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:
As per model, the optimal value of alpha for ridge regression is 20.0 and the optimal value of alpha for lasso regression is 0.001. After doubling the value of alpha for both ridge and lasso are respectively 40.0 and 0.002.

Ridge Regression when alpha=20:
R2 Score for Train: 0.9192
R2 Score for Test: 0.8737

Ridge Regression when alpha=40:
R2 Score for Train: 0.9184
R2 Score for Test: 0.8748

Lasso Regression when alpha=0.001:
R2 Score for Train: 0.9179
R2 Score for Test: 0.8752

Lasso Regression when alpha=0.002:
R2 Score for Train: 0.9161
R2 Score for Test: 0.8764

If we increase alpha for Lasso Regression, R2 Score on Train decreases slightly but R2 Score on Test increases slightly.

If we increase alpha for Ridge Regression, R2 Score on Train decreases slightly but R2 Score on Test increases slightly.

The most important predictor variables after the change is implemented:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond

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: It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretably.

Ridge regression has a particular advantage over OLS when the OLS estimates have high variance, i.e., when they overfit. Regularization can significantly reduce model variance while not increasing bias much. The tuning parameter lambda helps us determine how much we wish to regularize the model. The higher the value of lambda, the lower the value of the model coefficients, and more is the regularization. Choosing the right lambda is crucial so as to reduce only the variance in the model, without compromising much on identifying the underlying patterns, i.e., the bias.

Ridge regression does have one obvious disadvantage. It would include all the predictors in the final model. This may not affect the accuracy of the predictions but can make model interpretation challenging when the number of predictors is very large.

The behavior of Lasso regression is similar to that of Ridge regression. With an increase in the value of lambda, variance reduces with a slight compromise in terms of bias. Lasso also pushes the model coefficients towards 0 in order to handle high variance, just like Ridge regression. But, in addition to this, Lasso also pushes some coefficients to be exactly 0 and thus performs variable selection.
This variable selection results in models that are easier to interpret.

Hence, I will choose Lasso Regression in our assignment for final modelling.

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: Earlier, the five most important predictor variables were:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. YearBuilt

After excluding the above variables, the five most important predictor variables:

1. BsmtFinSF1
2. TotRmsAbvGrd
3. BsmtUnfSF
4. FullBath
5. LotArea

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 needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.