

What is Scaling Data?



	feature0	feature1	feature2	feature3
value0	100000	23423432	6782345	0.00002
value1	200000	3243223	460000	0.00001
value2	30000032	23542323	4567000	0.0000163
value3	400000	35463242	23452000	0.00001678
value4	500000	598299	54740000	0.000568

**Do you notice an issue
with this dataset?**

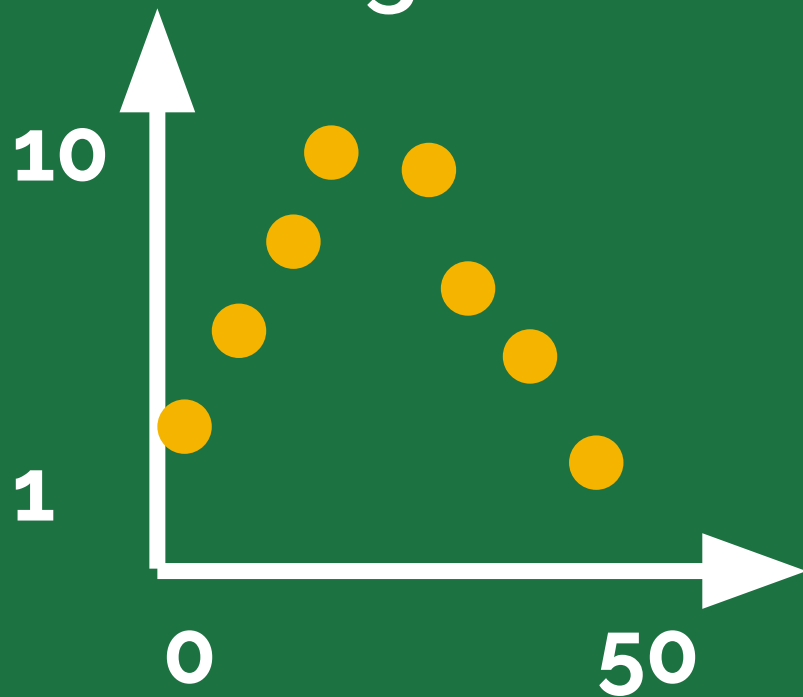
	feature0	feature1	feature2	feature3
value0	100000	23423432	6782345	0.00002
value1	200000	3243223	460000	0.00001
value2	300000032	23542323	4567000	0.0000163
value3	400000	35463242	23452000	0.00001678
value4	500000	598299	54740000	0.000568

- If features in a dataset have values on different scales, this can cause problems

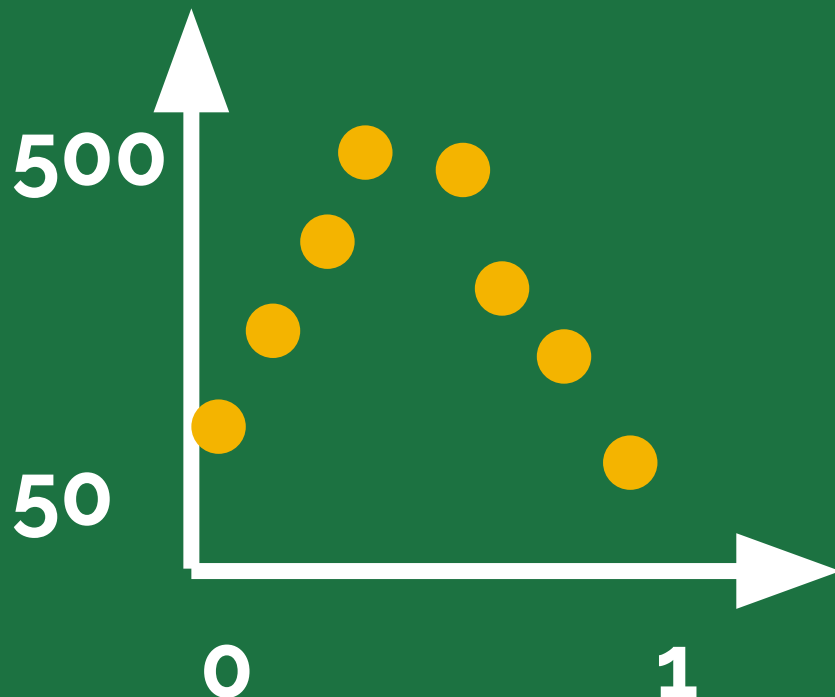
- **Scaling changes the range of values**
- **The shape of distribution stays the same!**
- **For example, a scale model of a city has the same proportions**

	feature0	feature1	feature2	feature3
value0	10000000	23423432	67823450	2000000
value1	20000000	32432230	4600000	1000000
value2	30000032	23542323	45670000	1630000
value3	4000000	35463242	23452000	1678000
value4	5000000	5982990	54740000	5680000

original



scaled



SciKit-Learn provides MinMaxScaler!



- The default range returned by `MinMaxScaler` is 0 to 1.
- The relative spaces between each feature's values are saved

- **MinMaxScaler doesn't reduce the importance of outliers!**
- **Use other scalers if you want a normal distribution or for outliers to have less influence**

**What is Scaling
Data? ✓**

**Let's
begin...**