Question 1
What is the optimal value of alpha for ridge and lasso regression?

Optimal Value of alpha :
Ridge : 10
Lasso : 0.001

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

For ridge regression when we doubled the lambda to 20, the evaluation metric values were :

Lambda 10	Lambda 20
R-Squared (Train) = 0.94	R-Squared (Train) = 0.93
R-Squared (Test) = 0.93	R-Squared (Test) = 0.93
RSS (Train) = 8.53	RSS (Train) = 9.37
RSS (Test) = 2.87	RSS (Test) = 2.82
MSE (Train) = 0.01	MSE (Train) = 0.01
MSE (Test) = 0.01	MSE (Test) = 0.01
RMSE (Train) = 0.09	RMSE (Train) = 0.09
RMSE (Test) = 0.10	RMSE (Test) = 0.10

For Lasso regression when we doubled the lambda to 0.002, the evaluation metric values were :

Lambda 0.001	Lambda 0.002
R-Squared (Train) = 0.92	R-Squared (Train) = 0.91
R-Squared (Test) = 0.93	R-Squared (Test) = 0.91
RSS (Train) = 11.29	RSS (Train) = 13.49
RSS (Test) = 2.92	RSS (Test) = 3.45
MSE (Train) = 0.01	MSE (Train) = 0.01
MSE (Test) = 0.01	MSE (Test) = 0.01
RMSE (Train) = 0.10	RMSE (Train) = 0.11
RMSE (Test) = 0.10	RMSE (Test) = 0.11

We can see that the R squared value for test has reduced in both scenarios than the optimal lambda values. For ridge the RMSE has not varied but for Lasso RMSE has increased

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

Ridge New variables - 20		Ridge Old variable -10	
GrLivArea	1.08	GrLivArea	1.09
OverallQual_8	1.07	OverallQual_9	1.08
OverallQual_9	1.07	OverallQual_8	1.08
Neighborhood_Crawfor	1.07	Neighborhood_Crawfor	1.08
Functional_Typ	1.06	OverallCond_9	1.08
Exterior1st_BrkFace	1.06	Functional_Typ	1.07
OverallCond_9	1.06	Exterior1st_BrkFace	1.07
TotalBsmtSF	1.05	SaleCondition_Alloca	1.07
CentralAir_Y	1.05	CentralAir_Y	1.06
OverallCond_7	1.04	TotalBsmtSF	1.05

Lasso New variables – 0.002		Lasso Old variable – 0.001	
GrLivArea	1.11	OverallQual_9	1.13
OverallQual_8	1.09	GrLivArea	1.11
OverallQual_9	1.08	OverallQual_8	1.11
Functional_Typ	1.07	Neighborhood_Crawfor	1.09
Neighborhood_Crawfor	1.07	Exterior1st_BrkFace	1.08
TotalBsmtSF	1.05	Functional_Typ	1.08
Exterior1st_BrkFace	1.05	CentralAir_Y	1.05
CentralAir_Y	1.04	Neighborhood_Somerst	1.04
YearRemodAdd	1.04	TotalBsmtSF	1.04
Condition1_Norm	1.03	Condition1_Norm	1.04

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?

From the output based on two models, I would choose Lasso regression as the business would need to decide the primary parameters to take a decision whether a property is good to act upon. Lasso reduces the total number of variable which will be needed to make a decision quickly

Ridge regression has more than 300 variables and would be cumbersome for the business to take a decision based on so many variables whereas in Lasso its 42 variables which has co eff not equal to zero

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?

The top 5 predictors for lasso regression model were OverallQual_9, GrLivArea, OverallQual_8,Neighborhood_Crawfor and Exterior1st_BrkFace

After removing these varables and running the lasso regrtession again, I got the top 5 variables as ： 2ndFlrSF, Functional_Typ, 1stFlrSF, MSSubClass_70, Neighborhood_Somerst

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?

To ensure that a model is robust and generalizable in the context of Ridge Regression and Lasso Regression, several strategies and considerations are essential:

1. Tuning the Regularization Parameter (λ):

- The regularization parameter (λ) controls the strength of the penalty applied to the coefficients in Ridge and Lasso regression. Selecting an optimal λ is crucial for balancing bias and variance.
- Use techniques like grid search or randomized search with cross-validation to find the optimal λ. You can also use algorithms like Elastic Net that combine L1 and L2 regularization to find the best balance.
- Proper tuning of λ ensures that the model is neither too simple (high bias) nor too complex (high variance), leading to better generalization and robustness. A well-tuned λ can improve the model's accuracy by preventing overfitting (small λ) or underfitting (large λ).

2. Regularization for Bias-Variance Tradeoff:

- Regularization techniques (Ridge and Lasso) control overfitting by penalizing large coefficients. The goal is to strike a balance between bias and variance.
- Adjust the regularization strength to ensure the model has neither too much variance (overfitting) nor too much bias (underfitting). Ridge helps reduce variance by shrinking coefficients, while Lasso also reduces variance and can perform feature selection.
- A well-regularized model (with the appropriate λ) is more likely to generalize well, leading to improved accuracy on unseen data. Too much regularization increases bias and decreases accuracy, while too little regularization can lead to overfitting and poor generalization.

Apart from the above 2 steps, we should follow the general methods of making a linear regression model more robust by following the process of :

1. Cross validation by dividing the data into multiple subsets and trained on multiple subsets and results averaged to get a generalized error to avoid over fitting
2. Feature Selection and engineering to make the model more interpretable and generalized. This can be done by feature reduction methods or automatic feature selection by methods like lasso regression
3. Assessing over an unsee test data to ensure out of sample testing which gives a good picture whether the model is stable for real scenarios