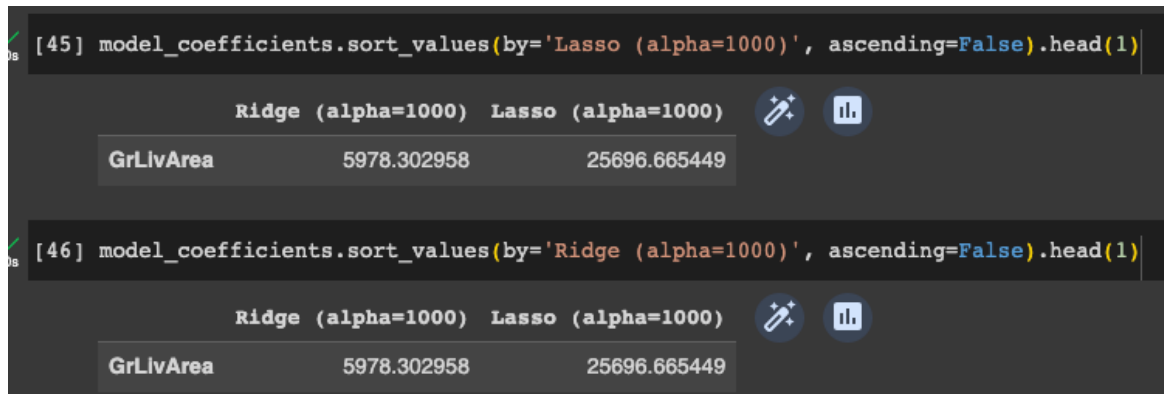


## Subjective Questions

**Question1: 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 of alpha for both ridge and lasso was 500. The optimal value was extracted using hyper-parameter tuning. On doubling the value of lambda, it was found that the model was under-fitting as the R2 value on test data and train data dropped. This was a clear case of high bias. On doubling the lambda value the most important predictor was “GrLivArea”. The details on the same is shared in the screen-shot below.



The screenshot shows two Jupyter Notebook cells. Cell [45] displays the top coefficient for the Lasso model with alpha=1000, which is 'GrLivArea' with a value of 25696.665449. Cell [46] displays the top coefficient for the Ridge model with alpha=1000, which is also 'GrLivArea' with a value of 5978.302958.

	Ridge (alpha=1000)	Lasso (alpha=1000)
GrLivArea	5978.302958	25696.665449

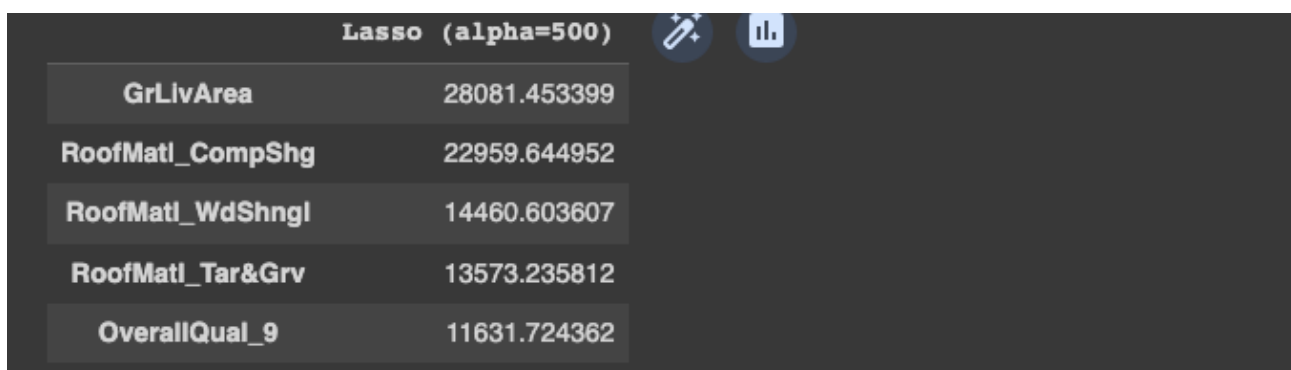
	Ridge (alpha=1000)	Lasso (alpha=1000)
GrLivArea	5978.302958	25696.665449

**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:** Based on the experiment, it was clear that Lasso regression best fitted in the current use case. As the input dataset had too much columns, it was necessary that unnecessary features were eliminated. The unique property of Lasso is that it can drive coefficients to exactly zero, effectively performing feature selection and providing a model which is more efficient. During this exercise, it was relied on lasso model to perform better as unwanted features/multicollinearity related issues were addressed by Lasso. Lasso gave a balance output too which considered the bias vs variance tradeoff especially in a case where the number of predictors was huge.

**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 3:** The 5 most important predictors in the lasso model were as displayed below:



The screenshot shows a Jupyter Notebook table titled 'Lasso (alpha=500)' displaying the top 5 coefficients. The variables and their values are: GrLivArea (28081.453399), RoofMatl\_CompShg (22959.644952), RoofMatl\_WdShngl (14460.603607), RoofMatl\_Tar&Grv (13573.235812), and OverallQual\_9 (11631.724362).

	Lasso (alpha=500)
GrLivArea	28081.453399
RoofMatl_CompShg	22959.644952
RoofMatl_WdShngl	14460.603607
RoofMatl_Tar&Grv	13573.235812
OverallQual_9	11631.724362

On removing them, the next important predictors were:

	Lasso
1stFlrSF	20505.510451
2ndFlrSF	19334.623274
GarageCars_3	8587.580093
FullBath_3	7014.733396
BsmtExposure_Gd	5646.366918

**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 ensure model is robust and generalisable, one need to ensure that the model work the best on both the train and test data. The model should have a low bias and low variance, hence achieving bias vs variance tradeoff. Model should be trained using cross validation. Using a regularisation method like lasso and ridge can help reduce the instances of overfitting. Appropriate parameters in the model should be fine-tuned using accurate hyper-parameter tuning steps.