House Price Prediction

Submitted by:

Raj Sharma

ACKNOWLEDGMENT

I wish to express my sincere thanks to the following companies, without whom I would have not got opportunity to work on this project; Data Trained Institute and Flip Robo Technology

INTRODUCTION

- ## 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. Our problem is related to one such housing company.

- ## Conceptual Background of the Domain Problem

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 data is provided in the CSV file below.

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 variable?
- How do these variables describe the price of the house?

- ## Review of Literature

Motivation for the Problem Undertaken

Housing and rental price always continue to rise. After the housing crisis of 2008, housing prices have recovered remarkably well. Some times it fluctuates based on the political party play. In order to maintain transparency among customers and also comparison can be made easy through this model. If customer finds the price of house at some website higher that the price predicted by the model, then he can reject that house.

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

House price prediction problem consist of 2 datasets Training and Testing dataset. Trained dataset to train the model and Testing to predict the model output.

First, the analysis is started with importing the data, this dataset contains lot off Null values which are cleaned by using mean(), median(), mode() functions and unnecessary columns which does not contribute for the target variable is removed.

Once data is cleaned outliers and skewness are checked, if present they are removed then in Data Pre-processing, Standard Scaler is used to standardize the data and by using VIF multicollinearity is removed.

Once this is all done then data is ready for modelling. As the target variable contains continuous data I did regression. 4 regressors – Decision Tree Regressor, Support Vector Regressor, KNeighbors Regressor, Linear Regression, 4 ensemble -Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, 3 metrics – r2_score, mean_squared_error, mean_absolute_error and 3 regularization – Lasso, Ridge, ElasticNet techinques are used in this project to build Regression model. From this the best model is identified and done Cross Validation technique and Hypertuning is done to increase accuracy. Then, Finally the best model is saved.

By using this train model, output is predicted for test dataset.

- ## Data Sources and their formats

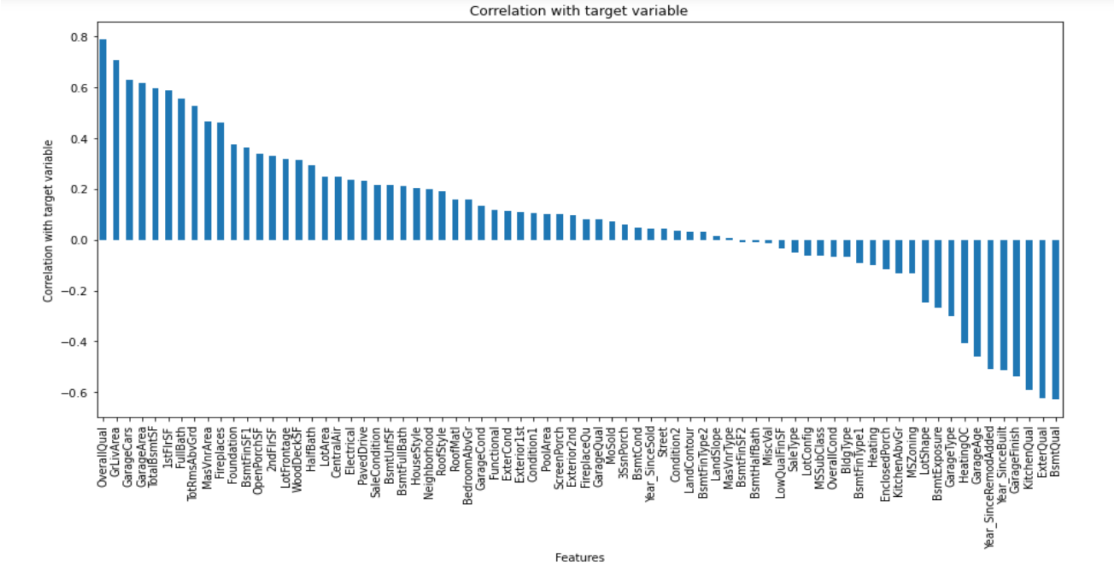
The data was provided by Surprise Housing Company which is in csv format.

- Data contains 1460 entries each having 81 variables.
- Here there are two dataset provided one is training and other is testing. Training dataset contains feature variable while the target housing price need to be predicted in testing data

- ## Data Preprocessing Done

In Machine Learning, data pre-processing refers to the process of cleaning and organising raw data in order to make it appropriate for creating and training Machine Learning models. In other words, anytime data is received from various sources, it is collected in raw format, which makes analysis impossible. Data pre-processing is a crucial stage in Machine Learning since the quality of data and the relevant information that can be gleaned from it has a direct impact on our model's capacity to learn; consequently, we must pre-process our data before feeding it into our model. As a result, it is the first and most important stage in developing a machine learning model.
Some of the techniques used in this project are listed below:

1. Started with importing the required libraries and then the train dataset which is in csv format.
2. As the dataset contains 81 variables to display all the columns so in this pd.options.display.max_columns method is used by which all the columns can be observed. Then the information of the columns are observed carefully.
3. The unique values is then found by which one can check if any unnecessary values like ? , / is present in the dataset and if present it can be dropped or replaced based on the use it has.
4. Few columns like 'Id', 'Utilities' is dropped as it has all unique values in it and does not contribute much in model building.
5. Then Null values is checked, where the dataset contains large number of null values. In this values which have more that 90% of zeros(0) is dropped as we cannot replace it with mean(), mode() values. Values in some columns are replaced with mean() and mode() as they are of numeric and categorical type.
6. Numeric data and Categorical data is then separated so that visualization are done with histplot for numeric data and countplot for categorical data.
7. Then all the categorical data is converted to numeric by using LabelEncoder so that analysis can be made in better way.
8. Data Description is made followed by correlation where positive and negative correlated data is checked.



In this project correlation is done with target variable 'SalePrice', by which we can observe that OverallQual, GrLivArea is highly positively correlated with target column, whereas BsmtQual, ExterQual is highly negatively correlated with target column.

9. Outliers and Skewness is checked and removed in order to avoid bias while model building.
10. Standard Scaler is used for data standardization and VIF is used to check Multicolinearity is checked to find if any data variable is correlated with each other and it is removed.

11. Once all these processes are done then data is ready for model building where various Machine Learning models is used to check the accuracy of data.

- # Data Inputs- Logic- Output Relationships

MSSubClass 60 – [2-STORY 1946 & NEWER] and 70- [2-STORY 1945 & OLDER] is the highest segment of building that is sold in the marker which means buyers mostly wish to buy these dwelling in the market, FV is Floating Village Residential which is being highly sold and RL- Residential Low Density being the costliest in the market. We can understand that low residential density which might be of posh residential area which are costlier in the market across all the other classifications of residents. IR1 - Slightly irregular being the costliest lot shape, followed by Reg – Regular. IR2 - Moderately Irregular are the highest sold lot shape. So, from above we can understand people are mostly interested in buying irregular shaped lot more, than the regular shaped lot as it is little costlier. And also, the availability of the Regular shaped plot is less compared to Irregular plot which may also be one of the reasons to note.

We can see one family type building being the costliest in the market and also has more buildings are sold. TwnhsE - Townhouse End Unit is second highest sold Building Type. According to the Neighborhood NoRidge that is Northridge is being sold high and also costliest in the market. NridgHt - Northridge Heights is the next costliest sold property with respect to neighbourhood.

Since, Utilities have only one values in all the columns it has no correlation. we will drop this column since it won't help in building the model.

Overall Quality yearbuilt year remodified have high corelation with sales prize.

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

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

SBrkr is Standard Circuit Breakers & Romex and also this is considerably safer than any other electrical circuits in industry. Standard Circuit Breakers & Romex is getting

sold higher and also the costliest across all the electrical systems. From above we can say that properties are built as with more safety and more reliable electrical equipment's over other. There was a saying the Quality of the kitchen is the beauty of the house, as similar to that we can see the excellent quality in kitchen will increase the cost of the property. And also, the excellent quality of kitchens is being mostly build. Good Quality in kitchen stands second in the order and also in number of units sold. From above we can say that people mostly preferred good quality kitchens and also good and excellent quality of kitchens are being costlier.

We can observe according to the correlations the scatterplot points are distributed. We can see high positive correlation of feature variables with the sales prize. We can see Garage Area, garage cars, garage area built, Fireplace, total rooms available, full bath, living area, 1stFlrSf are more positive corelated which means the increase in the above will also increase the cost and selling price of the property. KitchenQual, Garagetype, GarageFinish are negatively correlated and the increase in that will decrease the cost and the selling price of the property.

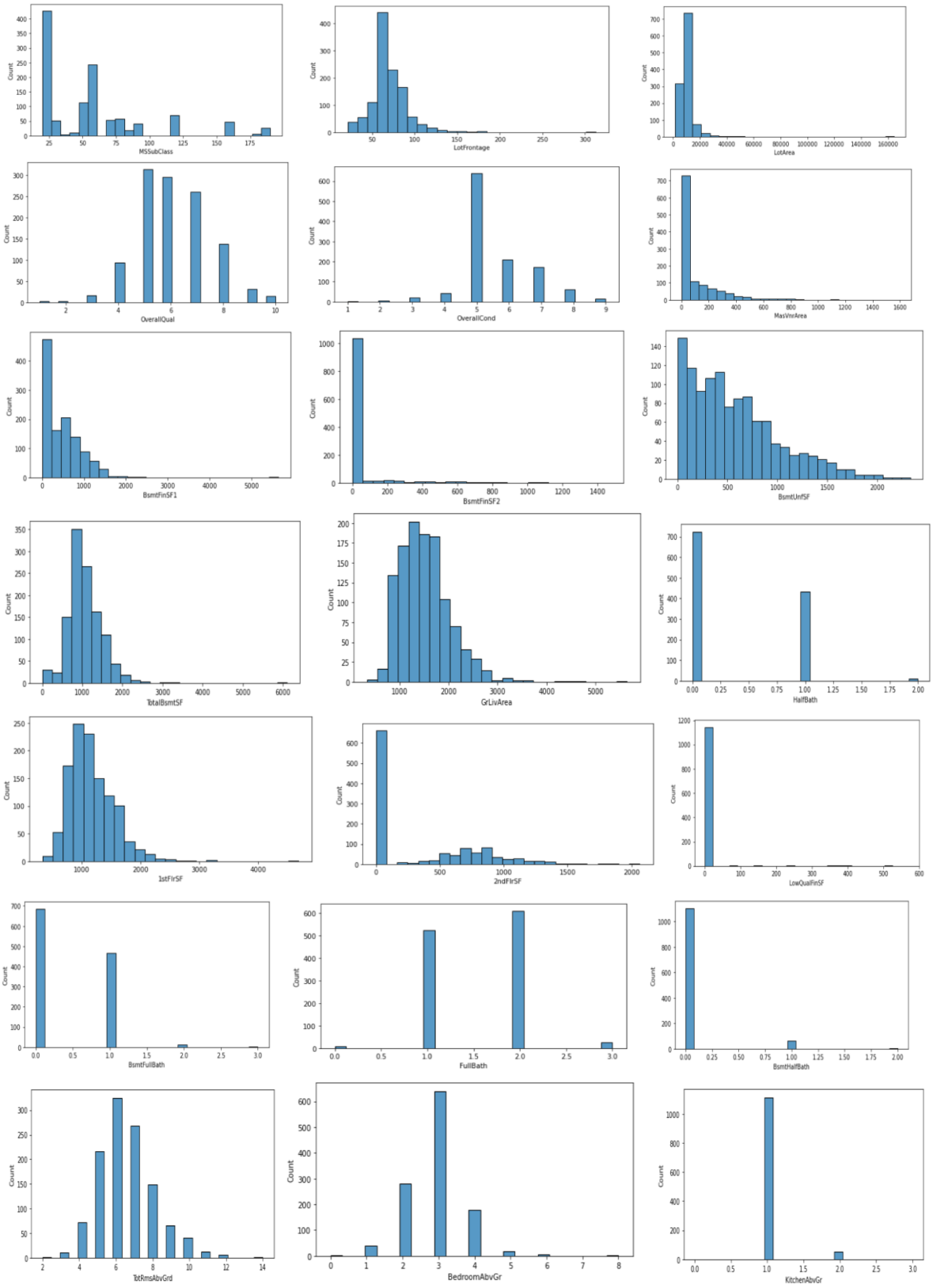
We also can see there are multiple variable those have more correlation among themselves than the target variable these variables will create the multi collinearity problem. To overcome those problem, I am dropping one of those variables. From above three categories the year sold, salestype and Sales condition year sold at 2007 with sale deed type Warranty Deed - Conventional Home was not completed when last assessed is costlier in market. Year build is 2007 Warranty Deed – Conventional and Normal Sale is costlier in the market. Home just constructed and sold sales type and Home was not completed when last assessed (associated with New Homes) sales condition are sold more. We can see more of a negative correlation in the with the target variable. WoodsDeckSF and OpenPorchSF have high correlation with sales prize.

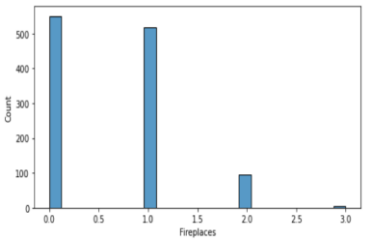
- # Hardware and Software Requirements and Tools Used

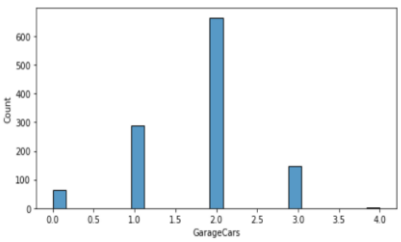
Tools Used:

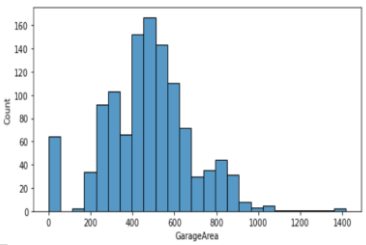
1. Python 3.8
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn
6. Data Science
7. Machine Learning

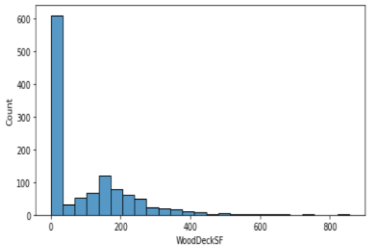
Data Visualization:
Visualization of Numeric Variables:

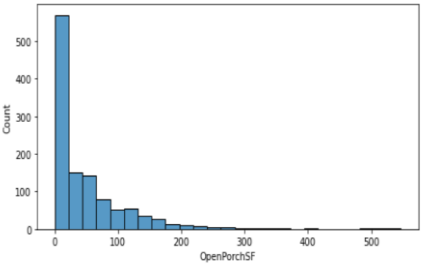


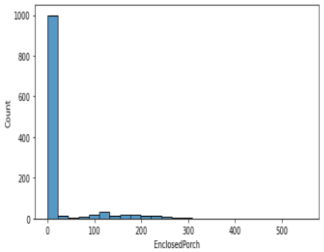

Count
Fireplaces

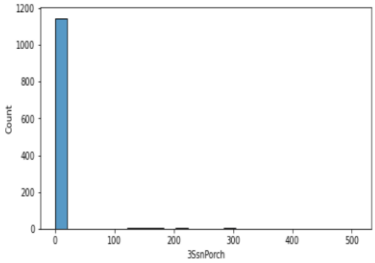

Count
GarageCars

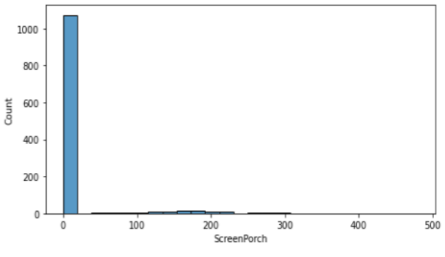

Count
GarageArea

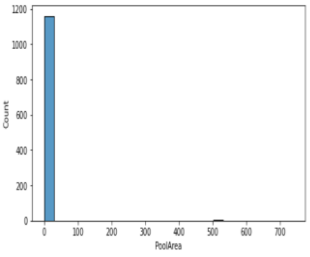

Count
WoodDeckSF

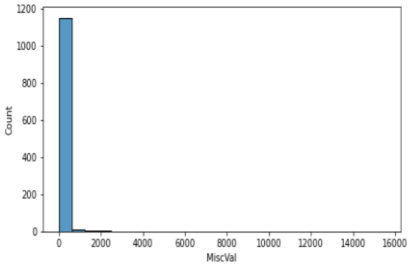

Count
OpenPorchSF

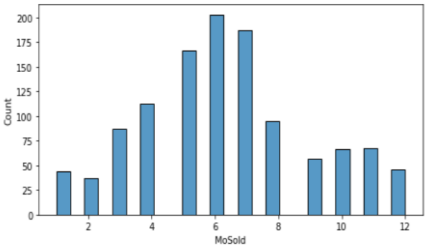

Count
EnclosedPorch

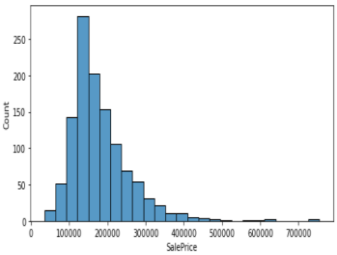

Count
3SsnPorch

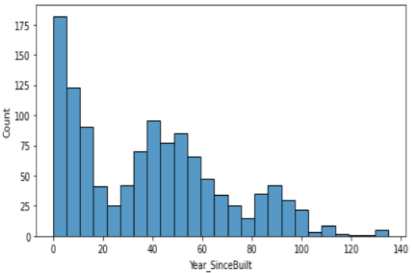

Count
ScreenPorch

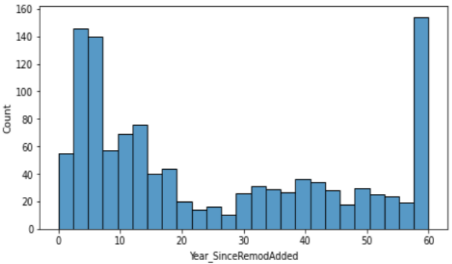

Count
PoolArea

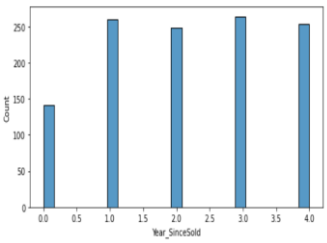

Count
MiscVal

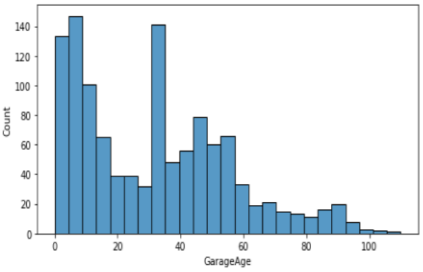

Count
MoSold


Count
SalePrice


Count
Year_SinceBuilt


Count
Year_SinceRemodAdded

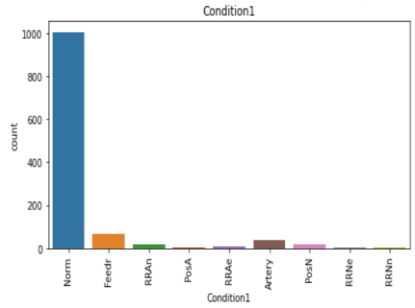

Count
Year_SinceSold

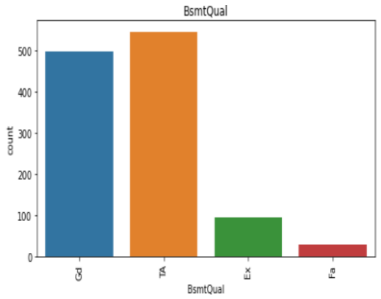

Count
GarageAge

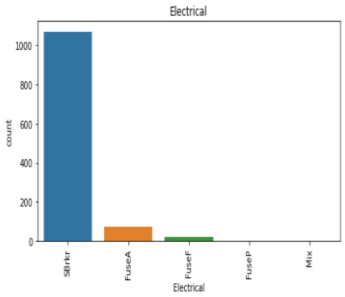
Key Observations:

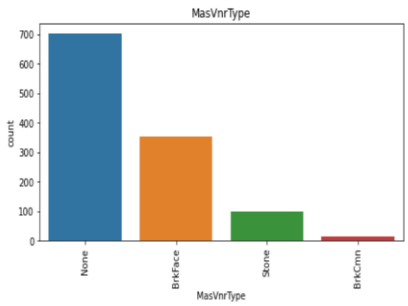
1. As MSSubClass is increasing, sales is decreasing. And it is seen between 25-75 there are a greater number of sales.
2. As LotFrontage increases, sales decreases. And more sales is between 25 and 100
3. As LotArea increases, sales decreases.
4. OverallQual is more between 4 and 8.
5. OverallCond is more at 5 and then sales decreases. Sales deponds on condition of the building if it is good sales increases otherwise it decreases.
6. AS MasVnrArea increases, sales decreases.
7. AS BsmtFinSF1 increases, sales decreases.
8. As BsmtFinSF2 increases, sales decreases.
9. As BsmtUnfSF increases, sales decrease gradually.
10. As TotalBsmtSF increases, sales decreases.
11. As 1stFlrSF increases, sales decreases.
12. Sales is more at 2ndFlrSF at 0-50. Then it gradually decreases.
13. GrLivArea increases, sales decreases.
14. HalfBath,Kitchen may prefers 1 room. As size increases sales decreases.
15. GarageCar mostly prefered onr is 2 car parking.
16. Mostly prefered GarageAre is between 400 and 800. As area increases sales decreases.
17. As OpenPorchSf increases, sales decreases.
18. Many Prefer ClosedPorch.
19. As age of building increases, sales decreases. Most of them prefer newly constructed building
20. Sales price ranges from 1.5-2lakhs in most cases. And there are even people who prefers upto 7 lakhs but number of counts who prefer is less.

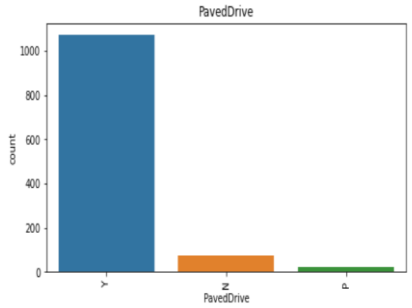
Visualization of categorical columns:

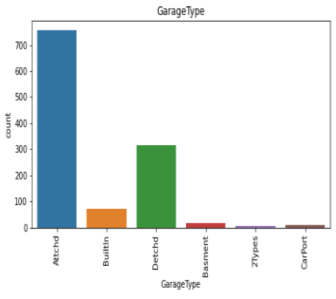


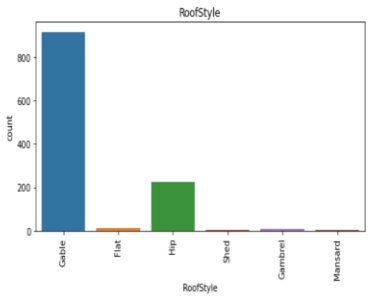


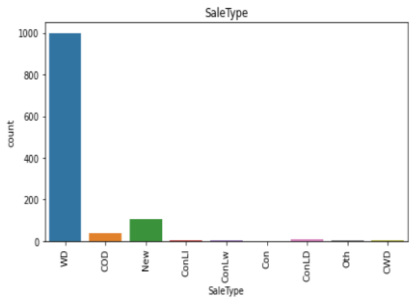


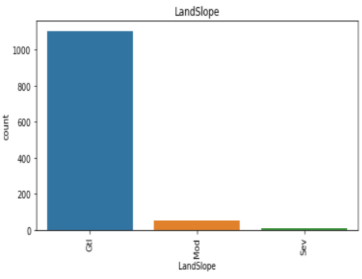


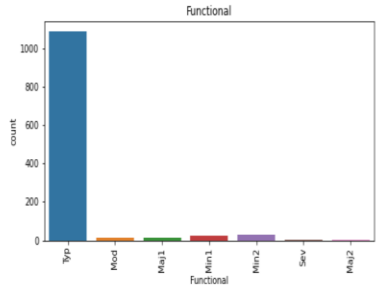


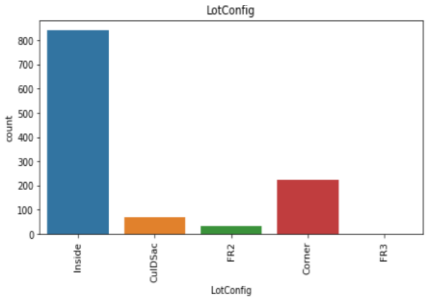


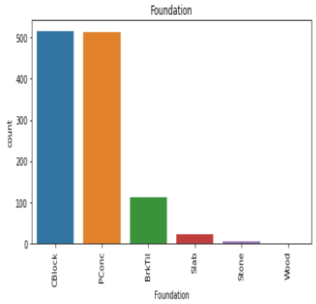


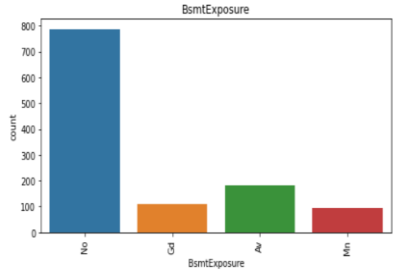


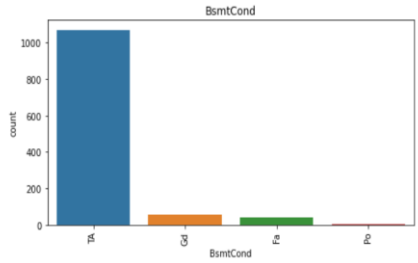


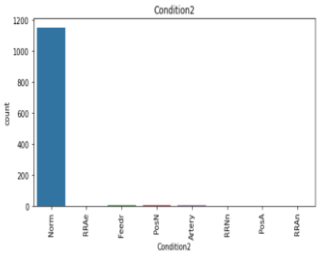


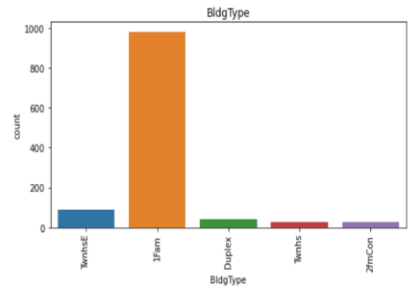


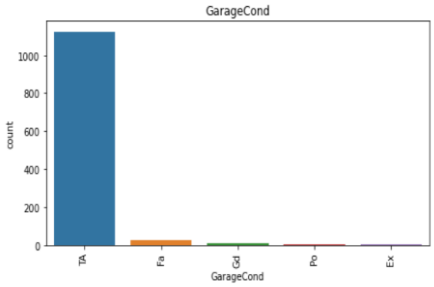


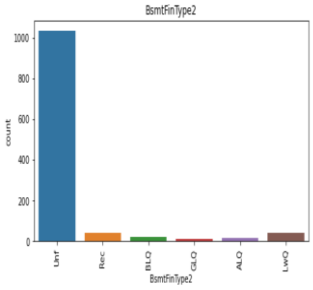


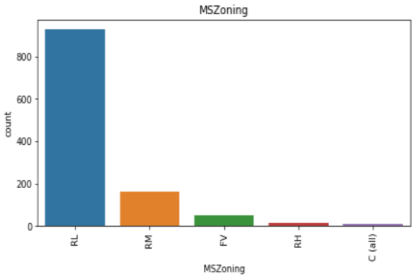


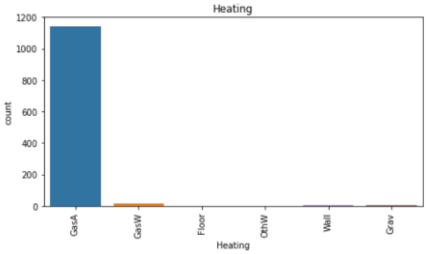


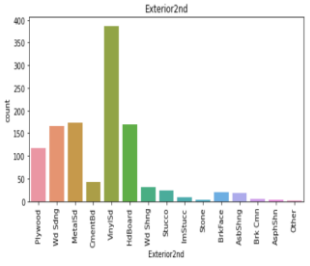

BldgType
count
TwnhsE
1Fam
Duplex
Twnhs
2fmCon
BldgType

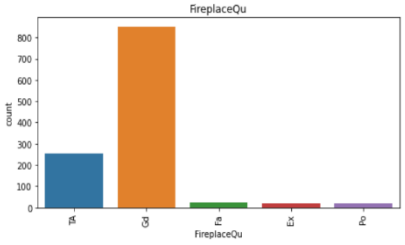

GarageCond
count
TA
Fa
Gd
Po
Ex
GarageCond

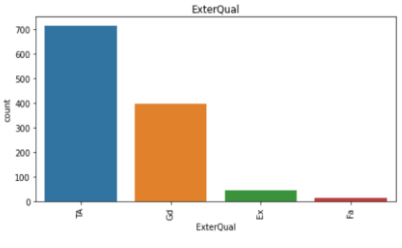

BsmtFinType2
count
Unf
Rec
BLQ
GLQ
ALQ
LwQ
BsmtFinType2

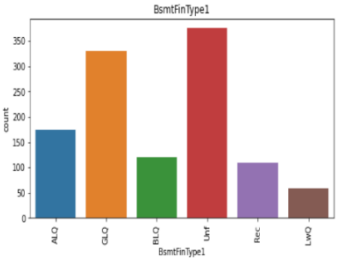

MSZoning
count
RL
RM
FV
RH
C (all)
MSZoning

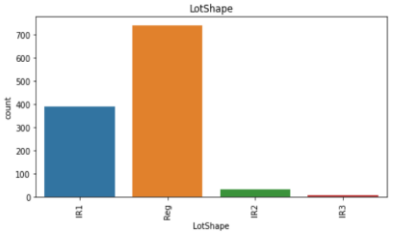

Heating
count
GasA
GasW
Floor
OthW
Wall
Grav
Heating

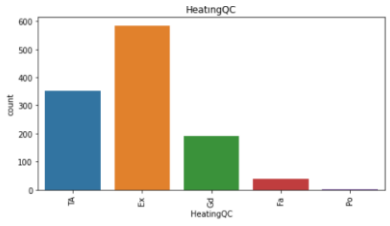

Exterior2nd
count
Plywood
Wd Sdng
MetalSd
CmentBd
VinylSd
HdBoard
Wd Shng
Stucco
ImStucc
Stone
BrkFace
AsbShng
Brk Cmn
AsphShn
Other
Exterior2nd

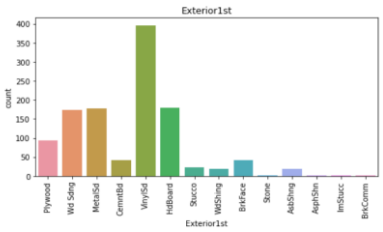

FireplaceQu
count
TA
Gd
Fa
Ex
Po
FireplaceQu

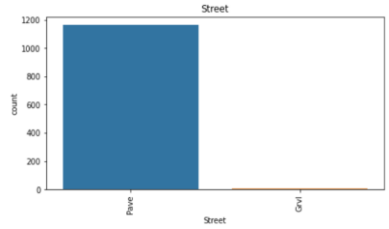

ExterQual
count
TA
Gd
Ex
Fa
ExterQual

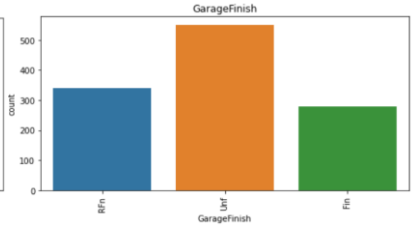

BsmtFinType1
count
ALQ
GLQ
BLQ
Unf
Rec
LwQ
BsmtFinType1

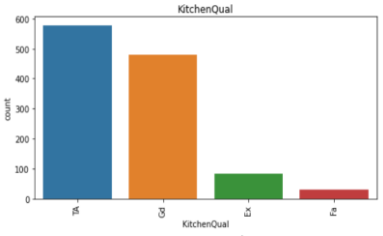

LotShape
count
IR1
Reg
IR2
IR3
LotShape

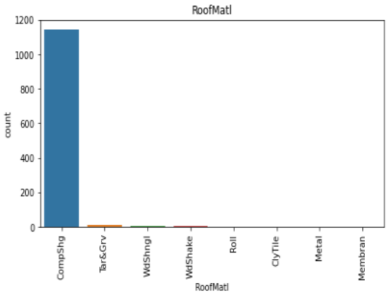

HeatingQC
count
TA
Ex
Gd
Fa
Po
HeatingQC

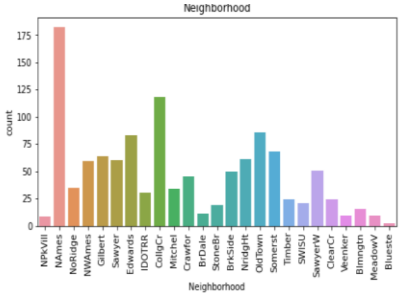

Exterior1st
count
Plywood
Wd Sdng
MetalSd
CmntBd
VinylSd
HdBoard
Stucco
WdShing
BrkFace
Stone
AsbShng
AsphShn
ImStucc
BrkComm
Exterior1st

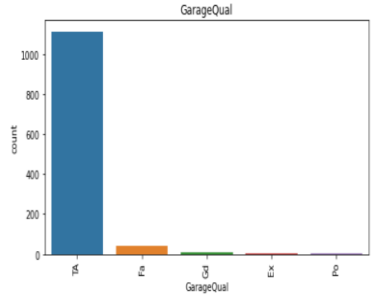

Street
count
Pave
Grvl
Street

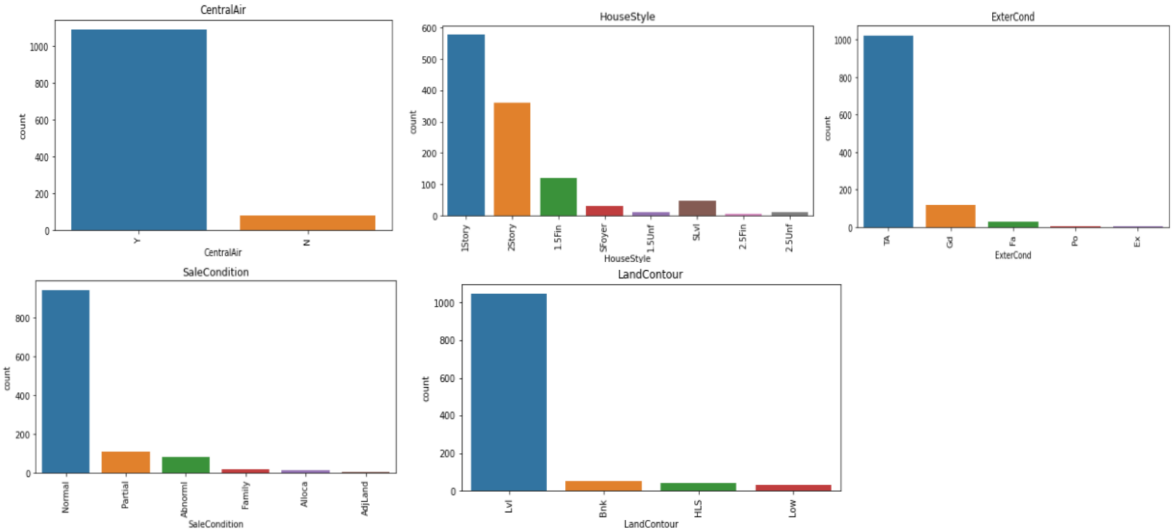

GarageFinish
count
RFn
Unf
Fin
GarageFinish


KitchenQual
count
TA
Gd
Ex
Fa
KitchenQual


RoofMatl
count
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
RoofMatl


Neighborhood
count
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


GarageQual
count
TA
Fa
Gd
Ex
Po
GarageQual



Key Observations:

1. Evaluates the quality of material on the exterior TA(average/Typical) has more sales followed by gd(Good)
2. Proximity to various conditions(if more than one is present) Norm(Normal) sales is more.
3. Many prefered Vinyl Sliding(VinylSd) followed by Hard Board(HDBoard),Metal Sliding(MetalSd), WoodSliding(Wd Sliding)
4. Garage Location - Attached to Home(Attchd) sees high sales followed by Detached from home(Detchd).
5. BsmtExposure(Walkout or Garden level walls) many sales is done with No Basement followed by Av(Average Exposure).
6. Functional: Many sales is done for Typical Funcionality(Typ).
7. More sales is done in North Ames(NAmes) followed by College Creek(CollgCr) which sees high sales. and lowest sales is seen in Bluestem(Bluestem).
8. In Rating of basement finished area(BsmtFinType2) many sales done with Unf(Unfinished), followed by LwQ(Low Quality)
9. Many sales done foe TA(Typical/Average) Garage Condition.
10. Many sales is seen with Unf(Unfinished) GarageFinish.
11. Sales Type is high in WD(Warranty Deed) followed by New(Home just constructed and sold).
12. LotShape(General Shape of Propert)- Reg(Regular) has seen more sales.
13. LandSlope(Slope of Property) -Gtl(Gentle Slope) has seen high sales.
14. Many sales are done with CentralAir conditioning.
15. Lvl(Near Flat/Level) has seen more slaes in LandContour(Flatness of Property)
16. MasVnrType(Masonry veneer type)- None has seen more slaes followed by BrkFace(Brick Face).
17. Foundation(Type of foundation) - CBlock(Cinder Block) has more sales followed by PConc(Poured Contrete)
18. BldgType(Type of Dwelling)- 1Fam(Single Family detached) has more sales, followed by TwnhsE(Townhouse End Unit)
19. MSZoning(General zoning classification)- more sales is done with RL(Residential Low Density) followed by RM(Residential Medium Density).

20. Electrical(Electrical System)- Many slaes is done for SBkr(Standard Circuit Breakers & Romex).
21. SaleCondition - Mostly Normal Sale is done followed by Partial.
22. Mostly 1Storey has ssen high sales followed by 2Storey in HouseStyle.
23. GasA(Gas forces warm air furnace) with excellent quality is prefered in Heating Type which has seen large number of slaes.
24. RoofStyle and Material most sales happed with CompShg(Standard(composite) shingle) with Gable material.
25. Most saled happed with PavedDrive
26. LotConfig(Lot Configuration) Inside lot has seen more sales.
27. Lot of sales is seen in Good Fireplace Quality.
28. More sales is seen with Pave Street than Gravel street.
29. Kitchen, ExternalQual, GarageQual sales is seen high in TA(Average).

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches

In this project, both Statistical and Analytical methods are used, in which Data Pre-processing, Exploratory Data Analysis is used after ensuring that data is cleaned.Here Target column is Sale Price, as it is continuous data Regression algorithm is used. This project consists of 4 regressors – Decision Tree Regressor, Support Vector Regressor, KNeighbors Regressor, Linear Regression, 4 ensemble -Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, 3 metrics – r2_score, mean_squared_error, mean_absolute_error and 3 regularization – Lasso, Ridge, ElasticNet techniques. Out of which Gradient Boosting Regressor gave high accuracy which is also chosen as best model in which there is least difference between r2 score and cv. Then with this Hypertuning is done in which accuracy is slightly reduced in order to correct the over fitting of the model. Once all this is done, best model is saved using joblib.
Then by using this saved train model the output of test model is determined which predicted the SalePrice of the house.

- ## Testing of Identified Approaches (Algorithms)

The approaches followed in this project are:

1. Find the best random state of a model.
2. Use all the other models to find accuracy score, mean squared error, mean absolute error and r2 score.
3. Then find Cross Validation of all models to find the best accuracy, which is the least difference between r2 score and cv score.
4. With the best model accuracy, we need to do hyper tuning using GridSearchCV.
5. Then finally saving the best model and by keeping this we need to predict the SalePrice of test dataset.

- Run and evaluate selected models

```
score = []
mean_squared_err = []
mean_absolute_err = []
r2 = []

for m in model:
    m.fit(x_train,y_train)
    m.score(x_test,y_test)
    predm = m.predict(x_test)
    print("Accuracy Score of ",m," is ",m.score(x_train,y_train))
    score.append(m.score(x_train,y_train))

    print("Mean Squared Error is ",mean_squared_error(y_test,predm))
    mean_squared_err.append(mean_squared_error(y_test,predm))
    print("Mean Absolute Error is ",mean_absolute_error(y_test,predm))
    mean_absolute_err.append(mean_absolute_error(y_test,predm))
    print("R2 Score is ",r2_score(y_test,predm))
    r2.append(r2_score(y_test,predm))
    print("\n\n")
```

```
Accuracy Score of  LinearRegression()  is  0.8320847573445096
Mean Squared Error is  1385808343.0923483
Mean Absolute Error is  22679.732617364814
R2 Score is  0.8081006677622541
```

Linear Regression gives 83% accuracy and r2_score 0.80

```
Accuracy Score of  Lasso()  is  0.8320847332606152
Mean Squared Error is  1385688687.559255
Mean Absolute Error is  22678.16203852448
R2 Score is  0.8081172370209209
```

Lasso gives 83% accuracy and r2_score 0.808

```
Accuracy Score of  Ridge()  is  0.832084140142928
Mean Squared Error is  1385755686.3146753
Mean Absolute Error is  22676.528604373292
R2 Score is  0.808107959391377
```

Ridge give 83% accuracy score and r2_score 0.808

```
Accuracy Score of  ElasticNet()  is  0.8141221573280071
Mean Squared Error is  1489435944.8422742
Mean Absolute Error is  22462.784624940607
R2 Score is  0.7937508713590695
```

ElasticNet gives 81% accuracy score and r2_score 0.79

```
Accuracy Score of  KNeighborsRegressor()  is  0.8225983926850005
Mean Squared Error is  1827702336.8393776
Mean Absolute Error is  24082.444357976656
R2 Score is  0.7469095494213864
```

KNeighborsRegressor gives 82% accuracy and r2 score 0.74

```
Accuracy Score of  DecisionTreeRegressor()  is  1.0
Mean Squared Error is  1980738485.003891
Mean Absolute Error is  27628.25680933852
R2 Score is  0.7257179215982525
```

DecisionTreeRegressor gives 100% accuracy and r2_score 0.725

```
Accuracy Score of  SVR()  is  -0.05801826281672717
Mean Squared Error is  7810474667.171102
Mean Absolute Error is  59414.50462636244
R2 Score is  -0.08155278510261166
```

SVR gives 5% accuracy and 0.08 r2_score

```
Accuracy Score of  RandomForestRegressor()  is  0.974807439106156
Mean Squared Error is  795424429.9379529
Mean Absolute Error is  17070.5219844358
R2 Score is  0.8898538764674538
```

RandomForestRegressor gives 97% accuracy and 0.88 r2_score

```
Accuracy Score of  AdaBoostRegressor()  is  0.8456615414547114
Mean Squared Error is  1389743292.519671
Mean Absolute Error is  26391.11801619872
R2 Score is  0.8075557769977726
```

AdaBoostRegressor gives 84% accuracy score and 0.80 r2_score

```
Accuracy Score of  GradientBoostingRegressor()  is  0.96657445221
47823
Mean Squared Error is  656221566.9059403
Mean Absolute Error is  16200.909701663772
R2 Score is  0.9091299449040295
```

Gradient Boosting Regressor gives 96% accuracy and 0.90 r2_score

Cross Validation:

```
Score of  LinearRegression()  is  [0.78335581 0.76175025 0.608587
81 0.82766633 0.78446133]
Mean score is :  0.7531643064551607
Standard Deviation is :  0.07538920669452026
```

CV Score of LinearRegressor is 0.75

```
Score of  Lasso()  is  [0.78338703 0.76189508 0.608689   0.827691
57 0.78454285]
Mean score is :  0.7532411029350798
Standard Deviation is :  0.07536796448868065
```

CV Score of Lasso is 0.75

```
Score of  Ridge()  is  [0.78346762 0.76234052 0.6098054  0.827753
66 0.78505411]
Mean score is :  0.7536842601379498
Standard Deviation is :  0.07501126525149635
```

CV Score of Ridge is 0.753

```
Score of  ElasticNet()  is  [0.78626938 0.77562527 0.69047169 0.8
2529161 0.83531361]
Mean score is :  0.7825943133888549
Standard Deviation is :  0.051285545153442844
```

CV Score of ElasticNet is 0.78

```
Score of  KNeighborsRegressor()  is  [0.7273917  0.76180962 0.694
91028 0.74696118 0.74139509]
Mean score is :  0.7344935714053101
Standard Deviation is :  0.022656180902048878
```

CV Score of KNeighborsRegressor is 0.73

```
Score of  DecisionTreeRegressor()  is  [0.76446138 0.63968556 0.7
4545299 0.75776065 0.64068073]
Mean score is :  0.7096082605736136
Standard Deviation is :  0.057013243041671234
```

CV Score of DecisionTreeRegressor is 0.70

```
Score of  SVR()  is  [-0.03499496 -0.13003483 -0.02727118 -0.1134
5164 -0.00300291]
Mean score is :  -0.061751104350369165
Standard Deviation is :  0.05038198233334458
```

CV Score of SVR is 0.06

```
Score of  RandomForestRegressor()  is  [0.89234937 0.78897492 0.8
3388256 0.89300093 0.82871008]
Mean score is :  0.8473835703649139
Standard Deviation is :  0.040113372995800256
```

CV Score of RandomForestRegressor is 0.84

```
Score of  AdaBoostRegressor()  is  [0.84880251 0.78725394 0.80249
244 0.83847242 0.73443774]
Mean score is :  0.8022918102114384
Standard Deviation is :  0.04074956233792602
```

CV Score of AdaBoostRegressor is 0.80

```
Score of  GradientBoostingRegressor()  is  [0.89955547 0.75705978 0.8938133  0.9092137  0.86261357]
Mean score is :  0.864451164881398
Standard Deviation is :  0.05591854690832243
```

CV Score of GradientBoostingRegressor is 0.86

By this we can conclude that GradientBoostingRegressor is having accuracy score of 86%. which can be sent for Hypertuning.

HyperTuning:

```
GBR = GradientBoostingRegressor()
parameters = {'learning_rate':[0.01,0.02,0.03,0.04],
              'subsample' : [0.9,0.5,0.2,0.1],
              'n_estimators':[100,150,200,250],
              'max_depth': [4,6,8,10]}

grid_GBR = GridSearchCV(estimator=GBR,param_grid=parameters,cv=2,n_jobs=1)
grid_GBR.fit(x_train,y_train)
```

```
GridSearchCV(cv=2, estimator=GradientBoostingRegressor(), n_jobs=1,
             param_grid={'learning_rate': [0.01, 0.02, 0.03, 0.04],
                         'max_depth': [4, 6, 8, 10],
                         'n_estimators': [100, 150, 200, 250],
                         'subsample': [0.9, 0.5, 0.2, 0.1]})
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

```
print("Results from Grid Search : ")
print("\n The best estimator across all search params :",grid_GBR.best_estimator_)
print("\n The best score for all searched params :",grid_GBR.best_score_)
print("\n The best parameters for all searched params : ",grid_GBR.best_params_)
```

```
Results from Grid Search :

 The best estimator across all search params : GradientBoostingRegressor(learning_rate=0.03, max_depth=8, n_estimators=150,
                           subsample=0.2)

 The best score for all searched params : 0.8327192987025529

 The best parameters for all searched params :  {'learning_rate': 0.03, 'max_depth': 8, 'n_estimators': 150, 'subsample': 0.2}
```

After hypertuning we have 83% accuracy

```
In [107]: final_grid_gbr = grid_GBR.best_estimator_

In [108]: final_grid_gbr.fit(x_train,y_train)

Out[108]: GradientBoostingRegressor(learning_rate=0.03, max_depth=8, n_estimators=150,
                                    subsample=0.2)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

Accuracy in hypertuning is reduced which is due to the model tried to reduce overfitting.

```
In [109]: final_grid_gbr.score(x_test,y_test)

Out[109]: 0.8962883565842157
```

After Hypertuning we have 89.62% accuracy score which is improved.

Saving the best model:

The best model is saved using joblib.

```
import joblib
joblib.dump(grid_GBR,"House_Price_Prediction.obj")
```

```
['House_Price_Prediction.obj']
```

Predicting the output:

```
Housing_Price = joblib.load('House_Price_Prediction.obj')
prediction = Housing_Price.predict(X)
prediction
```

```
array([138902.38223644, 293231.91007271, 245890.13323817, ...,
       146676.26556115,  57663.03888074, 180929.14538157])
```

```
Housing_Price_Prediction = pd.DataFrame()
Housing_Price_Prediction["House_Price"] = prediction
Housing_Price_Prediction.head(20)
```

The House Price is predicted using train dataset and is displayed in dataframe for better interpreting the results.

	House_Price
0	138902.382236
1	293231.910073
2	245890.133238
3	192315.115089
4	232985.982499
5	224134.440113
6	153240.474251
7	160155.551585
8	137160.101872
9	111341.649633
10	118440.054170
11	218321.671691
12	210086.991640
13	123855.733817
14	141974.146709
15	140668.756998
16	131075.435220
17	181765.361479
18	164332.700733
19	107742.090375

```
Housing_Price_Prediction.to_csv("Housing_Price_Prediction.csv")
```

The SalePrice is predicted for test dataset and then saved to csv format.

- # Key Metrics for success in solving problem under consideration

Some of key metrics for the evaluation of this project are

1. R2_score
2. Mean_squared_error
3. Mean_absolute_error
4. Cross_validation
5. Hypertuning

R2 Score:

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. Closer the r2_square value to 1 better the model fits.

Mean Squared Error:

The average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated is measured by the MSE of an estimator (of a process for estimating an unobserved variable). MSE is a risk function that represents the squared error loss's anticipated value. The Root Mean Squared Error is abbreviated as RMSE.

Mean Absolute Error:

MAE is a statistic that assesses the average magnitude of mistakes in a set of forecasts without taking into account their direction. It's the average of the absolute differences between forecast and actual observation over the test sample, where all individual deviations are given equal weight.

Cross Validation:

Cross-validation aids in determining the model's overfitting and underfitting. The model is constructed to run on several subsets of the dataset in cross validation, resulting in numerous measurements of the model. If we fold the data five times, it will be separated into five parts, each representing 20% of the whole dataset. During the Cross-validation, the first part (20%) of the 5 parts will be left out as a holdout set for validation, while the rest of the data will be utilised for training. We'll acquire the initial estimate of the dataset's model quality this way.
Further rounds are produced in the same way for the second 20% of the dataset, which is kept as a holdout set while the remaining four portions are utilised for training data during the process. We'll acquire the second estimate of the dataset's model quality this way. During the cross-validation procedure, these stages are repeated to obtain the remaining estimate of model quality.

Hypertuning:

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

- # Interpretation of the Results

This project consists of two dataset train and test. We have predicted the SalePrice of the house using train dataset. This can help Surprise Housing company to start a housing business in Australia.

CONCLUSION

- ## Key Findings and Conclusions of the Study

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

- ## Learning Outcomes of the Study in respect of Data Science

1. The above study helps one to understand how the price is changing across the Properties with respect to amenities like swimming pool, lawn size, play area, pavement etc.

2. With the help of the above analysis, one can sketch the needs of a property buyer and according to need we can predict the price of the property.

- ## Limitations of this work and Scope for Future Work

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