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?

Ans:

Optimal value of alpha for Ridge: 1.0

Optimal value of alpha for Lasso: 0.001

If you choose to double the value of alpha for both Ridge and Lasso, then there will will be a change in R2 score, RSS and MSE.

I used alpha values 2.0 and 0.002 respectively for Ridge and Lasso and observed a change in R2 score, RSS and MSE.

Metrics for Ridge and Lasso with optimal/original alpha values:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.391512e-01	0.938177	0.930310
1	R2 Score (Test)	-1.225865e+15	0.908864	0.912556
2	RSS (Train)	5.744127e+01	58.360993	65.786903
3	RSS (Test)	4.696814e+17	34.917981	33.503667
4	MSE (Train)	2.466755e-01	0.248642	0.263988
5	MSE (Test)	3.405447e+07	0.293628	0.287620

Metrics for Ridge and Lasso with double the alpha value:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.391512e-01	0.937200	0.925001
1	R2 Score (Test)	-1.225865e+15	0.909955	0.909506
2	RSS (Train)	5.744127e+01	59.283424	70.799133
3	RSS (Test)	4.696814e+17	34.500198	34.672102
4	MSE (Train)	2.466755e-01	0.250600	0.273860
5	MSE (Test)	3.405447e+07	0.291866	0.292592

In Lasso, the most important predictor variables are:-

1. OverallQual_Excellent
2. OverallQual_Very Excellent
3. GrLivArea
4. SaleCondition_Partial
5. OverallQual_Very Good

In Ridge, the most important predictor variables:

1. OverallQual_Excellent
2. OverallQual_Very Excellent
3. Functional_Typ
4. OverallQual_Very Good
5. SaleType_CWD

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:
By using the optimal values of alpha for ridge and lasso, we observed that Lasso is performing better because it has high r2_score value compared to ridge model. Also, MSE in test/unseen data is less compared to Ridge model.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.391512e-01	0.938177	0.930310
1	R2 Score (Test)	-1.225865e+15	0.908864	0.912556
2	RSS (Train)	5.744127e+01	58.360993	65.786903
3	RSS (Test)	4.696814e+17	34.917981	33.503667
4	MSE (Train)	2.466755e-01	0.248642	0.263988
5	MSE (Test)	3.405447e+07	0.293628	0.287620

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:

Five most important variables based on the coefficient values are→

For optimal alpha in original model:

1. OverallQual_Excellent
2. OverallQual_Very Excellent
3. SaleCondition_Partial
4. GrLivArea
5. OverallQual_Very Good

For model after removing 5 features:

1. SaleType_CWD
2. Neighborhood_NridgHt
3. SaleType_New
4. 1stFlrSF
5. Functional_Typ

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:

We cannot trust a model for predictive analysis, which is not robust. We have to make a model robust and generalizable to ensure that the model is not impacted by the outliers present in the training data. In addition to that we have to make sure that the model is generalizable so that the difference between train and test accuracy is not more. Also, the model should not be confined to only those data sets which we have used during the training it should also be accurate to unseen data.

Let's discuss about the accuracy of the model. We need to handle the outliers properly. We should not give a lot of weightages to the outliers, this will impact the accuracy predicted by the model. Hence, we have to only consider those outliers which are relevant to the model and all those irrelevant outliers need to be removed from the data set. This will help in increasing accuracy of the model. In addition to that in order to help standardize the prediction made by the model we can also use confidence intervals usually 3 standard deviation.