Read the house.csv data into a dataframe using the read_csv function.

df=pandas.read_csv("house.csv")

Let us divide the data into a training set and a test set in the ratio 60:40. We will split the data randomly to avoid any biases in sampling. To repeat the splits (and for all learners to get the same split) it is important to set a random state. (if you don't use a random state, every time you run the codes you get a different set of 60% data into train and 40% into test)

train, test = train_test_split(df, test_size=0.4,random_state=1)

Let us build a model on the training data using all variables given.

price ------->>>>> bedrooms+bathrooms+sqft_lot+floors+condition+yr_built

1) Check the performance of the Model on training data.

The Adjusted R2 of the model is BLANK (use 2 decimal places without rounding off)

0.35

2) How many dummy variables will we have into the Linear Regression Model for the variable condition?
Note:- Condition Variable Level average is set as a reference and not considered into the model to avoid redundancy.

4

3) The co-efficient of the variable conditionFair is -6.024e+04. This means that a house in Fair condition is priced -6.024e+04 $ lesser (because it has a negative sign) than a Average condition house. (since average is taken as the reference level)

The co-efficient of the YearBuilt is -3871. This means that for 1 unit increase in Year (or an house one year older) the price decreases by -3871 $ (because it has a negative sign) if all other variables are kept same.

Is the above a correct interpretation of Regression Co-efficients? Answer is True (Correct) or False (Not correct)

True

4) Calculate the value of SSE (Sum of Squared Error) on the training data.

377992536277877.06

5) The R2 for the test data is

0.30