Question 1

Q1: Part A - What is the optimal value of alpha for ridge and lasso regression?

As per the models developed, the optimal values of alpha are as follows:

- Ridge: 0.01
- Lasso: 0.0001

Q1: Part B - What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso?

Ridge:

For Ridge Regression, please refer to the table below to see how the R2 score has changed for the different alpha values (i.e. when we change the alpha from 0.01 to 0.02)

Metric	Ridge Regression - 0.01	Ridge Regression - 0.02
R2 Score (Train)	86.56%	86.53%
R2 Score (Test)	70.15%	71.94%

Lasso:

For Lasso Regression, please refer to the table below to see how the R2 score has changed for the different alpha values (i.e. when we change the alpha from 0.0001 to 0.0002)

Metric	Lasso Regression - 0.0001	Lasso Regression - 0.0002
R2 Score (Train)	84.83%	83.50%
R2 Score (Test)	82.98%	83.69%

For both Lasso and Ridge, we see that the r2 score increases for test data (i.e. unseen data) as the values of alpha increase. Also, we see that the r2 score decreases for the training data as the value of alpha decreases.

Q1: Part C - What will be the most important predictor variables after the change is implemented?

The most important predictor variables after the changes are implemented are as follows.

For Ridge:

We have changed the alpha value from 0.01 to 0.02. The top 10 predictor variables remain the same as when the alpha was 0.01. However, there is a change in the coefficient values.

Column_Name	Coefficient
GrLivArea	0.147653
1stFlrSF	0.146276
OverallQual	0.138767
LotArea	0.089661
MasVnrArea	0.076236
2ndFlrSF	0.074543
TotalBsmtSF	0.070361
BsmtFinSF1	0.064736
OverallCond	0.061409
RoofMatl	0.051473

For Lasso:

We have changed the alpha value from 0.0001 to 0.0002. Refer to the below table to see the top 10 predictor variables that we get with the new alpha (0.0002). Most of the variables are the same as before (except for that one has been removed "TotalBsmtSF"). "Functional" is a new entrant now.

Column_Name	Coefficient
GrLivArea	0.32001
OverallQual	0.172687
GarageCars	0.054642
MasVnrArea	0.054445
BsmtFullBath	0.040313
RoofMatl	0.035938
LandSlope	0.029859
Fireplaces	0.028317
OverallCond	0.023023
Functional	0.022957

Question 2

Q2: Part A - You have determined the optimal value of lambda for ridge and lasso regression during the assignment.

For this assignment, the optimal values of lambda are as follows,

- Ridge: 0.01
- Lasso: 0.0001

Q2: Part B - Now, which one will you choose to apply and why?

I will plan to proceed with the model developed using Lasso Regression.

Lasso regression is the best model to be used in this case as it gives a good accuracy for both training set data (~85%) and test set data (~ 83%). The test accuracy using Lasso Regression is very good compared to the other model developed using Ridge Regression.

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	86.56%	84.83%
R2 Score (Test)	70.15%	82.98%

Question 3

After building the model, you realized 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?

We have removed the five important predictor variables as per the initially developed Lasso Regression model.

columns_to_drop = ['GrLivArea', 'OverallQual', 'MasVnrArea', 'GarageCars', 'RoofMatl']

After this, we developed the model again. The five important predictor variables now are,

Column_Name	Coefficient
1stFlrSF	0.308429
2ndFlrSF	0.170193
TotalBsmtSF	0.116894
GarageArea	0.066253
OverallCond	0.04765

Question 4

Q4: Part A - How can you make sure that a model is robust and generalizable?

The assessment/usability of a machine learning model is decided based on its performance on unseen test data. This is referred to as generalizability. The practices that we should follow to ensure that a model is generalizable are,

- Implementing cross-validation techniques
- Applying regularization to avoid overfitting
- Maintaining an optimal level of model complexity, considering a good balance between bias and variance
- Hyperparameter Tuning
- Evaluating the model using various Performance Metrics

Q4: Part B - What are the implications of the same for the accuracy of the model and why?

The implications of having a robust and generalizable model are,

- Good accuracy on new/unseen data

- Achieving a good balance of understanding the patterns in the test data and also providing high accuracy with unseen data

Overall, a generalized model would be more reliable and practical.