Subjective Questions:

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

The optimal value of alpha for ridge is 2 and for lasso it is 0.0001.
The R2 of the model with these alpha values is approximately 0.82.
After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.82 but there is a small change in the co-efficient values.

The R2 Score of the model on the test dataset for doubled alpha is 0.8259998671982055.

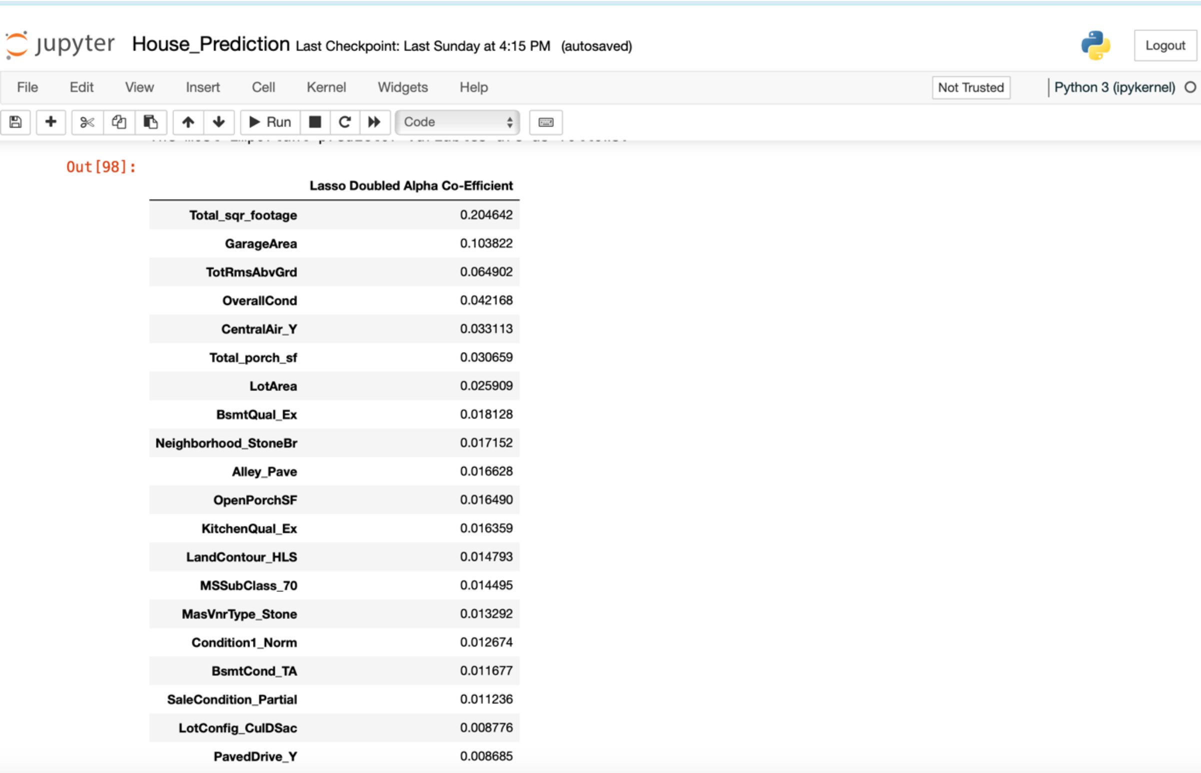
The Mean Squared Error (MSE) of the model on the test dataset for doubled alpha is 0.0018622905336132 81.
The important predictor variables are as follows:

Ridge model:

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
Neighborhood_Veenker	0.020098
BsmtQual_Ex	0.019949
KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.018952
MasVnrType_Stone	0.018388
PavedDrive_P	0.017973
RoofMatl_WdShngl	0.017856
PavedDrive_Y	0.016840

Lasso Model:

The R2 Score of the model on the test dataset for doubled alpha is 0.8237798637847479.
The Mean Squared Error (MSE) of the model on the test dataset for doubled alpha is 0.0018860508105446822.
The most important predictor variables are as follows:



	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.204642
GarageArea	0.103822
TotRmsAbvGrd	0.064902
OverallCond	0.042168
CentralAir_Y	0.033113
Total_porch_sf	0.030659
LotArea	0.025909
BsmtQual_Ex	0.018128
Neighborhood_StoneBr	0.017152
Alley_Pave	0.016628
OpenPorchSF	0.016490
KitchenQual_Ex	0.016359
LandContour_HLS	0.014793
MSSubClass_70	0.014495
MasVnrType_Stone	0.013292
Condition1_Norm	0.012674
BsmtCond_TA	0.011677
SaleCondition_Partial	0.011236
LotConfig_CulDSac	0.008776
PavedDrive_Y	0.008685

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?

The optimal values of lambda for Ridge model are 2 and for Lasso model is 0.0001. The mean squared error value for Ridge and Lasso model is almost same ~0.0018. Since, Lasso model helps in feature prediction (as the co-eff value of some of the feature become zero), this model has better edge over Ridge model and can be used as the final model.

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?

The five most important predictor variables in the current lasso model are:

1. Total_sqr_footage
2. GarageArea
3. TotRmsAbvGrd
4. OverallCond
5. LotArea

The R2 Score of the model on the test dataset is 0.7330077964268464
The Mean Squared Error (MSE) of the model on the test dataset is 0.002857 5670906482546
The most important predictor variables are as follows:

Out[102]:

	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

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?

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons: -

▪ Simpler models are usually more 'generic' and are more widely applicable.
▪ Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
▪ Simpler models are more robust.
    o Complex models tend to change wildly with changes in the training data set
    o Simple models have low variance, high bias and complex models have low bias, high variance
    o Simpler models make more errors in the training set. Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples.

Therefore, to make the model more robust and generalizable, m**ake the model simple but not simpler** which will not be of any use.

Regularization:

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use.

For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, making a model simple lead to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance refers to the degree of changes in the model itself with respect to changes in the training data. **Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error.**