Question 2: Clustering

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

k-means Clustering	**Hierarchical Clustering**
k-means, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance.	Hierarchical methods can be either divisive or agglomerative.
K Means clustering needed advance knowledge of K i.e. no. of clusters one want to divide your data.	In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.
In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible in Hierarchical clustering

K- means clustering a simply a division of the set of data objects into non- overlapping subsets (clusters) such that each data object is in exactly one subset).	A hierarchical clustering is a set of nested clusters that are arranged as a tree.
K Means clustering is found to work well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical clustering don't work as well as, k means when the shape of the clusters is hyper spherical.

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

At a high level, a simplistic interpretation is explained:

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 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.

Using the elbow method **:** The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point**.**
Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.

Unfortunately, there is no definitive answer to this question. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning.

Use cases for k-means clustering in business

1. Customer segmentation
2. Retail delivery optimisation
3. Document sorting & grouping
4. Customer retention
5. Discount analysis

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

Standardization is an important step of Data preprocessing.
it controls the variability of the dataset, it converts data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

It isn't strictly necessary to standardise, whether it is required or not may depend on the distance metric you choose

e) Explain the different linkages used in Hierarchical Clustering.

A hierarchical clustering is often represented as a dendrogram.

Single Linkage

In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.

Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points

Average Linkage

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other.