Problem 1

1. Distribution of AveBedrms: This histogram represents the average number of bedrooms per home. It skews sharply to the right, with most of the data concentrated at the low end, indicating that the average number of bedrooms in most homes is relatively small.
2. Population distribution: The population histogram is also right-skewed, showing that most regions (or groups) have small populations and decline sharply as population size increases.
3. Distribution of AveOccup: This histogram represents the average occupancy rate, i.e. the number of people per dwelling. The distribution is extremely skewed to the right, with the vast majority of data points indicating low average occupancy.
4. MedHouseVal's distribution: The data in the graph is concentrated in a range of 1 to 3, meaning that most properties are valued close to this median. As a result, we can see that most properties in the area range in value between 1 and 3.
5. Distribution of MedInc: This histogram represents the median income. The distribution is somewhat skewed to the right, suggesting that most data points cluster at the lower end of the income range, with fewer extending towards higher income values.
6. Distribution of Houseages: This histogram shows the ages of houses in the data set. The distribution seems fairly even, but there are a large number of homes in the lower age range (near 0) and two larger spikes near the 10-year and 50-year marks.
7. Distribution of AveRooms: This histogram shows the average number of rooms per house. The histogram is heavily skewed to the right. So we can know that most houses have a small number of rooms.

Problem 2:

The model's moderate R-squared value indicates that while median income is a significant predictor of median house value, it does not capture all the variability. Other factors likely influence house values, and incorporating additional features might improve the model's performance. The MSE values suggest the model's predictions are reasonably close to the

actual values, but there is room for improvement, especially for houses with higher median incomes where the variance is more pronounced.

Problem 3:

MAE (Mean Absolute Error): 0.5317.

R-squared: 0.6076. An R-squared of 0.6076 means that approximately 60.76% of the variance in the housing price can be predicted from the features used in the model.

Based on the MAE and R-squared values, we can infer that the Multiple Linear Regression model has a reasonable predictive power, with over half of the variance in the target variable being explained by the model. For instance, in some real estate markets, a small prediction error could still translate to a significant difference in price, so the context of the application is crucial to fully evaluate these results.

Problem 4:

- The predicted value using the Locally Weighted Linear Regression (LWLR) model with a bandwidth of 10 is approximately 4.035. Interestingly, this prediction does not change when increasing the bandwidth to 20, suggesting that the weights at this range are relatively stable or the additional points do not significantly influence the regression for this particular prediction point.
- When the bandwidth is set to a very narrow width of 1, the predicted value becomes negative and large in magnitude (-13.700). This result indicates that a very small bandwidth causes the model to be overly influenced by nearby data points, which may not be representative of the broader trend. This result could be due to overfitting to the very local noise in the data.
- As the bandwidth is increased to 5, the model's prediction appears more reasonable (4.373), and then as mentioned earlier, it stabilizes around 4.035 for bandwidths of 10 and 20.
- Further increasing the bandwidth to 30 yields a predicted value of approximately 3.942, which is slightly lower than the stable values obtained with a bandwidth of 10 and 20. This could indicate that as more distant points are included (by increasing bandwidth), their influence begins to alter the prediction, possibly due to capturing more global trends or noise.
- The Mean Squared Error (MSE) for the model, when evaluated over a set of prediction points, is 0.185. This metric is relatively low, suggesting that the

model's predictions are, on average, close to the actual values for the given dataset.

Adjustment of wide parameters in LWLR has a significant impact on predictions. Bandwidth that is too small may lead to overfitting, as seen with a bandwidth of 1. A more modest bandwidth of 10 seems to provide stable and reasonable predictions, while very large bandwidths start to include potentially less relevant data points, changing the predictions slightly. MSE provides a quantitative measure of model performance, with lower values indicating a better fit to the data.

Problem 5

In this project, we discussed three regression models: Simple Linear Regression, Multiple Linear Regression, and Locally Weighted Linear Regression (LWLR). Then, we use Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared to evaluate the accuracy of these three models respectively.

Model Performance

- Training MSE: 0.709
- Testing MSE: 0.667
- R-squared: 0.486

Multiple Linear Regression:

- MAE: 0.533
- R-squared: 0.599

Locally Weighted Linear Regression (LWLR):

- MSE for a subset of points: 0.185
- R-squared: Variable, dependent on the bandwidth chosen

Recommendations:

For general purposes, Multiple Linear Regression is the best model due to its balanced approach. Its accuracy is better than simple linear regression, and its computational overhead is also smaller than LWLR.

.