Project Report On

Housing: Price Prediction

Submitted By: Arun Sharma | Internship 28

ACKNOWLEDGEMENT

The following research papers helped me understand the problem of house prices, the various factors affecting house prices, fluctuations in house prices and their causes & helped me in my model building & predictions:

1."Comparison of Data Mining Models to Predict House Prices"- Stephen O'Farrell

Buying a house is commonly the most important financial transaction for the average person. The fact that most prices are negotiated individually (unlike a stock exchange system)creates an environment that results in an inefficient system. Most people buying houses are inexperienced amateurs with limited information about the market. A housing bubble cannot exist if individuals are making rational decisions. The objective of this paper is to evaluate the performance of a stacked regression model compared to several sub models based on predicting house prices. House characteristics and the final house price was gathered from King County,Washington, USA during the period of May 2014 and May 2015. The observed result indicates that combining the sub algorithms using a general linear model did not significantly improve results.

2. "Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study" - Shashi Bhushan Jha, Rajesh Kumar Jha

Developing an accurate prediction model for housing prices is always needed for socio-economic development and well-being of citizens. In this paper, a diverse set of machine learning algorithms such as XGBoost, CatBoost, Random Forest, Lasso, Voting Regressor, and others, are being employed to predict housing prices using publicly available datasets. The housing datasets of 62,723 records from January 2015 to November 2019 are obtained from Florida Volusia County Property Appraiser website. The records are publicly available and include the real estate or economic database, maps, and other associated information. The database is usually updated weekly according to the State of Florida regulations. Then, the housing price prediction models using machine learning techniques are developed and their regression model performances are compared. Finally, an improved housing price prediction model for assisting the housing market is proposed. Particularly, a house seller or buyer, or a real estate broker can get insight in making better-informed decisions considering the housing price prediction. The empirical results illustrate that based on prediction model performance, Coefficient of Determination (R2), Mean Square Error (MSE), Mean Absolute Error (MAE), and computational time, the XGBoost algorithm performs superior to the other models to predict the housing price.

3. "Housing Price pr Housing Price prediction Using Suppor ediction Using Support Vector Regr or Regression" - Jiao Yang Wu

The relationship between house prices and the economy is an important motivating factor for predicting house prices. Housing price trends are not only the concern of buyers and sellers, but it also indicates the current economic situation. Therefore, it is important to predict housing

prices without bias to help both the buyers and sellers make their decisions. This project uses an open source dataset, which include 20 explanatory features and 21,613 entries of housing sales in King County, USA. We compare different feature selection methods and feature extraction algorithms with Support Vector Regression (SVR) to predict the house prices in King County, USA. The feature selection methods used in the experiments include Recursive Feature Elimination (RFE), Lasso, Ridge, and Random Forest Selector. The feature extraction method in this work is Principal Component Analysis (PCA). After applying different feature reduction methods, a regression model using SVR was built. With log transformation, feature reduction, and parameter tuning, the price prediction accuracy increased from 0.65 to 0.86. The lowest MSE is 0.04. The experimental results show there is no difference in performance between PCA-SVR and feature selections-SVR in predicting housing prices in King County, USA. The benefit of applying feature reductions is that it helps us to pick the more important features, so we will not overfit the model with too many
features.

4. "BANGALORE HOUSE PRICE PREDICTION" - Amey Thakur, Mega Satish

We propose to implement a house price prediction model of Bangalore, India. It's a Machine Learning model which integrates Data Science and Web Development. We have deployed the app on the Heroku Cloud Application Platform. Housing prices fluctuate on a daily basis and are sometimes exaggerated rather than based on worth. The major focus of this project is on predicting home prices using genuine factors. Here, we intend to base an evaluation on every basic criterion that is taken into account when establishing the pricing. The goal of this project is to learn Python and get experience in Data Analytics, Machine Learning, and AI.

5. "House Price Forecasting Using Machine Learning" - Alisha Kuvalekar, Shivani Manchewar, Sidhika Mahadik, Shila Jawale

The real estate market is a standout amongst the most focused regarding pricing and keeps fluctuating. It is one of the prime fields to apply the ideas of machine learning on how to enhance and foresee the costs with high accuracy. The objective of the paper is the prediction of the market value of a real estate property. This system helps find a starting price for a property based on the geographical variables. By breaking down past market patterns and value ranges, and coming advancements future costs will be anticipated. This examination means to predict house prices in Mumbai city with Decision tree regressor. It will help clients to put resources into a request without moving towards a broker. The result of this research proved that the Decision tree regressor gives an accuracy of 89%.

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 modeling, Market mix modeling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

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. House Price prediction is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improving Real Estate efficiency.

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted, cost of property depending on the number of attributes considered.

Now as a data scientist our work is to analyze the dataset and apply our skills towards predicting house prices.

Conceptual Background of the Domain Problem

The real estate market is one of the most competitive markets in terms of pricing and the same tends to vary significantly based on a number of factors, forecasting of the price of a property/house is an important factor in the decision making processes for both the buyers and investors in deciding budget allocation, and determining suitable policies.

A 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. You 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 company wants to know:

- Which variables are important to predict the price of a variable?
- 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.

Review of Literature

Machine learning has been used in prediction for some time now with increasingly better results that were put in practice and changed the economic landscape. Practically every economic domain now benefits from machine learning prediction models, and the current models are becoming more accurate given the computational power available for processing

immense sets of data. In this research, the housing price problem is analyzed using several machine learning techniques such as XGBoost, , Random Forest, Linear Regression, and Boosting Regressor.

Park and Bae (2015) addressed the house price prediction problem considering the housing data available for Virginia's Fairfax County. To solve the problem, the authors have employed machine learning techniques such as Naive Bayesian, AdaBoost, and RIPPER to develop a house price classification model.

He and Xia (2020) studied the housing price problem stressing on heterogeneous traders and a healthy urban housing market. Their paper covered the speculative investment effects on house price and economic disturbance and proposed a dynamic stochastic general equilibrium model to solve the problem.

Nam and Seong (2019) studied stock market prediction problems by analyzing media housing market information considering unstructured data and utilizing the asymmetric relationships of firms.

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. Khamis compares the performance of predict house price between Multiple Linear Regression model and Neural Network model in New York.

Díaz et. al. (2019) considered the prediction problem of Spanish day-ahead electricity price. To solve the problem, a regression tree-based approach has been proposed. Moreover, in this problem the dataset, particularly, the model variables are obtained from publicly accessible energy consumption datasets.

In another research, the prediction of the daily bitcoin exchange rate was considered and the behavior of financial markets was studied (Mallqui & Fernandes, 2019). Authors proposed a method including the machine learning features to solve the bitcoin exchange rate prediction problem.

In another study, Gu et. al. (2011) studied the housing price problem with the aim of forecasting a house price model. A hybrid of genetic algorithms and a support vector machine method was proposed to solve the model. The model dealt with a housing dataset that was collected in China, during the 1993-2002 period.

Plakandaras et. al. (2015) also addressed the U.S. real estate house price index problem. In their research, a novel hybrid forecasting method was proposed combining the ensemble empirical mode decomposition (EEMD) with Support Vector Regression (SVR). The obtained solutions of their proposed model are compared with Random Walk (RW), Bayesian Vector Autoregressive, and Bayesian Autoregressive models.

Motivation for the Problem Undertaken

In this project, I have to build a model that calculates 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..

This is my second internship project and the first one in which I have to build a machine learning model. By doing this project I have got an idea about how to deal with data exploration & model building, where I used my analytic skills to predict the house price using ML models.

Further, the model will be a good way for the management to understand the pricing dynamics of a new market. The relationship between house prices and the economy is an important motivating factor for predicting house prices.

ANALYTICAL PROBLEM FRAMING

Mathematical/Analytical Modeling of the Problem

This particular problem has two datasets, one is the train dataset and the other is the test dataset. I have built a model using the train dataset and predicted SalePrice for the test dataset. By looking into the target column, I came to know that the entries of the SalePrice column were continuous and this was a Regression problem so I have to use all regression algorithms while building the model.

- I observed some unnecessary entries in some of the columns & in some columns I found more than 80% null values and more than 85% zero values so I decided to drop those columns. If I choose to keep those columns as it is, it will create high skewness in the model.
- While checking the data for null values, I found many columns with NaN values and I replaced those NaN values with suitable entries like mean for respective numerical columns and mode for respective categorical columns.
- To get a better insight on the features I have used various plots like distribution plot, bar plot, scatter and strip plot. With these plots I was able to understand the relation between the features in a better manner.
- Also, I found outliers and skewness in the dataset so I removed outliers using z-score method and I removed skewness using Power Transformer method.
- I have used all the regression models while building models, then tuned the best model and saved the best model. At last I have predicted the sale price for the test dataset using the saved model of train dataset.

Data Sources & 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 Technologies and it is in Excel format(which I converted into 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 concatenate both train and test data, but this may cause data leakage so I decided to process both the data separately.

Data preprocessing done

- As a first step I have imported required libraries and I have imported both the datasets which were in csv format.
- Then I did all the statistical analysis like checking shape, nunique, value counts, info etc.
- While checking the info of the datasets I found some columns with more than 80% null values, so these columns will create skewness in datasets so I decided to drop those columns.
- Then while looking into the value counts I found some columns with more than 85% zero values this also creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 85% zero values.
- While checking for null values I found null values in most of the columns and I have used an imputation method to replace those null values (mode for categorical column and mean for numerical columns).
- In Id and Utilities column the unique counts were 1168 and 1 respectively, which means all the entries in Id column are unique and ID is the identity number given for particular asset and all the entries in Utilities column were the same so these two columns will not help us in model building. So I decided to drop those columns.
- Next as a part of feature extraction I converted all the year columns to their respective age. Thinking that age will help us more than year.
- And all these steps were performed to both train and test datasets simultaneously.

Data Inputs- Logic- Output Relationships

To analyze the relation between features and target I have done EDA where I analyzed the relation using many plots like bar plot, reg plot, scatter plot, line plot, swarm plot, strip 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 features.
Important features that affect SalePrice positively and negatively.

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, Year_SinceRemodAdded, GarageAge, Year_SinceBuilt, GarageFinish.

Hardware and Software Requirements & Tools used

To build the machine learning projects it is important 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 - language used Python 3

Libraries Used:

- import numpy as np
- import pandas as pd
- import seaborn as sns
- import matplotlib.pyplot as plt
- from sklearn.preprocessing import LabelEncoder
- from sklearn.preprocessing import StandardScaler
- from statsmodels.stats.outliers_influence import variance_inflation_factor
- from sklearn.linear_model.LinearRegression
- from sklearn.ensemble import RandomForestRegressor
- from sklearn.ensemble import GradientBoostingRegressor
- from sklearn.model_selection import cross_val_score
- from sklearn.model_selection import GridSearchCV

DATA ANALYSIS & VISUALIZATION

Identification of possible problem-solving approaches (methods)

I have used imputation methods to replace null values. To encode the categorical columns I have used Label Encoding. To remove outliers I have used the z-score method. And to remove skewness I have used the Power Transformer method. Use of Pearson's correlation coefficient to check the correlation between dependent and independent features. Also I have used standardization. Then followed by model building with regression algorithms.

Testing of Identified Approaches (Algorithms)

Since SalePrice was my target and it was a continuous column so this particular problem was a regression problem. And I have used all regression algorithms to build my model. By looking into the r2 score and cross validation score I found Gradient Boosting Regressor as a best model with high scores. Also to get the best model we have to run through multiple models and to avoid the confusion of overfitting we have to go through cross validation.
Below is the list of regression algorithms I have used in my project.

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor
- Bagging Regressor

Key Metrics for success in solving problem under consideration

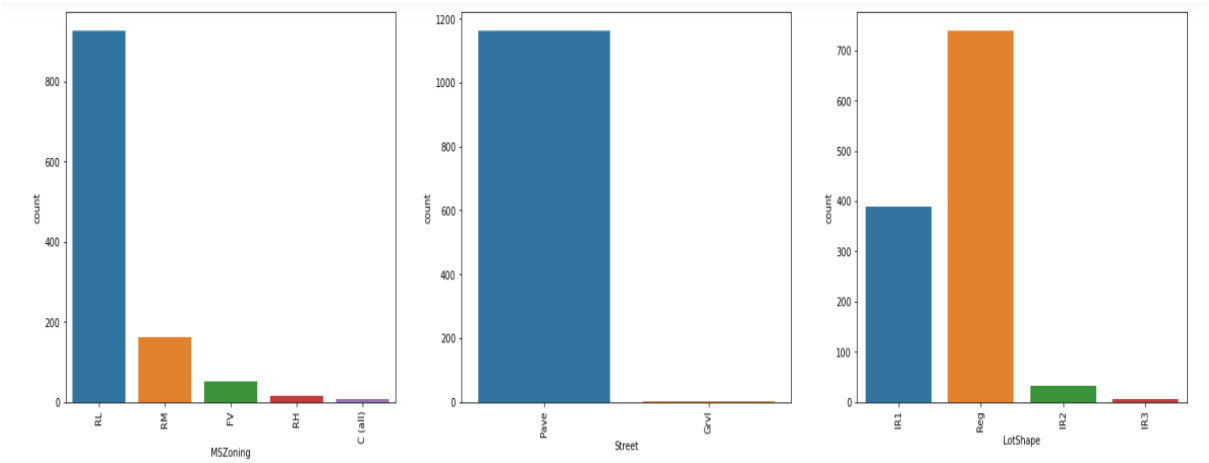
I have used the following metrics for evaluation:

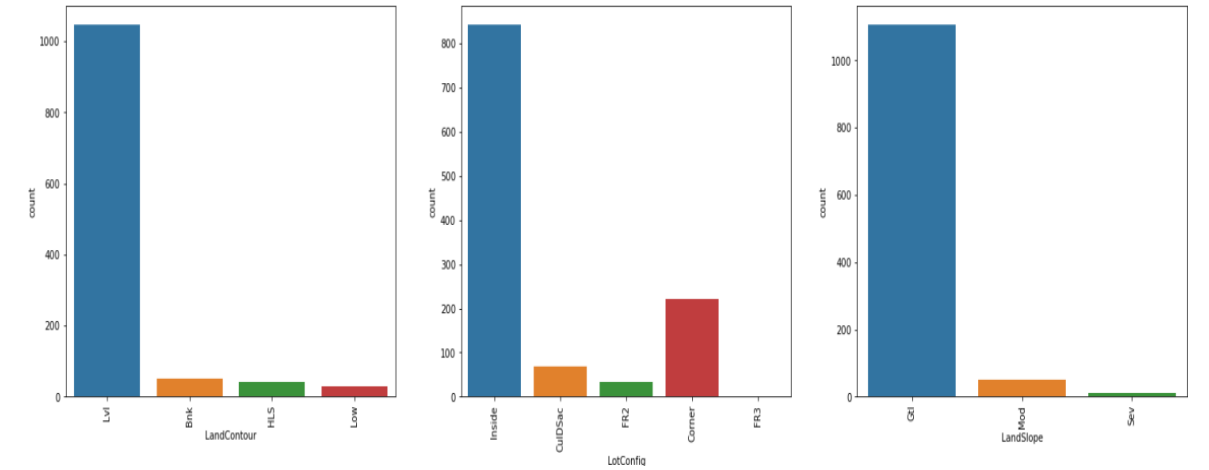
- I have used an r2 score which tells us how accurate our model is.
- I have used mean absolute error which gives a magnitude of difference between the prediction of an observation and the true value of that observation.
- I have used root mean square deviation as one of the most commonly used measures for evaluating the quality of predictions.

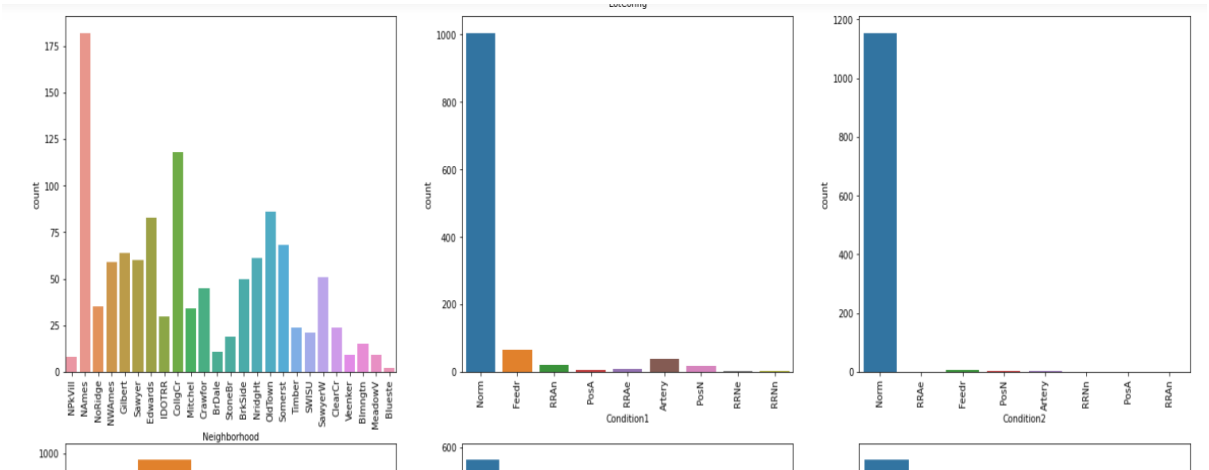
Visualizations

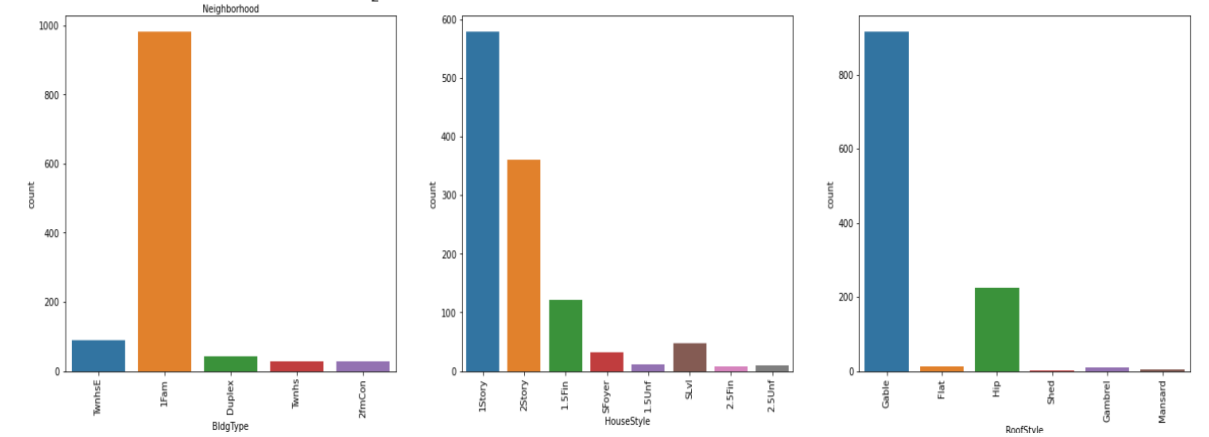
I have used bar plots to see the relation of categorical features and I have used 2 types of plots for numerical columns one is strip plot for ordinal features and other is reg plot for continuous features.

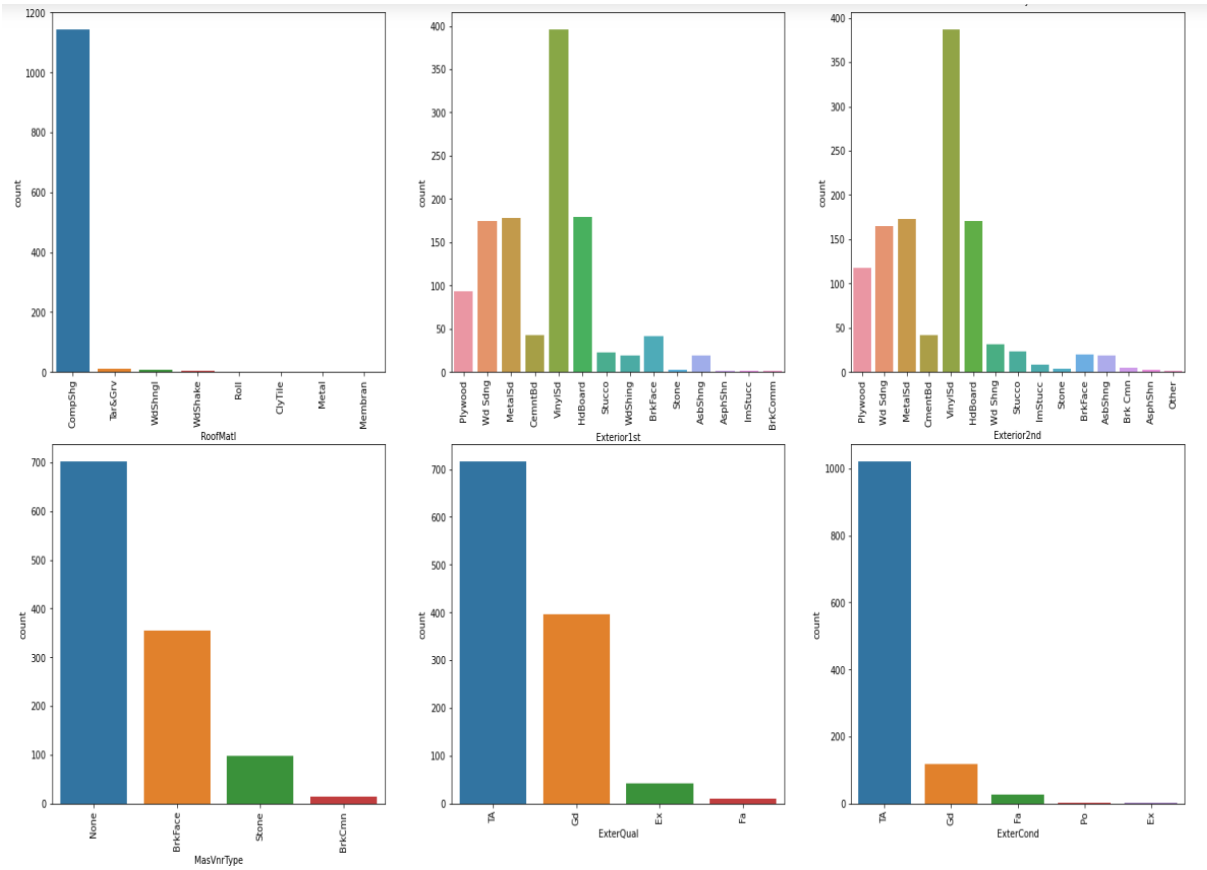
CATEGORICAL COLUMNS:

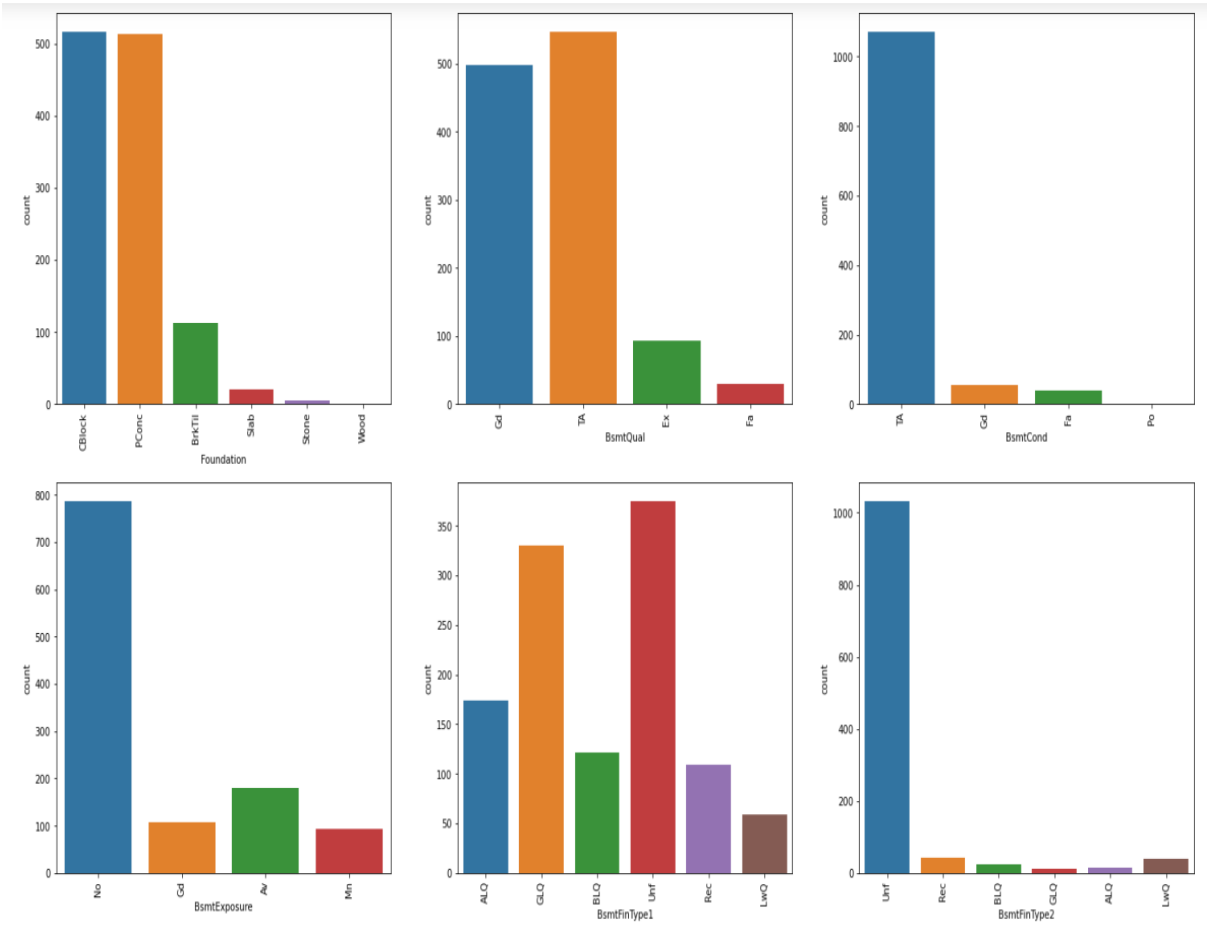

count
MSZoning
RL
RM
FV
RH
C (all)
Street
Pave
Grvl
LotShape
IR1
Reg
IR2
IR3

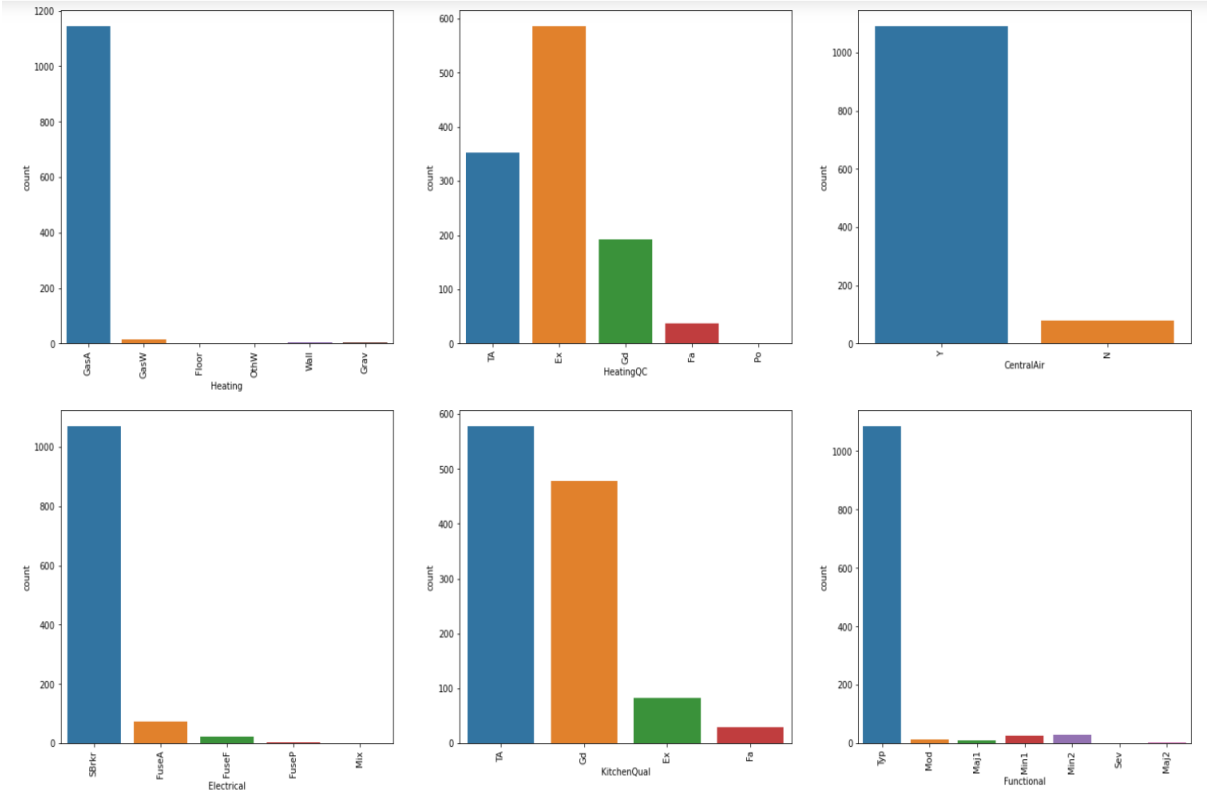

count
LandContour
Lvl
Bnk
HLS
Low
LotConfig
Inside
CulDSac
FR2
Corner
FR3
LandSlope
Gtl
Mod
Sev

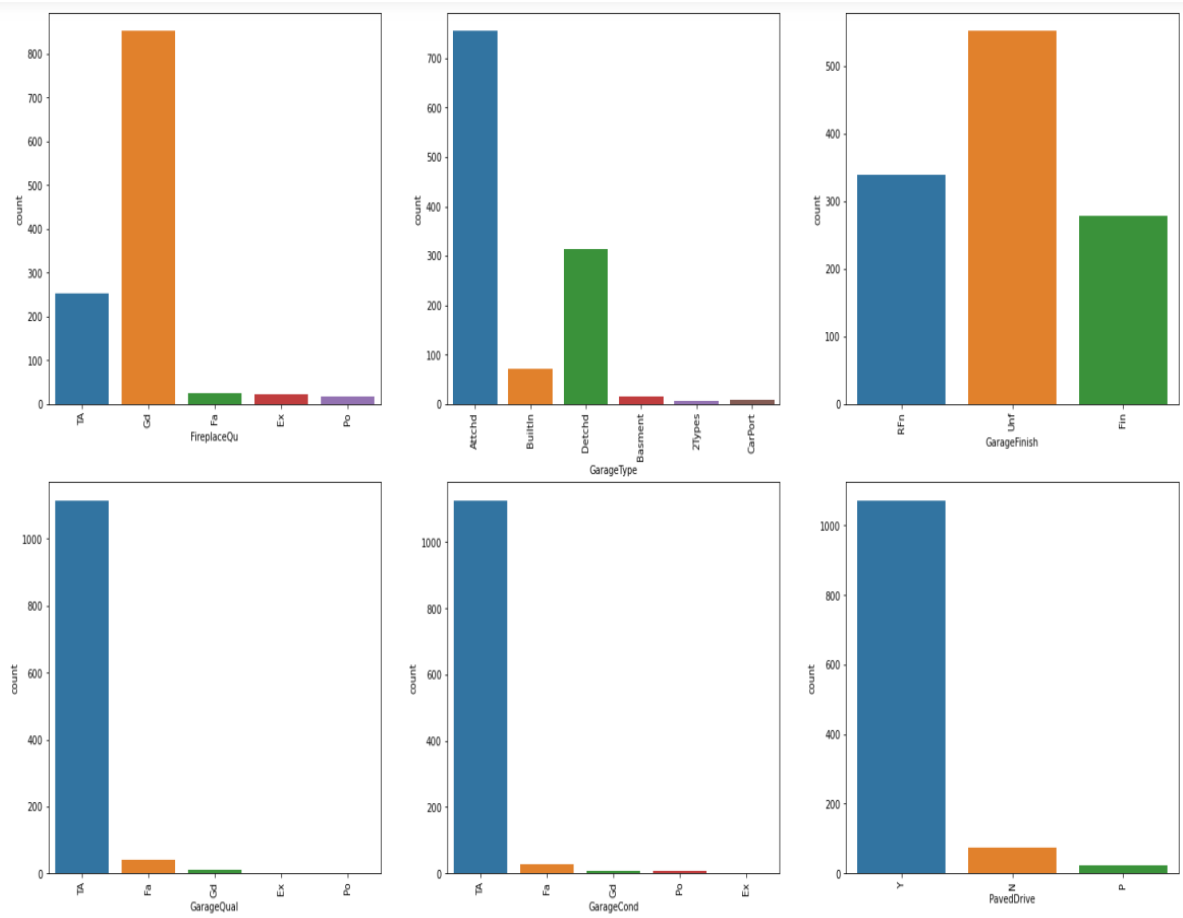

count
Neighborhood
NPkVill
NAmes
NoRidge
NWAmes
Gilbert
Sawyer
Edwards
IDOTRR
CollgCr
Mitchel
Crawfor
Blmngtn
Blueste
BrkSide
NridgHt
OldTown
Somerst
Timber
SWISU
SawyerW
ClearCr
Veenker
Blmngtn
MeadowV
Blueste
Condition1
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
Condition2
Norm
RRAe
Feedr
PosN
Artery
RRNn
PosA
RRAn

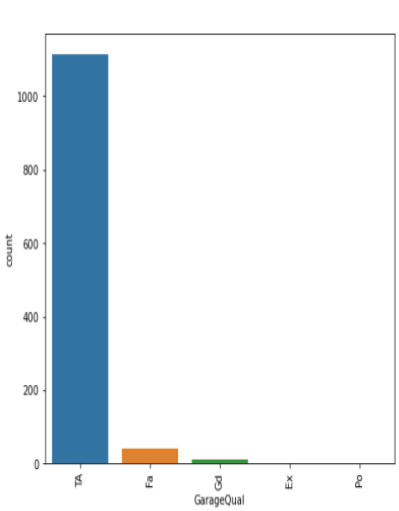

count
BldgType
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
HouseStyle
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
RoofStyle
Gable
Flat
Hip
Shed
Gambrel
Mansard

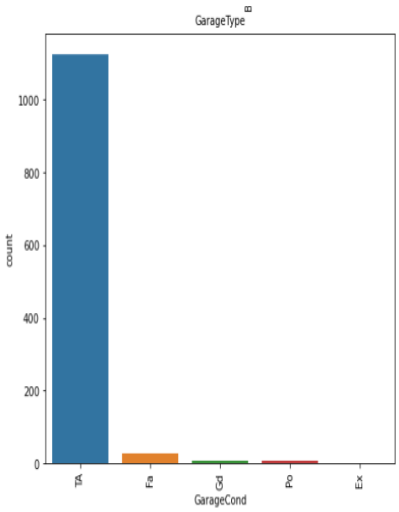

count
RoofMatl
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
Exterior1st
Plywood
Wd Sdng
MetalSd
CemntBd
VinylSd
HdBoard
Stucco
WdShing
BrkFace
Stone
AsbShng
AsphShn
ImStucc
BrkComm
Exterior2nd
Plywood
Wd Sdng
MetalSd
CmentBd
VinylSd
HdBoard
Wd Shng
Stucco
ImStucc
Stone
BrkFace
AsbShng
Brk Cmn
AsphShn
Other
MasVnrType
None
BrkFace
Stone
BrkCmn
ExterQual
TA
Gd
Ex
Fa
ExterCond
TA
Gd
Fa
Po
Ex

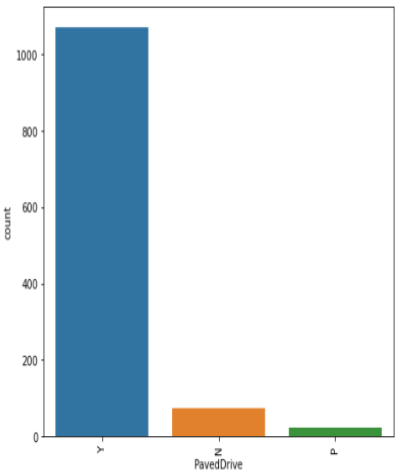

count
Foundation
CBlock
PConc
BrkTil
Slab
Stone
Wood
BsmtQual
Gd
TA
Ex
Fa
BsmtCond
TA
Gd
Fa
Po
BsmtExposure
No
Gd
Av
Mn
BsmtFinType1
ALQ
GLQ
BLQ
Unf
Rec
LwQ
BsmtFinType2
Unf
Rec
BLQ
GLQ
ALQ
LwQ

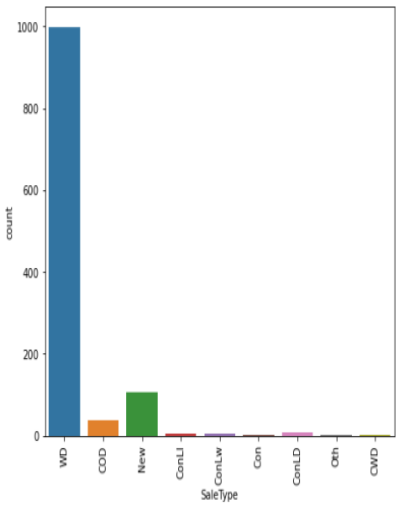

count
Heating
GasA
GasW
Floor
OthW
Wall
Grav
HeatingQC
TA
Ex
Gd
Fa
Po
CentralAir
Y
N
Electrical
SBrkr
FuseA
FuseF
FuseP
Mix
KitchenQual
TA
Gd
Ex
Fa
Functional
Typ
Mod
Maj1
Min1
Min2
Sev
Maj2

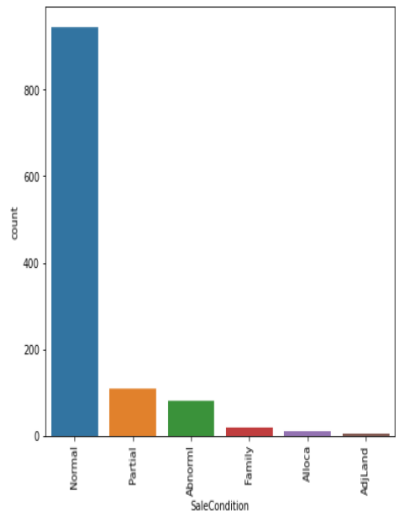

count
FireplaceQu
TA
Gd
Fa
Ex
Po
GarageType
Attchd
BuiltIn
Detchd
Basment
2Types
CarPort
GarageFinish
RFn
Unf
Fin
GarageQual
GarageCond
PavedDrive
Y
N
P

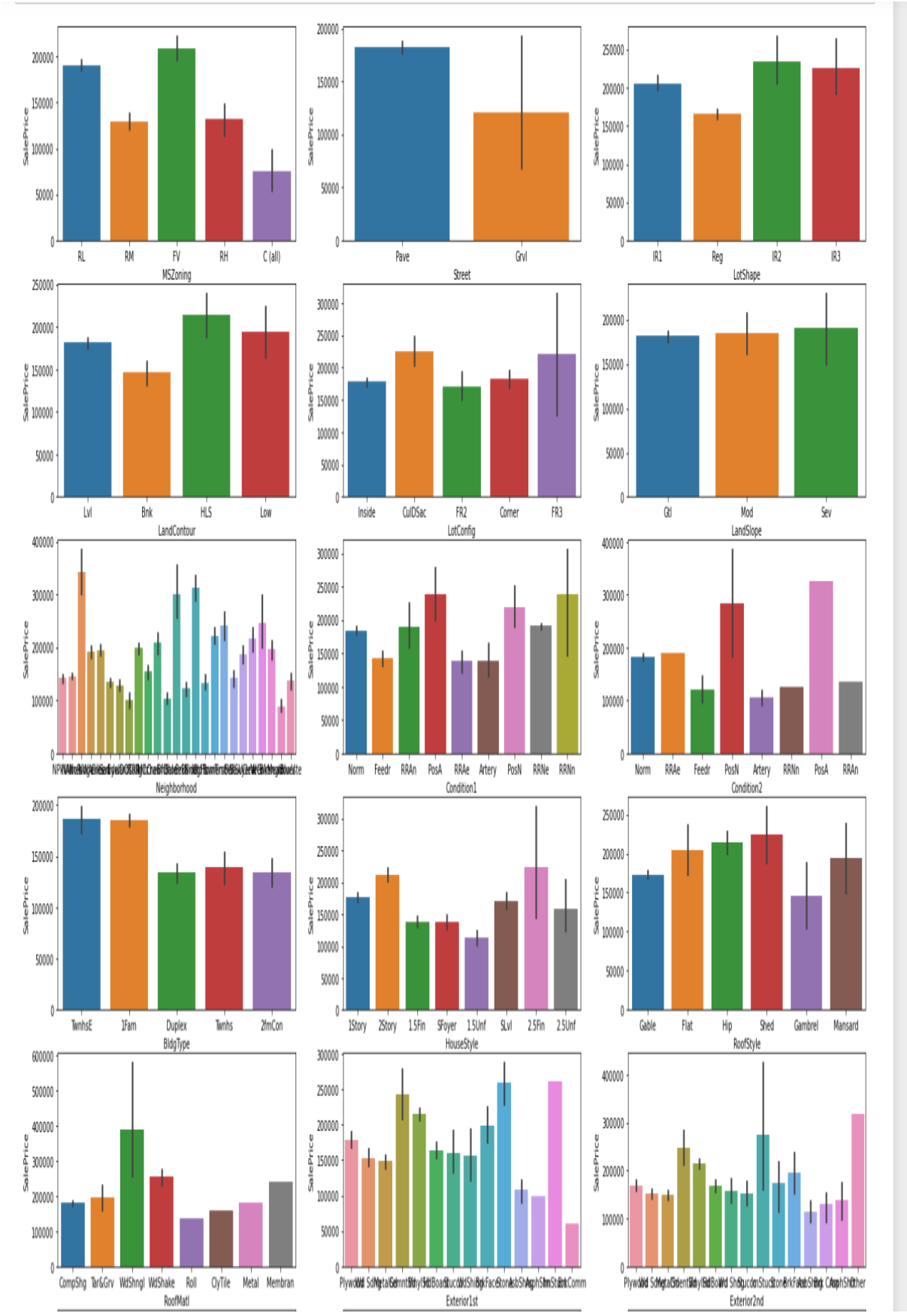

count
1000
800
600
400
200
0
TA
Fa
Gd
Ex
Po
GarageQual

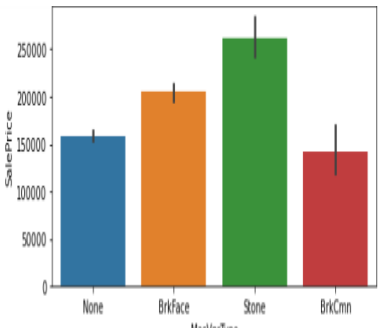

GarageType
count
1000
800
600
400
200
0
TA
Fa
Gd
Po
Ex
GarageCond

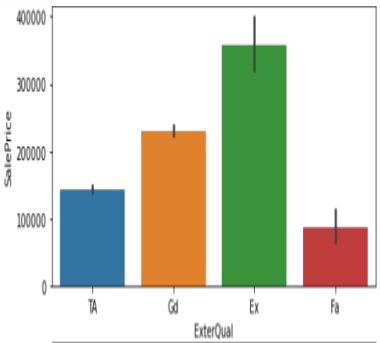

count
1000
800
600
400
200
0
Y
N
P
PavedDrive

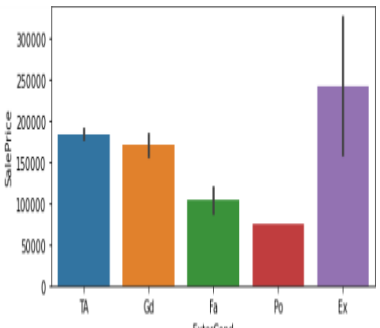

count
1000
800
600
400
200
0
WD
COD
New
ConLI
ConLw
Con
ConLD
Oth
CWD
SaleType

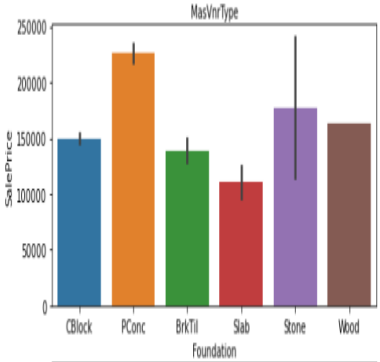

count
800
600
400
200
0
Normal
Partial
Abnorml
Family
Alloca
AdjLand
SaleCondition

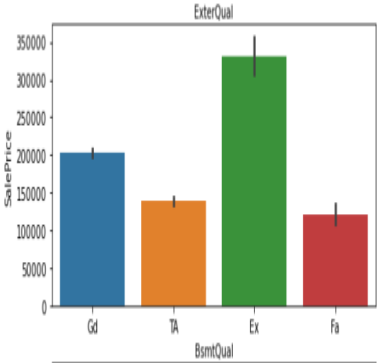

SalePrice
MSZoning
RL
RM
FV
RH
C (all)
Street
Pave
Grvl
LotShape
IR1
Reg
IR2
IR3
LandContour
Lvl
Bnk
HLS
Low
LotConfig
Inside
CulDSac
FR2
Corner
FR3
LandSlope
Gtl
Mod
Sev
Neighborhood
Condition1
Norm
Feedr
RRAn
PosA
RRAe
Artery
PosN
RRNe
RRNn
Condition2
Norm
RRAe
Feedr
PosN
Artery
RRNn
PosA
RRAn
BldgType
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
HouseStyle
1Story
2Story
1.5Fin
SFoyer
1.5Unf
SLvl
2.5Fin
2.5Unf
RoofStyle
Gable
Flat
Hip
Shed
Gambrel
Mansard
RoofMatl
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
Exterior1st
Exterior2nd

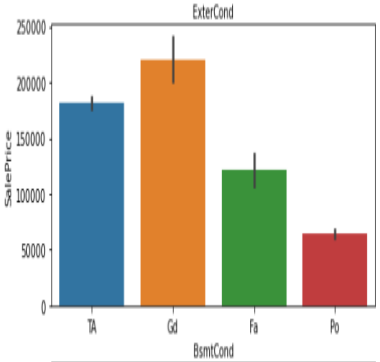

SalePrice
None
BrkFace
Stone
BrkCmn
MasVnrType

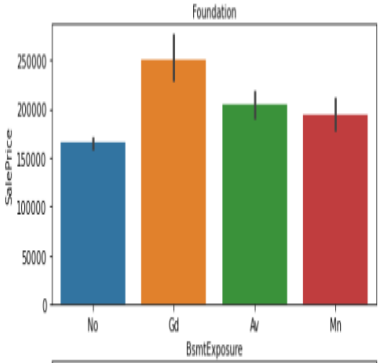

SalePrice
TA
Gd
Ex
Fa
ExterQual

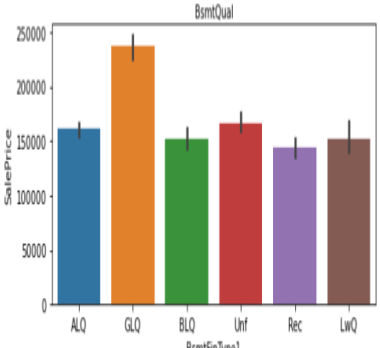

SalePrice
TA
Gd
Fa
Po
Ex
ExterCond

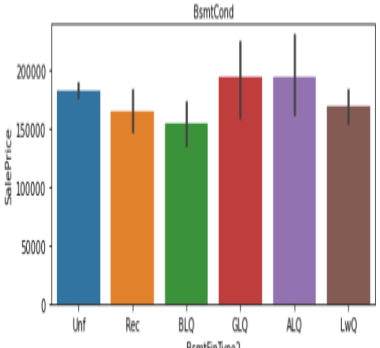

SalePrice
CBlock
PConc
BrkTil
Slab
Stone
Wood
Foundation

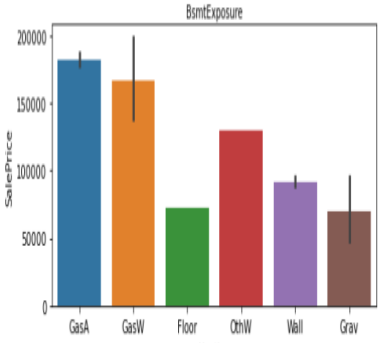

SalePrice
Gd
TA
Ex
Fa
BsmtQual

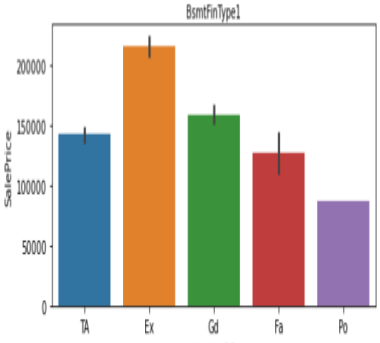

SalePrice
TA
Gd
Fa
Po
BsmtCond

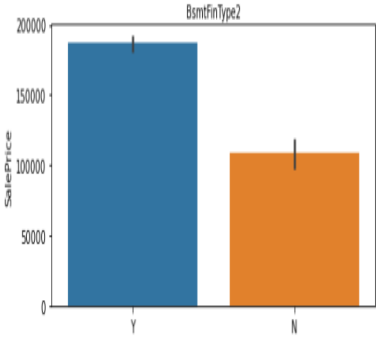

SalePrice
No
Gd
Av
Mn
BsmtExposure

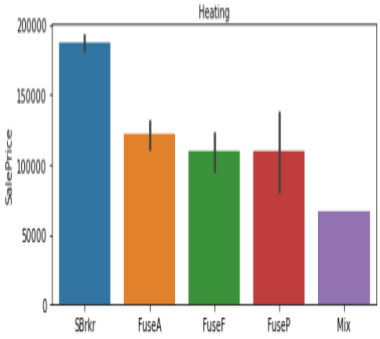

SalePrice
ALQ
GLQ
BLQ
Unf
Rec
LwQ
BsmtFinType1

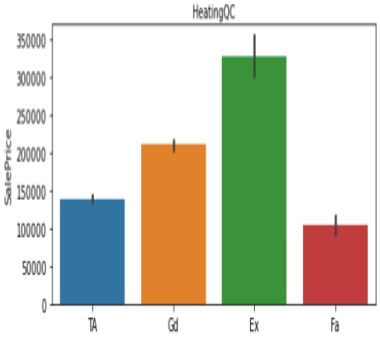

SalePrice
Unf
Rec
BLQ
GLQ
ALQ
LwQ
BsmtFinType2

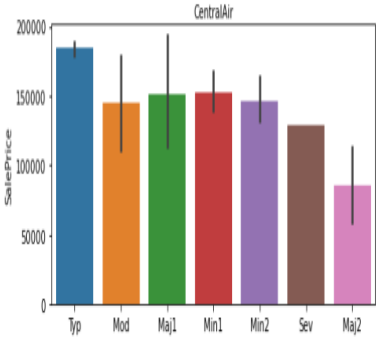

SalePrice
GasA
GasW
Floor
OthW
Wall
Grav
Heating

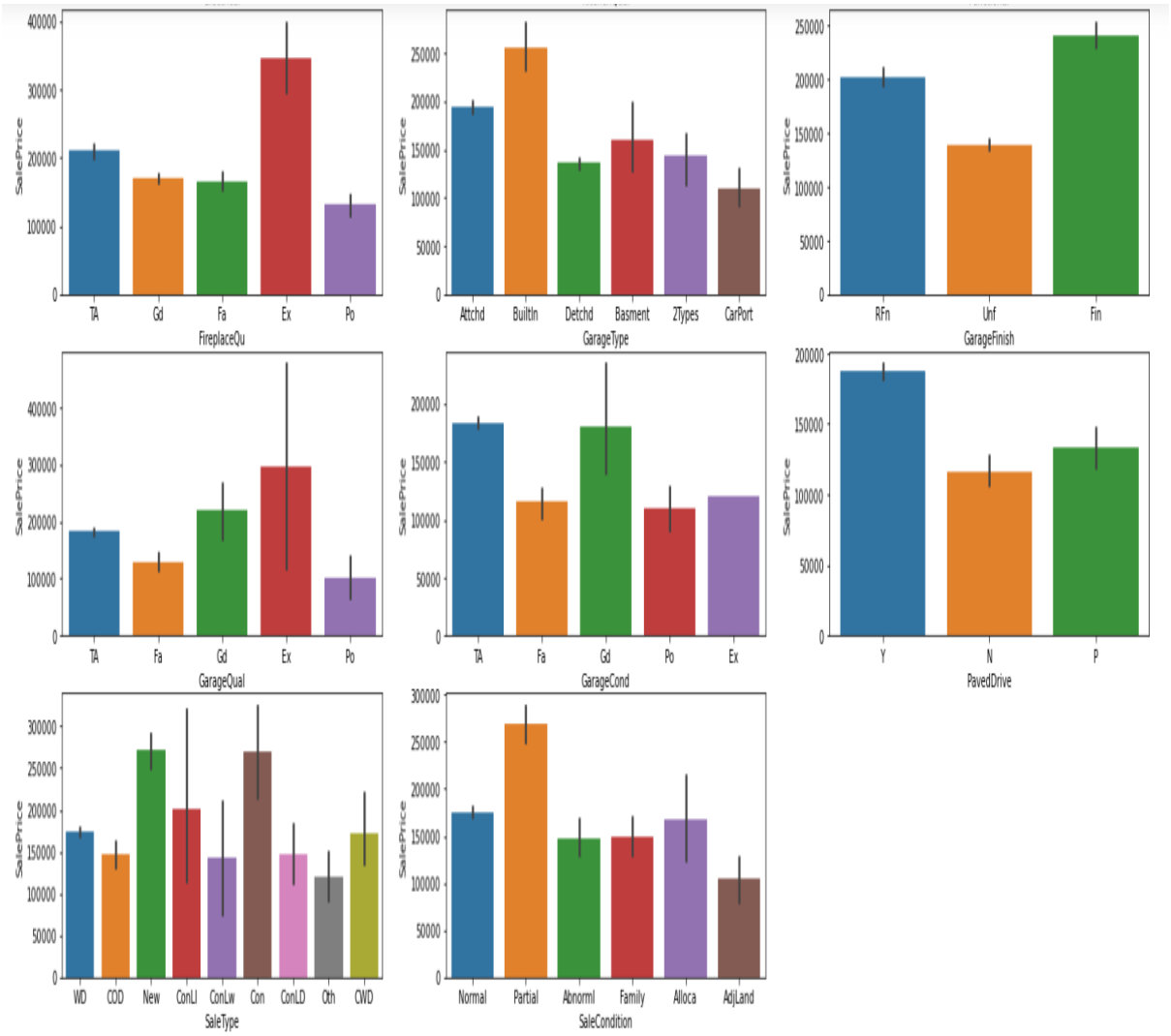

SalePrice
TA
Ex
Gd
Fa
Po
HeatingQC


SalePrice
Y
N
CentralAir


SalePrice
SBrkr
FuseA
FuseF
FuseP
Mix
Electrical


SalePrice
TA
Gd
Ex
Fa
KitchenQual


SalePrice
Typ
Mod
Maj1
Min1
Min2
Sev
Maj2
Functional



Observations from the above visualizations:

SalePrice vs MSZoning: Most of the houses belong to Floating Village Residential followed by Residential Low Density. The houses from this zone have high sale price compared to other zones.

SalePrice vs Street: By observing the bar plot, it is obvious that the property of house with Paved type of road have high SalePrice and the the houses in gravel roads have very less sale price.

SalePrice vs LotShape: Most of the houses having moderately irregular and irregular shape of property have high sale price and houses with regular type of property have less sale price compared to others.

SalePrice vs LandContour: The houses having the hillside and depression property flatness have high sale prices compared to others.

SalePrice vs LotConfig: Most of the houses with Frontage on 3 sides of property have high sale prices compared to others.

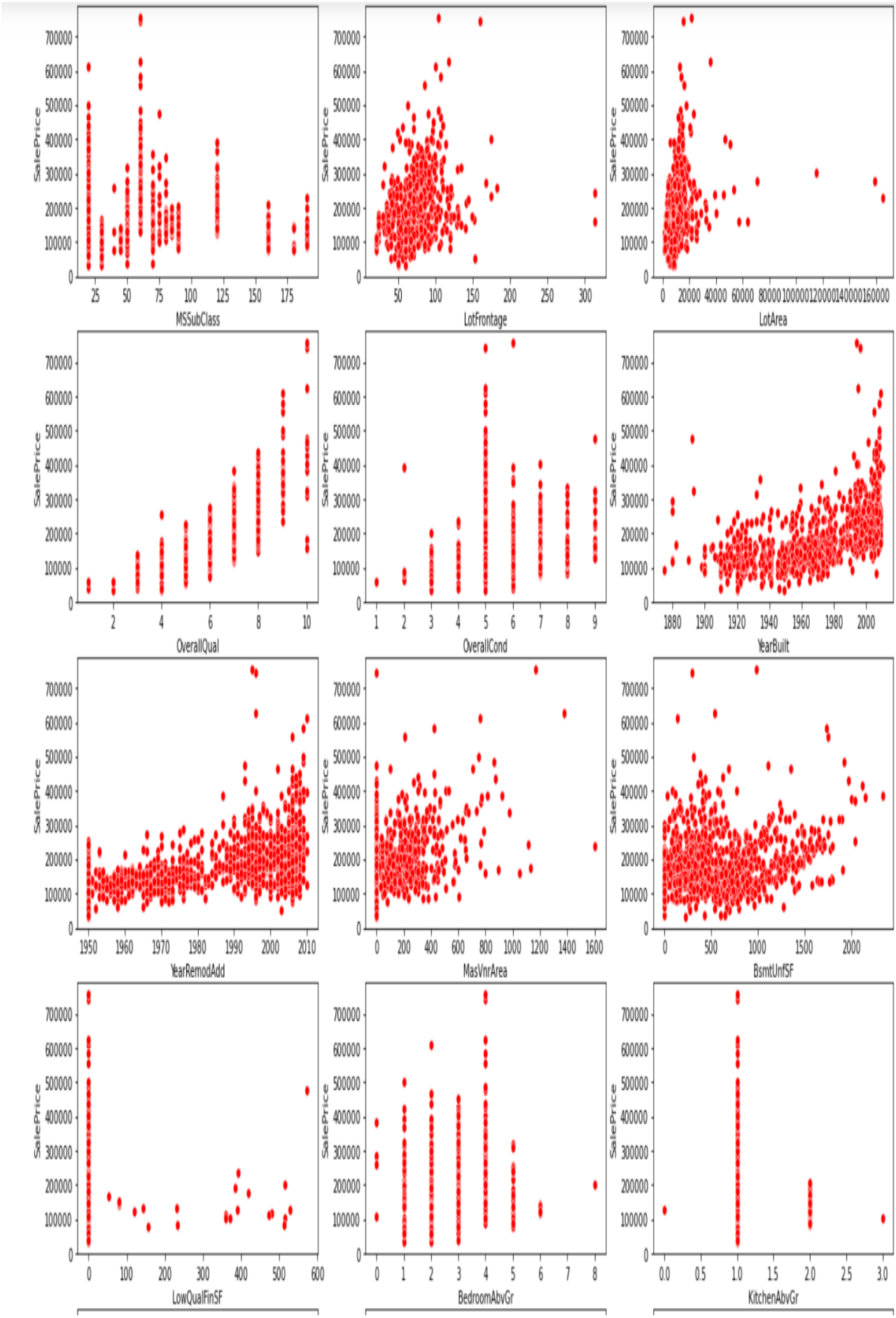
SalePrice vs LandSlope: There is no significant difference between the slope of the property. As we can observe the houses having Gentle slope, Moderate Slope and Severe Slope have the same sale price.

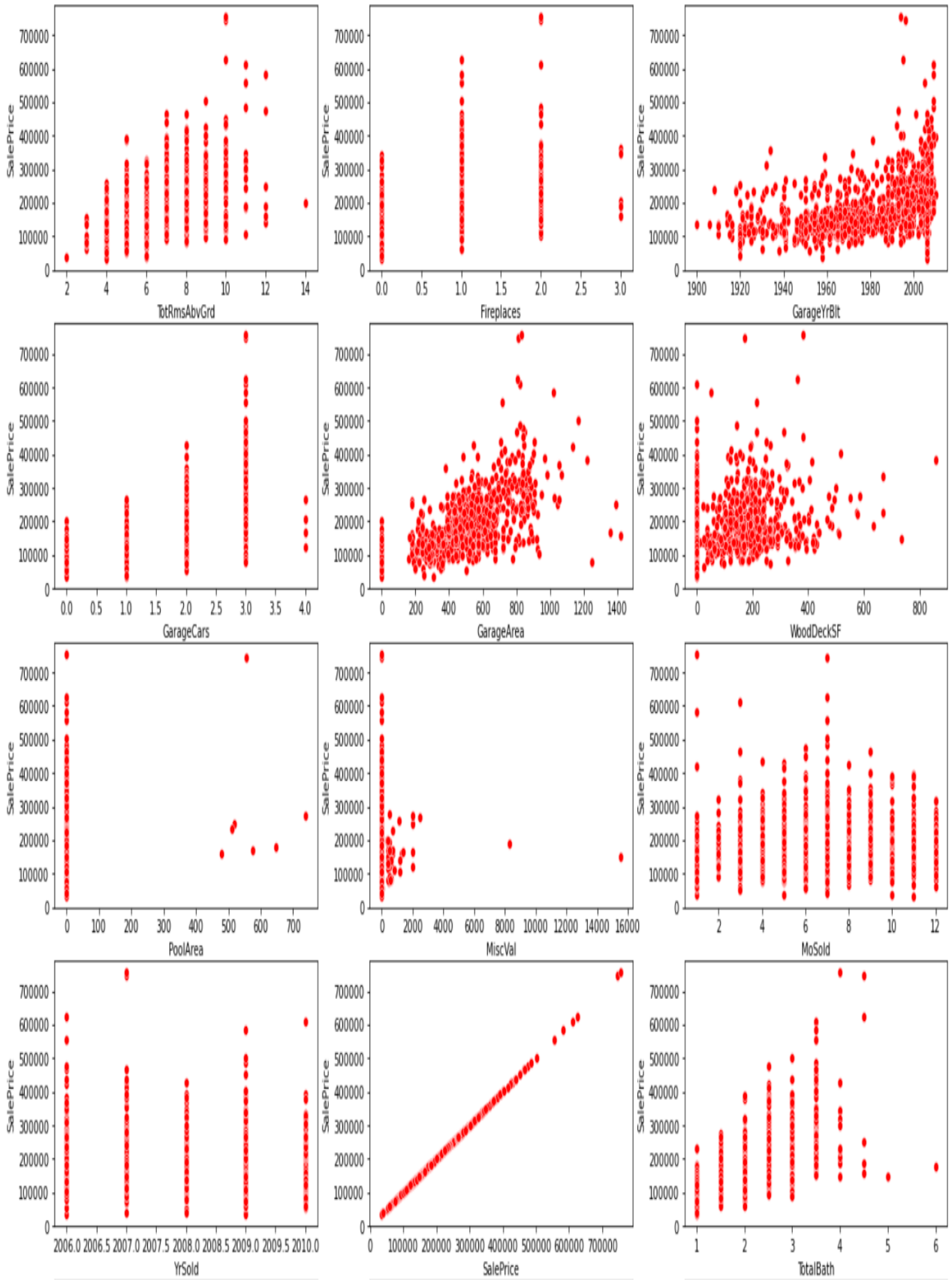
SalePrice vs Neighborhood: The houses which are located near Northridge have high sale prices compared to others.
SalePrice vs Condition1: The houses having the conditions adjacent to positive off-site features and houses within 200' of North-South Railroad have high sale price compared to others.
SalePrice vs Condition2: The houses having the conditions near positive off-site feature park, greenbelt, etc and adjacent to positive off-site feature have high sale price.
SalePrice vs BldgType: Most of the houses are Single-family Detached and Townhouse End Unit and they have higher sale price compared to other categories.
SalePrice vs HouseStyle: Houses which are having style of dwelling 2nd level finished and Two story have high sale price compared to other types.
SalePrice vs RoofStyle: The houses having the roof style Flat, Hip and Shed have high sale price and the houses having gambrel roof style have less sale price.
SalePrice vs RoofMatl: Houses with Wood Shingles root materials have high sale prices.
SalePrice vs Exterior1st: Houses having Imitation Stucco, Stone and Cement Board as 1st exterior cover have high sale price.
SalePrice vs Exterior2nd: Houses having Imitation Stucco and others as 2nd cover have high sale prices.

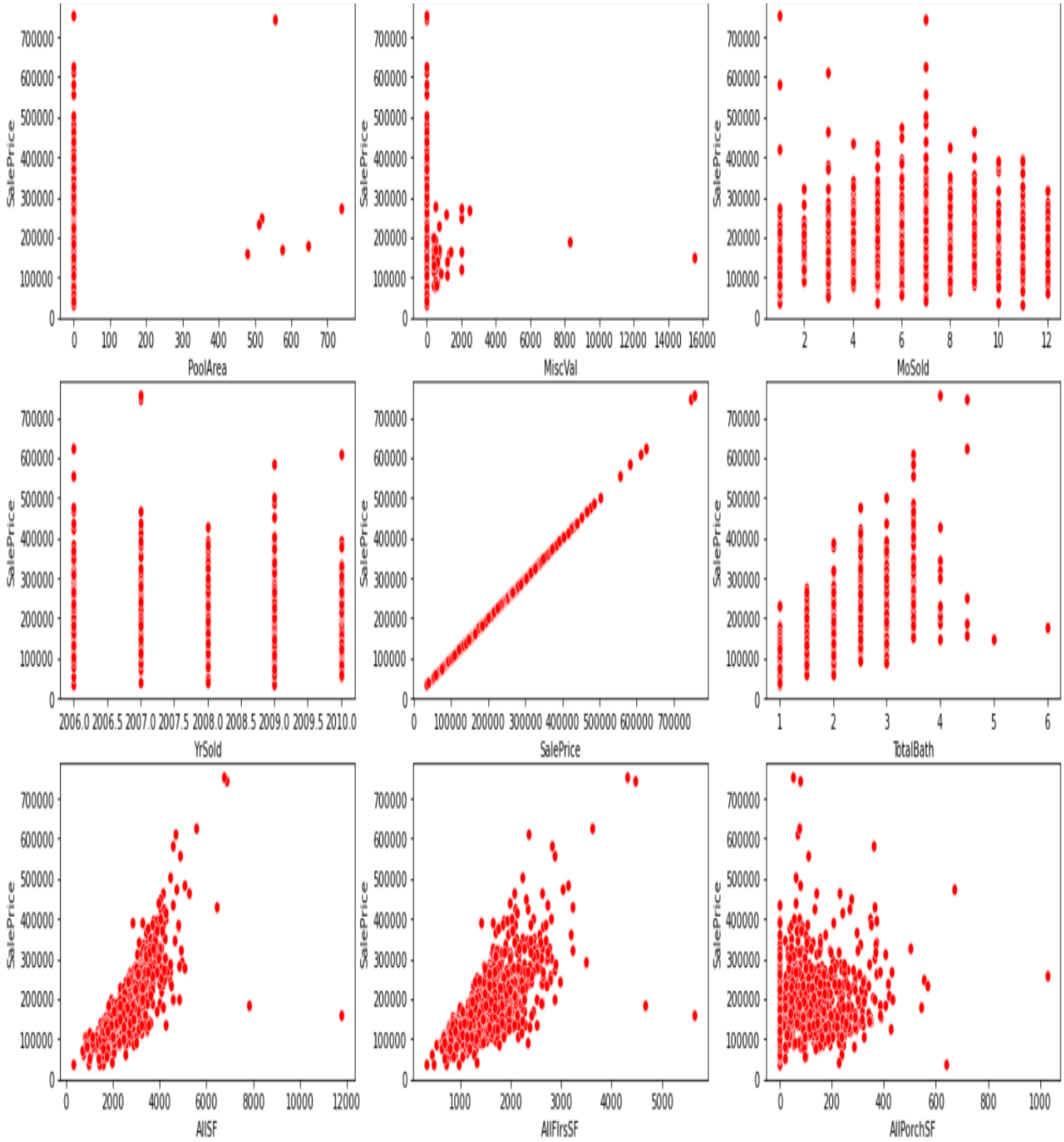
Observations from the above visualizations:
SalePrice vs MasVnrType: Houses having Stone Masonry veneer type have high sale price than other types.
SalePrice vs Foundation: Houses having Poured Concrete as foundation type have high sale prices compared to other types.
SalePrice vs BsmtExposure: Houses having good walkout or garden level walls have high sale prices compared to others.
SalePrice vs BsmtFinType1: The sale price is high for the houses containing good living quarters and basement finished area.
SalePrice vs BsmtFinType2: The sale price is moderately high for the houses having good living quarters and average living quarters.
SalePrice vs Heating: The houses having the heating type gas forced warm air furnace and gas hot water or steam heat have high sale price.
SalePrice vs CentralAir: Most of the houses have central air conditioning so it is obvious that these houses have high sale prices.
SalePrice vs Electrical: Most of the houses having standard circuit breakers & romex have high sale price compared to others.
SalePrice vs Functional: The houses having the typical functionality have maximum sales price and others have average sale price.
SalePrice vs FireplaceQu: The houses having excellent exceptional masonry fireplace quality have high sale price and the houses having poor fireplace quality have very less sale price compared to others.
SalePrice vs GarageType: The houses having built-in garage have high sale price compared to others.
SalePrice vs GarageFinishv: Garages located inside the house which is finished have a high sale price.
SalePrice vs PavedDrive: Houses having paved driveways have high sale prices.
SalePrice vs SaleType: Many houses having sale types as just constructed and sold and Contract 15 % Down payment regular terms have high sale prices.
SalePrice vs SaleCondition: Houses having partial sale condition that is home was not completed when last assessed have high sale price.

Observations from the above visualizations:
SalePrice vs ExterQual: Houses having excellent quality of the material on the exterior have high sale price and houses having fair quality have very less sale price.
SalePrice vs ExterCond: Houses having excellent condition of the material on the exterior have high sale price and the houses having poor condition of the material on the exterior have very less sale price compared to others.
SalePrice vs BsmtQual: The houses which evaluate the excellent quality of height of the basement have high sale price compared to others.
SalePrice vs BsmtCond: The houses which evaluate the good quality of general condition of the basement have high sale price compared to others.
SalePrice vs OverallQual: The houses which have very excellent overall quality like material and finish of the house have high sale price. Also we can observe from the plot as the overall quality of the house increases, the sale price also increases. That is there is a good linear relation between SalePrice and OverallQual.
SalePrice vs OverallCond: The houses having overall condition as excellent and average have very high sale price compared to others.
SalePrice vs HeatingQC: Most of the houses having excellent heating quality and condition have high sale prices.
SalePrice vs KitchenQual: Houses having excellent quality of the kitchen have high sale prices compared to others.
SalePrice vs GarageQual: The sale price of the house is high for the houses having excellent garage quality.
SalePrice vs GarageCond: Houses having typical/average garage condition have high sale price and the houses having good garage condition also have high sales price compared to others.

NUMERICAL COLUMNS:


SalePrice
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtUnfSF
LowQualFinSF
BedroomAbvGr
KitchenAbvGr


SalePrice
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
PoolArea
MiscVal
MoSold
YrSold
SalePrice
TotalBath



Observations from the above visualizations:

LotFrontage: As Linear feet of street connected to property is increasing sales are decreasing and the SalePrice is ranging between 0-3 lakhs.

LotArea: As Lot size in square feet is increasing sales are decreasing and the sale price is in between 0-4 lakhs.

MasVnrArea: As Masonry veneer area in square feet is increasing sales are decreasing and sale price is ranging between 0-4 lakhs.

BsmtFinSF1: As Type 1 finished square feet are increasing , sales are decreasing and the sale price is in between 0-4 lakhs.

BsmtUnfSF: As Unfinished square feet of basement area is increasing sales is decreasing and the sale price is in between 0-4 lakhs. There are some outliers also.
TotalBsmtSF: As Total square feet of basement area is increasing sales is decreasing and the sale price is in between 0-4 lakhs.
1stFlrSF: As First Floor square feet is increasing sales are decreasing and the sale price is in between 0-4 lakhs.
2nd FrSF: As Second floor square feet is increasing sales are increasing in the range 500-1000 and the sale price is in between 0-4 lakhs.
GrLivArea: As Above grade ground living area square feet is increasing sales is decreasing and the sale price is in between 0-4 lakhs.
GarageArea: As Size of garage in square feet is increasing sales is increasing and the sale price is in between 0-4 lakhs.
WoodDeckSF: As Wood deck area in square feet is increasing sales are decreasing and the sale price is in between 0-4 lakhs.
OpenPorchSF: As Open porch area in square feet is increasing sales is decreasing and the sale price is in between 0-4 lakhs.
Year_SinceBuilt: As Year_SinceBuilt is increasing sales are decreasing and the sale price is high for newly built buildings and the sales price is in between 0-4 lakhs.
Year_SinceRemodAdded: As Since Remodel date ,same as construction date if no remodeling or additions, sales are decreasing and the sale price is in between 1-4 lakhs.
GarageAge: As Year garage was built is increasing sales is decreasing and the sale price is in between 0-4 lakhs.

Observations from the above visualizations:
MSSuubClass: For 1-STORY 1946 & NEWER ALL STYLES(20) and 2-STORY 1946 & NEWER(60) types of dwelling the sales are good and SalePrice is also high.
OverallQual: As Rates the overall material and finish of the house is increasing linearly sales is also increasing And SalePrice is also increasing linearly.
OverallCond: For 5, Average overall condition of the house the sales is high and SalePrice is also high.
BsmtFullBath: For 0 and 1 Basement full bathrooms the sales as well as SalePrice is high.
BsmtHalfBath: For 0 Basement half bathrooms the sales as well as SalePrice is high.
FullBath: For 1 and 2 Full bathrooms above grade the sales as well as SalePrice is high.
HalfBath: For 0 and 1 Half baths above grade the sales as well as SalePrice is high.
BedroomAbvGr: For 2, 3 and 4 Bedrooms above grade ,does not include basement bedrooms, the sales as well as SalePrice is high.
KitchenAbvGr: For 1 Kitchens above grade the sales as well as SalePrice is high.
TotRmsAbvGrd: For 4-9 Total rooms above grade, does not include bathrooms, the sales as well as SalePrice is high.
Fireplaces: For 0 and 1 Number of fireplaces the sales as well as SalePrice is high.
GarageCars: For 1 and 2 Size of garage in car capacity the sales is high and for 3 Size of garage in car capacity the SalePrice is high.
MoSold: In between april to august for Month Sold the sales are good with SalePrice.
Year_SinceSold: For all the Year_SinceSold the sale price and sales both are same.

MODEL BUILDING & PREDICTION

Run and Evaluate selected models

Regression Models:

LINEAR REGRESSION:

linear regression

```
In [113]:
1 lr = LinearRegression()
2 lr.fit(x_train,y_train)

Out[113]:
LinearRegression
LinearRegression()

In [114]:
1 y_pred=lr.predict(x_test)

In [115]:
1 a=r2_score(y_test,y_pred)

In [116]:
1 c=cross_val_score(lr,x_scaled,y,cv=5).mean()

In [117]:
1 print("r-2 score : ",a,"\n","cross validation score :",c)

r-2 score :  0.8887280691750609
 cross validation score : 0.8742891322217456
```

The Linear Regression model gave us an R2 Score of 88.87 %.
CV score is 87.4%

BAGGING REGRESSOR:

BaggingRegressor

```
In [101]:
1 bg=BaggingRegressor()
2 bg.fit(x_train,y_train)

Out[101]:
BaggingRegressor
BaggingRegressor()

In [102]:
1 y_pred=bg.predict(x_test)

In [103]:
1 a=r2_score(y_test,y_pred)

In [104]:
1 c=cross_val_score(bg,x_scaled,y,cv=5).mean()

In [105]:
1 print("r-2 score : ",a,"\n","cross validation score :",c)

r-2 score :  0.8825879865686066
 cross validation score : 0.8550233190686122
```

Bagging Regressor gives us R2 score of 88.25
CV score is 85.5 %

RANDOM FOREST REGRESSOR:

RandomForestRegressor

In [107]:
```
rf=RandomForestRegressor()
rf.fit(x_train,y_train)
```

Out[107]:
```
RandomForestRegressor
RandomForestRegressor()
```

In [108]:
```
y_pred=rf.predict(x_test)
```

In [109]:
```
a=r2_score(y_test,y_pred)
```

In [110]:
```
c=cross_val_score(rf,x_scaled,y,cv=5).mean()
```

In [111]:
```
print("r-2 score : ",a,"\n","cross validation score :",c)
```

```
r-2 score :  0.890304596368356
 cross validation score : 0.874249513105178
```

The Random Forest Regression Model gave us a R2 Score of 89.03 %.
CV score is 87.4%

GRADIENT BOOSTING REGRESSOR:

GradientBoostingRegressor

```
In [95]:
1 gb = GradientBoostingRegressor()
2 gb.fit(x_train,y_train)
```

```
Out[95]:
▾ GradientBoostingRegressor
GradientBoostingRegressor()
```

```
In [96]:
1 y_pred = gb.predict(x_test)
```

```
In [97]:
1 a=r2_score(y_test,y_pred)
```

```
In [98]:
1 c=cross_val_score(gb,x_scaled,y,cv=5).mean()
```

```
In [99]:
1 print("r-2 score : ",a,"\n","cross validation score :",c)
```

```
r-2 score :  0.9066691268179792
 cross validation score : 0.8901505591292402
```

The Gradient Boosting Regressor Model gave us a R2 Score of 90.66 %.

CV score is 89%

Hyper Parameter Tuning:

Since the R2 Score & Cross Validation Score are both highest in Gradient Boosting Regressor we shall consider it for hyper parameter tuning.

We will use GridSearchCV for hyper parameter tuning.

Hyper Parameter Tuning

In [123]:
```
param_search ={"n_estimators":[500,600,700],
               "random_state":[20,40,50],
               "max_features":[20,25,17]}
```

In [124]:
```
grid_search = GridSearchCV(gb,param_grid=param_search,cv=4)
```

In [125]:
```
grid_search.fit(x_train,y_train)
```

Out[125]:

GridSearchCV
estimator: GradientBoostingRegressor
GradientBoostingRegressor

In [126]:
```
grid_search.best_params_
```

Out[126]:
```
{'max_features': 17, 'n_estimators': 500, 'random_state': 20}
```

In [127]:
```
gb = GradientBoostingRegressor(n_estimators=500,random_state=20,max_features=17)
```

In [128]:
```
gb.fit(x_train,y_train)
```

Out[128]:

GradientBoostingRegressor
GradientBoostingRegressor(max_features=17, n_estimators=500, random_state=20)

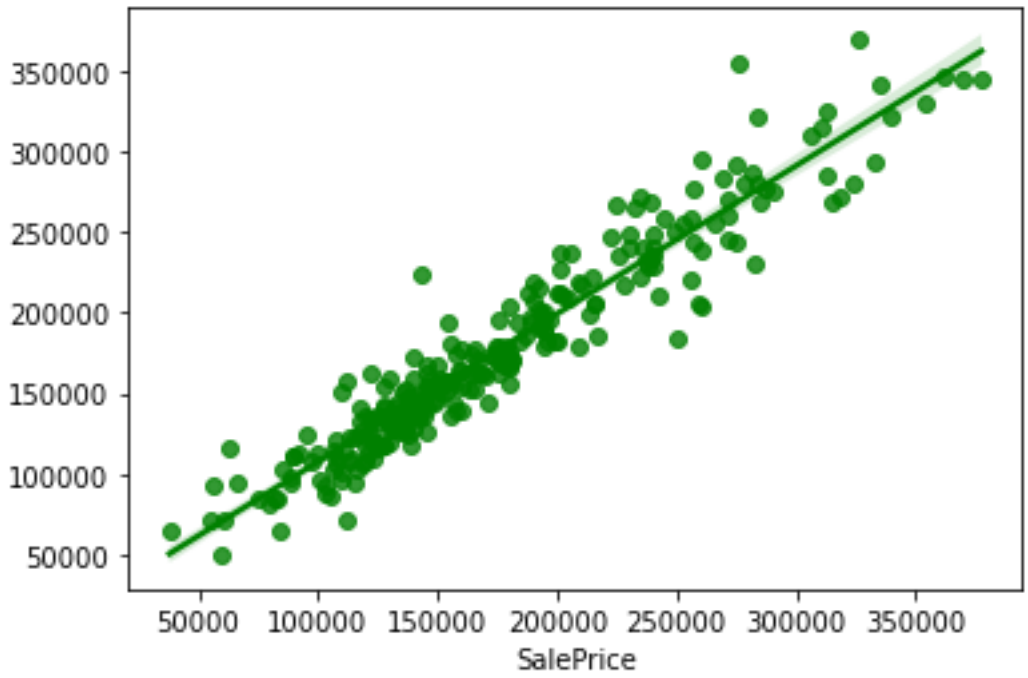
In [129]:
```
y_pred = gb.predict(x_test)
```

In [130]:
```
print("r-2 score : ",r2_score(y_test,y_pred),"\n","cross validation score :",cross_val_score(gb,x_scaled,y,cv=4).mean())
```

```
r-2 score :  0.912038865116276
 cross validation score : 0.8923767067158646
```

After Hyper Parameter Tuning, we have got a better R2 score of 91.2 %.

Visualizing the Predicted values :



Saving the final model and predicting the saved model

Now we shall save the best model.
Load the saved model and predict the sale price.

saving model

```
In [133]:
1 import pickle
```

```
In [134]:
1 import pickle
2 filename='house_rent.pk1'
3 pickle.dump(gb,open(filename,'wb'))
```

```
In [135]:
1 predicted_price  = gb.predict(df_test)
```

```
In [136]:
1 # saving the price in dataframe
2 Price_predicted = pd.DataFrame(predicted_price)
```

```
In [137]:
1 Price_predicted
```

10	361050.841262
11	320212.562966
12	341108.924621
13	325202.431272
14	336771.122846
15	332585.802623
16	372477.225365
17	354458.519728
18	328807.877802
19	330076.228217
20	341349.616315
21	351111.045109
22	347010.912968

Predicting SalePrice of house for test dataset using saved trained model

Above are the predicted sale price of the house for the test data set. We have created a data frame for the predicted result which is as above.

Interpretation of the Results

This dataset was very special as it had separate train and test datasets. We had to work with both datasets simultaneously. Firstly, the datasets contained null values and zero entries in a large number of columns, so we had to be careful while going through the statistical analysis of the datasets.

Secondly, proper plotting using various visualization techniques for the different types of features helped us to get better insight on the data. There were a lot of numerical continuous columns in linear relationship with the target column.

Thirdly, I noticed a huge amount of outliers and skewness in the data so we chose the proper methods to deal with the outliers and skewness. If we ignore these outliers and skewness we may end up with a bad model which has less accuracy.

Then, scaling both train and test dataset has a good impact like it will help the model not to get baised.

Lastly, we have to use multiple regression algorithms while building various models using a train dataset to get the best model out of it. And we have to use multiple metrics like mae, mse, rmse and R2 Score which will help us to decide the best model.

I found ExtraTreesRegressor as the best model with 87.39 % as its R2 Score. Then, I improved the accuracy of the best model by running hyper parameter tuning which increased the R2 Score slightly to 87.82 % .

Finally, I have predicted the SalePrice for the test dataset using the saved model of train dataset. I was able to get the predictions near to actual values.

CONCLUSION

Key Findings and Conclusions of the Study

In this project report, we have used machine learning algorithms to predict the house prices. We have mentioned the step by step procedure to analyze the dataset and find the correlation between the features. Thus, helping us select the features which are not correlated to each other and are independent in nature. These feature sets were then given as an input to 8 algorithms. We calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the data frame of predicted prices of the test dataset.

Learning Outcomes of the Study in respect of Data Science

New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation. It has made me understand what the data is trying to say.

Data cleaning is one of the most important steps to remove missing values and to replace null values and zero values with their respective mean, median or mode. This study is an exploratory attempt to use 8 machine learning algorithms in estimating housing prices, and then compare their results.

The application of machine learning in property research is still at an early stage. We hope this study has helped move a small step ahead in providing some methodological and empirical contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices.

Limitations of this work and Scope for Future Work

LIMITATIONS:

- First drawback is the data leakage when we merge both train and test datasets.
- Second drawback was the null values and the zero values, which was dealt with using imputation techniques.
- Third drawback was the number of outliers and the amount of skewness; these two tend to reduce our model accuracy. We have tried our best to deal with outliers & skewness.
- The dataset has many limitations, the main limitation is that we have no information about 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 analyze the data more clearly.
- So it looks quite good that we have achieved an accuracy of 87..82 % even after dealing with all these drawbacks.
- Also, this model doesn't predict future prices of the houses mentioned by the customer.
- Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process.

FUTURE WORK:

- One of the major future scopes is adding an estate database of more cities,which will provide the user to explore more estates and reach an accurate decision.