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 : The optimal value for the regressions are:

1. Ridge Regression alpha value : **2.0**
    a. R2 value for training data : **0.8923**
    b. R2 value for testing data : **0.8743**
    c. MAE : **14779**
2. Lasso Regression alpha value : **0.9**
    a. R2 value for training data : **0.8994**
    b. R2 value for testing data : **0.8716**
    c. MAE : **15403**

If I double the alpha value for ridge and lasso regression then it does not make much difference in the training and testing data for ridge and lasso and also MAE approx the same.

After Double the values of alpha in Ridge and Lasso Regression:

1. Ridge Regression alpha value : **4.0**
    a. R2 value for training data : **0.8861**
    b. R2 value for testing data : **0.8708**
    c. MAE : **14820**
2. Lasso Regression alpha value : **1.8**
    a. R2 value for training data : **0.8993**
    b. R2 value for testing data : **0.8722**
    c. MAE : **15349**

These are the results I got after double the alpha value of ridge and lasso regression, though results are approximately the same, won't make much difference in the r2 score of training and testing data and for MAE also.

Important predictor variables after the change is implemented is shown below along with the coefficient value :

```
for coefficient, column_name in zip(top_5_coefficients, top_5_column_names):
    print(f"Column: {column_name}, Coefficient: {coefficient}")
```

```
Column: TotRmsAbvGrd, Coefficient: 74406.99414622923
Column: GarageType_BuiltIn, Coefficient: 60150.28658793467
Column: OverallQual_9, Coefficient: 55638.833854578195
Column: GarageType_Basment, Coefficient: 54192.61493527966
Column: GarageType_Attchd, Coefficient: 53072.966824875606
```

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 value of alpha for ridge is **2.0** and for Lasso is **0.9**, the metrics we got after building our model with ridge and lasso regression is:

1. Ridge Regression alpha value : **2.0**
    a. R2 value for training data : **0.8923**
    b. R2 value for testing data : **0.8743**
    c. MAE : **14779**
2. Lasso Regression alpha value : **0.9**
    a. R2 value for training data : **0.8994**
    b. R2 value for testing data : **0.8716**
    c. MAE : **15403**

According to the metrics that we got after building a model using ridge and lasso regression, r2 score is quite the same for training and testing data on ridge and lasso regression, but if we check more closely we have less difference between training and testing in Ridge Regression than Lasso Regression.

And also if we check the **Mean Absolute Error(MAE)** for ridge and lasso regression, the good MAE value is for Ridge regression with alpha value 2.0.

I will choose the **Ridge Regression** model according to the current scenario.

Question 3: **After building the model, you realized 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: After building the model the five most important predictor variables is as follows:

```
[810] for coefficient, column_name in zip(top_5_coefficients, top_5_column_names):
          print(f"Column: {column_name}, Coefficient: {coefficient}")

      Column: TotRmsAbvGrd, Coefficient: 74361.32825328382
      Column: GarageType_BuiltIn, Coefficient: 63122.90266549021
      Column: GarageType_Basment, Coefficient: 57114.0137616815
      Column: OverallQual_9, Coefficient: 56099.64065350058
      Column: GarageType_Attchd, Coefficient: 55930.00941867394
```

After excluding above 5 variables during training our model, we got these 5 important features listed below:

```
for coefficient, column_name in zip(top_5_coefficients, top_5_column_names):
    print(f"Column: {column_name}, Coefficient: {coefficient}")

Column: FullBath, Coefficient: 71352.9386551076
Column: GarageCars, Coefficient: 49270.312747478696
Column: Neighborhood_StoneBr, Coefficient: 32745.547961641554
Column: Exterior1st_Stone, Coefficient: 30574.468142442074
Column: Foundation_Wood, Coefficient: 29234.683399207068
```

These Screenshots I took from the code itself.

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 : To make our model robust and generalisable we have some steps that need to follow listed below:

- We have to use **Sufficient and Diverse data** for training purposes to generate good models.
- We need to perform **Proper Data Processing** for a robust and generalisable model.
- Do **Proper Feature Engineering and Avoid Overfitting**
- **Validation and Evaluation** need to perform Accurately to increase model performance.
- **Hyperparameter Tuning**
- **Test on Unseen Data.**

With our dataset I created 3 models and all of them listed below:

1. Linear Regression Model without RFE
    a. R2 value for training data : **0.9340**
    b. R2 value for testing data : **0.9092**
    c. MAE : **13479**

2. Ridge Regression with RFE
    a. R2 value for training data : **0.8923**
    b. R2 value for testing data : **0.8743**
    c. MAE : **14779**

3. Lasso Regression with RFE
    a. R2 value for training data : **0.8994**
    b. R2 value for testing data : **0.8716**
    c. MAE : **15403**

By checking above models, best model perform for this dataset is the **Linear Regression without RFE**, this model is **not overfit** and we check the **condition of overfitting** like this : **Difference between r2 value for training and testing data is not more than 5%** in this model the difference is quite **less than 5%**, hence this model is not overfit.
In the Linear Regression model the **MAE value is also quite good which is 13479** less than the other models, and works better in **Unseen data** as well.
Linear regression model works better than Ridge and Lasso Regression.