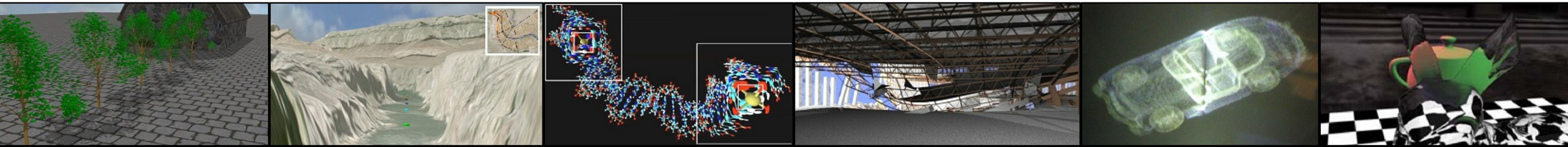
CIS 4930/6930-002
DATA VISUALIZATION

Histograms & Correlation

Paul Rosen
Assistant Professor
University of South Florida



HISTOGRAMS

Bar chart-based visualization that allows evaluating distribution of values.

Given: $X=\{x_0,\ldots,x_n\}$

Select: k bins

$$bin_i = k * (x_i - \min X) / (\max X - \min X)$$

X={1,2.5,3,4}

k = 3

$$X=\{1,2.5,3,4\}$$

$$k = 3$$

$$X=\{1,2.5,3,4\}$$

$$k = 3$$

$$bin_i = floor(k * (x_i - min\ X) / (max\ X - min\ X))$$

$$X=\{1,2.5,3,4\}$$

$$k = 3$$

$$bin_i = floor(3 * (x_i - 1) / (4 - 1))$$

X={1,2.5,3,4}

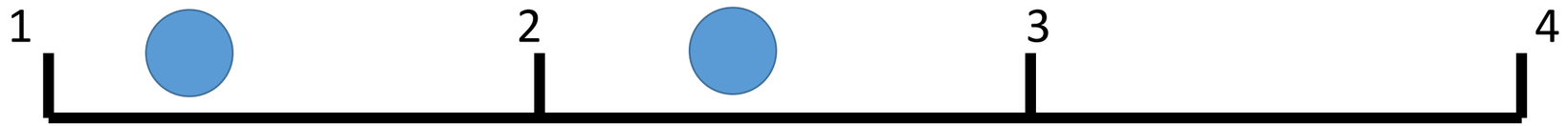
k = 3

1 -> floor(3 * (1 – 1) / (4 – 1)) = Bin 0

$$X=\{1,2.5,3,4\}$$

$$k = 3$$

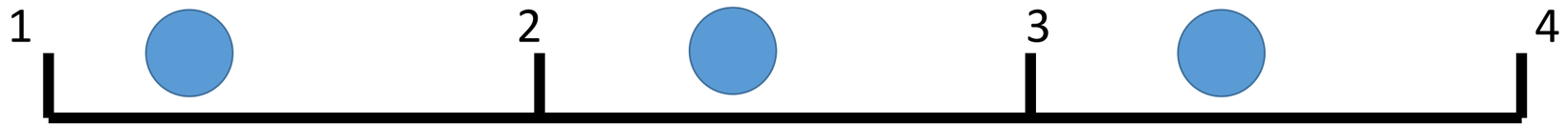
$$2.5 \rightarrow \text{floor}(3 * (2.5 - 1) / (4 - 1)) = \text{Bin } 1$$



X={1,2.5,3,4}

k = 3

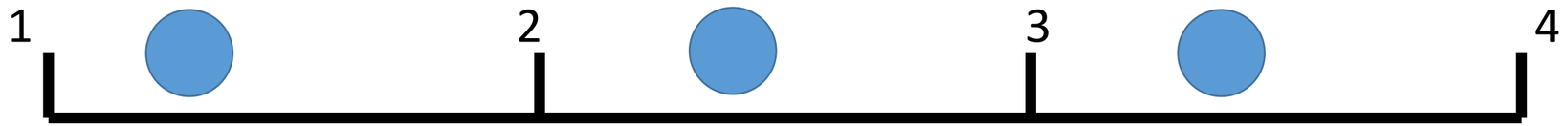
3 -> floor(3 * (3 – 1) / (4 – 1)) = Bin 2



X={1,2.5,3,4}

k = 3

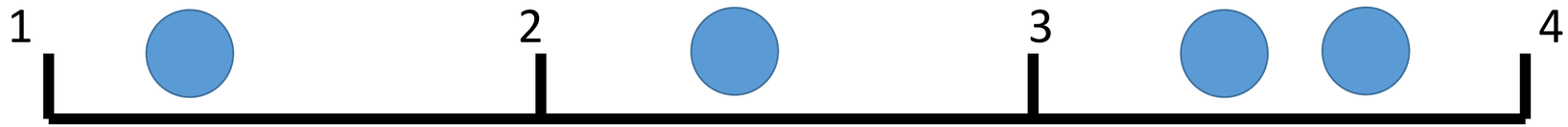
4 -> floor(3 * (4 – 1) / (4 – 1)) = Bin 3?



X={1,2.5,3,4}

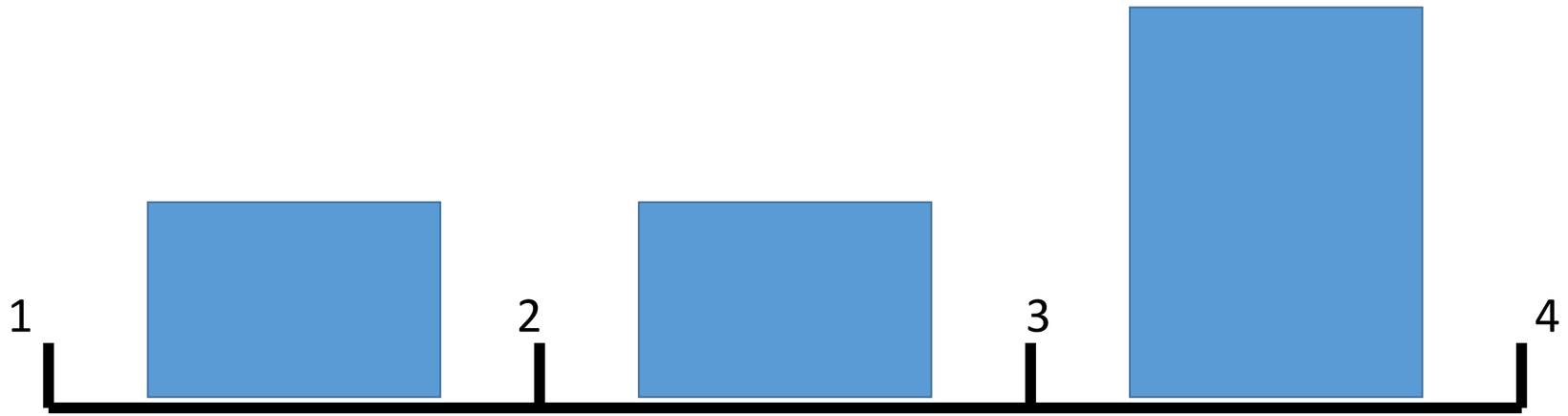
k = 3

4 -> floor(3 * (4 – 1) / (4 – 1)) = Bin 2



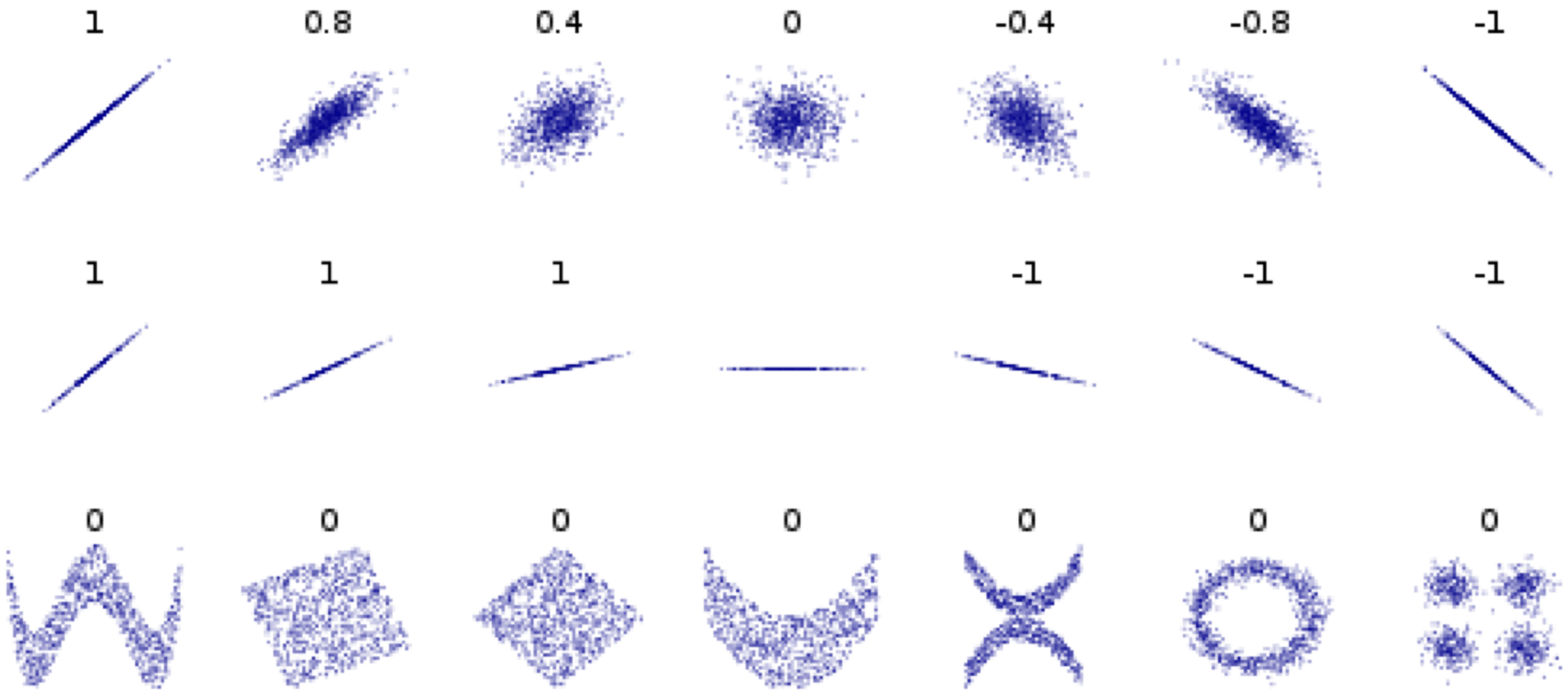
X={1,2.5,3,4}

k = 3



Pearson Correlation Coefficient

A measure of the linearity between 2 sets


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

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

where:

- **cov** is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n, x_i, y_i are defined as above
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Given: $X=\{x_0,\ldots,x_n\}, Y=\{y_0,\ldots,y_n\}$

Calculate mean(X), mean(Y), stdev(X), stdev(Y)

mean(X) = $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

stdev(X) = $\sigma_X = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$

$$r = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_X \sigma_Y}$$

$$X=\{1,2.5,3,4.5\}$$
$$Y=\{2,2.5,3.5,4\}$$

$$\text{mean}(X) = 2.75, \text{mean}(Y) = 3$$

$$\text{stdev}(X)= \text{sqrt}((1-2.75)^2 + (2.5-2.75)^2 + (3-2.75)^2 + (4.5-2.75)^2 / 4) = 1.25$$
$$\text{stdev}(Y)= \text{sqrt}((2-3)^2 + (2.5-3)^2 + (3.5-3)^2 + (4-3)^2 / 4) = 0.79$$

X={1,2.5,3,4.5}

Y={2,2.5,3.5,4}

mean(X) = 2.75, mean(Y) = 3

stdev(X)= 1.25, stdev(Y)= 0.79

$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ = 1/4 * (1-2.75)(2-3) + (2.5-2.75)(2.5-3) +

(3-2.75)(3.5-3) + (4.5-2.75)(4-3)

= 3.75 / 4 = 0.94

X={1,2.5,3,4.5}

Y={2,2.5,3.5,4}

mean(X) = 2.75, mean(Y) = 3

stdev(X)= 1.25, stdev(Y)= 0.79

$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ = 0.94

r = 0.94 / (1.25 * 0.79) = 0.95