Machine learning

Standard deviation 🡪 is a number that describes how spread out the values are

High standard deviation: most of the numbers are far from the mean

Low standard deviation 🡪 most numbers are close to the mean

Variance is another number that indicates how spread out the values are

If you take the square root of the variance you get the standard deviation

Standard Deviation is often represented by the symbol Sigma: σ

Variance is often represented by the symbol Sigma Square: σ2

What are Percentiles?

Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.

Example: Let's say we have an array of the ages of all the people that lives in a street.

What is the 75. percentile? The answer is 43, meaning that 75% of the people are 43 or younger.

## Data Distribution

Earlier in this tutorial we have worked with very small amounts of data in our examples, just to understand the different concepts.

In the real world, the data sets are much bigger, but it can be difficult to gather real world data, at least at an early stage of a project.

### **How Can we Get Big Data Sets?**

To create big data sets for testing, we use the Python module NumPy, which comes with a number of methods to create random data sets, of any size.