Project Proposal

Research question: Which features are the most relevant in predicting house sale price, and which type of model best predicts the house price as a function of those features?

The housing market has always been a trending topic; people are sensitive about how tight the current sales market is and whether it is a good time to buy or sell. According to the PD&R, there are 6120 thousand existing home sales and 771 thousand new home sales in 2021. The average sale price is $347 thousand and $397 thousand for each, approximately 20% higher than the price in 2020. With this highly active and fluctuating market, researchers are making lots of forecasts of pricing, supply, and demand every day based on economic conditions and demographic trends. As for individual builders and mortgagees, housing inventory characteristics are crucial in determining the sale price for each specific property. Road access, configuration, and accessories all affect how much people are willing to pay. Uncovering the key features in determining house sale prices is a difficult task. But we are excited to give house price prediction a shot using the skills we have learned in Math 243.

Our candidate dataset is Ames Housing dataset compiled by Dean De Cock for use in data science education. These data were gathered from Ames Iowa during a period spanning 2006 to 2010. The response variable, house sale price, is quantitative. The dataset includes 80 features, with more than half of our predictors categorical. The predictors range from geographical information about the property and the neighborhood, to interior measurements and quality of amenities, to sales statistics. There are 1460 observations recorded in the training dataset and 1459 observations recorded in the test dataset. It should be noted that this test set does not contain response values. It is intended for use in generating predictions to submit to a Kaggle competition. Hence, in our analysis, we will use the training dataset to train and test our models. At the end of the project we plan on submitting final predictions on the test set to Kaggle to see how we rank.

We are excited to be working with a dataset that is more current and expansive than some of the other famous and accessible datasets containing home price information. As an example, the Boston House Prices dataset that is used in many statistics classes (we even used it in our class!) is sourced from data compiled in 1970. In addition to being a relatively recent dataset, the response variable, home sale price, is incredibly relevant to the lives of people all across the country. While none of us are in a position to buy a house, there are millions of people in the US who are either homeowners or are considering buying. Predicting house prices is useful not only for buyers looking for a new home, but also for owners evaluating their assets and deciding whether to sell. Any insight we gain into the factors that lead to an increase in home price will be exciting, in light of the importance of home pricing to so many people.

The Ames housing dataset contains over a thousand observations and a large number of predictors. A dataset with these characteristics is useful in learning how to put our data analysis skills into practice. In analyzing the data, we will have to make careful use of the variable selection tools that we have learned about in Math 243 in order to select out superfluous variables. The dataset is large enough that we can try out various algorithms from the class without sample size concerns. Since the dataset is currently in use in an ongoing intermediate level Kaggle competition, we also can be fairly confident that many of the statistical methods we have learned in this class will work with the dataset. Overall, this

dataset presents an exciting opportunity that will allow us to put Math 243 skills to the test. As a fun bonus, we also get to submit our final model to the Kaggle leaderboard at the end of the project.

Since the data set is sourced from a Kaggle competition which provides full public access to the dataframe and codebook, we don't have to worry about data acquisition. As for data analysis, one problem is that over half of the variables in our data set are categorical, which means that we might have many dummy variables if we choose to do linear regression. The resulting linear model would be very complex and hard to interpret. To address this issue, we could turn this into a classification problem by breaking the response variable down into intervals. Predicting a sale price range may be more useful and accurate for practical purposes, and we won't have to create dummy variables for categorical features anymore. We can then compare the performance of classification models with regression models. Secondly, model selection may be computationally expensive due to the large number of observations and features. This could be an obstacle for certain variable selection procedures. For example, best subset selection is not going to be an option. Another potential problem could be the presence of outliers skewing the model. House price could potentially be a right-skewed distribution, where most houses cost a moderate amount and a few houses cost extremely large amounts. The dataset also contains many NA values which might be problematic for model training, so we need to decide on how to deal with those observations. For features where the majority values are NA, it might be more appropriate to disregard that feature entirely. Finally, the fact that these data were collected during a time period overlapping the 2008 financial crisis suggests that the results obtained from this data set may not be generalizable to current market conditions. However, we feel that we will still be able to learn from the general insights uncovered by our analysis of the Ames house price dataset.