### 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:

Optimal value of alpha for ridge and lasso regression are:

Ridge regression	3
Lasso Regression	100

Changes in the model if you choose double the value of alpha for both ridge and lasso:

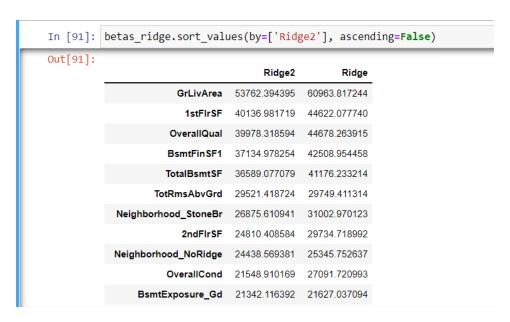
Ridge regression for alpha=3		ha=3	Ridge regression for alpha =6		
R2_score_train	=	0.9382415248870304	R2_score_train = 0.9335605422803847		
R2 score test	=	0.9141125243557431	R2_score_test = 0.9111806712392462		
RSS train	=	314235160827.6926	RSS_train = 338052609680.5071		
RSS test	=	210081184626.8556	RSS_test = 217252511659.64386		
MSE_train	=	336800815.46376485	MSE_train = 362328627.73902154		
MSE test	=	525202961.56713897	MSE_test = 543131279.1491096		
_					

Lasso regression for alpha=100	Lasso regression for alpha =200
R2_score_train = 0.9312366017154 R2_score_test = 0.9165593652143 RSS_train = 349877121795.64 RSS_test = 204096199944.19 MSE_train = 375002274.16468 MSE_test = 510240499.86047	219 R2_score_test = 0.9118871137890757 72 RSS_train = 394335003662.0322 055 RSS_test = 215524549734.74432 084 MSE_train = 422652737.0439788

Hence, by comparing both the values we can conclude that, when the optimal value of alpha is doubled,

- 1. R2 score in both cases (Ridge and Lasso) for train as well as test data is decreased.
- 2. RSS (Residual Sum of Squares) and MSE (Mean square Error) has increased.

The most important predictor variables after the change is implemented are compared in below image:



Hence, most important variable in **Ridge regression** for alpha=6 is **GrLivArea** with highest coefficient value = **53762.394395** 

In [97]:	<pre>betas_lasso.sort_values(by=['Lasso2'], ascending=False)</pre>			
Out[97]:		Lasso	Lasso2	
	GrLivArea	153253.870765	163597.847946	
	OverallQual	60390.760239	65343.682202	
	TotalBsmtSF	56728.318294	50211.534087	
	BsmtFinSF1	32833.197828	34881.600000	
	GarageCars	24902.593725	27715.668635	
	Neighborhood_StoneBr	30198.584422	21899.468216	
	SaleCondition_Partial	18966.405915	20946.652090	
	Neighborhood_NoRidge	25297.818143	20603.883679	
	BsmtExposure_Gd	20115.486527	19608.548168	
	$Neighborhood\_NridgHt$	17762.720709	18519.702491	
	OverallCond	26708.468281	17655.583816	
	- " 17	44740 450040	4.4554.000540	

Hence, most important variable in **Lasso regression** for alpha=200 is **GrLivArea** with highest coefficient value = **163597.847946** 

### **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:**

- For Optimal values of alpha, the R2 score for Ridge and Lasso is 0.9382415248870304 and 0.9312366017154192 respectively which are almost same.
- Hence I'll go for **lasso regression** because along with good R2 score it also have done the feature selection and thus with less features than ridge my model will be simple.

#### •

#### **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:**

The five most important predictor variables are:

Features	Coeff values		
OverallCond	26708.468281		
Neighborhood_NoRidge	25297.818143		
GarageCars	24902.593725		
KitchenQual_TA	-20792.825652		
KitchenQual_Gd	-20716.687472		

The next 5 features mentioned above has highest absolute value of coefficient and thus they are the most important predictor variables.

## **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:**

- A model is said to be robust when the performance of the model is not affected by any variations in the data. A generalizable model means it performs well on unseen data from the same dataset used for model building.
- In order to meet expectation of robustness and generasilablity, our model should not be too complex.

- Generally, complex model tends to overfit i.e., they performs well on seen (train) data but fails to perform well on unseen data.
- Hence, the model should be simple.

# Implications of the same for the accuracy of the model:

- Generally complex models have good accuracy as complex model try to fit in every possible datapoints.
- When we want our model to be robust and generalise, the accuracy may get affected since we need to decrease the variance which in result add some bias. Hence, accuracy get affected (decrease).
- To overcome these issues we go for Regularisation techniques like Ridge and Lasso.