EDA and Analysis of House Prices in India

Project Report for IITB DS 203: Programming for Data Sciene

Sahil Dharod
Department of Electrical Engineering
Indian Institute of Technology, Bombay
210070026@iitb.ac.in

Azeem Motiwala
Department of Electrical Engineering
Indian Institute of Technology, Bombay
210070018@iitb.ac.in

Chanakya Varude
Department of Electrical Engineering
Indian Institute of Technology, Bombay
210070092@iitb.ac.in

***Abstract*—The following document is a report of our project which aims at analysing and predicting house prices in major cities of India. We have shown the dataset using a variety of plots and graphs, which gives the essential comprehension of all the variables affecting house prices. With the help of this EDA, we have tried to find out parameters that influence house prices and then apply appropriate ML models to predict the price.**

I. Introduction

What decides a house's price in a city? Whether it be house sellers or buyers or builders or any businesses which work in the following field need the answer to the above question. There must be an answer to this question, this is where data science and machine learning arises.

Data science would help us identify useful parameters or variables in deciding house prices. It would also help us find relations between these variables, which would further reduce redundancy. We do this by performing exploratory data analysis on the targeted data.

Machine learning would finally help us predict house prices when certain inputs of variables are given. We do this by building various models on the modified data and choosing one of these to attain maximum accuracy in the prediction. In this project we primarily worked on the following things :

- We have collected data from Kaggle, a division of Google LLC.
- To make the data ready for prediction and to identify correct relations between parameters, we have performed data cleaning and filtering by removing the outliers and unnecessary columns.
- We started with basic EDA like data statistics, and correlation heatmap and went on to make graphs and plots to analyse the impact of variables like area, construction status, number of bedrooms, etc. on the target variable (House Price) and to identify relations between different variables in the data.
- We have used basic machine learning techniques like Multiple Linear Regression, Lasso and Ridge Regression and Random Forest Regression to predict the price as accurately as possible

II. Dataset and Pre-Processing

A. Dataset

The selected dataset consists of about 30000 rows of data with some essential factors affecting house prices such as area, construction status, number of bedrooms, RERA approval, location in the form of latitudes and longitude and address. The shortcomings in our data are a lesser number of features and limited data being available after appropriate cleaning, which is done by taking all the features into consideration. We have analysed only valid and sensible data and considered the same for training in our ML models.

B. Pre-Processing

- Examining each column's datatype and null values
- Splitting the column 'Address' which is of the form ('City','Locality') into separate columns to easily analyse data for a particular city and increase correlation between different columns by avoiding variations among cities.
- We replaced the so-called cities 'Lalitpur' and 'Maharashtra' with 'Mumbai' in the column for cities as their localities matched those of Mumbai.
- We clubbed and sorted all of the cities in alphabetical order to get better insights from the data. Further, we again sorted the data according to the number of rows(data). From this sorted data, we find the top 6 major cities (major in our terms is defined according to the number of available data). These top 6 Major cities are **Mumbai, Bangalore, Pune, Noida, Kolkata, Chennai**.
- We plotted the correlation heatmaps for all of the top 6 major cities.
  Here are the heatmaps for the top 3 major cities amongst the 6:

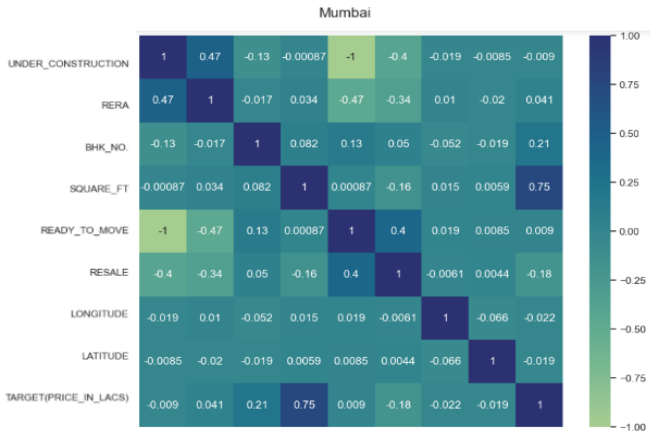


Fig. 1

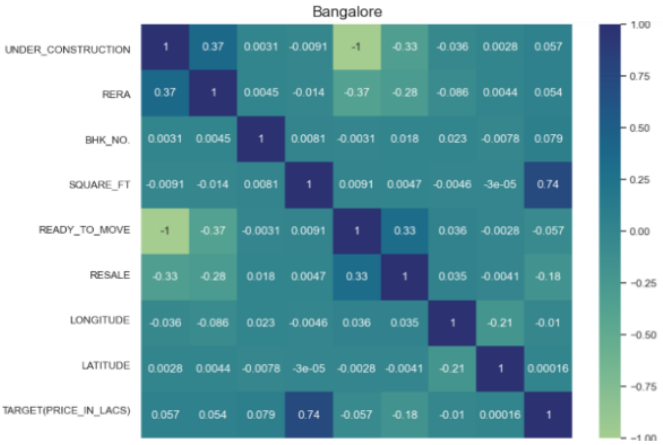


Fig. 2

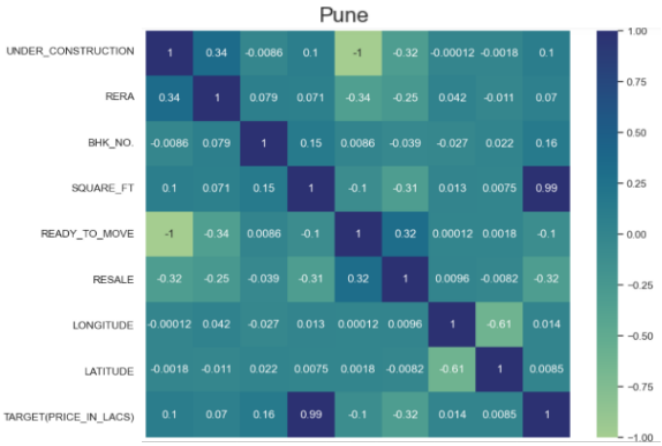


Fig. 3

We plot these individual city specific correlation heatmaps to identify correlations between variables in our data.

Conclusions:

- We conclude the variable square-ft is highly correlated to the target variable (House Prices in lacs).
- We observe the location that Latitude and Longitude are weakly correlated with all of the other variables, this is because firstly the latitude and longitude are almost the same for each city, and there is a very unclear distribution of other variables with this variable.
- We observe that the variable Ready to move and Under construction are complementary, thus to avoid redundancy we can avoid usage of one of these variables.

- We plotted the histogram and density plots for house prices, which gave us better insights into the price range of Houses. The following plot also helped us identify outliers, these were located at unusually higher prices. This was ”Part 1’ of identifying the outliers.

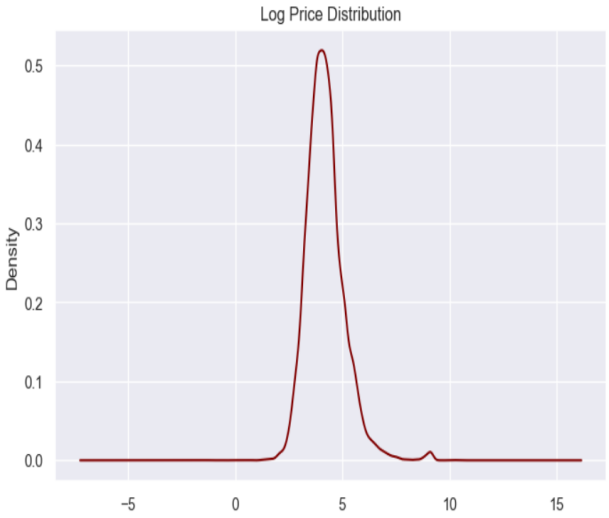
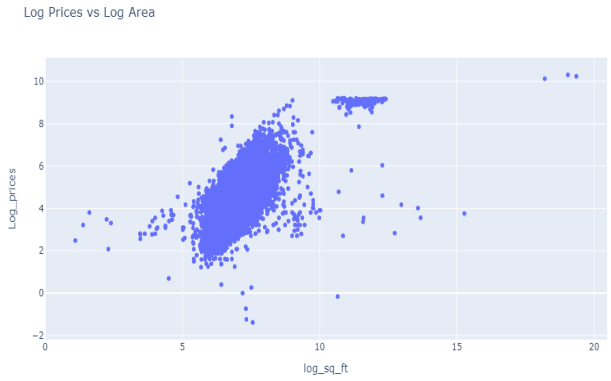


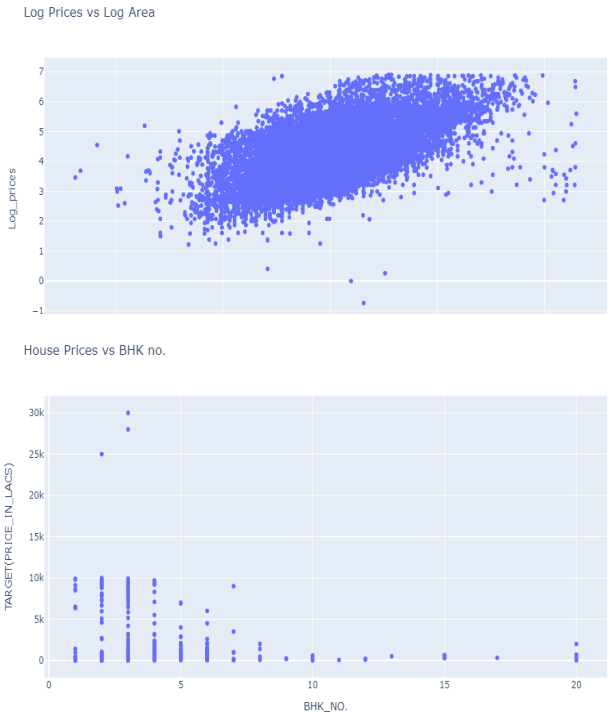
Fig. 4: Log Price Distribution

.

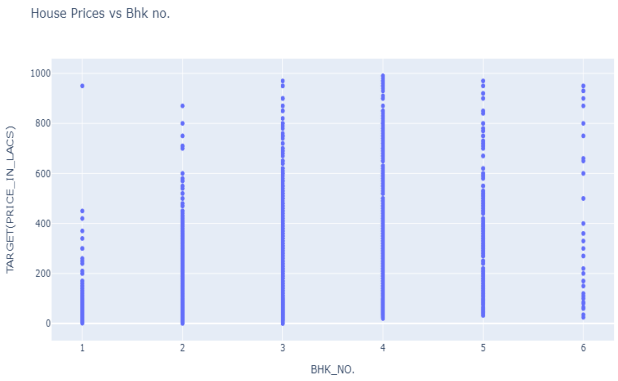
- We plotted a scatter plot of log(price) and log(area) to eliminate rows with houses having a very small or very large area (exceptions), these are nothing but the outliers of the data. We also plotted a similar plot with BHK-NO. on the x-axis and price on the y-axis and found some data which had higher number of BHK no(s). This small data acted as outliers. These outliers were cleaned from the data by putting good thresholds, decided by the plots and by logical interpretation.



(a)



(a)



(b)

Fig. 6: Price vs No. of Bedrooms

- We also used .shp files for all the cities to get their correct latitudes and longitudes. Using this information we removed rows which are out of the nearby range of the decided thresholds.
- Entropy is related to the randomness of the data being processed, that is, the higher the entropy higher is the randomness in the data (meaningful only for discrete datatypes). The higher the randomness(entropy) of a variable, the harder it is to draw conclusions from the variable. This lets us know the randomness of all of the variables in the dataset.

III. Data Visualization

Boxplot

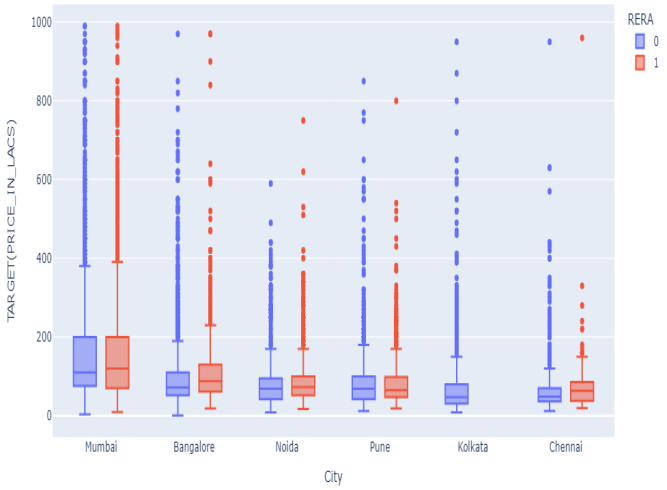


Fig. 7

Observations :

- From the boxplots, we observe that the average prices of houses that have been registered and approved under RERA (Real Estate Regulatory Authority) are higher than those which have been not, in Mumbai, Bangalore, Chennai and Noida and almost same in Pune.
- This is because getting the approval adds to compliance cost and increases the overall cost of the project by roughly 10%.

Line Plot

This is a multiple line plot showing average house price per bhk. Our dataset consists of inadequate data for 5-6 bhk for the cities Chennai and Noida.

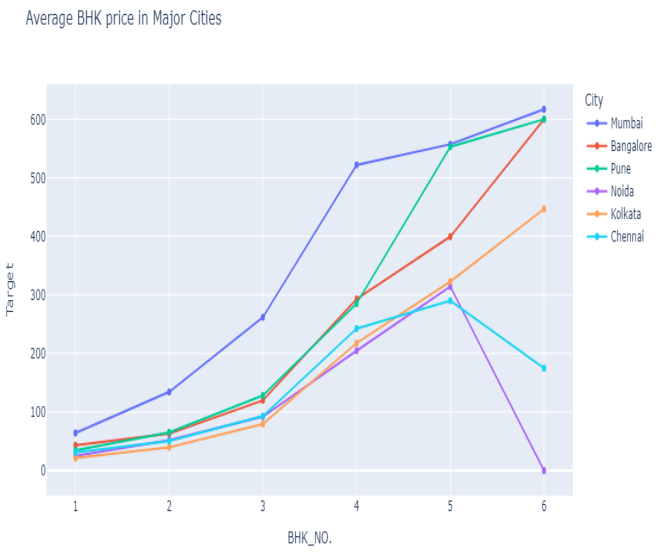


Fig. 8

Observations

- We conclude that the house prices generally increase with increase in number of bedrooms.
- We observe that house prices of the cities follows the following order : Mumbai > Pune > Bangalore > Chennai Kolkata Noida
- Due to a lack of good data in the major cities: Chennai and Noida the line plot drops down to zero or lower prices at higher BHK no(s)

Stacked Bar Plot

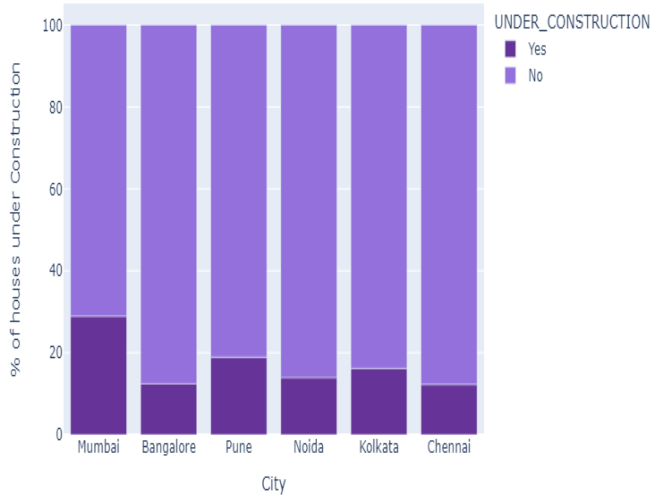


Fig. 9

Observations

- We can observe that only 13-29 % of houses are under construction in the major cities.
- Particularly Mumbai city has the highest percentage 29 % of houses under construction, and Bangalore having the least 13% as according to the given data.

Pie Plot

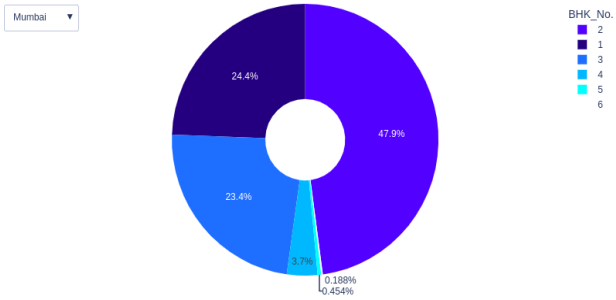


Fig. 10

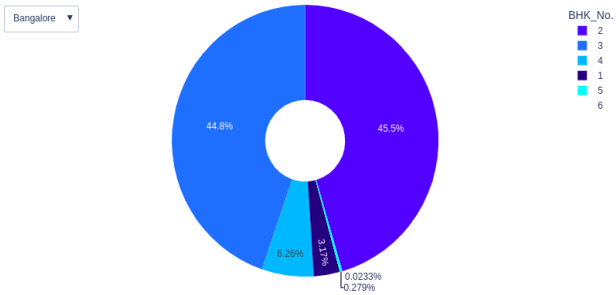


Fig. 11

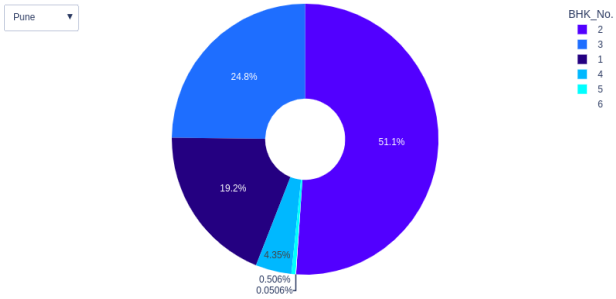


Fig. 12

Observations

- From the pie plots it can be concluded that the number of 5 and 6-BHK houses are fewer in number.
- Generally upto 85-90 % of the houses are 1,2 or 3-BHK.
- in all the cities, approximately 50 % of the houses are 2-BHK except in Noida.

- While all the cities have very few or no 6-BHK houses , Mumbai has 12 6-BHK houses.

Geographical Maps

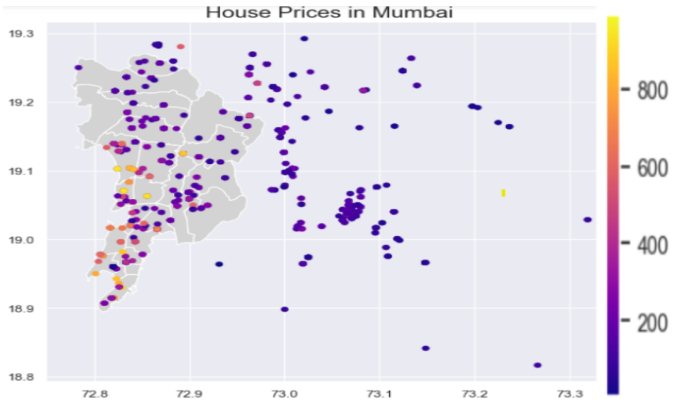


Fig. 13

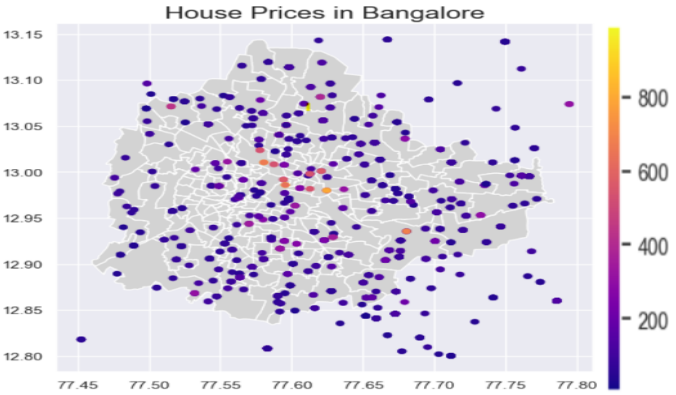


Fig. 14

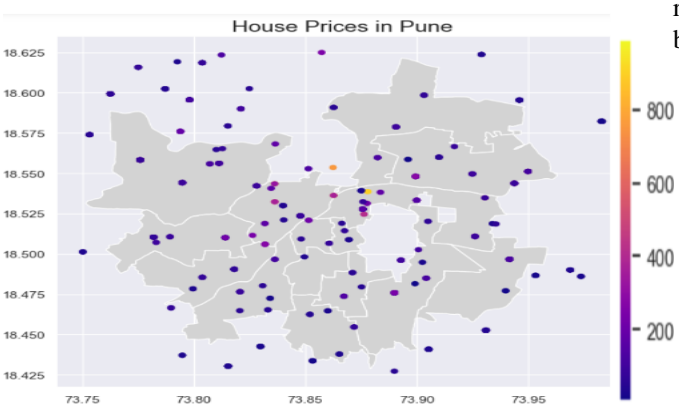


Fig. 15

Observations

- We use Geopandas to plot points on the shp file for each of the major cities, using CRS of latitudes and longitudes. The color of the points describe the house price according to the color scale given in the plot.
- We observe that in Mumbai most of the high valued houses (price) are located in the southern part of mumbai (600 to 850 lacs). Also there are points present out of the map, this is because the shp file does not contain Navi Mumbai and Thane, and houses here are generally cheaper than that in current map.
- We observe that in Bangalore most of the high valued houses(price) are located in the central part of Bangalore (600 - 800 lacs), also house prices comparatively are lower than that of Mumbai and are concentrated between 50 - 300 lacs.
- We observe that in Pune high valued houses are located in the top central part of Pune (700-850 lacs) , also that the average house prices is closer to that of Bangalore (we comment on this by looking at the color of the points). It can also be seen data for Pune is less.
- We observe that in Kolkata majority of houses lie in the range of 0 to 250 lacs, and some of the high priced houses are in the central part of kolkata having price in the range of 600-800 lacs.
- We observe that in Chennai high priced houses are located in the central right region of chennai, though less in number (because of less amount of data).
- For Noida we used the shp file of Delhi (unavailability of Noida shp file), we observe that higher priced houses are located in the west of Noida (closer to Delhi) and are less in number.

Violin Plot

This map shows the price density of houses which are for resale grouped by the person who has set the price i.e. owner, buyer or dealer.

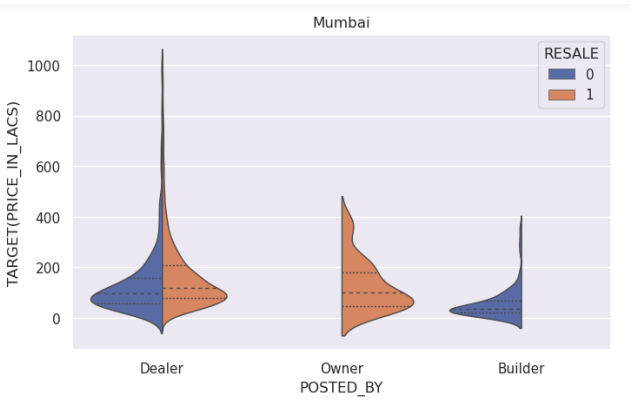


Fig. 16

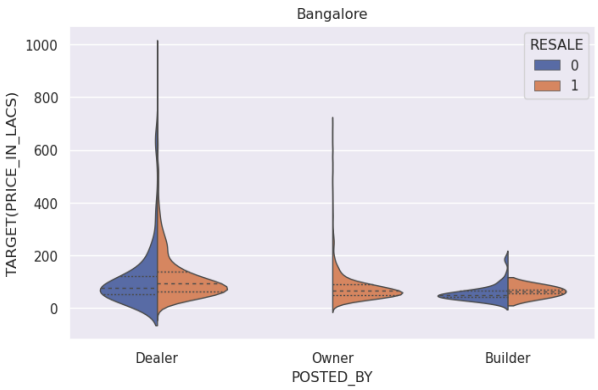


Fig. 17

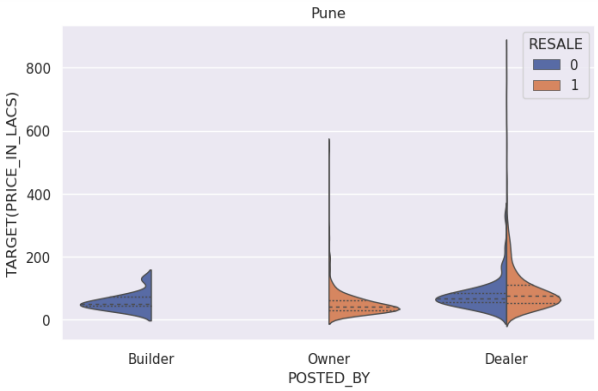


Fig. 18

Observations

- There is no or very less resale at higher prices for major cities

Bubble Plot

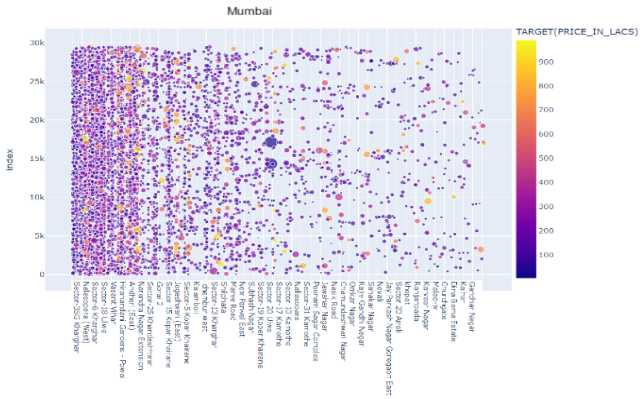


Fig. 19

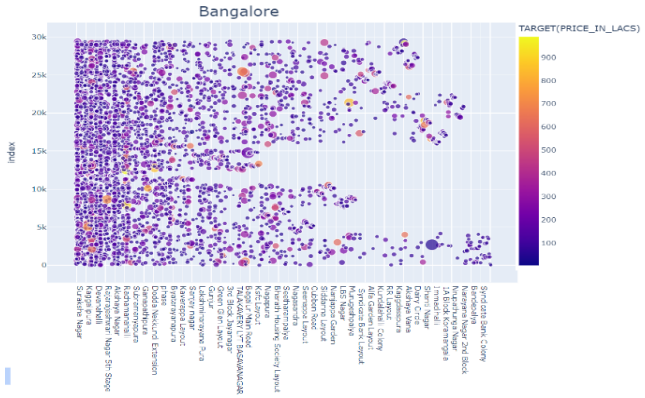


Fig. 20

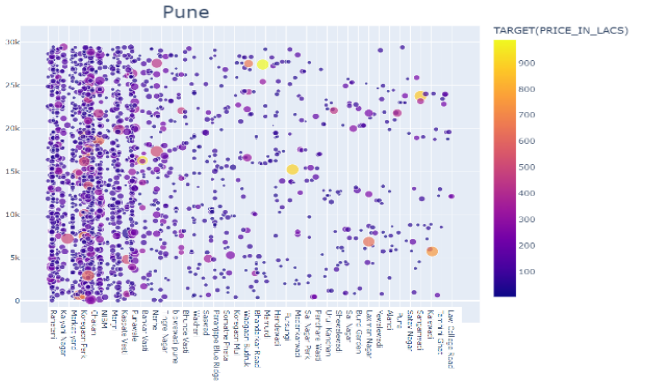


Fig. 21

Observations

- From the bubble plots, we can comment upon which localities have larger houses.
- We can also take a look at the range of prices in a specific locality which helps the buyer to select the locality as per his/her budget.
- Generally, higher the area, higher is the price of the house and house prices are highly dependent on localities.

IV. Machine Learning Models

After performing exploratory data analysis, we trained a number of machine learning models on our data set after standardising it to stop the models from concentrating on some factors over others.

Model optimisation, error functions, and decision processes are all used in machine learning algorithms. The model begins by estimating trends in our data. The accuracy of the model's prediction is measured by the error function. Then, in order to get the highest level of prediction accuracy, we can employ a variety of models or adjust the hyper-parameters of a single model. In an effort to better suit the training set's data points, the model modifies its weights. It assesses and works to make the best forecasts possible. Until a suitable

accuracy criterion has been reached, this process is repeated.

We used R-squared score in order to calculate our error. R-squared is a goodness-of-fit measure for regression models. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values. The various machine learning models used and their results are described below:

A. Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. Our analysis for the top three cities is shown in the table below :

Cost function for Multiple linear regression

$$J(\mathbf{w},b) = \frac{1}{2m}\sum_{i=0}^{m-1}(f_{\mathbf{w},b}(\mathbf{x}^{(i)}) - y^{(i)})^2$$

where:

$$f_{\mathbf{w},b}(\mathbf{x}^{(i)}) = \mathbf{w}\cdot\mathbf{x}^{(i)} + b$$

There are no Hyperparamters for Multiple Regression Model

City	R2-Score	MSE	MAE
Pune	0.97	0.012	0.7
Mumbai	0.90	0.1	0.13
Bangalore	0.76	0.23	0.10

B. Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. We have performed hyper-parameter tuning using GridSearchCV module. In hyperparameter tuning, we fed different values of hyper-parameter to our model and we obtain the best hyper-parameter which best fits the model.

Cost function for regularized linear regression or Lasso Regression

$$J(\mathbf{w},b) = \frac{1}{2m}\sum_{i=0}^{m-1}(f_{\mathbf{w},b}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \frac{\beta}{2m}\sum_{j=0}^{n-1}|w_j|$$

where:

$$f_{\mathbf{w},b}(\mathbf{x}^{(i)}) = \mathbf{w}\cdot\mathbf{x}^{(i)} + b$$

Hyperparameter = Lamda (which is also known as penalty in multiple regression) for Lasso Regression. Our analysis for the top three cities is shown in the table below :

City	R2-Score	MSE	MAE
Pune	0.91	0.07	0.19
Mumbai	0.89	0.11	0.13
Bangalore	0.71	0.29	0.11

C. Ridge Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated.Ridge regression performs L2 regularization, which adds a penalty equal to the square value of the magnitude of coefficients. We have performed hyper-parameter tuning similar to Lasso Regression.
Cost function for regularized linear regression or Lasso Regression

$$J(\mathbf{w},b) = \frac{1}{2m}\sum_{i=0}^{m-1}(f_{\mathbf{w},b}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \frac{\beta}{2m}\sum_{j=0}^{n-1}(w_j)^2$$

where:

$$f_{\mathbf{w},b}(\mathbf{x}^{(i)}) = \mathbf{w}\cdot\mathbf{x}^{(i)} + b$$

Hyperparameter = Lamda (which is also known as penalty in multiple regression) for Ridge Regression. Our analysis for the top three cities is shown in the table below :

City	R2-Score	MSE	MAE
Pune	0.98	0.014	0.05
Mumbai	0.89	0.10	0.13
Bangalore	0.76	0.23	0.10

D. Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. We have selected the hyper-parameter value (number of trees) to be 100. Our analysis for the top three cities is shown in the table below :

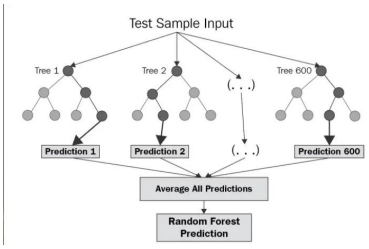


Fig. 22: Random Forest

City	R2-Score	MSE	MAE
Pune	0.94	0.06	0.05
Mumbai	0.78	0.22	0.21
Bangalore	0.95	0.05	0.06

Conclusion

- The best model for predicting house prices in Pune is Ridge Regression model as it has very high value of r2 score value compared to other models.
- The best model for predicting house prices in Mumbai is Lasso Regression model as it has very high value of r2 score value compared to other models.
- The best model for predicting house prices in Bangalore is Random Forest Regression model as it has very high value of r2 score value compared to other models.

ACKNOWLEDGMENT

We thank our professors, Prof. Amit Sethi, Prof. Manjesh K. Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan for teaching us all the Data Science techniques used in this project, and much more, and enabling us to conduct this study. We also thank our TAs for their constant efforts to solve our queries and doubts related to the assignments and concepts timely.

REFERENCES

[1] https://geopandas.org/en/stable/gallery/create_geopandas_from_pandas.html
[2] https://github.com/datameet/Municipal_Spatial_Data
[3] https://www.kaggle.com/datasets/anmolkumar/house-price-prediction-challenge
[4] https://stackoverflow.com/
[5] https://www.towardsdatascience.com/
[6] https://mindmajix.com/ridge-regression
[7] https://www.investopedia.com/terms/m/mlr.asp
[8] https://www.geeksforgeeks.org/multiple-linear-regression-with-scikit-learn/