

**HOUSING: PRICE PREDICTION**

Submitted by:

MEHUL SONTHALIA

**ACKNOWLEDGMENT**

Data Sources:- The dataset provided as train.csv & test.csv, Data Description.txt

The learning, practice and evaluation projects and the study material provided at DataTrained Academy as well as Flip Robo Technologies have helped in successful completion of the project

**INTRODUCTION**

* **Business Problem Framing**

A US-Based housing company named Surprise Housing has decided to enter the Australian Market. It is looking at prospective properties to buy houses to enter the market.

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.

Houses are one of the necessary needs of each and every person around the globe, 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.

* **Conceptual Background of the Domain Problem**

The domain related concepts that will be useful for better understanding of the project would be:- Sales, Statistical Analysis, Data Analysis, Machine Learning.

* **Motivation for the Problem Undertaken**

The objective behind the project is to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

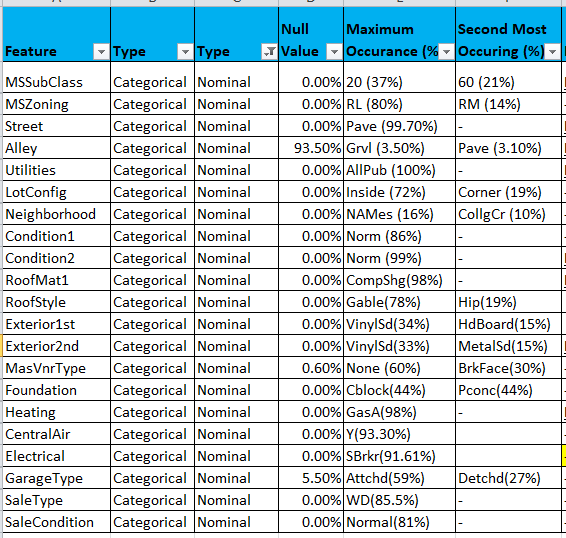
**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

1. Dataset Analysis :-
   1. Dividing the dataset’s variables in 4 data types:-
      1. Categorical (Nominal)
      2. Categorical (Ordinal)
      3. Numerical (Discrete)
      4. Numerical (Continuous)
   2. Noting down each column’s Null Values, Value Counts and understanding each feature’s Distribution.
   3. From the above initial analysis, deciding the unimportant features to be dropped.

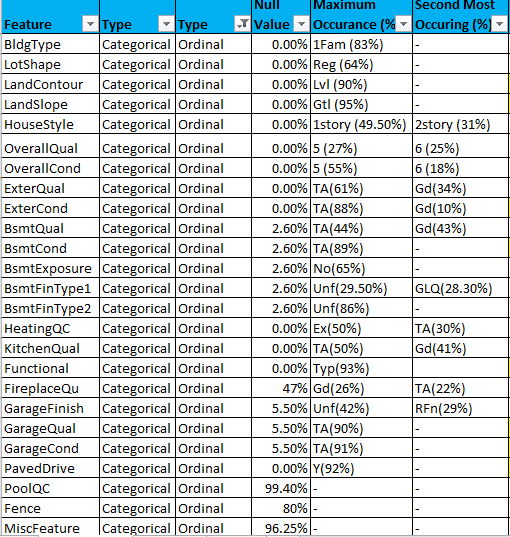
**(A)Dividing the dataset’s variables in 4 data types**

* + 1. **NOMINAL VARIABLES**



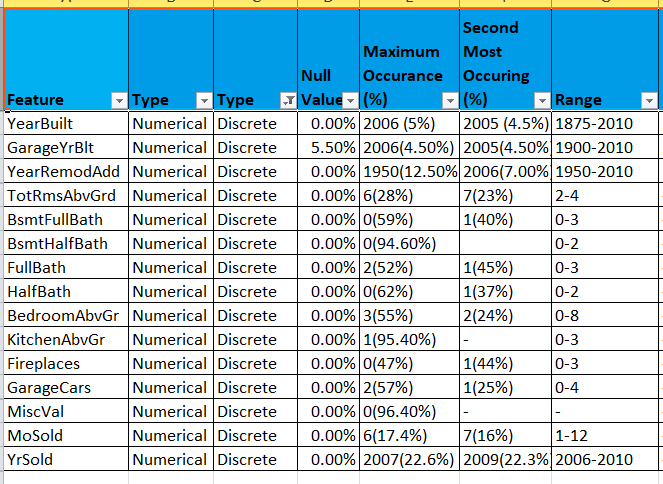
**(A)Dividing the dataset’s variables in 4 data types**

**ii. ORDINAL VARIABLES**



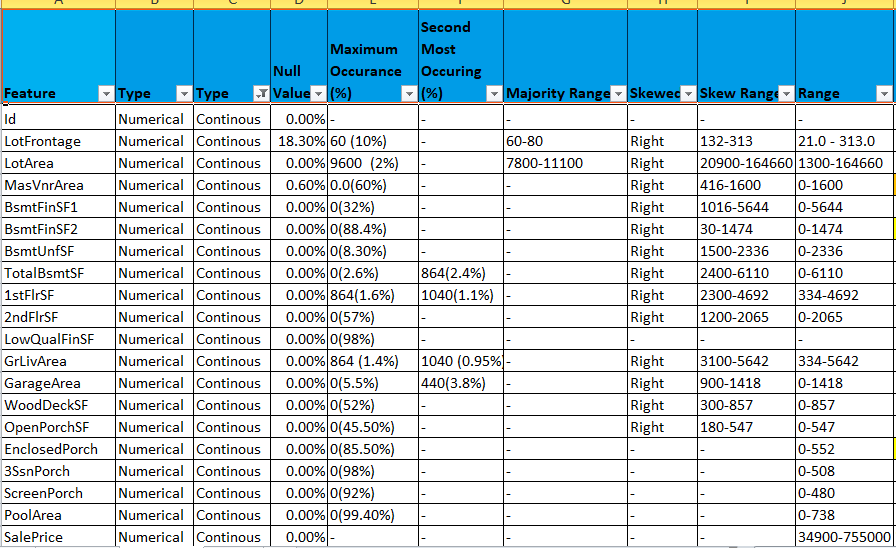
**(A)Dividing the dataset’s variables in 4 data types**

**iii.DISCRETE VARIABLES**

****

**(A)Dividing the dataset’s variables in 4 data types**

**iv. CONTINUOUS VARIABLES**

****

**(B)Noting down each column’s Null Values, Value Counts and understanding each feature’s Distribution.**

**(C)From the above initial analysis, deciding the unimportant features to be dropped.**

OBSERVATIONS :-

1) Id

Numerical Variable

This is just a pointer variable and not useful in predictive analysis.

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

2) MSSubClass

Categorical Variable

Identifies the type of dwelling involved in the sale.

Has no null values

Maximum Occurances :-

20 (1-STORY 1946 & NEWER ALL STYLES) 37%

60 (2-STORY 1946 & NEWER) 21%

50 (1-1/2 STORY FINISHED ALL AGES) 10%

Majority Dwelling = 1-story & 2-story 1946 & NEWER

3) MSZoning

Categorical Variable

Identifies the general zoning classification of the sale.

No Null Values

Maximum Occurances :-

RL (Residential Low Density) 80%

RM (Residential Medium Density) 14%

Majority Zoning = Residential Low Density Zone

4) LotFrontage

Continous Variable

Linear feet of street connected to property

No Null Values

Maximum Occurance = 60.0 (10%)

Majority Street Size = 60 to 80 feet

Right skewed values = 132-313 feet

5) LotArea

Continous Variable

Lot size in square feet

No Null Values

Maximum Occurance = 9600 (2%)

Majority Lot Size = 7800 to 11100 square feet

Right Skewed Values = 20900 to 164660 sq feet

6) Street

Categorical Variable

Type of road access to property

No Null Values

Maximum Occurance = Pave (Paved) (99.7%)

Almost all the roads are paved

7) Alley

Categorical Variable

Type of alley access to property

Null Values = 1091 (93.50%)

Grvl = 41 (3.5%)

Pave = 36 (3.1%)

Since the imputation of values to be done is 93.50%

It is better to drop this column than imputing 93.50% of aritifical values.

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

8) LotShape

Categorical Variable

General shape of property

No Null Values

Maximum Occurance = Reg (Regular) (64%)

Majority Shape = Regular & Slightly Irregular

9) LandContour

Categorical Variable

Flatness of the property

No Null Values

Maximum Occurance = Lvl (Near Flat/Level) (90%)

Most of the properties are Near Flat/Level

10) Utilities

Categorical Variable

Type of utilities available

No Null Values

Maximum Occurance = AllPub (All public Utilities (E,G,W,& S)) 100%

Since all the properites have all the public utilities available

It will make no difference in the predictive analysis.

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

11) LotConfig

Categorical Variable

Lot configuration

No Null Values

Maximum Occurance =

Inside (Inside lot) (72%)

Corner (Corner lot) (19%)

Most of the properties are Inside or Corner Lot

12) LandSlope

Categorical Variable

Slope of property

No Null Values

Maximum Occurance = Gtl (Gentle slope) (95%)

Almost all the properites have a gentle slope

13) Neighborhood

Categorical Variable

Physical locations within Ames city limits

No Null Values

Maximum Occurance =

NAmes (North Ames) (16%)

CollgCr (College Creek) (10%)

14) Condition1

Categorical Variable

Proximity to various conditions

No Null Values

Maximum Occurance = Norm (Normal) (86%)

Most of the properites has a normal proximity

15) Condition2

Categorical Variable

Proximity to various conditions (if more than one is present)

No Null Values

Maximum Occurance = Norm (Normal) (99%)

Almost all the properites have a normal proximity (secondary)

16) BidgType

Categorical Variable

Type of dwelling

No Null Values

Maximum Occurance = 1Fam (Single-family Detached) (83%)

Most of the properites has a single family detached dwelling

17) HouseStyle

Categorical Variable

Style of dwelling

No Null Values

Maximum Occurance =

1story (One story) (49.50%)

2Story (Two story) (31.00%)

Most of the properties have One & Two Stories

18) OverallQual

Categorical Variable (Ordinal)

Rates the overall material and finish of the house

No Null Values

Maximum Occurance =

5 (Average) (27%)

6 (Above Average) (25%)

Majority Material Quality = 4 to 8

19) OverallCond

Categorical Variable (Ordinal)

Rates the overall condition of the house

No Null values

Maximum Occurance =

5 (Average) (55%)

6 (Above Average) (18%)

7 (Good) (15%)

Majority Condition Quality = 5 to 7

20) YearBuilt

Continous Variable (Discrete)

Original construction date

No Null values

Maximum Occurance =

2006 (5%)

2005 (4.50%)

2007 (3.50%)

2004 (3.20%)

2003 (2.80%)

Majority Construction Range = 2002 to 2007

21) YearRemodAdd

Continous Variable (Discrete)

Remodel date (same as construction date if no remodeling or additions)

No Null values

Maximum Occurance =

1950 (12.50%)

2006 (7.00%)

2007 (5.50%)

2005 (5.00%)

2004 (4.00%)

Majority Remodel Range = 1950 & 2004-2007

22) RoofStyle

Categorical Variable

Type of roof

No Null Values

Maximum Occurance =

Gable (78%)

Hip (19%)

Almost all the properties have Gable & Hip roof

23) RoofMat1

Categorical Variable

Roof material

No Null Values

Maximum Occurance = CompShg (Standard (Composite) Shingle) (98%)

Almost all the properites have standard Roof Material

24) Exterior1st

Categorical Variable

Exterior covering on house

No Null Values

Maximum Occurance =

VinylSd (Vinyl Siding)(34%)

HdBoard (Hard Board)(15%)

MetalSd (Metal Siding)(15%)

Wd Sdng (Wood Siding) (15%)

Plywood (Plywood) (8%)

25) Exterior2nd

Categorical Variable

Exterior covering on house (if more than one material)

No Null Values

Maximum Occurance =

VinylSd (Vinyl Siding)(33%)

MetalSd (Metal Siding)(15%)

HdBoard (Hard Board)(14.50%)

Wd Sdng (Wood Siding) (14%)

Plywood (Plywood) (10%)

26) MasVnrType

Categorical Variable

Masonry veneer type

Null Values = 7 (0.6%)

Maximum Occurance =

None (None) (60%)

BrkFace (Brick Face) (30%)

27) MasVnrArea

Continous Variable

Masonry veneer area in square feet

Null Values = 7 (0.6%)

Maximum Occurance = 0.0 (60%)

Right Skewed Range = 416 to 1600

Since, the veneer type is none in 60% properites

the area of veneer in 60% properites is 0

28) ExterQual

Categorical Variable (Ordinal)

Evaluates the quality of the material on the exterior

No Null Values

Maximum Occurance =

TA Average/Typical (61%)

Gd Good (34%)

Most of the properites have good or average quality exterior

29) ExterCond

Categorical Variable (Ordinal)

Evaluates the present condition of the material on the exterior

No Null Values

Maximum Occurance =

TA Average/Typical (88%)

Gd Good (10%)

Majority of the properites have average quality exterior

30) Foundation

Categorical Variable

Type of foundation

No Null Values

Maximum Occurances =

CBlock Cinder Block (44%)

PConc Poured Contrete (44%)

Majority of the properites have CBlock/PConc Foundation

31) BsmtQual

Categorical Variable (Ordinal)

Evaluates the height of the basement

Null Values = 30 (2.6%)

Maximum Occurance =

TA Typical (80-89 inches) (44%)

Gd Good (90-99 inches) (43%)

Height of basement of majority of the houses is 80-99 inches

32) BsmtCond

Categorical Variable (Ordinal)

Evaluates the general condition of the basement

Null Values = 30 (2.6%)

Maximum Occurance =

TA Typical - slight dampness allowed (89%)

Majority of the houses have a typical quality basement

33) BsmtExposure

Categorical Variable

Refers to walkout or garden level walls

Null Values = 31 (2.6%)

Maximum Occurance = No (No Exposure) (65%)

Majority of houses have No Exposure to walkout

34) BsmtFinType1

Categorical Variable (Ordinal)

Rating of basement finished area

NUll Values = 30 (2.6%)

Maximum Occurances =

Unf (Unfinshed) (29.50%)

GLQ (Good Living Quarters) (28.30%)

35) BsmtFinSF1

Continous Variable

Type 1 finished square feet

No Null values

Maximum Occurance = 0 (32%)

Right Skewed Range = 1016 to 5644

Left Skewed at 0

36) BsmtFinType2

Categorical Variable (Ordinal)

Rating of basement finished area (if multiple types)

Null Values = 31 (2.6%)

Maximum Occurance = Unf (Unfinshed) (86%)

37) BsmtFinSF2

Continous Variable

Type 2 finished square feet

No NUll values

Maximum Occurance = 0 (88.4%)

Right Skewed Range = 30 to 1474

38) BsmtUnfSF

Continous Variable

Unfinished square feet of basement area

No NULL Values

Maximum Occurance = 0 (8.30%)

Right Skewed Range = 1500 to 2336

39) TotalBsmtSF

Continous Variable

Total square feet of basement area

No NULL Values

Maximum Occurances=

0 (2.6%)

864 (2.4%)

Left Skewed at 0

Right Skewed Range = 2400 to 6110

40) Heating

Categorical Variable

Type of heating

No NULL Values

Maximum Occurance = GasA (Gas forced warm air furnace) (98%)

Almost all the houses have GasA heating type

41) HeatingQC

Categorical Variable (Ordinal)

Heating quality and condition

No NULL Values

Maximum Occurances=

Ex (Excellent) (50%)

TA (Average/Typical) (30%)

Half of the houses have excellent and 30% average heating quality

42) CentralAir

Categorical Variable

Central air conditioning

No NULL Values

Maximum Occurance = Y (Yes) (93.3%)

Almost all the houses have central ac

43) Electrical

Categorical Variable

Electrical system

No NULL Values

Maximum Occurance = SBrkr (Standard Circuit Breakers & Romex) (91.61%)

Almost all the houses have Standard circuit electrical system

44) 1stFlrSF

Countinous Variable

First Floor square feet

No NULL Values

Maximum Occurances =

864 (1.6%)

1040 (1.1%)

912 (1.1%)

894 (1%)

Right Skewed Range = 2300 to 4692

45) 2ndFlrSF

Countinous Variable

Second Floor square feet

No NULL Values

Maximum Occurances = 0 (57%)

Left skewed at 0

Right skewed range = 1200 to 2065

More than 56% houses do not have a second floor

46) LowQualFinSF

Countinous Variable

Low quality finished square feet (all floors)

No NULL Values

Maximum Occurances = 0 (98%)

only 2% houses have low quality finish

47) GrLivArea

Countinous Variable

Above grade (ground) living area square feet

No NULL Values

Maximum Occurances =

864 (1.4%)

1040 (0.95%)

894 (0.86%)

Right Skewed Range = 3100 to 5642

48) BsmtFullBath

Numerical Variable (discrete)

Basement full bathrooms

No NULL Values

Maximum Occurances =

0 (59%)

1 (40%)

49) BsmtHalfBath

Numerical Variable (discrete)

Basement half bathrooms

No NULL Values

Maximum Occurance = 0 (94.60%)

50) FullBath

Numerical Variable (discrete)

Full bathrooms above grade

No NULL Values

Maximum Occurances =

2 (52%)

1 (45%)

51) HalfBath

Numerical Variable (discrete)

Half baths above grade

No NULL Values

Maximum Occurances =

0 (62%)

1 (37%)

52) BedroomAbvGr

Numerical Variable (discrete)

Bedrooms above grade (does NOT include basement bedrooms)

No NULL Values

Maximum Occurances =

3 (55%)

2 (24%)

4 (15%)

53) KitchenAbvGr

Numerical Variable (discrete)

Kitchens above grade

No NULL Values

Maximum Occurance = 1 (95.40%)

54) KitchenQual

Categorical variable (Ordinal)

Kitchen quality

No NULL Values

Maximum Occurances =

TA (Typical/Average) (50%)

Gd (Good) (41%)

55) TotRmsAbvGrd

Numerical Variable (Discrete)

Total rooms above grade (does not include bathrooms)

No NULL Values

Maximum Occurances=

6 (28%)

7 (23%)

5 (19%)

8 (13%)

56) Functional

Categorical Variable (Ordinal)

Home functionality (Assume typical unless deductions are warranted)

No NULL Values

Maximum Occurance = Typ (Typical Functionality) (93%)

57) Fireplaces

Numerical Variable (Discrete)

Number of fireplaces

No NULL Values

Maximum Occurances=

0 (47%)

1 (44%)

58) FireplaceQu

Categorical Variable (Ordinal)

Fireplace quality

Null Values = 551 (47%)

Maximum Occurances=

Gd (Good - Masonry Fireplace in main level) (26%)

TA (Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement) (22%)

59) GarageType

Categorical Variable

Garage location

Null Values = 64 (5.5%)

Maximum Occurances=

Attchd (Attached to home) (59%)

Detchd (Detached from home) (27%)

60) GarageYrBlt

Numerical Variable (Discrete)

Year garage was built

Null Values = 64 (5.5%)

Maximum Occurances=

2006 (4.50%)

2005 (4.50%)

2007 (3.50%)

2003 (3.20%)

2004 (3.10%)

Maximum Range = 2001 to 2008

61) GarageFinish

Categorical Variable

Interior finish of the garage

Null Values = 64 (5.5%)

Maximum Occurances =

Unf (Unfinished) (42%)

RFn (Rough Finished) (29%)

Fin (Finished) (24%)

62) GarageCars

Numerical Variable (Discrete)

Size of garage in car capacity

No NULL Values

Maximum Occurances =

2 (57%)

1 (25%)

6 (13%)

63) GarageArea

Continous Variable

Size of garage in square feet

No NULL Values

Maximum Occurances =

0 (5.5%)

440 (3.8%)

576 (3.4%)

240 (2.7%)

528 (2.3%)

Left Skewed at 0

Right Skewed Range = 900 to 1418

64) GarageQual

Categorical Variable (Ordinal)

Garage quality

Null Values = 64 (5.5%)

Maximum Occurance = TA (Typical/Average) (90%)

65) GarageCond

Categorical Variable (Ordinal)

Garage condition

Null Values = 64 (5.5%)

Maximum Occurance = TA (Typical/Average) (91%)

66) PavedDrive

Categorical Variable (Ordinal)

Paved driveway

No NULL Values

Maximum Occurance= Y (Paved) (92%)

67) WoodDeckSF

Continous Variable

Wood deck area in square feet

No NULL Values

Maximum Occurance = 0 (52%)

Left Skewed at 0

Right Skewed Range = 300 to 857

68) OpenPorchSF

Continous Variable

Open porch area in square feet

No NULL Values

Maximum Occurance = 0 (45.50%)

Left Skewed at 0

Right Skewed Range = 180 to 547

69) EnclosedPorch

Continous Variable

Enclosed porch area in square feet

No NULL Values

Maximum Occurance = 0 (85.50%)

70) 3SsnPorch

Continous Variable

Three season porch area in square feet

No NULL Values

Maximum Occurance = 0 (98%)

71) ScreenPorch

Continous Variable

Screen porch area in square feet

No NULL Values

Maximum Occurance = 0 (92%)

72) PoolArea

Continous Variable

Pool area in square feet

No NULL Values

Maximum Occurance = 0 (99.40%)

7 other distinct values

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

73) PoolQC

Categorical Variable (Ordinal)

Pool quality

Null Values = 99.40%

3 distinct values with 7 entries only

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

74) Fence

Categorical Variable (Ordinal)

Fence quality

Null Values = 80%

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

75) MiscFeature

Categorical Variable

Miscellaneous feature not covered in other categories

Null Values = 96.25%

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

76) MiscVal

Numerical Value

$Value of miscellaneous feature

No NULL Values

Maximum Occurance = 0 (96.40%)

Since, the misc features does not contain much info

Also, MiscVal has above 96% 0 values

\*\*\*Hence, WE SHALL DROP THIS COLUMN\*\*\*

77) MoSold

Numerical Value (Discrete)

Month Sold (MM)

No NULL Values

Maximum Occurances =

6 (17.4%)

7 (16%)

5 (14.2%)

4 (9.7%)

78) YrSold

Numerical Value (Discrete)

Year Sold (YYYY)

No NULL Values

Maximum Occurances =

2007 (22.6%)

2009 (22.3%)

2006 (21.7%)

2008 (21.2%)

2010 (12.2%)

79) SaleType

Categorical Variable

Type of sale

No NULL Values

Maximum Occurance = WD (Warranty Deed - Conventional) (85.5%)

80) SaleCondition

Categorical Variable

Condition of sale

No NULL Values

Maximum Occurance = Normal (Normal Sale) (81%)

81) SalePrice (TARGET VARIABLE)

Continous Variable

No NULL Value

Range = 34900 to 755000

**Data Sources and their formats**

**Data Description**:-

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

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

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

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

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

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

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)

**Data Preprocessing Done**

* + 1. Dropping Unnecessary Features
    2. Handling Null Values
    3. Handling Null Valued Rows
    4. Imputation
    5. Encoding
    6. Dropping Highly Correlated Features
    7. Identifying & Removing Skewness
    8. Scaling the data

**(i)** **Dropping Unnecessary Features**

1) Id

This is just a pointer variable and not useful in predictive analysis.

6) Street

Type of road access to property

Maximum Occurance = Pave (Paved) (99.7%)

7) Alley

Type of alley access to property

Null Values = 1091 (93.50%)

10) Utilities

Type of utilities available

Maximum Occurance = AllPub (100%)

15) Condition2

Proximity to various conditions (if more than one is present)

Maximum Occurance = Norm (Normal) (99%)

23) RoofMat1

Roof material

Maximum Occurance = CompShg (Standard (Composite) Shingle) (98%)

40) Heating

Type of heating

Maximum Occurance = GasA (Gas forced warm air furnace) (98%)

46) LowQualFinSF

Low quality finished square feet (all floors)

Maximum Occurances = 0 (98%)

70) 3SsnPorch

Three season porch area in square feet

Maximum Occurance = 0 (98%)

71) ScreenPorch

Screen porch area in square feet

Maximum Occurance = 0 (92%)

72) PoolArea

Pool area in square feet

Maximum Occurance = 0 (99.40%)

73) PoolQC

Pool quality

Null Values = 99.40%

74) Fence

Fence quality

Null Values = 80%

75) MiscFeature

Miscellaneous feature not covered in other categories

Null Values = 96.25%

76) MiscVal

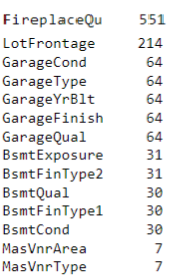
$Value of miscellaneous feature

Maximum Occurance = 0 (96.40%)

We shall drop the above mentioned features who either have 80-100% null values, or above 80% a single value

**(ii) Handling Null (Missing) Values**

* Following are the null values present in the features



**FireplaceQu**

-The most number of null values is in FireplaceQu feature

-That is because in all the 551 (47%) entries, the no of fire places as per the feature Fireplaces=0

-#The quality of a fireplace cannot be filled when there is no fireplace present

-We cannot impute any other value as the FireplaceQu, where there is no fireplace present

-Hence, in order to not misguide the algorithm with a value; We shall drop the feature FireplaceQu

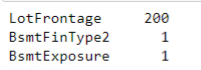
**(iii) Handling Null Valued rows**

Checking the location (rows) of the occurrence of null values in similar features

1. Features related to Garage
2. Features related to Basement
3. Features related to Veneer

* Using the following code to check the occurrence of null values for each feature
* df[df['GarageCond'].isnull() & df['GarageType'].isnull() & df['GarageYrBlt'].isnull() & df['GarageFinish'].isnull() & df['GarageQual'].isnull()]
* df[df['BsmtExposure'].isnull() & df['BsmtFinType2'].isnull() & df['BsmtQual'].isnull() & df['BsmtCond'].isnull() & df['BsmtFinType1'].isnull()]
* df[df['MasVnrType'].isnull() & df['MasVnrArea'].isnull()]
* The output of all the 3 codes states that the null values are occurring at the same row for all the similar features and hence we will be dropping the rows to avoid imputing arbitrary values for all the features.

**(iv) IMPUTATION**



* LotFrontage has 200 null values
* BsmtFinType2 & BsmtExposure each has 1 null value

**(iv) IMPUTATION**

1. **SIMPLE IMPUTER**

* BsmtExposure & BsmtFinType2 both has 1 Null value each , hence we shall impute the values in those entries.
* To impute the null values in this feature, we shall apply Simple Imputer; we will impute the null values with their mode/most frequent

from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='most\_frequent')

columns = ['BsmtFinType2','BsmtExposure']

for i in columns:

imputer = imputer.fit(df[[i]])

df[i] = imputer.transform(df[[i]])

**(iv) IMPUTATION**

1. **ITERATIVE IMPUTER**

* We will apply Iterative imputer for LotFrontage feature as there are 200 null entries in it and iterative imputer will apply regression to impute the values in it.
* We shall pass LotArea as the other parameter to the imputer

from sklearn.experimental import enable\_iterative\_imputer

from sklearn.impute import IterativeImputer

iter\_impute = IterativeImputer()

ite\_imp = pd.DataFrame(np.round(iter\_impute.fit\_transform(df[['LotFrontage','LotArea']])),columns=['LotFrontage','LotArea'])

df[['LotFrontage','LotArea']]=ite\_imp

**(v) ENCODING**

**(A) LABEL ENCODER**

- Label Encoder is applied where there is no particular order in each unique value in the column.

- For nominal variables there is no such order compulsion but in the case of ordinal variables there are orders to be maintained.

- Hence, we will apply label encoder for nominal variables.

- Let us first state nominal variables in a list.

nominal\_var = ['MSZoning','LotConfig','Neighborhood','Condition1','RoofStyle','Exterior1st','Exterior2nd','MasVnrType','Foundation','CentralAir','Electrical','GarageType','SaleType','SaleCondition']

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

for i in nominal\_var:

df[i] = le.fit\_transform(df[i])

**(v) ENCODING**

1. **ORDINAL ENCODER**

* We shall apply ordinal encoder over ordinal variables**.**
* from sklearn.preprocessing import OrdinalEncoder

ordinal\_var = ['LotShape','LandContour','LandSlope','ExterQual','ExterCond','BsmtQual','BsmtCond','BsmtFinType1','BsmtFinType2','BldgType','BsmtExposure','GarageFinish','HouseStyle','HeatingQC','KitchenQual','Functional','GarageQual','GarageCond','PavedDrive']

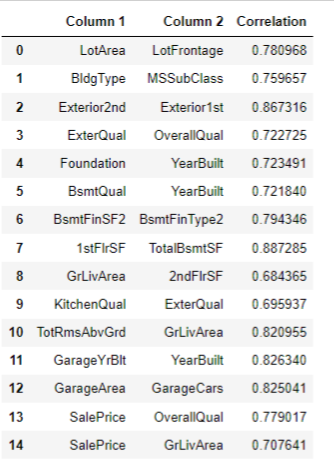
ord\_enc=OrdinalEncoder(categories=[['IR3','IR2','IR1','Reg'],['Low','HLS','Bnk','Lvl'],['Gtl','Mod','Sev'],['Fa','TA','Gd','Ex'],['Fa','TA','Gd','Ex'],['Fa','TA','Gd','Ex'],['Po','Fa','TA','Gd'],['Unf','LwQ','Rec','BLQ','ALQ','GLQ'],['Unf','LwQ','Rec','BLQ','ALQ','GLQ'],['1Fam','2fmCon','Duplex','TwnhsE','Twnhs'],['No','Mn','Av','Gd'],['Unf','RFn','Fin'],['1Story','1.5Fin','1.5Unf','2Story','2.5Fin','2.5Unf','SFoyer','SLvl'],['Po','Fa','TA','Gd','Ex'],['Fa','TA','Gd','Ex'],['Sev','Maj2','Maj1','Mod','Min2','Min1','Typ'],['Po','Fa','TA','Gd','Ex'],['Po','Fa','TA','Gd','Ex'],['N','P','Y']])

df1=ord\_enc.fit\_transform(df[ordinal\_var])

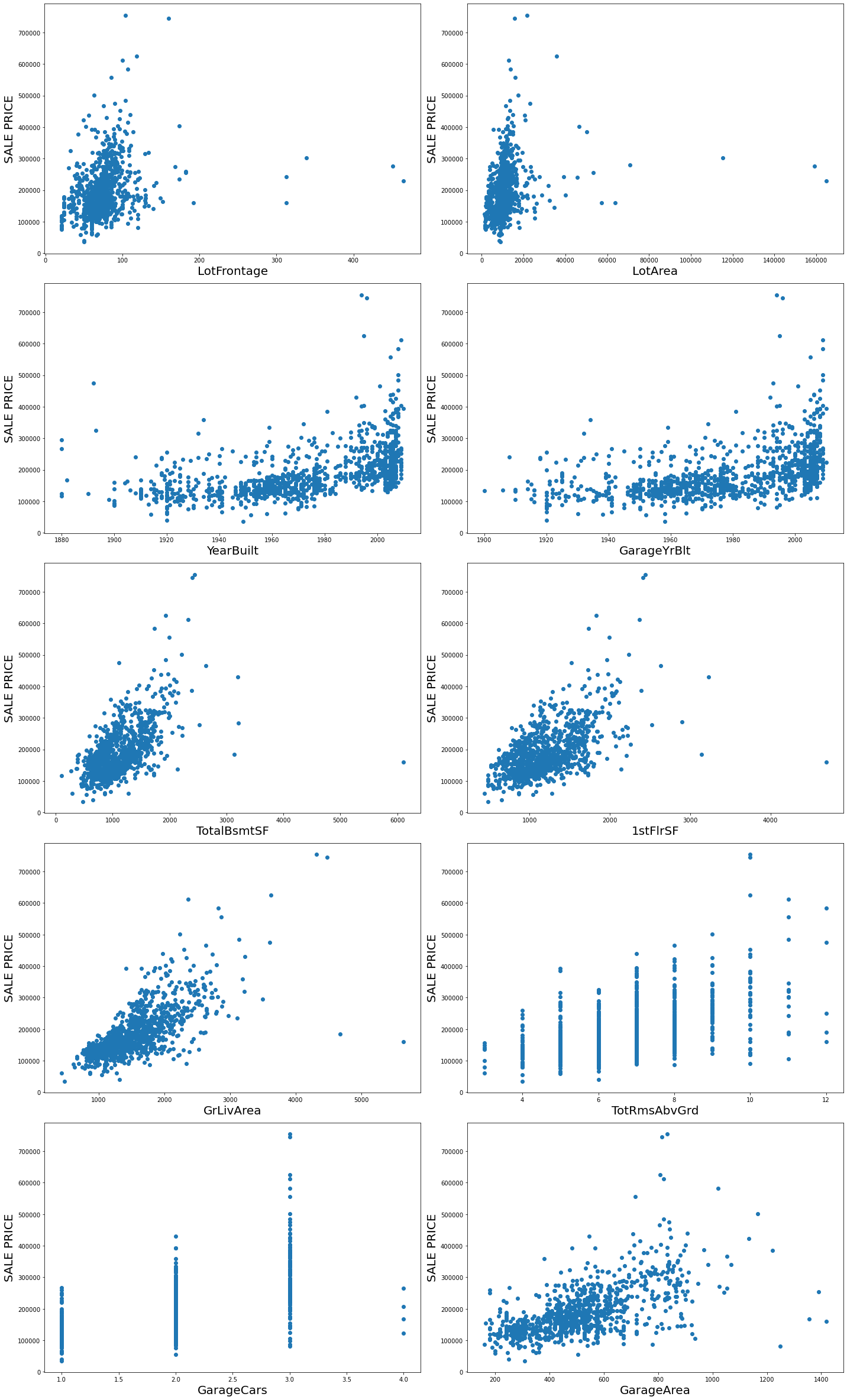
df[ordinal\_var]=df1

**(vi) DROPPING HIGHLY CORRELATED FEATURES**

* Let us check the correlation between each of the columns using df.corr() and heatmap



**(vi) DROPPING HIGHLY CORRELATED FEATURES**

* Checking the relation between highly correlated numerical features vs label :-
* 

**(vi) DROPPING HIGHLY CORRELATED FEATURES**

* Observations from the scatter plot b/w numerical variables and target variable :-

(#**1) LotArea LotFrontage 0.780968**

LotArea seems to have a better relationship with the target variable

(#2) **GarageYrBlt YearBuilt 0.826340**

Both have same relation,we will drop GarageYrBlt

as YearBuilt is about the entire house and GarageYrBlt is only about the Garage

(#3) **1stFlrSF TotalBsmtSF 0.887285**

TotalBsmtSF has a little better relation

(#4) **TotRmsAbvGrd GrLivArea 0.820955**

GrLivArea has better relationship

(#5) **GarageArea GarageCars 0.825041**

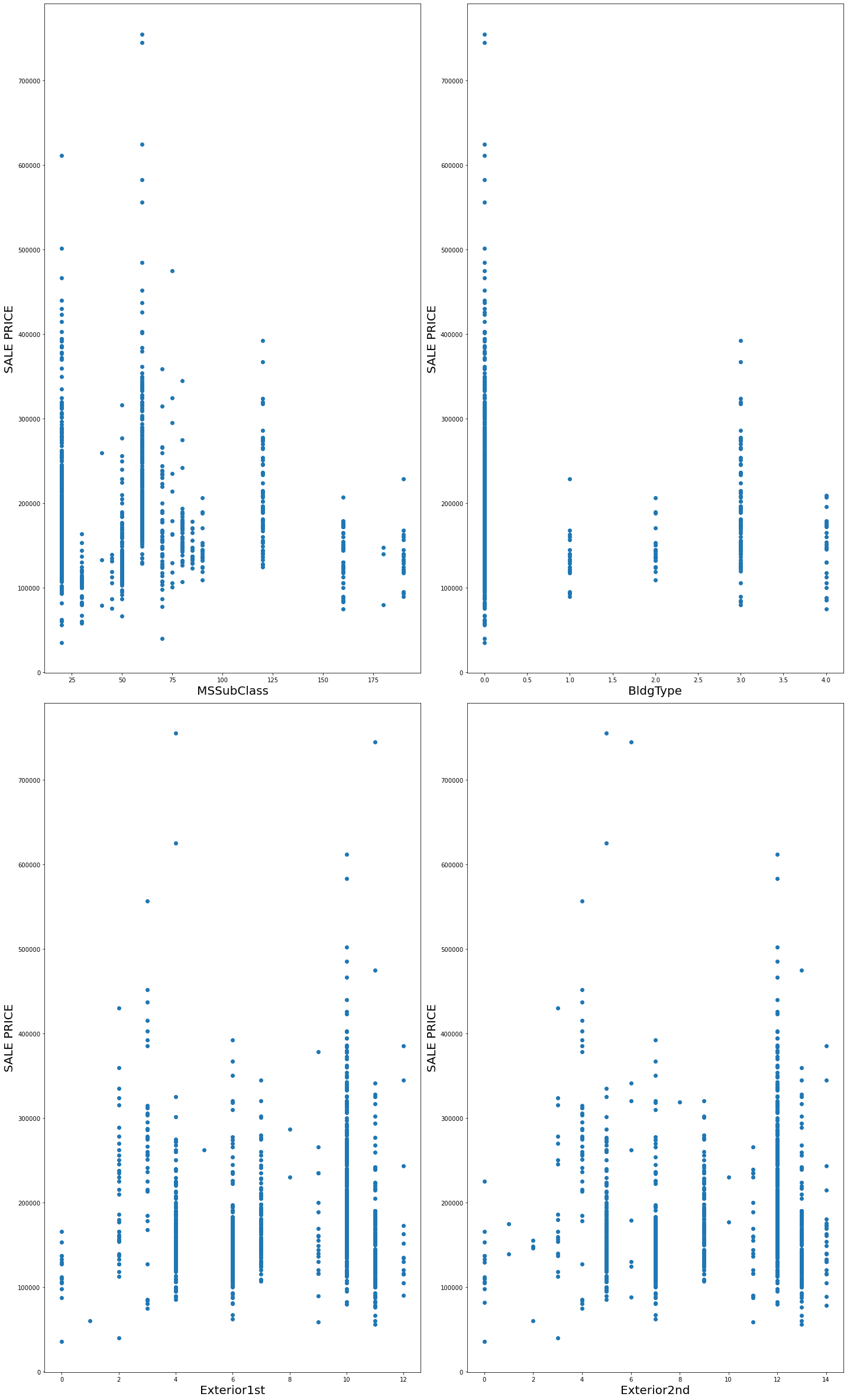
GarageArea has a better relationship

(#6) **BsmtFinSF2 BsmtFinType2 0.794346**

BsmtFinSF2 has a better relationship

**(vi) DROPPING HIGHLY CORRELATED FEATURES**

* Checking the relation between highly correlated categorical features vs label :-



Observations from the plot :-

#1 **BldgType MSSubClass 0.759657**

MSSubClass has a better relationship

#2 **Exterior2nd Exterior1st 0.867316**

Both have the same relationship, hence shall drop Exterior2nd

Just as they both denote similar thing

**(vi) DROPPING HIGHLY CORRELATED FEATURES**

* Cross Checking the important features to be selected by applying SelectKBest Feature Selection method

X=df.drop(columns=['SalePrice'])

y=df['SalePrice']

#Using SelectKBest feature seleciton Method

from sklearn.feature\_selection import SelectKBest,f\_classif

best\_features=SelectKBest(f\_classif, k=64)

fit=best\_features.fit(X,y)

df\_scores=pd.DataFrame(fit.scores\_)

df\_columns=pd.DataFrame(X.columns)

#concatenate dataframes

feature\_scores = pd.concat([df\_columns,df\_scores],axis=1)

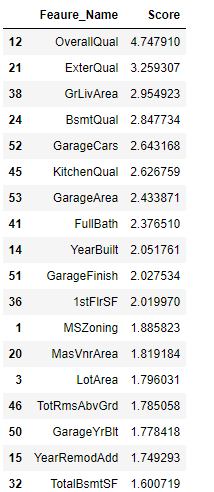
#name output columns

feature\_scores.columns = ['Feaure\_Name','Score']

#print 64 best features

top\_features = feature\_scores.nlargest(64,'Score')

print(feature\_scores.nlargest(64,'Score'))

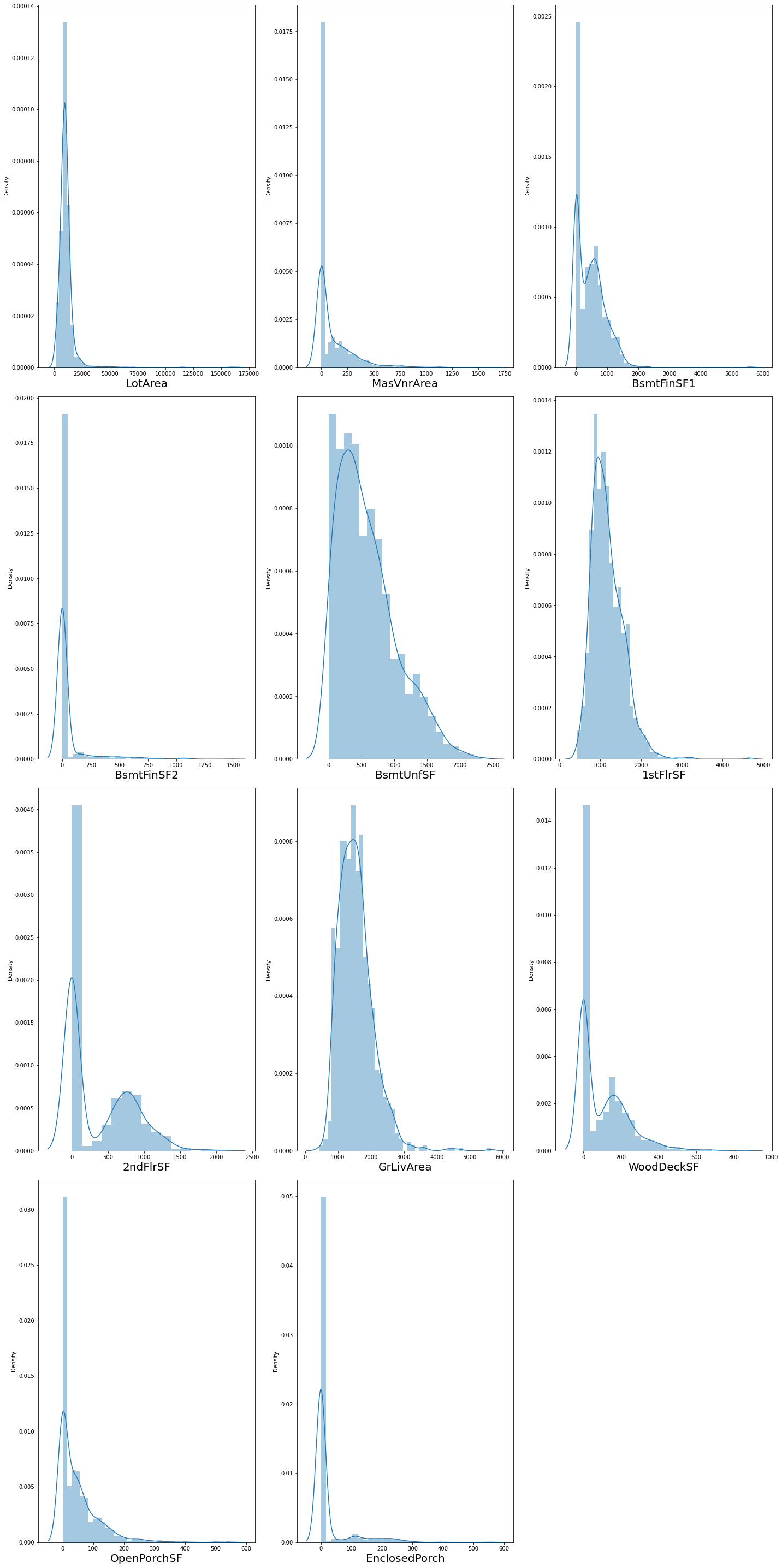


**(vi) DROPPING HIGHLY CORRELATED FEATURES**

* From the above observations taken from the correlation chart and feature importance by applying SelectKBest we are dropping the following features :-
* columns\_drop=['LotFrontage','GarageYrBlt','TotalBsmtSF','TotRmsAbvGrd','GarageArea','BsmtFinType2','MSSubClass','Exterior2nd']
* df.drop(columns=columns\_drop,inplace=True)
* df.shape
* (1073, 57)

**(vii) Identifying & Removing Skewness**

* Checking the distribution of the continuous variables via distribution plot



**(vii) Identifying & Removing Skewness**

* LotArea = High values of 0,right skewed
* MasVnrArea = High values of 0, righ skewed
* BsmtFinSF1 = High values of 0, right skewed
* BsmtFinSF2 = Very high values of 0, right skewed
* BsmtUnfSF = Slightly skewed
* 1stFlrSF = Slightly Skewed
* 2ndFlrSF = Very high values of 0, right skewed
* GrLivArea = right skewed
* WoodDeckSF = Very high values of 0, right skewed
* OpenPorchSF = Very high values of 0, right skewed
* EnclosedProch = Very high values of 0, right skewed

**(vii) Identifying & Removing Skewness**

* We will then check the skewness of the continuous variables by applying the .skew() method
* #Checking skewness of the features

df[cont\_columns].skew().sort\_values(ascending=False)



* As we can observe from the above output that
* LoArea has very high value of skewness
* BsmfFinSF2, EnclosedProch, MasVnrArea, OpenPorchSF, BsmtFinsF1, 1stFlrSF, GrLivArea, WoodDeckSF too have high skewness
* We will now apply Powertransformer() to reduce the skewness in the continuous variables.

#We can see skewness in few of our columns, we will remove the skewness using power\_transform function

from sklearn.preprocessing import power\_transform

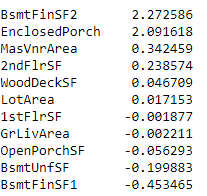
df\_cont\_new = power\_transform(df\_cont)

df\_cont=pd.DataFrame(df\_cont\_new,columns=df\_cont.columns)

**(vii) Identifying & Removing Skewness**

* Checking the skewness post transformation

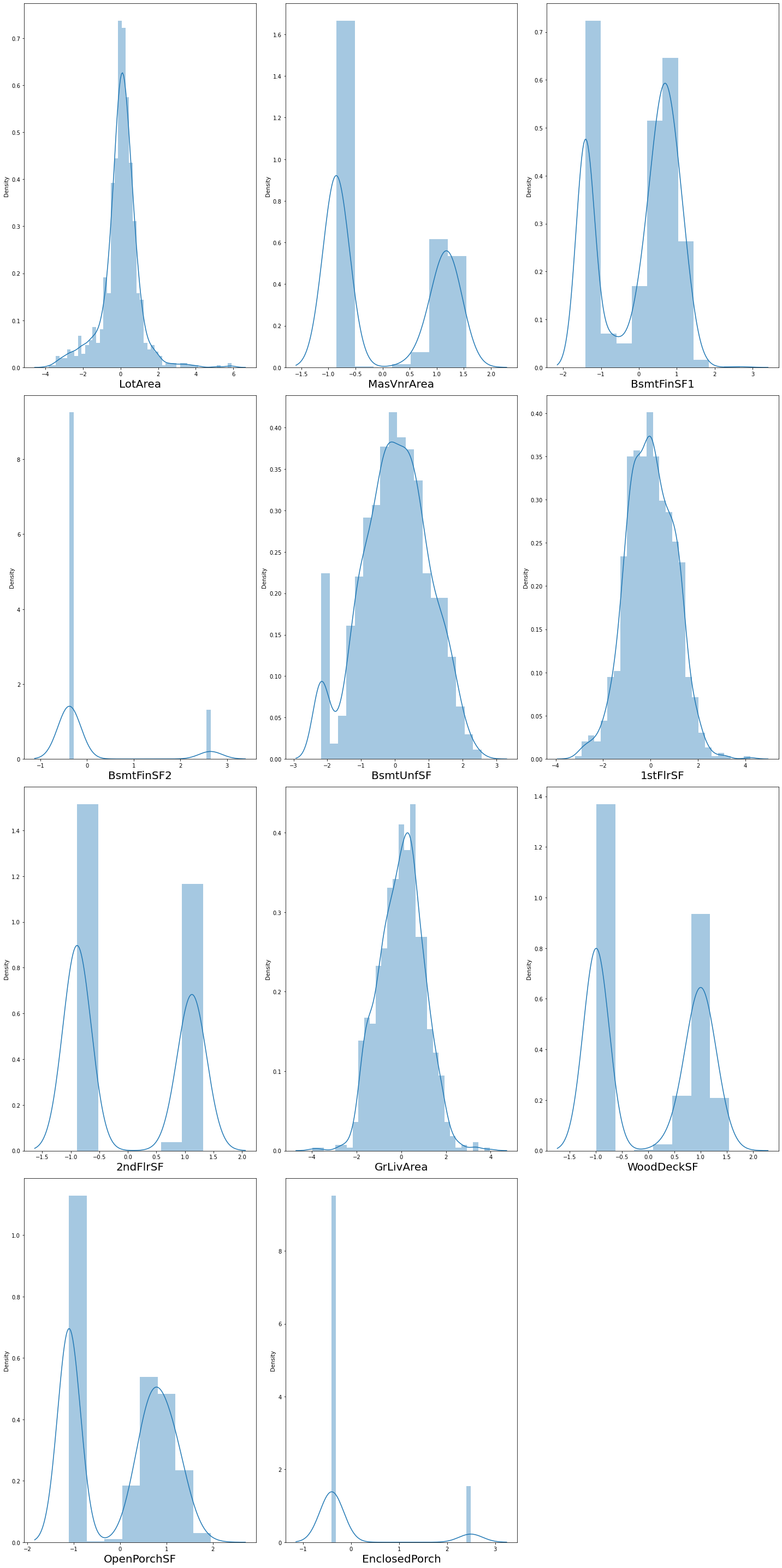
df\_cont.skew().sort\_values(ascending=False)



* BsmtFinsSF2, EnclosedPorch have skewness left only rest all the features do not have skewness left in them

**(vii) Identifying & Removing Skewness**

* Validating the skewness using distribution plot



**(viii)Scaling the data**

* Dividing into Feature and label

X=df.drop(columns=['SalePrice'])

y=df['SalePrice']

* Scaling the Features using StandardScaler()

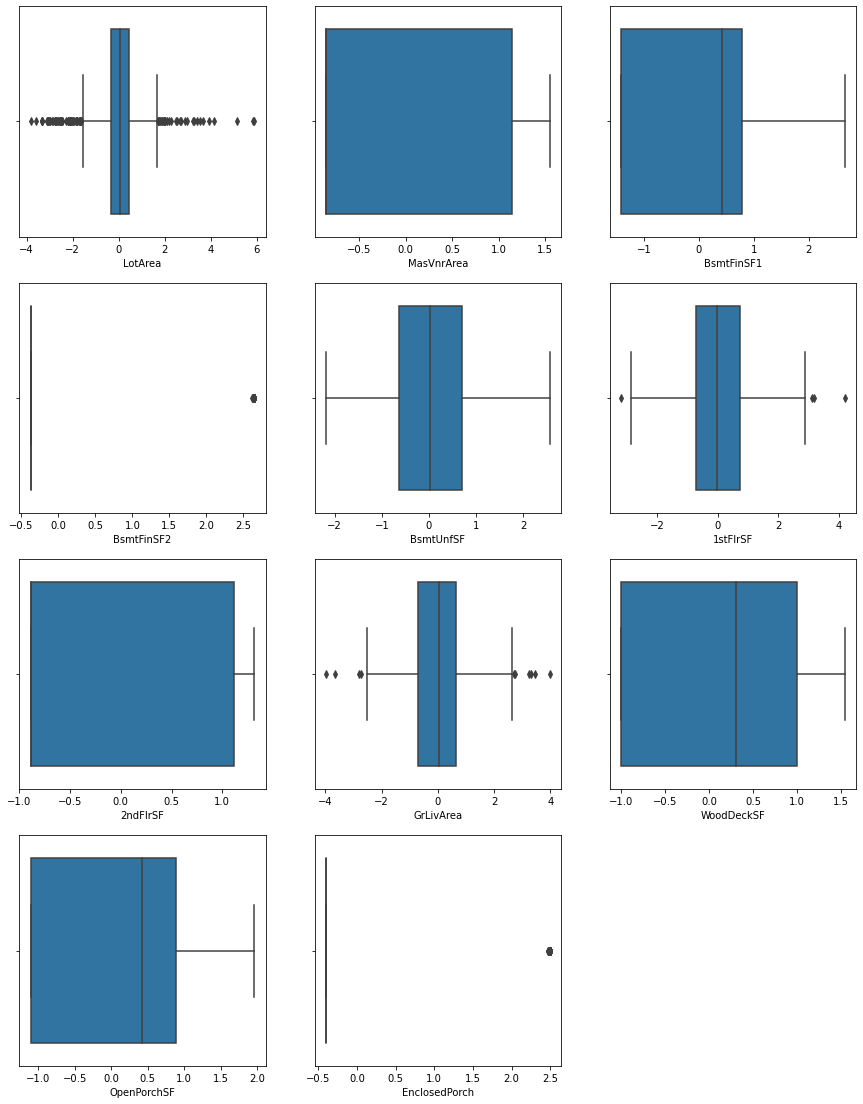
scaler = StandardScaler()

X= scaler.fit\_transform(X)

**Data Inputs- Logic- Output Relationships**

* As per the output of the correlation matrix , following two features are highly correlated with the output.
* 8 SalePrice OverallQual(4.75) 0.779017
* 9 SalePrice GrLivArea (2.95) 0.707641
* OverallQual has a positive correlation with SalePrice of 0.78
* GrLivArea has a positive correlation with SalePrice of 0.71
* According to the SelectKBest Feature Selection Method,
* OverallQual has an importance rating of 4.75
* GrLivArea has an importance rating of 2.95
* As per the above observation, both OverallQual & GrLivArea are important in predicting the SalePrice output and are also positive correlated with the output.

**State the set of assumptions (if any) related to the problem under consideration**



* LotArea has a lot of outliers in it, but because this feature describes the lot size in square feet there would be many houses with different lot sizes and hence we should not drop that data.
* That is why, we will move ahead with our modelling

**Hardware and Software Requirements and Tools Used**

* Following are the libraries imported for the successful implementation of the project

import pandas as pd

import numpy as np

from sklearn.linear\_model import LinearRegression

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import AdaBoostRegressor

import xgboost as xgb

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split,cross\_val\_score,GridSearchCV

from sklearn.feature\_selection import SelectKBest,f\_classif

import seaborn as sns

import matplotlib.pyplot as plt

import pickle

import warnings

warnings.filterwarnings('ignore')

* Pandas used to import the dataset
* Numpy used mathematical calculations
* LinearRegression, LogisticRegression, DecisionTreeRegressor, RandomForestRegressor, AdaBoostRegressor, XGBoostRegressor used for building Machine Learning Model
* StandardScaler for data scaling
* Train\_test\_split to split the dataset into training & testing data
* Cross\_val\_score used for cross validation of the ML Model
* GridSearchCV used for HyperParameterTuning of the Selected Model
* SelectKBest, f\_classif for feature selection
* Seaborn & matplotlib for visualization
* Pickle to save the model

**Model/s Development and Evaluation**

**Testing of Identified Approaches (Algorithms)**

* LinearRegression,
* LogisticRegression,
* DecisionTreeRegressor,
* RandomForestRegressor,
* AdaBoostRegressor,
* XGBoostRegressor

**Run and Evaluate selected models**

**Linear Regressor**

#Linear Regression

lr.fit(x\_train,y\_train)

y\_pred=lr.predict(x\_test)

r2score=r2\_score(y\_test,y\_pred)\*100

print("R2Score :",r2score)

scr = cross\_val\_score(lr,X,y,cv=5)

from sklearn import metrics

print('MAE:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('MSE:', metrics.mean\_squared\_error(y\_test, y\_pred))

rmse = np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred))

print('RMSE:',rmse )

model\_name.append('Linear Regression')

r2\_scores.append(r2score)

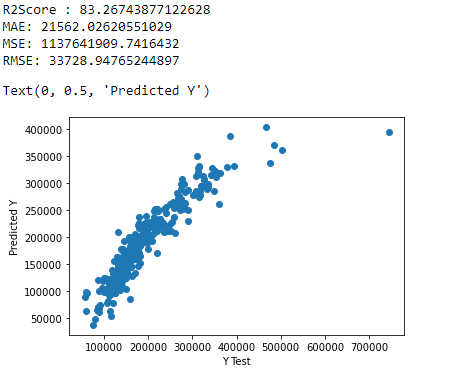
rmse\_value.append(rmse)

cvs.append(scr.mean())

plt.scatter(x=y\_test,y=y\_pred)

plt.xlabel('Y Test')

plt.ylabel('Predicted Y')

****

**Decision Tree Regressor**

#Decision Tree Regressor

from sklearn.tree import DecisionTreeRegressor

dtr=DecisionTreeRegressor()

dtr.fit(x\_train,y\_train)

y\_pred=dtr.predict(x\_test)

r2score=r2\_score(y\_test,y\_pred)\*100

scr2 = cross\_val\_score(dtr,X,y,cv=5)

print("R2Score:", r2score)

from sklearn import metrics

print('MAE:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('MSE:', metrics.mean\_squared\_error(y\_test, y\_pred))

rmse = np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred))

print('RMSE:',rmse )

model\_name.append('Decision Tree Regressor')

r2\_scores.append(r2score)

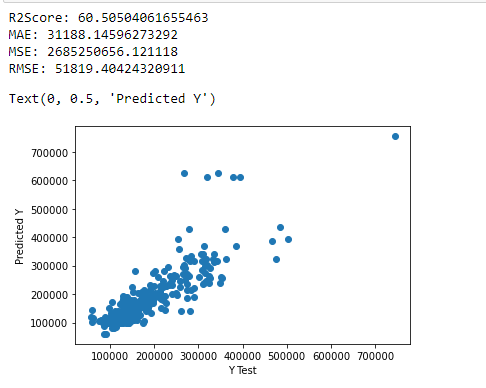
rmse\_value.append(rmse)

cvs.append(scr2.mean())

plt.scatter(x=y\_test,y=y\_pred)

plt.xlabel('Y Test')

plt.ylabel('Predicted Y')

****

**Random Forest Regressor**

#RandomForest Regressor

from sklearn.ensemble import RandomForestRegressor

rdr = RandomForestRegressor()

rdr.fit(x\_train,y\_train)

y\_pred=rdr.predict(x\_test)

r2score=r2\_score(y\_test,y\_pred)\*100

scr3 = cross\_val\_score(rdr,X,y,cv=5)

print("R2Score: ", r2score)

from sklearn import metrics

print('MAE:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('MSE:', metrics.mean\_squared\_error(y\_test, y\_pred))

rmse = np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred))

print('RMSE:',rmse )

model\_name.append('Random Forest Regressor')

r2\_scores.append(r2score)

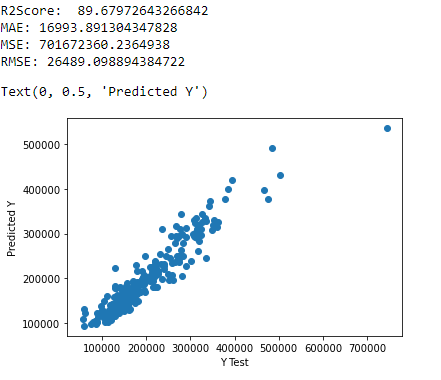
rmse\_value.append(rmse)

cvs.append(scr3.mean())

plt.scatter(x=y\_test,y=y\_pred)

plt.xlabel('Y Test')

plt.ylabel('Predicted Y')

****

**ADA Boost Regressor**

from sklearn.ensemble import AdaBoostRegressor

ada = AdaBoostRegressor()

ada.fit(x\_train,y\_train)

y\_pred=ada.predict(x\_test)

r2score=r2\_score(y\_test,y\_pred)\*100

print("R2 Score: ", r2score)

scr5 = cross\_val\_score(ada,X,y,cv=5)

from sklearn import metrics

print('MAE:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('MSE:', metrics.mean\_squared\_error(y\_test, y\_pred))

rmse = np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred))

print('RMSE:',rmse )

model\_name.append('ADA Boost')

r2\_scores.append(r2score)

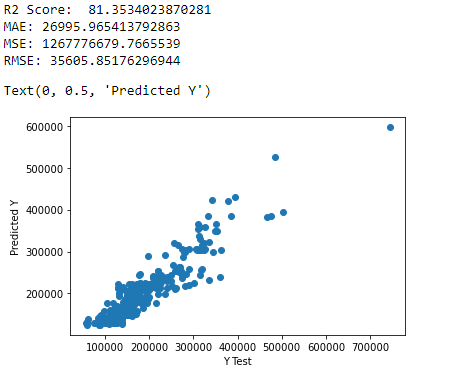
rmse\_value.append(rmse)

cvs.append(scr5.mean())

plt.scatter(x=y\_test,y=y\_pred)

plt.xlabel('Y Test')

plt.ylabel('Predicted Y')

****

**XGBoost Regressor**

#XGB

import xgboost as xgb

xgb = xgb.XGBRegressor()

xgb.fit(x\_train,y\_train)

y\_pred = xgb.predict(x\_test)

r2score=r2\_score(y\_test,y\_pred)\*100

print("R2 Score: ",r2score)

scr6 = cross\_val\_score(xgb,X,y,cv=5)

from sklearn import metrics

print('MAE:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('MSE:', metrics.mean\_squared\_error(y\_test, y\_pred))

rmse = np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred))

print('RMSE:',rmse )

model\_name.append('XGBoost')

r2\_scores.append(r2score)

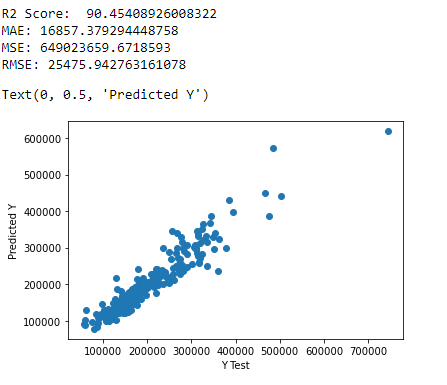
rmse\_value.append(rmse)

cvs.append(scr6.mean())

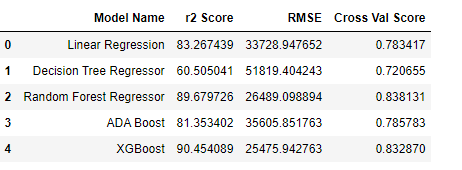
plt.scatter(x=y\_test,y=y\_pred)

plt.xlabel('Y Test')

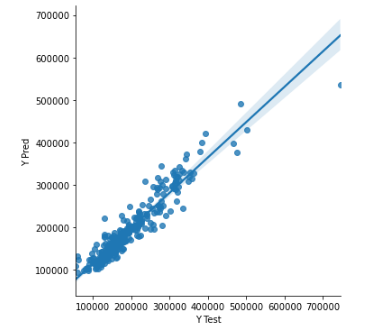
plt.ylabel('Predicted Y')

****

**MODEL DASHBOARD**

****

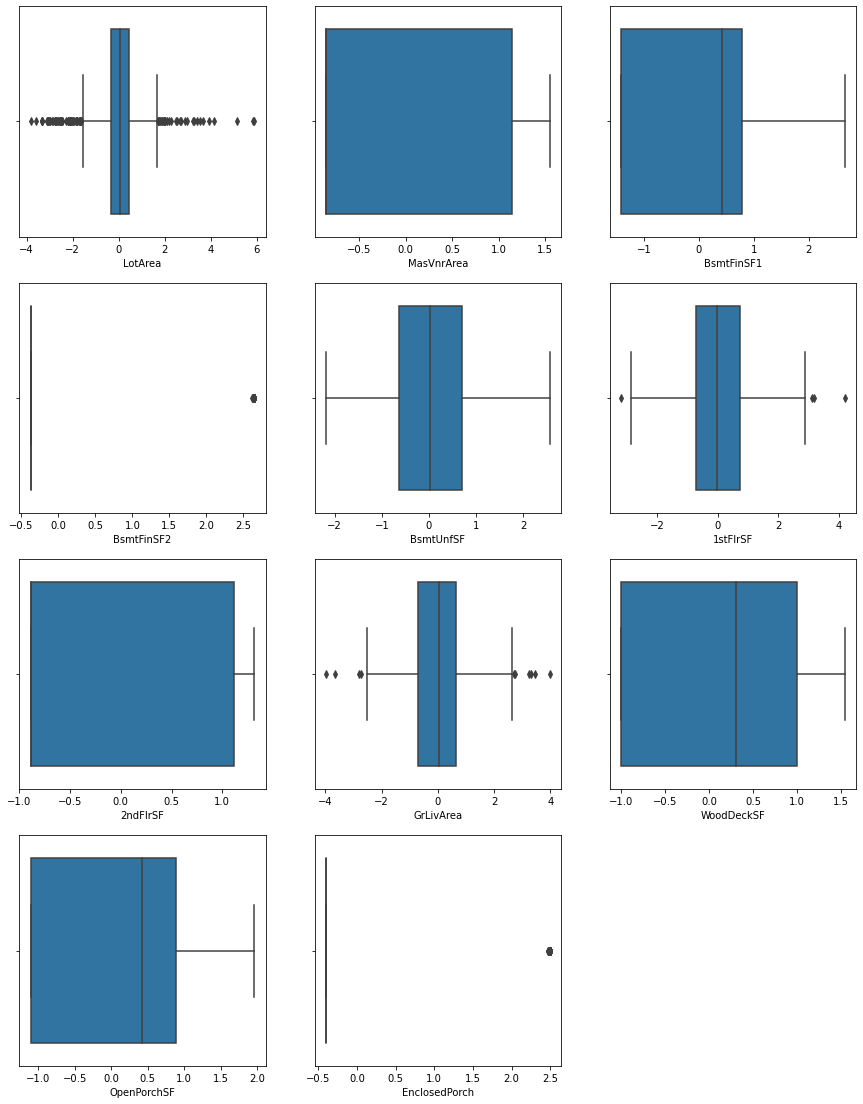
* **Random Forest Regressor has 89.68 r2score and 0.84 as its cv score.**



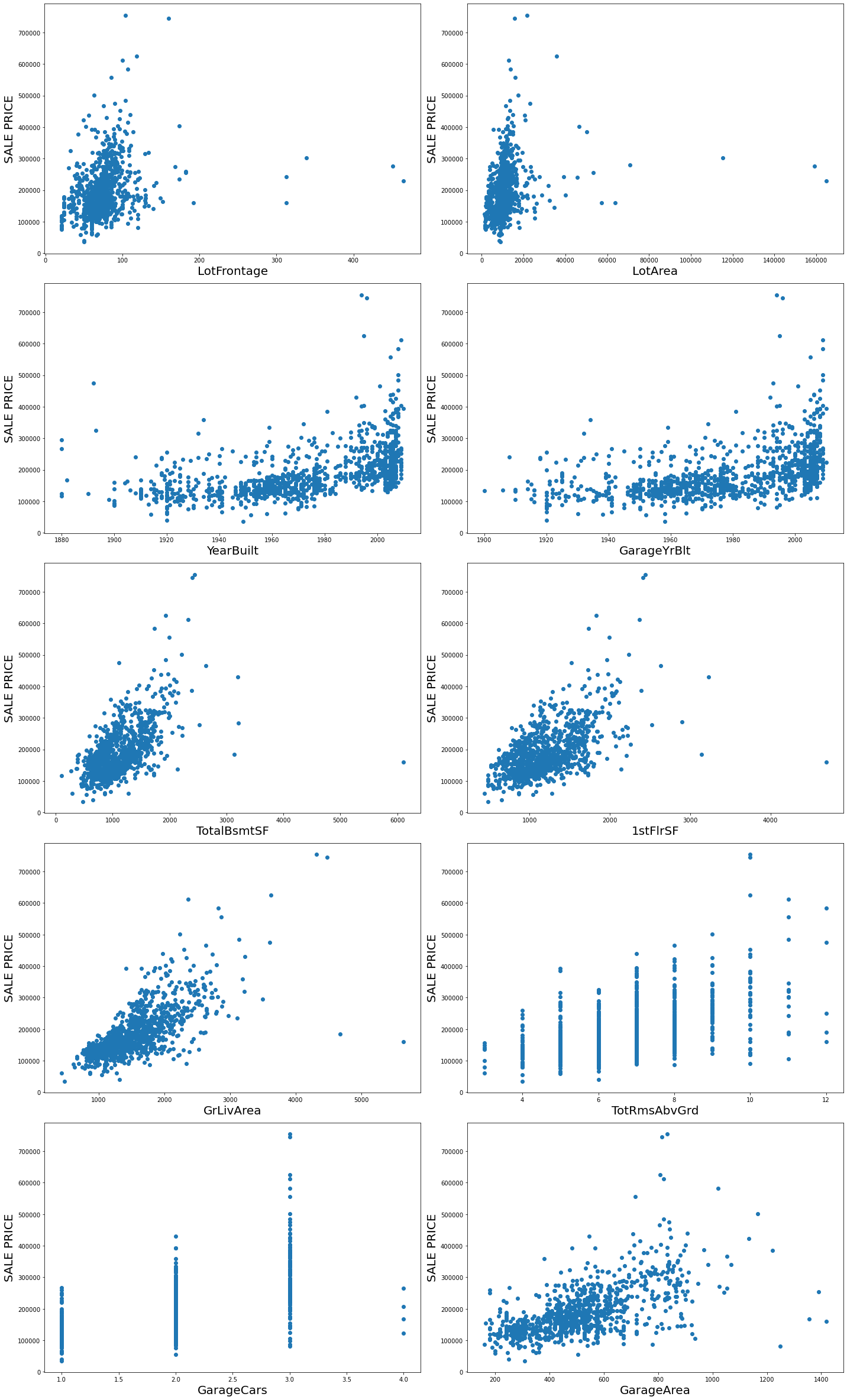
**Key Metrics for success in solving problem under consideration**

* **Data Cleaning**
* **Data Exploration & Analysis**
* **Feature Engineering**

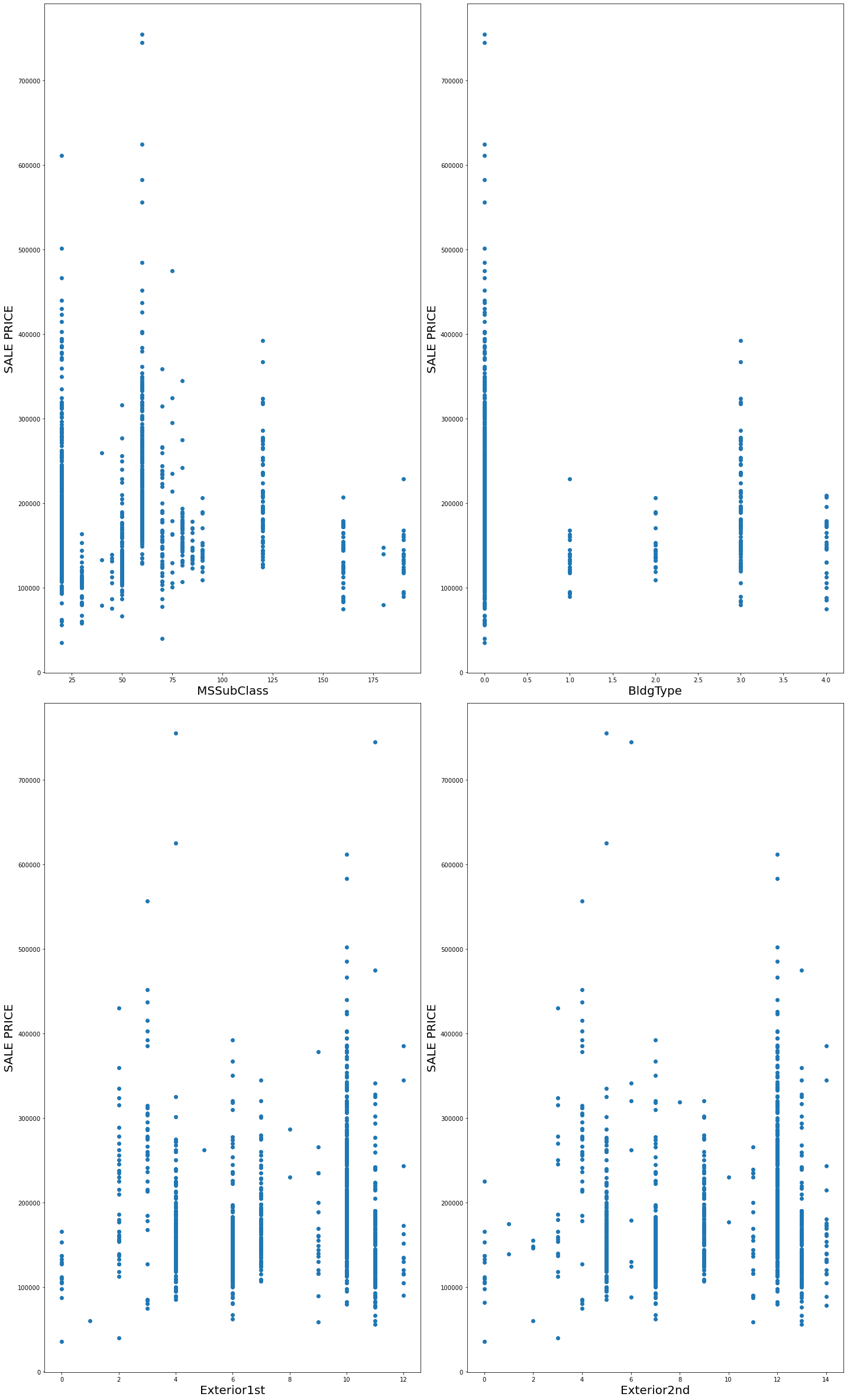
**Visualizations**

**- Box Plot**

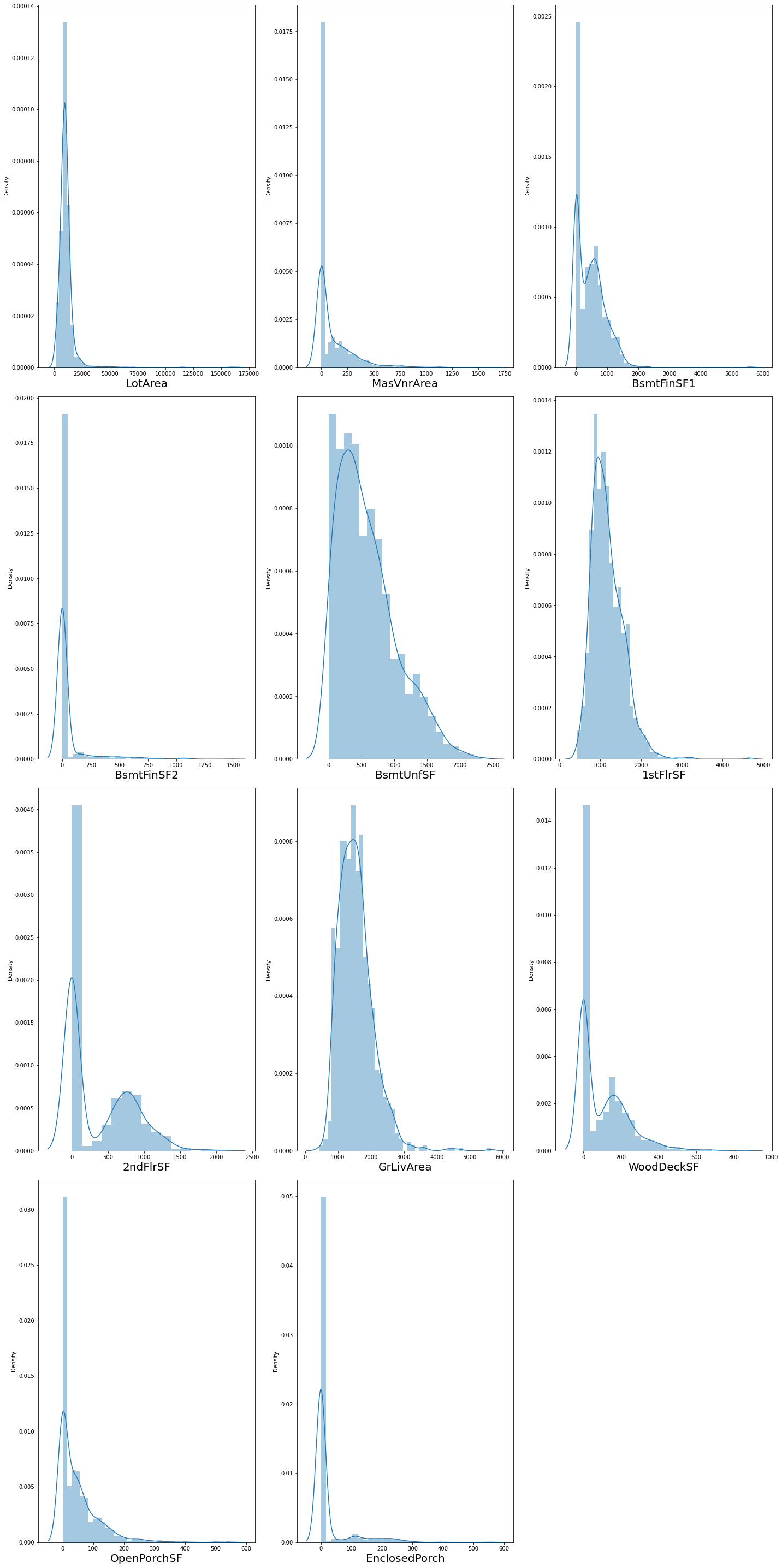
**Scatter Plot**



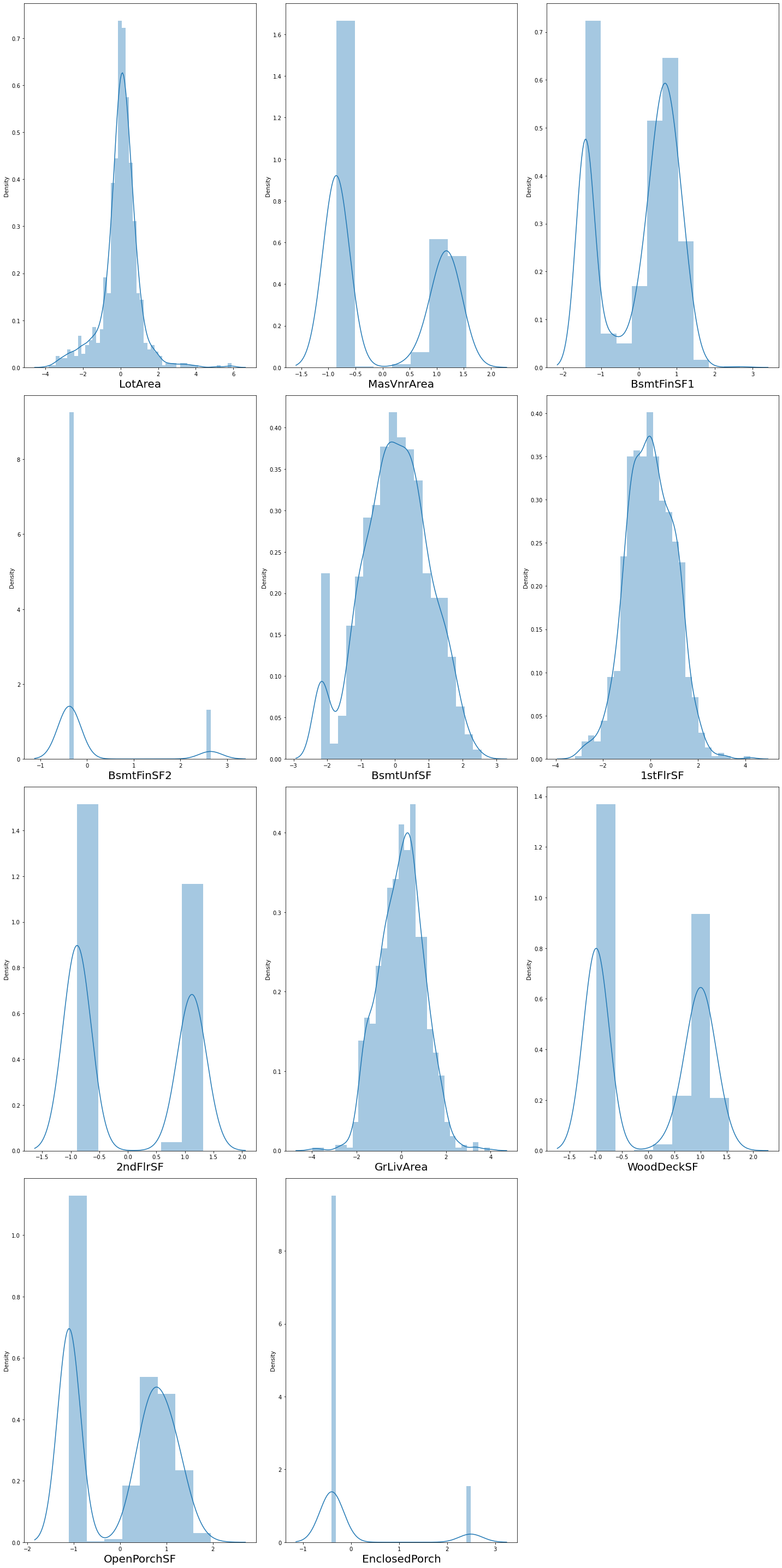
**Scatter Plot**

****

**Distribution Plot**

****

**Distribution Plot**

****

**CONCLUSION**

**Key Findings and Conclusions of the Study**

* OverallQuality of the house is the most important feature in predicting the SalesPrice
* RandomForestRegressor is the best Model for prediction

**Learning Outcomes of the Study in respect of Data Science**

Visualization:-

* From Visualization methods, we can easily check the skewness of the data, outliers and the relation between the features and label variable.
* From Visualization methods without even looking at the data entirely we can easily check many important features about the independent and dependent variables

Data Cleaning:-

* While Cleaning the data the entire knowledge about the dataset is very necessary as to understand what each column stands for, the datatype of each column
* In Data cleaning, it is very important to encode the data as per its datatype
* In Data Cleaning, handling null values and dropping columns is the most important aspect where the study, analysis of the dataset and domain knowledge comes into picture.
* Removing skewness and outliers from the dataset is the next most important aspect of handling the features, which can affect the model accuracy in many ways.

Challenges Faced:-

* The biggest challenge faced was while handling the test dataset.
* As the test dataset was also not cleaned and encoding the categorical features as per the same codes as done with training dataset was difficult.
* There were few distinct values in the columns absent in the training dataset present in the test dataset which needed to encoded similar to the training dataset.
* Hence, we could not encode the dataset directly as we did with the training dataset by using label encoder.
* The test dataset was first cleaned, handled missing values, encoded, scaled, feature engineering was done same as the training dataset and then the prediction of the output was done by using the same trained model with the training dataset.

**Limitations of this work and Scope for Future Work**

The solution provided is limited to the type and details of the dataset used in this project.

To further extend this study and improve the results, more such similar datasets can be used during the training and testing phase.