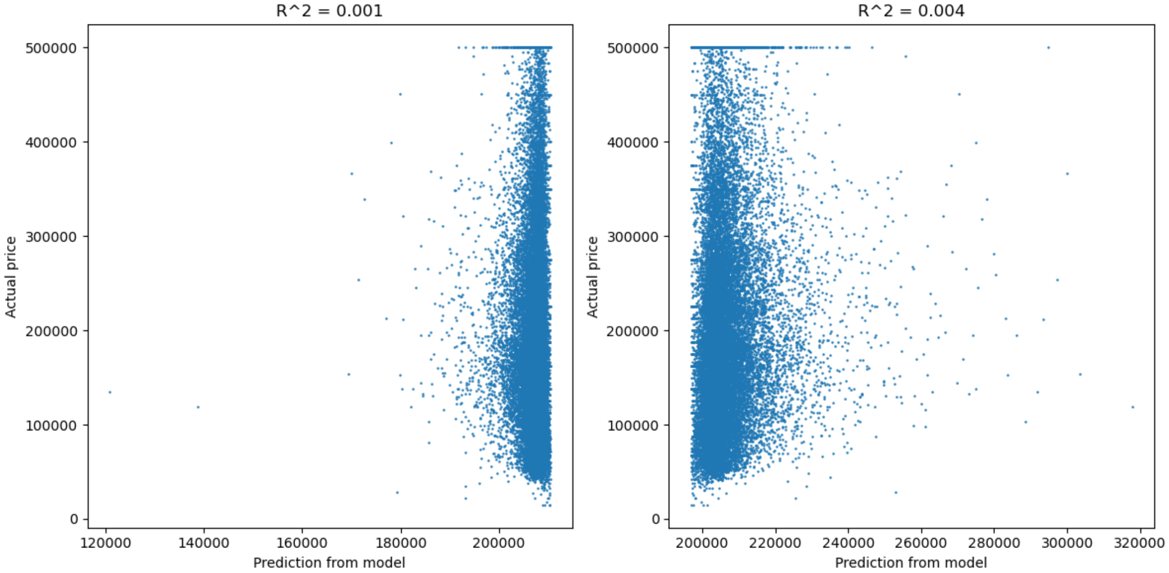
Homework 1

1. Two simple linear regression models are built using variables 4 or 5 to predict the median house price. R^2 values of these two models are calculated and the predicted prices are plotted against the actual prices.

Variables 2 and 3 largely reflect how big the whole block is, but according to common sense there usually is no clear relationship between the size of the whole block and how expensive each house in the block is – big and small blocks can both be cheap or expensive. But if we normalize them we may get a better understanding of how big the average individual house is likely to be, and the size of each house definitely determines its own price along with other factors. Variables 4 and 5 are not very useful by themselves for similar reasons – they are mainly decided by how big the whole block is, as larger blocks can contain more people and households, which does not reflect the value of the houses well.

R^2 when using population as predictor is about 0.001, and R^2 corresponding to number of households is about 0.004. The plot is below.

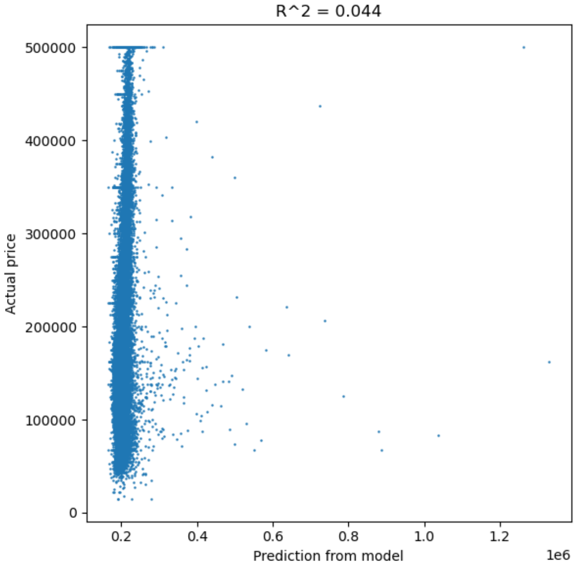


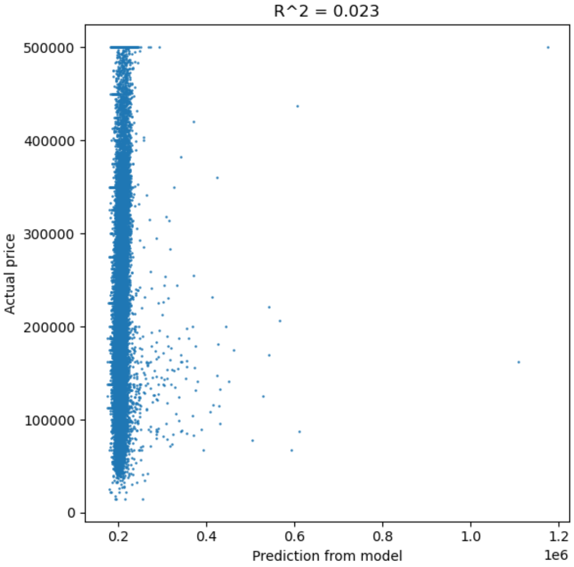
It can be seen that R^2 values are extremely small for both variables, so they don't work well in predicting. From the plots it can be seen that the simple linear regression models predict the price of most houses to be around 200000, while in reality it ranges from 0 to 500000 with a rather uniform distribution. Thus variables 4 and 5 are not very useful.

2. For each of the variables 2 and 3, two simple linear regression models are built, one using the variable normalized by population as the predictor and the other using the variable normalized by households to predict median prices. The R^2 values of the two models are calculated and the predicted prices are plotted against the actual prices.

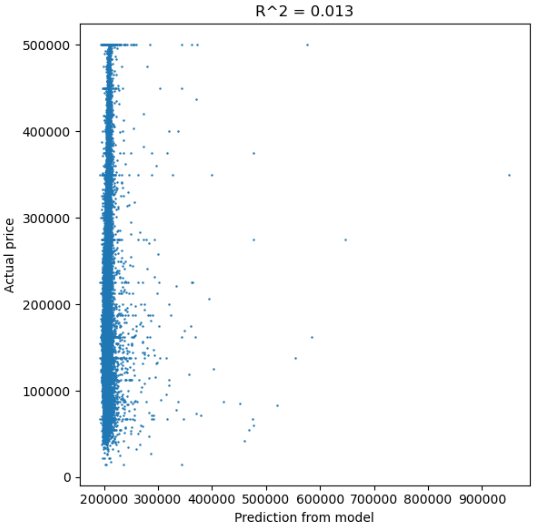
It's better to divide by the number of households because the ideal way to measure the average size of each house is to divide the number of rooms in a block by the number of houses in a block, but we don't have the number of houses. The number of households is a better approximation of the number of houses because in most cases each household would occupy one house (in rare cases two poor households may share a house or a family may have some guests living with them, but these should be negligible), so these two numbers should be roughly the same. Dividing by population shows how crowded the block is, but population density is not particularly useful in predicting how expensive an area is.

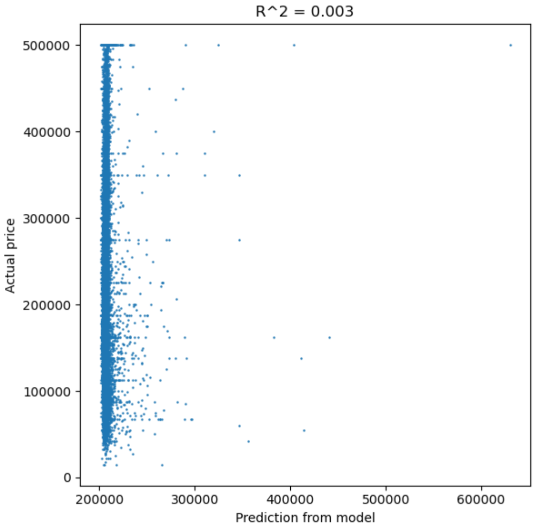
For total rooms, normalizing it by population yields a R^2 value of 0.044 and normalizing it by households has R^2 value of 0.023.





For total bedrooms, normalizing by population gives R^2 of 0.013 and normalizing by households produces R^2 value of 0.003.



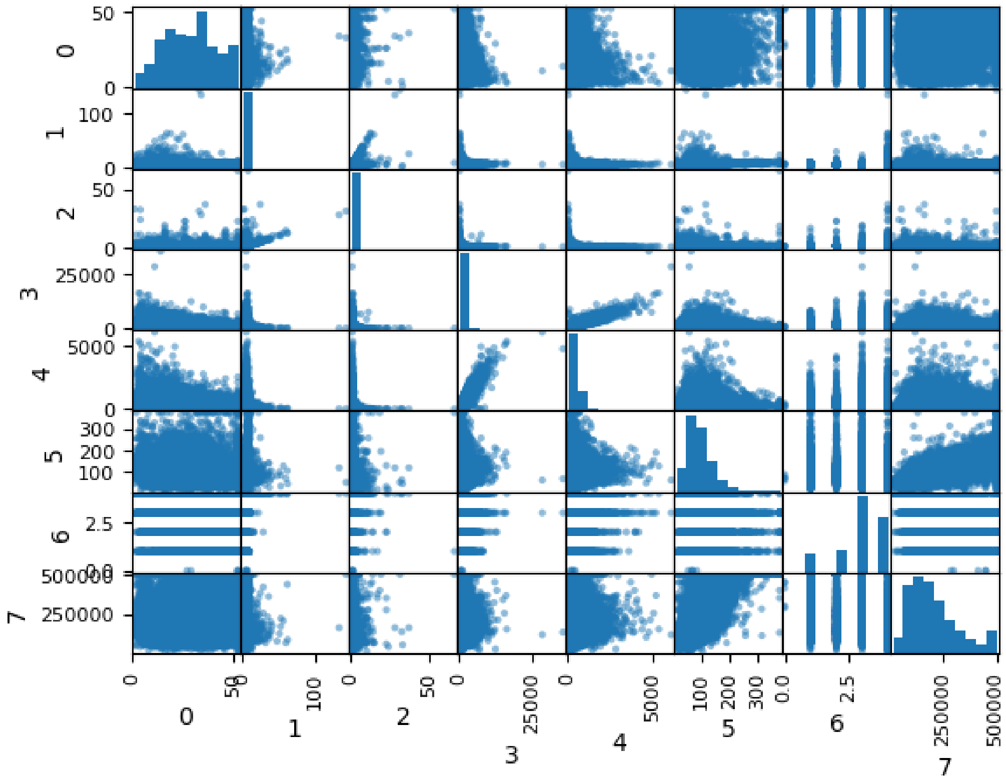


The results seem to imply that normalizing by population is better in both cases. But considering that the R^2 values are very small in all four cases suggesting a bad fit overall, and the fact that the plots are actually similar, I believe that the small discrepancies in R^2 values are not enough to override the original rationale, thus we should still rationalize by households.

3. All other variables that have not appeared so far (variable 1, 6, and 7) are used as predictors in simple linear regression models and their R^2 values are calculated. A scatterplot is drawn, and the last row is closely inspected because it plots the median price on the y-axis against all the variables on the x-axis.

The variable with the highest R^2 value should be the best predictor in a simple linear regression model. The scatterplot should be able to confirm this result by showing a relatively clear linear relationship existing between the variable and median prices.

R^2 for variables 1, 6 and 7 are respectively 0.011, 0.473 and 0.158. Combining with the R^2 results of previous questions, variable 6 (median income) has the highest R^2 value and variable 4 (population) has the lowest. The scatterplot is below.



Variable 6 (median household income) is the best predictor. This is very reasonable as people with higher incomes tend to live in more expensive areas. I think it won't be more effective even if we have more data on the number of houses, as this is already a median income and does not require normalizing. Variable 4 (population) is the worst predictor, for the reason explained in question 1.

4. A multiple regression model is built using all variables as predictors and the R^2 value is calculated.

This should normally enhance the R^2 of the models as we take more information into account. But it could cause some problems like collinearity.

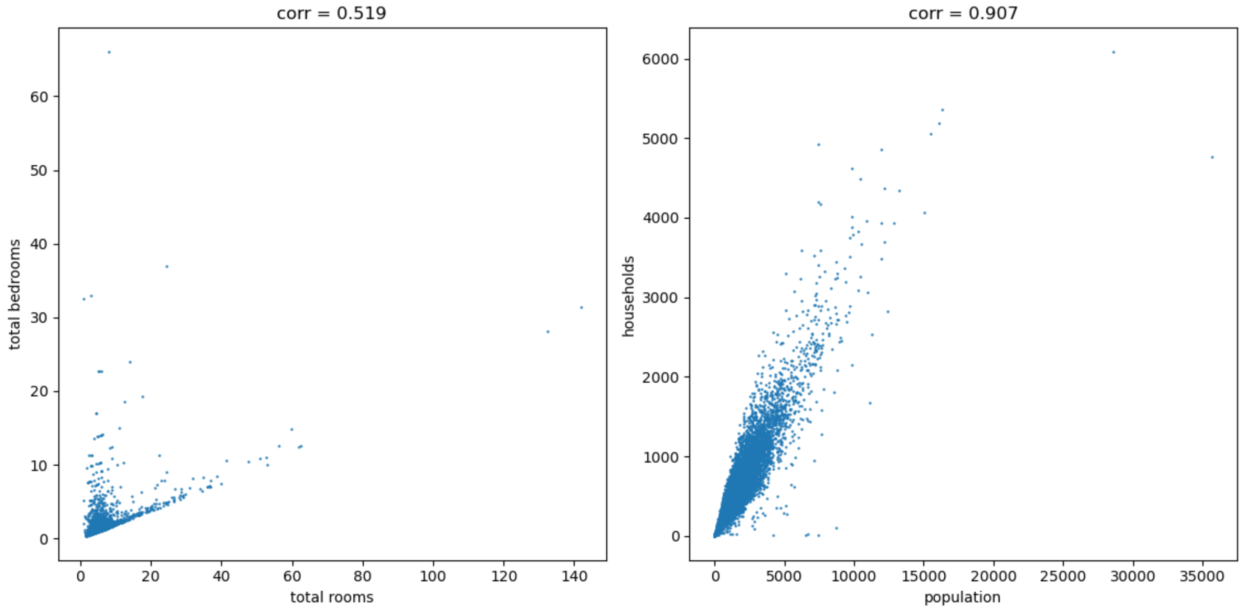
The R^2 value of the multiple linear regression model is 0.599.

The multiple linear regression model is better than the single best predictor (R^2=0.473).

5. The correlation coefficient between variables 2 and 3 and that between variables 4 and 5 are calculated.

If the correlation coefficient is high then there is likely to be collinearity, this should be confirmed by an observable linear relationship in the scatterplot shown in question 3. Generally bigger houses have both more rooms and more bedrooms, and more families (households) mean more people in total, so collinearity is expected to exist.

Correlation coefficient between variables 2 and 3 (standardized) is 0.519, for variables 4 and 5 it is 0.907. The scatterplot shows a slanted V-shaped relationship between 2 and 3 (left), and a clear linear relationship between 4 and 5 (right).



There is some level of correlation between 2 and 3, but the exact relationship is kind of complicated and whether it leads to collinearity may require further investigation, while for variables 4 and 5 there is no doubt that they are highly correlated and exhibit collinearity.

Extra Credit

(a) I think variable 1 (housing median age) may be normally distributed because on the scatterplot its histogram is higher in the middle part and has 2 lower tails. It is also somewhat symmetric.

(b) The outcome variable is heavily right skewed. This may mean that the houses in the survey tend to be cheaper ones.