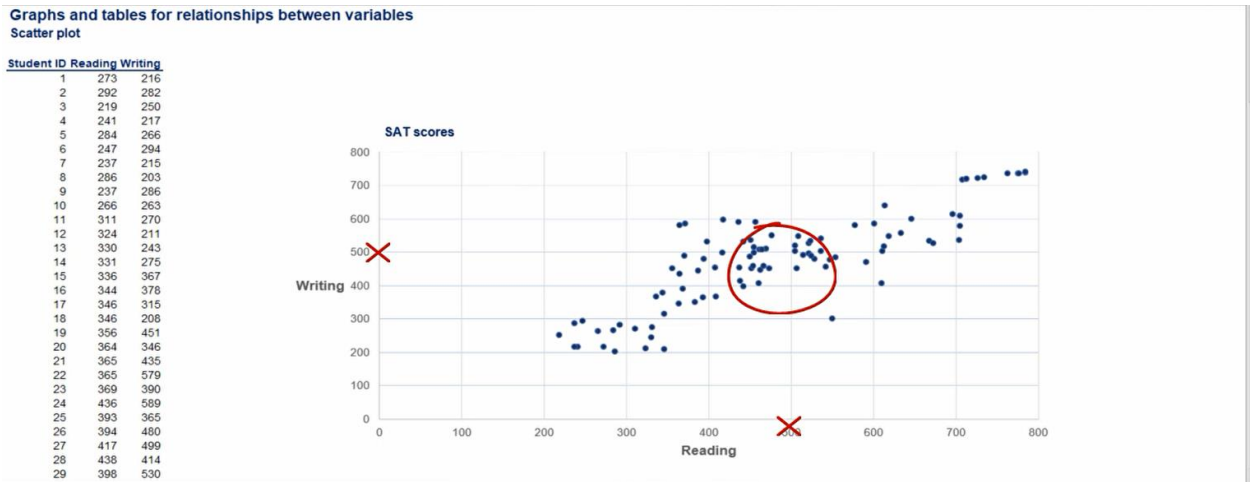
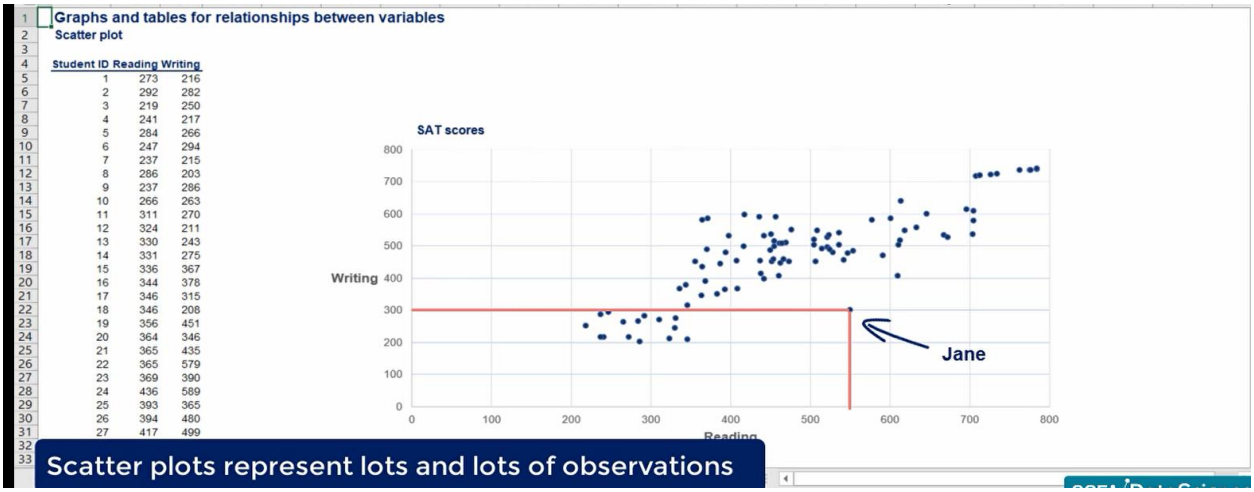
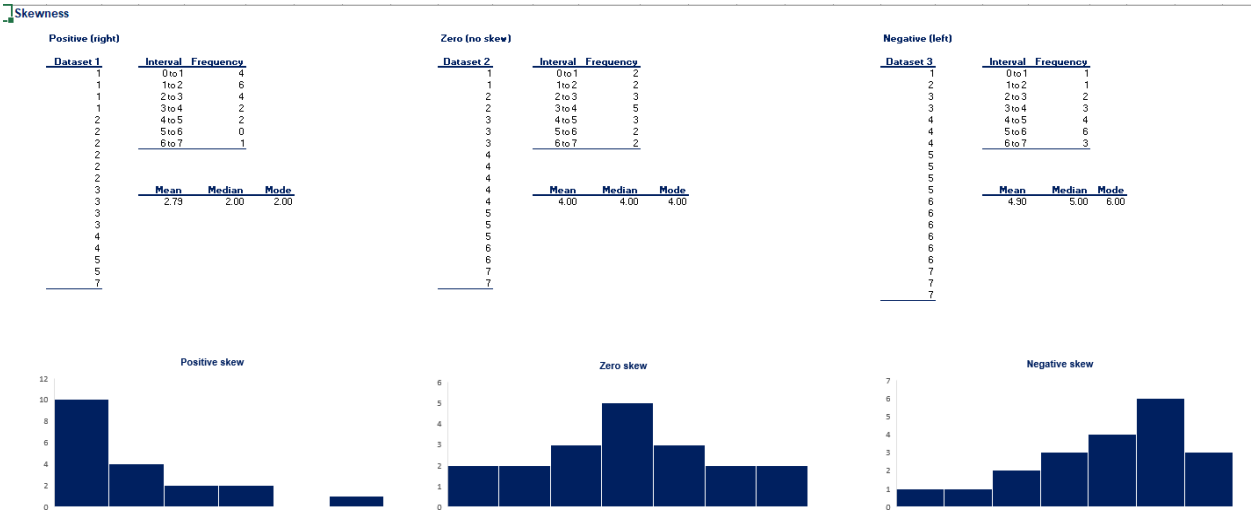
Scatter Plot

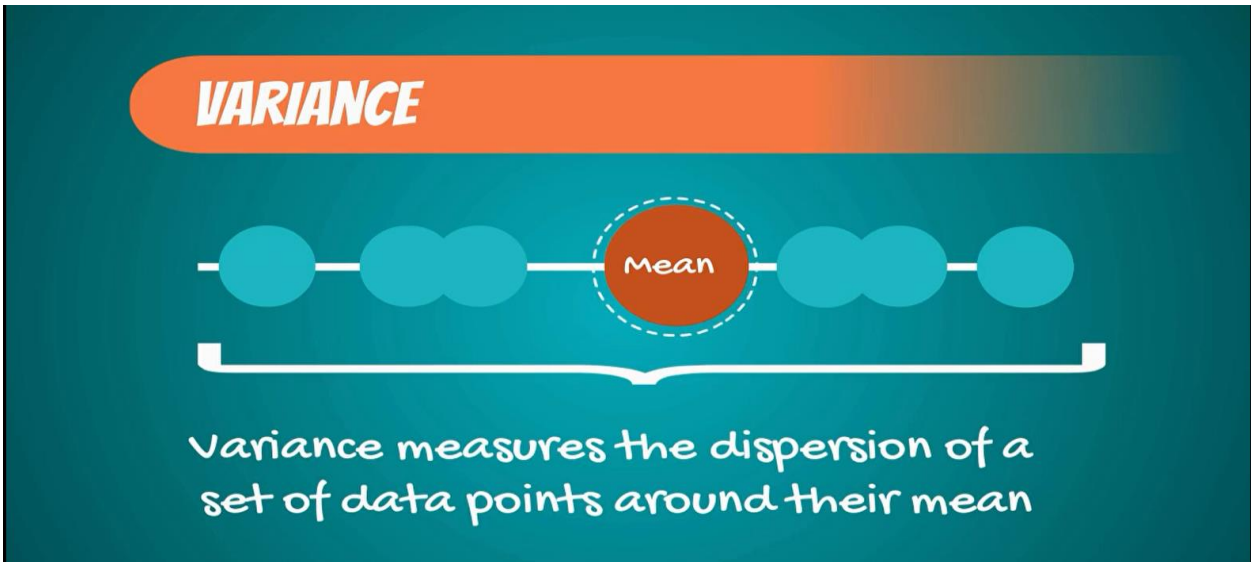


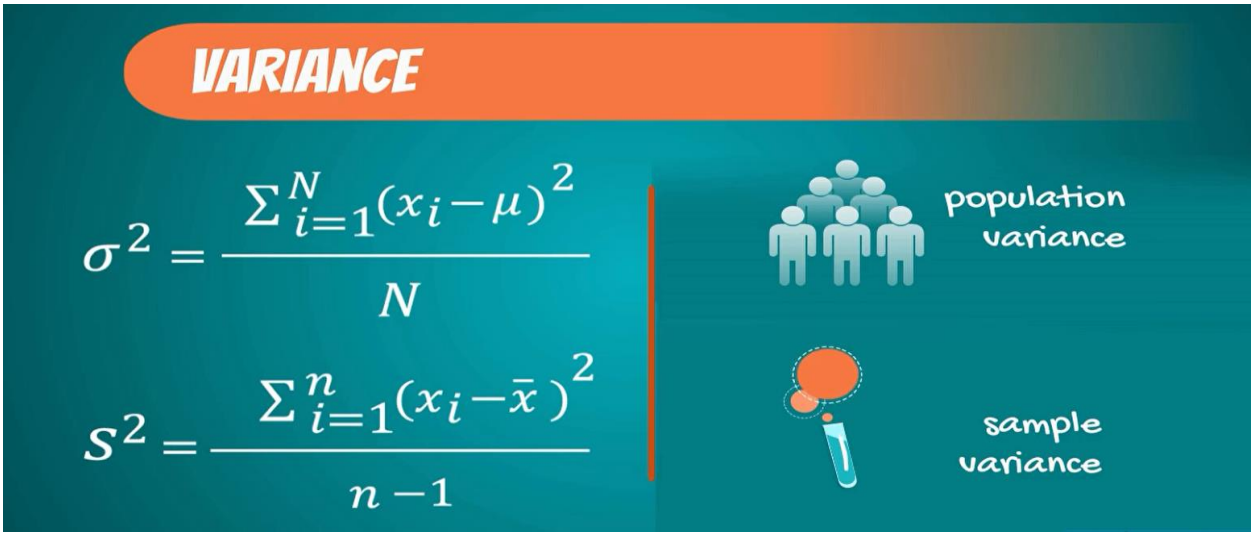


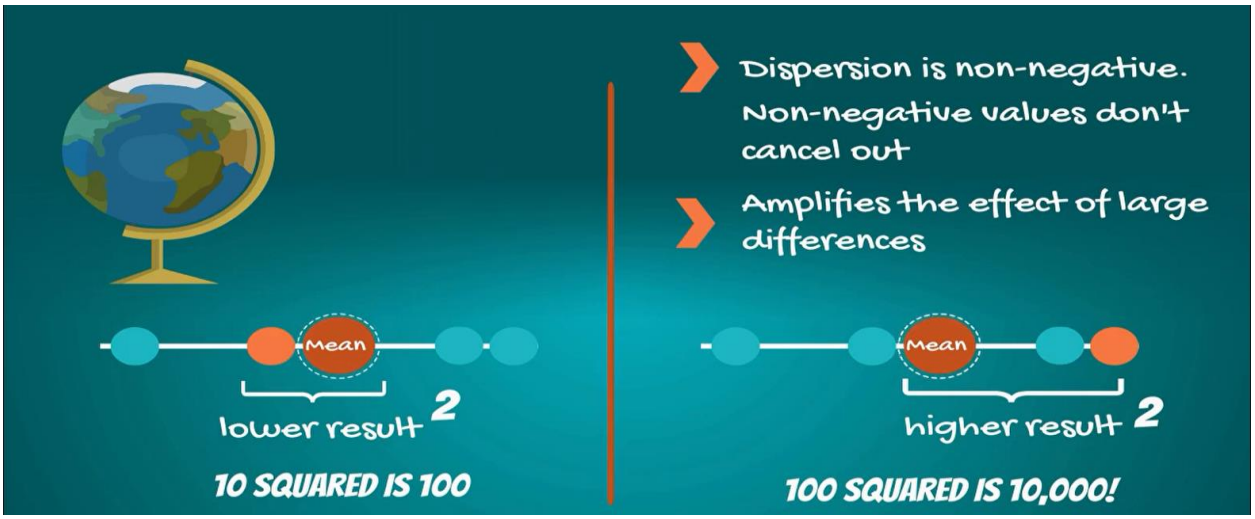
Skewness



Varience


VARIANCE
Mean
Variance measures the dispersion of a set of data points around their mean


VARIANCE
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
population variance
sample variance


Mean
lower result 2
10 SQUARED IS 100
Dispersion is non-negative. Non-negative values don't cancel out
Amplifies the effect of large differences
Mean
higher result 2
100 SQUARED IS 10,000!

Variance

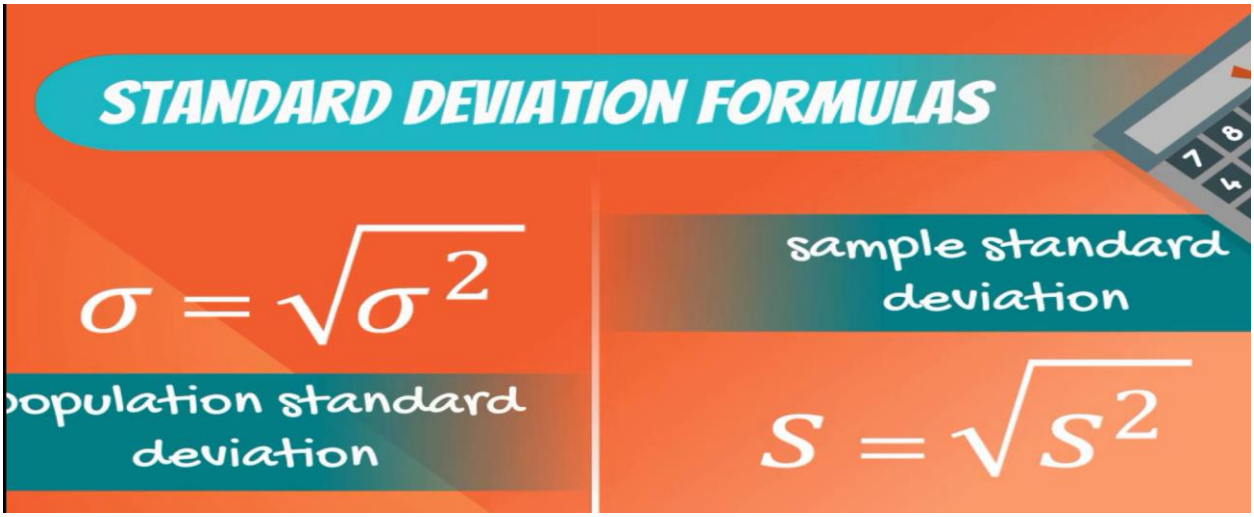
Population
1
2
3
4
5

Mean	3.00
Population variance	2.00
Sample variance	2.50

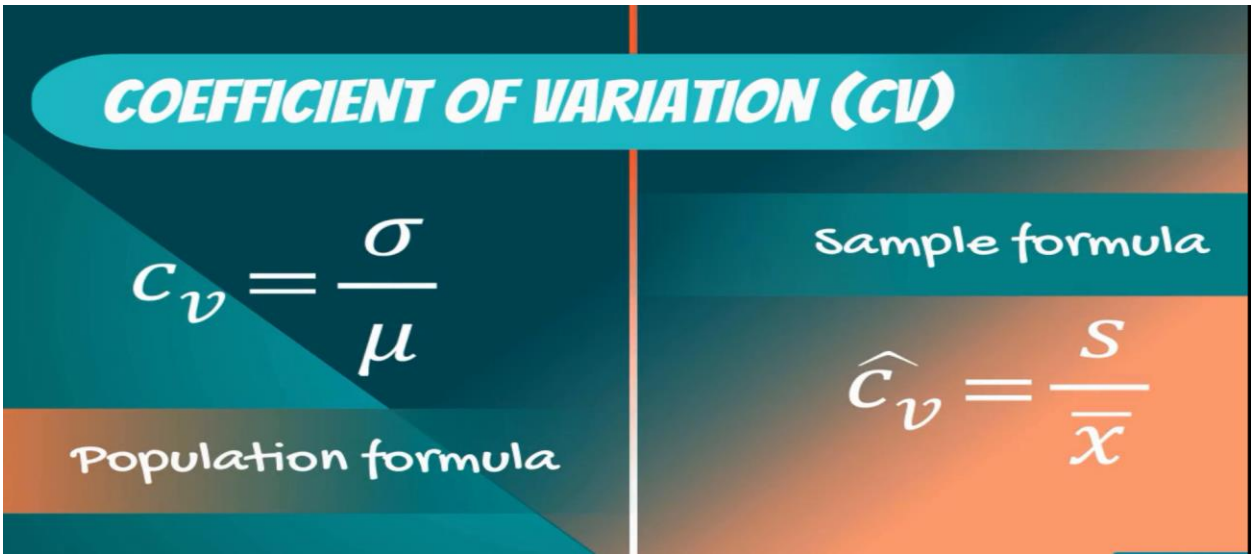
Imaginary population
1
1
1
2
3
4
5
5
5
5

Mean	3.20
Population variance	2.96

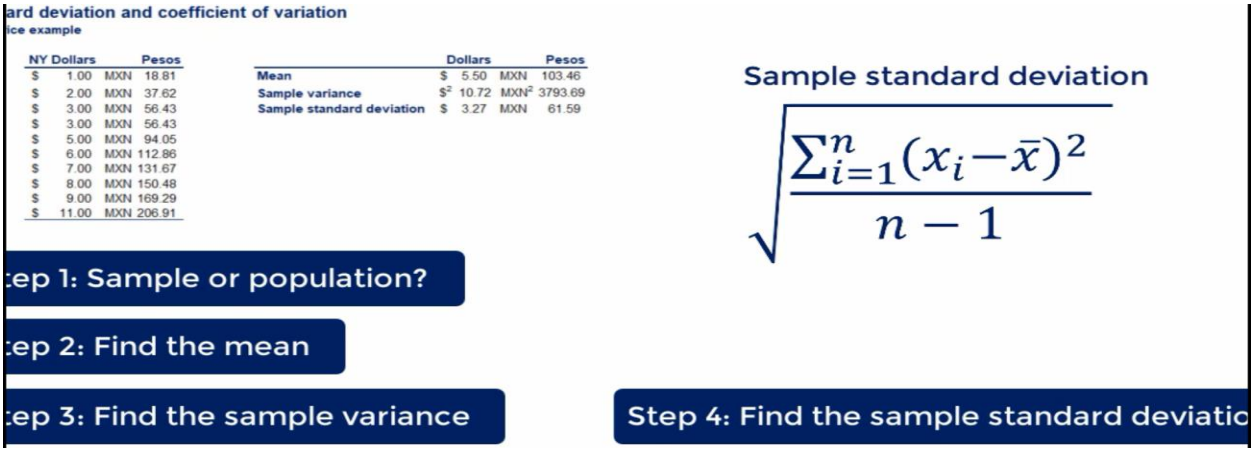
Standard Deviation



Coefficient of Variation (CV)



Comparing the standard deviations of two different data sets is meaningless but Comparing coefficient of coefficients is meaningful



ard deviation and coefficient of variation

ice example

NY Dollars		Pesos
$ 1.00	MXN	18.81
$ 2.00	MXN	37.62
$ 3.00	MXN	56.43
$ 3.00	MXN	56.43
$ 5.00	MXN	94.05
$ 6.00	MXN	112.86
$ 7.00	MXN	131.67
$ 8.00	MXN	150.48
$ 9.00	MXN	169.29
$ 11.00	MXN	206.91

	Dollars		Pesos
Mean	$ 5.50	MXN	103.46
Sample variance	$^2 10.72	MXN^2	3793.69
Sample standard deviation	$ 3.27	MXN	61.59
Sample coefficient of variation	0.60		0.60

- does not have a unit of measurement
- universal across datasets
- perfect for comparisons

Covariance and Liner corelation coefficient

Covariance

Housing data

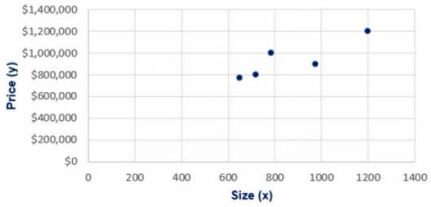
Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000

Sample formula

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population formula

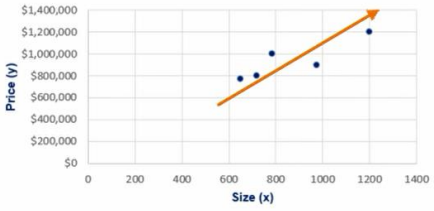
$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$



Covariance

Housing data

Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000

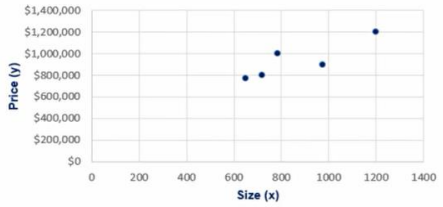


The two variables are correlated and the main statistic to measure this correlation is called covariance

Covariance

Housing data

	x	y		
	Size (ft.)	Price ($)		$(x-\bar{x})*(y-\bar{y})$
	650	772,000		34,776,000
	785	998,000		- 5,265,000
	1200	1,200,000		89,178,000
	720	800,000		19,418,000
	975	895,000		- 4,142,000
Mean	866	933,000	Sum	133,965,000
			Sample size	5
			Cov. Sample	33,491,250



Covariance gives a sense of direction

> 0, the two variables move together
< 0, the two variables move in opposite directions
= 0, the two variables are independent

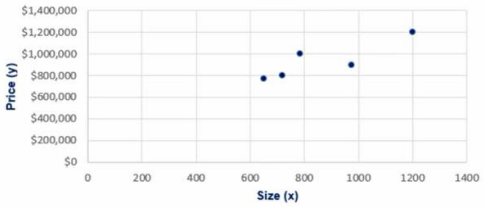
corelation coefficient

Correlation adjusts covariance, so that the relationship between the two variables becomes easy and intuitive to interpret

Correlation coefficient

Housing data

	Size (ft.)	Price ($)		$(x-\bar{x})*(y-\bar{y})$
	650	772,000		34,776,000
	785	998,000		- 5,265,000
	1200	1,200,000		89,178,000
	720	800,000		19,418,000
	975	895,000		- 4,142,000
Mean	866	933,000	Sum	133,965,000
Standard dev.	222	173,615	Sample size	5
			Cov. Sample	33,491,250



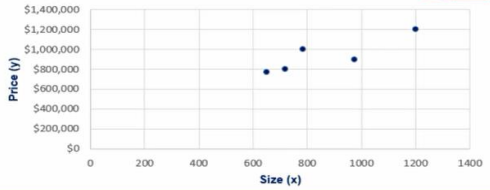
$$\frac{Cov(x,y)}{Stdev(x) * Stdev(y)}$$

$$\frac{S_{xy}}{S_x S_y} \qquad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$-1 \leq$ correlation coefficient ≤ 1

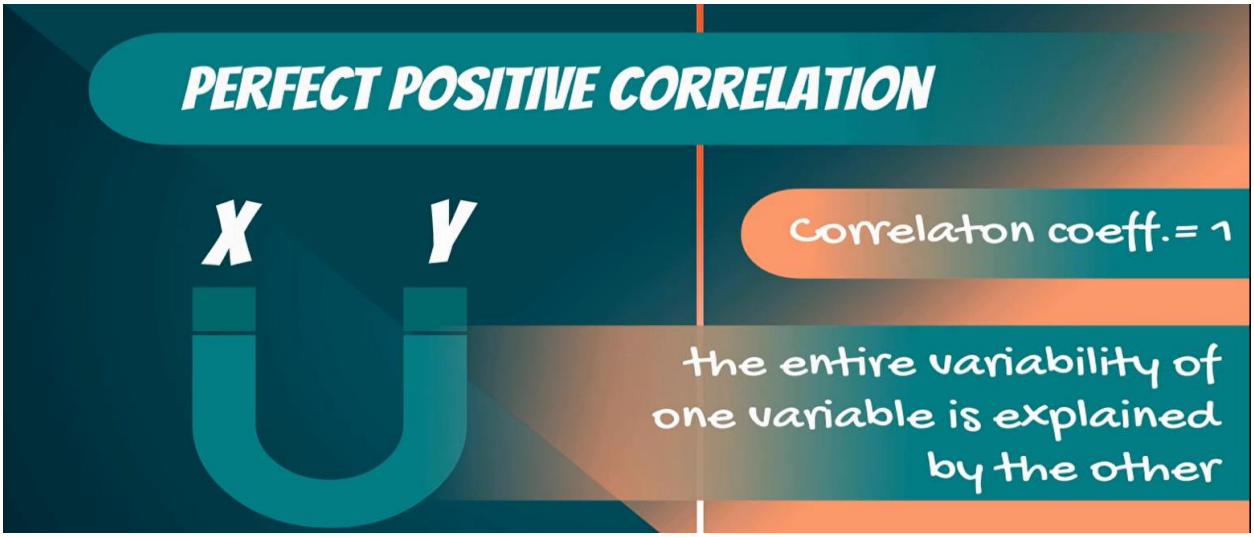
Correlation coefficient
Housing data

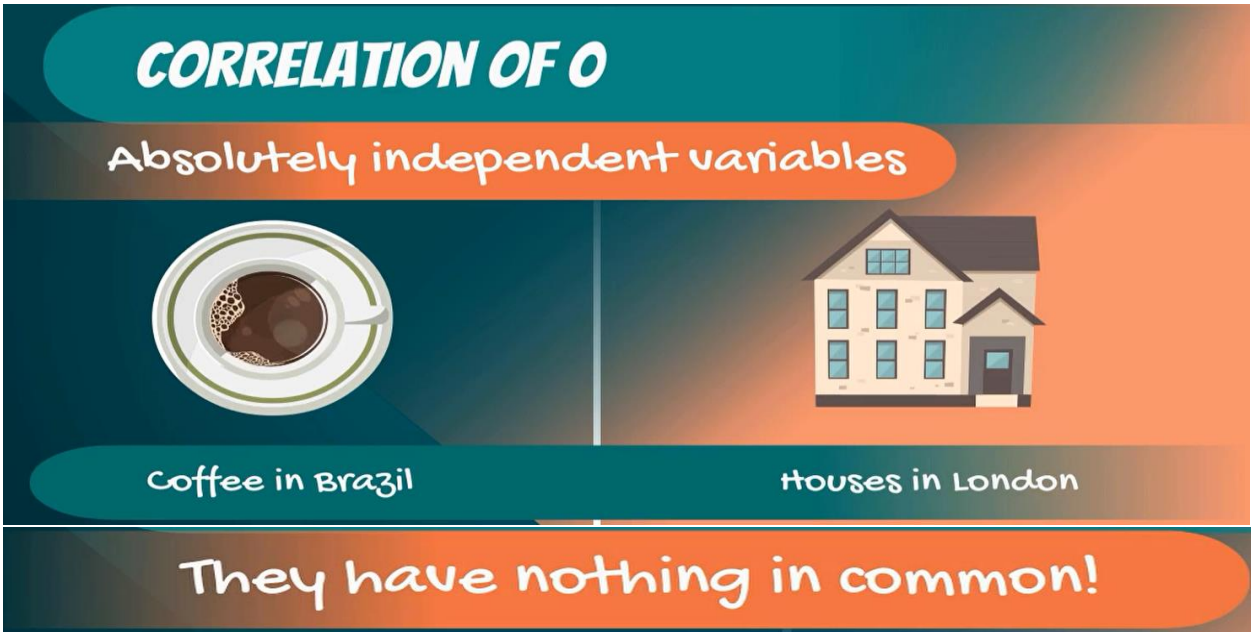
	Size (ft.)	Price ($)	(x-x̄)*(y-ȳ)
	650	772,000	34,776,000
	785	998,000	- 5,265,000
	1200	1,200,000	89,178,000
	720	800,000	19,418,000
	975	895,000	- 4,142,000
Mean	866	933,000	Sum 133,965,000
Standard dev.	222	173,615	Sample size 5
			Cov. Sample 33,491,250
			Correlation coeff. 0.87



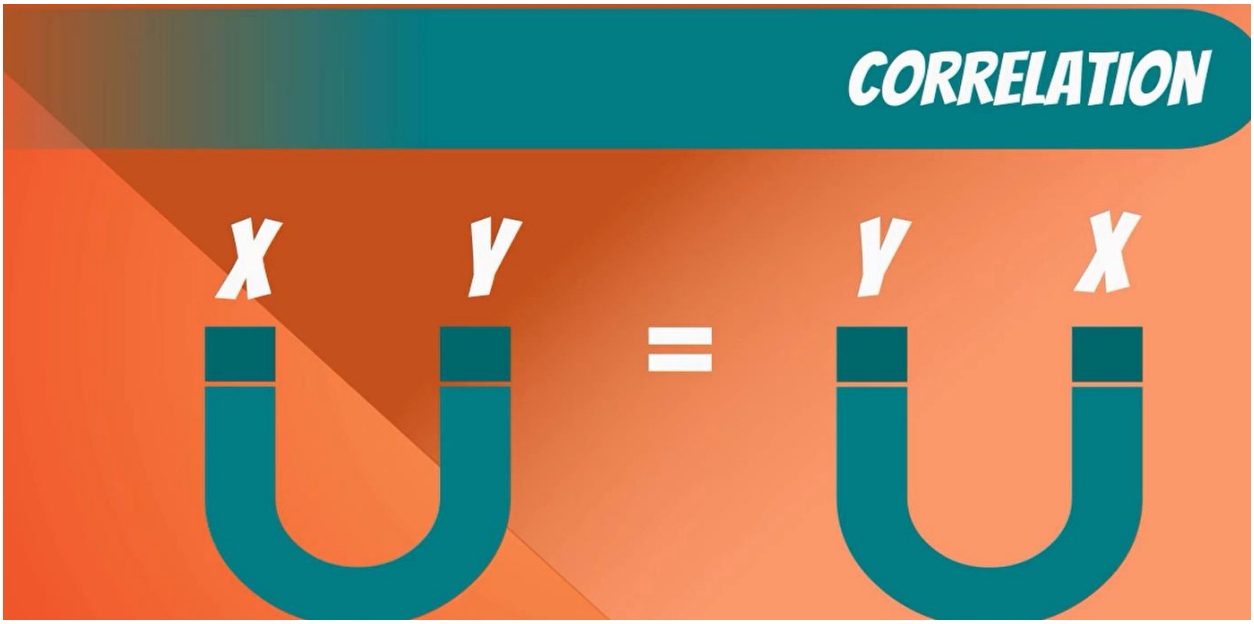
There is a STRONG relationship between the two variables

365√DataScience






NEGATIVE CORRELATION
Perfect negative correlation of - 1
Imperfect negative correlation: (- 1,0)


CORRELATION
X
Y
=
Y
X