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 to 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 regression is: 0.3
The optimal value of alpha for Lasso Regression is:0.0001
When the alpha value was doubled, on both the Ridge and Lasso Regression, the R2 score, RSS and MSE do not show any major differences.
The coefficients were getting moved close to zero. The negatively correlated coefficients moved in the positive direction on the number scale whereas the positively correlated features moved towards the negative direction. I am stating the direction of movement on the number scale and not the values changing from positive to negative or vice versa.

The table is given for reference:
Ridge Regression:

Features	**Alpha-0.1**	**Alpha-0.2**	**Alpha 0.1 - Alpha 0.2**
LotArea	0.190602	0.18204001	0.008562
1stFlrSF	0.135524	0.13840736	-0.002884
GrLivArea	0.242058	0.24124553	0.000812
OverallQual_Rank_3	-0.27267	-0.26434607	-0.008321
OverallQual_Rank_4	-0.2463	-0.24281258	-0.003485
OverallQual_Rank_5	-0.22949	-0.22645482	-0.003036
OverallQual_Rank_6	-0.20858	-0.2055565	-0.003027
OverallQual_Rank_7	-0.16926	-0.16628139	-0.002974
OverallQual_Rank_8	-0.09482	-0.09205688	-0.002764
Exterior1stBrkComm	-0.11348	-0.10427901	-0.009205
FullBath3	0.057087	-0.05792842	0.115015
GarageTypeAttchd	0.043925	0.04398151	-0.000057
GarageTypeBuiltIn	0.056177	0.05664127	-0.000464
GarageTypeDetchd	0.021239	0.02105105	0.000188

Lasso Regression:

Features	Alpha - 0.0001	Alpha - 0.0002	Alpha 0.0001 - Alpha 0.0002
LotArea	0.15386063	0.10712189	0.04673874
1stFlrSF	0.13731982	0.14270954	-0.00538972
GrLivArea	0.24706806	0.25124	-0.00417194
OverallQual_Rank_3	-0.24906407	-0.21642956	-0.03263451
OverallQual_Rank_4	-0.23541223	-0.22071909	-0.01469314
OverallQual_Rank_5	-0.21942594	-0.20580842	-0.01361752
OverallQual_Rank_6	-0.19859441	-0.18535898	-0.01323543
OverallQual_Rank_7	-0.15912627	-0.14583578	-0.01329049
OverallQual_Rank_8	-0.08486821	-0.07202864	-0.01283957
Exterior1stBrkComm	-0.028227	0	-0.028227
FullBath3	0.05602652	0.0558404	0.00018612
GarageTypeAttchd	0.03747531	0.03126287	0.00621244
GarageTypeBuiltIn	0.04956403	0.04359442	0.00596961
GarageTypeDetchd	0.01373459	0.00603018	0.00770441

Ground Floor Living Area (GrLivArea: Above grade (ground) living area square feet) is the most important predictor variable after making the change.

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:

	Linear	Ridge	Lasso
LotArea	0.173882	0.190602	0.153861
1stFlrSF	0.269374	0.135524	0.137320
GrLivArea	0.391603	0.242058	0.247068
OverallQual_Rank_3	-0.119447	-0.272667	-0.249064

OverallQual_Rank_4	-0.091068	-0.246298	-0.235412
OverallQual_Rank_5	-0.093244	-0.229491	-0.219426
OverallQual_Rank_6	-0.078962	-0.208584	-0.198594
OverallQual_Rank_7	-0.053212	-0.169256	-0.159126
OverallQual_Rank_8	0.005539	-0.094821	-0.084868
Exterior1stBrkComm	-0.136025	-0.113484	-0.028227
FullBath3	0.062346	0.057087	0.056027
GarageTypeAttchd	0.148852	0.043925	0.037475
GarageTypeBuiltIn	0.167515	0.056177	0.049564
GarageTypeDetchd	0.125664	0.021239	0.013735

Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.691814	0.775522	0.773494
1	R2 Score (Test)	0.701456	0.775205	0.775418
2	RSS (Train)	3.606666	2.627045	2.650776
3	RSS (Test)	1.465500	1.103482	1.102434
4	MSE (Train)	0.061009	0.052068	0.052303
5	MSE (Test)	0.059353	0.051503	0.051479

For the model, I chose, The Ridge and Lasso has performed the same way with respect to Explain ability of the model. From the coefficient's standpoint, the positive and negative correlations are matching between Ridge and Lasso. The coefficient values are mostly matching for the features just that Ridge predicts the correlation of LoTArea and Built In Garage type to be stronger than what Lasso has predicted. So I feel I would go with Ridge Regression.

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?

Answer:

From the Lasso Model, below are the five most important predictor variables ordered from lowest to highest.

GarageTypeBuiltIn

FullBath3

1stFlrSF

LotArea

GrLivArea

I removed the above five features and rerun the model with Lasso Regression. Here are the new features and coefficients.

Features	Coefficients

TotalBsmtSF	0.270843
NeighborhoodNoRidge	0.074316
Exterior1stStone	0.010829
Exterior1stImStucc	0
BedroomAbvGr5	0
Exterior1stBrkComm	-0.01607
FullBath2	-0.08668
OverallQual_Rank_8'	-0.09853
FullBath1	-0.11924
OverallQual_Rank_7	-0.17434
OverallQual_Rank_6	-0.2131
OverallQual_Rank_5	-0.23748
OverallQual_Rank_4	-0.26184
OverallQual_Rank_3	-0.29125

As per this Total Basement Square Feet, North Ridge in the Neighborhood, Exterior covered with Stone have positive correlation.

Exterior and the 1st floor covered with common brick and Full Bathrooms above Grade 2 are having strong negative correlation with Sale Price.

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:

A model should be able to maintain the accuracy even when it is run on unseen data. It should remain robust so that even if there are changes in one or more input variables, the forecasts or predictions made by the model should be consistently accurate.

We should handle overfitting issue of the model through Regularization. For eg., Lasso or Ridge Regression so that the model is robust and more generalizable.

The model will be accurate only if it is able to handle overfitting and any drastic changes in input variables due to unforeseen circumstances. Hence keeping the model robust and generalizable implies the model is accurate or accuracy of the model is consistent.