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 to double the value of alpha for both ridge and lasso?

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

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

For both Ridge the cost function = $\sum_{i=1}^{n}(y_i - \hat{y})^2 + \lambda \sum_{J=1}^{p} \beta_J^2$

For Lasso, cost function = $\sum_{i=1}^{n} (y_i - \hat{y})_{+}^{2}\lambda \sum_{j=1}^{b} |\beta_J|$

In both above equation alpha denotes the hyper parameter lambda which is also known as the penalty factor for regularization. If the penalty increases beyond the optimal value, then the co-efficient tend to become more towards zero. And this may 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 here because of the factors

1. Variable Sensitivity – House price as business case does not carry the importance of each and all variables like in a drug sampling. So, I would prefer to reduce the not so important variables as much possible.
2. Handles Multicollinearity - Lasso handles multicollinearity better than Ridge and thus can drop some of the feature variables which are collinear to each other.

3. Feature Selection – Lasso as an add on privilege provide the feature selection by setting the unimportant coefficient to exactly zero and thus reduce the number of feature variables and increase interpretability.

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.

```
# Top 5 predictors as per this new Lasso model
lasso_new_top5 =
pd.DataFrame(list(zip(X_train_del_top5.columns,lasso.coef_)),columns=['Feature
','Coef'])

# Sort this data frame based on the absolute values of the coefficients
lasso_new_top5['Coef_abs'] = np.abs(lasso_new_top5['Coef'])
lasso_new_top5.sort_values(by='Coef_abs',ascending=False,inplace=True)
lasso_new_top5.head()
```

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.