**Question 1**

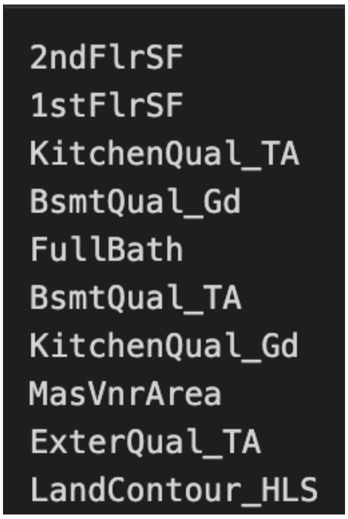
**What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

**Answer:**

The optimal value for alpha when using Ridge regression is 12 and when using the lasso regression it is 100. When we doubled up the alpha respectively, the R2 score of ridge regression on the test dataset becomes 0.825 and the R2 score of lasso regression would be 0.84.

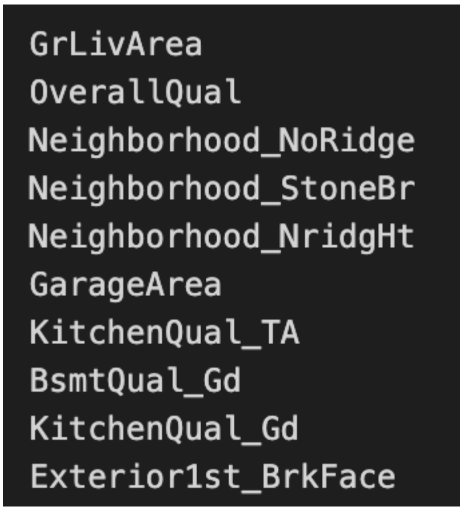
When changes are applied the following would be the important predictor variable:

Top features for Ridge Model



```
2ndFlrSF
1stFlrSF
KitchenQual_TA
BsmtQual_Gd
FullBath
BsmtQual_TA
KitchenQual_Gd
MasVnrArea
ExterQual_TA
LandContour_HLS
```

Top features for Lasso Model



```
GrLivArea
OverallQual
Neighborhood_NoRidge
Neighborhood_StoneBr
Neighborhood_NridgHt
GarageArea
KitchenQual_TA
BsmtQual_Gd
KitchenQual_Gd
Exterior1st_BrkFace
```

**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:**

According on the test data's R2 values, the Lasso model has a somewhat better R2 score (0.8417) than the Ridge model (0.8354). A higher R2 score shows that the Lasso model is doing marginally better on the test data as the R2 score is a measure of how well the model fits the data.

In addition to being effective, the Lasso regression also has the benefit of conducting feature selection by bringing some coefficients absolutely to zero. This feature selection characteristic might be useful when the dataset contains a large number of irrelevant or unimportant features. The Lasso model may provide a model that is simpler and easier to understand by removing unimportant elements, which is desired in many situations.

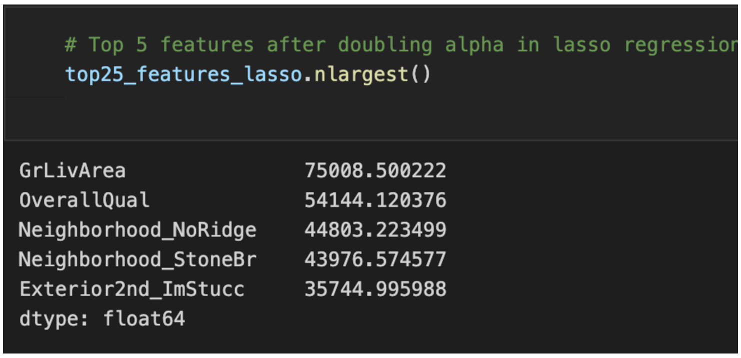
Nonetheless, the Lasso model's feature selection capability gives it a slight advantage over the Ridge model, making it a preferred choice in this scenario.

In our case, we have seen that when we performed the Lasso Regression the total number of features selected are 105. Where as the ridge model selected 199 features, this concludes that Lasso Regression is much optimized.

**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?**

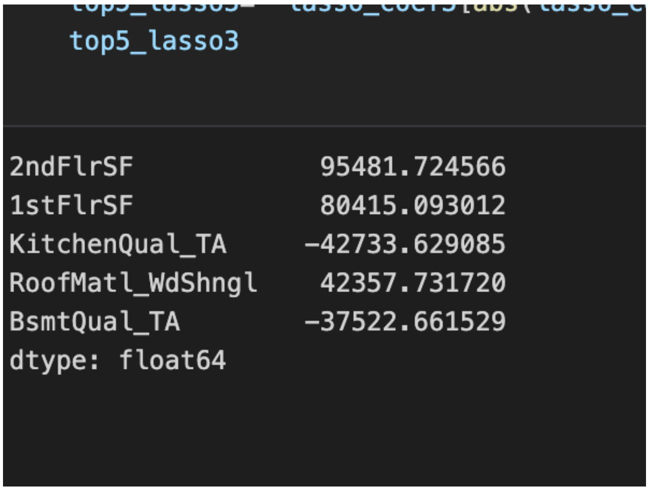
First Top 5 predictors in model



```
# Top 5 features after doubling alpha in lasso regressior
top25_features_lasso.nlargest()

GrLivArea               75008.500222
OverallQual             54144.120376
Neighborhood_NoRidge    44803.223499
Neighborhood_StoneBr    43976.574577
Exterior2nd_ImStucc     35744.995988
dtype: float64
```

Here, the Neighborhood columns are a dummy variable, we can delete the dummy variables of the Neighborhood column and the Neighborhood column itself. After performing this, the new top 5 features of the Lasso model are as below



```
top5_lasso3=  lasso_coefs[abs(lasso_c
top5_lasso3

2ndFlrSF           95481.724566
1stFlrSF           80415.093012
KitchenQual_TA    -42733.629085
RoofMatl_WdShngl   42357.731720
BsmtQual_TA       -37522.661529
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?**

To ensure that a model is robust and generalizable:

**Cross-validation:** Use techniques like k-fold cross-validation to assess the model's performance on multiple subsets of the data, reducing the risk of overfitting and providing a more accurate estimate of its generalization capabilities.

**Data Splitting:** Separate the dataset into training and testing sets to evaluate the model's performance on unseen data, ensuring that it generalizes well to new samples.

**Regularization:** Utilize techniques like Ridge and Lasso regression, which introduce regularization to prevent overfitting by penalizing large coefficients. Ridge regression uses L2 regularization, while Lasso regression employs L1 regularization.

**Implications for accuracy:**

**Ridge Regression:** Ridge helps to prevent overfitting and reduces the risk of high variance in the model. It tends to improve generalization, which may lead to slightly lower accuracy on the training data but can result in better performance on the test data.
**Lasso Regression:** Lasso performs both regularization and feature selection, as it can drive some coefficients to exactly zero. This can simplify the model and potentially lead to even better generalization, especially in datasets with many irrelevant features. The accuracy on the training data might decrease due to the feature elimination process, but it can often lead to improved performance on the test data.

In summary, utilizing regularization techniques like Ridge and Lasso, along with proper evaluation through cross-validation and data splitting, ensures that the model is more robust and capable of generalizing well to new, unseen data. Although these techniques may slightly impact the accuracy on the training data, they often lead to better overall performance on the test data, making the model more reliable for real-world predictions.