For this Writeup, please refer to the code as I wrote in-line responses for elements 3) and 4) (for **each** question) explaining findings and the interpretation of the findings such that it is relevant to the original question.

1. For Question 1, I split the answer into two parts. For the first part, it was entirely qualitative, to understand the dataset. Code wise, I used the describe function in Pandas to obtain a better view/understanding of the dataset. Then, I noticed that the unit of measurement is in blocks (more so stated directly in the Spec Sheet), and thus both bedrooms and rooms would be stated in terms of “per block,” and not made independent of the number of houses per block, as would be needed for the independent, identically distributed attribute required of predictors for linear regression. Ideally by normalizing/standardizing, the output would be dimensionless and be made of independent of the number of houses in a block (not be affected by differences in houses per block). However, because houses per block is not provided, a proxy is needed. Both population and households are viable choices, and the one to be used will be shown in later questions.

For part 2 of Question 1, I at first used Pandas to obtain a correlation matrix between households and population, the number shown in the notebook. Next, I ran a simple linear regression and plotted the results (with a label of median home price) to find the of both models, plotting them using matplotlib and fitting a red line of best fit. I did this to show the inter-predictor correlation indicative of potential to induce multicollinearity within the model, and then the predictive power (amount of variance explained by each simple linear regression) shown by the of each model.

1. For Question 2, I was stuck between intuition and data. My intuition told me that, to find a proxy for number of houses in a block, each house should contain a single household, which would make it more predictive than a more general population factor. However, upon analysis, this was shown to be false. First in code, I created four new predictors, “population normalized number of rooms,” “population normalized bedrooms,” “total household normalized number of rooms,” and “total household normalized number of bedrooms.” From this, two multiple regressions (using each population and household normalized pairs) were run, as well as four simple regressions, with each individual normalized x (bed)room pair as predictors. A 3d matplotlib plot was then rendered, and shown. From my understanding of the data, this was done due to the higher of the model using population as the normalizing factor. Therefore, more variance is explained though this process, it is the model better representative of the *linear relationship* between predictors and label.
2. For this Question, I used SK-Learn and Pandas to run 11 different linear regressions. The first 4 are each (bed)room normalized by population, household, and the last 7 are the original predictors from the dataset. I then graphed each plot (scatter) and line, as well as each model’s corresponding . This was done to show predictive ability of each model, with median income being the highest (number shown in notebook) and total population in block being the lowest. Of note, there was a second part of this question, for which I plotted the histogram for median hose value. This graph implies the capped maximum median house value of 500,000. I did this to show the abnormality in what can only be explained as an intentional/outside bounds of normal distribution kind of data handling.
3. I ran a complete multiple linear regression using the two normalized (by population) room, bedroom predictors. This was done to determine the , and thus how much variance is explained in this model. The value was then compared to the “best” simple linear regression model in the last question.
4. For this problem, I constructed a collinearity pairplot in Pandas, over all four pop\_norm\_rooms, pop\_norm\_bedrooms, population, and households. I then focused in on the correlation pairplot between pop\_norm\_rooms and pop\_norm\_bedrooms. Furthermore, I created an illustrated a correlation matrix between all four variables, noting the moderate correlation between the two population-normalized predictors. I did this to show the quantitative amount of collinearity (in correlation coefficient) between the two normalized variables, and how much less collinearity (in terms of the correlation coefficient) the normalized variables have compared to the unnormalized variables.
5. I plotted the histograms for all 12 potential variables used in this analysis, to visually determine the normality of each variables’ distribution. I considered running a Kolmogorov-Smirnoff test with the null hypothesis being a normal distribution, but as this was not covered in class – I thought it best not to. I then chose the variable most resembling a bell-curve, indicative of a normal distribution.
6. I plotted the histogram of the label in seaborn, with a kernel density estimation outline around the bins. I did this to see if any incongruous aspects of the label can be determined, as instead of running a KS test this seemed most reasonable.