<u>Question 1</u>

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?

<u>Answer</u>:

The optimal value of alpha is as follows:

<u>Ridge</u> – 2

<u>Lasso</u> – 50

The evaluation metric comparison for linear, ridge and lasso is as follows:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.507101e-01	9.440936e-01	9.388782e-01
1	R2 Score (Test)	-3.011816e+19	9.029818e-01	9.120512e-01
2	RSS (Train)	9.995070e+10	1.133676e+11	1.239436e+11
3	RSS (Test)	2.796918e+31	9.009579e+10	8.167350e+10
4	MSE (Train)	1.176588e+04	1.253071e+04	1.310218e+04
5	MSE (Test)	3.003717e+14	1.704792e+04	1.623154e+04

The top 20 predictor variables for Ridge are as follows:

In [138]:
```
# Lets look at the top 20 contributing factors for Ridge Regression Model (Both Positive and Neg correlation)
betas.sort_values(by = 'Ridge_abs', ascending = False).head(20)
```

Out[138]:

	Linear	Ridge	Lasso	Ridge_Pos_Neg_Corr	Lasso_Pos_Neg_Corr	Ridge_abs	Lasso_abs
GrLivArea	9.981496e+04	49533.091117	115081.097021	Pos	Pos	49533.091117	115081.097021
1stFlrSF	1.328171e+04	41439.399288	16411.590244	Pos	Pos	41439.399288	16411.590244
OverallQual_9	-6.884137e+14	28423.912540	47018.660851	Pos	Pos	28423.912540	47018.660851
2ndFlrSF	1.471365e+04	25047.112029	0.000000	Pos	Pos	25047.112029	0.000000
BsmtFinSF1	3.045451e+04	22630.694816	17206.296089	Pos	Pos	22630.694816	17206.296089
GarageArea	1.915356e+04	18690.368671	17763.786014	Pos	Pos	18690.368671	17763.786014
age_since_built	-3.074266e+04	-18323.170135	-33582.045463	Neg	Neg	18323.170135	33582.045463
BsmtQual_TA	-1.771620e+04	-17759.035322	-17946.336150	Neg	Neg	17759.035322	17946.336150
OverallQual_8	-6.884137e+14	16754.562940	26800.496768	Pos	Pos	16754.562940	26800.496768
Neighborhood_Crawfor	1.809021e+04	16644.850355	21150.244830	Pos	Pos	16644.850355	21150.244830
BsmtQual_Gd	-1.534889e+04	-16310.476936	-15783.870831	Neg	Neg	16310.476936	15783.870831
Neighborhood_Somerst	1.946633e+04	15387.626478	17967.770849	Pos	Pos	15387.626478	17967.770849
KitchenQual_Fa	-2.141658e+04	-13999.992042	-15048.892696	Neg	Neg	13999.992042	15048.892696
LotArea	1.680005e+04	13843.432151	13094.600959	Pos	Pos	13843.432151	13094.600959
Functional_Mod	-3.740040e+04	-13735.036488	-11087.898958	Neg	Neg	13735.036488	11087.898958
Exterior1st_BrkFace	-3.015000e+03	13570.456293	13462.356728	Pos	Pos	13570.456293	13462.356728
KitchenQual_TA	-1.858391e+04	-13476.392876	-13897.455456	Neg	Neg	13476.392876	13897.455456
BsmtExposure_Gd	1.340064e+04	13350.575858	13233.263642	Pos	Pos	13350.575858	13233.263642
OverallQual_10	-6.884137e+14	12418.400595	23930.272663	Pos	Pos	12418.400595	23930.272663
OverallQual_3	-6.884137e+14	-12017.117651	-6740.929714	Neg	Neg	12017.117651	6740.929714

The top 20 predictor variables for Ridge are as follows:

```
In [139]: betas.sort_values(by = 'Lasso_abs', ascending = False).head(20)
Out[139]:
```

	Linear	Ridge	Lasso	Ridge_Pos_Neg_Corr	Lasso_Pos_Neg_Corr	Ridge_abs	Lasso_abs
GrLivArea	9.981496e+04	49533.091117	115081.097021	Pos	Pos	49533.091117	115081.097021
OverallQual_9	-6.884137e+14	28423.912540	47018.660851	Pos	Pos	28423.912540	47018.660851
age_since_built	-3.074266e+04	-18323.170135	-33582.045463	Neg	Neg	18323.170135	33582.045463
OverallQual_8	-6.884137e+14	16754.562940	26800.496768	Pos	Pos	16754.562940	26800.496768
OverallQual_10	-6.884137e+14	12418.400595	23930.272663	Pos	Pos	12418.400595	23930.272663
Neighborhood_Crawfor	1.809021e+04	16644.850355	21150.244830	Pos	Pos	16644.850355	21150.244830
Neighborhood_Somerst	1.946633e+04	15387.626478	17967.770849	Pos	Pos	15387.626478	17967.770849
BsmtQual_TA	-1.771620e+04	-17759.035322	-17946.336150	Neg	Neg	17759.035322	17946.336150
GarageArea	1.915356e+04	18690.368671	17763.786014	Pos	Pos	18690.368671	17763.786014
BsmtFinSF1	3.045451e+04	22630.694816	17206.296089	Pos	Pos	22630.694816	17206.296089
MSSubClass_90	-1.227239e+15	-8823.378463	-16881.984295	Neg	Neg	8823.378463	16881.984295
1stFlrSF	1.328171e+04	41439.399288	16411.590244	Pos	Pos	41439.399288	16411.590244
BsmtQual_Gd	-1.534889e+04	-16310.476936	-15783.870831	Neg	Neg	16310.476936	15783.870831
MSSubClass_160	-1.455310e+04	-10992.059001	-15697.880903	Neg	Neg	10992.059001	15697.880903
KitchenQual_Fa	-2.141658e+04	-13999.992042	-15048.892696	Neg	Neg	13999.992042	15048.892696
OverallCond_9	2.492287e+15	11426.297177	14753.544834	Pos	Pos	11426.297177	14753.544834
KitchenQual_TA	-1.858391e+04	-13476.392876	-13897.455456	Neg	Neg	13476.392876	13897.455456
Exterior1st_BrkFace	-3.015000e+03	13570.456293	13462.356728	Pos	Pos	13570.456293	13462.356728
Functional_Typ	-6.659961e+03	7220.185767	13338.833183	Pos	Pos	7220.185767	13338.833183
BsmtExposure_Gd	1.340064e+04	13350.575858	13233.263642	Pos	Pos	13350.575858	13233.263642

As we can see there is a significant improvement in Ridge and Lasso with comparison to Linear Model and out of Ridge and Lasso, Lasso has performed slightly better maybe since not all predictors are important for predicting the Sale Price.

Also the top 20 predictors for Ridge and Lasso are almost the same.

If we double the alpha values for Ridge and Lasso:

Ridge – 4

Lasso – 100

The new evaluation metric comparison for linear, ridge and lasso is as follows:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.507101e-01	9.389447e-01	9.304224e-01
1	R2 Score (Test)	-3.011816e+19	9.013587e-01	9.126594e-01
2	RSS (Train)	9.995070e+10	1.238087e+11	1.410903e+11
3	RSS (Test)	2.796918e+31	9.160303e+10	8.110868e+10
4	MSE (Train)	1.176588e+04	1.309505e+04	1.397913e+04
5	MSE (Test)	3.003717e+14	1.718993e+04	1.617532e+04

There is not a major change in the R2 Score for Ridge and Lasso post doubling the alpha values. The R2 scores have reduced slightly for both Train data, and test data for Ridge, but it has increased very slightly for Lasso model.

The new top 20 predictor variables for Ridge are as follows:

```
betas_new.sort_values(by = 'Ridge_abs_new', ascending = False).head(20)
```

	Linear	Ridge	Lasso	Ridge_new	Lasso_new	Ridge_Pos_Neg_Corr	Lasso_Pos_Neg_Corr	Ridge_abs
GrLivArea	9.981496e+04	49533.091117	115081.097021	42741.713604	108689.440155	Pos	Pos	49533.091117
1stFlrSF	1.328171e+04	41439.399288	16411.590244	37471.411978	18119.877443	Pos	Pos	41439.399288
OverallQual_9	-6.884137e+14	28423.912540	47018.660851	22340.316666	41571.677665	Pos	Pos	28423.912540
BsmtFinSF1	3.045451e+04	22630.694816	17206.296089	20736.717912	17169.754510	Pos	Pos	22630.694816
2ndFlrSF	1.471365e+04	25047.112029	0.000000	20714.558326	0.000000	Pos	Pos	25047.112029
GarageArea	1.915356e+04	18690.368671	17763.786014	19089.414407	18670.484027	Pos	Pos	18690.368671
OverallQual_8	-6.884137e+14	16754.562940	26800.496768	17353.830295	26982.978750	Pos	Pos	16754.562940
BsmtQual_TA	-1.771620e+04	-17759.035322	-17946.336150	-16551.888360	-16254.446906	Neg	Neg	17759.035322
Neighborhood_Crawfor	1.809021e+04	16644.850355	21150.244830	15782.479700	20139.220852	Pos	Pos	16644.850355
BsmtQual_Gd	-1.534889e+04	-16310.476936	-15783.870831	-15420.109123	-14876.800660	Neg	Neg	16310.476936
age_since_built	-3.074266e+04	-18323.170135	-33582.045463	-13731.843777	-31612.171239	Neg	Neg	18323.170135
LotArea	1.680005e+04	13843.432151	13094.600959	13476.269714	12793.017032	Pos	Pos	13843.432151
Exterior1st_BrkFace	-3.015000e+03	13570.456293	13462.356728	13361.333284	13446.772754	Pos	Pos	13570.456293
BsmtExposure_Gd	1.340064e+04	13350.575858	13233.263642	12972.404655	12889.099927	Pos	Pos	13350.575858
Neighborhood_Somerst	1.946633e+04	15387.626478	17967.770849	12916.859151	17193.039945	Pos	Pos	15387.626478
FullBath	5.012664e+03	11731.888165	1296.046663	12626.131382	0.000000	Pos	Pos	11731.888165
KitchenQual_TA	-1.858391e+04	-13476.392876	-13897.455456	-12062.058240	-12013.641725	Neg	Neg	13476.392876
KitchenQual_Fa	-2.141658e+04	-13999.992042	-15048.892696	-11822.775489	-11635.341939	Neg	Neg	13999.992042
OverallQual_3	-6.884137e+14	-12017.117651	-6740.929714	-10712.030271	-5699.078470	Neg	Neg	12017.117651
MSSubClass_160	-1.455310e+04	-10992.059001	-15697.880903	-10582.102826	-12951.295535	Neg	Neg	10992.059001

The new top 20 predictor variables for Lasso are as follows:

```
betas_new.sort_values(by = 'Lasso_abs_new', ascending = False).head(20)
```

	Linear	Ridge	Lasso	Ridge_new	Lasso_new	Ridge_Pos_Neg_Corr	Lasso_Pos_Neg_Corr	Ridge_abs
GrLivArea	9.981496e+04	49533.091117	115081.097021	42741.713604	108689.440155	Pos	Pos	49533.091117
OverallQual_9	-6.884137e+14	28423.912540	47018.660851	22340.316666	41571.677665	Pos	Pos	28423.912540
age_since_built	-3.074266e+04	-18323.170135	-33582.045463	-13731.843777	-31612.171239	Neg	Neg	18323.170135
OverallQual_8	-6.884137e+14	16754.562940	26800.496768	17353.830295	26982.978750	Pos	Pos	16754.562940
Neighborhood_Crawfor	1.809021e+04	16644.850355	21150.244830	15782.479700	20139.220852	Pos	Pos	16644.850355
GarageArea	1.915356e+04	18690.368671	17763.786014	19089.414407	18670.484027	Pos	Pos	18690.368671
1stFlrSF	1.328171e+04	41439.399288	16411.590244	37471.411978	18119.877443	Pos	Pos	41439.399288
Neighborhood_Somerst	1.946633e+04	15387.626478	17967.770849	12916.859151	17193.039945	Pos	Pos	15387.626478
BsmtFinSF1	3.045451e+04	22630.694816	17206.296089	20736.717912	17169.754510	Pos	Pos	22630.694816
BsmtQual_TA	-1.771620e+04	-17759.035322	-17946.336150	-16551.888360	-16254.446906	Neg	Neg	17759.035322
BsmtQual_Gd	-1.534889e+04	-16310.476936	-15783.870831	-15420.109123	-14876.800660	Neg	Neg	16310.476936
Functional_Typ	-6.659961e+03	7220.185767	13338.833183	7635.915294	13868.626395	Pos	Pos	7220.185767
Exterior1st_BrkFace	-3.015000e+03	13570.456293	13462.356728	13361.333284	13446.772754	Pos	Pos	13570.456293
MSSubClass_160	-1.455310e+04	-10992.059001	-15697.880903	-10582.102826	-12951.295535	Neg	Neg	10992.059001
BsmtExposure_Gd	1.340064e+04	13350.575858	13233.263642	12972.404655	12889.099927	Pos	Pos	13350.575858
LotArea	1.680005e+04	13843.432151	13094.600959	13476.269714	12793.017032	Pos	Pos	13843.432151
KitchenQual_TA	-1.858391e+04	-13476.392876	-13897.455456	-12062.058240	-12013.641725	Neg	Neg	13476.392876
MSSubClass_90	-1.227239e+15	-8823.378463	-16881.984295	-7195.107714	-11815.768273	Neg	Neg	8823.378463
KitchenQual_Fa	-2.141658e+04	-13999.992042	-15048.892696	-11822.775489	-11635.341939	Neg	Neg	13999.992042
KitchenQual_Gd	-1.741086e+04	-11701.622786	-11969.837152	-10206.034162	-9750.056242	Neg	Neg	11701.622786

The top 20 predictor variables also almost remain the same with GrLivArea as the most important predictor across all.

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:

Choice –

Lasso regression with alpha – 50

Reason –

1. The R2 score is slightly better than Ridge Regression, in terms of the difference between train and test, which means that not all predictors were having some sort of contribution to the Sale Price.
2. Lasso successfully eliminated some of the irrelevant predictors by reducing the coefficients to 0.
3. Lower RSS for Lasso.
4. Lesser features making it a simpler 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?

Answer:

Top 5 for Lasso Regression

GrLivArea

OverallQual_9

age_since_built - (custom - negative correlation - so the newer the price rises)

OverallQual_8

OverallQual_10

Top 5 for Ridge Regression

GrLivArea,

1stFlrSF

OverallQual_9

2ndFlrSF

BsmtFinSF1

Based on above we will remove the following predictors:

['GrLivArea', 'OverallQual', 'age_since_built', '1stFlrSF', '2ndFlrSF', 'BsmtFinSF1']

Post removal these the final evaluation metric comparison is as follows:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.197061e-01	9.023405e-01	9.027040e-01
1	R2 Score (Test)	-3.836798e+22	8.377787e-01	8.397557e-01
2	RSS (Train)	1.628210e+11	1.980351e+11	1.972980e+11
3	RSS (Test)	3.563036e+34	1.506466e+11	1.488106e+11
4	MSE (Train)	1.501712e+04	1.656161e+04	1.653076e+04
5	MSE (Test)	1.072085e+16	2.204442e+04	2.190969e+04

Ridge Alpha - 4

Lasso Alpha - 50

It is clear visible that the Ridge and Lasso regression R2 score even though much better than the linear model, the test R2 score has reduced a significant amount.

Lasso is still slightly better in terms of Ridge still because of the slightly lesser gap in between Train and Test R2 Score.

New Top 20 for Ridge :

```
betas_q3.sort_values(by = 'Ridge_abs', ascending = False).head(20)
```

	Ridge	Lasso	Ridge_Pos_Neg_Corr	Lasso_Pos_Neg_Corr	Ridge_abs	Lasso_abs
GarageArea	34399.538969	41107.742157	Pos	Pos	34399.538969	41107.742157
FullBath	30043.628108	40526.503690	Pos	Pos	30043.628108	40526.503690
BsmtQual_TA	-22163.252979	-30830.413140	Neg	Neg	22163.252979	30830.413140
BsmtQual_Gd	-21279.511707	-27968.196602	Neg	Neg	21279.511707	27968.196602
Fireplaces	20625.971342	23523.295307	Pos	Pos	20625.971342	23523.295307
Neighborhood_StoneBr	19679.382588	33880.385752	Pos	Pos	19679.382588	33880.385752
Neighborhood_Crawfor	19307.119143	25871.719069	Pos	Pos	19307.119143	25871.719069
Exterior1st_BrkFace	18669.307203	22981.635828	Pos	Pos	18669.307203	22981.635828
LotArea	17555.438879	18555.406920	Pos	Pos	17555.438879	18555.406920
LotFrontage	16772.642008	15539.221221	Pos	Pos	16772.642008	15539.221221
BsmtExposure_Gd	16394.926443	17435.131332	Pos	Pos	16394.926443	17435.131332
BedroomAbvGr	15808.924264	17104.505123	Pos	Pos	15808.924264	17104.505123
HalfBath	15031.993511	18502.841343	Pos	Pos	15031.993511	18502.841343
KitchenQual_TA	-14638.696107	-15628.117751	Neg	Neg	14638.696107	15628.117751
OpenPorchSF	14632.658747	13979.236170	Pos	Pos	14632.658747	13979.236170
age_since_remod	-14262.650188	-14025.013892	Neg	Neg	14262.650188	14025.013892
MSSubClass_160	-13685.063171	-20308.133324	Neg	Neg	13685.063171	20308.133324
BsmtQual_Fa	-13553.915583	-25352.572141	Neg	Neg	13553.915583	25352.572141
KitchenQual_Fa	-13151.816741	-15347.140715	Neg	Neg	13151.816741	15347.140715
MasVnrType_Stone	12768.159781	9989.364917	Pos	Pos	12768.159781	9989.364917

New Top 20 for Lasso:

```
betas_q3.sort_values(by = 'Lasso_abs', ascending = False).head(20)
```

	Ridge	Lasso	Ridge_Pos_Neg_Corr	Lasso_Pos_Neg_Corr	Ridge_abs	Lasso_abs
GarageArea	34399.538969	41107.742157	Pos	Pos	34399.538969	41107.742157
FullBath	30043.628108	40526.503690	Pos	Pos	30043.628108	40526.503690
Neighborhood_StoneBr	19679.382588	33880.385752	Pos	Pos	19679.382588	33880.385752
BsmtQual_TA	-22163.252979	-30830.413140	Neg	Neg	22163.252979	30830.413140
BsmtQual_Gd	-21279.511707	-27968.196602	Neg	Neg	21279.511707	27968.196602
Neighborhood_Crawfor	19307.119143	25871.719069	Pos	Pos	19307.119143	25871.719069
SaleType_Con	11092.569365	25491.843076	Pos	Pos	11092.569365	25491.843076
BsmtQual_Fa	-13553.915583	-25352.572141	Neg	Neg	13553.915583	25352.572141
Fireplaces	20625.971342	23523.295307	Pos	Pos	20625.971342	23523.295307
Condition1_PosA	12377.411764	23385.842669	Pos	Pos	12377.411764	23385.842669
Exterior1st_BrkFace	18669.307203	22981.635828	Pos	Pos	18669.307203	22981.635828
MSSubClass_160	-13685.063171	-20308.133324	Neg	Neg	13685.063171	20308.133324
LotArea	17555.438879	18555.406920	Pos	Pos	17555.438879	18555.406920
HalfBath	15031.993511	18502.841343	Pos	Pos	15031.993511	18502.841343
Neighborhood_Somerst	12186.434794	18351.525176	Pos	Pos	12186.434794	18351.525176
Foundation_Wood	10070.986817	17796.692580	Pos	Pos	10070.986817	17796.692580
BsmtExposure_Gd	16394.926443	17435.131332	Pos	Pos	16394.926443	17435.131332
BedroomAbvGr	15808.924264	17104.505123	Pos	Pos	15808.924264	17104.505123
KitchenQual_TA	-14638.696107	-15628.117751	Neg	Neg	14638.696107	15628.117751
LotFrontage	16772.642008	15539.221221	Pos	Pos	16772.642008	15539.221221

New Top 5 as per Ridge:

1. GarageArea
2. FullBath
3. BsmtQual_TA
4. BsmtQual_Gd
5. Fireplaces

New Top 5 as per Lasso:

1. GarageArea
2. FullBath
3. Neighborhood_StoneBr
4. BsmtQual_TA
5. BsmtQual_Gd

GarageArea is the most important predictor in both.

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:

A model is both robust when it can handle variations in the data and generalisable when it can perform well on unseen data after it has learnt from seen (or training data) i.e. It understands the patterns in the data rather than memorising the data.

There are various things we can do to make sure that these 2 are met:

1. Try to have training data that has all data variations.
2. Data modification or transformation, including feature engineering wherever required to help the model understand the data better for eg, in this case I used age_since_built, instead of year built to be able to use it as a purely continuous variable.
3. Cross Validation can also be done to ensure model learns from different data variations and helps with generalisation. Helps with overfitting and ensures consistency across data subsets.
4. Regularisation as we have implemented here, helps with overfitting issues hence helping with generalisation.

Implications on Accuracy:

A generalised model may not always guarantee high accuracy, specially on the training dataset , but the important thing is that it should be performing equally well on test or unseen data as on training data.

If we do not take any of the steps as mentioned above the potential risks could be:

Overfitting – In this case model understands training data really well but does not perform well on test data, as in our case without regularisation.

Underfitting – In trying to make a very simple model the model may not even understand the patterns in the data well.