

NAME OF THE PROJECT

Submitted by:

ROHIT RAJAN

**INTRODUCTION**

* Business Problem Framing

From my understanding the problem is about making prediction of house prices for the company which is about to buy that property & to help make decision that whether it would be profitable or not to buy it on the basis of other variables given so that we could make this venture more profitable for the company.

* Conceptual Background of the Domain Problem

We have to study every feature’s behavious present in dataset

& make observation from its behaviour about how every

feature is giving the signs about that whether it would be

profitable to buy the property or not & building predictive

model on the basis of those features information to reduce the

loss & increase the profit of the company.

**Analytical Problem Framing**

• Origin of dataset & data types of every features

Dataset is provided by the company & we have to import it

using various libraries necessary for the project to get

completed. Also data types of features are both continuous

& categorical.

* Mathematical/ Analytical Modeling of the Problem

For visualization we only use four plots most of the times that were countplot, boxplot, distplot & scatterplot & for model building we use Linear Regression, lasso regression & Bagging Regression models to opt best out of them to work on dataset.

* Data Sources and their formats

• Assumptions related to problem statement No assumptions were made while working on the dataset.

• Libraries & Tools used We used numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle, scipy.stats, sklearn.metrics, sklearn.metrics,sklearn.linear ,scipy.special, sklearn.model\_selection

sklearn.preprocessing & warnings libraries for this task

* Hardware and Software Requirements and Tools Used

**Model/s Development and Evaluation**

STEPS TAKEN FOR THE TASK

1. Importing Libraries for the task Numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickles & warnings were imported for task to get completed.

2. Importing Dataset using libraries Imported the datasets using pandas library in jupyter notebook.

3. Checking Dimension of dataset By checking dimension of train dataset we get to know that it contains 1161 rows & 81 columns & for test dataset we get to know that it contains 292 rows & 80 columns.

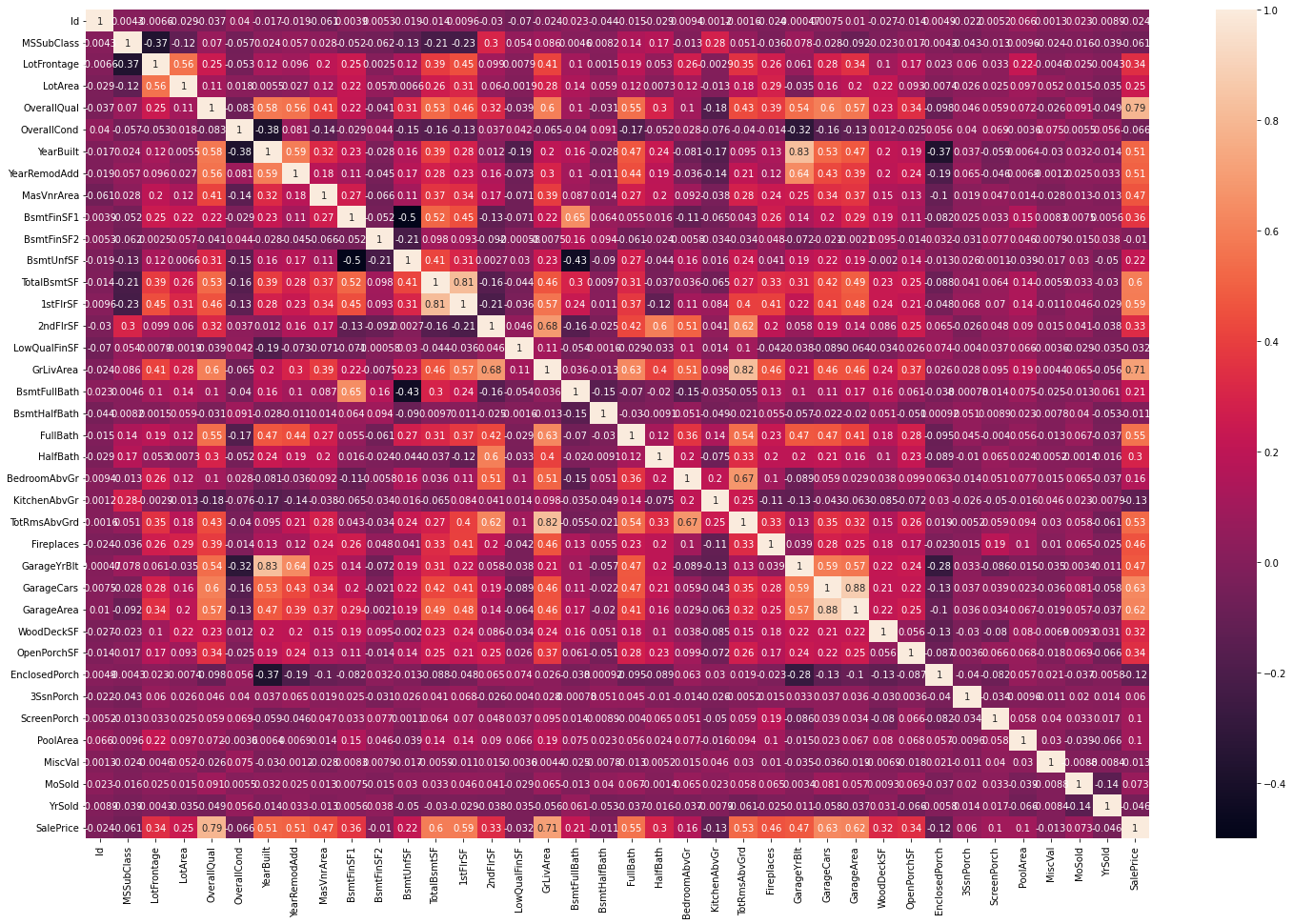
4. Checking Description of dataset From description we find mean, min value, max value, etc of every column which contains continuous data in them

5. Checking for presence of null values in dataset We checked for the presence of null values in every column of dataset by doing it repeatedly for every column as doing it repeatedly for every column & null values were present. Then we filled them randomly to avoid data bias.

6. Identifying Target variable By looking at dataset we identified target variable which is named SalePrice.

7. Performing EDA on whole Train dataset Id

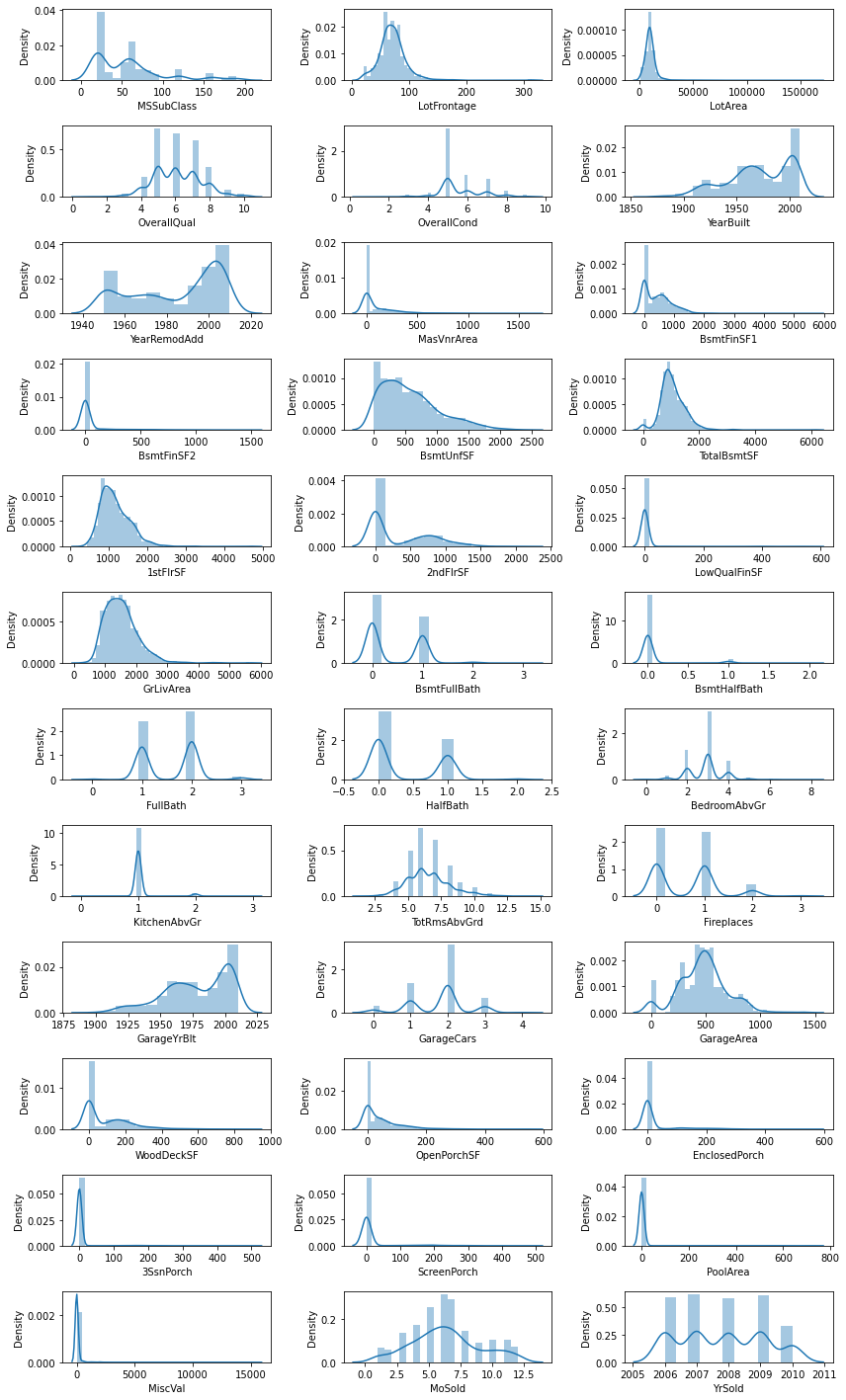
Making heatmap



Information:

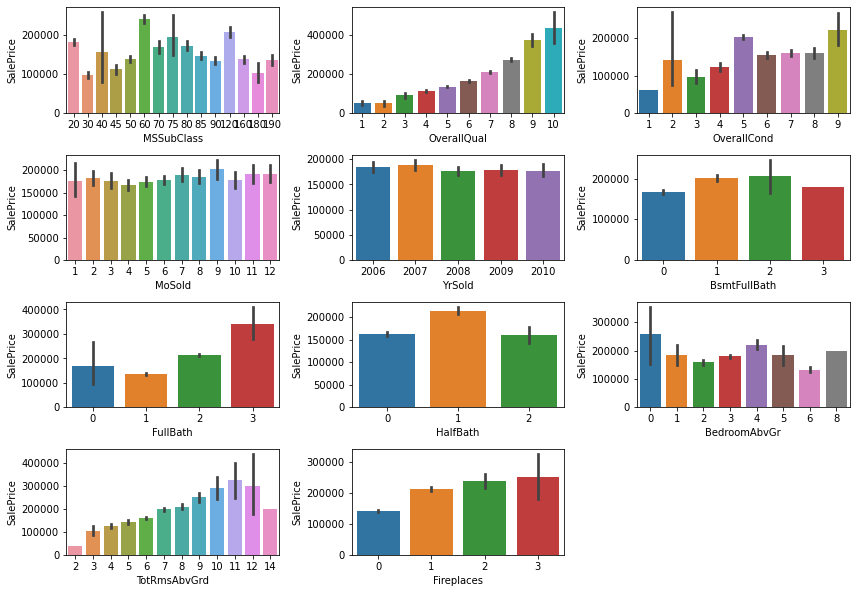
* Some features have high correlation with each other:
* GarageCars and GarageArea (0.88)
* GarageYrBlt and YearBuilt (0.83)
* TotRmsAbvGrd and GrLivArea (0.83)
* TotalBsmtSF and 1stflrSF (0.82)
* 1stflrSF and GrLivArea (0.69)
* TotRmsAbvGrd and BedroomAbvGrd (0.68)
* Will drop one feature from each of these pair data visualization.
* **Univariate Analysis**

Univariate analysis is **perhaps the simplest form of statistical analysis**. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved. Univariate analysis can yield misleading results in cases in which multivariate analysis is more appropriate



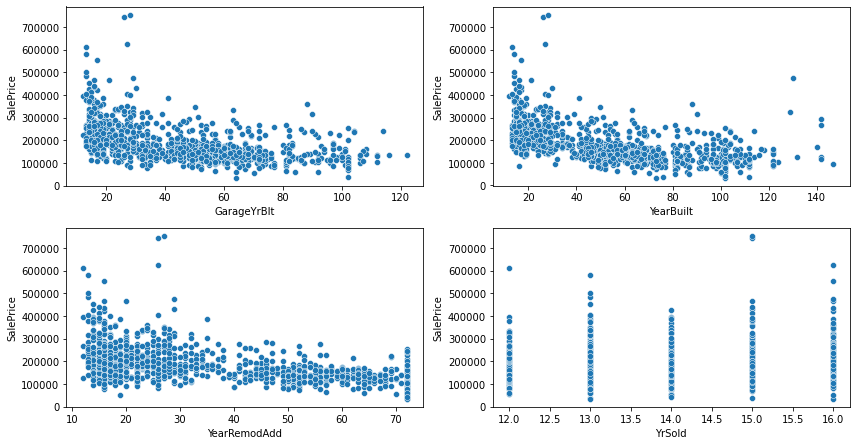
Information:

* 'YearBuilt', 'OverallCond', 'OverallQual', 'BsmtFullBath', 'BsmtHalfBath', 'HalfBath', 'FullBath', 'MoSold', 'Yrsold', 'BsmtQual', 'ExterCond','BsmtExposure', 'BsmtFinType2', 'TotRmsAbvGrd', 'GarageCond', 'GarageQual', 'KitchenAbvGr', 'LowQualFinSF', 'Fireplaces' are categorical.
* Some features have majority 1 unique values: BsmtFinSF2, LowQualFinSF, KitchenAbvGr, EnclosedPorch, PoolArea, ScreenPorch, 3SsnPorch, MiscVal. These will create class imbalance so better to remove after EDA.



Information:

* 'OverallQual' : More the rating of this feature, more the SalePrice (target variable)
* 'OverallCond' : SalePrice is highest for rating 5
* 'MoSold' and 'YrSold': SalePrice does not show a strong trend depending on month and year on which realty is sold
* 'FullBath' = 3 and 'HalfBath' = 1 have highest SalePrice
* 'TotRmsAbvGrd' : More the number of total rooms above grade more the Sale Price but after a certain value it decreases.



Information:

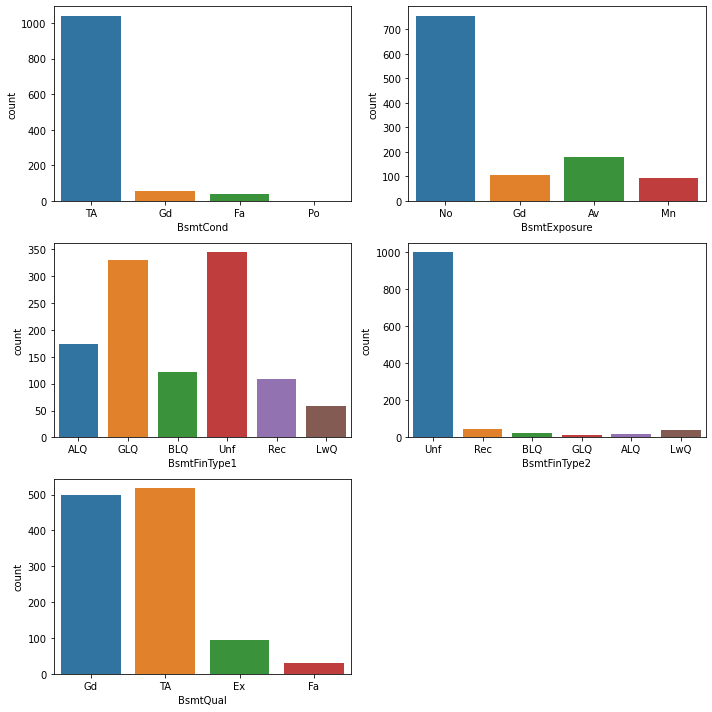
* For most the realty properties Garage is built within last 20-30 years, SalePrice is more for recently built garages
* SalePrice is more for recently build houses
* Recently remodelled houses (lower value of YearRemodAdd) have higher SalePrice
* YrSold still does not show any significant trend

### EDA on Categorical columns

1) Sorting the colums and values using python (sort) function

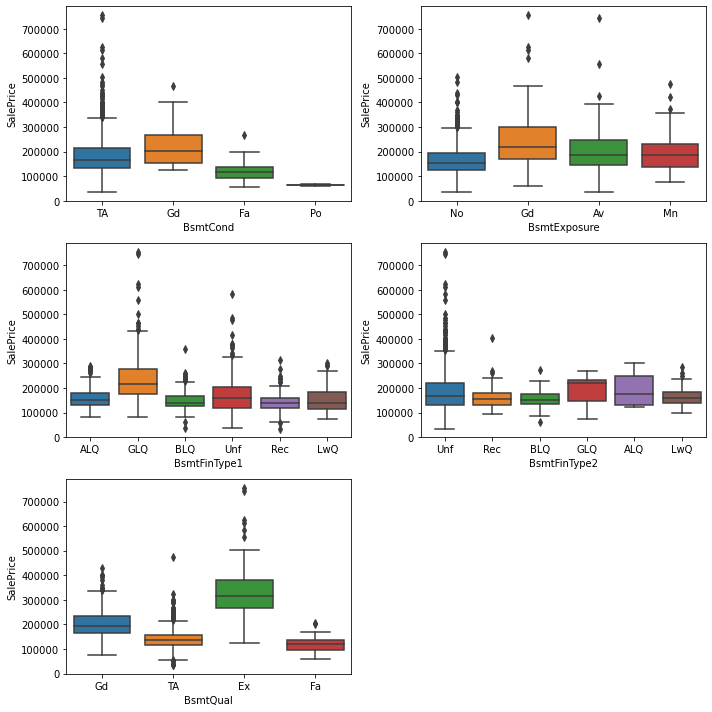
2) Dividing categorical features into sub classes

### Basement features



Information:

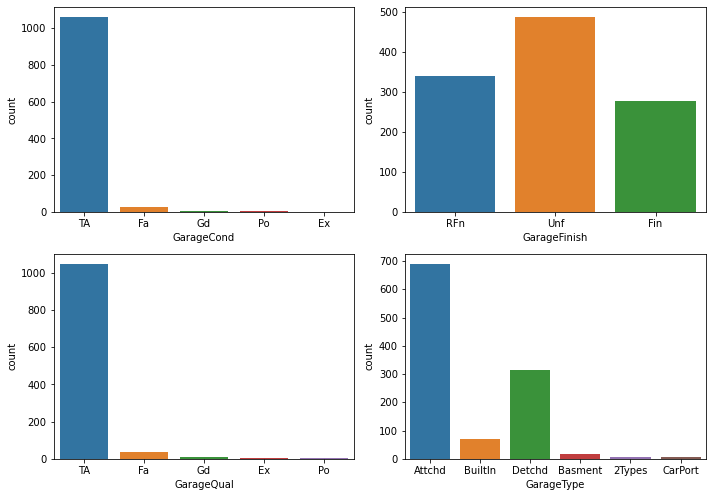
* BsmtCond in most houses are average with no basement exposure.
* Most hauses have 1 Good/ unfurnished basemet with second one as unfurnished.
* Basement quality has been mostly equal or above average.



Information:

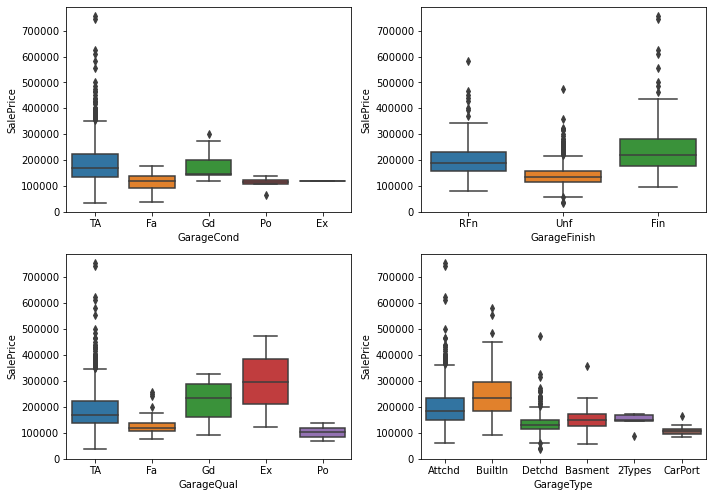
* Houses with good quality Bsmt Condition, Bsmt exposure, Bsmt Furnishing have high sale price .
* Houses with Excellent Basement height have high sale price.

### Garage features



Information:

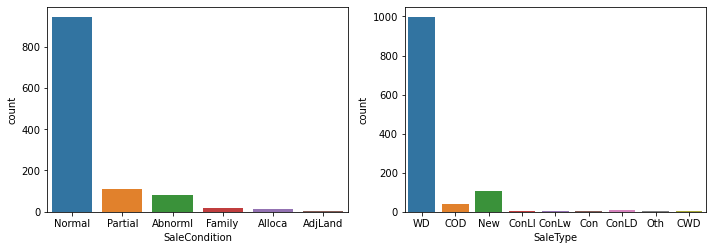
* Garage condition, garage quality in most houses are average with unfurnished garages.
* Most houses have attached garage type followed by deattached.



Information:

* Houses with good quality Bsmt Condition, Bsmt exposure, Bsmt Furnishing have high sale price .
* Houses with Excellent Basement height have high sale price.

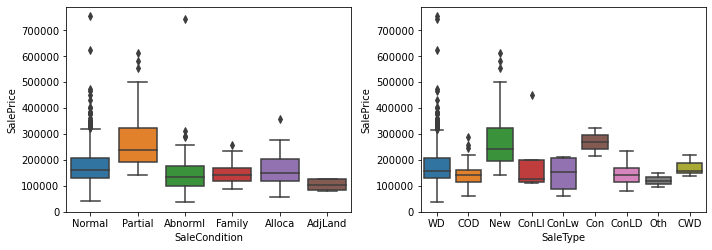
### Sale features



Information:

* Mostly houses have a normal sale condition, Sale type as WD:Warranty Deed - Conventional.

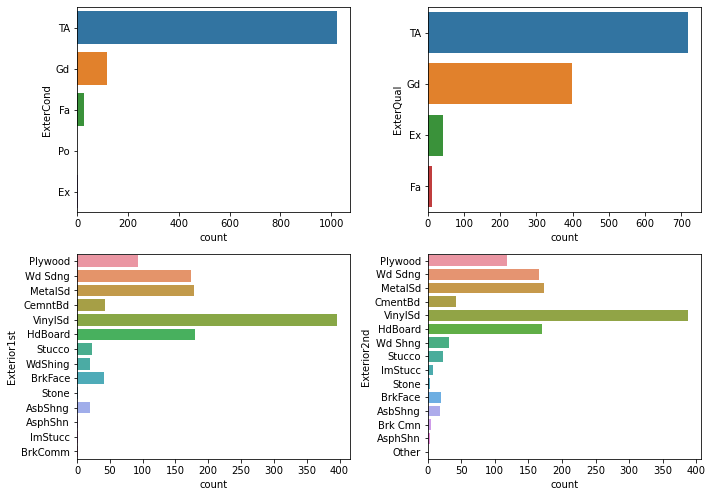




Information:

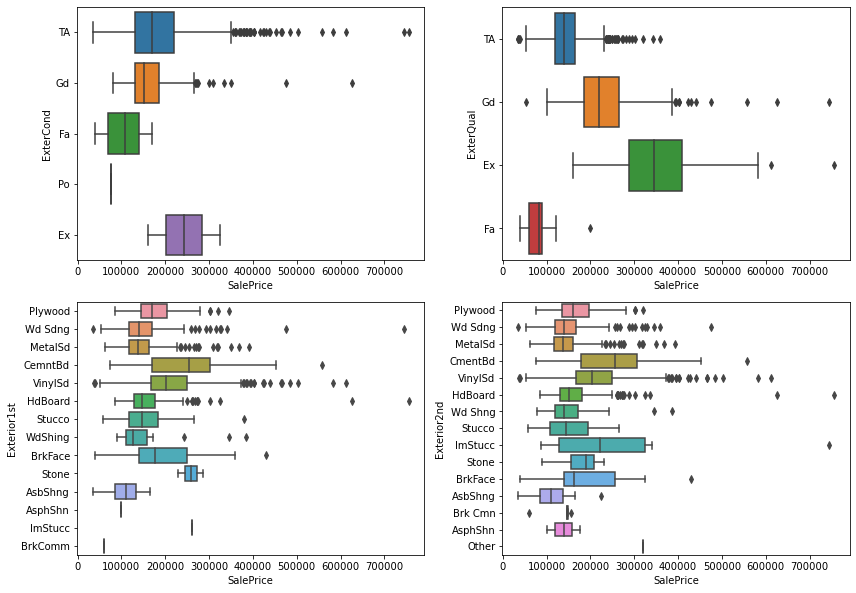
* Houses with Parial sale condition (new houses) and new Sale Type have high sale price.

### Exterior features



Information:

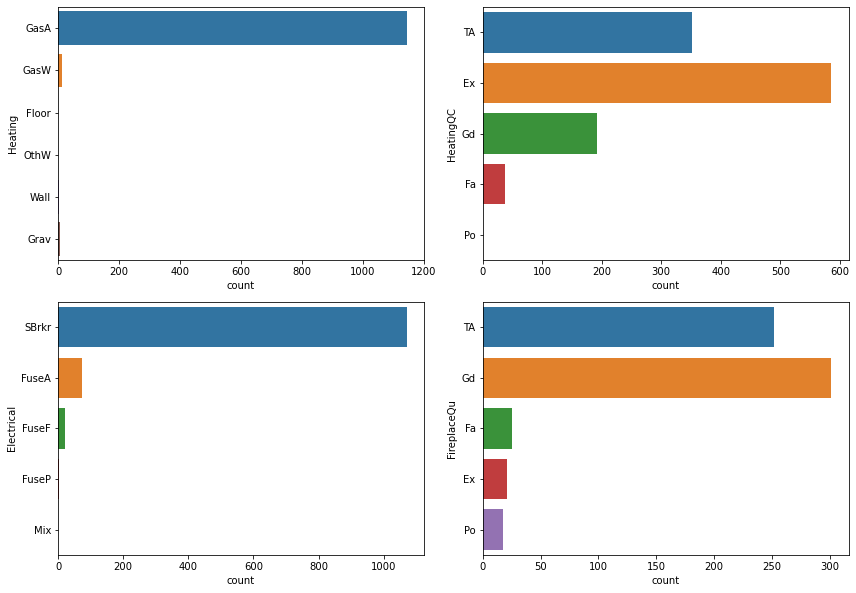
* Most houses have average exterior conditions and quality.
* Vinyl Siding is mostly used in houses



Information:

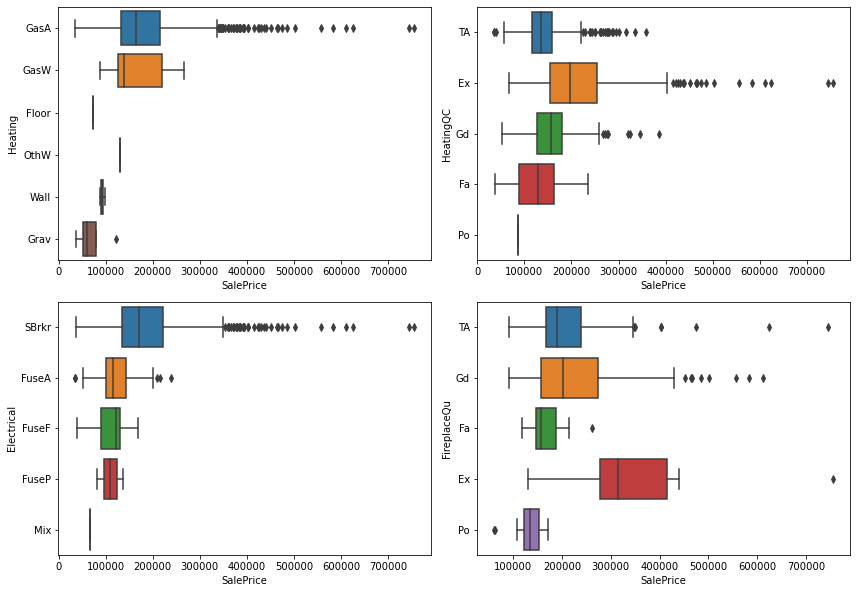
* Houses having excellent exterior conditions and quality have generally higher selling price.
* Houses with Exterior covering made of Cement Board have high selling price.

### Interior features



Information:

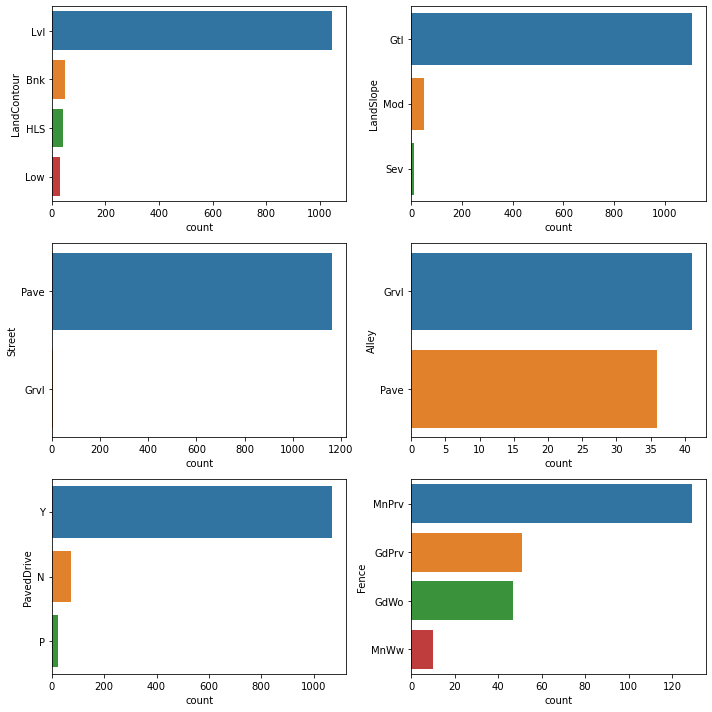
* Most houses have Gas forced warm air furnace and excellent heating quality with good fireplace.
* Most houses have Standard Circuit Breakers.



Information:

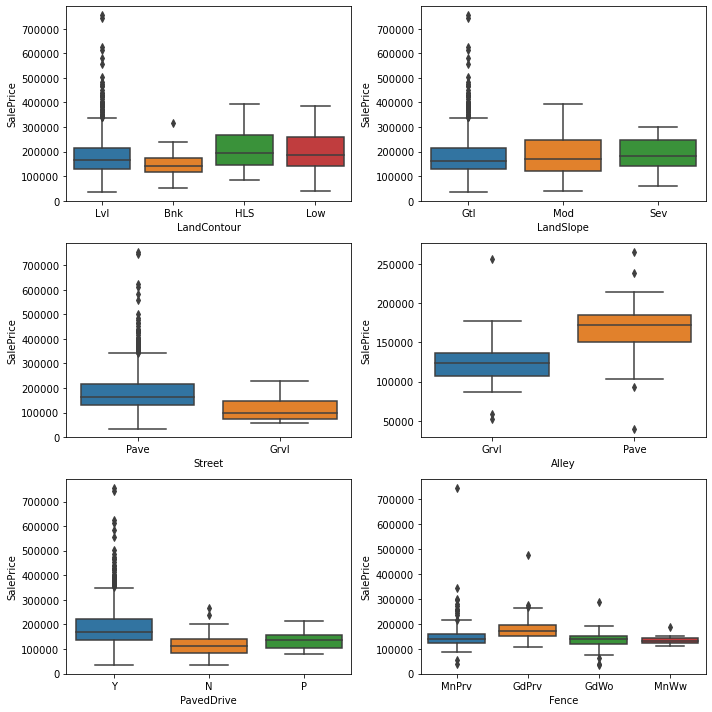
* Houses having Gas hot water or steam heat and excellent heating quality have generally higher selling price.
* Houses having Standard Circuit Breakers, good fireplace have higher selling price.

### Land features



Information:

* Most houses have level land contours with gentle land slope. Streets having pave are highly common.
* Paved drive are very common. Fences with min privacy are mostly there in houses.



Information:

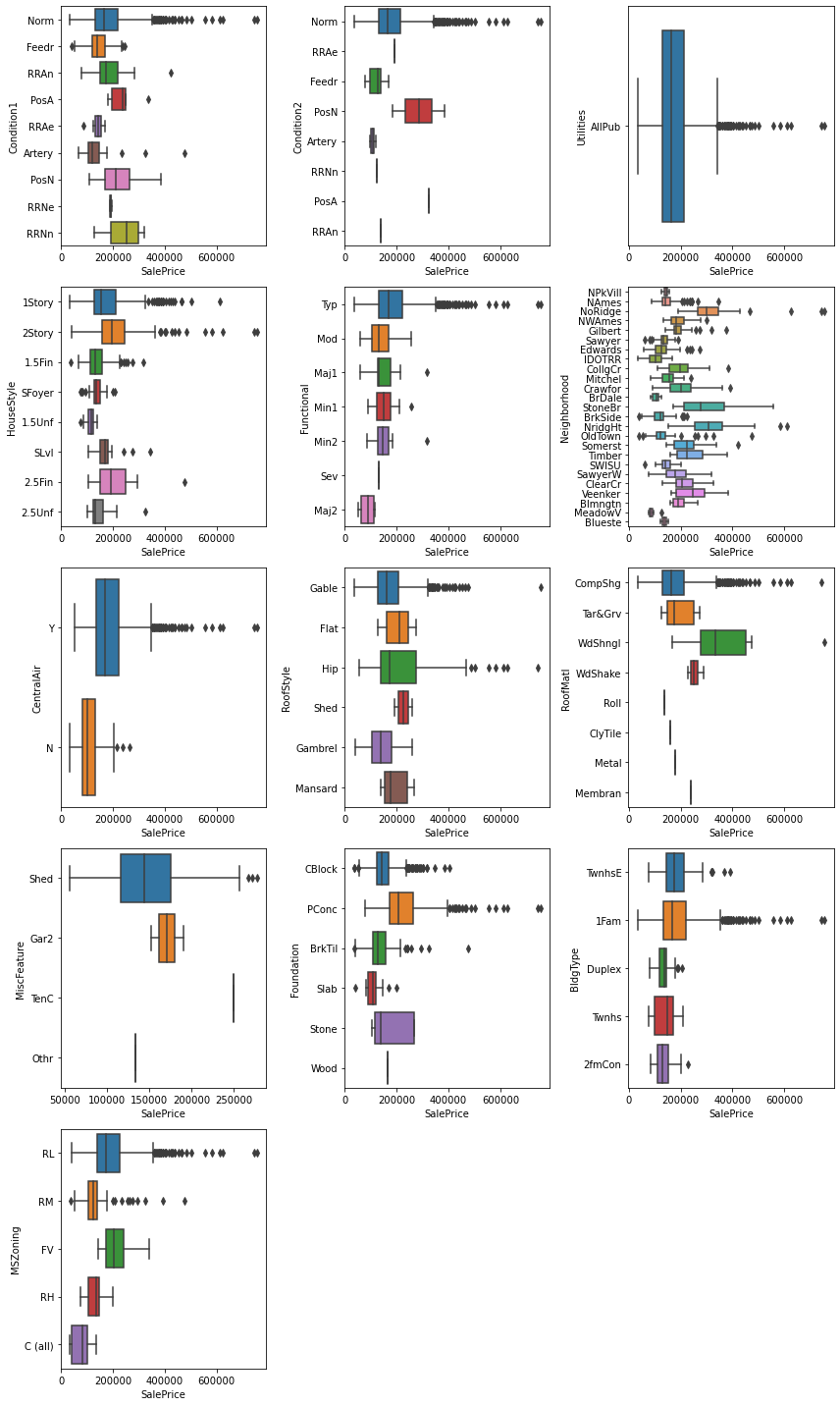
* Houses having level land contours and modelerate land slope have generally higher selling price.
* Houses with street having pave and alley with pave have high selling price.
* Houses having paved drive with good privacy have higher selling price.

### Rest Features



Information:

* Most houses have normal conditions with All public Utilities.
* 1 story house style are most common . Most of them have central AC.
* Shed is most common misc feature in houses. Gable and CompShg are most common under roof style and matterial
* Most homes are located near Northwest Ames and College creek.



Information:

* Houses having 2 Story with have generally higher selling price.
* Houses with central AC, hip style of roof with Wood Shingles have high selling price.
* Houses with StoneBr, NridgHt as neighbourhood have high selling price.

### Featues selections and Handling missing values

Removing one of the high correlated from pairs checked using heatmap

Thus we dropped ('GarageCars', 'GarageYrBlt', 'GrLivArea', 'TotalBsmtSF', 'BedroomAbvGr')from columns

Removing numerical features having 1 unique value as occuring more than 95%

### Checking % of missing values

1 Since MasVnrArea has only 1% data missing, dropping rows with NULL values in MasVnrArea

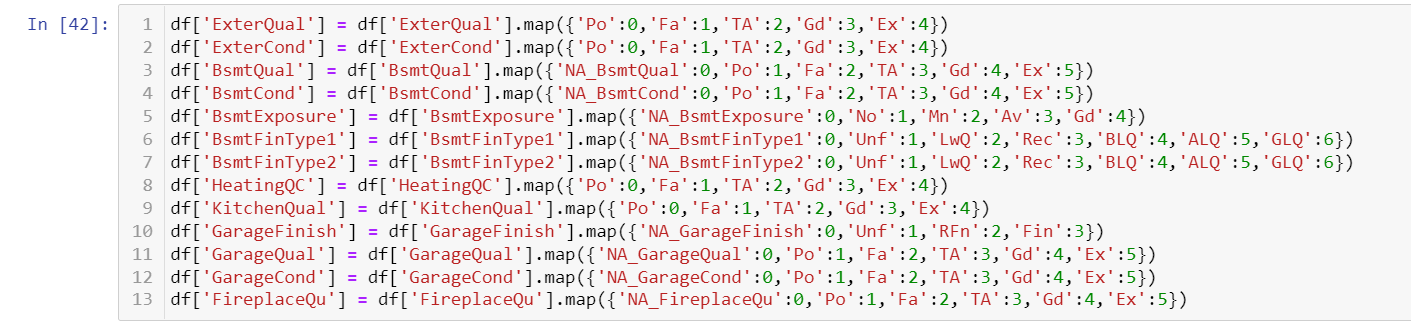
2 Dropping Id column as it does not contribute towards predicting SalePrice

3 Dropping 'PoolQC' for very high percentage of missing value and highly imbalance data (if missing value is imputed)

4 Some categorical features have NAN values which denotes to a particular class. By using the fillna method we will fill all the NAN values

* 2 types of categorical features are present:
* Encoding For Categorical Variables Ordered Features: to be label encoded: 'ExterQual', 'ExterCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'HeatingQC', 'KitchenQual','Functional','FireplaceQu', 'GarageFinish', 'GarageQual', 'GarageCond', 'CentralAir'
* Unordered Features: to be one hot encoded: 'MSZoning', 'Street', 'Alley', 'LandContour', 'LotConfig', 'Neighborhood', 'Condition1' , 'Condition2', 'BldgType', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'Foundation', 'Heating', 'Electrical', 'GarageType','PavedDrive', 'Fence','MiscFeature', 'SaleType','SaleCondition', 'LotShape', 'Utilities', 'LandSlope', 'HouseStyle'

### Encoding Ordinal Categorical Features



Information:

* For 'Alley', Nan means 'No access to alley'
* For 'BsmtQual', 'BsmtCond', BsmtExposure, BsmtFinType1, BsmtFinType2 Nan means 'No basement'
* For GarageType, GarageFinish, GarageQual, GarageCond Nan means 'No garage'
* For 'FireplaceQu' and 'Fence' Nan means 'No Fire place' and 'No fence' respectively
* MiscFeature - Nan means no additional features mentioned.
* All these features can be imputed by making them into one category in place of missing data.

In [43]:

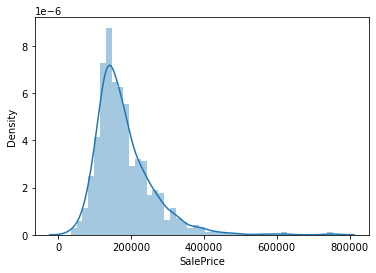
### Encoding on Nominal features

Encoding is **the process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage**.

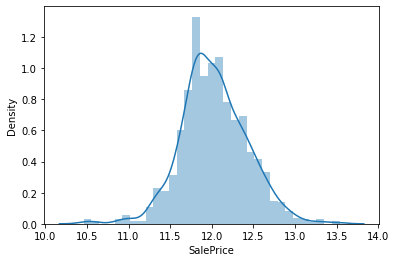
A dummy variable is **a binary variable that indicates whether a separate categorical variable takes on a specific value**.

Explanation: As you can see dummy variables are created for the categorical values of the temperature attribute. We can create dummy variables in python using get\_dummies() method.

After using it we are adding the dummy variables to the original dataframe. Then dropping the redundant columns.



### Log transformation to removing skewness[¶](http://localhost:8888/notebooks/Untitled%20Folder%202/Housing%20Project/housing%20prediction.ipynb#Log-transformation-to-removing-skewness)



### Train Test Split

From sklearn.model\_selection import train\_test\_split

You can split the dataset into train and test set **using the train\_test\_split() method of the sklearn library**. It accepts one mandatory parameter. – Input Dataset – It is a sequence of array-like objects of the same size. Allowed inputs are lists, NumPy arrays, scipy-sparse matrices, or pandas data frames.

### Outlier Treatment

We tend to delete the outlier if they are due to data entry errors caused due to human error, data processing errors.

Information:

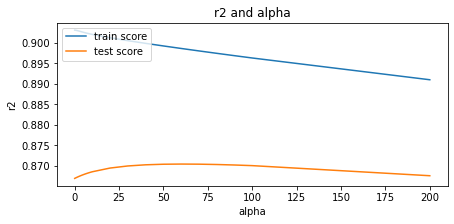
* Many features have outliers
* Dropping all the outliers will cause loss of information.
* Hence reassigning fixed minimum and maximum values to those rows where feature value is outside the range of [25th percentile - 1.5 IQR, 75th percentile + 1.5 IQR]
* IQR or Inter Quartile Range = Difference between 75th percentile and 25th percentile values of a feature.
* Target column 'SalePrice' is excluded in this. Some other features are also excluded since those are ordered categorical type which are labelled encoded to numeric form.

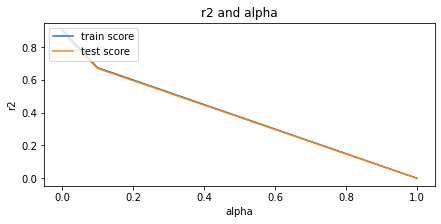
### Scaling[¶](http://localhost:8888/notebooks/Untitled%20Folder%202/Housing%20Project/housing%20prediction.ipynb#Scaling)

It is **a step of Data Pre Processing that is applied to independent variables or features of data**. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm. Package Used: sklearn.preprocessing.

### Modelling, hyperparams tuning for ridge and lasso, evalution

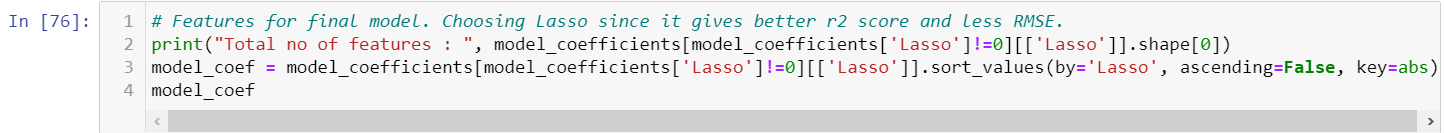
**Choosing a set of optimal hyperparameters for a learning algorithm**. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

Lasso regression is **a type of linear regression that uses shrinkage**. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).



**Model/s Development and Evaluation**

Features for final model. Choosing Lasso since it gives better r2 score and less RMSE.



Using **regression equations** to make predictions. Regression equations are a crucial part of the statistical output after you fit a model. The coefficients in the equation define the relationship between each independent variable and the dependent variable.

### Model Building with doubling the value of lambda

Ridge regression and Lasso regression are two popular techniques that make use of regularization for predicting. Both the techniques work by **penalizing the magnitude of coefficients of features along with minimizing the error between predictions and actual values or records**.

Saving Model

After making prediction we saved the model using pickle library.

CONCLUSION

In the end we conclude that model is quite accurate in predicting the price of houses & price is getting affected positively because of the following variables (OverallQual, YearBuilt, YearRemodAdd, ExterQual,