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?

First, to visualize the data, I drew scatterplots for the row data of predictors 2 and 3 and the standardized data of them. To better understand the relationship between them, I calculated the correlation coefficient between them and the median value, showing that the standardized predictors correlate closer to the median house value.

Then, I did a linear regression to examine this standardization, showing that this model doesn't explain much of the variation. In this way, the standardized data performed better in predicting the house value with higher R^2.

Since row total rooms and total bedrooms were misleading as block sizes vary, Dividing by household count gives a better measure of house size. That's why it is better to standardize/normalize the predictors 2 and 3. However, $R^2 = 0.0237$ ($\approx 2.37\%$) → This is very low, meaning the model explains only about 2.37% of the variability in house values.

This suggests that rooms per household and bedrooms per household alone are not strong predictors of house prices. Other factors (like income, location, etc.) likely also play an important role. Still, comparatively, standardized data is better based on the R^2 value we calculated.

Since predictors 4 and 5 don't tell us about individual house characteristics. Two blocks with the same population might have very different house sizes and prices. Therefore, they are not very useful by themselves as predictors.

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)?

For each standardized variable, I fit a separate linear regression model predicting median house value. Then, I computed the R^2 score, which tells us how well each predictor explains housing values.

Normalization is necessary because the absolute number of rooms or bedrooms alone does not provide meaningful information. Specifically, rooms per person had the highest R^2 score, 0.0439, meaning it was the best predictor among the four. On the other hand, bedrooms per household, 0.0034, had the lowest R^2 score, meaning it performs badly in predicting the median house value. Therefore, Normalizing by population (rooms_per_person, bedrooms_per_person) consistently produced higher R^2 scores than normalizing by households.

To further explain, rooms per person was the best predictor, suggesting that the availability of space per individual matters more for house value than the number of rooms per household. This makes sense because housing demand is often driven by individual living space rather than total household size.

Thus, for a more accurate prediction of housing values, rooms and bedrooms should be normalized by population rather than the number of households.

3. Which of the seven variables is most and least predictive of housing value, from a simple linear regression perspective?

First, we computed correlations between each predictor and housing value (median_value), sorting them in descending order to see which variables have the strongest relationships. Next, we calculated R^2 values for each predictor separately by fitting simple linear regression models and measuring how much variance in housing prices each variable explains individually.

The R^2 scores were stored and sorted to identify the most and least predictive variables. A set of scatterplots was also generated to visually inspect the relationships between housing value and key predictors, such as income, rooms per person, bedrooms per household, proximity to the ocean, and house age. This visualization helps confirm whether the relationships appear linear or if additional patterns exist.

From the findings, median income is likely the most predictive variable, as it tends to have the highest R^2 and correlation with housing prices. This makes sense, as higher-income areas typically have higher home values. Conversely, population is the least predictive with the least R^2, suggesting that while it may influence desirability, it does not predict house prices as strongly as income or room counts.

However, from the graph, there's a **"flattening" effect** at the top (where house values seem to stop increasing beyond a certain income), which suggests data truncation (e.g., house values might be capped at $500,000). If house prices were not capped, income might be even more predictive than shown.

These results directly answer the original question by identifying median income best predicts housing value and population contributes the least explanatory power. However, while simple linear regression provides a useful baseline, multiple regression would be needed to account for interactions between predictors and improve overall predictive accuracy.

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.?

For this question, I performed a multiple regression model and compared it with the single linear regression model I've done in the previous question. Moreover, I used stepwise multiple regression approach which gradually adds one predictor at a time to a regression model predicting housing value, computing and visualizing the resulting R^2 scores at each step.

The full multiple regression model might improve R^2 slightly from 0.473 to 0.602, but not dramatically, meaning that multiple regression predicts the variation of this model better.

Compared to the highest R^2 value for single regression (median income), regression performs better. Therefore, other predictors, like proximity to the ocean and house age, still play a role in predicting the median house value.

By examining how the R^2 changes, we see the incremental predictive power each new variable contributes. The left scatter plots compare actual vs. predicted values to visualize model performance, and the right histograms show the distribution of each newly added predictor. Together, these plots help us evaluate both the importance and the distribution of each variable. Ultimately, the final R^2 (with all predictors) reveals how well they collectively explain housing values.

Based on the stepwise regression approach, we observe that the predictive power (R^2) increases as more variables are added, indicating that these predictors together explain a significant portion of housing price variance. However, if the final R^2 is not very close to 1, it suggests that while these features are useful, additional external factors also influence housing value.

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?

I made two scatter plots and calculated correlation coefficients to check for collinearity between (rooms per person, bedrooms per person) and (population, households). Scatter plots help us to visualize relationships, while correlation values quantify them.

This was done to detect high correlation, which can cause multicollinearity in regression models. When predictors are highly correlated, they provide redundant information, leading to unstable coefficients and unreliable predictions.

There isn't high collinearity between standardized 2 and 3 (with $r = 0.641$). However, there is a high correlation, though slightly less extreme, concern for variables 4 and 5, since their high correlation ($r = 0.907$) could lead to multicollinearity issues in my regression model, indicating they are kind of redundant. The 0.907 correlation between population and households is also high, suggesting strong dependence.

These findings confirm collinearity concerns for the latter pair, so dropping one would be a better approach.

Extra credit

1) By visualizing the distribution of all the predictors, none of the variables in this dataset perfectly follow a normal distribution. In particular, total rooms, total bedrooms, population, households, median income, and median house value all show signs of right skewness. Median age might be the closest to normal if its histogram is relatively symmetric, but overall, the predictors and the outcome do not conform to a normal distribution.

2) A histogram and a density plot were generated to examine the distribution of the median house value. These visualizations help assess whether the data follows a normal distribution or if there are any skewness or outliers.

Understanding the distribution of the outcome variable is crucial because regression models assume that residuals are normally distributed. If the outcome variable is heavily skewed, it could affect the model's predictions and the validity of conclusions drawn from the analysis.

The distribution of median house value is left-skewed, meaning there are more high-value houses, and fewer low-value ones. This suggests that the data is not symmetrically distributed, which could impact linear regression assumptions.

The left skewness indicates that standard regression models may not perform optimally without transformations. This could affect the accuracy of predictions and inference, potentially leading to biased conclusions. It may be useful to apply a transformation (e.g., log transformation) to make the distribution more symmetric.