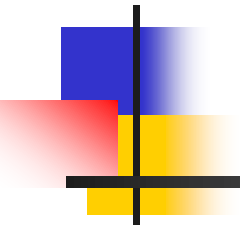
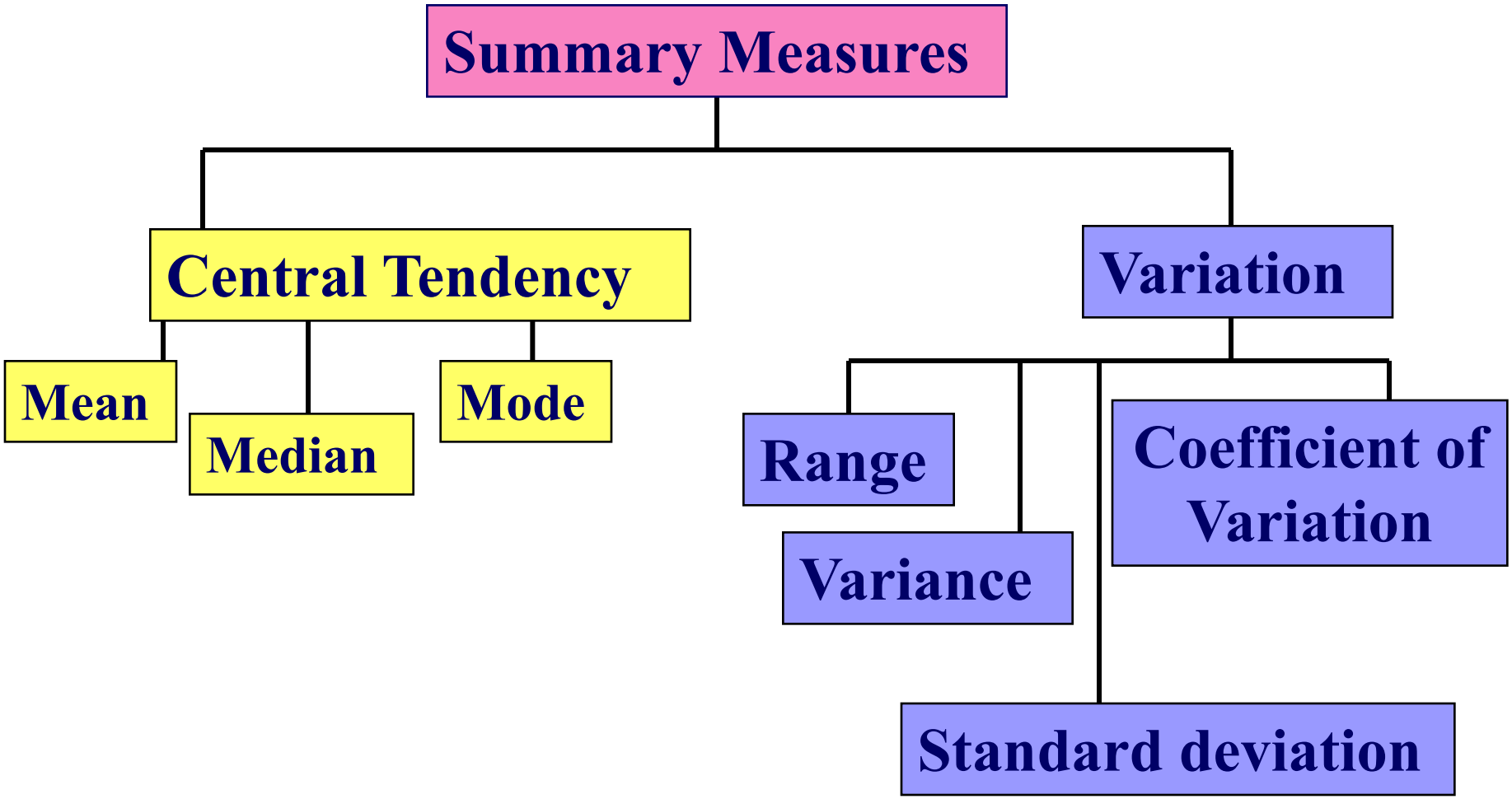


Exploratory Data Analysis

EDA- Exploratory Data Analysis

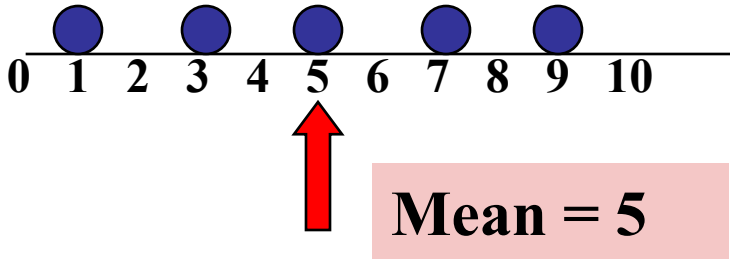
- Measures of Central tendency
- Measures of dispersion
- Covariance
- Correlation

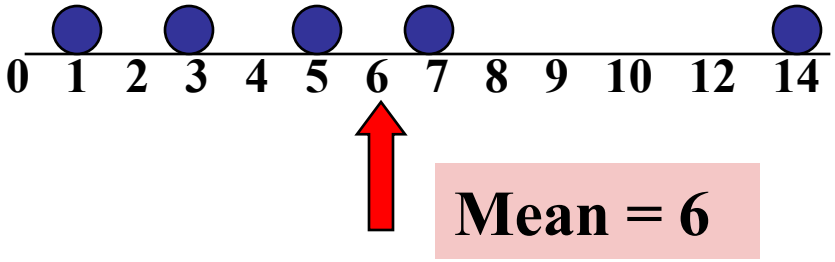
Summary Measures



Mean (Arithmetic Mean) *(continued)*

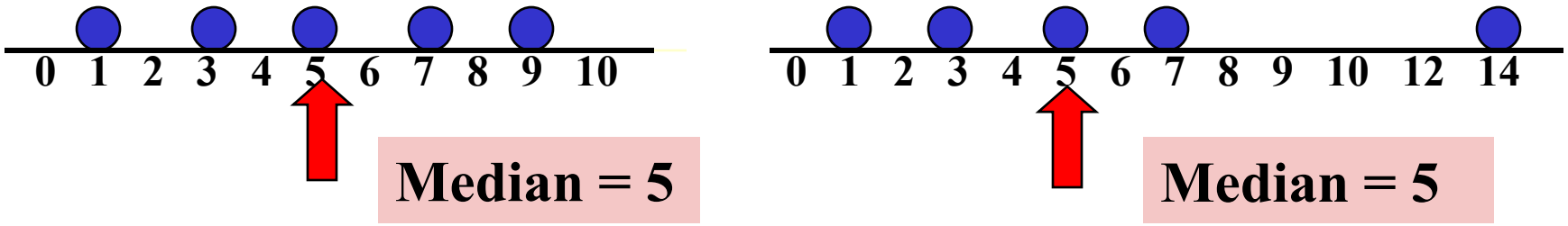
- The most common measure of central tendency
- Affected by extreme values (outliers)





Median

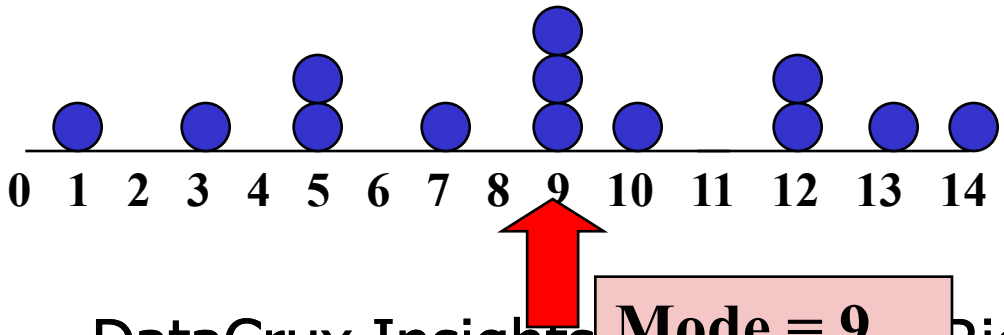
- Robust measure of central tendency
- Not affected by extreme values

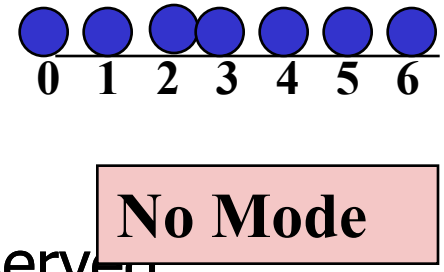


- In an ordered array, the median is the "middle" number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the two middle numbers

Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes



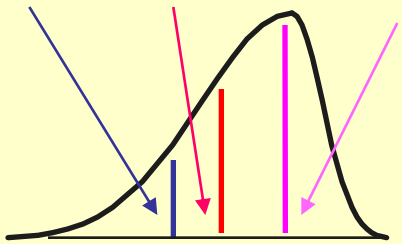


Shape of a Distribution

- Describes how data is distributed
- Measures of shape
  - Symmetric or skewed

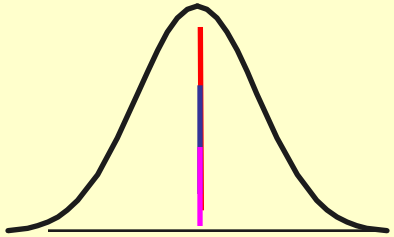
Left-Skewed

Mean < Median < Mode



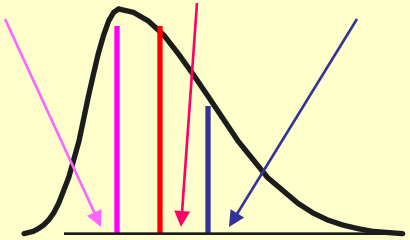
Symmetric

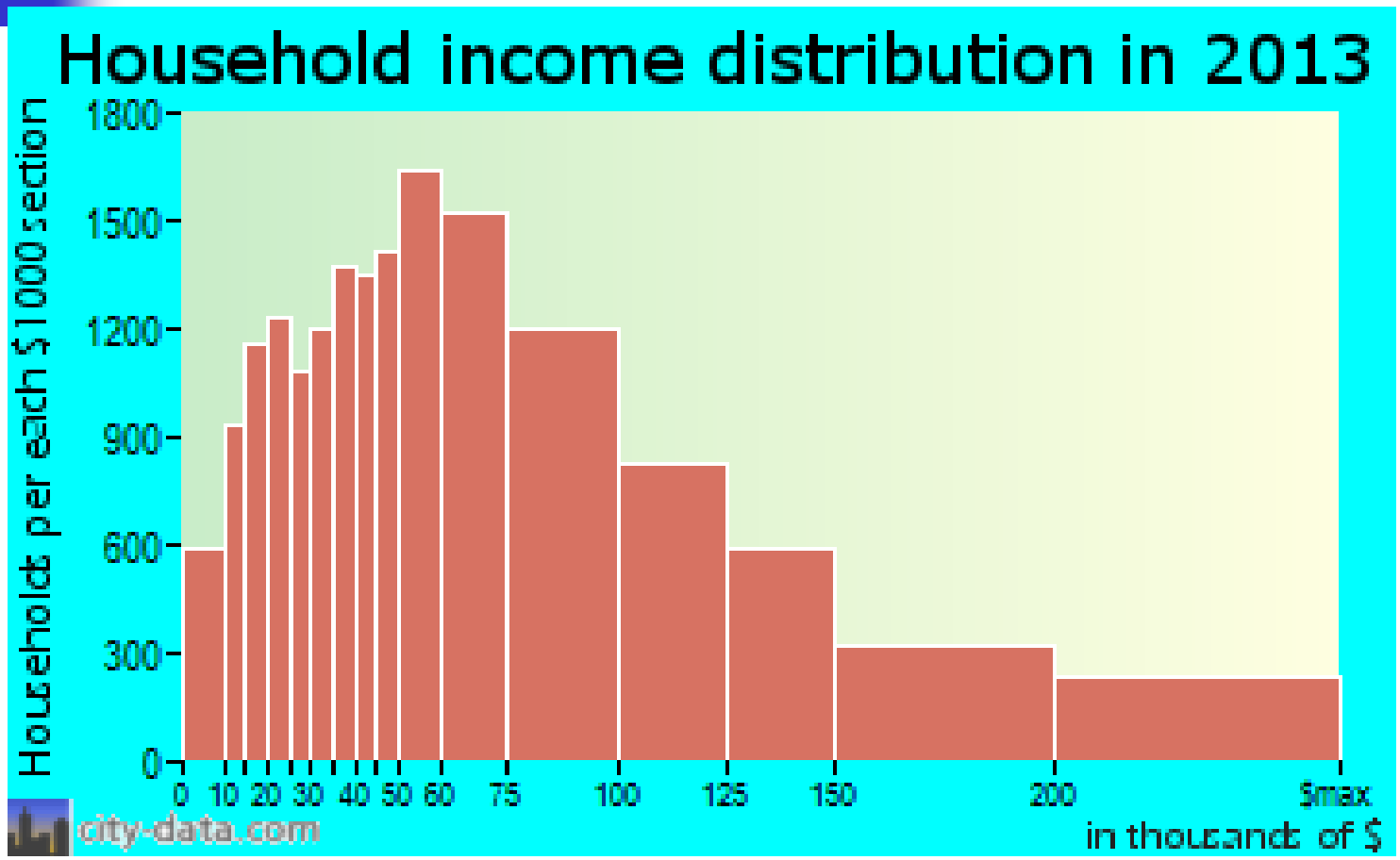
Mean = Median =Mode

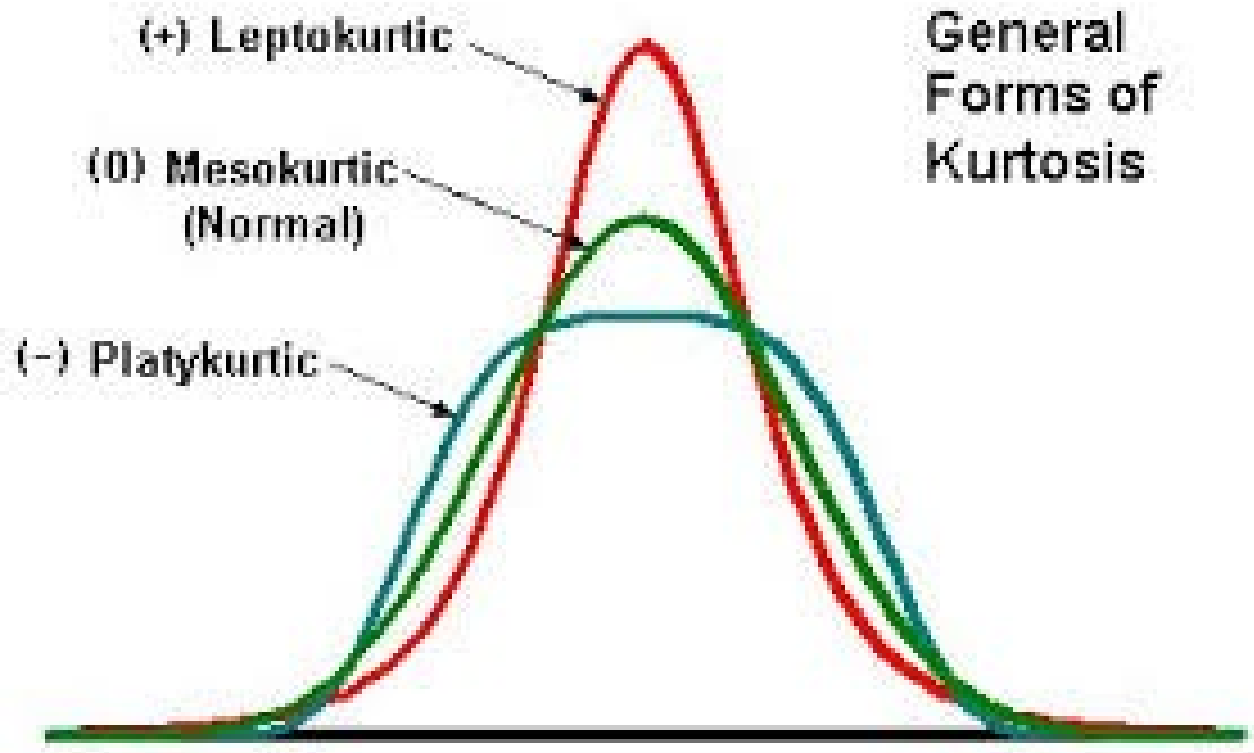


Right-Skewed

Mode < Median < Mean





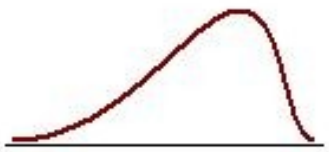


$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

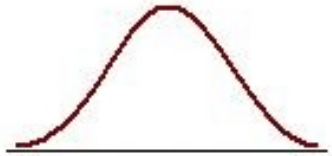
$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Skewness

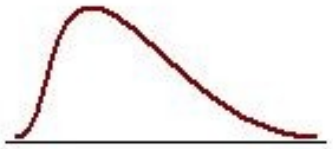
The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.



Negatively skewed distribution
or Skewed to the left
Skewness <0



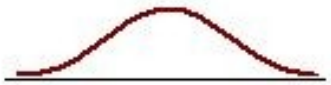
Normal distribution
Symmetrical
Skewness = 0



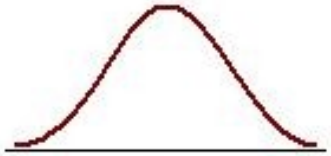
Positively skewed distribution
or Skewed to the right
Skewness > 0

Kurtosis

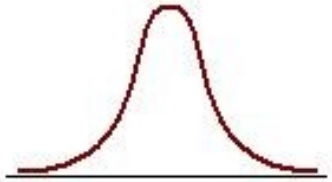
The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.



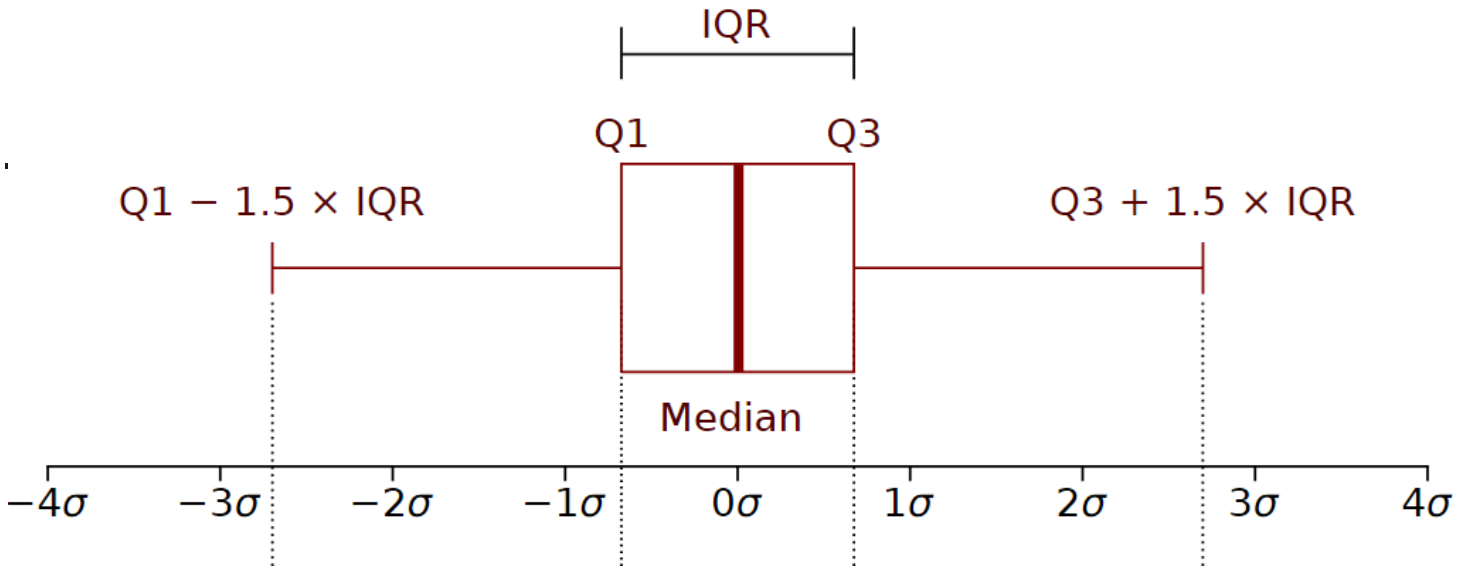
Platykurtic distribution
Low degree of peakedness
Kurtosis <0

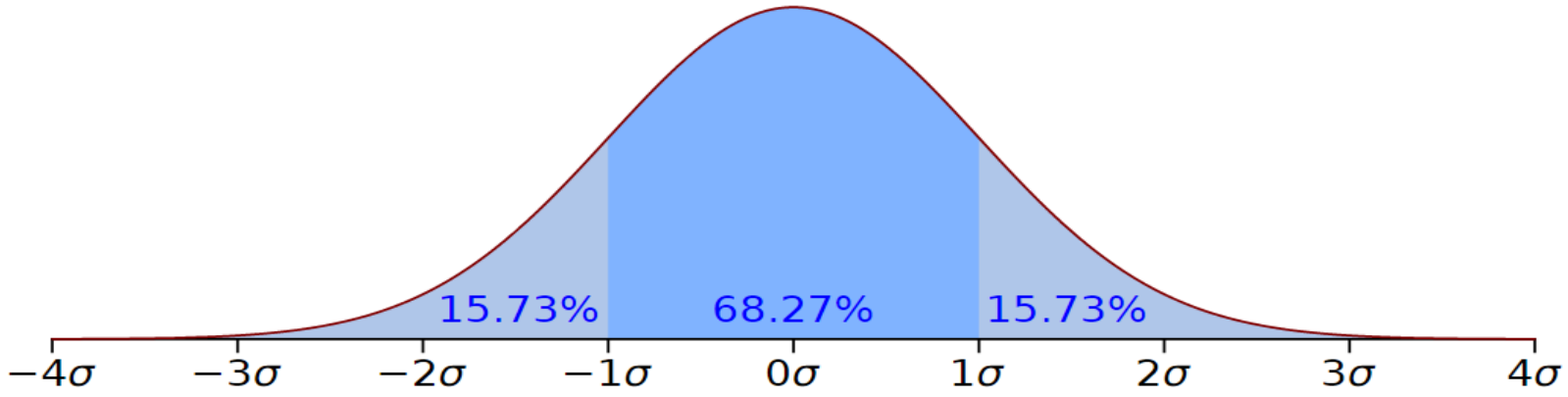


Normal distribution
Mesokurtic distribution
Kurtosis = 0

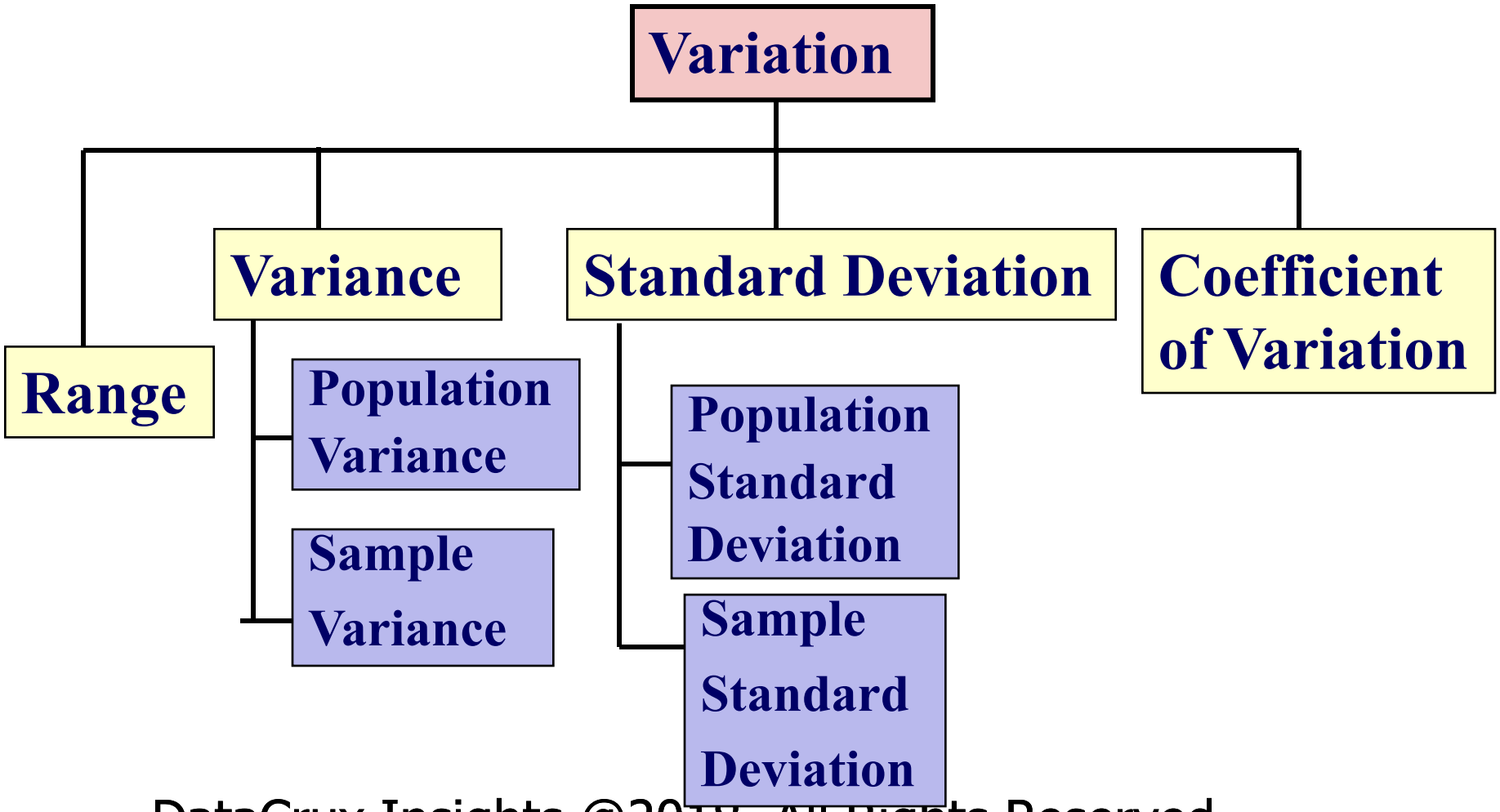


Leptokurtic distribution
High degree of peakedness
Kurtosis > 0





Measures of Variation



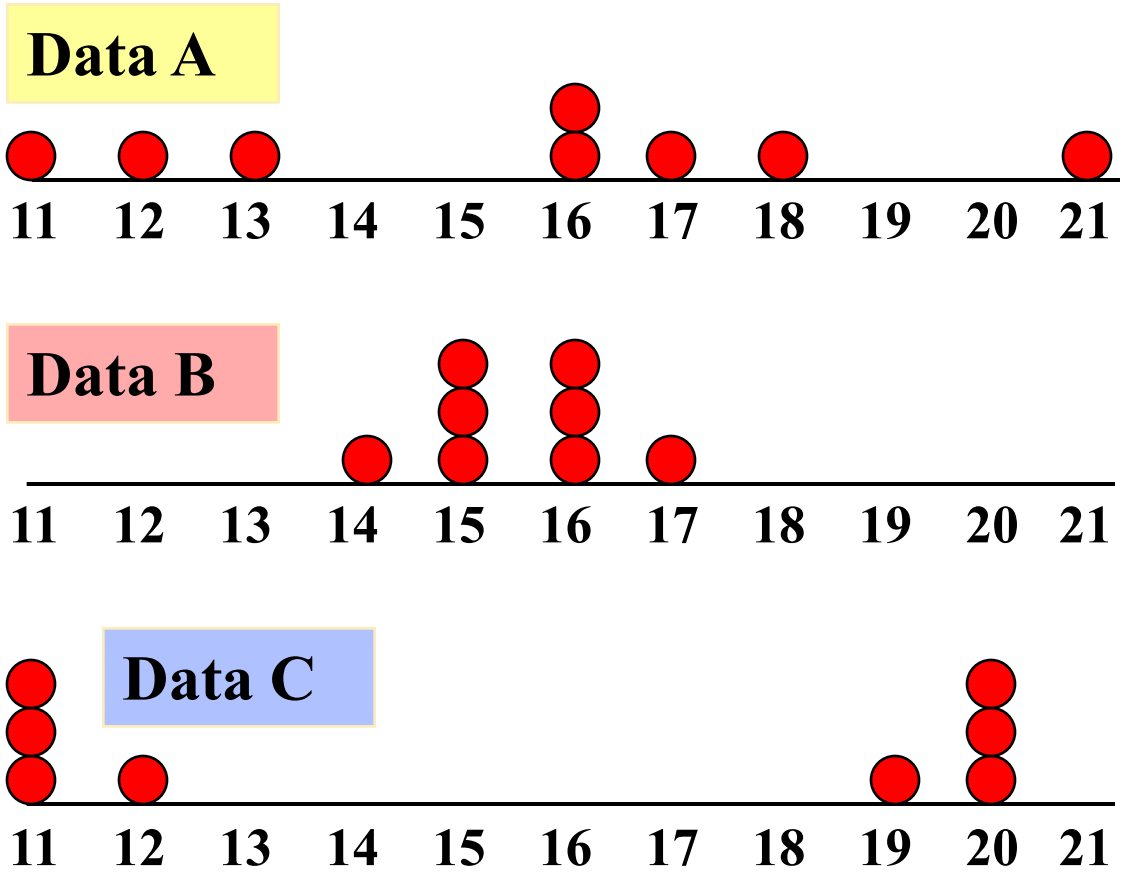
Range

- Measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$

- Ignores the way in which data are distributed

Comparing Standard Deviations



Mean = 15.5
s = 3.338

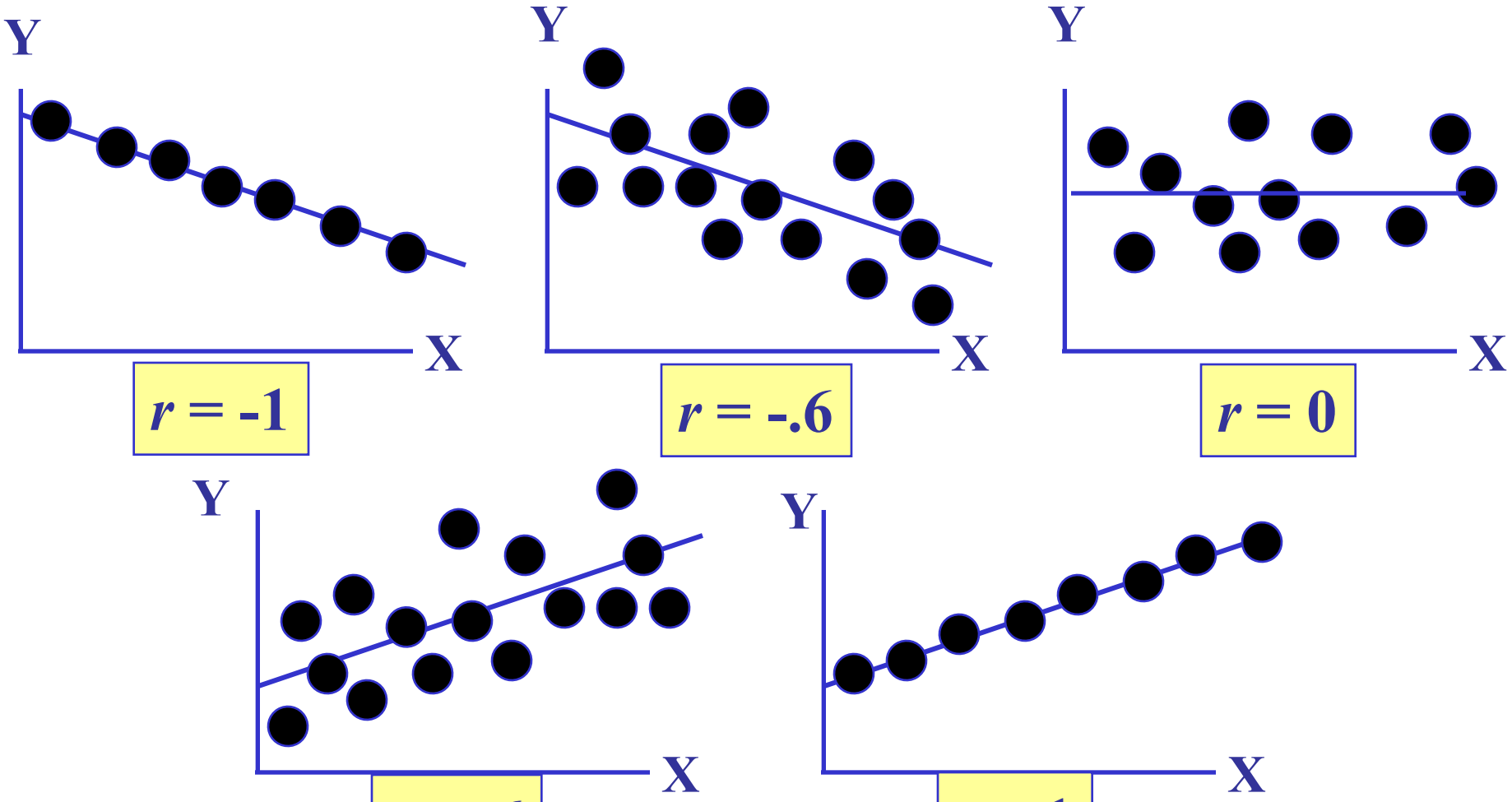
Mean = 15.5
s = .9258

Mean = 15.5
s = 4.57

Features of Correlation Coefficient

- Unit free
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients



$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$