Q1. 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?

Ans:-

- The optimal value of alpha for Ridge is 5 and for Lasso is 0.001
- With these alpha values the R-squared value of the model was approx. 0.92
- After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.92 but there is a small change in the co-efficient values
- The new model is created and demonstrated in the Jupyter notebook. Below are screenshots of the same (the top features) –

Ridge Regression Model –

	Ridge Alpha Co-Efficient
Neighborhood_Crawfor	0.101732
SaleCondition_Normal	0.071787
OverallQual	0.067289
GrLivArea	0.064910
Neighborhood_StoneBr	0.060831
Foundation_PConc	0.056440
Functional_Typ	0.056080
SaleCondition_Partial	0.053383
MSZoning_FV	0.052047
Condition1_Norm	0.051553
CentralAir_Y	0.051108
OverallCond	0.048153
BsmtExposure_Gd	0.047994
MSZoning_RL	0.045777
2ndFlrSF	0.042180
YearBuilt	0.038322
1stFlrSF	0.036583
TotalBsmtSF	0.035705
Street_Pave	0.035687
Exterior1st_BrkFace	0.034829

	Ridge Doubled Alpha Co-Efficient
Neighborhood_Crawfor	0.101732
SaleCondition_Normal	0.071787
OverallQual	0.067289
GrLivArea	0.064910
Neighborhood_StoneBr	0.060831
Foundation_PConc	0.056440
Functional_Typ	0.056080
SaleCondition_Partial	0.053383
MSZoning_FV	0.052047
Condition1_Norm	0.051553
CentralAir_Y	0.051108
OverallCond	0.048153
BsmtExposure_Gd	0.047994
MSZoning_RL	0.045777
2ndFlrSF	0.042180
YearBuilt	0.038322
1stFlrSF	0.036583
TotalBsmtSF	0.035705
Street_Pave	0.035687
Exterior1st_BrkFace	0.034829

Lasso Regression Model –

	Lasso Alpha Co-Efficient
MSZoning_FV	0.411518
MSZoning_RL	0.369185
MSZoning_RH	0.344318
MSZoning_RM	0.338656
Neighborhood_Crawfor	0.164373
Neighborhood_StoneBr	0.101397
SaleCondition_Partial	0.091374
Street_Pave	0.088058
GarageCond_Po	0.083437
SaleCondition_Normal	0.078976
GrLivArea	0.071109
Condition1_Norm	0.071083
Heating_Wall	0.070294
Foundation_PConc	0.069823
Exterior1st_BrkFace	0.063716
Condition1_PosN	0.061585
HouseStyle_2.5Unf	0.058809
Functional_Typ	0.058135
Condition1_RRNn	0.057835
OverallQual	0.056746

	Lasso Doubled Alpha Co-Efficient
MSZoning_FV	0.237814
MSZoning_RL	0.203854
MSZoning_RH	0.168271
MSZoning_RM	0.167057
Neighborhood_Crawfor	0.155627
GrLivArea	0.115767
SaleCondition_Partial	0.100004
Neighborhood_StoneBr	0.090456
SaleCondition_Normal	0.083943
Condition1_Norm	0.064707
Foundation_PConc	0.062271
Street_Pave	0.061264
OverallQual	0.060589
BsmtExposure_Gd	0.054586
Functional_Typ	0.052865
Exterior1st_BrkFace	0.051396
Condition1_PosN	0.049902
OverallCond	0.045228
CentralAir_Y	0.044762
Neighborhood_BrkSide	0.042177

Q.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?

Ans – On evaluation, it was observed that R-squared and mean squared error for both Ridge and Lasso are almost similar but since Lasso helps in feature reduction as it eliminates the variables that doesn't contribute to the model prediction by making the coefficient to zero. Therefore, Lasso has a better edge over Ridge and should be used as the final model.

Q3. 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?

Ans – Five most important predictor variables after dropping the previous 5 predictor variables are (demonstration done in jupyter notebook) –

- Street_Pave
- SaleCondition_Partial
- GarageCond_Po
- Condition1_Norm
- SaleCondition_Normal

Q4. 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?

Ans –

- We can assure the model robustness and generalisable by checking its performance on both seen and unseen data
- Model should perform good on training and testing data
- We can check this by looking at the accuracy of the model.
- It should be almost similar in both.
- The model should not overfit or underfit
- It shouldn’t be affected by the outliers