To Owner of Real Estate,

Greetings and hope you are doing great with good health!

Respected owner, I have implemented different techniques of Data Cleaning, Data Analysis, Data Visualization, and Machine Learning algorithms on the real estate data you provided for the analysis. There were various anomalies in your data like ‘total square feet are not in the proper format, it is given in a range (2223-1807) like this. Another feature ‘size’ is in various forms for the same purpose (2 BHK or 2 Bedroom) both have the same meaning but in different forms and there are many other issues also which are addressed during the analysis. Below are the following issues that were focused during the analysis.

Firstly, I have filtered some features. There are some features which do not have too much effect on the price like ‘availability’ and ‘balcony’. Most of the values for availability are ‘Ready to Move’ and some are given with dates (17-July) that have no impact on price. Secondly, the ‘balcony’ has three values (1, 2, and 3) which do not have too much effect on the price therefore these features are removed from the evaluation.

Secondly, I focused on total square feet. Is this a problem if the square feet are given in a range? Yes, it is, to know how this is an issue I calculated the price per square foot for a different location and add new columns in the data. As shown in the picture below. Now I checked the length of range some time it is around 500. To tackle this problem I replaced it with an average of the given range. However, it is even then can give a huge loss because if I replace it with the mean and it is very close to the top range then it makes a huge amount.

Let Given range is (1125-1594), then it will be replaced by mean then it will be 1359. If the actual value is 1593 then you have a loss of 234sqft. As the below graph shows most of the data fall in the range of 5000 per square foot. So, the above cost will be around 12 lac, which is a huge amount.

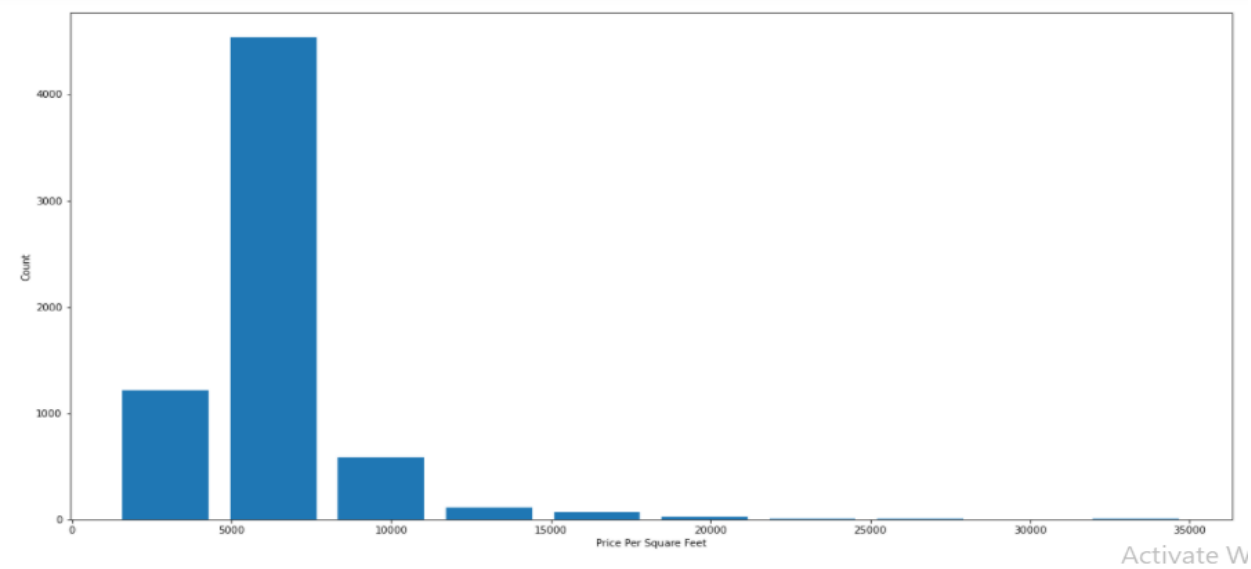

Figure 1. Price per Square foot and count

Another issue that I found in your data is about the number of rooms in a specific area. In figure 2, you can see that there is some record in which there are 43 rooms and 40 bathrooms in just a 2400sqft area, which I think is very close to impossible. I have filtered such data based on the condition that is total square foot per bedroom is less than 300 then removed it. You can reconstruct the building or can include some rooms to others by removing the middle wall to get rid of reconstruction costs.

	location	size	total_sqft	bath	price	bhk
1718	2Electronic City Phase II	27 BHK	8000	27.0	230.0	27
4684	Munnekollal	43 Bedroom	2400	40.0	660.0	43

Figure 2. Information about bedroom on a given area.

The next issue is related to the number of bathrooms. As shown in figure 3 below, there are some data in which the number of bathrooms is greater than the number of rooms which is not required. So, for such data, you can convert these extra bathrooms to rooms or include them in the attached room to increase the area.

	location	size	total_sqft	bath	price	bhk	price_pr_sqft
448	BTM 1st Stage	9 Bedroom	3300.0	14.0	500.0	9	15151.515152
984	Chikkabanavar	4 Bedroom	2460.0	7.0	80.0	4	3252.032520
2785	Nagasandra	4 Bedroom	7000.0	8.0	450.0	4	6428.571429
3188	Sathya Sai Layout	6 BHK	11338.0	9.0	1000.0	6	8819.897689
9289	other	3 BHK	1806.0	6.0	116.0	3	6423.034330

Figure 3. Information about extra bathrooms

Another issue I found in the data is about price. There are some data in which 3 bedroom area is cheaper than that of 2 bedrooms on the same location. If the provided data is correct then you can increase the rate of those with a higher number of rooms. Because, if a person is buying 2 bedroom apartment for the same price then definitely he/she will buy 3 bedroom apartment with a higher price.

After doing the above analysis on data I have applied machine learning algorithms on the above data and it gives around 80% results. There are more than 1300 locations in the data but the provided number of data is less than 10 for many locations. So, to improve the accuracy please collect some more data for such location.

Thanks and regards,

Irfan Ali