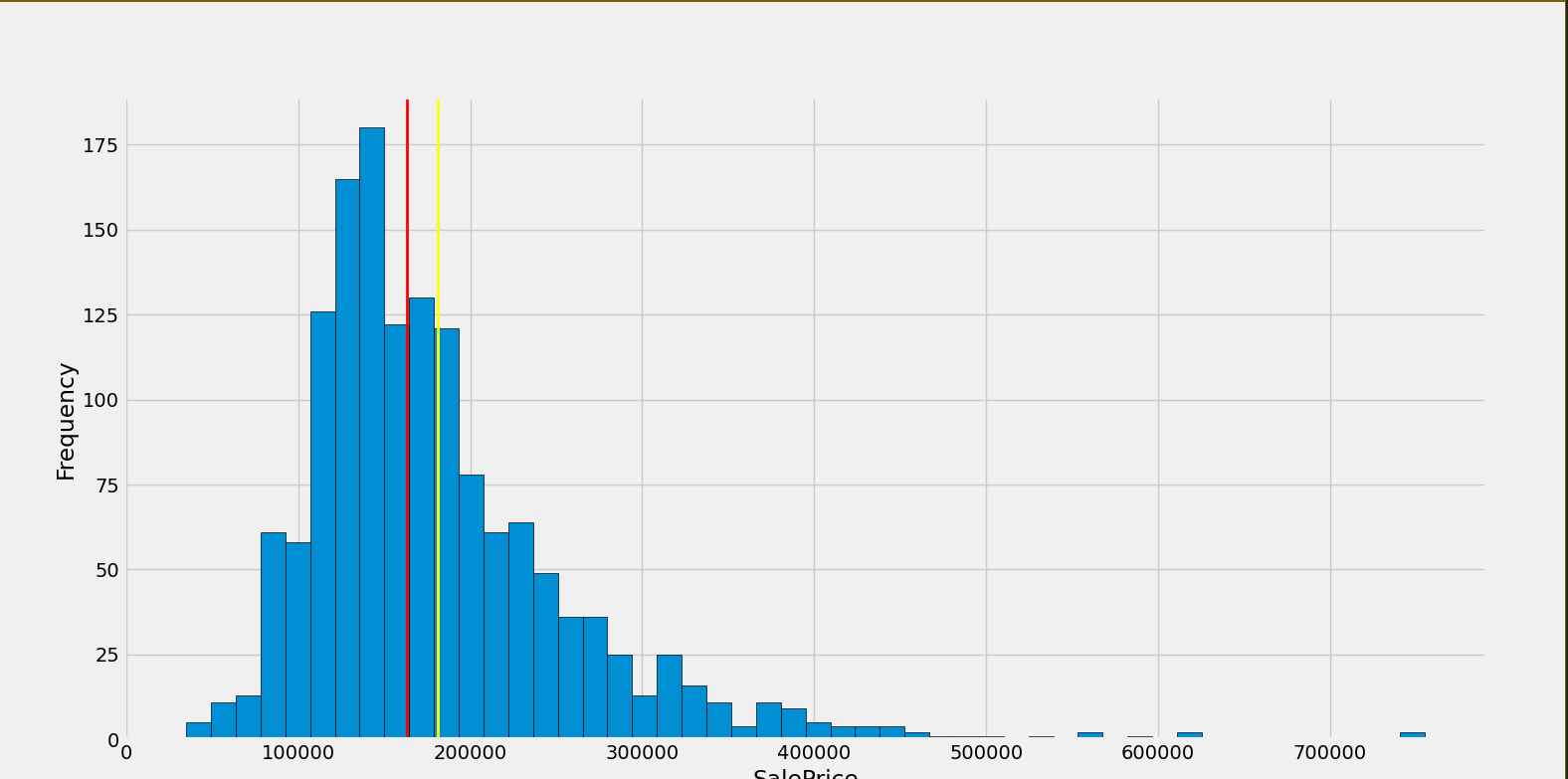
**SalePrice Description**

****

count 1460.000000

mean 180921.195890

std 79442.502883

min 34900.000000

25% 129975.000000

50% 163000.000000

75% 214000.000000

max 755000.000000

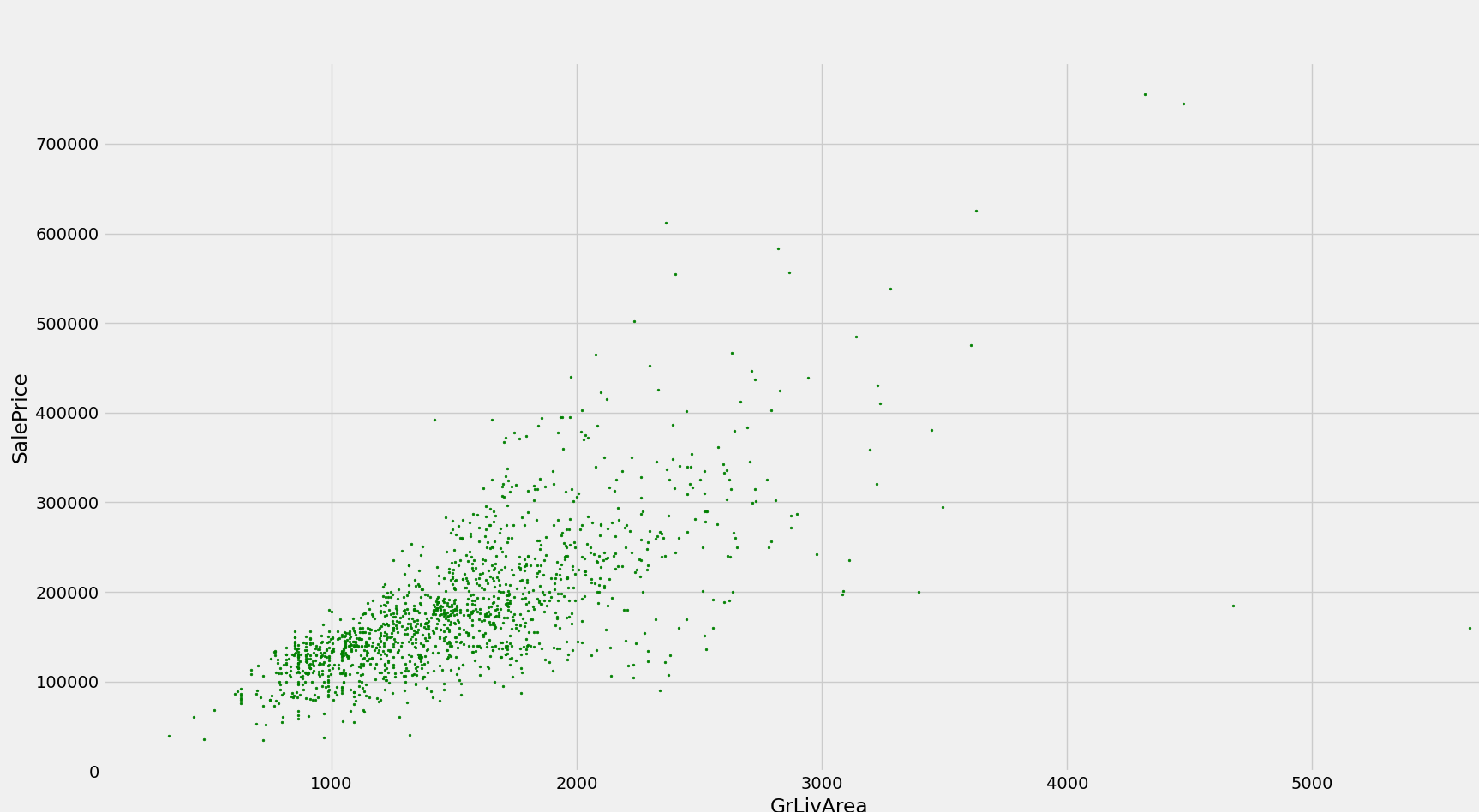
Skewness 1.8828757597682129

Kurtosis: 6.536281860064529

Conclusion:

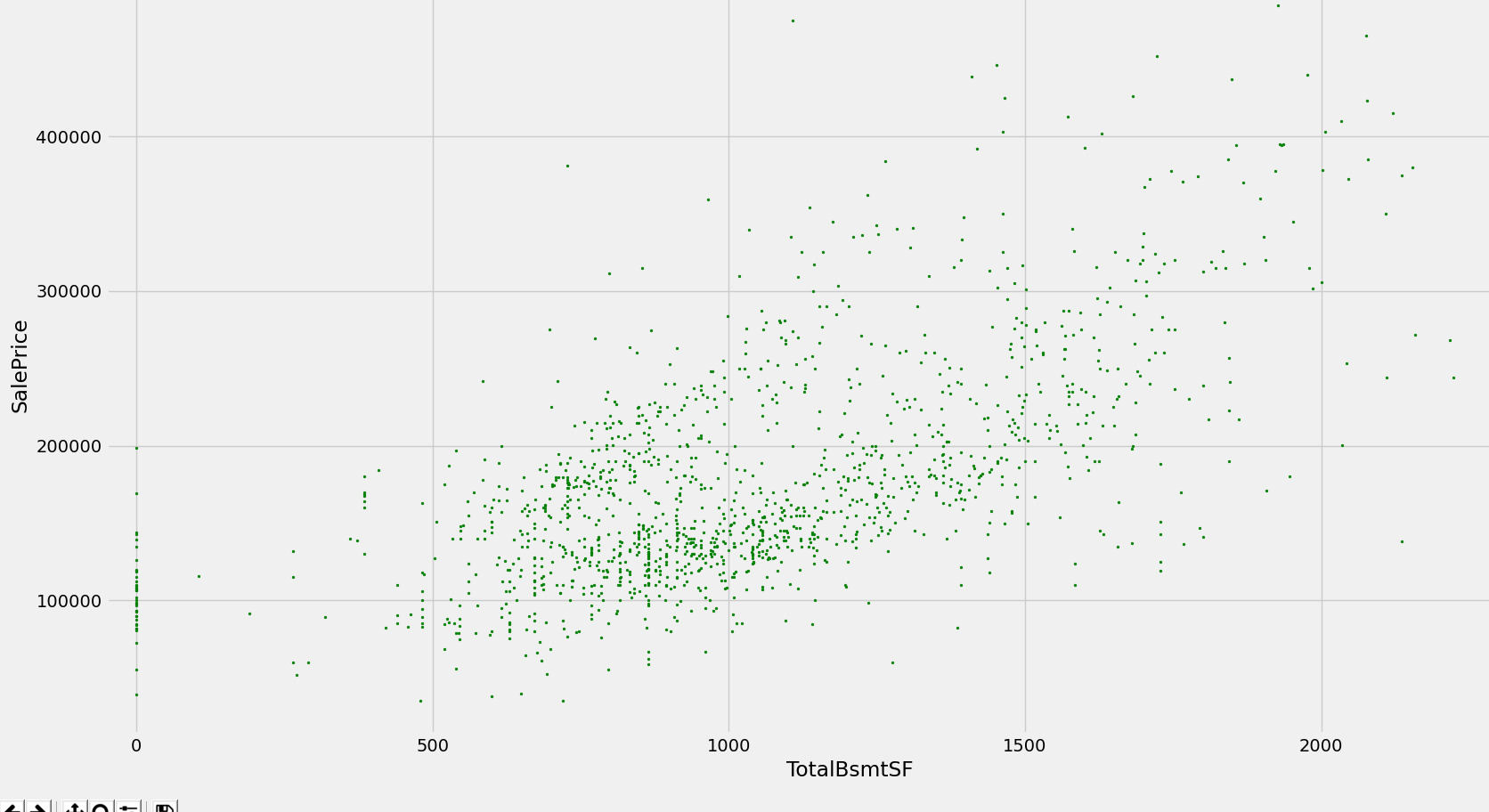
1. Positive skewness
2. Show peakedness
3. Deviate from normal distribution

GrLiveArea vs SalePrice:



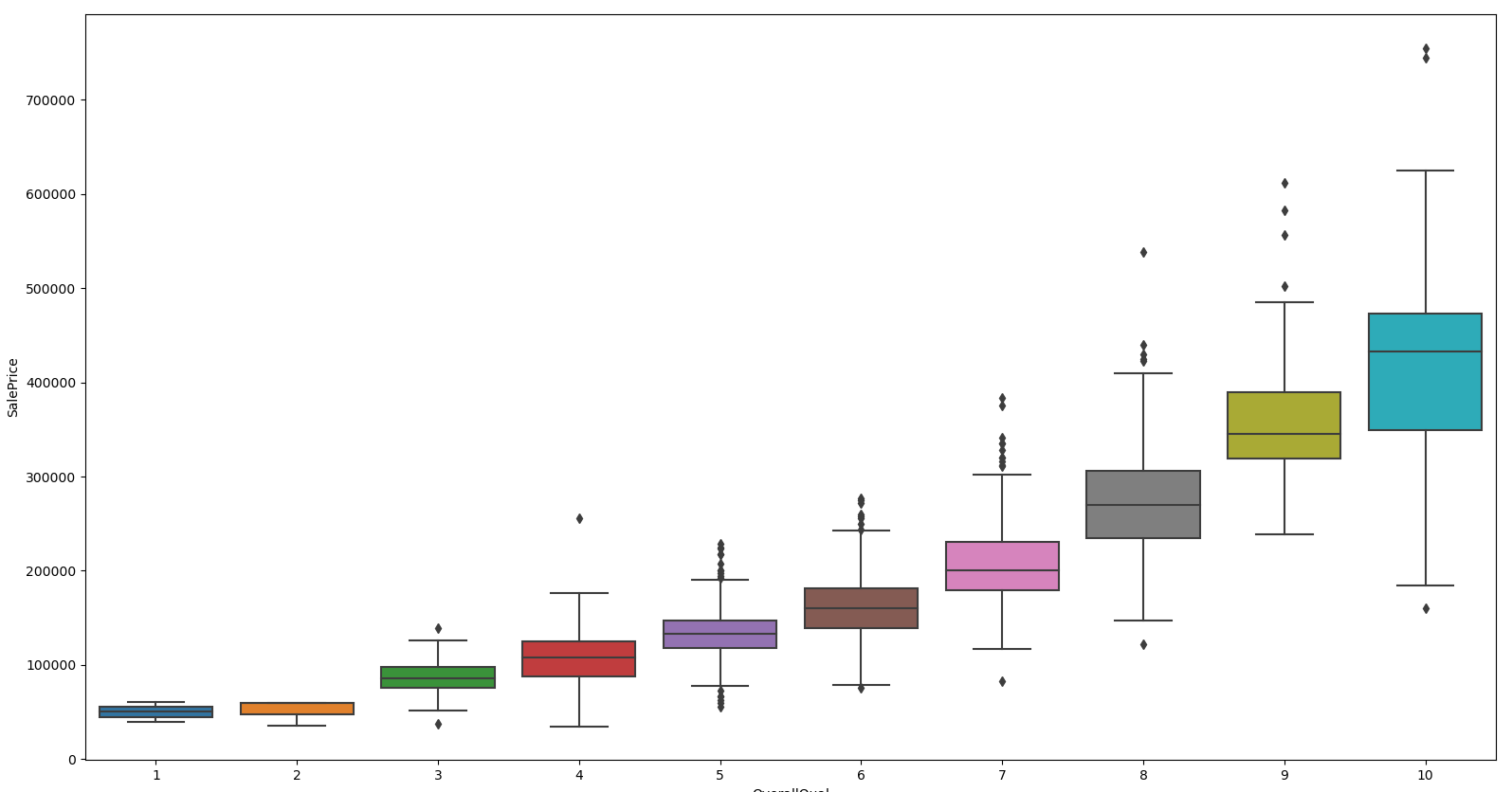
Linear Relationship

TotalBsmtSF vs SalePrice



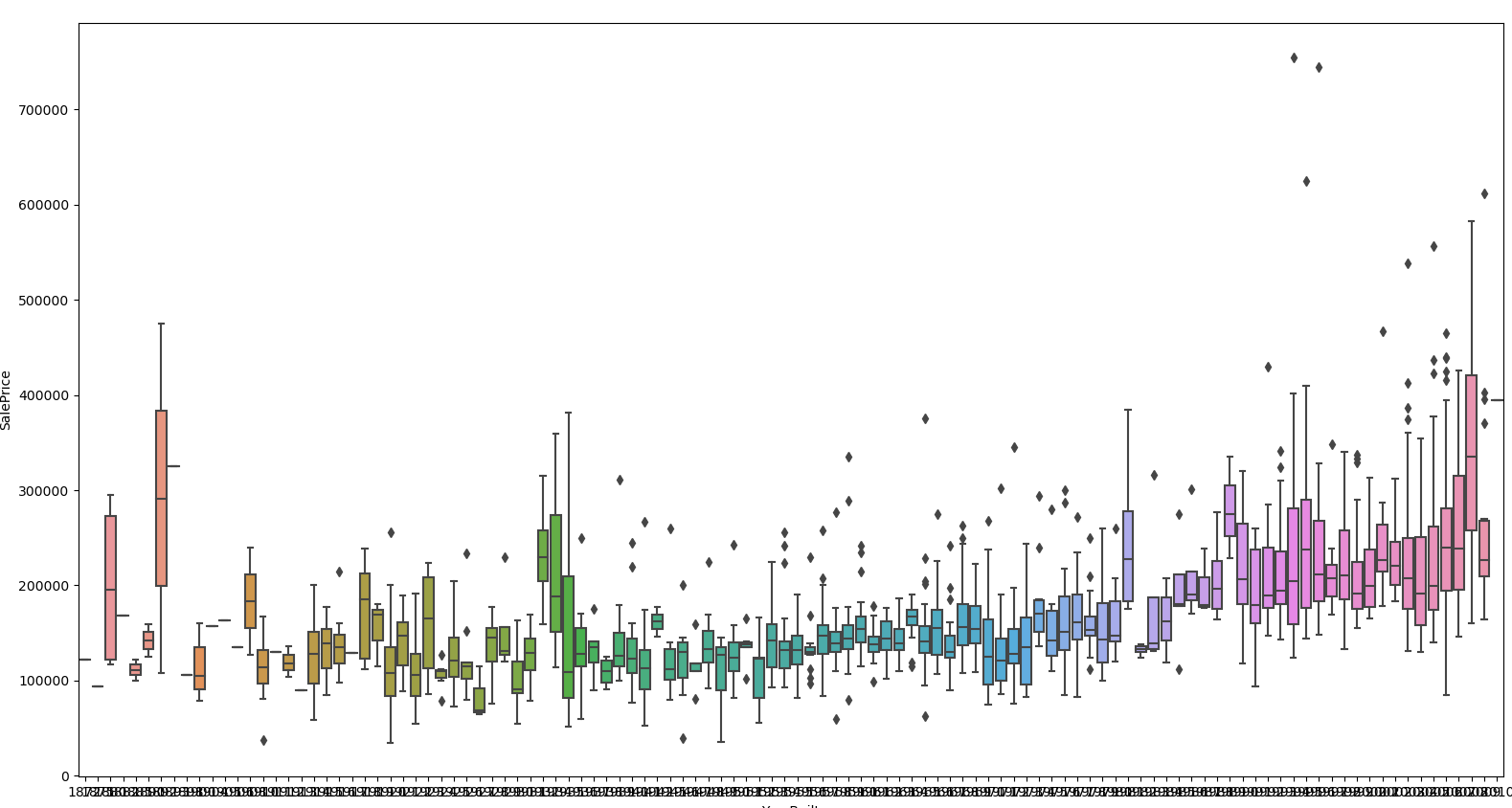
Linear relationship

OverallQual vs SalePrice



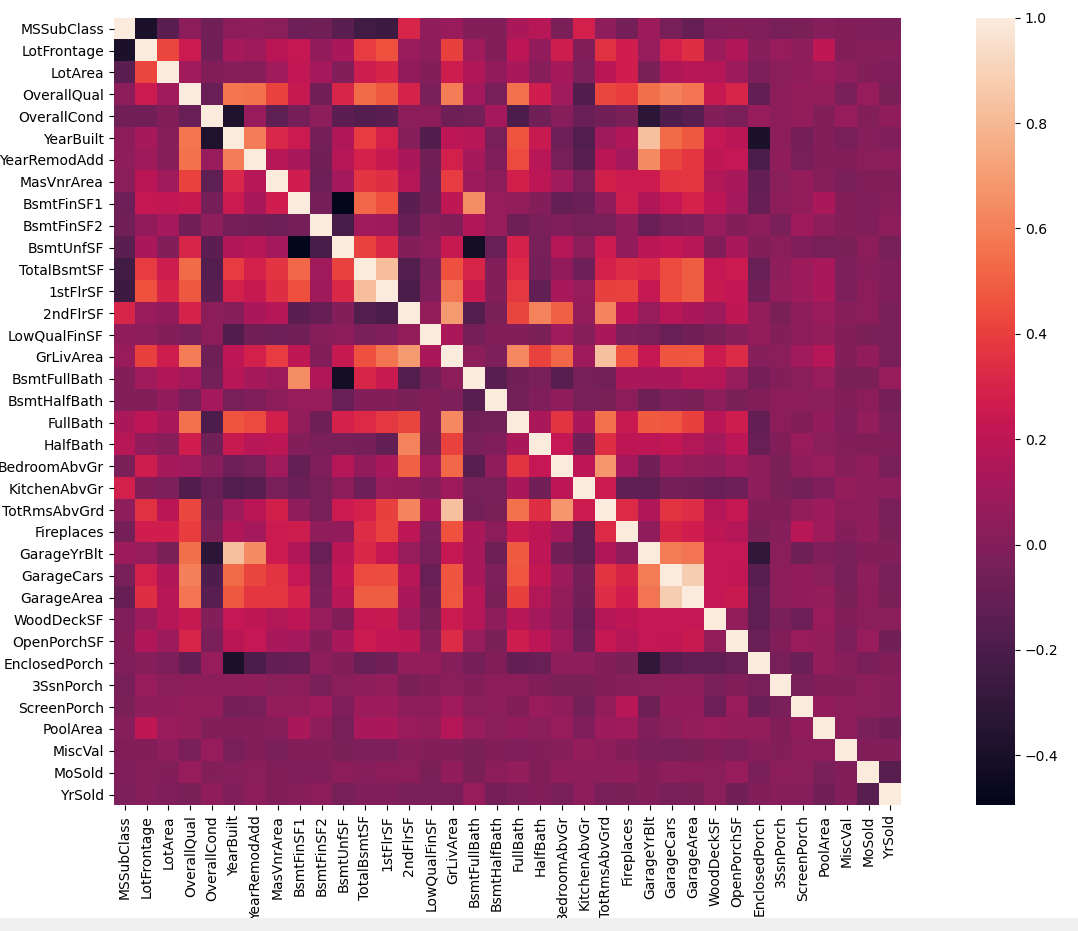
Good relation, strong tendency

YearBuilt Vs SalesPrice



Not strong tendency ,however new houses tends to have greater SalesPrice

Study Correlation Matrix



Highly correlated variables

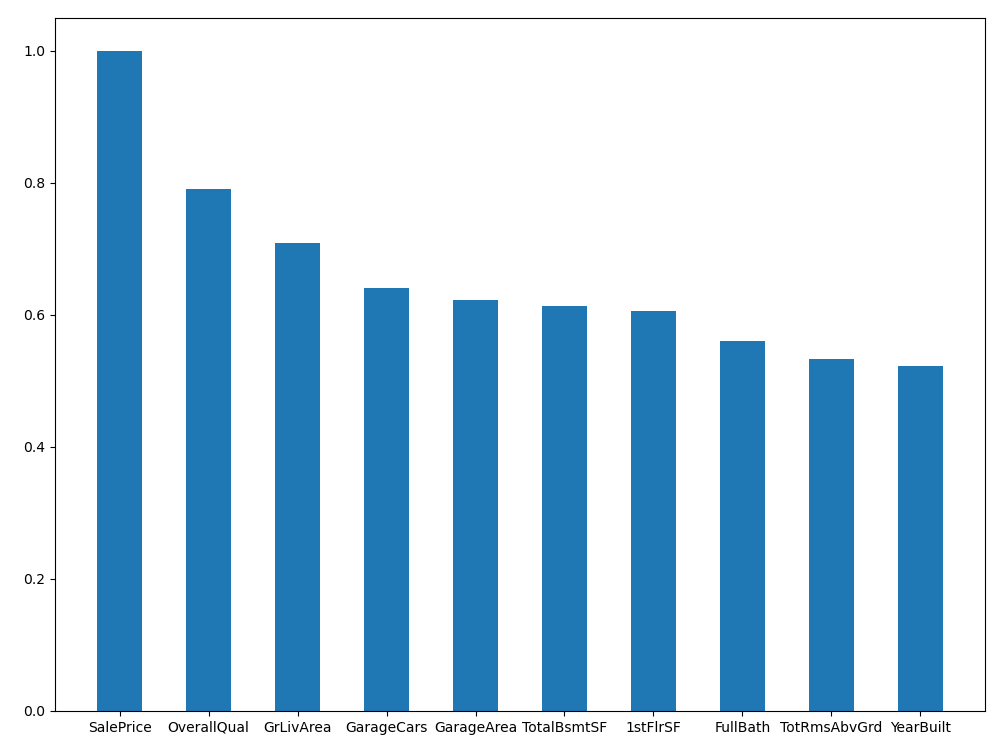
Yearbuilt, GarageYrBlt

GrLiveArea, TotRmsAbvGrd

GarageCars, GarageArea

TotalBsmtSF,1stFlrSF

SalePrice Correlation Matrix



OverallQual, GrLivArea, GarageCars highly correlated

ScatterPlot Between most correlated variables

Missing Data:

PoolQC 0.995205

MiscFeature 0.963014

Alley 0.937671

Fence 0.807534

FireplaceQu 0.472603

LotFrontage 0.177397

GarageCond 0.055479

GarageType 0.055479

GarageYrBlt 0.055479

GarageFinish 0.055479

GarageQual 0.055479

BsmtExposure 0.026027

BsmtFinType2 0.026027

BsmtFinType1 0.025342

BsmtCond 0.025342

BsmtQual 0.025342

MasVnrArea 0.005479

MasVnrType 0.005479

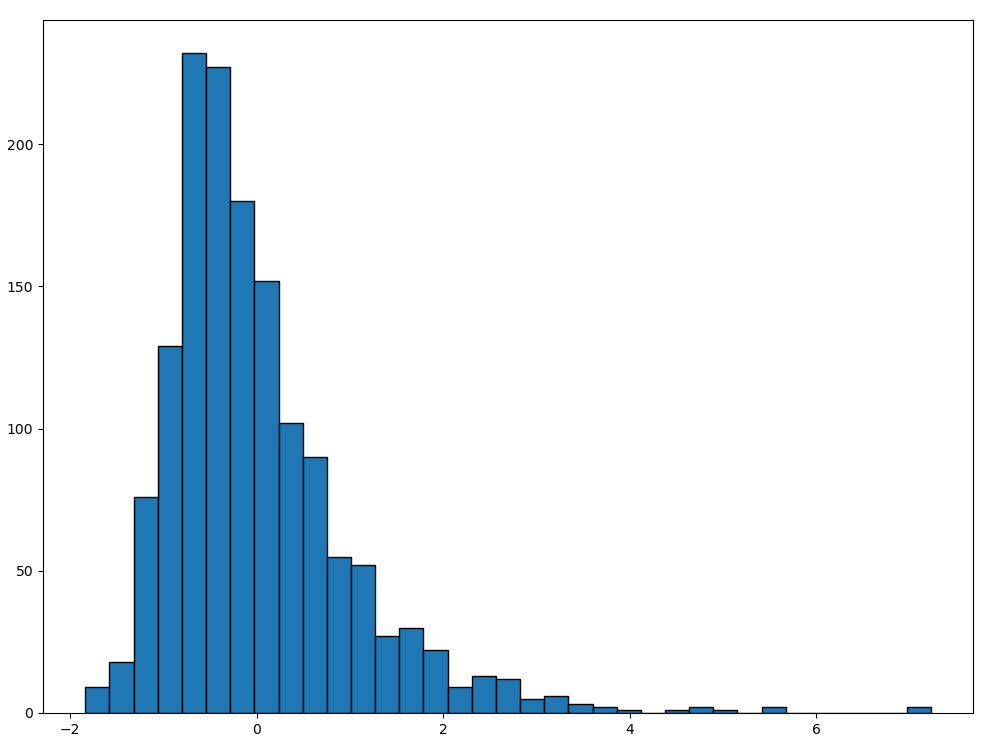
Electrical 0.000685

Drop columns with missing percentage above 20%

 the most important information regarding garages is expressed by 'GarageCars' and considering that we are just talking about 5% of missing data, I'll delete the mentioned 'GarageX' variables same logic applies to BsmtX

check ‘MasVnrArea' and 'MasVnrType'

Scaling SalePrice:



After Standard Scaling:

count 1.460000e+03

mean 1.362685e-16

std 1.000343e+00

min -1.838704e+00

25% -6.415162e-01

50% -2.256643e-01

75% 4.165294e-01

max 7.228819e+00

Outlier Analysis:

SalePrice after scaling

mean 1.362685e-16

std 1.000343e+00

min -1.838704e+00

25% -6.415162e-01

50% -2.256643e-01

75% 4.165294e-01

max 7.228819e+00

The min values are negative and close to zero

The larger values are positive and far from zero

**Min 10 SalePrice**

495 -1.838704

916 -1.833528

968 -1.800928

533 -1.783299

30 -1.774484

710 -1.623380

1337 -1.617084

705 -1.585604

1325 -1.585604

812 -1.573100

**Max 9 SalePirce**

178 4.040982

769 4.496348

440 4.710413

1046 4.730321

803 5.062146

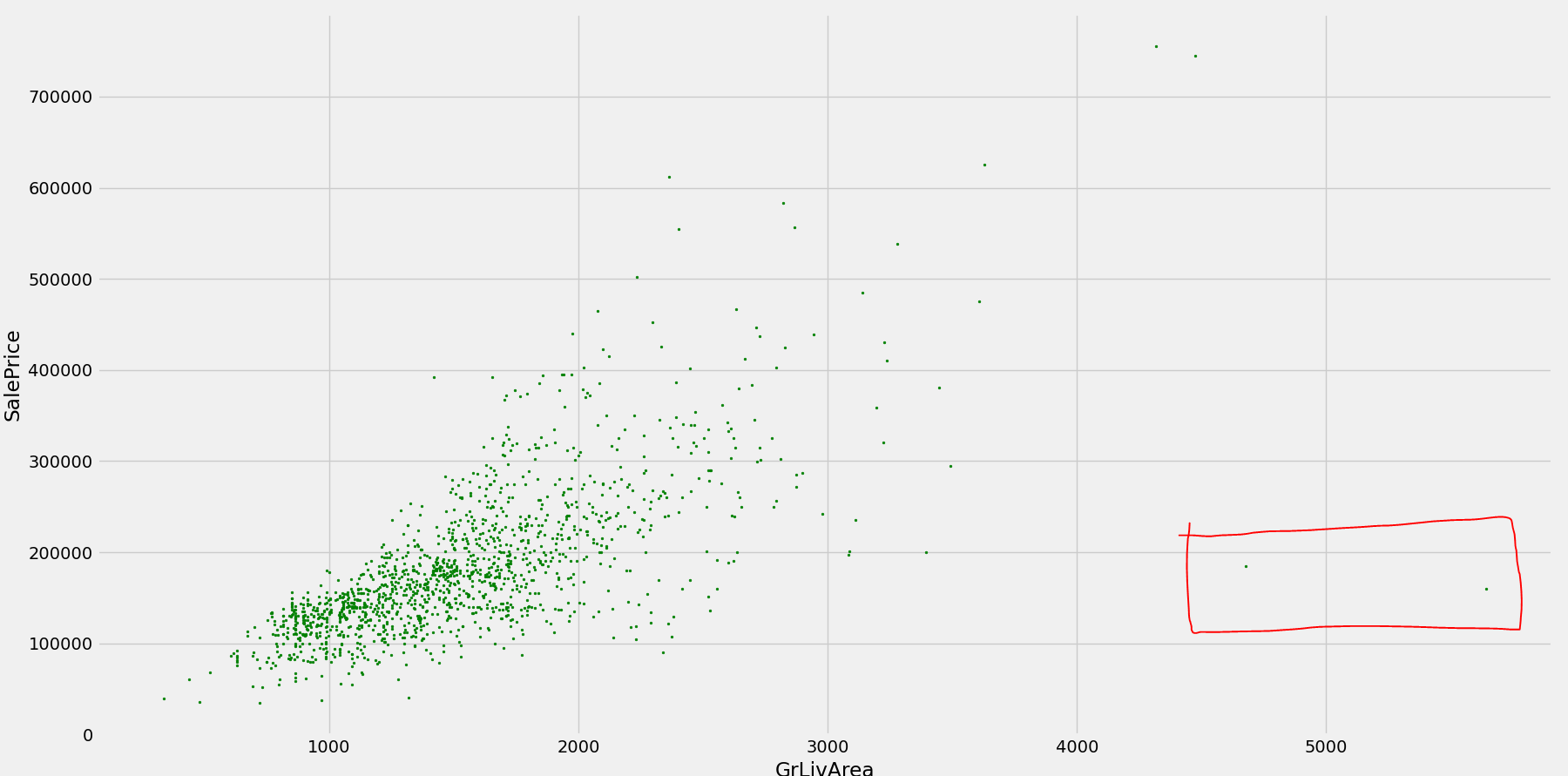
898 5.423840

1169 5.591855

1182 7.102899

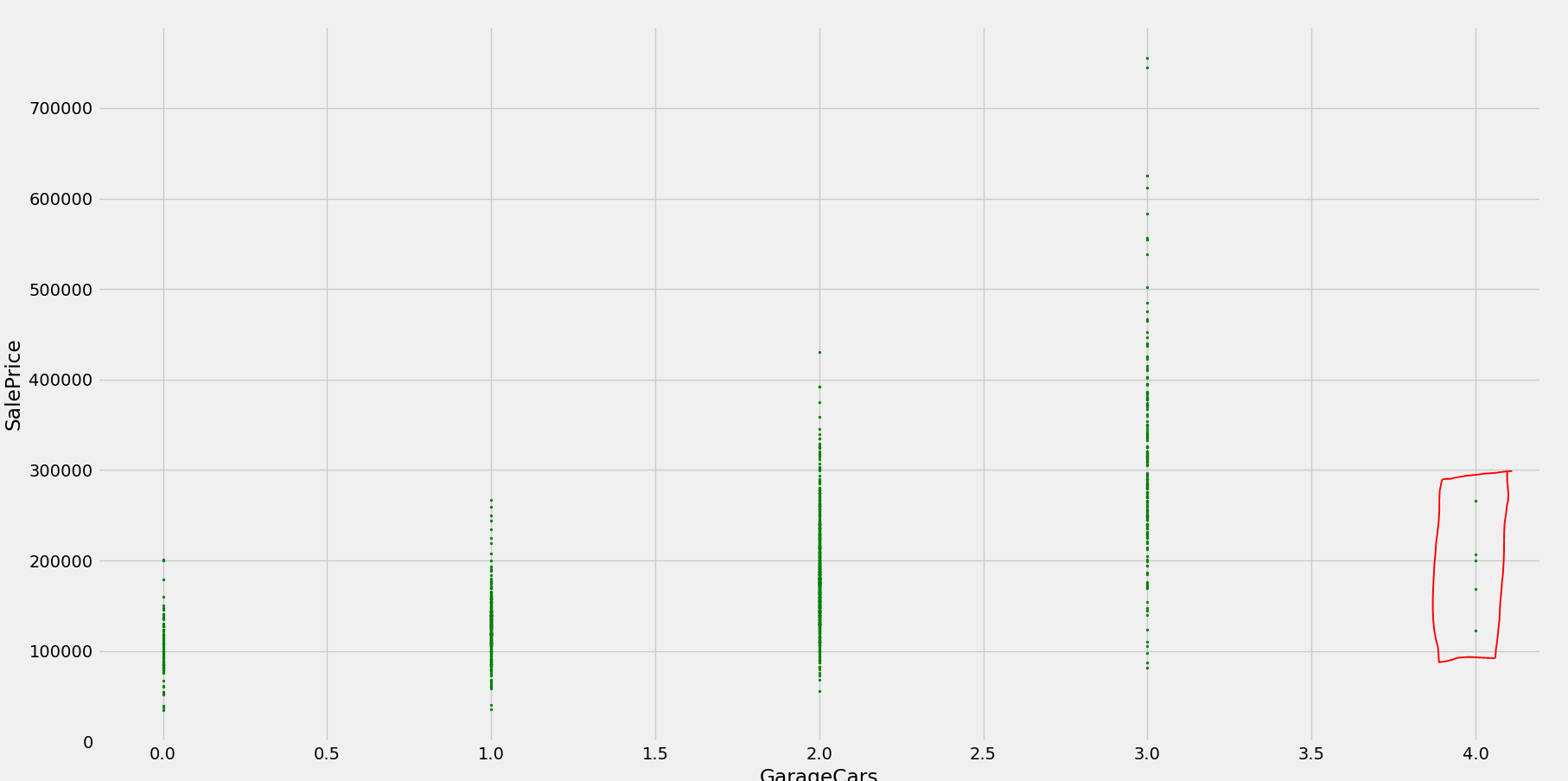
691 7.228819

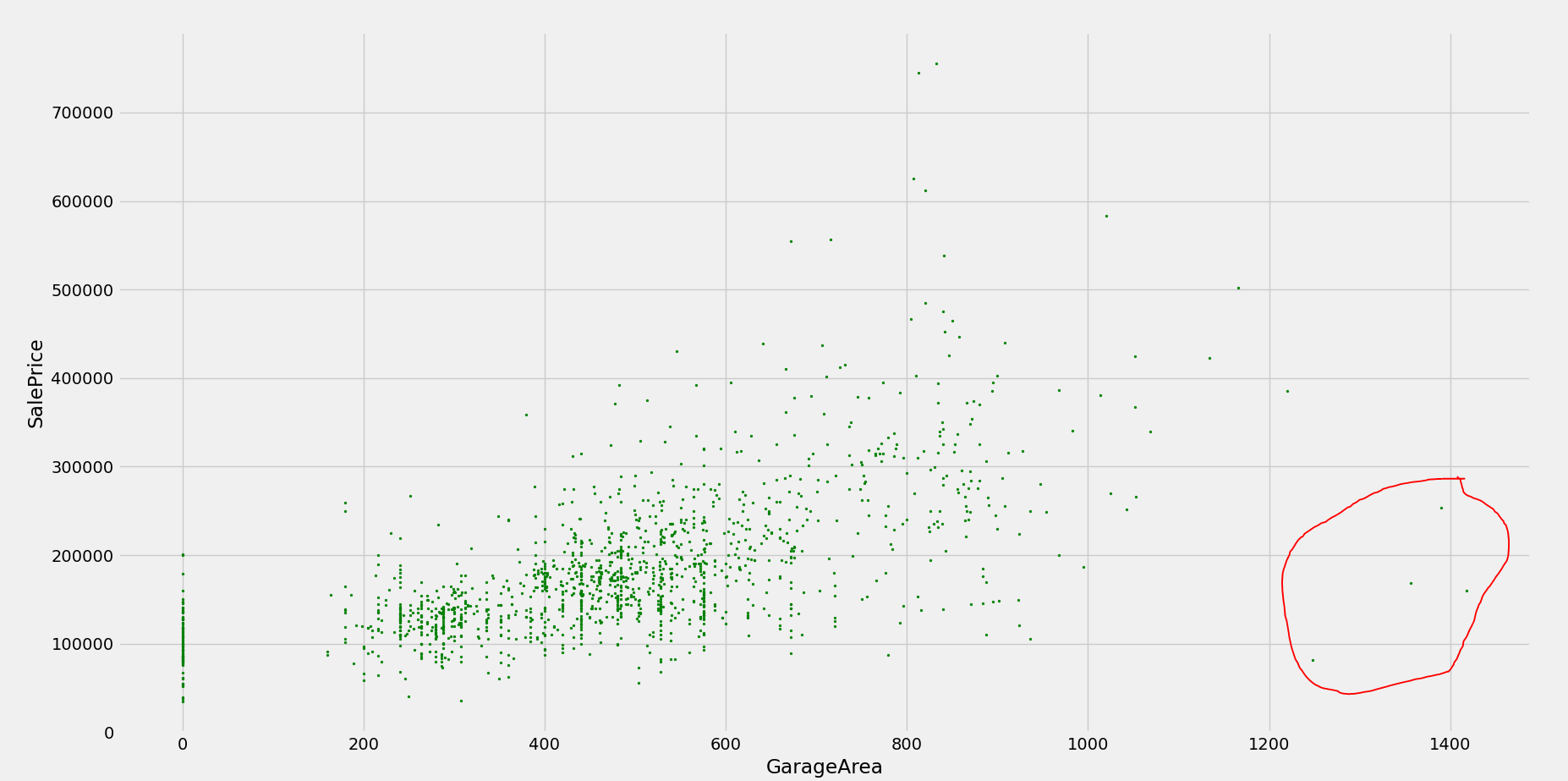
Bivariate analysis for outliers

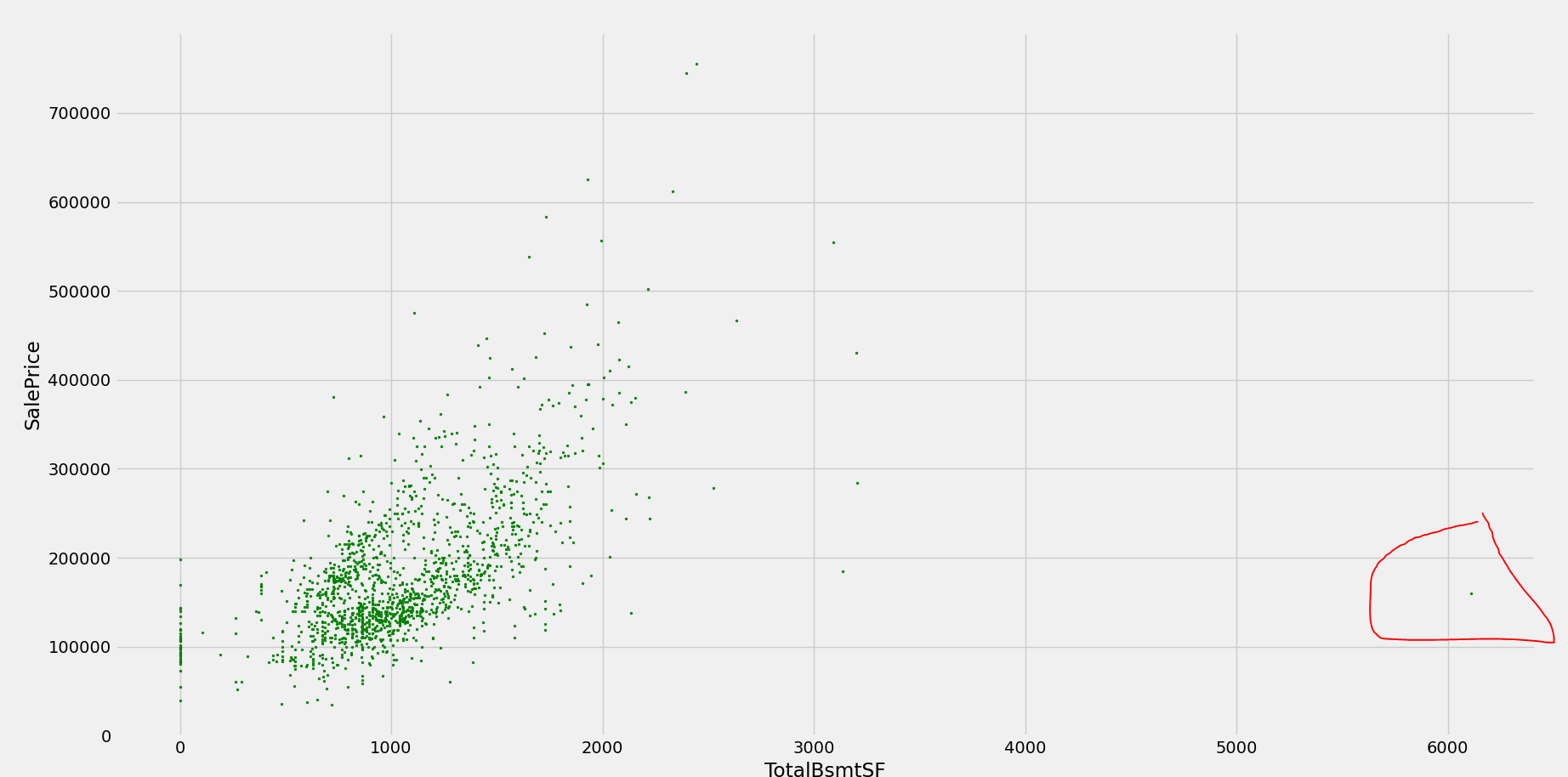


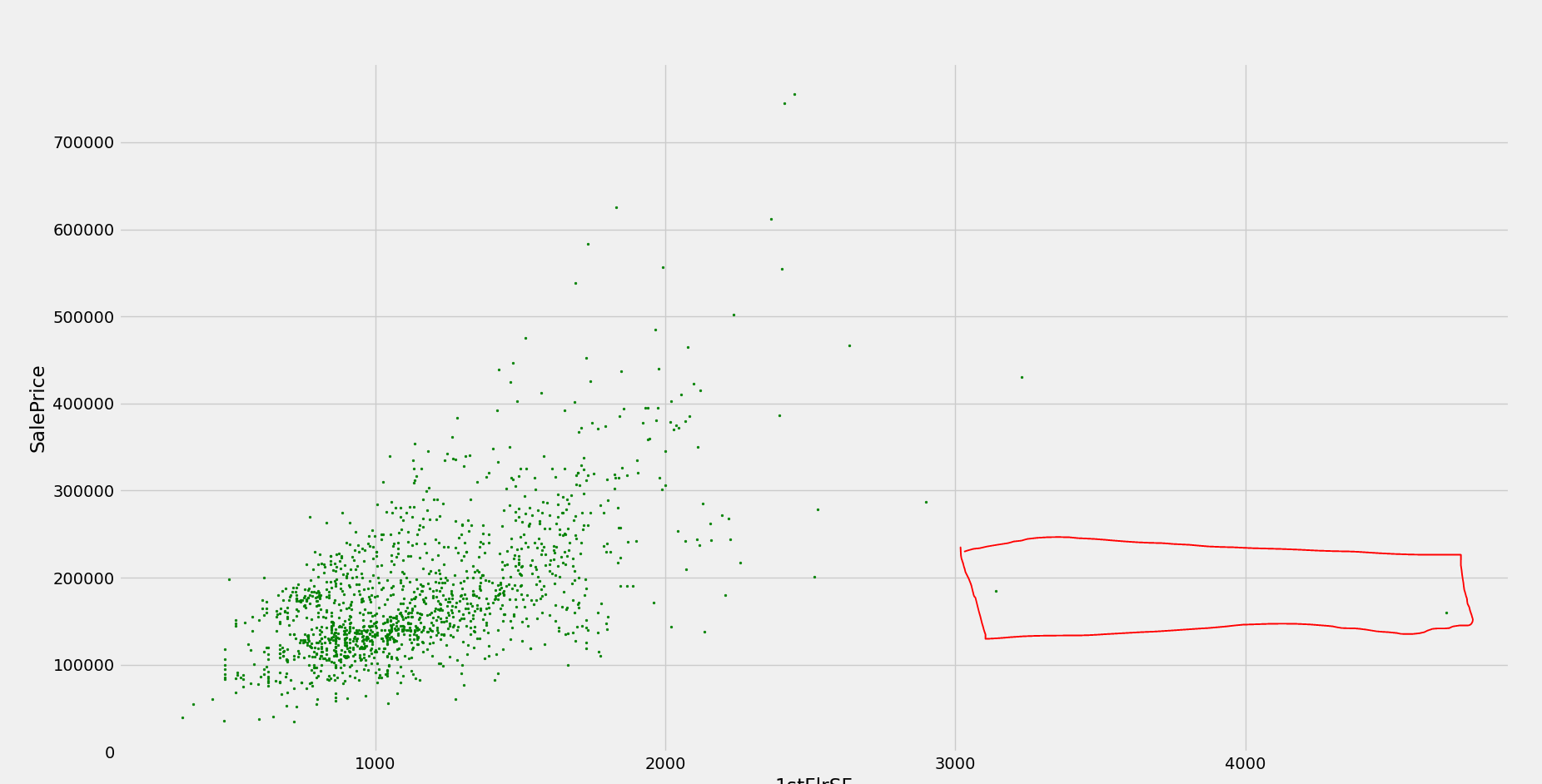
As the above ground living area (GrLivArea) increases the price of the house increase however when the GrLivArea get extreme the prices get lower. This can be due to the fact the plots with large area may agricultural land.

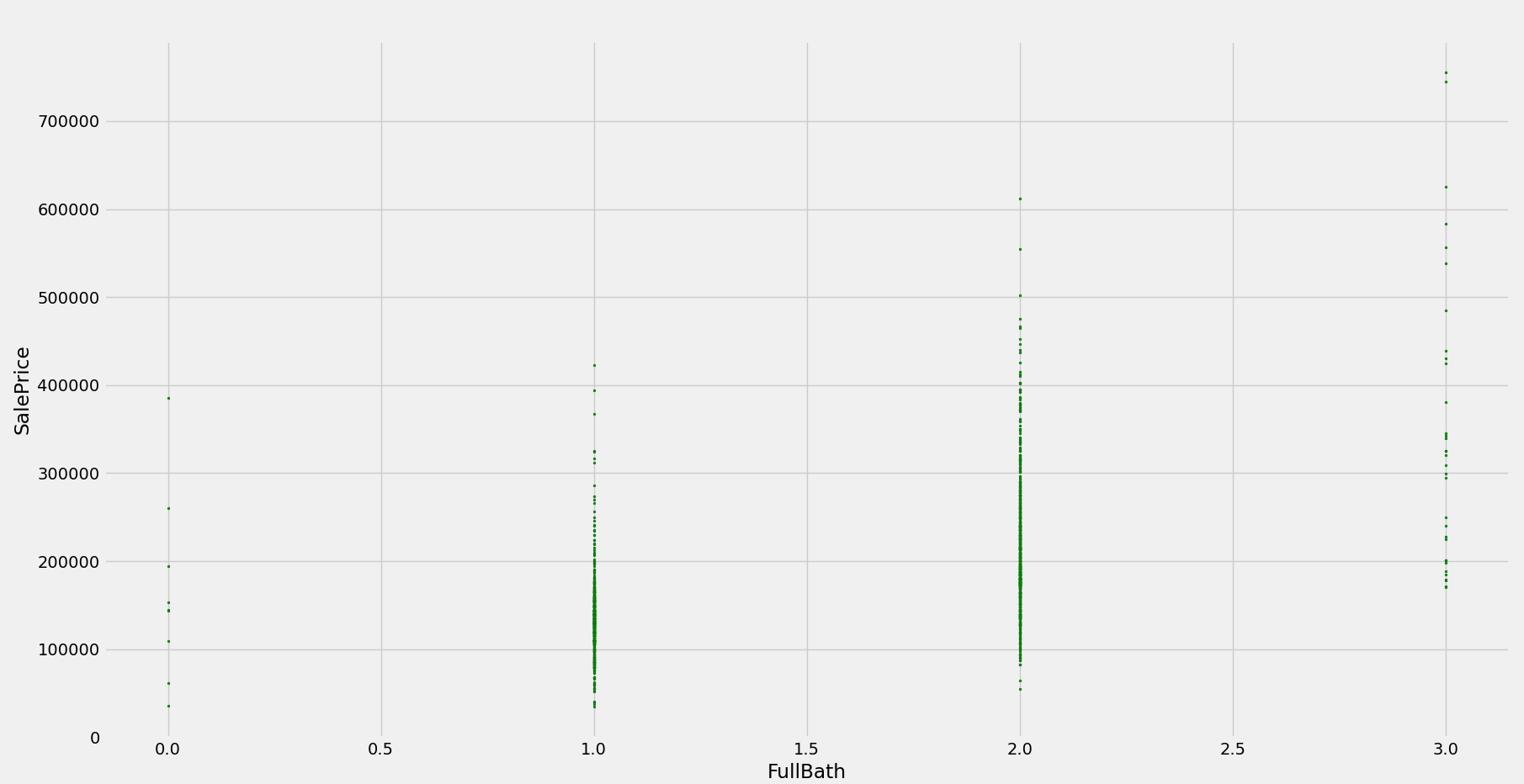
Remove the 2 observations on the top of the screen

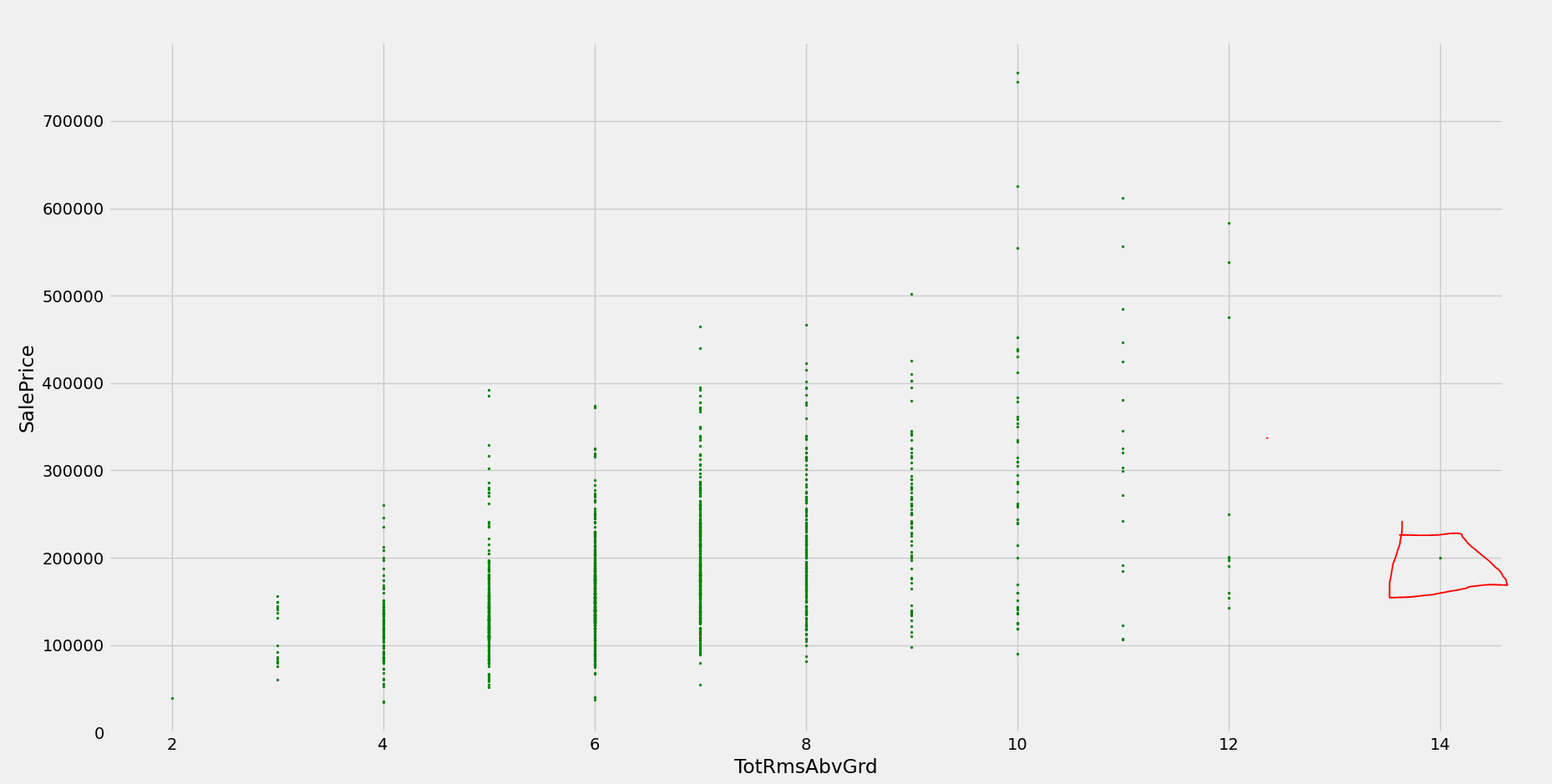


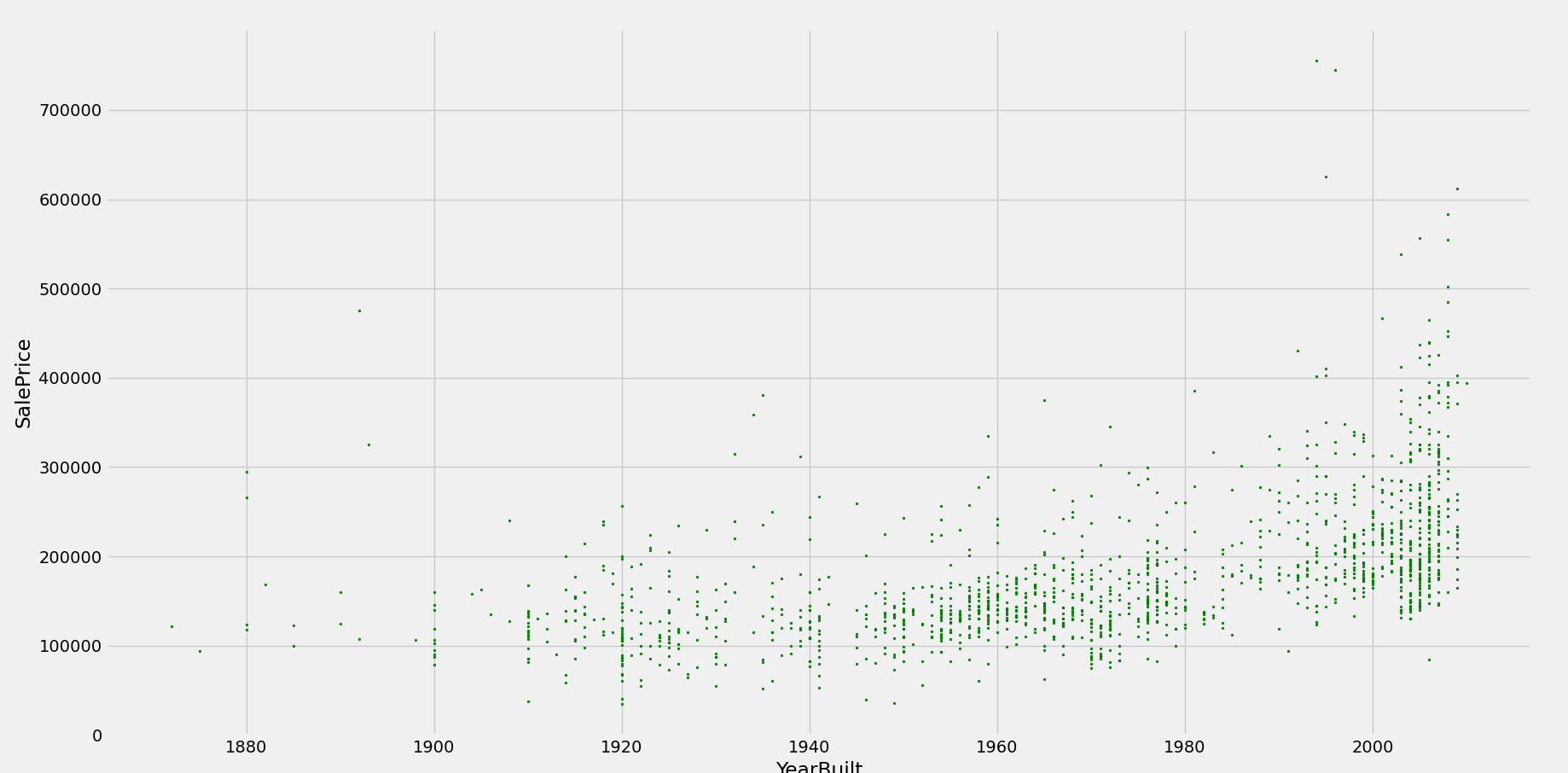








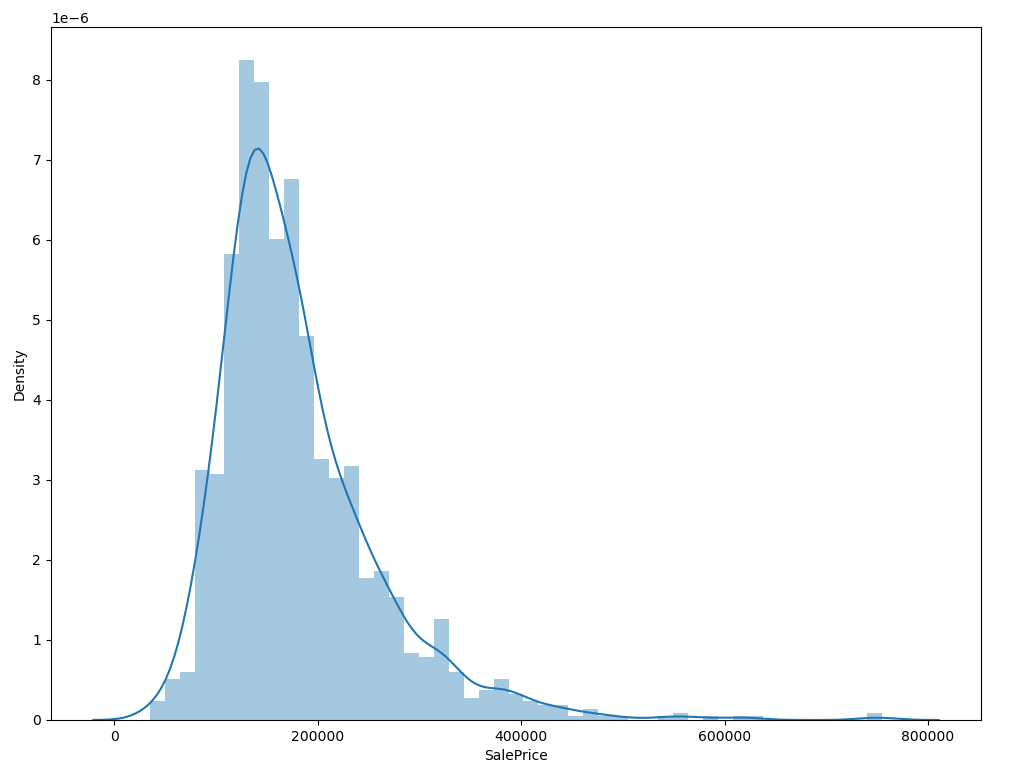


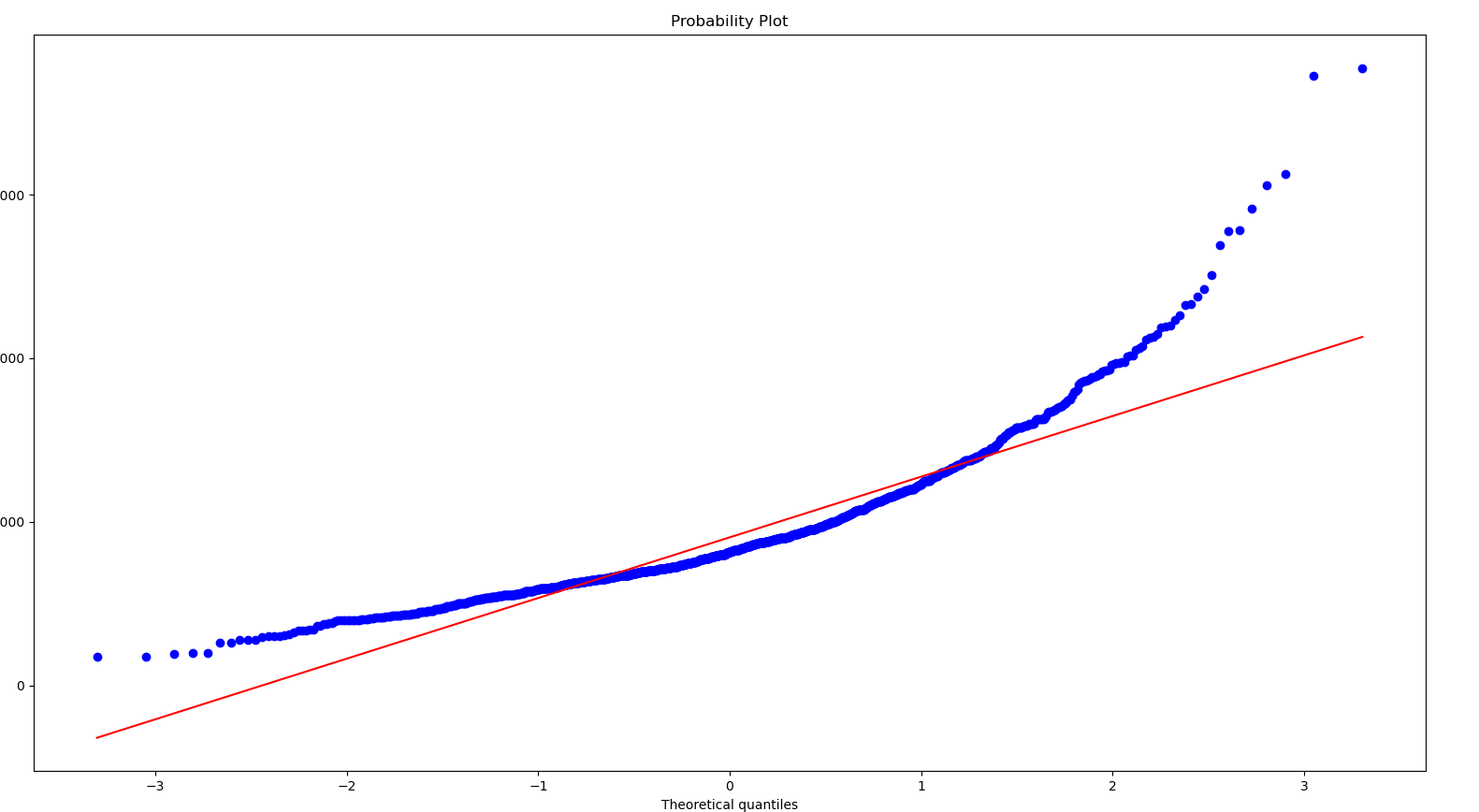


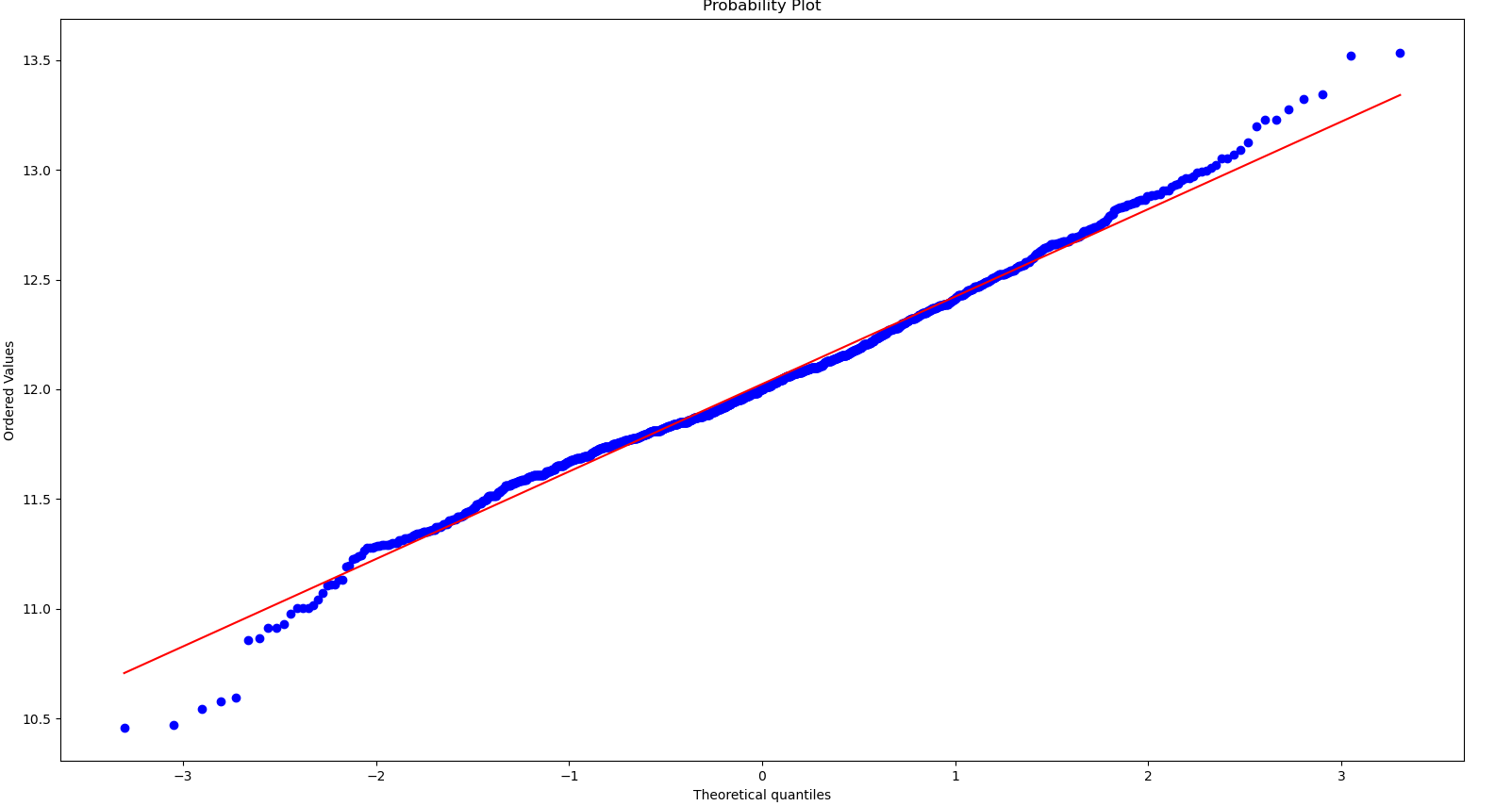
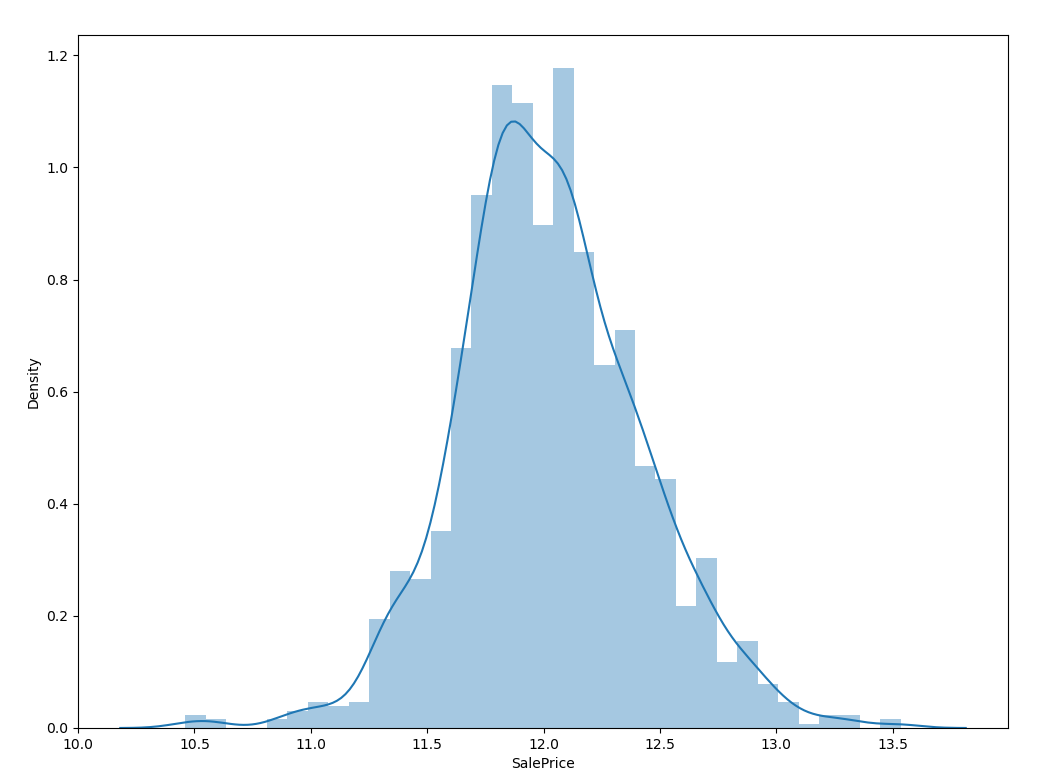
**Normality:**

The variable is not normally distributed (right skewed, peakedenss) the log transformation can be applied to normalize the data.

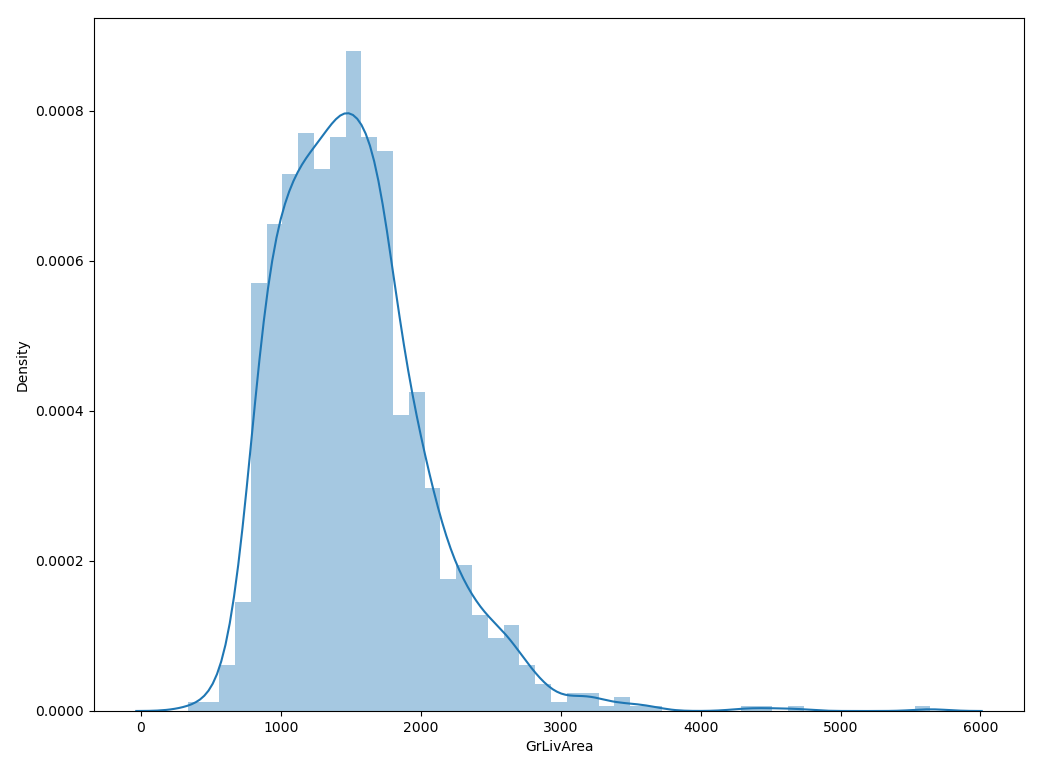
**SalePrice:**

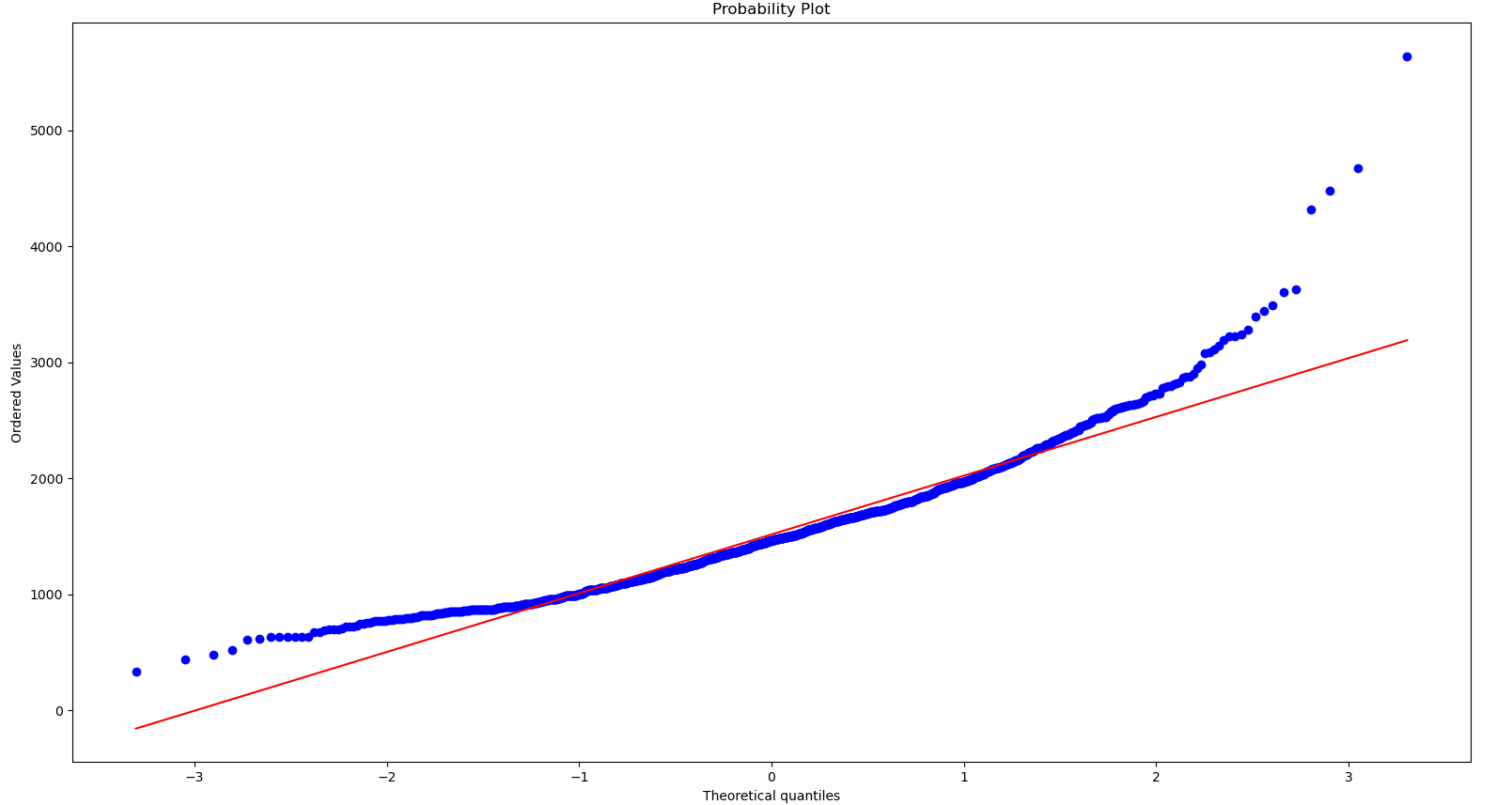




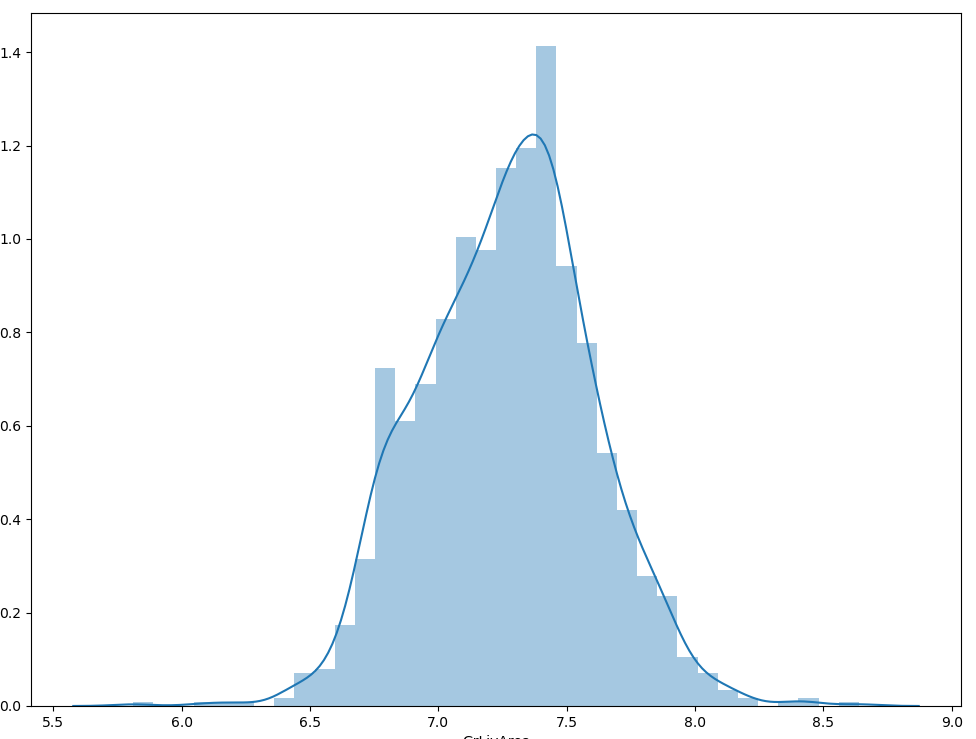
**After Log Transform**

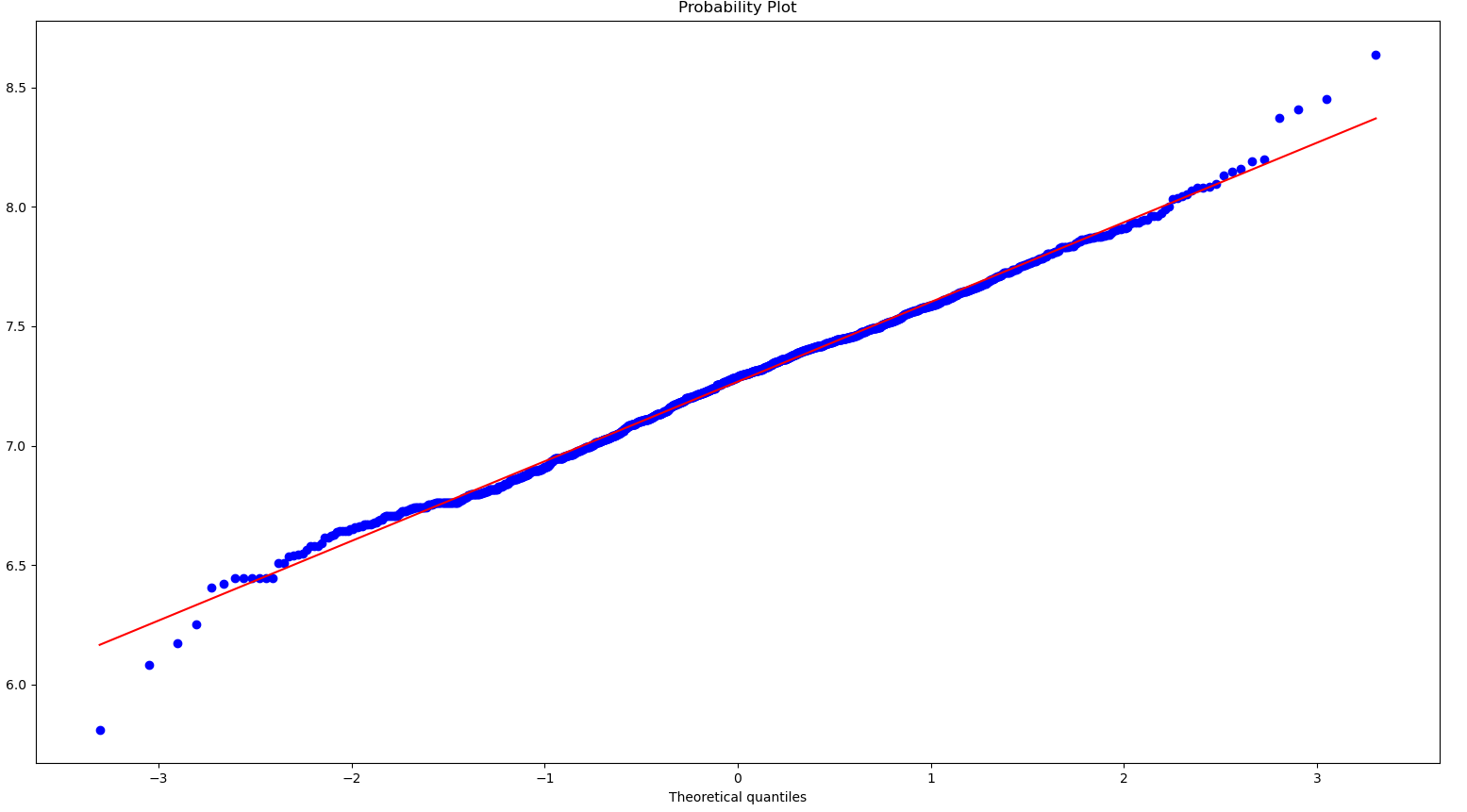
**GrLiveArea:**



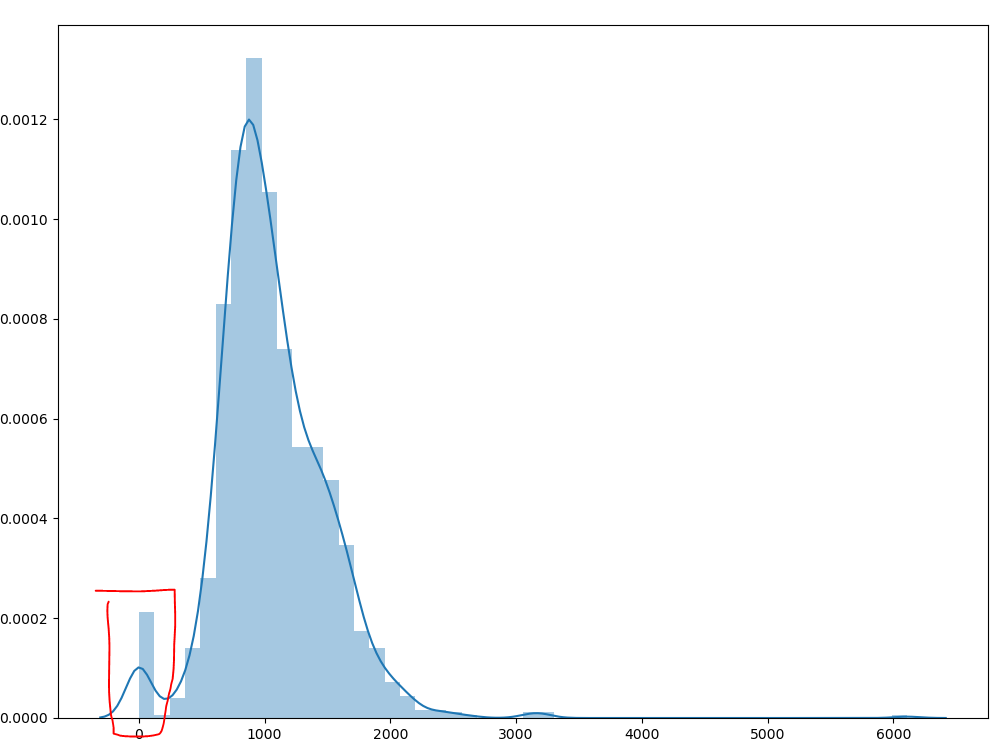


After Log Transform

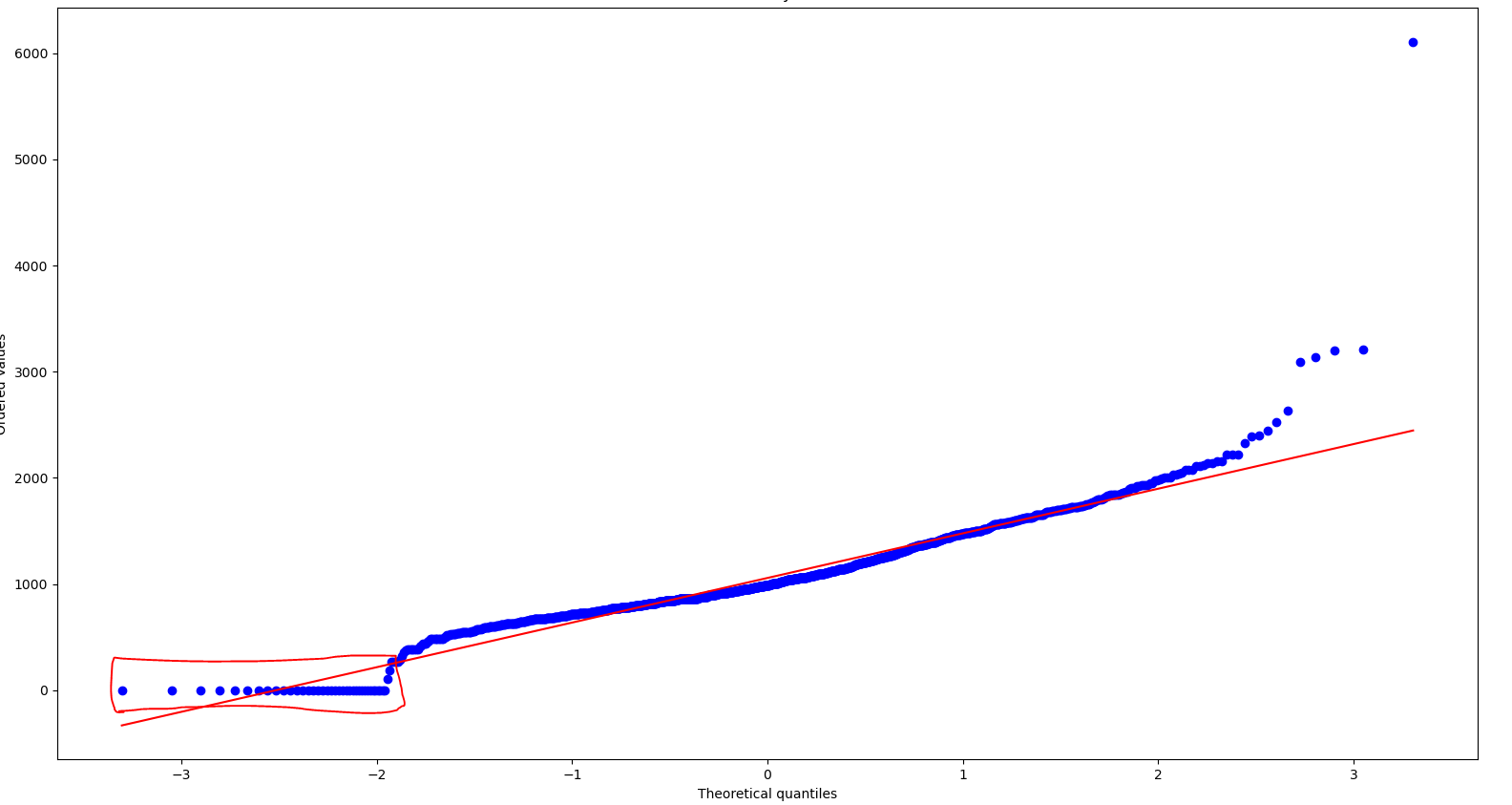




TotalBsmtSF

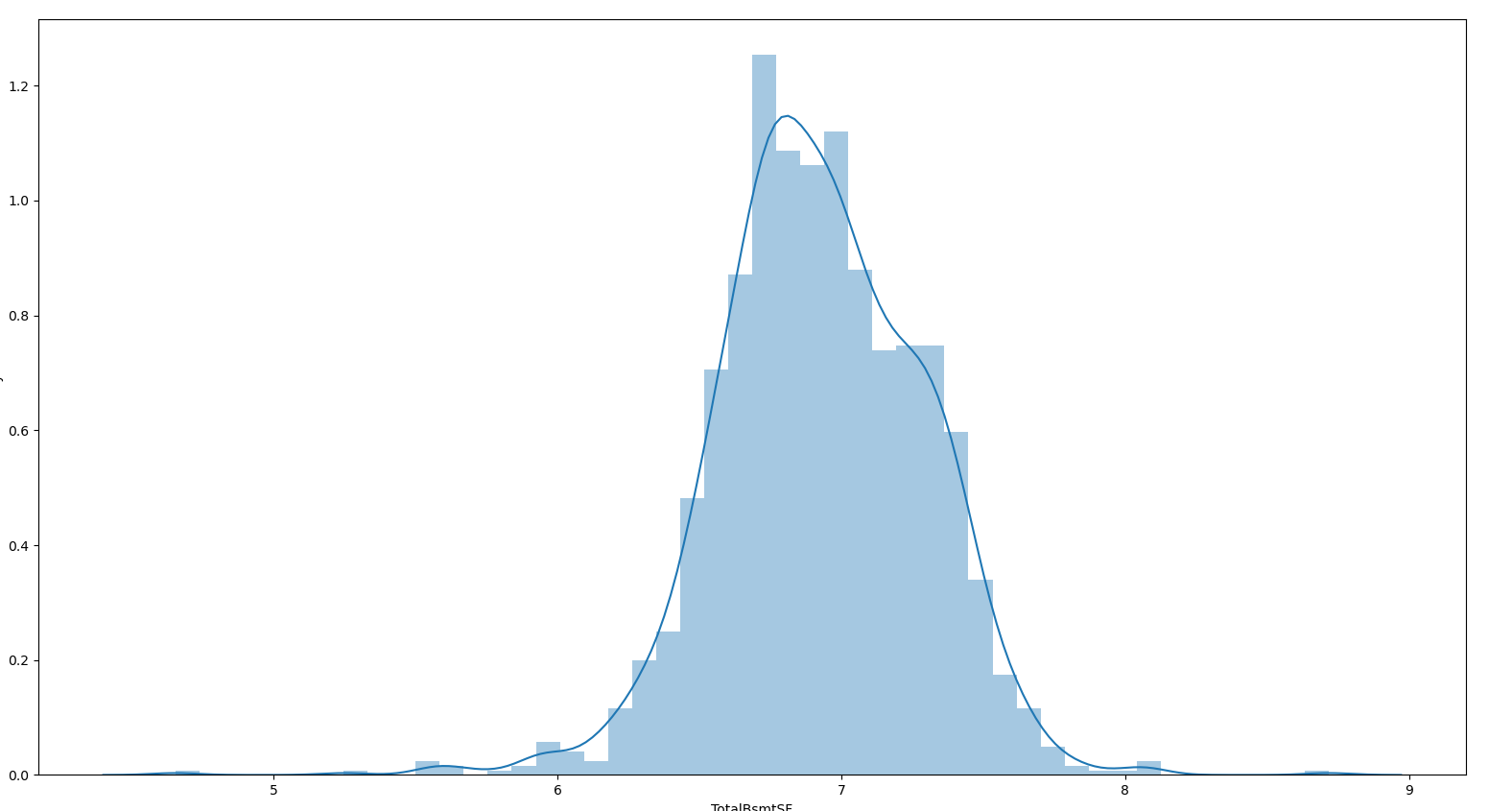


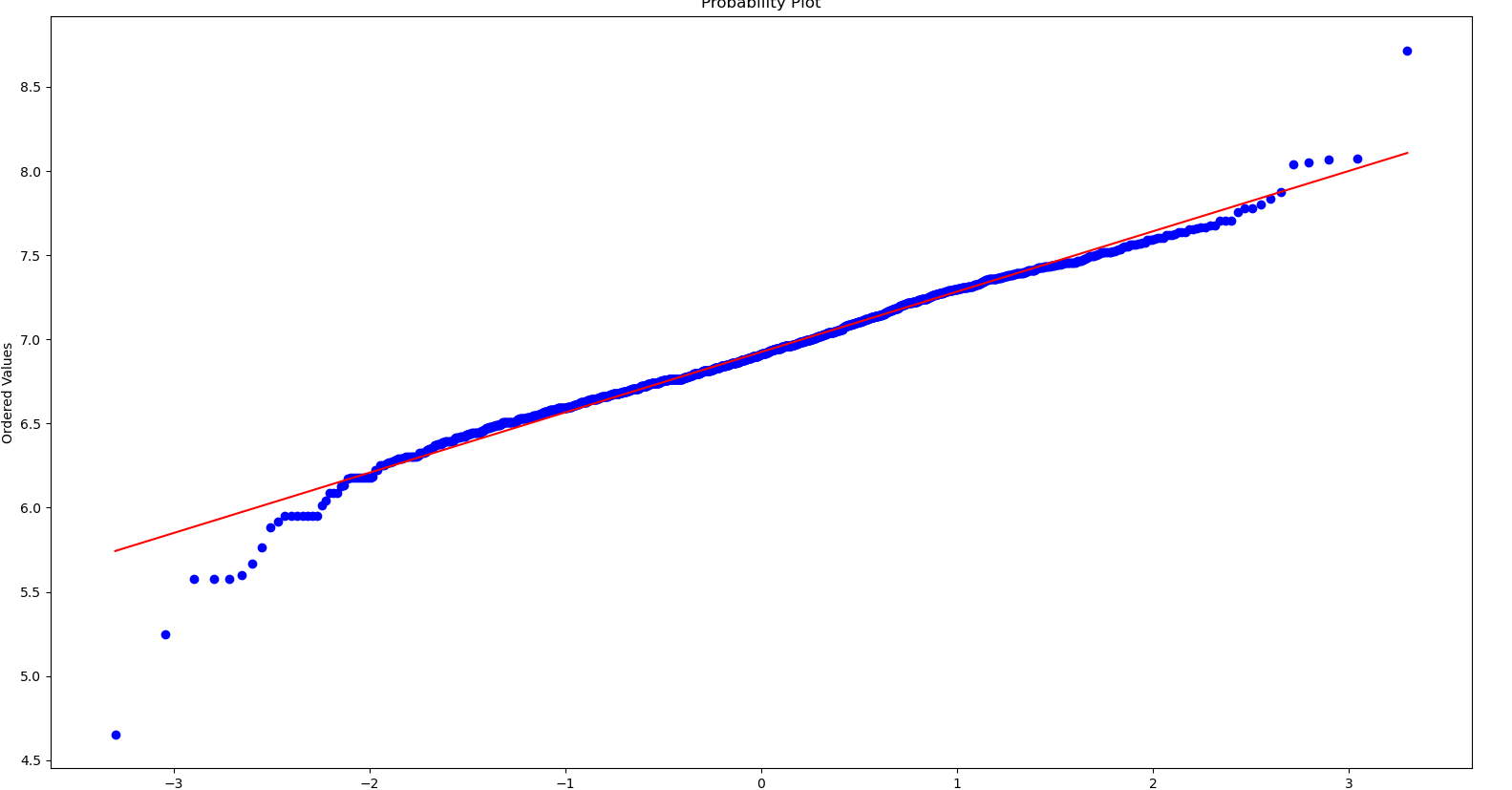
Shows skewdness and zero values



On the zero-values log transformation cannot be applied.

After Log transform of non-zero values





### Checking homoscedasticity