Statistics for Machine Learning

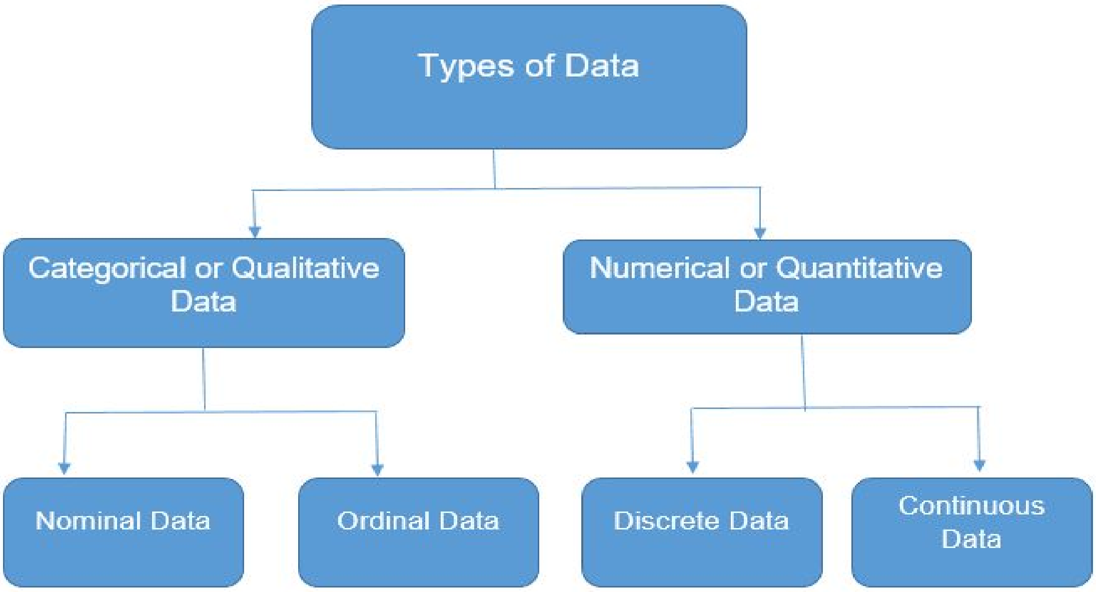
What is Statistics?

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.

What is Data?

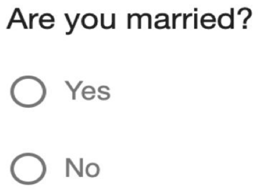
Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

Types of Data:



1. **Categorical or Qualitative Data:**
   Categorical data represents characteristics. Therefore it can represent things like a person's gender, language,age group etc.

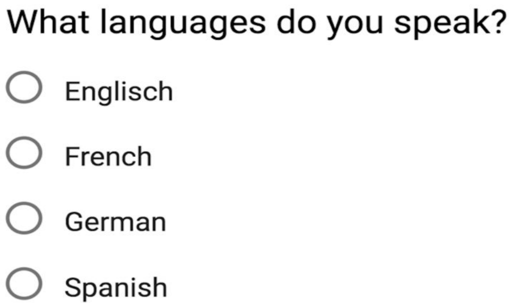
   - **Nominal Data:**
     Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as **labels**. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features below:



Are you married?

○ Yes

○ No



What languages do you speak?

○ Englisch

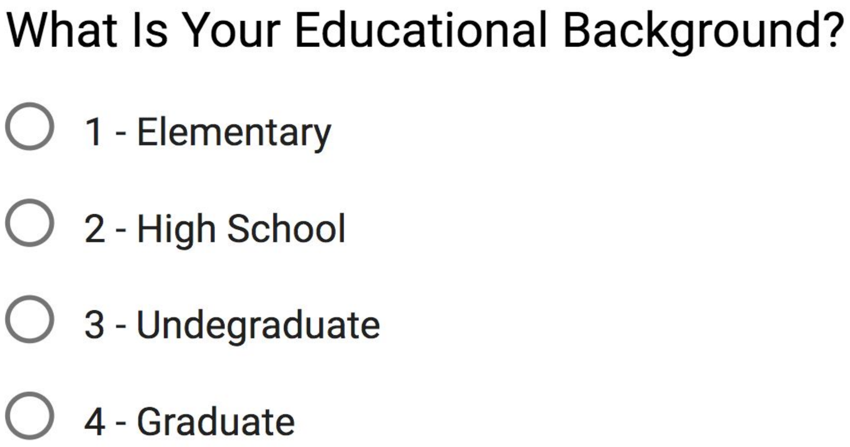
○ French

○ German

○ Spanish

The left feature that describes if a person is married would be called **dichotomous**, which is a type of nominal scales that contains only two categories.

- **Ordinal Data:**
  Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:



What Is Your Educational Background?

○ 1 - Elementary

○ 2 - High School

○ 3 - Undegraduate

○ 4 - Graduate

Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

2. **Numerical or Quantitative Data:**

- **Discrete Data:**
  We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data can't be measured but it can be counted. It basically

represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.

- **Continuous Data:**
  Continuous Data represents measurements and therefore their values can't be counted but they can be measured. An example would be the height of a person, which you can describe by using intervals on the real number line.

Types of Statistics

Statistics are broken into two categories **Descriptive statistics** and **Inferential statistics**.