FLIP ROBO

Housing Project

Submitted by:

Rakesh Chaudhary

ACKNOWLEDGMENT

I am over helmed in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

I would like to express my special thanks of gratitude to my **FlipRobo Team** as well as our **Internship24 SME** who gave me the golden opportunity to do this wonderful project on the topic (Autralian Housing Price Prediction), which also helped me in doing a lot of Research and I came to know about so many new things. I am really thankful to them.

Any attempt at any level can 't be satisfactorily completed without the support and guidance of **DataTrained Tutor** and **My FlibRobo Team**.

I would like to thank **DataTrained Team** who helped me a lot in gathering different information, collecting data and guiding me from time to time in making this project, despite of their busy schedules, they gave me different ideas in making this project unique.

##Data Source

The Source of Data is from Ames Housing dataset

INTRODUCTION

- Business Problem Framing

AUS-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.

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.

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.

Conceptual Background of the Domain Problem

> At a time when demand for homes is so high, and supply so low, it can be a battle to upgrade to a bigger, better property.

However, especially with COVID-19 forcing us to spend more time indoors, there are more prospective buyers than ever dreaming of the respite of a lovely backyard, more space at home to give us a break from our nearest and dearest, and extra room for a study area to allow us to work comfortably. And also our children are getting bigger and it would be nice if we had more space so they could bring home their teenage friends and have their own space. COVID delayed things a bit but, in the end, we took on a buyer's advocate to help us.

Review of Literature

Another strand of research has examined the housing market as part of structural macroeconometric models. Examples include Jovanoski, Stoney and Downes (1997), Powell and Murphy (1997) or the Reserve Bank of Australia's (RBA's) new MARTIN model (Cusbert and Kendall 2018). Our approach differs in that we look at the housing market in more detail, including variables that these models often exclude, such as building approvals, completions or the vacancy rate. The macroeconomic models are designed to settle down to an explicit steady state with simple properties, such as constant relative prices or expenditure shares. That facilitates the models' application to a wide range of questions. However, in discussing housing-specific issues, it is not desirable to have important results driven by assumptions for which evidence is weak. We do not constrain the steady state of our model except when the data suggest this is realistic.

These alternative approaches have advantages relative to our approach. Larger models allow more variables to be endogenous, allow for feedback and can answer a wider range of questions. Smaller models are less reliant on chains of causation, which can be as fragile as their weakest link. And more focused models allow examination of specific estimates in more detail. These different approaches are complementary. When a range of different approaches support similar results, we have more confidence in the conclusions. Accordingly, we discuss specific points of agreement and disagreement where they arise in the discussion.

In common with most previous Australian research on housing markets, and most macroeconomic forecasting, we focus on single-equation least squares estimates using aggregate quarterly data. Identification is typically through lags and *a priori* reasoning. For example, rents are explained by lagged vacancies and contemporaneous income, on the assumption that the right-hand side variables are weakly exogenous. Given that the future does not determine the past, that strikes us as plausible for most lagged variables. This argument is less compelling when the right-hand side variables are forward-looking, such as investment or asset prices. However, we are not aware of clear evidence or strong arguments that weak exogeneity fails in the relevant equations or of useful instruments that might rectify this. One alternative to our approach would be to assume model-consistent expectations, but that seems unrealistic. We recognise that finding ‘X regularly precedes Y’ does not prove that ‘X causes Y’. However, in the absence of reasonable arguments to the contrary, the latter statement seems a natural hypothesis to maintain pending more definitive tests. This approach – essentially that of structural macroeconometric modelling – rests on methodological assumptions that are controversial, but that debate is outside the scope of this paper.

Background: The company wants to enter the Australian Market and hence are looking at prospective properties to buy. They want to understand what are the factors affecting the prices and how exactly are those factors influencing it. The company would then manipulate the strategy of the firm and concentrate on areas that will yield high return.

• Motivation for the Problem Undertaken

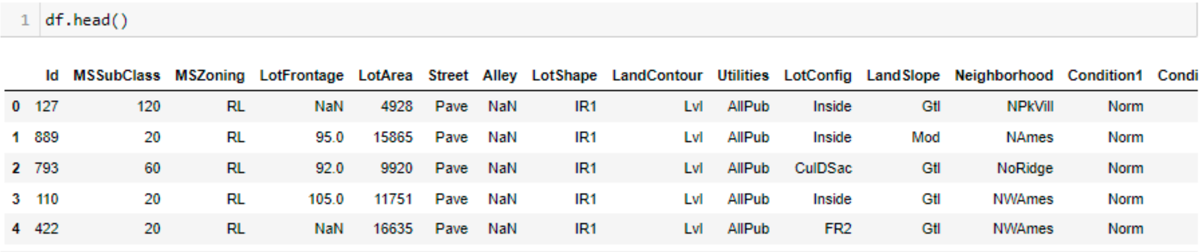
Growing unaffordability of housing has become one of the major challenges for metropolitan cities around the world. In order to gain a better understanding of the commercialized housing market we are currently facing, we want to figure out what are the top influential factors of the housing price. Apart from the more obvious driving forces such as the inflation and the scarcity of land, there are also a number of variables that are worth looking into. Therefore, we choose to study the house prices predicting problem. Our object is to discuss the major factors that affect housing price and make precise predictions for it. We use 80 explanatory variables including almost every aspect of residential homes in Australian market. So that to maintain the transparency among customers and also the comparison can be made easy through by this model. If customers find the price of house at some given website higher than the price predicted by the model, so he can reject the house.

Methods of both statistical regression models and machine learning regression models are applied and further compared according to their performance to better estimate the final price of each house. The model provides price prediction based on similar comparables of people’s dream houses, which allows both buyers and sellers to better negotiate home prices according to market trend.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

Our Data:-

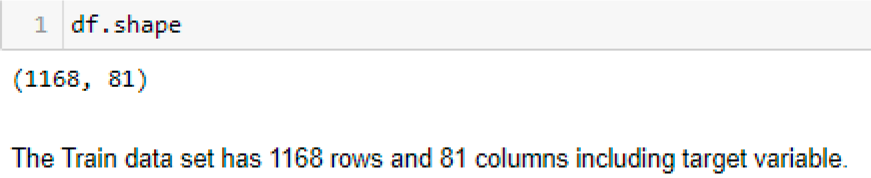


```
df.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condi
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	

Statistical or Mathematical exoploration is the first step of data analysis and typically involves summarizing the main characteristics of the data set, including it's size,accuracy and initial patterns in the data and other attributes.

Checking shape:-

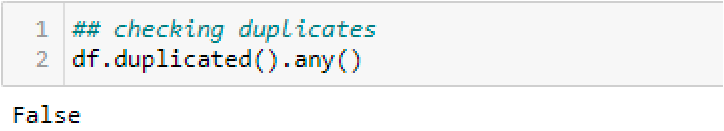


```
df.shape
```

```
(1168, 81)
```

The Train data set has 1168 rows and 81 columns including target variable.

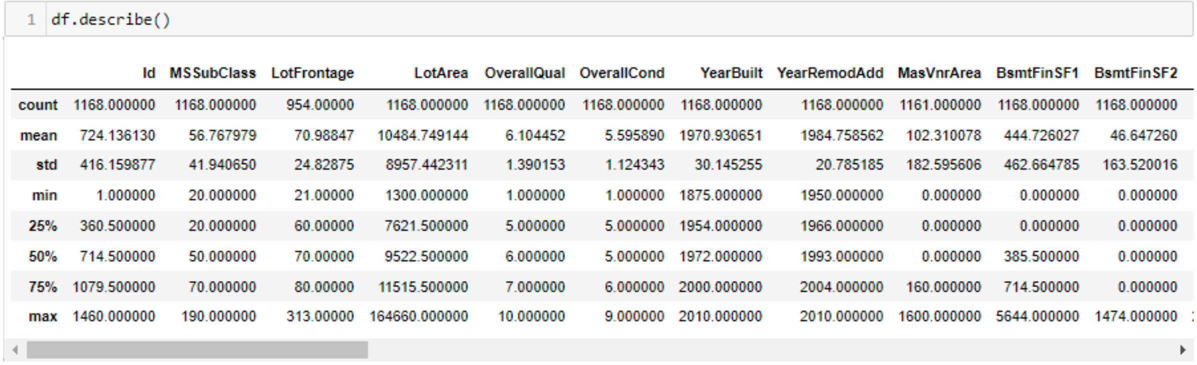
Checking Duplicates:-



```
## checking duplicates
df.duplicated().any()
```

```
False
```

UnderStanding Summary of Stats :-



```
df.describe()
```

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2
count	1168.000000	1168.000000	954.00000	1168.000000	1168.000000	1168.000000	1168.000000	1168.000000	1161.000000	1168.000000	1168.000000
mean	724.136130	56.767979	70.98847	10484.749144	6.104452	5.595890	1970.930651	1984.758562	102.310078	444.726027	46.647260
std	416.159877	41.940650	24.82875	8957.442311	1.390153	1.124343	30.145255	20.785185	182.595606	462.664785	163.520016
min	1.000000	20.000000	21.00000	1300.000000	1.000000	1.000000	1875.000000	1950.000000	0.000000	0.000000	0.000000
25%	360.500000	20.000000	60.00000	7621.500000	5.000000	5.000000	1954.000000	1966.000000	0.000000	0.000000	0.000000
50%	714.500000	50.000000	70.00000	9522.500000	6.000000	5.000000	1972.000000	1993.000000	0.000000	385.500000	0.000000
75%	1079.500000	70.000000	80.00000	11515.500000	7.000000	6.000000	2000.000000	2004.000000	160.000000	714.500000	0.000000
max	1460.000000	190.000000	313.00000	164660.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	1474.000000

Observations of Statistical Analysis:

- There are some features have missing value so we have to treat them accordingly.
- We can see the the first feature 'ID' is nomial data type i.e. it is only a unique id number and it is only for name-sake. So we will drop them going ahead.
- 'MS SubClass','LotFrontage','LotArea' features maximum value didn't match with minimum value. The differece is very high in both of them, seems there is some outliers present in our dataset so we will take care of it later.

• Data Sources and their formats

AUS-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.

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?

- PoolQC : data description says NA means “No Pool”.
- MiscFeature: data description says NA means “no misc feature”.
- Alley: data description says NA means “no alley access”.
- Fence : data description says NA means “no fence”.
- FireplaceQu: data description says NA means “no fireplace”.

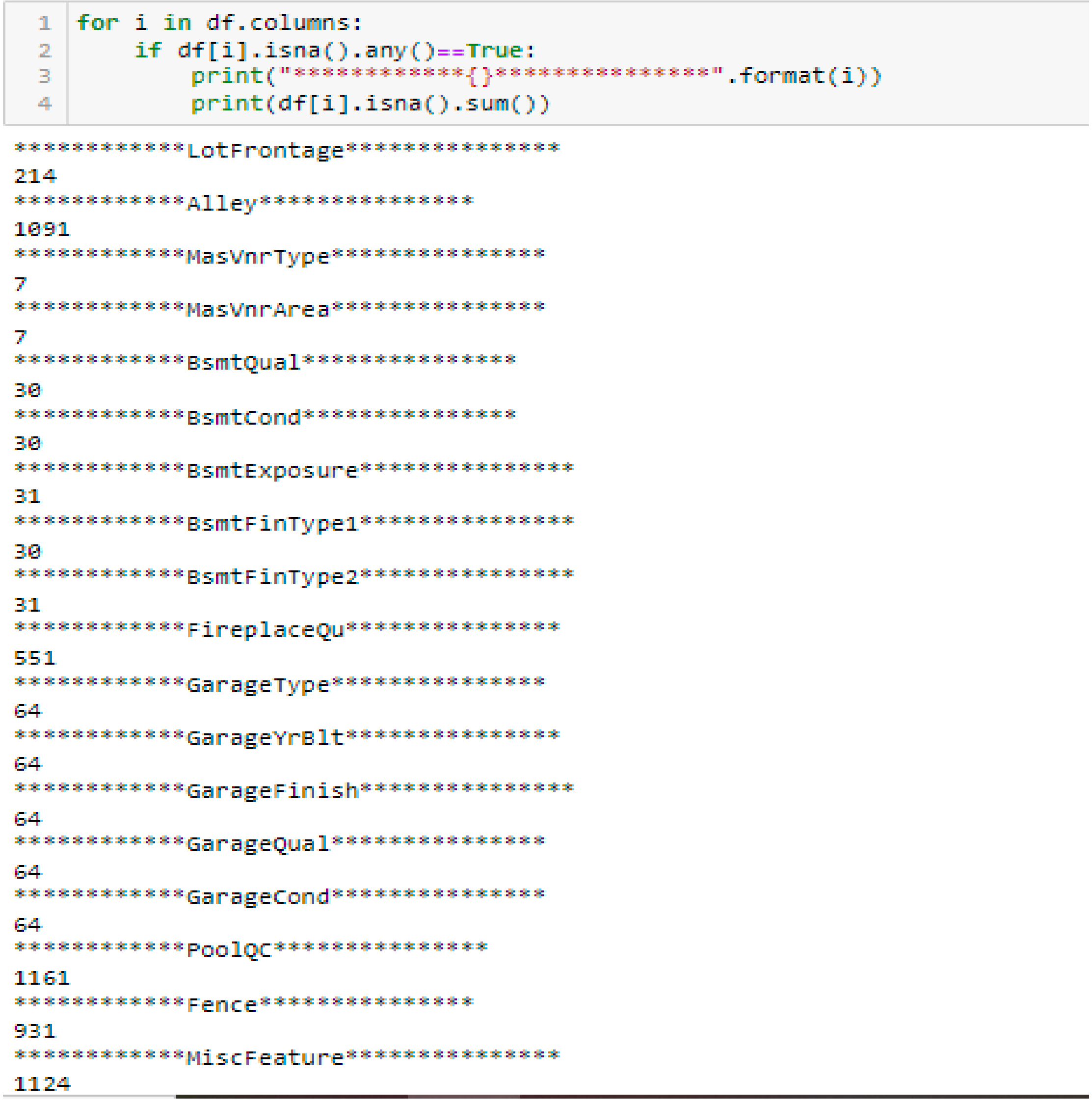
All Data Description Link:-
https://github.com/rakesh93619/Aus_HousingPrice_Pridiction/blob/main/Data_Description%20OR%20Features_Details.txt

• Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

Checking Missing values:-

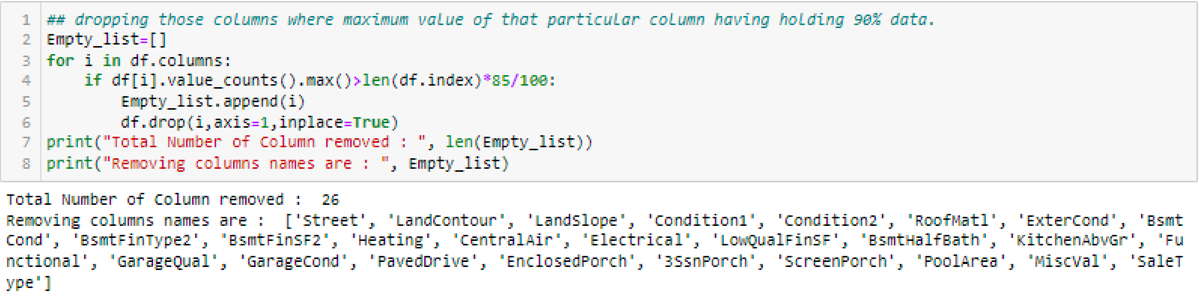
W



```
for i in df.columns:
    if df[i].isna().any()==True:
        print("************{}***************".format(i))
        print(df[i].isna().sum())
```

```
************LotFrontage***************
214
************Alley***************
1091
************MasVnrType***************
7
************MasVnrArea***************
7
************BsmtQual***************
30
************BsmtCond***************
30
************BsmtExposure***************
31
************BsmtFinType1***************
30
************BsmtFinType2***************
31
************FireplaceQu***************
551
************GarageType***************
64
************GarageYrBlt***************
64
************GarageFinish***************
64
************GarageQual***************
64
************GarageCond***************
64
************PoolQC***************
1161
************Fence***************
931
************MiscFeature***************
1124
```

- ## We will remove those features who have 90 or more than 90% of unique value since they are not adding any value so we will drop them.



```
## dropping those columns where maximum value of that particular column having holding 90% data.
Empty_list=[]
for i in df.columns:
    if df[i].value_counts().max()>len(df.index)*85/100:
        Empty_list.append(i)
        df.drop(i,axis=1,inplace=True)
print("Total Number of Column removed : ", len(Empty_list))
print("Removing columns names are : ", Empty_list)
```

```
Total Number of Column removed :  26
Removing columns names are :  ['Street', 'LandContour', 'LandSlope', 'Condition1', 'Condition2', 'RoofMatl', 'ExterCond', 'Bsmt
Cond', 'BsmtFinType2', 'BsmtFinSF2', 'Heating', 'CentralAir', 'Electrical', 'LowQualFinSF', 'BsmtHalfBath', 'KitchenAbvGr', 'Fu
nctional', 'GarageQual', 'GarageCond', 'PavedDrive', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'SaleT
ype']
```

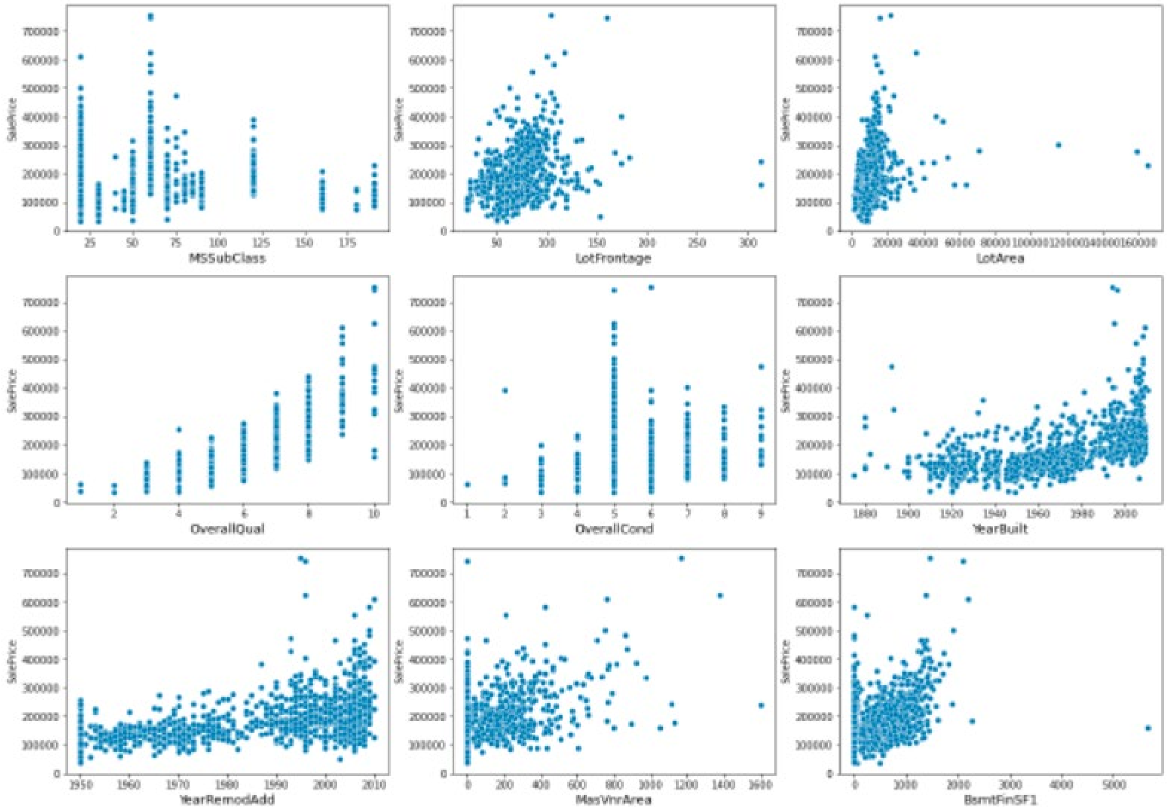
There are 26 columns having 90% data with one unique value.

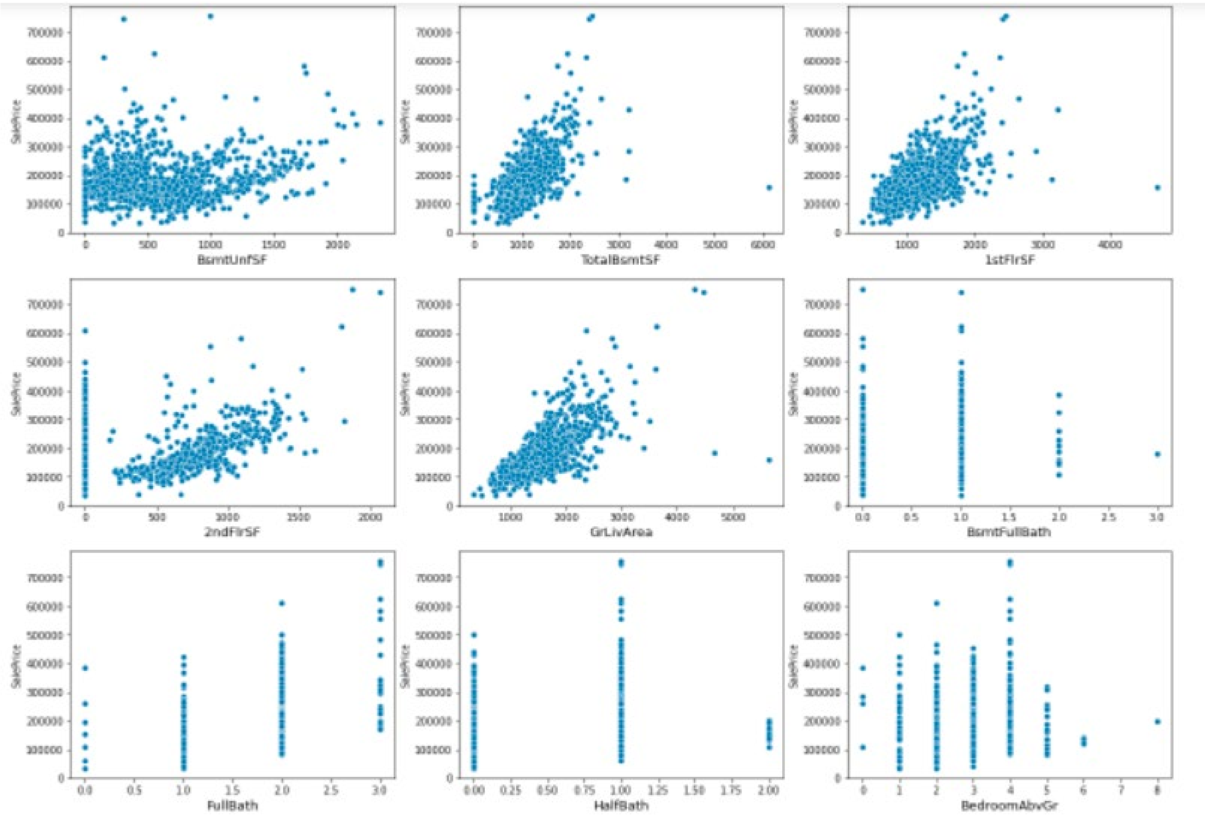
Filling NaN

- GarageType, GarageFinish, Replacing missing data with “None”. Which is most accuring value of them.
- lot frontage: Since the area of each street connected to the house property most likely have a similar area to other houses in its neighborhood, we can fill in missing values by the median LotFrontage of the neighborhood.
- MasVnrArea and MasVnrType: NA most likely means no masonry veneer for these houses. We can fill 0 for the area and None for the type. So we will fill Nan for them with Mean.
- BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, and BsmtFinType2: For all these categorical basement-related features, NaN means that there isn’t a basement. So we can simply fill Nan with Mode.

- Data Inputs- Logic- Output Relationships

## Numeric feature relation with Target:-

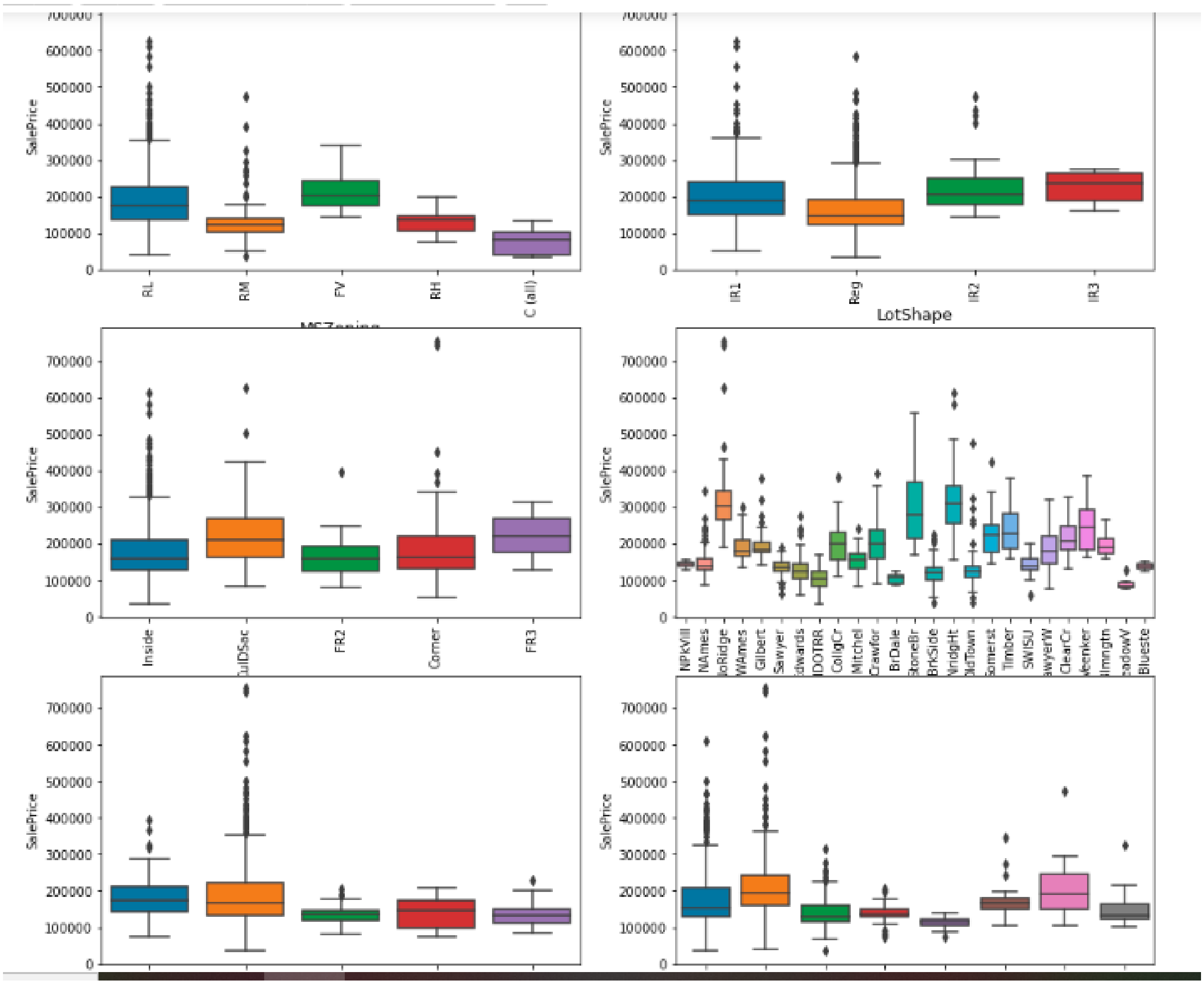


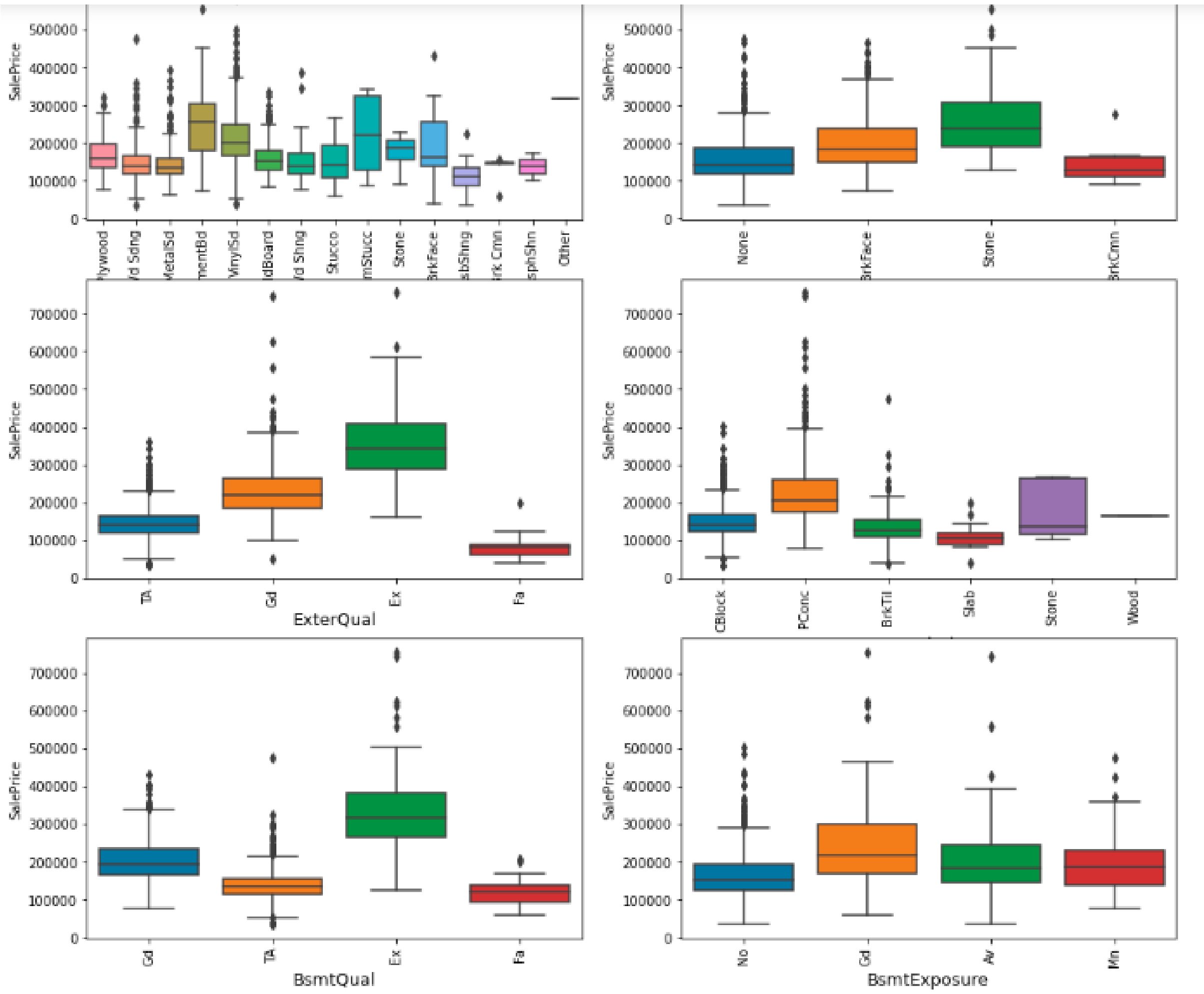


## observation:-

- BsmtUnfSF,TotalBsmtSF,lstFirSF,GrLivArea,GarageArea , these features are showing high correlation with the target variable. We can see in the graph, as these quantity increases the Saleprice is also increase.
- BSmtFinSF1,BsmtUnfSF,TotalBsmtSF,2ndFlrSF, BsmtFullBath,Fireplaces and OpenPorchSF, we can see that At '0' these feature are showing strange relationship with target variable but as quantity increasing the saleprice is also increasing. We assume that there are some lower outlier in these features at point 0. We will see that later.

## Categorical Feature relation with Target:-


SalePrice
700000
600000
500000
400000
300000
200000
100000
0
RL
RM
FV
RH
C (all)
MSZoning
IR1
Reg
IR2
IR3
LotShape
Inside
CulDSac
FR2
Corner
FR3
NPkVill
NAmes
loRidge
NWAmes
Gilbert
Sawyer
Edwards
IDOTRR
CollgCr
Mitchel
Crawfor
BrDale
StoneBr
BrkSide
NridgHt
OldTown
Somerst
Timber
SWISU
awyerW
ClearCr
/eenker
Blmngtn
eadowV
Blueste

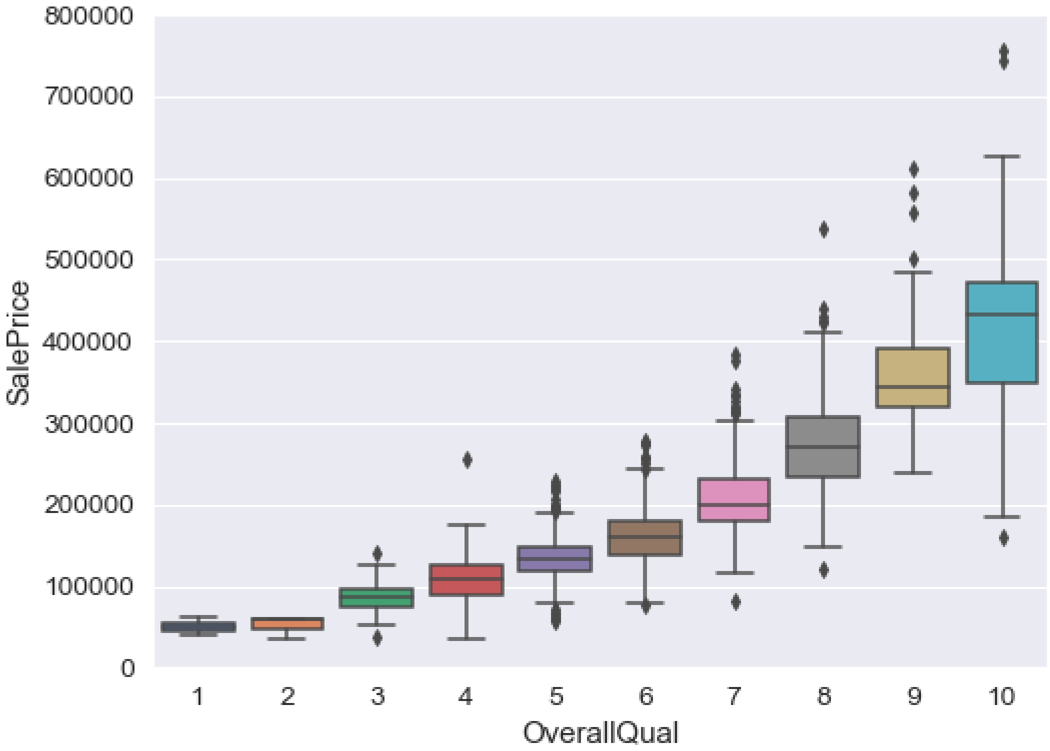


### Observations:-

- High number of SalePrice for MSSubClass 80 and then 120.
- MSZoning Floating Village Residential has highest number of SalePrice.
- 2Story and 1Story buildings have High Number of SalePrice
- Positive correlation of Overallquality and SalePrice.
- Neighborhood_Nridght has high value of SalePrice.
- Building type TwinSE and IFam have high value of SalePrice.
- ExternalQuality has high value in SalePrice.
- BasementQuality has high value in SalePrice.
- House with execellent Kitchenquality and HeatingQC have HighestSaleprice

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

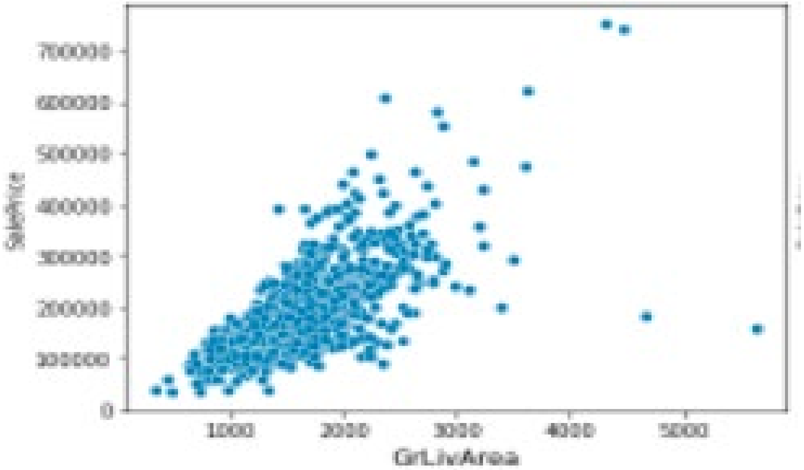
### Overall Quality Vs Sale Price



### Observations:-

- The plot defines we can state that as OverallQual increases, the SalePrice also increases.
- The above assumption is true on normal assumption also, As OverallQual is more the SalePrice is also more.

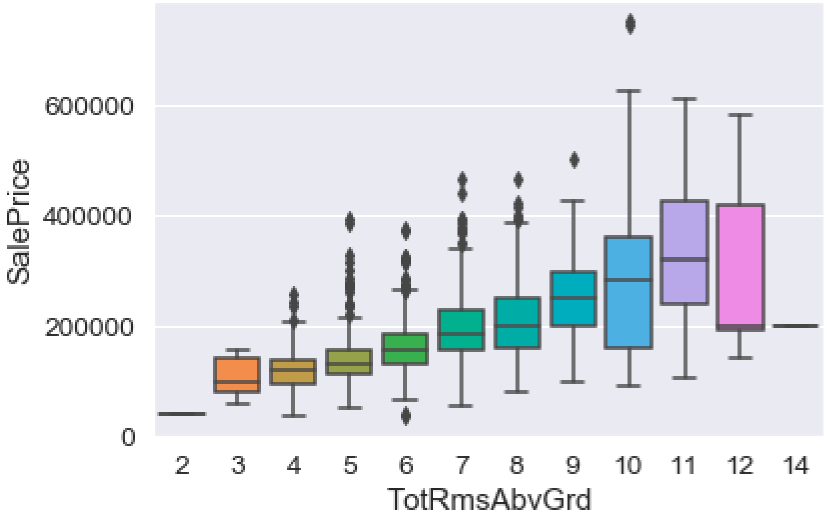
### GrLivArea Vs SalePrice:



### observations:-

- People pay more for more living area.
- In the above plot, there is a value which has the least cost for more living area. It is better to remove it.

##TotRmsAbvGrd Vs SalePrice



###Observation

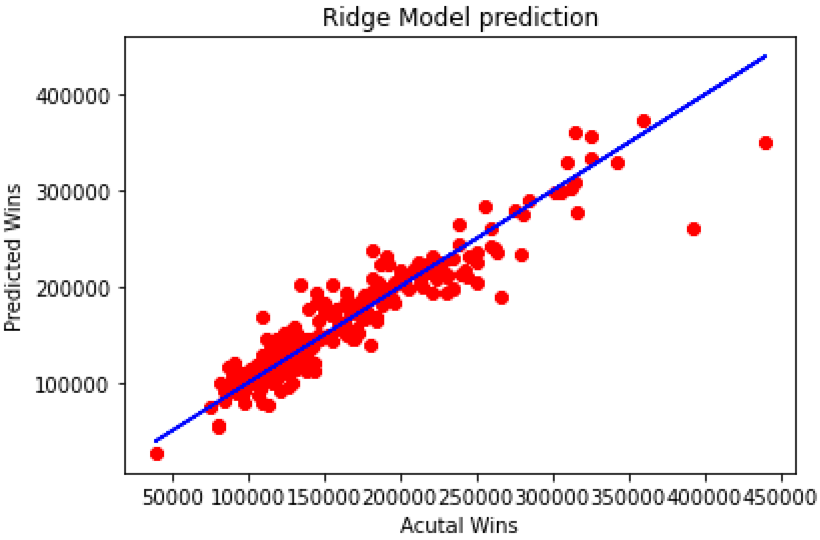
- From the above plot, we can say that for TotRmsAbvGrd having more than 11 rows has less weight.
- May be those are old enough due to which they cost less, But it is just an assumption.

- Hardware and Software Requirements and Tools Used
- **Pandas & Numpy** :-For Numerical Analysis and Importing the data .
- **Matplotlib & Seaborn**:- Both are used for Visualization.
- **MinMax Scaler**::- For scaling all the data.
- **Train Test Split:-** For spliting our dataset for traing and testing.
- **LinearRegression,KneighborsRegressor,DecisionTreeRegressor,RandomForestRegressor,AdaBoostRegressor,GradientBoostingRegressor,SVM,XGBRegressor :-** These model are used for predict the target.
- **R_Squared,MeanSquaredError,MeanAbsoluteError :-** These are for Evaluations.
- **GridSearchCV:-** Its for finding the best parameter for our model.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  ## RidgeRegression:-

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

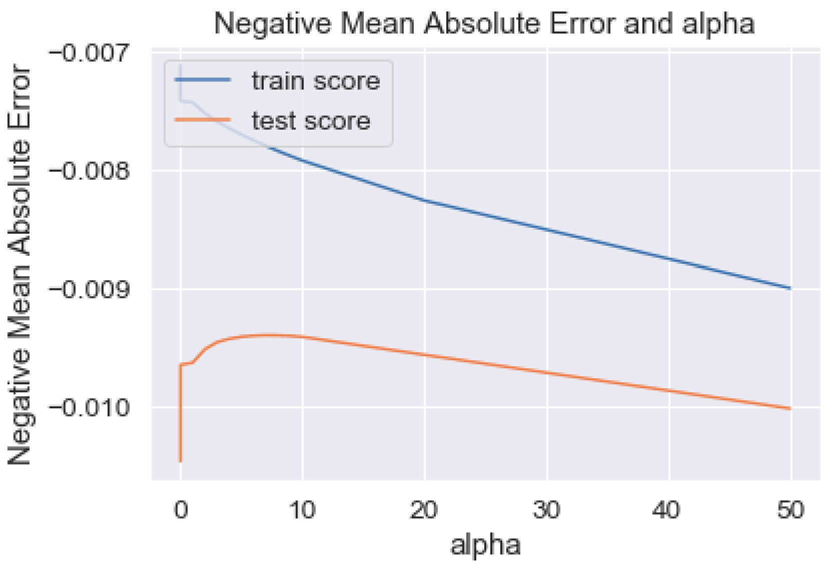


## Points:-

- The above plot defines the way to decide the Best fit line for model.
- The point in which train and test score has less gap between them is the value which we take as an optimum value of alpha.
- The R2 value for optimum alpha value. 87%

## Lasso Regression:-

Lasso regression analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that cause regression coefficients for some variables to shrink toward zero. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model. Variables with non-zero regression coefficients variables are most strongly associated with the response variable. Explanatory variables can be either quantitative, categorical, or both.



Observation

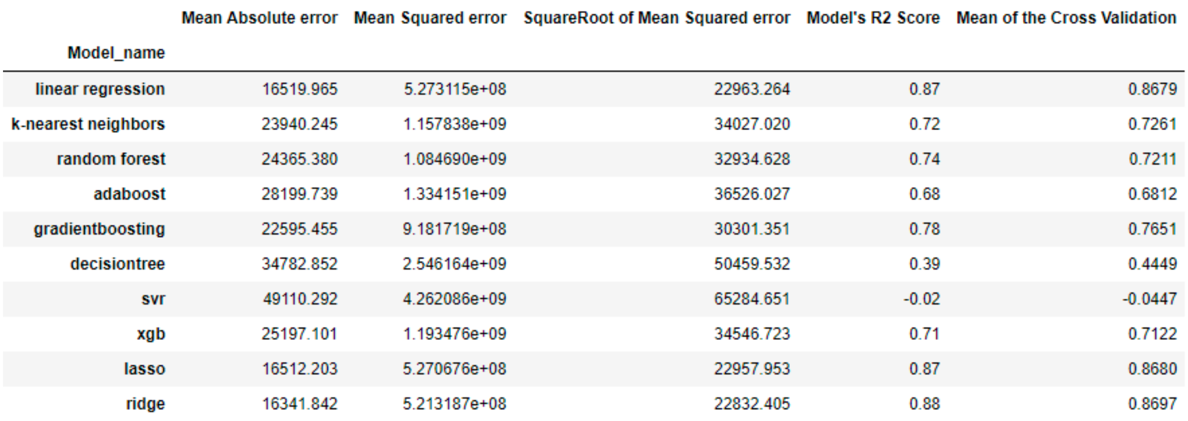
- The above plot defines the way to decide the optimum value of alpha.
- The point in which train and test score has less gap between them is the value which we take as an optimum value of alpha
- From the above plot, we came to know that the value with alpha = 0.0001 has a minimum gap between the test and the training score.
- The R2 value for optimum alpha value: 88%

- Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

- **LinearRegression()**
- **KNeighborsRegressor(),**
- **RandomForestRegressor(),**
- **AdaBoostRegressor(),**
- **GradientBoostingRegressor(),**
- **DecisionTreeRegressor(),**
- **SVR(),**
- **XGBRegressor(),**
- **Lasso(),**
- **Ridge()**

- Run and Evaluate selected models

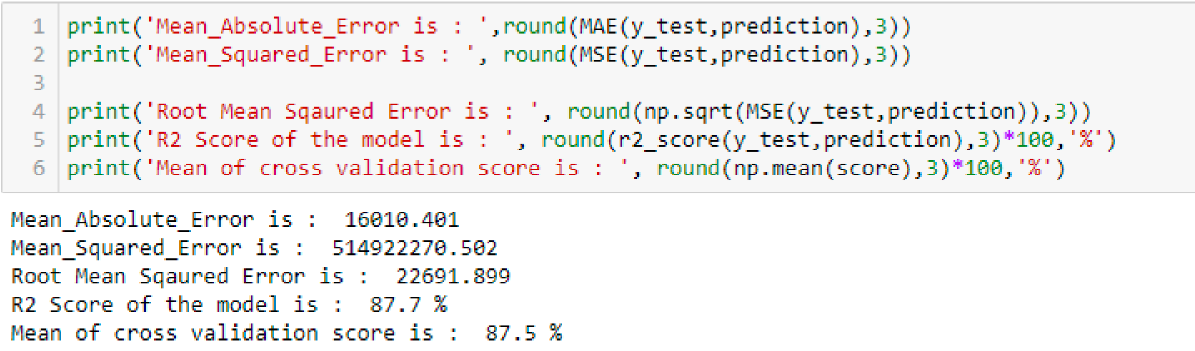


Model_name	Mean Absolute error	Mean Squared error	SquareRoot of Mean Squared error	Model's R2 Score	Mean of the Cross Validation
linear regression	16519.965	5.273115e+08	22963.264	0.87	0.8679
k-nearest neighbors	23940.245	1.157838e+09	34027.020	0.72	0.7261
random forest	24365.380	1.084690e+09	32934.628	0.74	0.7211
adaboost	28199.739	1.334151e+09	36526.027	0.68	0.6812
gradientboosting	22595.455	9.181719e+08	30301.351	0.78	0.7651
decisiontree	34782.852	2.546164e+09	50459.532	0.39	0.4449
svr	49110.292	4.262086e+09	65284.651	-0.02	-0.0447
xgb	25197.101	1.193476e+09	34546.723	0.71	0.7122
lasso	16512.203	5.270676e+08	22957.953	0.87	0.8680
ridge	16341.842	5.213187e+08	22832.405	0.88	0.8697

Lasso and Ridge give us the Best accuracy.

- Key Metrics for success in solving problem under consideration

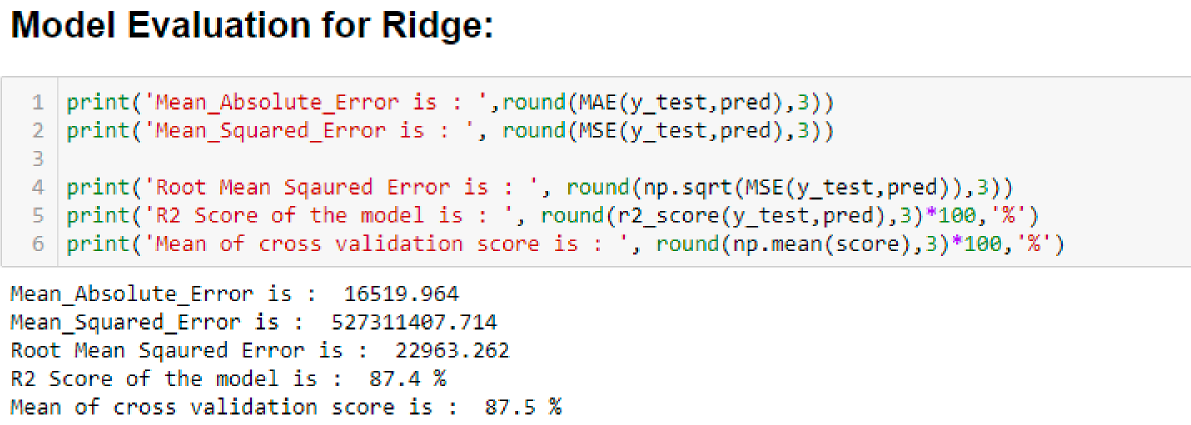
### Model Evaluation of lasso Model:-



```
print('Mean_Absolute_Error is : ',round(MAE(y_test,prediction),3))
print('Mean_Squared_Error is : ', round(MSE(y_test,prediction),3))

print('Root Mean Sqaured Error is : ', round(np.sqrt(MSE(y_test,prediction)),3))
print('R2 Score of the model is : ', round(r2_score(y_test,prediction),3)*100,'%')
print('Mean of cross validation score is : ', round(np.mean(score),3)*100,'%')
```

```
Mean_Absolute_Error is :  16010.401
Mean_Squared_Error is :  514922270.502
Root Mean Sqaured Error is :  22691.899
R2 Score of the model is :  87.7 %
Mean of cross validation score is :  87.5 %
```

### Model Evaluation of Ridge Model:-

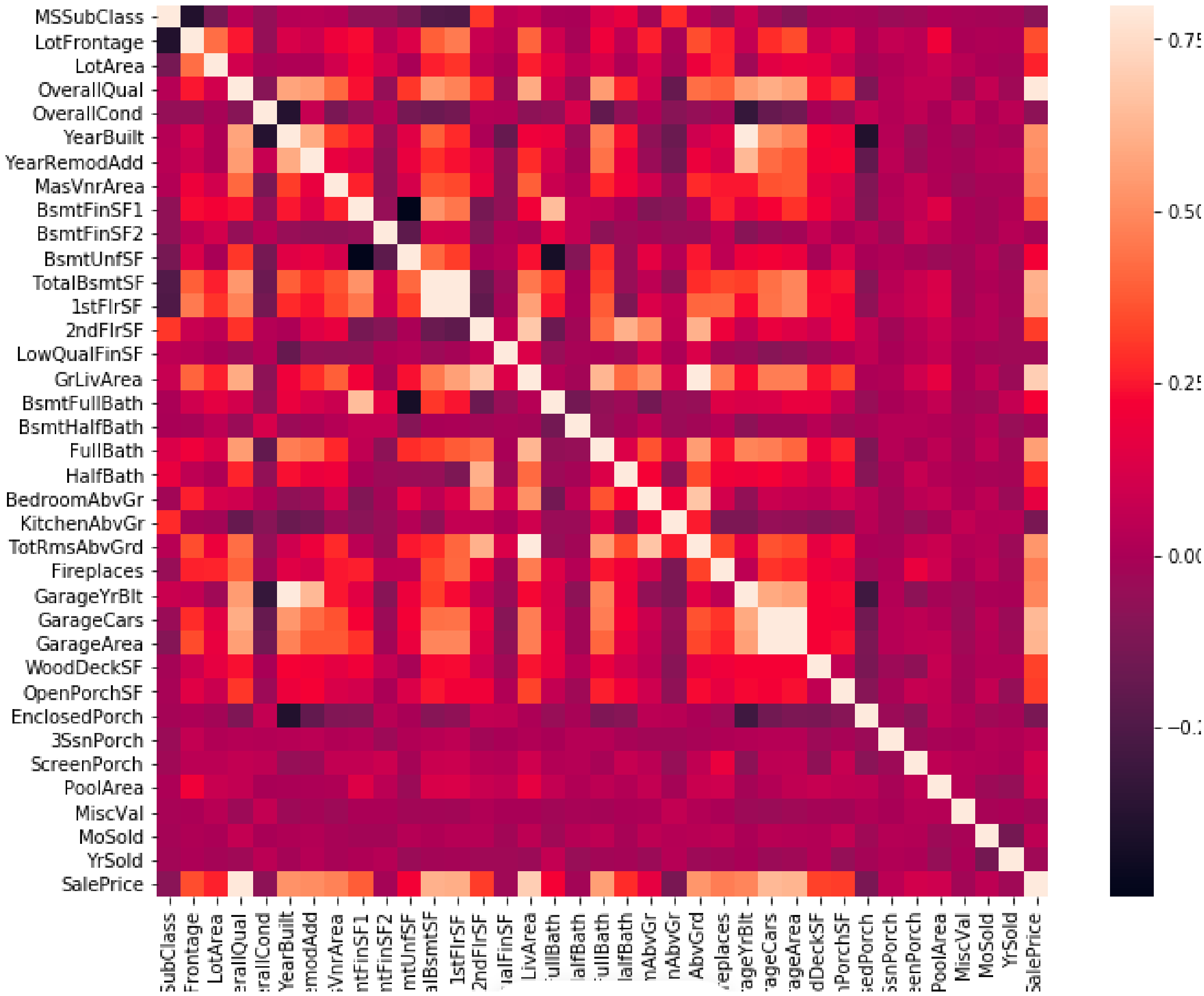


Model Evaluation for Ridge:

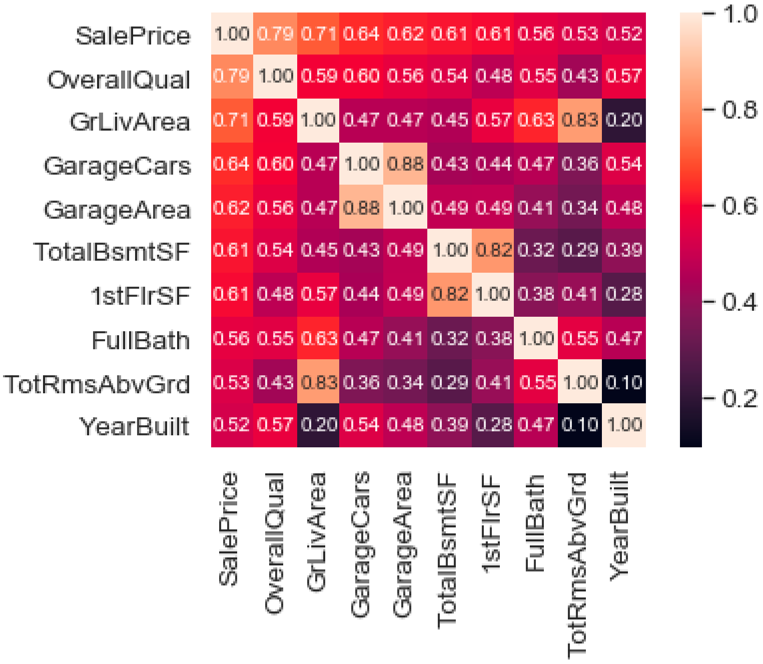
```
print('Mean_Absolute_Error is : ',round(MAE(y_test,pred),3))
print('Mean_Squared_Error is : ', round(MSE(y_test,pred),3))

print('Root Mean Sqaured Error is : ', round(np.sqrt(MSE(y_test,pred)),3))
print('R2 Score of the model is : ', round(r2_score(y_test,pred),3)*100,'%')
print('Mean of cross validation score is : ', round(np.mean(score),3)*100,'%')
```

```
Mean_Absolute_Error is :  16519.964
Mean_Squared_Error is :  527311407.714
Root Mean Sqaured Error is :  22963.262
R2 Score of the model is :  87.4 %
Mean of cross validation score is :  87.5 %
```

- Visualizations



### Top 10 Correlation Features:-

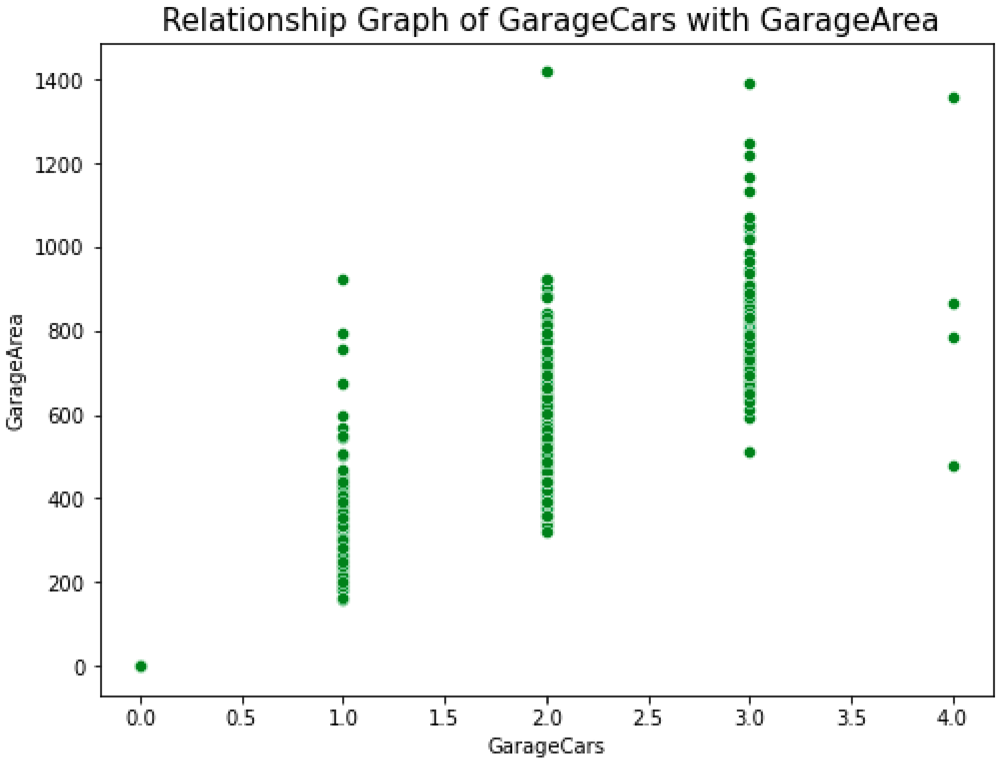


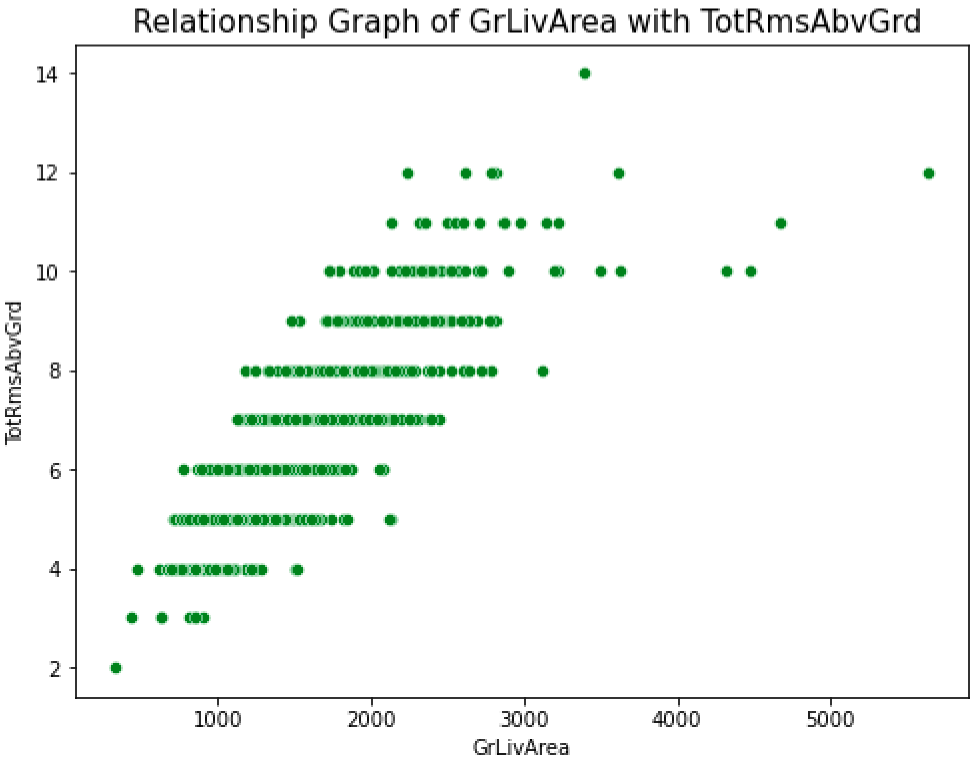
	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1.00	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20
GarageCars	0.64	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10
YearBuilt	0.52	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00

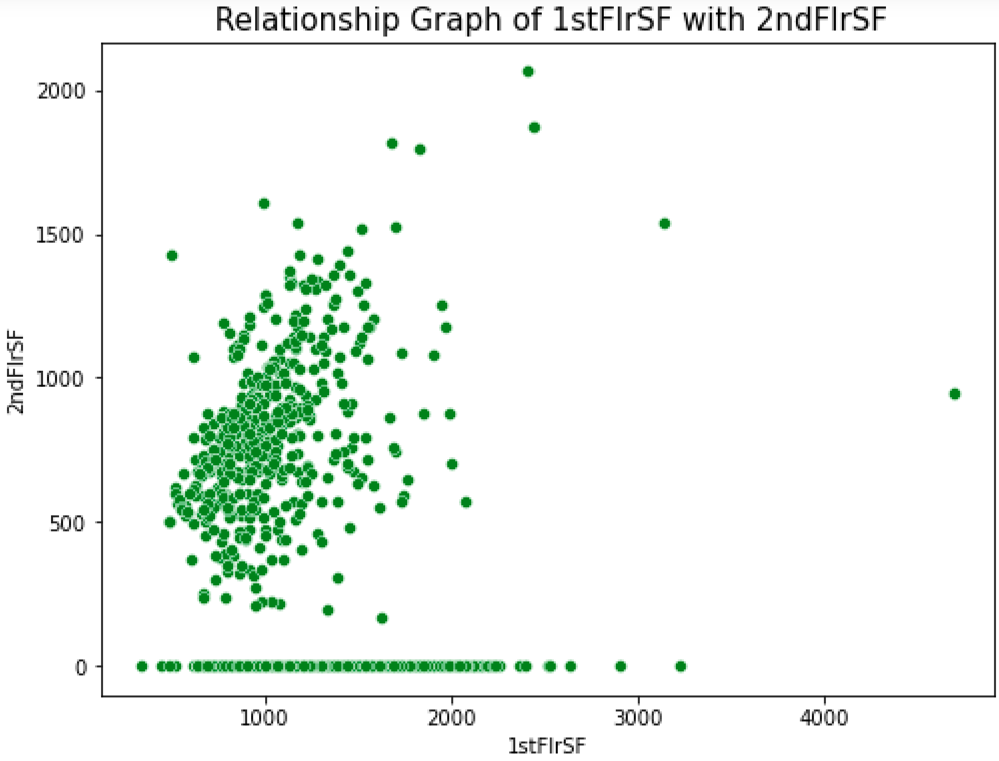
### observation of Heatmap:-

- GarageCars and GarageArea is highly correlated with each other. Typically we can say that our data set have multicollinearity problem. And also both are equally correlate with target variable.
- GLivArea and TotRmsAbvGrd features are also highly correlate with each other with 80% correlation.
- 1stFlrSF and 2ndFlrSF features are also correalte with each other with 77% correlation.

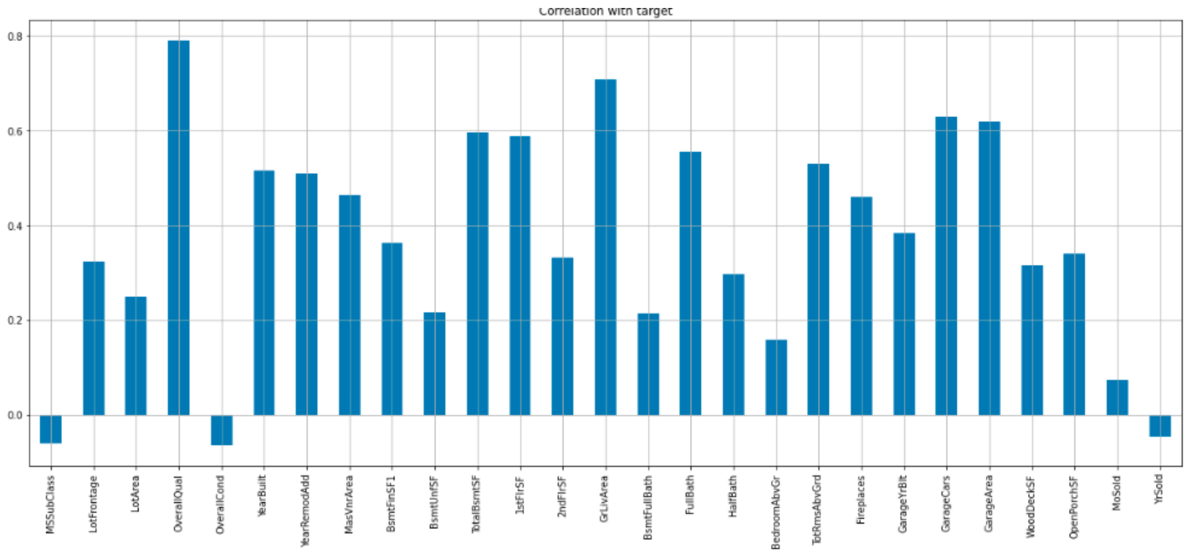
### Multicollinearity problem:-







Correction With Target:-



Interpretation of the Results

Multicollinearity Conclusion:-

- As we assume, all of them showing tight relation with eachother, so we have to drop any 1 feature from every pair.
- GarageCars and GarageArea both are equally correlate with target variable. So we can drop any of them.
- GrLivArea and TotRmsAbvGrd are correalted with target variable as 69% and 49% respectively. So we will drop TotRmsAbvGrd feature as it is less correlate with target with compare to GrLiveArea.
- 1stFlrSF and 2ndFlrSF are correalted with target variable as 55% and 30% respectively. So we will drop 2ndFlrSF feature as it is less correlate with target with compare to 1stFlrSF.

Observations:-

- MSSubClass, OverallCond, YrSold feature are negative correlated with target and in other hand all the features are positive correlated with the target variable.
- Overallcond feature is showing highest correlation with target variable.
- MoSold feature is showing least correlation with target variable.

CONCLUSION

Key Findings and Conclusions of the Study:-

- Our main results, which we do not repeat, were stated in the introduction. Several limitations of our model and directions for further research are worth noting.

- Some of our equations are better than others. The stability and precision of our estimates give us confidence in the responsiveness of construction to interest rates and prices and in the response of rents to vacancies. We are much less certain about the response of vacancies to completions and household formation. More broadly, we know more about the response of quantities to prices than we do about the response of prices to quantities.

• Learning Outcomes of the Study in respect of Data Science

To keep our analysis relatively transparent and manageable, we have ignored some minor variables (like taxes, lending restrictions, the approvals pipeline and first home owner grants). And we have assumed that many relevant variables are exogenous. A more complicated analysis would relax those assumptions. A priority is to allow household size to depend on income and rents. This would then affect household formation and housing quality. It is not clear that adding taxes to the user cost would substantially change the behaviour of the model, but it would be useful for policy discussions.

• Limitations of this work and Scope for Future Work

We are unaware of any aggregate time series data on the cost of new houses that includes land costs. The series we used for these exercises were constructed using ABS data on the cost of building new houses and data from a number of sources on land prices.

Actual household size is highly endogenous, adjusting so as to keep demand and supply in approximate balance, so lacks predictive power.

Technically, we show a reduction in real rents, whereas the text above refers to changes in nominal rents. Empirically, changes in the price level are quite small relative to the changes shown in Figure 12, so the distinction can be ignored.