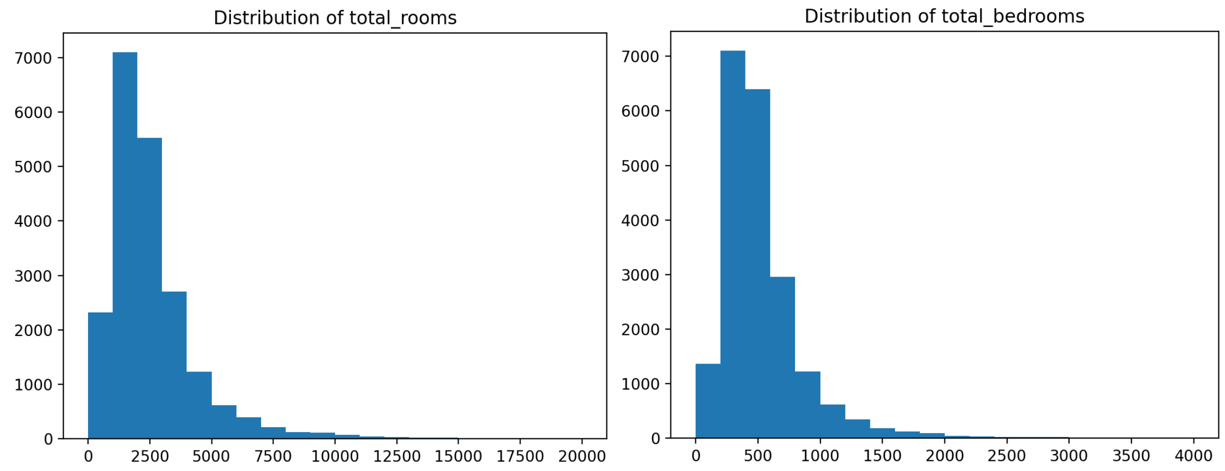
# FML Homework 1 - Erin Choi

## Question 1

I examined the statistics and distributions of each of the columns for predictor variables 2 through 5 using the describe() method and histograms. I also created a correlation matrix of the variables in the dataset, viewed each predictor's correlation with the outcome variable, and computed $r^2$ values for the population and household variables against the outcome.

The describe() method is a simple way to gain summary statistics for all columns of interest, which would allow me to numerically show the range of values for each variable. The histograms show how values are distributed and what ends may contain values that are outliers. The correlation matrix shows how each variable is correlated with the outcome and produces r values that can be used to calculate $r^2$ values.


Distribution of total_rooms

Distribution of total_bedrooms

The histograms for number of rooms (variable 2) and number of bedrooms (variable 3) show that the distributions for both variables are heavily skewed and tend towards lower values. I take this into account when evaluating the summary statistics of both variables. Looking at the statistics for total_rooms, the maximum is likely an outlier, so looking at the interquartile range instead (approximately 3148 - 1448 = 1700) shows that the number of rooms in a block varies a lot. Similarly, the maximum for total_bedrooms is, again, probably an outlier, but the interquartile range of this column is 630 - 324 = 306, which is also quite large. Thus the summary statistics of both columns show there is a lot of variation in the number of rooms and number of bedrooms in a block.

Looking at the correlations of all predictors with the outcome, it is evident that the population and households variables have the weakest correlations with median_house_value: about -0.0247 and 0.0658, respectively. Squaring these correlation coefficients gives even smaller numbers: about 0.00061 and 0.00434; these $r^2$ values represent the proportion of variation in the outcome that can be explained by population or number of households.

It is good to normalize the total numbers of rooms and bedrooms in a block (variables 2 and 3) because, as previously described, the numbers of rooms and bedrooms vary greatly, since block sizes and compositions vary a lot, especially within a state as large as California with  many different kinds of neighborhoods and houses. The number of rooms and bedrooms in a block depends very much on the size and density of the block, so the numbers should be standardized based on some measure of size or density to be made useful for predicting the median house value for the block.

The variables for population and number of households in the block (predictors 4 and 5) are not very useful alone for predicting median house values because these numbers by themselves don't have much to do with the median value of a house on the block. Numbers of people or households on a block aren’t very related to the characteristics of houses that determine their value. As individual predictors, they have very weak correlations with median_house_value, and the variation in their values explain basically none of the variation in the outcome (very small $r^2$). They are, however, indicators of how large or dense a block is, so they can be useful as divisors for total_rooms or total_bedrooms in a standardization process to get total rooms/bedrooms per person or household on the block.

## Question 2

Ideally, we would like to divide the number of rooms and the number of bedrooms (variables 2 and 3) by the number of houses on the block to get an idea of how big the average house is. This data is not available, and intuitively, the number of households is more similar to the number of houses than population. However, using numerical analysis, I found doing the opposite and dividing by population could be more useful.

I “normalized” both variables 2 and 3 by dividing each by variables 4 and 5, then calculated correlations and $r^2$ for each of them with the outcome variable. Computing $r^2$ is a simple way to see how much variation in the outcome can be explained by the variation in individual predictor variables. Determining which variable explains the outcome better as the only predictor in a regression can help in choosing which variables to keep for multiple regression.

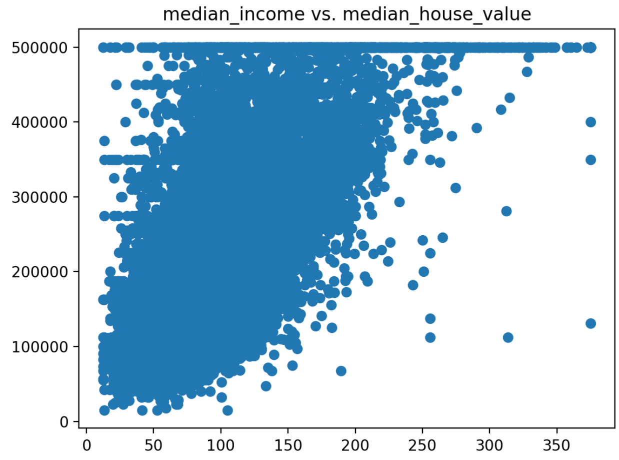
I found that the correlations of the outcome with rooms_per_person (0.2095) and bedrooms_per_person (0.1131), normalized by dividing by population, are stronger than those with their corresponding variables normalized by dividing by household (0.1519 and 0.0583, respectively). The $r^2$ values for standardizing by population - 0.0439 and 0.0129 - are, in turn, higher than standardizing by household as well (0.0231 and 0.0034). Thus I chose to continue using the variables standardized by population. I later dropped the population variable from the dataset in Question 4 to avoid collinearity between the normalized variables and population.

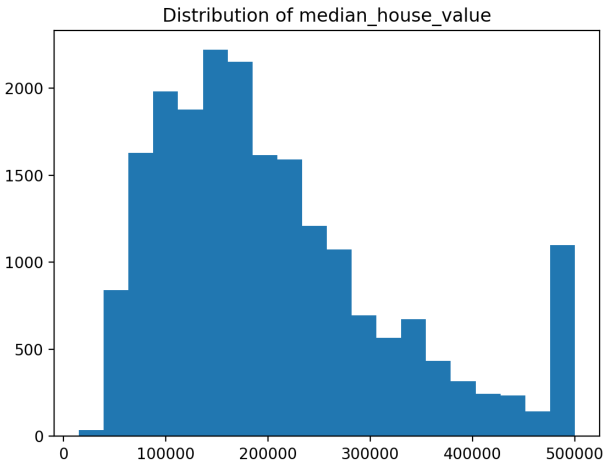
## Question 3

I checked the updated correlation matrix of the outcome variable with all predictor variables, including the normalized variables. I then performed a train/test split on the data, with 20% of the data being allocated to the test set, and fit a linear regression on each predictor variable with the outcome variable to compute the $R^2$ for each model.

The correlation matrix was the first step in seeing which predictors had the strongest and weakest relationships with median house value. The linear regressions helped check which variables produced the models that predicted the outcome best individually.

Both computing correlations and performing simple linear regressions gave the same results. Median income had the highest correlation with housing value (0.6881) and produced the model with the highest $R^2$ (0.4467). Population had the weakest correlation with the outcome (-0.0247) and the model with the lowest $R^2$ (-0.0002); the negative $R^2$ indicates the model is guessing worse than just predicting the mean of the outcome. Therefore I concluded that median income is the most predictive of housing value, and population is the least predictive.





I plotted the median income values against median house values and noticed a strange line of data points at about the maximum value of the outcome, which prompted me to examine the distribution of the outcome with a histogram. This confirmed that there are many values belonging to the rightmost bin, and applying value_counts() to the median house value column showed that there were 965 blocks with a housing value of $500,001.

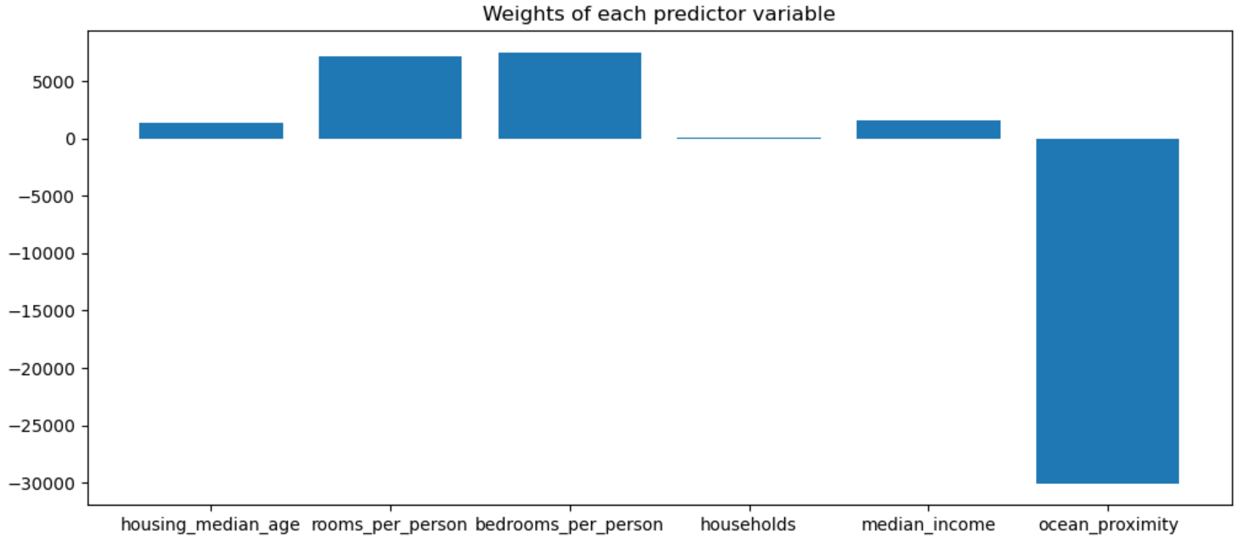
I suspect that there was some kind of capping that occurred during data collection; it seems that blocks with housing value medians greater than $500,000 were listed to have medians of $500,001. Having more accurate data over the $500,000 mark for median housing value would have allowed for a stronger correlation between this predictor and the outcome, which would let it be even more predictive of the outcome.

## Question 4

As prefaced in Question 2, I dropped the population column before performing multiple regression to avoid collinearity. I went back to the simple regression using only median income (the best predictor) to compute predictions and generate the root mean squared error (RMSE) of the model along with $R^2$. I then performed multiple regression using all predictor variables and computed $R^2$ and RMSE for that model as well before graphing the weights of the predictors.

While there is a tradeoff between $R^2$ and RMSE, I still wanted another metric to compare the two regressions by, so I chose to compute both metrics for each model. Graphing the weights allowed me to see how important the predictors are to the multiple regression model relatively, which can be compared to their performance as predictors just on their own.

Linear regression using only median income yielded a model with an $R^2$ of 0.4467 and RMSE of 84941.05. Meanwhile, multiple regression with all predictors produced a model with an $R^2$ of 0.5664 and RMSE of 75194.81. According to the chart below, the ocean_proximity variable had the greatest weight in the full model, while the households variable had the least.



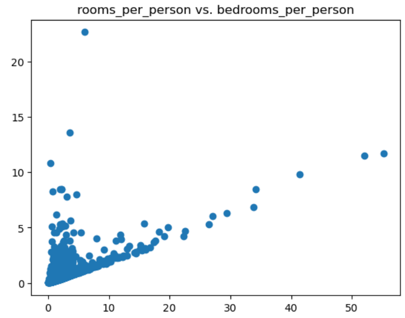
The full model outperformed the other model in both measures. It predicts the outcome about 57% of the time versus about 45% of the time for the single-predictor linear regression model. Its RMSE is lower than that of the simpler model. Ocean_proximity had the greatest weight in the model by a large margin, though it was the second-best predictor identified in Question 3 based on correlation and simple linear regression. It was also the only variable with a negative weight. The best predictor from Question 3, mean income, was actually one of the variables with less weight relative to the other predictors.

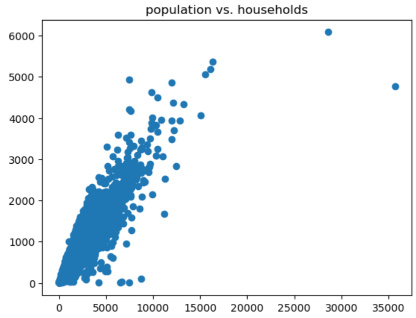
## Question 5

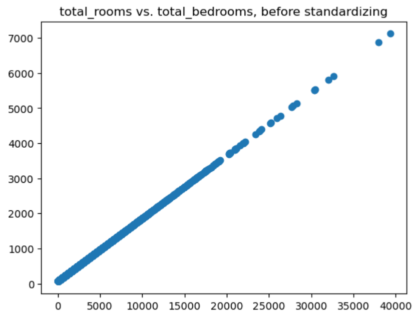
I computed the correlations and plotted the relationships between the standardized variables 2 and 3 and between variables 4 and 5. I also examined the relationship between the unstandardized variables 2 and 3.

Having both a number and visualization to evaluate correlation is important as there could be relationships that are strong but nonlinear, and looking at a relationship before and after standardizing provides another level of understanding of potential collinearity.

The correlation between rooms per person and bedrooms per person (variables 2 and 3) is 0.6415, and the correlation between population and household is 0.9072. The correlation between total rooms and total bedrooms (before standardizing) is extremely close to 1. The plotted relationships between all pairs of variables are shown below.



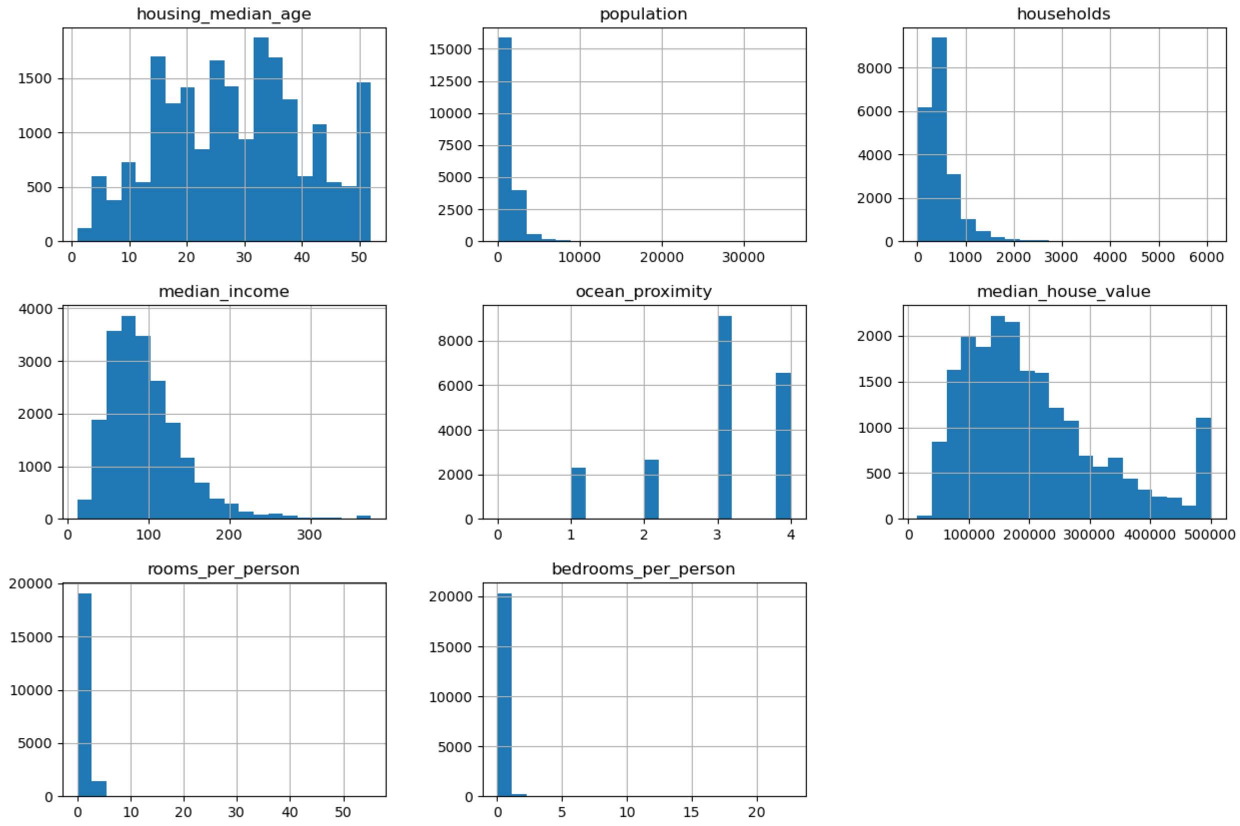




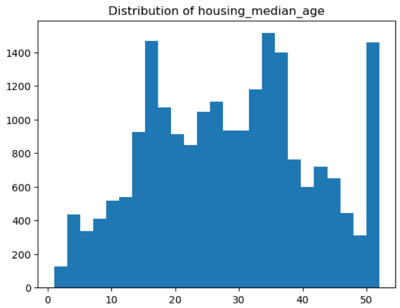
Considering the almost perfect correlation between predictor variables 2 and 3 before standardizing, the relationship between the standardized variables is not as strong in terms of linear correlation; the relationship looks as though a nonlinear model could fit it better post-normalization. Standardizing the variables did decrease the strength of their linear relationship, but there may still be a collinearity concern since r=0.6415 still represents a decently strong linear relationship. If population was included in the full model, there would also be another concern when looking at its relationship with the households variable, as their relationship is visually strong and linear and has a strong positive correlation of 0.9072.

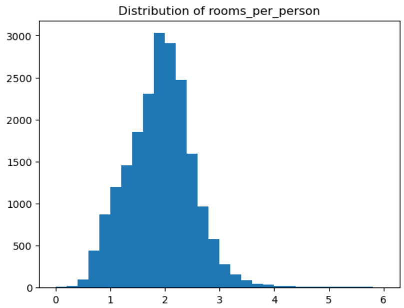
## Extra Credit Part A

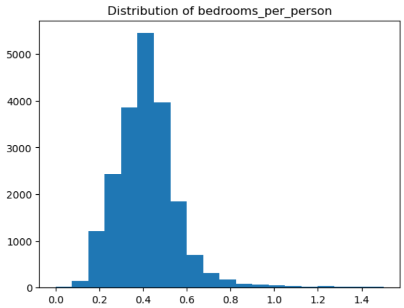
I initially plotted histograms for all variables to take a rough look at their distributions, then plotted some of them individually, changing the graphs' ranges and number of bins to inspect them more closely. Some graphs appear very zoomed out due to outliers, so to see the actual shape of the majority of the data, the histograms' ranges had to be adjusted.



Many of the initial histograms (population, households, median income, median house value) look right-skewed. Ocean proximity is not a continuous variable. I had predicted that housing_median_age would have an approximately normal distribution, but upon closer inspection, it appears closer to bimodal when the peak at the rightmost end of the distribution is not considered. The distributions of the standardized variables were hard to see, so I decreased the histograms' ranges. Compared to the unstandardized versions of the variables, the normalized ones are certainly closer to being normally distributed. Of all the variables, these two variables' distributions are closest to a normal distribution, but they still have long right tails.



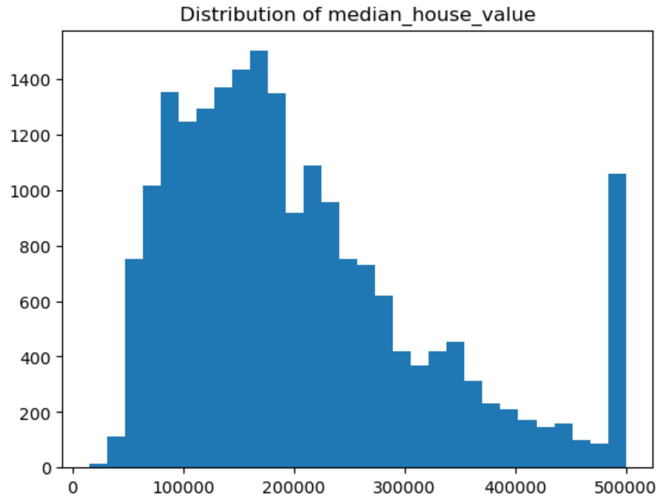




## Extra Credit Part B

I plotted a histogram to examine the distribution of the median house value variable. The distribution is heavily right-skewed except for the high number of large values at the very end of the tail. I know from Question 3 that there is a high number of blocks recorded to have a housing value of $500,001 but that likely have higher actual housing values.

This limitation from the data (specifically from the data collection process) prevents predictors, like median income, from having stronger relationships with the outcome variable and thus limits how predictive the predictor variables can be. This may affect the weight of each predictor included in the full model and impact the evaluation metrics that result from fitting both the full model and the single-best-predictor model. Therefore the suspicious characteristic in this distribution may be negatively affecting the validity of my results and conclusions.


Distribution of median_house_value