HW Assignment-4

# Question 1

We can detect outliers after hierarchical clustering with the following approaches:

* Distance-based analysis:
  + We calculate the distance between every instance and its nearest neighbor.
  + Instance farther away from all centers considering a threshold that instance can be considered an outlier.
* Visual Inspection:
  + Visualize the dendrogram.
  + Look for instances far away from other clusters or do not fit in well with any cluster.
* Silhouette analysis:
  + Silhouette plot measures how close each point is to those in the neighboring cluster.
  + Silhouette score near +1 indicates a point far from the neighboring cluster.
  + Silhouette score of 0 indicates that point is close to the decision boundary of neighboring clusters and may be wrongly labeled. These points may or may not be outliers.

# Question 2

* Median over Mean:
  + The mean is sensitive to outliers because it considers the magnitude of each observation, whereas the median only considers the order of the values.
  + In the presence of outliers, the mean can be significantly influenced by their extreme values, causing it to deviate from the actual central tendency of the data.
  + Median is less affected by outliers because it only considers the value in the middle of the distribution, regardless of their magnitude.
* Minimizing absolute error over squared errors
  + Squared error gives greater weight to significant errors, whereas the absolute error gives equal weight to all errors.
  + In the presence of outliers, the squared error can be heavily influenced by their large deviations, causing it to prioritize fitting the outliers at the expense of most of the data.
  + Absolute error is less affected by outliers because it treats all errors equally, making it more resistant to the influence of extreme values.