CORRELATION, COVARIANCE, AND INDEPENDENCY.

CORRELATION

- Correlation is a measure of association of two variables.

- There are several definitions for correlation. The most important are:
  - Pearson (or linear) correlation (numeric data only)
  - Spearman (or rank) correlation (numeric and ordinal data)

- If we have a data set with more than two variables, we compute the correlation for each pair of variables.

- We speak about the correlation between two variables. That is, correlation between X and Y is the same as correlation between Y and X. It is a symmetric quantity.

CORRELATION

The correlation only make sense for two variables that are part of a multidimensional dataset, so that for each each observation/ dataset row/ multidimensional datapoint, we have values for both variables.

X	Y	Z
4.428183	16.76798	21.19616
4.266916	15.43106	19.69797
2.937835	12.39779	15.33563
1.754835	5.375972	7.130807
7.681462	26.95491	34.63638
1.328777	8.24138	9.570158
2.569961	8.249133	10.81909
6.596999	21.89212	28.48911
1.357858	5.798073	7.155931
2.426349	8.780154	11.2065
0.418517	1.651663	2.07018
8.485556	29.857	38.34256
9.562872	32.2684	41.83127
0.006935	4.273806	4.280741
5.211658	16.99656	22.20822
1.549561	10.45515	12.00471
3.302038	13.41002	16.71206
3.853931	12.75443	16.60836
0.513752	7.152144	7.665896
5.052185	18.52358	23.57576
9.608876	32.1961	41.80498
9.913533	33.19059	43.10412
9.595957	31.40554	41.00149
1.319216	5.724218	7.043434
5.205118	18.55512	23.76024
6.566597	20.61485	27.18144
7.878595	23.97968	31.85827
4.865895	17.17571	22.04161

For this dataset, we can compute the correlation:

- between, X and Y
- between X and Z
- between Y and Z

For these datasets, it does not make sense to talk about correlation, or association.

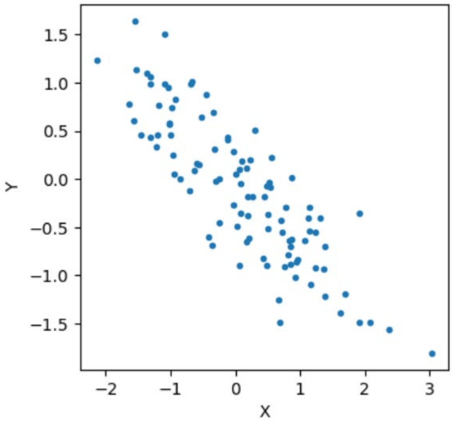
Values1
9
0
8
0
0
5
5
0
2
1
1
8
6
4
6
5
5
1
1
8

Values2
-2.26715
-0.62968
0.067861
0.751434
-0.86123
-0.99071
1.384102

INTUITION

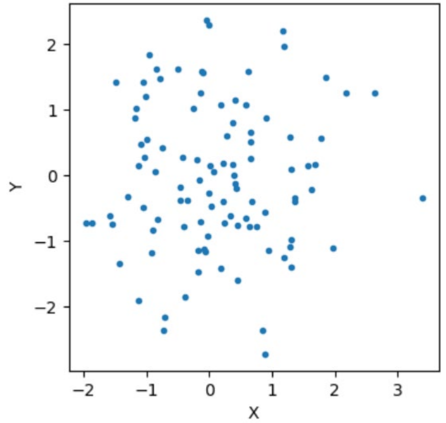
Suppose that you have a data set with two numeric variables. Let’s make a scatter plot

1



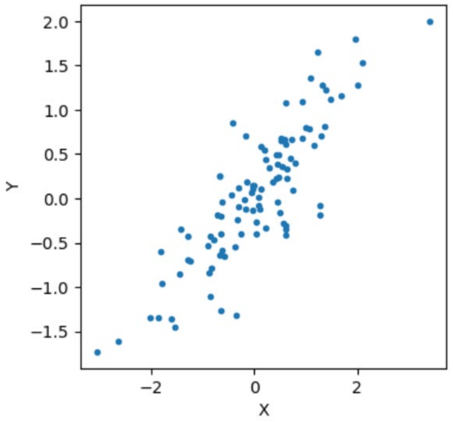
Negative correlation

2



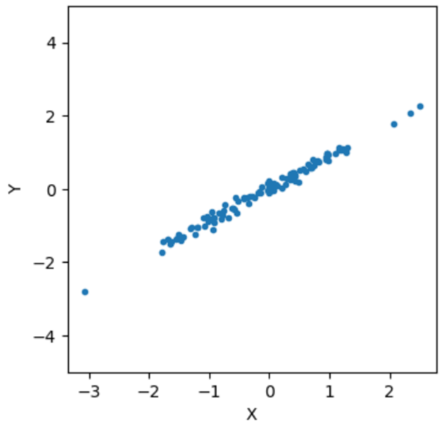
Very small or
no correlation

3



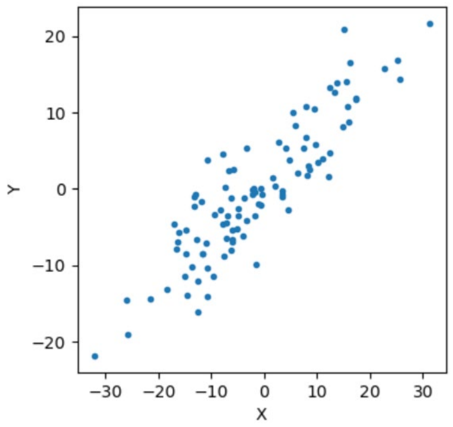
Positive correlation

4



Almost perfect correlation

5



Similar correlation to figure 3,
but more absolute (look at
values of X and Y variables)

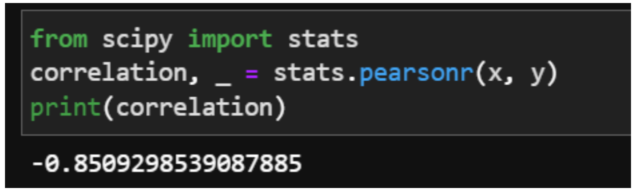
More covariance

PROPERTIES OF CORRELATION

- The correlation between two variables is a number between -1 and 1
- It is usually denoted by the symbol ***ρ*** (*rho*), or with the letter ***r***
- If nothing is specified, it means Pearson correlation
- The correlation is computed via a complicated formula, but also very easily using python code

Using python

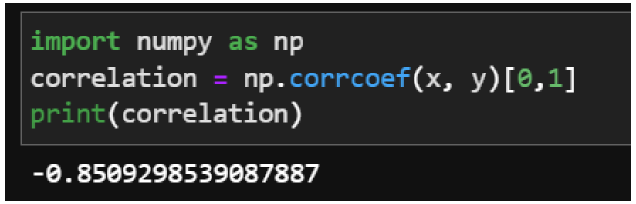
Method 1



```
from scipy import stats
correlation, _ = stats.pearsonr(x, y)
print(correlation)
```

```
-0.8509298539087885
```

Method 2



```
import numpy as np
correlation = np.corrcoef(x, y)[0,1]
print(correlation)
```

```
-0.8509298539087887
```

Using formula (for Pearson correlation)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

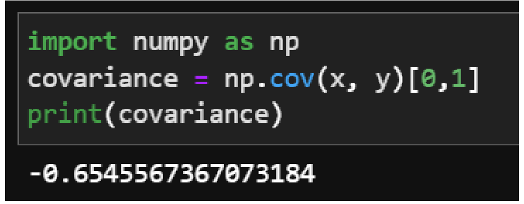
CORRELATION AND COVARIANCE

- Let X and Y be two numeric variables, with correlation $\boldsymbol{\rho}$ and with standard deviations $\sigma_x$and σ_y respectively

- The covariance between x and y (denoted ($Cov_{x,y}$) is defined as

$$Cov_{x,y} = \boldsymbol{\rho}\ \sigma_x\ \sigma_y$$

Using python

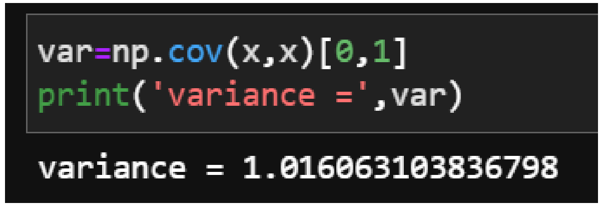


```
import numpy as np
covariance = np.cov(x, y)[0,1]
print(covariance)

-0.6545567367073184
```

- The covariance of a variable with itself is equal to the variance

$$Cov_{x,x} = \sigma_x^2$$

Using python



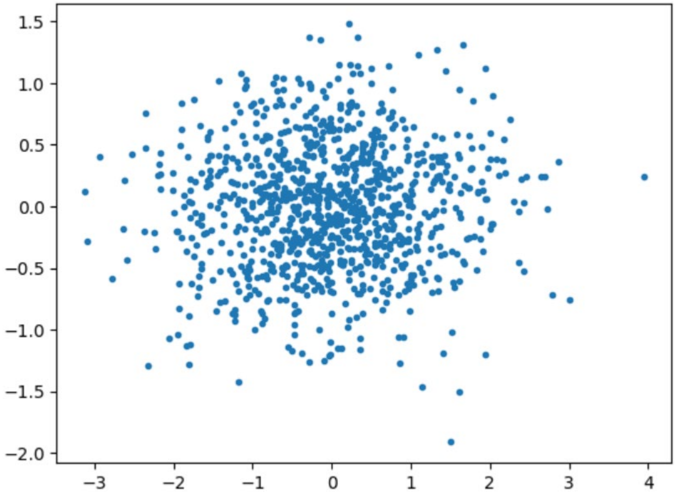
```
var=np.cov(x,x)[0,1]
print('variance =',var)

variance = 1.016063103836798
```

INDEPENDENCY

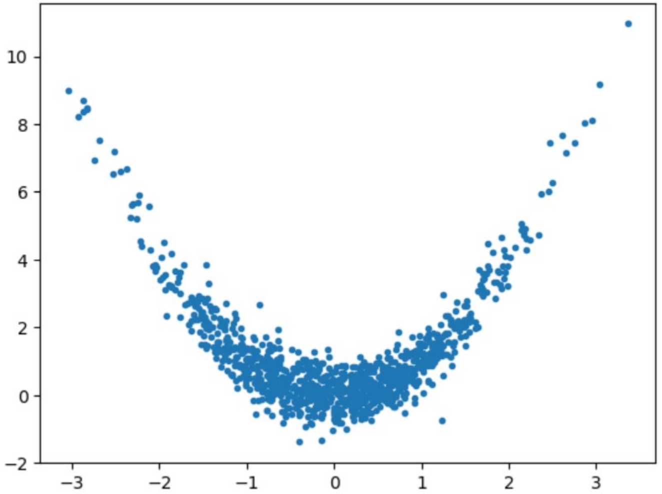
- Two variables are independent if knowing the value of one of them, does not give some indication about the value of the second.

- If two variables are independent, then the correlation must be zero
- The converse is NOT true

These variables are independent, and the correlation, of course, is zero



These variables are NOT independent, but the correlation is still zero



10
8
6
4
2
0
−2
−3
−2
−1
0
1
2
3