Subjective Questions

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?

Ans:

- Optimal value of alpha for Ridge regression is 4 and for and Lasso regression is 0.0001
- Doubling the value of alpha for Ridge and Lasso regression does not have much impact on the overall model, as the value of alpha is relatively small even after doubling. Only few variable coefficients got changed

<table>
<tr><td></td><td>Original</td><td>Double</td></tr>
<tr><td>Scores Ridge</td><td>R2 Score (Train): 0.9177
R2 Score (Test): 0.91
MSE (Train): 0.0031
MSE (Test): 0.0035</td><td>R2 Score (Train): 0.9096
R2 Score (Test): 0.9028
MSE (Train): 0.0034
MSE (Test): 0.0037</td></tr>
<tr><td>Scores Lasso</td><td>R2 Score (Train): 0.921
R2 Score (Test): 0.9111
MSE (Train): 0.003
MSE (Test): 0.0034</td><td>R2 Score (Train): 0.9162
R2 Score (Test): 0.91
MSE (Train): 0.0032
MSE (Test): 0.0035</td></tr>
<tr><td>Coefficients Ridge</td><td>
<table>
<tr><th></th><th>Feaure</th><th>Coef</th></tr>
<tr><td>4</td><td>OverallCond</td><td>0.177514</td></tr>
<tr><td>14</td><td>BsmtFullBath</td><td>0.158376</td></tr>
<tr><td>12</td><td>2ndFlrSF</td><td>0.132886</td></tr>
<tr><td>9</td><td>BsmtUnfSF</td><td>0.106383</td></tr>
<tr><td>11</td><td>1stFlrSF</td><td>0.099609</td></tr>
<tr><td>3</td><td>OverallQual</td><td>0.097416</td></tr>
<tr><td>13</td><td>GrLivArea</td><td>0.07795</td></tr>
<tr><td>5</td><td>YearBuilt</td><td>0.072816</td></tr>
<tr><td>6</td><td>YearRemodAdd</td><td>0.068456</td></tr>
<tr><td>25</td><td>WoodDeckSF</td><td>0.066429</td></tr>
</table>
</td><td>
<table>
<tr><th></th><th>Feaure</th><th>Coef</th></tr>
<tr><td>4</td><td>OverallCond</td><td>0.14922</td></tr>
<tr><td>14</td><td>BsmtFullBath</td><td>0.133575</td></tr>
<tr><td>12</td><td>2ndFlrSF</td><td>0.112319</td></tr>
<tr><td>9</td><td>BsmtUnfSF</td><td>0.101844</td></tr>
<tr><td>11</td><td>1stFlrSF</td><td>0.087729</td></tr>
<tr><td>3</td><td>OverallQual</td><td>0.084024</td></tr>
<tr><td>25</td><td>WoodDeckSF</td><td>0.066441</td></tr>
<tr><td>13</td><td>GrLivArea</td><td>0.064381</td></tr>
<tr><td>24</td><td>GarageArea</td><td>0.064366</td></tr>
<tr><td>21</td><td>Fireplaces</td><td>0.059508</td></tr>
</table>
</td></tr>
</table>

Coefficient Lasso		Feaure	Coef		Feaure	Coef
	14	BsmtFullBath	0.412989	14	BsmtFullBath	0.434021
	4	OverallCond	0.241747	4	OverallCond	0.257981
	11	1stFlrSF	0.135308	11	1stFlrSF	0.116751
	6	YearRemodAdd	0.122261	6	YearRemodAdd	0.111505
	3	OverallQual	0.112172	3	OverallQual	0.104172
	5	YearBuilt	0.096480	9	BsmtUnfSF	0.093742
	9	BsmtUnfSF	0.091181	5	YearBuilt	0.090535
	24	GarageArea	0.068852	51	Neighborhood_Edwards	0.068406
	158	GarageType_Attchd	0.067541	24	GarageArea	0.062794
	67	Neighborhood_Timber	0.067411	61	Neighborhood_OldTown	0.058646

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?

Ans:

As R2 Score and Mean Square Error for both the Ridge and Lasso is the same, we would choose Lasso as it does feature reduction by making the coefficients as zero. This helps in simplifying the model

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?

Ans:

Top 5 predictor in the current Lasso model is

	Feaure	Coef
14	BsmtFullBath	0.412989
4	OverallCond	0.241747
11	1stFlrSF	0.135308
6	YearRemodAdd	0.122261
3	OverallQual	0.112172

After removing these predictor and rebuilding the Lasso model we get the next top 5 predictors as

	Feaure	Coef
10	TotalBsmtSF	0.490116
8	BsmtFinSF1	0.169520
4	OverallCond	0.122421
6	YearRemodAdd	0.103936
3	OverallQual	0.093818

With scores as

```
R2 Score (Train):  0.9067
R2 Score (Test):  0.8917
MSE (Train):  0.0035
MSE (Test):  0.0042
```

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?

Ans:

As per Occam's Razor, given two models that shows similar performance within finite set of training and test data, we should select a simpler model

Simpler models are more generic and require few training samples hence easier to train. It has low variance and higher bias

Complex model changes wildly with changes in a training data. It has low bias and higher variance. Complex model may lead to overfitting

Therefore, to make a model robust and generalizable, make the model simple but not simpler.

Regularization can be used to make the model simple but not too naïve. It makes sure that the model lies in between bias-variance trade-off