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:

The optimal values of alpha for ridge and lasso regression are 0.0001 and 0.00001 respectively. We have chosen Lasso as it is giving lowest test R2 with this alpha. The top 10 features, in this case will be 'GrLivArea', 'OverallQual', 'LotArea', 'OverallCond', '1stFlrSF', 'age', 'GarageArea', 'BsmtFinSF1', '3SsnPorch' and 'Fireplaces'

If we double the alphas, below are resulting train/test R2 and RMSE. Based on these values, we are again choosing Lasso as it is giving higher R2 and lower RMSE.

```
Model Evaluation : Ridge Regression alpha=.0002
R2 score (train) :  0.8791020706823203
R2 score (test) :  0.8847823335235915
RMSE (train) :  0.045715080472932185
RMSE (test) :  0.04283561681194803
=============
Model Evaluation : Lasso Regression alpha=0.00002
R2 score (train) :  0.8789856817448308
R2 score (test) :  0.8856834276080051
RMSE (train) :  0.04573708022808691
RMSE (test) :  0.04266778366948434
```

After implementing this change, top 10 features will be like below. Their coefficient values has been decreased but there no change in list of top 10 feature, which we got using the original lasso model.

	Lasso alpha=0.00002
GrLivArea	0.326470
OverallQual	0.195657
LotArea	0.144838
OverallCond	0.110205
1stFlrSF	0.103020
age	0.076970
GarageArea	0.075121
BsmtFinSF1	0.062481
Fireplaces	0.049059
3SsnPorch	0.045566

Here, we are considering both positive and negative impact of features on target variable.

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: From the below metrics comparison, we see that the Linear regression is as good as Ridge regression (alpha = .0001), basically no regularisation is applied in Ridge. However, as all metrics are working slightly better in Lasso regression (alpha=.00001) on unseen test data, hence, choosing Lasso as the final model.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.879102	0.879102	0.879073
1	R2 Score (Test)	0.884782	0.884782	0.885267
2	RSS (Train)	2.133756	2.133756	2.134270
3	RSS (Test)	0.803681	0.803682	0.800304
4	MSE (Train)	0.045715	0.045715	0.045721
5	MSE (Test)	0.042836	0.042836	0.042746

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: The five most important variables present in original Lasso model was

['GrLivArea', 'OverallQual', 'LotArea', 'OverallCond', '1stFlrSF']

If we rebuild the model, after removing these features, below will be the five most important predictor variables.

	Lasso
BsmtFinSF1	0.234023
BedroomAbvGr	0.167870
GarageArea	0.150012
Fireplaces	0.118172
FullBath	0.109272

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: A model becomes robust and generalizable when it is following the general pattern instead of following noise present in the data. Such models show similar fair accuracy metrics on both training and unseen data. So basically, a model should not overfit the data and be as simple as it can be at the same time. A robust and generalized model should have a fair test R2 values which is not very less than the train R2 and test RMSE should not be very high than the train RMSE. Our selected model shows both the properties of being a robust and generalizable one.