**Q1. Compare and Contrast the K-Means Clustering and Hierarchical Clustering.**

Clustering is an unsupervised learning technique, which creates different groups or clusters of the given set o f inputs and is also able to put a new input into the appropriate cluster. While doing clustering, the basic objective is to group input points in such a way as to maximize the inter-cluster variance and minimize the intra-cluster variance.

**K-Means** **Clustering -**

1. K-Means starts with identified initial value of ‘K’ and then identifying k-observations as centroids.
2. Then the iterative process starts
   1. All the data points are assigned to the nearest centroid.
   2. The new centroids for each group are calculated.

The process goes on till the centroids converge i.e. there is no change in centroid values.

**Hierarchical Clustering –**

As the name suggests, it builds the hierarchy of clusters. This can be done using 2 approaches

1. Agglomerative – Every data point is its own cluster. We then start merging the clusters based on linkage between the clusters. This goes on till we have 1 single cluster.
2. Divisive – In this approach, all the data points are initially in the same cluster. We then start splitting it till each data point is in its own cluster.

The result of Hierarchical clustering is shown using dendrogram. Number of clusters is decided by interpreting the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

**Key Differences between K-Means and Hierarchical clustering**

1. Hierarchical clustering is not efficient with large amounts of data. Linkage calculation is a O(n2) operation.

K-means on the other hand is a linear operation in each iteration and hence more efficient on large amount of data.

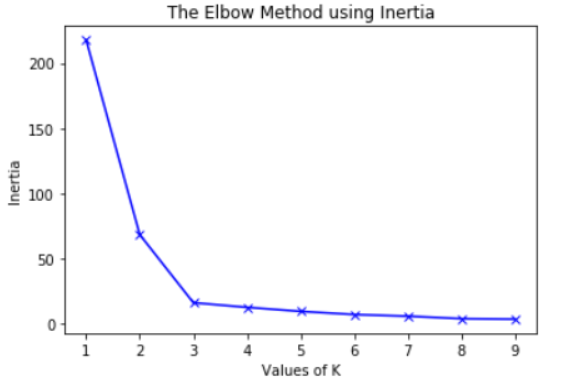
1. K-means starts with random choice of initial centroids so different iterations can produce different results. On the other hand, the results of Hierarchical clustering can be reproduced across multiple runs of algorithm.
2. K-Means requires prior knowledge of K i.e. the number of clusters you want. Elbow curve method or silhouette score can be used to get a starting value of ‘K’ but in hierarchical, dendrogram can be used to decide the value of K.

**Q2. Explain the steps involved in K-Means Algorithm.**

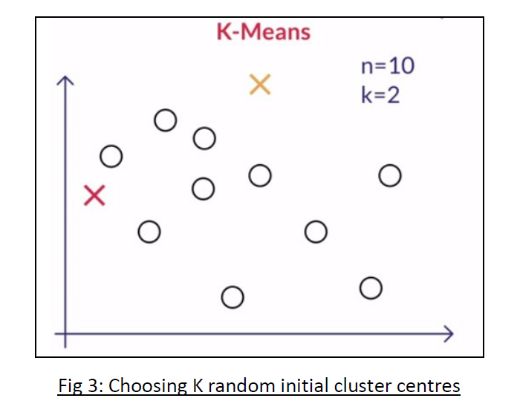
A2. K-Means is a clustering algorithm. Main steps involved in K-Means algorithm are as follows.

1. Specify the number of Clusters – ‘K’.
2. Initialize the ‘K’ centroid points by randomly selecting datapoints.
3. Assign all the data points to 1 centroid based on Euclidean distance or other metric as decided.
4. Re-compute the centroids for each group again.
5. Iterate step 3, 4 till the centroid values do not change.
6. **Specify the number of Clusters- K**

The first requirement is to determine the number of clusters the data has to split into. 2 most common methods to identify a suitable ‘k’ are:

* 1. **Elbow curve**- We plot the **Inertia** - sum of square distances of samples to their closest cluster center, for each value of K in ascending order. The Point where the curve smoothens is likely to be the optimal k. In image below – k=3 looks a decent one to start with.
  2. **Silhouette Score -** The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

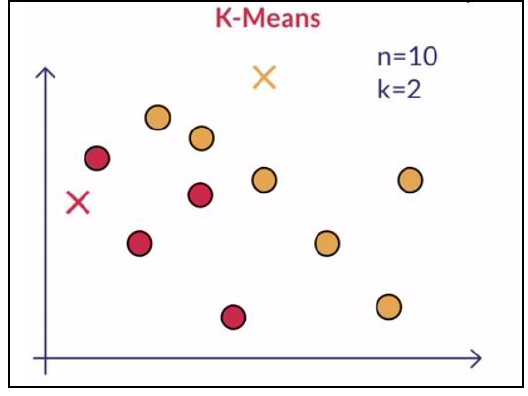
a(i) – average distance from points in own cluster. b(i) – average distance from points in another cluster. i– ith data point.

We take the ‘K’ with highest average silhouette score.

1. **Initialize the ‘K’ centroid points by randomly selecting datapoints.**

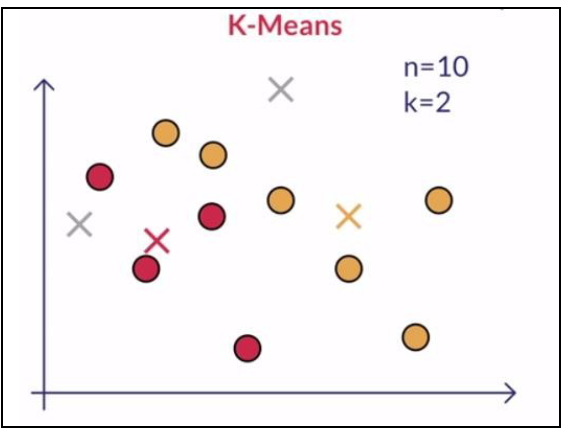
We can choose the initial ‘K’ centroids randomly. But this choice of centroids can impact the result and there can be different results for different runs of the algorithm.

We can also use K-means++ algorithm to initialize the centroids.



1. **Assign all the data points to 1 centroid based on Euclidean distance or other metric as decided.**

For each given point we calculate Euclidean distance between the point and all the centroids calculated earlier. Euclidean distance between 2 points p, q is given by.

1. **Re-compute the centroids for each group again.**

Centroid is calculated by taking mean of each dimension of the

points in given cluster.

1. **Iterate step 3, 4 till the value of centroids do not change**.

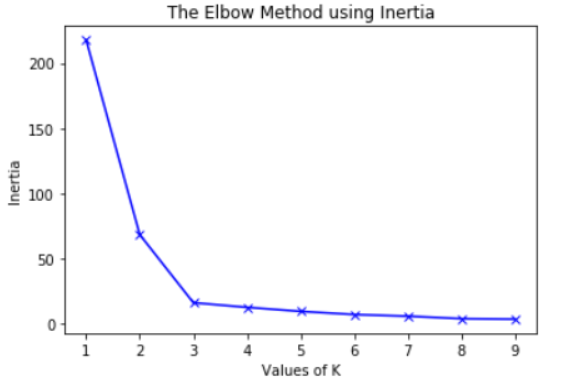
At this point we have our required clusters. We can analyze the various parameters of these clusters and make decisions regarding business significance accordingly.

**Q3. How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.**

**A3.** First Requirement in K-means Algorithm is choosing the value of K i.e. the number of clusters to be formed in the data. Clusters should follow following properties –

1. Cohesion of data points within the cluster – all the data points inside a cluster should be similar to each other.
2. Coupling of data points across clusters – The data points in 2 clusters should not be very similar to each other.

For example, in a retail store customer segmentation problem, the customers in each cluster should have similar buying habits or shown interest in a category of products, have a similar buying timeline etc. However, customers in different clusters should be inherently different in their characteristics.

****Initial value of ‘K’ can be chosen using following methods

1. **Elbow curve**- We plot the **Inertia**.

Inertia calculates the sum of square distances of samples to their closest cluster center for each cluster. Total inertia is the sum of all these values calculated above. So, Inertia is the sum of all intra-cluster distances. Hence, value of inertia should be as much minimum as possible.

We plot the values of inertia for each value of ‘K’.

The Point where the curve smoothens is likely to be the optimal k. This is because further granular grouping/clustering does not reduce the inertia value significantly.

In image below – k=3 looks a decent one to start with.

1. **Silhouette Score -** The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

a(i) – average distance from points in own cluster. b(i) – average distance from points in another cluster. i– ith data point.

We take the ‘K’ with highest average silhouette score.

Choice of K will have impact on the business decisions.

**Q4.  Explain the necessity for scaling/standardization before performing Clustering.**

**A4.** Standardization of the data means converting them into values such that mean of resulting values is 0 and SD = 1.

Formula for standardization is - , is the mean and is the standard deviation.

Standardization does not impact the outliers in the data. It is necessary to perform the standardization in before clustering because of 2 main reasons.

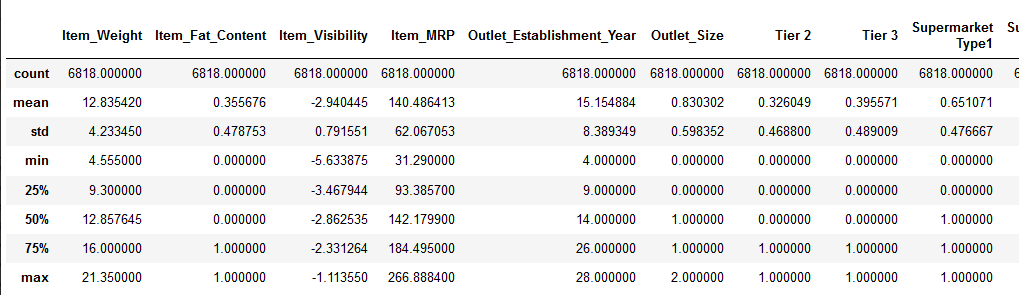
1. We use Euclidean distance between the data points.

As it is evident from the equation for the Euclidean distance, the magnitude and units of the various data points will impact the result.

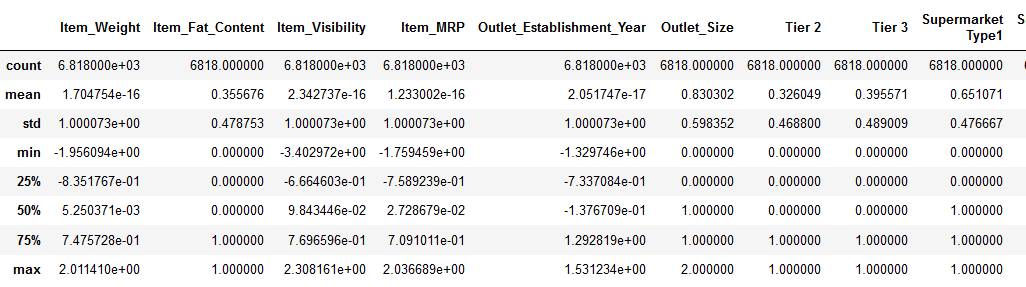
The attributes with larger range of values will weigh out the attributes with small range. Therefore, scaling down is required to avoid this.

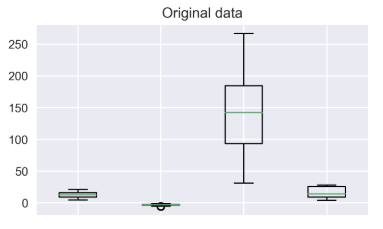
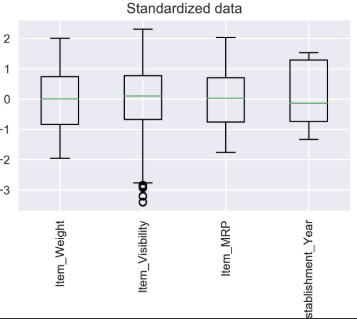
1. The standardization also helps in making the attributes unit free and uniform.

Example: Retail store data.

Original statistics:

Standardized data statistics:



Let us look at some box plots. This helps us visualize how standardization can help in data reading and analyzing the data.

**Q5. Explain the different linkages used in Hierarchical Clustering.**

**A5.** Different types of linkages in Hierarchical Clustering.

1. Single Linkage:

Distance between 2 clusters is defined as shortest distance between any 2 points in the two clusters.

for all points ‘i’ in cluster u and ‘j’ in cluster v. This approach is also known as Nearest Point Algorithm.

1. Complete Linkage:

Distance between 2 clusters is defined as maximum distance between any 2 points of the 2 clusters.

for all points ‘i’ in cluster u and ‘j’ in cluster v. This approach is also known as Farthest Point Algorithm or Vor Hees Algorithm.

1. Average Linkage:

Distance between 2 clusters is defined as average distance between every point of one cluster to every other point of another cluster.

for all points i, j in both clusters. |u|, |v| are cardinalities of clusters u, v respectively.

1. Weighted Linkage: Also known as Weighted Pair Group Method with Arithmetic Mean is used in Agglomerative hierarchical clustering method.

At each step, the nearest clusters ‘i’, ‘j’ are merged to form . The distance of new cluster to another cluster k is given by following equation:

1. Centroid Linkage:

Distance between the 2 clusters is defined as the distance between centroids of the 2 clusters.

cu, cv are the centroids of the clusters u, v.

When the clusters are merged or a new element is added to the cluster, the new centroids are calculated with all the elements currently in the cluster. In case of merging, distance becomes the distance between centroid of merged cluster and a centroid of a remaining cluster ‘v’ in the forest. This can some time lead to inversion in dendrogram, as smaller clusters may be more similar to larger cluster than to their individual clusters.

1. Median Linkage: When 2 clusters say – ‘s’, ‘t’ are merged to form a new cluster ‘u’ average of centroids of ‘s’, ‘t’ give the centroid of new cluster ‘u’. Rest calculations are done as in centroid linkage.
2. Ward Linkage: This uses ward variance minimization algorithm. This is used in agglomerative hierarchical clustering. At each step in the algorithm, we merge the clusters such that it leads to minimum increase in the within-cluster variance. Initial distance is set as squared Euclidean distance between the points.