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?

The optimum value of alpha of ridge and lasso regression found during the initial model building was 20 and 0.0001 respectively. Coming to the r2_score, for ridge regression, the train and test r2_scores are 87.17 and 87.07 respectively, while at the same time, for lasso regression, the scores are 89.47 and 87.39 respectively. Both the models provide more or less equal r2_scores, but lasso, which is slightly better than ridge, which is about 0.31%. So, we can go with lasso regression.

If the value of alpha is doubled, in both the cased, training accuracy is less than testing accuracy.

Important Variables:

1. Ridge Regression
    a. NoRidge – Northridge (Physical Location)
    b. GrLivArea - Above grade (ground) living area square feet
    c. 2ndFlrSF - Second floor square feet
    d. MasVnrArea - Masonry veneer area in square feet
    e. NridgHt - Northridge Heights
2. Lasso Regression
    a. GrLivArea - Above grade (ground) living area square feet
    b. WdShngl - Wood Shingles
    c. MasVnrArea - Masonry veneer area in square feet
    d. NoRidge – Northridge (Physical Location)
    e. NridgHt - Northridge Heights

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?

The r2_score, for ridge regression, the train and test r2_scores are 87.17 and 87.07 respectively, while at the same time, for lasso regression, the scores are 89.47 and 87.39 respectively. Both the models provide more or less equal r2_scores, but lasso, which is slightly better than ridge, which is about 0.31%. So, we can go with lasso regression.

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?

Initially, the most important predictors are,

1. Lasso Regression
    a. GrLivArea - Above grade (ground) living area square feet
    b. WdShngl - Wood Shingles
    c. MasVnrArea - Masonry veneer area in square feet
    d. NoRidge – Northridge (Physical Location)
    e. NridgHt - Northridge Height

After removing the important predictors, the important predictors are

1. Lasso Regression
    a. 1stFlrSF: First Floor square feet
    b. 2ndFlrSF - Second floor square feet
    c. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
    d. GarageYrBlt: Year garage was built
    e. LotArea: Lot size in square feet

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?

By comparing the model, it is obvious that, training and test scores does not vary much, with respect to r2_score, which represents there is no overfitting in the data and underfitting. And it should also follow the assumptions of regression

1. Non-Linearity of errors
2. Normality of error terms
3. Heteroskedasticity

If the testing accuracy is very low than training and training accuracy also very less, than it is considered as under-fitting, if it is low, it is considered as over fitting (does not explain the test data but explains train data). By keeping it in mind, from the above we can say that, our model is generalised to our data and works well with both train and test data.