Question1:

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?

Answer:

The optimal alpha values for Ridge and Lasso are:

1. Ridge: 10
2. Lasso: 316.23

Updated alpha values are:

1. Ridge: 20
2. Lasso: 632.46

After we choose to double the alpha value for both Lasso and Ridge, R2 score for both test and train is decreasing.

The most important predictor variables after the changes are implemented are:

```
Ridge Regression with alpha=20
Most important predictors after doubling the alpha value:
                 Variable   Coefficient
228        TotRmsAbvGrd_10  27833.546671
120          OverallQual_9  27166.736629
22              GarageCars  27038.462784
12                2ndFlrSF  26323.607518
121         OverallQual_10  25357.733328
77    Neighborhood_NoRidge  25320.548754
14               GrLivArea  25225.083723
21              Fireplaces  23708.329805
17                FullBath  23653.409088
188            BsmtQual_TA -20433.307276
84    Neighborhood_StoneBr  20383.507466
23              GarageArea  19612.633758
220         KitchenQual_TA -18703.732828
187            BsmtQual_Gd -18316.904381
11                1stFlrSF  17678.074732
```

```
Lasso Regression with alpha=632.46
Most important predictors after doubling the alpha value:
                  Variable    Coefficient
14               GrLivArea  154678.900410
120          OverallQual_9   59660.912457
22              GarageCars   55284.421878
121         OverallQual_10   45070.403672
21              Fireplaces   35916.398528
119          OverallQual_8   27838.559909
77    Neighborhood_NoRidge   21564.690471
228        TotRmsAbvGrd_10   20160.634532
192        BsmtExposure_Gd   15835.519298
4             YearRemodAdd   15795.785626
188            BsmtQual_TA  -14814.591854
267  SaleCondition_Partial   13971.210717
78    Neighborhood_NridgHt   13097.051272
187            BsmtQual_Gd  -11388.291038
176           ExterQual_TA  -11329.796072
```

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?

Answer:

We selected the Lasso Regression model as it is the most suitable for this dataset. Its robust performance on the training set indicates a good fit, and it also outperforms the other models on the test set, demonstrating strong generalisation to new data. Additionally, Lasso's ability to perform feature selection by setting some coefficients to zero enhances interpretability, making it particularly beneficial for our needs.

However, choosing between Ridge and Lasso depends on specific needs. If the goal is to perform feature selection and simplify the model by reducing the number of features, Lasso is preferred since it can set some coefficients to zero, thereby excluding those features from the model.

Conversely, if all features are considered important and should be retained in the model, Ridge Regression may be the better choice. Ridge is effective when there are many parameters with comparable values, meaning most predictors influence the response. It generally offers better predictive accuracy than Linear Regression when multicollinearity exists among the predictor variables. Additionally, Ridge Regression is more stable and less likely to overfit compared to Lasso when predictors are highly correlated.

Question 3:

After building the model, you realised 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?

Answer:

We excluded the five most important predictor variables, which are:

1. GrLivArea
2. OverallQual_10
3. OverallQual_9
4. GarageCars
5. Fireplaces

After creating another model excluding the five most important predictor variables,

The five most important predictor variables for new model are:

```
1stFlrSF                 125924.638196
2ndFlrSF                  83108.233741
GarageArea                50829.213281
Neighborhood_NoRidge      40560.554650
Neighborhood_StoneBr      36733.618876
dtype: float64
```

Question 4:

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

Answer:

Ensuring a model's robustness and ability to generalize typically involves a combination of the following strategies:

1. Cross-validation: This method involves dividing the data into multiple subsets and then training and testing the model on various combinations of these subsets. It helps verify that the model performs consistently across different data sections and isn't overly tailored to specific data characteristics.

2. Regularization: Techniques such as Ridge and Lasso are employed to mitigate overfitting by adding a penalty term to the loss function. This approach discourages the development of overly complex models by effectively limiting the number of features used. Selecting the right regularization parameter (alpha) is crucial.

3. Feature Selection: This strategy includes using statistical tests to identify the most informative features or applying methods like Lasso that incorporate feature selection during the training process.

4. Performance Evaluation on a Separate Test Set: Assessing the model's performance on a distinct test set helps confirm its ability to generalize to new data. The model's test set performance is a reliable indicator of how well it will handle unseen data.

The implications for the model's accuracy are as follows:

Overfitting: A model that is overfit may show high accuracy on the training data but will likely underperform on unseen data due to being overly tailored to the training set and lacking generalization.

Generalization: A well-generalized model will exhibit similar performance on both training and unseen data.

Based on the performance table in the output:

Performance Table

	Regression	Dataset	RSS	R2	Adj. R2	MSE	NRMSE
0	Linear	Train	2.009376e+12	0.673322	0.670413	1.968047e+09	-44362.677856
1	Linear	Test	8.879912e+11	0.709465	0.703356	2.027377e+09	-45026.405696
2	Ridge	Train	1.568644e+12	0.744974	0.742704	1.536380e+09	-39196.686645
3	Ridge	Test	7.345883e+11	0.759656	0.754602	1.677142e+09	-40952.928115
4	Lasso	Train	1.382797e+12	0.775189	0.773188	1.354355e+09	-36801.567592
5	Lasso	Test	5.997267e+11	0.803780	0.799654	1.369239e+09	-37003.230857

Ridge and Lasso Performance: Both Ridge and Lasso models demonstrate superior performance (higher R2 and lower RMSE) on both training and test datasets compared to the Linear Regression model.

Consistency: The Ridge and Lasso models show comparable performance across training and test datasets (similar R2 and RMSE values), indicating good generalization.

Lasso's Superiority: The Lasso model slightly outperforms the Ridge model on both training and test data, making it the preferred choice based on the given results.

The optimal alpha values for Ridge and Lasso are 10.0 and 316.23, respectively, indicating these parameters provide the best balance between bias and variance for each model.

In summary, the Lasso model is the most robust and generalizable based on the output, with high R2 values and low RMSE on both datasets. However, continuous monitoring and validation are necessary to ensure consistent performance over time and with new data.