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: The optimal Value for alpha for Ridge and Lasso Regression are 10 and 0.001 respectively.
On doubling the value of alphas for regularisation technique, there isnt any significant impact on the model. Model score changes , however, the change is very small.

For Ridge Regression, after implementing the change, following variables are important to the model:

Neighborhood_Crawfor
Neighborhood_NridgHt
OverallQual
Condition1_Norm
Neighborhood_Somerst
BsmtExposure_Gd
Neighborhood_ClearCr
MSZoning_RL
GarageCars
CentralAir_Y
OverallCond

And others

For Lasso, after implementing the change, following variables are important to the model
Neighborhood_Crawfor
OverallQual
Neighborhood_NridgHt
Neighborhood_Somerst
Condition1_Norm
OverallCond
GarageCars
BsmtExposure_Gd
BsmtFinType1_GLQ
MSZoning_RL
CentralAir_Y
and others

Which suggests that houses with good neighbourhoods and have good overall condition and quality and probably spacious garage is more expensive.

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: In this case, comparing both regression techniques, we notice that there is hardly any difference in scores for Ridge and Lasso, however, Ridge slightly performs better than Lasso as Ridge regression gives the score of **0.9154059**
And 0.8890314 respectively on train and test dataset

Lasso gives a score of **0.9001001 and 0.8860912** on the train and test dataset respectively. Additionally, both models provide more or less similar coefficient values.

Both the techniques have their own advantages, however, considering the number of parameters, I preferred Lasso here as it is very helpful for feature selection by reducing it to 0 which leaves the model with significant features to focus upon.

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?

Ans : The previous model build used below were the top 5 features:
Neighborhood_Crawfor
Neighborhood_Somerst
Neighborhood_NridgHt
OverallQual
Condition1_Norm

After re-building the model after dropping top five features, below are the most important features now:
MSZoning_FV
BsmtExposure_Gd
OverallCond
MSZoning_RL
BsmtFinType1_GLQ
SaleType_New

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 ensure a model is robust and generalizable is by making sure the model is not overfitting, which can be checked by different metrics and evaluation techniques. Depending on the model built ,Linear Regression in this case, Mean Square Error, R2_score, RSS score etc can be checked on both train and test data. Similarly, confusion matrix, specificity, accuracy etc should be checked,
taking care of Bias Variance trade-off. Overfit models usually have higher accuracy on the data on which it is trained, however, fails or accuracy drops, considerably on test data. Under-fit models usually have lesser accuracy on train and test data both.
Balancing this and achieving better accuracy on both train and unseen data would make a model robust and generalisable.