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?

For ridge, optimal value of alpha is 10 while for LASSO, it is 100.

If we double the optimal values of alpha, obviously there is a slight dip in the train metrics of LASSO and Ridge. But, for Ridge, there is a slight improvement in test RMSE.

For Ridge, top 5 variables remain unchanged, but there is a change in order.

Neighborhood_StoneBr (StoneBrook in neighborhood) replaces Condition2_PosN in top 5 for LASSO.

Important Variables after change is implemented:

Ridge:

```
OverallQual             40366.288869
2ndFlrSF                30797.231563
TotRmsAbvGrd            30334.047114
Neighborhood_NoRidge    29580.317402
GrLivArea               29492.518679
```

LASSO:

```
GrLivArea              126701.483872
OverallQual             98644.504683
GarageCars              48499.478575
Neighborhood_NoRidge    45061.801658
Neighborhood_StoneBr    38899.134677
```

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?

As I described in the notebook, LASSO model is more generalizable of the two because of lower error values of train and test.

LASSO:

```
{'Train R-squared': 0.890860040874442,
 'Test R-squared': 0.8754379259793063,
 'Train RMSE': 25641.854755181736,
 'Test RMSE': 29448.872328543894}
```

Ridge:

```
{'Train R-squared': 0.8812002148063836,
 'Test R-squared': 0.8613153676412659,
 'Train RMSE': 26752.56144084877,
 'Test RMSE': 31073.48213441446}
```

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?

In the absence of the previous important variables ("GrLivArea", "OverallQual", "Condition2_PosN", "Neighborhood_NoRidge", "GarageCars") for LASSO, the new top5 are:

```
1stFlrSF              138498.431302
2ndFlrSF              111448.749328
RoofMatl_WdShngl       53181.129272
GarageArea             46644.483697
Exterior2nd_ImStucc    46286.602150
```

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?

In order to make a model robust and generalizable, we need to strike the balance between train and test losses. In other words, the model should neither overfit nor underfit.

For an underfit model, both the train and the test losses are high, while for an overfit model, the test loss is much higher than the train loss.

Choosing a balance means we would have to compromise on the train loss, but at the cost of a more generalizable and robust model. For linear models like Linear Regression especially, this is achieved by tuning the regularization parameters of L1 and/or L2.