FLIP ROBO

House Price Prediction

Submitted by:

Dishant Doshi

ACKNOWLEDGMENT

I complete this project but its not possible without helping from the organization and take help from site htpp://scikit-learn.org

I would like to thank who helped me to complete this project. I also like to thanks Flip Robo, Benglore for giving me this project and their support.

I would also like to thank Mr. Shubham Yadav for guide me during my whole project and solve all the query.

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.

We have to build model that will predict the price of the House. And price of the house is depending on which features. According to that Company will decide their strategy and try to build the house.

➢ 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 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?

➢ Motivation for the Problem Undertaken: -

House is very primary and necessary requirement of the anyone. Our main objective for this project is to make model that will predict the house price. And on which features house price is depends.so that company will fix the price. For model preparation client provides some data according we have to create model. Client wants some prediction for fixing the house price and from that he will decide the price and strategy to attract the customers. Usually, house price is very from the location, area. But client want other assumption too.

So, considering this Objective, with help of machine learning I will create model that will help to predict the house price. I will use many regressions model to get the optimistic result.

Analytical Problem Framing

➢ Mathematical/ Analytical Modelling of the Problem: -

Here I had done EDA process first to understand the data and identify the hidden pattern or information from the data by using various charts. After that I will check the weather, my data is right skewed, left skewed or not. Also, I will find the outliers of the data. And after that I will do different Data Pre-process which will be useful to built the model.

From that I will be build the regression model to predict the house price.

➢ Data Sources and Their Formats: -

Here my data is in csv format which contain 1168 rows and 81 features.

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	Mc
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	

5 rows × 81 columns

Data Information: -

From this information we can check the Datatype of my features and how much null data is present in my data set.

38 Numerical and 43 Categorical Features

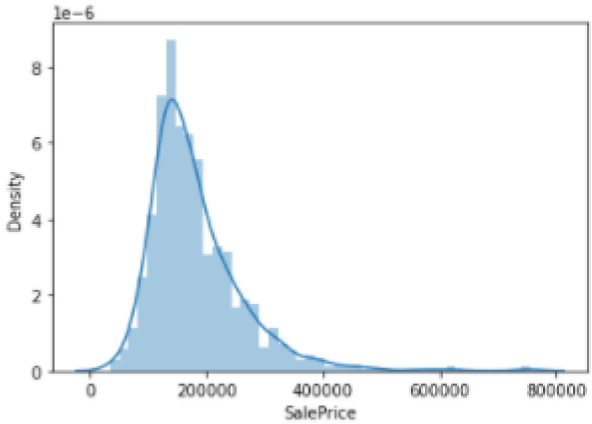
```
#Missinf Features of Train Dataset
missing_features=[features for features in df.columns if df[features].isnull().sum()>1]

for feature in missing_features:
    print(feature,np.round(df[feature].isnull().mean()*100,4),'% missing values')
```

```
LotFrontage 18.3219 % missing values
Alley 93.4075 % missing values
MasVnrType 0.5993 % missing values
MasVnrArea 0.5993 % missing values
BsmtQual 2.5685 % missing values
BsmtCond 2.5685 % missing values
BsmtExposure 2.6541 % missing values
BsmtFinType1 2.5685 % missing values
BsmtFinType2 2.6541 % missing values
FireplaceQu 47.1747 % missing values
GarageType 5.4795 % missing values
GarageYrBlt 5.4795 % missing values
GarageFinish 5.4795 % missing values
GarageQual 5.4795 % missing values
GarageCond 5.4795 % missing values
PoolQC 99.4007 % missing values
Fence 79.7089 % missing values
MiscFeature 96.2329 % missing values
```

Features like 'MiscFeature', 'Fence', 'PoolQC', 'Alley' have more than 50% null data.

```
: sns.distplot(df['SalePrice'])
: <AxesSubplot:xlabel='SalePrice', ylabel='Density'>
```



From Our Target Variable SalePrice Distribution chart we can see that data is right skews present and heavy outliers also present in the SalePrice.

Data Pre-processing: -

Data Pre-processing is the important the step in Data Science Model. Data is usually in the unstructured format so that we have convert into structured data for that there are many steps

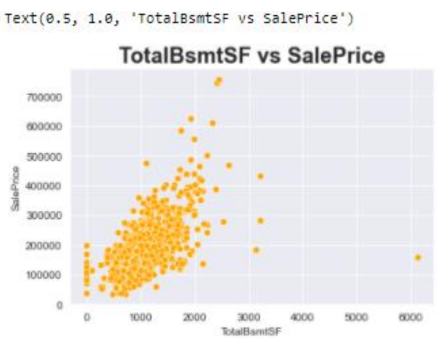
- Handling Missing Values
- Features Selection
- Handling Multicollinearity
- Removing Outliers
- Removing Skewness

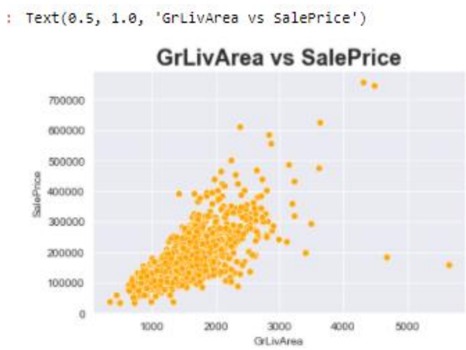
Need Of Pre-Processing: -

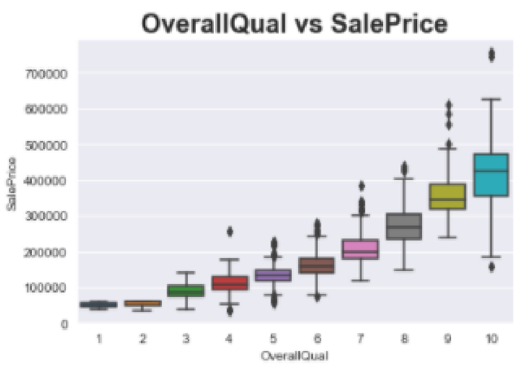
For achieving best result from the applied machine learning we have to apply the data to model in proper manner. For example, Random Forest cannot run for the null values. We cannot apply the null data in to the random forest regresior.for that we have to manage the null data from origin. And data should be in one format than it can apply in all the machine learning algorithms so that we can select best result comes out from the diff diff models.

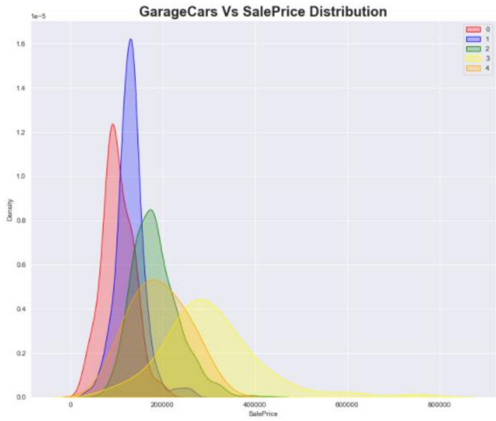
➢ Data Inputs- Logic- Output Relationships: -

To find out the relation between the target variable and input variable I used EDA process in which used visualization. From the charts we can see that how the price is affected on the features. For that I used graphs like Scatter plot, Box Plot, Bar Plot, Joint plot.

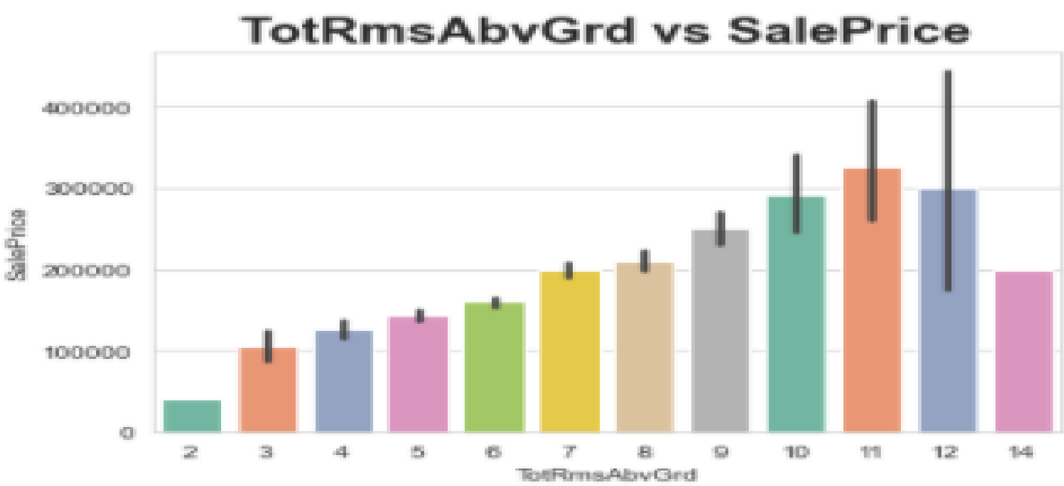








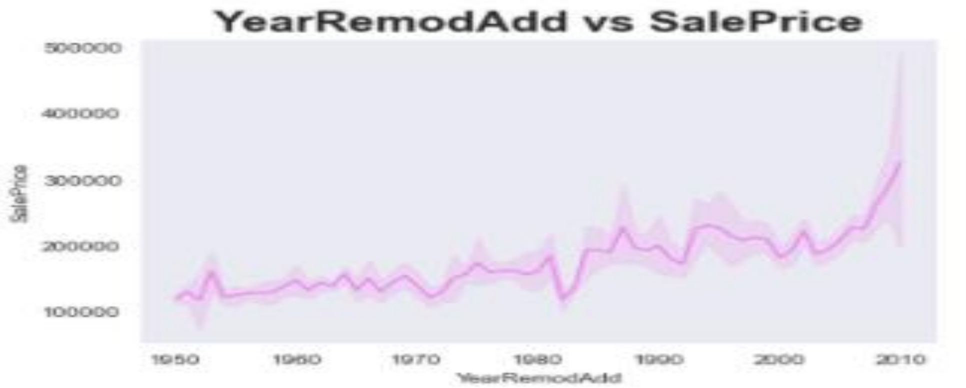
.



```
Text(0.5, 1.0, 'Fireplaces vs SalePrice')
```


```
Text(0.5, 1.0, 'YearBuilt vs SalePrice')
```


```
Text(0.5, 1.0, 'YearRemodAdd vs SalePrice')
```



➢ Observation: -

- As The TotalBsmtSF means square feet of Basement increases price of the House also Increases.
- GrLivArea means above ground living area increases with that price also increases
- With Increases in OverallQual house price also increases.
- No of rooms 2-11 price increases as no of room increases but after 11 price decreases.
- Price in 1880-1990 increases after sudden it decreases and never reach at previous price.
- As the House is Remodel or if not change than consider their construction data price also increases.

➢ Hardware and Software Requirements and Tools Used: -

I used my Corei3 processor and 8gb Ram as Hardware. For software I used Python3

For tool I used below list of libraries: NumPy, Pandas, Matplotlib, Seaborn, Sklearn, scikitplot

Model/s Development and Evaluation

Here my Target variable is House Price so it continues variable so I have to use Regression model. Here I used many algorithms like linear regression, Random Forest, AdaboostRegressor LGBM Regression etc. From all them I get good score and good performance metrics in **LGBM Regression**. Some models like lasso, linear and random forest. There is overfitting or underfitting in this model.

Here I have small dataset so model didn't learn too much if I had large dataset so model can perform well.

For checking model is in overfitting or not I used KFold method.my dataset is small so I used 5 kFold but if my dataset would big than I used 10 KFold.

➢ Testing of Identified Approaches (Algorithms)

➢ Following list of Algorithms: -

- Linear Regression
- Decision Tree Regressors
- KNearest Neighbours Regressor
- Random Forest Regressor
- AdaBoost Regressor
- Gradient Boosting Regressor
- LGBM Regressor

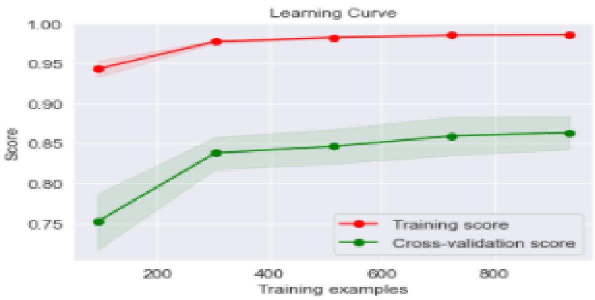
➢ Run and evaluate selected models: -

Result of all the model given below

	Model Name	CV Score	R2 Score	Mean Absoluate Error	Root Mean Squared Error
0	Linear Regression	0.823	85.13	0.104	0.156
1	DTC	0.687	67.25	0.053	0.231
2	KNN	0.619	61.85	0.180	0.250
3	Random Forest Regressor	0.855	83.78	0.107	0.163
4	AdaBoost Regressor	0.800	79.66	0.134	0.182
5	Gradient Boosting Regressor	0.862	86.34	0.099	0.150
6	LGBM Regressor	0.860	86.27	0.098	0.146
7	Lasso	0.720	73.41	0.043	0.208
8	Ridge	0.839	85.13	0.104	0.156

From all the result of all the model I get the best result in **the LGBM Regressor.**

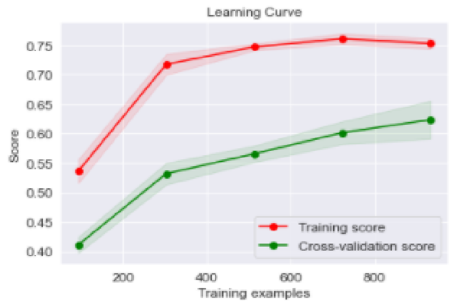
```
---- LGBMRegressor() ----
Taining Score:- 98.54264613574392
Mean Absolute Error 0.09882460957298 91
Mean Squared Error 0.02251315011592213
Test Root Mean Squared Erro 0.1500438273169614
Cross Validation Score 0.860554939524485
R2 Score 86.27288203948528
Test Score 86.27288203948528
Model Peformance Cure
```




```
---- LinearRegression() ----
Taining Score:- 88.70483776932747
Mean Absolute Error 0.10489570257397161
Mean Squared Error 0.02438085116581712
Test Root Mean Squared Erro 0.15614368756314526
Cross Validation Score 0.8237244055332059
R2 Score 85.13407416520414
Test Score 85.13407416520414
Model Peformance Cure
```


```
---- DecisionTreeRegressor() ----
Taining Score:- 100.0
Mean Absolute Error 0.15857631460796578
Mean Squared Error 0.04966287066031362
Test Root Mean Squared Erro 0.22285167861228602
Cross Validation Score 0.6879901718263592
R2 Score 69.71867196275802
Test Score 69.71867196275802
Model Peformance Cure
```

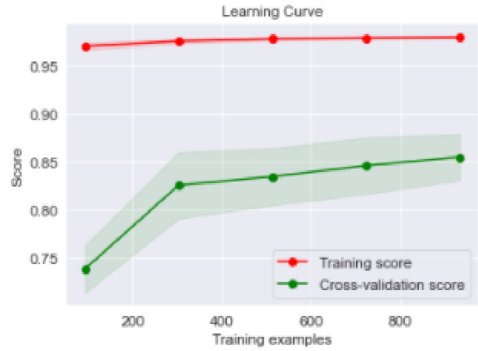
```
---- KNeighborsRegressor() ----
Taining Score:- 75.082254288179
Mean Absolute Error 0.18091989215225326
Mean Squared Error 0.06256620969288276
Test Root Mean Squared Erro 0.25013238433454144
Cross Validation Score 0.6196414008646592
R2 Score 61.85101878794441
Test Score 61.85101878794441
Model Peformance Cure
```




```
---- Lasso() ----
Taining Score:- 75.09683826376731
Mean Absolute Error 0.1377955918856526
Mean Squared Error 0.043603989095831484
Test Root Mean Squared Erro 0.20881568211183632
Cross Validation Score 0.72213709025341
R2 Score 73.4130009001204
Test Score 73.4130009001204
Model Peformance Cure
```

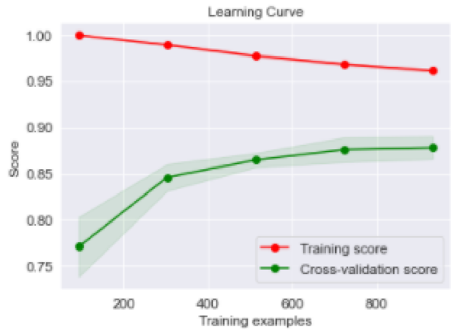

```
---- Ridge() ----
Taining Score:- 88.67814940738063
Mean Absolute Error 0.10436437184030613
Mean Squared Error 0.02438450828479228
Test Root Mean Squared Erro 0.15615539787273536
Cross Validation Score 0.8395100759482375
R2 Score 85.13184428163346
Test Score 85.13184428163346
Model Peformance Cure
```


```
---- RandomForestRegressor() ----
Taining Score:- 97.96552074568137
Mean Absolute Error 0.10716030697911005
Mean Squared Error 0.026498030052547798
Test Root Mean Squared Erro 0.16278215520304368
Cross Validation Score 0.855077558067163
R2 Score 83.84315023088058
Test Score 83.84315023088058
Model Peformance Cure
```




```
---- AdaBoostRegressor() ----
Taining Score:- 86.99544223730298
Mean Absolute Error 0.1385538671467355
Mean Squared Error 0.03514609412104412
Test Root Mean Squared Erro 0.1874729156999595
Cross Validation Score 0.8000356363563821
R2 Score 78.5700989258844
Test Score 78.5700989258844
Model Peformance Cure
```


```
---- GradientBoostingRegressor() ----
Taining Score:- 96.44253380396977
Mean Absolute Error 0.10009299878001891
Mean Squared Error 0.023099530737955096
Test Root Mean Squared Erro 0.15198529776907732
Cross Validation Score 0.8621601815898282
R2 Score 85.91534362629301
Test Score 85.91534362629301
Model Peformance Cure
```



.

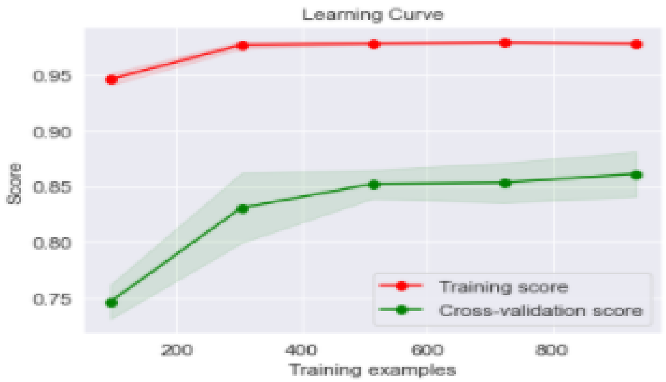
➢ Key Metrics for success in solving problem under consideration: -

- Absolute Error: - It gives the difference the true value and measured value.
- Squared Error: -It tells how close the regressor line to the points.it takes the distance that points to the regression line and that difference is the error.
- R2-Score: -R2 is the graphical representation of the how close the data are fitted to the regression line.

➢ Hyper Tunning for Best Score: -

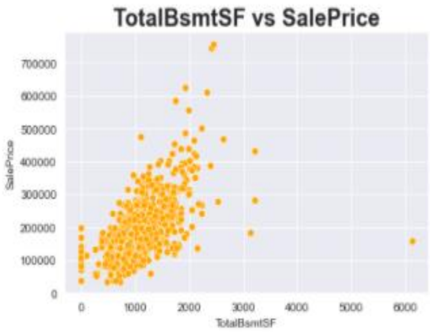
```
fun(lgbm1)
```

```
---- LGBMRegressor(importance_type='mse', max_depth=8) ----
Taining Score:- 97.64843231443152
Mean Absolute Error 0.09967149265441701
Mean Squared Error 0.022385013235386997
Test Root Mean Squared Erro 0.14961621982721993
Cross Validation Score 0.8615233814391188
R2 Score 86.35101193535242
Test Score 86.35101193535242
Model Peformance Cure
```

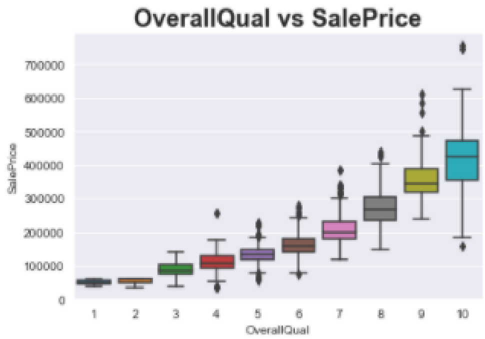


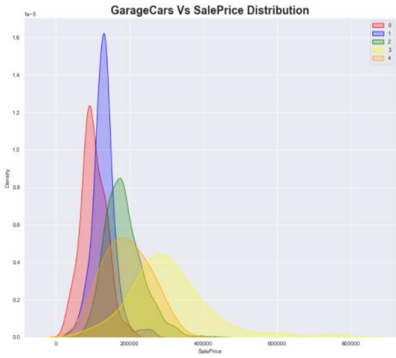
➢ Visualizations: -

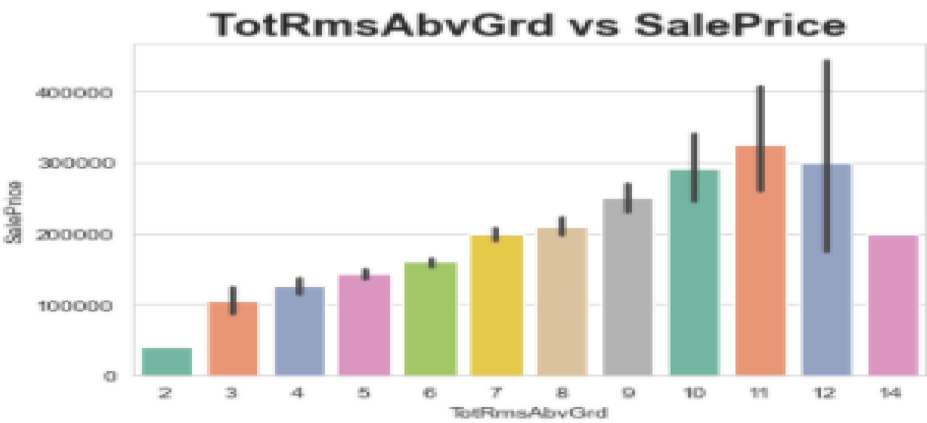
```
Text(0.5, 1.0, 'TotalBsmtSF vs SalePrice')
```




```
Text(0.5, 1.0, 'GrLivArea vs SalePrice')
```



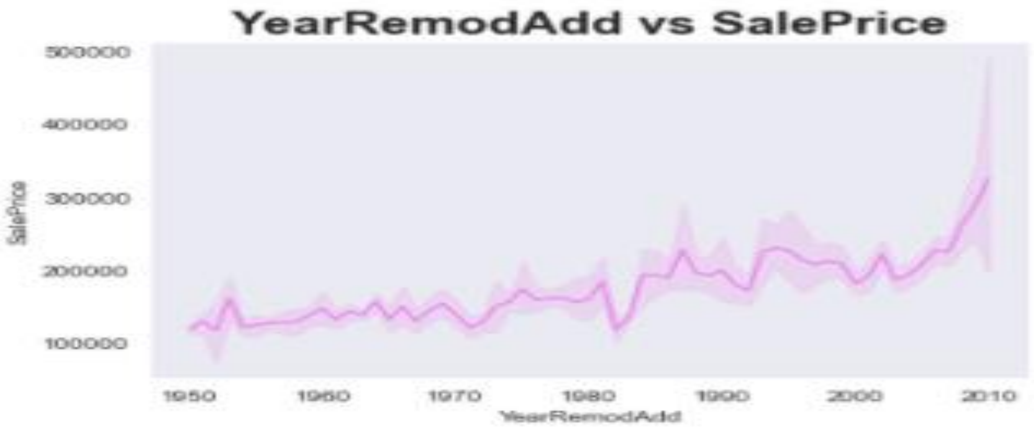


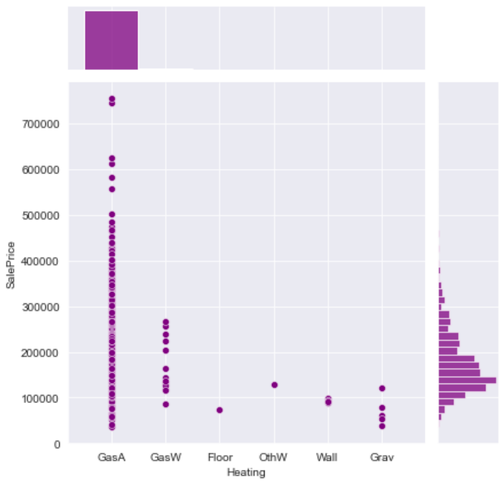


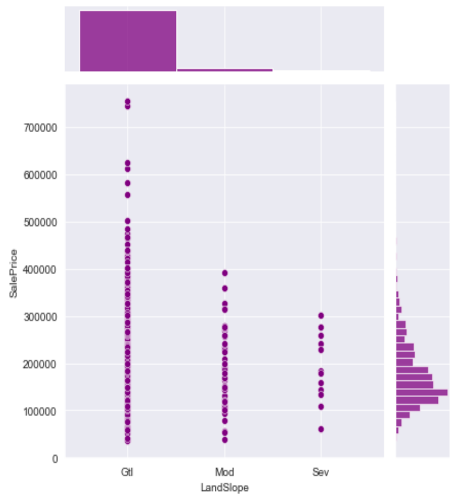

```
Text(0.5, 1.0, 'YearBuilt vs SalePrice')
```

```
Text(0.5, 1.0, 'Fireplaces vs SalePrice')
```

```
Text(0.5, 1.0, 'YearRemodAdd vs SalePrice')
```







➢ Interpretation of the Results: -

- As The TotalBsmtSF means square feet of Basement increases price of the House also Increases.
- GrLivArea means above ground living area increases with that price also increases
- With Increases in OverallQual house price also increases.
- No of rooms 2-11 price increases as no of room increases but after 11 price decreases.
- Price in 1880-1990 increases after sudden it decreases and never reach at previous price.
- As the House is Remodel or if not change than consider their construction data price also increases.

CONCLUSION

➢ Key Findings and Conclusions of the Study: -

From the Project I learn many things like which Data Pre processing works and its important. How to handle the skewed data and outliers.

1-Total Rooms Above Ground-As the room no. increasing the average price is also increasing till 11th room after that price start decreasing

2-Bedroom Above Ground-For the 0,4,8 Bedroom price is high and price is very less for 6 and 2

3-Kitchen Above Ground-as the no of kitchen is increasing the price is reducing and mostly people take one kitchen only

4-In Basement full bathroom and half bathrooms as the bathroom size increasing the price is also increasing

5-Fireplaces-As the fireplaces increasing the sale price is also increasing

6-PoolArea-as big the pool the more costly the house

7-YRsold-the price was high in 2006 as compare to old year prices described in 2008-10

8-MOSold-most of the people who sold their home in 09 month they got high price and people who sold there home on 4th month got less price

9-Electrical-Most of the properties have standard circuit breakers and having highest average sale price of 170000.

10-Properties with poor fuse box system and mixed system have less than 10000 sale prices.

11-Heating-Heating in the wall or hot water/steam is associated with very low houses prices. Gas Formed warm air appears to drive a higher sales price

12-Central AC- The properties which have AC will have higher price that the ones which don’t have

13-LandSlope- From this chart we can see that People like to live in General Slope area and very few people live in Serve Slope Are.so density of the people increases in the General slope so prices are high there and serve slope like valley area few people live there so prices are usually low there.

14-Street-Most of the people like to live in that type of house whose connected street should be paved.

These are some few things.

➢ Learning Outcomes of the Study in respect of Data Science: -

- From this Project I learned tree base algorithms works good as compare to others. Model like lasso and ridge are going in overfitting.
- I got good all over result like cv score, r2score and less RSME in LGBM Regressor.
- Here data also have many features which having too much null values. But I overcome this problem at last.
- Here data set is very less but at last I built model which perform very well.
- From the model I learn how the price is depends on Basement area, living area and fire places.

➢ Limitations of this work and Scope for Future Work: -

- Limitation of this project is this model can apply on that area only where data is coming from. We cannot use globally.
- If we had or we can add the geological inputs like Longitude and Latitude than with the help of Maps we can visualize very well in future