# Housing project

**Submitted by:**

**Pooja Rajpal**

**Internship Batch 33**

# Acknowledgement

I would like to express my appreciation to my institute

(Data Trained Education)

And to my respected teacher(Shankar sir) for his teachings and notes he provided to the students.I would like to thank my institute which is always available to answer our queries.

This project has been source of learning and bring our theoretical knowledge into real time projects .

I would really acknowledge my teacher’s help and guidance.

# Introduction

## Business problem statement:

This problem statement is based on predicting the purchase of houses at in low prices and selling them at higher prices to decide whether to purchase or not .As we all know that housing is the basic necessity for each one of us in the world. Housing market is playing an important role in world economy.There are so many companies working in this real estate market .This problem is related to one such housing company who has decided to enter the Australian market .The company is looking at some properties to buy some houses to enter the market.Building the model to predict the actual value of properties and decide whether to invest in them or not.In the real world this model is the way for the company to decide the pricing dynamics of market.

## Background of the problem:

To understand the project it is necessary to understand what the project is about .This project is about the us based company who want to enter the real estate market in Australia .Before entering the company should know about basic features so that the company can earn profit .

Housing project

1) Ms sub class:-

This includes

- 1story (new and old)
- 2 story(new and old)
- Finished
- Unfinished
- Multilevel
- Family conversion

2) Ms zoning

Includes whether it is

- Agricultural

- Commercial
- Village residential
- Industrial
- Residential  high density
- Residential  low density

3) Lot Frontage area  means length of the plot
4) Lot Area :lot size
5) Street :which type of road to get the access to the property
6) Alley:narrow lane reserved for pedestrians
7) Lot shape:shape of the property
8) LandContour:vertical distance or difference in elevation between contourlines
9) Utilities:what all the facilities are available near that property
10) LotConfig:a method for locating buildings
11) Land slope :slope of the property
12) Neighborhood:physical locations nearby
13) Condition 1
14) Condition 2
15) Bldg. type:which type of building:single family,duplex
16) House style:one story,2 story
17) Overallqual:which type of material used to finish the products

For better understanding of the project one must know the detail of all the features

## Review of literature:

This is a project about housing in which there are many features and only one label ie sale price .all the features are responsible for predicting the target variable .This is a regression problem because the label is continuous data.there is a detailed information about the features in the data description All the features are responsible for building a model .This model is build to determine the actual value of the property for a us based company who want to enter in Australian market in real estate world .And want to earn profit by buying the houses on low price and selling in high prices.So various features are there from which we can take help to build a model.

In this dataset there are so many features .we will check the shape of the dataset to know the number of rows and columns.After that check for nulls and treat them and all the EDA to analyse the data that includes data visualization ,checking for skewness ,outliers and many more .After the EDA is over than we will start building a model .

But before starting it is necessary to observe and study the data .

## Motivation behind the problem:

Building project will improve my skills and make things better.Building successful projects gives the feeling of pride .And for that I need to build more and more projects passionately and try to do it with more accuracy.

However completing the project requires complex steps.Things get really difficult at some stages.but we need to adjust our vision .And try once again.Also its my passion to build a model with high accuracy and will definitely try to learn more n more for that.

# Analytic Problem Framing

- Mathematical/ Analytical Modeling of the Problem
  The first step here is to import necessary libraries.And load the dataset in jupyter notebook to analyse it.

  Statistical Modelling is necessary because it summarizes the results can be observed by the evaluators .It is relationship between variables.

  **Checking for nulls is another step in this dataset and impute them using imputers to clean the data.**

  **Statistical techniques such as mean,median ,standard deviation ,interquartile ranges .These are simple and we can start it as starting point for EDA.**
  **The goal is to collect the data and make predictions about real world.**

  **Two types statistics (Descriptive and inferential)**
  **Eg(weather forecasting ,stock market,loan approvaland fraud detection ,housingetc)**

  **Mean ,median and mode are the measures of central tendency and is descriptive statistics.**

  **Sampling of data and infer the results to describe entire population.**

# Housing project

**Submitted by:**

**Pooja Rajpal**

**Internship Batch 33**

# Acknowledgement

I would like to express my appreciation to my institute

(Data Trained Education)

And to my respected teacher(Shankar sir) for his teachings and notes he provided to the students.I would like to thank my institute which is always available to answer our queries.

This project has been source of learning and bring our theoretical knowledge into real time projects .

I would really acknowledge my teacher’s help and guidance.

# Introduction

## Business problem statement:

This problem statement is based on predicting the purchase of houses at in low prices and selling them at higher prices to decide whether to purchase or not .As we all know that housing is the basic necessity for each one of us in the world. Housing market is playing an important role in world economy.There are so many companies working in this real estate market .This problem is related to one such housing company who has decided to enter the Australian market .The company is looking at some properties to buy some houses to enter the market.Building the model to predict the actual value of properties and decide whether to invest in them or not.In the real world this model is the way for the company to decide the pricing dynamics of market.

## Background of the problem:

To understand the project it is necessary to understand what the project is about .This project is about the us based company who want to enter the real estate market in Australia .Before entering the company should know about basic features so that the company can earn profit .

Housing project

18) Ms sub class:-

This includes

- 1story (new and old)
- 2 story(new and old)
- Finished
- Unfinished
- Multilevel
- Family conversion

19) Ms zoning

Includes whether it is

- Agricultural

- Commercial
- Village residential
- Industrial
- Residential  high density
- Residential  low density

20) Lot Frontage area  means length of the plot
21) Lot Area :lot size
22) Street :which type of road to get the access to the property
23) Alley:narrow lane reserved for pedestrians
24) Lot shape:shape of the property
25) LandContour:vertical distance or difference in elevation between contourlines
26) Utilities:what all the facilities are available near that property
27) LotConfig:a method for locating buildings
28) Land slope :slope of the property
29) Neighborhood:physical locations nearby
30) Condition 1
31) Condition 2
32) Bldg. type:which type of building:single family,duplex
33) House style:one story,2 story
34) Overallqual:which type of material used to finish the products

For better understanding of the project one must know the detail of all the features

## Review of literature:

This is a project about housing in which there are many features and only one label ie sale price .all the features are responsible for predicting the target variable .This is a regression problem because the label is continuous data.there is a detailed information about the features in the data description All the features are responsible for building a model .This model is build to determine the actual value of the property for a us based company who want to enter in Australian market in real estate world .And want to earn profit by buying the houses on low price and selling in high prices.So various features are there from which we can take help to build a model.

In this dataset there are so many features .we will check the shape of the dataset to know the number of rows and columns.After that check for nulls and treat them and all the EDA to analyse the data that includes data visualization ,checking for skewness ,outliers and many more .After the EDA is over than we will start building a model .

But before starting it is necessary to observe and study the data .

## Motivation behind the problem:

Building project will improve my skills and make things better.Building successful projects gives the feeling of pride .And for that I need to build more and more projects passionately and try to do it with more accuracy.

However completing the project requires complex steps.Things get really difficult at some stages.but we need to adjust our vision .And try once again.Also its my passion to build a model with high accuracy and will definitely try to learn more n more for that.

# Analytic Problem Framing

- Mathematical/ Analytical Modeling of the Problem
  The first step here is to import necessary libraries.And load the dataset in jupyter notebook to analyse it.

  Statistical Modelling is necessary because it summarizes the results can be observed by the evaluators .It is relationship between variables.

  **Checking for nulls is another step in this dataset and impute them using imputers to clean the data.**

  **Statistical techniques such as mean,median ,standard deviation ,interquartile ranges .These are simple and we can start it as starting point for EDA.**
  **The goal is to collect the data and make predictions about real world.**

  **Two types statistics (Descriptive and inferential)**
  **Eg(weather forecasting ,stock market,loan approvaland fraud detection ,housingetc)**

  **Mean ,median and mode are the measures of central tendency and is descriptive statistics.**

  **Sampling of data and infer the results to describe entire population.**

# Housing project

**Submitted by:**

**Pooja Rajpal**

**Internship Batch 33**

# Acknowledgement

I would like to express my appreciation to my institute

(Data Trained Education)

And to my respected teacher(Shankar sir) for his teachings and notes he provided to the students.I would like to thank my institute which is always available to answer our queries.

This project has been source of learning and bring our theoretical knowledge into real time projects .

I would really acknowledge my teacher’s help and guidance.

# Introduction

## Business problem statement:

This problem statement is based on predicting the purchase of houses at in low prices and selling them at higher prices to decide whether to purchase or not .As we all know that housing is the basic necessity for each one of us in the world. Housing market is playing an important role in world economy.There are so many companies working in this real estate market .This problem is related to one such housing company who has decided to enter the Australian market .The company is looking at some properties to buy some houses to enter the market.Building the model to predict the actual value of properties and decide whether to invest in them or not.In the real world this model is the way for the company to decide the pricing dynamics of market.

## Background of the problem:

To understand the project it is necessary to understand what the project is about .This project is about the us based company who want to enter the real estate market in Australia .Before entering the company should know about basic features so that the company can earn profit .

Housing project

35) Ms sub class:-

This includes

- 1story (new and old)
- 2 story(new and old)
- Finished
- Unfinished
- Multilevel
- Family conversion

36) Ms zoning

Includes whether it is

- Agricultural

- Commercial
- Village residential
- Industrial
- Residential  high density
- Residential  low density

37) Lot Frontage area  means length of the plot
38) Lot Area :lot size
39) Street :which type of road to get the access to the property
40) Alley:narrow lane reserved for pedestrians
41) Lot shape:shape of the property
42) LandContour:vertical distance or difference in elevation between contourlines
43) Utilities:what all the facilities are available near that property
44) LotConfig:a method for locating buildings
45) Land slope :slope of the property
46) Neighborhood:physical locations nearby
47) Condition 1
48) Condition 2
49) Bldg. type:which type of building:single family,duplex
50) House style:one story,2 story
51) Overallqual:which type of material used to finish the products

For better understanding of the project one must know the detail of all the features

## Review of literature:

This is a project about housing in which there are many features and only one label ie sale price .all the features are responsible for predicting the target variable .This is a regression problem because the label is continuous data.there is a detailed information about the features in the data description All the features are responsible for building a model .This model is build to determine the actual value of the property for a us based company who want to enter in Australian market in real estate world .And want to earn profit by buying the houses on low price and selling in high prices.So various features are there from which we can take help to build a model.

In this dataset there are so many features .we will check the shape of the dataset to know the number of rows and columns.After that check for nulls and treat them and all the EDA to analyse the data that includes data visualization ,checking for skewness ,outliers and many more .After the EDA is over than we will start building a model .

But before starting it is necessary to observe and study the data .

## Motivation behind the problem:

Building project will improve my skills and make things better.Building successful projects gives the feeling of pride .And for that I need to build more and more projects passionately and try to do it with more accuracy.

However completing the project requires complex steps.Things get really difficult at some stages.but we need to adjust our vision .And try once again.Also its my passion to build a model with high accuracy and will definitely try to learn more n more for that.

# Analytic Problem Framing

- Mathematical/ Analytical Modeling of the Problem
  The first step here is to import necessary libraries.And load the dataset in jupyter notebook to analyse it.

  Statistical Modelling is necessary because it summarizes the results can be observed by the evaluators .It is relationship between variables.

  **Checking for nulls is another step in this dataset and impute them using imputers to clean the data.**

  **Statistical techniques such as mean,median ,standard deviation ,interquartile ranges .These are simple and we can start it as starting point for EDA.**
  **The goal is to collect the data and make predictions about real world.**

  **Two types statistics (Descriptive and inferential)**
  **Eg(weather forecasting ,stock market,loan approvaland fraud detection ,housingetc)**

  **Mean ,median and mode are the measures of central tendency and is descriptive statistics.**

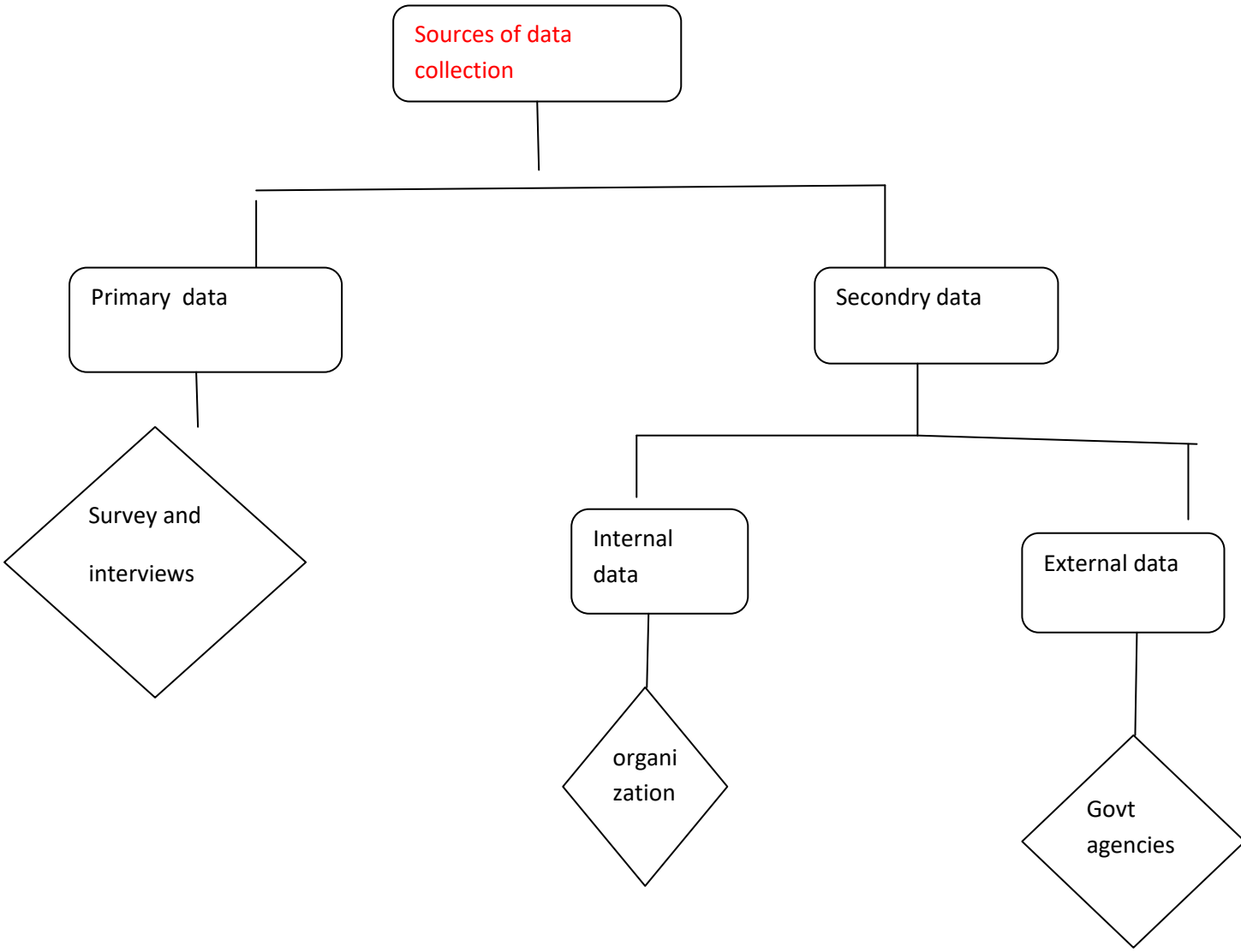
  **Sampling of data and infer the results to describe entire population.**

Central limit theorem that normalises the non normal distribution .if the sample size is >30 distribution start looking normal

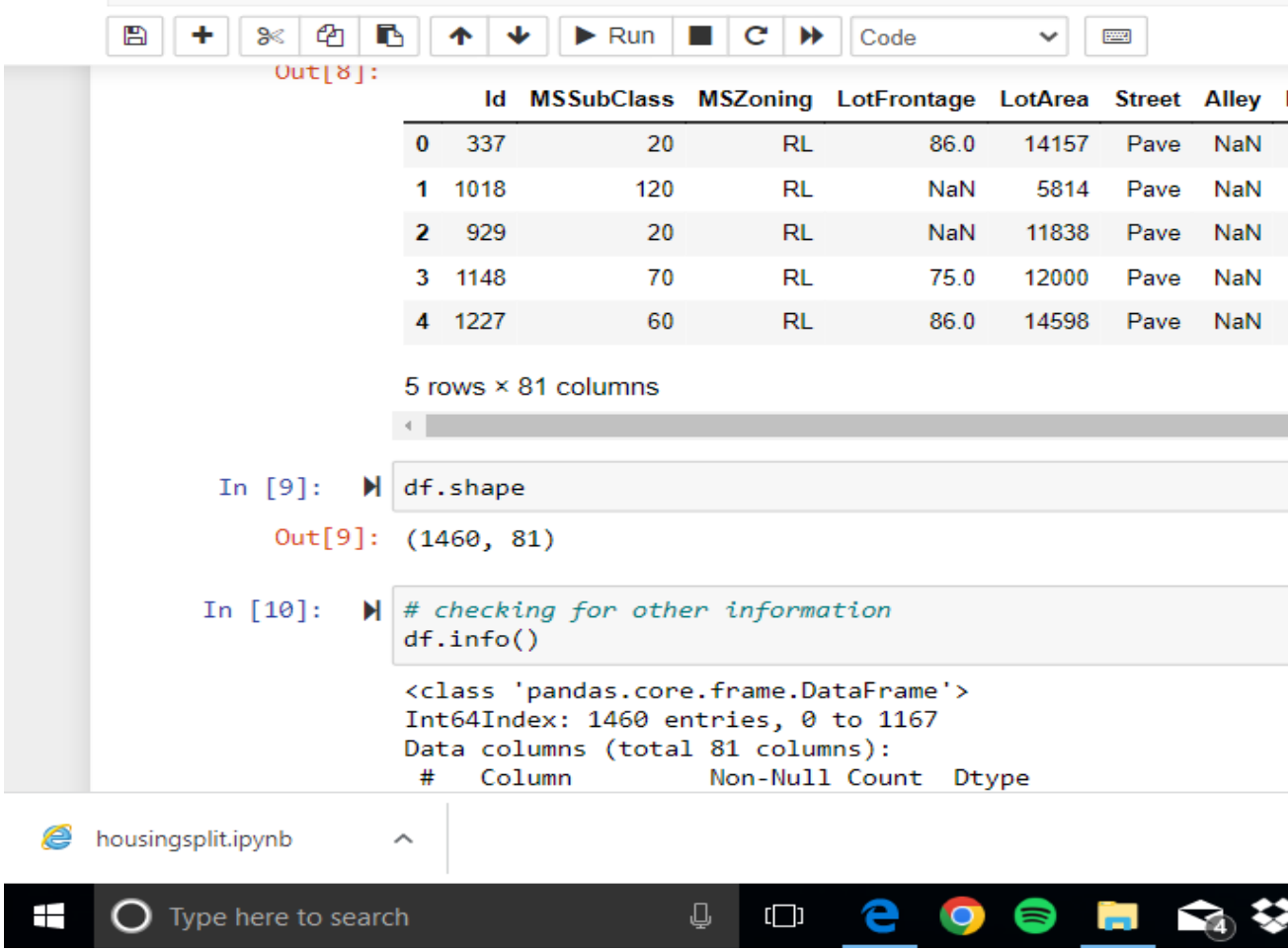
**Normal distribution, Bernoulli distribution and binomial distribution, uniform distribution are types of distribution**

- Data Sources and their formats

  **The data is collected is raw data which is not useful ,cleaning of raw data and utilizing the data for further analysis.Data collection is the process of collecting data whether in structured format or unstructured format**



**This data is internal data in the form of CSV file given to us by our organization which need to be cleaned to make a model.**



Out[8]:

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley |
|---|---|---|---|---|---|---|---|
| 0 | 337 | 20 | RL | 86.0 | 14157 | Pave | NaN |
| 1 | 1018 | 120 | RL | NaN | 5814 | Pave | NaN |
| 2 | 929 | 20 | RL | NaN | 11838 | Pave | NaN |
| 3 | 1148 | 70 | RL | 75.0 | 12000 | Pave | NaN |
| 4 | 1227 | 60 | RL | 86.0 | 14598 | Pave | NaN |

5 rows × 81 columns

In [9]:
```
df.shape
```

Out[9]: (1460, 81)

In [10]:
```
# checking for other information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1460 entries, 0 to 1167
Data columns (total 81 columns):
 #   Column         Non-Null Count  Dtype
```

- Data Preprocessing Done

**Step1:collection of data and load the dataset in jupyter notebook using pandas and for that we need to import pandas library .**

**Step2:data cleaning that include checking for nulls and treat them using imputers.**

**Step3:checking for datatypes and if object datatype convert them into integer as computers only understand numeric data**

**Step 4:data visualisation**

**Step 5:plotting boxplot for columns to check for outliers and treating them using outlier detection**

**Step6:checking for skewness and treat them using Power transformer**

**Step 7:when the data is cleaned we will check for collinearity using heatmap**

**At the end split the dataset into features and label**

- Data Inputs- Logic- Output Relationships:
  All the other variables except sale price are input variables for eg Lot frontage,Lot area,street,Alley ,utilities etc are input variables.The column on which we have to predict for  new input data is called output variable.All the input variables have more or less relationship with output variable.Input variables are necessary for predicting output variable.

  - Hardware and Software Requirements and Tools Used:
    Laptop and
    Anaconda navigator is desktop graphical user interface that allows to launch applications
    Jupyter notebook is open source software.

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

Regularization techniques,hyper parameter tuning to increase the accuracy.

- Testing of Identified Approaches (Algorithms):
  As it is a regression problem Linear Regression and RandomForestRegression algorithm is used

- Run and Evaluate selected models
  With linear regression

```
In [266]:  # lets import necessary libraries to build a model
           from sklearn.linear_model import LinearRegression
           from sklearn.model_selection import train_test_split,GridSearchCV,c
           from sklearn.metrics import accuracy_score,r2_score,classification_
           from sklearn.linear_model import Lasso
           from sklearn.ensemble import RandomForestRegressor

In [267]:  lr=LinearRegression()

In [268]:  for i in range(0,50):
               x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.
               lr.fit(x_train,y_train)
               pred_train=lr.predict(x_train)
               pred_test=lr.predict(x_test)
               print(f"at random state{i}the training accuracy is{r2_score(y_t
               print(f"at random state{i}the testing accuracy is{r2_score(y_te
               print("\n")

           at random state0the training accuracy is0.6793829056307057
           at random state0the testing accuracy is0.6035217893278018
```

housingsplit.ipynb

Type here to search

```
print("\n")
```

```
at random state6the training accuracy is0.6940911806241895
at random state6the testing accuracy is-7.640579006482606e+19

at random state7the training accuracy is0.6603131342518904
at random state7the testing accuracy is-8.058457420888148e+19

at random state8the training accuracy is0.6907013022554843
at random state8the testing accuracy is0.5832339797748557

at random state9the training accuracy is0.6828889536591483
at random state9the testing accuracy is0.5955182564245193

at random state10the training accuracy is0.6859434349953792
at random state10the testing accuracy is0.5662166110132427
```

In [269]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random
```

usingsplit.ipynb

Type here to search

jupyter housingsplit Last Checkpoint: Last Saturday at 19:44 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

```
pred_test=lr.predict(x_test)
print(r2_score(y_test,pred_test))
```

```
0.6618073265019697
```

In [272]:
```
#regularization
parameters={'alpha':[.0001,.001,.01,.1,1],
            'random_state':list(range(0,10))}
ls=Lasso()
clf=GridSearchCV(ls,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'alpha': 1, 'random_state': 0}
```

In [279]:
```
ls=Lasso(alpha=.0001,random_state=0)
ls.fit(x_train,y_train)
```

Out[279]:
```
Lasso
Lasso(alpha=0.0001, random_state=0)
```

In [280]:
```
ls_score_training=ls.score(x_train,y_train)
pred_ls=ls.predict(x_test)
```

singsplit.ipynb

Type here to search

### With random forest regressor

```
Out[280]: 66.18073316626602

In [281]: #ensemble technique
parameters={'criterion':['mse','mae'],
            'max_features':["auto","sqrt","log2"]}
rf=RandomForestRegressor()
clf=GridSearchCV(rf,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)

{'criterion': 'mse', 'max_features': 'log2'}

In [282]: rf=RandomForestRegressor(criterion='mse',max_features='log2')
rf.fit(x_train,y_train)
rf.score(x_train,y_train)
pred_decision=rf.predict(x_test)
rfs=r2_score(y_test,pred_decision)
print('r2_score',rfs*100)

r2_score 71.69934561914127

In [289]: # lets save the model
```

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

Metrics are used to measure the performance of the model. I used the accuracy_score to find the accuracy of the model.because according to me it gives most reliable results and its easy to determine what is the level of accuracy of the model built.

- Visualizations

- Distplot to visualise whether the data is right, left skewed to make it

```
       'PavedDrive', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch',
       'ScreenPorch', 'PoolArea', 'Fence', 'MiscVal', 'MoSold', 'YrSold',
       'SaleType', 'SaleCondition', 'SalePrice'],
      dtype='object')
```

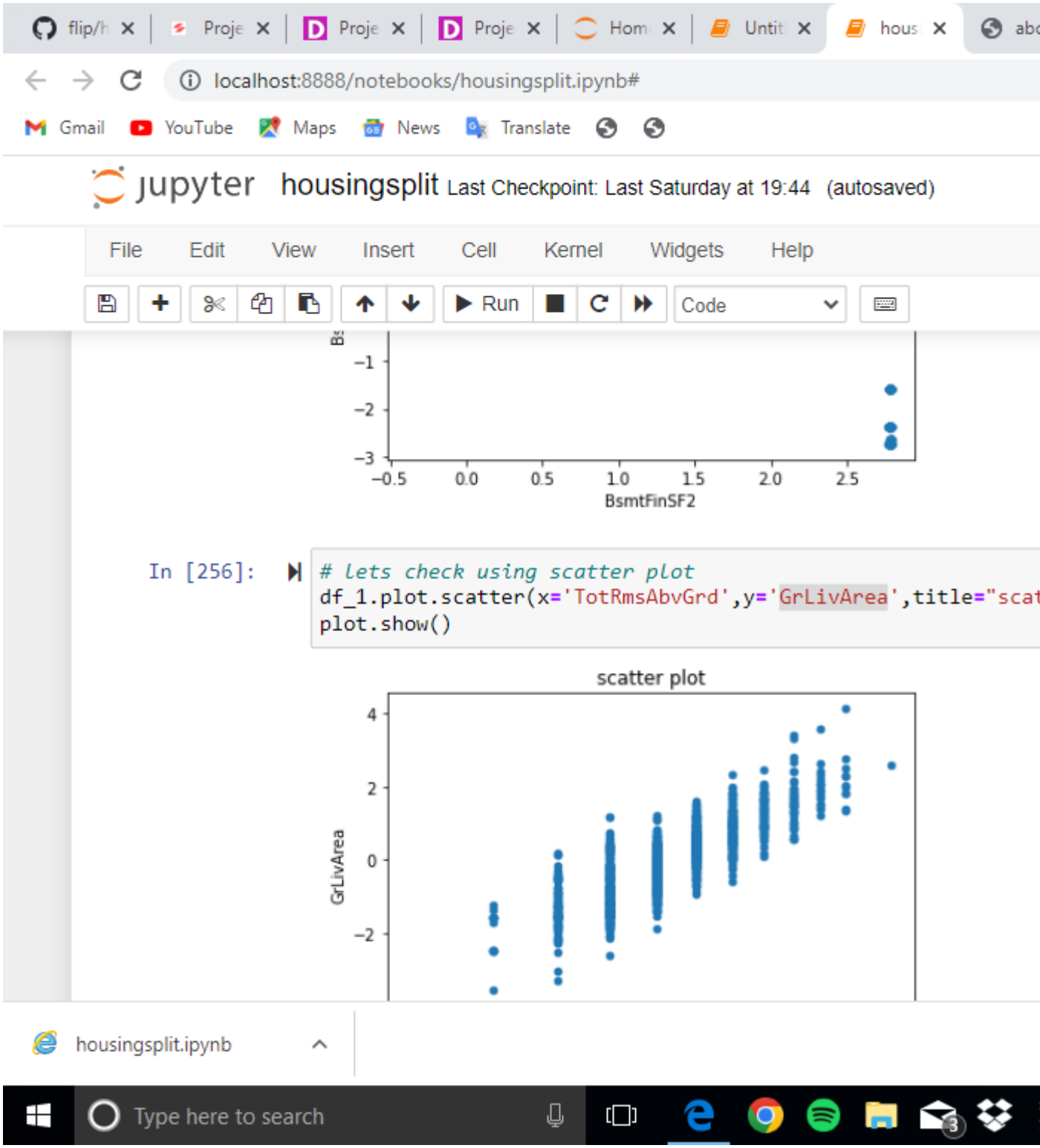
In [183]:
```
df_1.shape
```
Out[183]:
```
(1460, 78)
```

In [ ]:

In [98]:
```
# data visualization
```

In [184]:
```
#data visualization
plt.figure(figsize=(20,15))
plotnumber=1
for column in df_1:
    if plotnumber<=78:
        ax=plt.subplot(13,6,plotnumber)
        sns.distplot(df_1[column])
        plt.xlabel(column,fontsize=20)
    plotnumber+=1
plt.show()
```

## Scatter plot to see the relationship between features



- Interpretation of the Results

RandomForest Regressor works well with this project .As this is a regression problem because the label is continuous data .Preprocessing is necessary to clean the data
Visualization because it is easily understood by us to observe the data in graphical form.

# *conclusion*

**This is housing dataset which required data cleaning and Exploratory Data Analysis to analyse the data.After that checking the correlation between the features because it can affect the accuracy of the model .**

**Hyperparameter tuning is necessary to increase the accuracy.**

**Then building model and checking the accuracy.**

- Learning Outcomes of the Study in respect of Data Science
  I faced lot of challenges in making this project
  Also learnt lot of things As it has lot of columns found difficulty in finding the correlation between features.
  In regression problem the output variable must be continuous and in classification it must be descrete.

- Limitations of this work and Scope for Future Work
  I think I should have worked harder to increase the accuracy .
  Thank you