# Data Selection Directions to Instructors

The **learning objectives** for the Data Selection activity are:

1. Learn how to classify data into qualitative and quantitative and their different subcategories.
2. Select discrete data to be used in 5-point summary and correlational analysis
3. Learn how to navigate the database https://www.gapminder.org/tools/

The first step of this module is to guide students through finding adequate data. While navigating through the Data selection mini-lecture and worksheet, make sure you explore the different possibilities offered by the bubble graphs at <https://www.gapminder.org/tools/>. Note that students can select most of the Health folder data in Gapminder, with a few exceptions of data that they should avoid (data that are already presented as average). Here are examples of data that are **averages in the HEALTH folder and we are NOT interested in**:

* Bad teeth per child
* Body mass index (BMI)
* Fat in blood
* Food supply
* Babies per woman
* Health spending
* Life expectancy

In the second step of this module we will learn how to calculate a 5-point summary, where you need to find the median, quartiles, maximum and minimum. Later we will draw a hypothesis for correlation of 2 selected datasets and perform a correlational analysis. **Quantitative discrete datasets** are perfect for these types of exploration. Let’s understand why discrete datasets can be used in five-point summaries and correlational analysis and are useful in our activity:

1. In the Gapminder database, there are multiple datasets that are “headcounts,” which are discrete data (the type of data we are interested in!).
2. In the Gapminder database, discrete data are not average numbers (so we can use discrete data to calculate the median of raw numbers).
3. These aforementioned pieces of data can be ratio or interval types (positive or negative numbers)

For a deeper understanding and in case students ask, below are the reasons why we are not going to include continuous and qualitative data in our range of possibilities:

1. Continuous data: In the Gapminder database, continuous data are calculated averages, so they would not be interesting to us. The reason for this is that in the 5-point summary, we would have to calculate the median of these averages, which is a mathematically incorrect way to treat numbers.

1. Qualitative data: none of the subtypes of qualitative data will be part of our choice of data, since we cannot calculate the mean and correlate data such as colors of candies or small, medium and large candies.

After the Data Selection section, you can proceed to the 5-point summary & box and whisker plot section.

This material is based upon work supported by the National Science Foundation under Grant No.1919613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.