Tomisin Adeyemi
Fundamentals of 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?

This is more of a conceptual question, so I began answering it intuitively, later I found evidence in the data from my initial exploratory data analysis to back up my points. (I know this is kind of a taboo, but it is what it is). The main evidence to support my point was from the correlation matrix I had initially plotted to understand the data better.

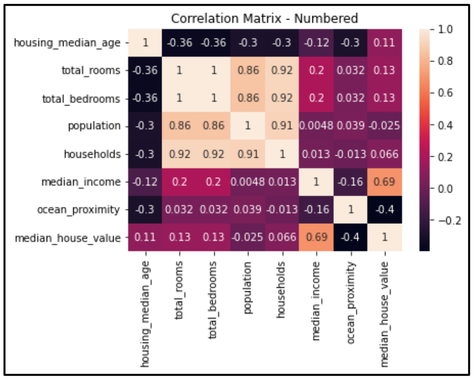
Number of rooms/bedrooms in each block

There isn't a "standard" block amount: each block has a different size. As a result, raw total number of rooms in a block is not going to be a very meaningful predictor of median house price. This is because the total number of rooms/bedrooms is dependent on the block size itself. One would expect that a housing block with more people will have more rooms, and housing blocks with less people would have less rooms (except the edge case that densely populated areas are more impoverished, and scarcely populated areas are more affluent, but let us ignore this for now). Because of this possible correlation between number of rooms/bedrooms and population/number of households (correlation matrix supports this by the way) in a block, we need to think of the number of rooms/bedrooms as a function of the population or number of households. It would be better to standardize by population or number of households to get a better idea of the effect of the relative number of rooms on median house value.

4 and 5 – Population/Number of Households

Like the number of rooms/bedrooms above, the total population/number of households in a block depends on the size of the housing block, which varies in this dataset. Like above, population and number of households are not very useful unless they are standardized.

In the correlation matrix below, we can see that the correlation between total rooms/bedrooms and population/households are 0.86 and 0.92 respectively, suggesting a very strong relationship between the variables.



	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
housing_median_age	1	-0.36	-0.36	-0.3	-0.3	-0.12	-0.3	0.11
total_rooms	-0.36	1	1	0.86	0.92	0.2	0.032	0.13
total_bedrooms	-0.36	1	1	0.86	0.92	0.2	0.032	0.13
population	-0.3	0.86	0.86	1	0.91	0.0048	0.039	-0.025
households	-0.3	0.92	0.92	0.91	1	0.013	-0.013	0.066
median_income	-0.12	0.2	0.2	0.0048	0.013	1	-0.16	0.69
ocean_proximity	-0.3	0.032	0.032	0.039	-0.013	-0.16	1	-0.4
median_house_value	0.11	0.13	0.13	-0.025	0.066	0.69	-0.4	1

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)?

To answer this question, I added 4 new columns to my DataFrame: `rooms_per_household, bedrooms_per_household, rooms_per_person, and bedrooms_per_person`. Each of these columns were the result of taking the total number of rooms/bedrooms in each housing block and dividing them by the total population/number of households in the block. This was done automatically in Pandas. After, I found the correlation between each of these new 4 features and the median house value using the pandas corr() function. In addition, I ran linear regression on these 4 variables and retrieved the R^2 value. I did all the linear regression and evaluation using the scikit-learn LinearRegression function.

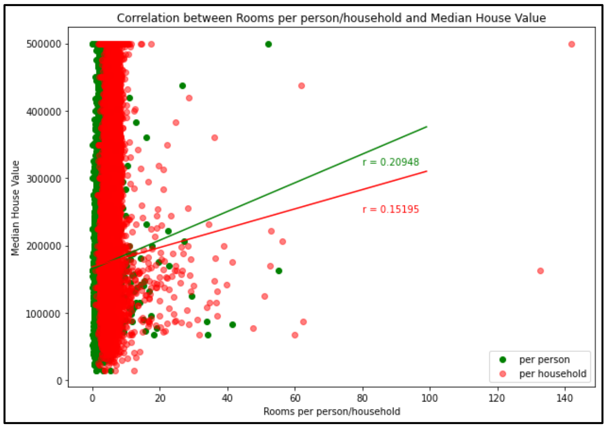
I used the correlation first because it is a good way to describe the relationship between 2 variables. Additionally, a strong correlation between a predictor and an outcome could (possibly) mean that the predictor is a good predictor of the outcome. To further ascertain my suspicions, I also computed the R^2 value to examine how much of the variance in median house value that each variable accounted for. A more intuitive approach could have been to just use the number of households in the block as the standardizer because we are trying to predict median household value. Although this is intuitive, the data suggests that population is a better standardizer.

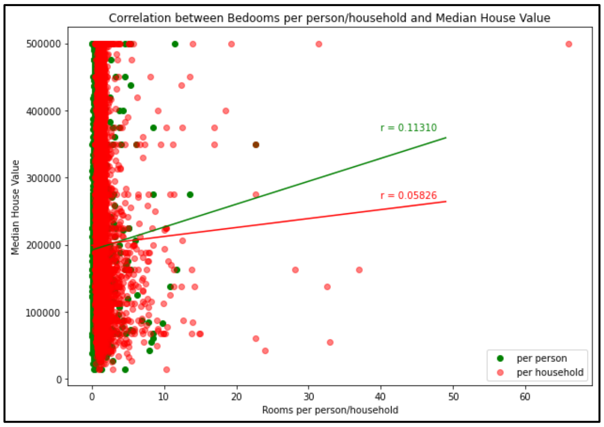
As to what was found numerically, and visually:

The respective correlations between the newly created variables and `median_house_value` are as follows:

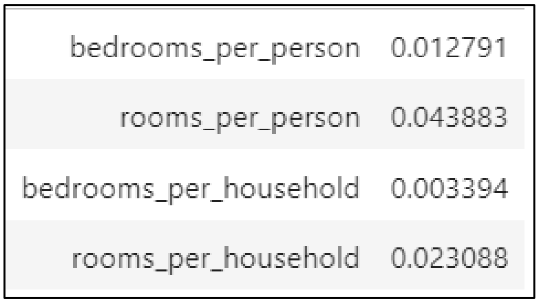
```
rooms_per_household       0.151948
bedrooms_per_household    0.058260
rooms_per_person          0.209482
bedrooms_per_person       0.113095
```

In addition, the correlation plots are as follows:





Finally, the R^2 value for each of the values are shown in the table below:



bedrooms_per_person	0.012791
rooms_per_person	0.043883
bedrooms_per_household	0.003394
rooms_per_household	0.023088

As to my interpretation of the findings:

The better pick, according to the data, is to standardize per population as there are stronger correlations and R^2 values than their household counterparts.

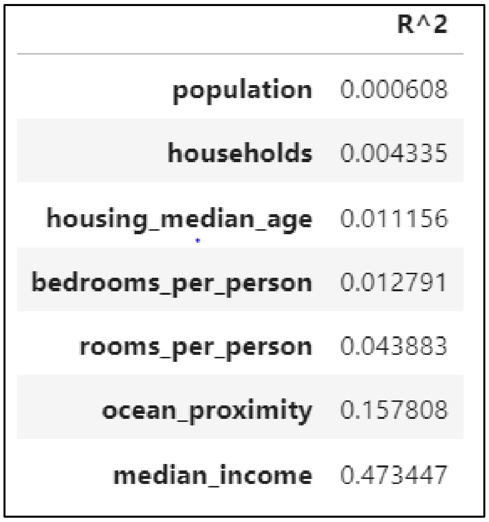
3. Which of the seven variables is most *and* least predictive of housing value, from a simple linear regression perspective? [Hints: a) Make sure to use the standardized/normalized variables from 2. above; b) Make sure to inspect the scatter plots and comment on a potential issue – would the best predictor be even more predictive if not for an unfortunate limitation of the data?

First, I dropped `total_rooms, total_bedrooms, rooms_per_household, and bedrooms_per_household` from the dataset. Then I created the X variable by dropping median house value from the dataset and isolating median house value as the y variable. Afterwards, I made linear regression models for each predictor variable using scikit-learn. I also computed the R^2 values and plotted the actual vs predicted outcome for each predictor variable. I also extracted the R^2 values for each column and created a DataFrame of the sorted values.

As explained in number 2, R^2 explains the portion of the variance in median house value that each variable accounts for. While the correlation, r, could hint at one predictor being better than another, R^2 quantifies the effect of one variable on another based on Linear Regression. R^2 is also a metric commonly associated with linear regression specifically, r, not necessarily.

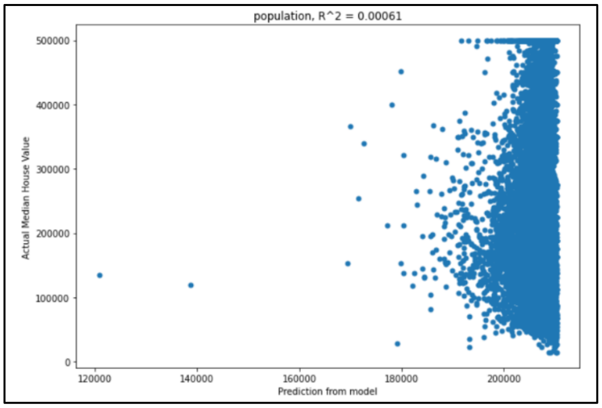
As to what was found numerically, and visually:

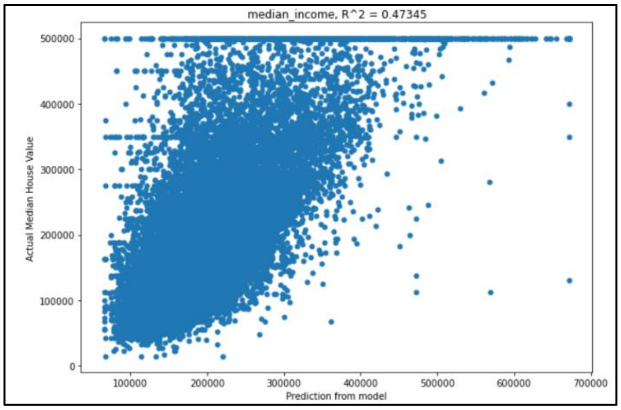
The table of correlation values in order of increasing R^2 values:



	R^2
population	0.000608
households	0.004335
housing_median_age	0.011156
bedrooms_per_person	0.012791
rooms_per_person	0.043883
ocean_proximity	0.157808
median_income	0.473447

From this table, we can see that population is the least predictive from a linear regression perspective, and median income the most predictive from a linear regression perspective. In addition, here are the graphs of the actual vs predicted values for population and median income:



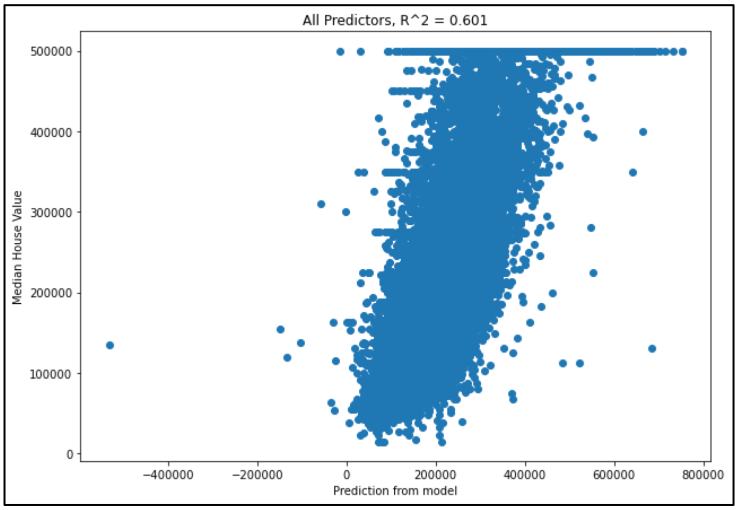


Inspecting the plots, we can see that the predictions from the model using population are mostly concentrated around the 200000 mark, despite the actual values ranging from 15000 to 500000. As with median income, the scatter plot shows that the model does a pretty decent job at predicting the median house value. However, it seems like there are an unusual amount of houses worth 500000. This could be a result of some corruption/imputation in the data. If these $500000 median house values were distributed more normally, the median income could perhaps be an even stronger predictor.

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.?

Like number 3, I extracted the relevant X and y variables from the dataset and input them into scikit-learn's LinearRegression function. I then plotted the graph of the actual vs predicted values. To make the comparison fair, I decided to continue with the same method used in 3 to make comparison easier.

The R^2 value derived from using all the predictors is 0.6006645246293567, suggesting that the model with all the features as predictors accounts for more of the variance than the model with just the median income as the predictor. This model, however, seems to have more extreme predictions, with a couple of negatives predictions, and higher maximum predictions (around 800000). We still see the unusual concentration of houses valued at 500000 in the plot. Here is the graph below:

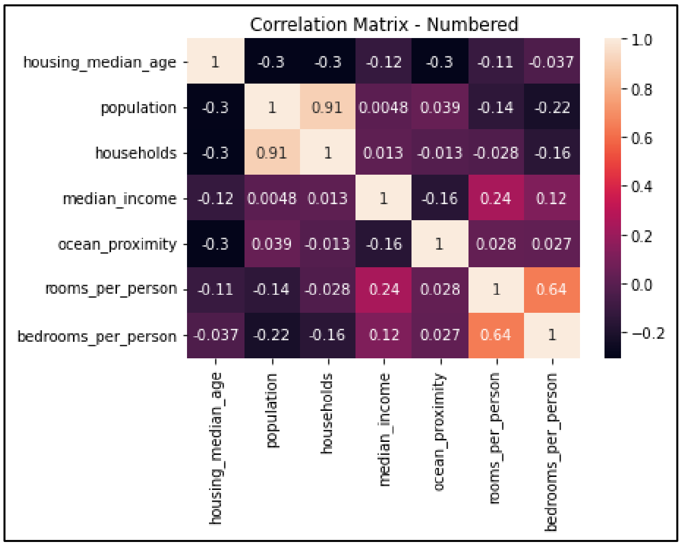


5. Considering the relationship between the (standardized) variables 2 and 3, is there potentially a concern regarding collinearity? Is there a similar concern regarding variables 4 and 5, if you were to include them in the model?

First, I computed the correlation between rooms per person and household per person, and the correlation between population and households. I used the corr() function in Pandas to do this. I then computed the correlation matrix to see how these correlations compare to others and graphed the scatter plots of the variables to visualize the correlation.

Collinearity occurs because of high correlation between predictor variables, leading to an unstable model that allocates betas wrongly. Hence, I decided to look at the correlation between these variables to determine if there was a potential concern regarding collinearity.

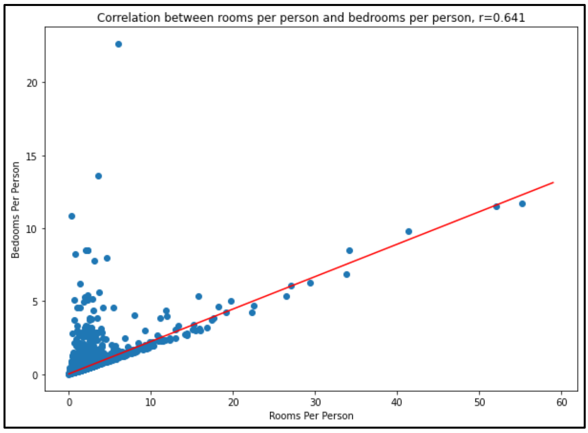
The correlation between rooms per person and bedrooms per person was found to be 0.6414637002481975, while the correlation between household and population was found to be 0.9072222660959659. The picture below shows the correlation matrix, showing that these 2 correlations are the highest among all the predictors in the matrix.



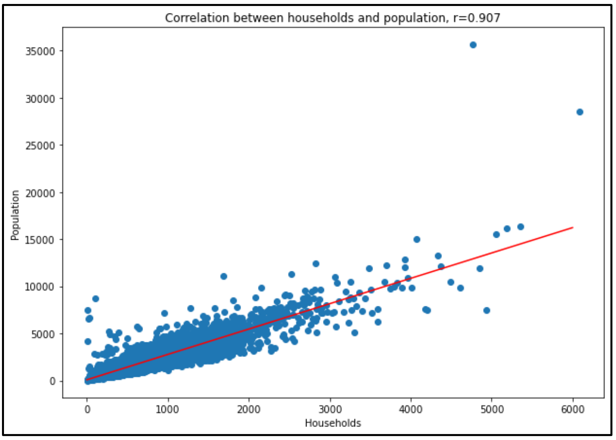
Correlation Matrix - Numbered

	housing_median_age	population	households	median_income	ocean_proximity	rooms_per_person	bedrooms_per_person
housing_median_age	1	-0.3	-0.3	-0.12	-0.3	-0.11	-0.037
population	-0.3	1	0.91	0.0048	0.039	-0.14	-0.22
households	-0.3	0.91	1	0.013	-0.013	-0.028	-0.16
median_income	-0.12	0.0048	0.013	1	-0.16	0.24	0.12
ocean_proximity	-0.3	0.039	-0.013	-0.16	1	0.028	0.027
rooms_per_person	-0.11	-0.14	-0.028	0.24	0.028	1	0.64
bedrooms_per_person	-0.037	-0.22	-0.16	0.12	0.027	0.64	1

The graphs below illustrate the correlation between these variables:



Correlation between rooms per person and bedrooms per person, r=0.641

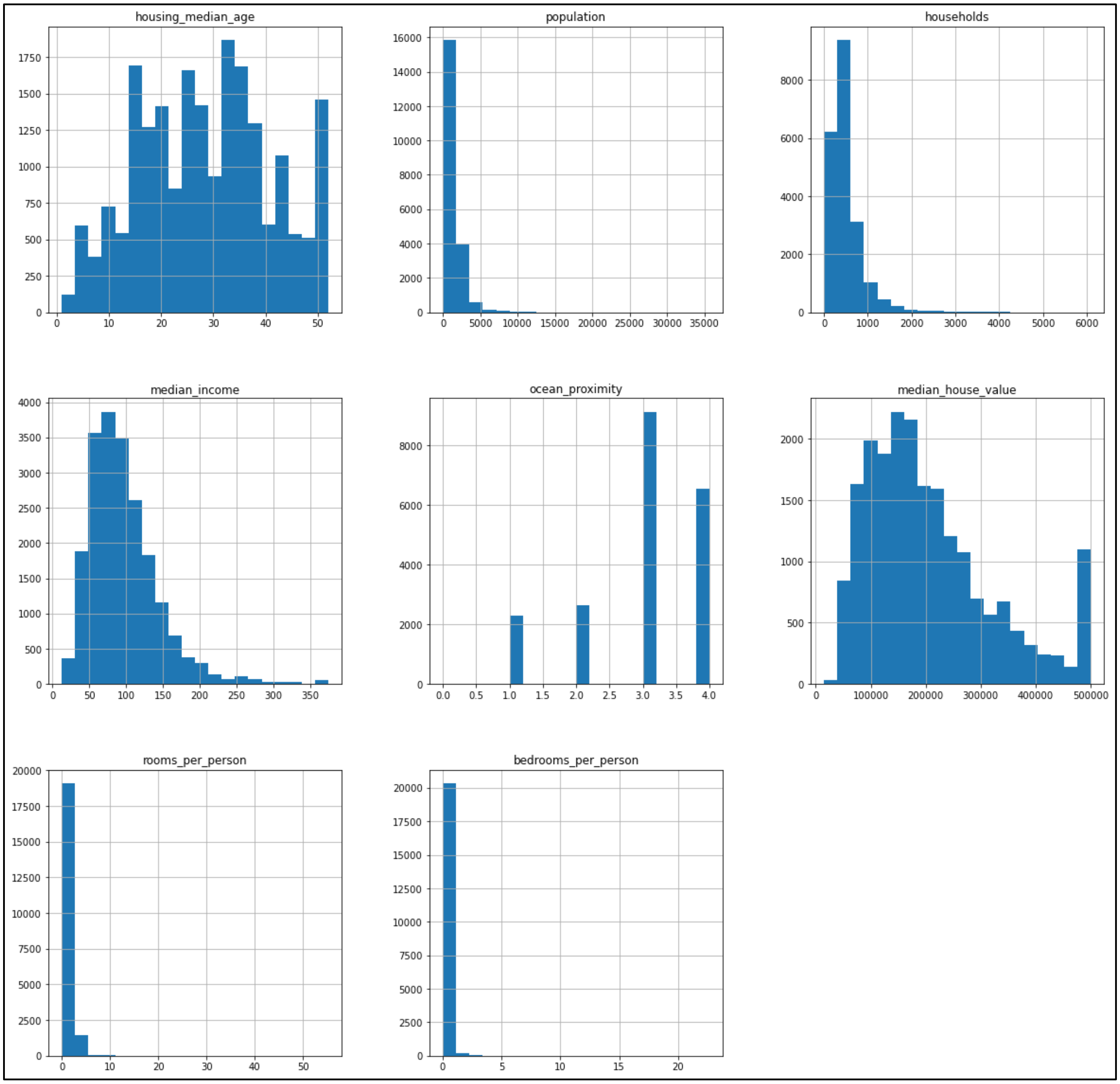


Correlation between households and population, r=0.907

Since both correlations are above 0.5, I'd say there is a possibility of collinearity. For the households and population, we should be very concerned as there is an extremely strong relationship. It also makes sense intuitively; we'd expect blocks with more households to have a higher population. Even though the correlation between rooms and bedrooms per person is not as strong, it is also still stronger than the rest of the variables in the dataset, so should be investigated as well.

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

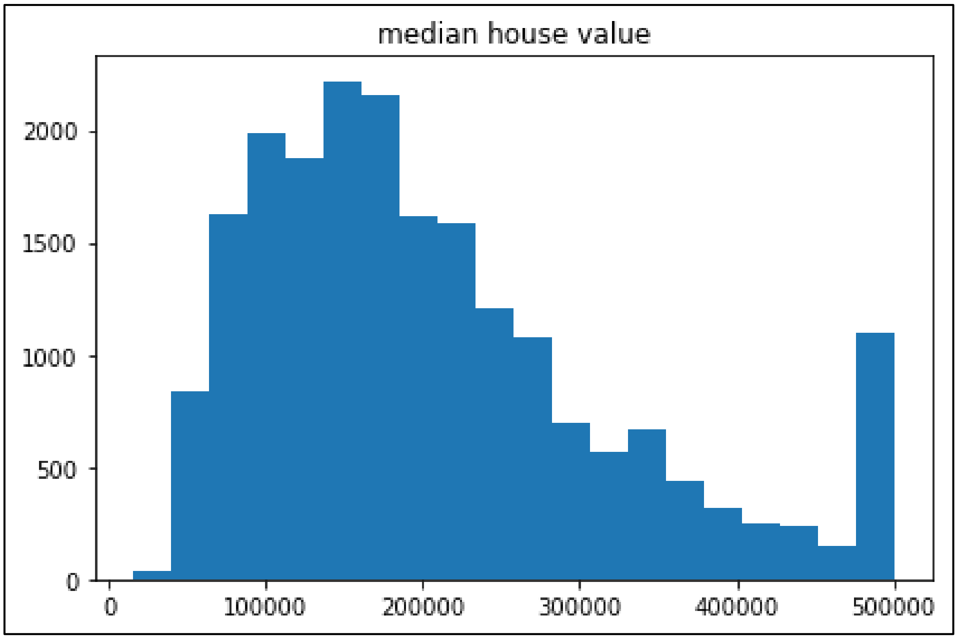
I plotted histograms for each variable in the dataset and visually examined them to determine if any looked similar to a normal distribution. I did this because histograms signify how data is distributed, so this was one of the most intuitive approaches to take to answer this question. The plots are below:



Most of the distributions are left skewed. There is no variable whose distribution resembles a normal distribution.

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 the histogram of the outcome variable to determine the distribution. As explained above, this is because the histogram is the easiest way to examine the distribution visually. The plot is below:



This plot brings us back to the problem encountered in question #3: the data has a scarily fat tail at the 500000 mark, suggesting that there might be something about this data that we don’t know about.