**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** : After executing RFE on top_40, top_45, top_50 and top_55, I selected top_55 being the best adj R2 of 0.913.

Optimal value of alpha for ridge : 10.0

Optimal value of alpha for ridge : 0.0001

Ridge :

```
Optimum aplha for ridge is 10.000000
ridge Regression with 10.0
==============================
R2 Score (train) :  0.9168908447027774
R2 Score (test) :  0.873014948602806
RMSE (train) :  0.11289771548276484
RMSE (test) :  0.1523521619609823
```

Lasso :

```
Optimum aplha for lasso is 0.000100
lasso Regression with 0.0001
==============================
R2 Score (train) :  0.9169731799366677
R2 Score (test) :  0.8725229295812081
RMSE (train) :  0.1128417784253933
RMSE (test) :  0.15264703009942346
```

After doubling the optimal value i.e,

Optimal value of alpha for ridge : 20.0

Optimal value of alpha for ridge : 0.0002 .Ridge and Lasso R2 square and RMSE becomes

Ridge :

```
Model evaluation : Ridge regression with alpha = 20.0
R2 Score (Train):  0.91665
R2 Score (Test):  0.87368
RMSE (Train):  0.11306
RMSE (Test):  0.15195
```

Lasso :

```
Model evaluation : Lasso regression with alpha = 0.0002
R2 Score (Train):  0.917
R2 Score (Test):  0.87193
RMSE (Train):  0.11283
RMSE (Test):  0.153
```

I have got very similar values for both Ridge and Lasso.

After changing the optimal value alpha,below are the mentioned becomes the most important feature.

- **1stFlrSF**
- **2ndFlrSF**
- **OverallQual**
- **OverallCond**
- **SaleCondition_Partial**

## 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** : After model evaluation I have got slight similar values for both Ridge and Lasso at alpha 10.0 and 0.0001 respectively.

Ridge regression doesn't zero any of the coefficient (not able to remove insignificant variable).

Lasso is better option since it helps in feature elimination by making the coefficient zero.

So I will go for Lasso.

## 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** : After removing the below five most important predictor variables in the lasso model

- **1stFlrSF**
- **2ndFlrSF**
- **OverallQual**
- **OverallCond**
- **SaleCondition_Partial**

The new top 5 predictor will be :

- **FullBath**
- **GarageArea**
- **KitchenQual**
- **Fireplaces**
- **LotArea**

## 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** : **The model is robust and generalizable when** :

1. Model works well on unseen data.
2. Test accuracy is not much lesser than the training score.
3. The model should not be affected by the outliers : Outlier treatment is the most important to get the robust model. We can detect outliers in the dataset using box plots, Z score etc.
4. The predicted variables should be significant :
   Model significance can be determined by p-value, R2 and adj. R2

   Always a simple model will be more robust.

**Implications of Accuracy of a model** :

1. Fix missing values and outliers:
   If data has missing values and outliers, it may lead to an inaccurate model. Outliers can affect the mean, median that we are imputing to a continuous variables.

2. Get more data as much as you can :
   Having more data allows model to train itself by looking at the different pattern.

3. **Feature Emgineering or newly derived columns:**
   We can always extract data from existing data eg: Using YearBuilt feature we can derive age or no of years since the house was built.
   Scaling the values: eg: One value is in meters and the other in kilo meters, it is important to scale these feature into one standardized unit.

4. **Feature selection** :
   It is purely based on the domain knowledge, which can help to select the important features that have good impact on the target variable.

5. **Apply the right algorithm**:
   Choosing the right machine learning algorithm is very important to get accurate model.

6. **Cross validation**:
   Some times more accuracy will cause overfitting then we can use cross validation technique, i.e. leave a partition of data which you do not train the model and test the model on this partition before got the final model.