Final Report: Housing Prices Prediction using Advanced regression

Math 189/289C

Homework 4 Report

April 4th, 2017

By:

- Emma Roth, 4th Year Computer Science/Bioinformatics Major
- Ileena Mitra, 1st year Bioinformatics PhD
- Megan Lee, 3rd Year Mathematics/Economics Major
- Jing Gu, 2nd year Chemistry MS
- Keven Nguyen, 4th Year Mathematics/Economics Major
- Adithya Bharadwaj Balaji, 1st year Graduate student in Electrical and Computer Engineering

TABLE OF CONTENTS: 2

INTRODUCTION 3

GOAL 4

DATA 5

DATA CLEANING 5

EXPLORATORY DATA ANALYSIS 6

Correlation 6

Summary of Key Variables 14

RESULTS 23

Multivariate Regression 23

Random Forest 27

Gradient Boosting 33

CONCLUSION 35

APPENDIX 36

Detailed Data Description 36

Variable Selection 50

R code for analysis: 56

INTRODUCTION

The purpose of this report is to create a model to accurately predict sale prices in Ames, Iowa. In order to do so, it is essential to gather real-world data about Ames so we have an accurate picture of its social, economic, political, and geographic climate.Ames is located in central Iowa and has a population of about 65,000. The average household income of residents of Ames, Iowa is $56,399 and the average age is 31 (*Neighborhood Scout*). Ames boasted a 97.4% high school graduation rate and a 2.1% unemployment rate in September 2015. While 62% of all adults over age 25 hold a college degree, roughly 31.5% of the population lives under the poverty line. In 2014, the fraction of the population who rented houses was 48%, higher than the national average of 35% ("Quick Facts: Resident Demographics.").

Buying a house is a very involved process on the customer's part as the decision concerns an incredibly wide range of structural, environmental, and neighborhood characteristics, as well as the psychological wants and needs of the buyer. To predict the sale price of a home is a challenge for a number of reasons, but especially due to the sheer number of variables that describe the intrinsic and extrinsic properties of a house. However, much research has been done to determine universal features that buyers are willing to pay highly for. For example, many studies have proved that access to a quality education deeply affects the price of houses in an area. For example, the real estate company Redfin did a study in 2013 analyzing real estate markets in high- quality school districts. "Using a huge database of about 407,000 home sales and nearly 11,000 elementary school districts in 57 metropolitan markets, the study concluded that, on average, buyers pay $50 more per square foot for homes in top-rated school districts compared with homes served by average-rated schools" (Harney).

Additionally, researchers at Ohio State University found that "each percentage point increase in the pass rate of ninth grade students on a statewide proficiency exam increases house prices by one-half percent" (Haurin 16). The average home value in Ames is $207,752; in comparison the national average home value is $186,000 (*The Economist*). This may be explained by the fact that Ames is home to Iowa State University.

Along with access to education, neighborhood is also a huge factor in both determining the asking price of a house and the willingness of a customer to buy a house. "When determining the location of their residence, consumers will prioritise the environment and public facilities and services offered by the house's neighborhood. After selecting an appropriate neighborhood environment or the desired public facilities and services, customers will further consider the house site," writes Lee Chun Chang in *International Journal of Management and Stability* (31).

In a paper for the American Society of Business and Behavioral Sciences Conference, Shahidul Islam of Grant MacEwan University analyzed the impact of neighborhood characteristics on house prices. He determined that both tangible and intangible factors, such as number of rooms and average household income within the neighborhood, affect the sale prices of homes. The most significant result of his analysis was that crime variables have a negative influence on house prices, but overall has little impact on the final sale price of a home. Islam's study shows that the final determination of house price is influenced by many characteristics, and in aggregate certain characteristics are outweighed by others.

Unsurprisingly, the features of the house affect its value. A 2014 U.S. News article studied return on investment of different types of home remodels and found that entry door replacement, deck addition, attic bedroom remodel, and minor kitchen remodel had the greatest percentage return on investment, while home office remodel, bathroom addition, and master suite addition yielded the smallest percentage returns (Weigley). Because the housing market is driven by supply and demand, housing trends will also affect what buyers are willing to pay more for. In 2013, USA Today published a list of features prospective buyers said they were willing to pay more for based on survey data, and these features included central air conditioning, home less than 5 years old, granite countertops, and new kitchen appliances (DiClerico). Another economic factor that plays into demand for houses is the overall market climate—houses may sell for much more during a boom and much less during a bust. It has been well documented that the housing market operates in a general cycle, so knowing what year houses were sold in is useful for further refining current house price predictions.

A popular method to determine real estate prices is hedonic regression. This method breaks down a good being sold (in our case, a house) into its constituent parts, assigns a value to each of the parts, and estimates how much influence each feature has on the house. “A primary reason for undertaking hedonic analysis of the housing markets is to understand the structure of demand for housing attributes and environmental amenities. Such understanding is essential for predicting the response to changes in the housing market and for providing welfare estimates of the costs and benefits associated with these changes.” (Sheppard 5). Hedonic pricing is useful when assessing real estate because houses and buildings can be drastically different and therefore creating a generalized price model for structures as a whole may not be accurate.

Based on knowledge accumulated from MATH 189 as well as an understanding of hedonic demand theory, our group creates a model to predict sale prices in Ames, Iowa. We begin by correlating all variables in our data set with sale price of the training data and discard features that do not appear to have any relevant influence on sale price. With our reduced subset of features, we then run multivariate linear regression, random forest, and gradient boosting.

GOAL

This project is about analyzing data about houses in an effort to predict house sale prices in Ames, Iowa, United States. Our goal is to build a robust and accurate predictive model that uses information about a house as input data and outputs the estimated sale price of the house. To build our model, we use multiple features about the house (see more details in sections below). It is important to be able to accurately predict the house sale prices to gauge the house retail market. Our successful model could potentially be used by home sellers and buyers to estimate the house bidding price.

DATA

The data given for analysis and prediction of the housing prices is of two types: “train.csv” and “test.csv”. Here, we take into consideration the “train.csv” file for analysis. The dataset consists of 81 variables and 1460 entries. A brief description of the variables and the explanation the categorical variables is given in the APPENDIX 1.

DATA CLEANING

On observing the given dataset, we can see that there are 19 variables with missing values in their records. First and foremost we must try to clean up these NA’s present in the data as it would make it much more easier to work with in terms of exploratory analysis.
The 19 variables with missing values and the corresponding number of missing entries are:

LotFrontage	Alley	MasVnrType	MasVnrArea	BsmtQual	BsmtCond
259	1369	8	8	37	37
BsmtExposure	BsmtFinType1	BsmtFinType2	Electrical	FireplaceQu	GarageType
38	37	38	1	690	81
GarageYrBlt	GarageFinish	GarageQual	GarageCond	PoolQC	Fence
81	81	81	81	1453	1179
MiscFeature					
1406					

Instead of randomly imputing the missing values we tried to take logic into consideration while imputing.

- LotFrontage: Linear feet of street connected to property. This variable cannot be 0 (as this would be a property without access), so we imputed the square root of the Lot Area on observing that there is a correlation between the two variables for the non missing entries.
- Alley: We presumed that these properties just don't have an alley access so assigned “NoAccess”.
- MasVnrType/MasVnrArea: Both these variables have the same entries missing. So we set them as "None"/0.
- Bsmt Variables: A number of variables in connection with the basement and all the corresponding entries for the various variables have Na’s, so these houses are assumed to not have a basement. We filled in the categorical variables with “None” and the numerical variables with “0”.
- Electrical: Just one missing value so we just imputed the most frequent one.
- FireplaceQu: We assumed that the properties with missing values just don't have a fireplace. So we replaced the missing values with “None”.
- Garage Variables: A number of variables in connection with the garage and all the corresponding entries for the various variables have Na’s, so these houses

are assumed to not have a garage. We filled in the categorical variables with "None" and the numerical variables with "0".

- PoolQC - On checking with the corresponding entry of the PoolArea variable, we can see that the pool area is 0. So these houses don't have a pool and we assign "None" to PoolQC.
- Fence: There are lots of missing entries for Fence and we assign "None" to them.
- MiscFeature: We assign "None" to the missing values as these houses are assumed to have no special features.

We had 81 variables to start off with and we reduced that to 66 variables, by combining variables and removing the redundant variables. Out of these 66 variables remaining, 31 are categorical and 35 are numerical.

Firstly, we removed the variables which had more than 50% of the entries missing. So, we ended up removing: Alley, FireplaceQu, PoolQC, Fence and MiscFeature.

We considered the Categorical variables: BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2. We assigned numerical values to their levels and multiplied these five variables to obtain their interaction effects. We then saved the result under the variable BsmtQual and removed the other variables.

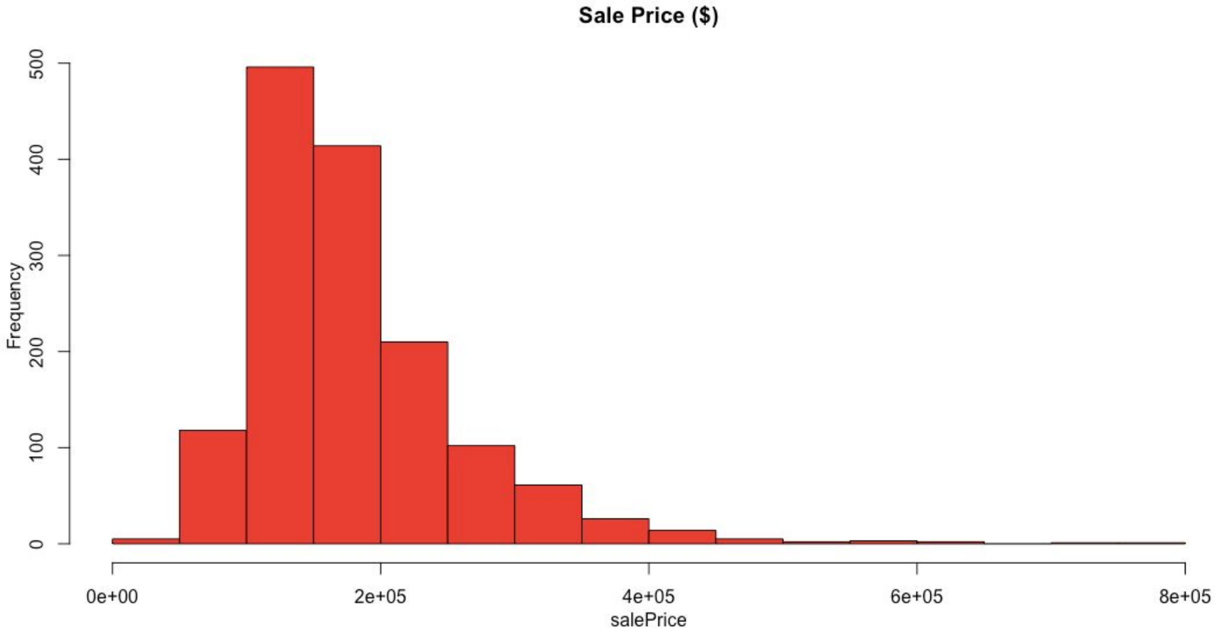
Similarly, we considered GarageCond and GarageQual and assigned numerical values to its levels. We multiplied the two variables and saved the resultant under the title GarageQual and removed the GarageCond variable.

EXPLORATORY DATA ANALYSIS

Correlation

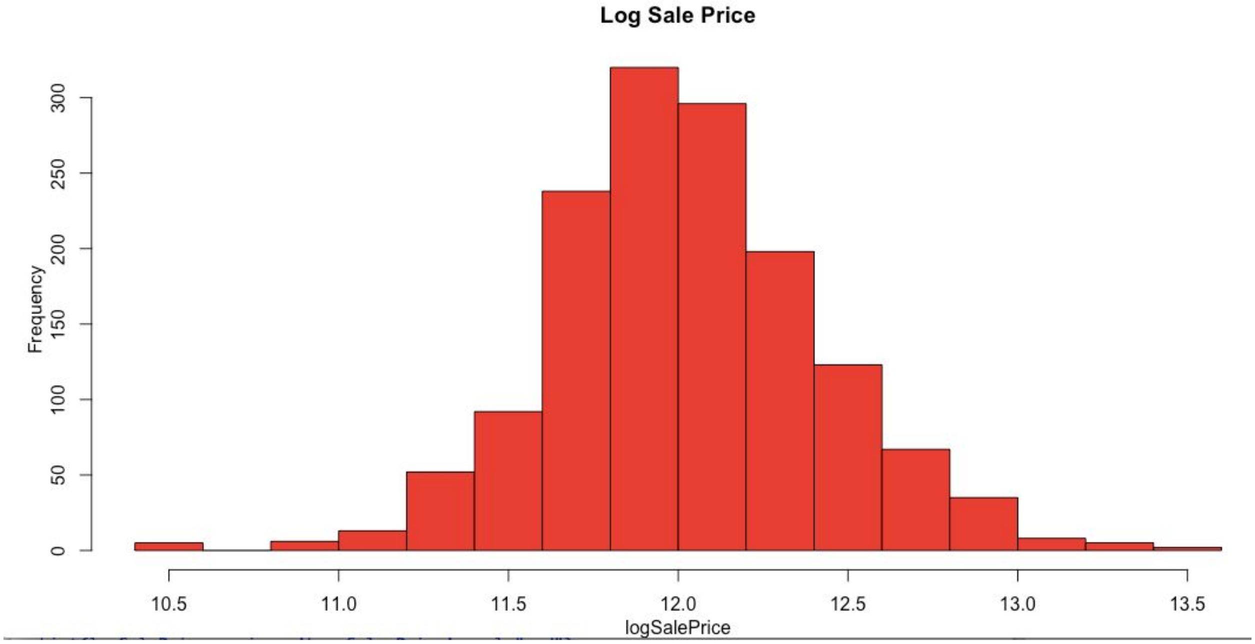
To begin to reduce the number of variables for our model and identify the variables that most influence house sale price, we graphically explored the the distribution of our data as well as any correlations present between variables. First, we examine the distribution of all sale prices in our data set. A histogram of sale prices indicates that the data is not normally distributed and is right skewed.

Figure 1: Histogram of Sale Price



We then transformed our data by plotting the log of each sale price, and found that this data looks more normally distributed.

Figure 2: Histogram of Log of Sale Prices



Next, we explored the relationship between variables and sale price. We split the data into 2 subsets: numerical and categorical. For the numerical data, we calculated Pearson's

Correlation Coefficient between each variable and sale price, and plotted this coefficient (Figures 3 and 4.) Pearson correlation coefficients show the strength of the linear relationship between two data sets, with a coefficient of +1 being a perfect positive correlation and -1 being a perfect negative correlation. The closer to +1 or -1 the coefficient is, the stronger the relationship. Note that even though the Id of each data entry is numeric, we removed it because any correlation with Id is irrelevant.

Figure 3: Correlation coefficients for columns 1-19

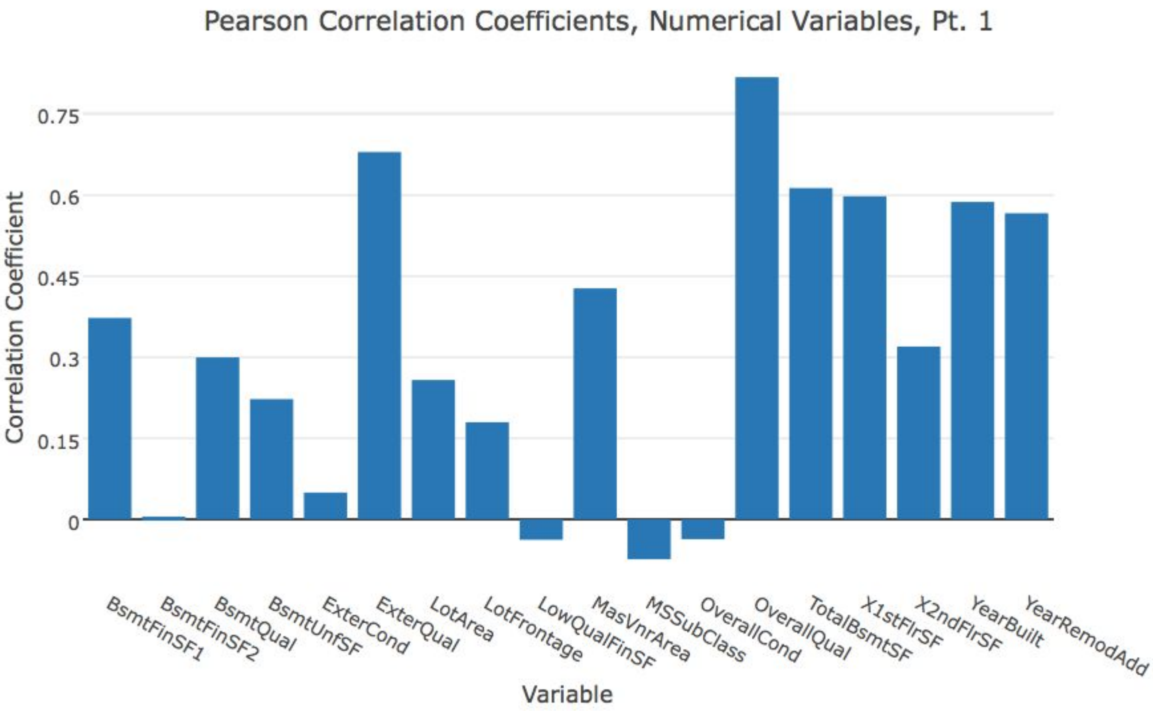
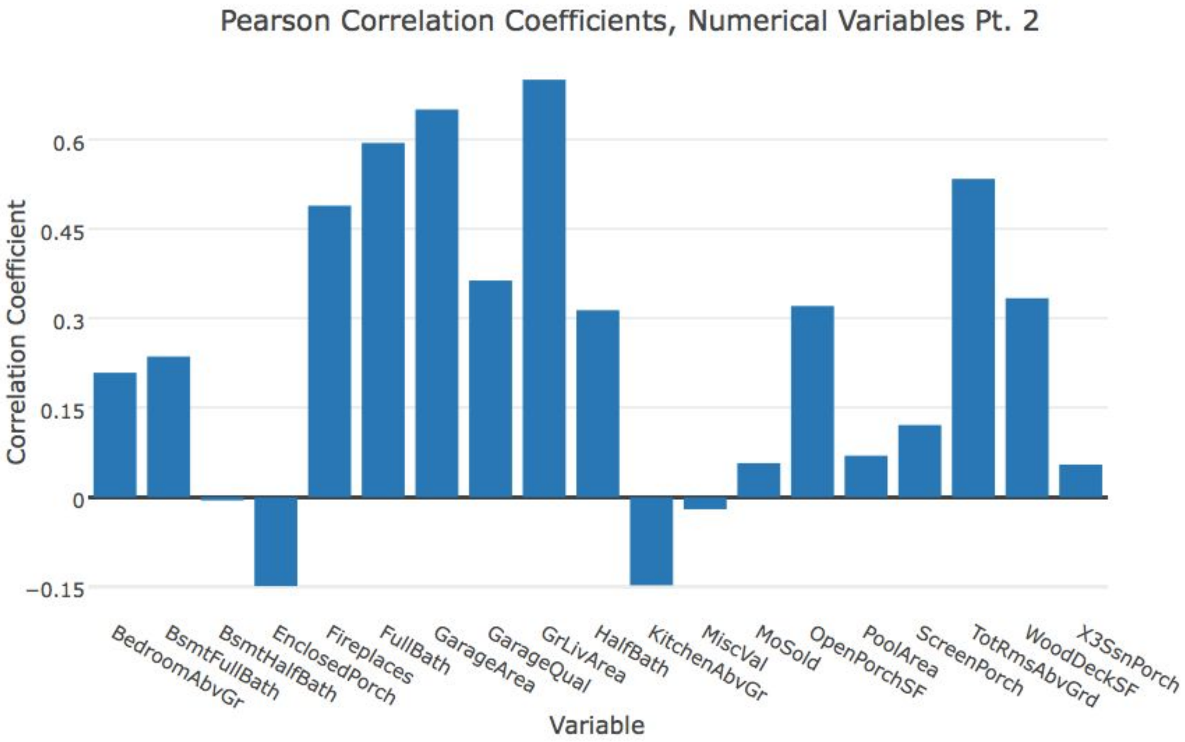


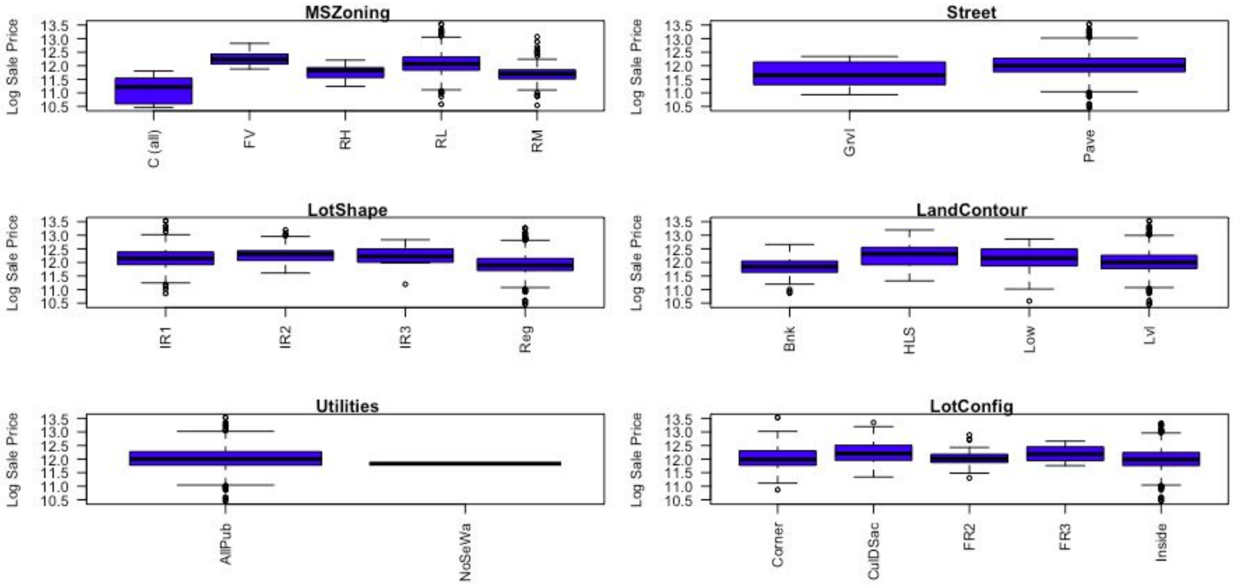
Figure 4: Correlation coefficients for data columns 20-37



The above figures show that there are very few numerical variables that have a negative correlation with house sale price--the majority positively correlate with sale price. Further, the variables that do correlate negatively with sale price have weak negative relationships with sale price: the largest negative correlation coefficient is only -0.15. This figure also suggests that Above Ground Living Area, Garage Area, Overall Quality, and Exterior Quality are strongly positively correlated with sale price. This observation is reasonable, as we would expect buyers to pay more money for higher quality houses and larger houses.

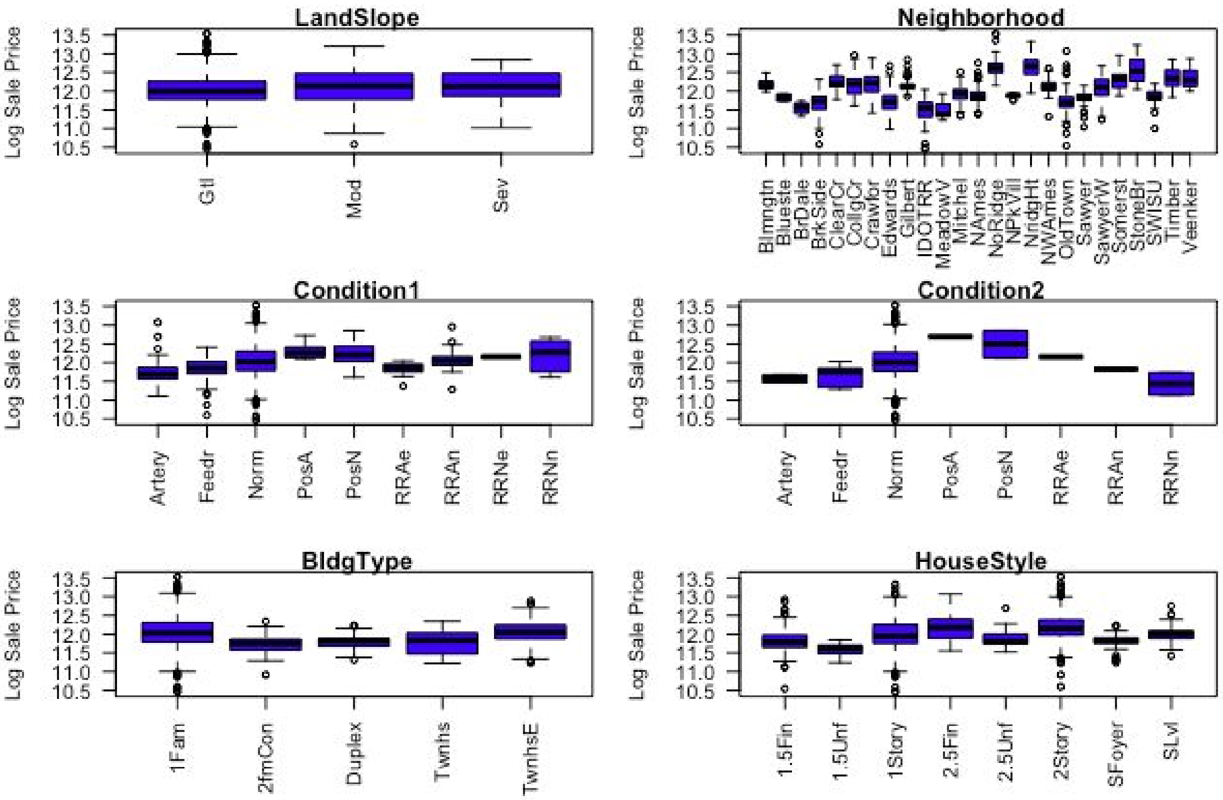
To examine the relationship between the categorical variables in the dataset and sale price, we plotted comparison boxplots. The purpose of doing this is to examine how the distribution of a numerical variable (sale price) changes between categories.

Figure 5: Boxplots of Categorical Variables, Part 1.



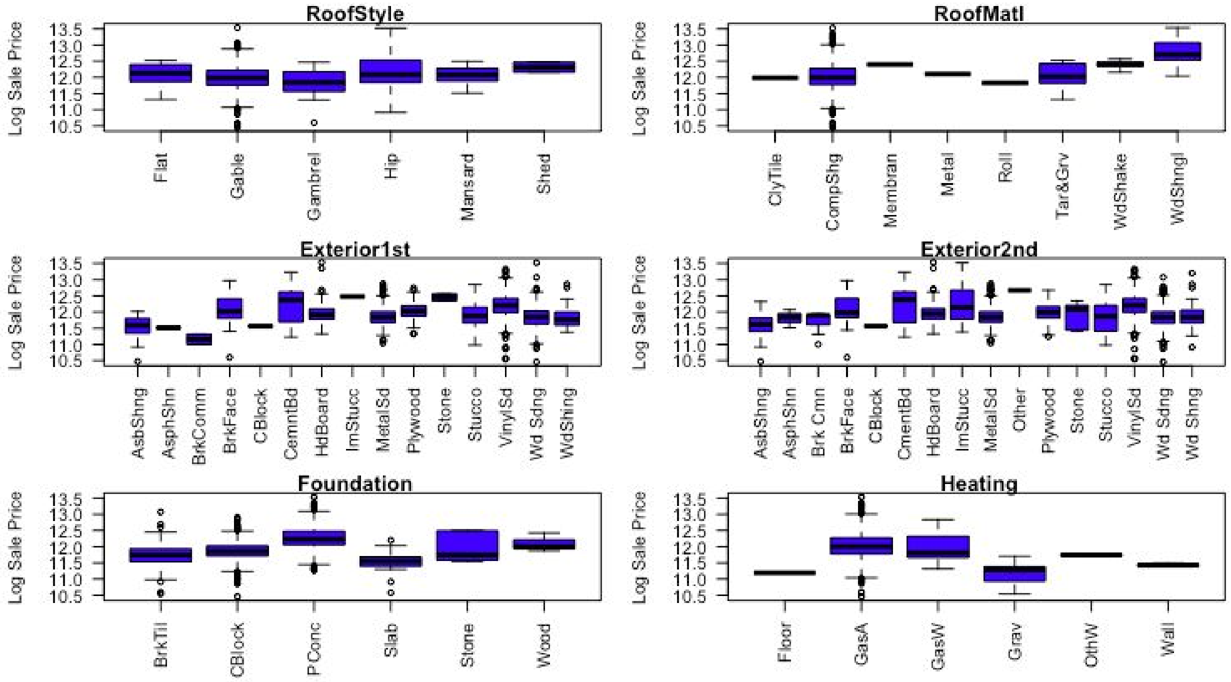
From this group of plots, we can learn that houses in residential low density areas as well as houses on paved streets generally draw higher prices, Additionally, hillside properties (HLS) and low-lying properties(Low) also draw higher prices.

Figure 6: Boxplots of Categorical Variables, Part 2.



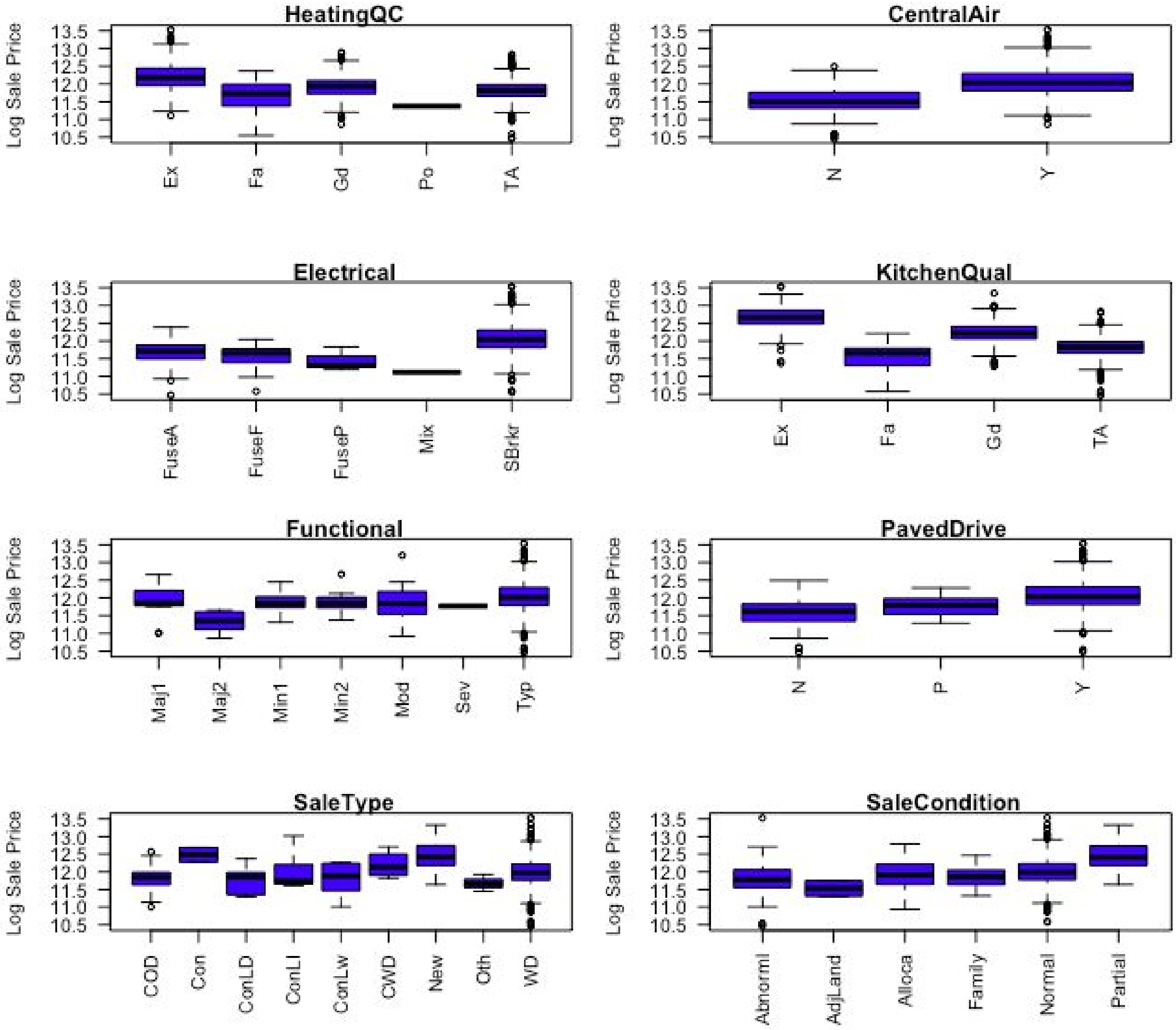
We can also make some observations about relationships with log sale price of a house based on figure 7. The neighborhood-price relationship is detailed in a later section. For the Condition1 plot, the data suggests the highest-priced houses are located within 200 feet of the North-South Railroad (RRn). Condition2 data suggests the houses near parks are associated with higher price (PosN). The HouseStyle plot indicates that 2 story and 2.5 story finished are associated with higher price.

Figure 7: Boxplots of Categorical Variables, Part 3.



Here, we see that high price is associated with hip roofs, wood shingles, and houses with stone, imitation stucco, cement board, and vinyl siding. Additionally, stone and poured concrete foundations as well as as gas forced warm air furnace (GasA) appear to have a positive association with price.

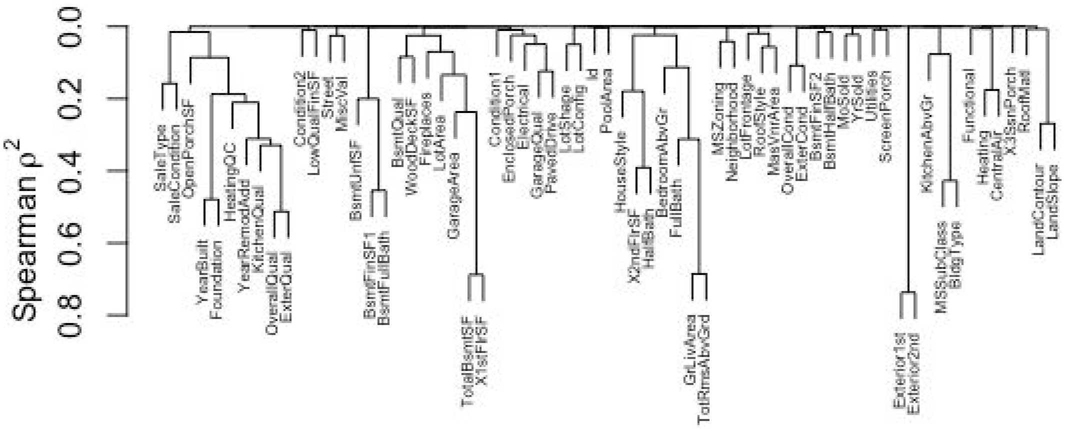
Figure 8: Boxplots of Categorical Variables, Part 4.



Finally, from figure 8, the boxplot for HeatingQC indicates that heating systems in excellent quality(Ex) are associated with higher price. Having a central air system draws a higher price as well, as does a paved drive. As expected, new houses(New) draw higher prices, as well as houses that have a 15% downpayment stipulation(Con). It is important to note that partial sale condition, meaning house was not completed at the time of sale and is most associated with new homes, is also associated with a higher price.

Finally, we examined correlations between all variables except sale price. The purpose of this was to determine if there is any multicollinearity between variables that may negatively influence regression models. We used rank-based Spearman Correlation Coefficient to visualize possible relationships between all variables, both numerical and categorical. Spearman coefficient measures the strength and direction between two ranked variables. We cannot use a Pearson Correlation Coefficient on categorical data because the assumption that all variables are continuous is violated. Unlike a Pearson Coefficient, Spearman does not measure the strength of the linear association between two variables. Rather, Spearman is concerned with monotonic relationships, which are less strict than linear relationships and have two cases: both variables increase together, or one variable increases and the other variable decreases. To plot the below tree, we changed all categorical variables to ranked integers (R method as.integer()) to be able to view relationships between all variables of all types.

Figure 9: Spearman Tree Plot



3 pairs of variables stand out in this tree: Total Basement Square Feet and 1st Floor Square Feet, Above Ground Living Area and Total Rooms Above Ground and, and Exterior 1st and Exterior 2nd. This suggest the members of these pairs are highly related to each other.

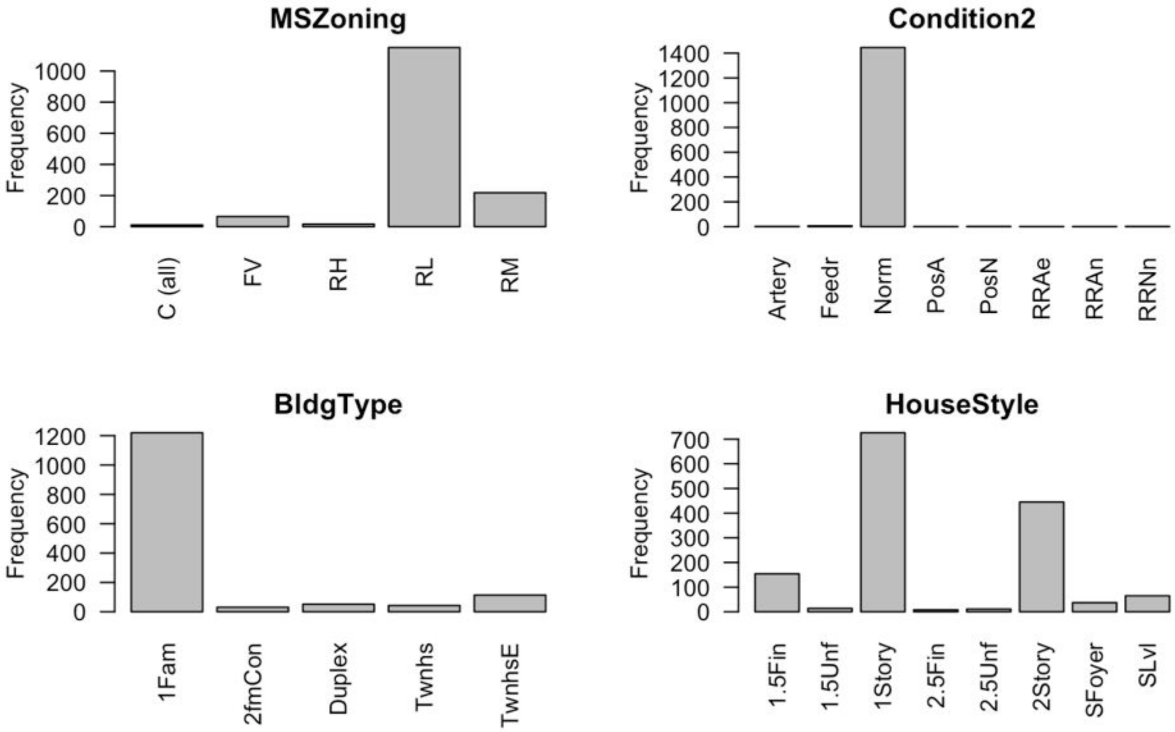
The above explorations in the relationships between variables will allow us to decrease the complexity of our model by identifying key variables important to price prediction and also identify variables that may complicate or have no effect on our model that we can consider discarding from our analysis.

Summary of Key Variables

Categorical variables

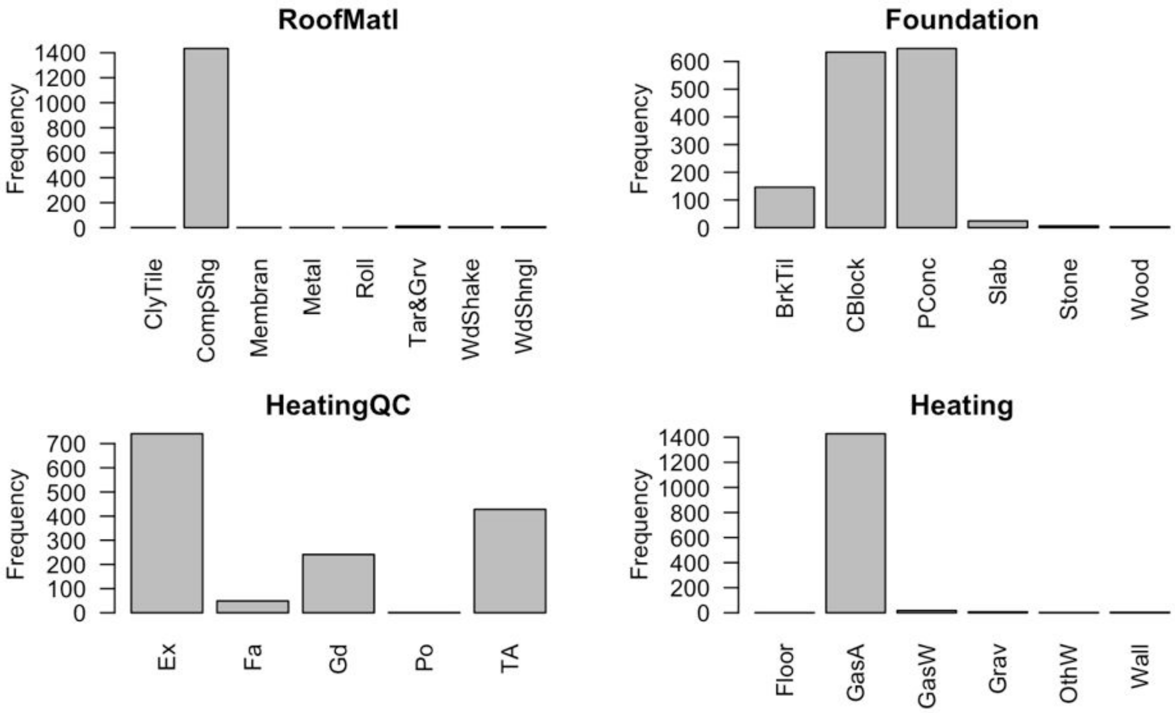
We have identified 27 key categorical variables that are statistically significant to sales price. Barplots are used to visualize each of the categorical variable within their own group.

Figure 10: Histograms of Categorical Data, Part 1



From this block of barplots, we learn that most of the houses in our data set are 1 family homes, 1 story, and are in areas classified as "residential low density" (MSZoning). Condition 2 tells us that most houses are not near a railroad or a park.

Figure 11: Histograms of Categorical Data, Part 2.



Above, we can see there is little to no variety in type of roof among the houses in our dataset. Nearly all houses have standard shingle roofs (see RoofMatl). As for house foundation, there is a mix of houses with brick and tile, cinderblock, and poured concrete foundations. In terms of heating quality (HeatingQC), about half of the houses have been deemed excellent, while the rest are mostly categorized “good” (Gd) and “typical/average” (TA). In terms of heating type, most houses have a forced warm air furnace (GasA).

Figure 12: Number of houses in each neighborhood.

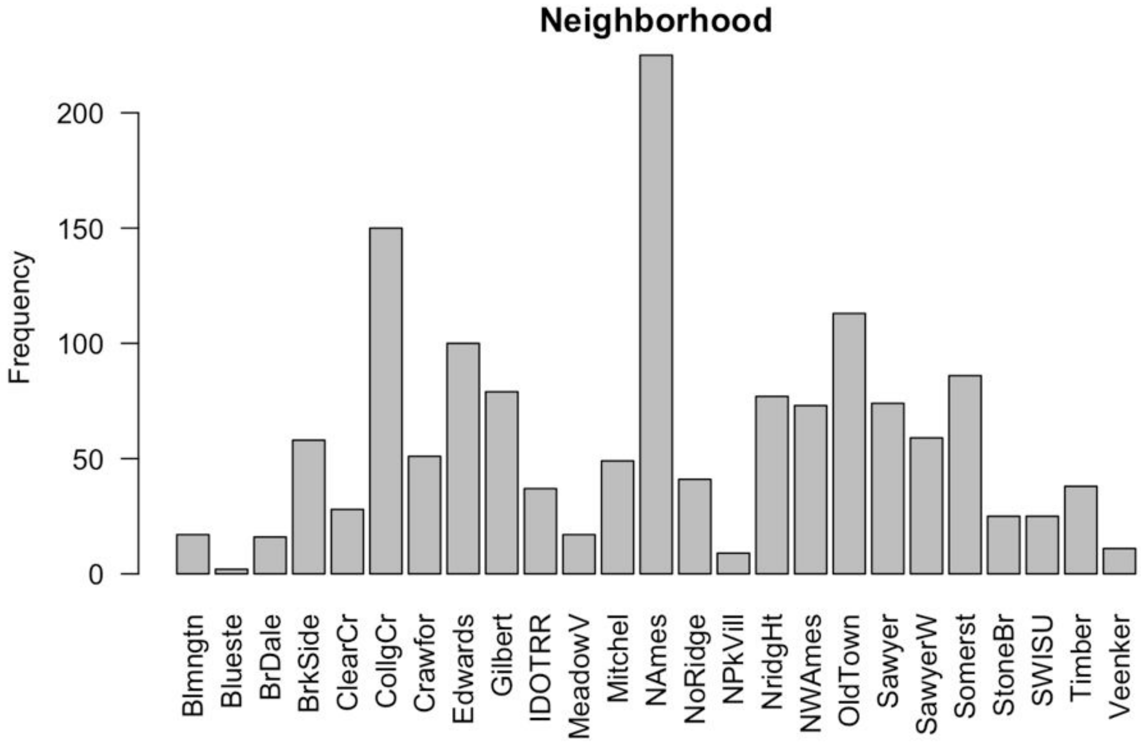
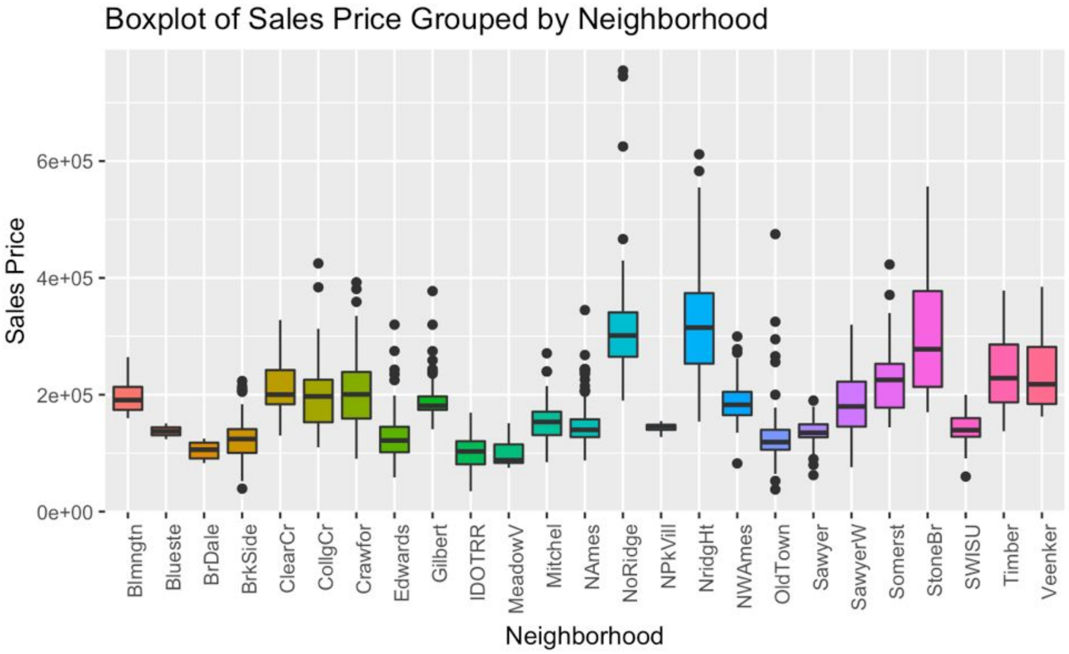


Figure 13: Analysis of Price of Houses in each Neighborhood



Figures 12 and 13 give us insight into neighborhood demographics of Ames. The most frequent neighborhoods in our data set are North Ames and College Creek. The data suggests that the most expensive neighborhoods are Northridge, Northridge Heights, and Stone Brook, while the least expensive neighborhoods are Briardale, Iowa DOT and Railroad, and Meadow Village.

Numerical Variables

From the previous sections we identified 18 statistically significant numerical variables which are summarized in the following table.

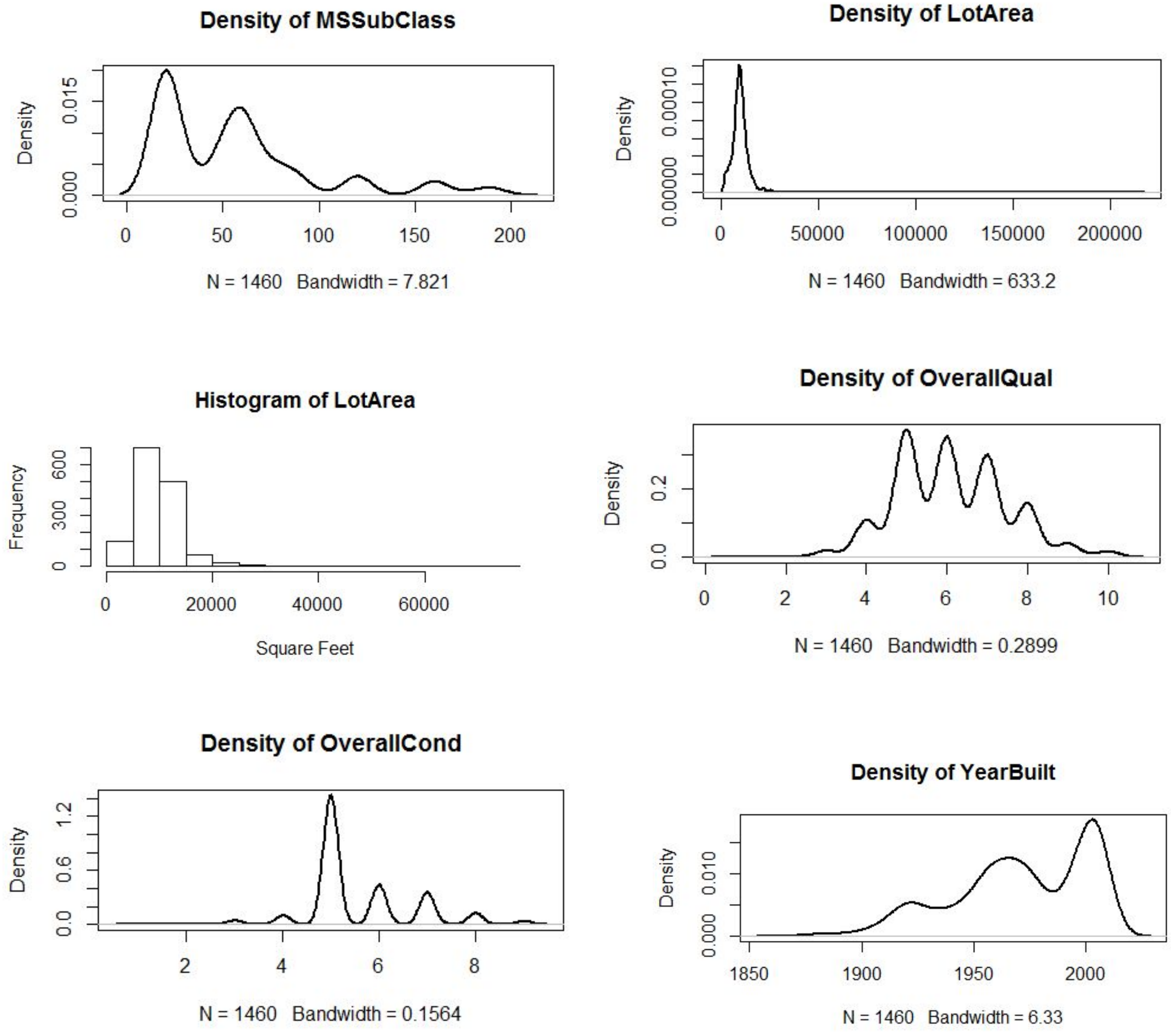
Table 1: Summary Statistics of Key Numerical Variables

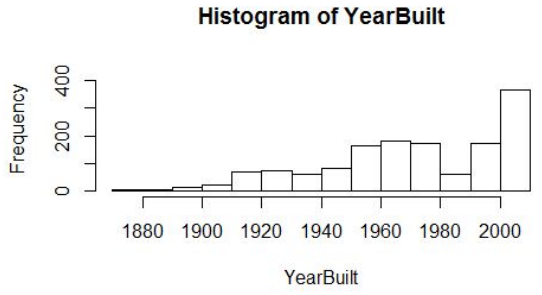
Variables	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
MSSubClass	20.0	20.0	50.0	56.9	70.0	190.0
LotArea	1300	7554	9478	10520	11600	215200
OverallQual	1.000	5.000	6.000	6.099	7.000	10.000
OverallCond	1.000	5.000	5.000	5.575	6.000	9.000
YearBuilt	1872	1954	1973	1971	2000	2010
YearRemodAdd	1950	1967	1994	1985	2004	2010
ExterQual	2.000	3.000	3.000	3.396	4.000	5.000
BsmtQual	0.00	0.00	0.00	46.61	45.00	1350.00
FirstFlrSF	334	882	1087	1163	1391	4692
SecondFlrSF	0	0	0	347	728	2065
BsmtFullBath	0	0	0	0.4253	1.000	3.000
FullBath	0	1.000	2.000	1.565	2.000	3.000
Fireplaces	0	0	1.000	0.613	1.000	3.000
GarageArea	0	334.5	480.0	473.0	576.0	1418.0
GarageQual	0	9.000	9.000	8.392	9.000	25.000
WoodDeckSF	0	0	0	94.24	168.00	857.00

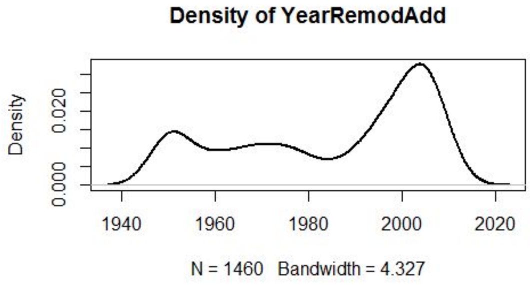
ScreenPorch	0	0	0	15.06	0	480.00
PoolArea	0	0	0	2.759	0	738.00

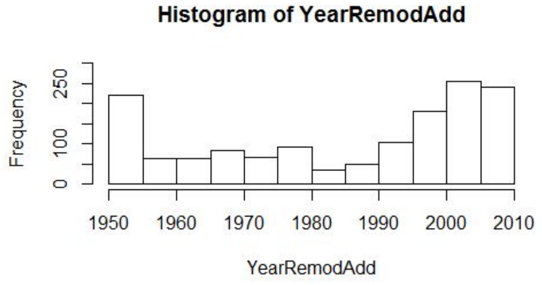
We also utilize density plots and histograms to graphically summarize these key variables. The density plots allow us to visualize the distributions of the variables over their respective continuous intervals. The histograms allow us to extract unique trends from specific variables that tell us more about the housing features.

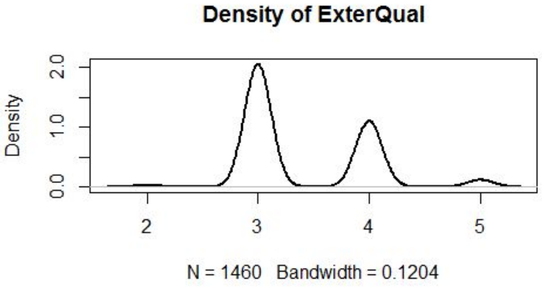
Figure 14: Graphical Analysis of Numerical Data, Part 1.

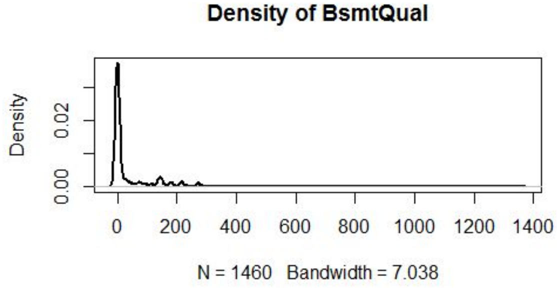


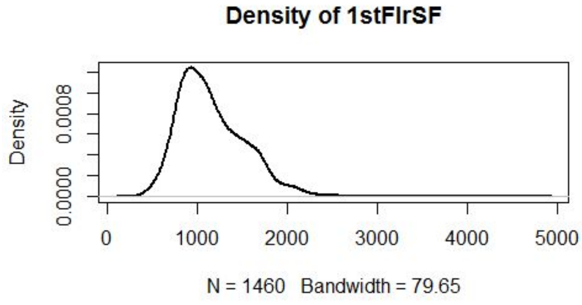

Histogram of YearBuilt

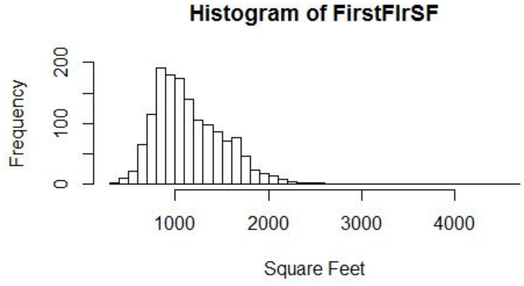

Density of YearRemodAdd

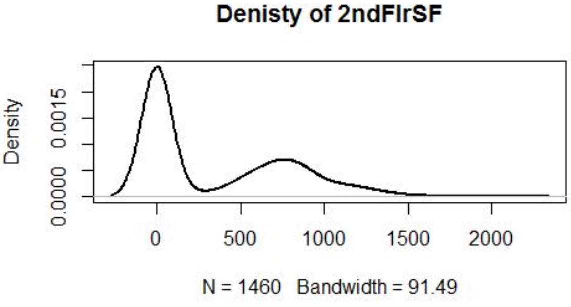

Histogram of YearRemodAdd

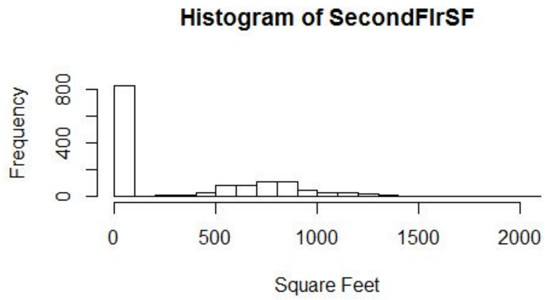

Density of ExterQual

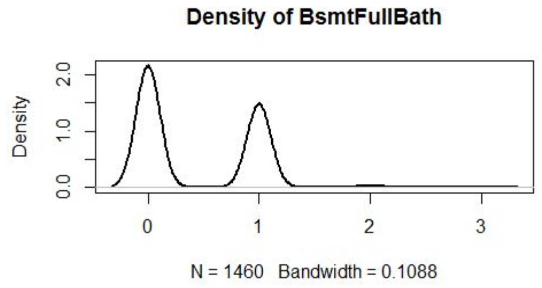

Density of BsmtQual

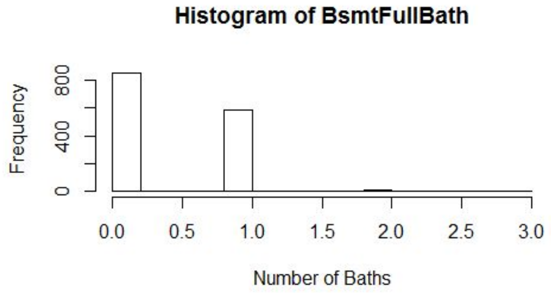

Density of 1stFlrSF

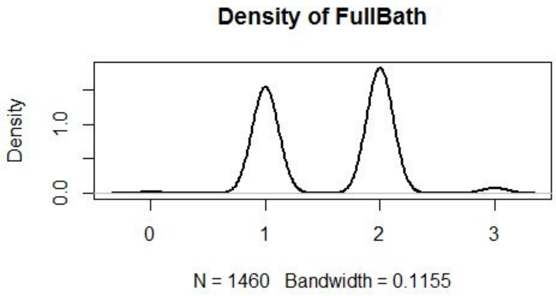

Histogram of FirstFlrSF

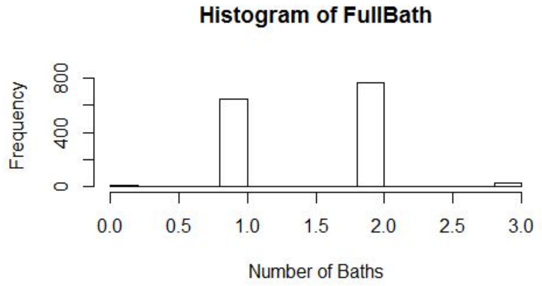

Denisty of 2ndFlrSF

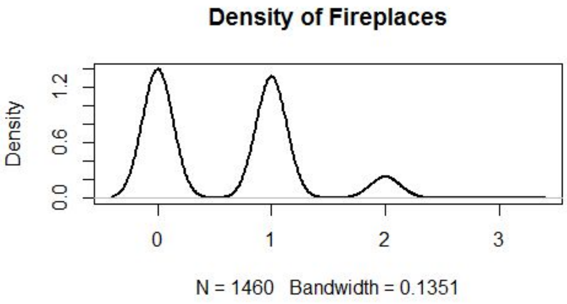

Histogram of SecondFlrSF
Frequency
0
400
800
0
500
1000
1500
2000
Square Feet

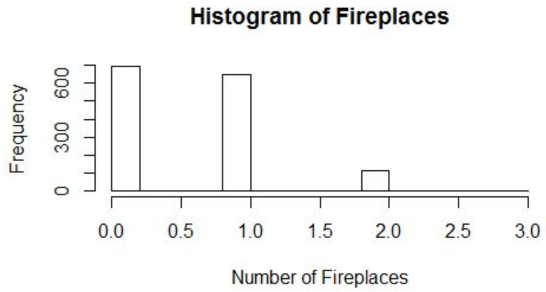

Density of BsmtFullBath
Density
0.0
1.0
2.0
0
1
2
3
N = 1460 Bandwidth = 0.1088

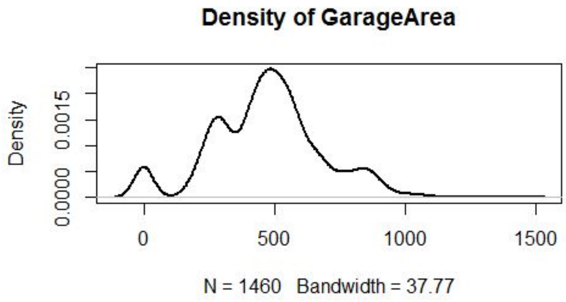

Histogram of BsmtFullBath
Frequency
0
400
800
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Number of Baths

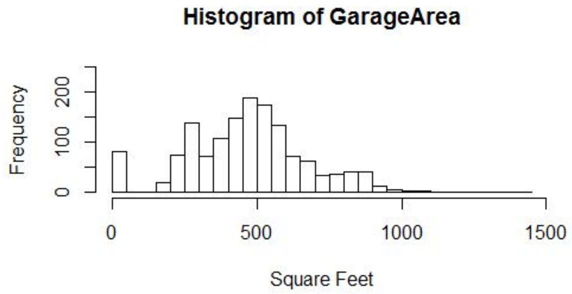

Density of FullBath
Density
0.0
1.0
0
1
2
3
N = 1460 Bandwidth = 0.1155

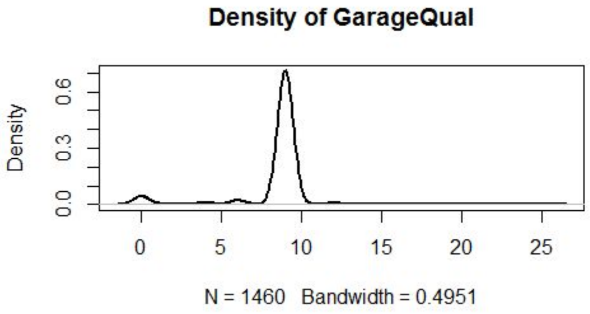

Histogram of FullBath
Frequency
0
400
800
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Number of Baths

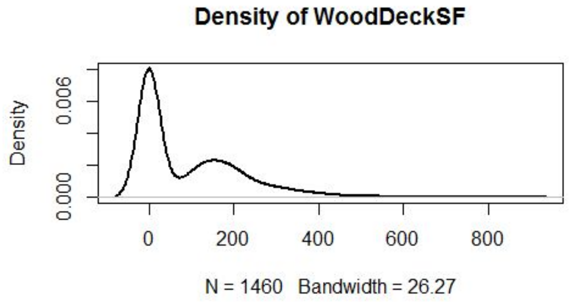

Density of Fireplaces
Density
0.0
0.6
1.2
0
1
2
3
N = 1460 Bandwidth = 0.1351

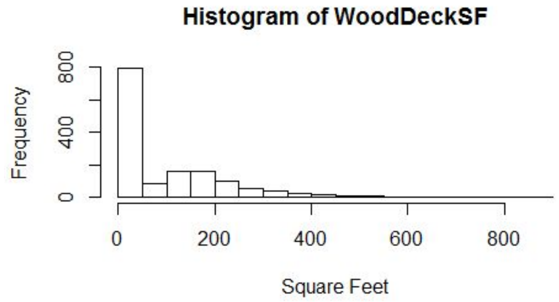

Histogram of Fireplaces
Frequency
0
300
600
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Number of Fireplaces

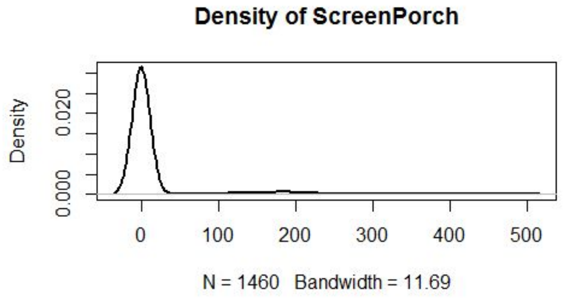

Density of GarageArea
Density
0.0000
0.0015
0
500
1000
1500
N = 1460 Bandwidth = 37.77

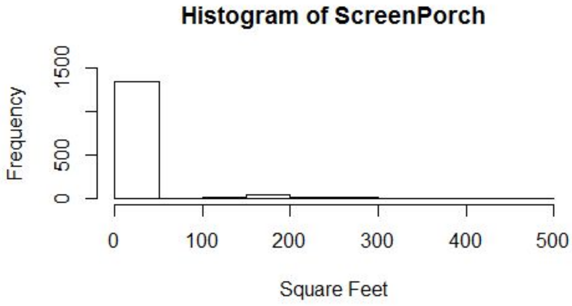


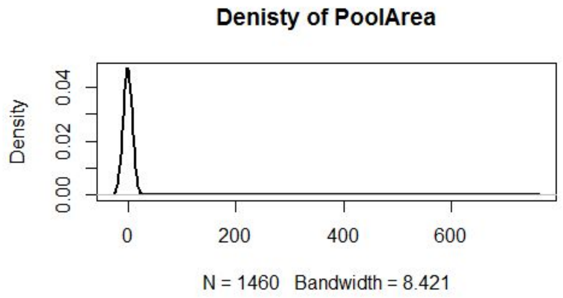


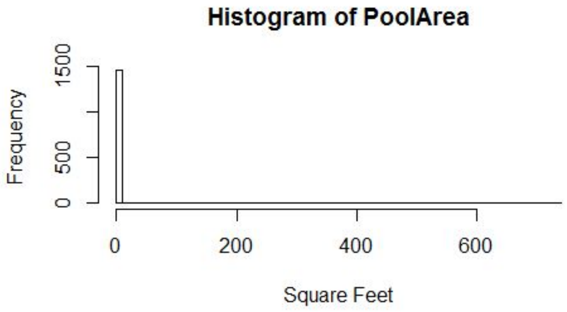












From these various plots we are able to extract several key factors that differentiate the houses in Ames, Iowa. For instance, the histogram of LotArea tells us that the majority and average lot size is about 10,000 square feet. The density plots of OverallQual, OverallCond, and YearBuilt show that the houses in Ames vary greatly by building material, condition, and year constructed respectively. The plots of YearRemodAdd show that remodels were done frequently in the 1950's and resurged around 2000. The plots of SecondFlrSF reveal that the majority of the houses in Ames are one story homes. About half the houses have a bathroom in the basement. About half the houses have two full baths and a fireplace. Looking at the plots of PoolArea shows that pools are very rare in the dataset, likely a result of the colder climate in the city. These features provide us with a baseline understanding of the average home in Ames, Iowa.

RESULTS

For all three models tested below we did the following procedures. The cleaned training dataset (see above section "Data Cleaning" for more information) was split randomly (seed set as 888) into two groups with ⅔ as training data with 973 observations and ⅓ as testing data with 487 observations. We used the 973 training data points to select variables and build the model, and the 487 testing data points to test and validate the model. We then use the testing dataset (test.csv) given by kaggle for submission of our model to the kaggle contest. The square root of the mean errors are commonly used to evaluate how different estimators are from the observed values. In the context of regression, root-mean-square-error (RMSE) can be a measure of standard deviation of the residuals, which has the following equation:

$$\text{RMSE} = \sqrt{\tfrac{1}{n}(\sum_{1}^{n} Y i - \widehat{Y} i)^2}$$

Using the R package "ModelMetrics", we calculated the root-mean-square-error (RMSE), and normalized RMSE (NRMSE) as a validation check (see equation for calculation) of the model.

Equation: NRMSE = RMSE / (max(test$LogSalePrice) - min(test$LogSalePrice))

1. Multivariate Regression

1. Theory

A multivariate regression model allows for the relationship among multiple variables, rather than one single variable as in linear regression, and a dependent variable to be analyzed. For instance given multiple independent variables aX_1, bX_2, cX_3,... and a dependent variable Y, a unit change in X_1 produces an a unit change in Y, holding all else constant. A literal interpretation of the intercept term may not make sense; it's main purpose it to determine the height of the regression line. The independent variables in a multivariate regression model may suffer from collinearity when they are correlated with each other. For example, suppose X_1 and

X_2 are highly, but not perfectly correlated. This complicates the model as it becomes more difficult to determine a meaningful change in Y when X_1 changes holding X_2 constant and vice versa. Perfect multicollinearity, however, does invalidate the model as it makes the interpretation of variables indeterminable. The R^2 in the multivariate model has the same interpretation as in the linear model, it accounts for the fraction of the variance in Y caused by the regression model. This becomes significant with multivariate regression as adding more variables always increases the R^2 value, but is not indicative of a good model incorporating well chosen variables. The adjusted R^2 improves upon the R^2 as it does not necessarily increase as more regressors are added. The adjusted R^2 is captured by the formula adjusted $R^2 = 1 - \frac{(n-1)}{(n-k-1)} \frac{SSR}{TSS}$, where n is the sample size, k is the number of regressors, SSR is the sum of squared residuals, and TSS is the total sum of squares. For the multivariate model to hold certain conditions must be fulfilled. These include the residuals being close to normal, the residuals having constant variability, the residuals being independent, and none of the regressors being perfectly multicollinear with one another. To check that these conditions hold, graphical methods such as Q-Q plots, histograms, and scatterplots can be utilized. Multivariate regression is a key method when trying to build a model to predict an extensive range of outcomes. In the next section we utilize multivariate regression to build a model for the Ames, Iowa housing dataset.

2. Results

We begin testing the dataset by building different models and analyzing which reflects the dataset with the most accuracy.

The first model we build will utilize multivariate linear regression. Analyzing the model with all 79 variables included provides the followings results:

```
Call:
lm(formula = log(SalePrice) ~ ., data = cleandummy)
Residuals:
    Min       1Q   Median      3Q      Max
-0.69226 -0.04325  0.00196  0.04702  0.69226
Residual standard error: 0.09712 on 1130 degrees of freedom

Multiple R-squared:  0.9465,  Adjusted R-squared:  0.9347
F-statistic: 80.57 on 248 and 1130 DF,  p-value: < 2.2e-16
```

Including all variables in the model generates a high R^2 value, however this is not indicative of a good model. The high R^2 is a result of overfitting, as introducing more variables into a model always increases the R^2. To find a better predictive model we turn back to the feature reduction done in previous sections which identified the most statistically significant variables.

Analyzing the variables with the selected variables under the multivariate log-linear model yields the following results:

```
Residuals:
   Min      1Q  Median      3Q     Max 
-378564  -15557   -1766   13688  268055 
Residual standard error: 0.1412 on 920 degrees of freedom
Multiple R-squared:  0.8792,  Adjusted R-squared:  0.8724 
F-statistic: 128.8 on 52 and 920 DF,  p-value: < 2.2e-16
```

The analysis of the selected variables is shown in Appendix 2; these variables were selected using linear regression and were determined to be the most statistically significant. We utilize the log-linear model because, as mentioned in prior sections, the data appears to be skewed and by logging SalePrice, the data becomes more normalized. This model indicates that 87.92% of the variability in SalePrice can be explained by the regression model. Plotting the residuals against the fitted data points in Figure 15 indicates that the data appears to be well modeled under the multivariate model using the selected variables from the feature selection. The plotted residuals form an approximate horizontal band which shows that constant variance and independence hold. Figure 16 shows a Q-Q plot of the residuals, Figure 17 shows a histogram of the residuals, and Figure 18 shows a scatterplot of the residuals; all of which reinforce the validity of the multivariate log-linear regression model. The Q-Q plot does not significantly deviate from the normal distribution, however the data points at the lower end of the distribution and the data points at the upper end of the distribution do deviate from normality. Given the large sample size of the dataset, this deviation is not too disconcerting. The histogram and scatterplot of the regression residuals follows an approximate normal distribution and is symmetric around zero.

Figure 15: Residuals vs Fitted Plot

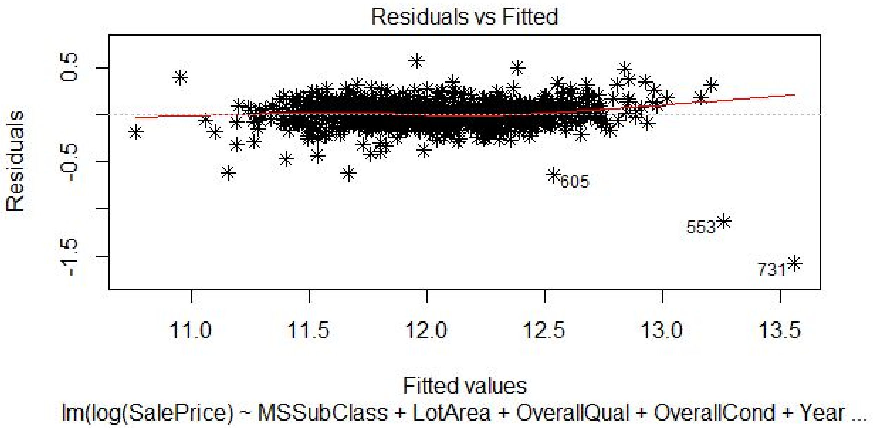


Figure 16: Q-Q Plot of Regression Residuals

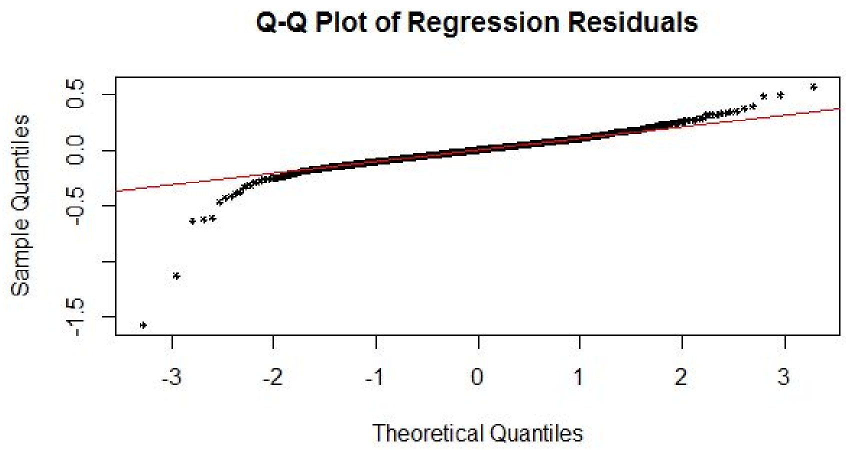


Figure 17: Histogram of Regression Residuals

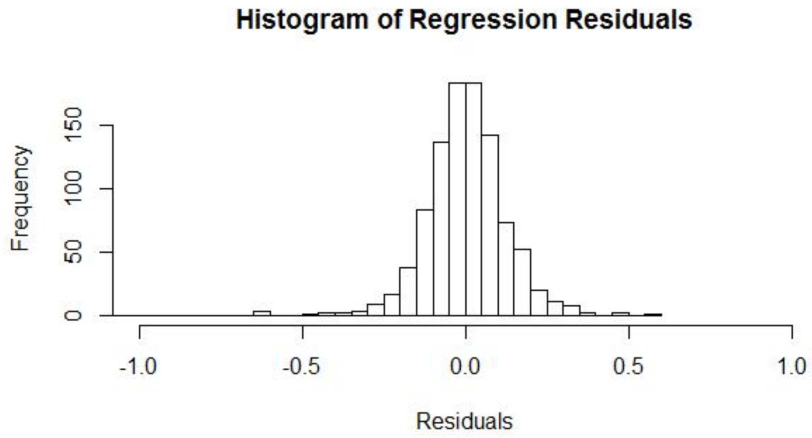
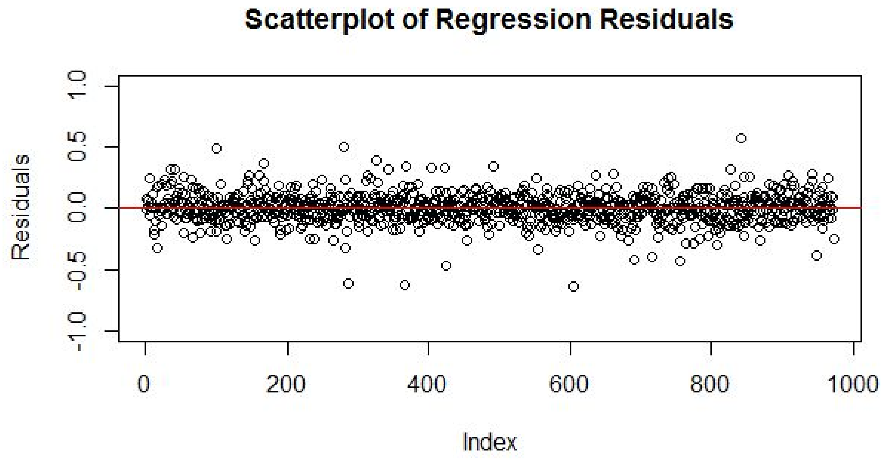


Figure 18: Scatterplot of Regression Residuals



Next we use the test data to see how accurately the multivariate model with the feature selected variables predicts sales price. Table 2 shows the summary statistics of the testing dataset, which consists of 487 observations, and is measured against Table 3, which shows the summary statistics of the actual data. The average sale price derived from the model deviates from the actual average sale price by about $2700. The normalized RMSE was found by taking the mean of the squared residuals and taking its square root; this produced a RMSE value of 0.48. In the following sections we will build alternative models to more accurately predict the sale prices of homes in Ames, Iowa.

Table 2: Summary statistics of modeled data

Summary of modeled data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22310	130400	173200	183600	224200	418700

Table 3: Summary statistics of actual data

Summary of actual data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
34900	130000	163000	180900	214000	755000

2. Random Forest

1. Theory

A random forest consists of many uncorrelated decision trees. Decision trees produce estimates that are low in bias and high in variance, which are two properties that greatly benefit from averaging(bagging). The crux of the random forest algorithm is that it grows *N* trees by sampling *N* cases at random with replacement from the original data. Random forest also differs from simple decision trees in that *m* predictors are selected at random out of *M* total predictors for each tree - with *m* as an unchanging parameter for the entirety of the forest. The parameter, *m,* is positively correlated with correlation and strength, and so an optimal range of *m* exists. The randomization of the training subsets and chosen predictors for each tree makes them uncorrelated. Then, a simple voting system starts when a random forest is given data to estimate. Each tree in the forest will take the data and run it through its leaves and nodes until a classification is achieved. Once all decision trees provide their predictions, the most occurring prediction, or the prediction with the most "votes", is chosen as the prediction of the random forest.

2. *Variable Selection*

For all the the below random forest procedures, we used the R package “randomForest”, which implements Breiman and Cutler's Random Forests for Classification and Regression method. (Please see Appendix 3 for R code.)

First, it was necessary to select the key variables to use in our prediction model. To do this we implemented the random forest decision classification method using 40 numerical variables and 152 categorical variables with binary values in our cleaned dataset. (For more information on the cleaned dataset, please see previous section “Data Cleaning”.) We ordered the all the variables by the metric Mean decrease in accuracy (%MSE), and selected the top 50 most important variables to build our model (see below section). For clarity, only the top 15 are shown in Figure 19 and Table 4.

Figure 19: Top 15 variables selected using random forest. The graphs show the Mean decrease in accuracy (%MSE) and Mean decrease in node impurity (NodePurity).

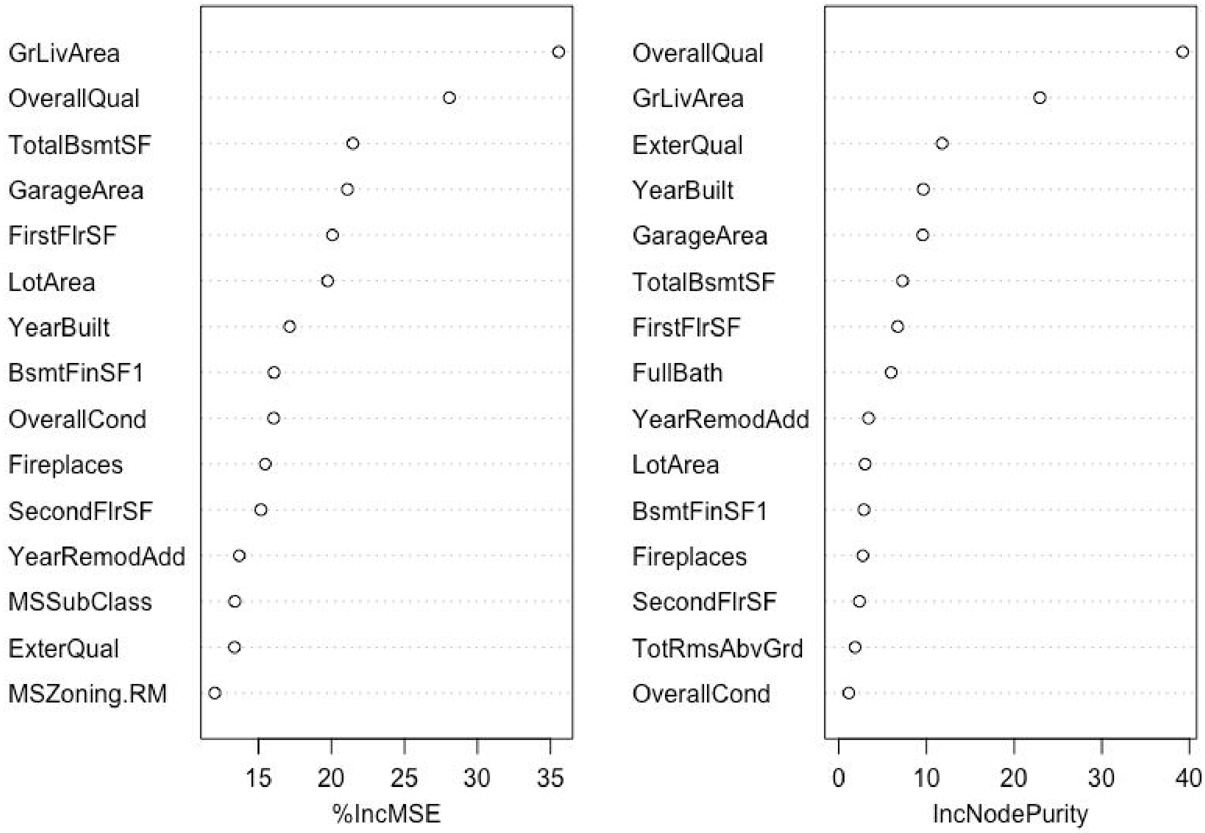


Table 4: Top 15 variables selected using random forest

Variables	**Mean decrease in accuracy (%MSE)**	**Mean decrease in node impurity (NodePurity)**
GrLivArea	35.56710279	22.94910838
OverallQual	28.07172706	39.23361039
TotalBsmtSF	21.47341478	7.269046382
GarageArea	21.09195847	9.58898421
FirstFlrSF	20.06570801	6.733576371
LotArea	19.74085193	2.991338412
YearBuilt	17.14145579	9.653213551
BsmtFinSF1	16.07205615	2.864444751
OverallCond	16.0531503	1.155099945
Fireplaces	15.46307161	2.736911917
SecondFlrSF	15.17722216	2.370891411
YearRemodAdd	13.68314828	3.396436847
MSSubClass	13.39272425	0.618763857
ExterQual	13.3669434	11.78141911
MSZoning.RM	12.00919433	0.632195075

3. *Building the model*

3. 1 Random Forest model formula:

```
LogSalePrice ~ GrLivArea + OverallQual + TotalBsmtSF + GarageArea +
                  FirstFlrSF + LotArea + YearBuilt + BsmtFinSF1 + OverallCond +
                  Fireplaces + SecondFlrSF + YearRemodAdd + MSSubClass +
               ExterQual + MSZoning.RM + FullBath + BsmtQual + TotRmsAbvGrd +
               BedroomAbvGr + MSZoning.RL + KitchenQual.Gd + BsmtUnfSF +
               WoodDeckSF + OpenPorchSF + LotShape.Reg + KitchenQual.TA +
               BsmtFullBath + Exterior1st.VinylSd + Neighborhood.Crawfor +
               MasVnrArea + GarageQual + HalfBath + Neighborhood.NAmes +
               HouseStyle.1Story + CentralAir.Y + HouseStyle.2Story + KitchenAbvGr
               + Foundation.PConc + Exterior2nd.VinylSd + Neighborhood.Edwards +
                  Foundation.CBlock + Neighborhood.MeadowV + HeatingQC.TA +
                  Neighborhood.Gilbert + Condition1.Norm + LandSlope.Mod +
                  Neighborhood.NWAmes + Neighborhood.CollgCr +
               Neighborhood.ClearCr + BldgType.Duplex
```

3.2 Random Forest Analysis:

```
 randomForest(x = train[, key.vars], y = train$LogSalePrice, xtest = test[,      key.vars],
ytest = test$LogSalePrice, ntree = 300, keep.forest = T,      corr.bias = F, formula =
rf.form.key, data = train)
                    Type of random forest: regression
                          Number of trees: 300
     No. of variables tried at each split: 16

               Mean of squared residuals: 0.01838725
                         % Var explained: 88.22
                           Test set MSE: 0.02
                         % Var explained: 86.6
```

Figure 20: Random Forest mean squared error (MSE) versus the number of trees.

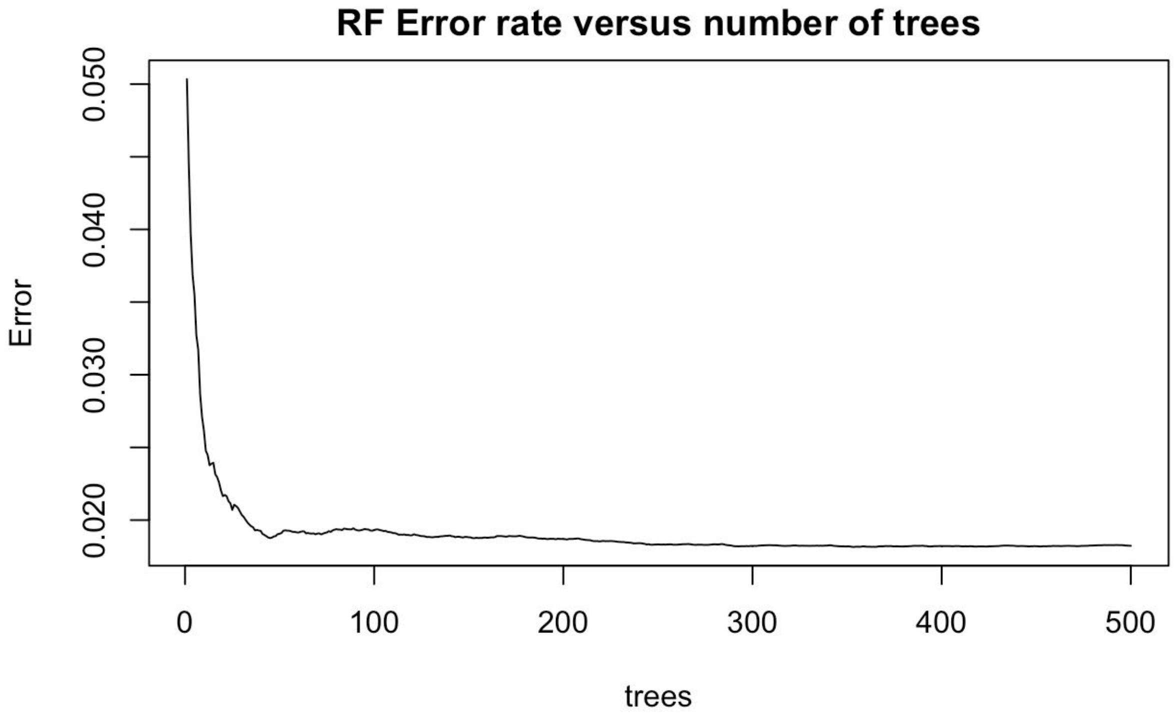
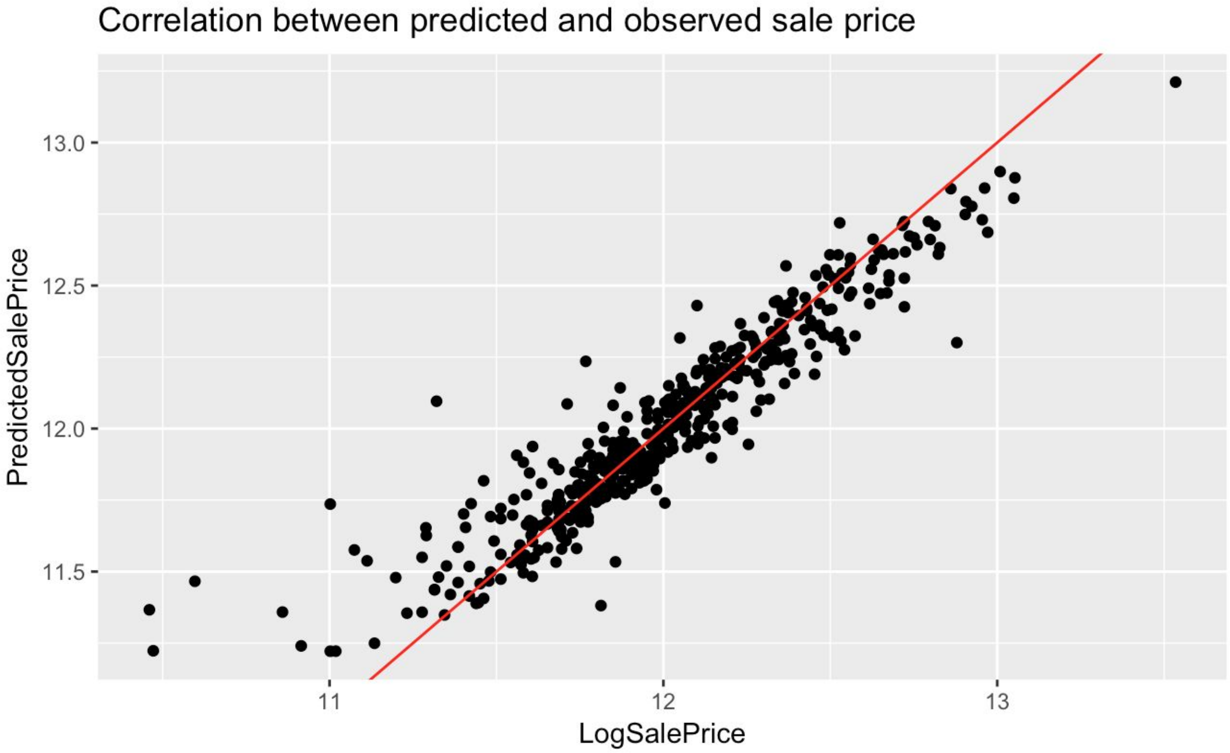


Figure 21: Correlation between predicted and observed sale price for 487 testing data points



3.3 Spearman's rank correlation rho

Data: testing set - LogSalePrice and PredictedSalePrice

N: 487 data points

S = 976,540

P-value < 2.2e-16

Alternative hypothesis: true rho is not equal to 0

Sample estimates: rho = 0.949271

3.4 Root Mean Squared Error

RMSE = 0.1487058

NRMSE = 0.04837172 (RMSE / (13.53447 - 10.46024)

After the 50 key features of the model were selected, we build a regression formula (3.1 above) and implemented the Random Forest regression method with 300 trees (3.2 above) in R. We choose 300 trees because the mean squared error rate plateaued after 300 trees and did not seem to decrease with more trees (Figure 20). We conducted a Spearman correlation test between predicted sale price and actual sale price in the testing dataset, and found a significantly strong relationship between the two (3.3) . The reason we used the Spearman correlation test is because our sale prices do not follow the Pearson assumptions of a normal distribution. We also calculated the RMSE and NRSME values to determine the spread of the predicted sale prices (3.4). We found the RMSE values to indicate the predicted sale prices in the test data were concentrated around the regression line, meaning our random forest model is quite successful.

3. Gradient Boosting

1. Theory

Considering a probability distribution P(x,y), where y is the output variable and x is a vector of input variables. The objective of most supervised learning techniques is to use a training set $\{(x_1, y_1), ..., (x_n, y_n)\}$ to find the approximation of the loss function that minimizes the expected value of the loss function $L(y, F(x))$.

$$\widehat{F} = arg\ min_F E_{x,y}[L(y, F(x))]$$

The special feature of gradient boosting is that it takes a real-valued y and seeks an approximation $\widehat{F}(x)$ in the form of weighted sum of functions $h_i(x)$ from some class which is called base learners:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x)\ +\ constant$$

It uses the method of empirical risk minimization to find an approximation $\widehat{F}(x)$ that minimizes the average value of the loss function on the training set. It starts with a model having a constant function $F_0(x)$ and increments it in a greedy fashion by using the following equations:

$$F_0(x) = arg\ min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

$$F_m(x) = F_{m-1}(x) + arg\ min_f \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + f(x_i))$$

where function f is restricted from the class of base learner functions.

From our lectures, we learned that least absolute deviation (LAD) can also be estimators other than least square errors which are the maximum likelihood estimators (MLE) when errors are normally distributed. LAD actually is a MLE for Laplace distribution, also called as double -exponential distribution. Laplace distribution was used in gradient boosting to optimize the absolute loss function. The advantage to select absolute loss function is to avoid tendency of dominance effect from outlier.

2. Basic Tuning Parameters

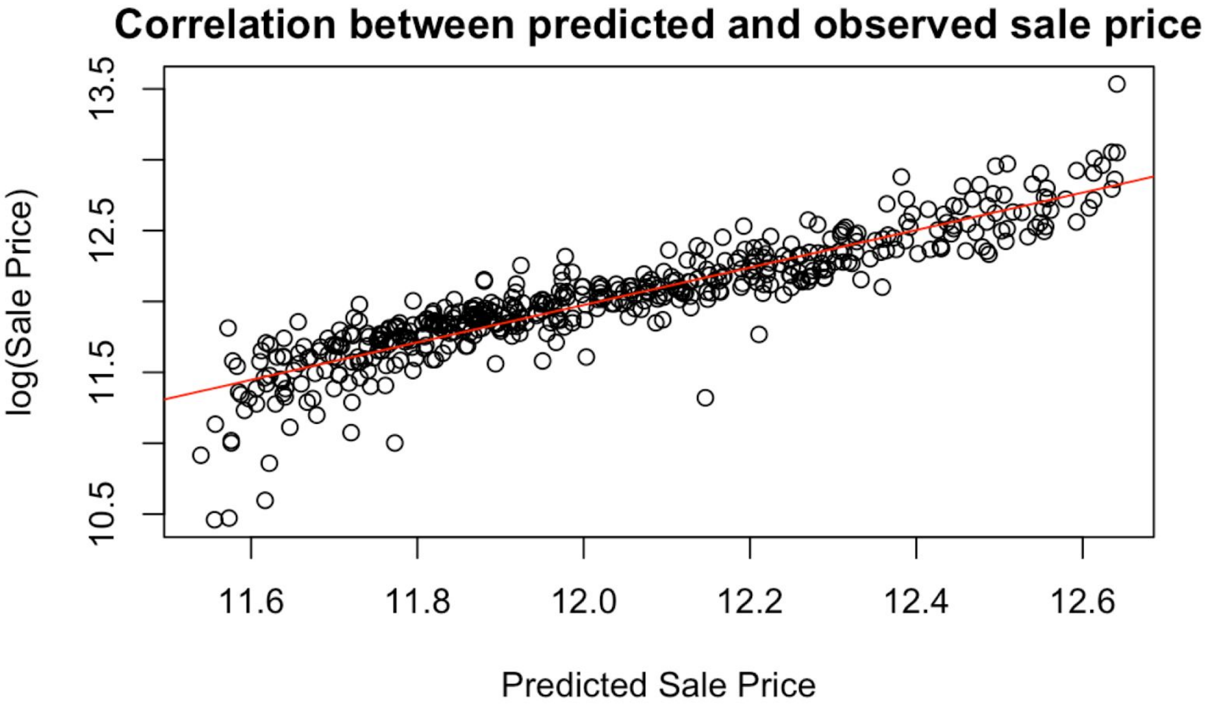
In gradient boosting, there are many model tuning parameters which can affect the performance of the model. A shrinkage parameter represents the learning rate for each tree in the expansion. The maximum shrinkage was determined to be 0.01 for our dataset. Interaction depth stands for the maximum depth of variable iterations, which was determined to be 8 for our training dataset. We started the total number of trees from 100 and then gradually increased.

The larger number of trees to be run will cause the computing time to be longer. In addition, higher number of trees may also lead to overfitting. In the end, 300 was chosen to be the number of trees for lower RMSE values. All other parameters remain as default. Definitely, there are ways to better optimize tuning parameters. We then used previously selected key variables in the correlation section to fit into gradient boosting model, which eventually gives 0.194 as RMSE value.

3. Gradient Boosting Analysis

The reason for choosing gradient boosting is that it is a supervised learning algorithm with fewer hyperparameters to tune to. Also because we tried to use the fact that a well tuned gradient boosting model can outperform the random forest model. But we couldn't find the exact set of parameters to achieve this.

Figure 22: Gradient Boosting Method: Correlation between predicted and observed sale price for 487 testing data points



3. 1 Spearman's rank correlation rho

```
data:  prediction and log(test$SalePrice)
S = 1162500, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:rho = 0.9396097
```

We further examined the correlation between predicted sale price and actual sale price in the testing dataset. Spearman correlation test shows they are significantly correlated.

CONCLUSION

Data imputation, feature engineering, and feature selection were of utmost importance for the ‘House Prices’ dataset. Feature selection in particular resembled the real life process of purchasing a house. Aspects of a home include square footage, number of bedrooms and bathrooms, presence of a pool, overall condition, neighborhood, and many other intrinsic and extrinsic features. Feature selection comes into play because the selling price of a home is based on the willingness of a customer to compromise or not compromise on certain aspects of a house. This means that each buyer of a home weighs(prices) the importance of each attribute of a home differently. Feature selection reduces the amount of variables in our model, speeds up the training process, allows us to obtain better average predictions, and makes our models more interpretable. Data imputation was also used in efforts of obtaining a better dataset. Afterwards, we ran linear regression on all the variables to obtain the coefficients and significance factors. We then chose the variables that were the most significant.

The three pre-modeling processes alone allowed us to obtain good performance on a simple multivariate linear regression model, with Root Mean Squared Error of 0.48 for our linear regression model. In pursuit of a better prediction accuracy we explored non-linear models that could better model the non-linear relationships between the attributes and sale price. This led us to explore the random forest algorithm and stochastic gradient boosting. While stochastic gradient boosting decreased Root Mean Squared Error to 0.194, random forest performed the best out of the three models we trained, with a Root Mean Squared Error of 0.1487058. We can attribute this to the fact that the algorithm itself is robust. It is not sensitive to specific hyperparameters and its predictions tend to be the least biased and variating due to its bootstrap and bagging process. Thus, based on our results, we believe that random forest is a good model to estimate house prices in Ames, Iowa data because its predictions are only off by an average of roughly 15%

However, our model is not perfect as our best-performing model had non-zero error. In our background research, we found many variables that in past studies have proven to have a significant influence on house price. For example, buyers are willing to pay more for houses in high quality school districts. Further data on the location and quality of local schools may refine our model. Additionally, statistics like local land availability and rental vs. ownership rates have potential to improve our model, because prices increase when demand is higher than supply, and if the price-to-rent ratio is high, buyers may be more likely to rent than buy homes. Finally, the housing market operates in a cycle. Our dataset had a relatively small range of years sold--4, between 2006 and 2010. That is not enough data to analyze market trends, and it is especially important to note that this range falls during the massive housing crisis of the 2000's. Our model could improve with more data about sale price and year sold so we could examine past market cycles and predict future cycles. Further investigations into improving predictions of house prices in Ames should include addition of more data to the current dataset.

APPENDIX

1. Detailed Data Description

MSSubClass: Identifies the type of dwelling involved in the sale. (Categorical)

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale. (Categorical)

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density

LotFrontage: Linear feet of street connected to property (Numerical)

LotArea: Lot size in square feet (Numerical)

Street: Type of road access to property (Categorical)

Grvl	Gravel
Pave	Paved

Alley: Type of alley access to property (Categorical)

Grvl	Gravel
Pave	Paved
NA	No alley access

LotShape: General shape of property (Categorical)

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular

LandContour: Flatness of the property (Categorical)

Lvl	Near Flat/Level
Bnk	Banked - Quick and significant rise from street grade to building
HLS	Hillside - Significant slope from side to side
Low	Depression

Utilities: Type of utilities available (Categorical)

AllPub	All public Utilities (E,G,W,& S)
NoSewr	Electricity, Gas, and Water (Septic Tank)
NoSeWa	Electricity and Gas Only
ELO	Electricity only

LotConfig: Lot configuration (Categorical)

Inside	Inside lot
Corner	Corner lot
CulDSac	Cul-de-sac
FR2	Frontage on 2 sides of property
FR3	Frontage on 3 sides of property

LandSlope: Slope of property (Categorical)

Gtl	Gentle slope
Mod	Moderate Slope
Sev	Severe Slope

Neighborhood: Physical locations within Ames city limits (Categorical)

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames

OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

Condition1: Proximity to various conditions (Categorical)

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present) (Categorical)

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

BldgType: Type of dwelling (Categorical)

1Fam	Single-family Detached
2FmCon	Two-family Conversion; originally built as one-family dwelling
Duplx	Duplex
TwnhsE	Townhouse End Unit

TwnhsI	Townhouse Inside Unit

HouseStyle: Style of dwelling (Categorical)

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

OverallQual: Rates the overall material and finish of the house (Categorical)

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

OverallCond: Rates the overall condition of the house (Categorical)

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

YearBuilt: Original construction date (Numerical)

YearRemodAdd: Remodel date (same as construction date if no remodeling) (Numerical)

RoofStyle: Type of roof (Categorical)

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

RoofMatl: Roof material (Categorical)

ClyTile	Clay or Tile
CompShg	Standard (Composite) Shingle
Membran	Membrane
Metal	Metal
Roll	Roll
Tar&Grv	Gravel & Tar
WdShake	Wood Shakes
WdShngl	Wood Shingles

Exterior1st: Exterior covering on house (Categorical)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone

Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material) (Categorical)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

MasVnrType: Masonry veneer type (Categorical)

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone

MasVnrArea: Masonry veneer area in square feet (Numerical)

ExterQual: Evaluates the quality of the material on the exterior (Categorical)

Ex	Excellent
Gd	Good

TA	Average/Typical
Fa	Fair
Po	Poor

ExterCond: Evaluates the present condition of the material on the exterior (Categorical)

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

Foundation: Type of foundation (Categorical)

BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Contrete
Slab	Slab
Stone	Stone
Wood	Wood

BsmtQual: Evaluates the height of the basement (Categorical)

Ex	Excellent (100+ inches)
Gd	Good (90-99 inches)
TA	Typical (80-89 inches)
Fa	Fair (70-79 inches)
Po	Poor (<70 inches)
NA	No Basement

BsmtCond: Evaluates the general condition of the basement (Categorical)

Ex	Excellent
Gd	Good
TA	Typical - slight dampness allowed
Fa	Fair - dampness or some cracking or settling
Po	Poor - Severe cracking, settling, or wetness
NA	No Basement

BsmtExposure: Refers to walkout or garden level walls (Categorical)

Gd	Good Exposure
Av	Average Exposure (split levels or foyers typically score average or above)
Mn	Mimimum Exposure
No	No Exposure
NA	No Basement

BsmtFinType1: Rating of basement finished area (Categorical)

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF1: Type 1 finished square feet (Numerical)

BsmtFinType2: Rating of basement finished area (if multiple types) (Categorical)

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF2: Type 2 finished square feet (Numerical)

BsmtUnfSF: Unfinished square feet of basement area (Numerical)

TotalBsmtSF: Total square feet of basement area (Numerical)

Heating: Type of heating (Categorical)

Floor	Floor Furnace

GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace

HeatingQC: Heating quality and condition (Categorical)

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

CentralAir: Central air conditioning (Categorical)

N	No
Y	Yes

Electrical: Electrical system (Categorical)

SBrkr	Standard Circuit Breakers & Romex
FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix	Mixed

1stFlrSF: First Floor square feet (Numerical)

2ndFlrSF: Second floor square feet (Numerical)

LowQualFinSF: Low quality finished square feet (all floors) (Numerical)

GrLivArea: Above grade (ground) living area square feet (Numerical)

BsmtFullBath: Basement full bathrooms (Numerical)

BsmtHalfBath: Basement half bathrooms (Numerical)

FullBath: Full bathrooms above grade (Numerical)

HalfBath: Half baths above grade (Numerical)

Bedroom: Bedrooms above grade (does NOT include basement bedrooms) (Numerical)

Kitchen: Kitchens above grade (Numerical)

KitchenQual: Kitchen quality (Categorical)

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms) (Numerical)

Functional: Home functionality (Assume typical unless deductions are warranted) (Categorical)

Typ	Typical Functionality
Min1	Minor Deductions 1
Min2	Minor Deductions 2
Mod	Moderate Deductions
Maj1	Major Deductions 1
Maj2	Major Deductions 2
Sev	Severely Damaged
Sal	Salvage only

Fireplaces: Number of fireplaces (Numerical)

FireplaceQu: Fireplace quality (Categorical)

Ex	Excellent - Exceptional Masonry Fireplace
Gd	Good - Masonry Fireplace in main level
TA	Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa	Fair - Prefabricated Fireplace in basement
Po	Poor - Ben Franklin Stove

NA	No Fireplace

GarageType: Garage location (Categorical)

2Types	More than one type of garage
Attchd	Attached to home
Basment	Basement Garage
BuiltIn	Built-In (Garage part of house - typically has room above garage)
CarPort	Car Port
Detchd	Detached from home
NA	No Garage

GarageYrBlt: Year garage was built (Numerical)

GarageFinish: Interior finish of the garage (Categorical)

Fin	Finished
RFn	Rough Finished
Unf	Unfinished
NA	No Garage

GarageCars: Size of garage in car capacity (Numerical)

GarageArea: Size of garage in square feet (Numerical)

GarageQual: Garage quality (Categorical)

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor
NA	No Garage

GarageCond: Garage condition (Categorical)

Ex	Excellent
Gd	Good
TA	Typical/Average

Fa	Fair
Po	Poor
NA	No Garage

PavedDrive: Paved driveway (Categorical)

Y	Paved
P	Partial Pavement
N	Dirt/Gravel

WoodDeckSF: Wood deck area in square feet (Numerical)

OpenPorchSF: Open porch area in square feet (Numerical)

EnclosedPorch: Enclosed porch area in square feet (Numerical)

3SsnPorch: Three season porch area in square feet (Numerical)

ScreenPorch: Screen porch area in square feet (Numerical)

PoolArea: Pool area in square feet (Numerical)

PoolQC: Pool quality (Categorical)

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
NA	No Pool

Fence: Fence quality (Categorical)

GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

MiscFeature: Miscellaneous feature not covered in other categories (Categorical)

Elev	Elevator
Gar2	2nd Garage (if not described in garage section)
Othr	Other
Shed	Shed (over 100 SF)
TenC	Tennis Court
NA	None

MiscVal: $Value of miscellaneous feature (Numerical)

MoSold: Month Sold (MM) (Numerical)

YrSold: Year Sold (YYYY) (Numerical)

SaleType: Type of sale (Categorical)

WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold
COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest
ConLI	Contract Low Interest
ConLD	Contract Low Down
Oth	Other

SaleCondition: Condition of sale (Categorical)

Normal	Normal Sale
Abnorml	Abnormal Sale - trade, foreclosure, short sale
AdjLand	Adjoining Land Purchase
Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family	Sale between family members
Partial	Home was not completed when last assessed (associated with New Homes)

2. Variable Selection

39 NUMERICAL VARS REGRESSION ONLY

```
Call:
glm(formula = log(SalePrice) ~ ., family = gaussian, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.98094  -0.06482   0.00336   0.07233   0.51366

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.861e+01  5.952e+00   3.128 0.001799 **
MSSubClass    -7.154e-04  1.141e-04  -6.271 4.74e-10 ***
LotFrontage   -1.537e-04  1.213e-04  -1.267 0.205424
LotArea        1.673e-06  4.312e-07   3.881 0.000109 ***
OverallQual    7.812e-02  5.308e-03  14.718  < 2e-16 ***
OverallCond    4.662e-02  4.589e-03  10.158  < 2e-16 ***
YearBuilt      2.595e-03  2.627e-04   9.878  < 2e-16 ***
YearRemodAdd   1.145e-03  2.851e-04   4.017 6.20e-05 ***
MasVnrArea     2.804e-06  2.510e-05   0.112 0.911051
ExterQual      4.384e-02  1.092e-02   4.014 6.27e-05 ***
ExterCond     -1.451e-02  1.211e-02  -1.198 0.231305
BsmtQual       1.778e-04  4.915e-05   3.617 0.000309 ***
BsmtFinSF1     6.329e-05  1.972e-05   3.210 0.001357 **
BsmtFinSF2     2.666e-05  3.108e-05   0.858 0.391310
BsmtUnfSF      5.160e-05  1.782e-05   2.896 0.003838 **
TotalBsmtSF          NA         NA      NA       NA    #####correlated
`1stFlrSF`     1.881e-04  2.428e-05   7.747 1.78e-14 ***
`2ndFlrSF`     1.625e-04  2.090e-05   7.774 1.45e-14 ***
LowQualFinSF   1.556e-04  8.296e-05   1.875 0.060935 .
GrLivArea            NA         NA      NA       NA    ##### correlated
BsmtFullBath   6.308e-02  1.109e-02   5.689 1.55e-08 ***
BsmtHalfBath   2.266e-02  1.726e-02   1.313 0.189294
FullBath       4.845e-02  1.185e-02   4.089 4.57e-05 ***
HalfBath       3.147e-02  1.125e-02   2.798 0.005208 **
BedroomAbvGr   3.017e-03  7.253e-03   0.416 0.677462
KitchenAbvGr  -3.403e-02  2.202e-02  -1.546 0.122420
```

```
TotRmsAbvGrd   1.588e-02  5.213e-03   3.046 0.002365 **
Fireplaces     4.815e-02  7.453e-03   6.460 1.43e-10 ***
GarageArea     1.569e-04  2.813e-05   5.577 2.92e-08 ***
GarageQual     7.935e-03  2.091e-03   3.795 0.000154 ***
WoodDeckSF     1.147e-04  3.371e-05   3.403 0.000684 ***
OpenPorchSF   -6.695e-05  6.396e-05  -1.047 0.295386
EnclosedPorch  1.540e-04  7.102e-05   2.168 0.030317 *
`3SsnPorch`    2.195e-04  1.321e-04   1.662 0.096742 .
ScreenPorch    3.596e-04  7.238e-05   4.969 7.54e-07 ***
PoolArea      -3.860e-04  9.997e-05  -3.861 0.000118 ***
MiscVal       -4.832e-06  7.827e-06  -0.617 0.537087
MoSold         4.645e-04  1.452e-03   0.320 0.749061
YrSold        -7.726e-03  2.956e-03  -2.613 0.009060 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02140011)

    Null deviance: 232.801  on 1459  degrees of freedom
Residual deviance:  30.452  on 1423  degrees of freedom
AIC: -1430.9

Number of Fisher Scoring iterations: 2
```

26 CATEGORICAL VARS REGRESSION ONLY:

```
Call:
lm(formula = log(SalePrice) ~ ., data = dat, family = gaussian)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7157 -0.1120  0.0000  0.1071  0.7059 

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         10.7950163  0.3859032  27.973  < 2e-16 ***
MSZoningFV           0.3607380  0.0916338   3.937 8.69e-05 ***
MSZoningRH           0.4516202  0.0924738   4.884 1.17e-06 ***
MSZoningRL           0.4481203  0.0785556   5.705 1.44e-08 ***
MSZoningRM           0.4814035  0.0736416   6.537 8.97e-11 ***
```

```
StreetPave          0.0072340  0.0848246   0.085 0.932050
LotShapeIR2          0.0487032  0.0320237   1.521 0.128539
LotShapeIR3          0.1264233  0.0665028   1.901 0.057519 .
LotShapeReg         -0.0293685  0.0125112  -2.347 0.019055 *
LandContourHLS       0.0543087  0.0396410   1.370 0.170918
LandContourLow       0.0126561  0.0476464   0.266 0.790570
LandContourLvl       0.0350025  0.0284716   1.229 0.219150
UtilitiesNoSeWa     -0.3995829  0.1941703  -2.058 0.039798 *
LotConfigCulDSac     0.0452416  0.0250488   1.806 0.071126 .
LotConfigFR2        -0.0983426  0.0311360  -3.158 0.001622 **
LotConfigFR3        -0.1398336  0.0987405  -1.416 0.156963
LotConfigInside     -0.0323643  0.0136090  -2.378 0.017542 *
LandSlopeMod         0.0781463  0.0302693   2.582 0.009939 **
LandSlopeSev        -0.0024816  0.0701075  -0.035 0.971768
NeighborhoodBlueste  -0.1272824  0.1455751  -0.874 0.382093
NeighborhoodBrDale   -0.3738055  0.0814996  -4.587 4.94e-06 ***
NeighborhoodBrkSide  -0.3478224  0.0667248  -5.213 2.16e-07 ***
NeighborhoodClearCr  -0.1016236  0.0687937  -1.477 0.139857
NeighborhoodCollgCr  -0.1792263  0.0530404  -3.379 0.000749 ***
NeighborhoodCrawfor  -0.0488259  0.0616502  -0.792 0.428515
NeighborhoodEdwards  -0.3717546  0.0571526  -6.505 1.11e-10 ***
NeighborhoodGilbert  -0.2237159  0.0573789  -3.899 0.000102 ***
NeighborhoodIDOTRR   -0.4670412  0.0767739  -6.083 1.54e-09 ***
NeighborhoodMeadowV  -0.5810496  0.0808121  -7.190 1.09e-12 ***
NeighborhoodMitchel  -0.2423093  0.0597053  -4.058 5.23e-05 ***
NeighborhoodNAmes    -0.2510708  0.0562476  -4.464 8.75e-06 ***
NeighborhoodNoRidge   0.1857861  0.0609475   3.048 0.002348 **
NeighborhoodNPkVill  -0.0757571  0.1067449  -0.710 0.478015
NeighborhoodNridgHt   0.1536471  0.0538719   2.852 0.004412 **
NeighborhoodNWAmes   -0.0953534  0.0597444  -1.596 0.110726
NeighborhoodOldTown  -0.4587083  0.0672308  -6.823 1.36e-11 ***
NeighborhoodSawyer   -0.2899072  0.0590964  -4.906 1.05e-06 ***
NeighborhoodSawyerW  -0.1442010  0.0575623  -2.505 0.012362 *
NeighborhoodSomerst  -0.0104563  0.0664320  -0.157 0.874955
NeighborhoodStoneBr   0.1448514  0.0620951   2.333 0.019813 *
NeighborhoodSWISU    -0.2564234  0.0698047  -3.673 0.000249 ***
NeighborhoodTimber   -0.0213517  0.0615204  -0.347 0.728597
NeighborhoodVeenker   0.0436358  0.0792646   0.551 0.582065
Condition1Feedr      -0.0229052  0.0382604  -0.599 0.549500
```

```
Condition1Norm       0.0518702  0.0314399   1.650 0.099219 .
Condition1PosA       0.1682169  0.0768031   2.190 0.028683 *
Condition1PosN       0.1250374  0.0566248   2.208 0.027405 *
Condition1RRAe      -0.0119587  0.0702144  -0.170 0.864787
Condition1RRAn       0.0464061  0.0524286   0.885 0.376249
Condition1RRNe      -0.1379849  0.1375774  -1.003 0.316065
Condition1RRNn       0.0730654  0.0984445   0.742 0.458100
Condition2Feedr      0.1299720  0.1704193   0.763 0.445804
Condition2Norm       0.1027186  0.1461832   0.703 0.482386
Condition2PosA       0.8161868  0.2442271   3.342 0.000856 ***
Condition2PosN      -0.2312974  0.2042250  -1.133 0.257606
Condition2RRAe       0.1600362  0.3334123   0.480 0.631311
Condition2RRAn      -0.0030759  0.2408223  -0.013 0.989811
Condition2RRNn       0.1246113  0.2025170   0.615 0.538455
BldgType2fmCon       0.0473350  0.0390758   1.211 0.225975
BldgTypeDuplex       0.0727995  0.0332057   2.192 0.028528 *
BldgTypeTwnhs       -0.2631321  0.0406821  -6.468 1.40e-10 ***
BldgTypeTwnhsE      -0.1979347  0.0263031  -7.525 9.76e-14 ***
HouseStyle1.5Unf    -0.1503098  0.0571431  -2.630 0.008628 **
HouseStyle1Story    -0.1095875  0.0206121  -5.317 1.24e-07 ***
HouseStyle2.5Fin     0.2100386  0.0719152   2.921 0.003553 **
HouseStyle2.5Unf     0.2153847  0.0669998   3.215 0.001338 **
HouseStyle2Story     0.0155787  0.0217402   0.717 0.473757
HouseStyleSFoyer    -0.1348832  0.0379013  -3.559 0.000386 ***
HouseStyleSLvl      -0.0456907  0.0312421  -1.462 0.143852
RoofStyleGable      -0.1919339  0.1394833  -1.376 0.169046
RoofStyleGambrel    -0.1525624  0.1520590  -1.003 0.315897
RoofStyleHip        -0.0952287  0.1398747  -0.681 0.496110
RoofStyleMansard    -0.1800702  0.1634161  -1.102 0.270703
RoofStyleShed       -0.0240862  0.2654997  -0.091 0.927729
RoofMatlCompShg      0.5342172  0.2065620   2.586 0.009810 **
RoofMatlMembran      0.5917212  0.3217252   1.839 0.066110 .
RoofMatlMetal        0.2372540  0.3185198   0.745 0.456488
RoofMatlRoll         0.5679884  0.2812899   2.019 0.043667 *
RoofMatlTar&Grv      0.4924917  0.2494340   1.974 0.048542 *
RoofMatlWdShake      0.6742358  0.2381576   2.831 0.004711 **
RoofMatlWdShngl      1.0157916  0.2213463   4.589 4.88e-06 ***
Exterior1stAsphShn  -0.0806170  0.2549777  -0.316 0.751922
Exterior1stBrkComm  -0.3825147  0.2139792  -1.788 0.074068 .
```

```
Exterior1stBrkFace    0.2002721  0.0946714   2.115 0.034581 *
Exterior1stCBlock    -0.1254318  0.1960853  -0.640 0.522493
Exterior1stCemntBd    0.1491182  0.1451343   1.027 0.304399
Exterior1stHdBoard    0.0116821  0.0960995   0.122 0.903264
Exterior1stImStucc   -0.0571245  0.2161394  -0.264 0.791595
Exterior1stMetalSd    0.0948599  0.1096353   0.865 0.387070
Exterior1stPlywood    0.1041090  0.0950483   1.095 0.273575
Exterior1stStone      0.1997694  0.1790100   1.116 0.264641
Exterior1stStucco     0.1342783  0.1042399   1.288 0.197916
Exterior1stVinylSd   -0.0361710  0.0998011  -0.362 0.717089
Exterior1stWd Sdng    0.0007439  0.0915867   0.008 0.993521
Exterior1stWdShing   -0.0006293  0.0994769  -0.006 0.994954
Exterior2ndAsphShn    0.0868878  0.1695406   0.512 0.608395
Exterior2ndBrk Cmn    0.0935107  0.1551238   0.603 0.546737
Exterior2ndBrkFace   -0.0523087  0.0985676  -0.531 0.595725
Exterior2ndCBlock            NA         NA      NA       NA
Exterior2ndCmentBd    0.0953529  0.1430568   0.667 0.505184
Exterior2ndHdBoard    0.0885939  0.0925248   0.958 0.338484
Exterior2ndImStucc    0.1322879  0.1079928   1.225 0.220807
Exterior2ndMetalSd    0.0064160  0.1072552   0.060 0.952308
Exterior2ndOther      0.2867918  0.2123662   1.350 0.177102
Exterior2ndPlywood    0.0833887  0.0896038   0.931 0.352213
Exterior2ndStone     -0.1090217  0.1287278  -0.847 0.397197
Exterior2ndStucco    -0.0084659  0.1018655  -0.083 0.933778
Exterior2ndVinylSd    0.1527217  0.0964720   1.583 0.113648
Exterior2ndWd Sdng    0.1097454  0.0885114   1.240 0.215234
Exterior2ndWd Shng    0.0173364  0.0926527   0.187 0.851602
FoundationCBlock      0.0230597  0.0229723   1.004 0.315659
FoundationPConc       0.1050097  0.0247753   4.238 2.41e-05 ***
FoundationSlab       -0.2060116  0.0496687  -4.148 3.57e-05 ***
FoundationStone       0.2417977  0.0832240   2.905 0.003730 **
FoundationWood        0.0384592  0.1114054   0.345 0.729986
HeatingGasA           0.3699536  0.1921900   1.925 0.054454 .
HeatingGasW           0.6792378  0.1974940   3.439 0.000602 ***
HeatingGrav           0.2143335  0.2089151   1.026 0.305111
HeatingOthW           0.6203092  0.2371861   2.615 0.009018 **
HeatingWall           0.4957611  0.2225628   2.228 0.026083 *
HeatingQCFa          -0.1219413  0.0357002  -3.416 0.000656 ***
HeatingQCGd          -0.0578300  0.0158684  -3.644 0.000279 ***
```

```
HeatingQCPo          0.0566384  0.2089234   0.271 0.786360
HeatingQCTA         -0.0832613  0.0156070  -5.335 1.13e-07 ***
CentralAirY          0.1816583  0.0292457   6.211 7.04e-10 ***
ElectricalFuseF     -0.0115945  0.0450779  -0.257 0.797057
ElectricalFuseP      0.0399945  0.1306550   0.306 0.759572
ElectricalMix        0.1536180  0.2137952   0.719 0.472560
ElectricalSBrkr      0.0265129  0.0225305   1.177 0.239506
KitchenQualFa       -0.4436795  0.0437431 -10.143  < 2e-16 ***
KitchenQualGd       -0.2147330  0.0237107  -9.056  < 2e-16 ***
KitchenQualTA       -0.3417447  0.0263551 -12.967  < 2e-16 ***
FunctionalMaj2      -0.4358872  0.1101018  -3.959 7.93e-05 ***
FunctionalMin1       0.0024103  0.0639643   0.038 0.969947
FunctionalMin2      -0.0237346  0.0627932  -0.378 0.705506
FunctionalMod       -0.0073053  0.0763200  -0.096 0.923758
FunctionalSev       -0.2713592  0.2063435  -1.315 0.188712
FunctionalTyp       -0.0106475  0.0532525  -0.200 0.841555
PavedDriveP          0.0663881  0.0414153   1.603 0.109179
PavedDriveY          0.0787806  0.0248983   3.164 0.001591 **
SaleTypeCon          0.1107391  0.1391166   0.796 0.426167
SaleTypeConLD        0.1405556  0.0746693   1.882 0.060007 .
SaleTypeConLI       -0.1086182  0.0898545  -1.209 0.226950
SaleTypeConLw       -0.1252482  0.0929026  -1.348 0.177838
SaleTypeCWD          0.1266847  0.1002982   1.263 0.206785
SaleTypeNew          0.2004492  0.1205721   1.662 0.096655 .
SaleTypeOth         -0.0634126  0.1128414  -0.562 0.574238
SaleTypeWD          -0.0681273  0.0323338  -2.107 0.035308 *
SaleConditionAdjLand 0.0241678  0.1109008   0.218 0.827523
SaleConditionAlloca  0.2055804  0.0646309   3.181 0.001503 **
SaleConditionFamily  0.0444343  0.0477545   0.930 0.352297
SaleConditionNormal  0.1034536  0.0223942   4.620 4.22e-06 ***
SaleConditionPartial -0.0998031  0.1165138  -0.857 0.391835
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1834 on 1308 degrees of freedom
Multiple R-squared:  0.811,    Adjusted R-squared:  0.7892
F-statistic: 37.18 on 151 and 1308 DF,  p-value: < 2.2e-16
```

3. **R code for analysis:**
    a. For all R scripts related to the analysis in this report please see the Github repo here: https://github.com/ileenamitra/kaggle_predict_house_prices
    b. For more information about the randomForest R package please see documentation here: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf or online description here: https://www.stat.berkeley.edu/~breiman/RandomForests/
    c. For more information about the gradient boosting package,please see document "Generalized Boosted Models: A guide to the gbm package" written by Greg Ridgeway.

REFERENCES

"American House Prices: Realty Check." The Economist. The Economist Newspaper, 24 Aug. 2016. http://www.economist.com/blogs/graphicdetail/2016/08/daily-chart-20. Accessed 21 March 2017.

"Ames,." Ames Housing Market Data and Appreciation Trends - NeighborhoodScout. N.p., n.d. Web. https://www.neighborhoodscout.com/ia/ames/real-estate. Accessed 21 March 2017.

Chang, Lee Chun and Hui-Yu Lin. "The impact of neighborhood characteristics on housing prices—an application of hierarchical linear modeling." International Journal of Management and Sustainability, vol. 1, no. 2, pp. 31-44. http://pakacademicsearch.com/pdf-files/ech/11/31-44%20Vol%201%20issue%202%20December%20%202012.pdf. Accessed 25 March 2017.

DiClerico, Daniel. "8 Ways to Boost Your Home Value." *Consumer Reports*. 9 Feb 2016. http://www.consumerreports.org/home-improvement/8-ways-to-boost-your-home-value/ Accessed 30 March 2017.

Harney, Kenneth R. "School quality is tied to home prices in new study." *Washington Post*, 4 October 2013. https://www.washingtonpost.com/realestate/school-quality-is-tied-home-prices-in-new-study-but-other-factors-may-affect-values/2013/10/02/f7b12e24-2aa4-11e3-8ade-a1f23cda135e_story.html?utm_term=.c3dcef372559. Accessed 25 March 2017.

Haurin, Donald R. *The Impact of School Quality on House Prices: Interjurisdictional Effects*. Dissertation, Ohio State University, 1996. http://ecolan.sbs.ohio-state.edu/pdf/haurin/haurin.pdf. Accessed 25 March 2017.

Islam, Shahidul. "IMPACT OF NEIGHBOURHOOD CHARACTERISTICS ON HOUSE PRICES." Proceedings of ASBBS 19.1 (2012): n. pag. http://asbbs.org/files/ASBBS2012V1/PDF/I/IslamS.pdf. Accessed 21 March 2017.

Sheppard, Stephen. “Hedonic Analysis of Housing Markets.” *Handbook of Regional and Urban Economics 3* (1999): 1595-1635.

“Quick Facts: Ames city, Iowa.” *United States Census Bureau*. https://www.census.gov/quickfacts/table/RHI105210/1901855. Accessed 25 March 2017.

“Quick Facts: Resident Demographics.” *National Multifamily Housing Council Online*, Septmber 2016. http://www.nmhc.org/Content.aspx?id=4708. Accessed 30 March 2017.

Weigley, Samuel. “11 home features buyers will pay extra for.” *USA Today*, 28 April 2013. http://www.usatoday.com/story/money/personalfinance/2013/04/28/24-7-home-features/2106203/. Accessed 25 March 2017.