Viva Questions for K-Means Clustering

1. **What is K-Means clustering, and how does it work?**

   Answer: K-Means clustering is an unsupervised learning algorithm used to partition a dataset into `k` distinct clusters. The algorithm works by:

   - Initializing `k` cluster centroids randomly.
   - Assigning each data point to the nearest centroid.
   - Recomputing the centroids based on the mean of the points assigned to each cluster.
   - Repeating the assignment and centroid update steps until convergence or a maximum number of iterations is reached.

2. **What does the `n_clusters` parameter specify in K-Means?**

   Answer: The `n_clusters` parameter specifies the number of clusters to form. The algorithm will partition the data into this number of clusters.

3. **What is the purpose of the `init` parameter in K-Means, and what does "k-means++" do?**

   Answer: The `init` parameter specifies the method for initializing the cluster centroids. "k-means++" is a strategy that selects initial centroids in a way that spreads them out to improve convergence and reduce the risk of ending up in a local minimum.

4. **How does the `max_iter` parameter affect the K-Means algorithm?**

   Answer: The `max_iter` parameter specifies the maximum number of iterations the algorithm will perform during a single run. This limits the number of times the algorithm updates centroids and assigns points, helping to prevent excessive computation.

5. **What role does the `random_state` parameter play in K-Means clustering?**

   Answer: The `random_state` parameter controls the randomness of the centroid initialization. Setting a fixed value ensures that the results are reproducible, as the same initial centroids will be used each time the algorithm is run.

6. **How do you interpret the cluster centers in K-Means clustering?**

   Answer: Cluster centers (or centroids) are the mean positions of the data points assigned to each cluster. They represent the central point of each cluster in the feature space and are used to define the clusters.

7. **What is the purpose of visualizing the clusters and centroids?**

   Answer: Visualizing the clusters and centroids helps to:

   - Understand how the data is partitioned into clusters.

   - See the distribution of data points within each cluster.
   - Identify the locations of the cluster centroids.
   - Validate if the clusters make sense based on the data distribution.

8. **What can you infer from the scatter plot of clusters and centroids?**

   Answer: The scatter plot shows the data points colored by their cluster assignment and the centroids marked distinctly. It helps to:

   - Observe how well-separated the clusters are.
   - Determine if the clusters overlap or if any points are far from their centroids.
   - Evaluate the spread and density of the data points within each cluster.

9. **Why might you use the `fit_predict` method instead of `fit` and `predict` separately?**

   Answer: The `fit_predict` method combines the fitting of the model and the prediction of cluster labels into a single step, which is more efficient and convenient for obtaining the cluster assignments immediately after fitting the model.

10. **How do you determine the optimal number of clusters for K-Means?**

    Answer: The optimal number of clusters can be determined using methods such as:

    - **Elbow Method:** Plotting the within-cluster sum of squares (WCSS) against the number of clusters and looking for the "elbow" point where the rate of decrease slows.
    - **Silhouette Score:** Measuring how similar an object is to its own cluster compared to other clusters.

11. **What are some limitations of the K-Means algorithm?**

    Answer: Some limitations include:

    - **Choosing the number of clusters:** K-Means requires specifying the number of clusters beforehand.
    - **Sensitivity to initial centroids:** The results can vary depending on the initial placement of centroids.
    - **Assumption of spherical clusters:** K-Means assumes clusters are spherical and equally sized, which may not always be true for real-world data.
    - **Sensitivity to outliers:** Outliers can significantly affect the position of centroids and the clustering results.