Python for Data Science

Introduction to K Means Clustering

Introduction to K Means Clustering

Reading Assignment

Chapter 10 of

Introduction to Statistical Learning

By Gareth James, et al.

K Means Clustering

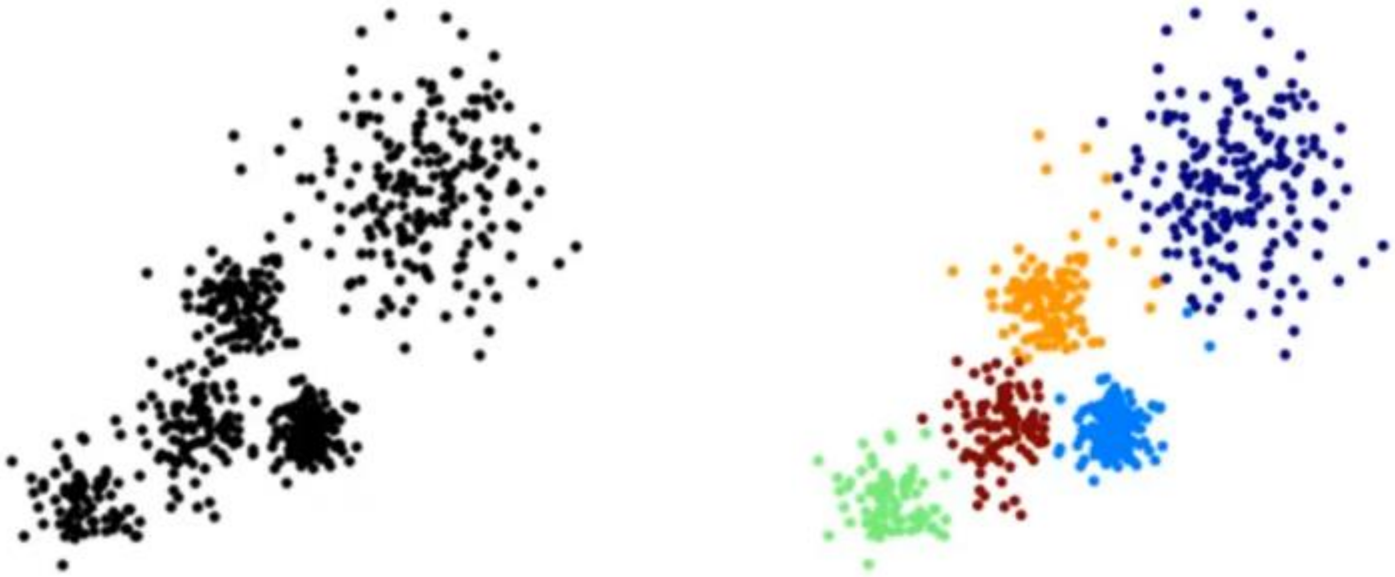
K Means Clustering is an unsupervised learning algorithm that will attempt to group similar clusters together in your data.

So what does a typical clustering problem look like?

- Cluster Similar Documents
- Cluster Customers based on Features
- Market Segmentation
- Identify similar physical groups

K Means Clustering

- The overall goal is to divide data into distinct groups such that observations within each group are similar

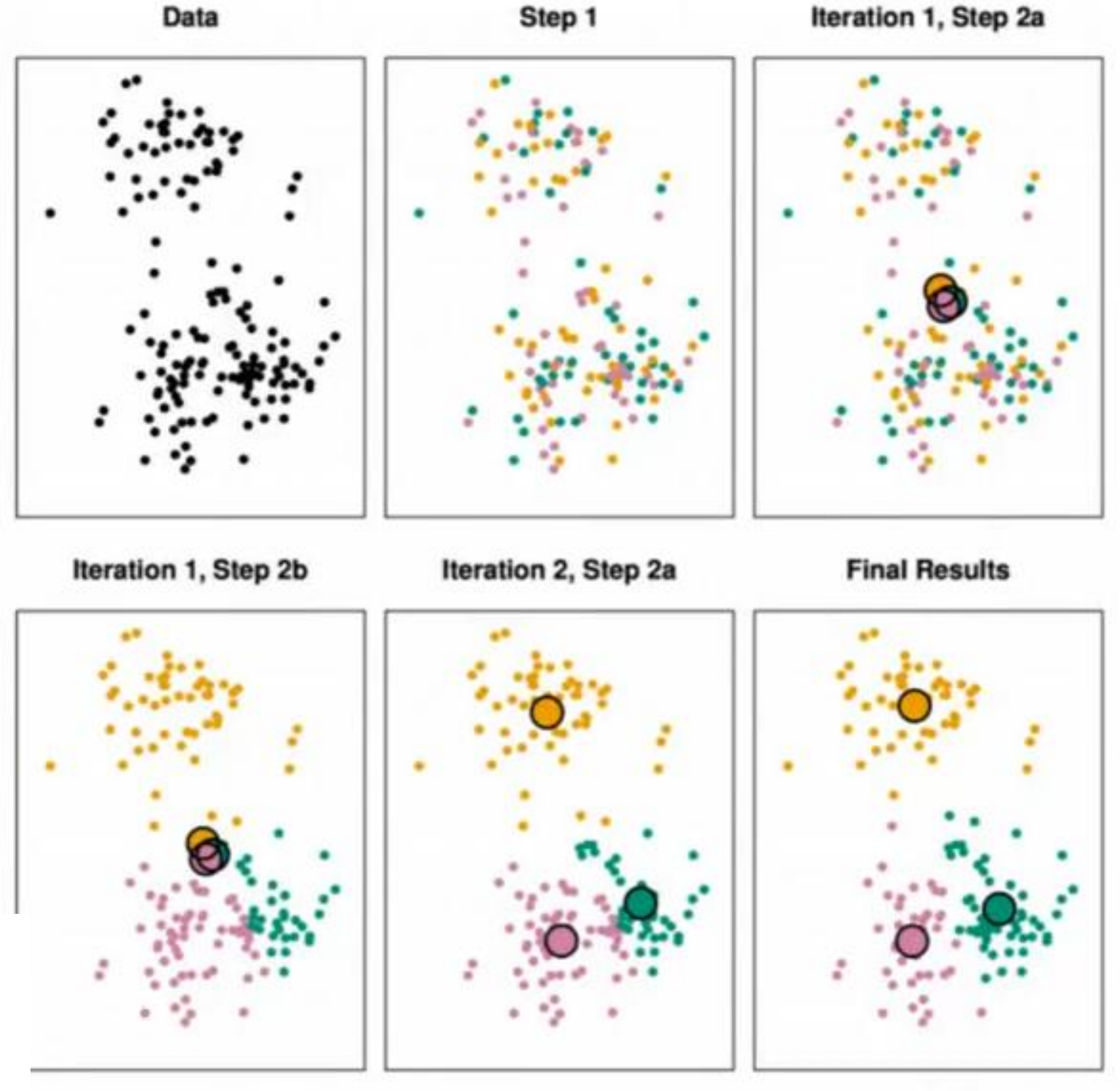


K Means Clustering

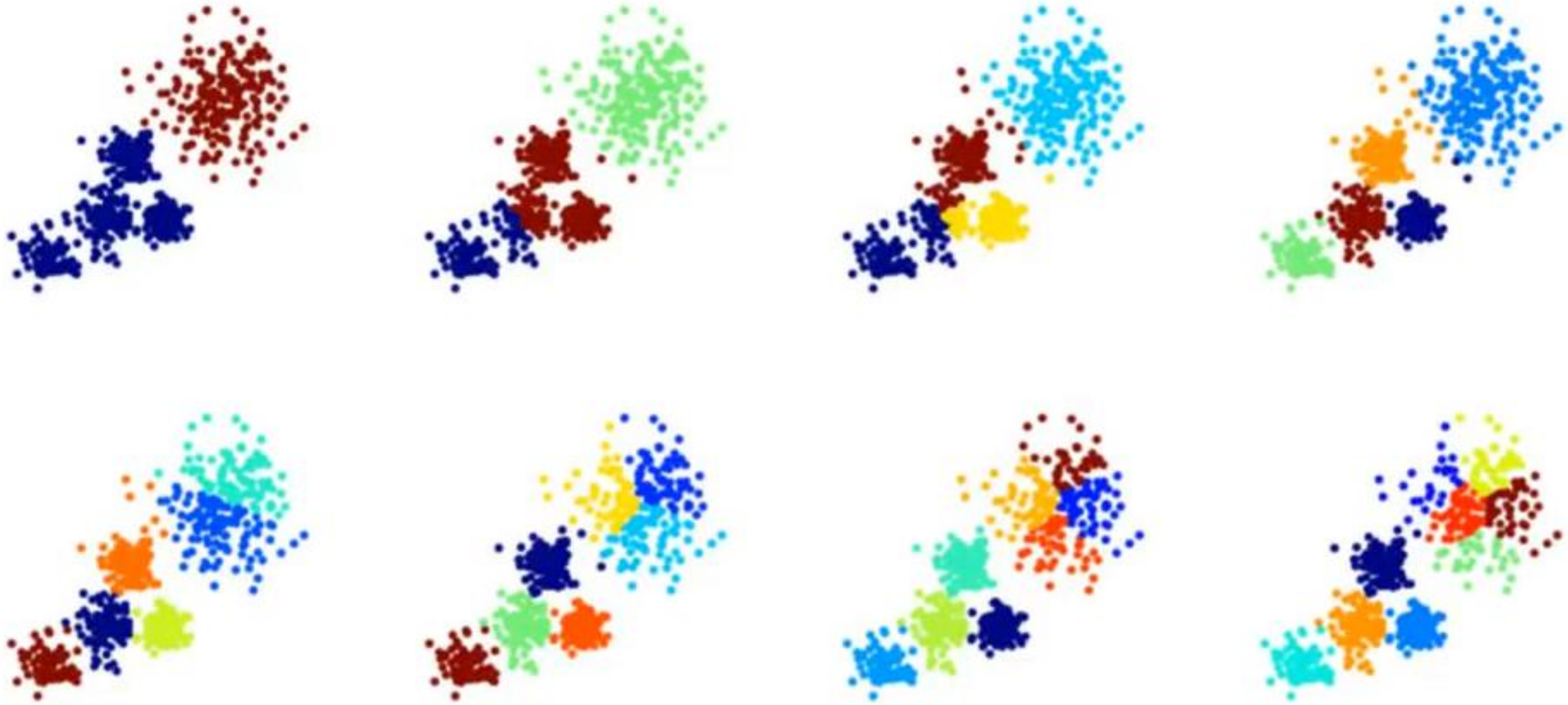
The K Means Algorithm

- Choose a number of Clusters “K”
- Randomly assign each point to a cluster
- Until clusters stop changing, repeat the following:
  - For each cluster, compute the cluster centroid by taking the mean vector of points in the cluster
  - Assign each data point to the cluster for which the centroid is the closest

K Means Clustering



Choosing a K Value



Choosing a K Value

- There is no easy answer for choosing a “best” K value
- One way is the elbow method

First of all, compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.).

The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid.

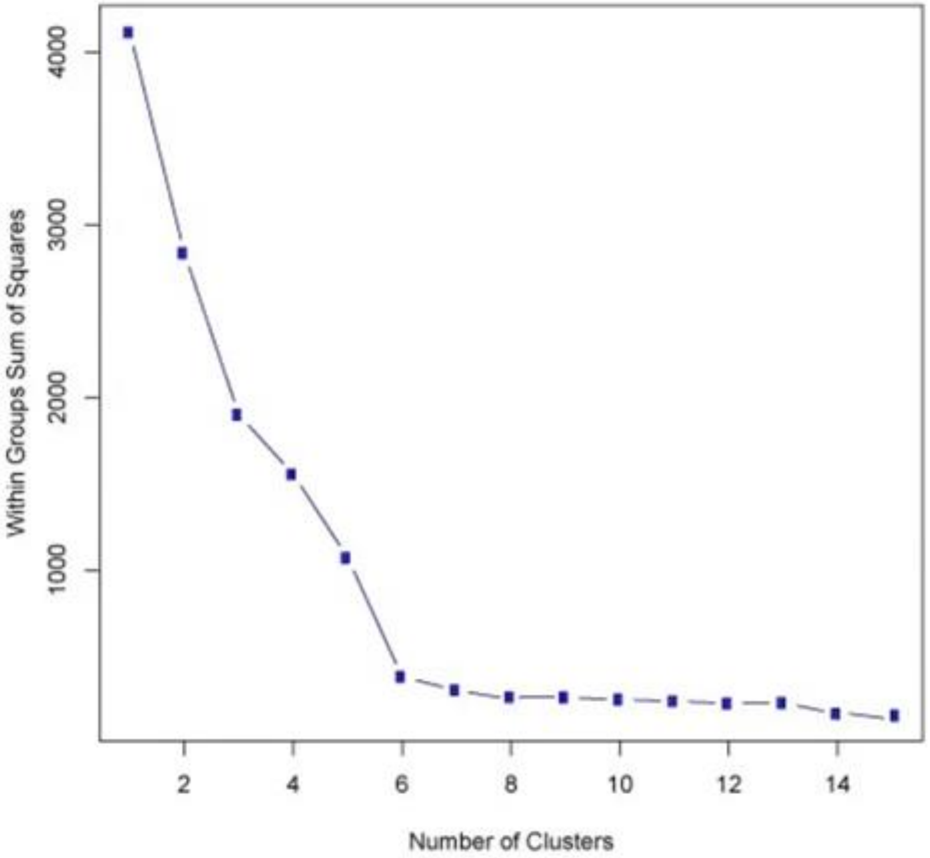
Choosing a K Value

If you plot k against the SSE, you will see that *the error decreases as k gets larger*; this is because when the number of clusters increases, they should be smaller, so distortion is also smaller.

The idea of the elbow method is to choose the k at which the SSE decreases abruptly.

This produces an "elbow effect" in the graph, as you can see in the following picture:

Choosing a K Value



Example with Python

We’ll start by showing you how to use SciKit-Learn’s own built in data generation tools to create some clusters and then test the K-means algorithm.

Then for your portfolio project you’ll work with real world data and try to cluster universities into groups based off their features and see if you can distinguish between Private and Public schools!

K Means Clustering with Python

Thanks!

Any questions ?