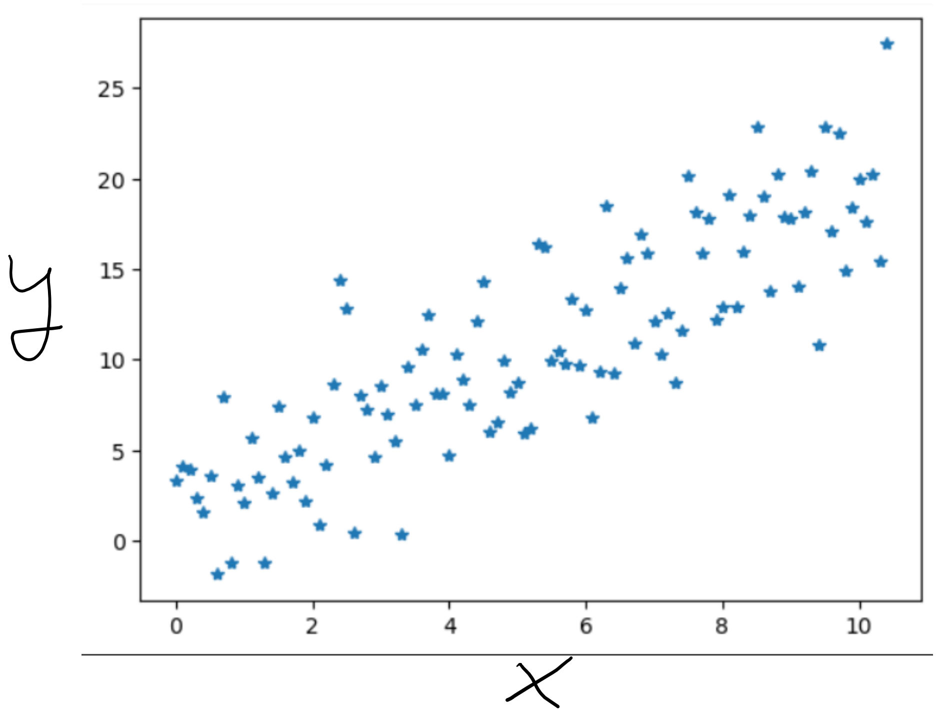
LINEAR VS NONLINEAR MODELS

What is a model?

A model is a mathematical representation of a simplified aspect of reality, that allows us to understand, measure and predict quantities of interest.

Given certain inputs, the model serves to calculate or estimate the output.

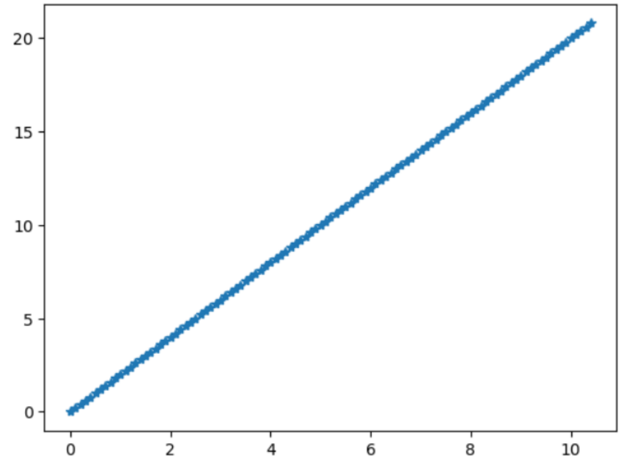


In this example the model that generated the points on the graph was:

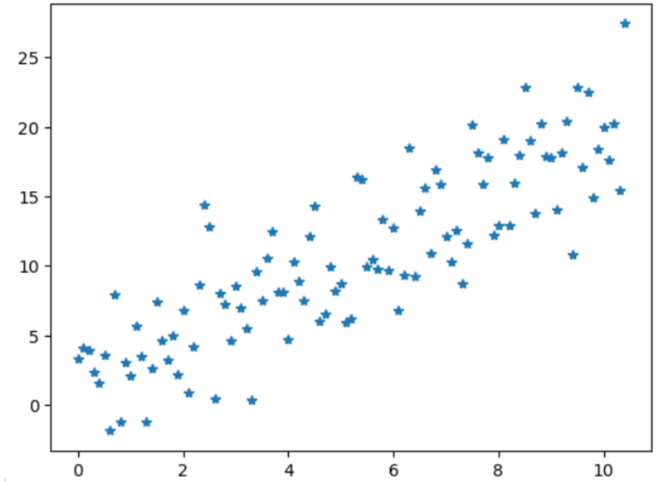
$$y = 2x + (noise)$$

Typically, we have some data (inputs and outputs) and we want to find the model that 'better fits' my real data, hoping that the model will also be good in the future, so I can use it as a predictor.

Deterministic vs. statistical (noisy) model



Deterministic model: Knowing the inputs, we can calculate **exactly** the output.
- Easy to use, but often it simplifies reality.



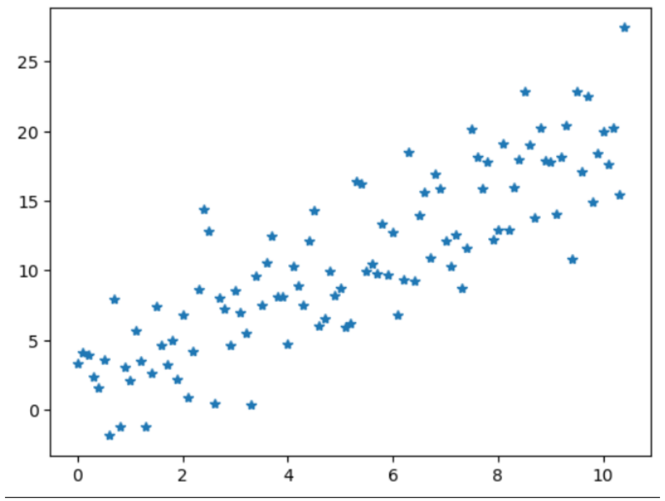
Statistical model: Knowing the inputs, we can only **estimate** a range, or a distribution, in which we can find the output.

The noise in the model usually encodes the information that we do not know (variables that affect the output but that are not in the model).

Correlation vs. causation

Correlation DOES NOT imply causation.

Again, using the previous model, $y = 2x + (noise)$
Consider that *x* represent weekly business sales (in thousands of dollars) for camping business and *y* represents the average temperature of the week.



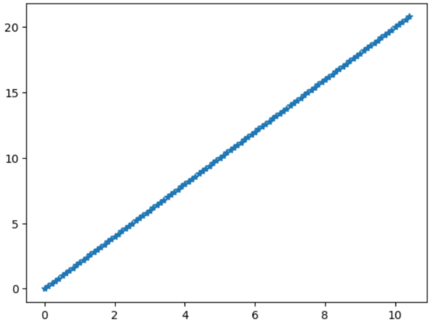
The model tells us (at least graphically) that both quantities are correlated, but high temperatures are not caused by higher sales.

We can be tempted to say that there is causation in the other direction. That is, higher temperatures can explain that more people go camping, but this is not something that we know from the model alone.

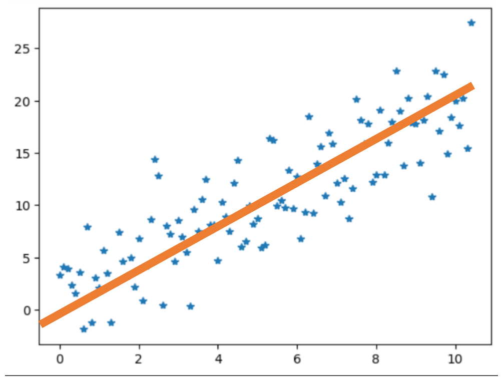
Linear vs. Nonlinear?

Linear model: We say that the model is linear if the graph is a straight line. When the model is noisy, a linear model refers to one in which the average predicted value (the mean of the distribution for a given input value) is a straight line.

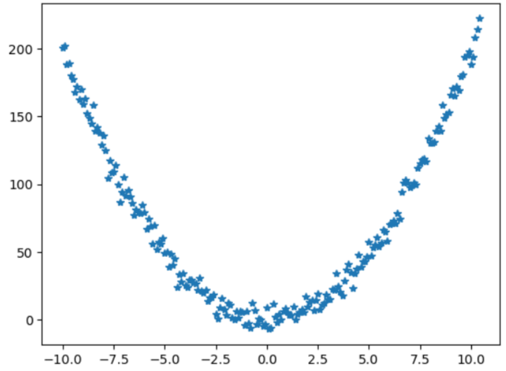
Nonlinear model: any model which does not satisfy the previous conditions.



Linear deterministic



Linear noisy



Nonlinear noisy

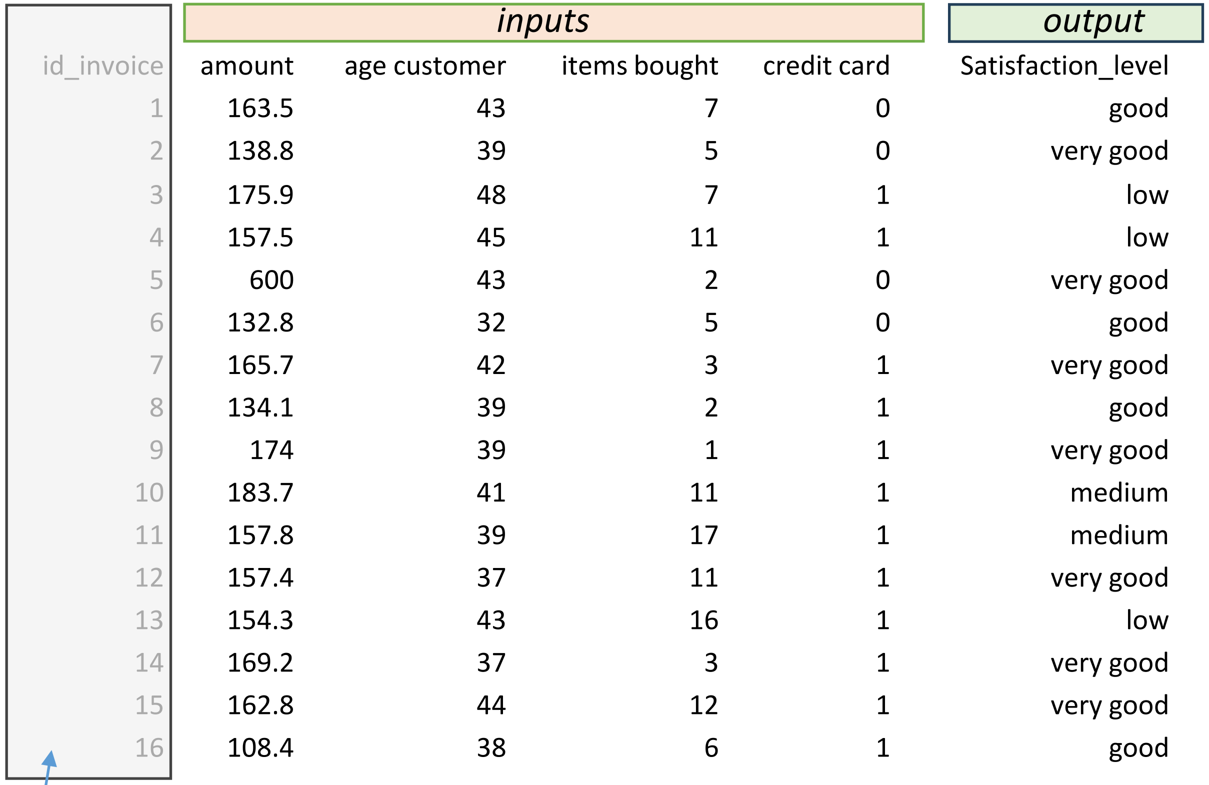
Univariate vs. Multivariate

Univariate model : The input consist in only one variable

Multivariate model: The input consists in more than 1 variable. For instance, consider a dataset with many variables like the following one.

We could build a model that predicts the Satisfaction level, based on the variables amount, age, items bought and credit card used.

Some models used in business have hundreds of inputs



	inputs				*output*
id_invoice	amount	age customer	items bought	credit card	Satisfaction_level
1	163.5	43	7	0	good
2	138.8	39	5	0	very good
3	175.9	48	7	1	low
4	157.5	45	11	1	low
5	600	43	2	0	very good
6	132.8	32	5	0	good
7	165.7	42	3	1	very good
8	134.1	39	2	1	good
9	174	39	1	1	very good
10	183.7	41	11	1	medium
11	157.8	39	17	1	medium
12	157.4	37	11	1	very good
13	154.3	43	16	1	low
14	169.2	37	3	1	very good
15	162.8	44	12	1	very good
16	108.4	38	6	1	good

No information in this variable

Univariate vs. Multivariate

For the same dataset, we can decide to make a completely different predictive model, depending on what we want to predict or understand.

Inputs and outputs depend on what we want to analyse.

	output	*inputs*			
id_invoice	amount	age customer	items bought	credit card	Satisfaction_level
1	163.5	43	7	0	good
2	138.8	39	5	0	very good
3	175.9	48	7	1	low
4	157.5	45	11	1	low
5	600	43	2	0	very good
6	132.8	32	5	0	good
7	165.7	42	3	1	very good
8	134.1	39	2	1	good
9	174	39	1	1	very good
10	183.7	41	11	1	medium
11	157.8	39	17	1	medium
12	157.4	37	11	1	very good
13	154.3	43	16	1	low
14	169.2	37	3	1	very good
15	162.8	44	12	1	very good
16	108.4	38	6	1	good