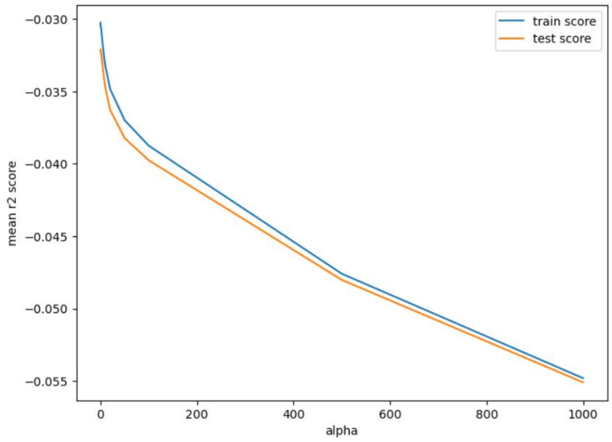
Question 1

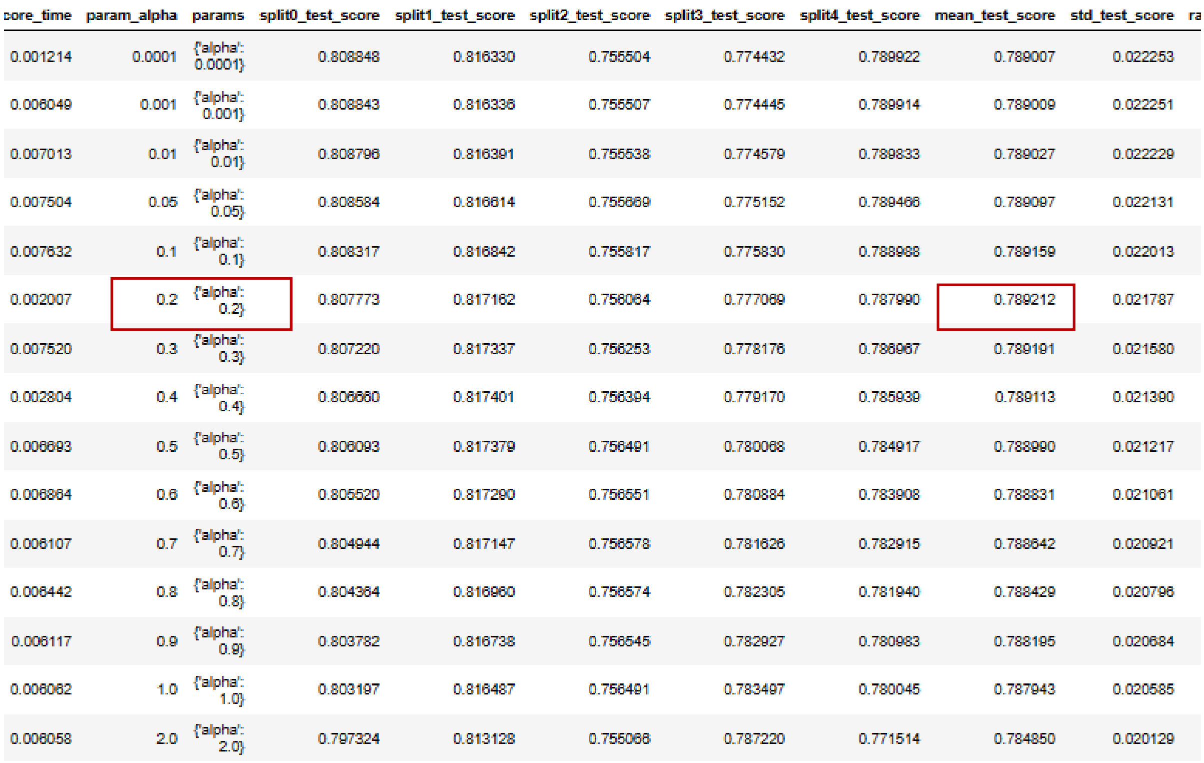
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The Optimal value of alpha for ridge and lasso regression is:

For Ridge Regression: The Optimal value of alpha for ridge is 0.2. This is based on the high R2 score for the test data as explained below.





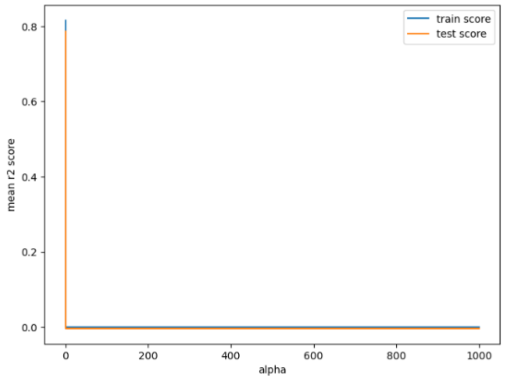
core_time	param_alpha	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score	ra
0.001214	0.0001	{'alpha': 0.0001}	0.808848	0.816330	0.755504	0.774432	0.789922	0.789007	0.022253	
0.006049	0.001	{'alpha': 0.001}	0.808843	0.816336	0.755507	0.774445	0.789914	0.789009	0.022251	
0.007013	0.01	{'alpha': 0.01}	0.808796	0.816391	0.755538	0.774579	0.789833	0.789027	0.022229	
0.007504	0.05	{'alpha': 0.05}	0.808584	0.816614	0.755669	0.775152	0.789466	0.789097	0.022131	
0.007632	0.1	{'alpha': 0.1}	0.808317	0.816842	0.755817	0.775830	0.788988	0.789159	0.022013	
0.002007	0.2	{'alpha': 0.2}	0.807773	0.817162	0.756064	0.777069	0.787990	0.789212	0.021787	
0.007520	0.3	{'alpha': 0.3}	0.807220	0.817337	0.756253	0.778176	0.786967	0.789191	0.021580	
0.002804	0.4	{'alpha': 0.4}	0.806660	0.817401	0.756394	0.779170	0.785939	0.789113	0.021390	
0.006693	0.5	{'alpha': 0.5}	0.806093	0.817379	0.756491	0.780068	0.784917	0.788990	0.021217	
0.006864	0.6	{'alpha': 0.6}	0.805520	0.817290	0.756551	0.780884	0.783908	0.788831	0.021061	
0.006107	0.7	{'alpha': 0.7}	0.804944	0.817147	0.756578	0.781626	0.782915	0.788642	0.020921	
0.006442	0.8	{'alpha': 0.8}	0.804364	0.816960	0.756574	0.782305	0.781940	0.788429	0.020796	
0.006117	0.9	{'alpha': 0.9}	0.803782	0.816738	0.756545	0.782927	0.780983	0.788195	0.020684	
0.006062	1.0	{'alpha': 1.0}	0.803197	0.816487	0.756491	0.783497	0.780045	0.787943	0.020585	
0.006058	2.0	{'alpha': 2.0}	0.797324	0.813128	0.755066	0.787220	0.771514	0.784850	0.020129	

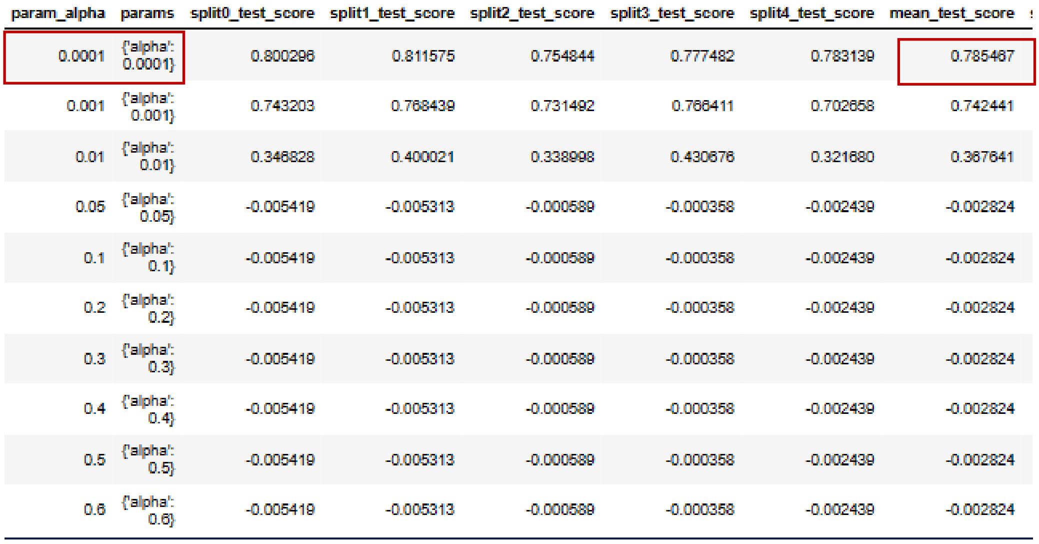
Train Score Analysis: With the increasing alpha, the R2 train score is decreasing, which means the error is increasing due to model becoming less overfitting and more generalised. Thus , by increasing the alpha the model becomes more simple for the training data.

Train Score Analysis: With the increasing value of alpha , the error started decreasing and it reached to a peak at alpha=2. But after alpha > 0.2 the R2 test score is decreasing, which means the error is increasing due to model becoming less overfitting and more generalised. Thus , by increasing the alpha >0.2 the model becomes more simple for the test data.

Selecting the optimal alpha value when the test score peaks up and in this case it is for alpha=2. So, for the optimum alpha = 2, we will have a right balance between the error and the generalisation of the model for creating a simpler model.

For Lasso Regression: The Optimal value of alpha for Lasso is 0.0001. This is based on the high R2 score for the test data as explained below.





param_alpha	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score
0.0001	{'alpha': 0.0001}	0.800296	0.811575	0.754844	0.777482	0.783139	0.785467
0.001	{'alpha': 0.001}	0.743203	0.768439	0.731492	0.766411	0.702658	0.742441
0.01	{'alpha': 0.01}	0.346828	0.400021	0.338998	0.430676	0.321680	0.367641
0.05	{'alpha': 0.05}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824
0.1	{'alpha': 0.1}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824
0.2	{'alpha': 0.2}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824
0.3	{'alpha': 0.3}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824
0.4	{'alpha': 0.4}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824
0.5	{'alpha': 0.5}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824
0.6	{'alpha': 0.6}	-0.005419	-0.005313	-0.000589	-0.000358	-0.002439	-0.002824

Train and Test Score Analysis: With the increasing value of alpha, the R2 train score is decreasing drastically, which means the error is increasing due to model becoming less overfitting and more generalised.

Train Score Analysis: With the increasing value of alpha, the R2 test score started decreasing drastically, which means the error is increasing due to model becoming less overfitting and more generalised.

Selecting the optimal alpha value when the test score peaks up and in this case it is for alpha=0.0001. So, for the optimum alpha = 0.0001, we will have a right balance between the error and the generalisation of the model for creating a simpler model.

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso:

For Ridge: The R2 Score for Test is increased from 0.7987% to 0.7995%.

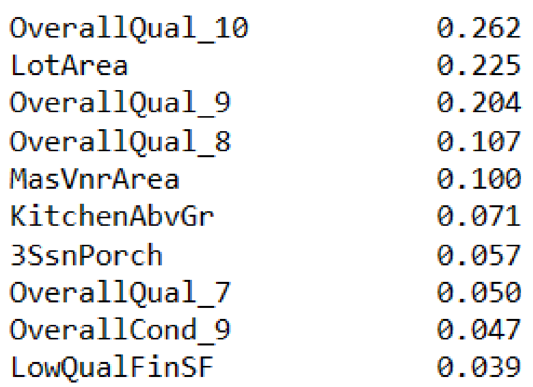
For Lasso: The R2 Score for Test is decreased from 0.8006% to 0.7925%.

The most important predictor variables after the change is implemented:

Important features remained same after the alpha is changed, but the coefficient values are changed as shown below.

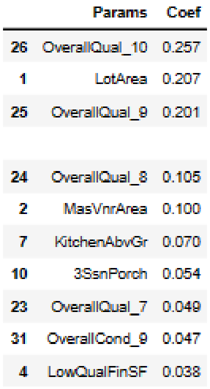
For Ridge:

With Alpha = 0.2



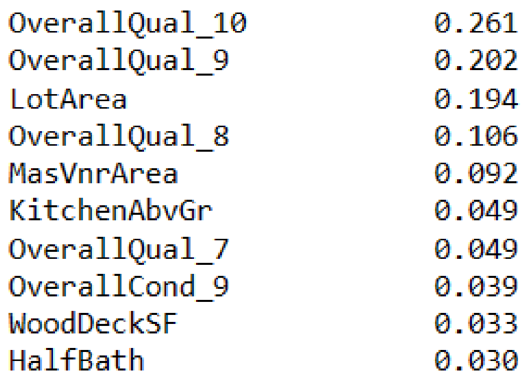
```
OverallQual_10      0.262
LotArea             0.225
OverallQual_9       0.204
OverallQual_8       0.107
MasVnrArea          0.100
KitchenAbvGr        0.071
3SsnPorch           0.057
OverallQual_7       0.050
OverallCond_9       0.047
LowQualFinSF        0.039
```

With Alpha = 0.4



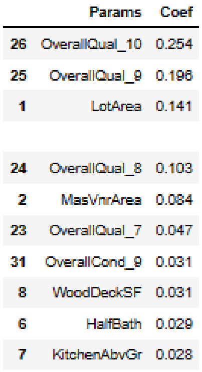
	Params	Coef
26	OverallQual_10	0.257
1	LotArea	0.207
25	OverallQual_9	0.201
24	OverallQual_8	0.105
2	MasVnrArea	0.100
7	KitchenAbvGr	0.070
10	3SsnPorch	0.054
23	OverallQual_7	0.049
31	OverallCond_9	0.047
4	LowQualFinSF	0.038

With Alpha = 0.0001



```
OverallQual_10      0.261
OverallQual_9       0.202
LotArea             0.194
OverallQual_8       0.106
MasVnrArea          0.092
KitchenAbvGr        0.049
OverallQual_7       0.049
OverallCond_9       0.039
WoodDeckSF          0.033
HalfBath            0.030
```

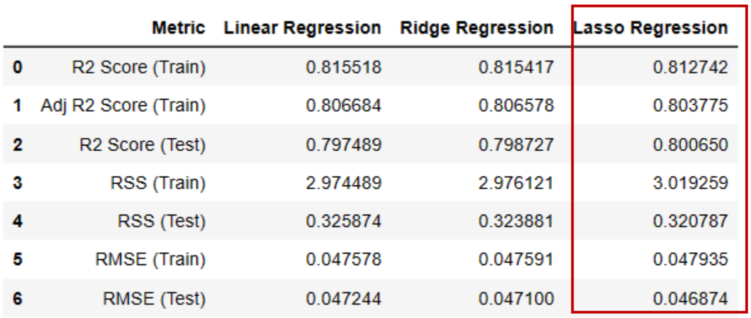
With Alpha = 0.0002



	Params	Coef
26	OverallQual_10	0.254
25	OverallQual_9	0.196
1	LotArea	0.141
24	OverallQual_8	0.103
2	MasVnrArea	0.084
23	OverallQual_7	0.047
31	OverallCond_9	0.031
8	WoodDeckSF	0.031
6	HalfBath	0.029
7	KitchenAbvGr	0.028

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?

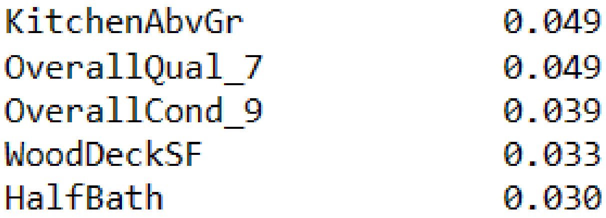
Ans: Based on the best estimated Alpha from the assignment, I chose to go with the Lasso Regression, as the Test R2 score and Root Mean Square is better for Lasso when compared to Linea r and Ridge Regression as shown below.



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.815518	0.815417	0.812742
1	Adj R2 Score (Train)	0.806684	0.806578	0.803775
2	R2 Score (Test)	0.797489	0.798727	0.800650
3	RSS (Train)	2.974489	2.976121	3.019259
4	RSS (Test)	0.325874	0.323881	0.320787
5	RMSE (Train)	0.047578	0.047591	0.047935
6	RMSE (Test)	0.047244	0.047100	0.046874

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?

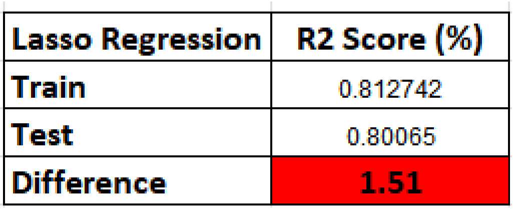
Ans: The next few important variables that I choose to go with are shown below with the corresponding coefficient values.



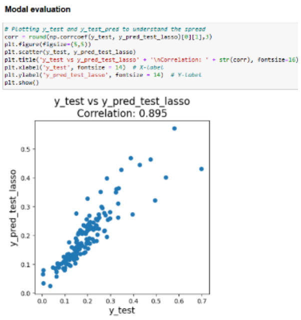
KitchenAbvGr	0.049
OverallQual_7	0.049
OverallCond_9	0.039
WoodDeckSF	0.033
HalfBath	0.030

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?

Ans: The Lasso Regression model is considered robust and can be considered for any inferences based on the validation of the below mentioned assumptions and difference between train to test R2 score which is well within 5%, confirming No model overfitting.



Lasso Regression	R2 Score (%)
Train	0.812742
Test	0.80065
Difference	1.51

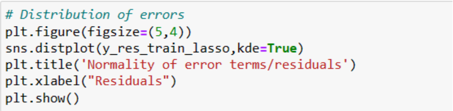


Modal evaluation

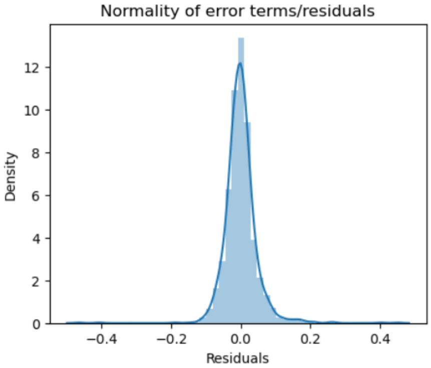
```
# Plotting y_test and y_test_pred to understand the spread
corr = round(np.corrcoef(y_test, y_pred_test_lasso)[0][1],3)
plt.figure(figsize=(5,5))
plt.scatter(y_test, y_pred_test_lasso)
plt.title('y_test vs y_pred_test_lasso' + '\nCorrelation: ' + str(corr), fontsize=16)
plt.xlabel('y_test', fontsize = 14)  # X-Label
plt.ylabel('y_pred_test_lasso', fontsize = 14)  # Y-Label
plt.show()
```

(1) *Normality assumption*: As shown below, the error terms, ε(i), are normally distributed. Implying that the If the model is able to explain the relation in the data without losing randomness.

(2) *Zero mean assumption*: Residuals have a zero mean value i.e., the error terms are normally distributed around zero.

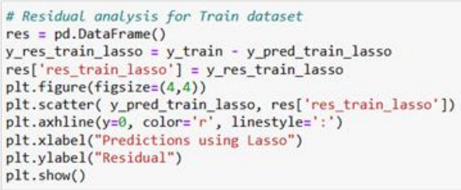


```
# Distribution of errors
plt.figure(figsize=(5,4))
sns.distplot(y_res_train_lasso,kde=True)
plt.title('Normality of error terms/residuals')
plt.xlabel("Residuals")
plt.show()
```

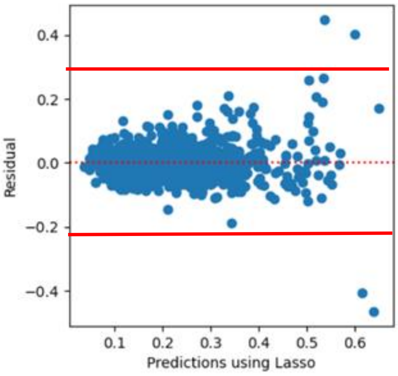


(3) Constant variance assumption: Residual terms have the same (but unknown) variance, σ2 with average value around zero.

(4) Independent error assumption: Residual terms are independent of each other, i.e., their pair-wise co-variance is zero. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves.



```
# Residual analysis for Train dataset
res = pd.DataFrame()
y_res_train_lasso = y_train - y_pred_train_lasso
res['res_train_lasso'] = y_res_train_lasso
plt.figure(figsize=(4,4))
plt.scatter( y_pred_train_lasso, res['res_train_lasso'])
plt.axhline(y=0, color='r', linestyle=':')
plt.xlabel("Predictions using Lasso")
plt.ylabel("Residual")
plt.show()
```


Residual
Predictions using Lasso