(a) One of the features is the ID of the entry. Please remove this feature from the data. This will reduce the feature by one.
Question: Is it a good idea to use this feature in predicting the price of the house? why?

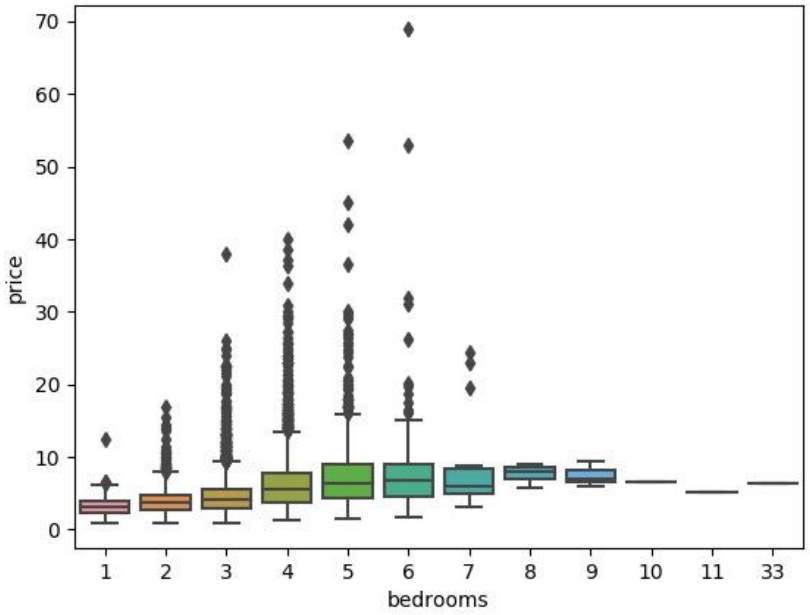
Answer: No, it is not a good idea to use this ID feature. The ID feature is only the "stamp" of each data entry, which provides no solid information. You can even use any other arbitrary numbers to replace it.

(b) One of the features the date. Please split the date feature into three separate numerical features: month, day, and year. This will increase the dimension of the data by two.
Question: 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?

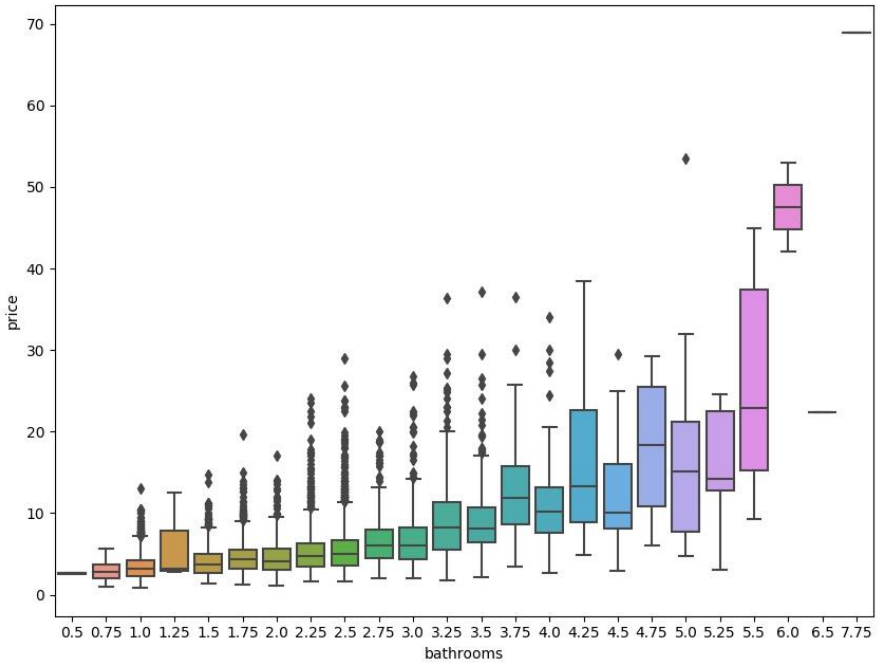
Answer: I think the date feature is useful for this problem. Another way to use date feature is to convert date to timestamp. Timestamp is only a sequence of characters (numbers), so it can be directly visualized or manipulated for many different purposes.

(c) Several features are listed as numerical but have only a small number of discrete values. This includes "bedrooms", "bathrooms" and "floors". For each of these three features, identify the unique values that are observed (you can use the unique function from Pandas). For each of the 3 features, generate a boxplot (you can use the boxplot function from Pandas) of the price. Specifically, you should generate one plot for each feature grouped by the feature value.

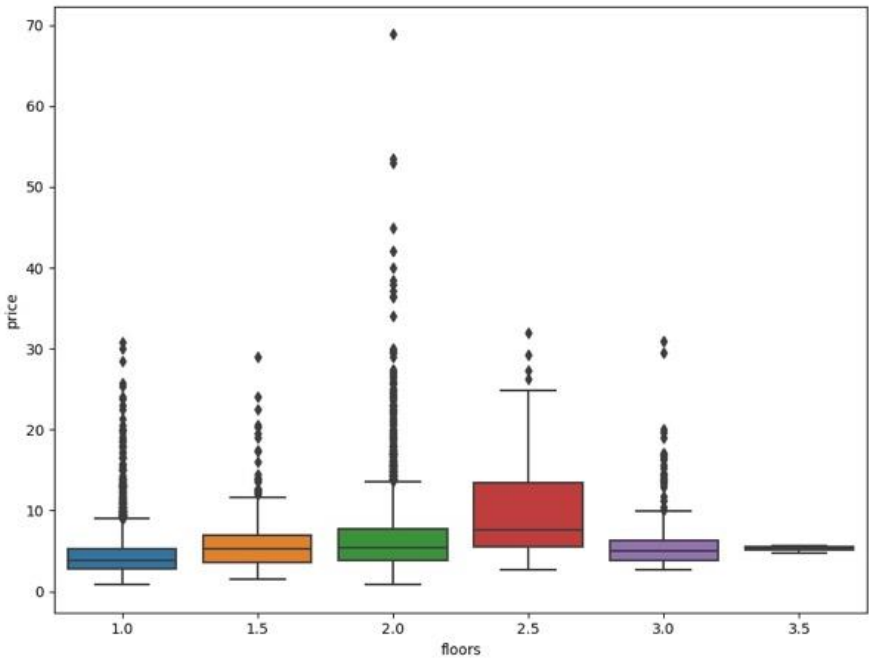
Answer:
Bedrooms:



Bathrooms:



Floors:

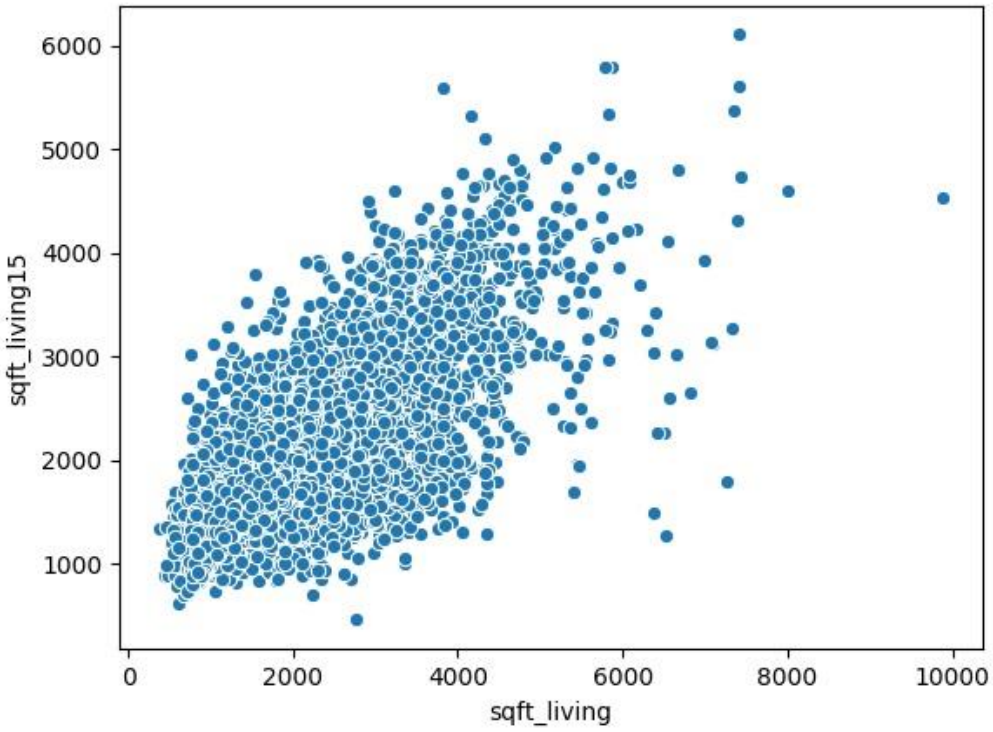


(d) Consider the following numerical features: sqrt_living, sqrt_lot, sqrt_living15, sqrtot15. Calculate the co-variance matrix of these four features. Generate a scatter plot of the data using sqrt_living and sqrt_living15, and another scatter plot using sqrt_lot with sqrt_lot15.

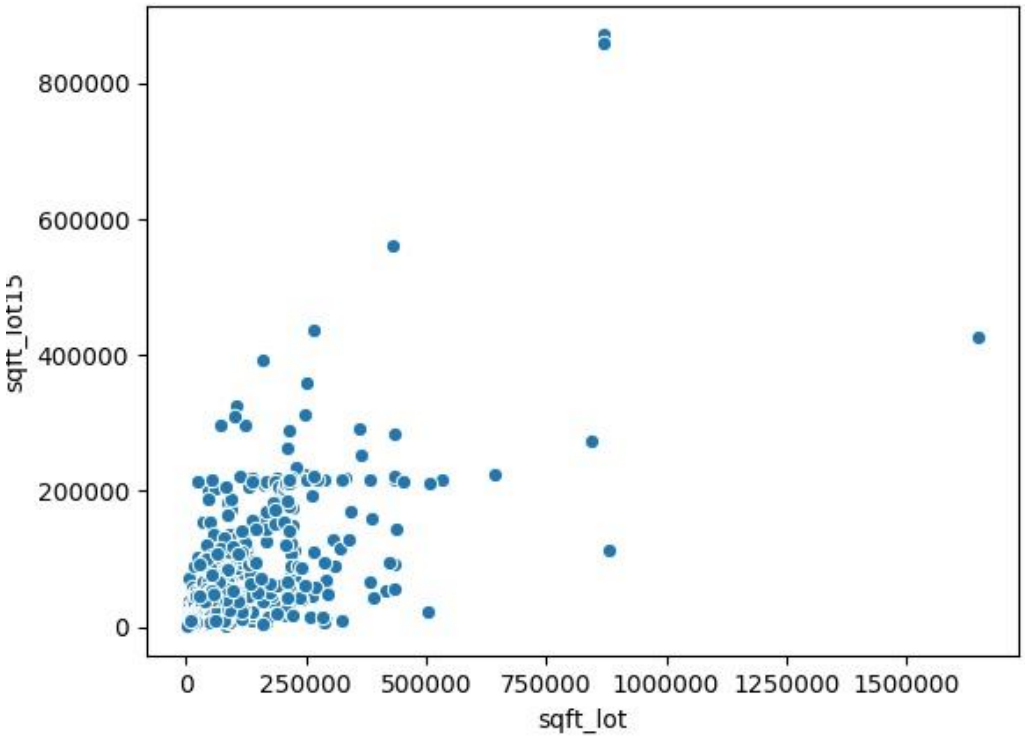
co-variance matrix:

	sqft_living	sqft_living15	sqft_lot	sqft_lot15
sqft_living	8.305303e+05	4.839029e+05	6.473942e+06	4.836731e+06
sqft_living15	4.839029e+05	4.787260e+05	4.160910e+06	3.568584e+06
sqft_lot	6.473942e+06	4.160910e+06	1.697761e+09	8.924357e+08
sqft_lot15	4.836731e+06	3.568584e+06	8.924357e+08	7.975678e+08

sqrt_living vs sqrt_living15:



sqrt_lot vs sqrt_lot15:



Question: what do you observe from the scatter plot? Are these features redundant?

Answer: In both scatter plots, the “x values” and “y values” are positive correlated, i.e. the “y values” will increase while “x values” increase. Although it is not very clear in the second plot, it is due to far away outliers. The observation can also be explained in the covariance matrix, where the covariance values are both positive for sqrt_living vs sqrt_living15 and sqrt_lot vs sqrt_lot15.

Theoretically, these features can be considered as redundant because they are positive correlated, then you can kick out one feature from each feature pair. However, in real cases, you’ll never know if a redundant feature can contribute to the model training. Thus, I’ll probably try to not abandon these features at the beginning, then train the model to see the performance differences comparing to the model trained without redundant features. Then if the results show that the feature is indeed redundant, I’ll abandon it for sure.