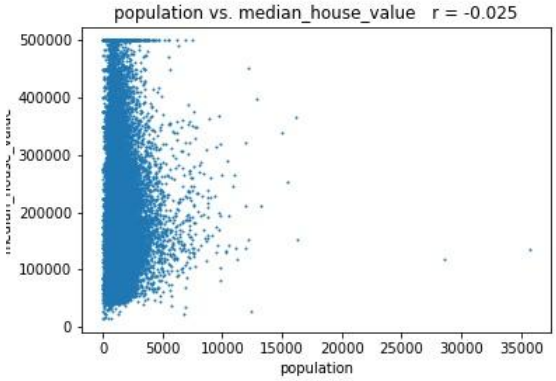
1. Why is it a good idea to standardize/normalize the predictor variables 2 and 3 and why are predictor variables 4 and 5 probably not very useful by themselves to predict median house values in a block?

To answer why we should normalize variables 2 and 3, I simply examined the data. As their names suggest, they present the total number of rooms and bedrooms in all houses in a block, which does not say much about an individual house. I don't think other more advanced methods are necessary.

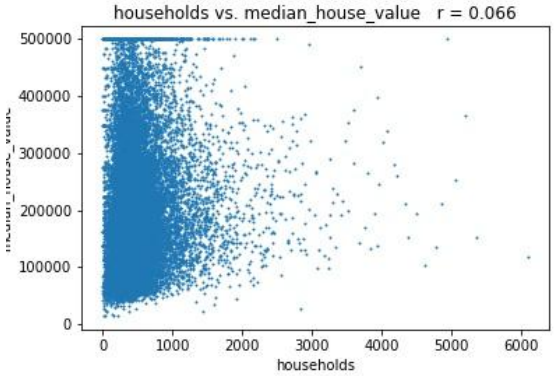
I then computed a correlation matrix and found that variables 4 and 5 had negligible correlation with the median house value, -0.0245 and 0.0658, respectively, which suggests that they are not very useful to predict the target. I chose to examine the correlation matrix because correlation represents the degree of association between two variables. Low correlation means little association.

Also from examining the scatter plots, we can see that neither population nor number of households have any association with median house value.

Var 4 vs. target:



Var5 vs. target:



2. 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)?

From analyzing the correlations, it is better to normalize both variables by population.:
Correlation between roomsPerPerson and median_house_value: **0.209**
Correlation between roomsPerHousehold and median_house_value: 0.152
Correlation between bedroomsPerPerson and median_house_value: **0.113**
Correlation between bedroomsPerHousehold and median_house_value: 0.058

However, this is against my intuition: the value of a house should depend more on the number of rooms in it, instead of how many people live in its rooms.

More importantly, none of these standardized variables nor the original ones has considerable correlation with the target, perhaps because they are collinear with each other.

3. Which of the seven variables is most *and* least predictive of housing value, from a simple linear regression perspective? [Hints: a) Make sure to use the standardized/normalized variables from 2. above; b) 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?]

The most predictive variable is median income, and the least predictive is population. Since this is a linear model, I looked at the (linear) correlation matrix because it tells us about the degree of linear association between variables. The correlation between median income and median house value is 0.688. Population vs. median house value is -0.025.

Perhaps median income can be more predictive if collinear variables— total rooms, total bedrooms— are removed from the data set so that median income can gain more weight in the linear regression model.

This finding makes sense because richer people tend to live in more expensive homes.

4. 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.?

To answer this question, I used the linear regression function from sklearn, and fit all the features and target variable to build the full model. Then do the same thing with variable 3 and target variable. To see which model performs better, I used the model.score() function to compare each model's R^2s. R^2 indicates how much variation in median_house_value is explained by the features. The higher the R^2, the more accurate the models is.

Below are the results.

All predictors R^2: 0.602

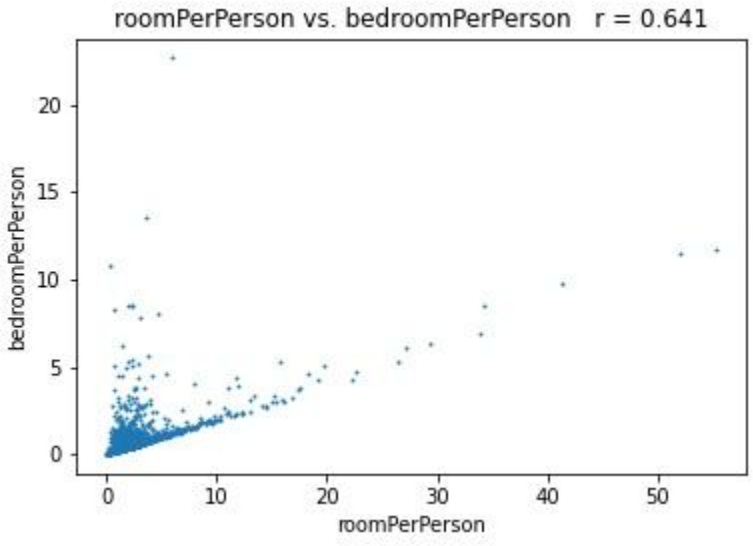
Median_income R^2: 0.473

The full model performs better because it produces higher R^2, which means that it produces lower error.

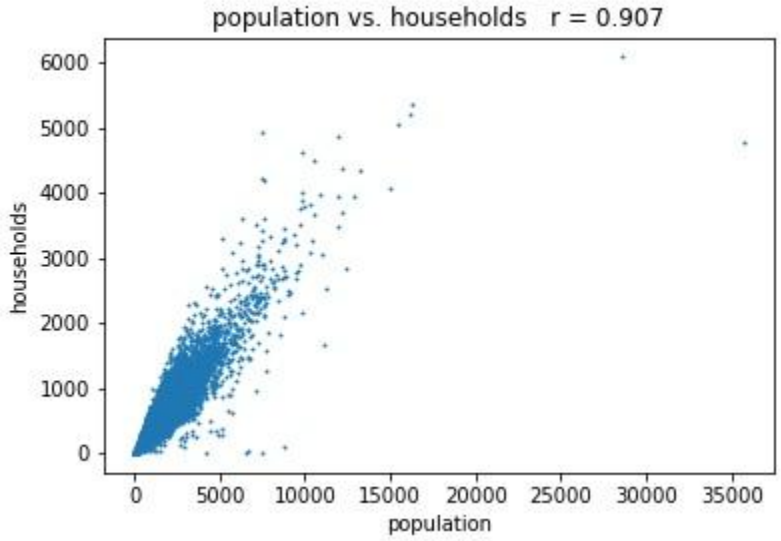
5. Considering the relationship between the (standardized) variables 2 and 3, is there potentially a concern regarding collinearity? Is there a similar concern regarding variables 4 and 5, if you were to include them in the model?

From examining the correlation matrix, I find that variables 2 and 3 have a high correlation of 0.641. Variables of 4 and 5 have a very high correlation of 0.907. Therefore, collinearity is an issue for variables 2 and 3, and variables 4 and 5.

room/person vs. bedroom/person:



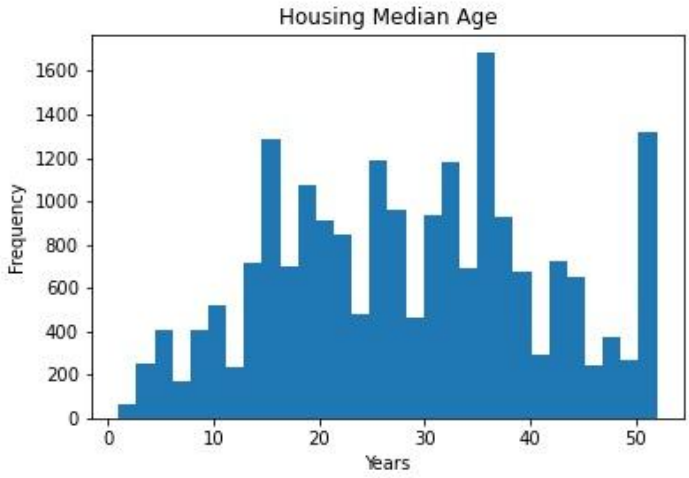
Population vs. number of households:



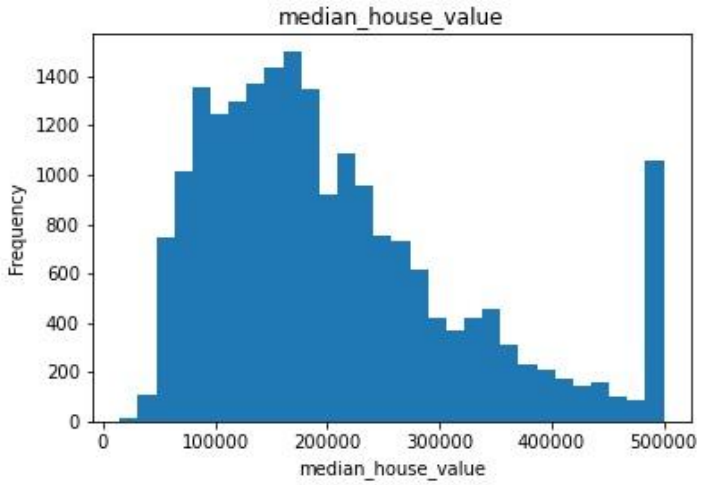
Extra credit:

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

Housing median age slightly resembles a normal distribution. Below is a histogram of the distribution:



b) 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.



Median house value does not seem to follow the normal distribution. The median house value distribution does not peak at the center. It is right skewed instead. Furthermore, at 500000 there is another spike, indicating that it is a potential outlier.