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: There is no standard optimal value for alpha in lasso and ridge regression. The best practice for this will be to obtain alpha value using Gridsearch cross validation by passing a parameter grid. This can either be integers or logspace. For better accuracy, it is recommended to use logspace.

As shown in the syntax below, this will perform the regression based on the estimator(ridge or lasso based on the algorithm used) and return the model.

```
folds = 5

model_cv = GridSearchCV(estimator = ridge/lasso,

                    param_grid = params,

                    scoring= 'neg_mean_absolute_error',

                    cv = folds,

                    return_train_score=True,

                    verbose = 1)
```

model_cv.best_params_ will then return the optimal value of alpha based on the data.

If Alpha is doubled under Ridge regression, bias of the model will increase and variance will be reduced and as a result, multicollinearity is handled much more aggressively

If Alpha is doubled under Lasso regression, more coefficients are driven towards zero. This will cause the model to be sparse and it will perform feature selection aggressively. In the end, only dominant predictors will survive.

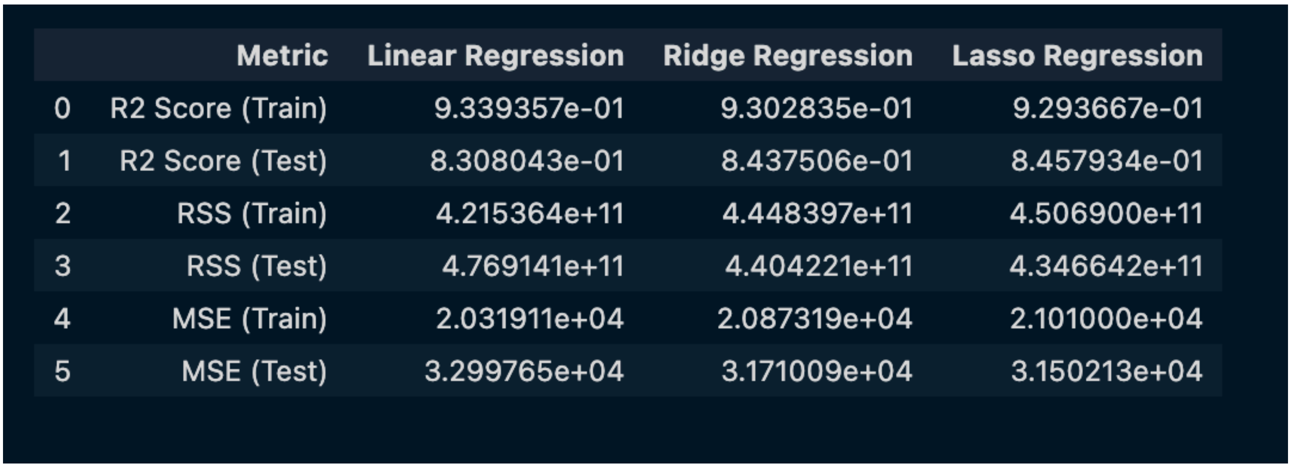
Even for lasso, bias will increase and variance will reduce

The most important predictor variables will be the variables that exhibit high predictive strength, low multicollinearity, and stable coefficient estimates across samples.

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: Based on the assignment, I will chose lasso as it performed the feature selection and in the final output, lasso regression had the least difference between R2 for test and train. Following is the result matrix



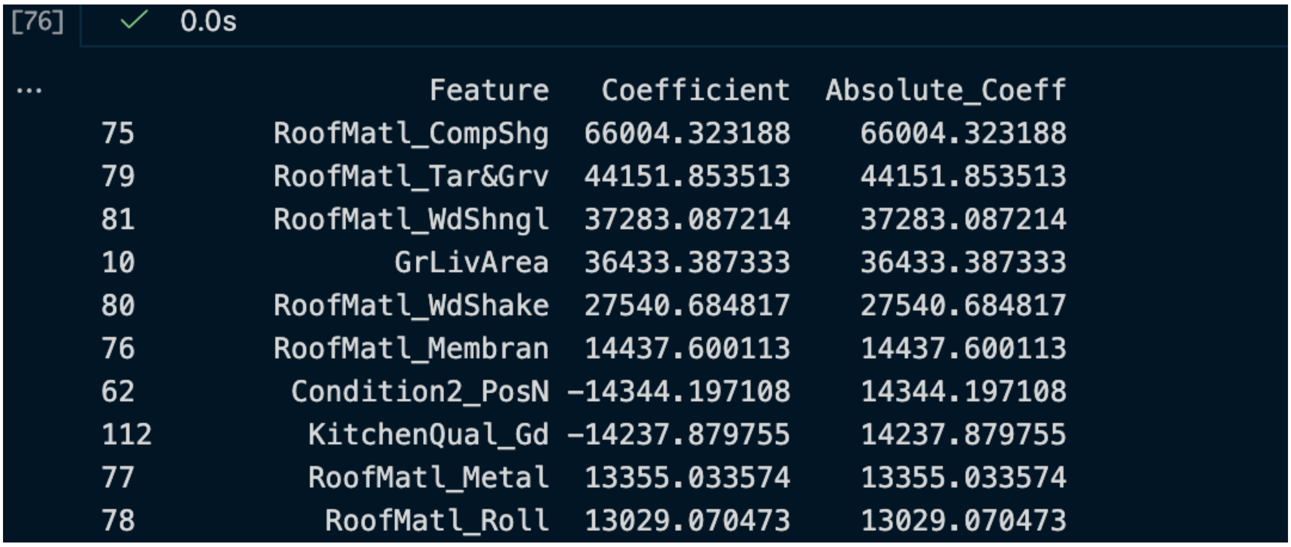
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.339357e-01	9.302835e-01	9.293667e-01
1	R2 Score (Test)	8.308043e-01	8.437506e-01	8.457934e-01
2	RSS (Train)	4.215364e+11	4.448397e+11	4.506900e+11
3	RSS (Test)	4.769141e+11	4.404221e+11	4.346642e+11
4	MSE (Train)	2.031911e+04	2.087319e+04	2.101000e+04
5	MSE (Test)	3.299765e+04	3.171009e+04	3.150213e+04

As it is evident, lasso has the best R2 score on test data and the difference between test and train is the least for lasso regression

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?

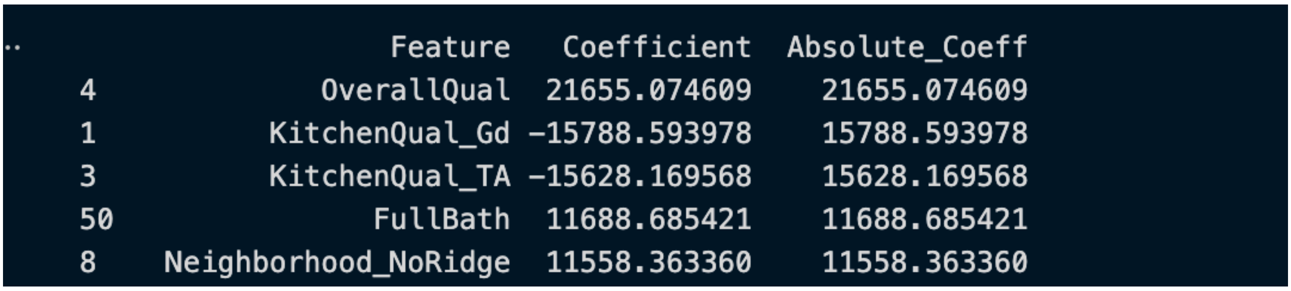
Answer: The top 10 coefficients based on lasso model are given below



```
[76]   ✓  0.0s

...                   Feature   Coefficient  Absolute_Coeff
     75        RoofMatl_CompShg  66004.323188    66004.323188
     79        RoofMatl_Tar&Grv  44151.853513    44151.853513
     81        RoofMatl_WdShngl  37283.087214    37283.087214
     10               GrLivArea  36433.387333    36433.387333
     80        RoofMatl_WdShake  27540.684817    27540.684817
     76        RoofMatl_Membran  14437.600113    14437.600113
     62         Condition2_PosN -14344.197108    14344.197108
     112          KitchenQual_Gd -14237.879755    14237.879755
     77          RoofMatl_Metal  13355.033574    13355.033574
     78           RoofMatl_Roll  13029.070473    13029.070473
```

Out of these, if the top 5 predictor variables are not available in the incoming data, the next 5 important variables after running lasso regression again, will be



```
..                    Feature   Coefficient  Absolute_Coeff
    4             OverallQual  21655.074609    21655.074609
    1          KitchenQual_Gd -15788.593978    15788.593978
    3          KitchenQual_TA -15628.169568    15628.169568
    50               FullBath  11688.685421    11688.685421
    8     Neighborhood_NoRidge  11558.363360    11558.363360
```

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: A model can be made robust and generalisable by using cross-validation, regularisation, proper data splitting, and feature selection. These techniques intentionally reduce training accuracy to control variance and prevent overfitting. As a result, although training accuracy may decrease, test accuracy becomes more reliable and representative of real-world performance