

Why is it a good idea to standardize/normalize the predictor variables 2 and 3, and why are predictor variables 4 and 5 probably not very useful by themselves to predict median house values in a block?

I calculated the correlations between predictors 2 and 3 and the outcome using numpy in order to see whether predictor variables 2 and 3 have a linear relationship with the outcome. The correlations of predictor variables 2 and 3, and median house value in the block, are 0.134 and 0.134 respectively (rounded to the thousandth place). This shows a very weak linear relationship, which has the potential to be fixed by standardization. In addition to this, fitting a simple linear regression separately using predictors 2 and 3 and a 60/40 training/test split both gave R^2 s of 0.019, which is very low, showing that predictor variables 2 and 3 are not good linear predictors of median house values in a block unnormalized, and only marginally better than just predicting the mean median house value per block.

I calculated the correlation between the predictor variables 4 and 5 using numpy in order to see whether they might cause issues with collinearity. Predictor variables 4 and 5 have a correlation of 0.907, meaning that they are very highly correlated. They measure basically the same trait, so using both predictor variables 4 and 5 is very similar to using only one of them. Also, fitting a simple linear regression separately using predictors 4 and 5 and a 60/40 training/test split gave R^2 s of 0.0006 and 0.005, showing that they are not good predictors.

To meaningfully use predictor variables 2 (number of rooms) and 3 (number of bedrooms), you will need to standardize/normalize them. Using the data, is it better to normalize them by population (4) or number of households (5)?

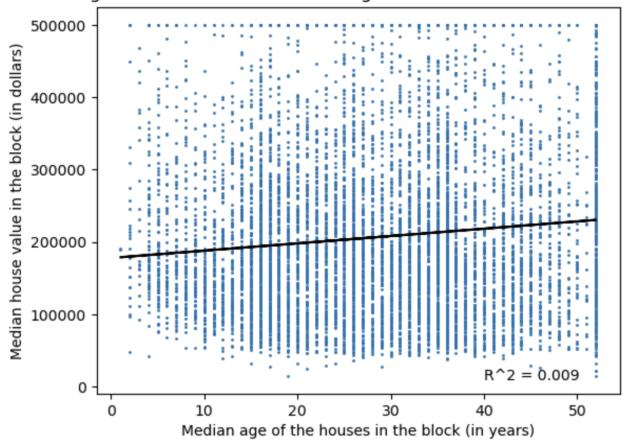
I calculated the R^2 value for predictor variables 2 and 3 normalized by population and number of households and the outcome variable in order to test which models predicted better using the same method as the previous question. This gave R^2 s of 0.026 and 0.010 when normalizing by population and 0.019 and 0.003 when ordering by number of households, showing that it is better to normalize by population. This makes sense, as the size of a household can vary greatly and the total amount of rooms per household makes a household with 2 rooms and 2 people look the same as a household with 2 rooms and 10 people, which likely affects the value of the house. All of these R^2 values are low, but normalizing by population gives higher ones than unnormalized.

Which of the seven variables is most *and* least predictive of housing value, from a simple linear regression perspective?

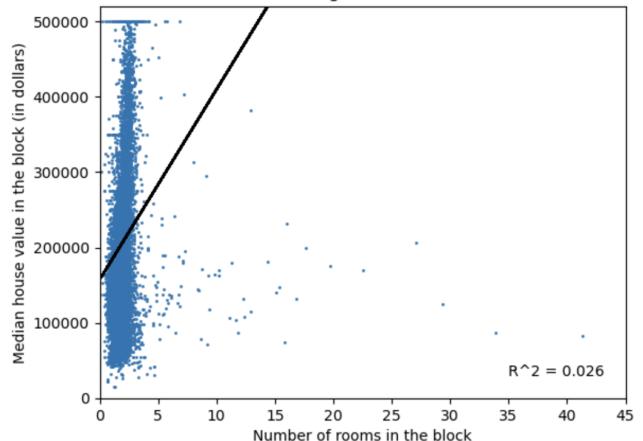
- Make sure to inspect the scatter plots and comment on a potential issue –would the best predictor be even more predictive if not for an unfortunate limitation of the data?

I graphed each of the seven predictor variables against the outcome variable and calculated their R^2 values in order to find which variable is most predictive of housing value, as well as check for any other possible issues using the graph. I did this using a 60/40 training/test data split. This gave these:

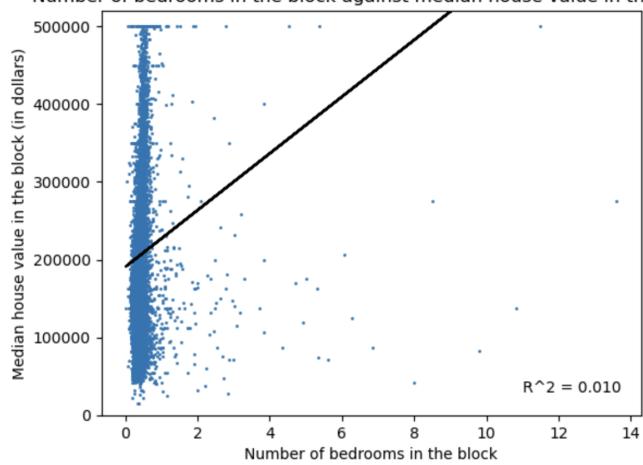
Median age of the houses in the block against median house value in the block



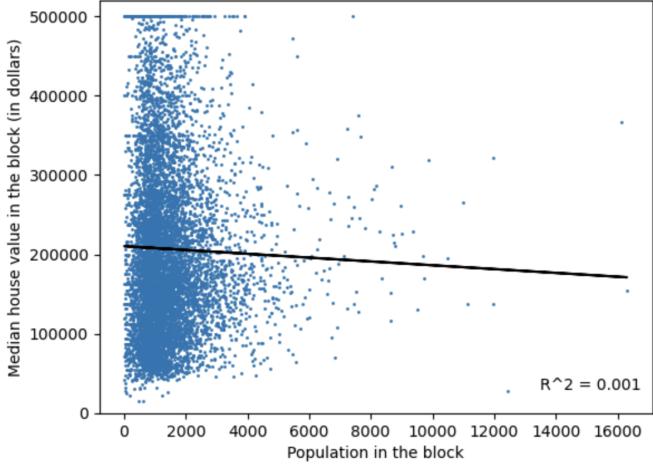
Number of rooms in the block against median house value in the block



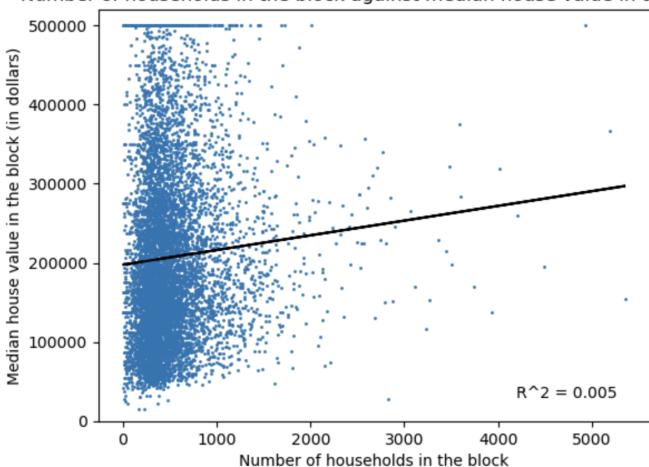
Number of bedrooms in the block against median house value in the block



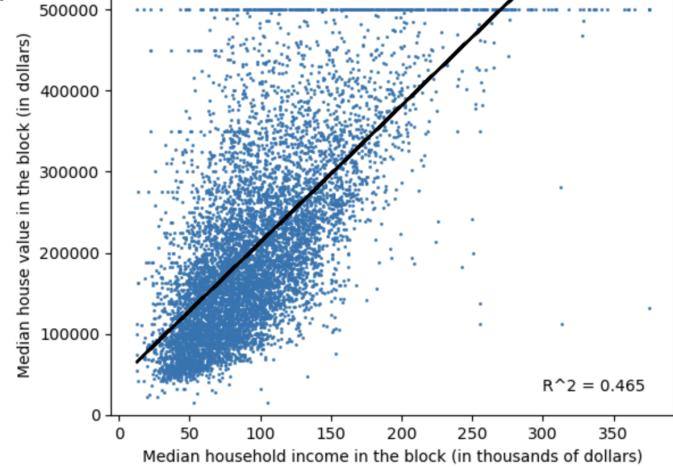
Population in the block against median house value in the block

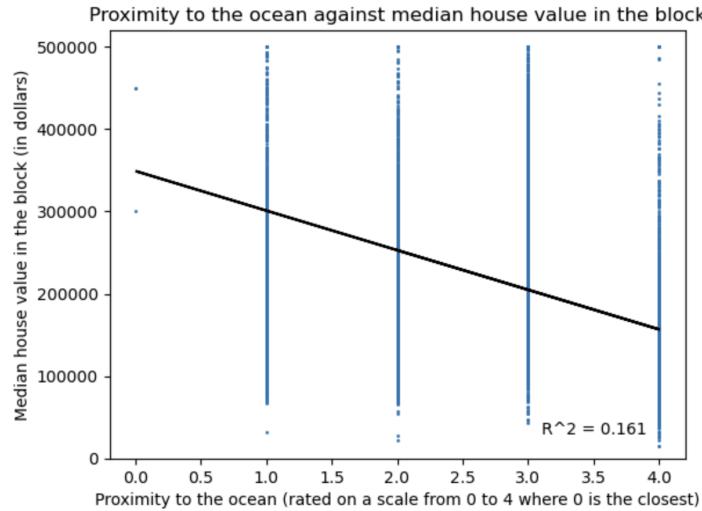


Number of households in the block against median house value in the block



Median household income in the block against median house value in the block

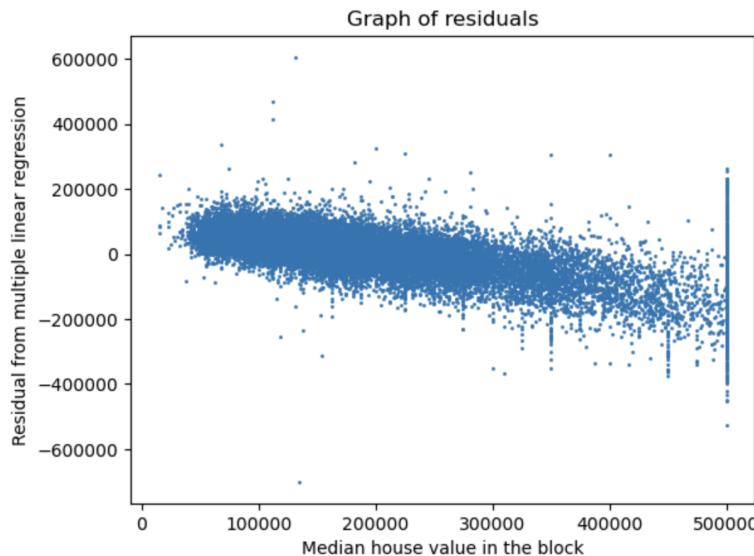




The highest R^2 comes from the model predicting using median household income in the block, which shows that it is the best predictor of housing value. The lowest R^2 comes from the model predicting using the population in the block, showing that it is the least predictive of housing value. This dataset appears to place all median house values above \$500,000 at \$500,000, which decreases the predictive ability of every model.

Putting all predictors together in a multiple regression model –how well do these predictors taken together predict housing value? How does this full model compare to the model that just has the single best predictor from 3.?

I fit a multiple linear regression using a 60/40 training/test data split in order to create a multiple regression model. Together, the predictors can predict median housing value in the block moderately well, with an R^2 of 0.551. Graphing the actual values against the residuals gives the following graph. From the graph of the residuals, there is a linear shape. This suggests that the data is not suited to linear regression. This model is better than the single best predictor from the previous question, but not by a very large amount.



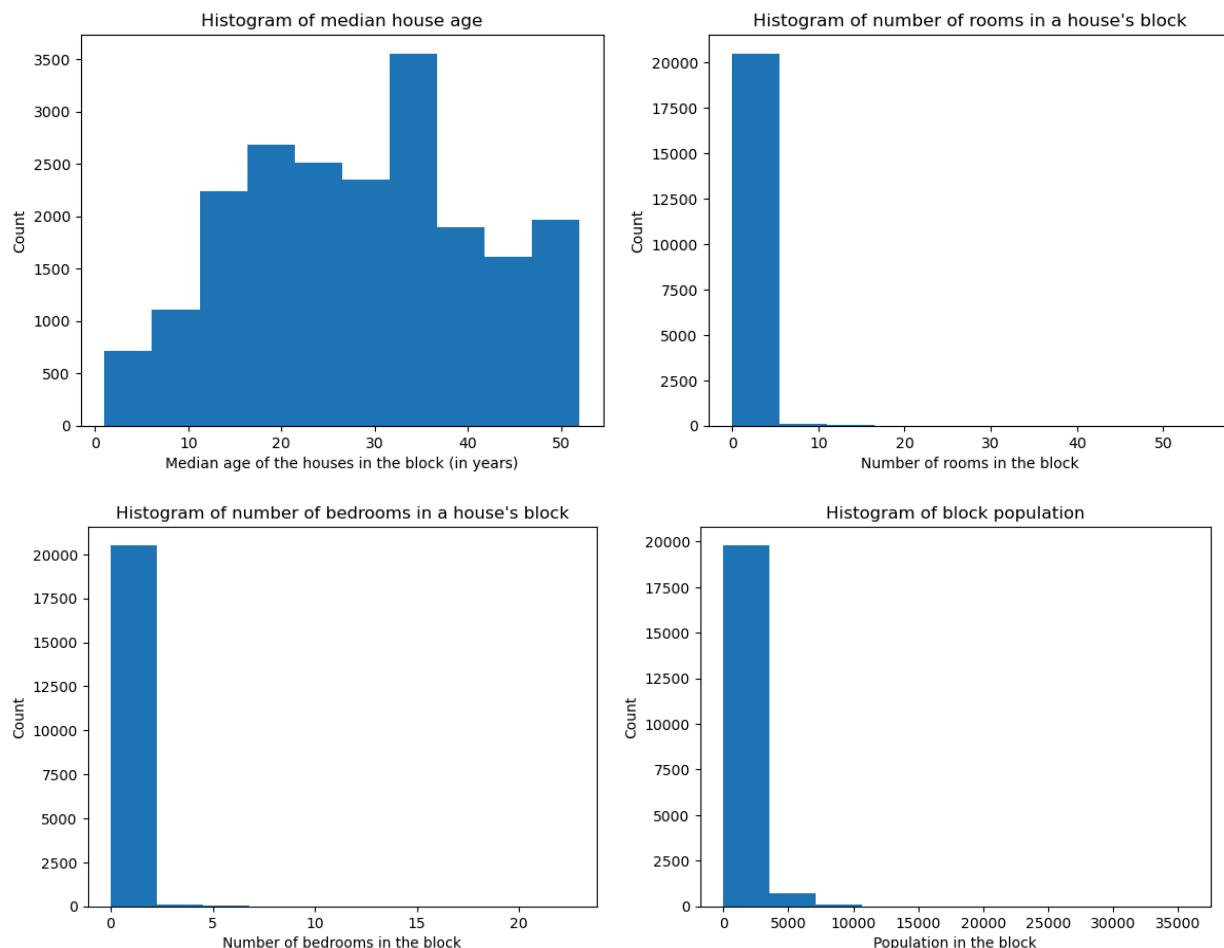
Considering the relationship between the (standardized) variables 2 and 3, is there potentially a concern regarding collinearity? Is there a similar concern regarding variables 4 and 5, if you were to include them in the model?

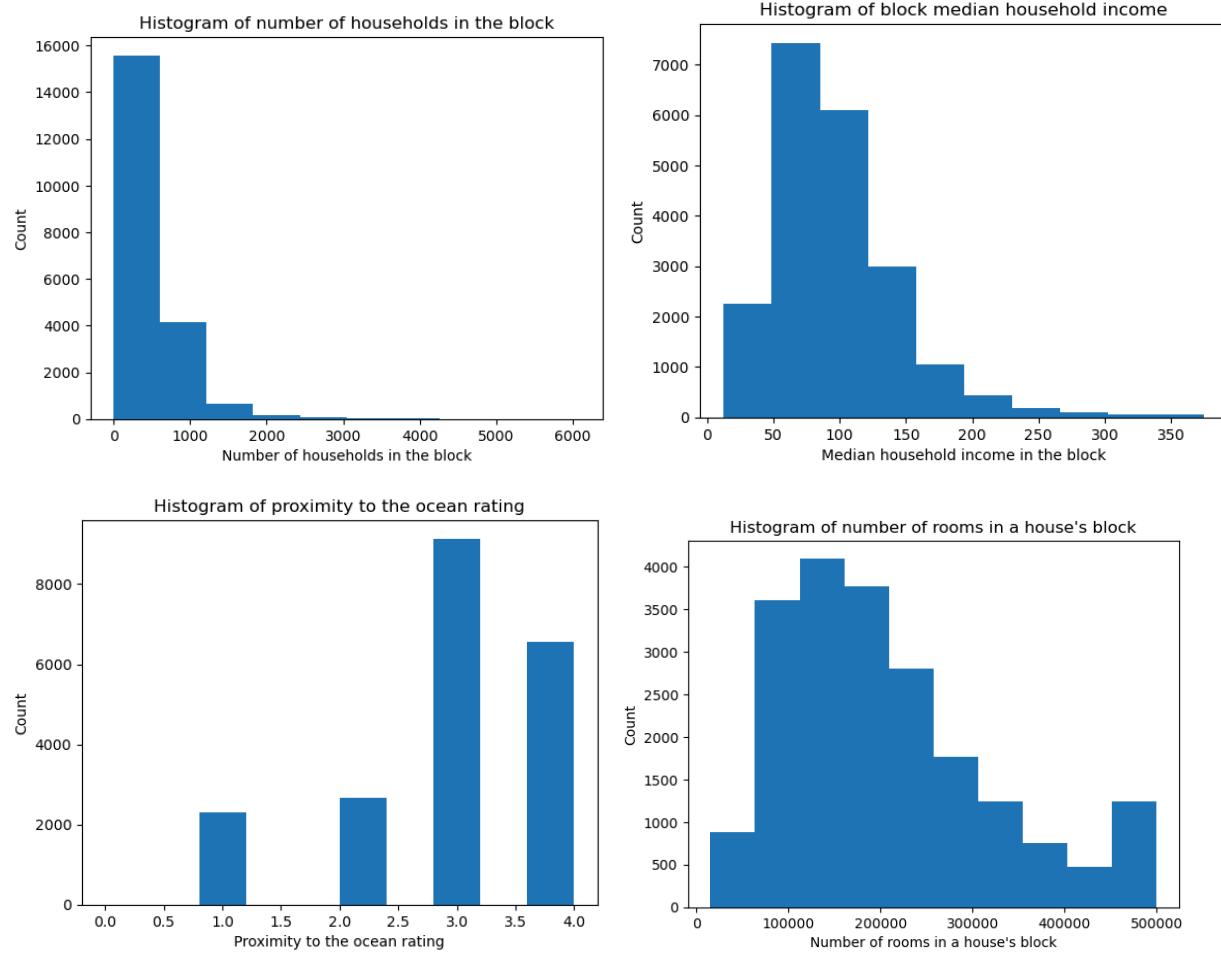
I calculated the correlation for both pairs of variables in order to see if they are strongly correlated and have the potential to cause collinearity issues. Predictor variables 2 and 3 are moderately strongly correlated, with a correlation of 0.641. Predictor variables 4 and 5 are very strongly correlated, with a correlation of 0.907. This means that these variables measure very similar features about the houses, and do not provide very different information from each other, causing the model to appear better at predicting than it actually is.

Extra credit:

Does any of the variables (predictor or outcome) follow a distribution that can reasonably be described as a normal distribution?

I plotted a histogram of each (standardized) variable's distribution. The only one that appears potentially normally distributed is the first predictor variable.





Examine the distribution of the outcome variable. Are there any characteristics of this distribution that might limit the validity of the conclusions when answering the questions above? If so, please comment on this characteristic.

This distribution does not appear to have any characteristics that may limit the validity of the conclusions besides the aforementioned cutoff at \$500,000, which is especially visible in the above histogram.