For this assignment, we used the cleaned dataset that we created in Project A.

Linear Regression

The correlation coefficients between the training data features (that contained numerical values) and the target column, Sale Price, was calculated. If the absolute value of the coefficient was greater or equal to 0.5, we would mark this as a feature to be used in the linear regression model.

The linear regression model was then created using the default implementation provided by sklearn and our model had a root-mean-square-error (RMSE) that was approximately 50% lower than the baseline model's RMSE.

KNN Model

First, we created the target column, called After1970, which contains either 0 or 1 to mark whether it was false or true that a house was built after 1970. Similar to what we did for the Linear Regression model, we calculated the correlation coefficients between the training data features (that contained numerical values) and the target column (After1970.)

If the absolute value of the coefficient was greater or equal to 0.45, we would mark this as a feature to be used in the KNN model. These features were then normalized to between 0 and 1, since KNN involves the calculation of Euclidean distances and non-normalized values would result in distances heavily skewed towards the features with larger values.

The model used 5 neighbours in the computation and performs 65% better than the baseline model on average.