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 1
Lasso Lambda optimal value is 100

Below are screenshots of Ridge and Lasso before doubling the optimal values

Ridge(left side) & Lasso(right side)

```
{'GrLivArea': 38432.23171739467,
 'MSSubClass_160': -40136.85612724523,
 'MSZoning_FV': 29175.500389792327,
 'LotShape_IR3': -30041.871930686782,
 'Neighborhood_NoRidge': 53047.08895184835,
 'Neighborhood_NridgHt': 61763.9155255193,
 'Neighborhood_StoneBr': 66605.03842014009,
 'HouseStyle_2.5Fin': -62138.52463593441,
 'Exterior1st_BrkComm': -37606.93623304896,
 'Exterior2nd_ImStucc': 35744.80354695586,
 'ExterQual_Gd': 18002.939247370312,
 'BsmtQual_Fa': -95857.44454374244,
 'BsmtQual_Gd': -57988.529795619215,
 'BsmtQual_NA': -103602.55190114137,
 'BsmtQual_TA': -79281.82709362333,
 'BsmtExposure_Gd': 35942.587649853616,
 'KitchenQual_Fa': -27243.8814777096}
```

```
{'GrLivArea': 38387.02058151359,
 'MSSubClass_160': -37316.21378448333,
 'MSZoning_FV': 26059.535558591422,
 'LotShape_IR3': -21094.186181945806,
 'Neighborhood_NoRidge': 51937.30959056557,
 'Neighborhood_NridgHt': 60786.178868601986,
 'Neighborhood_StoneBr': 64502.781721585845,
 'HouseStyle_2.5Fin': -54461.94490709967,
 'Exterior1st_BrkComm': -0.0,
 'Exterior2nd_ImStucc': 24754.503453591842,
 'ExterQual_Gd': 17943.91442348818,
 'BsmtQual_Fa': -96027.94072305528,
 'BsmtQual_Gd': -57674.60447437879,
 'BsmtQual_NA': -104090.97741397016,
 'BsmtQual_TA': -79595.66106585,
 'BsmtExposure_Gd': 35354.2601898522,
 'KitchenQual_Fa': -24271.174552296263}
```

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.

Ridge(left side) & Lasso (right side)

```
{'GrLivArea': 38790.35314463018,
 'MSSubClass_160': -38836.25140607578,
 'MSZoning_FV': 28029.399114943302,
 'LotShape_IR3': -26792.70883107708,
 'Neighborhood_NoRidge': 50861.45562898567,
 'Neighborhood_NridgHt': 62523.80040443886,
 'Neighborhood_StoneBr': 64101.53659754174,
 'HouseStyle_2.5Fin': -53218.11397673868,
 'Exterior1st_BrkComm': -25037.06251958483,
 'Exterior2nd_ImStucc': 32391.702394084587,
 'ExterQual_Gd': 18914.631569128804,
 'BsmtQual_Fa': -87980.8296903716,
 'BsmtQual_Gd': -53817.71047705766,
 'BsmtQual_NA': -95553.7665634428,
 'BsmtQual_TA': -74641.33059163962,
 'BsmtExposure_Gd': 36415.56386562139,
 'KitchenQual_Fa': -27204.617028477885}
```

```
{'GrLivArea': 38699.69972818102,
 'MSSubClass_160': -33196.75271087468,
 'MSZoning_FV': 21827.112959586506,
 'LotShape_IR3': -7854.749616189218,
 'Neighborhood_NoRidge': 48634.43737443285,
 'Neighborhood_NridgHt': 61153.6532851357,
 'Neighborhood_StoneBr': 59911.037583243124,
 'HouseStyle_2.5Fin': -34099.227596742705,
 'Exterior1st_BrkComm': -0.0,
 'Exterior2nd_ImStucc': 9533.601716158777,
 'ExterQual_Gd': 18948.100372965368,
 'BsmtQual_Fa': -86658.82272958137,
 'BsmtQual_Gd': -52316.542151666225,
 'BsmtQual_NA': -94871.53549418844,
 'BsmtQual_TA': -74167.5977718491,
 'BsmtExposure_Gd': 35413.04890677426,
 'KitchenQual_Fa': -21628.06702174387}
```

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 have seen Regularisation helps us in the below situations:

When it comes to training models, there are two major problems one we can encounter: **Overfitting** and **Underfitting**.

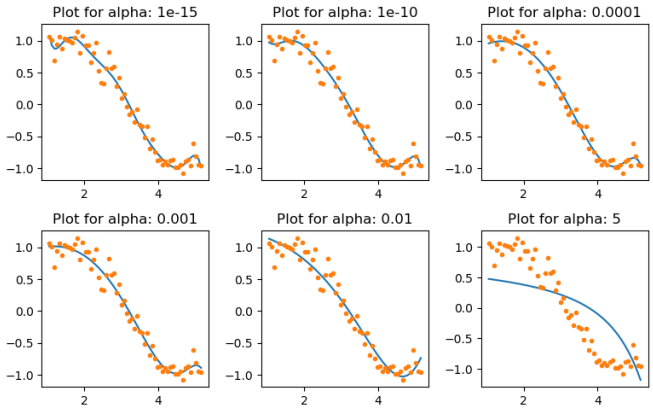
Overfitting happens when the model performs well on the training set but not so well on unseen test data.
Underfitting happens when it neither performs well on the train set nor on the test set.

With regularization, the number of features used in training is kept constant, yet the magnitude of the coefficients is reduced.

For us Ridge Lambda optimal value is 1 and for lasso it's 100.

As selecting a good value of lambda helps us maintain a good balance between bias and variance, based on the graphs plots for various lambda values we can decide on which has the expected output and based on that we can start picking up the lambda value.

The best lambda value can also be obtained by using the below method:
best_params_



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 input data:
1.TotalHouseAge
2.2ndFlrSF
3.GarageYrBlt
4.bsmtEposure
5.Exterior1st

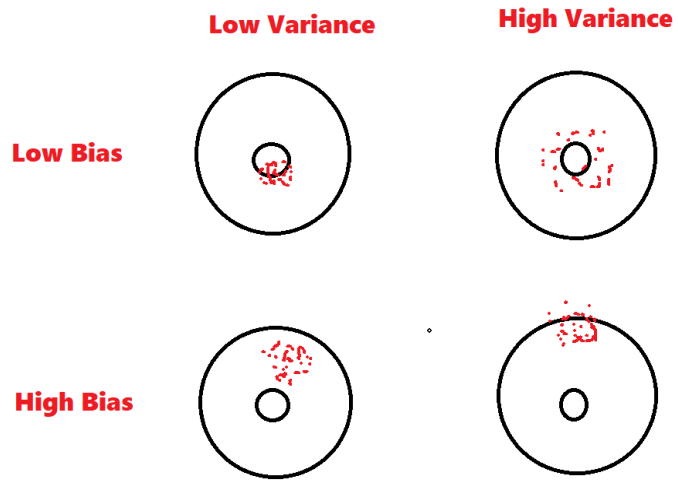
4.How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Implemented model should be kept as simple as possible, we can expect a decrease in accuracy but the model will be more robust and generalised.

Let's try to understand the concept of Bias-Variance trade-off

Bias: Bias is error in model, it can be explained as the model is unable to correctly learn pattern from the data. Model performs poor on training and testing data because it failed to learn the patterns well so gives out bad results which not accurate.

Variance: Variance can be explained as the situation when model tries to over learn or under learn from data. High variance indicates that the model performs exceptionally well on training data as it is very well trained on training data but performance is very poor on testing data



We have to maintain a balance between both Bias and Variance to avoid overfitting and underfitting of data being used for modelling
So, it is always better to build simpler models, though we might have more bias, the variance would be less when applied to the unseen data and so it would be more generalised model.

Finally, the implications would be that the model with simpler form has to be encouraged, as they lead to better accuracy when compared to complex model with hundreds of independent variables and complex equations.