What to do when there is no data

Earlier, you learned how you can still do an analysis using proxy data if you have no data. You might have some questions about proxy data, so this reading will give you a few more examples of the types of datasets that can serve as alternate data sources.

Proxy data examples

Sometimes the data to support a business objective isn't readily available. This is when proxy data is useful. Take a look at the following scenarios and where proxy data comes in for each example:

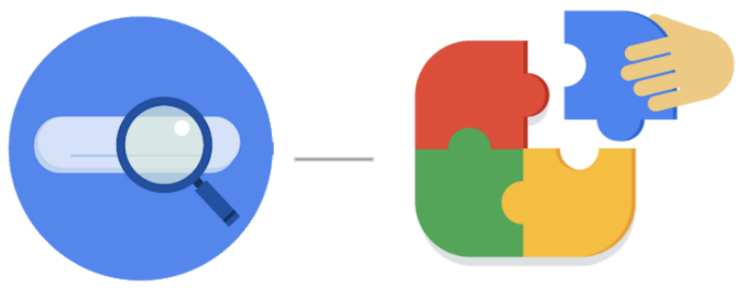
Business scenario	How proxy data can be used
A new car model was just launched a few days ago and the auto dealership can't wait until the end of the month for sales data to come in. They want sales projections now.	The analyst proxies the number of clicks to the car specifications on the dealership's website as an estimate of potential sales at the dealership.
A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years.	The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years.
The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren't publicly available yet.	The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier.

Open (public) datasets

If you are part of a large organization, you might have access to lots of sources of data. But if you are looking for something specific or a little outside your line of business, you can also make use of open or public datasets. (You can refer to this Towards Data Science article for a brief explanation of the difference between open and public data.)

Here's an example. A nasal version of a vaccine was recently made available. A clinic wants to know what to expect for contraindications, but just started collecting first-party data from its patients. A **contraindication** is a condition that may cause a patient not to take a vaccine due to the harm it would cause them if taken. To estimate the number of possible contraindications, a data analyst proxies an open dataset from a trial of the injection version of the vaccine. The analyst selects a subset of the data with patient profiles most closely matching the makeup of the patients at the clinic.

There are plenty of ways to share and collaborate on data within a community. Kaggle (kaggle.com) which we previously introduced, has datasets in a variety of formats including the most basic type, Comma Separated Values (CSV) files.



CSV, JSON, SQLite, and BigQuery datasets

- CSV: Check out this Credit card customers dataset, which has information from 10,000 customers including age, salary, marital status, credit card limit, credit card category, etc. (CC0: Public Domain, Sakshi Goyal).
- JSON: Check out this JSON dataset for trending YouTube videos (CC0: Public Domain, Mitchell J).
- SQLite: Check out this SQLite dataset for 24 years worth of U.S. wildfire data (CC0: Public Domain, Rachael Tatman).
- BigQuery: Check out this Google Analytics 360 sample dataset from the Google Merchandise Store (CC0 Public Domain, Google BigQuery).

Refer to the Kaggle documentation for datasets for more information and search for and explore datasets on your own at kaggle.com/datasets.

As with all other kinds of datasets, be on the lookout for duplicate data and 'Null' in open datasets. Null most often means that a data field was unassigned (left empty), but sometimes Null can be interpreted as the value, 0. It is important to understand how Null was used before you start analyzing a dataset with Null data.

Mark as completed

Like **Dislike** **Report an issue**