Question 1

1. I plotted predictor variable 2 against predictor variables 4 and 5 and predictor variable 3 also against predictor variables 4 and 5 on scatterplots. I also calculated the correlation between predictor variable 2 and predictor variables 4 and 5 and predictor variable 3 and predictor variables 4 and 5. Finally, I calculated the correlation between predictor variables 4 and 5 and the outcome variable.
2. I did what I did with predictor variables 2 and 3 to find how correlated predictor variables 2 and 3 were with predictor variables 4 and 5 as this would potentially show if these predictor variables to be modified before being used if they were highly correlated. Furthermore, I did what I did with predictor variables 4 and 5 to observe the relationship between them and the outcome variable.
3. I found that both predictor variables 2 and 3 were highly positively correlated with both predictor variables 4 and 5 (getting correlation coefficients greater than 0.85 for all calculations). This means that as the total number of rooms in a block increases, so does the population and the number of households and that as the total number of bedrooms in a block increases, so does the population and the number of households. Furthermore, I found that predictor variables 4 and 5 had very low correlations with the outcome variable, both resulting in very low absolute values (less than 0.1).
4. It is a good idea to standardize/normalize predictor variables 2 and 3 because these values alone are not meaningful predictor variables of median house value since not all blocks are of equal size in regards to number of houses – a block having a large/small number of rooms/bedrooms does not necessarily indicate the houses in that block having a large/small number of rooms/bedrooms (generally, the total number of rooms/bedrooms increases as the population/number of households increases) which is what would be more important in predicting house value. Standardizing/normalizing these predictor variables would make them more useful in our prediction. Predictor variables 4 and 5 are not very useful by themselves to predict median house values in a block because they aren't giving any information about houses themselves, they are simply just describing numbers of people/groups (which is shown by the very level of correlation between them and the median house value). In summary, predictors 2, 3, 4, and 5 as is all do not correlate with the outcome variable.

Question 2

1. I tried standardizing/normalizing predictor variable 2 by both predictor variable 4 and predictor variable 5 and predictor variable 3 by both predictor variable 4 and 5. From there, I then fit regression models with each of these standardized/normalized predictor variables (so four models total) and calculated the r-squared model for each model.
2. I did this to find out what proportion of the variance in the outcome variable would be explained if I standardized predictor variables 2 and 3 by predictor variable 4 versus predictor variable 5 as whatever explains a higher proportion will be the predictor variable that we should standardize by.

3. I found that when we standardize/normalize by predictor variable 4, we get an r-squared value of 0.044 for predictor variable 2 and an r-squared value of 0.013 for predictor variable 3. Furthermore, I found that when we standardize/normalize by predictor variable 5, we get an r-squared value of 0.023 for predictor variable 2 and an r-squared value of 0.003 for predictor variable 3.
4. Since r-squared values tell us how well the model explains the data, the higher the r-squared value, the better whatever the inputs are of the model predict whatever you are trying to predict (which in this case is housing value). Given this fact, it is better to standardize/normalize the data by population (predictor variable 4). This is because when we standardized/normalized both predictor variables 2 and 3 by predictor variables 4 and 5, standardizing/normalizing by predictor variable 4 produces a higher r-squared value for both predictor variables 2 and 3.

Question 3

1. I fit seven linear regression models with each of the seven predictor variables individually as the inputs and the actual median house value as the output for all the models. I also calculated the r-squared value for each model. To visualize all seven models, I plotted the predicted values against the actual values on scatter plots.
2. I did this to see how well each predictor variable by itself did at predicting housing values.
3. I found that the linear regression model with median household income as the input had the highest r-squared value of 0.473 and that the linear regression model with population as the input had the lowest r-squared value of 0.001. When looking at the scatter plots, I found that a potential issue is the fact that there seems to be a cap on the median house value shown by an anomalous number of data points when the median house value equals 500k.
4. Since r-squared values tell us how well the model explains the data, the higher the r-squared value, the better whatever the inputs are of the model predict whatever you are trying to predict (which in this case is housing value). Since median household income has the highest r-squared value by far, it is the most predictive of housing value. Furthermore, since population has the lowest r-squared value of all the predictor variables, it is the least predictive of housing value. The best predictor, which is median household income, would be even more predictive if not for the outcome variable being cut-off at 500k since this causing the distribution of median house values in the data to not be representative of the real distribution of median house values.

Question 4

1. I fit a regression model with all seven predictor variables as inputs and the actual median house value as the output. I also calculated the r-squared value for the model. To visualize the model, I plotted the predicted values against the actual values on scatter plots.
2. I did this to see how well all of the predictor variables together did at accurately predicting housing values.

3. I found that the linear regression model had a decently high r-squared value of 0.601.
4. Since r-squared values tell us how well the model explains the data, the higher the r-squared value, the better whatever the inputs are of the model predict whatever you are trying to predict (which in this case is housing value). Since the model has decently high r-squared value of 0.601, it does a relatively good job of predicting housing value. The full model does a slightly better job of predicting housing value compared to the model that just has the single best predictor as the r-squared values have a notable difference of 0.128. However, it should be noted that this improvement of 0.128 for the r-squared value came from the addition of six more predictor variables while the best predictor alone could explain a good proportion of the outcome data by itself since the model using it alone already had an r-squared value of 0.473.

Question 5

1. I plotted predictor variable 2 against predictor variable 3 and predictor variable 4 against predictor variable 5 on scatterplots. I also calculated the correlation between predictor variable 2 and predictor variable 3 and predictor variable 4 and predictor variable 5.
2. I did this to observe the relationship between predictor variables 2 and 3 and predictor variables 4 and 5. I specifically was looking for how correlated these variables were with each other.
3. I found that predictor variables 2 and 3 have a correlation coefficient of 0.6414637002481952 and predictor variables 4 and 5 have a correlation coefficient of 0.9072222660959617.
4. Since predictor variables 2 and 3 only have a moderately high correlation, while their relationship should be noted, it is not a major concern regarding collinearity. However, predictor variables 4 and 5 have a very high correlation so it is potentially a major concern regarding collinearity.

Extra Credit 1

1. I plotted the distribution of all the predictor variables and the outcome variable using histograms.
2. I did this to compare how the distribution of all the variables looked to the shape of a normal distribution function.
3. Thie histogram of predictor variable 1 shows that most of the data points cluster around the middle of the range with the rest of the points falling on the extremes asymmetricaly (with a greater number of data points falling on the higher extreme). The histograms of variables 2-6 show that most of the data points cluster towards the lower extreme with the rest of the points tapering off as we increase along the range of values. The histogram of variable 7 shows that the majority of data points take on the discrete higher values of 3 and 4 with the remainder taking on values 0, 1, or 2 (which makes sense since this predictor variable only takes on these discrete values). The histogram of the outcome variable shows that most of the data points cluster around the lower end of the

range of values with the rest of the data points for the most part tapering off as we increase along the range of values.

4. Since a normal distribution is described as having most of the data points falling in the middle of the range of values (mean and median generally in the center) with the rest of the data points symmetrically tapering off to the lower and higher extremes (resulting in a bell-shaped symmetrical curve), none of of the variables can be reasonably described as a normal distribution since none of them fit that description.

Extra Credit 2

1. I plotted the distribution of the outcome variable using a histogram.
2. I plotted the distribution to be able to properly visualize and see if there were any special characteristics I could identify visually.
3. I found that the data was right skewed with an abnormally large number of data points at the maximum value of the outcome variable.
4. This “cap” at 500k for the outcome variable might limit the validity of the conclusions we drew to earlier questions in that the betas calculated may be off (since it’s likely that any blocks that had a median house value greater than 500k were classified at just 500k). Although we can see that there is a clear strong correlation between median income and median house value, because of the cap which affects the regression, we can’t assume that the exact values that come from our regression are correct.