Summary: In this exercise, the housing prices were predicted by: 1) KNN, feature-weighted KNN (self-invinted, with PSO optimizer), class-based linear regression and Maximum likelihood Kernel Smooth Regression (MLR), implemented in *Matlab*; 2)Support Vector Regression (SVR), implemented in *R*; 3)Random Forrest Regression (RFR), implemented in *Python*. The data was sampled for fast prototyping. The performances of models were evaluated using root of mean-square-error (RMSE) for convenient performance analysis using statistics of the original data. The result shows:

- KNN classfication (Matlab): £2.238 * 10^5
- feature-weighted KNN (Matlab):£ 3.175 * 10^5
- MLR (Matlab): £3.046 * 10^5
- SVR (R): £2.643 *10^5
- RFR (Python): £2.920 * 10^5

A case-based linear regression experiment shows that there are a few types of data (particular property type, location and lease duration) are missing. Improvements of the prediction can be gained by using larger sampled data, tuning model parameters, using more features, and error analysis to get rid of over-fiting and under-fitting. The present comparison results can’t draw any reliable conclusion due to lack of samples, features, and parametric tunings. The running version of code is available at: https://github.com/cwangED/HousingPriceExercise.

Problem: One unclear point in the exercise requirement is whether we should treat the date as time-series. If not, we have to treat it as a classification problem. In this case KNN become one of few possible solutions. But if date is allowed to be considered, many regression tools can be used.

Answer of questions:

1. If you are struggling to implement something that deals with this volume of data, do you know of a way to deal with it in theory?

Using iterative training and testing implementation. Each time just input one row of data to the model. Using big data techniques to run the experiment on a server.