Our goal is to continue our work over the summer and try to make better predictions of rental housing prices. So the variable that we’re predicting is the price.

Using Python, we will be practicing the modeling methods we learned in this class for continuous outcome variables. Namely, we will be using Linear Regression, GAM, KNN, and Random Forest. Then, we will be learning and implementing new methods of Boosting Trees and Neural Network. We are also considering applying Support Vector Regression.

Our preliminary results is based on a crude method of filling NAs with averages of each corresponding variable. The results will likely change in our final draft, where we would be using PCA matrix completion to fill the NAs. For Linear Regression, our current out-of-sample error is 342. For GAM, our oos error is 297. For KNN, our oos error is 341. For RF, our oos error is 268.

Our deliverable will be a short paper with an introduction, a method description with a table of models we used and their respective expected prediction error (both out-of-sample and in-sample) and computation time, a result part with our selected model and a visualization of that model’s predictions, and a discussion part explaining potential reasons why our selected model best suits our problem.

By Dec. 14th, our plan is to first use PCA to complete missing data. Then we will be adding an in-sample EPE to our preliminary methods through CV. Then, we will implement boosting trees and neural network, and calculate in-sample and oos EPE and record all results into a table. Then, we will select our best model and visualize its predictions. Finally, we will discuss why this model was expected to perform best, as well as where we can improve.