Report

***Q1.a:* What method should we use to solve this problem and why?**

***A*: The method we can use is the k-means clustering algorithm. To determine the optimal number of clusters, one can use the Elbow Method.**

Because this involves running k-means clustering on the dataset for a range of values of k. and then for each value of k compute the sum of squared distances from each point to its assigned center.

***Q1.b:* Are these data affected by the curse of dimensionality? Explain.**

A:* Not really. *Because the curse of dimensionality refers to the exponential increase in volume associated with adding extra dimensions to a mathematical space. It can cause various issues, such as: Distance metrics start losing meaning. Increased computational cost.

***Q2:* Implement two alternative solutions to Q1 (c), one of which uses DBSCAN. compare and report Discover.**

***A*: In my code, I'm using the "Agglomerative Clustering". Because this method is repeated until there's only one single cluster left, containing all data points. I decided to cut the**

dendrogram to obtain 5 clusters, it would intersect 5 vertical lines, corresponding to the 5 clusters I've chosen.

The silhouette score measures how similar an object is to its cluster compared to other clusters, with values ranging from -1 to 1. A high silhouette score indicates well-defined clusters.

Q3:** **Evaluate the quality of the groupings you report as solutions to Q1 (c) and Q2. based on the evaluation results, report the best solution and interpret the results.

***A:	For the most suitable clustering solution for your dataset, it's essential to consider both the quantitative metrics and qualitative insights.***

Quantitative Metrics: The silhouette score computed for Agglomerative Clustering gives a measure of how well-defined the clusters are.

A score closer to 1 suggests that the clusters are well apart from each other and clearly distinguished.

The optimal clustering solution would be the one that offers a balance between a high silhouette score, meaningful and interpretable clusters.

Q4:** **Create two 2D plots to display relationships between these new variables and explain the plot.

***A:** PCA (Principal Component Analysis) Plot: Look for any visible patterns, clusters, or groupings. The direction and spread of the data points can give insights into the primary sources of variance in the dataset.*

T-SNE (T-distributed neighbor embedding) Plot: t-SNE is known for its ability to reveal clusters or groupings in the data. If there are distinct groupings or clusters, it suggests that there are inherent structures or subgroups within your data.

***Q5:** Will the conversion done in Q4 result in loss of information? explain your answer with evidence.*

***A:** For t-SNE doesn't provide a straightforward way to measure the amount of variance it retains, as it's a non-linear method focused on preserving local structures. Instead, t-SNE is best evaluated by visual inspection of the resulting plots and the structures they reveal. To conclude, by checking the explained variance in PCA, you can determine the loss of information due to the transformation.*