

**HOUSING: PRICE PREDICTION**

Submitted by:

KOMMERA AMITH REDDY

**ACKNOWLEDGMENT**

I would like to thank FlipRobo for giving this challenging assignment and also I would like to thank Data trained for making me understand about Data cleaning and modelling of a data set for required outcomes.

**INTRODUCTION**

* **Business Problem Framing**

The objective of our project is to predict 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.

* **Conceptual Background of the Domain Problem**

Houses are one of the necessary need of each and every 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.

* **Motivation for the Problem Undertaken**

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes challenges you to predict the final price of each home.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The main models we used are:

1. Linear Regression model
2. Support Regression model
3. Decision Tree Regressor
4. Random Forest Regressor

My approach consists in taking accuracy and RMSE score into account before finalising the model, and also the model is Hypertuned with Gridsearch CV and the best parameters with model are considered while finalising.

* **Data Sources and their formats**

Data was obtained from FlipRobo and was given to us as an assignment, Data types are of mix type and are of form float, int and there are also null values, below table shows the data information

RangeIndex: 1168 entries, 0 to 1167

Data columns (total 81 columns):

Id 1168 non-null int64

MSSubClass 1168 non-null int64

MSZoning 1168 non-null object

LotFrontage 954 non-null float64

LotArea 1168 non-null int64

Street 1168 non-null object

Alley 77 non-null object

LotShape 1168 non-null object

LandContour 1168 non-null object

Utilities 1168 non-null object

LotConfig 1168 non-null object

LandSlope 1168 non-null object

Neighborhood 1168 non-null object

Condition1 1168 non-null object

Condition2 1168 non-null object

BldgType 1168 non-null object

HouseStyle 1168 non-null object

OverallQual 1168 non-null int64

OverallCond 1168 non-null int64

YearBuilt 1168 non-null int64

YearRemodAdd 1168 non-null int64

RoofStyle 1168 non-null object

RoofMatl 1168 non-null object

Exterior1st 1168 non-null object

Exterior2nd 1168 non-null object

MasVnrType 1161 non-null object

MasVnrArea 1161 non-null float64

ExterQual 1168 non-null object

ExterCond 1168 non-null object

Foundation 1168 non-null object

BsmtQual 1138 non-null object

BsmtCond 1138 non-null object

BsmtExposure 1137 non-null object

BsmtFinType1 1138 non-null object

BsmtFinSF1 1168 non-null int64

BsmtFinType2 1137 non-null object

BsmtFinSF2 1168 non-null int64

BsmtUnfSF 1168 non-null int64

TotalBsmtSF 1168 non-null int64

Heating 1168 non-null object

HeatingQC 1168 non-null object

CentralAir 1168 non-null object

Electrical 1168 non-null object

1stFlrSF 1168 non-null int64

2ndFlrSF 1168 non-null int64

LowQualFinSF 1168 non-null int64

GrLivArea 1168 non-null int64

BsmtFullBath 1168 non-null int64

BsmtHalfBath 1168 non-null int64

FullBath 1168 non-null int64

HalfBath 1168 non-null int64

BedroomAbvGr 1168 non-null int64

KitchenAbvGr 1168 non-null int64

KitchenQual 1168 non-null object

TotRmsAbvGrd 1168 non-null int64

Functional 1168 non-null object

Fireplaces 1168 non-null int64

FireplaceQu 617 non-null object

GarageType 1104 non-null object

GarageYrBlt 1104 non-null float64

GarageFinish 1104 non-null object

GarageCars 1168 non-null int64

GarageArea 1168 non-null int64

GarageQual 1104 non-null object

GarageCond 1104 non-null object

PavedDrive 1168 non-null object

WoodDeckSF 1168 non-null int64

OpenPorchSF 1168 non-null int64

EnclosedPorch 1168 non-null int64

3SsnPorch 1168 non-null int64

ScreenPorch 1168 non-null int64

PoolArea 1168 non-null int64

PoolQC 7 non-null object

Fence 237 non-null object

MiscFeature 44 non-null object

MiscVal 1168 non-null int64

MoSold 1168 non-null int64

YrSold 1168 non-null int64

SaleType 1168 non-null object

SaleCondition 1168 non-null object

SalePrice 1168 non-null int64

dtypes: float64(3), int64(35), object(43)

memory usage: 739.2+ KB

* **Data Preprocessing Done**

The first task was data cleaning, as ever. I immediately dropped five variables ( PoolQC','MiscFeature','Alley','Fence','FireplaceQu') that had less than 500 observations each. The “LotFrontage” (linear feet of street connected to property) variable, which I thought could be important, was missing 214 observations. I filled the missing values with the average “LotFrontage” when grouped by their respective “LotShape” categories.

Graphical user interface, text, application

Description automatically generated

For the remaining variable in which there are null values an assumption is made that the values are either ‘NA’ or 0 depending upon the variable

A close-up of a document

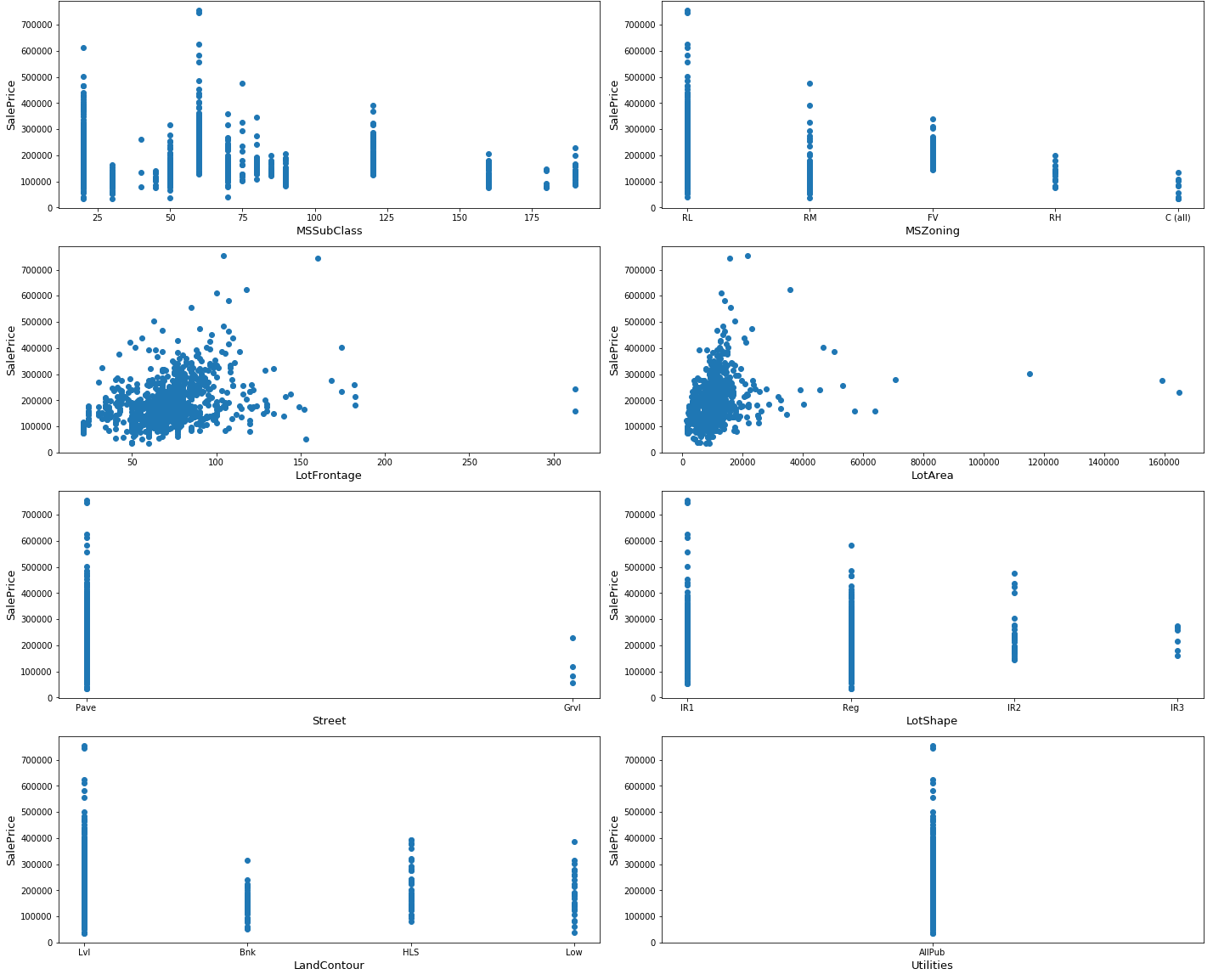
Description automatically generated with low confidence

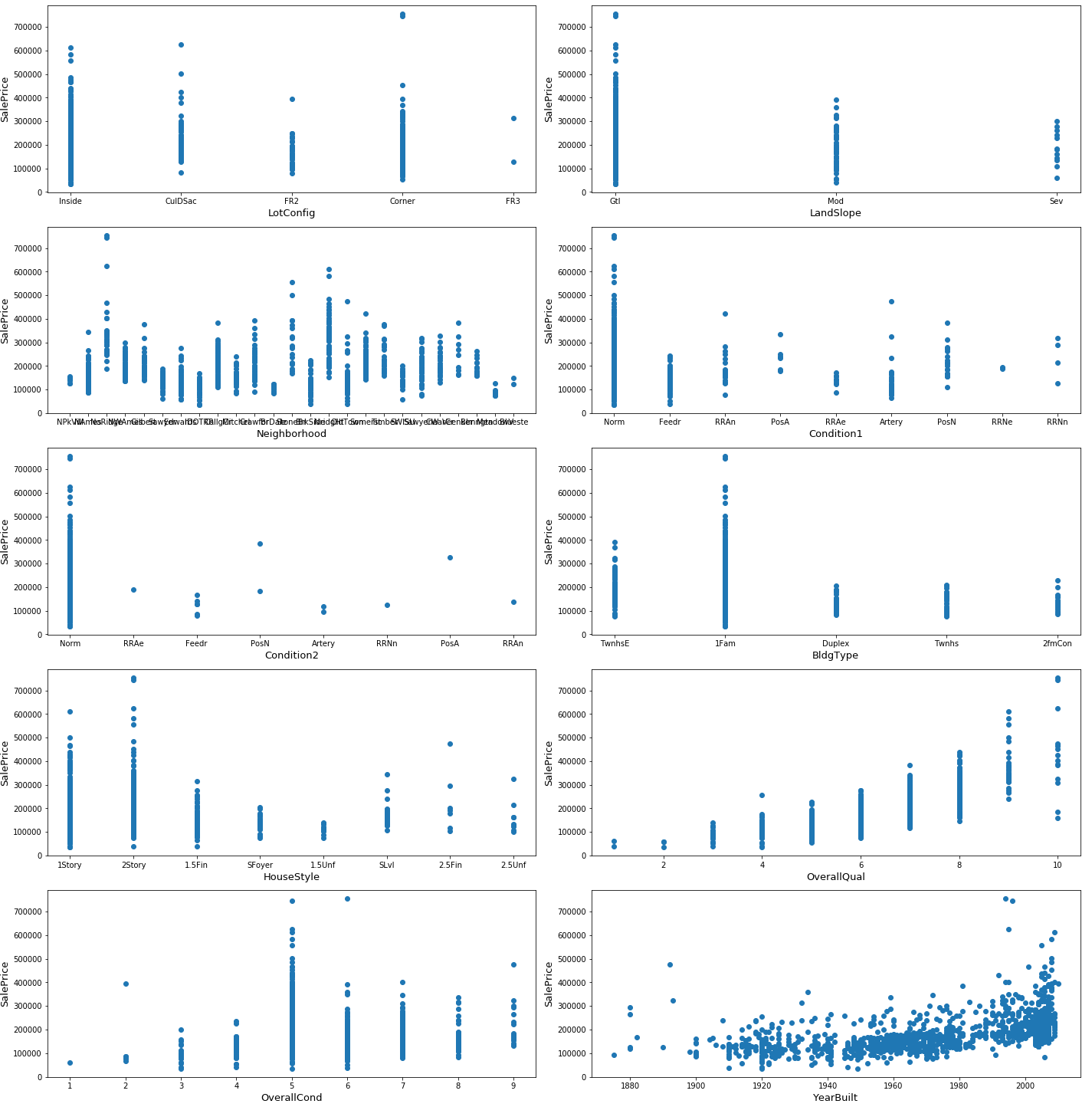
Also outliers in the features were also removed as per the visualization plots.

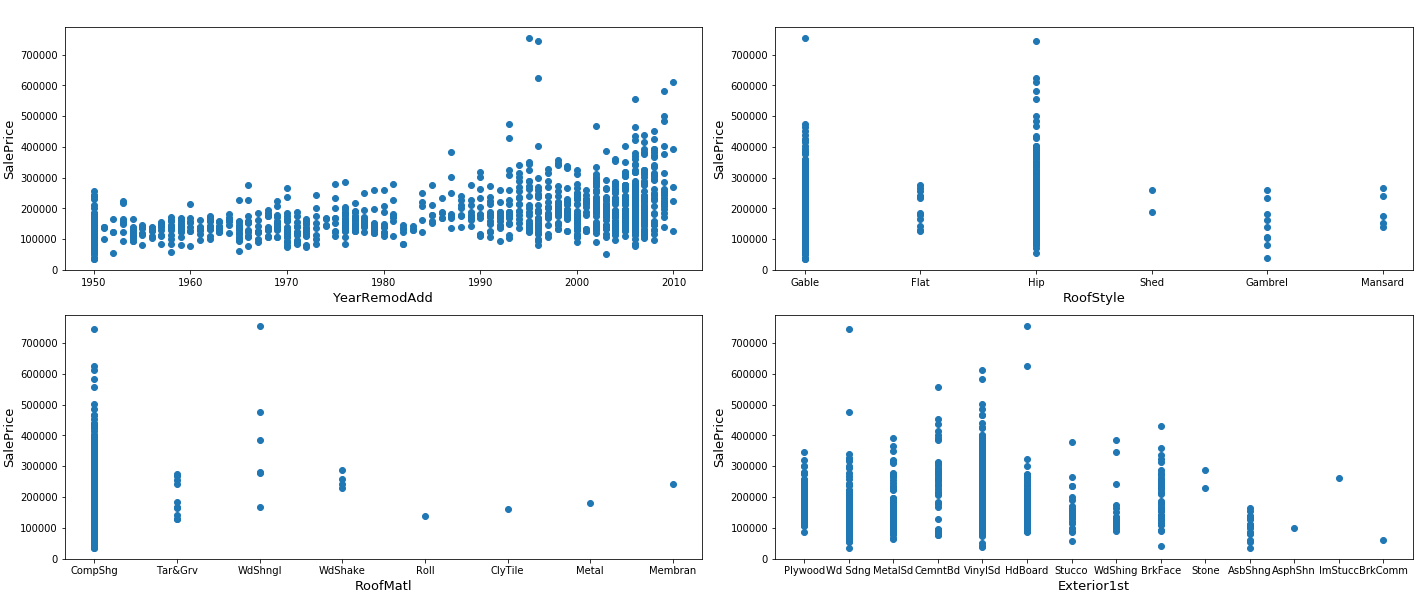
A picture containing text

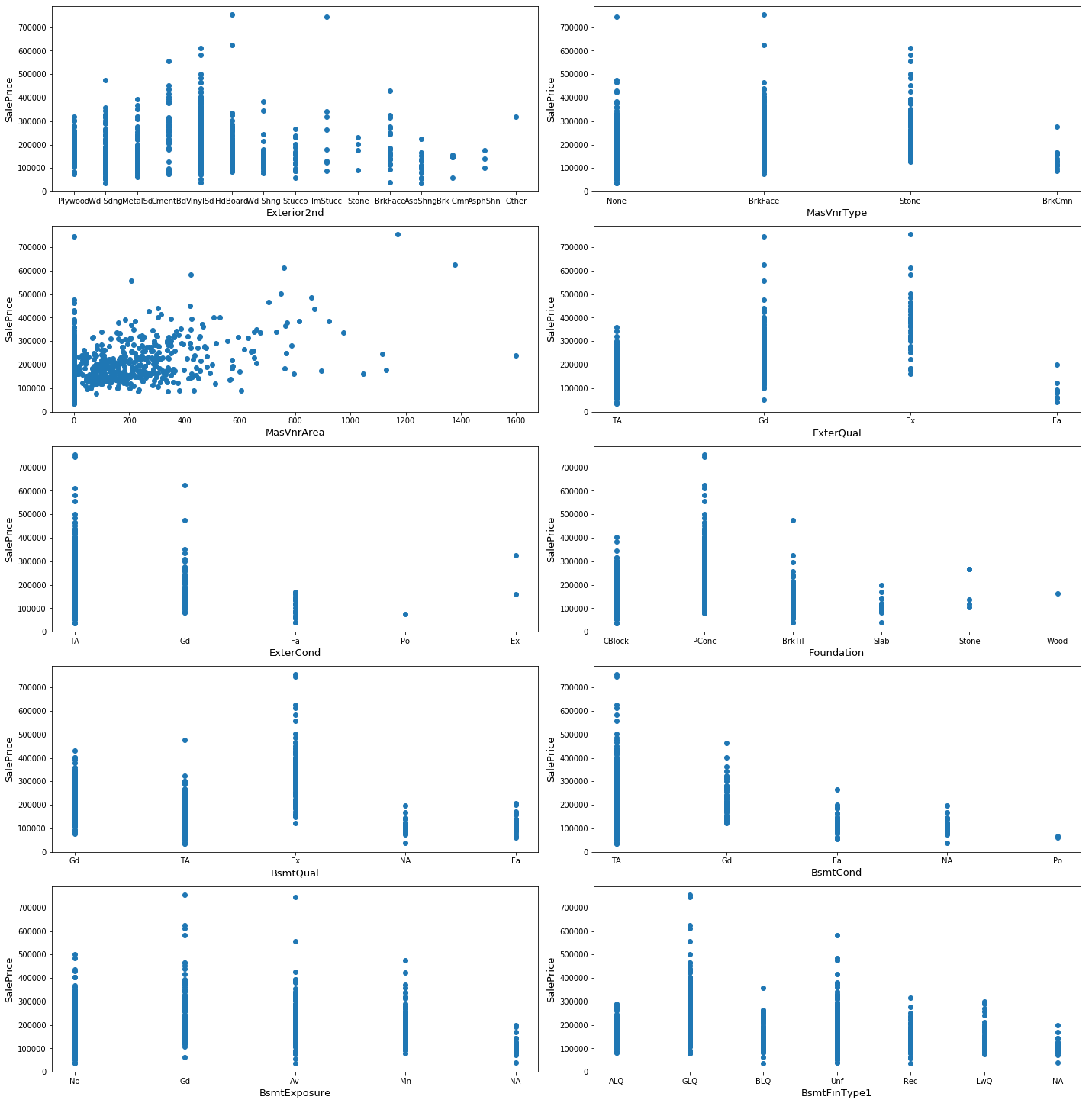
Description automatically generated

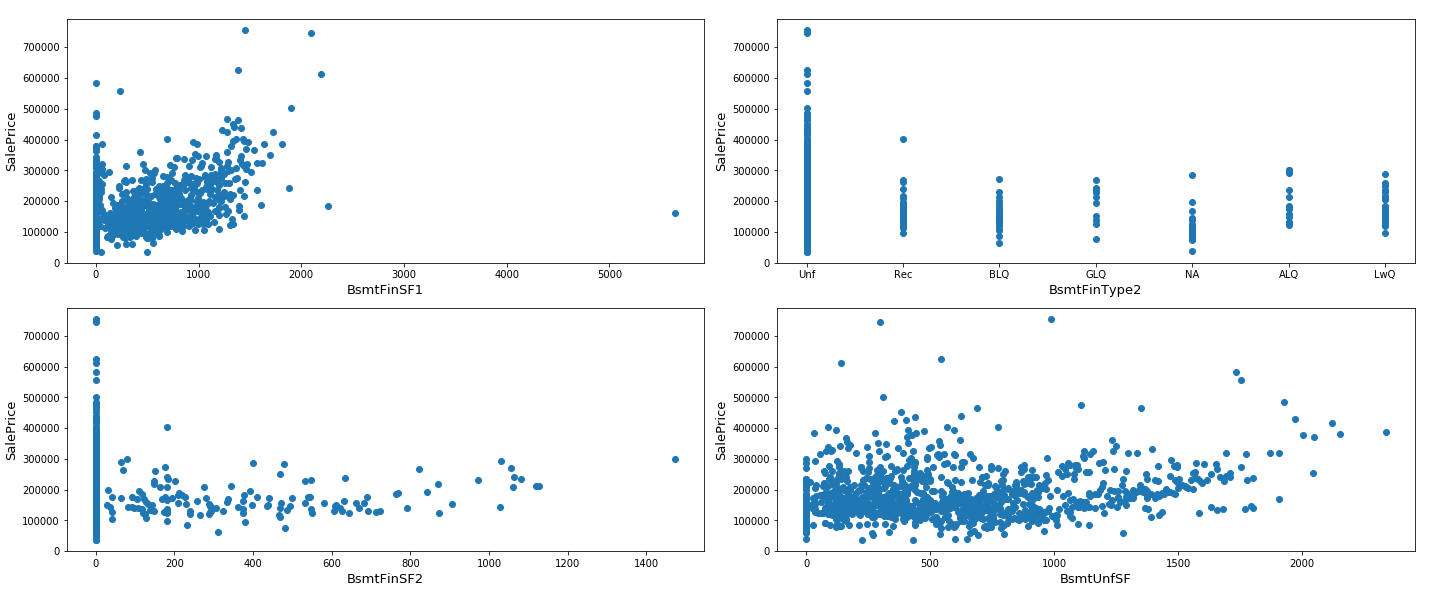
Later explorative data visualization plots were plotted against each feature and extreme outliers are removed

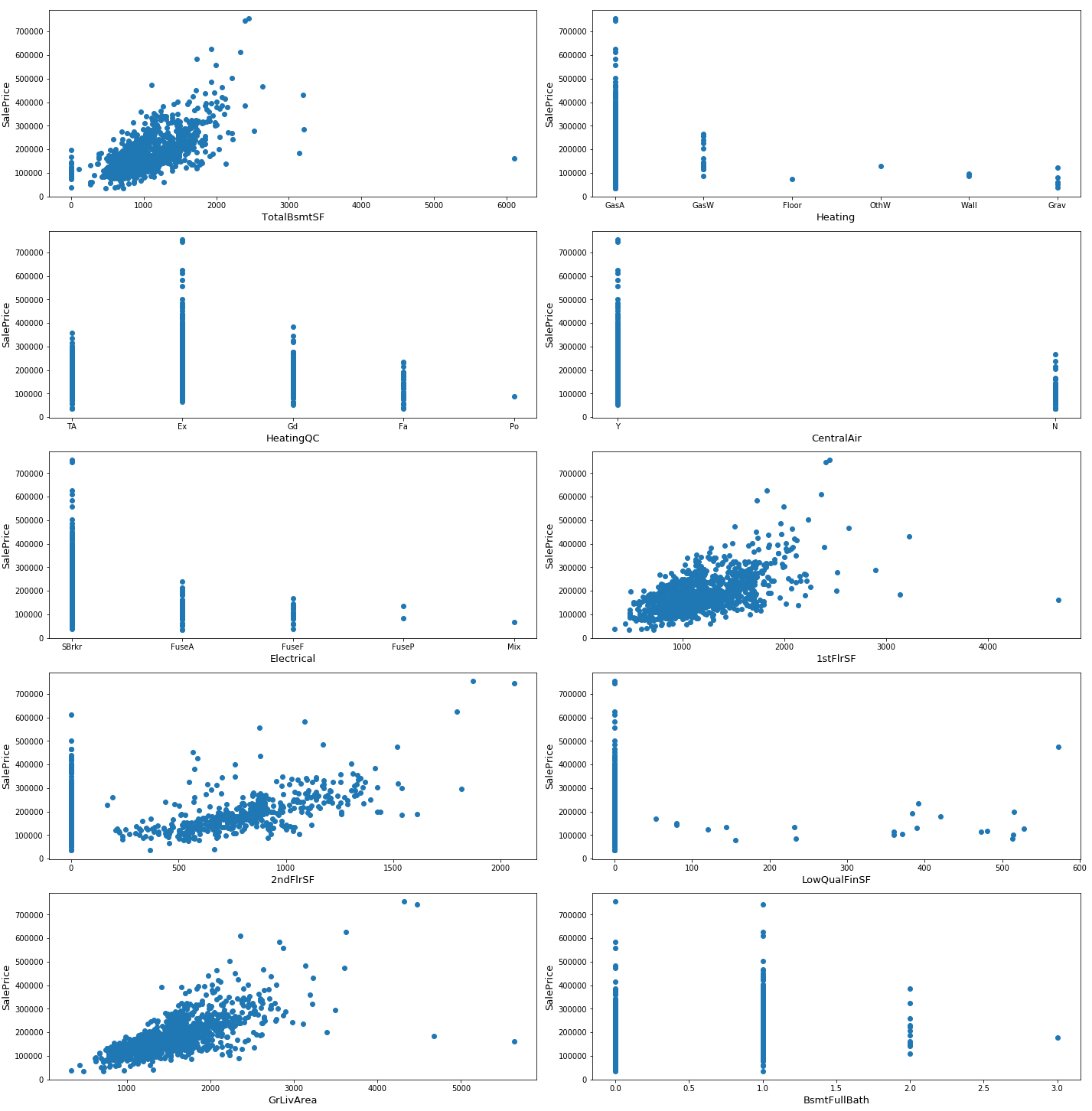


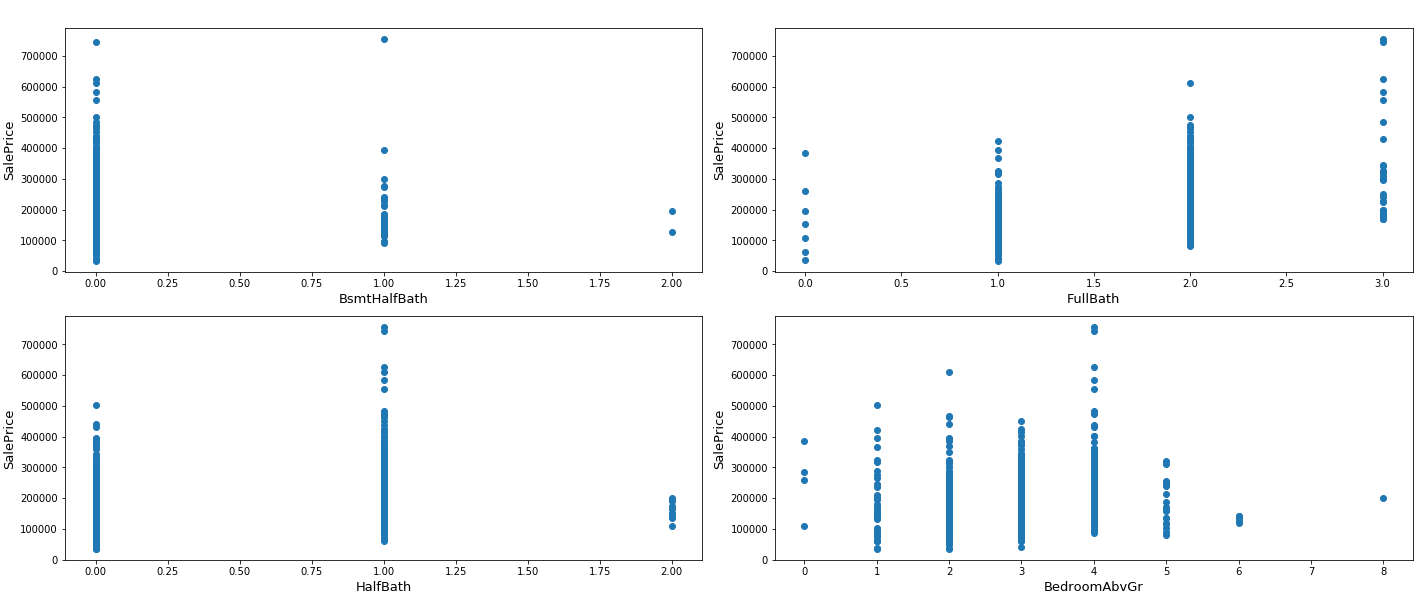


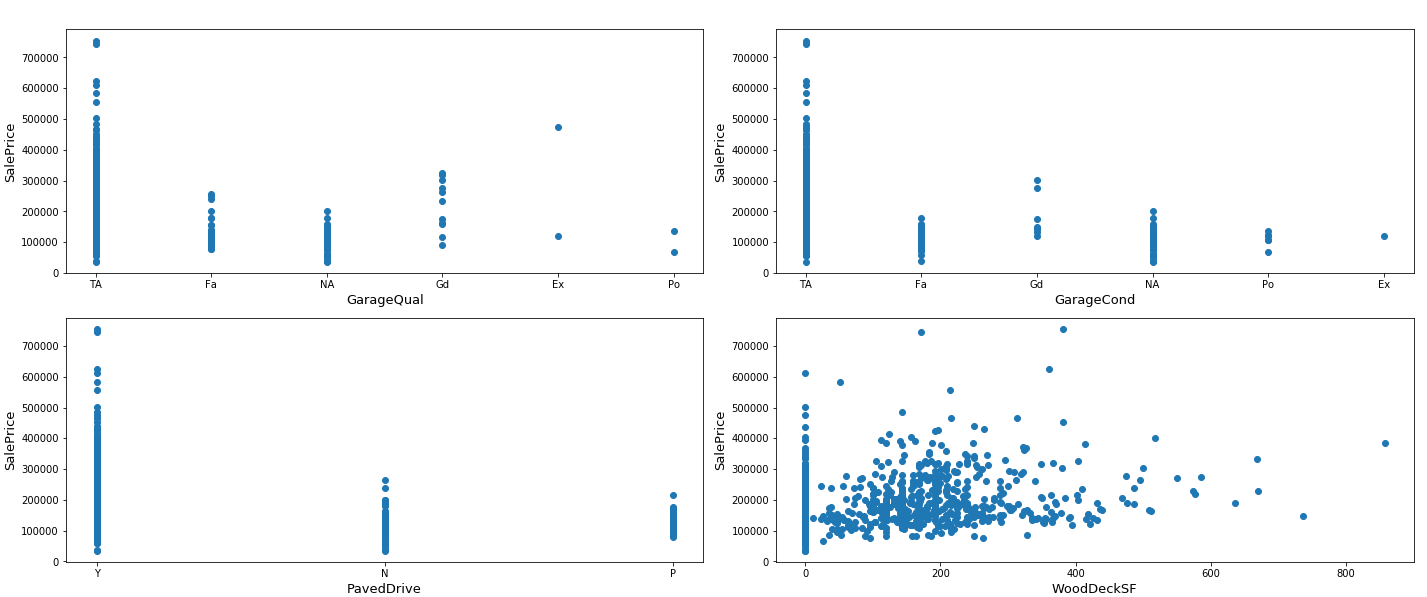
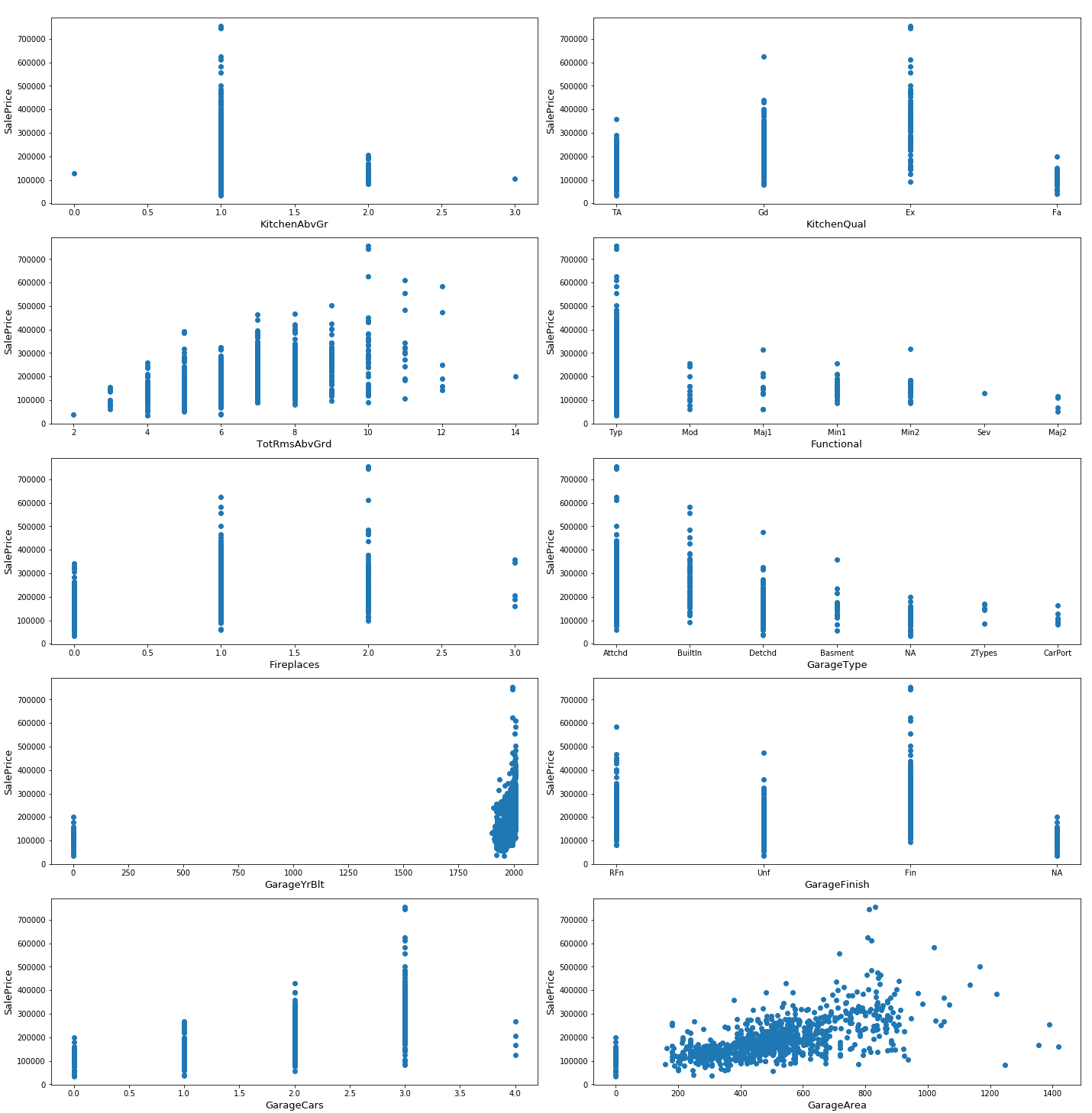














Text

Description automatically generated with medium confidence

**Explorative Visualizations showing how some variables are affecting Sale Price:**

Chart, scatter chart

Description automatically generated

From the above graph we can see that

* OverallQual causes different SalePrice even for the same GrLivearea
* OverallQual is also having linear relationship with saleprice

Graphical user interface, website

Description automatically generated

From the above graph we can see that

* 2ndFlrSF depressed the power of GrLivArea toward SalePrice
* TotalBsmtSF had no impact on SalePrice

Graphical user interface

Description automatically generated with low confidence

Chart, box and whisker chart

Description automatically generated

From the above 2 graphs we can see that

* SalePrice is directly proportional to the number of baths
* Also for Bath(2,1) linearity is improved and spreadness is also decreased for SalePrice-GrLivArea

Chart, treemap chart

Description automatically generated

From Heat map we can say that most people are looking for houses having total rooms of range (4,10) and bedrooms of range (2,4)

Graphical user interface

Description automatically generated

Most of the houses sold are of 3 bedrooms and bedroom itself is not proportional to sale price

Chart

Description automatically generated with medium confidence

Total rooms is having a linear relationship with the sale price, more the rooms saleprice also increases

A screenshot of a computer

Description automatically generated with medium confidence

From the above graphs we can see that

* Most of the Houses are having 2 cars
* Garage area is having a linear relationship with the saleprice
* 0 Cars and 1 Cars has no difference in SalePrice
* From Unit Garagearea we can wee that higher sale prices are maintaining the unit garage area relationship

Chart, bar chart

Description automatically generated

PoolArea, ScreenPorch, 3SsnPorch, Enclosed porch are almost available for every house sold

Chart, bar chart

Description automatically generated

Good quality houses are having more outside amenities

A picture containing chart

Description automatically generated

From the above graph we can see that

* + After 1950 no houses are remodeled
  + Most houses built in 1950 are remodeled
  + Amount of trade is increased in the month of range (Mar-Aug) i.e. Spring & summer season
* **State the set of assumptions (if any) related to the problem under consideration**

One assumption which was considered while filling null values we considered that feature as not available

* **Hardware and Software Requirements and Tools Used**

I have used Jupyter for programming in Python

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I have considered following 3 models for solving the problem and they are.

1. Linear Regression model
2. Support Regression model
3. Decision Tree Regressor
4. Random Forest Regressor

* Testing of Identified Approaches (Algorithms)

For testing and training the algorithm I have use the most common one sklearn.model\_selection and since the data were having different values ranging on a very high note I used standard scalar for scaling the data

Text

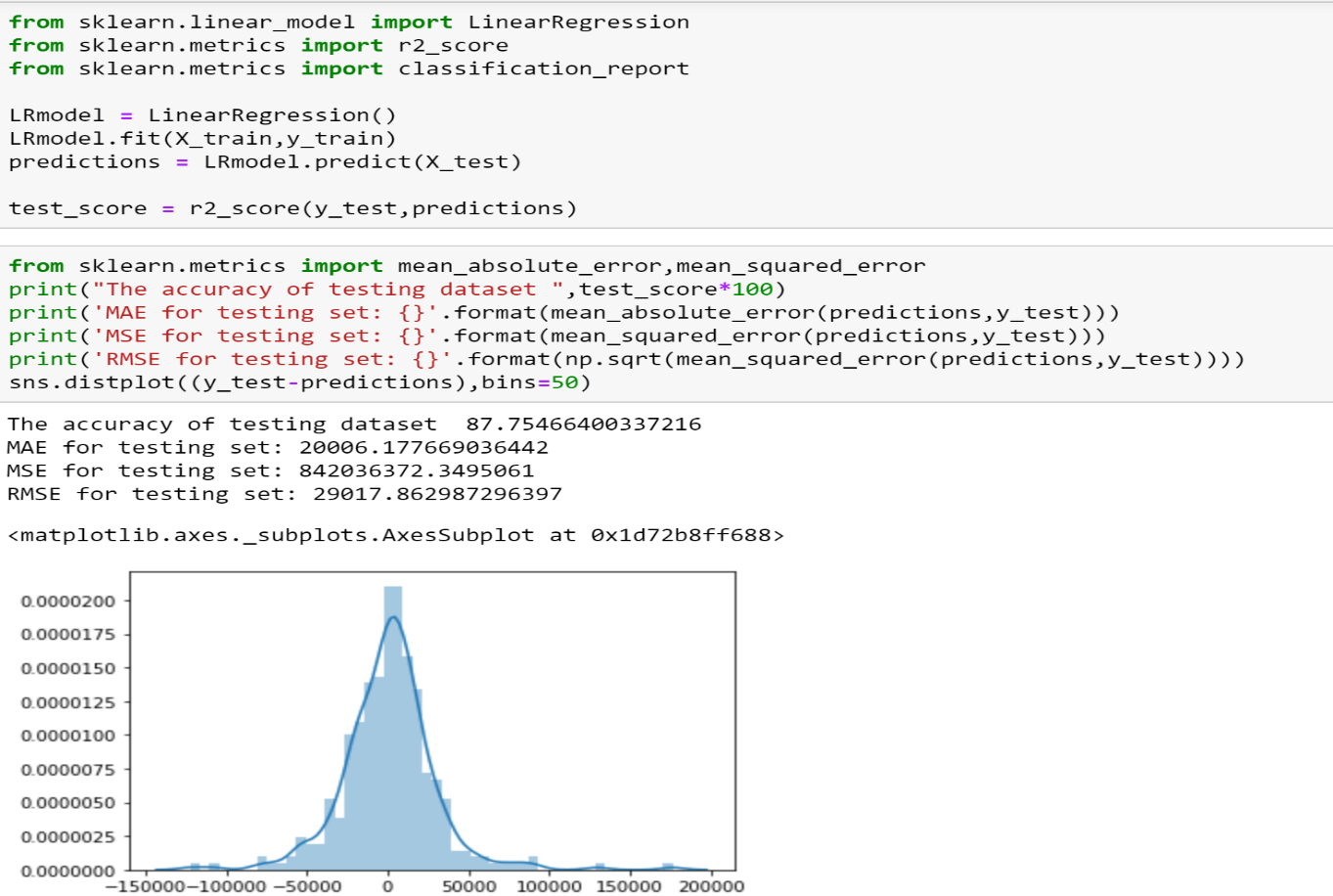
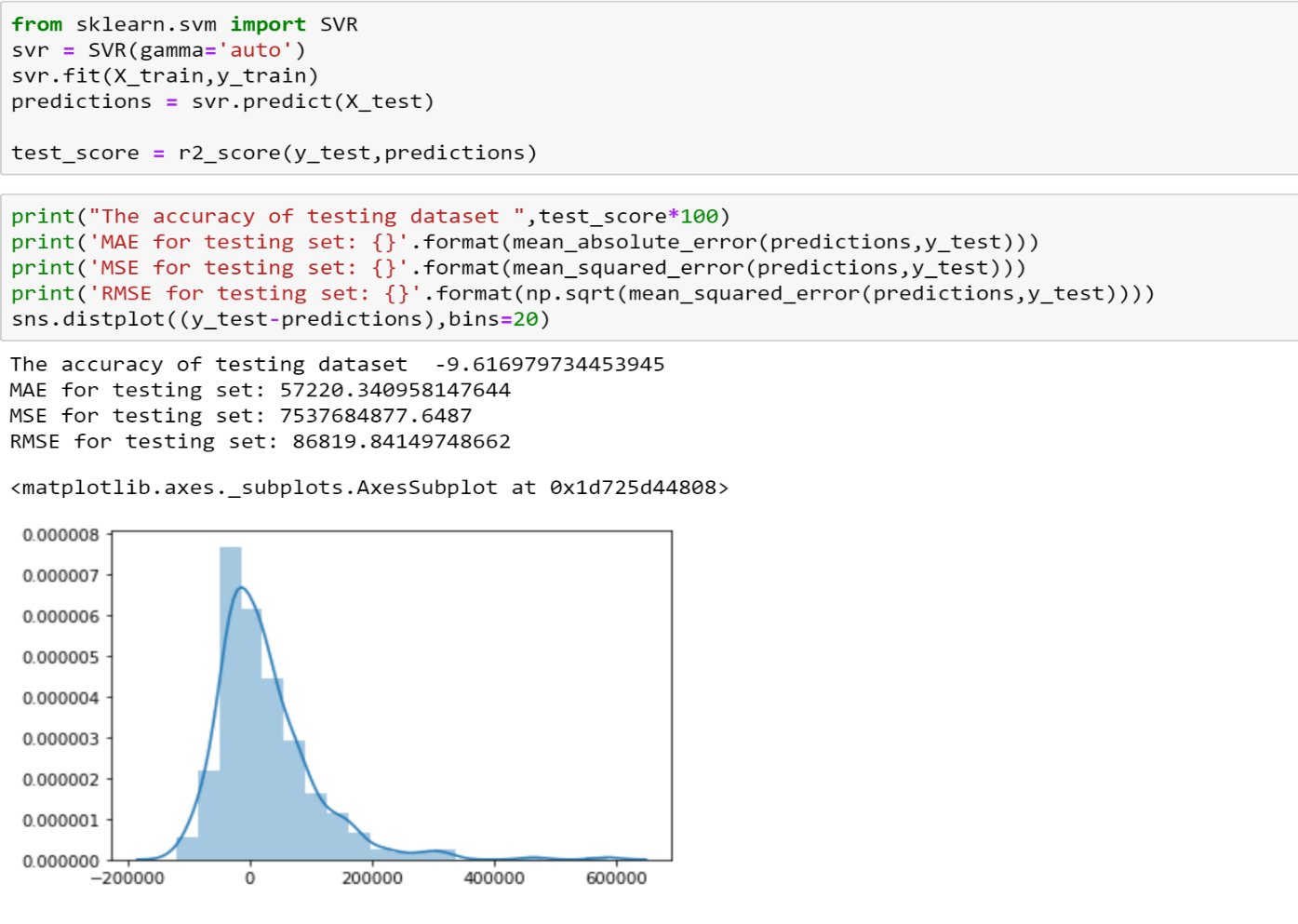
Description automatically generated

Text

Description automatically generated

* **Run and Evaluation of selected models**

Following models were evaluated and their results were given below

1. **Linear Regression Model**
2. **Support Vector Regression**
3. **Decision Tree Regression**

A picture containing chart

Description automatically generated

1. **Random Forest Regression**

A picture containing chart

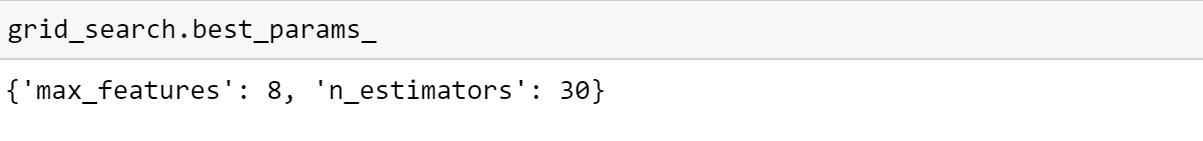
Description automatically generated

Since Random Forest Regression was having high accuracy and also RMSE score is low compared to others this model is selected for Hyperparameter tuning

* **Hyperparameter Tuning**

Text

Description automatically generated



**CONCLUSION**

* Key Findings and Conclusions of the Study
  + The RMSE is 29220 which is comparatively less compared to others
  + {'max\_features': 8, 'n\_estimators': 30} gives the optimum result which should be used for testing
  + Overall model could be improved with more data
  + The dollar impact of a one-unit change in each explanatory variable on the average house price in Ames is listed in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Value** |  | **Feature** | **Value** |
| Street | 57255 |  | Fireplaces | 2435 |
| MSZoning\_C (all) | 25084 |  | HalfBath | 2224 |
| LandContour\_Low | 21604 |  | LotShape\_Reg | 1688 |
| KitchenAbvGr | 16874 |  | GarageCars | 1610 |
| MSZoning\_FV | 14113 |  | TotRmsAbvGrd | 1514 |
| LandContour\_HLS | 13017 |  | BsmtFullBath | 1255 |
| MSZoning\_RH | 8775 |  | LotShape\_IR1 | 475 |
| LandContour\_Bnk | 7279 |  | MoSold | 442 |
| OverallQual | 7249 |  | YrSold | 380 |
| MSZoning\_RL | 6071 |  | CentralAir | 323 |
| OverallCond | 5948 |  | YearBuilt | 288 |
| LotShape\_IR3 | 5640 |  | BsmtHalfBath | 175 |
| LotShape\_IR2 | 4427 |  | GarageYrBlt | 132 |
| FullBath | 4375 |  | YearRemodAdd | 94 |
| BedroomAbvGr | 3948 |  | LotFrontage | 69 |
| MSZoning\_RM | 3876 |  |  |  |