Assignment Based Subjective Questions

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 optimal values of alpha for ridge and lasso regression are 1.55 and 0.001 respectively. After doubling the value of alpha for both ridge and lasso our R2 Score remains same for both train and test which is approx 0.88 (R2 is same for both lasso and ridge). For the exact difference on top features can be seen on the below image provided.

Best Lasso

	Feature	Coef
45	YearBuilt_Updated	0.555471
30	KitchenQual	0.512864
28	BedroomAbvGr	0.401170
78	Neighborhood_Sawyer	0.385873
14	BsmtFinType2	0.370103
27	HalfBath	0.359513
23	GrLivArea	0.218707
69	Neighborhood_MeadowV	0.191489
2	LotShape	0.188181
38	WoodDeckSF	0.182314

Doubled Lasso

	Feature	Coef
30	KitchenQual	0.443736
14	BsmtFinType2	0.367227
78	Neighborhood_Sawyer	0.360468
28	BedroomAbvGr	0.351406
27	HalfBath	0.315973
45	YearBuilt_Updated	0.308491
2	LotShape	0.195525
23	GrLivArea	0.187727
69	Neighborhood_MeadowV	0.172243
20	1stFlrSF	0.142357

Best Ridge

	Feature	Coef
45	YearBuilt_Updated	0.622993
30	KitchenQual	0.570810
28	BedroomAbvGr	0.437022
27	HalfBath	0.397470
78	Neighborhood_Sawyer	0.372305
16	BsmtUnfSF	0.323408
35	GarageArea	0.301067
69	Neighborhood_MeadowV	0.275385
49	MSZoning_FV	0.271276
38	WoodDeckSF	0.254848

Doubled Ridge

	Feature	Coef
30	KitchenQual	0.521498
45	YearBuilt_Updated	0.501936
28	BedroomAbvGr	0.407814
27	HalfBath	0.377407
78	Neighborhood_Sawyer	0.356680
16	BsmtUnfSF	0.251057
69	Neighborhood_MeadowV	0.249619
23	GrLivArea	0.225322
38	WoodDeckSF	0.224817
14	BsmtFinType2	0.210702

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?

The optimal values of lambda for ridge and lasso regression are 1.55 and 0.001 respectively. And R2 Score is also the same for both which is 0.88 (For both train and test).
As we know lasso regression will penalize more and help in feature elimination also, so my go to model will be that only.

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?

Our five most important predictor variables in the lasso model were-

- YearBuilt_Updated
- KitchenQual
- BedroomAbvGr
- Neighborhood_Sawyer
- BsmtFinType2

After building another lasso model and removing these 5 features. Now we have the follow top five important predictor variables as -

- HalfBath
- GrLivArea
- Neighborhood_MeadowV
- LotShape
- WoodDeckSF

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?

When two models perform the same on the train and test data, we should choose the one which gives fewer predictions on test data because they are usually simple models means more generic, applicable to many scenarios. They will need less data to train hence they are not complex and are more generalized. Simple models will have low variance and high bias also.Complex model tends to overfit also , that's not the case in simple models.

Hence the model should be generic and robust but not simple enough that it doesn't learn anything.
Good model is one which is not overfitting and underfitting during training. And We have to find the good trade- off in bias and variance. Further we can reduce the complexity of a model using regularization also.