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?

The optimum values of alpha obtained are **1 for ridge and 0.0001 for Lasso**

Original values

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.919236	0.919190
1	R2 Score (Test)	0.882839	0.882459
2	RSS (Train)	12.958133	12.965575
3	RSS (Test)	8.477038	8.504503
4	MSE (Train)	0.012692	0.012699
5	MSE (Test)	0.019310	0.019372
6	RMSE (Train)	0.112657	0.112689
7	RMSE (Test)	0.138960	0.139185

On doubling the alpha

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.917914	0.917914
1	R2 Score (Test)	0.883121	0.883121
2	RSS (Train)	13.170236	13.170236
3	RSS (Test)	8.456653	8.456653
4	MSE (Train)	0.012899	0.012899
5	MSE (Test)	0.019263	0.019263
6	RMSE (Train)	0.113575	0.113575
7	RMSE (Test)	0.138793	0.138793

Observation: There is a minute decrease (3rd order decimal) to R2 in the train and test set for both Ridge and Lasso.

Coefficients changes for Predictor variables

Original values

```
OverallQual_Excellent                0.213995
Neighborhood_NridgHt                 0.149678
Neighborhood_NoRidge                 0.138238
Neighborhood_Crawfor                 0.132234
OverallQual_Very Good                0.131093
Neighborhood_StoneBr                 0.119630
Neighborhood_Somerst                 0.119350
OverallCond_Excellent                0.109616
OverallQual_Very Excellent           0.104470
Neighborhood_ClearCr                 0.086901
MSSubClass_2-1/2 STORY ALL AGES      0.072413
SaleCondition_Partial                0.068016
Neighborhood_Veenker                 0.066420
GrLivArea                            0.063446
BsmtExposure_Gd                      0.060882
OverallCond_Very Good                0.054710
SaleCondition_Normal                 0.054673
OverallQual_Good                     0.052662
Exterior2nd_MetalSd                  0.049731
2ndFlrSF                             0.047862
Name: Ridge, dtype: float64
```

On doubling the alpha

```
OverallQual_Excellent                0.200143
Neighborhood_NridgHt                 0.135041
OverallQual_Very Good                0.126250
Neighborhood_Crawfor                 0.122665
Neighborhood_NoRidge                 0.119018
Neighborhood_Somerst                 0.109239
OverallCond_Excellent                0.104502
Neighborhood_StoneBr                 0.102728
OverallQual_Very Excellent           0.083833
Neighborhood_ClearCr                 0.073615
MSSubClass_2-1/2 STORY ALL AGES      0.066412
SaleCondition_Partial                0.066247
GrLivArea                            0.064248
BsmtExposure_Gd                      0.060328
OverallCond_Very Good                0.056056
Neighborhood_Veenker                 0.055531
SaleCondition_Normal                 0.054291
OverallQual_Good                     0.049735
2ndFlrSF                             0.048289
MSSubClass_1-STORY 1946 & NEWER ALL STYLES    0.041306
Name: Ridge, dtype: float64
```

```
OverallQual_Excellent                0.225407
Neighborhood_NridgHt                 0.152167
Neighborhood_NoRidge                 0.143192
Neighborhood_Crawfor                 0.141590
OverallQual_Very Good                0.136055
Neighborhood_Somerst                 0.134324
Neighborhood_StoneBr                 0.121012
OverallQual_Very Excellent           0.109521
OverallCond_Excellent                0.109516
Neighborhood_ClearCr                 0.092407
Neighborhood_Veenker                 0.068806
SaleCondition_Partial                0.065214
BsmtExposure_Gd                      0.061631
2ndFlrSF                             0.059537
MSSubClass_2-1/2 STORY ALL AGES      0.058001
OverallQual_Good                     0.055116
OverallCond_Very Good                0.053875
SaleCondition_Normal                 0.052154
GrLivArea                            0.047632
1stFlrSF                             0.046092
Name: Lasso, dtype: float64
```

```
OverallQual_Excellent                0.219288
OverallQual_Very Good                0.132459
Neighborhood_NridgHt                 0.130515
Neighborhood_Crawfor                 0.129373
Neighborhood_NoRidge                 0.118820
Neighborhood_Somerst                 0.114878
GrLivArea                            0.108221
OverallCond_Excellent                0.105081
Neighborhood_StoneBr                 0.096759
OverallQual_Very Excellent           0.085727
Neighborhood_ClearCr                 0.075824
SaleCondition_Partial                0.060788
BsmtExposure_Gd                      0.060766
OverallCond_Very Good                0.056168
OverallQual_Good                     0.052386
SaleCondition_Normal                 0.050059
Neighborhood_Veenker                 0.049828
MSSubClass_2-1/2 STORY ALL AGES      0.046595
GarageCars                           0.040859
LotConfig_CulDSac                    0.040108
Name: Lasso, dtype: float64
```

Predictor variables are still the same but the order has changed.

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?

In this case where we have too many variables it's better to use Lasso.

Question 3

After building the model, you realized 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?

```
OverallCond_Excellent              0.127606
MSSubClass_2-1/2 STORY ALL AGES    0.090685
2ndFlrSF                           0.072037
BsmtExposure_Gd                    0.067025
1stFlrSF                           0.058205
Name: Lasso, dtype: float64
```

Now we have overall condition to be excellent, MSSubCalss- 2-½ Story, 2nd floor Square ft,Good Basement Exposure, 1st floor SF to be Top 5 parameters

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?

We can ensure to keep the model robust and generalisable by ensuring that overfitting is avoided to maximum extent. This is possible by evaluating residual error of the model against predictors to check whether they fit the linear regression conditions (heteroskedasticity, no noticeable pattern between predictors and residuals). Further by employing strategies like Cross validate (GridSearchCV etc.), robust feature engineering (here we introduced Age as feature) and ensuring R2 scores are nominal in both train and test. Too complex model will have high accuracy but will have decreased variance and thereby low bias. By generalizing models, we are to a few extent compromising on accuracy by introducing bias. However, we ensure as we saw here that RMSEs are not overly increased and are fairly regulated to keep the model generalized and thereby avoiding overfitting.