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 1:

For the **ridge** regression algorithm, I have used GridSearchCV model, *which* allowed us to automatically perform the 5-fold cross-validation to find the optimal value of alpha. In our case the optimal value of alpha is **2**.

To find the optimal value of alpha, I have used **lasso** linear model with iterative fitting along a regularization path. The best model is selected by cross-validation. In our case the optimal value of alpha is **.0001**.

When we double the value of alpha for our ridge regression, we will take the value of alpha equal to 4 the model will apply more penalty. When alpha is **4** some coefficient value also changed. Similarly, when we increase the value of alpha for lasso **(.0002)** we try to penalize more our model, when we increase the value of our r2 square also decreased. Some features and coefficient value got changed.

The most important variable after the changes has been implemented for **ridge regression** are as follows: -

1. MSSubClass
2. Neighborhood_Crawfor
3. MSZoning_FV
4. Neighborhood_NridgHt
5. Neighborhood_StoneBr
6. OverallCond
7. MSZoning_RL
8. YearRemodAdd
9. Exterior1st_BrkFace
10. Exterior1st_Stucco

The most important variable after the changes has been implemented for **lasso regression** are as follows: -

1. MSSubClass
2. BsmtFullBath
3. OverallCond
4. YearRemodAdd
5. Neighborhood_Crawfor
6. YearBuilt
7. 1stFlrSF

8. WoodDeckSF
9. BsmtFinSF2
10. d_SaleCondition

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 2:

- The optimal lambda value in case of Ridge and Lasso is as below:
  - Ridge - 2
  - Lasso - 0.0001
- The r2_score(train and test) in case of Ridge and Lasso are:
  - Ridge - 0.9258386180186315, 0.8915642098870991
  - Lasso - 0.9157976942856315, 0.8918513721348732
- The Mean Squared error in case of Ridge and Lasso are:
  - Ridge - 0.00011103633128697509
  - Lasso - 0.00011074228222398805

In Ridge, the penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

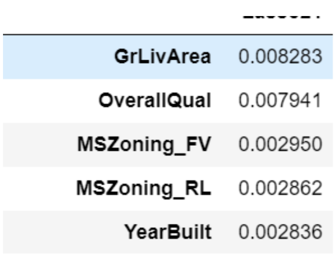
Also R2_score in case of Lasso is little less as comparison to Ridge. I will choose Ridge.

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 3:

I created another model excluding the five most important predictor variables and the following is the result.



GrLivArea	0.008283
OverallQual	0.007941
MSZoning_FV	0.002950
MSZoning_RL	0.002862
YearBuilt	0.002836

The five most predictor variables are defined below:

- GrLivArea
- OverallQual
- MSZoning_FV
- MSZoning_RL
- YearBuilt

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 4:

A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. Model should also be generalisable so that the test accuracy is not lesser than the training score. Outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. In case outliers does not make sense in the context of the analysis then those should be taken out so that future predictions are correct and accurate. If the model is not robust, it cannot be trusted for predictive analysis.