TOOL

Analysis and Variable Types

When you answer a question using your data, you should start by identifying the type of question you are asking, then identify the type of variable(s) with which you are working. Considering this information will help you decide which kind of summarization and visualization tools will best help you answer your question.

Use this tool to help you determine:

1. **Analysis type**
   Are you answering a univariate, bivariate, or multivariate question?

2. **Variable type**
   Are you working with categorical variables, numerical variables, or both?

The examples below are based on the `titanic` data set you examined in this course. You can use the `titanic` data set to ask many different questions that have varying levels of complexity.

Data Set Information

The `titanic` data set contains demographic information of different passengers on the RMS Titanic, which sank in the Atlantic Ocean in 1912. The `titanic` data set has data on each passenger in the rows and on passenger characteristics in the columns:

	Name	PClass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29.00	female	Yes
2	Allison, Miss Helen Loraine	1st	2.00	female	No
3	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	No
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	No
5	Allison, Master Hudson Trevor	1st	0.92	male	Yes
6	Anderson, Mr Harry	1st	47.00	male	Yes

Analysis Type

Univariate

A question that only requires you to work with one variable (column) in your data set. Some examples of univariate questions you could ask about the `titanic` data set are:

- What proportion of passengers survived the accident?
- What proportion of passengers were male?

Bivariate

A question that requires you to work with two variables *simultaneously*. This means that you need to know the two variables simultaneously for each observation in the data. Bivariate questions are typically designed to find some form of association between two variables. For example, when one variable changes, do we expect the other variable to change as well? Some bivariate questions you could ask about the `titanic` data set are:

- What is the association between a passenger's sex and their survival when the Titanic sank? To answer this question, you need to know the variables `Sex` and `Survived` for each passenger. Answering this question can tell you, for example, the proportion of female passengers that survived the sinking of the Titanic.
- What is the association between a passenger's class and their survival when the Titanic sank? To answer this question, you need to know the variables `PClass` and `Survived` for each passenger. Answering this question can tell you, for example, the proportion of first-class passengers that survived.

Multivariate

A question that requires you to work with more than two variables *simultaneously*. Multivariate questions can take many forms. In this course, we focus on a particular style of multivariate question that often comes up when you study the association between two variables. In this style of multivariate question, you see that Variables *X* and *Y* are associated in a bivariate analysis. You then want to know if a third categorical variable, *Z*, affects the nature or strength of the association. To check this, you can focus on different subgroups of the data where the value of *Z* is the same and answer the bivariate question separately for each subgroup of *Z*.

A multivariate question of this type that you could ask about the **titanic** data set is:

- Does passenger class influence the association between passenger survival and passenger sex?
  To answer this question, you need to know the variables **Sex, Survived,** and **PClass** for each passenger. Answering this question can tell you, for example, whether the difference in the proportion of male and female passengers that survived in first class varied from the difference in the proportion of male and female passengers that survived in the third class.

If you see that the bivariate association differs across subgroups, you suspect that the association between *X* and *Y* is not causal in nature. Instead, what looks like an association between *X* and *Y* is explained by the confounding variable *Z*. In the previous example, if you see that the relationship between **Survived** and **Sex** differs depending on **PClass**, then **PClass** is a confounding variable that influences the relationship between **Survived** and **Sex**.

Variable Type

Categorical

A variable that takes a small number of distinct values that represent different categories. These categories often describe levels or groupings of a characteristic, such as year in school or country of origin.

Some examples of categorical variables in the **titanic** data set include:

- **Sex**: In the **titanic** data set, this can be either male or female.
- **PClass**: In the **titanic** data set, this can be first, second, or third.

Different levels of a categorical variable cannot be added or subtracted like numbers; for example, adding second class to third class does not equal fifth class.

Numerical

A variable that is made up of numerical values, either from a fixed set of numbers or from a range of real numbers, such as time spent doing an activity. Unlike categorical variables, you can perform mathematical operations on these variables. For example, you can add or subtract different values of a continuous variable or calculate their average.

Age is the one numerical variable in the **titanic** data set.