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 optimal value of alpha for ridge = 5.5. The optimal value of alpha for lasso = 200.0. There is a slight reduction in r2 score for ridge and lasso both when the alpha value is doubled. For Ridge, when the alpha is 5.5, the most important predictor variable is the overall condition with the coefficient 58245.7. This however changes when the alpha is doubled. The new most important predictor variable changes to lot frontage with the coefficient 81737.5. For Lasso, when alpha is 200, the most important predictor variable is Overall Condition with the coefficient 112140.4. Upon changing the alpha to 400.0, the most important predictor remains the same, however the coefficient reduces to 107552.7.

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?

- From calculations and analysing the coefficients, Lasso seems a better approach to solving this problem. Not only the train and test r2 score is better, but lasso has a lot of predictor variables with zero as coefficient. This means, the number of predictor variables have reduced. This would in turn reduce the complexity of the model and also the computational power for machine learning algorithm.

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?

- Fireplaces, LotFrontage, GarageQual, ExterCond and TotRmsAbvGrd are the new top five predictor variables.

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?

- The machine learning model should be as simple as possible. It is a trade-off between accuracy and simplicity, but it would lead to better robustness and also would be generalisable. One important aspect would be using Bias-Variance trade-off. Simpler is the model, more is the bias, but less is the variance, and more generalizable the model becomes. A model which is more robust and generalizable, would perform well on both training and test data.

Bias is the error in model, when the model is weak to learn from the data provided. A model with high bias would be unable to learn details using the data.

Variance is the error in the model, when the model is trying to overlearn from the provided data. High variance would mean that the model performs exceptionally well on the training data, but poorly on the test data.

Introducing a trade-off between bias and variance is important to avoid under-fitting and over-fitting of data.