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?

Ans-

Optimal value for -

Ridge Regression - **8.**

Lasso regression – 0.001

Changes after doubling the alpha value -

Ridge Regr- alpha = 16

And Lasso Reg- alpha = 0.002

Observed changes are decrease in R2 score and difference in RSS between Train and Test for both the regression algorithm. In Summary the metrics after doubling the alpha value is –

Metric	Ridge Regression	Lasso Regression
R2 Score Train	0.917546	0.896914
R2 Score Test	0.886285	0.881985
RSS Train	12.349241	15.439341
RSS Test	9.413000	9.768989
MSS Train	0.012083	0.015107
MSS Test	0.021491	0.022304

Predictor variables does not change much. Still the important variables are-

1. Area of living room
2. Overall Quality of material used
3. Garage size in terms of number of cars
4. Certain Neighborhoods like Stone Brook, Northridge Heights effects positively and other locations like Edwards, Old Town effect sale price negatively.
5. Sale type of the house like newly constructed house and then sold off immediately
6. Overall condition of house and area of the lot.

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?

Ans- Optimal value for lasso is chosen and applied as the R2 score for lasso has less difference between training and test data. Although the difference is very less compared to ridge but still that means the overfitting is comparatively less in Lasso as Compared to Ridge.

Lasso regression also helps in feature elimination.

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?

Ans-

Top feature for the model are – **GrLivArea, OverallQual , OverallCond , GarageCars, Neighborhood_Edwards.**

Now if these predictor variables are dropped then the model accuracy for Ridge and Lasso drops to 86.5 % and 86.1 % respectively. The top features now are- **Area of 1st Floor, Location like Stone Brook, NorthRidge Heights and Crawford, Basement Quality, Average Kitchen Quality etc.**

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?

Ans -

To derive a generalized and robust model below things needs to be considered

a. **Outliers**- Outliers should be analyzed and if not useful should be removed. Presence of outlier effect the performance of model on unseen data
b. Performance Metrics-
   Accuracy- Should have high accuracy. There should not be much difference between the test and train accuracy, because the difference shows overfitting in model. In our case the accuracy is 91 %.
   RMSE and Adjusted R2- These values can be used to check the robustness of the model. The low training and test errors points to a robust model. Also, there should not be much variations in values of training and test data.