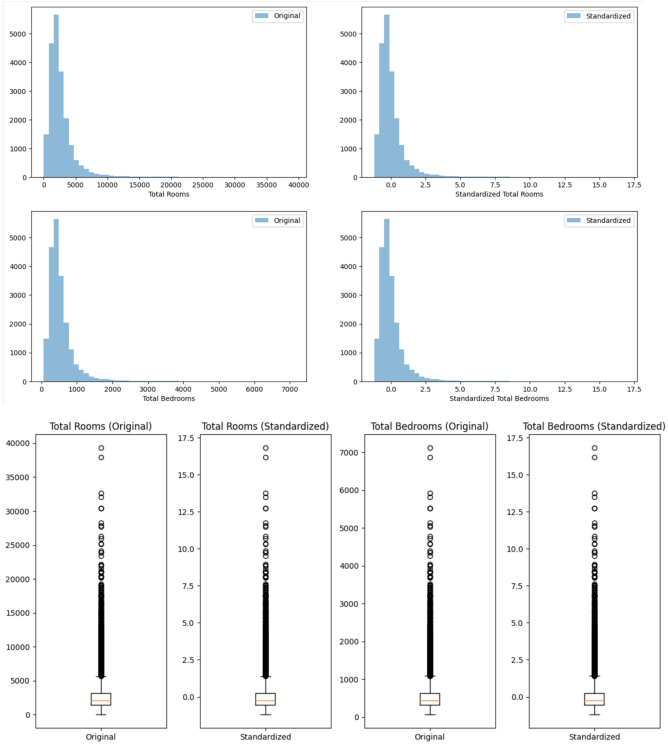
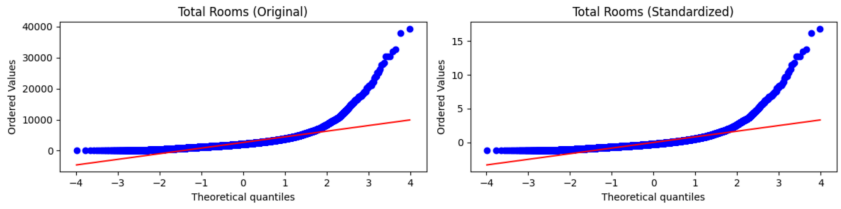
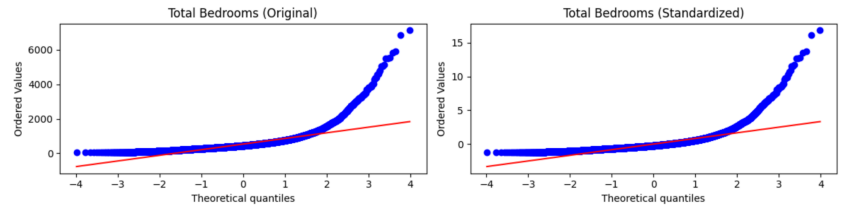
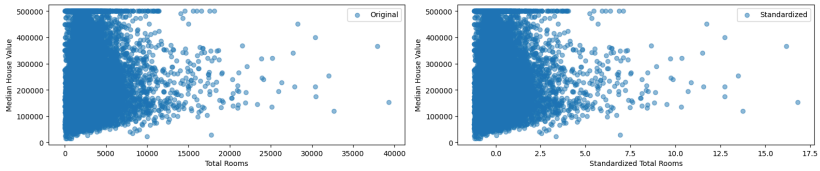
Question 1:

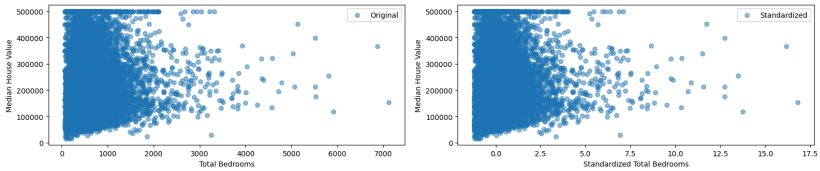
In order to understand why it is a good idea to standardize the predictor variables 2 and 3, I did a lot of plotting for comparison to show the differences before and after standardizing, both by themselves and in relation to the target variable (median house values). The logic behind why predictor variables 4 and 5 are not very useful by themselves to predict median house values in a block is essentially the same as why it is a good idea to standardize predictor variables 2 and 3. This is also verified later in Question 3 when we do simple linear regressions. Whichever predictor variable, standardized is always better than raw. I plotted histograms, box plots, and QQ plots in order to visualize the differences between the distributions by themselves before and after standardizing. The histograms show the difference in the general scale and distribution; the box plots show the difference in central tendency, spread, and skewness; the QQ plot show the difference in the variable's distribution to a standard normal distribution. I also plotted scatter plots and density plots in order to visualize the differences of the relationship between the predictor variables and the target variable before and after standardization. Like histograms, they do a similar job of showing scale differences. I also calculated the correlation coefficients between the predictor variables and the target variable in order to numerically show the differences before and after standardizing.

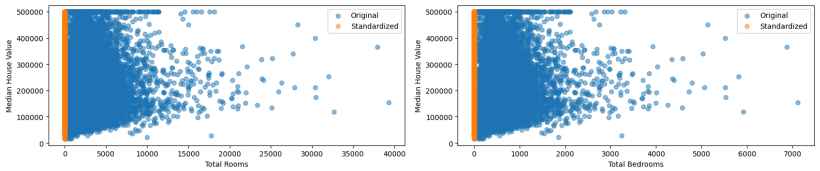


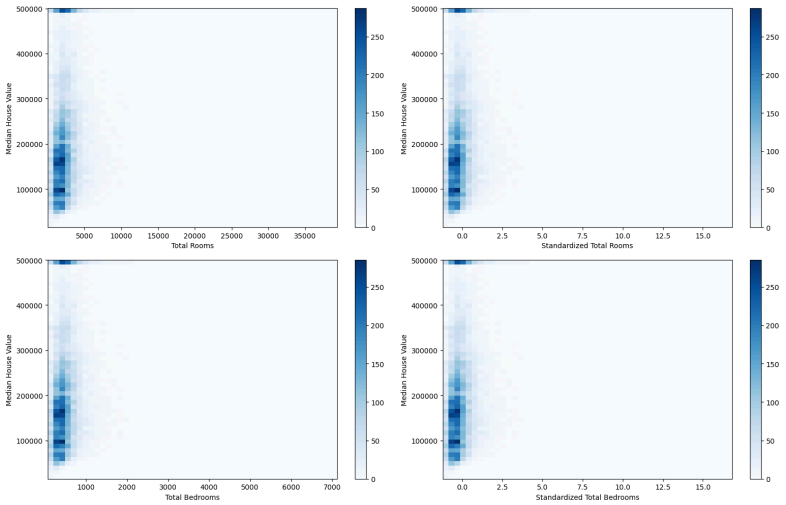






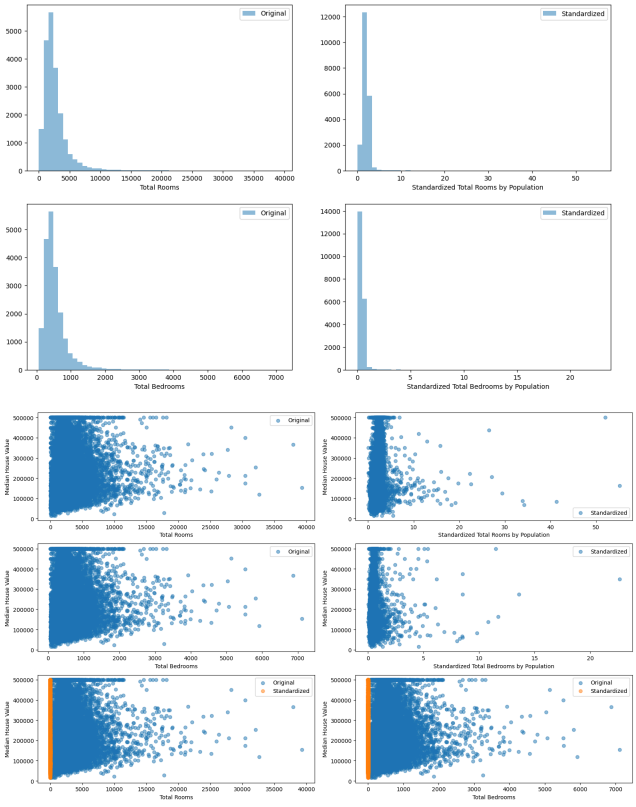







```
Correlation coefficient (Original Total Rooms vs. Median House Value): 0.1341531138065631
Correlation coefficient (Standardized Total Rooms vs. Median House Value): 0.1341531138065631
Correlation coefficient (Original Total Bedrooms vs. Median House Value): 0.1341536985700889
Correlation coefficient (Standardized Total Bedrooms vs. Median House Value): 0.13415369857008877
```

As we can see, the histograms look like the same. In the box plots, there is a slight downward shift after standardizing, which suggests a reduction in the spread of the data. This reduction indicates that standaradization helps make the distribution denser as the data points are closer to the mean and less widely spread out. This shift towards a mean of 0 and a standardi deviation of 1 also makes it easier to compare and analyze variables with each other that originally had different scales. The QQ plots and the density plots also have no huge differences. From our QQ plots, we only know that our data deviates from the theoretical normal distribution starting from quintile 2, which might indicate that our data is heavily tailed. From the scatter plots, we can see how standardization improves the interpretation of scatter plots by removing the effect of scale differences. We can also see that the correlation between the predictor variables and the target variable becomes slightly stronger after standardizing. So far we are studying the standardizing the variables by themselves. Let's look at how standardizing by another variable is an even better idea. For example, let's standardize the two variables by another variable population. I believe this is necessary because the raw number of rooms in a block is not going to be a very meaningful predictor of median house price compared to if we divide it by the number of people to better predict the median house value. When we take the population into account, it starts to get meaningful as we are doing our best to try to mimic the ideal case where the blocks would have equal sizes. We are trying to get a sense of how big the average house is by doing this dividing.

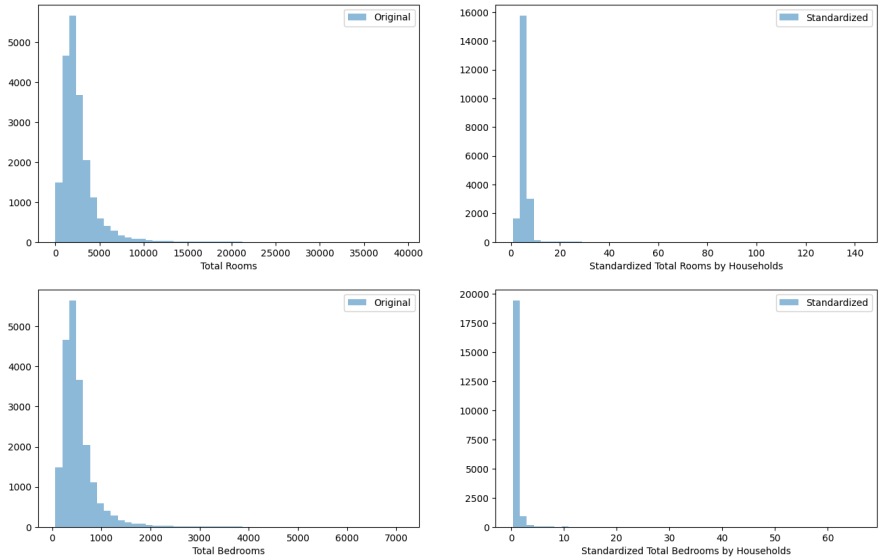


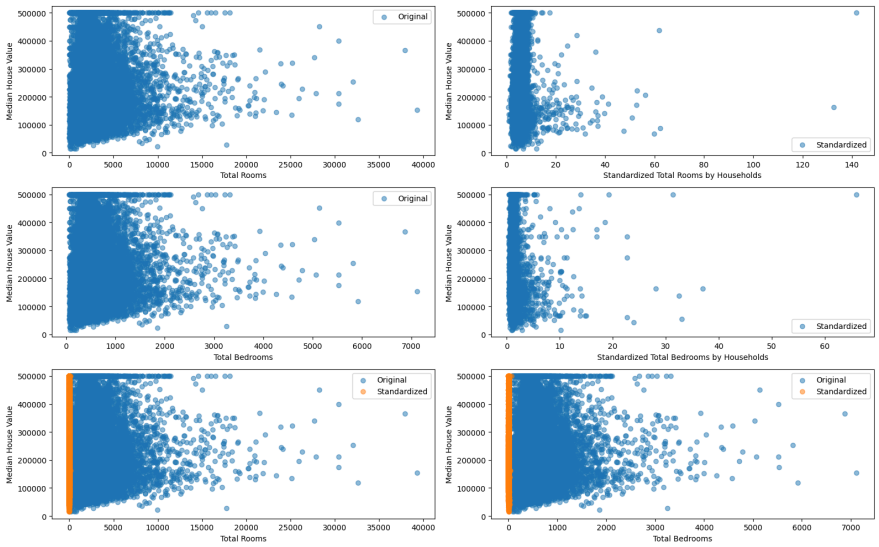

```
Correlation coefficient (Original Total Rooms vs. Median House Value): 0.1341531138065631
Correlation coefficient (Standardized Total Rooms by Population vs. Median House Value): 0.20948196900668967
Correlation coefficient (Original Total Bedrooms vs. Median House Value): 0.1341536985700889
Correlation coefficient (Standardized Total Bedrooms by Population vs. Median House Value): 0.11309509846221796
```

Similarly, I plotted some histograms and scatter plots and calculated the correlation coefficients to compare before and after standardizing. This time, they differ greatly. There is a much more condensed distribution, suggesting that standardizing by another variable more effectively reduces variability and makes them meaningfully comparable across blocks, making it easier and more logical to identify patterns or relationships. The substantial increase in correlation after standardizing total rooms by population (compared to standardizing by itself) suggests it becomes a stronger predictor of the median house value. The slight decrease in correlation after standardizing total bedrooms by population may indicate that the relationship between total bedrooms and median house value is weaker when predicting median house value. One possible reason is that the number of bedrooms per person may not be as strongly related to median house value as the number of rooms per person. This is perfectly understandable because houses have different sizes and room arrangements and people have different living preferences, let alone other socio-economic factors that influence housing demand. Overall, standardizing is so beneficial that we can ignore this slight decrease given the logic behind this variable. All the evidence shows that the variables become more meaningful and relatively stronger predictors of median house value after standardizing. The same reasoning applies to predictor variables 4 and 5. Just like variables 2 and 3, the raw variables by themselves are not so meaningful if not standardized by other variables. We can see how much difference standardizing makes from the evidence above. From a social perspective, the size of the population and households does not really or directly impact the prices of the house because of all those other factors. So we have to take those factors into account. One way is standardizing. Why and how predictor variables 4 and 5 by themselves are not very useful at predicting median house value is later illustrated in Question 3 as well.

Question 2:

Like what I did at the end of Question 1, I plotted the histograms and scatter plots as well as calculated the correlation coefficient of standardizing variables 2 and 3 with 5 in order to compare to standardizing them with 4 to know which way is better.





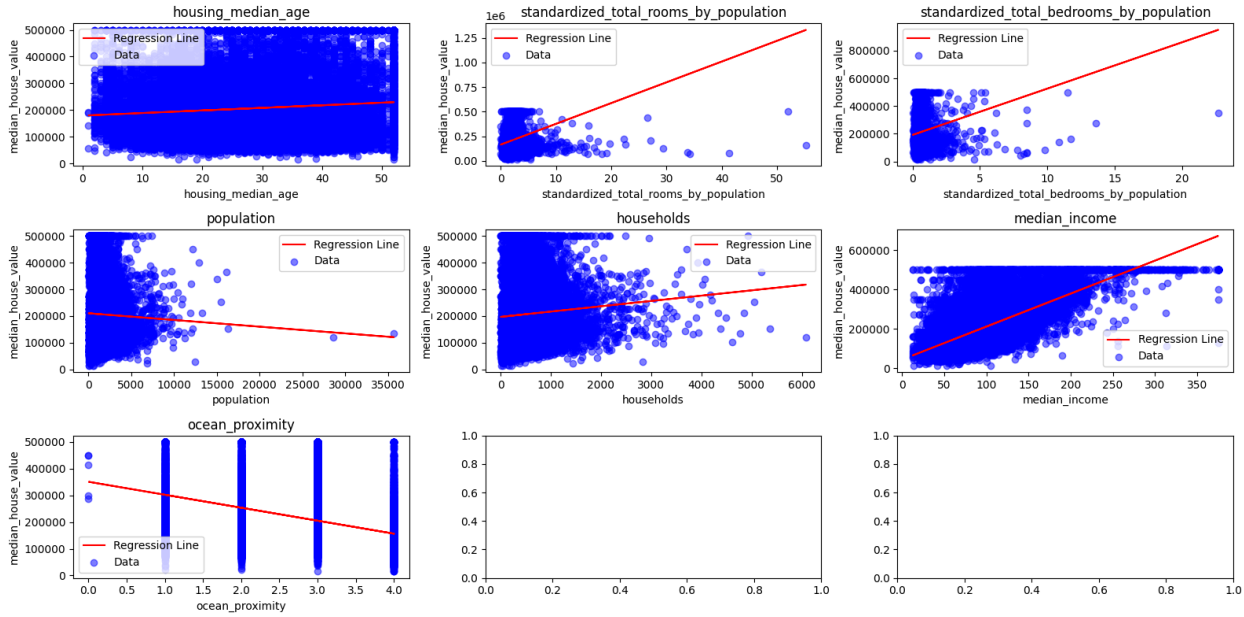

```
Correlation coefficient (Original Total Rooms vs. Median House Value): 0.1341531138065631
Correlation coefficient (Standardized Total Rooms by Households vs. Median House Value): 0.15194828974145796
Correlation coefficient (Original Total Bedrooms vs. Median House Value): 0.1341536985700889
Correlation coefficient (Standardized Total Bedrooms by Households vs. Median House Value): 0.0582604339126752
```

As we can see, the histograms and scatter plots of variables 2 and 3 standardized by 4 and 5 look similar. So we have to look into details by looking at the correlation coefficient. When standardizing by 4, the correlation coefficient of 2 and median_house_value increases by 0.07 after standardizing while the correlation coefficient of 3 and median_house_value decreases by 0.02. In contrast, when standardizing by 5, the correlation coefficient of 2 and median_house_value increases by 0.02 while the correlation coefficient of 3 and median_house_value decreases by 0.08. Since standardizing by 4 causes a greater increase and lower decrease than standardizing by 5, we would safely conclude that it is better to normalize 2 and 3 by 5 than 4.

Question 3:

I did simple linear regression for each of the seven predictor variables. For 2 and 3, I used the better standardized ones from Question 2. I fitted a simple linear regression model using each predictor variable and median_house_value. I then plotted the scatter plot of each predictor variable against median_house_value, along with the regression line. I finally calculated the coefficient of determination (R-squared) to quantify the predictiveness of each variable.

R-squared measures the proportion of the variance in median_house_value that's explained by the predictor variable. A higher R-squared indicates a higher proportion being explained by the predictor variable, which suggests a better fit of the model to the data. Using this logic, the one that has the highest R-squared value among the seven variables is the most predictive variable. Conversely, the one that has the lowest R-squared value among the seven variables is the least predictive variable.



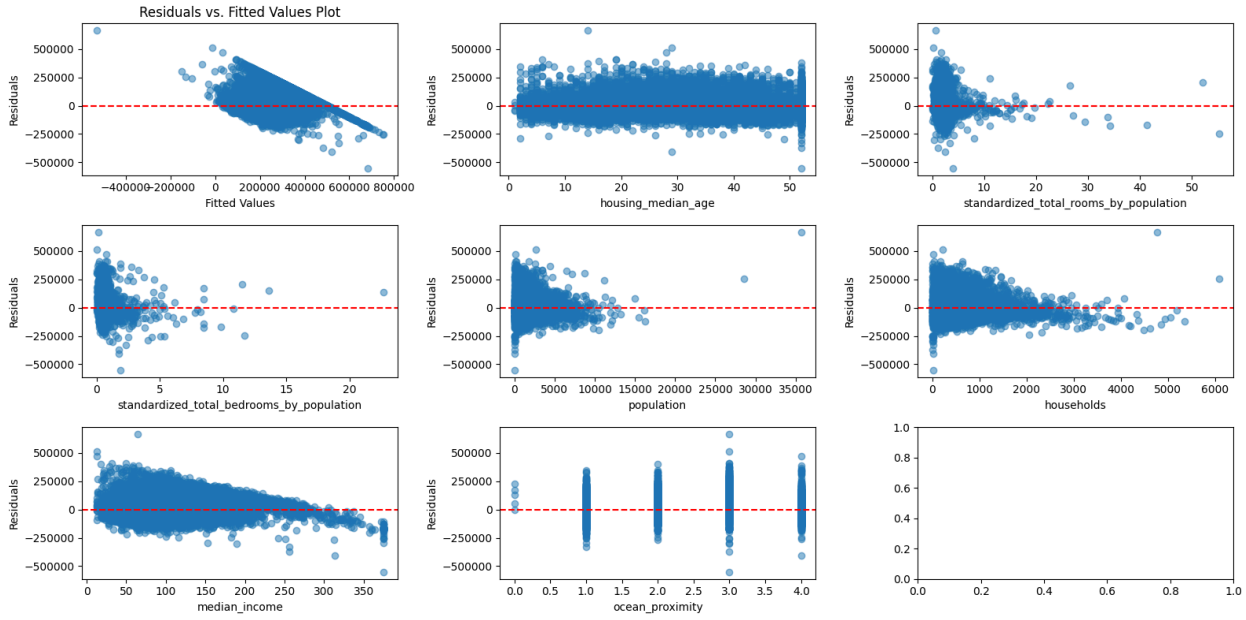

```
R-squared value for housing_median_age: 0.011156305266710853
R-squared value for standardized_total_rooms_by_population: 0.043882695338919864
R-squared value for standardized_total_bedrooms_by_population: 0.012790501296178869
R-squared value for population: 0.0006076066693256887
R-squared value for households: 0.0043352546340906795
R-squared value for median_income: 0.47344749180719903
R-squared value for ocean_proximity: 0.15780848616855125
```

According to the scatter plots and the calculated R-squared values, the median income is the most predictive variable because it has the highest R-squared value. Conversely, the population is the least predictive variable because it has the lowest R-squared value. And the fact that population and households both have a R-squared value near 0 validates our argument in Question 1 that predictor variables 4 and 5 by themselves are not very useful in predicting the median house value. Notice that the scatter plot of median income looks a little odd compared to the other scatter plots. It has some interesting irregularities. All the data points are capped at a maximum of 50,000. No data points exceed 50,000. It looks like a piece of flatland. This can occur for various reasons. For example, it's possible that there is data collection bias from a potential reporting threshold. It could be also possible that individuals who buy houses value exceeding 50,000 are less likely to participate get counted in this study. It could also be a natural cap due to income inequality or legal constraints such as tax law, social welfare programs, etc. Admittedly, whatever cause it is, such truncation of the data posed quite a challenge for the linear regression. However, we cannot be sure of whether eliminating such limitation would cause this median income predictor variable to be even better or to be no longer the best predictor since we don't know which case or maybe even a combination of cases caused this limitation of data. Some might help make the linear regression more precise but some might make the data deviate away from prediction. For example, if it's simply a data collection mistake (missing data from house value above 50,000), it would result in getting better results; but if it's a difference in subjective purchase preference or pattern from high income to low income, then it may result in more inaccurate results.

Question 4:

I did a multiple regression for the seven predictor variables together. Again, for 2 and 3, I used the better standardized ones from Question 2. I fitted the multiple regression model using the seven predictor variables and median_house_value. Since we cannot directly plot a graph to visualize the relationship between multiple predictors and median_house_value, I chose to plot the scatter plot of residuals vs. fitted values and residuals vs. predictor for each predictor variables, along with the regression line. I finally calculated the R-squared value to quantify the predictiveness of this multiple regression model in order to compare to the simple linear regressions we did in Question 3.

The residuals vs. fitted values plot shows the relationship between the predicted values (fitted values) and the residuals (the differences between the observed and predicted values). It shows whether there are patterns or trends in the residuals, which can indicate violations of the model assumptions. The residuals vs. predictor plots show the relationship between each predictor variable and the residuals. It also identifies any patterns or trends in the residuals that may suggest nonlinearity. For both graphs, the residuals should ideally be randomly scattered around the horizontal line at y=0 with no clear pattern or trend. If there is such a pattern, trend, or irregularity, it suggests that the relationship between the predictors and median_house_value is not sufficiently captured by the multiple regression model.

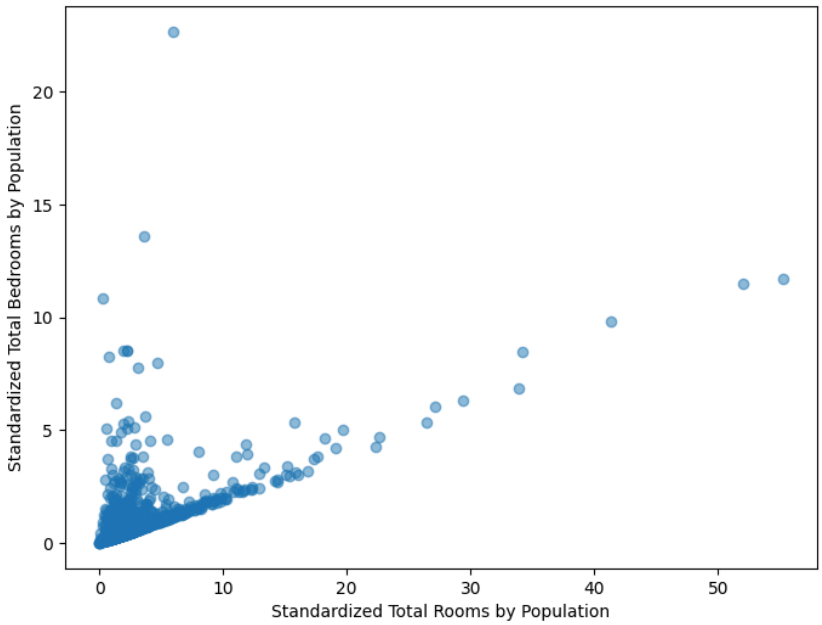


```
R-squared for multiple regression model: 0.6006645246293567
R-squared values for individual predictor variables:
housing_median_age: 0.011156305266710853
standardized_total_rooms_by_population: 0.043882695338919864
standardized_total_bedrooms_by_population: 0.012790501296178869
population: 0.0006076066693256887
households: 0.0043352546340906795
median_income: 0.47344749180719903
ocean_proximity: 0.15780848616855125
```

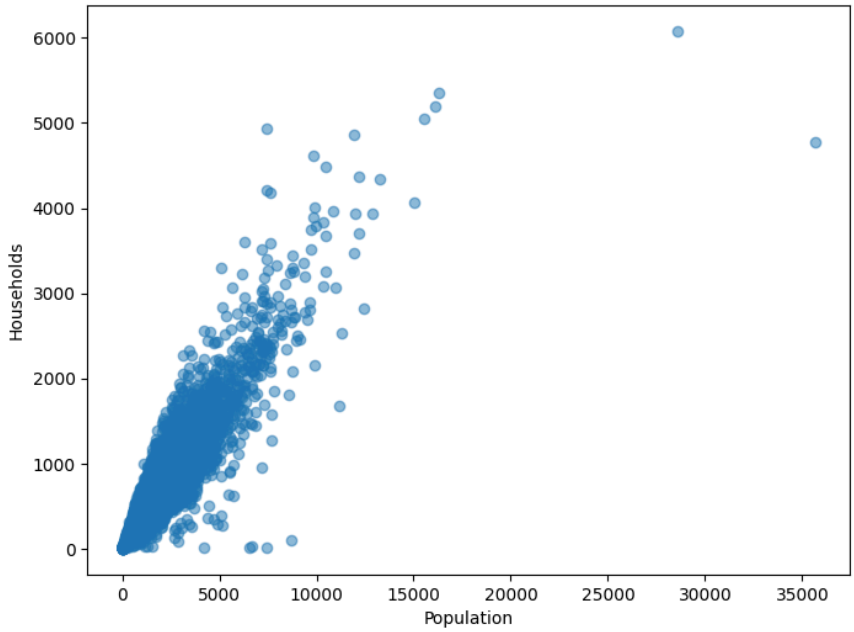
The first plot shows the multiple regression. As we can see from the scatter plots, it appears as a tilted bulk. Although it mostly follows a random pattern, the tiltness is indicating some variables are not so well predicted by the multiple regression. Looking at the individual plots into details, we can see that most irregularities are shown in total rooms, total bedrooms, and population. The households graph is moderately regular. Overall, the multiple regression model appears to be strong at predicting median_house_value. From the R-squared value, we can see that the multiple regression model has a value of 0.6. This is doing pretty well at predicting median_house_value. It is even higher value than the best predictor median income. Therefore, this multiple regression model is better than the simple linear regression models.

Question 5:

I plotted a scatter plot to examine the relationship between standardized variable 2 and 3. I calculated the correlation coefficient in order to check if there is potentially a concern regarding collinearity. Similarly, I did these for variables 4 and 5.



```
Correlation coefficient between standardized variables 2 and 3: 0.6414637002481954
```

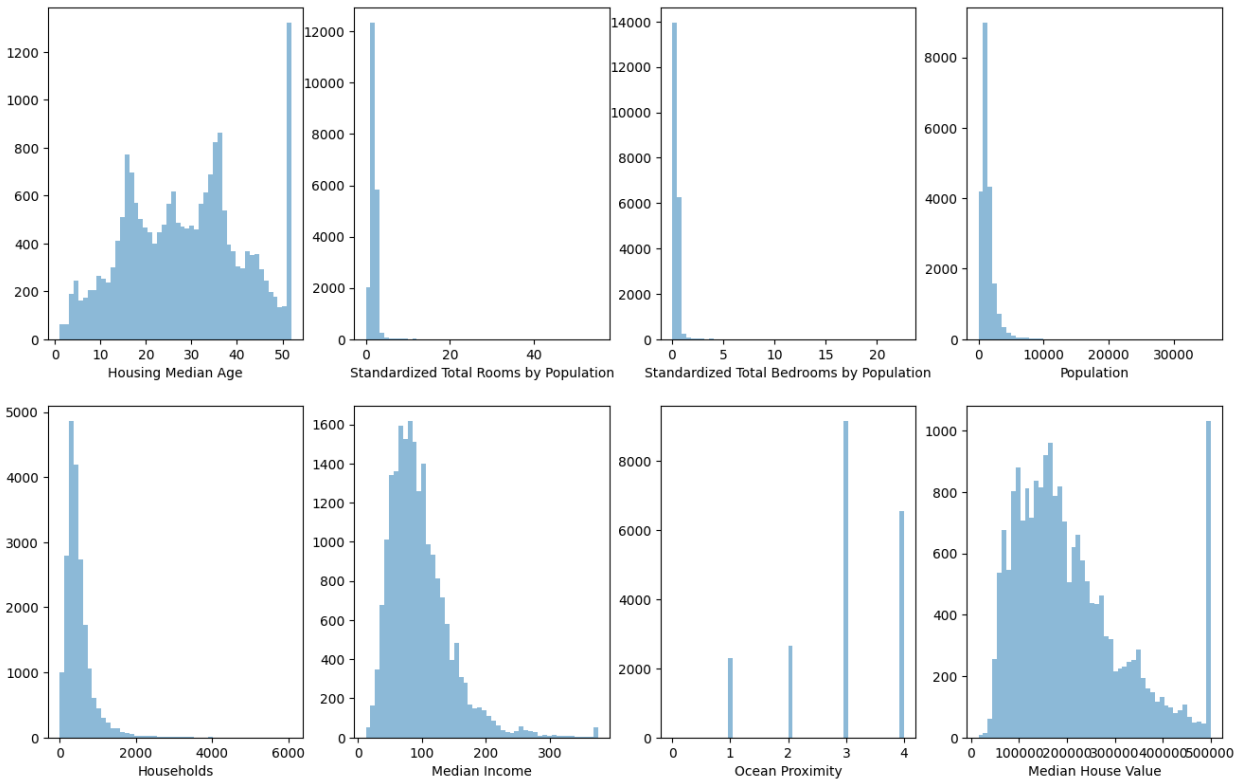



```
Correlation coefficient between variables 4 and 5: 0.9072222660959618
```

As we can see, the 2 and 3 plot appears to be showing two looming linear trends, one at 30 degree and the other at 80 degree approximately. This is verified by the correlation coefficient of 0.64. Such a correlation coefficient suggests a moderate correlation, which does not really imply collinearity. On the other hand, the 4 and 5 plot shows a major linear trend, which is also verified by the correlation coefficient of 0.90. It's a large correlation coefficient that implies collinearity between variables 4 and 5.

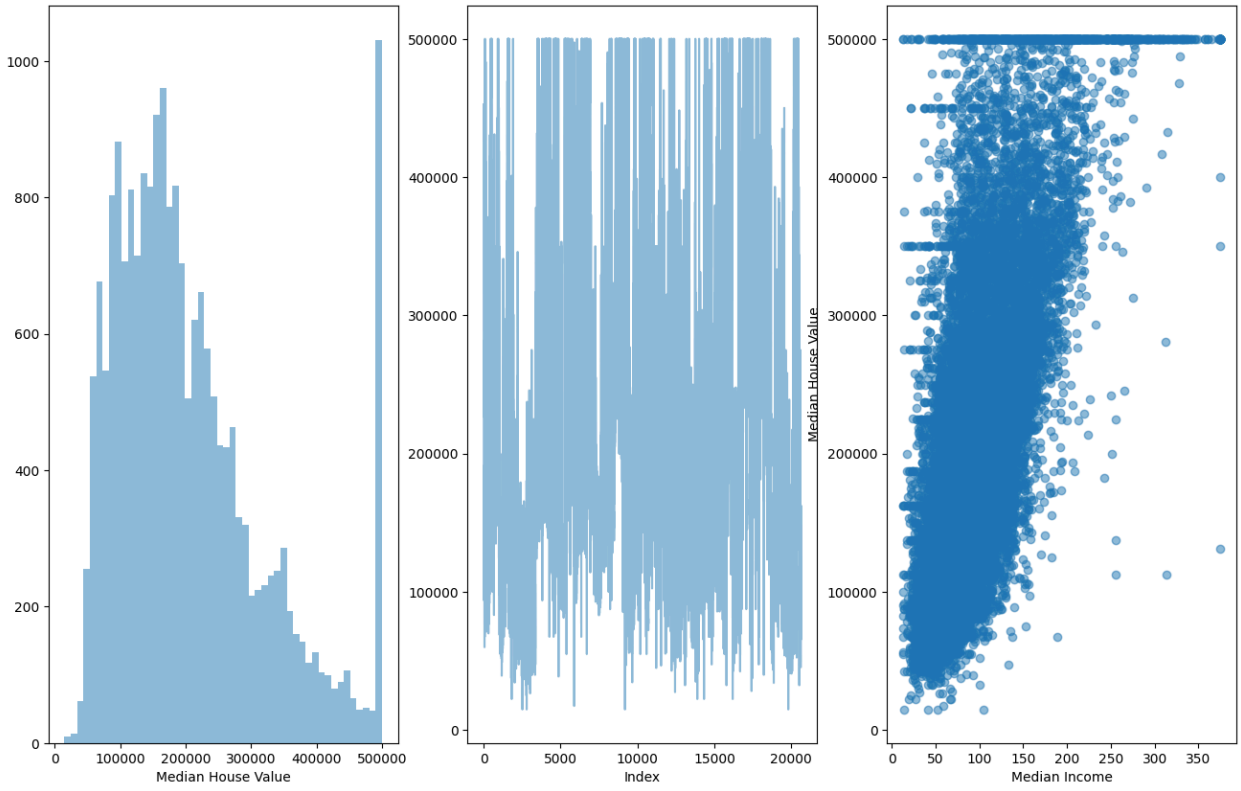
Extra credit:
a) Does any of the variables (predictor or outcome) follow a distribution that can reasonably be described as a normal distribution?

I plotted the histograms of all eight variables (predictor or outcome). As we can see, none of them can reasonably be described as a normal distribution. The housing median age and median house value variables have a distribution that looks a little like that of a normal 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.

I plotted the histogram and scatter plot of the outcome variable to examine its distribution. Just like the irregularity from the median income vs. median house value plot from Question 3 (which is shown to the right), the outcome variable median house value by itself also has this irregularity, with a weird cap at 50,000. This makes sense because it's essentially the same issue that's causing this pattern. The data points in the dataset have a maximum of 50,000. Again, this could have various kinds of reasons. One natural reason is that the collection of data at the beginning of the study does not include housing values over 50,000. This of course limits the validity of the conclusions because they are not valid for those expensive houses.


1000
800
600
400
200
0
0
100000
200000
300000
400000
500000
Median House Value
500000
400000
300000
200000
100000
0
0
5000
10000
15000
20000
Index
Median House Value
500000
400000
300000
200000
100000
0
0
50
100
150
200
250
300
350
Median Income