AI Boot Camp

Visualization and Statistics

Module 7 Day 3

Class Objectives

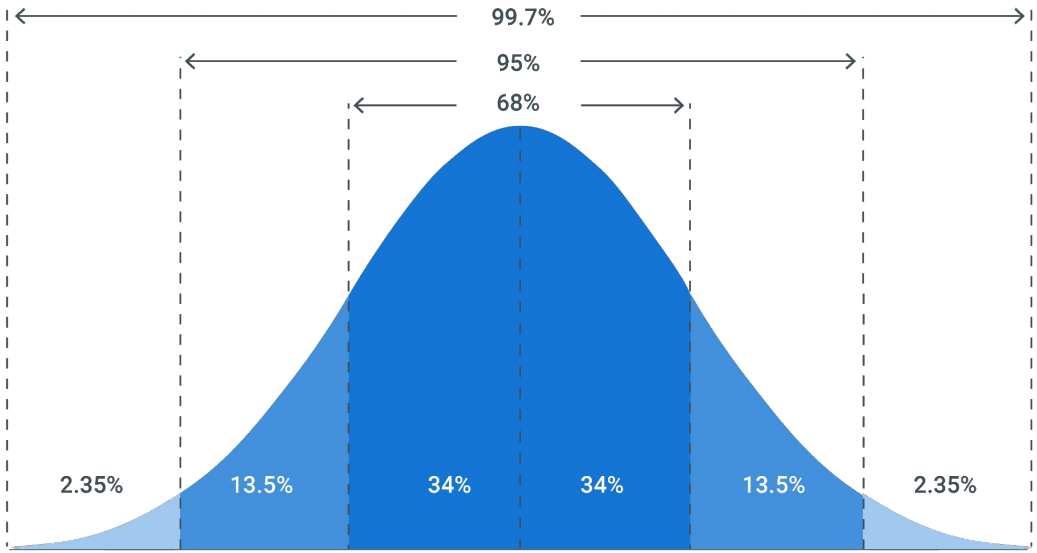
By the end of class, you will be able to:

1. Recall how to calculate and interpret summary statistics by using Python.
2. Identify potential outliers in a dataset.
3. Differentiate between a sample and a population in regard to a dataset.
4. Define and quantify correlation between two factors.
5. Make predictions about data by using linear regression.

Instructor **Demonstration**

Stats Review

The Normal Distribution: 68-95-99.7 Rule



The **68–95–99.7** rule states that roughly 68% of all values in normally distributed data fall within one standard deviation of the mean (in either direction). Additionally, 95% of the values fall within two standard deviations, and 99.7% of the values fall within three standard deviations.

Instructor **Demonstration**

Box Plots

Quantiles, Quartiles, and Outliers

Cybersecurity professionals also need to clearly communicate technical topics for other reasons, including:

01

Quantiles divide our data into well-defined regions based on their order in a ranked list. The 2 most commonly used quantiles are **quartiles** and **percentiles.**

02

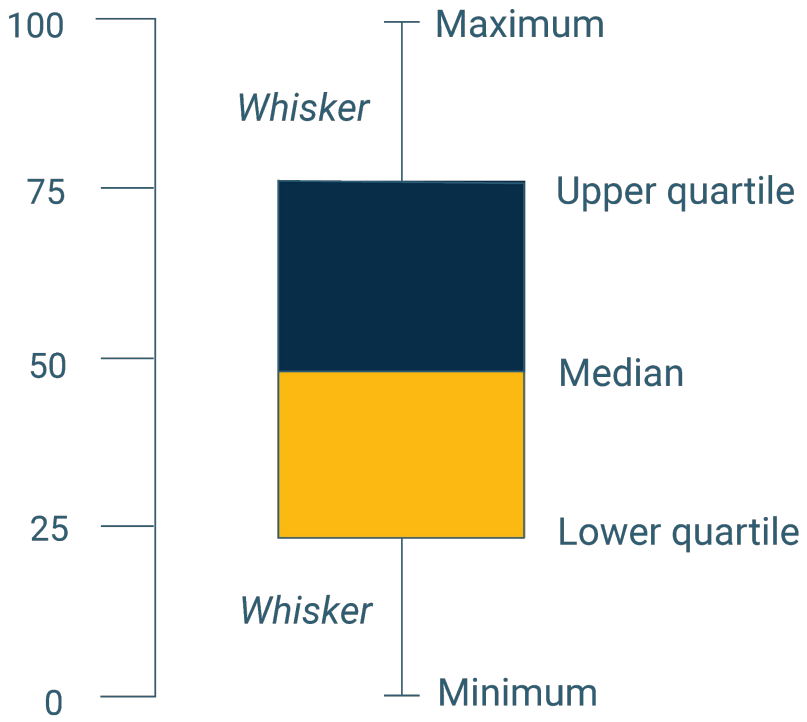
Quartiles divide the sorted data into four equal-sized groups, and the median is known as the second quartile.

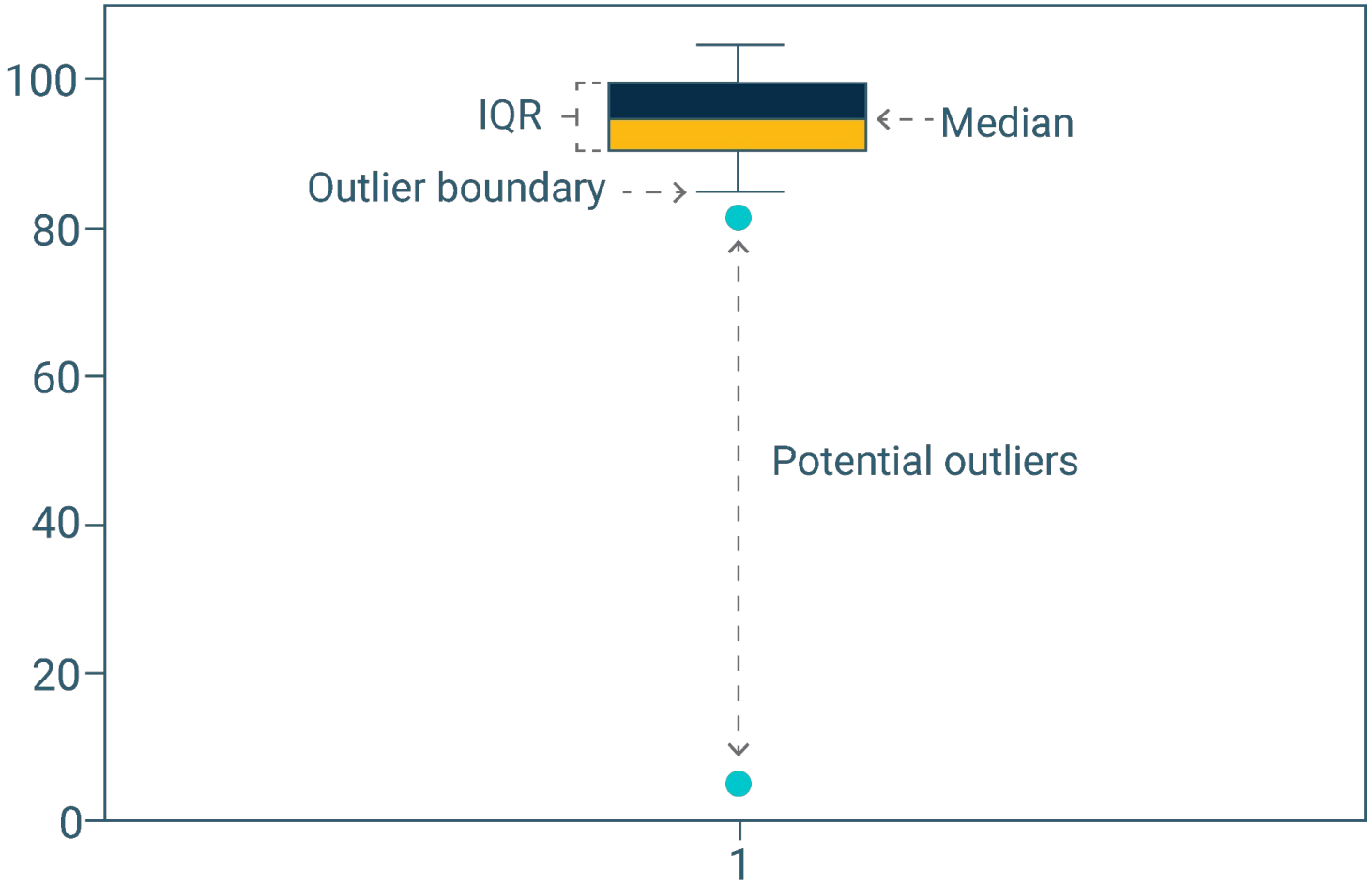
03

An outlier is an extreme value that can skew a dataset. An **outlier** is typically identified as a value that is 1.5 **IQR (interquartile range)** beyond the first and third quartiles.

The Box Plot

A box plot provides a visual summary of the 25th, 50th, and 75th percentile using a box. Two lines extend on either side of the box to mark the upper and lower outlier boundaries. These two lines make up the “whiskers” portion of the box and whisker plots.




100
80
60
40
20
0
1
IQR
Median
Outlier boundary
Potential outliers

Activity:

Temperature Outliers

In this activity, you will search for outliers in a dataset that contains National Oceanic and Atmospheric Administration temperature measurements taken at the Los Angeles International (LAX) airport.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Instructor **Demonstration**

Sample, Population, and SEM

Activity:

SEM and Error Bars

In this activity, you will work with a partner to characterize sample data from a California housing dataset.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

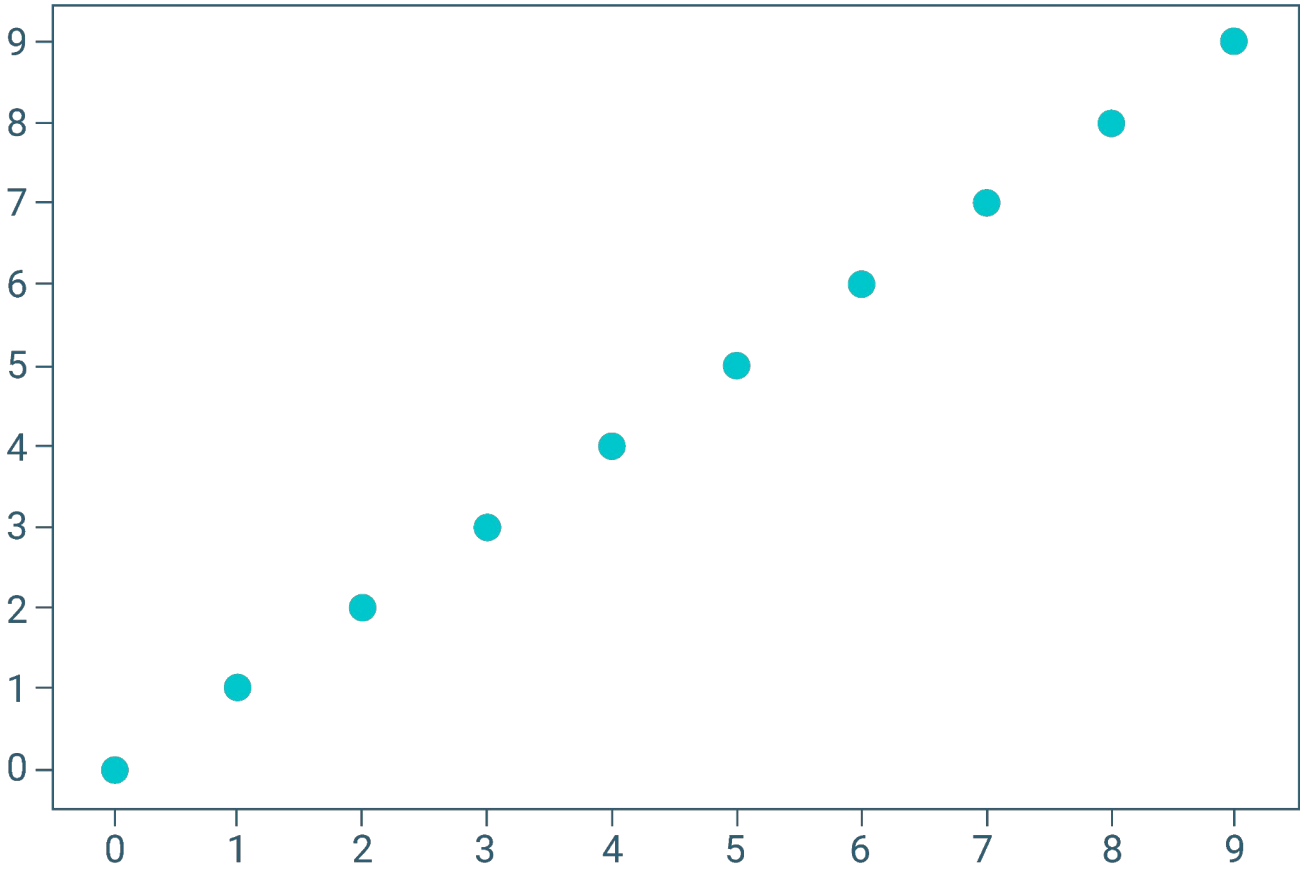
Break

15 mins

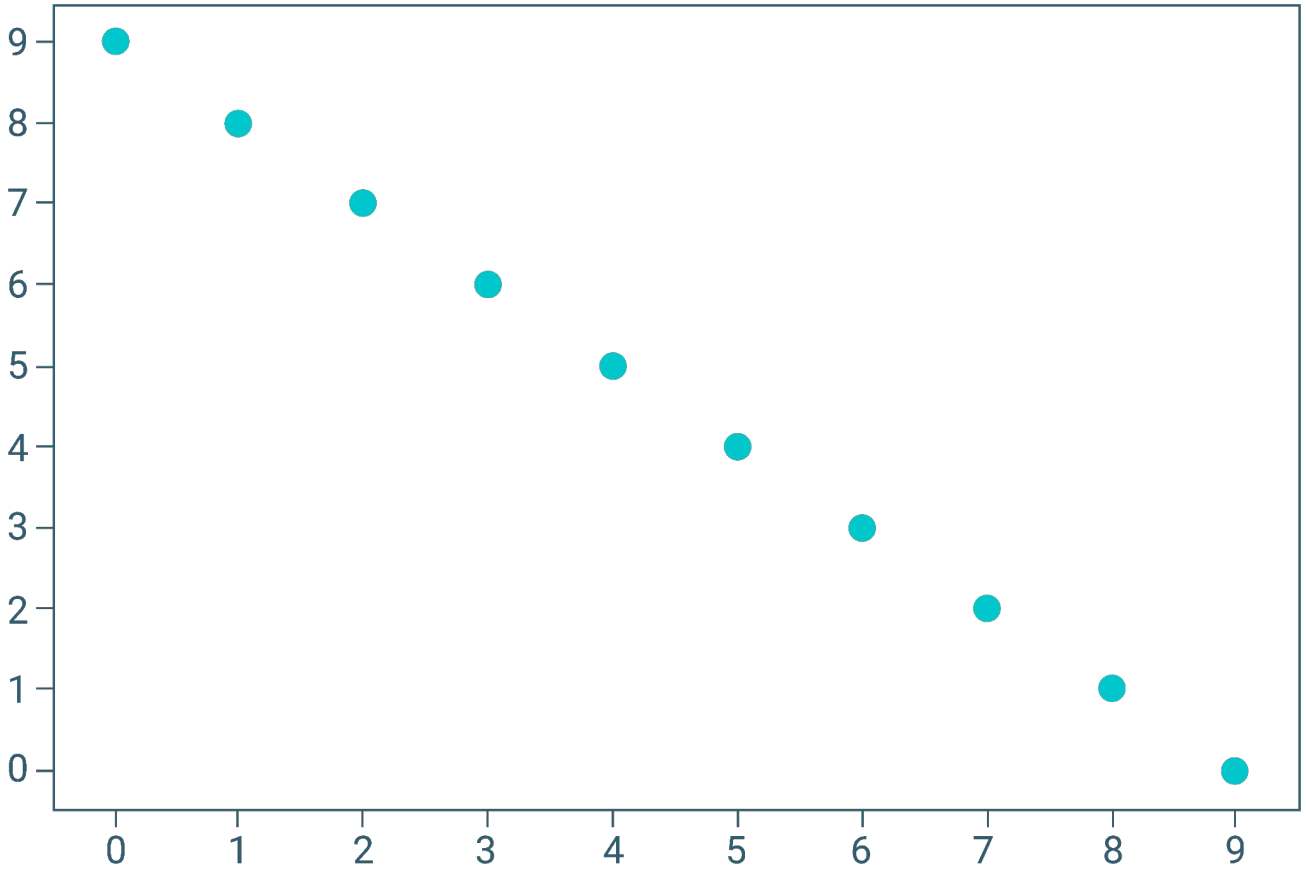
Instructor **Demonstration**

Correlation Conundrum

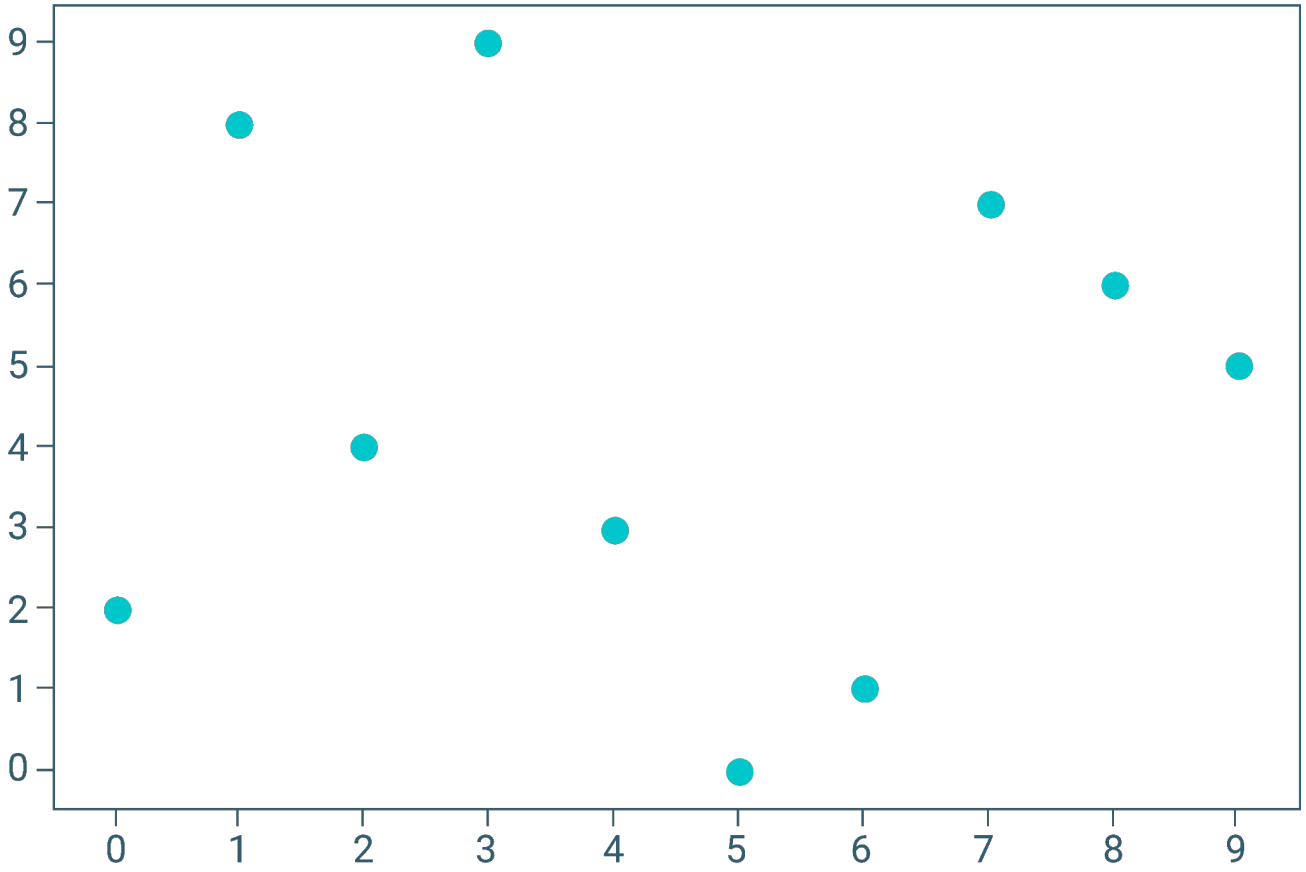
Positive Correlation



Negative Correlation



No Correlation



Calculating Correlations

1. In statistics, we can calculate the degree of correlation using the **Pearson correlation coefficient**.

2. The **Pearson correlation coefficient** is a quantitative measure that describes the simultaneous movement (variability) of two variables.

Coefficient Range	Meaning
$-1 \leq r < 0$	This coefficient indicates a **negative correlation**.
$r = 0$	This coefficient means that there is **no correlation**.
$0 < r \leq 1$	This coefficient indicates a **positive correlation**.

The tidy positive and negative correlation figures we saw earlier were examples of perfect correlation.

Activity:

Correlation Conquerors

In this activity, you will have an opportunity to use SciPy to compare variables across Scikit-learn's wine recognition dataset.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Instructor **Demonstration**

Fits and Regression

Equation of a Line

The equation of a line is _y_ = _mx_ + _b_

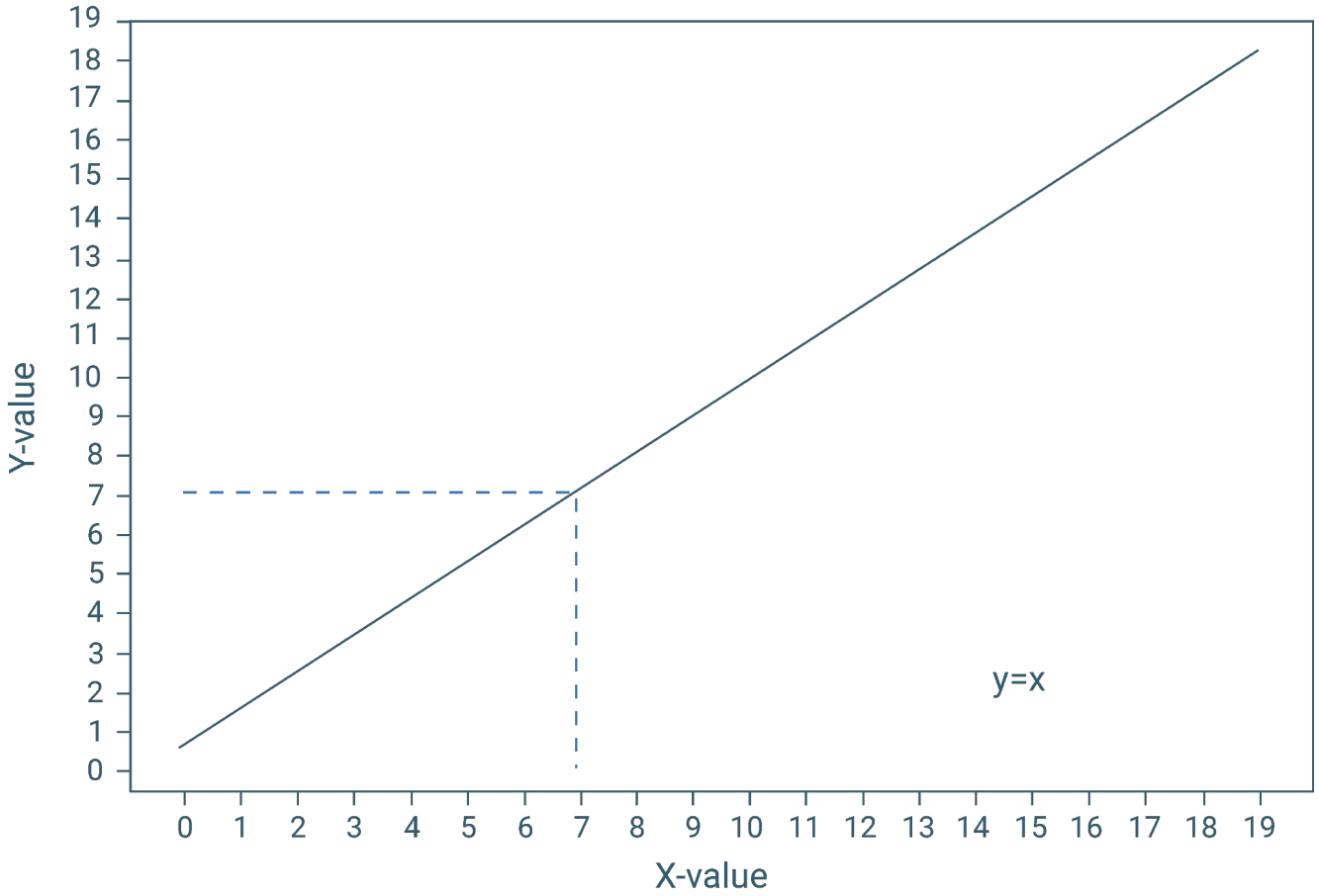
1. The equation of a line defines the relationship between x-values and y-values.

2. When it comes to variables in the equation, we refer to the _x_ in the equation as the **independent variable**, and the _y_ as the **dependent variable**.

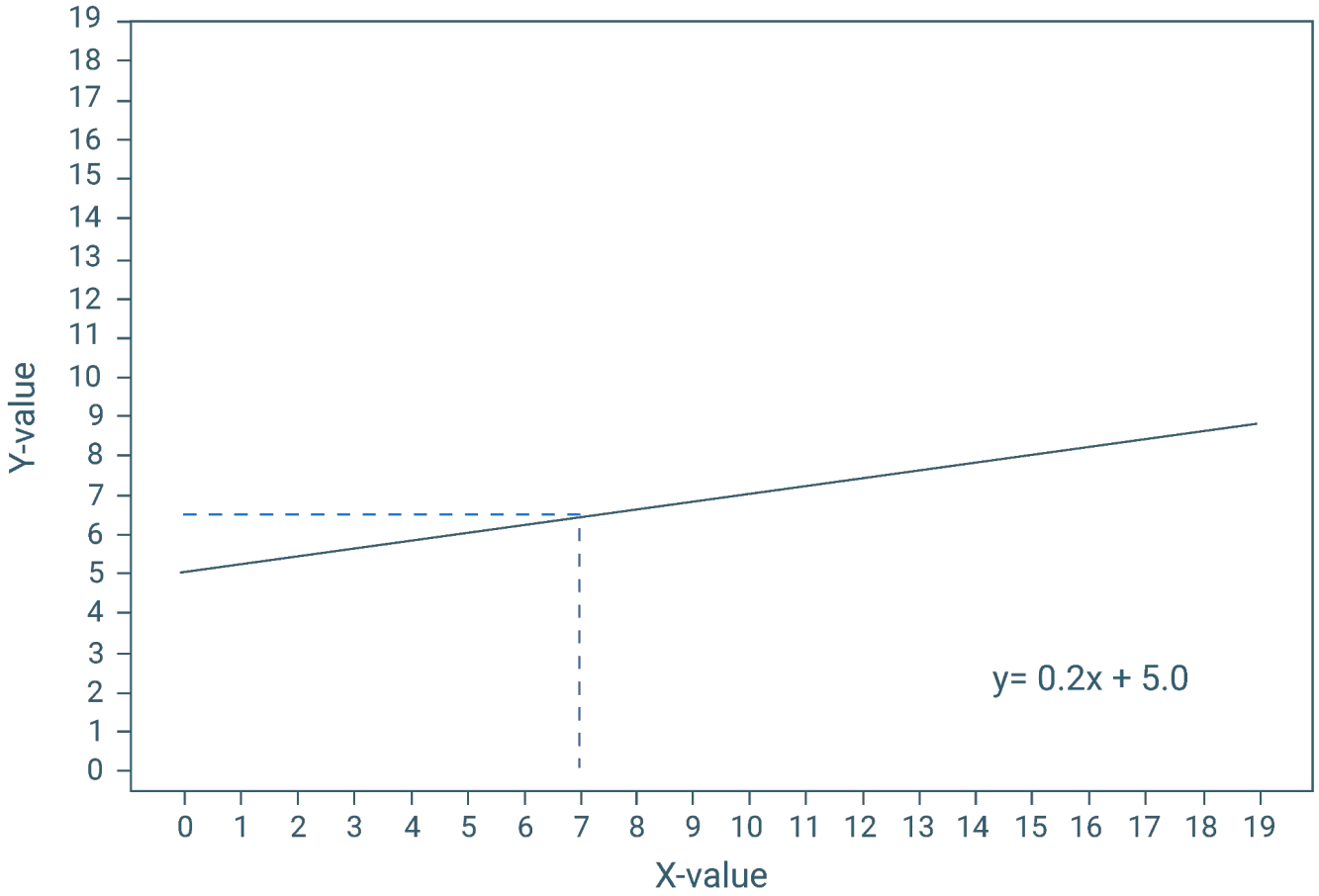
3. The **slope** of a line is denoted as _m_ in the equation, and the **_y_ intercept** is denoted as _b_.

4. Knowing the slope and _y_ intercept of a line, we can determine any value of _y_ given the value for _x_. This is why we say _y_ is dependent on _x_.

Y = x


Y-value
X-value
y=x
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Y = 0.2x + 5


Y-value
X-value
y= 0.2x + 5.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Linear Regression

1. **Linear regression** is used in data science. In particular, we use it in machine learning to model and predict the relationship between two variables.

2. Linear regression is a powerful tool: it provides us with a way to predict house prices, stock market movements, and the weather based on other data.

Activity:

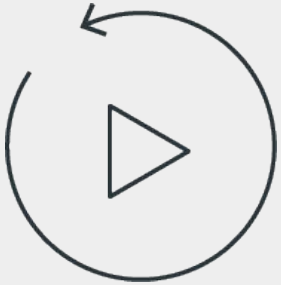
Fits and Regression

In this activity, you will have an opportunity to use SciPy to compare variables across Scikit-learn's wine recognition dataset.

Suggested Time:
5 Minutes

Time’s up!

Let’s review



Let’s **recap**

Recap

In today's lesson, you learnt how to:

1. Calculate and interpret summary statistics.
2. Identify potential outliers in a dataset.
3. Differentiate between a sample and a population in regard to a dataset.
4. Define and quantify correlation between two factors.
5. Make predictions about data by using linear regression.

Questions?

The End