FLIP ROBO

Project Name: -

“HOUSING: PRICE PREDICTION”

Submitted by:

Vikas Hanmant Bandgar

ACKNOWLEDGMENT

I would like to express my special gratitude to the “Flip Robo” team, who has given me thisopportunity to deal with a beautiful dataset and it has helped me to improve my analysing skills. Also, I want to express my huge gratitude to Mrs . Sapna Verma (SME FlipRobo), who has helped me overcome difficulties within this project and others. and also encouraged me a lot with his valuable words and with his unconditional support I have ended up with a beautiful Project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life.

And also thank you for many other persons who has helped me directly or indirectly to complete the project.

INTRODUCTION

- Business Problem Framing

  Describe the business problem and how this problem can be related to the real world.

- Conceptual Background of the Domain Problem

  Describe the domain related concepts that you think will be useful for better understanding of the project.

- Review of Literature

  This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

- Motivation for the Problem Undertaken

  Describe your objective behind to make this project, this domain and what is the motivation behind.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- Data Sources and their formats

  What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

- 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?

- Data Inputs- Logic- Output Relationships

  Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

  Here, you can describe any presumptions taken by you.

- Hardware and Software Requirements and Tools Used

  Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving of this problem.

- Testing of Identified Approaches (Algorithms)

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

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

- Key Metrics for success in solving problem under consideration

  What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

- Visualizations

  Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

  If different platforms were used, mention that as well.

- Interpretation of the Results

  Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

CONCLUSION

- Key Findings and Conclusions of the Study

  Describe the key findings, inferences, observations from the whole problem.

- Learning Outcomes of the Study in respect of Data Science

  List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

- Limitations of this work and Scope for Future Work

  What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

INTRODUCTION:- ¬

Business Problem Framing:- 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 largemarket 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. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality.

Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency. The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analysing previous market trends and price ranges, and also upcoming developments future prices will be predicted. cost of property depending on number of attributes considered. Now as a data scientist our work is to analyse the dataset and apply.

Conceptual Background of the Domain Problem:-

The real estate market is one of the most competitive in terms of pricing andsame tends to vary significantly based on numerous factors; forecasting property price is an important module in decision making for both the buyersand investors in supporting budget allocation, finding property finding stratagems and determining suitable policies.

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?

Why is house price prediction important?

House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency.

There are three factors that influence the price of a house which include physical conditions, concept and location. Hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Therefore, in this project report we present various important features to use while predicting housing prices with good accuracy. While using features in a regression model some feature engineering is required for better prediction.

Review of Literature: -

The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data is to be considered. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector,the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned.

The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. Therefore, in this project report, we present various important features to use while predicting housing prices with good accuracy.

We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. The primary aim of this report is to use these Machine Learning Techniques andcurate them into ML models which can then serve the users. The main objective of a Buyer is to search for their dream house which has all the amenities they need. Furthermore, they look for these houses/Real estates with a price in mind and there is no guarantee that they will get the product for a deserving price and not overpriced. Similarly, A seller looks for a certain number that they can put on the estate as a price tag and this cannot be just a wild guess, lots of research needs to be put to conclude a valuation of a house.

Motivation for the Problem Undertaken: -

I have to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market. The relationship between house prices and the economy is an important motivating factor for predicting house prices.

2.Analytical Problem Framing: -

Mathematical/ Analytical Modeling of the Problem :-

This perticular problem has two datasets one is train dataset and the other is test dataset. I have built model using train dataset and predicted SalePrice for test dataset. By looking into the target column, I came to know that the entries of SalePrice column were continuous and this was a Regression problem so I have to use all regression algorithms while building the model. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 80% null values and more than 85% zero values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. While checking the null values in the datasets I found many columns with nan values and I replaced those nan values with suitable entries like mean for numerical columns and mode for categorical columns. To get better insight on the features I have used ploting like distribution plot, bar plot, reg plot and strip plot. With these ploting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using percentile method and I removed skewness using yeo-johnson method. I have used all the regression models while building model then tunned the best model and saved the best model. At last I have predicted the sale price fot test dataset using the saved model of train dataset.

Data Sources and their formats:-

The data was given by my internship company – Flip Robo technologies in csv (comma separated values) format. Here I was having two datasets one is train and other is test. I have built model using train dataset and predicted Sale Price for test dataset. My train dataset was having 1168 rows and 81 columns including target, and my test dataset was having 292 rows and 80 columns excluding target. In this particular datasets I have object, float and integer types of data. I can merge these two datasets and perform my analysis, but I have not done that because of data leakage issue. This is how my datasets looks for me when I import those datasets to my python.

Data Preprocessing Done: -

- As a first step I have imported required libraries and I have importedboth the datasets which were in csv format.

- Then I did all the statistical analysis like checking shape, nunique, valuecounts, info etc.....

- While checking the info of the datasets I found some columns with more than 80% null values, so these columns will create skewness in datasets.so I decided to drop those columns.

- Then while looking into the value counts I found some columns withmore than 85% zero values this also creates skewness in the model andthere are chances of getting model bias so I have dropped thosecolumns with more than 85% zero values.

- While checking for null values I found null values in most of the columnsand I have used imputation method to replace those null values (modefor categorical column and mean for numerical columns). • In Id and Utilities column the unique counts were 1168 and 1 respectively, which means all the entries in Id column are unique and ID is the identity number given for perticular asset and all the entries in Utilities column were same so these two column will not help us in model building. So I decided to drop those columns.

- Next as a part of feature extraction I converted all the year columns to there respective age. Thinking that age will help us more than year.

- And all these steps were performed to both train and test datasets separately and simultaneously.

Data Inputs- Logic- Output Relationships:-

- I have used box plot for each pair of categorical features that shows the relation with the median sale price for all the sub categories in each categorical feature.

- And also for continuous numerical variables I have used reg plot to show the relationship between continuous numerical variable and target variable.

- I found that there is a linear relationship between continuous numerical variable and SalePrice.

Hardware and Software Requirements and Tools Used:-

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software…

Hardware required: -

1. Processor — core i5 and above

2. RAM — 4 GB or above

3. SSD — 250GB or above

Software required: -

Anaconda

Libraries required :-

- **import pandas as pd:-** pandas is a popular Python-based data analysis toolkit which can be imported using import pandas as pd. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.

- **import numpy as np:-** NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

- **Import seaborn as sns:-** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

- **Import matplotlib.pyplot as plt:-** matplotlib.pyplot is a collection offunctions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

some other libraries which are used:-

from sklearn. preprocessing import OrdinalEncode

from sklearn.preprocessing import StandardScaler

from statsmodels.stats.outliers_influence import variance_inflation_factor

from sklearn.ensemble import RandomForestRegressor

from sklearn.tree import DecisionTreeRegressor

from xgboost import XGBRegressor

from sklearn.ensemble import GradientBoostingRegressor

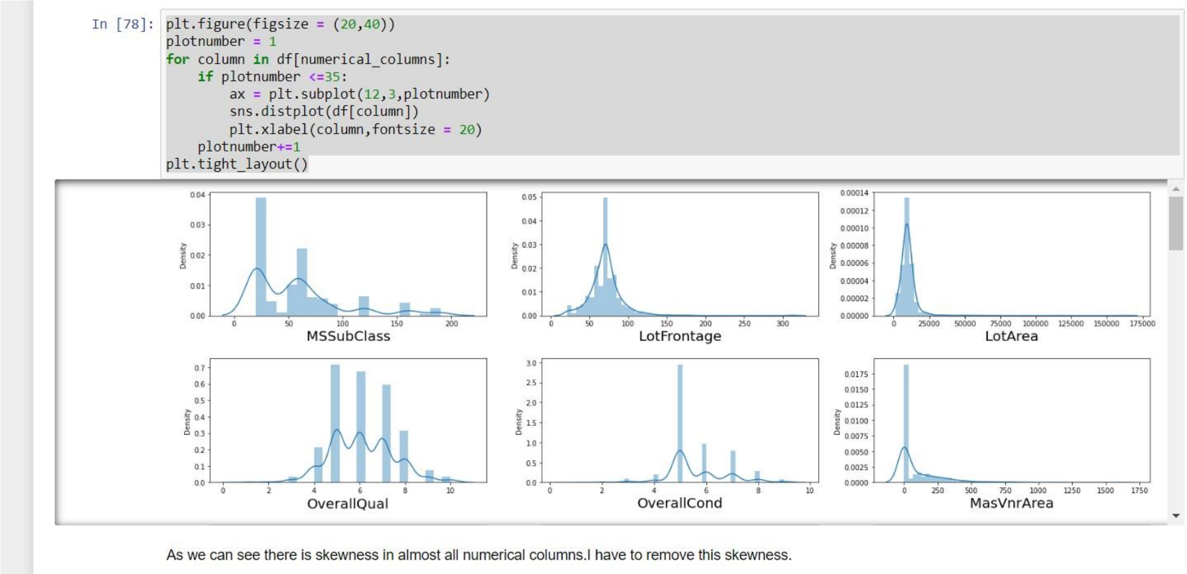
from sklearn.ensemble import ExtraTreesRegressor

from sklearn.metrics import classification_report from sklearn.model_selection import cross_val_score

3.Data Analysis and Visualization: -

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

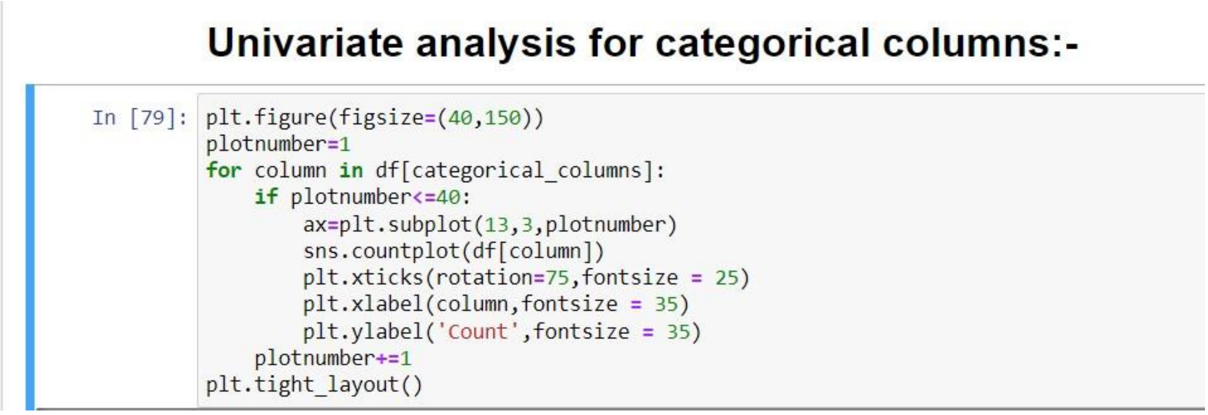
Here I have used imputation method to replace null values.



```
In [78]: plt.figure(figsize = (20,40))
         plotnumber = 1
         for column in df[numerical_columns]:
             if plotnumber <=35:
                 ax = plt.subplot(12,3,plotnumber)
                 sns.distplot(df[column])
                 plt.xlabel(column,fontsize = 20)
             plotnumber+=1
         plt.tight_layout()
```

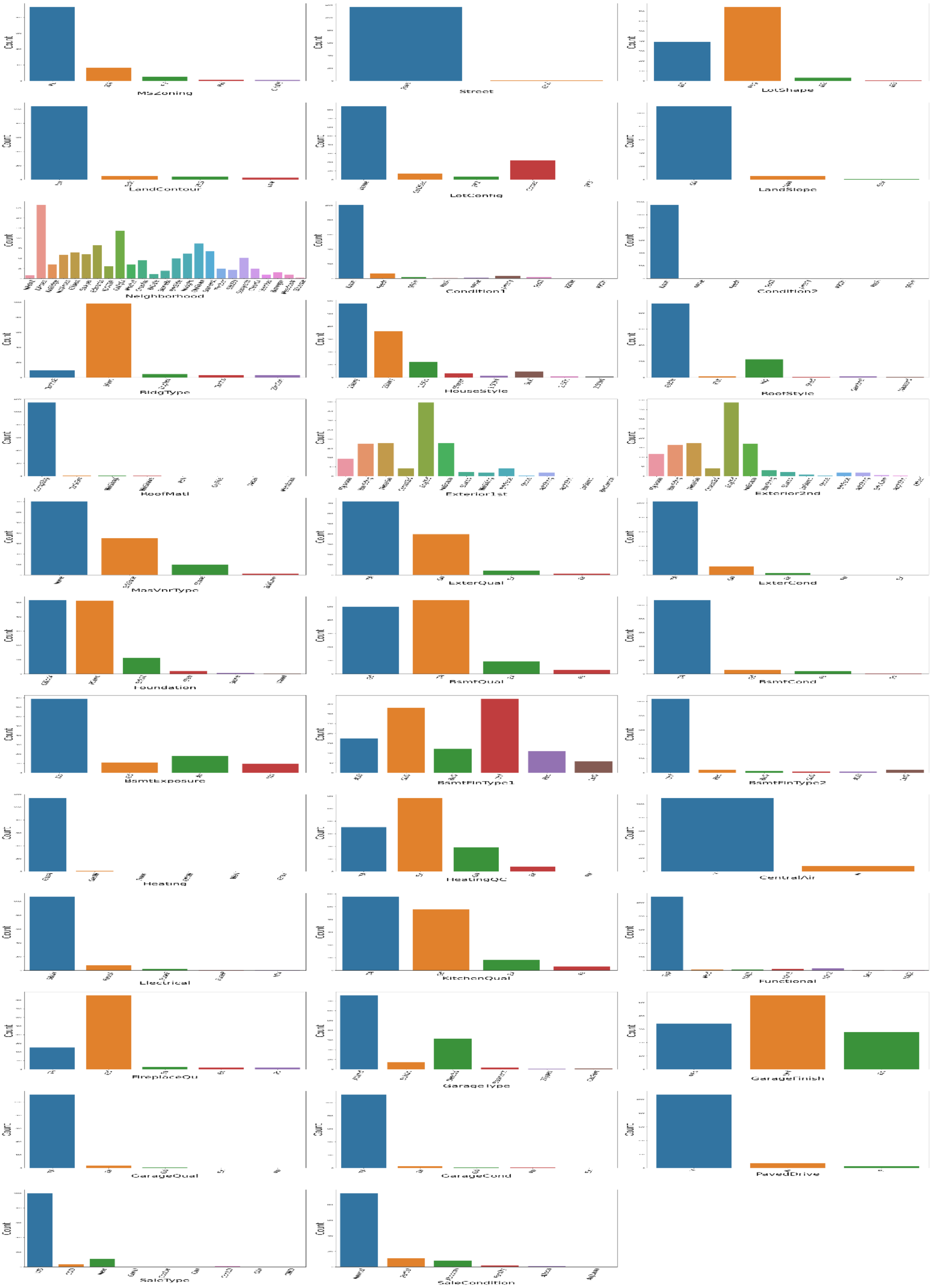
As we can see there is skewness in almost all numerical columns.I have to remove this skewness.

For remove outliers I have used percentile method. And to remove skewness I have used yeo-johnson method.

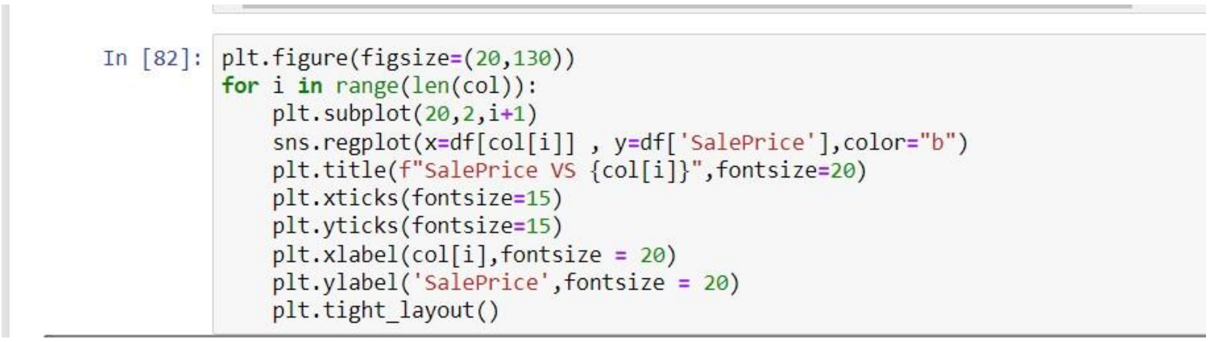


Univariate analysis for categorical columns:-

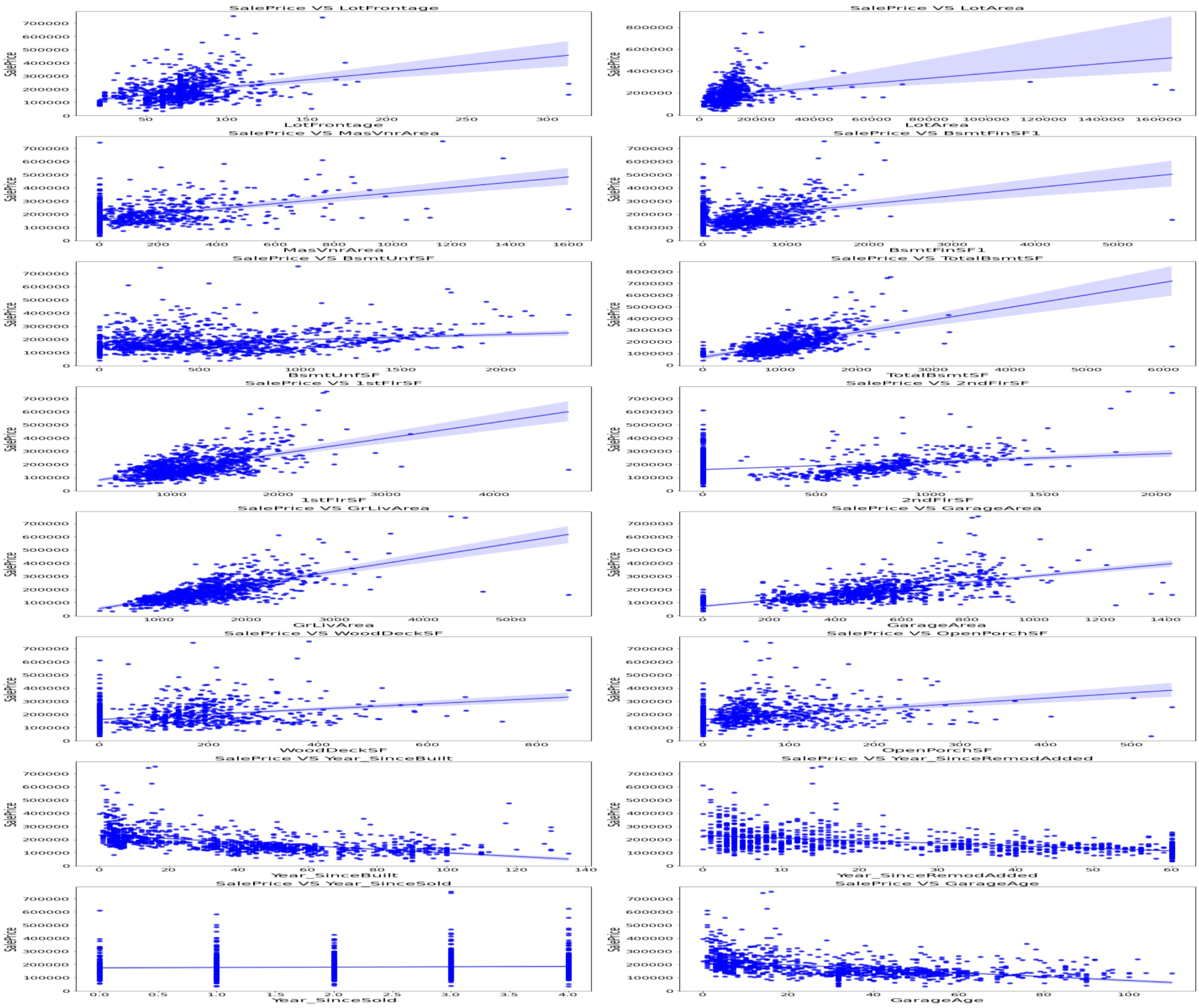
```
In [79]: plt.figure(figsize=(40,150))
         plotnumber=1
         for column in df[categorical_columns]:
             if plotnumber<=40:
                 ax=plt.subplot(13,3,plotnumber)
                 sns.countplot(df[column])
                 plt.xticks(rotation=75,fontsize = 25)
                 plt.xlabel(column,fontsize = 35)
                 plt.ylabel('Count',fontsize = 35)
             plotnumber+=1
         plt.tight_layout()
```


Count
MSZoning
Street
LotShape
LandContour
LotConfig
LandSlope
Neighborhood
Condition1
Condition2
BldgType
HouseStyle
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
Heating
HeatingQC
CentralAir
Electrical
KitchenQual
Functional
FireplaceQu
GarageType
GarageFinish
GarageQual
GarageCond
PavedDrive
SaleType
SaleCondition

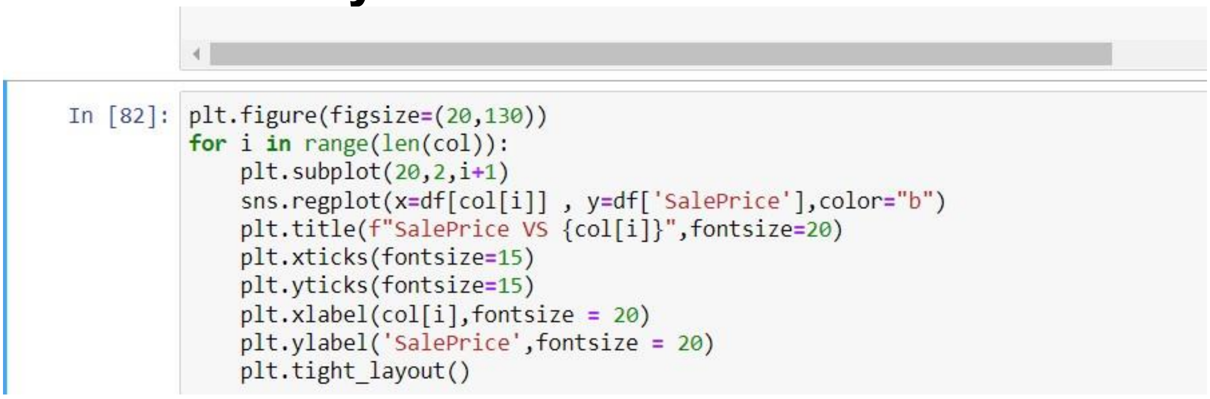
for numeric column:-



```
In [82]: plt.figure(figsize=(20,130))
         for i in range(len(col)):
             plt.subplot(20,2,i+1)
             sns.regplot(x=df[col[i]] , y=df['SalePrice'],color="b")
             plt.title(f"SalePrice VS {col[i]}",fontsize=20)
             plt.xticks(fontsize=15)
             plt.yticks(fontsize=15)
             plt.xlabel(col[i],fontsize = 20)
             plt.ylabel('SalePrice',fontsize = 20)
             plt.tight_layout()
```

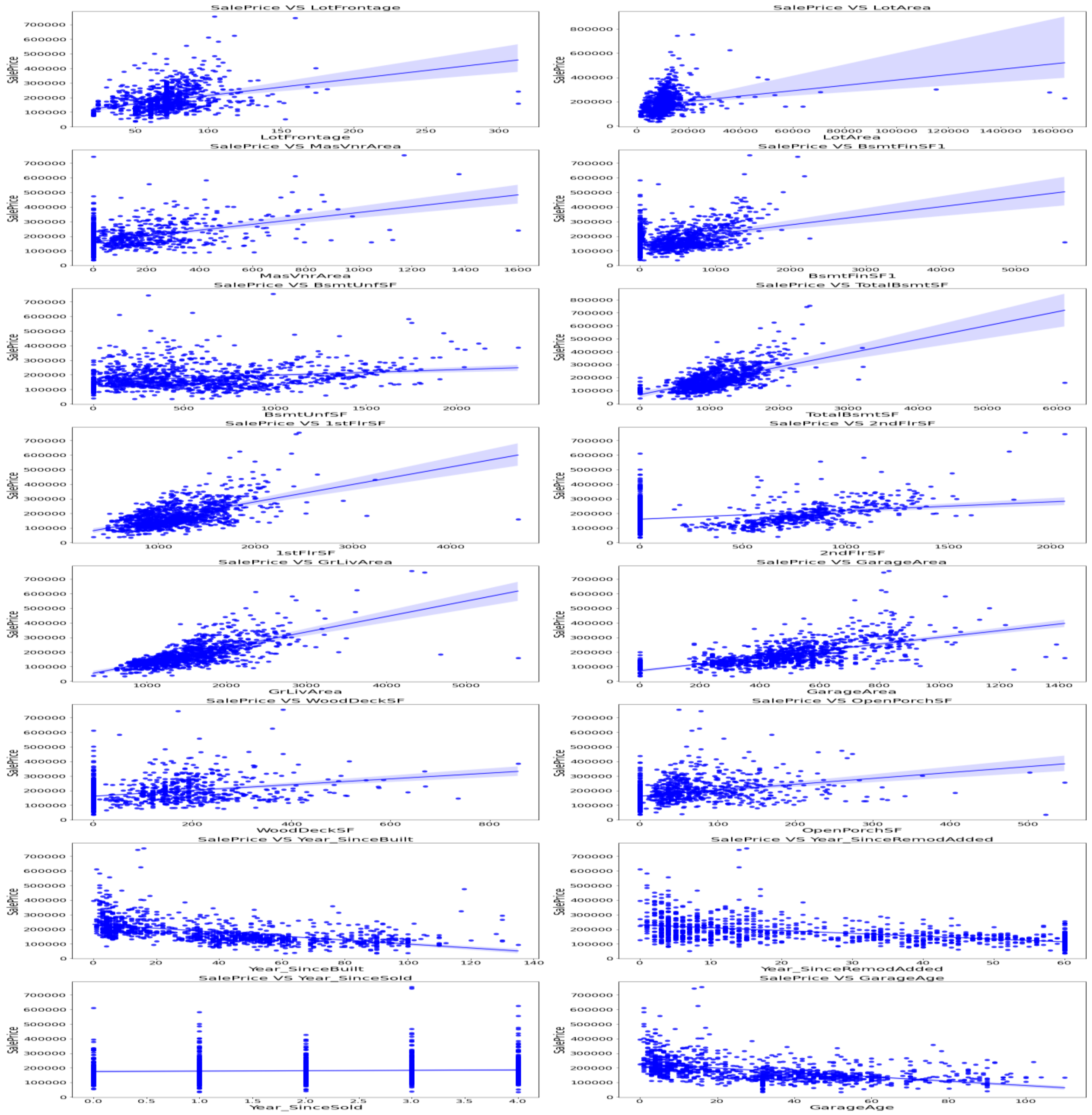


Bivariate Analysis:-

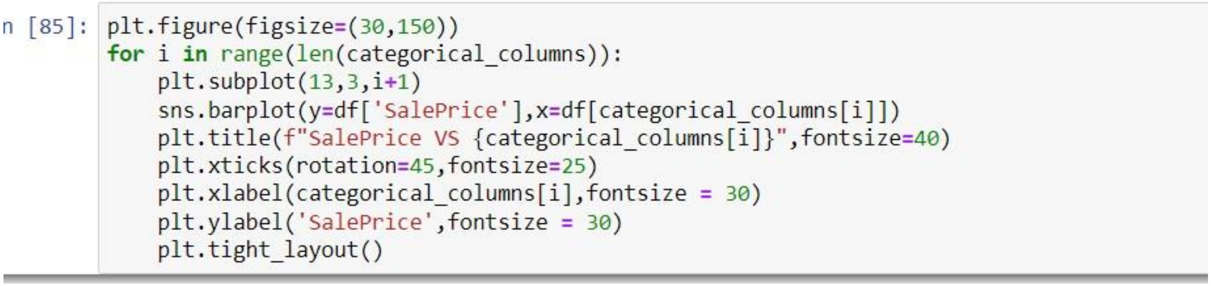


In [82]:

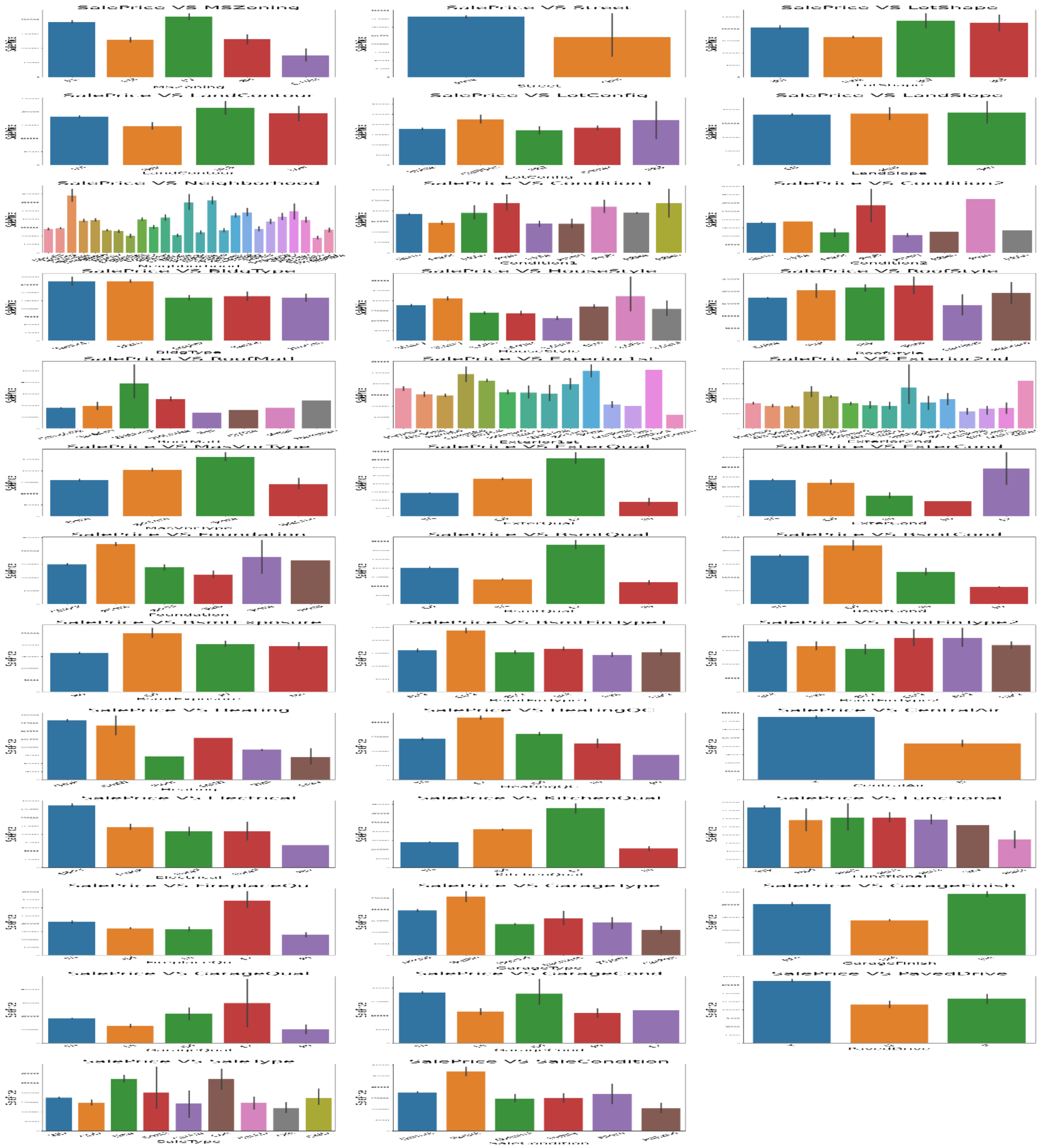
```
plt.figure(figsize=(20,130))
for i in range(len(col)):
    plt.subplot(20,2,i+1)
    sns.regplot(x=df[col[i]] , y=df['SalePrice'],color="b")
    plt.title(f"SalePrice VS {col[i]}",fontsize=20)
    plt.xticks(fontsize=15)
    plt.yticks(fontsize=15)
    plt.xlabel(col[i],fontsize = 20)
    plt.ylabel('SalePrice',fontsize = 20)
    plt.tight_layout()
```



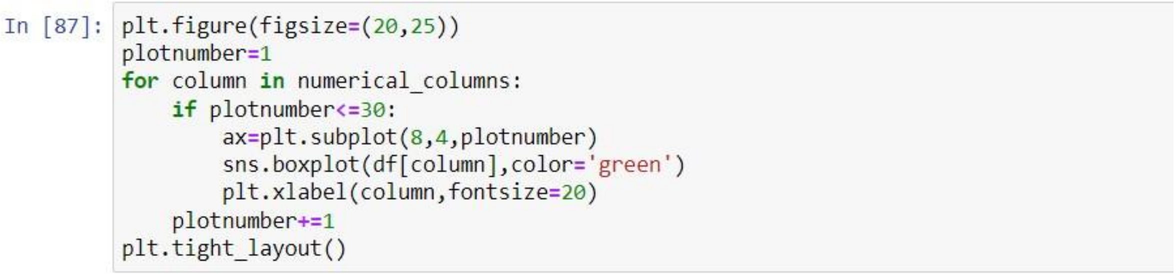
Bivariate Analysis for Categorical Columns:-



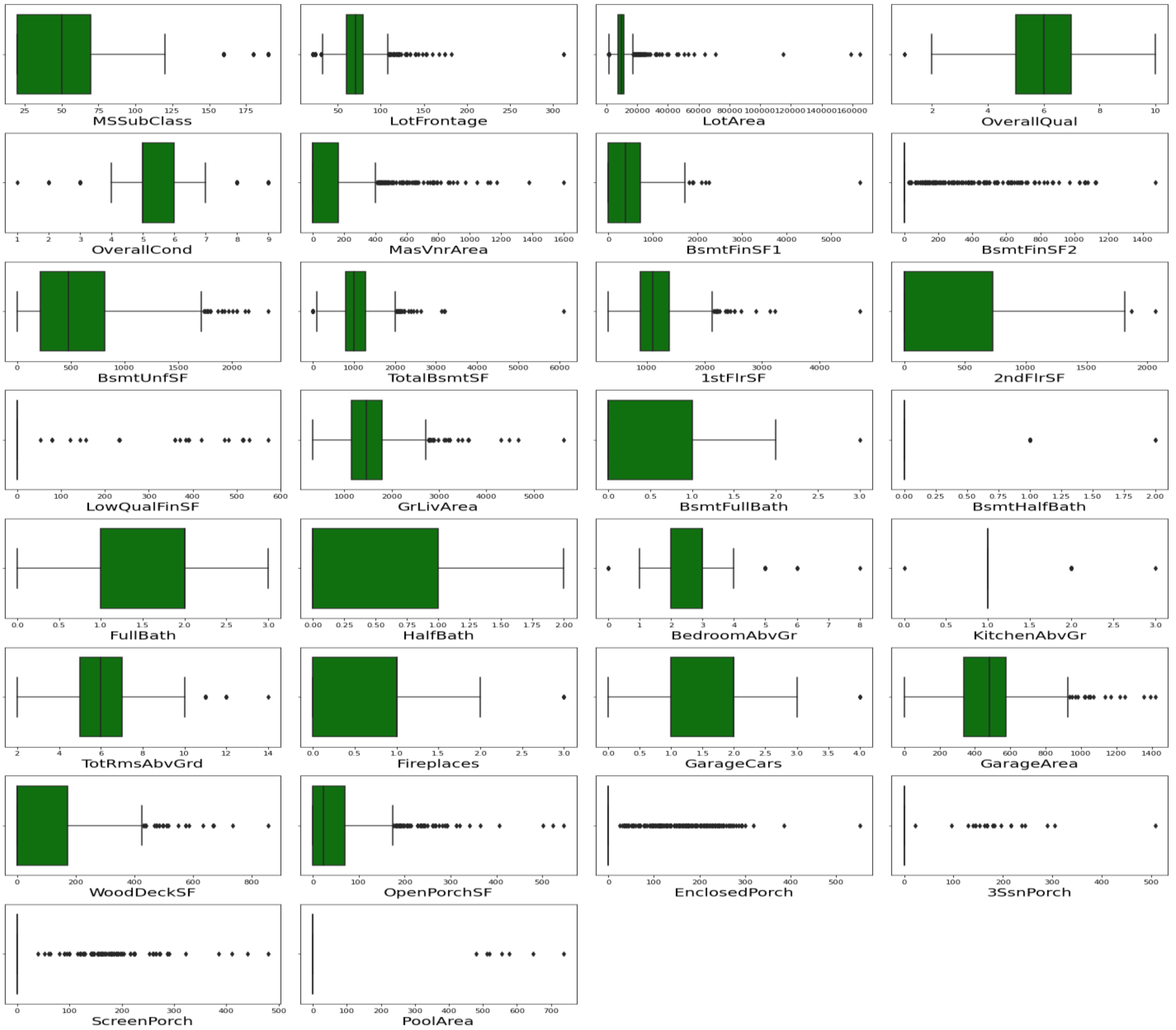
```
n [85]: plt.figure(figsize=(30,150))
        for i in range(len(categorical_columns)):
            plt.subplot(13,3,i+1)
            sns.barplot(y=df['SalePrice'],x=df[categorical_columns[i]])
            plt.title(f"SalePrice VS {categorical_columns[i]}",fontsize=40)
            plt.xticks(rotation=45,fontsize=25)
            plt.xlabel(categorical_columns[i],fontsize = 30)
            plt.ylabel('SalePrice',fontsize = 30)
            plt.tight_layout()
```



Checking for outliers:-

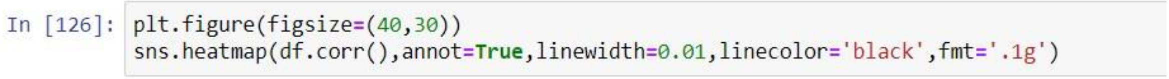


```
In [87]: plt.figure(figsize=(20,25))
         plotnumber=1
         for column in numerical_columns:
             if plotnumber<=30:
                 ax=plt.subplot(8,4,plotnumber)
                 sns.boxplot(df[column],color='green')
                 plt.xlabel(column,fontsize=20)
             plotnumber+=1
         plt.tight_layout()
```

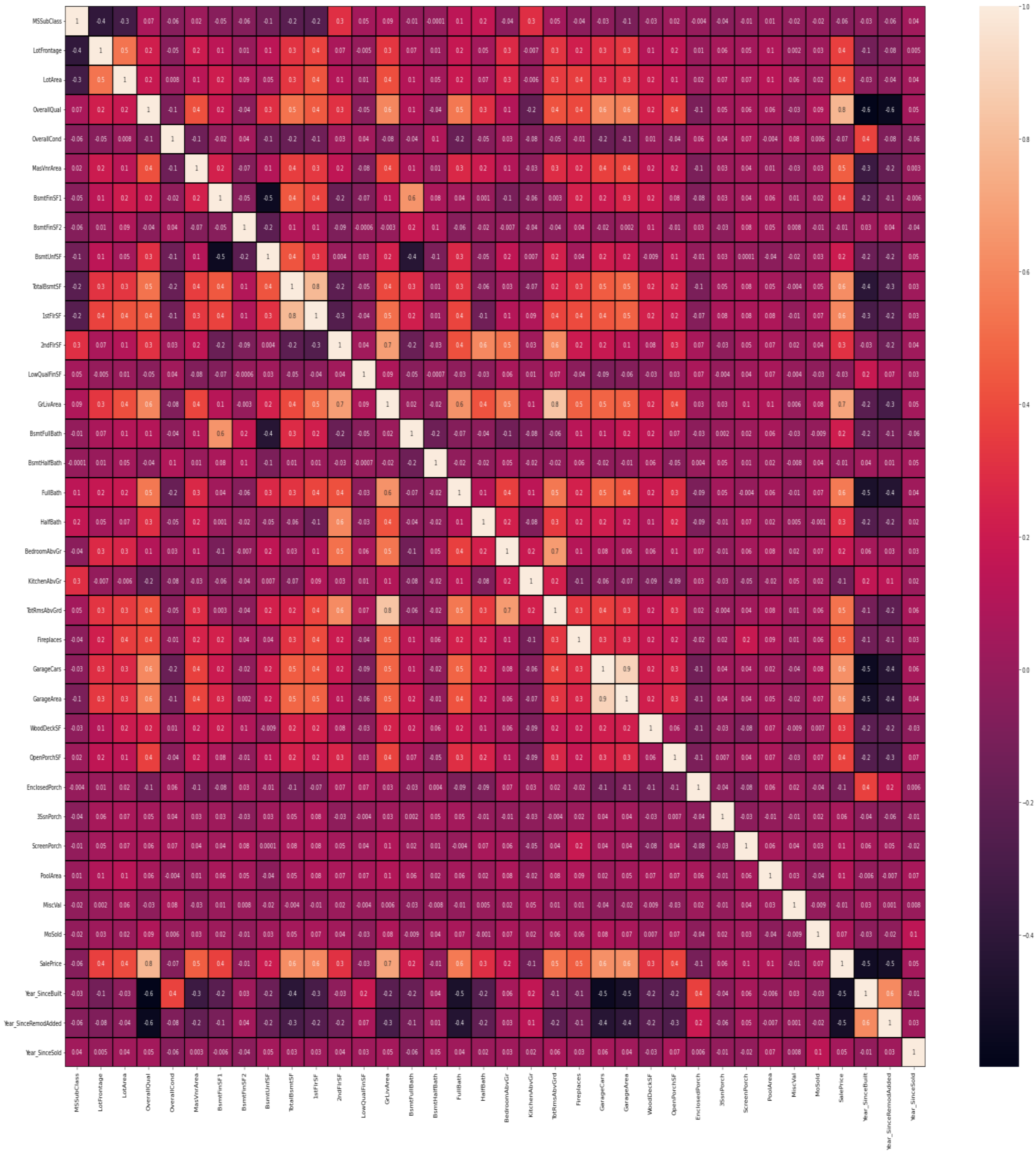


Observation:- SalePrice is my target i should not remove outliers from this column.And MSSubClass, OverallQual and OverallCond are seems to be categorical so let me not remove outliers in this columns.

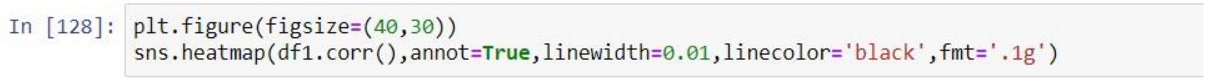
Checking correlation



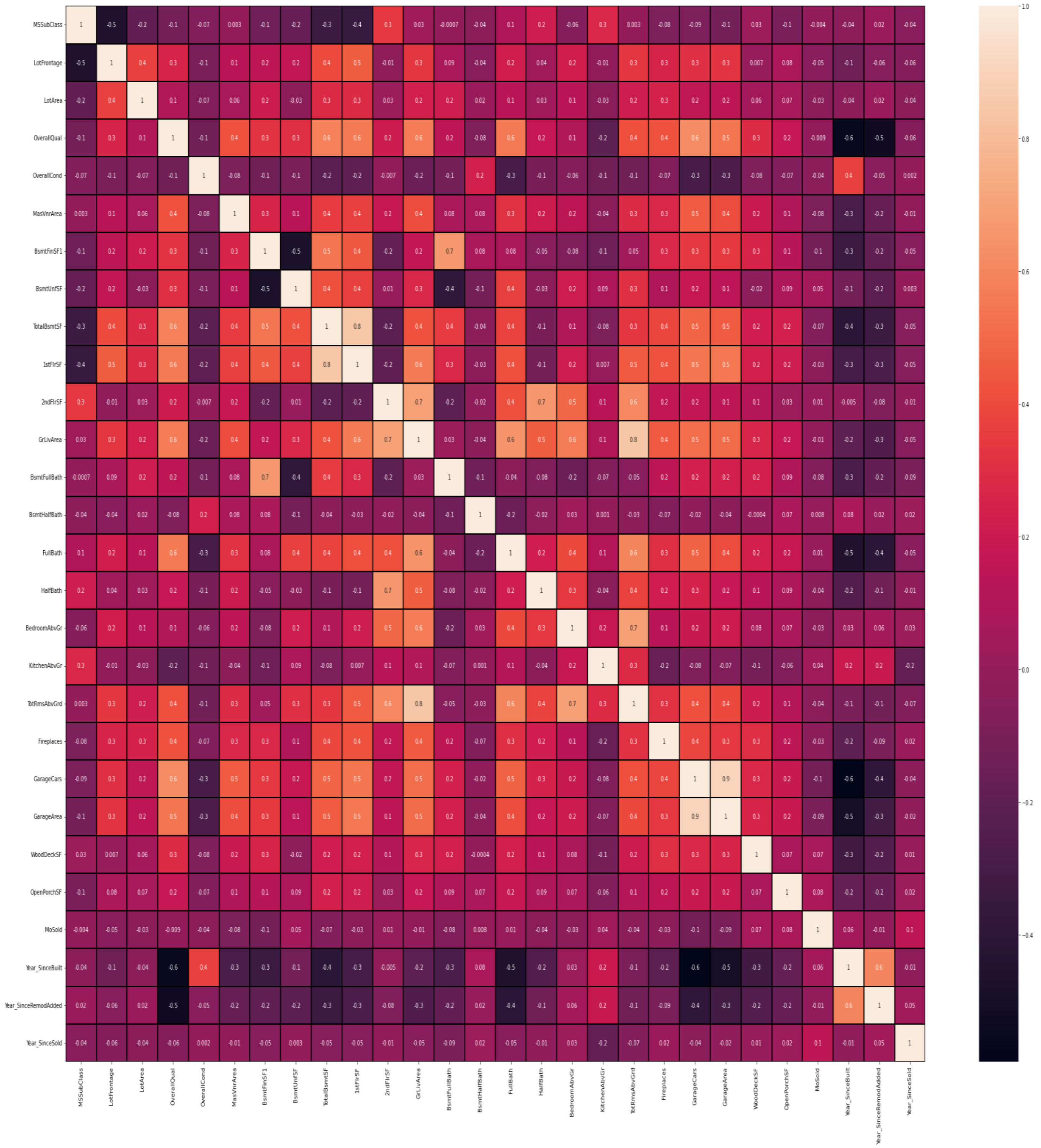
```
In [126]: plt.figure(figsize=(40,30))
          sns.heatmap(df.corr(),annot=True,linewidth=0.01,linecolor='black',fmt='.1g')
```



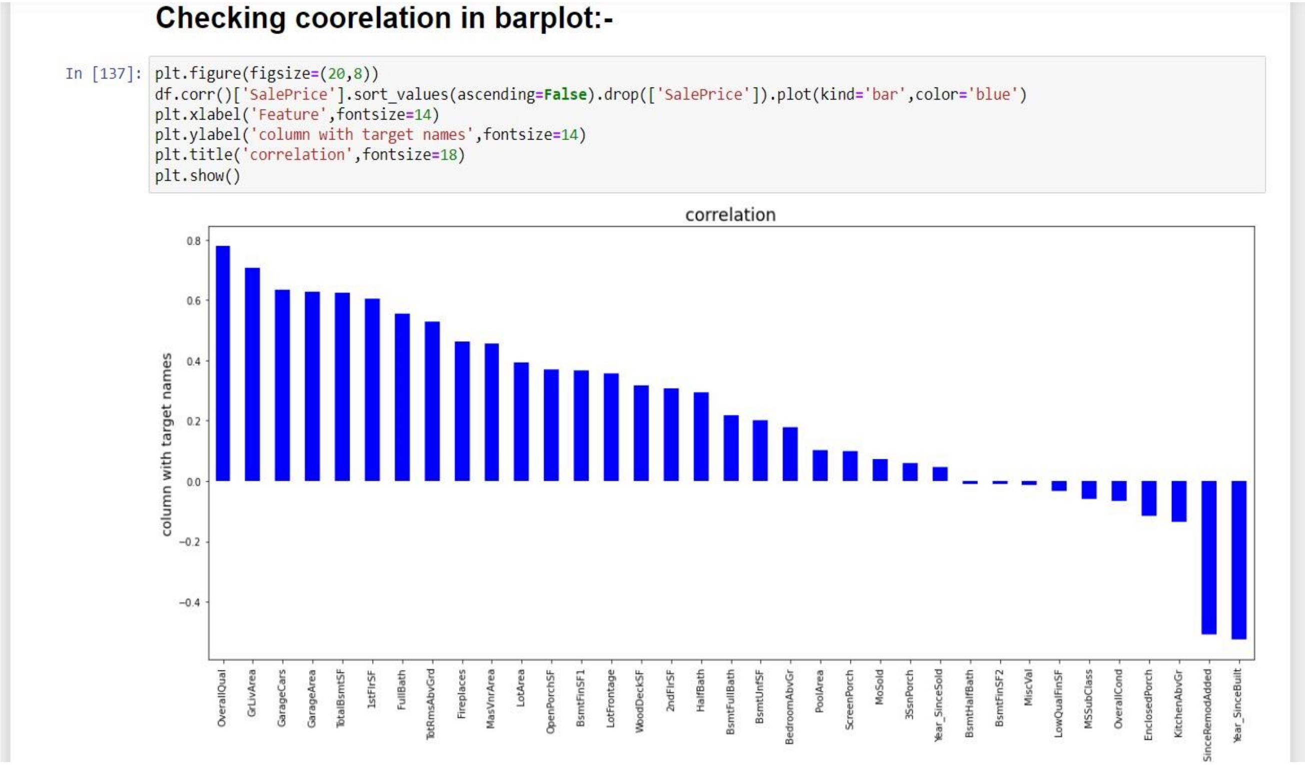
Observation:- Here I can clearly observe a multicolinearity issue in some of the features of train dataset so i have to check VIF and Let me plot a bar graph to get better insight on targets correlation with other features.



```
In [128]: plt.figure(figsize=(40,30))
          sns.heatmap(df1.corr(),annot=True,linewidth=0.01,linecolor='black',fmt='.1g')
```



Observation:- I can clearly observe a multicolinearity issue in some of the features of test dataset so i have to check VIF and Let me plot a bar graph to get better insight on targets correlation with other features.



Checking coorelation in barplot:-

In [137]:
```
plt.figure(figsize=(20,8))
df.corr()['SalePrice'].sort_values(ascending=False).drop(['SalePrice']).plot(kind='bar',color='blue')
plt.xlabel('Feature',fontsize=14)
plt.ylabel('column with target names',fontsize=14)
plt.title('correlation',fontsize=18)
plt.show()
```

Observation:- Here i can clearly see thatthat MasVnrType and ForeplaceQu are less correlated with target but let me keep those columns as it is and continue.

Interpretation of the Results:-

- This dataset was very special as it had separate train and test datasets. We have to work with both datasets simultaneously.
- Firstly, the datasets were having null values and zero entries in maximum columns so we have to be careful while going through the statistical analysis of the datasets.
- And proper ploting for proper type of features will help us to get better insight on the data. I found maximum numerical continuous columns were in linear relationship with target column.

- I notice a huge amount of outliers and skewness in the data so we have choose proper methods to deal with the outliers and skewness. If we ignore this outliers and skewness we may end up with a bad model which has less accuracy. Then scaling both train and test dataset has a good impact like it will help the model not to get baised.
- We have to use multiple models while building model using train dataset as to get the best model out of it.
- And we have to use multiple metrics like mae, mse, rmse and r2_score which will help us to decide the best model.
- I found ExtraTreesRegressor as the best model with 89.66% r2_score. Also I have improved the accuracy of the best model by running hyper parameter tunning.
- At last I have predicted the SalePrice for test dataset using saved model of train dataset. It was good!! that I was able to get the predictions near to actual value

CONCLUSION:-

1 Key Findings and Conclusions of the Study:-

In this project report, we have used machine learning algorithms to predict thehouse prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the features. Thus we can select the features which are not correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a csv file was generated consisting of predicted house prices. Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the dataframe of predicted prices of test dataset.

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

I found that the dataset was quite interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove missing value and to replace null value and zero values with there respective mean, median or mode. This study is an exploratory attempt to use five machine learning algorithms in estimating housing prices, and then compare their results.

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

- First draw back is the data leakage when we merge both train and test datasets.Followed by more number of outliers and skewness these two will reduce our model accuracy.
- Also, we have tried best to deal with outliers, skewness, null values and zero values. So it looks quite good that we have achieved a accuracy of 89% even after dealing all these drawbacks.
- Also, this study will not cover all regression algorithms instead, it is focusedon the chosen algorithm, starting from the basic regression techniques tothe advanced ones.
- This model doesn't predict future prices of the houses mentioned by the customer. Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process