Worksheet 5: K-Means++

1. What's the main limitation of Farthest First Traversal?
   - Farthest First Traversal (FFT) is a clustering initialization method where new cluster centers are chosen to be as far as possible from existing centers.
   - Its main limitations are:
     - **Sensitivity to Outliers**: It can place cluster centers far from dense regions if an outlier is chosen, leading to poor clustering.
     - **Ignores Density Variations**: It selects centers based only on distance, potentially leading to poor cluster assignments in datasets with varying densities.
     - **Suboptimal for Non-Spherical Clusters**: It assumes well-separated spherical clusters, performing poorly on elongated or irregularly shaped clusters.
2. What is the difference between K means and K means ++?
   - Both are clustering algorithms, but they differ in **how initial centroids are selected**:
   - **K-Means**: Randomly selects K initial centroids, which can lead to poor clustering and local minima.
   - **K-Means++**: Uses a probabilistic initialization that spreads centroids apart by selecting new centers with a probability proportional to their squared distance from already chosen centers.
     - **Advantage**: Reduces the chances of bad initialization and improves convergence speed.
     - **Result**: More stable and often achieves better clustering than standard K-Means.
3. What are some limitations of Kmeans/ Kmeans++?
   - **Assumes Spherical Clusters**: Performs poorly on clusters that are not well-separated or have complex shapes.
   - **Sensitive to Outliers**: A single outlier can drastically shift a centroid.

- **Fixed Number of Clusters**: Requires manually specifying K, which is often unknown in real-world data.
- **Poor Performance on Unequal Cluster Sizes/Densities**: It struggles when clusters have varying densities or sizes.
- **Computational Complexity (for large data)**: Though K-Means++ improves initialization, K-Means still requires multiple iterations, making it expensive for very large datasets.

4. Explain why we need silhouette scores

- The **Silhouette Score** is a metric used to **evaluate the quality of clustering** by measuring how well each point fits within its assigned cluster compared to other clusters.
- It is needed because:
  - **No Ground Truth in Clustering**: Since clustering is unsupervised, silhouette scores provide an objective way to measure clustering effectiveness.
  - **Evaluates Separation & Cohesion**: A high score (close to 1) indicates well-separated and compact clusters, while a low score (close to 0 or negative) suggests poor clustering.
  - **Helps Determine Optimal K**: By comparing silhouette scores for different values of K, we can estimate the best number of clusters.