In the examples previous to this lesson, we completed one iteration of the data analysis process. In some cases, a single iteration may be sufficient, but in most real-life cases, you'll need to iterate at least a few times. From the examples above, there are still some things left to do:

- **Price Survey Data**: We ended the example by fitting a Gamma distribution model. But how does that fit the data? What would we expect from the data if they truly followed a Gamma distribution (we never made that plot)? Is there a better way to capture that spike in the distribution right around $10?
- **Ozone and Temperature**: The smoother suggested a nonlinear relationship between temperature and ozone, but what is the reason for this? Is the nonlinearity real or just a chance occurrence in the data? Is there a known physical process that explains the dramatic increase in ozone levels beyond a certain temperature and can we model that process?

Ultimately, you might be able to iterate over and over again. Every answer will usually raise more questions and require further digging into the data. When exactly do you stop the process then? Statistical theory suggests a number of different approaches to determining when a statistical model is "good enough" and fits the data well. This is not what we will discuss here, but rather we will discuss a few high-level criteria to determine when you might consider stopping the data analysis iteration.

Are you out of data?

Iterative data analysis will eventually begin to raise questions that simply cannot be answered with the data at hand. For example, in the ozone/temperature analysis, the modeling suggested that there isn't just a simple relationship between the two variables, that it may be nonlinear. But the data can't explain precisely why such a nonlinear relationship might exist (although they can suggest certain hypotheses). Also, you may need to collect additional data to determine whether what you observe is real or simply a fluke or statistical accident. Either way, you need to go back out into the world and collect new data. More data analysis is unlikely to bring these answers.

Another situation in which you may find yourself seeking out more data is when you've actually completed the data analysis and come to satisfactory results, usually some interesting finding. Then, it can be very important to try to *replicate* whatever you've found using a different, possibly independent, dataset. In the ozone/temperature example, if we concluded that there were a nonlinear relationship between temperature and ozone, our conclusion might be made more powerful if we could show that this relationship were present in other cities besides New York. Such independent confirmation can increase the strength of evidence and can play a powerful role in decision making.

Do you have enough evidence to make a decision?

Data analysis is often conducted in support of decision-making, whether in business, academia, government, or elsewhere, we often collect an analyze data to inform some sort of decision. It's important to realize that the analysis that you perform to get yourself to the point where you can make a decision about something may be very different from the analysis you perform to achieve other goals, such as writing a report, publishing a paper, or putting out a finished product.

That's why it's important to always keep in mind the *purpose* of the data analysis as you go along because you may over- or under-invest resources in the analysis if the analysis is not attuned to the ultimate goal. The purpose of a data analysis may change over time and there may in fact be multiple parallel purposes. The question of whether you have enough evidence depends on factors specific to the application at hand and your personal situation with respect to costs and benefits. If you feel you do not have enough evidence to make a decision, it may be because you are out of data, or because you need to conduct more analysis.

Can you place your results in any larger context?

Another way to ask this question is "Do the results make some sort of sense?" Often, you can answer this question by searching available literature in your area or see if other people inside or outside your organization have come to a similar conclusion. If your analysis findings hew closely to what others have found, that may be a good thing, but it's not the only desirable outcome. Findings that are at odds with past results may lead down a path of new discovery. In either case, it's often difficult to come the the right answer without further investigation.

You have to be a bit careful with how you answer this question. Often, especially with very large and complex datasets, it's easy to come to a result that "makes sense" and conforms to our understanding of how a given process *should* work. In this situation, it's important to be hyper-critical of our findings and to challenge them as much as possible. In our experience, when the data very closely match our expectation, it can be a result of either mistakes or misunderstandings in the analysis or in the data collection process. It is critical to question every aspect of the analysis process to make sure everything was done appropriately.

If your results do *not* make sense, or the data do not match your expectation, then this is where things get interesting. You may simply have done something incorrectly in the analysis or the data collection. Chances are, that's exactly what happened. For every diamond in the rough, there are 99 pieces of coal. However, on the off-chance that you've discovered something unusual that others have not yet seen, you'll need to (a) make sure that the analysis was done properly and (b) replicate your findings in another dataset. Surprising results are usually met with much scrutiny and you'll need to be prepared to rigorously defend your work.

Ultimately, if your analysis leads you to a place where you can definitively answer the question "Do the results make sense?" then regardless of how you answer that question, you likely need to **stop your analysis and carefully check every part of it**.

Are you out of time?

This criterion seems arbitrary but nevertheless plays a big role in determining when to stop an analysis in practice. A related question might be "Are you out of money?" Ultimately, there will be both a time budget and a monetary budget that determines how many resources can be committed to a given analysis. Being aware of what these budgets are, even if you are not necessarily in control of them, can be important to managing a data analysis. In particular, you may need to argue for more resources and to persuade others to given them to you. In such a situation, it's useful to know when to stop the data analysis iteration and prepare whatever results you may have obtained to date in order to present a coherent argument for continuation of the analysis.

NOTE: Parts of this reading were taken from *The Art of Data Science* by Peng & Matsui.