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 1:

The optimal value of alpha for ridge regression is 100

The optimal value of alpha for lasso regression is 0.001.

The top 10 important predictor variables at optimal value of alpha are:

```
In [69]: a = pd.DataFrame(X_train.columns)
         b = pd.DataFrame(ridge.coef_)

In [70]: coefficients = pd.concat([a,b], axis=1)
         coefficients.columns = ['Features','Coefficient']

In [71]: coefficients.sort_values(by='Coefficient', ascending=False).head(10)
```

Out[71]:

	Features	Coefficient
12	GrLivArea	0.059387
1	OverallQual	0.053578
10	1stFlrSF	0.043126
9	TotalBsmtSF	0.038409
2	OverallCond	0.036185
11	2ndFlrSF	0.029444
19	GarageArea	0.027710
27	MSZoning_RL	0.027653
4	LotArea	0.025752
7	BsmtFinSF1	0.022943

Upon doubling the value of alpha:
Test r2 score decreased by 1% because more features were eliminated in Lasso, while in the Ridge, R2 score slightly decreased.

The MSE / RMSE values went up slightly and the magnitude of coefficients decreased.

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:
Between Lasso and Ridge, I'll choose the Ridge regression model because the test r2 score was little bit higher than Ridge model. Also, the RMSE value was lower with Ridge regression.

With optimal lambda value, using Lasso we had 164 features and after doubling the lambda, we had 110 features only. Since Ridge does not eliminate features (only minimize the coefficients), there wasn't much difference.

Doubling the lambda in our case, will generalize little better on unseen data than optimal value of lambda.

Another thing to keep in mind is that Lasso model is computationally heavy. It depends on current company's computational power if Lasso is the right model for them because there is not significant difference between r2_score and RMSE of Lasso and Ridge model.

Question 3

After building the model, you realized 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:

After removing top 5 features, the new model's top 5 features are Basement finish SF1, Basement unfinish SF, garage area, 2nd floor square feet and full bath.

The r2 score decreased by 1% and RMSE increased little bit.

In [79]:
```
c = pd.DataFrame(X_train_new.columns)
d = pd.DataFrame(ridge_new.coef_)
```

In [80]:
```
coefficients = pd.concat([c,d], axis=1)
coefficients.columns = ['Features','Coefficient']

coefficients.sort_values(by='Coefficient', ascending=False)
```

Out[80]:

	Features	Coefficient
5	BsmtFinSF1	0.064823
6	BsmtUnfSF	0.053967
14	GarageArea	0.045039
7	2ndFlrSF	0.044560
9	FullBath	0.037988
...	...	...
48	Neighborhood_IDOTRR	-0.022645
148	BsmtQual_TA	-0.022818
185	KitchenQual_Gd	-0.024400
49	Neighborhood_MeadowV	-0.024793
186	KitchenQual_TA	-0.039994

Question 4

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

Answer 4:

By doing train test split, we measure the training model on test data set (unseen). If the difference between train and test accuracy score is very wide, let's say > 0.10, the model seems to be overfitting and will not generalize well. The model will learn train data set perfectly and won't able to perform well on unseen data.

Contrary to it, the closer the difference in train and test accuracy score, the more generalizable is the model.

R2 score denotes how much variation in a model is explained by current set of features. Generally speaking, the test r2 score should be lower than train r2 score. The closer the gap between them, better is the model.

In our current assignment, the r2 score on train and test data set is very close (0.94 on train and 0.92 on test) with a significant number of feature elimination. This makes our model more robust and generalizable.