House price prediction

Submitted by:-

DUMPALA ADITYA

Acknowledgement

References:-google

INTRODUCTION

- **Business Problem Framing:**

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

- **Motivation for the Problem Undertaken:**

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy, and house is essential need and necessary to every one this is the motivation

Analytical Problem Framing

- Data Sources and their formats

  Data set is in the form of csv file and it has two sets of the data one is training and another is test and also shape of the sets are (1168,81) and (292,80) respectively.

- Data Preprocessing Done

  ➔1) checking shape of the data
  ➔2) checking null values of the sets
  ➔3)checking information of the data
  ➔4) checking statistical information of the data it shows that all the information of the sets like mean median and mode standard deviation and minimum value max value.
  ➔5)in the both the sets there are no of null values are present in each feature so we have to fill them by using appropriate data like if it is continuous we can use mean to impute or if it is categorical mode is the best value to fill nan values
  ➔6)after fill the nan data again I am checking info of the both the sets, there are some of the features are in the form of object dtype so ml model did not understand object type data only model understands numerical data so we have to convert object data into numerical data there are many techniques to convert object into numeric.

  In the object type features some are in the form of ordinal data so we have to use ordinal encoder to encode the data

  And get dummies/one hot encoding is suitable for remaining features
  ➔7) after encode the features the next step is to check correlation by using heatmap

- Data Inputs- Logic- Output Relationships:

In the data set we have so many inputs are there, some of important inputs are more Important to out put prediction, like lot area

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

➔In the data set years data is also there iam substract the year of sold from year of built it is easy to find the how old the house, from this iam comparing with sale price by using scatter plot, as years increased the sale price is decreased, it indicates that new house price is more than old house

- Hardware and Software Requirements and Tools Used

➔libraries and packages are used are used in this project

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
pd.set_option('display.max_columns',None)
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
from sklearn.ensemble import RandomForestRegressor,GradientBoostingRegressor, GradientBoostingRegressor
from xgboost import XGBRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Lasso
import joblib
```

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

➔Our target is continuous so regression model is suitable for this problem so I am checking with all the regression models and pick up the best model from all the models

➔ some of the features are skewed right side we have to convert them into normally distributed so I can apply log transformation to convert into normally distributed

- Testing of Identified Approaches (Algorithms)
  ➔LinearRegression
  ➔RandomForestRegressor
  ➔GradientBoostingRegressor
  ➔ BaggingRegressor
  ➔ DecisionTreeRegressor
  ➔ XGBRegressor
- Run and Evaluate selected models
  from the above models two were giving the best score they are randomforest and gradientboostingregressor so I have to check the evaluation metrics for this two models

```
#evaluation metrics for randomforestregressor
from sklearn.metrics import mean_absolute_error,mean_squared_error
print('mean absolute error is:',mean_absolute_error(y_test,pred))
print('\n')
print('mean squared error is:',mean_squared_error(y_test,pred))
print('\n')
print('Root mean suared error is:',np.sqrt(mean_squared_error(y_test,pred)))
```

```
mean absolute error is: 5482.590684931507


mean squared error is: 68759522.72222398


Root mean suared error is: 8292.136197761345
```

```
y_pred1=gbr.predict(x_train)
pred1=gbr.predict(x_test)
```

```
#evaluation metrics gradientboostingregressor
print('mean absolute error is:',mean_absolute_error(y_test,pred1))
print('\n')
print('mean squared error is:',mean_squared_error(y_test,pred1))
print('\n')
print('Root mean suared error is:',np.sqrt(mean_squared_error(y_test,pred1)))
```

```
mean absolute error is: 9431.514845080941


mean squared error is: 161143728.9367134


Root mean suared error is: 12694.239990511973
```

Form the above code we can see clearly rfc error rate is less than gbr so I choose rfc is the final model

- Key Metrics for success in solving problem under consideration

  Metrics used in this project are:-

  ➔R2-score = it is used to evaluate the performance of the model

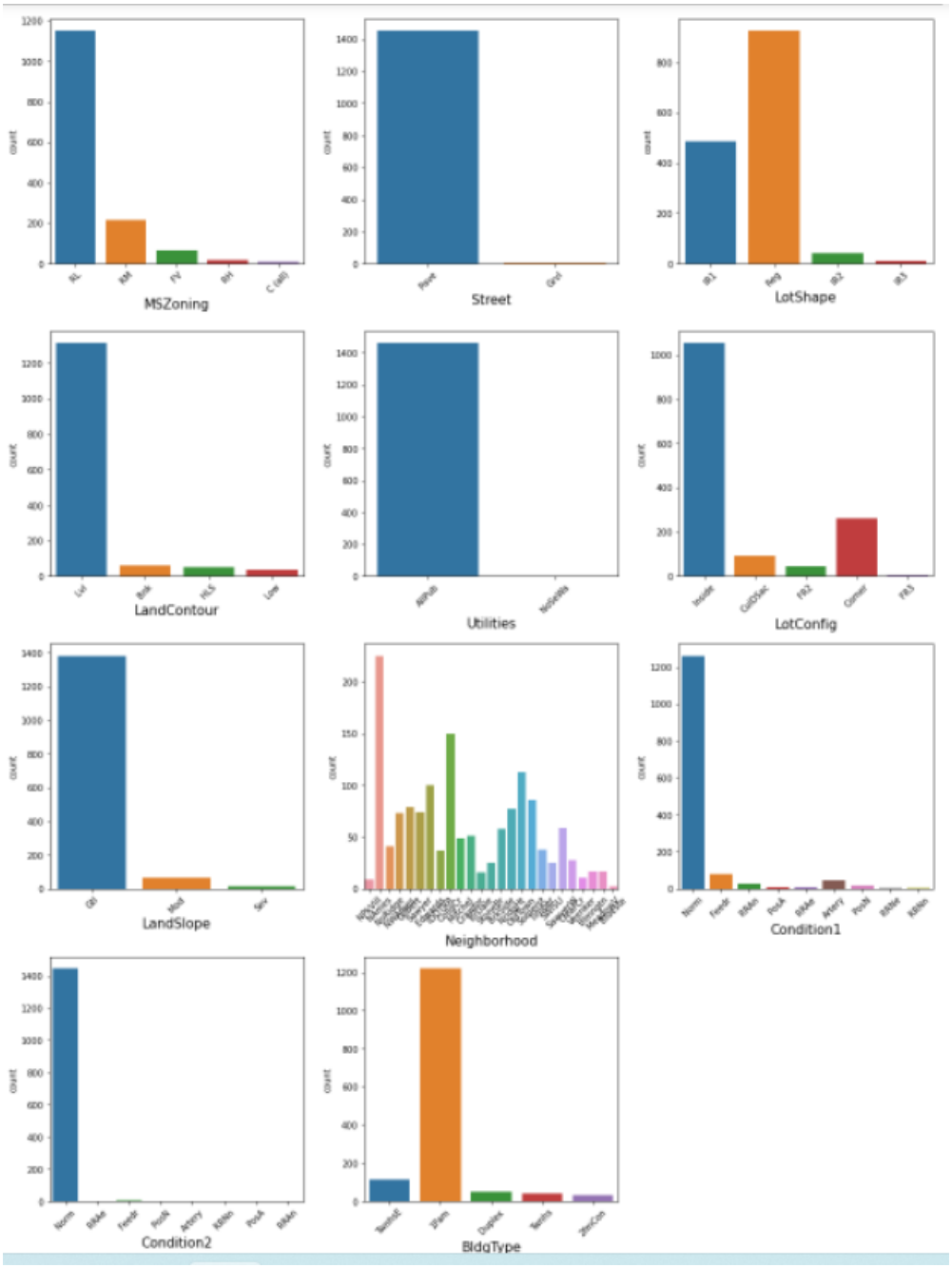
  ➔Cross validation= Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

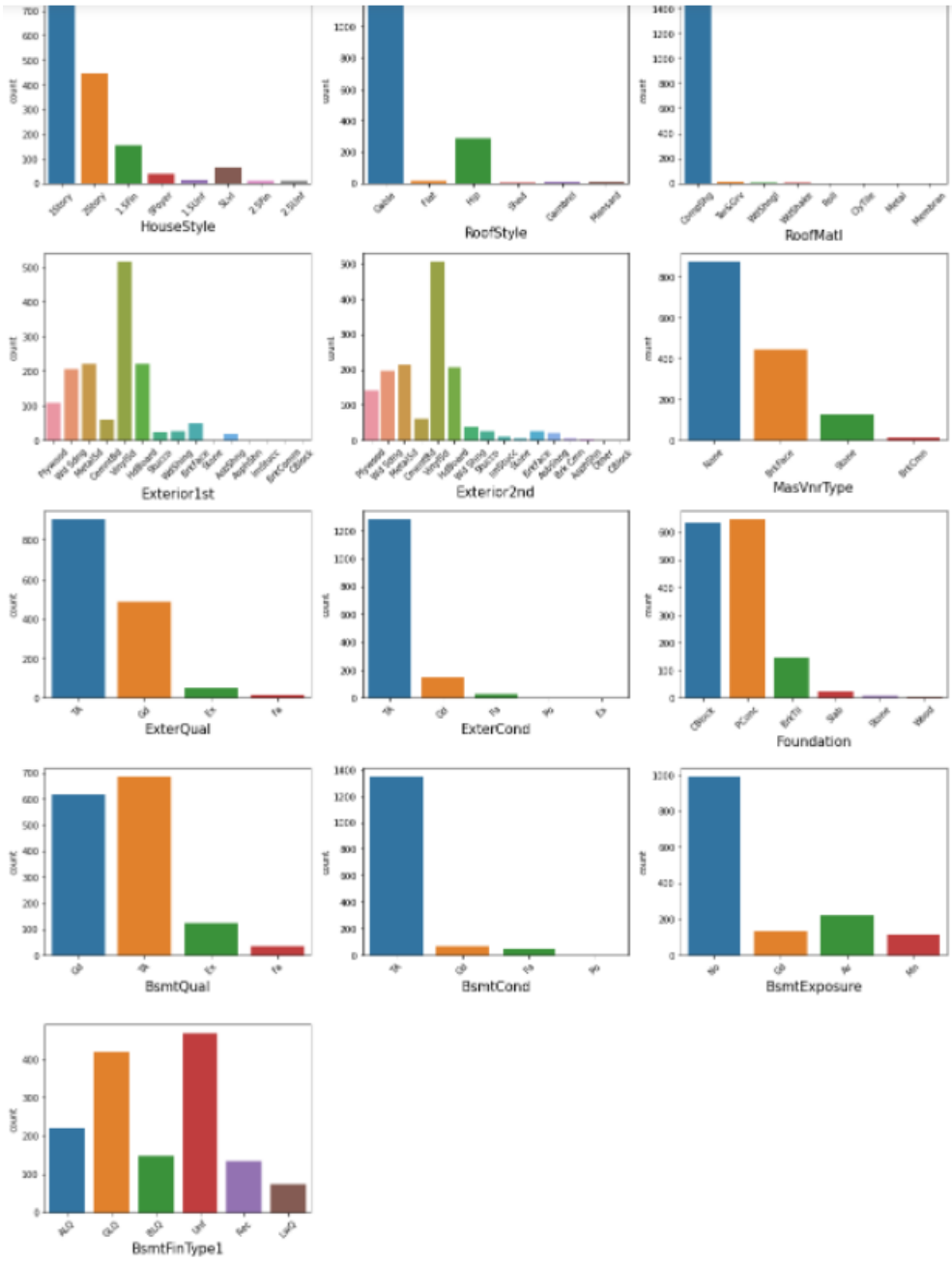
  The three steps involved in cross-validation are as follows :

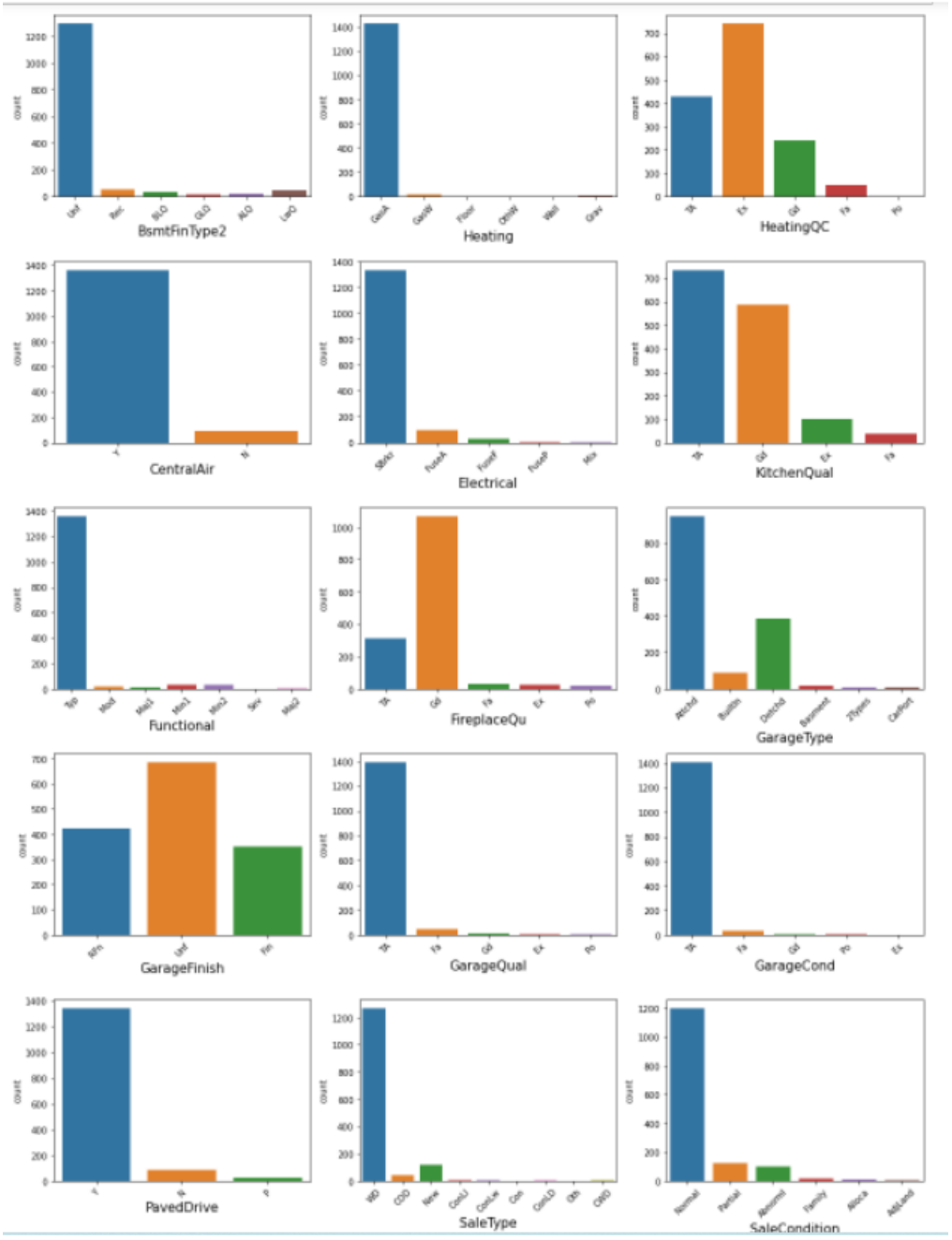
  1. Reserve some portion of sample data-set.
  2. Using the rest data-set train the model.
  3. Test the model using the reserve portion of the data-set.

- Visualizations:

  Univariate analysis

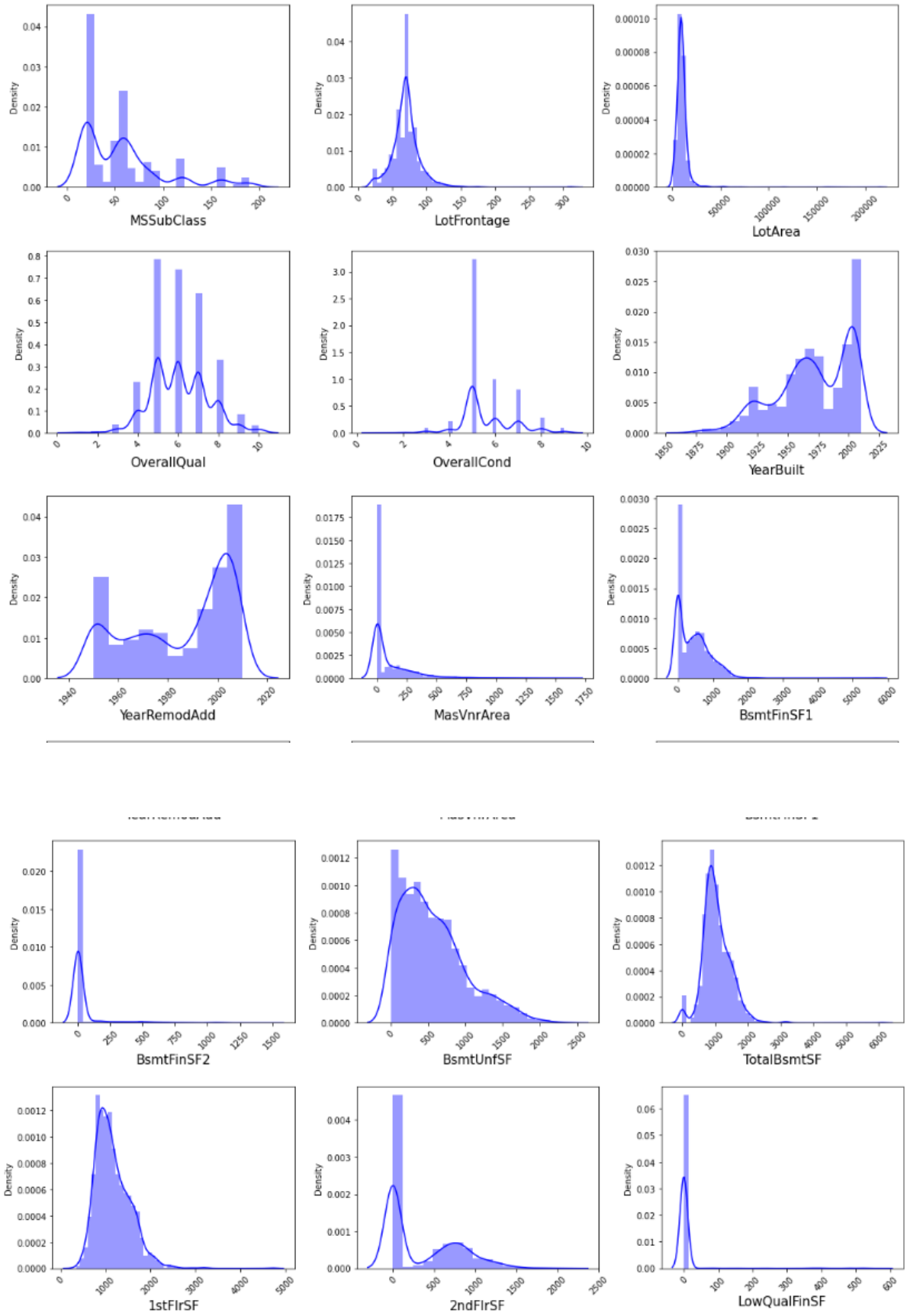

MSZoning
Street
LotShape
LandContour
Utilities
LotConfig
LandSlope
Neighborhood
Condition1
Condition2
BldgType

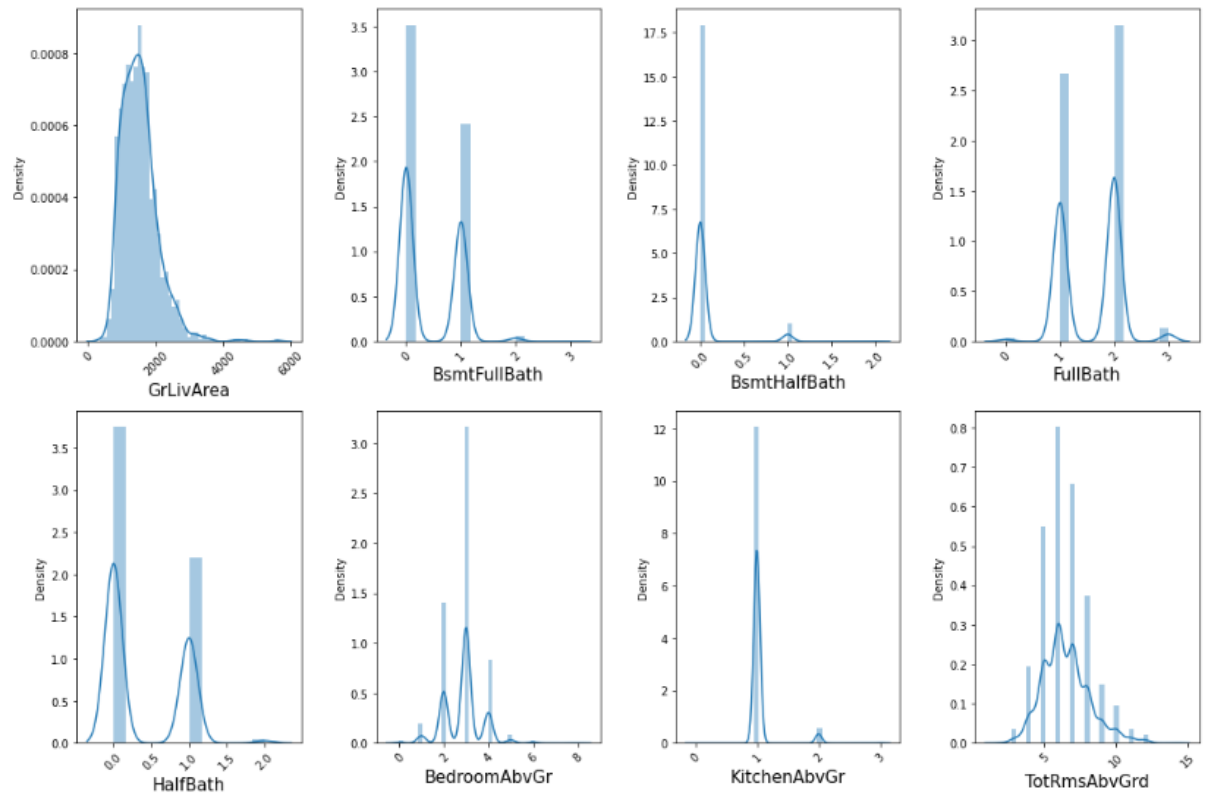

HouseStyle
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1

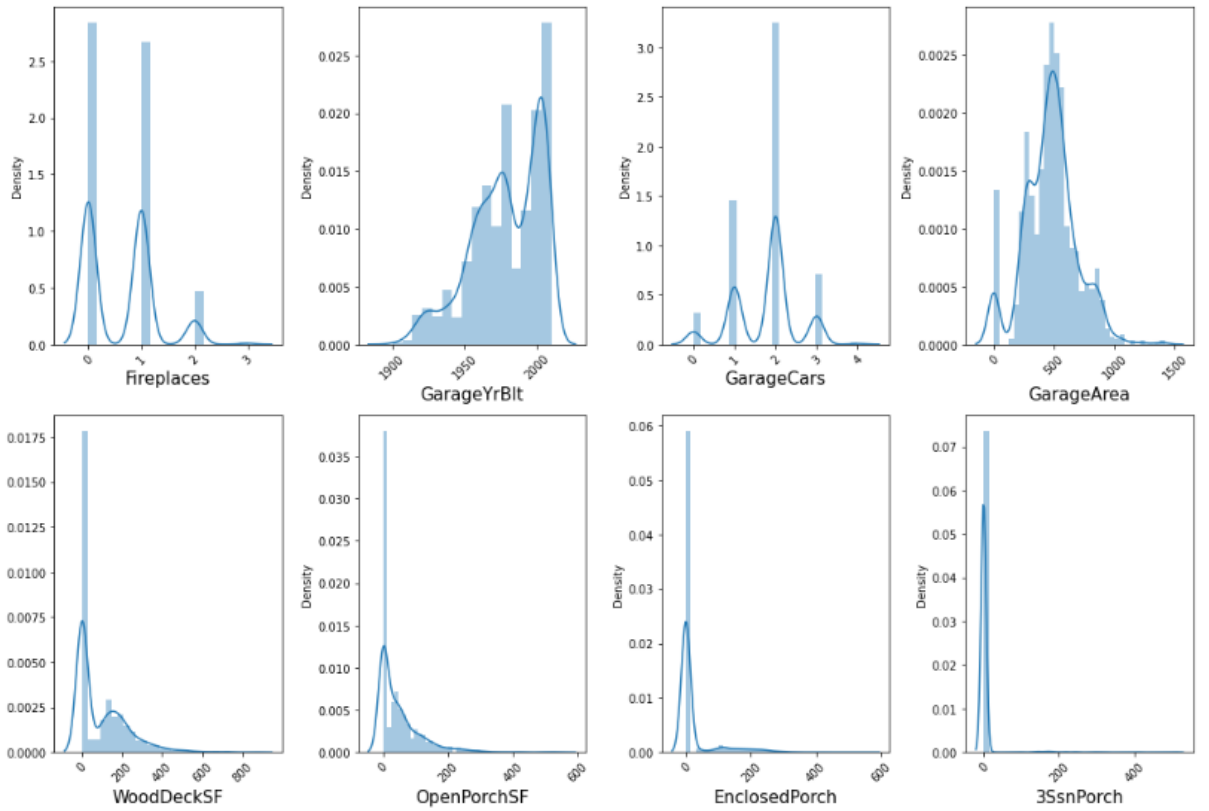


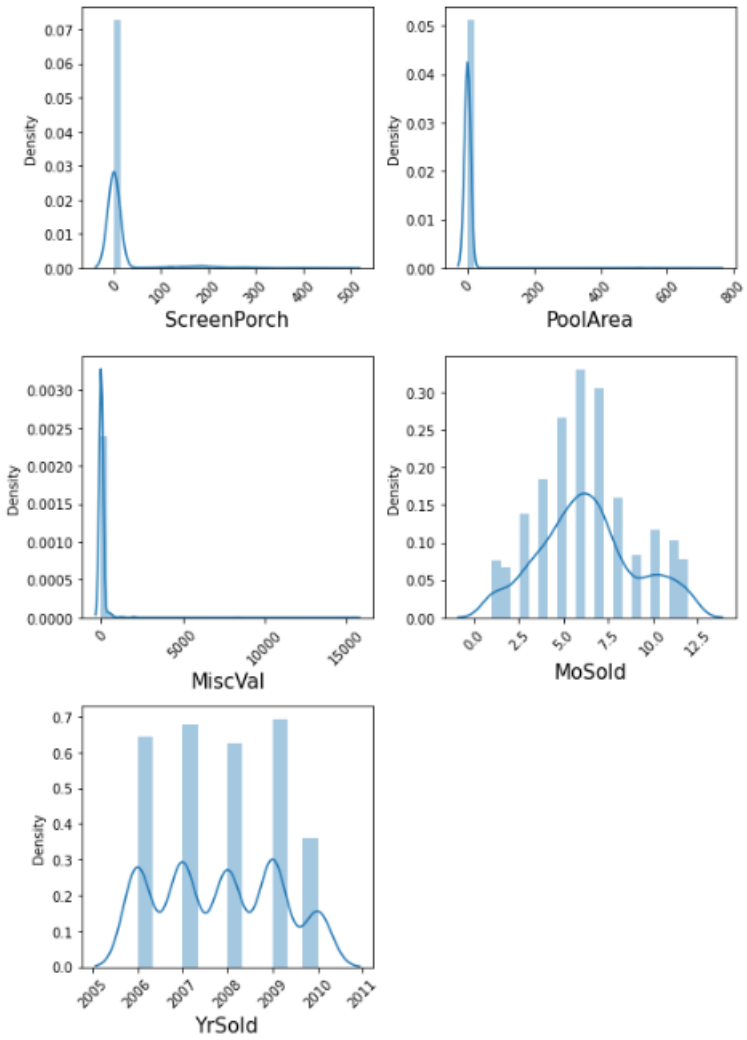
numerical data:

distplot is better for numerical/continuous data


Density
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF


Density
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd

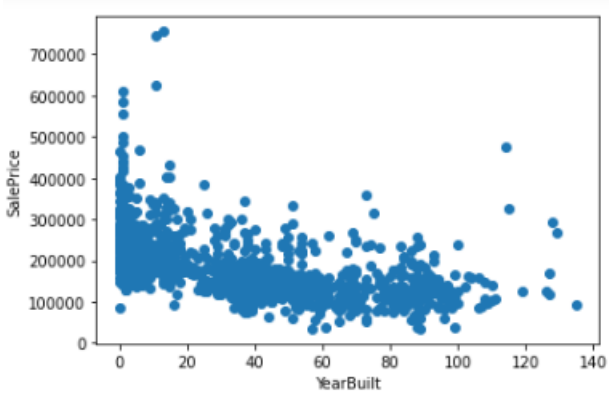

Density
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch

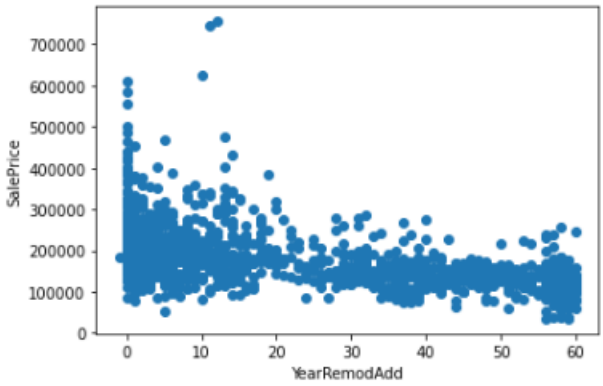


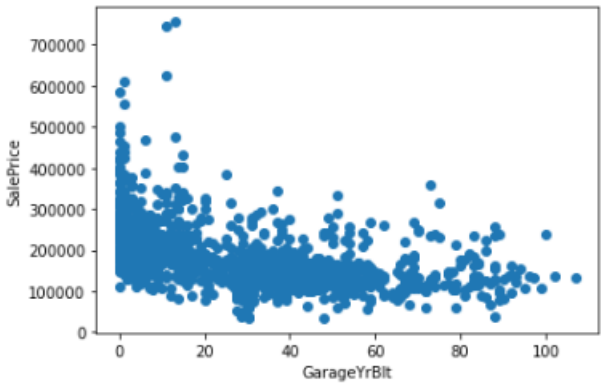
Observations:

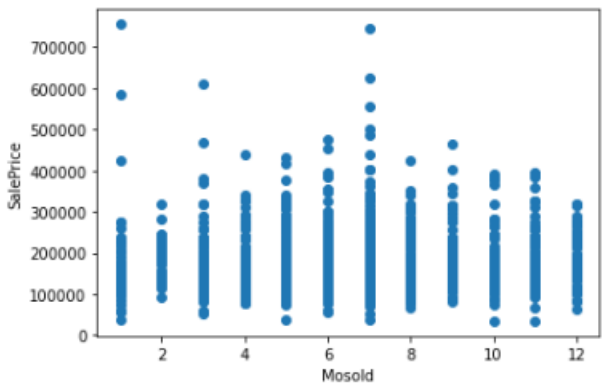
➔most of the features are skewed right side we can see except discrete data

Bivariate analysis:





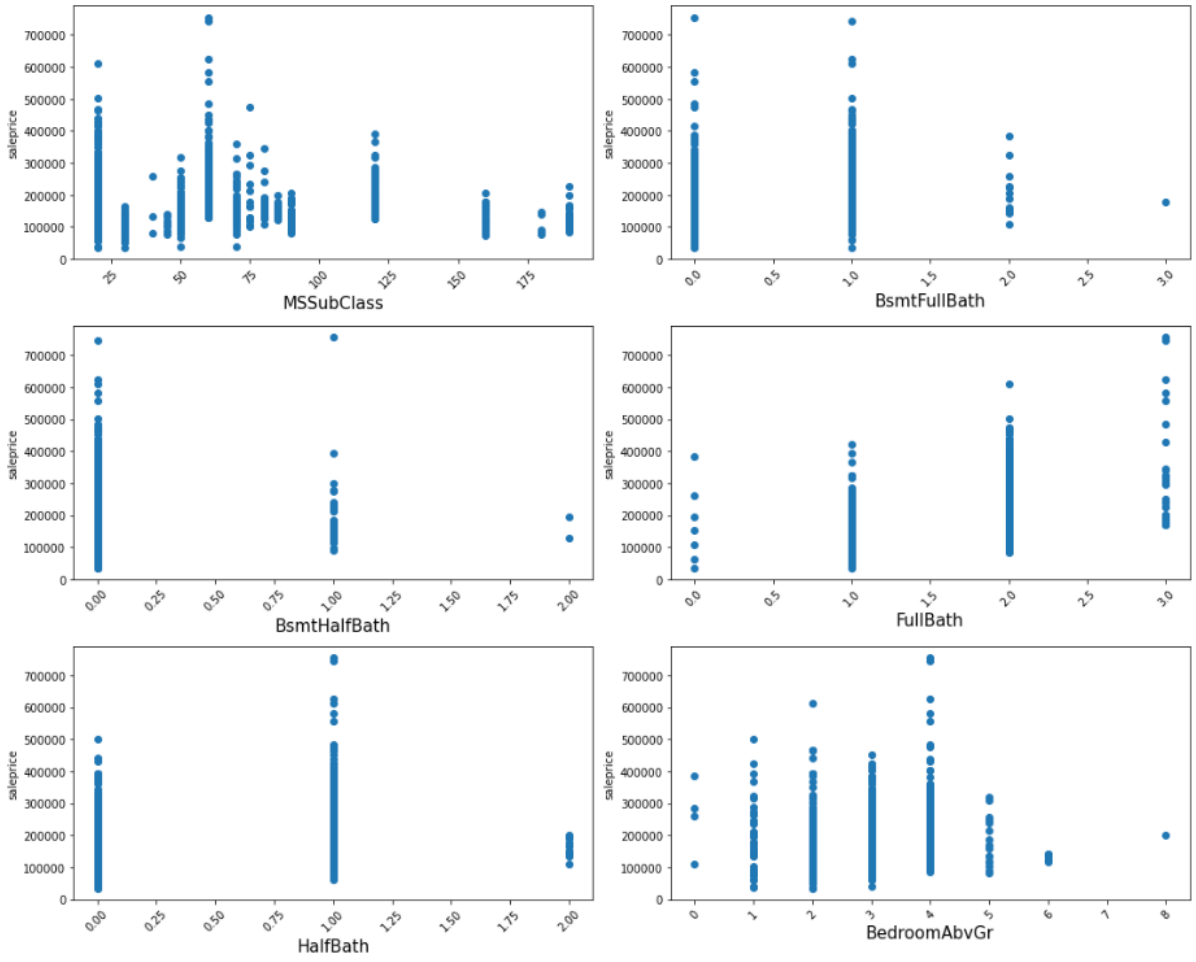


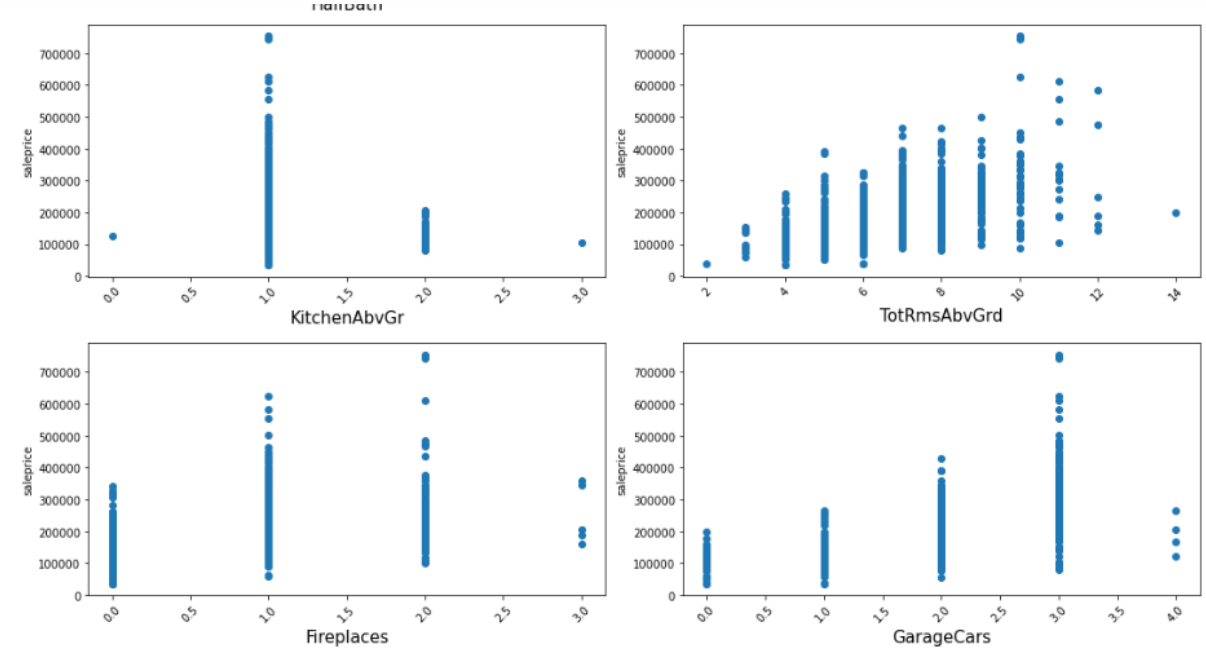


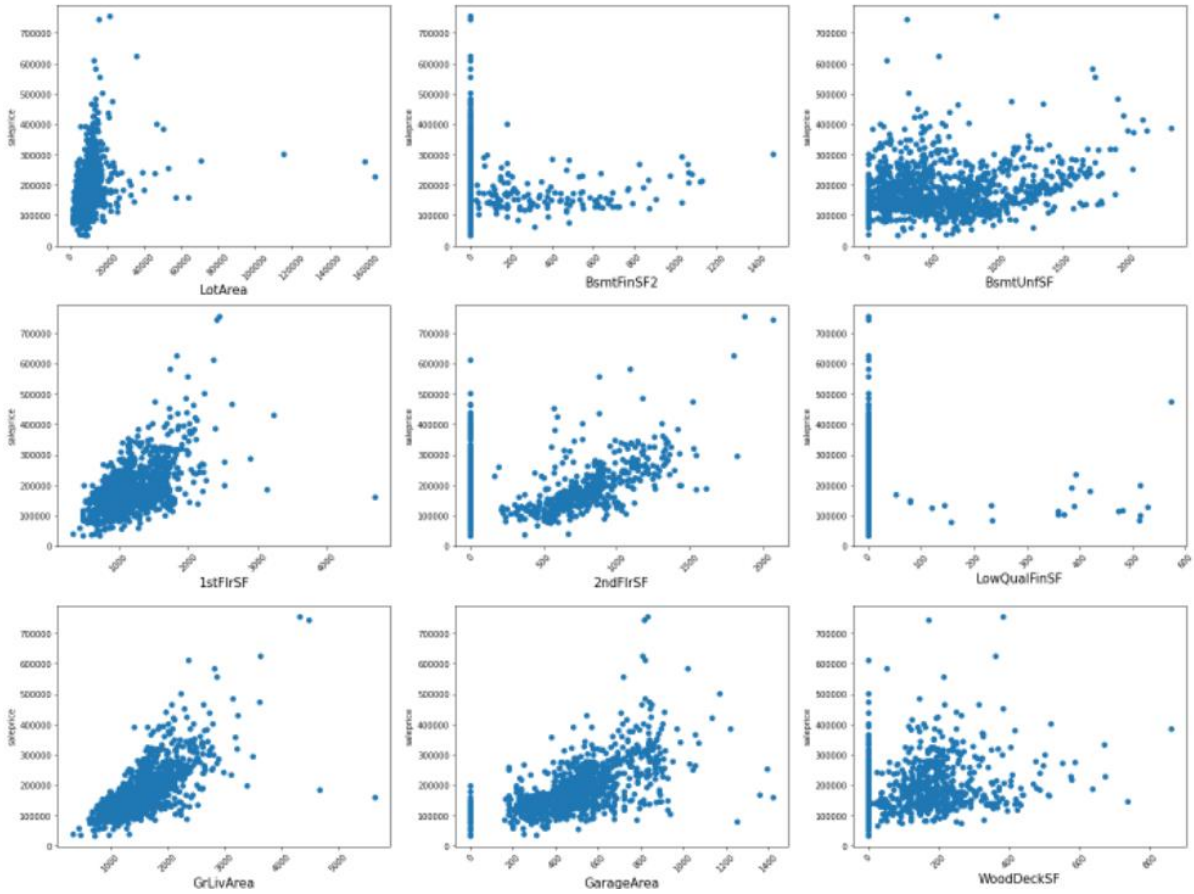
Observations:

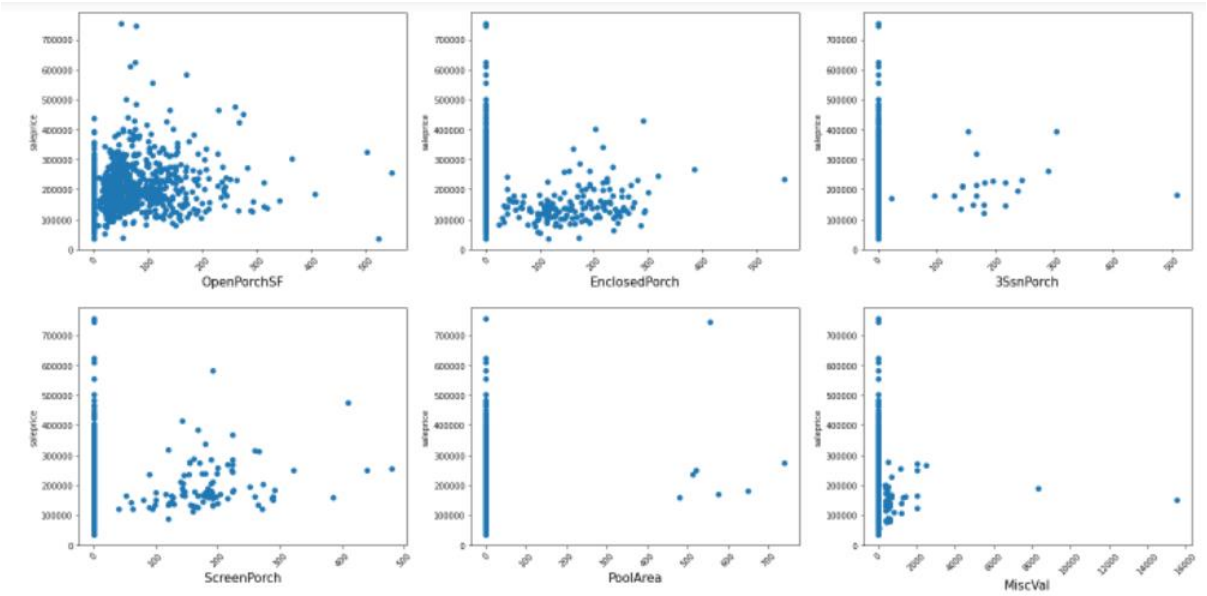
➔no of years increases the price of the house is decreases and it is same in year remodification, garage yearbuilt

➔in the month of 1 and 7 sales are high compared to another months we can see in the Mosold




HalfBath
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageCars
saleprice


LotArea
BsmtFinSF2
BsmtUnfSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
GarageArea
WoodDeckSF
saleprice



Observations:

➔ when the area is increased price is also increases we can see

➔ and also there is good relationship with sales price we can see in each feature In the x-axis value increases the price in the y-axis is also increases so this indicates that there is a good relationship between them

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

- Interpretation of the Results

➔from the visualization we can see when features compared with label by using scatterplot there is good relationship

➔after dealing with null values and encoding the categorical features the total no of features are become 203

CONCLUSION

- Key Findings and Conclusions of the Study

➔From the whole problem we can observe that house price is high when the year of built is low I,e recent constructed house price is high.

➔if area is increases the price of the house is also increases

➔more than 90% of the values are null those features are not useful for our model so we can remove those

➔quality is also contributes more to the output(saleprice)

➔coming to the statistical observations most of the features are having right skewed data due to high values present in the features some of the features are also having more no of zeros

➔In the evaluation metrics the error rate is low in the random forest regressor model so it is better to choose this model for final model

- Learning Outcomes of the Study in respect of Data Science

  Learning objectives:

  1)fill null values
  2)using different ways to fill nan values
  3)using different metrics to convert the object data into numerical data(simple imputer and onehot encoding)
  4)using transformation techniques
  5)finding correlation of the features
  6)scaling the data(independent data)
  7)applying various ml models to the given data
  8)applying various metrics to check the performance of the model
  9)checking evaluation metrics
  10)finalise the model based on error rate

11)perform hyperparameter tunning for improve the accuracy of the model

12)saving the model by using joblib

➔this is the one of the best and toughest problem I have been attempt

➔my learnings form this problem is:

1)finding null values and removing those are more than 90%

2)knowing about visualization tools seaborn and matplotlib

3)I learnt fill null values with appropriate data i.e if the data is categorical iam using mode and if the data is continuous I can choose mean or median

4)and also dealing with skewed data in our problem most of the data is skewed right side so applying log transformation is better for dealing with right skewed data,and also most of the features are having zeros applying log to the zero is going to print infinite so to overcome this iam adding constant value 1.

5)after applying getdummies and simple imputer no of features are increased more, this becomes very tough to me to plot heat map and finding correlation within the features.

6)after applying minmax scaler or scaling the data iam checking with all the models to trian and test data from all those models two were given best score they are randomforest and gradientboosting but here I have to choose one of them only,for this iam going to check evaluation metrics i.e M,S,E , M,A,E , R,M,S,E after checking rfc is given less error so iam choosing randomforest is the final model

➔Challenges:

1)Due to more no of features it is difficult to finding the correlation between the features

To overcome this situation iam using triangular heatmap (mask)

To plot the heatmap it plots only lower triangle of the map it helps us to see the features some more clearly

2)two models are giving best score but after checking evaluation metrics model with high accuracy is given high error and model with little low accuracy given low error rate so we don't trust before checking evaluation metrics of the each model

- Limitations of this work and Scope for Future Work

  Limitations of this solution is this is the model which predicts out come only when the inputs of the features which is already contains inside

  Due to developing technology more variety of the models of the houses are made so it is difficult to predict with new independent features i.e which are not included in the builten model