**CITY CLASSIFICATION USING MACHINE**

**LEARNING**

**A** MajorProject

Submitted in partial fulfillment for the

Completion of the

**Verzeo Data Science Internship**

**(Sept-Oct Batch)**

Submitted by

|  |  |
| --- | --- |
| **MEHUL RAJ** | [**mehulraj1995@gmail.com**](mailto:mehulraj1995@gmail.com) |
| **HIMANSHU SINGH** | [**himanshusinghgn2502@gmail.com**](mailto:himanshusinghgn2502@gmail.com) |
| **SHRAVYA M** | [**Shravya1903@gmail.com**](mailto:Shravya1903@gmail.com) |
| **HAUZAN GIRACH** | [**hauzangirach20@gmail.com**](mailto:Hauzangirach20@gmail.com) |
| **KARAN B** | [**Bkaran034@gmail.com**](mailto:Bkaran034@gmail.com) |

**Contents:**

* **INTRODUCTION**
* **EXPLORATORY DATA ANALYSIS ALGORITHM**
* **LOGISTIC REGRESSION ALGORITHM**
* **ML ALGORITHMS USED**
* **RANDOM FOREST ALGORITHM**
* **SUPPORT VECTOR MACHINE**
* **Logistic Regression**
* **KNN**
* **ENSEMBLE MACHINE LEARNING MODEL**

**INTRODUCTION**

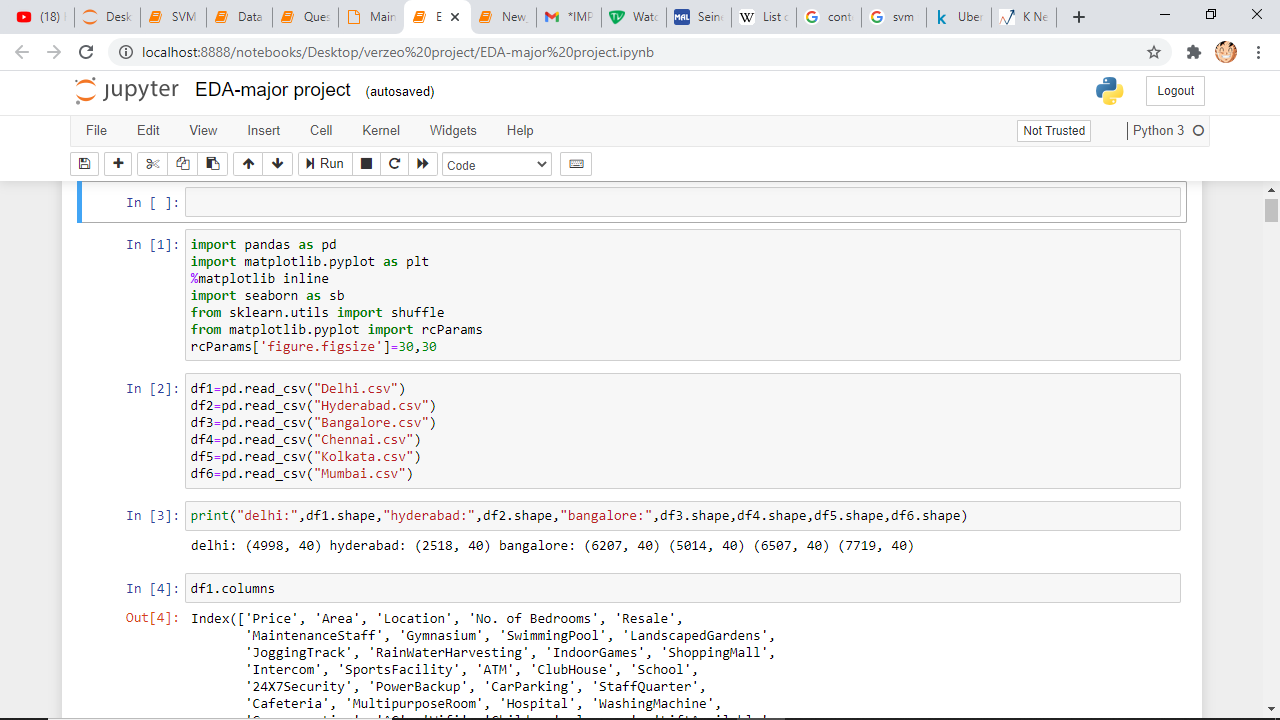
We have a collection of total 32963 Datasets in the main.csv file. Which covers nearly most of the Metropolitan cities in all the Indian states. We are classifying the city where the particular house with facilities given is actually located.

We are going to visualize the dataset and answer two questions using three different algorithms respectively and find which algorithm provides the output with more accuracy.

* **EXPLORATORY DATA ANALYSIS ALGORITHM**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

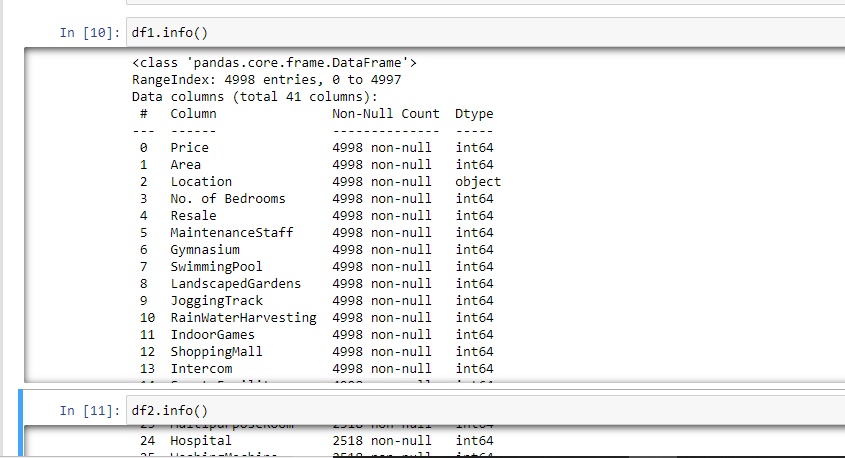
To start with, we imported necessary libraries (for this example **pandas**, **numpy**, **matplotlib** and **seaborn**) and loaded the data set.

****

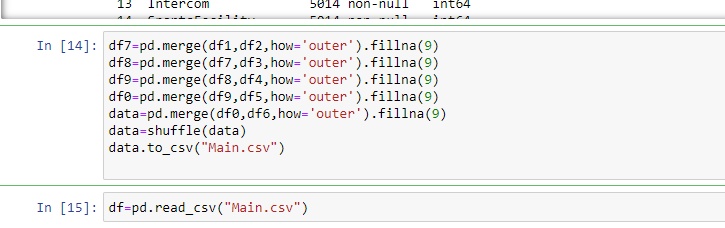
We used the variable df1, df2, df3, df4, df5 and df6 to access the files Delhi.csv, Hyderabad.csv, Bangalore.csv, Chennai.csv, Kolkata.csv and Mumbai.csv respectively.

****

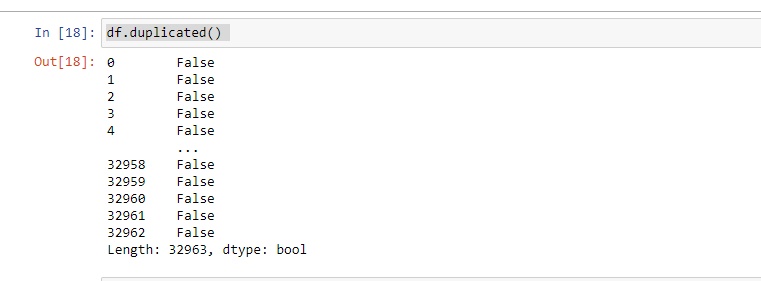
We used “**df1.columns**” to display all the column names. Likewise we displayed all the df2, df3, df4, df5 and df6 respectively.

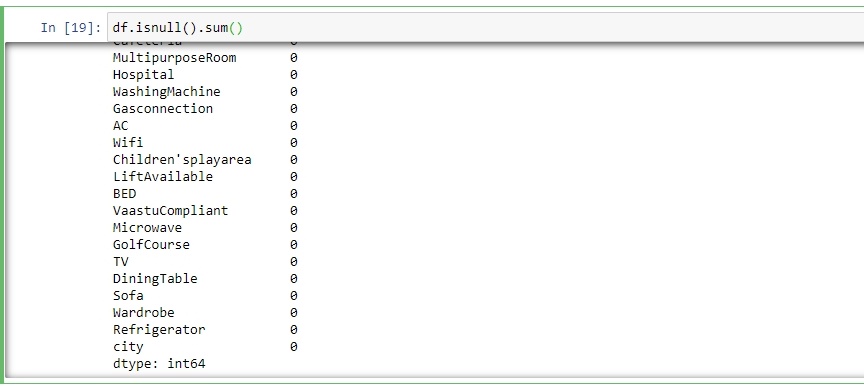
****

We used “**df1.info()**” function to check whether there is any null value. Likewise did for df2, df3, df4, df5 and df6 respectively.

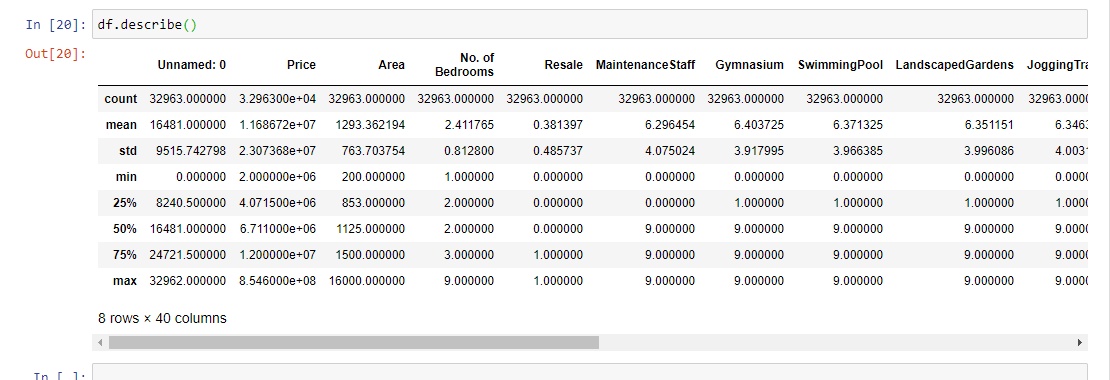
****

Using the above code we shuffled and merged all the datasets df1, df2, df3, df4, df5 and df6 into single variable “**df**” (Main.csv).

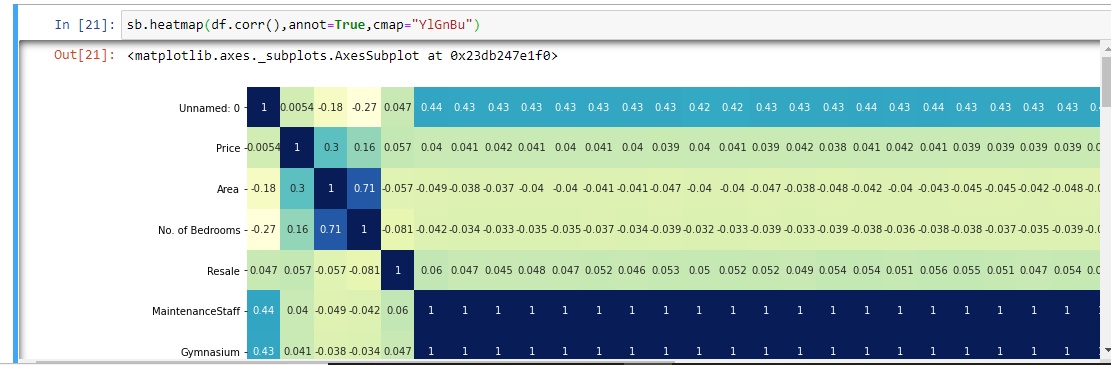
****We used the function “**.duplicate()**” to check whether there are any repeated values are there in the dataset.

****

The function **df.isnull().sum()**returns the number of missing values in the data set

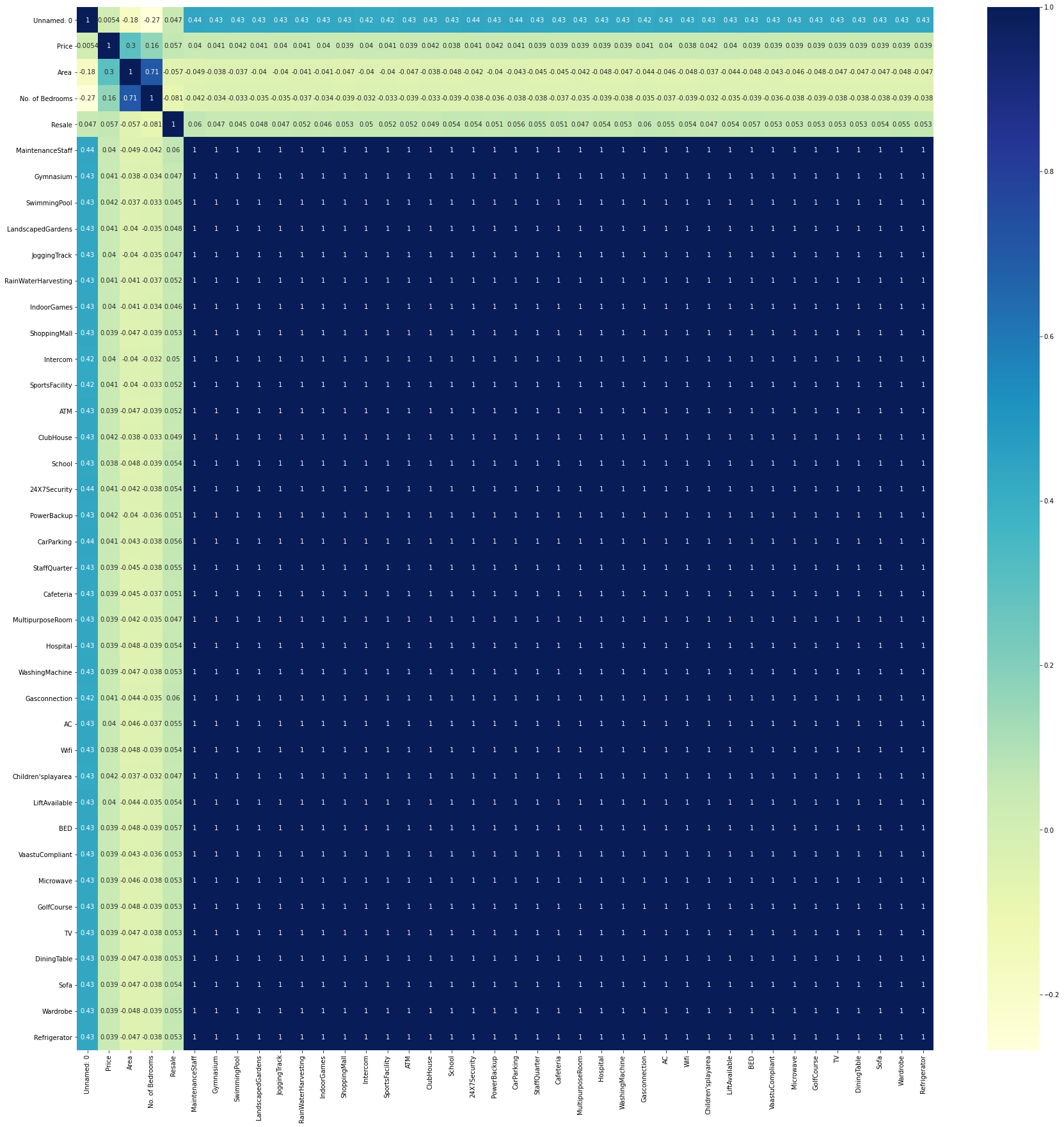
****

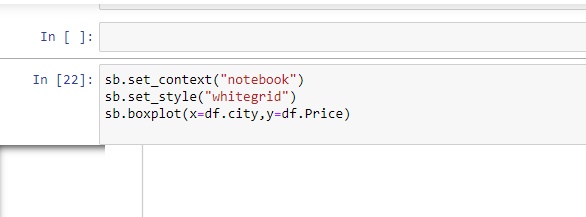
The  “**.describe()”** method is used for calculating some statistical data like percentile, mean and standard of the numerical values of the Series or **DataFrame**. It analyzes both numeric and object series and also the **DataFrame** column sets of mixed data types.

****

We used the “**.heatmap()**” function to create a Plot rectangular data as a color-encoded matrix.

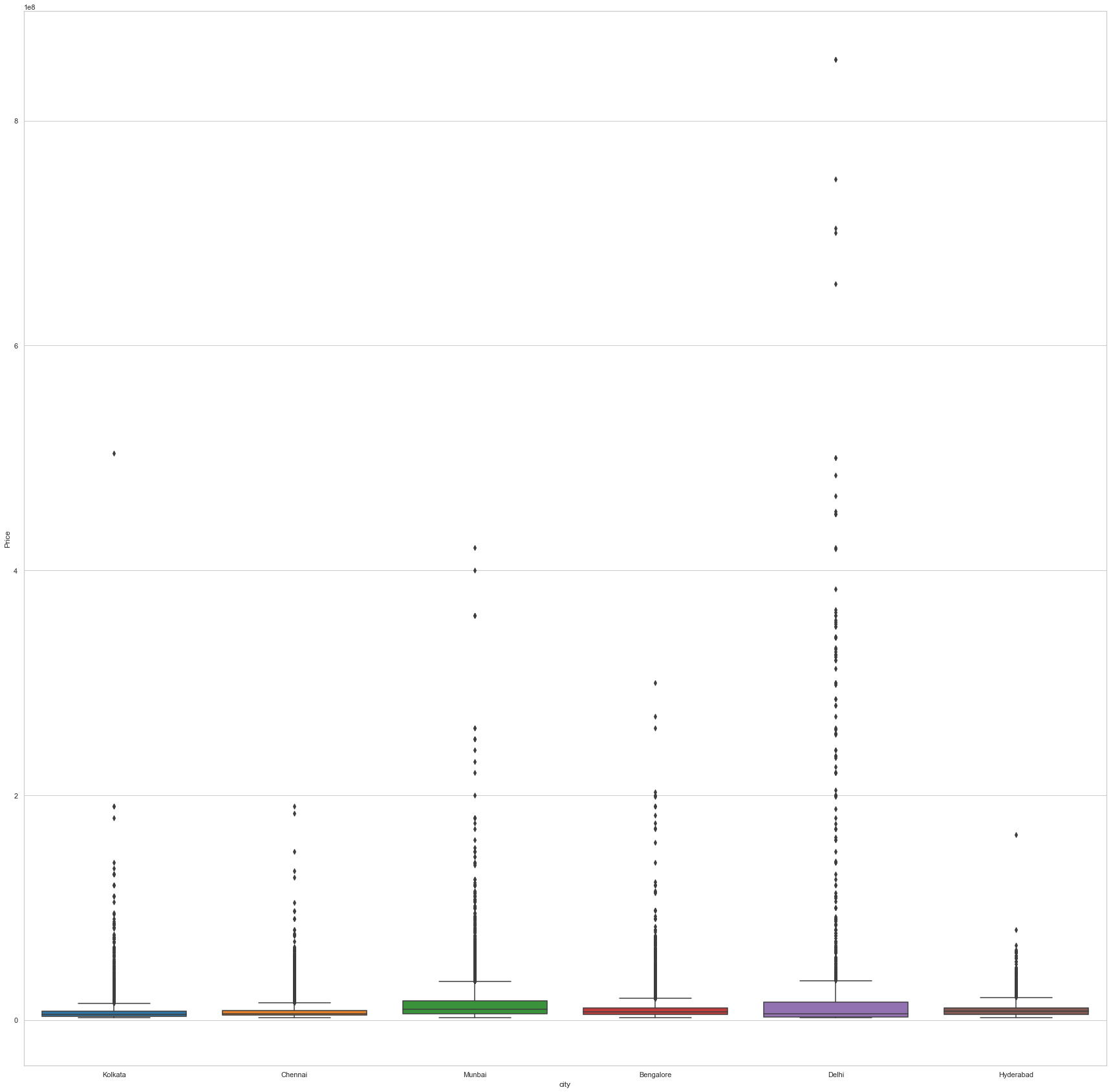
**OUTPUT:**

****

****

We used the “**.boxplot()**” function to Plot a graph with **city** as x-axis and **Price** as y-axis

**OUTPUT:**

****

* **Which city is being unreasonable for pricing as in having more values out of their normal range?**

We can identify from the above graph that the city which is unreasonable for pricing is **Delhi**.

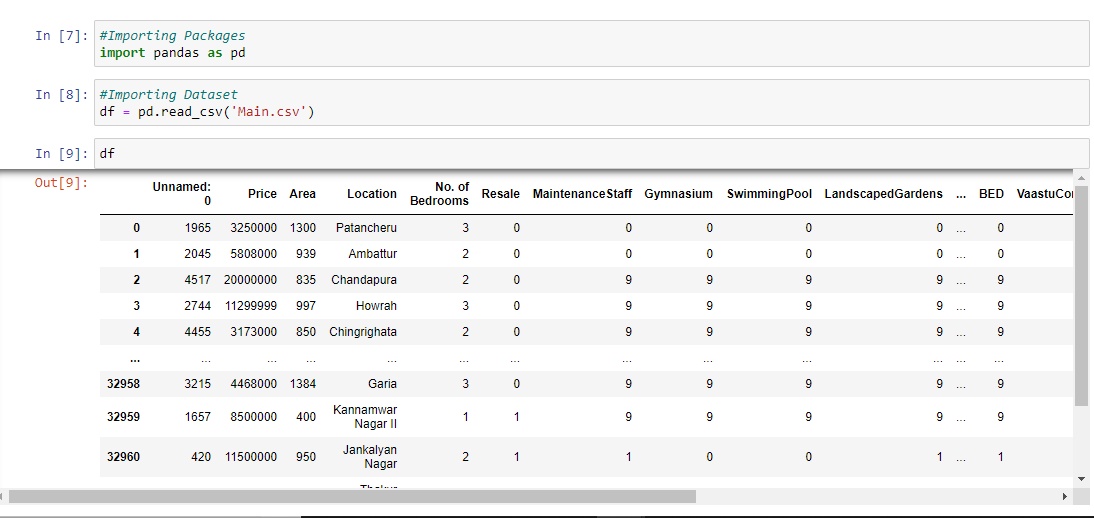
* **LOGISTIC REGRESSION ALGORITHM**

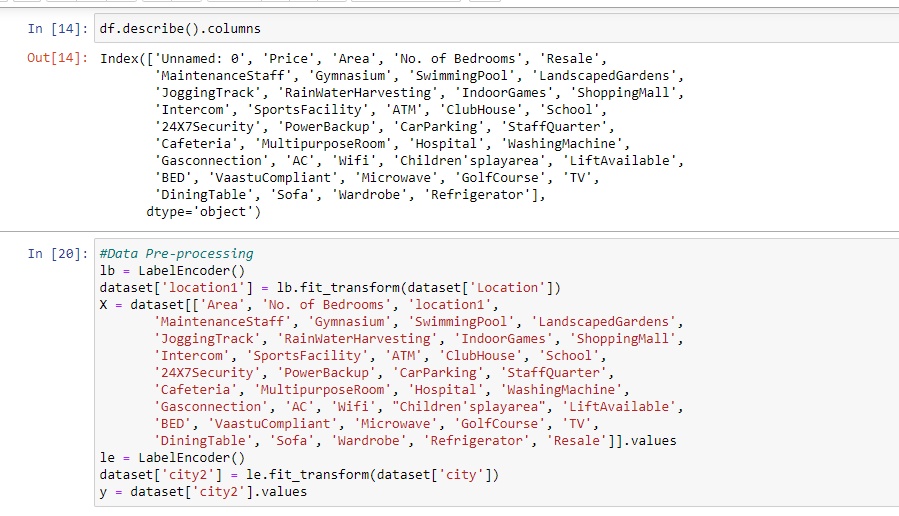
Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

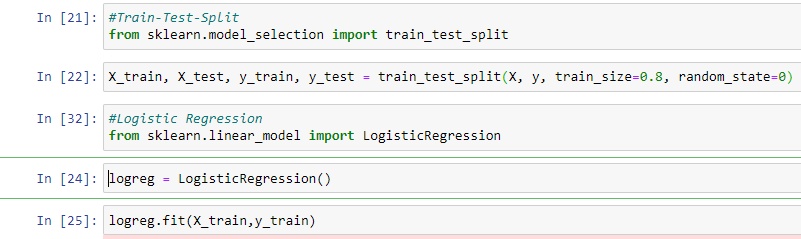
Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

We imported **pandas** and created a variable **df** to access the file Main.csv

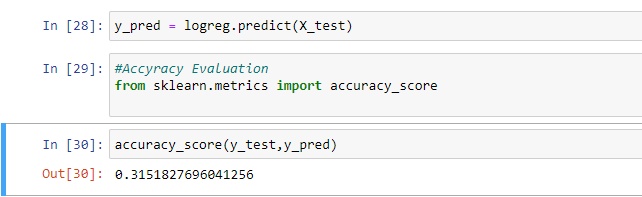




  We used **LabelEncoder** to normalize labels.



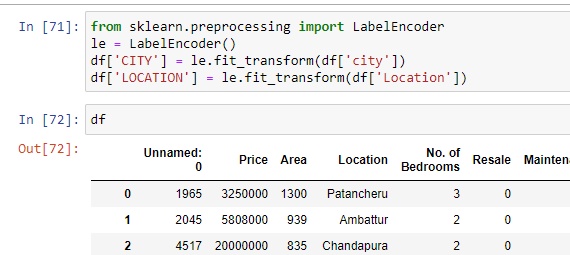
We used **train\_test\_split** to split arrays or matrices into random train and test subsets.



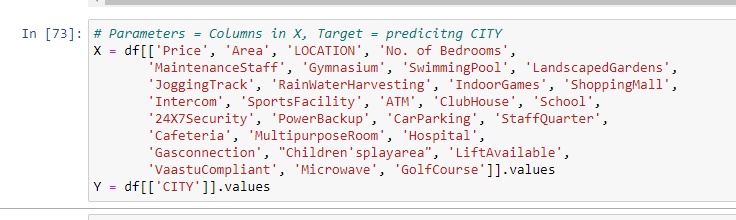
The accuracy obtained by the Logistic Regression algorithm is 0.315

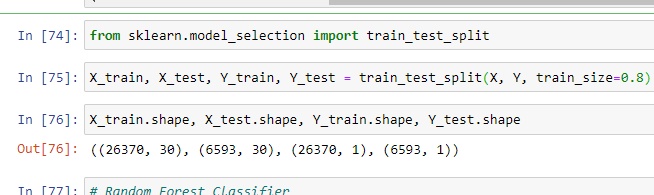
* **RANDOM FOREST ALGORITHM**

Random forest is a supervised learning algorithm. The “forest” it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

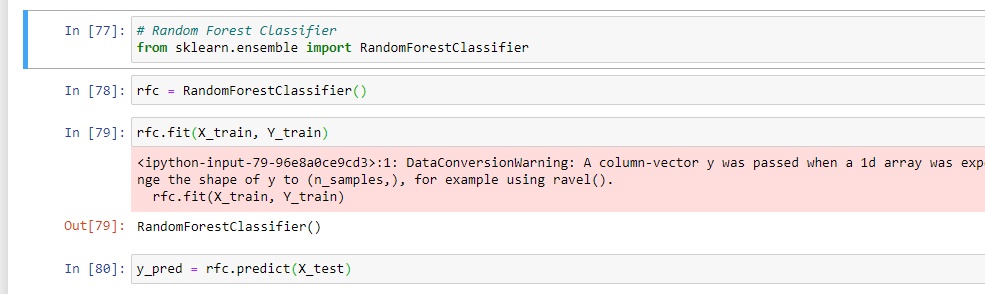


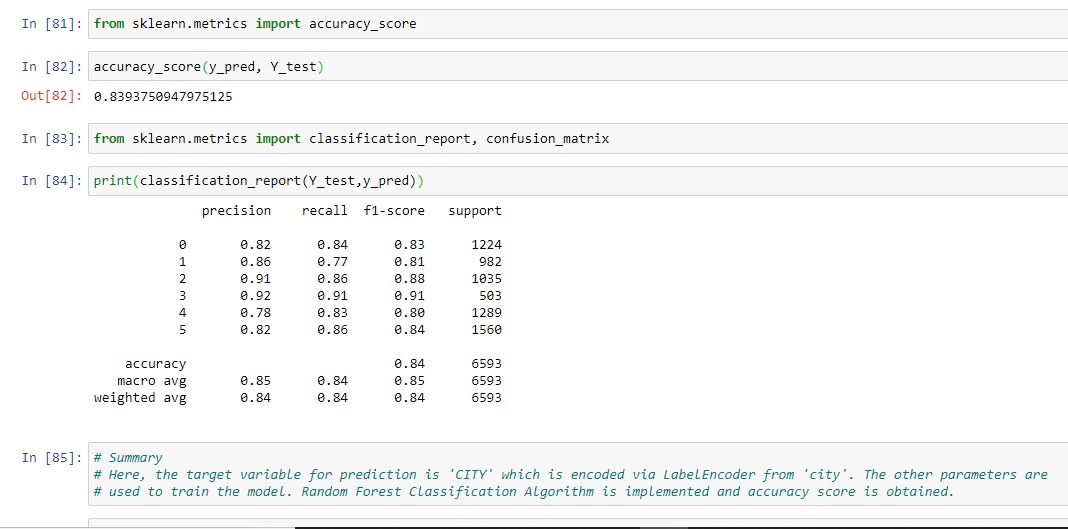
We used **LabelEncoder** to normalize labels.





We used **train\_test\_split** to split arrays or matrices into random train and test subsets.

****

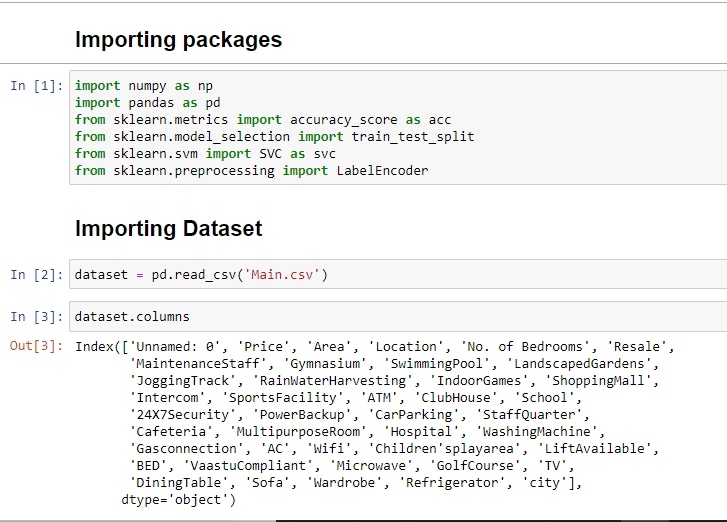
****

Here, the target variable for prediction is ‘**CITY**’ which is encoded via LabelEncoder from ‘**city**’. The other parameters are used to train the model. Random Forest Classification Algorithm is implemented and accuracy score is obtained.

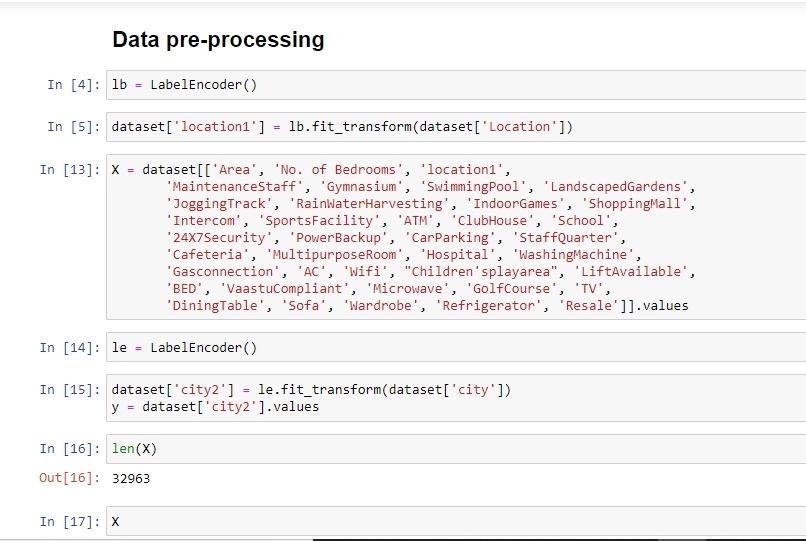
Hence the accuracy obtained by using Random Forest Algorithm is 0.84.

* **SUPPORT VECTOR MACHINE**

It analyzes the data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences.

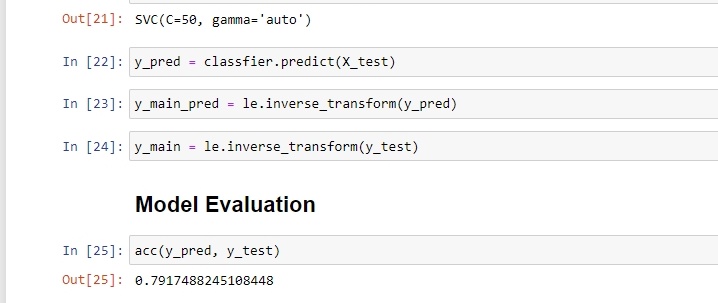
****

By using the above code we have imported packages and dataset.

****

In the Data Pre-Processing We used **LabelEncoder** to normalize labels. We used **train\_test\_split** to split arrays or matrices into random train and test subsets.

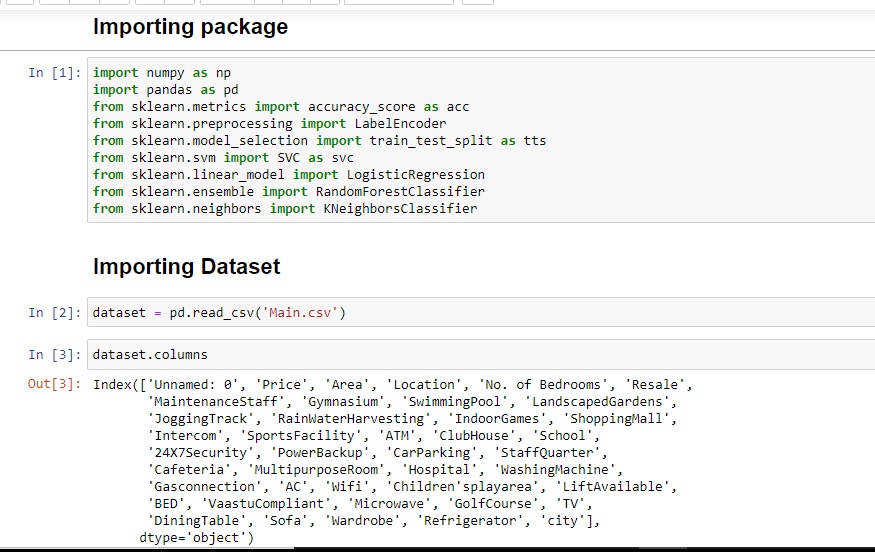
****

****

Hence the accuracy obtained by the Support Vector Model Algorithm is 0.79

* **ENSEMBLE MACHINE LEARNING MODEL**

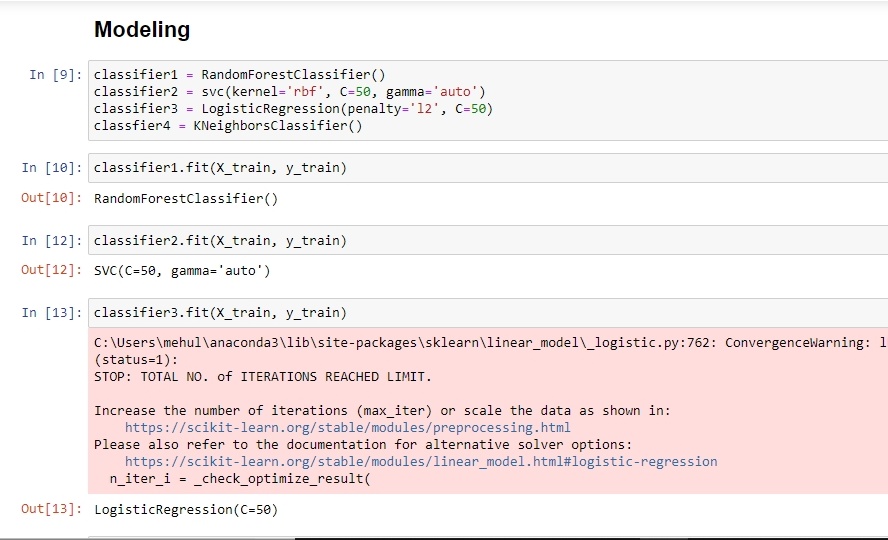
**Ensemble** method **is** a **machine learning** technique that combines several base models in order to produce one optimal predictive model.

****

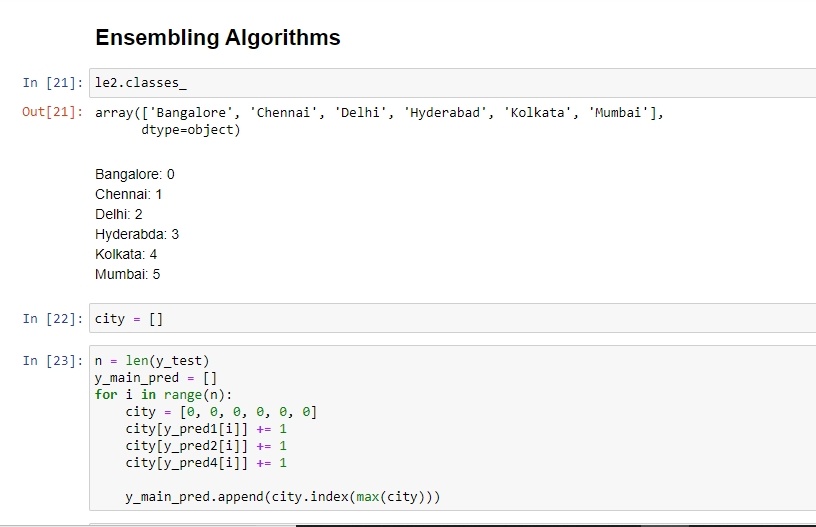
By using the above code we have imported packages and dataset.

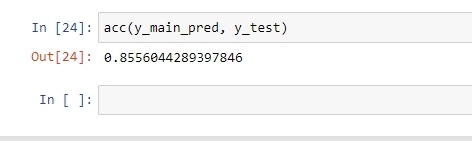
****

In the Data Pre-Processing We used **LabelEncoder** to normalize labels. We used **train\_test\_split** to split arrays or matrices into random train and test subsets.

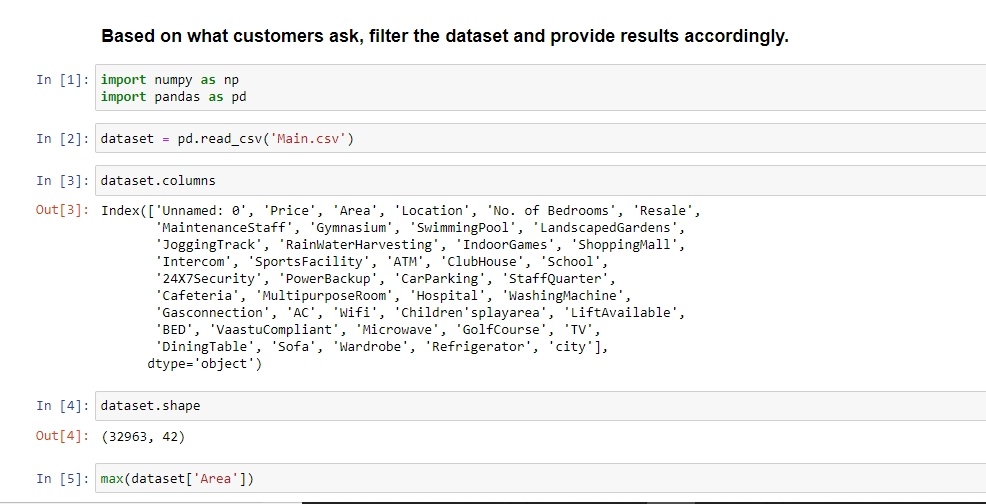
****

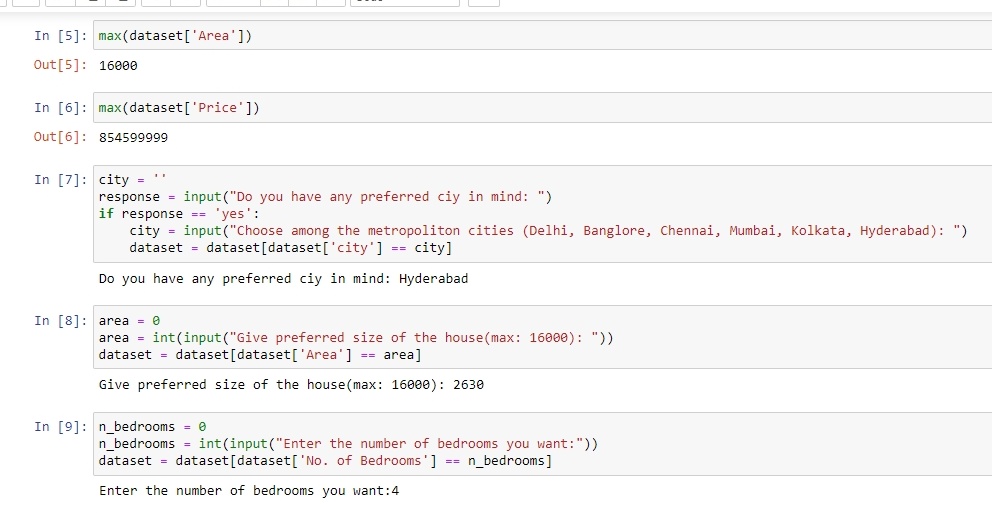
****

****

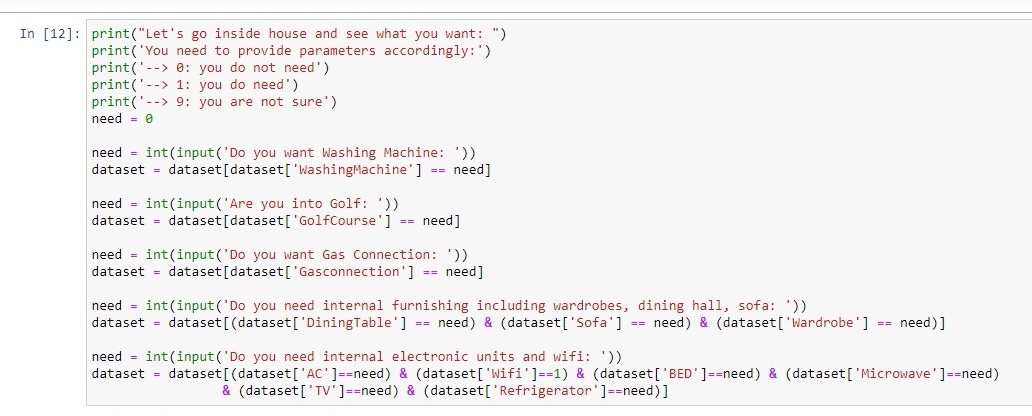
****

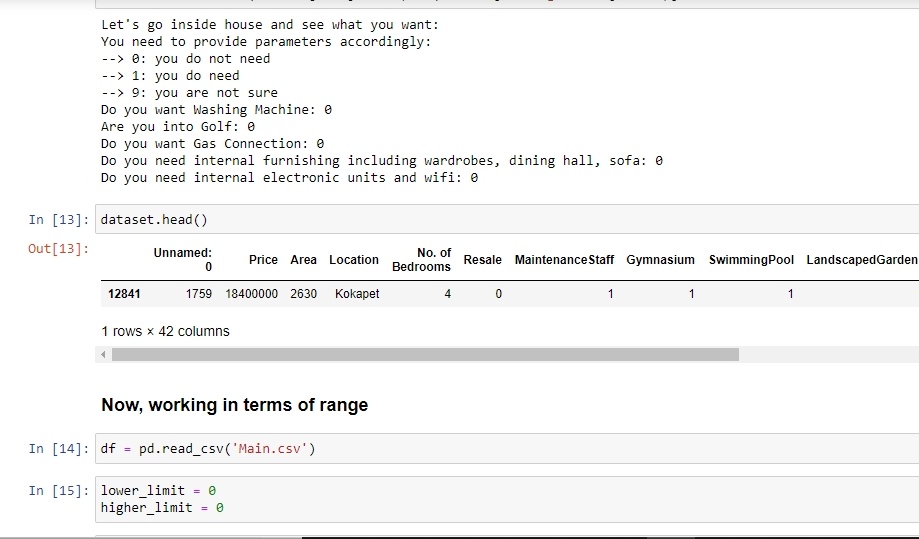
* **QUESTIONS ON THE DATASETS**

****

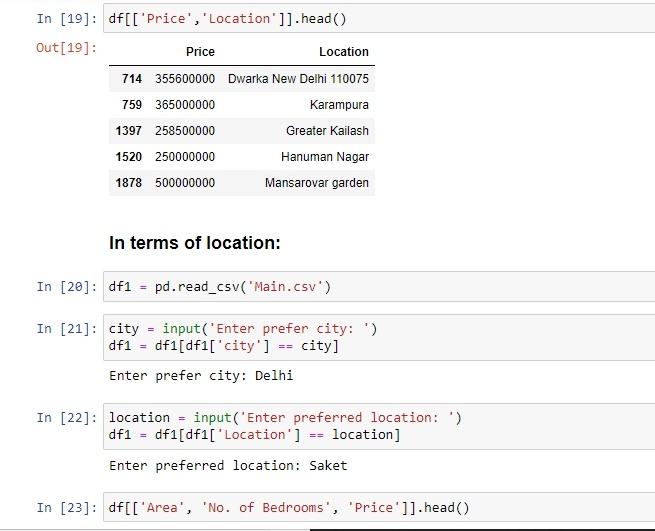
****

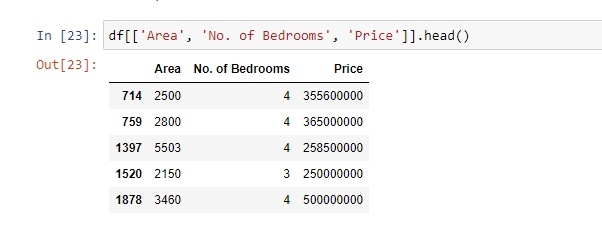
****

****

****

****

****

****