Interpreting Aggregates: Populations and Samples

When you compute aggregates of columns in a real-world dataset, you need to be mindful of whether the dataset describes a *population* or a *sample*. There are implications for your conclusions and how you might want to phrase them.

The Difference Between Populations and Samples

Some datasets have a row describing each and every item in a *population*. An example of this is a table of customers, which has one row for each customer. In a dataset like this, every customer is described in the data. The *population* is all the business's customers, so the dataset contains the full population.

Other datasets describe a *sample* from a larger population. An example of this is data collected from a survey or poll. Typically, the respondents to a survey or poll comprise only a small proportion of the population they are drawn from. For example, if you took a survey of people in Brazil, you might have 1000 survey respondents (the sample), but you'll use those 1000 respondents to represent the 200 million people who live in Brazil (the population).

In many cases, it is impractical or impossible to collect data describing an entire population. This is true even when the population is fairly small. For example, a company that manufactures airplane wings needs to perform tests to determine the amount of force that the wings can endure before breaking. They cannot test every wing they manufacture, because testing a wing requires breaking it. So they must test a small sample of the wings, and they must take care to ensure that the sample is representative of the full population.

None of the datasets on the VM describe samples, but many real-world datasets do.

Phrasing Conclusions Appropriately

When you analyze a dataset that describes a sample, it's important to phrase your observations appropriately. In particular, you should make it clear whether a sample or a population is involved, and you should describe that sample or population accurately so the results are not overgeneralized.

Example Scenario

Imagine you were querying a dataset that described 300 responses to a survey of North American air travelers. Your query results showed that 79% of the 300 respondents would prefer traveling by high-speed rail over traveling by plane.

Inappropriate Conclusions

The following conclusion is inappropriate because it conflates the sample with the population:

- "79% of North American air travelers prefer traveling by high-speed rail over traveling by plane."

The following conclusion is even worse; it does not describe the population correctly and it overgeneralizes the results:

- "79% of North Americans prefer traveling by high-speed rail over traveling by plane."

Appropriate Conclusions

The following conclusion is appropriate:

- "79% of respondents to a survey of North American air travelers said they prefer traveling by high-speed rail over traveling by plane."

The following conclusion is appropriate if you are confident that the sample is representative of the population:

- “Our survey results suggest that 79% of North American air travelers prefer traveling by high-speed rail over traveling by plane.”

Aggregate Functions for Samples

Depending on what SQL engine you’re using, you might have noticed some built-in aggregate functions with “_samp” or “_pop” in their names. For purposes of this course, you should disregard these functions. They are described in a later course in this specialization.