Question 1-

1.1. What is the optimal value of alpha for ridge and lasso regression?
1.2. What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?
1.3. What will be the most important predictor variables after the change is implemented?

Answer 1-

1.1. The optimal value of alpha(lambda) for Ridge is 0.05 and for Lasso is 0.0001.

1.2.

- **The results that we got with actual alpha values for ridge and lasso are-**

Models	r2_score	rss	rmse
Linear_Train	0.871348	2.290580	0.048785
Linear_Test	0.852227	0.901205	0.048785
Ridge_Train	0.871314	2.291198	0.048785
Ridge_Test	0.853775	0.891764	0.048785
Lasso_Train	0.866726	2.372879	0.048785
Lasso_Test	0.854459	0.887589	0.048785

- **The results that we got by doubling the alpha values in ridge and lasso-**

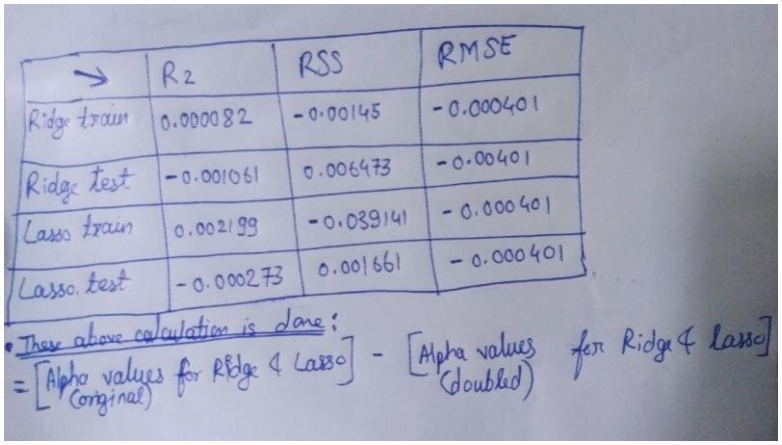
Models	r2_score	rss	rmse
Linear_Train	0.871348	2.290580	0.049186
Linear_Test	0.852227	0.901205	0.049186
Ridge_Train	0.871232	2.292648	0.049186
Ridge_Test	0.854836	0.885291	0.049186
Lasso_Train	0.864527	2.412020	0.049186
Lasso_Test	0.854732	0.885928	0.049186

Observation- 1. R2-Score: We can see that there is a very marginal decrease in the r2 score of Ridge and Lasso in train data and a very marginal increase in the r2 score of Ridge and Lasso test data. (Refer Fig:1)

Observation- 2. There is a marginal increase in the RSS value of the Ridge and Lasso train set and a marginal decrease in the RSS value of Ridge and Lasso test set. (Refer Fig:1)

Observation- 3. The RMSE value increases as we double the alpha values for both Ridge and Lasso. (Refer Fig:1)

Fig: 1



→	R_2	RSS	RMSE
Ridge train	0.000082	-0.00145	-0.000401
Ridge test	-0.001061	0.006473	-0.00401
Lasso train	0.002199	-0.039141	-0.000401
Lasso test	-0.000273	0.001661	-0.000401

• These above calculation is done:
= [Alpha values for Ridge & Lasso (Original)] - [Alpha values for Ridge & Lasso (doubled)]

1.3. After the changes are implemented, the following will be the list of most important variables- (All the features with non-zero values in Lasso column) Refer Fig: 2.

Fig: 2

constant	-0.273693	-0.240140	-0.075594
LotArea	0.070336	0.069269	0.050207
YearBuilt	0.089035	0.089865	0.092366
BsmtFinSF1	0.074040	0.074702	0.074310
TotalBsmtSF	0.377031	0.371487	0.354552
2ndFlrSF	0.179709	0.178858	0.175219
GarageArea	0.128455	0.128069	0.123508
MSZoning_FV	0.081285	0.076612	0.012144
MSZoning_RH	0.070915	0.065896	0.000000
MSZoning_RL	0.070025	0.065869	0.008197
MSZoning_RM	0.065896	0.061279	0.000000
Neighborhood_Crawfor	0.070414	0.070473	0.066601
BldgType_Duplex	-0.064239	-0.062901	-0.045865
OverallQual_Excellent	0.193387	0.193804	0.193695
OverallQual_Very Excellent	0.248952	0.248088	0.239626
OverallQual_Very Good	0.072110	0.072724	0.074569
OverallCond_Fair	-0.067881	-0.067070	-0.051118
Foundation_Slab	0.105080	0.102390	0.074695
Foundation_Wood	-0.079177	-0.075454	-0.000000
GarageType_Attchd	0.137567	0.109884	0.014668
GarageType_Basment	0.144345	0.114865	0.000000
GarageType_BuiltIn	0.148231	0.120320	0.022962
GarageType_CarPort	0.114305	0.084819	-0.000000
GarageType_Detchd	0.124558	0.096637	-0.000000
SaleType_Con	0.091806	0.083326	0.000000
SaleCondition_AdjLand	0.106288	0.098607	0.000000

We can see from above table that few coefficients have changed and some have become even 0.

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 2.- For this we can check the R2 score of both Ridge and Lasso:

- The R2 score of Ridge for train set is-
  0.8713135111736388
- And the R2 score of Ridge for test that set is –
  0.8537746139373138

DIFFERENCE = 0.175

- The R2 score of Lasso for train set is-
  0.8667258903336928
- The R2 score of Lasso for test that we got is –
  0.854459296286021

DIFFERENCE = 0.123

As we can see that the R2 score is more in Lasso compared to Ridge.
Also the difference between test and train R2 score is less in Lasso comparatively.
IN addition to that Lasso also helps in eliminating the features that are unwanted by making them equal to 0.
It seems that we will choose to apply Lasso.

Question 3-

After building the model, you realised that the five most important predictor variables in the lasso model are n ot available in the incoming data. You will now have to create another model excluding the five most importan t predictor variables. Which are the five most important predictor variables now?

Answer 3-

From table, our top 5 features in the Lasso model are-
['TotalBsmtSF', 'OverallQual_Very Excellent', 'OverallQual_Excellent', '2ndFlrSF', 'GarageArea']

When we remove these top 5 features and rebuild the model, we got the following results-

Fig: 3

```
lasso_parameters = [i for i in lasso_parameters]

cols = X_train_rfe.columns
cols = cols.insert(0,'constant')
lasso_param_list = list(zip(cols, lasso_parameters))
lasso_param_list
```

Out[89]:
```
[('constant', 0.24972948501300943),
 ('LotArea', 0.0),
 ('YearBuilt', 0.0),
 ('BsmtFinSF1', 0.0),
 ('TotalBsmtSF', 0.0),
 ('2ndFlrSF', -0.0),
 ('GarageArea', 0.0),
 ('MSZoning_FV', -0.0),
 ('MSZoning_RH', 0.0),
 ('MSZoning_RL', -0.0),
 ('MSZoning_RM', 0.0),
 ('Neighborhood_Crawfor', -0.0),
 ('BldgType_Duplex', -0.0),
 ('OverallQual_Excellent', -0.0),
 ('OverallQual_Very Excellent', 0.0),
 ('OverallQual_Very Good', -0.0),
 ('OverallCond_Fair', 0.0),
 ('Foundation_Slab', -0.0),
 ('Foundation_Wood', -0.0),
 ('GarageType_Attchd', 0.0),
 ('GarageType_Basment', -0.0)]
```

As we can see from above (Fig:3), we are getting all the coefficient values as 0.
Also, the R-squared value is 0 with RSS of 17.80 and RMSE of 0.0178.

```
For Training Set....
r2 value is: 0.0
rss value is: 17.80449794099225
rmse value is: 0.017858072157464645
```

Question 4-

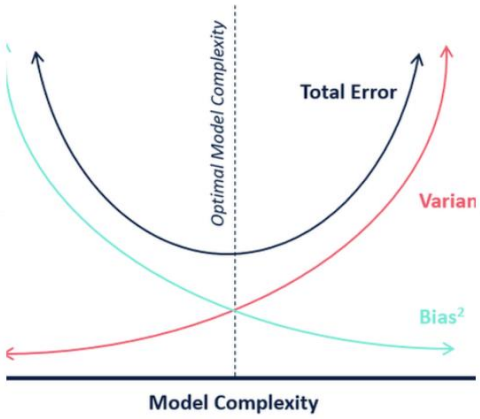
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer 4-

The following points should be considered to make sure that any model is robust and generalizable-

1. There should neither be presence of any null values and nor there should be any outliers to make sure our further predictions and interpretations is correct.
2. The model should correctly predict and should have high accuracy on the unseen data set. The testing accuracy should be generally within the 5% range of training accuracy.
3. The model should neither be very simple to under fit (which has high bias and low variance) and nor too complex to over fit (which has low bias and high variance). It should be just right that is required for the specific business application.

Fig: 4



4. Model performing very good on training set and very poor on testing set is not desirable as it is the case of overfitting and the model won't be able to generalize the results on the unseen data set.
5. Model performing is unable to learn from training data and also not able to generalise the results on testing data is also not desirable as it is the case of underfitting.