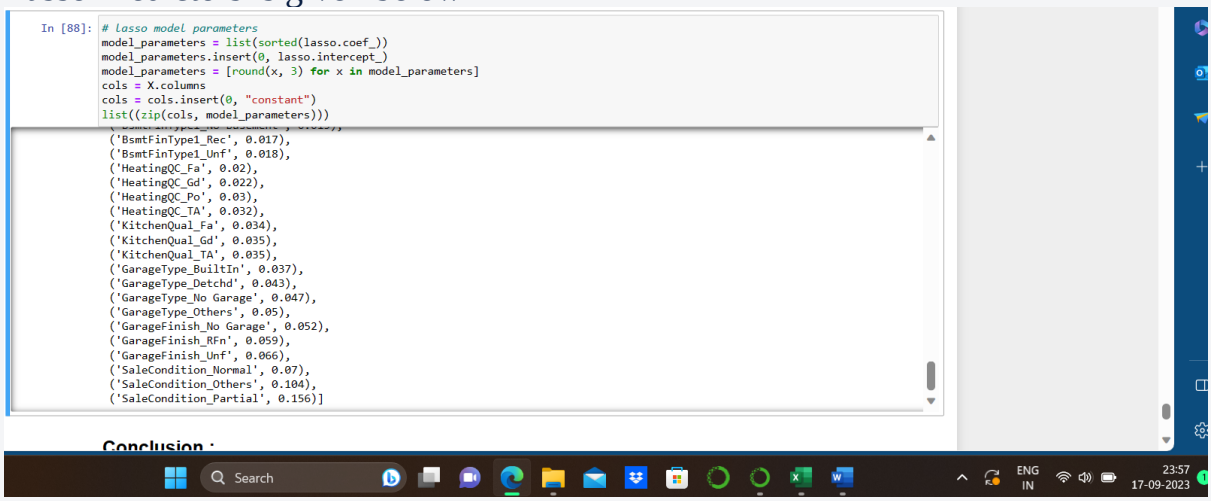
Question 1

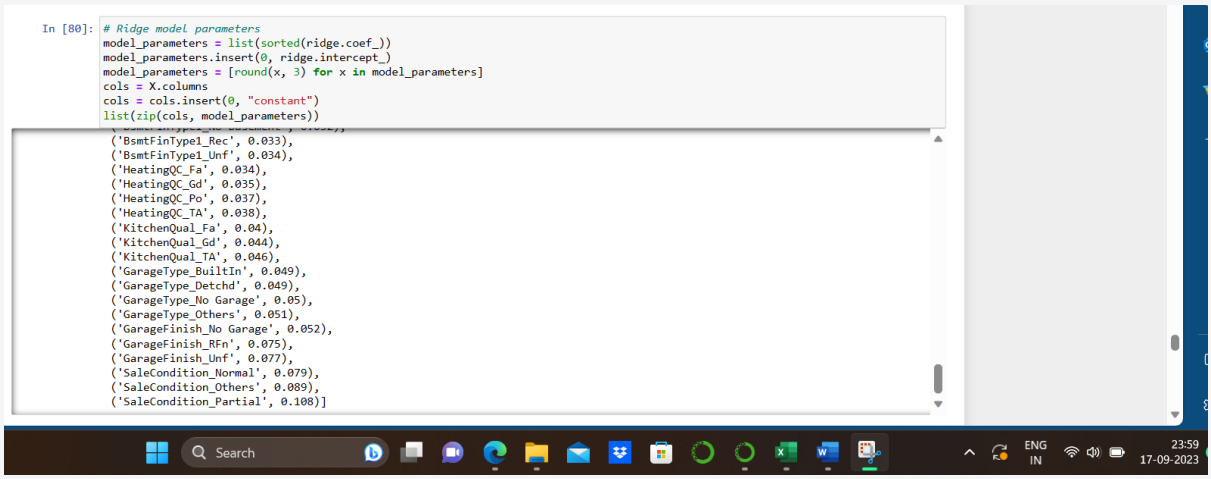
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

1. For Ridge the optimal value chosen is 10
2. For Lasso the optimal value chosen is 0.001

Value of alpha if doubled, we need to choose 20 for Ridge and 0.002 for Lasso.

Lasso Predictors is given below

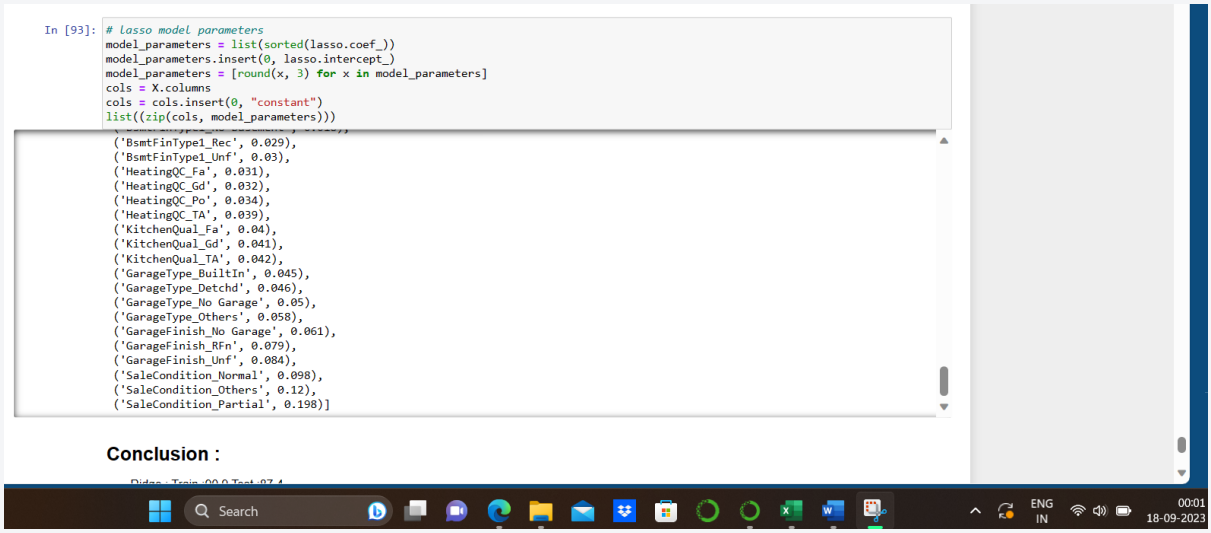




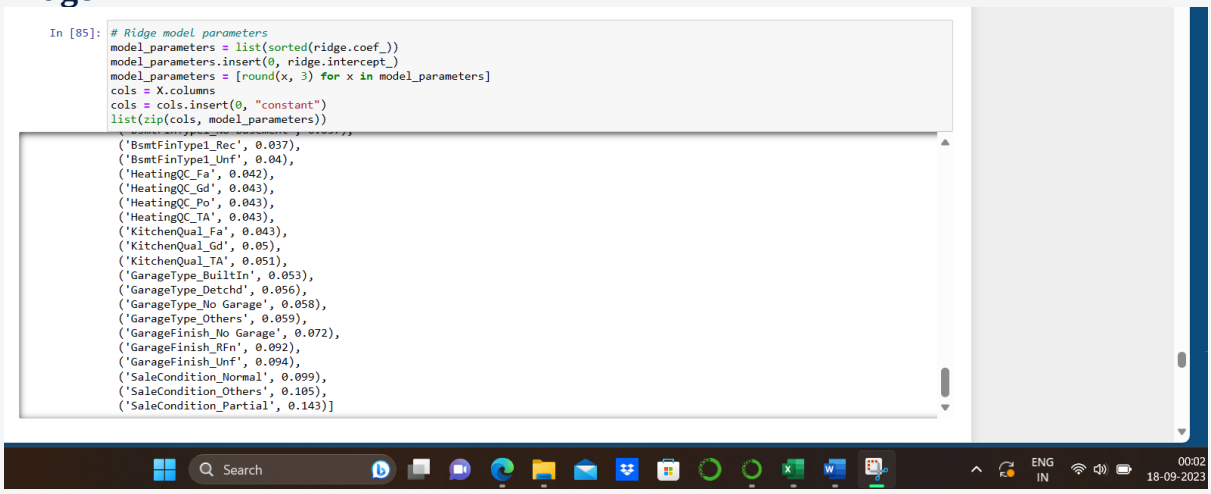
In addition, in case of ridge the coefficients are very low and in case of Lasso coefficients is turning 0.

Before change, the values are described below:

Lasso



Ridge



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 obtained are given below:

1. For Ridge the optimal value chosen is 10
2. For Lasso the optimal value chosen is 0.001

Lasso Score we got is

```
In [91]: lasso.score(X_train,y_train)
Out[91]: 0.898288939025357

In [92]: lasso.score(X_test,y_test)
Out[92]: 0.8646575331441892
```

Ridge score we got is

```
In [83]: ridge.score(X_train,y_train)
Out[83]: 0.9092068605070023

In [84]: ridge.score(X_test,y_test)
Out[84]: 0.8744204967072813
```

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?

We will now remove the following:

Lasso

Top 5 most significant variables in Lasso are:

- 1 SaleCondition_Partial - 0.198
- 2 SaleCondition_Others - 0.12

- 3 SaleCondition_Normal - 0.098
- 4 GarageFinish_Unf - 0.084
- 5 GarageFinish_RFn - 0.079

Post removing the above variables, the top five variables now are as given below

```
'GarageFinish_No Garage', 0.193)
'GarageType_Others', 0.118),
'GarageType_No Garage', 0.093),
'GarageType_Detchd', 0.086),
'GarageType_BuiltIn', 0.084),
```

lasso.score(X_train,y_train)

0.8971962713506288

lasso.score(X_test,y_test)

0.8621945956996743

Ridge

Top 5 most significant variables in Ridge are:

- 1 SaleCondition_Partial - 0.143
- 2 SaleCondition_Others - 0.105
- 3 SaleCondition_Normal - 0.099
- 4 GarageFinish_Unf - 0.094
- 5 GarageFinish_RFn - 0.092

Post removing the above variables, the top five variables now are as given below

```
('GarageFinish_No Garage', 0.141)]
('GarageType_Others', 0.103),
('GarageType_No Garage', 0.097),
('GarageType_Detchd', 0.094),
('GarageType_BuiltIn', 0.09),
```

ridge.score(X_train,y_train)

0.9074415548618695

ridge.score(X_test,y_test)

0.8714690856626222

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?

1. We can tweak the values of alpha and try with multiple combinations to evaluate the scores of both ridge and lasso.
2. Check the correlation chart and evaluate if we can drop some more columns and then build the model.
3. For different variables with numerical values, we can try to normalize or apply log values and build the model. There are multiple outliers in numerical variables, those can be improved with this approach.