Q1. 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?

From my created model alpha value are below:

Lasso: 0.001

Ridge: 3

CASE1

In case making alpha double for Ridge, i,e 6

Value of R2 decreased in training and test data set as mentioned below:

Ridge = 3

R2 Score on train data: 0.9149300630462407
R2 Score on test data: 0.9227217093381243

Ridge = 6

R2 Score on train data: 0.9101661567017255
R2 Score on test data: 0.9218317277072438

Left with alpha: 6, Right with alpha 3

	Feaure	Coef
39	EnclosedPorch	0.292054
15	BsmtFinType2	0.289780
23	GrLivArea	0.282540
88	Condition1_RRNe	0.257760
2	LotArea	0.226545
57	LotConfig_Inside	0.226122
30	KitchenQual	0.219182
96	BldgType_2fmCon	0.206300
17	BsmtUnfSF	0.204765
11	BsmtCond	0.191865

	Feaure	Coef
15	BsmtFinType2	0.413511
39	EnclosedPorch	0.339518
17	BsmtUnfSF	0.314169
23	GrLivArea	0.307043
88	Condition1_RRNe	0.297470
30	KitchenQual	0.259851
57	LotConfig_Inside	0.246144
96	BldgType_2fmCon	0.222523
2	LotArea	0.217375
37	WoodDeckSF	0.208735

Most important predictor after change in alpha: **EnclosedPorch**

CASE2

In case making alpha double for Lasso, i,e 0.002

Value of R2 decreased in training and test data set as mentioned below

Lasso = 0.001

R2 Score on train data: 0.9123454803572383
R2 Score on test data: 0.9240928710053207

Lasso = 0.002

R2 Score on train data: 0.9054618835340047
R2 Score on test data: 0.9229668461008835

Left with alpha: 0.002, Right with alpha 0.001

	Featuere	Coef
11	BsmtCond	0.338454
23	GrLivArea	0.270549
2	LotArea	0.247376
96	BldgType_2fmCon	0.243869
57	LotConfig_Inside	0.209273
15	BsmtFinType2	0.207768
88	Condition1_RRNe	0.198567
3	LotShape	0.153827
77	Neighborhood_SawyerW	0.150922
100	HouseStyle_1.5Unf	0.150003

	Featuere	Coef
11	BsmtCond	0.334686
23	GrLivArea	0.298685
88	Condition1_RRNe	0.298289
15	BsmtFinType2	0.288403
57	LotConfig_Inside	0.251372
96	BldgType_2fmCon	0.247226
2	LotArea	0.233744
30	KitchenQual	0.208040
17	BsmtUnfSF	0.189826
86	Condition1_RRAe	0.189166

Most important predictor after change in alpha: **BsmntCond**

Q2. 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?

Optimum value of lambdas being calculated after multiple value of it, iterated through model. With increase in λ, the polynomial regression model becomes progressively simpler as the coefficients are pushed down towards zero.

As Lasso regression penalize more on the dataset, intern helping in more feature elimination, we will consider Lasso Regression model as the final model.

Q3. 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?

Below are the five most important predictor variables in lasso model:

1) BsmtCond
2) GrLivArea
3) Condition1_RRNe
4) BsmtFinType2
5) LotConfig_Inside

Rebuild the model without above variable, the most important variables are below:

1) BldgType_2fmCon
2) LotArea
3) KitchenQual
4) BsmtUnfSF
5) WoodDeckSF

Q4. 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?

This is very challenging to make the model robust and generalizable but we can try to maintain the training data is comprehensive, diverse, and representative of the real-world scenarios that the model will encounter. This helps the model learn patterns that are more likely to generalize well to unseen data, on the other hand techniques like k-fold cross-validation to assess the model's performance on multiple subsets of the data. This helps in detecting overfitting and ensures that the model's performance is consistent across different data samples.

Optimize model hyperparameters using techniques like grid search or random search to find the configuration that yields the best generalization performance.