When you find an issue with your data

When you are getting ready for data analysis, you might realize you don't have the data you need or you don't have enough of it. In some cases, you can use what is known as proxy data in place of the real data. Think of it like substituting oil for butter in a recipe when you don't have butter. In other cases, there is no reasonable substitute and your only option is to collect more data.

Consider the following data issues and suggestions on how to work around them.

Data issue 1: no data

Possible Solutions	Examples of solutions in real life
Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data.	If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees.
If there isn't time to collect data, perform the analysis using proxy data from other datasets. *This is the most common workaround.*	If you are analyzing peak travel times for commuters but don't have the data for a particular city, use the data from another city with a similar size and demographic.

Data issue 2: too little data

Possible Solutions	Examples of solutions in real life
Do the analysis using proxy data along with actual data.	If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors.
Adjust your analysis to align with the data you already have.	If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: *this conclusion applies to adults 25 years and older only*.

Data issue 3: wrong data, including data with errors*

Possible Solutions	Examples of solutions in real life
If you have the wrong data because requirements were misunderstood, communicate the requirements again.	If you need the data for female voters and received the data for male voters, restate your needs.
Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors.	If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values.
If you can't correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won't cause systematic bias.	If your dataset was translated from a different language and some of the translations don't make sense, ignore the data with bad translation and go ahead with the analysis of the other data.

***Important note:** Sometimes data with errors can be a warning sign that the data isn't reliable. Use your best judgment.*

Use the following decision tree as a reminder of how to deal with data errors or not enough data: