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?

Alpha is nothing but the tuning parameter of the cost function which we minimize to get best fitted value of betas using Ridge and Lasso regression. The greatest the value of alpha, the model tends to underfit and the lowest the value of alpha, the overfitting is not addressed. If we doubled the value of alpha, it means model tend to be a bit more underfit and thus the value of r2_core will decrease for both test and train data set.

After change is implemented. Most important predictor will be

	Ridge	Lasso
Neighborhood_NoRidge	55060.622046	85082.944816
Neighborhood_NridgHt	43264.224807	65232.485007
MasVnrArea	40426.509266	73538.331842
BsmtExposure_Gd	37891.184584	42092.034372
WoodDeckSF	28126.235755	40838.121415

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?

In OLS regression we minimize the RSS (Residual sum of square) to get the best coefficients value. But in Lasso and Ridge, the cost function is little bit different, we add a penalty with RSS and get cost function. Now the cost function has a tuning parameter lambda which helps us to determine how much wish to regularize the model. I tried multiple value of lambda with GridsearchCV to get the best lambda for the data set and used that value to perform further analysis. If the value of lambda is too high, it will lead to underfitting and if the value is too low, it will not handle the over fitting. My chosen lambda is 9.0 for Ridge and 100 for Lasso. I choose Ridge regression as value of lambda is not too large or two low compared to lambda I got in Lasso and more over the model accuracy r2_score for test and train data is better in Ridge compared to Lasso.

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?

Following 5 important variable are

1. MasVnrArea: Masonry veneer area in square feet
2. Neighborhood: Physical locations within Ames city limits
3. WoodDeckSF: Wood deck area in square feet
4. OpenPorchSF: Open porch area in square feet
5. BsmtExposure: Refers to walkout or garden level walls

Question 4

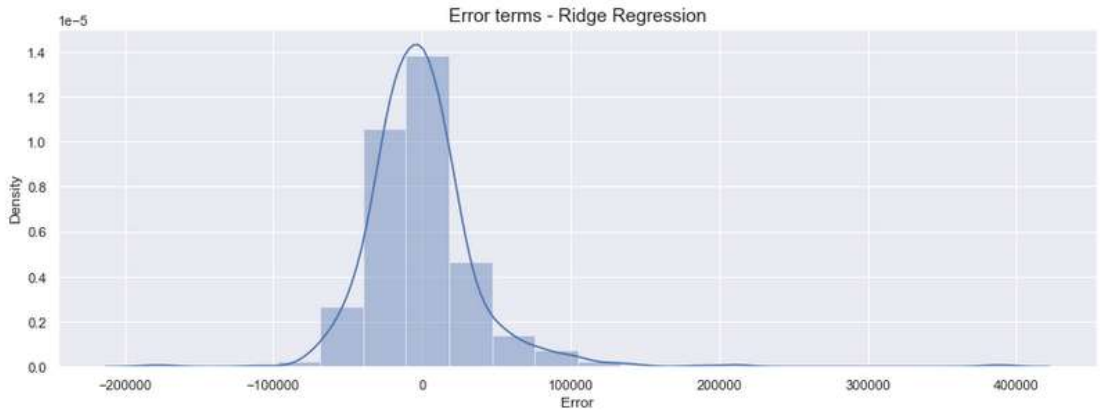
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

We can compare the value of R2_score and RMSE (root mean square error) for test and train data set to determine the accuracy and generalizability. Also the error terms are normalized in test data set.

Ridge:

```
r2_score train : 0.7633920384901599
r2_score test : 0.7414525848115794
```

```
1 sns.distplot((y_train - y_pred_train), bins = 20)
2 plt.title("Error terms - Ridge Regression", fontsize=15)
3 plt.xlabel("Error")
4 plt.show()
```



Lasso:

```
r2_score train : 0.775675727595701
r2_score test : 0.7375982926403913
```

In [93]:

```
sns.distplot((y_test - y_pred_test), bins = 20)
plt.title("Error terms - Lasso Regression", fontsize=15)
plt.xlabel("Error")
plt.show()
```