Introduction to Statistics

FinTech

Outline

- Summary statistics, such as mean, median, mode, variance, and standard deviation
- Plotting, characterizing, and quantifying a normally distributed dataset.
- Qualitatively and quantitatively identifying potential outliers in a dataset.
- Differentiating between a sample and a population in regard to a dataset.
- Defining and quantifying correlation between two factors.
- Calculating and plotting a linear regression.

What are the three measures of central tendency?

Measure of Central Tendency = Center of a Dataset

The three most common measures are **mean**, **median**, and **mode.**

Mean

Mean is the sum of all values divided by the number of elements in a dataset.

Median

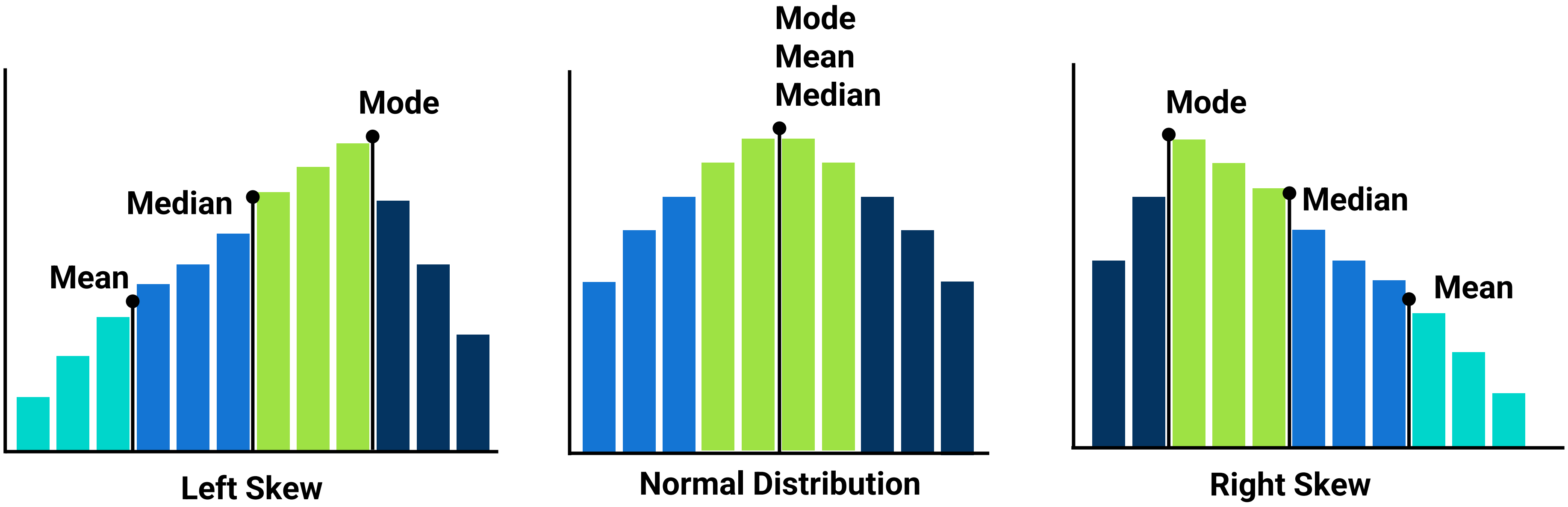
Median is the middle value in a sorted dataset.

Mode

Mode is the most frequently occuring value(s) in a dataset.

The mean, median and mode.

The mean, median and mode.



The **mean** can be used to determine the average value of a portfolio or stock over time.

What are the measures of central tendency used for?

A

Metrics used to describe the center of a data set.

How do you describe the variability of a data set?

Variability of a Data Set

Three summary statistics metrics for describing variability:

01 Variance

02 Standard Deviation

03 Z-Score

What are variance and standard deviation?

Variance and Standard Deviation

Variance and standard deviation describe variability of data.

Variance is the measurement of how far each value is from the mean of the dataset.

Standard deviation is the square root of variance.

Variance

- Used to describe how far values in the data set are from the mean
- Describes how much variation exists in the data
- Variance considers the distance of each value in the data set from the center of the data

The value of the one observation **The mean value of all observations**

Sample variance

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

The number of observations

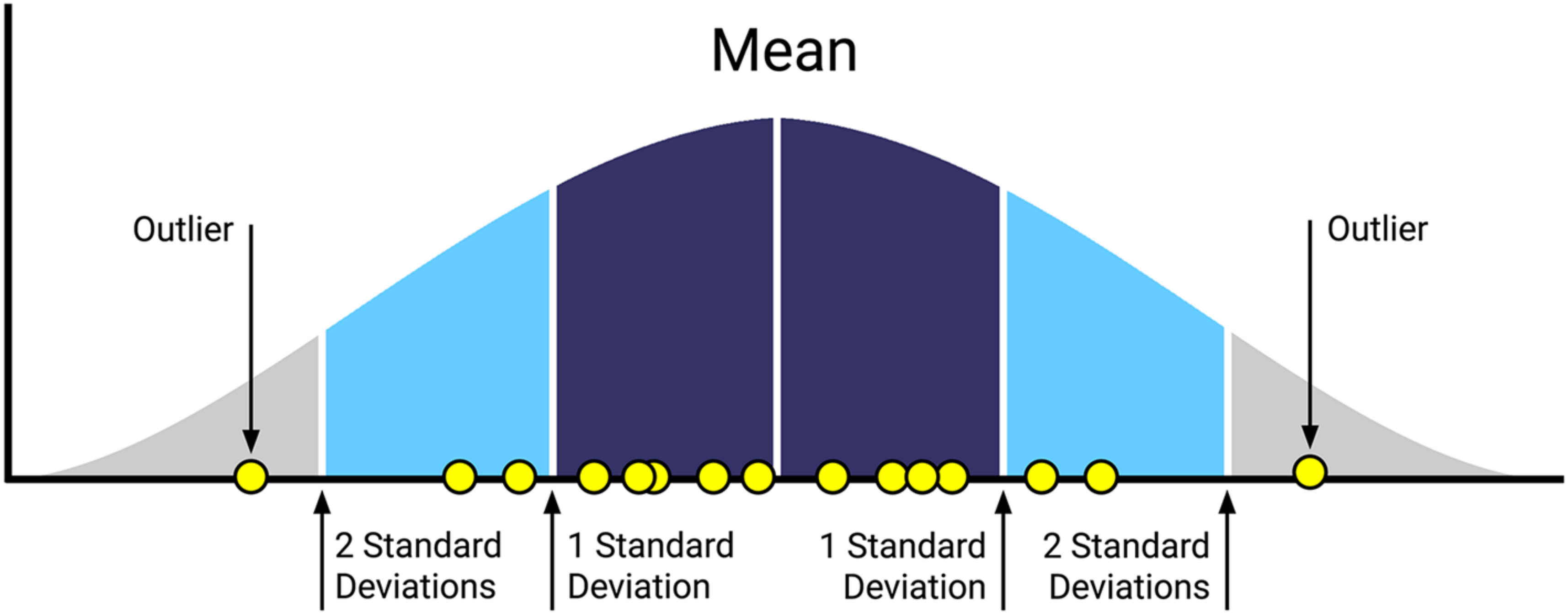
Standard Deviation

- Describes how spread out the data is from the mean
- Calculated from the square root of the variance
- In the same units of measurement as the mean

Standard deviation $\sigma = \sqrt{S^2}$ The variance

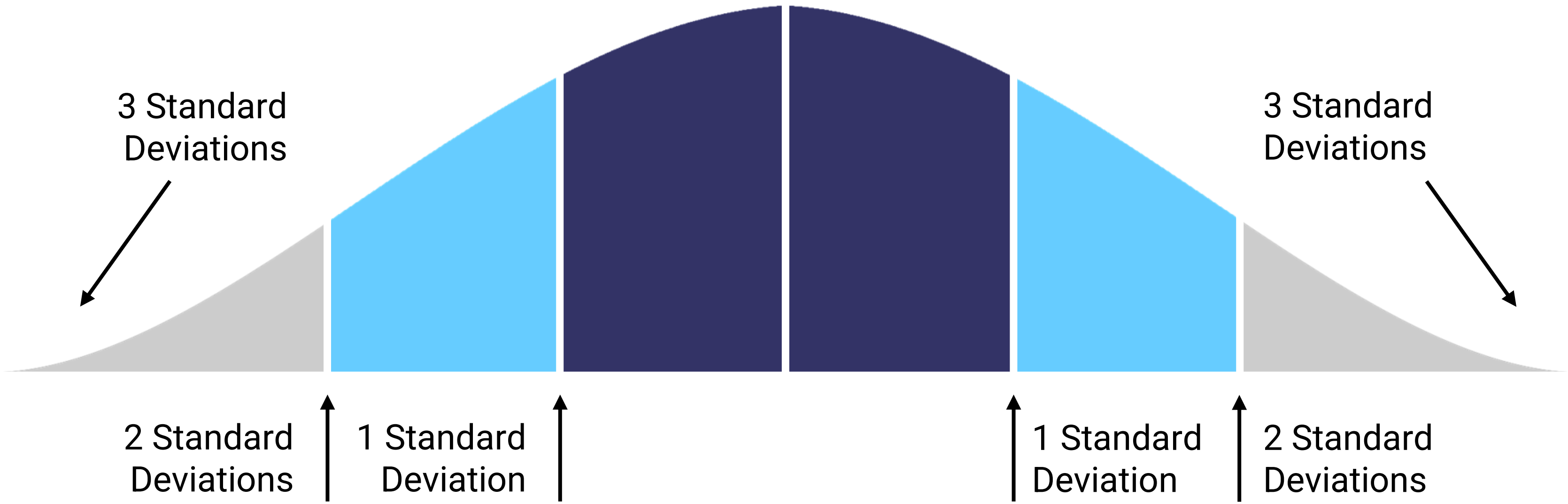
Standard Deviation

Square root of the variance; a measure used to quantify the dispersion of a set of observations.



Standard Deviation

The `std` Pandas function is used to calculate standard deviation for a DataFrame. Standard deviation can be used to determine the risk associated with an investment. Standard deviation is also used to calculate how much returns have been distributed from the average.



Standard Deviation and Risk

Standard deviation identifies precisely how far away a value is from the average.

The greater the standard deviation, the greater the risk (and the potential) for a larger payout.

Z-Score

Z-Score describes a single value’s distance from the mean of the data set
The distance is in terms of standard deviations. Can be positive or negative:

If negative	the value is less than the mean
If positive	the value is greater than the mean.

The smaller the z-score, the closer the value is to the mean

$$z = \frac{\text{A single value } X - \mu \text{ The mean of the dataset}}{\sigma \text{ The standard deviation of the dataset}}$$

Real-World Data

Be careful when describing real-world data:

- Real world data can contain extreme values
- Some summary statistics such as the mean take into account all values of a data set
- Extreme values can skew these statistics!

But how can we summarize real-world data?

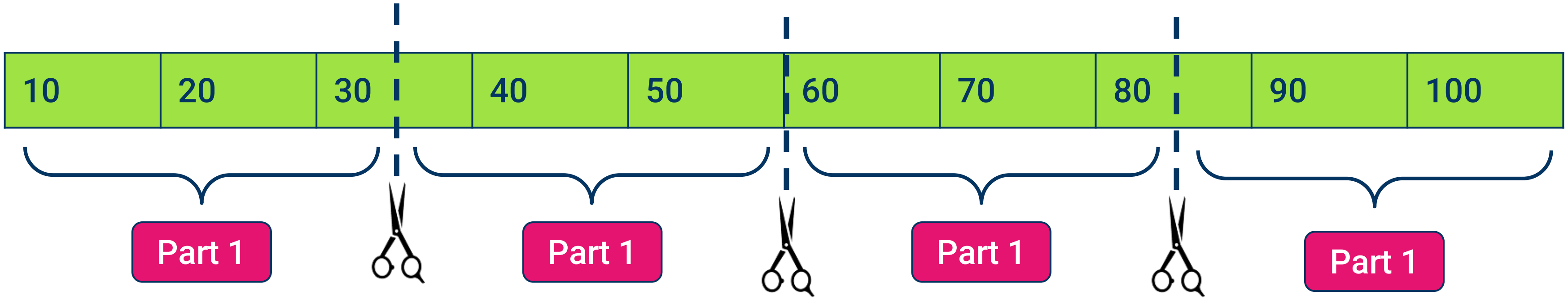
Quantiles: Used to Describe Segments of a Dataset

Quantiles separate a sorted dataset into equally sized fragments.

The two most popular types of quantiles are quartiles and percentiles .

01 **Quartiles** divide the dataset into four equally sized parts.

02 **Percentiles** divide the dataset into 100 equally sized parts.



Quantiles, Quartiles, and Outliers

Quantiles, quartiles, and outliers describe a dataset.

Quantiles

Quantiles divide data into well-defined regions based on a sorted dataset.

Quartiles

Quartiles are a specific type of quantile where a sorted dataset is split into four equal parts.

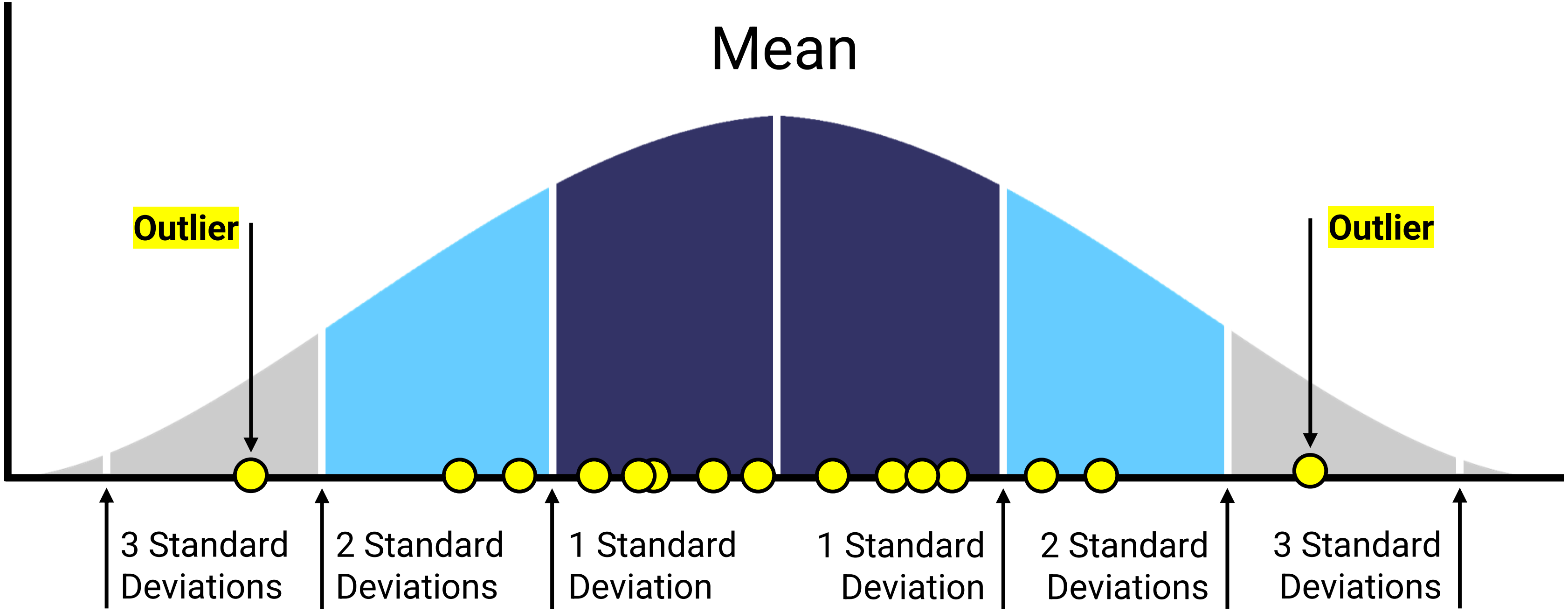
Q1	25% of the data
Q2	50% of the data
Q3	75% of the data

Quartiles

Outliers are extreme values in a dataset that can skew calculations and results.

Outliers

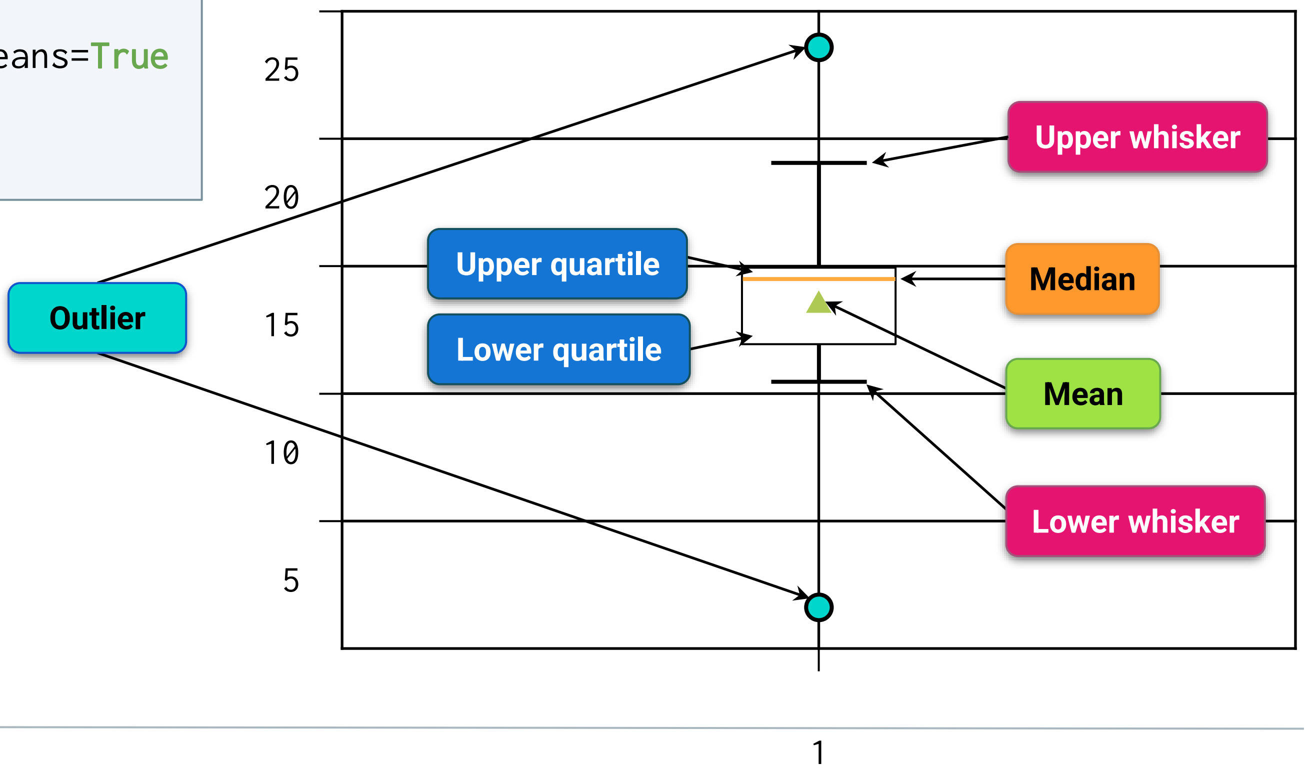
Suspicious values are called potential outliers. An outlier is a data point that differs from the rest of a data set. Outliers can inaccurately skew a data set.



How to Identify Potential Outliers: Qualitatively

Use box-and-whisker plots to visually identify potential outlier data points.

```
# Create box plot
plt.boxplot(arr, showmeans=True)
plt.grid()
plt.show()
```



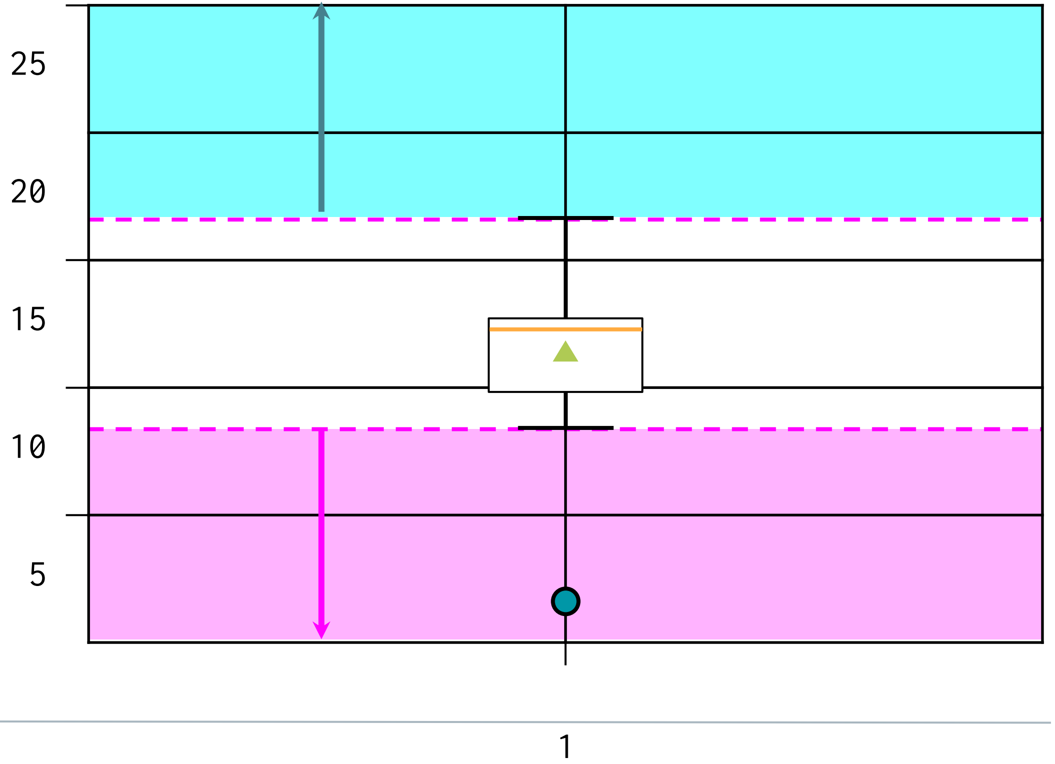
How to Identify Potential Outliers: Quantitatively

Determine the outlier boundaries in a dataset by using the **1.5 × IQR rule**.

The IQR is the range between the first and the third quartile.

Anything less than, or below, Quartile 1 – (1.5 × IQR) might be an outlier.

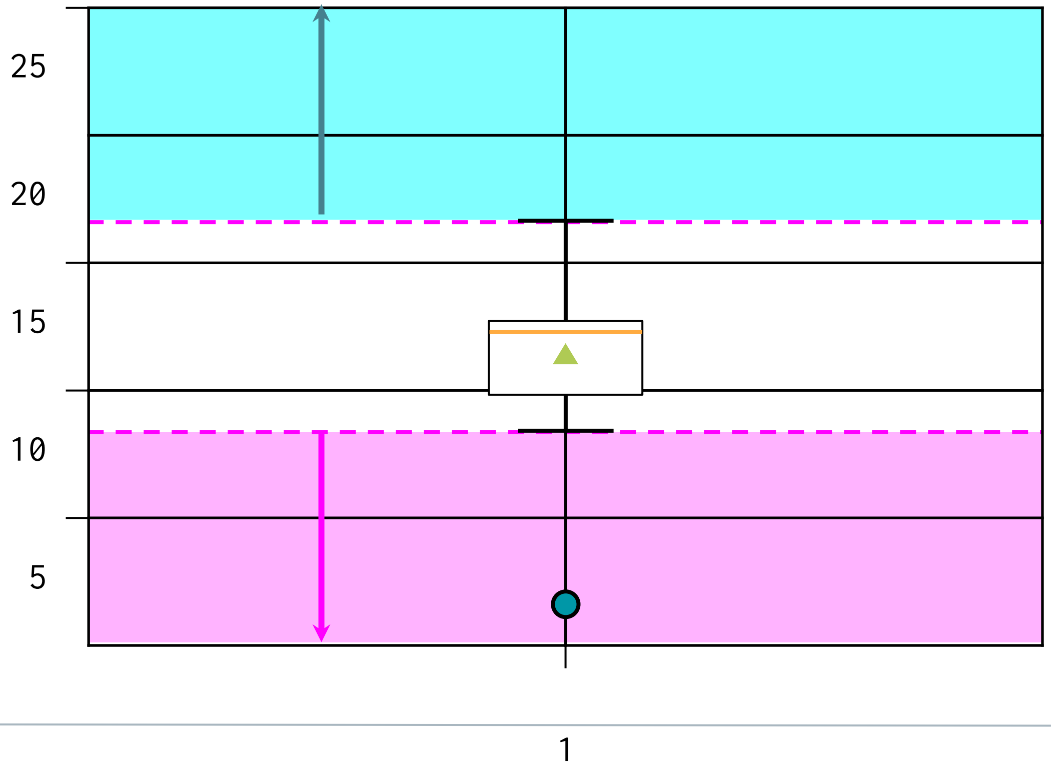
Anything greater than, or above, Quartile 3 + (1.5 × IQR) might be an outlier.



How to Identify Potential Outliers in Python: Quantitatively

Use Pandas' `series.quantile` function to calculate the quantile.

Calculate the outlier boundaries.



When new data comes along,
you must plot it!

Why Plot Data?

01 To determine if the data is normally distributed.

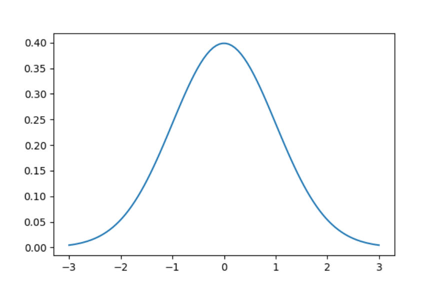
02 To determine if the data is multimodal.

03 To characterize clusters in the dataset.

What Is Normally Distributed Data?

01

The distribution of data follows a bell-curve shape.



02

We can quantitatively test if a dataset is normal using SciPy.

```
stats.normaltest()
```

03

Some statistical tests assume normally distributed data.

Sample and Population Datasets

Let’s think about
the following scenario…

Predicting the City Election

Weeks before Election Day, a local newspaper wants to predict the winner of the mayoral election. The newspaper will poll voters for their intended candidate. Consider the following:

- It would be prohibitively expensive to poll all voters.
- It is logistically impossible to know who will actually go out to vote on Election Day.
- Therefore, the newspaper must predict the outcome of the election using data from a subset of the population.
- This calls for the use of a sample dataset in place of a population dataset.

Population Dataset vs. Sample Dataset

Population Dataset

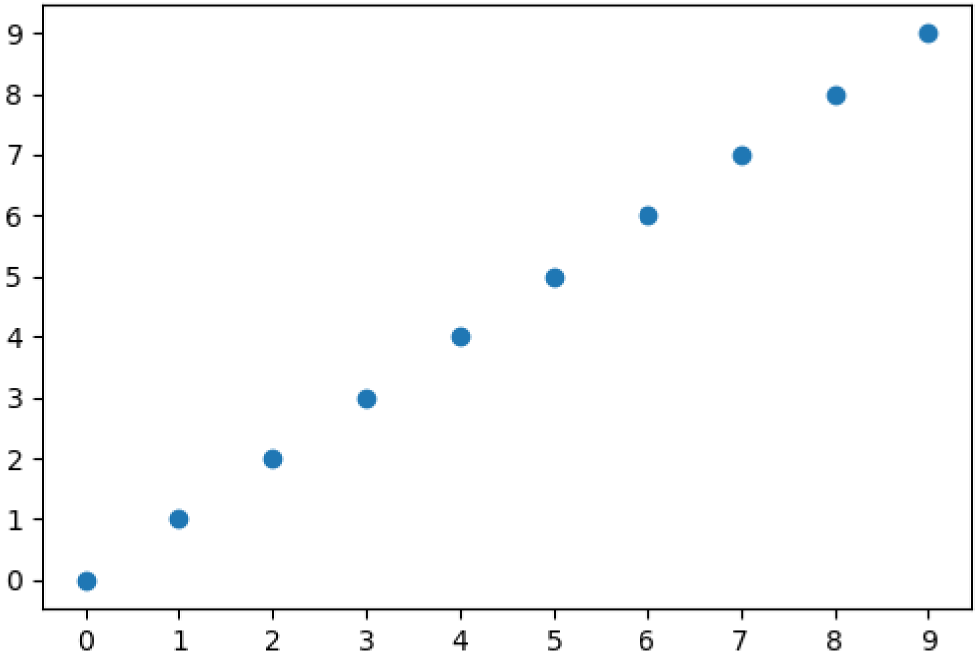
- Dataset containing all possible elements of an experiment or study.
- In statistics, “population” does not mean “people.”
- Any complete set of data is a population dataset.

Sample Dataset

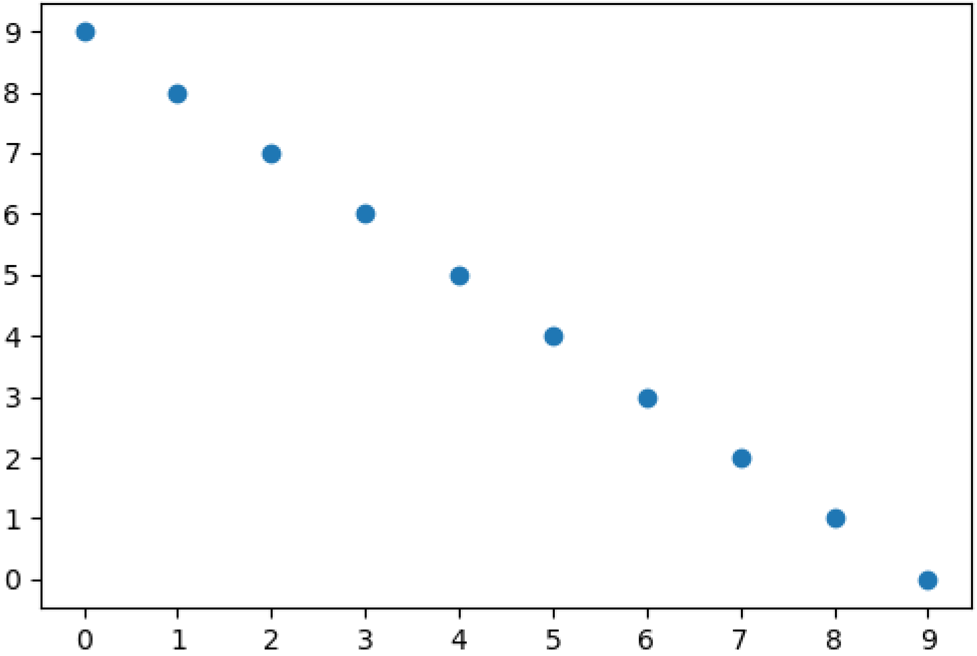
- A subset of population data.
- A sample dataset can be selected randomly from the population or selected with bias.

Correlation describes the question,
“Is there a relationship between A and B?”

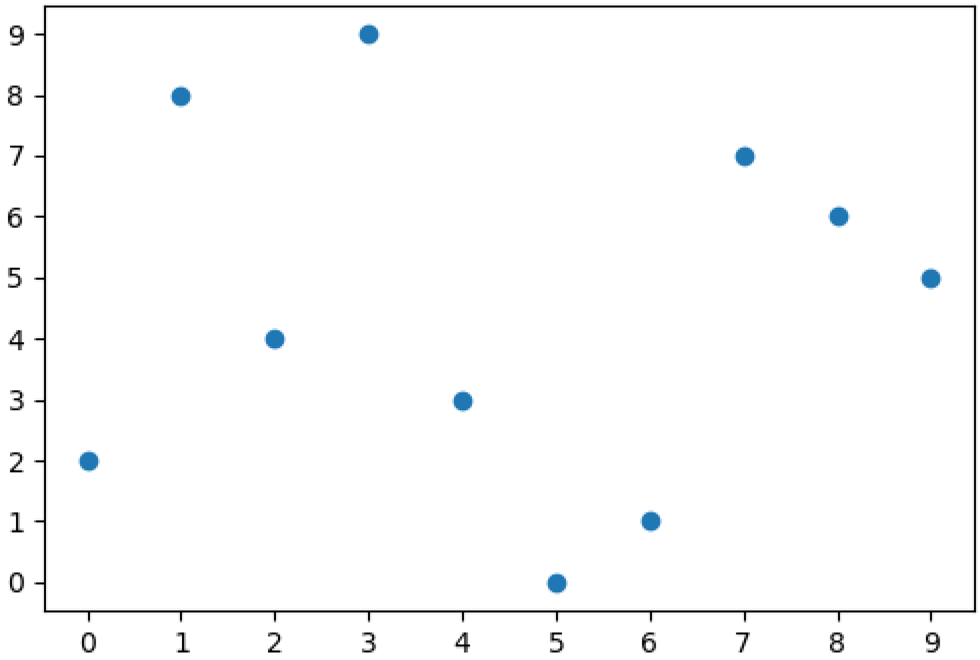
Positive Correlation



Negative Correlation



No Correlation

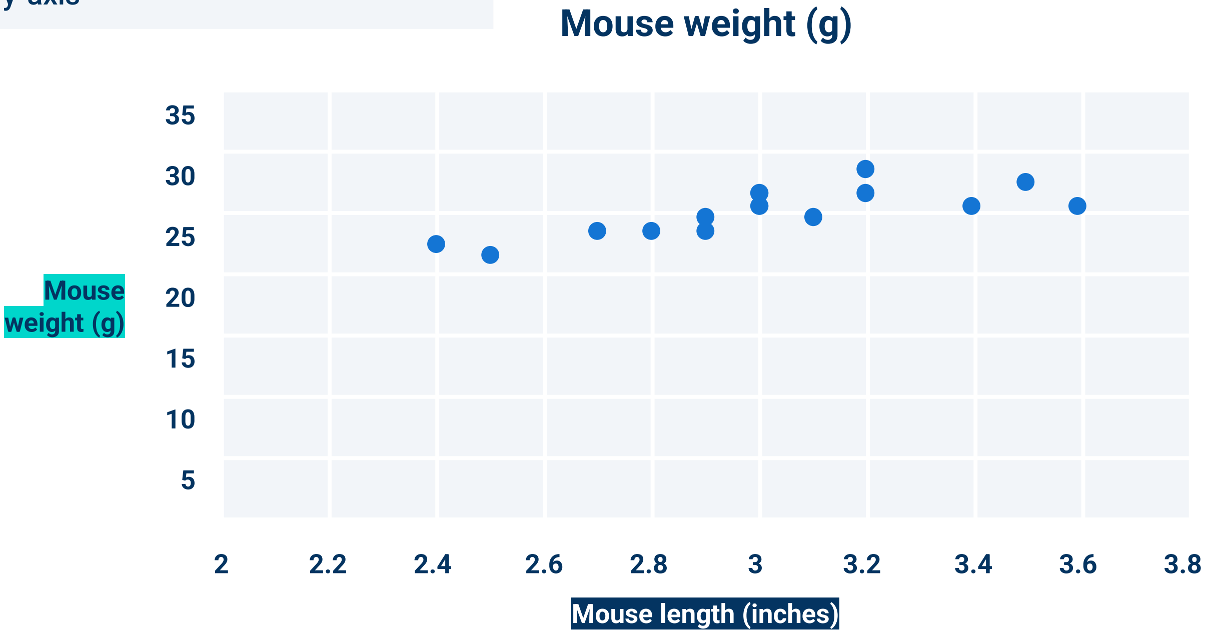


Scatter plots are a powerful visualization tool!

Visualizes the comparison between two variables:

One variable	is located on the x-axis
Another variable	is plotted on the y-axis

- Each data point represents a pair of measurements
- Measurements on a scatter plot are independent
- Scatter plots can help to identify positive or negative relationships between two variables
- Adding a trend line to a scatterplot can visualize this relationship even easier!



What is the equation of a line?

The equation of a line is:

$$y = mx + b$$

Dependent variable

Slope

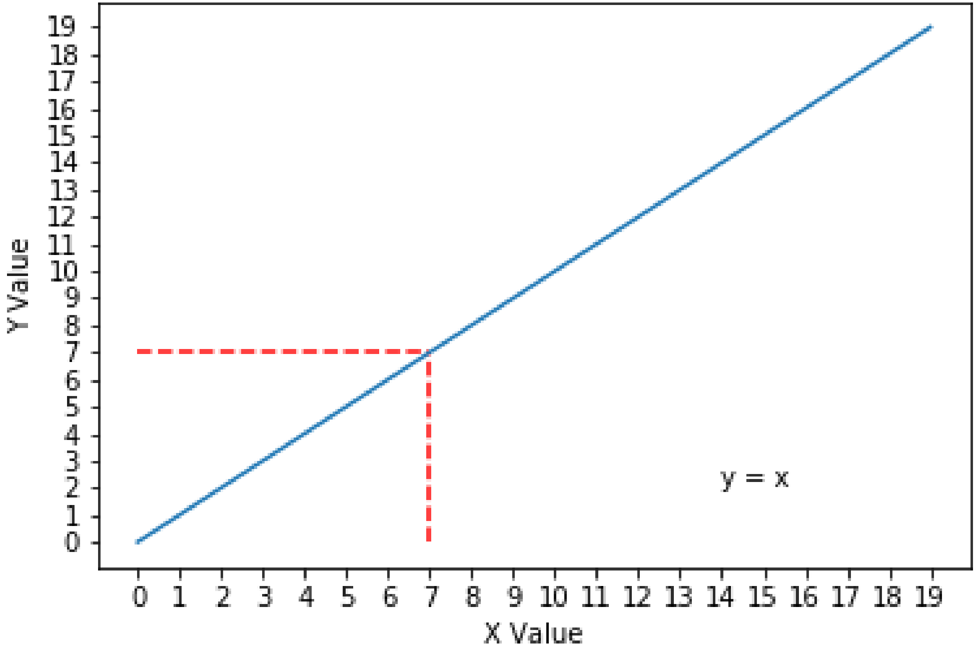
Independent variable

y-intercept

The Equation of a Line Determines *y* Values Given *x*

In this example:

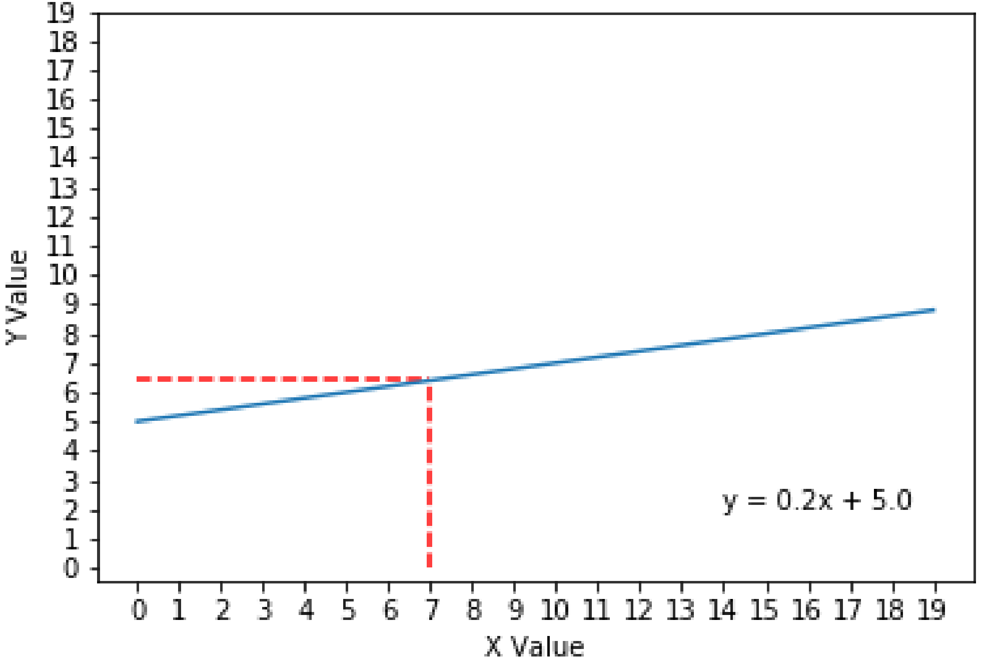
- Slope = 1
- y intercept = 0
- Whatever *x* is, the value of *y* is the same.



The Equation of a Line Determines *y* Values Given *x*

In this example:

- Slope = 0.2
- y intercept = +5
- If x = 7, then y = 6.4



The End