

## Distributed Computing Report

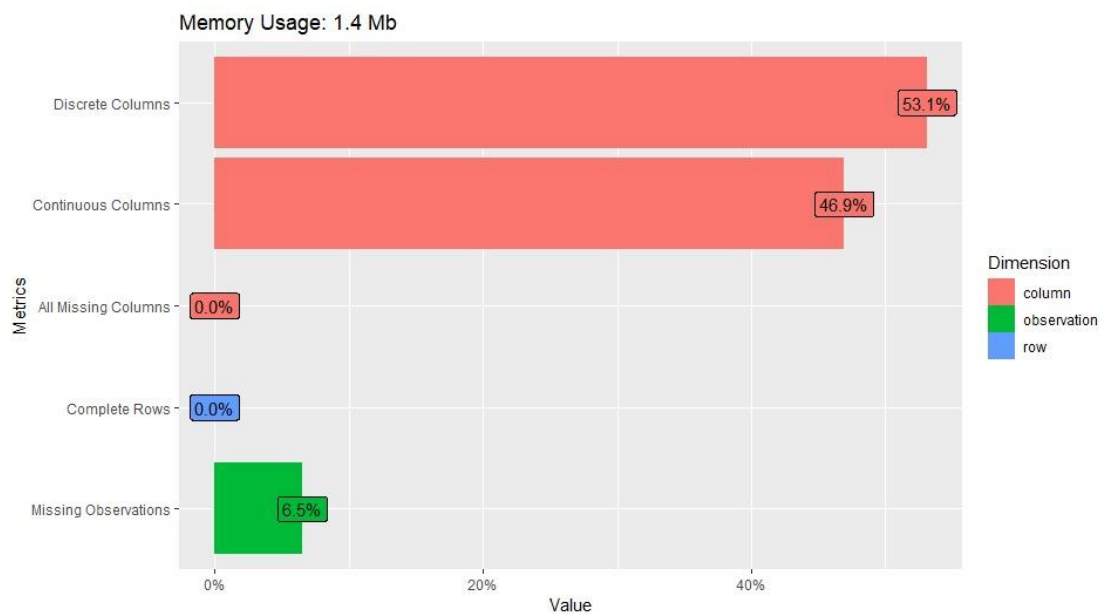
Name	ID
محمد احمد عبدالرحمن	20191700494
ماهر احمد علي	20191700483
يوسف سامح رشدي	20191700769
يحيى حسانين محمد محمود	20191700753

### Exploring Data:

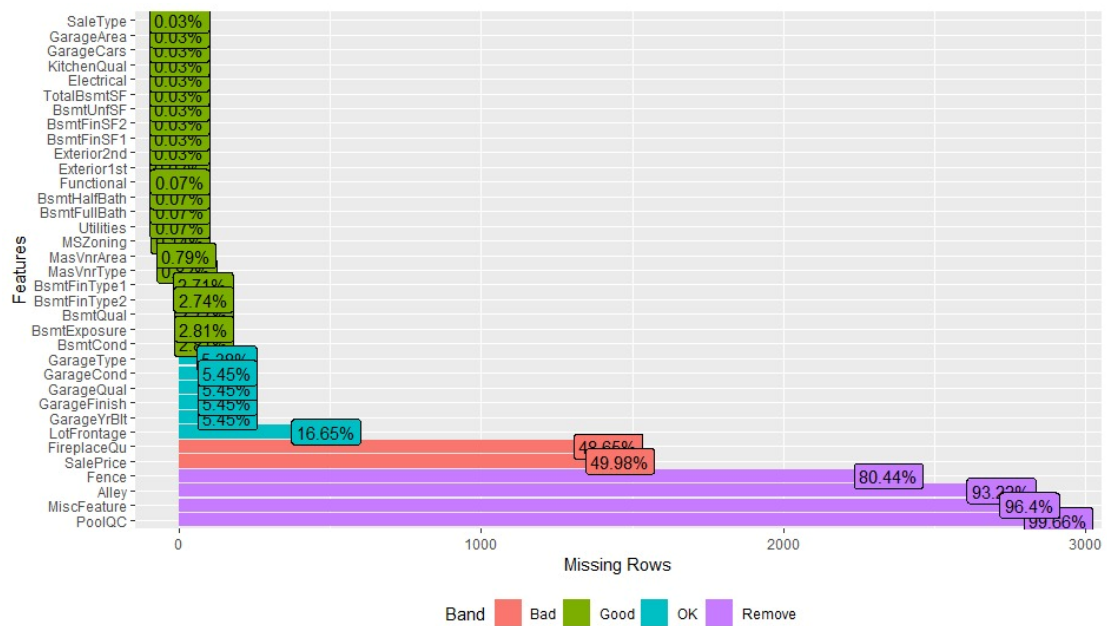
Train Data Dimension: 1460 row x 81 column.

Test Data Dimension: 1459 row x 80 column.

Main Dataframe: 2919 row x 81 column.

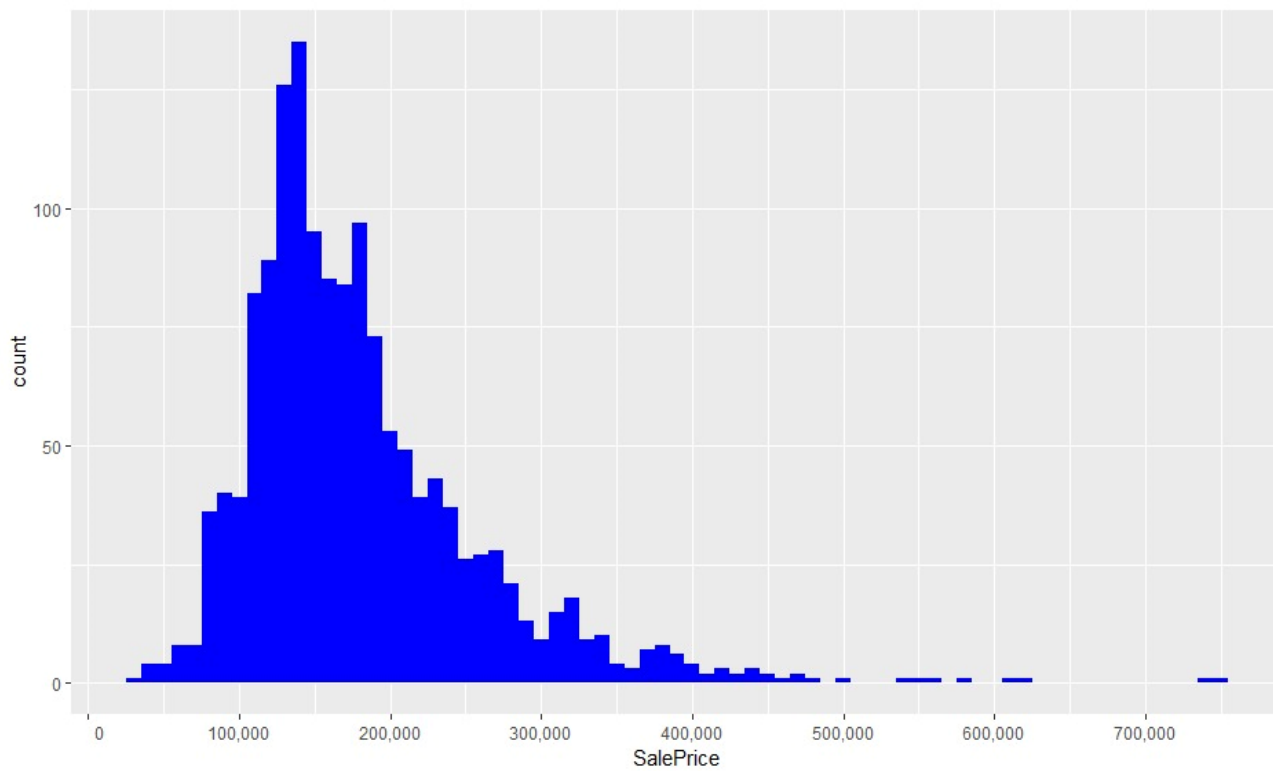


There are 38 numeric variables, and 43 categoric variables.

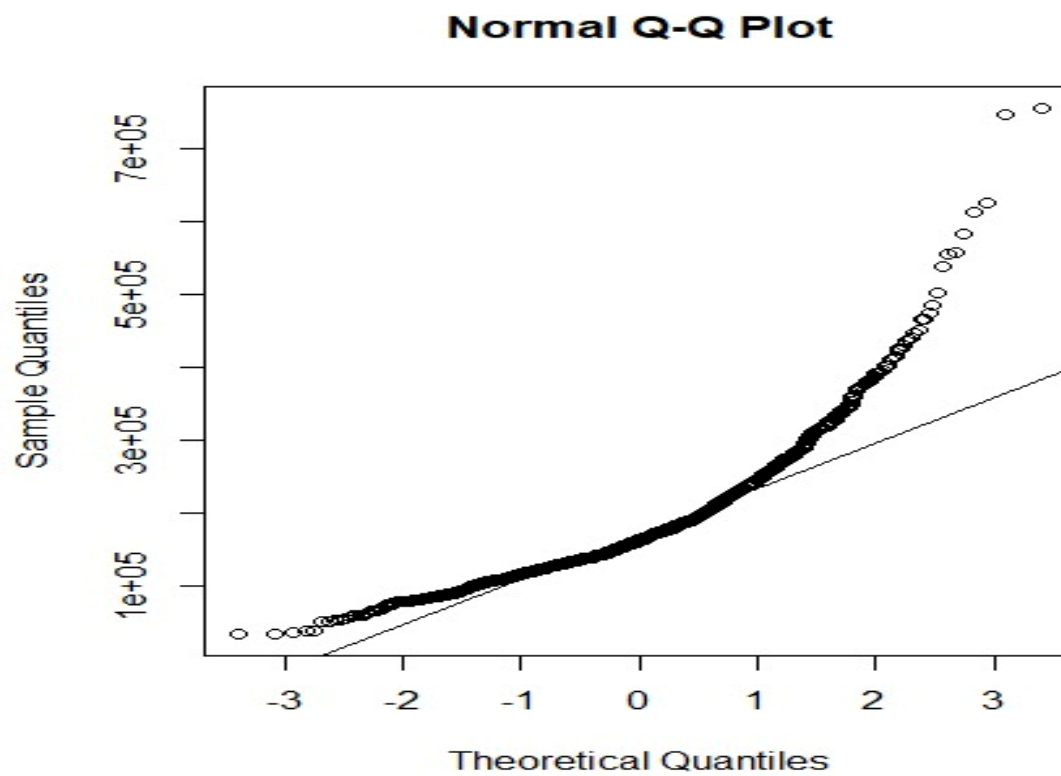


## Analyzing Target feature:

Min	. 1st Qu	Median	Mean	3rd Qu	Max
755000	214000	180921	163000	129975	34900



Skew: 1.87900



## **Handling missing data:**

Drop these columns due too much missing data

(PoolQC, MiscFeature, Alley, Fence, Utilities).

-> Replaced missing values in numeric columns with (0).

-> Replaced missing values in Characteristic columns with (none).

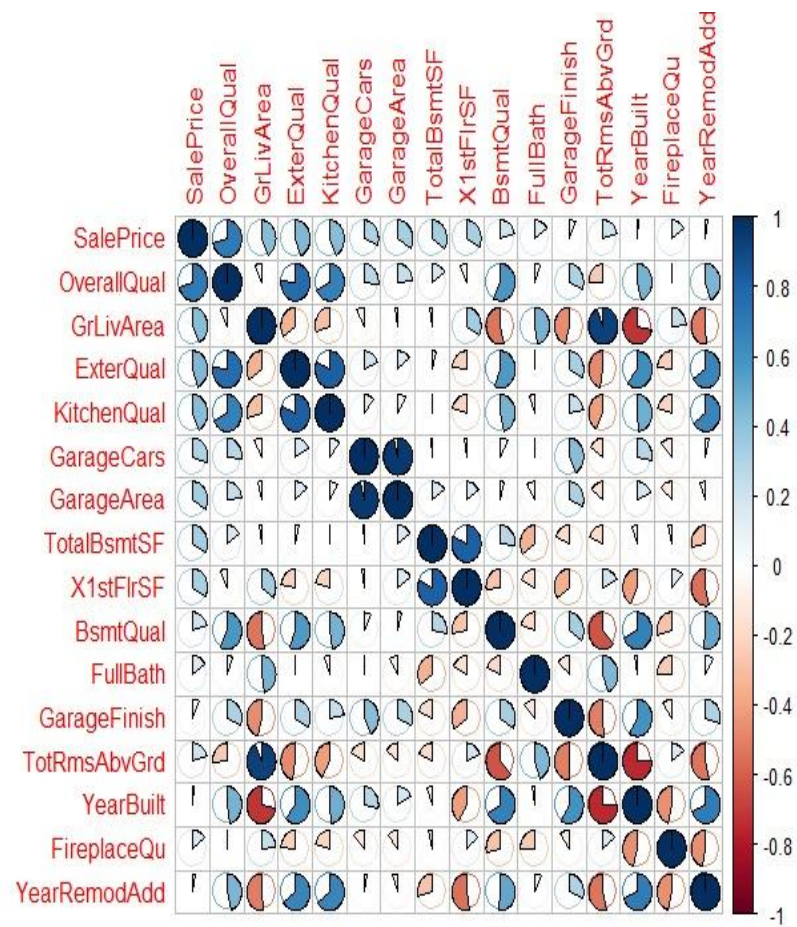
## **Encoding & Factorizing:**

We made revalue on this columns (FireplaceQu , LotShape , GarageType , GarageFinish , GarageQual , GarageCond , BsmtQual , BsmtCond , BsmtExposure BsmtExposure , BsmtFinType1, BsmtFinType2 , HeatingQC, PavedDrive, LandSlope, KitchenQual ExterCond ,ExterQual ,Functional ,MasVnrType TotalBsmtSF, BsmtUnfSF, BsmtFinSF2, BsmtFinSF1 BsmtFullBath, BsmtHalfBath)

We made factorization on this columns (MSSubClass, MoSold, Heating, Foundation, RoofMatl, RoofStyle ,HouseStyle, BldgType, Condition2, Condition1, Neighborhood, LandContour, SaleCondition, SaleType, Electrical, MSZoning, LotConfig).

After handling missing data, encoding and factorization there are 58 numeric variables, and 18 categoric variables.

### Correlation:



## **Feature Engineering:**

->**Total\_Bathrooms**: There are 4-bathroom variables. Individually, these variables are not very important. However, we assume that if we add them up into one .predictor, this predictor is likely to become a strong one

$\text{BsmtHalfBath} + \text{BsmtFullBath} + \text{HalfBath} + \text{FullBath}$

->**Age**:  $\text{YrSold} - \text{YearRemodAdd}$

->**Is new**: It is a condition that checks if the year built = year sold

->**Total\_Sq\_Feet**: As the total living space generally is very important when people buy houses, I am adding a predictors that adds up the living space above and below ground equals  $\text{TotalBsmtSF} + \text{GrLivArea}$

->**Total\_Home\_Quality**:  $\text{OverallCond} + \text{OverallQual}$

->**Neighborhood\_Class**: The highest-ranked(Sale price) neighborhoods are given a value of 2, followed by a value of 1 for middle-ranked neighborhoods, and a value of 0 for lowest-ranked neighborhoods

$[\text{MeadowV}, \text{IDOTRR}, \text{BrDale}] = 0$

$[\text{StoneBr}, \text{NridgHt}, \text{NoRidge}] = 1$

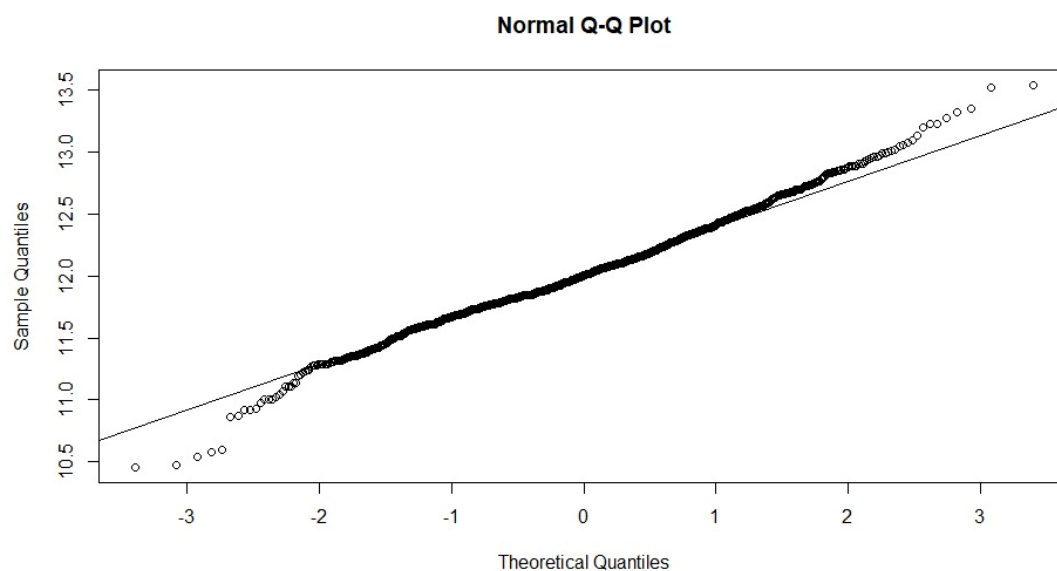
$[\text{StoneBr}, \text{NridgHt}, \text{NoRidge}] = 2$

We applied normalization on numeric columns and one-hot-encoding on categorical data.

### **Fixing target feature:**

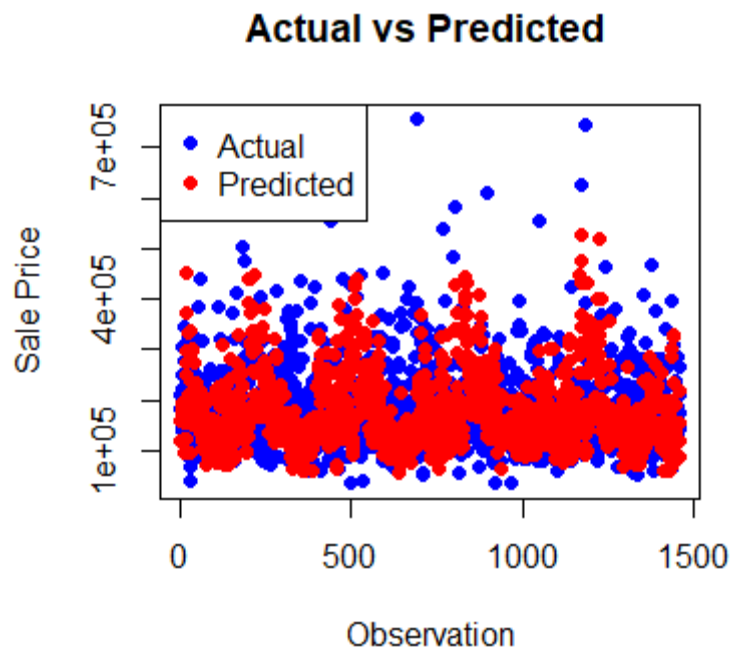
We used a log transformation to reduce the skewness of the target feature, which improved its distributional properties for modeling purposes.

Skew after log transformation = 0.1210859

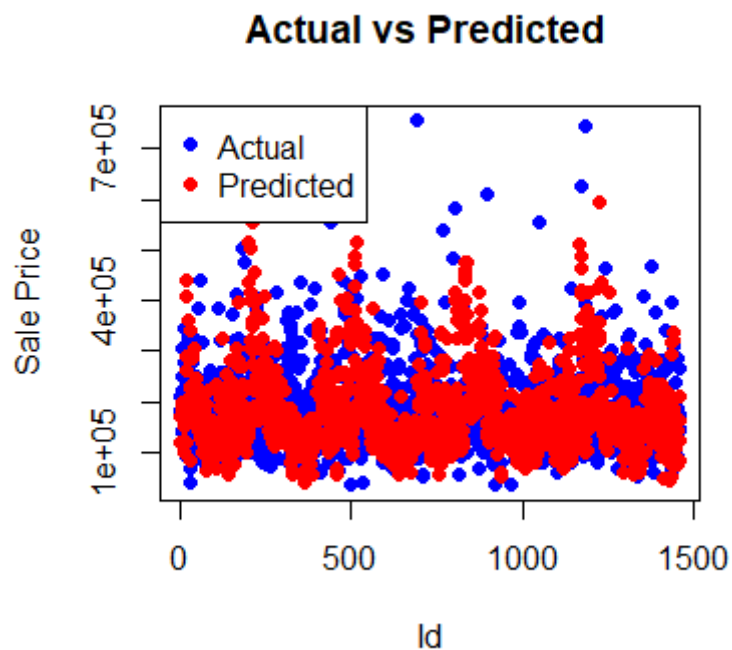


### **Modeling:**

Random forest: Random Forest model Root mean squared error: 0.5362701

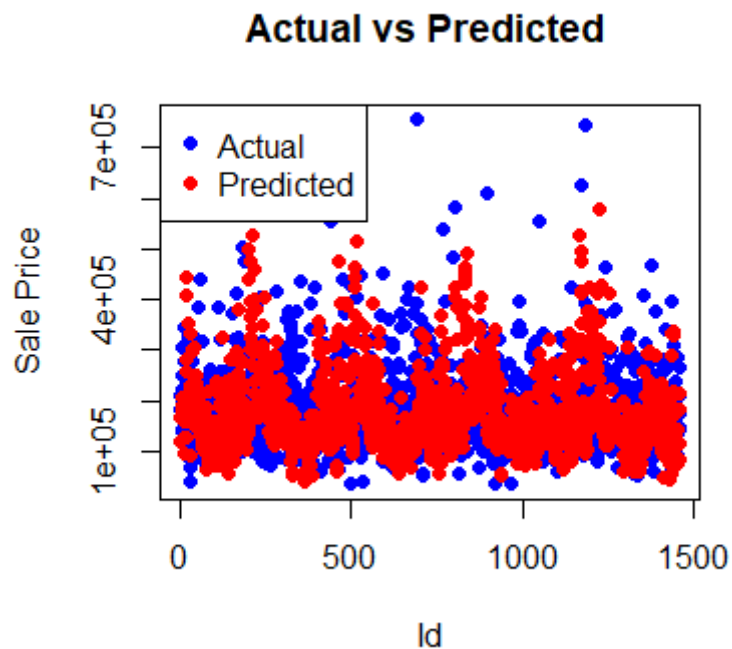


Gradient Boosting1: Gradient Boosting model Root mean squared error = 0.556455 , shrinkage = 0.05, n.trees=1000

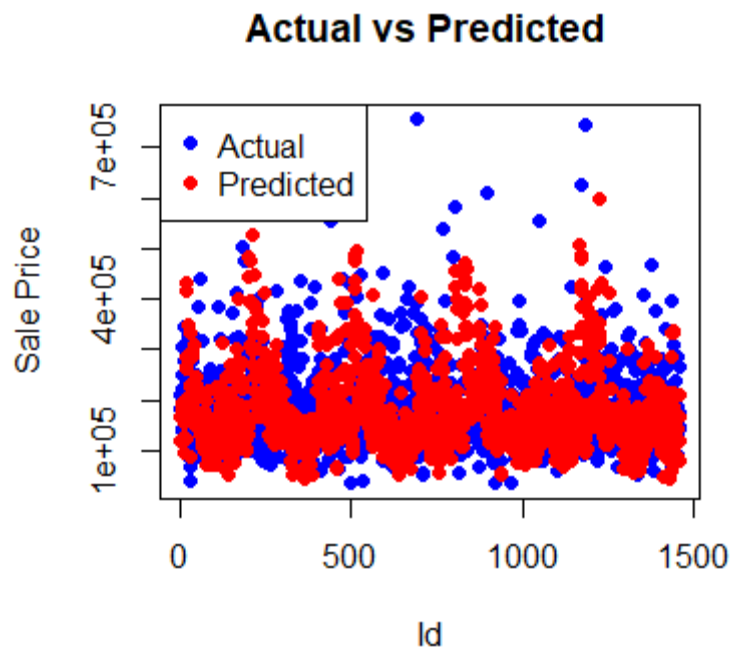




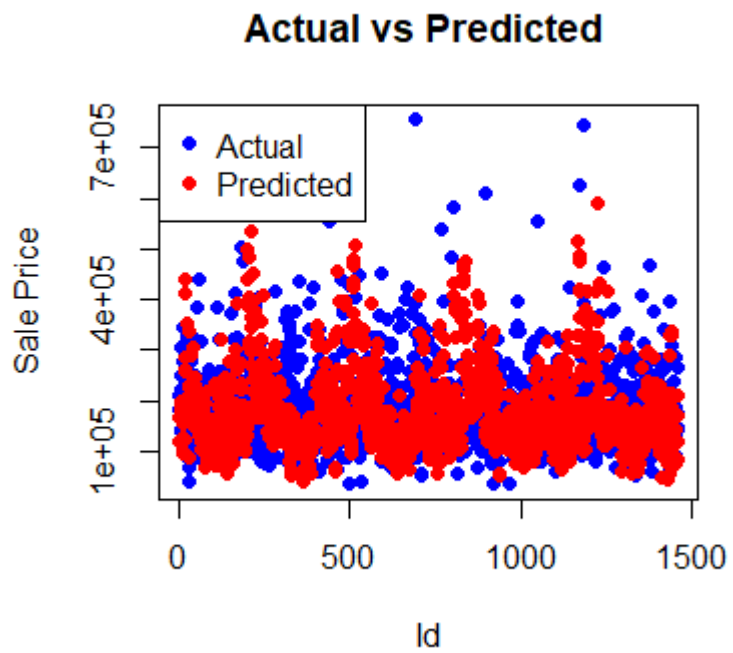
Gradient Boosting2: Gradient Boosting model Root mean squared error = 0.5581766 , shrinkage = 0.07, n.trees = 1500



Gradient Boosting3: Gradient Boosting model Root mean squared error = 0.5581766, shrinkage = 0.05, n.trees = 2000








Average Gradient Boosting: Root mean squared error  
=0.556431



# Accuracy at Kaggle competition:

Evaluation: RMSE

Submission and Description		Public Score 
	<b>M_M_GBM_1_2_3_v2.csv</b> Complete · 5d ago	<b>0.1222</b>
	<b>GBM_3_v2.csv</b> Complete · 5d ago	<b>0.12244</b>
	<b>GBM_2_v2.csv</b> Complete · 5d ago	<b>0.12416</b>
	<b>GBM_1_v2.csv</b> Complete · 5d ago	<b>0.1255</b>