PIAS
PHENIKAA INSTITUTE FOR ADVANCED STUDY

PHENIKAA UNIVERSITY
Realizing your potential

DATA ANALYSIS

STATISTICAL ANALYSIS

Anh-Tuan Nguyen
Phenikaa School of Computing

EXAMPLE

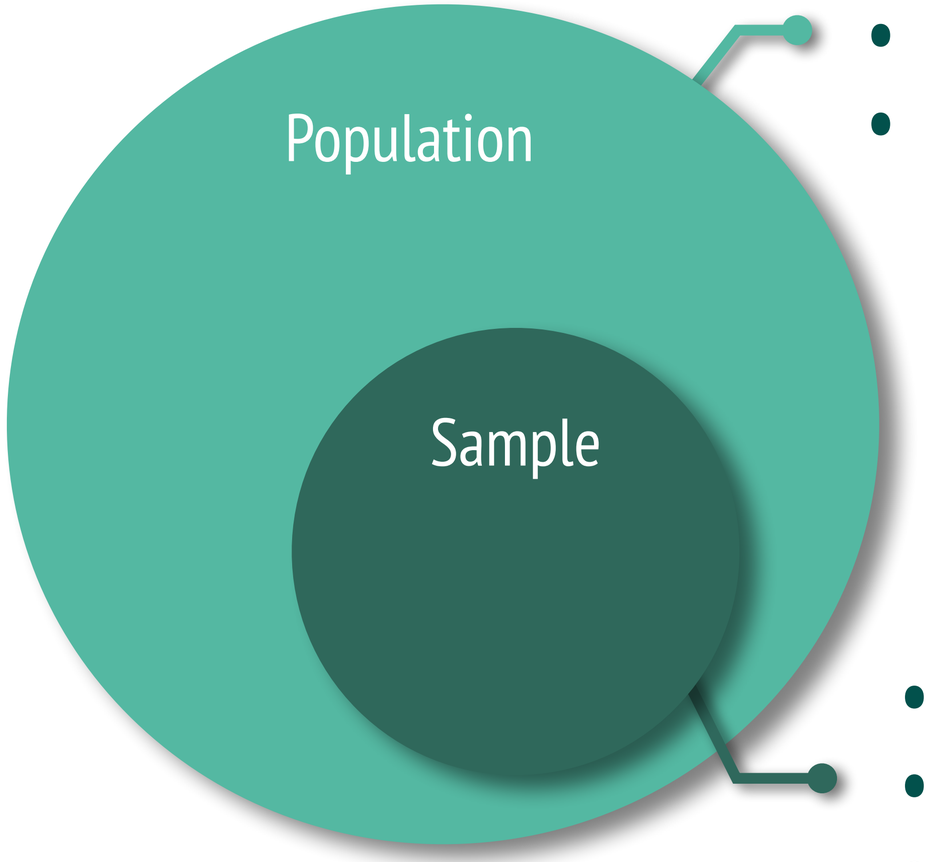
- Male/Female Ratio?
- Number of consumers?
- Average of age?
- Correlation between income and age?
- Correlation between income and purchasing capacity?
- Will customers buy goods are males or females?
- Anticipating a customer's purchasing ability?
- ...

→ **Synthesis report? Building business scenarios from data?**

OUTLINE

- Statistical values
- Correlation analysis
- Analysis of variance
- Example
- Assignment

SAMPLE/ POPULATION



- The population is a complete set
- Reports are a true representation of opinion
- Contains all member of a specific group

- The sample is a subset of the population
- Reports have a margin of error and confidence interval
- A subset represents the entire population

MIN/ MAX

- MinValue is also known as infimum. MinValue is typically used to find the smallest possible values given constraints.

- MaxValue is also known as supremum. MaxValue is typically used to find the largest possible values given constraints.

MEAN

- This is sometimes known as the average of the data.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

μ mean
N number of data points
x_i data points

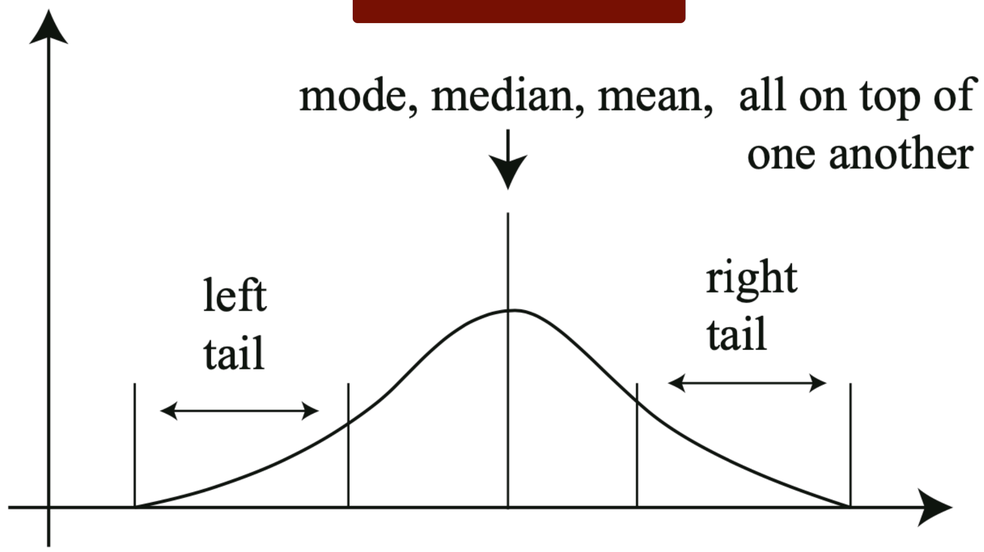
- The mean is sensitive to outliers in the data.

MEDIAN/MODE

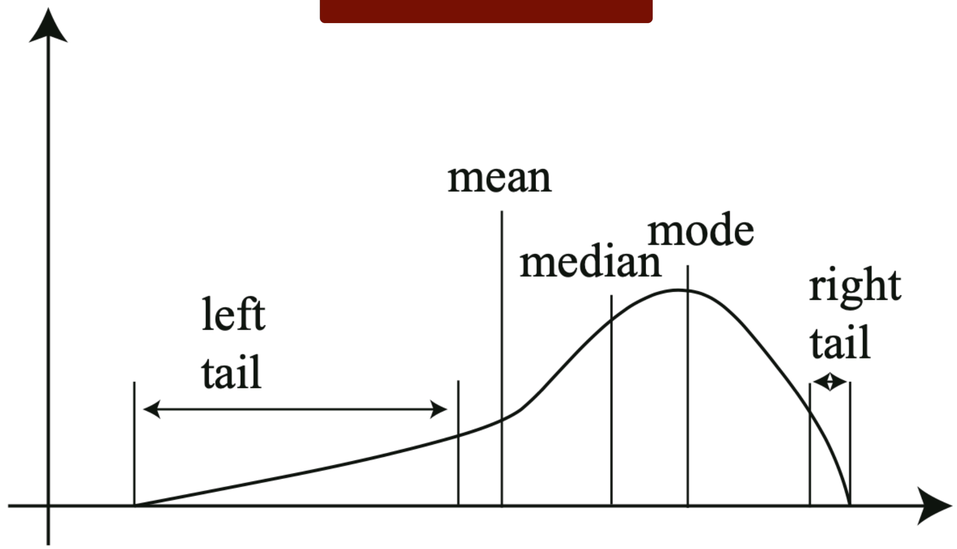
- Median is the midpoint of the data, and is calculated by either arranging it in ascending or descending order. If there are N observations.

- Mode is the most repetitive data point in the data

Symmestrical distribution



Asymmestrical distribution

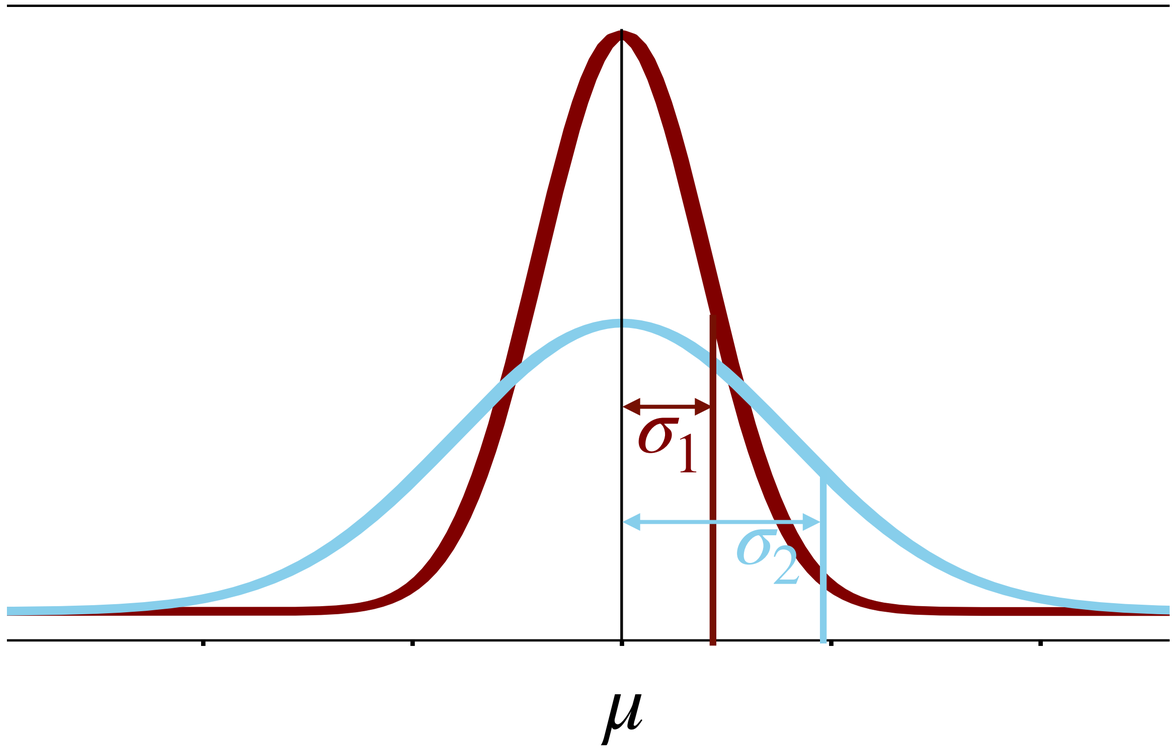


STANDARD DEVIATION/ VARIANCE

- We would also like to know the extent to which data items are close to the mean. This information is given by the standard deviation, which is the root mean square of the offsets of data from the mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$var = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$



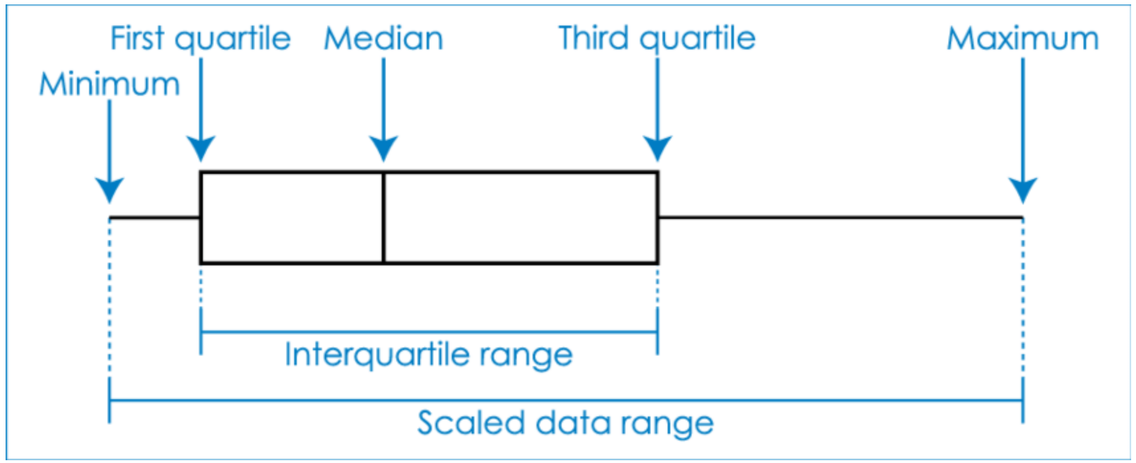
INTERQUARTILE RANGE

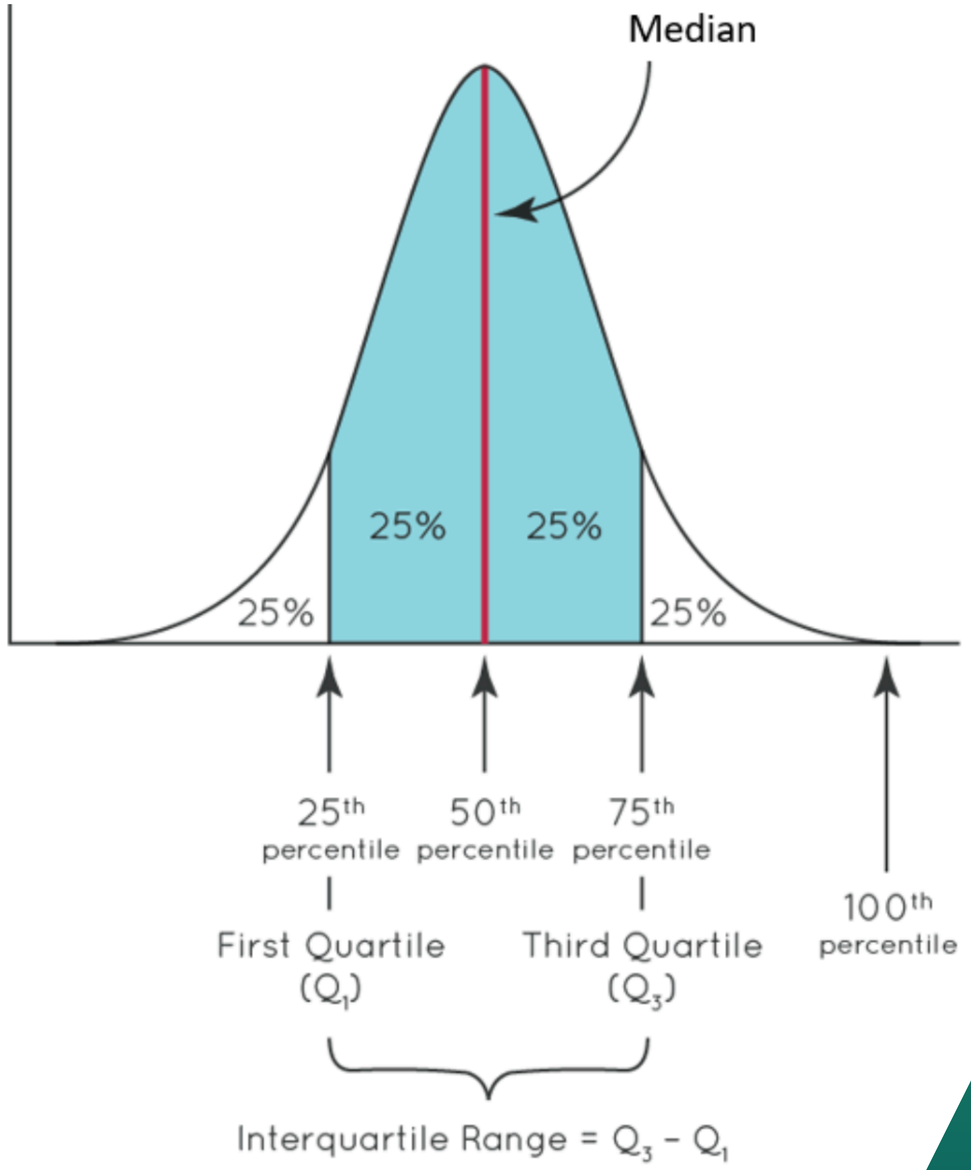
Percentile

The k'th percentile is the value such that k% of the data is less than or equal to that value.

Quartiles

k = 25, 50, 75, 100

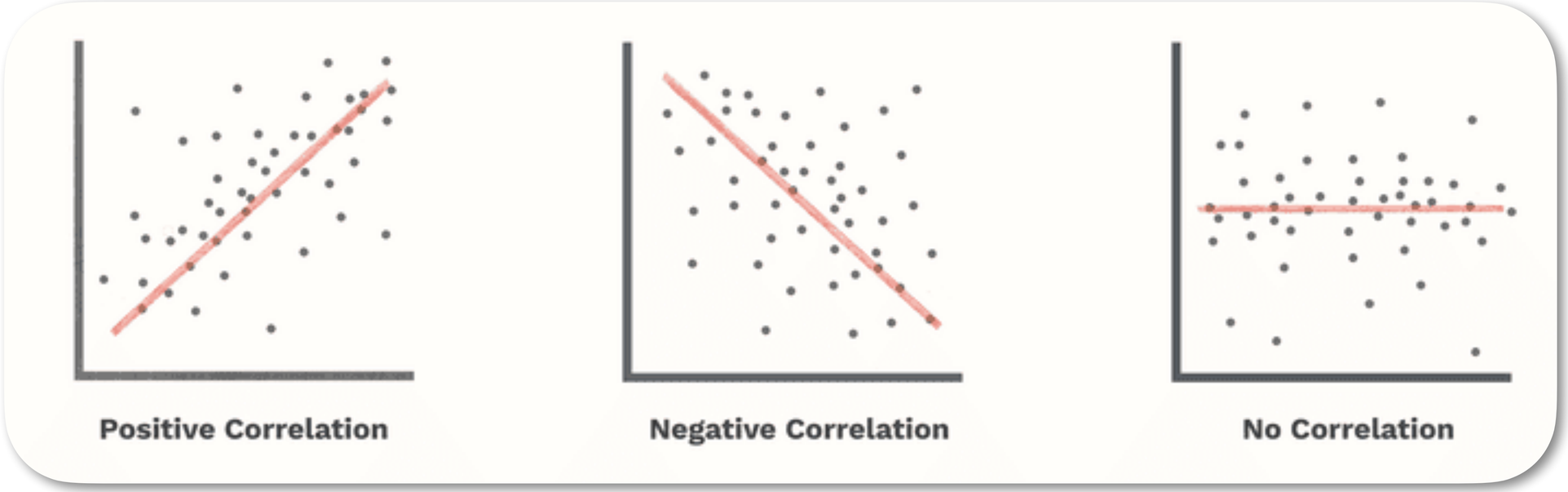




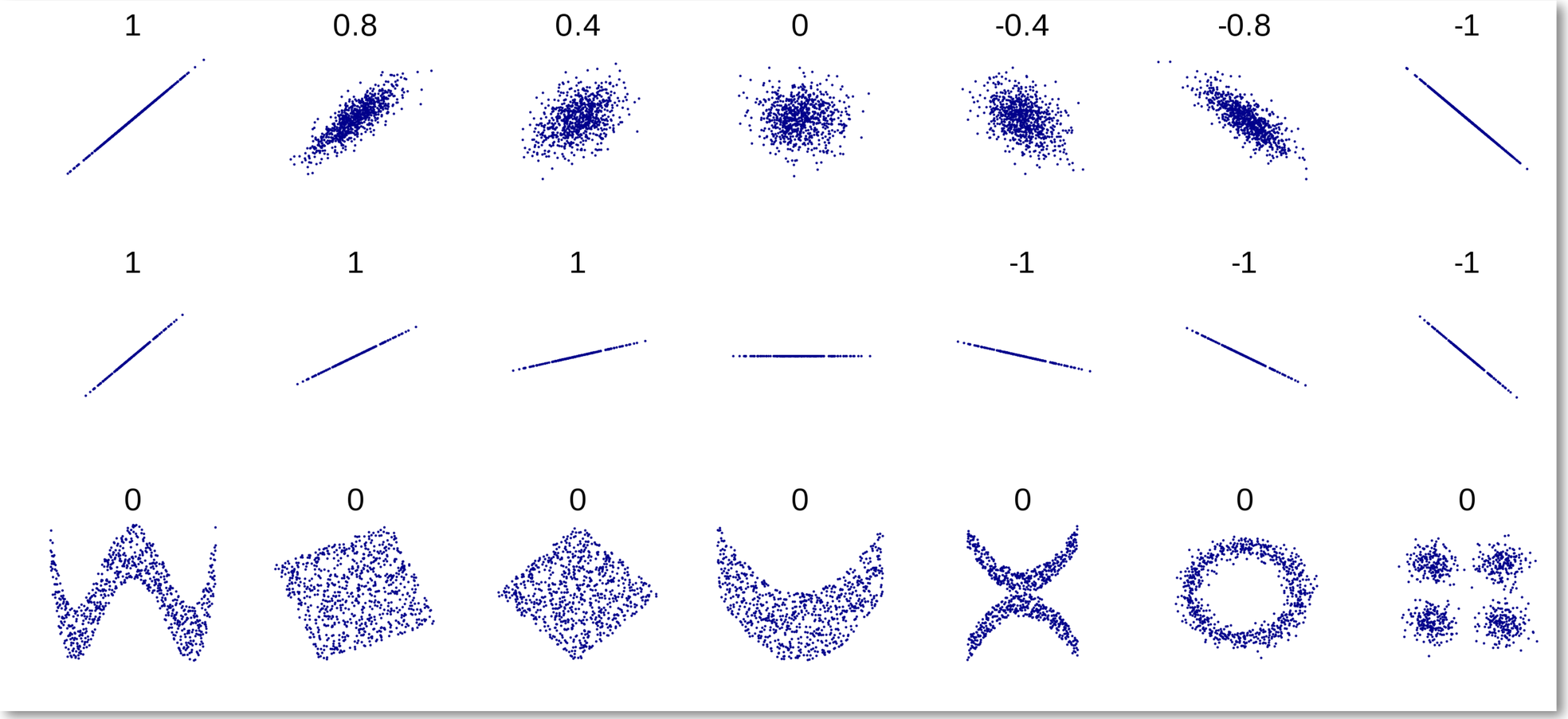
CORRELATION COEFFICIENT/ COVARIANCE

- Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y

- Covariance is a measure of how two variables change together



CORRELATION COEFFICIENT/ COVARIANCE



CORRELATION COEFFICIENT/ COVARIANCE

$$Cov(x, y) = \sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

$$corr = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N} (x_i - \mu_x)^2 (y_i - \mu_y)^2}}$$

$$-1 < corr < 1$$

CORRELATION COEFFICIENT/ COVARIANCE

PHENIKAA UNIVERSITY
Realizing your potential

- Covariance matrix?

The covariance matrix is also known as the variance-covariance matrix, as the diagonal values of the covariance matrix show variances and the other values are the covariances.

$$C = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T$$

$$C = \begin{bmatrix} \sigma_{(x,x)} & \sigma_{(x,y)} \\ \sigma_{(y,x)} & \sigma_{(y,y)} \end{bmatrix}$$

ASSIGNMENT

PHENIKAA UNIVERSITY
Realizing your potential

Using The Boston housing prices dataset from sckit learn, calculate basic statistic values with and without numpy for one feature/attribute only. With corr, chose two features

Plot histogram, boxplot

```
Boston house prices dataset
---------------------------

**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's
```