Applied Data Analytics

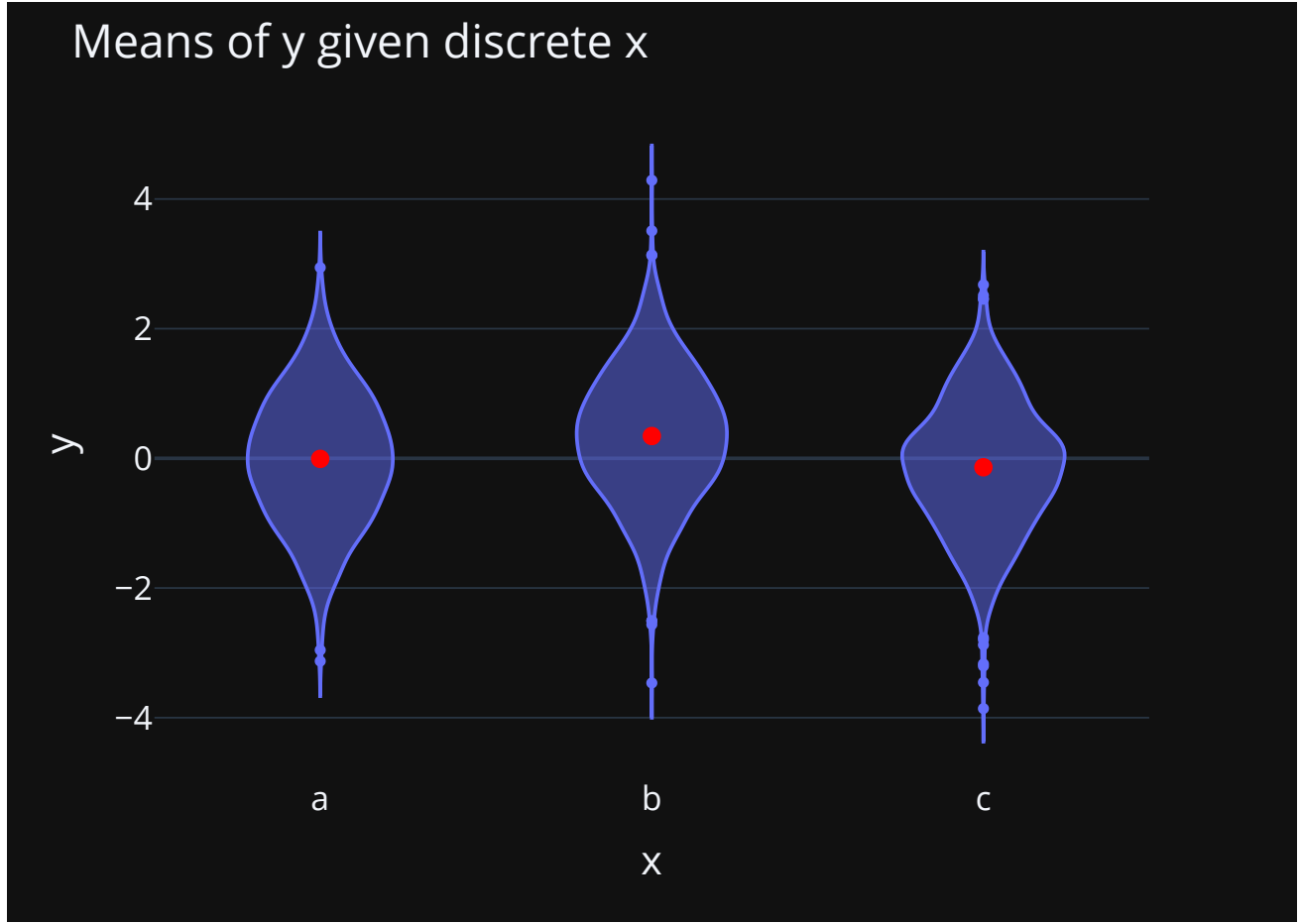
Statistics — Measures for bivariate data

Regression methods: Intuition

Hans-Martin von Gaudecker and Aapo Stenhammar

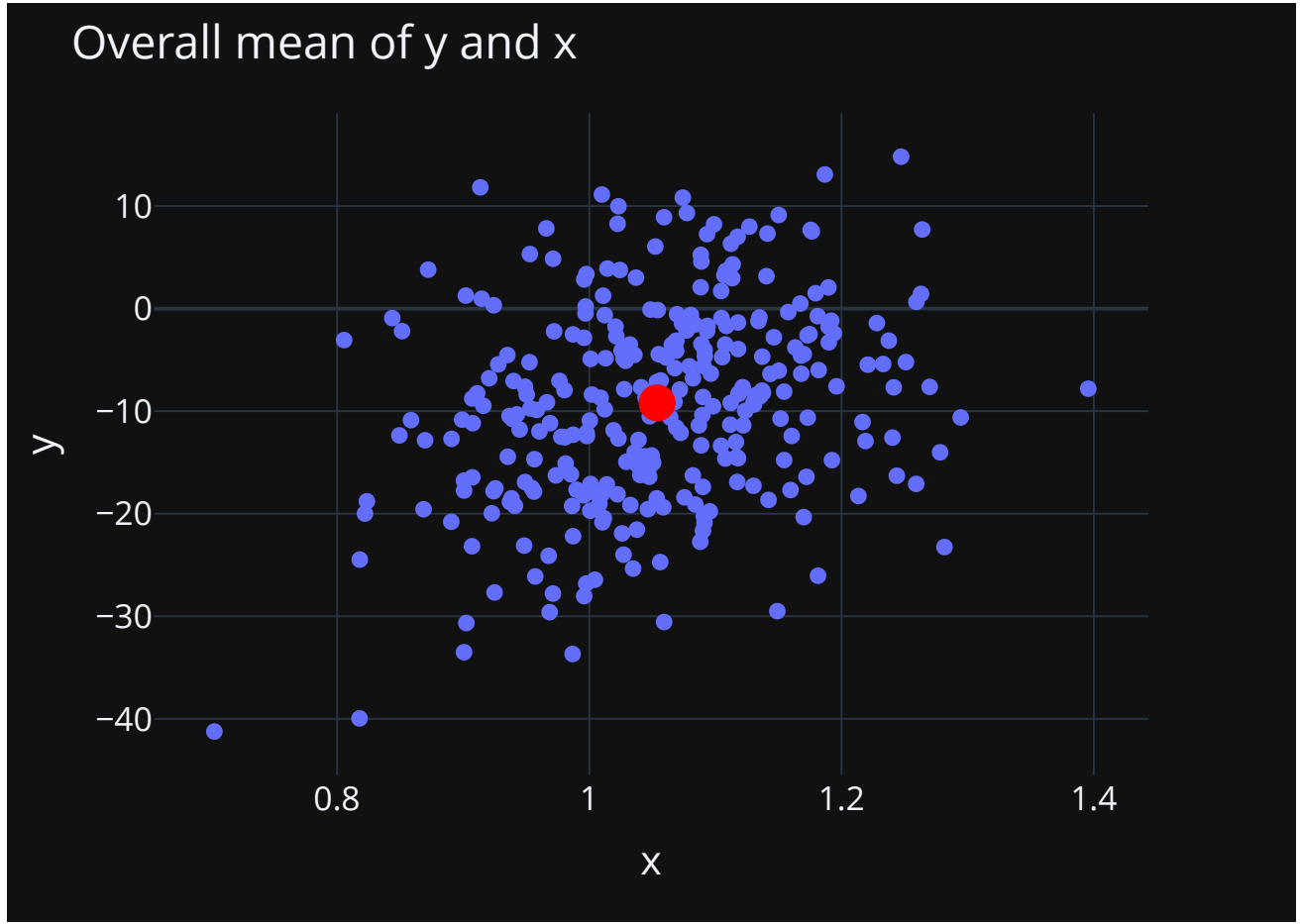
Conditional mean function

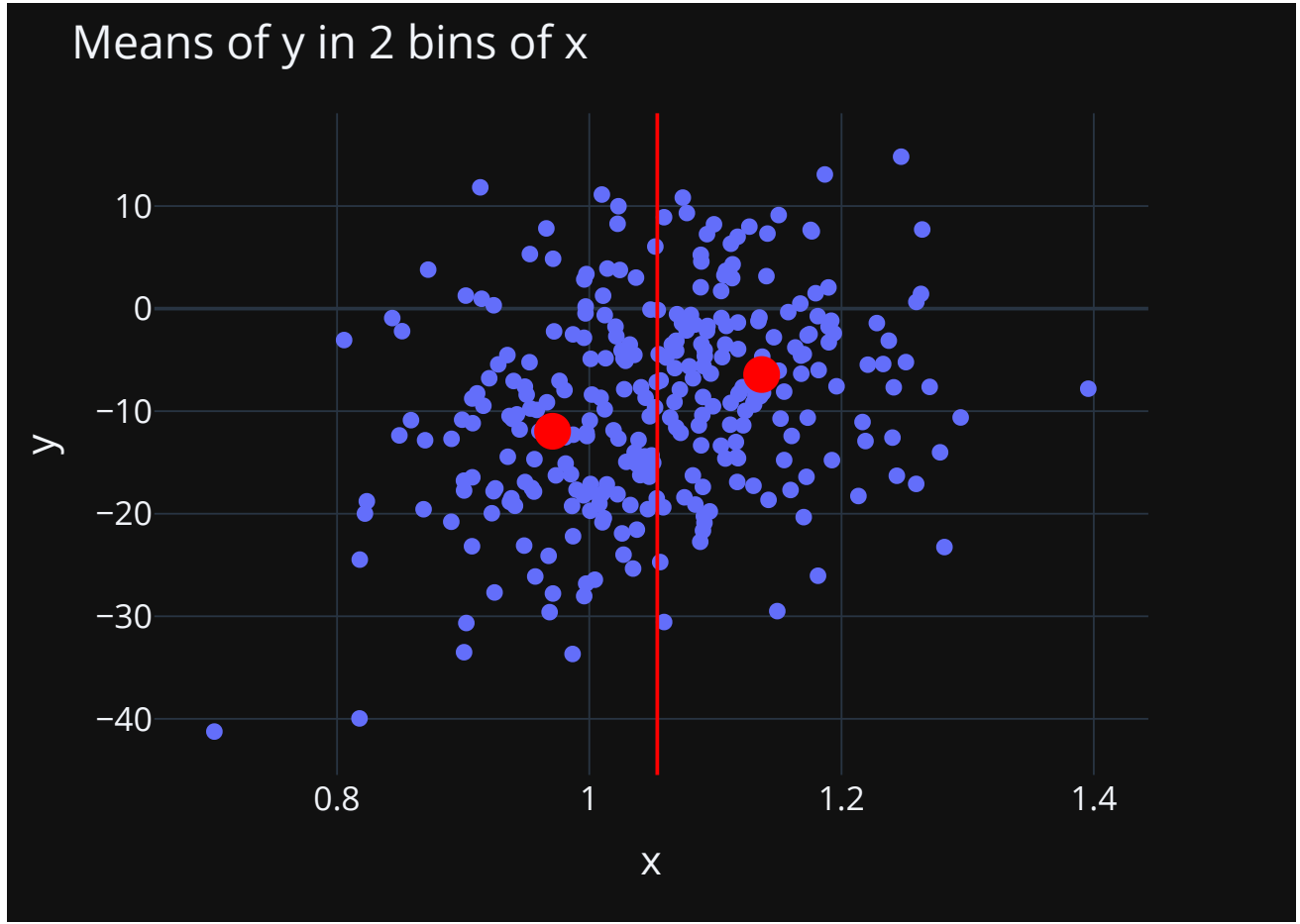
- Given that I know X, what is the mean value of Y?
- We saw this for X discrete and measured Y (GDP per capita for different countries)
- Now look $\bar{Y}|X = x$
  - Discrete X: `df.groupby("x").mean()["y"]`

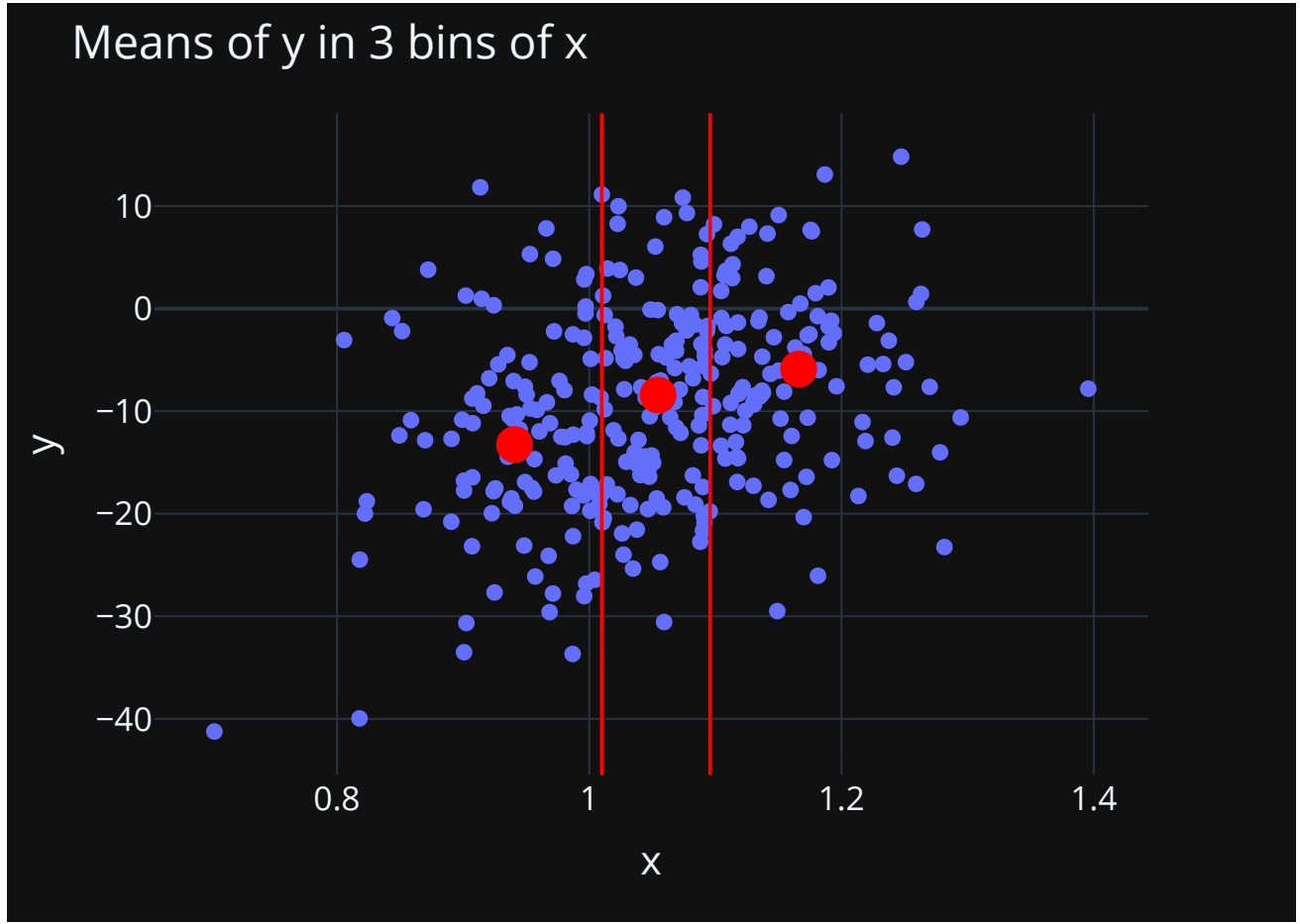

Means of y given discrete x
4
2
0
−2
−4
y
a
b
c
x

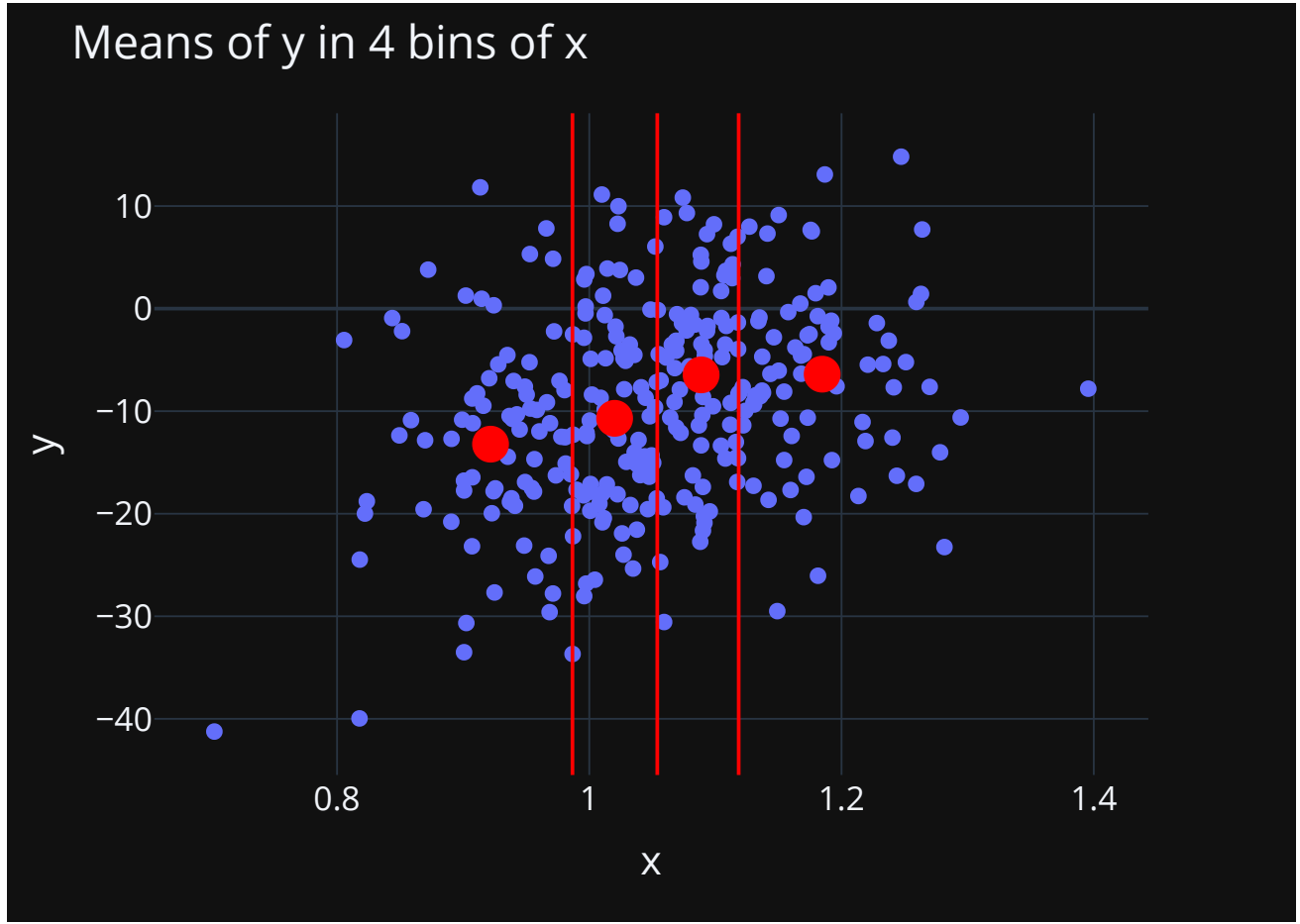
Conditional mean function

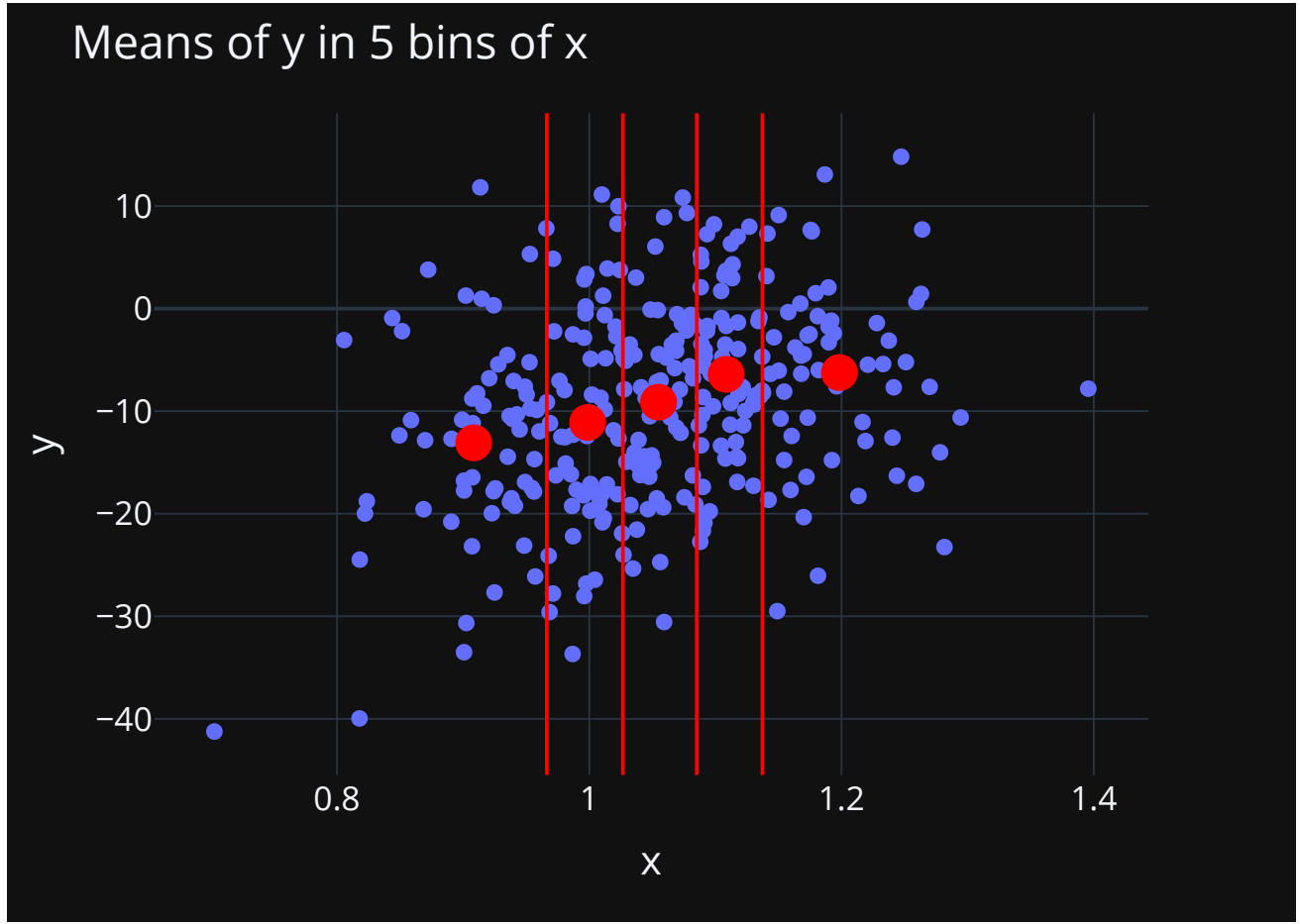
- Given that I know X, what is the mean value of Y?
- We saw this for X discrete and measured Y (GDP per capita for different countries)
- Now look $\bar{Y}|X = x$
  - Discrete X: `df.groupby("x").mean()["y"]`
  - Continuous X: Bin the data and calculate the mean for each bin

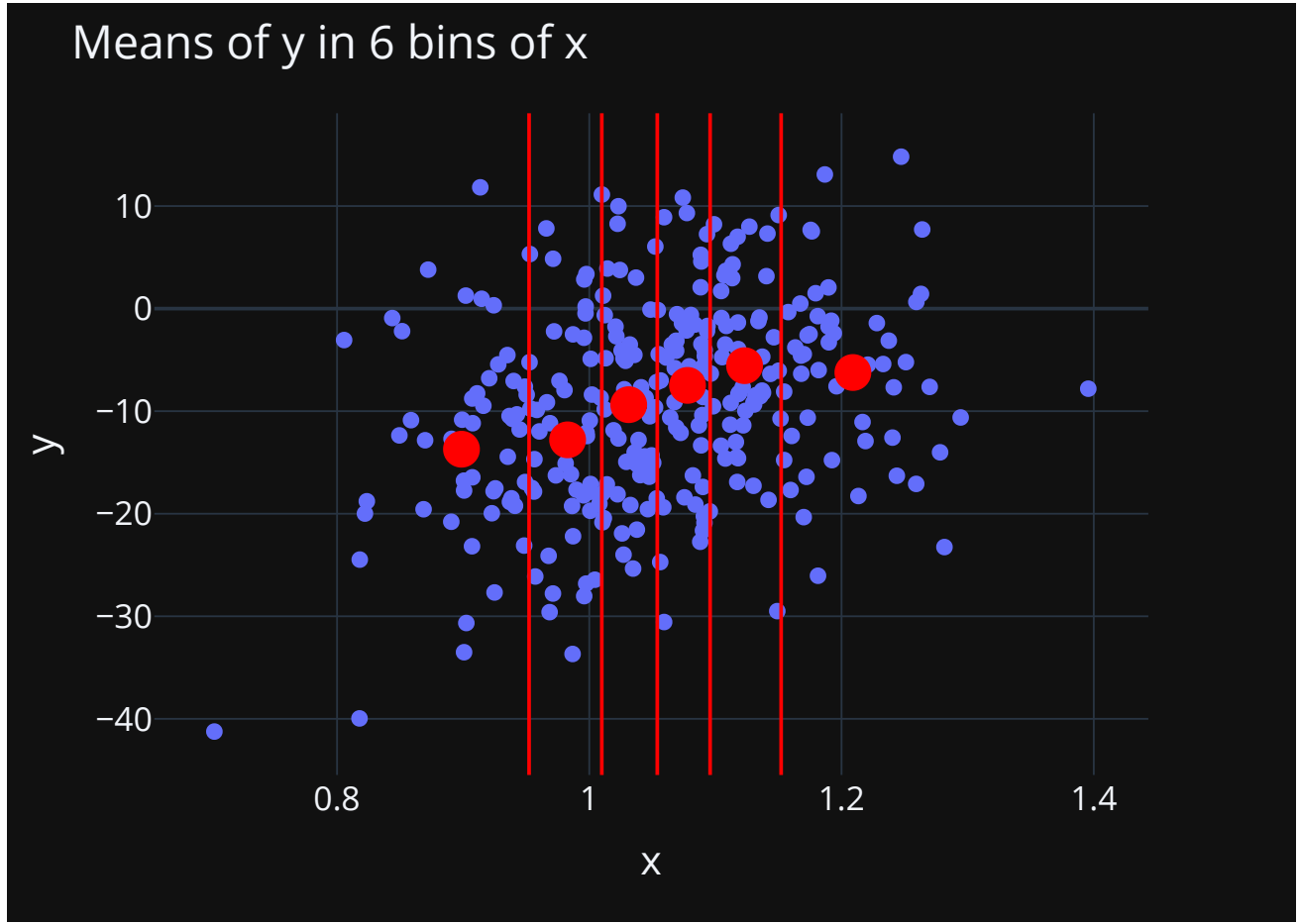

Overall mean of y and x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Means of y in 2 bins of x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Means of y in 3 bins of x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

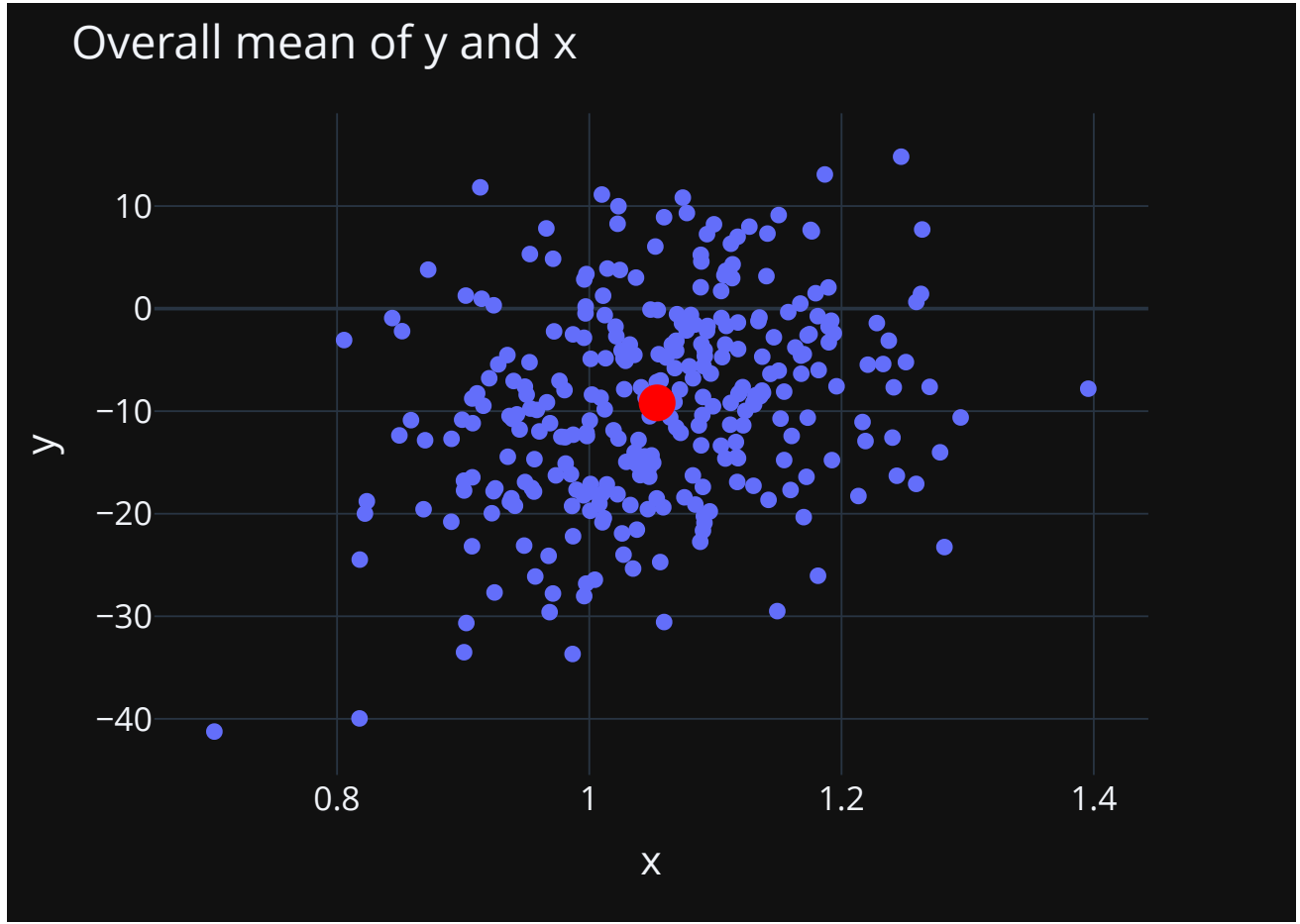

Means of y in 4 bins of x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

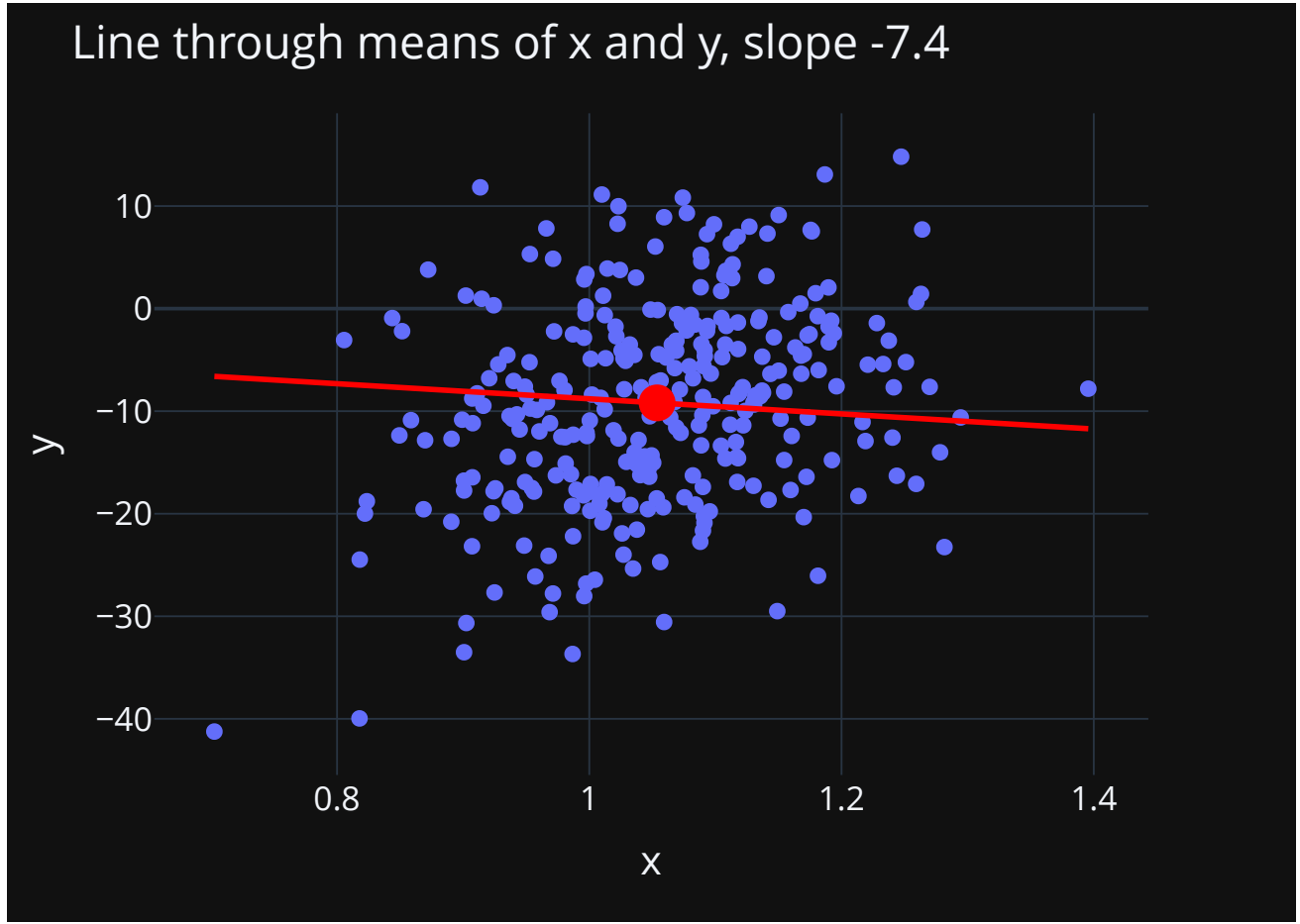

Means of y in 5 bins of x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

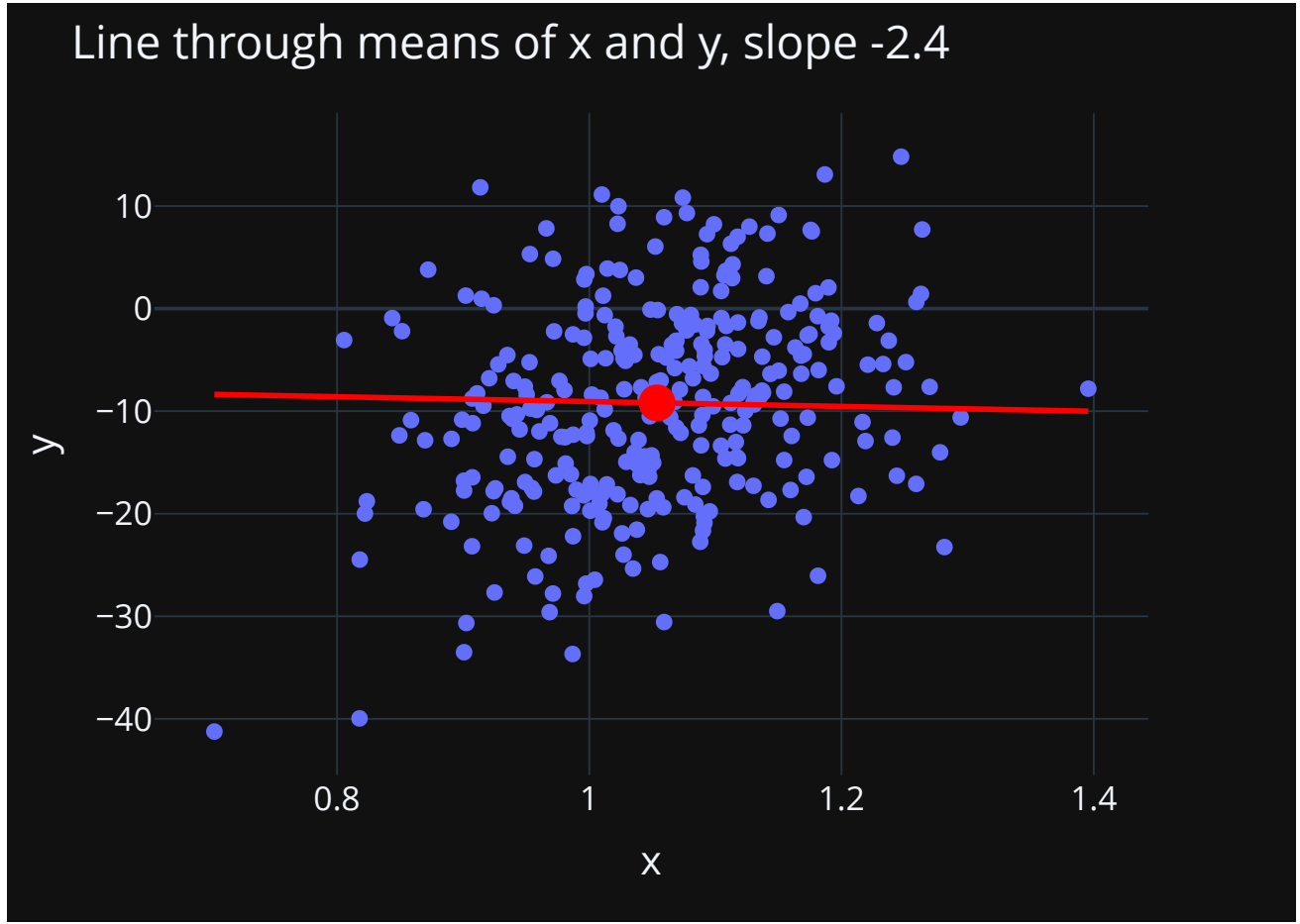

Means of y in 6 bins of x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

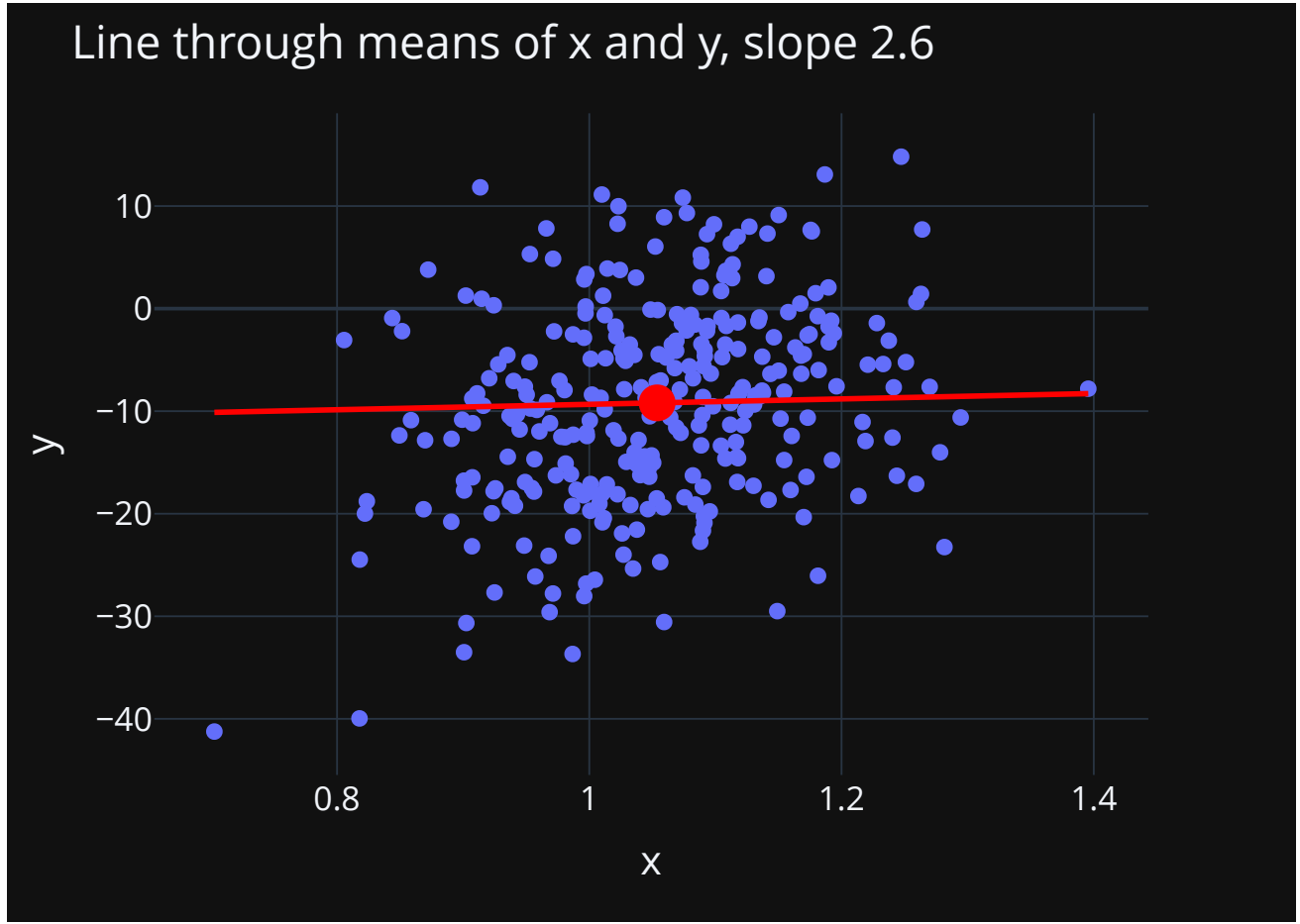
Continuous X, alternative

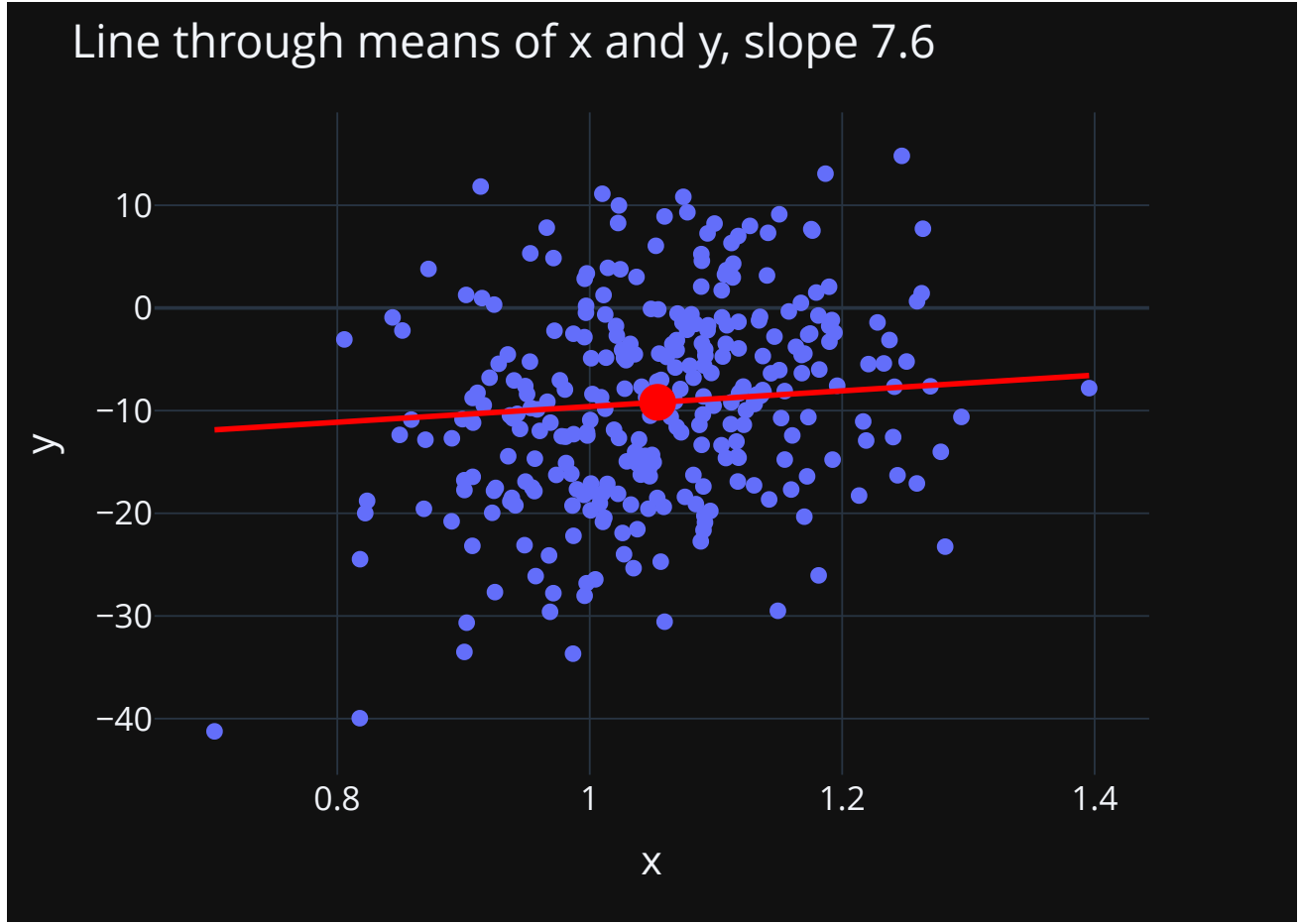
- Assume a functional form for the relationship between X and Y
- Linear relationship: $Y_i = \beta_0 + \beta_1 \cdot X_i + U_i$
- $\bar{Y}|X = x$ becomes $\beta_0 + \beta_1 \cdot x$

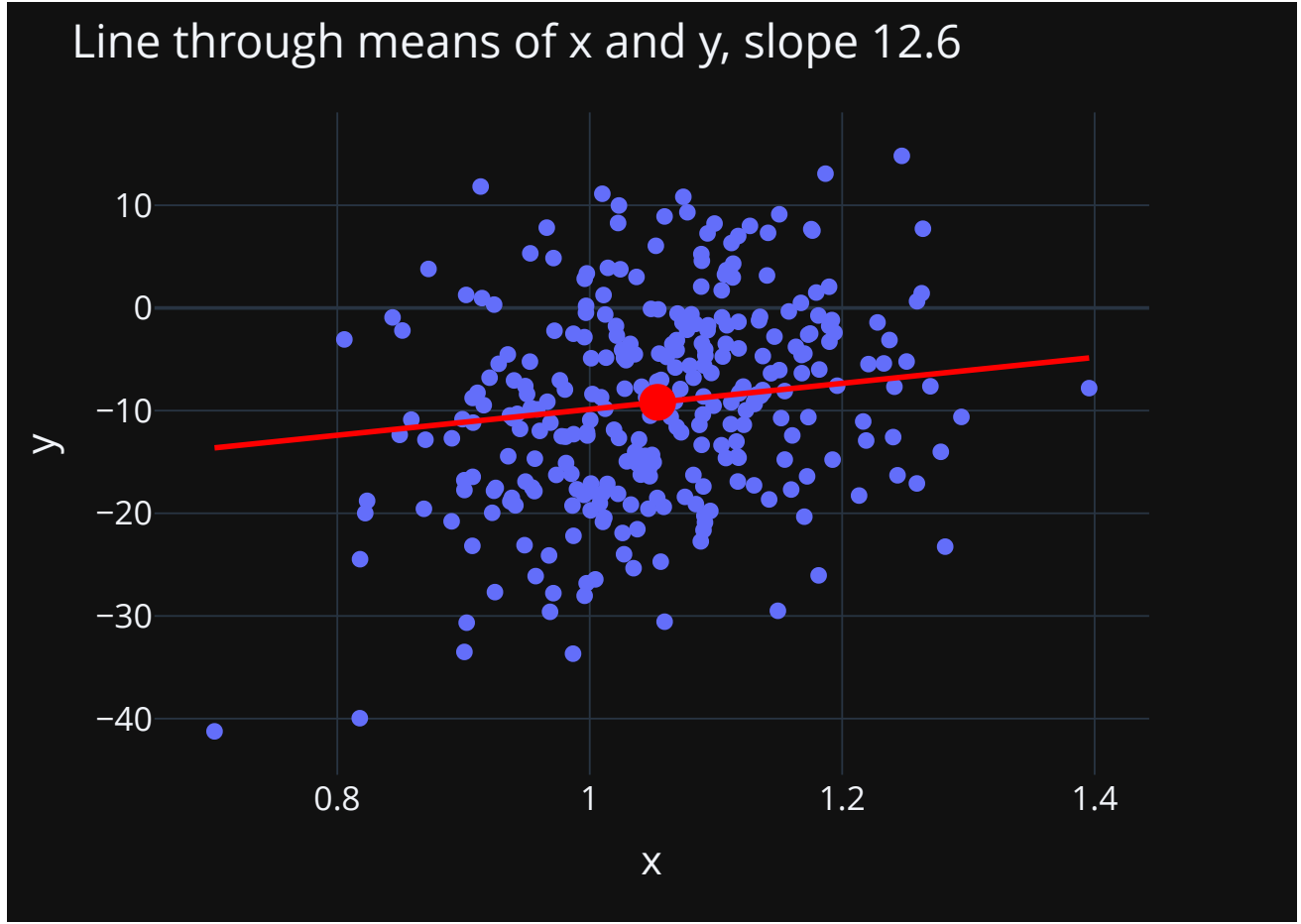

Overall mean of y and x
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

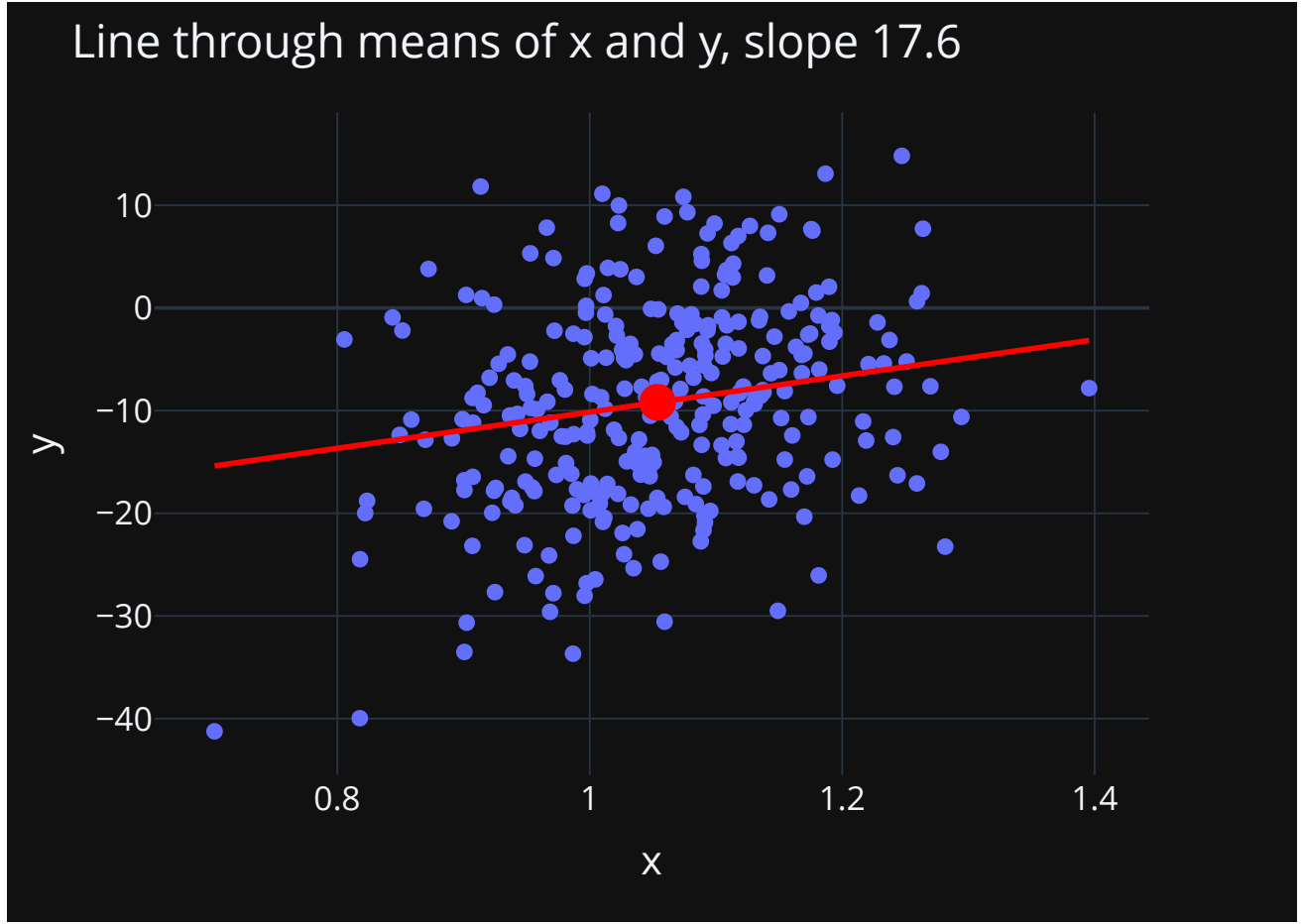

Line through means of x and y, slope -7.4
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

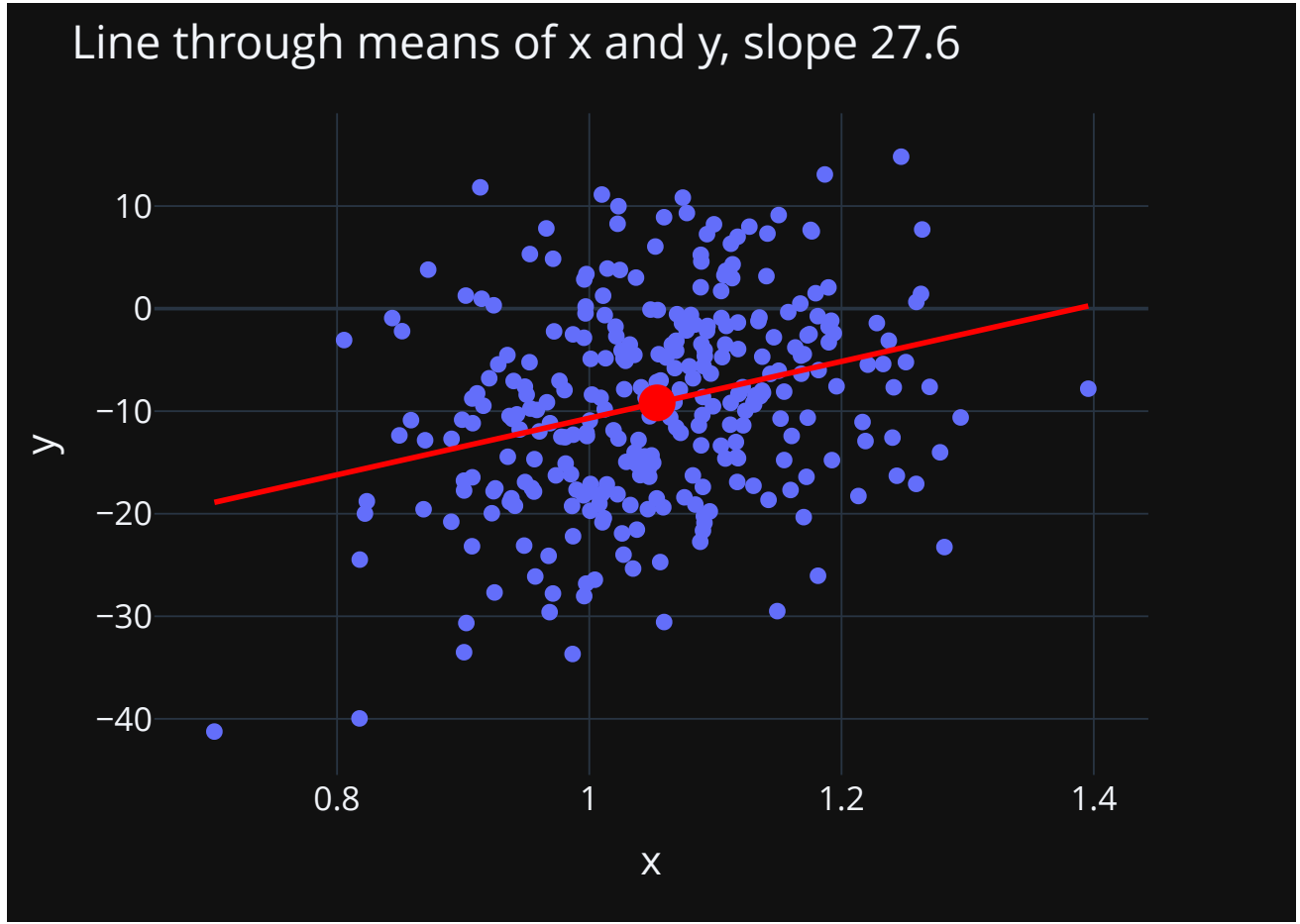

Line through means of x and y, slope -2.4
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

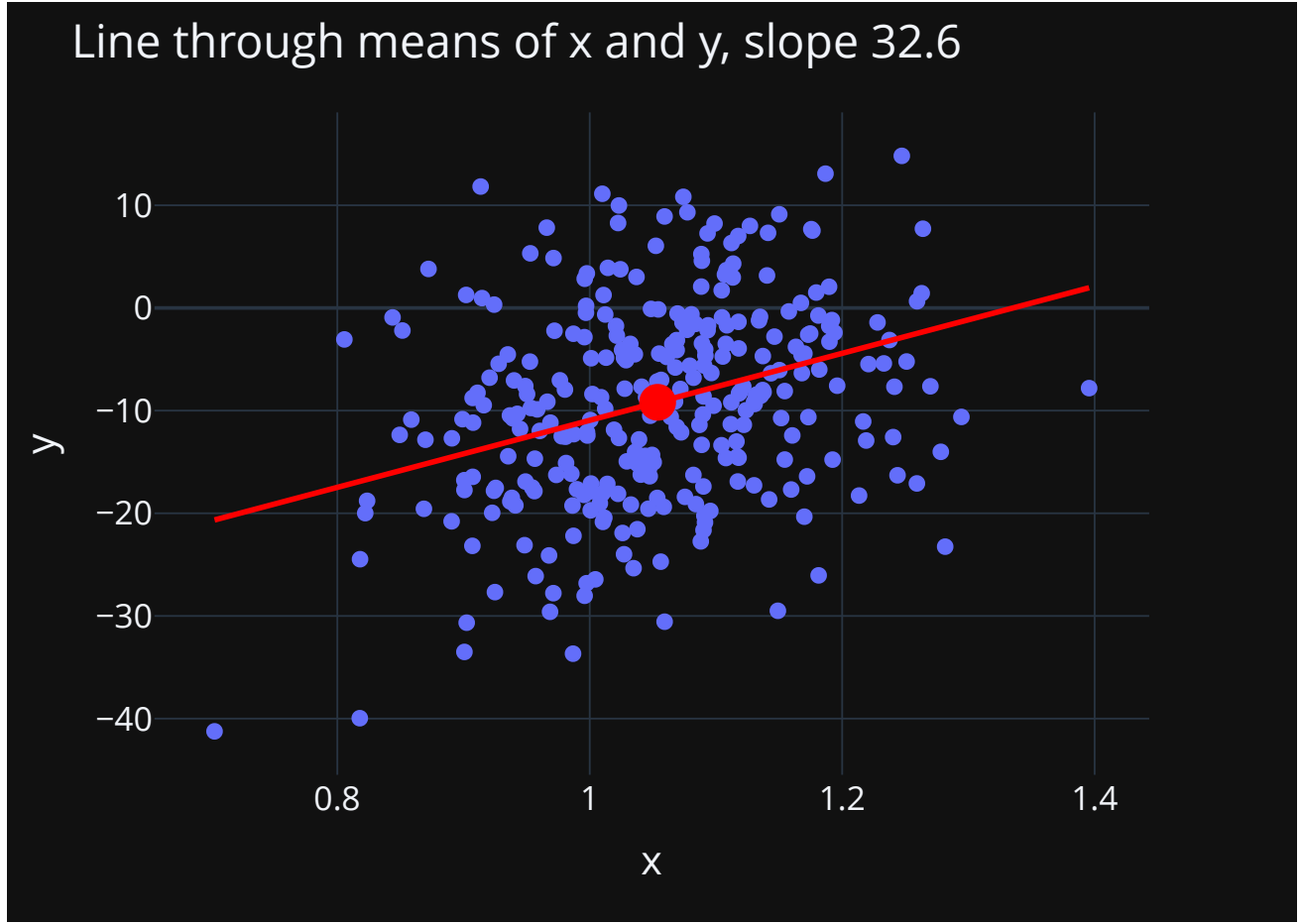

Line through means of x and y, slope 2.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

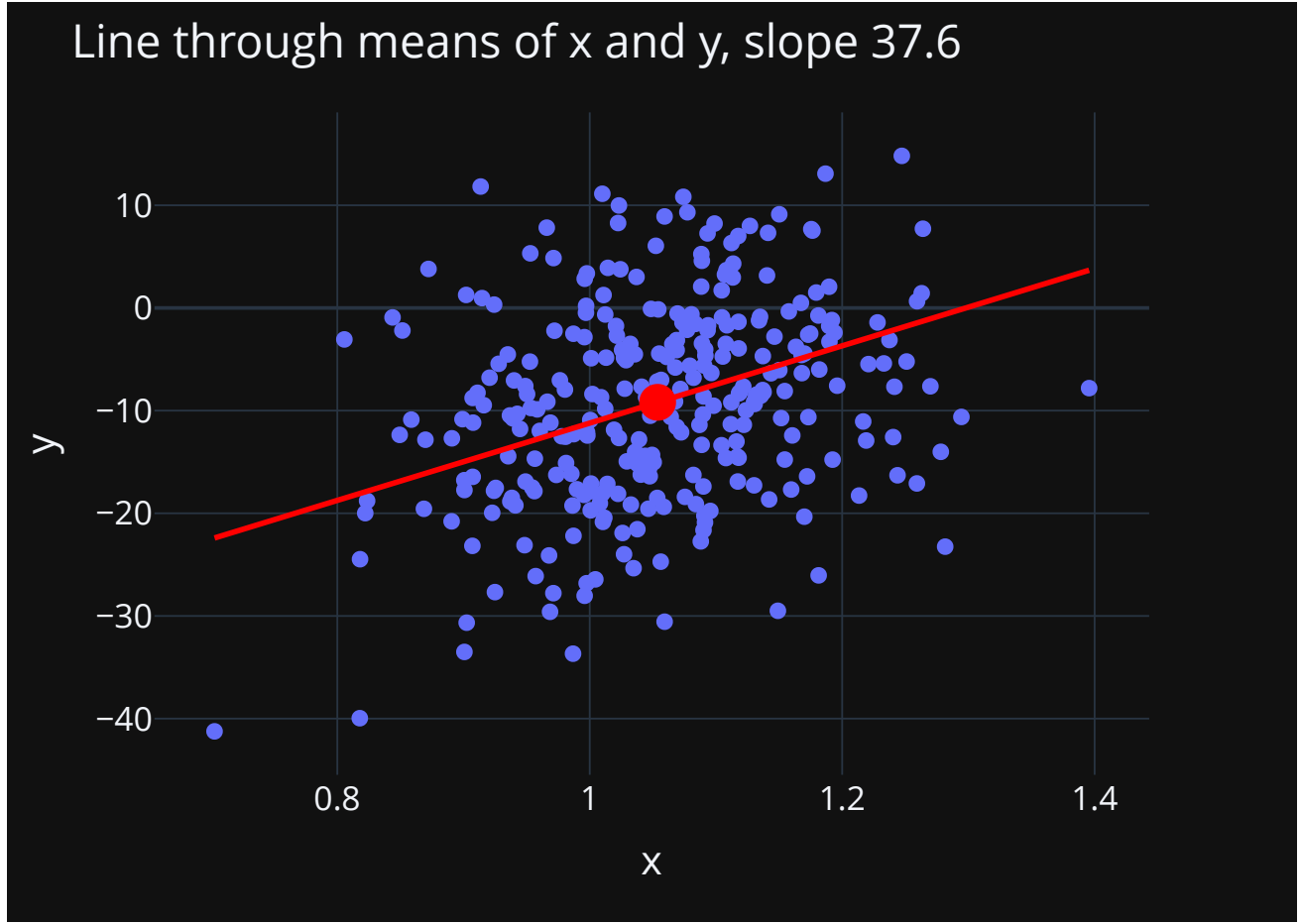

Line through means of x and y, slope 7.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

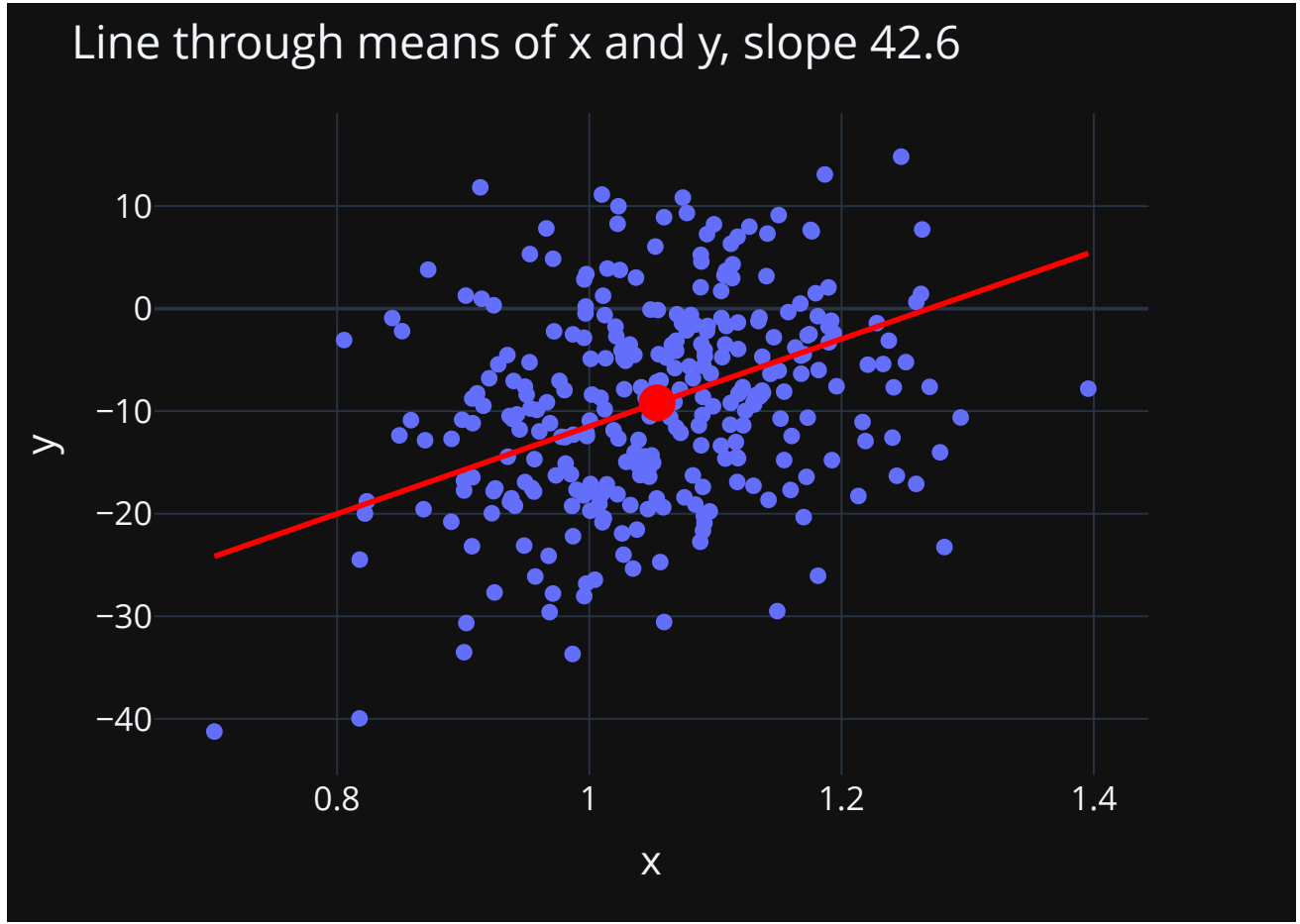

Line through means of x and y, slope 12.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

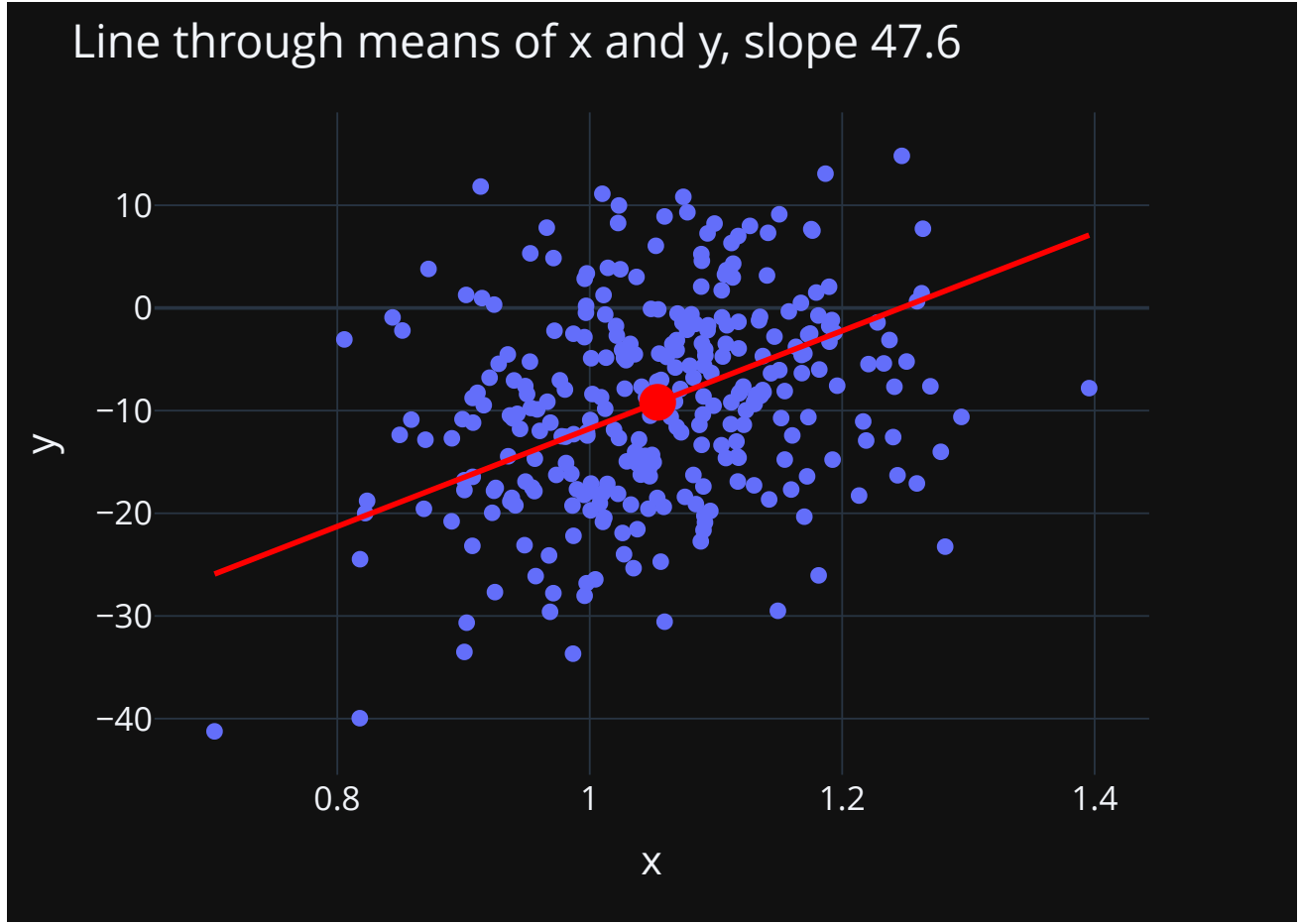

Line through means of x and y, slope 17.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

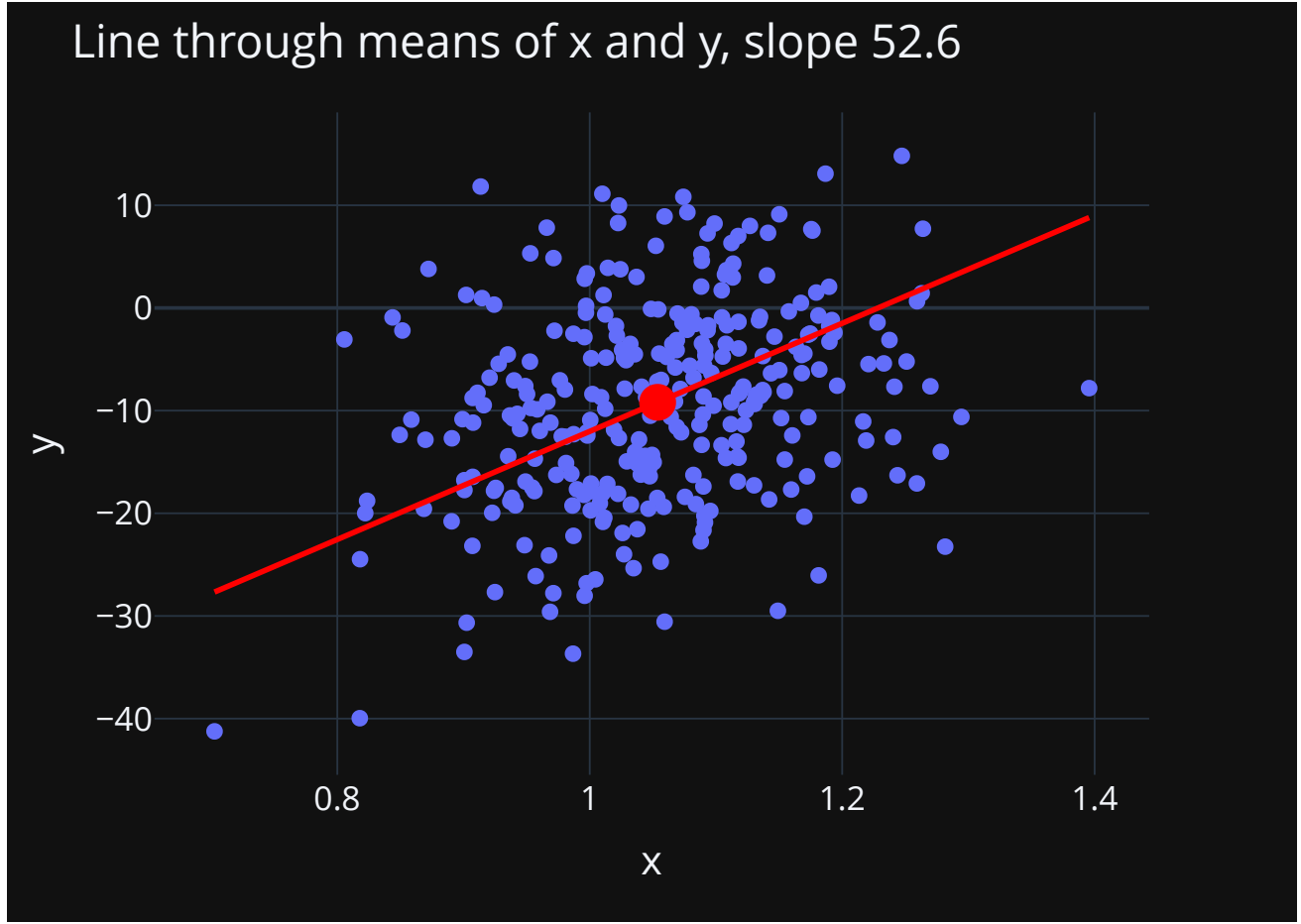

Line through means of x and y, slope 27.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

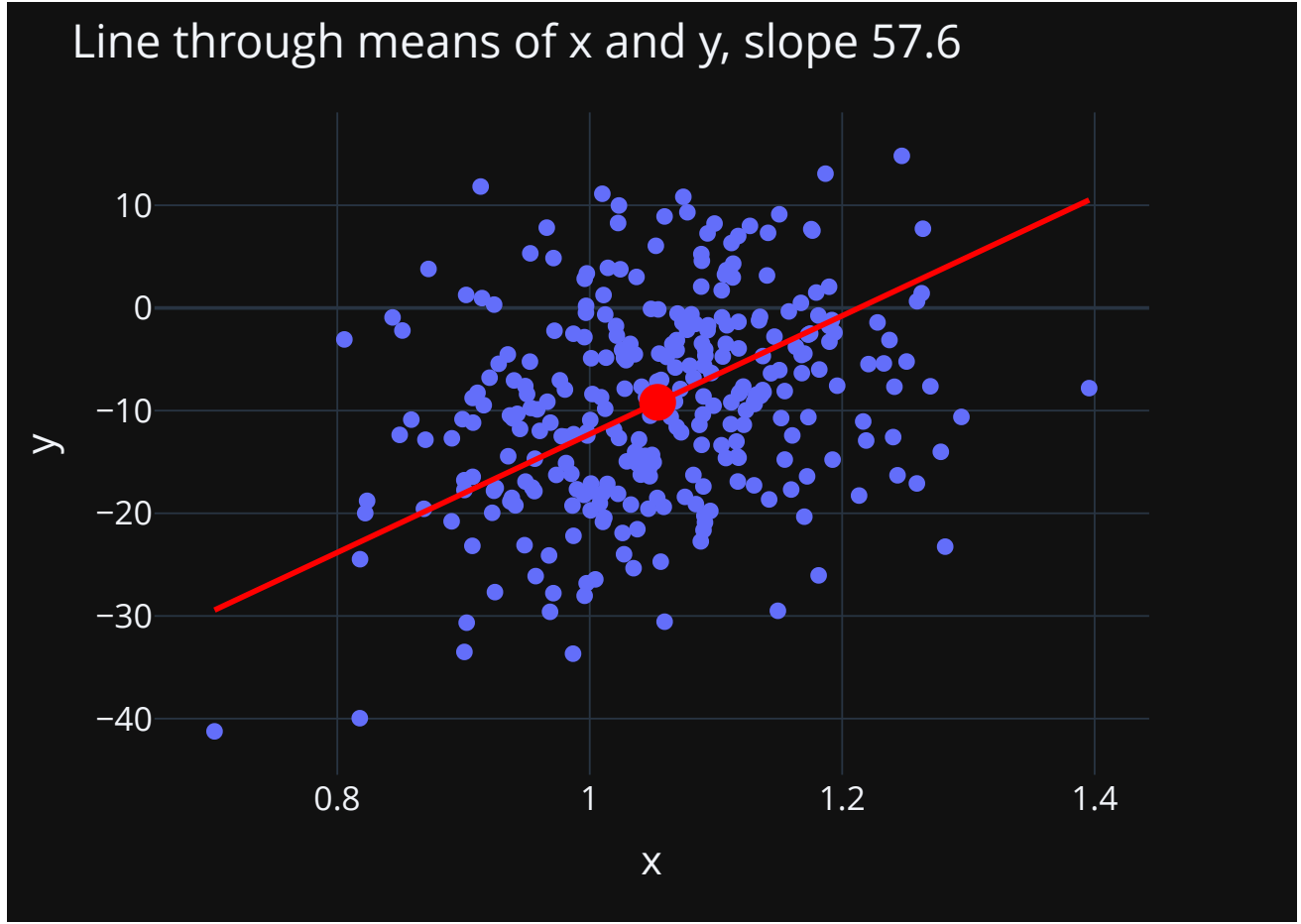

Line through means of x and y, slope 32.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x

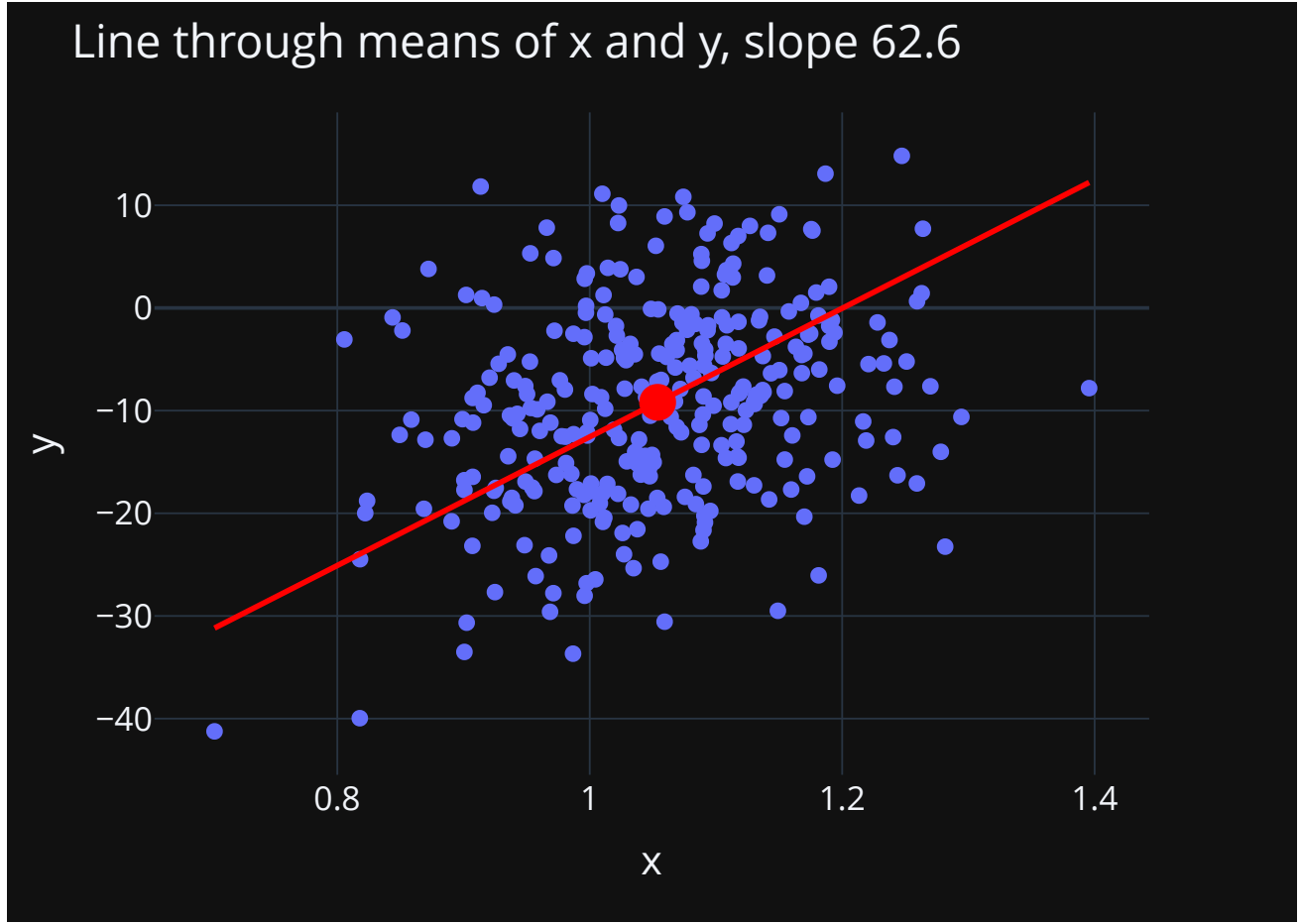

Line through means of x and y, slope 37.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Line through means of x and y, slope 42.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Line through means of x and y, slope 47.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Line through means of x and y, slope 52.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Line through means of x and y, slope 57.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x


Line through means of x and y, slope 62.6
10
0
−10
−20
−30
−40
y
0.8
1
1.2
1.4
x