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?

Answer:

- **For Ridge Regression,**

  The Optimal value of alpha for ridge regression is 5.0.

  If we choose to double the value of alpha for both Ridge and Lasso, the regularization penalty will become stronger, which will lead to higher shrinkage of the coefficient values. As a result, the model will become more conservative and simpler, with lower variance but potentially higher bias. Specifically, the effect of doubling the value of alpha will be:

  Regarding Rige regression, the coefficient values will be shrunk more towards zero, resulting in a flatter slope of the regression line. This will decrease the flexibility of the model and make it less prone to overfitting. The optimal value of alpha for Ridge regression is usually determined by cross-validation, and a value that is too high can lead to underfitting, which can result in poor predictive performance. The double value of the optimal lamda is 10.

```
Coefficients for alpha = 10.0
              Feature   Coefficient
93   RoofMatl_CompShg  50585.924741
5           GrLivArea  32156.497381
97   RoofMatl_Tar&Grv  31986.150848
99   RoofMatl_WdShngl  28832.946105
98   RoofMatl_WdShake  20833.369908
..                ...           ...
171    KitchenQual_TA  -8539.037110
170    KitchenQual_Gd  -8860.214831
142       BsmtQual_Gd  -9235.741749
143       BsmtQual_TA  -9416.572904
64    Condition2_PosN -14089.311209

[212 rows x 2 columns]
```

The most important predictor variables after the change is implemented:

```
Most important predictor variables:
93     RoofMatl_CompShg
5             GrLivArea
97     RoofMatl_Tar&Grv
99     RoofMatl_WdShngl
98     RoofMatl_WdShake
           ...
171      KitchenQual_TA
170      KitchenQual_Gd
142         BsmtQual_Gd
143         BsmtQual_TA
64      Condition2_PosN
Name: Feature, Length: 209, dtype: object
```

- Lasso Regression:

The optimal value of alpha for lasso regression is 500.

Similar to Ridge regression, doubling the value of alpha will increase the strength of the regularization penalty, resulting in higher shrinkage of the coefficient values.

However, Lasso regression has the additional property of performing feature selection, which means that it can set some of the coefficient values exactly to zero, effectively eliminating the corresponding predictors from the model. Therefore, doubling the value of alpha in Lasso regression may lead to more feature selection and a simpler model. If the optimal value of alpha is already very high, then doubling it further may lead to underfitting and poor performance.

The double value of the optimal lamda is 1000. Therefore, the coefficients as as follows:

```
Coefficients for alpha = 1000
              Feature   Coefficient
5           GrLivArea  26606.817519
86      OverallQual_9  14521.521643
87     OverallQual_10  12582.283293
85      OverallQual_8  10957.473001
10         GarageCars   8839.623172
..                ...           ...
69    BldgType_Duplex  -2809.252495
131      ExterQual_TA  -2903.888445
188  GarageFinish_Unf  -3118.286251
154  BsmtFinType1_Unf  -3142.787188
64    Condition2_PosN -11227.527134

[212 rows x 2 columns]
```

The most important predictor variables after the change is implemented:

```
Most important predictor variables:
5                 GrLivArea
86            OverallQual_9
87           OverallQual_10
85            OverallQual_8
10               GarageCars
99          RoofMatl_WdShngl
147          BsmtExposure_Gd
43     Neighborhood_NoRidge
0               OverallCond
204            SaleType_New
44     Neighborhood_NridgHt
151         BsmtFinType1_GLQ
34     Neighborhood_Crawfor
2                TotalBsmtSF
93          RoofMatl_CompShg
84             OverallQual_7
49     Neighborhood_Somerst
9                 Fireplaces
23         LotConfig_CulDSac
6                    FullBath
```

```
73         HouseStyle_1Story
164              CentralAir_Y
14                MSZoning_RL
126     Exterior2nd_VinylSd
102     Exterior1st_BrkFace
156            Heating_GasW
32     Neighborhood_ClearCr
146              BsmtCond_TA
28             LandSlope_Sev
137         Foundation_PConc
120     Exterior2nd_ImStucc
55           Condition1_PosA
22           LandContour_Lvl
7                   HalfBath
179          FireplaceQu_Gd
52     Neighborhood_Veenker
118     Exterior2nd_CmentBd
98          RoofMatl_WdShake
94          RoofMatl_Membran
17              LotShape_IR2
21           LandContour_Low
200             SaleType_Con
```

```
50     Neighborhood_StoneBr
116     Exterior2nd_BrkFace
62          Condition2_Norm
168        Electrical_SBrkr
106     Exterior1st_ImStucc
15              MSZoning_RM
209    SaleCondition_Family
38     Neighborhood_MeadowV
185      GarageType_CarPort
40       Neighborhood_NAmes
176          Functional_Sev
57          Condition1_RRAe
161            HeatingQC_Gd
37      Neighborhood_IDOTRR
110      Exterior1st_Stucco
81            OverallQual_4
189           GarageQual_Fa
138         Foundation_Slab
142             BsmtQual_Gd
163           HeatingQC_TA
75        HouseStyle_2.5Unf
129           ExterQual_Fa
186       GarageType_Detchd
88          RoofStyle_Gable
158           Heating_OthW
143            BsmtQual_TA
171          KitchenQual_TA
128     Exterior2nd_Wd Shng
18              LotShape_IR3
4               LowQualFinSF
70           BldgType_Twnhs
68          BldgType_2fmCon
71          BldgType_TwnhsE
53         Condition1_Feedr
149         BsmtExposure_No
125      Exterior2nd_Stucco
45     Neighborhood_OldTown
35     Neighborhood_Edwards
69           BldgType_Duplex
131            ExterQual_TA
188        GarageFinish_Unf
154        BsmtFinType1_Unf
64          Condition2_PosN
Name: Feature, dtype: object
```

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?

Answer:

We create a table which contain all the metrics as follows:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.484219e-01	9.455464e-01	9.261185e-01
1	R2 Score (Test)	-3.391040e+22	8.283313e-01	8.414244e-01
2	RSS (Train)	3.291046e+11	3.474517e+11	4.714153e+11
3	RSS (Test)	9.558368e+34	4.838847e+11	4.469790e+11
4	MSE (Train)	1.795371e+04	1.844737e+04	2.148765e+04
5	MSE (Test)	1.477253e+16	3.323792e+04	3.194526e+04

The table aove shows the performance metrics for the three regression models: Linear Regression, Ridge Regression, and Lasso Regression. The performance metrics used are R2 Score, Residual Sum of Squares (RSS), and Mean Squared Error (MSE). The R2 Score is a measure of how well the model fits the data, with a maximum value of 1.0 indicating a perfect fit. The RSS and MSE measure the difference between the predicted values and the actual values, with lower values indicating better performance.

From the table, we can see that the Linear Regression model has the highest R2 Score on the training data, indicating the best fit. However, on the test data, the model performs poorly with an extremely low R2 Score and very high RSS and MSE values. This suggests that the model is overfitting the training data and does not generalize well to new data.

The Ridge Regression and Lasso Regression models perform better on the test data, with higher R2 Scores and lower RSS and MSE values. This is because these models use regularization to prevent overfitting, which helps them to generalize better to new data. The Lasso Regression model performs slightly better than the Ridge Regression model in terms of R2 Score and MSE on the test data, but the Ridge Regression model has a slightly lower RSS.

A table of some coefficient values for Linear, Ridge and Lasso regularization as follows:

	Linear	Ridge	Lasso
OverallCond	6.773692e+03	6402.957544	6010.341099
BsmtUnfSF	-7.733017e+03	-5875.421462	-3253.526955
TotalBsmtSF	1.955298e+04	14364.323928	8682.750816
1stFlrSF	-7.600989e+03	-4931.324842	0.000000
LowQualFinSF	-2.820525e+03	-2632.603537	-1996.025268
GrLivArea	4.007838e+04	35491.665748	28209.095727
FullBath	1.849859e+03	2808.690480	2352.279252
HalfBath	1.892424e+02	741.781745	841.760035
TotRmsAbvGrd	-2.601749e+03	-1552.390080	-0.000000
Fireplaces	4.856562e+02	836.561508	2446.013247
GarageCars	2.260336e+03	4583.072990	6940.915501
MSZoning_FV	7.211762e+03	4807.459088	0.000000
MSZoning_RH	4.531708e+03	3257.620284	0.000000
MSZoning_RL	1.627247e+04	11242.268690	1529.866863
MSZoning_RM	1.314944e+04	9302.925160	-0.000000
Street_Pave	-6.554423e+02	-471.598820	-0.000000
LotShape_IR2	9.795724e+02	900.163118	659.754084
LotShape_IR3	1.524410e+03	437.729795	-218.084236
LotShape_Reg	-1.199496e+02	-84.362307	-0.000000
LandContour_HLS	-1.011737e+02	536.649919	0.000000
LandContour_Low	4.639210e+02	988.608886	501.597813
LandContour_Lvl	2.084567e+03	2969.260392	1152.580480
LotConfig_CulDSac	1.435778e+03	1677.249084	2185.650773

LotConfig_FR3	1.616541e+02	-9.766732	-0.000000
LotConfig_Inside	-8.605518e+02	-946.787368	-104.129674
LandSlope_Mod	7.759871e+02	850.742419	1.839374
LandSlope_Sev	9.302222e+02	1418.671882	1256.196069
Neighborhood_Blueste	1.936058e+02	136.613980	-0.000000
Neighborhood_BrDale	8.928426e+02	513.193639	0.000000
Neighborhood_BrkSide	1.531796e+03	556.994378	-0.000000
Neighborhood_ClearCr	1.376150e+03	1190.705273	1203.754427
Neighborhood_CollgCr	2.316718e+03	1693.026141	200.211553
Neighborhood_Crawfor	5.622603e+03	5366.450958	4069.018545
Neighborhood_Edwards	-4.755385e+02	-1587.448185	-2385.243510
Neighborhood_Gilbert	2.324178e+03	1602.044509	0.000000
Neighborhood_IDOTRR	1.019217e+03	71.273750	-657.137080
Neighborhood_MeadowV	-2.142409e+03	-2316.876451	-1266.811734
Neighborhood_Mitchel	3.402707e+01	-547.520504	-171.280468
Neighborhood_NAmes	1.559787e+02	-694.094629	-845.430806
Neighborhood_NPkVill	1.112676e+03	836.345423	0.000000
Neighborhood_NWAmes	7.352985e+02	181.119046	-0.000000
Neighborhood_NoRidge	5.331821e+03	5862.351980	5506.805365
Neighborhood_NridgHt	3.997626e+03	4364.571044	4154.778399
Neighborhood_OldTown	-4.550958e+02	-1558.405126	-2459.967722
Neighborhood_SWISU	6.547679e+02	661.146810	-0.000000
Neighborhood_Sawyer	6.476532e+02	15.279772	-184.535094
Neighborhood_SawyerW	2.154074e+03	1783.335154	323.638371
Neighborhood_Somerst	5.654899e+03	5250.451214	3605.350298
Neighborhood_StoneBr	1.843067e+03	2011.361923	854.880612

Neighborhood_Timber	1.146250e+03	725.664368	0.000000
Neighborhood_Veenker	1.461481e+03	1395.212867	657.310506
Condition1_Feedr	-1.541075e+02	-740.676743	-1382.752052
Condition1_Norm	4.502880e+03	4243.303886	2867.249061
Condition1_PosA	1.186070e+03	1350.165292	861.539527
Condition1_PosN	5.914784e+02	681.044993	0.000000
Condition1_RRAe	-3.721969e+02	-511.067669	-565.833278
Condition1_RRAn	9.654489e+02	1169.573924	500.782493
Condition1_RRNe	1.705699e+02	71.910767	-0.000000
Condition1_RRNn	-4.429563e+02	-281.158043	-0.000000
Condition2_Feedr	1.142947e+03	1054.439377	-0.000000
Condition2_Norm	1.935050e+03	1671.822537	560.021202
Condition2_PosA	-1.997256e+02	85.699427	-0.000000

Looking at the coefficient values, we can see that regularization has shrunk the coefficient values towards zero for both Ridge and Lasso Regression models. The Lasso Regression model has even set some of the coefficients to zero, effectively performing feature selection.

Finally, we can see that the Ridge Regression and Lasso Regression models have significantly smaller coefficients compared to the Linear Regression model, which indicates that regularization is indeed helping to prevent overfitting by shrinking the coefficients towards zero. In particular, the Lasso Regression model is able to reduce many coefficients to exactly zero, effectively performing feature selection and identifying the most important features for predicting the target variable.

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?

Answer:

To create a new model excluding the five most important predictor variables, we need to first identify these variables. Based on the Lasso model with alpha=500, we can see that the five most important predictor variables (with the largest absolute coefficients) are:

['RoofMatl_CompShg', 'GrLivArea', 'RoofMatl_Tar&Grv', 'RoofMatl_WdShngl', 'RoofMatl_WdShake']

To create a new model excluding these variables, we can simply remove them from the feature set used to train the model.

To identify the five most important predictor variables in the new model, we can fit the Lasso model on the updated X_train dataset and extract the coefficients using the .coef_ attribute. We can then sort the coefficients in descending order and select the top five variables.

Please find Question 3 in the jupyter HousePricePrediction for the python code.

Now, the five most important predictor variables are:

['OverallQual_9', 'OverallQual_10', 'OverallQual_8', '1stFlrSF', 'Condition2_PosN']

	Variable	Coef
85	OverallQual_9	14888.508652
86	OverallQual_10	14477.233964
84	OverallQual_8	12219.839257
3	1stFlrSF	11212.237075
63	Condition2_PosN	-11101.839424

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?

Answer:
To make sure that a model is robust and generalizable, we can perform the following steps:

- **Splitting the data:** Splitting the dataset into training and testing sets is a standard practice. The model is trained on the training data and validated on the testing data. This helps to ensure that the model does not overfit the training data and can be generalized to new data.
- **Cross-validation:** Cross-validation is another technique that can be used to evaluate the model's performance. In this technique, the dataset is divided into k-folds, and the model is trained and validated on each of these folds. This helps to ensure that the model is robust and can perform well on new data.
- **Regularization techniques** such as Ridge and Lasso regression can help to prevent overfitting by adding a penalty term to the model's cost function. This penalty term helps to reduce the magnitude of the model coefficients, which in turn reduces the model's sensitivity to changes in the training data and improves its ability to generalize to new data.
- **Feature selection** is another technique that can help to improve a model's robustness and generalizability. By selecting only the most important features, we can reduce the complexity of the model and avoid overfitting to noise in the data.

The implications of having a robust and generalizable model are that it will perform well not only on the training data but also on new, unseen data. This means that the model will be more reliable in making predictions and can be used with confidence in real-world applications. However, it is important to note that improving the robustness and generalizability of a model may come at the cost of some reduction in accuracy, as the model is designed to be more conservative in its predictions to avoid overfitting to the training data.