MACHINE LEARNING APPLIED: PRICE PREDICTION OF THE PROSPECTIVE REAL ESTATE PROPERTIES.

Submitted by:

S. Dilip Kumar

ACKNOWLEDGMENT

Real estate is property consisting of land and the buildings on it, along with its natural resources such as crops, minerals or water; immovable property of this nature; an interest vested in this (also) an item of real property, (more generally) buildings or housing in general. Real estate is different from personal property, which is not permanently attached to the land, such as vehicles, boats, jewellery, furniture, tools and the rolling stock of a farm.

Residential real estate.

Residential real estate may contain either a single family or multifamily structure that is available for occupation or for non-business purposes. Any property used for residential purposes. Examples include single-family homes, condos, cooperatives, duplexes, townhouses, and multifamily residences with fewer than five individual units.

Residences can be classified by and how they are connected to neighbouring residences and land. Different types of housing tenure can be used for the same physical type.

For example, connected residences might be owned by a single entity and leased out, or owned separately with an agreement covering the relationship between units and common areas and concerns

Major categories.

- Attached / multi-unit dwellings
- Apartment (American English) or Flat (British English)
- Multi-family house
- Terraced house (a. k. a. townhouse or rowhouse).
- Condominium (American English).
- Cooperative (a. k. a. co-op).
- Semi-detached dwellings
- Duplex.
- Detached dwellings
- Detached house or single-family detached house
- Portable dwellings
- Mobile homes or residential caravans.
- Houseboats.
- Tents.

The size of an apartment or house can be described in square feet or meters. In the United States, this includes the area of "living space", excluding the garage and other non-living spaces. The "square meters" figure of a house in Europe may report the total area of the walls enclosing the home, thus including any attached garage and non-living spaces, which makes it important to inquire what kind of surface area definition has been used. It can be described more roughly by the number of rooms. A studio apartment has a single bedroom with no living room (possibly a separate kitchen). A one-bedroom apartment has a living or dining room separate from the bedroom. Two bedroom, three bedroom, and larger units are

common. (A bedroom is a separate room intended for sleeping. It commonly contains a bed and, in newer dwelling units, a built-in closet for clothes storage.)

Other categories

- Chawls
- Villas
- Havelis

The size of these is measured in Gaz (square yards), Quila, Marla, Beegha, and acre.

(James Chen, 2019)

In the following blog we are going to work with the data given by Real estate company "Surprise Housing" a US based housing and Real Estate company, who have decided to enter into Australian Real Estate Market. The Company have collected the data from the sales of houses in Australia. Using the collected the data with Data analytics and with the help of Machine Learning company is trying to plot the required factors that affect the cost of the property. Further the company is looking at the prospective properties to buy at a cheaper rate and sell it a best price. These data are provided by the "Surprise housing" as in CSV format. Using these data. We have to predict and find.

- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

Tools Used. (Hardware and Software Requirements):

1. Python 3.8
2. NumPy
3. Pandas
4. Matplotlib
5. Seaborn
6. Data science
7. Machine learning
8. CPU with RAM of 8GB
9. Anaconda Environment
10. Jupyter Notebook

Insights from the Data.

From all the Data available, we can bring out some neat insights or conclusions such as

- ✓ How the location of the lot decides the selling price of the property.
- ✓ How the amenities like Garages, Swimming pool, parking lot, fence type, insulation type and quality increase the selling prices.
- ✓ How the number of rooms directly increases the cost and size of the property.
- ✓ What is the type of building and the year built mostly comes for the sales?
- ✓ The year the building build versus the cost of the property.
- ✓ The year the modified versus the cost of the property.
- ✓ Nearby neighbourhood versus cost of the property.

INTRODUCTION

Business Problem Framing.

Surprise Housing is a US based Real estate and housing company who is trying to entry into Australian Real estate market. 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.

For this Surprise Housing wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

Business Goal:

We are required 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.

Conceptual Background of the Domain Problem.

KEY TAKEAWAYS

- The most common way to make money in real estate is through appreciation—an increase in the property's value that is realized when you sell.
- Location, development, and improvements are the primary ways that residential and commercial real estate can appreciate in value.
- Inflation can also play a role in increasing a property's value over time.
- You can also make money in the form of income from rents for both residential and commercial properties, and companies may pay you royalties on raw land, for example, for any discoveries, such as minerals or oil.

Real Estate Profits from Increasing Property Value

The most common way real estate offers a profit: It appreciates—that is, it increases in value. This is achieved in different ways for different types of property, but it is only realized in one way: through selling. However, you can increase your return on investment on a property in several ways. One way—if you borrowed money to buy the property—is to refinance the loan at lower interest. This will lower your cost basis for the property, thus increasing the amount you clear from it.

How Money is Made Through Real Estate

(Emily Roberts {Copyright} Investopedia, 2019.)

The most obvious source of appreciation for undeveloped land is, of course, developing it. As cities expand, land outside the limits becomes increasingly valuable because of the potential for it to be purchased by developers. Once developers build houses or commercial buildings, it raises that value even further.

Appreciation in land can also come from discoveries of valuable minerals or other commodities—provided the buyer holds the rights to them. An extreme example of this would be striking oil, but appreciation can also come from gravel deposits, trees, and other natural resources.

When looking at residential properties, location is often the biggest factor in appreciation. As the neighbourhood around a home evolves, adding transit routes, schools, shopping centres, playgrounds, and more, these changes cause the home's value to climb. Of course, this trend can also work in reverse, with home values falling as a neighbourhood decay.

Home improvements can also spur appreciation. Putting in an extra bathroom, heating a garage, and remodelling a kitchen with state-of-the-art appliances are just some of the ways a property owner may try to increase the value of a home.

Commercial property gains value for the same reasons as raw land and residential real estate: location, development, and improvements. The best commercial properties are perpetually in demand.

The Role of Inflation in Property Values

When considering appreciation, you have to factor in the economic impact of inflation. An annual inflation rate of 10% means that your dollar can only buy about 90% of the same goods the following year, and that includes property. If a piece of land was worth $100,000 in 1970 and it sat dormant and undeveloped for decades, it would still be worth many times more today. Because of runaway inflation throughout the 1970s and a steady pace since, it would likely take more than $500,000 to purchase that land now, assuming $100,000 was fair market value at the time. Thus, inflation alone can lead to appreciation in real estate, but it is a bit of a Pyrrhic victory. While you may get five times your money due to inflation when you sell, many other goods cost five times as much to buy too, so purchasing power in your current environment is still a factor.

Real Estate Profits from Income

The second big way real estate generates wealth is by providing regular payments of income. Generally referred to as rent, income from real estate can come in many forms.

(James Chen, 2019)

Review of Literature

To know about the review of the literature we first need to thoroughly study about the data we have. The data given by the "Surprise Housing" have 81 columns with 1168 records. Let's see the definition about them in details.

About the Data Set.

Size of training set: 1168 records

Size of test set: 292 records

Columns/Features:

1) **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

2) **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

3) **LotFrontage**: Linear feet of street connected to property
4) **LotArea**: Lot size in square feet

5) **Street**: Type of road access to property

Grvl	Gravel
Pave	Paved

6) **Alley**: Type of alley access to property

Grvl	Gravel
Pave	Paved
NA	No alley access

7) **LotShape**: General shape of property

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular

8) **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

9) **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

10) **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

11) **LandSlope**: Slope of property

Gtl	Gentle slope
Mod	Moderate Slope
Sev	Severe Slope

12) **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

13) **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

14) **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

15) **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

16) **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

17) **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

18) **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

19) **YearBuilt**: Original construction date

20) **YearRemodAdd**: Remodel date (same as construction date if no remodelling or additions)

21) **RoofStyle**: Type of roof

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

22) **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

23) **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

24) **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

25) **MasVnrType**: Masonry veneer type

BrkCmn Brick Common
BrkFace Brick Face
CBlock Cinder Block
None None
Stone Stone

26) **MasVnrArea**: Masonry veneer area in square feet

27) **ExterQual**: Evaluates the quality of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

28) **ExterCond**: Evaluates the present condition of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

29) **Foundation**: Type of foundation

BrkTil Brick & Tile
CBlock Cinder Block
PConc Poured Contrete
Slab Slab
Stone Stone
Wood Wood

30) **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

31) **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

32) **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

33) **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 Unfinished
NA No Basement

34) **BsmtFinSF1**: Type 1 finished square feet
35) **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 Unfinished
NA No Basement

36) **BsmtFinSF2**: Type 2 finished square feet
37) **BsmtUnfSF**: Unfinished square feet of basement area
38) **TotalBsmtSF**: Total square feet of basement area
39) **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

40) **HeatingQC**: Heating quality and condition

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

41) **CentralAir**: Central air conditioning

N No
Y Yes

42) **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

43) **1stFlrSF**: First Floor square feet
44) **2ndFlrSF**: Second floor square feet
45) **LowQualFinSF**: Low quality finished square feet (all floors)
46) **GrLivArea**: Above grade (ground) living area square feet
47) **BsmtFullBath**: Basement full bathrooms
48) **BsmtHalfBath**: Basement half bathrooms
49) **FullBath**: Full bathrooms above grade
50) **HalfBath**: Half baths above grade
51) **Bedroom**: Bedrooms above grade (does NOT include basement bedrooms)
52) **Kitchen**: Kitchens above grade
53) **KitchenQual**: Kitchen quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor

54) **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
55) **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

56) **Fireplaces**: Number of fireplaces
57) **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

58) **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

59) **GarageYrBlt**: Year garage was built

60) **GarageFinish**: Interior finish of the garage

Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage

61) **GarageCars**: Size of garage in car capacity
62) **GarageArea**: Size of garage in square feet

63) **GarageQual**: Garage quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

64) **GarageCond**: Garage condition

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

65) **PavedDrive**: Paved driveway

Y Paved
P Partial Pavement
N Dirt/Gravel

66) **WoodDeckSF**: Wood deck area in square feet
67) **OpenPorchSF**: Open porch area in square feet
68) **EnclosedPorch**: Enclosed porch area in square feet
69) **3SsnPorch**: Three season porch area in square feet
70) **ScreenPorch**: Screen porch area in square feet
71) **PoolArea**: Pool area in square feet
72) **PoolQC**: Pool quality

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
NA No Pool

73) **Fence**: Fence quality

GdPrv Good Privacy
MnPrv Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA No Fence

74) **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

75) **MiscVal**: $Value of miscellaneous feature
76) **MoSold**: Month Sold (MM)
77) **YrSold**: Year Sold (YYYY)

78) **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

79) **SaleCondition**: Condition of sale

Normal	Normal Sale
Abnorml	Abnormal Sale - trade, foreclosure, short sale
AdjLand	Adjoining Land Purchase
Alloca	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)

80) **SalePrice:** (Target Variable)
The price of every property with the above-mentioned details.

Above is the detailed explanation of every Feature variable along with the Target variable further we will see the relationship between the Feature variable and the Target variable and also decide what are the Feature variable that is important in building most efficient model.

Motivation for the Problem Undertaken

We are required 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.

For this Surprise Housing wants to know:

- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?
- Depending upon the Customer requirements how to project the selling price of the property?
- What is the relationship between Feature Variable and Target Variable?
- How the property amenities decide the cost of the property?

Analytical Problem Framing

Data Sources and their formats.

As like what we saw above the data was given to us by “Surprise Housing” and these data was collected from the Databases of Sale of houses in Australia. These data were given to us in CSV format.

- Data contains 1460 entries each having 81 variables.
- Data contains Null values. We need to treat them using the domain knowledge and your own understanding.
- Extensive EDA has to be performed to gain relationships of important variable and price.
- Data contains numerical as well as categorical variable. We need to handle them accordingly.
- We have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.
- We need to find important features which affect the price positively or negatively.
- Two datasets i.e., Train and Test.

Exploratory Data Analysis (EDA) & Data Pre-processing pipeline.

We have started our analysis by importing the required libraries and I have also imported the data which was given in CSV format. On importing data with the initial play around analysis and also with the above explanation I have figured out that the data have Null values and no duplicate records.

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

importing the Data

In [2]:
```
df_train = pd.read_csv("C:/Users/Friday/Downloads/Project-Housing/Project-Housing_splitted/train.csv")
df_train
```

Out[2]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0

I have decided to drop Id column since it has unique values in all the records and also, I have split the data into five parts to make the analysis much easier.

In [6]:
```
def Drop_id(df): ###droping the id column.
    """Function that drops Id column"""
    df.drop('Id', axis=1, inplace = True)
    return df
```

In [7]:
```
df_train = Drop_id(df_train)
```

In [8]:
```
df_train1, df_train2, df_train3, df_train4, df_Target = df_train.iloc[:,:20], df_train.iloc[:,20:40],
                                                        df_train.iloc[:,40:64], df_train.iloc[:,64:79],
                                                        df_train['SalePrice']

###Splitting the Table.
```

Post splitting the Data to fill the NaN values I have used the technique of iterative imputer which as of now still used as experimental and also with the data explanations we already have using which I have filled NaN values as follows.

In [10]:
```
def Filling_NaN1(df): ####### Filling NaN values in the First table
    df['Alley'].replace(np.nan, 'No_access', inplace = True)
    from sklearn.experimental import enable_iterative_imputer
    from sklearn.impute import IterativeImputer
    imp = IterativeImputer(max_iter=10, random_state=0)
    imp.fit(df_train1[['LotArea','LotFrontage']])
    df[['LotArea','LotFrontage']] = np.round(imp.transform(df[['LotArea','LotFrontage']]))
    return df
```

In [11]:
```
df_train1 = Filling_NaN1(df_train1)
```

In [12]:
```
df_train1.isnull().sum()
```

Out[12]:
```
MSSubClass      0
MSZoning        0
LotFrontage     0
LotArea         0
Street          0
Alley           0
LotShape        0
LandContour     0
Utilities       0
LotConfig       0
LandSlope       0
Neighborhood    0
Condition1      0
Condition2      0
BldgType        0
HouseStyle      0
OverallQual     0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
dtype: int64
```

In [14]:
```
def Filling_NaN2(df): ####### Filling NaN values in the third table
    for i in ['BsmtQual', 'BsmtCond', 'BsmtExposure','BsmtFinType1','BsmtFinType2']:
        df[i].replace(np.nan, 'No_Basement', inplace = True)

    df['MasVnrType'].replace(np.nan,df['MasVnrType'].describe().top, inplace = True)
    from sklearn.experimental import enable_iterative_imputer
    from sklearn.impute import IterativeImputer
    imp = IterativeImputer(max_iter=10, random_state=0)
    df[['MasVnrArea', 'TotalBsmtSF']] = np.round(imp.fit_transform(df[['MasVnrArea', 'TotalBsmtSF']]))
    return df
```

In [15]:
```
df_train2 = Filling_NaN2(df_train2)
```

In [16]:
```
df_train2.isnull().sum()
```

Out[16]:
```
RoofStyle       0
RoofMatl        0
Exterior1st     0
Exterior2nd     0
MasVnrType      0
MasVnrArea      0
ExterQual       0
ExterCond       0
Foundation      0
BsmtQual        0
BsmtCond        0
BsmtExposure    0
BsmtFinType1    0
BsmtFinSF1      0
BsmtFinType2    0
BsmtFinSF2      0
BsmtUnfSF       0
TotalBsmtSF     0
Heating         0
HeatingQC       0
dtype: int64
```

In [19]:
```
def Filling_NaN3(df): ####### Filling NaN values in the second tabel.
    for i in ['GarageType', 'GarageYrBlt', 'GarageFinish','GarageQual','GarageCond']:
        df[i].replace(np.nan, 'No_Garage', inplace = True)
    df_train3['GarageYrBlt'] = df_train3['GarageYrBlt'].astype(str)
    df['FireplaceQu'].replace(np.nan, 'No_Fireplace', inplace = True)
    return df
```

In [20]:
```
df_train3 = Filling_NaN3(df_train3)
```

In [21]:
```
df_train3.isnull().sum()
```

Out[21]:
```
CentralAir      0
Electrical      0
1stFlrSF        0
2ndFlrSF        0
LowQualFinSF    0
GrLivArea       0
BsmtFullBath    0
BsmtHalfBath    0
FullBath        0
HalfBath        0
BedroomAbvGr    0
KitchenAbvGr    0
KitchenQual     0
TotRmsAbvGrd    0
Functional      0
Fireplaces      0
FireplaceQu     0
GarageType      0
GarageYrBlt     0
GarageFinish    0
GarageCars      0
GarageArea      0
GarageQual      0
GarageCond      0
dtype: int64
```

In [23]:
```
def Filling_NaN4(df): ####### Filling NaN values in the forth table
    df['PoolQC'].replace(np.nan, 'No_Pool', inplace = True)
    df['Fence'].replace(np.nan, 'No_Fence', inplace = True)
    df['MiscFeature'].replace(np.nan, 'None', inplace = True)
    return df
```

In [23]:
```
def Filling_NaN4(df): ####### Filling NaN values in the forth table
    df['PoolQC'].replace(np.nan, 'No_Pool', inplace = True)
    df['Fence'].replace(np.nan, 'No_Fence', inplace = True)
    df['MiscFeature'].replace(np.nan, 'None', inplace = True)
    return df
```

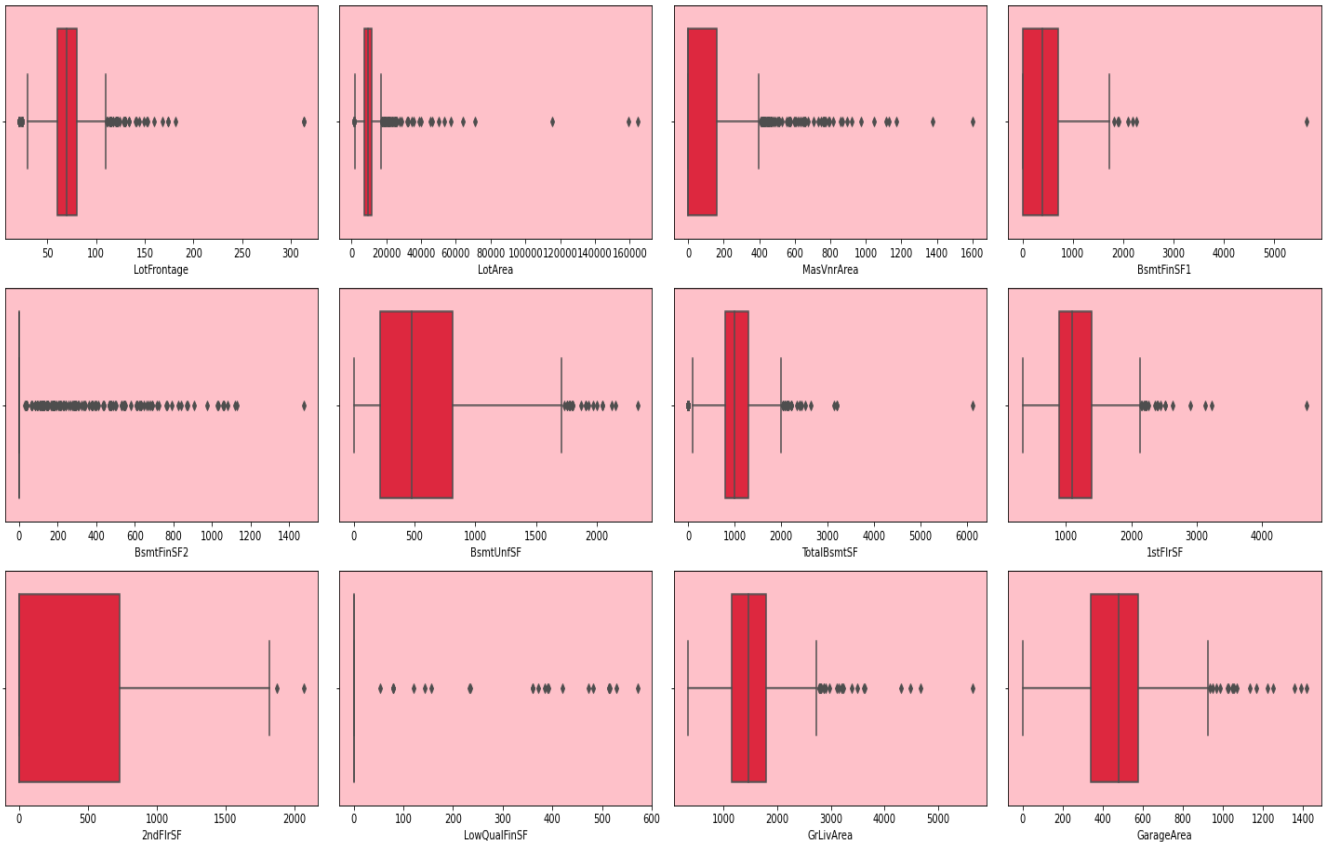
In [24]:
```
df_train4 = Filling_NaN4(df_train4)
```

In [25]:
```
df_train4.isnull().sum()
```

Out[25]:
```
PavedDrive       0
WoodDeckSF       0
OpenPorchSF      0
EnclosedPorch    0
3SsnPorch        0
ScreenPorch      0
PoolArea         0
PoolQC           0
Fence            0
MiscFeature      0
MiscVal          0
MoSold           0
YrSold           0
SaleType         0
SaleCondition    0
dtype: int64
```

I have made note of all the steps that is been followed in the pre-processing of this training data which later will create a function that will make the data ready to run in the model.

Post the imputation of filling the NaN values I also have separated the continuous data for checking the outliers and skewness present in the data. Also, with the visualization of skewness and outliers in the data with matplot.lib



There were totally 12 continuous variables in the data set, in the visualization with box plot there were to many outliers I have tried to remove them with Z-Score method. But the loss of data was 100%.

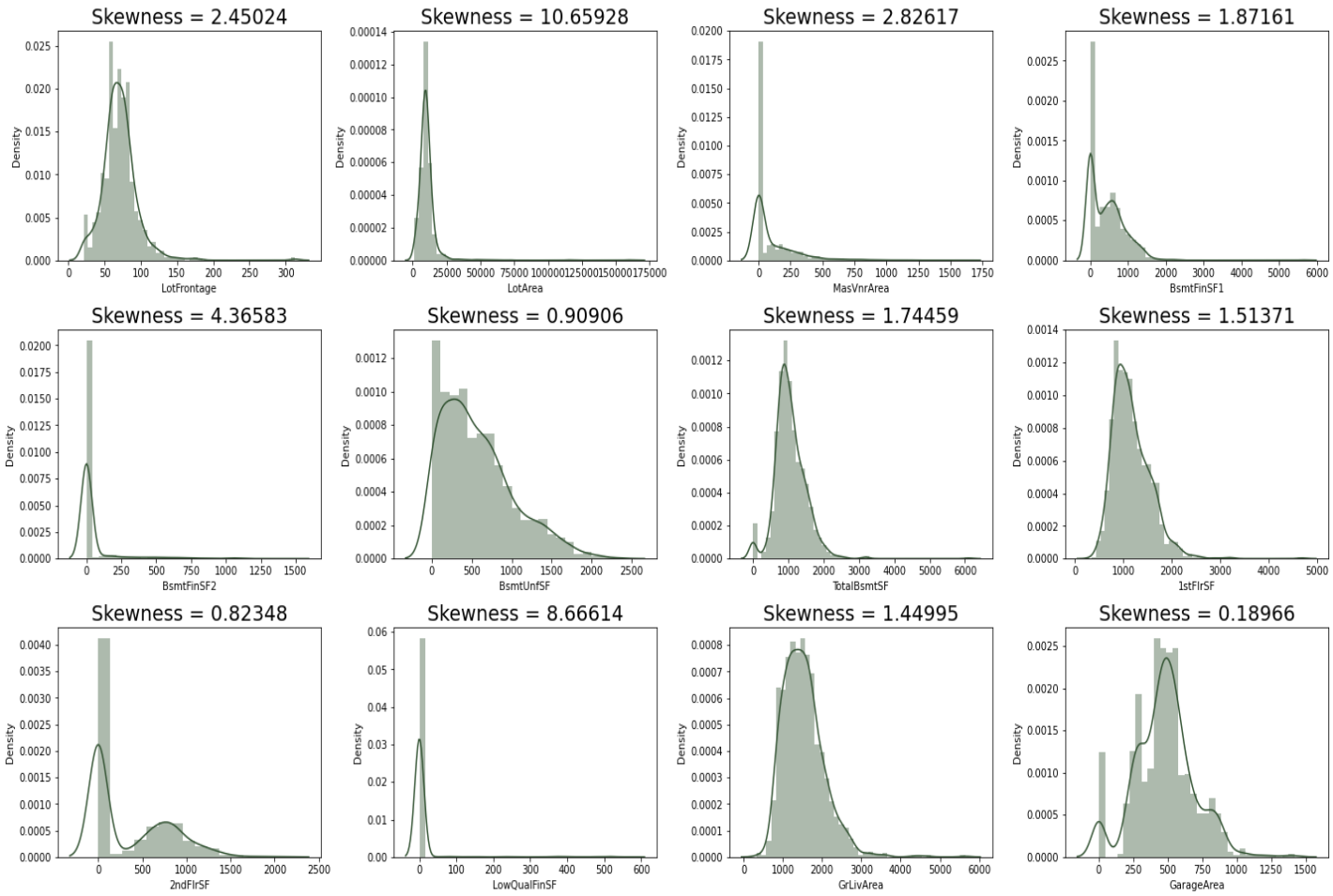
In [63]:
```
from scipy.stats import zscore

z= np.abs(zscore(DF_cv))
threshold= 3
df_new = DF_cv[(z < 3).all(axis=1)]
```

In [64]:
```
print(f"Orginal Data {DF_cv.shape}\nAfter Removing outliers {df_new.shape}\nThe percentage of data loss {((1168-0)/1168)*100}%")
```

```
Orginal Data (1168, 12)
After Removing outliers (0, 12)
The percentage of data loss 100.0%
```

As we observe that removing Outliers we will completely loss the data so we have to work with outliers present in the data

Before deciding on correcting the outliers I have also analysed and checked for skewness in the data. Like following.



The distribution plot shows the how the data is distributed and also the skewness present in the data. So, to over come the skewness in the data and also the outliers I have selected sklearn preprocessing powertransformation method. I have also made note of this step to

create the preprocessing pipeline function, but before the creation of the function I also wanted to know what are the variables that are important in creating this model and what are the variables that are not important so that I can decide to drop those variables and also to avoid multicollinearity problem from the data.

Mathematical/ Analytical Modelling of the Problem.

I have already split the data into 5 parts using that we will analysis the data for easier purpose. I have done Univariate and Multivariate analysis as follows.

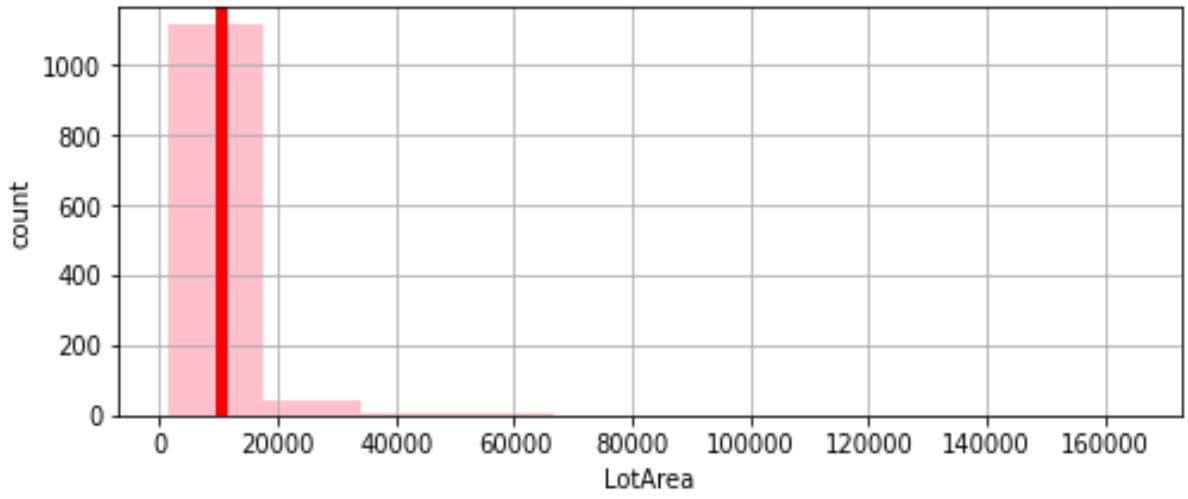
UNIVARIATE ANALYSIS:

TABLE 1:

MATHEMATICAL SUMMARY OF LotFrontage:

count	1168.000000
mean	72.678938
std	29.781705
min	21.000000
25%	60.000000
50%	70.000000
75%	80.000000
max	472.000000

Name: LotFrontage, dtype: float64



MATHEMATICAL SUMMARY OF LotArea:

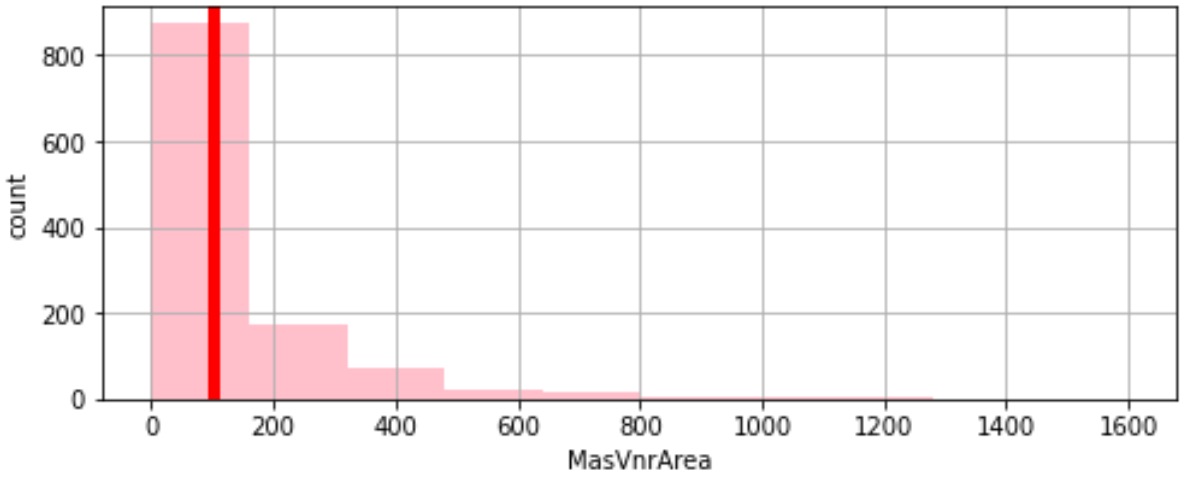
count	1168.000000
mean	10484.749144
std	8957.442311
min	1300.000000
25%	7621.500000
50%	9522.500000
75%	11515.500000
max	164660.000000

Name: LotArea, dtype: float64

KEY OBSERVATION:

1. Gable Roof type standard composite is high in number and exterior covering to house is Vinyl Siding and cement board.
2. Masonry veneer type is mostly none brick face, exterior quality and condition is mostly average/typical.
3. Cinder Block type foundation and basement quality is mostly good and condition is average/typical and basement mostly have no exposure.
4. basement finished type is mostly unfinished.
5. Gas forced warm air furnace heating type and Excellent in quality is high.

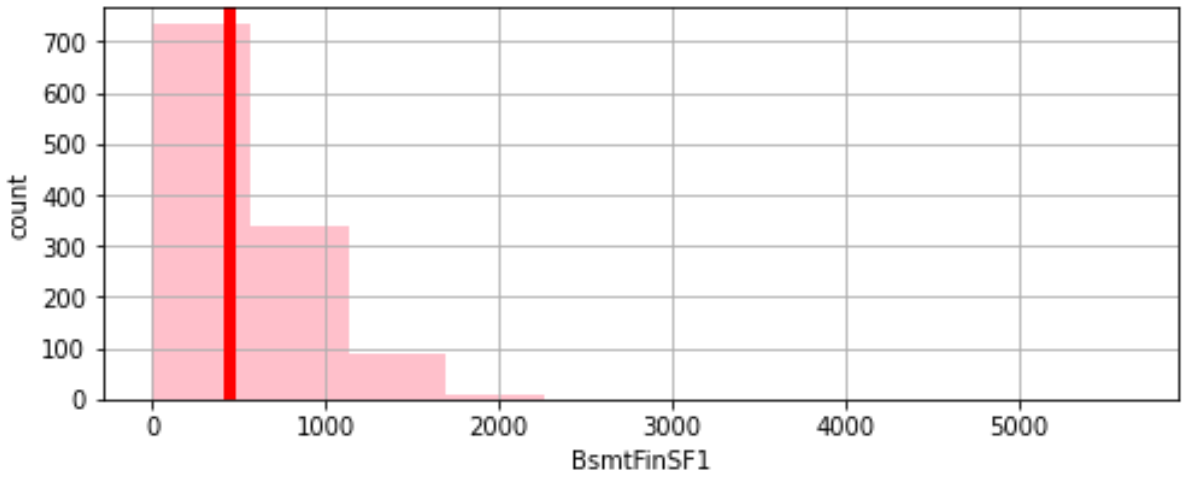
TABLE 2:



MATHEMATICAL SUMMARY OF MasVnrArea:

count	1168.000000
mean	102.590753
std	182.168633
min	0.000000
25%	0.000000
50%	0.000000
75%	160.000000
max	1600.000000

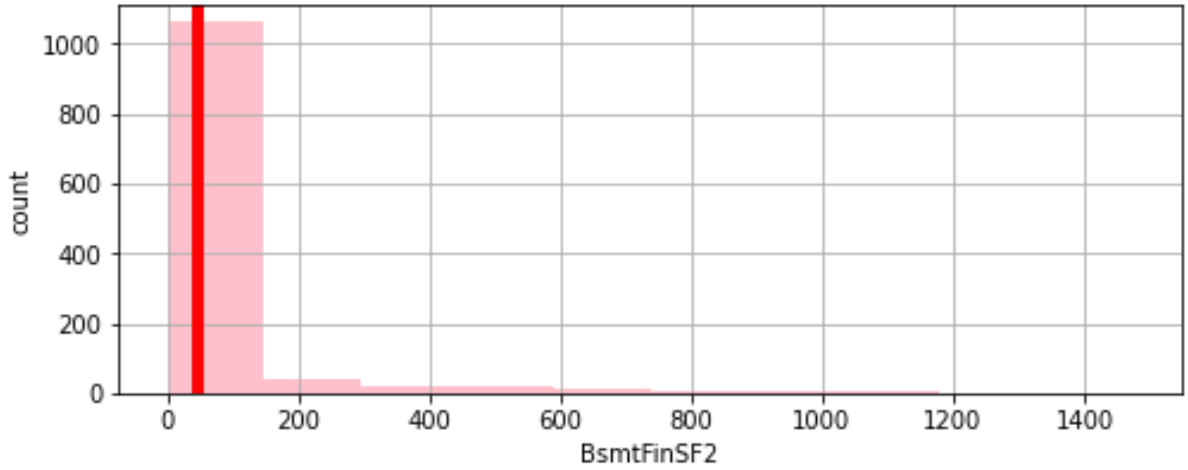
Name: MasVnrArea, dtype: float64



MATHEMATICAL SUMMARY OF BsmtFinSF1:

count	1168.000000
mean	444.726027
std	462.664785
min	0.000000
25%	0.000000
50%	385.500000
75%	714.500000
max	5644.000000

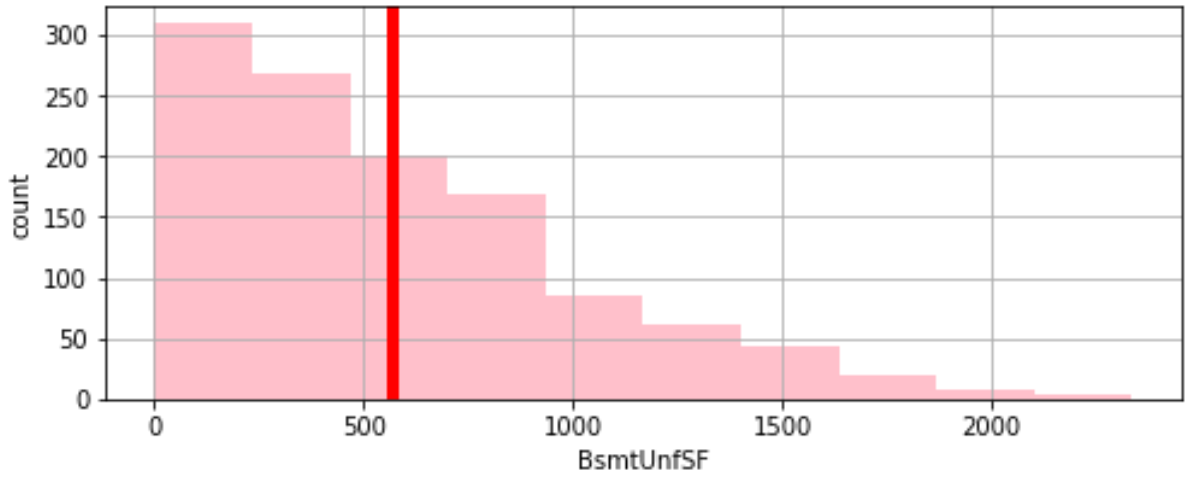
Name: BsmtFinSF1, dtype: float64



MATHEMATICAL SUMMARY OF BsmtFinSF2:

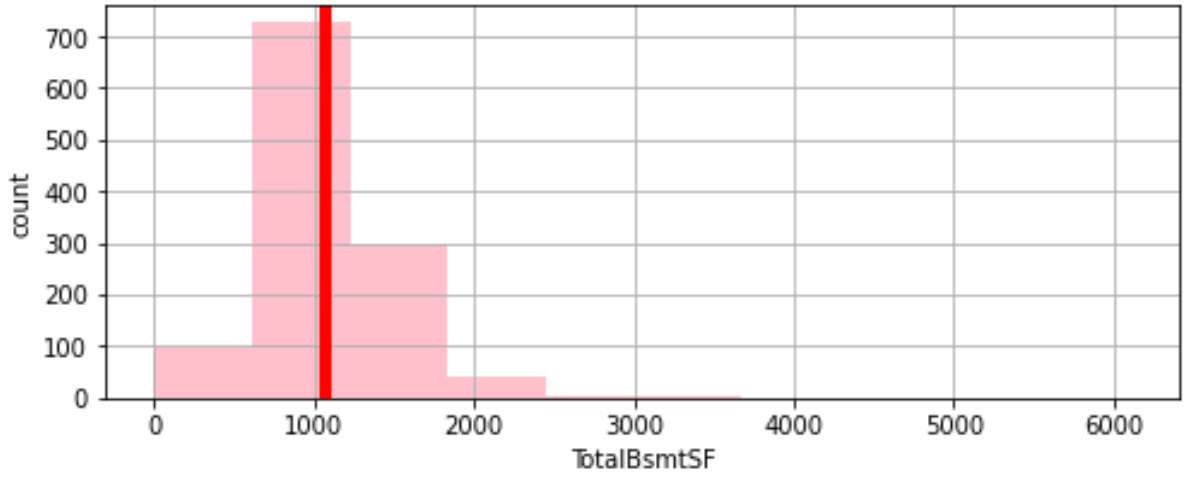
count	1168.000000
mean	46.647260
std	163.520016
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1474.000000

Name: BsmtFinSF2, dtype: float64



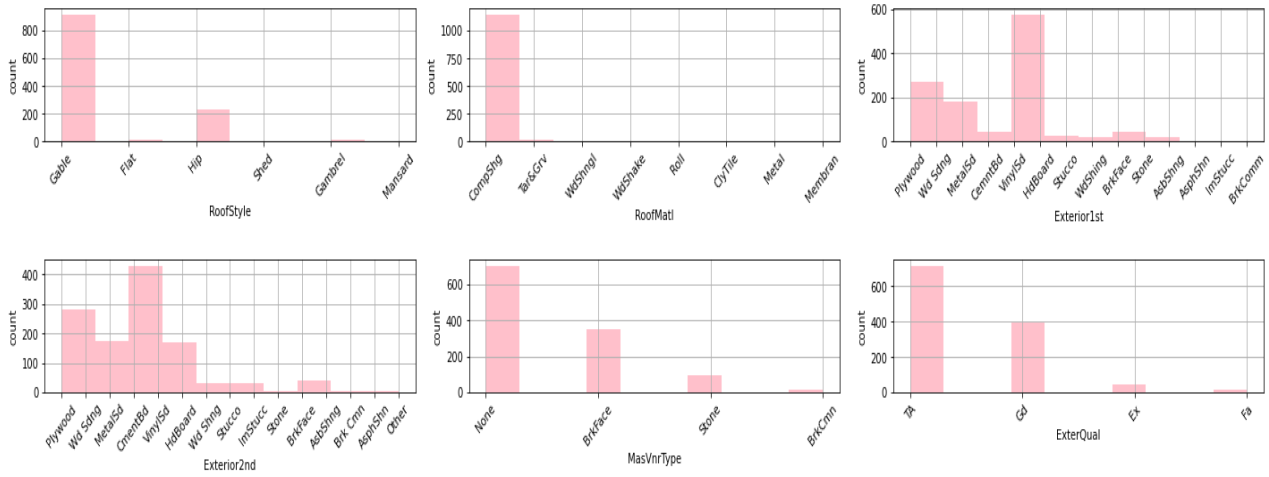
MATHEMATICAL SUMMARY OF BsmtUnfSF:

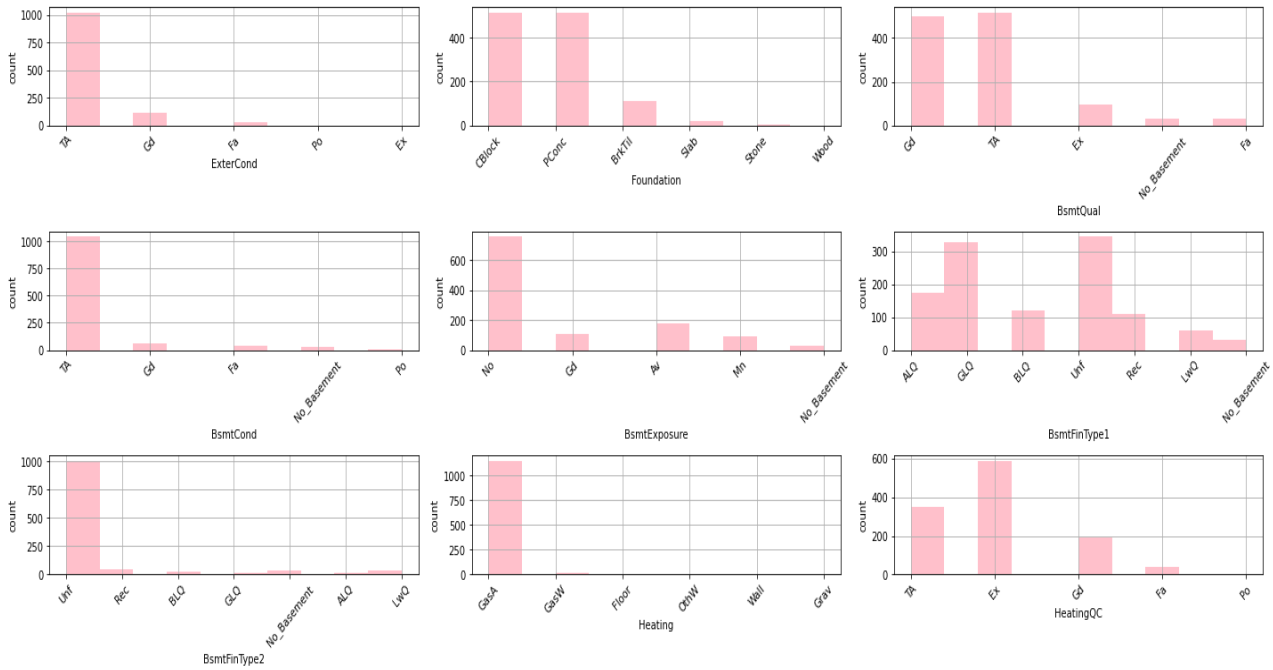
count 1168.000000
mean 569.721747
std 449.375525
min 0.000000
25% 216.000000
50% 474.000000
75% 816.000000
max 2336.000000
Name: BsmtUnfSF, dtype: float64



MATHEMATICAL SUMMARY OF TotalBsmtSF:

count 1168.000000
mean 1061.095034
std 442.272249
min 0.000000
25% 799.000000
50% 1005.500000
75% 1291.500000
max 6110.000000
Name: TotalBsmtSF, dtype: float64

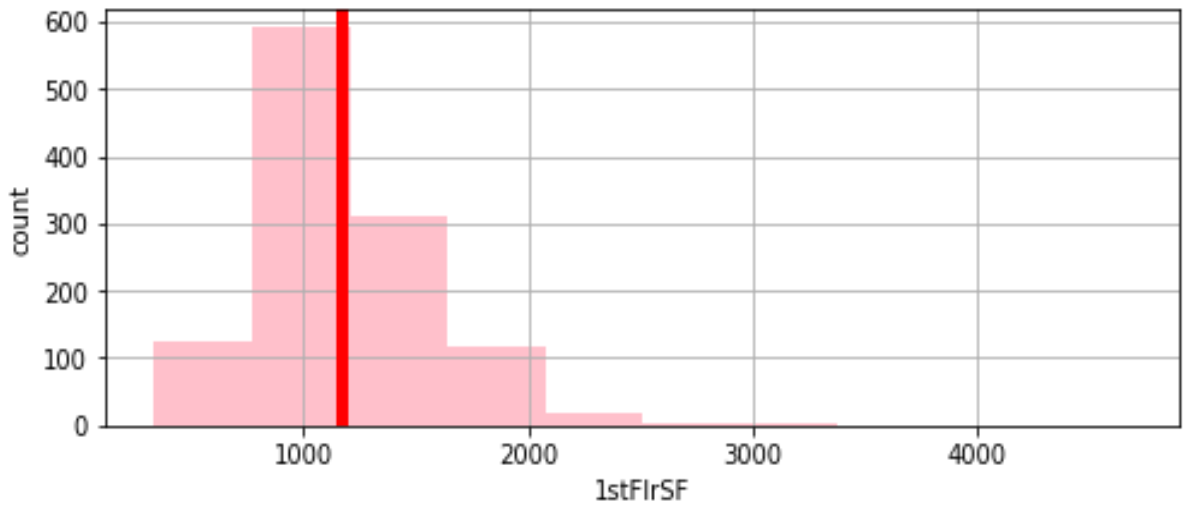




KEY OBSERVATIONS:

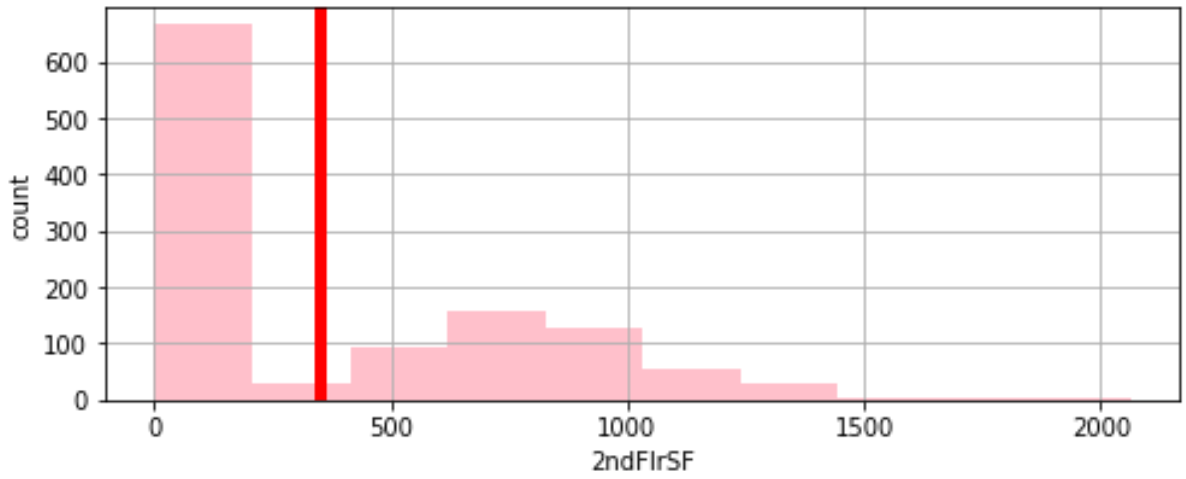
1. Gable Roof type standard composite is high in number and exterior covering to house is Vinyl Siding and cement board.
2. Masonry veneer type is mostly none brick face, exterior quality and condition is mostly average/typical.
3. Cinder Block type foundation and basement quality is mostly good and condition is average/typical and basement mostly have no exposure.
4. basement finished type is mostly unfinished.
5. Gas forced warm air furnace heating type and Excellent in quality is high.

TABLE 3:



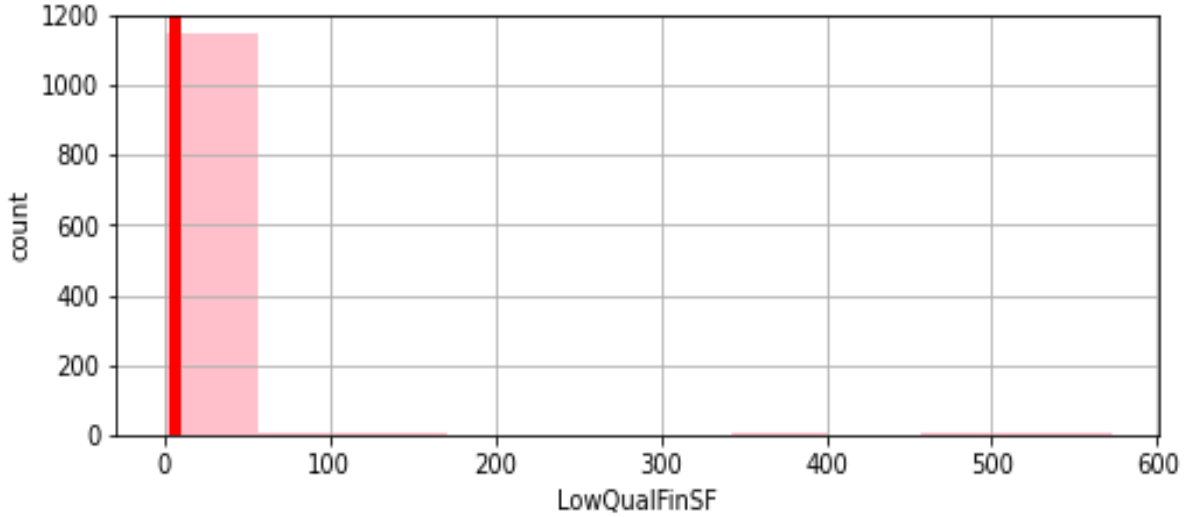
MATHEMATICAL SUMMARY OF 1stFlrSF:

count 1168.000000
mean 1169.860445
std 391.161983
min 334.000000
25% 892.000000
50% 1096.500000
75% 1392.000000
max 4692.000000
Name: 1stFlrSF, dtype: float64



MATHEMATICAL SUMMARY OF 2ndFlrSF:

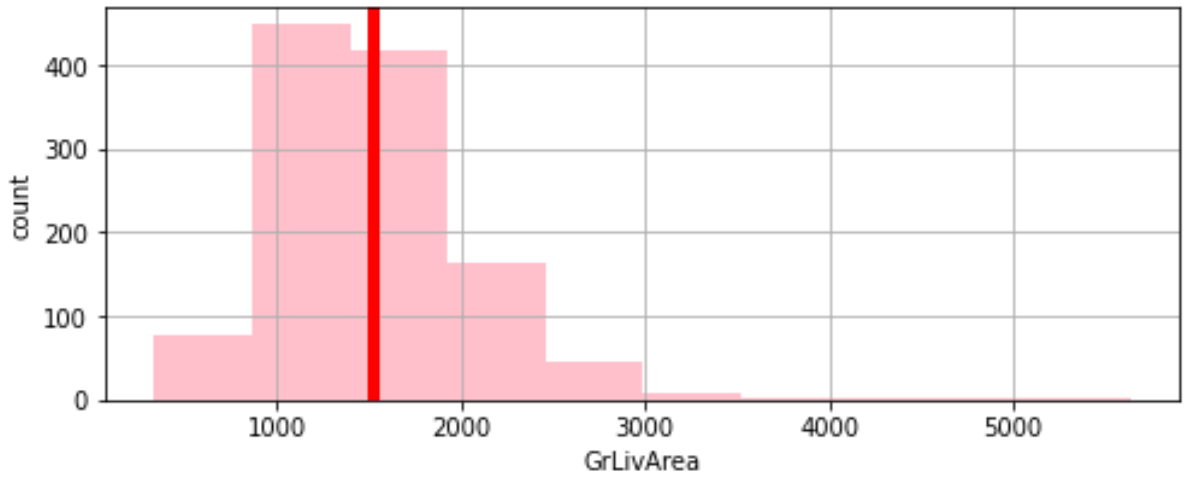
count 1168.000000
mean 348.826199
std 439.696370
min 0.000000
25% 0.000000
50% 0.000000
75% 729.000000
max 2065.000000
Name: 2ndFlrSF, dtype: float64



MATHEMATICAL SUMMARY OF LowQualFinSF:

count 1168.000000
mean 6.380137
std 50.892844
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 572.000000

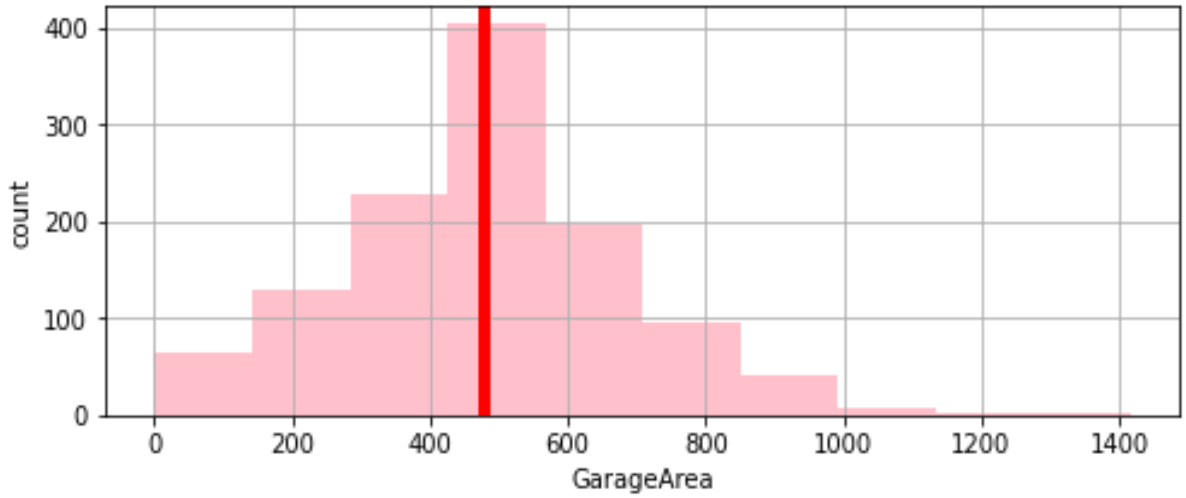
Name: LowQualFinSF, dtype: float64



MATHEMATICAL SUMMARY OF GrLivArea:

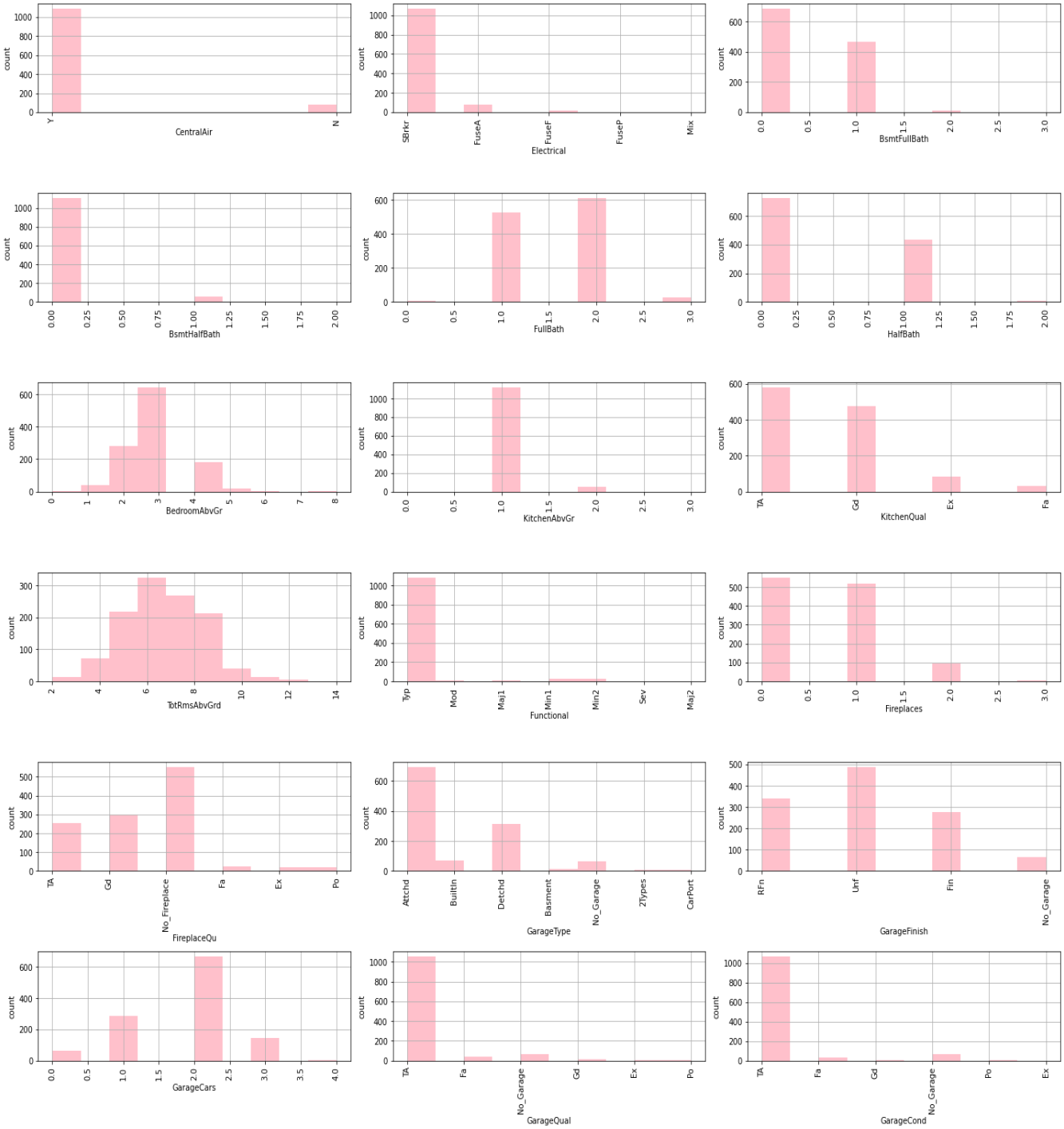
count 1168.000000
mean 1525.066781
std 528.042957
min 334.000000
25% 1143.250000
50% 1468.500000
75% 1795.000000
max 5642.000000

Name: GrLivArea, dtype: float64



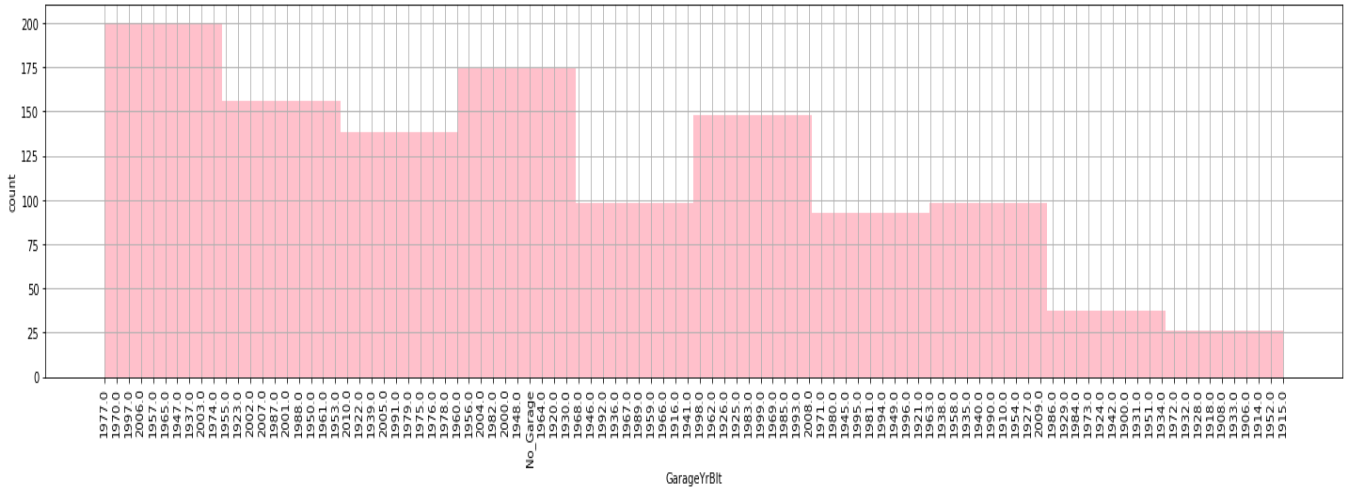
MATHEMATICAL SUMMARY OF GarageArea:

count 1168.000000
mean 476.860445
std 214.466769
min 0.000000
25% 338.000000
50% 480.000000
75% 576.000000
max 1418.000000
Name: GarageArea, dtype: float64



KEY OBSERVATIONS:

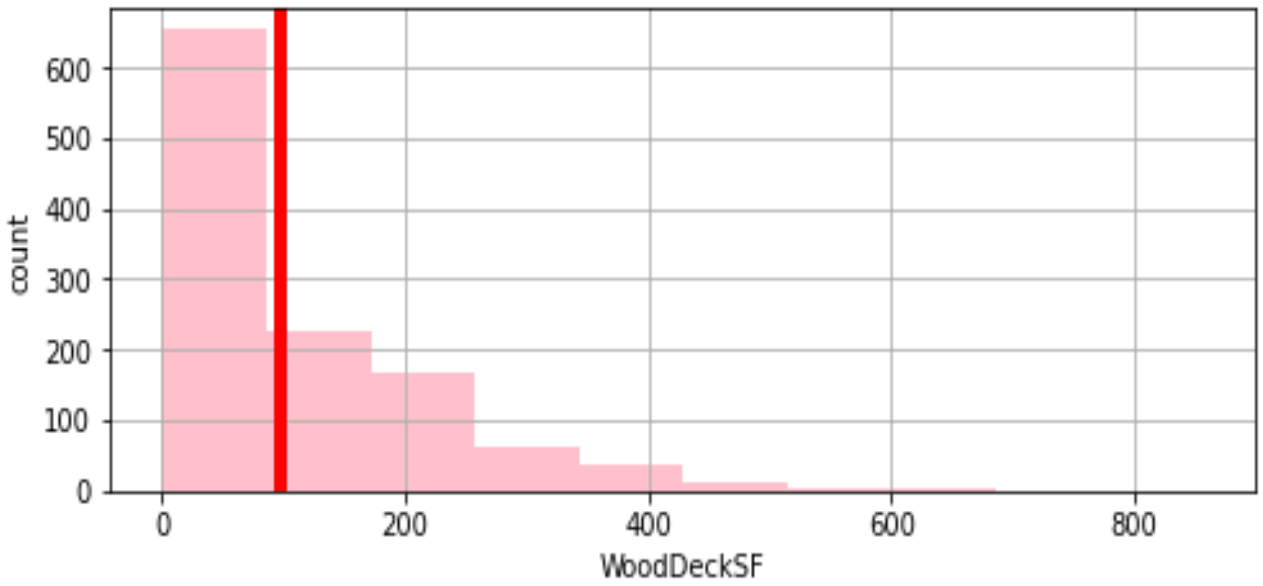
1. Property mostly has central air conditioning with mostly electrical system of Standard Circuit Breakers & Romex
2. Kitchen Quality is mostly Typical/Average. home functionality is mostly Typical Functionality and fire place is between 0-1.
3. Garage type is attached and mostly unfinished 2 cars parking quality and condition is Typical/Average.



KEY OBSERVATION:

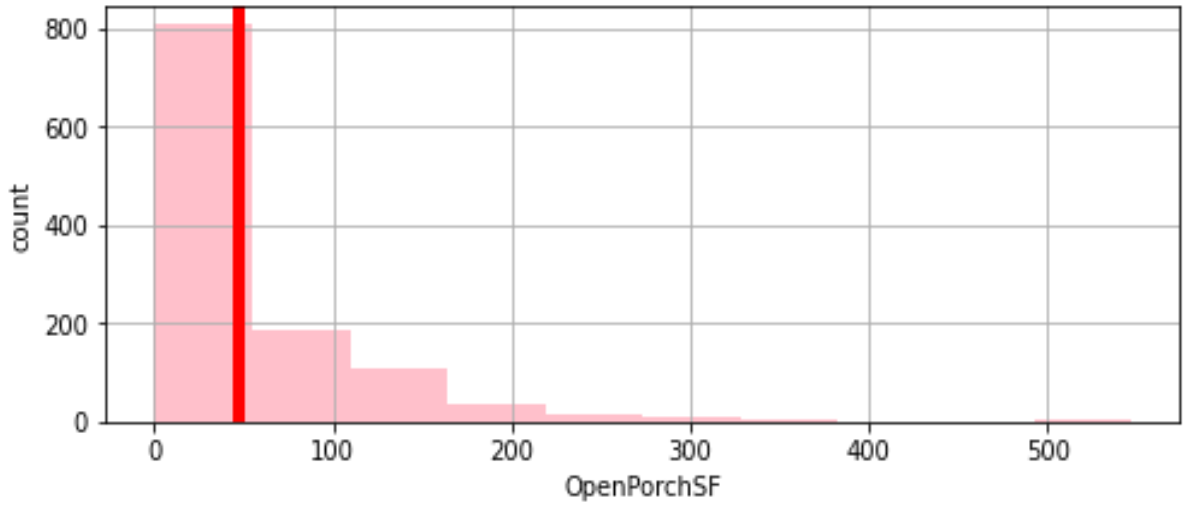
1. Properties out for sale are mostly built-in year 1977, 1970. 1997, 2006, 1957, 1965, 1947, 1937, 2003, 1974... which means from 1945 to 2006 in all frequency we have properties largely out for sale.

TABLE 4:



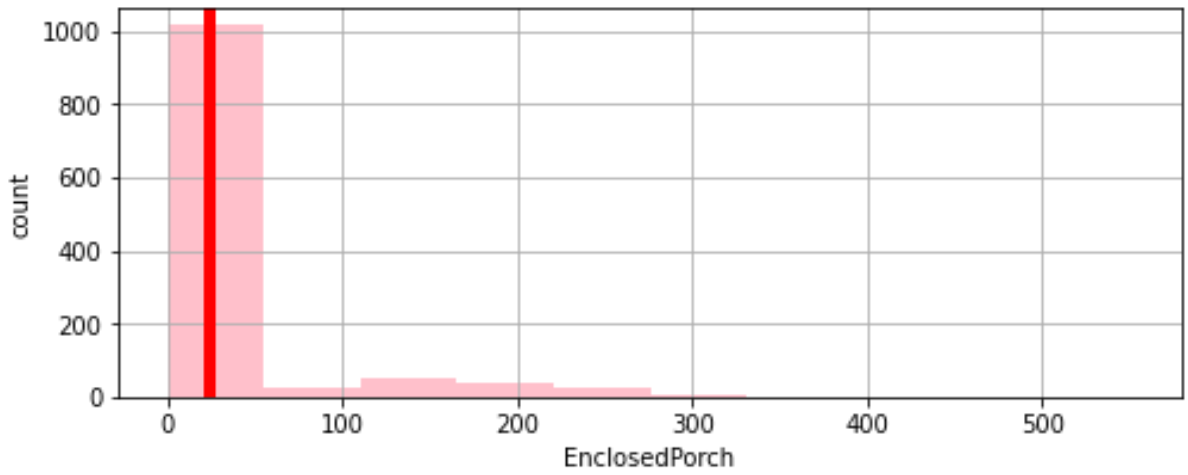
MATHEMATICAL SUMMARY OF WoodDeckSF:

count 1168.000000
mean 96.206336
std 126.158988
min 0.000000
25% 0.000000
50% 0.000000
75% 171.000000
max 857.000000
Name: WoodDeckSF, dtype: float64



MATHEMATICAL SUMMARY OF OpenPorchSF:

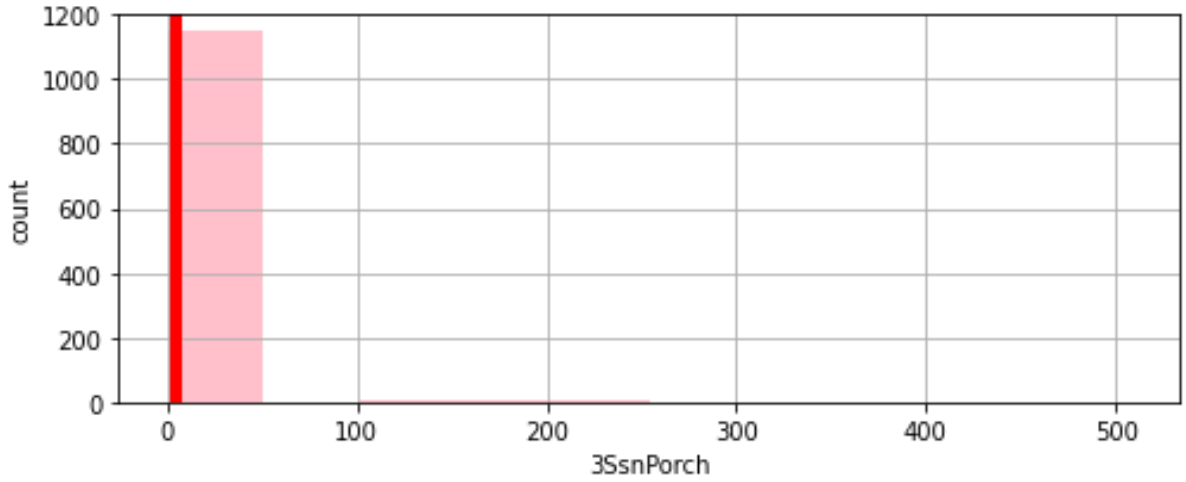
count 1168.000000
mean 46.559932
std 66.381023
min 0.000000
25% 0.000000
50% 24.000000
75% 70.000000
max 547.000000
Name: OpenPorchSF, dtype: float64



MATHEMATICAL SUMMARY OF EnclosedPorch:

count 1168.000000
mean 23.015411
std 63.191089
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 552.000000

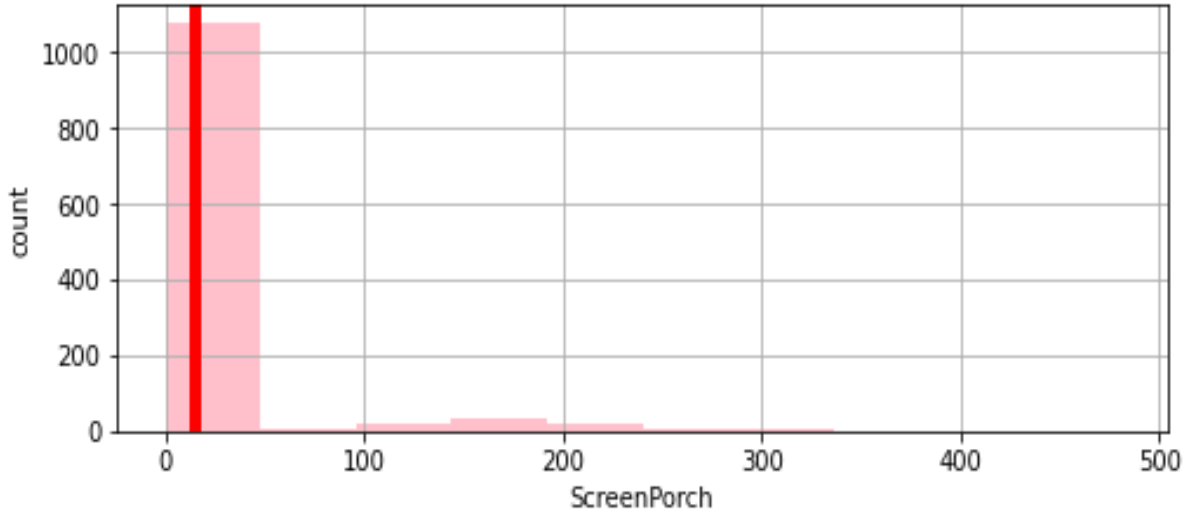
Name: EnclosedPorch, dtype: float64



MATHEMATICAL SUMMARY OF 3SsnPorch:

count 1168.000000
mean 3.639555
std 29.088867
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 508.000000

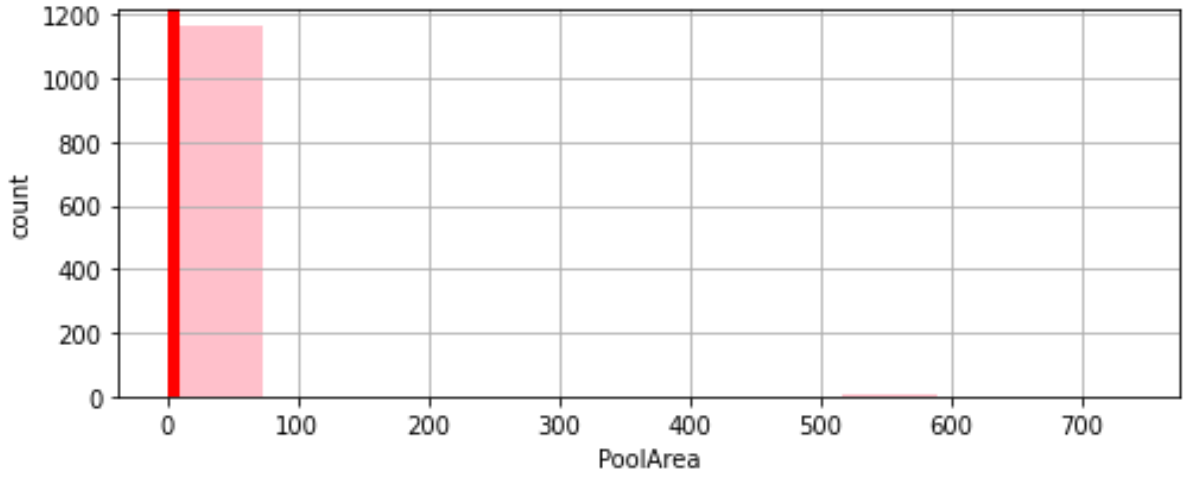
Name: 3SsnPorch, dtype: float64



MATHEMATICAL SUMMARY OF ScreenPorch:

count 1168.000000
mean 15.051370
std 55.080816
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 480.000000

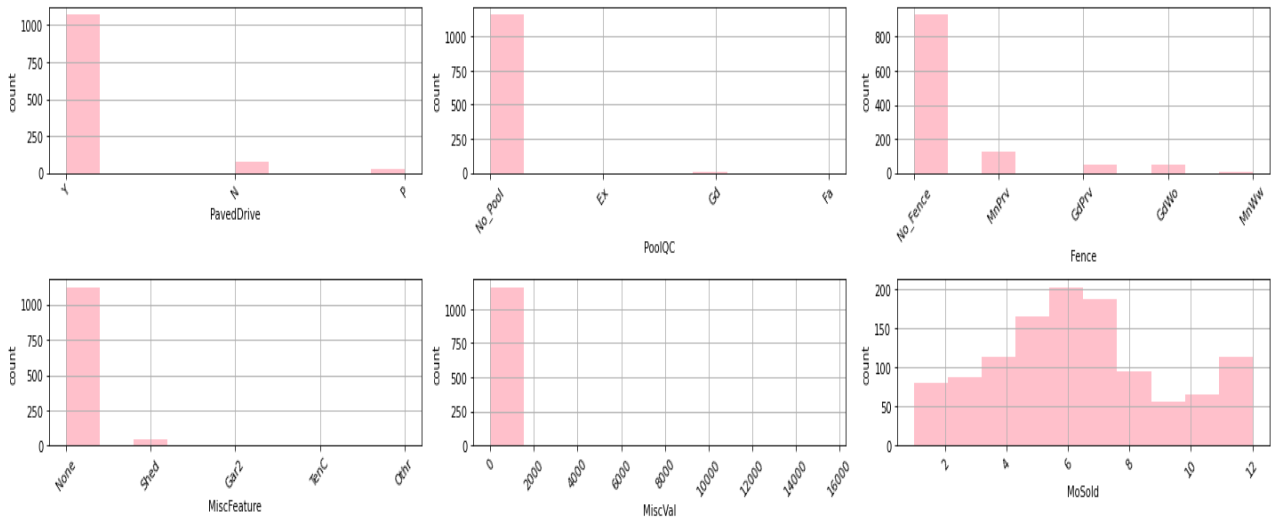
Name: ScreenPorch, dtype: float64

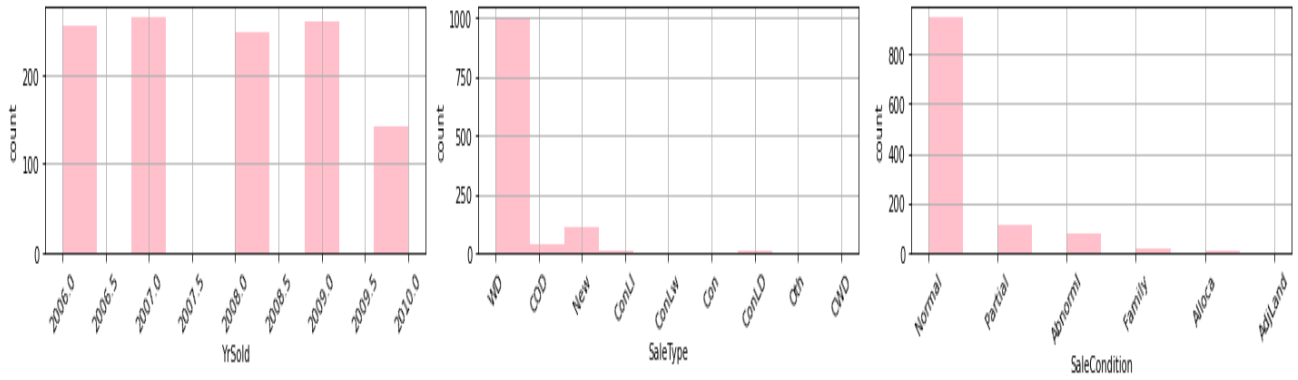


MATHEMATICAL SUMMARY OF PoolArea:

count 1168.000000
mean 3.448630
std 44.896939
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 738.000000

Name: PoolArea, dtype: float64



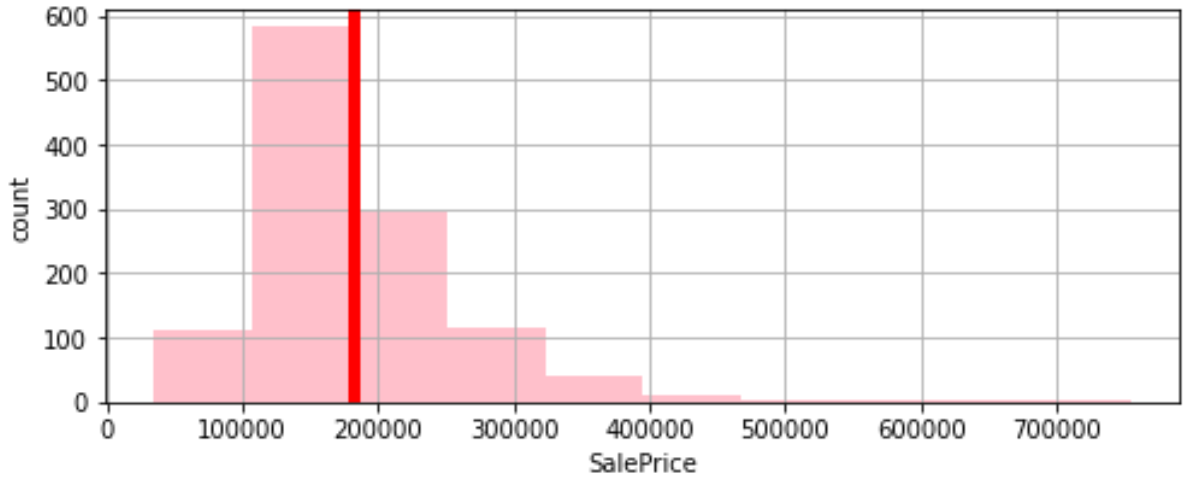


KEY OBSERVATIONS:

1. Paved driveway is mostly Paved
2. Sales type is Warranty Deed - Conventional condition is normal year sold is mostly 2007 with no fence and no pool.

TABLE 5:

Table 5 is with only one variable that is the target variable.

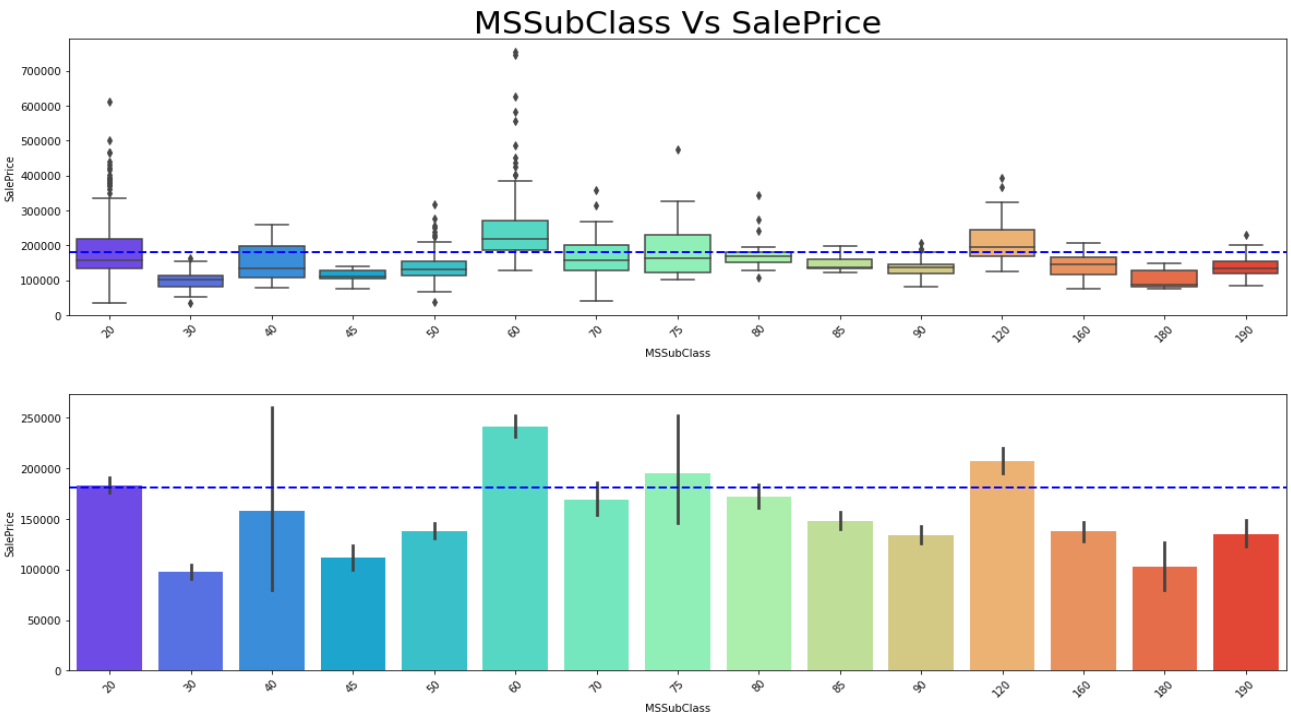


MATHEMATICAL SUMMARY OF VOTES:

count 1168.000000
mean 181477.005993
std 79105.586863
min 34900.000000
25% 130375.000000
50% 163995.000000
75% 215000.000000
max 755000.000000
Name: SalePrice, dtype: float64

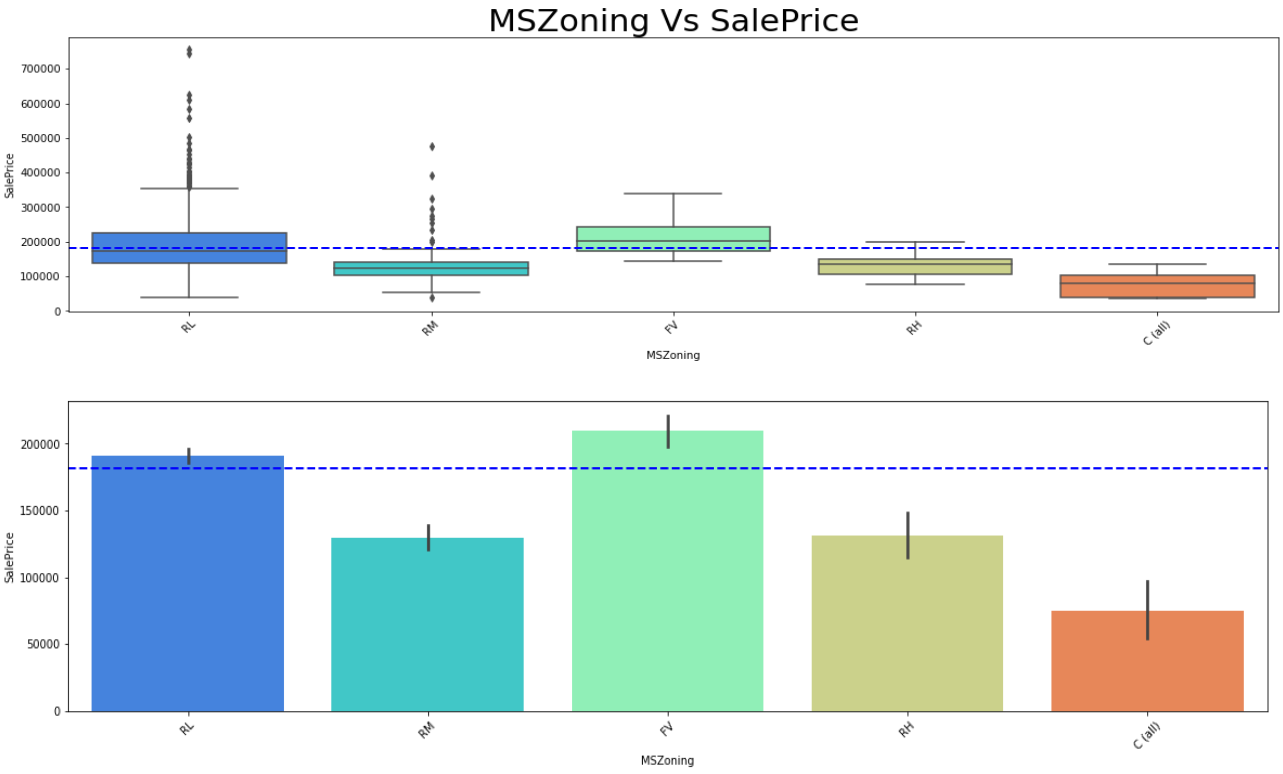
We have seen about each and every variable and its features separately let's do the multivariate analysis of those variable along with the correlation of every variable. From which we can analysis the relation ship of every variable with the target variable.

MULTI VARIATE ANALISIS: TABLE 1.



KEY OBSERVATION:

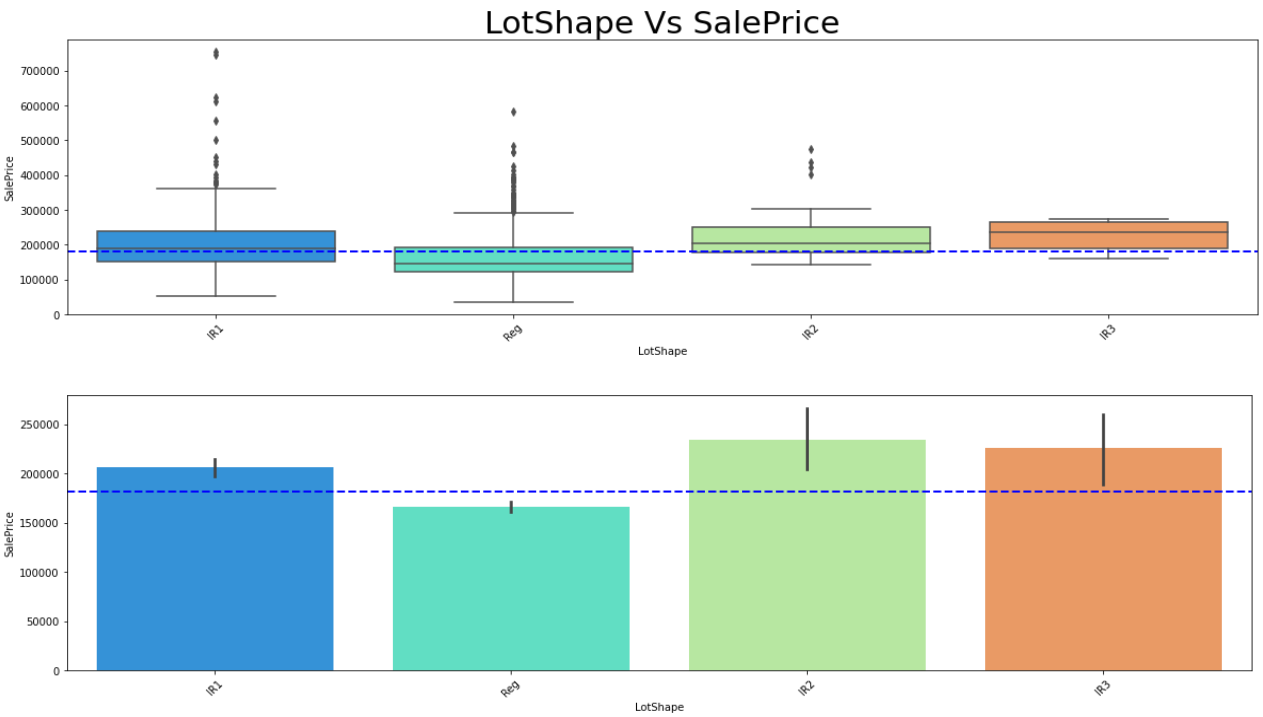
1. We can see that MSSubClass **60 – [2-STORY 1946 & NEWER]** and **70- [2-STORY 1945 & OLDER]** is the highest segment of building that is sold in the marker which means buyers mostly wish to buy these dwelling in the market



KEY OBSERVATION:

1. FV is **Floating Village Residential** which is being highly sold and RL- **Residential Low Density** being the costliest in the market.

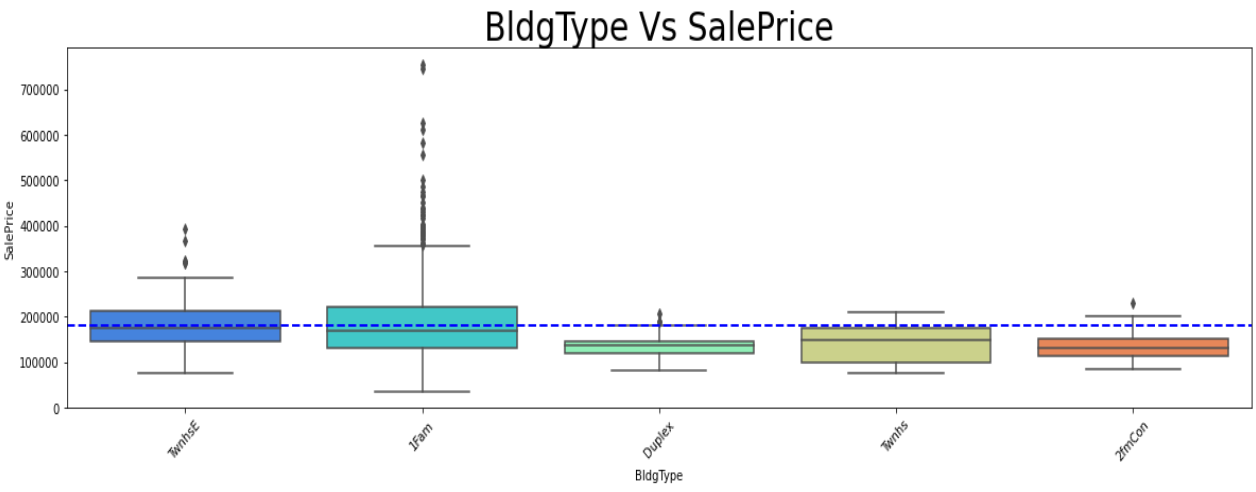
We can understand that low residential density which might be of more posh residential area which are costlier in the market across all the other classifications of residents

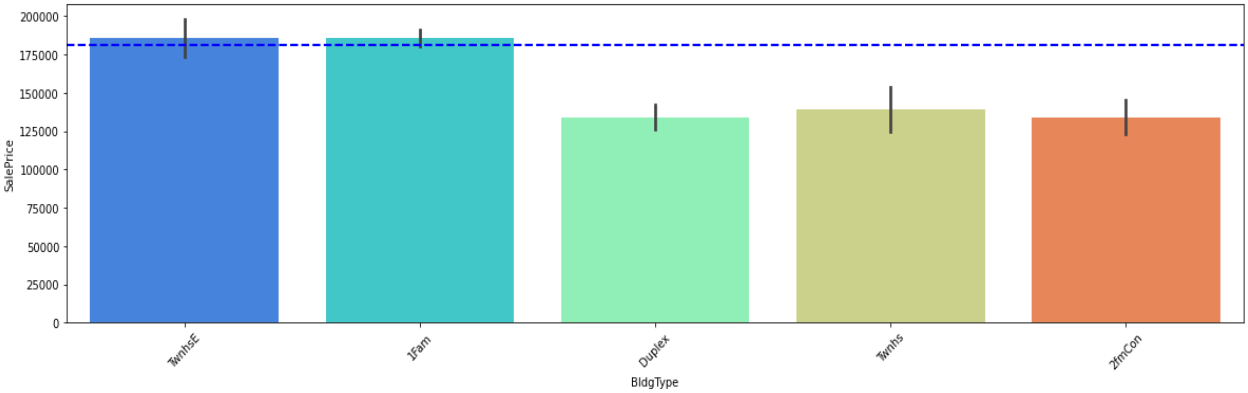


KEY OBSERVATION:

1. IR1 - **Slightly irregular** being the costliest lot shape followed by Reg – **Regular.**
2. IR2 - **Moderately Irregular** are the highest sold lot shape.

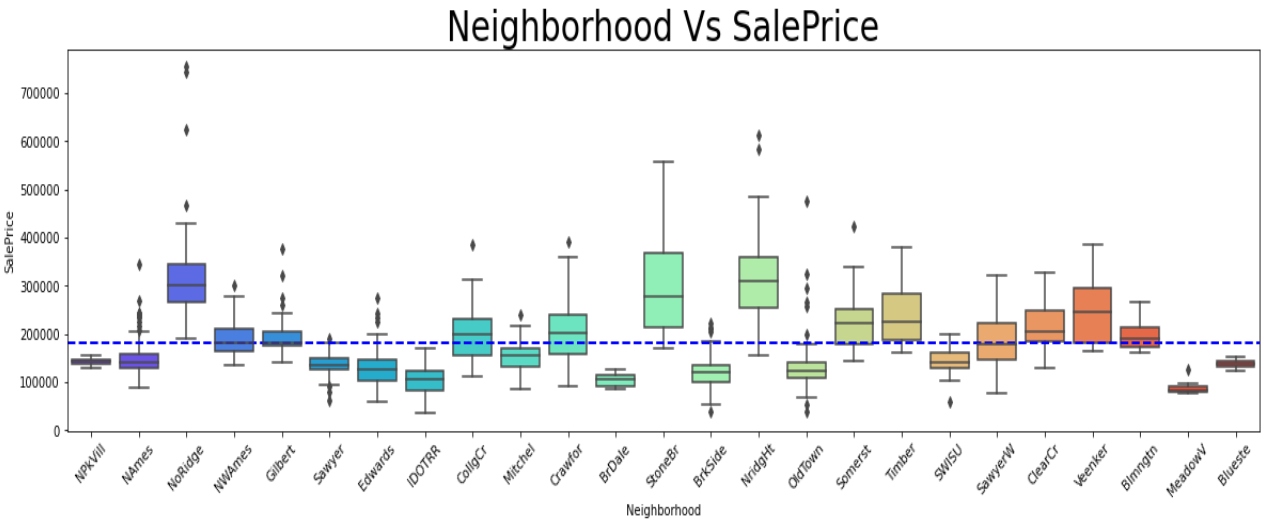
So, from above we can understand people are mostly interested in buying irregular shaped lot more than the regular shaped lot and since people are not buying the regular shaped lot the cost of the regular shaped lot is lesser than slightly irregular plot. And also the availability of the Regular shaped plot is lesser that might be also one of the reason to note.

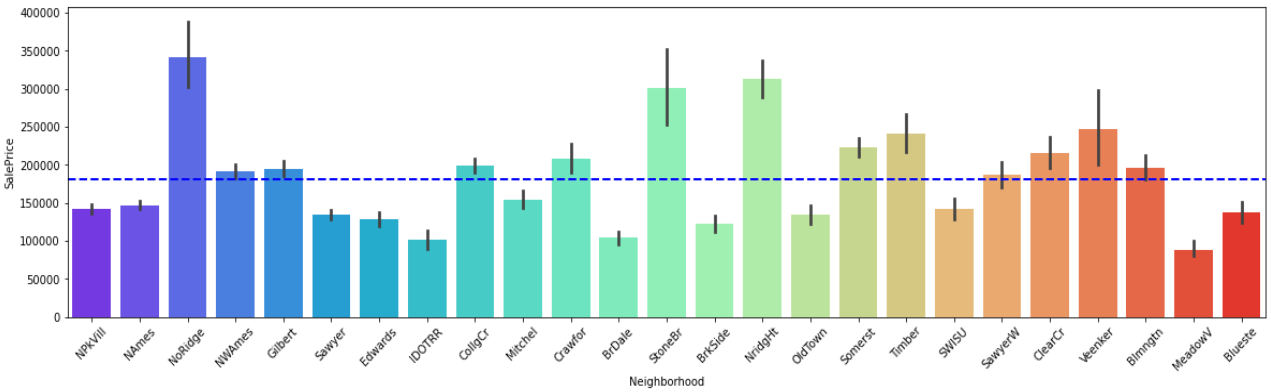




KEY OBSERVATIONS:

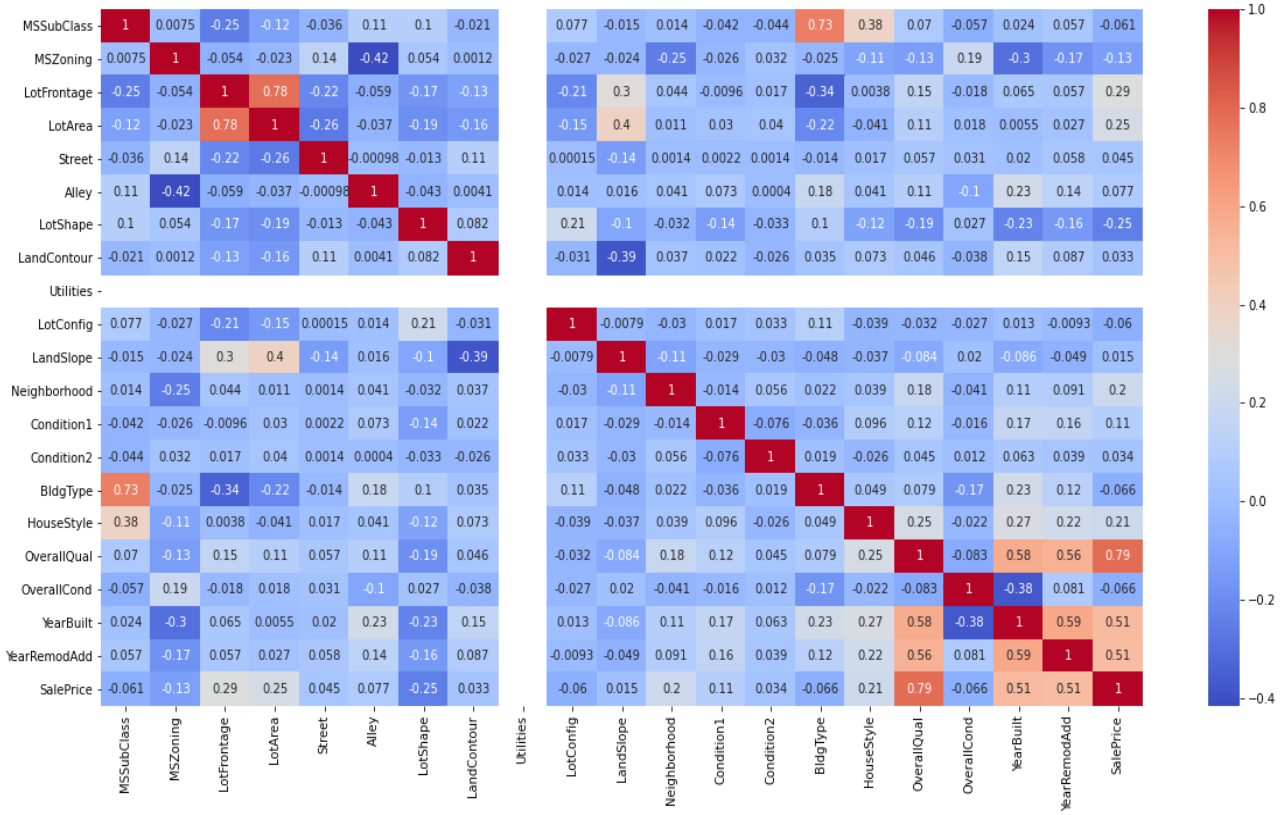
1. We can see one family type building being the costliest in the market and also getting sold higher.
2. TwnhsE - Townhouse End Unit is second highest sold Building Type.





KEY OBSERVATION:

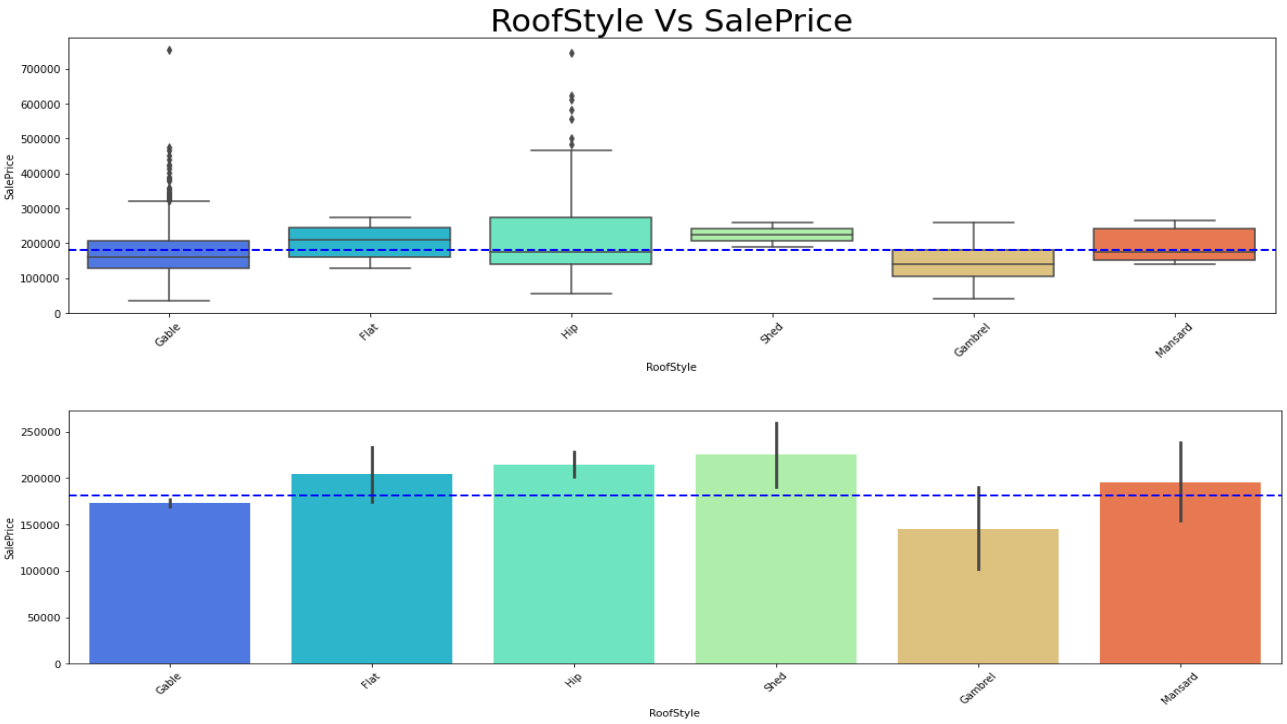
1. According to the Neighborhood NoRidge that is Northridge is being sold high and also costliest in the market.
2. NridgHt - Northridge Heights is the next costliest sold property with respect to neighbourhood.



KEY OBSERVATIONS:

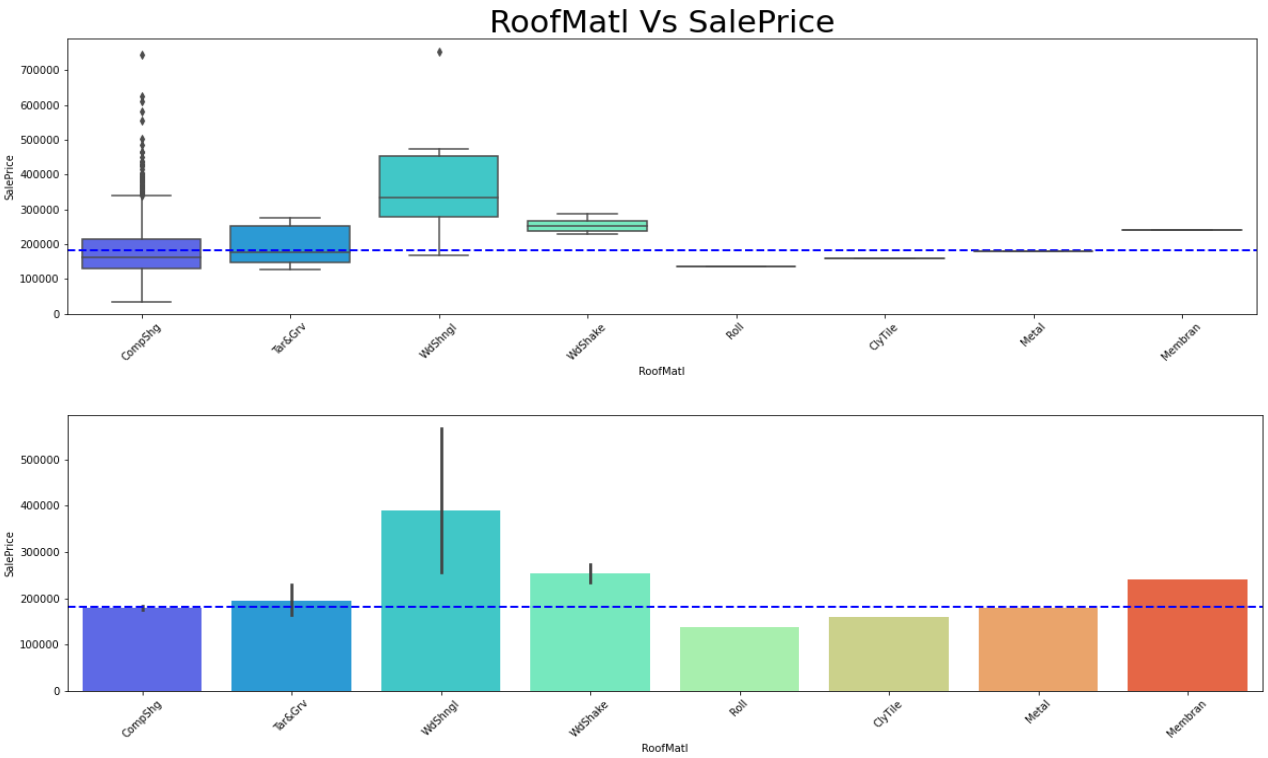
1. Since Utilities have only one values in all the columns it has no correlation. we will drop this column since it won't help in building the model.
2. The table have more positive correlation at the bottom. Overall Quality yearbuilt year remodified have high corelation with sales prize.

MULTI VARIATE ANALISIS: TABLE 2.



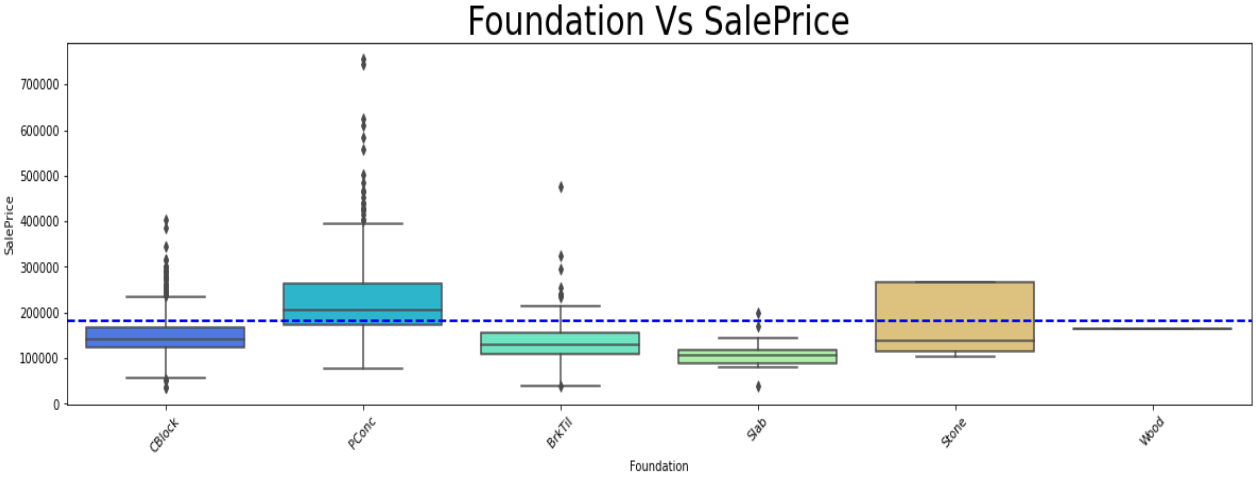
KEY OBSERVATION:

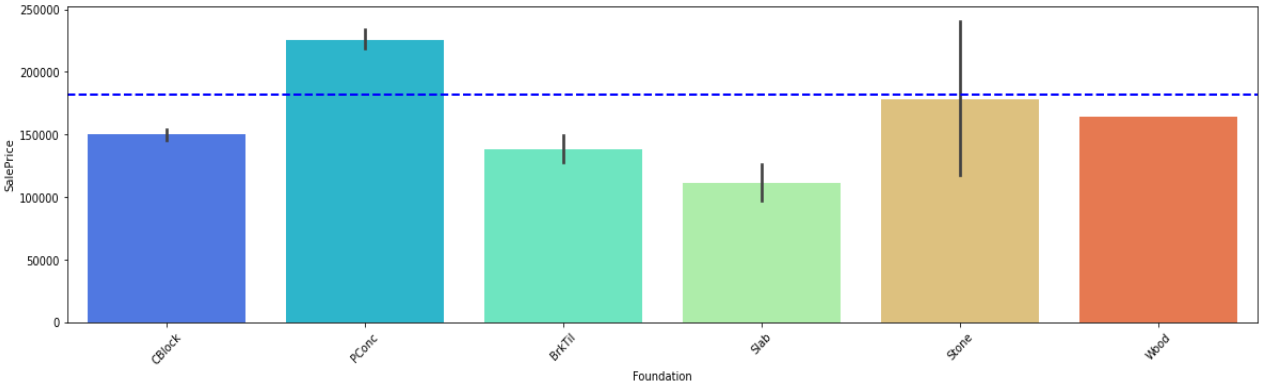
1. Roof style is Gable which is being the costliest in the type of roof style followed by hip.
2. But shed being the highest sold roof type.
3. From above we can analysis that the Gable roof style is costlier so mostly people prefer to buy shed roof type.



KEY OBSERVATION:

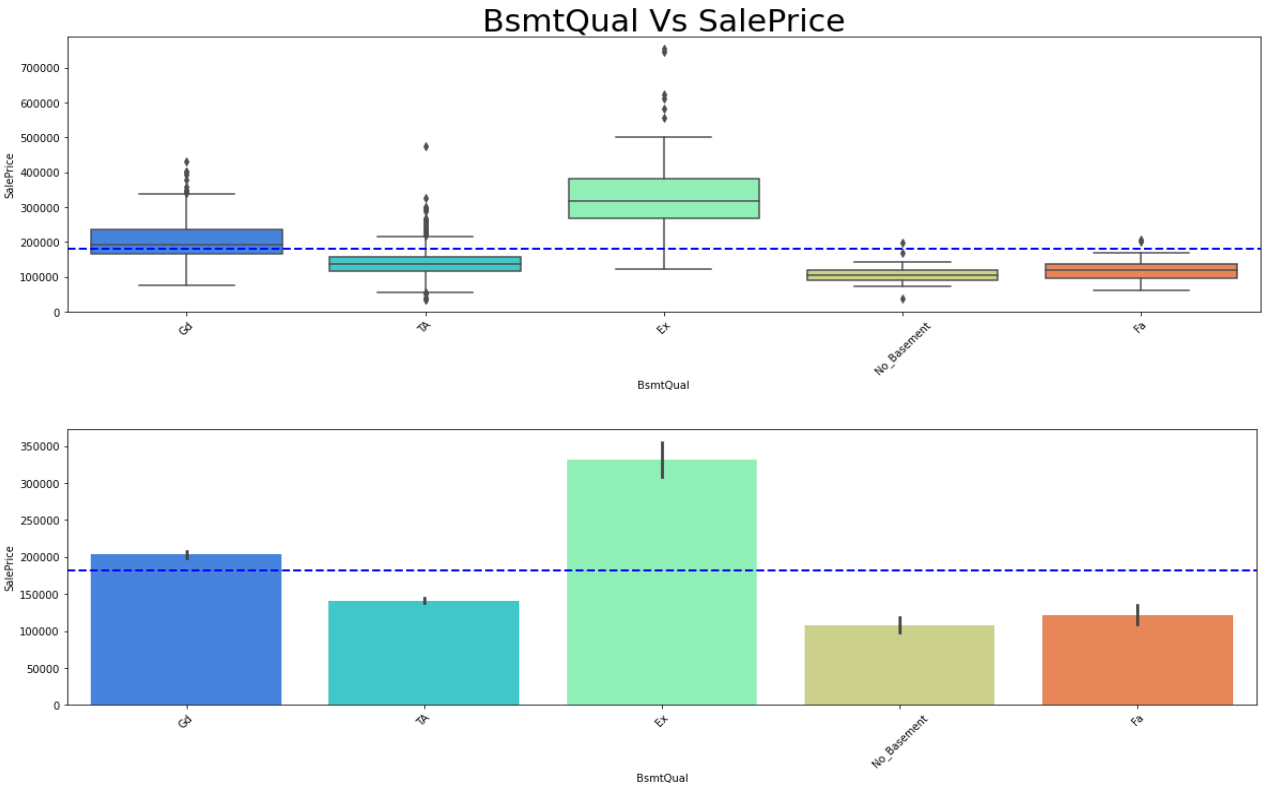
1. Roof material Standard (Composite) Shingle being the costliest but soled comparatively lesser than Wood Shingles.
2. Wood Shingles have many varieties of cost and it has been also sold higher and costlier than the other roofing materials.





KEY OBSERVATION:

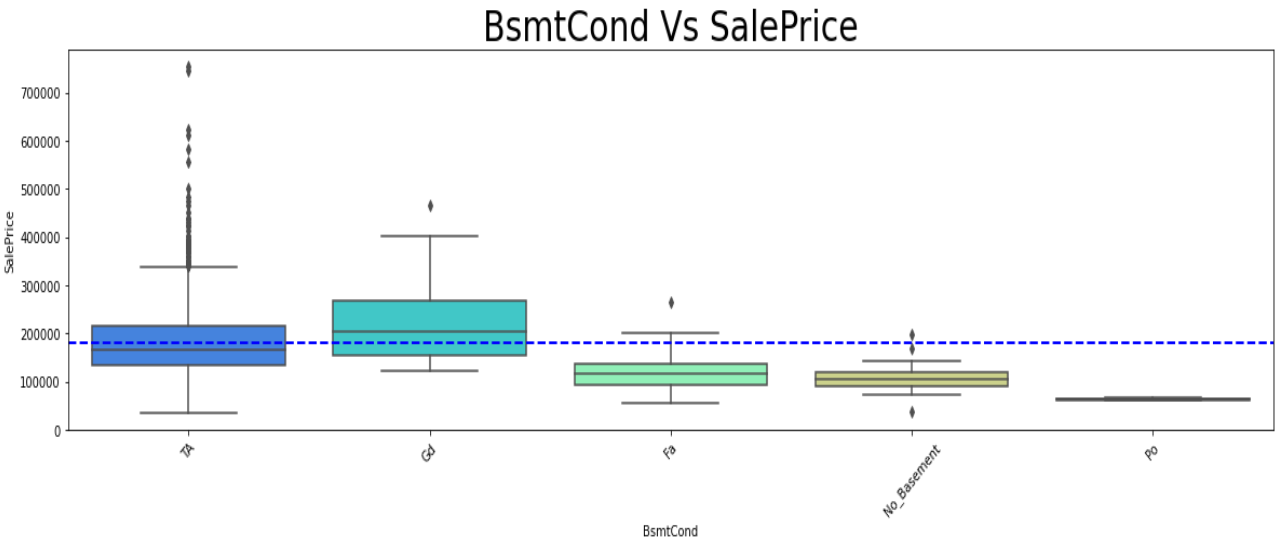
1. PConc - Poured Concrete type foundation is being costlier than the other and also sold in large number.
2. Stone foundations are next costliest foundation in type of foundation.
3. From above we can understand people are more interested in building with concrete or stone foundation and same type of building being build higher.

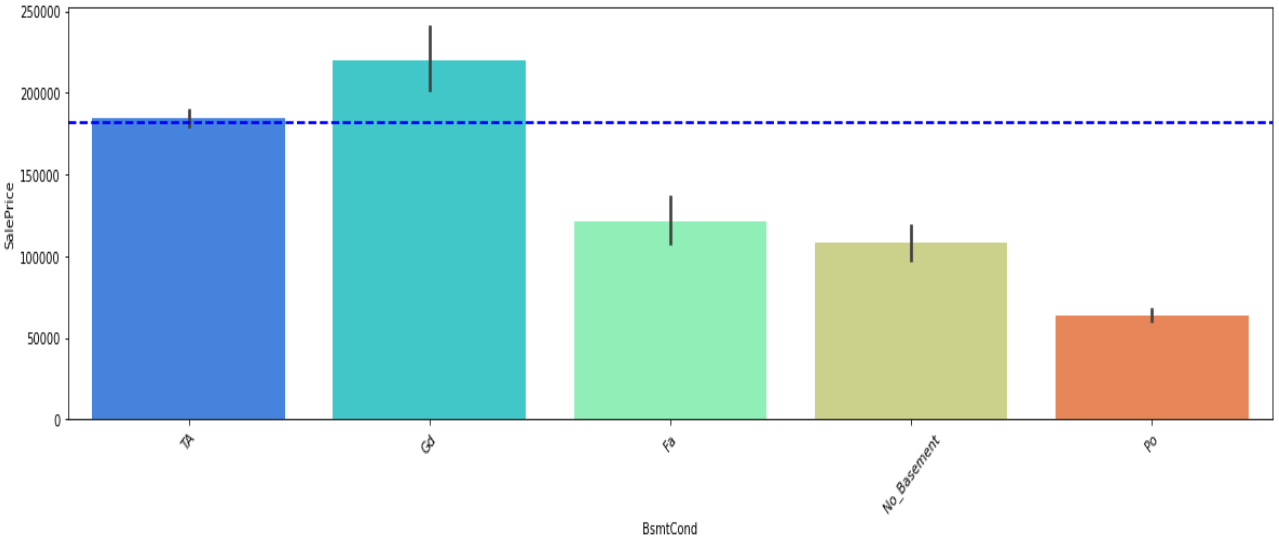


KEY OBSERVATION:

1. Ex - Excellent (100+ inches) being the highest sold and also costliest as like said before people are more interested in building with strong foundation.
2. Gd - Good (90-99 inches) good is the second highly preferred foundation type and also it is the second highly sold foundation in quality wise.

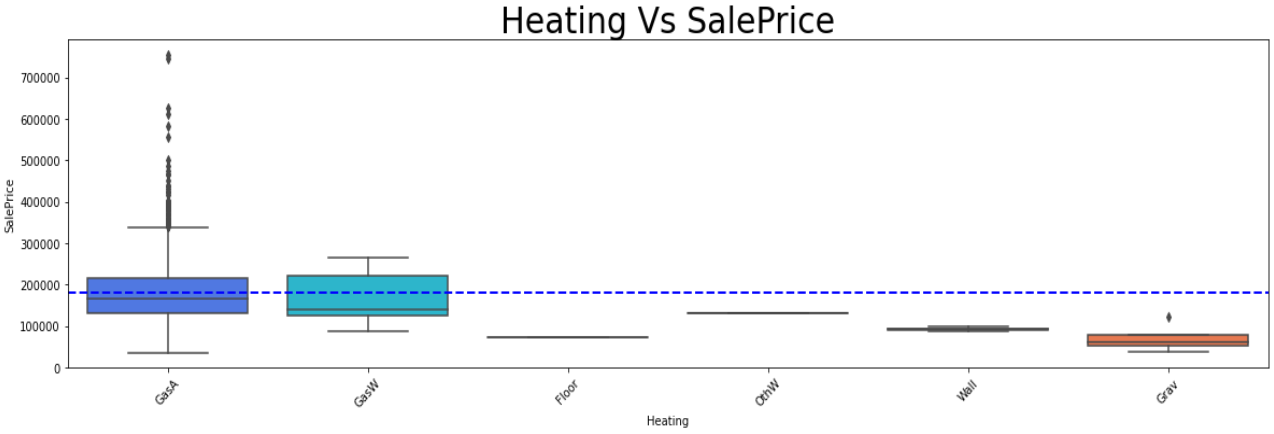
From above two we can say buildings are more build above good foundation type.

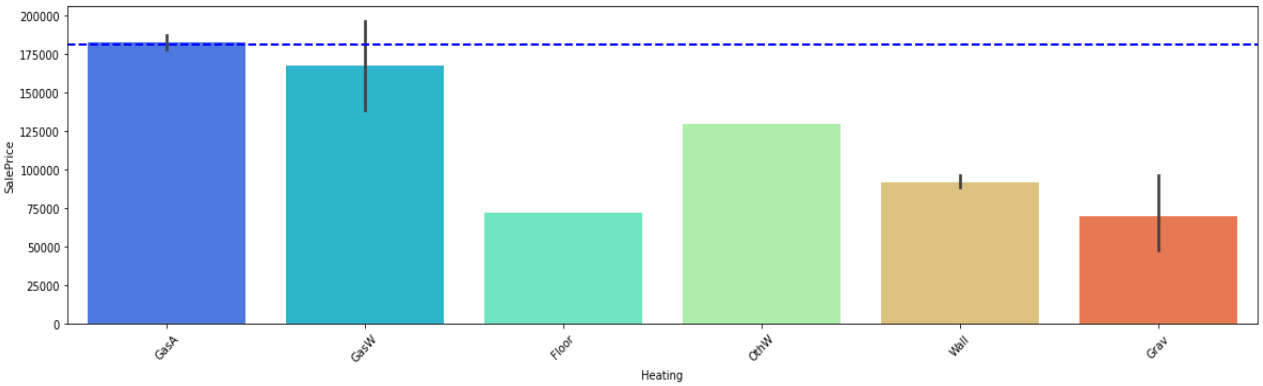




KEY OBSERVATION:

1. TA - Typical - slight dampness allowed is the costliest basement type in the property and this type of basement is second largely sold.
2. Gd- Good condition Basement type are largely sold but same is not costliest in the market. The remaining basement type are all sold below the average sale price of the property.

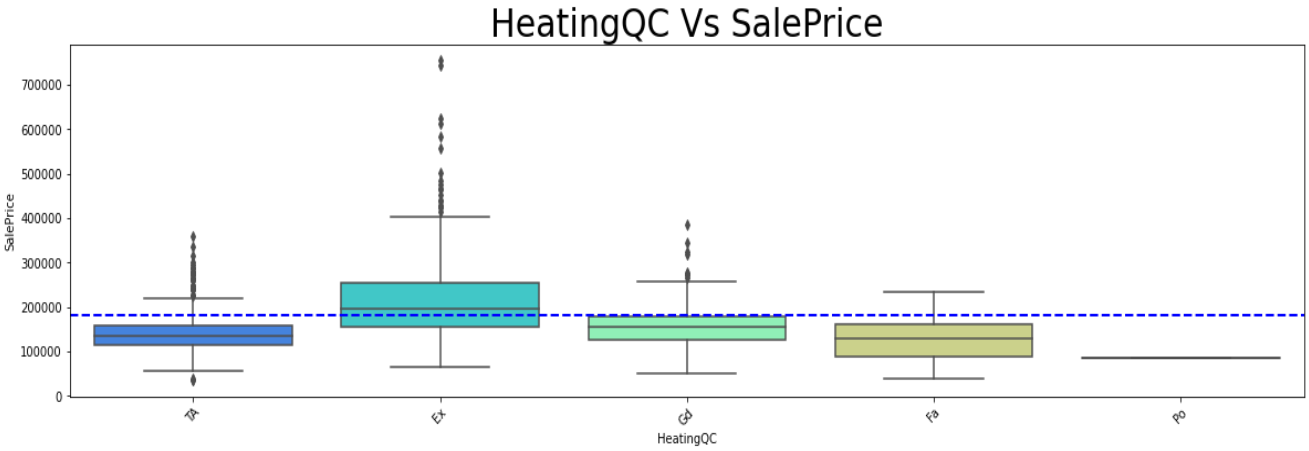


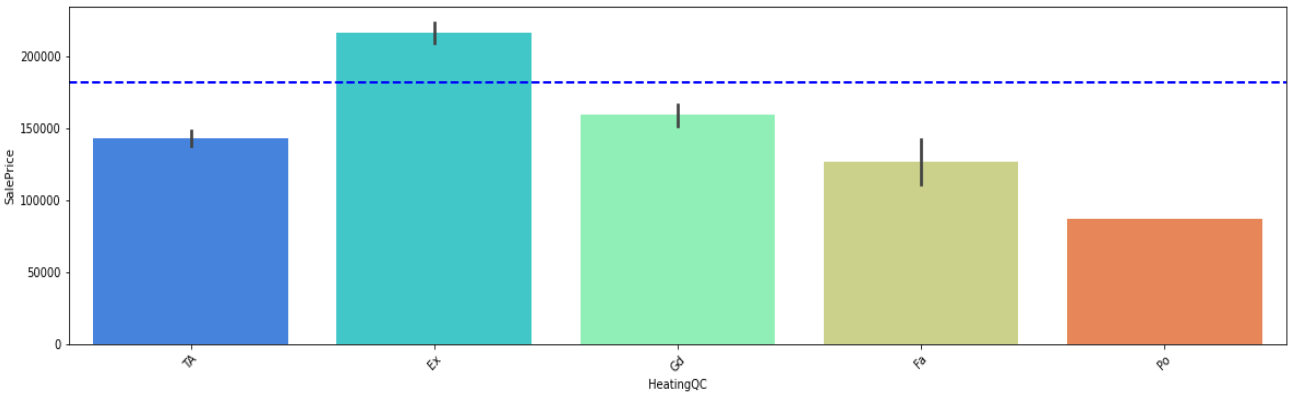


KEY OBSERVATION:

1. Heating is one of the most preferred quality of the property in cold countries and also more insulated houses saves more electricity from using heaters.
2. GasA - Gas forced warm air furnace is most sold heating facility and also being the costliest.
3. GasW - Gas hot water or steam heat is the second highly sold and second costliest but the difference between GasA and GasW is large.

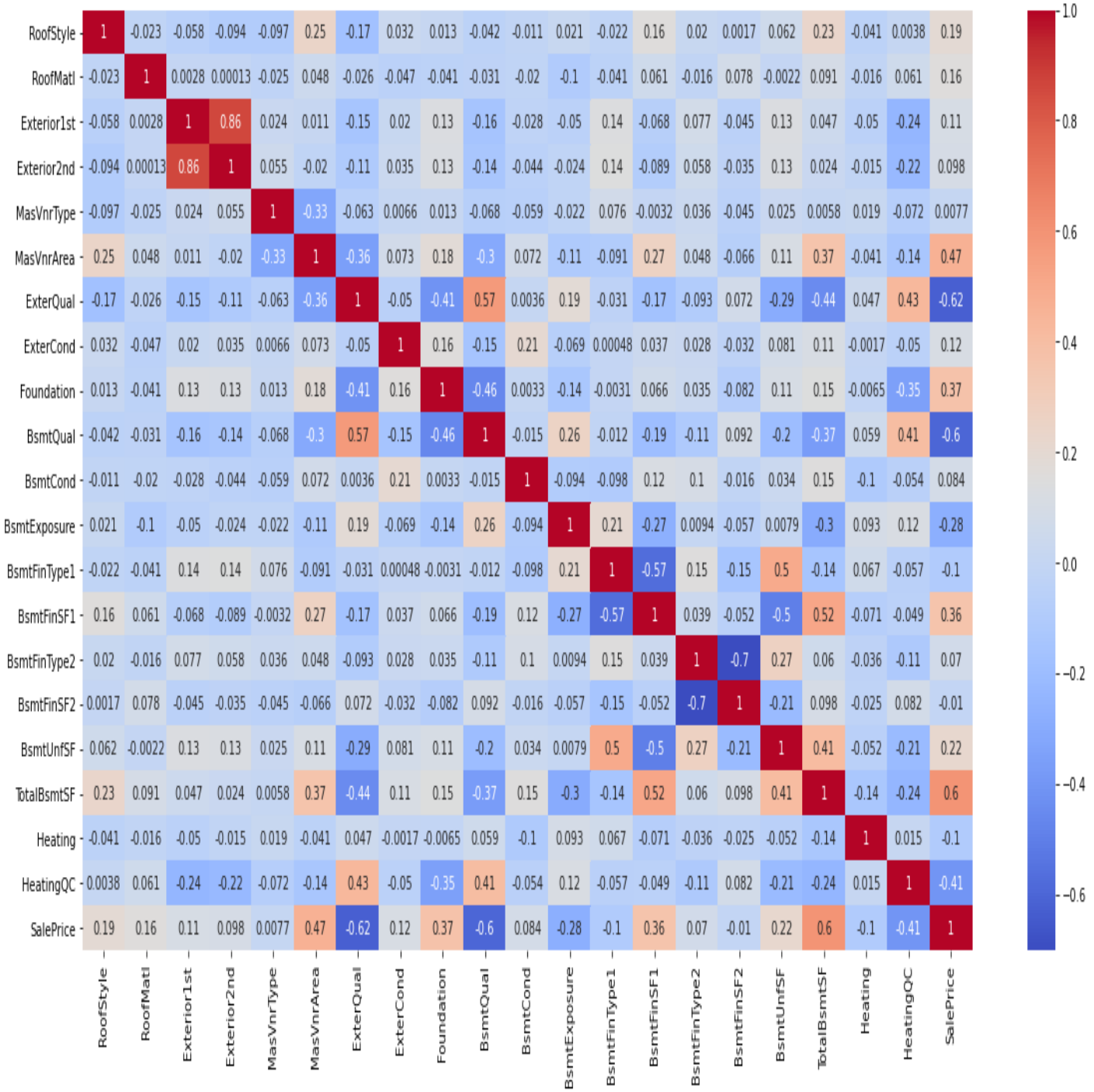
The above observation defines that a good quality residence has Gas forced warm air furnace but water or steam heated are also used in few residences.





KEY OBSERVATIONS:

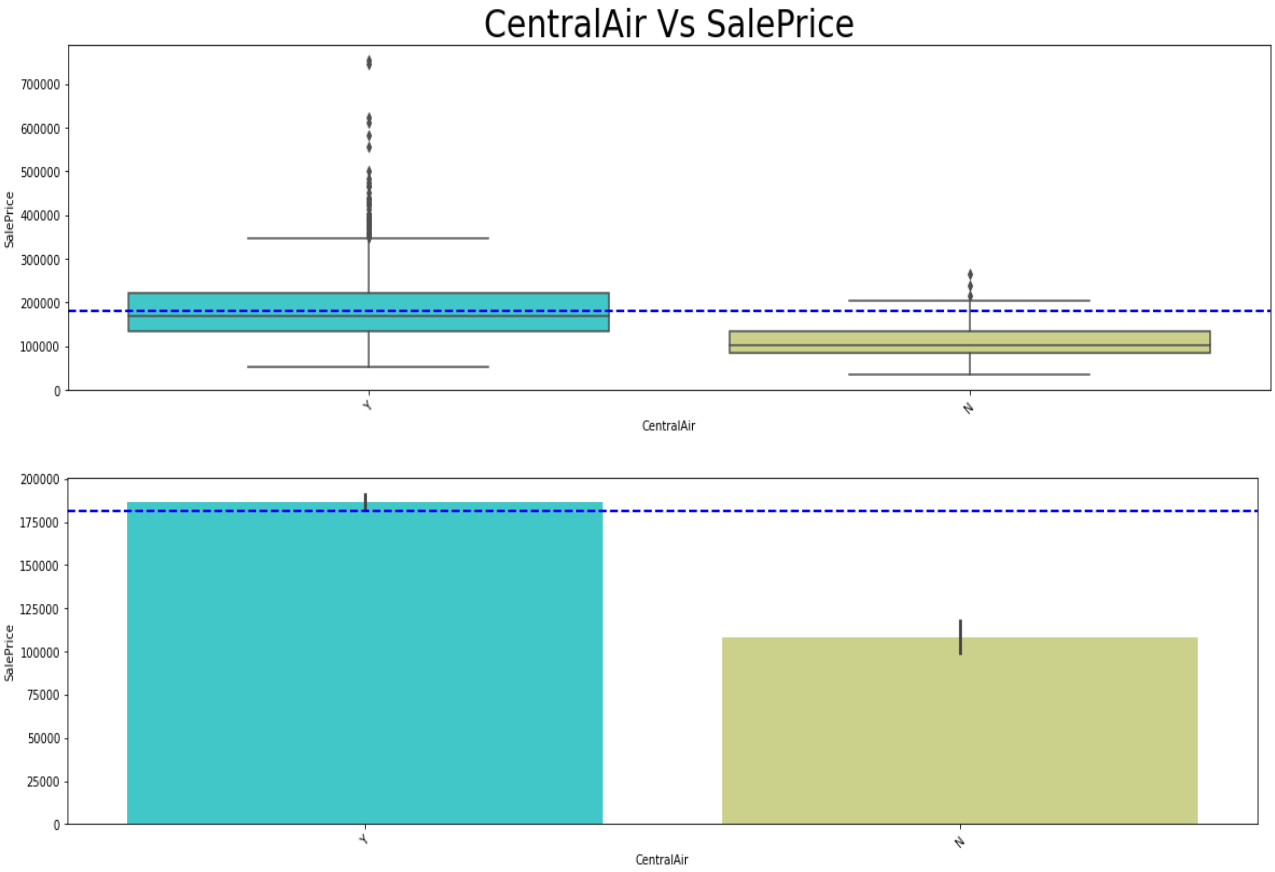
1. Ex- Excellent heating Quality are mostly preferred in good properties.
2. Excellent heating is mostly with GasA - Gas forced warm air furnace we can say this since because the cost is matching.



KEY OBSERVATIONS:

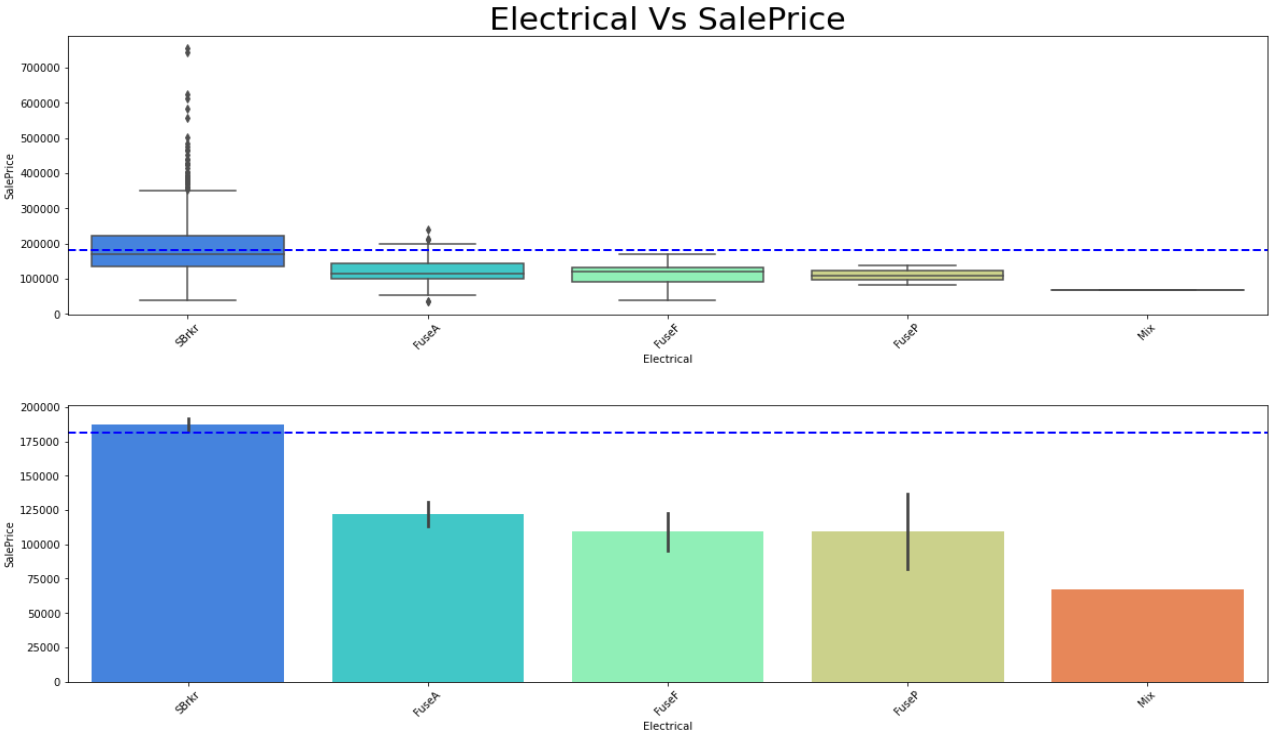
1. We can observe the feature variable have lesser correlation among themselves but they have high correlation with target variable.
2. Exterior1st and Exterior2nd have high correlation with themselves to avoid multicollinearity we will drop Exterior2nd.
3. The feature variables have high positive al well as negative corelations as we know that positive corelation increases the price od the property and the negative correlation will reduce the price of the property.

MULTI VARIATE ANALISIS: TABLE 3



KEY OBSERVATIONS:

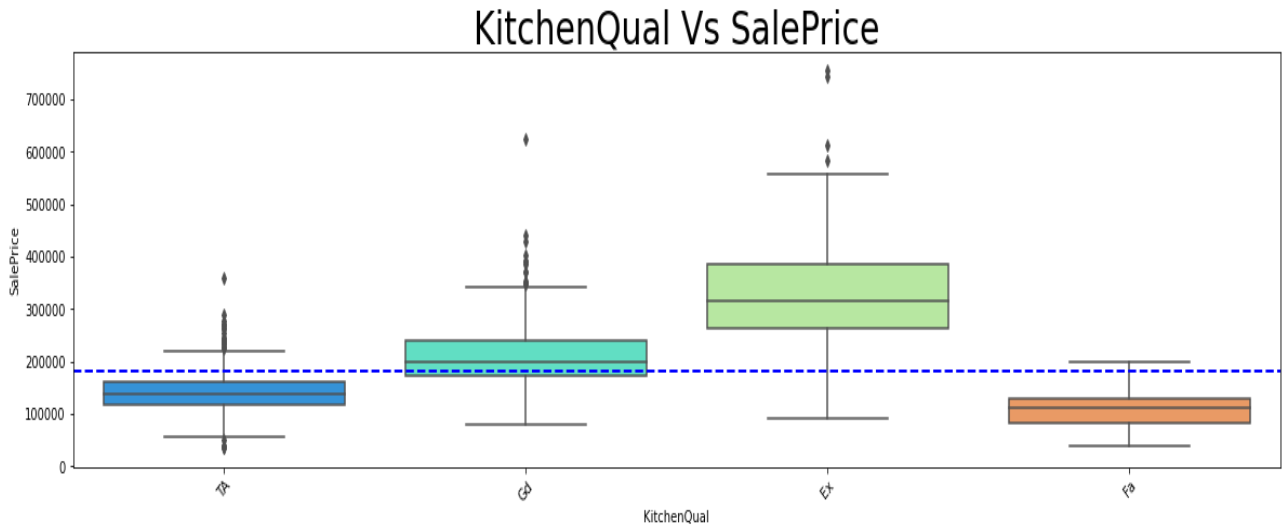
1. We can observe the most property is been preferred with central air conditioning and also being the costliest.

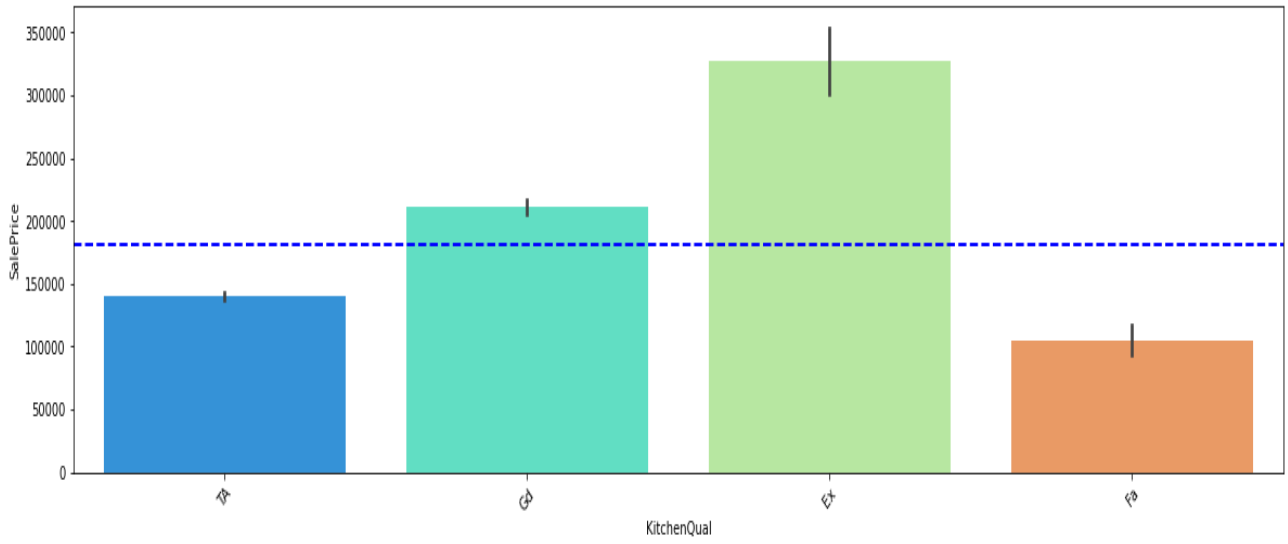


KEY OBSERVATIONS:

1. SBrkr is **Standard Circuit Breakers & Romex** and also this is considerably safer than any other electrical circuits in industry.
2. **Standard Circuit Breakers & Romex** is getting soled higher and also the costliest across all the electrical systems.

From above we can say that properties are build as with more safety and more reliable electrical equipment's over other.

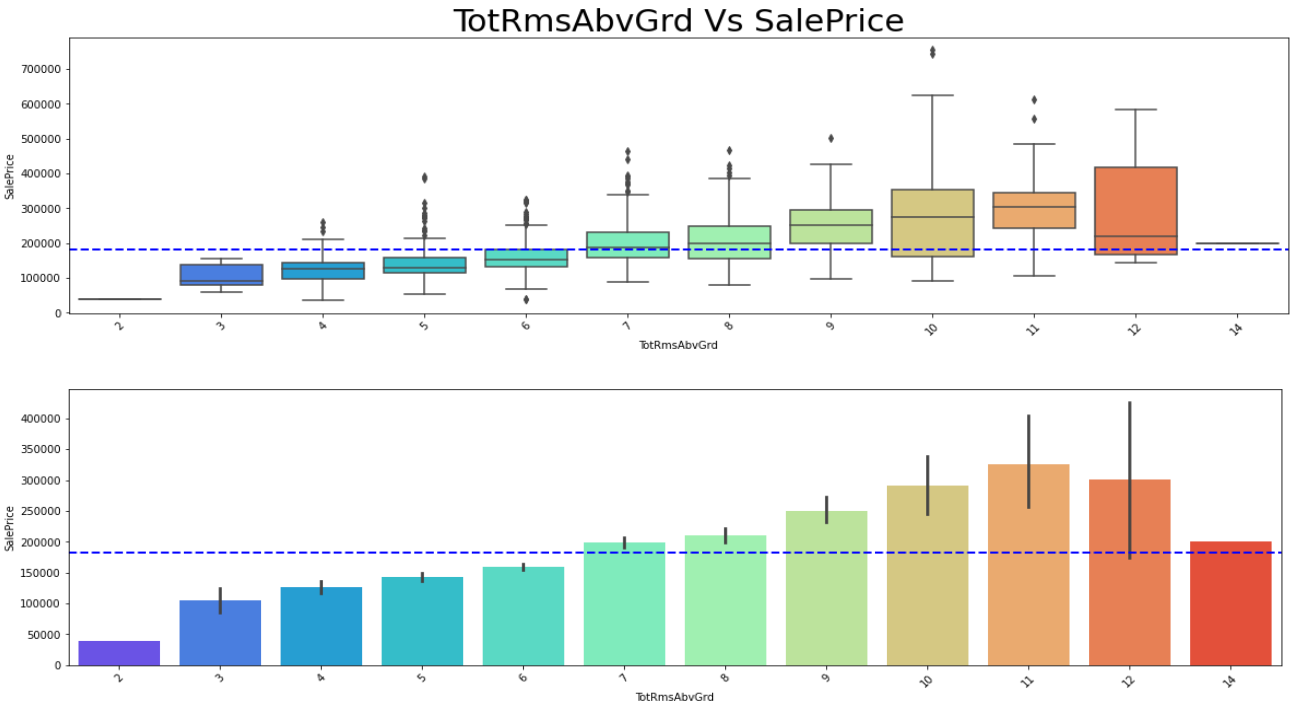




KEY OBSERVATIONS:

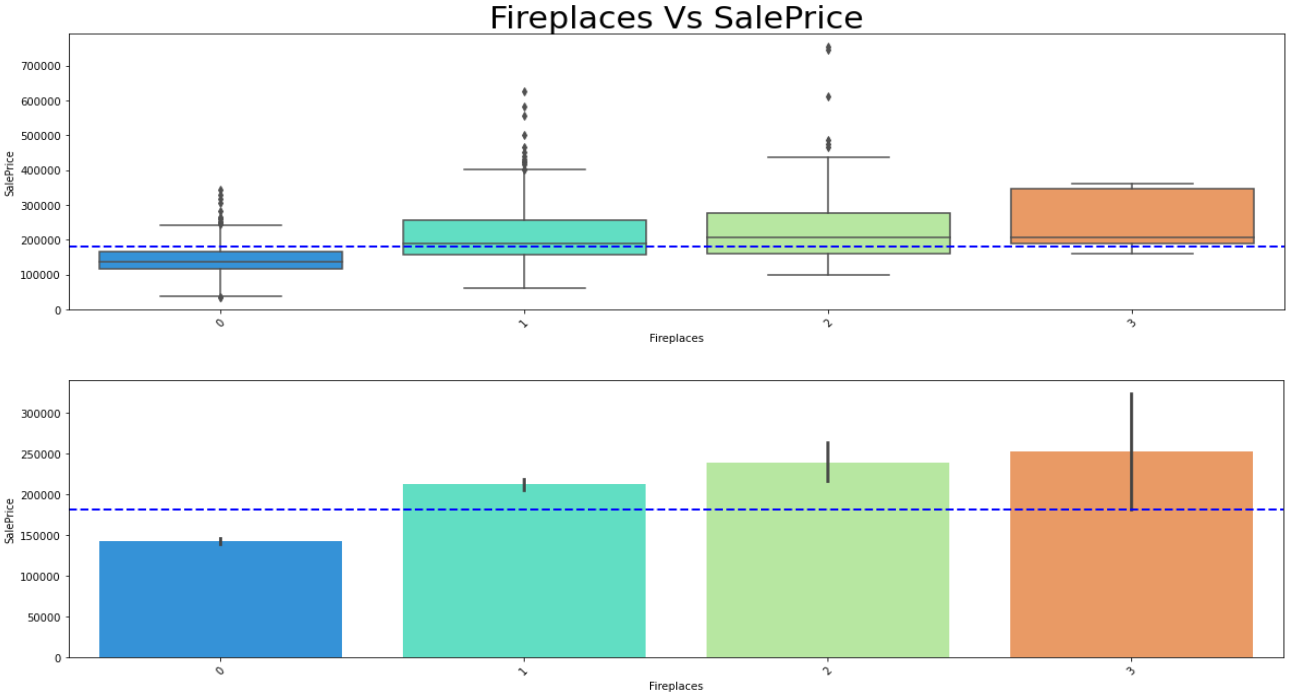
1. There was a saying the Quality of the kitchen is the beauty of the house, as similar to that we can see the excellent quality in kitchen will increase the cost of the property.
2. And also, the excellent quality of kitchens is being mostly build.
3. Good Quality in kitchen stands second in the order and also in number of units sold.

From above we can narrate a story that people mostly preferred good quality kitchens and also good and excellent quality of kitchens are being costlier.



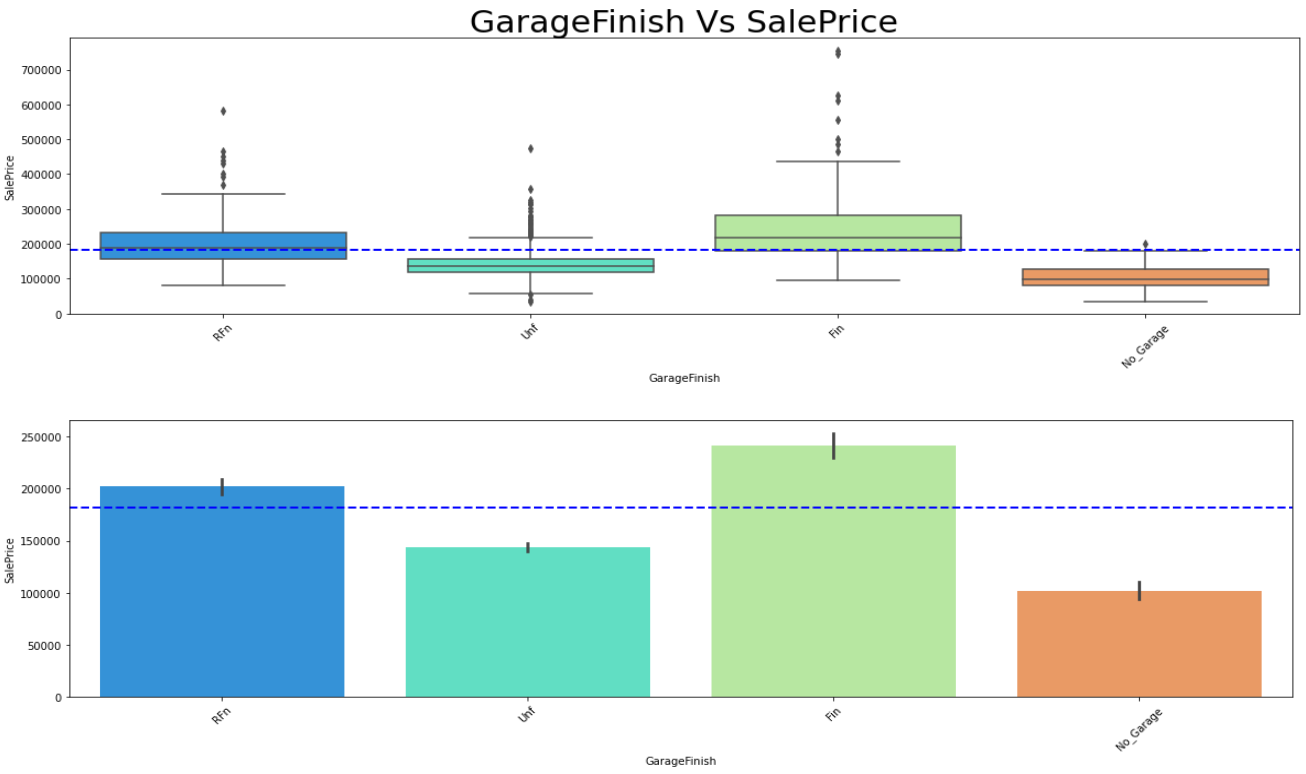
KEY OBSERVATIONS:

1. We can observe that the increase in number of rooms also increases the cost of the property, the highest sold property is 11 and 12 rooms property.
2. And costliest of the properties is 10 rooms sold property.



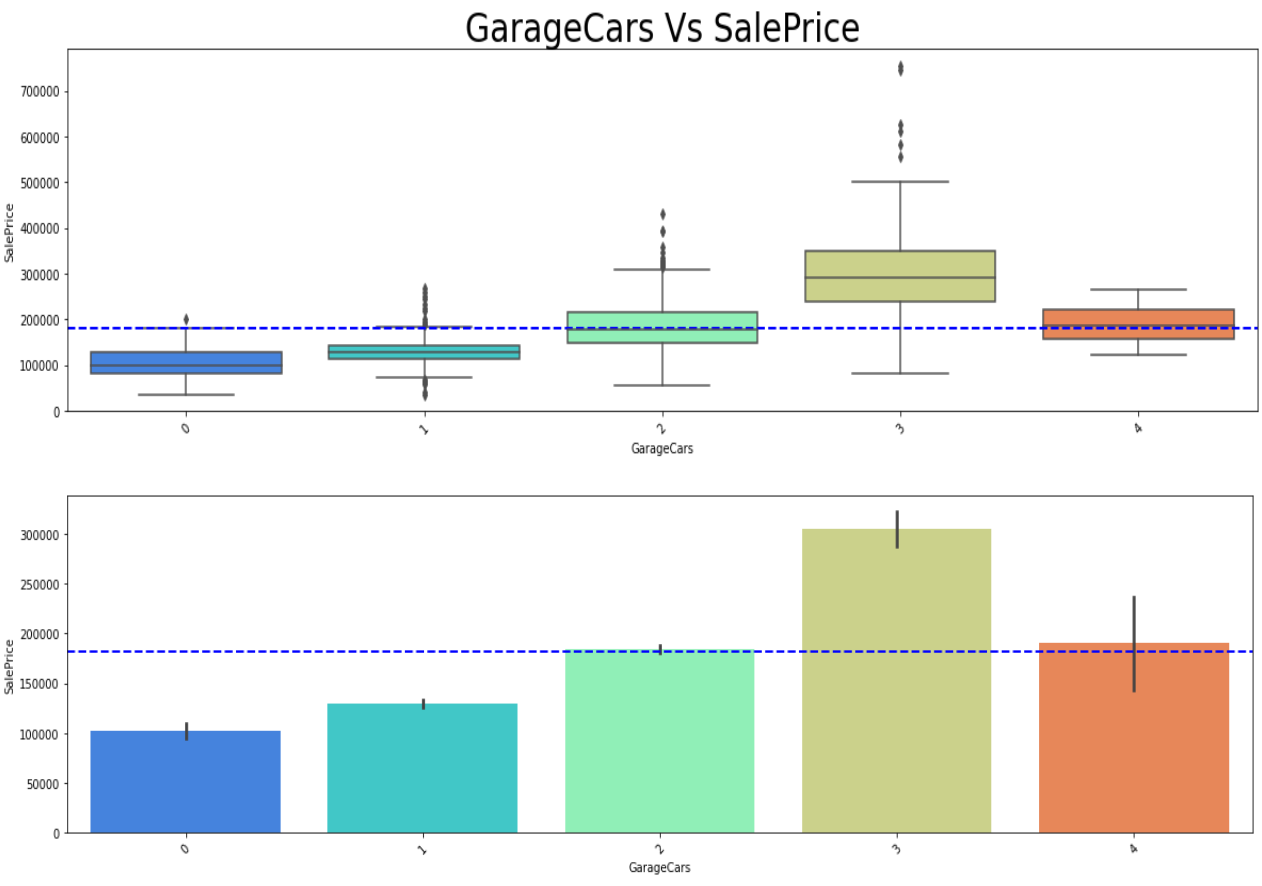
KEY OBSERVATIONS:

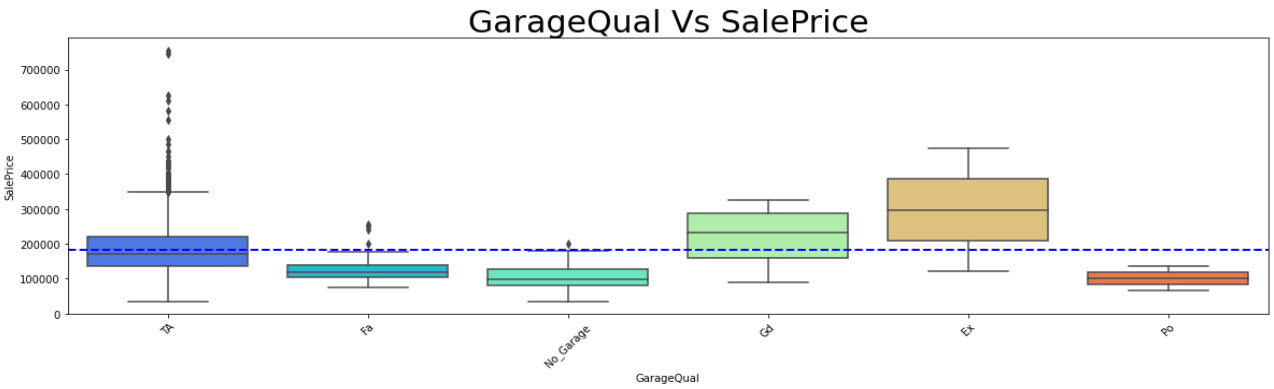
1. We can observe that the property that have two fire place is costliest of all and with out fireplace is also reduces the case of the property below average.

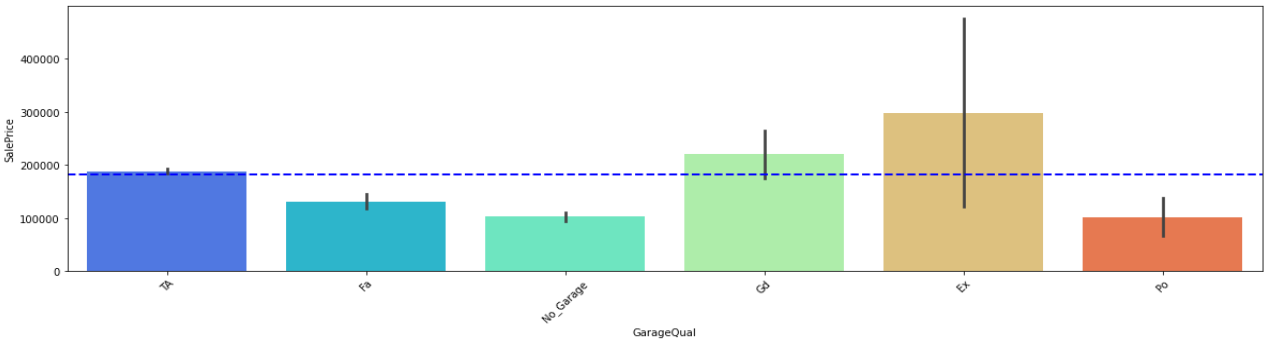


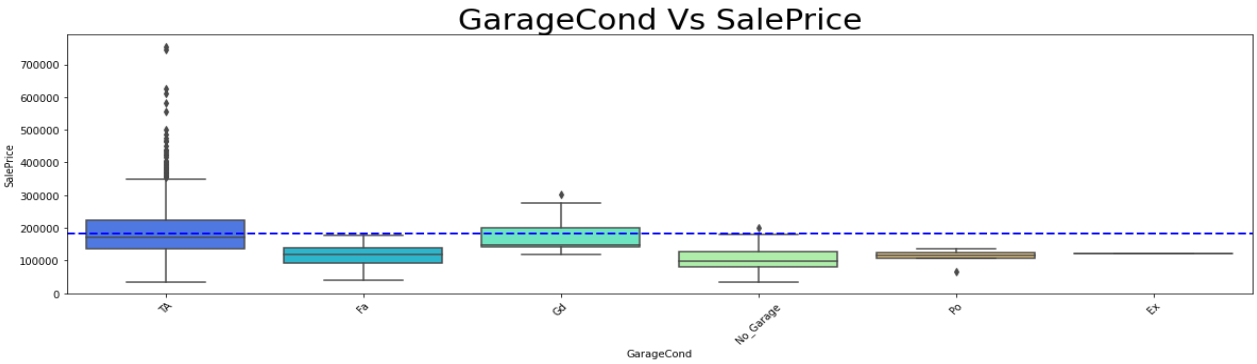
KEY OBSERVATIONS:

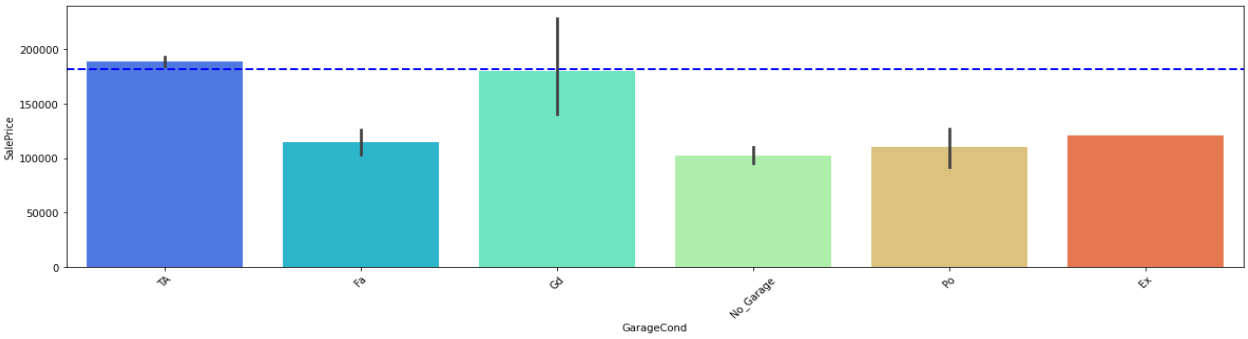
1. We can observe the feature variable have lesser correlation among themselves but they have high correlation with target variable.





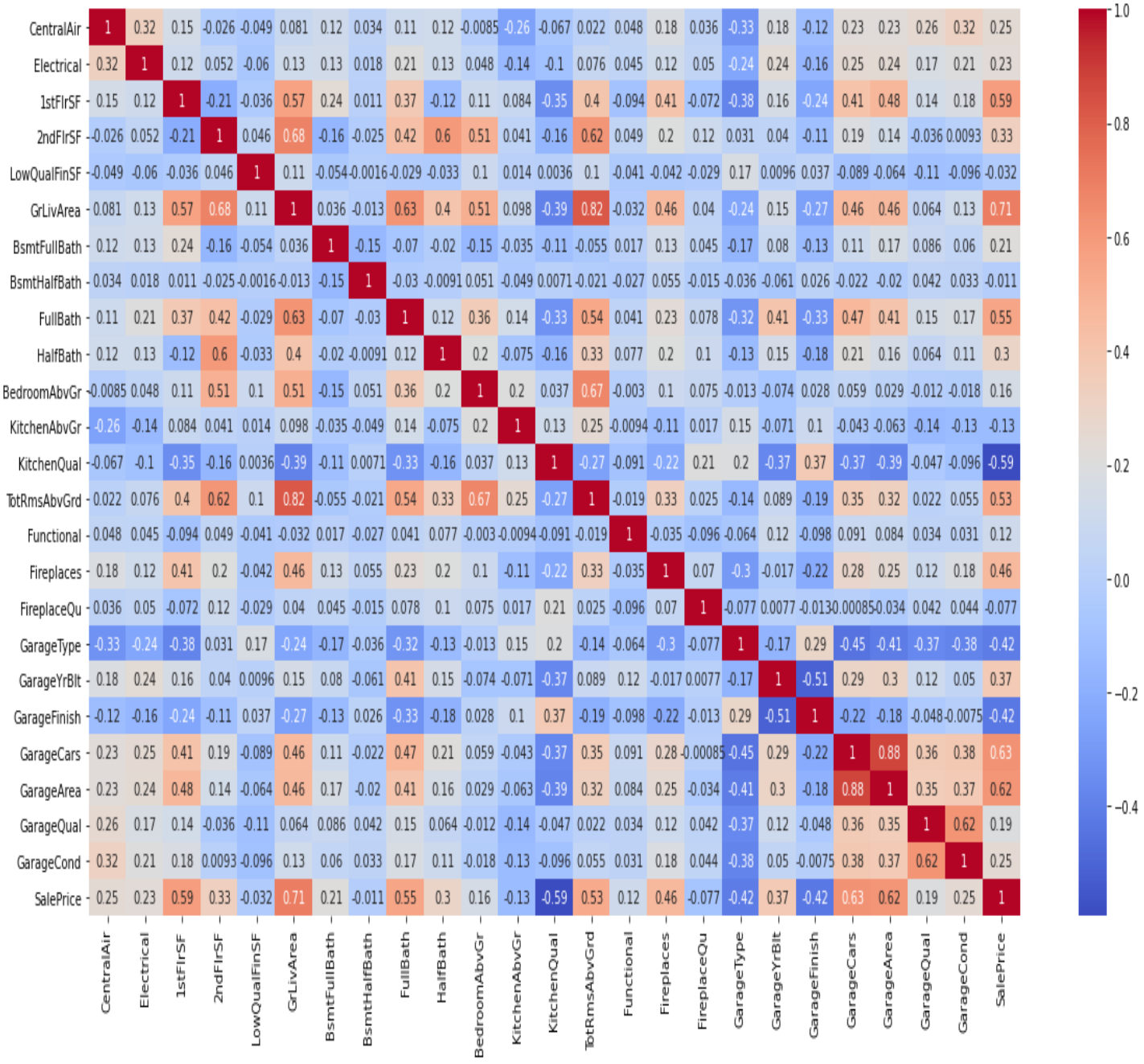






KEY OBSERVATIONS:

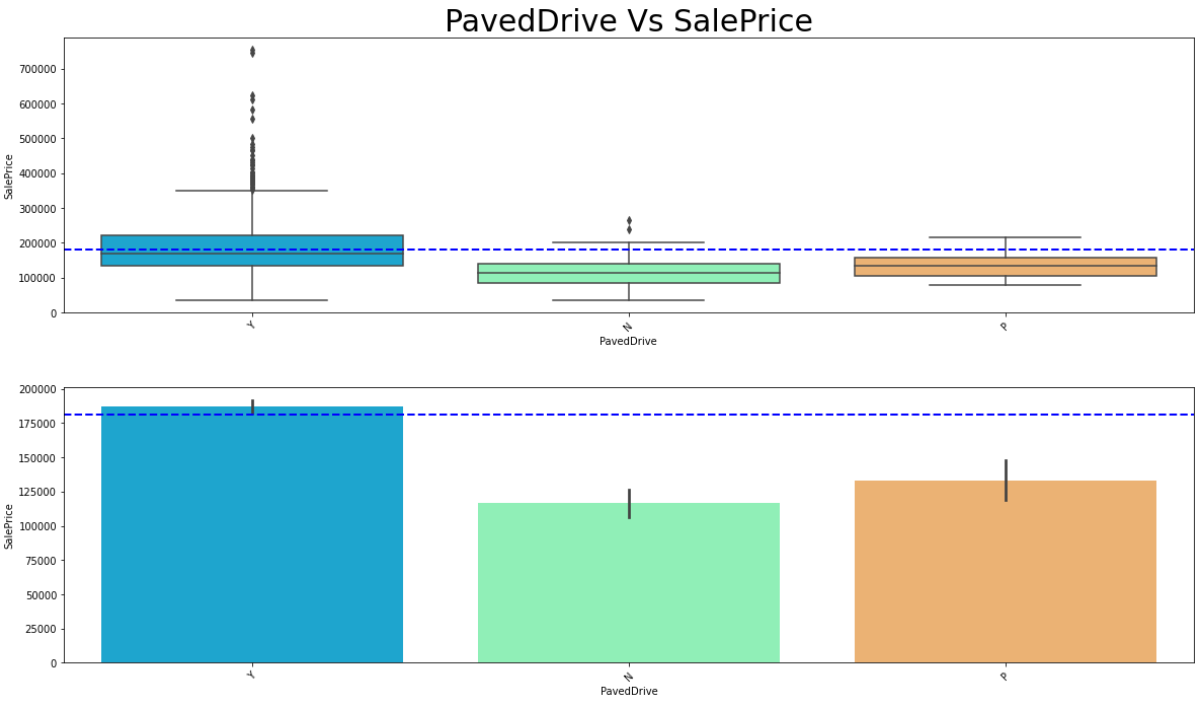
1. From the above three observations that is Garage cars Garage Quality and garage Conditions we can see three cars garage are costliest since because of the area of accommodation.
2. People mostly preferred excellent quality with good condition which is also not being costliest.
3. Typical/Average quality and conditioned associated property or house being costliest of all the property.



KEY OBSERVATIONS:

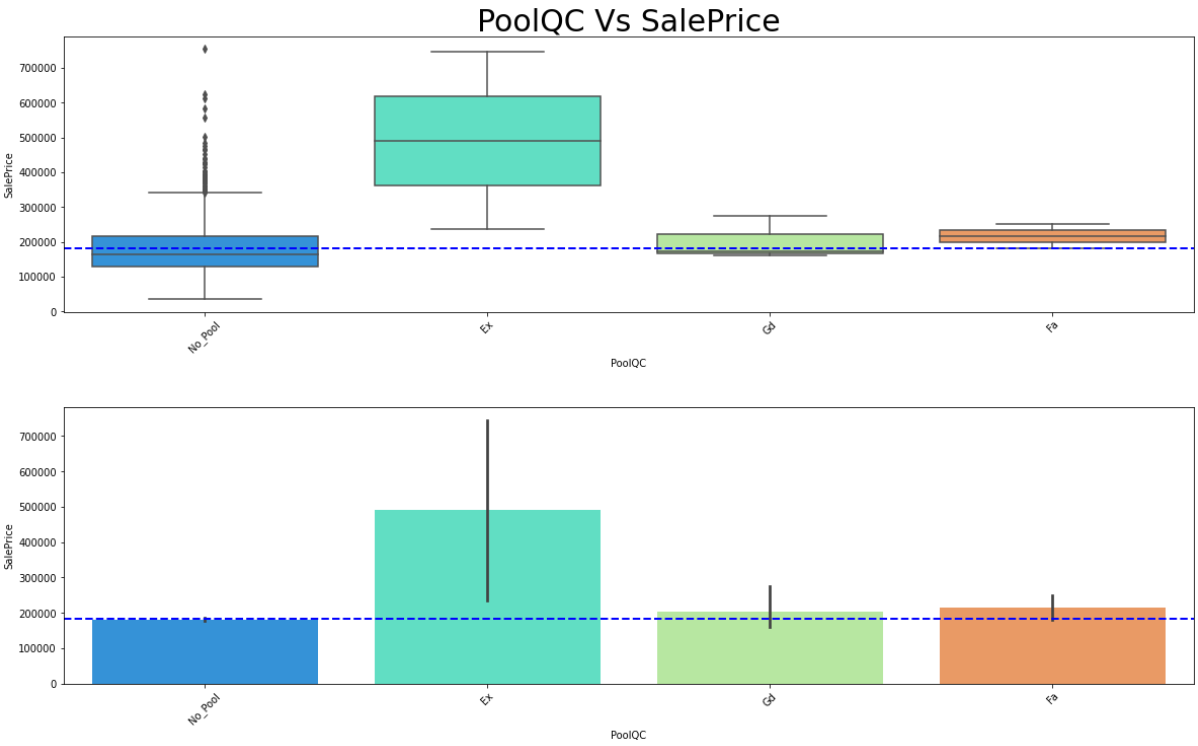
1. We can observe according to the correlations the scatterplot points are distributed.
2. We can see high positive correlation of feature variables with the sales prize.
3. We can see Garage Area, garage cars, garage area built, Fireplace, total rooms available, full bath, living area, 1stFlrSf are more positive corelated which means the increase in the above will also increased the cost and selling price of the property.
4. KitchenQual, Garagetype, GarageFinish are negatively correlated and the increase in that will decrease the cost and the selling price of the property.
5. We also can see there are multiple variable those have more correlation among themselves than the target variable these variables will create the multi collinearity problem. To overcome those problem, I am dropping one of those variables.
6. Garagecars, TotalrmsAbvGround can be removed to avoid multicollinearity.

MULTI VARIATE ANALISIS: TABLE 4



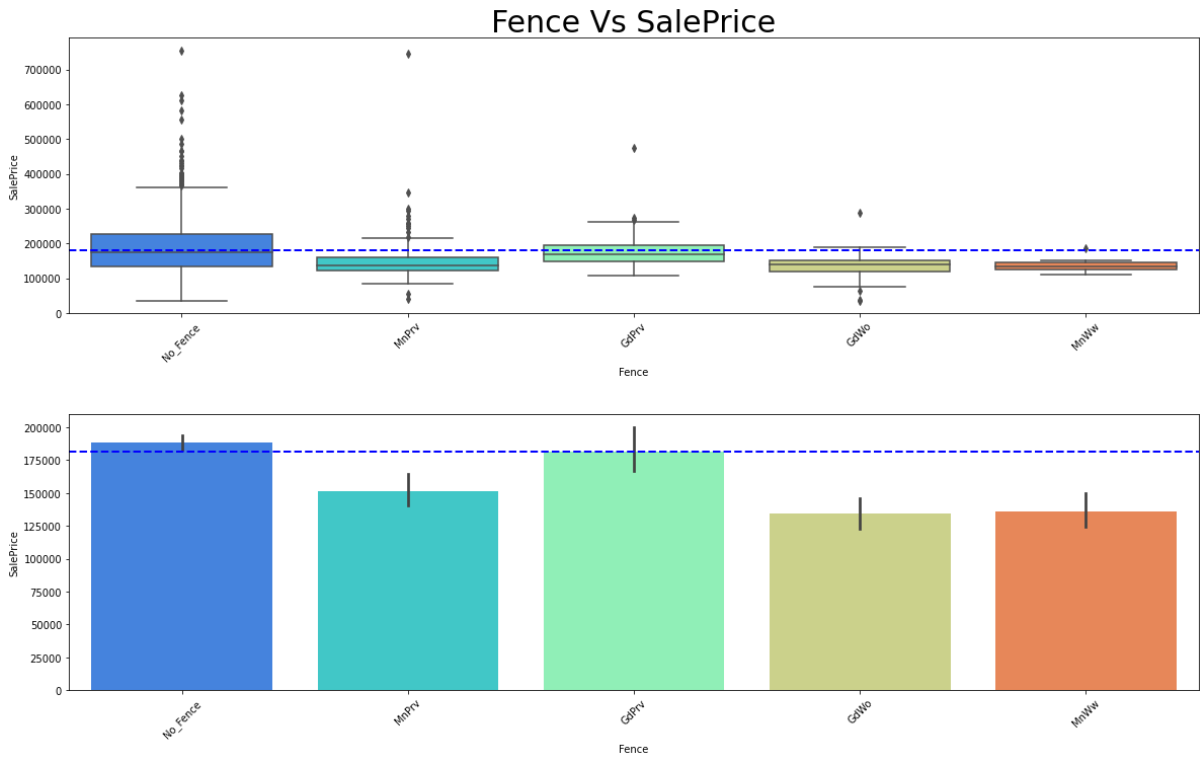
KEY OBSERVATIONS:

1. We can observe property with paved drive sold higher and costlier.



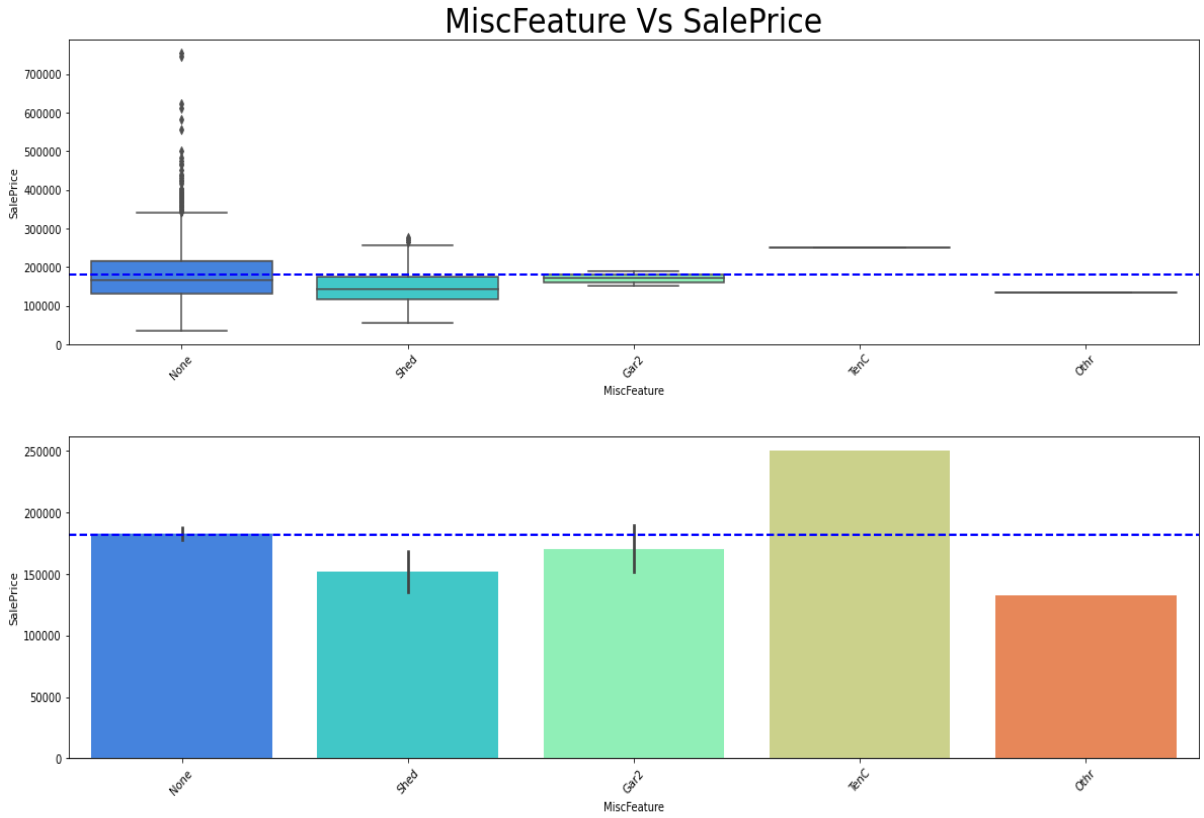
KEY OBSERVATIONS:

1. We can observe can observe two things excellent pool conditioned properties and no pool properties are soled costlier, but excellent pool conditioned properties are sold higher.



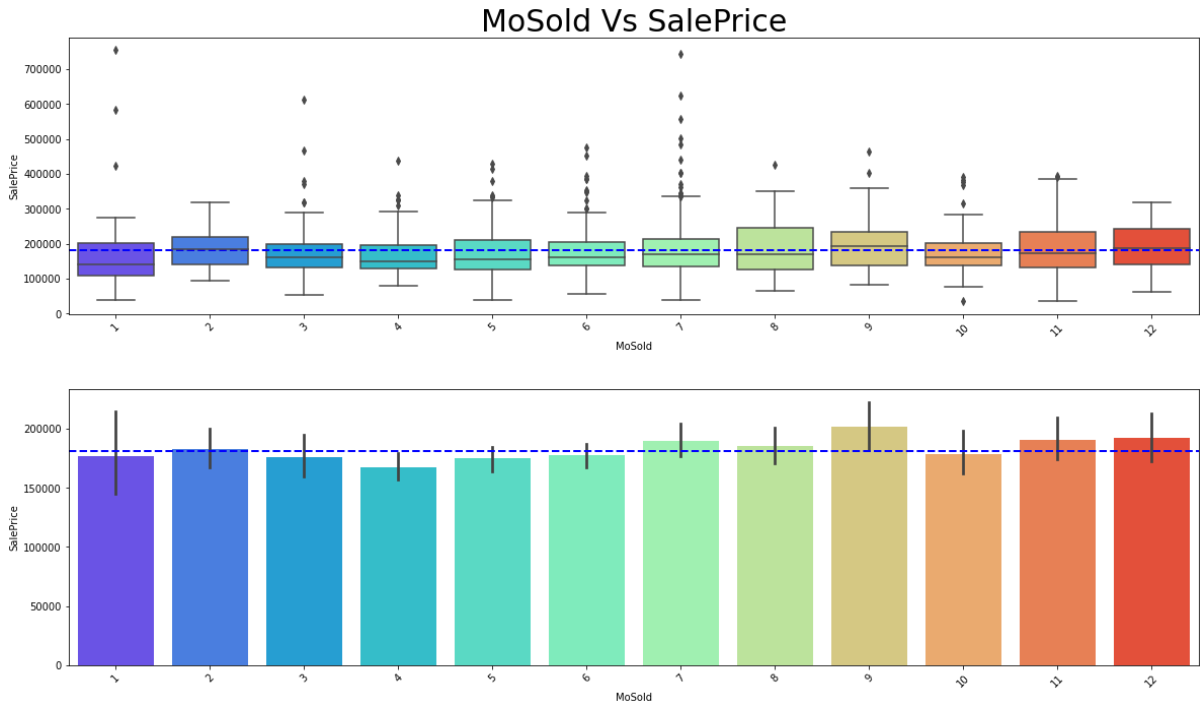
KEY OBSERVATIONS:

1. No Fence and Minimum Wood/Wire fence are sold higher and costlier above all other fence type.
2. Most of the costliest properties are in low neighbourhood in open area so fence is not being built around is an assumption.



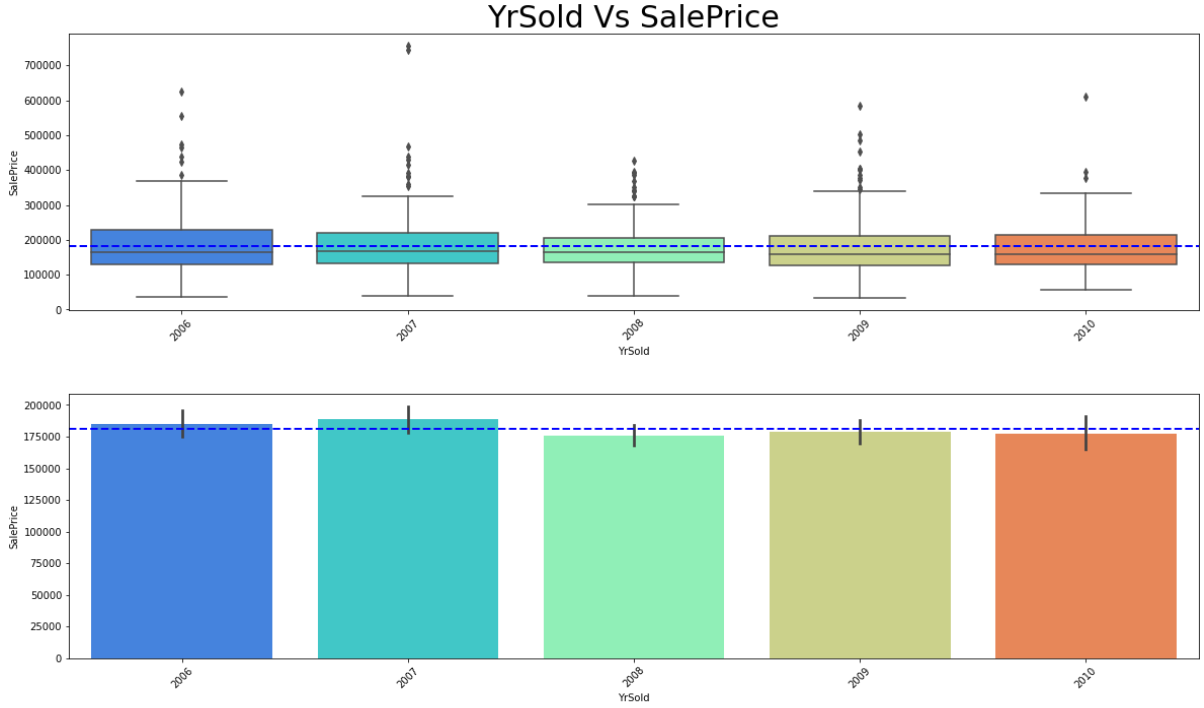
KEY OBSERVATIONS:

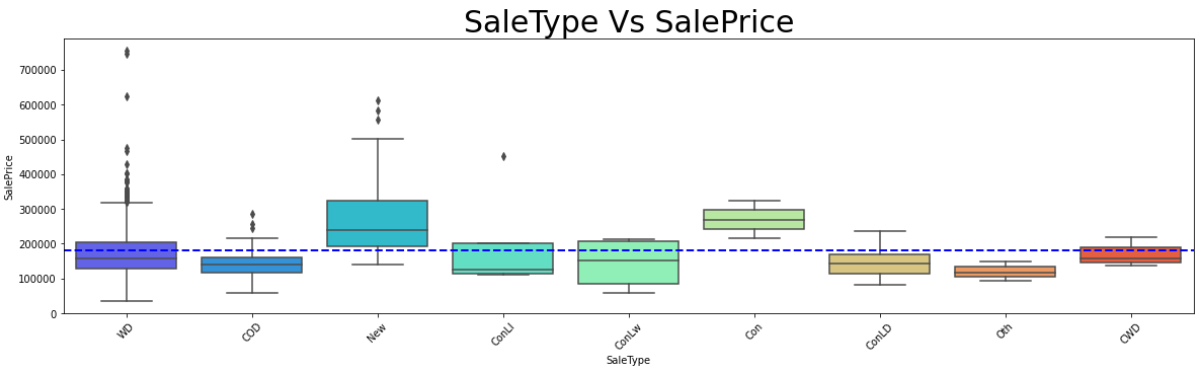
1. No other features associated with the properties are sold more.
2. Tennis court amenities with the property are sold costlier followed by two garage and shed over 100 sf are sold costlier.

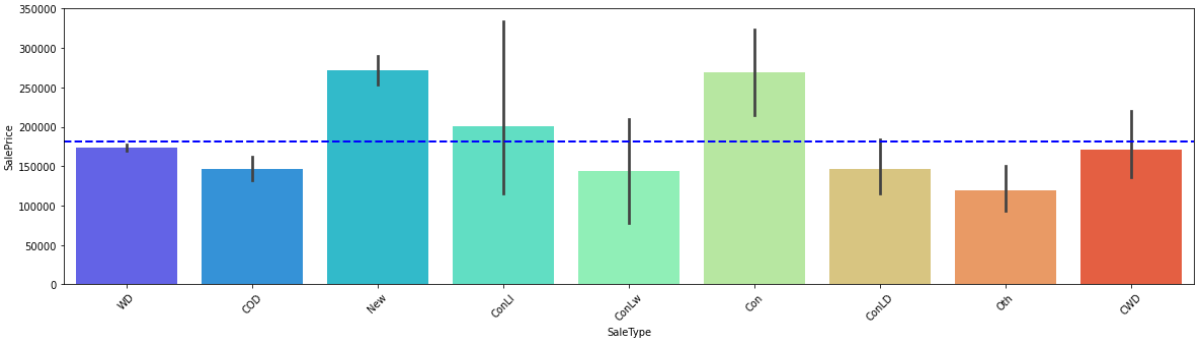


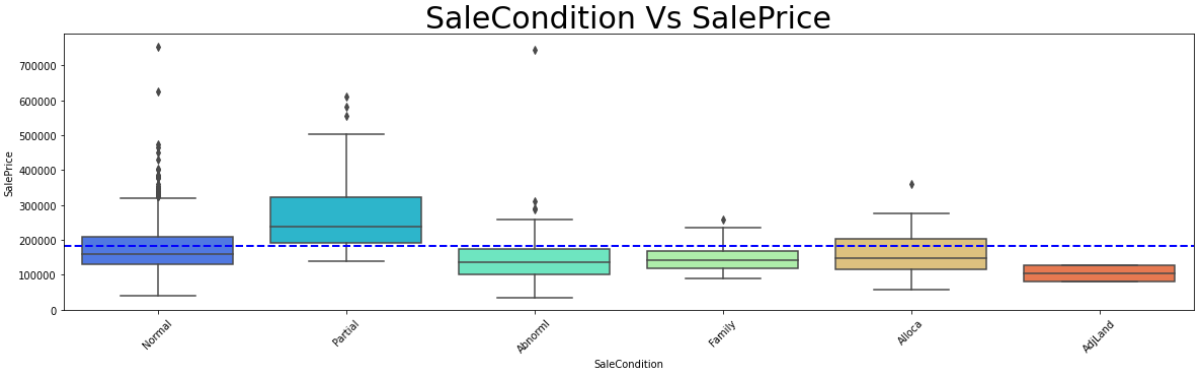
KEY OBSERVATIONS:

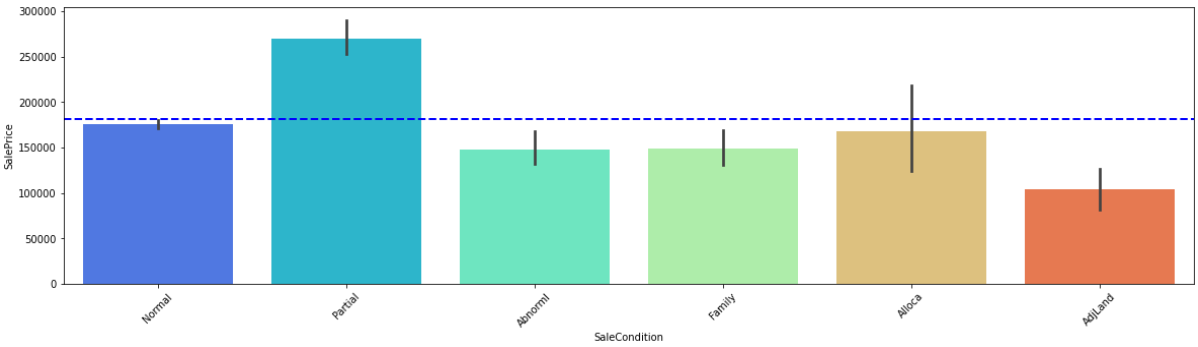
1. 1 month old are sold higher and costlier followed by 7 months old.
2. Howsoever 9 months old or more than that are being sold higher, assumption from this is to sell a property post the building of properties takes at least more than 9 months.





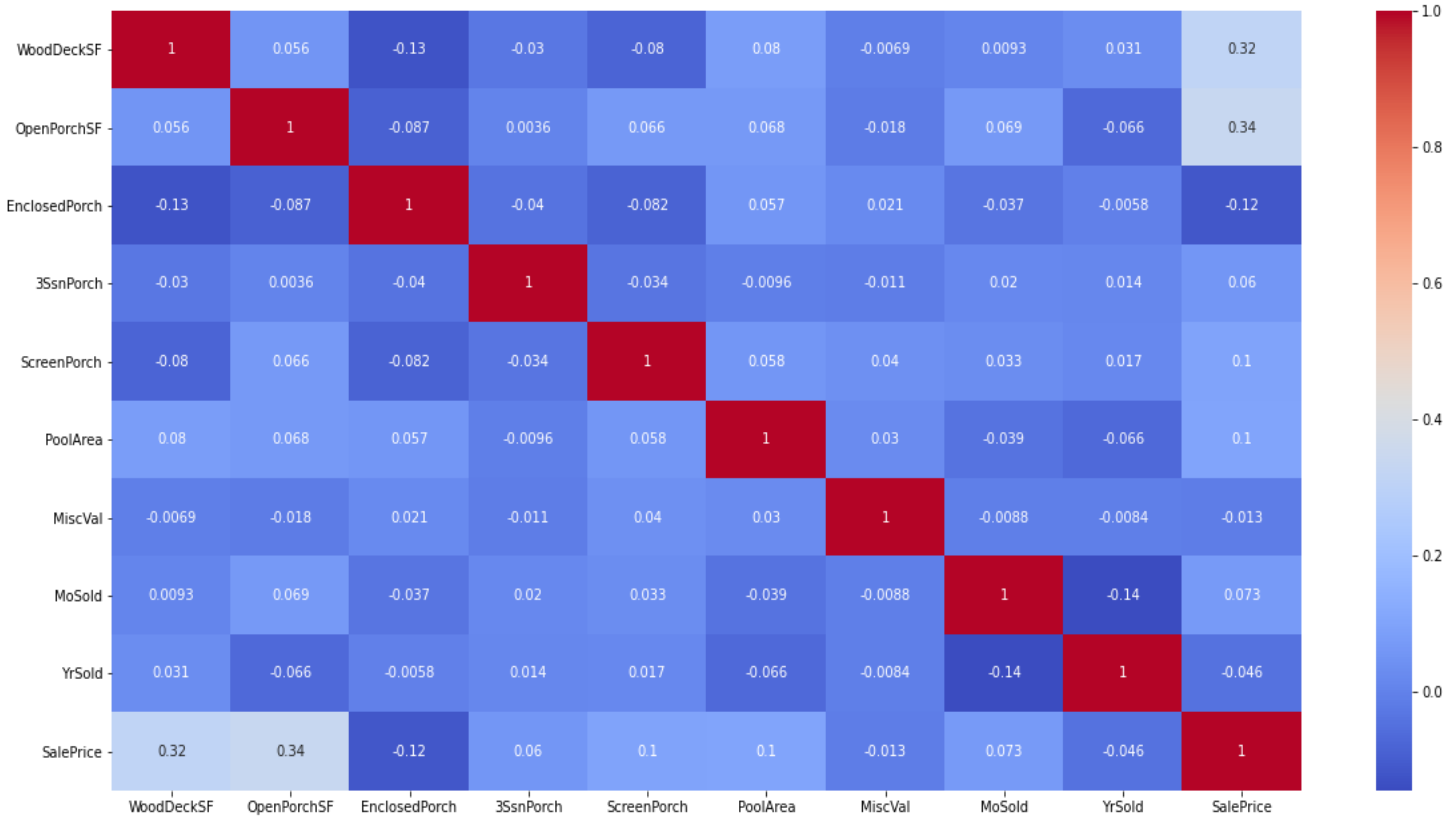






KEY OBSERVATIONS:

1. From above three categories the year sold, sales type and Sales condition year sold at 2007 with sale deed type Warranty Deed - Conventional Home was not completed when last assessed is costlier in market.
2. Year build is 2007 Warranty Deed – Conventional and Normal Sale is costlier in the market.
3. Home just constructed and sold sales type and Home was not completed when last assessed (associated with New Homes) sales condition are sold more.



KEY OBSERVATIONS:

1. We can see more of a negative correlation in the with the target variable.
2. WoodsDeckSF and OpenPorchSF have high correlation with sales prize

From all the above analysis we can clearly chart the following points.

Data Inputs- Logic- Output Relationships

MSSubClass 60 – [2-STORY 1946 & NEWER] and **70- [2-STORY 1945 & OLDER]** is the highest segment of building that is sold in the marker which means buyers mostly wish to buy these dwelling in the market, FV is Floating Village Residential which is being highly sold and RL- Residential Low Density being the costliest in the market. We can understand that low residential density which might be of more posh residential area which are costlier in the market across all the other classifications of residents. **IR1 - Slightly irregular** being the costliest lot shape followed by **Reg – Regular**. **IR2 - Moderately Irregular** are the highest sold lot shape. So, from above we can understand people are mostly interested in buying irregular shaped lot more than the regular shaped lot and since people are not buying the regular shaped lot the cost of the regular shaped lot is lesser than slightly irregular plot.

And also, the availability of the Regular shaped plot is lesser that might be also one of the reasons to note. We can see one family type building being the costliest in the market and also getting sold higher. **TwnhsE - Townhouse** End Unit is second highest sold Building Type. According to the **Neighborhood NoRidge** that is Northridge is being sold high and also costliest in the market**. NridgHt - Northridge Heights** is the next costliest sold property with respect to neighbourhood. Since, Utilities have only one values in all the columns it has no

correlation. we will drop this column since it won't help in building the model. The table have more positive correlation at the bottom. Overall Quality **yearbuilt** year remodified have high corelation with sales prize.

Roof style is Gable which is being the costliest in the type of roof style followed by hip. But shed being the highest sold roof type. From above we can analysis that the Gable roof style is costlier so mostly people prefer to buy shed roof type.

Roof material **Standard (Composite) Shingle** being the costliest but soled comparatively lesser than Wood Shingles. Wood Shingles have many varieties of cost and it has been also sold higher and costlier than the other roofing materials. We can observe the feature variable have lesser correlation among themselves but they have high correlation with target variable. Exterior1st and Exterior2nd have high correlation with themselves to avoid multicollinearity we will drop Exterior2nd. The feature variables have high positive as well as negative corelations as we know that positive corelation increases the price of the property and the negative correlation will reduce the price of the property.

SBrkr is Standard Circuit Breakers & Romex and also this is considerably safer than any other electrical circuits in industry. Standard Circuit Breakers & Romex is getting soled higher and also the costliest across all the electrical systems. From above we can say that properties are built as with more safety and more reliable electrical equipment's over other. There was a saying the Quality of the kitchen is the beauty of the house, as similar to that we can see the excellent quality in kitchen will increase the cost of the property. And also, the excellent quality of kitchens is being mostly build. Good Quality in kitchen stands second in the order and also in number of units sold. From above we can narrate a story that people mostly preferred good quality kitchens and also good and excellent quality of kitchens are being costlier.

We can observe according to the correlations the scatterplot points are distributed. We can see high positive correlation of feature variables with the sales prize. We can see Garage Area, garage cars, garage area built, Fireplace, total rooms available, full bath, living area, 1stFlrSf are more positive corelated which means the increase in the above will also increase the cost and selling price of the property. **KitchenQual, Garagetype, GarageFinish** are negatively correlated and the increase in that will decrease the cost and the selling price of the property. We also can see there are multiple variable those have more correlation among themselves than the target variable these variables will create the multi collinearity problem. To overcome those problem, I am dropping one of those variables. **Garagecars, TotalrmsAbvGround** can be removed to avoid multicollinearity. From above three categories the year sold, **salestype** and Sales condition year sold at **2007** with sale deed type Warranty Deed - Conventional Home was not completed when last assessed is costlier in market. Year build is 2007 Warranty Deed – Conventional and Normal Sale is costlier in the market. Home just constructed and sold sales type and Home was not completed when last assessed (associated with New Homes) sales condition are sold more. We can see more of a negative correlation in the with the target variable. **WoodsDeckSF** and **OpenPorchSF** have high correlation with sales prize.

Creating Pre-Processing Pipeline Function:

With all above collected information I have built a model to pre-process and power transform our data without much loss in data records and also in future same function can be also used in testing data to pre-process the testing data.

So before creating this function I have split the data into target and feature variable as follows.

In [66]:
```
x_1=DF_train.drop(["SalePrice"], axis = 1)
y_1=DF_train.SalePrice
```

The pre-processing function are as follows.

In [67]:
```
def Data_Preparation(df): ### Creating function for Preprocessing the data.
    """The Function that returns the preprocessed Data to train"""

    ### Droping the unwanted data----------(1)
    df.drop(['Id', 'Utilities', 'Exterior2nd', 'GarageCars', 'TotRmsAbvGrd'], axis=1, inplace = True)

    ### Imputing the missing Values----------(2)
    df['Alley'].replace(np.nan, 'No_access', inplace = True)

    from sklearn.experimental import enable_iterative_imputer
    from sklearn.impute import IterativeImputer
    imp = IterativeImputer(max_iter=10, random_state=0)
    imp.fit(df_train1[['LotArea','LotFrontage']])
    df[['LotArea','LotFrontage']] = np.round(imp.transform(df[['LotArea','LotFrontage']]))

    for i in ['BsmtQual', 'BsmtCond', 'BsmtExposure','BsmtFinType1','BsmtFinType2']:
        df[i].replace(np.nan, 'No_Basement', inplace = True)

    df['MasVnrType'].replace(np.nan,df['MasVnrType'].describe().top, inplace = True)
    df[['MasVnrArea', 'TotalBsmtSF']] = np.round(imp.fit_transform(df[['MasVnrArea', 'TotalBsmtSF']]))

    for i in ['GarageType', 'GarageFinish','GarageQual','GarageCond']:
        df[i].replace(np.nan, 'No_Garage', inplace = True)

    df['GarageYrBlt'].replace(np.nan, 0, inplace = True)
    df['FireplaceQu'].replace(np.nan, 'No_Fireplace', inplace = True)
    df['Electrical'].replace(np.nan,df['Electrical'].describe().top, inplace = True)

    df['PoolQC'].replace(np.nan, 'No_Pool', inplace = True)
    df['Fence'].replace(np.nan, 'No_Fence', inplace = True)
    df['MiscFeature'].replace(np.nan, 'None', inplace = True)

    ###Encoding the Data converting to numerical value----------(3)
    clos = df.columns
    nums_clos = df._get_numeric_data().columns

    Cat_col = list(set(clos)-set(nums_clos))

    from sklearn.preprocessing import LabelEncoder
    le = LabelEncoder()

    for i in Cat_col:
        df[i] = le.fit_transform(df[i])

    ###power_transform the Data to reduce the skewness and outliers----------(4)
    from sklearn.preprocessing import power_transform
    df=power_transform(df,method="yeo-johnson")
    df

    return df
```

I have followed four steps in the pre-processing function (Data_preparation) these steps are as follows.

1. Dropping the unwanted columns or feature variables as per our observation ('Id', 'Utilities', 'Exterior2nd', 'GarageCars', 'TotRmsAbvGrd')
2. Imputing all the missing values.
3. Converting all the categorical columns into numerical values with the help of label Encoder.
4. Finally Transforming the data into small vectors with Power Transformation technique.

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

Since 2013, real estate has ranked as the top investment pick for the majority (35%) of Americans, according to Gallup's annual Economy and Personal Finance survey, conducted in early April 2020. That puts real estate ahead of stocks and mutual funds (21%), savings accounts (17%), gold (16%), and bonds (8%) as the most favoured investment.

It may be the top investment pick, but is real estate investing safe? Just like any investment, real estate investing has risks. Here are seven real estate investment risks to watch out for when you're thinking about buying an investment property.

The Real Estate Market Is Unpredictable

Leading up to the 2008 Great Recession, many investors (wrongly) believed the real estate market could only move in one direction—up. The basic assumption was that if you bought a property today, you could sell it for a lot more later on.

While real estate values do tend to rise over time, the real estate market is unpredictable—and your investment could depreciate. Supply and demand, the economy, demographics, interest rates, government policies, and unforeseen events all play a role in real estate trends, including prices and rental rates. You can lower the risk of getting caught on the wrong side of a trend through careful research, due diligence, and monitoring of your real estate holdings.

Choosing a Bad Location

The location should always be your first consideration when buying an investment property. After all, you can't move a house to a more desirable neighbourhood—nor can you move a retail building out of an abandoned strip mall.

Location ultimately drives the factors that determine your ability to make a profit—the demand for rental properties, types of properties that are in the highest demand, tenant pool, rental rates, and the potential for appreciation. In general, the best location is the one that will generate the highest return on investment. You have to do some research to find the best locations, however.

(James Chen, 2019)

Model/s Development and Evaluation

Identification of possible problem-solving approaches.

Post splitting the data I have devised loop function to get the best randomstate which gives the highest accuracy as follows.

Selecting parameters for training

In [70]:
```
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score, cross_val_predict, cross_validate
from sklearn.metrics import r2_score, mean_absolute_error,mean_squared_error

accu = 0
for i in range(0,1000):
    x_train_1, x_test_1, y_train_1, y_test_1 = train_test_split(x_1,y_1,test_size = .25, random_state = i)
    mod = LinearRegression()
    mod.fit(x_train_1,y_train_1)
    y_pred_1 = mod.predict(x_test_1)
    tempacc = r2_score(y_test_1,y_pred_1)
    if tempacc> accu:
        accu= tempacc
        best_rstate=i

print(f"Best Accuracy {accu*100} found on randomstate {best_rstate}")
```

```
Best Accuracy 88.35540389186862 found on randomstate 454
```

In [71]:
```
x_train, x_test, y_train, y_test = train_test_split(x_1,y_1,test_size = .25, random_state = best_rstate, shuffle = True)
```

In [72]:
```
from sklearn.linear_model import LinearRegression, Lasso, Ridge, ElasticNet
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor
```

Testing of Identified Approaches (Algorithms)

As we can see the best accuracy is 88.35 found at 454 randomstate wising which we will split and train our model. I have trained our data in 9 different algorithmic model which are. "LinearRegression", "Lasso", "Ridge", "ElasticNet", "SVR", "KNeighborsRegressor",

	MODEL	SCORE	CV_mean_score	CV_STD	MBE	MSE	RMSE	R2
8	RandomForestRegressor	0.976888	0.848460	0.032119	18323.940959	6.479206e+08	25454.285476	0.890212
7	AdaBoostRegressor	0.874343	0.808786	0.040002	25181.907319	1.089128e+09	33001.945572	0.815451
3	ElasticNet	0.797024	0.776164	0.054707	19202.666335	7.950430e+08	28196.506014	0.865283
2	Ridge	0.817014	0.749972	0.090076	19620.380354	6.950183e+08	26363.200302	0.882232
1	Lasso	0.817019	0.749361	0.090910	19645.371134	6.961904e+08	26385.419354	0.882033
0	LinearRegression	0.816070	0.749324	0.090958	19405.813744	6.872142e+08	26214.771210	0.883554
5	KNeighborsRegressor	0.801678	0.738769	0.019570	24837.043151	1.410387e+09	37555.125053	0.761015
6	DecisionTreeRegressor	1.000000	0.710734	0.061716	28637.479452	1.626216e+09	40326.363958	0.724444
4	SVR	-0.041009	-0.061789	0.050386	57169.267411	6.216638e+09	78845.658062	-0.053387

"DecisionTreeRegressor", "AdaBoostRegressor", "RandomForestRegressor" and with the metrics such as cross_val_score, R2 Score, MEAN ABSOLUTE ERROR, MEAN SQUARED ERROR and ROOT MEAN SQUARED ERROR as follows.

And I have sorted the model with highest Cross Validation score as we can see Random Forest Regressor is the best model with model score of 97% with CV of 84 % further, I have tried to hyper parameter tune the same model for increasing the performance.

Run and evaluate selected models

I have declared a variable as rf to define Random Forest model.

In [78]:
```
rf = RandomForestRegressor(random_state=42)
```

In [79]:
```
rf.fit(x_train, y_train)
```

Out[79]:
```
RandomForestRegressor(random_state=42)
```

In [80]:
```
rf.score(x_test,y_test), rf.score(x_train,y_train)
```

Out[80]:
```
(0.890212193748729, 0.9768879289793189)
```

Let's see all the parameter in random forest regressor.

In [81]:
```
rf.get_params()
```

Out[81]:
```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'mse',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 42,
 'verbose': 0,
 'warm_start': False}
```

Selecting parameters bootstrap, criterion, n_estimators, max_features, min_samples_split from above I have tried to tune the model as follows.

HYPERPARAMETER TUNING....

In [82]:
```
param_grid = [
    {"bootstrap": [True, False],
     "criterion": ["mse", "mae"],
     "n_estimators"      : [10,20,30,50,100],
     "max_features"      : ["auto", "sqrt", "log2"],
     "min_samples_split" : [2,4,8],
    }
]

rf = RandomForestRegressor(random_state=42)

grid_search = GridSearchCV(estimator = rf, param_grid = param_grid,
                           cv = 5, n_jobs = -1, verbose = 2)
```

```
In [83]: grid_search.fit(x_train,y_train)

         Fitting 5 folds for each of 180 candidates, totalling 900 fits

         [Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
         [Parallel(n_jobs=-1)]: Done  33 tasks      | elapsed:   17.2s
         [Parallel(n_jobs=-1)]: Done 164 tasks      | elapsed:   25.1s
         [Parallel(n_jobs=-1)]: Done 409 tasks      | elapsed:  2.3min
         [Parallel(n_jobs=-1)]: Done 706 tasks      | elapsed:  3.8min
         [Parallel(n_jobs=-1)]: Done 900 out of 900 | elapsed:  6.5min finished

Out[83]: GridSearchCV(cv=5, estimator=RandomForestRegressor(random_state=42), n_jobs=-1,
                      param_grid=[{'bootstrap': [True, False],
                                   'criterion': ['mse', 'mae'],
                                   'max_features': ['auto', 'sqrt', 'log2'],
                                   'min_samples_split': [2, 4, 8],
                                   'n_estimators': [10, 20, 30, 50, 100]}],
                      verbose=2)
```

As like above the best params for Random Forest Regressor is sorted by Grid Search CV and seeing the best CV score and model score I have fitted the data to the model.

```
In [84]: grid_search.best_score_

Out[84]: 0.8504290189576658

In [85]: grid_search.best_params_

Out[85]: {'bootstrap': False,
          'criterion': 'mae',
          'max_features': 'sqrt',
          'min_samples_split': 2,
          'n_estimators': 50}

In [86]: grid_search.best_estimator_

Out[86]: RandomForestRegressor(bootstrap=False, criterion='mae', max_features='sqrt',
                               n_estimators=50, random_state=42)

In [87]: reg_final_model = grid_search.best_estimator_

In [88]: reg_final_model.fit(x_train,y_train)

Out[88]: RandomForestRegressor(bootstrap=False, criterion='mae', max_features='sqrt',
                               n_estimators=50, random_state=42)

In [89]: reg_final_model.score(x_train,y_train), reg_final_model.score(x_test,y_test)

Out[89]: (0.9999994772684867, 0.8934153259916571)
```

We can see the increase in CV score to 85% and also the model score increased from 97% to 99.99% further am saving the same model for future usage.

Saving the model:

```
In [90]: import joblib
         joblib.dump(reg_final_model,"HOUSE-PRICEPREDICTION.obj")

Out[90]: ['HOUSE-PRICEPREDICTION.obj']
```

Then as per the “Surprise Housing” client’s requirement I have imported test data which was also given to us in CSV format I have pre-processed and predicted the output (Sales Price) which again I have declared as Data Frame and saved in CSV format for client usage.

Key Metrics for success in solving problem under consideration

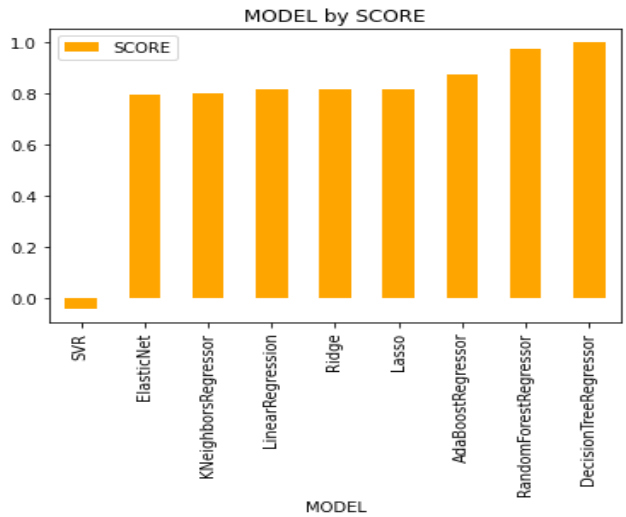
As we saw above the metrics that I used to evaluate the model is Cross validation score and R2 Score and MEAN ABSOLUTE ERROR, MEAN SQUARED ERROR and ROOT MEAN SQUARED ERROR, have trained the data with 7 different algorithmic model and sorted with highest CV score to be the best model.

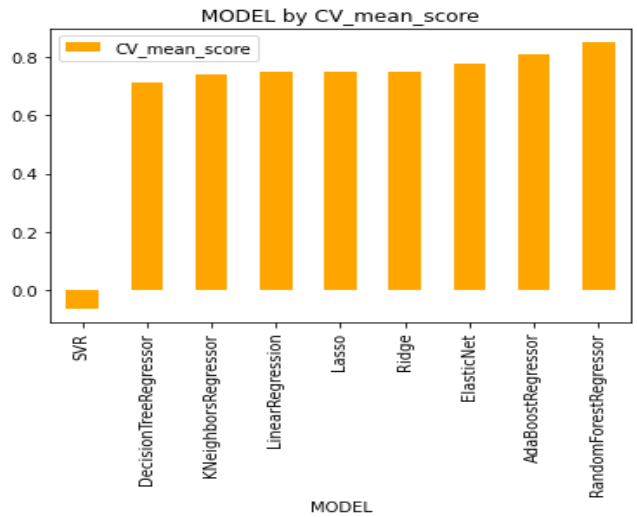
	MODEL	SCORE	CV_mean_score	CV_STD	MBE	MSE	RMSE	R2
8	RandomForestRegressor	0.976888	0.848460	0.032119	18323.940959	6.479206e+08	25454.285476	0.890212
7	AdaBoostRegressor	0.874343	0.808786	0.040002	25181.907319	1.089128e+09	33001.945572	0.815451
3	ElasticNet	0.797024	0.776164	0.054707	19202.666335	7.950430e+08	28196.506014	0.865283
2	Ridge	0.817014	0.749972	0.090076	19620.380354	6.950183e+08	26363.200302	0.882232
1	Lasso	0.817019	0.749361	0.090910	19645.371134	6.961904e+08	26385.419354	0.882033
0	LinearRegression	0.816070	0.749324	0.090958	19405.813744	6.872142e+08	26214.771210	0.883554
5	KNeighborsRegressor	0.801678	0.738769	0.019570	24837.043151	1.410387e+09	37555.125053	0.761015
6	DecisionTreeRegressor	1.000000	0.710734	0.061716	28637.479452	1.626216e+09	40326.363958	0.724444
4	SVR	-0.041009	-0.061789	0.050386	57169.267411	6.216638e+09	78845.658062	-0.053387

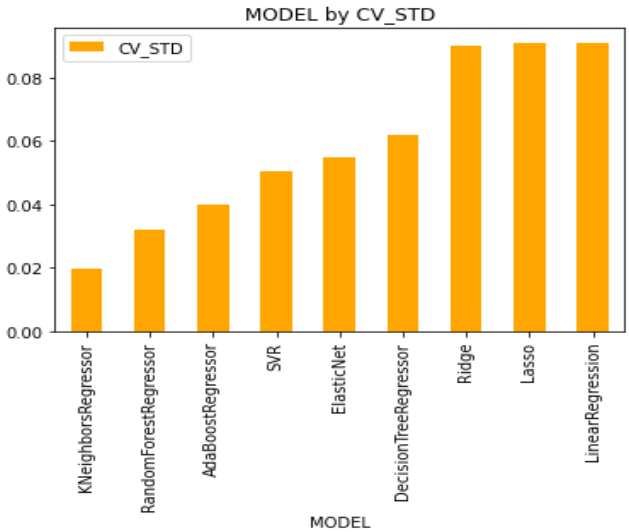
We can see two model scoring high in the chart Random Forest regressor and Decision tree model. Random forest regressor with model score of 97% but the Decision tree Regressor have the model score of 100%. However, we can see that in Decision Tree Regressor the CV score is only 71% which clearly says the Decision tree is over fitting with the data, on tuning with correct parameters the Decision Tree Regressor scores will reduce, but in the Random Forest regressor the CV score is 84.84% which is considerably better than Decision Tree Regressor. So, basis on the same metrics I am choosing Random Forest Regressor as best across our models. Further we are going to visualize the performance of these models.

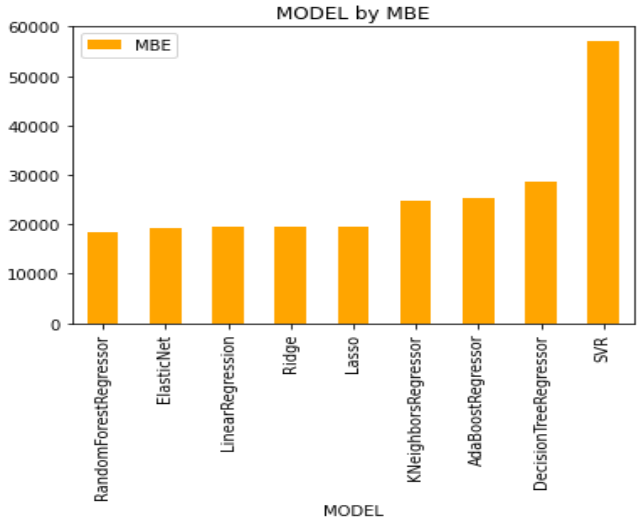
Visualizations

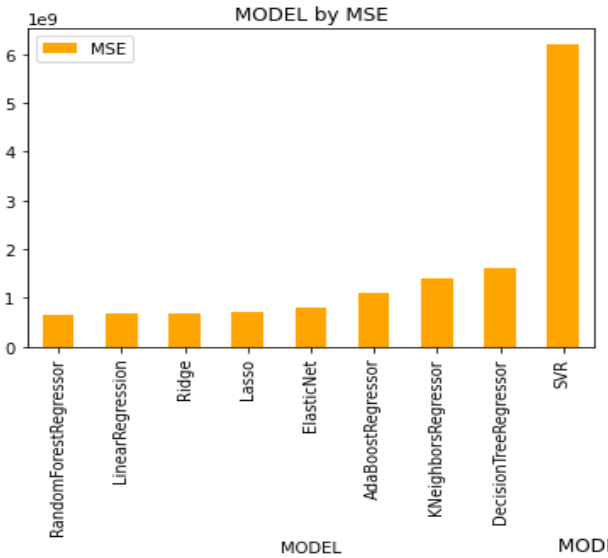
We have seen the mathematical expression of the performance of the model now we are going to visualize the same in bar plot.

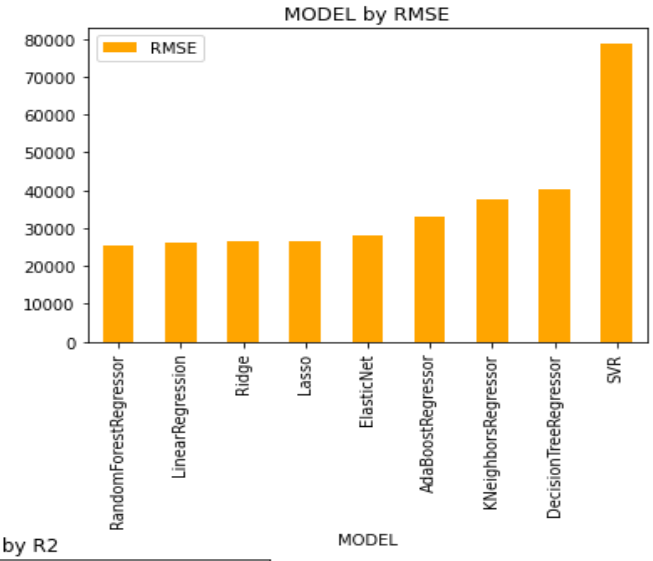


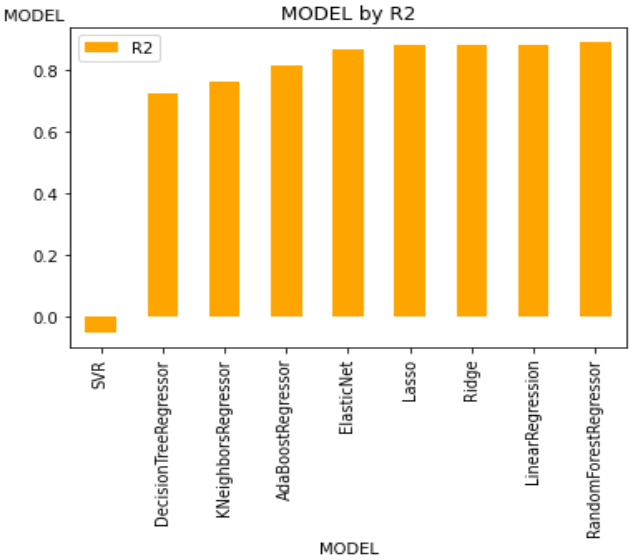












From above we can see that Random Forest regressor and Decision tree regressor have high scores of 97% and 100% but cross validation average in Random Forest is 84% which is higher than 70% of Decision tree regressor. It is evident that Random Forest is the best model with Score of 0.97 let's try in Hyper tuning the same for improved performance

Interpretation of the Results

Then as per the "Surprise Housing" client's requirement I have imported test data which was also given to us in CSV format I have pre-processed and predicted the output (Sales Price) which again I have declared as Data Frame and saved in CSV format for client usage.

Have denoted 'x_t' as test data variable. And also have passed the data to the pre-processing function that we have already created to make it ready for prediction.

Lets import and clean Test data.

```
In [91]: DF_test= pd.read_csv("C:/Users/Friday/Downloads/Project-Housing/Project-Housing_splitted/test.csv")

In [92]: x_T = Data_Preparation(DF_test)

In [93]: Price_Prediction= joblib.load('HOUSE-PRICEPREDICTION.obj')
         predi= Price_Prediction.predict(x_T)
         prediction = [round(i) for i in predi]
```

Post preparing the data for prediction (to run in the model) I have called back our saved model and run the prediction of the data and saved in the variable named prediction.

```
In [94]: Price_Prediction.score(x_T,predi)
Out[94]: 1.0

In [95]: HOUSE_PRICE_Predicted=pd.DataFrame({"HOUSE-PRICE_Predicted":prediction})
         HOUSE_PRICE_Predicted.head()
```

Out[95]:

	HOUSE-PRICE_Predicted
0	317369.0
1	203012.0
2	260685.0
3	172166.0
4	236337.0

Post that I have created the data frame and stored our variable and also saved our data frame in csv format.

```
In [96]: HOUSE_PRICE_Predicted.describe()
```

Out[96]:

	HOUSE-PRICE_Predicted
count	292.000000
mean	180570.078767
std	64863.243561
min	82090.000000
25%	135660.000000
50%	167074.000000
75%	207489.750000
max	415144.000000

```
In [97]: HOUSE_PRICE_Predicted.to_csv("HOUSE-PRICE_Predicted.csv") #we are saving our model in CSV format.
```

We can observe the data description of the predicted data and also, I have tried in evaluating the prediction with the model and the score of the test data comes as 100%.

CONCLUSION

Key Findings and Conclusions of the Study

The benefits of investing in real estate are numerous. With well-chosen assets, investors can enjoy predictable cash flow, excellent returns, tax advantages, and diversification—and it's possible to leverage real estate to build wealth.

Thinking about investing in real estate? Here's what you need to know about real estate benefits and why real estate is considered a good investment as per our above study.

Key Findings.

1. What are the types of the building that are build most along with the foundation type and basement quality?
2. What is the most preferred type of neighbourhood.?
3. What are the other amenities that increases the cost of property.?
4. Which electrical systems are more preferred on the building?
5. What is Quality and the type of Fence, Garage, Heating systems that is mostly preferred.?
6. How many months old building sales type and sales conditions that increases the cost.?

Building types and attachments:

As from the analysis we can surely say **that 2-STORY 1946 & NEWER, 2-STORY 1945 & OLDER** are the highly preferred dwellings with preferred near road access on **Paved** type roads. And also because of the lot availability slightly irregular lots are mostly preferred by the builders and also by the people. Lot Configuration the lot which is inside is slightly cheaper than the corner lot with **gentle slope**. These are the basic built type that comes out for sales more.

Builders and people mostly preferred a very strong foundation with a good quality foundation that are built with **Stones** and **Poured Concrete** are preferred more than the **Cinder Block** or **Brick & Tile.** And mostly depending on the construct of the building the cost of the foundation is also decided. Basement with **Excellent** (100+ inches) quality and also with Typical - slight dampness condition is sold most.

Neighbourhood

There are 24 variety of neighbourhood associated with the property mostly. But most preferred and costliest is **Northridge** and **Northpark Villa** type neighbourhood property with low density of people living are being costlier. The type of neighbourhood will also define the safety, security and other immediate needs that can be fulfilled.

Amenities.

At least fifty presents of the property have more amenities which increases the cost of the property amenities like **Tennis Court** is being the costliest and there are also other amenities such as Swimming pool, shed, elevator, 2ed garage.

Electric Systems:

Electric Systems plays a vital role in the safety of the houses and most preferred electrical systems are Standard Circuit Breakers & Romex that is mostly preferred and also being sold as costliest in the property.

Security Privacy heating systems and Garage models.

The above three are the most important factors on which the sales of the property are decided mostly. Fence with Minimum Wood/Wire are mostly preferred but good privacy fence is being the costliest off all. The type of heating systems Gas forced warm air furnace is highly preferred reason being the conception of electricity is lesser than any other furnace and quality and working condition being good on the same. Property which is build with garages are also increases the cost of the property garages with 2 cars parking are mostly prepared but two garages or garages with multiple car parking are the costliest garage.

Sales type and Sales conditions of the property and Month Sold.

To build a property and sell it to the preferred customer it takes the marketing team at least of 9 – 10 months as per our above analysis so on a average the most property sold at more than 9 months but built lesser than 1 month are being the costliest. Warranty Deed – Conventional with Home was not completed when last assessed (associated with New Homes) are being the costliest and normal sales conditioned property are sold higher.

Learning Outcomes of the Study in respect of Data Science.

- The above study helps one to understand the business of real estate. How the price is changing across the Properties.
- With the Study we can tell how multiple real estate amenities like swimming pool, garage, pavement and lawn size of Lot Area, and type of Building raise decides the cost.
- With the help of the above analysis, one can sketch the needs of a property buyer and according to need we can project the price of the property.

Limitations of this work and Scope for Future Work

The real estate industry is likely just at the beginning of a significant shift towards greater use of data and data-driven decision making. There are huge opportunities that are now starting to be unlocked by various start-ups and forward-thinking institutions. There is a range of concrete methods — as outlined above — to apply data science to real estate, to help move from millions of rows of data to granular understandings of past, present, and future real estate submarket performance, and make superior investment and business decisions. **However**, the required skills may often be absent across a good percentage of the industry. There is now the opportunity to learn these techniques and methods — specifically for real estate — and investing the time to upgrade could benefit a range of participants. Real estate researchers could begin to use data and machine learning to produce game-changing insights and unlock the value of large datasets. Those in the Protect industry (or even investing in Protect) could do well to understand these methods better and build (or invest in) disruptive activities. Finally, real estate investors who learn these methods could use data-driven approaches to find exceptional opportunities and beat the market.