Huong Phan

Assignment 2 - Linear Regression

- First, I use a heat map to determine the strength of correlation of house price against different continuous variables, and also correlation between any continuous variables. I chose the following variables: 'price', 'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'grade', 'sqft_above', 'lat', 'long', 'sqft_living15', 'sqft_lot15', ‘sqft_basement’. The heat map shows that price against ‘sqft_living’ has the strongest correlation (r = 0.7) out of all correlations of price against another continuous variables, while ‘long’ has the weakest (r = 0.022) correlation out of all correlations of price against another continuous variables.
- After making some preliminary scatterplots of price against sqft_living, I realize that there are outliers that affect the linear regression model, so I remove any house price outliers with z-score of above 5 or below -5.
- Then, I began to make scatterplots of price against sqft_living, and also make scatterplots for different house categories, for example, a scatterplot of price against sqft_living for houses that have been renovated and a scatterplot of price against sqft_living for houses that have not been renovated. I found that prices of houses with the same sqft_living are generally higher when they have been renovated, and fit linear regression to each scatterplot that would predict the price of houses by sqft_living, either renovated or not.
- I made another scatterplot for price against grade, but this time I fit polynomial, ridge and lasso regression into the scatterplot.
- Results can be seen when the code is run.