Applied Data Analytics

Statistics — Measures for bivariate data

Correlations

Hans-Martin von Gaudecker and Aapo Stenhammar

(Co)variance & Correlation

Variance:

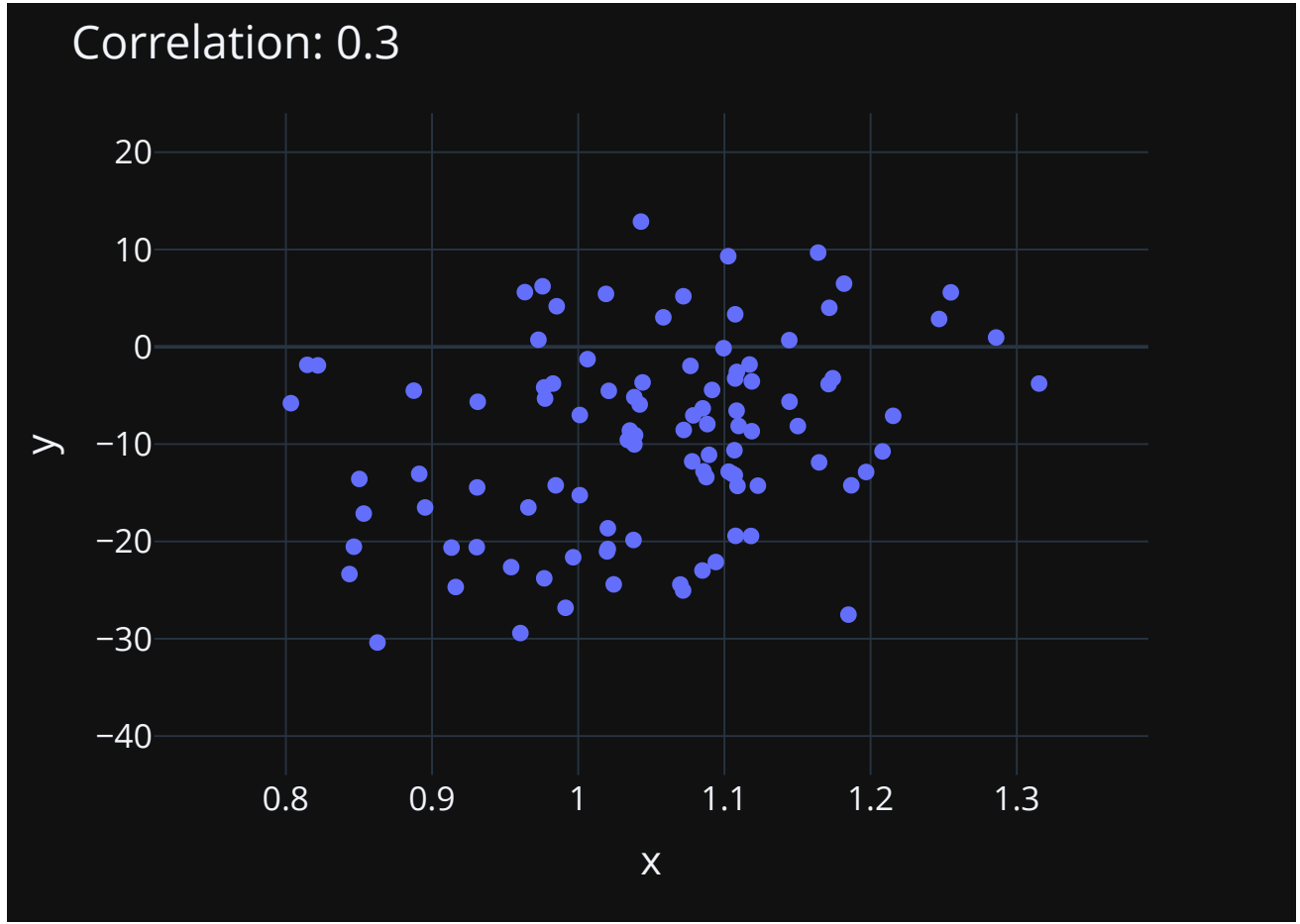
$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

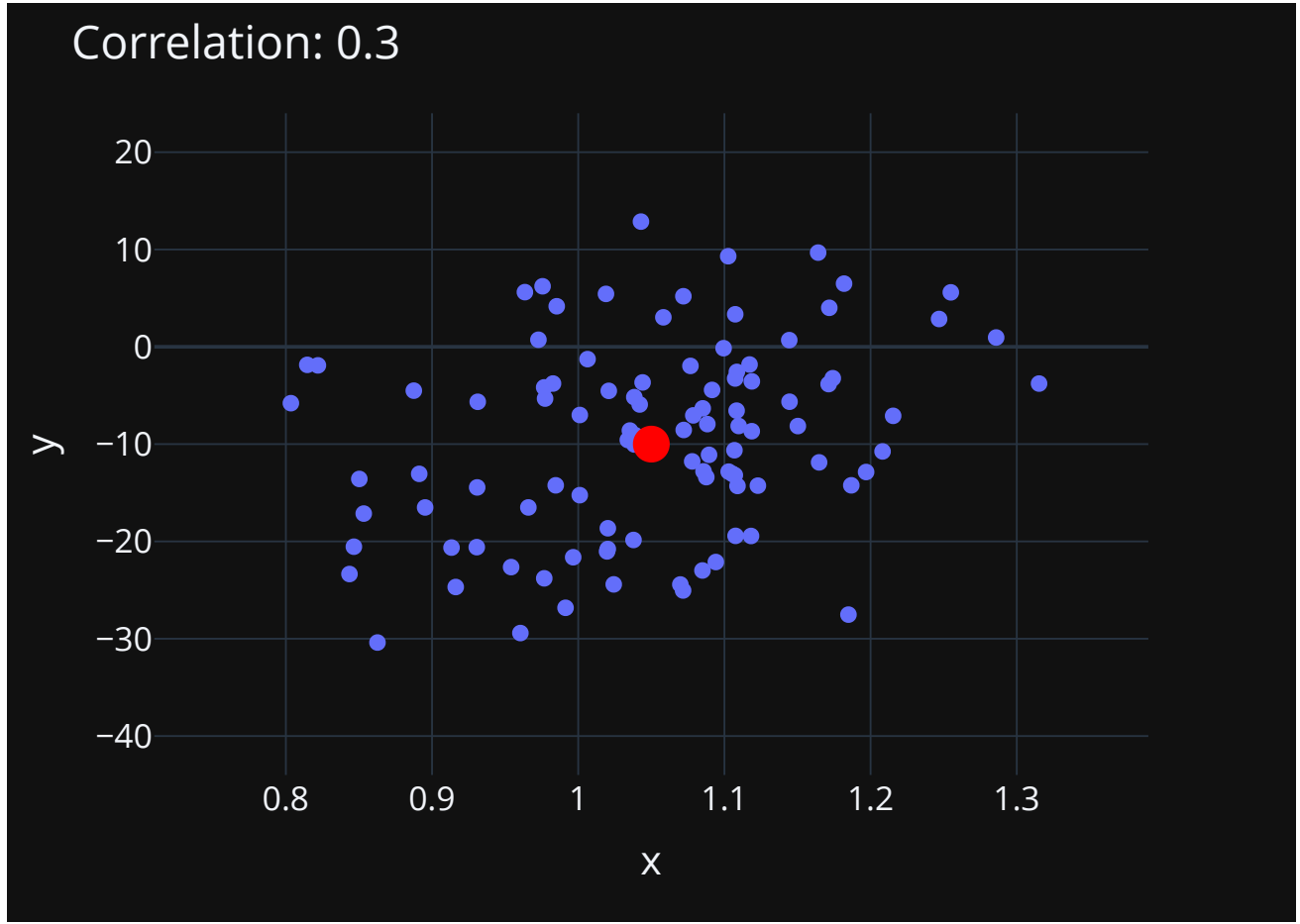
Covariance:

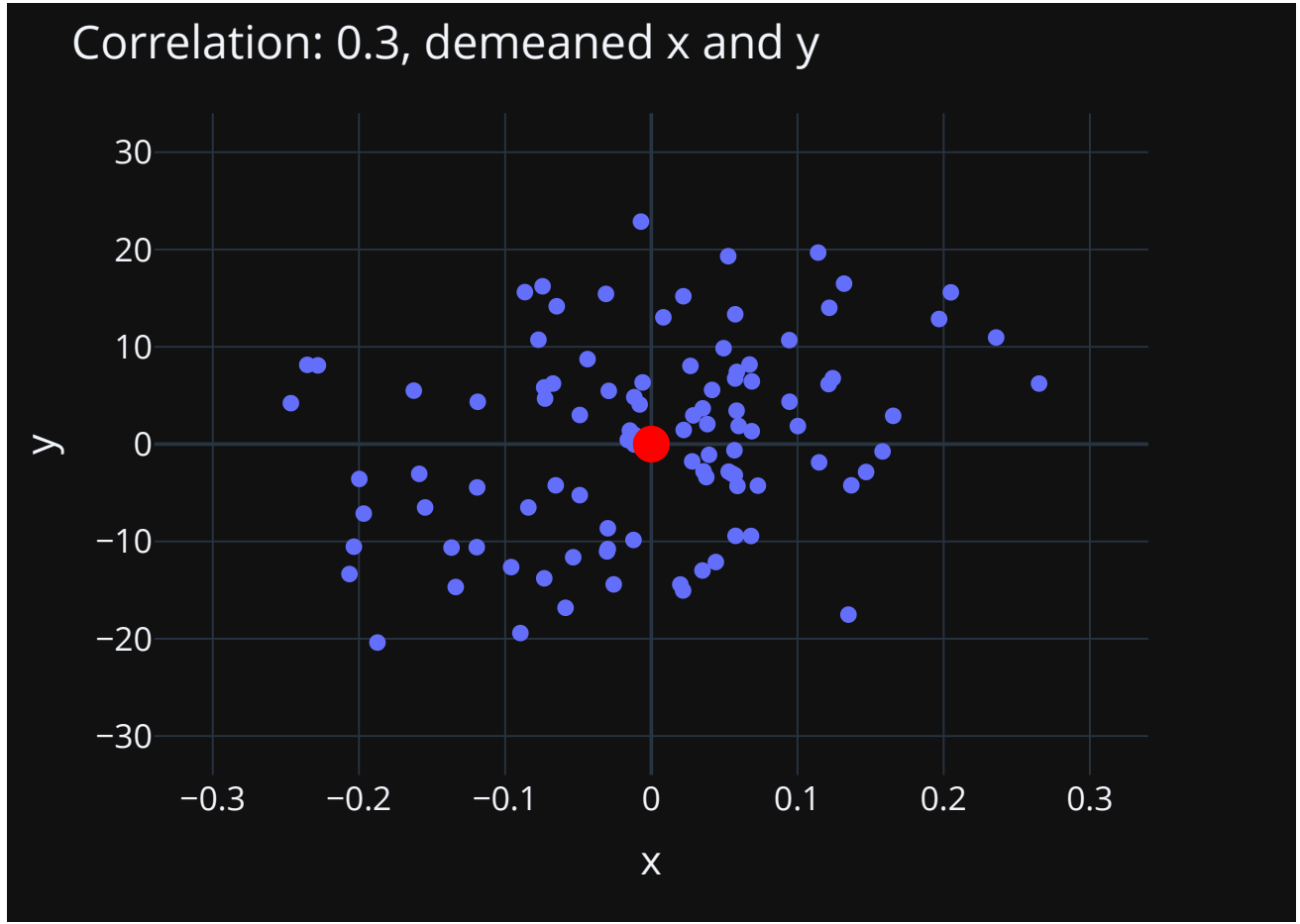
$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})$$

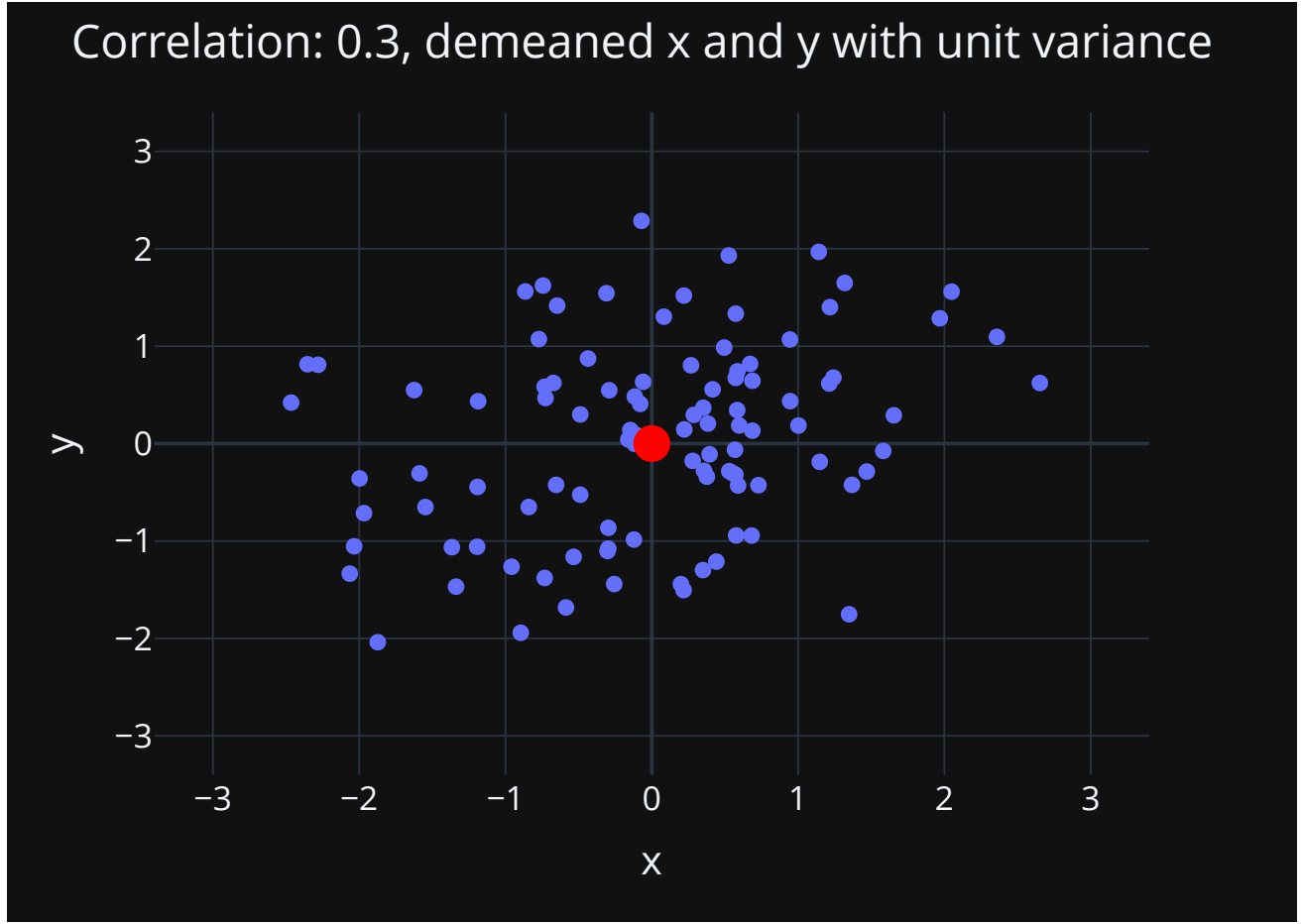
Correlation (Pearson's ρ):

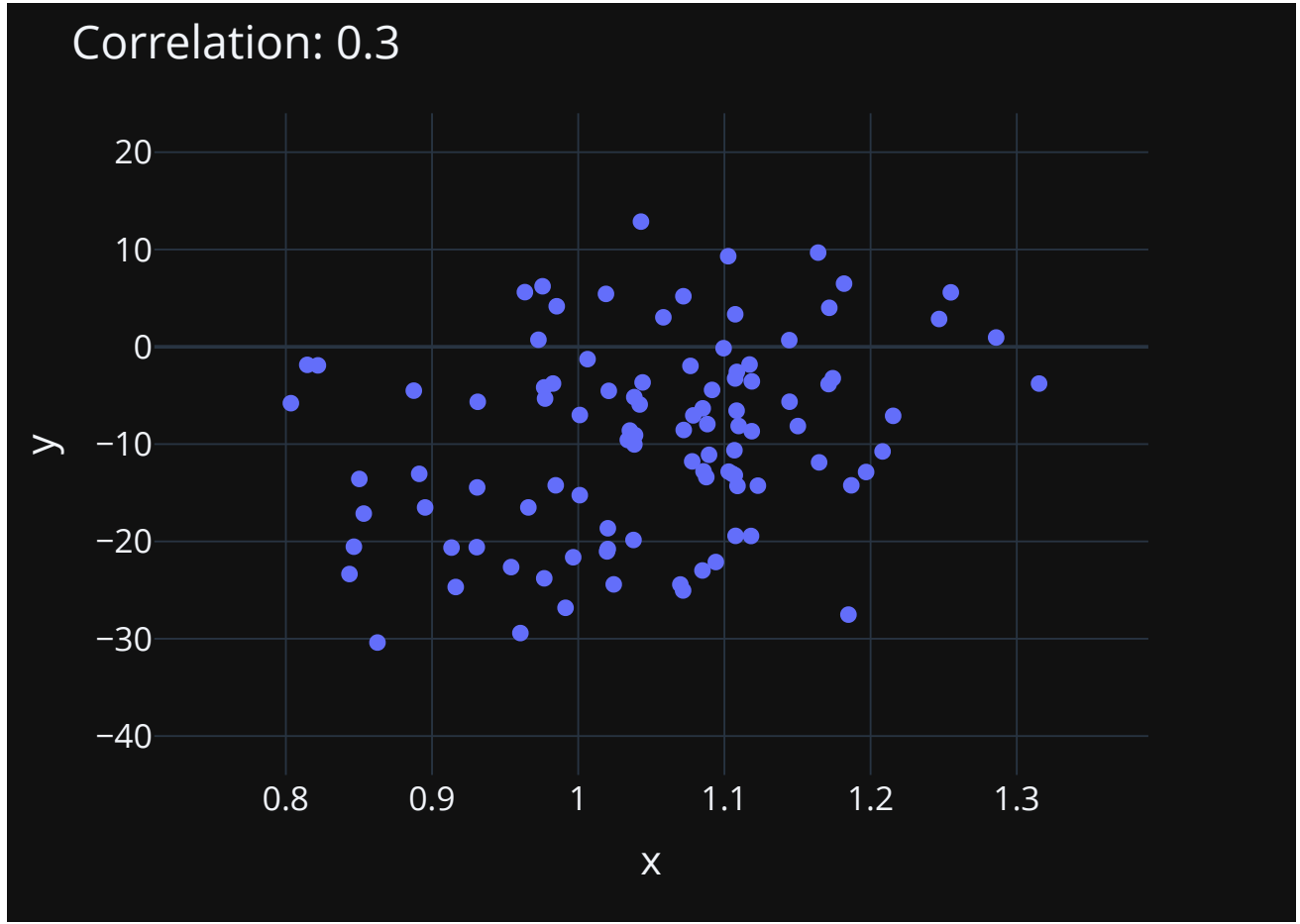
$$\rho_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}$$

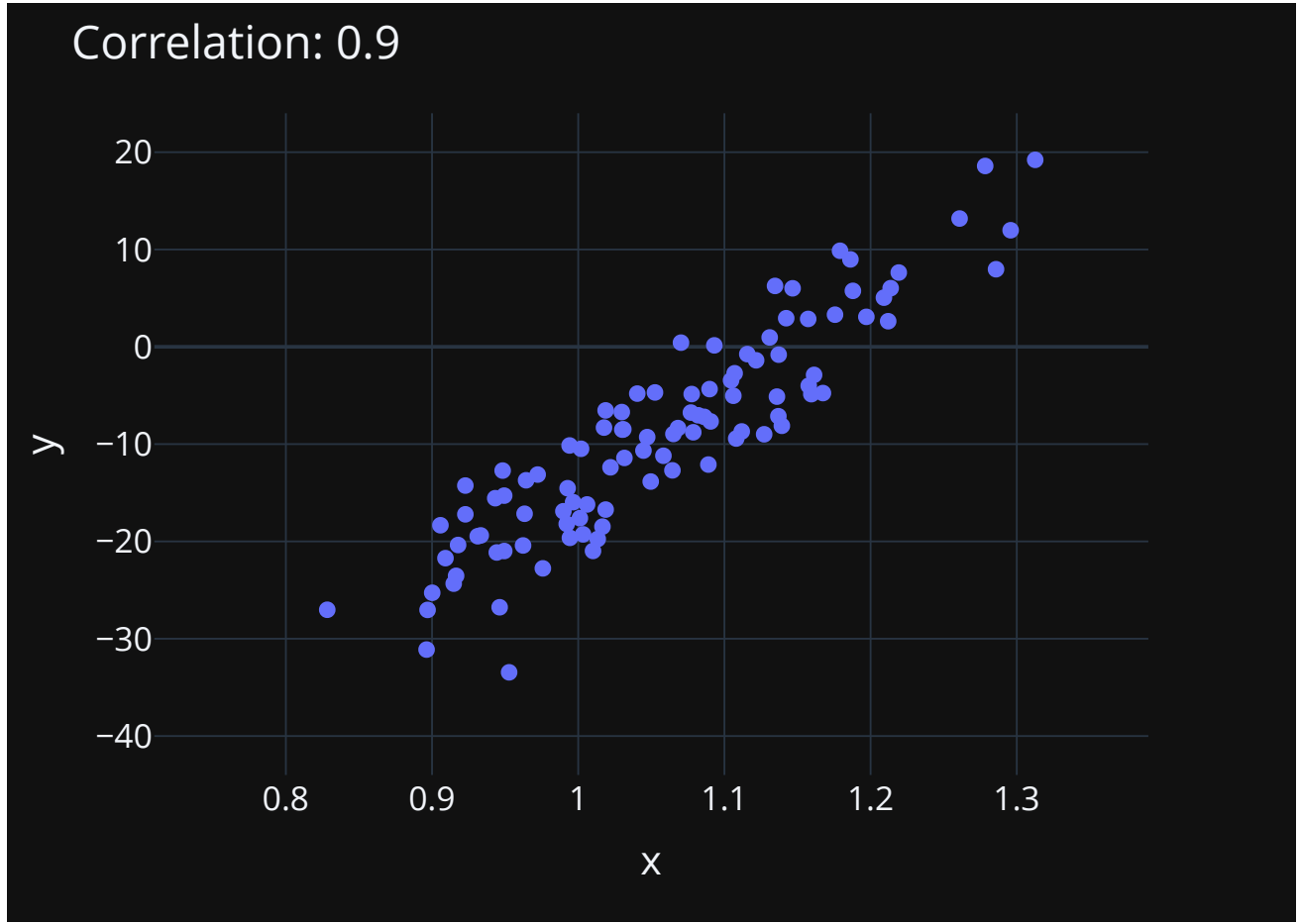

Correlation: 0.3
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x

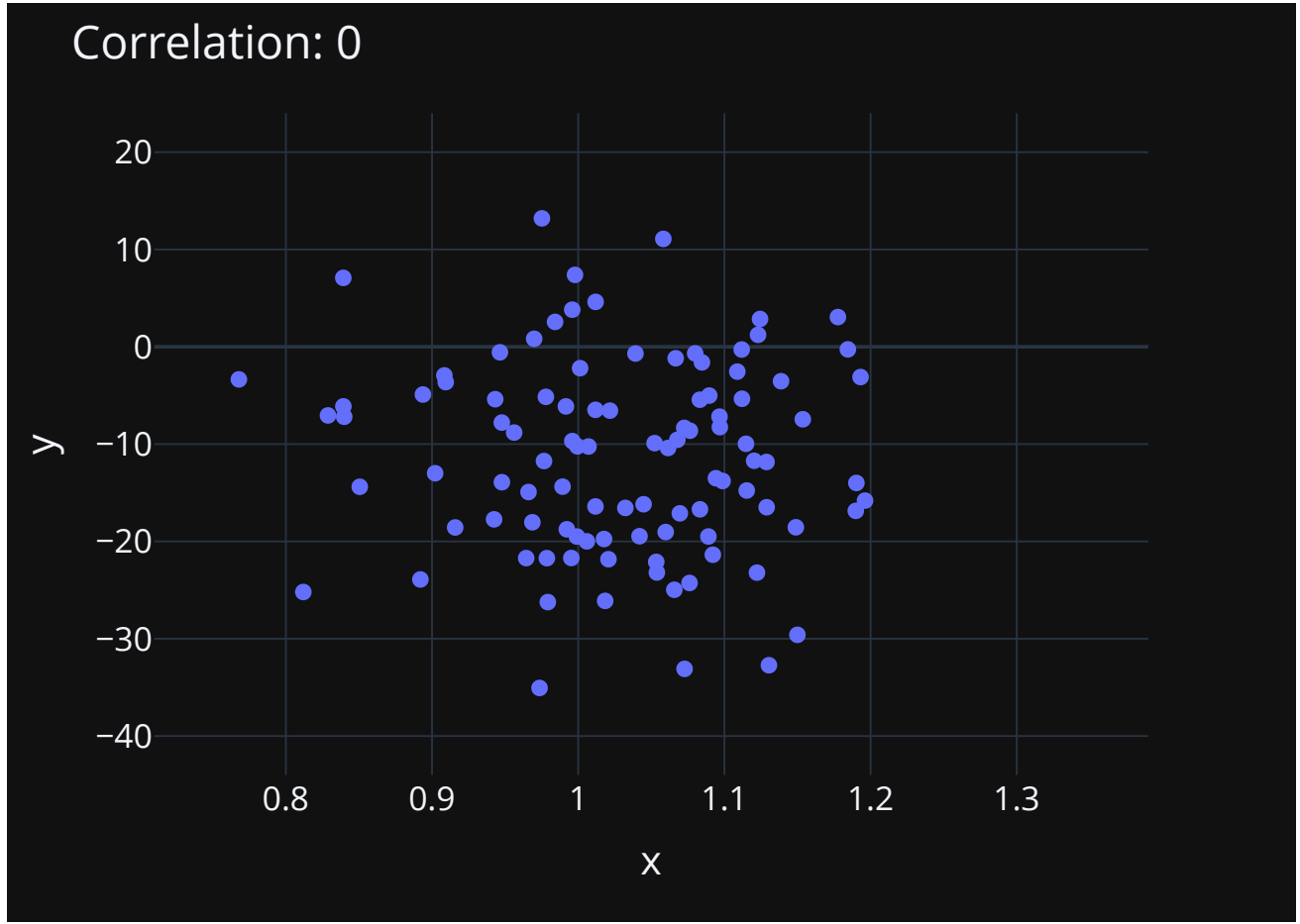

Correlation: 0.3
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x

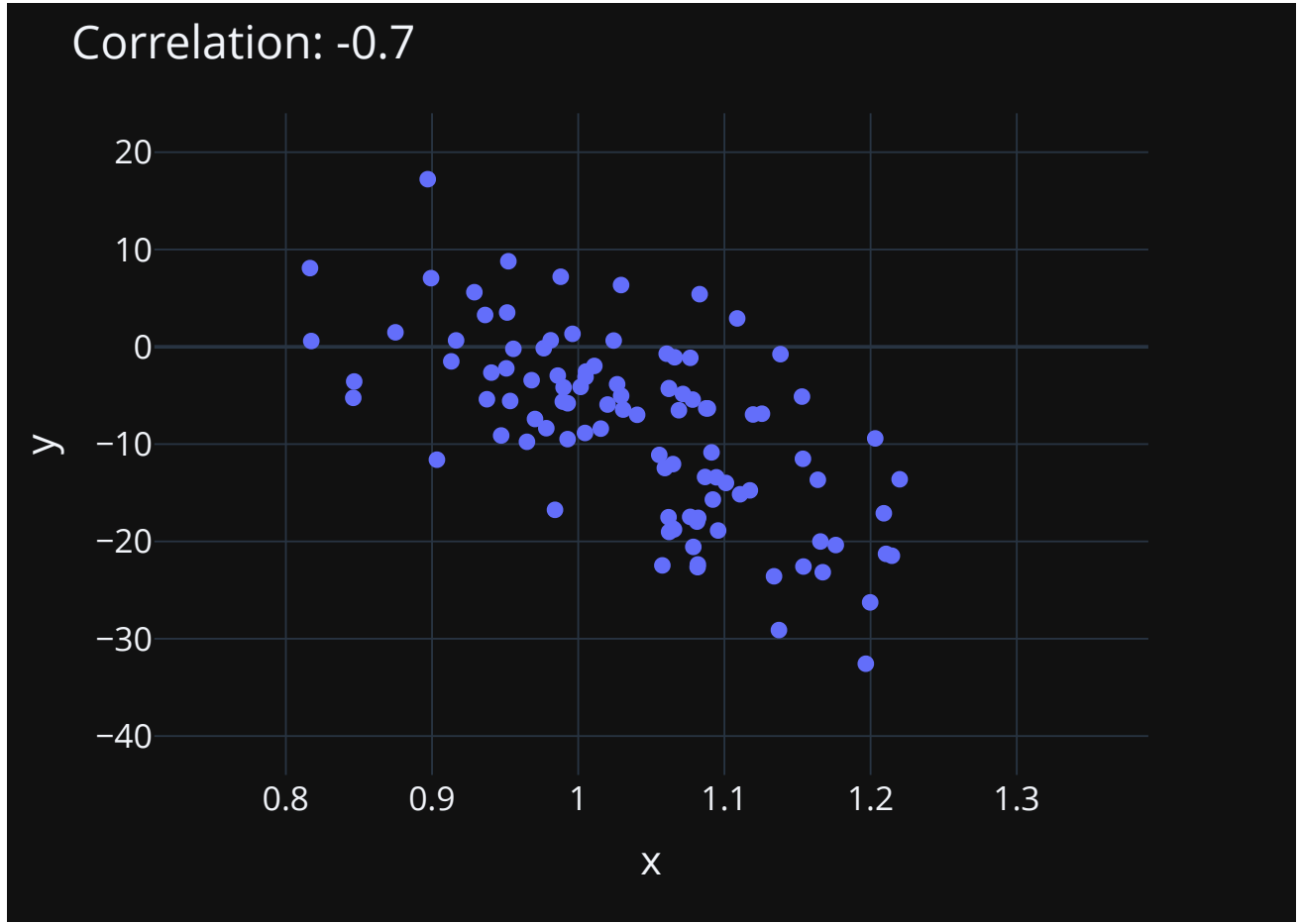

Correlation: 0.3, demeaned x and y
30
20
10
0
−10
−20
−30
y
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
x

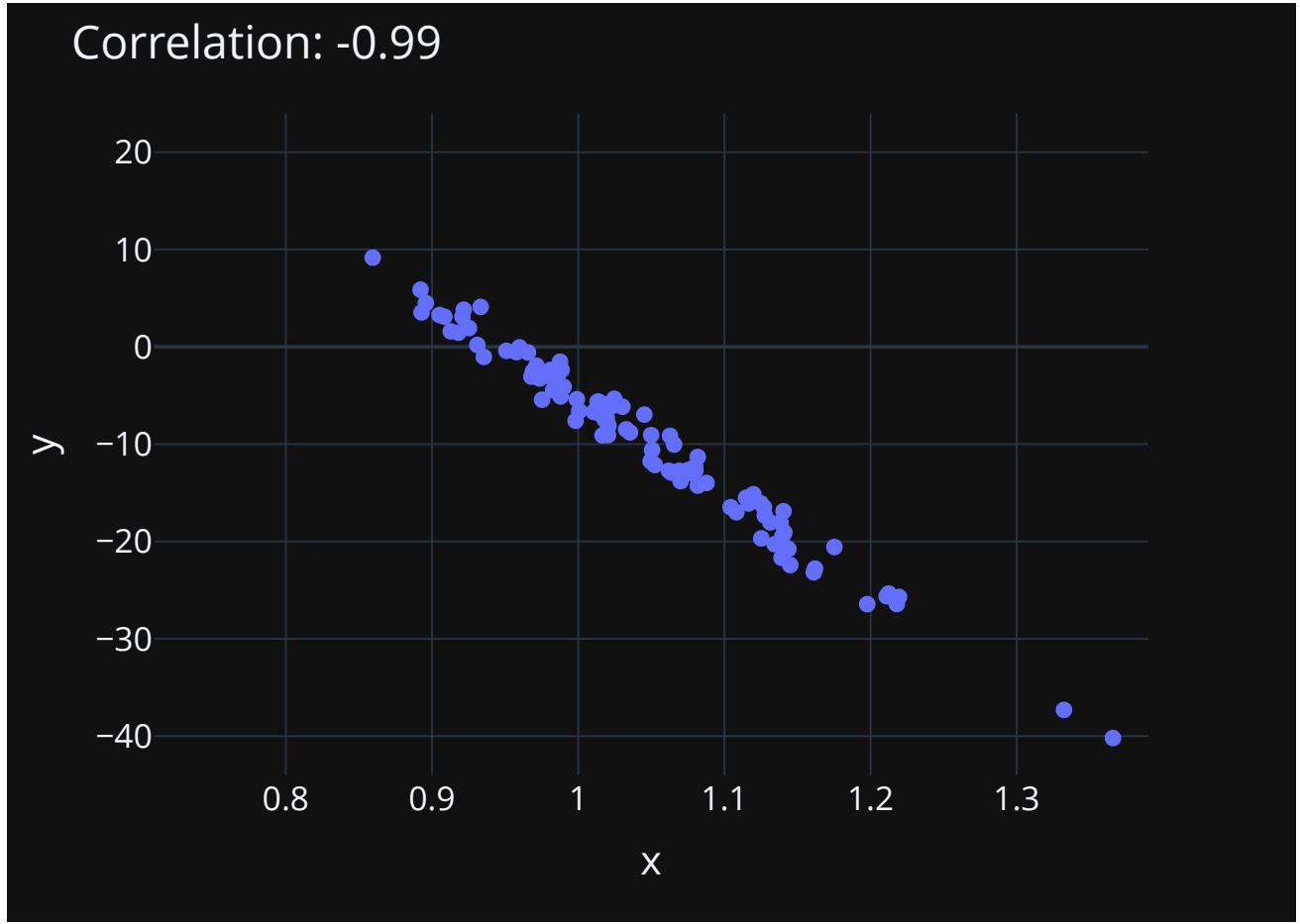

Correlation: 0.3, demeaned x and y with unit variance
3
2
1
0
−1
−2
−3
y
−3
−2
−1
0
1
2
3
x


Correlation: 0.3
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x


Correlation: 0.9
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x


Correlation: 0
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x


Correlation: -0.7
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x


Correlation: -0.99
20
10
0
−10
−20
−30
−40
y
0.8
0.9
1
1.1
1.2
1.3
x

Correlation: Properties

1. $-1 \leq \rho_{x,y} \leq 1$
2. Both x and y need to vary
3. -1 or 1 means perfect linear relationship, i.e., all points on a straight line
   (2. guarantees that slope is nonzero and finite)