scikit-learn

1. scatter plot
2. line plot
3. bar chart
4. histogram
5. box plot
6. piechart

Clustering

Overview

(EDA)

- **Clustering** is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data
- It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different
- In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance

→ distance

- The decision of which similarity measure to use is application-specific
- Unlike supervised learning, clustering is considered an unsupervised learning method since we don't have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance

Applications of Clustering

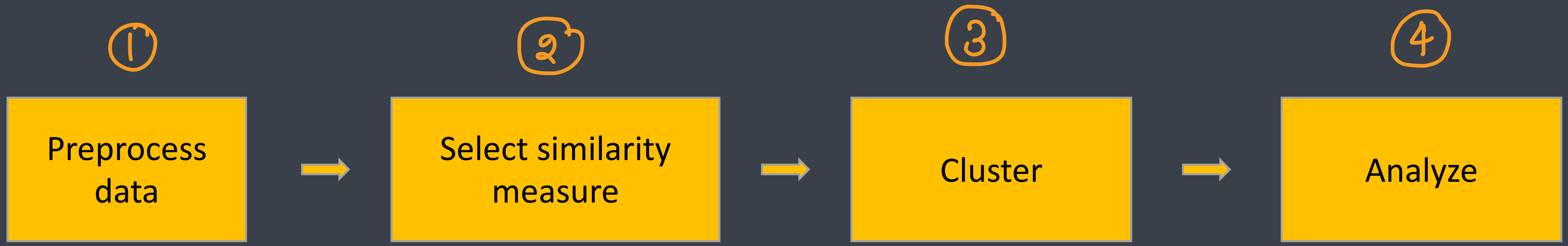
- Marketing
  - Customer segment discovery
- Library
  - To cluster different books based on topics and information
- Biology
  - Classification among different species of plants and animals
- City planning
  - Analyze the value of houses based on location
- Document Analysis
  - Various research data and documents can be grouped according to certain similarities
  - Labeling large data is really difficult. Clustering can be helpful in these cases to cluster text & group it into various categories
  - Unsupervised techniques like LDA are also beneficial in these cases to find hidden topics in a large corpus

Issues

- The results may be less accurate since data isn’t labeled in advance and input data isn’t known
- The learning phase of the algorithm might take a lot of time as it calculates and analyses all possibilities
- Without any prior knowledge the model is learning from raw data
- As the number of features increases, complexity increases
- Some projects involving live data may require continuous data feeding to the model, resulting in time-consuming and inaccurate results

Overview

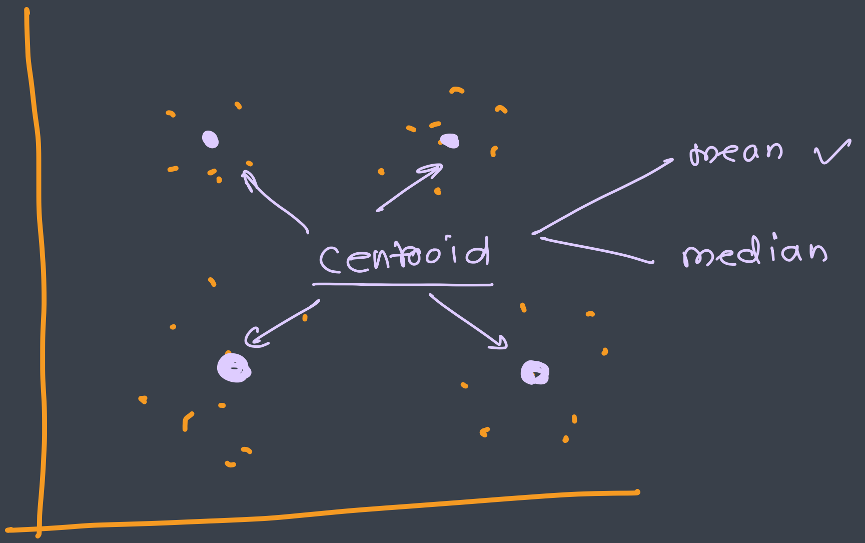
EDA



Clustering Types

Centroid-based Clustering

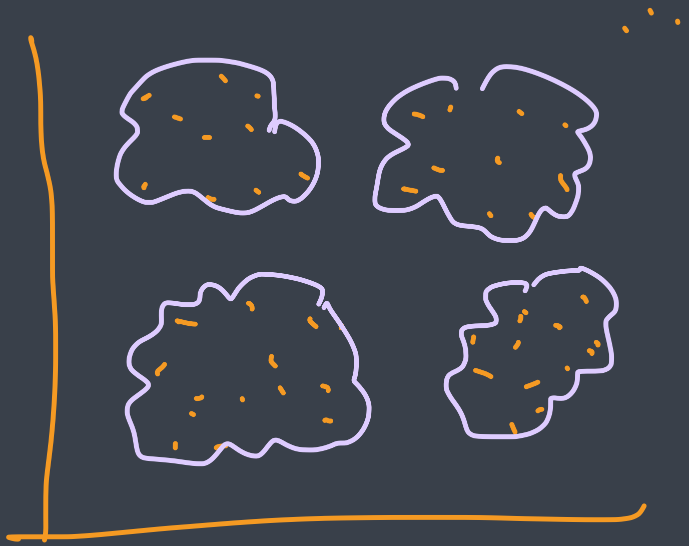
- Centroid-based clustering organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering defined below
- Centroid-based algorithms are efficient but sensitive to initial conditions and outliers
- K-Means is the most widely-used centroid-based clustering algorithm



Density based Clustering

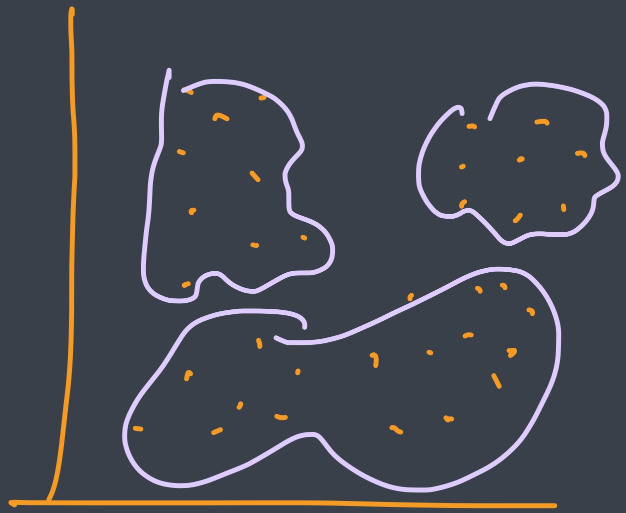
- Density-based clustering connects areas of high example density into clusters
- This allows for arbitrary-shaped distributions as long as dense areas can be connected
- These algorithms have difficulty with data of varying densities and high dimensions
- Further, by design, these algorithms do not assign outliers to clusters

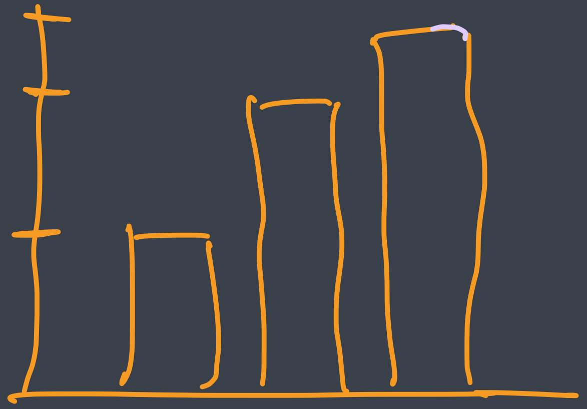
→ DBScan



Distribution based Clustering

- This clustering approach assumes data is composed of distributions, such as Gaussian distributions
- The distribution-based algorithm clusters data into three Gaussian distributions
- As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases
- The bands show that decrease in probability.
- When you do not know the type of distribution in your data, you should use a different algorithm

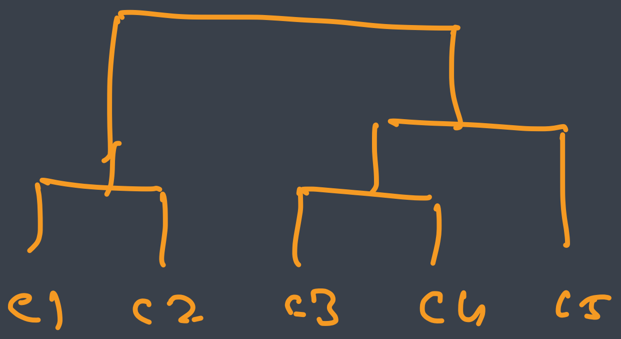




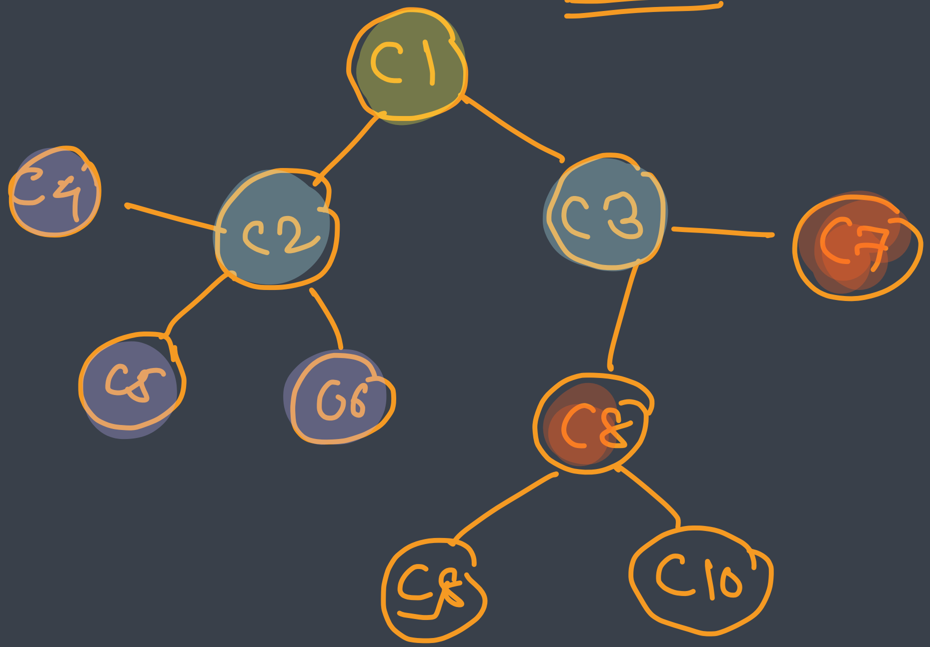
Hierarchical Clustering

- Hierarchical clustering creates a tree of clusters
- Hierarchical clustering, not surprisingly, is well suited to hierarchical data, such as taxonomies
- In addition, another advantage is that any number of clusters can be chosen by cutting the tree at the right level

dendrogram



Neo4J



k-means

Overview

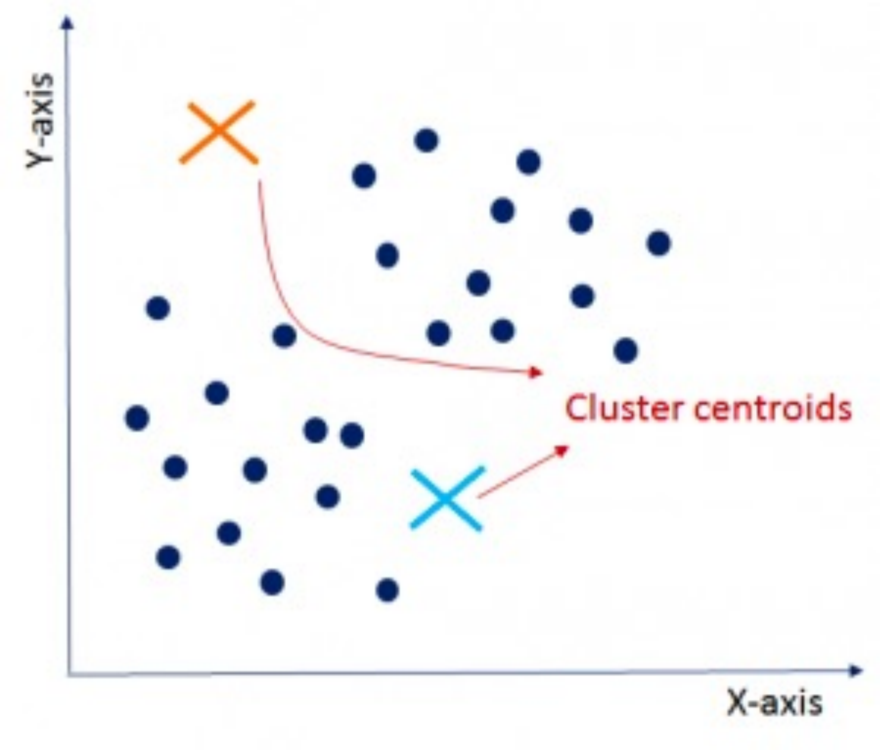
- **k-means** algorithm is an iterative algorithm that tries to partition the dataset into distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**
- It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible
- It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum
- The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster

How does it work?

- Specify number of clusters *K*
- Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing
- Compute the sum of the squared distance between data points and all centroids
- Assign each data point to the closest cluster (centroid)
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster

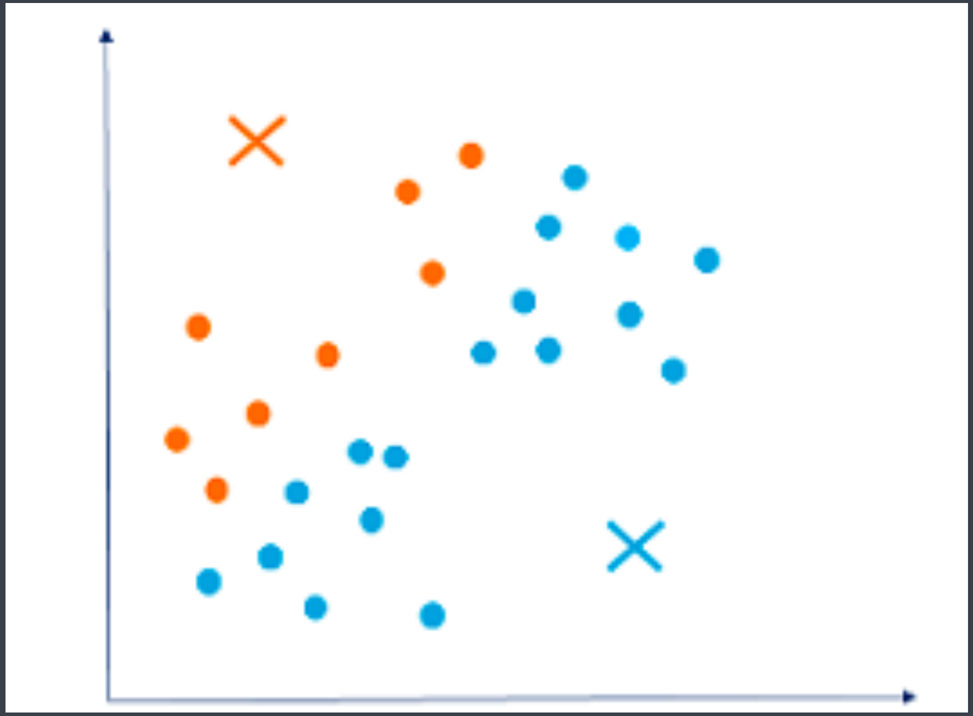
K-Means Clustering - Algorithm

- **Initialization**
  - randomly initialise two points called the cluster centroids



K-Means Clustering - Algorithm

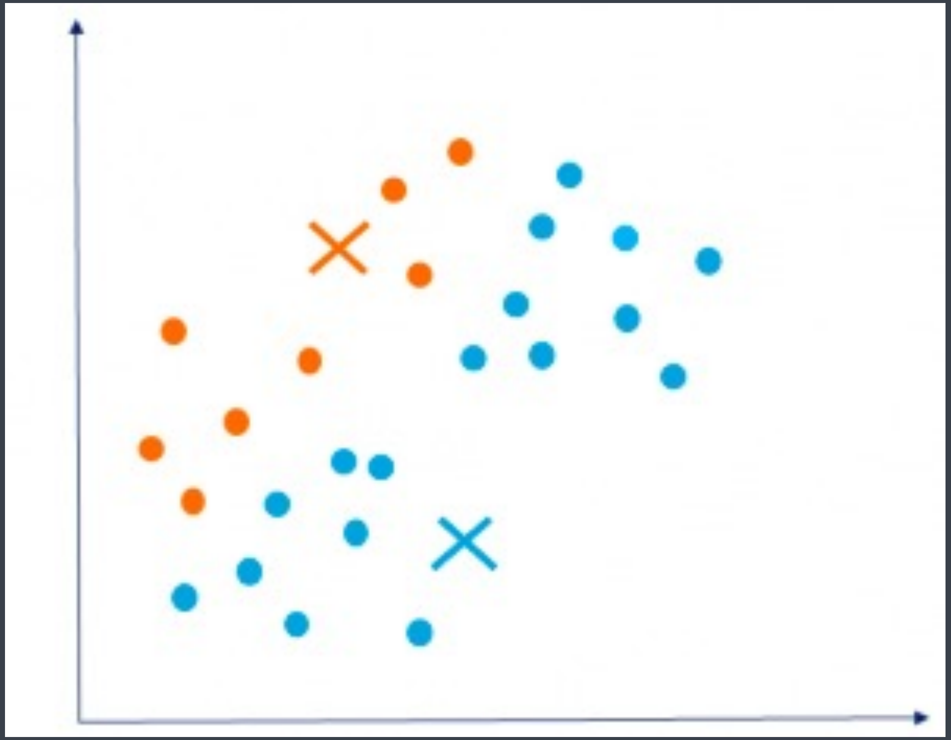
- **Cluster Assignment**
  - Compute the distance between both the points and centroids
  - Depending on the minimum distance from the centroid divide the points into two clusters



K-Means Clustering - Algorithm

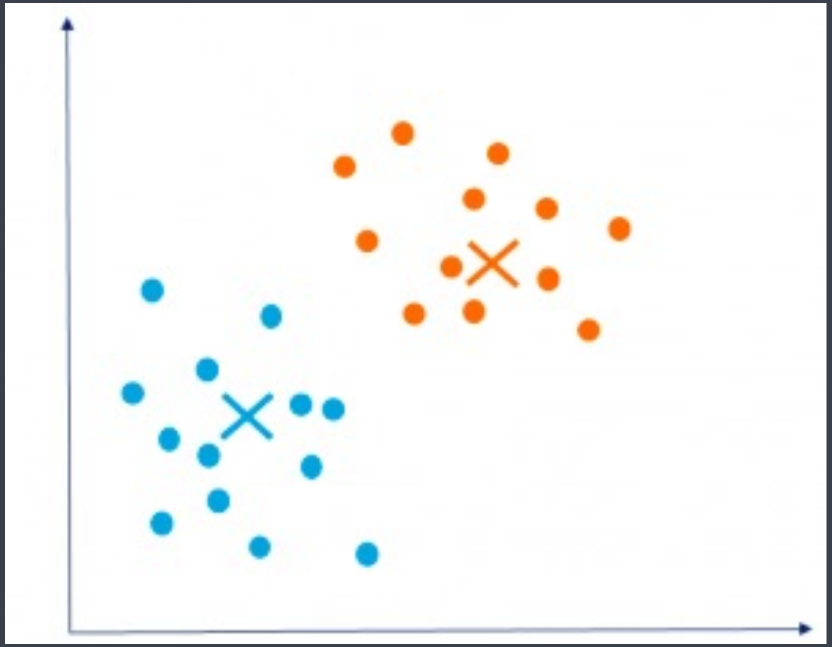
- **Move Centroid**
    - Consider the older centroids are data points
    - Take the older centroid and iteratively reposition them for optimization

- **Optimization**
    - Repeat the steps until the cluster centroids stop changing the position



K-Means Clustering - Algorithm

- **Convergence**
  - Finally, k-means clustering algorithm converges and divides the data points into two clusters clearly visible in multiple clusters



K-Means Clustering - Example

- Suppose we want to group the visitors to a website using just their age (one-dimensional space) as follows:

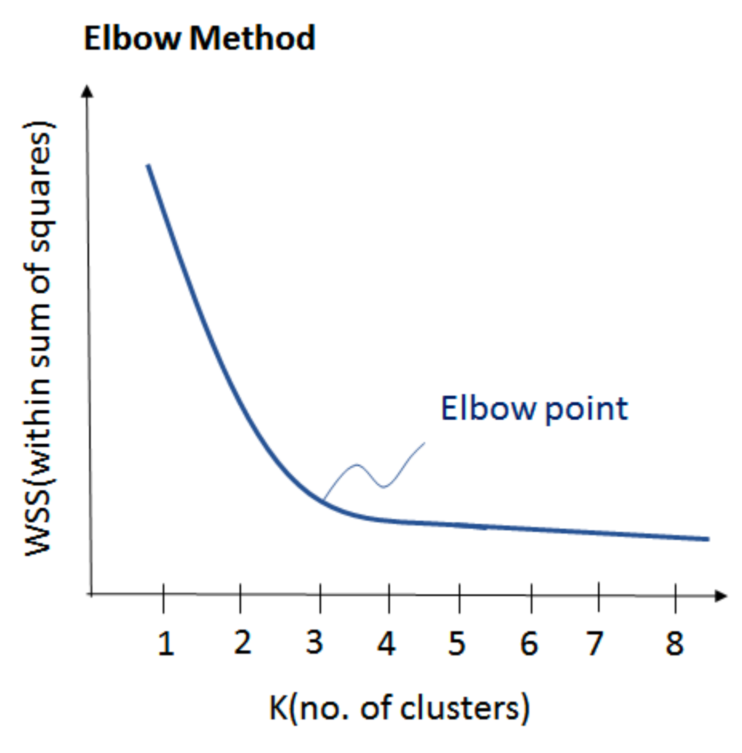
15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65

N = 19

Optimization

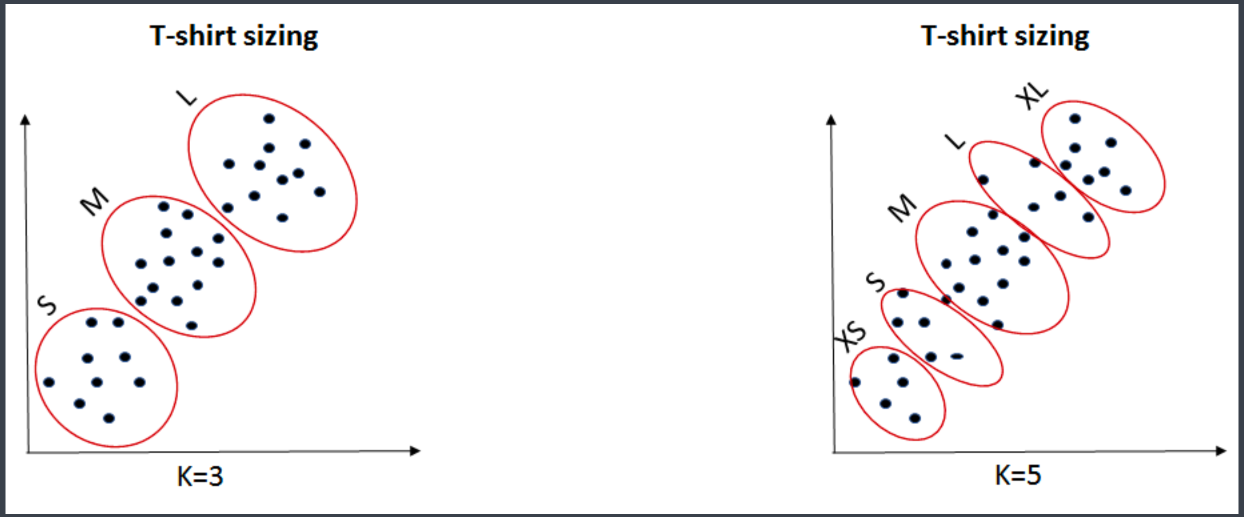
Elbow Method

- Total within-cluster variation
    - Also known as Within Sum of Squares (WSS)
    - The sum of squared distances (Euclidean) between the items and the corresponding centroid
- Draw a curve between WSS (within sum of squares) and the number of clusters
- It is called elbow method because the curve looks like a human arm and the elbow point gives us the optimum number of clusters



Purpose Method

- Get different clusters based on a variety of purposes
- Partition the data on different metrics and see how well it performs for that particular case



- K=3: If you want to provide only 3 sizes(S, M, L) so that prices are cheaper, you will divide the data set into 3 clusters.
- K=5: Now, if you want to provide more comfort and variety to your customers with more sizes (XS, S, M, L, XL), then you will divide the data set into 5 clusters.

Advantages

- It’s straightforward to implement
- It’s scalable to massive datasets and also faster for large datasets
- It adapts to new examples very frequently

Disadvantages

- K-Means clustering is good at capturing the structure of the data if the clusters have a spherical-like shape. It always tries to construct a nice spherical shape around the centroid. This means that the minute the clusters have different geometric shapes, K-Means does a poor job clustering the data.
- Even when the data points belong to the same cluster, K-Means doesn’t allow the data points far from one another, and they share the same cluster
- K-Means algorithm is sensitive to outliers
- As the number of dimensions increases, scalability decreases