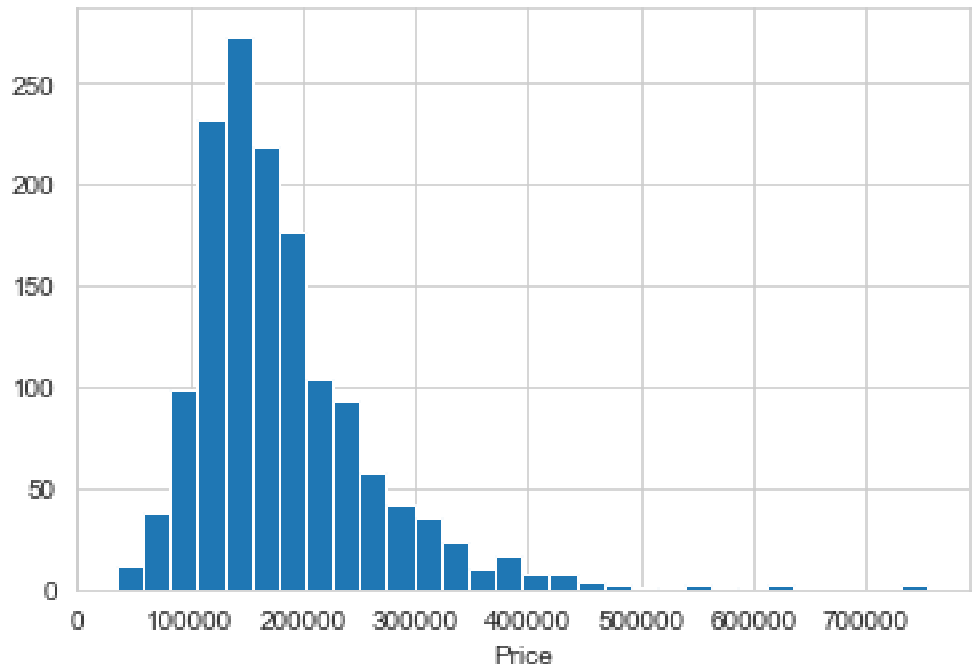
Housing Prices Exploratory Data Analysis- Ayooluwa Adedipe- Stutern Graduate Accelerator 07

1. Target Variable Univariate Analysis

The target variable to be predicted in this competition is the price at which houses may be sold.



Above is a Histogram representing the housing prices. At first glance, the distribution can be seen to be positively skewed with lots of data clustering towards the left side.

The measures of central tendency support this as the mean price ($180,921) is greater than the median ($163,000)

```
count     1460.000000
mean    180921.195890
std      79442.502883
min      34900.000000
25%     129975.000000
50%     163000.000000
75%     214000.000000
max     755000.000000
```

While there are a few outlier values, they are not so extreme compared to the average house price, and so this distribution is not so fat-tailed.

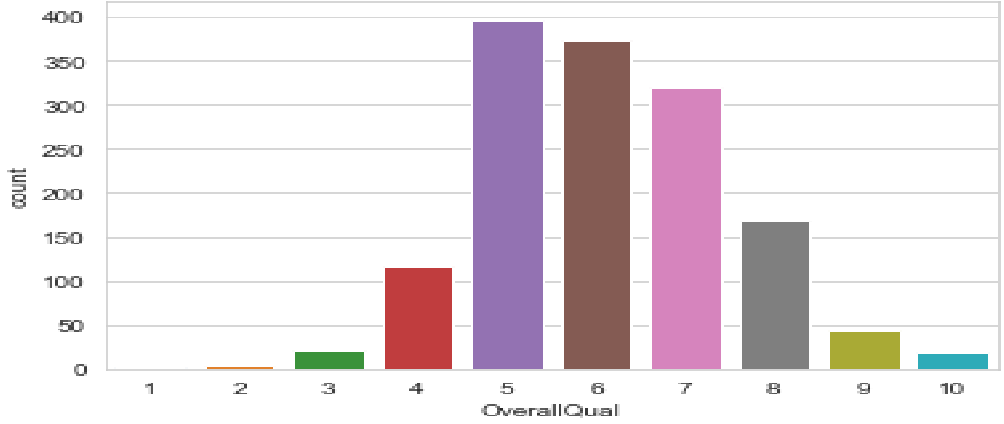
2. Important Feature Variables

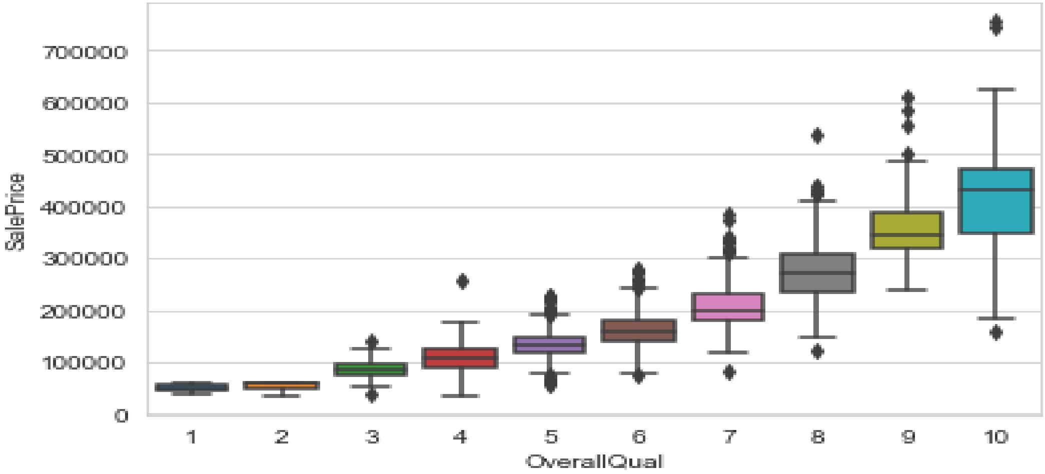
The correlation coefficient matrix for SalePrice is used to identify which feature variables most affect the target variable. Strongly correlated values (above 0.3 or less than 0.3) are considered

OpenPorchSF	0.315856
2ndFlrSF	0.319334
WoodDeckSF	0.324413
LotFrontage	0.351799
BsmtFinSF1	0.386420
Fireplaces	0.466929
MasVnrArea	0.477493
GarageYrBlt	0.486362
YearRemodAdd	0.507101
YearBuilt	0.522897
TotRmsAbvGrd	0.533723
FullBath	0.560664
1stFlrSF	0.605852
TotalBsmtSF	0.613581
GarageArea	0.623431
GarageCars	0.640409
GrLivArea	0.708624
OverallQual	0.790982
SalePrice	1.000000

- OverallQuality

The feature variable with the highest correlation to SalePrice is the overall quality of the houses with the barplot below

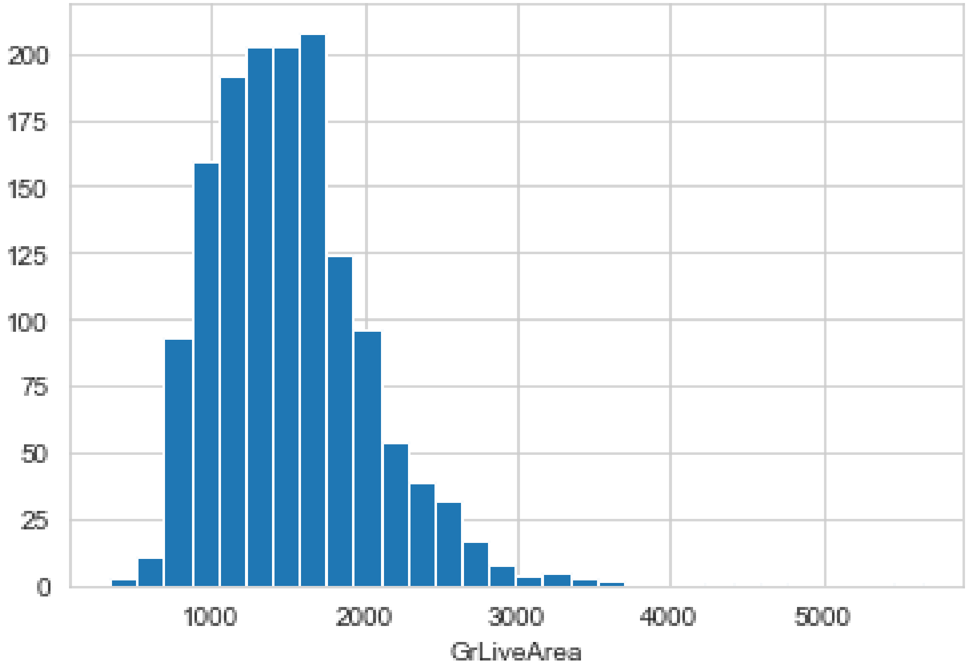




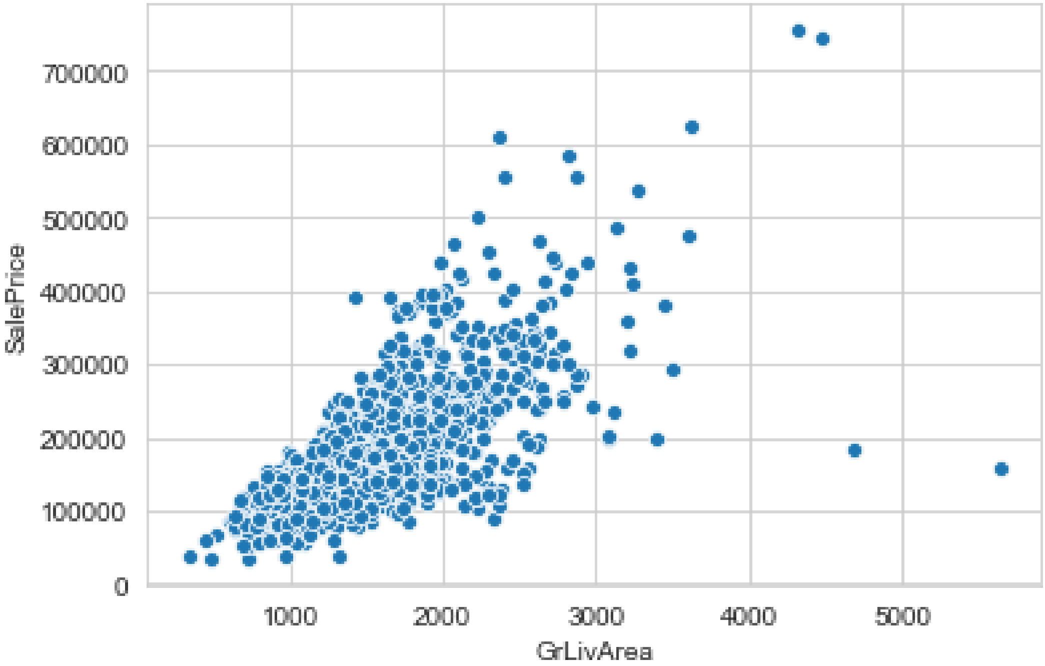
From the bar chart, it can be seen that most of the houses in the dataset are of average quality and as expected, the boxplot shows that price rises with increase in quality.

- GrLivArea

The next important feature variable is the GrLivArea, which is the Above Grade Living Area in square feet

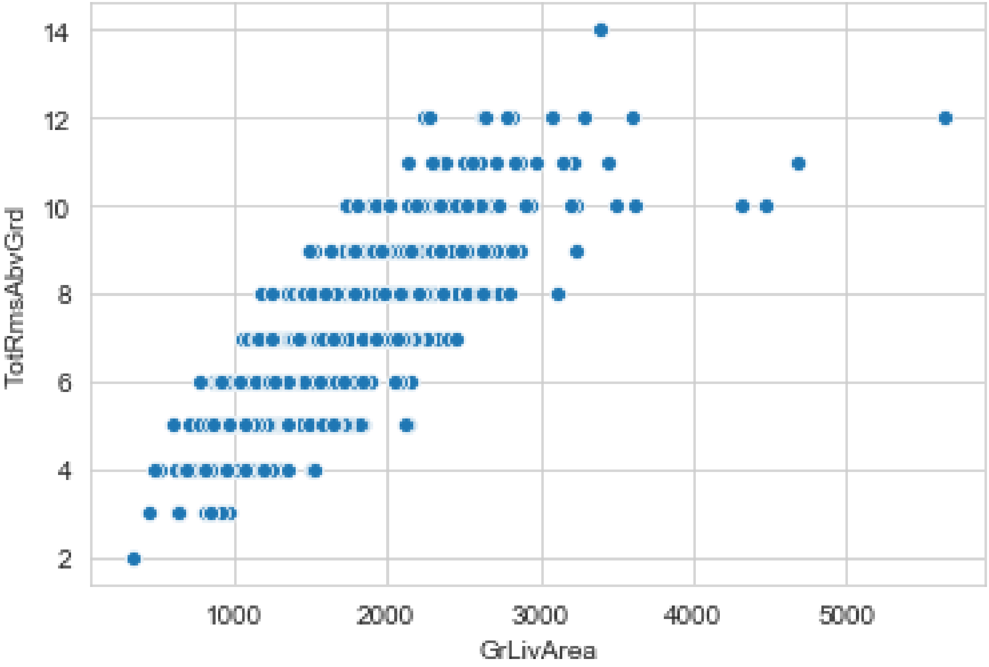


Just like Sales price, GrLivArea is positively skewed as well, and its graphical relationship with Sales Price is shown below:



The general trend shows increase in Sales price as Above Grade Living Area increases, but there are two observations that have high GrLivArea values but low SalePrice. These anomalous values may be due to an error in data collection or perhaps a promo sale.

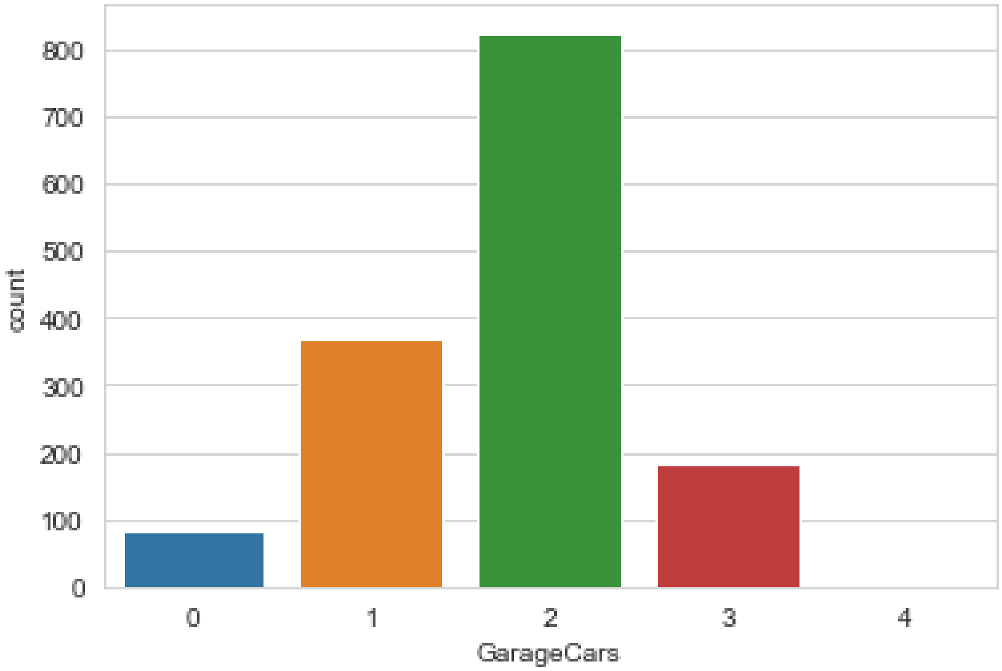
The correlation matrix also shows that GrLivArea is highly correlated with the variable, TotRmsAbvGrd, which is the total number of rooms above ground.



And that makes sense since the living area gets bigger if there are more rooms, so we can ignore one of the variables. We can choose to keep GrLivArea since it has the higher correlation with Sale Price.

- GarageCars

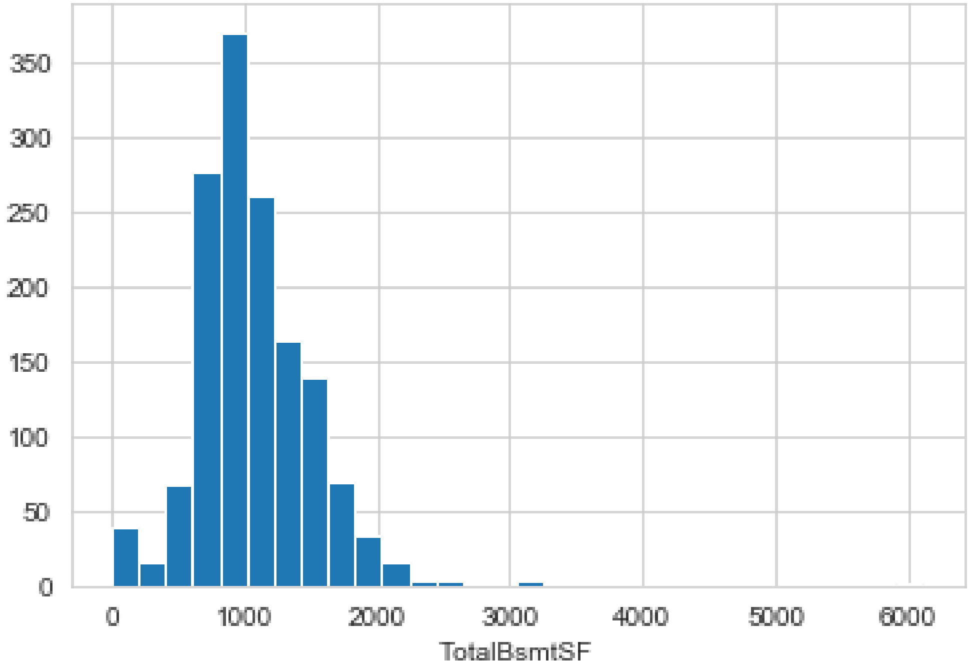
The next correlated variable is Garage Cars, which is number of cars that the Garage can hold



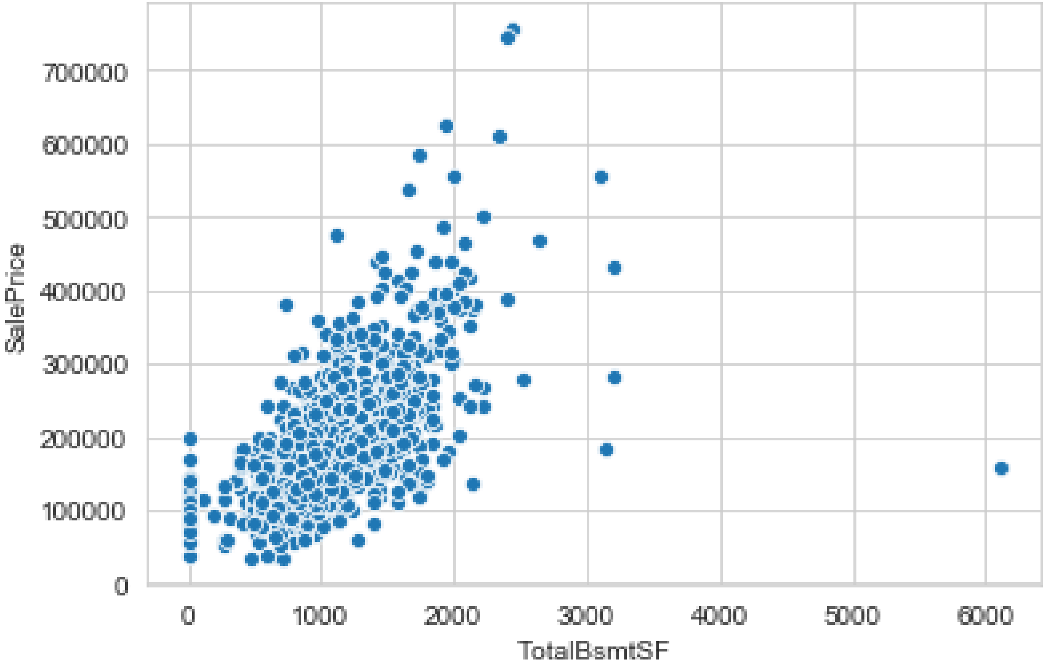
Again, since Garage Cars and Garage Area are so highly correlated (0.88 correlation coefficient), and that's because, the more the number of cars, the bigger the garage will be. Therefore, we can drop Garage Area and keep Garage Cars alone.

- Total BsmtSF

This refers to the total basement area in square feet.

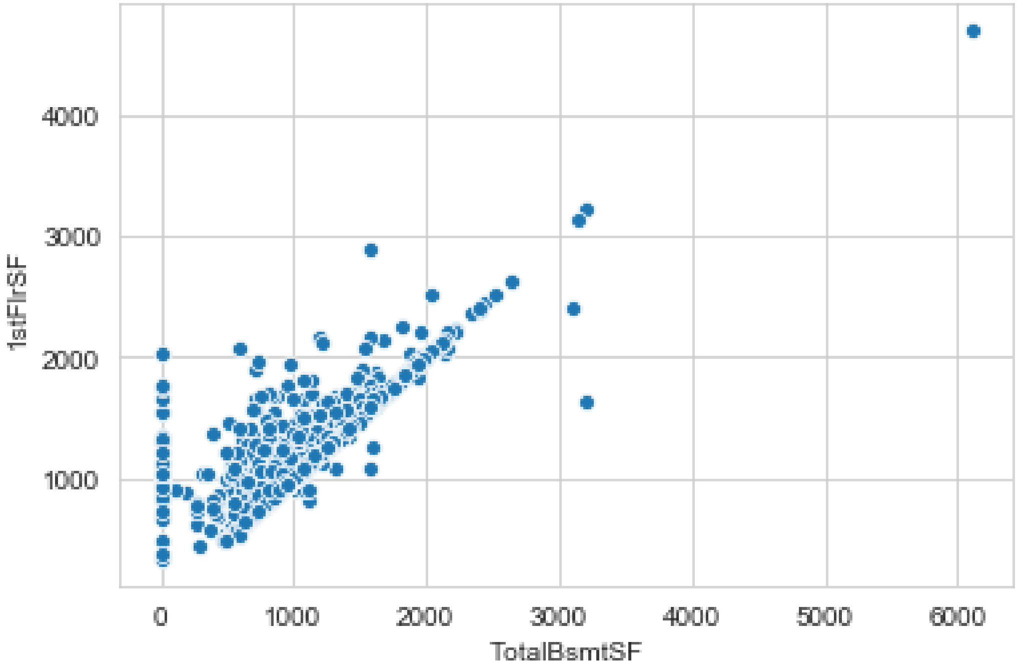


The histogram of this distribution shows a negative skewness also.



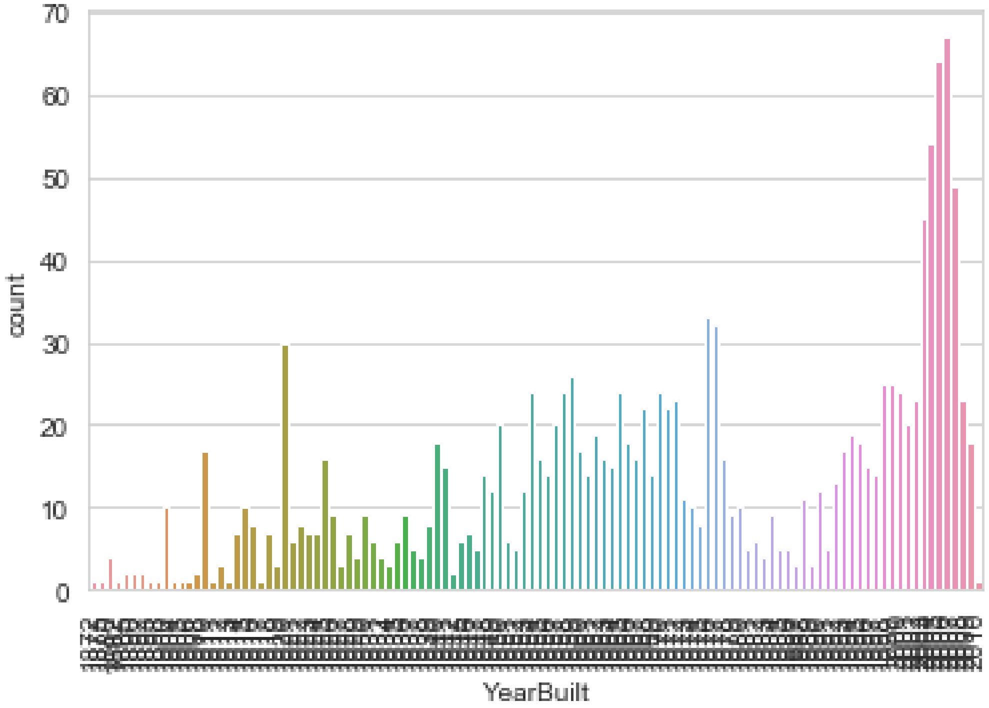
We observe an outlier that defies the regular trend with a Basement area of more than 6,000 square feet but a low price of less than 200,000.

Also, this variable is highly correlated with 1stFlrSF (coefficient of 0.819). The scatterplot even shows an almost perfectly linear relationship for houses where basementSF is above zero. So we can drop 1stFlrSF as well. (The observations where TotalBsmtSF has zero values are most likely for houses that don’t have a basement, just the living area)

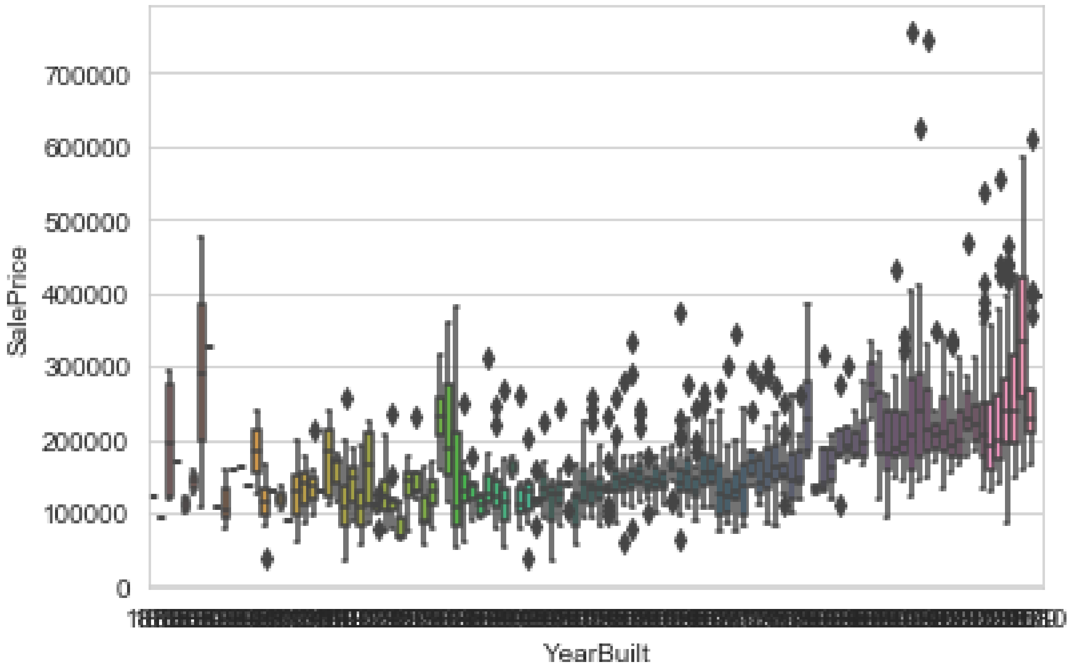


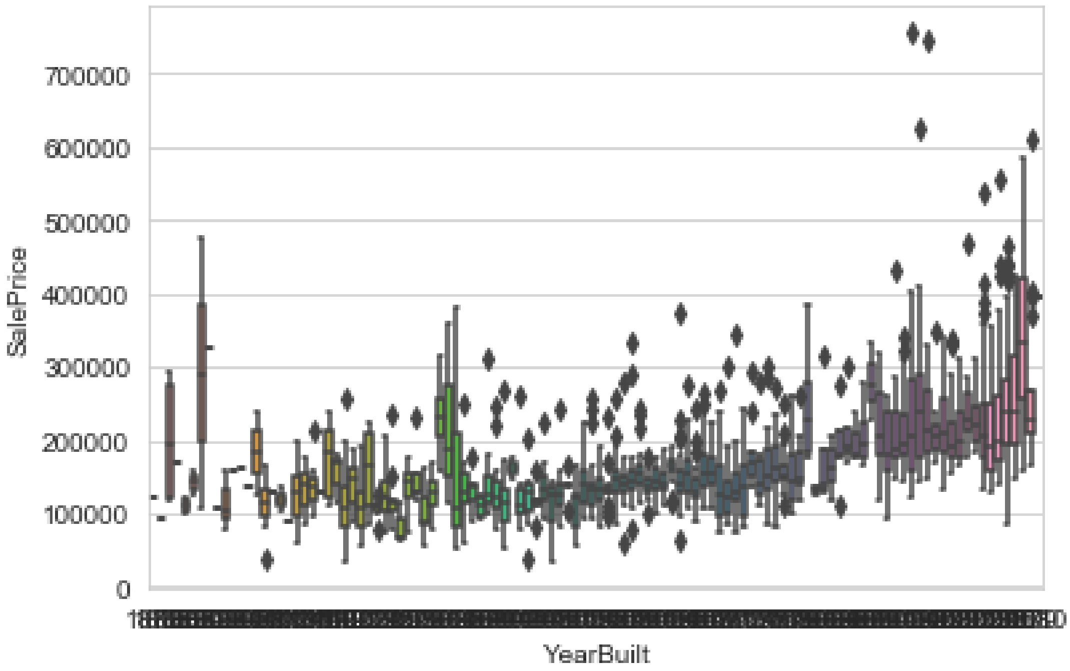
- YearBuilt

Another feature variable that correlates well with SalePrice is the Year in which the house was built, and its distribution is shown in the barchart below



This plot shows that most of the houses in this dataset were built recently. For its relationship with Sales price, we use a boxplot.


700000
600000
500000
400000
300000
200000
100000
0
SalePrice
YearBuilt


700000
600000
500000
400000
300000
200000
100000
0
SalePrice
YearBuilt

Missing Values

Below are the variables with missing data, and the number of observations that are missing

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