Boston Housing dataset statistics (in $1000's):

Total number of houses: 506
Total number of features: 13
Minimum house price: 5.0
Maximum house price: 50.0
Mean house price: 22.533
Median house price: 21.2
Standard deviation of house price: 9.188

Question 1:

The attributes that have the higher influence on the price of the property are CRIM, RM and DIS. CRIM as safer areas are characterized by houses with a higher value than those in a neighbour with a high crime rate.
RM: the price of a house is strongly related to how big it is.
DIS: Areas that are more convenient to more emplyment center have a higher demand and thus properties are more expensive.

Question 2:

**Answer: ** CRIM = 11.95, RM = 5.609, DIS = 1.385

Question 3:

If we were to test the model on the same data we trained it onto we would likely obtain great results that however could not be very representative with new data(data leakage). In order to verify that the model is able to generalize there is the need of a set of data that is indepentent from the training data.

Question 4:

Since this is a regression problem I used MSE. I preferred it over MAE as it emphasizes big error over small ones and is derivable.

Question 5:

It's an exhaustive search approach used to find the value of the model's parameters that best approximate the data. It is applicable all the time that cross-validation is. This method needs a finite number of parameters on which iterate and compare.

Question 6:

Cross-validation is a technique used to exploit the totality of the data for both training and testing. It consists in dividing the data in two complementary sets: one for testing and one for training. This operation is repeated several times on different partition of the data. The final result from the training is the mean of the errors obtained with each partition. If this method wasn't used in combination with grid search then the result of grid search would be of finding the combination of parameters that work best with that specific set of data, again, this might not be representative of a more general situation.

Question 7:

max_ depth = 1. As the training set size increases the training error raises to a bias value, this happens because the model is too simple to generalize well a large number of data. Test error rapidly decreases to an almost constant level with the growth of training points, this happens again because the model is too simple and misinterprets the data. Notable the variance on the test error is very low compared to other depths.

Question 8:

The model suffer from high bias when depth is 1. This means that the model is underfit: it doesn't take into account all the important features of the data so it doesn't represent them correctly. This results in a high and costant training and test error (bias). Those values are similar because both dataset are bad interpreted by the model.

When the depth is 10 the model suffers of great variance which means that it is highly sensitive to different training set. The model was trained to score perfectly on one specific training set but it is not able to generalize its prediction to test sets. This phenomenon is called overfitting. A tipical behaviour of overfitting that can be found in the graph is a very low training error along a much higher testing error.

Question 9:

The training error is invertially proportional to the model depth: this happens because by increasing the complexity the regression fits tighter and tighter around training data. This phenomenon however makes it difficult for the model to generalize to independent data, as can be seen in the evolution of the testing error. The testing error initially decreases because the model gets more complex and it's able to interpret the training data. However when the depth becomes too big the model it's not able to correctly work for the test set and the testing error rises.
According to the graph I am seeing the best depth is 4 as there is a clear minimum in the testing error.

Question 10:

I run the code ten times. Median value was 6 while the average was 4.6. The average value is consistent with my prediction (4).

Question 11:

After 10 runs the average best selling price is 20.839, median 20.766. Those values are lower of both median (21.2) and mean (22.533) selling price of the dataset but are well within the data standard deviation range (9.188).

Question 12:

In order to make this decision I need to understand what is the error that is applicable to my prediction. The testing mean squared error ranges between 15 and 20, I'll use the higher value so to have a more conservative estimate. An MSE of 20 means an error on the measure of +-4.47 which is roughly 20% of the sum. This means that my prediction has an error bar of overall 40%: the correct price could be 20% higher or 20% lower. The error is pretty wide and I think it's caused by the small test set we are using. That being said i wouldn't use the model to define the exact price of a property, however if I didn't have any other reference I would use it at least to get an idea of what the price could be.