Homework 1- Jenell Lewis

(1) The size of the training set is 1168 and the size of the test set is 292.

(2) The top ten features by score are: 'OverallQual', 'ExterQual\_Coded', 'GrLivArea', 'KitchenQual\_Coded','TotalBsmtSF','1stFlrSF', 'GarageCars', 'GarageArea', 'BsmtQual\_Coded', 'GarageFinish\_Coded'

Below is a list of the top ten features, their scores and rmse values:

1. OverallQual (0.6454631197278843, 49018.43578812499)
2. ExterQual\_Coded (0.4991753471267154, 58260.14865535067)
3. GrLivArea (0.444289666781406, 61369.55515927838)
4. KitchenQual\_Coded (0.43970293723342646, 61622.30113827634)
5. TotalBsmtSF (0.42491266111423953, 62430.33167581401)
6. 1stFlrSF (0.4166679048480356, 62876.2562386846)
7. GarageCars (0.41193116659665807, 63131.022100901646)
8. GarageArea (0.4073263621599066, 63377.71018105261)
9. BsmtQual\_Coded (0.3409700148262619, 66831.5181312807)
10. GarageFinish\_Coded (0.3162849544224505, 68071.65468394081)

(3) The top ten paired features, their scores and rmse values are listed below:

('OverallQual', '1stFlrSF') (0.7173882583039417, 43764.68831784155)

('OverallQual', 'TotalBsmtSF') (0.706791108152437, 44577.66409611638)

('OverallQual', 'GrLivArea') (0.692198677904023, 45673.467009316504)

('OverallQual', 'GarageArea') (0.6813946935673894, 46468.13490082075)

('OverallQual', 'KitchenQual\_Coded') (0.6730869536074611, 47070.07230442364)

('OverallQual', 'GarageCars') (0.6723247870365776, 47124.910051666164)

('OverallQual', 'ExterQual\_Coded') (0.6680147343521529, 47433.82463056837)

('OverallQual', 'BsmtQual\_Coded') (0.6566660974647345, 48237.75463112187)

('OverallQual', 'GarageFinish\_Coded') (0.6519020465859777, 48571.27166300008)

('ExterQual\_Coded', '1stFlrSF') (0.6364054863411375, 49640.64317074208)

(4) Using all features, the score is 0.8127547098527903 and the rmse value is 35623.30345000739. From my results, it is seen that using all the features was a better predictor because it has a higher score and lower rmse value than the top 10 pairs.

(5) Using the 5NN regressor, the score is 0.6015421069563245 and the rmse value Is 51966.07983368139.

Using the 10NN regressor, the score is 0.5949812006956909 and the rmse value is 52392.16317256475.

When comparing KNN to linear regression, linear regression is better at inference due to the higher score and lower rmse value.