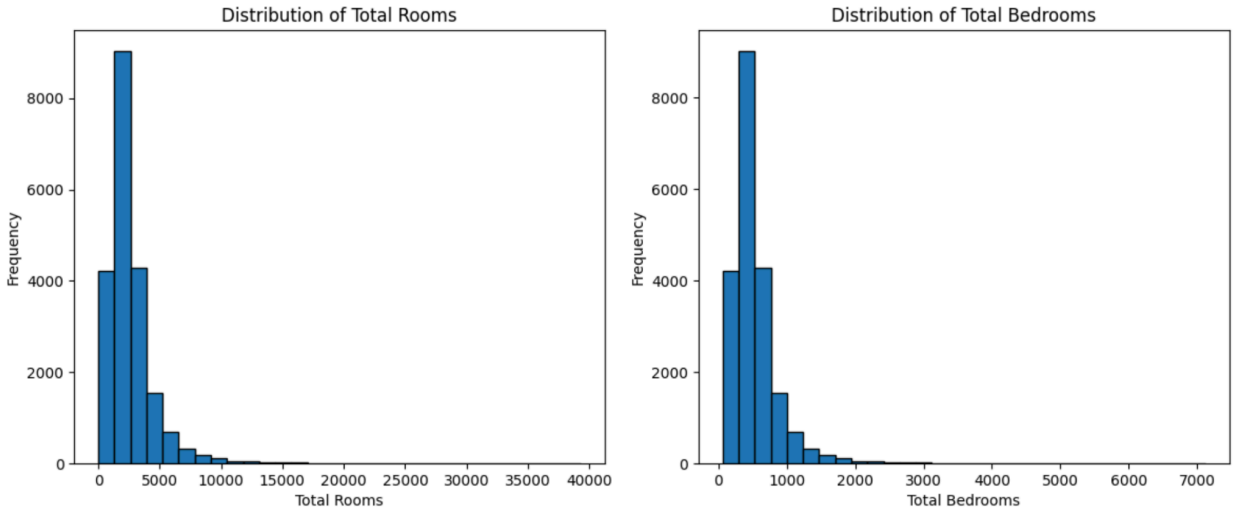
Question 1: Why is it a good idea to standardize/normalize the predictor variables 2 and 3 and why are predictor variables 4 and 5 probably not very useful by themselves to predict median house values in a block?

To determine why Total Rooms (predictor 2) and Total Bedrooms (predictor 3) should be normalized, I made a histogram of these variables as well as of Population (predictor 4) and Households (predictor 5) to check their distribution. I found all these distributions were skewed right, indicating that most housing blocks have a few rooms and bedrooms. Because both Total Rooms and Total Bedrooms are skewed right, it is difficult to compare their values by different sized housing blocks. Normalization is needed to account for differing block sizes and make these variables meaningful predictors.

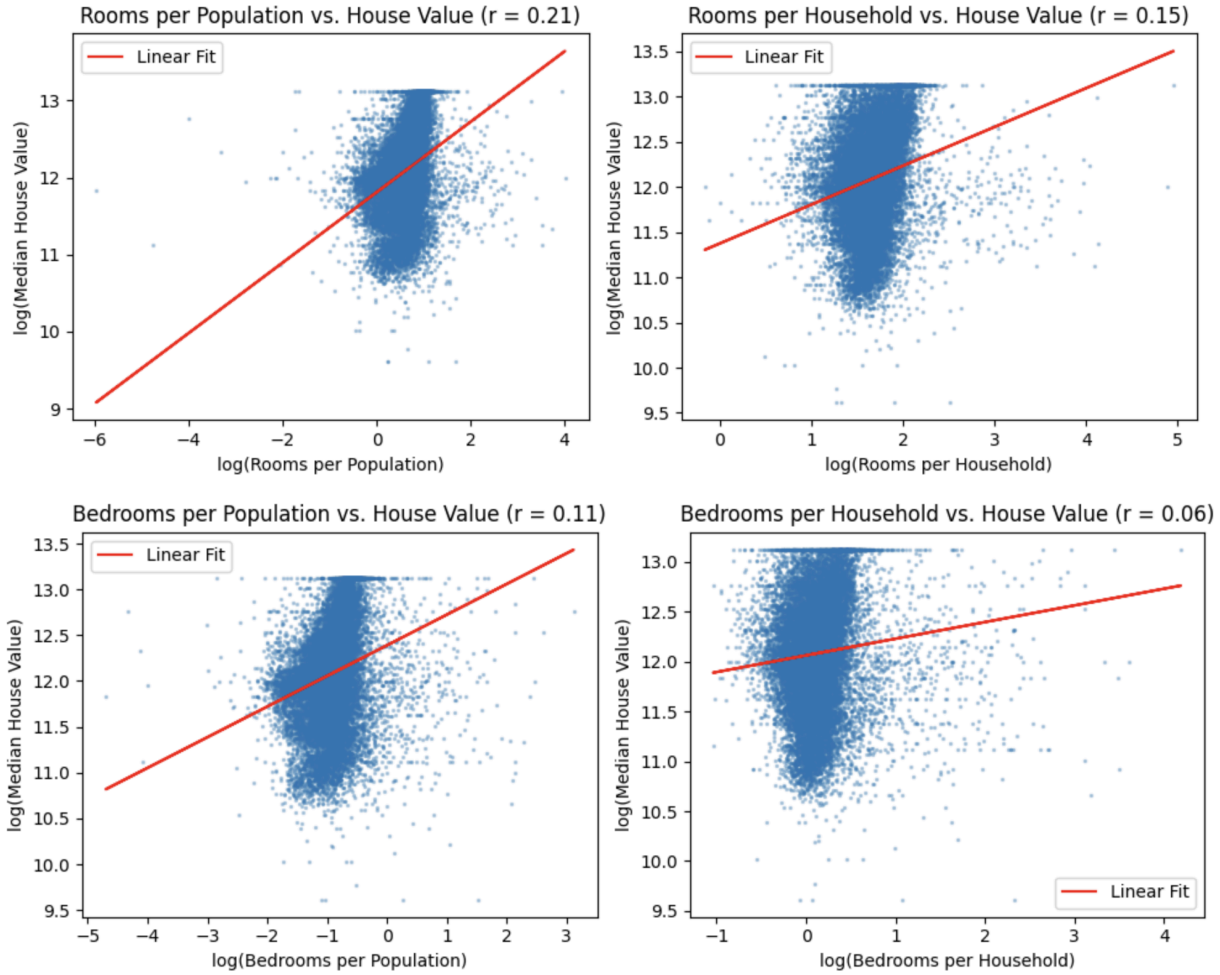
I also checked the correlation between Population and Median Housing Value as well as Households and Median Housing Value. This was done to check if there was an underlying linear relationship between these predictor variables and the outcome variable before running any linear regression. I found that the correlation between Population and Median House Value is -0.025 and that between Households and Median House Value is 0.066. This ultimately means that these predictors are not good indicators of Median House Value, as a larger population and/or household doesn't necessarily imply a larger Median House Value.



Question 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)?

I normalized both Total Rooms & Total Bedrooms by Population & Household and correlated each value with Median House Value. This was done in order to see which normalization method (Population or Household) yielded the higher correlation. Holistically, this was done to help make fair comparisons across blocks of different sizes. I found that normalizing both Total Rooms as well as Total Bedrooms by Population led to higher correlations with House Value

than did normalizing by Household. The correlation for Rooms per Population vs. House Value is 0.209 while that for Rooms per Household vs. House Value is 0.152. The correlation for Bedrooms per Population vs. House Value is 0.113 while that for Bedrooms per Household vs. House Value is 0.058. This means that the size of the population in a block is a stronger factor in determining house value than the number of households.

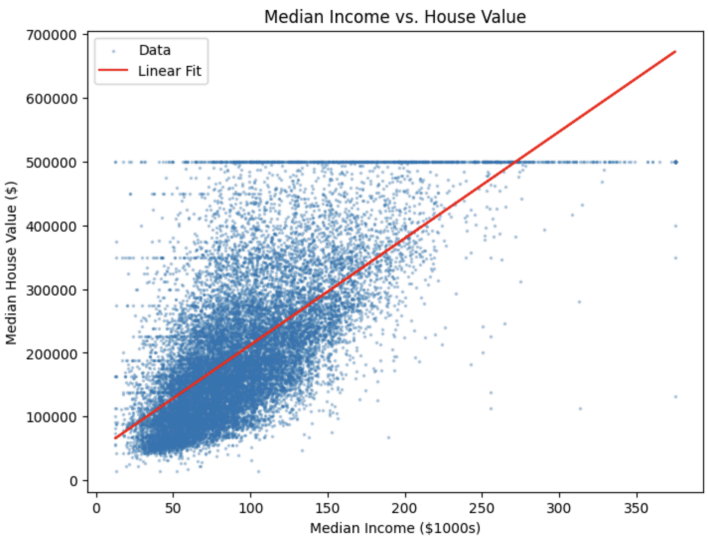


*Question 3: Which of the seven variables is most *and* least predictive of housing value, from a simple linear regression perspective?*

I ran a simple linear regression model using each of the given predictors (Predictors 1-7) with Median House Value as the outcome variable of interest. This was done to see which of the 7 predictor variables has the highest R^2 value (and is most predictive of Median House Value) and which has the lowest R^2 value (and is least predictive of Median House Value). I chose a simple linear regression model and not a multiple linear regression model because we are trying to investigate the effect of singular variables in predicting Median House Value alone. I found that Median Income has the highest R^2 value of 0.473, while Population has the lowest R^2 value of 0.001. This implies that of the 7 predictor variables, Median Income explains the

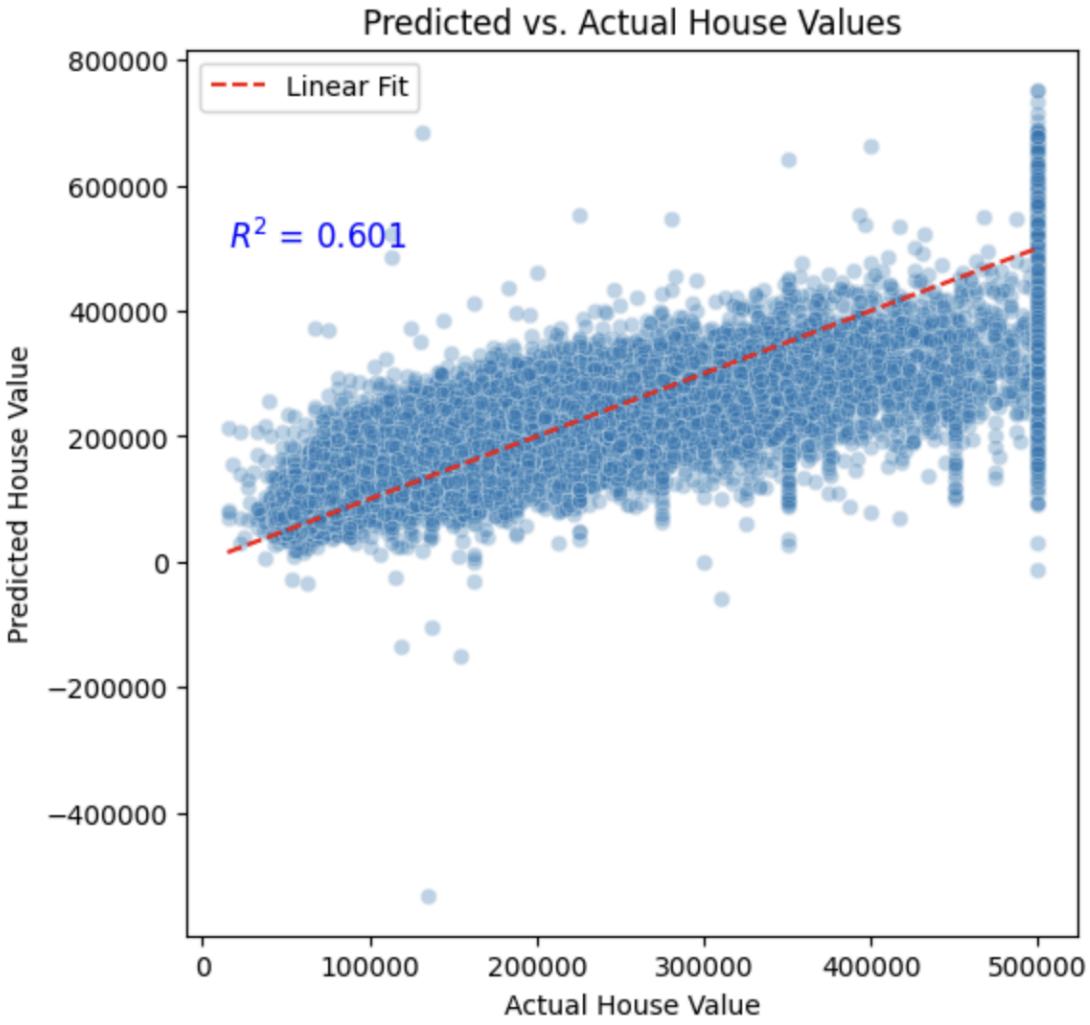
variation in Median House Value the best, while Population explains the variation in Median House Value the worst.

A potential issue with the scatterplot is that it has an artificial Median House Value cut-off at $500,000. This means that higher-income areas might actually have much higher house values, but they are artificially cut off at $500,000. This limits the predictive power of median income because we can't see the true relationship for high-income neighborhoods. If the dataset didn't have this cap, we would likely see an R^2 value greater than 0.473 and a steeper regression line, indicating that the best predictor, Median Income, would be even more predictive without this artificial cut-off.



Question 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?

I ran a multiple regression model using all of the predictor variables with Median House Value as the outcome variable of interest. This was done to see the overall effect of all the independent variables together in predicting Median House Value, and ultimately, seeing if this drastically alters R^2. I chose a multiple regression model because we are trying to investigate the effect of multiple variables in predicting Median House Value simultaneously. The R^2 for this model improved to 0.601, meaning more variance in Median House Value is explained by all the predictors than the best predictor variable alone. This model also improved R^2 by 0.127 from the R^2 of the previous model utilizing the best individual predictor of Median Income.

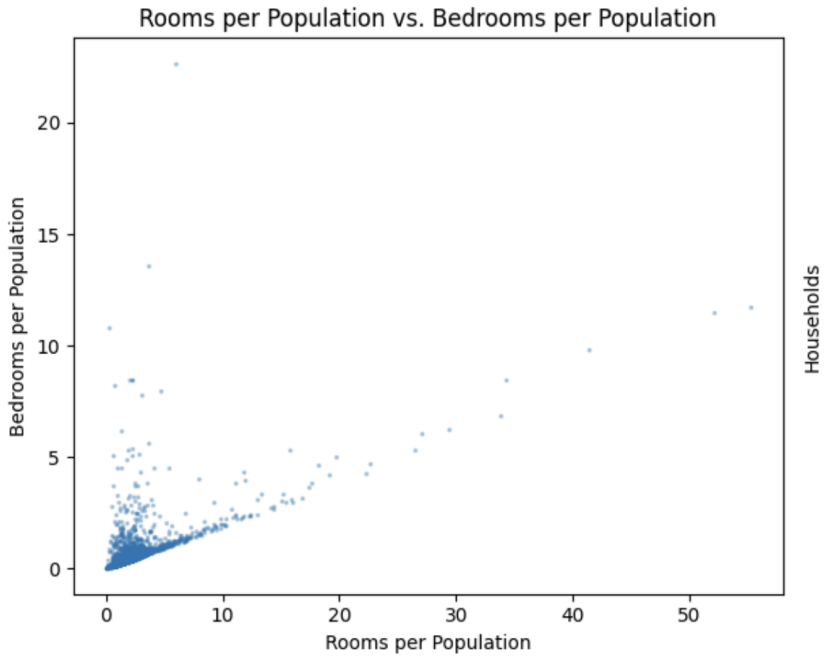


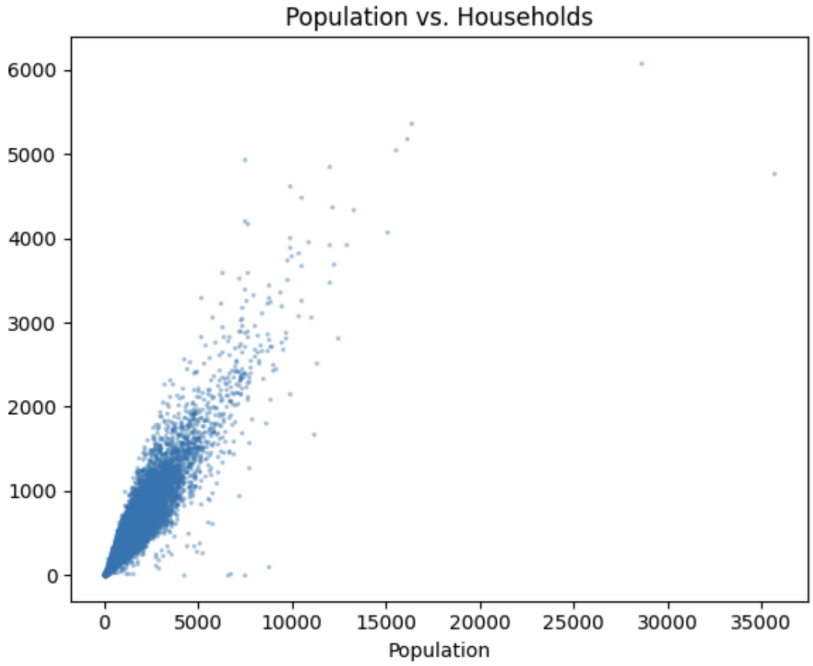

```
Feature Importance (from strongest to weakest):
Proximity to Ocean: -28170.297
Bedrooms per Population: 7124.588
Rooms per Population: 2404.258
Median Income: 1591.701
Median Age: 1308.469
Households: 124.350
Population: -34.806
```

Question 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?

I found the correlation between Rooms per Population and Bedrooms per Population is 0.641. The correlation between Population and Households is 0.907. I did this because having two collinear predictors in a multiple regression model doesn't improve R^2, makes beta estimates for each predictor unstable, and adds unnecessary variance to the model. Rooms per Population and Bedrooms per Population are moderately correlated at 0.641. Given that this

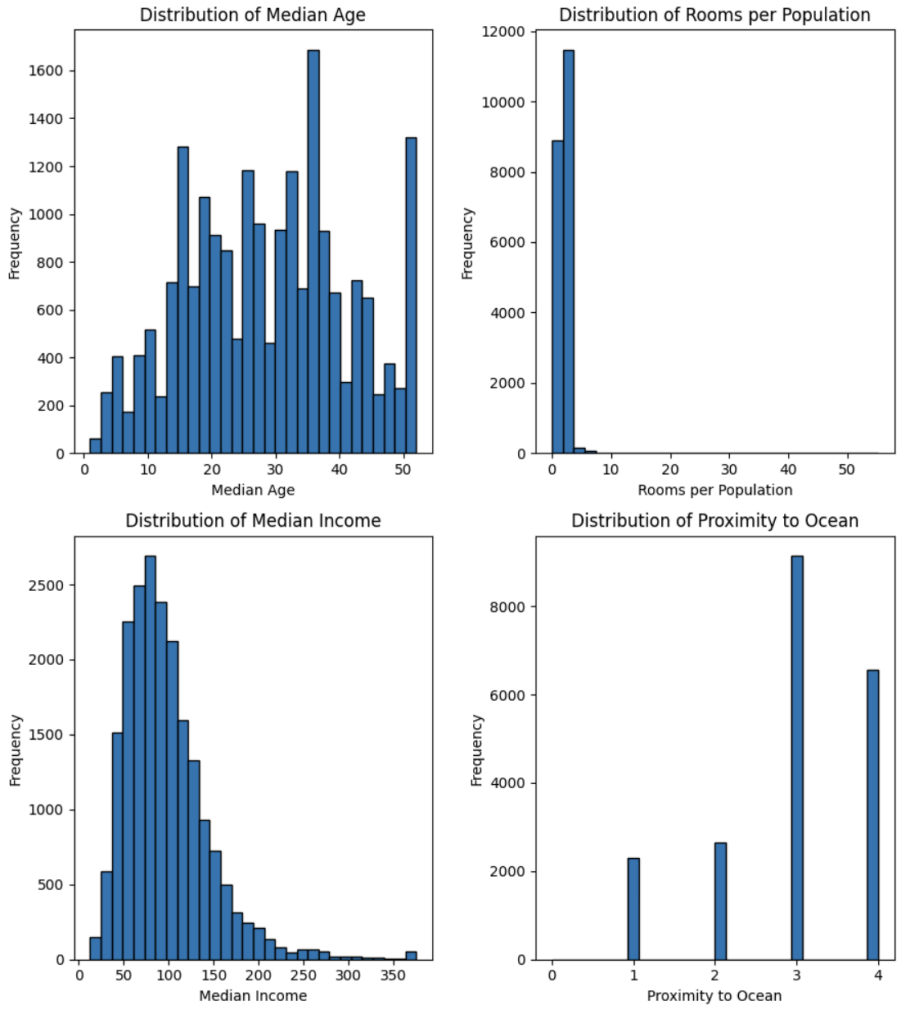
R^2 is below the threshold of 0.8 or 0.9 for highly collinear variables, Rooms per Population and Bedrooms per Population can still be used in the model. The correlation of 0.907 between Population & Households, however, is a cause for concern as these variables are extremely collinear. Since they have similar predicting power to each other, it would be a good idea to exclude Population as this predictor has the weakest feature importance based on the figure in Question 4.

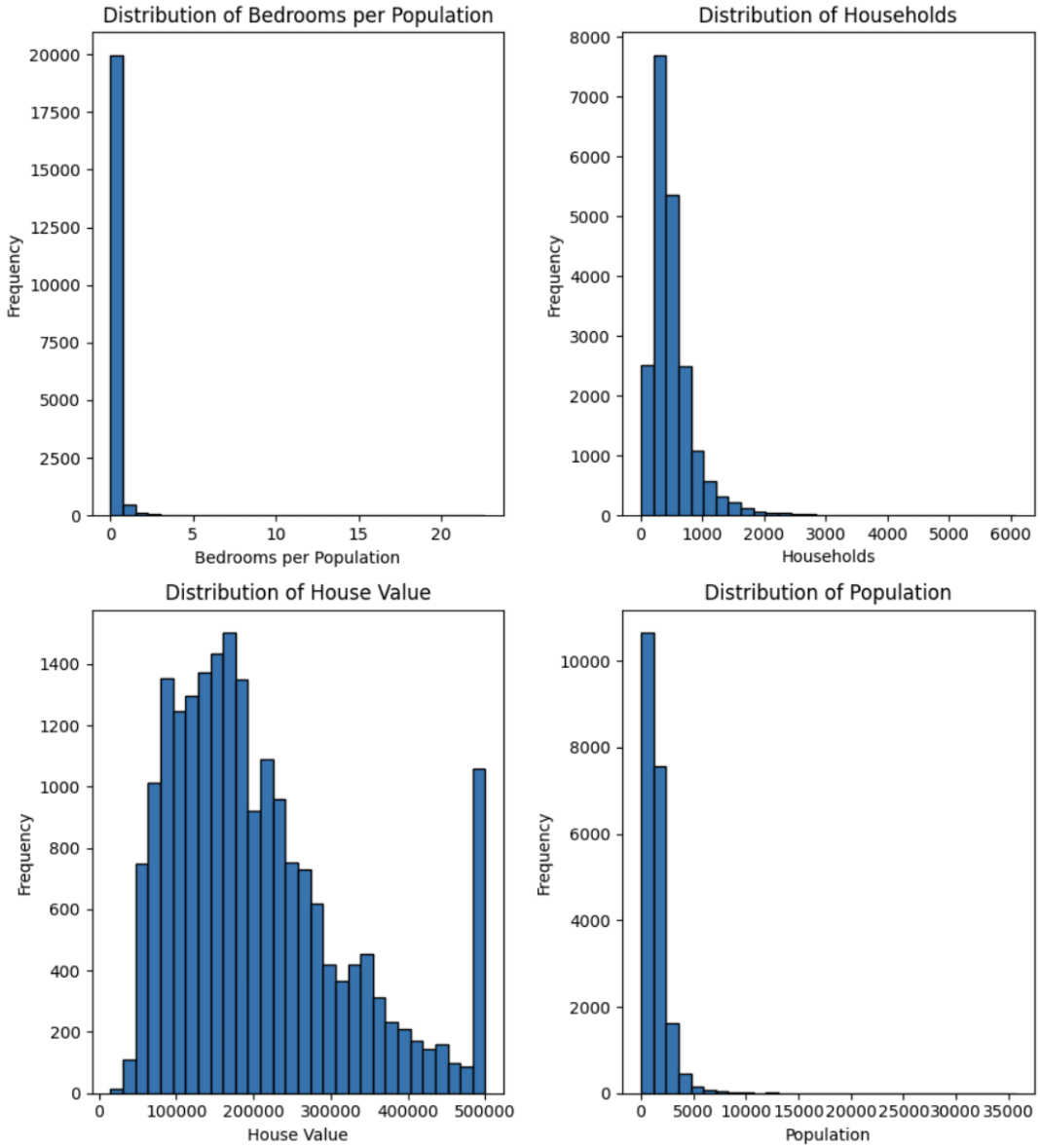




Extra Credit A: Does any of the variables (predictor or outcome) follow a distribution that can reasonably be described as a normal distribution?

I created a histogram for each predictor as well as the outcome variable. I also computed skewness and kurtosis scores for each distribution. I did this to have a visual and numerical representation of each distribution in order to determine which variables are normally distributed. While skewness detects the symmetry of the distribution, kurtosis detects the extremity of the tails. I found that none of the variables fit the criteria for distributions that are approximately normal: skewness approximately equal to 0 and kurtosis approximately equal to 3. Consequently, none of the predictor variables are distributed normally.





Median Age - Skewness: 0.060, Kurtosis: -0.801
Rooms per Population - Skewness: 17.774, Kurtosis: 600.523
Bedrooms per Population - Skewness: 20.697, Kurtosis: 763.614
Households - Skewness: 3.410, Kurtosis: 22.052
Median Income - Skewness: 1.647, Kurtosis: 4.951
Proximity to Ocean - Skewness: -0.731, Kurtosis: -0.290
House Value - Skewness: 0.978, Kurtosis: 0.328
Population - Skewness: 4.935, Kurtosis: 73.535

Extra Credit 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.

I plotted a histogram of the distribution for House Value in order to examine any unusual behavior or patterns in the data. I found an outlier at $500,000 due to an artificial cutoff. This limits the model's ability to make accurate housing valuations in high income areas because there are likely some Houses that are valued higher than $500,000. The model may predict a House Value higher than this amount, which, in reality may be accurate, but is deemed inaccurate in the model because no House Values above $500,000 exist in our data. Consequently, this limits the predicting power of any of the above models because we are unable to establish the true relationship between our predictors and House Value.