Question 1:

What is the optimal value of alpha for ridge and lasso regression?

Ridge: alpha = 6

Lasso: alpha = 0.001

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

When lambda value is increased:

- The R square value decreases on Test dataset for both ridge and lasso
- The mean square error increases for both ridge and lasso
- Fewer fields are chosen on lasso (coefficients almost zero)

In general, too high lambda leads to underfitting.

Optimal:
Ridge: alpha = 6
Lasso: alpha = 0.001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.952426	0.932703	0.919312
1	R2 Score (Test)	0.873460	0.889879	0.890947
2	RSS (Train)	7.633061	10.797501	12.946007
3	RSS (Test)	9.155622	7.967668	7.890353
4	MSE (Train)	0.086464	0.102837	0.112604
5	MSE (Test)	0.144415	0.134720	0.134065

Sub Optimal:
Ridge: alpha = 12
Lasso: alpha = 0.002

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.952426	0.925850	0.902194
1	R2 Score (Test)	0.873460	0.887292	0.879986
2	RSS (Train)	7.633061	11.896939	15.692457
3	RSS (Test)	9.155622	8.154839	8.683432
4	MSE (Train)	0.086464	0.107946	0.123975
5	MSE (Test)	0.144415	0.136294	0.140641

Ridge: alpha = 24
Lasso: alpha = 0.004

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.952426	0.916402	0.876928
1	R2 Score (Test)	0.873460	0.882047	0.855832
2	RSS (Train)	7.633061	13.412865	19.746188
3	RSS (Test)	9.155622	8.534343	10.431064
4	MSE (Train)	0.086464	0.114617	0.139068
5	MSE (Test)	0.144415	0.139429	0.154146

What will be the most important predictor variables after the change is implemented?

The below 4 variables with a reasonably high coefficient continues to be significant predictors. Whereas the variables with contending coefficients have switched positions in the ranking.

After increasing the lambda below are the predictors:

- Ground Living Area
- Overall Quality
- Lot Area
- First Floor Square Feet

	optimal lambda		
Fields	Linear	Ridge	Lasso
GrLivArea	0.39	0.25	0.37
OverallQual	0.23	0.20	0.27
Neighborhood_Crawfor	0.12	0.10	0.11
Functional_Typ	0.20	0.09	0.09
Neighborhood_NoRidge	0.07	0.09	0.09
KitchenQual_Ex	0.07	0.06	0.08
Neighborhood_NridgHt	0.02	0.06	0.07
LotArea	0.10	0.07	0.06
1stFlrSF	0.08	0.11	0.06
Neighborhood_Somerst	0.02	0.06	0.06
Condition1_Norm	0.04	0.05	0.06
Exterior1st_BrkFace	0.12	0.06	0.06
RoofMatl_ClyTile	-1.47	-0.15	-0.23

	double lambda		
	Linear	Ridge	Lasso
GrLivArea	0.39	0.21	0.37
OverallQual	0.23	0.17	0.28
Neighborhood_Crawfor	0.12	0.09	0.08
Functional_Typ	0.20	0.08	0.07
KitchenQual_Ex	0.07	0.06	0.07
LotArea	0.10	0.07	0.07
1stFlrSF	0.08	0.12	0.06
Condition1_Norm	0.04	0.05	0.05
Neighborhood_NoRidge	0.07	0.08	0.05
BExp_Good	0.06	0.06	0.04
Neighborhood_NridgHt	0.02	0.05	0.03
Neighborhood_Somerst	0.02	0.05	0.03
RoofMatl_ClyTile	-1.47	-0.08	0.00

	triple lambda		
	Linear	Ridge	Lasso
GrLivArea	0.39	0.18	0.36
OverallQual	0.23	0.13	0.25
LotArea	0.10	0.07	0.08
1stFlrSF	0.08	0.11	0.05
KitchenQual_Ex	0.07	0.06	0.04
Functional_Typ	0.20	0.06	0.04
Condition1_Norm	0.04	0.04	0.04
BsmtQual	0.01	0.02	0.03
FireplaceQu	0.01	0.02	0.03
BExp_Good	0.06	0.05	0.02
Neighborhood_Crawfor	0.12	0.07	0.02
BsmtFinSF1	0.01	0.01	0.01
RoofMatl_ClyTile	-1.47	-0.04	0.00

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?

Lasso is better.

- Test R2 score is slightly better in Lasso compared to Ridge and LR
- Test MSE is also better in Lasso compared to Ridge and LR
- Lasso has helped with Feature Selection (reduced 50+ variables). Robust model.

Ridge: alpha = 6
Lasso: alpha = 0.001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.952426	0.932703	0.919312
1	R2 Score (Test)	0.873460	0.889879	0.890947
2	RSS (Train)	7.633061	10.797501	12.946007
3	RSS (Test)	9.155622	7.967668	7.890353
4	MSE (Train)	0.086464	0.102837	0.112604
5	MSE (Test)	0.144415	0.134720	0.134065

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?

	optimal lambda		
Fields	Linear	Ridge	Lasso
GrLivArea	0.39	0.25	0.37
OverallQual	0.23	0.20	0.27
Neighborhood_Crawfor	0.12	0.10	0.11
Functional_Typ	0.20	0.09	0.09
Neighborhood_NoRidge	0.07	0.09	0.09
KitchenQual_Ex	0.07	0.06	0.08
Neighborhood_NridgHt	0.02	0.06	0.07
LotArea	0.10	0.07	0.06
1stFlrSF	0.08	0.11	0.06
Neighborhood_Somerst	0.02	0.06	0.06
Condition1_Norm	0.04	0.05	0.06
Exterior1st_BrkFace	0.12	0.06	0.06
RoofMatl_ClyTile	-1.47	-0.15	-0.23

Top 5 Removed	
Fields	Lasso
1stFlrSF	0.35
SaleType_ConLD	0.21
ExterCond_Ex	0.17
Fuse_Poor	0.15
HouseStyle_2.5Unf	0.11
RoofMatl_WdShngl	0.11
Foundation_Stone	0.10
KitchenQual_Ex	0.10
LotArea	0.10
Exterior1st_BrkFace	0.09
RoofMatl_Membran	0.09
TotRmsAbvGrd	0.09
RoofMatl_ClyTile	-1.56

- **1stFlrSF** 0.353652
- **SaleType_ConLD** 0.205254
- **ExterCond_Ex** 0.165834
- **Fuse_Poor** 0.154333
- **HouseStyle_2.5Unf** 0.110008

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?

Below tasks can be performed to make a model robust.

- **Remove Outliers.**
  - This will help achieve better accuracy of the model. When a variable with high coefficient has got outliers the model will not be able to predict accurately when tested on new dataset.
- **Ensure Test R2 score is at par with Train R2 score.**
  - This will imply that the built model is robust and covers all significant predictors.
- **Feature elimination**
  - This will imply that the model is not overfit/underfit.
- **Regularization**: Model should not be too complex.
  - Trade-off between Bias & Variance (Accuracy vs Complexity)