1/

Partitioning the data into training and validation sets is important to evaluate the performance of a machine learning model. The training set is used to train the model, which involves finding the optimal parameters that minimize the error on the training set. The validation set is used to evaluate the model's performance on unseen data, which helps to estimate how well the model will perform on new data.

The purpose of the training set is to train the model to learn the patterns and relationships in the data so that it can make accurate predictions on new data. During training, the model learns the optimal parameters that minimize the error on the training set. These parameters are then used to make predictions on the validation set.

The validation set, on the other hand, is used to evaluate the performance of the model on unseen data. The model is applied to the validation set, and the predictions are compared to the actual values in the validation set. This allows us to estimate how well the model will perform on new data that it has not seen before. By comparing the performance of different models on the validation set, we can select the best model for making predictions on new data.

2/ Question2.py

Equation:

MEDV = X\*CRIM + Y\*CHAS + Z\*RM + Intercept

With:

CRIM: per capita crime rate by town

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

RM: average number of rooms per dwelling

Coefficients = [X Y Z]

3/

Using the estimated regression model, the predicted median house price for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6 can be calculated as:

MEDV = 4.012 + (-0.202) x 0.1 + 7.381 x 0 + 8.115 x 6

MEDV = 4.012 + (-0.0202) + 48.69

MEDV = 52.6828

Therefore, the predicted median house price for this tract is approximately $52,683.

4/

INDUS, NOX, and TAX are likely to be measuring related aspects of the housing market.

INDUS represents the proportion of non-retail business acres per town, NOX represents nitric oxide concentration (parts per 10 million), and TAX represents the full-value property-tax rate per $10,000. These three predictors are likely to be measuring related aspects of the housing market since areas with a high proportion of non-retail business acres may also have higher levels of air pollution and higher property tax rates. Therefore, it may be redundant to include all three predictors in a regression model, and it may be more appropriate to choose one of these predictors to represent this aspect of the housing market.

In summary, it may be useful to reduce the number of predictors by selecting the most important predictors based on their relationship with the target variable, and by removing predictors that are highly correlated with other predictors. In this case, INDUS, NOX, and TAX are likely measuring related aspects of the housing market, and therefore, it may be appropriate to remove one of these predictors to reduce the number of predictors in the model.

5/Questrion5.py

This will print the correlation matrix, which is a table showing the correlation coefficients between all pairs of numerical variables. Correlation coefficients range from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

To search for highly correlated pairs, we can look for correlation coefficients that are close to 1 or -1. Pairs with coefficients close to 1 indicate strong positive correlation, while pairs with coefficients close to -1 indicate strong negative correlation.

Based on the correlation matrix, we can decide which predictors to remove to reduce multicollinearity. One common rule of thumb is to remove one of the predictors in a pair that has a correlation coefficient above 0.7 or below -0.7. Another approach is to use a statistical method such as principal component analysis (PCA) to transform the predictors into a set of uncorrelated variables.

Note that the decision of which predictors to remove should also take into account the context and the goals of the analysis. Some predictors may be important for the research question, even if they are highly correlated with other predictors.

6/