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 of alpha for ridge regression is “10” and that for lasso regression is “0.001”.

The models were built again after doubling the alpha values.

The Metrics before the changes are :

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.94	0.92
R2 Score (Test)	0.93	0.93
RSS (Train)	8.53	11.29
RSS (Test)	2.87	2.92
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.09	0.10
RMSE (Test)	0.10	0.10

The Metrics after the changes are :

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.93	0.91
R2 Score (Test)	0.93	0.91
RSS (Train)	9.37	13.49
RSS (Test)	2.82	3.45
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.09	0.11
RMSE (Test)	0.10	0.11

So, we can see that after the changes, the R2 Score for Ridge regression changed from 0.94 to 0.93 for train data and remained 0.93 for test data.

For Lasso, the R2 Score changed from 0.92 to 0.91 for train data and 0.93 to 0.91 for test data.

The other metrics changes have also been shown in the above tables.

Most important predictors after doubling the alpha are :

- GrLivArea
- OverallQual_8
- OverallQual_9
- Functional_Typ
- Neighborhood_Crawfor
- Exterior1st_BrkFace
- TotalBsmtSF
- CentralAir_Y

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:

In this case, the requirement is to find out the variables which are significant in predicting the price of the house. So, our primary goal is feature selection. We will go with “Lasso”.

If the requirement would be to reduce coefficient magnitude, then we would go for “Ridge”.

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:

In order to answer this question, we will drop the top 5 predictor variables and rebuild the model.

Steps have been done in the python notebook. The top 5 predictor variables.

```
betas['Lasso'].sort_values(ascending=False)[:5]

2ndFlrSF               0.10
Functional_Typ         0.07
1stFlrSF               0.07
MSSubClass_70          0.06
Neighborhood_Somerst   0.06
Name: Lasso, dtype: float64
```

Top 5 predictors are the following :

- 2ndFlrSF
- Functional_Typ
- 1stFlrSF
- MSSubClass_70
- Neighborhood_Somerst

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:

A model is “robust” if its predictions are consistent even if the input variables are changed due to any unforeseen circumstances.

A model is said to be generalizable, if it is able to adapt well to unseen data. It should be able to predict based on new data with the same level of accuracy as that when it used the training data.

Implications on accuracy:

If a model is not robust, that means, it will have high variance. So any changes in the features will result in major changes in the output variable. This might be a case of overfitting in case of very complex models. Thus, the model which is not robust might be accurate in case of training data, but will not be accurate when using unseen data.

Similarly, if there is overfitting, the model will not be generalizable. So, it will not give accurate results for unseen data. However, it will be accurate in case of training data.