18.2.1

## Steps for Preparing Data

**After** digging into unsupervised learning a bit, you realize that your first step in convincing Accountability Accountants to invest in cryptocurrency is to preprocess the data.

You and Martha open up the dataset to get started preprocessing it. Together, you will want to manage unnecessary columns, rows with null values, and mixed data types before turning your algorithm loose.

# Data Selection

Before moving data to our unsupervised algorithms, complete the following steps for preparing data:

1. Data selection
2. Data processing
3. Data transformation

Data selection entails making good choices about which data will be used. Consider what data is available, what data is missing, and what data can be removed. For example, say we have a dataset on city weather that consists of temperature, population, latitude and longitude, date, snowfall, and income. After looking through the columns, we can readily see that population and income data don't affect weather. We might also notice some rows are missing temperature data. In the data selection process, we would remove the population and income columns as well as any rows that don't record temperatures.

# Data Processing

Data processing involves organizing the data by formatting, cleaning, and sampling it. In our dataset on city weather, if the date column has two different formats—mm-dd-yyyy (e.g., 01-23-1980) and month-data-year (e.g., jan-23-1980)—we would convert all dates to the same format.

## Data Transformation

Data transformation entails transforming our data into a simpler format for storage and future use, such as a CSV, spreadsheet, or database file. Once our weather data is cleaned and processed, we would export the final version of the data as a CSV file for future analysis.