Scatter Diagrams and Correlation

Gaston Sanchez

So far …

- Frequency Tables
- Summary measures
- Histograms & Barcharts

descriptive statistics
for one single variable

Two variables

X	Y
quantitative	quantitative
quantitative	qualitative
qualitative	qualitative

Interested in the association of two quantitative variables

Two quantitative variables

X	Y
Height	Weight
High School GPA	Undergrad GPA
Yrs of Study	Income
Area Size	House Price
Price of gas	# car accidents

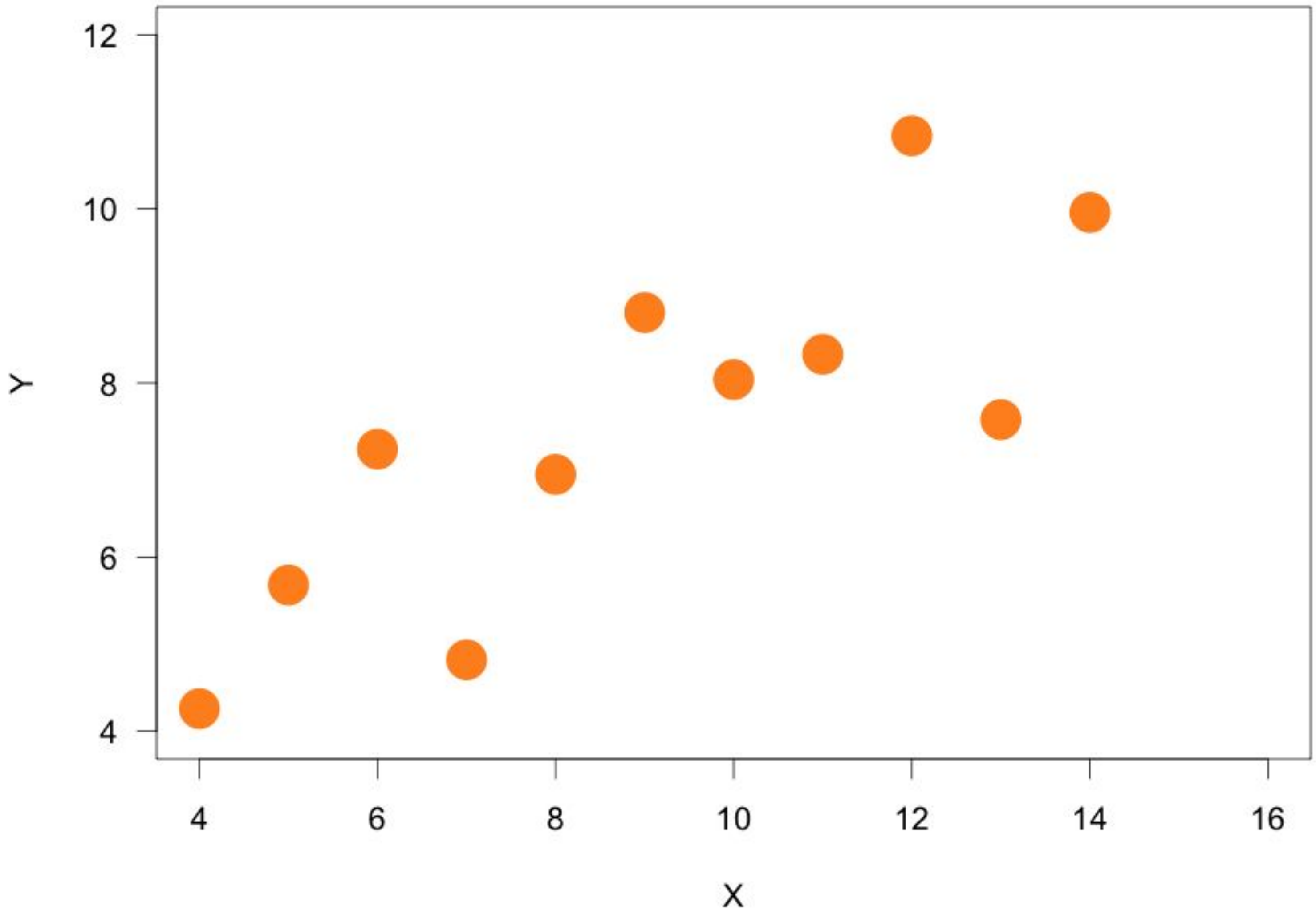
Study the relationship between X and Y

Scatter Diagrams

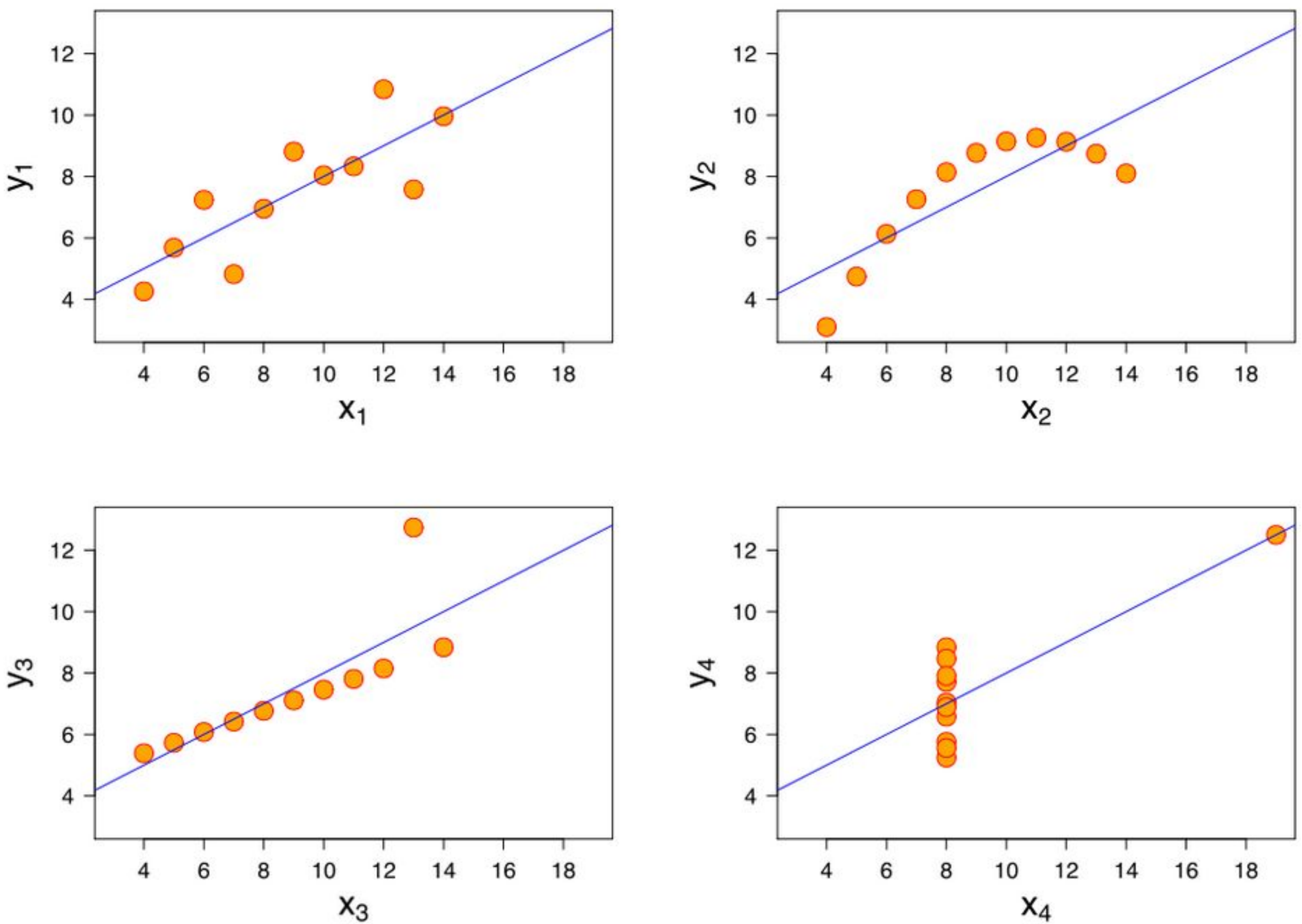
X	Y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

How are **X** and **Y** related?

Scatter Diagram



dataset 1		dataset 2		dataset 3		dataset 4	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

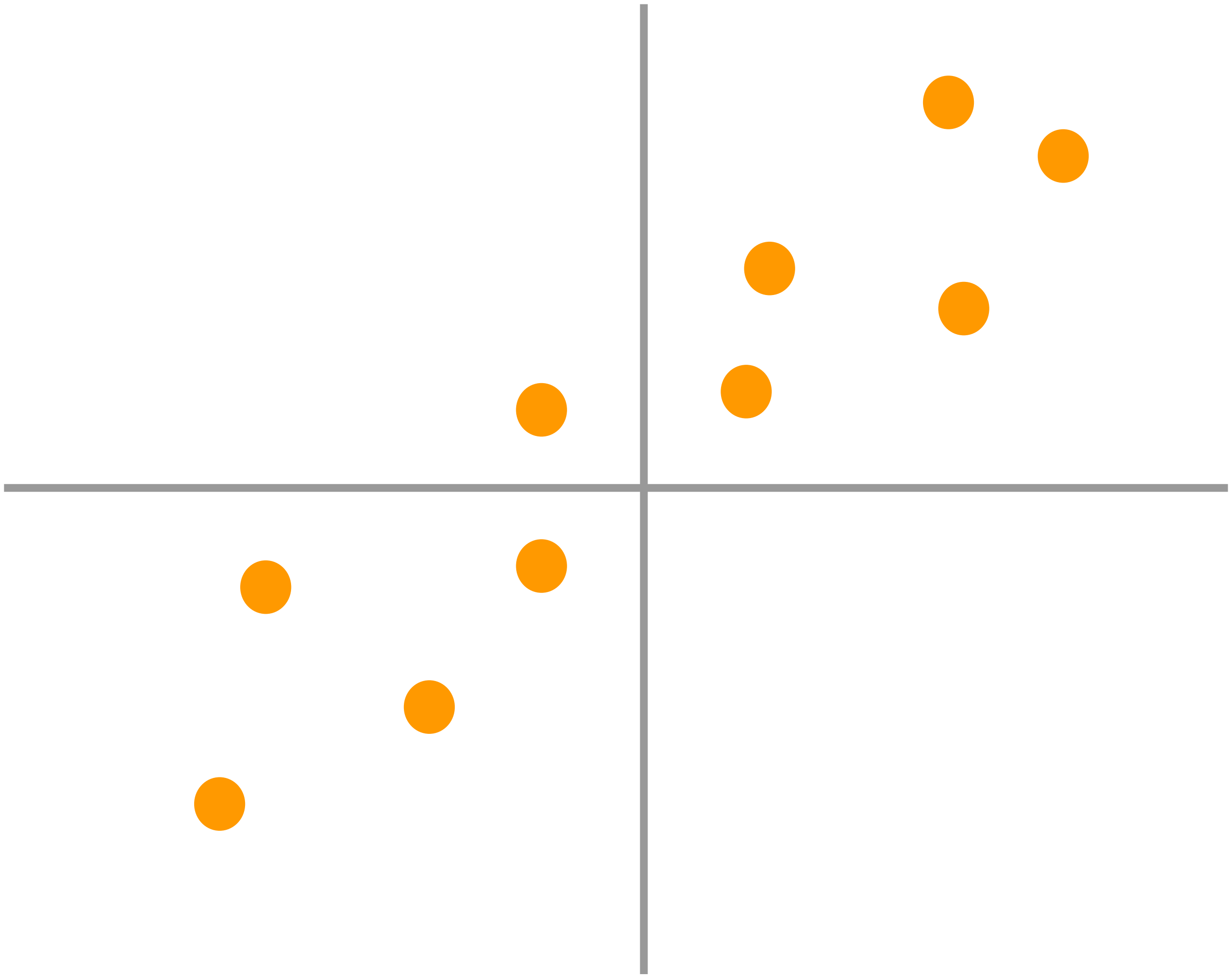


4
6
8
10
12
14
16
18
4
6
8
10
12
x1
y1
x2
y2
x3
y3
x4
y4

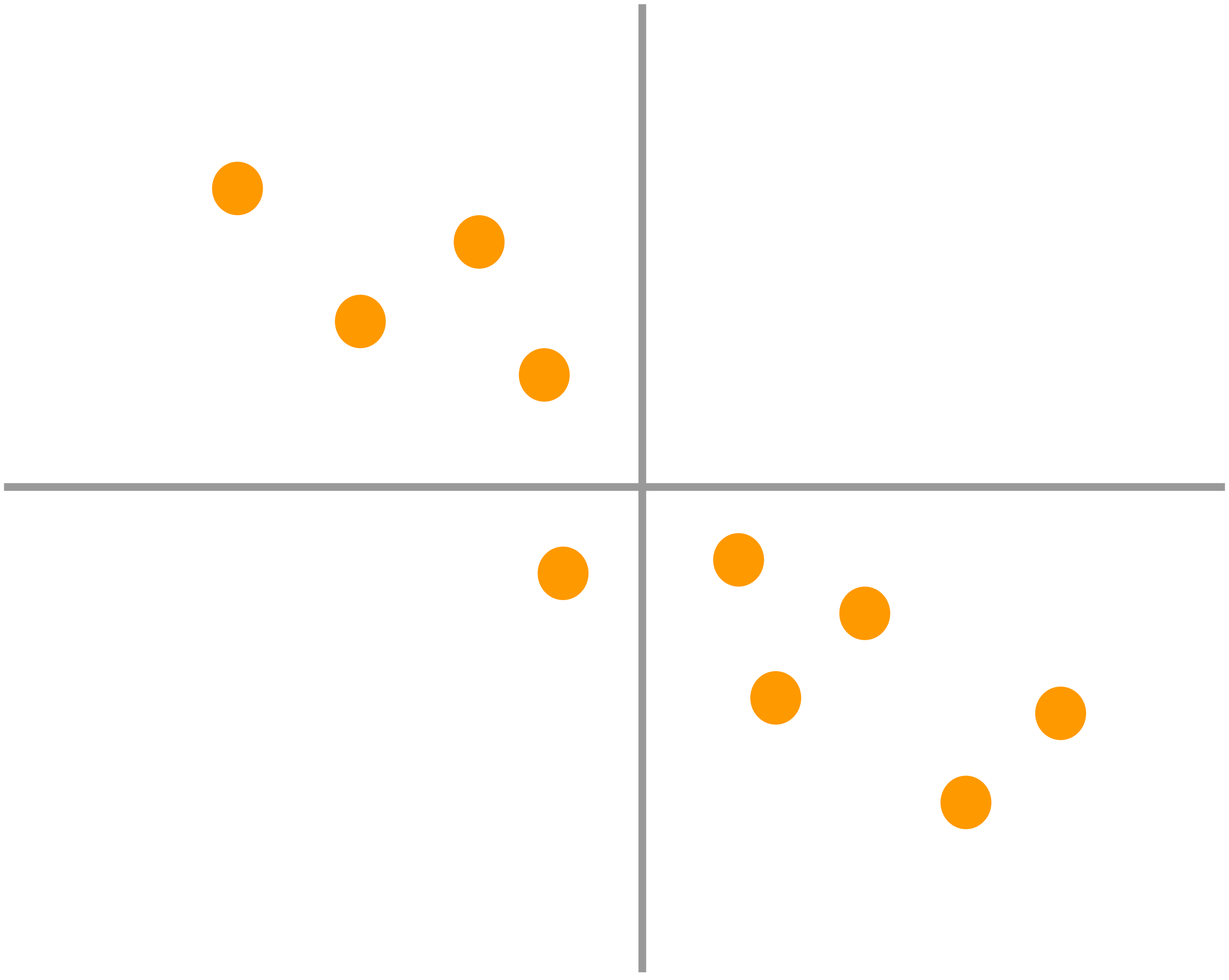
Types of Relations

Linear relationship between X and Y

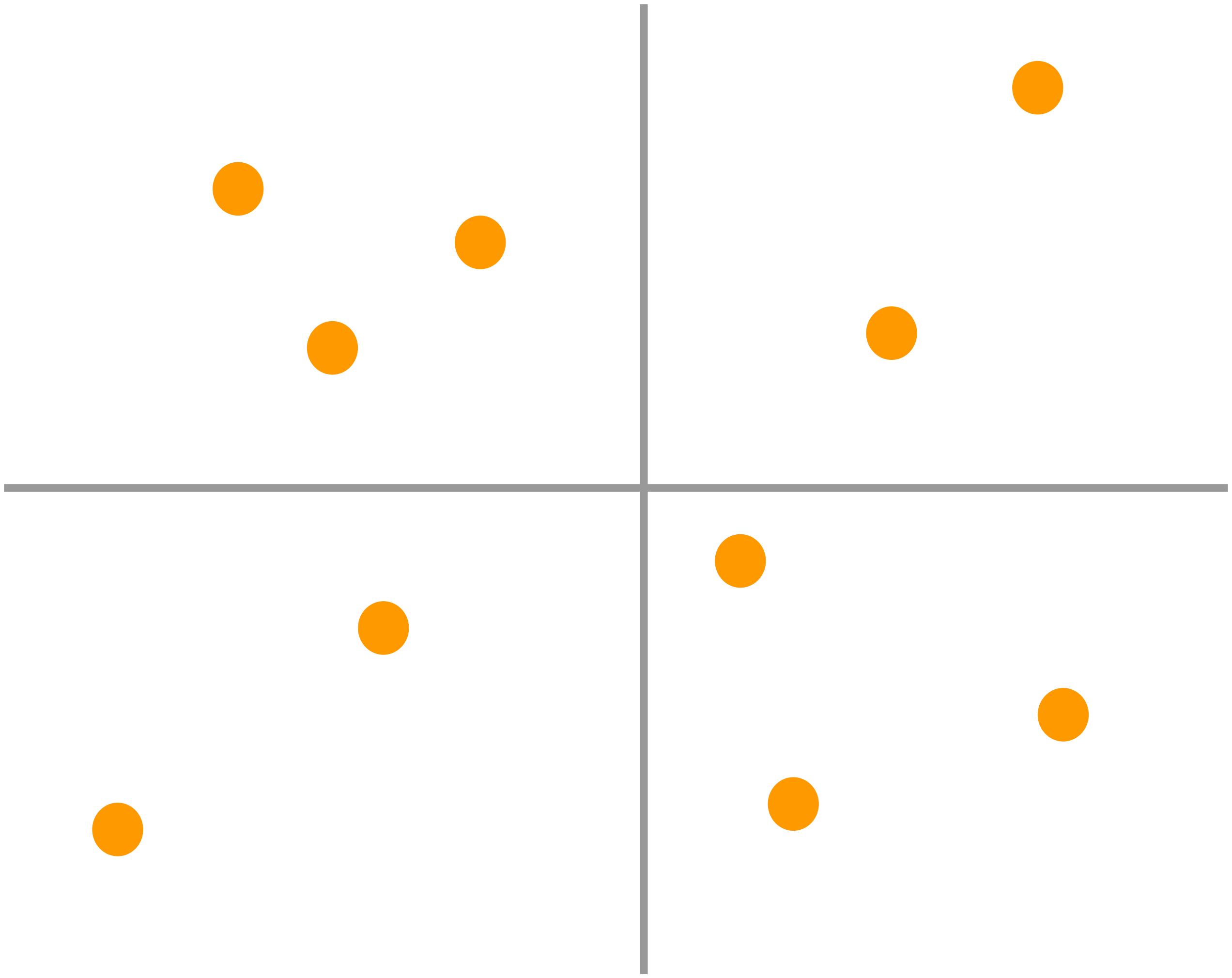
Positive linear relation



Negative linear relation



Little or no linear relation



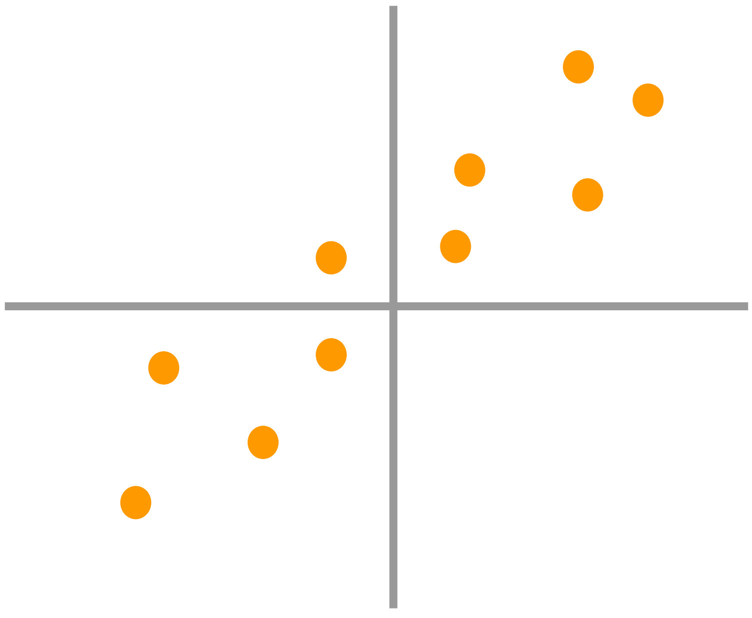
Correlation Coefficient

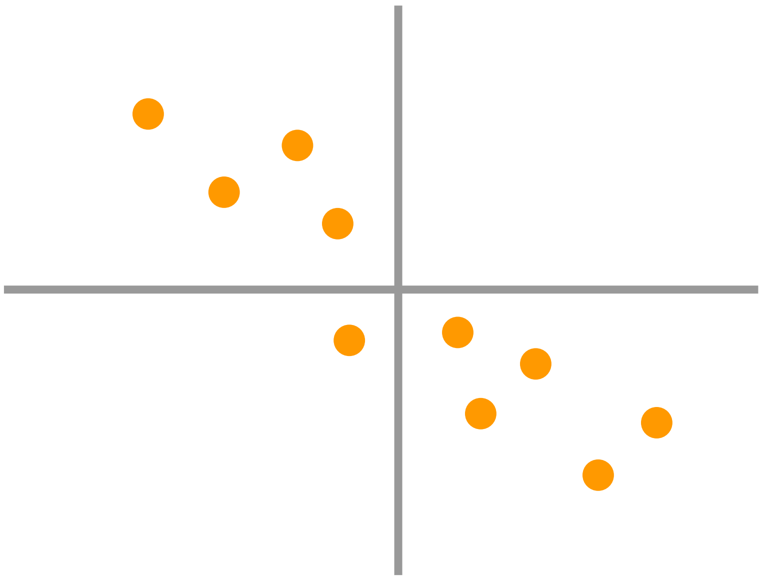
Correlation Coefficient

Summary statistic that measures the strength of a linear relationship between two variables X and Y

In simpler terms ...

The Correlation coefficient measures how the points in a scatter diagram are close to a line how close the points in a scatter diagram are close to a line



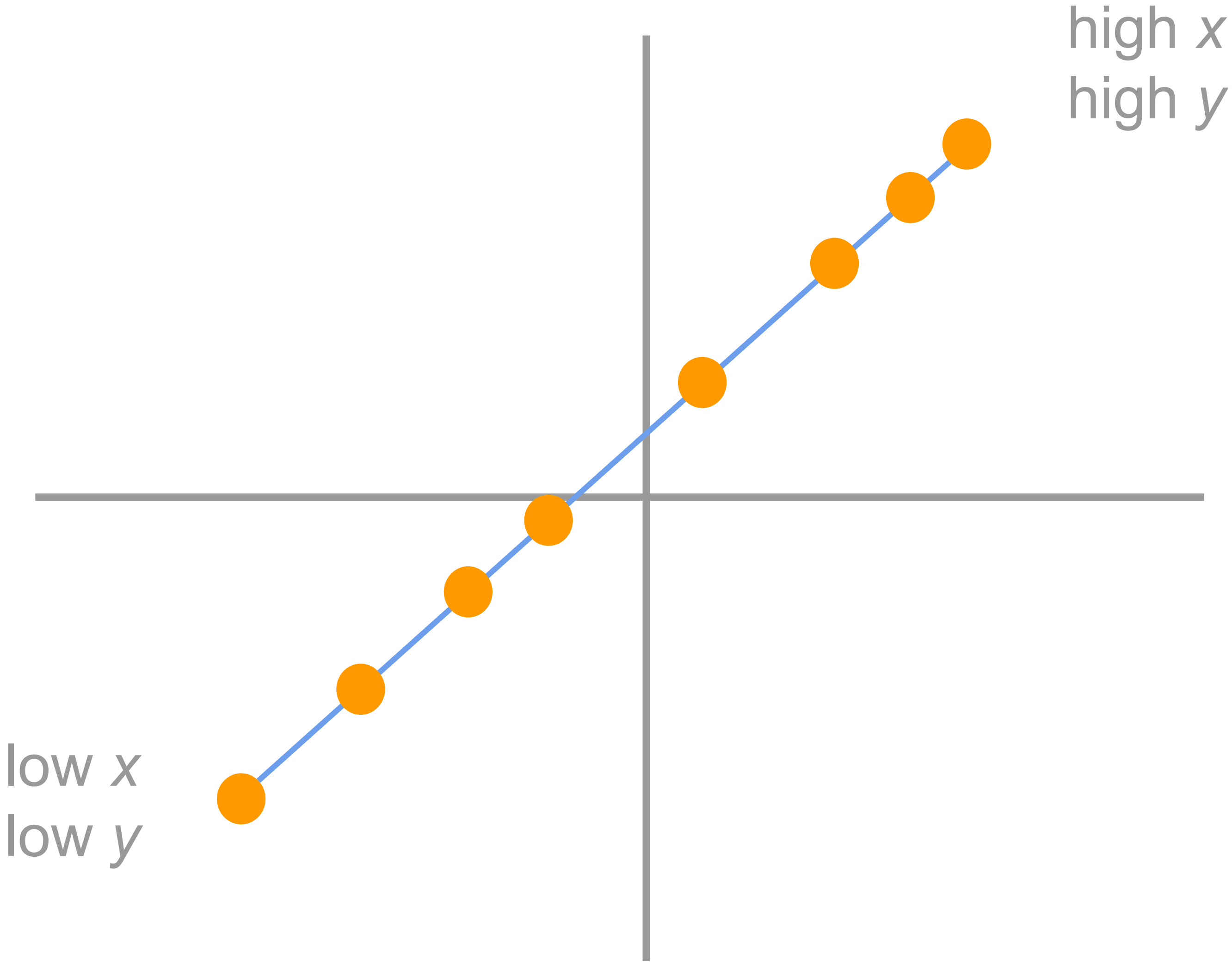


About the Correlation Coefficient *r*

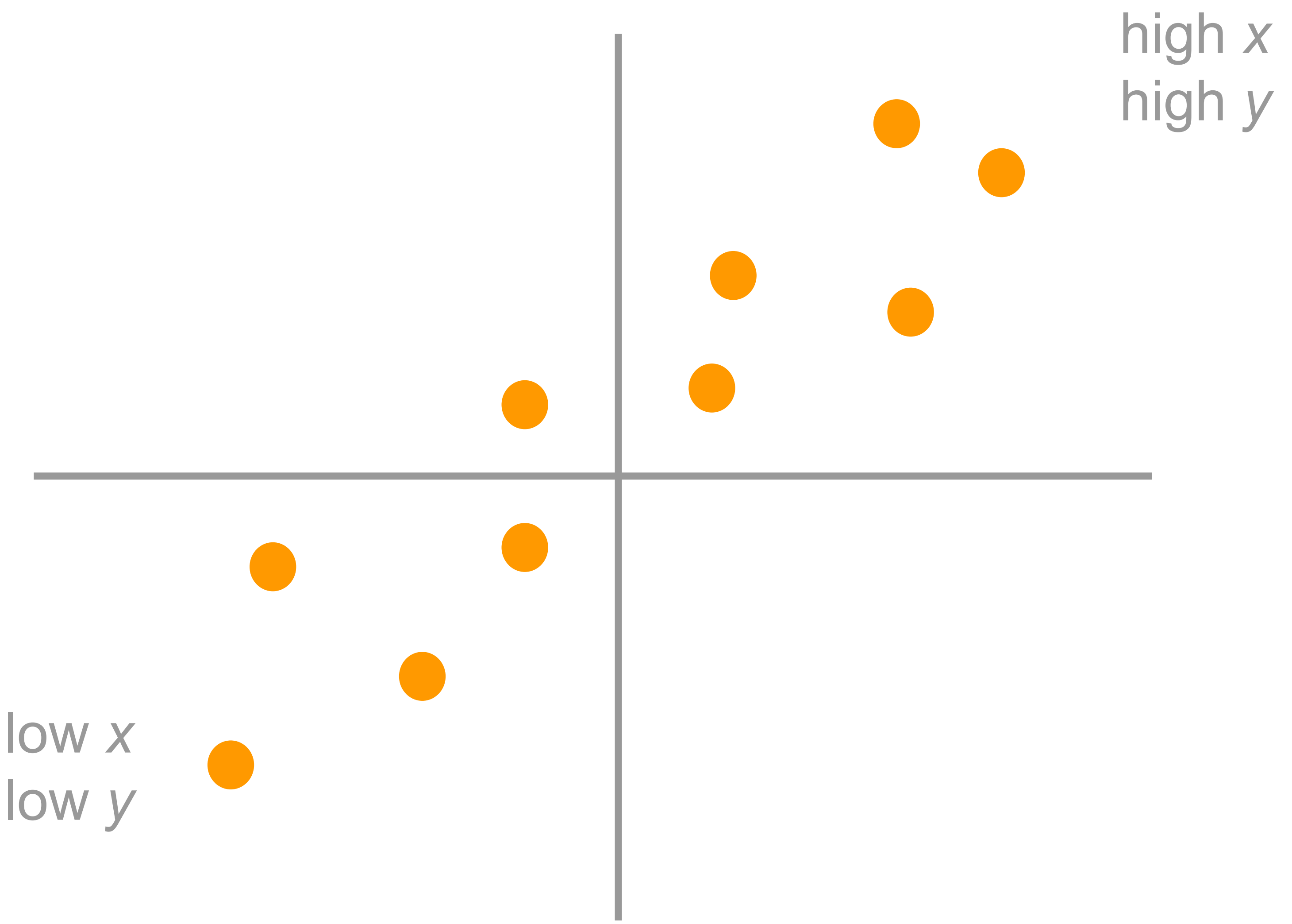
$$-1 \leq r \leq 1$$

$r = 1$	points on a line with positive slope
$r = -1$	points on a line with negative slope
$r = 0$	no linear relation

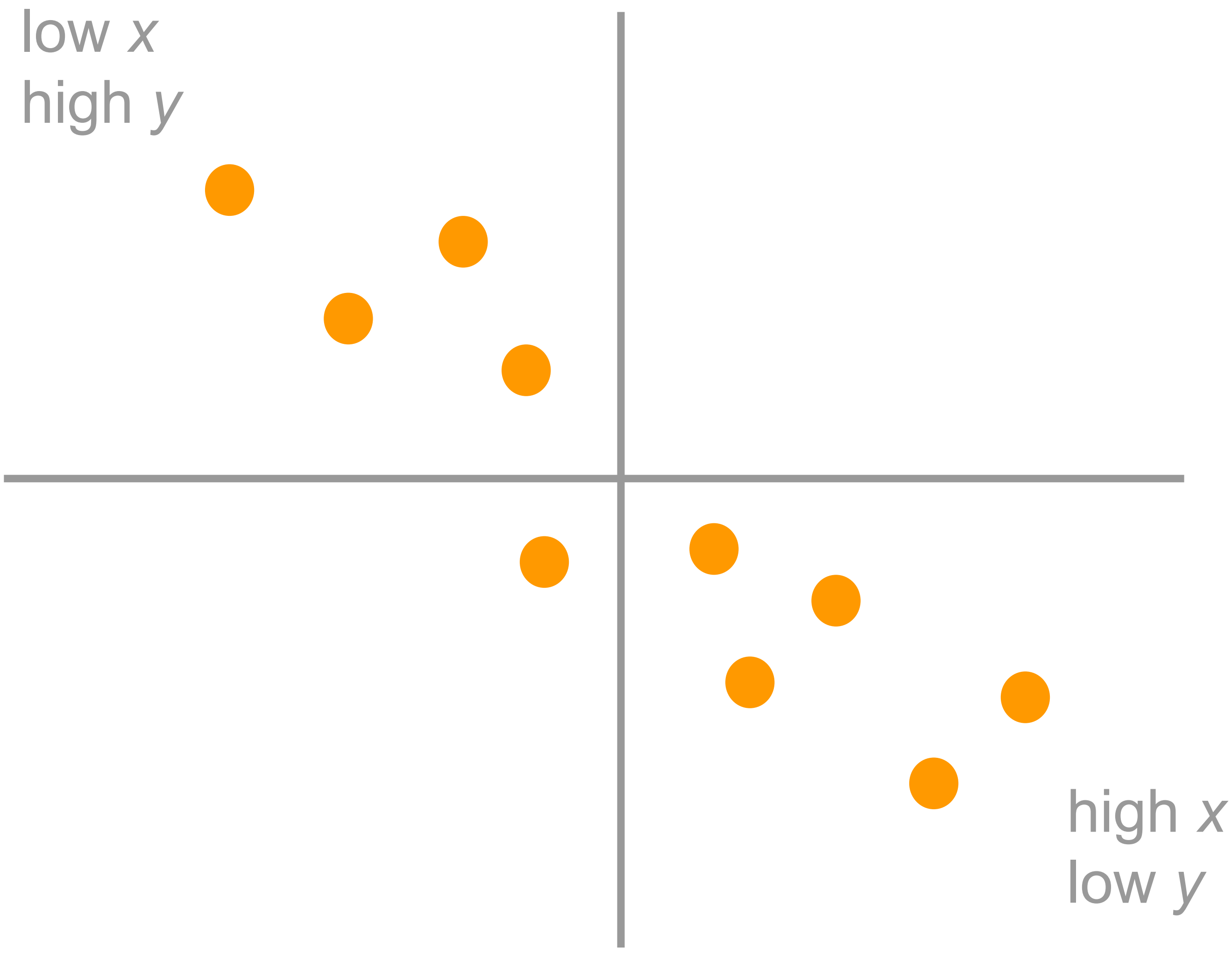
Correlation r = 1



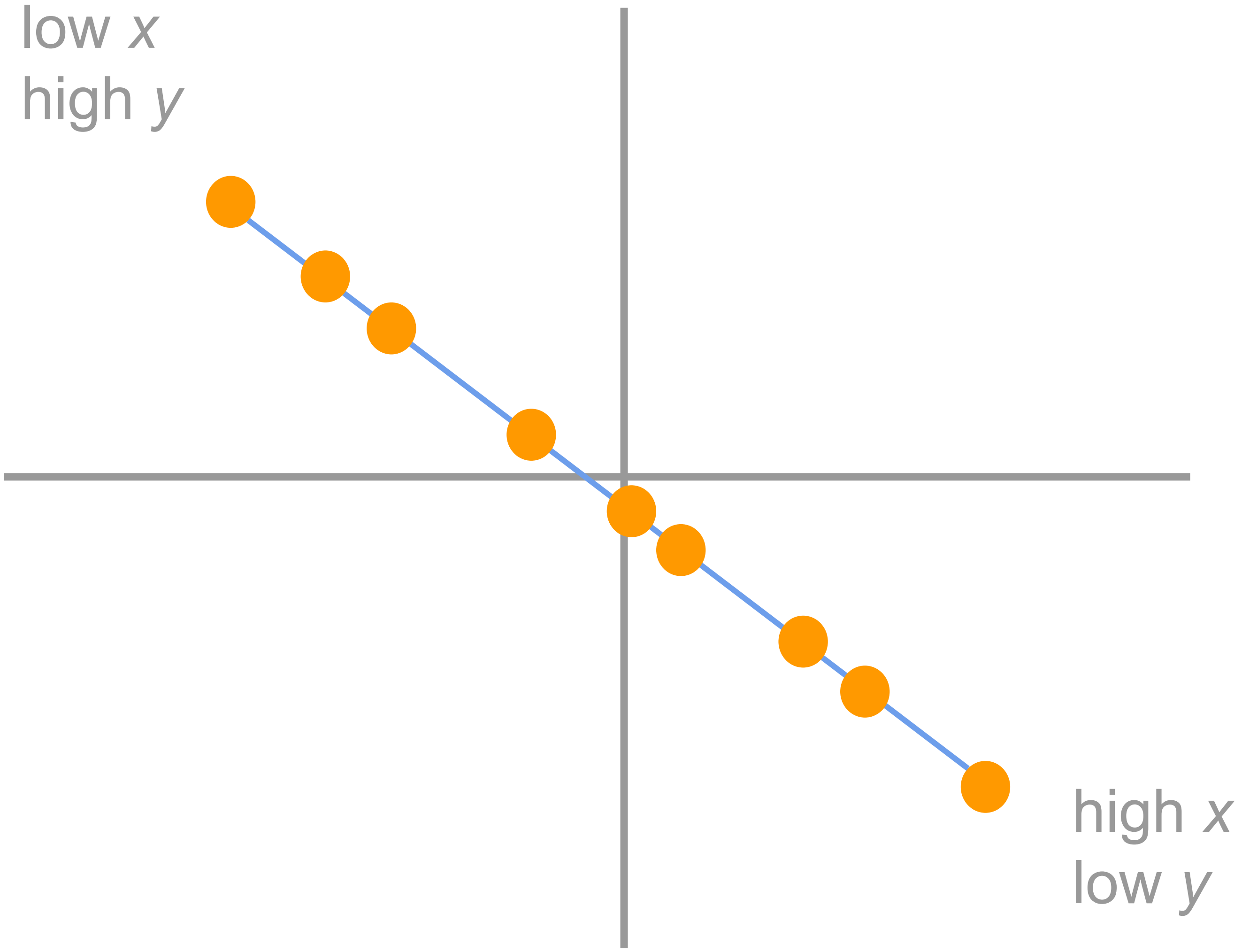
Positive Correlation: **r** close to 1



Negative correlation: **r** close to -1

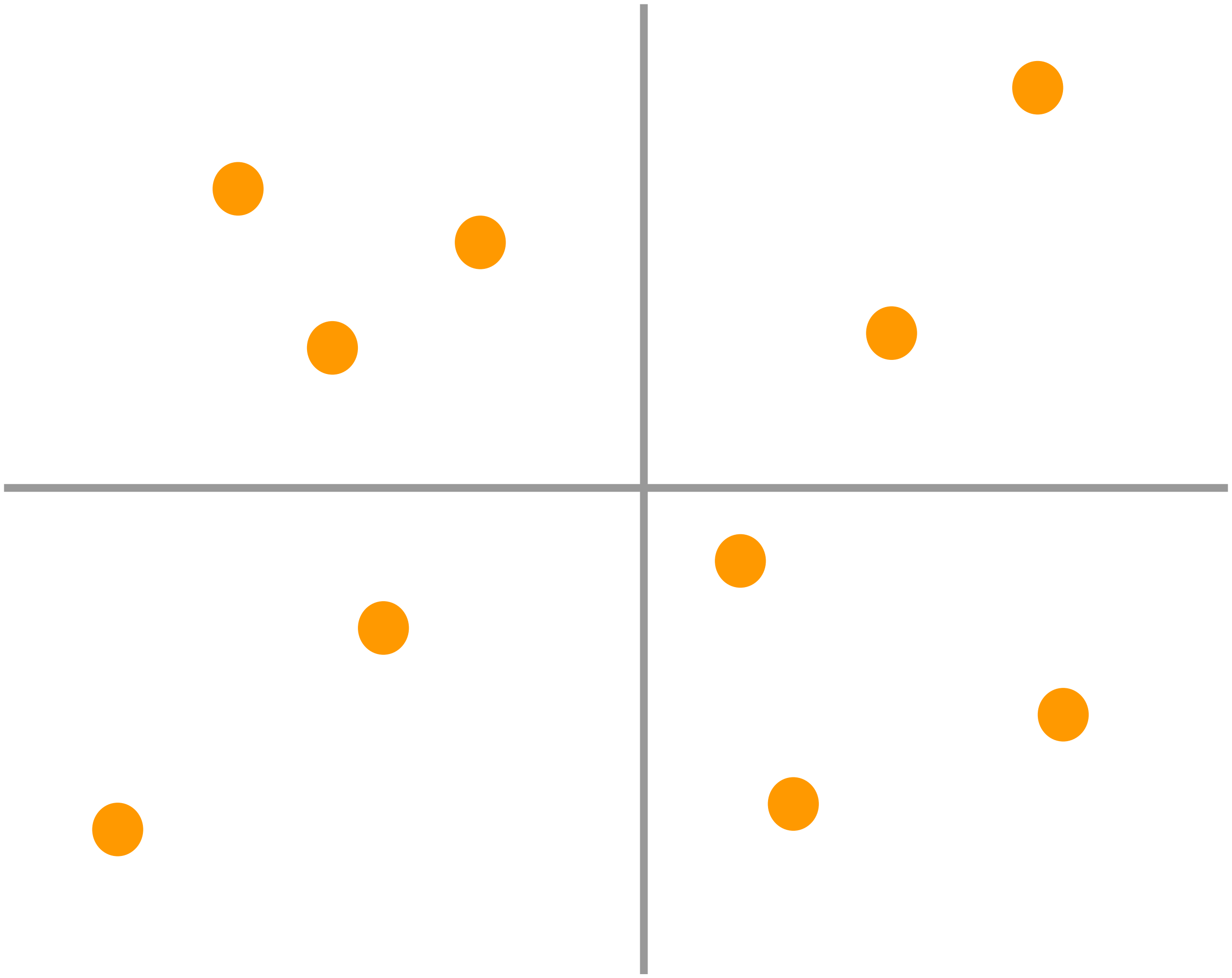


Correlation **r** = -1



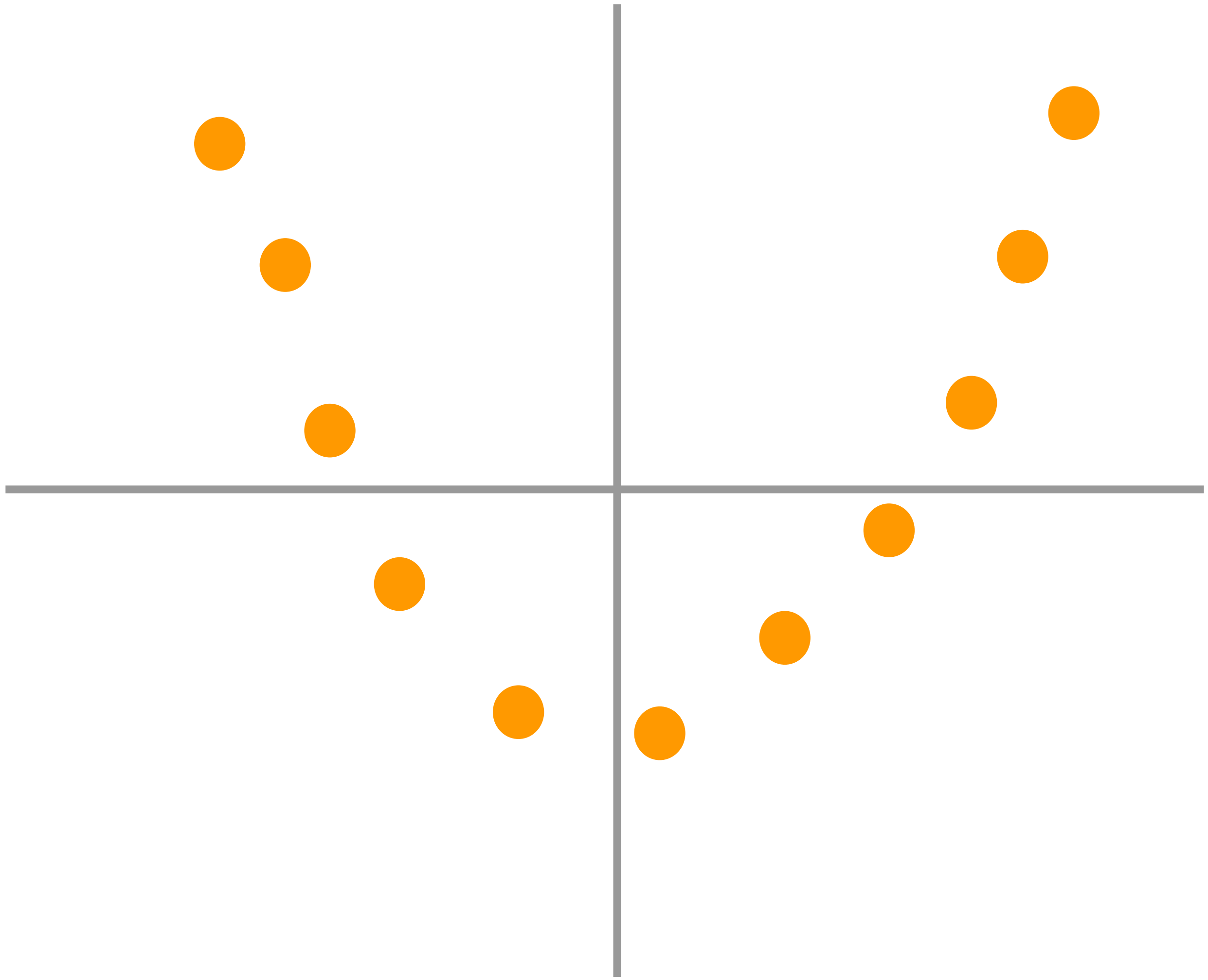
Gaston Sanchez

Little or no linear relation

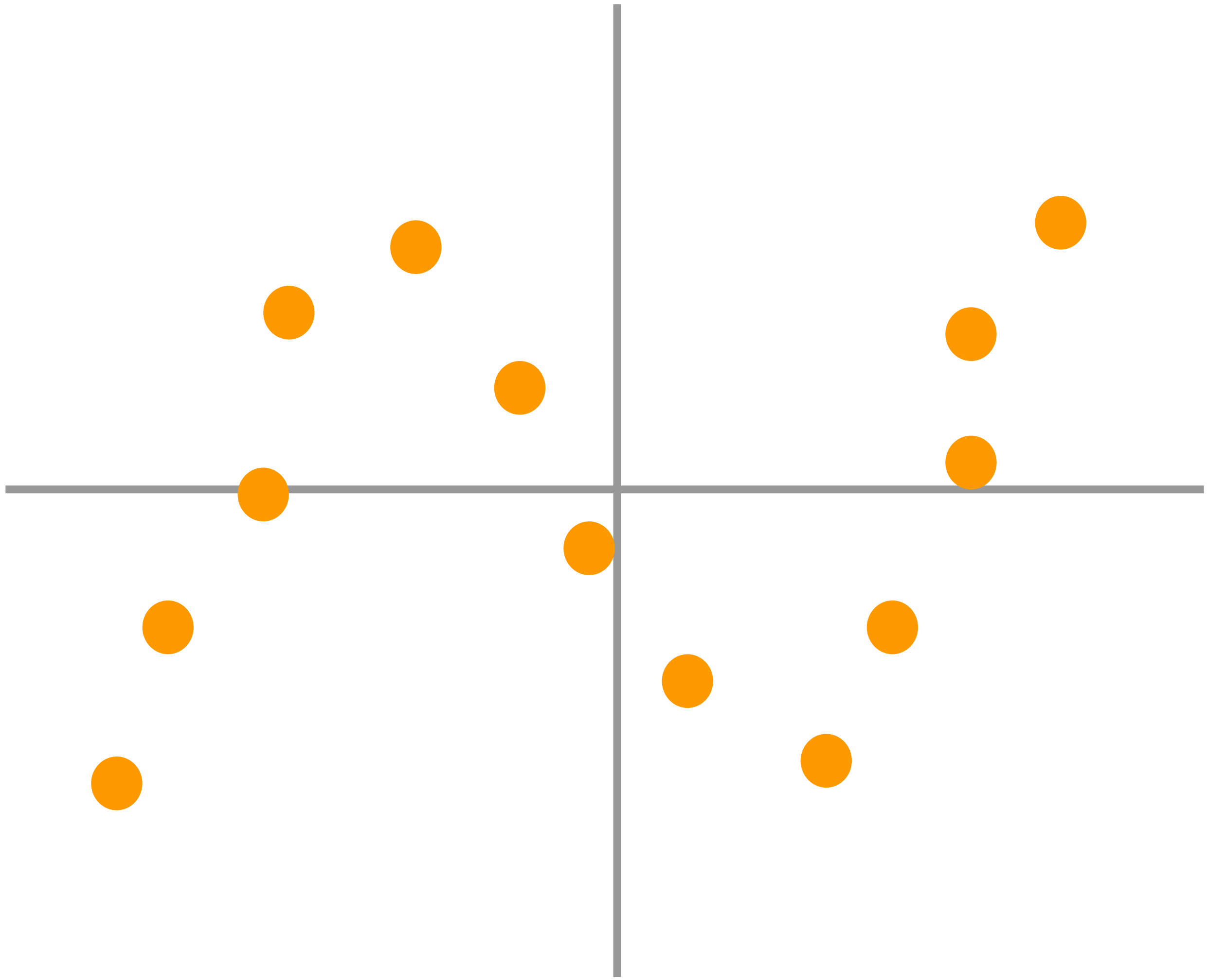


Gaston Sanchez

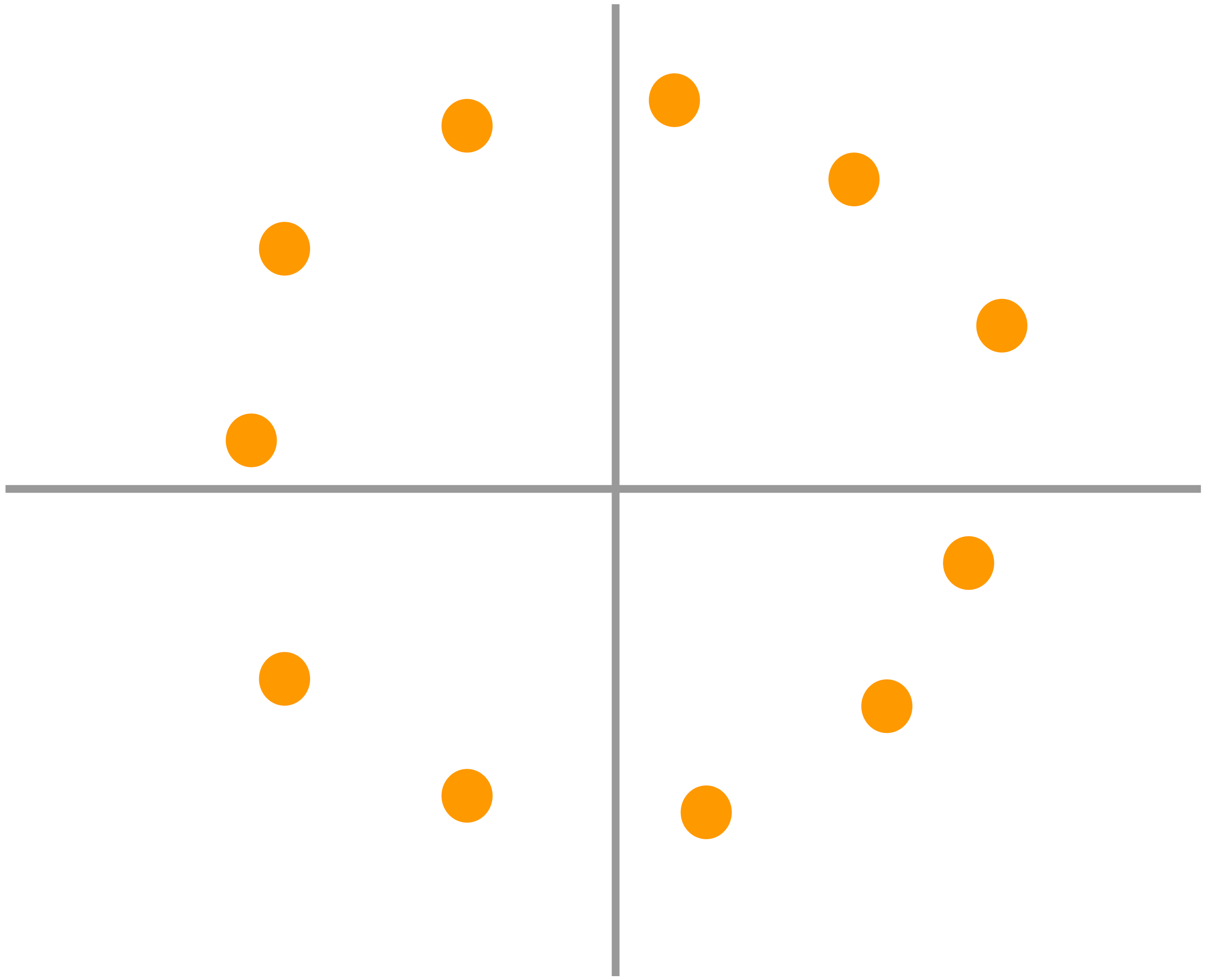
Non-linear correlation r = 0



Non-linear correlation r = 0



Non-linear correlation r = 0



About the Correlation Coefficient *r*

$$-1 \leq r \leq 1$$

$r = 1$	perfect positive linear relation
$0 < r < 1$	positive linear relation
$r = 0$	no linear relation
$0 > r > -1$	negative correlation
$r = -1$	perfect negative linear relation

Calculating Correlation Coefficient

Finding the correlation coefficient

1. Convert **X** and **Y** into standard units
   (find the average, find the SD)

2. Take products of **SU(x)** and **SU(y)**

3. Take average of products

Reminder: Standard Units

SU: Measures how many SDs a value is above or below the average

$$SU = \frac{\text{value - average}}{\text{SD}}$$

Coefficient of Correlation *r*

$$\mathbf{r} = \text{correlation}(X, Y)$$

$$\mathbf{r} = \text{average of } \left\{ \begin{bmatrix} \text{X in std} \\ \text{units} \end{bmatrix} \times \begin{bmatrix} \text{Y in std} \\ \text{units} \end{bmatrix} \right\}$$

Toy dataset

X	Y
1	5
3	9
4	7
5	1
7	13

average X = 4
SDx = 2

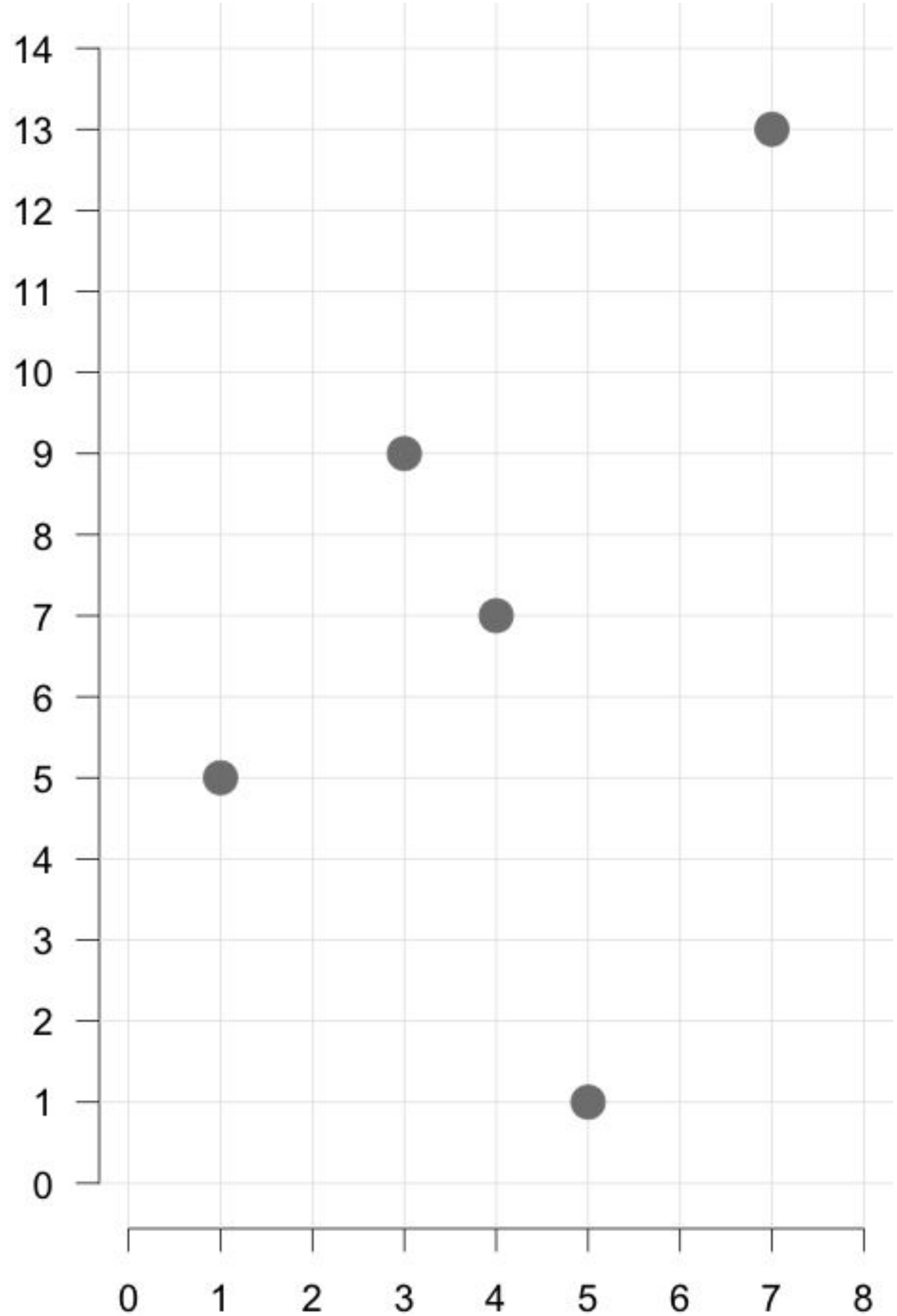
average Y = 7
SDy = 4

X	Y	X std units	Y std units	product (SUx)(SUy)
1	5	-1.5	-0.5	0.75
3	9	-0.5	0.5	-0.25
4	7	0.0	0.0	0.00
5	1	0.5	-1.5	-0.75
7	13	1.5	1.5	2.25

r = average of (x in std units) x (y in std units)

r = (0.75 - 0.25 + 0.00 - 0.75 + 2.25) / 5 = 0.40

Gaston Sanchez



r = 0.40

Properties of correlation

Standard Units

$$SU = \frac{\text{value - average}}{\text{SD}}$$

original units cancel out

About the correlation coefficient

r is unitless

hard to compare different **r**'s

r = 0.5 doesn't mean that 50% of data is around a line

$r_1 = 0.2$ vs $r_2 = 0.4$ doesn't mean that points in line 2 are twice as clustered as those in line 1

Some remarks

r is not affected if you switch X and Y

Coefficient of Correlation *r*

r = average of { [X in std units] x [Y in std units] }

correlation(X, Y) = correlation(Y, X)

X	Y	X std units	Y std units	product (SUx)(SUy)	product (SUy)(SUx)
1	5	-1.5	-0.5	0.75	0.75
3	9	-0.5	0.5	-0.25	-0.25
4	7	0.0	0.0	0.00	0.00
5	1	0.5	-1.5	-0.75	-0.75
7	13	1.5	1.5	2.25	2.25

r = average of (y in std units) x (x in std units)

r = (0.75 - 0.25 + 0.00 - 0.75 + 2.25) / 5 = 0.40

Gaston Sanchez

Some remarks

r is not affected if you add the same number to all the values of one variable

X	3+X	Y	X std units	Y std units	product (SUx)(SUy)
1	4	5	-1.5	-0.5	0.75
3	6	9	-0.5	0.5	-0.25
4	7	7	0.0	0.0	0.00
5	8	1	0.5	-1.5	-0.75
7	10	13	1.5	1.5	2.25

r = average of (x in std units) x (y in std units)

r = (0.75 - 0.25 + 0.00 - 0.75 + 2.25) / 5 = 0.40

Gaston Sanchez

Some remarks

r is not affected if you multiply all the values of one variable by the same positive number

X	2X	Y	X std units	Y std units	product (SUx)(SUy)
1	2	5	-1.5	-0.5	0.75
3	6	9	-0.5	0.5	-0.25
4	8	7	0.0	0.0	0.00
5	10	1	0.5	-1.5	-0.75
7	14	13	1.5	1.5	2.25

r = average of (x in std units) x (y in std units)

r = (0.75 - 0.25 + 0.00 - 0.75 + 2.25) / 5 = 0.40

Gaston Sanchez

What about multiplying all the values by a negative number?

X	-2X	Y	X std units	Y std units	product (SUx)(SUy)
1	-2	5	1.5	-0.5	-0.75
3	-6	9	0.5	0.5	0.25
4	-8	7	0.0	0.0	0.00
5	-10	1	-0.5	-1.5	0.75
7	-14	13	-1.5	1.5	-2.25

r = average of (x in std units) x (y in std units)

r = (-0.75 + 0.25 + 0.00 + 0.75 - 2.25) / 5 = - 0.40

Gaston Sanchez

Some remarks

r is not affected by changes of linear scale (add the same number and multiply by the same positive number)

X	2X + 3	Y	X std units	Y std units	product (SUx)(SUy)
1	5	5	-1.5	-0.5	0.75
3	9	9	-0.5	0.5	-0.25
4	11	7	0.0	0.0	0.00
5	13	1	0.5	-1.5	-0.75
7	17	13	1.5	1.5	2.25

r = average of (x in std units) x (y in std units)

r = (0.75 - 0.25 + 0.00 - 0.75 + 2.25) / 5 = 0.40

Gaston Sanchez