Data Types

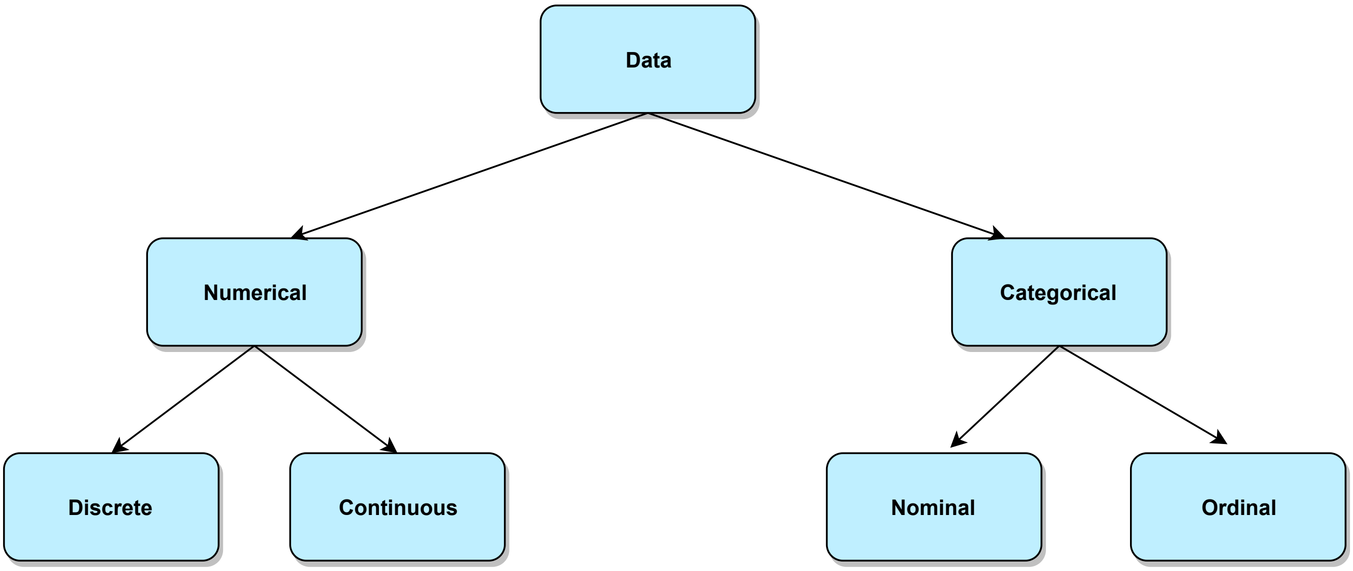
This lesson will introduce the learner to the different data types in which data can exist.

WE'LL COVER THE FOLLOWING

Introduction to Data types

Data is basically raw information. It can be in any form. We will look at data from the perspective of a data scientist who is going to clean and analyze the data. We need to know in what form the data is present to analyze it properly and apply different statistical methods on it. To a data scientist, data can take two basic forms:

- **Numerical**
- **Categorical**



Numerical data

Numerical data is data that has some meaning as a measurement, such as the height of a person, the price of a product, the IQ of a person, the number of lessons in this course, etc. It is also known as *Quantitative Data*. It can be broken down into two types.

1. Discrete data

Discrete data is data that can take separate and distinct values. It can take only a certain number of values. It cannot be divided into smaller meaningful parts. For instance, the number of heads in 100 tosses of a coin flip, the number of students in a classroom, the number of cars in a showroom, etc.

2. Continuous data

Continuous data cannot be counted, but it can be measured. It represents measurements. It includes quantities that do not have an end to them such as money, the height of a person, the amount of rainfall, the speed of a car, etc. It can be divided into further meaningful parts.

We can use statistical methods such as mean, median, quartiles, Box plots, and Histograms to describe numerical data.

Example of box plot and histogram

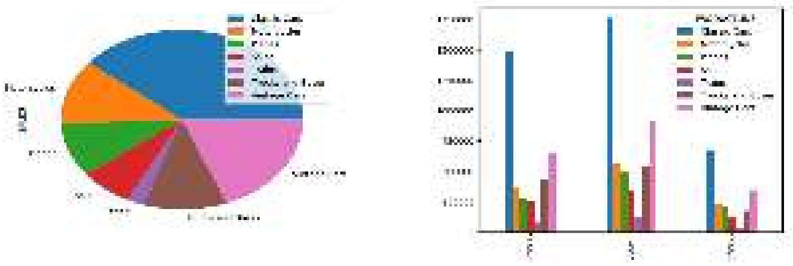
Categorical data

Categorical data as the name suggests, represent categories or characteristics such as gender, language, level of education, marital status, the genre of a movie, etc. It is also known as *Qualitative Data*. We can associate numerical values with categorical data, but they would not have any mathematical meaning, e.g., 0/1 for male/female.

1. Ordinal

Ordinal data is categorical data that has a sense of order to it. For instance, the happiness level of a customer, level of education, or rating of a movie on a scale of $0 - 5$.

We can summarize ordinal data with percentiles, frequencies, median, mean, etc. For visualization, we can use pie charts and bar charts.



Example of pie chart and bar plot

2. Nominal data

Nominal data is categorical data that has no order. It can be thought of as *labels*. For instance, the gender of a person as male orfemale, the language a person speaks, etc. Nominal Data can be dealt with using frequencies, proportions, pie charts, bar plots, etc.

Figuring out Data types in Pandas

In pandas, it is very easy to figure out the data types of the variables. We can use the `info` function on our dataframe.

```
import pandas as pd
df = pd.read_csv('sales_data.csv')
# print information on data types of each column
print(df.info())
```

The output of `df.info` tells us the count and type of each column that we have in our dataset. `int64` means it is an integer, which means it is a discrete or ordinal variable. `float64` means it is a number with a fractional part, therefore, a continuous variable. We can infer from `object` that it will be a nominal variable.

However, sometimes looking at the output of `df.info` is not enough. There are instances when nominal data is written in numbers, such as *gender* is written as $1/2$ or $0/1$ instead of “male” and “female”. So, we have to look out for those cases as well.

Now that we know how our data is available to us, in the next lesson, we will dive into analyzing individual variables to get insights.