
Estonia	1	1,318,005	**0.76**
Mongolia	2	2,736,800	**0.73**
New Zealand	3	4,432,620	**0.68**
Georgia	3	4,469,200	**0.67**
Australia	14	22,880,619	**0.61**

Table 5.4 shows a completely different set of "winners" than we saw in Chapter 2 where our tables just reported raw counts of the medals. Jamaica now emerges as the leader (thanks in part to the extraordinary Usain Bolt) while Australia is the only country that appears on both the per capita calculation of the 2008 Olympics and the more common representations found in Chapter 2.

And, of course, population is not the only relevant measure. A country's relative wealth also matters since wealthier countries have more resources to invest in their athletes. Thus, we could also report medals by a country's Gross Domestic Product (GDP), a standardized measure of wealth. We could also choose to examine a country's performance against previous years' accomplishments, measuring which country gained the most.

Our ability to make different arguments does ***not*** mean that we can say whatever we want. For instance, no amount of manipulation will ever make Venezuela's one bronze medal in 2008 come out on top. But we do have a wide range of choices that need to be considered. Critical readers of data need to be able to imagine other ways that data may be presented—just as critical readers of verbal or written arguments need to imagine the different ways a quotation might be paraphrased.

Figure 5.8 below shows the choices writers have when deciding how to present their data. As we discussed above, data can be reported different ways, such as raw counts, percentages, and averages. It can be broken down and summarized by subgroups such as gender, age, or building type. New variables can be created. And existing data can be combined with reliable data from other sources to create a more detailed and nuanced story.

[2] Data from http://www.medalspercapita.com/#golds-per-capita:2008

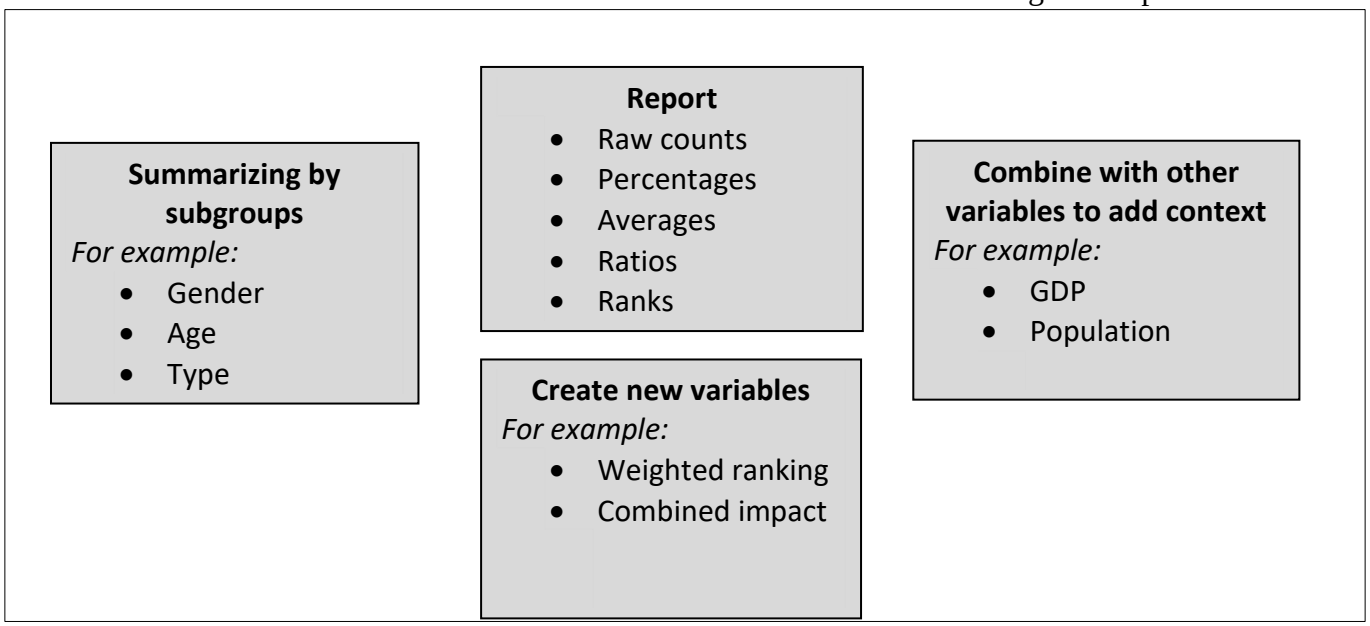


Figure 5.8: Options for counting data include reporting totals, percentages or averages; combining with other variables; creating new variables; and summarizing by subgroups.

Of course, you will not always have complete flexibility. In many technical and academic contexts, data is consistently counted and presented the same way. Such standardization helps readers quickly interpret your data and compare it to that collected by others. However, when you read and write about data in less controlled circumstances, you should be very attentive to the choices authors make and how these choices in turn shape the stories writers tell.

Exercise 5.2: 2012 Olympic Data

The Table below shows selected data from the 2012 London summer Olympics. If your last name begins

- A-H: Create a visualization that presents **Cuba** in the best possible light
- I-N: Create a visualization that presents **Hungary** in the best possible light
- O-Z: Create a visualization that presents **China** in the best possible light

Your visualization should follow the principles covered in Chapter 4.

Experiment with presenting the data both credibly and less credibly. Experimenting with ways to manipulate numbers to support a specific purpose or story prepares you to see how others might make similar manipulations.

A downloadable spreadsheet version of this table can be found at _____

Country	Population	GDP	Athletes	Gold	Silver	Bronze
US	309,349,000	15,094,000,000,000	531	46	29	29
China	1,338,300,000	7,298,100,000,000	371	38	27	23
UK	62,232,000	2,431,590,000,000	556	29	17	19
Russia	141,750,000	1,857,770,000,000	435	24	26	32
South Korea	48,875,000	1,116,250,000,000	255	13	8	7
Hungary	10,000,000	140,029,000,000	158	8	4	5
Cuba	11,258,000	64,100,000,000	110	5	3	6

Summary

Our options for presenting data are not limited to the type of visual representation we chose. We can also perform basic mathematical operations to summarize, condense, or combine data. The most common operations are to:

- Average data across groups
- Report percentages
- Create new variables
- Combine with other data (such as population counts)

Performing these basic mathematical operations can help you put your story into relief. You may not know at the beginning of your data analysis which story you want to tell. Thus, you should experiment not only with different visualizations but also different ways to combine different values so that you can tell a story that is both clear and credible.

Exercise 5.3: The Browser Wars

Read Appendix A: "Browser Performance Tests" by the web development company Midas.com. Midas has compiled some very useful data that can help users decide which of the five most popular web browsers may help them with their productivity. However, their report could be more effective.

a. Do you find Midas' measurements credible? Do you have questions about how they collected or reported their data that would change your overall interpretation of the story?

b. In their final summary, Midas evaluates the browsers by ranking them in each category and then summing those rankings. What are the pros and cons of this decision to use relative rankings to determine the "winner"? What are some other ways they might have summarized or reported this data to obtain a different "winner"?

c. By placing data in sixteen different charts, Midas makes it difficult for readers to compare across different categories. Table 4.8 below is a data dump of all the different metrics Midas reported. Reorganize this table to tell a clearer story. You can report the variables differently, rearrange the order in which they appear, combine variables, or create new variables. Be creative while also still being credible. (a copy of this data in spreadsheet form can be found at….)

d. Midas concludes that Chrome is the first place winner followed by Opera. Does your revised table suggest a different outcome? Make additional formatting changes to your table to emphasize the browser you think should be the "winner"

e. Write 2-8 sentences describing the primary story of your new visualization and any secondary stories.

Table B: Browser performance on nine different tasks

Task	Chrome 31	Firefox 25	IE 11	Opera 17	Safari 5.1
Cold start	11.35s	3.37s	3.66s	11.98s	7.82s
Non-cold start	0.98s	1.40s	0.01s	4.10s	0.52s
Page load time (non-cached load)	3.869s	11.091s	3.588s	3.526s	6.272s
Page load time (reload from cache)	1.638s	5.179s	1.966s	1.685s	3.354s
Base memory usage (blank tab)	99.5mb	49.1mb	29.5mb	91.7mb	35.0mb
Memory Usage (10 open tabs)	423.1mb	163.1mb	259.0mb	308.6mb	224.7mb
HTML 5 Compliance	93%	82%	70%	88%	56%
CSS3 Compliance	57%	53%	57%	58%	45%
JavaScript performance (Sunspider: lower is better)	702.2	902.6	769.4	871.9	965.7
JavaScript performance (Dromaeo: higher is better)	465.84	335.51	245.48	417.56	238.36
JavaScript performance (Speed-Battle: higher is better)	166.65	170.3	127.4	135.92	66.37
JavaScript performance (Peacekeeper: higher is better)	667	439	324	677	332
JavaScript performance (Octane: higher is better)	3183	2848	2288	3113	684