Question 1

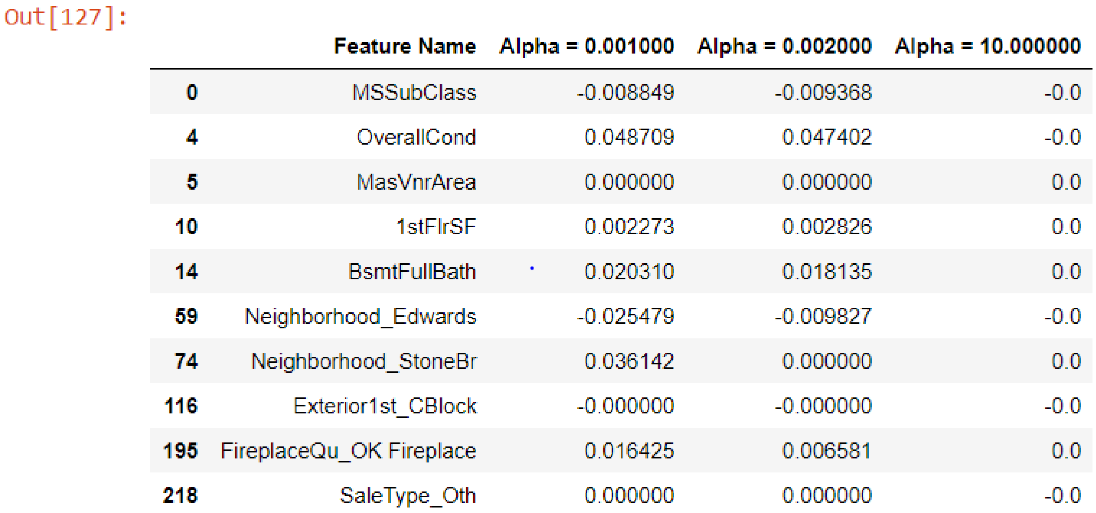
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer 1) The optimal value for alpha are:-

- Lasso = 0.001
- Ridge = 10.0

We observed following changes in Lasso after making alpha value double i.e 0.002

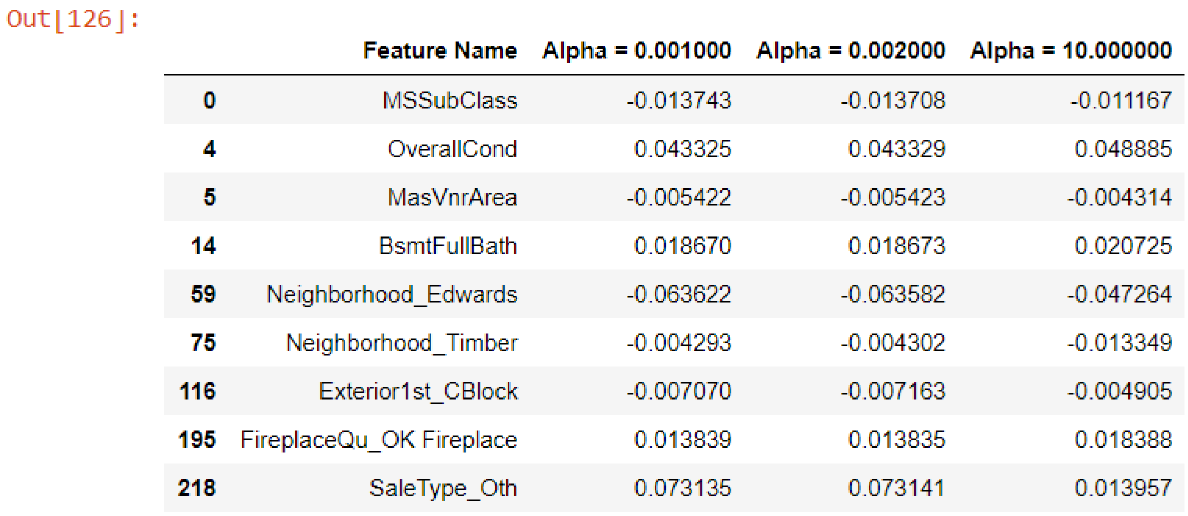
- Firstly there is change in R squared value as it has decreased from 93% to 92% in training and 87% to 86% in test
- The coefficients of the features reduced and will reach to 0 as the alpha keeps on increasing.
- See the output below



Out[127]:

	Feature Name	Alpha = 0.001000	Alpha = 0.002000	Alpha = 10.000000
0	MSSubClass	-0.008849	-0.009368	-0.0
4	OverallCond	0.048709	0.047402	-0.0
5	MasVnrArea	0.000000	0.000000	0.0
10	1stFlrSF	0.002273	0.002826	0.0
14	BsmtFullBath	0.020310	0.018135	0.0
59	Neighborhood_Edwards	-0.025479	-0.009827	-0.0
74	Neighborhood_StoneBr	0.036142	0.000000	0.0
116	Exterior1st_CBlock	-0.000000	-0.000000	-0.0
195	FireplaceQu_OK Fireplace	0.016425	0.006581	0.0
218	SaleType_Oth	0.000000	0.000000	-0.0

- In case of Ridge R square value for training data decreased from 94% to 93% but test data value is almost similar about 87%
- The coefficients of the features reduced but do not become exact 0 rather close to the zero as the alpha keeps on increasing. This is clear as ridge regression doesn’t eliminate features.
- See the output below



Out[126]:

	Feature Name	Alpha = 0.001000	Alpha = 0.002000	Alpha = 10.000000
0	MSSubClass	-0.013743	-0.013708	-0.011167
4	OverallCond	0.043325	0.043329	0.048885
5	MasVnrArea	-0.005422	-0.005423	-0.004314
14	BsmtFullBath	0.018670	0.018673	0.020725
59	Neighborhood_Edwards	-0.063622	-0.063582	-0.047264
75	Neighborhood_Timber	-0.004293	-0.004302	-0.013349
116	Exterior1st_CBlock	-0.007070	-0.007163	-0.004905
195	FireplaceQu_OK Fireplace	0.013839	0.013835	0.018388
218	SaleType_Oth	0.073135	0.073141	0.013957

Top features after changes

Lasso	Ridge
MSSubClass	MSSubClass
BsmtFullBath	OverallCond
OverallCond	Neighborhood_Edwards
FireplaceQu_OK Fireplace	BsmtFullBath
Neighborhood_Edwards	FireplaceQu_OK Fireplace
MasVnrArea	MasVnrArea
SaleType_Oth	SaleCondition_Partial
1stFlrSF	LowQualFinSF
OverallQual	Exterior1st_CBlock
BsmtFinSF2	Neighborhood_Timber

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) As we can see the R squared value for both the models is very similar hence will choose Lasso over ridge because LASSO does both parameter shrinkage and variable selection automatically. Ridge regression can't zero out the coefficients.

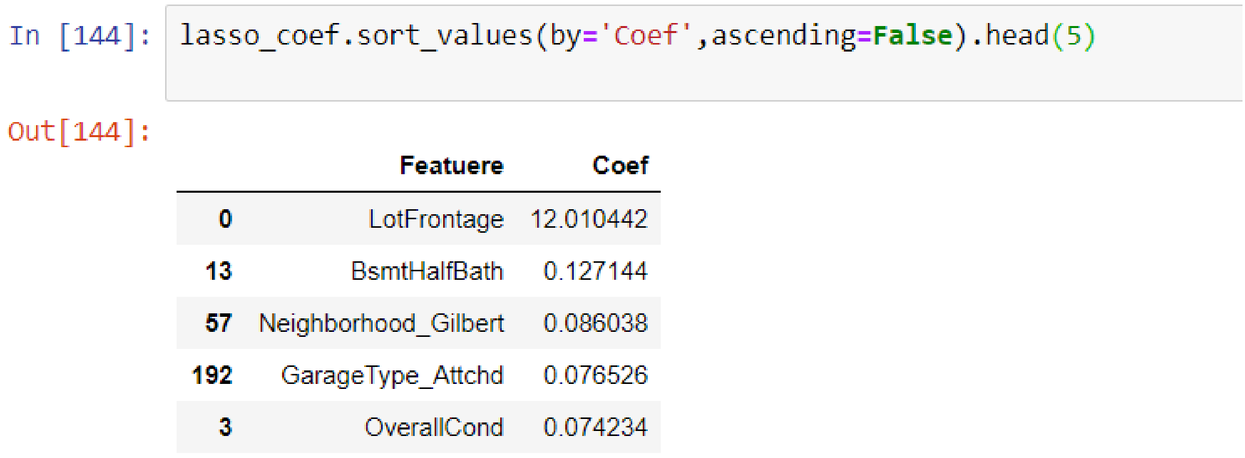
Question 3

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) After removing the top 5 predictors from lasso new model predicts:

R squared value for train to be 92% and test to be 87%

The five most important variables are as follows:=



```
In [144]: lasso_coef.sort_values(by='Coef',ascending=False).head(5)
```

Out[144]:

	Featuere	Coef
0	LotFrontage	12.010442
13	BsmtHalfBath	0.127144
57	Neighborhood_Gilbert	0.086038
192	GarageType_Attchd	0.076526
3	OverallCond	0.074234

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) Robustness is the property that tested on a training sample and on a similar testing sample, the performance is close.

Generalization is the ability of the learned model to fit unseen instance.

The ability of a model to generalize is crucial to the success of a model. If a model is trained too well on training data, it will not be able to generalize, case of overfitting. Thus, one needs to know when to stop training the model so that it doesn't overfit.

In order to make sure that model is robust and less variant we can split the training and test data set using “cross validation”.

Also to make model robust we handle outliers in the features.

We can also perform transformation of features from which we cannot remove outliers for example our target variable hence we can use log transformation on such features.

There is a tradeoff between accuracy and generalization of the model. The model that gives very high accuracy on the training data fails to generalize on the unseen data.

Accurate models can thus lead to overfitting whereas robust one will learn from the data. So we can prefer robust one over accurate models.