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?

Below are the optimal alpha values for both models and their R2_scores are nearly 0.91 on training data and 0.86 test data.

- Ridge regression - 100
- Lasso regression - 0.001

If we double the alpha values in both models, we will have the below alpha values

- Ridge regression - 200
- Lasso regression - 0.002

While doubling the alpha values, we are having drip in the R2 scores on both Ridge and Lasso regression models. Also, there are changes in the co-efficient values for each feature.

Below are the top features for both Ridge and Lasso regression models for the both optimal alpha and doubled alpha values.

Ridge Regress Model:

Features	Co-efficient with optimal alpha
OverallQual	0.059396
GrLivArea	0.049624
GarageCars	0.042117
OverallCond	0.041071
1stFlrSF	0.033092
2ndFlrSF	0.029292
FullBath	0.028227
MSZoning_RL	0.026743
TotRmsAbvGrd	0.023595
CentralAir_Y	0.018787
Neighborhood_NridgHt	0.018774
Condition1_Norm	0.018757
SaleCondition_Normal	0.018575
Neighborhood_Somerst	0.018031
BsmtFullBath	0.01703
LotArea	0.015253
BsmtExposure_Gd	0.014601
HalfBath	0.013891
MSSubClass_2-STORY 1945 & OLDER	0.013777
Foundation_PConc	0.013223

Features	Co-efficient with doubled alpha
OverallQual	0.054219
GrLivArea	0.042365
GarageCars	0.036511
OverallCond	0.036391
1stFlrSF	0.03082
FullBath	0.026261
TotRmsAbvGrd	0.024441
2ndFlrSF	0.022706
MSZoning_RL	0.018707
Neighborhood_NridgHt	0.018241
CentralAir_Y	0.017759
Condition1_Norm	0.015883
BsmtFullBath	0.015096
Neighborhood_Somerst	0.014926
SaleCondition_Normal	0.014529
BsmtExposure_Gd	0.014395
LotArea	0.01415
GarageArea	0.014027
HalfBath	0.01337
MSSubClass_2-STORY 1945 & OLDER	0.013096

Lasso Regression Model:

Features	Co-efficient with optimal alpha
GrLivArea	0.117108
OverallQual	0.069362
MSZoning_RL	0.062392
GarageCars	0.050071
OverallCond	0.046426
MSZoning_RM	0.03355
MSZoning_FV	0.028411
FullBath	0.024095
Neighborhood_Somerst	0.022655
MSSubClass_1-STORY 1946 & NEWER ALL	0.020909
SaleCondition_Normal	0.020275
BsmtFullBath	0.019884
Condition1_Norm	0.019866
Neighborhood_NridgHt	0.019159
CentralAir_Y	0.018084
LotArea	0.017239
Neighborhood_others	0.017191
Foundation_PConc	0.015414
BsmtExposure_Gd	0.014593
TotRmsAbvGrd	0.014238

Features	Co-efficient with doubled alpha
GrLivArea	0.11563
OverallQual	0.078751
GarageCars	0.049222
OverallCond	0.045094
MSZoning_RL	0.024277
Neighborhood_Somerst	0.021374
BsmtFullBath	0.020452
FullBath	0.020381
Neighborhood_NridgHt	0.020194
SaleCondition_Partial	0.020109
MSSubClass_1-STORY 1946 & NEWER ALL	0.019598
SaleCondition_Normal	0.018092
Condition1_Norm	0.017236
Foundation_PConc	0.016921
CentralAir_Y	0.016896
LotArea	0.016577
Neighborhood_others	0.016311
BsmtExposure_Gd	0.014523
LotConfig_CulDSac	0.013354
ScreenPorch	0.010976

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?

Below are the optimal alpha values for both models.

- Ridge regression – 100
- Lasso regression - 0.001

R2_scores and mean squared error values are almost same.

	Ridge	Lasso
train_R2_score	0.914307	0.917804
test_R2_score	0.865082	0.868639
train_MSE	0.012362	0.011858
test_MSE	0.019535	0.01902

Lasso model will helps feature reduction by assigning co-efficient values as zero for some features. So, we can choose Lasso over Ridge model in our final application.

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?

Below are the top five important features with the original predictor columns.

Features	Lasso model with all features
GrLivArea	0.117108
OverallQual	0.069362
MSZoning_RL	0.062392
GarageCars	0.050071
OverallCond	0.046426

While we removing these columns from model building, we are getting the same optimal alpha as 0.001. Also compared with previous model, the R2 score decreased in both training and test data set.

Below are the new top five important predictor variables with the new Lasso model.

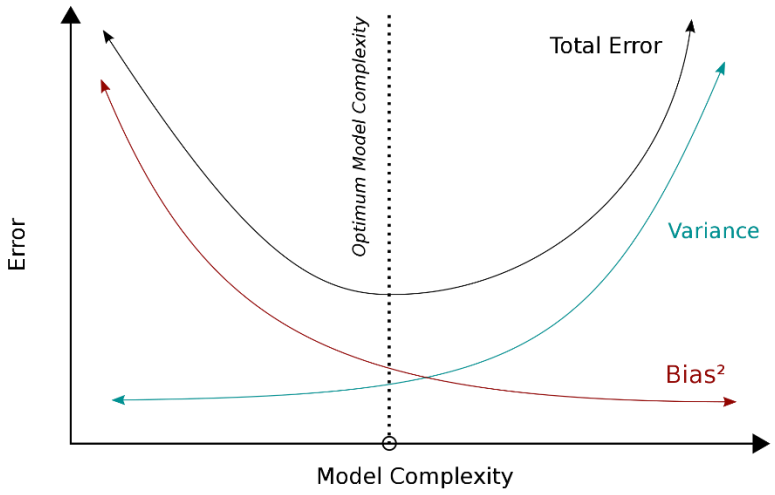
Features	Lasso model without top5 features
2ndFlrSF	0.103209
1stFlrSF	0.097001
FullBath	0.035169
GarageArea	0.034473
SaleCondition	0.0281

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?

- A model needs to be made robust and generalisable so that we will not have any impact in predicting the output variable with the new dataset which are not used in the training phase.
- A robust and generalized model will be consistently accurate with the test dataset even if the input dataset having high variance.
- If the model is having high complexity, will lead to have huge errors with the new dataset which are having outliers.

- So, we always need to build the model by considering the optimal bias-variance trade off. This will prevent the model being over fitted/under fitted. Also, this will reduce the errors in the predictions.
- We can trust the predicted values of a model only if the model is robust and generalised.



Above Image reference from Wikipedia.