HOUSING SALE PRICE PROJECT

Submitted by:

NEETAL TIWARI

ACKNOWLEDGMENT

This is the house sale price prediction project report.

Suppose you have a house and you want to sale it. Through House sale price prediction, you can predict the price from previous sell price history**.**

We have a Housing Dataset to predict the sales which contains 1168 rows and 81 columns, such as; Lot Area, Lot Shape, House Style, LowQualFinSF, KitchenQual, Functional, GarageType, GarageQual, PoolQC, SaleCondition, SaleType, MiscFeature and many more……

To predict house sale price first filled the null values than did the EDA (Exploratory Data Analysis) than checked the skewness and removed that with the help of sklearn_preprocessing, power_transform method, checked the outliers and removed that with the help of SciPy. stats, Zscore.

After doing all preprocessing, I use 4 algorithms to predict House saleprice Linear Regression, RandomForestRegressor, KNeigbhourRegressor and DecisionTreeRegressor.

RandomForestRegressor is the best algorithm for House sale price.

INTRODUCTION

Business Problem Framing

In today's time, building and selling a house has become a business, people make building and then sell it. But how to decide the price to sale the house, for that we look at many things how much the area of the house, which floor is the house on, does it have a parking area, when is the house built and many more things….

Nowadays so many buildings are being built. It becomes difficult to decide the sale price of the house, so with the help of all the features the sale price of the house is decided by using the machine learning algorithms.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

In most of the independent variables High Variance in 75% and 100% in many columns and also variance mean and median value, it means independents variables data are skewed. OverallCond (Rates the overall condition of the house) and EnclosedPorch (Enclosed porch area in square feet) are very less correlated and GrLivArea (Above grade (ground) living area square feet) are high correlated with the sale price. It means Enclosed porch area and overall condition of the house not very much affect the house sale price.

Data Sources and their formats

OverallQual (Rates the overall material and finish of the house), 1stFlrSF (First Floor square feet), GarageArea (Size of garage in square feet), GarageCars (Size of garage in car capacity), All these data sources greatly affect the sale price. Because it is natural that whenever someone goes to get the house or anything he first sees its rating whose rating is good that sells well, it means rating is very important data source for any house or anything. Also, people see that which floor the house is in, what is the area of the house and many more…

All these data source or independent features are very important to determine the house sale price.

Data Pre-processing Done

In Data pre-processing I use Label Encoding method to encode the objects to int, there are 38 objects in housing dataset.

Removed the skewness of the data by using power transform, Power Transformer, most of the independents features such as poolarea, 3Ssnporch, lowQualFinSF are high skewed.

Then remove the outliers by using zscore method. In 680 rows outliers were present.

After that I scaled the Data by using StandardScaler method.

Model/s Development and Evaluation

Testing of Identified Approaches (Algorithms)

I had used linear Regression, Random Forest Regressor, KNeighborsRegressor and Decision Tree Regressor to predict the sale price of the house.

Run and evaluate selected models

I had train x and y than I had predict x_test. In LinearRegressin model r2 score was 89.19%.

LinearRegression

```
]: li=LinearRegression()
   li.fit(x_train,y_train)
   lipred=li.predict(x_test)

   print('Mean absolute error:',mean_absolute_error(y_test,lipred))
   print('Mean squared error:',mean_squared_error(y_test,lipred))
   print('Root mean squraed Error:',np.sqrt(mean_squared_error(y_test,lipred)))
   print(r2_score(y_test,lipred)*100)
```

```
Mean absolute error: 15268.645344478418
Mean squared error: 369463929.2027042
Root mean squraed Error: 19221.444513945986
89.19144303618182
```

In RandomforestRegressor I got 92.31% of accuracy which is very good.

RandomForestRegressor ¶

```
]: rf=RandomForestRegressor()
rf.fit(x_train,y_train)
rfpred=rf.predict(x_test)

print('Mean absolute error:',mean_absolute_error(y_test,rfpred))
print('Mean squared error:',mean_squared_error(y_test,rfpred))
print('Root mean squraed Error:',np.sqrt(mean_squared_error(y_test,rfpred)))
print(r2_score(y_test,rfpred)*100)
```

```
Mean absolute error: 12174.42857142857
Mean squared error: 262625027.68144077
Root mean squraed Error: 16205.709724706314
92.31698320876734
```

In KNeighbors Regressor I got 86.09 of accuracy.

KNeighborsRegressor

6]:
```
knn=KNeighborsRegressor()
knn.fit(x_train,y_train)
knnpred=knn.predict(x_test)
print('Mean absolute error:',mean_absolute_error(y_test,knnpred))
print('Mean squared error:',mean_squared_error(y_test,knnpred))
print('Root mean squraed Error:',np.sqrt(mean_squared_error(y_test,knnpred)))
print(r2_score(y_test,knnpred)*100)
```

```
Mean absolute error: 16328.212244897957
Mean squared error: 475315694.7812245
Root mean squraed Error: 21801.736049709998
86.09478123094114
```

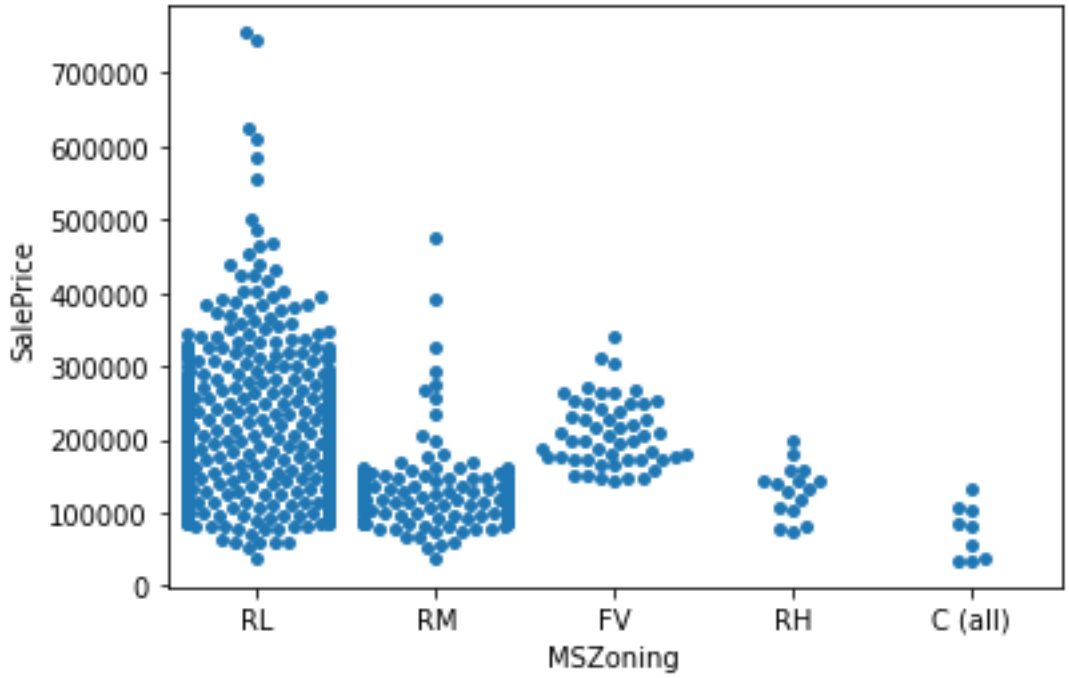
In Decision Tree regressor I got 73.09% of accuracy.

DecisionTreeRegressor

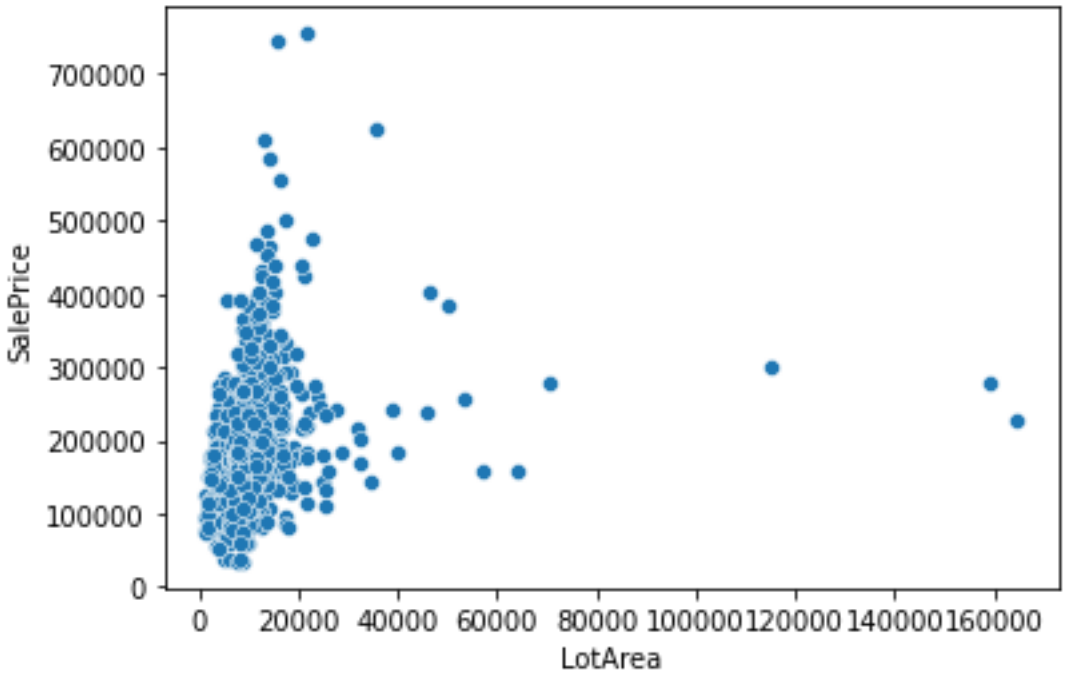
```
]: dtr=DecisionTreeRegressor()
   dtr.fit(x_train,y_train)
   dtrpred=dtr.predict(x_test)
   print('Mean absolute error:',mean_absolute_error(y_test,dtrpred))
   print('Mean squared error:',mean_squared_error(y_test,dtrpred))
   print('Root mean squraed Error:',np.sqrt(mean_squared_error(y_test,dtrpred)))
   print(r2_score(y_test,dtrpred)*100)
```

```
Mean absolute error: 18803.979591836734
Mean squared error: 919528250.244898
Root mean squraed Error: 30323.724214629343
73.09947551832813
```

Visualizations

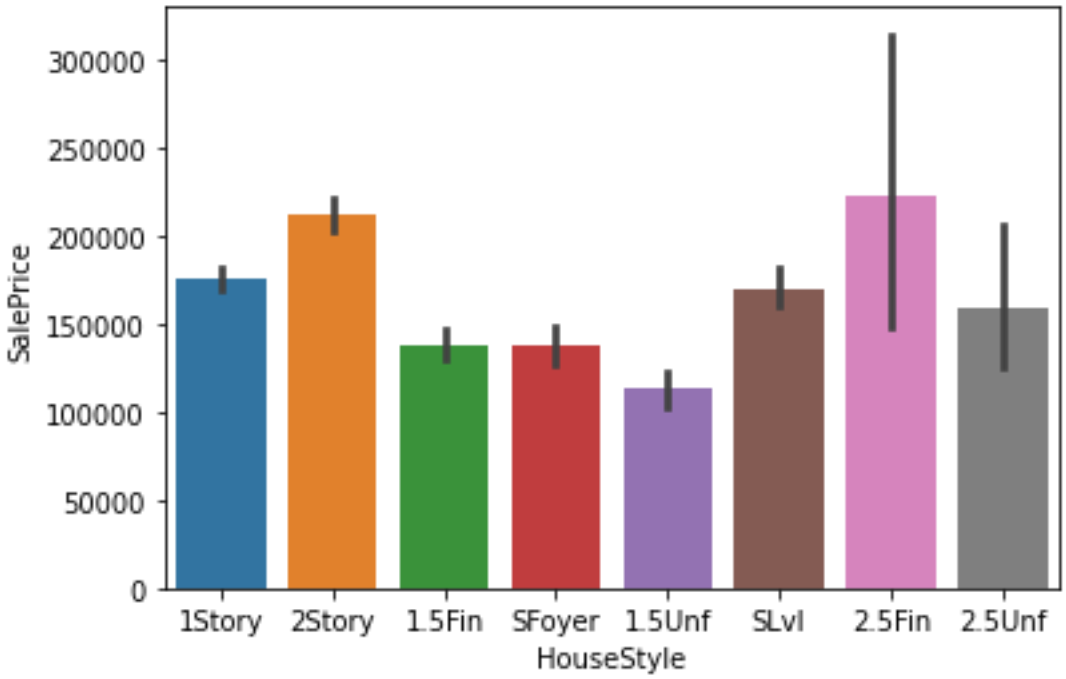


Sale Price of RL (Residential Low Density) is highest.



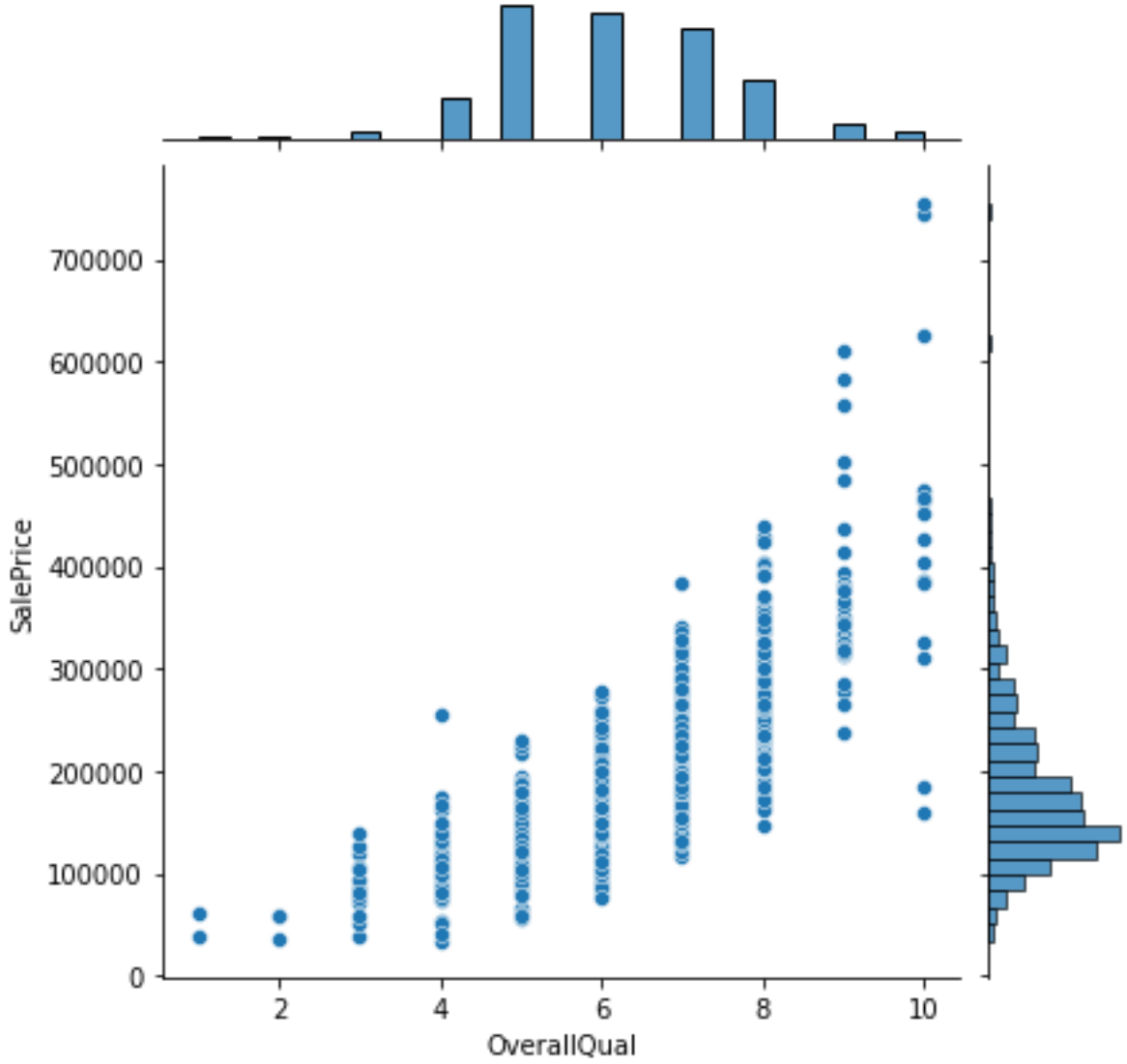
Observation;

Sale Price is increasing with Lot Area (Lot size in square feet).



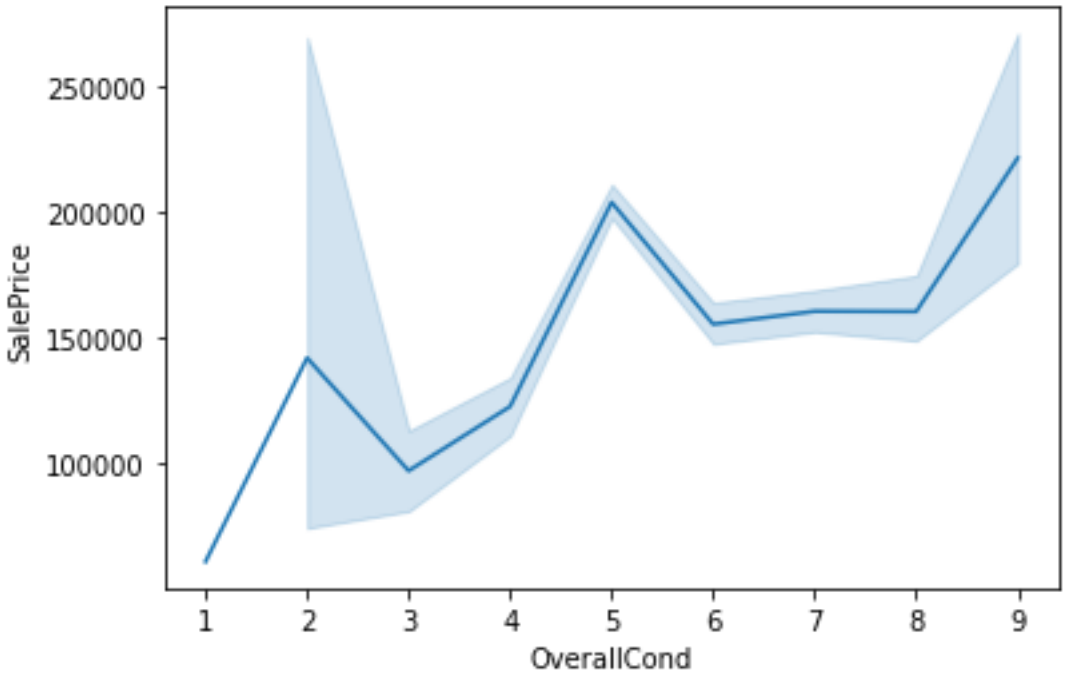
Observation;

Sale Price is All type of House Style is Good.



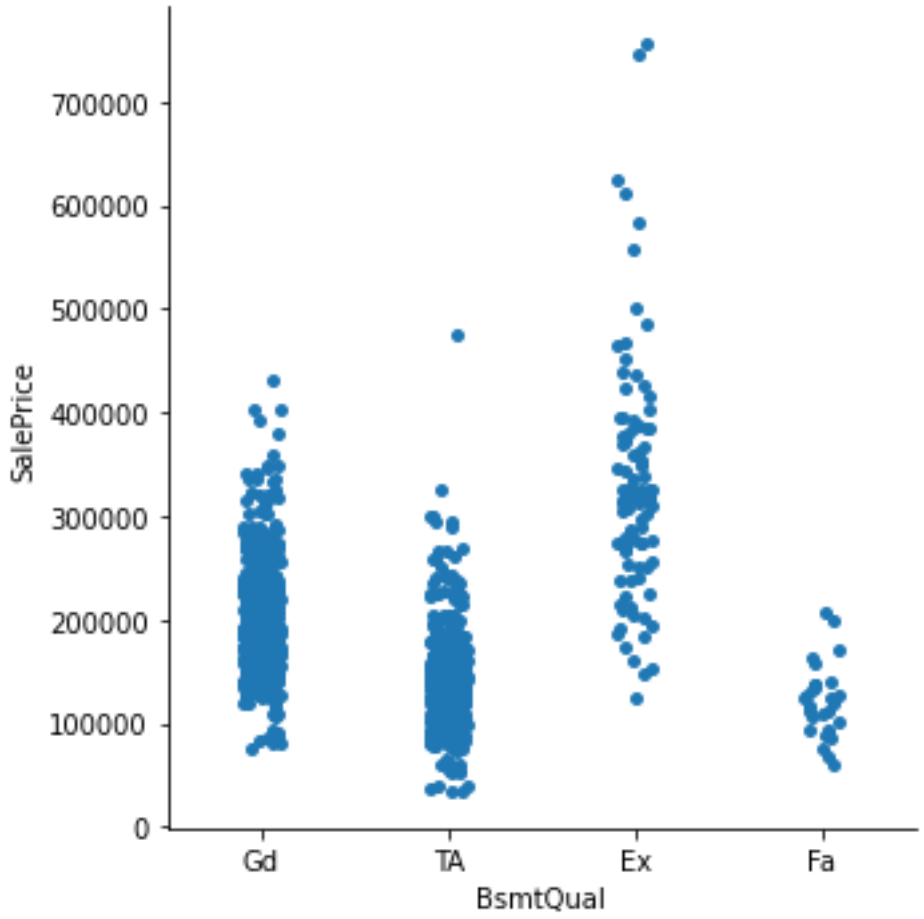
Observation;

Sale Price is increasing With the OverallQual (Rates the overall material and finish of the house).

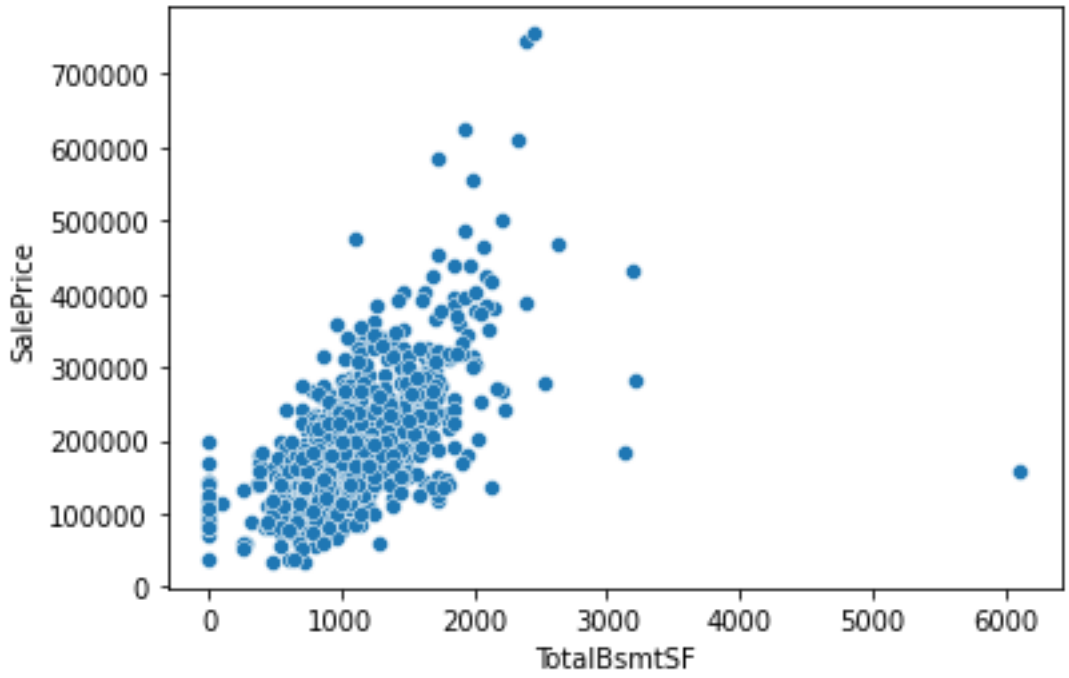


Observation;

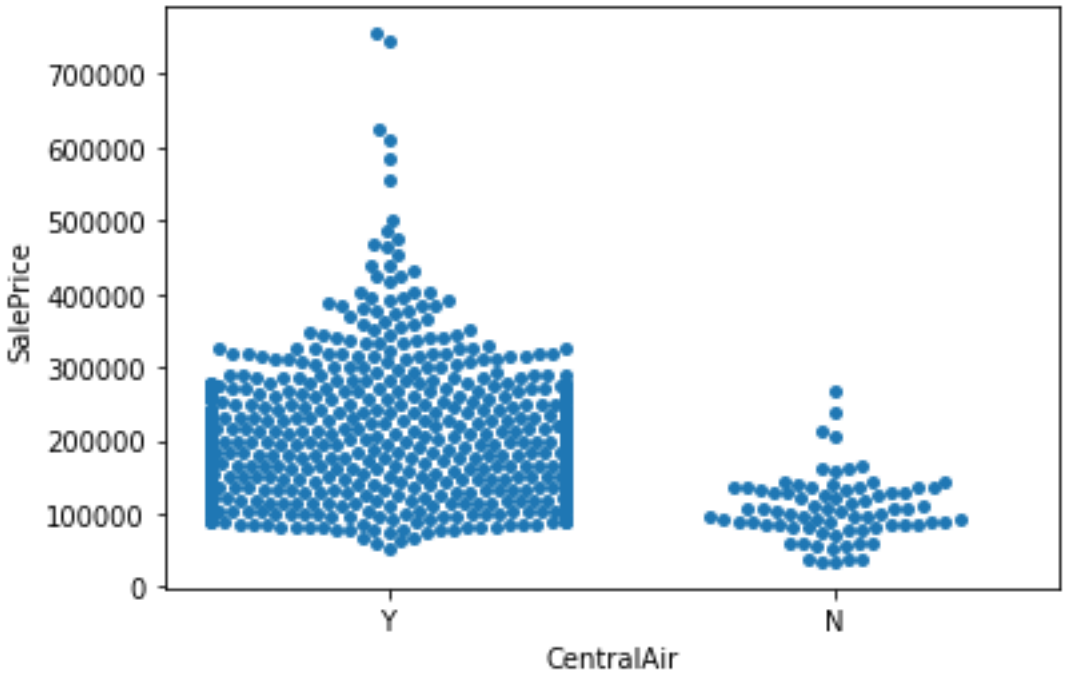
Sale Price is increasing with the OverallCond (Rates the overall condition of the house).



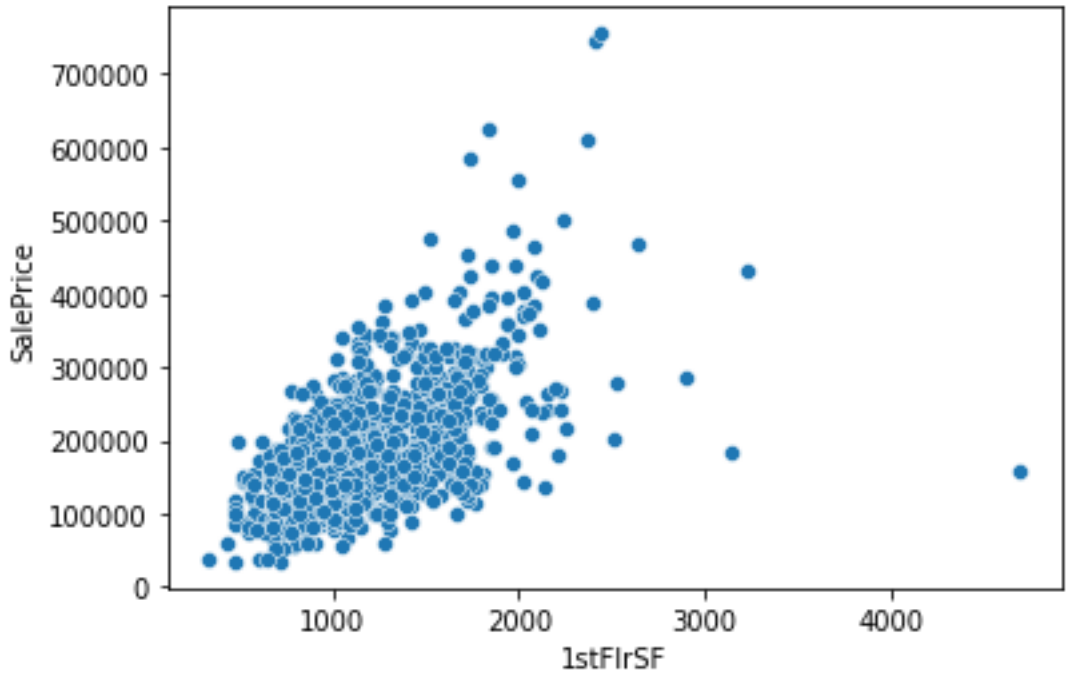
Sale Price is higher Ex (Excellent (100+ inches)) type of BsmtQual (Evaluates the height of the basement).



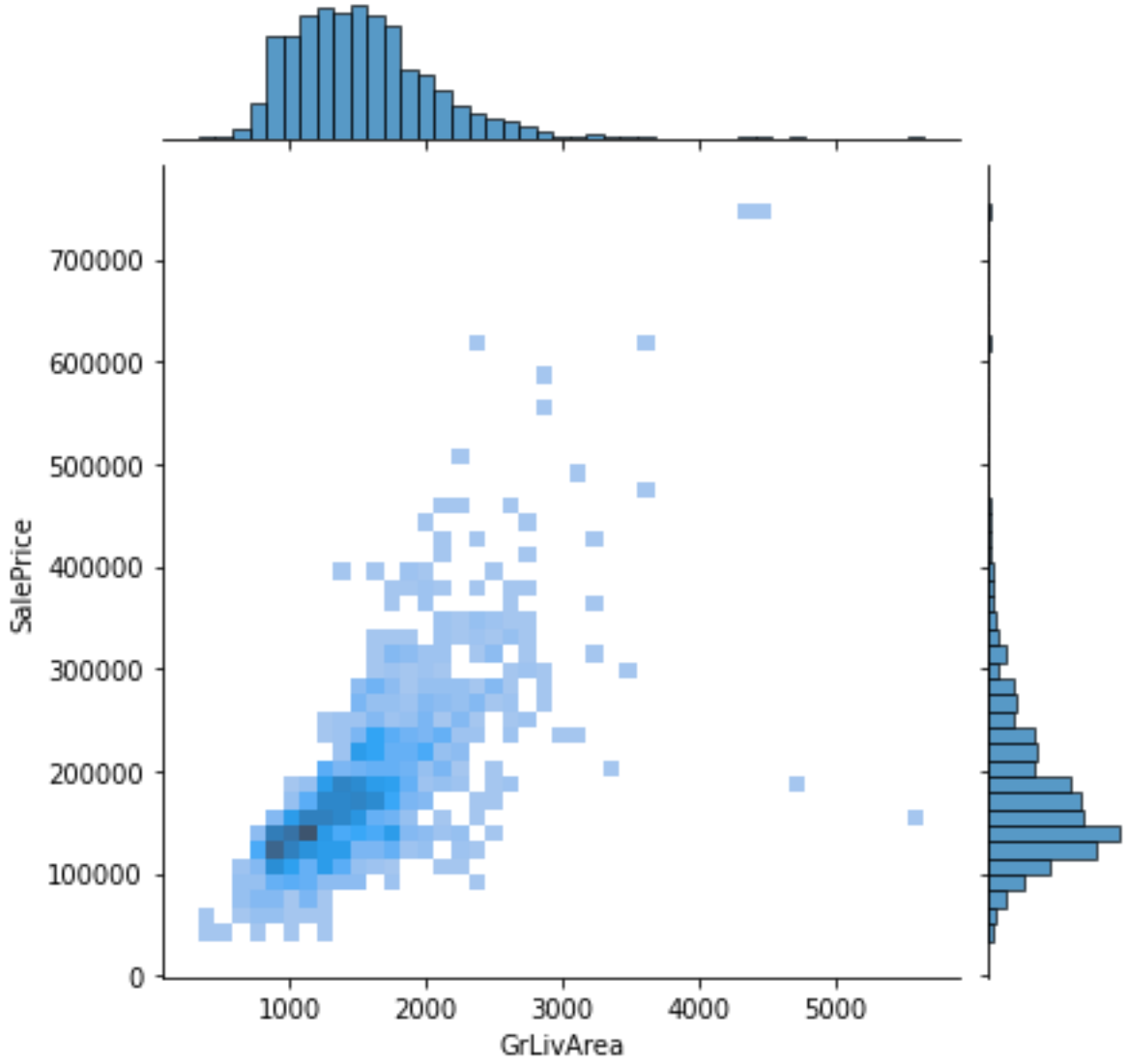
Sale Price is increasing with the TotalBsmtSF (Total square feet of basement area).



Sale Price of Y(Yes) type of CentralAir (Central air conditioning) is highest.



Sale Price is increasing with the 1stFlrSF (First Floor square feet).



Sale Price is increasing with the GrLivArea (Above grade (ground) living area square feet).

Interpretation of the Results

We can see by the all graphs that how independent features affect the sale price of house. Whose ratings is good has higher sale price excellent type of basement has a higher sale price than other type of basement. And so many independent variables are there who set the price for the house...

CONCLUSION

In Housing sale price project 81 columns and 1168 were there, if we see by excel sheet it is very hard to set the price of house but by using ML it is easy to predict the sale price of house or buy price of the house as well…

It is very hard to analysis that which of the independent variable affect the house sale price more and which independent variable affect the sale price less, but by using all the visualization tools like matplotlib and seaborn we can easily see by the graph that how sale price increases and decreases.

Predict sell price is a regression problem that is why I used the all-regression algorithm,

Random forest regressor model works good for housing project with 92.316% of accuracy, which is very good accuracy for the prediction.