Assignment 2

Use the Boston Housing Data (from Canvas).

Perform cluster analysis of the data. (GRADING – You will be given 1 point for each of the 10 tasks below).

1. Select the numeric variables that you think are appropriate and useful. Use kmeans and Gaussian Mixture models.
2. Scale the data – using standard scaling or 0-1 minmax scaling. R scale function does standard scaling.
3. Generate the K-means solution. Extract 2-10 k-means clusters using the variable set. Present the Variance-Accounted-For (VAF or R-square). Remember: the local optima problem is big for all the clustering and mixture models. So remember to run them from at least 50-100 random starts.
4. Perform Scree tests to choose appropriate number of k-means clusters
5. Show the scree plot.
6. Choose 1 K-means solution to retain from the many solutions that you have generated
    a. Use the criteria of VAF.
    b. Interpretability of the segments
    c. Doing well in Test. For Test, use the centers (means) generated from the training set k-means solution, as the starting point for performing k-means in test. Use VAF and relative cluster sizes as measures of stability.
7. Generate 3-5 Gaussian Mixtures (GM). Remember to start from 50-100 random starts.
8. Choose one solution based on the mean values and interpretation of the generated mixtures.
9. Compare the chosen k-means training data solution with the chosen GM solution in the training data from an interpretability perspective.
10. Summarize results and interpret the clusters/segments you choose as your final solution.

Please be precise. Be verbose only if you need to.