Springboard Project 11.1 (Linear Regression)

Using Boston Housing data

The first three-quarters of this notebook consists mostly of guided exercises from the Harvard CS109 course, which provided the notebook. I create histograms and scatter plots as preliminary steps to linear regression and, in a few cases, enhance plots with steps like adjusting bins in histograms.

Next, I follow guided steps to perform a linear regression. These steps include using the scikit-learn package, fitting a model, and capturing metrics like mean squared error (MSE) to assess performance. I also obtain pieces of the linear model including the intercept and coefficients. Finally, through graphs of predicted values from the model, I assess results.

In the last fourth of the notebook, where the Relationship between PTRATIO and Housing Price heading appears, I begin more rigorous independent work in the Your Turn section. I start by generating a scatter matrix to look for linear relationships between the three independent variables (crime rate, rooms per dwelling, and pupil-teacher ratio) and the dependent variable (price). All predictors show linear relationships to varying degrees, but rooms per dwelling is of particular note because it has a clear, strong positive relationship with price.

The next step is fitting a model and capturing the intercept and coefficients. With that done, I generate a plot of the predicted versus actual prices to gauge the model's performance. It looks good overall but does better with higher prices. I finish by calculating the residual sum of squares and MSE.

After the regression exercise, I follow some guided steps to using training and test data, which help ensure that a model is generalizable. In the independent work at the end, I compare MSE values to determine that the model generalizes fairly well. An overlaid scatter plot of residuals vs. fitted values also indicates a generalizable model, but it is worth noting that outliers occur and that the residual patterns are not uniformly random. (A vague cone shape emerges). With that in mind, further model assessment would be useful in a more in-depth project.