Introduction:

Housing has always been seen as one of the major stepping stones in adulthood; people graduate college, get a job, start a family, and save enough money to buy a house. The question though has always been how much does it cost to buy a house and what factors into the price of a house. In this project, we looked into dataset from 1460 houses bought in Ames Iowa to create a regression model that best predicts the price of a house in Ames, Iowa. The variable of interest that we study is Sales Price and in our dataset we have eighty predictor variables that are utilized to predict the cost of a house. Of the eighty predictors, we have 23 nominal, 23 ordinal, 14 discrete, and 20 continuous and these range from everything from Total Plot Area to Fire Place and Pool. Our interest in this project is to select the most important variables to study and create a model that best predicts the Sales Price of a house. In evaluating our model, we will look into the bias of our model to the actual sales price, the maximal deviation, mean absolute deviation, and mean square error to conclude how accurately we created a model to predict the Sales Price. We will use 1060 observations in our training data set and compare our model to the 400 observations in the test data set to see how our modeled Sales Price compares to the actual Sales Price of the 400 observations.

From our dataset, we will look into normalizing the Sales Price by using the log function, and we will convert some of our predictors into dummy variables. From the list of variables, we will focus on several that we believe are highly correlated with SalesPrice, and then variable selection techniques also to see the maximum number of variables for predicting a highly accurate model without overfitting the data. The purpose is to create a model that can best predict the Sales Price from our validation data set, but also that could be used on other data sets and still get a good prediction.