**Real Estate Price Prediction Model**

Minor project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology in

**Computer Science and Engineering**

By

Ananya Mishra (191213)

Prishita Singh (191223)

**Under the Supervision of**

Dr. Jagpreet Sidhu



Department of Computer Science & Engineering and Information Technology

**Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA**

**Table of Contents**

| **Title** | **Page No.** |
| --- | --- |
| DECLARATION | i |
| CERTIFICATE | ii |
| ACKNOWLEDGEMENT | iii |
| ABSTRACT | iv |
| **CHAPTER 1: Introduction** | **1** |
| 1.1 Introduction | 1 |
| 1.2 Objectives | 1 |
| 1.3 Motivation of the Project | 2 |
| 1.4 Language Used | 2 |
| 1.5 Technologies Used | 2 |
| 1.6 Deliverables of the Minor Project | 3 |
| **CHAPTER 2: Feasibility Study, Requirements Analysis and Design** | **4** |
| 2.1 Feasibility Study | 4 |
| 2.2 Requirements on Minor Project | 5 |
| 2.3 Use Case Diagram of Minor Project | 6 |
| 2.4 DFD Diagram | 6 |
| 2.5 State Transition Diagram | 7 |
| **CHAPTER 3: Implementation** | **8** |
| 3.1 Data Set used in the Project | 8 |
| 3.2 Dataset Features | 9 |
| 3.3 Design of Problem Statement | 10 |
| 3.4 Algorithm/Pseudo Code of the Project | 10 |
| 3.5 Flow Graph of the Project | 11 |
| 3.6 Screenshots of the various stages of the Project | 11 |
| **CHAPTER 4: Results** | **30** |
| 4.1 Discussion on the Results | 30 |
| 4.2 Applications of the Project | 30 |
| 4.3 Limitations of the Project | 31 |
| 4.4 Future Work | 31 |
| **CONCLUSION** | **31** |
| **REFERENCES** | **32** |

**Declaration**

We declare that the work supplied in the project report titled "Real Estate Price Prediction and Sales Insights Analyser" presented by Ananya Mishra (191213) and Prishita Singh(191223) in partial fulfillment of the requirements for the grant of the degree of Bachelor of Technology in Computer Science and Engineering is an authentic record of our own work carried out during a period from January 2022 to May 2022 under the supervision of **Dr. Jagpreet Sidhu** (Assistant Professor(SG) – CSE & IT).

We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

**Supervised by**

**Dr. Jagpreet Sidhu**

Assistant Professor (SG)

Department of Computer Science and Engineering & Information Technology

**Submitted by**

Ananya Mishra Prishita Singh

191213 191223

i

**Certificate**

This is to confirm that the work provided in the project report titled "Real Estate Price Prediction Model" is in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering filed in the Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an accurate record of our own work done under the supervision of Dr. Jagpreet Sidhu (Assistant Professor – CSE & IT) from January to May 2022

**Ananya Mishra Prishita Singh**

**191213 191223**

The above statement is correct to the best of my knowledge.

**Dr. Jagpreet Sidhu**

**Assistant Professor(SG)**

**Department of Computer Science and Engineering & Information Technology**

ii

**Acknowledgement**

Firstly, I express my heartiest thanks and gratefulness to Almighty God for His divine blessing that made it possible for us to complete the project work successfully.

I am really grateful and wish my profound indebtedness to **Dr. Jagpreet Sidhu,** Assistant Professor(SG), Department of CSE Jaypee University of Information Technology, Wakhnaghat. His deep knowledge & keen interest in the field of Data Science helped us to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would also generously thank each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. I would also like to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

**Ananya Mishra(191213)**

**Prishita Singh(191223)**

iii

**Abstract**

The project “Real Estate Price Prediction Model” is a prediction and analysis model that predicts the price of a flat or building in the city of Bangalore and analyses the sales. The project uses the techniques of machine learning to build the model. The project shall be presented to the customer in the form of a web page offering the selection constraints like the locality, number of bedrooms, and sq. ft. area.

The customer is made to select the specifications of his future flat and as per the entire selection list, the model predicts the price of the flat.

The project is a prediction and analysis model. It uses the techniques of machine learning and data analysis.

iv

**Chapter 1: Introduction**

**1.1 Introduction**

“Real Estate Price Prediction Model” is a consumer-friendly website that will help the user to predict the price of real estate in the city of Bangalore depending on the specifications of the area, location, and other constraints.

The model is designed keeping in mind the business scenario and the demand of the market. The consumer would be able to choose his flat or building using the website and when the sales are conducted, the property dealer shall be able to analyze the sales himself without making consistent calls to his data analyst every third day.

The prediction model is built using machine learning techniques and the mathematical skills of regression.

**1.2 Objectives**

The objectives of the project are discussed below:

i. Implementing our knowledge of the basic concepts of machine learning and data science into a project

ii. Creating a consumer-friendly and convenient utility of the model in the form of a webpage

iii. Creating an easy-to-handle platform for the property dealers to analyze the sales insights, since they might not have any prior technical background

1

**1.3 Motivation of the Minor Project**

The model is designed keeping in mind the business scenario and the demand of the market. The customers demand ease of searching the desired location with their other specifications and wish to know how much would it cost if they purchase the property within an area. The model offers a convenient and efficient way of real estate price prediction to the customers. Moreover, many times the property dealers do not have a technical background and hire a data scientist to deal with the analysis of the sales. Hence, this model shall help the property dealers to analyze the sales themselves without making consistent calls to the data analyst every third day and also to keep a check on the authenticity of his work.

**1.4 Language Used**

The machine learning model is designed using python due to the availability of numerous machine learning libraries and methods.

**1.5 Technologies Used**

Following are the technologies and software used in the project:

1.Python PyCharm

2.Anaconda

3.NumPy and Pandas (Data Cleaning)

4.Matplotlib (Data Visualization)

5.Sklearn for model building

6.Python flask for http server

7.Notepad++

8.HTML/CSS/Javascript for UI

9.MySQL server

2

**1.6 Deliverables of the Minor Project**

Designing the regression model and using the techniques of web development we shall finally have a web page ready for the consumer to predict the price of his desired flat or mansion in a specific locality in the city of Bangalore. The outcomes of the regression model accuracy shall be visualized and displayed.

**Chapter 2: Minor Project SDLC**

**2.1 Feasibility Study :**To evaluate the project’s potential, a feasibility analysis over different constraints is put forward.

i. Technical Feasibility: The project “Real Estate Price Prediction Model” is a technically feasible model. Using the specific technologies, the model shall work up to the user’s expectations. The model has created a consumer-friendly UI and uses Anaconda which is the most feasible software of the regression models. All the technologies used are stable.

ii. Operational Feasibility: The project is an operational feasible model and would benefit the customers and the sellers at the same time. It is a user-friendly model. The model is easy to operate and functions to display the authentic and true outcomes.

iii. Scheduling Feasibility: The project has been completed as per the schedule and meets the constraints of the schedule feasibility for a model.

iv. Market Feasibility: The project has particularly been designed to meet the market demands of an easy and manageable platform for the customer to explore the real estate and the dealers to analyze their sales timely. It meets the constraints of market feasibility.

v. Economic Feasibility: The software used in the making of the project is free of cost and does not involve any expenses. The website is presently run on a localhost server, in order to publicly make it available in the cyber market we would need to buy the domain for the website, which is an economically feasible expense.

4

**2.2 Requirements on Minor Project**

**2.2.1 Functional Requirements**

The requirements which are fundamentally required and the end-user particularly demands from the system[1].

The functional requirements of the project are listed below:

i. The consumer has a convenient interface to operate.

ii. The property dealer has his system under authentication privacy and password protection.

iii. Once the user/customer fills in the specifications of the flat then the webpage offers a button to submit the information details and predict the price.

**2.2.2 Non-Functional Requirements**

The requirements of the project are not mandatory but enhance the quality of the project[1].

A few non-functional requirements of the project are shared below:

i. The credentials of the customer are not asked or shared before he starts his dealings with the property dealer.

iii. The web page loads and displays the predicted price within a time span of 4-5 seconds.

iii. The model is security feasible taking into consideration the password protection for the property dealers.

5

**2.3 Use Case Diagram of the Minor Project**

To graphically depict the interaction of the customer/user with the UI by filling the specifications of the flat and predicting the price, the case diagram is shown below. Additionally, the dealer’s interaction with his sales analyser is also presented. The SQL file will contain the sales data and it is used for analysis.

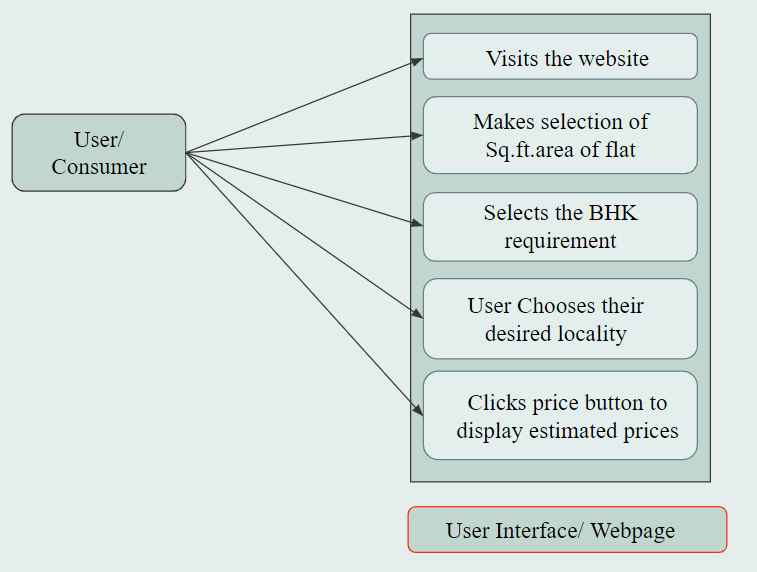


Fig. 1: Use Case Diagram of the project

**2.4 DFD Diagram**

The Diagram is the visual representation of the data flow in the model. The diagram is the physical and logical flow of data in our model from the user making the server request to predicting the price of the real estate.

6

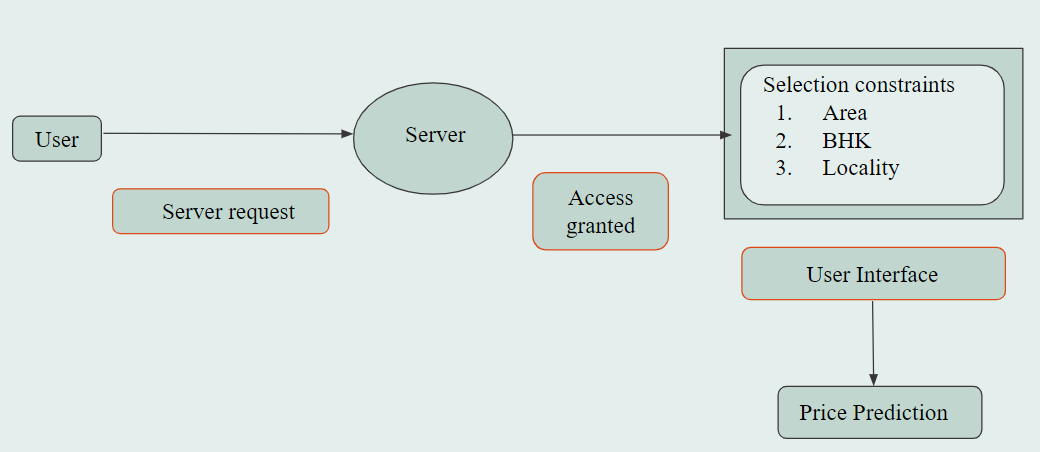


Fig. 2: Data Flow Diagram of the model

**2.5 State Transition Diagram**

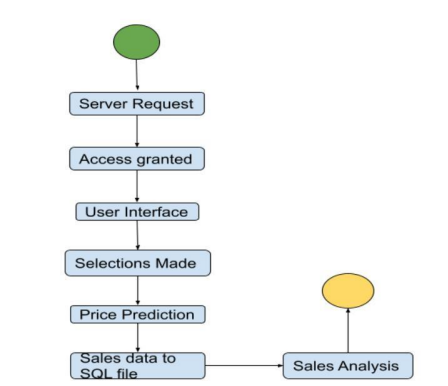
The state diagram below depicts the various states of the model and the simple transitions between them. The diagram shows the behavioral model of how one state stimulates according to the previous states.

Fig. 3: State Transition Diagram of the model

7

**Chapter 3: Implementation**

**3.1 Data Set used in the Minor Project**

The data set used for building the regression model for the price prediction of the real estate has been imported from Kaggle. The data set is the data of Bangalore city, India. The data set is a CSV file. It consists of various fields and the data size is large which makes it easy to handle and make operations on the data.[2]

The various column fields of the data set are area type, locality, number of bedrooms, total sq. ft. area, price, etc. This data set is cleaned and thereafter used for building the regression model.[3]

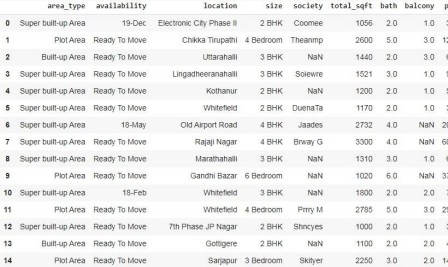


Fig. 4: Dataset used in the model

**3.2 Data Set Features**

**3.2.1 Types of Dataset**

8

The dataset used in the project is a combination of various categories:

i. Numerical Data Set: The data set contains various numerical fields for the evaluation, hence the dataset used is numerical data.

ii. Multivariate Data Set: The data set consists of various independent fields which collectively determine the price of real estate.

iii. Correlation Data Set: The data set contains the two fields namely, total\_sqft and size, which depend on each other. The data is unstructured so there might exist uncorrelated rows, which will be structured to display correlation after data cleaning.

**3.2.2 Number of attributes, fields, description of data set**

The attributes in our data set can be categorized into the following types:

i. Quantitative attributes: The data set has

1. Numeric attributes: In the data set the attributes namely, total\_sqft, bath, etc. are the numeric attributes.

2. Discrete attributes: The data set consists of attributes like bath, balcony, etc. have discrete values.

3. Continuous attributes: In the data set, certain attributes like price acquire continuous values.

ii. Qualitative attributes: The data set that we have used only has a single category of qualitative attributes, i.e., nominal attributes. In the data set the attributes like area\_type, and location has nominal values (related to names).

The unstructured dataset of our project contains 9 fields/attributes.

Please note in the database, the field refers to the data member and attribute is another term reference for the field to be used publicly.

The data set is the data collection of the real estate of Bangalore city, India. The dataset has been deployed from Kaggle. The dataset initially is unstructured and contains plenty of fields. The size of the dataset is nearly 13000 rows. The large size and variety of fields of the dataset help in the evaluation and better data cleaning. The dataset is cleaned and used to build the regression model for the price prediction of the real estate in Bangalore.

9

**3.3 Design of Problem Statement**

To build a model and design a platform that helps to predict the price of the real estate. The model additionally must help the property dealer to analyze the sales in a visualized manner.

**3.4 Algorithm/Pseudocode of the model**

The project is a regression-based model and the analysis model is a SQL based model. To design the regression model is not a tedious task. The most important part is the data cleaning of the dataset. The dataset is unstructured data deployed randomly from an online platform.

1. The dataset is read in the ipynb notebook.

2. The data is cleaned. (*data cleaning shall be discussed in the stages of the project since it is not part of the algorithm of the model*)

3. Building of the regression model - The regression model is built using the sklearn train test split.

X = dataframe.drop(['price'],axis='columns') #x is all the independent variables y = df12.price

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.2,random\_state=10)

4. Linear Regression is used to design the price prediction model[4][5].

from sklearn.linear\_model import LinearRegression

lr\_clf = LinearRegression()

5. The efficiency of the model is evaluated after fitting in the model.[5]

lr\_clf.fit(X\_train,y\_train)

lr\_clf.score(X\_test,y\_test)

6. This model is then tested on certain specific data.

7. The tested model is exported to a pickle file that shall further help the flask server to generate the model request.

8. The python flask server imports the tested model file and creates a server request to display it on http online portal.

10

9. The web page is designed using the web development techniques (HTML, CSS, Java script) in order to offer the customer a better User Interface to explore.

**3.5 Flow Graph of the model/project**

The flow graph of the project shows the various stages of the project.

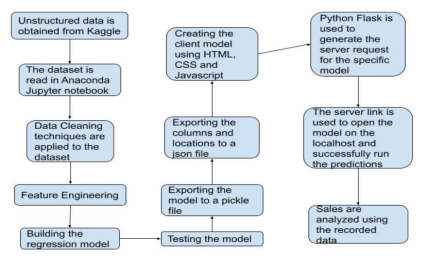
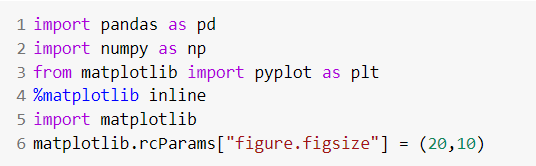


Fig. 5: Flow diagram of the model

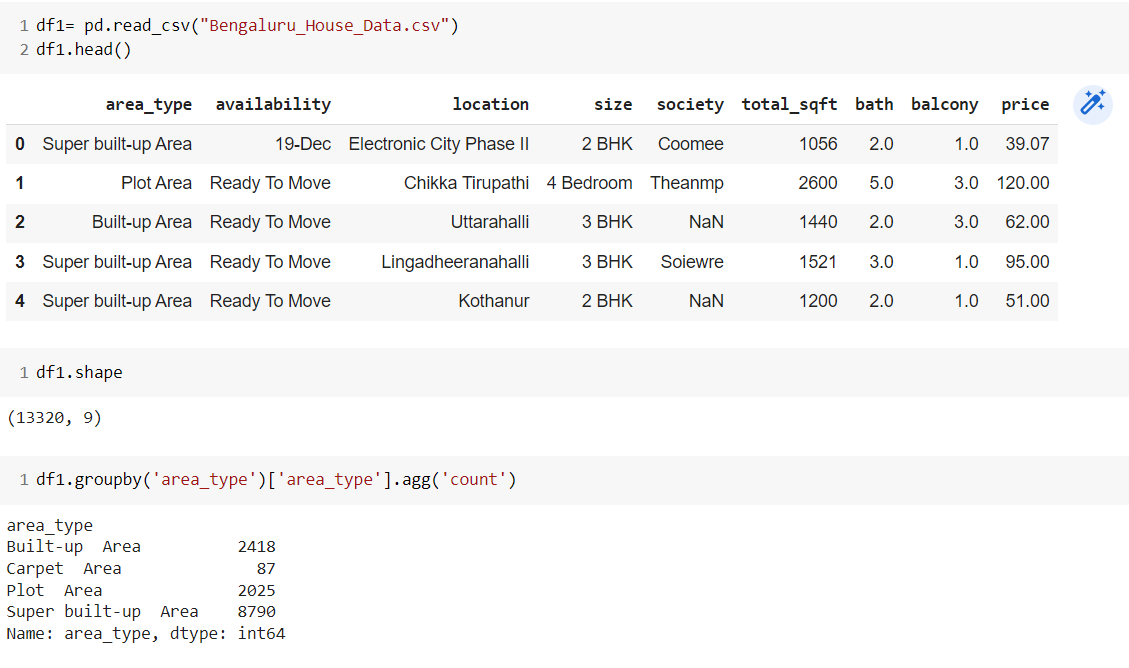
**3.6 Screenshots of the various stages of the project**

**1. Libraries are imported**

****

11

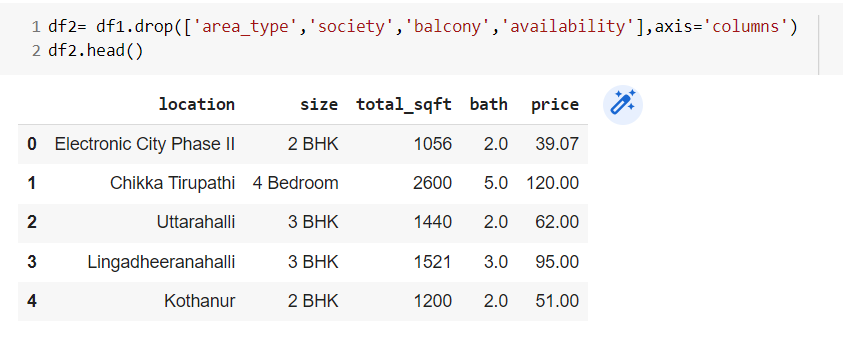
**2. The data is loaded into model**

****

**3. Data Cleaning**

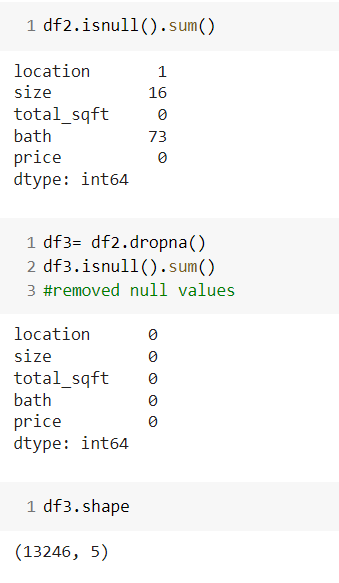
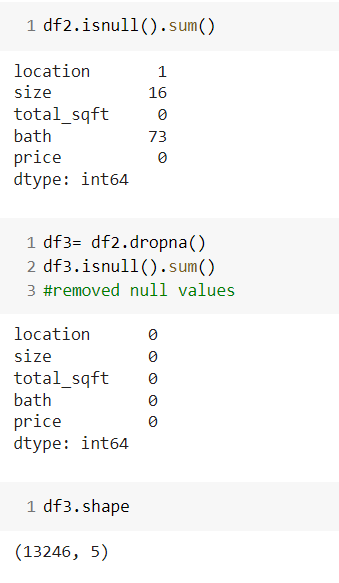
This is the most important part of model designing.

**3.1** The features that are not required are dropped

****

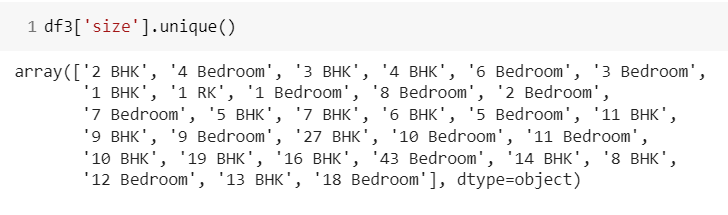
**3.2** Check for the null values and drop them.

12

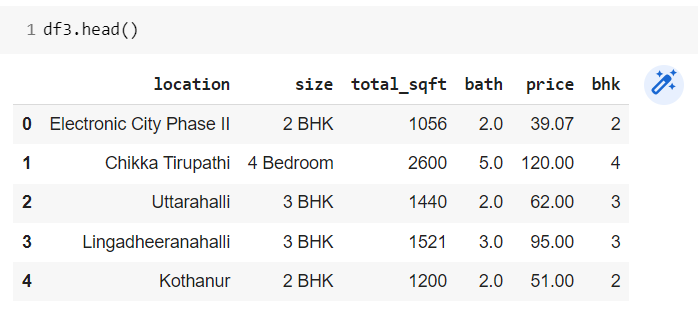
** **

(Checking for null values) (dropping null values)

**3.3** We observe a unique pattern in the ‘size’ column of our dataset.

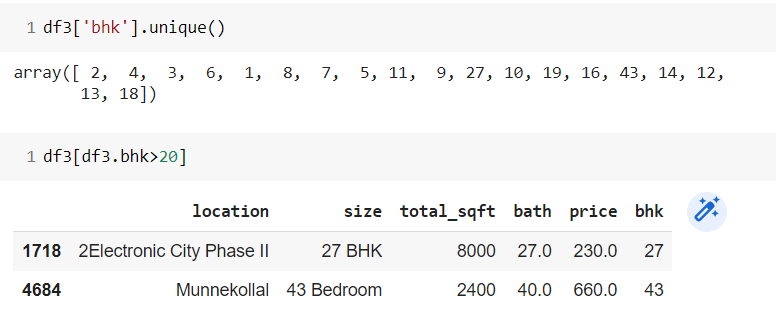


The field contains attribute values with varying units labeled as ‘BHK’, and ‘Bedroom’, and the machine learning algorithms shall consider them as different entities. So this needs to be worked on. We shall work on the technique of “**Feature Engineering**”. We would add a new feature ‘BHK’ which will replace this field.

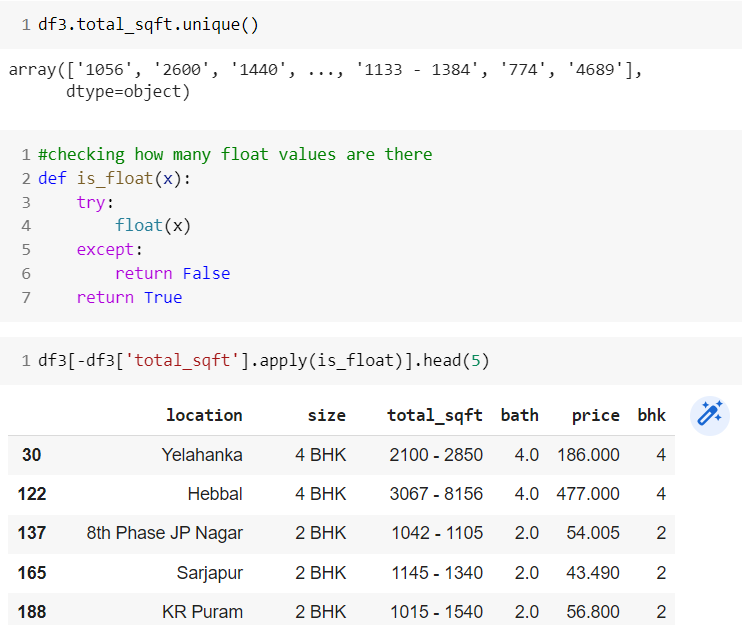


13

After this feature has been added to our dataset, we need to improve it more. We made yet another observation, the feature field ‘BHK’ contains certain values more than 20, which is practically not a feasible flat to be purchased by a normal living family.



Secondly, there were certain values displayed in the form of intervals and it is very clear that range values are not accepted. The dataset demands accurate numeric values in fields like “BHK”.



14

**3.4** The next observation was made in the field of “total\_sqft”, it contains certain range values and a few values with different units. Since we have a large size of data, we ignore the rows with the different units since the number of such values is very low. However, the range values need a solution. We prefer to take the middle value of the interval in the case of range values.



Finally, this is what we obtain.



**3.5** Feature Engineering

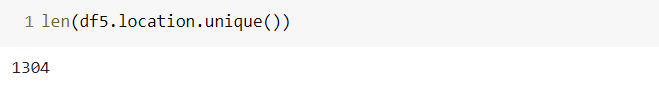
We will introduce another feature “price\_per\_sqft”. The major reason behind this feature engineering technique is to make the predictions better and fast. It evaluates the influencing constraints and predicts the price using the price per sq. ft. area.

15



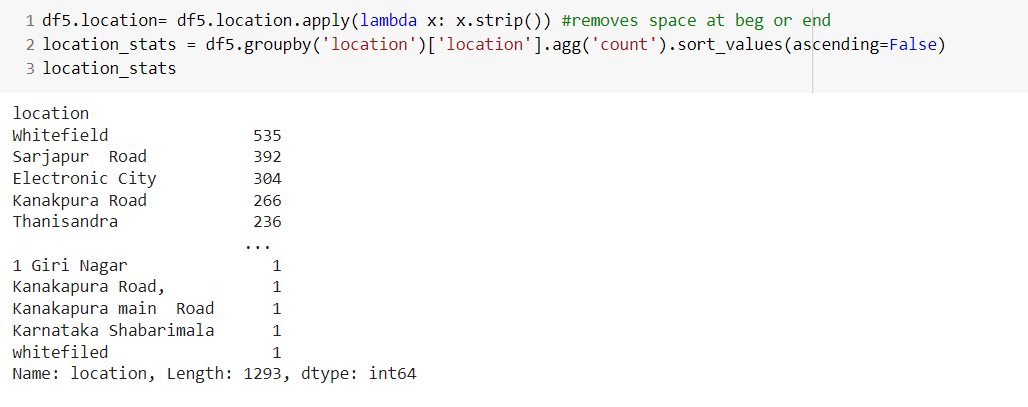
**3.6** Dimensionality reduction

We check the number of columns.



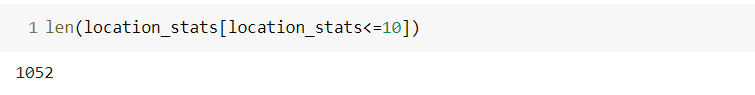
Since the number of columns is 1304, which is very high. Such a high number of columns is known as the dimensionality curse, which is a problem. To resolve this, we shall combine the locations having less than 10 data points into a single value.

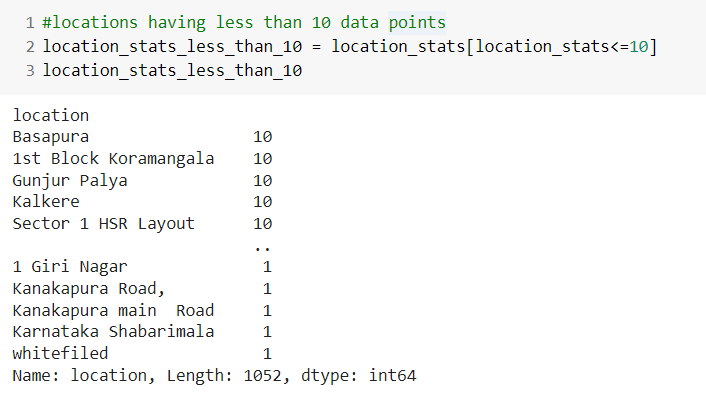
We then check the number of data points for each location.



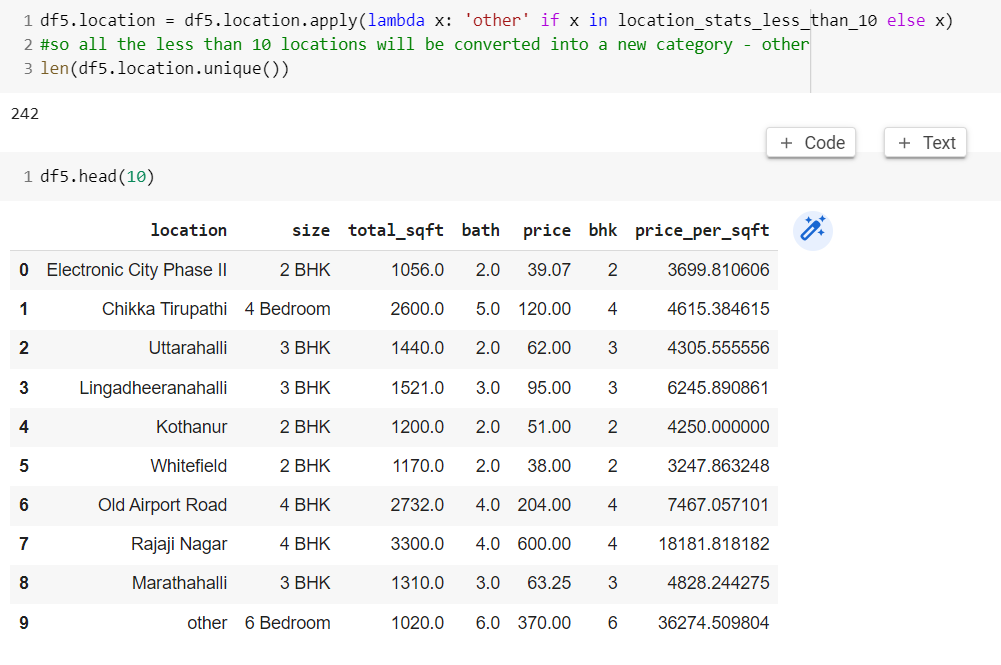
16

The number of locations with less than 10 data points is found.



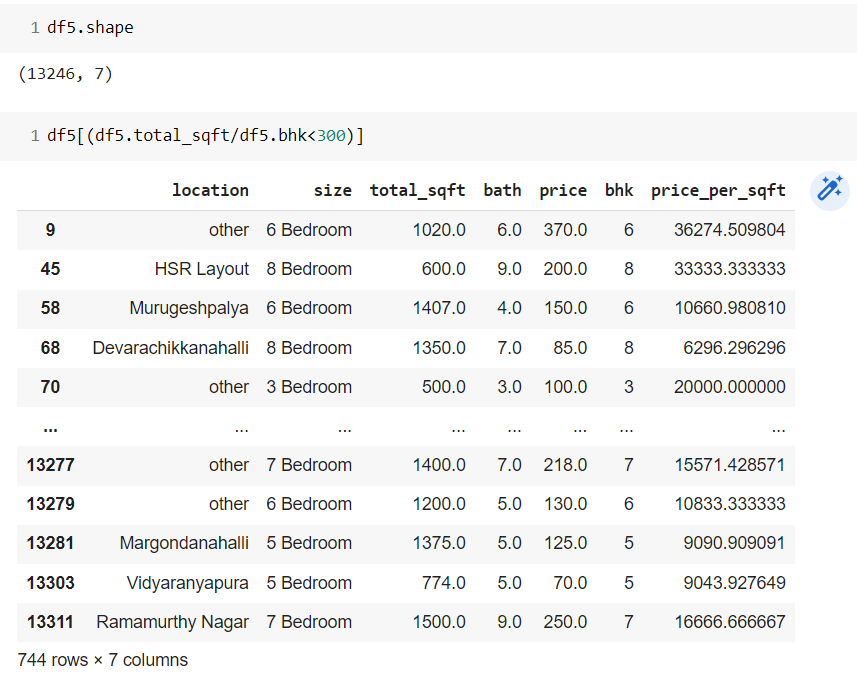


All the locations that have less than 10 data points are now combined in a new location value called other.

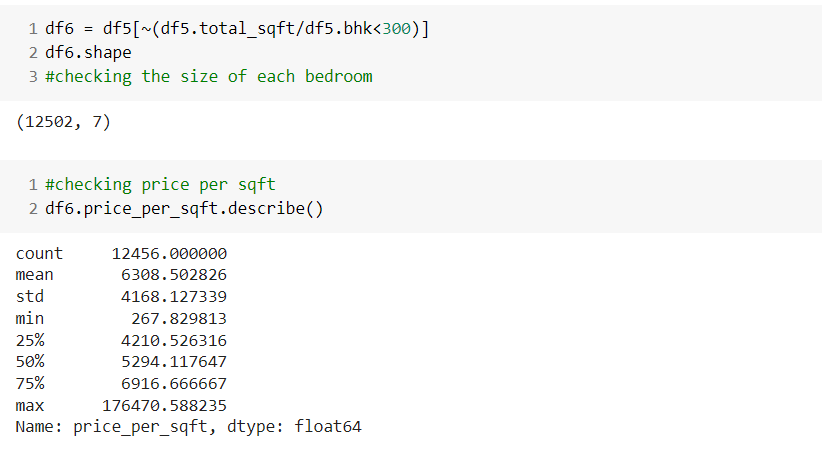


17

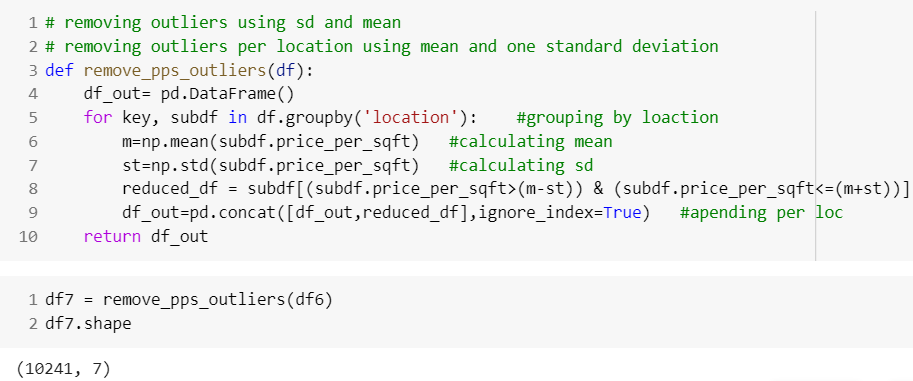
**3.7** Outliers Removal Using Business Logic



Removal of outliers using mean and standard deviation.

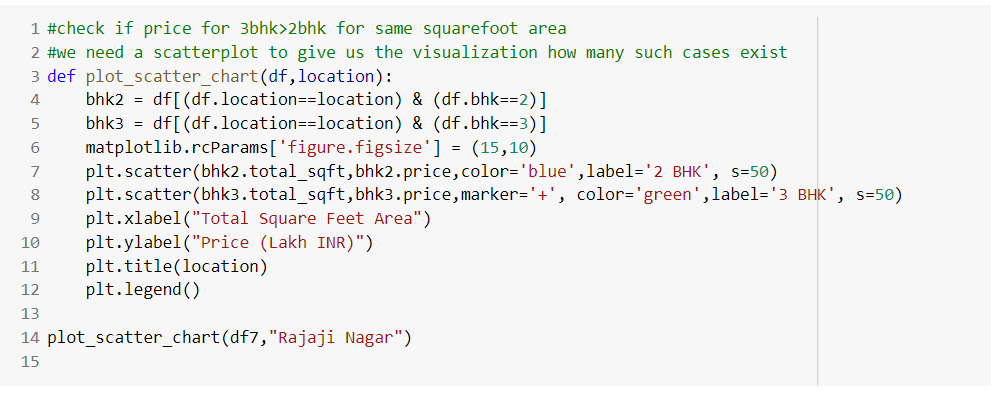


18



Before Removing Outliers

We shall check if the price for a 3bhk flat is more than that of a 2bhk flat for the same square-foot area.



19

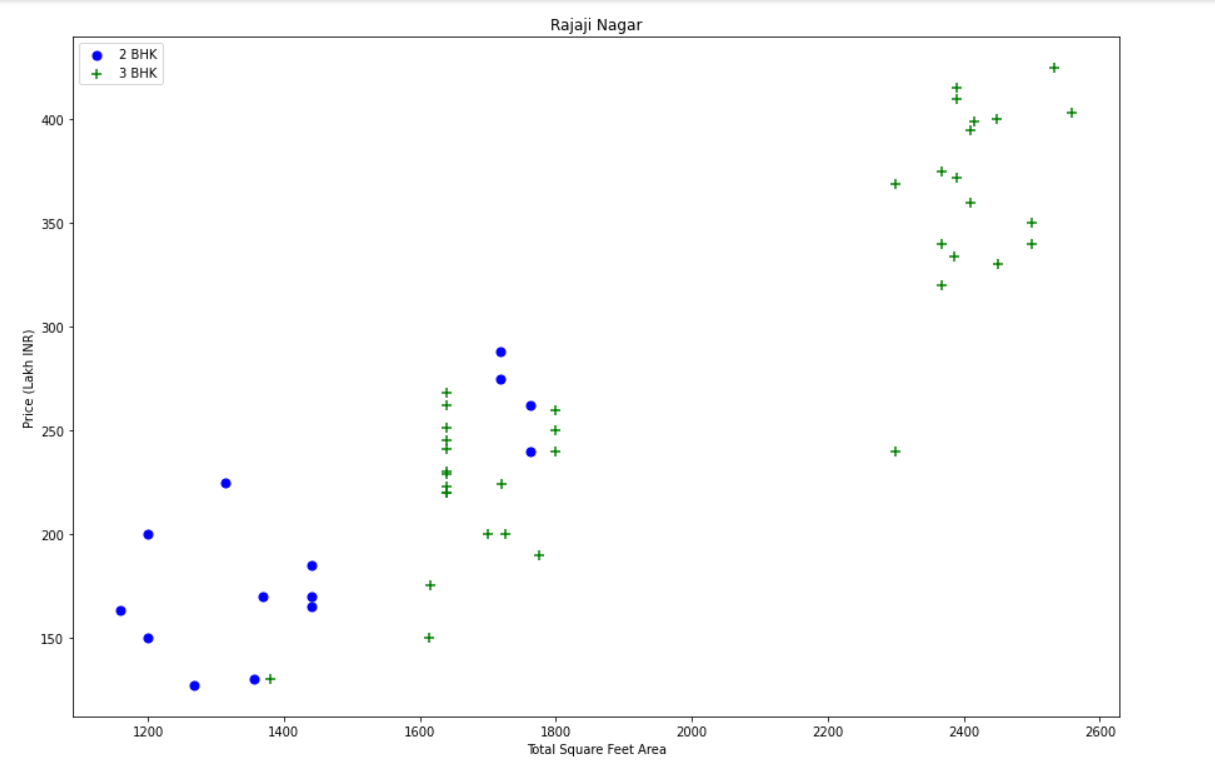


Fig. 6: Plot of variation of price per square feet with total square feet area with outliers

The major outliers have been observed that for the same sq. ft area, 2 BHK price is more than 3 BHK.

After removing outliers:



20

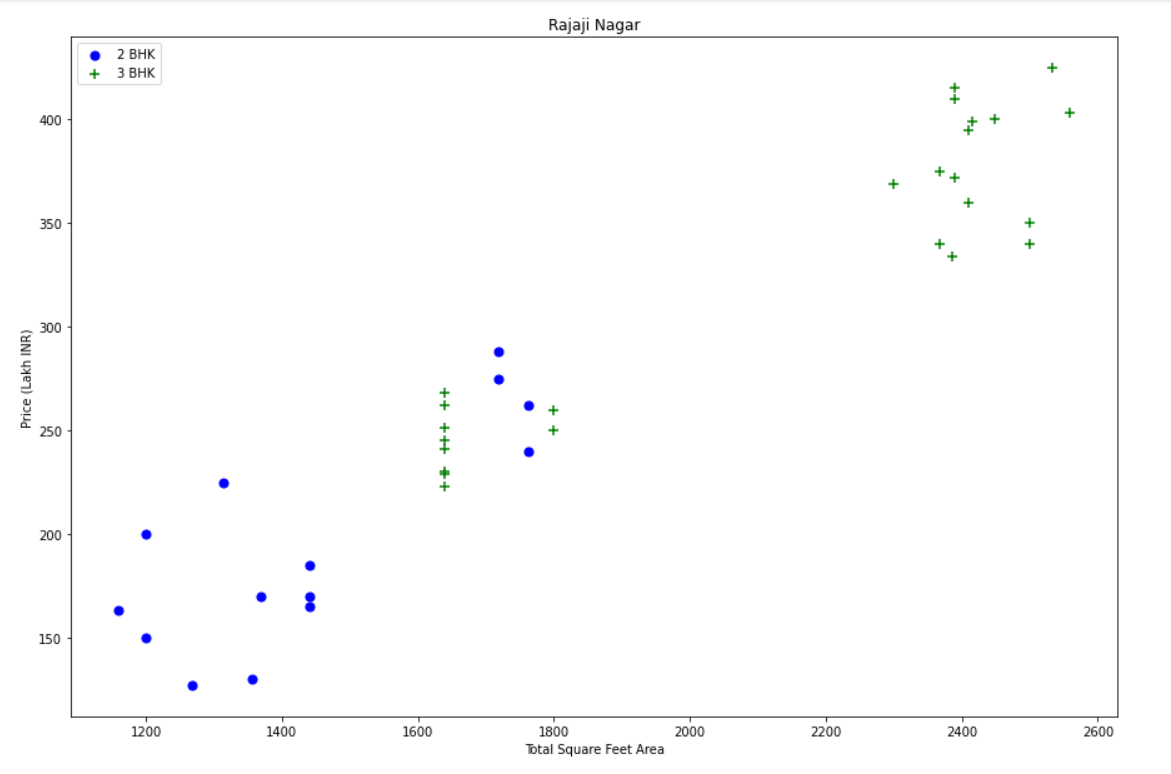


Fig. 7: Plot of the variation of price per square feet with total square feet area with outliers removal

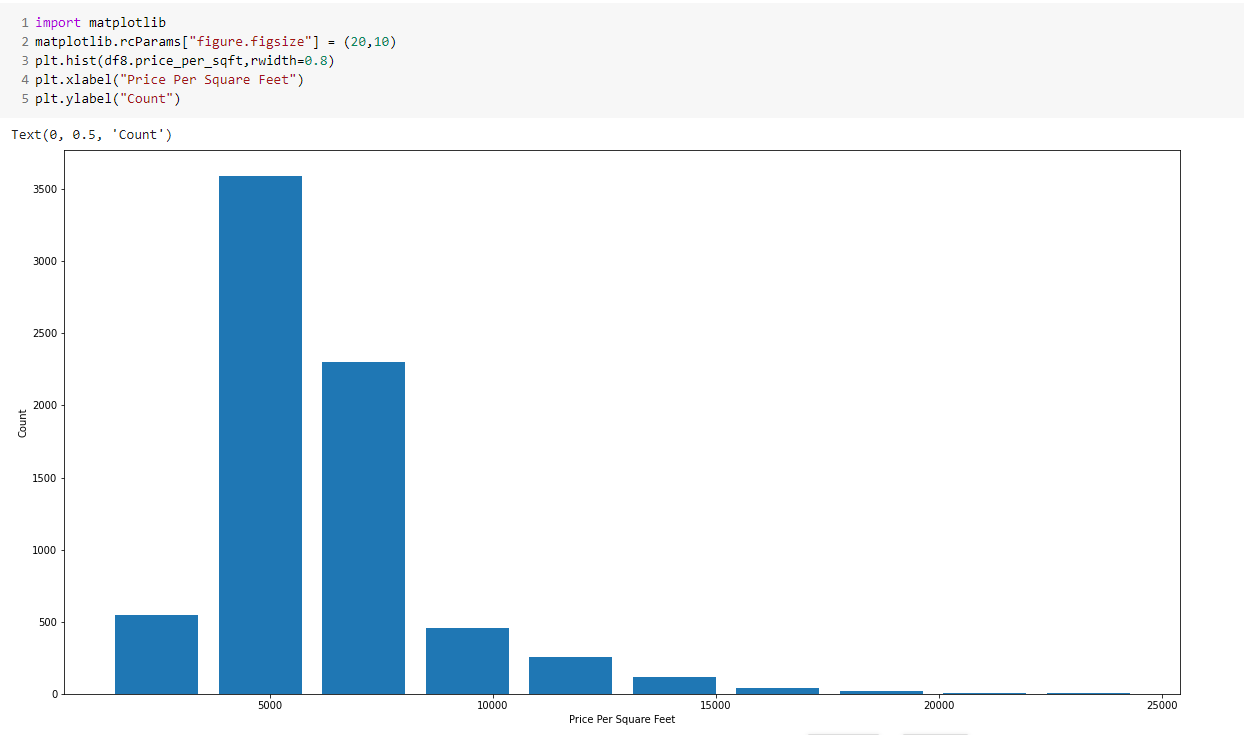
To visualize the count of flats with the price per square feet area, we prefer to use histogram: 

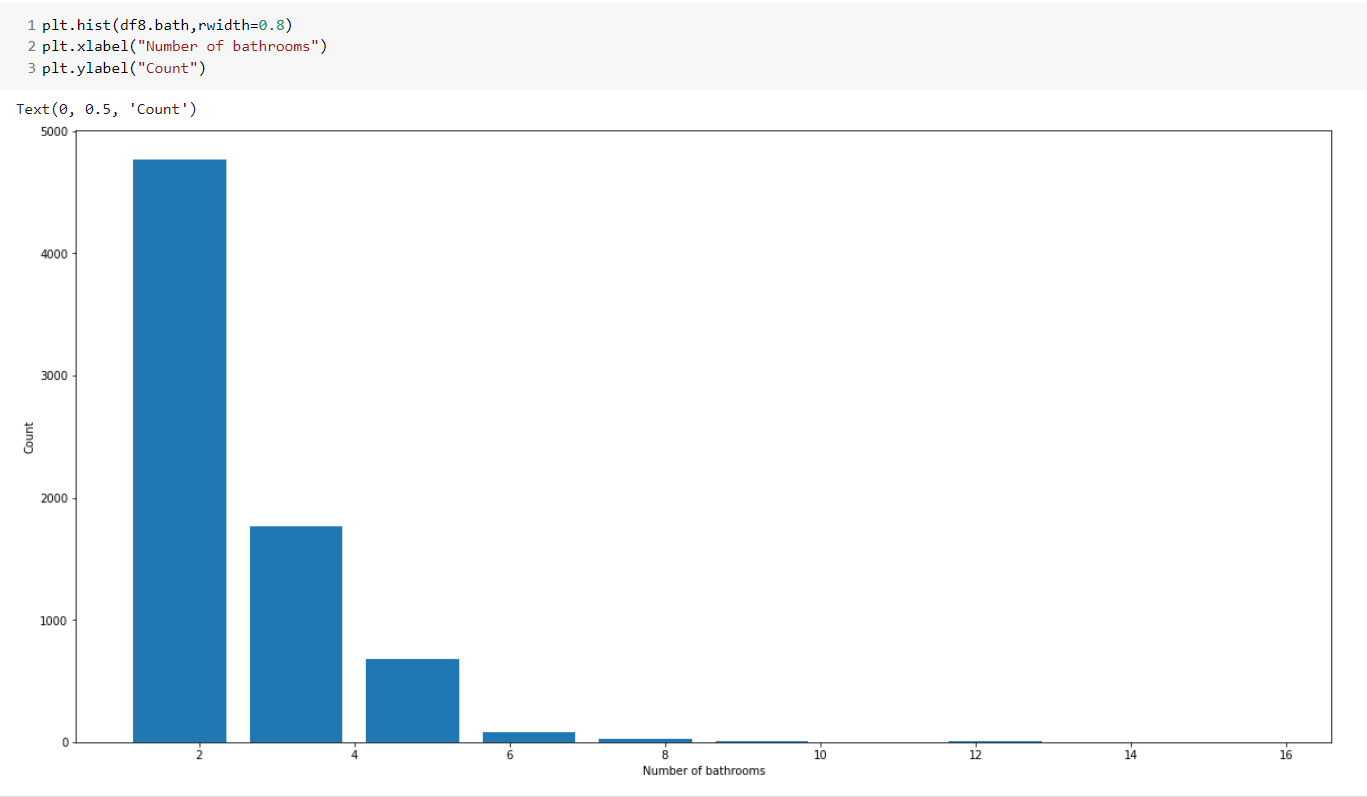
Fig. 8: Plot of the variation of count with price per square feet

21

From the histogram we can observe that a majority of the data points are in 0 to 10000 sq ft range. Hence, it is a normal distribution.

Next, we will remove the bathroom outliers. It is not practically common to have the number of bathrooms twice more than the number of bedrooms.

So, we shall plot a histogram to check the number of bathrooms.



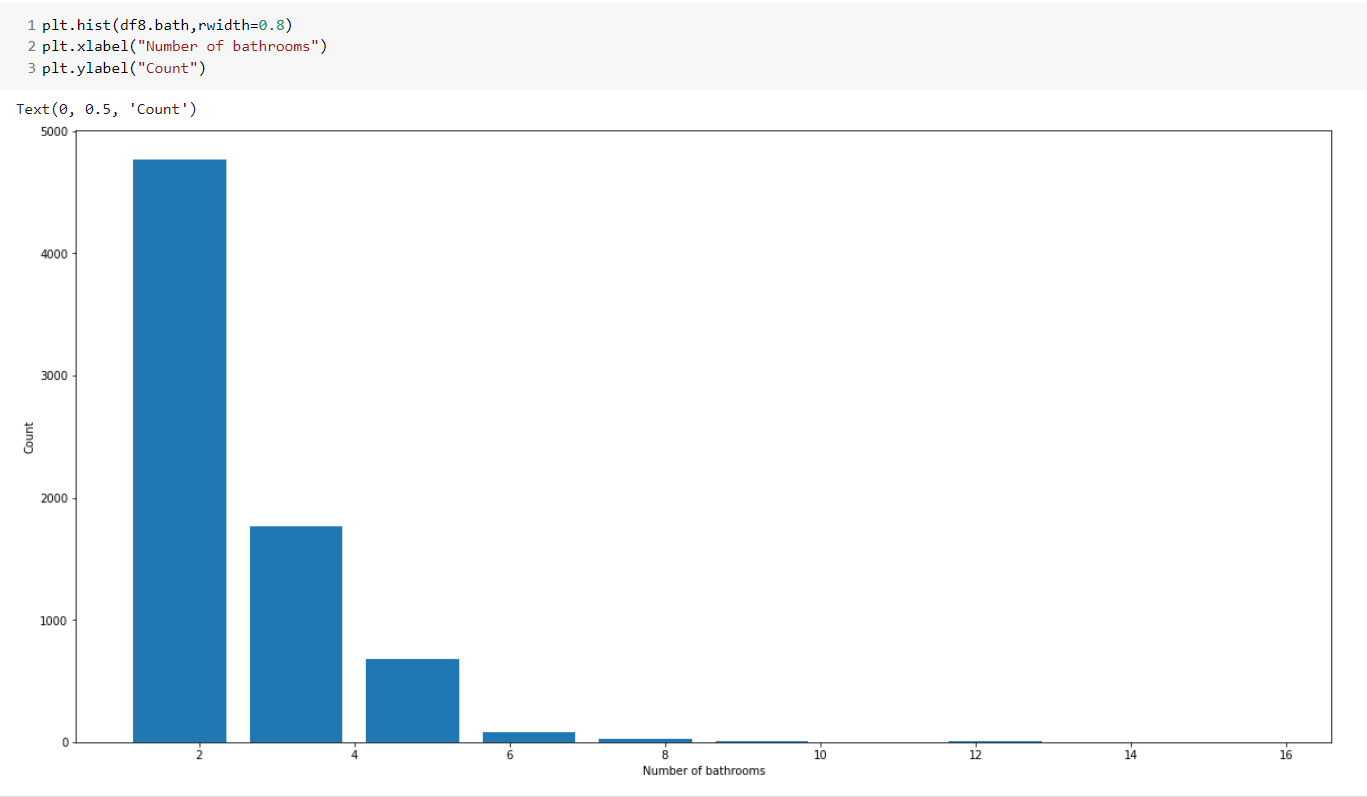
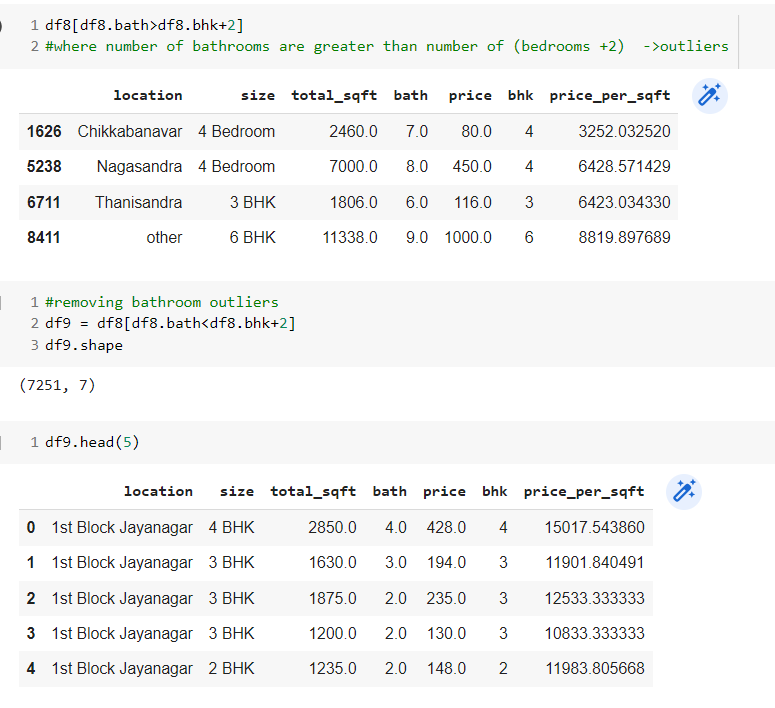


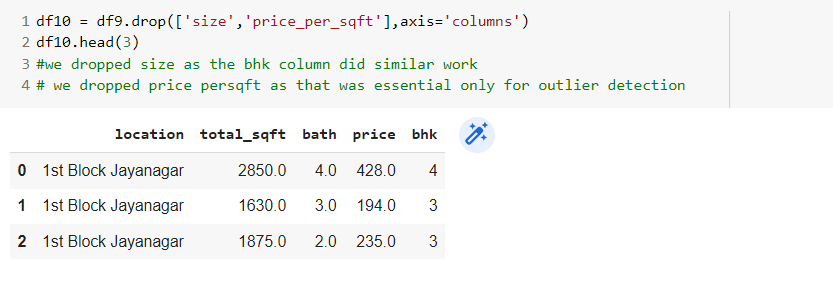
Fig. Plot of Number of bathrooms vs count

Hence, the data points where the number of bathrooms > the number of bedrooms +2, will be removed.

22



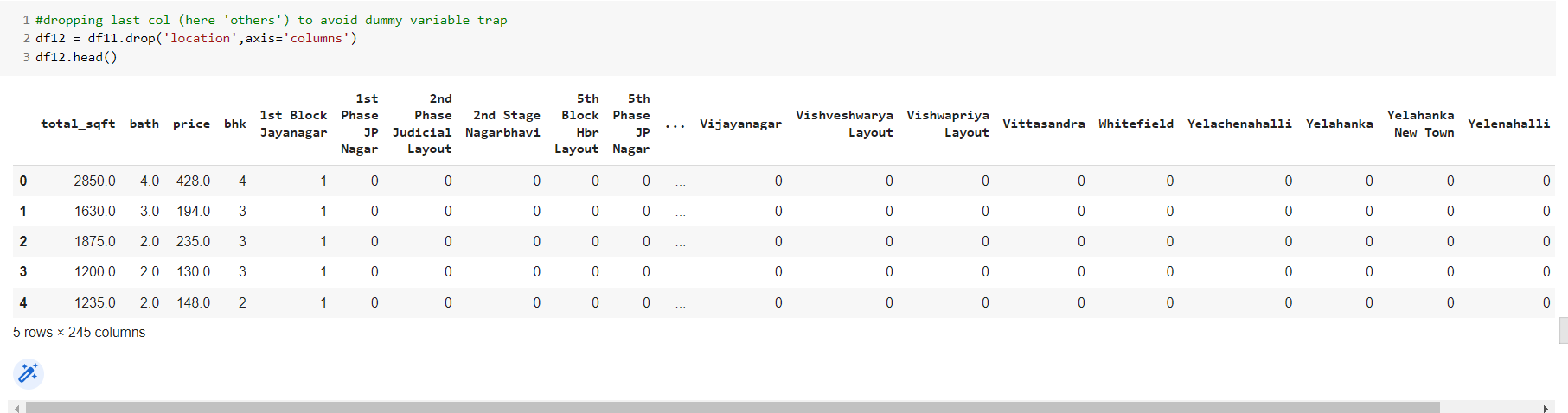
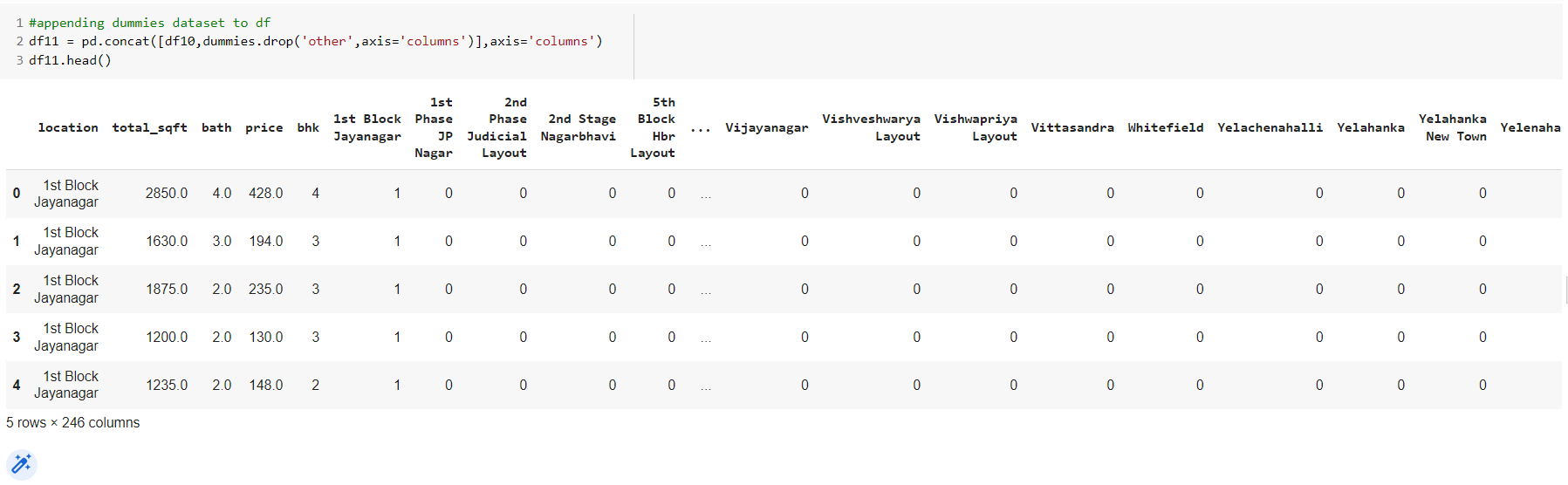
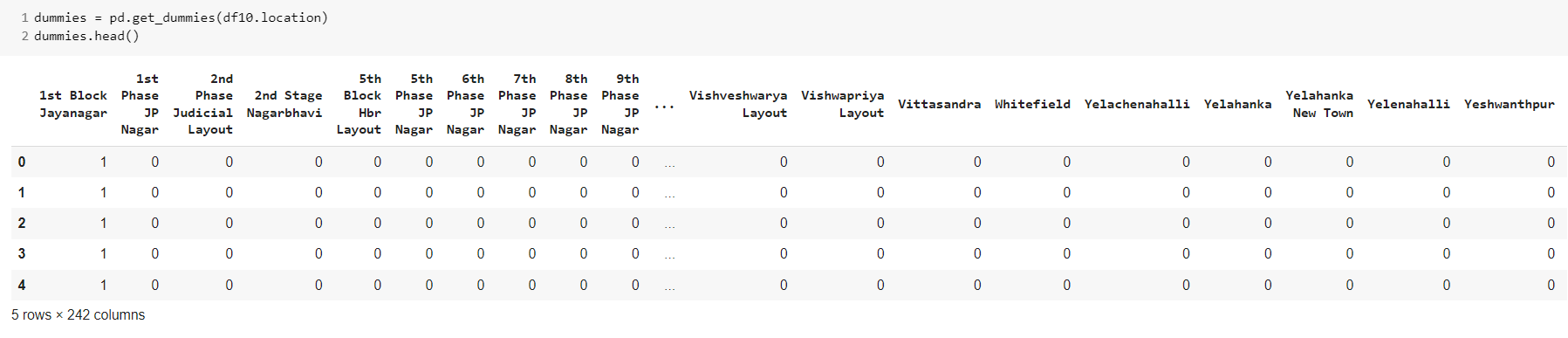
Lastly we drop the price\_per\_sqft column as it was needed only for outlier detection.



**3.8** One Hot Encoding

Machine Learning Model Cannot Interpret Text Data, So we use dummies for locations.

23

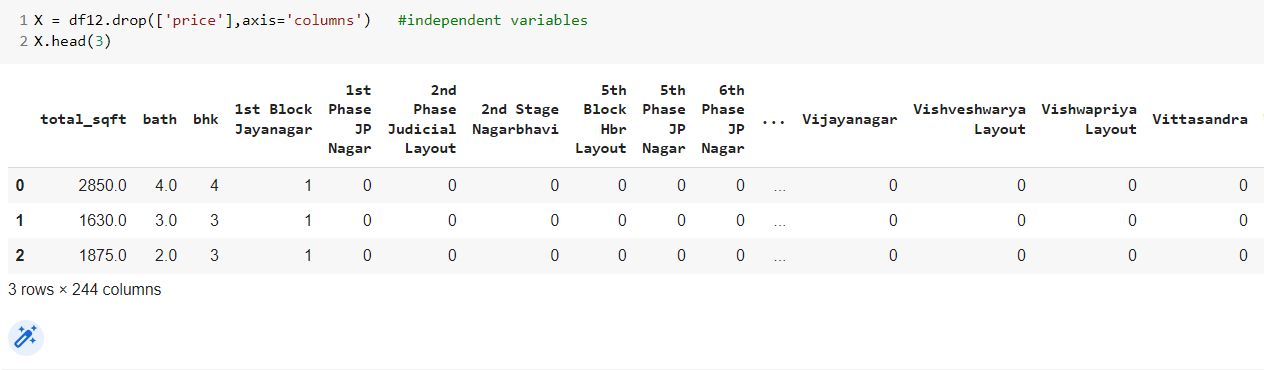
****

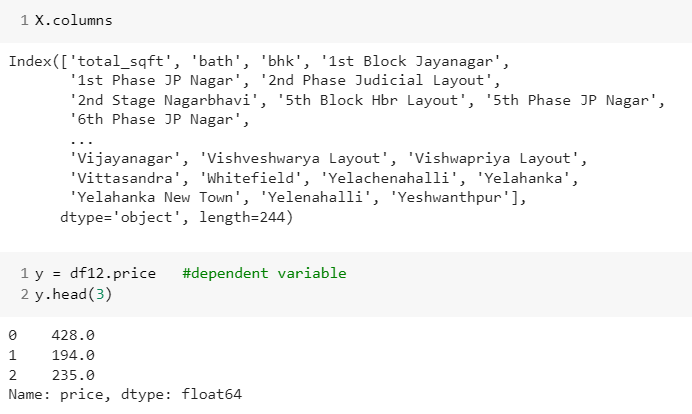
The data is cleaned using the above techniques, now the next stage is to build the regression model.

24

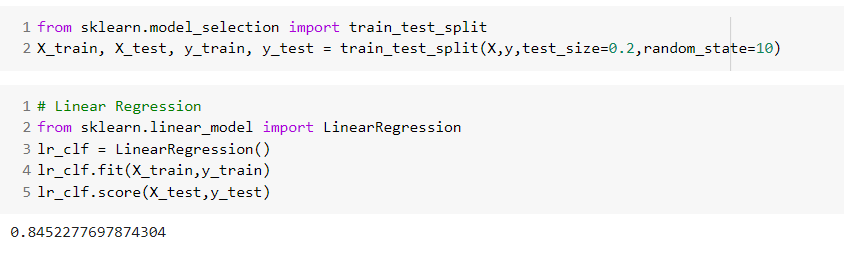
**4. Building Model**

Price column is dropped and the entire list of independent variables is stored in X and price in Y. The linear regression model will design and build our model using sklearn.

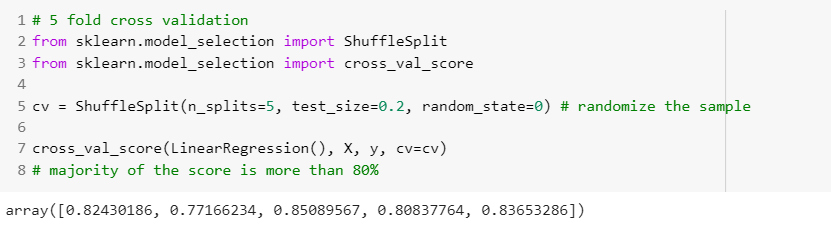




25



**5. K-Fold Cross Validation to measure the accuracy of our Linear Regression Model**

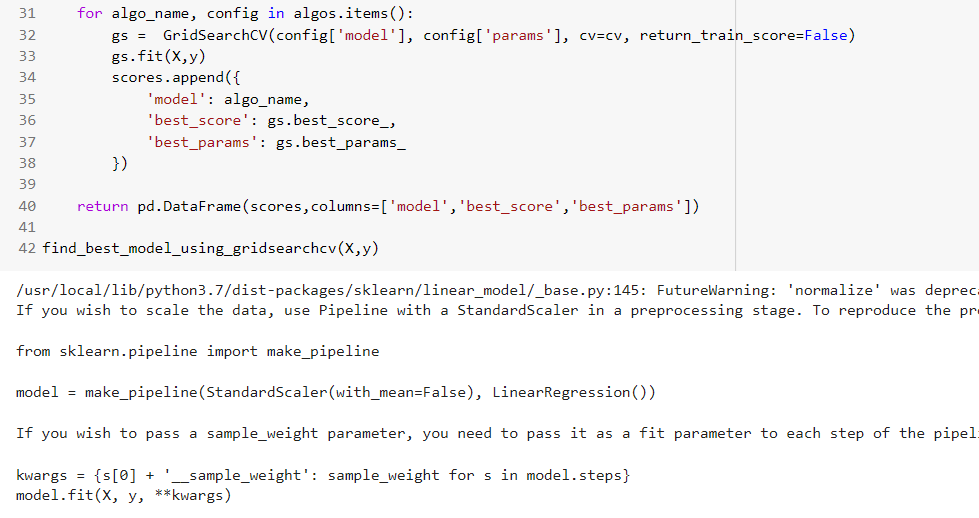
****

**6. GridSearchCV is implemented.**

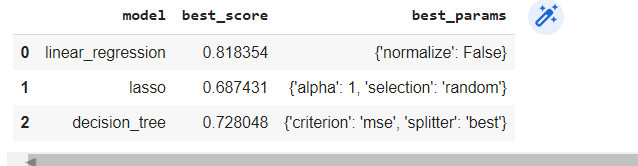
The accuracy of various datasets is then checked with GridSearchCV.

The machine learning models used here are Linear Regression, Lasso Regression, and Decision Tree. GridSearchCV provides us with the model accuracy and the parameters that should be used to get the accuracy.

26

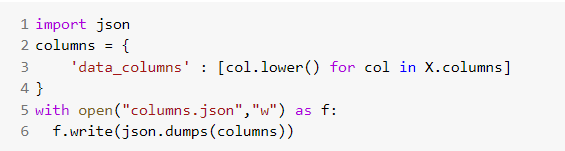
****

27

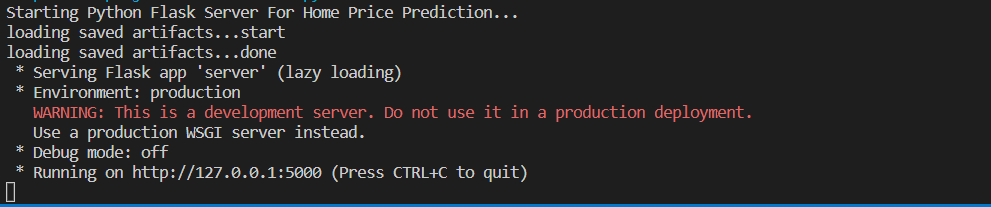
****

Based on the above result we can say that Linear Regression gives the best accuracy score. Hence, we shall use Linear Regression to predict the real estate prices(in lacs.

**7. The model is pickled to be used for the generation of server requests in the flask. **



**8. Flask server request is generated**

****

28

**9. The web page is ready to be operated.**

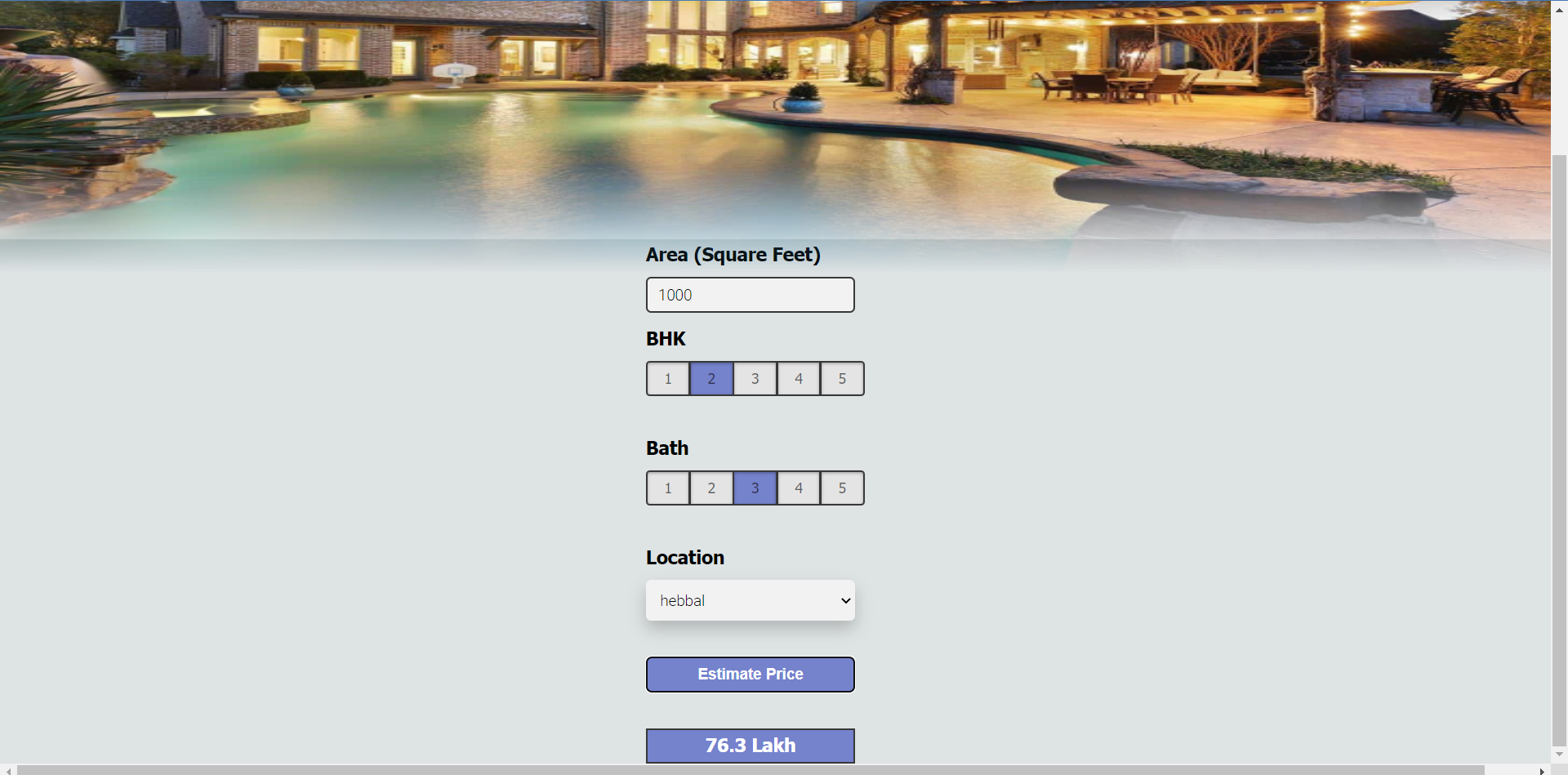
****

Fig. 11: Visual of the web page

The location will give a drop-down of all the locations in the city and once the ‘Estimate Price’ button is clicked, we shall obtain the price of the real estate of our specifications.

29

**Chapter 4: Results**

**4.1 Discussion on the Results Achieved**

The model has successfully been able to estimate the price of real estate. On cleaning the data, it becomes structured and hence it becomes easy to perform manipulations on the dataset. The efficiency of the linear regression model is much higher as compared to other models.

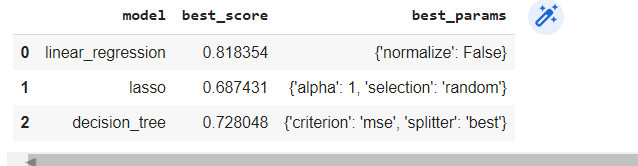
****

Fig. 15: Comparison of the efficiency scores of different regression models

The web page runs on the flask-generated server. The UI designed for the customer is feasible content to work on.

The price is estimated in a span of time and the interface offers the customer to make multiple selections and predict the price of the real estate.

**4.2 Application of the Project**

The project incurs to the present demand of the market. The real estate customers wish to know the price of the flat or the building while just exploring the different locations in a city. The model offers the customer to select the specifications of his desired flat and estimate the price accordingly. He could maintain the authenticity of the record shared by his data analyst and offer relevant amendments by observing the visual analysis.

30

**4.3 Limitations of the Project**

No model can be a perfect implementation of the idea in the first attempt, our model has certain limitations too.

i. The webpage runs on the server request generated by the flask for localhost and is not a domain purchased model.

ii. The price estimation of the real estate on the User Interface is not an immediate process but takes a span of 4-5 seconds.

**4.4 Future Work**

We have a list of certain improvements and future work to be performed on the project:

i. The sales analysis model shall be worked upon to make a Graphical UI for better visualization and easy operation.

ii. The website shall be made live and globally operative by purchasing a domain name.

iii. The website shall be made responsive.

**Conclusion**

The project that we have created shall find a suitable position in the cyber market world catering to the need for a platform to explore the desired property in a city and the owner or the dealer gets an appropriate platform to keep a check on his sales chronologically.

This opportunity provided to us has accomplished its motive of letting us explore our knowledge and implement it.

31

**References**

[1] https://www.geeksforgeeks.org/functional-vs-non-functional-requirements/

[2] https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data

[3] Rana, V., Mondal, J., Sharma, A., & Kashyap, I. (2020). *House Price Prediction Using Optimal Regression Techniques*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9362864&isnumber=9362727

[4] Fan, C., Cui, Z., & Zhong, X. (2018). *House Prices Prediction with Machine Learning Algorithms*. Association for Computing Machinery. https://doi.org/10.1145/3195106.3195133

[5] https://realpython.com/linear-regression-in-python/

32