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 1: The alpha value is 10 for alpha regression and for ridge regression is 1. For both, the score on training data decreases but increases on testing data.

The important predictor variables are PoolArea, FullBath, Neighbourhood_Gilbert, KtichenAbvGr, Neighbourhood_IDOTRR, BsmtHalfBath, 1stFlrSF, OpenPorchSF, BsmtFinType2, BsmtUnfSF.

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 2: The r2_scores are almost same for both of them but as Lasso will penalize more on the dataset and can also help in feature elimination I am going to consider Lasso as my final model.

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 3: PoolArea, FullBath, Neighbourhood_Gilbert, KtichenAbvGr, Neighbourhood_IDOTRR.

Question 4: How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer 4: The model should be generalized 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 importance should not be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers 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 much be removed from the dataset. If the model is not robust, it cannot be trusted for predictive analysis.