1. What is cluster analysis and its types?

* Cluster analysis is an EDA which tries to identify structures within the data.
* Cluster analysis tries to identify homogenous groups of cases if grouping is not previously known.
* Different types of clustering analysis

1. Hierarchical clustering

* This groups similar objects in clusters.
* The endpoint is set of clusters where each cluster is distinct from other clusters and objects in each cluster are broadly similar to each other.
* Main output of hierarchical clustering is dendrogram.
* Euclidean distance can be used as parameter to decide clusters.

1. Partition clustering

* Partition clustering is just division of data points based on similarity in non-overlapping clusters.
* K means is a type of partition clustering

1. Exclusive clustering

* Exclusive clustering is a clustering type in which each data point can exist in on cluster only.

1. Overlapping clustering

* This is a contradiction of exclusive clustering
* A datapoint can exist in more than one clusters in this case.
* A data point possessing qualities of more than one clusters can be divided successfully on this basis.

1. Fuzzy clustering

* In **fuzzy Clustering**, each object belongs to each cluster with a membership weight that is between 0 and 1.
* We usually set the additional constraint, and the sum of weights for each object must be equal to 1.

1. Complete clustering

* **Complete clustering** allocates each object to a cluster.

1. How is cluster quality measured?

* Cluster quality depends on how well the clusters are separated and how compact they are.
* Silhouette Coefficient is a measure of calculating the cluster quality
* For a data set, D, of n objects, suppose D is partitioned into k clusters, C1, …, Ck. For each object **o** ∈ D, we calculate a (**o**) as the average distance between **o** and all other objects in the cluster to which **o** belongs. Similarly, b(**o**) is the minimum average distance from **o** to all clusters to which **o** does not belong. Formally, suppose **o** ∈Ci (1 ≤ i ≤ k); then Silhouette Coefficient is given by



* Value of silhouette coefficient is between -1 an 1..
* The value of **a(o)** reflects the compactness of the cluster to which ***o*** belongs. The smaller the value, the more compact the cluster. The value of b(***o***) captures the degree to which ***o*** is separated from other clusters. The larger b(***o***) is, the more separated ***o*** is from other clusters.

1. How is cluster analysis calculated

* The process of doing cluster analysis includes the following steps
* Perform the descriptive analysis.
* In k means clustering we need to specify the number of clusters that need to be formed.
* We can tweak the number of clusters to get a better performing model
* The methods used to get the optimal number of clusters:
* **The elbow method:**

the sum of squares at each number of clusters is calculated and graphed

a change of slope from steep to shallow is observed to get the optimal number of clusters

* **The Gap Statistic:**

The [gap statistic](http://www.web.stanford.edu/~hastie/Papers/gap.pdf) compares the total within intra-cluster variation for different values of **k** with their expected values under null reference distribution of the data.

* **The silhouette method:**

Average silhouette method computes the average silhouette of observations for different values of k

The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for **k**

* **The sum of squares method:**

This method involves choosing the optimal number of clusters by minimizing the within cluster sum of squares and maximizing the between-cluster sum of squares