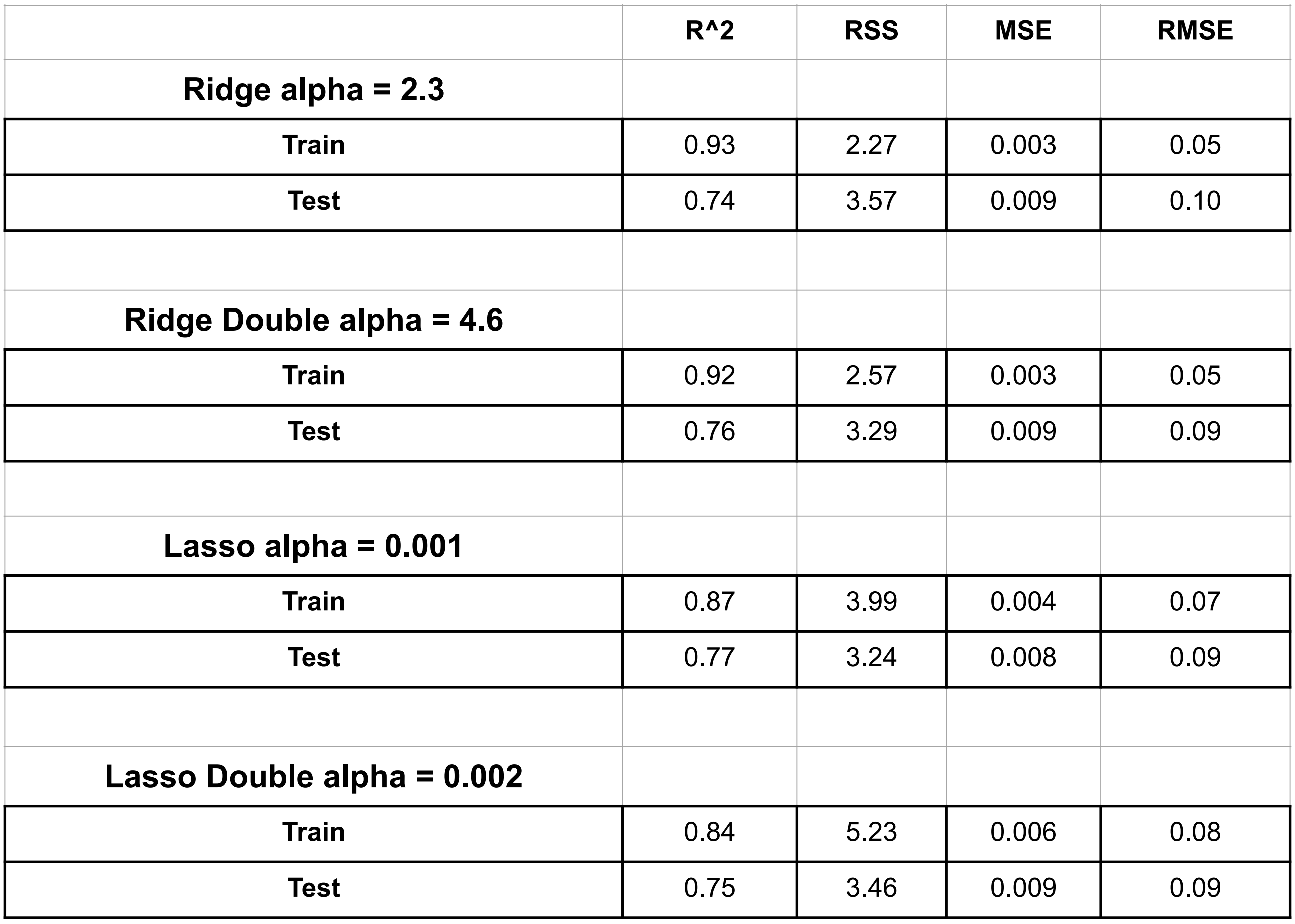
Question-1:

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:

a) Optimal value of Ridge alpha = 2.3 and Lasso alpha = 0.001

b) After doubling Alpha the Model changes are as in the table below:



	R^2	RSS	MSE	RMSE
Ridge alpha = 2.3				
Train	0.93	2.27	0.003	0.05
Test	0.74	3.57	0.009	0.10
Ridge Double alpha = 4.6				
Train	0.92	2.57	0.003	0.05
Test	0.76	3.29	0.009	0.09
Lasso alpha = 0.001				
Train	0.87	3.99	0.004	0.07
Test	0.77	3.24	0.008	0.09
Lasso Double alpha = 0.002				
Train	0.84	5.23	0.006	0.08
Test	0.75	3.46	0.009	0.09

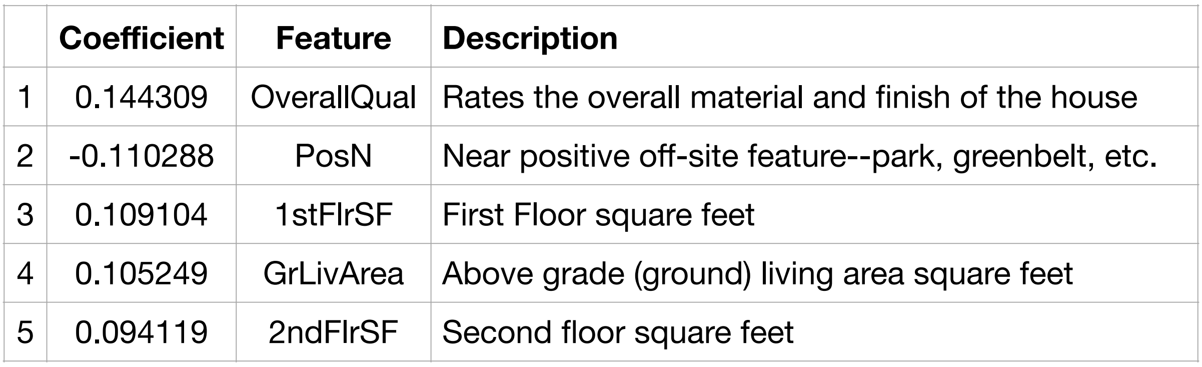
After doubling in general there is more regularisation in Ridge as well as Lasso, as expected the coefficients are further reduced towards zero. Also the models have moved towards more bias and less variance in general due to reduction in complexity that is reduction in features.

c) The models important predictor variables for optimal alpha and after doubling the alpha are;

Ridge important predictor variables for optimal alpha:

	Coefficient	**Feature**	**Description**
1	-0.196204	PosN	Near positive off-site feature--park, greenbelt, etc.
2	0.164374	OverallQual	Rates the overall material and finish of the house
3	0.129407	1stFlrSF	First Floor square feet
4	0.126207	GrLivArea	Above grade (ground) living area square feet
5	0.114766	2ndFlrSF	Second floor square feet

Ridge important predictor variables after doubling alpha:



	Coefficient	Feature	Description
1	0.144309	OverallQual	Rates the overall material and finish of the house
2	-0.110288	PosN	Near positive off-site feature--park, greenbelt, etc.
3	0.109104	1stFlrSF	First Floor square feet
4	0.105249	GrLivArea	Above grade (ground) living area square feet
5	0.094119	2ndFlrSF	Second floor square feet

Lasso important predictor variables for optimal alpha:

	Coefficient	Feature	Description
1	0.263613	OverallQual	Rates the overall material and finish of the house
2	0.226042	GrLivArea	Above grade (ground) living area square feet
3	0.099747	TotalBsmtSF	Total square feet of basement area
4	0.093736	GarageArea	Size of garage in square feet
5	0.066573	TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)

Lasso important predictor variables after doubling alpha:

	Coefficient	**Feature**	**Description**
1	0.251120	OverallQual	Rates the overall material and finish of the house
2	0.240555	GrLivArea	Above grade (ground) living area square feet
3	0.079671	TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
4	0.077119	GarageArea	Size of garage in square feet
5	0.049733	GarageCars	Size of garage in car capacity

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:

I will choose Lasso because there is a huge reduction in number of features (Ridge has 238 variables and Lasso has 54 non-zero coefficients) which translates to huge reduction in complexity/Variance. Lasso is simpler model.

The difference between R^2 of Train and Test data is less in case of lasso regression. The metrics data shows that though Lasso accuracy is low on training data it is more generalisable and performs better on unseen data.

Ridge alpha = 2.3	**R^2**
Train	0.93
Test	0.74
Lasso alpha = 0.001	
Train	0.87
Test	0.77

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?

Answer:

The 5 most important predictor variables after rebuilding the lasso model.

	Coefficient	Feature	Description
1	0.406979	1stFlrSF	First Floor square feet
2	0.224993	2ndFlrSF	Second floor square feet
3	0.103522	GarageCars	Size of garage in car capacity
4	0.082820	NridgHt	Location in Northridge Heights Neighborhood
5	0.074936	Somerst	Location in Somerset Neighborhood

Lasso alpha = 0.001	R^2	RSS	MSE	RMSE
Train	0.86	4.47	0.00	0.07
Test	0.76	3.36	0.01	0.09

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?

Answer:

A model is robust if it can handle noise, variations, and uncertainties in the input data like our lasso model which performs well even after the 5 important features were not available. A model is generalisable if it can make accurate predictions on data it has not seen during training.

We can make sure that model is robust and generalisable by;

1) Cleaning data: Removing noisy data and outliers will help the accuracy and prevent biased model.
2) Creating test set: By setting aside a portion of available data for testing the model helps evaluate how well the model performs on unseen data. A large difference in the accuracies of test and train data indicates that the model is overfit.
3) Regularisation: Regularising methods like Ridge and Lasso adds penalties to the complexity of models encouraging simple models. It prevents overfitting. It reduces the variance helping the model generalise.
4) Cross Validation: It helps identify how well the model generalise on different subsets of the data preventing the model overfitting to a specific set.
5) Hyperparameter tuning: Tuning hyperparameters by using techniques like grid search helps find the optimal generalisable model which is a balance of bias and variance.

Robust and generalisable models have implications on the accuracy of model. Trained models can have high accuracy on the training data but may not perform so well on unseen test data which indicates overfitting, that means the model is not robust/generalisable.

Implementing the above methods to make model robust and generalisable can slightly decrease the accuracy on training data but it will make the model have good accuracy consistently on training data as well as unseen data. We need to strike balance between fitting on train data and generalising on unseen data.