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?

Answer1: The optimal value of alpha for
Ridge : 8
Lasso : 20

Changes in the model if you choose double the value of alpha for both ridge and lasso?

	RIDGE		LASSO	
	alpha = 8	alpha =16	alpha = 20	alpha =40
R2 score(train)	0.89	0.87	0.94	0.91
R2 score(test)	0.86	0.85	0.84	0.85
	R2 score on training and test data has decreased		R2 score of training data has decrease and increased on testing data	

What will be the most important predictor variables after the change is implemented?

•[79]:
```python
## View the top 10 coefficients of Ridge in descending order
betas['Ridge2'].sort_values(ascending=False)[:10]
```

[79]:
```
Neighborhood_NoRidge    35225.208601
OverallQual_10          34251.021203
GrLivArea               33806.716980
2ndFlrSF                33206.993721
OverallQual_9           28073.401057
FullBath                28043.867049
TotRmsAbvGrd            26060.038636
GarageCars              25822.850786
1stFlrSF                25116.727664
Fireplaces              22881.603542
Name: Ridge2, dtype: float64
```

[80]:
```python
## View the top 10 coefficients of Lasso in descending order
betas['Lasso40'].sort_values(ascending=False)[:10]
```

[80]:
```
GrLivArea               243390.729792
RoofMatl_WdShngl        129605.759375
OverallQual_10           96397.321428
RoofMatl_CompShg         63708.769520
OverallQual_9            63658.822700
LotArea                  46818.180672
RoofMatl_WdShake         41485.110754
2ndFlrSF                 39472.332179
Neighborhood_NoRidge     38919.035846
GarageCars               35924.595265
Name: Lasso40, dtype: float64
```

Analysis : Most of the predictor variables remain same ; only the coefficients has changed

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?

Answer2: According to the below table

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.560765e-01	8.923725e-01	9.434261e-01
1	R2 Score (Test)	-4.931206e+20	8.677785e-01	8.491067e-01
2	RSS (Train)	2.802628e+11	6.867386e+11	3.609807e+11
3	RSS (Test)	1.389965e+33	3.726943e+11	4.253249e+11
4	MSE (Train)	1.656799e+04	2.593480e+04	1.880309e+04
5	MSE (Test)	1.781414e+15	2.917019e+04	3.116186e+04

Lasso regression model demonstrates good performance on both the training and test sets, with reasonable R2 Score. Additionally, Lasso regression introduces sparsity by automatically selecting a subset of the most important features, which can be beneficial in real-world scenarios with a large number of features. Therefore, I would choose to apply the Lasso regression model over the Ridge regression model due to its better performance and ability to handle overfitting while also providing feature

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?

Answer3: Top 5 previous predictors
RoofMatl_WdShngl
GrLivArea
OverallQual_10
LotArea
2ndFlrSF

After removing above predictors from the data set. The new 5 predictors after running lasso model are as below

1stFlrSF
Neighborhood_NoRidge
FullBath
TotRmsAbvGrd
GarageCars
Neighborhood_Crawfor

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?

Answer4: Ensuring a model's robustness and generalizability involves several important considerations:

- **Data Quality and Diversity**: Robust models are trained on high-quality, diverse datasets that accurately represent all scenarios. This helps prevent the model from becoming overly specialized and ensures it can perform well on unseen data.
- **Feature Engineering**: Thoughtful selection and transformation of features help enhance the model's ability to generalize by focusing on relevant information and reducing noise.
- **Regularization Techniques**: Methods like Ridge and Lasso and dropout are used to prevent overfitting by constraining the model's complexity during training, encouraging it to learn more generalizable patterns.
- **Cross-Validation**: Techniques such as k-fold cross-validation evaluate the model's performance across different data subsets, providing insights into its stability and generalization capabilities.
- **Ensemble Learning**: Combining multiple models through ensemble methods improves generalization by reducing the risk of individual model errors and enhancing overall performance on new data.
- **Evaluation on Test Data**: Robust models are evaluated on independent test datasets to assess their generalization ability and ensure reliable performance in real-world scenarios.

Implications for Model Accuracy:
While prioritizing robustness and generalization, there's typically a trade-off with model accuracy. Overfitting to the training data can lead to high accuracy on that specific dataset but may result in poor performance on new data. By focusing on robustness and generalization, the model's accuracy on the training data may decrease slightly, but it's more likely to perform well in diverse real-world situations. The aim is to find a

balance between accuracy and generalization, ensuring the model can handle varied inputs effectively while still achieving satisfactory performance.