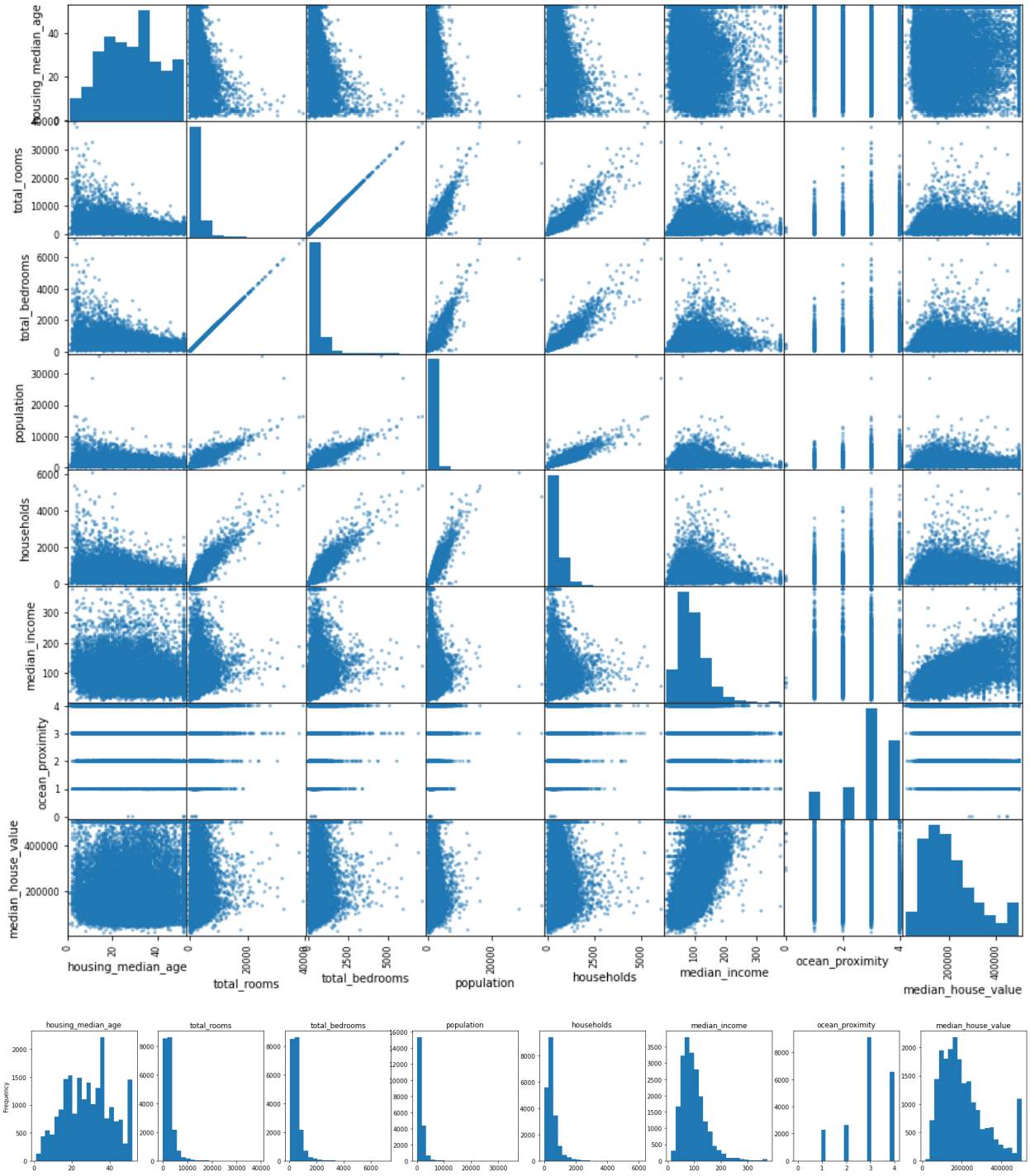
Simone Rittenhouse
Prof. Pascal Wallisch
Into to Machine Learning
16 February 2022

Intro to Machine Learning: Homework 1

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?**
    a. For this question, I performed exploratory analysis on the data set to see each variable's distribution and the relationships between them. Specifically, I used the Pandas' plotting.scatter_matrix() method to generate scatter plots between all the variables as well as histograms describing their distributions. I then found the Pearson correlation coefficient between all of the variables using Pandas' df.corr() method. Lastly, to more closely inspect the shape of each variable's distribution, I plotted individual histograms for the eight variables.
    b. I found the correlation coefficients between the predictor variables and housing value because I wanted to assess how strong the linear relationship was between predictors and the outcome. For linear regression models, it's assumed that the predictors have a linear relationship with their outcome. I was interested in assessing whether or not variables 4 and 5 had a strong correlation with housing value. Additionally, I wanted to see how variables 2 and 3 correlate with the other variables to determine why we need to standardize them. I also made scatter plots and histograms to better visualize the relationships between variables (through scatter plots) and their individual distributions (through histograms).
    c. I found that variables 2 and 3 were both highly correlated with variables 4 and 5. Both variables had r = 0.857 with population and r = 0.918 with households. This linear relationship could also be seen in the scatter plots between them, and the histograms of variables 2 and 3 also resemble variables 4 and 5 in distribution. Out of the seven predictor variables, 4 and 5 had the lowest correlations with median housing value (variable 4 had r = -0.025 and variable 5 had r = 0.066).

Index	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
housing_median_age	1	-0.361262	-0.361259	-0.296244	-0.302916	-0.119034	-0.295012	0.105623
total_rooms	-0.361262	1	1	0.857126	0.918484	0.19805	0.0315864	0.134153
total_bedrooms	-0.361259	1	1	0.857118	0.918481	0.198046	0.0315846	0.134154
population	-0.296244	0.857126	0.857118	1	0.907222	0.00483435	0.0394148	-0.0246497
households	-0.302916	0.918484	0.918481	0.907222	1	0.0130331	-0.0128726	0.0658427
median_income	-0.119034	0.19805	0.198046	0.00483435	0.0130331	1	-0.163755	0.688075
ocean_proximity	-0.295012	0.0315864	0.0315846	0.0394148	-0.0128726	-0.163755	1	-0.397251
median_house_value	0.105623	0.134153	0.134154	-0.0246497	0.0658427	0.688075	-0.397251	1

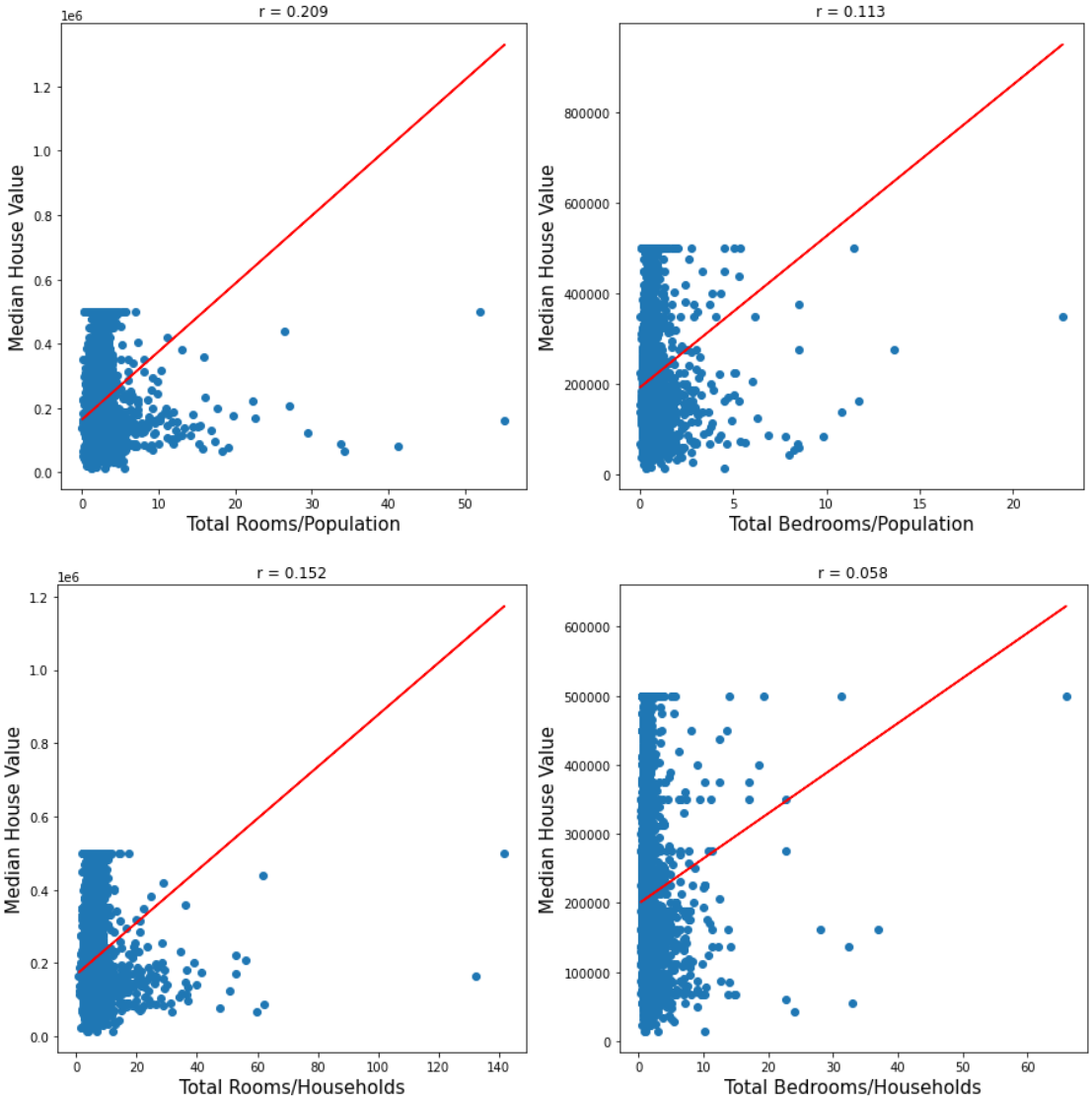


d. The low correlation between variables 4 and 5 and the outcome variable suggests that there is not a strong linear relationship between them, which means that they will be poor predictors of median housing value in a linear regression model. The lack of their linear relationship is further evidenced in the scatter plots between them. Additionally, it's a good idea to standardize variables 2 and 3 because they

are highly correlated with the population and number of households in a particular block. This suggests that the total number of rooms and bedrooms in a block depends largely on the size of the block itself (i.e. how many people, and houses, are represented in each block). This means that we will not get a useful prediction of housing value from the aggregated number of rooms or bedrooms since we aren't capturing the number of rooms per house, but rather an amount of rooms that varies based on the size of the block itself. It's therefore necessary to standardize by either population or by the number of households in a block to get the number of rooms/bedrooms per person or per household. This ensures that variables 2 and 3 vary block to block based on the houses in the block rather than based on the size of the block itself. Standardization will also reduce the correlation between variables 2 and 3 with variables 4 and 5, which helps prevent collinearity (since they are currently not independent of one another).

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)?**
    a. In order to decide whether or not to standardize by population or number of households, I created two copies of my original data set and divided the number of rooms and the number of bedrooms by population and number of households respectively. I then found the correlation between both versions of the standardized variables 2 and 3 and the outcome variable and created scatter plots to show the relationship between the two predictor variables and the outcome. Lastly, I ran two multiple regression models using all predictors and both forms of the standardized data and found the R-squared for both of them.
    b. I tried both methods of standardization because I wanted to see which method would produce a more informative regression model. I checked the correlation coefficients and R-squared for the data set after both to see how dividing by (i.e. standardizing by) the variable affected the strength of a linear regression model. This would help me choose the best method of standardization, because the method that produced the highest R-squared would be a better overall predictor of the median house value, as it would explain more of the variance in the outcome.
    c. I found that standardizing by population produced a higher correlation between both variables 2 and 3 and the outcome variable ($r = 0.209$ for total_rooms and $r = 0.113$ for total_bedrooms standardized by population vs. $r = 0.152$ for total_rooms and $r = 0.058$ for total_bedrooms standardized by household). Because the correlation coefficients were higher standardized by population, I also found that the R-squared of a multiple regression model using all predictors with variables 2 and 3 standardized by population was higher than when I ran a multiple regression model using standardization by household (R-squared = 0.599

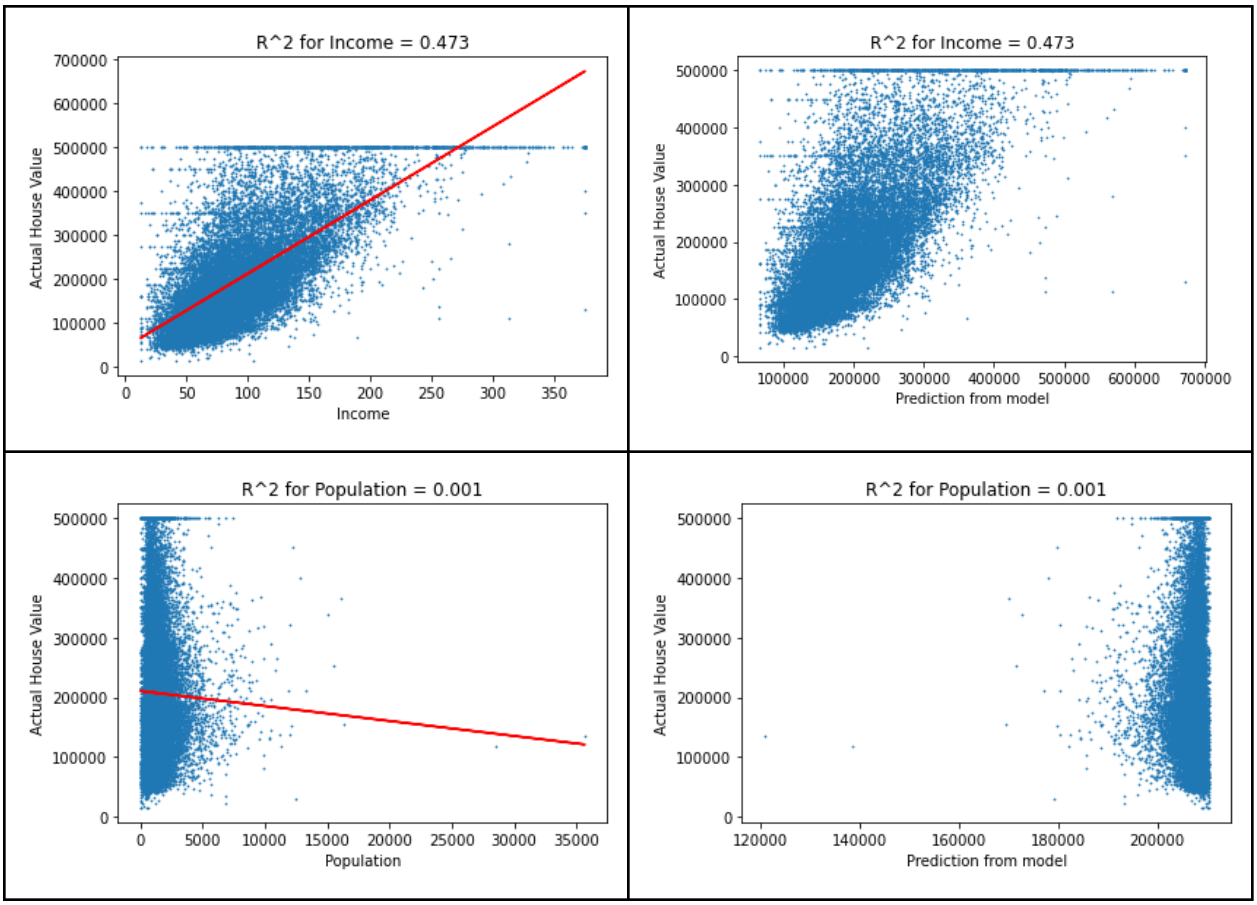
standardized by household vs. R-squared = 0.601 standardized by population). The scatter plots below demonstrate the increased correlation between both predictors and outcome after standardizing by population (as compared with household).



d. A higher R-squared in the multiple regression model and higher correlation coefficients between the standardized predictors and the outcome variable after standardizing by population (as compared with standardizing by the number of households) suggests that it is better to standardize by population. This will lead to a regression model that explains more of the variance in the outcome, leading to better predictions. This also makes sense intuitively as household size is variable. For example, 3 bedrooms per household could mean the house is large (and likely more valuable) if the average household in a block is made up of 2

people, but could mean the opposite if it's 3 bedrooms per household for an average household of 7 people. Rooms per person gives a much more consistent scale of the size of a typical house in each block. Therefore, in the rest of my analysis, total_rooms and total_bedrooms have been standardized by population.

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?]**
    a. After standardizing variables 2 and 3, I iterated through the seven predictor variables and ran seven simple linear regression models. After fitting a model to each feature, I calculated the R-squared and stored this value so I could compare it to that of the other six models. I also created seven scatter plots of each model's predictions against the actual values of the median house value. To further compare predictors, I created scatter plots between the predictors themselves and median housing value along with the line of best fit.
    b. I regressed median house value on each of the seven predictors individually because I could then compare each of the model's R-squared to see which variable is least and most predictive of housing value. The model with the highest R-squared would be most predictive because it would explain the most variance in housing value. Likewise, the model with the lowest R-squared would be least predictive. Additionally, plotting the predictions against the actual outcome values allowed me to examine the performance of each model and see if the model was consistently over- or under-estimating housing value. The scatter plots between the predictors themselves and housing value allowed me to see if the underlying relationship between each feature and the outcome was linear to begin with, as linear regression assumes this and it will make for a more accurate predictive model.
    c. Regressing housing value on income had the highest R-squared of 0.473. Regressing housing value on population had the lowest R-squared of 0.001. The scatter plot for income shows a positive relationship between the feature and housing value; however, there is a lot of variation in income across values of the outcome - particularly at higher housing values. The scatter plot for population does not suggest a strong linear relationship. Population remains relatively low (between 0 and 10,000) across all values of the outcome with a few extreme outliers (> 25,000).

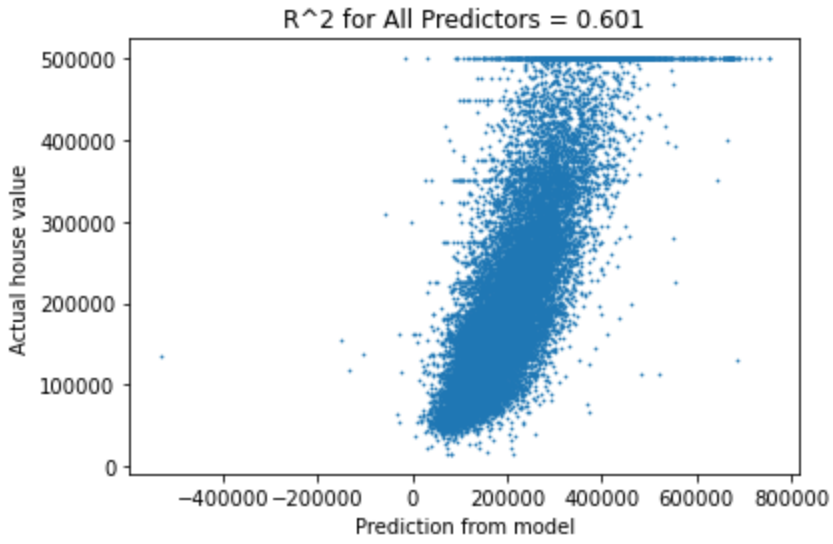


	R-Squared
housing_median_age	0.011156
total_rooms	0.043883
total_bedrooms	0.012791
population	0.000608
households	0.004335
median_income	0.473447
ocean_proximity	0.157808

d. Because the simple regression model with median income had the highest R-squared, income is the most predictive of housing value. This is because it explains the most variance in the outcome variable (47.3%). Population is therefore the least predictive of housing value, because it has the lowest R-squared and only explains 0.1% of variance in the outcome variable. The low R-squared for population could be explained by the lack of a strong linear relationship between population and housing value (as shown in the scatter plot). Additionally, outliers strongly impacted the model because the errors are squared. The strongest predictor - income - also had a few outliers which affected the model. Another concern for this model is that housing value is only measured to $500,000. Linear regression produces continuous, unbounded predictions, but

because housing value is capped, some predictions were much higher than the actual value. Income would likely be a stronger predictor if housing value was measured beyond $500,000.

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.?**
    a. For this question, I included all seven predictors (with the standardized versions of 2 and 3) in a multiple regression model to predict housing value. I found the R-squared, intercept, and beta coefficients and output them for analysis. Lastly, I created a scatter plot between the model's predictions and the actual values of the outcome variable.
    b. I included all predictors in my model to see how well all of them together could predict housing value. Additionally, I computed the R-squared so that I could compare this model to the strongest simple linear regression model (that regressed median housing value on median income) in terms of how well both explain the variance in the outcome variable. I also outputted the beta coefficients so that I could examine how each predictor performs within the model. When the units of all predictors are standardized, a larger absolute value of beta coefficient means that the associated feature has a stronger effect in predicting the outcome variable - although importantly in this case the units between predictors differ.
    c. I found that the multiple regression model explains 60.1% of the variance in median housing value (R-squared = 0.601). This is compared with an R-squared of 0.473 in my strongest single linear regression model from question 3. Additionally, the scatter plot shows some of the same issues as found in the previous question. Namely, there is still an effect of outliers and the effect of a bounded outcome variable.



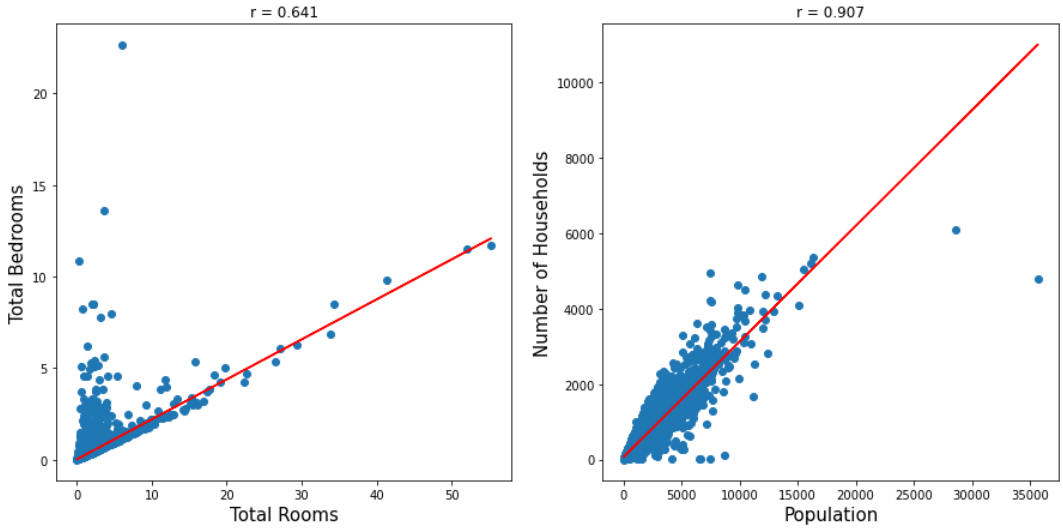

```
All Predictors in Multiple Regression Model
R-squared: 0.6006645246293567
Intercept: 78483.12861017353
Betas: [  1308.46936341   2404.25830438   7124.58804627    -34.80646359
    124.34961588   1591.70114514 -28170.29653102]
```

d. Because the R-squared in the multiple regression model is higher than that computed from using the most predictive feature in a single regression model, this model is a better predictor of housing value than the model using income from question 3. This is to be expected, because adding additional features to a multiple linear regression model will always increase the R-squared of that model. However, it's also worth noting that this model is also affected by extreme outliers and a bounded outcome variable. This led to some predictions that were much higher than the upper limit of the median housing value (predicted value $800,000 vs. actual value $500,000) and some impossibly low (a predicted median housing value < -$400,000).

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?**
    a. For this question, I found the correlation matrix for all variables using my population standardized variables 2 and 3 and the Pandas' .corr() method. I then created two scatter plots. One to show the relationship between the standardized variables 2 and 3, and the other to show the relationship between variables 4 and 5. I added best fit lines to each plot and found the Pearson's correlation coefficient for these two relationships (which can be seen in the title of my scatter plots).
    b. I first plotted scatter plots between variables 2 and 3 as well as between variables 4 and 5 because it would allow me to assess the strength of their relationship. If

the variables exhibited collinearity, I would expect to see highly correlated data - which is observed if the scatter plot shows a strong linear relationship. I included best fit lines to assess how well the variables' relationship can be fit linearly. Additionally, I looked at the correlation matrix of all variables to see if variables 2 and 3 as well as variables 4 and 5 were significantly more highly correlated than the other variables. If there is collinearity between these two pairs of variables, I would again expect to see a higher correlation between them than with other predictors.

c. Total rooms and total bedrooms, after being standardized by population, had a correlation coefficient of r = 0.641. Population and the number of households in a block had an extremely high correlation coefficient of r = 0.907. Both scatter plots show fairly linear relationships with some outliers. Additionally, the correlation matrix that included all variables showed that population and the number of households had the highest correlation out of any variable pairing. The total number of rooms and the total number of bedrooms was also the third highest correlation, with the second being between income and the outcome variable.



Index	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
housing_median_age	1	-0.107926	-0.0372396	-0.296244	-0.302916	-0.119034	-0.295012	0.105623
total_rooms	-0.107926	1	0.641464	-0.14052	-0.0284734	0.237828	0.0283277	0.209482
total_bedrooms	-0.0372396	0.641464	1	-0.219651	-0.160157	0.116347	0.0272215	0.113095
population	-0.296244	-0.14052	-0.219651	1	0.907222	0.00483435	0.0394148	-0.0246497
households	-0.302916	-0.0284734	-0.160157	0.907222	1	0.0130331	-0.0128726	0.0658427
median_income	-0.119034	0.237828	0.116347	0.00483435	0.0130331	1	-0.163755	0.688075
ocean_proximity	-0.295012	0.0283277	0.0272215	0.0394148	-0.0128726	-0.163755	1	-0.397251
median_house_value	0.105623	0.209482	0.113095	-0.0246497	0.0658427	0.688075	-0.397251	1

d. Because variables 2 and 3 are highly correlated with r = 0.641 and show a correlation significantly stronger than most of the other predictors do (when

correlated with other predictor variables), there is a concern regarding collinearity. Similarly, variables 4 and 5 are extremely highly correlated with an r close to 1. This also suggests that there is collinearity between them. Overall, there is evidence to suggest that variables 2 and 3 are not independent of each other, and variables 4 and 5 are also not independent of one another. This means that our beta estimates of variables 2, 3, 4, and 5 in a multiple regression model will be unreliable and higher than their true values should we include all four predictors.

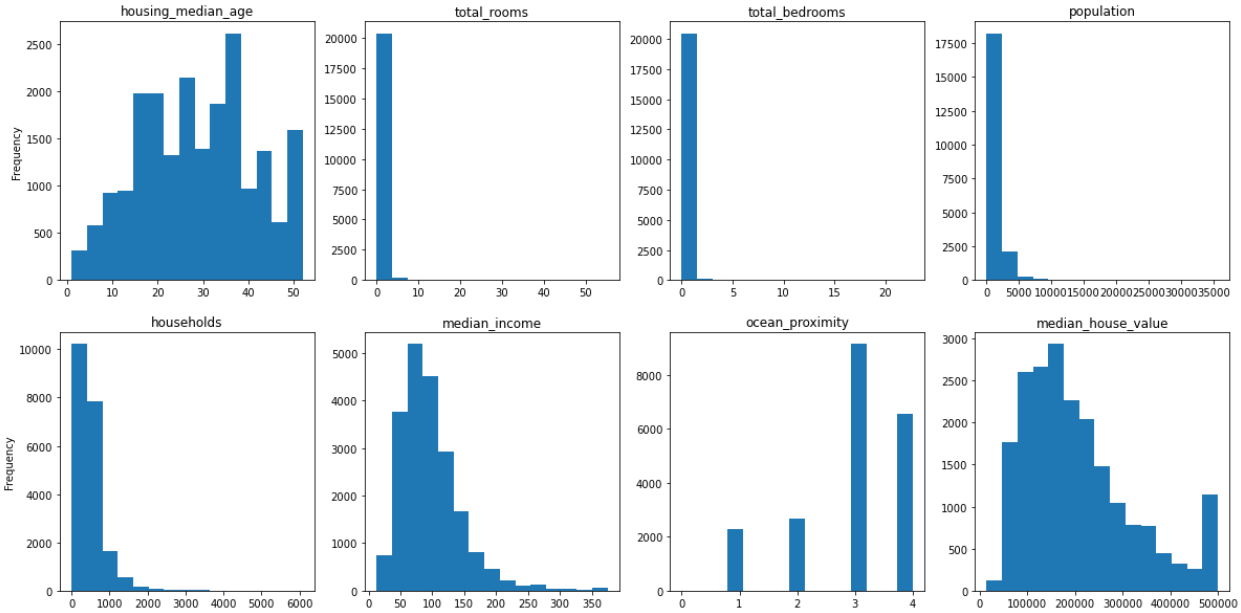
Extra Credit:

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

a. In order to find the distributions of each variable, I plotted histograms of all eight variables - using the standardized versions of variables 2 and 3. I divided each variable into 15 bins because, after testing several options, I found that it gave a good overall impression of the shape of each variable's distribution.

b. I decided to use histograms to examine the shape of each variable's distribution because histograms plot the frequency of each value of a particular variable. This means that we can use them to determine the underlying distribution of a variable since the variable's distribution simply shows how often a particular value of the variable is expected to occur. To see if a variable is normally distributed, we would expect a bell shaped curve with symmetrical tails on either side of the distribution's peak.

c. Several of the variables had peaks as expected in a normal distribution, but did not have symmetrical tails. In order words, there was skew in several distributions that's more characteristic of a gamma or log-normal distribution (rather than a normal distribution). Both median income and median housing value appear to be one of these distributions. Total rooms, total bedrooms, population, and households appear to follow power-law distributions, as they peak at lower values before quickly tapering off. Ocean proximity appears to have left-skew. Lastly, median house age has significant variability - which may be due to noise. However, it also seems to follow a bell curve with a peak around 25 years.



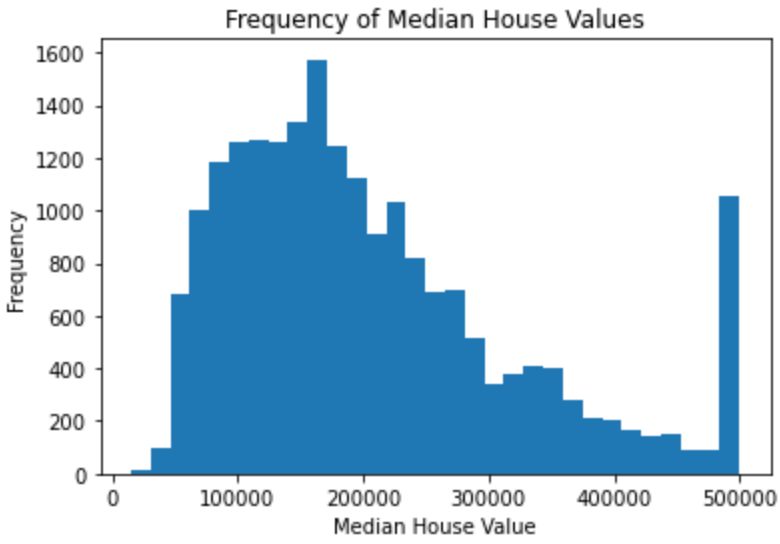
d. Despite the variability around the frequency of values of median house age, it is the only variable that doesn't show significant skew. It seems reasonable to think that the underlying distribution may be a normal distribution, as there seems to be a peak between 20 and 30 years with symmetric tails on either side. In contrast, the other variables seem to follow different distributions.

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.

a. In order to determine if the distribution of median housing value had any limiting characteristics, I plotted the histogram of the outcome variable on its own. I divided the outcome variable into 31 bins in this plot to more closely assess the shape of its distribution.

b. Similarly to the previous question, plotting the outcome variable's histogram allows me to examine its distribution. Additionally, I used more bins than in the previous question because it allowed me to examine the distribution of median housing value on a more granular level.

c. I found that median housing value appears to follow a right-skewed distribution (possibly a gamma or log-normal distribution). Additionally, there is a spike in frequency at the upper limit of median housing value ($500,000). This is unexpected because the tail of a right-skewed distribution should continue to decrease in frequency rather than showing a second peak.



d. One characteristic of the distribution that might limit the validity of my previous conclusions is that the median housing value is not measured beyond $500,000. It's likely that in certain blocks the median housing value is greater than $500,000, but was marked as $500,000 since this appears to be the greatest accepted value in the data set. This would explain the spike in frequency, as that frequency actually captured the total number of blocks that had a median housing value greater than or equal to the cutoff. If this were true, the distribution of median housing values would continue beyond $500,000 and the frequency of $500,000 would be significantly lower. It would also mean that certain models that have overestimated housing value based on this data set may be closer to the true value than they appear - as the true value of our outcome variable could be greater than $500,000.