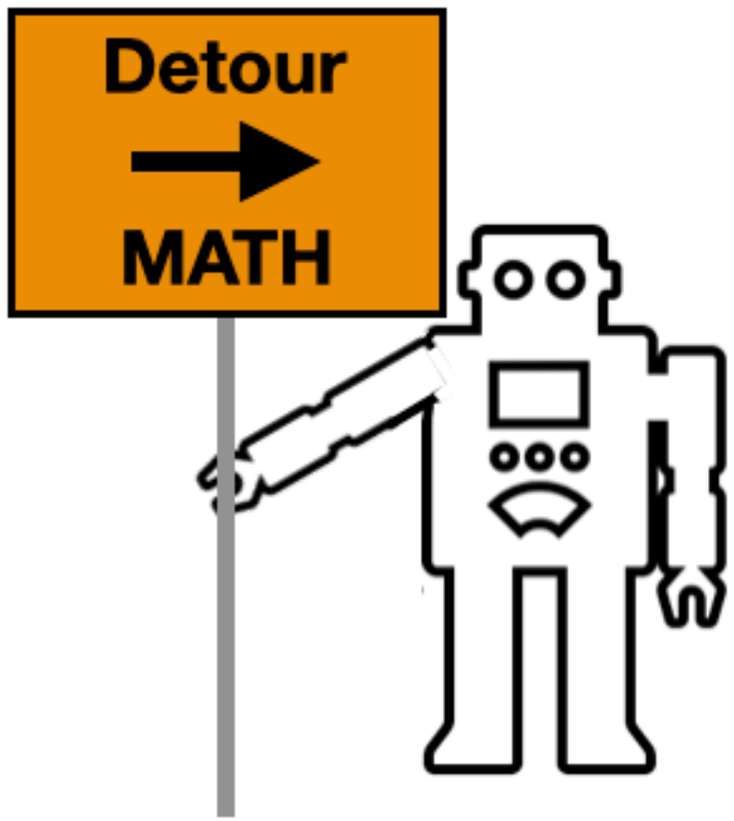
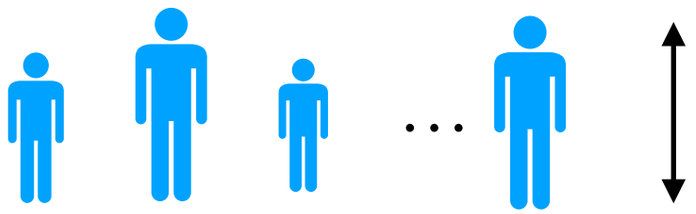
Multi-variable stats review



Multiple features



i	1	2	3		N	
h_i	h_1	h_2	h_3		h_N	ft
						weight
w_i	w_1	w_2	w_3		w_N	lb

shape = $2 \times N$

Transpose

sample set (np array or dataframe)

Feature-1
height

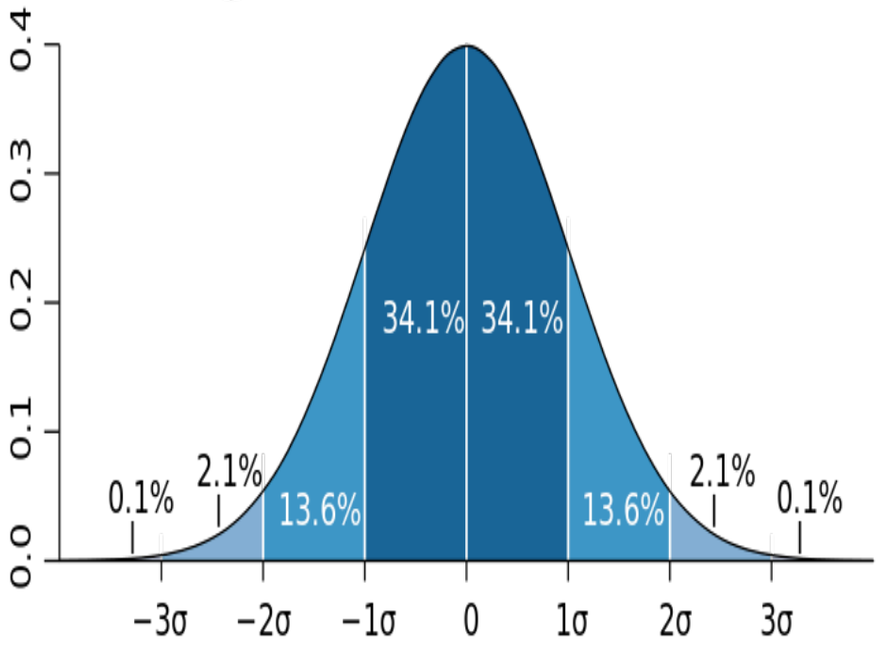
Feature-2
height

$$\mathbf{X} = \begin{pmatrix} h_1 & w_1 \\ h_2 & w_2 \\ \vdots & \\ h_N & w_N \end{pmatrix}$$

sample: i=1

sample: i=2

Height distribution



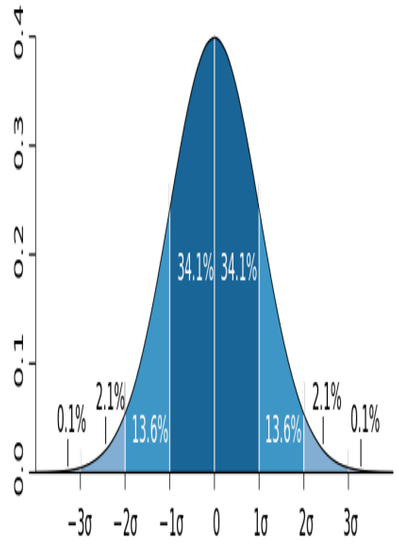
μ_h

σ_h

Weight distribution

μ_w

σ_w



Single variable summary stats

(Still analyzing individually)

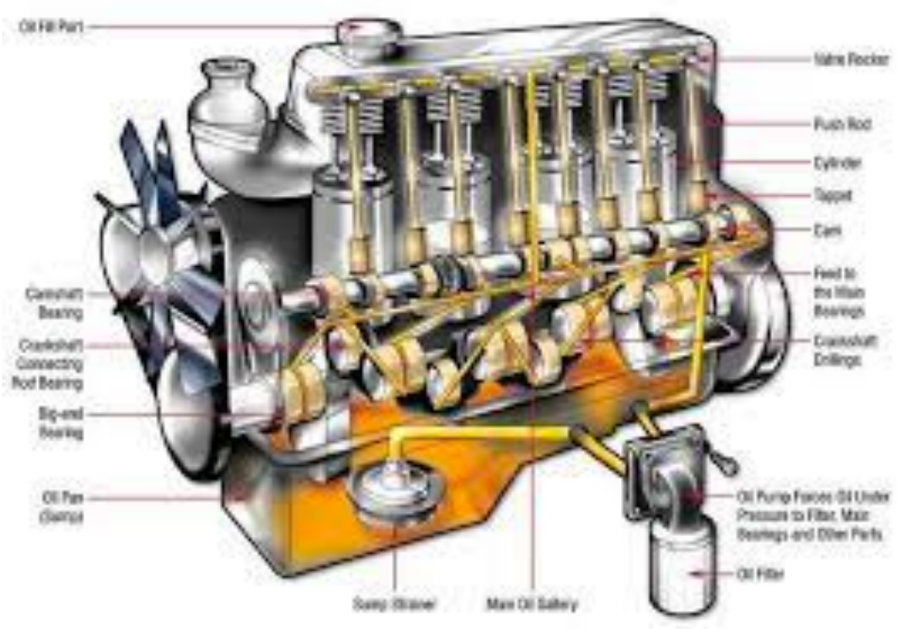
MPG Dataset

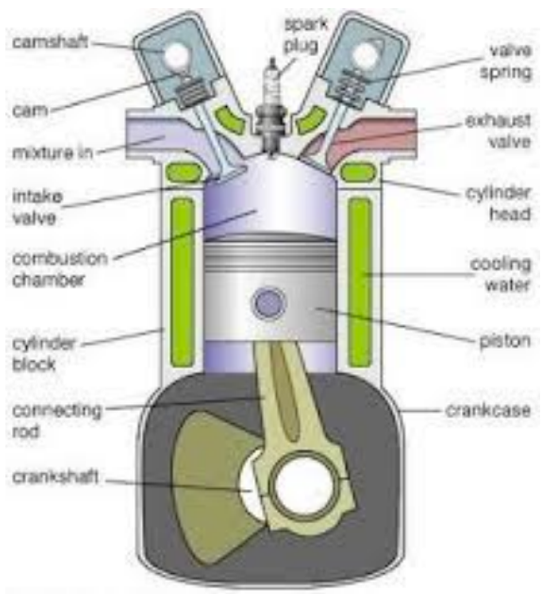
https://www.tensorflow.org/tutorials/keras/regression

DATA FRAME

	numeric continuous	numeric discrete	numeric continuous				Categorical		
ID	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Europe	Japan	USA
393	27.0	4	140.0	86.0	2790.0	15.6	82	0	0	1
394	44.0	4	97.0	52.0	2130.0	24.6	82	1	0	0
395	32.0	4	135.0	84.0	2295.0	11.6	82	0	0	1
396	28.0	4	120.0	79.0	2625.0	18.6	82	0	0	1
397	31.0	4	119.0	82.0	2720.0	19.4	82	0	0	1

number of cyclinders **volume of cyclinders**

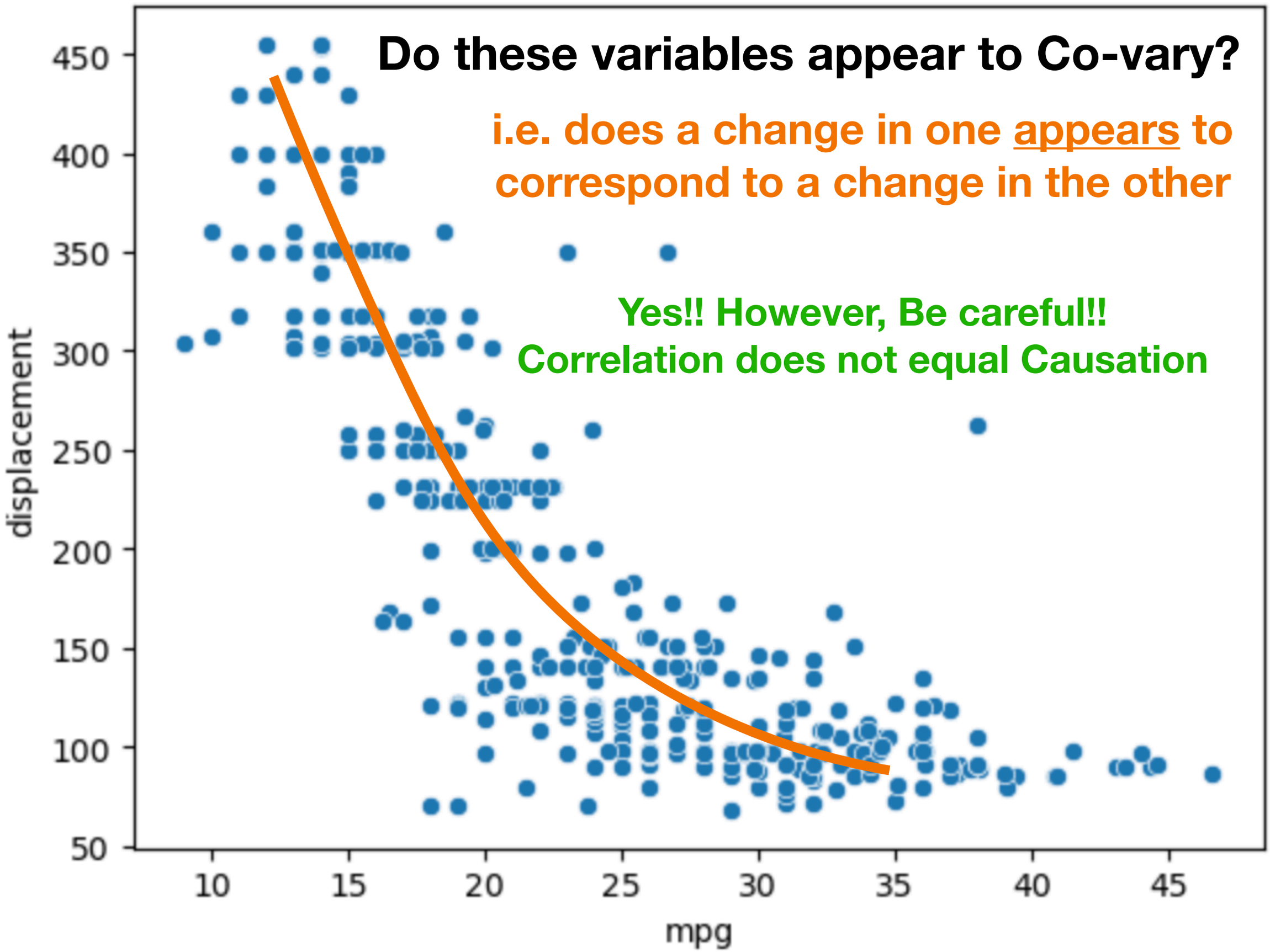






Scatter plot

```
sns.scatterplot(data=df, x="mpg", y="displacement")
```



Covariance-1

sample set (matrix or dataframe)

$$\mathbf{X} = \begin{pmatrix} h_1 & w_1 \\ h_2 & w_2 \\ \vdots & \\ h_N & w_N \end{pmatrix}$$

sample: i=1

sample: i=2



height deviation: $\Delta h_i = h_i - \mu_h$ **(distance from mean)**

weight deviation: $\Delta w_i = w_i - \mu_w$

variance: mean square deviation

$$var(h) = \sigma_h^2 = \left\langle \Delta h_i^2 \right\rangle = \left\langle \Delta h_i \Delta h_i \right\rangle = \frac{1}{N}\sum_{i=1}^{N}(h_i - \mu_h)(h_i - \mu_h)$$

co-variance

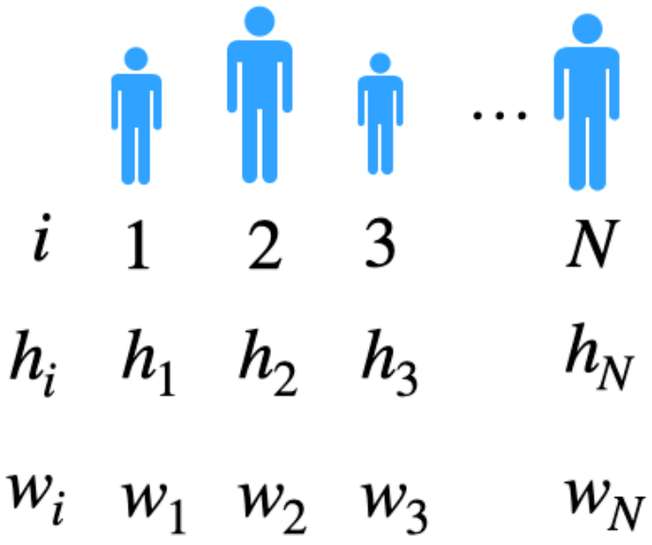
$$\mathrm{cov}(h,w) = \langle \Delta h \Delta w \rangle = \frac{1}{N}\sum_{i=1}^{N}(h_i - \mu_x)(h_i - \mu_h)$$

Pearson correlation coefficent (normalized covariance):

condenses pair-plot down to a single number from -1 to 1

$$\rho_{hw} = \frac{cov(h,w)}{\sigma_h \sigma_w} = \frac{\langle \Delta h \Delta w \rangle}{\sigma_h \sigma_w}$$

Covariance-2



$$\begin{pmatrix} h_1 & w_1 \\ h_2 & w_2 \\ \vdots & \\ h_N & w_N \end{pmatrix}$$

single variable statistics:

$$\mu_h = \langle \mathbf{h} \rangle \qquad \sigma_h^2 = \langle \mathbf{\Delta h}\, \mathbf{\Delta h} \rangle$$

$$\mu_w = \langle \mathbf{w} \rangle \qquad \sigma_w^2 = \langle \mathbf{\Delta w}\, \mathbf{\Delta w} \rangle$$

Covariance (un-unormalized) $\mathrm{cov}(h,w) = \langle \mathbf{\Delta w}\, \mathbf{\Delta h} \rangle = \big\langle (w_i - \mu_w)\,(h_i - \mu_h) \big\rangle$

Pearson's correlation coefficient **normalized co-variance**

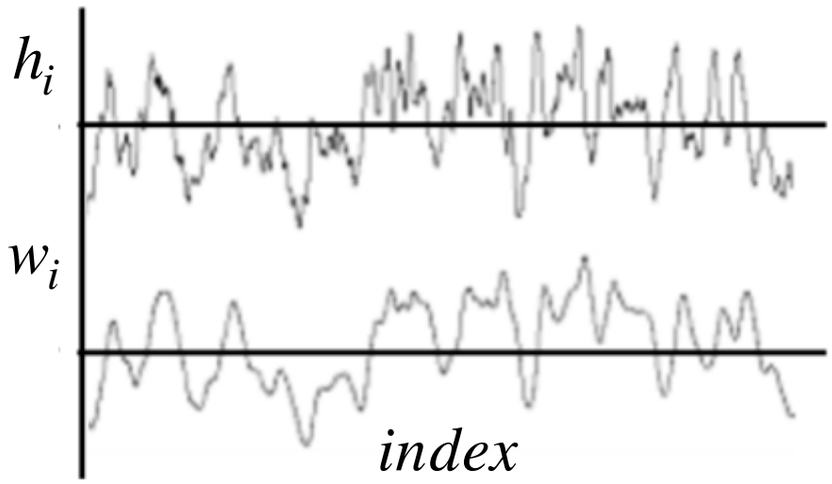
$$\rho_{\mathbf{i,j}} = \frac{\mathrm{cov}(x_i, x_j)}{\sigma_i \sigma_j} = \left\langle \frac{(x_i - \mu_i)}{\sigma_i} \frac{(x_j - \mu_j)}{\sigma_j} \right\rangle \in [-1,1]$$

co-variance matrix $\mathbf{C}_{hw} = \begin{pmatrix} \sigma_h^2 & cov(h,w) \\ cov(w,h) & \sigma_w^2 \end{pmatrix}$ $\qquad \rho_{hw} = \begin{pmatrix} 1 & \rho_{h,w} \\ \rho_{w,h} & 1 \end{pmatrix}$

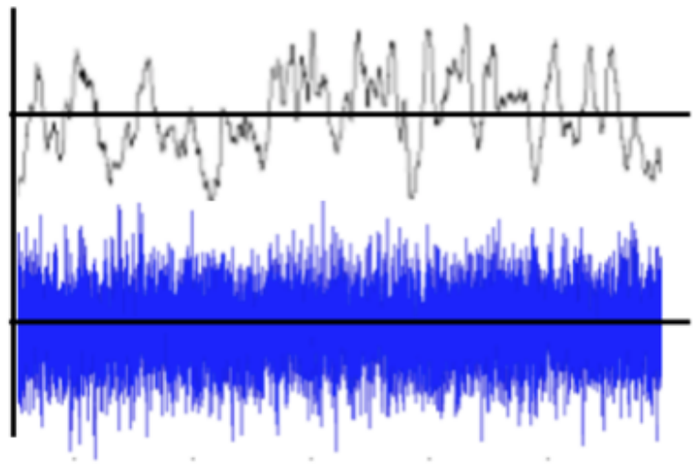
Possible cases

Stacked series plots (similar to "timeseries")

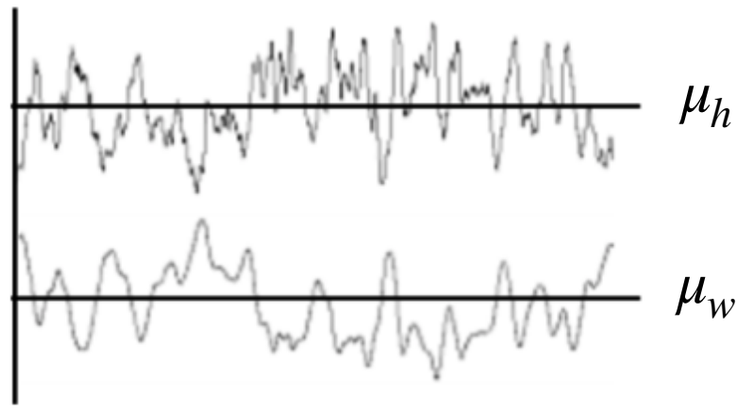
$$\mathrm{cov}(h, w) = \langle \mathbf{\Delta w\, \Delta h} \rangle = \left\langle (w_i - \mu_w)\,(h_i - \mu_h) \right\rangle$$



correlated



uncorrelated

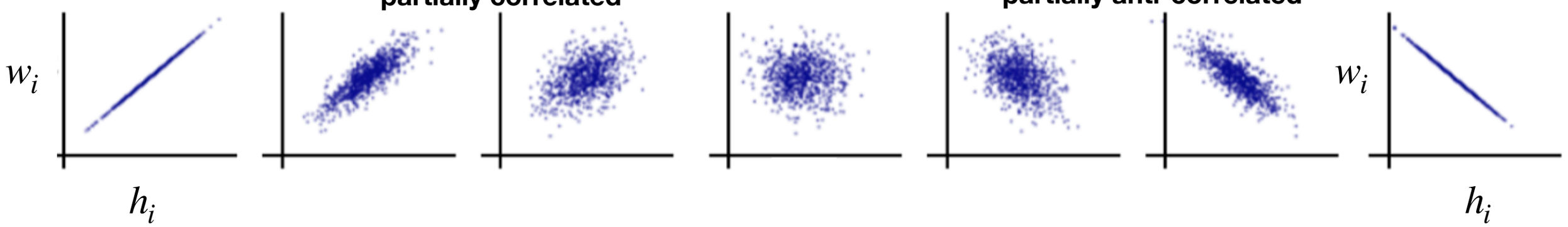


anti-correlated

Different possibilties for how w and h COULD be are "related"

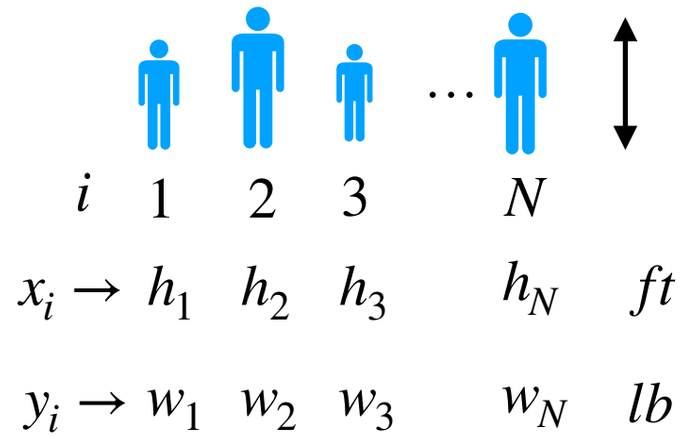
pair plot: plot weight as a function of height

correlated **uncorrelated** **anti-correlated**



$\rho_{hw} = 1$ $\rho_{hw} = 0$ $\rho_{hw} = -1$

Pearsons's correlation coefficient



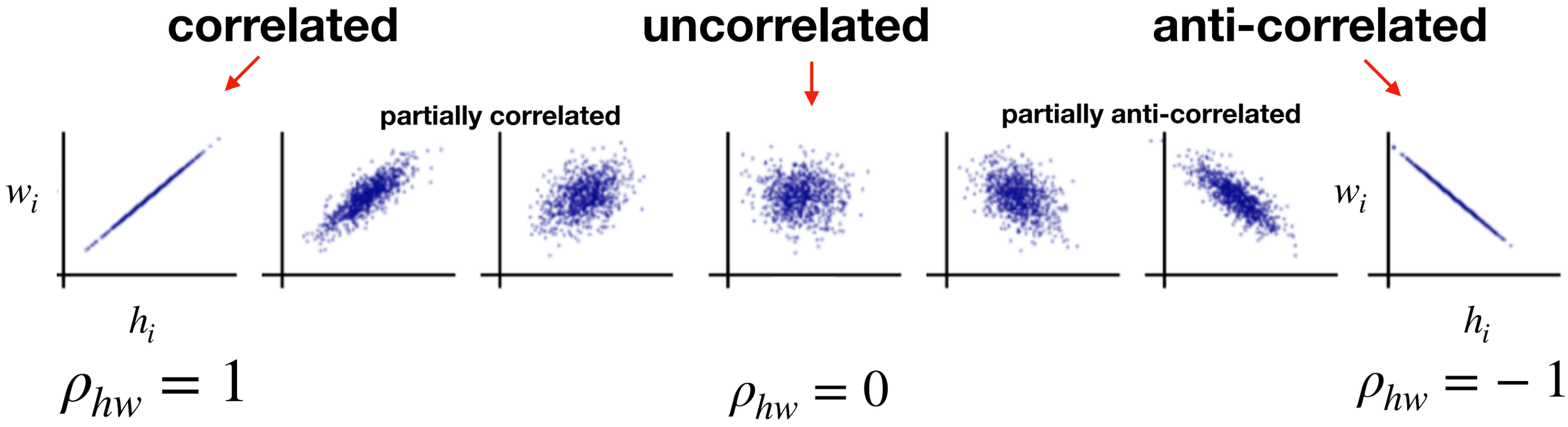
The correlation coefficient is calculated as follows:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

where m_x is the mean of the vector x and m_y is the mean of the vector y.

Under the assumption that x and y are drawn from independent normal distributions (so the population correlation coefficient is 0), the probability density function of the sample correlation coefficient r is ([1], [2]):

pair plot: possible cases for weight as a function of height



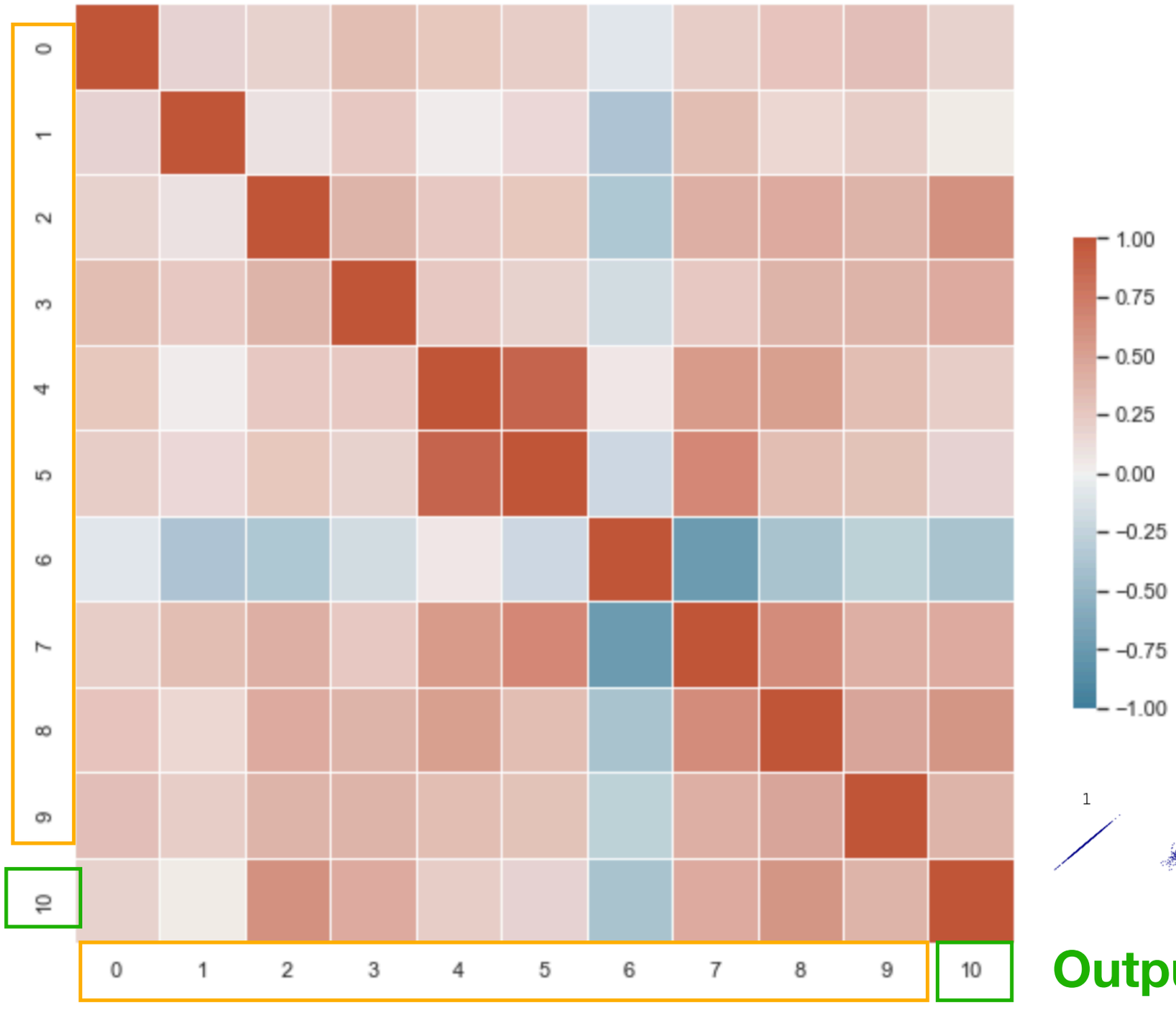
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

Visualizing correlation matrix

Correlation matrix heatmap

```
sns.set_theme(style="white")
f, ax = plt.subplots(figsize=(11, 9))  # Set up the matplotlib figure
cmap = sns.diverging_palette(230, 20, as_cmap=True)     # Generate a custom diverging colormap
# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr,  cmap=cmap, vmin=-1, vmax=1, center=0,
        square=True, linewidths=.5, cbar_kws={"shrink": .5})
plt.show();
```

Inputs: x



Basic feature selection

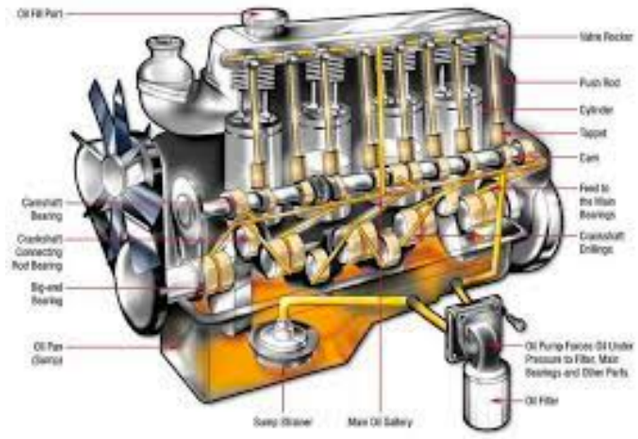
$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$



Outputs: y

Example

```
print("------------------------")
print("CORRELATION MATRIX:")
print("------------------------")
print(df.corr())
```



```
------------------------
CORRELATION MATRIX:
------------------------
                    mpg   cylinders  displacement  horsepower      weight
mpg            1.000000   -0.775396     -0.804203   -0.778427 -0.831741
cylinders     -0.775396    1.000000      0.950721    0.842983  0.896017
displacement  -0.804203    0.950721      1.000000    0.897257  0.932824
horsepower    -0.778427    0.842983      0.897257    1.000000  0.864538
weight        -0.831741    0.896017      0.932824    0.864538  1.000000
acceleration   0.420289   -0.505419     -0.543684   -0.689196 -0.417457
model_year     0.579267   -0.348746     -0.370164   -0.416361 -0.306564
```

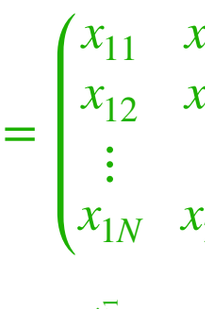
Visualizing-3

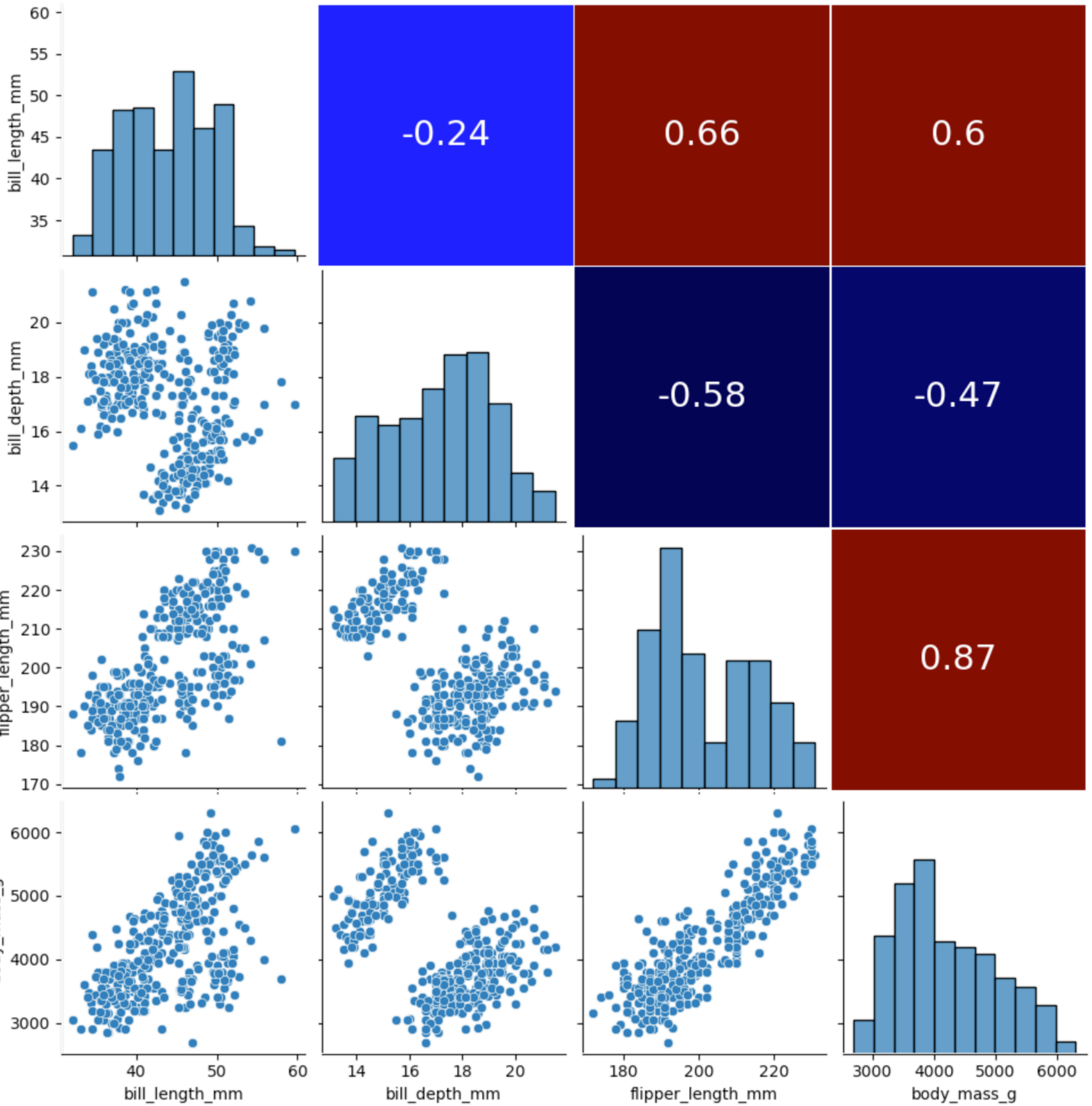
sample set (matrix or dataframe)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \\ \vdots & & & \\ x_{1N} & x_{2N} & x_{3N} & x_{4N} \end{pmatrix} \begin{matrix} \text{sample-1} \\ \text{sample-2} \\ \\ \text{sample-N} \end{matrix}$$

feature-1 $\rightarrow x_1$ **feature-2** $\rightarrow x_2$ **feature-3** $\rightarrow x_3$ **feature-4** $\rightarrow x_4$

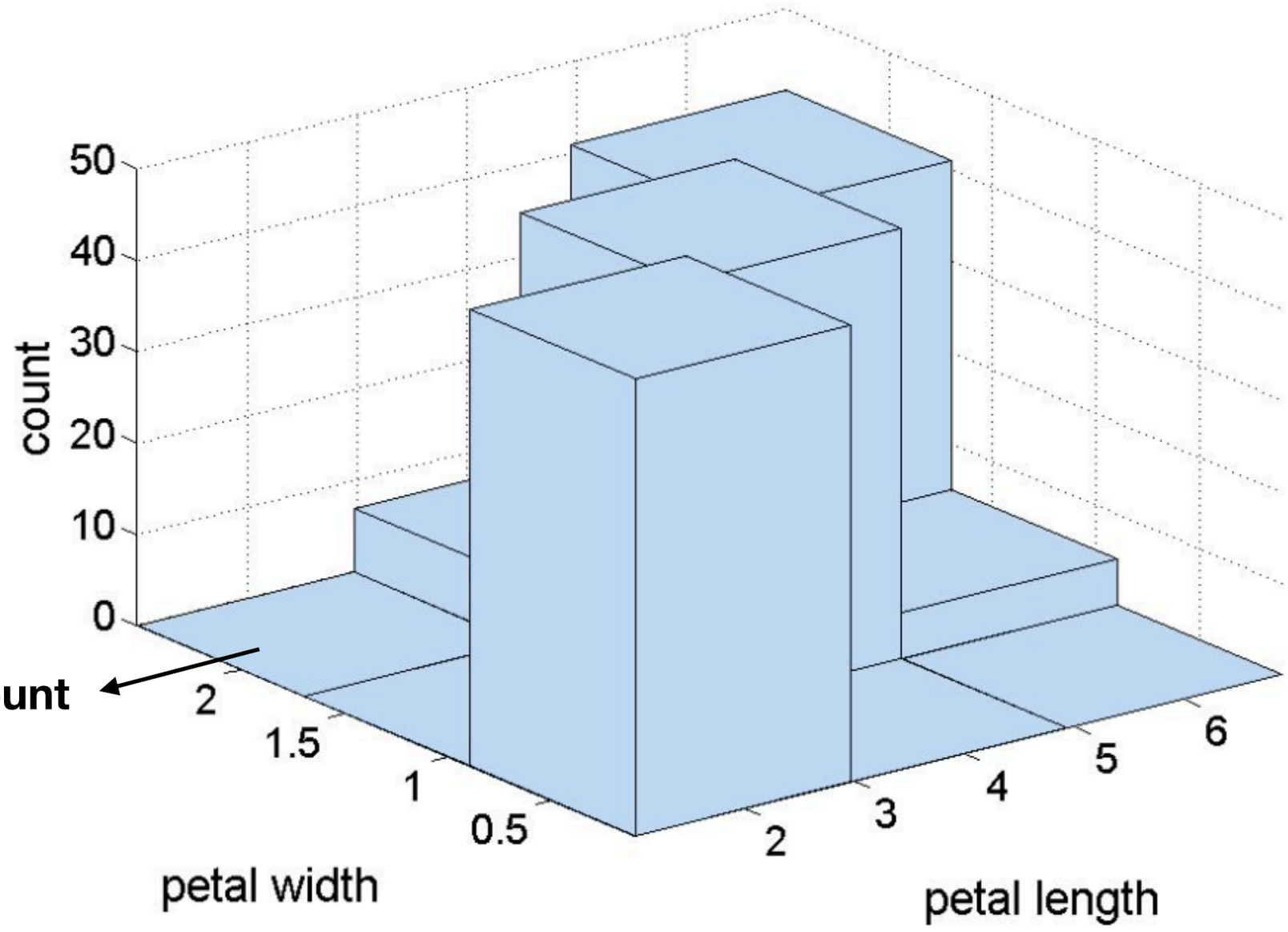
bill_length_mm bill_depth_mm flipper_length_mm body_mass_g





Two - Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?

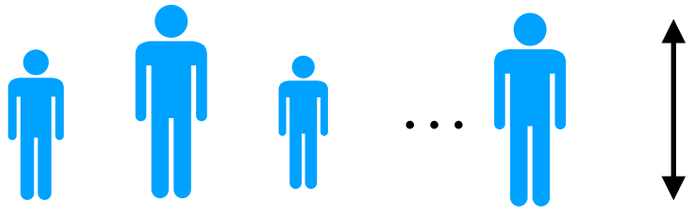


form 2D mesh and count

Add mesh slides

Source: Dr. Faisal Kamiran: https://github.com/FaizSaeed/Data-Science-Course/tree/master/Lectures

Multi-variable PDF



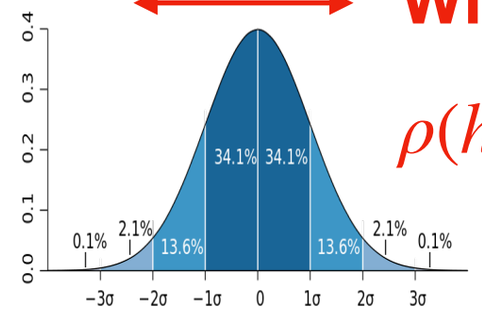
i	1	2	3		N	
h_i	h_1	h_2	h_3		h_N	ft
w_i	w_1	w_2	w_3		w_N	lb

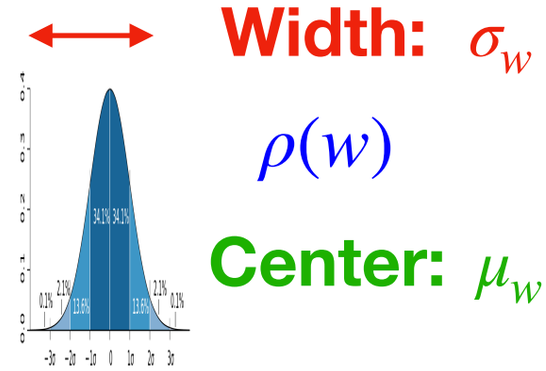
Height distribution

weight distribution

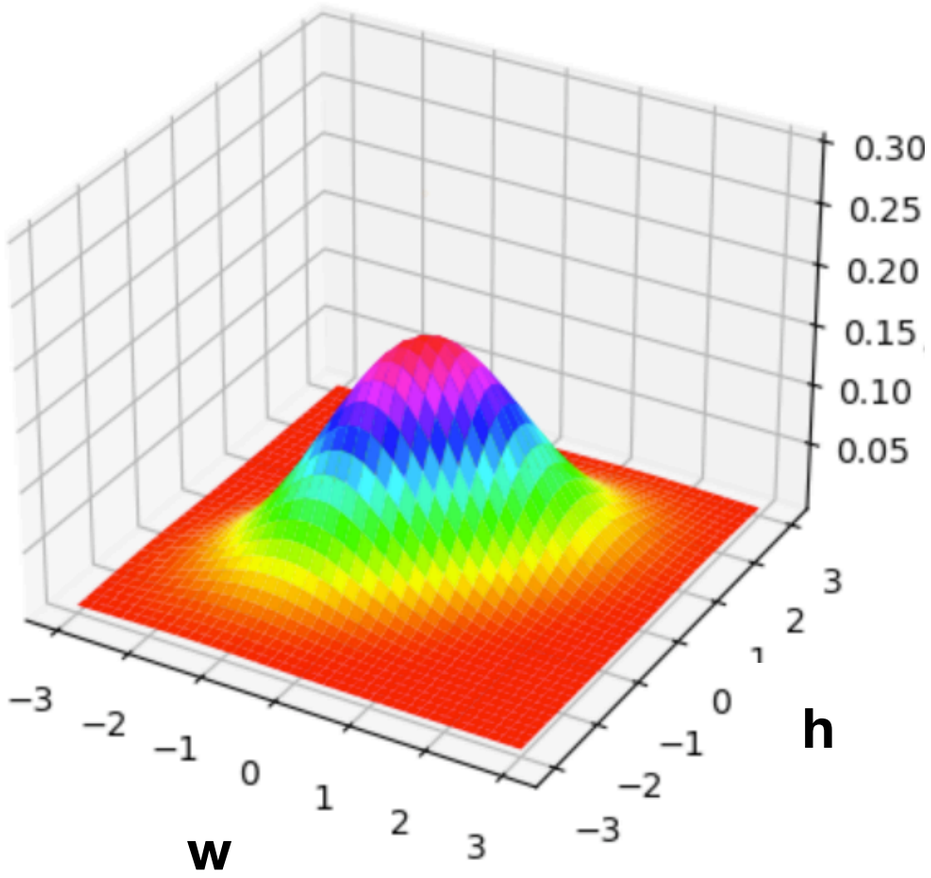
1D projections of a 2D data-set

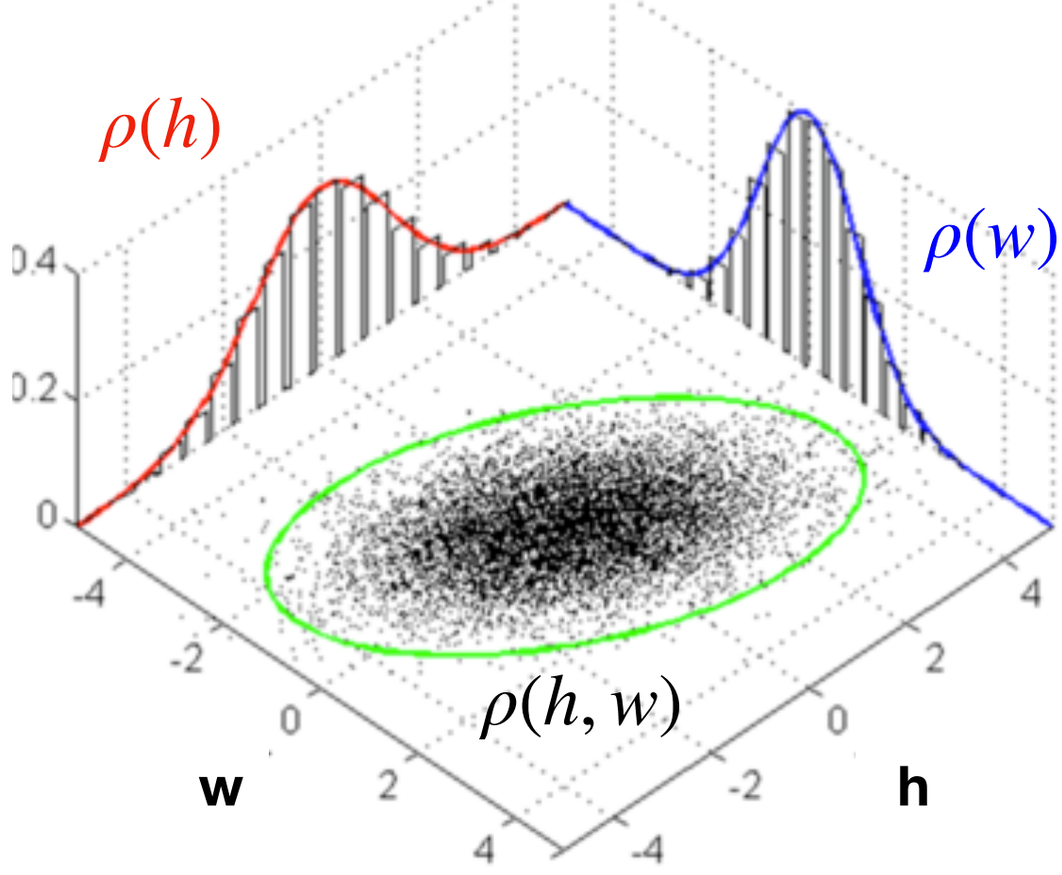
Single variable distribtions:





Multi-variable PDF





combined (joint) distribution

2D Normal distribution

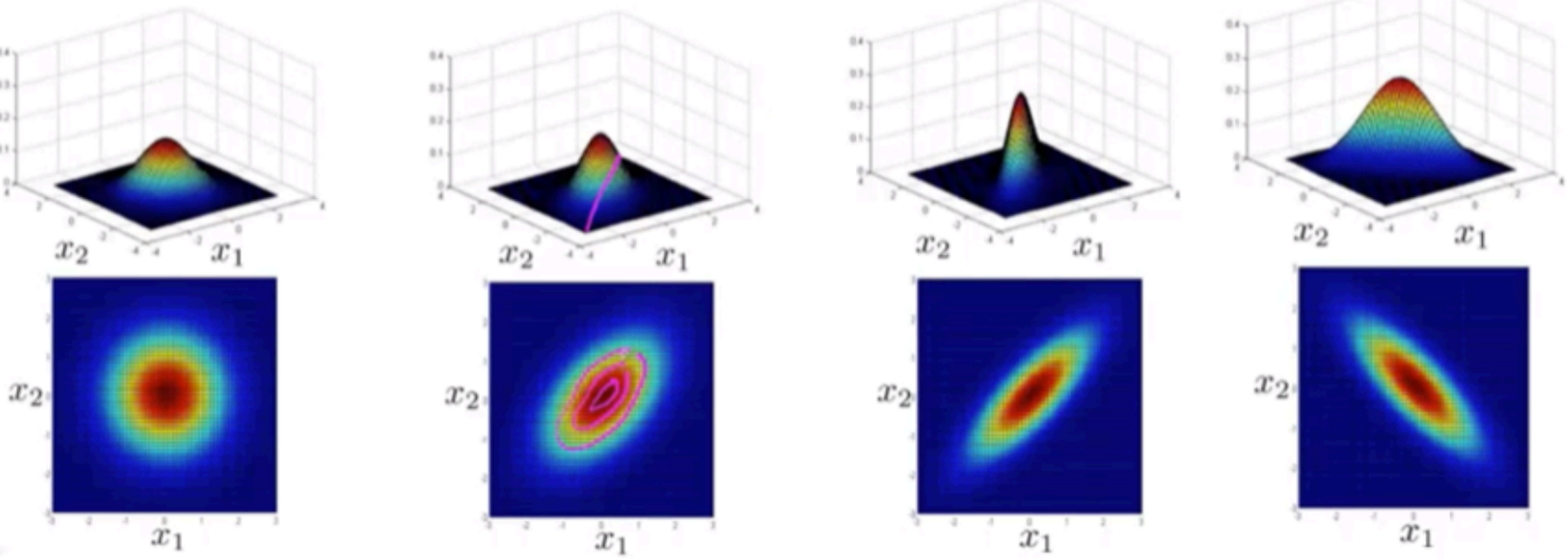
1D: $p(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$ **two parameters:** (μ, σ)

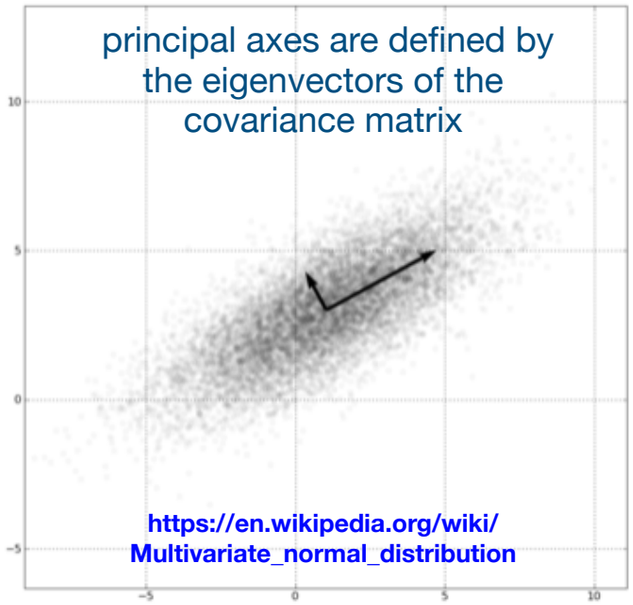
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

2D: $f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}$

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}$$

Image source: Andrew Ng





uncorrelated

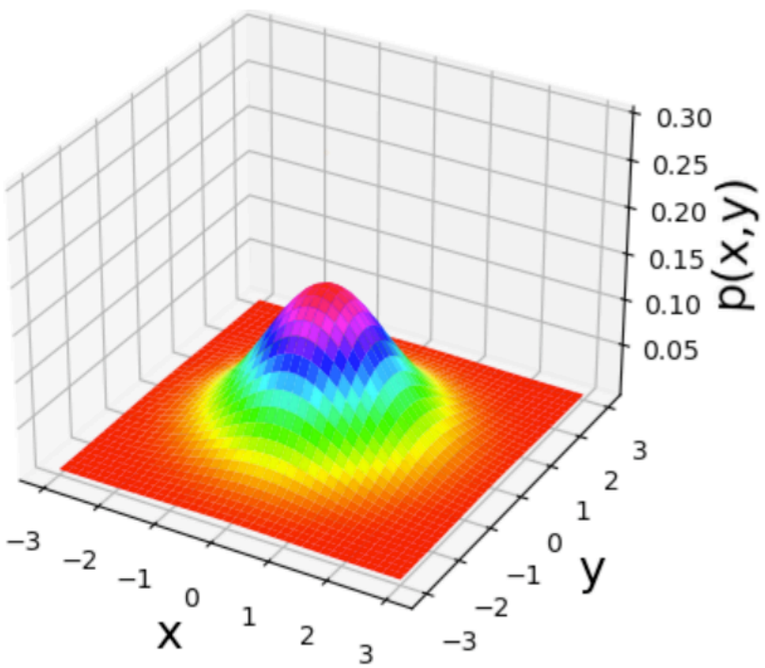
$$\boldsymbol{\mu} = \begin{bmatrix}0 \\ 0\end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}$$

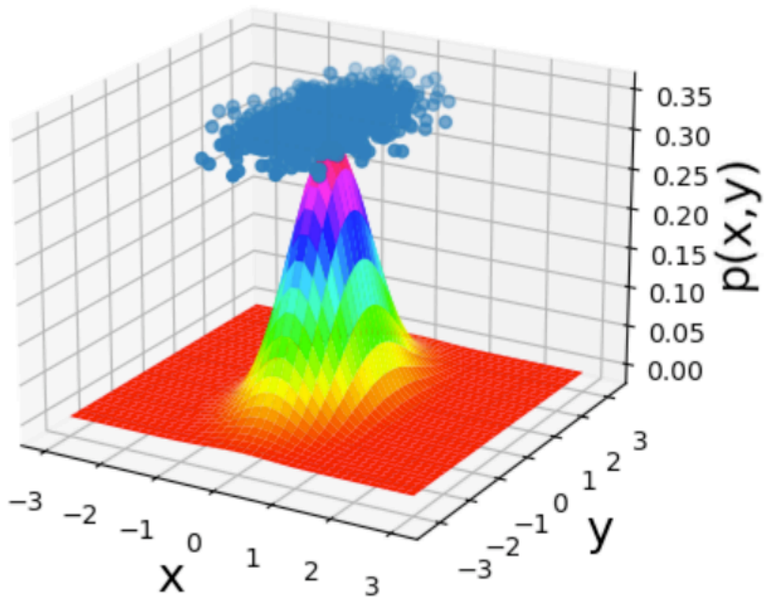
correlated: non-zero covariance on the off diagonal

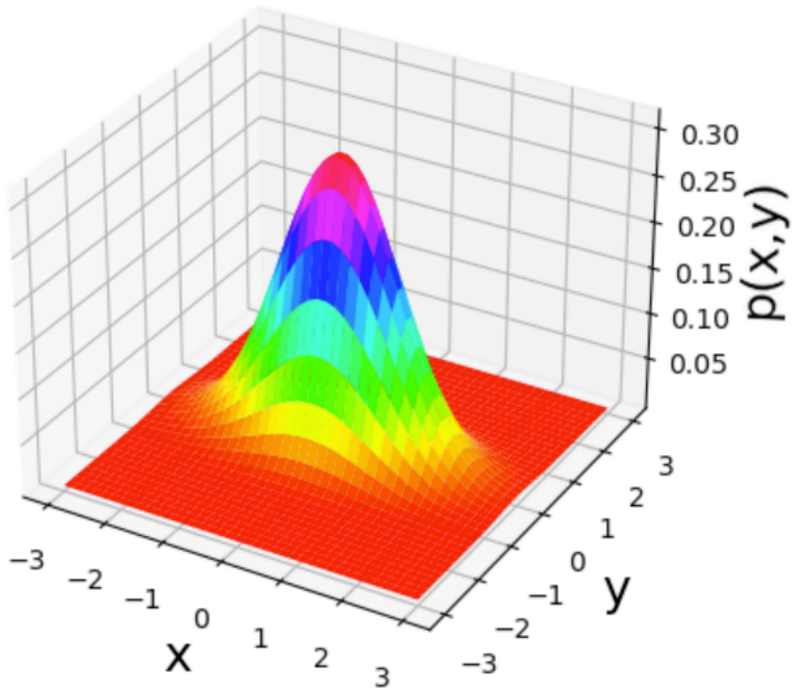
$$\boldsymbol{\mu} = \begin{bmatrix}0 \\ 0\end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix}1 & A \\ A & 1\end{bmatrix}$$

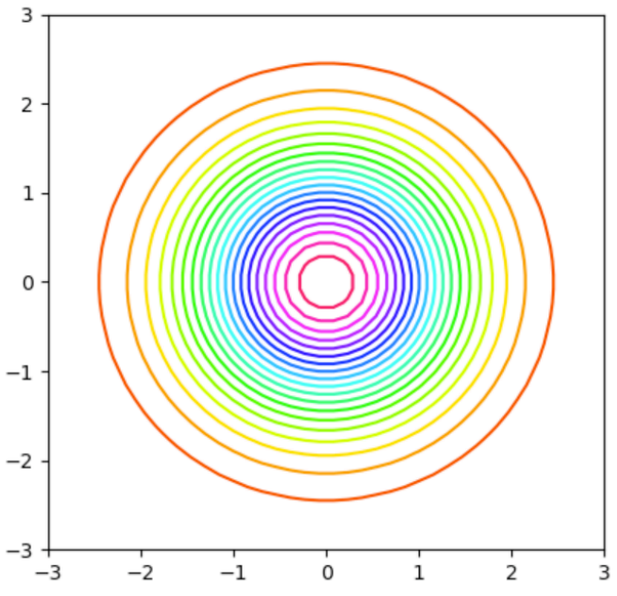
k-D: $f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$ **arbitary dimension:**

Bi-variate (uncorrelated) LECTURE-CODES/WEEK5/2D-PLOTS.py

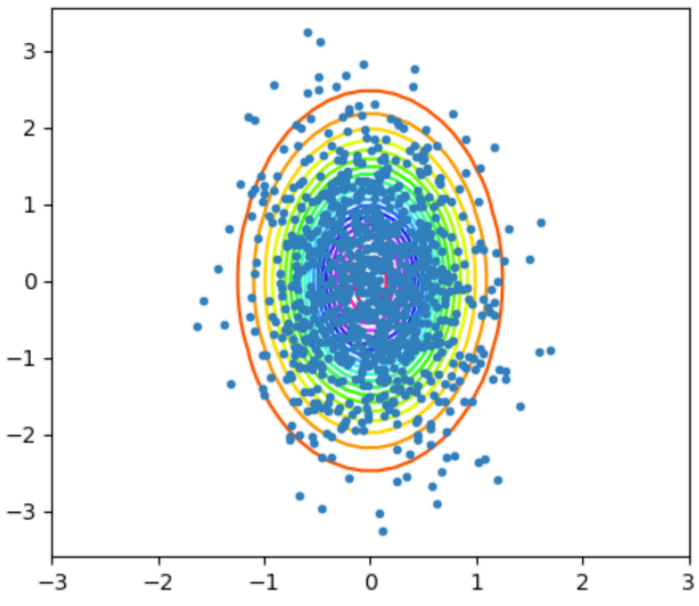




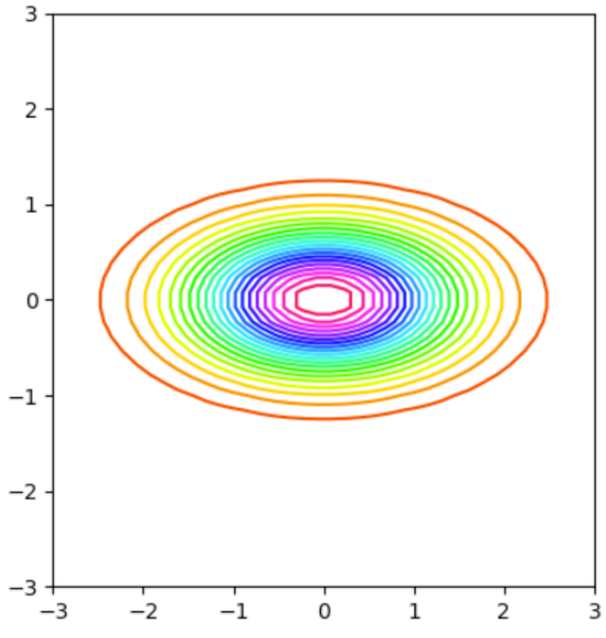




$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



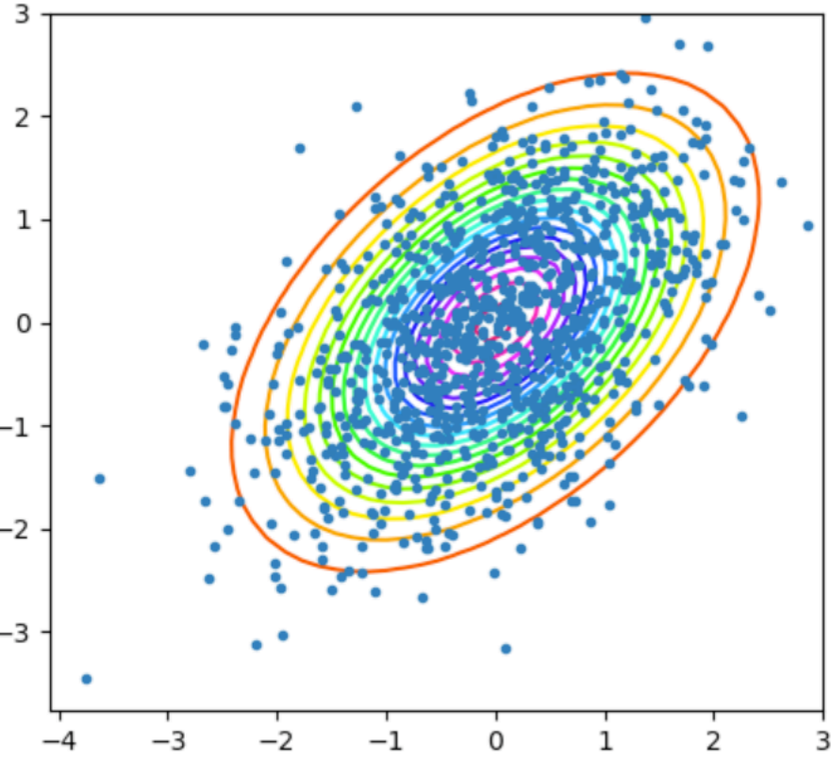
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}$$

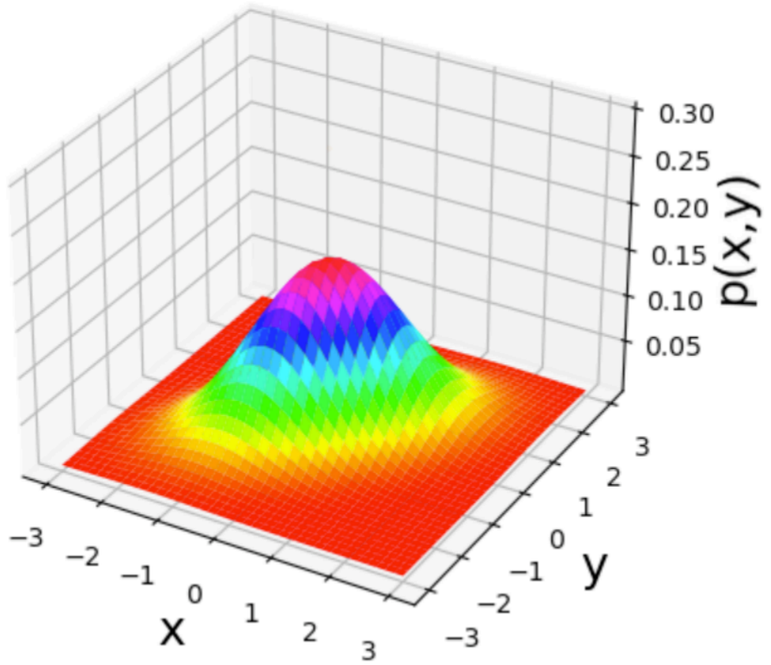


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

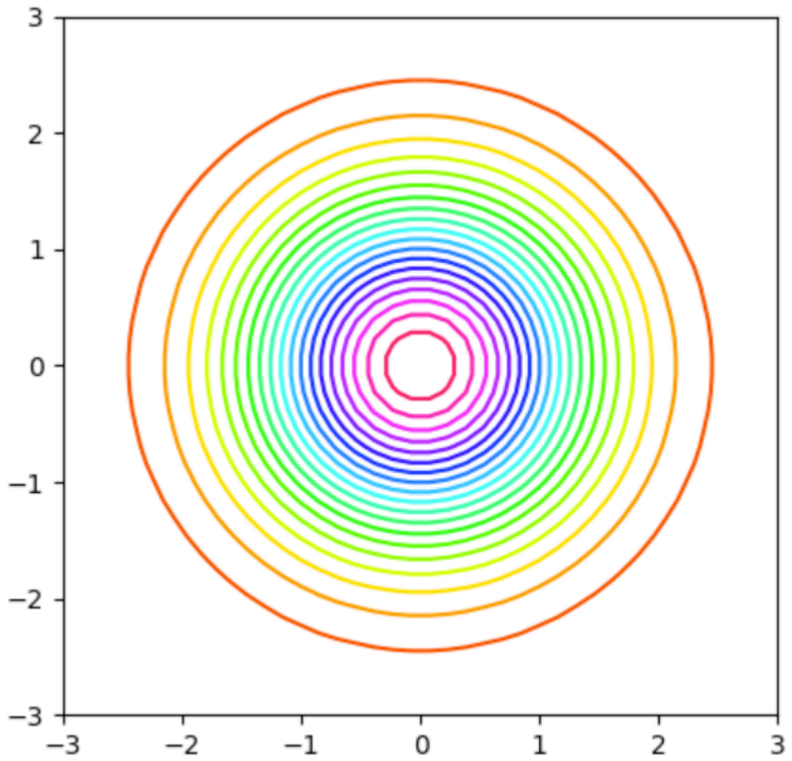
Bi-variate (correlated)

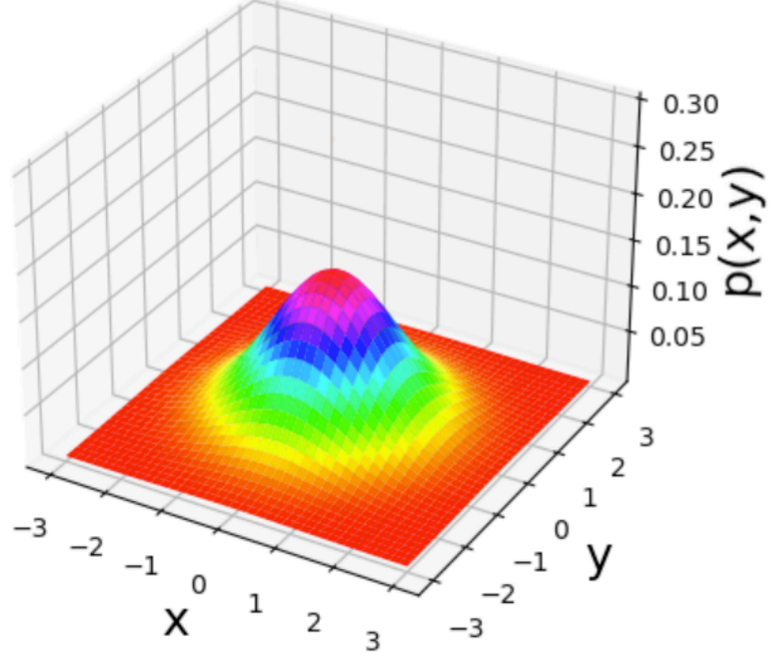
LECTURE-CODES/WEEK5/2D-PLOTS.py





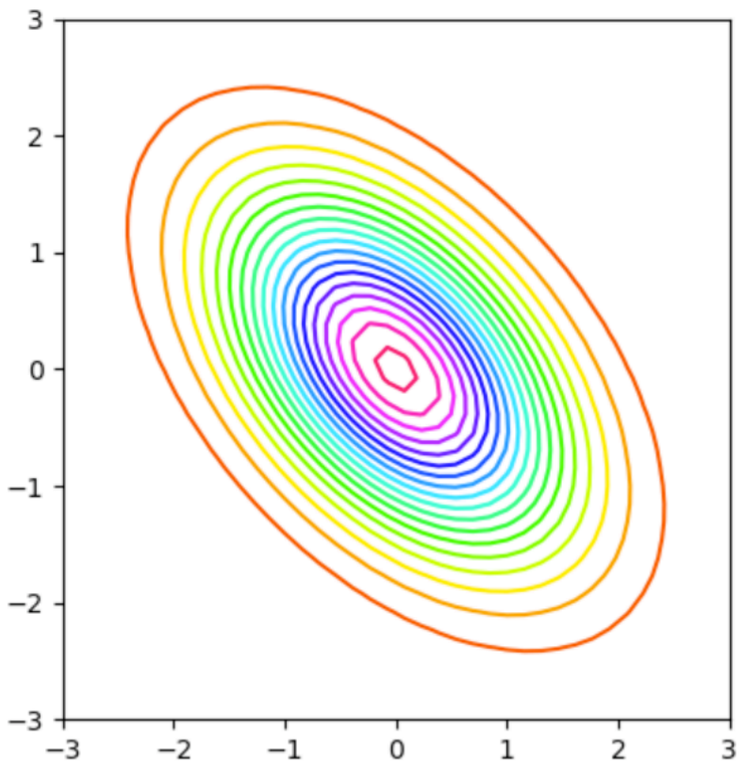
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

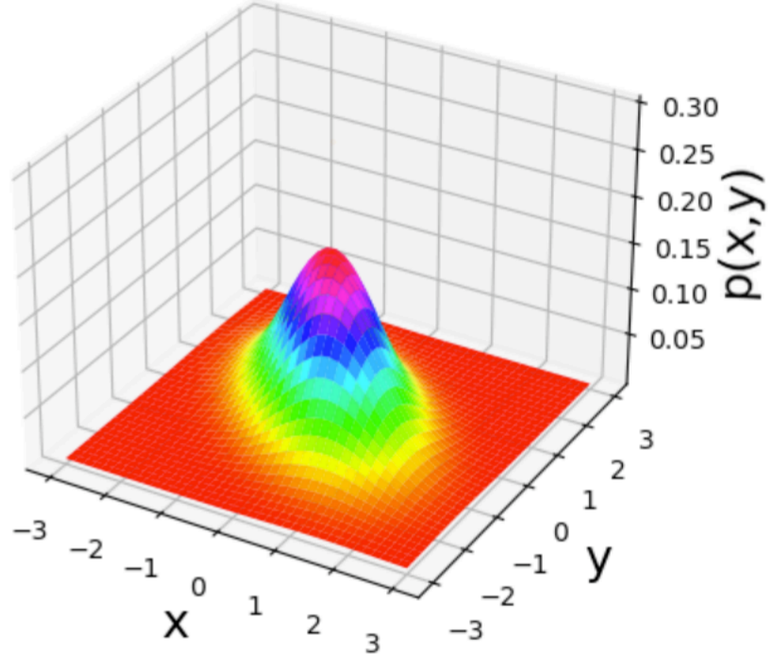




$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

un-correlated





$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$