K Means Clustering

# Introduction

* Helps to classify or group unlabeled data in an unsupervised learning algorithm.
* Clustering can be used to group documents, market segmentation, etc.
* K means cluster runs by
* randomly assign each point to a number of cluster s (K),
* calculating the distance of each point to the centers (centroid) of the clusters and reassigning it to the closest centroid’s cluster
* repeats the process until there are no new changes in the classification.
* Choosing a k value (number of expected clusters) is difficult but the elbow method could be used.
* The elbow method involves calculating the Sum of Squared Errors (SSE) within groups and plotting K against SSE. The point in the graph where the line stops dropping significantly is probably the best option. Note: Domain knowledge might come in handy as well for this purpose.

# K means clustering with python

* Note that, k means is not predictive, it simply groups.
* You can create an artificial data with make\_blobs imported from sklearn.datsets e.g., df = make\_blobs(n\_samples = a, n\_features = b, centres = c, cluster\_std = d, random\_state = e) where a, b, c, and e are all integers and d is a number. Make sure to assign the simulated data to a variable.
* You can explore the artificial data.
* To run the model:
* From sklearn.cluster import KMeans
* Instantiate the KMeans with number of expected clusters
* Fit with features
* You can then grab a lot of methods and attributes such as cluster centers, labels. Note: if you are working with real unlabeled data, this is the end but with some sort of labeled or artificial data, you can go on to verify its accuracy.