Assignment-based Subjective Questions

Answer for Question 1 –

For Lasso and Ridge, the ideal value of alpha is 0.0001 and 0.1, respectively.

The table of metrics for alpha and 2*alpha is shown below to show how the ridge and lasso models have changed over time.

The table below shows that between Ridge Regression (for alpha) and Ridge Regression 2 (for 2*alpha), there are no notable differences in the metrics values other than the RSS values. The results for Lasso Regression (for alpha) and Lasso Regression 2 (for 2*alpha) are also comparable.

When the alpha is larger, the RSS values have risen for both Ridge and Lasso.

	Metric	Linear Regression	Ridge Regression	Ridge Regression_2	Lasso Regression	Lasso Regression_2
0	R2 Score (Train)	0.866939	0.848032	0.832905	0.799707	0.784254
1	R2 Score (Test)	0.753414	0.773528	0.781570	0.780122	0.787400
2	RSS (Train)	1.637317	1.869966	2.056114	2.464610	2.654755
3	RSS (Test)	1.340397	1.231059	1.187348	1.195217	1.155655
4	MSE (Train)	0.040045	0.042796	0.044876	0.049132	0.050992
5	MSE (Test)	0.055320	0.053015	0.052066	0.052238	0.051366

The table below summarises the comparison of the Ridge & Lasso predictors for the two periods:

	Ridge	Ridge2	Lasso	Lasso2
LotArea	0.142824	0.137836	0.108932	7.009392e-02
OverallQual	0.289363	0.299267	0.319999	3.294737e-01
BsmtFinSF1	0.210838	0.192301	0.164018	1.487186e-01
TotalBsmtSF	0.246730	0.198390	0.088737	3.299499e-02
1stFlrSF	0.190310	0.196307	0.082993	1.171108e-01
2ndFlrSF	0.091859	0.083342	0.000000	0.000000e+00
LowQualFinSF	-0.033000	-0.034663	-0.047813	-3.755106e-02
GrLivArea	0.191167	0.192601	0.364314	3.382306e-01
MSSubClass_90	-0.019589	-0.018269	-0.026769	-2.237499e-02
Condition2_PosN	-0.537665	-0.474707	-0.427799	-2.980990e-01
BldgType_2fmCon	-0.035461	-0.034309	-0.028189	-2.151567e-02
BldgType_Duplex	-0.019589	-0.018269	-0.000449	-3.915159e-04
RoofMatl_CompShg	0.572179	0.390827	0.105851	2.990357e-02
RoofMatl_Membran	0.560986	0.362890	0.050939	0.000000e+00
RoofMatl_Metal	0.543066	0.346559	0.029342	0.000000e+00
RoofMatl_Roll	0.515678	0.321207	0.000000	0.000000e+00
RoofMatl_Tar&Grv	0.548480	0.365704	0.080785	0.000000e+00
RoofMatl_WdShake	0.533769	0.347554	0.050798	0.000000e+00
RoofMatl_WdShngl	0.663178	0.480103	0.205640	1.145397e-01
BsmtQual_none	0.017661	0.014436	0.017226	1.918569e-03
BsmtCond_none	0.017661	0.014436	0.000000	2.955205e-18
BsmtFinType1_none	0.017661	0.014436	0.000000	0.000000e+00
Heating_OthW	-0.147493	-0.136538	-0.068203	-0.000000e+00
BedroomAbvGr_8	-0.035184	-0.026839	-0.000000	-0.000000e+00
SaleType_New	0.022665	0.021760	0.040506	3.860928e-02
SaleCondition_Partial	0.022665	0.021760	0.000000	0.000000e+00

Prior to the update, Ridge's top 5 predictions were:

1. RoofMatl WdShngl 0.66
2. RoofMatl CompShg 0.57
3. RoofMatl Membran 0.56
4. RoofMatl Tar&Grv 0.55
5. RoofMatl_Metal 0.54

Top 5 indicators for Ridge following the modification are:

1. RoofMatl WdShngl 0.48
2. RoofMatl CompShg 0.39
3. RoofMatl Membran 0.36
4. RoofMatl Tar&Grv 0.37
5. RoofMatl Metal 0.35

Therefore, we may conclude that although the most significant predictor variables do not change when the alpha is doubled, the value of the coefficients in the model with 2*alpha for the Ridge Regression has decreased, thereby reducing the model complexity to address overfitting.

Prior to the adjustment, the top 5 predictors for Lasso were:

1. GrLivArea 0.36
2. sOverallQual 0.32
3. RoofMatl WdShngl 0.21
4. sBsmtFinSF1 0.16
5. sLotArea 0.11

Prior to the adjustment, the top 5 predictors for Lasso were:

1. GrLivArea 0.34
2. OverallQual 0.33
3. RoofMatl WdShngl 0.11
4. sBsmtFinSF1 0.15
5. sLotArea 0.07

Therefore, we may conclude that although the most significant predictor variables do not change when the alpha is doubled, the value of the coefficients in the model with 2*alpha for the Lasso Regression has decreased, thereby reducing the model complexity to address overfitting.

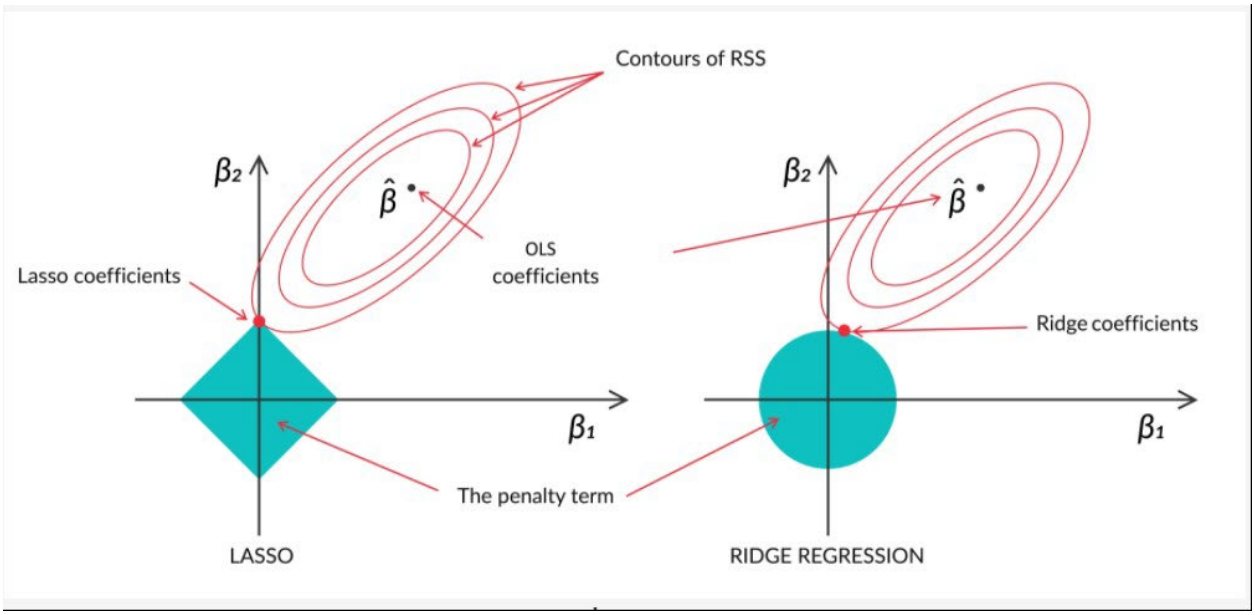
Answer for Question 2 –

Following is a list of some of the main distinctions between Ridge and Lasso:

- Ridge regression does have a clear drawback. It would be a part of the final model and contain every predictor. Even if it might not have an impact on the forecasts' accuracy, a large number of predictors can make model interpretation difficult.
- The Lasso decreases the coefficient estimate toward zero, similar to Ridge regression.
- The best choice of the lambda value is essential for Lasso regression because:
    - The penalty in Lasso forces some of the coefficient estimates to be exactly equal to zero, performing feature scaling;
    - Models produced by Lasso are typically easier to interpret than those produced by Ridge regression
    - In addition, in Lasso, as the lambda increases, the variance decreases and the bias increases, moving it from being overfitting to underfitting.

Their penalty term is the main distinction between Lasso and Ridge regression. The total absolute values of all the model's coefficients make up the penalty term in this case. Lasso regression lowers the coefficient estimates toward 0 like Ridge regression does. There is, however, one distinction. With Lasso, if the tuning parameter lambda, is large enough, the penalty forces some of the coefficient estimations to be exactly 0.

Lasso conducts feature selection as a result. Here, too, picking the right lambda value is crucial. As a result, models created by Lasso are simpler to interpret than models created by Ridge. Additionally, standardisation of variables is required for Lasso just as it is for Ridge regression.



Ridge and Lasso's optimal alpha values for this assignment were found using the grid search to be - Lasso Regression: 0.0001 and Ridge Regression: 0.10

The table below lists the R2 square, RSS, and RMSE values that were determined for the various models.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.866939	0.848032	0.799707
1	R2 Score (Test)	0.753414	0.773528	0.780122
2	RSS (Train)	1.637317	1.869966	2.464610
3	RSS (Test)	1.340397	1.231059	1.195217
4	MSE (Train)	0.040045	0.042796	0.049132
5	MSE (Test)	0.055320	0.053015	0.052238

The coefficient values for the predictor variables are shown in the following table.

	Linear	Ridge	Lasso
LotArea	0.146960	0.142824	0.108932
OverallQual	0.258376	0.289363	0.319999
BsmtFinSF1	0.260282	0.210838	0.164018
TotalBsmtSF	0.402327	0.246730	0.088737
1stFlrSF	0.158413	0.190310	0.082993
2ndFlrSF	0.114345	0.091859	0.000000
LowQualFinSF	-0.029342	-0.033000	-0.047813
GrLivArea	0.189714	0.191167	0.364314
MSSubClass_90	-0.023549	-0.019589	-0.026769
Condition2_PosN	-0.649300	-0.537665	-0.427799
BldgType_2fmCon	-0.038184	-0.035461	-0.028189
BldgType_Duplex	-0.023549	-0.019589	-0.000449
RoofMatl_CompShg	1.111425	0.572179	0.105851
RoofMatl_Membran	1.158233	0.560986	0.050939
RoofMatl_Metal	1.137692	0.543066	0.029342
RoofMatl_Roll	1.107797	0.515678	0.000000
RoofMatl_Tar&Grv	1.092990	0.548480	0.080785
RoofMatl_WdShake	1.091415	0.533769	0.050798
RoofMatl_WdShngl	1.205003	0.663178	0.205640
BsmtQual_none	0.027609	0.017661	0.017226
BsmtCond_none	0.027609	0.017661	0.000000
BsmtFinType1_none	0.027609	0.017661	0.000000
Heating_OthW	-0.156225	-0.147493	-0.068203
BedroomAbvGr_8	-0.056839	-0.035184	-0.000000
SaleType_New	0.025116	0.022665	0.040506
SaleCondition_Partial	0.025116	0.022665	0.000000

Conclusion:

1. Looking at the metrics table, the R2 square for the regressions using Lasso and Ridge on the train and test data indicates that Lasso fits the data better than Ridge, which is an overfit.
2. In the test set, Lasso outperformed Ridge in terms of RSS and RMSE values.
3. When examining the coefficients table, it becomes clear that the Lasso performed feature selection and assisted in identifying a smaller collection of the most crucial predictor variables than Ridge.
4. Given the information above, it is clear that the Lasso model performs better. As a result, I will select this model, which has an alpha value of 0.0001 that was acquired using GridSearch.

Answer for Question 3 –

The coefficient values for the Lasso model's predictor variables are shown in the following table.

List of Features	Coefficients
GrLivArea	0.364
OverallQual	0.320
RoofMatl_WdShngl	0.206
BsmtFinSF1	0.164
LotArea	0.109
RoofMatl_CompShg	0.106
TotalBsmtSF	0.089
1stFlrSF	0.083
RoofMatl_Tar&Grv	0.081
RoofMatl_Membran	0.051
RoofMatl_WdShake	0.051
SaleType_New	0.041
RoofMatl_Metal	0.029
BsmtQual_none	0.017
2ndFlrSF	0.000
RoofMatl_Roll	0.000
BsmtCond_none	0.000
BsmtFinType1_none	0.000
BedroomAbvGr_8	0.000
SaleCondition_Partial	0.000
BldgType_Duplex	0.000
MSSubClass_90	-0.027
BldgType_2fmCon	-0.028
LowQualFinSF	-0.048
Heating_OthW	-0.068
Condition2_PosN	-0.428

The top 5 features the model predicted, according on our observations, are those that are stated below.

List of Features	Coefficients
GrLivArea	0.364
OverallQual	0.320
RoofMatl_WdShngl	0.206
BsmtFinSF1	0.164
LotArea	0.109
RoofMatl_CompShg	0.106

The essential scores for the next metric are also provided, which aids in our understanding of how the model is doing.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.866939	0.848032	0.799707
1	R2 Score (Test)	0.753414	0.773528	0.780122
2	RSS (Train)	1.637317	1.869966	2.464610
3	RSS (Test)	1.340397	1.231059	1.195217
4	MSE (Train)	0.040045	0.042796	0.049132
5	MSE (Test)	0.055320	0.053015	0.052238

Now, in accordance with the question, the top 5 variables are not present in the incoming data, and we must determine what the top 5 variables will be in their place.

After removing the top 5 variables from the predictor variable data frame, we built a new model, ran the analysis, and made the observations listed below. Following the elimination of the top 5 variables from the previous Lasso model, the coefficients obtained for the predictor variables are listed below.

List of Features	Coefficients
1stFlrSF	0.521
TotalBsmtSF	0.444
2ndFlrSF	0.238
SaleType_New	0.062
BsmtQual_none	0.029
RoofMatl_CompShg	0.020
RoofMatl_Membran	0.000
RoofMatl_Metal	0.000
RoofMatl_Roll	0.000
RoofMatl_Tar&Grv	0.000
RoofMatl_WdShake	0.000
BsmtCond_none	0.000
BsmtFinType1_none	0.000
BedroomAbvGr_8	0.000
SaleCondition_Partial	0.000
BldgType_Duplex	-0.002
Heating_OthW	-0.003
LowQualFinSF	-0.053
BldgType_2fmCon	-0.057
MSSubClass_90	-0.069
Condition2_PosN	-0.480

Here is a list of the new top 5 predictor variables predicted by the updated model.

List of Features	Coefficients
1stFlrSF	0.521
TotalBsmtSF	0.444
2ndFlrSF	0.238
SaleType_New	0.062
BsmtQual_none	0.029

What we have seen is that the order of the new top five variables has altered from the prior model; that is, the variables from 6 to 10 are not the same.

Additionally, we saw from the subsequent statistic that the new model's model accuracy was much worse than the old model.

```
R2 score_train:  0.6726215321554745
R2 score_test:  0.6494544531225319
RSS score_train:  4.028401363174951
RSS score_test:  1.905502920459972
RMSE score_train:  0.003945544919857934
RMSE score_test:  0.004350463288721397
```

The top 5 factors from the earlier model, which are not present in the incoming data, were indeed the most important predictors of the sale price, as this confirms.

Answer for Question 4 –

Overfitting, or a model that is overly complex, is a prevalent issue in machine learning. When given training data, these models perform very well, but when faced with real data, they perform poorly. These models are known to have low bias and high variance. Low bias models are particularly robust because they do not make a lot of errors in training data. Substantial variance because the mistakes in unobserved data will be very high and therefore not generalizable.

When the model is blatantly straightforward, the inverse is also true; this is the underfitting scenario. In this case, the training set's mistakes will be quite high (high bias), yet the model won't perform well in an unknown data set since it hasn't picked up the pattern from the training set (low variance). The model can be generalised in this situation.

To solve this problem, it's critical that we can create a durable and generalizable model by finding the ideal balance between overfitting and underfitting.

Here comes regularisation, where the loss function of the algorithm is modified by the addition of a penalty term. The weights of the model are altered as a result of minimising the loss function.

The three most used regularisation methods are Elastic Net, Lasso, and Ridge.

As a result, we can draw the conclusion that regularisation is the method by which we can make the model reliable and generalizable and avoid overfitting the train dataset.

Regularization greatly lowers the model's variance while only slightly increasing its bias. The impact on bias and variance is therefore controlled by the tuning parameter, which is employed in the regularisation procedures discussed above. Because of this, balancing the accuracy of the train and test data sets will help to improve the model's overall accuracy.