



## M6-02: Correlation

Part of the "Towards Machine Learning" Learning Badge

Video Walkthrough: <https://discovery.cs.illinois.edu/m6-02/>

### Scatter Plots: Python Documentation

Using Python, we can easily create scatter plots with datasets:

<b>Python:</b>	<p><b>General Syntax:</b></p> <pre>df.plot.scatter()</pre> <p><b>Options:</b></p> <pre>x = 'Column Name' y = 'Column Name'</pre> <p><b>Example Shown:</b></p> <pre>df = pd.read_csv('diamonds.csv') df.plot.scatter(x='carat', y='price')</pre>	
----------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

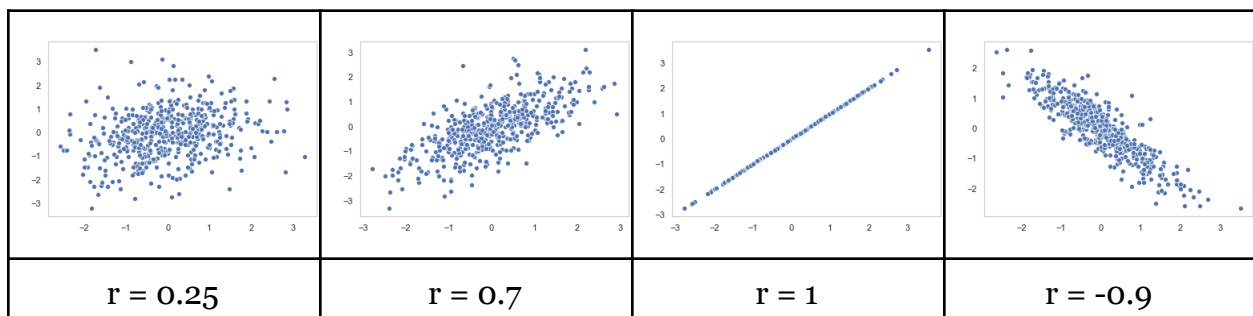
**CORRELATION COEFFICIENT (  $r$  )** – measures the strength of the **linear** association between two variables (X and Y). It measures how tightly points are clustered around a line. (It does not measure clustering around a curve). It is relevant when the scatter plot forms a linear trend.

**Puzzle #1:** Draw a few examples of scatterplots where  $r$  would be an appropriate summary statistic.

**The correlation coefficient is always between  $-1$  and  $1$ .**

The closer the points hug a line with a positive slope, the closer  $r$  is to  $+1$ . The closer the points hug a line with a negative slope the closer  $r$  is to  $-1$ . If there is no association between  $x$  and  $y$  then the correlation coefficient is  $0$  and the scatter plot has no linear pattern.

- ★ A correlation of  $1$  or  $-1$  means you can perfectly predict one variable knowing the other.
- ★ A correlation of  $0$  means that knowing one variable gives you no information about the other.





## M6-02: Correlation

Part of the "Towards Machine Learning" Learning Badge

Video Walkthrough: <https://discovery.cs.illinois.edu/m6-02/>

### How to mathematically calculate the correlation coefficient:

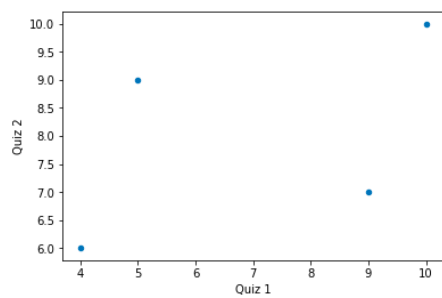
#### In words:

1. Convert  $x$ -values and  $y$ -values to standard units ( $z$ -scores).  $Z$ -scores tell you how many SDs a value is above or below average
2. Multiply each  $x$ -value (in standard units) by each corresponding  $y$ -value (in standard units)
3. The correlation coefficient is the average of the products.

#### In symbols:

**Puzzle #2:** Using the dataset below, calculate the correlation coefficient ( $r$ ).

Quiz 1	Quiz 2	Zx	Zy	Zx*Zy	
10	10				<b>r = _____</b>
9	7				
5	9				
4	6				

	Average	SD	
Quiz 1			 <pre>df.plot.scatter(x='Quiz 1', y='Quiz 2')</pre>
Quiz 2			