Question 1

What is the optimal value of alpha for ridge and lasso regression?

For ridge the optimal value of alpha came up as 100, for lasso the optimal value if alpha is .001.

What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso?

What will be the most important predictor variables after the change is implemented?

In both above equation alpha denotes the hyper parameter lambda. If the penalty increases beyond the optimal value, then the co-efficient tend to become more towards zero and lead to underfitting of the model with low variance and with very high bias.

By doubling alpha the most important predictor variables itself will not alter, only their values will be further reduced.

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?

I will choose Lasso over Ridge as explained below:

1. Variable Sensitivity:
   - In house price prediction, not all variables are equally important. Unlike drug sampling, where each variable might carry critical significance, house price prediction can have many less impactful features.
   - Lasso allows you to reduce the impact of less important variables by shrinking their coefficients towards zero, simplifying the model.
2. Handling Multicollinearity:
   - Multicollinearity occurs when predictor variables are highly correlated. Ridge regression mitigates multicollinearity by shrinking coefficients, but it doesn't eliminate variables.
   - Lasso, however, can drop some feature variables entirely if they are collinear. This feature selection capability simplifies the model.
3. Feature Selection:
   - Lasso sets unimportant coefficients to exactly zero, effectively removing those features from the model.
   - This reduction in feature variables improves interpretability and prevents overfitting.

In summary, Lasso strikes a balance between regularization and feature selection, making it suitable for house price prediction. The choice between Lasso and Ridge depends on dataset characteristics and modeling goals.

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?

Here we created a new lasso model after removal of the top 5 columns. This new model resulted below when we finally create a data frame with the feature variables and the sorted orders of the betas.

	Feature	Coef	Coef_abs
14	BsmtFinSF1	0.104461	0.104461
16	BsmtUnfSF	0.080710	0.080710
19	2ndFlrSF	0.070871	0.070871
0	LotArea	0.045425	0.045425
5	YearBuilt	0.045377	0.04537

So, now the top 5 predictor variables are

1. BsmtFinSF1: Type 1 finished square feet
2. BsmtUnfSF: Unfinished square feet of basement area
3. 2ndFlrSF: Second floor square feet
4. LotArea: Lot size in square feet
5. YearBuilt: Original construction date

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?

A model's robustness is highly influenced by its variance factor. If the model's variance is higher then may perform good on the train data set but as soon the train data changes the model accuracy may significantly drop.

A model as it gets more complex tries to memorize the train data and thus loses it's capability of being generalized over unseen data.

A robust model may compromise to some extent with the accuracy of it's prediction as it is not overfitted.

Regularization is introduced to impart a penalty factor on the model's cost function and thus prevent it from being overfitted.

So, a proper balance between the model accuracy and generalization is considered to build an efficient model.