DSCI 445 Project Presentation

Megan Dunnahoo, Mandey Brown, Emma Hamilton

Motivation

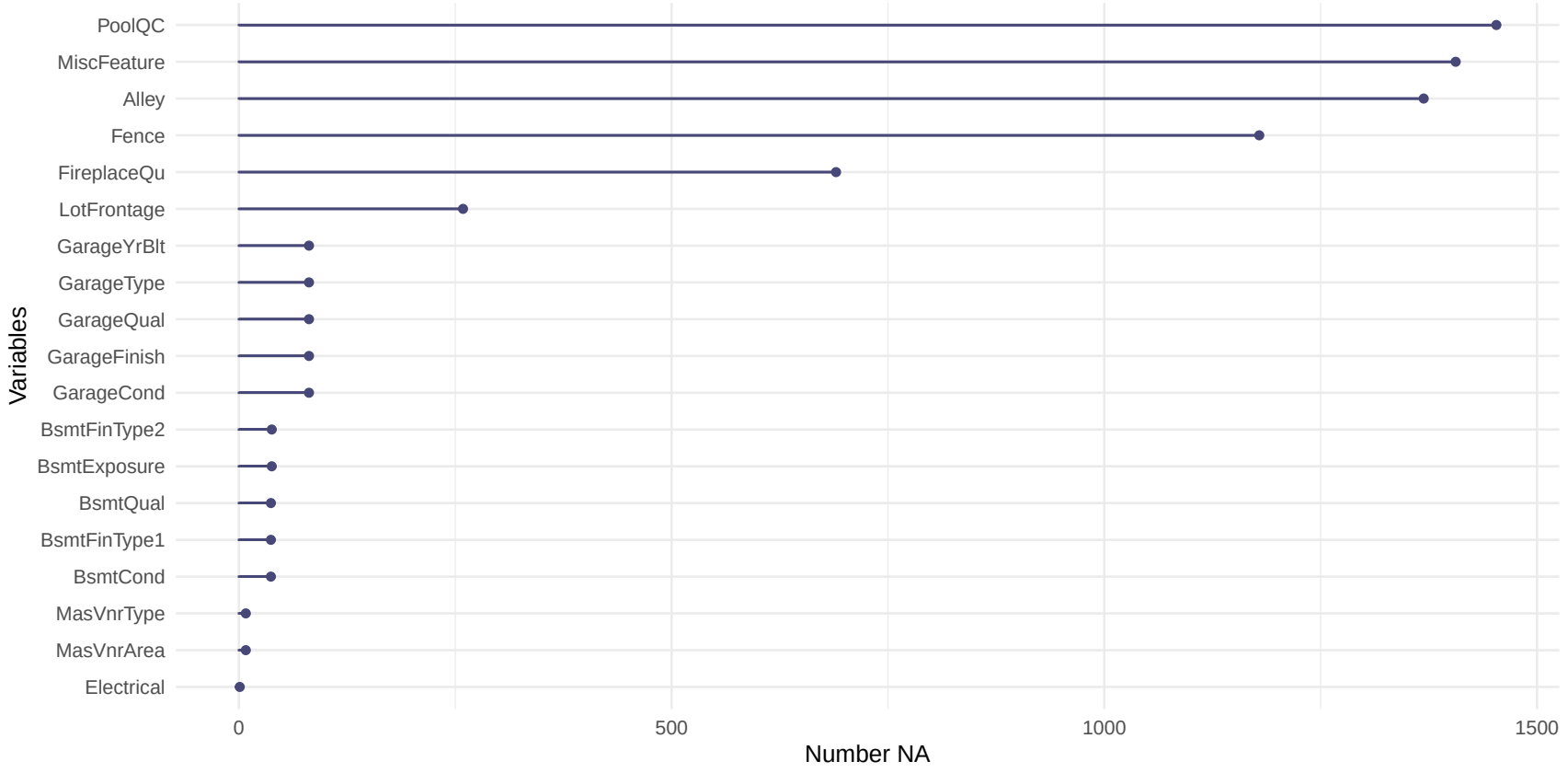
- Housing prices affects U.S. economy
  - 65-percent of houses are owner occupied
  - Housing is a profitable investment
  - Housing Market accounts for 13.3-percent of GDP
- Many professions and industries would benefit
  - Appraisers
  - Tax assessors
  - Mortgage lenders
  - Insurers
  - Home Builders
    - Worked with co-owner of Deluxe Homes LLC
    - Anticipate building costs
    - Need for flexible predictions based on specifics of home

Methodology

- Kaggle Knowledge Competition
  - "House Prices - Advanced Regression Techniques"
  - Residental property of Ames, Iowa from 2006 to 2010
  - Dataset includes 79 variables
    - 23 nominal, 23 ordinal, 14 discrete, and 20 continuous
  - Exploratory Analysis
  - Preprocessing
  - Advanced Regression Techniques
    - Decision Tree
    - Random Forest
    - LASSO
    - Boosting

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

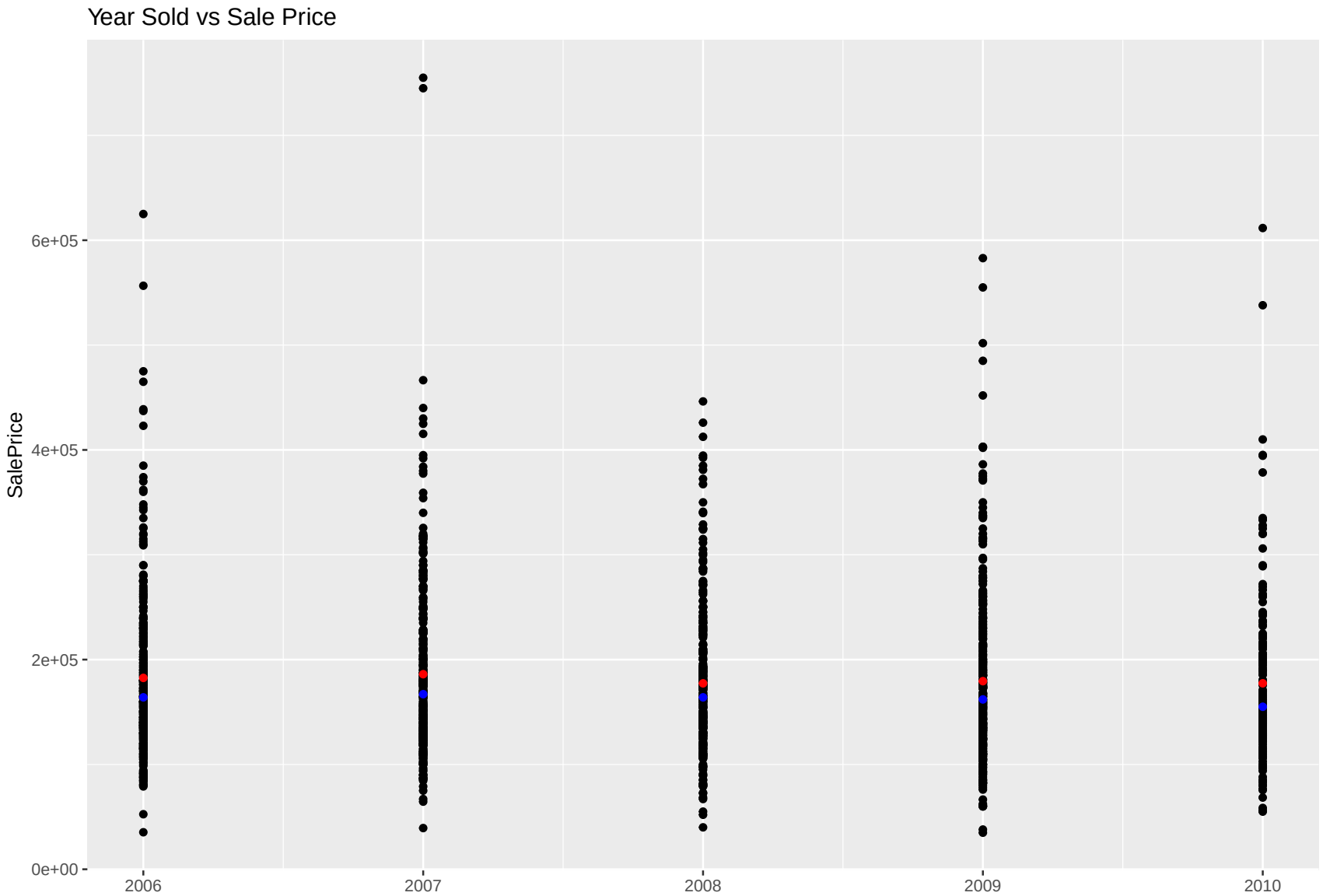
The categorical variables that have missing values are PoolQC, Fence, MiscFeature, Alley, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, BsmtQual, BsmtExposure, BsmtFinType1, BsmtFinType2, BsmtCond, MasVnrType, and Electrical. NAs for all of these variables, except for Electrical, likely 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

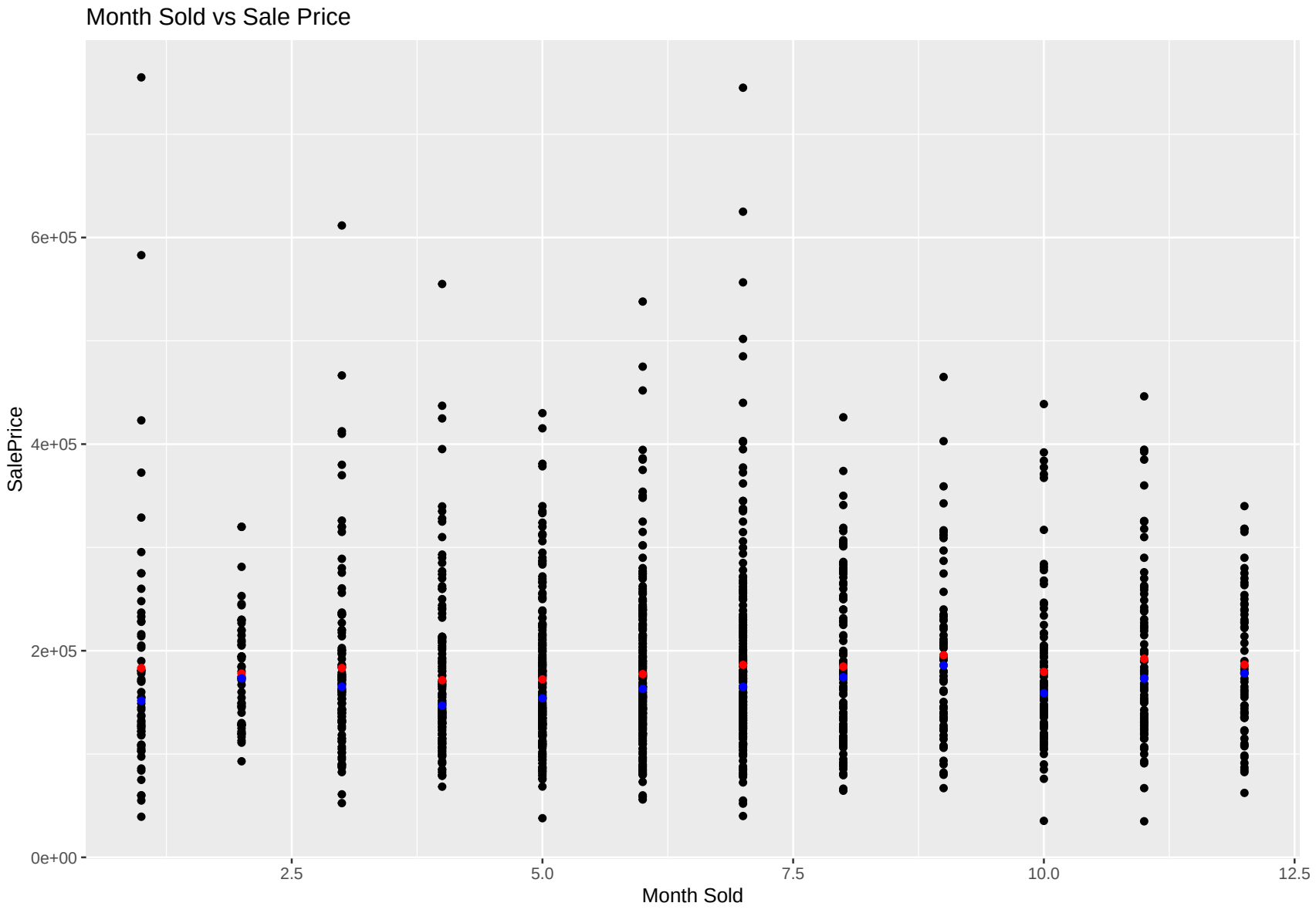
The numerical variables that have missing values are LotFrontage, GarageYrBuilt, and MasVnrArea. The missing values for these variables, similarly, likely mean that there is no garage, masonry veneer, or street connected to the property. Therefore, we replaced these missing values with 0.

Exploratory Data Analysis

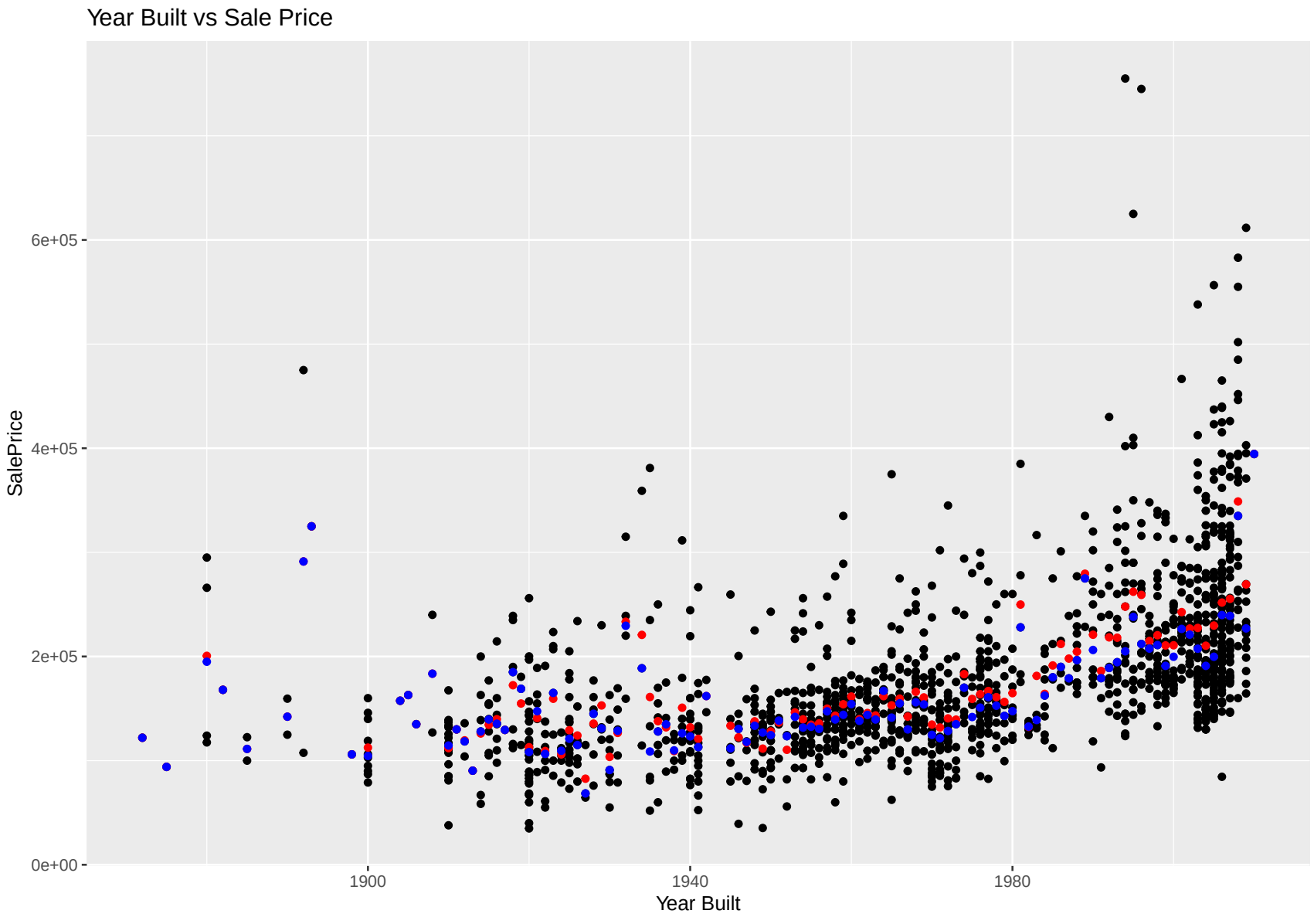
We first looked at some of the time variables vs Sale Price. We included red dots for the mean and blue dots for the median.



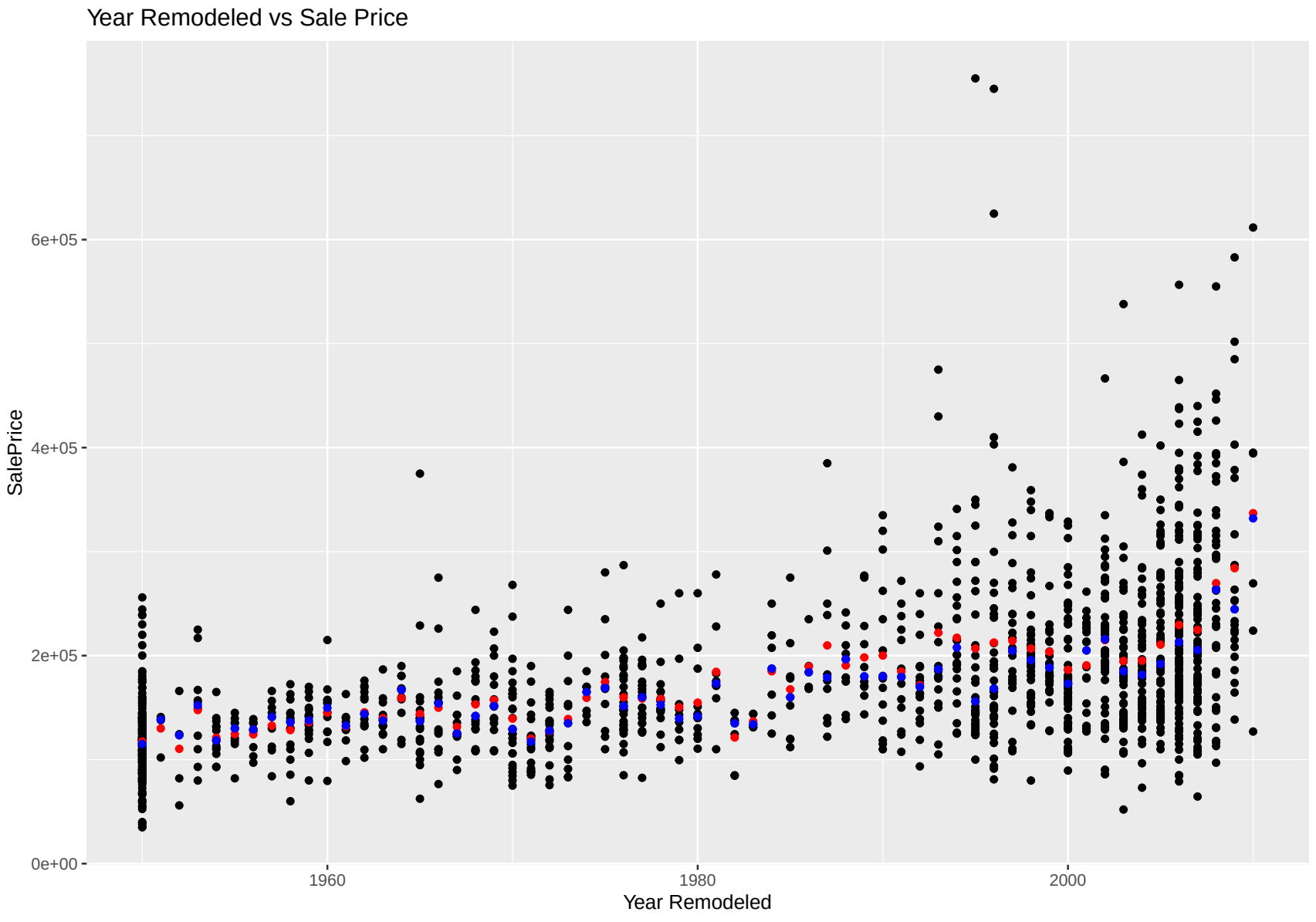
Exploratory Data Analysis Cont.



Exploratory Data Analysis Cont.

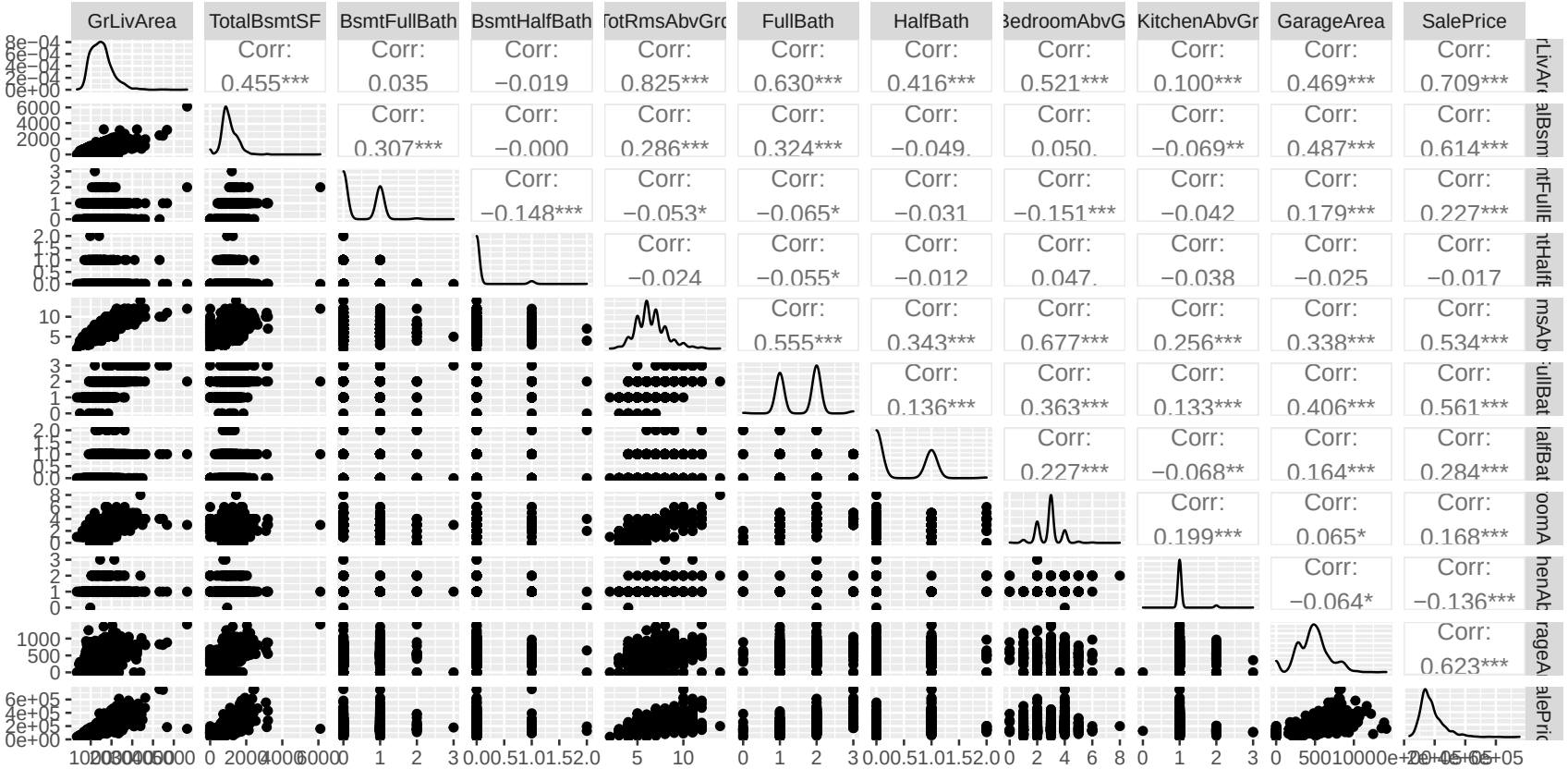


Exploratory Data Analysis Cont.



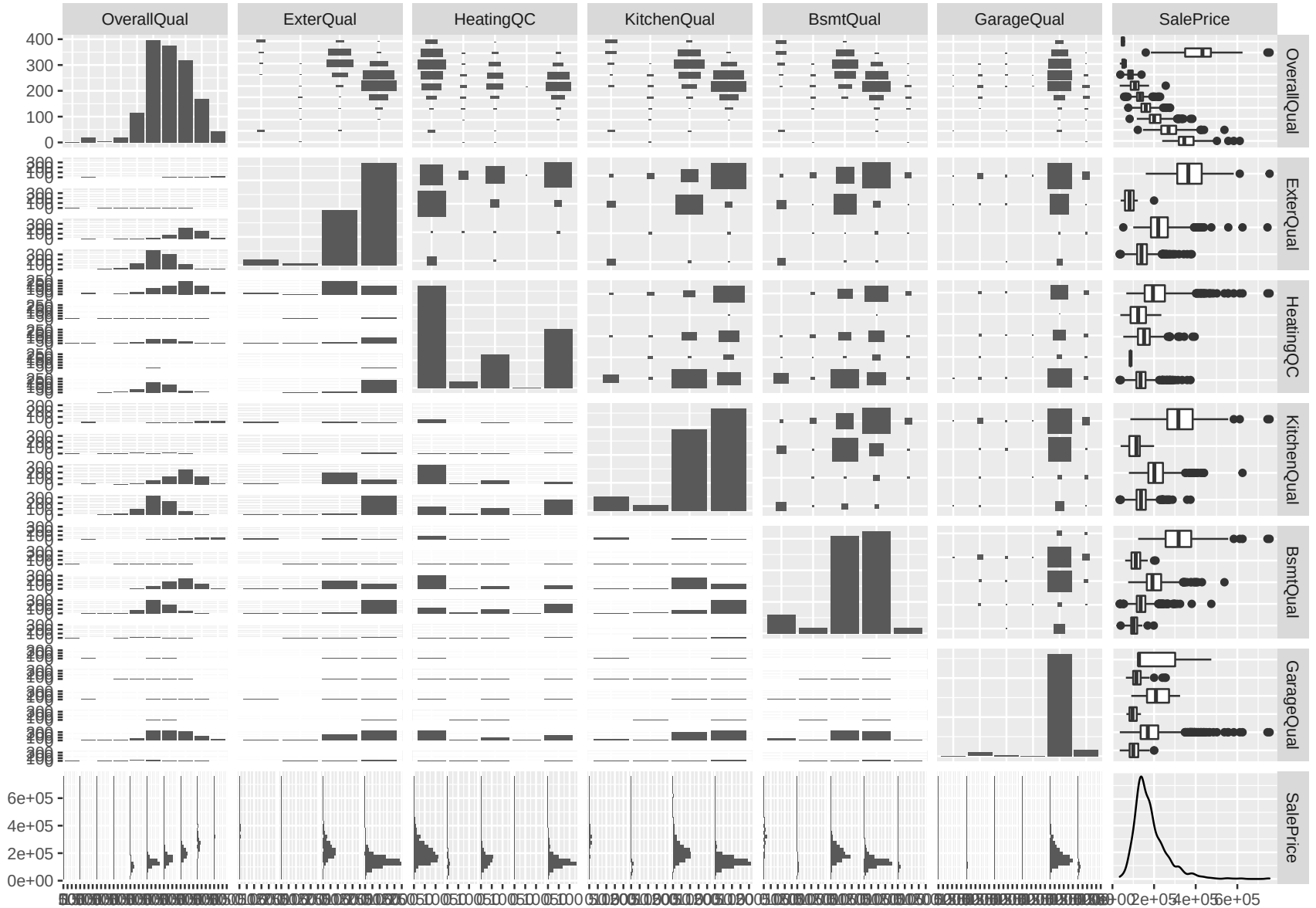
Exploratory Data Analysis Cont.

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.



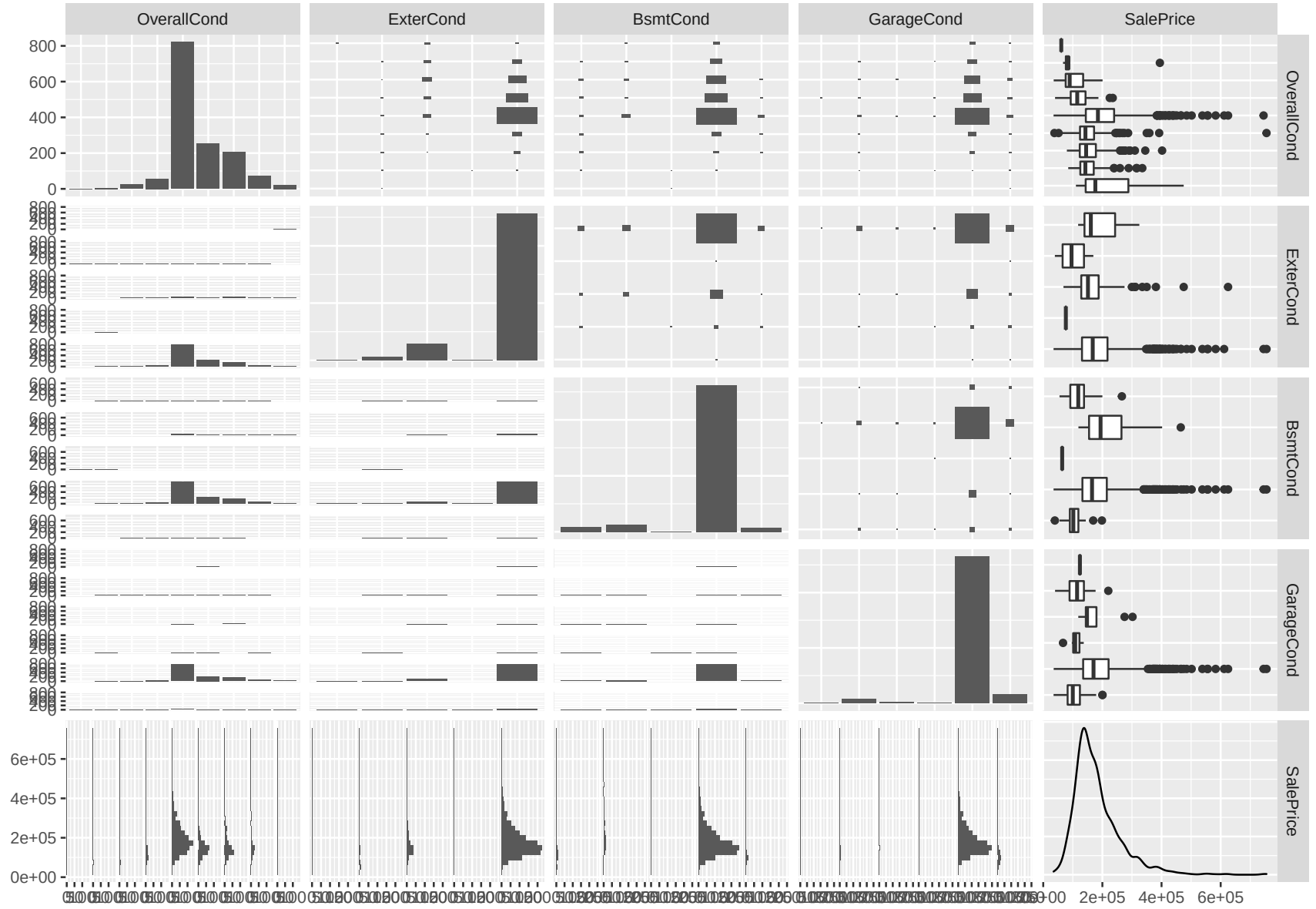
Exploratory Data Analysis Cont.

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



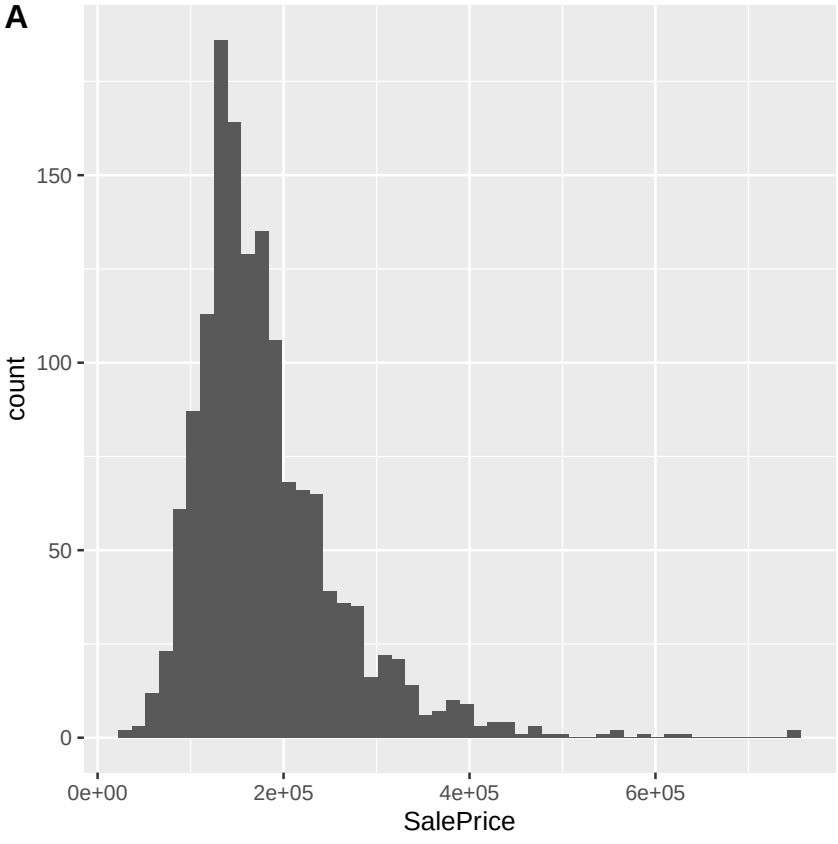
Exploratory Data Analysis Cont.

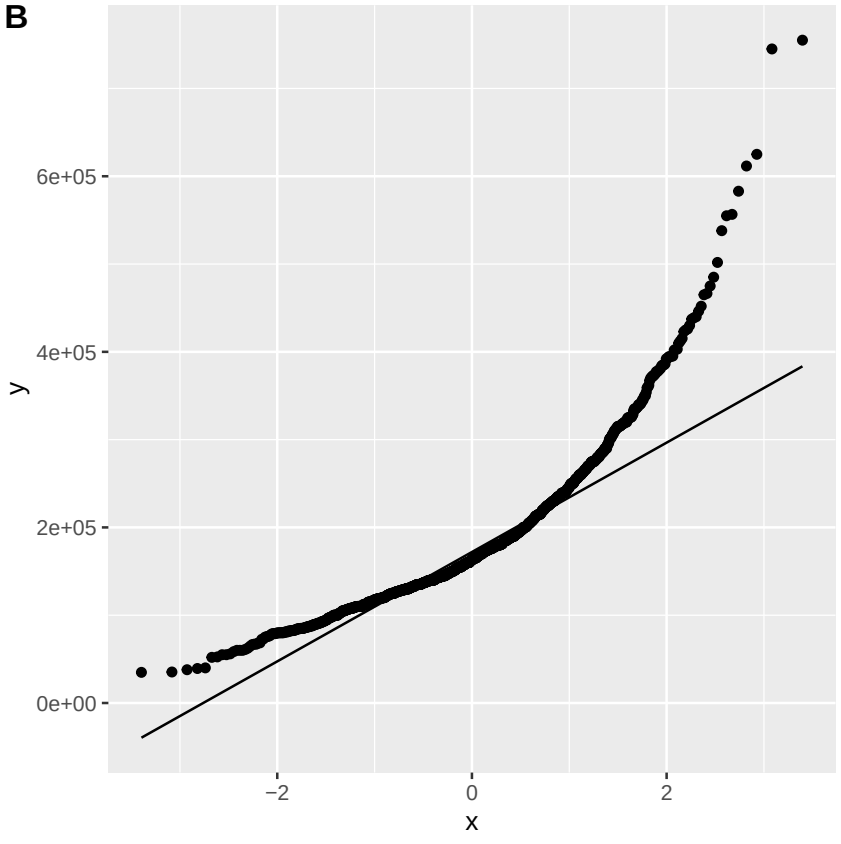
This plot shows variables which indicate the condition of the house.



Log Transform the Data

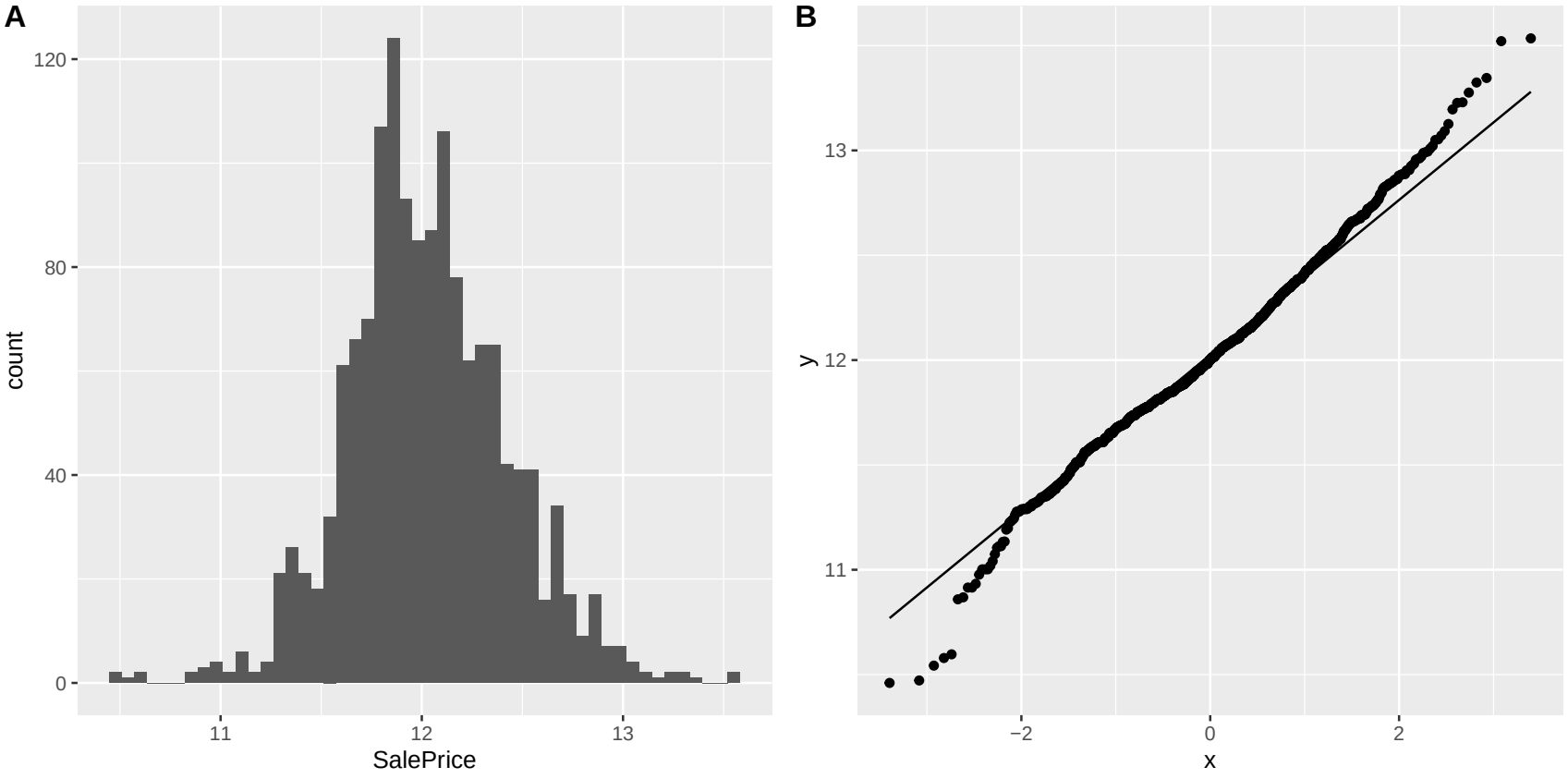
We decided to log transform Sale Price as it violates the assumption of normality. These are the plots showing Sale Price before the transformation.





Log Transform the Data

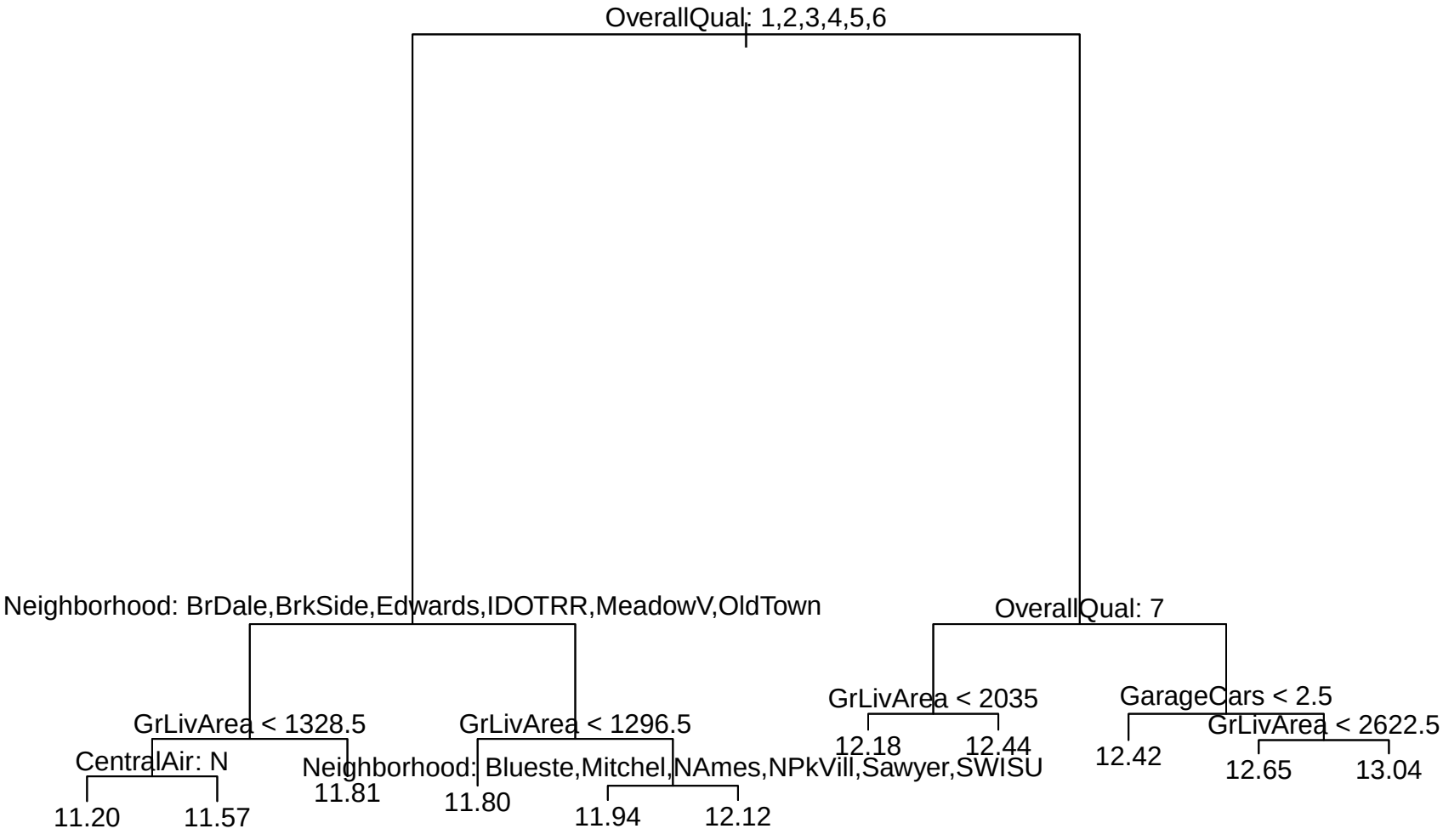
These are the plots showing Sale Price after the log transformation.



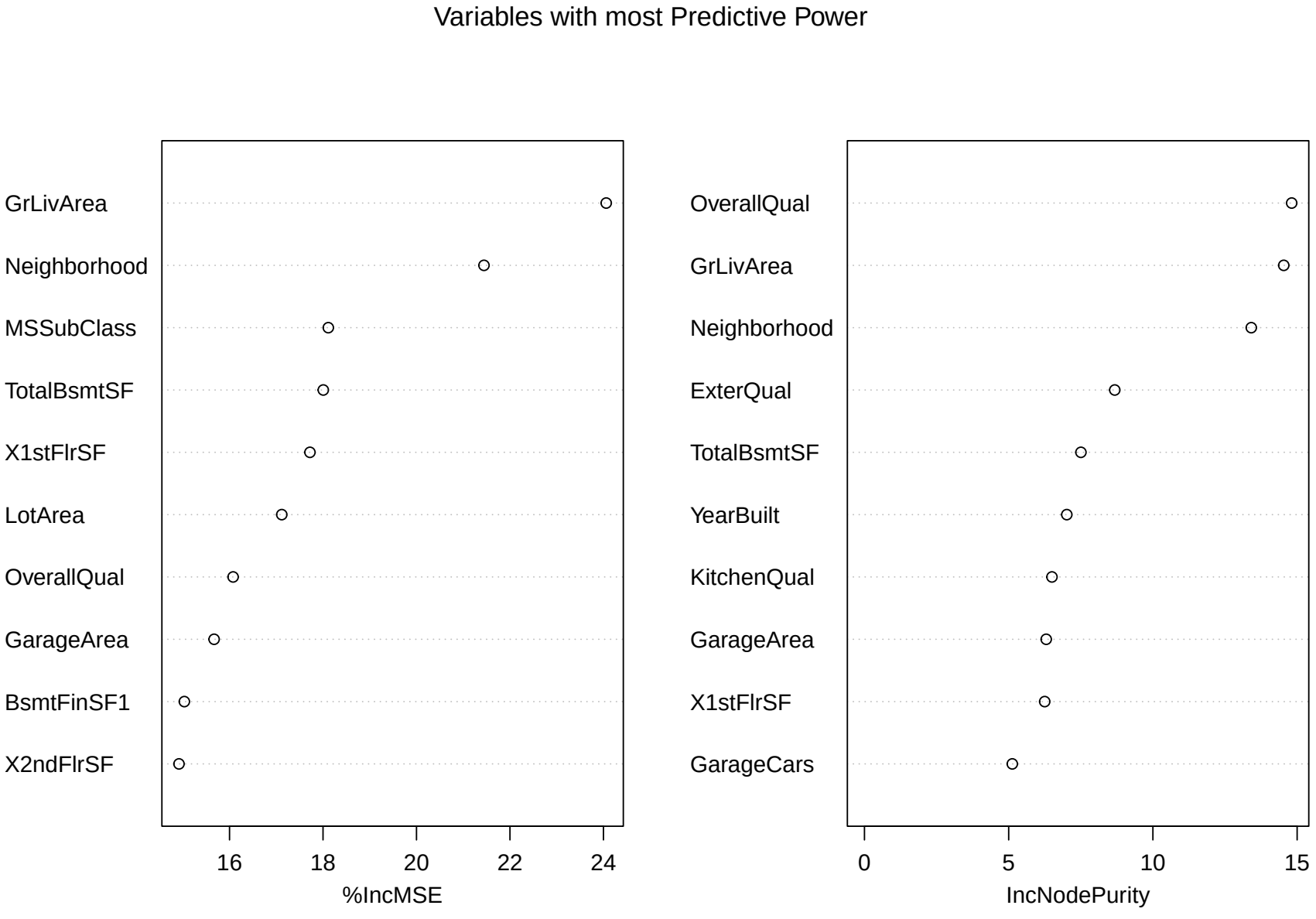
Tree-Based Methods

- Variety of Trees-Based Methods used
  - Curious which model might perform better
  - Allow for good prediction
    - Goal is to predict Sale Price
  - Allow for ease of interpretation
  - Good visualization
  - Suggest which variables are most significant

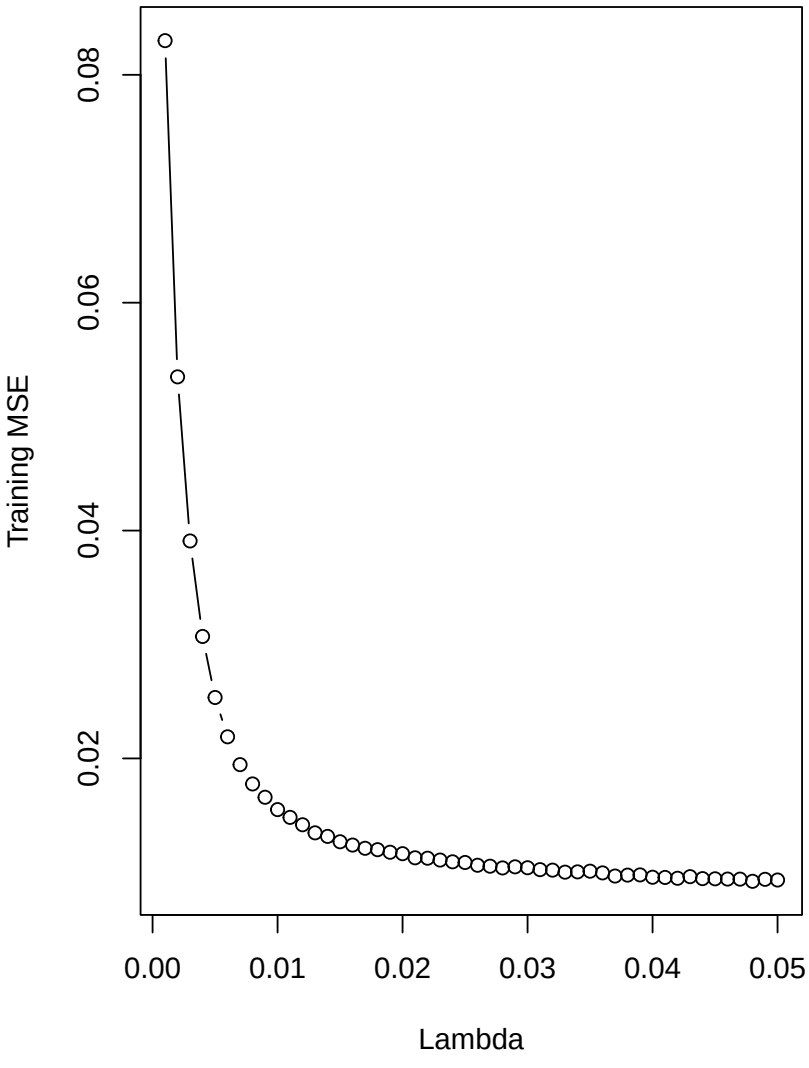
Decision Tree

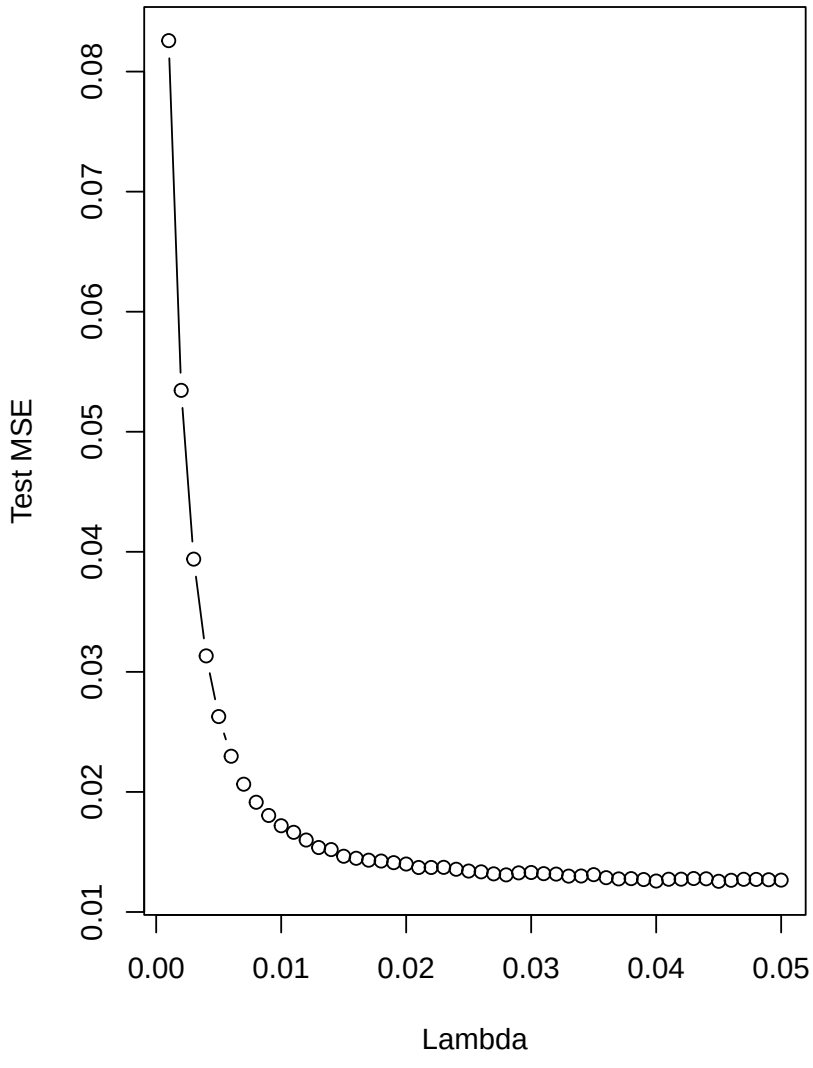


Random Forest

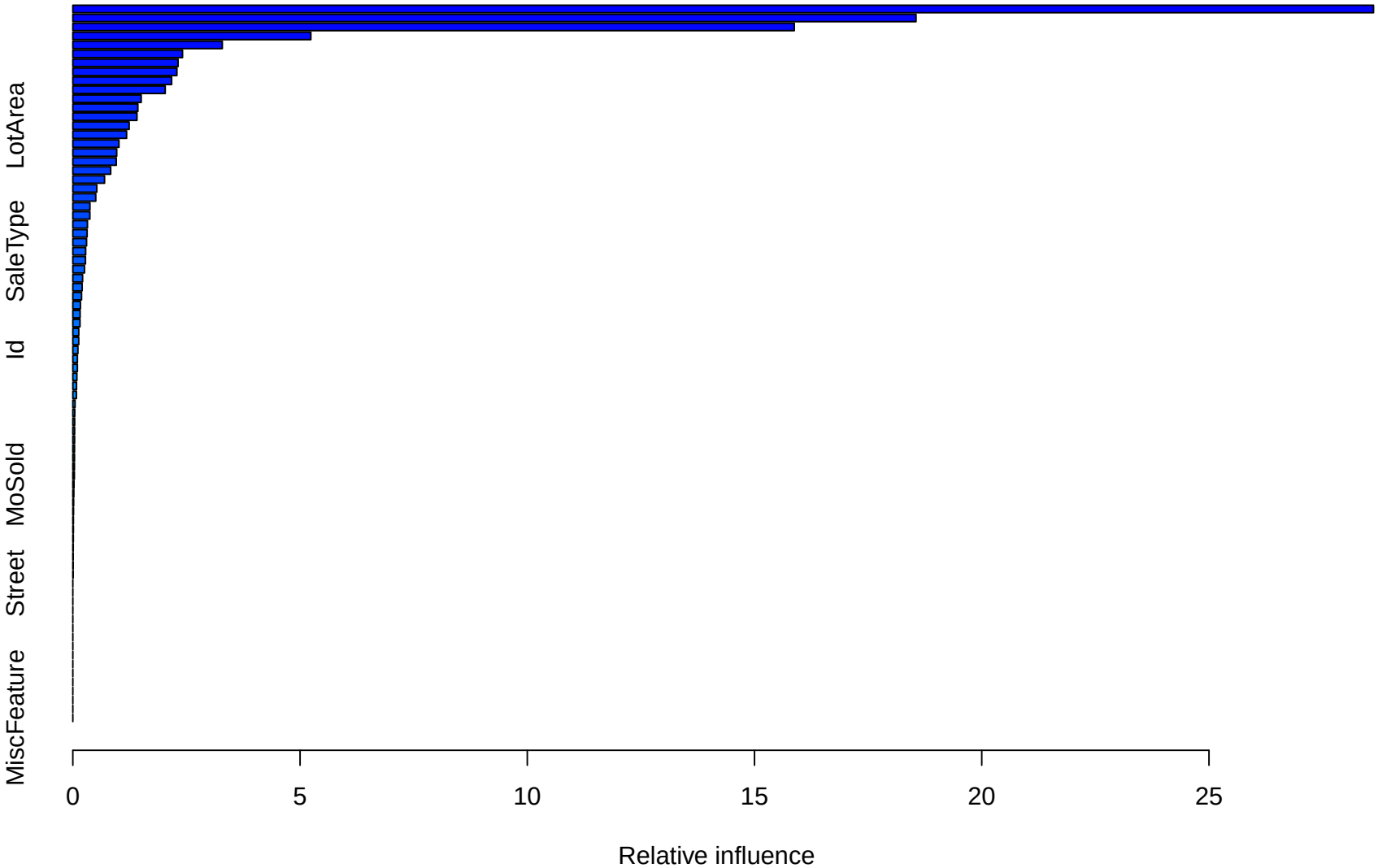


Boosting


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.



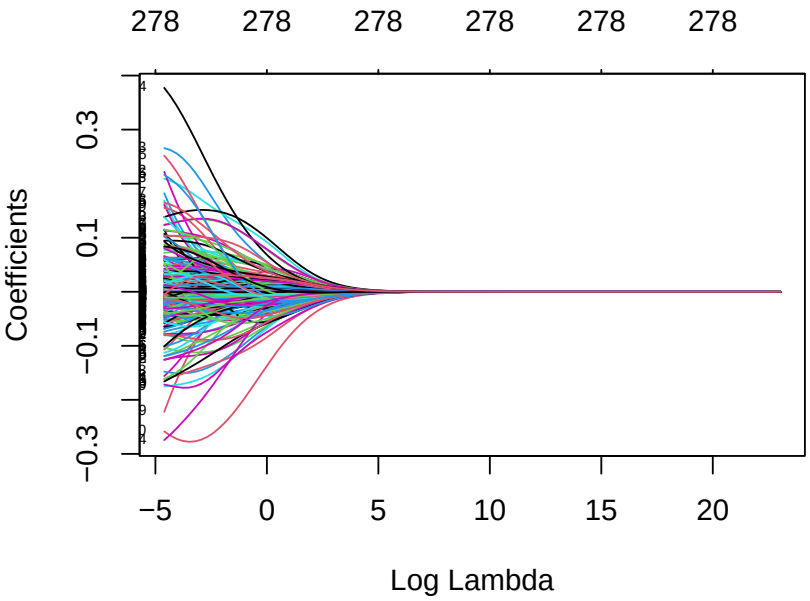

```
##                          var      rel.inf
## OverallQual      OverallQual 28.620280312
## GrLivArea          GrLivArea 18.553677330
```

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

- Many predictor variables (72 total)
- Desire to determine which subset to use
- Allow for ease of interpretation



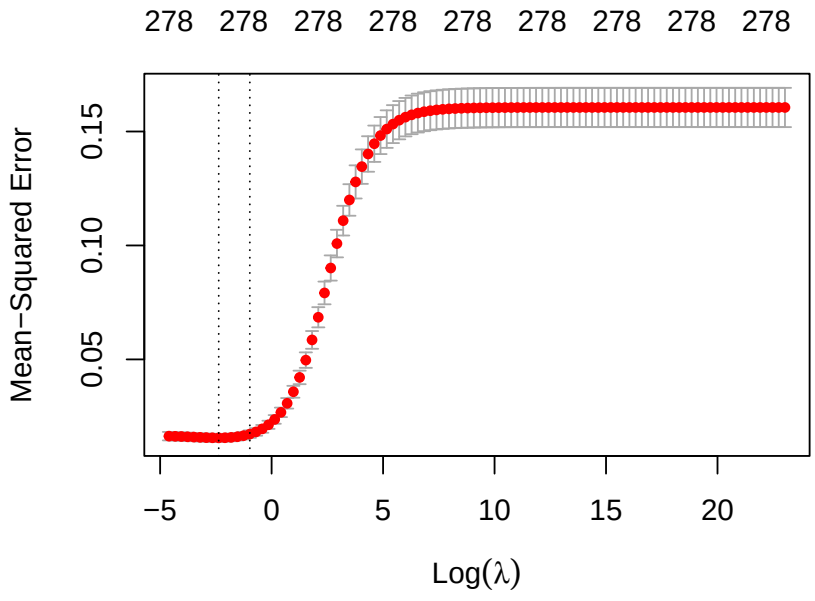


Table of Test MSE Values

Table 1: Test MSE Values for Different Models

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

Outside Exploration

- Wanted to attempt to use best model to predict sale prices in Northern Colorado
- Worked with co-owner of Deluxe Homes LLC
- Some issues were discovered:
  - Using models:
    - Overall Quality and Neighborhood are significant
    - Quality is subjective
    - Neighborhood is very different from Ames, Iowa
  - Additional aspects not considered:
    - Supply and price of building materials can change
    - Soil quality will impact price of foundation
- Data snapshot in time

References

Ames, Iowa: Alternative to the Boston Housing Data as an . . . http://jse.amstat.org/v19n3/decock.pdf.

"Convert Character to Factor in R: Vector, Data Frame Columns & Variable." Statistics Globe, 14 June 2021, https://statisticsglobe.com/convert-character-to-factor-in-r.

"House Prices - Advanced Regression Techniques." Kaggle, https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Sprecher, Stu. "Deluxe Homes LLC Housing Prices." 1 Dec. 2021.