

HOUSING PRICE PREDICTION

**Submitted By:**

**Meet Mehta**

ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot.

Some of the reference sources are as follows:

* Medium.com
* StackOverflow

# INTRODUCTION

BUSINESS PROBLEM FRAMING

This is a real estate issue where a US based real estate company called Surprise Housing decided to invest in the Australian market. Their plan is to buy homes in Australia at prices below their actual market value and sell them at high prices for a profit. To do this, this company uses data analysis to decide which property to invest in.

The ​company has collected data from homes already sold in Australia and using this data they want to know the value of potential properties to decide whether it will be appropriate to invest in the properties or not.

## To know the value of properties, the company provided us with data to perform data analysis and extract information on important attributes to predict house prices. They want a machine learning model that can predict house prices as well as the meaning of each important attribute in house prediction - how and with what intensity each variable affects house prices.

## Conceptual Background OF the Domain Problem

In real estate the value of property usually increases with time as seen in many countries. One of the causes for this is due to rising population.

The value of property also depends on the proximity of the property, its size its neighborhood and audience for which the property is subjected to be sold. For example, if audience is mainly concerned of commercial purpose. Then the property which is located in densely populated area will be sold very fast and at high prices compared to the one located at remote place. Similarly, if audience is concerned only on living place, then property with less dense area having large area with all services will be sold at higher prices.

The company is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

## Review OF Literature

Houses are one of the necessary needs of each person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price.

We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

## Motivation OF The Problem Undertaken

To understand real world problems where Machine Learning and Data Analysis can be applied to help organizations in various domains to make better decisions with the help of which they can gain profit or can be escaped from any loss which otherwise could be possible without the study of data. One of such domains is Real Estate.

Houses are one of the necessary needs of each person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

# Analytical Problem Framing

## Mathematical/ Analytical Modeling OF the Problem

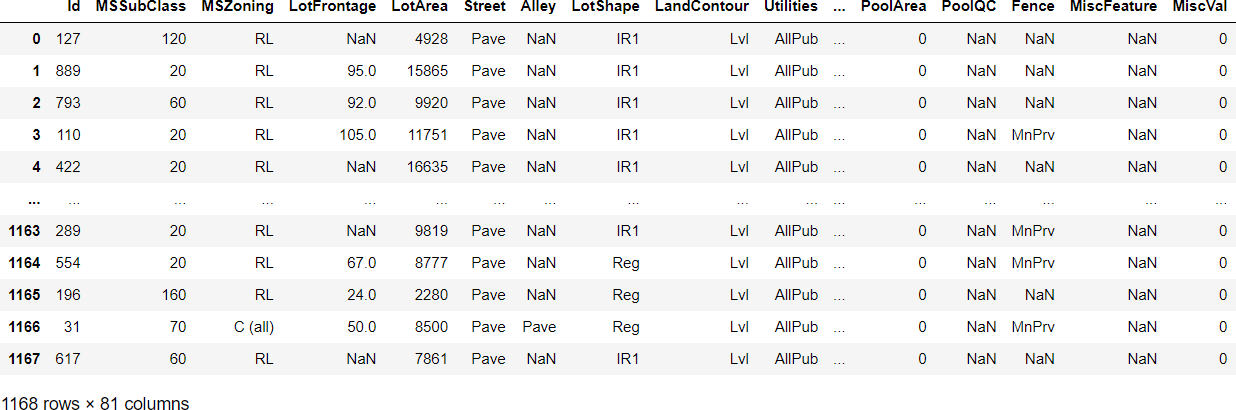
This particular problem contains two datasets train dataset and test dataset. Model are built using train dataset. This model is then used to predict the Sale Price for test dataset. By analyzing into the target column. After analysis it was concluded that the data entries of sale Price column contains data points of continuous nature, it is a Regression problem, hence all regression algorithms were used while building the model. While checking for the null values in the datasets, many columns with nan values were found and null values were replaced with suitable entries like mean for numerical columns and specific value for categorical columns. For further analyses graph plot like distribution plot, bar plot, reg plot and scatter plot were used. With these plots, the relation between the feature columns and target column was visualized. Upon analyzing outliers and skewness were found in the dataset and were removed. Outliers were removed using percentile method and Skewness using Yeo-Johnson method. All the regression models were iterated to find the best model and then further Hyper-tune the best model and save the best model. Finally, Sale Price was predicted for test dataset using the saved model built from train dataset.

## Data Sources & Data Formats

The data was provided in csv (comma separated values) format.

It contained two datasets train and test dataset. Model was built using train dataset and then used to predict sale Price for test dataset. Train dataset contains 1168 rows and 81 columns including target variable, and test dataset was having 292 rows and 80 columns excluding target variable. Dataset’s columns have data types objects, float and integer type s of data.

Dataset was imported using Panda’s library and then transformed into data-frame.

Train Dataset

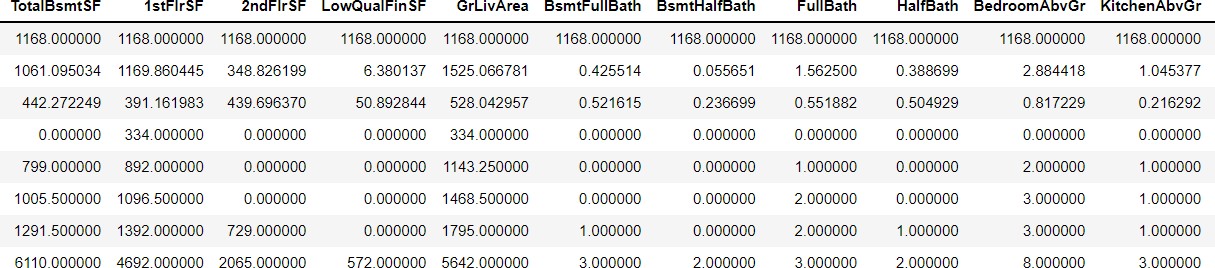
## Data Pre-processing

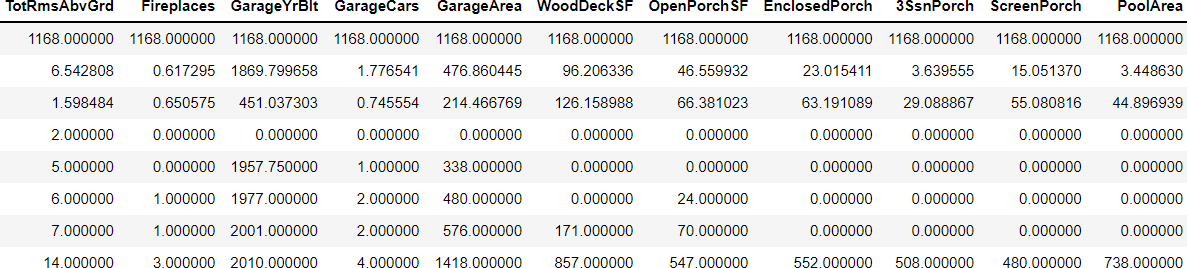
Current dataset is raw data. By proper Data Transformation methods, a lot of valuable insights can be gained. Also, we need to convert categorical info into numerical data type. Categorical columns having categorical data point needs to be assigned a unique integer data point.

In ID and Utilities column the unique counts were 1168 and 1 resp, which concludes that all the data entries in ID column is unique. ID is the unique identity number given to every asset. Utilities column contained only one data point in whole dataset, hence these two columns were dropped.

In this project we have performed various mathematical and statistical analysis such as description or statistical summary of the data using describe, checked correlation using .corr() and also visualized it using heatmap. Then we have used Z-Score to plot outliers and remove them.









From this statistical analysis we make some of the interpretations that,

* + Maximum standard deviation of 8957.44 is observed in LotArea column.
  + Maximum SalePrice of a house observed is 755000 and minimum is 34900.
  + In the columns Id, MSSubclass, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfsF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, HalfBath, TotRmsAbvGrd, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, Miscval, salePrice mean is considerably greater than median so the columns are positively skewed.
  + In the columns FullBath, BedroomAbvGr, Fireplaces, Garagecars, GarageArea, YrSold Median is greater than mean so the columns are negatively skewed.

In the columns Id, MSSubClass, LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtHalfBath, BedroomAbvGr, ToRmsAbvGrd, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, SalePrice there is considerable difference between the 75 percentile and maximum, so outliers are present.

## Column Description

The variable features of this problem statement are as: MSSubClass: Identifies the type of dwelling involved in the sale MSZoning: Identifies the general zoning classification of the sale LotFrontage: Linear feet of street connected to property LotArea: Lot size in square feet

Street: Type of road access to property Alley: Type of alley access to property LotShape: General shape of property LandContour: Flatness of the property Utilities: Type of utilities available LotConfig: Lot configuration LandSlope: Slope of property

Neighborhood: Physical locations within Ames city limits Condition1: Proximity to various conditions

Condition2: Proximity to various conditions (if more than one is present) BldgType: Type of dwelling

HouseStyle: Style of dwelling

OverallQual: Rates the overall material and finish of the house OverallCond: Rates the overall condition of the house YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions) RoofStyle: Type of roof

RoofMatl: Roof material

Exterior1st: Exterior covering on house

Exterior2nd: Exterior covering on house (if more than one material) MasVnrType: Masonry veneer type

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior ExterCond: Evaluates the present condition of the material on the exterior Foundation: Type of foundation

BsmtQual: Evaluates the height of the basement BsmtCond: Evaluates the general condition of the basement BsmtExposure: Refers to walkout or garden level walls BsmtFinType1: Rating of basement finished area BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types) BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area TotalBsmtSF: Total square feet of basement area Heating: Type of heating

HeatingQC: Heating quality and condition CentralAir: Central air conditioning

Electrical: Electrical system

1stFlrSF: First Floor square feet 2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors) GrLivArea: Above grade (ground) living area square feet BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms FullBath: Full bathrooms above grade HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms) Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms) Functional: Home functionality (Assume typical unless deductions are warranted) Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality GarageType: Garage location GarageYrBlt: Year garage was built GarageFinish: Interior finish of the garage GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet GarageQual: Garage quality GarageCond: Garage condition

PavedDrive: Paved driveway

WoodDeckSF: Wood deck area in square feet OpenPorchSF: Open porch area in square feet EnclosedPorch: Enclosed porch area in square feet 3SsnPorch: Three season porch area in square feet ScreenPorch: Screen porch area in square feet PoolArea: Pool area in square feet

PoolQC: Pool quality Fence: Fence quality

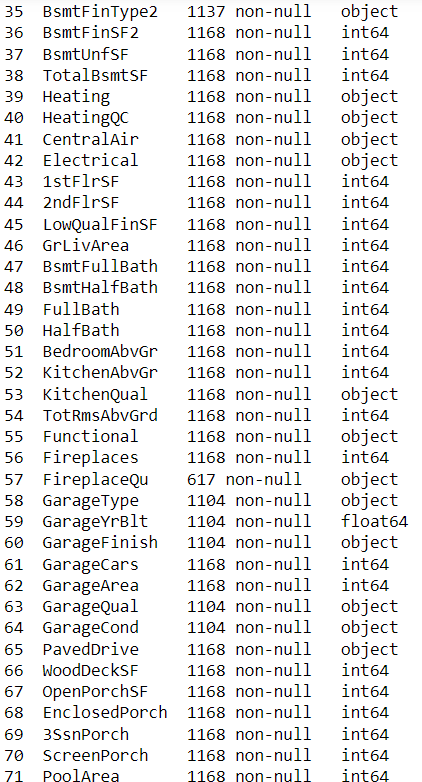
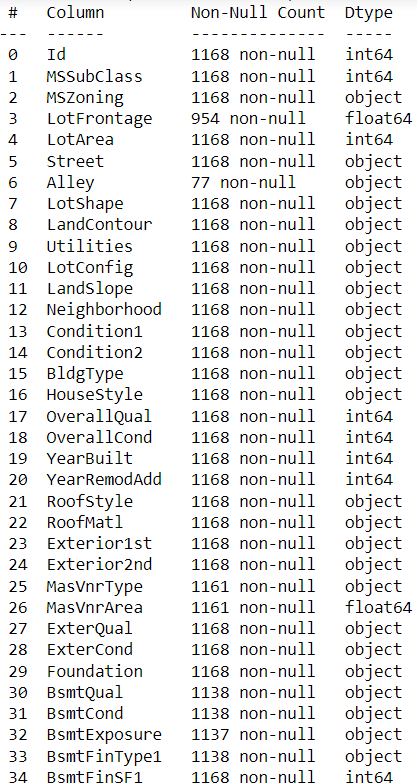
MiscFeature: Miscellaneous feature not covered in other categories MiscVal: $Value of miscellaneous feature

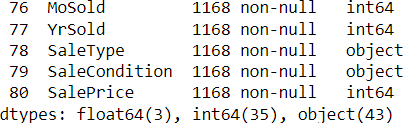
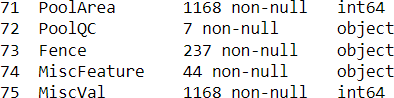
MoSold: Month Sold (MM) YrSold: Year Sold (YYYY) SaleType: Type of sale SaleCondition: Condition of sale

## Data Inputs-Logic-Output Relationship

SalePrice is our target variable i.e., output column. Rest all columns are to be used as feature column i.e., input column. We need to find which feature columns have positive correlation and which have negative correlation, accordingly to train our model. Columns which have no correlation have to dropped.

Execute command info to obtain descriptive summary of the train dataset





These 81 columns comprise of both dimensions (categorical value) and measures (numeric value)

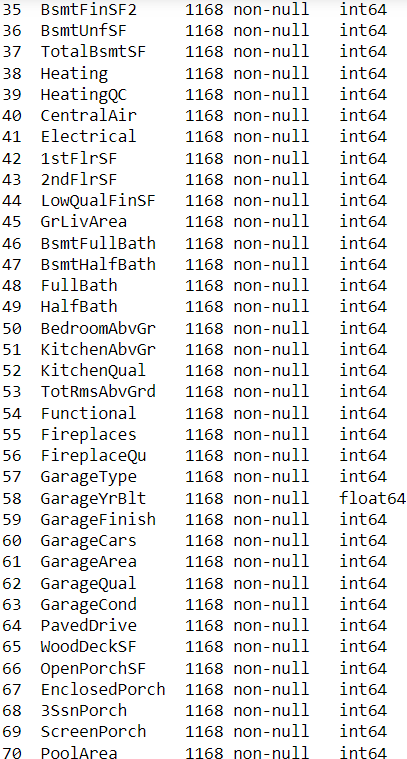
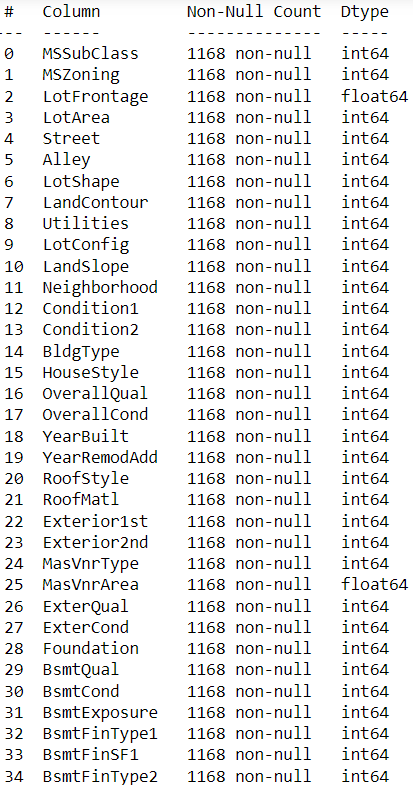
The dataset is not clean, i.e., consists of missing values as well

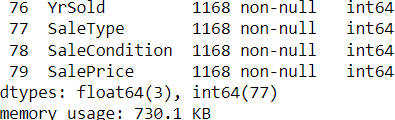
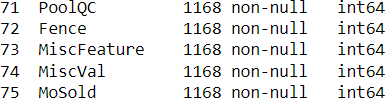
We can fill null values of categorical columns with 0 as it basically represents that, the feature

is not available for the property.

For missing values in numerical column, we can fill them mean of the resp column. Next step is to assign every categorical data a unique value.

Verify whether all categorical columns are converted into numerical column using info command.





## Hardware and Software Requirements and Tools Used

Hardware required:

* + - Processor — core i5 and above
    - RAM — 8 GB or above
    - SSD — 250GB or above

Software/s required: Anaconda

LIBRARIES:

The tools, libraries, and packages we used for accomplishing this project are pandas, numpy, matplotlib, seaborn, scipy, sklearn, mlxtend, xgboost, joblib.

Through pandas library we loaded our csv file ‘Data file’ into dataframe and performed

data manipulation and analysis.

With the help of numpy we worked with arrays.

With the help of matplotlib and seaborn we did plot various graphs and figures and done data visualization.

Train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train\_test\_split will make random partitions for the two subsets.

With sklearn’s standardscaler package we scaled all the feature variables onto single scale. As these columns are different in scale, they are standardized to have common scale while building machine learning model. This is useful when you want to compare data that correspond to different units.

With sklearn’s package we imported many regression models, we could obtain cross\_val\_score which is an accuracy metric used to evaluate model, we could obtain best parameters of a model using GridSearchCV or RandomizedSearchCV, we could reduce skewness using power transform library of sklearn.

With mlxtend package we could stack low performing models to obtain a high accuracy model.

# Model/s Development and Evaluation

## Identification of possible problem-solving approaches

Null values of numerical columns can be filled by replacing null values with mean of respective column. Null values can of categorical column can be either replaced by using mode value or if it’s for a column having ordinal datapoint, we can perform ordinal encoding.

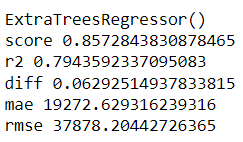
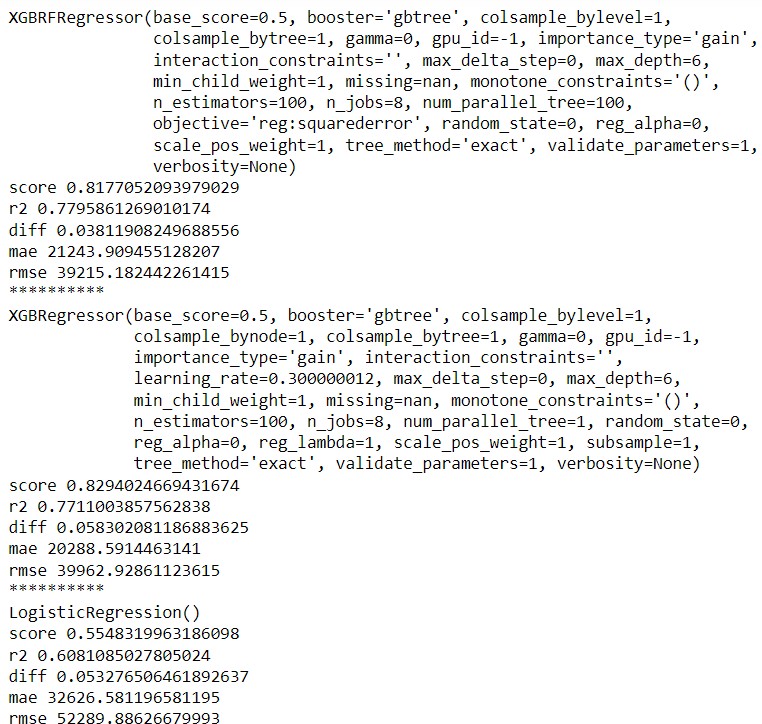
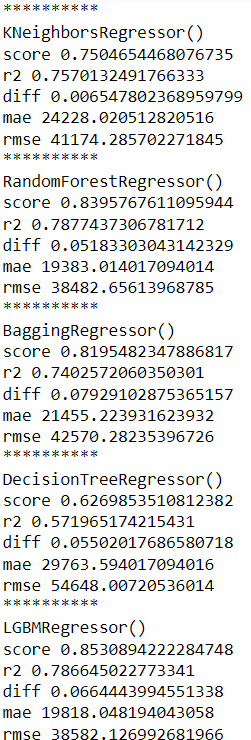
If outliers are present, we shall remove them using Z-Score, IQR method or by percentile method. If skewness exists, we shall remove them using Yeo-Johnson method.

If models have low accuracy, we shall fine tune them to improve accuracy but if accuracy is still low then we shall stack up our top performing models to boost accuracy by combining models.

## Testing of Identified Approaches (Algorithms)

We can check null values using info function. Outliers can be detected using Boxplot. Skewness can be detected using skew function. Our target variable is SalePrice which has datapoints of continous in nature, hence it is a regression problem. For that we shall use all regression algorithms to find & build the optimized model. By looking into the difference of r2 score and cross validation score of each model we can find our best model with least difference. To get the best model we have to run through multiple models and to avoid the confusion of overfitting we have go through cross validation.

## Run & evaluate selected models



Key Metrics for success in solving problem under consideration

Following metrics were used to evaluate our model:

--- Cross Val Score

--- R2 Score

--- Mean absolute error

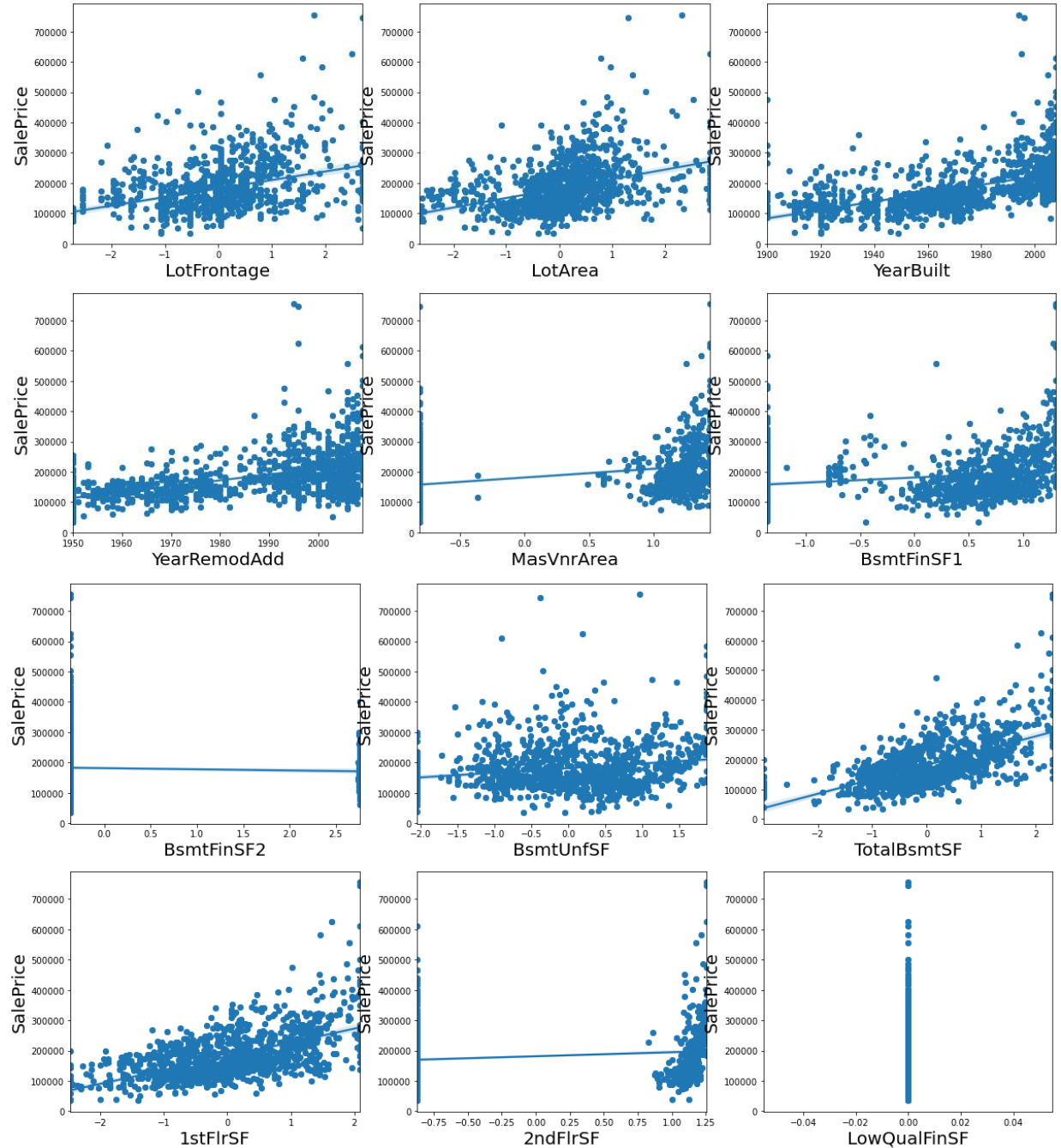
--- Mean squared error

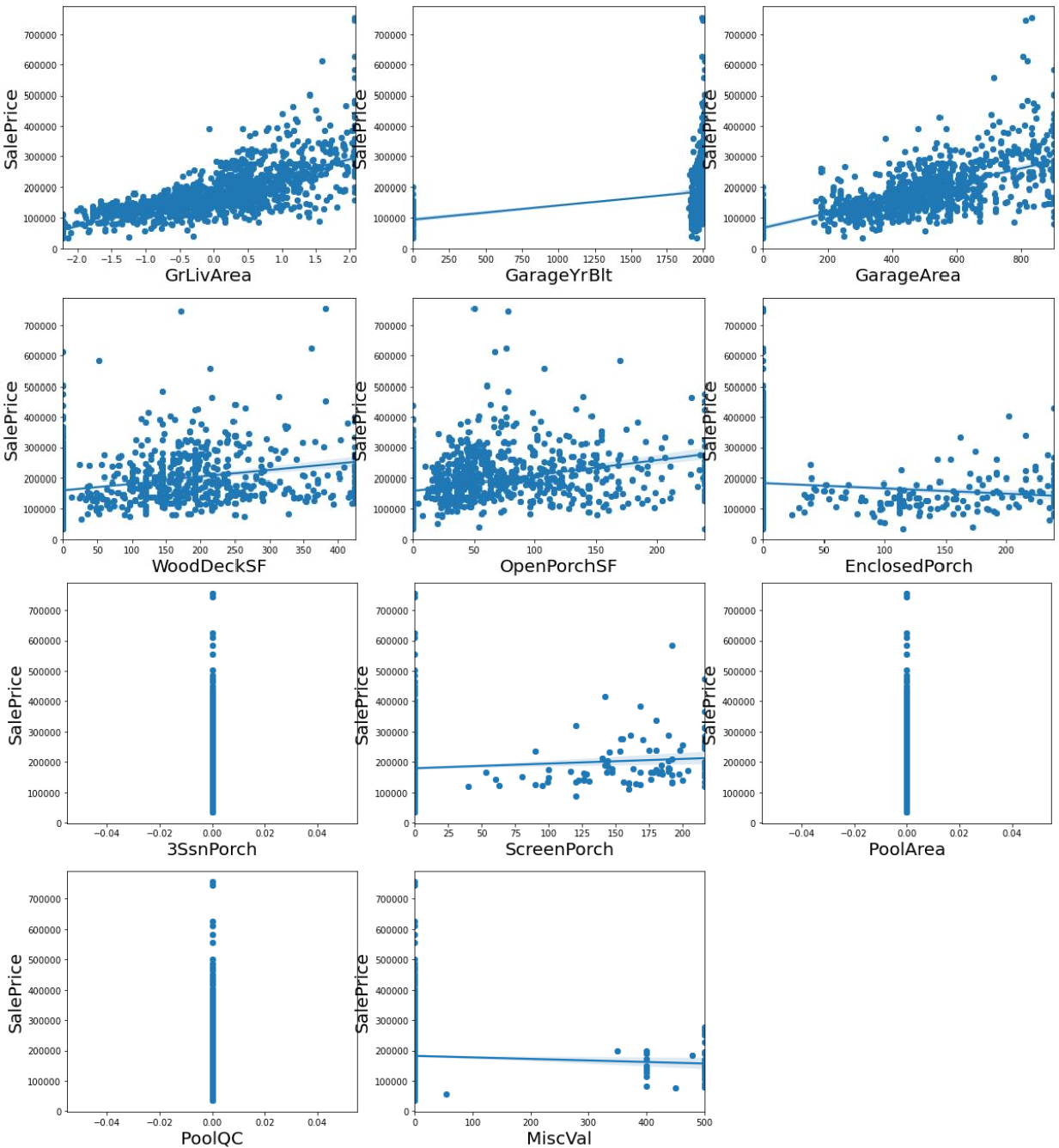
--- Standard deviation error

# Visualization

## Scatter Plot & Regression Plot

Plot graph for all columns having datapoints of continuous nature w.r.t target variable to find the relation between the feature column and target variable.



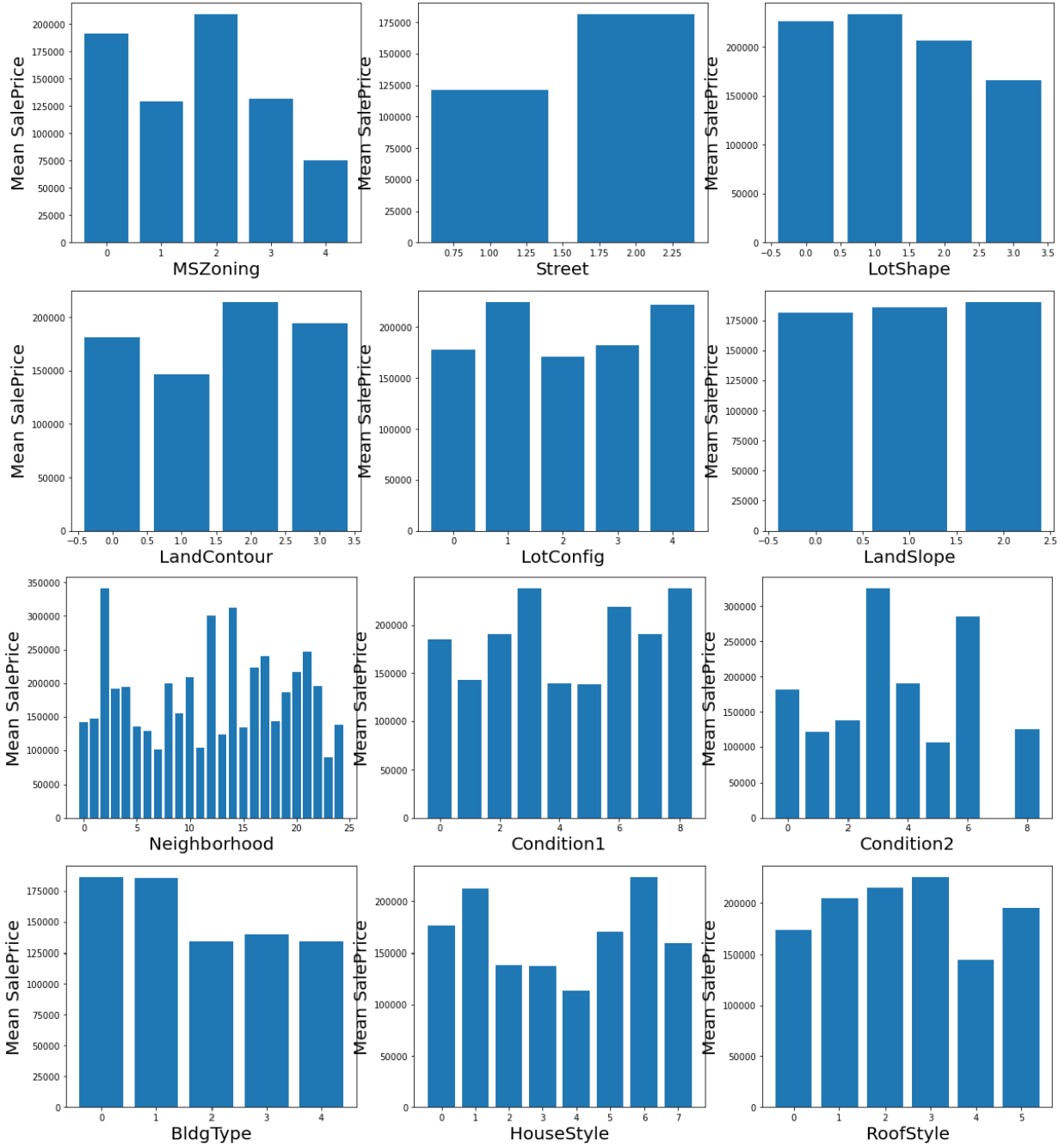


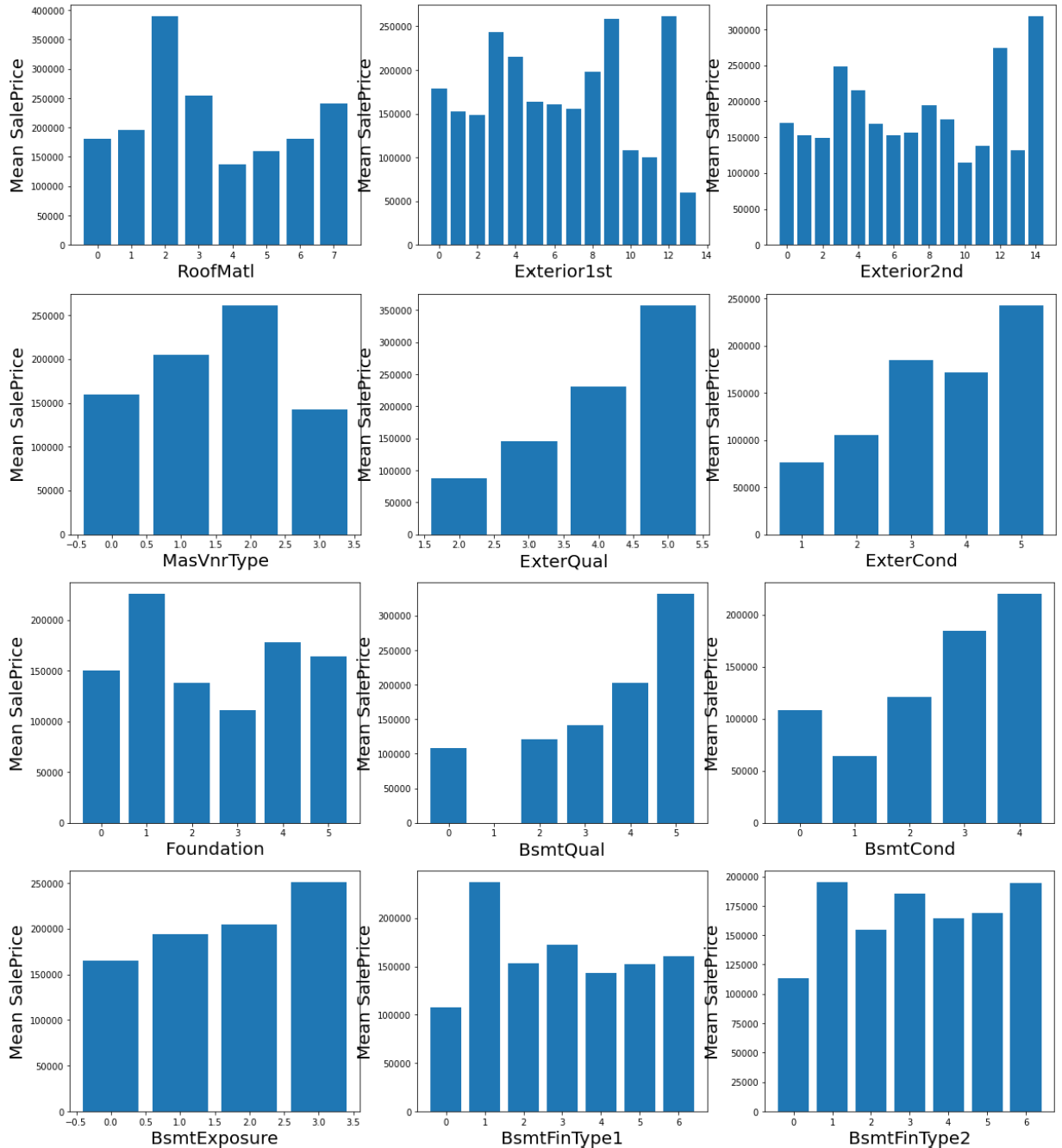
Observations:

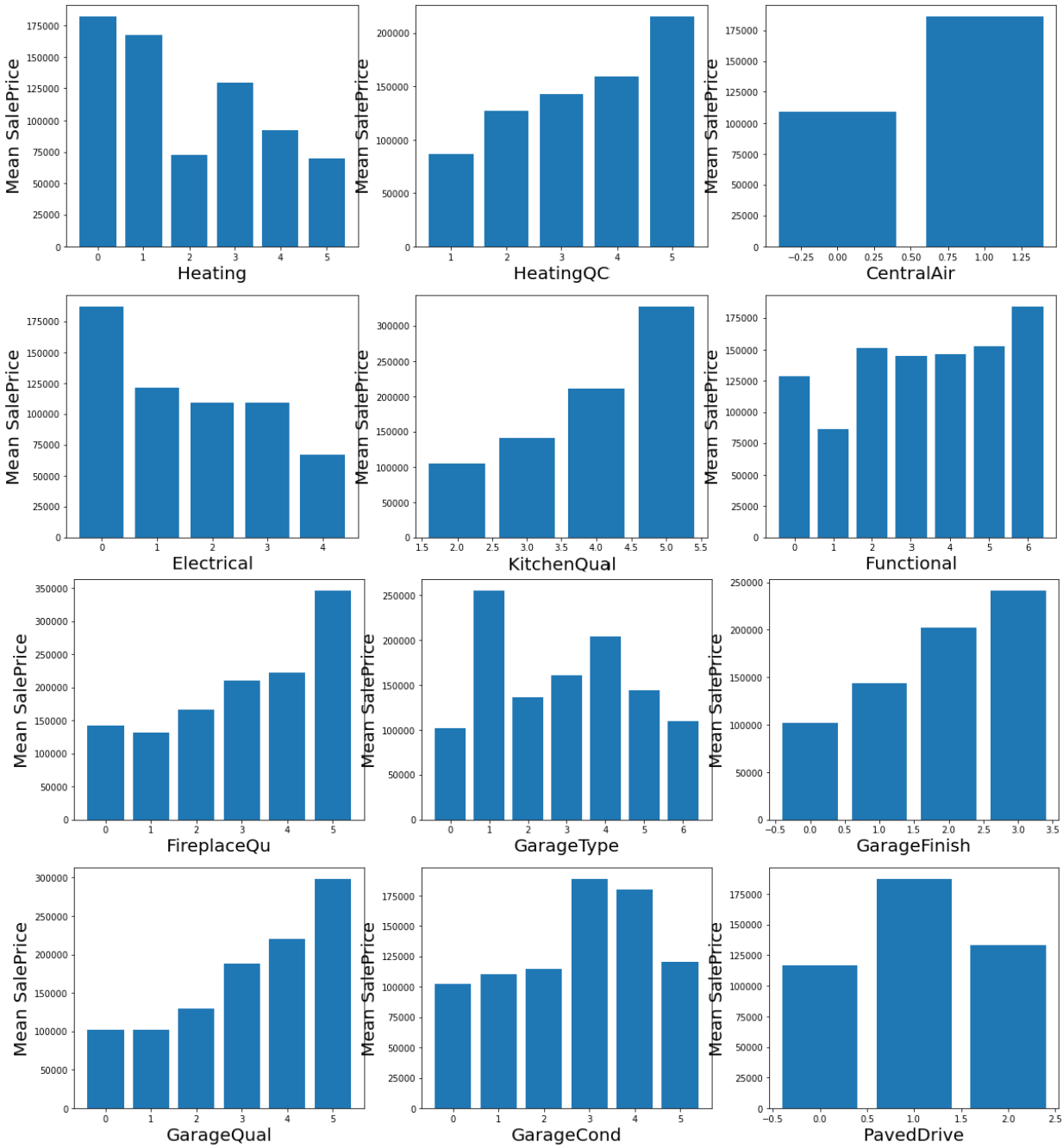
Columns such as OpenporchSF, WoodDeckSF, GarageArea, GarageYrBuilt, GrlivArea, 1stflrsf, 2ndflrsf, totalbsmtsf, YearremodAdd, Yearbuilt, lotfrontage, lotarea, masvnrarea, bsmntfinsf1 have positive correlation w.r.t salesprice

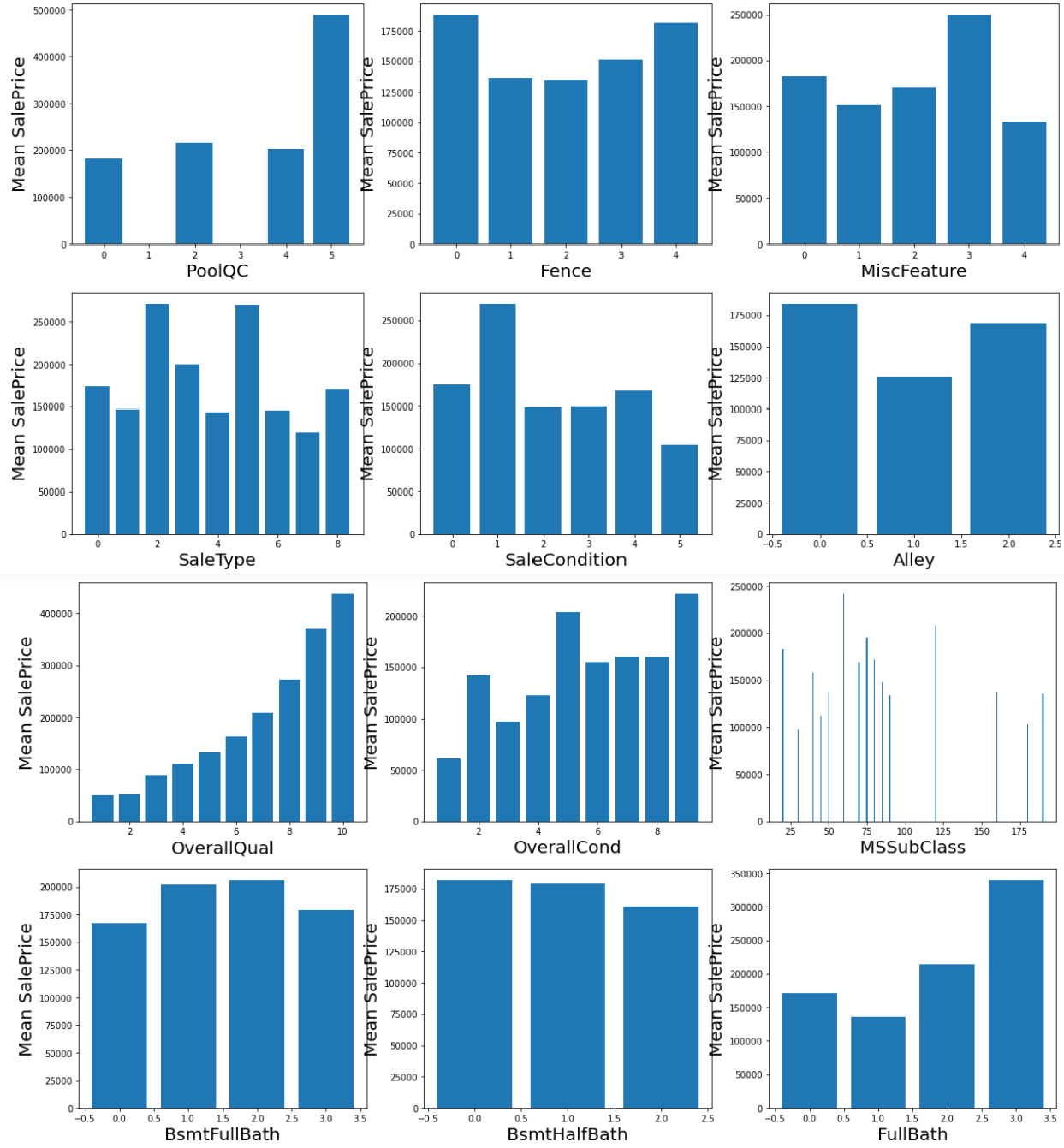
Columns such as MiscVal, Enclosedporch have slight negative correlation Remaining columns do not have any correlation w.r.t salesprice

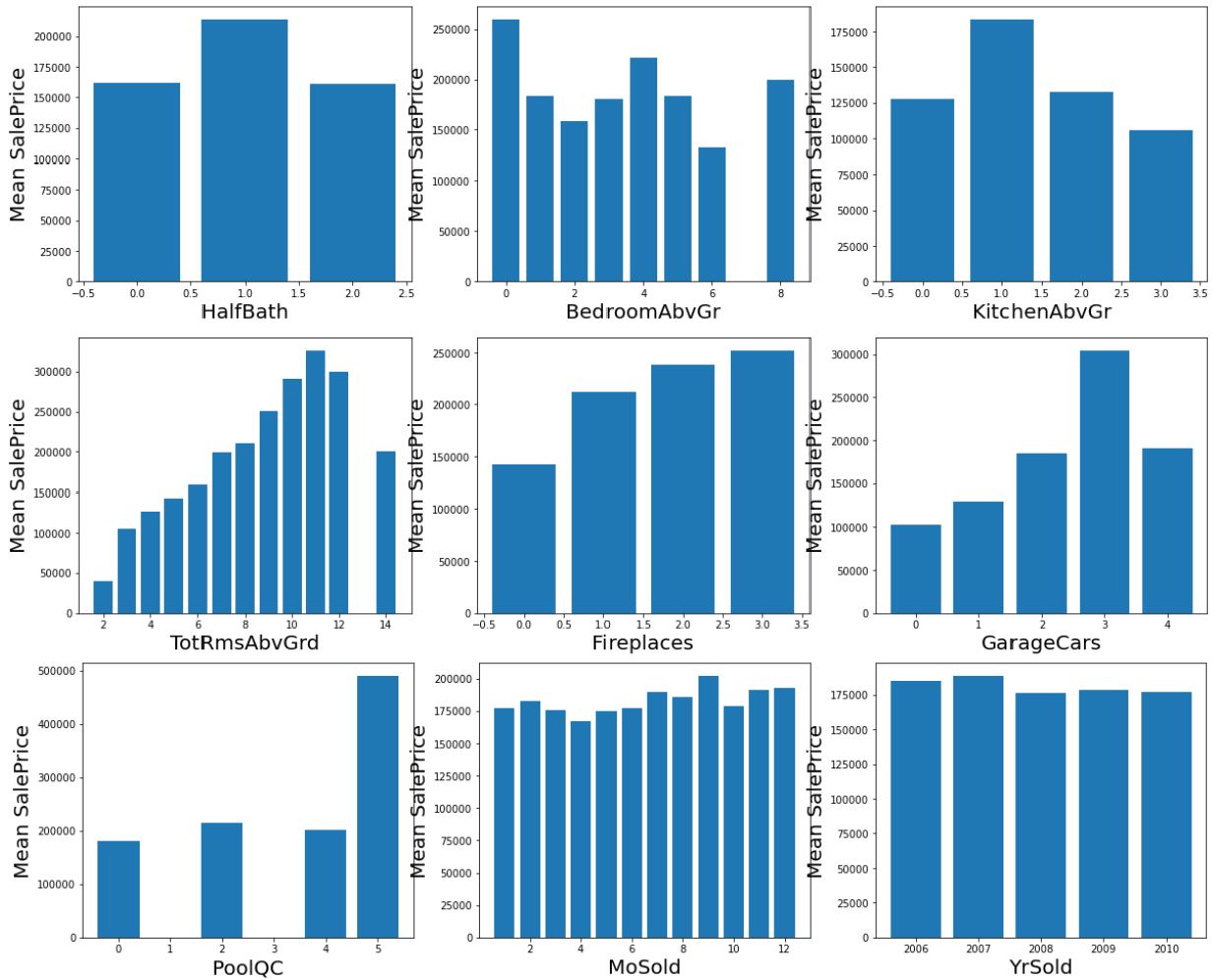
## Plot graph showing categorical columns value point and the respective mean value of salesprice for that category







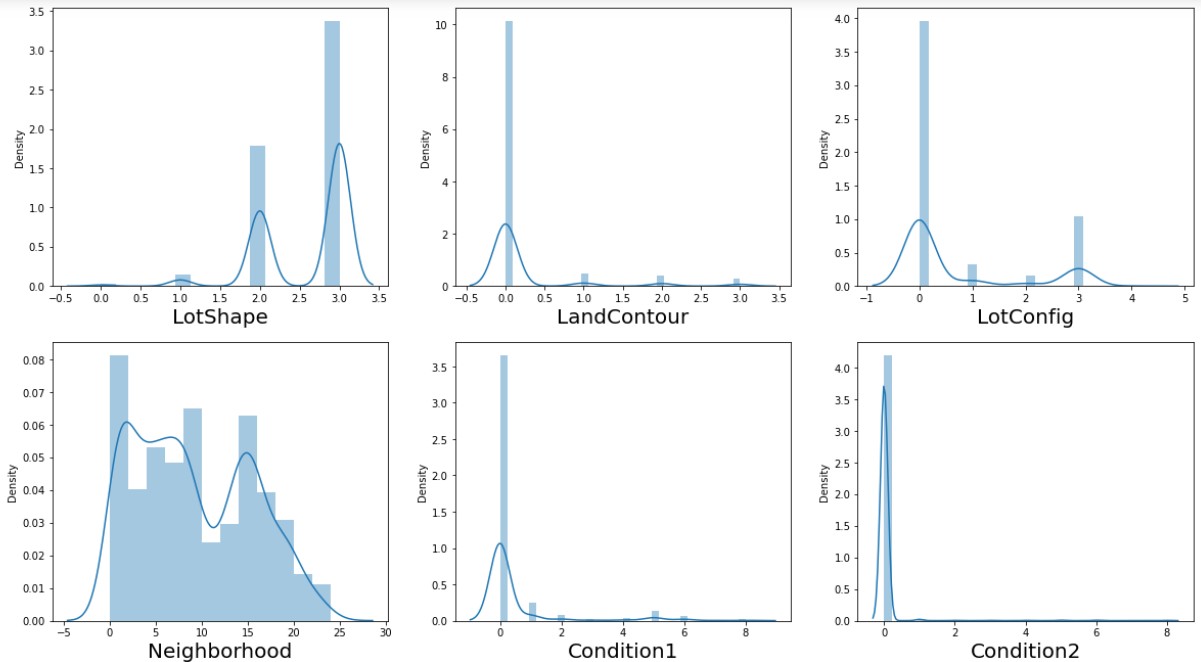
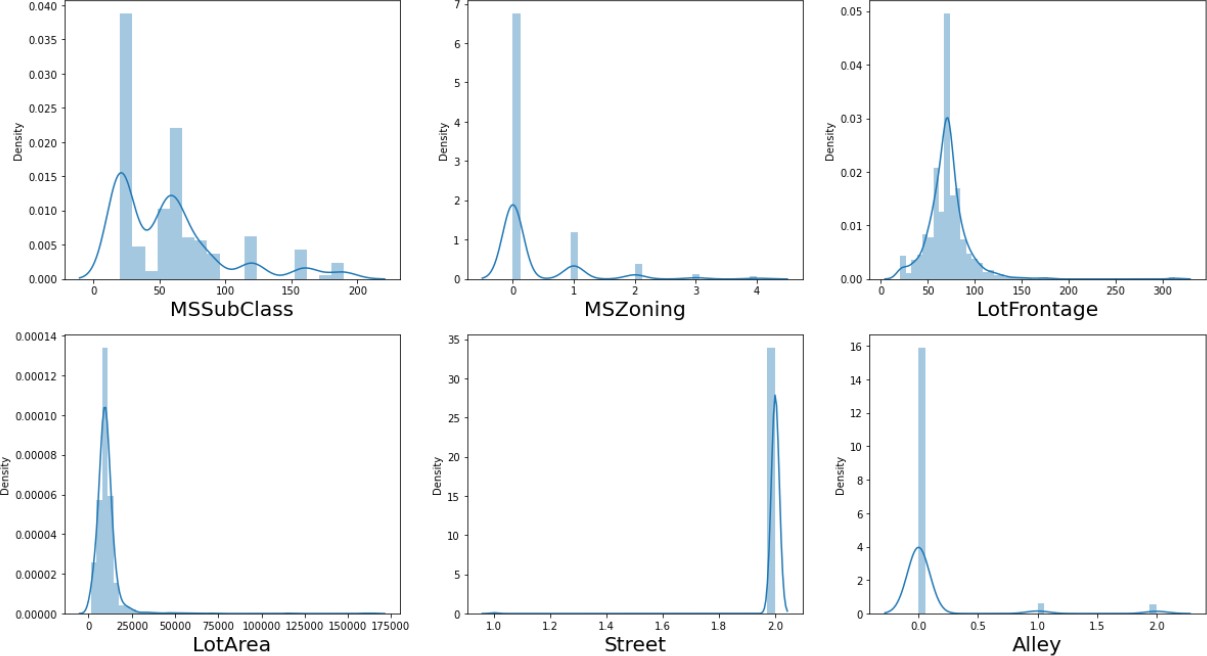


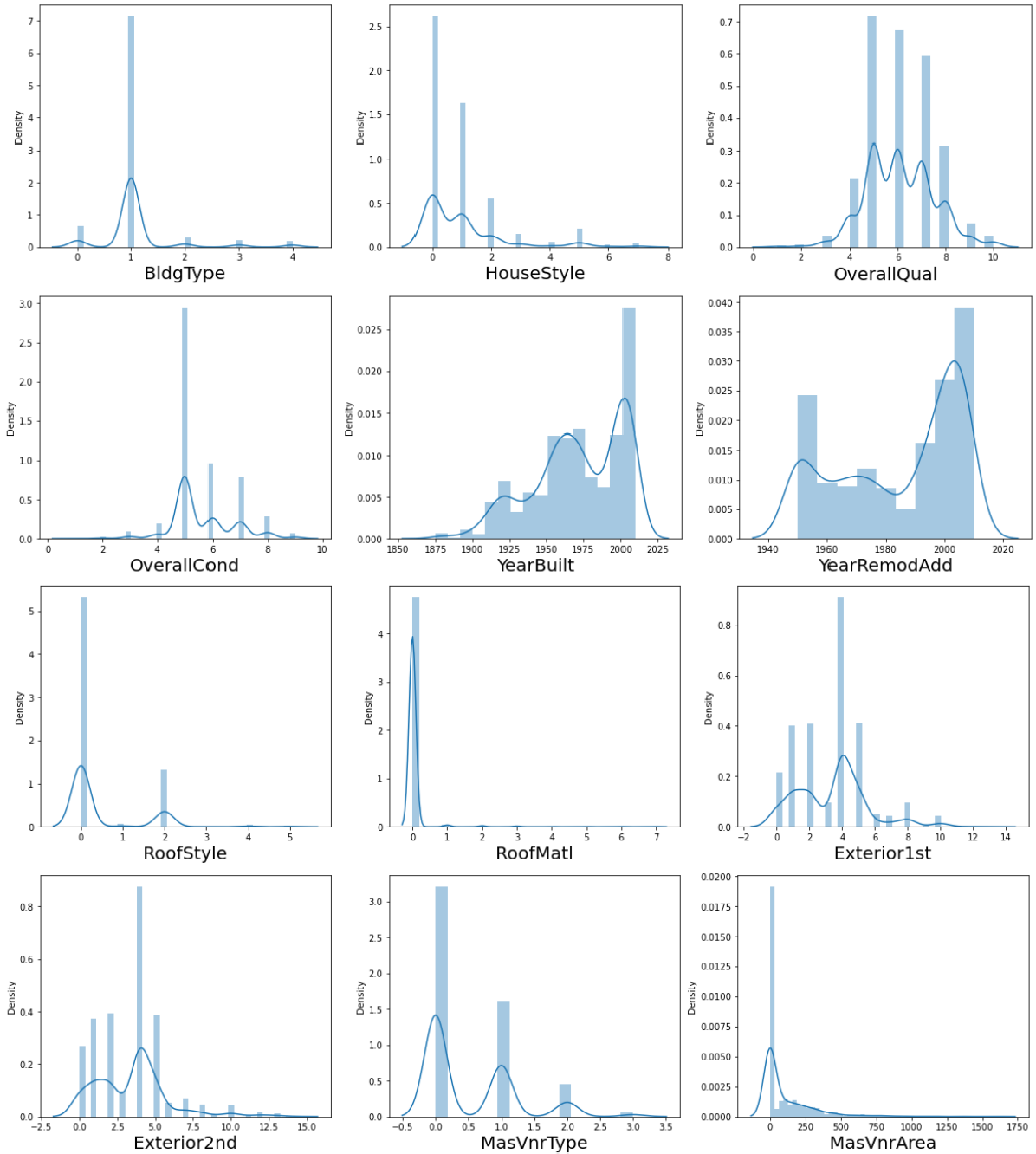


Observations:

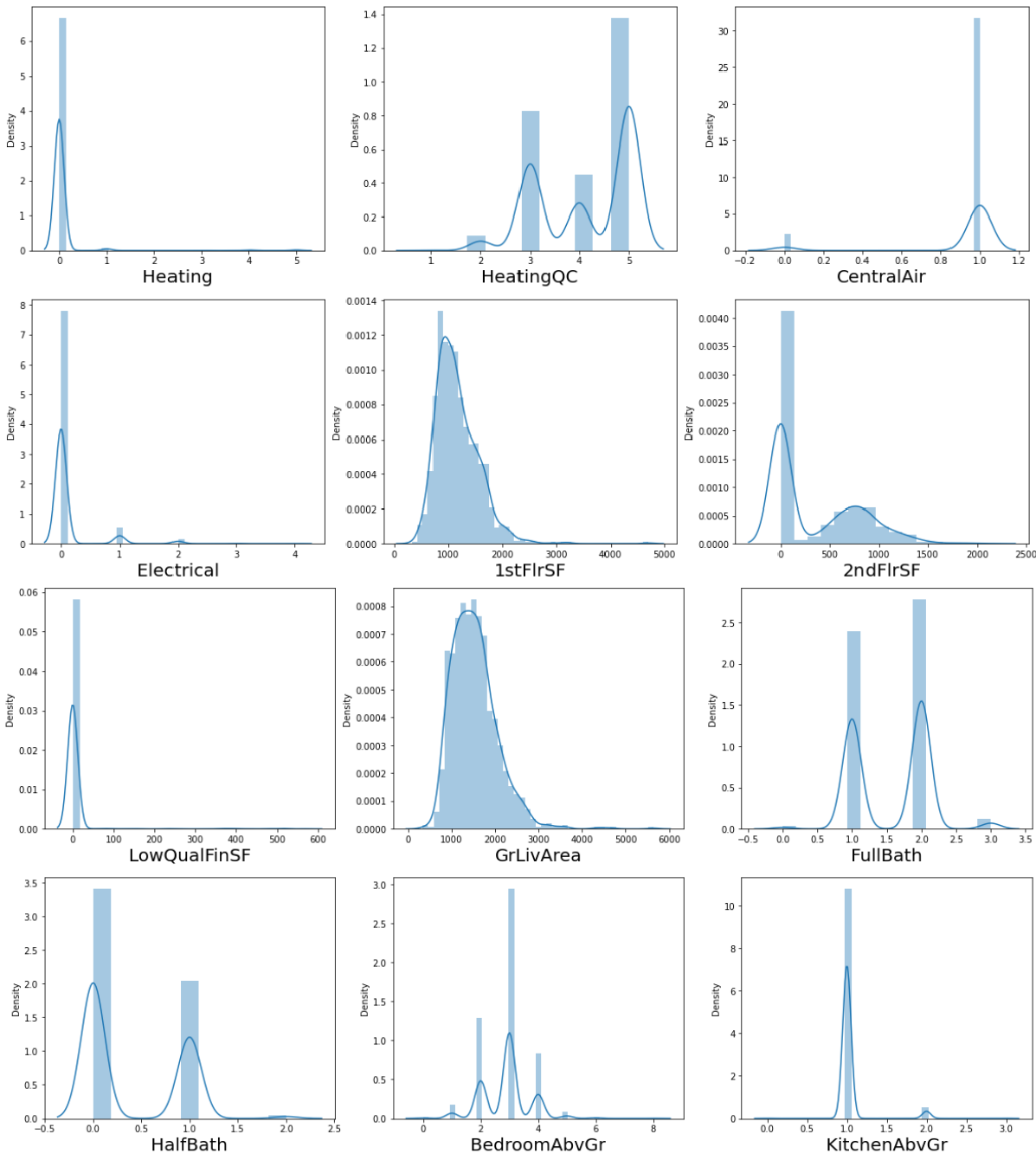
From above we can observe that categorical datapoints of columns LandSlope, MoSold, YrSold, BsmtFullBath, BsmtHalfBath have equal mean value of salesprice throughout their respective category, hence drop them.

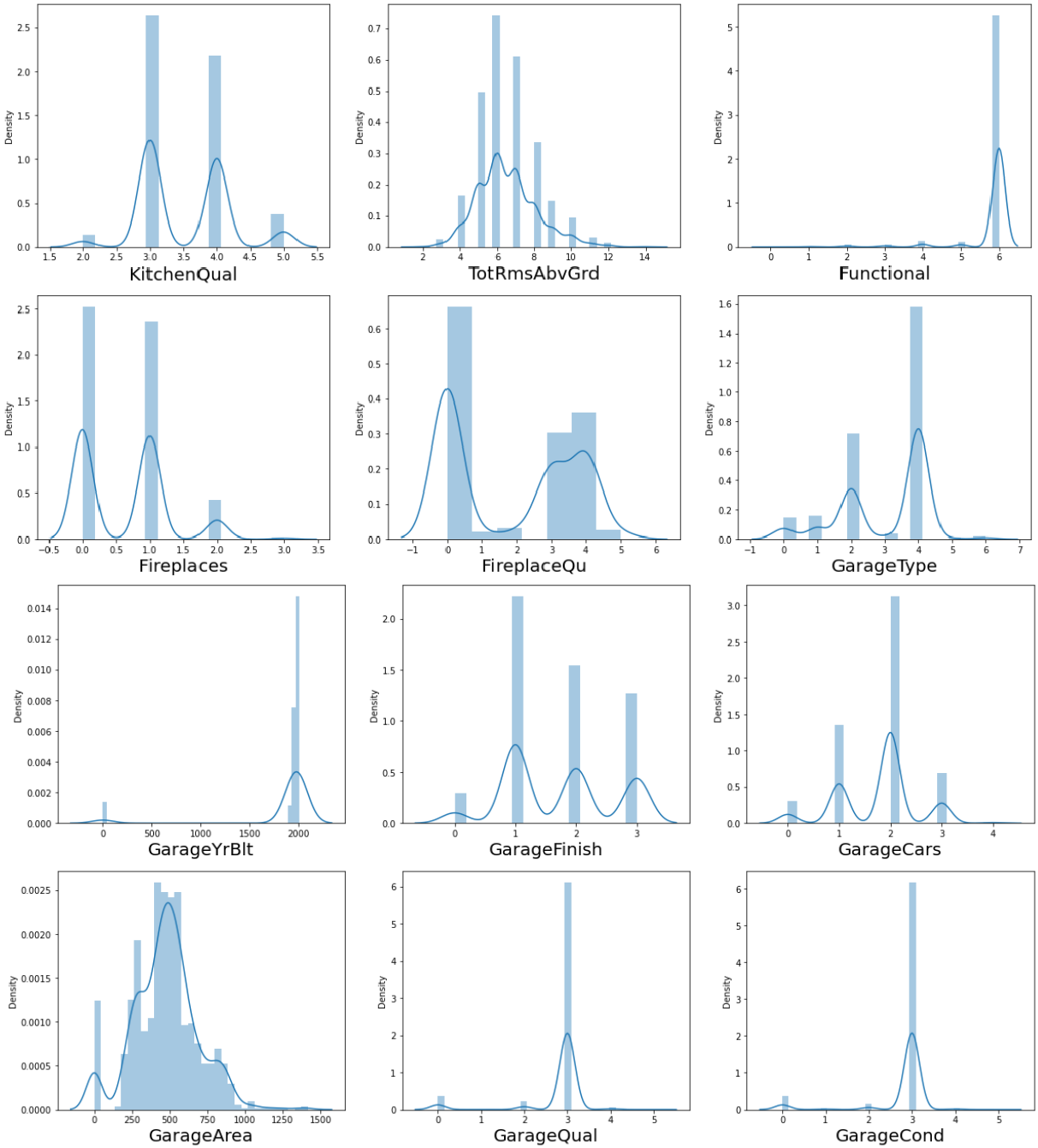
## Plot Distribution plot of each column

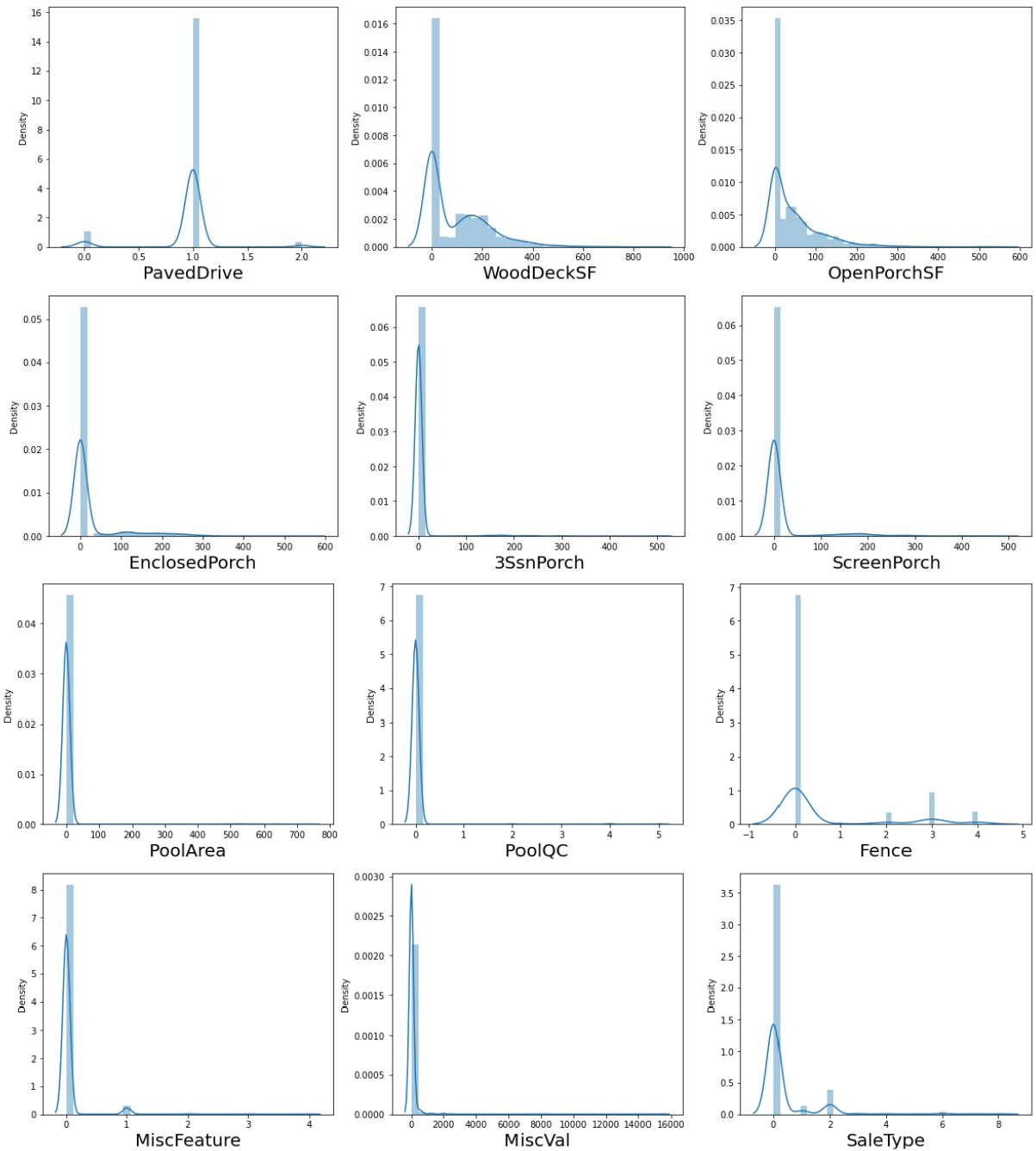


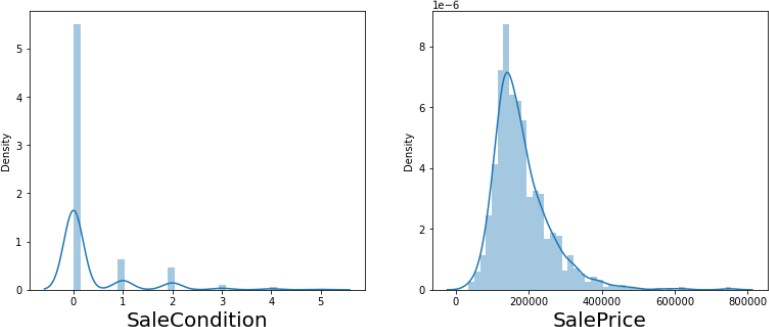












Observation:

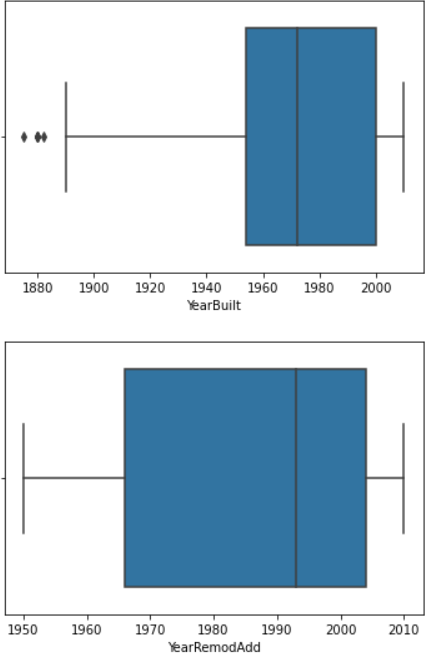
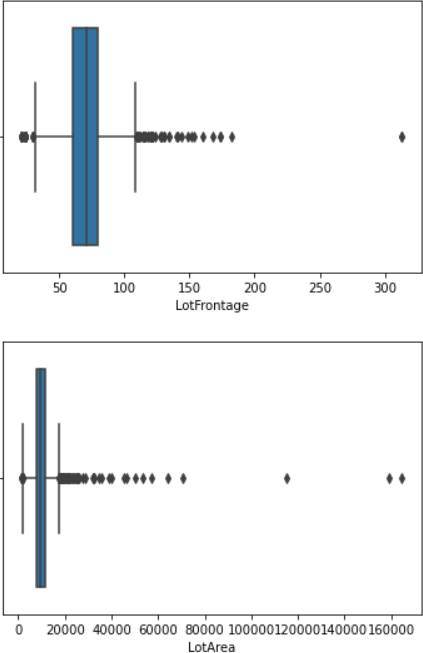
Columns having normal distribution plots are:

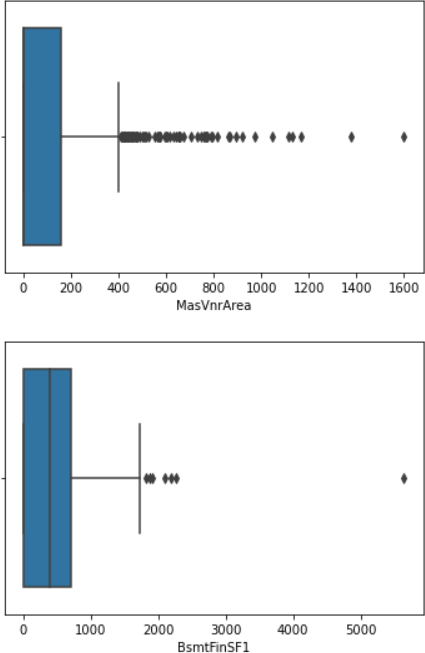
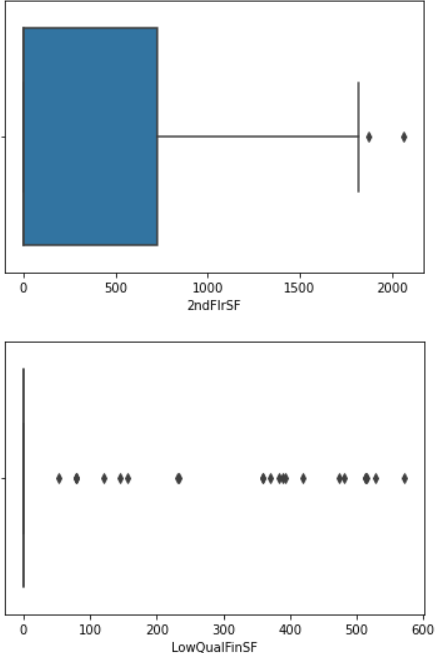
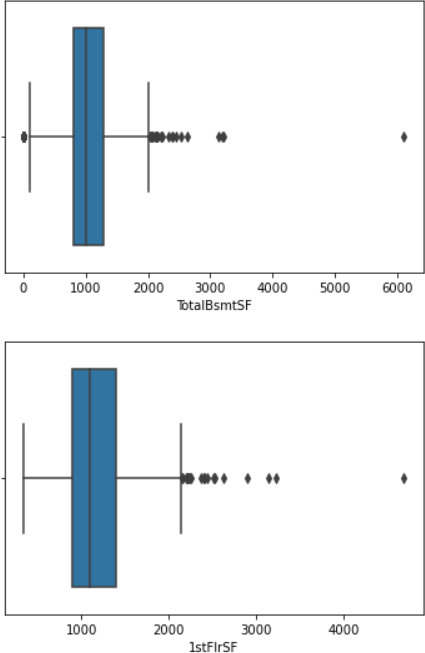
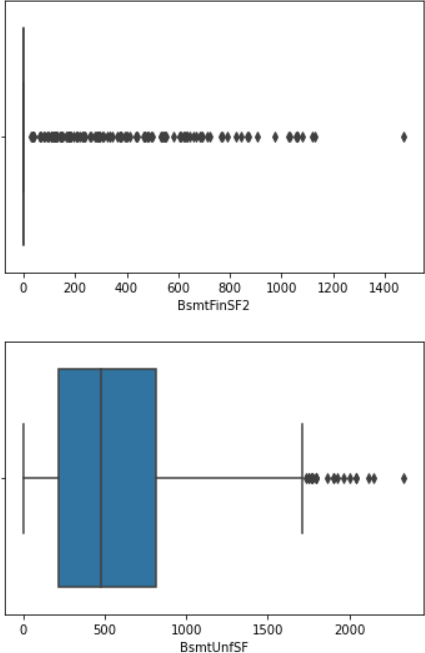
* + - LotFrontage
    - Alley
    - LandContour
    - Condition1
    - Conditional2
    - BldgType
    - RoofMatl
    - MasVnrArea
    - BsmtCond
    - BsmtFinType2
    - BsmtFinSF2
    - LotArea
    - BsmtUnfSF
    - TotalBsmtSF
    - Heating
    - CentralAir
    - Electrical
    - 1stFlrSF
    - LowQualFinSF
    - GrLivArea
    - KitchenAbvGr
    - Functional
    - GarageYrBlt
    - Street
    - GarageQual
    - GarageCond
    - PavedDrive
    - OpenPorchSF
    - EnclosedPorch
    - 3SsnPorch
    - ScreenPorch
    - PoolArea
    - PoolQC
    - Fence
    - MiscFeature
    - MiscVal
    - SaleType
    - SalePrice

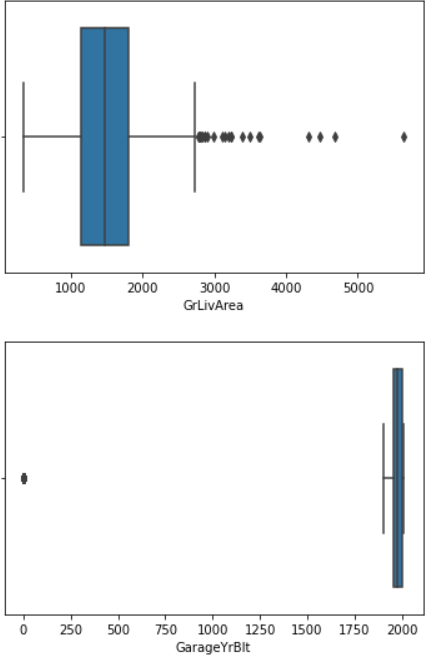
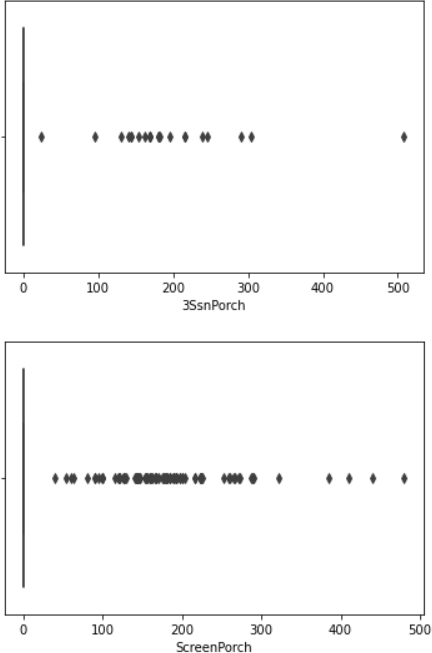
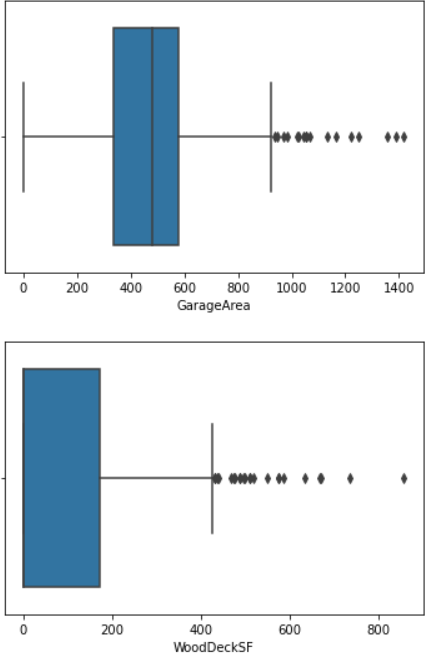
Rest all columns have bimodal type distribution plot

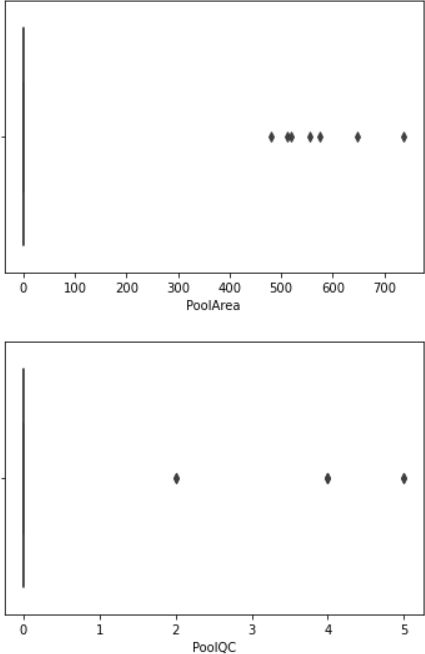
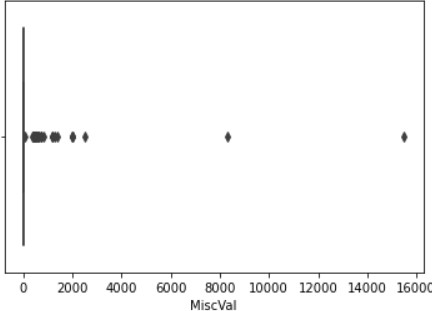
## Check For Outliers

Visualize Boxplot of every column having datapoints of continuous nature



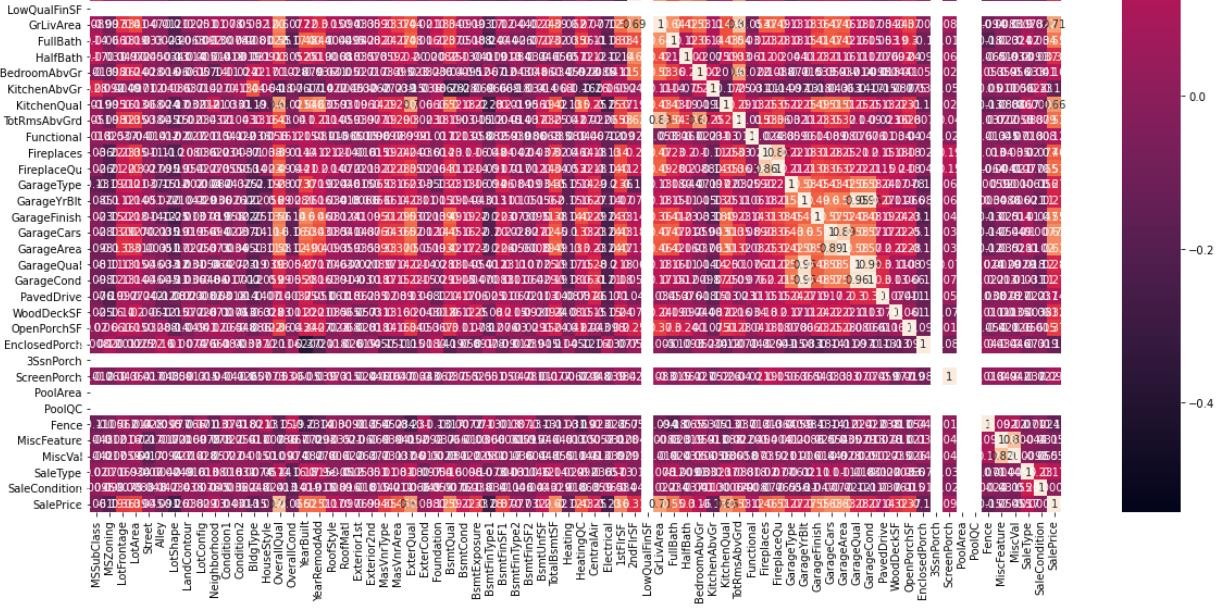
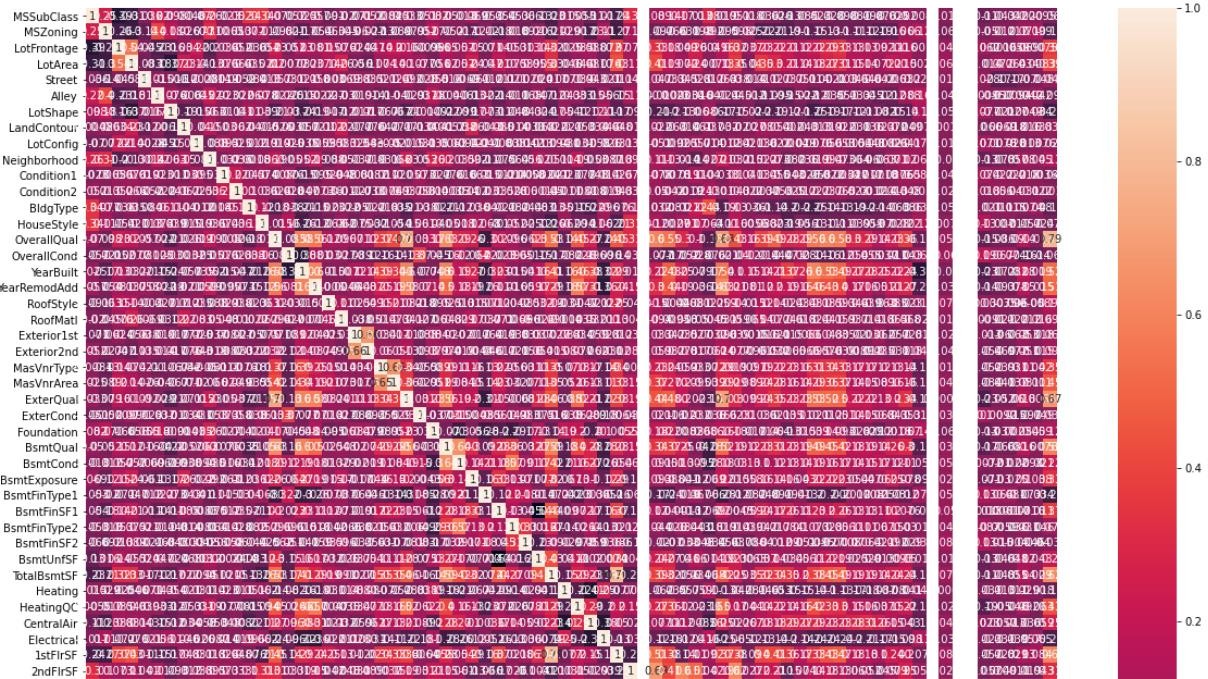
 

Outliers were detected in almost every column, remove them using percentile method

## Check For Correlation



Observation:

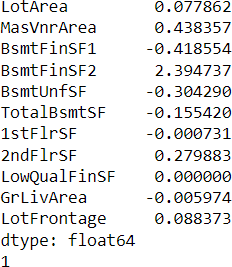
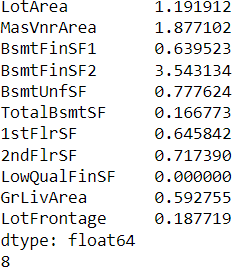
Very few columns have high correlation with column SalePrice but further dropping of column is not possible as we cannot drop more than 10% of columns w.r.t the original dataset.

## Perform Feature Scaling

Model Building

Before we start model building, we need to perform feature scaling on all columns, to avoid biasing of data.

Also check for skewness in data and remove it.

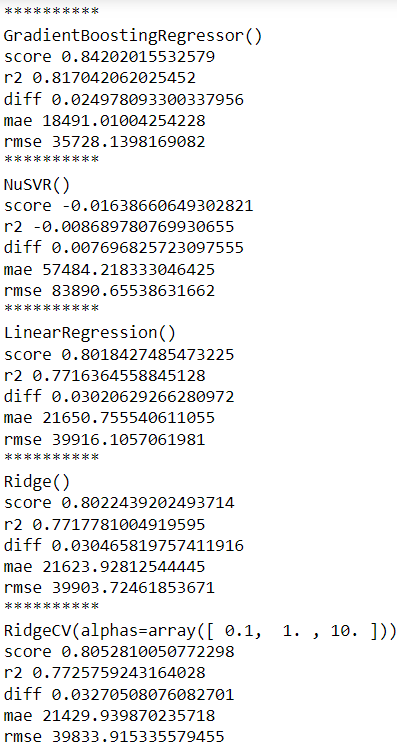
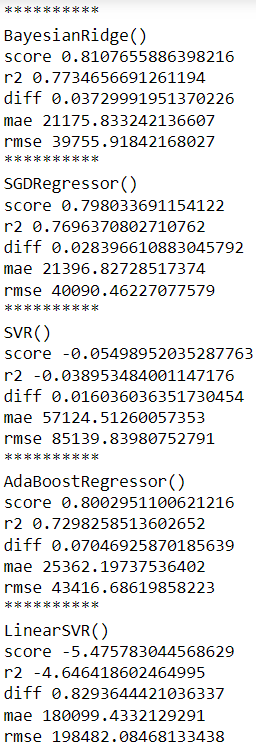


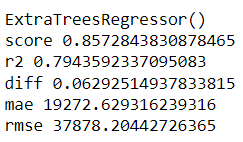
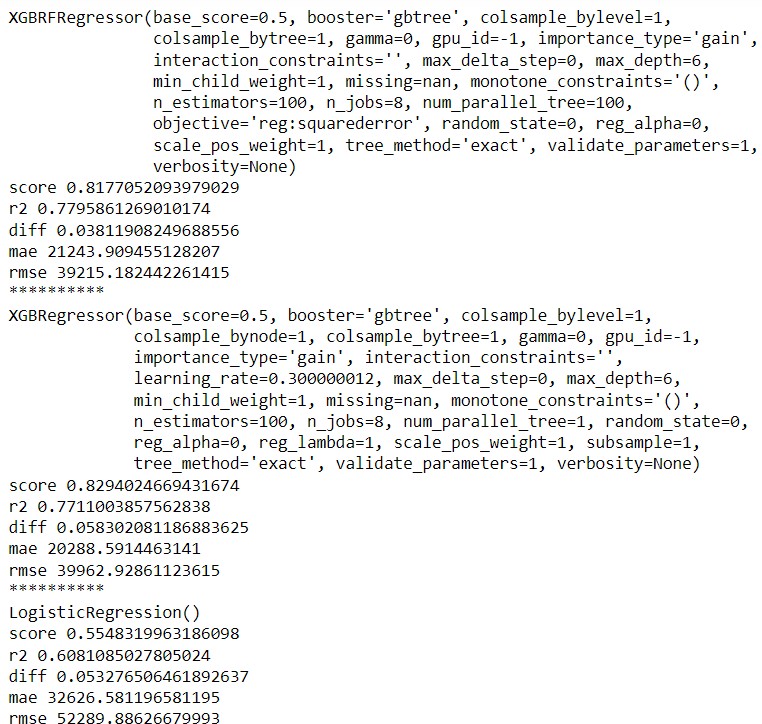
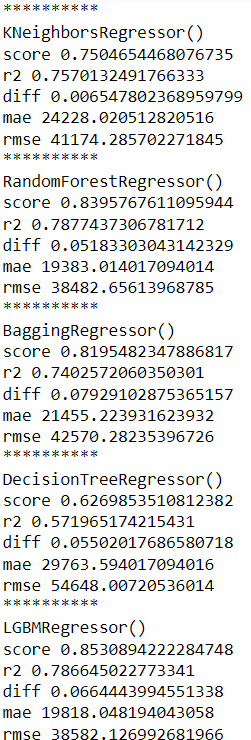
## Model Building

As we know, this is a regression problem we need to build a model using regression algorithm models.

First, we need to write a function which can find us best random state for train test split.

Then we shall iterate through all the models supporting regression algorithms to find the best models.



From above we get to know that the top 5 models are:

* + - ExtratreesRegressor
    - LGMBRegressor
    - GradientboostingRegressor
    - RandomForestRegressor
    - XGBRegressor

Fine tune all these models and find their best parameters to use. Next, find the best random state for train test split.

As we know from above output that our top models do not have accuracy above 90%, hence we will stack our top 5 models to build one model to obtain higher accuracy.

To stack models, we must use StackCVRegressor to combine all our fine tune models. After using StackCVRegressor we obtain test accuracy of more than 90%.

CV score of this model is more than 87%.

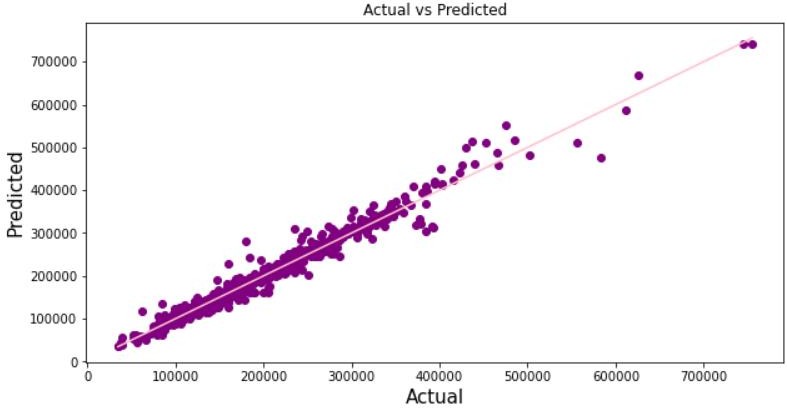
To analyze our model, we shall find the difference between actual and predicted value.

Value difference between actual and predicted value when actual value is greater than predicted value is 89259

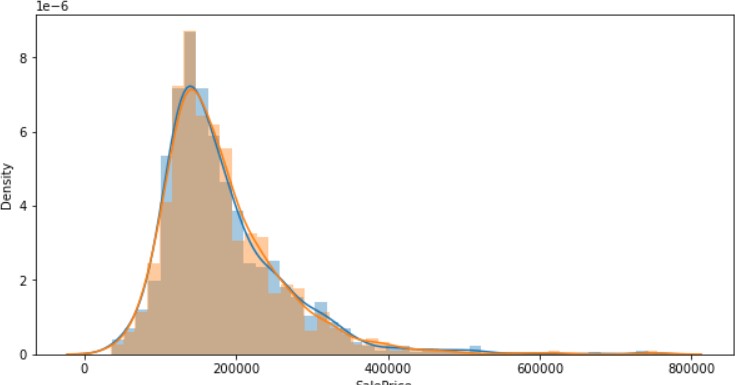
Value difference between actual and predicted value when actual value is less than predicted value is 95990

Avg difference between actual and predicted value is 193

## Regression Plot of Actual vs Predicted value



Compare Distribution plot of Actual vs Predicted Value



Observation:

From above 2 plots we can observe the closeness between actual and predicted value.

Hence, we can verify that the model built is able to predict SalePrice with a high accuracy. Save model for further use.

## Perform the same data processing steps for test dataset & predict the value for target variable using the existing Stacked Model



Save this dataset in a CSV file

Interpretation of the Results

Here we check the correlation between all our feature variables with target variable label

1. The column OverallQual is most positively correlated with SalePrice.
2. The column KitchenAbvGrd is most negatively correlated with SalePrice.

Set of assumptions related to the problem under consideration

By looking into the target variable datapoints we assumed that it was a Regression type of problem.

We also observed that only one single unique value was present in Utilities column so we assumed that we will be dropping these columns. ID column was also dropped as it contained all unique vales.

Columns such as LandSlope, MoSold, YrSold, BsmtFullBath, BsmtHalfBath have equal mean value of salesprice throughout their respective category which basically concludes that whatever the datapoint is SalePrice is not affected by it, hence drop them also.

Outliers were removed using percentile method. Skewness was reduced using Yeo- Johnson method.

Final model built is actually a combination of top 5 fine tuned model to achieve high accuracy.

# Conclusion

## Key Features and conclusion of the study

In this project we have tried to show how the house prices vary and what are the factors related to the changing of house prices. The best accuracy score was achieved by stacking our top 5 fine-tuned models.

## LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

Through different powerful tools of visualization, we were able to analyze and interpret different hidden insights about the data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The data was improper scaled, so we scaled it to a single scale using sklearns’s package StandardScaler.

There were lot of missing values present in different columns which we imputed on the basis of our understanding.

The columns were skewed due to presence of outliers which we handled through percentile technique.

Stacked Model was then built having accuracy more than 90% using train dataset.

Finally, we predicted the SalePrice for Test dataset using our stacked model and saved the data frame into a CSV file.

## LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

While we couldn’t reach out goal of minimum RMSE in house price prediction without letting the model to overfit, we did end up creating a system that can with enough time and data get very close to that goal.

As with any project there is room for improvement here.

The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result.

This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code.

This provides a great degree of modularity and versatility to the project.