Machine Learning Project

Ames House Prices Prediction

Yiling Lin

Project Information

Dataset:

Ames Housing dataset

79 explanatory variables describing every aspect of residential homes in Ames, Iowa

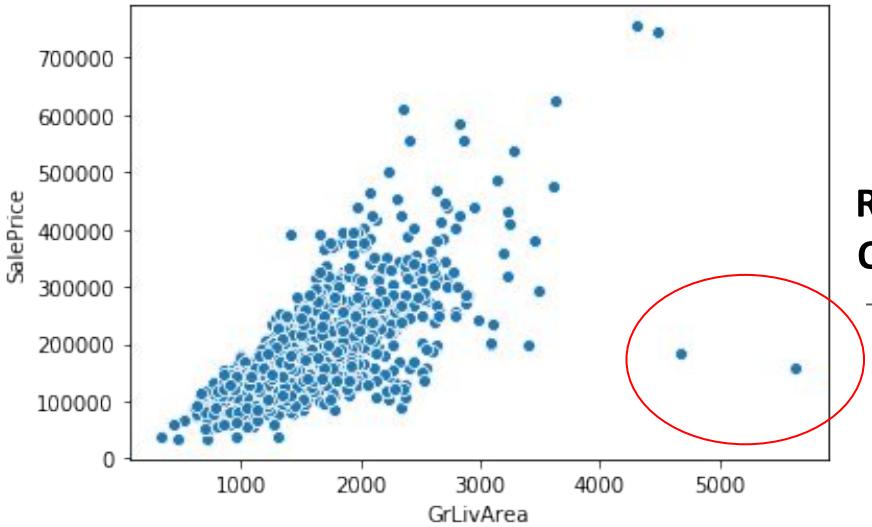
Part 1: Exploratory Data Analysis:

Does Remodeling Increase The Profit of Your Property Sale?

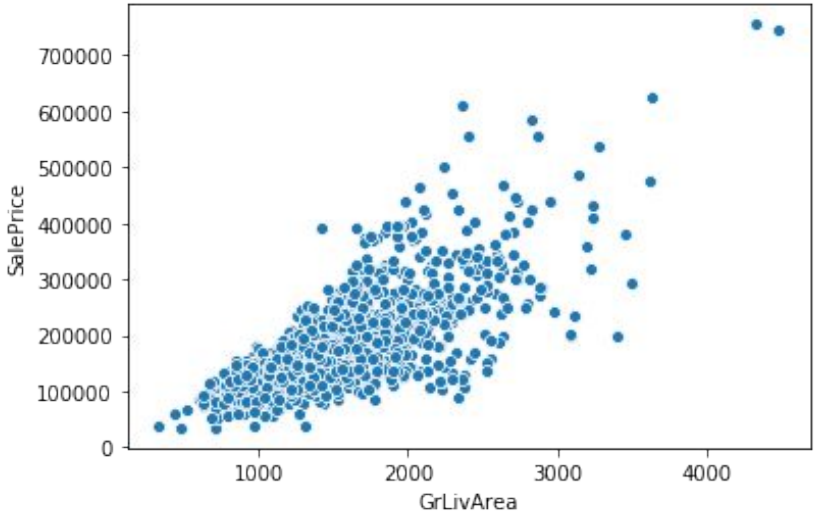
Part 2: Machine Learning:

Predict Sale Prices of houses in Ames, Iowa

Pre-Processing



Remove Outliers

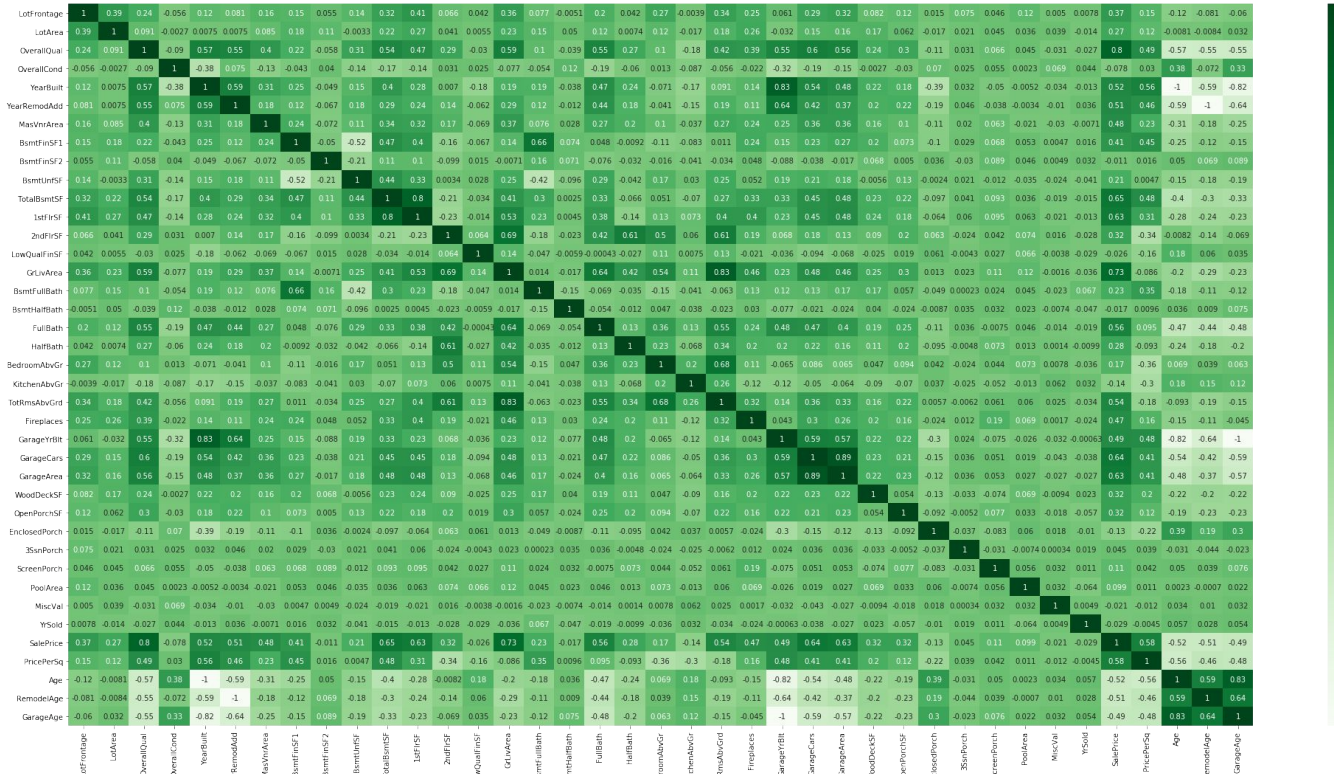


Feature Engineering

Columns Added

- **Age** = YrSold - YrBuilt
- **Remodel Age** - YrSold - YearRemodAdd
- **Remodeled** - if the house is remodeled
- **HasGarage** - if the house has a garage
- **GarageAge** = YrSold - GarageYrBlt
- **PricePerSq** = Price Per Square Foot
  (Sale Prices/Total Ground Living Area)

Correlation Matrix



Top 5 correlation coefficient:

Overall Quality vs Sale Price: 0.8

GrLivArea vs Sale Price: 0.73

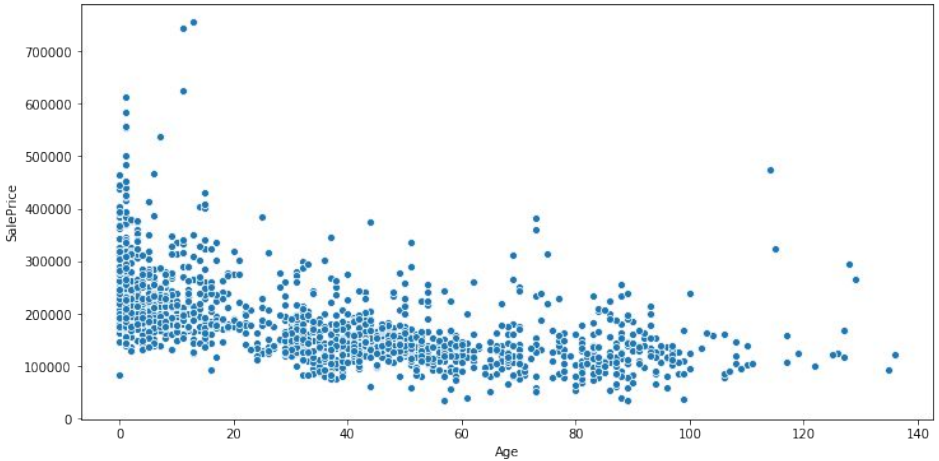
TotalBsmtSF vs Sale Price: 0.65

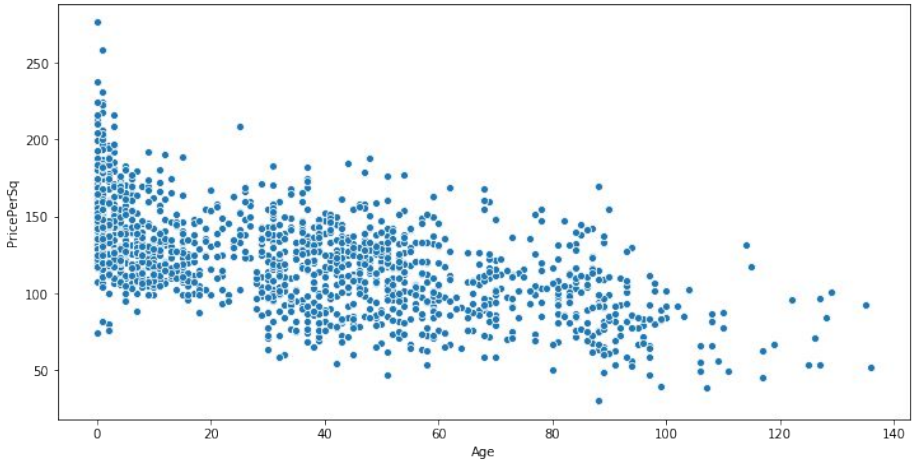
GarageCars vs Sale Price:0.64

GarageArea vs Sale Price:0.62

House Age, Remodel, Price

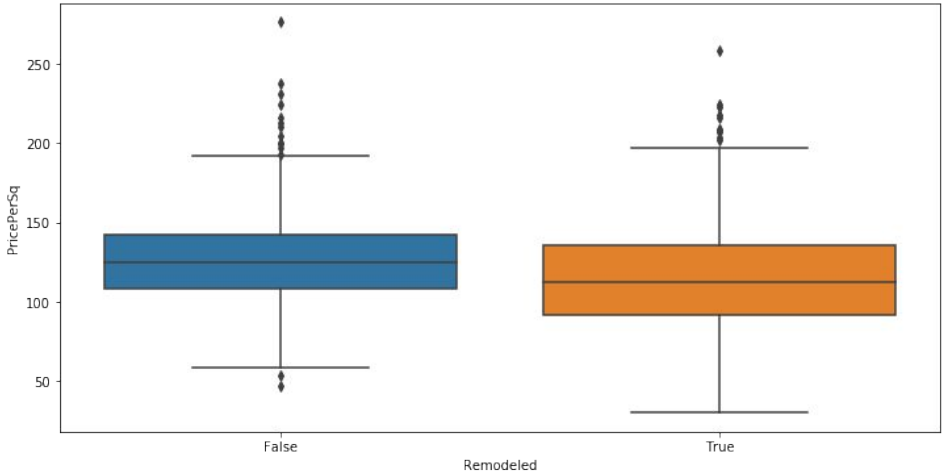
House Age vs Price

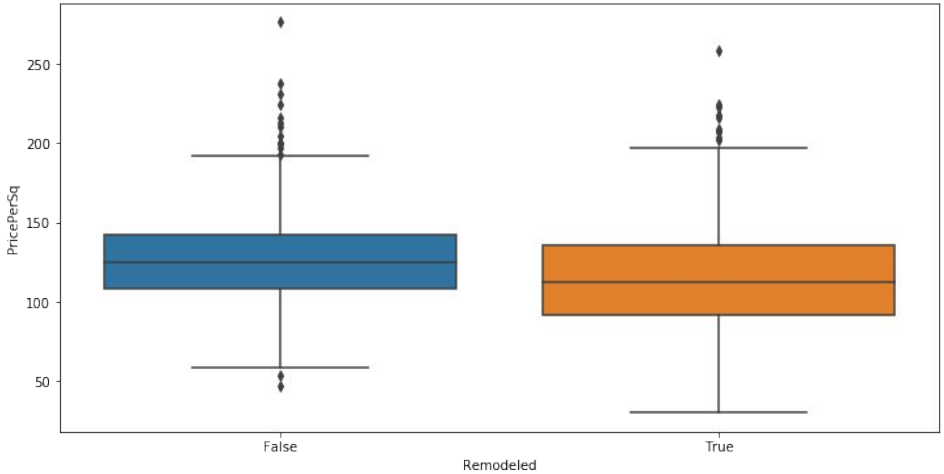




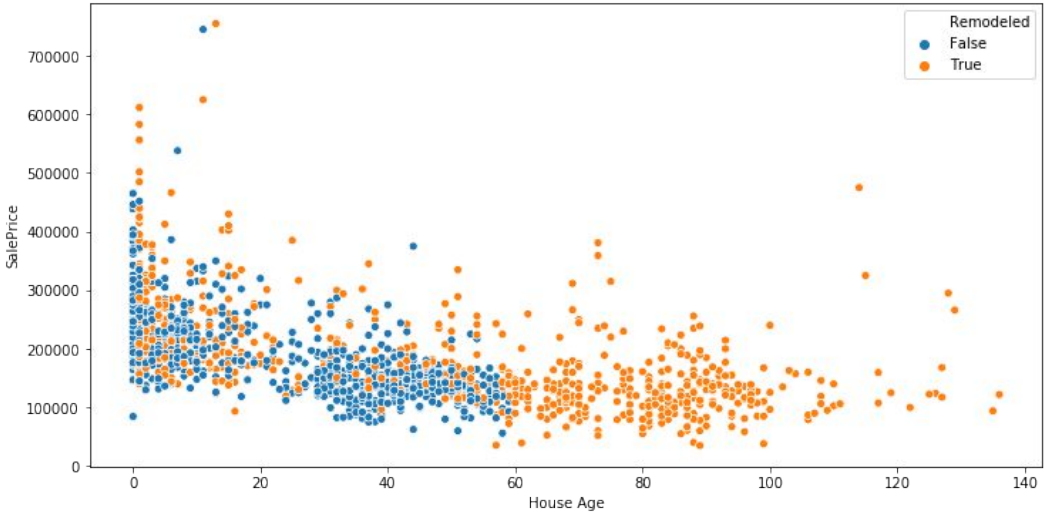
The older, the cheaper

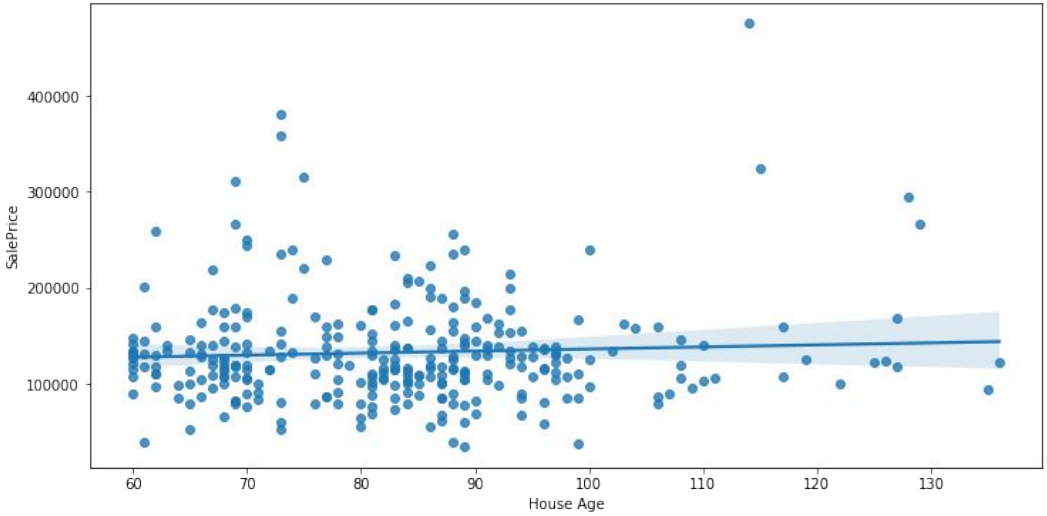
Remodeled Houses vs Sale Prices & Price Per Sqft





It seems that prices of remodeled houses are lower than non-remodeled houses? Why?

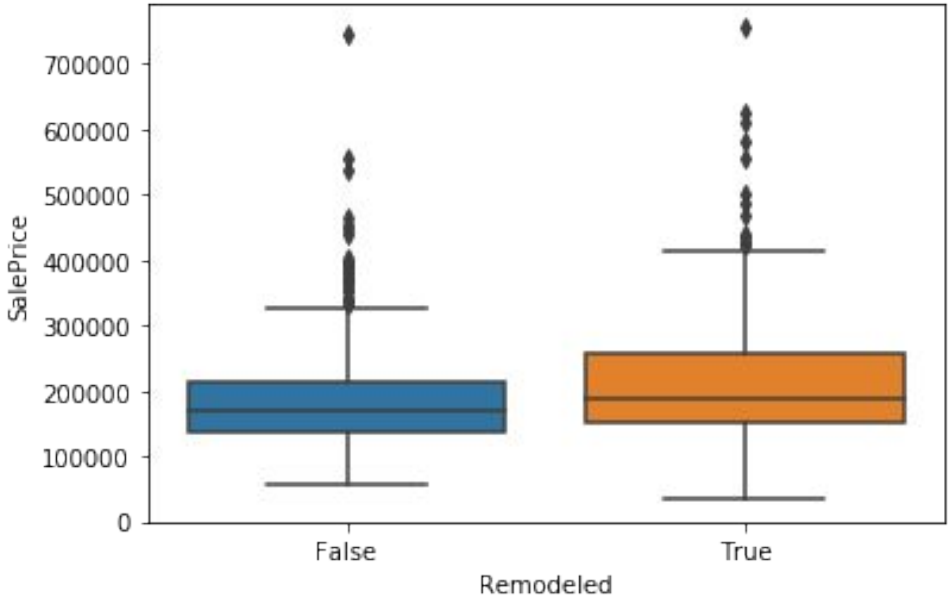


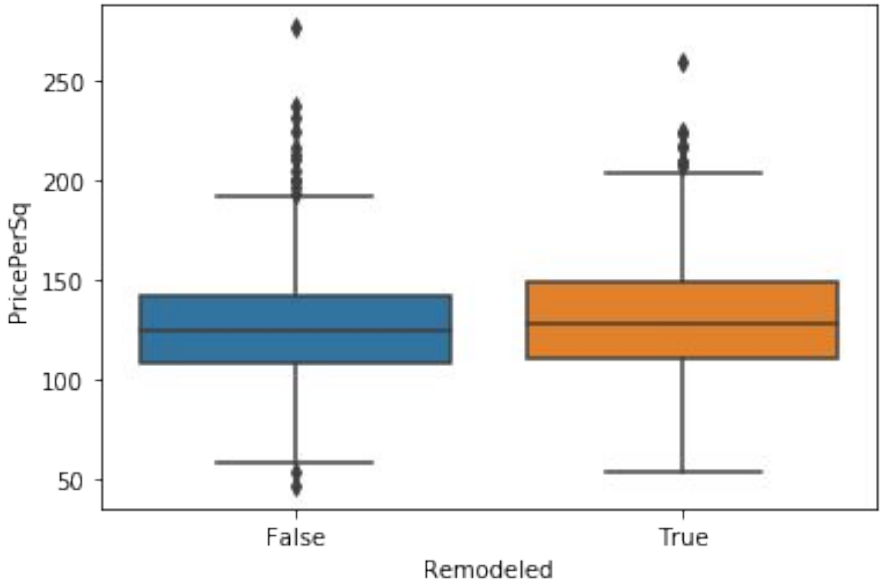


House Age, Remodel & Sale Price

- Around 44% of remodeled houses in the dataset are houses over 59 years old.
- Houses over 59 years old have no significant price changes.

Remodeled Houses vs Sale Prices & Price Per Sqft
0-59 Year-Old House





Averagely, when the age of a house is below 60, the price of a remodeled house is higher than a non-remodeled one.

House Age vs Avg. Price Per Sqft

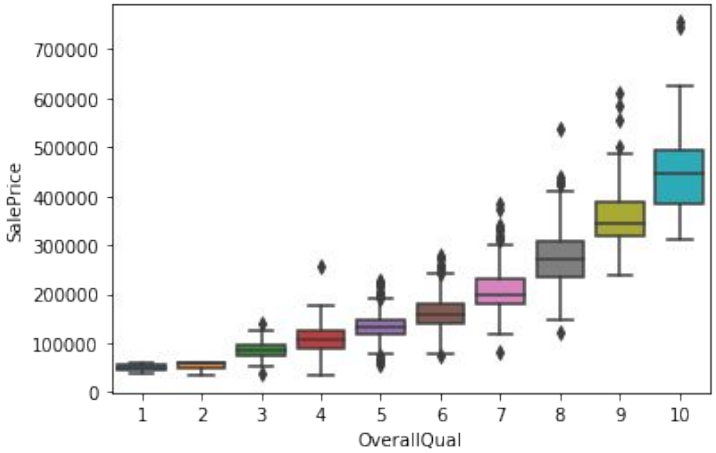
	0-19	20-39	40-59
Remodeled	$141.14	$126.20	$116.54
Non-Remodeled	$137.14	$118.43	$112.93
Increase %	3%	6.6%	3.2%

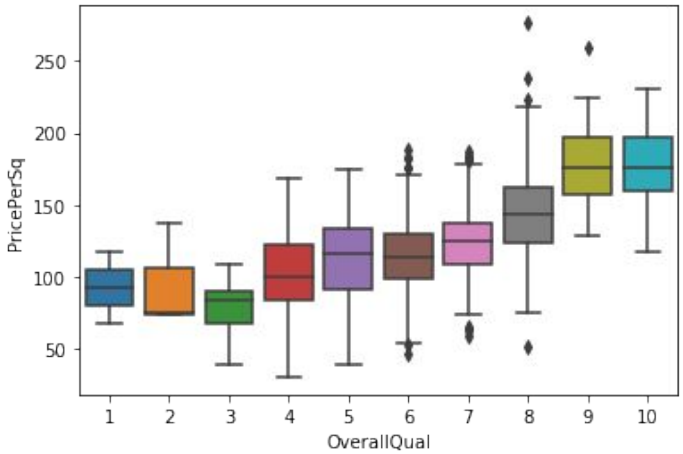
Remodeling brings the biggest impact on the price when the age of the house is between 20-39 years old.

House Age vs Avg. Sale Price

	0-19	20-39	40-59
Remodeled	$255613.63	$190699.10	156695.68
Non-Remodeled	$227679.55	$151715.08	139179.70
Increase %	12.3%	25.7%	12.6%

Quality & Condition

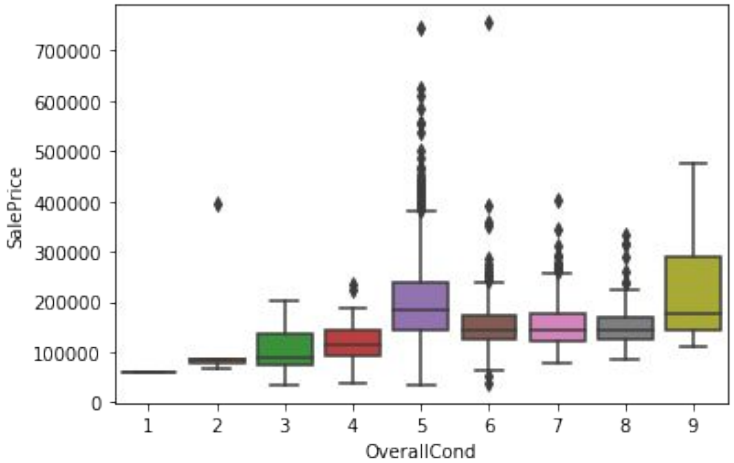


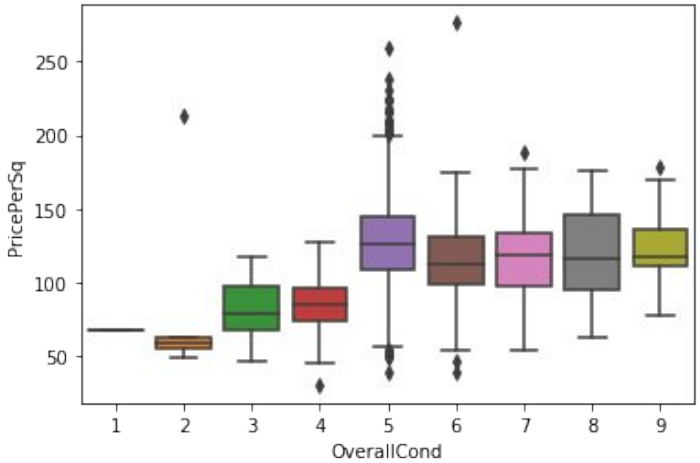


Overall Quality

Rates the overall material and finish of the house

- **Price increases when overall quality increase.**

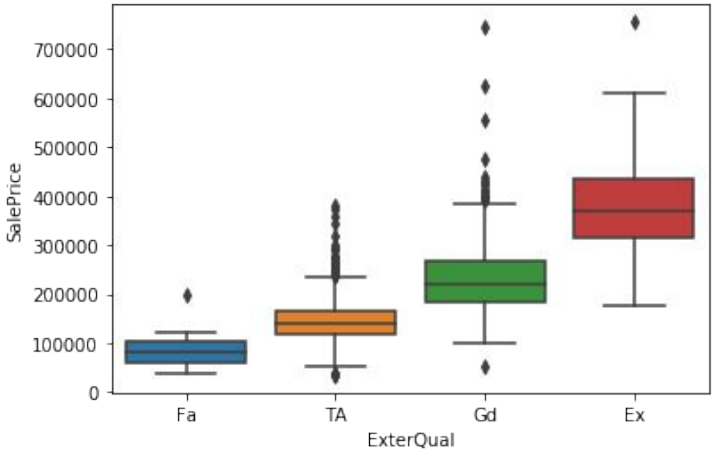


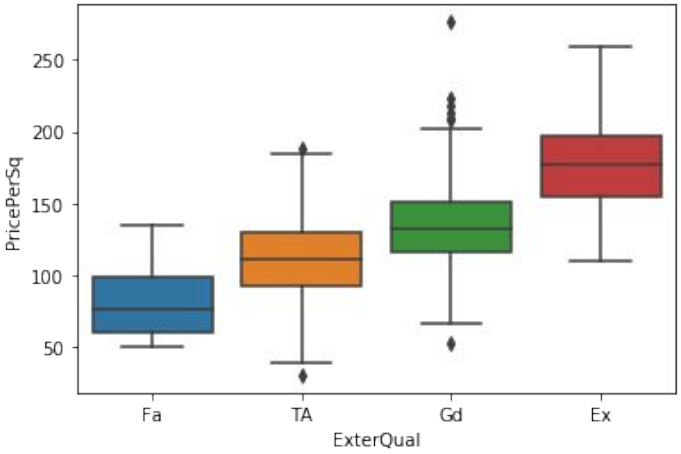


Overall Condition

Rates the overall condition of the house

- **Better overall condition does not guarantee a better price.**





Exterior Quality

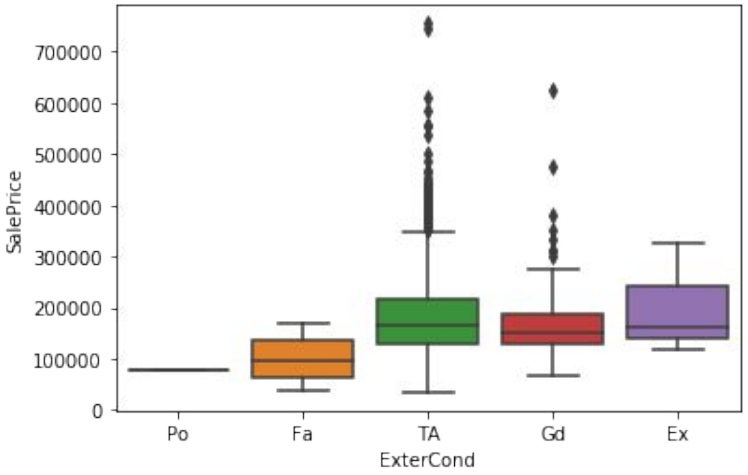
Evaluates the quality of the material on the exterior

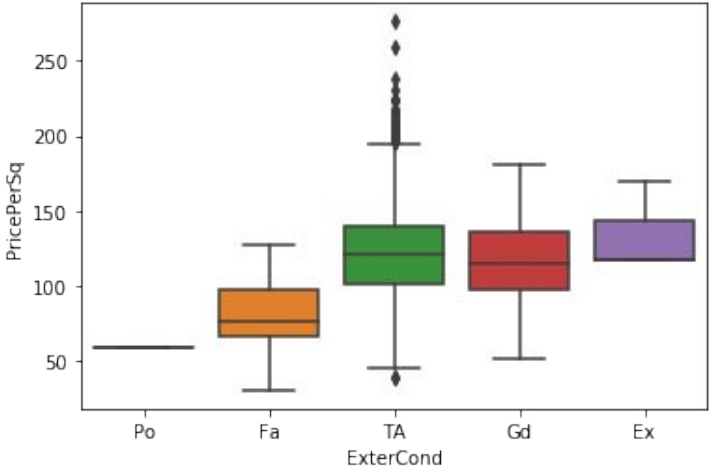
- **Price increases when exterior quality increases.**

Exterior Condition

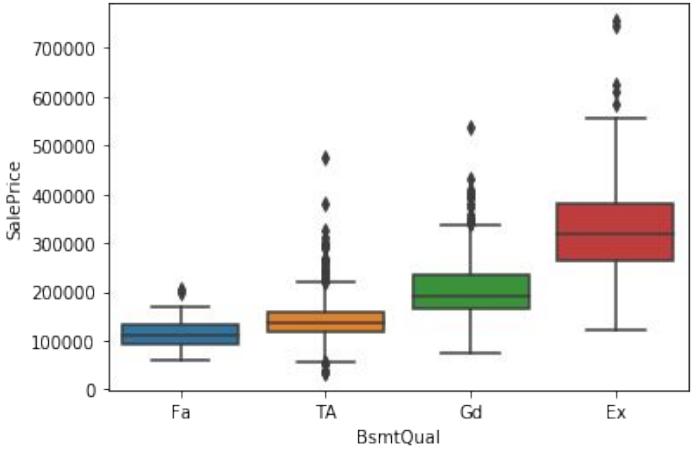
Evaluates the present condition of the material on the exterior

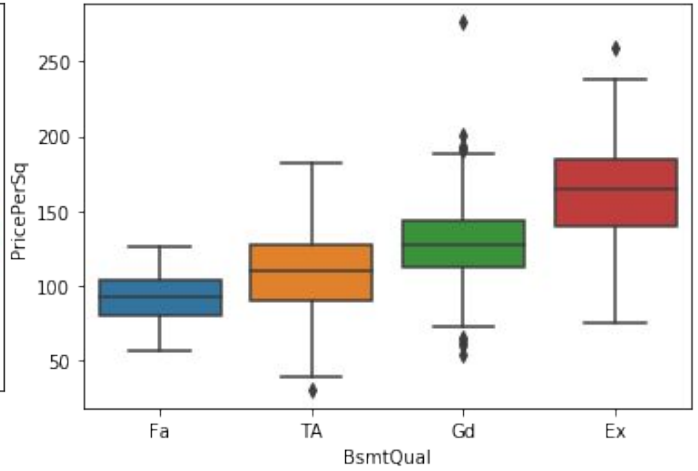
- **Better exterior condition does not guarantee a better price.**

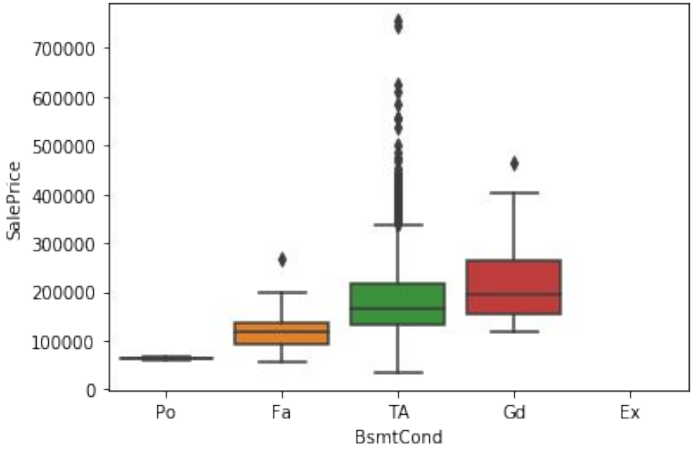


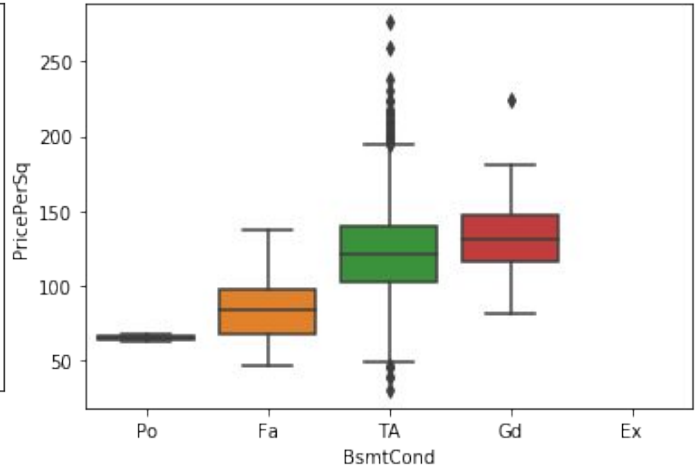


Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor







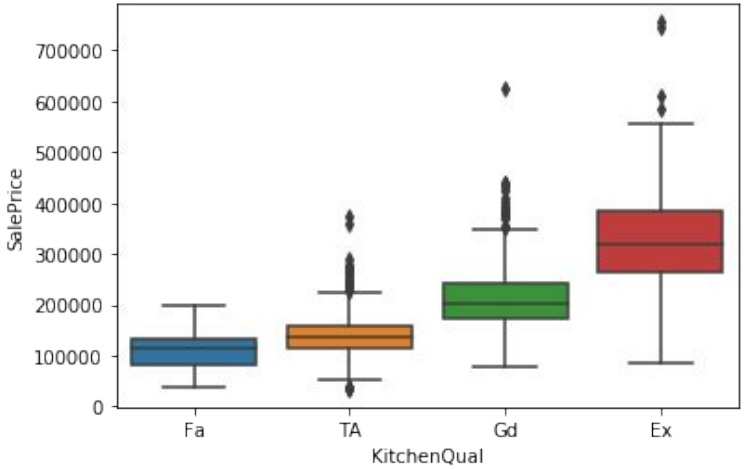


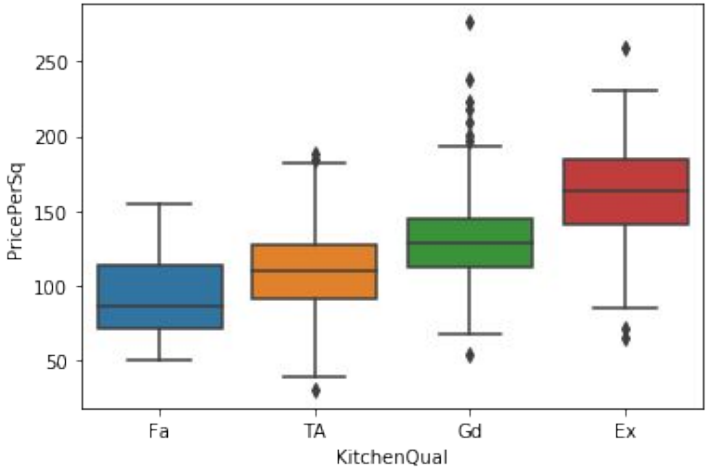
Basement Quality & Basement Condition

- **Both the height of a basement and the condition of a basement are positively correlated to house prices.**

Basement Quality	
Ex	Excellent (100+ inches)
Gd	Good (90-99 inches)
TA	Typical (80-89 inches)
Fa	Fair (70-79 inches)
Po	Poor (<70 inches
NA	No Basement

Basement Condition	
Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

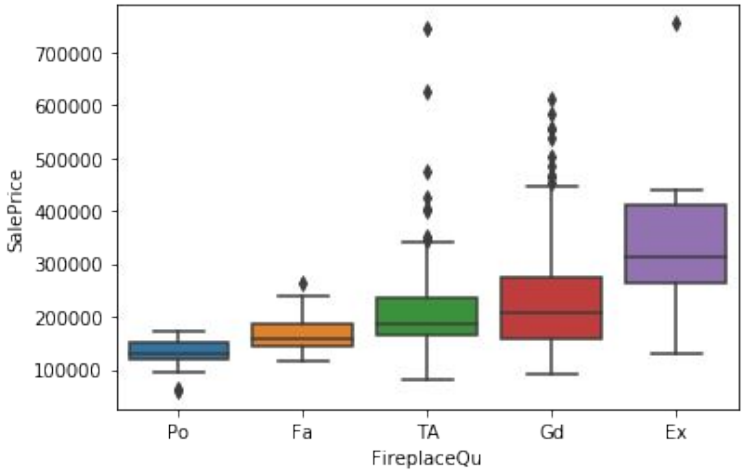


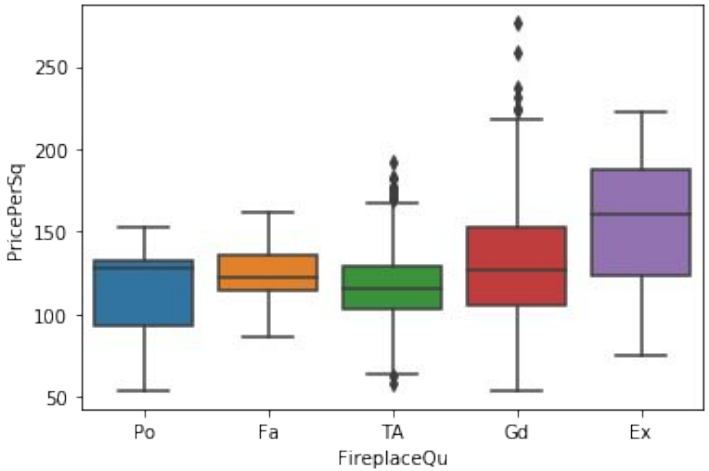


Kitchen Quality

- As kitchen quality increases, sale prices and price per sqft increase.

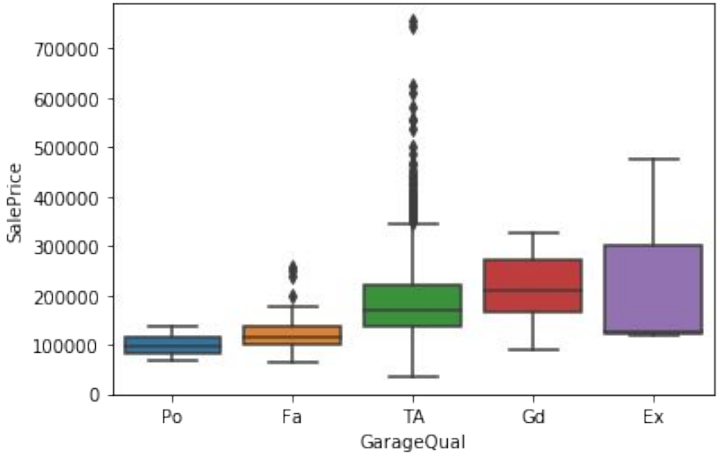
- Averagely, price per sqft increases by 20% when updating the quality to one level up.

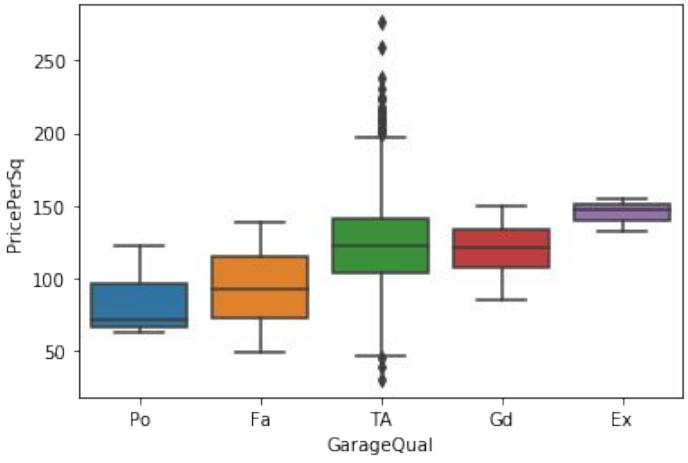


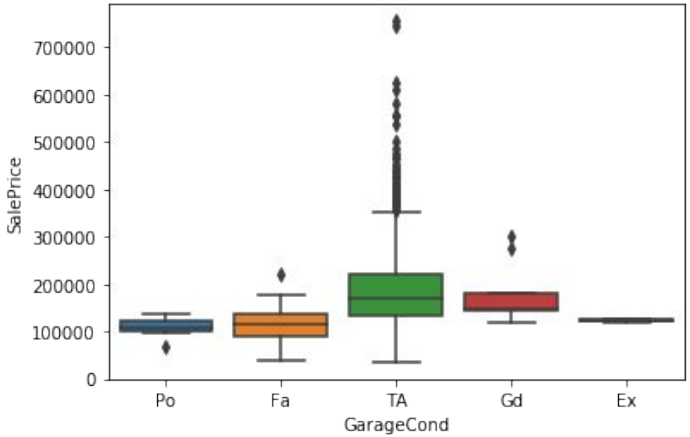


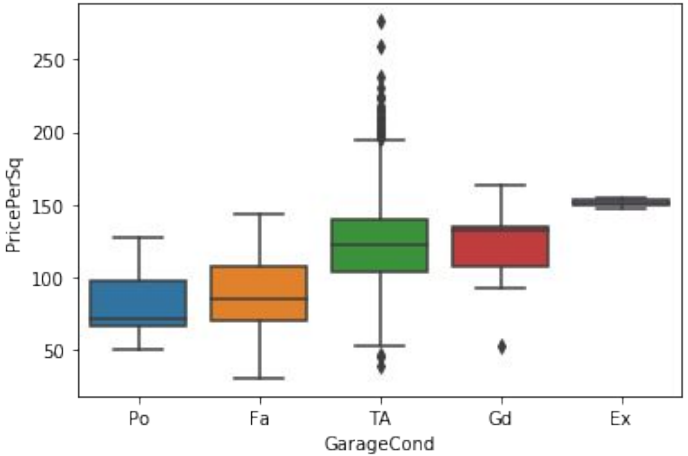
Fireplace Quality

- The distribution of fireplace quality is more skewed than kitchen quality. The avg price per sqft doesn't necessarily increase when quality improves.





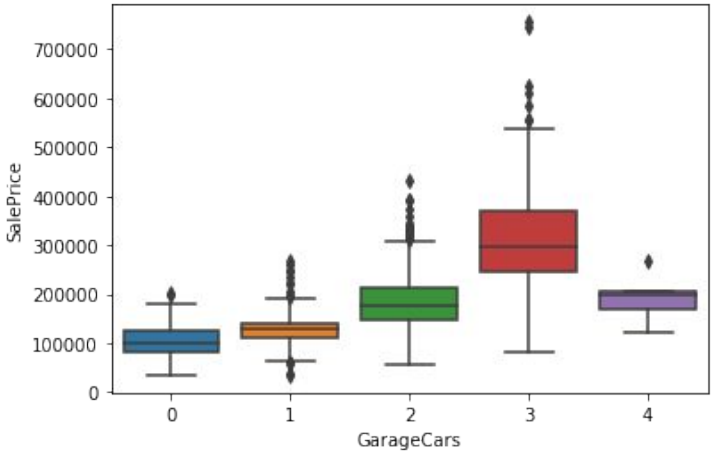


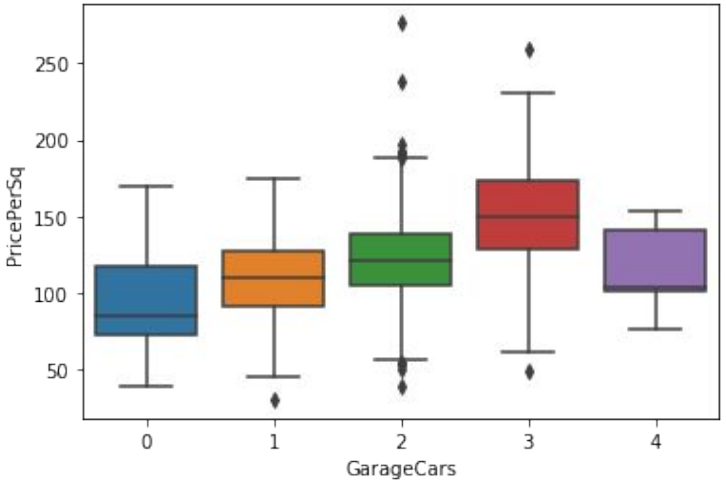


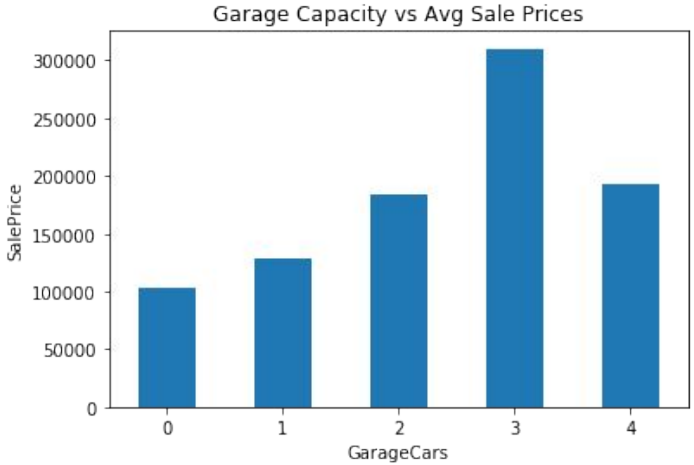
Garage Quality & Garage Condition

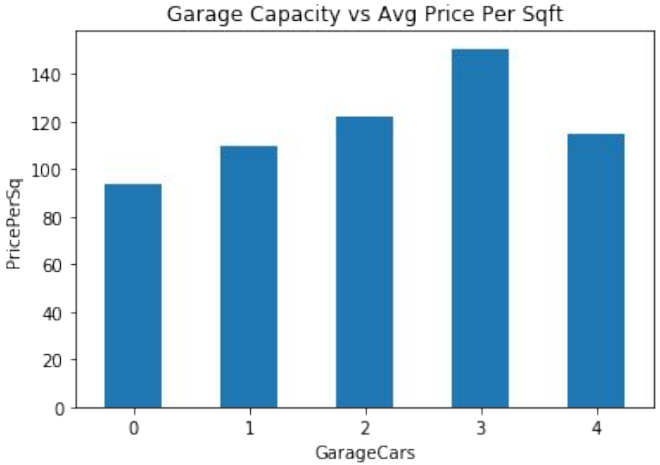
- If you have a poor or fair quality garage, updating the garage to at least TA(Average/Typical) level will help increase the house price.
- Price Per Sqft doesn't increase from TA Quality to Gd Quality.
- Price Per Sqft increases by 30% from fair quality to typical/average quality. (from poor to fair, price per sqft only increases by 9%).

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor









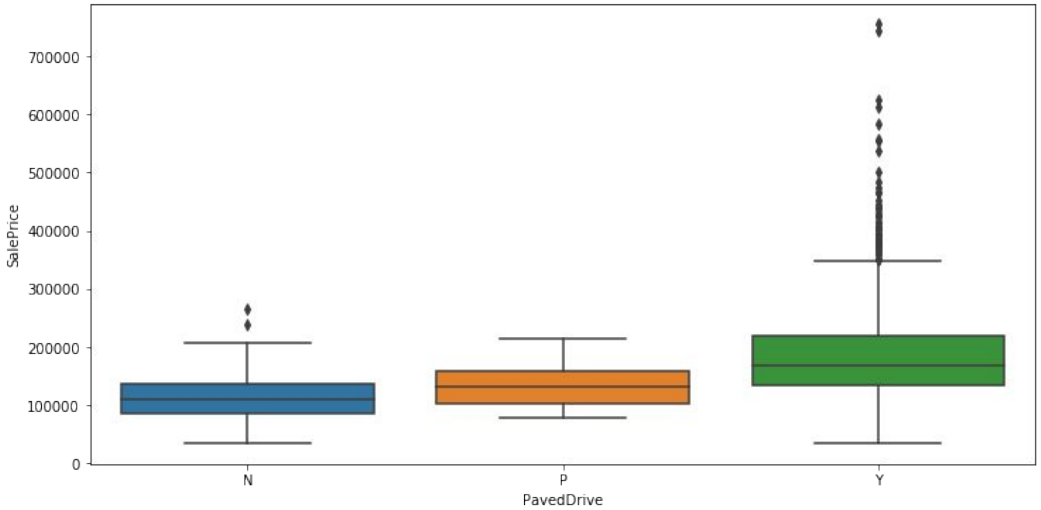
- **Garage capacity increases prices up to three cars:**

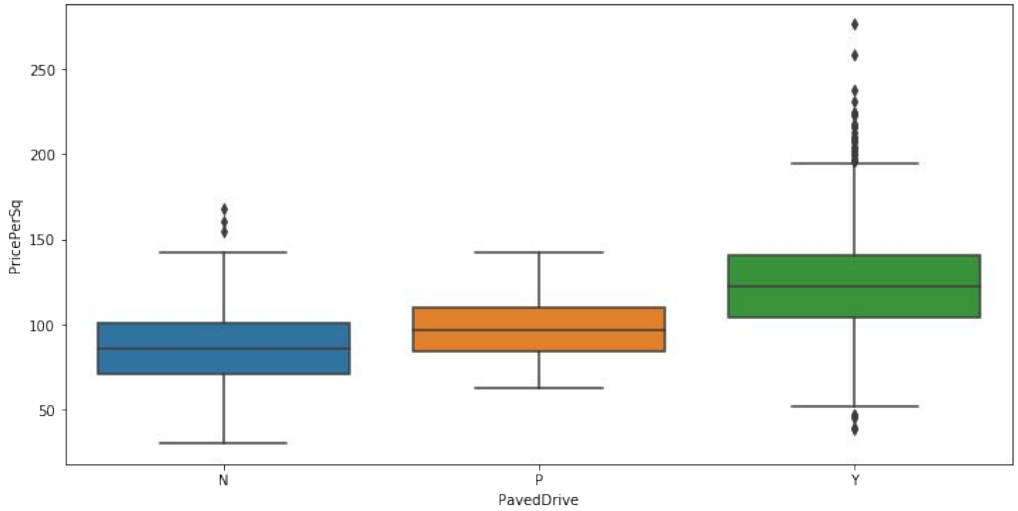
0 - 1 car:
Price per sqft increases by 16%

1-2 cars:
Price per sqft increases by 12%.

2-3 cars:
Price per sqft increases by 23%

Additional Amenities



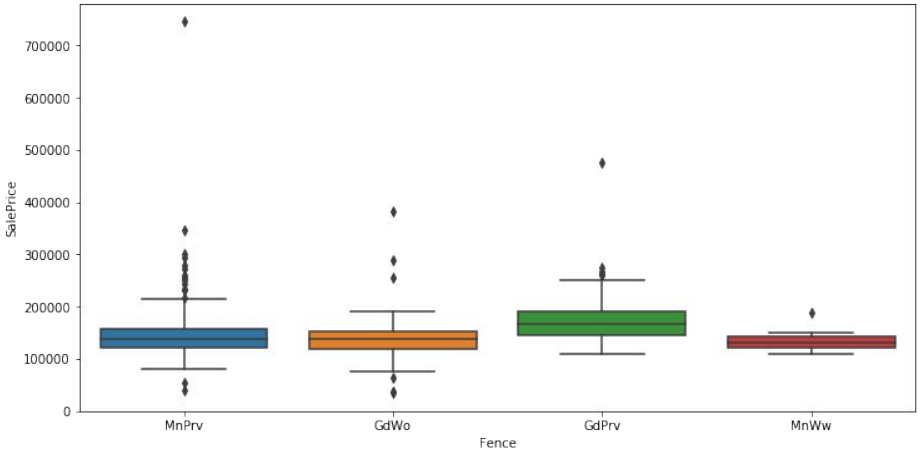


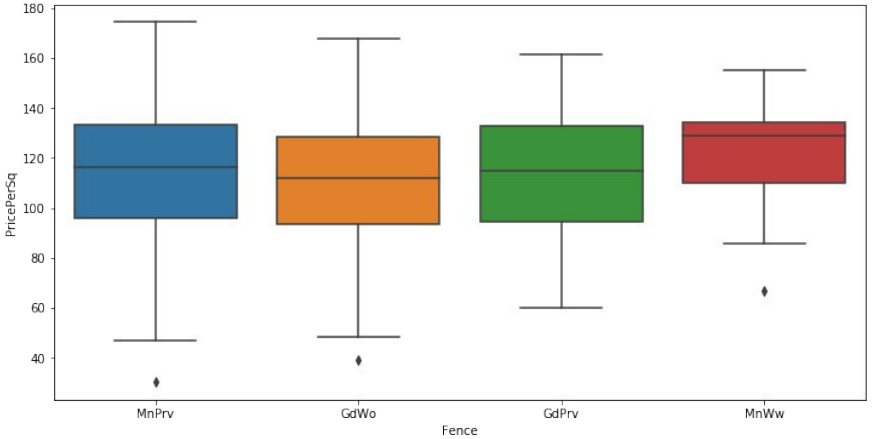
Paved Driveway vs Price

- The price of a house is higher when driveway is paved.
- From no paved driveway to a partial paved driveway, the avg price per sqft increases by 10%.
- From a partial paved driveway to a paved driveway, the avg price per sqft increases by 26%

Y	Paved
P	Partial Pavement
N	Dirt/Gravel

Fence vs Price

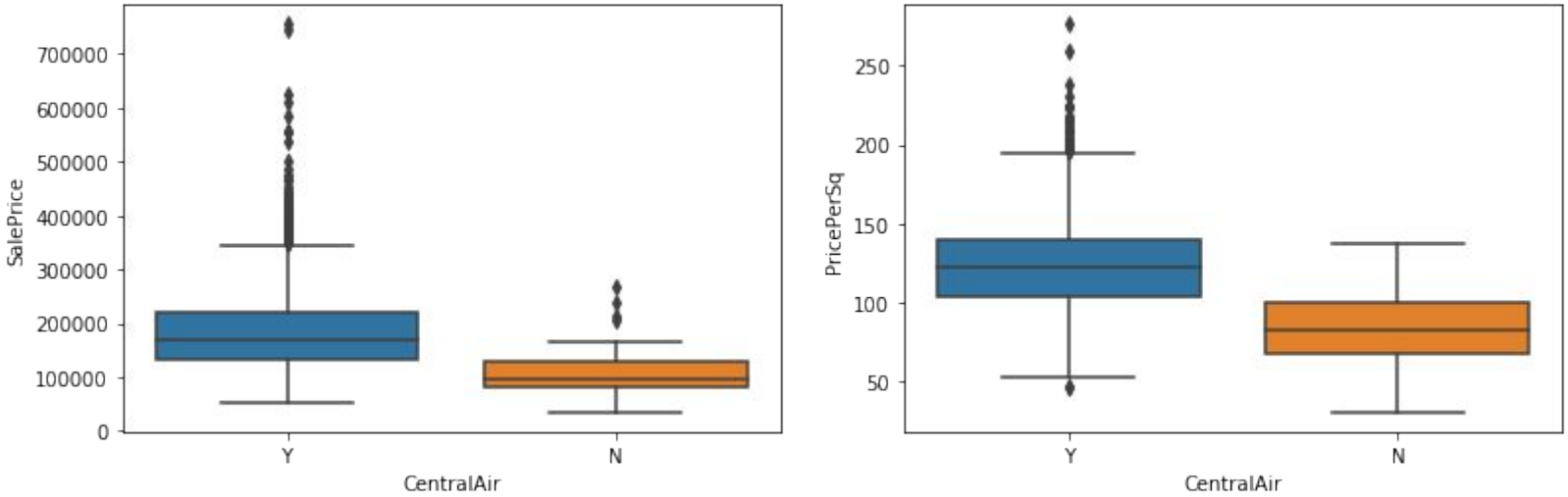




GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

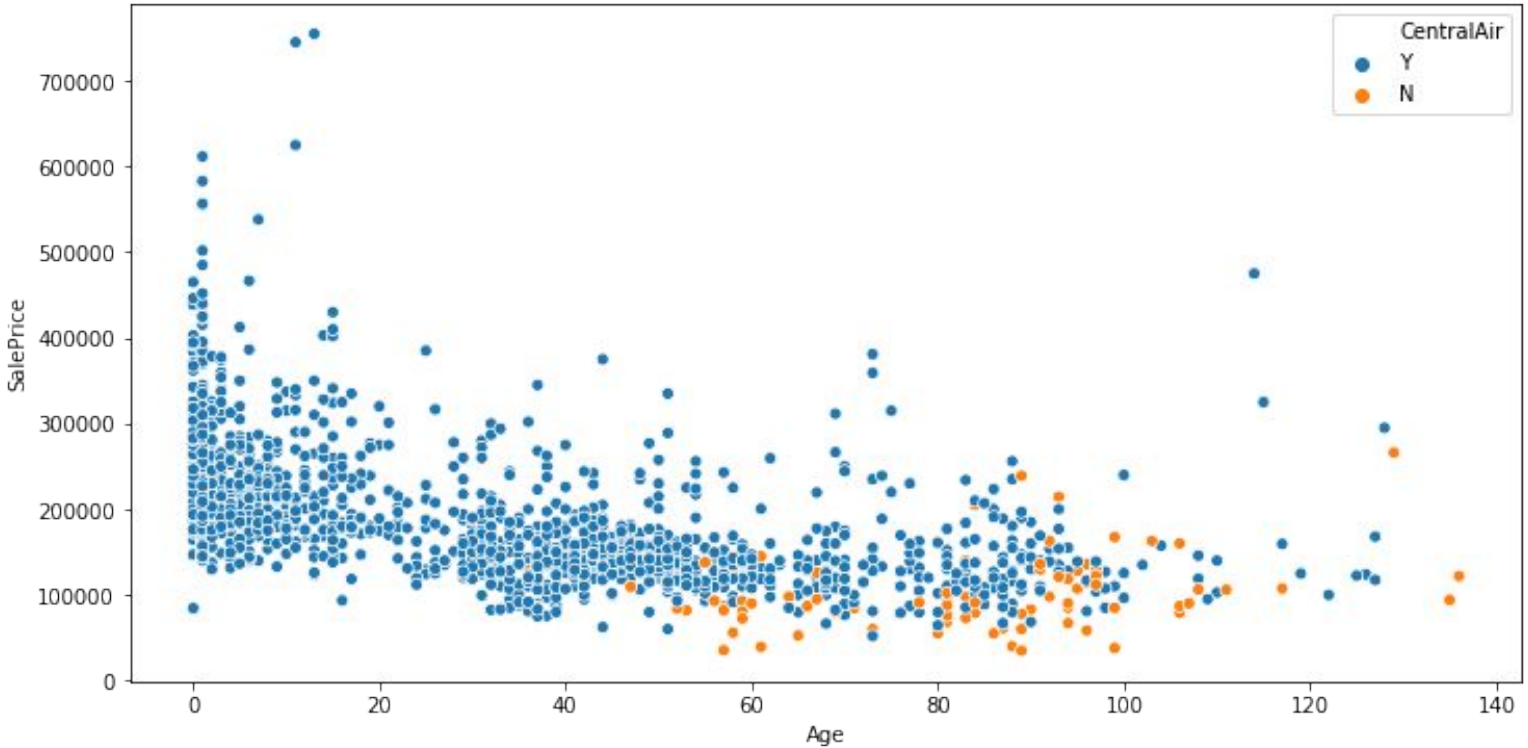
Averagely, the price of a house with a better fence quality is not necessarily higher.

Central Air



Averagely, the sale price of a house featuring central air is 70% higher than the price of a house that has no central air. Price per sqft increase by 46%.

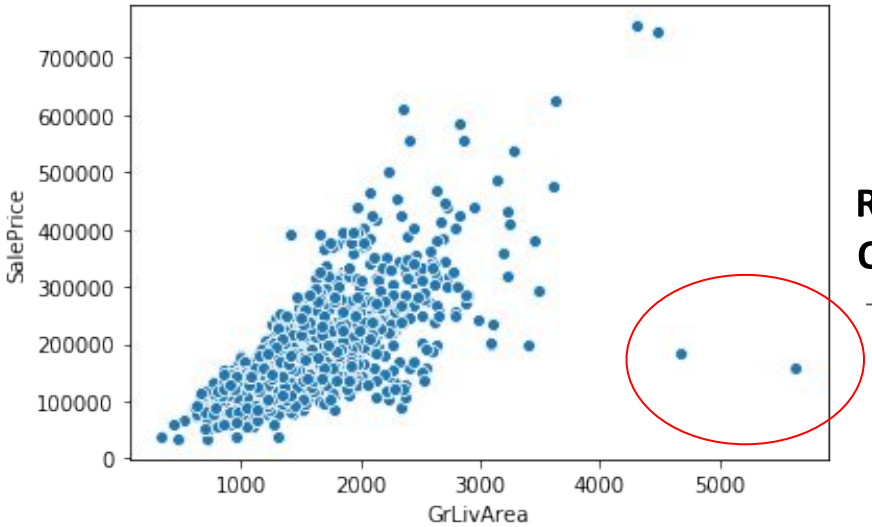
Central Air



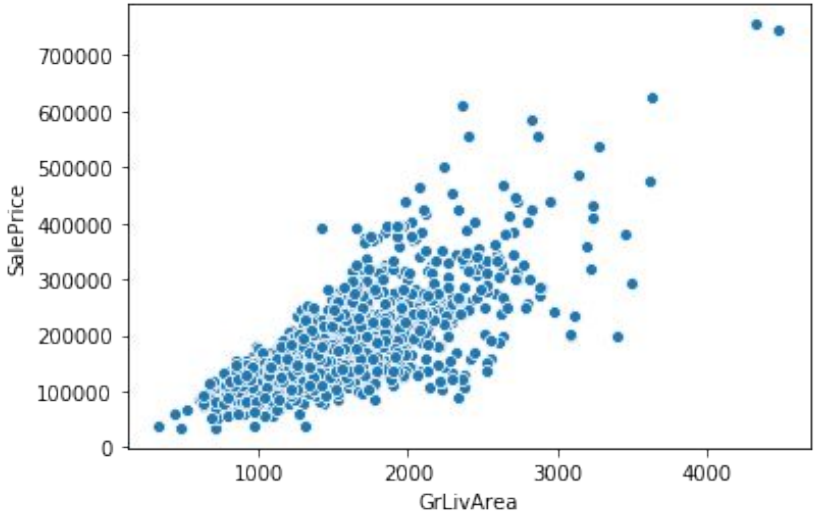
Most of the houses (94%) in the dataset have central air except for houses over 50 years old. However among the same age range, the prices of houses with central air are still higher than houses without central air.

Machine Learning

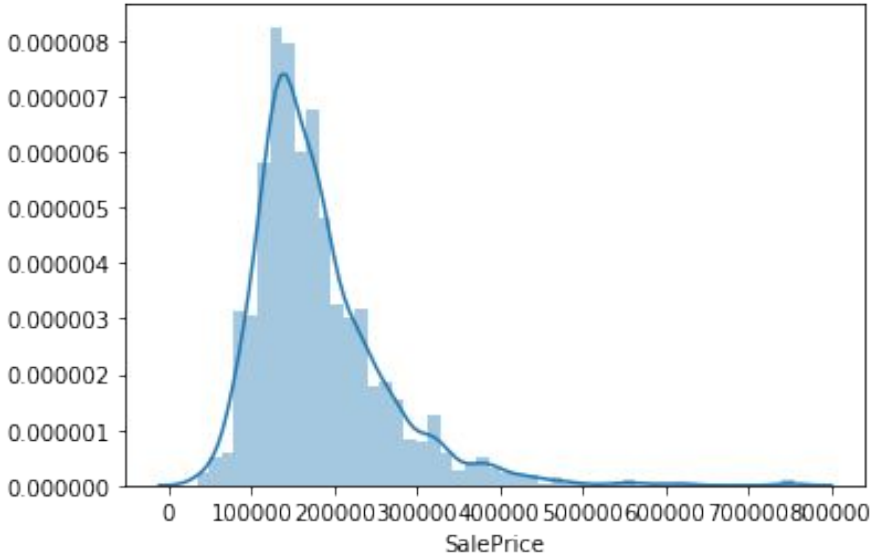
Pre-Processing



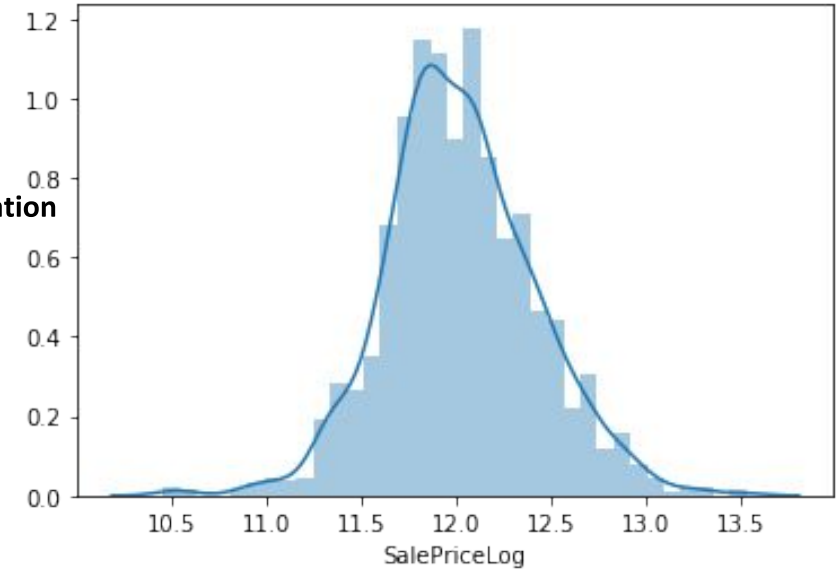
Remove Outliers



Pre-Processing

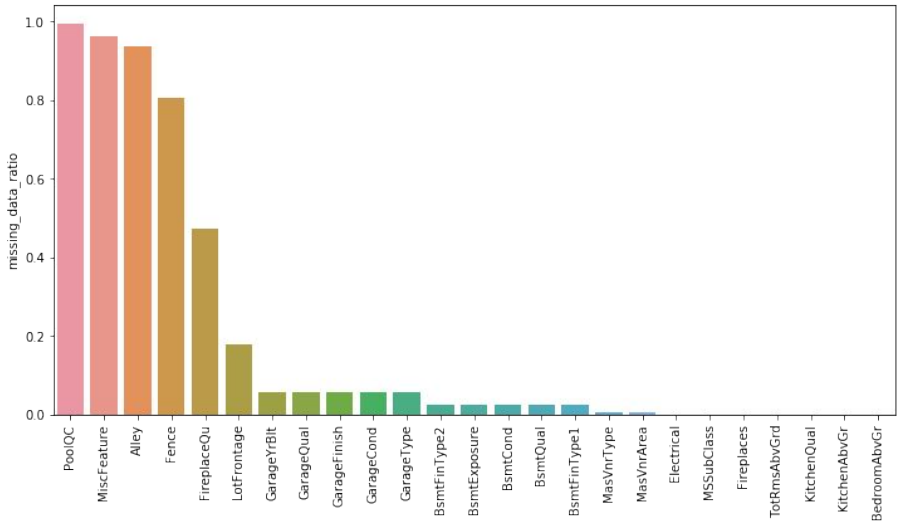


Log Transformation

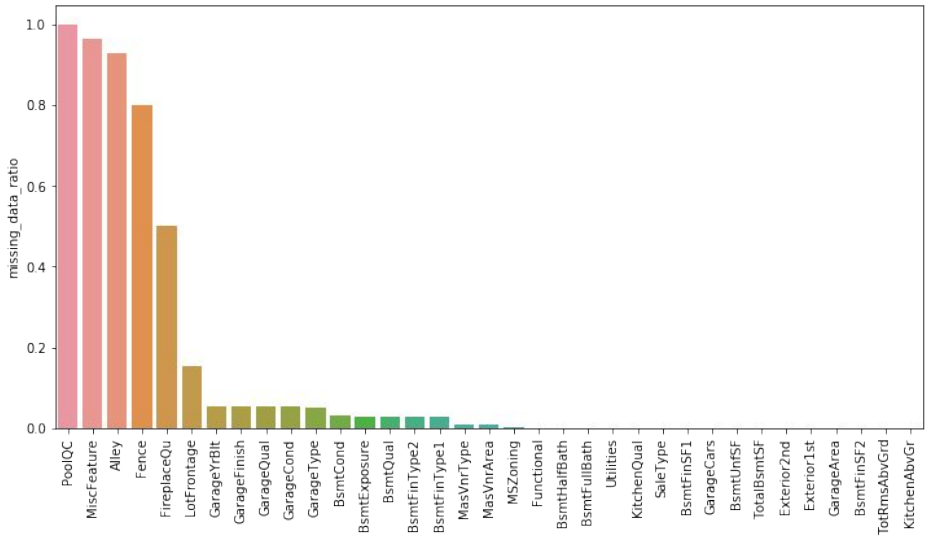


Impute Missing Values

Train



Test



Some of the missing values in these datasets simply mean that a house does not have a specific house feature, such as pool, alley, garage, basement, etc.

Impute Missing Values

No..(NoPool, NoBasement, NoAlley, etc...)	Median	Mode	0
PoolQC MiscFeature Alley Fence FireplaceQu GarageType GarageFinish GarageCond GarageQual BsmtFinType2 BsmtExposure BsmtCond BsmtQual BsmtFinType1 MasVnrType	LotFrontage BsmtFinSF1 BsmtUnfSF TotalBsmtSF	Electrical MSZoning Functional BsmtHalfBath BsmtFullBath Utilities KitchenQual SaleType Exterior1st Exterior2nd	GarageYrBlt MasVnrArea GarageArea BsmtFinSF2 GarageCars

Pre-Processing

Feature Engineering

Columns Added

- **Age** = YrSold - YrBuilt
- **Remodel Age** - YrSold - YearRemodAdd
- **Remodeled** - if the house is remodeled
- **HasGarage** - if the house has a garage
- **GarageAge** = YrSold - GarageYrBlt

Columns Dropped

- YrSold, GarageYrBlt, BsmtFinSF2, BsmtUnfSF, BsmtHalfBath, 2ndFlrSF, MoSold , MiscVal, 3SsnPorch, EnclosedPorch, BedroomAbvGr, LowQualFinSF, MiscFeature, Condition2

Assign numerical numbers to quality variables

- One Hot Encoder does not do a good job
- No -0, Poor - 1, Fair -2,
- TA -3, GD- 4, Ex - 4

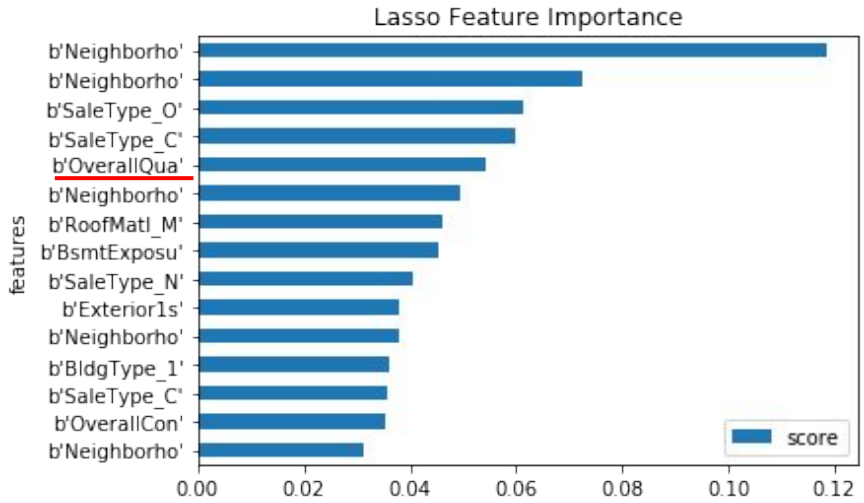
Dummify - get.dummies

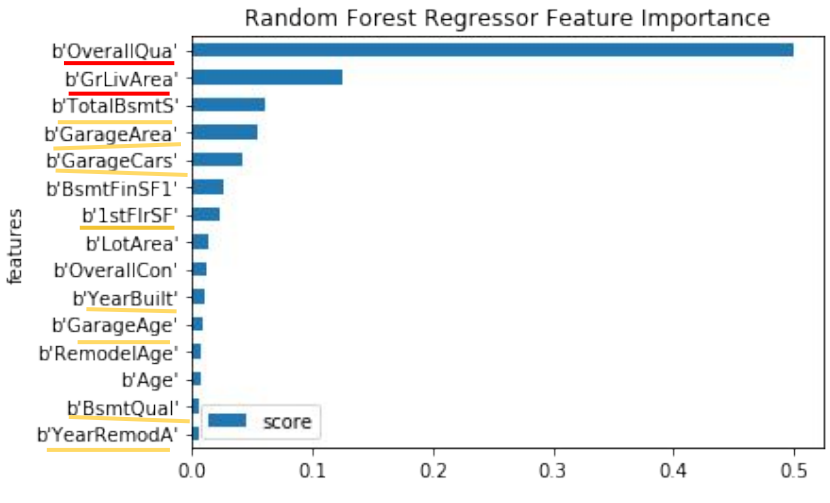
- Categorical variables

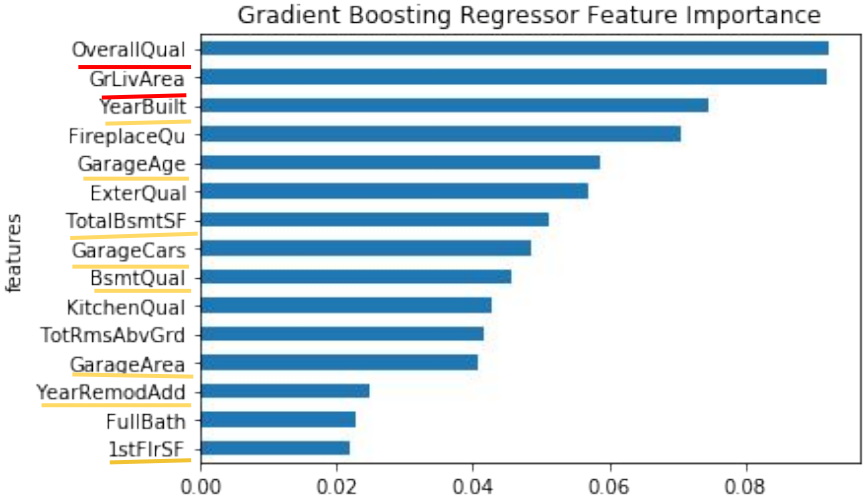
Machine Learning Models & Scores

	Ridge	Lasso	Random Forest	Gradient Boosting
Train Score	0.9110	0.9205	0.8786	0.9114
Test Score	0.9132	0.9144	0.8682	0.9099
RMSE	0.1141	0.1133	0.1407	0.1163

Gradient Boosting have the best performance on Kaggle.







- **Overall Quality, Above Ground Living Area(GrLivArea) are the most important indicators when predicting sale prices.**

- **Important indicators to prediction:**
  Year Built, Garage Area, Garage Cars, Basement Quality, Remodel Year, Total Basement Area, and 1stFlrSF are some of the most important features in both random forest and gradient boosting model.

- **Lasso- neighborhood is a good indicator in the lasso model.**

Recommendation

- House quality outweighs house condition (quality of house materials is highly important).

- Remodeling a 20 to 39-year-old house makes you have more leverage and possibilities to increase returns.

- It is not necessary to update your garage to a 4-car garage. A 4-car garage does not guarantee a higher house price. Updating a garage to at least the TA(Typical/Average) level will improve the price. Price Per Sqft increases by 30% from fair quality to typical/average quality.

- Central Air is a necessity: Central Air is commonly equipped in the house nowadays. The price per sqft of a house featuring central air is 46% higher than the price per sqft of a house that has no central air.

- Spend money improving the quality in this order: kitchen > driveway > fireplace> fence.