Name: Chengxu Xu

AI534

IA0

a)

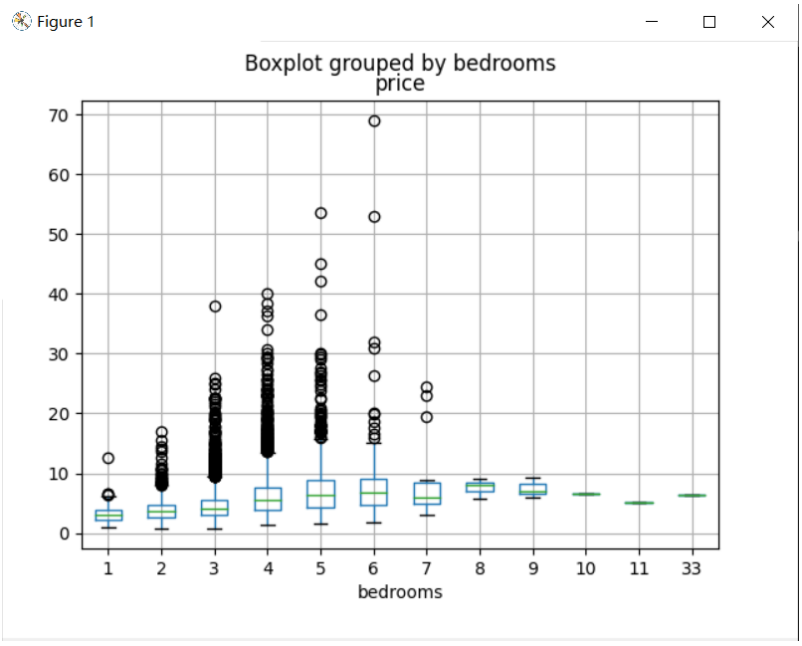
I don't think it's a good idea to use ids to predict house prices because every house has a different id and the ids are only there to distinguish each house from the others, so it doesn't have any effect on the price of the house.

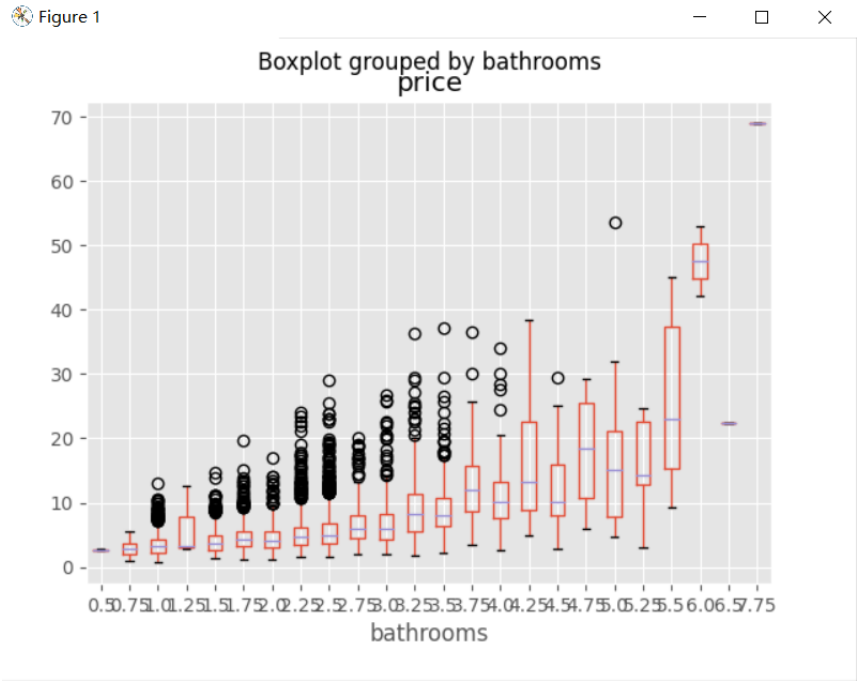
b)

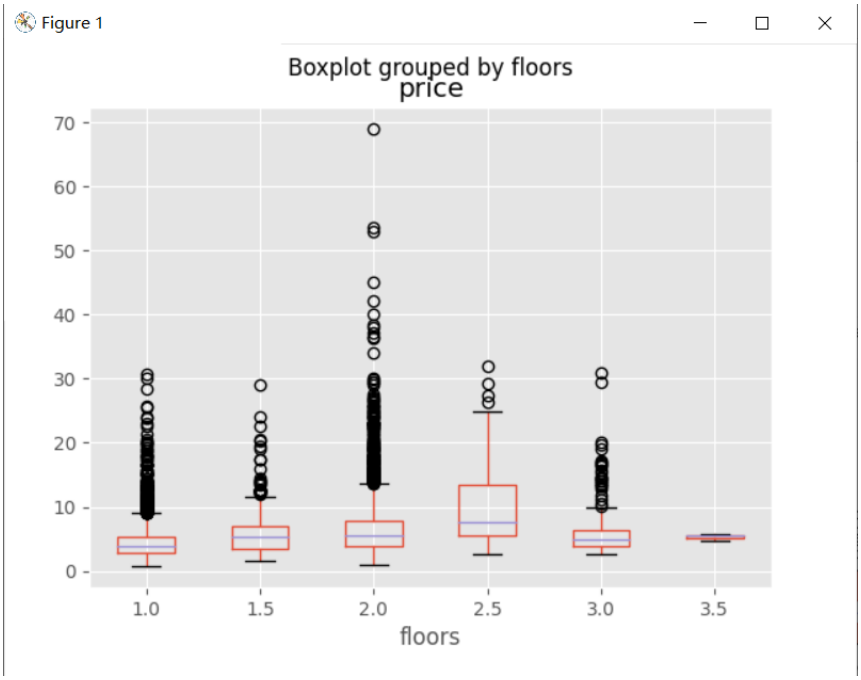
I think date feature is useful for this problem because the year of the house is a very important reference factor for predicting the price of a house, for example those houses that are relatively old tend to have more potential problems and so they will be relatively cheaper.

I think this is a good way to break down the date data, but in the case of houses, dates down to specific days don't make much sense and add to the calculations. So, I would just keep the year of the house to predict the value of the home.

c)







d)

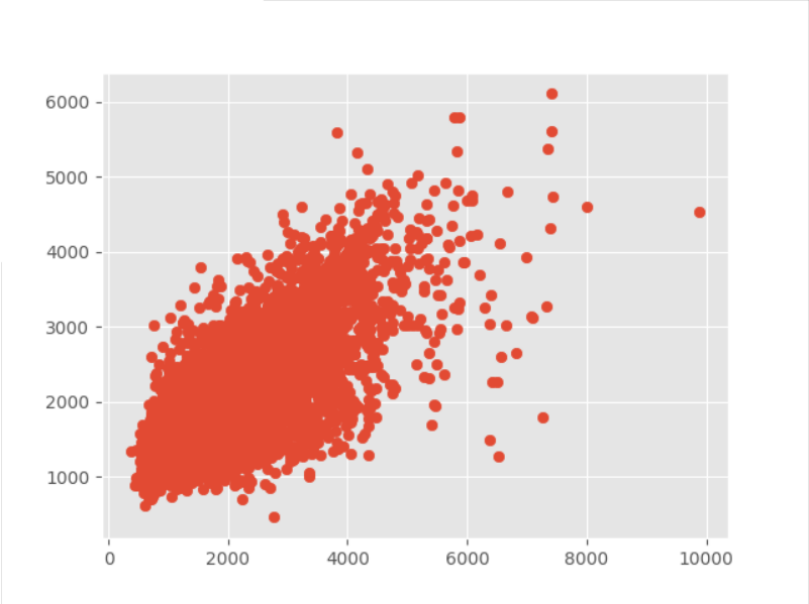
The co-variance matrix of these four features are:

[[8.30530313e+05 6.47394178e+06 4.83902913e+05 4.83673113e+06]

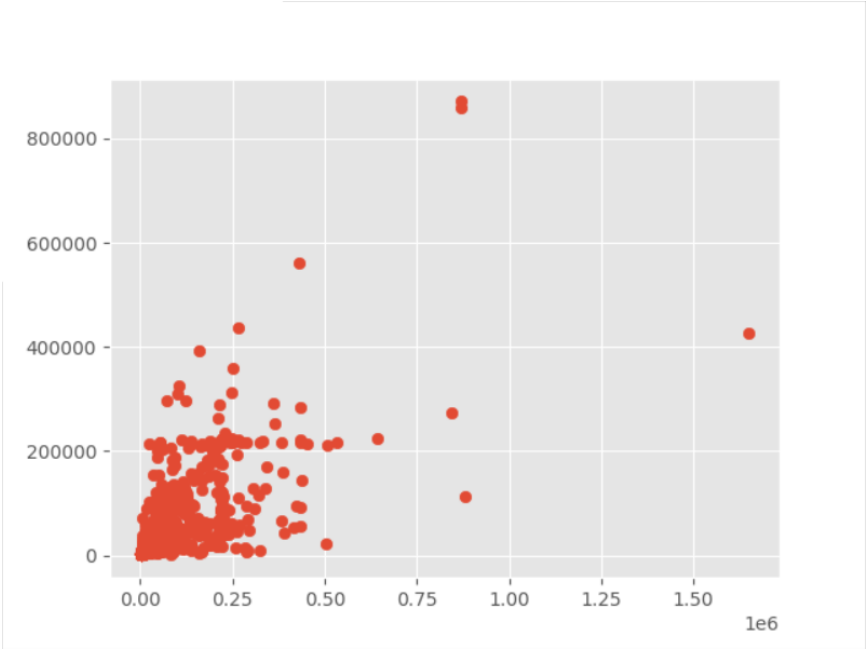
[6.47394178e+06 1.69776096e+09 4.16090960e+06 8.92435675e+08]

[4.83902913e+05 4.16090960e+06 4.78726027e+05 3.56858420e+06]

[4.83673113e+06 8.92435675e+08 3.56858420e+06 7.97567809e+08]]



From the scatter plot, it is observed that for sqrt_living with sqrt_living15, the data are mainly concentrated in the interval of (0,6000), which means that the data are more closely associated in this interval.



While the distribution status of sqrt_lot with sqrt_lot15 is mainly distributed in the range of sqrt_lot15≤200000, which is perhaps the maximum value of sqrt_lot in general, and the connection between sqrt_lot and sqrt_lot15 is closer in this interval. But there are also some pairs with errors due to noise.

These features are heavily redundant within a given interval.