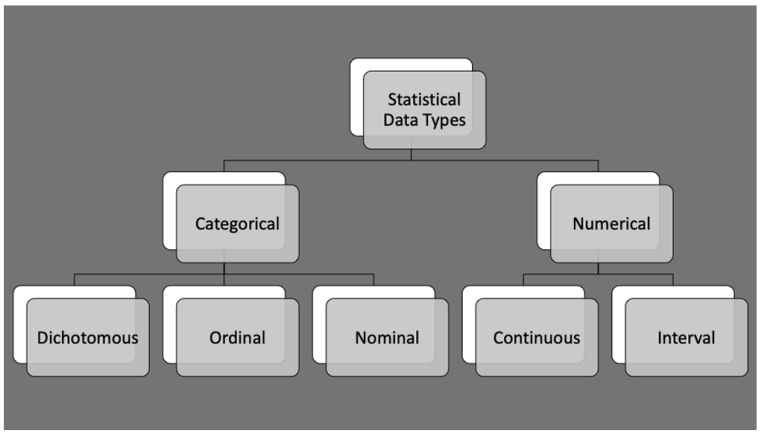
15.4.2

## Identify Different Data Types

**Now** that Jeremy feels comfortable with R, he's excited to jump into statistics. Colleen has been wanting to brush up on her statistics knowledge as well, so she suggests that their new team start an internal study group. Their first session is about different data types.

There are two major data types in statistics: categorical and numerical. Within these types are several subtypes, each with its own use cases. In this section, we'll describe these types and explain how to analyze data effectively. First, take a look at the following diagram to visualize how data is categorized.




# Categorical Data

**Categorical data** represents data characteristics or qualitative descriptions. Generally, categorical data is any data that is not measured, also known as qualitative data. Categorical data can be collected in the form of strings, true/false Boolean values, or even encoded numbers as categories (such as one for red, two for blue, three for green, etc.). Several statistical tests use categorical data to inform which groups to compare. Categorical data has three subtypes: dichotomous, ordinal, and nominal

## Dichotomous Data

**Dichotomous data** is collected from either one of two categories. For example, an online survey might collect member/non-member or demographic information. Dichotomous data can be collected in the form of true/false Boolean values, 0 or 1 binary values, or two strings. Later in the module, we'll use dichotomous data to help perform many of our comparative statistical tests.

## Ordinal Data

**Ordinal data** has a ranked order. Although ordinal data has a sequence, we don't necessarily know the value between each ordinal data point. Data that is collected on a value scale (e.g., movie rankings, survey results, and the **Likert scale (https://en.wikipedia.org/wiki/Likert_scale)**) are common forms of ordinal data. Ordinal data combines the qualitative properties of labels to the quantitative properties of scale to allow for comparative analyses. Ordinal data is very popular with research and survey groups because it allows for quantitative analysis without the need of machinery and tools to obtain measurements.

NOTE

There are statistical tests such as the Mann-Whitney U test and the Kruskal-Wallis H test that compare ordinal datasets. These statistical tests are more advanced versions of basic comparative statistics tests and are outside the scope of this course. However, once you master the basics of statistical testing, it is not difficult to apply more advanced statistical models based on your specific data needs. Remember—Google is your best friend!

## Nominal Data

**Nominal data** is data used as labels or names for other measures. Nominal data can be as individual as an identification number or can be as general as a list of three options. Unlike ordinal data, nominal data has no ranking. Therefore, nominal data is often used with a more quantitative data type to perform an analysis. Often nominal data will be transformed using a grouping function to decrease the complexity of the data.

## Numerical Data

Typically, **numerical data** is obtained by taking a measurement from an instrument (such as a ruler, measuring scale, sensor, etc.) or by counting. In statistics, numerical data is used to perform quantitative analysis that can produce the

probability of an outcome or quantify the relationship between categories. Within numerical data there are two primary data types to consider: continuous and interval.

---

## Continuous Data

**Continuous data** can be subdivided infinitely. For example, if you want to describe the thickness of window glass, you could measure it in x number of centimeters, millimeters, nanometers, picometers, and so on. Continuous data is typically recorded with decimal places to match the precision of the measurement. Almost all statistical tests and models use continuous data to generate precise results.

---

## Interval Data

**Interval data** is spaced out evenly on a scale. Also known as integer data, interval data does not use decimal places and can't be subdivided. Interval data also can't be multiplied or divided. Because interval data is spaced out evenly, it can be grouped together or bucketed easily. For example, a set of integers 15, 4, 18, 10, 3, and 5 could be collected as a group that is less than 20. Due to this property, interval data can be treated as a numerical data type or transformed into a nominal data type.

Additionally, interval data can be generated through rounding continuous data at the cost of losing precision of the measurement. Therefore, interval data can be used by most statistical models as either a quantitative or qualitative variable, depending on the use case.

Now that we understand the different statistical data types and how to identify each by its characteristics, we should be able to classify any tabular dataset.

---

## Getting Oriented with Data

The easiest means of orienting ourselves on a new dataset in R is to use the `head()` function, which shows us the first few rows of our data frame. At any point when looking at the first few rows, we can use bracket notation (or the `$` operator) to select an individual column to explore.

Alternatively, if we're using RStudio, we can explore any data frame by clicking on it in our environment pane. By navigating through each column and classifying each data type, we can determine which columns provide measurement results, and which columns provide characteristics about our subjects.

If we're fortunate to have context provided for a given dataset via documentation or from the data collector, we should be able to identify columns and metrics of interest. However, we have not finished characterizing our data just yet—we still need to understand how values in our data are distributed.

Retake