Supervised learning is fairly simple: We use it when we know what we're looking for or what our output should be. Two of the most popular techniques are classification and regression. For example, if we want to see how Bitcoin performs over time, or if we want to figure out what kinds of coins a user is most likely to buy, we use supervised learning. However, we use unsupervised learning when we don't yet know the question we're asking of the data. In other words, we just want to figure out if there is anything at all the data can tell us.

**REWIND**

In supervised learning, first a model is initiated, or a template for the algorithm is created. Then it will analyze the data and attempt to learn patterns, which is also called fitting and training. After the data has been fit and trained, it will then make predictions.

In supervised learning, the input data already has a paired outcome, which is plugged in to train the model to predict outcomes in new datasets. For example, we want to build a model that, when given unfamiliar data, can accurately predict the outcomes.

In **unsupervised learning**, there are two key differences from the above approach:

* There are no paired inputs and outcomes.
* The model uses a whole dataset as input.

Unsupervised learning is used in one of the following two ways:

* Transform the data to create an intuitive representation for analysis or to use in another machine learning model; or
* Cluster or determine patterns in a grouping of data, rather than to predict a classification.

For example, imagine a store owner is planning to stock up on back-to-school supplies. The owner's challenges are selecting from a wide range of school supplies, devoting considerable space to the inventory, and maintaining a fast inventory turnover rate, from stocking to selling.

The owner wants to predict the best time to start selling school supplies so staff can make room for the next big items to sell. The owner decides to focus on the previous year's data, which includes the number of items sold, when they were sold, and the area schools' start dates for 10, 20, and 30 days prior to the items' sell dates.

Now, imagine the owner notices that when customers come in to buy school supplies, they also purchase non-school items, giving other categories a bump in sales. There are a wide variety of products to offer different consumers, so the owner wants to know which consumers might buy more products and what items they might want to buy. We don't really know how to group the data, but we know that grouping the data will be instructive. You'll need to factor in many data points, such as age, race, gender, location, and much more.

**Challenges of Unsupervised Learning**

**IMPORTANT**

Unsupervised learning isn't the solution for every data analytic challenge. Just because supervised learning might not work for one situation doesn't mean unsupervised learning will work instead. Understanding the data and what can be done with it is an important first step before choosing an algorithm.

Recall that unsupervised learning does not take in any pairing of input and outcomes from the data—it only looks at the data as a whole. This can cause some challenges when running the algorithm. Since we won't know the outcome it's predicting, we might not know that the result is correct.

This can lead to issues where we're trying to decide if the model has provided any helpful information that we can use to make decisions in the real world. For example, our store owner might run a model that ends up grouping the type of people by how much they're buying. This could be useful in some contexts—for example, knowing who the top spenders are—but it might not help the store owner better organize the store for maximum purchases per person, or understand the differences in product preferences between top purchasers.

The only way to determine what an unsupervised algorithm did with the data is to go through it manually or create visualizations. Since there will be a manual aspect, unsupervised learning is great for when you want to explore the data. Sometimes you'll use the information provided to you by the unsupervised algorithm to transition to a more targeted, supervised model.

As with supervised learning, data should be preprocessed into a correct format with only numerical values, null value determination, and so forth. The only difference is unsupervised learning doesn't have a target variable—it only has input features that will be used to find patterns in the data. It's important to carefully select features that could help to find those patterns or create groups.

The next section will cover data preprocessing and data munging, and provide a refresher on Pandas and data cleaning. First, you'll need to install the necessary libraries for practice.