Linear Regression and Unsupervised Learning

FinTech

Lesson 13.1

Class Objectives

By the end of this lesson, you will be able to:

Recognize the differences between supervised and unsupervised machine learning.

Use the linear regression algorithm to find the relationship between two variables, as well as to forecast future scenarios.

Apply the k-means algorithm to identify clusters on a given dataset.

Demonstrate how the k-means algorithm is useful for customer segmentation.

Speed up machine-learning algorithms by using principal component analysis.

In this unit, we will deploy machine learning in the cloud using Amazon Web Services (AWS).

Thanks to the capabilities of the cloud, which has more computing power than your personal computer, you will be able to optimize the performance of your machine learning models.

AWS Services

In this unit, you will get hands-on experience with the following AWS services:

Amazon SageMaker:	To deploy machine-learning models
Amazon Lex:	To create conversational interfaces
Amazon S3:	To store files in the cloud
AWS Lambda:	To create serverless applications

Cloud services are fantastic tools for fintech professionals.

The skills you have already learned in this course—Python, APIs, time series, classification algorithms, and NLP—will empower the cloud applications that you create.

Linear Regression

Linear regression is a model that is intended to describe the relationship between a numerical response and one or more variables that explains that response.

Linear Regression

In statistics and machine learning:

Dependent variable

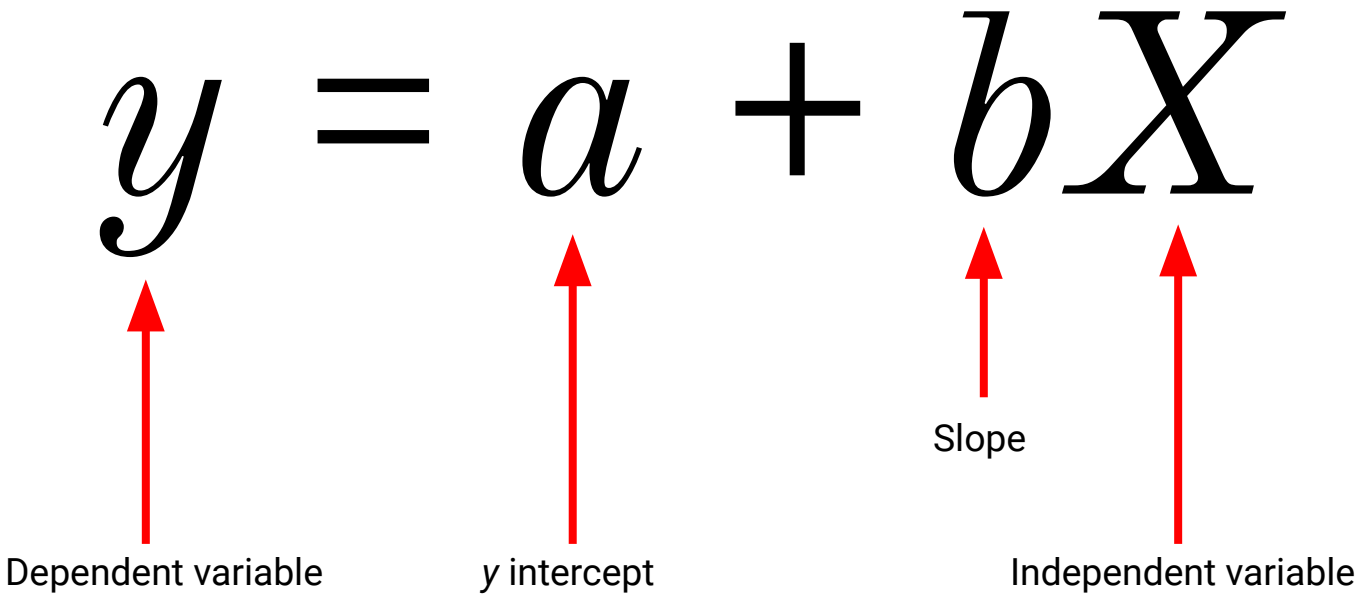
The numerical response is known as **dependent variable**, because its value depends on other variables.

Independent variable

These other variables that explain the dependent variable are known as **independent variables**.

Linear Regression

We will explore simple linear regression, where we have one explanatory (independent) variable. This type of linear regression is represented by the following formula:



$$y = a + bX$$

This linear relationship implies the following:

- As *X* increases, *y* increases.
- How fast *y* increases in relation to *X* is called the slope.
- The slope is represented by the letter *b* in the formula.
- The value of *y* when *X* is 0 is called the *y*-intercept. It is represented by the letter *a*.

From the perspective that the linear regression model can predict the value of *y* based on historical data, this model is considered as a supervised learning model.

Linear Regression

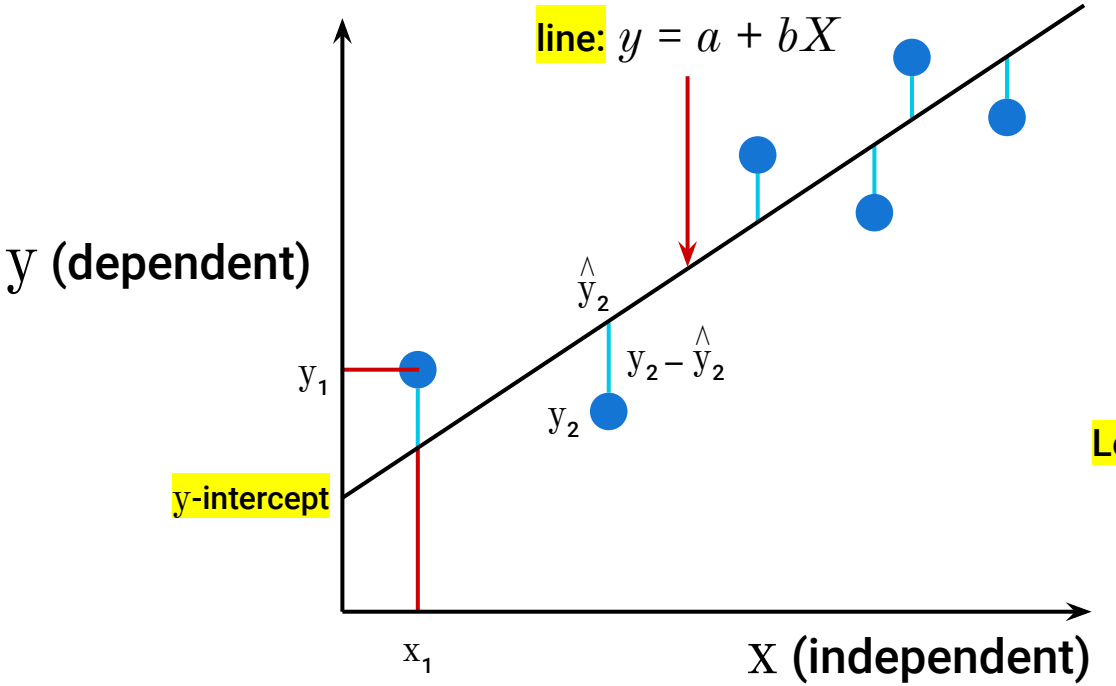
Let’s implement a simple linear regression model by using scikit-learn.

Instructor Demonstration

Linear Regression Model

Linear Regression Model

The linear regression model is mathematically constructed to minimize the sum of all the errors after they have been squared.



Minimize: $\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$

Least squares method

Summary

Key points of linear regression:

It models data with a linear trend. It is not useful when the data does not follow a linear trend, e.g., exponential trends.

Based on the *X* values, it predicts *y* values.

It does not do a good job of describing nonlinear patterns.

We will cover techniques to model nonlinear data later in the course.

One way to assess the accuracy of a linear regression model is to observe the errors.

Linear Regression Model

The linear regression model is mathematically constructed to minimize the sum of all the errors after they have been squared, as the following image shows.

mean squared error (MSE)	The average of the square of the errors of the dataset. It is the variance of the errors in the dataset.
root mean square error (RMSE)	The square root of the MSE. It is therefore the standard deviation of the errors in the dataset.
correlation coefficient	A numerical description of the extent to which the two variables move together. It ranges from -1 to 1.
R2 or r-square value	The square of the correlation coefficient. It describes the extent to which a change in one variable is associated with the change in the other variable. It ranges from 0 to 1.

Low MSE and RMSE scores indicate a more accurate model.

Questions?

Activity: Predicting Sales with Linear Regression

In this activity, you will apply linear regression to predict sales based on historical data.

Suggested Time:

15 Minutes

Time's Up! Let's Review.

Questions?

Introduction to Unsupervised Learning and Clustering

Machine Learning

Machine learning uses algorithms, which are essentially a block of instructions designed for a computer, to train a model that analyzes and makes predictions about something that we expect to happen.

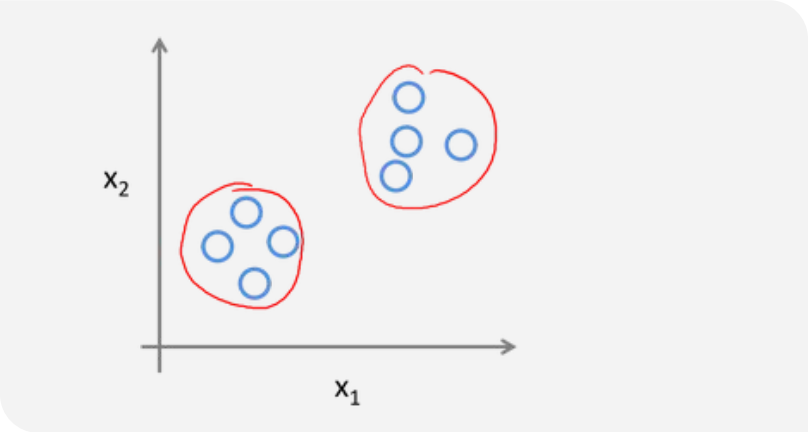
Machine Learning

We can use machine learning to help categorize a model that defines borrowers as likely or unlikely to default on their mortgage loans.

If machine learning can help determine the set of characteristics that define each of those groups, a fintech lender might be better able to focus their advertising dollars on individuals that fall in the favorable category.

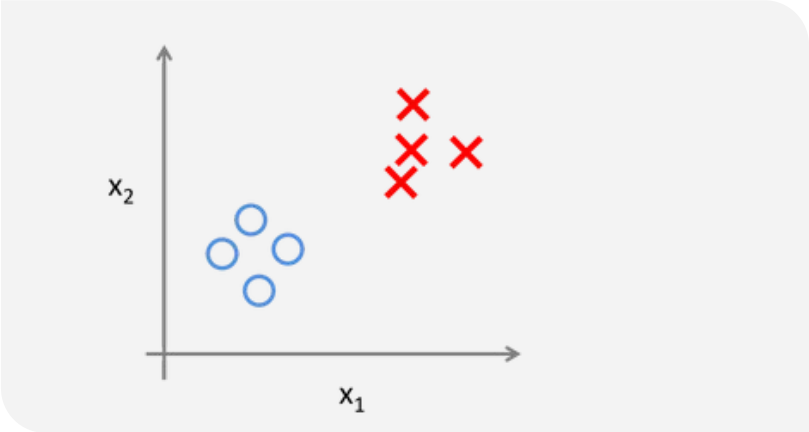
Unsupervised vs. Supervised Learning

Unsupervised Learning



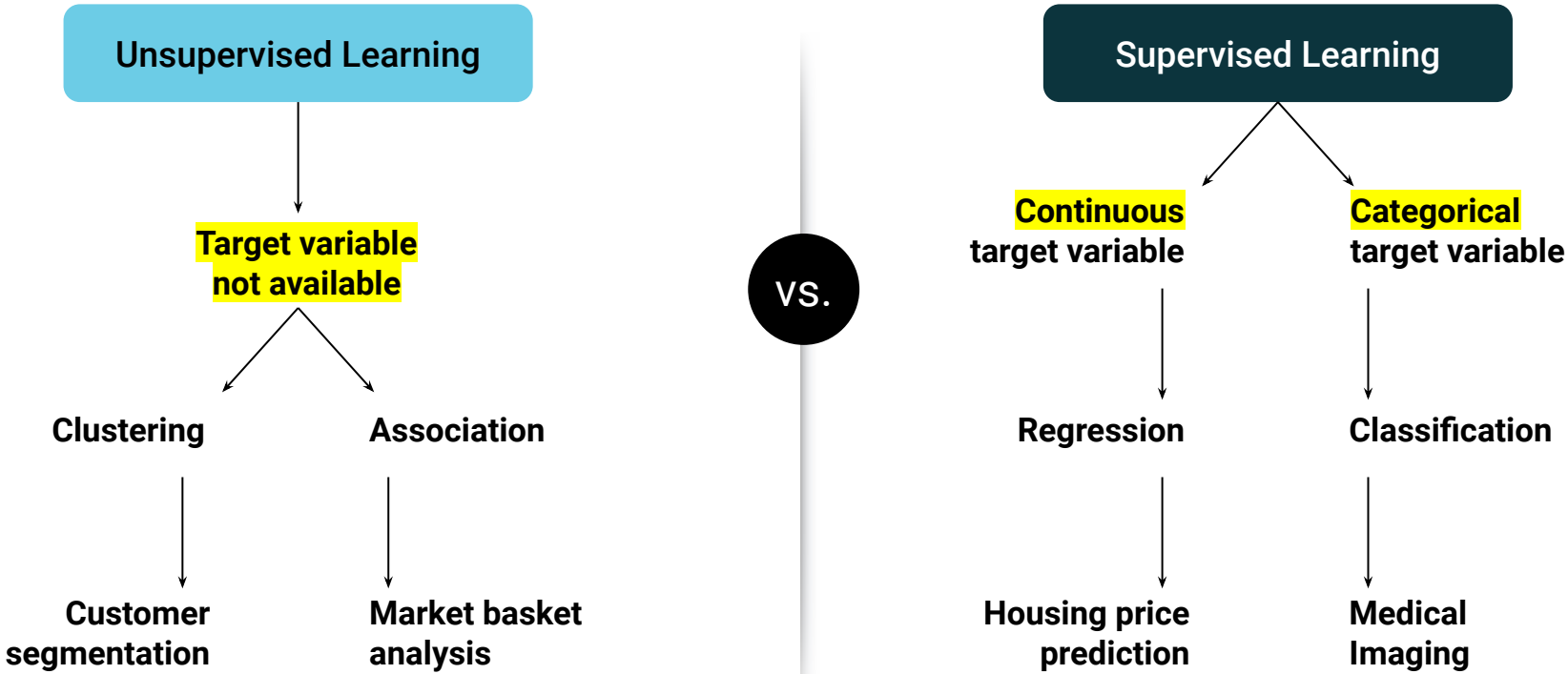
- Input data is unlabeled.
- Only uses input datasets.
- **Goal:** determine patterns or groups of data, called data clusters.

Supervised Learning



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

Unsupervised Learning vs. Supervised Learning



Unsupervised Learning

Common uses include customer segmentation tasks like:

Grouping customers by spending habit

Finding fraudulent credit card charges

Identifying unusual data points (outliers) within a dataset

How might clustering be used by fintech businesses?

One possible answer:

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

How might anomaly detection be used by credit card companies?

One possible answer:

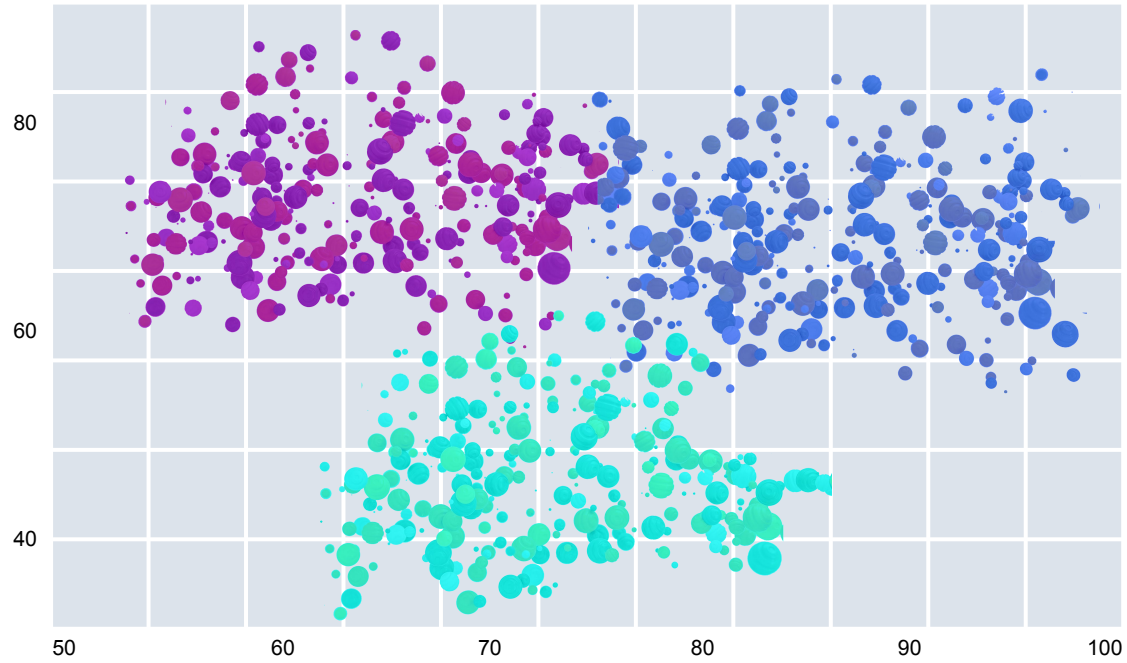
Anomaly detection can be used to detect potentially fraudulent credit card transactions by grouping transactions into normal and abnormal.

Introduction to Clustering

Clustering is grouping together data that is similar or related in some way.

Introduction to Clustering

Unsupervised learning is usually applied via a clustering algorithm.



Instructor Demonstration

Introduction to Clustering

Introduction to Clustering

The process of clustering data points into groups is called **centering**.

In advanced analytics, centering helps to determine the number of classes or groups to create.

Centering improves the performance of logistic regression models by ensuring that all data points share the same starting mean value.

Data points with the same starting mean value are clustered together.

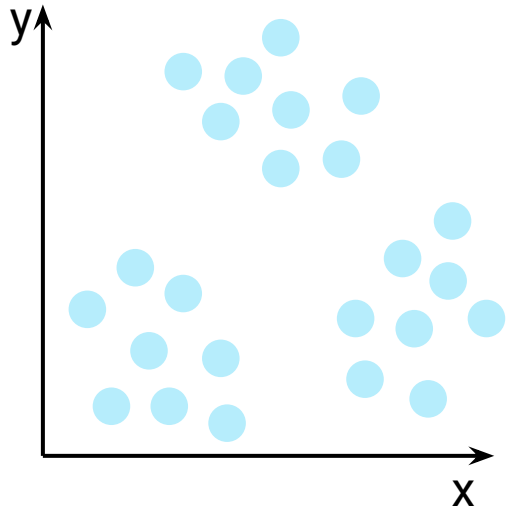
The datasets you will work with as fintech professionals do not have a defined set of clusters at the outset.

Introduction to Clustering

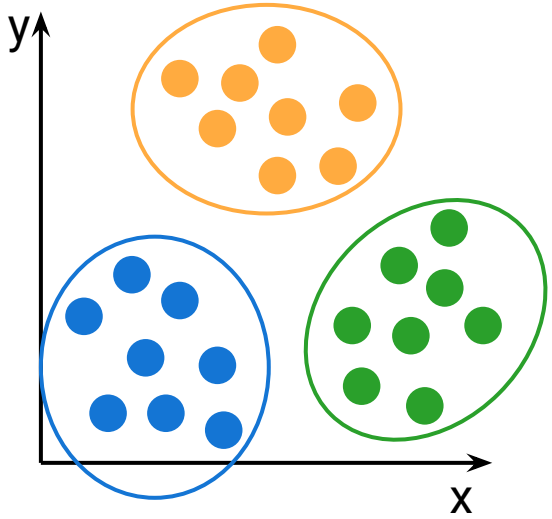
The purpose of the unsupervised learning algorithm is to identify the number of distinct sets of clusters in the dataset.

This is accomplished with the **K-means algorithm**, which we'll learn about next.

Before K-means



After K-means



Questions?

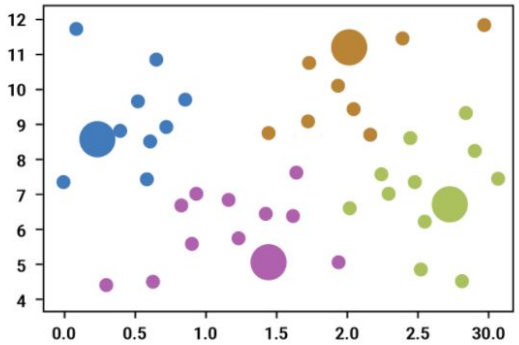
The K-means Algorithm

The **k-means algorithm** is the simplest and most common algorithm for grouping data points into clusters.

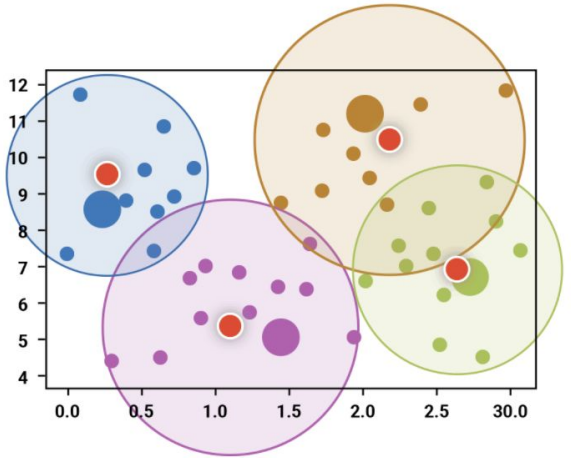
The K-means Algorithm

K-means takes a predetermined amount of clusters and assigns each data point to one of those clusters. It works by performing two steps:

01 **The algorithm assigns points to the closest cluster center.**



02 **The algorithm re-adjusts the cluster's center by setting each center as the mean of all the data points contained within that cluster.**



Instructor Demonstration

The K-means Algorithm

Activity: Segmenting Customers

In this activity, you will use the K-means algorithm to segment customer data for mobile versus in-person banking service ratings.

You will evaluate the data by segmenting the data into three and four clusters.

Suggested Time:

15 Minutes

Time’s Up! Let’s Review.

Questions?


Break

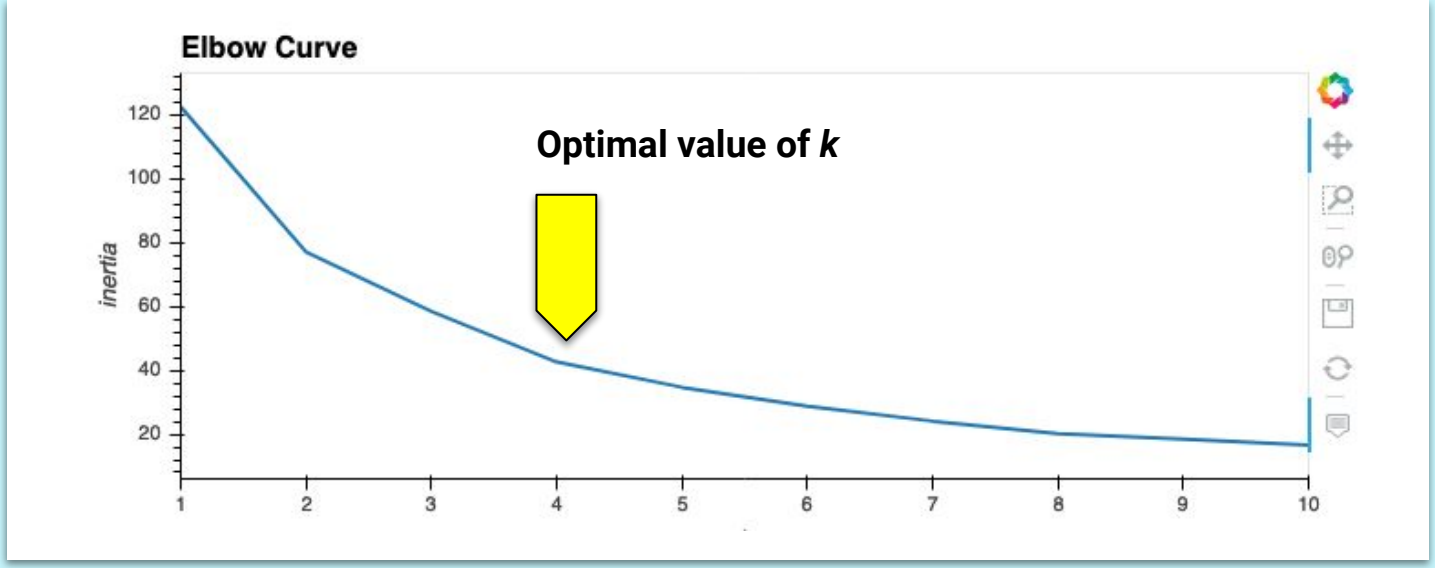
The Elbow Method

Because the K-means algorithm requires the number of clusters ahead of time, we need to be sure the number of clusters that we choose is correct.

The Elbow Method

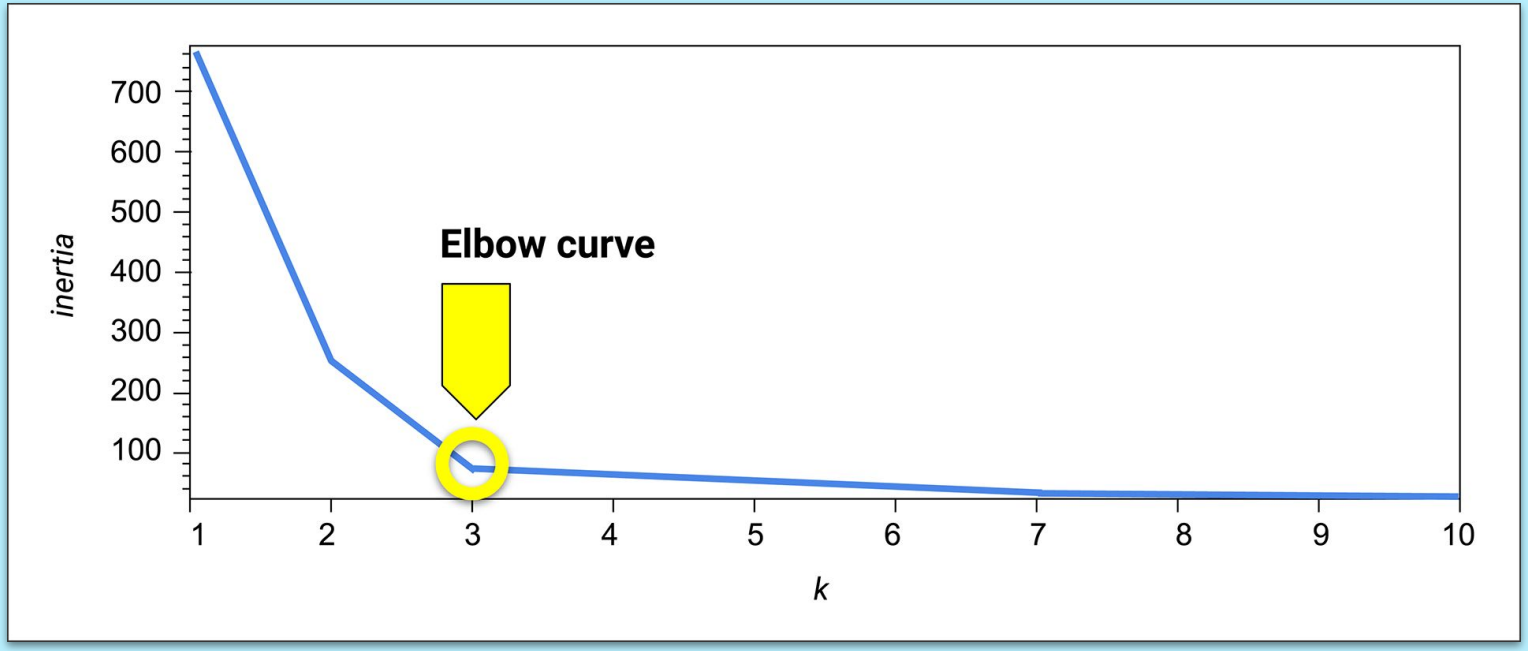
One method for determining the optimal value of k, or the number of clusters in a dataset, is the **elbow method**.

- The elbow method runs the K-means algorithm for a range of possibilities for k, or the number of clusters.
- The resulting elbow curve plots the number of clusters, x, vs. an objective function called inertia.



Elbow Curve

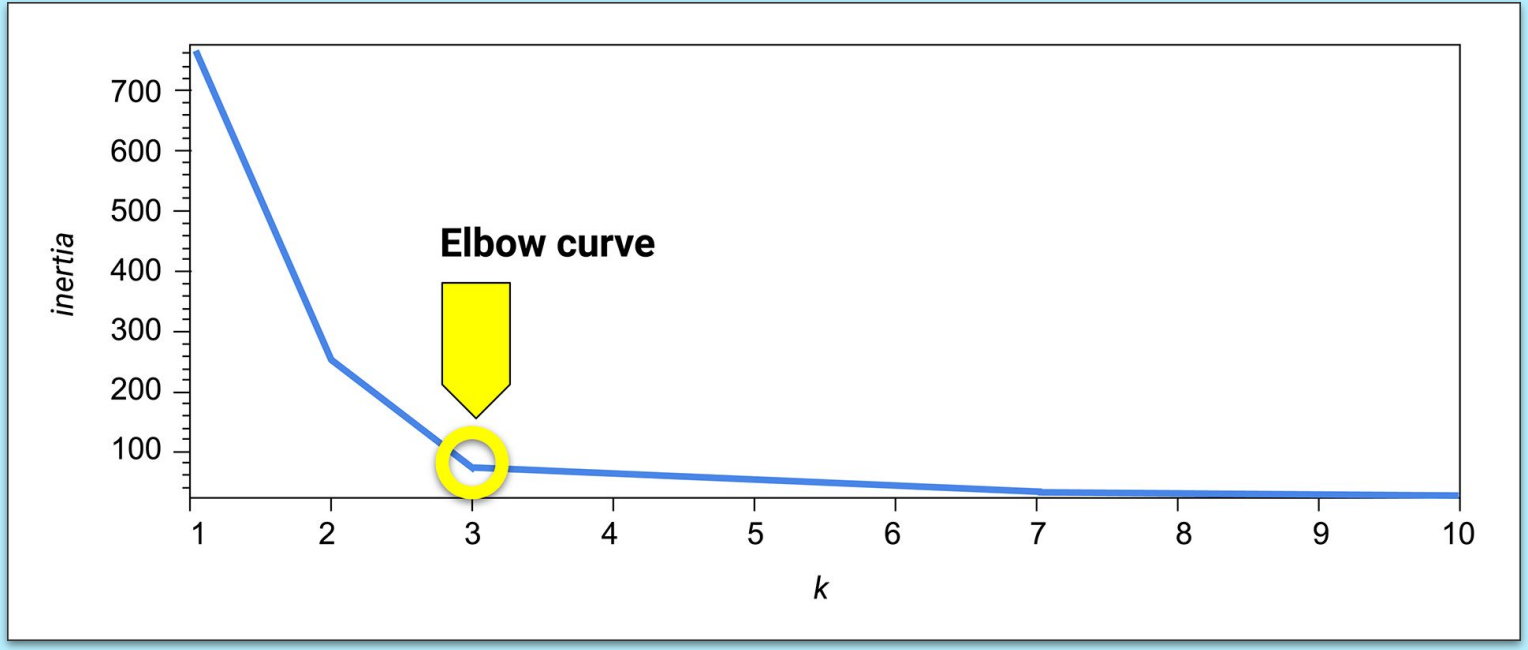
The **elbow curve** is commonly used to figure out the best value of k. It is essentially used to determine the number of clusters at which the data points become tightly clustered.



Elbow Curve

On the elbow curve, the x axis is the value of clusters, while the y axis is a metric used to assess the value of k.

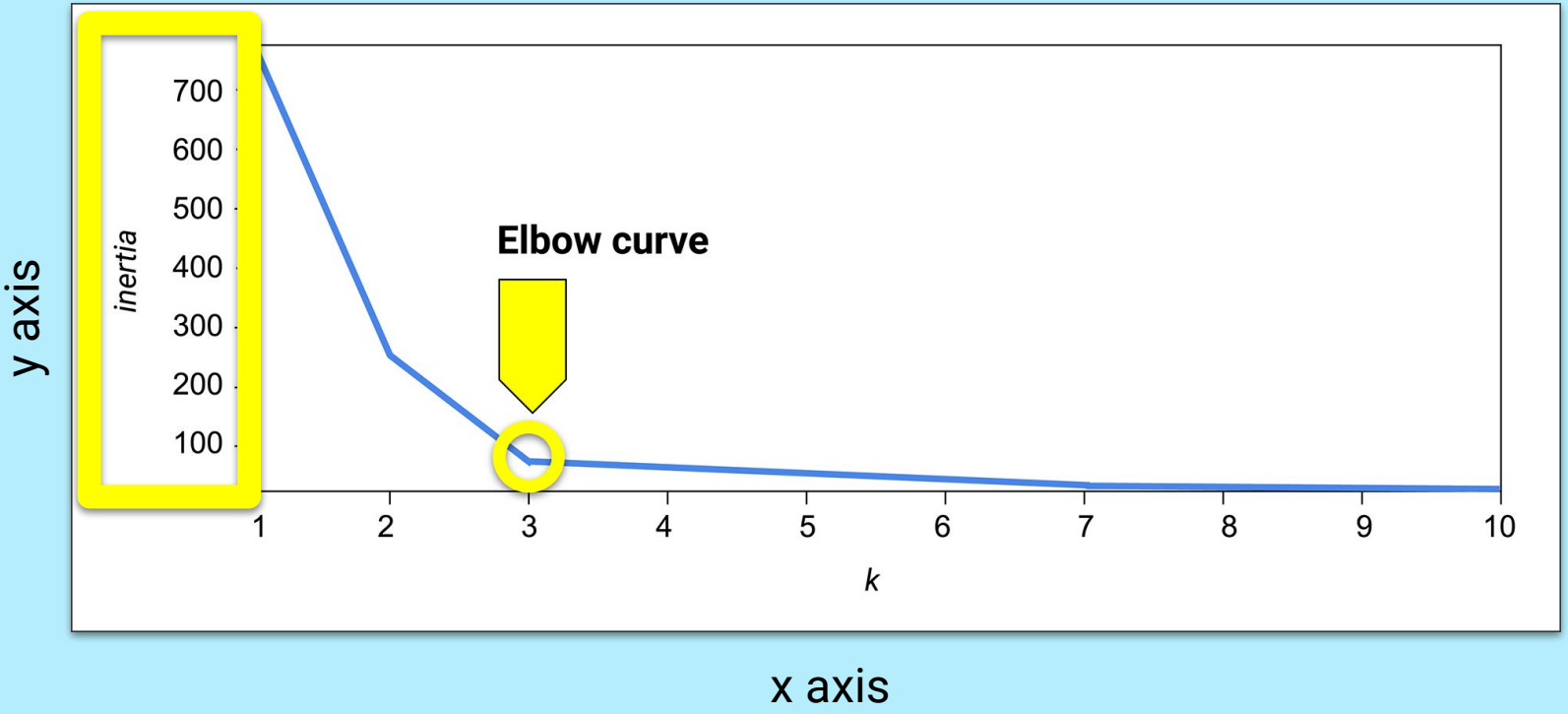
y axis



x axis

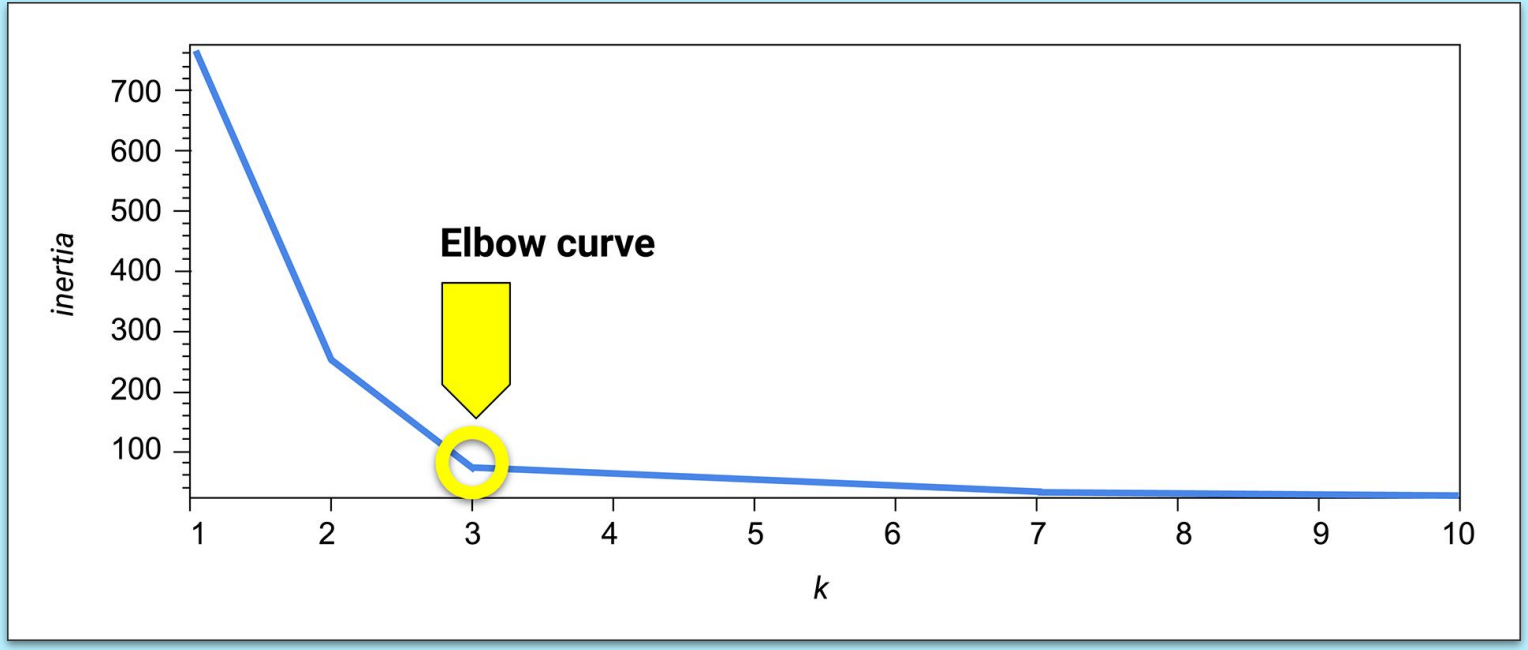
Elbow Curve

The **inertia** is commonly used as an objective function. It is the sum of squared distances of samples to their closest cluster center.



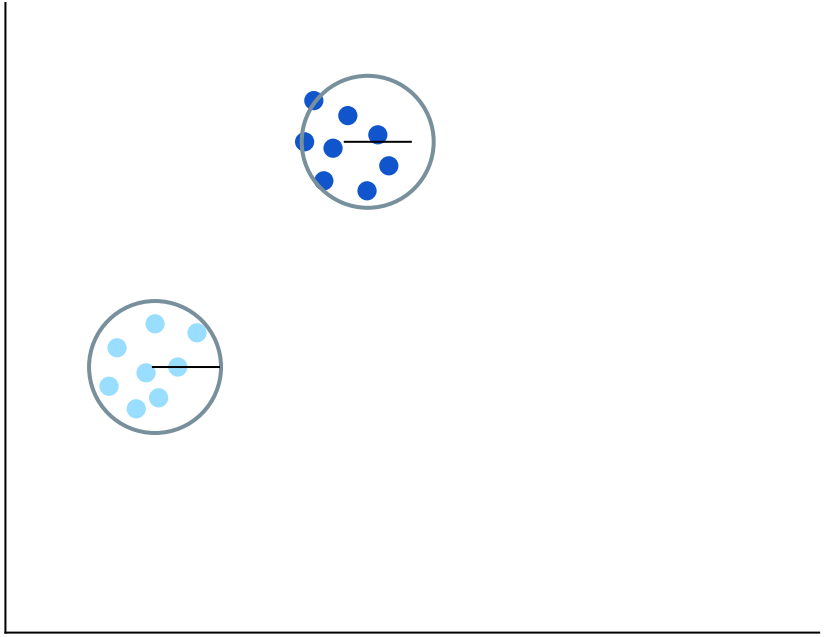
Elbow Curve

A low inertia value means that the data points are tightly clustered around the cluster center.



Low Inertia

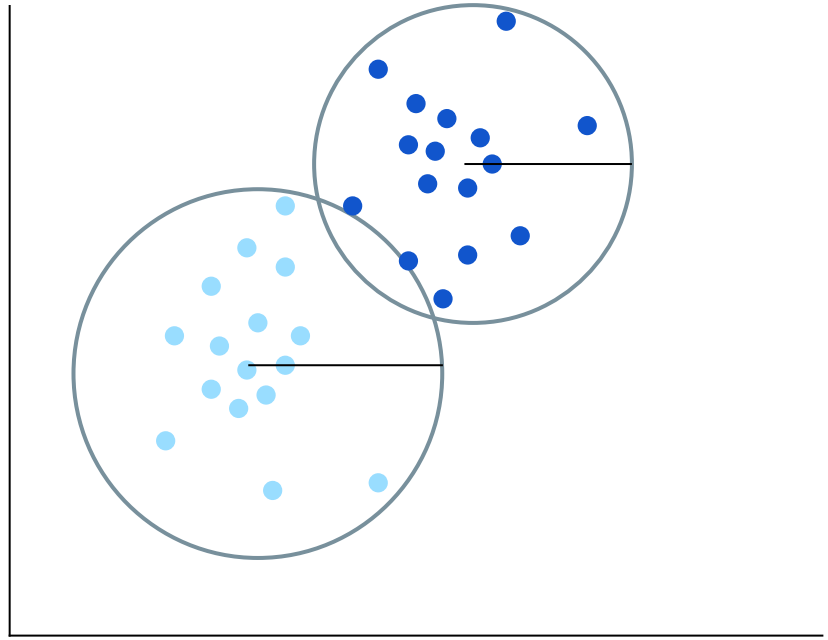
Radius of circle is small = small standard deviation from cluster mean



vs.

High Inertia

Radius of circle is large = large standard deviation from cluster mean



Instructor Demonstration

The Elbow Method

Questions?

Activity: Finding k

In this activity, you will use the elbow method to determine the optimal value for k, and then run the K-means algorithm to visualize the clusters based on that value.

Suggested Time:

15 Minutes

Time's Up! Let's Review.

Questions?

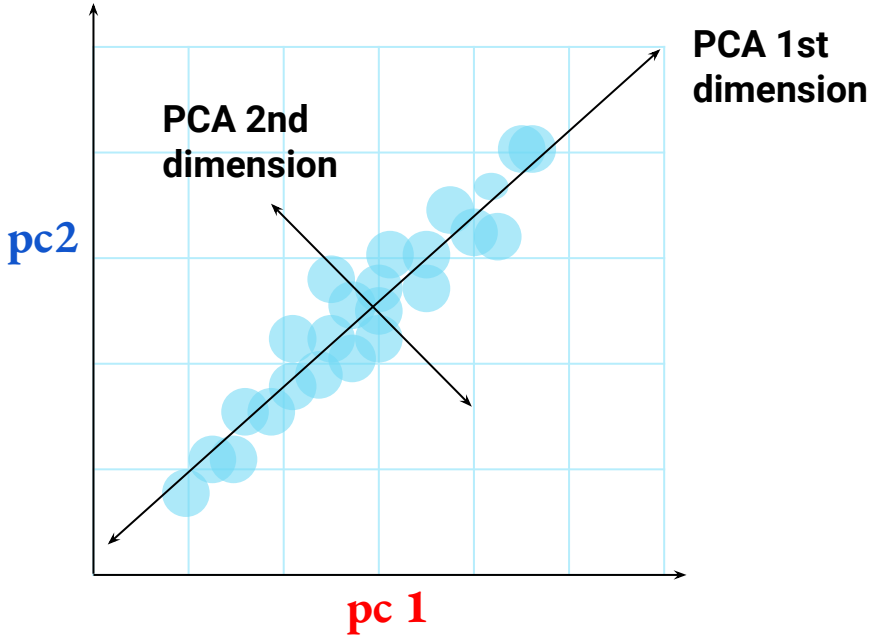
Instructor Demonstration

Introducing PCA

Principal component analysis (PCA) is a statistical technique that is used to streamline the machine learning process when too many factors exist in the data.

Principal Component Analysis

PCA reduces the number of factors by transforming a large set of features into a smaller one that contains MOST of the information of the original larger dataset.



Introducing PCA

PCA is a dimensional-reduction method that:

Looks at all of the dimensions (or data columns) in a dataset

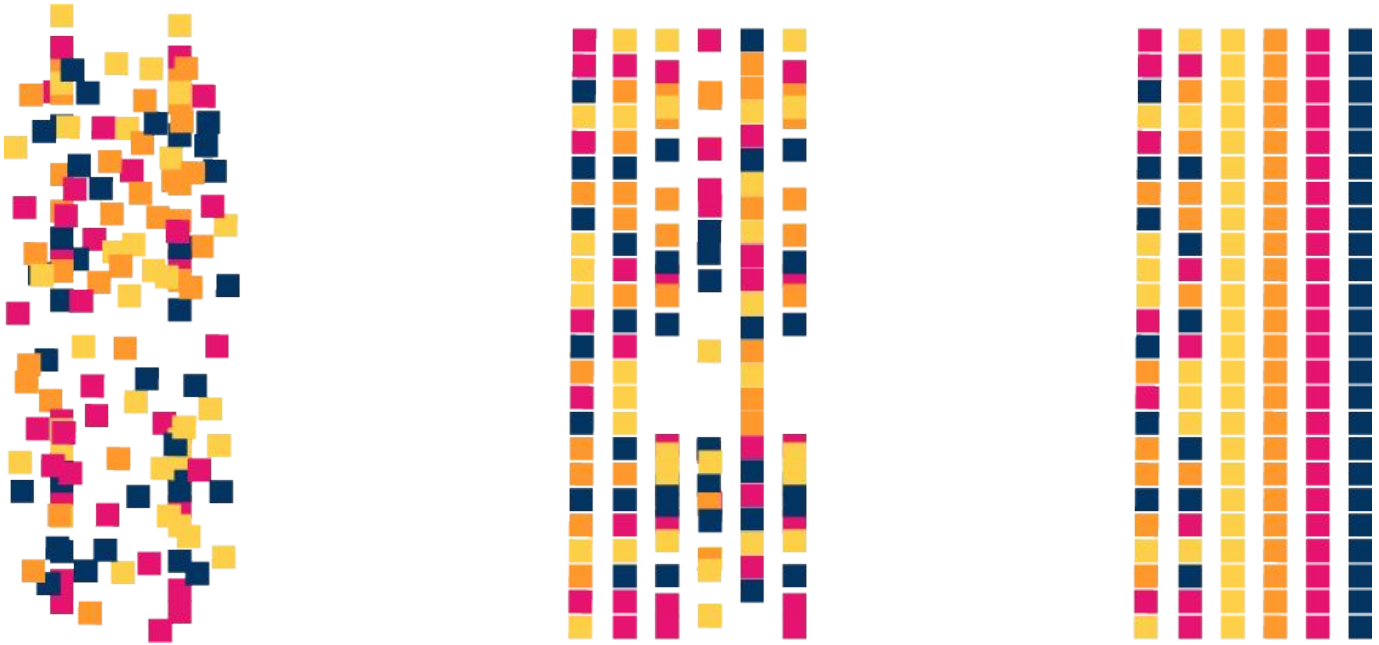
Analyzes the weight of their contribution to the variance in the dataset

Reduces them to a smaller set of dimensions that still contains as much of the information, or the maximum variance, of the original dataset as possible.

PCA will NOT capture all of the information from the original dataset, but it will capture as much as possible to maintain the predictive power and the meaning of the original dimensions.

Introducing PCA

Reducing the number of factors, or **dimensional reduction**, comes at the expense of some accuracy, but the goal is to trade a little accuracy for simplicity.



Activity: Segmenting with PCA

In this activity, you will use your knowledge of PCA to reduce the dimensionality of a dataset, and then compare that result to the segmentation of the data by using all the factors.

Suggested Time:

15 Minutes

Time's Up! Let's Review.

Questions?


The End