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?

Ans -

1. The optimal value of alpha
    a. Ridge -> 2.0
    b. Lasso -> 100
2. If we double the alpha
    a. ridge alpha from 2.0 to 4.0

Alpha=2	Alpha=4
R2 -Train : 0.8733343102496596	R2 -Train : 0.8680756683671782
R2- Test : 0.8421736827875954	R2- Test : 0.8407498931175879
RSS-Train 209182234427.9912	RSS-Train 217866626082.9234
RSS-Test 106764077839.38605	RSS-Test 107727222604.08981
MSE-Train 396930236.10624516	MSE-Train 413409157.65260607
MSE-Test 472407424.0680798	MSE-Test 476669126.56676906

R2 score on training data decreases and also there is slight decrease in test data R2

    b. lasso alpha from 100 to 200

Alpha=100	Alpha=200
R2 -Train : 0.8720951204887207	R2 -Train : 0.8679022045771799
R2- Test : 0.8500816934602986	R2- Test : 0.8496027147896281
RSS-Train 211228696130.32275	RSS-Train 218153093106.91144
RSS-Test 101414580480.97708	RSS-Test 101738593085.35182
MSE-Train 400813465.1429274	MSE-Train 413952738.34328544
MSE-Test 448737081.7742349	MSE-Test 450170765.8643886

R2 score on training data decreases and also there is slight decrease in test data R2

3. After change is implemented the most important predictor variable are

Column1	Ridge	Lasso
OverallQual	80393.72232	9.97E+04
OverallCond	26100.65832	2.79E+04
YearBuilt	59238.2128	6.12E+04
BsmtFinSF1	35222.81321	2.98E+04
BsmtUnfSF	8087.271073	0.00E+00
TotalBsmtSF	33058.21629	3.23E+04
1stFlrSF	41123.15317	1.54E+02
2ndFlrSF	37501.1101	0.00E+00

OverallQual, OverallCond, YearBuilt, BsmtFinSF1, BsmtUnfSF, TotalBsmtSF, 1stFlrSF 2ndFlrSF

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?

Ans - R2 score for Train data set in Ridge is almost similar to that lasso. However in case of test data set lasso R2 is slightly higher, so lasso will be the best choice to proceed.

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?

```
6 coef_reg.head(100)
```

Out[80]:

	Lasso
BsmtUnfSF	0.000000
1stFlrSF	34470.477698
2ndFlrSF	0.000000
LowQualFinSF	-20933.909612
GrLivArea	153954.224757
GarageArea	101945.371711
RoofStyle_Gable	2841.811116
RoofStyle_Gambrel	-0.000000
RoofStyle_Hip	-0.000000
RoofStyle_Mansard	0.000000
Exterior1st_BrkComm	-39001.520445
Exterior1st_CemntBd	24146.072197
Exterior1st_Stucco	-3251.535475
Exterior2nd_Brk Cmn	-0.000000
Exterior2nd_CmentBd	52.070785
Exterior2nd_Stucco	-0.431538
GarageCond_Fa	-25103.764681
GarageCond_Gd	-0.000000
GarageCond_Po	-44306.524616
SaleType_Con	46862.520451

Five most important predictor variables are GrLivArea, GarageArea, SaleType_Con, 1stFlrSF ,Exterior1st_CemntBd

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?

The model accuracy will be generalized so that the model accuracy is not less than the training set accuracy. However the model accuracy shouldn’t be dropped with other dataset. The key areas is to identify the good data (means outlier should be treated well in advance) so that the model is more robust for different data sets