Project Report on

Housing Price Prediction

Submitted by:

Sankalp Mahapatra

Internship-29

ACKNOWLEDGMENT

I would like to express my sincere thanks of gratitude to my mentors from Data Trained academy and FlipRobo Technologies Bangalore for letting me work on this project. Their suggestions and directions have helped me in the completion of this project successfully. All the required information & the dataset are provided by FlipRobo Technologies.

Finally, I would like to thank my family and friends who have helped me with their valuable suggestions and guidance and have been very helpful in various stages of project completion.

TABLE OF CONTENTS

1. INTRODUCTION

Thousands of houses are sold every day. There are some questions every buyer asks himself like: What is the actual price that this house deserves? Am I paying a fair price? Also is it the location? Is it the overall quality of the house? Is it the size? Could it be sold at a good price in future? All these questions come in to our mind when we decide to purchase a house.

In this study, a machine learning model is proposed to predict a house price based on data related to the house (its size, the year it was built in, etc.). During the development and evaluation of our model, we will show the code used for each step followed by its output. This will facilitate the reproducibility of our work.

1.1 Business Problem Framing:

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

The project endeavours to extensive data analysis and implementation of different machine learning techniques in python for having the best model with most important features of a house on insight of both business value and realistic perspective.

Business goal:

With the help of available independent variables, we need to model the price of the houses. 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.

1.2 Conceptual Background of the Domain Problem

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

The problem statement is related to the US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company is looking at prospective properties to buy houses to enter the market. It is 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 company wants to know:

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

In this section, we evaluate widely used regression technologies like Linear Regression, regularization, bagging and boosting and many more ensemble techniques to predict the house sale price result.

1.3 Review of Literature

The relationship between house prices and the economy is an important motivating factor for predicting house prices (Pow, Janulewicz, & Liu, 2014). There is no accurate measure of house prices (Pow, Janulewicz, & Liu, 2014). Pow states that Real Estate property prices are linked with economy (Pow, Janulewicz, & Liu, 2014). He also states there is no accurate measure of house prices. A property's value is important in real estate transactions. Pow tries to predict the sold and asking prices of real estate values without bias to help both buyers and sellers make their decisions. A property's value is important in real estate transactions.

Housing market is important for economic activities (Khamis & Kamarudin, 2014). Traditional housing price prediction is based on cost and sale price comparison. So, there is a need for building a model to efficiently predict the house price.

Based on the sample data provided to us from our client database where we have understood that the company is looking at prospective properties to buy houses to enter the market. The data set explains it is a regression problem as we need 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.
House prices trends are not only the concerns for buyers and sellers, but they also indicate the current economic situations. Therefore, it is important to predict the house prices without bias to help both buyers and sellers make their decisions.

1.3 Motivation for the Problem Undertaken

I have gone through many projects before, but this project has given me an idea to handle large number of attributes. By doing this project I have got an idea about how to deal with data exploration where I have used all my analyzation skills to predict the house price using ML models. The model will be a good way for both buyers and sellers to understand the pricing dynamic of a new market.

The main objectives of this study are as follows:

- To apply data pre-processing and preparation techniques in order to obtain clean data.
- To build machine learning models able to predict house price based on house features.
- To analyse and compare models' performance in order to choose the best model.

By processing the above objects, I will be able to find which variables are important to predict the price of house? And how do these variables describe the price of the house? The relation between house prices and the economy is an important factor for predicting house prices.

2. ANALYTICAL PROBLEM FRAMING

2.1 Mathematical/ Analytical Modelling of the Problem:

The house price model is based on a demand function for housing services and a standard life-cycle model of utility for a representative household. This is a common approach in academic research into house prices. The study is to predict the sale price of the house and analysing which features are important and how they contribute in the prediction.

There are two datasets. One is train dataset which is supervised and another one is test dataset which is unsupervised. The target variable is “SalePrice” and it is a regression type problem. I have used train dataset to build machine learning models and then by using this model I made prediction for the test dataset.

I have observed some columns having more than 85% of zero entries and 70% of null values so, I decided to drop those columns. I have analysed the categorical and numerical features using categorical plots and numerical plots respectively to get better insights from the data.

In this project I have done various mathematical and statistical analysis such as describing the statistical summary of the columns, feature engineering, treating null values, removing outliers, skewness, encoding the data etc. Checked for correlation between the features and visualized it using heat map. Also, I built many regression algorithms while building machine learning models, used hyper tuning method for best model and saved the best model. Finally, I predicted the sale price of the house using the saved trained model.

2.2 Data Sources and their formats

A US-based housing company named Surprise Housing has collected the dataset from the sale of houses in Australia and the data is provided by Flip Robo Company and it is in csv format. There are 2 data sets:

1. Train dataset
2. Test dataset

- Train dataset will be used for training the machine learning models. The dataset contains 1168 rows and 81 columns, out of 81 columns, 80 are independent variables and remaining 1 is dependent variable (SalePrice).
- Test dataset contains all the independent variables, but not the target variable. We will apply the trained model to predict the target variable for the test data. The dataset contains 292 rows and 80 columns.
- The dataset contains both numerical and categorical data. Numerical data contains both continuous and discrete variables and categorical data contains both nominal and ordinal variables.
- I can keep both the train and test datasets in one data frame but this may cause data leakage so I have decided to process both the data separately.

2.3 Data Pre-processing Done

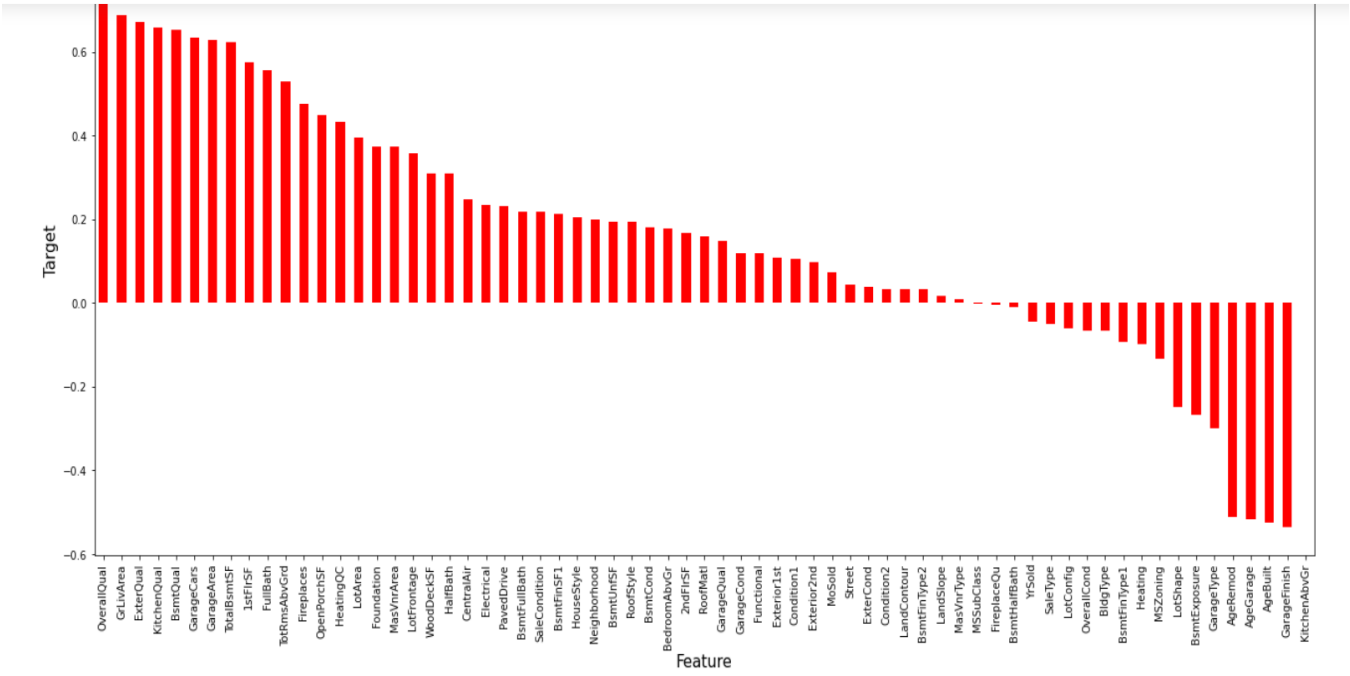
- Firstly, I have imported the necessary libraries and imported both train and test datasets which were in csv format. And process both datasets simultaneously.
- I have done some statistical analysis like checking shape, nunique, column names, data types of the features, info about the features, value counts etc. for both train and test data.
- I have dropped the columns “Id” and “Utilities” from both the datasets. Since Id is the unique identifier which contains unique value throughout the data also all the entries in Utilities column were unique. They had no significance impact on the prediction.
- While looking into the value count function I found some of the columns having more than 85% of zero values so, I dropped those columns from both the datasets as they might create skewness which will impact my model.
- Also, I have done some feature extraction as the datasets contained some time variables like YearBuilt, YearRemodAdd, GarageYrBlt and YrSold. Converting them into age seems more meaningful as they offer more information about the longevity of the features. It is analogous to the fact that, the statement “Mr. X died at the age of 66 years” holds more information for us than the statement “Mr. X died in the year 2019”. So, I have extracted age information from the date time variables by taking the difference in year between the year the house was built and year the house was sold and dropped the year columns.

- I checked the null values and found them in some of the columns. So, I imputed null values present in categorical and numerical columns using mode and mean methods respectively. I found some columns having more than 80% of null values so, I dropped those columns to overcome with the skewness.
- Described statistical summary of both train and test datasets.
- Visualized each feature using seaborn and matplotlib libraries by plotting several categorical and numerical plots.
- Identified outliers using box plots in both datasets. I tried to remove them using both Zscore and IQR method and got huge data loss of around 19% and 35% respectively, so removed outliers using percentile method by setting data loss to 2%.
- Checked for skewness and removed skewness in numerical columns using power transformation method (yeo-johnson). Also dropped KitchenAbvGr columns as it contains zero values throughout the data after using power transformation.
- Encoded both train and test data frames using Ordinal Encoder. Also replaced some categorical columns having ratings by numbers based on specific condition.
- Used Pearson's correlation coefficient to check the correlation between label and features.
- While checking the correlation I came across multicollinearity problem, I checked VIF values and removed GrLivArea to overcome with the multicollinearity issue.
- Scaled both the datasets using Standard Scalar method and used regression algorithms to build ML models.
- All these steps were performed to both train and test datasets simultaneously.

2.4 Data Inputs- Logic- Output Relationships

- To analyse the relation between features and target I have done EDA where I analysed the relation using many plots like bar plot, reg plot, scatter plot, line plot, swarm plot, strip plot, violin plot etc. And found some of the columns like OverallQual, TotalRmsAbvGrd, FullBath, GarageCars etc have strong positive linear relation with the label.

- I have checked the correlation between the target and features using heat map and bar plot. Where I got the positive and negative correlation between the label and feature .



Features having positive and negative impact on the label.

Features having high Positive correlation with label

- OverallQual
- GrLivArea
- ExterQual
- KitchenQual
- BsmtQual
- GarageCars
- GarageArea
- TotalBsmtSF
- 1stFlrSF
- FullBath
- TotRmsAbvGrd

Features having high Negative correlation with label

- Heating
- MSZoning
- LotShape

- BsmtExposure
- GarageType
- AgeRemod
- AgeGarage
- AgeBuilt
- GarageFinish

2.5 Hardware & Software Requirements & Tools Used

To build the machine learning models we need to have the following hardware and software requirements and tools.

Hardware required:

• Processor: core i5 or above

• RAM: 8 GB or above

• ROM/SSD: 250 GB or above

Software required:

• Anaconda 3- language used Python 3

Libraries required:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
import pickle

import warnings
warnings.filterwarnings('ignore')
```

```
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
```

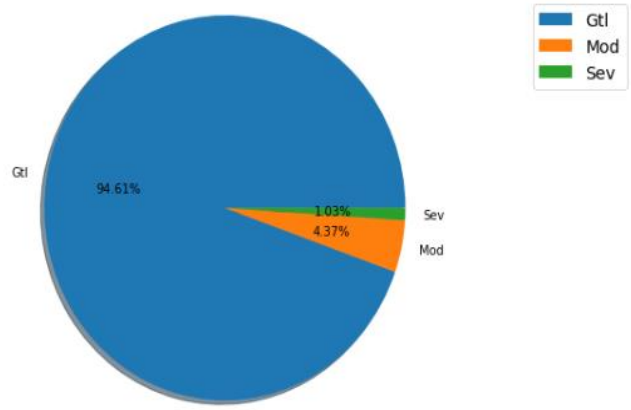
3.MODEL/S DEVELOPMENT AND EVALUATION

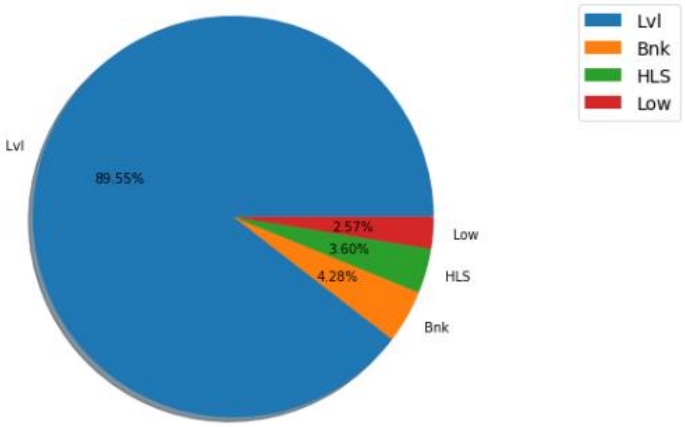
3.1 Identification of possible Problem-solving approaches (Methods):

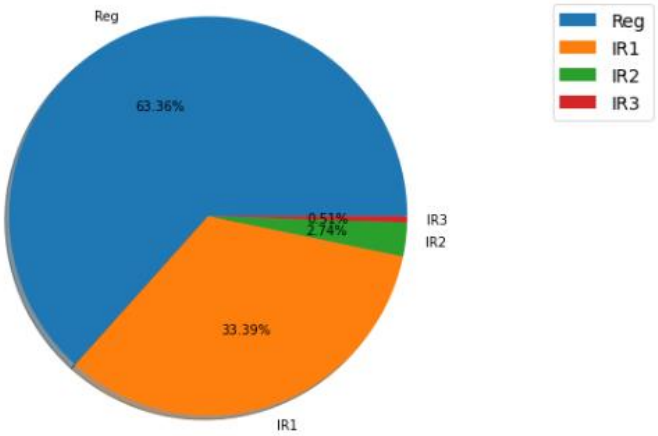
- I have used imputation methods to treat the null values.
- Used percentile method to remove outliers.
- Removed skewness using power transformation (yeo-johnson) method.
- Encoded the object type data into numerical using Ordinal Encoder.
- I have used Pearson's correlation coefficient method to check the
- Correlation between the dependent and independent variables.
- I have scaled the data using Standard Scalar method to overcome with the data biasness.
- Used many machine learning models to predict the sale price of the house.

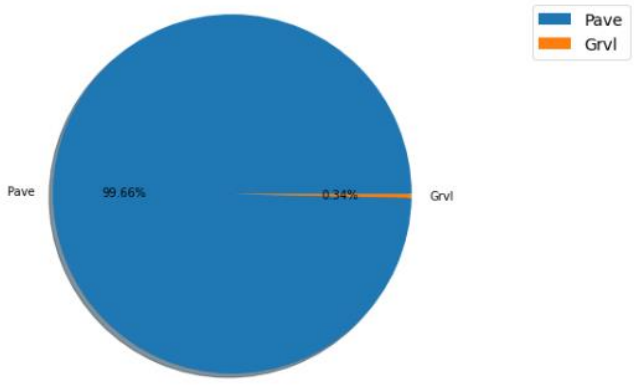
3.2 Visualizations

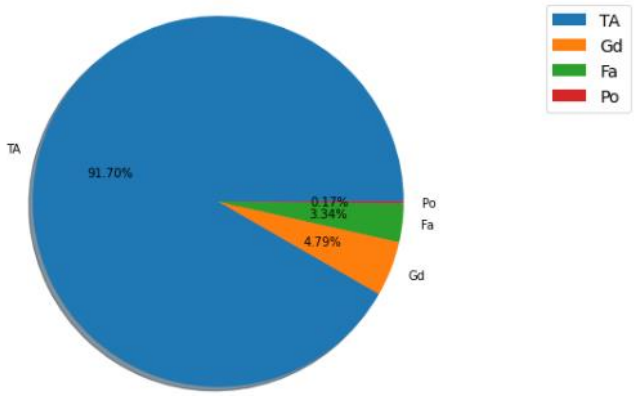
I have analysed the data by plotting the relationship between the features and labels as well as the relationship among the features. I have also used box plots to find the outliers. I have used pie plots, count plots and distribution plot and used reg plots, scatter plots and heatmap plot.. These plots have given good pattern.

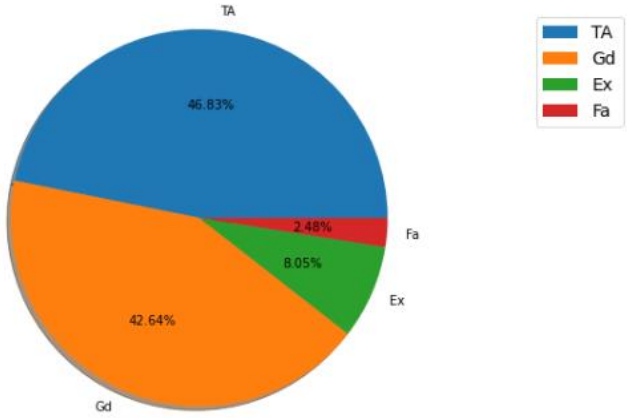

Gtl
Mod
Sev
Gtl
94.61%
1.03%
4.37%
Sev
Mod

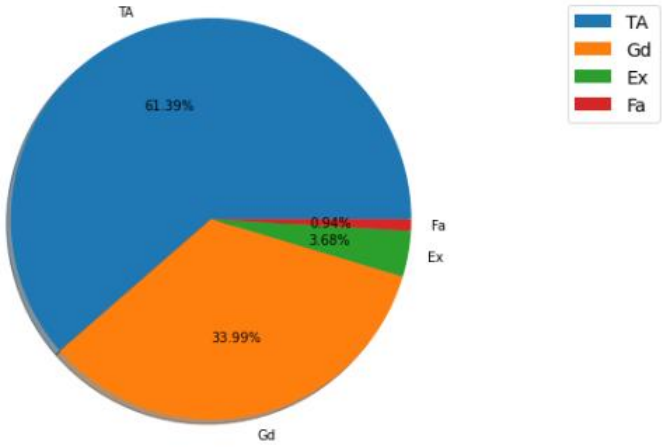

Lvl
Bnk
HLS
Low
Lvl
89.55%
2.57%
3.60%
4.28%
Low
HLS
Bnk

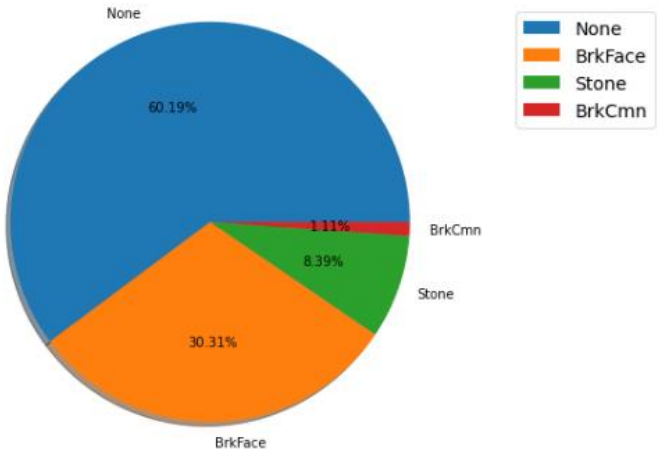

Reg
IR1
IR2
IR3
Reg
63.36%
0.51%
2.74%
33.39%
IR3
IR2
IR1

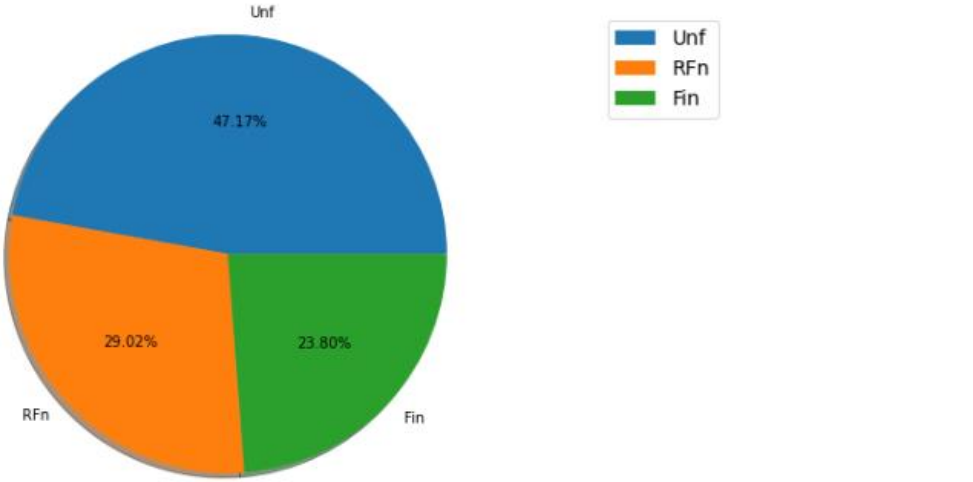

Pave
Grvl
Pave
99.66%
0.34%
Grvl

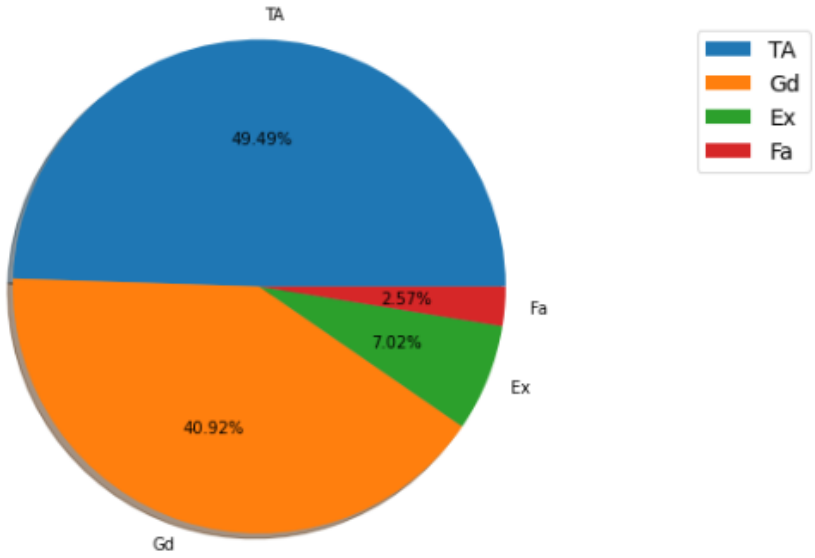

TA
Gd
Fa
Po
TA
91.70%
0.17%
3.34%
4.79%
Po
Fa
Gd

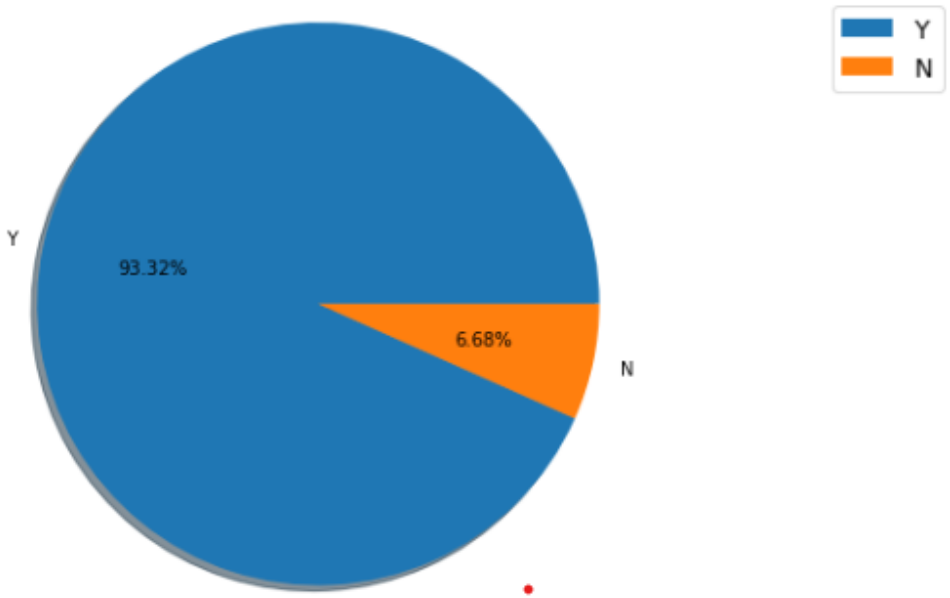

TA
Gd
Ex
Fa
TA
46.83%
2.48%
8.05%
42.64%
Fa
Ex
Gd

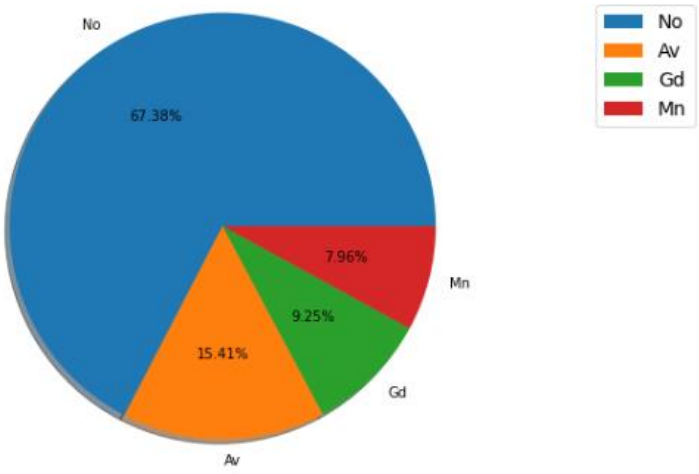

TA
Gd
Ex
Fa
TA
61.39%
0.94%
3.68%
33.99%
Fa
Ex
Gd

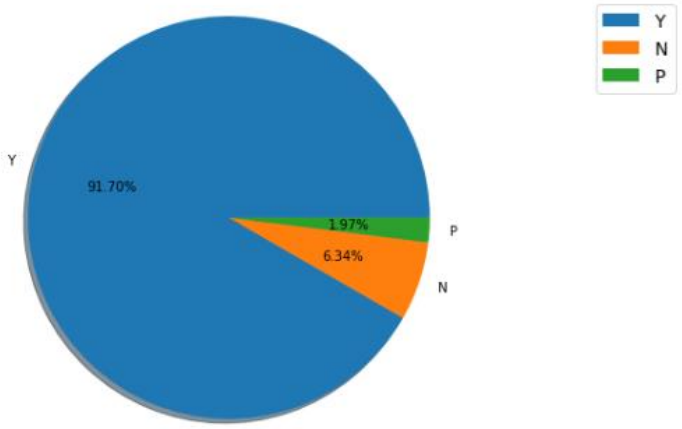

None
BrkFace
Stone
BrkCmn
None
60.19%
1.11%
8.39%
30.31%
BrkCmn
Stone
BrkFace


Unf
47.17%
29.02%
23.80%
RFn
Fin
Unf
RFn
Fin


TA
49.49%
2.57%
7.02%
40.92%
Fa
Ex
Gd
TA
Gd
Ex
Fa


Y
93.32%
6.68%
N
Y
N





Observation from the Pie plots:

- The count of road access to the property Paved is 1164 which covers around 99.66% of the property where Graved type has count 4 that is only 0.34%.

- The count is high for the property having the shape regular.

- The total number of flatness of the property for level is high which has 89.55%.

- The slope of the property Gentle slope has very high count of 1105 i.e, 94.61%.

- Around 60% of the houses do not have Masonry veneer type and 30% of the houses contain Brick Face type of Masonry veneer.
- Around 61% of the houses evaluates typical/average quality of the material on the exterior, 34% of the houses have good quality of the material on the exterior. Only a few have excellent quality.

- Most of the houses evaluate typical/average and good quality of height of the basement.

- Around 91% of the houses have typical/average condition of the basement.

- Around 67% of the houses do not contain any walkout or garden level walls.

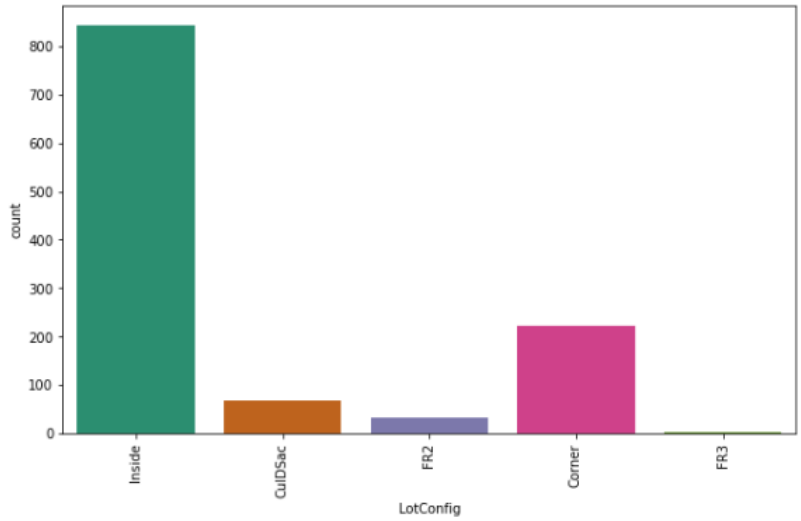
- 93.32% of the houses have central air conditioning.

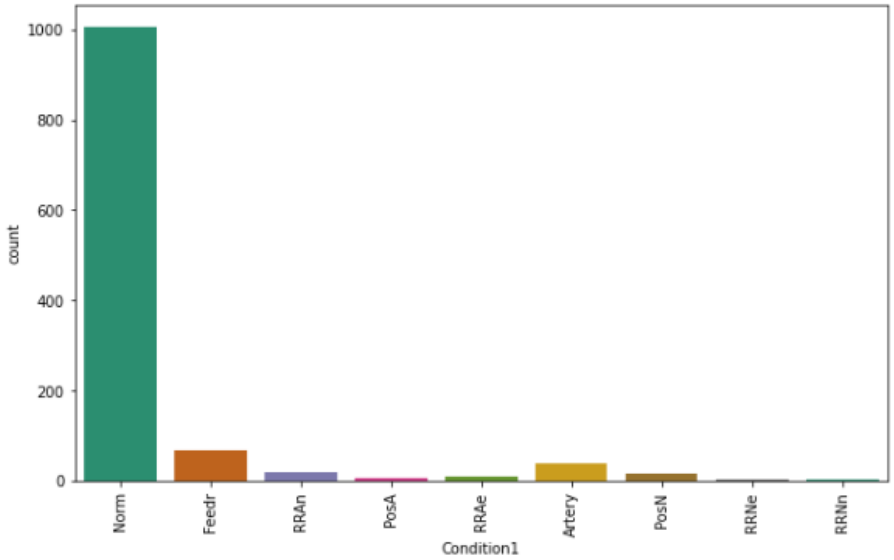
- 49% of the houses contain typical/average kitchen quality and 40% of the houses have good kitchen quality. The count for excellent kitchen quality is very low and is around 2%.

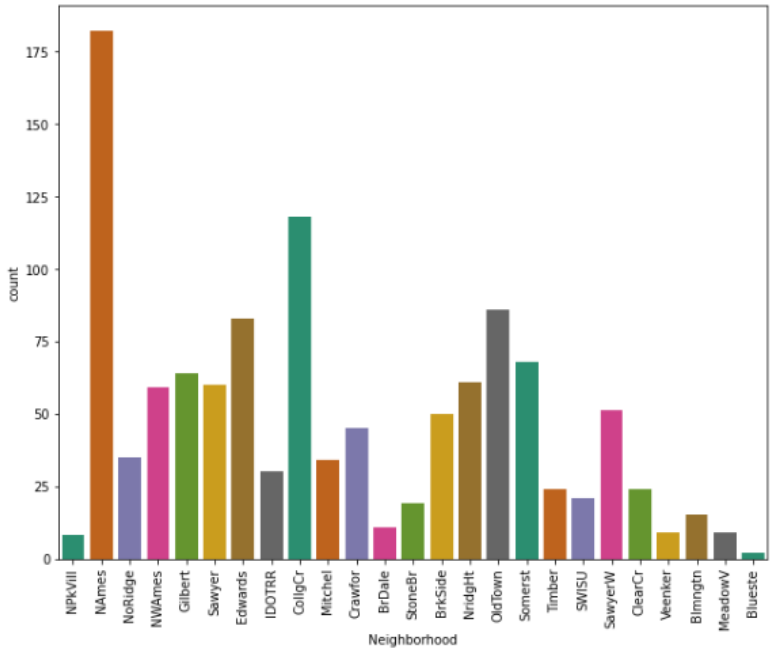
- 47% Of the houses have unfinished garage interior, 29% rough finished and only 23% of the houses' interior garage has finished.

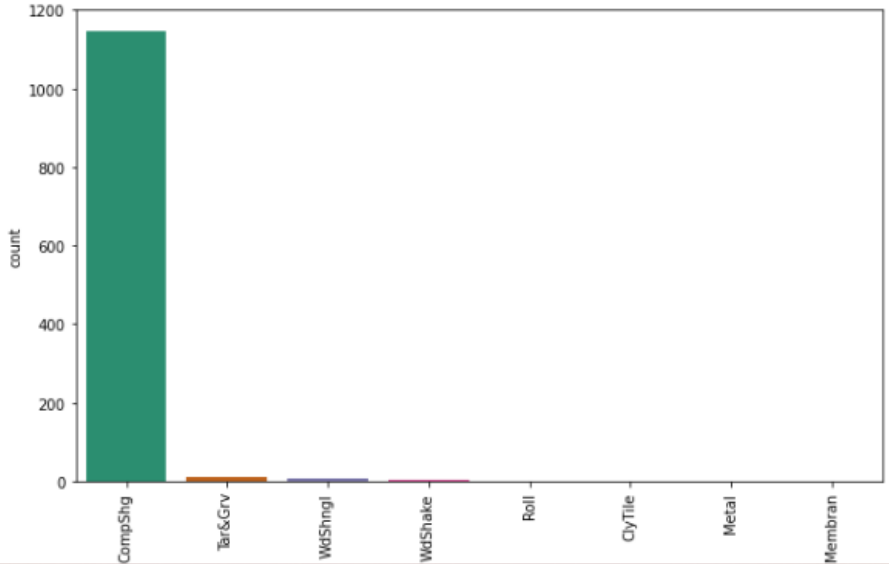
- 91.70% of the houses contains the paved drive way.

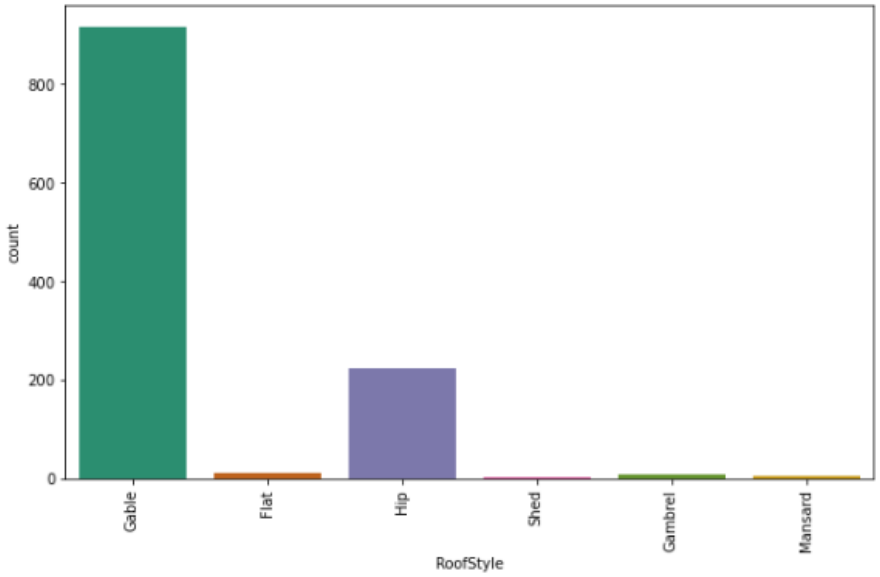
Now let’s plot some more graphs to understand the data better,

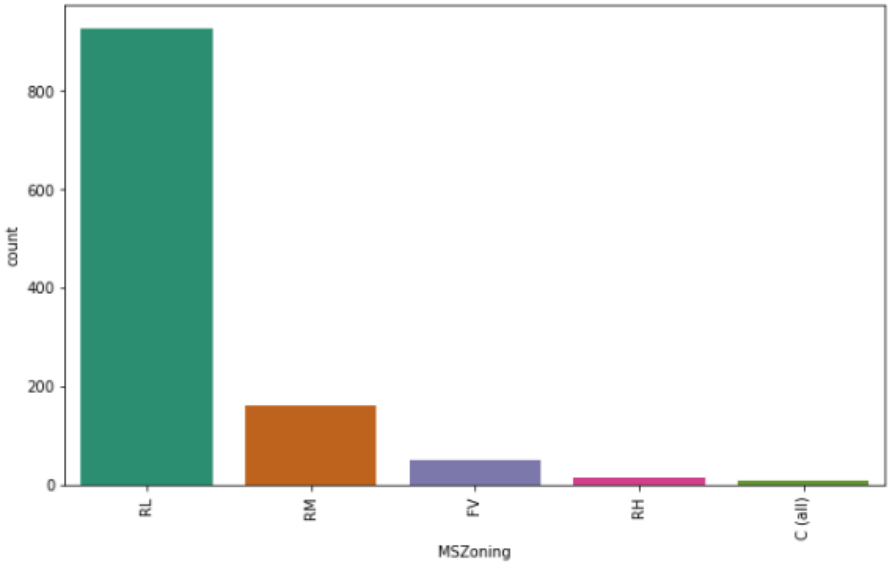

count
800
700
600
500
400
300
200
100
0
Inside
CulDSac
FR2
Corner
FR3
LotConfig

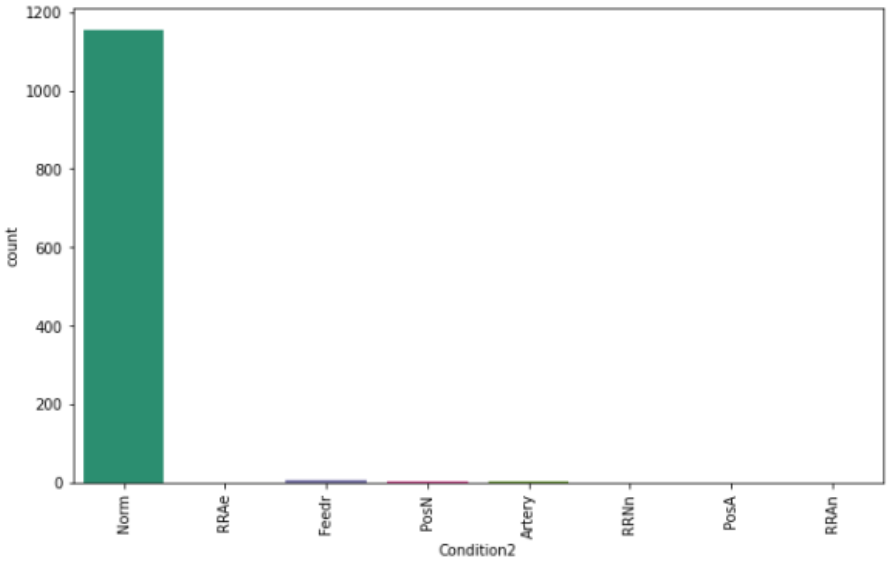

count
1000
800
600
400
200
0
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
Condition1

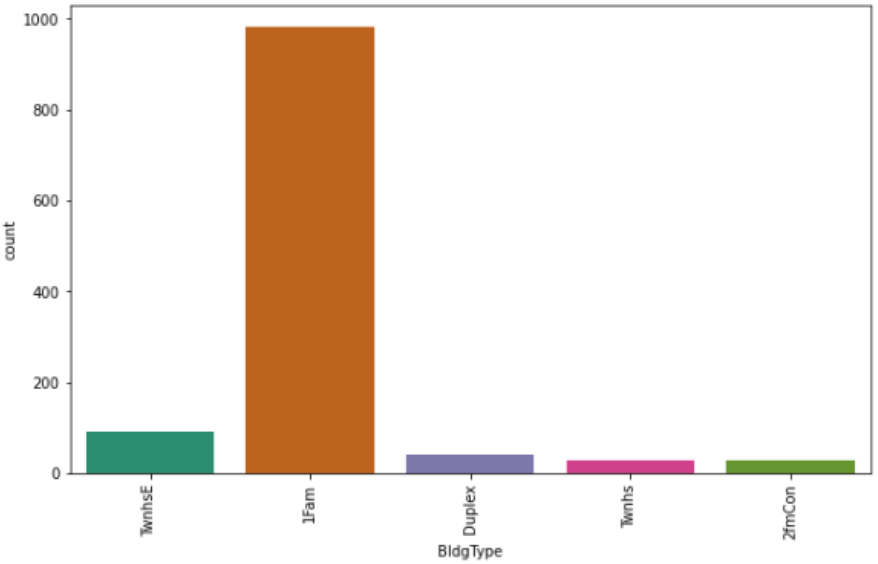

count
175
150
125
100
75
50
25
0
NPkVill
NAmes
NoRidge
NWAmes
Gilbert
Sawyer
Edwards
IDOTRR
CollgCr
Mitchel
Crawfor
BrDale
StoneBr
BrkSide
NridgHt
OldTown
Somerst
Timber
SWISU
SawyerW
ClearCr
Veenker
Blmngtn
MeadowV
Blueste
Neighborhood

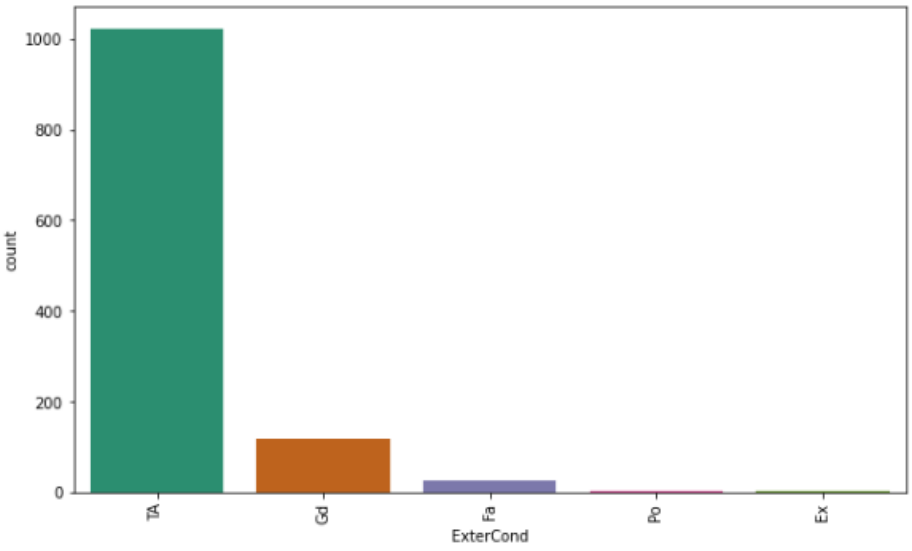

1200
1000
800
600
400
200
0
count
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran

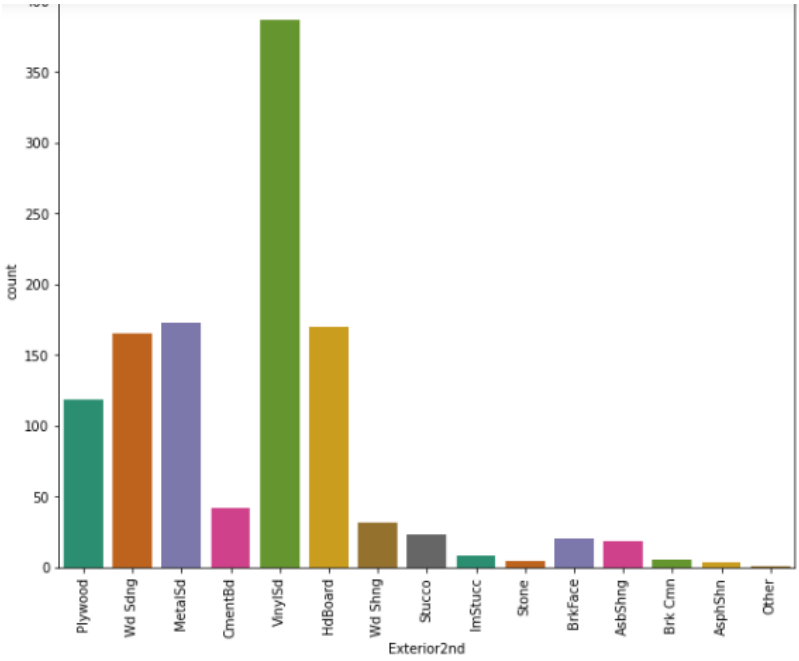

800
600
400
200
0
count
Gable
Flat
Hip
Shed
Gambrel
Mansard
RoofStyle

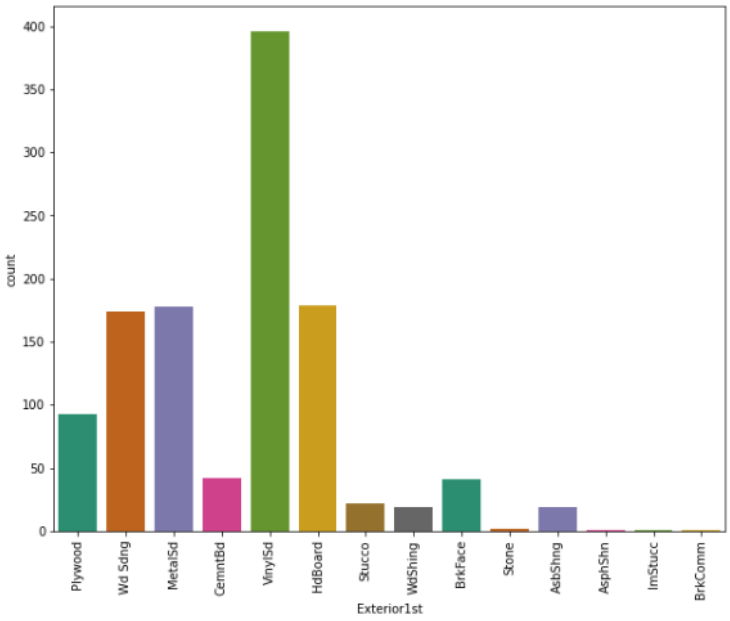

800
600
400
200
0
count
RL
RM
FV
RH
C (all)
MSZoning

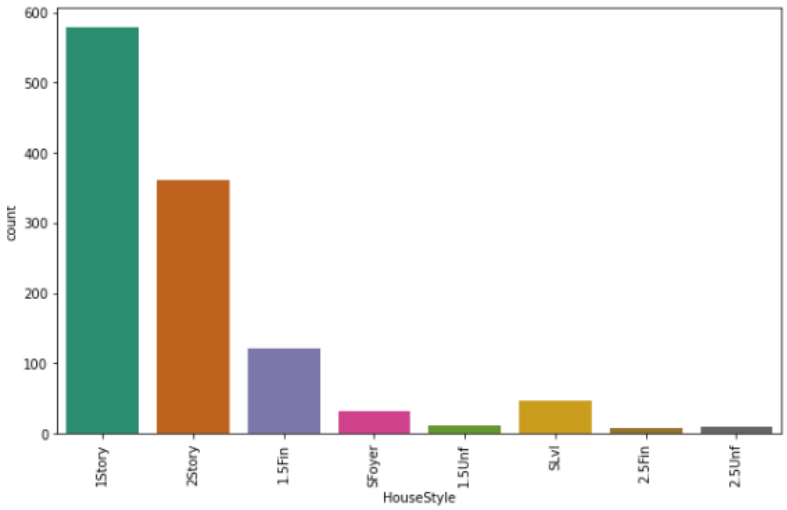

1200
1000
800
600
400
200
0
count
Norm
RRAe
Feedr
PosN
Artery
RRNn
PosA
RRAn
Condition2

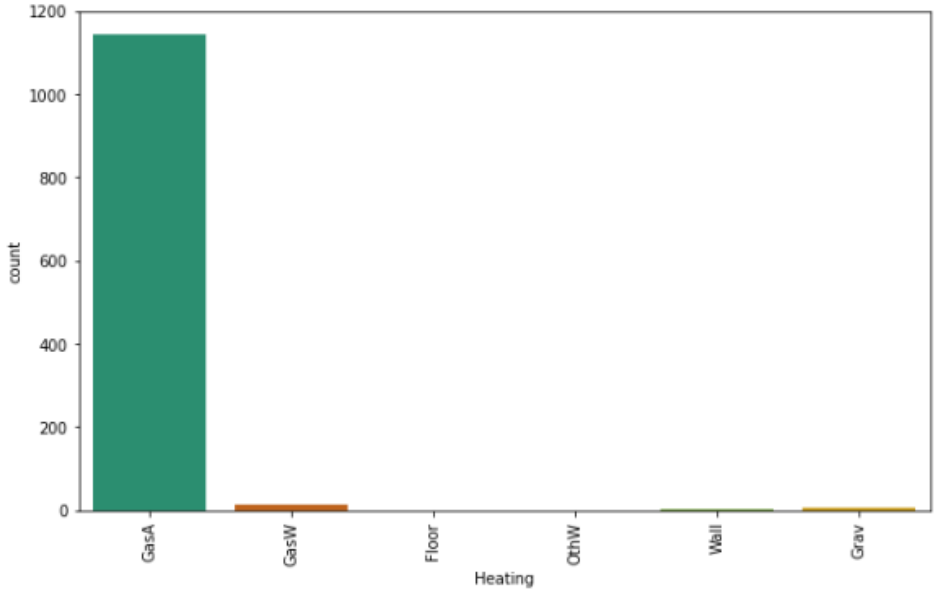

1000
800
600
400
200
0
count
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
BldgType

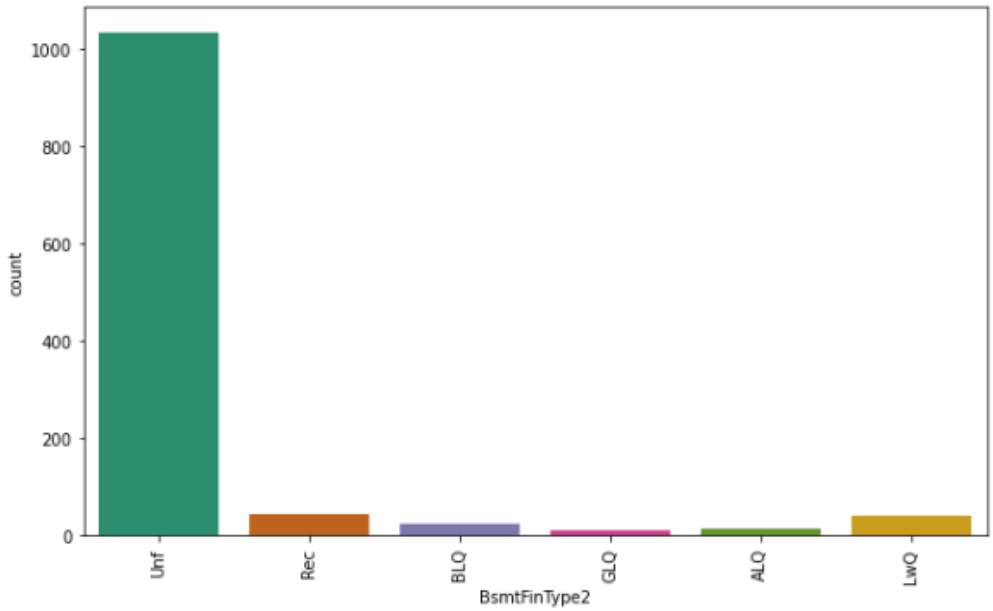

1000
800
600
400
200
0
count
TA
Gd
Fa
Po
Ex
ExterCond

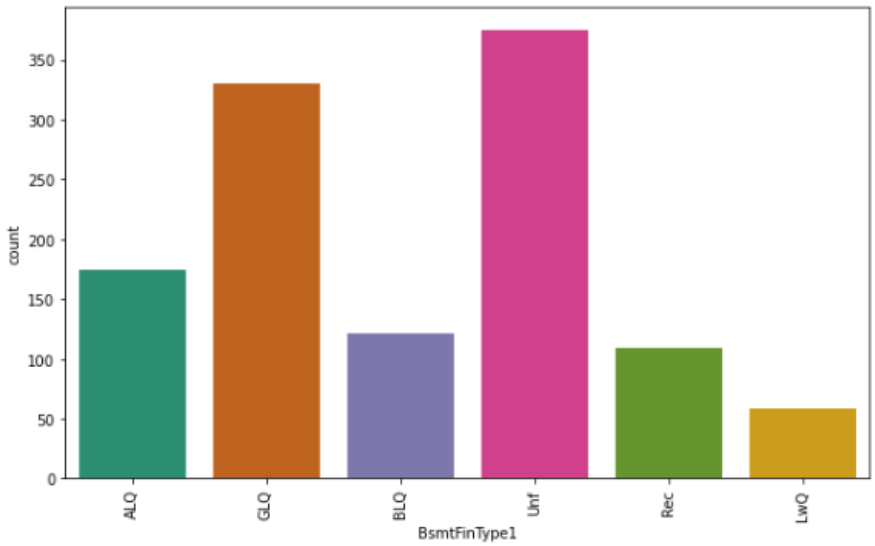

count
0
50
100
150
200
250
300
350
Plywood
Wd Sdng
MetalSd
CmentBd
VinylSd
HdBoard
Wd Shng
Stucco
ImStucc
Stone
BrkFace
AsbShng
Brk Cmn
AsphShn
Other
Exterior2nd

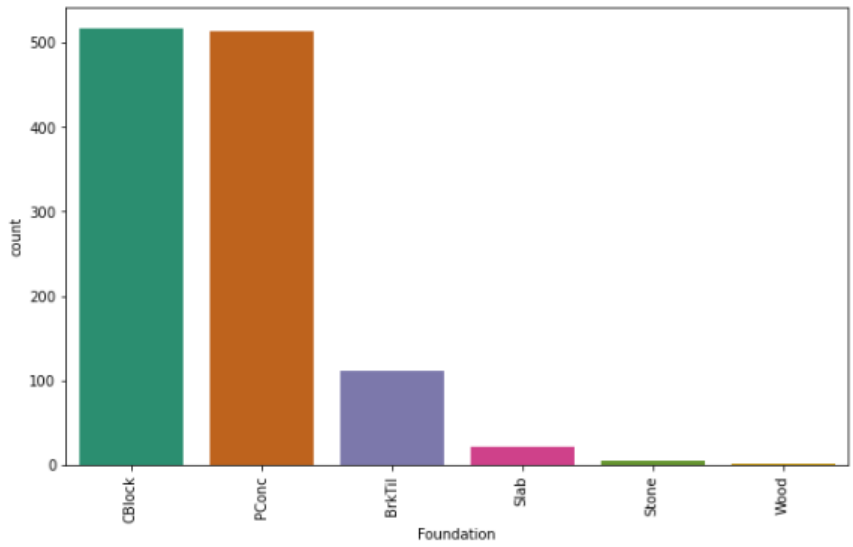

count
0
50
100
150
200
250
300
350
400
Plywood
Wd Sdng
MetalSd
CemntBd
VinylSd
HdBoard
Stucco
WdShing
BrkFace
Stone
AsbShng
AsphShn
ImStucc
BrkComm
Exterior1st

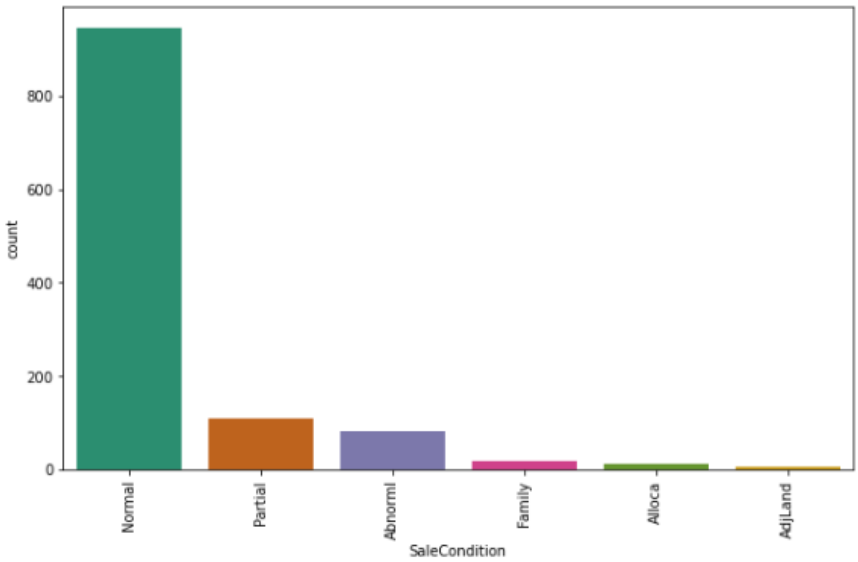

count
0
100
200
300
400
500
600
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
HouseStyle

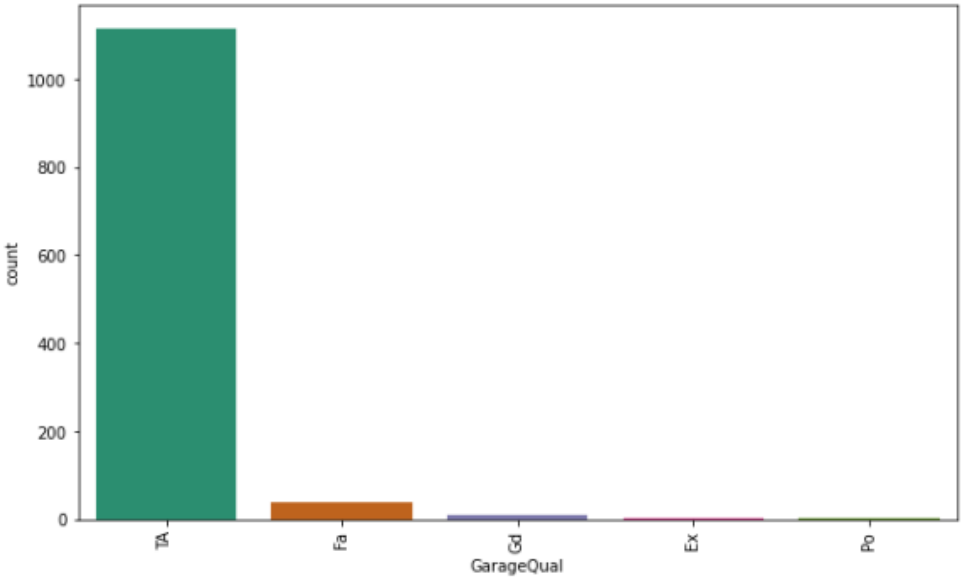

1200
1000
800
600
400
200
0
count
GasA
GasW
Floor
OthW
Wall
Grav
Heating

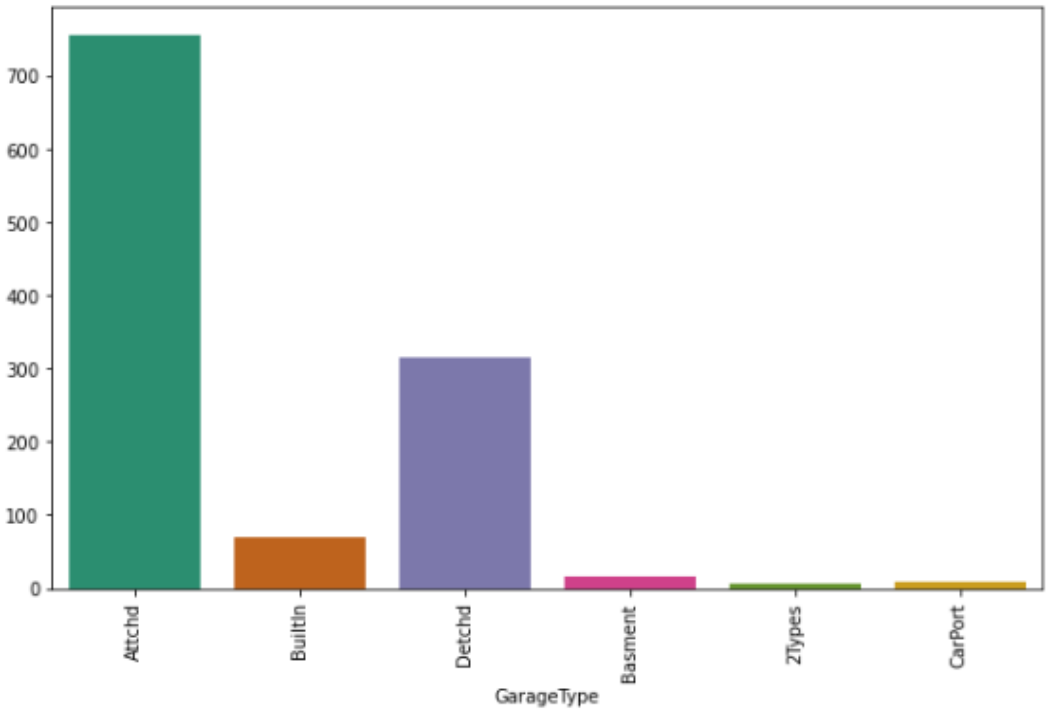

1000
800
600
400
200
0
count
Unf
Rec
BLQ
GLQ
ALQ
LwQ
BsmtFinType2

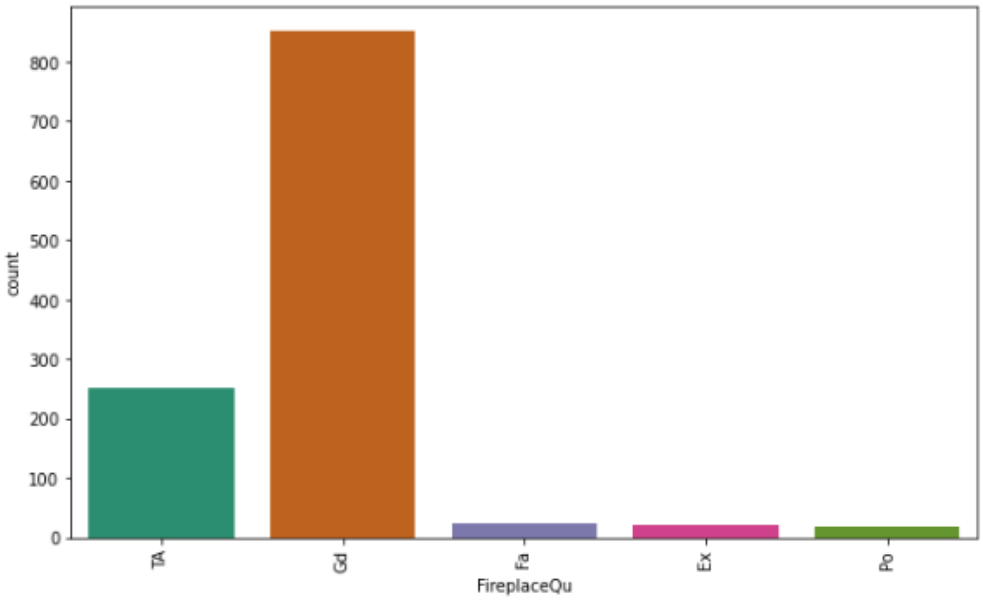

350
300
250
200
150
100
50
0
count
ALQ
GLQ
BLQ
Unf
Rec
LwQ
BsmtFinType1


count
500
400
300
200
100
0
CBlock
PConc
BrkTil
Slab
Stone
Wood
Foundation


count
800
600
400
200
0
Normal
Partial
Abnorml
Family
Alloca
AdjLand
SaleCondition


count
1000
800
600
400
200
0
TA
Fa
Gd
Ex
Po
GarageQual





Observations from the count plots

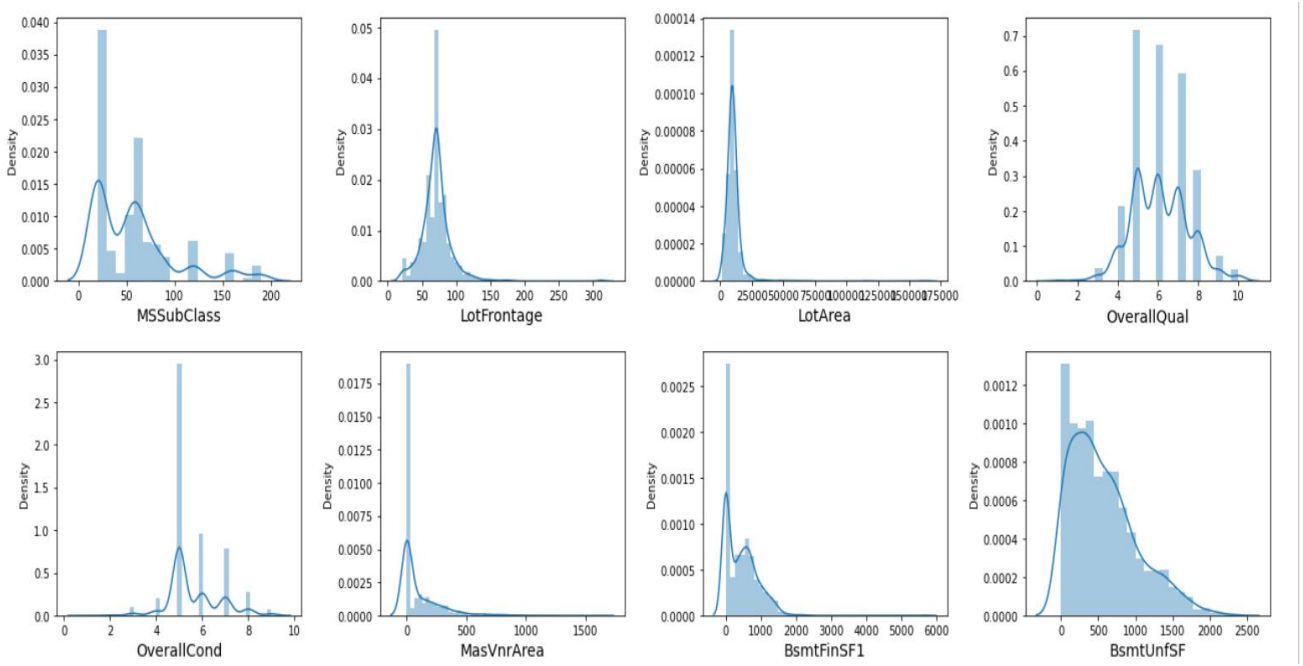
- The houses having Residential Low Density zoning of the sale have high count and commercial zoning sale have very less count compared to others.

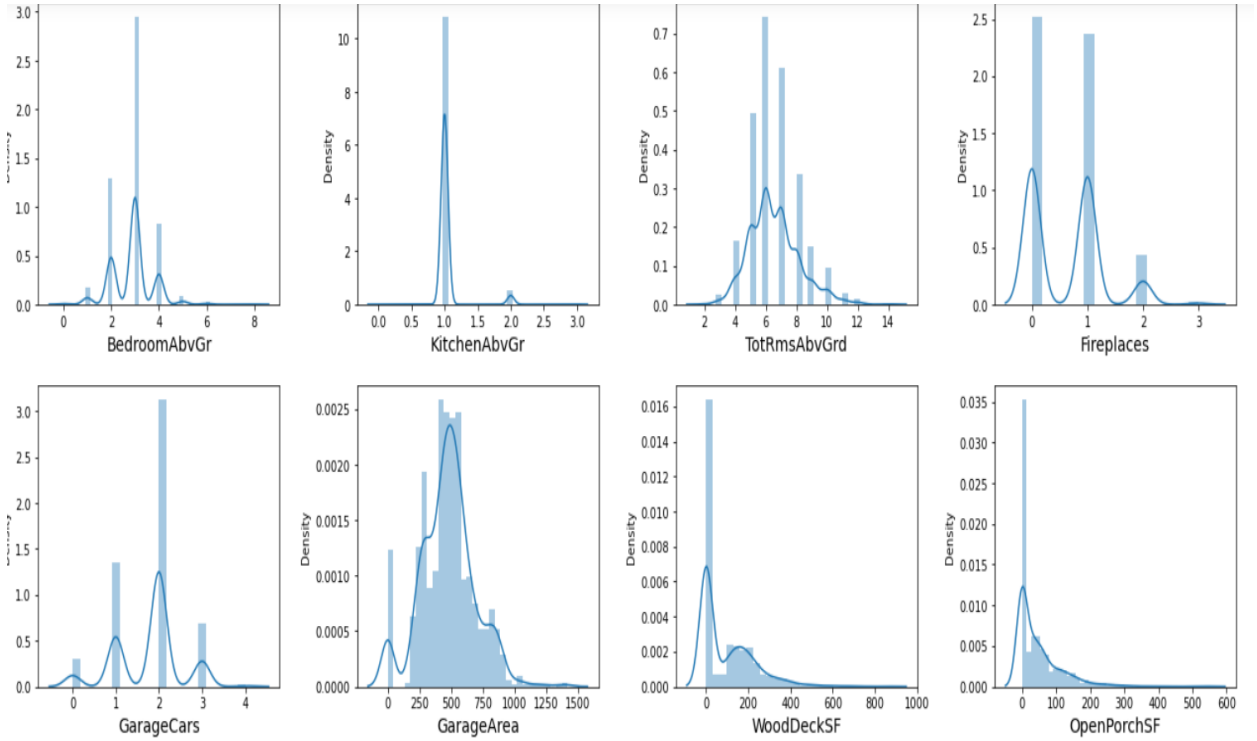
- Inside lot configiration has high count and Frontage on 3 sides of property have very less count compared to others.
- The count is high for the North Ames cities floowed by college creek and count is very low for Bluestem.
- The count is high for the Normal proximity condition apart from this all the others have very less count.
- Similar to condition1, in comdition2 also Normal proximity has very high count compared to others.
- Single-family detached dwelling type has very high counts compared to other types have very less count.
- 1 story style of dwelling has high count followed by 2 story and others have very less count.
- The flat type roof has high count and shed has very less count.
- The roof material type Standard (Composite) Shingle has highest count and others have very less counts.
- Most of the houses have Vinyl Siding exterior covering materials followed by hard board also Brick Common, Asphalt Shingles and Imitation Stucco have very count which means there are no more houses with these types.
- Similar to Exterior1st, here also most of the houses have Vinyl Siding exterior covering materials.
- The present condition of the material on the exterior for most of the houses are Average/Typical.
- Most of the houses have Cinder Block and Poured Contrete type of foundation.
- The count is high for the houses having unfinished basement area. Also some houses have Good Living Quarters.
- Similar to BsmtFinType1, here also the count is high for unfinished basements.

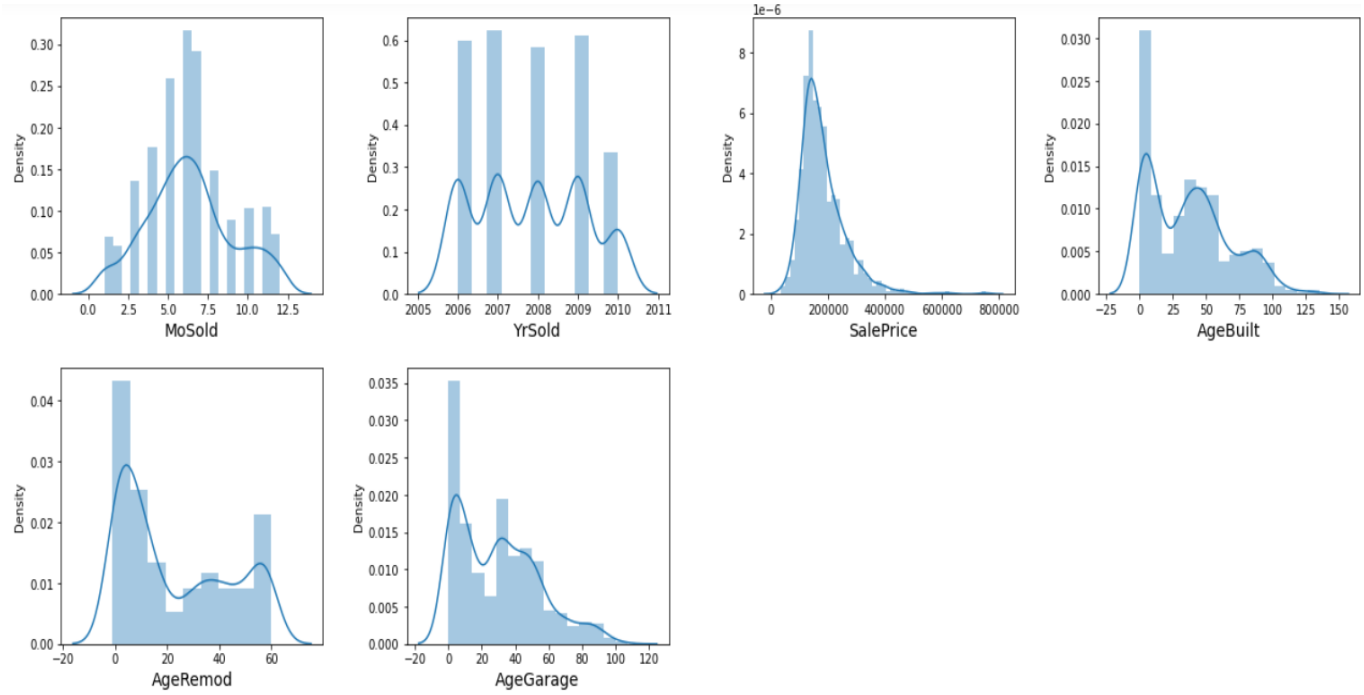
- Most of the houses have Gas forced warm air furnace heating type.
- Most of the houses have excellent heating quality and condition also some houses have typical/average HeatingQC and only 1% of the houses have poor heating quality and condition.
- The electrical system of the type Standard Circuit Breakers & Romex has very high count which means most of the houses have this facility.
- The total number of home functionality of the property for typical functionality have high count compared to others.
- The houses with Masonry Fireplace in main level have good quality compared to others.
- The garage location attached to home has high count also the garage locations detached to home have moderate level of counts. Only few houses have car port and more than one type of garage.
- Houses with typical/average garage quality have high count compared to others.
- Houses with typical/average garage condition have high count compared to others.
- Warranty Deed type of sale has high count followed by Home just constructed and sold(new).
- Normal sale has high count also the home which was not completed when last assessed also has average sale condition.

Now let's visualize the type of distribution of all the features.

Visualizing the data distribution of all features


Density
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
MasVnrArea
BsmtFinSF1
BsmtUnfSF

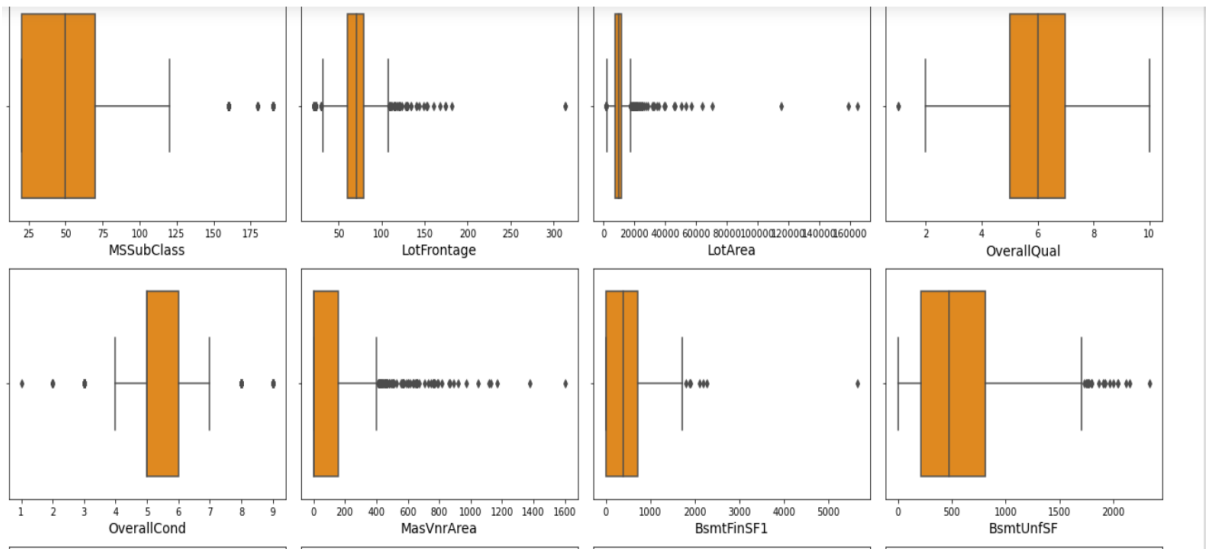

Density
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF

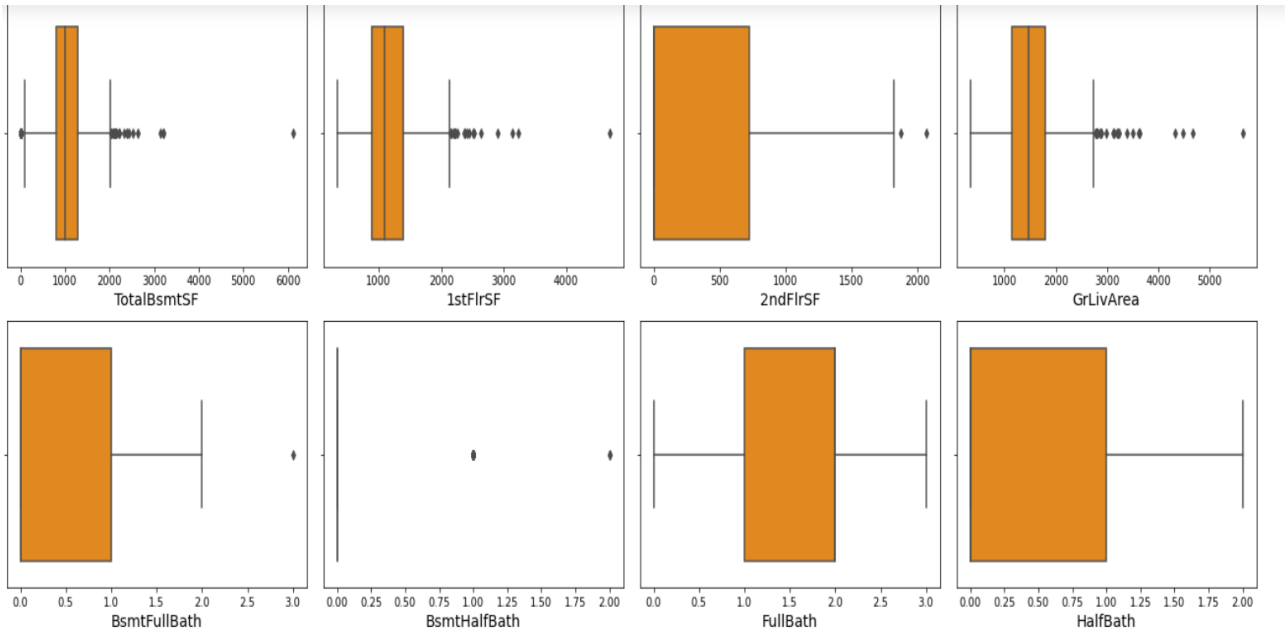


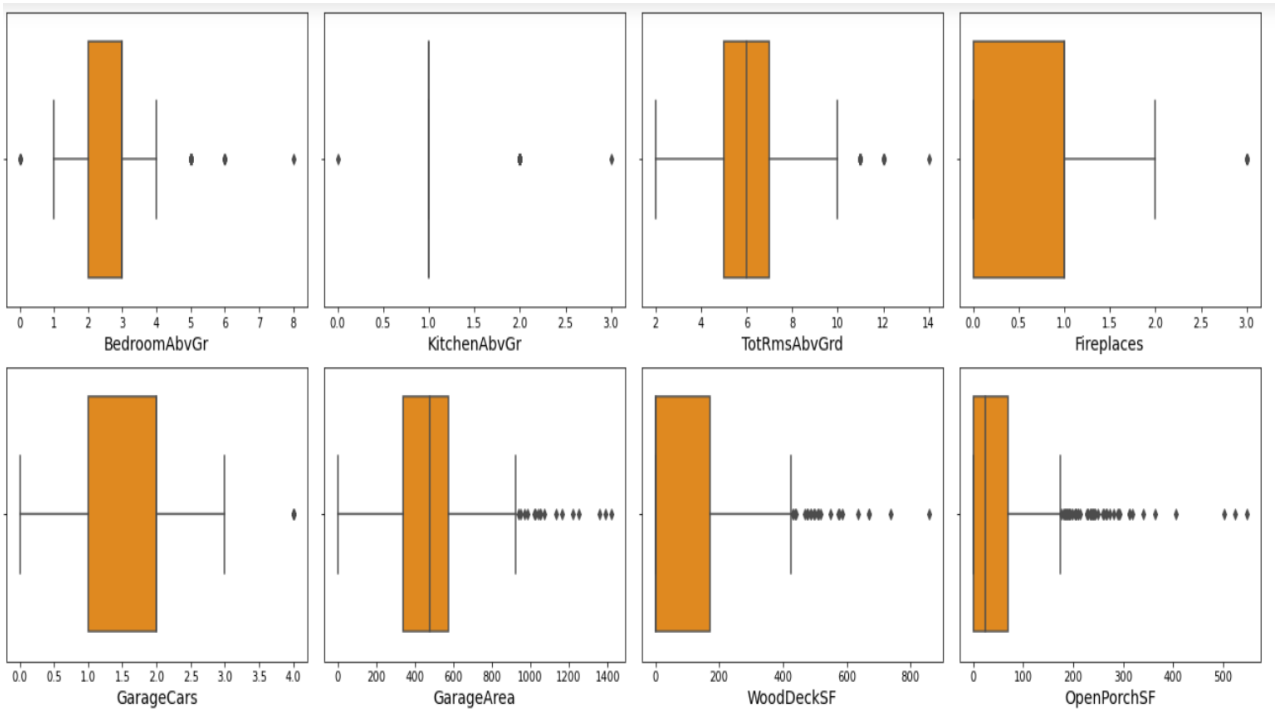
Observations from the distplots:

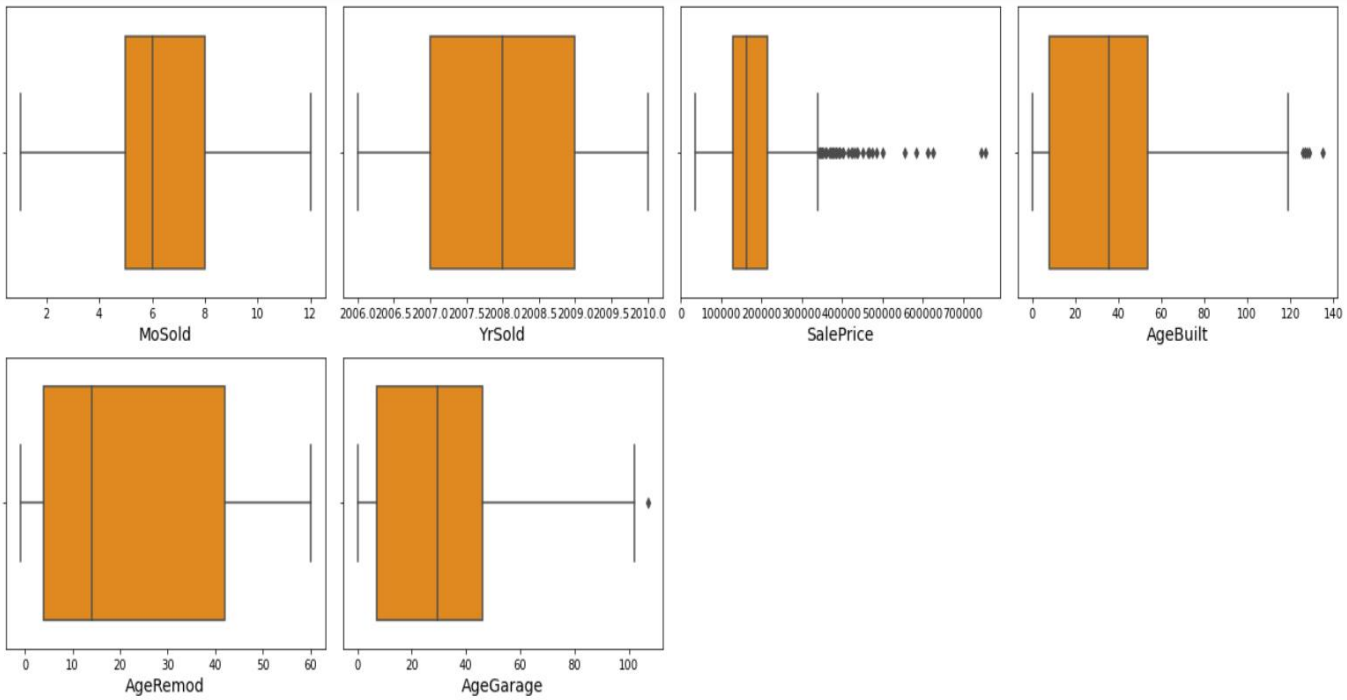
- From the above distribution plot we can observe most of the columns are not normally distributed only some of the columns are almost distributed normally.
- Almost all the columns have skewness and are skewed to right. We need to remove this skewness before building our machine learning models.

Identifying the outliers by using Boxplots

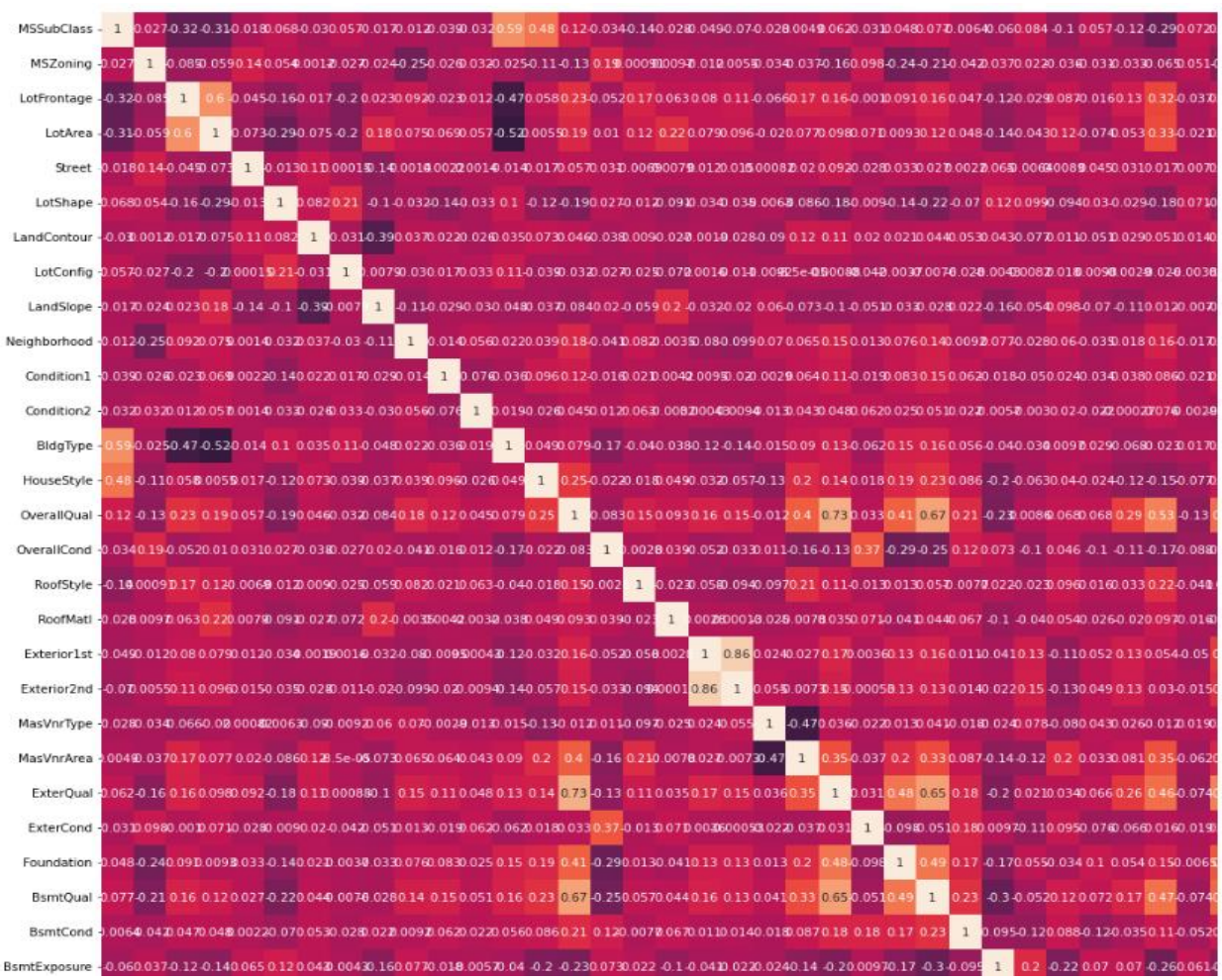



TotalBsmtSF
1stFlrSF
2ndFlrSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath


BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF



Checking multi-colinearity by using Heatmap



3.4 Testing of Identified Approaches (Algorithms)

In this problem SalePrice is my target variable which is continuous in nature, from this I can conclude that it is a regression type problem hence I have used following regression algorithms to predict the sale price of the house. After the pre-processing and data cleaning I left with 67 columns including target and I used these features for prediction.

1. Linear Regression

2. Random Forest Regressor

3. Support vector Regressor

4. Decision Tree Regressor

5. K Nearest Neighbour Regressor

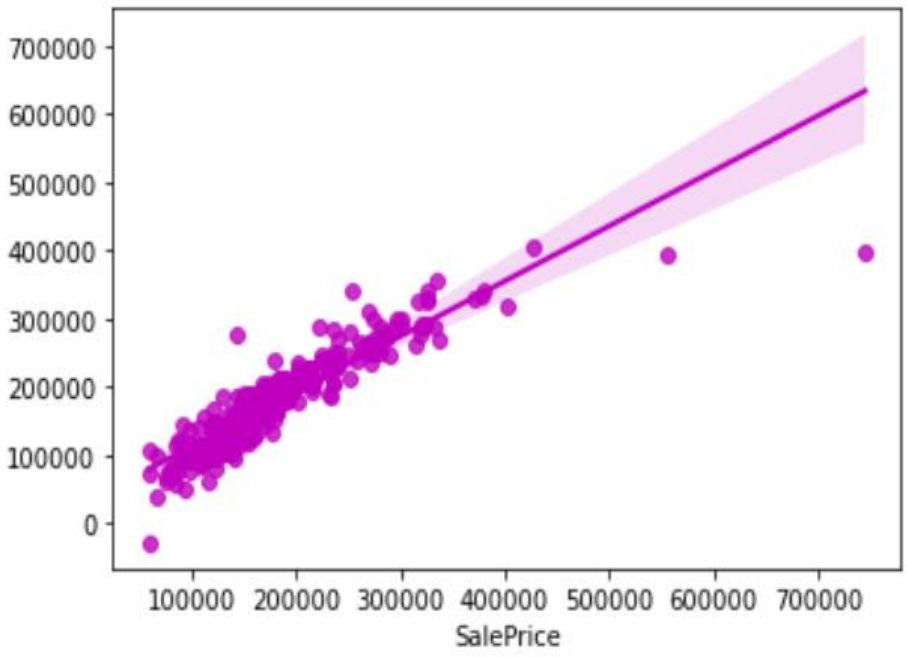
```
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
```

```
regression=LinearRegression()
knn=KNeighborsRegressor()
rf=RandomForestRegressor()
svr=SVR()
dtc=DecisionTreeRegressor()
```

3.4 Run and evaluate selected models

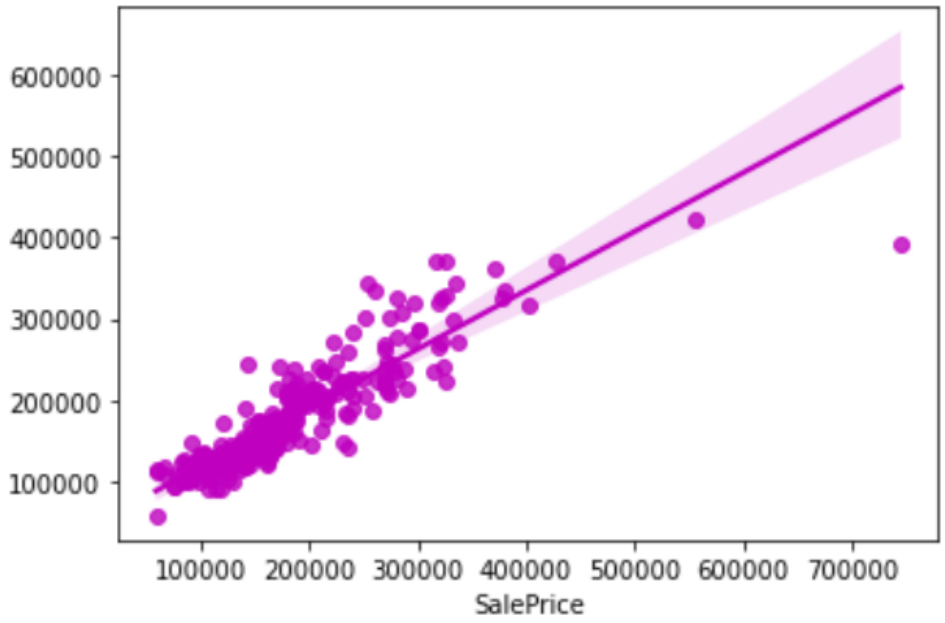
- **LinearRegression Model**

```
LinearRegression()
adjusted R2 score for training data------ 0.830734384951073
adjusted R2 score for testing data------ 0.8173363475569375
mean absolute error------ 20666.47324987143
mean squared error------ 1140920609.8048487
root mean squared error------ 33777.51633564624
```



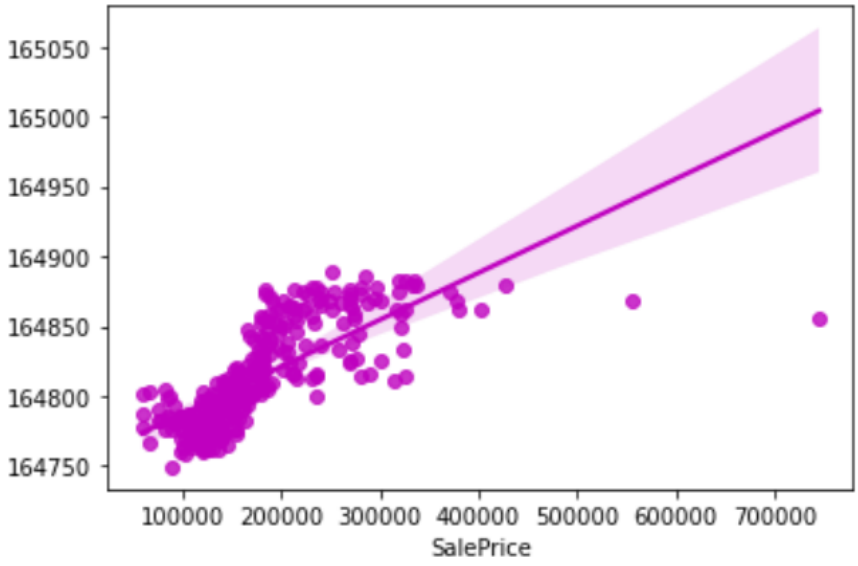
- **KNeighborsRegressor Model**

```
KNeighborsRegressor()
adjusted R2 score for training data------ 0.8128694914401822
adjusted R2 score for testing data------ 0.7838500202348965
mean absolute error------ 22827.45410958904
mean squared error------ 1350076840.272192
root mean squared error------ 36743.39179052734
```



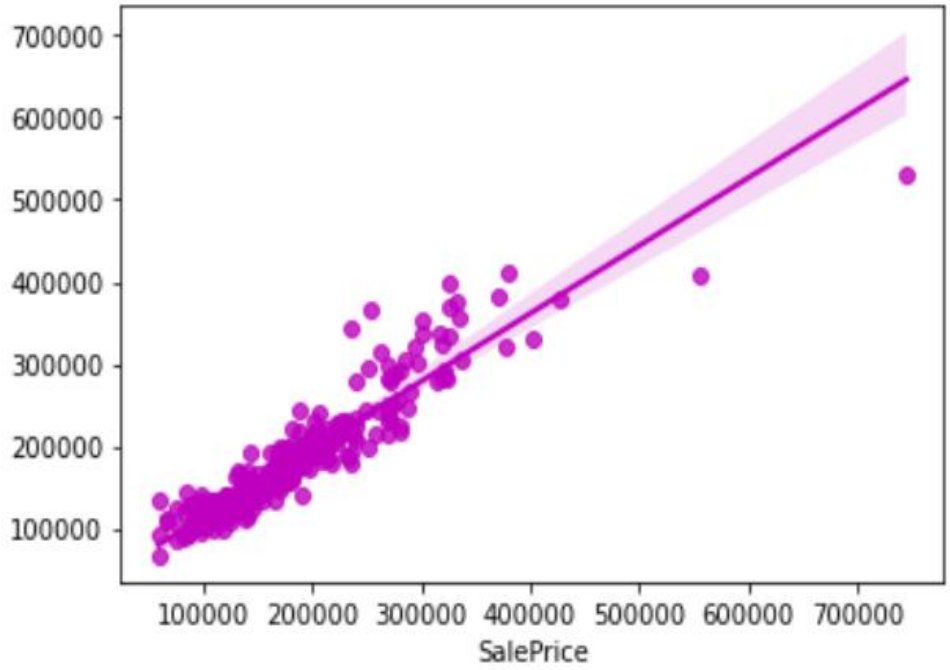
- **SVR Model**

```
SVR()
adjusted R2 score for training data------ -0.04728498637688827
adjusted R2 score for testing data------ -0.03393316982929773
mean absolute error------ 55190.04999460517
mean squared error------ 6457966031.237678
root mean squared error------ 80361.47106193165
```



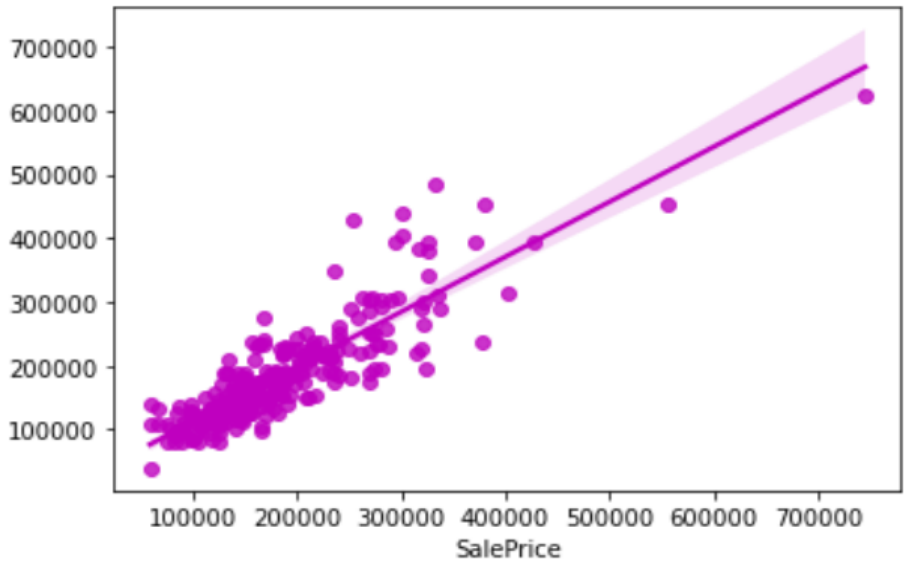
- **RandomForestRegressor Model**

```
RandomForestRegressor()
adjusted R2 score for training data------ 0.9737284137383279
adjusted R2 score for testing data------ 0.8655706665445837
mean absolute error------ 18844.537979452056
mean squared error------ 839648145.9244907
root mean squared error------ 28976.68279711276
```



DecisionTreeRegressor Model

```
DecisionTreeRegressor()
adjusted R2 score for training data------ 1.0
adjusted R2 score for testing data------ 0.7442932917793881
mean absolute error------ 27808.75684931507
mean squared error------ 1597148910.4280822
root mean squared error------ 39964.34548979981
```



Cross-Validation Scores for all the Models

```
1 # Cross validation scores for all models
2 from sklearn.model_selection import cross_val_score
3 for m in models:
4     c_v= cross_val_score(m, x_scaled, y, cv = 10)
5     print ('Cross Validation Score for ',m, ' is :', c_v.mean())
6     print (' ')
```

```
Cross Validation Score for  LinearRegression()  is : 0.8070324237682417

Cross Validation Score for  KNeighborsRegressor()  is : 0.7438954640638589

Cross Validation Score for  SVR()  is : -0.06088659516394675

Cross Validation Score for  RandomForestRegressor()  is : 0.8314994527794763

Cross Validation Score for  DecisionTreeRegressor()  is : 0.574229067871293
```

Choosing the best Model

- After analyzing all the models we have concluded that RandomForestRegressor() model gives the best R2 score and cross validation score. And based on the R2 score we have chosen RandomForestRegressor() as the best model. We will use RandomForestRegressor() model for further analysis.

Hyperparameter Tuning of RandomForestRegressor() model using GridSearchCV

```
# Hyperparameter Tuning
from sklearn.model_selection import GridSearchCV
param_grid={'min_samples_split': [2, 5, 10],'max_features':['auto', 'sqrt'],'min_samples_leaf':[1,2,4,6],'n_estimators': [2,
```

```
gridsearch=GridSearchCV(estimator=rf , param_grid=param_grid )
```

```
gridsearch.fit(x_train,y_train)
```

```
GridSearchCV(estimator=RandomForestRegressor(),
             param_grid={'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [1, 2, 4, 6],
                         'min_samples_split': [2, 5, 10],
                         'n_estimators': [2, 5, 8, 10, 17]})
```

```
gridsearch.best_params_
```

```
{'max_features': 'sqrt',
 'min_samples_leaf': 2,
 'min_samples_split': 5,
 'n_estimators': 17}
```

```
rf=RandomForestRegressor(max_features='sqrt', min_samples_leaf=2,min_samples_split=5,n_estimators=17)
```

```
rf.fit(x_train,y_train)
```

```
RandomForestRegressor(max_features='sqrt', min_samples_leaf=2,
                      min_samples_split=5, n_estimators=17)
```

```
y_pred=rf.predict(x_test)

```

```
y_pred
```

```
array([155926.20868347, 159002.39215686, 125890.90252101, 170074.26470588,
       109577.32352941, 275368.24607843, 195104.28688142, 144455.90616246,
       177241.04575163, 115262.26470588, 139023.56325864, 145229.90569561,
       164249.07200316, 158296.21848739, 121394.18113912, 125710.32142857,
       132968.92156863, 120868.54761905, 156468.93093965, 174466.83006536,
       101087.01988796, 224661.30252101,  91506.59313725, 186310.78431373,
       313691.94509804, 287555.06218487, 103479.94215686, 132426.43641457,
       134624.14565826, 139416.97712418, 144712.21078431, 108499.1713352 ,
       169775.57189542, 137847.2248366 , 141909.27829132, 179894.40373483,
       152781.49159664, 129197.57012138, 329588.09012605, 222317.61951447,
       113595.74369748,  97779.58053221, 161720.0270775 , 112932.5857737 ,
```

```
rf.score(x_train,y_train)
```

```
0.931495566813185
```

```
rf.score(x_test, y_test)

```

```
0.8219097130822915
```

- After using Hyperparameter tuning the accuracy score for RandomForestRegressor Model has increased.

Predicting SalePrice of house for test dataset using our best model.

```
prediction = rf.predict(test)
prediction
```

```
array([327152.7875817 , 232767.03641457, 257653.78039216, 166909.23319328,
       221593.33263305,  97407.77310924, 143675.8767507 , 317514.55784314,
       234699.7805789 , 169635.53711485,  88548.07352941, 144891.52661064,
       122377.55882353, 217821.        , 306989.91561625, 133403.66218487,
       119373.16526611, 125661.94467787, 164296.56582633, 179350.01876751,
       163081.51260504, 156153.43137255, 153816.01493931, 107968.63305322,
       101187.54201681, 126237.34052288, 175903.01470588, 145748.91456583,
       191612.52917834, 106682.49719888, 134219.15919701, 214839.7535014 ,
       241985.15336134, 160962.58403361, 127808.08123249, 177720.27394958,
       203051.33053221, 116196.81372549, 159830.35947712, 149340.65837104,
       115824.7797619 , 278452.58496732, 199951.92810458, 189729.35154062,
       147221.49859944, 134450.81680672, 131577.94677871,  97193.72794118,
       213002.27310924, 323804.20070028, 135080.88235294, 188649.67973856,
       106586.71176471, 105342.10294118, 262238.83921569, 127429.3697479 ,
       147200.33921569, 188074.73295985, 112925.27521008, 257396.29796919,
        95876.64565826, 192485.18814192, 129506.74005602, 146828.85154062,
       197925.34033613,  96789.32072829, 162012.20821662, 212176.05182073,
       140760.821662  , 148021.61531279, 261289.37605042, 171448.18431373,
```

Saving the Final model

```
#saving the machine learning model
import pickle
filename='finalized_model.pickle17'
pickle.dump(rf,open('finalized_model.pickle17','wb'))
```

4. CONCLUSION

In this study, we have used multiple machine learning models to predict the house sale price. We have gone through the data analysis by performing feature engineering, finding the relation between features and label through visualizations. And got the important feature and we used these features to predict the price by building ML models. We have got good prediction results.

Key Findings and Conclusions of the Study

- The houses which have very excellent overall quality like material and finish of the house have high sale price. Also we have observed from the plot that as the overall quality of the house increases, the sale price also increases. That is there is good linear relation between SalePrice and OverallQual. So, if the seller builds the house according to these types of qualities that will increase the sale price of the house.
- There is a linear relation between the SalePrice and 1stFlrSF. As we have seen as the 1st floor area increases, sales price also increases moderately. So, people like to live in the houses which have only 1-2 floors and the cost of the house also increases in this case.
- Also, we have seen the positive linear relation between the SalePrice and GarageArea. As size of garage area increases, sale price also increases.
- There is positive linear relation between sale price and TotalBsmtSF. As total basement area increases, sale price also increases.
- Using the features which have some relation with target we have built many ML models also seen the increase in accuracy of the best model.

4.2 Learning Outcomes of the Study in respect of Data Science

While working on this project I learned more things about the housing market and how the machine learning models have helped to predict the price of house which indeed helps the sellers and buyers to understand the future price of the house. I found that the project was quite interesting as the dataset contains several types of data. I used several types of plotting to visualize the relation between target and features. This graphical representation helped me to understand which features are important and how these features describe the sale price. Data cleaning was one of the important and crucial things in this project where I replaced all the null values with imputation methods and dealt with features having zero values and time variables.

Finally, our aim is achieved by predicting the house price for the test data, I hope this will be further helps for sellers and buyers to understand the house marketing. The machine learning models and data analytic techniques will have an important role to play in this type of problems. It helps the customers to know the future price of the houses.

4.3 Limitations of this work and scope for future work

Limitations:

- In case of processing train and test dataset, I felt concatenation is not suitable as it causes data leakage. The dataset contains some irrelevant columns, zero values, null values, so it is need to increase the dataset size by filling these values.

- The dataset has many limitations, the main limitation is that we have no information potential buyers and environment of the sale. The factors such as auctions can have an influence on the price of the house.

- The dataset does not capture many economic factors. Collecting more accurate and important details about the houses from the buyers will help to analyse the data more clearly.

Future work:

- One of the major future scopes is adding estate database of more cities which will provide the user to explore more estates and reach an accurate decision.
- As a recommendation, I advise to use this model by the people who want to buy a house in the area covered by the dataset to have an idea about the actual price. The model can be used also with datasets that cover different cities and areas provided that they contain the same features. I also suggest that people take into consideration the features that were deemed as most important as seen in this study might help them estimate the house price better.