LAB 4 - Linear Regression Using K-fold Cross-Validation

In this lab, we will perform linear regression using the K-fold cross-validation technique. We will perform the same regression as the previous lab (i.e. we will investigate the effects of age, height, mental and skill to the player's salary), but this time we will perform it multiple times, once for each *fold*. A *fold* is a portion of the data, where each fold is (almost) equal in size. We will split our dataset into k equal-sized folds. For each fold, after performing linear regression, we will compute the MSE and store it. We will take the average of all these MSE values, which will be our cross-validation error. We will compare this technique to the usual train-test splitting method (also referred to as the "validation" method).

Follow these instructions:

- Form your input matrix X and output vector y like you did in LAB 3. Copy your functions from the previous lab which calculated the regression coefficients and computed the MSE.

- (55 pts) Implement a new function which accepts 3 parameters: The input matrix X, the output vector y, and an integer k. This function should return the cross-validation MSE. Inside the function:

    - Implement a loop which iterates through each *fold*. This means you should know the size of a fold beforehand (**Warning: the sizes of different *folds* might be different from each other depending on the value of k!**). Inside this loop:

        - Split X and y into train and test sets. The current *fold* should be the test set, and the remaining sections of X should be the train set.
        - Perform linear regression using train data.
        - Calculate predictions & compute MSE using test data.

    - After the loop ends, take the average of these MSE values and return it.

- (45 pts) In a loop, do the following 10 times:
    - Shuffle the data. (**Warning: X and y must be shuffled in the same exact order!**) You can use `numpy.random.shuffle()` for this purpose.
    - Calculate and display the "validation MSE":
        - Split the data 80 to 20 (first 80 train, last 20 test), perform multiple linear regression and calculate the MSE (We have already done this last week).
    - Calculate and display the cross-validation MSE, using 8-fold CV.

Here's the console output (try to replicate this output exactly):

```
Validation MSE      8-fold CV MSE
--------------      -------------
17694869.46         18892637.43
22427800.23         17651637.56
17172217.74         17686502.67
10092111.19         18253748.81
13253743.33         18384006.05
10496698.57         17078808.16
16878871.96         17206852.46
11111507.43         17040001.56
21442072.16         18026621.67
25057863.41         16917962.95
```

Numbers will be different each time you run the code, since shuffling is done randomly.

Important note: These instructions require you to implement the calculations manually. Any submissions which bypass such calculations will receive 0 points from corresponding parts.