There are five key characteristics of a good question for a data analysis, which range from the very basic characteristic that the question should not have already been answered to the more abstract characteristic that each of the possible answers to the question should have a single interpretation and be meaningful. We will discuss how to assess this in greater detail below.

As a start, the question should be of **interest** to your audience, the identity of which will depend on the context and environment in which you are working with data. If you are in academia, the audience may be your collaborators, the scientific community, government regulators, your funders, and/or the public. If you are working at a start-up, your audience is your boss, the company leadership, and the investors. As an example, answering the question of whether outdoor particulate matter pollution is associated with developmental problems in children may be of interest to people involved in regulating air pollution, but may not be of interest to a grocery store chain. On the other hand, answering the question of whether sales of pepperoni are higher when it is displayed next to the pizza sauce and pizza crust or when it is displayed with the other packaged meats would be of interest to a grocery store chain, but not to people in other industries.

You should also check that the question has **not already been answered.** With the recent explosion of data, the growing amount of publicly available data, and the seemingly endless scientific literature and other resources, it is not uncommon to discover that your question of interest has been answered already. Some research and discussion with experts can help sort this out, and can also be helpful because even if the specific question you have in mind has not been answered, related questions may have been answered and the answers to these related questions are informative for deciding if or how you proceed with your specific question.

The question should also stem from a **plausible** framework. In other words, the question above about the relationship between sales of pepperoni and its placement in the store is a plausible one because shoppers buying pizza ingredients are more likely than other shoppers to be interested in pepperoni and may be more likely to buy it if they see it at the same time that they are selecting the other pizza ingredients. A less plausible question would be whether pepperoni sales correlate with yogurt sales, unless you had some prior knowledge suggesting that these should be correlated.

If you ask a question whose framework is not plausible, you are likely to end up with an answer that's difficult to interpret or have confidence in. In the pepperoni-yogurt question, if you do find they are correlated, many questions are raised about the result itself: is it really correct?, why are these things correlated- is there another explanation?, and others. You can ensure that your question is grounded in a plausible framework by using your own knowledge of the subject area and doing a little research, which together can go a long way in terms of helping you sort out whether your question is grounded in a plausible framework.

The question, should also, of course, be **answerable**. Although perhaps this doesn't need stating, it's worth pointing out that some of the best questions aren't answerable - either because the data don't exist or there is no means of collecting the data because of lack of resources, feasibility, or ethical problems. For example, it is quite plausible that there are defects in the functioning of certain cells in the brain that cause autism, but it not possible to perform brain biopsies to collect live cells to study, which would be needed to answer this question.

Specificity is also an important characteristic of a good question. An example of a general question is: Is eating a healthier diet better for you? Working towards specificity will refine your question and directly inform what steps to take when you start looking at data. A more specific question emerges after asking yourself what you mean by a "healthier" diet and when you say something is "better for you"? The process of increasing the specificity should lead to a final, refined question such as: "Does eating at least 5 servings per day of fresh fruits and vegetables lead to fewer upper respiratory tract infections (colds)?" With this degree of specificity, your plan of attack is much clearer and the the answer you will get at the end of the data analysis will be more interpretable as you will either

recommend or not recommend the specific action of eating at least 5 servings of fresh fruit and vegetables per day as a means of protecting against upper respiratory tract infections.

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