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?

Optimal values of alpha which I got is

Ridge = 0.7

Lasso = 0.0001

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.927501	0.933712
1	R2Score Test	0.898518	0.905344
2	RSS Train	1.424779	1.302719
3	RSS Test	0.820164	0.764999
4	MSE Train	0.001438	0.001315
5	MSE Test	0.001925	0.001796

When double the value of alpha

Ridge = 1.4

Lasso = 0.0002

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.926564	0.930370
1	R2Score Test	0.898008	0.903099
2	RSS Train	1.443208	1.368409
3	RSS Test	0.824285	0.783147
4	MSE Train	0.001456	0.001381
5	MSE Test	0.001935	0.001838

Above ridge given the very close result of R2 score for both train and test. The most important feature of ridge regression after double the value of alpha is :

1. Fireplaces
2. OverallCond
3. BsmtFinSF1
4. BsmtFinSF2
5. OverallQual
6. GrLivArea
7. WoodDeckSF
8. LotArea
9. Neighborhood_MeadowV
10. SaleType_Oth

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 prefer Lasso regression because it would help in feature elimination and the model will be more robust and it also showing minimum difference between R2 of test and train, good value for RSS, MSE and RMSE.

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?

I have dropped five most important predictor variables in lasso model as below

- BsmtFullBath
- OverallCond
- MasVnrArea
- HeatingQC
- OverallQual

Created another model and

	Featuere	Coef
7	BsmtFinSF1	0.332896
0	MSSubClass	0.225616
6	BsmtFinType1	0.111499
3	ExterCond	0.085780
9	BsmtFinSF2	0.083012
1	LotArea	0.064734
19	GarageFinish	0.058287
24	MoSold	0.057154
57	Neighborhood_SWISU	0.039218
8	BsmtFinType2	0.035935

Got below five most important variables

- BsmtFinSF1,
- MSSubClass,
- BsmtFinType1,
- BsmtFinSF2 and
- ExterCond

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?

1] The model is robust and generalizable when the difference between the train and test score is minimum

2] The model should not impact by the outliers therefore outlier treatment is important to get the robust model.

3] We can detect outliers in the dataset using boxplot. treating the outliers will not affect the mean, median, etc. so that we can input correct values to the missing values. Outliers' analysis should be done to avoid bias.

4] The predicted features should be significant and this significance can be determined by p-values, R-square and adjusted R2

Implications of accuracy of a model:

1] Get more data as we can.

2] Fix missing values and outliers.

3] Derive new variables from existing one and also standardize the values.

4] Feature selection using RFE.

5] Applying the right algorithm (selecting right ML algorithm is important to get accurate model).

6] Cross validation.