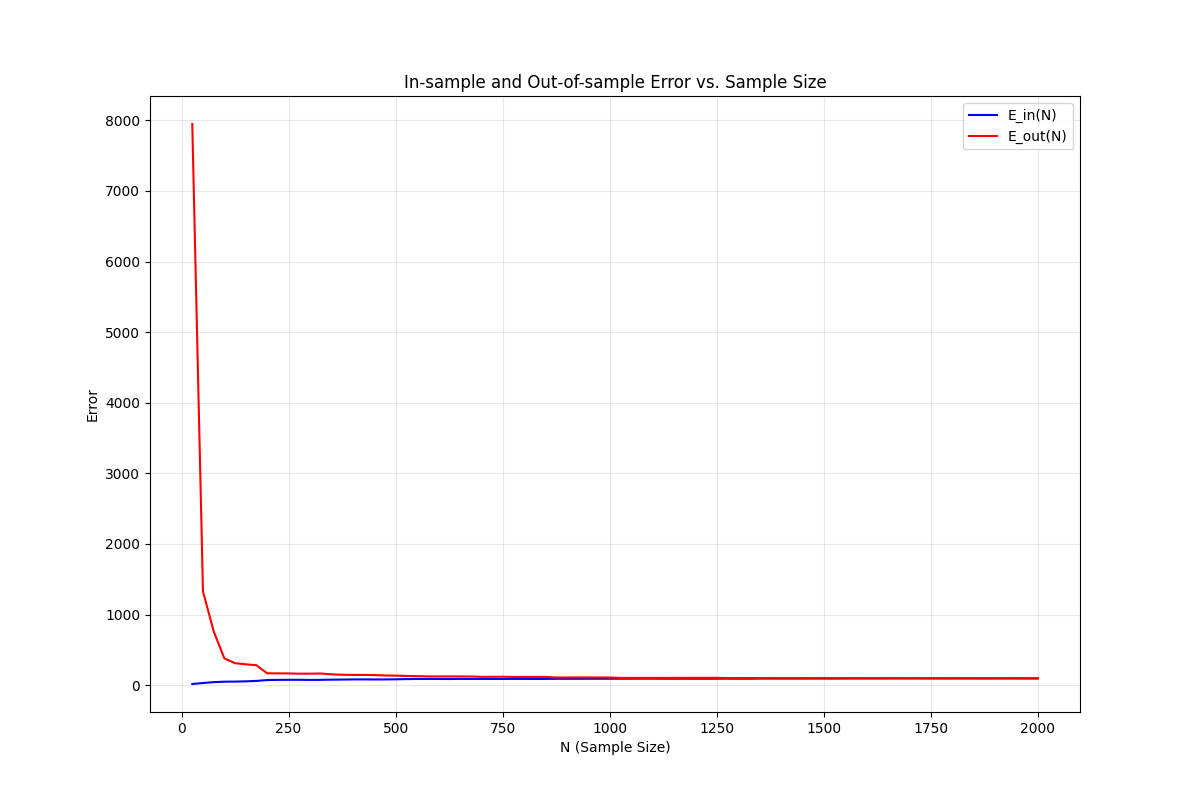
# Question 11: Report



## Explanations:

In this problem, we gradually increase the sample size (until 2000, which is nearly of our dataset), and calculate the average in sample error and out of sample error over only 16 experiments.

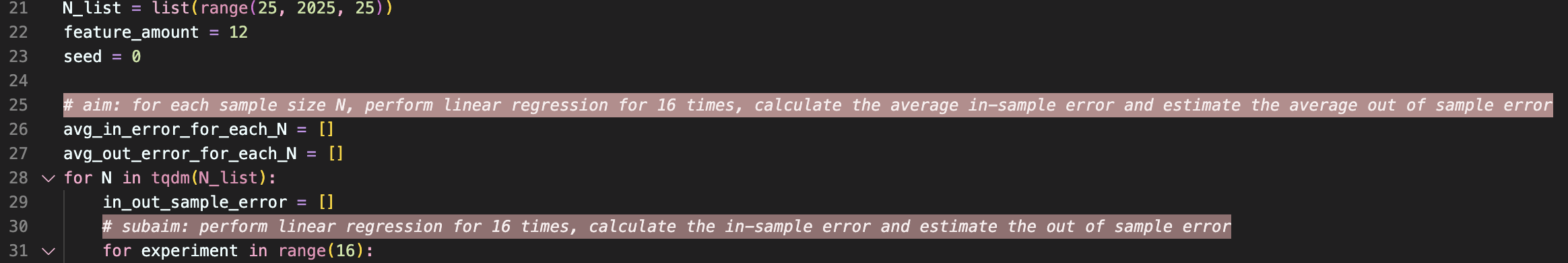
From the graph we can see that when the sample size is under about 100, we cannot generalize well since this is represented by the high out of sample error. But after we met a certain sample size (which is maybe 100 or 150 or some amount around them), we have enough data to represent the whole dataset by the sample, so that we have enough information to construct our model.

Also, if we really look at the examples, we could find that the number of features is 12, which is enough to provide sufficient degrees of freedom to fit the training data. In a geometrical view, we can think of finding a 12 dimensional hyperplane to split the data points, in a 13 dimensional space (since we add 1 to the input vector.) In lower dimensional spaces, like 2D or 3D spaces, the degree of freedom is smaller, therefor it is hard to find a line or a 2D plane that can pass through all the points, but for a 12 dimensional hyperplane, it is easier to do so. Thus, it is possible to find the “optimal” hyperplane to fit best to the sample points, producing nearly 0 in sample error in all cases.

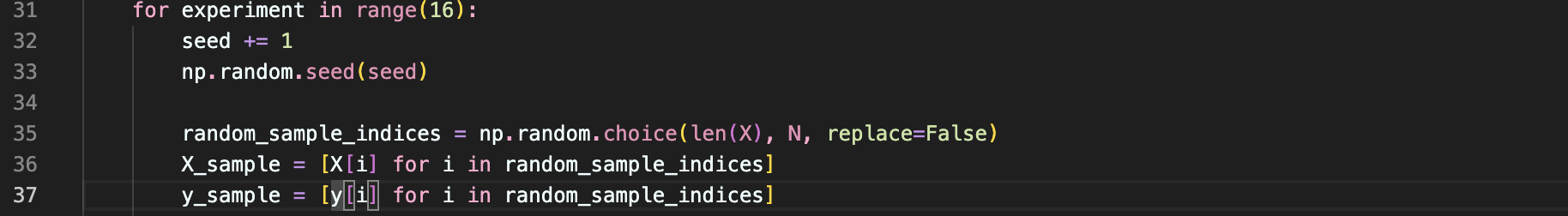
## Code:

(The reading data part is ignored in this report since it’s the same in the previous question.)

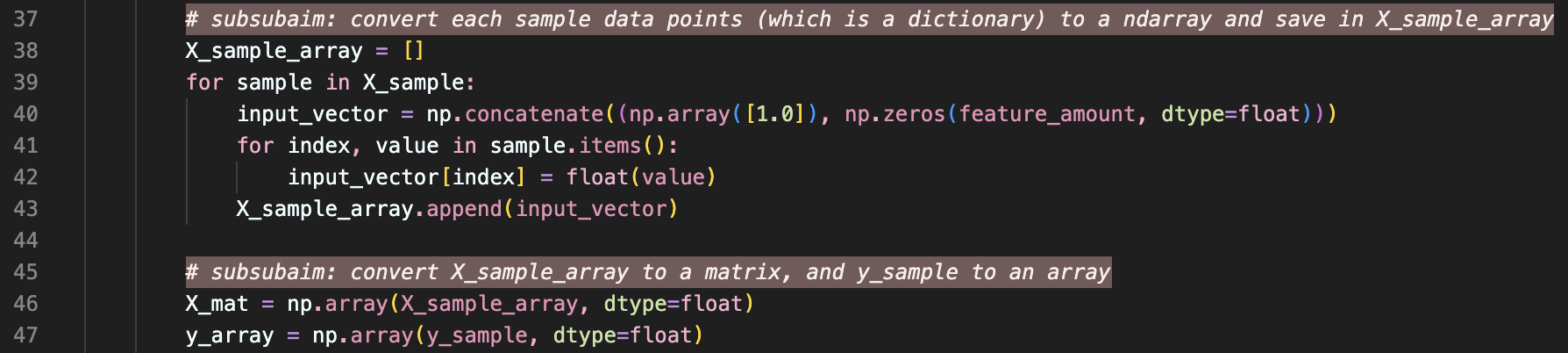
The first thing to do is to define a list of all possible s that we need to iterate through (25, 50, 75,…,2000), then for each value of , we conduct the experiment 16 times.



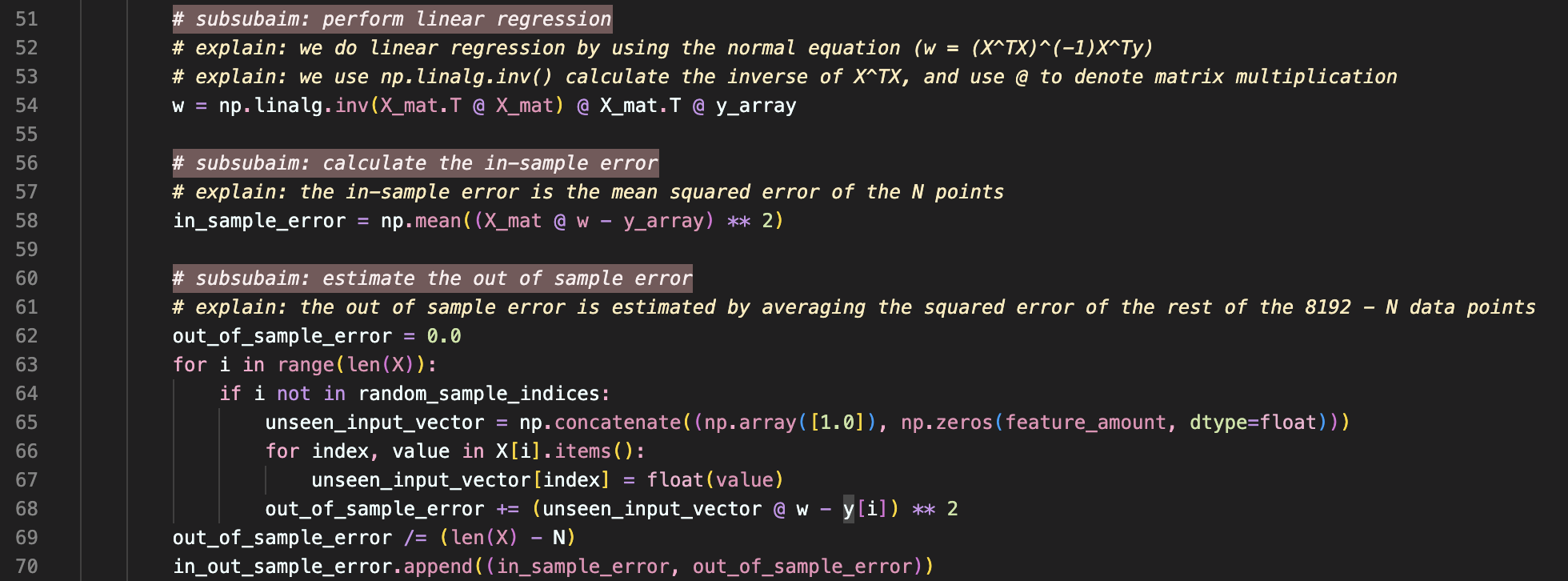
The random choosing process is done without replacement. We save the chosen examples in X\_sample and their corresponding in y\_sample:



Then convert data type:

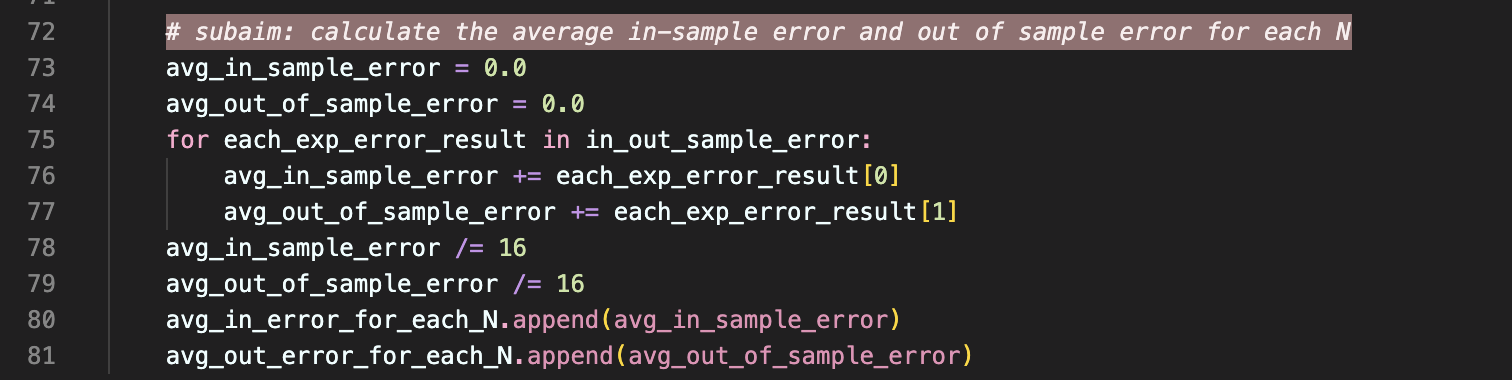


Calculate in sample error and out of sample error:



The in\_out\_sample\_error is a list containing the result of the 16 experiments for any given .

We then calculate and :



The last part is to plot the result, but it’s similar to the previous question.