House Prices Prediction With Elastic Net and Neural Network

Group 12
Sharifa Rahmani
Zifan Huang
Mengshi Ge
Laisi Ma

Technique
R Programming

Background

“Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.”

Data Source

The Ames City Assessor's Office

The initial Excel file contained 113 variables describing 3970 property sales that had occurred in Ames, Iowa between 2006 and 2010

Simplified by Dean De Cock

79 variables remained that were directly related to property
sales.
The 79 variables focus on the quality and quantity of many physical attributes of the property.

Most of the variables are the type of information that a typical home buyer would want to know about a potential property

Introduction

- Two Datasets → Train & Test

- Difference: No “SalesPrice” in Test dataset.

- Dependent variable: SalesPrice（Y）
- Independent variables: 79 + 1 (Feature)（X1，X2，X3…）

- The Number of Observations in our dataset: 1459 (test) +1460 (train) =2919

Description of Variables

SalePrice: the property's sale price in dollars. This is the target variable that we're trying to predict

LotFrontage: Linear feet of street connected to property

LotShape: General shape of property

Utilities: Type of utilities available

YearBuilt: Original construction date

RoofStyle: Type of roof

ExterQual: Exterior material quality

Foundation: Type of foundation

BsmtFinSF1: Type 1 finished square feet

HeatingQC: Heating quality and condition

KitchenQual: Kitchen quality

SaleCondition: Condition of sale

Three steps for the project

1. Data Cleaning
2. Feature Engineering
3. Model Training

Data Cleaning

Missing Values

Totally 2919 rows from train and test datasets.

For variables whose count of missing value is less than 100 missing values, just ignore them.

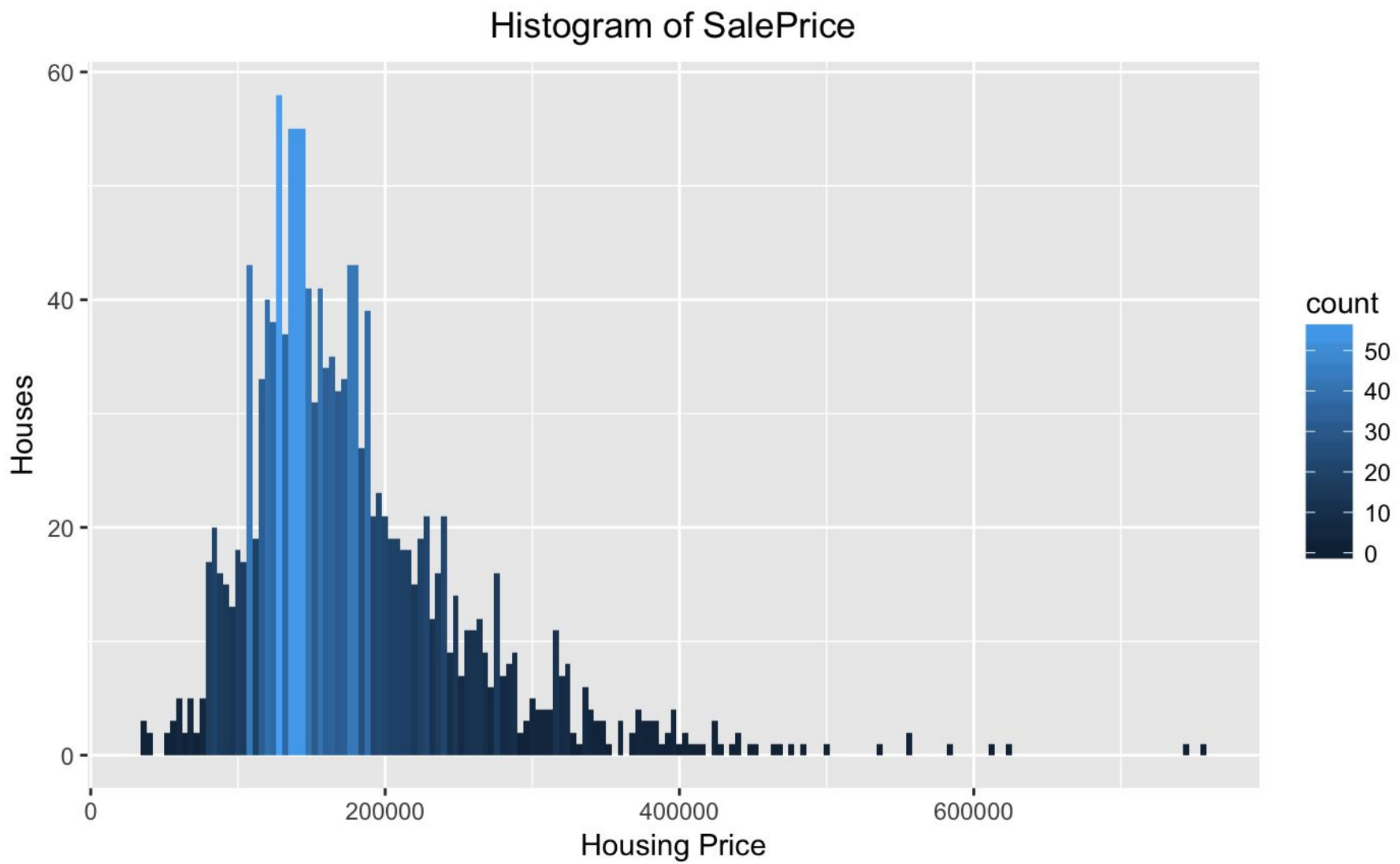
For categorical variables that have larger than 100 missing values, insert ’missing’, so missing value as a new level.

For numeric variables that have larger than 100 missing values(only LotFrontage), we fill them with 0 and add a new indicator feature, if LotFrontage is missing, the feature=1, if LotFrontage is not missing, the feature=0.

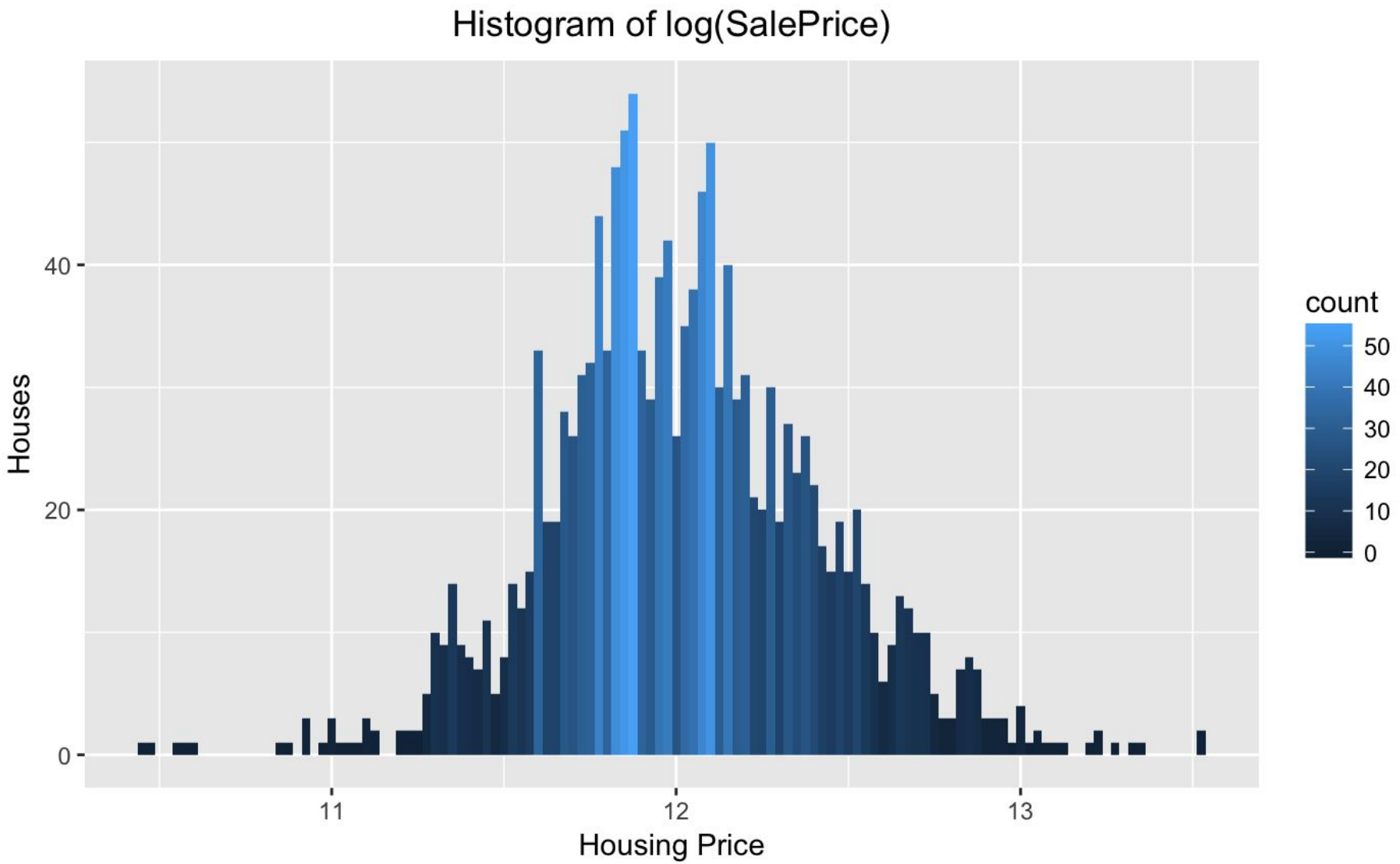
```
##feature engineering
numofRow <- length(train[,1])
trainNoid$Alley <- as.factor(sapply(1:numofRow, function(i) ifelse(is.na(trainNoid$Alley[i]),"missing", trainNoid$Alley[i])))
trainNoid$FireplaceQu <- as.factor(sapply(1:numofRow, function(i) ifelse(is.na(trainNoid$FireplaceQu[i]),"missing", trainNoid$FireplaceQu[i])))
trainNoid$PoolQC <- as.factor(sapply(1:numofRow, function(i) ifelse(is.na(trainNoid$PoolQC[i]),"missing", trainNoid$PoolQC[i])))
trainNoid$MiscFeature <- as.factor(sapply(1:numofRow, function(i) ifelse(is.na(trainNoid$MiscFeature[i]),"missing", trainNoid$MiscFeature[i])))
trainNoid$Fence <- as.factor(sapply(1:numofRow, function(i) ifelse(is.na(trainNoid$Fence[i]),"missing", trainNoid$Fence[i])))
```

Feature?

PoolQC	MiscFeature	Alley	Fence	SalePrice
2909	2814	2721	2348	1459
FireplaceQu	LotFrontage	GarageYrBlt	GarageFinish	GarageQual
1420	486	159	159	159
GarageCond	GarageType	BsmtCond	BsmtExposure	BsmtQual
159	157	82	82	81
BsmtFinType2	BsmtFinType1	MasVnrType	MasVnrArea	MSZoning
80	79	24	23	4
Utilities	BsmtFullBath	BsmtHalfBath	Functional	Exterior1st
2	2	2	2	1
Exterior2nd	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF
1	1	1	1	1
Electrical	KitchenQual	GarageCars	GarageArea	SaleType
1	1	1	1	1
Id	MSSubClass	LotArea	Street	LotShape
0	0	0	0	0
LandContour	LotConfig	LandSlope	Neighborhood	Condition1
0	0	0	0	0
Condition2	BldgType	HouseStyle	OverallQual	OverallCond
0	0	0	0	0
YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	ExterQual
0	0	0	0	0
ExterCond	Foundation	Heating	HeatingQC	CentralAir
0	0	0	0	0
X1stFlrSF	X2ndFlrSF	LowQualFinSF	GrLivArea	FullBath
0	0	0	0	0
HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd	Fireplaces
0	0	0	0	0
PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch
0	0	0	0	0
ScreenPorch	PoolArea	MiscVal	MoSold	YrSold
0	0	0	0	0
SaleCondition				
0				

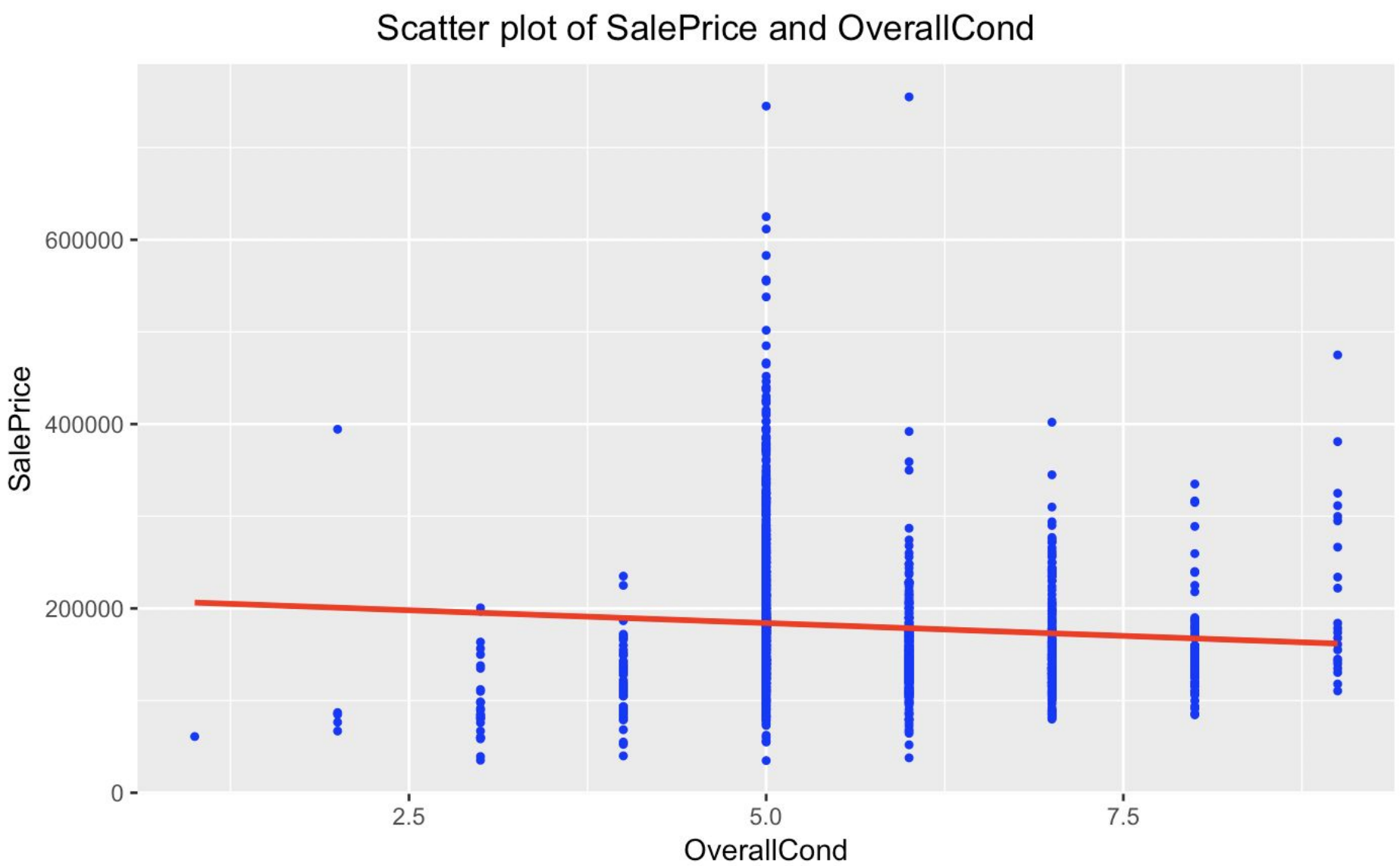


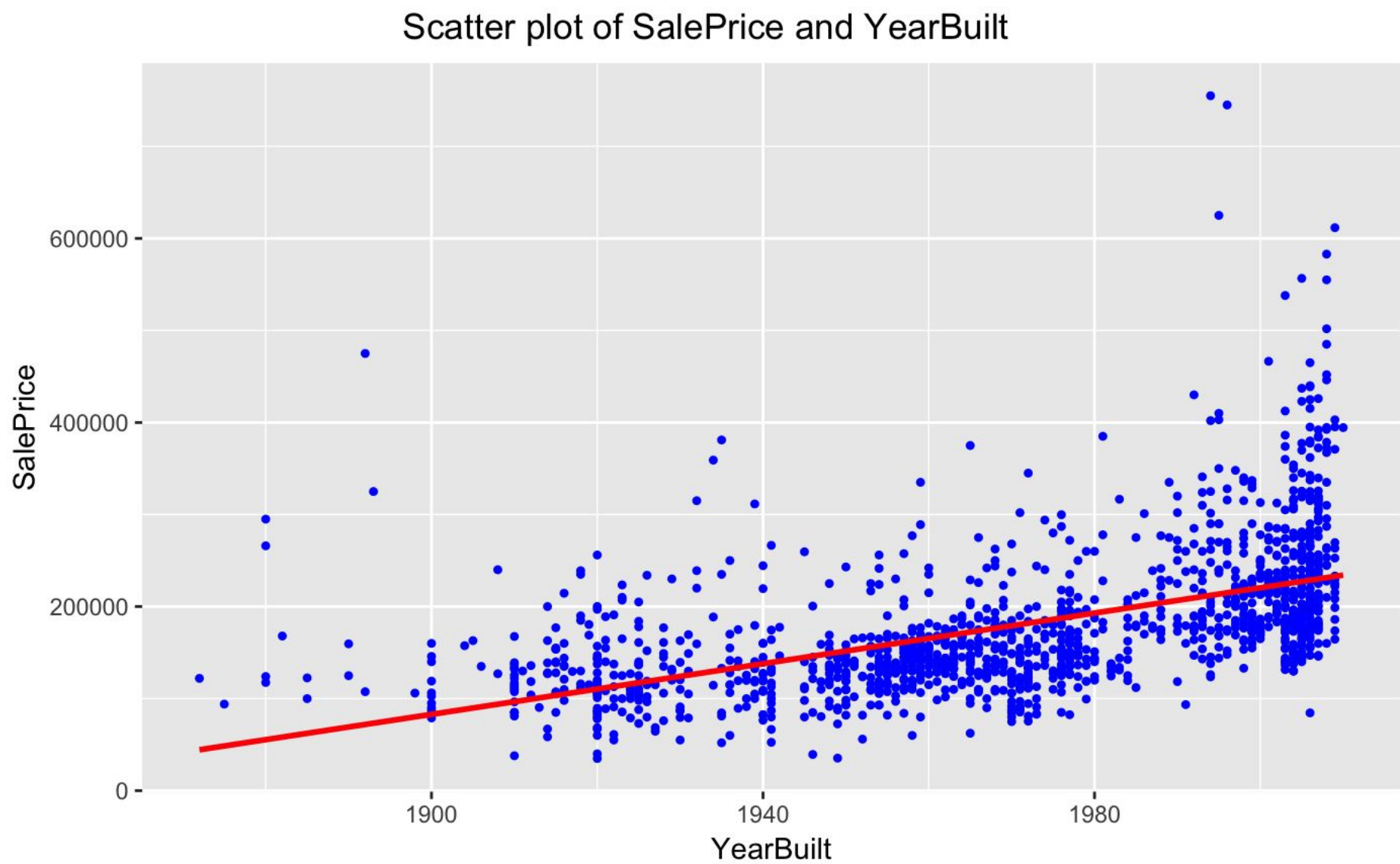
Right skewed, heavy tailed.

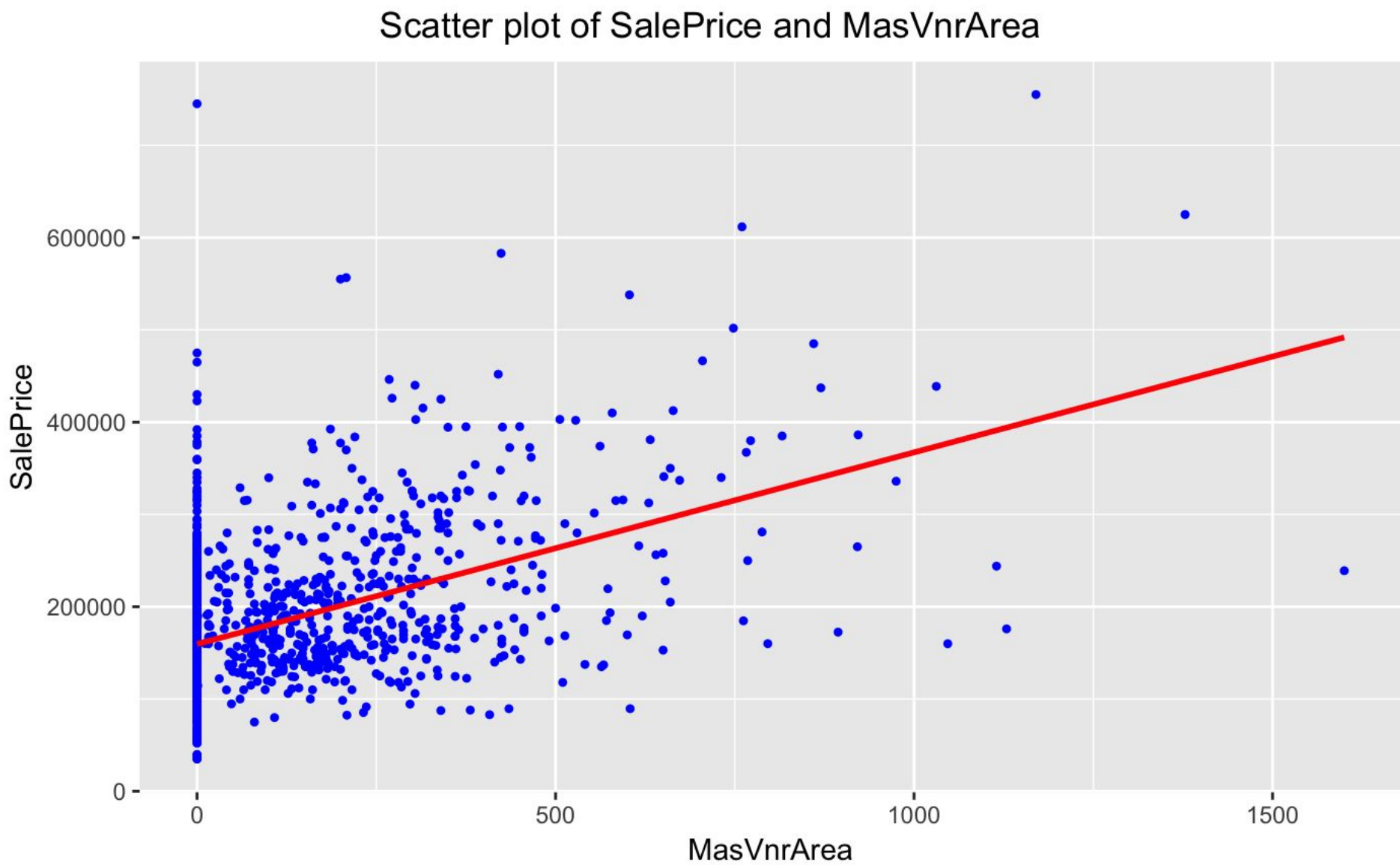


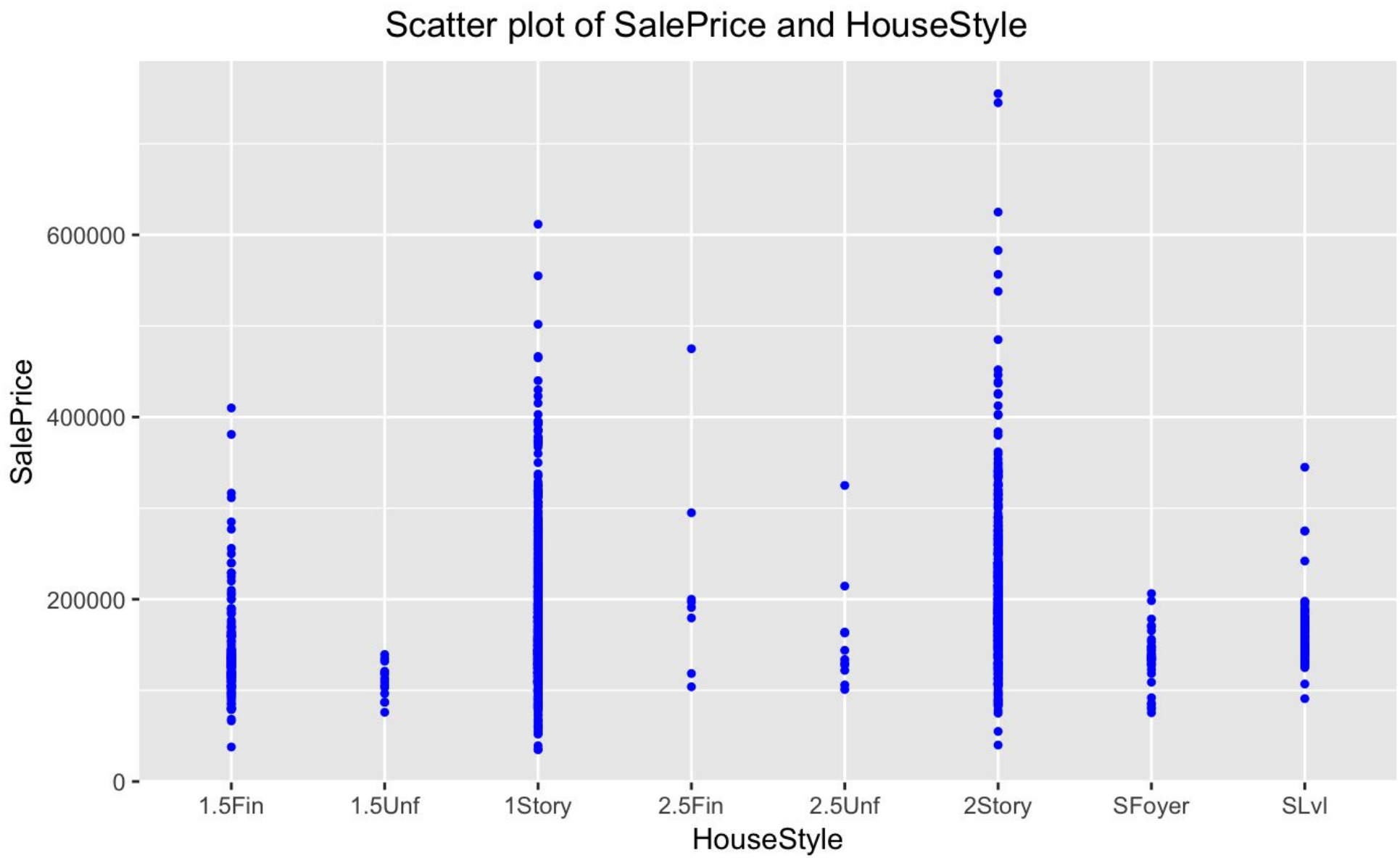
Much more symmetric.

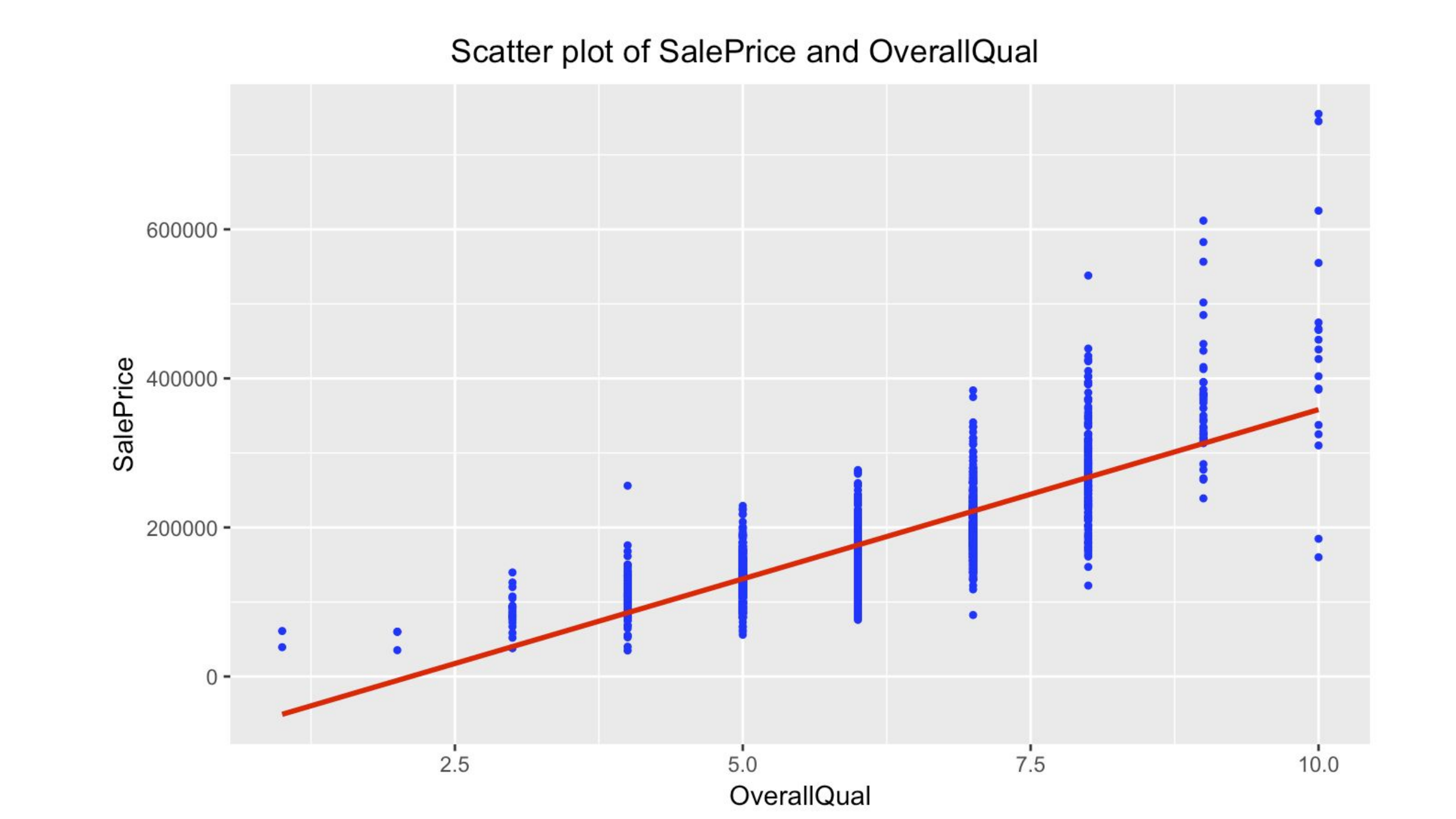
Let’s take a look of some scatterplots


Scatter plot of SalePrice and OverallCond
SalePrice
600000
400000
200000
0
2.5
5.0
7.5
OverallCond


Scatter plot of SalePrice and YearBuilt
SalePrice
600000
400000
200000
0
1900
1940
1980
YearBuilt


Scatter plot of SalePrice and MasVnrArea
SalePrice
600000
400000
200000
0
0
500
1000
1500
MasVnrArea


Scatter plot of SalePrice and HouseStyle
SalePrice
600000
400000
200000
0
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle


Scatter plot of SalePrice and OverallQual
SalePrice
600000
400000
200000
0
2.5
5.0
7.5
10.0
OverallQual

Some are numerical and some are categorical

For Regression with neural network, normalization is the best practice. (If you don’t do it…Good luck!)

2.Feature Engineering

- Normalization
  - Some variables have large magnitude therefore to get a more accurate model.

Before normalization

Huge magnitude difference

```
                (Other):   7
> summary(trainNoid)
   MSSubClass       MSZoning     LotFrontage        LotArea        Street         Alley      LotShape  LandContour
 Min.   : 20.0   C (all):  10   Min.   :  0.00   Min.   :  1300   Grvl:   6   1      :  50   IR1:484   Bnk:  63
 1st Qu.: 20.0   FV     :  65   1st Qu.: 42.00   1st Qu.:  7554   Pave:1454   2      :  41   IR2: 41   HLS:  50
 Median : 50.0   RH     :  16   Median : 63.00   Median :  9478               missing:1369   IR3: 10   Low:  36
 Mean   : 56.9   RL     :1151   Mean   : 57.62   Mean   : 10517                              Reg:925   Lvl:1311
 3rd Qu.: 70.0   RM     : 218   3rd Qu.: 79.00   3rd Qu.: 11602
 Max.   :190.0                  Max.   :313.00   Max.   :215245

  Utilities      LotConfig    LandSlope   Neighborhood   Condition1     Condition2     BldgType      HouseStyle
 AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445   1Fam  :1220   1Story :726
 NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6   2fmCon:  31   2Story :445
               FR2    :  47   Sev:  13   OldTown:113   Artery :  48   Artery :   2   Duplex:  52   1.5Fin :154
               FR3    :   4              Edwards:100   RRAn   :  26   PosN   :   2   Twnhs :  43   SLvl   : 65
               Inside :1052              Somerst: 86   PosN   :  19   RRNn   :   2   TwnhsE: 114   SFoyer : 37
                                         Gilbert: 79   RRAe   :  11   PosA   :   1                 1.5Unf : 14
                                         (Other):707   (Other):  15   (Other):   2                 (Other): 19
  OverallQual      OverallCond      YearBuilt     YearRemodAdd    RoofStyle      RoofMatl     Exterior1st
 Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13   CompShg:1434   VinylSd:515
 1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141   Tar&Grv:  11   HdBoard:222
 Median : 6.000   Median :5.000   Median :1973   Median :1994   Gambrel:  11   WdShngl:   6   MetalSd:220
 Mean   : 6.099   Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286   WdShake:   5   Wd Sdng:206
 3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7   ClyTile:   1   Plywood:108
 Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2   Membran:   1   CemntBd: 61
                                                                              (Other):   2   (Other):128
  Exterior2nd    MasVnrType    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual   BsmtCond     BsmtExposure
 VinylSd:504   BrkCmn : 15   Min.   :   0.0   Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
 MetalSd:214   BrkFace:445   1st Qu.:   0.0   Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
 HdBoard:207   None   :864   Median :   0.0   Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
 Wd Sdng:197   Stone  :128   Mean   : 103.7   TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
 Plywood:142   NA's   :  8   3rd Qu.: 166.0             TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
 CmentBd: 60                 Max.   :1600.0                       Wood  :  3
 (Other):136                 NA's   :8
 BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2       BsmtUnfSF      TotalBsmtSF      Heating     HeatingQC
 ALQ :220     Min.   :   0.0   ALQ :  19    Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741
 BLQ :148     1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49
 GLQ :418     Median : 383.5   GLQ :  14    Median :   0.00   Median : 477.5   Median : 991.5   GasW :  18   Gd:241
 LwQ : 74     Mean   : 443.6   LwQ :  46    Mean   :  46.55   Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
 Rec :133     3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
 Unf :430     Max.   :5644.0   Unf :1256    Max.   :1474.00   Max.   :2336.0                    Wall :   4
```

After normalization

no magnitude difference

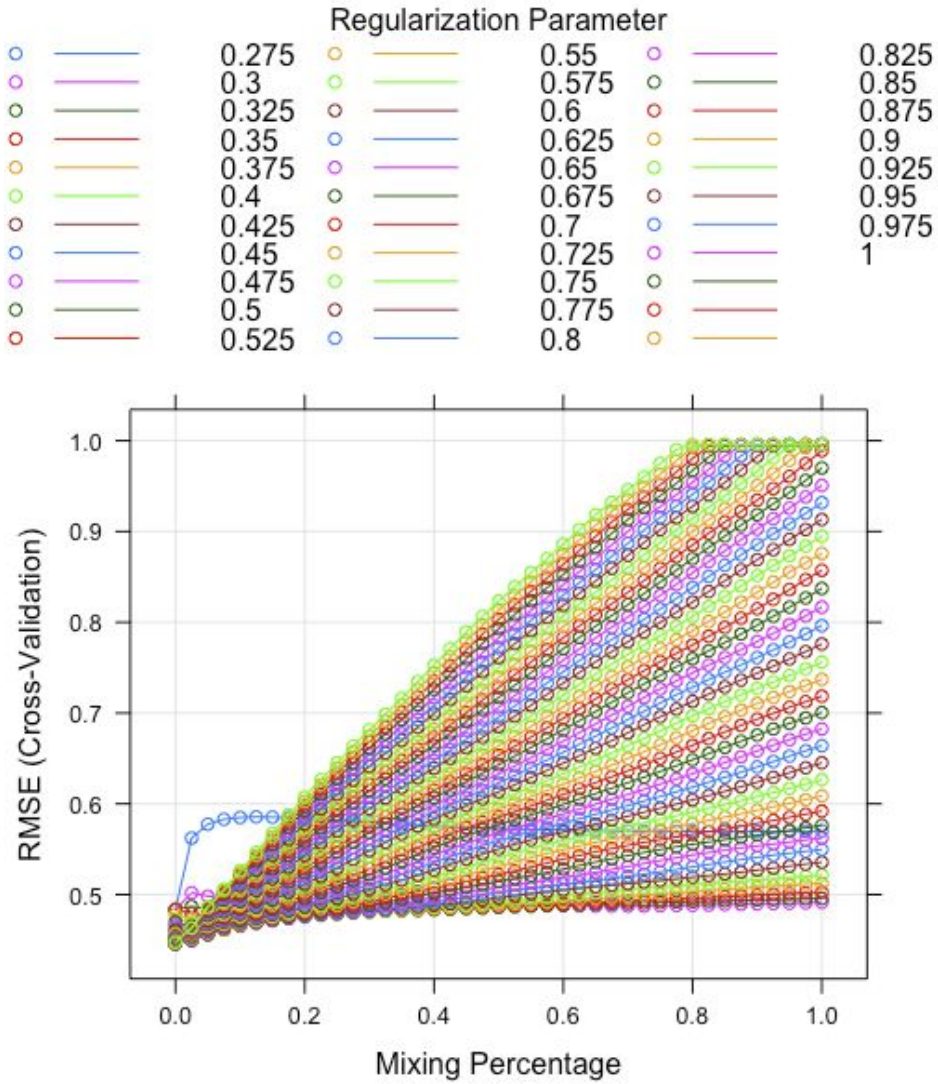
```
> summary(trainNnet)
   MSSubClass         LotFrontage          LotArea          OverallQual       OverallCond         YearBuilt
 Min.   :-0.878447   Min.   :-1.649841   Min.   :-0.918036   Min.   :-3.1105   Min.   :-3.31304   Min.   :-3.159531
 1st Qu.:-0.878447   1st Qu.:-0.479128   1st Qu.:-0.287804   1st Qu.:-0.8786   1st Qu.:-0.53834   1st Qu.:-0.572212
 Median :-0.147873   Median : 0.177614   Median :-0.106285   Median :-0.1346   Median :-0.53834   Median : 0.108662
 Mean   : 0.001554   Mean   : 0.002171   Mean   : 0.001912   Mean   : 0.0289   Mean   : 0.01328   Mean   : 0.007523
 3rd Qu.: 0.339176   3rd Qu.: 0.634477   3rd Qu.: 0.105040   3rd Qu.: 0.6094   3rd Qu.: 0.38655   3rd Qu.: 0.959753
 Max.   : 3.261470   Max.   : 7.287555   Max.   :20.006078   Max.   : 2.8413   Max.   : 3.16125   Max.   : 1.266146

  YearRemodAdd       MasVnrArea         BsmtFinSF1         BsmtFinSF2         BsmtUnfSF          TotalBsmtSF
 Min.   :-1.72793   Min.   :-0.590625   Min.   :-0.98889   Min.   :-0.29365   Min.   :-1.285     Min.   :-2.22092
 1st Qu.:-0.84790   1st Qu.:-0.590625   1st Qu.:-0.98889   1st Qu.:-0.29365   1st Qu.:-0.726     1st Qu.:-0.58082
 Median : 0.44770   Median :-0.590625   Median :-0.08907   Median :-0.29365   Median :-0.183     Median :-0.11787
 Mean   : 0.01594   Mean   : 0.007387   Mean   : 0.02256   Mean   : 0.00507   Mean   : 0.026     Mean   : 0.05296
 3rd Qu.: 0.91216   3rd Qu.: 0.352233   3rd Qu.: 0.60813   3rd Qu.:-0.29365   3rd Qu.: 0.552     3rd Qu.: 0.57625
 Max.   : 1.20550   Max.   : 8.079326   Max.   :11.30789   Max.   : 8.65256   Max.   : 3.976     Max.   :11.55842

   X1stFlrSF          X2ndFlrSF         LowQualFinSF         GrLivArea          BsmtFullBath
 Min.   :-1.914270   Min.   :-0.804295   Min.   :-0.103339   Min.   :-2.109225   Min.   :-0.83456
 1st Qu.:-0.731114   1st Qu.:-0.804295   1st Qu.:-0.103339   1st Qu.:-0.718928   1st Qu.:-0.83456
 Median :-0.201807   Median :-0.804295   Median :-0.103339   Median :-0.102729   Median :-0.83456
 Mean   : 0.001151   Mean   : 0.008488   Mean   : 0.001564   Mean   : 0.008158   Mean   : 0.01913
 3rd Qu.: 0.618100   3rd Qu.: 0.880937   3rd Qu.:-0.103339   3rd Qu.: 0.497101   3rd Qu.: 1.11132
 Max.   : 9.123336   Max.   : 3.895238   Max.   :14.109149   Max.   : 7.911702   Max.   : 3.05720

  BsmtHalfBath         FullBath           HalfBath         BedroomAbvGr       KitchenAbvGr       TotRmsAbvGrd
 Min.   :-0.247266   Min.   :-2.88227   Min.   :-0.7857   Min.   :-3.661380   Min.   :-0.19463   Min.   :-2.233566
 1st Qu.:-0.247266   1st Qu.:-1.05537   1st Qu.:-0.7857   1st Qu.:-1.106149   1st Qu.:-0.19463   1st Qu.:-0.975981
 Median :-0.247266   Median : 0.77154   Median :-0.7857   Median : 0.171467   Median :-0.19463   Median :-0.347189
 Mean   : 0.006098   Mean   :-0.00401   Mean   : 0.0164   Mean   :-0.001364   Mean   :-0.04122   Mean   :-0.001306
 3rd Qu.:-0.247266   3rd Qu.: 0.77154   3rd Qu.: 1.2055   3rd Qu.: 0.171467   3rd Qu.:-0.19463   3rd Qu.: 0.281604
 Max.   : 8.123145   Max.   : 2.59844   Max.   : 3.1967   Max.   : 4.004314   Max.   :10.06834   Max.   : 3.425566

   Fireplaces        GarageYrBlt         GarageCars         GarageArea         WoodDeckSF         OpenPorchSF
 Min.   :-0.99446   Min.   :-3.174469   Min.   :-1.377686   Min.   :-1.830642   Min.   :-0.77042   Min.   :-0.72317
 1st Qu.:-0.99446   1st Qu.:-0.663220   1st Qu.:-1.377686   1st Qu.:-0.660909   1st Qu.:-0.77042   1st Qu.:-0.72317
 Median : 0.55663   Median : 0.065852   Median : 0.206884   Median :-0.088144   Median :-0.72308   Median :-0.29229
 Mean   : 0.01178   Mean   : 0.009001   Mean   : 0.003187   Mean   : 0.005703   Mean   : 0.01362   Mean   : 0.01213
 3rd Qu.: 0.55663   3rd Qu.: 0.956941   3rd Qu.: 0.206884   3rd Qu.: 0.444286   3rd Qu.: 0.60620   3rd Qu.: 0.35402
 Max.   : 3.65880   Max.   : 1.280973   Max.   : 3.376024   Max.   : 4.934982   Max.   : 5.99040   Max.   : 7.69432
```

3. Model Training - Elastic Net

Linear regression with regularization

Hyper parameter tuning with
10-fold cross validation

- The best result for elastic net parameter:
- mixing percentage ：$\alpha = 0$
- regularization parameter ：$\lambda = 0.725$

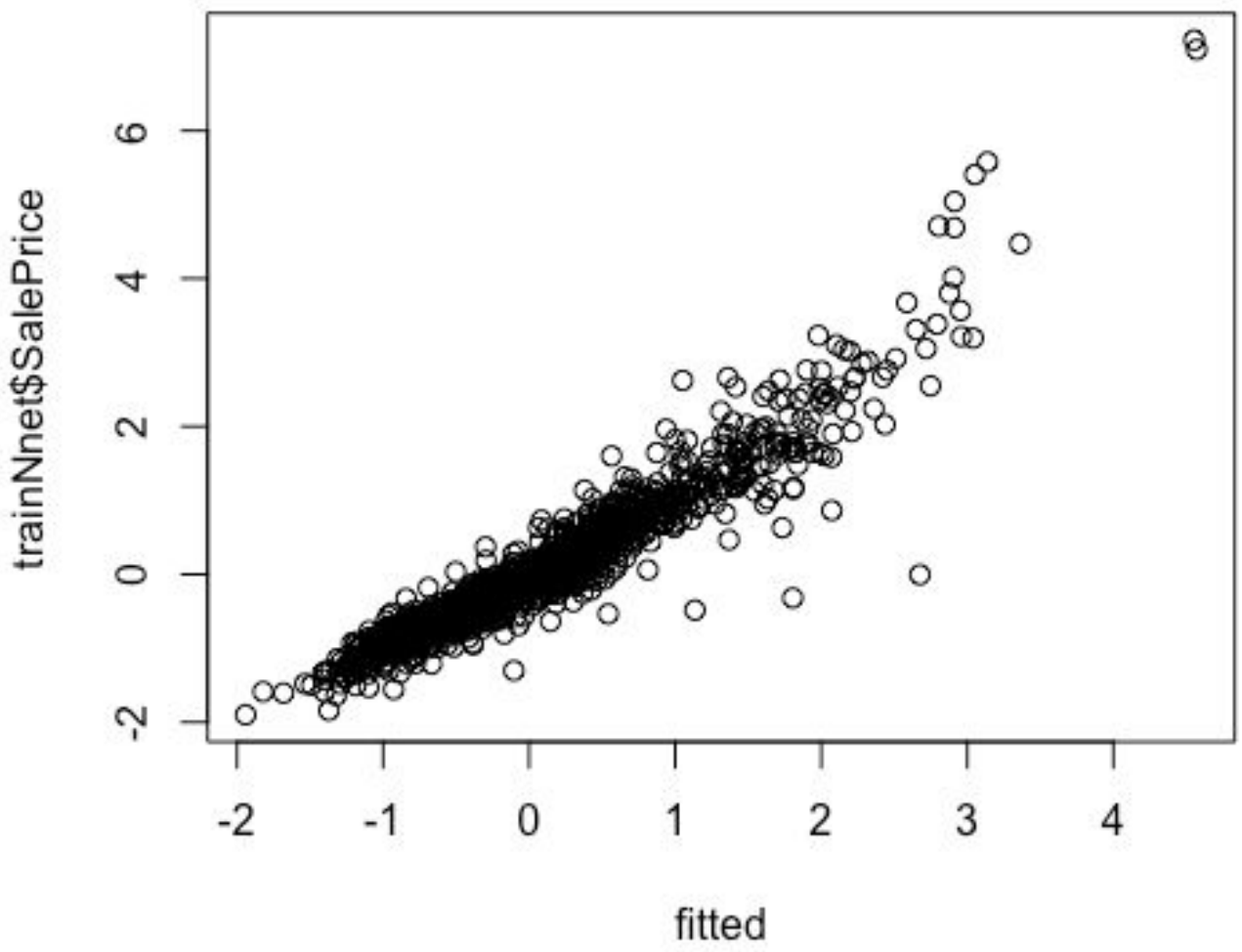



```
glmnetFinalmodel <- glmnet(SalePrice ~ ., data = trainNnet, lambda = 0.725, alpha = 0, family = "gaussian")
```

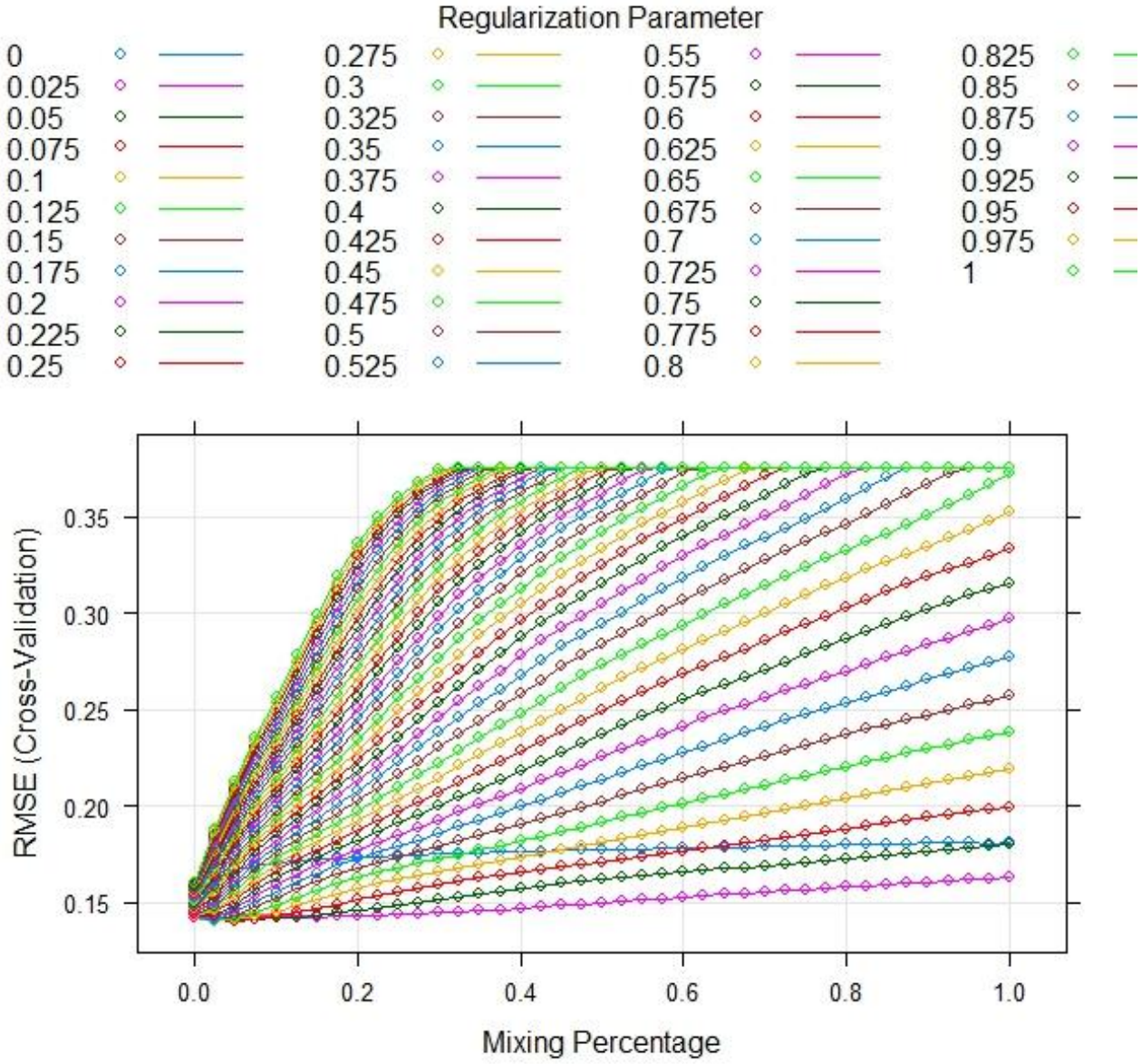
Results

	alpha	lambda	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
30	0	0.725	0.4457456	0.8043806	0.2408655	0.1278541	0.1182548	0.02057472

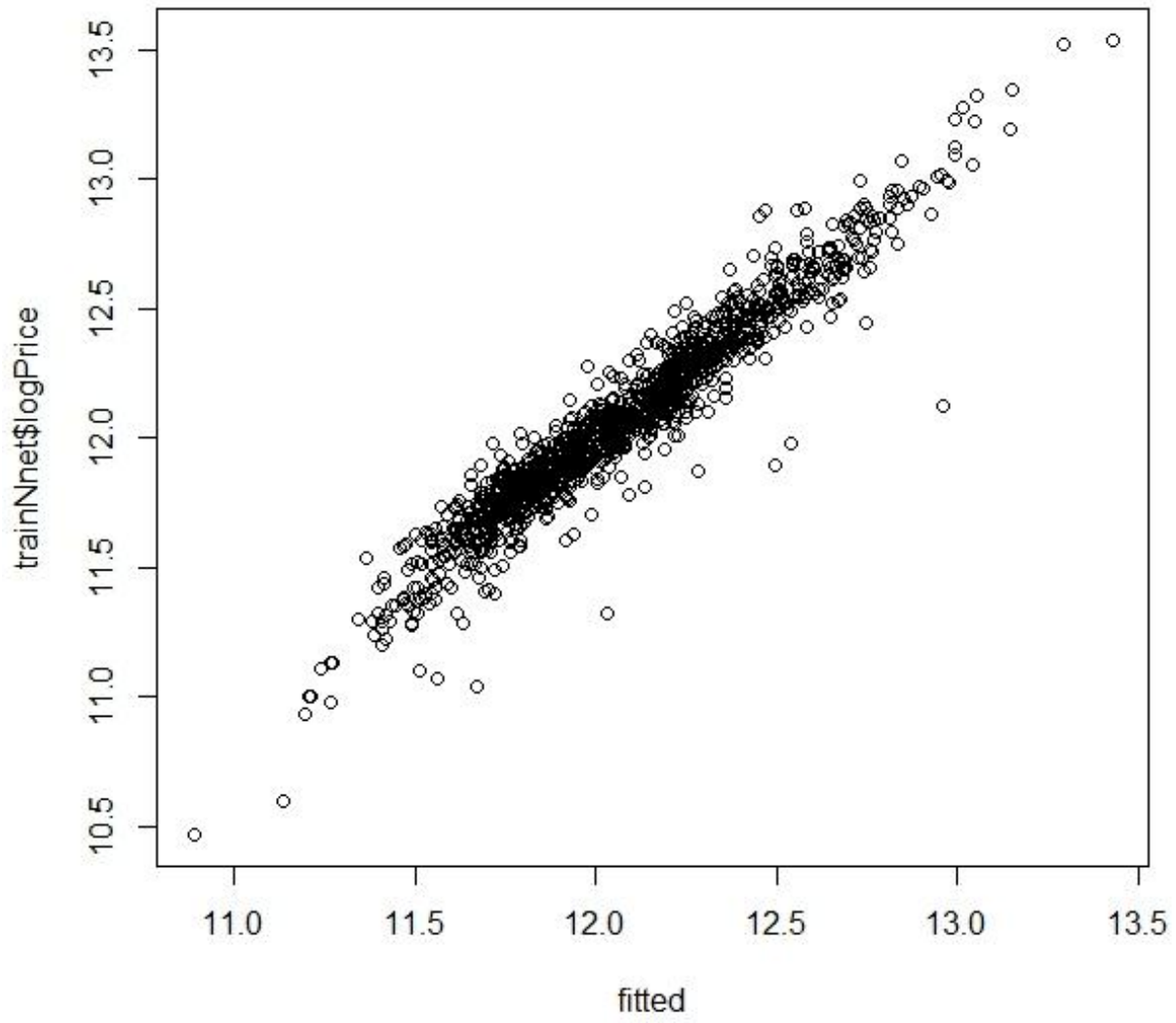
A little bit cone shape, so if we take log of the Salesprice it could be better.



Hyperparameter tuning after taking log of salesprice



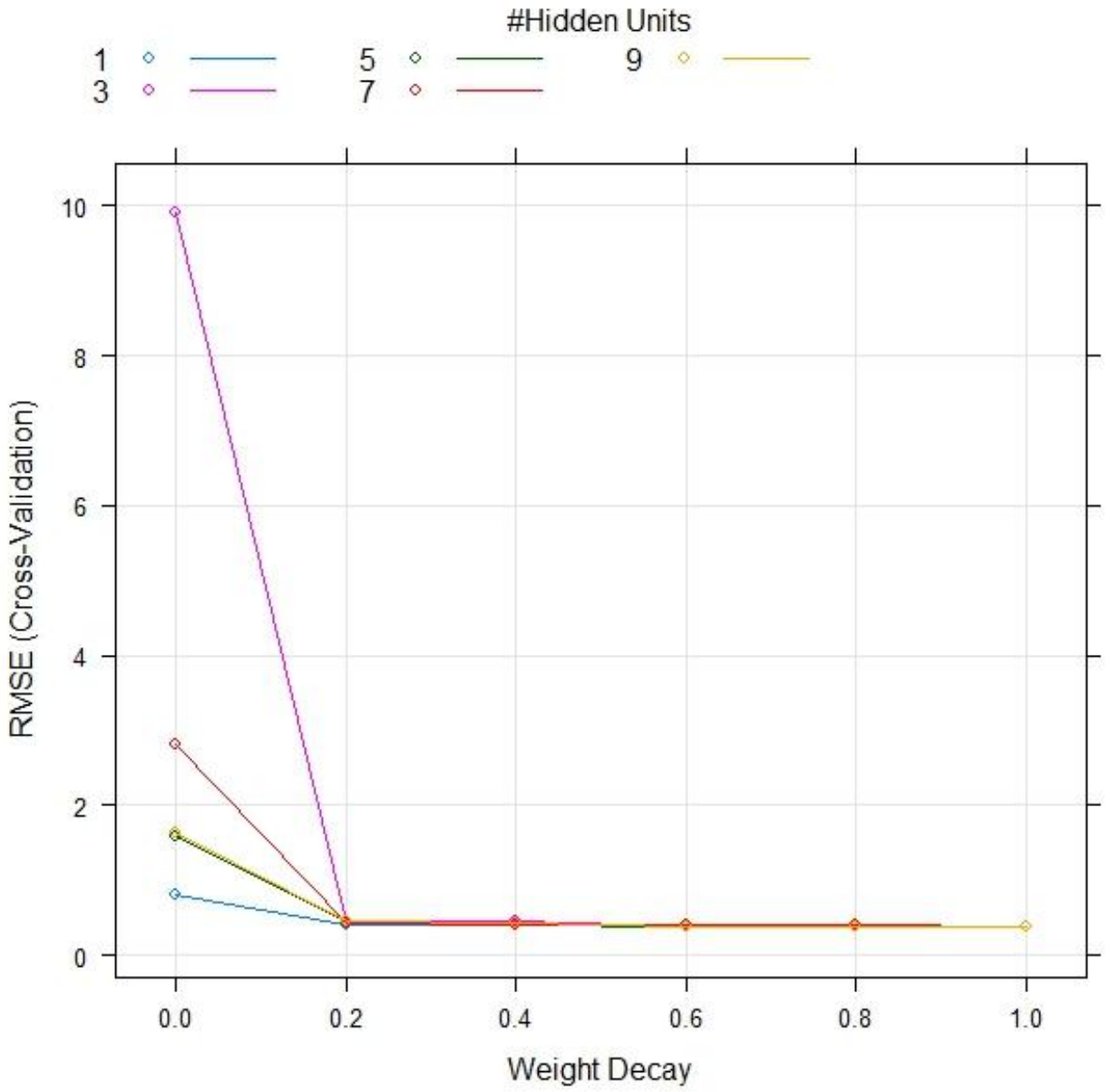
Scatter Plot of log(Saleprice) against fitted value.



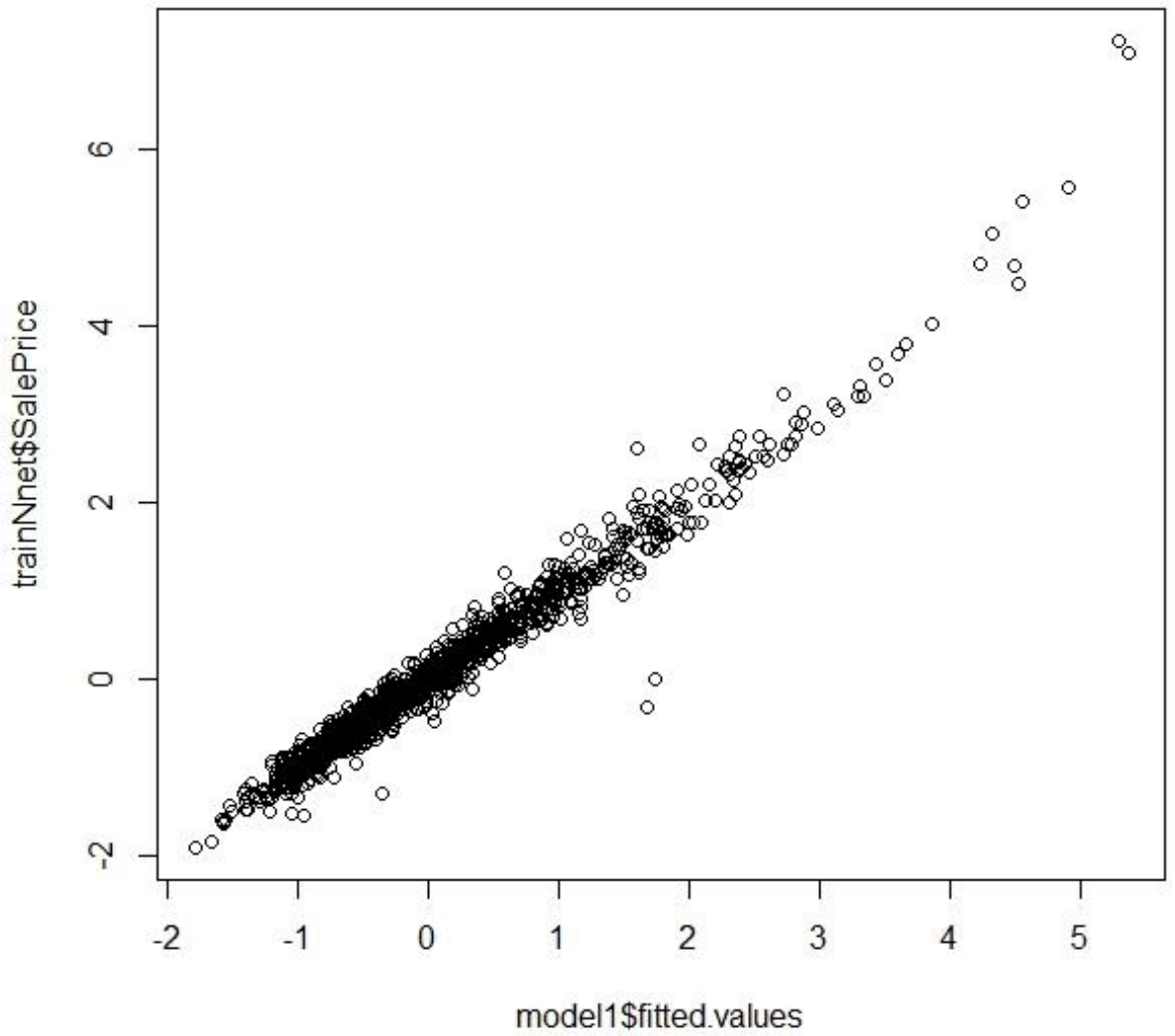
3. Model Training

Neural Network Tuning

The best parameter :
Size = 5 i.e. 5 nodes on hidden layer
Decay = 1 regularization parameter
No activation function - because it’s regression.



Result
-No more cone shape.



Results

size	decay	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
5	1	0.3761174876	0.8605903355	0.2088225498	0.1023479498	0.06214833052	0.02458673067

Compare RMSE

Elastic Net	0.44
Neural Network	0.37
Elastic Net(log Saleprice)	0.12

```
> glmnetModel$beta
295 x 1 sparse Matrix of class "dgCMatrix"
                              s0
MSSubClass           -0.0032546597
LotFrontage           0.0001003035
LotArea               0.0134725970
OverallQual           0.0466032401
OverallCond           0.0223424303
YearBuilt             0.0103836389
YearRemodAdd          0.0158688338
MasVnrArea            0.0072609595
BsmtFinSF1            0.0141887693
BsmtFinSF2            0.0009436890
BsmtUnfSF             .
TotalBsmtSF           0.0234638895
X1stFlrSF             0.0273830818
X2ndFlrSF             0.0192776771
LowQualFinSF          .
GrLivArea             0.0417723758
BsmtFullBath          0.0090441296
BsmtHalfBath          .
FullBath              0.0161438232
HalfBath              0.0098153078
BedroomAbvGr          0.0052310356
KitchenAbvGr         -0.0079827851
TotRmsAbvGrd          0.0237742149
Fireplaces            0.0131700551
GarageYrBlt           0.0010746513
GarageCars            0.0209264340
GarageArea            0.0192536410
WoodDeckSF            0.0092703213
OpenPorchSF           0.0078170602
EnclosedPorch         .
```

Descriptive Recommendations

```
NeighborhoodStoneBr    0.0928962317
```

```
NeighborhoodCrawfor    0.0912275662
```


Tak
Obrigado
감사합니다
Gracias
Merci
شكرا لك
Thank you
Danke
Grazie
ありがとう