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?

Part 1

Optimal value of alpha for ridge regression – 10

Optimal value of alpha for lasso regression – 500

Part 2

	Ridge Reg @ 10.00	Ridge Reg @ 20.00
R2 Score (Train)	0.903	0.890
R2 Score (Test)	0.860	0.854
RSS (Train)	616,909,057,841	703,998,589,170
RSS (Test)	394,942,891,888	411,637,065,462
MSE (Train)	604,220,429	689,518,697
MSE (Test)	901,696,100	939,810,652

	Lasso Reg @ 500.00	Lasso Reg @ 1000.00
R2 Score (Train)	0.859	0.815
R2 Score (Test)	0.831	0.784
RSS (Train)	899,827,197,897.071	1,182,275,639,286.440
RSS (Test)	475,672,452,728.439	607,649,464,211.608
MSE (Train)	881,319,488.636	1,157,958,510.565
MSE (Test)	1,086,010,166.047	1,387,327,543.862

For both ridge and lasso regression the metrics start to degrade if Alpha is doubled though the difference is not that high.

Part 3

The most important predictor variable with double alpha with Ridge Regression

1. FullBath_3
2. TotRmsAbvGrd_10
3. Neighborhood_NoRidge
4. GrLivArea
5. OverallQual_10

The most important predictor variables with double alpha with Lasso Regression

1. GrLivingArea
2. Bsmt_Quality
3. FullBath_3
4. OverallQual_9
5. Neighborhood_NoRidge

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?

I tried regression using RFE, Linear regression, Ridge regression and Lasso regression. All models gave importance to different variables. However, based on my analysis Lasso regression made most sense from a business logic perspective that is the predictor variables picked makes most business sense. Out of all these models, I will go with Lasso Regression.

I have used different values of lambda. I will choose the lambda that will balance between rsquared and other metrics for training and test data.

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?

After remove the five most important predictor variables, the next set of important variables are

1. 1stFlrSF
2. 2ndFlrSF
3. GarageCars_3
4. TotRmsAbvGrd_10
5. Fireplaces_2

It is interesting to note that GrLivingArea was replaced with 1stFlrSF and 2ndFlrSF while they had zero value for coefficients in the original model. Basically, the model is looking for area of the house.

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?

Few ways to make the model robust and generalisable

1. Implement various algorithms to understand the important features. As seen in the current assignment, each model gave some unique set of features. Pick the key variables and create a model using those variables.
2. Build the model using cross-validation approach.
3. Ensure that the residual analysis metrics are similar for training and testing data.
4. Ensure that each variable is clearly understood from business perspective. There are 80 variables in the current assignment with several categorical variables. Each categorical variable must be studied closely.
5. Run through the initial results with the business experts and then recreate the final model based on their feedback.

The accuracy will improve for each case above. For example

1. If we pick the right features, the rsquared will improve.
2. Cross validation will give us more data to train the model.
3. Similar metrics for training and testing data will ensure there is no overfitting
4. Business variables are critical to understand for better accuracy. It is possible that an important feature may be eliminated. Domain experts can provide feedback on these variables.