

2.2: Sanity Checking and Data Cleaning

Regardless of where you obtain your data, it is important to do some sanity checks to ensure that nothing is drastically flawed. For instance, you can check the minimum and maximum values of key input parameters (i.e., columns) of your data to see if anything looks obviously wrong. One of the exercises in Chapter 8 encourages you explore other approaches for verifying your data. R also provides good plotting functions to quickly obtain a visual indication of some of the key relationships in your data set. We will see some examples of these functions in [Section 3.1](#).

If you discover obvious errors or flaws in your data, you may have to eliminate portions of that data. For instance, you may find that the performance reported for a few system configurations is hundreds of times larger than that of all of the other systems tested. Although it is possible that this data is correct, it seems more likely that whoever recorded the data simply made a transcription error. You may decide that you should delete those results from your data. It is important, though, not to throw out data that looks strange without good justification. Sometimes the most interesting conclusions come from data that on first glance appeared flawed, but was actually hiding an interesting and unsuspected phenomenon. This process of checking your data and putting it into the proper format is often called data cleaning.

It also is always appropriate to use your knowledge of the system and the relationships between the inputs and the output to inform your model building. For instance, from our experience, we expect that the clock rate will be a key parameter in any regression model of computer systems performance that we construct. Consequently, we will want to make sure that our models include the clock parameter. If the modeling methodology suggests that the clock is not important in the model, then using the methodology is probably an error. We additionally may have deeper insights into the physical system that suggest how we should proceed in developing a model. We will see a specific example of applying our insights about the effect of caches on system performance when we begin constructing more complex models in [Chapter 4](#).

These types of sanity checks help you feel more comfortable that your data is valid. However, keep in mind that it is impossible to prove that your data is flawless. As a result, you should always look at the results of any regression modeling exercise with a healthy dose of skepticism and think carefully about whether or not the results make sense. Trust your intuition. If the results don't feel right, there is quite possibly a problem lurking somewhere in the data or in your analysis.

This page titled [2.2: Sanity Checking and Data Cleaning](#) is shared under a [CC BY-NC 4.0](#) license and was authored, remixed, and/or curated by [David Lilja](#) ([University of Minnesota Libraries Publishing](#)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.