IAO Written Portion:

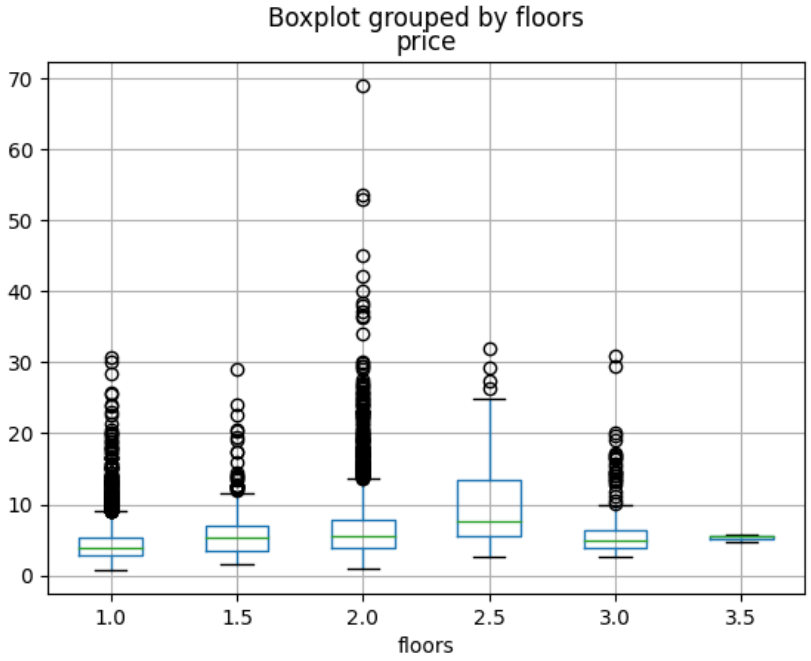
(a) Is it a good idea to use this feature (ID) in predicting the price of the house? Why?

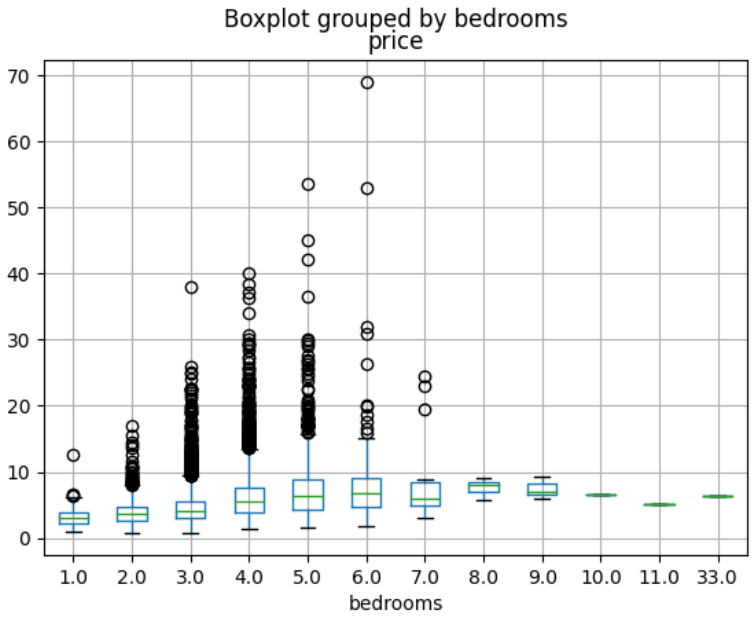
If ID's were assigned based on some factor there *may* be a reason to keep them (such as time that the house was built), but in general IDs shouldn't be factored in as they are not related to the actual object of interest. In this case, I think it makes sense to drop them since even if the IDs have some correlation to time, we should already capture that relationship with the date entry.

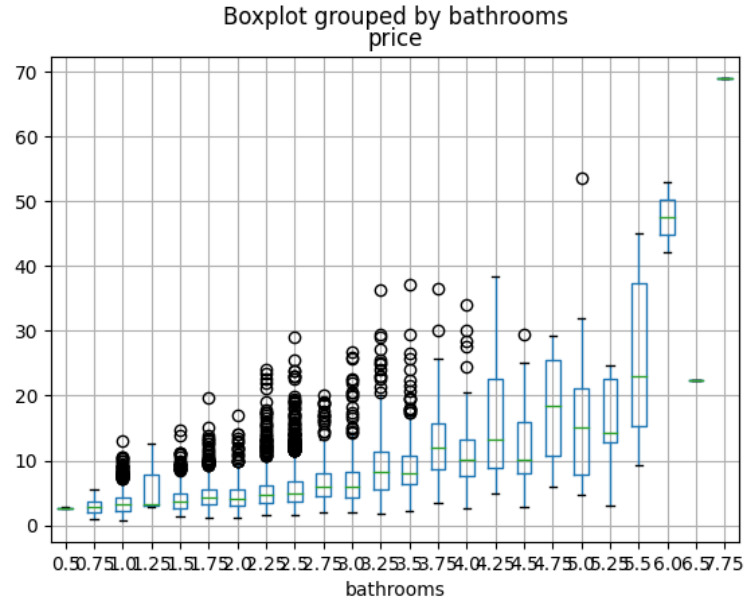
(b) Do you think the date feature is useful for this problem? Can you think of better ways of using this date feature than splitting them into three numerical features?

Dates in general can be useful in price estimations because they show the age of the thing you're trying to classify. However this requires an extra abstraction since the date does not directly relate the age. A possible better solution would be to compute the age based on the date, and replace the date with that value.

(c) Box plots:

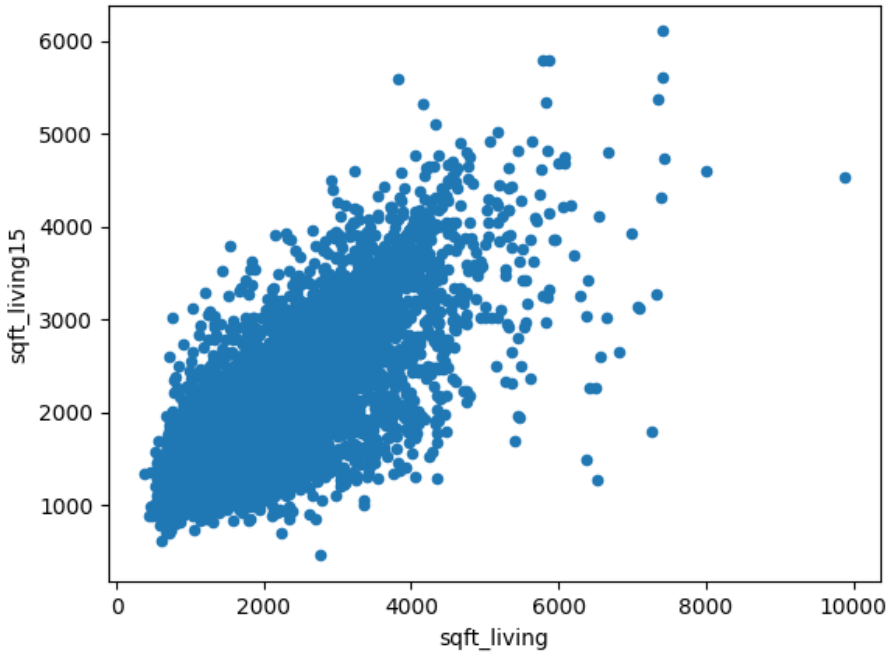


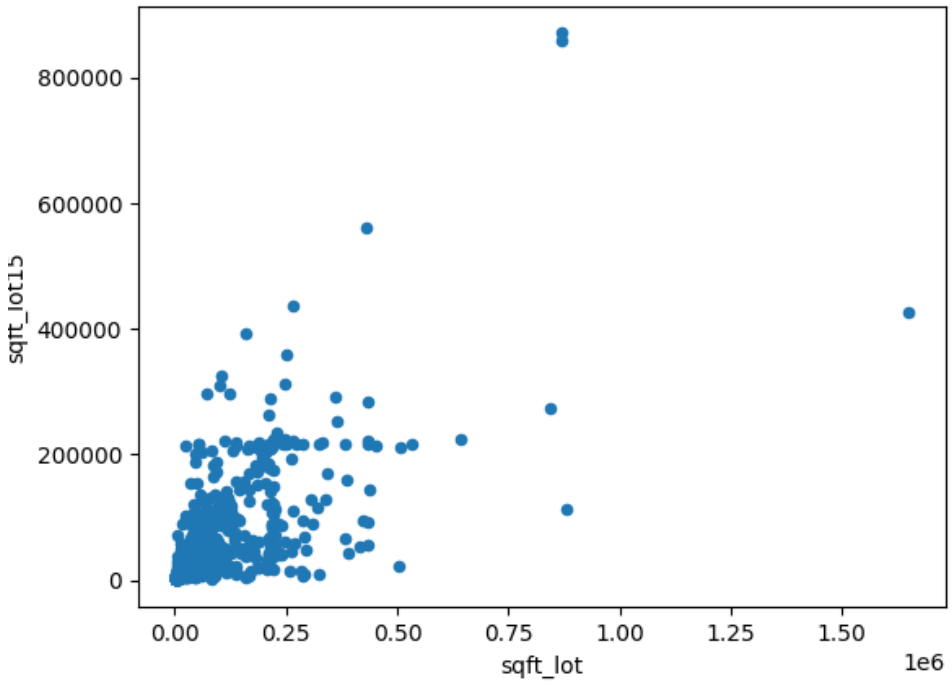




(d) What do you observe from the scatter plot? Are these features redundant?

So the features seem to be positively correlated with each in a general sense. This would normally lead me to want to drop them. But I think both are spread enough to justify their inclusion. If we needed to drop one. I would drop sqft_living15 since it's a bit tighter than sqft_lot.


6000
5000
4000
3000
2000
1000
sqft_living15
0
2000
4000
6000
8000
10000
sqft_living


800000
600000
400000
200000
0
sqft_lot15
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e6
sqft_lot