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?

- The optimal value of alpha for ridge regression is 500 and for lasso regression is also 500.
- The changes in the model of alpha =500 and alpha =1000 is mentioned in below table:-

For 'Alpha' = 500			
Metric	**Linear Regression**	**Ridge Regression**	**Lasso Regression**
R2 Score (Train)	0.951	0.891	0.925
R2 Score (Test)	-3.82E+22	0.860	0.851
RSS (Train)	3.10E+11	6.97E+11	4.81E+11
RSS (Test)	1.08E+35	3.95E+11	4.20E+11
MSE (Train)	1.74E+04	2.61E+04	2.17E+04
MSE (Test)	1.57E+16	3.00E+04	3.10E+04
For 'Alpha' = 1000			
R2 Score (Train)	0.9515	0.867	0.896
R2 Score (Test)	-3.82E+22	0.845	0.844
RSS (Train)	3.10E+11	8.50E+11	6.60E+11
RSS (Test)	1.08E+35	4.37E+11	4.40E+11
MSE (Train)	1.74E+04	2.89E+04	2.54E+04
MSE (Test)	1.57E+16	3.16E+04	3.17E+04

- The most important predictor variable after the change is implemented is : - '**GrLivArea**' {Above grade (ground) living area square feet}.

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?

- As it can be seen from the above table 'R2 score' of Ridge regression is slightly higher than lasso regression in the test data set, also difference in 'R2 Score' of train vs test data set is minimum in the case of Ridge regression.
  RSS and MSE value of test data is slightly less in Ridge than in lasso regression.
  Hence, in this case, Ridge regression is the better-suited model.

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?

- 1stFlrSF (First Floor square feet)
- 2ndFlrSF (Second floor square feet)
- OverallQual_10 (the overall material and finish of the house – 10 , Excellent)
- Neighborhood_NridgHt (Neighborhood - Northridge Heights)
- Neighborhood_NoRidge (Neighborhood - Northridge Northridge)

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?

- We can make sure that a machine learning model is robust and generalizable that it performs consistently across different datasets and is less prone to overfitting or being overly sensitive to variations in the training data. Below are few considerations for achieving model robustness and generalizability: -
  - **Cross-Validation:** Use techniques like k-fold cross-validation to assess the model's performance on multiple subsets of the data which ensure that the model's performance is consistent.
  - **Train-Test Split:** Divide the dataset into separate training and testing sets. This allows you to evaluate the model's performance on unseen data, which is a better measure of its generalization ability compared to using the training data alone.
  - **Regularization:** Regularization helps to control the model's complexity and reduce its sensitivity to noise in the training data.
  - **Feature Engineering:** Choose relevant features that capture the underlying patterns in the data.
  - **Hyperparameter Tuning:** Fine-tune hyperparameters using techniques like grid search or random search which helps in finding the best configuration that optimizes performance across various datasets.
  - **Outlier Detection and Handling:** Identify and handle outliers appropriately.
  - **Domain Knowledge:** Applying domain knowledge into the modelling process.
  - **Regular Monitoring:** Continuously monitor and evaluate the model's performance in real-world scenarios.

- The Implications for Model Accuracy:
  - **Overfitting:** A model that's overly complex can fit the training data extremely well but fail to generalize to new data resulting in high accuracy on the training set and poor accuracy on new data.
  - **Bias-Variance Trade-off**: To reduce overfitting, there is a need to sacrifice some accuracy on the training data to achieve better generalization performance.
  - **Unseen Scenarios:** Models that are trained with too much focus on specific patterns in the training data might fail to handle scenarios that differ from those patterns.

In conclusion, the model's aim is to maintain a balance between accuracy and generalization. A robust and generalizable model might not achieve the highest accuracy on the training data, but it is more likely to perform well on new, unseen data, which is the ultimate measure of success for machine learning models in practical applications.