CMR UNIVERSITY
Private University Established in Karnataka State by Act No. 45 of 2013

School of Engineering and Technology

Main Campus, Off Hennur-Bagalur Main Road, Chagalahatti, Bengaluru-562149

INTERNSHIP REPORT
on
"House price prediction using Artificial Intelligence"

submitted to,
School of Engineering and Technology, CMR University

in partial fulfilment of the requirement for the award of the degree of

Bachelor of Technology,
in
computer science & engineering

by
SWAPNIL SUNIL HERAGE
(USN: 21BBTCS241)

Internship Carried
out at Code
Clause Pune

Internal Guide	External Guide:
Prof. Priyanka.S	**Rupam Mankar**
Associate Professor	Program coordinator
Dept. of CSE , SOET	Code Clause Pune

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Off Hennur - Bagalur Main Road,
Near Kempegowda International Airport, Chagalahatti,
Bangalore, Karnataka - 562149
2022 - 2023

CMR UNIVERSITY

Private University Established in Karnataka State by Act No. 45 of 2013

School of Engineering and Technology, CMR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Internship work entitled “House price prediction using Artificial intelligence”, submitted to the CMR University, Bangalore, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer science and Engineering is a record of work done by Mr./Ms. SWAPNIL SUNIL HERAGE bearing university register number 21BBTCS241 during the academic year 2022-23 at School of Engineering and Technology, CMR University, Bangalore under my supervision and guidance. The Internship report has been approved as it satisfies the academic requirement in respect of internship work prescribed for the said degree.

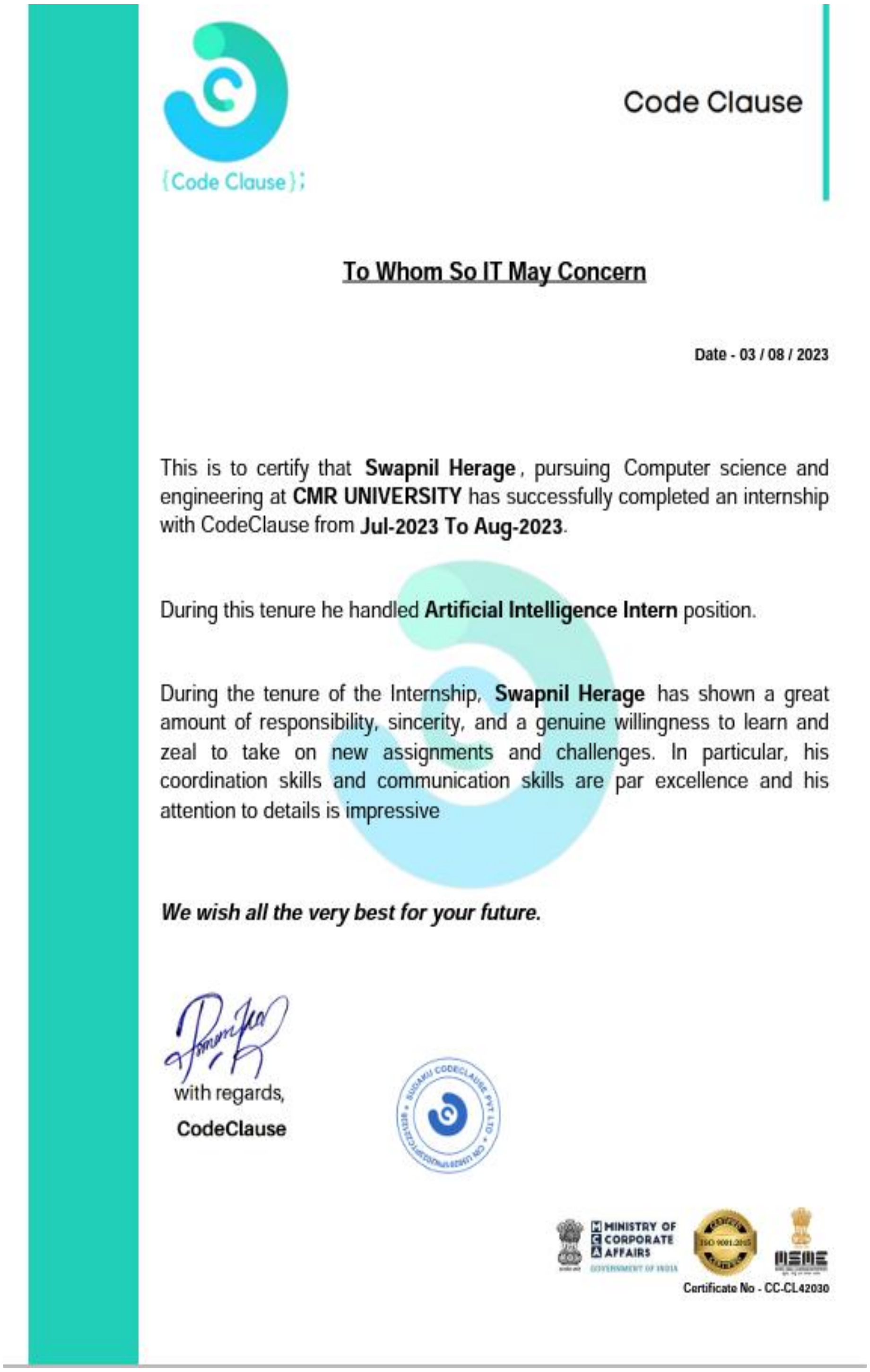
Prof. Priyanka.S	***Dr. Rubini P***	**Dr. V R Manjunath**
Signature of the Guide	*Signature of the HOD*	*Signature of the Dean*

Examiners

Name of the examiners **Signature with date**

1

2



{Code Clause};

Code Clause

To Whom So IT May Concern

Date - 03 / 08 / 2023

This is to certify that **Swapnil Herage** , pursuing Computer science and engineering at **CMR UNIVERSITY** has successfully completed an internship with CodeClause from **Jul-2023 To Aug-2023**.

During this tenure he handled **Artificial Intelligence Intern** position.

During the tenure of the Internship, **Swapnil Herage** has shown a great amount of responsibility, sincerity, and a genuine willingness to learn and zeal to take on new assignments and challenges. In particular, his coordination skills and communication skills are par excellence and his attention to details is impressive

We wish all the very best for your future.

with regards,

CodeClause

SUGAKU CODECLAUSE PVT LTD

MINISTRY OF CORPORATE AFFAIRS
GOVERNMENT OF INDIA

ISO 9001:2015

MSME

Certificate No - CC-CL42030

DECLARATION

I, SWAPNIL SUNIL HERAGE bearing USN 21BBTCS241, student of Bachelor of Technology, Computer science and Engineering, CMR University, Bengaluru, hereby declare that the internship work entitled " House Price Prediction using Artificial Intelligence" submitted by me, for the award of the Bachelor's degree in Computer Science And Engineering to CMR University is a record of bonafide work carried out independently by me under the supervision and guidance of prof. Priyanka Associate Professor, CSE Dept., CMR University.

I further declare that the work reported in this internship work has not been submitted and will not be submitted, either in part or in full, for the award of any other degree in this university or any other institute or University.

Date :03/08/2023

Place : Bangalore

SWAPNIL SUNIL HERAGE *(USN: 21BBTCS241)*

SIGNATURE OF STUDENT

ACKNOWLEDGEMENT

I express our foremost gratitude **to Dr V R Manjunath**, Dean for his constant support. We express our foremost gratitude to **Dr Rubini p**, Professor and HOD, Dept. of Computer science and Engineering for his constant support and guidance. We express our foremost gratitude to my guide **prof. Priyanka.S**, Assistant Professor Dept. of Computer Science & Engineering, CMR University, Bagalur for his/her inspiration, adroit guidance, constant supervision, direction and discussions in successful completion of this dissertation. Our sincere thanks to all teaching and non-Teaching Staff of Computer Science & Engineering for all the facilities provided, without which, we could not have progressed with my work. Thanks to our parents who have been a great source of strength in the completion of this dissertation.

SWAPNIL SUNIL HERAGE
(USN: 21BBTCS241)

ABSTRACT

The sales of the houses are determined on various factors like the location, area, population and some of the information to predict the indiviual housing price. In addition to these housing prices, the prediction of the housing prices can greatly assist in the prediction of the future housing prices of the real estate.This study uses the machine learning algorithms and technology as a research methodology to develop a housing price prediction. Many algorithms are used here to effectively increase the precentage of the prediction which is considered as the best models in the price prediction. This project shows us that the machine learning algorithm based on accuracy,consistency out performs the other in the performance of the housing price prediction. The project can be created using python(AI/ML),HTML,CSS,JSS.Python is used for writing the Machine Learing Algorithms .HTML,CSS and JS is used for designing the front end of the system. At last,I can conclude y saying that House Price Prediction system will be very helpful in detecting the prices of the houses and keeping the record of the high and low of the prices.So it will help the user to know the real price of the property,it could not be used for any fraud means.

ABOUT THE COMPANY

We offer reliable, efficient delivery with high-caliber engineers & finely-tuned software development processes.We Believe In Leadership to lead the technology to build a better future Integrity to follow truth and be real Accountability for our every commitment.

- We Imagine
- We Engineer
- We Modernize
- We Manage

Increasing revenue, improving efficiency, reducing cost—these are all accomplished by implementing innovative technology that’s purpose-built to solve the challenges holding your organization back.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

Earlier, it’s a very popular and common practice to price the property without the proper evaluation of the land, infrastructure etc. We need a proper prediction on the real estate and the houses in housing market we can see a mechanism that runs throughout the properties buying and selling buying a house will be a life time goal for most of the individual but, There are lot of people making mistake in india as most of the people are buying properties from the people they don’t know by seeing the news all around them.

In India, people buy properties which are too expensive but it’s not worth it. In the housing market 2016 the house sold in India was about 80 lakh but the real price according locality and size was about 60 lakh In earlier year, there was an economic collapse that give the clue to the impending disaster, this situation is currently happening and the prices of houses are getting higher compared to current economic situation of our country, the indian government fails to produce the data about the houses so it was very difficult for peoples to buy the properties. Therefore, the people searched on internet for the evidence for house price.

Many methods have been used in the price prediction like advance regression in this I am trying to predict the real estate price for the future using the machine learning techniques with the help of the previous works. I have used the multiple regression and more algorithms with different tools to predict the house price. The purpose of this paper is to establish the proper data prepocessing practices in order to increase the accuracy of machine learning algorithms.

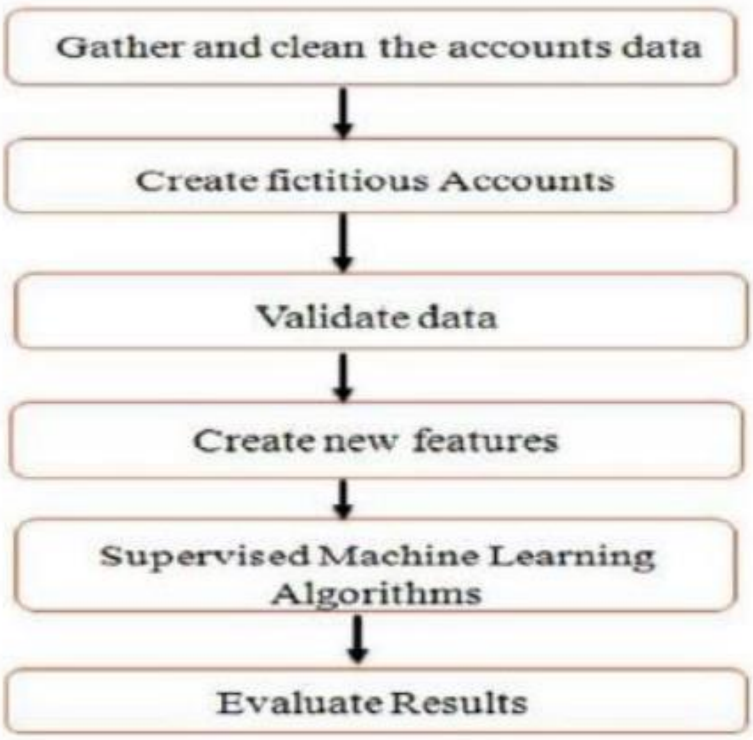

Gather and clean the accounts data
Create fictitious Accounts
Validate data
Create new features
Supervised Machine Learning Algorithms
Evaluate Results

Fig 1. Proposed Methodology

CHAPTER 2

LITERATURE SURVEY

The real estate market is one of the most competitive in terms of pricing and same tends to be vary significantly based on lots of factor, forecasting property price is an important modules in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. The literature review give the clear idea and it will serve as the support for the future projects.most of the authors have concluded that artificial neural network have more influence in predicting the but in real world there are other algorithms which should have taken into the consideration. Investors decisions are based on the market trends to reap maximum returns.

Developers are interested to know the future trends for their decision making, this helps to know about the pros and cons and also help to build the project. To accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled. It is inferred that establishing a simple Regression linear mathematical relationship for these time-series data is found not viable for prediction. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and predict future trends. As the real estate is fast developing sector, the analysis and prediction of land prices using mathematical modelling and other techniques is an immediate urgent need for decision making by all those concerned.

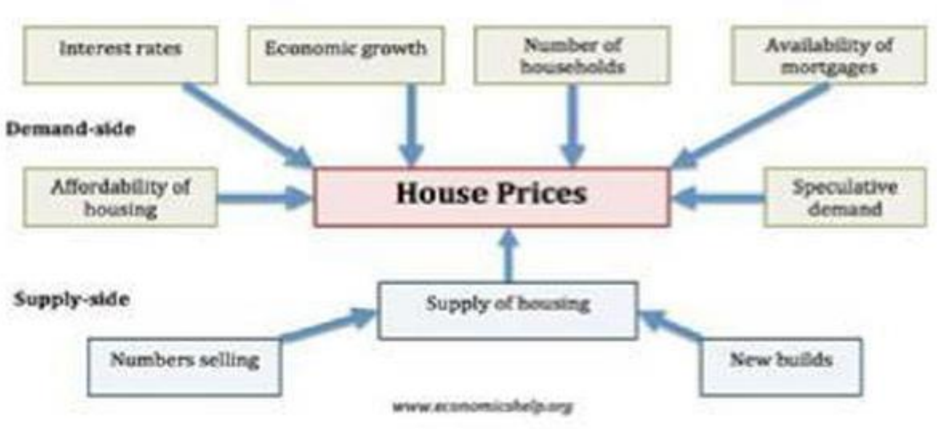


Fig. 1. Factors affecting real estate

CHAPTER 3

METHODOLOGY

A. Data Description

Each record in the database describes a Bangalore suburb or town. The data was drawn from the Bangalore Standard Metropolitan Statistical Area (SMSA) in 2017. The attri butes are defined as follows :-

1. Area type: Area of the Property in which they exist

2. Availability property's status as in 'ready to move' or still under construction.

3. Location: name of locality

4. Size: No. of Bedrooms along with 1 Hall and 1 kitchen.

5. Society: Name of the society

6. total sqft: Area of the Property in square feet

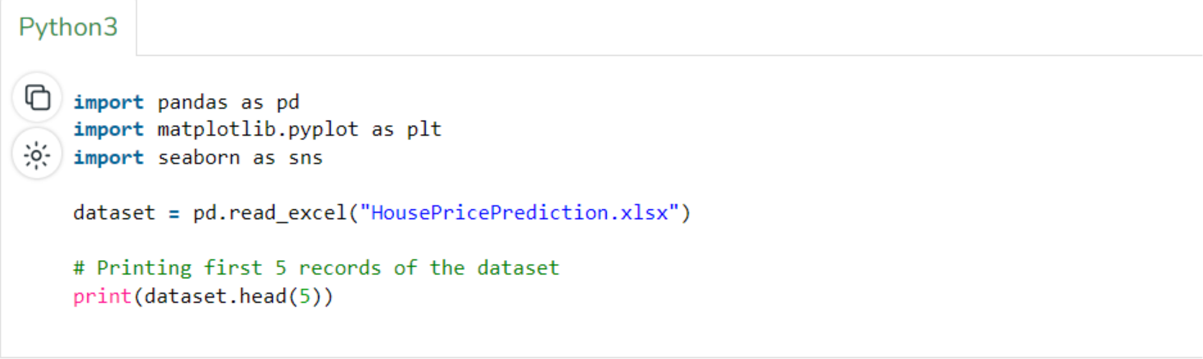
7. Bath: No. of bathrooms

8. Price: price in Inr

Importing Libraries and Dataset

Here we are using

- Pandas – To load the Data frame
- Matplotlib – To visualize the data features i.e. bar plot
- Seaborn – To see the correlation between features using heatmap

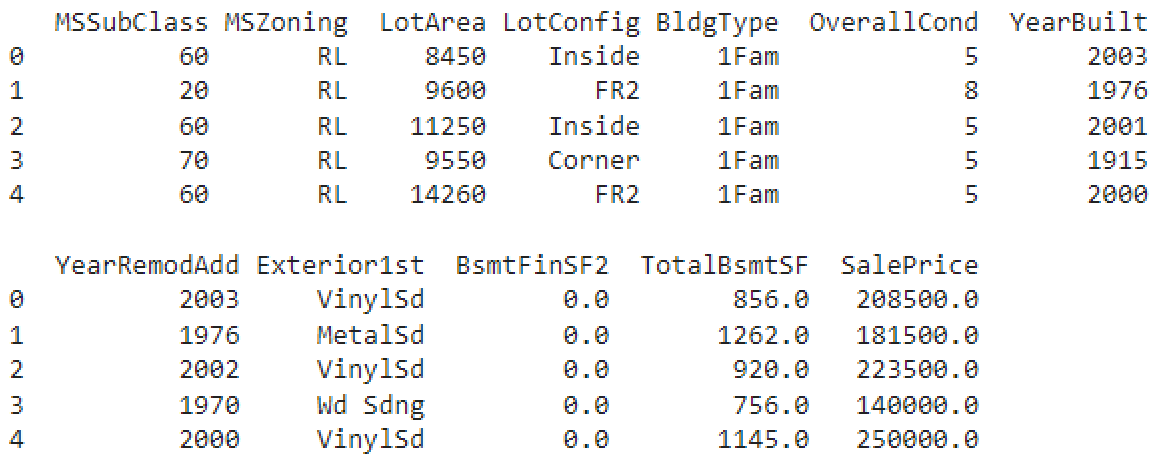


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

dataset = pd.read_excel("HousePricePrediction.xlsx")

# Printing first 5 records of the dataset
print(dataset.head(5))
```

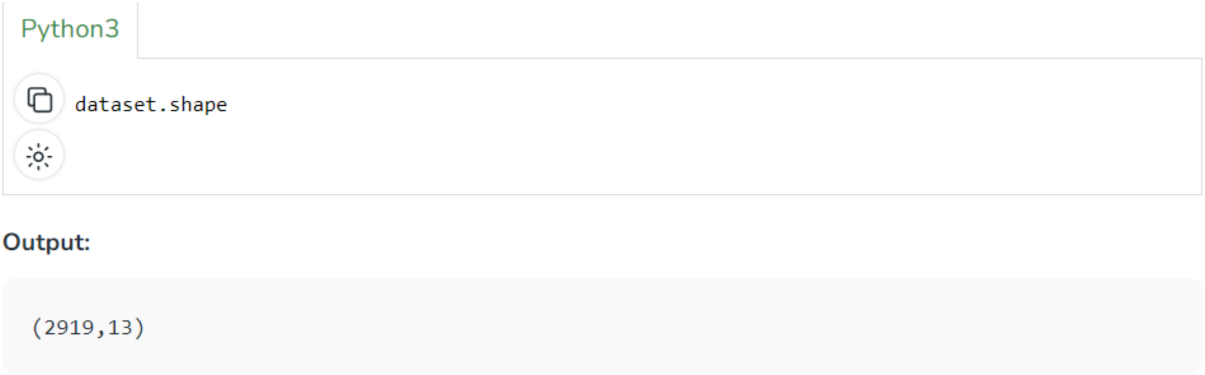
Output:



```
   MSSubClass MSZoning  LotArea LotConfig BldgType  OverallCond  YearBuilt
0          60       RL     8450    Inside     1Fam            5       2003
1          20       RL     9600       FR2     1Fam            8       1976
2          60       RL    11250    Inside     1Fam            5       2001
3          70       RL     9550    Corner     1Fam            5       1915
4          60       RL    14260       FR2     1Fam            5       2000

   YearRemodAdd Exterior1st  BsmtFinSF2  TotalBsmtSF  SalePrice
0          2003     VinylSd         0.0        856.0   208500.0
1          1976     MetalSd         0.0       1262.0   181500.0
2          2002     VinylSd         0.0        920.0   223500.0
3          1970     Wd Sdng         0.0        756.0   140000.0
4          2000     VinylSd         0.0       1145.0   250000.0
```

As we have imported the data. So shape method will show us the dimension of the dataset.



Python3

```
dataset.shape
```

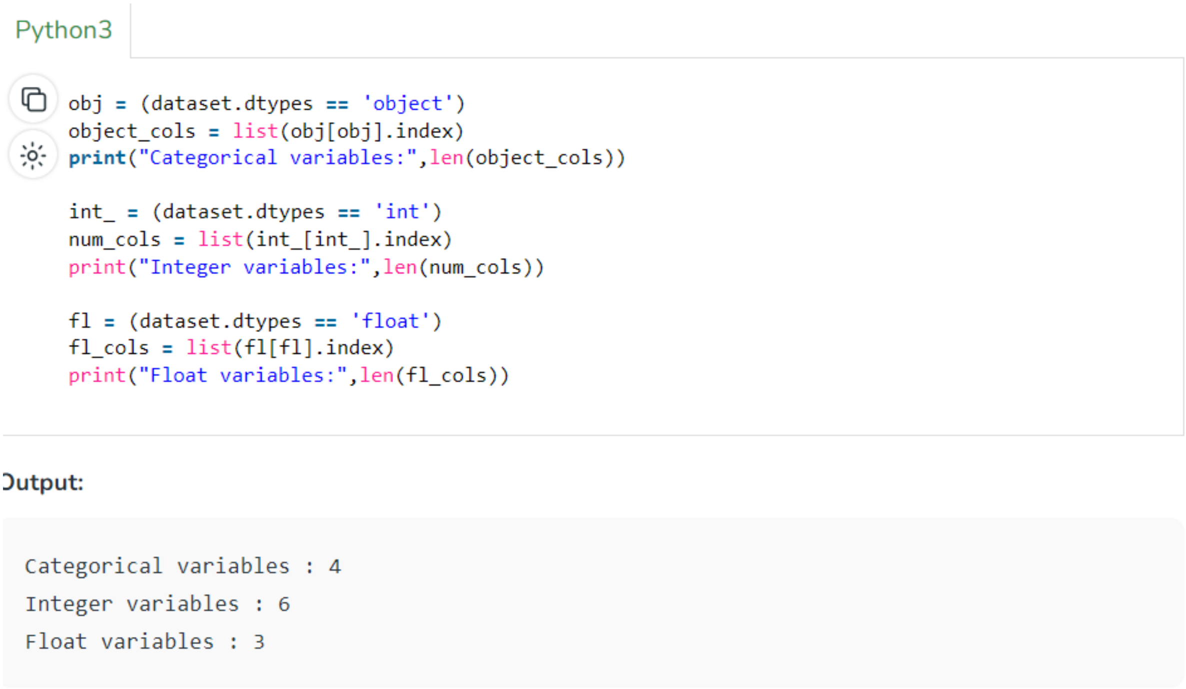
Output:

```
(2919,13)
```

B. Data Preprocessing

Now, we categorize the features depending on their datatype (int, float, object) and then calculate the number of them.

This is the process of transforming the data before it is fed to the algorithm. It is used to convert raw information into a clean data set. This is a information mining strategy that involves transferring raw information into a logical organization Enter raw information in a logical organization. The result of preprocessing data is the final dataset used for preparation and reason for testing.



Python3

```
obj = (dataset.dtypes == 'object')
object_cols = list(obj[obj].index)
print("Categorical variables:",len(object_cols))

int_ = (dataset.dtypes == 'int')
num_cols = list(int_[int_].index)
print("Integer variables:",len(num_cols))

fl = (dataset.dtypes == 'float')
fl_cols = list(fl[fl].index)
print("Float variables:",len(fl_cols))
```

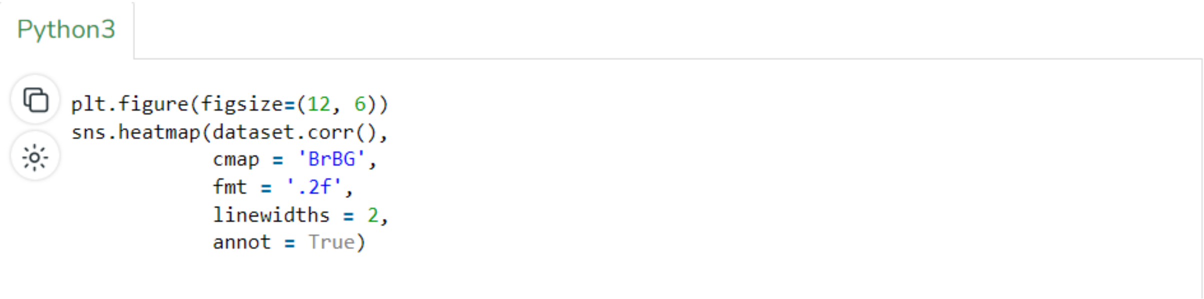
Output:

```
Categorical variables : 4
Integer variables : 6
Float variables : 3
```

C. Exploratory Data Analysis

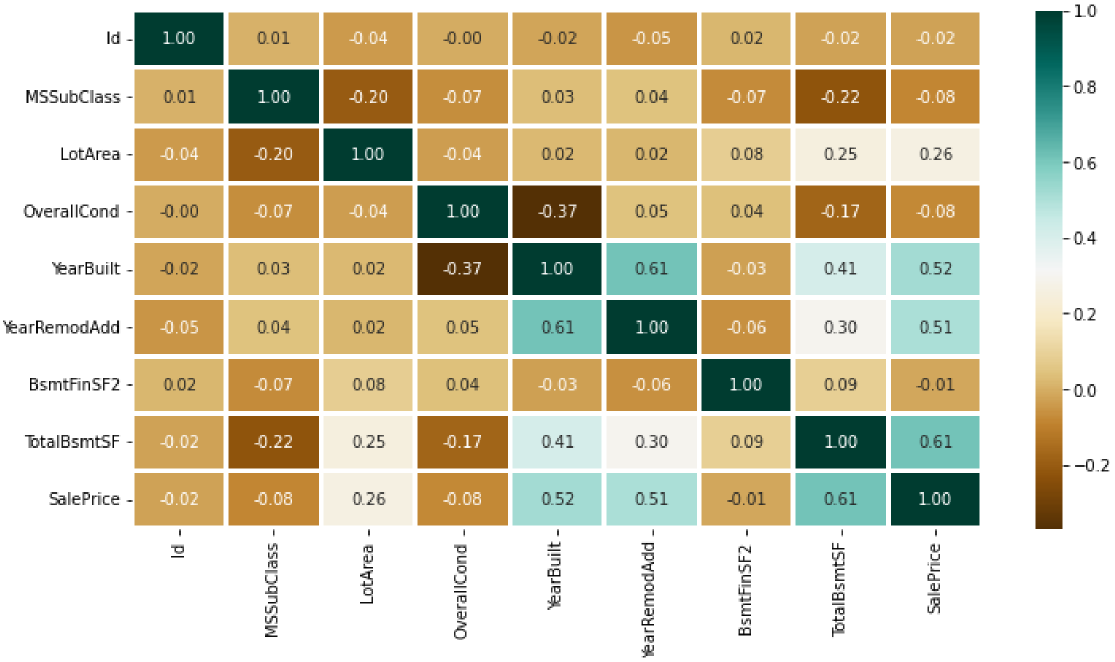
EDA refers to the deep analysis of data so as to discover different patterns and spot anomalies. Before making inferences from data it is essential to examine all your variables. So here let’s make a heatmap using seaborn library.

In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task.

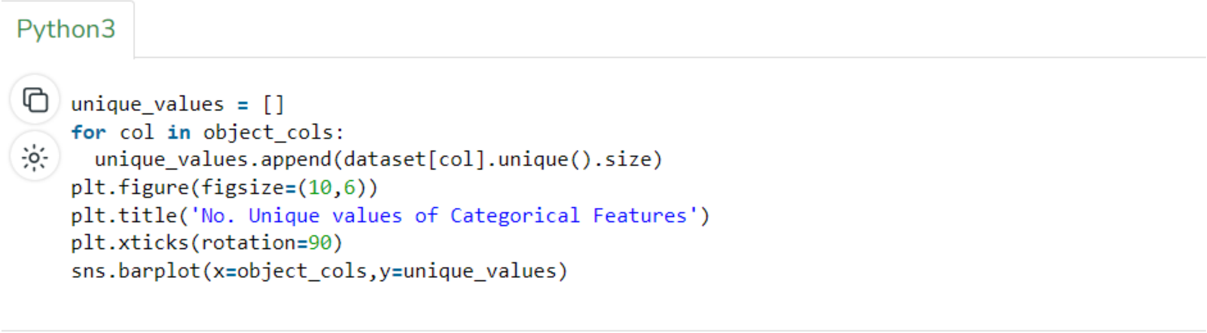


```
plt.figure(figsize=(12, 6))
sns.heatmap(dataset.corr(),
            cmap = 'BrBG',
            fmt = '.2f',
            linewidths = 2,
            annot = True)
```

Output:

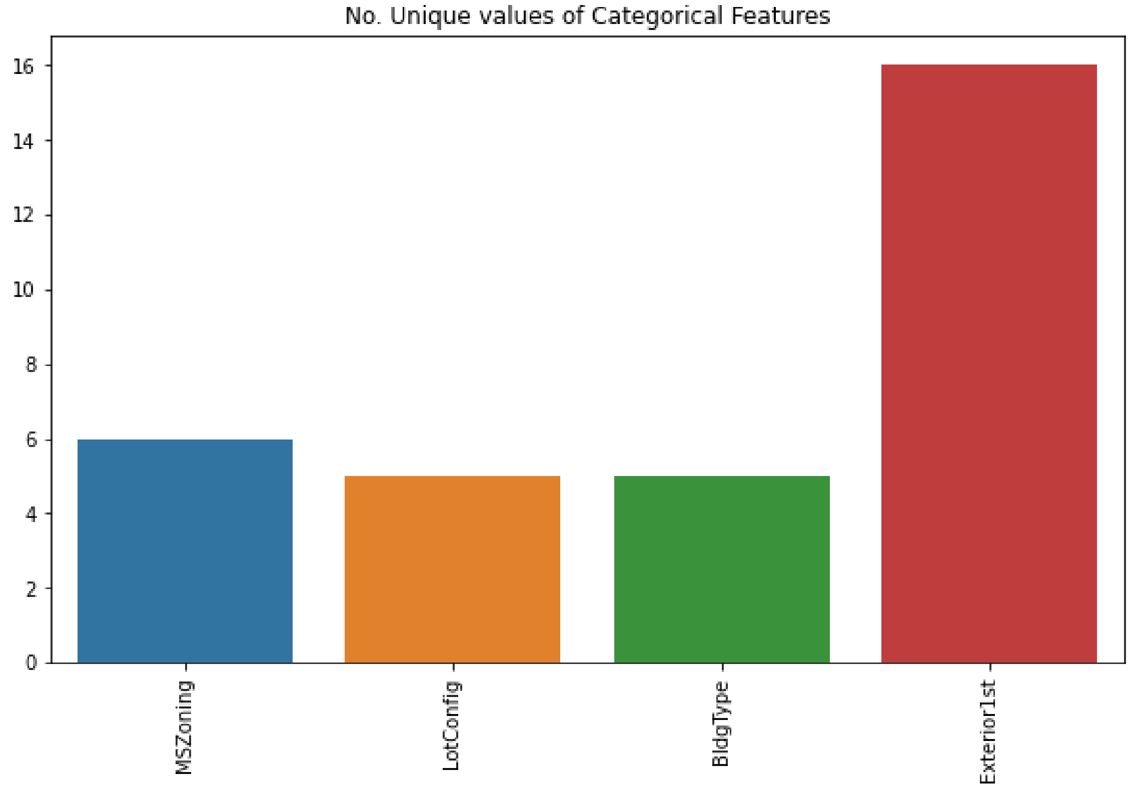


To analyze the different categorical features. Let’s draw the barplot.

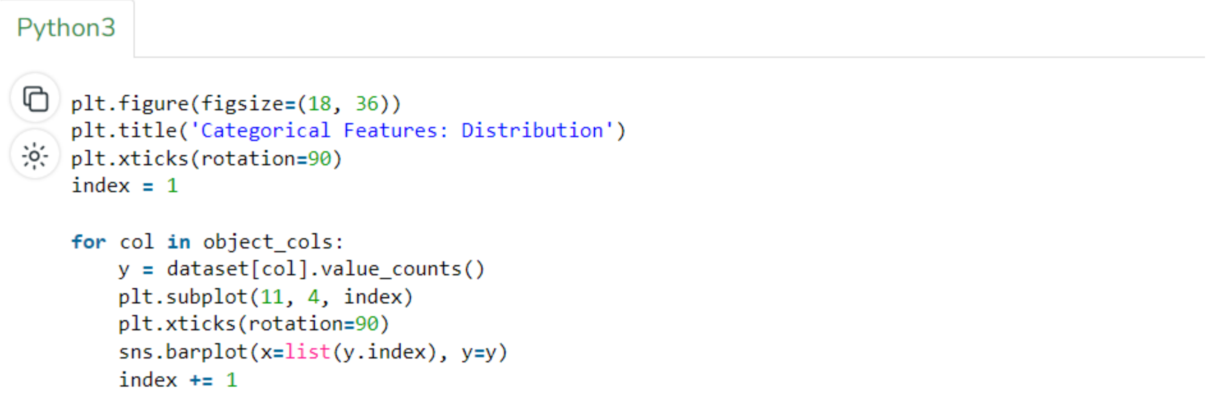


```
unique_values = []
for col in object_cols:
  unique_values.append(dataset[col].unique().size)
plt.figure(figsize=(10,6))
plt.title('No. Unique values of Categorical Features')
plt.xticks(rotation=90)
sns.barplot(x=object_cols,y=unique_values)
```

Output:



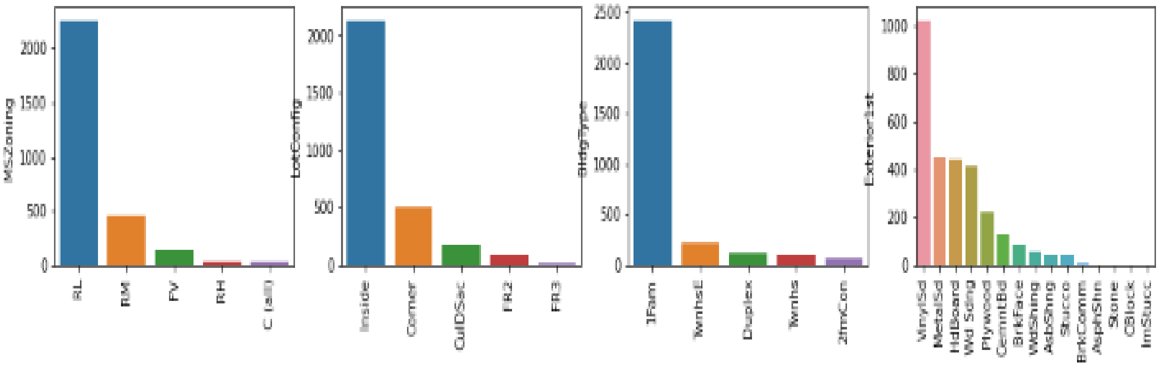
The plot shows that Exterior1st has around 16 unique categories and other features have around 6 unique categories. To find out the actual count of each category we can plot the bar graph of each four features separately.



```
plt.figure(figsize=(18, 36))
plt.title('Categorical Features: Distribution')
plt.xticks(rotation=90)
index = 1

for col in object_cols:
    y = dataset[col].value_counts()
    plt.subplot(11, 4, index)
    plt.xticks(rotation=90)
    sns.barplot(x=list(y.index), y=y)
    index += 1
```

Output:

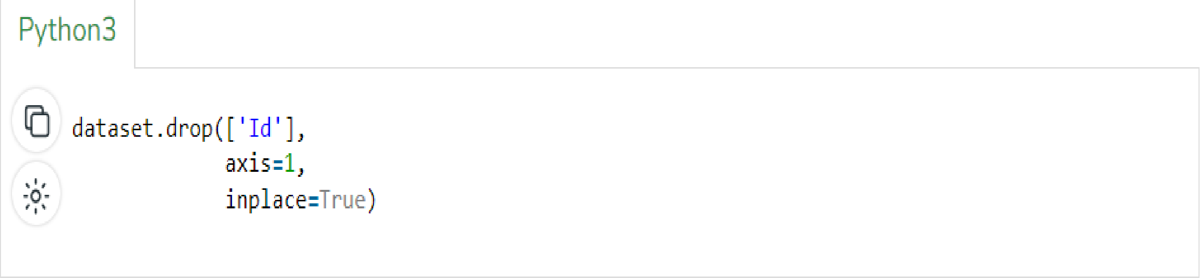


D. Data Cleaning

Data Cleaning is the way to improvise the data or remove incorrect, corrupted or irrelevant data. As in our dataset, there are some columns that are not important and irrelevant for the model training. So, we can drop that column before training. There are 2 approaches to dealing with empty/null values

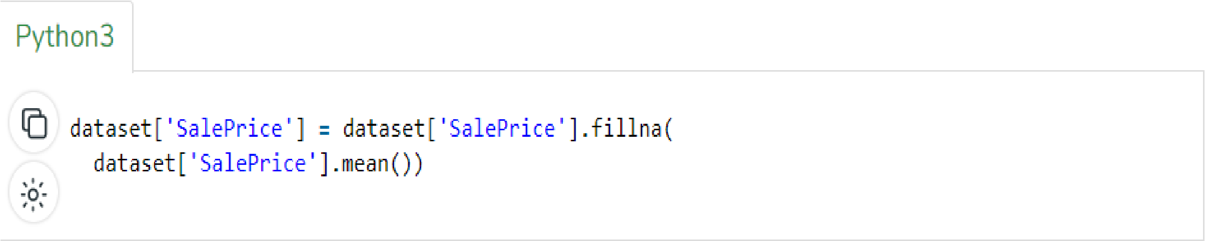
- We can easily delete the column/row (if the feature or record is not much important).
- Filling the empty slots with mean/mode/0/NA/etc. (depending on the dataset requirement).

As Id Column will not be participating in any prediction. So we can Drop it.



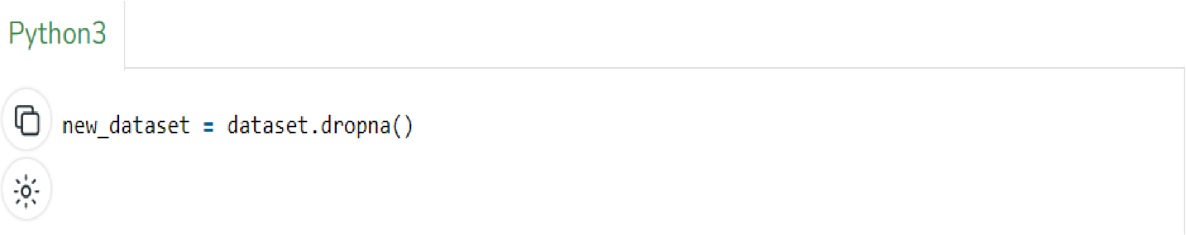
```
dataset.drop(['Id'],
             axis=1,
             inplace=True)
```

Replacing Sale Price empty values with their mean values to make the data distribution symmetric.



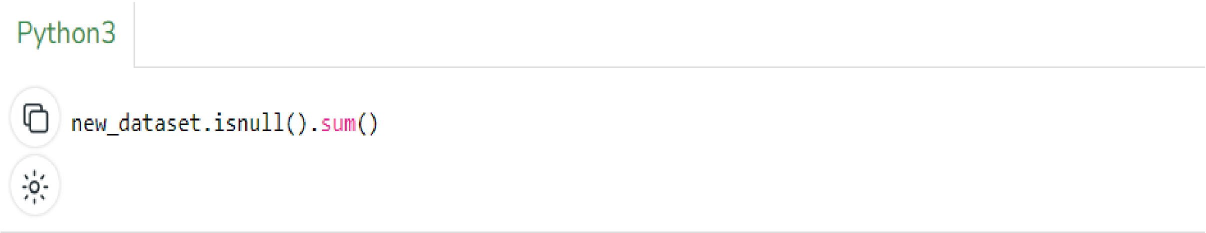
```
dataset['SalePrice'] = dataset['SalePrice'].fillna(
  dataset['SalePrice'].mean())
```

Drop records with null values (as the empty records are very less).



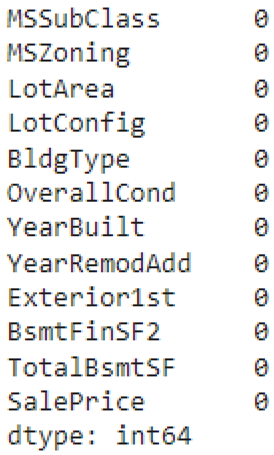
```
new_dataset = dataset.dropna()
```

Checking features which have null values in the new dataframe (if there are still any).



```
new_dataset.isnull().sum()
```

Output:



```
MSSubClass      0
MSZoning        0
LotArea         0
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     0
BsmtFinSF2      0
TotalBsmtSF     0
SalePrice       0
dtype: int64
```

E. ALGORITHM BRIEF OUTLINE

1. Import the python libraries that are required for house price prediction using linear regression. Example: numpy is used for convention of data to 2d or 3d array format which is required for linearregression model, matplotlib for plotting the graph, pandas for readingthe data from source and manipulation that data, etc.

2. First Get the value from source and give it to a data frame and then manipulate this data to required form using head (), indexing, drop ().

3. Next we have to train a model, its always best to spilt the data intotraining data and test data for modelling.

4. Its always good to use shape() to avoid null spaces which will cause error during modelling process.

5. It’s good to normalize the value since the values are in very large quantity for house prices , for this we may use minmax scaler to reducethe gap between prices so that its easy and less time consuming for comparing and values. range usually specified is between 0 to 1 using fittransform.

6. Then we have to make few imports from keras: like sequential for initializing the network lstm to add lstm layer, dropout to prevent overfitting of lstm layers, dense to add a densely connected network layer for output unit.

7. In lstm layer declaration its best to declare the unit, activiation,returnsequence.

8. To compile this model its always best to use adam optimizer and set the loss as required for the specific data.

9. We can fit the model to run for a number of epochs. Epochs are the number of times the learning algorithm will work through the entire training set. 24

10. Then we convert the values back to normal form by using inverse minimal scale by scale factor.

11. Then we give a test data(present data)to the trained model to get the predicted value(future data).

12. Then we can use matplotlib to plot a graph comparing the test andpredicted value to see the increase/decrease rate of values in each time of the year in a particular place. Based on this people will know when its best time to sell or buy a place in a given location.

CHAPTER 4

SYSTEM ANALYSIS

4.1 REQUIRED TOOLS

- 4.1.1 Hardware Requirements

    - Processor – (minmum)i3
    - Hard Disk – 2 GB
    - Memory – 1GB RAM

- 4.1.2 Software Requirements

    - Windows 7(ultimate, enterprise)
    - Visual studio (Latest)
    - Python
    - Jupyter

4.2 EXISTING SYSTEM

There are many existing approaches that can be used to determine the prices of the house,one of them is prediction analysis.

The first approach looks for the time-series data.The time-series approach is to look for the relationship between current prices and prevaling prices.The existing system calculates the price of the house without knowing the information for the future and necessary prediction.This project House Price Prediction helps the people who wanted to buy the house so they can know the price range in the future.

House price prediction also helps the property dealer to know the worth of the property in future.

4.3 PROBLEM FORMULATION

- Price of house/property is linked to our economy, Due to availability of the huge data we do not have the accurate prices.

● Wanted to help the people to know the worth of their owned property in future.

Therefore, the goal of this project is to use machine learning and to predict the selling prices of the houses based on locality and many more economic factors.

4.4 PROPOSED SYSTEM

The main aim or focus of our project is to predict the accurate price of the real estate properties present in India for the next upcoming years through different Algorithms.

A.) Linear Regression

It is a supervised learning technique and responsible for predicting the value of a dependant variable (Y) based on the given independent variable (X).It is the relationship between the input(X) and output(Y) .

B.) Multiple correlation

Analysis It helps to take out the maximum degree of linear relationship that can be obtained between two or more independent variables and a single dependant variable

C.) Classification Trees

Classification Trees are used to predict the objec into classes of a categorial dependant variable based on the one or more predictor variables.

Technology used:-

A.)Data Science:-

Data science is the first stage in which we take the dataset and will do the data cleaning on it. We will do the data cleaning to make sure that it provides the reliable predictions.

B.)Machine Learning:-

The cleaned data is fed into the machine learning model, and we do some of the algorithms like linear regression , regression trees to test out our model.

C.)Front End (UI)

The front end is basically the structure or a build up for a website. In this to receive an information for predicting the price.

It takes the form data entered by the user and executes the function which employee the prediction model to calculate the predicted price for the house.

Chapter 5

Conclusion

The main aim of this project is to determine the prediction of prices.In this paper, we have discovered many algorithms and application of machine learning techniques with the objective of buying the real estate properties and to predict the woth in the future of the owned real estate properties.

Price can be predicted through many factors like the surrounding, marketplaces and many related factors with the house. We have first cleaning and exploring of the input data. The predicted data can be stored in the database and app or website can be made for the people, as people can have the brief idea of the property.

We have performed ensembles of regression trees, k-nearest neighbors, multi-linear regression as we understand that parameterization of the can drive the significant result in the performance.

Clearly, SVM model is giving better accuracy as the mean absolute error is the least among all the other regressor models i.e. 0.18 approx. To get much better results ensemble learning techniques like Bagging and Boosting can also be used.

REFERENCES

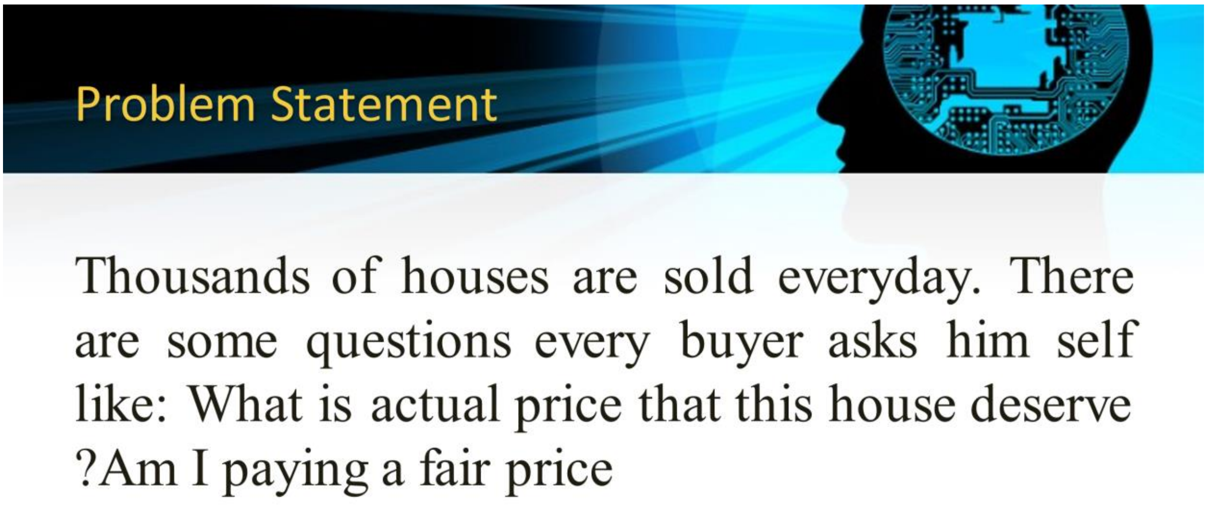
- https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/
- http://103.47.12.35/bitstream/handle/1/9651/BT3083_RPT%20-%20Amit%20Kumar.pdf?sequence=1&isAllowed=y
- https://www.jetir.org/papers/JETIR2204579.pdf
- file:///C:/Users/swapnil%20sunil%20herage/Downloads/internshipe%20report%20hous%20eprice%20prediction.pdf

"House Price Prediction Using Artificial Intelligence"

Presentation by:

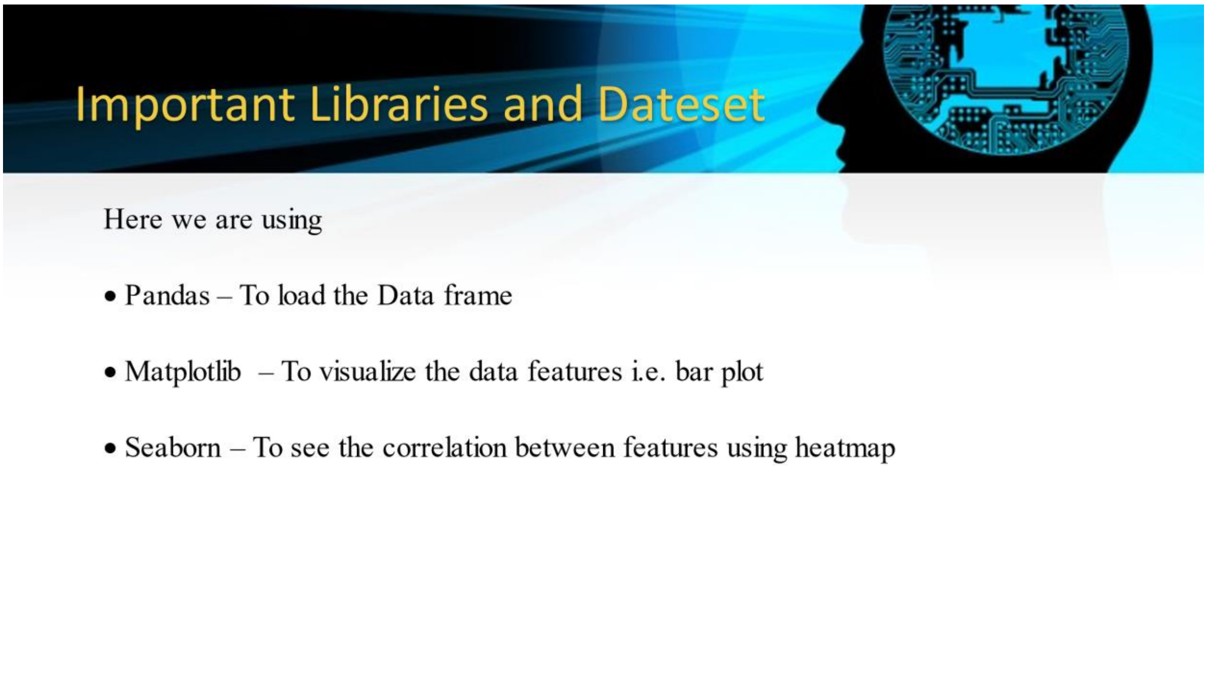
SWAPNIL SUNIL HERAGE

USN: 21BBTCS241



Problem Statement

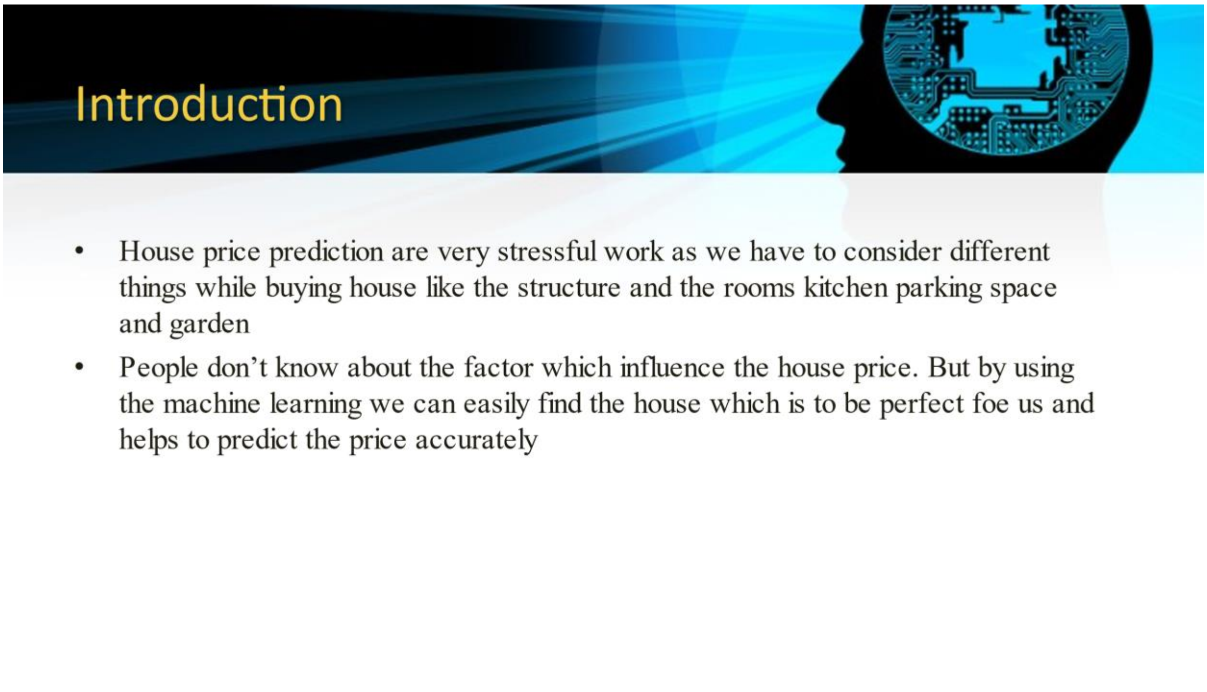
Thousands of houses are sold everyday. There are some questions every buyer asks him self like: What is actual price that this house deserve ?Am I paying a fair price



Important Libraries and Dateset

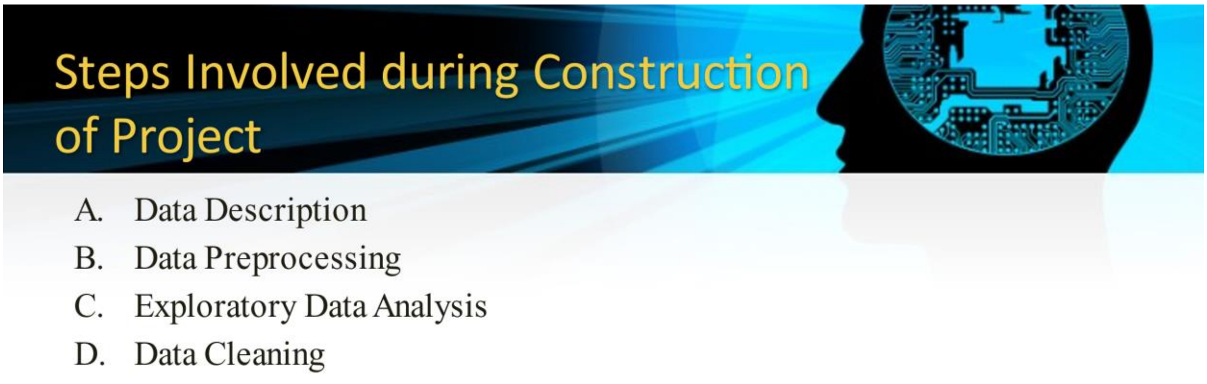
Here we are using

- Pandas – To load the Data frame
- Matplotlib – To visualize the data features i.e. bar plot
- Seaborn – To see the correlation between features using heatmap



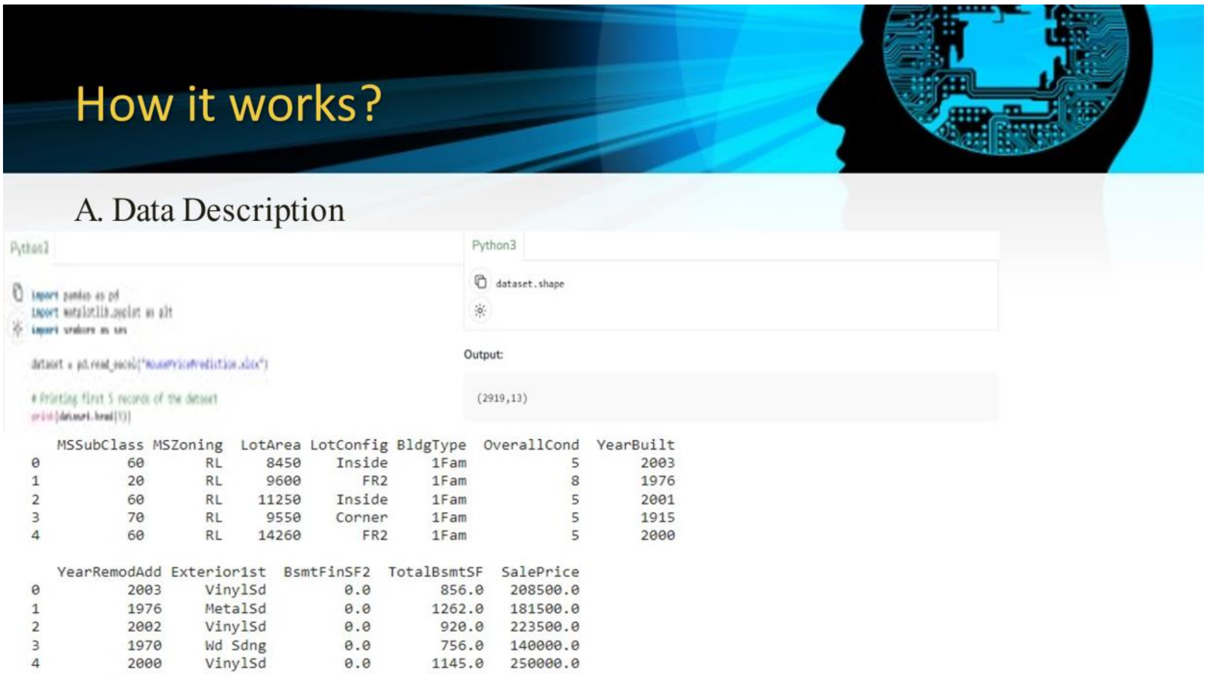
Introduction

- House price prediction are very stressful work as we have to consider different things while buying house like the structure and the rooms kitchen parking space and garden
- People don’t know about the factor which influence the house price. But by using the machine learning we can easily find the house which is to be perfect foe us and helps to predict the price accurately



Steps Involved during Construction of Project

A. Data Description
B. Data Preprocessing
C. Exploratory Data Analysis
D. Data Cleaning



How it works?

A. Data Description

Python3

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

dataset = pd.read_excel("HousePricePrediction.xlsx")

# Printing first 5 records of the dataset
print(dataset.head(5))
```

Python3

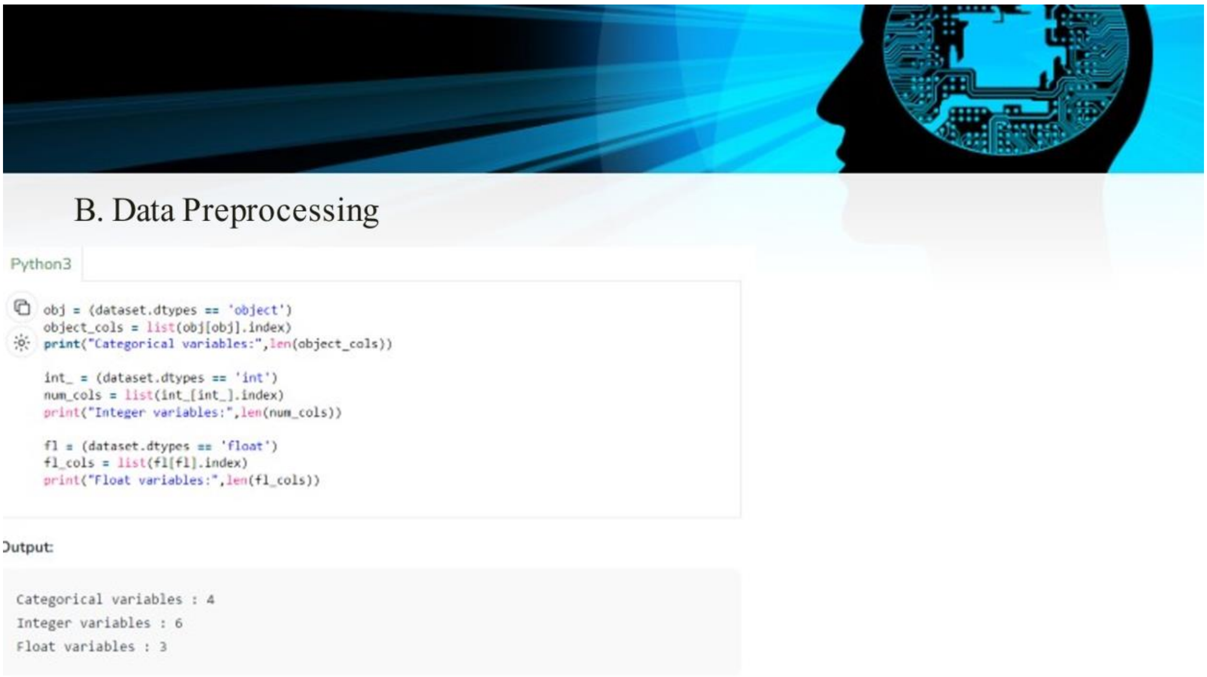
```
dataset.shape
```

Output:

```
(2919,13)
```

```
   MSSubClass MSZoning  LotArea LotConfig BldgType  OverallCond  YearBuilt
0          60       RL     8450    Inside     1Fam            5       2003
1          20       RL     9600       FR2     1Fam            8       1976
2          60       RL    11250    Inside     1Fam            5       2001
3          70       RL     9550    Corner     1Fam            5       1915
4          60       RL    14260       FR2     1Fam            5       2000

   YearRemodAdd Exterior1st  BsmtFinSF2  TotalBsmtSF  SalePrice
0          2003     VinylSd         0.0        856.0   208500.0
1          1976     MetalSd         0.0       1262.0   181500.0
2          2002     VinylSd         0.0        920.0   223500.0
3          1970     Wd Sdng         0.0        756.0   140000.0
4          2000     VinylSd         0.0       1145.0   250000.0
```



B. Data Preprocessing

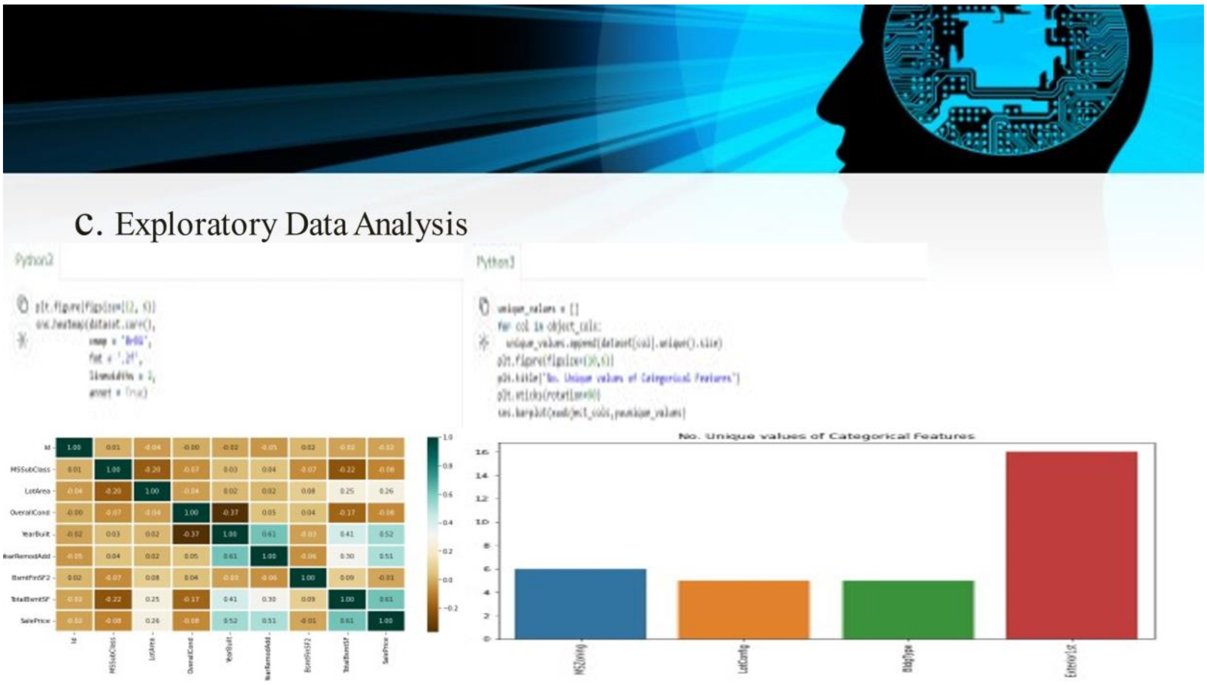
```
obj = (dataset.dtypes == 'object')
object_cols = list(obj[obj].index)
print("Categorical variables:",len(object_cols))

int_ = (dataset.dtypes == 'int')
num_cols = list(int_[int_].index)
print("Integer variables:",len(num_cols))

fl = (dataset.dtypes == 'float')
fl_cols = list(fl[fl].index)
print("Float variables:",len(fl_cols))
```

Output:

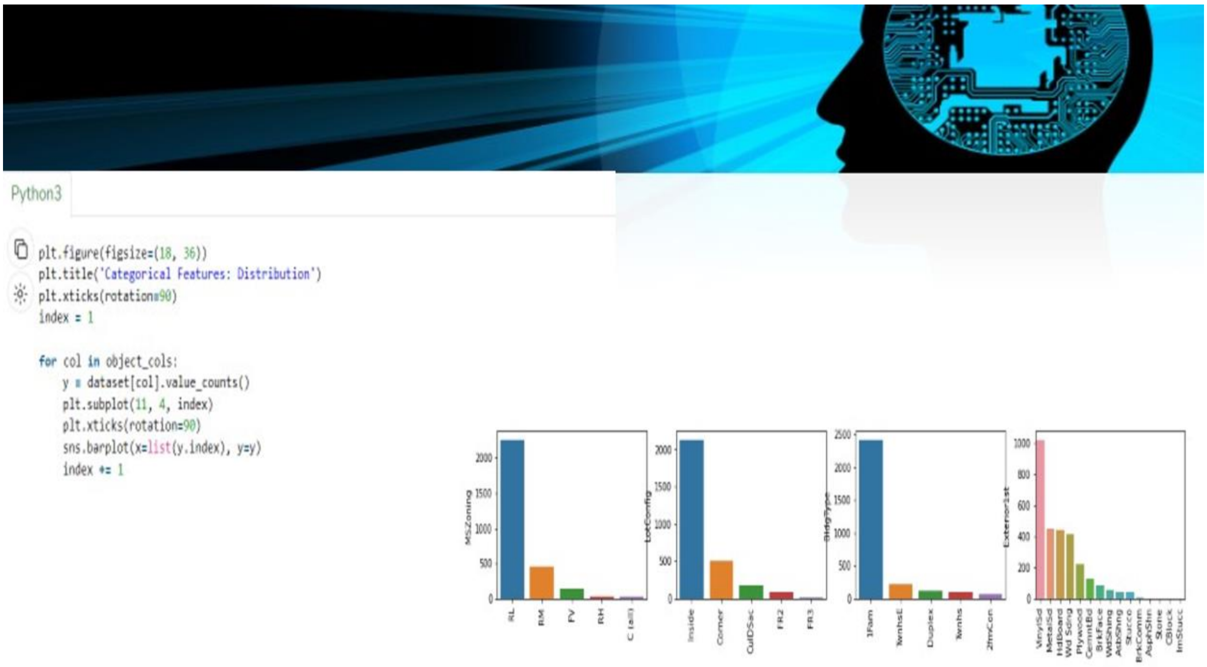
```
Categorical variables : 4
Integer variables : 6
Float variables : 3
```



C. Exploratory Data Analysis

```
plt.figure(figsize=(12, 6))
sns.heatmap(dataset.corr(),
            cmap = 'BrBG',
            fmt = '.2f',
            linewidths = 2,
            annot = True)
```

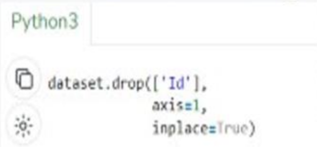
```
unique_values = []
for col in object_cols:
  unique_values.append(dataset[col].unique().size)
plt.figure(figsize=(10,6))
plt.title('No. Unique values of Categorical Features')
plt.xticks(rotation=90)
sns.barplot(x=object_cols,y=unique_values)
```



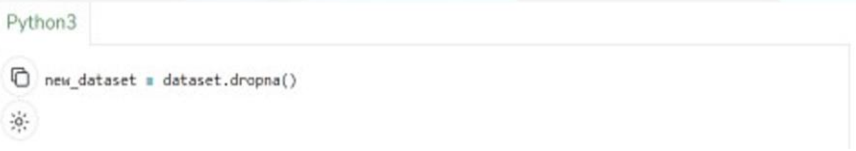
```
plt.figure(figsize=(18, 36))
plt.title('Categorical Features: Distribution')
plt.xticks(rotation=90)
index = 1

for col in object_cols:
    y = dataset[col].value_counts()
    plt.subplot(11, 4, index)
    plt.xticks(rotation=90)
    sns.barplot(x=list(y.index), y=y)
    index += 1
```

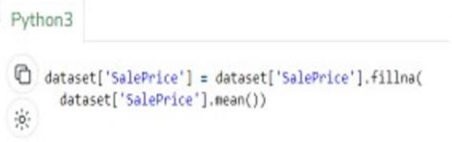
D. Data Cleaning



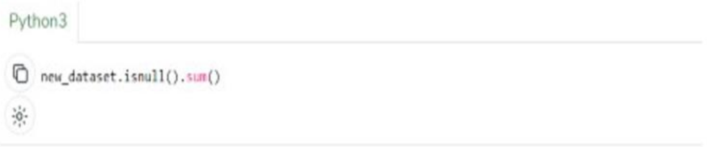
```
dataset.drop(['Id'],
             axis=1,
             inplace=True)
```



```
new_dataset = dataset.dropna()
```



```
dataset['SalePrice'] = dataset['SalePrice'].fillna(
  dataset['SalePrice'].mean())
```



```
new_dataset.isnull().sum()
```

THANK
YOU