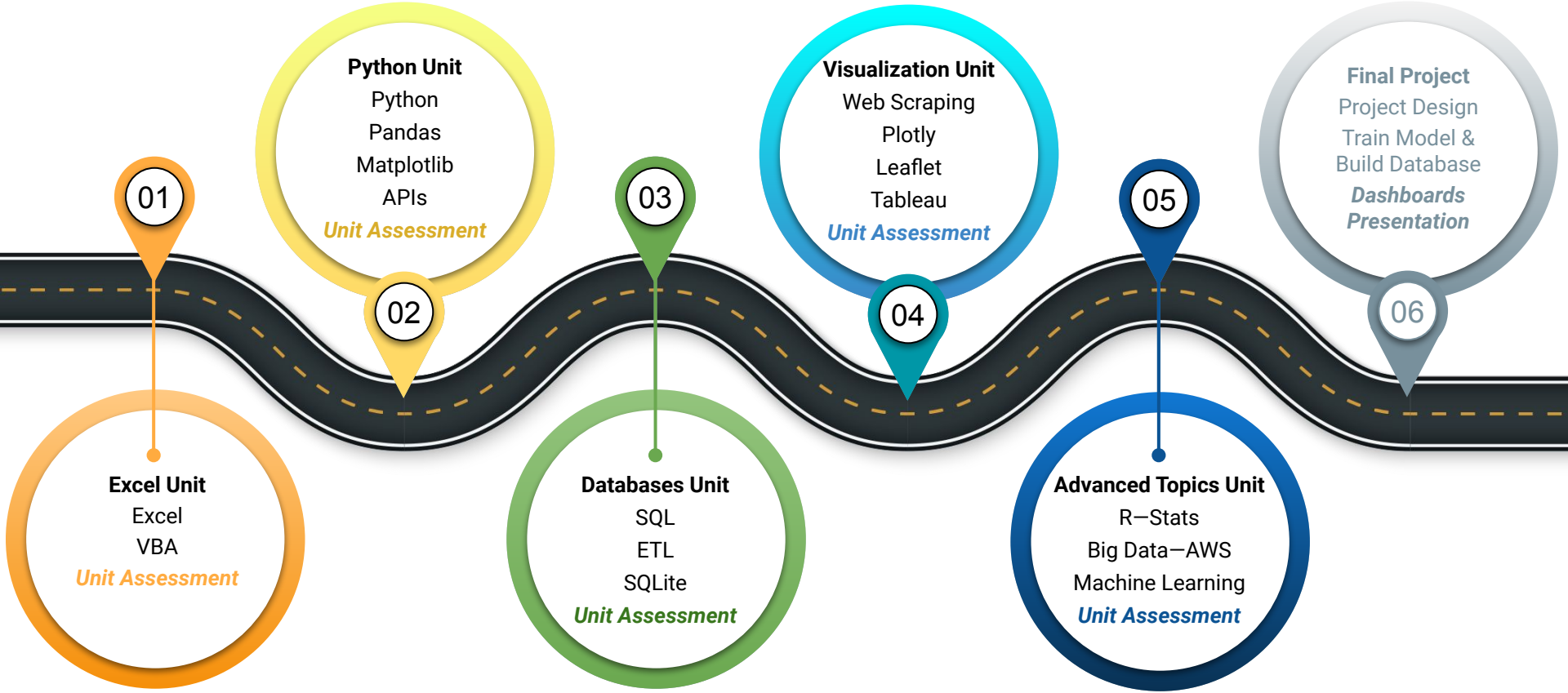
Introduction to Unsupervised Machine Learning

Data Boot Camp

Lesson 18.1

The Big Picture



Quick Tip for Success:

Take some time and review any past-due assignments you may be missing! The end of the course is coming up soon.

Module 18

This Week:
Unsupervised Machine Learning

This Week: Unsupervised Machine Learning

By the end of this week, you'll know how to:

- Describe the differences between supervised and unsupervised learning, including real-world examples of each
- Preprocess data for unsupervised learning
- Cluster data using the K-means algorithm
- Determine the best amount of centroids for K-means using the elbow curve
- Use principal component analysis (PCA) to limit features and speed up the model

This Week's Challenge

Using the skills learned throughout the week, you will analyze cryptocurrency data, reduce data dimensions with principal component analysis, cluster with K-means, and visualize the results.

Career Connection

How will you use this module's content in your career?

Module 18

How to Succeed This Week

Quick Tip for Success:

Today, we're continuing to build an understanding of the basics of machine learning. There are still many skills to learn!

Module 18

Today’s Agenda

Today's Agenda

By completing today's activities, you'll learn the following skills:

01 Preparing and processing data for unsupervised learning

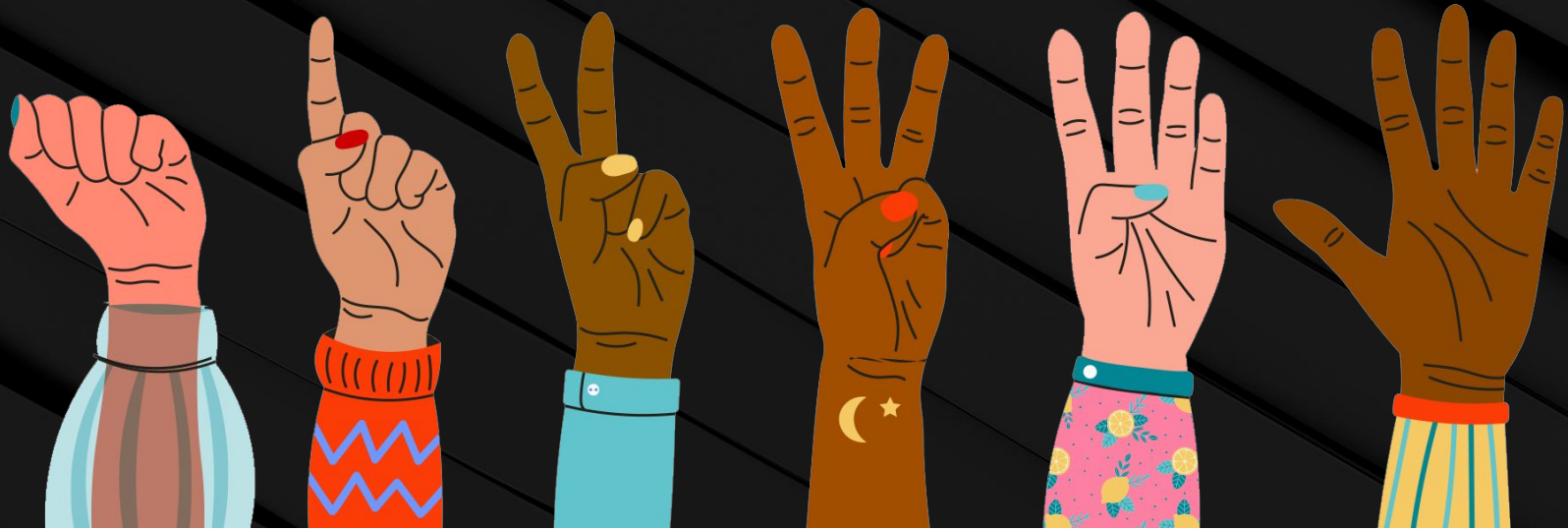
02 Applying K-means clustering

03 Using the elbow curve to determine optimal cluster number

Make sure you've downloaded any relevant class files!

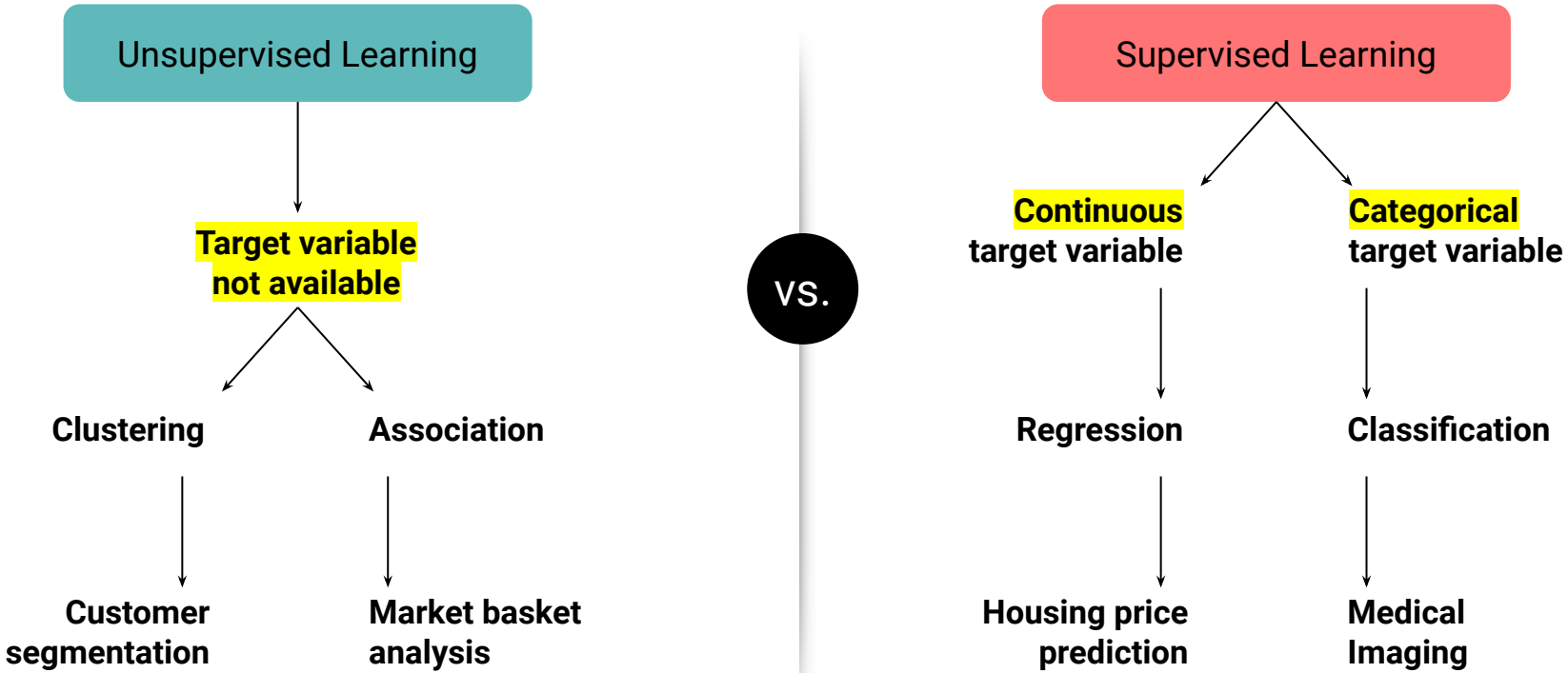
FIST TO FIVE:

How comfortable do you feel with this topic?



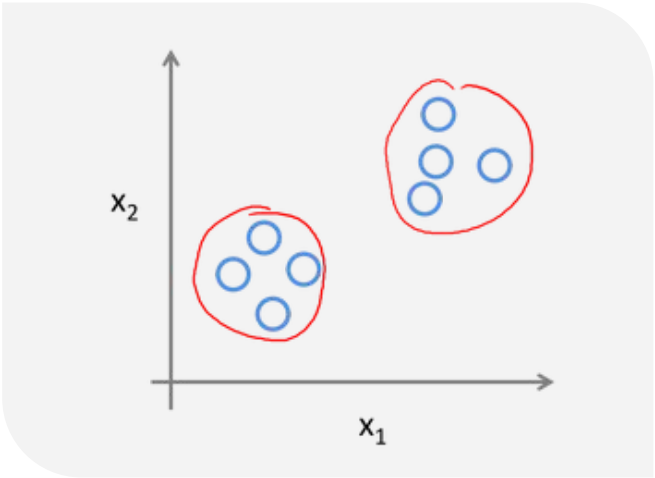
Intro to Unsupervised Machine Learning

Unsupervised Learning vs. Supervised Learning



Supervised vs. Unsupervised Learning

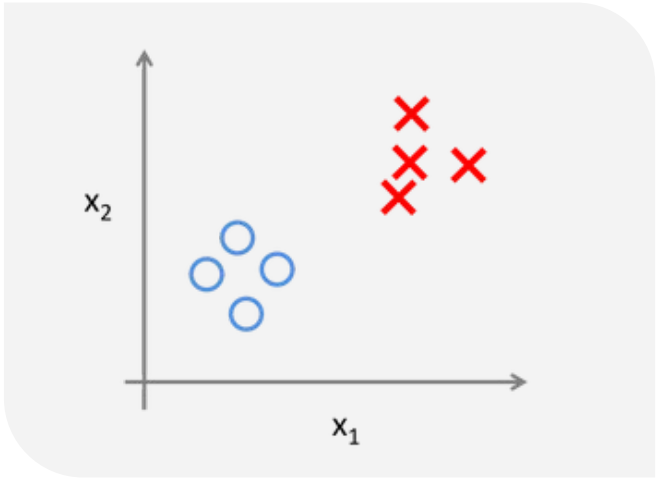
Supervised Learning



- Input data is labeled.
- Uses training datasets.
- **Goal:** Predict a class or value.

vs.

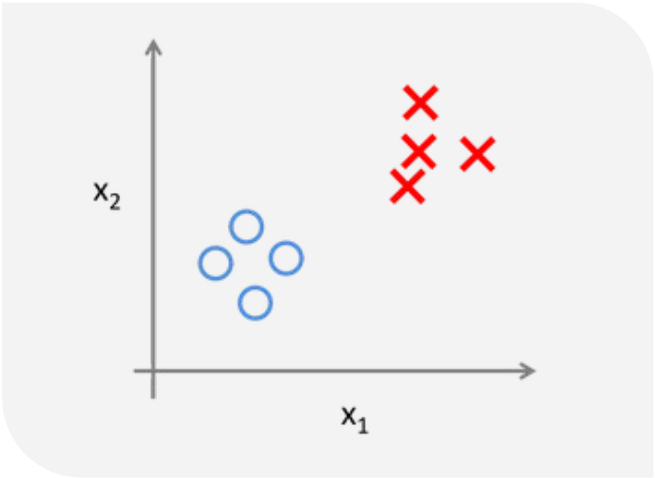
Unsupervised Learning



- Input data is unlabeled.
- Uses just input datasets.
- **Goal:** Determine patterns or grouping data.

Supervised vs. Unsupervised Learning

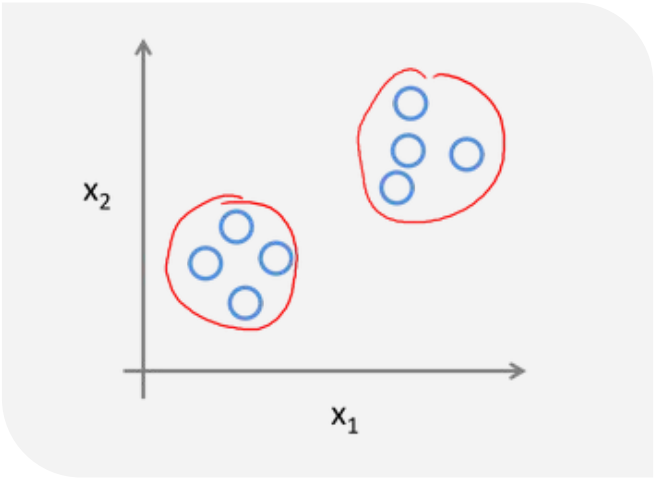
Supervised Learning



- Input data is labeled.
- Uses training datasets.
- **Goal:** Predict a class or value.

vs.

Unsupervised Learning



- Input data is unlabeled.
- Uses just input datasets.
- **Goal:** Determine patterns or grouping data.

Supervised vs. Unsupervised Learning

Supervised Learning

- Is this person satisfied?
- How much is this customer going to spend next month?

Upbeat Millennial
purchases Matcha tea drinks most often

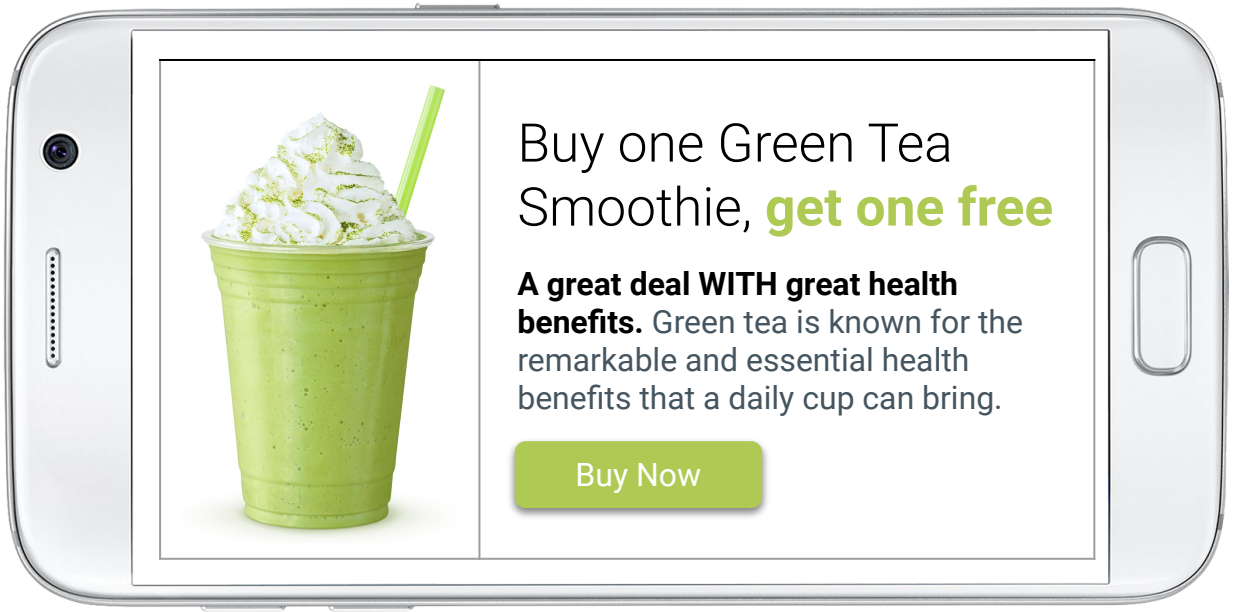
Behaviors
Average purchase per month: $64.32

Motivations
The most important factor in decision to purchase is knowing that the tea is sustainably sourced.

Supervised vs. Unsupervised Learning

Unsupervised Learning

How can I create an offer customized to customers?



Unsupervised Learning

Common uses of unsupervised learning:

- Grouping unlabeled data into distinct clusters (clustering).
- Detecting unusual data points in a dataset—i.e., anomaly detection.
- Reducing large datasets into smaller datasets while preserving most of the useful information (dimensionality reduction).

How is clustering used by retail businesses?

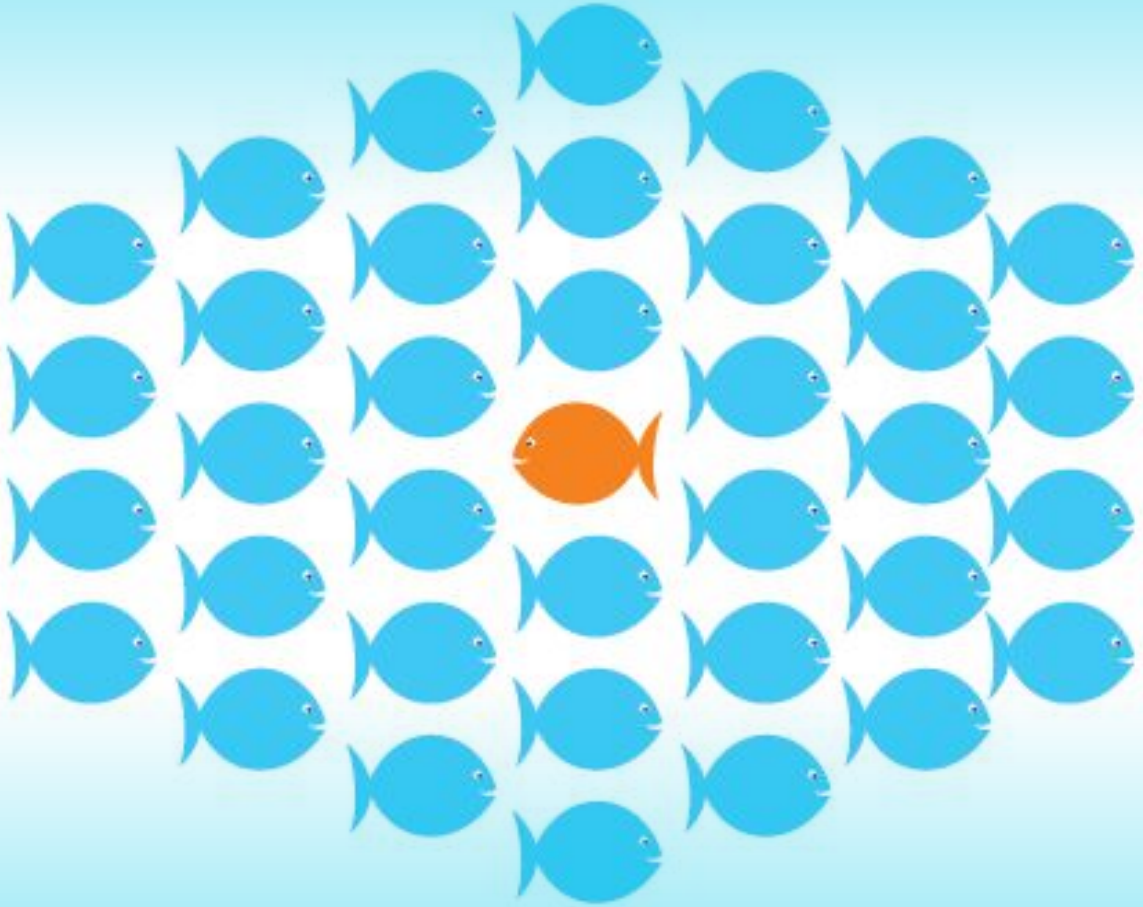
Clustering

Clustering can be used to group customers by shopping habits to create customized offers via email or mobile apps.

How is anomaly detection used by credit card companies?

Anomaly Detection

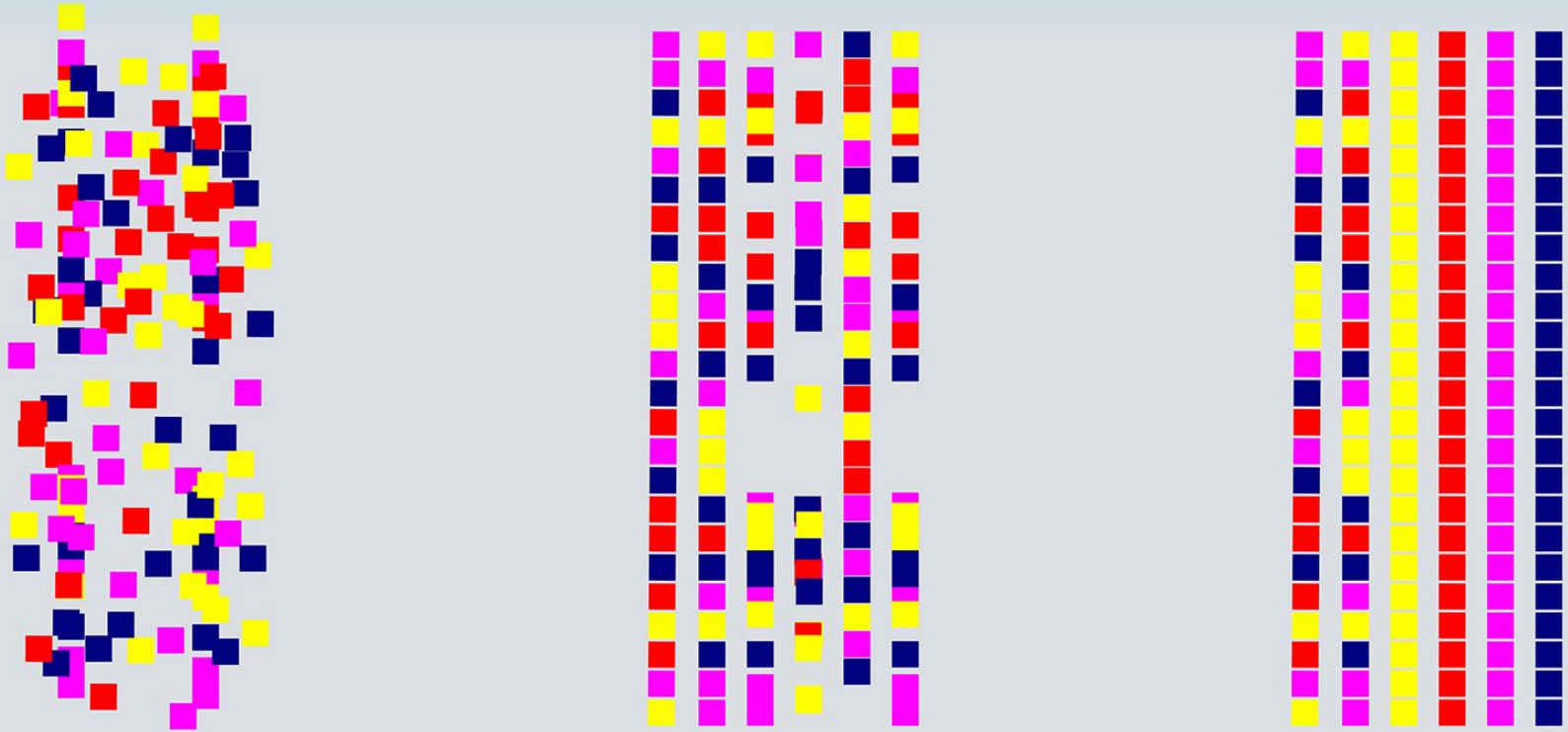
Anomaly detection can be used to identify anomalous and potentially fraudulent credit card transactions (by grouping transactions into normal vs. abnormal).



How is dimensionality reduction used in organizations?

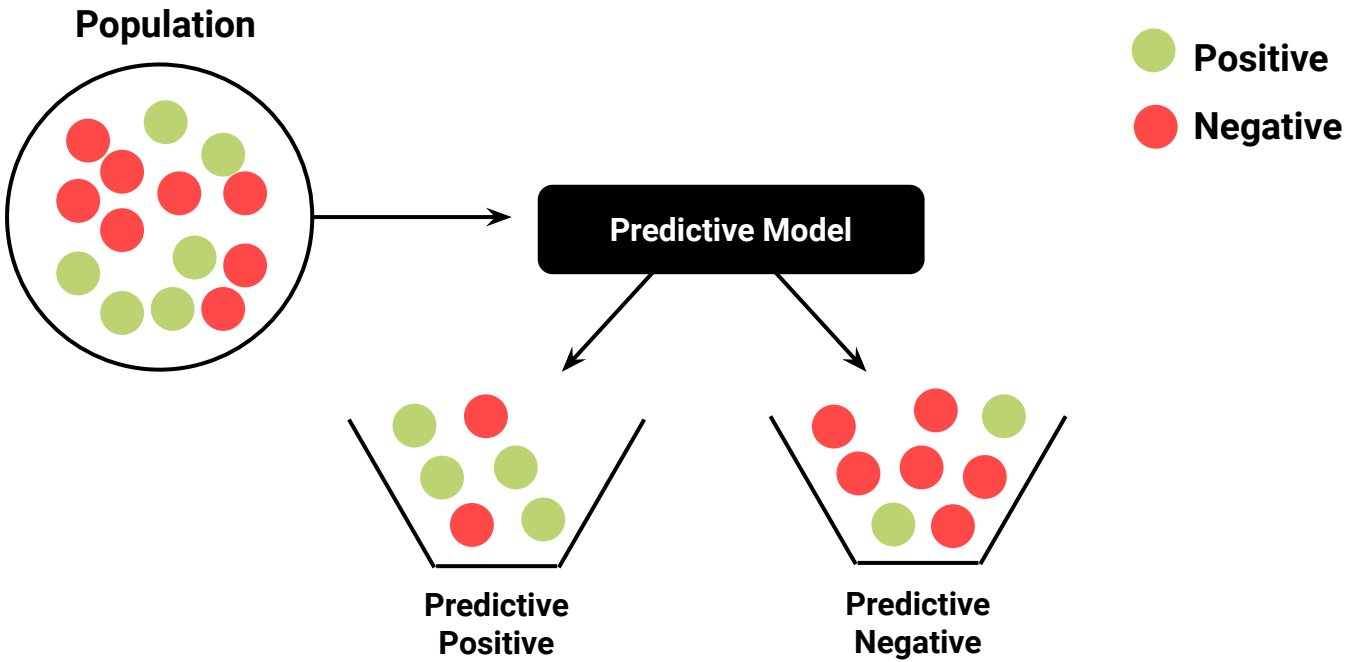
Dimensionality Reduction

It can speed up machine learning by reducing the size of large datasets.



Supervised Learning

Supervised learning is very helpful at predicting the future based on labeled historical data. **However, labeled data is not always available.**



Unsupervised learning allows us to cluster data to find hidden or unknown patterns in data.

Customer Segmentation

Customer segmentation is one of the most popular applications of unsupervised learning. It categorizes customers based on their demographic and behavioral traits.

Customer Segmentation

With unsupervised learning algorithms, we can group customer-based similarities such as:

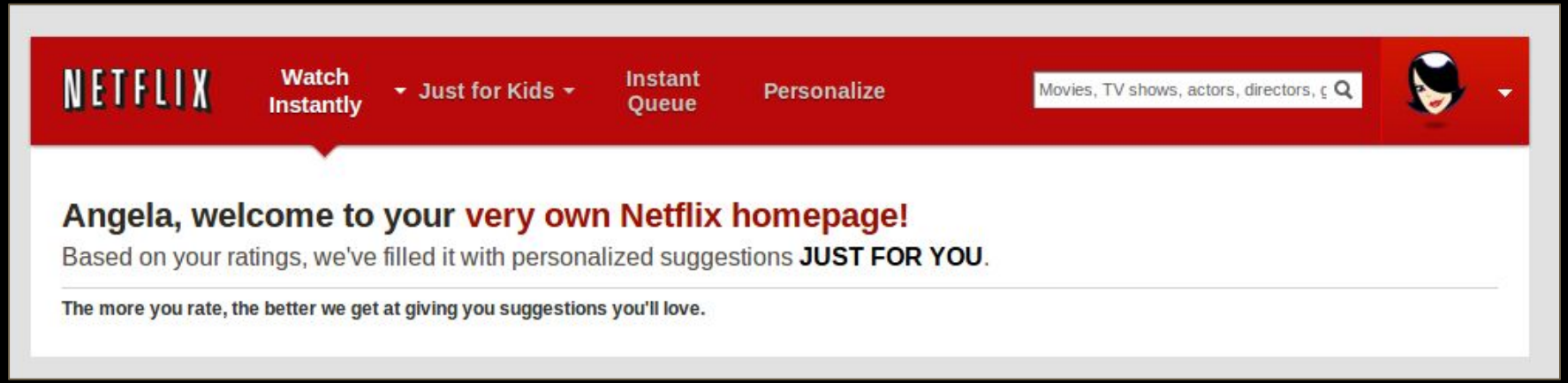
- Customer needs (e.g., a particular product can satisfy some of them)
- Responses to online marketing channels
- Buying habits (e.g., best day for buying, weekly spend, etc.)

Customer segmentation is a major revenue driver for leading companies like Netflix and Amazon.

Customer Segmentation

75% of Netflix viewer activity is driven by recommendation.

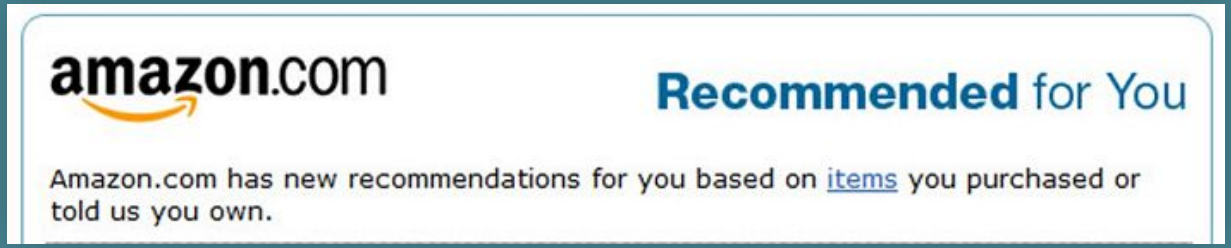
Netflix's recommendation system saves the company an estimated $1 billion per year through reduced churn.



wired.com

Customer Segmentation

35% of Amazon's sales are generated through their recommendation engine.



Data Preparation

Data Preparation

Data selection

Make a good choice of what data is going to be used. It is important to consider what is available in the dataset, what is missing, and what can be removed.

Data preprocessing

Organize the selected data by formatting, cleaning, and sampling it.

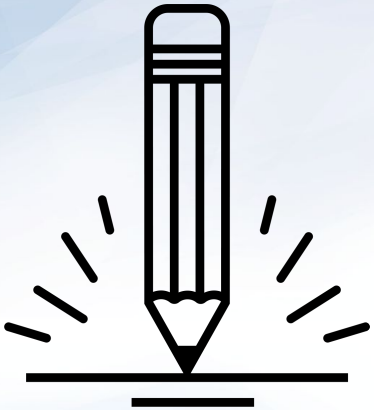
Data transformation

Transform the data to a format that eases its treatment and storage for future use (e.g., CSV file, spreadsheet, database).

Instructor Demonstration

Data Preparation for Unsupervised Learning

Questions?



Activity: Understanding Customers

In this activity, you will perform data preparation tasks on a dataset about the purchases made by 200 customers on an e-commerce website.

Suggested Time:
20 minutes

Activity: Understanding Customers

Annual Income	This feature should be regularized since it is on a different scale than the other features; dividing by `1000` is the simplest solution.
Previous Shopper	The `Previous Shopper` should be transformed to a numerical value; in this case, transforming `Yes` to `1` and `No` to `0` is a feasible solution.
CustomerID	Since this column is not relevant to the clustering algorithm, it should be dropped from the `DataFrame`.

Let's Review

Questions?

Boston Marathon Preprocessing

Suggested Time:

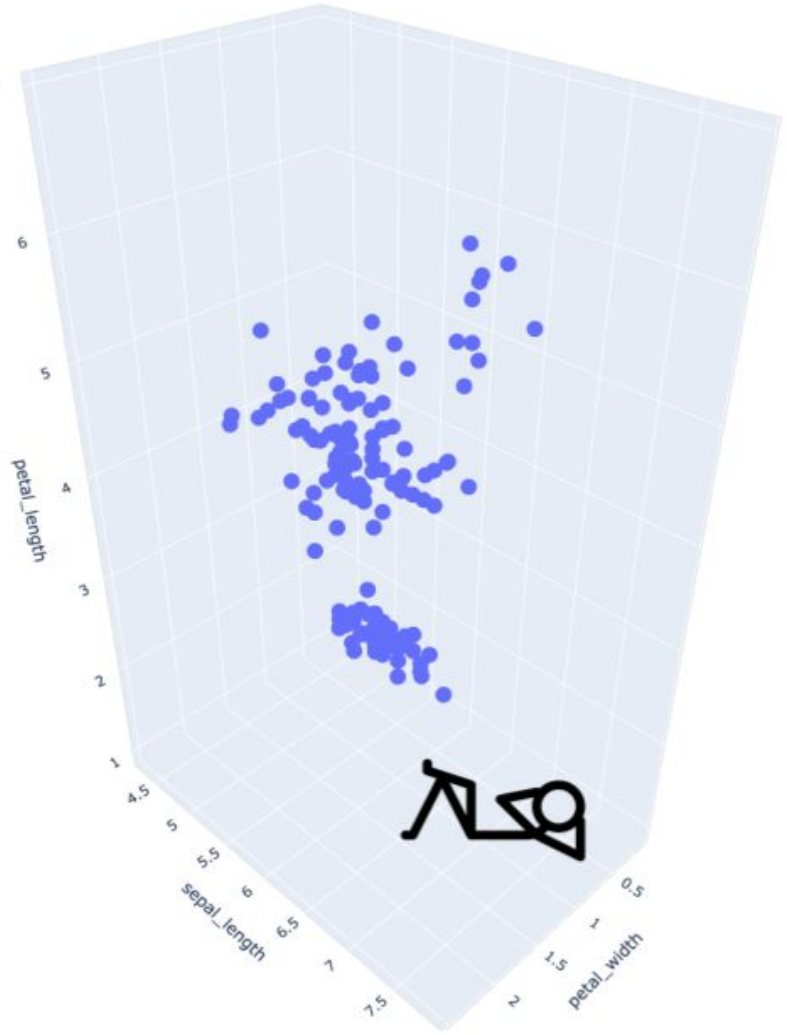
20 minutes

K-Means Algorithm

What Can K-Means Do for Me?

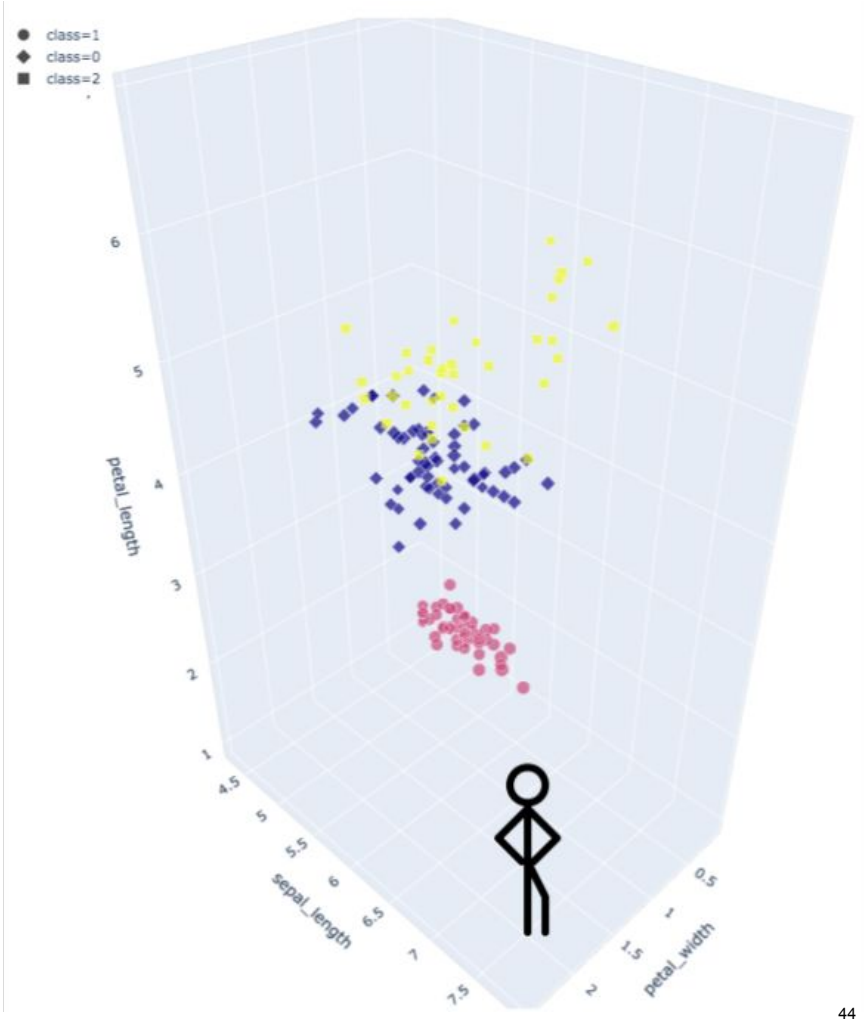
Imagine that you are in a room full of small spheres (data points).

Each sphere represents a flower (iris), and the axes represent features of flowers.

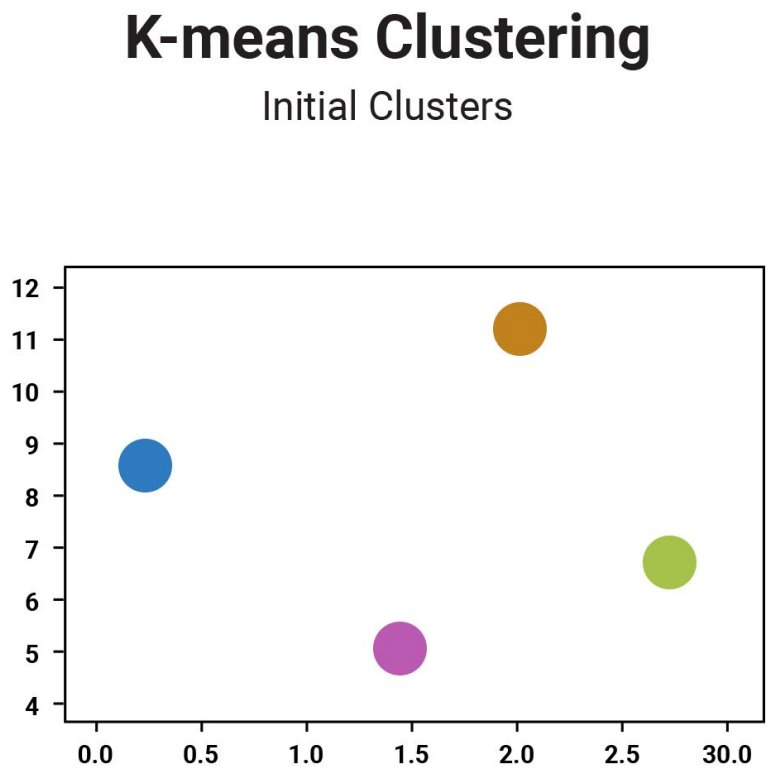


What Can K-Means Do for Me?

K-means is an unsupervised learning algorithm used to identify clusters.



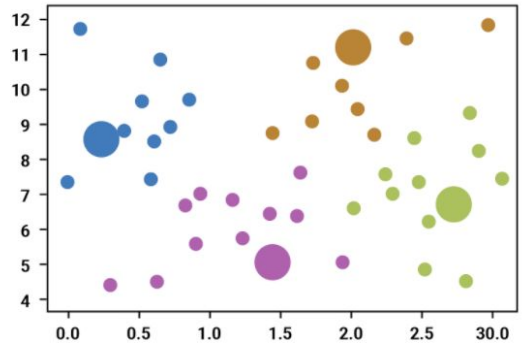
How the K-Means Algorithm Works



How the K-Means Algorithm Works

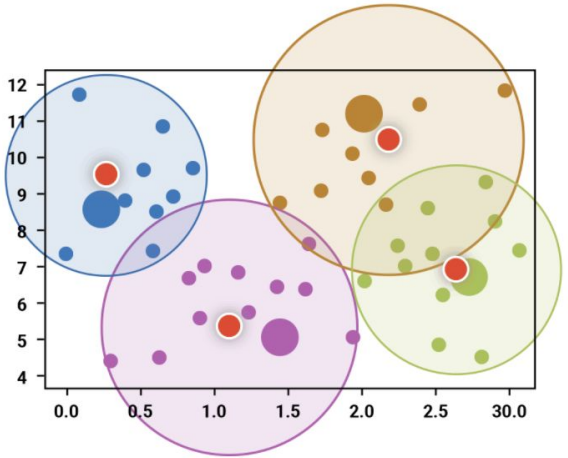
k clusters

The K-means algorithm groups the data into *k* clusters, where belonging to a cluster is based on some similarity or distance measure to a centroid.



Centroid

A centroid represents a data point in the arithmetic mean position of all the points on a cluster.



How the K-Means Algorithm Works

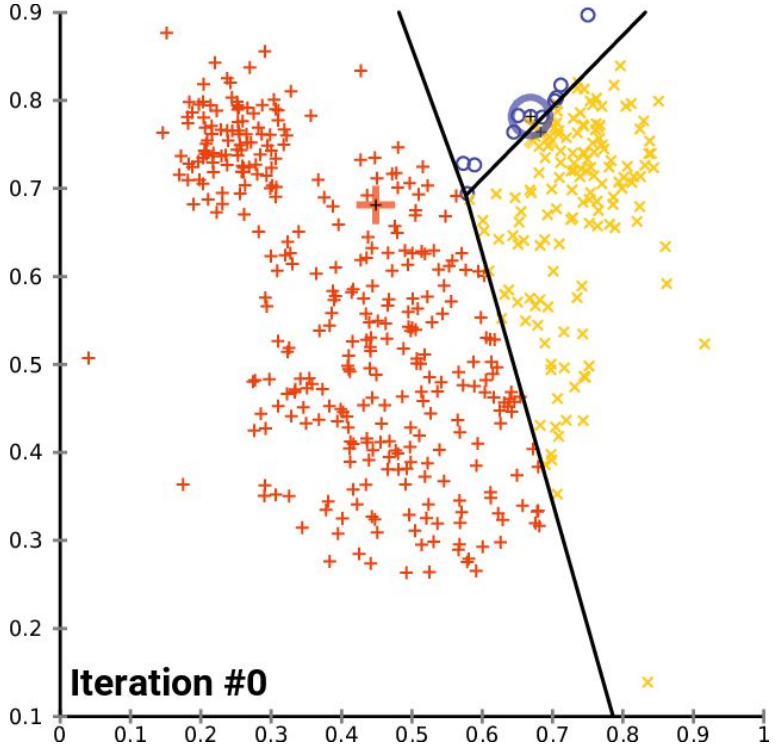
Algorithm at a glance:

01 Randomly initialize the *k* starting centroids.

02 Each data point is assigned to its nearest centroid.

03 The centroids are recomputed as the mean of the data points assigned to the respective cluster.

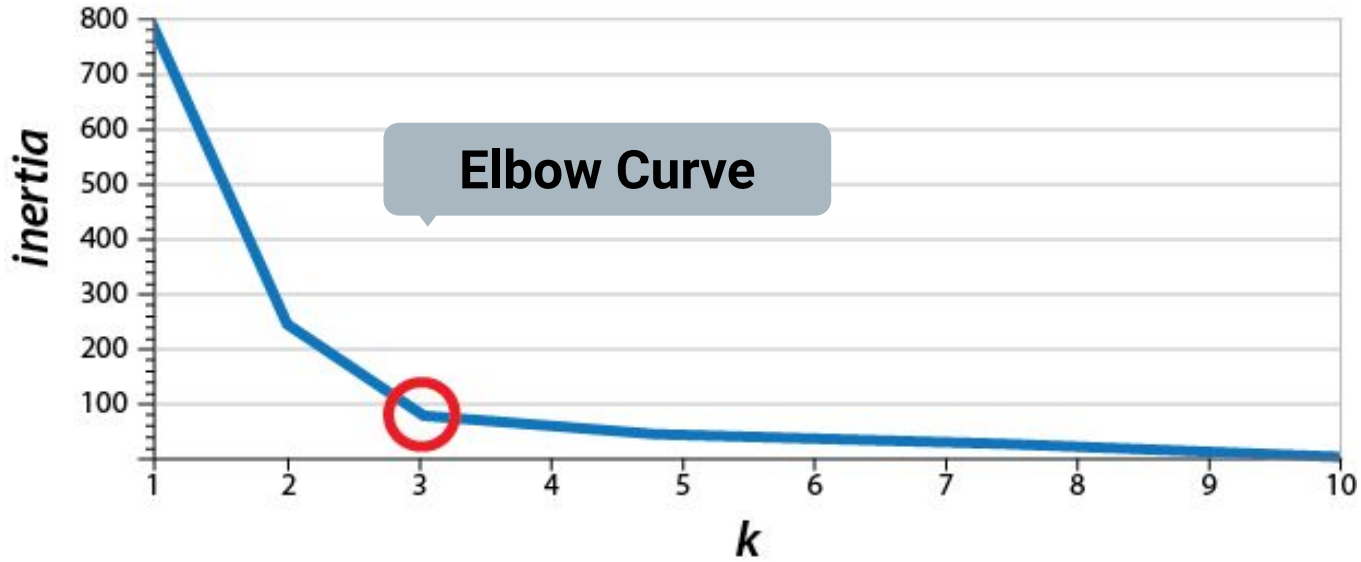
04 Repeat steps 1 through 3 until the stopping criteria is triggered.



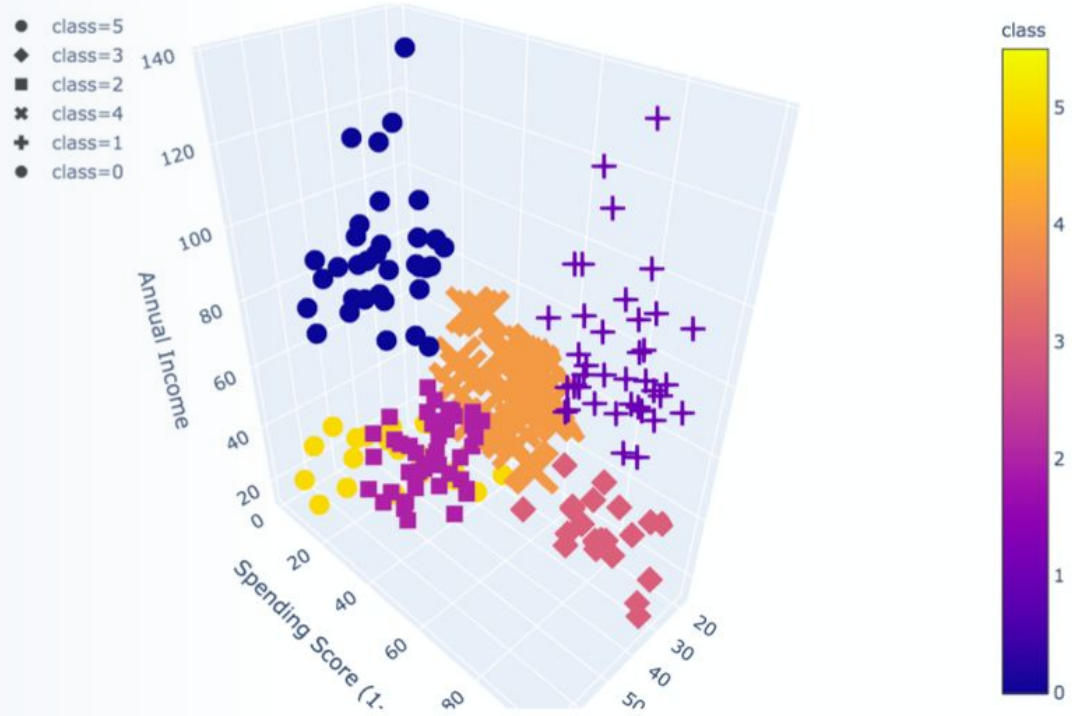
What Is the Best Number for *k*?

This is done using an **elbow curve** where the *x*-axis is the *k* value and the *y*-axis is some *objective function*.

A common objective function is the *inertia*.



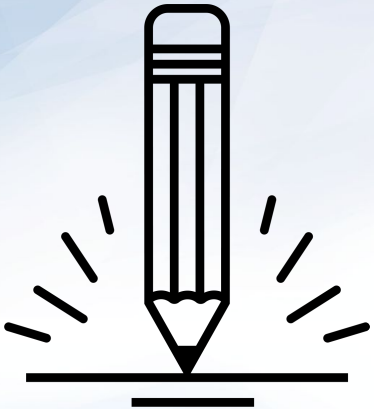
K-Means with 3 Features



Instructor Demonstration

Chicago Marathon K-Means

Questions?



Activity: K-Means

In this activity, you will will apply the K-means algorithm on a different marathon dataset.

Suggested Time:
20 minutes

Let's Review

Questions?