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 : Optimal value of alpha for ridge and lasso regression is 100 and 500 respectively. We tried changing the value of alpha to 2 times for both the regression models. The accuracy on test data decreases. Top predictor variables after the changes are implemented are :-

a. Ridge regression :
    i. GrLivArea
    ii. OverallQual_V_Good
    iii. TotalBsmtSF
    iv. 1stFlrSF
    v. YearRemodAdd
b. Lasso Regression :
    i. GrLivArea
    ii. OverallQual_V_Good
    iii. YearBuilt
    iv. YearRemodAdd
    v. TotalBsmtSF

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 : Optimal lambda value of ridge and lasso regression are 100 and 500. We will prefer the lasso regression model with 500 value of lambda as it simply our model by reducing many coefficients to zero and also improving accuracy on test data. On test data, R square value is 88.17% as compared to 86.96% of ridge regression.

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?

Answer: After removing top 5 predictor variables from lasso model and rebuilding it on rest of data set and target variable, below are the new 5 columns or important predictor variables

a. 1stFlrSF
b. 2ndFlrSF
c. GarageArea
d. YearRemodAdd
e. ExterQual_Gd

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 :
We can make sure that model is robust and generalized in following ways :

1. Cross Validation : We have built both ridge and lasso regression models using cross validation using 5 folds and ensured that our model is not based on any specific data split.
2. Train-Test Split : we have split our data set into a 70-30 ratio for running our model on unseen data
3. We also checked accuracy using key metrics like R square.
4. Regularization : After getting an overfit model using linear regression model, we use regularization using Ridge and Lasso models to implement regularization and maintain high accuracy and less complexity of model.

Ensuring robust and generalization impact directly to accuracy of the model :

1. Overfit : It ensure that model is not overfit on training data and hence ensure more accuracy on unseen data
2. Complexity : Higher complexity models generally reduce accuracy on unseen data set and hence we are able to simplify complexity using Lasso and Ridge and ensure huge improvement on accuracy on unseen/test data sets.
3. Linear regression also works well upon training data set but our goal for accuracy is on test data set that means models perform better on unseen data with decent accuracy.