Steve Su, Kristian Montoya

Kaggle King’s County Home Prices

**EDA**

Our first step was to examine the data by performing EDA. We plotted the response variable, price, to each predictor variable. We noticed some linear relation with some predictors but not with others. Were we could, we transformed the data to something more meaningful relative to price. For example, the variables, lat and long, themselves are not particularly good predictors themselves but combined together they determine where a house is located and any realator will tell you location, location, location. Knowing that house prices tend to be highest towards the city center, we combined lat and long to calculate distance from the most expensive neighborhoods in King County. We then plotted this distance against home prices to get a better linear relation compared to lat and long on their own. Correlation coefficients were also computed between predictors where we noticed a few with high correlation coefficients.

[talk about normalizing other data here]

Lastly, as part of our EDA we created box plots and threw out some of the outliers from our training data because we didn’t want our model to train on these unlikely observations.

**Linear Regression Model**

Now with our data prepared we split the data, 70% for training and 30% for test. We used multiple linear regression as our technique using libraries from statsmodel. To evaluate the performance of different models we used RMSE. For our first model, we ran all predictors but took out lat and long and replaced with distance (as described above). Our Rsqared value was 0.72 with a RMSE of 199305. With room for improvement, we tried other models. From the EDA, we noticed that our distance predictor seemed to have a squared relationship with prices, so in the model we added a distance squared term which dropped the RMSE by 10000. Smaller gains were made by taking out predictors which had high correlation with another or taking their product and putting it into the model. This only accounted for a drop of RMSE of 2000.

**Tree Model**