MIS 776 Assignment 3
Clustering and Model Improvement

This assignment is worth 20 points. You will submit a Jupyter Notebook file (ipynb) with your solutions to each problem. Points will be awarded only for fully executable solutions. In other words, if your solutions throw unresolved errors, the assignment cannot be graded and will result in a 0 score for that portion. Please provide comments in your code, explaining each step in your solution so that I can follow your logic. As a reminder, assignments are individual work and the use of generative AI is prohibited.

Data Scenario

The three most important features impacting housing process are said to be "location, location, and location." It is the characteristics of the neighborhood in which the house is located that will impact the house price as much as the characteristics of the house itself. In this assignment, we will test this concept by seeing if we can predict median home prices in various neighborhoods in Boston based on the characteristics of that neighborhood.

In this data set, we will attempt to predict the median house price (MEDV), which is a continuous variable, based on other continuous and categorical features of the neighborhood. The details of the data set are below. This data set is a modified version of the set published by Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', *J. Environ. Economics & Management*, vol.5, 81-102, 1978.

Data

The target classification (output) column is *MEDV*. All other columns are potential predictors.

CRIM: per capita crime rate by town
ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS: proportion of non-retail business acres per town
RIVER: Yes if the tract is bordered by the Charles river, No of not.
NOX: nitric oxides concentration (parts per 10 million)
RM: average number of rooms per dwelling
AGE: proportion of owner-occupied units built prior to 1940
DIS: weighted distances to five Boston employment centres
RAD: index of accessibility to radial highways
TAX: full-value property-tax rate per $10,000
PRATIO: pupil-teacher ratio by town
LSTAT: % lower status of the population
MEDV: Median value of owner-occupied homes in $1000's

1. Create a new Jupyter notebook called Assignment2_Clustering.ipynb. Import required packages and write the script to open the file into a Pandas data frame called **datHousing**. Run any necessary exploratory data analysis (EDA) to make sure that everything looks reasonable.

2. If there are any features that require dummy coding, modify the dataset to include the dummy coding, using the value in the first row as the reference value for the dummy coding. Remove the feature(s) from the data frame that were the source of the dummy coding so that the features are not duplicated.
3. For now, we are not going to break the data frame into training and testing partitions. Create a regression model on the entire data frame predicting MEDV using all of the other predictors. This is your baseline model. Report the R^2 value and MSE for this model.
4. Create a data frame called **datHousingSub** for clustering by generating a subset with all columns except MEDV. We want to exclude MEDV from the clustering since this is the value that we will try to predict later from the clusters. Do a summary on the subset to verify.
5. Using the KMeans algorithm, create a model with 2 clusters on the **datHousingSub** data frame. Report the centroid values for each cluster and the sizes of each cluster. Using the characteristics that are especially divergent between the two clusters, what would you name these clusters?
6. Repeat step 5 with 3 clusters. Do you think that adding another cluster helps to partition the data? Why or why not?
7. We will use the 2-cluster model going forward. Merge the cluster ids from this 2-cluster model into the **datHousing** data frame using the column name *Cluster* to store this cluster id. Look at the first few rows of data to verify.
8. Create a new data frame called **datHousingC1** which contains all of the rows from **datHousing** in cluster 1. Look at the first few rows of data to verify. Check the Cluster column to make sure that it only stores the value of 1
9. Create a new data frame called **datHousingC2** which contains all of the rows from **datHousing** in cluster 2. Use summary to verify the results. Check the Cluster column to make sure that it only stores the value of 2
10. Create a regression model predicting MEDV and the same predictors as the baseline model in step 3, using the data frame from cluster 1. What are the R^2 value and MSE? Is this higher or lower than the baseline model?
11. Create a regression model predicting MEDV and the same predictors as the baseline model in step 3, using the data frame from cluster 2. What are the R^2 value and MSE? Is this higher or lower than the baseline model?
12. Summarize your findings. Do you think that clustering might improve your ability to predict the MEDV value? If so, under what contexts or constraints?