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?**

Please find the below metrics that we have got during the assignment:
Ridge Lambda optimal value is 9
Lasso Lambda optimal value is 100

Below are screenshots of Ridge and Lasso before doubling the optimal values
Ridge(left side) & Lasso(right side)

```
16]: {'OverallQual': 34027.87491648577,
      '1stFlrSF': 16858.23280527155,
      'GarageCars': 16202.506654263729,
      'MSSubClass_160': -14027.438080069098,
      'MSZoning_FV': 12792.731617339166,
      'MSZoning_RM': -20804.55729435751,
      'LotConfig_CulDSac': 18395.32216322096,
      'Neighborhood_Crawfor': 20722.354124886977,
      'Neighborhood_Edwards': -16366.320497412153,
      'Neighborhood_Mitchel': -21373.590493559997,
      'Neighborhood_NAmes': -18088.1451690103,
      'Neighborhood_NWAmes': -12265.302628205334,
      'Neighborhood_NoRidge': 53811.27742670077,
      'Neighborhood_NridgHt': 28646.315332687882,
      'Neighborhood_Sawyer': -12936.086497140692,
      'Neighborhood_StoneBr': 21297.897654886256,
      'RoofStyle_Mansard': 7591.780781928544,
      'Exterior1st_BrkComm': -5350.685441059089,
      'Exterior1st_BrkFace': 14472.905739098773,
      'Exterior1st_CemntBd': 22606.78032849823,
      'Exterior1st_ImStucc': -6979.039166808433,
      'Exterior1st_Plywood': 1752.569783366725,
      'Exterior2nd_HdBoard': 4615.505579603989,
      'Exterior2nd_ImStucc': 18045.329482136774,
      'Exterior2nd_MetalSd': 5019.83104349529,
      'Exterior2nd_VinylSd': 11697.316787879756,
      'Foundation_Slab': -13783.3558731061,
      'BsmtQual_Gd': -16027.122265886226,
      'BsmtExposure_Gd': 26486.337556082,
      'BsmtFinType1_Unf': -14548.919931159942,
      'GarageType_BuiltIn': 35582.440647233554,
```

```
[127]: {'OverallQual': 32775.81694989505,
        '1stFlrSF': 16437.035052929394,
        'GarageCars': 15930.310041864734,
        'MSSubClass_160': -15258.957186963195,
        'MSZoning_FV': 15972.820123647263,
        'MSZoning_RM': -21226.757106383633,
        'LotConfig_CulDSac': 18532.578479624295,
        'Neighborhood_Crawfor': 24150.284453820987,
        'Neighborhood_Edwards': -17244.17864085186,
        'Neighborhood_Mitchel': -23404.527630247976,
        'Neighborhood_NAmes': -18017.83651713426,
        'Neighborhood_NWAmes': -9812.910495881059,
        'Neighborhood_NoRidge': 70671.80618517699,
        'Neighborhood_NridgHt': 34489.2557169669,
        'Neighborhood_Sawyer': -12827.998408840685,
        'Neighborhood_StoneBr': 28283.279704836426,
        'RoofStyle_Mansard': 1936.5915759302327,
        'Exterior1st_BrkComm': -0.0,
        'Exterior1st_BrkFace': 15832.424589150292,
        'Exterior1st_CemntBd': 24787.949420583584,
        'Exterior1st_ImStucc': -0.0,
        'Exterior1st_Plywood': 974.3794629196695,
        'Exterior2nd_HdBoard': 3782.156919641046,
        'Exterior2nd_ImStucc': 24468.30648746192,
        'Exterior2nd_MetalSd': 4283.115857524713,
        'Exterior2nd_VinylSd': 11975.655135171493,
        'Foundation_Slab': -15769.727621333237,
        'BsmtQual_Gd': -16799.49575676876,
        'BsmtExposure_Gd': 28770.596267051,
        'BsmtFinType1_Unf': -14548.896552509756,
        'GarageType_BuiltIn': 38372.162655665976,
```

Below are screenshots of Ridge and Lasso after doubling the optimal values (. i.e., Ridge lambda as 2 and Lasso lambda as 200)

We can observe that the coefficients of few variables are increasing while few of them are reducing in magnitude, which results in overfitting because as we keep increasing the lambda value, the coefficients will move close to zero and it will be same as linear regression then and the same has to be avoided.

The tuning parameter lambda controls the impact on bias and variance. As the value of lambda rises, it reduces the value of coefficients and thus reducing the variance. Till a point, this increase in lambda is beneficial as it is only reducing the variance (hence avoiding overfitting), without losing any important properties in the data. But after a certain value, the model starts losing important properties, giving rise to bias in the model and thus underfitting. Therefore, the value of lambda should be carefully selected.

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?

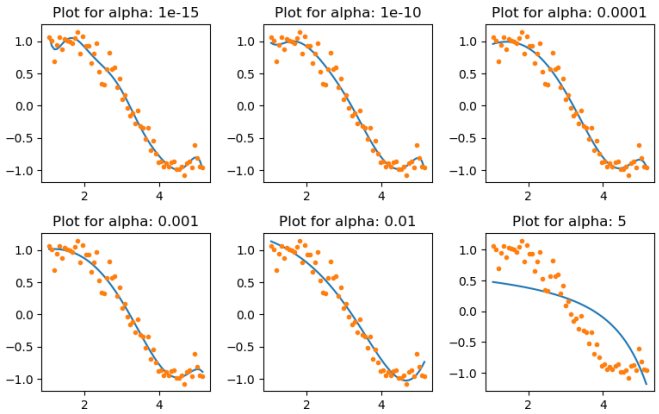
As we've seen, regularization is beneficial in the following circumstances:

There are two main issues with training models that we could run into: overfitting and underfitting.

When a model performs well on the training set but poorly on the unknown test data, this is known as overfitting.
When it performs poorly on both the test set and the train set, it is underfit.

Regularization reduces the magnitude of the coefficients while maintaining a constant number of training features.

Ridge Lambda's ideal value for us is 9, whereas for lasso it is 100.



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?

Below are the 5 most important predictors in case, when earlier one's doesn't exist in the inputdata:

1.Total House Age

2.2ndFlrSF

3.GarageYrBlt

4.bsmtEposure

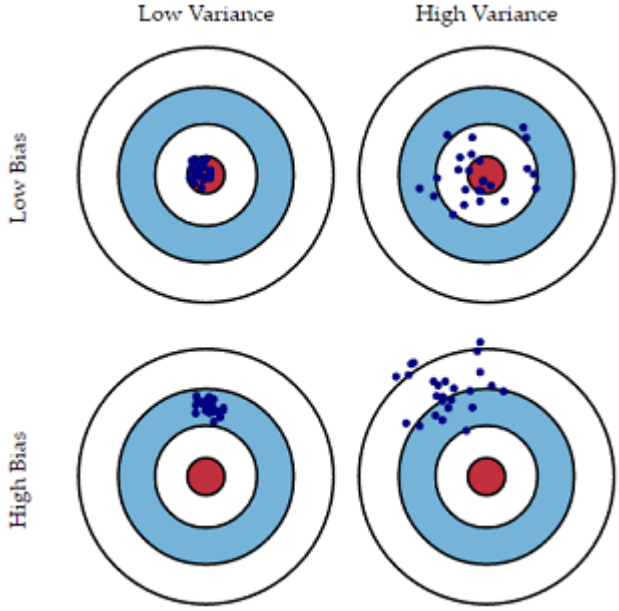
5.Exteroir1st

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

The implemented model should be kept as straightforward as feasible; while accuracy may suffer, the model will be more reliable and versatile.

Let's attempt to comprehend the bias-variance trade-off notion.

Bias: A model error that results from the model's inability to appropriately extract patterns from the data. Due to weak pattern recognition, the model performs poorly on both training and testing data, producing inaccurate results.



Variance: Variance is the result of a model trying to learn too much or too little from the data. High variance means the model performs incredibly well on training data since it was very well trained on training data, but it performs horribly on testing data.

To prevent overfitting and underfitting of data used for modeling, we must maintain a balance between bias and variance.

Therefore, it is always better to construct simpler models because, even though we may have more bias, the variance would be lower when applied to unobserved data, making the model more general.