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?

Optimal Values for Ridge and Lasso

- Ridge : 4.0
- Lasso : 0.001

ridge_train_r2	0.9385203516514612
ridge_test_r2	0.912144014899987
lasso_train_r2	0.9344669145816011
lasso_test_r2	0.9154968712244933

The value of alpha decides the level of regularization in the model. It helps to shrink the model parameters towards zero to reduce the overfitting.

If we increase the alpha above the optimal level, the bias starts to increase rapidly as compared to the decrease in variance and we end up with a simpler model with higher bias.

After Doubling the alpha Values for Ridge and Lasso

- Ridge : 8.0
- Lasso : 0.002

ridge_train_r2	0.93476228475573
ridge_test_r2	0.9097785318821927
lasso_train_r2	0.9279938375731456
lasso_test_r2	0.9129802283948922

As Shown in the table, the model score decreases when we double the value of alpha for Ridge and Lasso. i.e., Model becomes much simpler.

After Doubling the alpha Values, most important predictor variable : OverallQual

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?

The given dataset has large number of attributes with multi collinearity present for many of them.

Although the model performance for Ridge is a little better than Lasso, it considers all the variables in the model.

On the other hand, Lasso helps us to remove the insignificant, redundant features by making their coefficients to zero. This will get rid of multicollinearity present in the dataset.

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?

Top 5 positive and negative predictor variables:

```
MSSubClass_DUPLEX               -0.221598
OverallCond_Fair                -0.197726
KitchenQual_Fa                  -0.177264
BldgType_Twnhs                  -0.168710
KitchenQual_TA                  -0.165490
OverallQual_Very Excellent       0.988202
OverallQual_Excellent            0.757556
Neighborhood_StoneBr             0.414192
OverallQual_Very Good            0.367675
Neighborhood_Crawfor             0.306493
```

After Removing top 5 positive and negative predictor variables:

ExterQual_TA	-0.322627	**SaleCondition_Partial**	0.331441
BsmtQual_Gd	-0.271432	**OverallCond_Excellent**	0.261087
Neighborhood_Gilbert	-0.260739	**Exterior1st_BrkFace**	0.240641
BldgType_Duplex	-0.256853	**BsmtExposure_Gd**	0.229118
Neighborhood_Edwards	-0.222231	**Functional_Typ**	0.225796

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?

The model is robust and generalizable when it can perform well on unseen data. Its important to regularize models with the help of cross validation so that the model is not overfitting the train data available.

When we regularize the model, we may observe reduced accuracy on the train data because we are reducing the complexity of the model by adding penalty on loss function to refrain model from learning train data too closely.