Diagnostic Methods

Non Model Diagnostic Methods

K-Means Algorithm

Outline

- Overview
- K-means algorithm
  - Definition
  - K-means steps
  - Advantages/disadvantages
  - K-means in fault diagnosis
    - Medical applications

Overview

- Clustering is an approach used to identify groups of similar objects in datasets with two or more variable quantities .
- Clustering involves automatically discovering natural grouping in data.
- It can be divided into:
  - Partitioning clustering
  - Hierarchical clustering

Overview

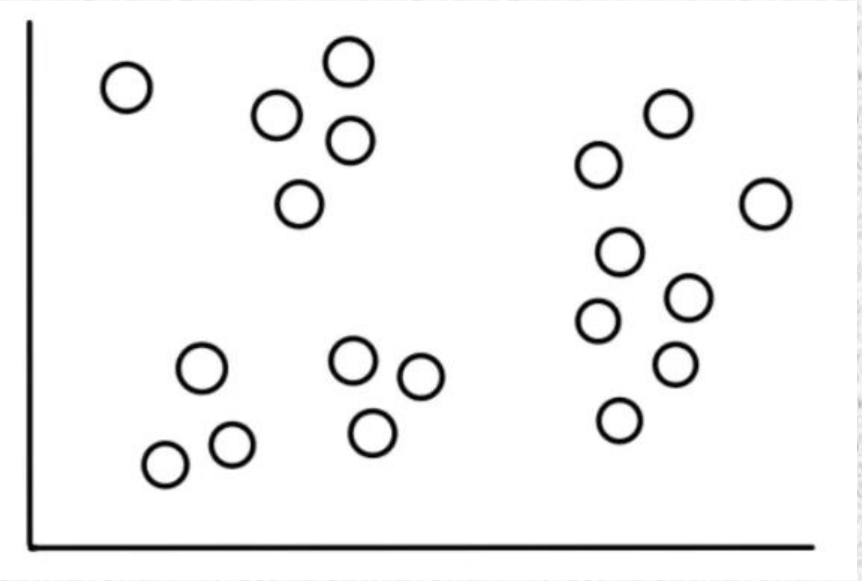
- Partitioning clustering starts with all data points and tries to divide them into a fixed number of clusters such as K-means algorithm.
- Hierarchical clustering does not require any input parameters, it involves creating clusters in a predefined order from top to bottom .

Definitions

- K-Means is a very popular clustering algorithm .
- It is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning.
- K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.
- The term 'K' is a number.

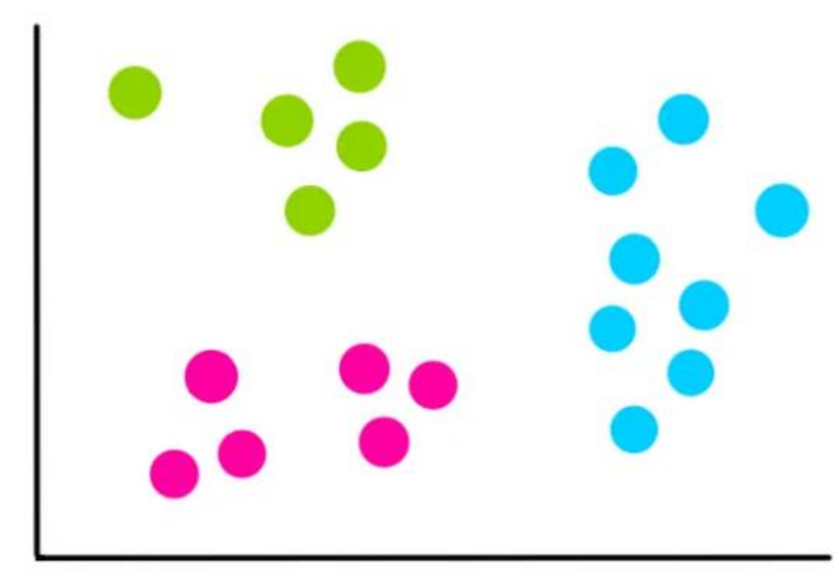
K-means Algorithm

- How does K-means algorithm work?
- Assume that we have 19 data points that look like this:



K-means Algorithm

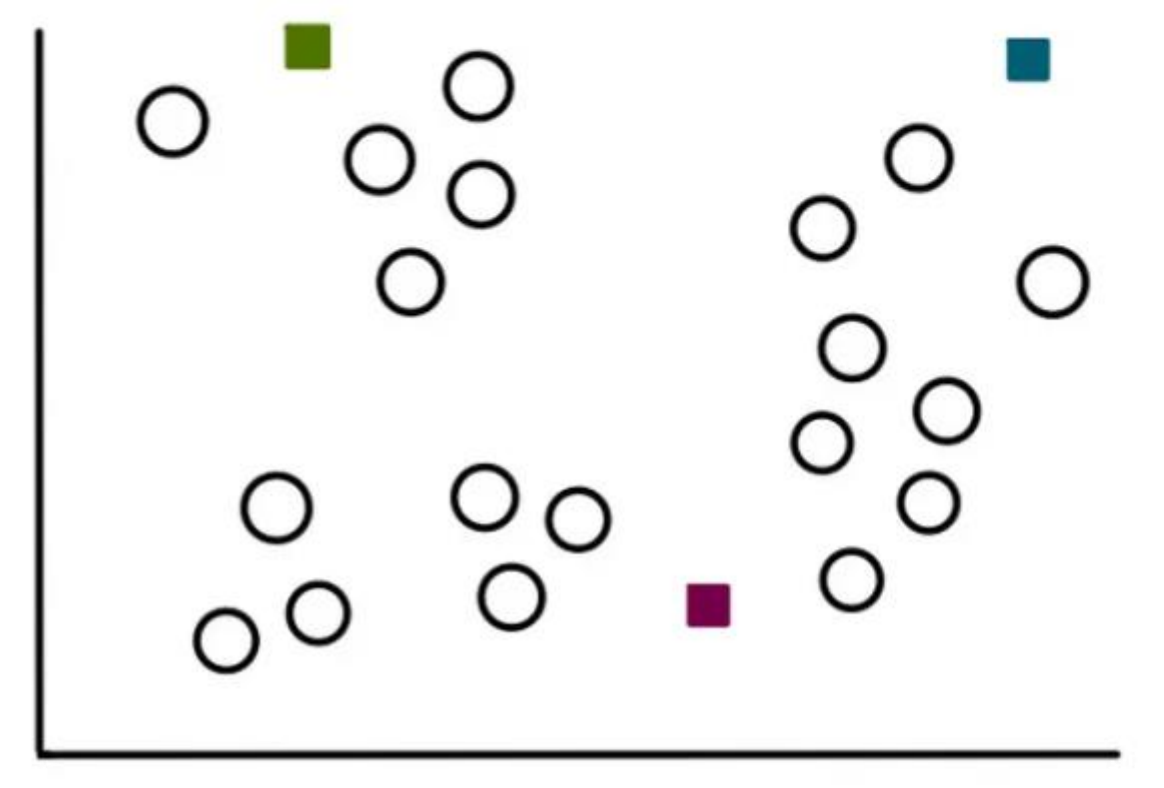
- We choose the number of clusters = 3



K-means Algorithm

- We apply k-means to perform clustering,
- Step 1: Number of Clusters, k
- Step 2: Select k Points at Random
  - We start the process of finding clusters by selecting 3 random points (not necessarily our data points).
  - These points will now act as *centroids,* or the center, of clusters that we are going to make

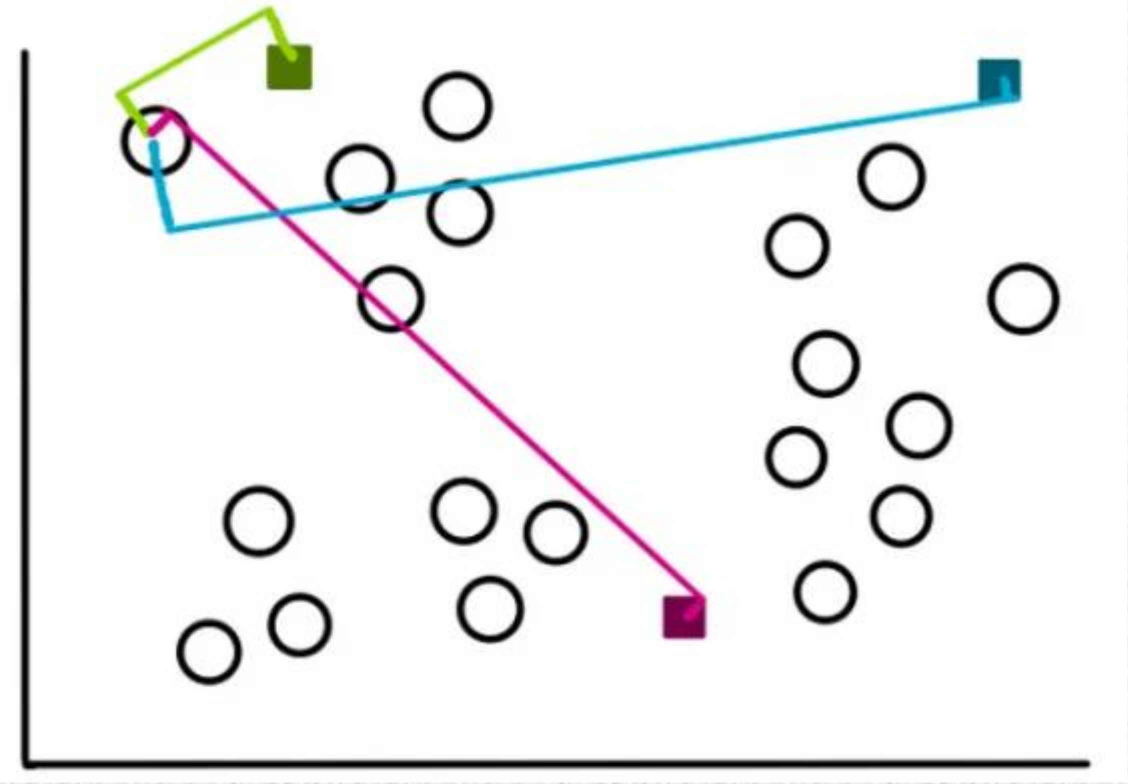
K-means Algorithm



K-means Algorithm

- Step 3: Make k Clusters
  - To make the clusters, we start by measuring the distance from each data point to each of the 3 centroids. And we assign the points to the cluster closest to it. So for a sample point, the distances will look like this:

K-means Algorithm



K-means Algorithm

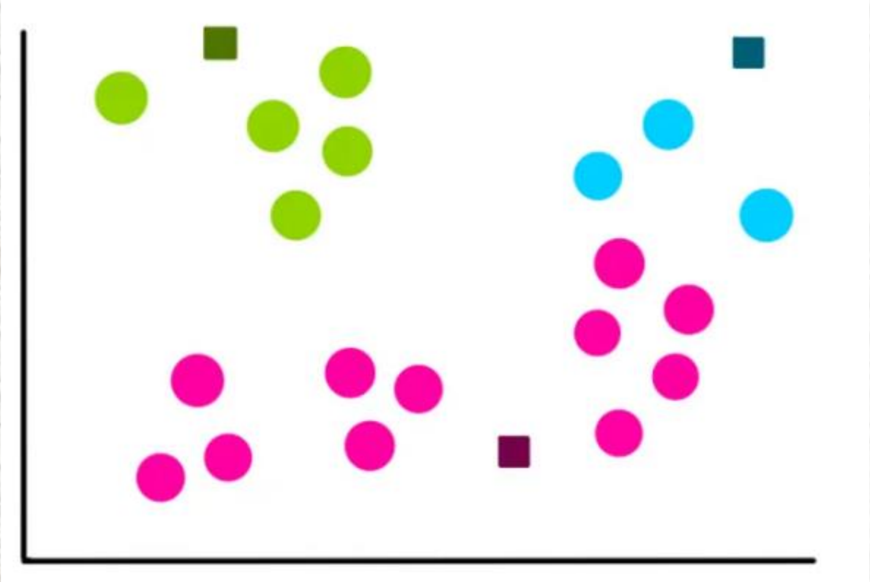
- We see that the distance from the point to the green centroid is the least, so we assign the point to the green cluster,
- Distance measure determines the similarity between two elements and influences the shape of clusters,

K-means Algorithm

- K-means clustering supports various kinds of distance measures, such as:
  - Euclidean distance measure
  - Manhattan distance measure
  - A squared euclidean distance measure
  - Cosine distance measure

K-means Algorithm

- Using one of the previous formulas, we repeat this process for the rest of the points and the clusters will look something like this:



K-means Algorithm

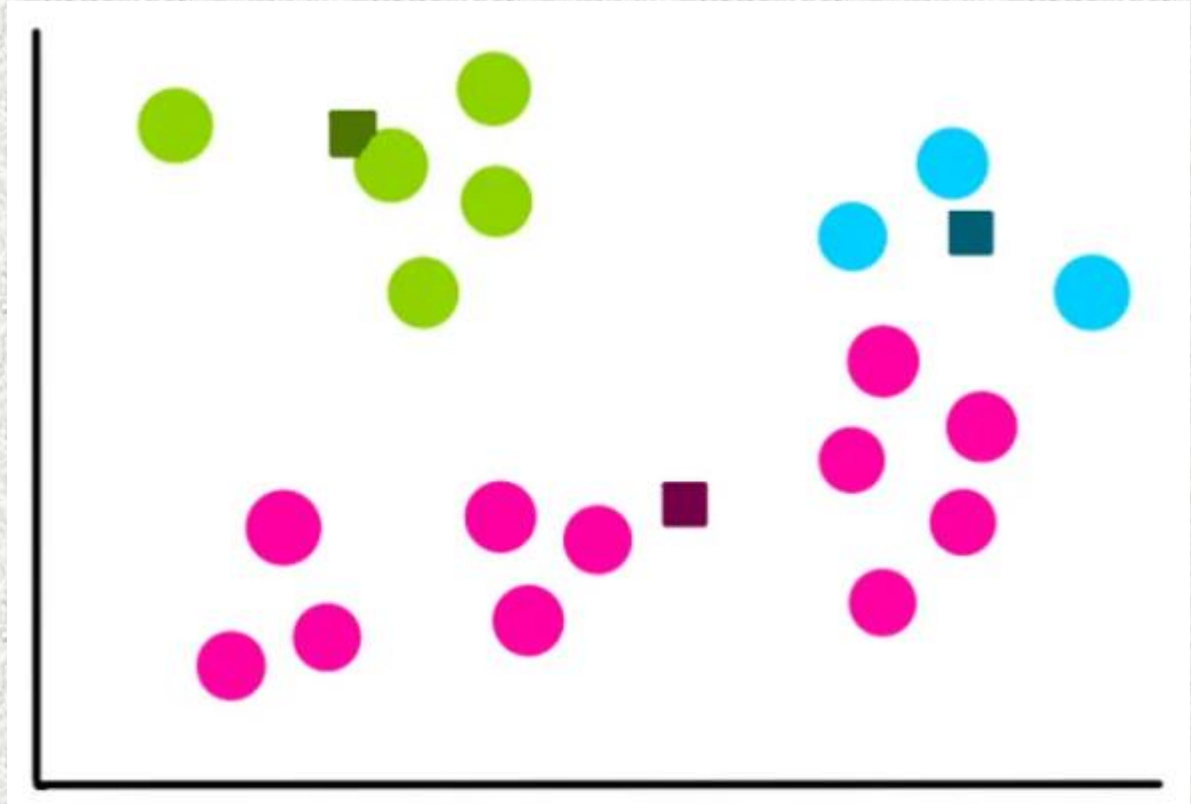
- Step 4: Compute New Centroid of Each Cluster
  - We find the new centroids formed by each of cluster. For example, the way we calculate the coordinates of the centroid of the blue cluster is:

$$(x', y') = \left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3} \right)$$

  - x1, x2, and x3 are the x-coordinates of each of the 3 points of the blue cluster. And y1, y2, and y3 are the y-coordinates of each of the 3 points of the blue cluster.
- The above formula is applied to all clusters

K-means Algorithm

- So, the new centroids look like this:



K-means Algorithm

- Step 5: Assess the Quality of Each Cluster
  - We measure the quality by finding the variation within all the clusters. *The basic idea behind k-means clustering is defining clusters so that the within-cluster variation is minimized.*
- For each value of K, we are calculating Within-Cluster Sum of Square (WCSS).

K-means Algorithm

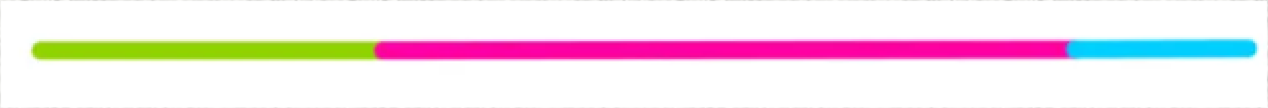
- WCSS is the sum of squared distance between each point and the centroid in a cluster.

$$\mathrm{WCSS} = \sum_{C_k}^{C_n} \Big(\sum_{d_i\, in\, C_i}^{d_m} distance(d_i, C_k)^2 \Big)$$

Where,

C is the cluster centroids and d is the data point in each Cluster.

- let's represent the variation visually like this:



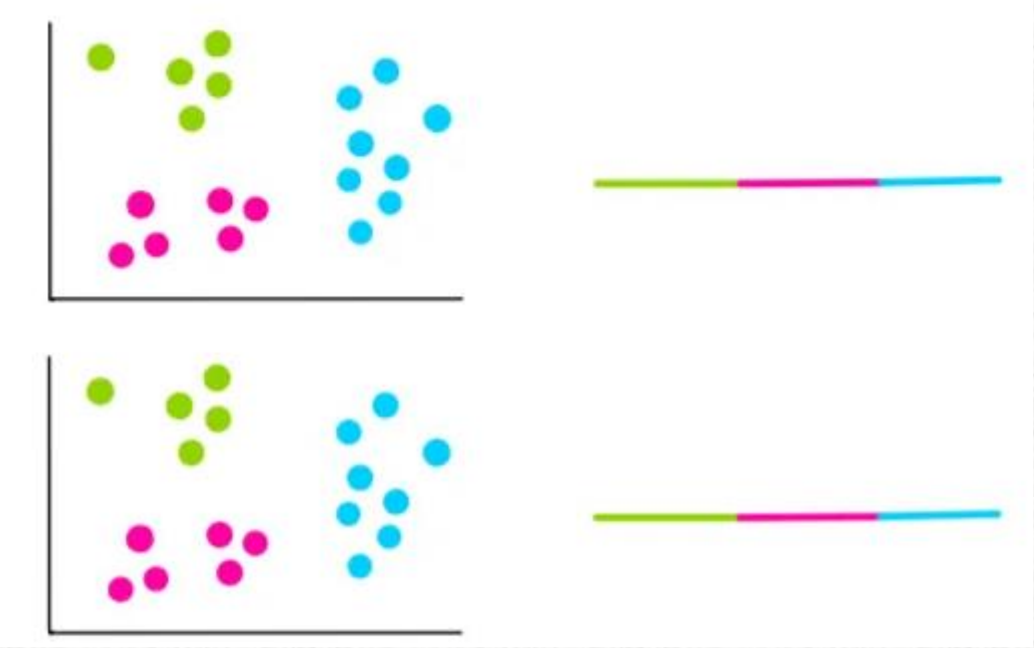
K-means Algorithm

- Step 6: Repeat Steps 3–5
- Once we have previous clusters and the variation stored, we start all over. But only this time we use the centroids we calculated previously to - make 3 new clusters, recalculate the center of the new clusters, and calculate the sum of the variation within all the clusters.
- We stop when the WCSS does not change,

K-means Algorithm

- From the last two iterations, we see that the clusters haven’t changed.
- This means that the algorithm has *converged* and we stop the clustering process.
- We then choose the clusters with the least WCSS

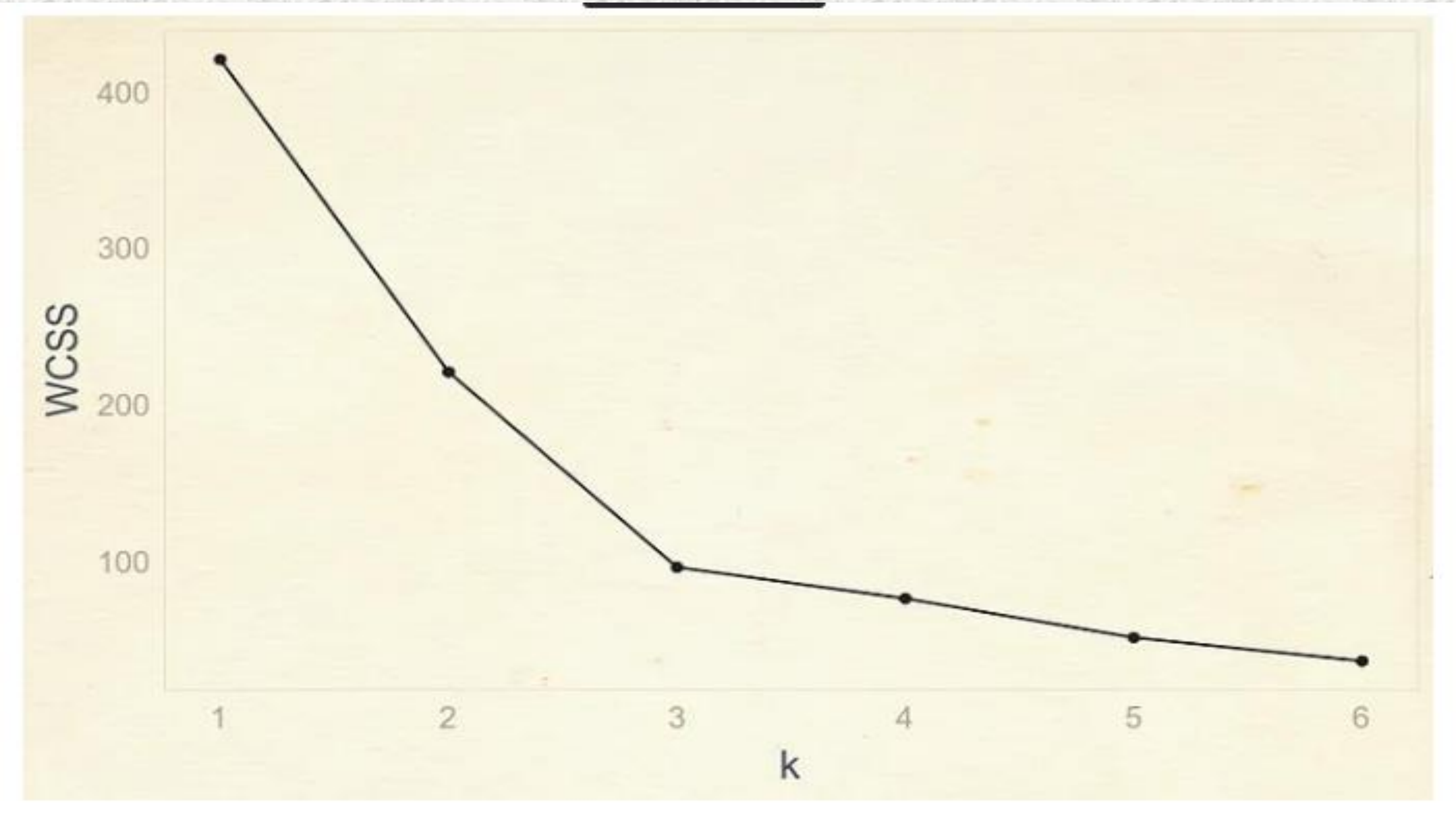
K-means Algorithm



K-means Algorithm

- The selection of k value is a critical issue
- We try multiple k values and calculate the WCSS.
- We notice that each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.
- We need to use something called an *elbow plot* to find the best k. It plots the WCSS against the number of clusters or k.

K-means Algorithm



K-means Algorithm

- This is called an elbow plot because we can find an optimal k value by finding the “elbow” of the plot, which is at 3.
- Until 3 you can notice a huge reduction in variation, but after that, the variation doesn't go down as quickly.

Advantages/ Disadvantages

- Advantage:
  - Simple to implement and use.
  - Scales to large data sets.
  - Guarantees convergence.
  - Easily adapts to new examples.
- Disadvantage:
  - Choosing k manually.
  - Curse of dimensionality
  - Clustering outliers.

K-means in fault diagnosis

- Medical applications:
  - The medical profession uses k-means in creating smarter medical decision support systems, especially in the treatment of liver ailments.
  - Decision Support in Heart Disease Prediction System (HDPS) .

References

- Shreya Rao, Published in Towards Data Science; 2022