HW1

Question 1

First, drop the "CAT..MEDV" column, because all of its values depend on the "MEDV" column

Then, do the train-test split, and use decision tree to predict "MEDV" value, use a if-statement to derive categorical values for classification.

We do this because we know the distribution function for this classification:

$$f(x) = \begin{cases} 0, \mathbf{w}^T\mathbf{x} \leq 30 \\ 1, \mathbf{w}^T\mathbf{x} > 30 \end{cases}$$

Therefore, we would not need to fit a Sigmoid to transform this into Logistic regression, and therefore we would not need to choose a threshold t as well.

The AUC for this decision tree is 0.8722729, and some of its confusion matrix statistics is the following:

```
Confusion Matrix and Statistics
          Reference
Prediction   0   1
         0 155   7
         1  11  30
               Accuracy : 0.9113
            Sensitivity : 0.9337
            Specificity : 0.8108
         Pos Pred Value : 0.9568
         Neg Pred Value : 0.7317
             Prevalence : 0.8177
         Detection Rate : 0.7635
   Detection Prevalence : 0.7980
      Balanced Accuracy : 0.8723
       'Positive' Class : 0
```

Question 2

First, perform train test split, and normalize the data. For training set, normalize using min-max scaler. For test/prediction sets, use the min max of the training set.

Then, iterate through 1 to 10, to find the k with the maximum AUC. The result is k = 3, with AUC = 0.8603061.

Question 3

As we've noticed, this classification does not require us to decide the threshold t for the sigmoid function, therefore, we would be interested in the AUC of the model, as we have no special preference towards a low False-Positive or False-Negative rate.

Therefore we choose the Decision Tree model.

Question 4

The only thing worth noticing is, using the str() command, we found out that the training dataset has specific datatypes. We need to adjust the prediction dataframe accordingly.

The result is that we predict the MDEV to be 21.31474, therefore it would be classified as 0.