Brendan Garrett & Sharon Greenblatt
Final Project Proposal

Project Idea:

When buying or selling a home, there are many different factors that can and do affect the price. Our project will use a data set from a Kaggle Competition; specifically House Prices: Advanced Regression Techniques. This competition is recommended for data science students who have experience with machine learning techniques but we want to fulfill the requirements of the competition while also exploring other avenues for feature engineering and model selection. The goal of the project is to accurately predict the final sale price of homes.

Data Set Description: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

There are 79 explanatory variables describing every aspect of residential homes. These variables are both categorical and continuous and will need to be scaled, normalized, or removed in accordance with their relevance. The data set contains variables including everything from lot size, to type of roof, to proximity to a main road or railroad. The data set contains training and testing sets which contain about 1400 instances each.

Code and Libraries:

This dataset of house prices includes a large amount of categorical and numerical attributes. In order to conquer this, we will take three steps. First, we will do EDA. This will include seeing if any of the features are directly related to each other. If the examples of one category are allows part of another category, we will include only one of the categorical features. We will see if categorical features seem to cluster together by running topic modeling. We will use clustering, like k-means and hierarchical, to see if the examples clump together in clusters that are related to the price of the house. An ANOVA can be run in order to determine if clusters significantly relate to price. After EDA, we will look at feature selection. Besides removing redundant features, we will remove features that do not correlate with price at all. After this, we will perform dimensionality reduction while keeping variance in features. First, we will try principal component analysis, then something nonlinear like kernel PCA. We will be using scikit-learn and scipy for feature selection, dimensionality reduction, and EDA

Now that we have a good idea of which features are relevant and how they relate to each other, we will build a model. Because of the mix of both categorical and numerical features, we will first try ensemble learning with either random forests or adaboost. These are implemented in scikit-learn. After that we may evaluate using another model, such as a neural network in tensorflow.

Evaluation Criteria:

This data set is part of an active Kaggle competition. This competition currently uses root mean squared logarithmic error. Due to this, we will be using this measurement to compare our models. We will also look at how our models perform in relationship to the price of the house itself. A hypothesis is that higher cost houses will be more complex, so our model may produce different rates of error than lower cost houses.

Planned Work Completed by MS1:

By MS1, we will complete the EDA as well as the feature selection and dimensionality reduction. We want to have graphs and models that explain our feature space by MS1. In addition, we will build at least a simple model to see how it performs compared to the current leader boards on Kaggel.

Collaboration Plan:

Brendan and Sharon will both work on the EDA and collaborate to evaluate feature selection. Then Brendan will focus on the theory behind the preprocessing techniques as well as the domains of the different libraries we will be using. Sharon will focus on the writing intensive aspects as well as logistics and final evaluations of the model and its error.