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:

- Optimal Alpha Value for Ridge is 20 and for Lasso is 0.001.
- If we double the alpha values then there will be overfitting introduced which can be confirmed by the R2 scores of train and test set
- R2 square increased on training but decresed on testing by doubling the alpha for lasso
  - `Alpha- 0.001,Train R2-0.9328610693783794`
  - `Alpha- 0.001,Test R2-0.8818397232646152`
  - `Alpha- 0.002,Train R2-0.9409894041793858`
  - `Alpha- 0.002,Test R2-0.8755950054873416`
- R2 square decreased on training but increased on testing by doubling the alpha for ridge
  - `Alpha- 20,Train R2- 0.908`
  - `Alpha- 20,Test R2- 0.8939`
  - `Alpha- 40,Train R2-0.901`
  - `Alpha- 40,Test R2-0.895`
- After the change the important variable are through ridge (changing alpha from 20 to 40)
  - OverallQual
  - Neighborhood_Crawfor
  - BsmtFullBath
  - Neighborhood_NridgHt
  - Condition1_Norm
- After the change the important variable are through lasso (changing alpha from 0.001 to 0.002)
  - `PoolQC_none            |   1.933 |`
  - `SaleCondition_Alloca   |   0.199 |`
  - `Neighborhood_StoneBr   |   0.134 |`
  - `Functional_Maj2        |  -0.181 |`
  - `Heating_Grav           |  -0.185 |`
  - `Condition2_PosN        |  -0.603 |`
  - `PoolQC_Gd              |  -1.11`

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:

I will apply the lasso regression as it makes the non significant features's coefficient 0, hence it helps with the feature selection as well.

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?

Answer 3:

The top 5 important variable are

- PoolQC_none | 3.632 |
- SaleCondition_Alloca | 0.207 |
- Functional_Maj2 | -0.232 |
- Condition2_PosN | -0.738 |
- PoolQC_Gd | -1.497

If we don't get these variables in the incoming data then the most important will be-

- | Neighborhood_Crawfor | 0.097 |
- | Neighborhood_StoneBr | 0.096 |
- | Neighborhood_NridgHt | 0.094 |
- | OverallQual | 0.07 |
- | Exterior1st_BrkFace | 0.068 |

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 4:

- The Model is robust and generaliasable if the R2 score is similar on testing data as well as training data which indicates there is no underfitting or overfitting.
- To make the model robust we do regularization using Ridge or Lasso
- Accuracy might suffer or drop a bit if we regularize but that is fine as long as as there is no significant drop and R2 scores are stable or similar for training and testing set as accuracy metrics is not reliable much as it doesn't account for overfitting and underfitting case