

Predicting House Prices – Bangalore, India

By Erick Gaitán

I. Background

It is common to find announces of data science jobs in Bangalore, India on websites like [LinkedIn](#), [indeed](#), [naukri](#), [monsterindia](#) and many others.

It is known that Data Science is experiencing a surge in Jobs across the World. India is one such country that is experiencing the “data explosion” [1]. The ones who started to learn about popular programming languages such as Python and R, about databases, and machine learning and deep learning algorithms in order to become a Data Scientist will find interesting that India is hungry for data scientists, the “*Sexiest Job of the 21st Century*”.

On February 2019, there was close to 97,000 data science and analytics jobs vacant in India, which represents the 45% over 2018 [2]. This country represents a strong option for career development for many emerging data scientists.

Bangalore is considered the IT hub of the country, it has some of the top Data Science companies in India [3] looking for:

Data Science aspirants	Data Science professionals	ML or AI aspirants	Cloud Computing professionals
Equifax Gramener Fractal Analytics	Accenture Amazon Deloitte Flipkart IBM Citrix LinkedIn	Braina AI Niki.AI Dell Juniper Network ABB Absentia VR Sig Tuple	IBM Citrix Salesforce Cisco Dell 42Gears Accenture

Bangalore has to be considered by those who want to start a career as a Data Scientist (or continue it). The main purpose of the project is to give information of house prices in Bangalore by building and applying different regression models.

II. Data

First, the Foursquare API is used for getting venues data of Bangalore, India. The dataframe contains five rows.



There are two Shopping malls, a snack place, a movie theater, and a historic site around 500 meters obtained from Foursquare; this map shows the State Bank of India, Reliance Mobile Store, the Govt Unani Hospital, a book store and more places around the Hanamkonda Chowarasta.

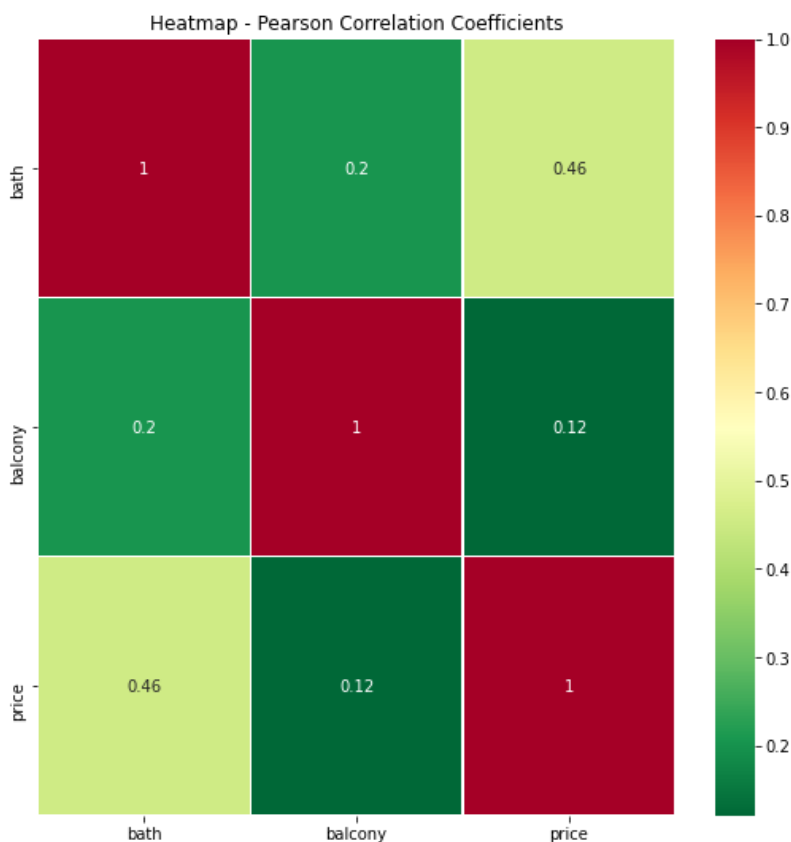
There are all kind of places near to this neighborhood, so no one has to be worried about having near hospitals or stores.

Then we have a dataset of house prices in Bangalore obtained from Kaggle. It is a CSV file that contains 13,320 rows of data. You can find the file [here](#).

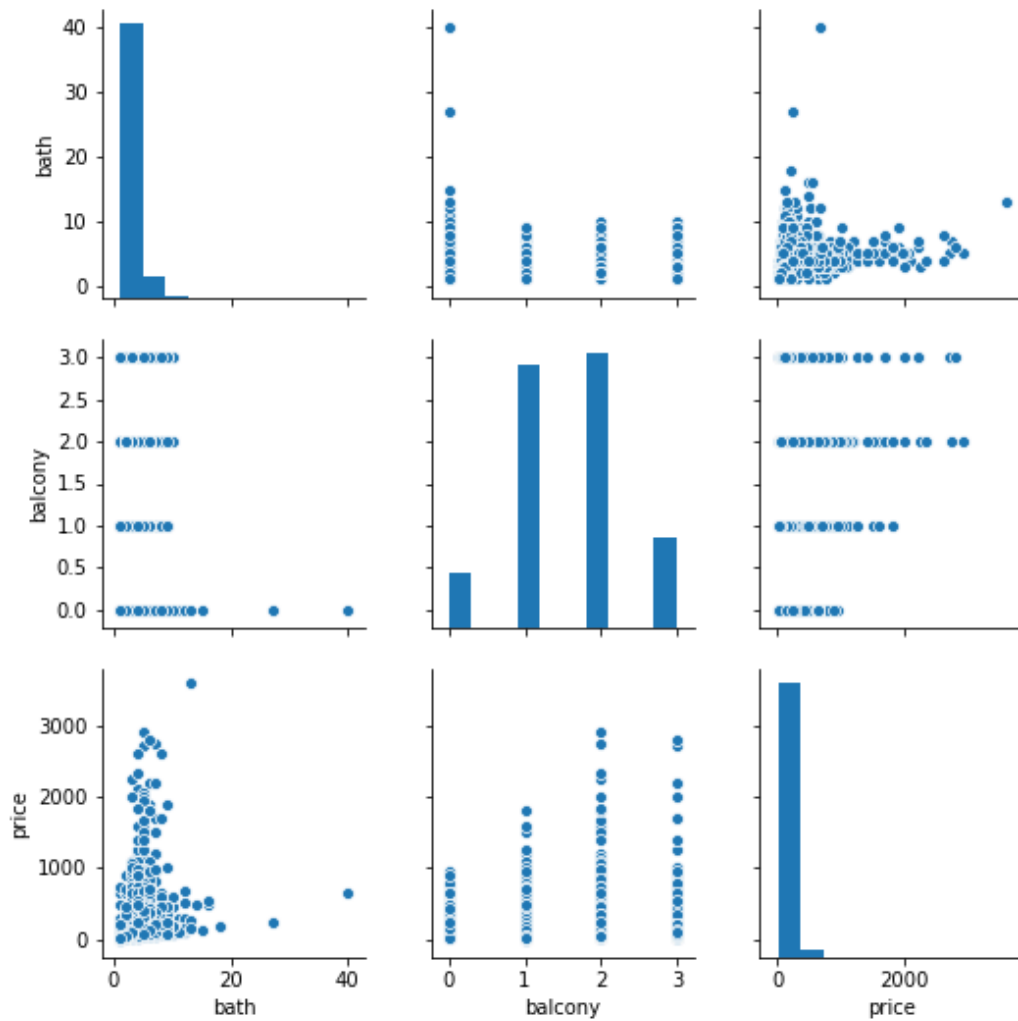
The dataframe contains nine columns, where the variable “price” will be the target. First, is necessary to remove empty values:

Column name	Data Type	#Distinct	NA Values
area_type	object	4	0
availability	object	81	0
location	object	1305	1
size	object	31	16
society	object	2688	5502
total_sqft	object	2117	0
bath	float64	19	73
balcony	float64	4	609
price	float64	1994	0

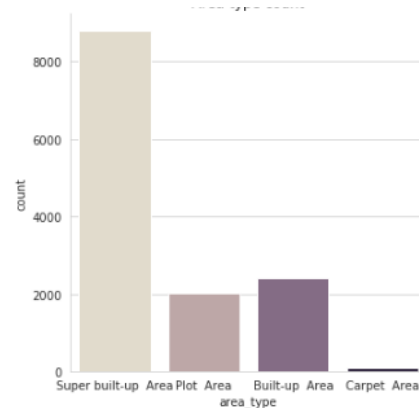
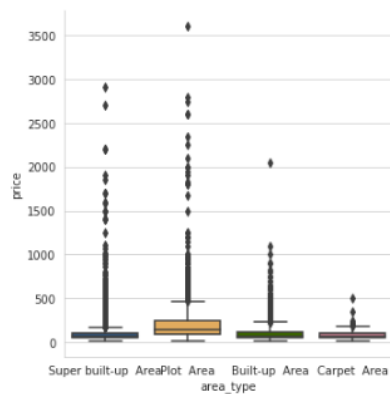
After removing those incomplete rows, the dataframe contains 7,496 rows. Now, it is moment to start describing those variables. Here is a heatmap that contains Pearson Correlation Coefficients, where it is shown that there are not strong correlations between variables.



Here are histograms and scatter plots of continuous variables. There is a positive relationship between variables bath and price.



Then we have the categorical variables, one of them is area type, which contains four unique values. Below we can see price boxplots by area type and a barplot. We can see that there are many outliers.



Here are some basic statistics of each continuous and categorical variables:

Continuous variables

	bath	balcony	price
count	13247.000000	12711.000000	13320.000000
mean	2.692610	1.584376	112.565627
std	1.341458	0.817263	148.971674
min	1.000000	0.000000	8.000000
25%	2.000000	1.000000	50.000000
50%	2.000000	2.000000	72.000000
75%	3.000000	2.000000	120.000000
max	40.000000	3.000000	3600.000000

Categorical variables

	area_type	availability	location	size	society	total_sqft
count	13320	13320	13319	13304	7818	13320
unique	4	81	1305	31	2688	2117
top	Super built-up Area	Ready To Move	Whitefield	2 BHK	GrrvaGr	1200
freq	8790	10581	540	5199	80	843

The target variable is price as is mentioned before. The features are used to build different regression models to predict house prices in Bangalore. The best one is selected by comparing various evaluation metrics of each model.

References

- [1] DATAFLAIR TEAM. (2019). Is there any Scope of Data Science in India? Take Expert's Opinion. Retrieved from <https://data-flair.training/blogs/scope-of-data-science/>.
- [2] Press Trust of India. (2019). 97,000 analytics, data science jobs vacant in India: Study. Retrieved from <https://www.thehindubusinessline.com/economy/97000-analytics-data-science-jobs-vacant-in-india-study/article26399660.ece#>.
- [3] Manipal ProLearn. (n.d.) Data Science Companies In Bangalore. Retrieved from <https://www.manipalprolearn.com/data-science-companies-in-bangalore>.