Assignment-based 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 double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

Answer: The optimal value of alpha for the ridge is 2 and the lasso is 0.0001.
The double value of alpha for the ridge is 4 and the lasso is 0.0002.
The following are the most important predictor variables for Ridge:

	Ridge After Doubled Alpha Value
Total_sqr_footage	0.149002
GarageArea	0.091904
TotRmsAbvGrd	0.068464
OverallCond	0.043421
LotArea	0.038783
Total_porch_sf	0.033754
CentralAir_Y	0.031951
LotFrontage	0.027438
Neighborhood_StoneBr	0.026545
OpenPorchSF	0.022839
MSSubClass_70	0.022118
Alley_Pave	0.021662
Neighborhood_Veenker	0.020244
BsmtQual_Ex	0.020020
KitchenQual_Ex	0.019799
HouseStyle_2.5Unf	0.019078
MasVnrType_Stone	0.018272
PavedDrive_P	0.017978
RoofMatl_WdShngl	0.017900
PavedDrive_Y	0.016811

The following are the most important predictor variable for Lasso:

	Lasso After Doubled Alpha Value
Total_sqr_footage	0.204561
GarageArea	0.103918
TotRmsAbvGrd	0.065118
OverallCond	0.042285
CentralAir_Y	0.033256
Total_porch_sf	0.030407
LotArea	0.025948
remodel_age	0.019040
BsmtQual_Ex	0.018200
Neighborhood_StoneBr	0.017116
OpenPorchSF	0.016846
Alley_Pave	0.016607
KitchenQual_Ex	0.016400
LandContour_HLS	0.014795
MSSubClass_70	0.014454
MasVnrType_Stone	0.013130
Condition1_Norm	0.012738
SaleCondition_Partial	0.011577
BsmtCond_TA	0.011522
LotConfig_CulDSac	0.008785

As the alpha value was very small, we see that there is not much change in the models and the R2, as well as MSE, remain the same. Also, it is noticed that the 'Central Air Conditioned - Yes' gains higher importance than 'Lot Area' in the Lasso Model although other important predictor variables remain the same.

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: Alpha values:

1. Ridge – 2
2. Lasso – 0.0001

MSE values:

1. Ridge - 0.0018423496270144383
2. Lasso - 0.0018672365777117732

As we see that there is no much difference in the MSE of ridge or lasso we can go ahead with lasso as the final model because it also helps in feature selection by making the coefficient value of features zero.

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: After creating the new lasso model post removing the 5 most important predictor variables we can see that below are the 5 most important predictors:

	Lasso 5 most important predictor after removal of top 5 predictors
LotFrontage	0.146470
Total_porch_sf	0.072205
HouseStyle_2.5Unf	0.063263
HouseStyle_2.5Fin	0.050827
Neighborhood_Veenker	0.042939

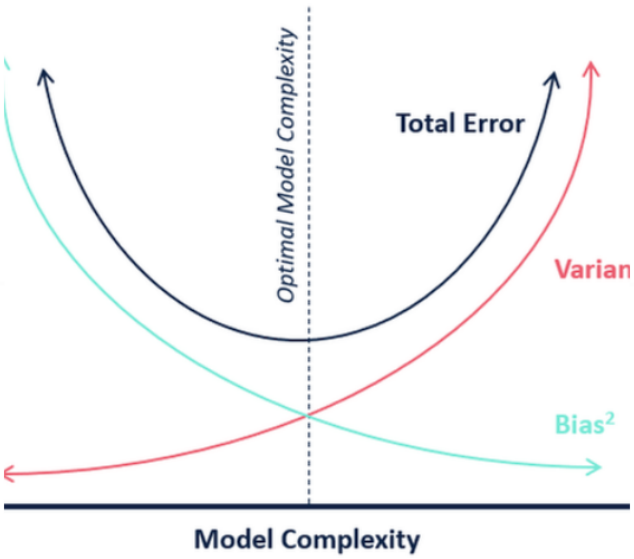
4. **How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?**

Answer: A model can be made robust and generalisable if the model performs on test data as good as it performs on the training data. As Occam's razor states that the model should be simple but not simpler which means it should find the balance and not overfit or underfit. Regularization is a very effective strategy to make the model simpler and avoid overfitting as we saw in this case study.

The accuracy of the model on the training data can decrease if we are regularizing it as it avoids overfitting.

However, the test accuracy improves as the model generalizes better which is what we want in the real world for our models to perform better on the unseen dataset.

The accuracy of the model is something that has to find a balance or sweet spot between bias and variance as it tries to minimize the error.


Optimal Model Complexity
Total Error
Varian
Bias²
Model Complexity