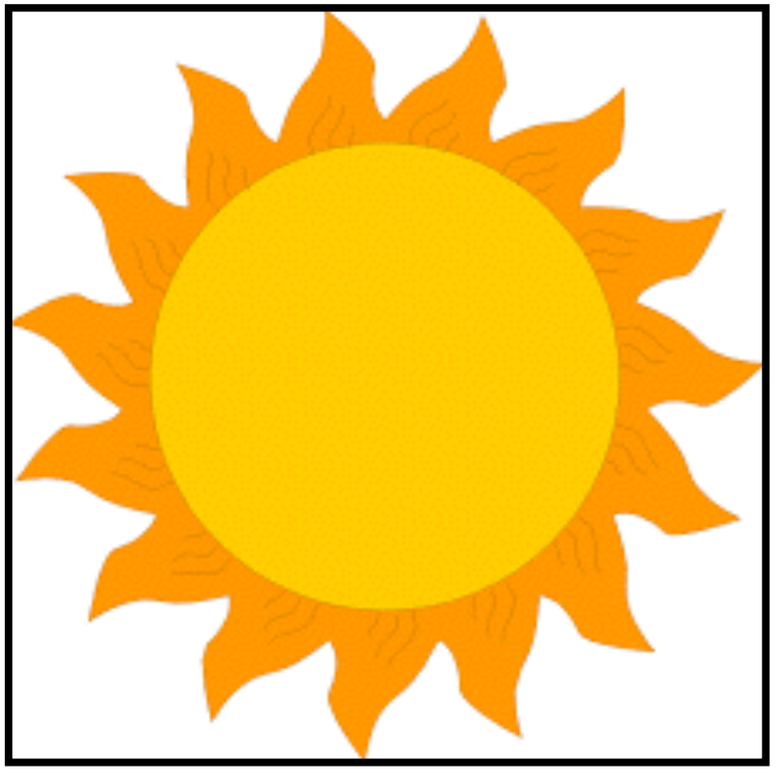
GENERAL ASSEMBLY

EXPLORATORY DATA ANALYSIS

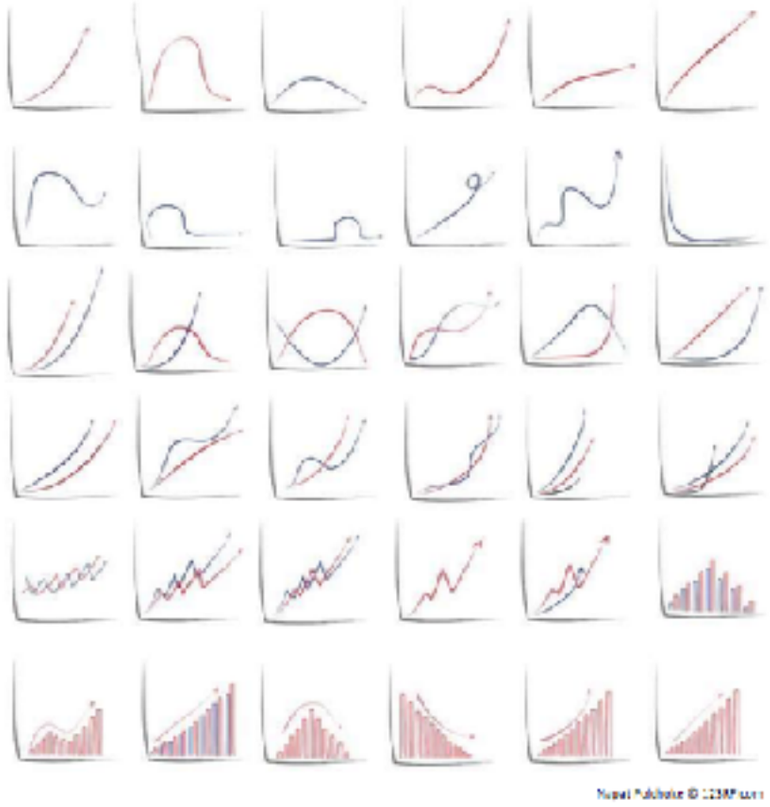
Gus Ostow

OBJECTIVES

- After this class you should be able to:
  - Compare and contrast measures of central tendency (mean, median, mode)
  - Use the pandas library to analyze datasets using basic descriptive statistics
  - Create data visualizations in pandas - including boxplots, histograms and scatter plots - to discern trends in the data

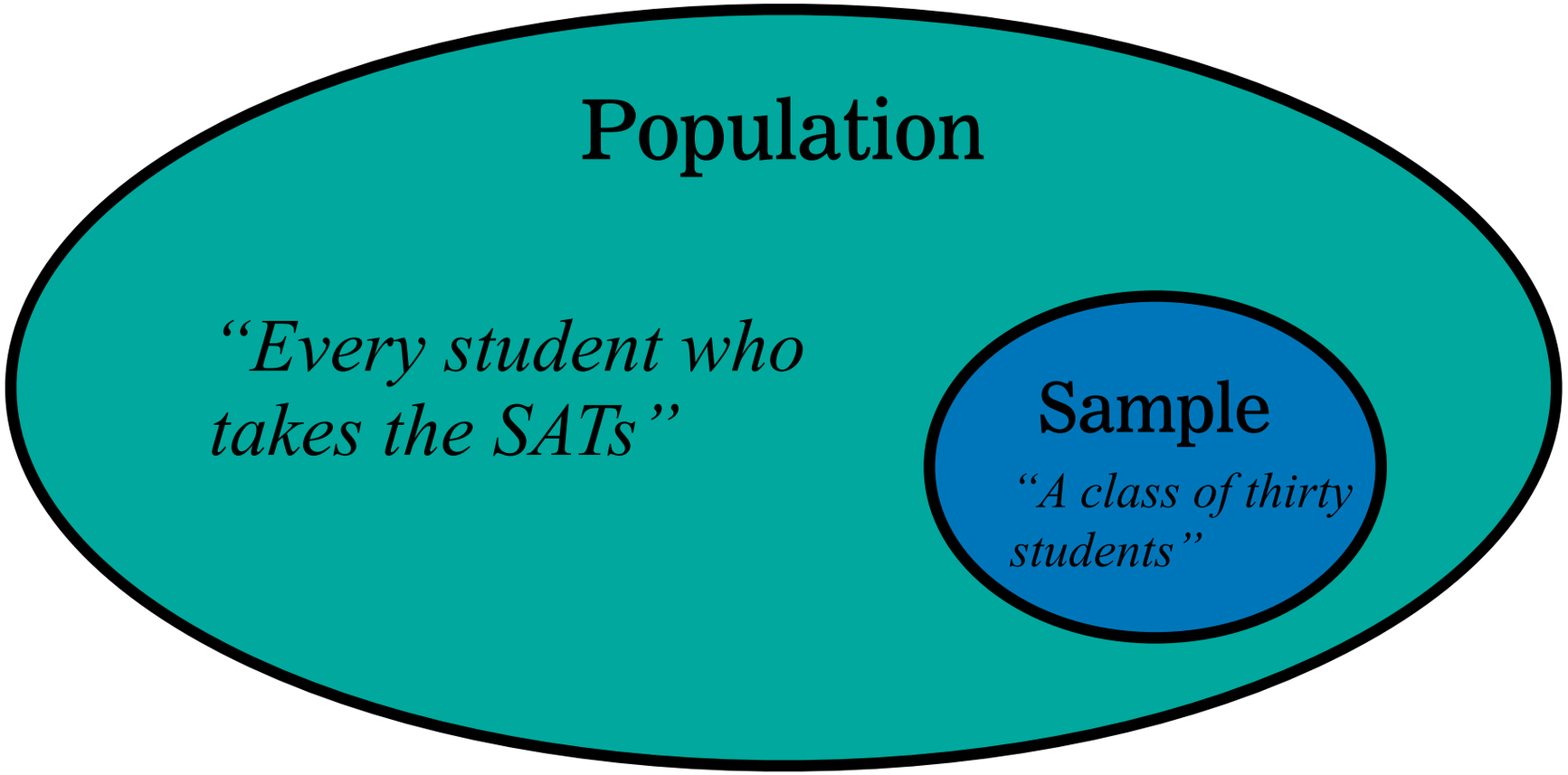


FOUNDATIONAL STATISTICS REQUIRED FOR EDA



- Descriptive statistics vs inferential statistics
- Measures of central tendency and spread
- Plots
  - Boxplot
  - Histogram
  - Scatterplots
- Correlation and covariance

POPULATIONS AND SAMPLES



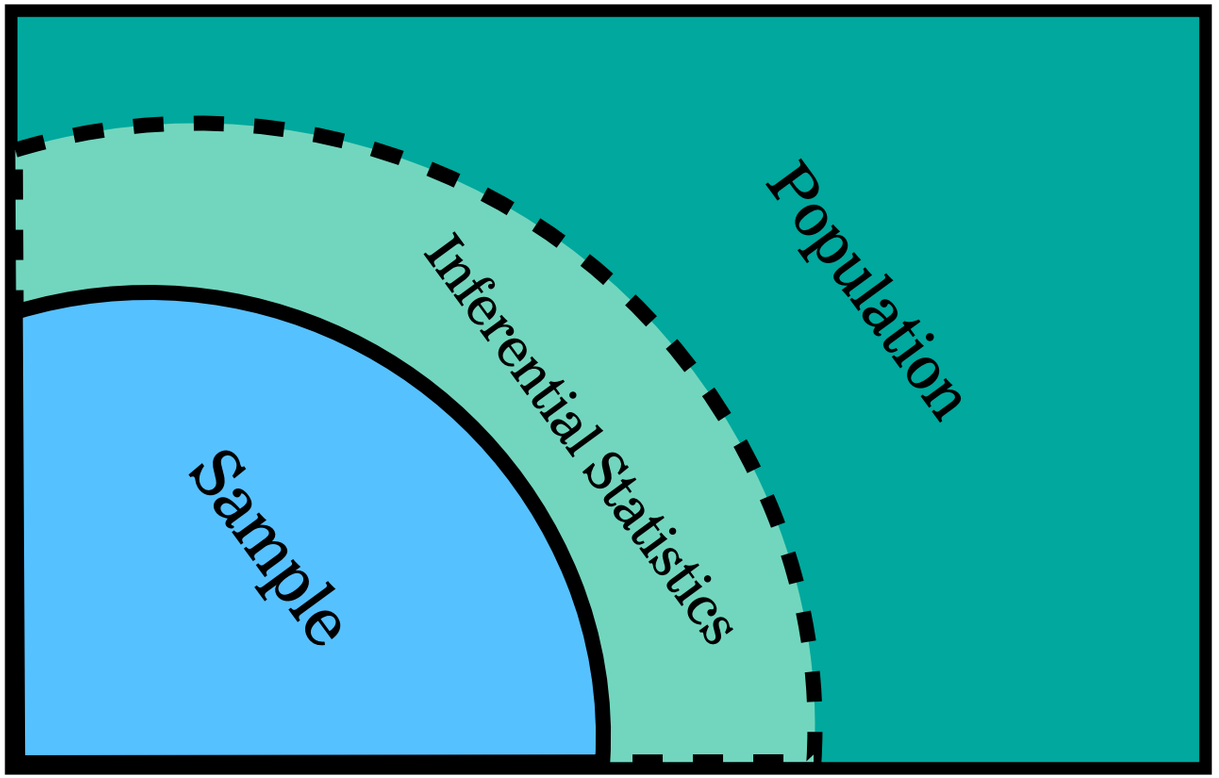
Descriptive Statistics

- Information actually observed about samples
  - “What were sales”
  - “Average score”
  - “How correlated were …”

Inferential Statistics

- Making statements about populations based on samples
  - “Layout A will result in more conversions than B”
  - “Smoking triples the odds of developing lung cancer”
  - “This trend is likely to continue”

EDA IS CONCERNED WITH SAMPLES

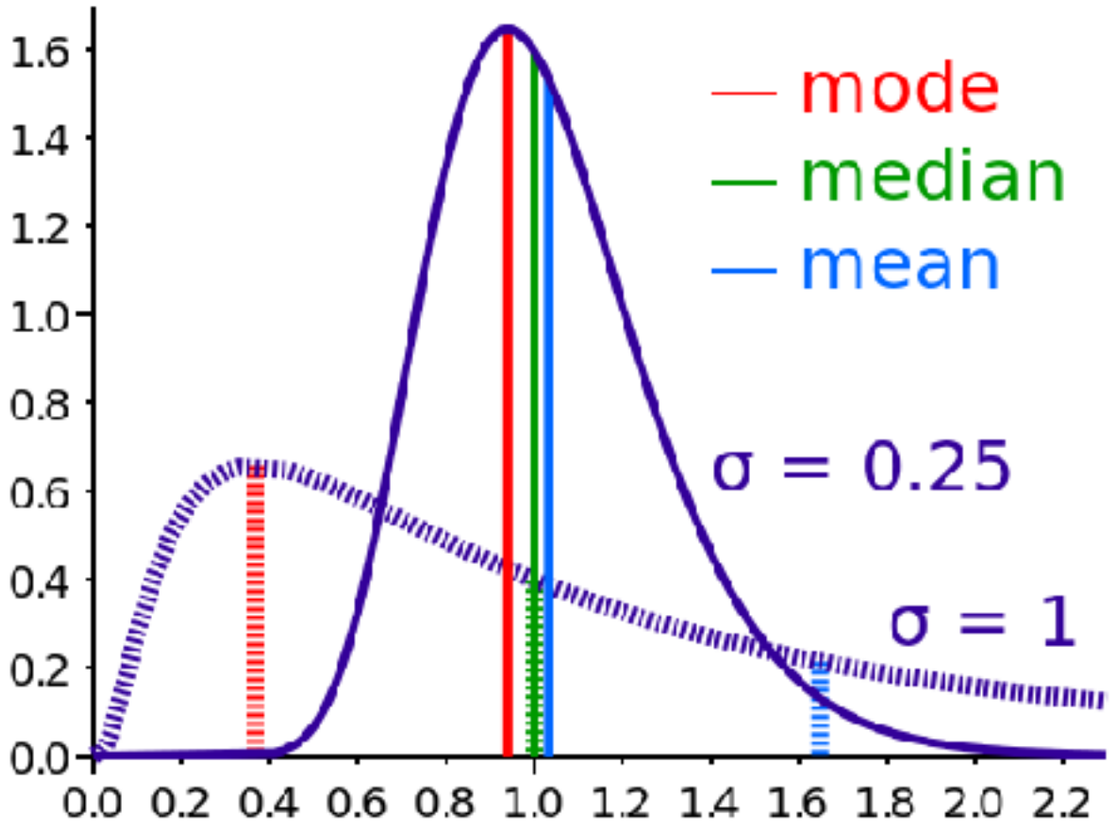


Exploratory data analysis is an open-ended investigate of the data as a sample, using descriptive statistics

MEASURES OF CENTRAL TENDENCY AND MEASURES OF SPREAD

If you want to describe the data with a single number

- Mean
  - “The average”
- Median
  - “The middle value”
- Mode
  - “The most common value”



If you want to describe the data with a single number

Mean

- Pros
  - Commonly used and well understood
  - Extensively used in probability and statistics (expected value)
- Cons
  - Sensitive to outliers
  - Value probably isn't in dataset

Median

- Pros
  - Robust against outliers
  - Value is in the dataset (with odd row counts)
- Cons
  - Less often used in inferential statistics
  - Sensitive to bimodal distributions

Mode

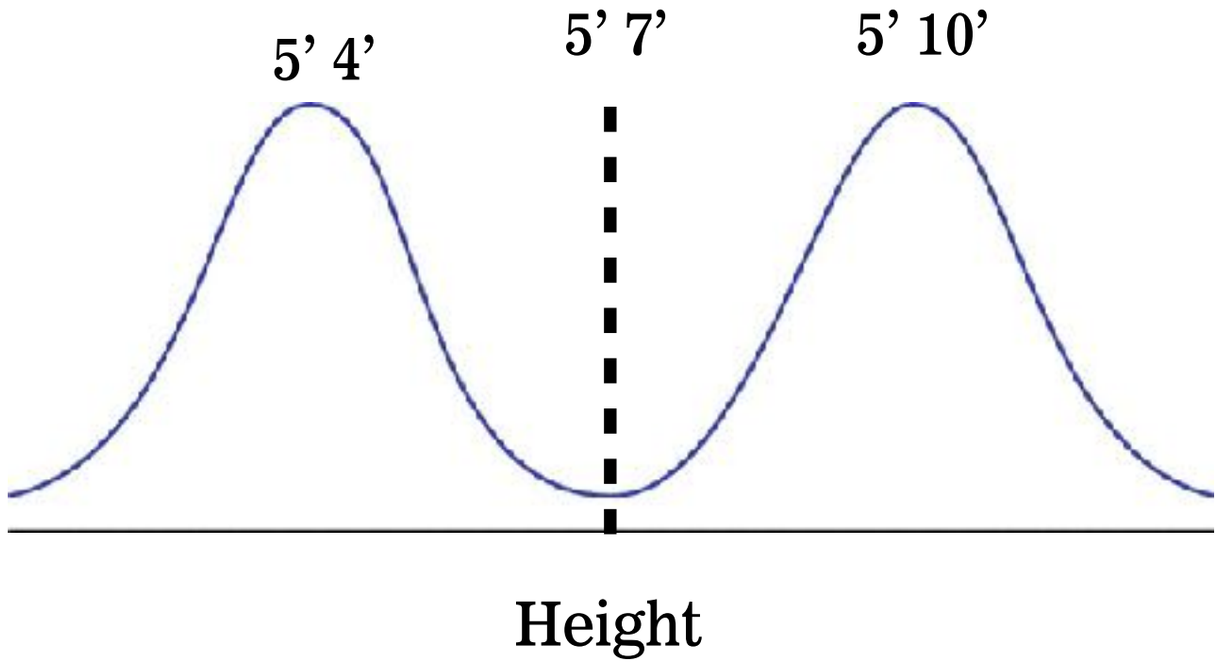
- Pros
  - Actually the most common value
- Cons
  - Might not be a clear mode

Stumbling point examples

- Mean
  - Issue
    - “The average number of limbs humans have is less than four”
  - Solution
    - “The median number of limbs humans have is four”

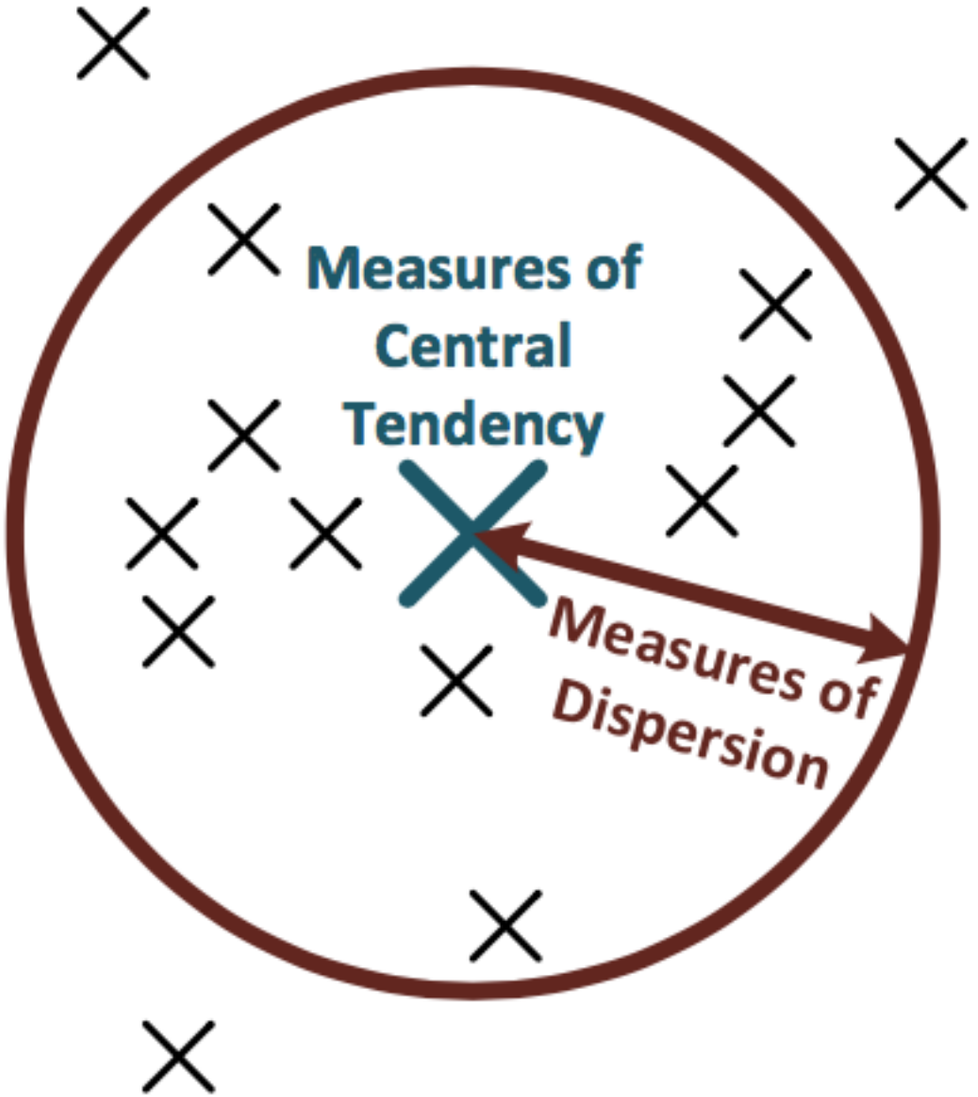
Stumbling point examples

- Median
  - Issue
    - The median human height is 5’ 7”
  - Solution
    - The median height of men is 5’ 10” and women is 5’ 4”
    - The modes of human height are 5’ 4” and 5’ 10”



SPREAD

- Measures of central tendency
  - “What is typical or common”
  - Mean, median, mode
- Measures of spread
  - “How far do values stray from the center”
  - Variance, standard deviation, range, Interquartile range



Variance and Standard Deviation

- Variance
  - "The average squared distance from the mean"

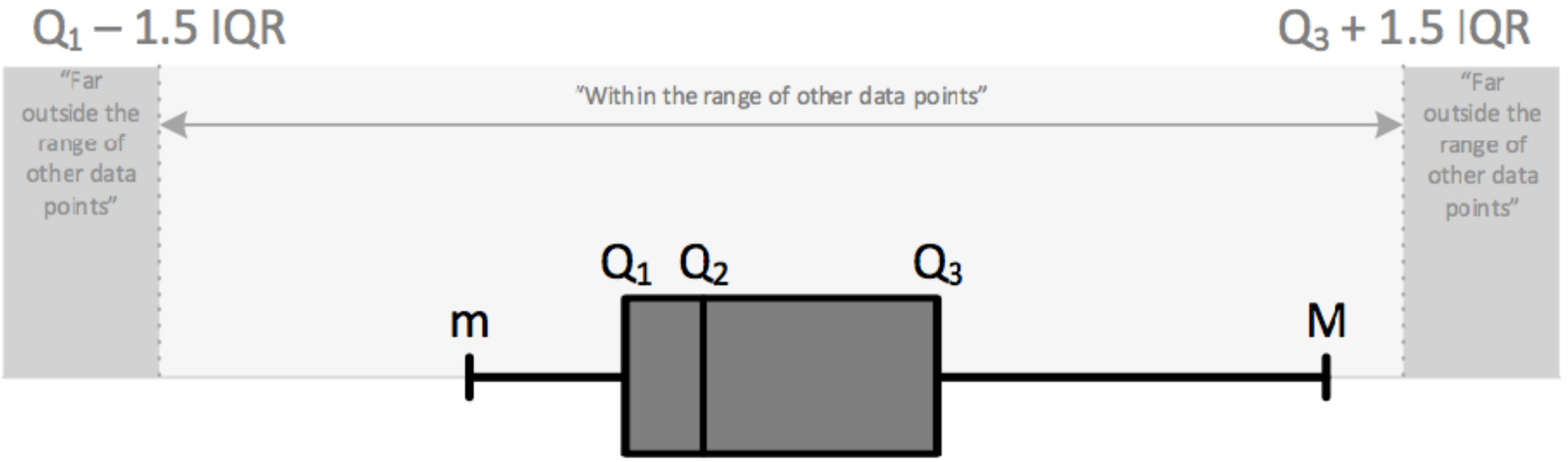
$$var = \sum_{i=0}^{n} \frac{(x_i - \mu)^2}{n}$$

- Standard deviation
  - "The square root of the variance"
  - In the same units as the original data

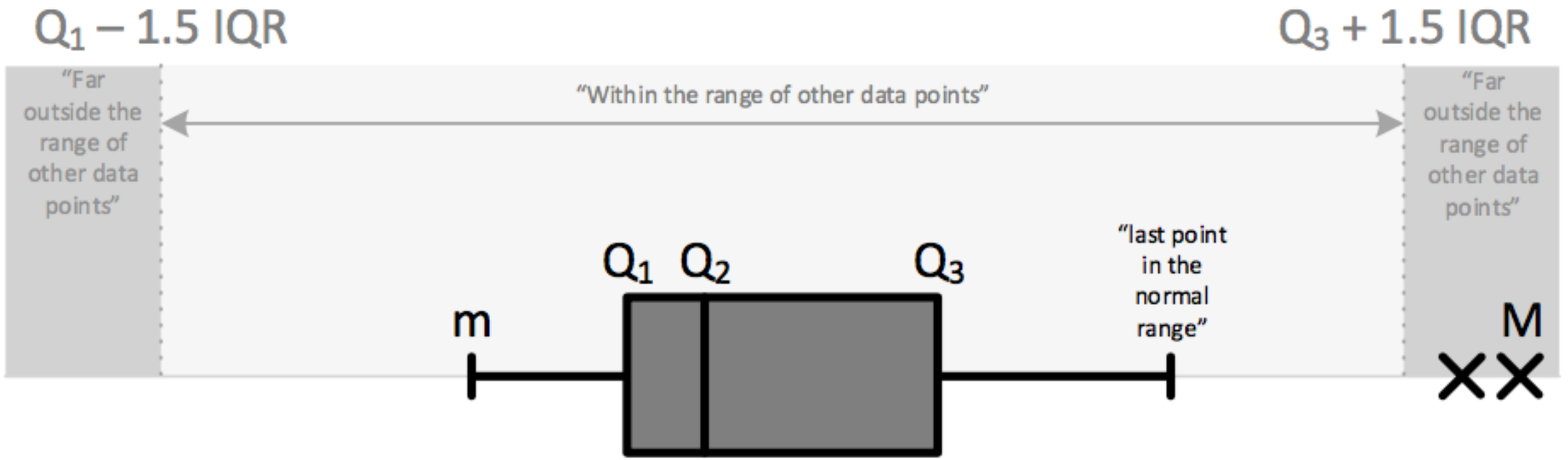
$$\sigma = \sqrt{var} = \sqrt{\sum_{i=0}^{n} \frac{(x_i - \mu)^2}{n}}$$

PLOT THE DATA – BOXPLOTS

Median, Range, Interquartile Range; no outliers

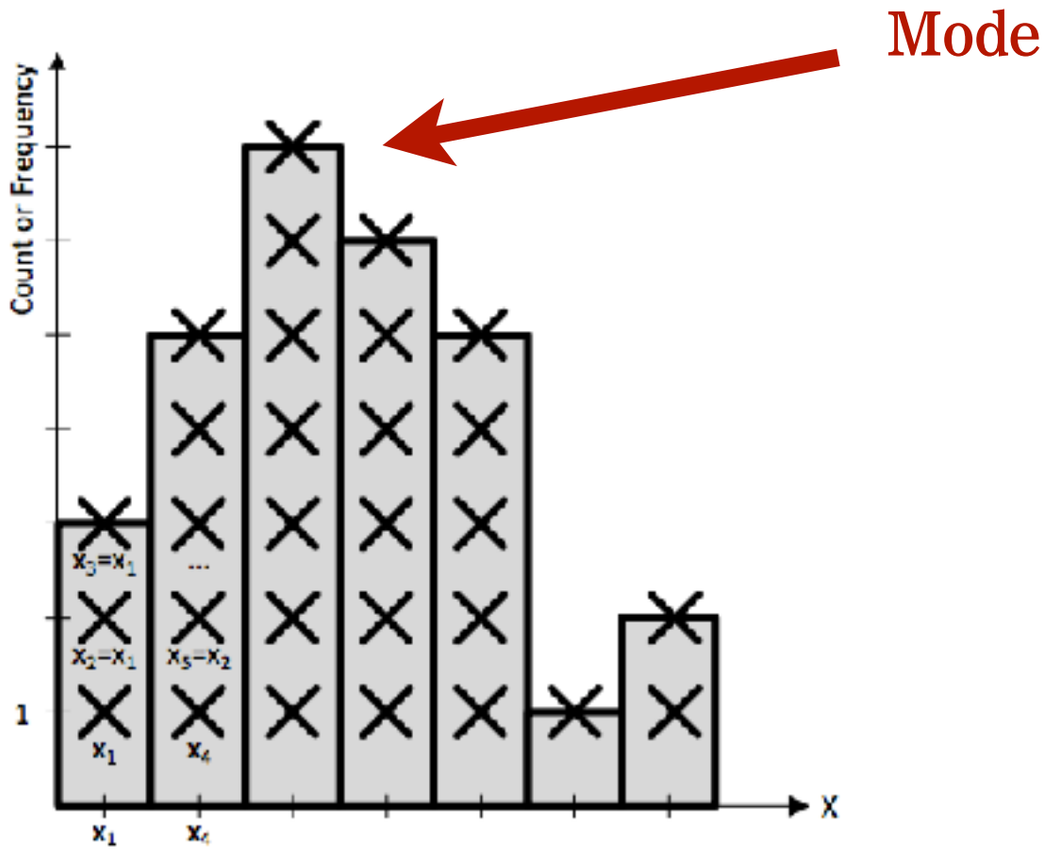


Median, Range, Interquartile Range; with outliers



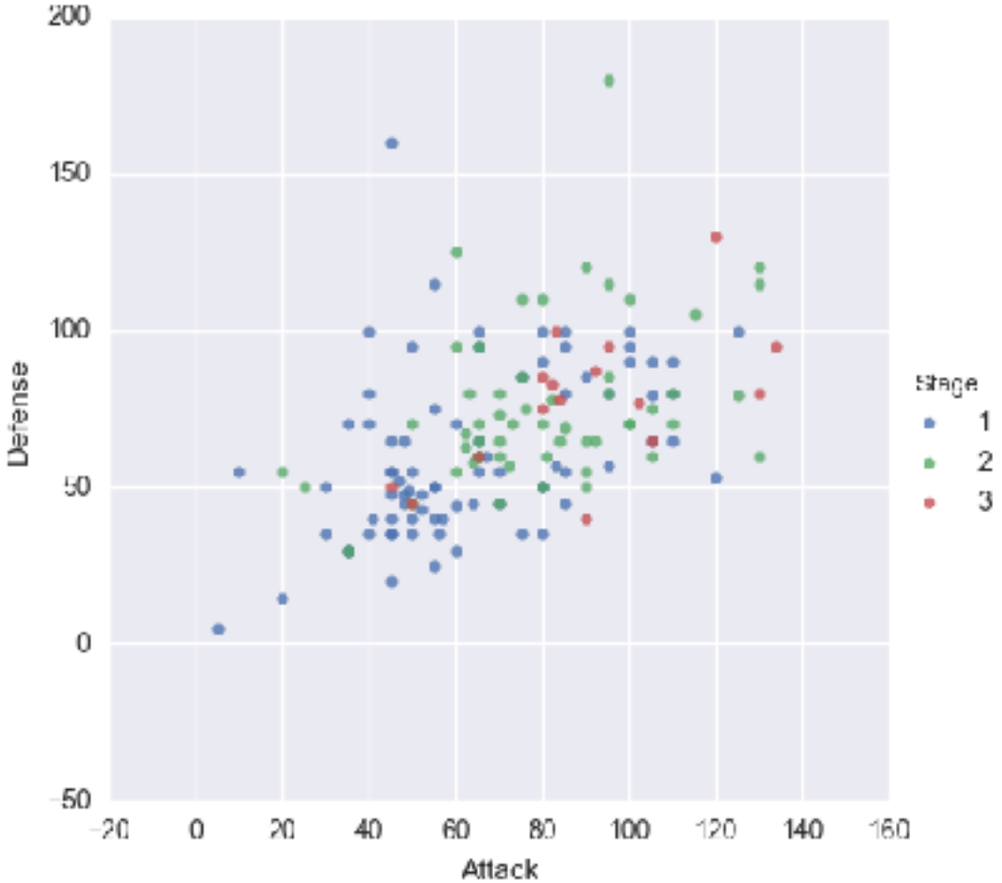
PLOT THE DATA – HISTOGRAMS

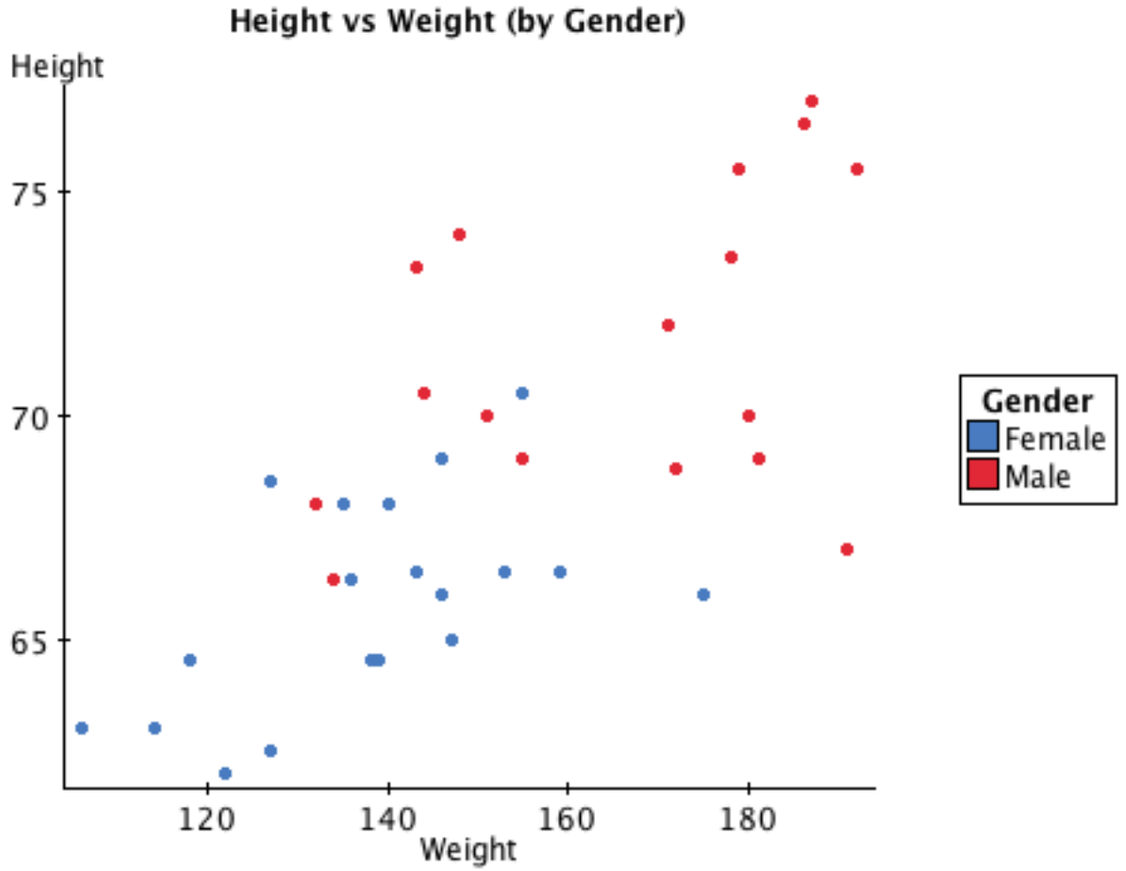
Display the true distribution when central tendency can be misleading



PLOT THE DATA - SCATTER PLOTS

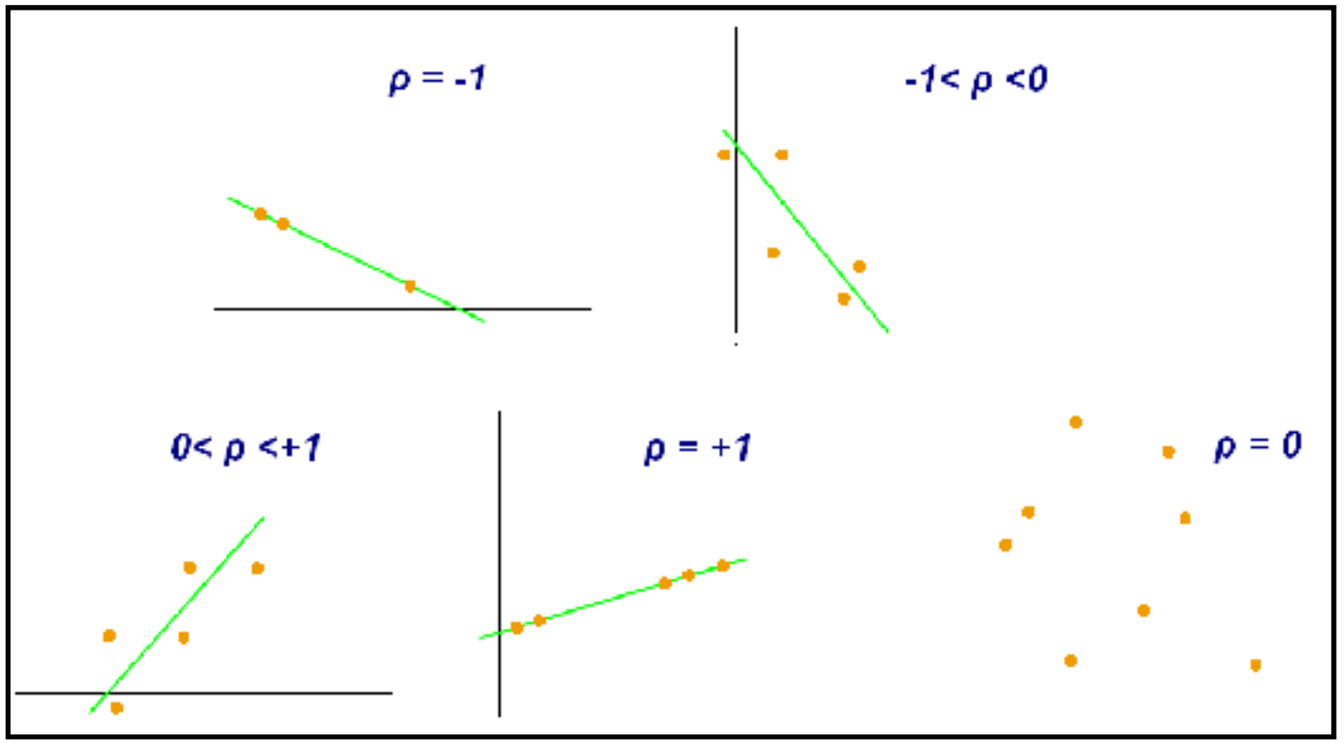
Scatter plots display the relationship between two continuous variables



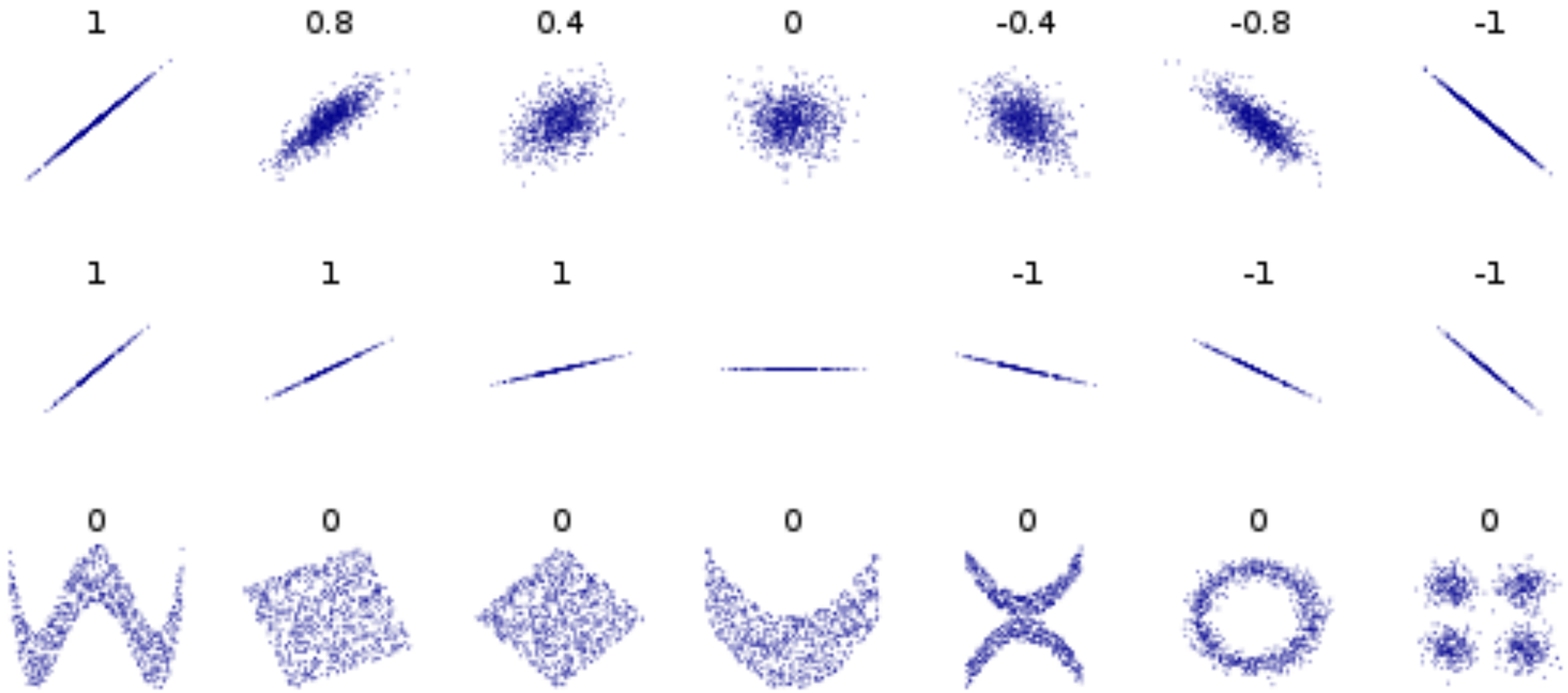


What is the (linear) association between two variables?

- Ranges from -1 to 1
  - 1: Both variables “move” together
  - 0: Variables don’t have any “linear relationship”
  - -1: Both variables move in opposite directions

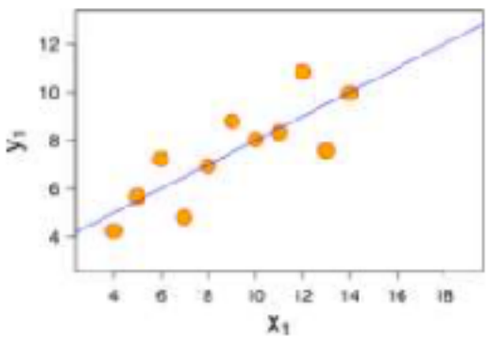


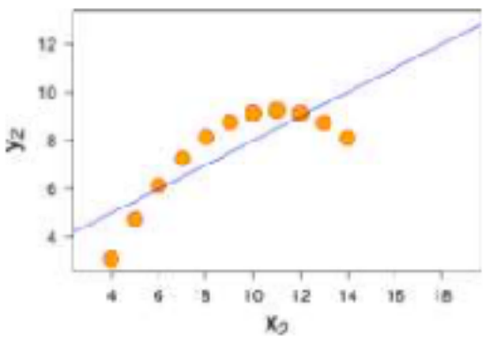
What is the (linear) association between two variables?



Don’t always rely on summary statistics - plot the data

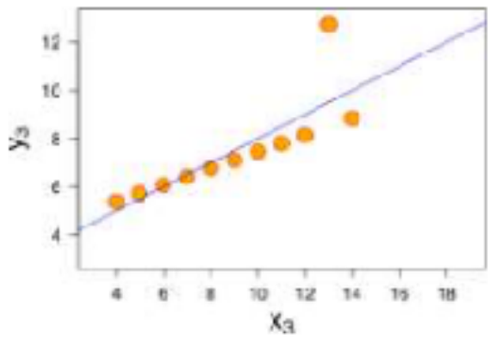
Scatter plot appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.

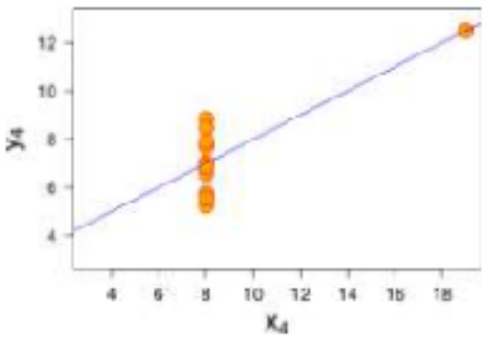




Not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the linear correlation is not relevant.

Distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line.





Example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

Property	Value
Mean of x	9
Sample variance of x_i	11
Mean of y	7.50
Sample variance of y_i	4.122 or 4.127
Correlation between x_i and y_i	0.816
Linear regression line in each case	$Y_i = 3.00 + 0.500\ x_i$

QUESTIONS?

LET'S DO THIS STUFF IN PANDAS