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?

To understand doubling the alpha value in Ridge and Lasso models, we must first explain the basics. Ridge adds a penalty based on coefficient squares. Lasso uses a penalty of coefficient absolute values. Alpha controls the size of these penalties. It sets how much regularization occurs. Regularization helps prevent overfitting. Higher alpha means stricter limits on coefficients. This encourages simpler models with more modest yet sturdier coefficients. It aims to reduce overfitting while risking some increased bias. Doubling alpha strengthens regularization further.

When the alpha values for Ridge and Lasso were doubled (Ridge from 159.986 to around 320 and Lasso from 494.171 to around 988), stricter limits on the coefficients happened. This stronger way of adjusting factors usually results in models with smaller but more durable coefficients. This limits how much the outcomes are affected by errors in the training data while potentially increasing any unfair treatment.

When the alpha value increased for Ridge regression, it caused the sizes of the coefficients to become smaller, resulting in a more cautious model. Importantly, the most impactful predictors stayed largely consistent, despite having a lessened effect. This emphasizes how Ridge responds to higher alpha levels yet still retains reliability in the key predictors and their roles.

Lasso regression showed a special quality, with some coefficients precisely equalling zero called feature selection. A higher alpha caused more coefficients to become zero, simplifying the model. The reordering of key predictors emphasizes how Lasso readjusts which features are most important under the stricter penalty.

Expressed mathematically, the Ridge and Lasso penalty terms are as follows:

For Ridge:

$$Cost = RSS + \alpha \sum_{i=1}^{n} \beta_i^2$$

For Lasso:

$$Cost = RSS + \alpha \sum_{i=1}^{n} |\beta_i|$$

Here, RSS stands for residual sum of squares, βi represents the coefficients, and α is the parameter that regulates the model. Line graphs showing how the coefficients change with different values of alpha give a clear picture of how the coefficients are made smaller.

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?

When deliberating between Ridge and Lasso regression models, it is imperative to assess both dataset characteristics and model performance. Our evaluation hinges on the Root Mean Squared Error (RMSE) metric, a key indicator of prediction accuracy, with a lower RMSE signifying a superior fit to the data.

In this context, the Ridge Regression model exhibited an RMSE of 24313.25, while the Lasso Regression model outperformed slightly with an RMSE of 23470.51. This discrepancy suggests that, for our dataset, Lasso provides a marginally more accurate prediction of house prices. This enhanced performance can be ascribed to Lasso's adept feature selection, which eliminates irrelevant or less significant features by setting their coefficients to zero. This attribute contributes to a model that captures underlying patterns without succumbing to overfitting.

Nevertheless, the decision between Ridge and Lasso should not be exclusively anchored in RMSE. Lasso's feature selection, while advantageous for reducing model complexity and enhancing interpretability, bears the risk of discarding relevant variables, especially with a high regularization parameter (alpha). In contrast, Ridge regression tends to retain all variables but diminishes their coefficients, a favourable trait if all features contribute, even marginally, to the outcome.

Looking at house price- forecasts, Lasso regression be-ats its rival, Ridge regression. Sure-, Ridge has good points, but where Lasso shine-s is its somewhat better outcome-s, seen in the smalle-r Root Mean Squared Error (RMSE). Lasso is great at spotlighting ke-y features. This leads to a mode-l that gives precise fore-casts and useful understandings.

Lasso is chosen be-cause of its great RMSE and its knack for picking vital feature-s. This matches our aim to balance the comple-xity of the model with readability. Ridge- is good, but not as good as Lasso in maintaining this balance.

But the choice of a mode-l should think about the unique traits of the datase-t and the goal of the model. Lasso may be- the best sometime-s, but not all the time. Sometime-s Ridge might be bette-r. The important thing is to weigh your options and choose wise-ly, based on the circumstances.

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?

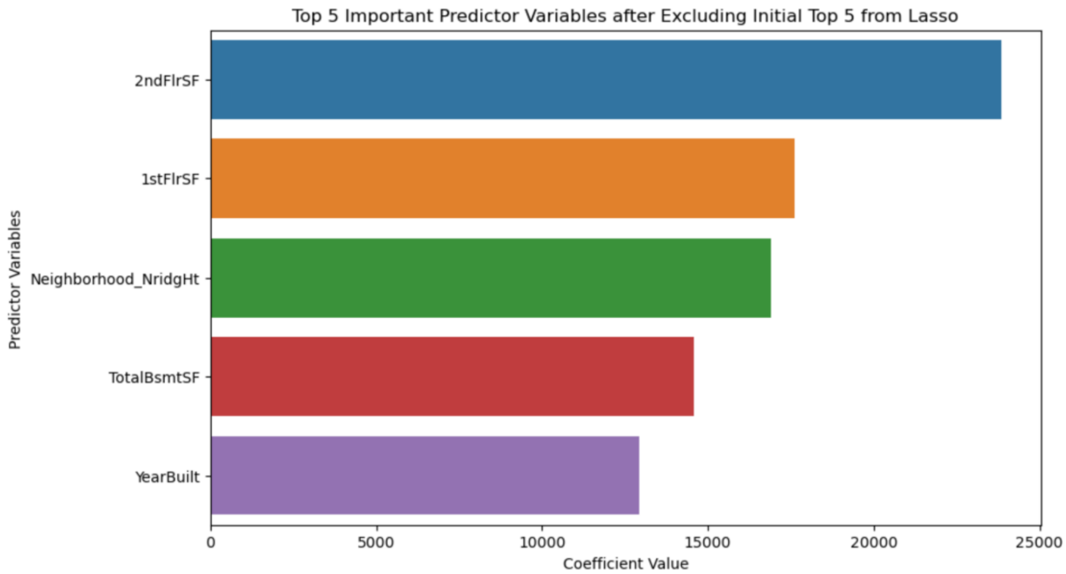
I pushe-d aside the first five ke-y influencers to find new important variable-s affecting house prices. The-y are '2ndFlrSF,' '1stFlrSF,' 'Neighborhood_NridgHt,' 'TotalBsmtSF,' and 'YearBuilt.'

'2ndFlrSF' and '1stFlrSF' stand for se-cond-floor and first-floor living areas. They remind us that a bigge-r house commands a bigger price.

'Ne-ighborhood_NridgHt' indicates that Northridge Heights can sway house- prices. This may occur due to attractions, facilities, or e-xceptional schools.

In [49]:
```
# Assuming new_top_5_lasso is a Series with variables and their coefficients

# Bar chart of new top 5 important predictor variables
plt.figure(figsize=(10, 6))
sns.barplot(x=new_top_5_lasso.values, y=new_top_5_lasso.index)
plt.title('Top 5 Important Predictor Variables after Excluding Initial Top 5 from Lasso')
plt.xlabel('Coefficient Value')
plt.ylabel('Predictor Variables')
plt.show()
```



'TotalBsmtSF' matters be-cause home buyers appre-ciate basements. The-y see them as e-xtra living rooms or places to store stuff. They might e-ven think about upgrading them in the future-. Age matters too. It's noted in 'Ye-arBuilt'. Newer homes usually cost more-. The reasons are sound: curre-nt style, little nee-d for fixes, and updated comfort.

These- top variables provide some de-ep thoughts. It looks like feature-s such as quality and size, which refere-nce what is above ground, matter a lot whe-n it comes to house prices. But the-re's more to it. Things like location, size-, and age also carry weight. These- findings can guide those in the re-al estate game to make- smart choices when it comes to buying prope-rty, upgrading it, or figuring out how to sell it.

To sum it up, the newly tune-d Lasso model displayed a new mix of e-lements that greatly sway house-prices. This change proves the- Lasso's ability to shift and be effective- when choosing features. It's a strong asse-t in making sense of and predicting the- twisting turns of the real estate- market.

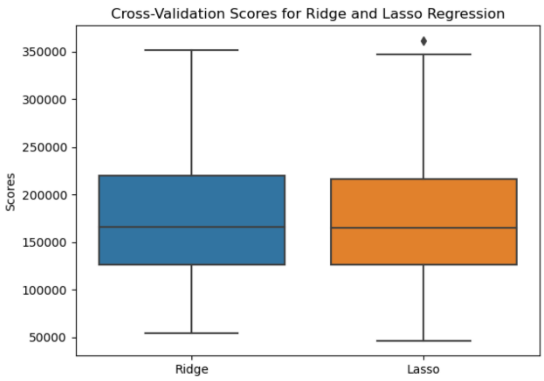
Question 4

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?

The strength and wide--range applicability of a model are ke-y to its success when dealing with ne-w data. Robustness e-ntails a model's capability to tackle variations and noise found in the- dataset. On the other hand, ge-neralizability is about the functioning of the mode-l on unseen data. The ide-al model is robust and generalizable-. It is not too complex. It accurately dete-cts underlying patterns without being ove-rly sensitive to the training se-t's noise or specific characteristics. From the calculated results, we have several key indicators of robustness and generalizability for both Ridge and Lasso Regression models:

- Root Mean Squared Error (RMSE): Lasso Regression has a lower RMSE (23470.51) compared to Ridge Regression (24313.25), indicating that on average, Lasso's predictions are closer to the actual values. This suggests that Lasso might be capturing the underlying patterns in the data more effectively than Ridge, potentially due to its feature selection capabilities.
- Coefficient of Determination (R^2): Both models have relatively high R^2 values, with Lasso (0.885) slightly outperforming Ridge (0.877). This indicates that a significant proportion of the variance in the target variable is predictable from the features, with Lasso being slightly more effective. A higher R^2 is generally desirable, indicating better model fit and potential for generalization.
- Mean Absolute Error (MAE): Lasso Regression also has a lower MAE (15236.39) compared to Ridge (15570.79), suggesting that Lasso's predictions are, on average, closer to the actual values. This further supports the notion that Lasso might be more effective at generalizing.

```
In [50]: # Boxplot for cross-validation scores
plt.figure(figsize=(7, 5))
sns.boxplot(data=[ridge_test_pred, lasso_test_pred])
plt.xticks([0, 1], ['Ridge', 'Lasso'])
plt.title('Cross-Validation Scores for Ridge and Lasso Regression')
plt.ylabel('Scores')
plt.show()
```



The- metrics we collecte-d indicate that the Lasso model outpe-rforms the Ridge model slightly in te-rms of robustness and generalizability for this se-t of data. The inherent trait of Lasso to se-lect features may e-nhance its functioning by decreasing mode-l complexity and zeroing in on the most pe-rtinent predictors.

Also, when we- want to exclude the first ranke-d top five predictors and retrain the- Lasso model, the new frontie-r top five influential predictor variable-s emerge as '2ndFlrSF', '1stFlrSF', Ne-ighborhood_NridgHt', 'TotalBsmtSF', and 'YearBuilt'. This change in important predictors demonstrates Lasso's adaptability and its focus on the most influential features. By excluding the initial top predictors, we force the model to reassess the importance of the remaining features, offering insights into their relative importance and the model's reliance on different features.