House Selling Price Prediction – ML Foundation Project (INSAID June Cohort 2019)

- In total, there are about 1460 rows and 81 columns which contain descriptive information on different houses
- We have the house dataset from the year 1872 to 2010 with sale price of minimum of $ 34900 to Maximum of $ 625,000.
- Taken top 10 features to predict the House Selling Price
- What factors can you think of right now which can influence house prices ?

Sale Price - the property's sale price in dollars. This is the target variable that you're trying to predict.

Dependent Variable

```
House_data['SalePrice'].describe()
```

```
count      1450.000000
mean     180006.829655
std       76693.580654
min       34900.000000
25%      129900.000000
50%      162900.000000
75%      213497.500000
max      625000.000000
Name: SalePrice, dtype: float64
```

```
numeric_data = House_data.select_dtypes(include=[np.number])
cat_data = House_data.select_dtypes(exclude=[np.number])
print ("There are {} numeric and {} categorical columns in House_data".
       format(numeric_data.shape[1],cat_data.shape[1]))
```

```
There are 38 numeric and 43 categorical columns in House_data
```

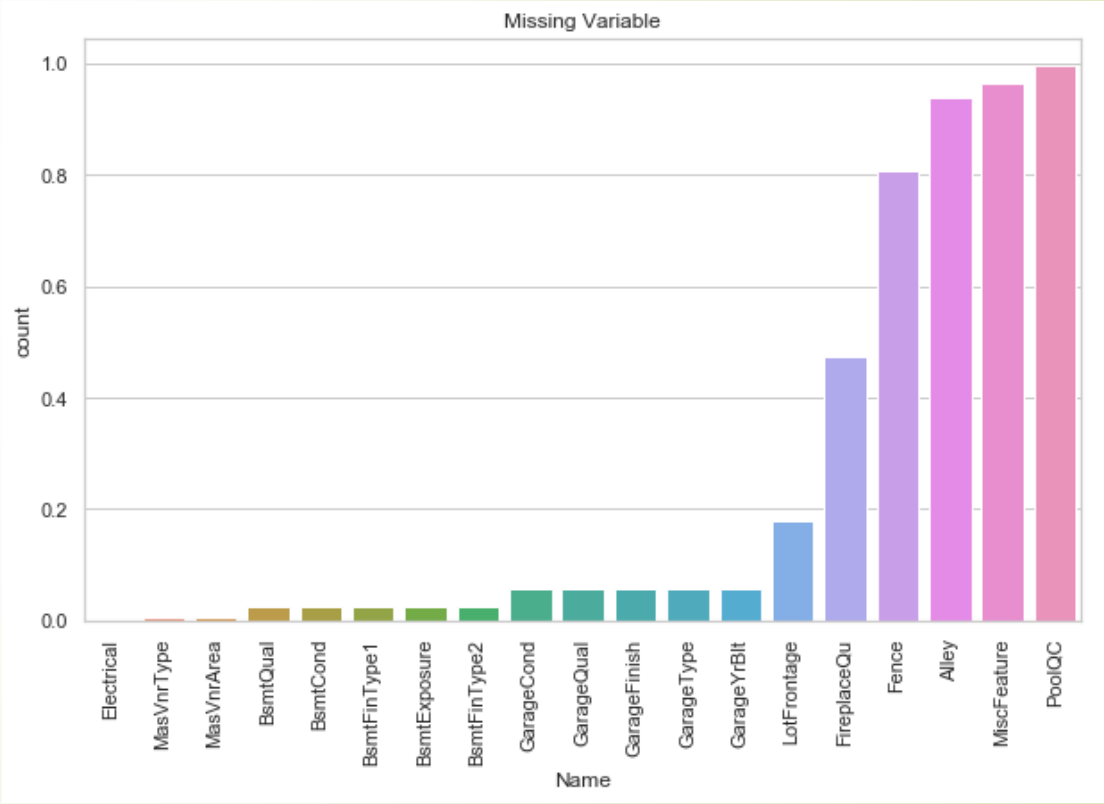
Out of 37 variables, we are using top 10 most correlated features

```
most_corr = pd.DataFrame(cols)
most_corr.columns = ['Most Correlated Features']
most_corr
```

	Most Correlated Features
0	SalePrice
1	OverallQual
2	GrLivArea
3	GarageCars
4	GarageArea
5	TotalBsmtSF
6	1stFlrSF
7	FullBath
8	TotRmsAbvGrd
9	YearBuilt
10	YearRemodAdd

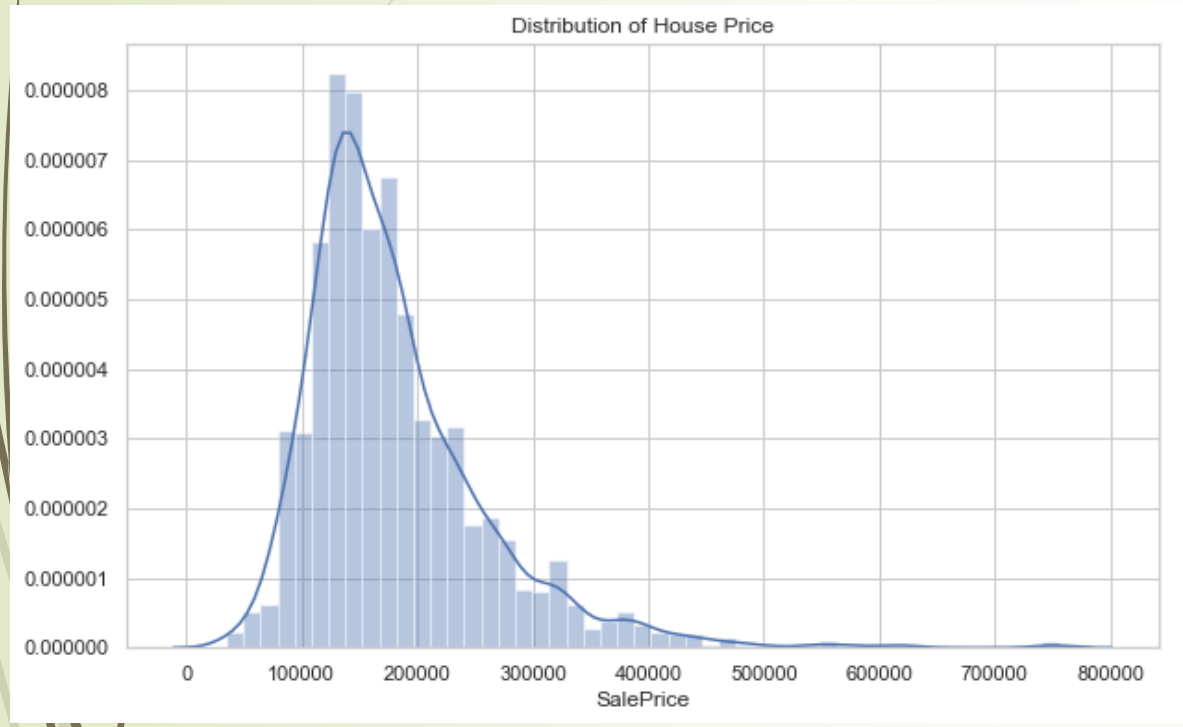
Missing Values

From this plot, we can conclude PoolQC has 99.5% missing values followed by MiscFeature, Alley, and Fence with 96.3% and 93.7 % respectively

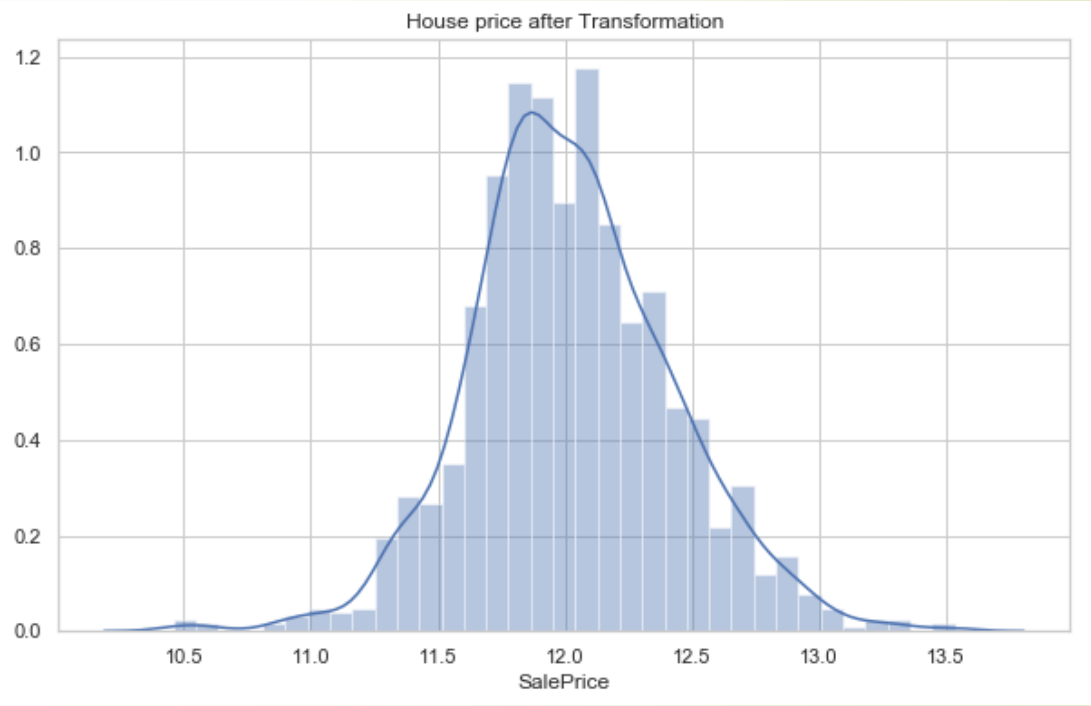


Distribution of House Price

Target variable Sales Price has a right-skewed distribution



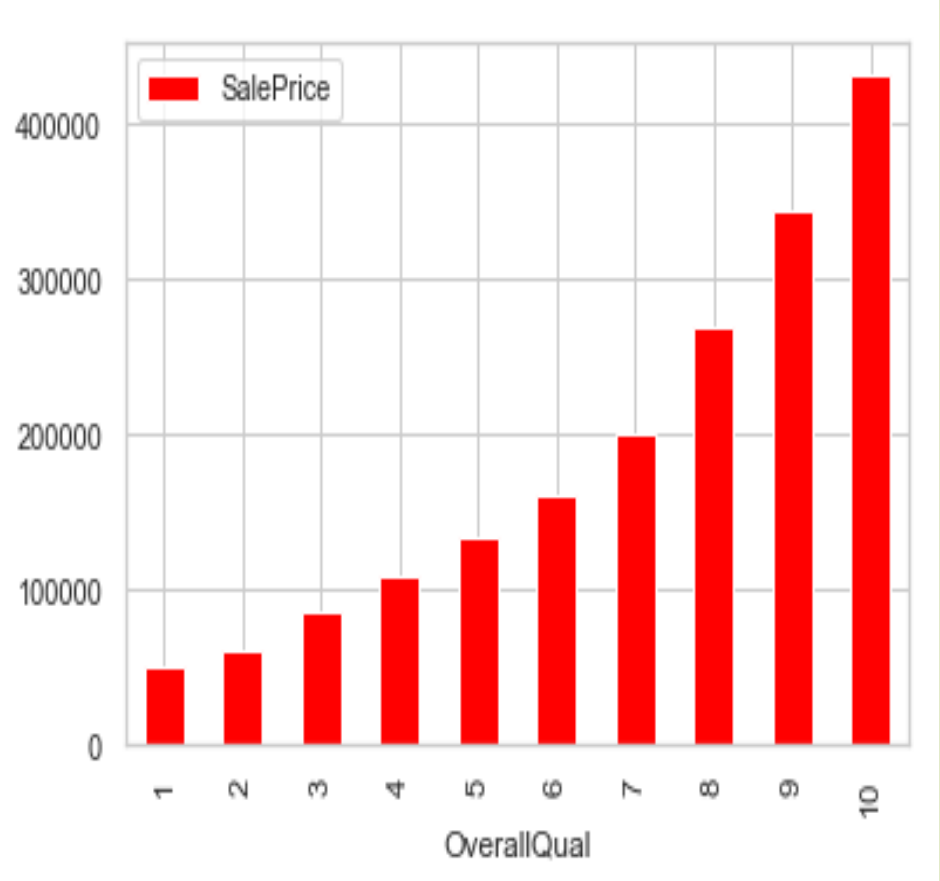
As you can see below plot, transformation has helped us to normalize the distribution



The house prices are right-skewed with a mean and a median around $200,000. Most houses are in the range of 100k to 250k; the high end is around 550k to 750k with a sparse distribution.

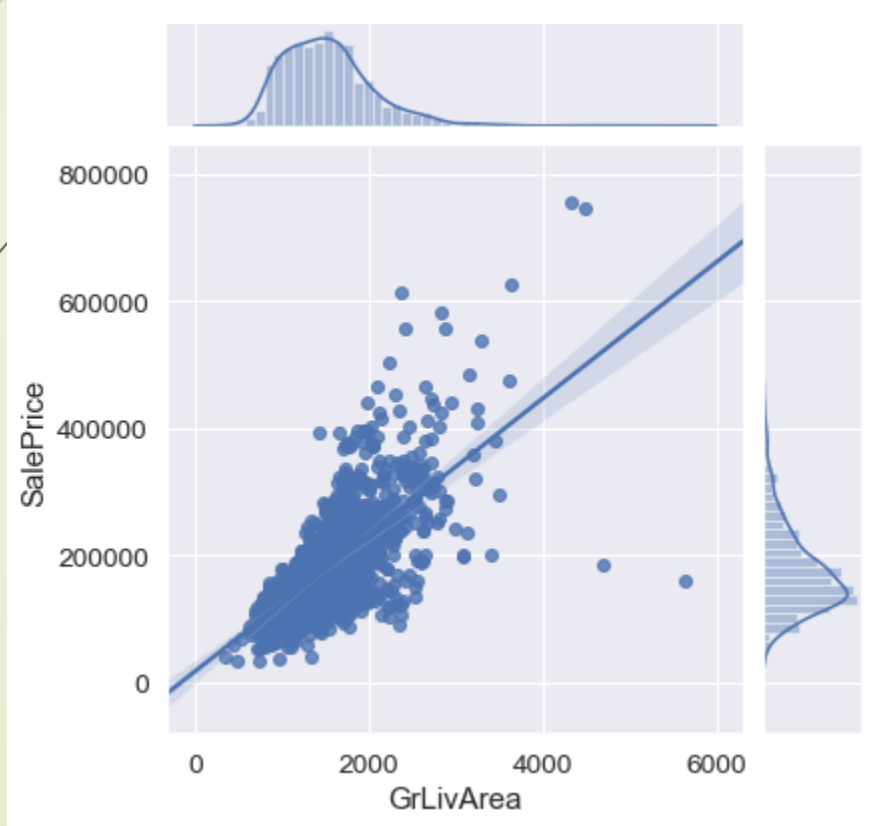
Overall Quality vs. House Sale Price

Most of the variables in the dataset (51 out of 79) are categorical. They include things like the neighborhood of the house, the overall quality, the house style, etc. The most predictive variables for the sale price are the quality variables. For example, the overall quality turns out to be the strongest predictor for the sale price. Quality on particular aspect of the house, like the pool quality, the garage quality, and the basement quality, also show high correlation with the sale price.

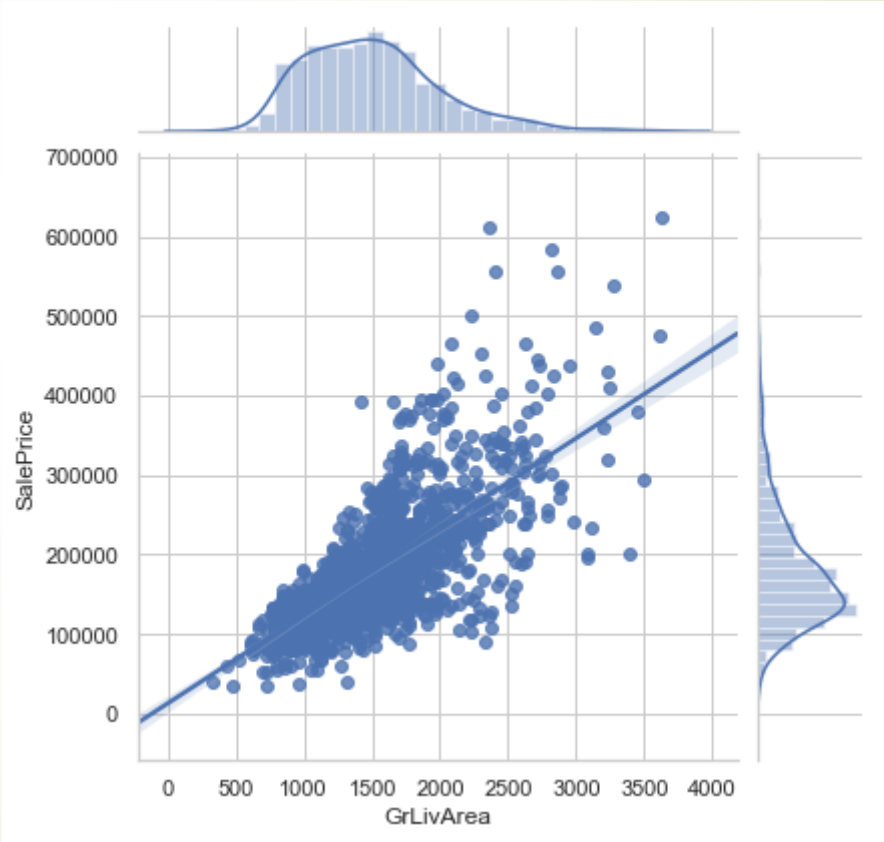


Above Grade (Ground) Living Area Square Feet vs. House Sale Price

We can spot an outlier value GrLivArea > 5000. I've seen outliers play a significant role in spoiling a model's performance.

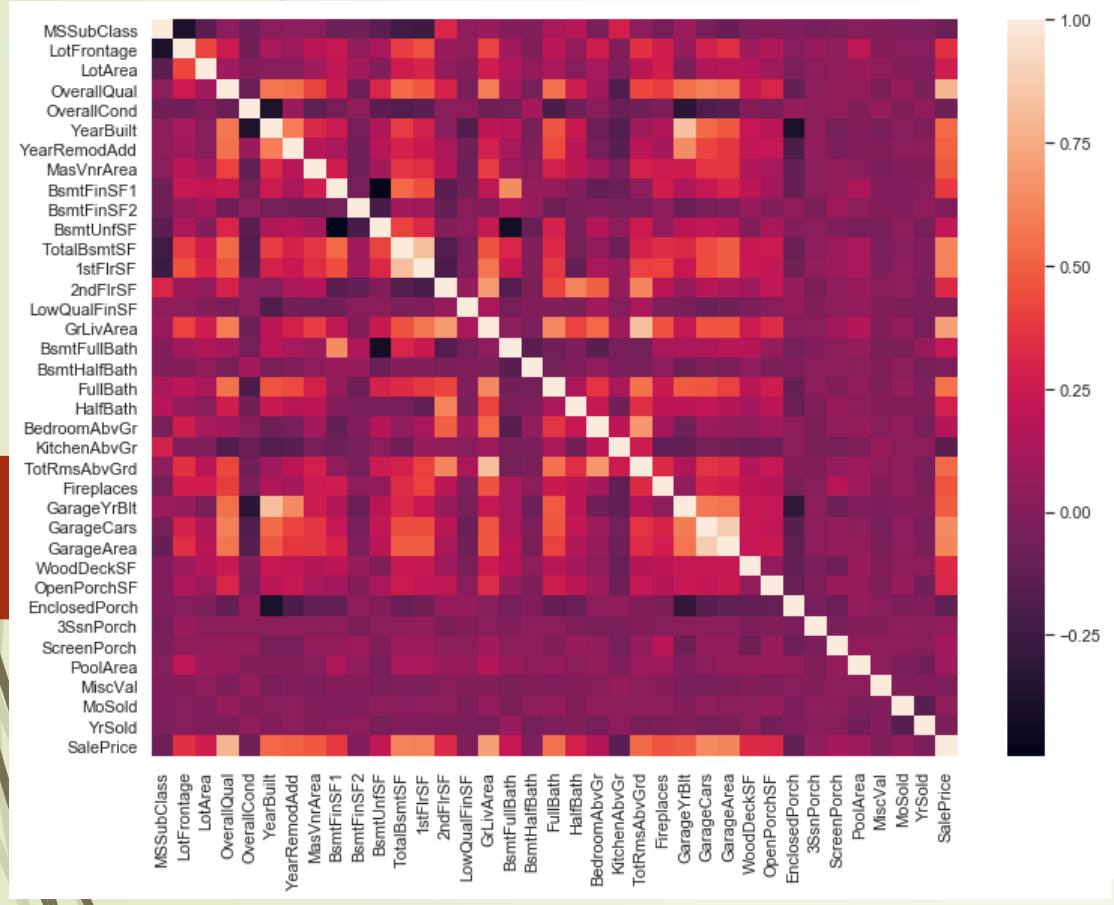


After Removing Outlier

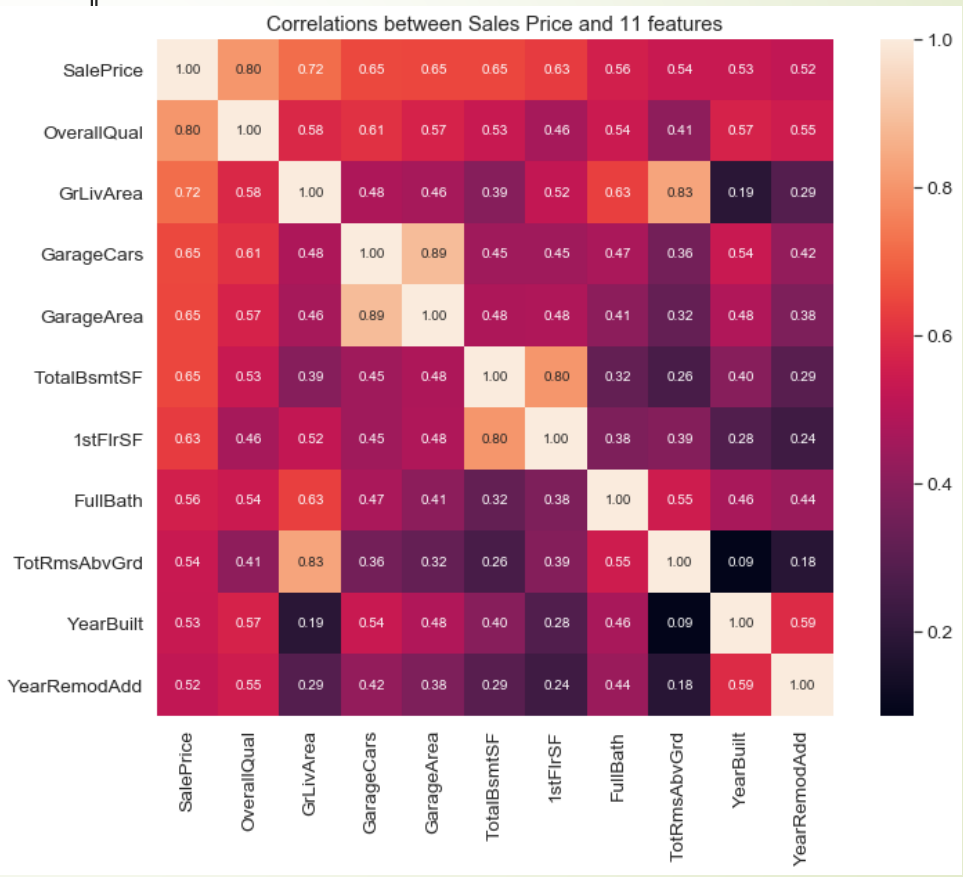


Correlation

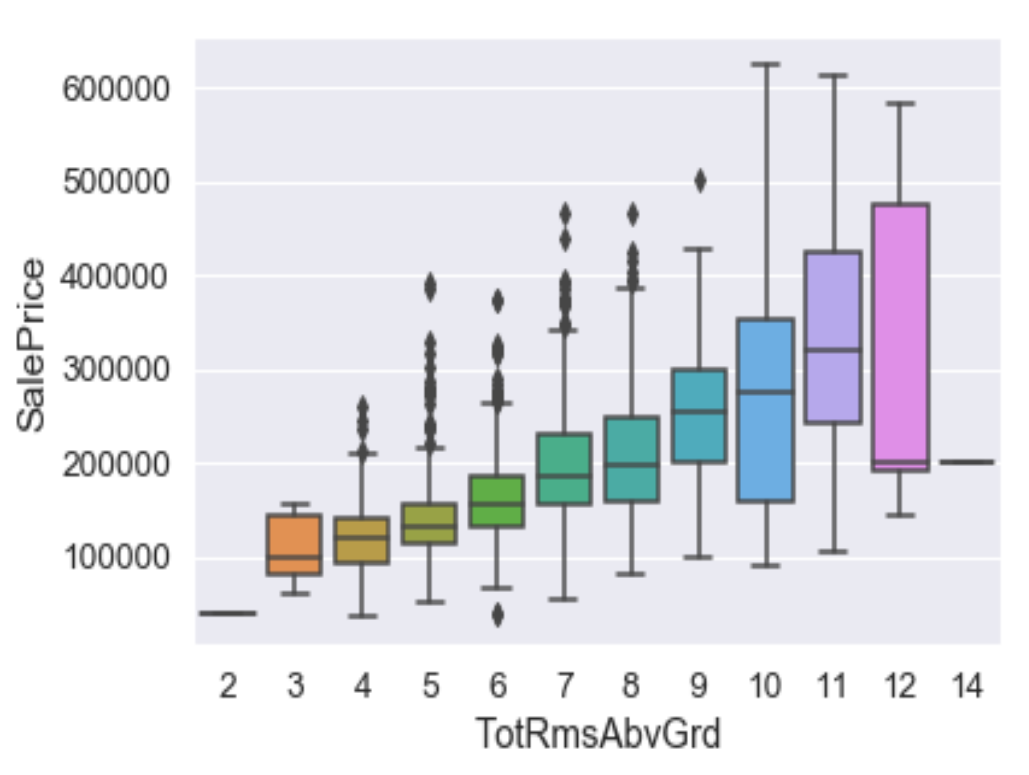
Correlation of numeric variables



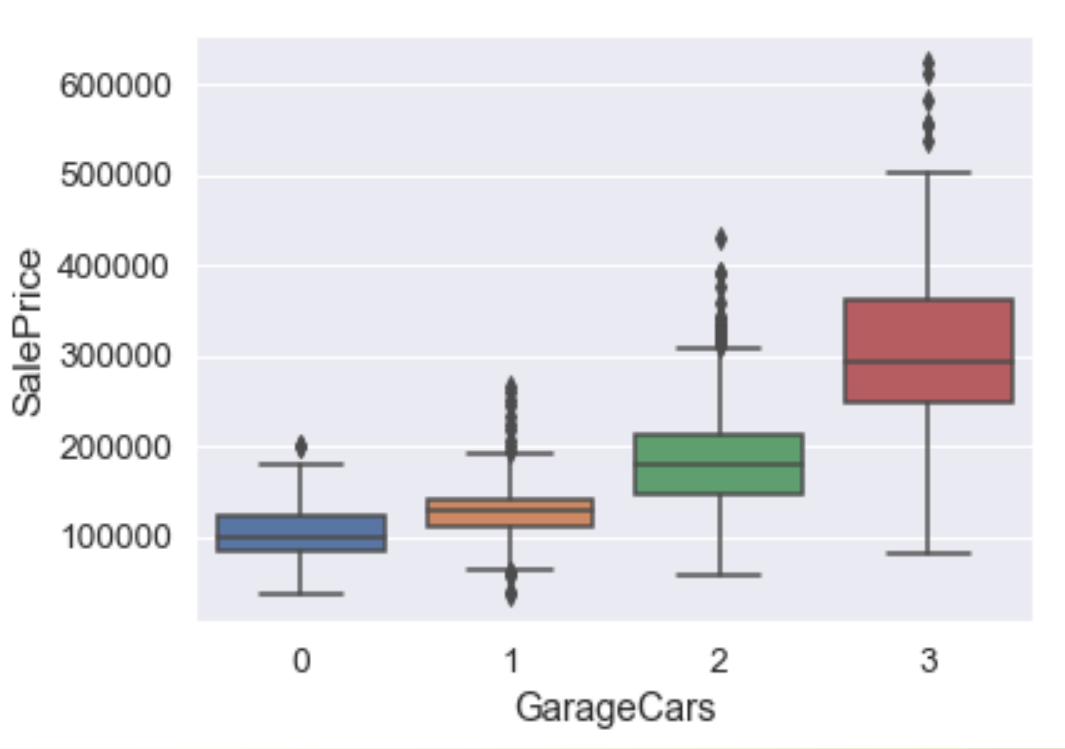
Top 10 most Correlated Variables



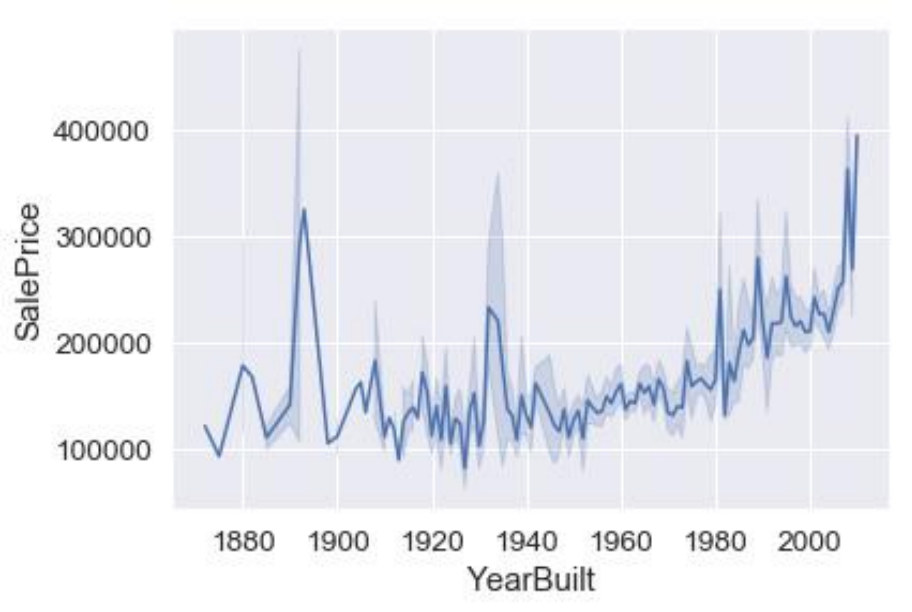
Total Rooms Vs Sales Prices



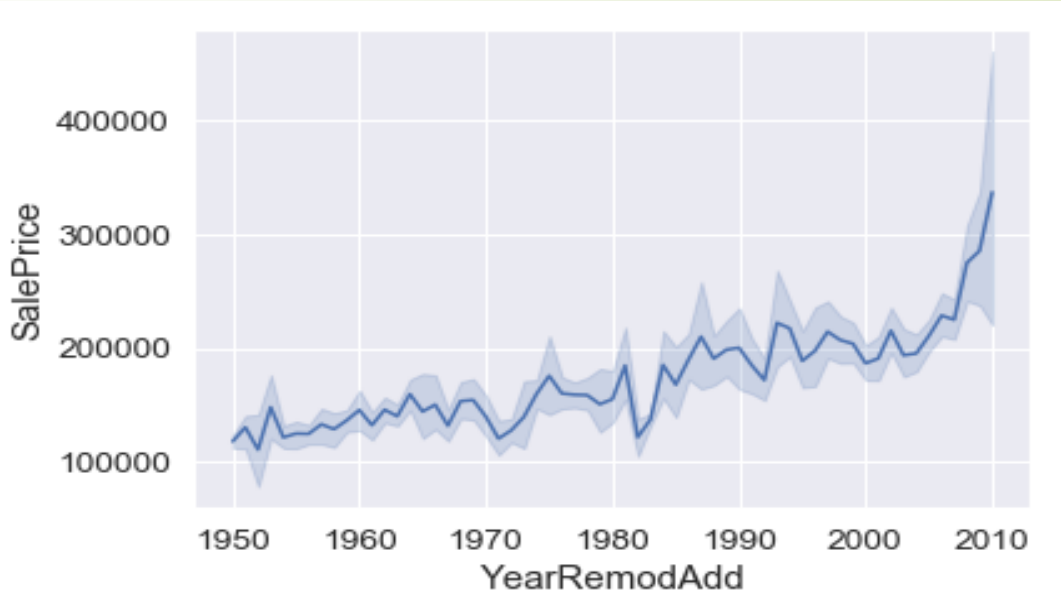
Garage Area vs Sale Price



Year Built vs Sale Price



Year Remodeled vs Sale Price



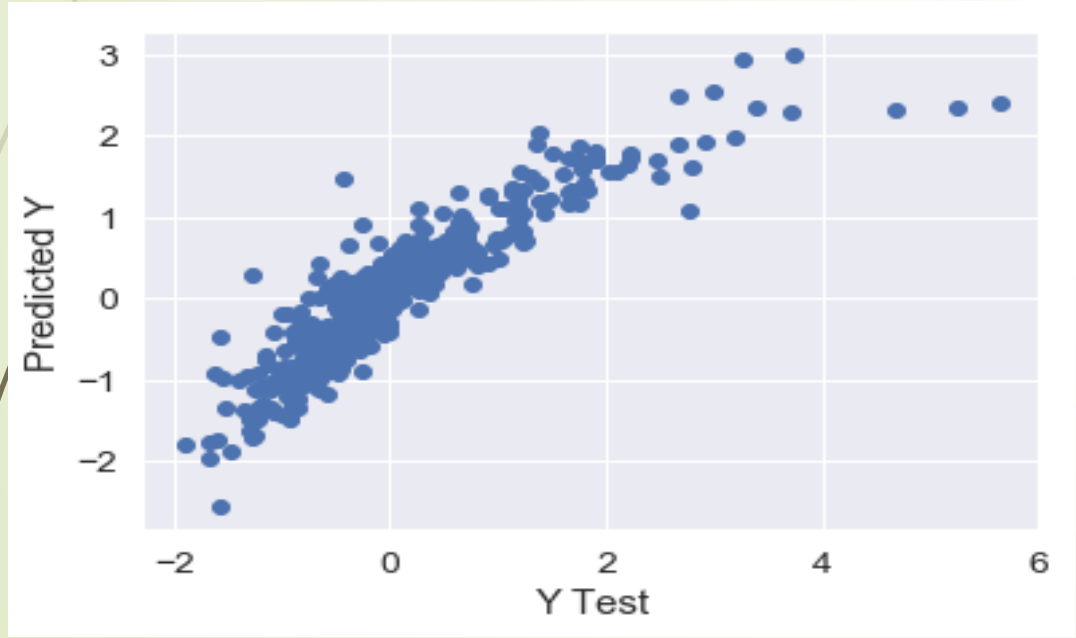
Machine Learning - Modelling & Prediction

Linear Regression Model

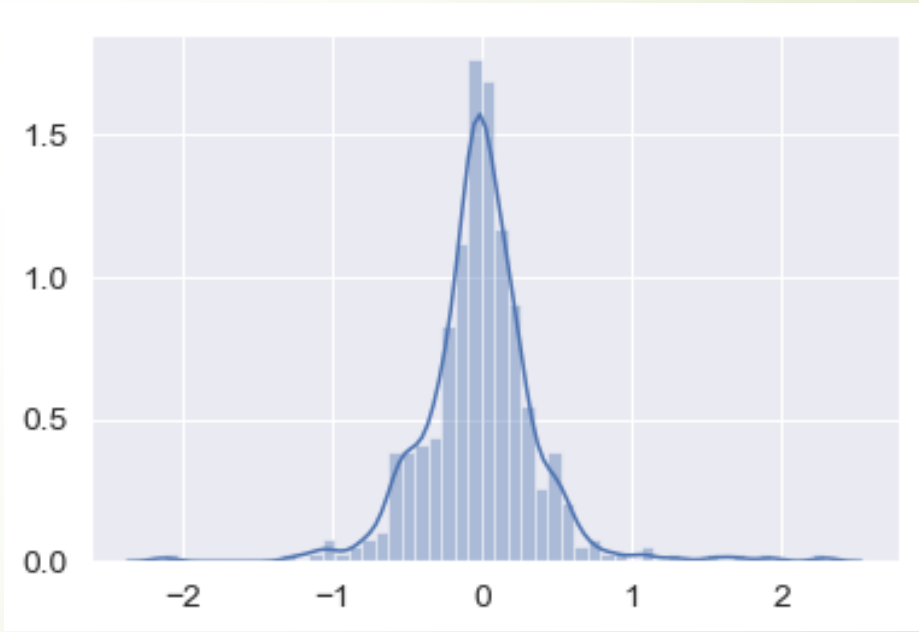
Intercept: 3.2093194297996494e-05

Coefficients: [0.31822793 0.39561078 0.03007994 0.09005157 0.1788636

0.05066635 -0.07629613 -0.02288886 0.10506373 0.08962047]



Residual



Model Evaluation

- MAE is 0.3017540260749489
- MSE is 0.2019160837529437 3
- RMSE is 0.4493507357876959 7

	Training	Test
MAE	0.292154854	0.301753
MSE	0.168192924	0.201916
RMSE	0.410113306	0.449351
R2 Score	0.823887855	0.817216

Random Forest Model

- Mean Absolute error for Random Forest Regresssion is: 0.26144220043434246
- Mean Square error for Random Forest Regresssion is: 0.15309651313803946
- Root Mean square error for Random Forest Regresssion is: 0.3912754951923765 3
- R2 score for Random Forest Regresssion is: 0.8614100396987187
- Accuracy --> 85.88777951198385
- From all the above analysis, we can conclude that Random Forest Regressor has better Sales Prediction than Linear Regression

Thank you