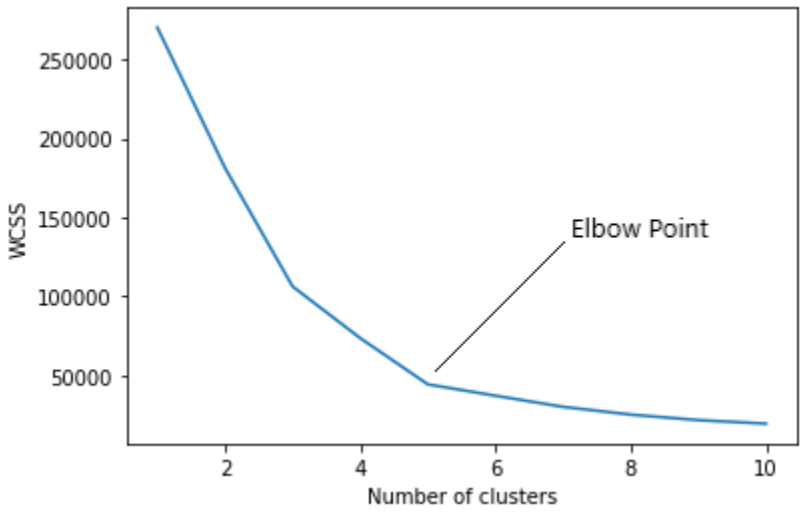
WORKSHOP II SOLUTION - POINT 3

What is the elbow method in clustering?

The Elbow Method is a technique that we use to determine the number of centroids(k) to use in a k-means clustering algorithm. In this method to determine the k-value we continuously iterate for k=1 to k=n (Here n is the hyperparameter that we choose as per our requirement). For every value of k, we calculate the Within-Cluster Sum of Square (WCSS*) value.

Now for determining the best number of clusters(k) we plot a graph of k versus their WCSS value. Surprisingly the graph looks like an elbow (which we will see later). Also, when k=1 the WCSS has the highest value but with increasing k value WCSS value starts to decrease. We choose that value of k from where the graph starts to look like a straight line.

*WCSS - It is defined as the sum of square distances between the centroids and each point.



And which flaws does it pose to assess quality?

The elbow method, while a useful tool for determining the optimal number of clusters in K-means clustering, has some drawbacks:

- Subjectivity: The choice of the "elbow point" can be subjective and might vary between individuals analyzing the same data.
- Non-Gaussian Data: It assumes that clusters are spherical and equally sized, which may not hold for complex datasets with irregularly shaped or differently sized clusters.

- Sensitivity to Initialization: K-means itself is sensitive to initial cluster centroids, which can affect the WCSS values and, consequently, the choice of the optimal K.
- Inefficient for Large Datasets: For large datasets, calculating WCSS for a range of K values can be computationally expensive and time-consuming.
- Unsuitable for All Distributions: The elbow method is not suitable for all data distributions, especially when clusters have varying densities or are non-convex.
- Limited to K-means: It specifically applies to K-means clustering and may not be suitable for other clustering algorithms with different objectives.