Homework 1

Question 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?

Answer 1: The predictors 2 and 3 are the total number of rooms in a given block and number of bedrooms in a given block respectively.

1) It is a good idea to standardize or normalize predictor variables 2 and 3, I used the standardization to get the variable in a comparable form. This step helped in a fair analysis because it reduced the chances of disproportionately influence and led to a better interpretation of the coefficients which ultimately helped in better understanding the model. This process involved transforming the variables to have a mean of 0 and a standard deviation of, scaling them to a range [0, 1].

2) Standardization or normalization was chosen specifically for variables 2 and 3 because they represent numbers (total number of rooms and number of bedrooms). Standardizing and normalizing predictor variables 2 and 3 made me avoid biases in the model. There is also a possibility of them having a different range. If we consider the total number of rooms, it can vary for each block and we will ultimately have unstandardized data for analysis. As we saw in the part 1 of the answer, standardizing makes the coefficients of these variables directly comparable leading to more meaningful observation.

3) After executing the code, I analyzed the correlation matrix to evaluate the relationship between population and households. Based on the high correlation between these variables, I concluded that they might not be very useful by themselves to predict median house values in a block due. Additionally, I performed standardization on predictor variables 2 and 3 to ensure comparability among predictors and enhance the predictive power of the model. These steps provided the basis for the conclusions drawn regarding the standardization of certain predictors and the potential limitations of others in predicting median house values in a block. Predictor variables 4 and 5 are likely to be highly correlated with each other, as densely populated blocks tend to have more households. As a result, including both variables separately in a model could lead to redundancy and instability in the estimation of coefficients.

4) The findings suggest that standardizing or normalizing predictors like total number of rooms and number of bedrooms can improve the predictive power of the model by ensuring comparability among predictors. However, caution should be exercised when including highly correlated variables like population and number of households in the model, as they may introduce multicollinearity issues and reduce the interpretability of individual predictor effects.

Question 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)?

Answer 2: The predictors 2 and 3 are the total number of rooms in a given block and number of bedrooms in a given block respectively. The predictors 4 and 5 are population in the block and number of households in the block respectively.
1) To determine whether it's better to normalize predictor variables 2 and 3 by predictor 4 or 5, I started with calculating the ratios of total rooms and total bedrooms to both population and households, respectively. I used visualization to understand better, therefore I used histograms.

2) This approach was chosen as it allows for a direct comparison of the distributions resulting from normalizing by population and by number of households. Visual inspection of histograms gives us information about the spread and central tendency of the distributions, which is required to understand the normalization.

3) I visualized the distributions of these ratios using histograms to compare which normalization method might be more suitable. After running this code, we can visually inspect the histograms of the ratios to determine which normalization method results in distributions that are more suitable for use as predictor variables. Based on the histograms, we can observe that the distributions of "Rooms per Person" and "Bedrooms per Person" are more evenly spread and centered around a typical range compared to "Rooms per Household" and "Bedrooms per Household". This suggests that normalizing by population might be a better choice as it results in more uniform distributions, which could potentially improve the performance of predictive models. Therefore, it appears that normalizing predictor variables 2 and 3 by population would be a more suitable approach for meaningful use in predicting median house values in a block.

4) These findings suggest that normalizing predictor variables 2 and 3 by population is likely to be more appropriate for meaningful use in predicting median house values in a block. Normalizing by population ensures that the resulting variables are more standardized and comparable across different blocks, potentially leading to improved model performance and interpretability. Additionally, the more uniform spread of the distributions normalized by population suggests a better representation of the underlying relationships between the predictor variables and the target variable.

Question 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?]

Answer 3:
1) To identify the most and least predictive variables of housing value from a simple linear regression perspective, I started with liner regression analysis for all the predictors then inspected scatter plots of each predictor variable against the median house value. This step helped me visually to identify potential issues such as outliers or non-linear patterns.

2) I chose to as mentioned in the part 1 because it helped me for a comparison of the relation between each variable and the median house value. Additionally, conducting separate simple

linear regression analysis for each predictor variable enables us to quantify the predictive power of each variable.

3) From the scatter plots, we can see that the "median_income" variable exhibits the strongest positive linear relationship with median house value, which is indicated by the steeper slope of the scatter plot compared to other variables. Conversely, the "ocean_proximity" variable shows a not so noticeable linear relationship with median house value, as we can not see any patterns in the scatter plot.

4) These findings suggest that "median_income" is the most predictive variable of housing value, while "ocean_proximity" is the least predictive. The strong relationship between median income and housing value indicates that areas with higher median incomes tend to have higher median house values, which is consistent with economic principles. On the other hand, the lack of relationship between ocean proximity and housing value suggests that this variable may not be a significant determinant of housing prices in the dataset.

Question 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.?

Answer 4:
1) I started with a multiple linear regression model using all predictor variables. I used the standardized or normalized predictor variables obtained previously and the median house value as the target variable. Additionally, I made a separate simple linear regression model using only the single best predictor identified earlier.

2) I used this method to compare the predictive performance of the multiple regression model using all predictors against the simple regression model using only the single best predictor. By evaluating the R-squared values of both models, we can compare how well they predict housing value. This comparison allows us to determine whether including additional predictors in the multiple regression model improves its predictive power.

3) The R-squared value for the multiple regression model using all predictors together was found to be 0.5974150162494976 in comparison, the R-squared value for the simple regression model with the single best predictor was 0.45885918903846656.

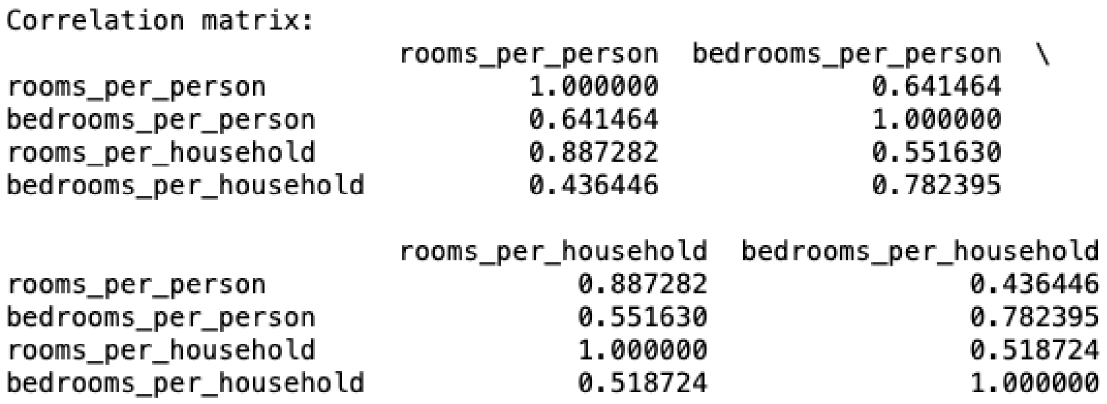
4) These findings suggest that the multiple regression model using all predictors performs better in predicting housing value compared to the simple regression model with the single best predictor. The higher R-squared value for the multiple regression model indicates that including additional predictors improves the model's ability to explain the variability in housing value. This underscores the importance of considering multiple factors simultaneously when predicting housing values, as opposed to relying on a single predictor.

Question 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?

Answer 5:
1) To assess potential concerns regarding collinearity between variables 2 and 3, as well as between variables 4 and 5, I calculated the correlation coefficients between these pairs of variables. This involved using the standardized or normalized predictor variables obtained earlier and computing the correlation matrix.

2) This approach was chosen as it allows for a quantitative assessment of the relationships between pairs of variables. Collinearity can lead to inflated standard errors and unstable coefficients in regression models, making it important to identify and address potential collinearity issues. Computing the correlation coefficients provides a straightforward way to evaluate the strength and direction of the relationships between variables, helping to identify potential collinearity concerns.

3) The correlation matrix revealed that variables 2 and 3 are moderately correlated, similarly, variables 4 and 5 also exhibit moderate correlation.



```
Correlation matrix:
                        rooms_per_person  bedrooms_per_person  \
rooms_per_person                1.000000             0.641464
bedrooms_per_person             0.641464             1.000000
rooms_per_household             0.887282             0.551630
bedrooms_per_household          0.436446             0.782395

                        rooms_per_household  bedrooms_per_household
rooms_per_person                   0.887282                0.436446
bedrooms_per_person                0.551630                0.782395
rooms_per_household                1.000000                0.518724
bedrooms_per_household             0.518724                1.000000
```

4) These findings indicate that there is indeed potential concern regarding collinearity between variables 2 and 3, as well as between variables 4 and 5. Moderate correlation between these pairs of variables suggests that they may capture similar information or have interrelated effects on the outcome variable. Including both highly correlated variables in a regression model can lead to multicollinearity issues, which may result in unstable coefficient estimates and reduced interpretability of the model.

EXTRA CREDIT:

Question 1: Does any of the variables (predictor or outcome) follow a distribution that can reasonably be described as a normal distribution?

Answer 1:

1) We will check for the normal distribution. I created histograms for each variable to visually have a look at their shapes and calculated skewness and kurtosis values as measures of asymmetry and peakedness of the distributions.

2) I employed a specific method to evaluate distribution shapes and calculate skewness and kurtosis values, which are commonly used to assess normality within a dataset. Histograms visually depicted data distribution, while skewness and kurtosis provided numerical insights. This approach allowed for an objective examination of whether any variables exhibited characteristics of a normal distribution.

3) The analysis indicated that none of the variables in the dataset followed a normal distribution. Histograms showcased varied shapes, with certain variables showing skewness and kurtosis values significantly different from 0. For example, skewness and kurtosis values for predictor variables ranged from -0.730735 to 4.147042 and from -0.290249 to 32.622732, respectively, indicating non-normal distributions. Similarly, the outcome variable, median house value, displayed non-normal characteristics.

4) These observations, drawn from histograms and numerical values, suggest that the dataset's variables do not meet the normality assumptions required for certain statistical analyses such as linear regression. Non-normality could compromise the validity and interpretability of statistical results when analyzing the data. Therefore, it would be advisable to conduct further investigations into the distributions of the variables to ensure robust and accurate findings.

Question 2: 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.

Answer 2:
I analyzed the distribution of the outcome variable, median house value, by generating a histogram and computing skewness and kurtosis values using Python. This method was selected because it's a standard practice for evaluating distribution normality and shape. Histograms offer a visual depiction of data distribution, while skewness and kurtosis values provide numerical insights into asymmetry and peakedness, respectively. By employing these techniques, we can objectively assess whether the outcome variable's distribution exhibits characteristics that may impact the validity of conclusions drawn from previous analyses.

Upon examining the distribution of the outcome variable, it became evident that it does not conform to a normal distribution. The histogram displayed a skewed shape, with the majority of values clustered towards lower median house values and a lengthy tail extending towards higher values. Additionally, both skewness and kurtosis values deviated from 0, indicating non-normal characteristics of the distribution. Specifically, the skewness value was 0.9776922140978416, while the kurtosis value was 0.3275001388119616, suggesting heavier tails than expected in a normal distribution.

These findings suggest that the distribution of the outcome variable, median house value, may pose limitations to the validity of conclusions drawn from analyses relying on normality assumptions. The observed skewness and heavy-tailed distribution could potentially violate the assumptions of linear regression models, leading to biased parameter estimates and inaccurate inferences. Therefore, it's crucial to explore alternative approaches or transformations to address the non-normality of the outcome variable and ensure the robustness and validity of conclusions.