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?

Answer: Optimal value of alpha Ridge is 2 as the test error is minimum and negative mean absolute error gets stabilized at alpha =2. For lasso, alpha is 0.01 as when the value of alpha increases the model penalizes more and makes most of the coefficients as 0.

For Ridge: When we double value of alpha i.e. 4 the model will apply more penalty on the curve and will try to make the model more generalized that is making model more simple and will try to fit every data of dataset. Chart below indicates more error for test and train.

Important Ridge variables after making changes (i.e. double alpha):

	Variable	Coeff
0	constant	11.872
50	Neighborhood_Crawfor	0.109
29	MSZoning_FV	0.099
31	MSZoning_RL	0.087
66	Neighborhood_StoneBr	0.081
210	SaleCondition_Partial	0.079
13	GrLivArea	0.076
209	SaleCondition_Normal	0.071
4	OverallQual	0.067
30	MSZoning_RH	0.063
70	Condition1_Norm	0.061
95	Exterior1st_BrkFace	0.060
32	MSZoning_RM	0.054
5	OverallCond	0.053
134	Foundation_PConc	0.048
166	CentralAir_Y	0.048
136	Foundation_Stone	0.047
11	2ndFlrSF	0.047

For lasso: when we double alpha the model will be penalized more and more coefficient of the variable will be reduced to 0. Also, r-square value also decreases.

R-square & RMSE after changes on train and test:

Train: 0.872579770020679

RMSE: 0.12768634359747102

Important lasso variables after making changes (i.e. double alpha):

Out[328]:

	Variable	Coeff
0	constant	12.154
13	GrLivArea	0.123
4	OverallQual	0.097
5	OverallCond	0.048
9	TotalBsmtSF	0.042
7	BsmtFinSF1	0.032
20	Fireplaces	0.026
21	GarageArea	0.025
2	LotFrontage	0.012
3	LotArea	0.011
14	BsmtFullBath	0.002
22	WoodDeckSF	0.001
1	MSSubClass	-0.002
28	PropertyAge	-0.004

Q2. 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: Ridge regression, uses tuning parameter lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum of squares should be small by using penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficient that have greater values gets penalized. AS the value of lambda increases the variance is model gets reduced and bias remains constant. **Ridge regression includes all variables in final model unlike Lasso.**

Lasso regression, uses tuning parameter lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increase Lasso shrinks the coefficients towards zero and makes variables exactly equal to zero. Lasso also does variable selection. When lambda is small simple Linear regression is performed and when its large shrinkage happens and variables with zero values are neglected by model.

In the model, model performance by Ridge was better in terms of R-square values of train and test but I chose Lasso ad it brings and assigns a 0 value to insignificant features, which helps to choose predictive variables.

R-square & RMSE for Ridge:

Train: 0.93645474997655

Test: 0.9106281303179837

RMSE: 0.11305786818560676

R-square & RMSE for Lasso:

Train: 0.8852986602638322

Test: 0.8937853852569423

RMSE: 0.12325158400977712

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?

Answer: Top 5 important predictor variables initially were:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1

After removing these 5 variables, next 5 important predictor variables are:

- GarageArea
- Fireplaces
- LotArea
- LotFrontage
- BsmtFullBath

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?

Answer: The model should be as simple as possible, though its accuracy will decrease but it will be more generalizable and robust. This can also be interpreted using Bias-Variance trade-off. The simpler the model the more bias but less variance and more generalizable. Its implication in terms of accuracy is : A robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias, in model refers to error in model. Higher bias means model is not able to learn details of data. Model will perform poorly on train and test data.

Variance, refers to consistency of model i.e. how the model reacts to inputs to model. High model means model performs exceptionally well on training data as it is very well trained on this data but performs very poor on testing data (as it unseen data).

Balance between Bias and variance is important to avoid **over-fitting** and **under-fitting**.