W:

In this lab, I will use linear regression to analyze the dataset of housing prices. I think a lot of people will be interested in housing prices, such as house buyers, real estate agents, and investors. It is very useful to manage one's financial situation if the potential buyer can effectively predict the housing price. In addition, real estate agents or investors can understand the trend of housing prices in a certain location by prediction. They can more easily become proficient in the real estate market to reduce risks and increase their assets. This dataset is interesting and useful because it includes almost all the factors that might affect house prices except for the region. Therefore, I can use some algorithms to understand the impact of these different factors and create a linear regression model to predict housing prices.

First, I need to have an understanding of the dataset and make some improvements to best fit the linear regression model. I found that there were three different statuses of furnishing in this dataset, so I used data visualization to look at the differences between the status and the number of them. I also created a scatter plot to understand the relationship between house sizes and prices. From most of the data points, increasing house size leads to increasing house prices. After a preliminary understanding of the dataset, I converted all the "yes" and "no" data in the dataset into 1 and 0 for analyzing purposes.

After splitting the data into test and train data, then fit them into the linear regression model. There was a lot of information I can find from statistical results. I can see the difference between the original price and the predicted price. I do not think the difference between the actual price and the predicted price is far apart after comparison, so the linear regression model for housing price prediction should be effective. In this linear regression model, when all the prediction variables are zero, the average value of the response variables, namely intercept, is 4,752,269. And then, I created a coefficient dataframe, we can see how different factors can cause to increase in the housing price. For example, each additional unit in the area will lead to an increase in the predicted price of $495,476. However, it can be seen from the data on different furnishing statuses that it has little or no influence on the housing price.

As a result of the linear regression metrics, the R squared, Mean Squared Error (MSE), results of Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) evaluate the validity of the model. The result of R Squared is 0.657, which is close to 66%. From this result, I think the model should perform well because 66% of the data matches the model. The MSE tells me the error between the predicted price and the actual price, which represents the distance of the regression line from a set of data points. In terms of the results, this error value appears to be relatively high which also leads to the high RMSE. In addition, MAE also represents the error value between the prediction price and the actual price. Although these metrics appear a large number, this does not mean that the linear regression model is not feasible. Therefore, I put the dataset of housing prices into the logistic regression for analysis to compare it with the linear regression model. The results show that MSE, RMSE, and MAE in the logistic regression model are lower than those in the linear regression model. The comparison results show that the linear regression model is more suitable for housing price prediction.