



Hou\$e Price\$

A MACHINE LEARNING PROJECT

Gradient De\$endant\$

Agenda

Project aims at predicting house prices (residential) in Ames, Iowa, USA based on data set provided by Kaggle between 2006 and 2010.

- Data Exploration
- Data Pre-processing
- Feature Engineering
- Model Training and Evaluation
- Analysis, Conclusions and Considerations



**SO YOU HAVE A
MACHINE THAT CAN**

**PREDICT HOME
PRICES IN IOWA**

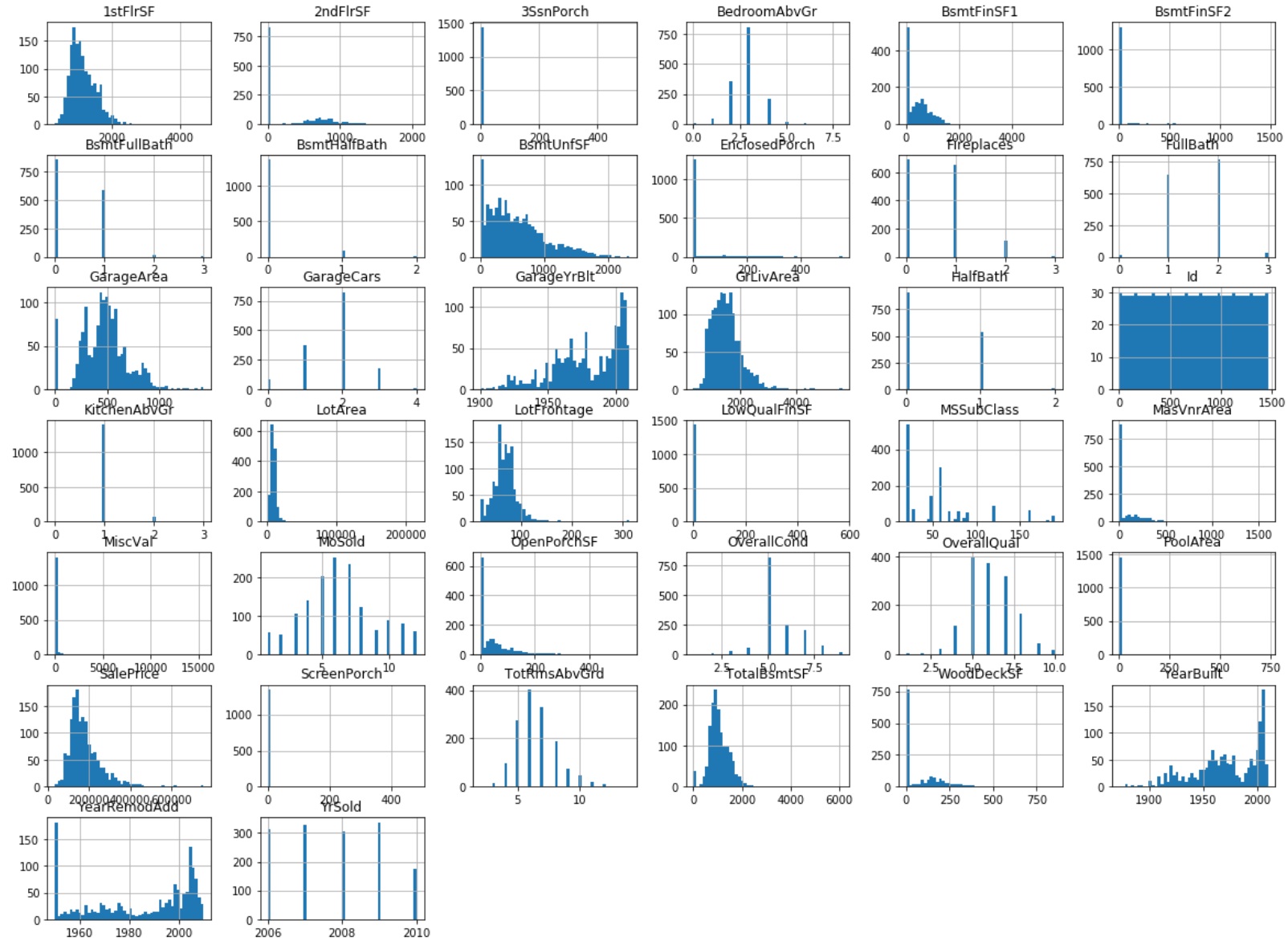


Data Exploration

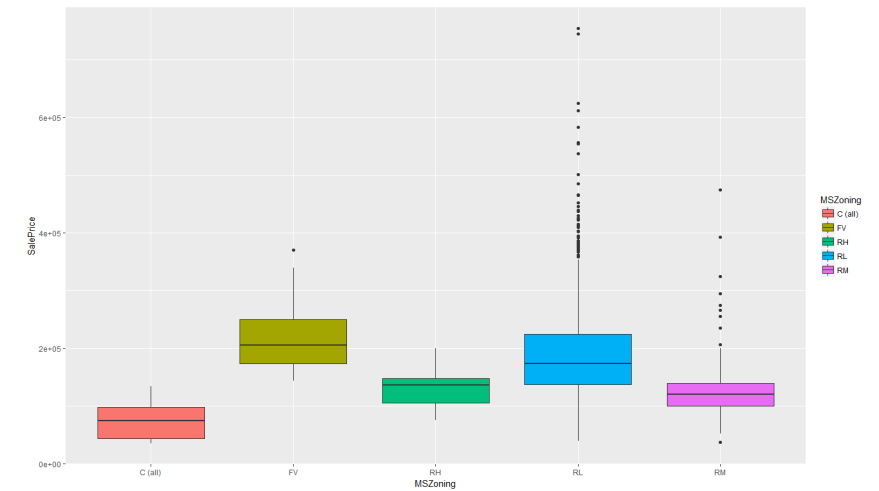
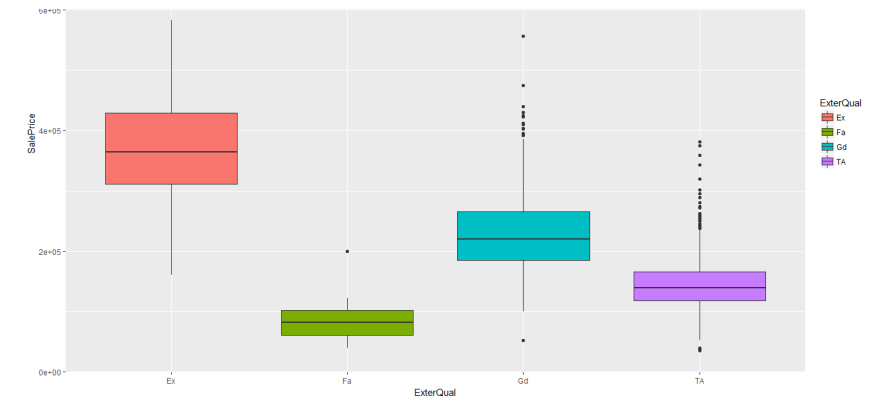
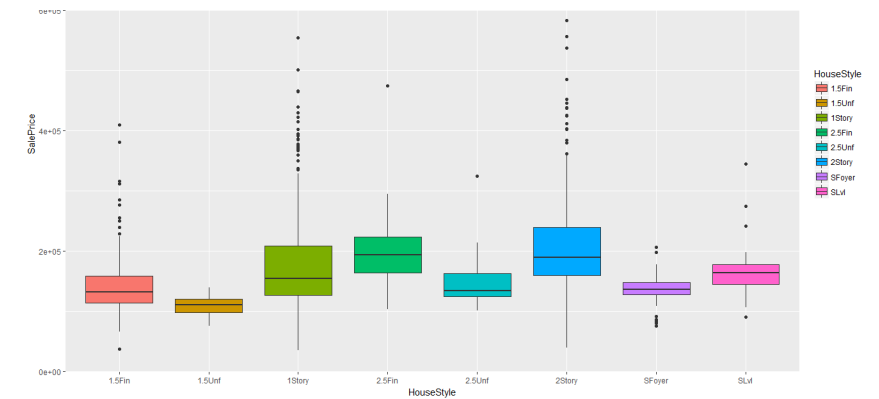
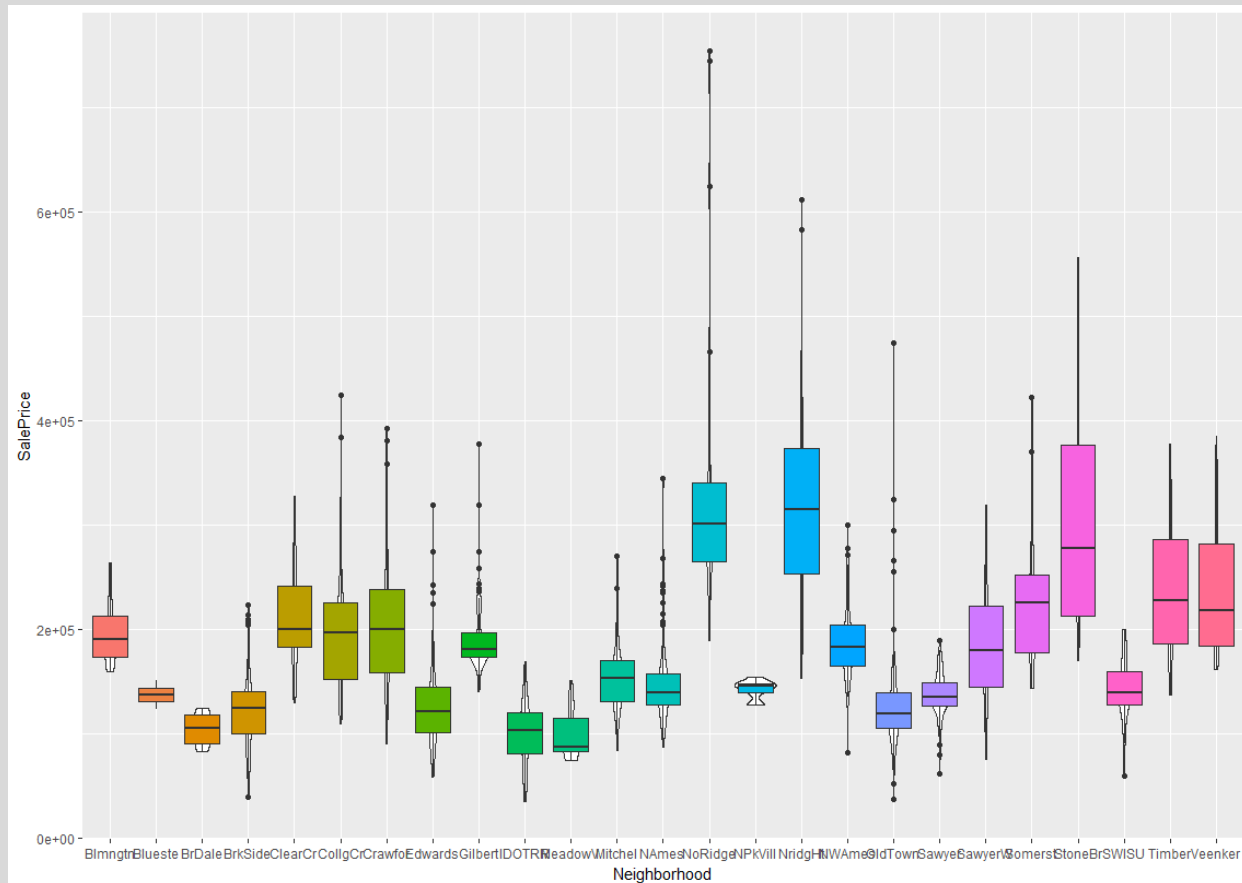
What factors do we believe to influence house prices?

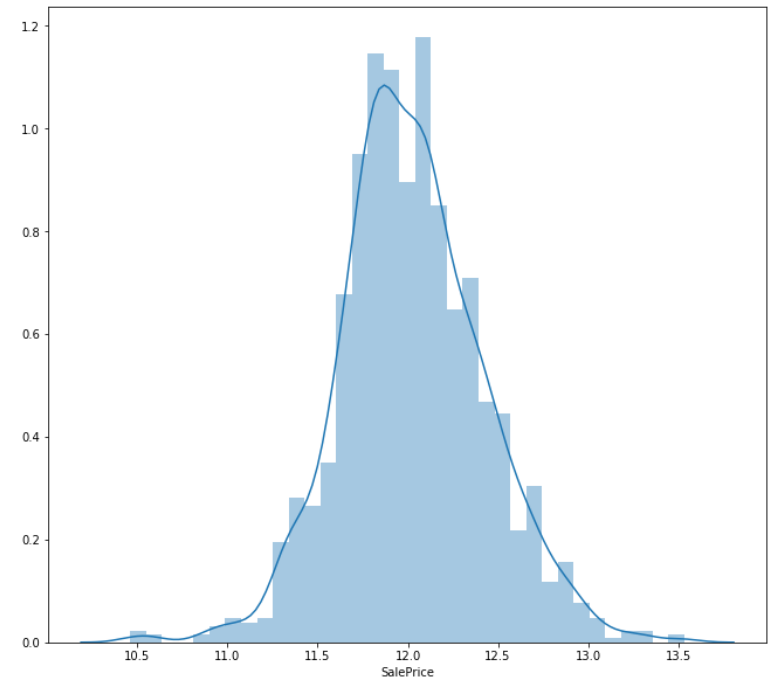
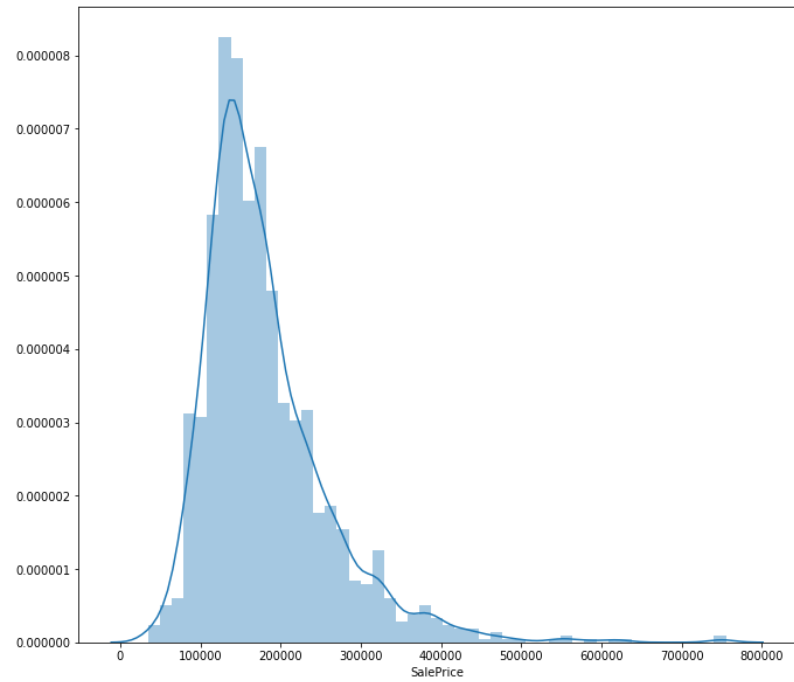
- The dataset contains 1460 observations in the training set and 1459 observations in the test set.
- There are 46 categorical variables including 23 nominal and 23 ordinal ones, and 33 numeric variables in the dataset.
- The training set has the sale price as response while the test set doesn't.

Univariate Analysis



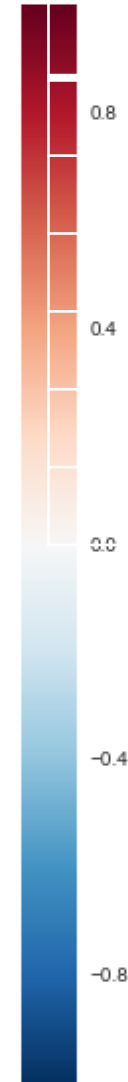
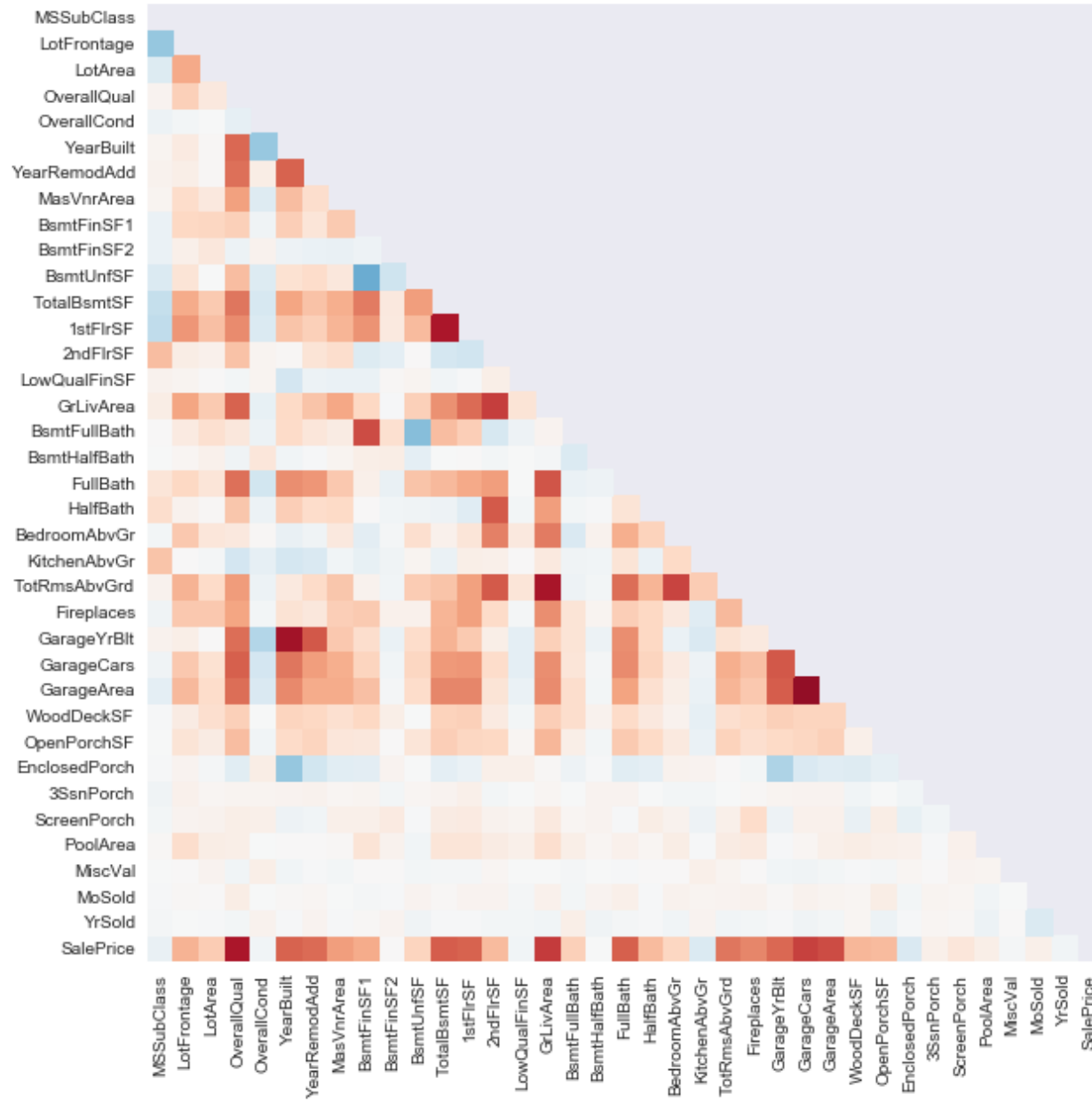
Qualitative vs Sale Price





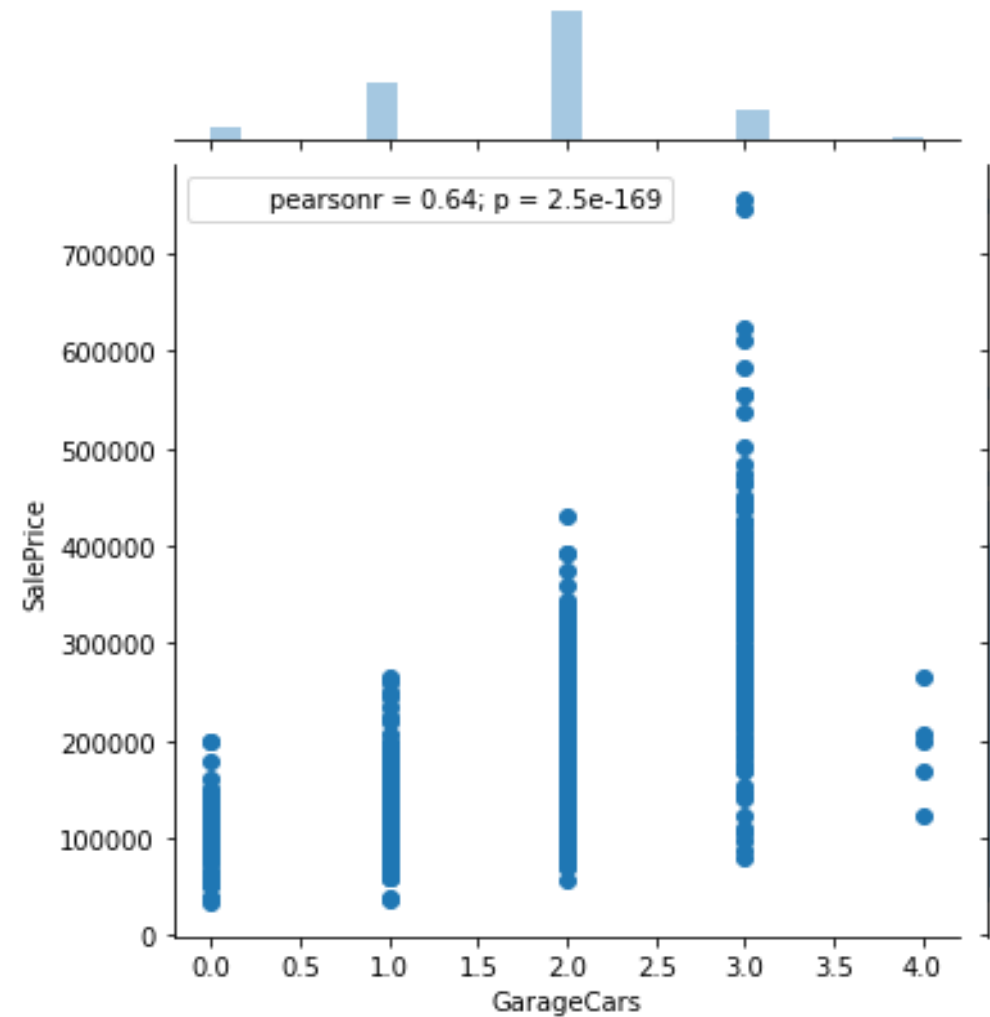
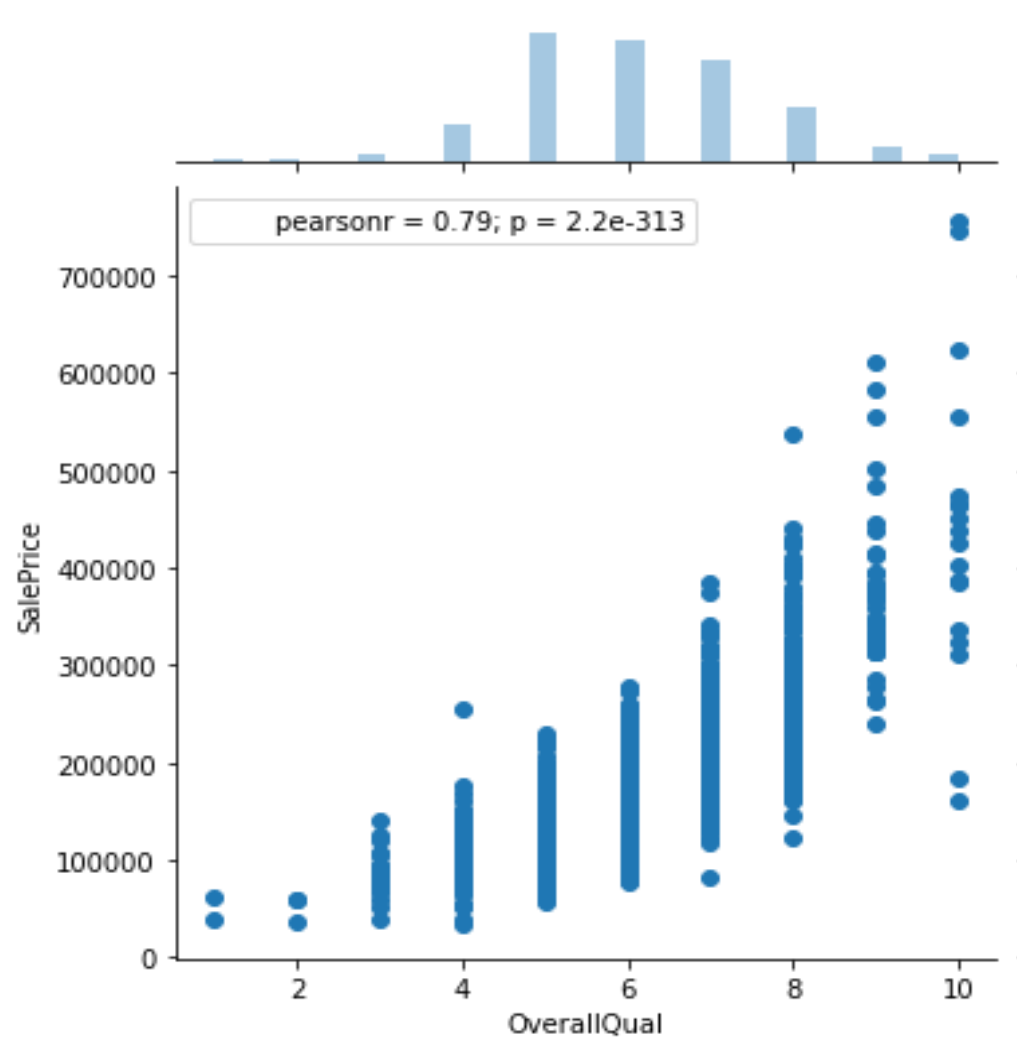
Log(Sale Price)

Correlation



Variables	Correlation
overallqual	0.79
Totalbsmtsf	0.61
1stflrsf	0.61
Grlivearea	0.71
Garagecars	0.64
Garagearea	0.62

Overall Quality & Garage

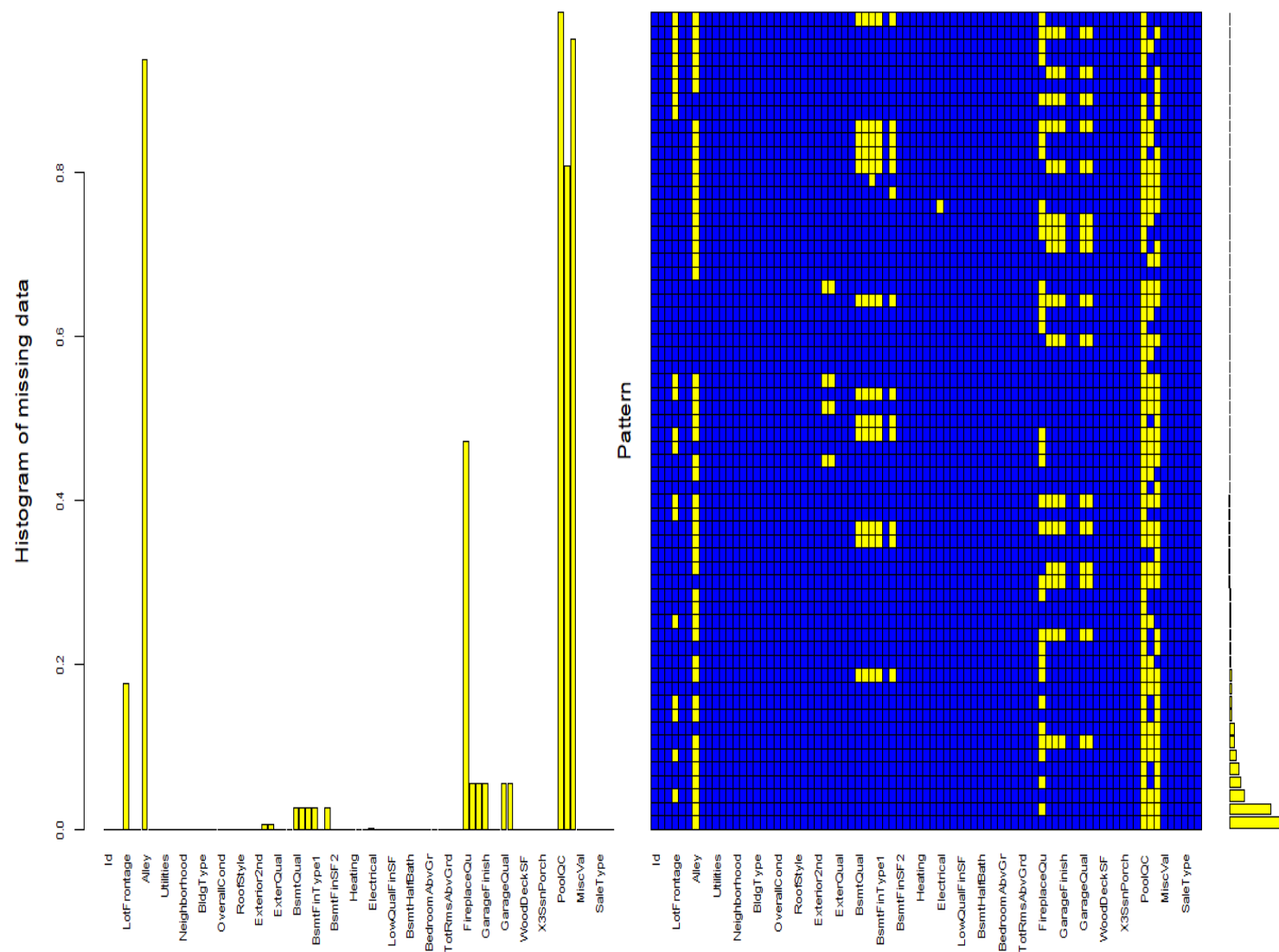




Data Pre-processing

I thought I told you to clean the basement!

Missing Values

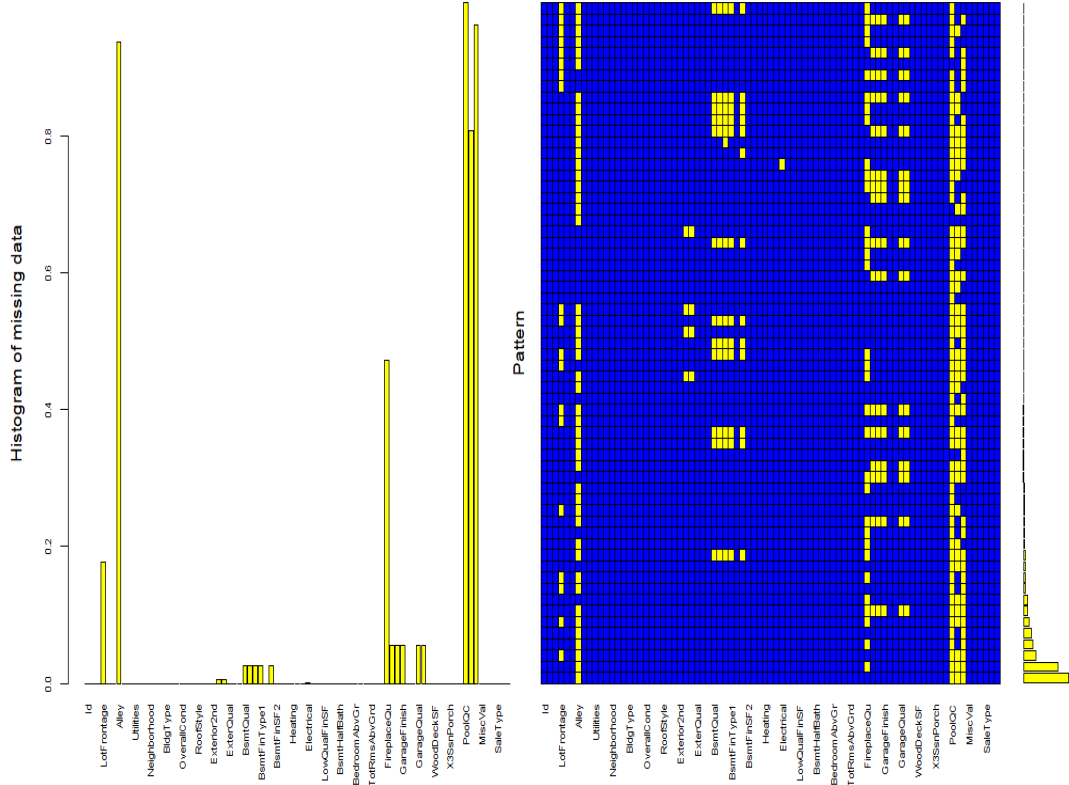


Variable	Missing
PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FireplaceQu	1179
LotFrontage	259
GargeType	81

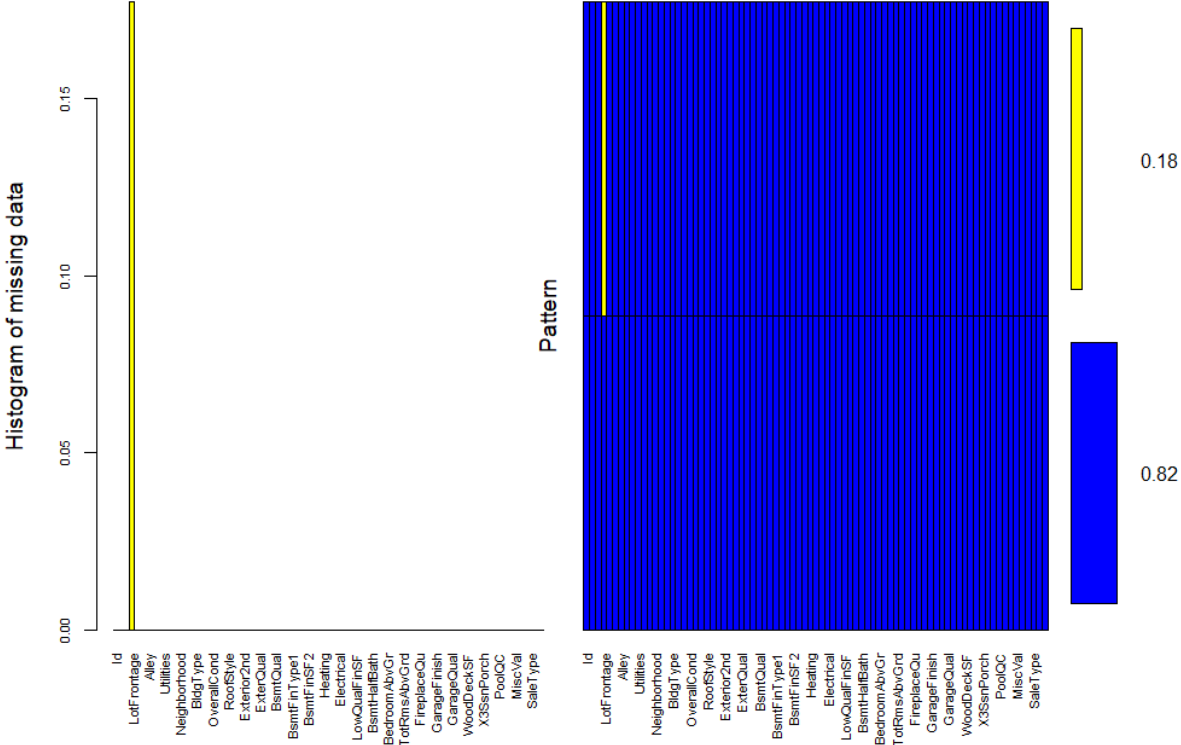
1. Categorical values to “None”
2. Quantitative values to 0

Missing Values

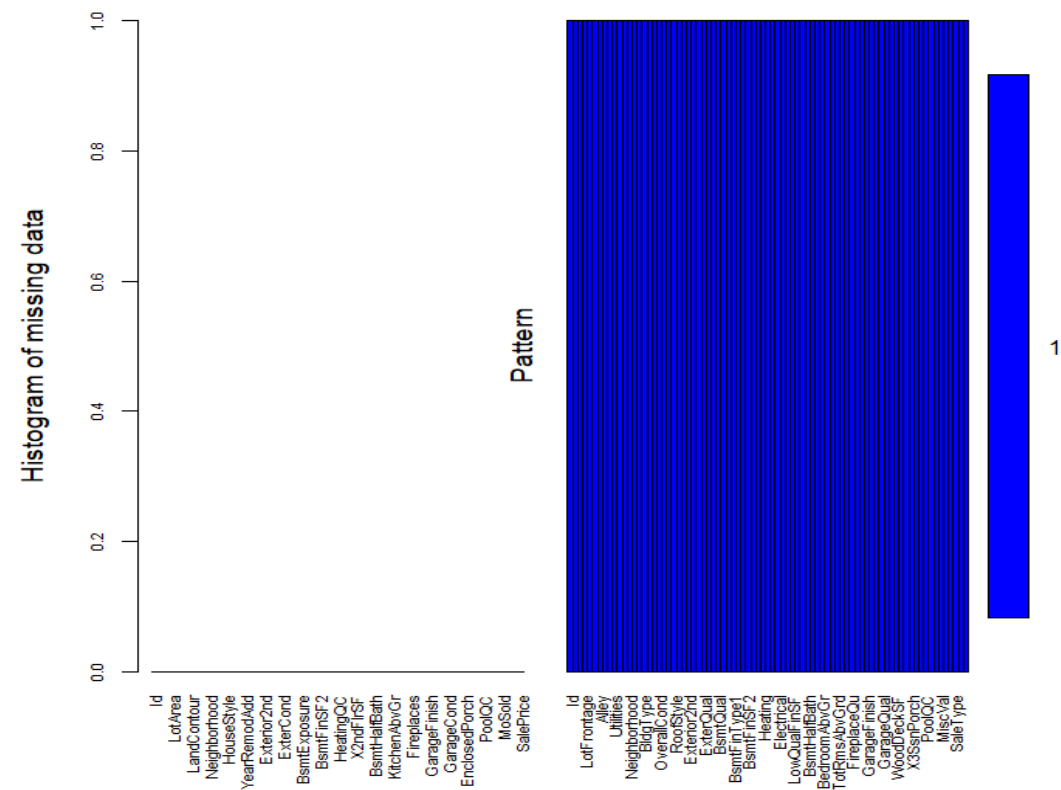
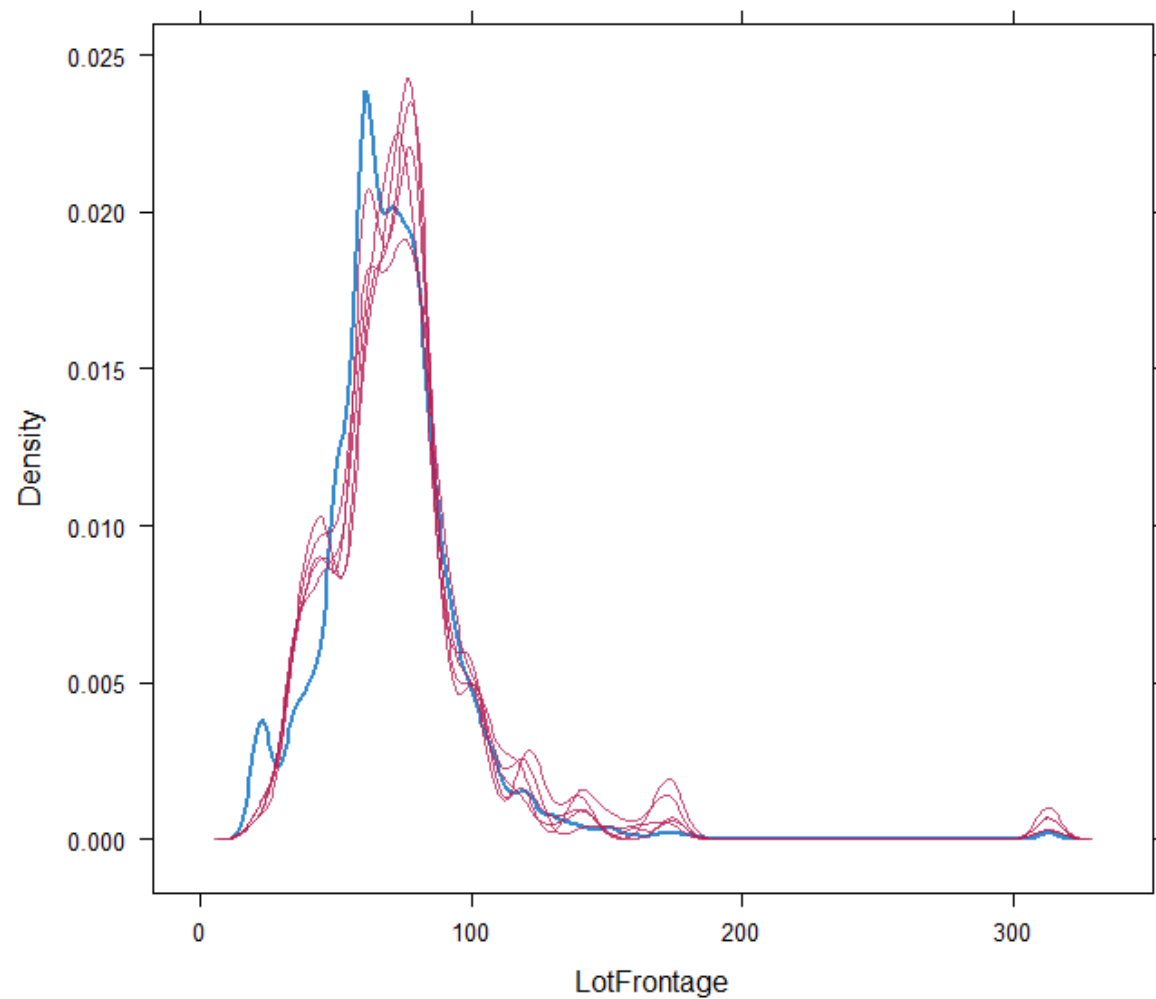
Before



After



Mice Imputations



Pre-Processing

- SalePrice transformation – LOG value
- Skewness
 - Scale all the variables with skewness above 0.75
- New features
 - IsRemodeled (0 or 1)
 - QtrSold = MoSold / 3
 - TotalSF = TotalBsmtSF + GrLivArea + GarageArea
- Sale Condition - Normal
- Data correction
 - GarageYrBlt in Test Data

Pre-Processing

- Anova
 - Street was found not to be significant and so dropped from the dataset
 - GarageType , GarageFinish and Garage Condition are more significant in predicting sale price
 - Interaction between GarageType and Garage Finish had a significant p-value
- Correlation plot
 - Sales price has strong correlation with Overall quality, TotalBsmtArea, x1stFlrSE, GrliviArea, Fullbath, TotalRoom, CarGarage, GarageArea
 - Sale price has a second level correlation with year built, year remodeled, MasVnrArea, GarageYrbuilt, Fireplace
 - SalesPrice negatively correlated with overallCondition, kitchenAbvGrd, enclosedPorch

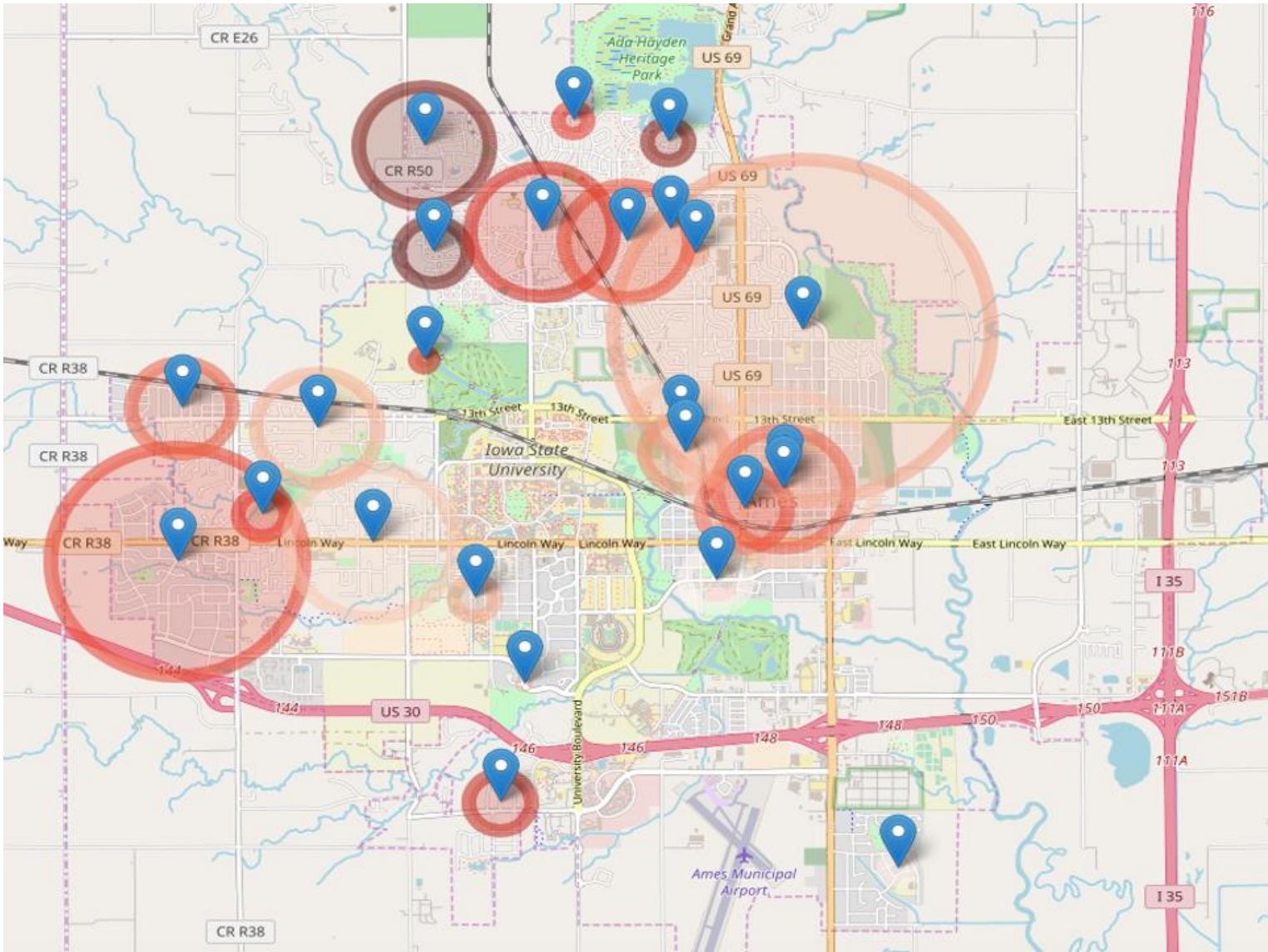


Features Engineering

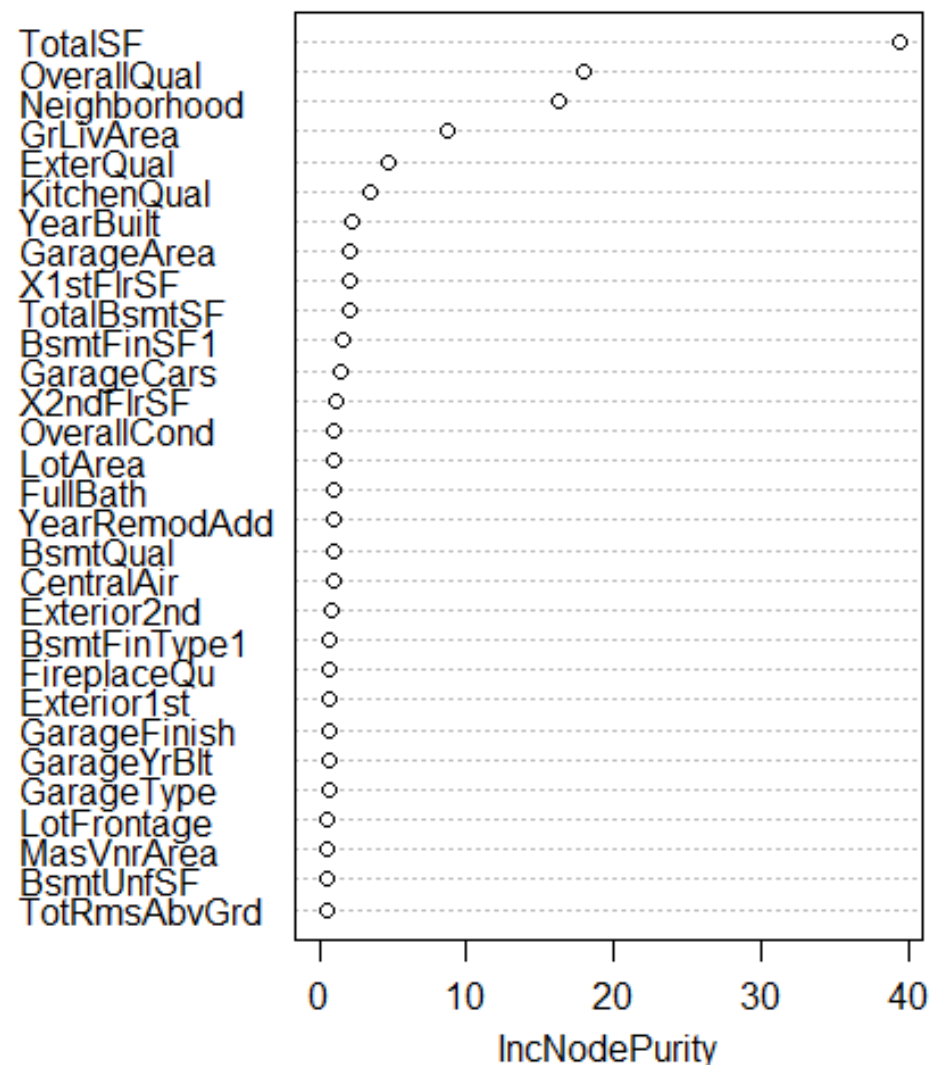
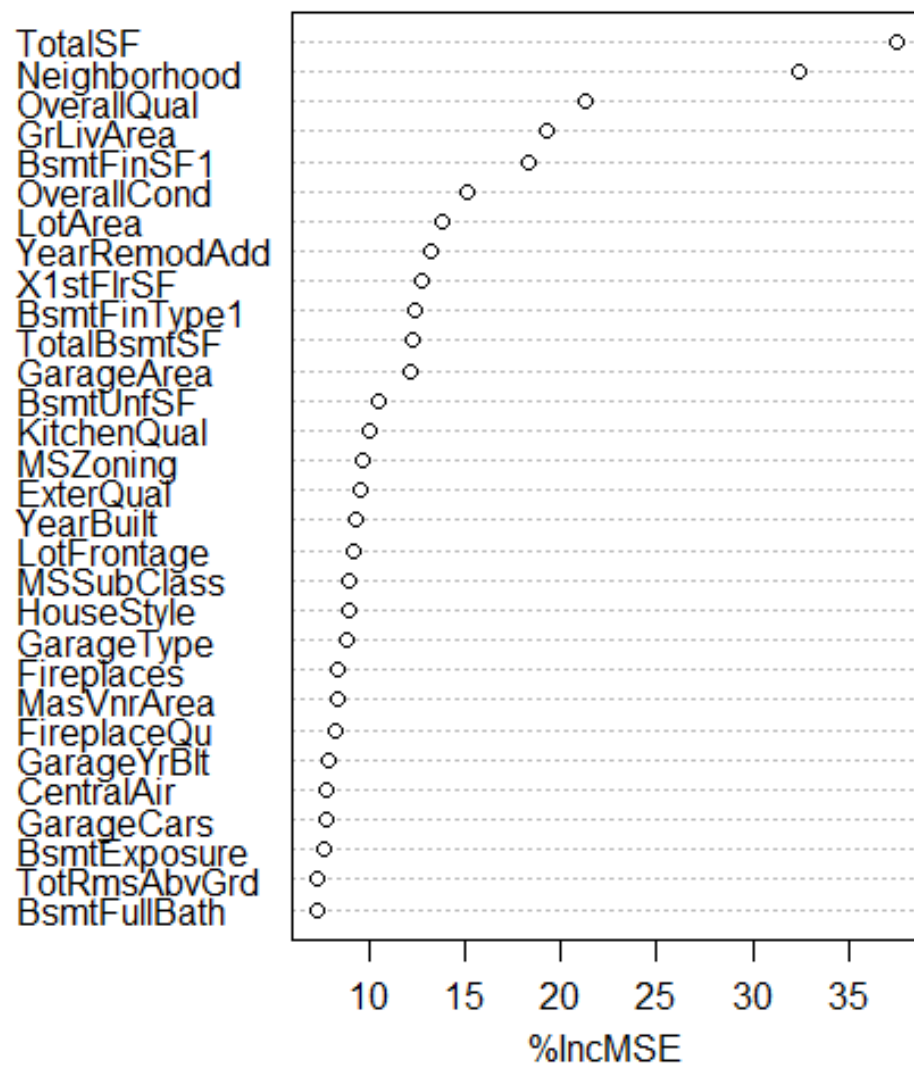
'If you torture the data enough, it will always confess'

- Ronald Coase, British Economist

Ames, IA Neighborhood

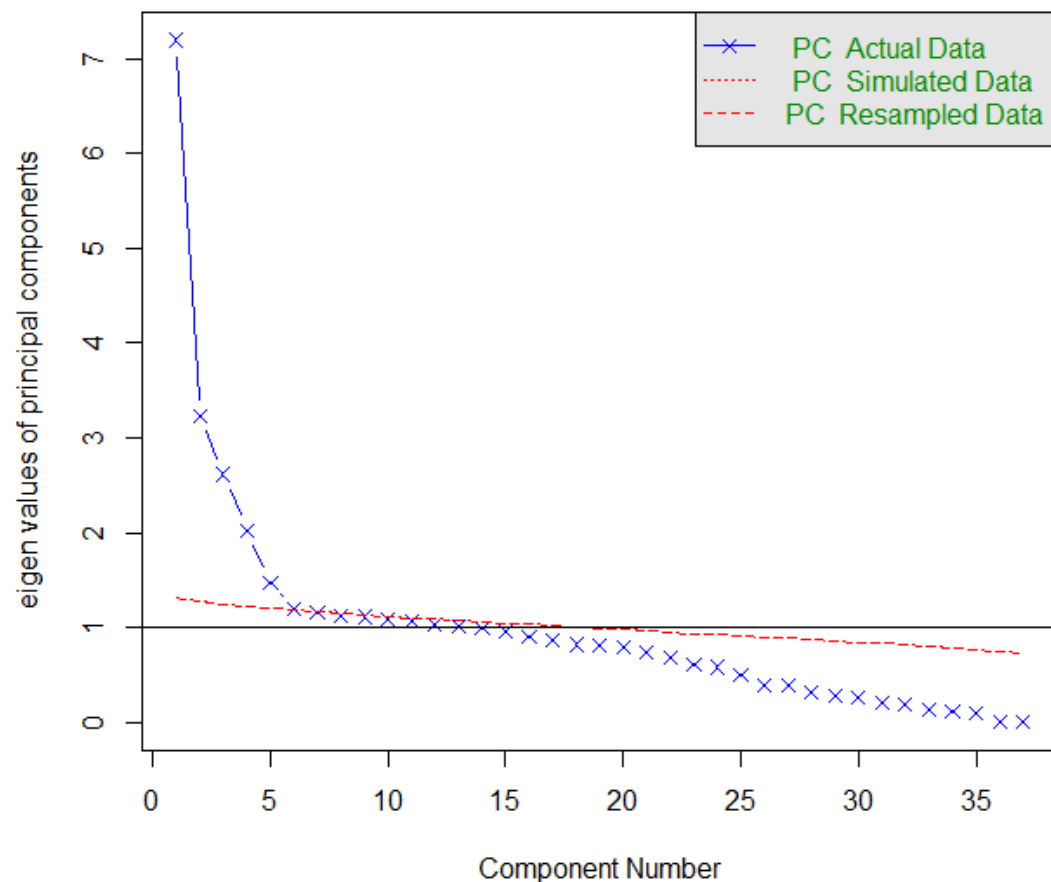


Random Forest For Best Variables



PCA – Component Analysis

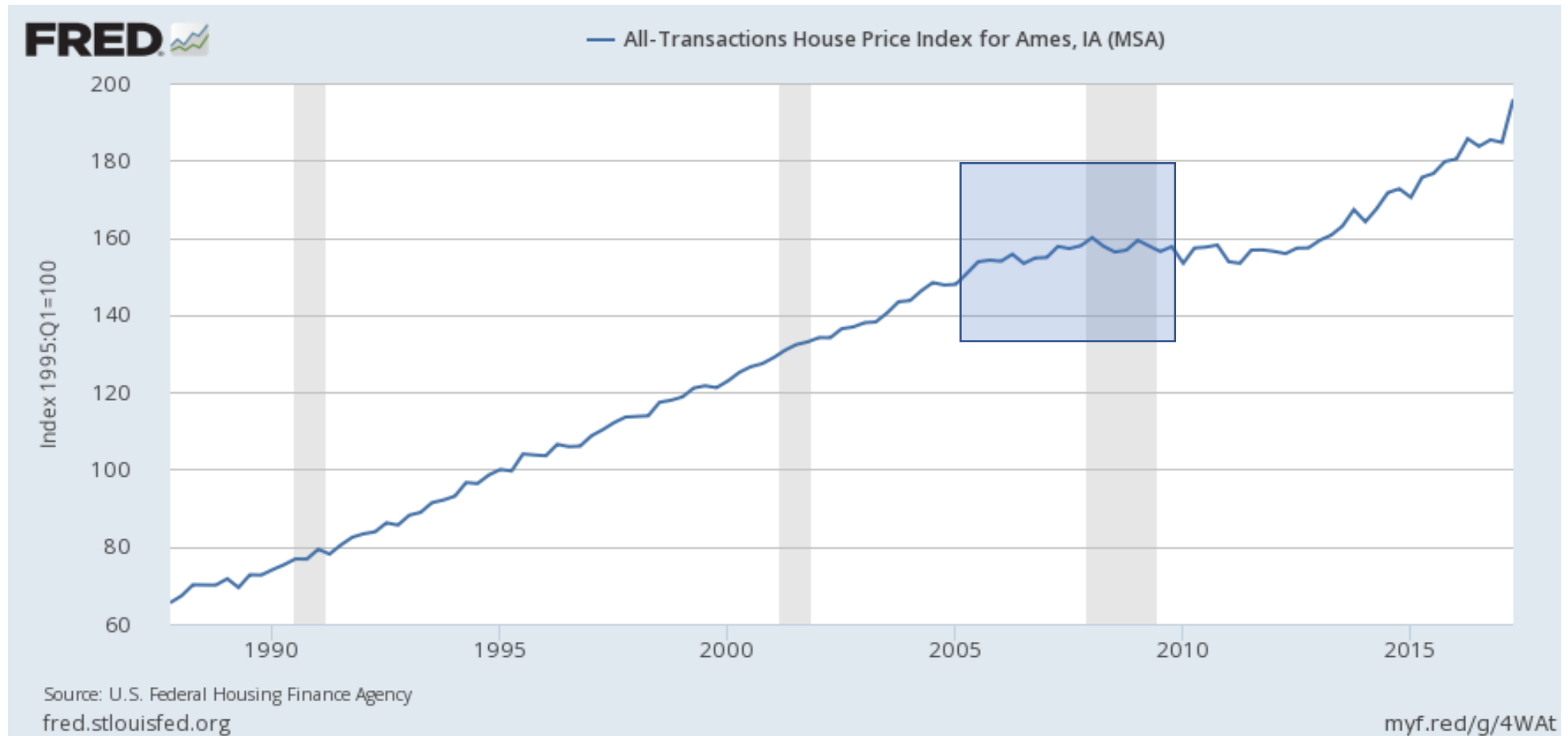
Parallel Analysis Scree Plots



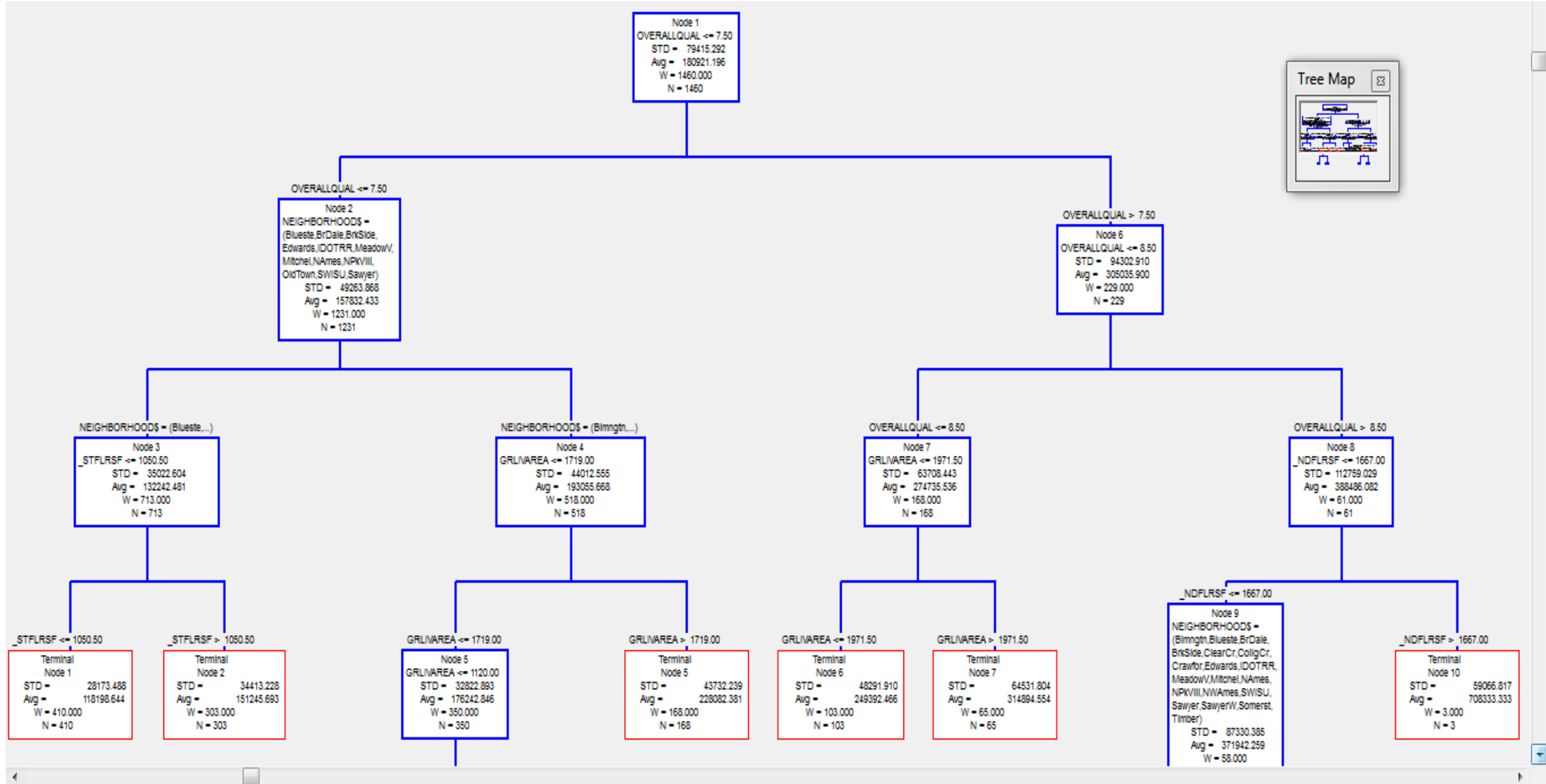
```
sp$finalModel$xNames  
[1] "PC1" "PC2" "PC3"
```

```
Principal Components Analysis  
Call: principal(r = houseTrain[, numericFeatures[1:37]], nfactors = 5,  
  rotate = "none")  
Standardized loadings (pattern matrix) based upon correlation matrix  
      PC1  PC2  PC3  PC4  PC5  h2  u2 com  
Id      -0.01  0.03  0.00  0.01 -0.06 0.0049 0.995 1.5  
MSSubClass -0.05  0.28 -0.38  0.42 -0.38 0.5410 0.459 3.8  
LotFrontage  0.42  0.03  0.52 -0.18  0.12 0.4945 0.505 2.3  
LotArea      0.29 -0.03  0.48  0.04  0.08 0.3272 0.673 1.7  
OverallQual  0.80 -0.03 -0.17 -0.01  0.14 0.7000 0.300 1.2  
OverallCond -0.22  0.11  0.20  0.13  0.48 0.3448 0.655 2.1  
YearBuilt    0.64 -0.38 -0.50  0.03 -0.03 0.8004 0.200 2.6  
YearRemodAdd 0.56 -0.17 -0.40  0.02  0.19 0.5352 0.465 2.3  
MasVnrArea   0.52 -0.03  0.00  0.07 -0.03 0.2765 0.723 1.1  
BsmtFinSF1   0.38 -0.50  0.37  0.49 -0.18 0.7969 0.203 4.0  
BsmtFinSF2  -0.03 -0.11  0.27  0.14  0.11 0.1192 0.881 2.3  
BsmtUnfSF    0.31  0.20 -0.15 -0.81  0.05 0.8168 0.183 1.5  
TotalBsmtSF  0.69 -0.35  0.33 -0.25 -0.10 0.7842 0.216 2.4  
X1stFlrSF    0.69 -0.24  0.45 -0.28 -0.18 0.8401 0.160 2.6  
X2ndFlrSF    0.35  0.76 -0.17  0.37  0.12 0.8852 0.115 2.1  
LowQualFinSF -0.04  0.23  0.15 -0.03 -0.02 0.0758 0.924 1.9  
GrLivArea    0.79  0.48  0.20  0.10 -0.03 0.9077 0.092 1.8  
BsmtFullBath 0.20 -0.50  0.27  0.52 -0.26 0.7007 0.299 3.4  
BsmtHalfBath -0.03  0.01  0.11  0.06  0.29 0.1003 0.900 1.4  
FullBath     0.70  0.27 -0.19 -0.10 -0.18 0.6359 0.364 1.7  
HalfBath     0.33  0.39 -0.26  0.44  0.25 0.5791 0.421 4.2  
BedroomAbvGr 0.30  0.66  0.21 -0.04 -0.08 0.5815 0.418 1.7  
KitchenAbvGr -0.05  0.32  0.08 -0.05 -0.70 0.6118 0.388 1.5  
TotRmsAbvGrd 0.63  0.61  0.18  0.01 -0.13 0.8227 0.177 2.3  
Fireplaces   0.49  0.04  0.30  0.15  0.21 0.4016 0.598 2.3  
GarageYrBlt  0.65 -0.32 -0.52 -0.01 -0.01 0.7964 0.204 2.4  
GarageCars   0.77 -0.13 -0.17 -0.04  0.01 0.6361 0.364 1.2  
GarageArea   0.75 -0.19 -0.07 -0.05  0.01 0.6096 0.390 1.1  
WoodDeckSF   0.36 -0.12  0.05  0.19  0.07 0.1864 0.814 1.9  
OpenPorchSF  0.39  0.08 -0.02  0.04  0.17 0.1943 0.806 1.5  
EnclosedPorch -0.20  0.22  0.29 -0.08 -0.05 0.1798 0.820 3.0  
X3SsnPorch   0.04 -0.05  0.01 -0.07  0.09 0.0178 0.982 3.1  
ScreenPorch  0.08  0.04  0.18  0.10  0.24 0.1079 0.892 2.5  
PoolArea     0.13  0.04  0.24  0.12  0.00 0.0917 0.908 2.2  
MiscVal      -0.03  0.05  0.08  0.03  0.02 0.0103 0.990 2.4  
MoSold       0.06  0.05 -0.01 -0.06  0.08 0.0159 0.984 3.8  
YrSold       -0.04 -0.06  0.01  0.07 -0.07 0.0144 0.986 3.6
```

Case-Shiller Ames Housing Index



Decision Tree for Variable Selection





Model Training and Evaluation

Multiple Linear Regression

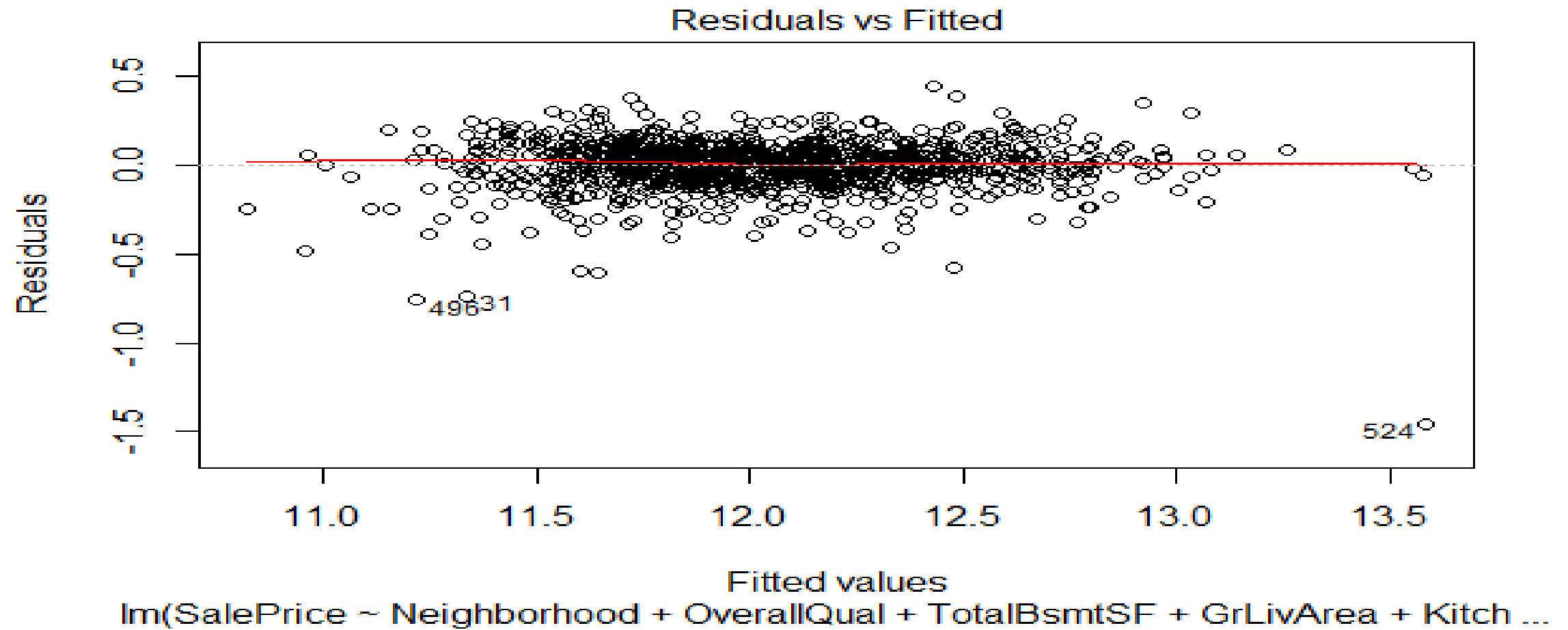
Summary Statistics

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.11e+01	6.14e-02	180.06	< 2e-16	***
NeighborhoodBlueste	-8.99e-02	9.54e-02	-0.94	0.34659	
NeighborhoodBrDale	-1.75e-01	6.30e-02	-2.77	0.00568	**
NeighborhoodBrkSide	-5.13e-02	4.66e-02	-1.10	0.27199	
NeighborhoodClearCr	8.50e-02	5.19e-02	1.64	0.10203	
NeighborhoodCollgCr	4.38e-03	4.26e-02	0.10	0.91818	
NeighborhoodCrawfor	1.11e-01	4.65e-02	2.39	0.01727	*
NeighborhoodEdwards	-8.14e-02	4.51e-02	-1.80	0.07153	.
NeighborhoodGilbert	6.99e-02	4.55e-02	1.54	0.12424	
NeighborhoodIDOTRR	-1.92e-01	5.12e-02	-3.76	0.00018	***
NeighborhoodMeadowV	-1.94e-01	5.75e-02	-3.37	0.00079	***
NeighborhoodMitchel	-4.10e-02	4.69e-02	-0.87	0.38250	
NeighborhoodNames	-3.85e-02	4.34e-02	-0.89	0.37490	
NeighborhoodNoRidge	5.02e-02	4.80e-02	1.05	0.29552	
NeighborhoodNPkVill	-9.07e-02	6.23e-02	-1.46	0.14587	
NeighborhoodNridgHt	2.13e-02	4.68e-02	0.46	0.64841	
NeighborhoodNWAmes	-2.48e-03	4.50e-02	-0.06	0.95600	
NeighborhoodOldTown	-1.56e-01	4.49e-02	-3.47	0.00054	***
NeighborhoodSawyer	-3.44e-02	4.55e-02	-0.76	0.45004	
NeighborhoodSawyerW	-3.58e-02	4.56e-02	-0.79	0.43262	
NeighborhoodSomerst	3.48e-02	4.54e-02	0.77	0.44347	
NeighborhoodStoneBr	1.26e-02	5.52e-02	0.23	0.81973	
NeighborhoodSWISU	-1.11e-01	5.06e-02	-2.20	0.02823	*
NeighborhoodTimber	-2.55e-02	5.21e-02	-0.49	0.62547	
NeighborhoodVeenker	1.27e-01	5.80e-02	2.20	0.02838	*
TotalBsmtSF	1.08e-04	2.33e-05	4.66	3.7e-06	***
GrLivArea	2.41e-04	1.21e-05	19.97	< 2e-16	***
KitchenQualFa	-2.37e-01	3.43e-02	-6.92	8.8e-12	***
KitchenQualGd	-7.35e-02	2.16e-02	-3.41	0.00068	***
KitchenQualTA	-1.45e-01	2.36e-02	-6.13	1.3e-09	***
GarageArea	2.37e-04	2.63e-05	9.01	< 2e-16	***
BsmtFinSF1	6.66e-05	1.63e-05	4.08	4.8e-05	***
BsmtFinType1BLQ	-3.35e-02	1.56e-02	-2.14	0.03241	*
BsmtFinType1GLQ	-1.79e-02	1.41e-02	-1.27	0.20607	
BsmtFinType1LwQ	-5.24e-02	2.03e-02	-2.58	0.00997	**
BsmtFinType1NB	-9.68e-02	3.54e-02	-2.73	0.00637	**
BsmtFinType1Rec	-5.75e-02	1.63e-02	-3.52	0.00046	***
BsmtFinType1Unf	-7.64e-02	1.67e-02	-4.57	5.4e-06	***
OverallQual	8.41e-02	5.16e-03	16.30	< 2e-16	***
LotArea	2.31e-06	3.95e-07	5.87	6.2e-09	***
X1stFlrSF	-1.25e-05	2.48e-05	-0.51	0.61367	

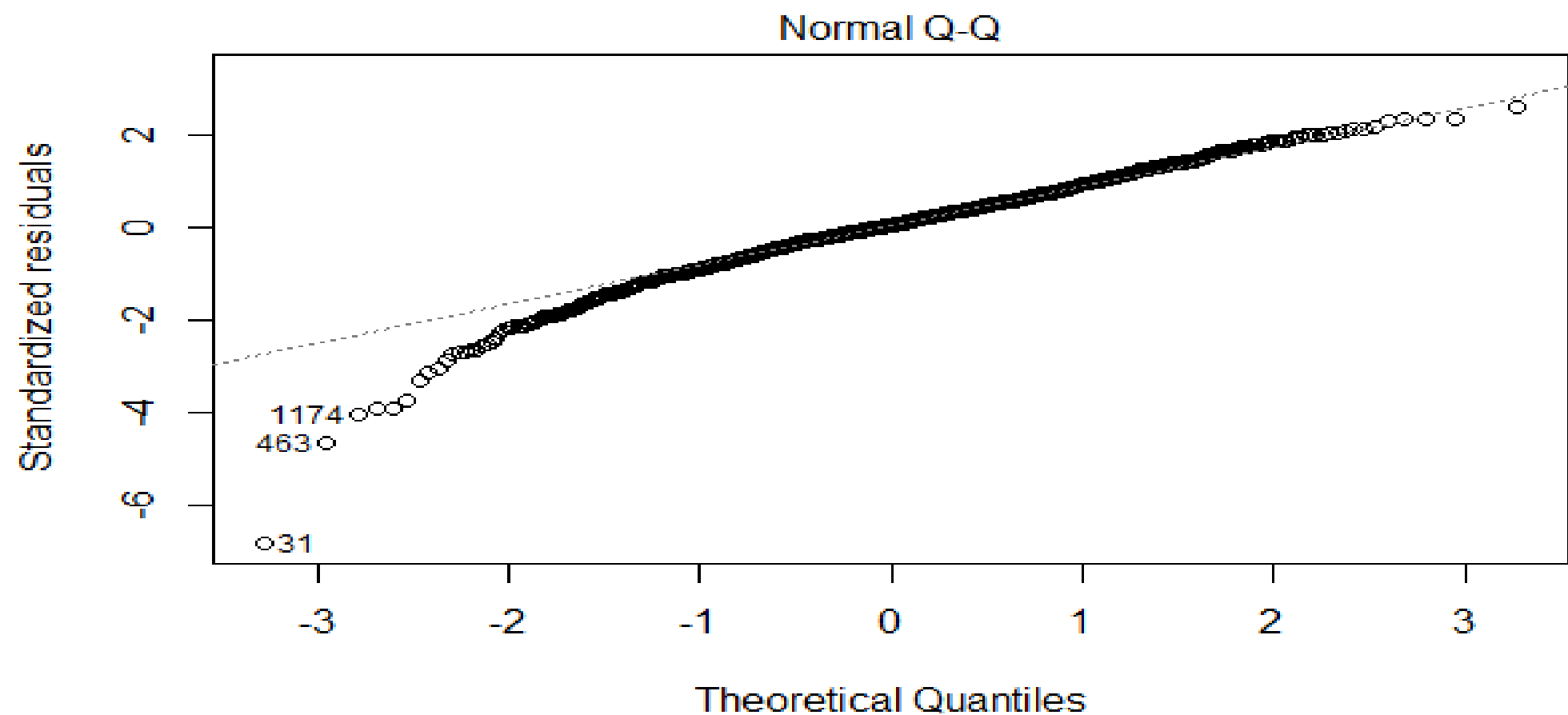
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 0.119 on 916 degrees of freedom					
Multiple R-squared: 0.896, Adjusted R-squared: 0.892					
F-statistic: 198 on 40 and 916 DF, p-value: <2e-16					

Assumptions

Constant Variance

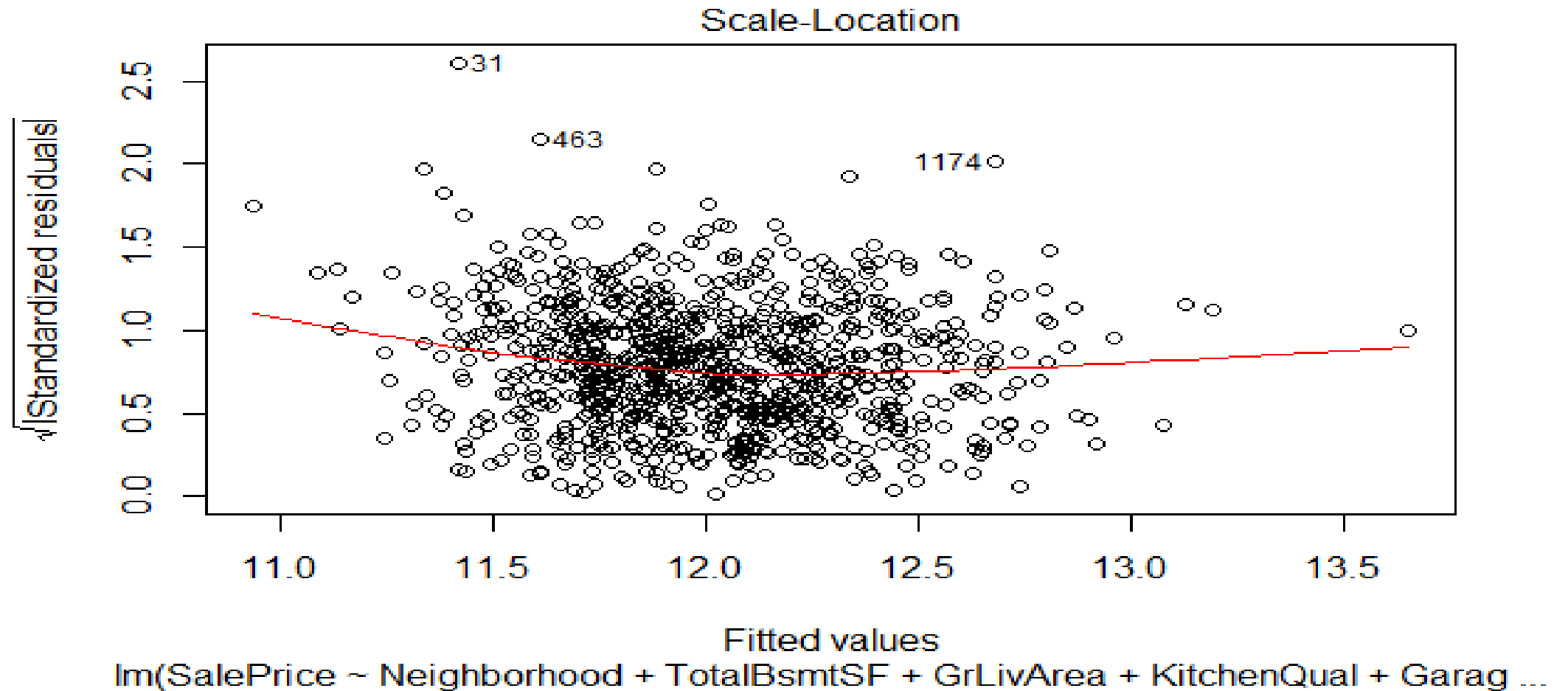


Normality

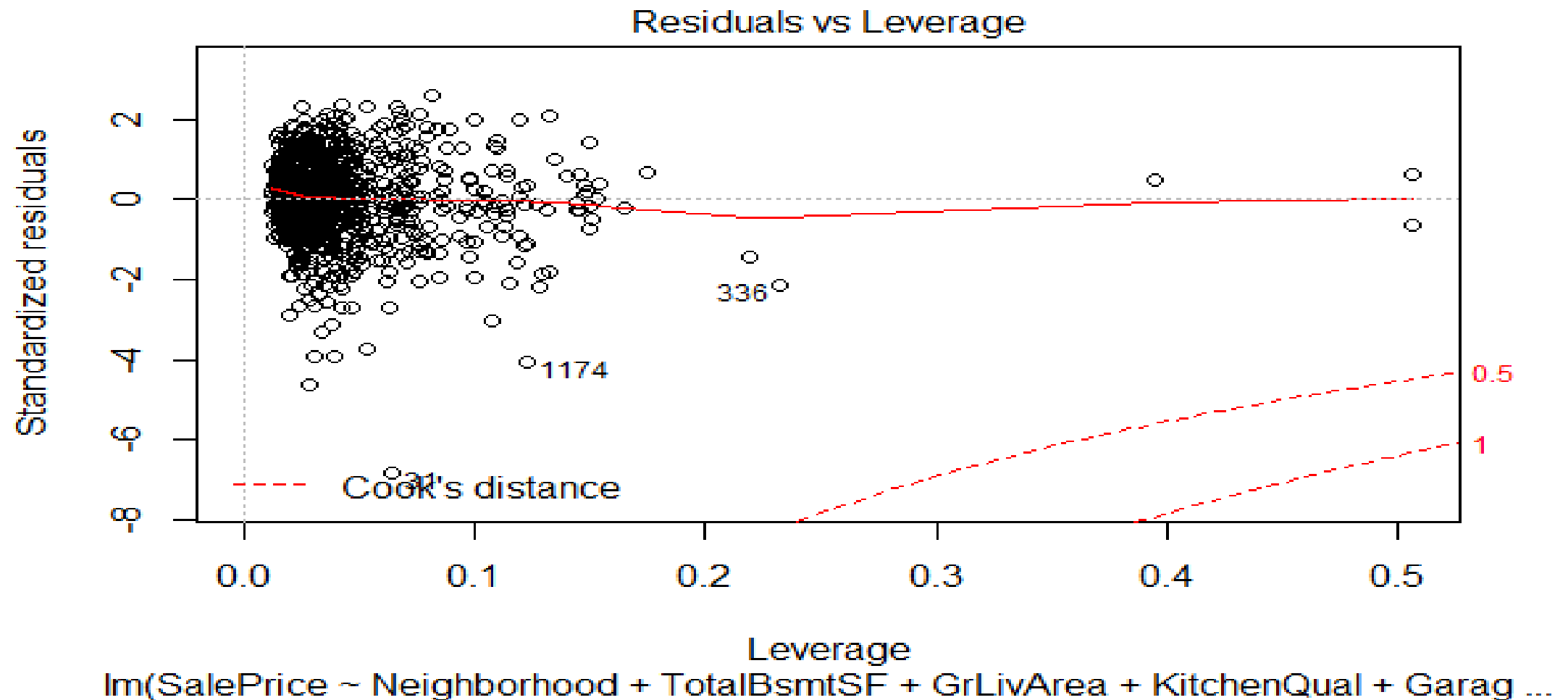


$\text{lm}(\text{SalePrice} \sim \text{Neighborhood} + \text{TotalBsmtSF} + \text{GrLivArea} + \text{KitchenQual} + \text{Garag} \dots)$

Independent Error

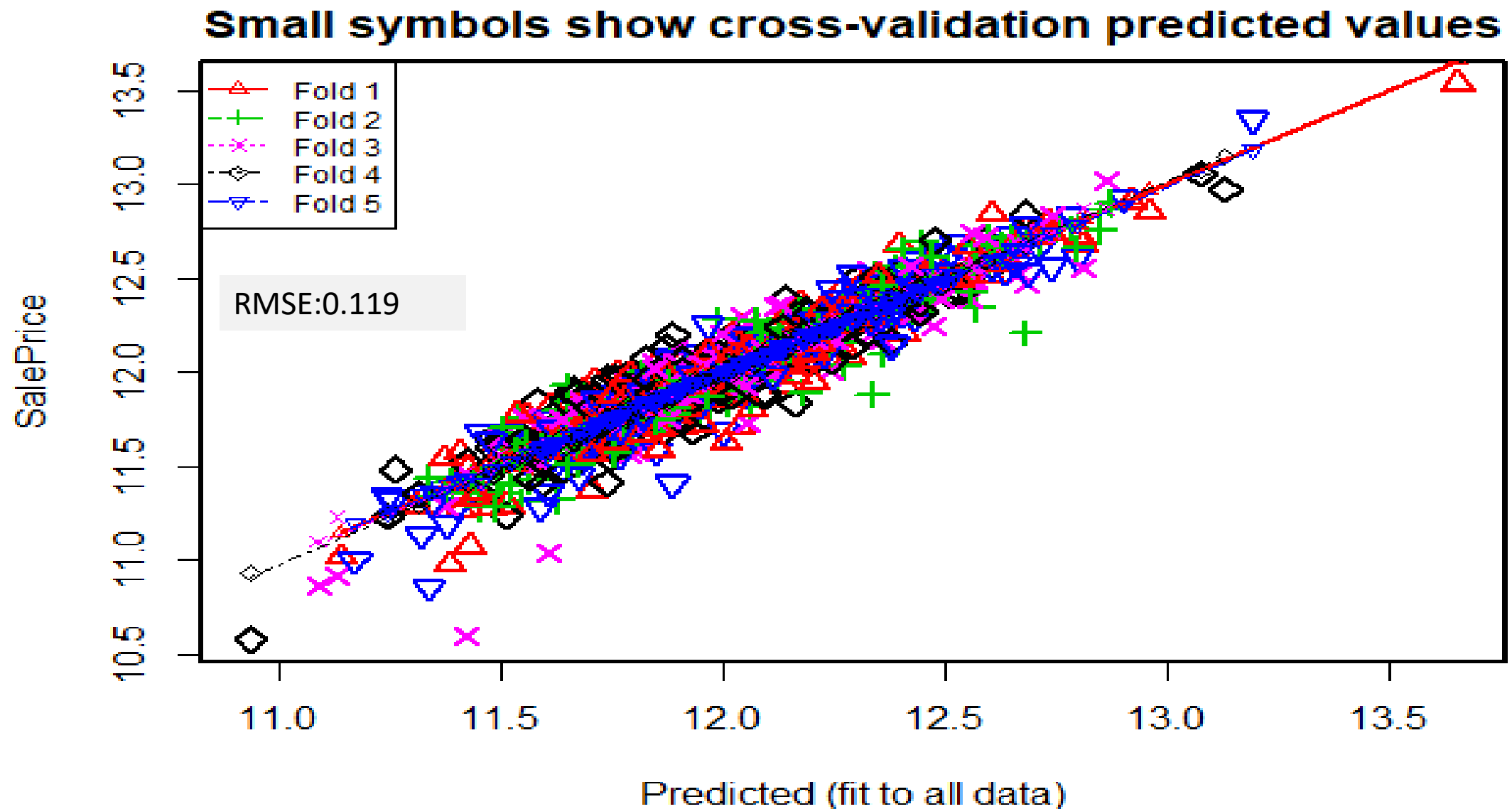


Residuals and Leverage



Cross Validation

Predicted vs Actuals





Analysis, Conclusions and Considerations

Extracting Additional Value from the Models

- Key Stakeholders: individuals (home seekers, investors), real estate developers, asset managers, realtors, contractors, creditors/banks
- Discovering investment opportunities by comparing fixed features vs changeable features
- Considering macro/micro economic factors
- Quantifying the opportunity and deriving actionable insights and conclusions

Statistically supported conclusions

1. Go Big or Go Home (pun intended)

- Total Square Footage(using basement + living + garage space) average (mean) of homes in Ames cost 68 sq/ft.
- 73% of sales price is implied by attributes related to size (R-sq)
- Sorry Guys, Size Matters!

2. Be Selective but not obsessive

- 6 variables including Size, Age (year built and year remodeled) , Location and Fireplaces account for 81% for price.
- 12 Variables that include (Lot Area, Year Built, YearRemodAdd, TotalBsmtSf, BsmtFinSFOne, GrLivArea, BedroomAbvGr, KitchenAbvGr, Fireplaces, Garage Area, Good Neighborhood, Troubled Neighborhood) tells 88% of the story (based on R-sq).

Flip or Flop – Investment opportunities by renovating?

- If 88% of price is accounted for by fixed features, only 12% (or \$19k of house value) remains to other features which limits upside potential to add value via renovation and repairs using “Quality” (ie kitchen quality) and “Condition” (ie garage condition) metrics, including roof, exterior features, heating, electric components, etc.
- If the goal is to add value via renovation, the most promising upside will come from:
 1. Adding a garage if land and zoning eligibility permits (worth approximately \$6k per car)
 2. Complete and unfinished basement (amounts to \$35/sq. ft.)
 3. Remodel/Upgrade the kitchen.
- Profit model relies on ability to source low cost contractors.
- Analysis suggests that the most prudent investment strategy would be investing in underpriced listings based on the criteria in the linear model vs “flipping” houses by renovating features and making improvements.



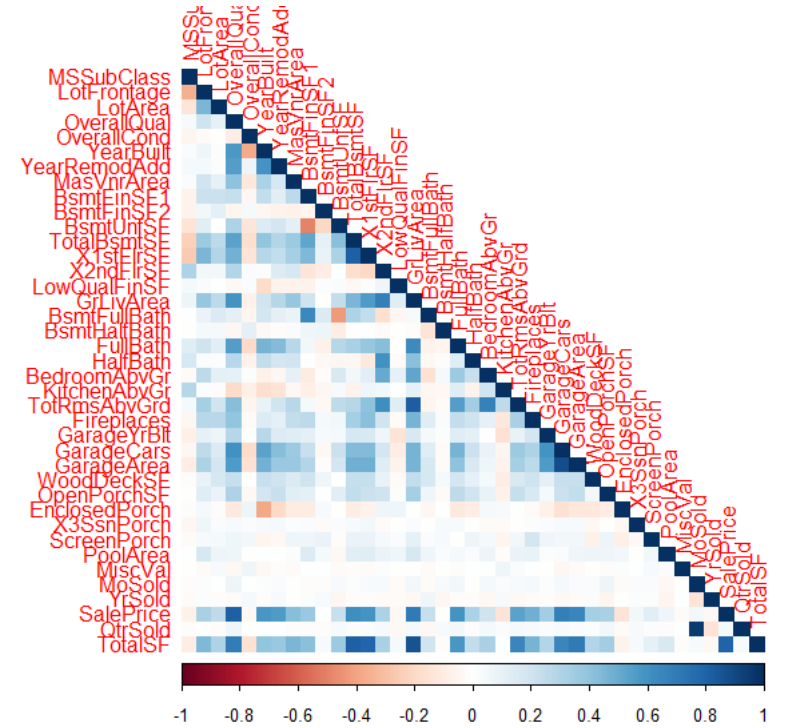
Thank You!

Appendix-1

File Edit Format View Help

	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
Id	1	0.07	0.075	6.913	0.008665	**
MSSubClass	1	1.27	1.267	116.842	< 2e-16	***
MSZoning	4	39.87	9.967	919.343	< 2e-16	***
LotFrontage	1	14.94	14.936	1377.695	< 2e-16	***
LotArea	1	2.71	2.714	250.351	< 2e-16	***
Street	1	0.55	0.555	51.152	1.48e-12	***
Alley	2	1.12	0.562	51.872	< 2e-16	***
LotShape	3	6.75	2.250	207.557	< 2e-16	***
LandContour	3	3.08	1.027	94.771	< 2e-16	***
Utilities	1	0.17	0.166	15.266	9.86e-05	***
LotConfig	4	1.04	0.260	23.945	< 2e-16	***
LandSlope	2	1.75	0.876	80.769	< 2e-16	***
Neighborhood	24	75.42	3.142	289.866	< 2e-16	***
Condition1	8	1.66	0.207	19.099	< 2e-16	***
Condition2	7	1.21	0.173	15.912	< 2e-16	***
BldgType	4	3.97	0.993	91.565	< 2e-16	***
HouseStyle	7	3.28	0.469	43.229	< 2e-16	***
OverallQual	1	28.30	28.296	2610.116	< 2e-16	***
OverallCond	1	1.29	1.285	118.538	< 2e-16	***
YearBuilt	1	2.62	2.624	242.043	< 2e-16	***
YearRemodAdd	1	0.77	0.775	71.458	< 2e-16	***
RoofStyle	5	0.54	0.109	10.052	1.93e-09	***
RoofMatl	7	1.44	0.206	18.985	< 2e-16	***
Exterior1st	14	2.03	0.145	13.343	< 2e-16	***
Exterior2nd	14	0.52	0.037	3.424	1.75e-05	***
MasVnrType	3	0.79	0.265	24.408	2.47e-15	***
MasVnrArea	1	0.49	0.487	44.914	3.16e-11	***
ExterQual	3	0.32	0.106	9.802	2.17e-06	***
ExterCond	4	0.31	0.078	7.158	1.07e-05	***
Foundation	5	0.88	0.175	16.143	1.98e-15	***
BsmtQual	4	0.90	0.225	20.775	< 2e-16	***
BsmtCond	3	0.05	0.017	1.524	0.206600	
BsmtExposure	4	0.92	0.231	21.268	< 2e-16	***
BsmtFinType1	5	1.35	0.270	24.889	< 2e-16	***
BsmtFinSF1	1	1.60	1.601	147.692	< 2e-16	***
BsmtFinType2	6	0.27	0.045	4.116	0.000425	***
BsmtFinSF2	1	0.16	0.156	14.381	0.000157	***
BsmtUnfSF	1	4.49	4.487	413.862	< 2e-16	***
Heating	5	0.60	0.119	10.996	2.27e-10	***
HeatingQC	4	0.27	0.068	6.271	5.40e-05	***
CentralAir	1	0.43	0.433	39.919	3.72e-10	***
Electrical	5	0.21	0.042	3.837	0.001877	**
X1stFlrSF	1	2.71	2.708	249.784	< 2e-16	***
X2ndFlrSF	1	2.24	2.240	206.651	< 2e-16	***
LowQualFinSF	1	0.05	0.049	4.550	0.033113	*
BsmtFullBath	1	0.13	0.127	11.750	0.000629	***
BsmtHalfBath	1	0.00	0.001	0.075	0.784360	
FullBath	1	0.00	0.000	0.042	0.836794	
HalfBath	1	0.13	0.132	12.173	0.000502	***
BedroomAbvGr	1	0.00	0.002	0.198	0.656181	
KitchenAbvGr	1	0.04	0.041	3.805	0.051335	.
KitchenQual	3	0.18	0.061	5.644	0.000769	***
TotRmsAbvGrd	1	0.03	0.030	2.786	0.095348	.
Functional	6	0.68	0.114	10.511	2.16e-11	***
Fireplaces	1	0.24	0.235	21.711	3.52e-06	***
FireplaceQu	5	0.09	0.017	1.595	0.158518	*
GarageType	6	0.70	0.117	10.824	9.33e-12	***
GarageYrBlt	1	0.10	0.105	9.656	0.001932	**
GarageFinish	2	0.03	0.013	1.174	0.309403	
GarageCars	1	0.36	0.357	32.945	1.20e-08	***
GarageArea	1	0.20	0.196	18.114	2.24e-05	***
GarageQual	4	0.14	0.035	3.211	0.012364	*
GarageCond	4	0.14	0.035	3.191	0.012795	*
PavedDrive	2	0.01	0.006	0.579	0.560401	
WoodDeckSF	1	0.06	0.059	5.481	0.019387	*
OpenPorchSF	1	0.00	0.003	0.260	0.610548	
EnclosedPorch	1	0.03	0.030	2.776	0.095958	.
X3SsnPorch	1	0.02	0.016	1.481	0.223822	
ScreenPorch	1	0.19	0.187	17.217	3.57e-05	***
PoolArea	1	0.03	0.034	3.112	0.077953	*
PoolQC	3	0.10	0.033	3.064	0.027234	*
Fence	4	0.10	0.025	2.318	0.055311	.
MiscFeature	4	0.01	0.002	0.200	0.938311	
MiscVal	1	0.00	0.000	0.044	0.833119	
MoSold	1	0.00	0.000	0.022	0.881130	
YrSold	1	0.02	0.017	1.568	0.210697	
SaleType	8	0.32	0.040	3.675	0.000302	***
SaleCondition	5	0.27	0.055	5.050	0.000138	***
IsRemodeled	1	0.01	0.006	0.566	0.451824	
QtrSold	1	0.00	0.001	0.063	0.801865	
Residuals	1204	13.05	0.011			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
GarageType	6	77.90	12.984	144.375	< 2e-16	***
GarageFinish	2	20.51	10.254	114.023	< 2e-16	***
GarageQual	4	1.32	0.330	3.669	0.00558	**
GarageCond	4	2.24	0.561	6.239	5.63e-05	***
GarageType:GarageFinish	6	1.81	0.226	2.513	0.01032	*
GarageType:GarageQual	6	0.50	0.084	0.934	0.46968	
GarageFinish:GarageQual	4	0.49	0.123	1.363	0.24461	
GarageType:GarageCond	3	0.25	0.085	0.941	0.42006	
GarageFinish:GarageCond	1	0.01	0.006	0.066	0.79777	
GarageQual:GarageCond	2	0.22	0.109	1.215	0.29712	
GarageType:GarageQual:GarageCond	1	0.02	0.018	0.200	0.65508	
Residuals	1418	127.52	0.090			

Appendix-2

MSZoning Street Alley LotShape LandContour Utilities
9.909275e-18 1.594118e-03 1.420779e-02 4.495140e-01 3.127842e-08 1.939245e-02
LotConfig Landslope Neighborhood Condition1 Condition2 BldgType
1.412358e-01 1.668988e-01 5.576778e-43 7.515057e-01 7.173339e-01 3.332217e-23
HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
1.626511e-06 6.468740e-08 8.029727e-03 2.028676e-28 5.394584e-24 9.246429e-31
ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
3.071252e-46 1.813475e-02 2.161004e-30 9.358370e-55 1.223155e-07 2.937449e-11
BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
1.909141e-15 5.928728e-04 9.990498e-01 8.132604e-22 2.620041e-04 3.363524e-23
KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
1.445136e-36 2.120077e-01 3.707982e-16 8.370676e-22 1.116818e-27 3.817652e-07
GarageCond PavedDrive PoolQC Fence MiscFeature SaleType
1.293424e-07 5.899112e-06 5.929498e-09 2.227592e-04 9.927046e-01 2.998978e-321 IsRemodeled 9.788271e-01