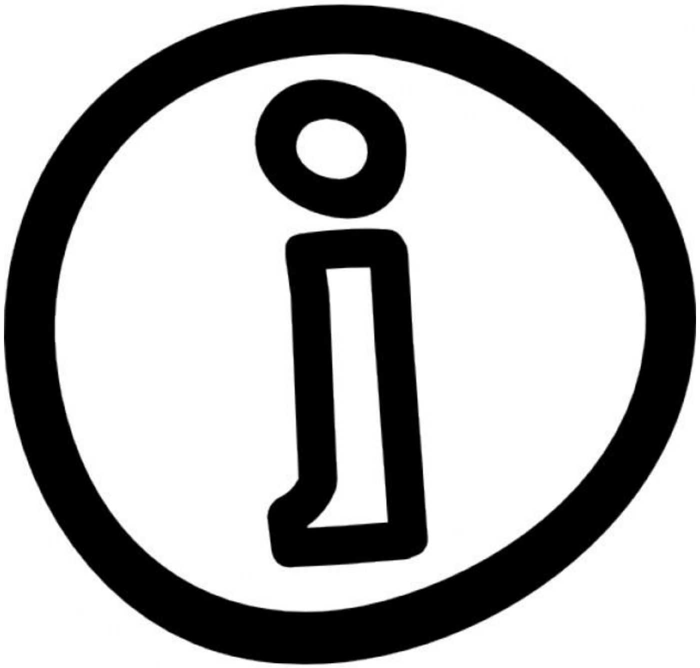
Predicting Housing Sale Prices

CKME 136 – Capstone Course

Spring/Summer 2017 | William Wong

RYERSON UNIVERSITY

Introduction



Research Question

How can we predict housing prices as accurately as possible?

- ***Dataset:*** Understanding the relationships between different variables.
- ***Feature Selection:*** Selecting a subset of relevant features for use in model construction
- ***Model Selection:*** Picking the model that fits the relation between your predictor and response variables. Selecting the model that gives you the best accuracy

Our Dataset

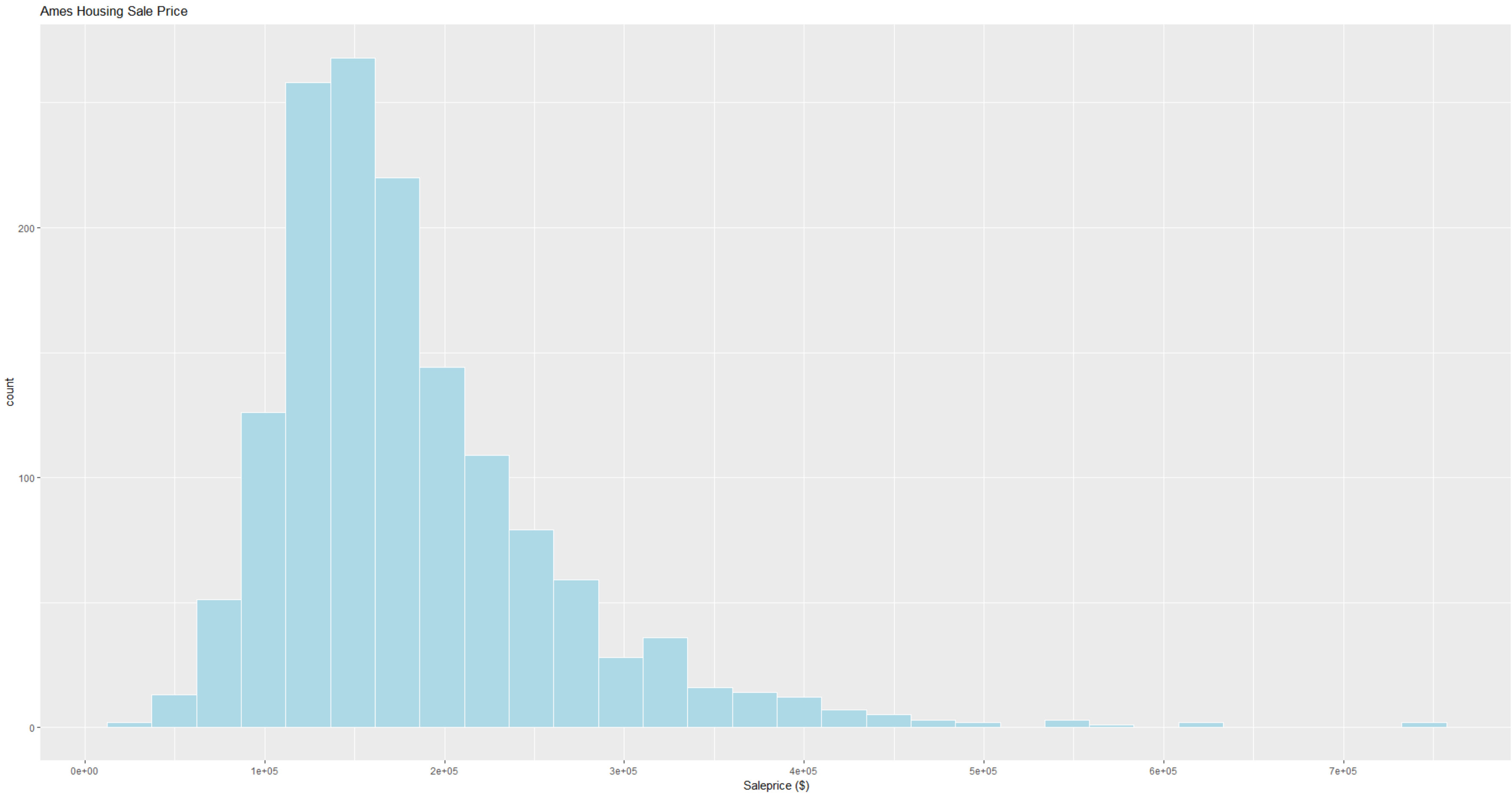
Ames Housing Dataset

- ***Description:*** Contains the sales of residential properties in Ames, Iowa, during the period from 2006 to 2010.
- ***From:*** The data was originally collected from Ames City Assessor's Office. Our data was retrieved from Kaggle's competition website.

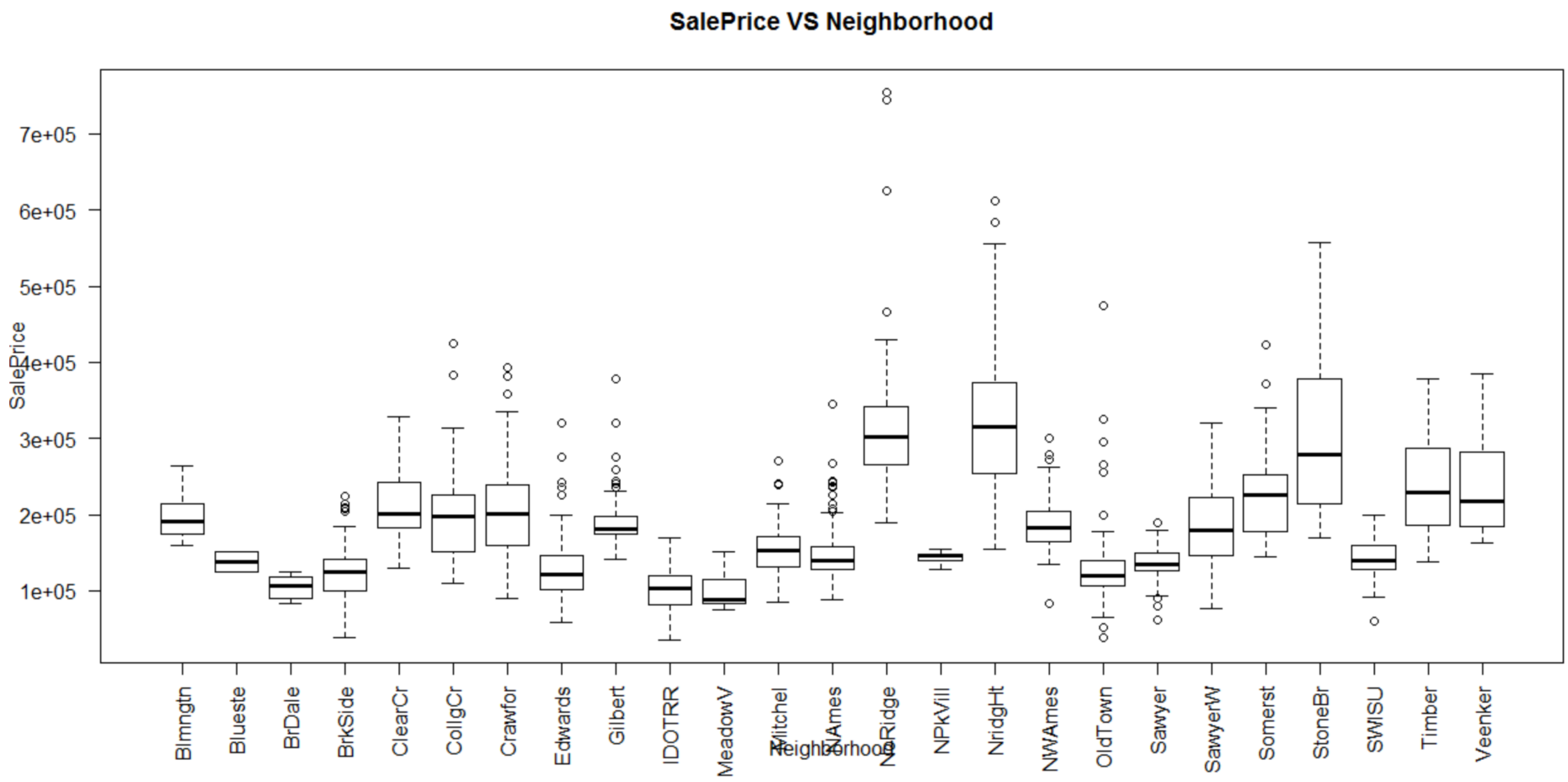
- ***Overview:*** There are 1460 observations and 81 attributes in this dataset.
- ***Data Type:*** Combination of 38 numeric and 43 categorical (nominal & ordinal) attributes.
- ***Response Variable:*** SalePrice!

Ames Housing Data

Response Variable: Sale Price



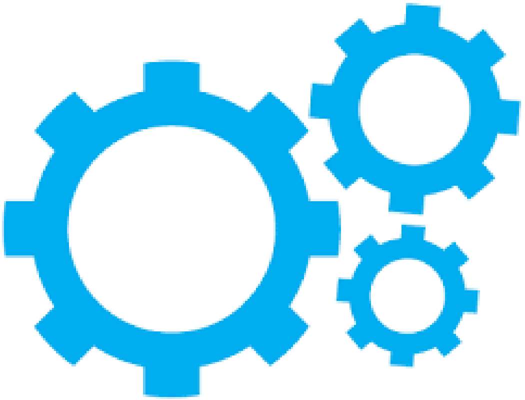
Ames Housing Data

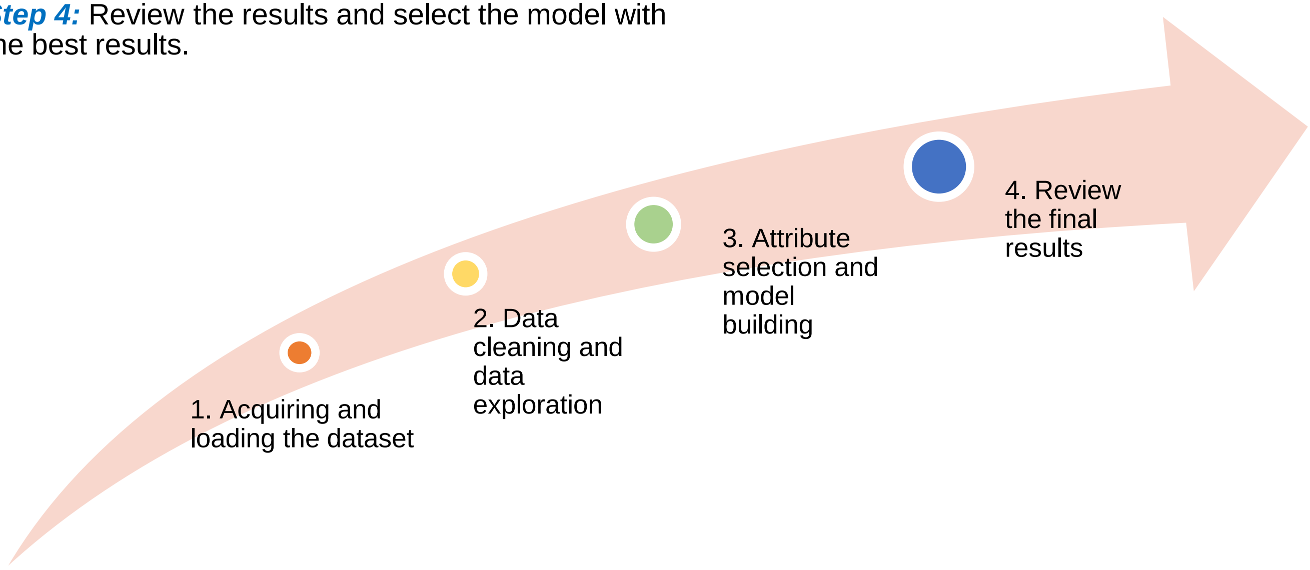


Our Approach

The Plan:

- ***Step 1:*** Download the data and uploading it to the R environment.
- ***Step 2:*** Explore the structure of the data. Examine the relationships between attributes. Prepare the data for modeling by filling in the NAs, removing outliers and transforming the data.
- ***Step 3:*** Select a set of attributes to model our predictions. The models used include decision tree, multiple regression and random forest.
- ***Step 4:*** Review the results and select the model with the best results.





Step 2: Data Cleaning & Data Exploration

Finding the number of NAs in our dataset

- There's a total of 19 attributes that have at least one missing value

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

Related values?

- There won't be a pool quality rating if the house does not have a pool.
- These can be filled with "0"s or "None".

```
> head(pool,5)
  PoolArea PoolQC
1        0   <NA>
2        0   <NA>
3        0   <NA>
4        0   <NA>
5        0   <NA>
```

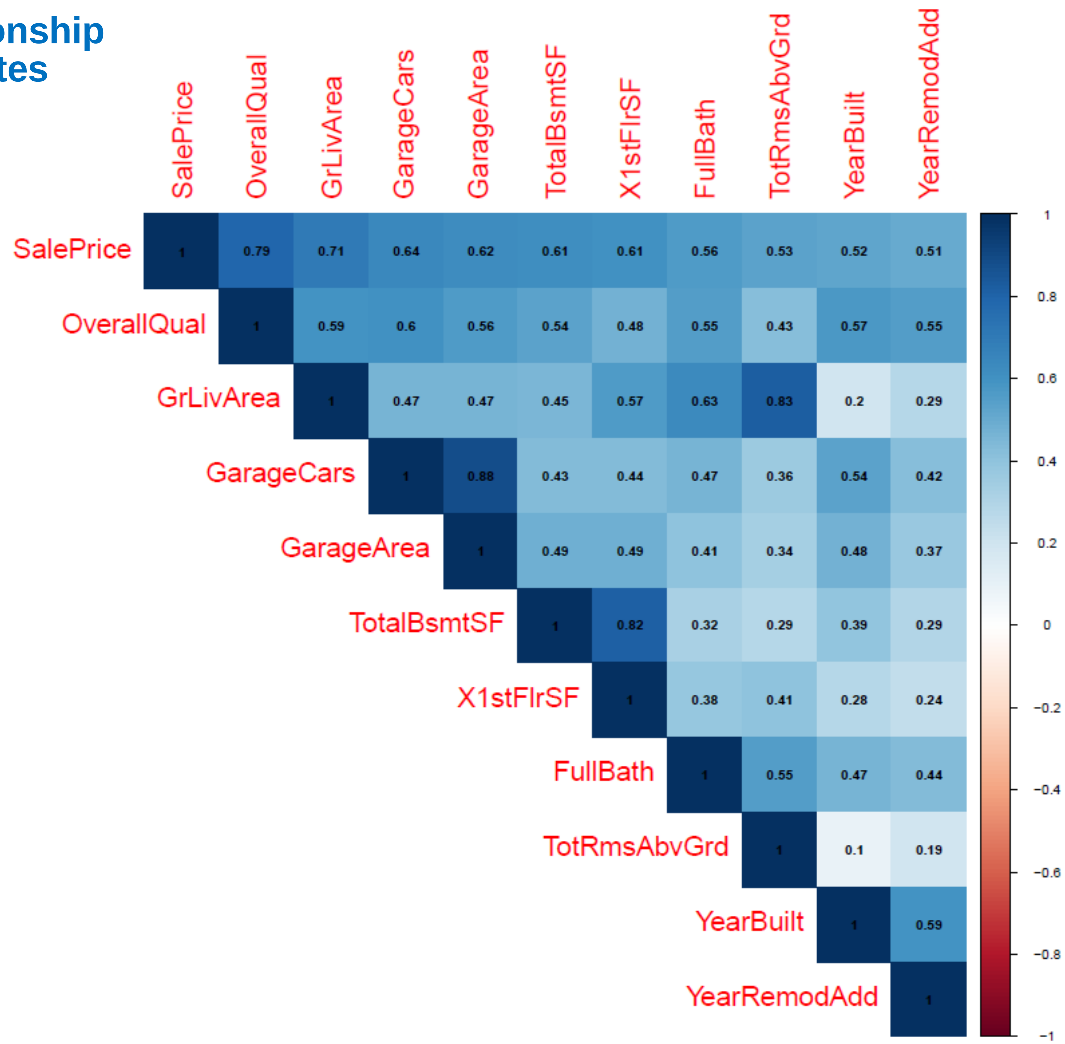
Real missing values

	LotFrontage	Neighborhood
8	NA	NWAmes
13	NA	Sawyer
15	NA	NAmes
17	NA	NAmes
25	NA	Sawyer
32	NA	Sawyer
43	NA	SawyerW
44	NA	CollgCr
51	NA	Gilbert
65	NA	CollgCr
67	NA	NAmes
77	NA	NAmes
85	NA	Gilbert

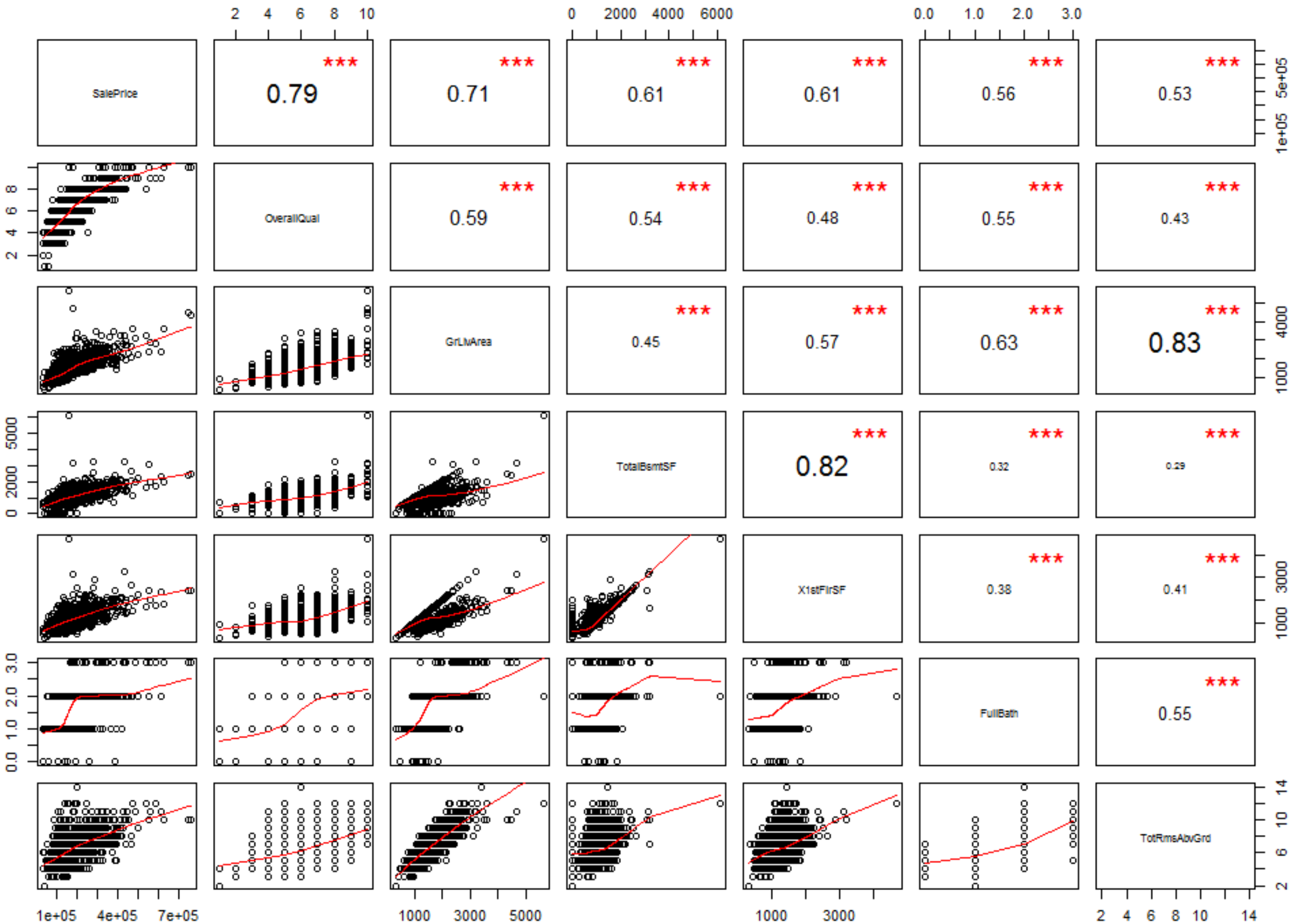
- Missing values can be filled in with the attribute's average value.
- ***Lot Frontage:*** Assumed lot frontage is similar within a specific neighborhood. Average lot frontage values were calculated for each neighborhood. Filled in the missing lot frontage values based on the neighborhood is in.

Step 2: Data Cleaning & Data Exploration

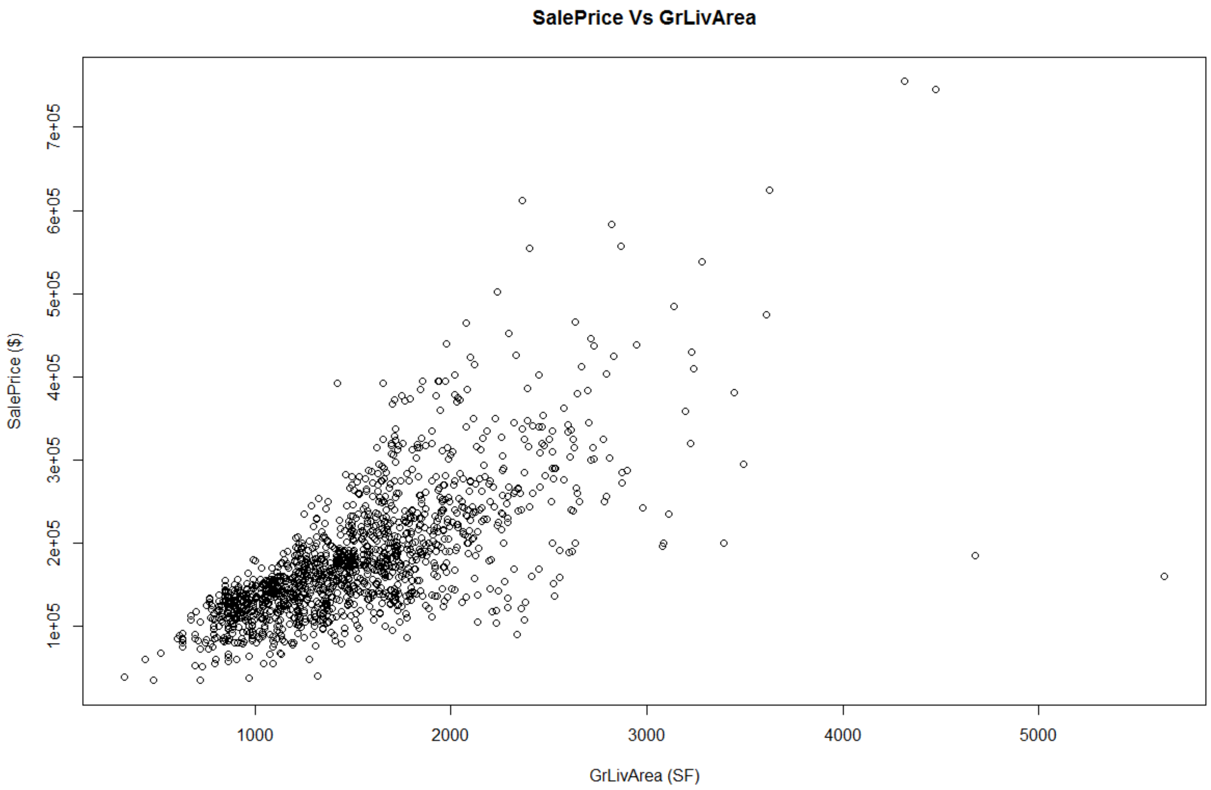
Examine the relationship between attributes



Step 2: Data Cleaning & Data Exploration



Step 2: Data Cleaning & Data Exploration

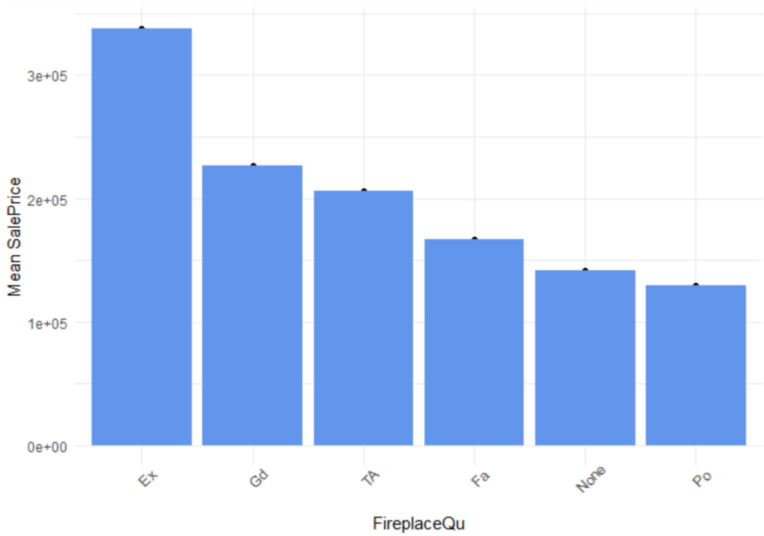


Removing Outliers

- There are 4 houses with an above grade area over 4000 square feet. These houses are significantly larger than the rest of the data set.
- Two of the largest houses are priced at a very low sale price.

Step 2: Data Cleaning & Data Exploration

Fireplace Quality Example:



	FireplaceQu	Mean.Price	Count
1	Po	129764.1	20
2	None	141331.5	690
3	Fa	167298.5	33
4	TA	205723.5	313
5	Gd	226351.4	380
6	Ex	337712.5	24

Transforming ordinal data into numeric data

- Houses with a better quality fire place will generally yield a higher selling price.
- We can assign a numeric value to each of the ordinal values to transform them into numeric data.

Rating	Assigned Value
None	0
Poor (Po)	1
Fair (Fa)	2
Average (TA)	3
Good (Gd)	4
Excellent (Ex)	5

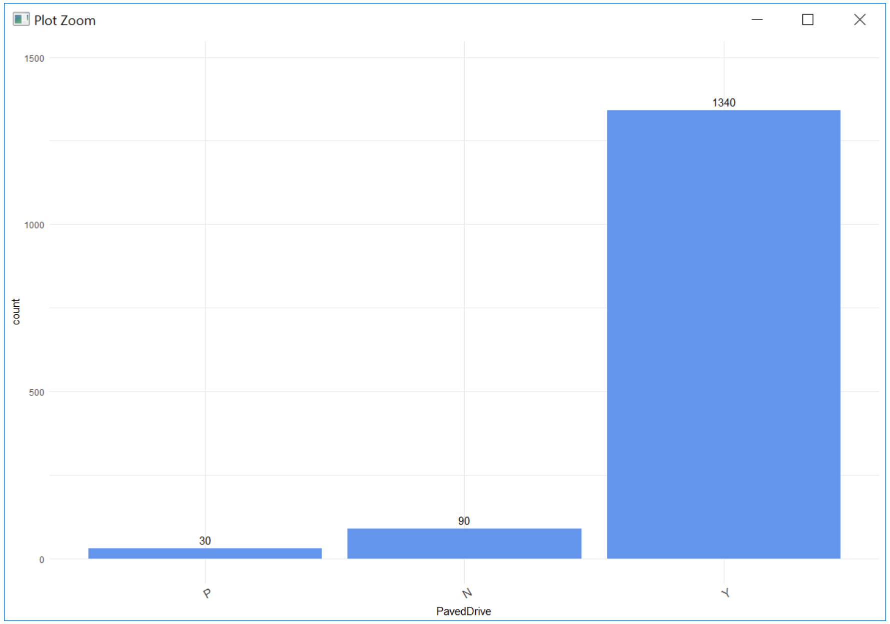
Step 2: Data Cleaning & Data Exploration

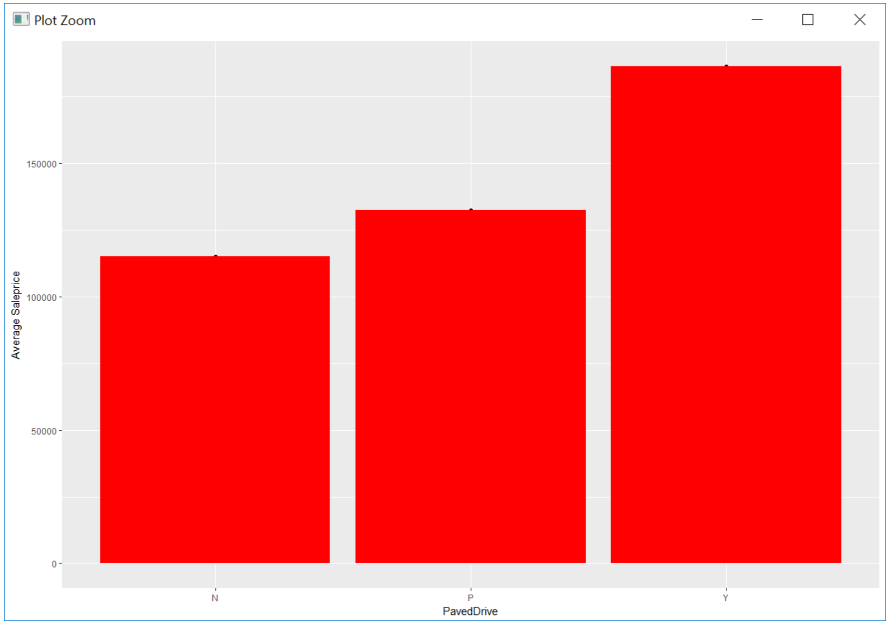
Transforming nominal data into numeric data

- Houses with a paved driveway are sold at a higher price than houses that have a partially paved or non-paved driveway.
- We can transform these variables into binary variables assigning the value 1 for paved and 0 for non-paved or partially paved.

Variable	Assigned Value
Paved Driveway (Y)	1
Non-Paved (N) or Partially Paved (P)	0

Paved Driveway Example:





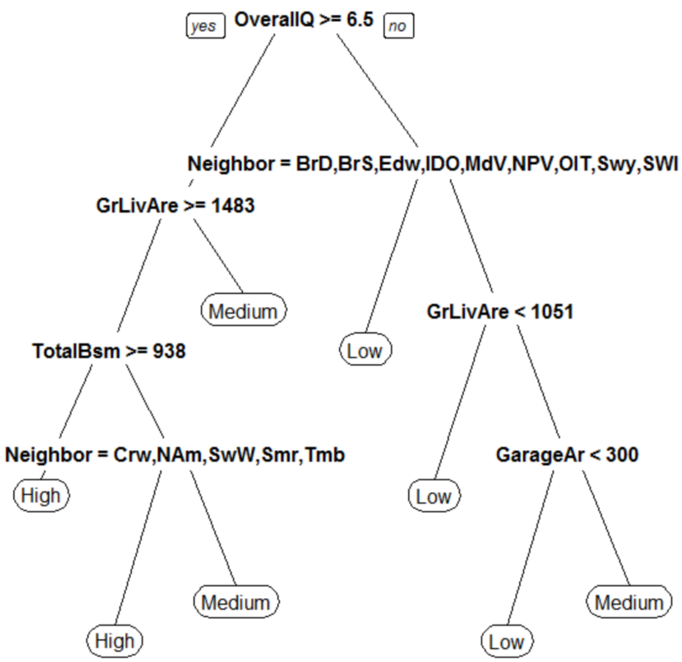
Step 3: Attribute Selection & Model Building

	Description/Attribute Selection	Evaluation
Decision Tree	The pricing level attribute was created and we classified each house based on the sale price range they fell into. A decision tree model will be used to predict the pricing level of the house. All attributes will be used in this model.	• Data split into 70/30 training and testing sets. • Model will be evaluated by accuracy and true positive rate.
Multiple Regression	There will be 3 variations of the multiple regression model. Each variation will run on a different set of attributes to predict the sale price. 1. Regression with 10 highest correlated variables 2. Regression with all 89 variables 3. Regression with stepwise attribute selection	• Data split into 70/30 training and testing sets. • Model will be evaluated by the root mean square error (RMSE) and coefficient of determination, R^2.
Random Forest	The random forest model will use all 89 attributes in the dataset to predict sale prices. The number trees in the model is 500. The total number of variables tried at each split is 29.	• Data split into 70/30 training and testing sets. • Model will be evaluated by the root mean square error (RMSE) and coefficient of determination, R^2.

Price Level	Sale Price Range
Low	Less than $140,000
Medium	$140,000 to $200,000
High	Greater than $200,000

Step 4: Review Results

Decision Tree Model



Confusion Matrix

Pricing Levels	Predicted: High	Predicted : Low	Predicted: Medium
Actual: High	99	0	12
Actual: Low	3	144	43
Actual: Medium	27	27	82

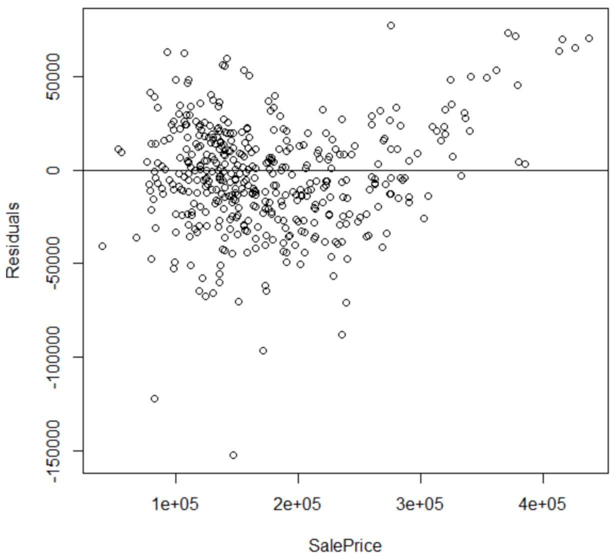
Results

Pricing Levels	High	Low	Medium
True Positive	0.8919	0.7579	0.6029
Model Accuracy	0.7437		

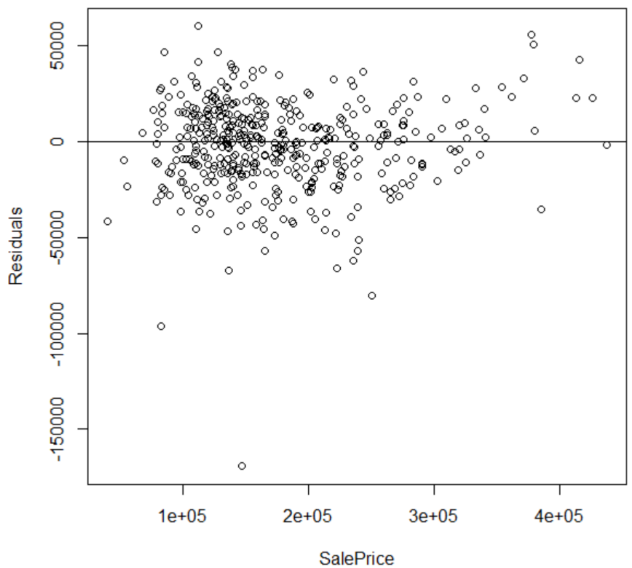
Step 4: Review Results

Multiple Regression Model

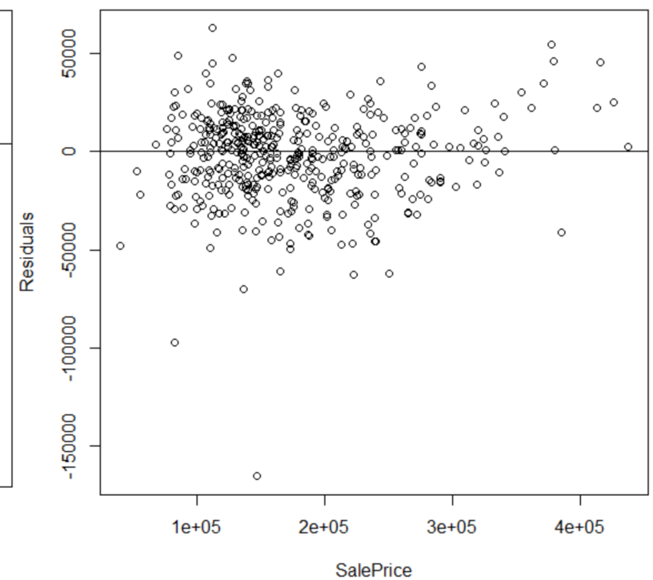
Residual Plot - Regression Model 1



Residual Plot - Regression Model 2



Residual Plot - Regression Model 3

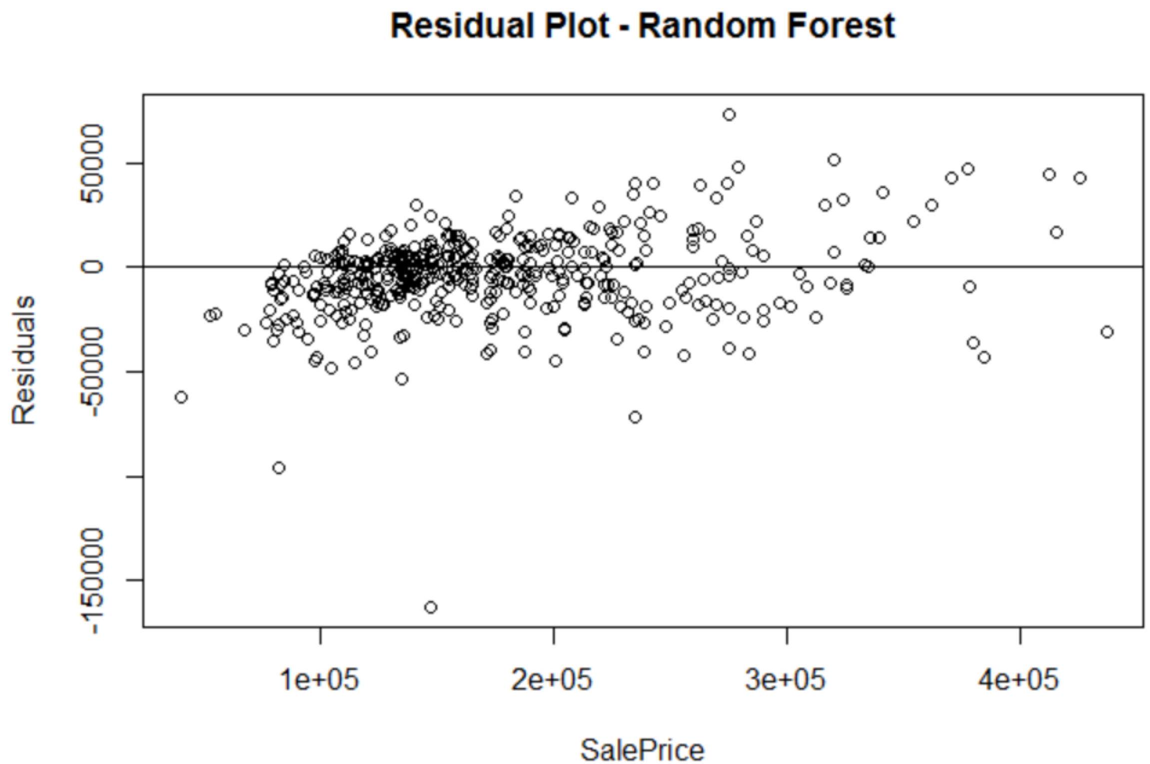


Results:

Model	RMSE	% of Variance Explained
Linear Regression (with 10 correlated variables)	$28,781.54	81.58
Linear Regression (with all variables)	$23,155.12	89.64
Linear Regression (with Stepwise regression)	$22,697.29	89.88

Step 4: Review Results

Random Forest Model



Results

Model	RMSE	% of Variance Explained
Random Forest	$20,365.42	89.22

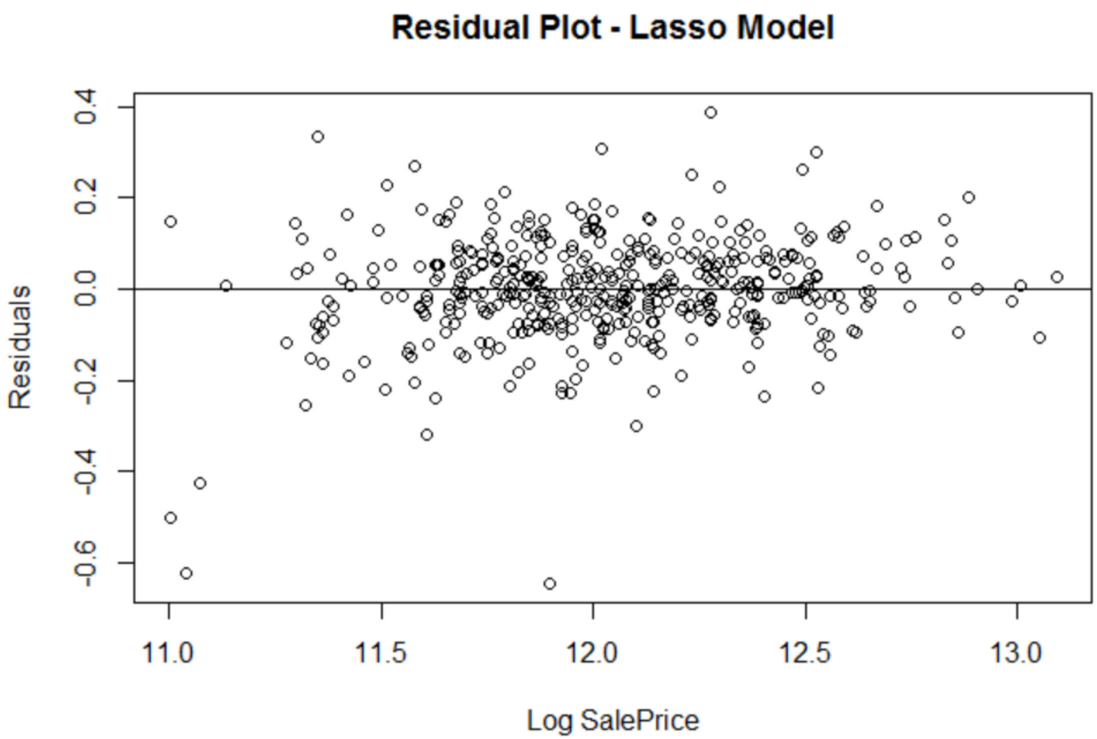
Step 4: Review Results

Final Results:

Model	RMSE	% of Variance Explained
Linear Regression (with 10 correlated variables)	$28,781.54	81.58
Linear Regression (with all variables)	$23,155.12	89.64
Linear Regression (with Stepwise regression)	$22,697.29	89.88
Random Forest	$20,365.42	89.22

Bonus!

LASSO Model:



Results

Model	RMSE ($)	RMSE (Log)
LASSO	$19,543.60	0.112658