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?

The best hyperparameter values for Ridge and Lasso regression, based on the top 120 variables selected through Recursive Feature Elimination (RFE), are as follows:

- **Ridge Alpha: 5.0**
- **Lasso Alpha: 0.001**

When we increase the alpha value for Ridge and Lasso, the model tends to become simpler, resulting in increased bias and reduced variance. Additionally, as we raise the lambda (λ), (aka alpha) value, the coefficients' magnitude decreases.

Before any alterations, the most important predictor variables can be divided into two groups:

- The first half consists of variables sorted by absolute Ridge coefficients.
- The second half consists of variables sorted by absolute Lasso coefficients.

```
                         Ridge     Lasso
OverallQual_9         0.548672  0.893827
Neighborhood_NoRidge  0.458254  0.495898
OverallQual_10        0.444709  0.810819
FullBath_3            0.402350  0.464406
TotRmsAbvGrd_11       0.389301  0.561941
BsmtQual_TA           0.384423  0.406338
BsmtQual_Fa           0.377461  0.438682
1stFlrSF              0.311118  0.302724
Neighborhood_NridgHt  0.291803  0.269859
BsmtExposure_Gd       0.290568  0.276800
Index(['OverallQual_9', 'Neighborhood_NoRidge', 'OverallQual_10', 'FullBath_3',
       'TotRmsAbvGrd_11', 'BsmtQual_TA', 'BsmtQual_Fa', '1stFlrSF',
       'Neighborhood_NridgHt', 'BsmtExposure_Gd'],
      dtype='object')
                         Ridge     Lasso
OverallQual_9         0.548672  0.893827
OverallQual_10        0.444709  0.810819
TotRmsAbvGrd_11       0.389301  0.561941
Neighborhood_NoRidge  0.458254  0.495898
FullBath_3            0.402350  0.464406
BsmtQual_Fa           0.377461  0.438682
BsmtQual_No_Basement  0.158990  0.427374
OverallQual_8         0.196693  0.413236
BsmtQual_TA           0.384423  0.406338
Fireplaces_3          0.274595  0.381024
Index(['OverallQual_9', 'OverallQual_10', 'TotRmsAbvGrd_11',
       'Neighborhood_NoRidge', 'FullBath_3', 'BsmtQual_Fa',
       'BsmtQual_No_Basement', 'OverallQual_8', 'BsmtQual_TA', 'Fireplaces_3'],
      dtype='object')
```

Primary Predictor Variables Following Modification:

- **Ridge Alpha: 10.0. (doubled)**
- **Lasso Alpha: 0.002 (doubled)**

The initial half of the variables is arranged based on their absolute Ridge coefficients, while the latter half is organized by their absolute Lasso coefficients.

```
                         Ridge     Lasso
OverallQual_9         0.468019  0.888518
Neighborhood_NoRidge  0.391507  0.445670
OverallQual_10        0.352474  0.768713
BsmtQual_TA           0.341262  0.349484
FullBath_3            0.340954  0.431942
1stFlrSF              0.322460  0.311226
BsmtExposure_Gd       0.295137  0.284642
BsmtQual_Fa           0.291203  0.343902
TotRmsAbvGrd_11       0.282698  0.424173
Neighborhood_NridgHt  0.277305  0.239794
Index(['OverallQual_9', 'Neighborhood_NoRidge', 'OverallQual_10',
       'BsmtQual_TA', 'FullBath_3', '1stFlrSF', 'BsmtExposure_Gd',
       'BsmtQual_Fa', 'TotRmsAbvGrd_11', 'Neighborhood_NridgHt'],
      dtype='object')
                         Ridge     Lasso
OverallQual_9         0.468019  0.888518
OverallQual_10        0.352474  0.768713
Neighborhood_NoRidge  0.391507  0.445670
FullBath_3            0.340954  0.431942
TotRmsAbvGrd_11       0.282698  0.424173
OverallQual_8         0.184961  0.414269
BsmtQual_No_Basement  0.140299  0.383242
BsmtQual_TA           0.341262  0.349484
BsmtQual_Fa           0.291203  0.343902
1stFlrSF              0.322460  0.311226
Index(['OverallQual_9', 'OverallQual_10', 'Neighborhood_NoRidge', 'FullBath_3',
       'TotRmsAbvGrd_11', 'OverallQual_8', 'BsmtQual_No_Basement',
       'BsmtQual_TA', 'BsmtQual_Fa', '1stFlrSF'],
      dtype='object')
```

Metric	R2 Score (Train)	R2 Score (Test)	RSS (Train)	RSS (Test)	RMSE (Train)	RMSE (Test)
Linear Regression	0.9	0.84	98.71	74.08	0.31	0.41
Ridge Regression	0.9	0.85	106.12	69.0	0.32	0.4
Lasso Regression	0.9	0.85	105.88	70.12	0.32	0.4
Ridge Regression Double Lambda	0.89	0.85	112.92	69.01	0.33	0.4
Lasso Regression Double Lambda	0.89	0.85	113.89	70.3	0.33	0.4

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 for Ridge and Lasso regression, utilizing the top 120 variables chosen through RFE, are as follows:

- **Ridge Alpha: 5.0**
- **Lasso Alpha: 0.001**

Metric	R2 Score (Train)	R2 Score (Test)	RSS (Train)	RSS (Test)	RMSE (Train)	RMSE (Test)
Linear Regression	0.9	0.84	98.71	74.08	0.31	0.41
Ridge Regression	0.9	0.85	106.12	69.0	0.32	0.4
Lasso Regression	0.9	0.85	105.88	70.12	0.32	0.4

R2 : There has been an improvement in the difference between the R-squared (r2) values for the training and test datasets. Linear Regression – Train (0.9), Test (0.84). Post regularization, Train (0.9), Test (0.85)

RSS: Furthermore, the RSS for the test data has decreased from 74.08 to 69.0 for Ridge and 70.12 for Lasso, indicating improved model performance (lower values are preferable).

RMSE : Similarly, the Root Mean Squared Error (RMSE) for the test data has reduced from 0.41 to 0.39 for Ridge and Lasso, indicating better predictive accuracy (lower values are preferable).

```
#sum of coefficents
betas.sum()
```

```
Linear    4.399577
Ridge     1.598996
Lasso     2.498147
dtype: float64
```

```
#Number of variables in model after feature elimination
betas[betas!=0].count()
```

```
Linear    121
Ridge     118
Lasso      80
dtype: int64
```

Moreover, Lasso regularization facilitates feature elimination by driving their coefficients to zero, simplifying the model, enhancing its robustness, and improving its generalizability.

In this dataset, Lasso eliminated 40 features from 120 (selected by RFE), demonstrating its efficacy in feature selection.

In contrast, Ridge regularization didn't perform any major feature elimination.

Evidently, Lasso Regularization yields a simpler model and is expected to deliver superior results on unseen data. Lesser variables helps better explanation.

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?

Metric	R2 Score (Train)	R2 Score (Test)	RSS (Train)	RSS (Test)	RMSE (Train)	RMSE (Test)
Linear Regression	0.9	0.84	98.71	74.08	0.31	0.41
Ridge Regression	0.9	0.85	106.12	69.0	0.32	0.4
Lasso Regression	0.9	0.85	105.88	70.12	0.32	0.4
Ridge Regression Double Lambda	0.89	0.85	112.92	69.01	0.33	0.4
Lasso Regression Double Lambda	0.89	0.85	113.89	70.3	0.33	0.4
Ridge Regression Drop Top 5	0.88	0.85	123.16	71.13	0.35	0.4
Lasso Regression Drop Top 5	0.88	0.85	122.81	69.11	0.35	0.4

A decrease in R-squared (r2) values has been observed for both the training and test datasets.

There has been an increase in Residual Sum of Squares (RSS) and Root Mean Squared Error (RMSE) values for both the training and test datasets.

After the adjustments, the top 5 predictors for Ridge and Lasso regression are as follows:

Using Ridge:

- 'BsmtQual_TA',
- 'OverallQual_4',
- 'BsmtQual_Fa',
- 'OverallQual_5',
- 'OverallQual_6',

Using Lasso:

- 'OverallQual_4',
- 'OverallQual_5',
- 'OverallQual_3',
- 'OverallQual_6',
- 'BsmtQual_Fa',

```
                     Ridge      Lasso
BsmtQual_TA        0.433544  0.467369
OverallQual_4      0.427662  0.624788
BsmtQual_Fa        0.415342  0.492583
OverallQual_5      0.414648  0.594444
OverallQual_6      0.384770  0.550162
1stFlrSF           0.355413  0.338224
KitchenQual_TA     0.339388  0.321421
TotRmsAbvGrd_10    0.334195  0.290988
KitchenQual_Fa     0.330739  0.360429
OverallQual_3      0.325835  0.564505
Index(['BsmtQual_TA', 'OverallQual_4', 'BsmtQual_Fa', 'OverallQual_5',
       'OverallQual_6', '1stFlrSF', 'KitchenQual_TA', 'TotRmsAbvGrd_10',
       'KitchenQual_Fa', 'OverallQual_3'],
      dtype='object')
                          Ridge      Lasso
OverallQual_4          0.427662  0.624788
OverallQual_5          0.414648  0.594444
OverallQual_3          0.325835  0.564505
OverallQual_6          0.384770  0.550162
BsmtQual_Fa            0.415342  0.492583
BsmtQual_TA            0.433544  0.467369
Fireplaces_3           0.302195  0.453422
BsmtQual_No_Basement   0.164897  0.421333
OverallQual_7          0.259060  0.398541
KitchenQual_Fa         0.330739  0.360429
Index(['OverallQual_4', 'OverallQual_5', 'OverallQual_3', 'OverallQual_6',
       'BsmtQual_Fa', 'BsmtQual_TA', 'Fireplaces_3', 'BsmtQual_No_Basement',
       'OverallQual_7', 'KitchenQual_Fa'],
      dtype='object')
```

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?

We can employ **Ridge** and **Lasso** Regularization techniques to ensure the model's robustness and generalizability.

A robust model maintains consistently accurate predictions even when one or more input variables undergo changes.

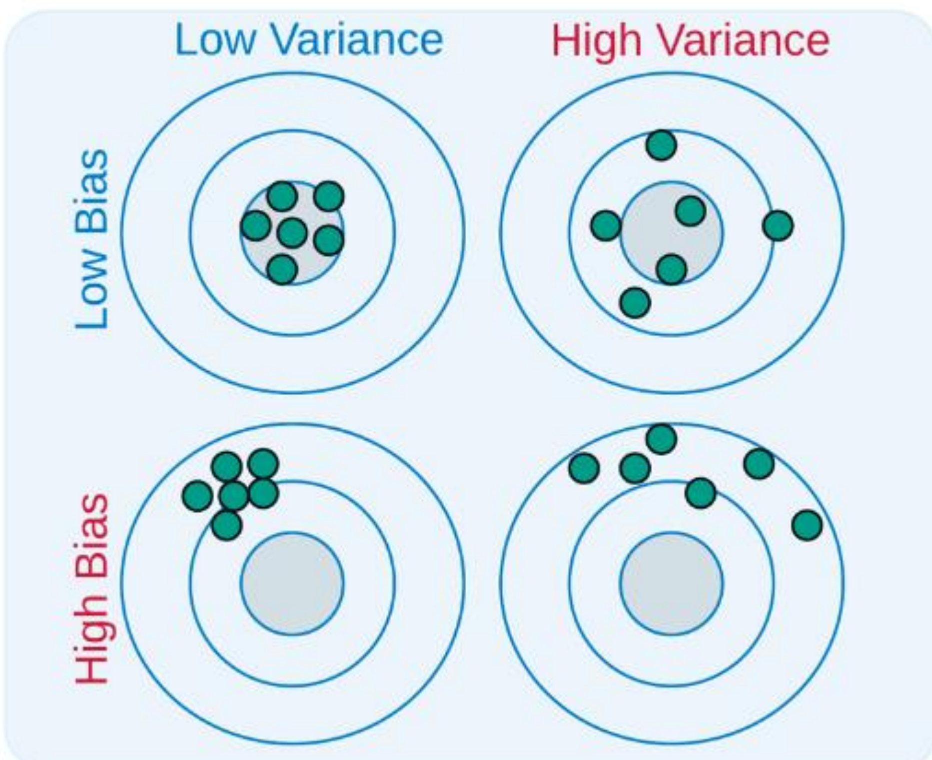
A generalizable model can effectively predict outcomes for the entire population using a model trained on a subset of the data.

Simpler models are generally more versatile, and simplicity contributes to robustness. Various ways to gauge a model's complexity include:

- The number of parameters required to fully define the model.
- The degree of the function, especially relevant for polynomial models.
- The size of the most compact representation of the model.
- The depth or size of a decision tree.

In the context of model evaluation, a small difference in R-squared (r2) values between the training and test datasets is desirable.

Balancing bias and variance, a key aspect of model development, involves minimizing both.



Avoiding excessive model complexity, such as reducing polynomial degree or tree depth, is crucial. Implications of simplifying the model include:

- Potential reduction in model accuracy on the training set.
- Potential improvement in model accuracy on the test set.