Housing: Price prediction

Submitted by:

Miral Dhrafani

Internship 28

ACKNOWLEDGMENT

The project housing price prediction is a project to predict the housing price. A US based housing company decided to enter the Australian Market. They collected a data-set from the sale of houses in Australia. This data-set was provided to me by fliprobo Technologies as a part of my internship project. My mentor Khushboo Garg guided in various aspect of this project, she was of great help in framing the business problem and addressing the solution

INTRODUCTION

- Business Problem Framing

Real estate is the world biggest industry. One of the biggest concern in investing in the real estate industry is to study the structure and dynamics of the residential or commercial space. The price of the property depends on various factor such location, space, neighbourhood, proximity to school/transportation etc and also Due to constant fluctuation in demand and supply it is very difficult to identify the real worth of a property. To address this issue, data science comes as a very important tool to solve problem by predicting the accurate price of the property by analyzing various features and comparing it with the label. Predict the selling price with the help of historical data

- Conceptual Background of the Domain Problem

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

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

Housing or real estate industry is one of the most attractive industry in the world, almost everyone wants to invest their money into real estate, either as a source of passive income or building equity. One of the biggest concern in investing in the real estate is to study the structure and dynamics of the residential or commercial space. The price of a property is determined by various factor such as land Area, Location, Proximity, so it is very crucial to find the accurate market value considering all the factor. Data Science plays a major role in predicting the selling price building a Machine learning model who predict the selling price of the property based on all the features available in the dataset. Which help major players in Market to invest the money in the real estate properties

- ## Motivation for the Problem Undertaken

With increase in the mobility of man and business through out the world, demand in Real estate property also increase drastically, and business from all over the world wants to invest in foreign land by purchasing, commercial, residential, agricultural or barren lands. But the strategy for buying properties can be different based on the geographical location. So it is very important to study the market thoroughly before investing, because unlike any other market such as stock market, real estate requires huge amount of money for investing. Similar problem was faced by A US-based housing company named Surprise Housing has decided to enter the Australian market. However they want to predict the accurate market value. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For increasing their overall revenue, profits, improving their marketing strategies.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

Various Mathematical/ Analytical Modeling was performed through out the project which includes identifying the null values. Dealing with null values based on the data-set. We also needed to identify best features because the data-set contains 80 features which is quite big so we selected only top 75% of the features who are strongly correlated with the label data. We also identified skewness and detected outliers in the dataset, which we removed using various mathemaical functional such as cuberoot and zscore technique

- Data Sources and their formats

Data source was Flip Robo technologies. The project was assigned to me by my Mentor Khushboo Garg.

The data was collected by Surprise Housing Company for making the machine learning model to predict the price for entering into the new Australian market for real estate. The data contains 1168 rows and 81 columns. They are both in integer and string forms.

```
# Importing the training dataset
data_train = pd.read_csv(r'D:\Data Science\Internship - Flip Robo\Project-Housing_splitted\train.csv')
data_train
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	Misc
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1163	289	20	RL	NaN	9819	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	
1164	554	20	RL	67.0	8777	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	
1165	196	160	RL	24.0	2280	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	
1166	31	70	C (all)	50.0	8500	Pave	Pave	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	
1167	617	60	RL	NaN	7861	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	

1168 rows × 81 columns

Data Preprocessing Done

Two data-set was provided training and test data-set, training data-set contains 81 columns which includes 80 features and 1 label columns, test data contains 80 columns all of them are features data and we need to predict label based on the model we build. Data provided was bit noisy there were lots of null values, we treated null values based on the observation we made. There we around 35 object columns so we encoded those columns. We also removed the skewness and outliers from the continuous columns

- ## Data Inputs- Logic- Output Relationships

Data inputs are given in the dataset we need to filter both train and test data we can say the train data was the input and test is the output because we were need to train the model on the behalf of train data and put them in the test data. So we have done all the data cleaning steps to the test data too. After getting the best model put that model on test.

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

The assumption we made while building this model is that the dataset contains only residential properties and no commercial or agricultural properties were included in dataset

- ## Hardware and Software Requirements and Tools Used

A PC with pre-install Anaconda Navigator has been used, Jupiter Notebook was used as a platform to run Python, different python libraries were used such as pandas for data manipulation, numpy for calculations and sea-born and matplotlib for visualization.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

In this data-set we need to predict the selling price of property which is a continuous data so we have used various regression algorithm and we will choose only one regression model which will give the best r2 Score. We will also fine tune the model selected to further improve the accuracy score of the model

- Testing of Identified Approaches (Algorithms)

Algorithms which are used for training and testing are:

Random Forest Regressor

Gradient Boost Regressor

Decision Tree Regressor

AdaBoost Regressor

Linear Regression

- Run and Evaluate selected models

Following are the Algorithm used and snapshot of there code

```
DecisionTreeRegressor()
fit score : 1.0
r2 score 0.7454408788321328
mean absolute error 22350.548717948717
root mean squered error 30484.08687950031


RandomForestRegressor()
fit score : 0.9824459942002917
r2 score 0.8545263921302702
mean absolute error 15435.306923076925
root mean squered error 23044.694978195


GradientBoostingRegressor()
fit score : 0.9744003294259012
r2 score 0.8670638494305472
mean absolute error 14048.034068358025
root mean squered error 22029.28535614636


AdaBoostRegressor()
fit score : 0.8961956388107729
r2 score 0.8175049785306732
mean absolute error 17291.956799671618
root mean squered error 25810.97260299813


LinearRegression()
fit score : 0.9151043841976304
r2 score 0.8552466990158691
mean absolute error 16404.881075283203
root mean squered error 22987.5717261543
```

- Key Metrics for success in solving problem under consideration

The key metrics which are considered for selecting the best models are:

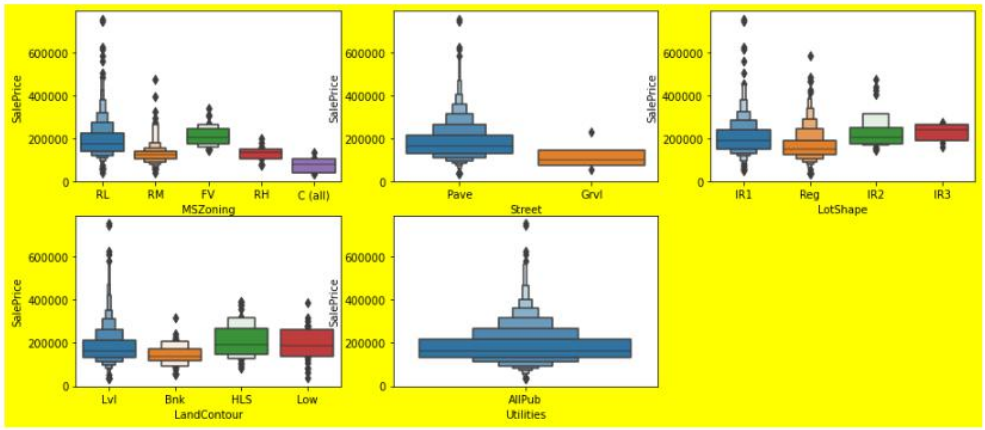
Fit score

R2_score.

- Visualizations

1) Distribution of Categorical Data

```
plt.figure(figsize=(15,45),facecolor='Yellow')
fignumber = 1
for i in data_train[cat_col[:5]]:
    if fignumber <= len(data_train[cat_col]):
        ax = plt.subplot((round(len(cat_col)/3,0)),3,fignumber)
        sns.boxenplot(data_train[i],y=data_train['SalePrice'])
    fignumber+=1
plt.show()
```



in the above image we have compared sales of price of the house with different categorical column:

Zoning - we can see that floating house are the most expensive houses with avergar sale price of 209479 and commercial properties are the cheapest with average selling price of 75209 dollars

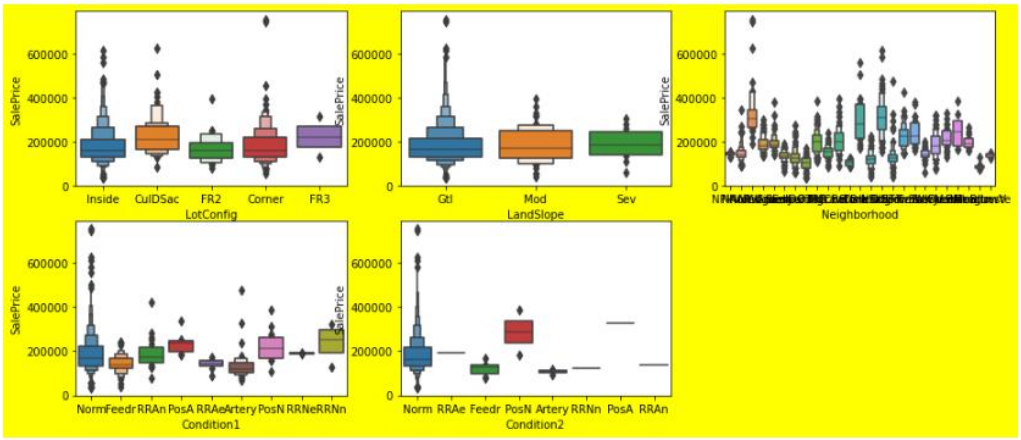
properties on the paved road are more expensive than gravel road

sales price of the house are inversely proportionate to its shape, the more irregular shape, more expensive houses are

Hillside and low depression houses are more expensive as compared to near flat and banked houses

there is only single variable on utilities column, so there is no point in keeping this column so we will delete this column

```
plt.figure(figsize=(15,45),facecolor='Yellow')
fignumber = 1
for i in data_train[cat_col[5:10]]:
    if fignumber <= len(data_train[cat_col]):
        ax = plt.subplot((round(len(cat_col)/3,0)),3,fignumber)
        sns.boxenplot(data_train[i],y=data_train['SalePrice'])
    fignumber+=1
plt.show()
```



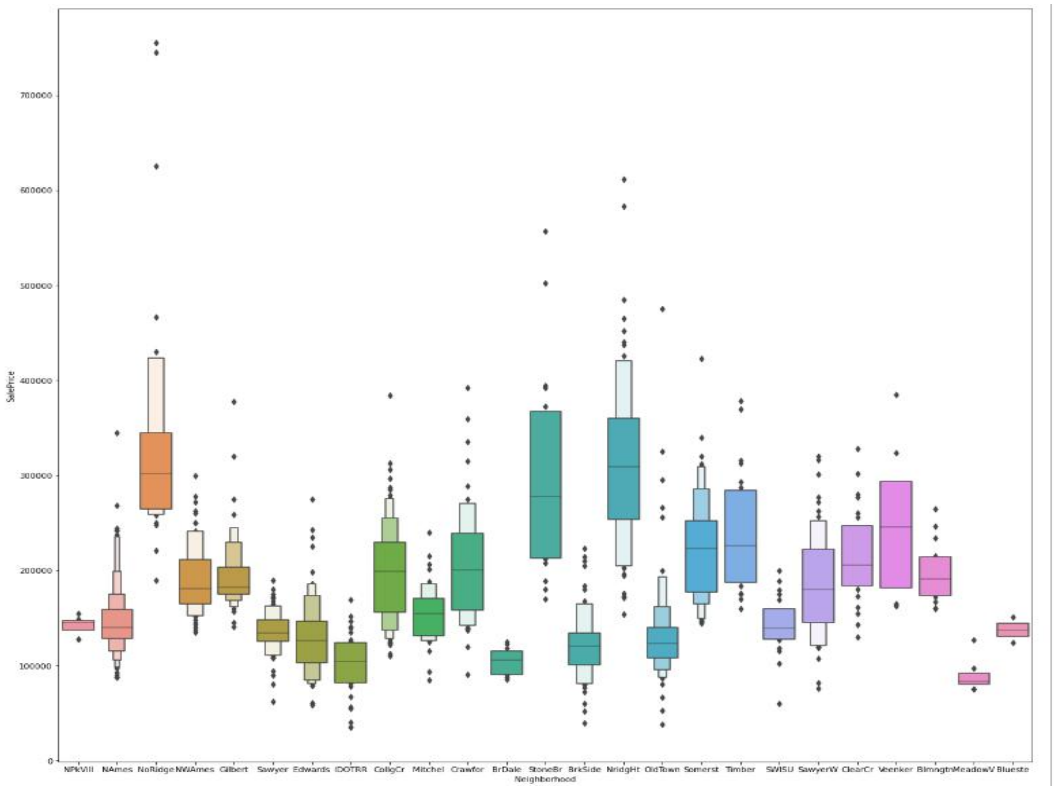
Following are the observation made from above graphs:

there is no significant impact of land configure and land slope in the sale price of the property

properties near railroad and feature park are more expensive as compared to other areas

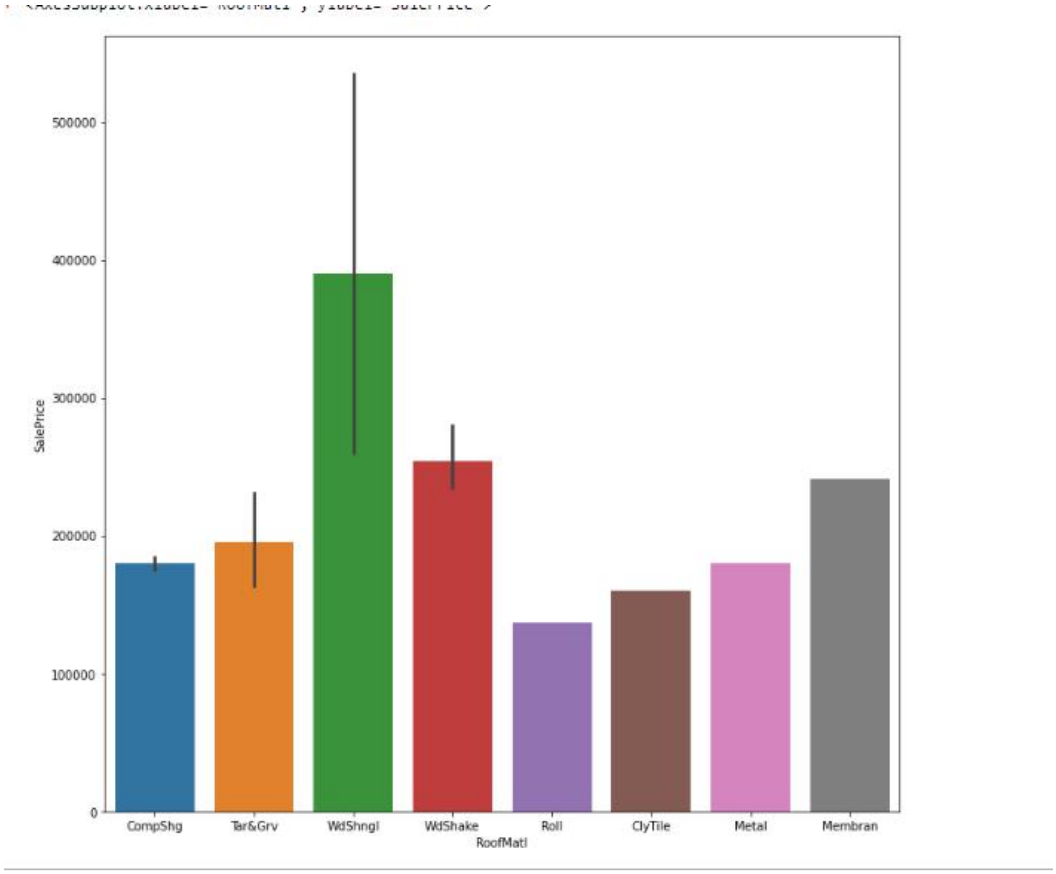
there are lot of variables in Neighbourhood features so we will plot it separately

2) Sales price and Neighbourhood



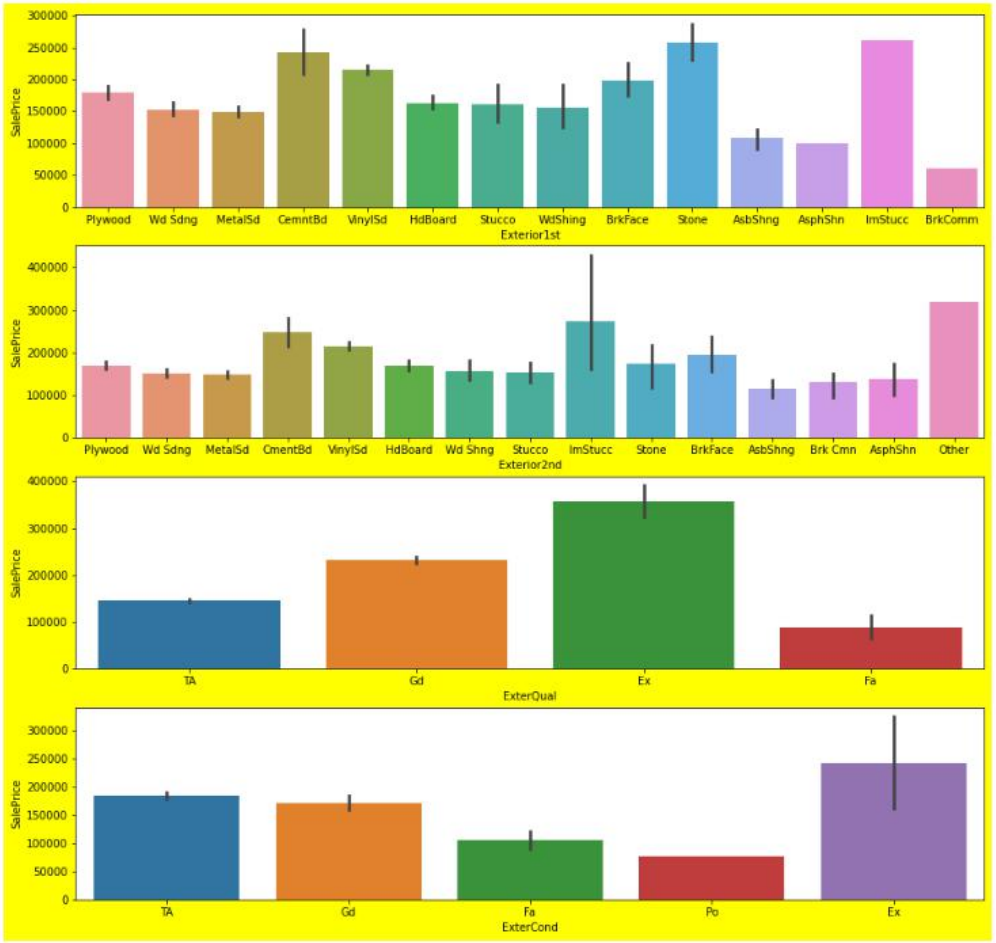
localities like Northridge Heights, Stone Brook, Northridge have highest selling prices, whereas localities like Meadow Village, Bluestem and Briardale are some of the cheapest locality

3) Roofing Material and Sales Price



Selling price for property is higher when wood is used as the primary material for roof, on the other hand when roll,clay or tiles are used, the selling price are tends to be less

4) Exterior and Sales Price

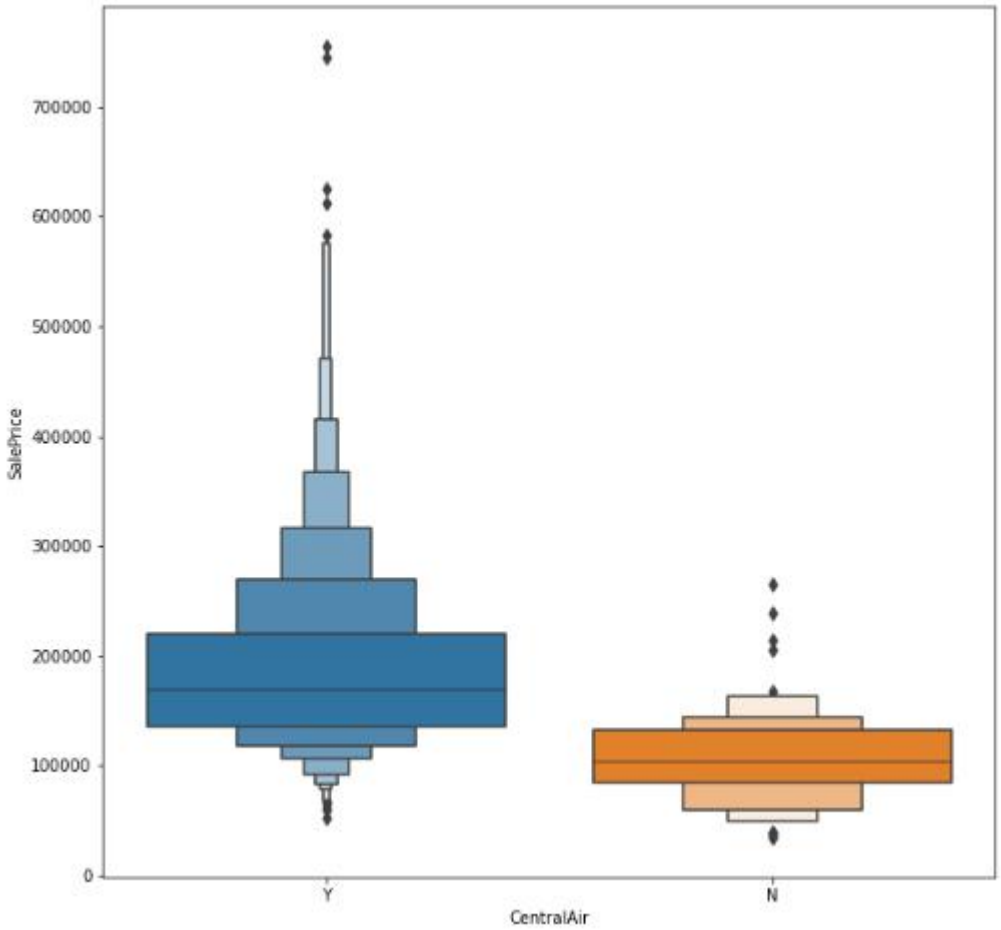


After analyzing all the expects of exterior work with selling price we found that

if sturdy material such as stone, cement, vinyl and brick face are used for exterior the selling price tends to up

exterior condition and quality are directly proportionate to the selling price, if the exterior condition is excellent, the selling price of those properties are highest, however if the condition or quality goes down the selling price

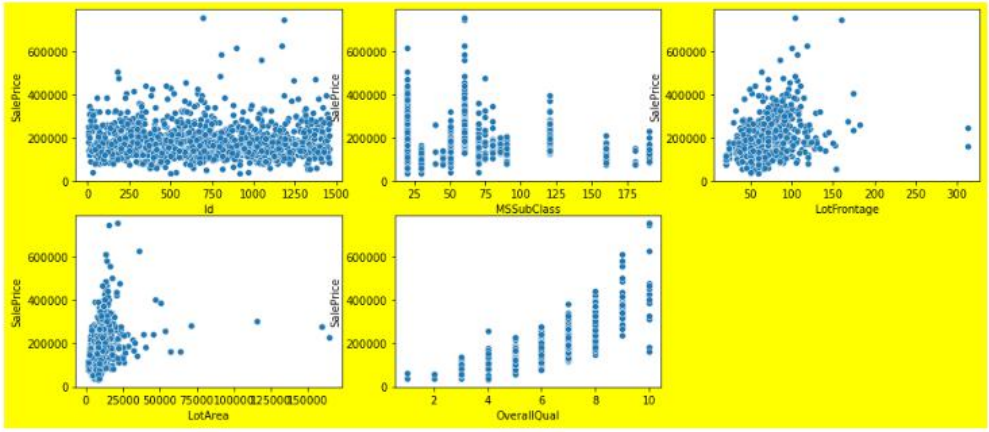
5) Air conditioning and Sales Price



As we can see, the unit that has centralized air conditioning are more expensive as compared to one without it

6) Continuous data and Sales Price

```
plt.figure(figsize=(15,45),facecolor='Yellow')
fignumber = 1
for i in data_train[cont_col[:5]]:
    if fignumber <= len(data_train[cat_col]):
        ax = plt.subplot((round(len(cat_col)/3,0)),3,fignumber)
        sns.scatterplot(data_train[i],y=data_train['SalePrice'])
    fignumber+=1
plt.show()
```



AS per the above scatter plot:

ID field is irrelevant so we will delete that column from the data-set

Type of dwelling in the property does not show any significant influence on the selling price

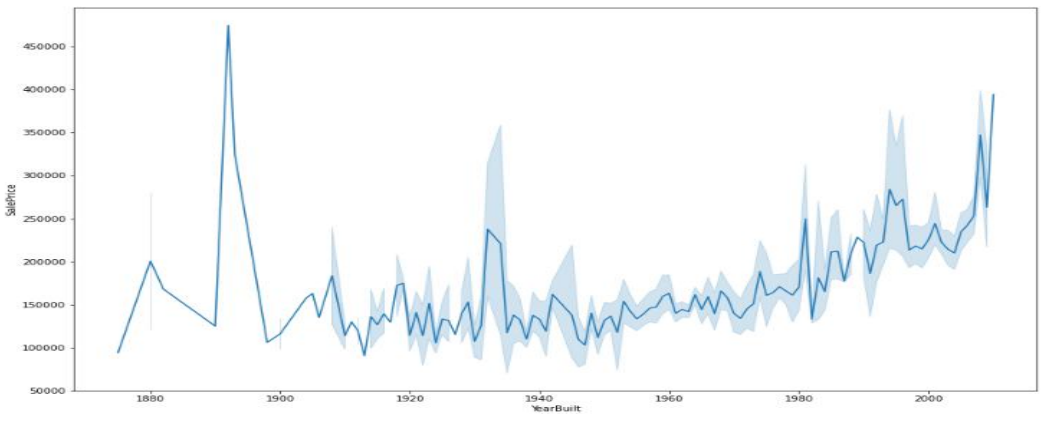
Lot fontage significant impact on the selling price, less lot font-age means cheaper selling price

Although lot area and selling price are positively correlated, but we can see just small increase in lot area have significant impact on price

As expected, selling price of the property depends on the overall quality of the house, poor quality of the property leads to less market value

7) Year Built and Sales Price

```
plt.figure(figsize=(15,10))
sns.lineplot(data_train['YearBuilt'],data_train['SalePrice'])
plt.show()
```

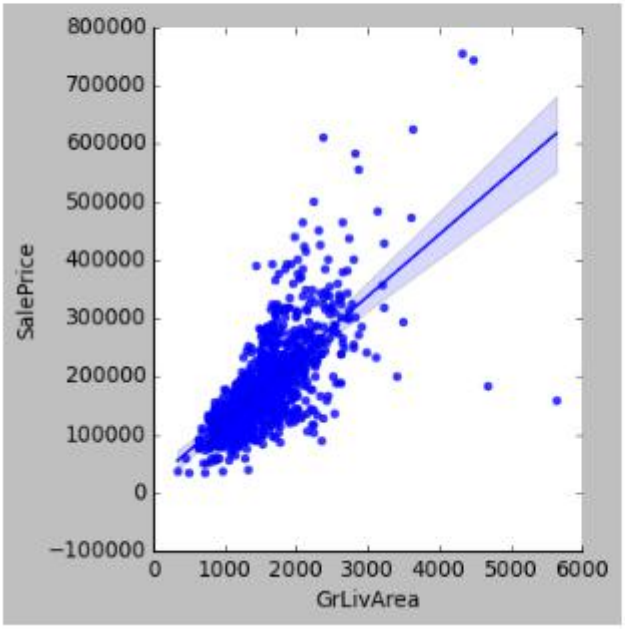


When we check if the year of built have any impact on he selling price through the line plot we can see the upward trend on the line which means that new properties are more likely to sell at higher price as compared to the older one,

8) Year Built and Sales Price

```
1 plt.figure(figsize=(30,20))
2 sns.lmplot(x='GrLivArea',y='SalePrice',data=data_train)
3 plt.show()
```

```
<Figure size 2400x1600 with 0 Axes>
```



Same upward trend can be seen on Ground living Area, it means that buyer are willing to pay higher price for more ground area

CONCLUSION

- ## Key Findings and Conclusions of the Study

Many factor such as exterior quality, Neighborhood, ground living area plays a major role in determining the prices, also the age of property is also considered by the buyer before buying any property, as people usually prefer newer property over older property . also air conditioning property are likely to sell at higher prices,

So to conclude we can conclude that not only material or land area determines the price but also the features provided like Air conditioning is also very important factor in evaluating the price of the property

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

Because the data is huge the number of rows and columns are also high so these make my understanding of data increased. I have tried many things to clean up the data than find with are used for prediction and make the model powerful I have correlate many useful independent variables with the targeted variable so that I can understand the data more. After visualization I have correlate the data with target in numbers to understand which independent variables gives how much impact on the targeted variable. So after all of these I used standard scaler to scale the data after scaling the data train the model than used the different machine learning models. And get the best score in random forest regressor than doing hyper parameter tuning etc. these steps improve my machine learning and model building skill. And these model surely will help the consumers, companies, property agents to estimate the price of a particular property..

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

Because we get the good r2 score and we have choose the best one model. We can assume the estimated price are almost correct. But we all know property prices are very dynamic it could change with many other factors. So we can took the estimated prices but also need to see the market trend.