## Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

A) The optimal value of alpha for ridge is 10 and lasso is 0.0002.

If we double the value we use 20 and 0.0004 we get

| | Metric | Linear Regression | Ridge Regression at alpha=10 | Ridge Regression at alpha=20 | Lasso Regression at alpha=0.0002 | Lasso Regression at alpha=0.0004 |
|---|---|---|---|---|---|---|
| 0 | R2 Score (Train) | 0.958938 | 0.930574 | 0.921423 | 0.950662 | 0.941862 |
| 1 | R2 Score (Test) | 0.724097 | 0.857320 | 0.849644 | 0.702000 | 0.737655 |
| 2 | RSS (Train) | 6.849184 | 11.580304 | 13.106808 | 8.229623 | 9.697401 |
| 3 | RSS (Test) | 18.188384 | 9.405945 | 9.911920 | 19.645095 | 17.294626 |
| 4 | MSE (Train) | 0.081904 | 0.106499 | 0.113301 | 0.089779 | 0.097457 |
| 5 | MSE (Test) | 0.203779 | 0.146543 | 0.150433 | 0.211782 | 0.198709 |

The R2 Score and RSS values show that the Linear Regression model has the highest R2 Score for training data, while the Ridge Regression model with alpha=10 has the highest R2 Score for testing data. The Lasso Regression model with alpha=0.0002 has the lowest MSE value for training data, while the Ridge Regression model with alpha=20 has the lowest MSE value for testing data.

**If we choose to double the value of alpha for both ridge and lasso regression, it would lead to stronger regularization and the model will become more constrained. As a result, the coefficients of the predictor variables will shrink further towards zero, leading to a simpler model with fewer predictor variables.**

**Question 2**

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

A) Based on the given summary, if we prioritize R2 score as the evaluation metric, then Ridge regression seems to perform better than Lasso regression as it has a higher R2 score on both train and test sets. However, if we prioritize minimizing the mean squared error (MSE), Lasso regression seems to be a better choice as it has a lower MSE on both train and test sets.

Overall, the choice between Ridge and Lasso regression will depend on the specific goals and priorities of the analysis, and further investigation and experimentation may be required to determine the most appropriate method.

**Question 3**

After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

The top five parameters are found using:

```
sorted_coef_indices = np.argsort(-np.abs(lasso.coef_))
```

```
sorted_coef = lasso.coef_[sorted_coef_indices]
predictor_names_sorted = X.columns[sorted_coef_indices]
```

```
betas_1 = pd.DataFrame(index=predictor_names_sorted)
betas_1.rows = predictor_names_sorted
betas_1['Lasso'] = sorted_coef
pd.set_option('display.max_rows', None)
betas_1
```

| | Lasso |
|---|---|
| PoolQC_Gd | -1.735128 |
| Condition2_PosN | -0.475938 |
| OverallCond_3 | -0.180774 |
| OverallQual_9 | 0.179395 |
| Neighborhood_MeadowV | -0.148699 |

Now deleting these five predictor variables and creating new model we get best alpha as 0.0006. Please check attached jupyter notebook end for the same.

The top five parameters now are:

| | Lasso |
|---|---|
| OverallQual_2 | -0.154800 |
| Fireplaces_3 | -0.136769 |
| Neighborhood_NridgHt | 0.124278 |
| Neighborhood_Crawfor | 0.113207 |
| MSSubClass_160 | -0.112823 |

**Question 4**

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A) To ensure that a model is robust and generalizable, we need to check its performance on both the training and testing datasets. In the given summary, we can see that the linear regression model has a higher R2 score on the training set than both the Ridge and Lasso regression models, indicating that it fits the training data better. However, the linear regression model has a significantly lower R2 score on the test set, indicating that it is overfitting to the training data and may not generalize well to new data.

On the other hand, the Ridge and Lasso regression models have lower R2 scores on the training set, but higher R2 scores on the test set, indicating that they are more generalizable and robust. Additionally, both models have lower RSS and MSE values on the test set, further supporting their superior generalizability compared to the linear regression model.

In summary, a robust and generalizable model should perform well on both the training and testing datasets, indicating that it is not overfitting to the training data and can generalize well to new data. The implications of this are that the model is more likely to be accurate when making predictions on new, unseen data, which is the ultimate goal of any predictive model.