a) Compare and contrast K-means Clustering and Hierarchical Clustering.

1. If there is a specific number of clusters in the dataset, but the group they belong to is unknown, choose K-means
2. If the distinguishes are based on prior beliefs, hierarchical clustering should be used to know the number of clusters
3. With a large number of variables, K-means compute faster
4. The result of K-means is unstructured, but that of hierarchal is more interpretable and informative
5. It is easier to determine the number of clusters by hierarchical clustering's dendrogram

b) Briefly explain the steps of the K-means clustering algorithm.

The algorithm works as follows:

1. First we initialize k points, called means, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters. Or till the mean became constant and not changing with iterations.

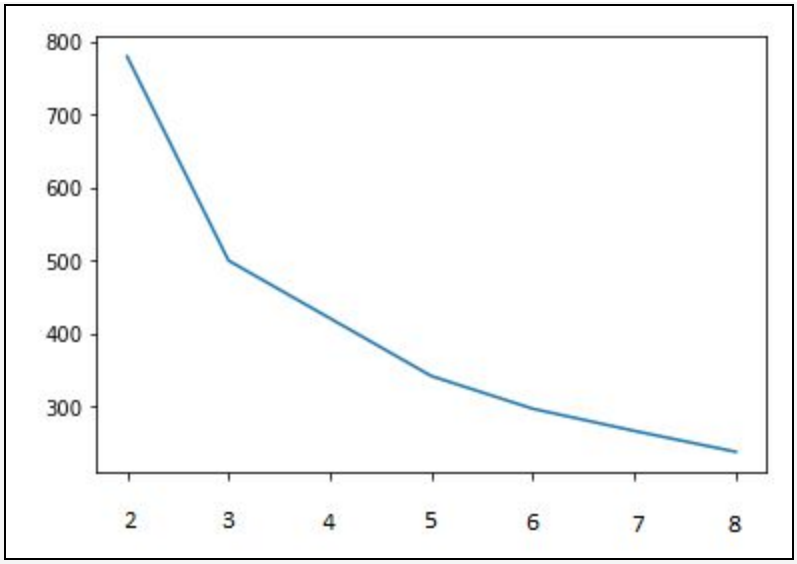
The "points" mentioned above are called means, because they hold the mean values of the items categorized in it. To initialize these means, we have a lot of options. An intuitive method is to initialize the means at random items in the data set. Another method is to initialize the means at random values between the boundaries of the data set (if for a feature x the items have values in [0,3], we will initialize the means with values for x at [0,3]).

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

To identify k we have two methods : SSD elbow and Silhoutte Score

The Elbow Method
Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.
How to get Within-Cluster-Sum of Squared Errors:

- The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
- The WSS score is the sum of these Squared Errors for all the points.
- Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.



The plot looks like an arm with a clear elbow at k = 3
Unfortunately, we do not always have such clearly clustered data. This means that the elbow may not be clear and sharp.

The Silhouette Method

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

The Silhouette Value s(i) for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Here, a(i) is the measure of similarity of the point i to its own cluster. It is measured as the average distance of i from other points in the cluster.
Similarly, b(i) is the measure of dissimilarity of i from points in other clusters.
d(i, j) is the distance between points i and j. Generally, Euclidean Distance is used as the distance metric.

d) Explain the necessity for scaling/standardisation before performing Clustering.

If we have two features, one where the differences between cases is large and the other small, are we prepared to have the former as almost the only driver of distance?
So for example if we clustered people on their weights in kilograms and heights in metres, is a 1kg difference as significant as a 1m difference in height? Does it matter that we would get different clusterings on weights in kilograms and heights in centimetres? If our answers are "no" and "yes" respectively then we should probably scale.
On the other hand, if we were clustering Canadian cities based on distances east/west and distances north/south then, although there will typically be much bigger differences east/west, we may be happy just to use unscaled distances in either kilometres or miles (though we might want to adjust degrees of longitude and latitude for the curvature of the earth).

e) Explain the different linkages used in Hierarchical Clustering.

Single-Linkage
Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage
Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

Average-Linkage
Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

Centroid-Linkage
Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn’t arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.