Question 1:

i. **What is the optimal value of alpha for ridge and lasso regression?**
Ridge: alpha = 3 | Lasso: alpha = 100

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

<u>Ridge</u>: alpha = 6
R2 score of training data reduced from 0.8718507878152617 to 0.8595904721349068.
R2 score of testing data reduced from 0.8382464361790748 to 0.8301857458631146.
<u>Lasso</u>: alpha = 200
R2 score of training data reduced from 0.8689176078722685 to 0.8595904721349068.
R2 score of testing data reduced from 0.8439859315028525 to 0.8349390025870937.

However, there is no change in the top 5 predictors of the model, except a minor change in their co-efficient values.

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

⇒ *LotFrontage - Linear feet of street connected to property*
⇒ *LotArea - Lot size in square feet*
⇒ *OverallQual - Rates the overall material and finish of the house*
⇒ *OverallCond - Rates the overall condition of the house*
⇒ *MasVnrArea - Masonry veneer area in square feet*
⇒ *BsmtFinSF1 - Type 1 finished square feet*
⇒ *1stFlrSF - First Floor square feet*
⇒ *2ndFlrSF - Second Floor square feet*
⇒ *LowQualFinSF - Low quality finished square feet (all floors)*
⇒ *GrLivArea - Above grade (ground) living area square feet*
⇒ *BsmtFullBath - Basement full bathrooms*
⇒ *FullBath - Full bathrooms above grade*

However, there is no change in the top 5 predictors of the model before & after the change.

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?

It's better to choose <u>Lasso</u> because of the following reasons:

⇒ *The difference between training R2 score & testing R2 score is less than 3% (~2.92) when compared to Ridge, which has a difference of ~3.98% .*
⇒ *It has slightly higher R2 score on test data (~0.844) compared to Ridge (~0.838).*

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?

- ⇒ *BsmtFinSF1 - Type 1 finished square feet*
- ⇒ *1stFlrSF - First Floor square feet*
- ⇒ *2ndFlrSF - Second Floor square feet*
- ⇒ *LowQualFinSF - Low quality finished square feet (all floors)*
- ⇒ *GrLivArea - Above grade (ground) living area square feet*

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?

- ⇒ *Outlier Analysis: Only relevant features which makes sense from business point of view should be retained, else, the model might try to overfit the datapoints and always return high accuracy.*

- ⇒ *Data Transformation: Helps in improving the data quality, ensures there is not much variance in the data and ensures data standardization.*

- ⇒ *Multiple Models: It's always important to build multiple models before finalizing one. With multiple models, it becomes easier to test the data and see the pros & cons of all, again, based on business needs, appropriate model could be chosen i.e., if less penalty needs to be imposed / higher r2 score is desired, then we can choose linear regression or ridge models. However, if the features needs to be eliminated which are not adding much value to the r2 score, then lasso could be used, as it does feature elimination automatically.*

- ⇒ *R2 Score of Train set v/s Test set: It's necessary to evaluate the r2 scores of both training set and test set and make sure the test data score is not off by a huge margin when compared with training data score, as this would result in poor prediction capabilities of the model.*

  So, it's always better to perform outlier analysis, data cleaning, transformation, scaling and choosing right threshold values of hyperparameters before model building as this will help in building a simple, stable, scalable, and a robust model.