DSCI 445 Project Presentation

Megan Dunnahoo, Mandey Brown, Emma Hamilton

Motivation

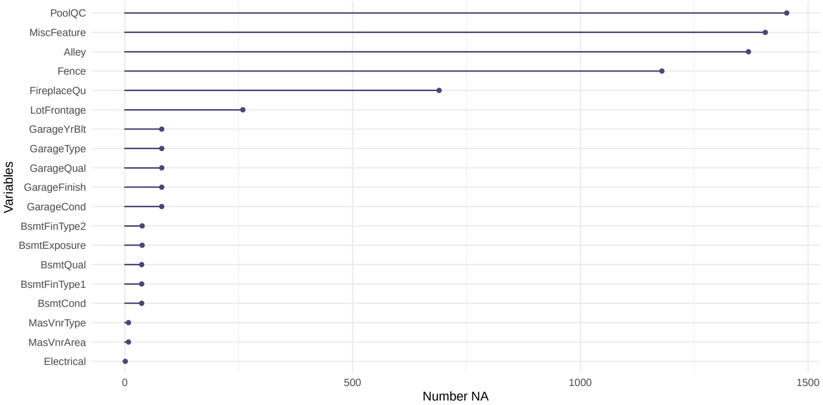
Housing prices play a central role in the U.S. economy. According to a *Congressional Research Service* article, *Introduction to U.S. Economy: Housing Market*, “at the individual level, roughly 65% of occupied housing units are owner occupied, homes are a substantial source of household wealth in the United States... housing accounts for a significant portion of all economic activity, and changes in the housing market can have broader effects on the economy.” The housing market is also incorporated into gross domestic product (GDP), which is considered the primary measure of economic activity for a country. Also, according to the article, *Introduction to U.S. Economy: Housing Market*, “as of 2020, spending on housing services was about $2.8 trillion, accounting for 13.3% of GDP. Taken together, spending within the housing market accounted for 17.5% of GDP in 2020.” The housing market not only affects the U.S. economy and GDP, up and coming college graduates will also soon be on the lookout for a place to live and knowing exactly what affects housing prices could prove to be incredibly useful.

Methodology

Housing prices influence the economy, but what influences house prices? Kaggle competition, *House Prices – Advanced Regression Techniques* provides a dataset of housing prices compiled by Dean De Cock in 2011, which describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The dataset includes 79 explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) all involved in evaluating home values. In order to find out which aspects of a home matter most when it comes to the final price of a home, the 79 available variables will be explored using a Decision Tree, Random Forest, PCR, Linear Regression, LASSO, and Ridge Regression models.

Handling Missing Values

We visualized the number of missing values for each variable and produced a plot which shows the number of NA values for variables with at least 1 NA value.



Missing Categorical/Character Variable Values

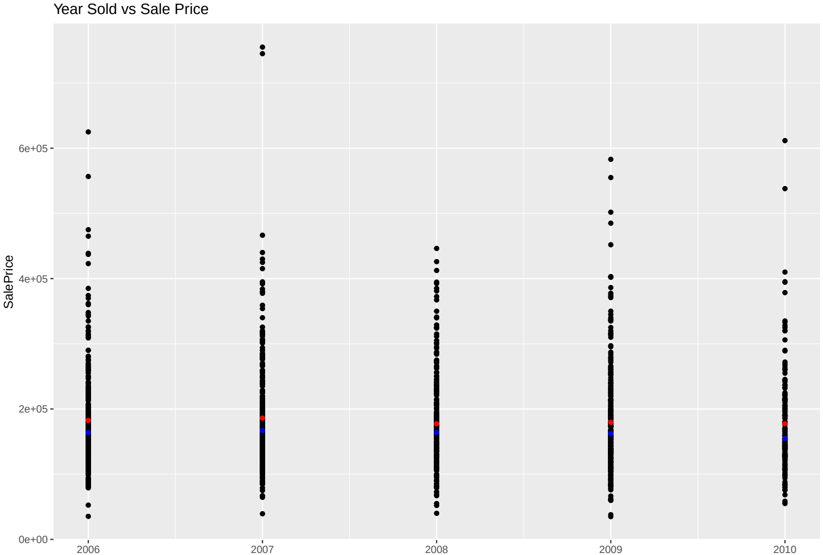
The categorical variables that have missing values are PoolQC, Fence, MiscFeature, Alley, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, BsmtQual, BsmtExposure, BsmtFinType1, BsmtFinType2, BsmtCond, MasVnrType, and Electrical. NAs for all of these variables, except for Electrical, likely represent the absence of a pool, fence, alley access, fireplace, garage, basement, etc. For these variables, we replaced the missing values with the level "None". For Electrical, there was only one missing value, which we replaced with the most common Electrical type.

Missing Numeric Variable Values

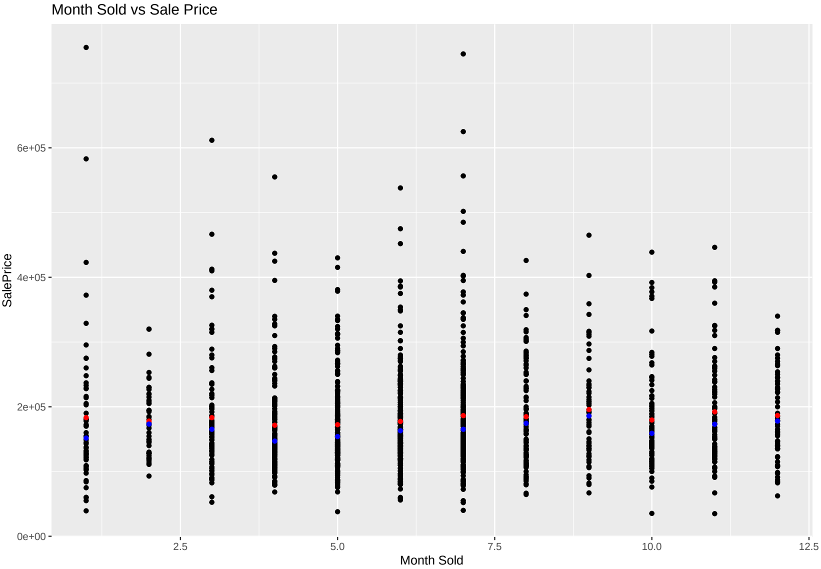
The numerical variables that have missing values are LotFrontage, GarageYrBuilt, and MasVnrArea. The missing values for these variables, similarly, likely mean that there is no garage, masonry veneer, or street connected to the property. Therefore, we replaced these missing values with 0.

Exploratory Data Analysis

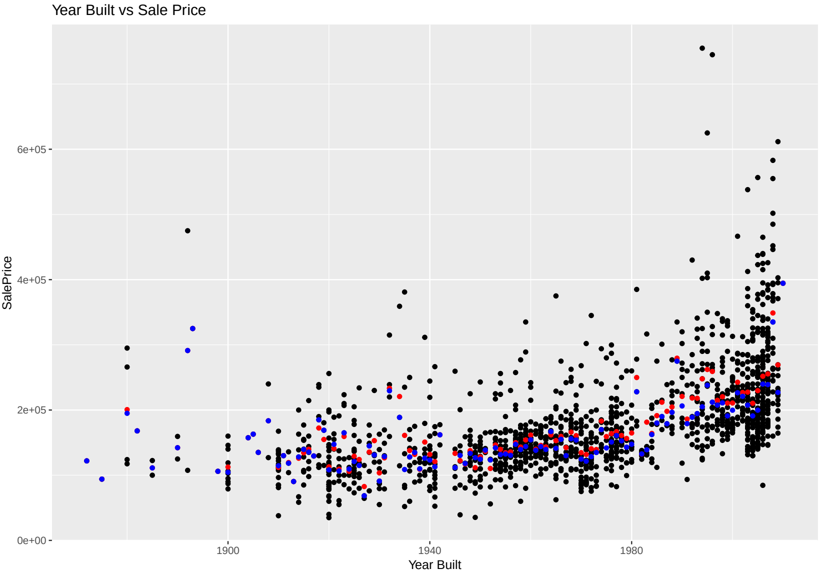
We first looked at some of the time variables vs Sale Price. We included red dots for the mean and blue dots for the median.



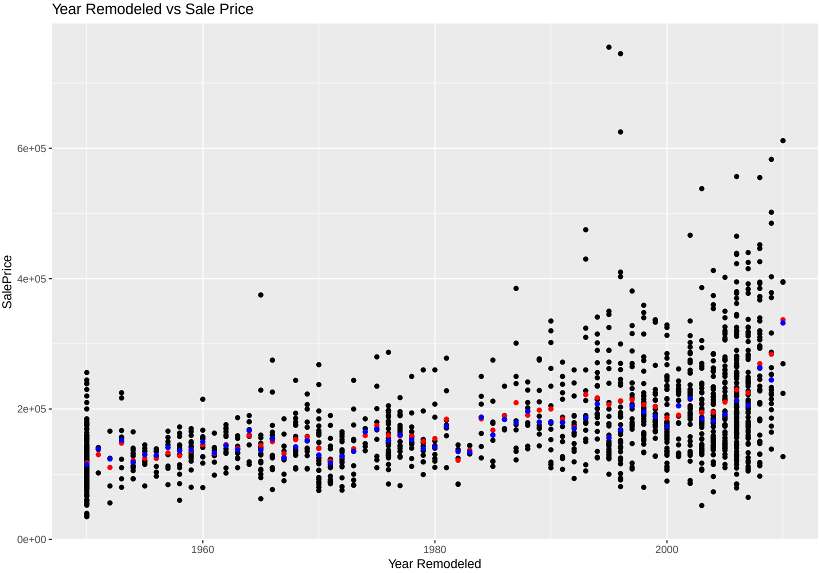
Exploratory Data Analysis Cont.



Exploratory Data Analysis Cont.

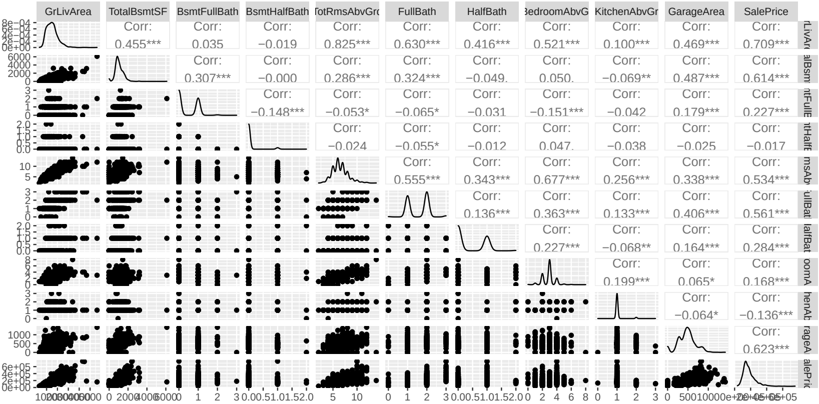


Exploratory Data Analysis Cont.



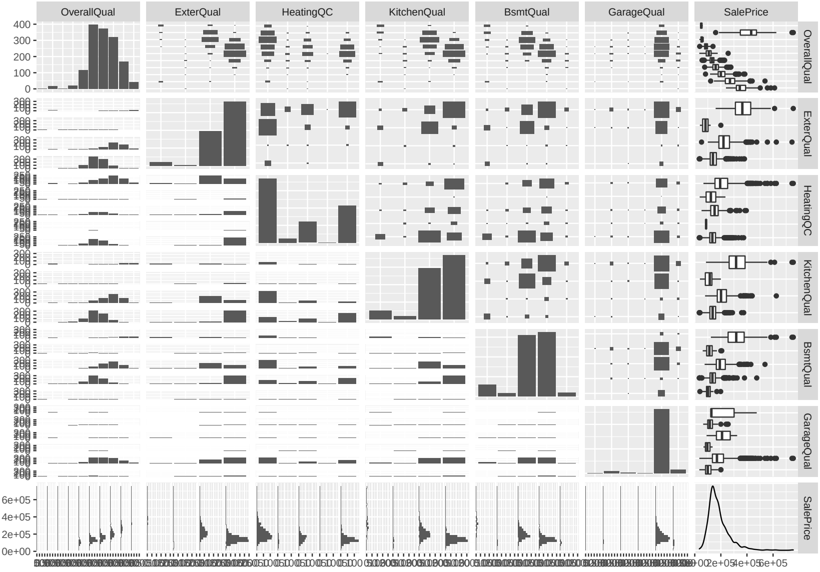
Exploratory Data Analysis Cont.

After this, we subsetted the variables into different general categories to make ggpair plots. This first plot shows variables which indicate size of the house. These include square feet of different living areas, number of bedrooms, number of bathrooms, etc.



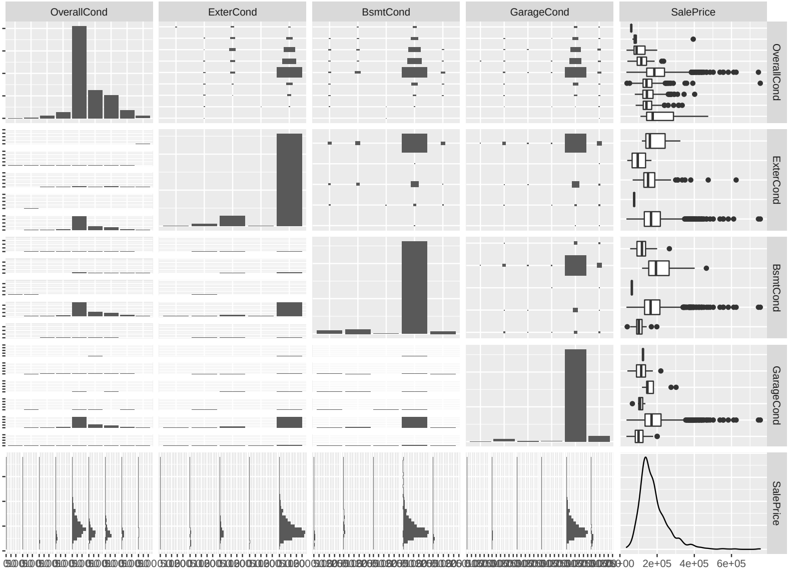
Exploratory Data Analysis Cont.

This plot shows variables which indicate the quality of the house.



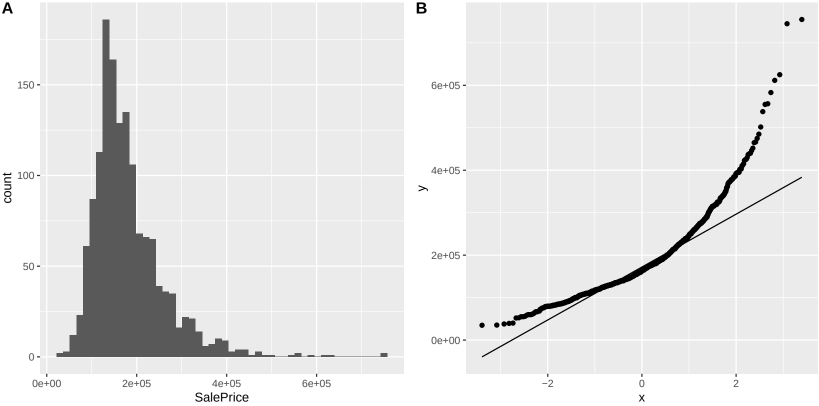
Exploratory Data Analysis Cont.

This plot shows variables which indicate the condition of the house.



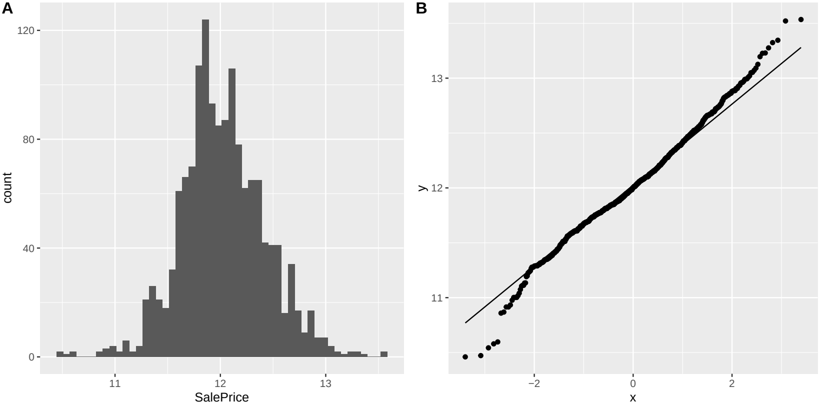
Log Transform the Data

We decided to log transform Sale Price as it violates the assumption of normality. These are the plots showing Sale Price before the transformation.

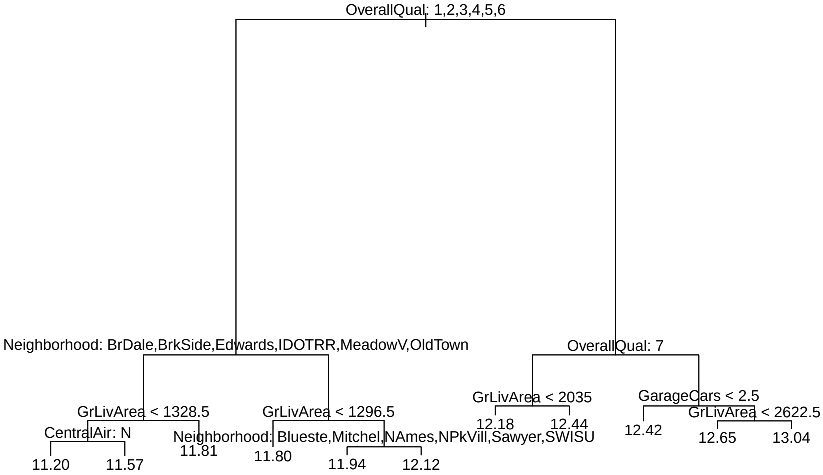


Log Transform the Data

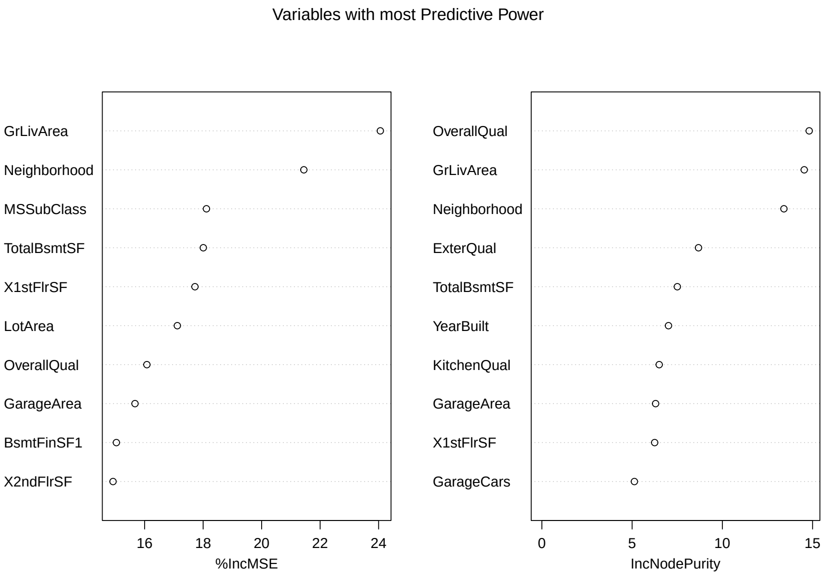
These are the plots showing Sale Price after the log transformation.



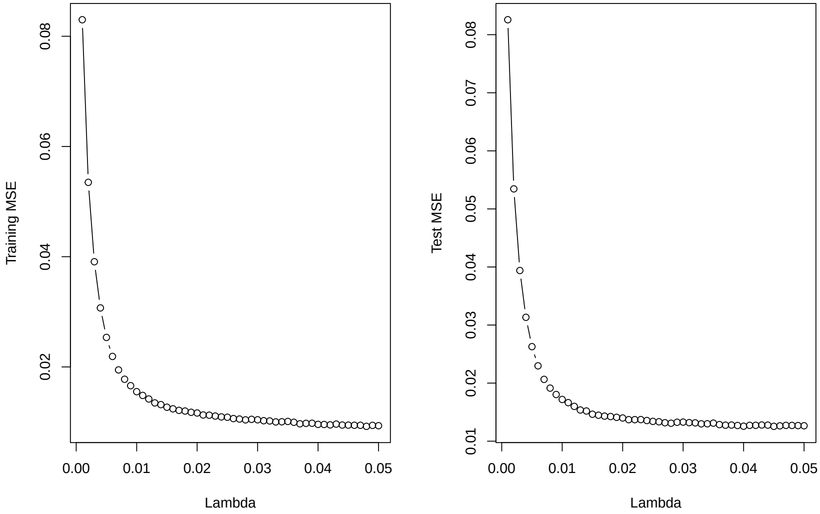
Decision Tree



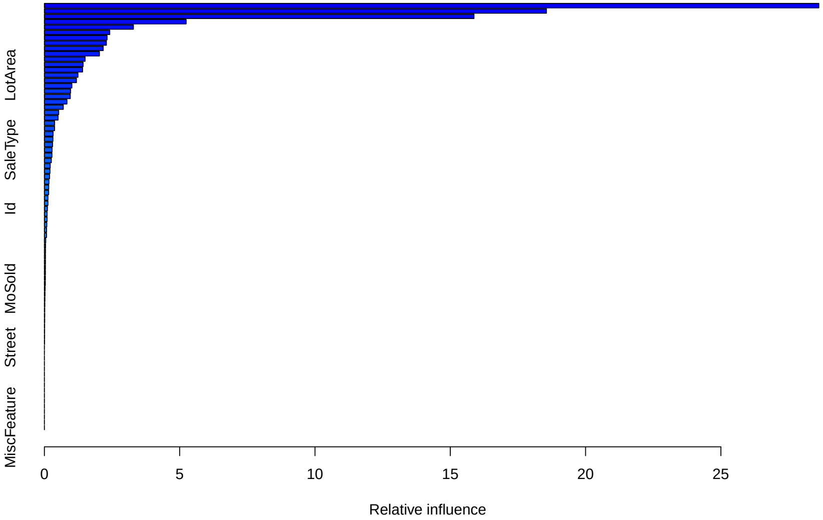
Random Forest



Boosting


Training MSE
0.02
0.04
0.06
0.08
0.00
0.01
0.02
0.03
0.04
0.05
Lambda
Test MSE
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.00
0.01
0.02
0.03
0.04
0.05
Lambda

Boosting Cont.



```
##                                    var      rel.inf
## OverallQual      OverallQual 28.620280312
## GrLivArea          GrLivArea 18.553677330
```

Boosting Cont.

```
##                           var   rel.inf
## OverallQual     OverallQual 28.620280
## GrLivArea         GrLivArea 18.553677
## Neighborhood   Neighborhood 15.875801
## TotalBsmtSF     TotalBsmtSF  5.237568
## KitchenQual     KitchenQual  3.289479
## OverallCond     OverallCond  2.414015
## X1stFlrSF         X1stFlrSF  2.315245
## GarageType       GarageType  2.290376
## GarageArea       GarageArea  2.174446
## ExterQual         ExterQual  2.032960
## YearRemodAdd   YearRemodAdd  1.504338
## CentralAir       CentralAir  1.431267
## BsmtFinSF1       BsmtFinSF1  1.411325
## LotArea             LotArea  1.239949
## SaleCondition SaleCondition  1.182974
```

LASSO

Anything to put here?

Table of MSE Values

```r
MSE_table <- data.frame(Model = c("Decision Tree", "Random
                        MSE = c(dt_MSE, rf_MSE, boost_MSE,

kable((MSE_table), caption = "MSE Values for Different Mode
```

Table 1: MSE Values for Different Models

Model	MSE
Decision Tree	0.0394
Random Forest	0.0097
Boosted Forest	0.0126
LASSO	0.0165