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?

Answer : After performing Ridge and Lasso regression on the provided housing data, below are the optimal values of alpha

- **Ridge regression:** alpha ≈ **214.61**
- **Lasso regression:** alpha ≈ **0.01**

These values were chosen because they provide a good bias–variance trade-off, and deliver strong test performance ensuring there is neither over-fitting nor under-fitting.

Effect of choosing double the alpha

- **Ridge :** Stronger L2 penalty. Coefficients will shrink further. Training and Test R square decreased. The gap between them as well increased. All predictors will still remain in the model, but with smaller magnitudes.
- **Lasso :** Stronger L1 penalty. More coefficients will shrink to zero. Fewer predictors will be selected. Training as well as Test R-square value dropped but their gap reduced even further. As given in the notebook, the number of coefficients shrunk to 26 on doubling alpha.

Most important predictors after doubling alpha

- **Ridge :** Important predictors mostly remain the same, but only effect will be that coefficients will be smaller. Variables with consistently large standardised coefficients will still remain most important. From the analysis in the notebook, the most important predictor variables are still : [GrLivArea, OverallQual, LotArea, GarageCars, BsmtFinSF1]
- **Lasso :** Only the important predictor variables will survive. Weaker or highly correlated predictors are likely to be removed. The top 5 variables changed from [‘GrLivArea', 'OverallQual', 'LotArea', 'YearBuilt', 'GarageCars’] to [‘OverallQual', 'GrLivArea', 'GarageCars', 'LotArea', 'YearBuilt’] I.e the order changed even though the variables are the same in this case.

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?

Answer : After determining the optimal regularization parameters, **Lasso regression** would be the preferred model.

Reasons are as follows:

- Lasso model resulted in a **higher Test R^2 (0.8839)(and adjusted R2)** compared to Ridge (0.8731).
- The **train–test gap is very small (~1.5%)**, indicating there is no overfitting to training data.(Generally less than 5 is preferred but less than 3 is excellent)
- Lasso selects **40 out of 100 predictors**, hence model complexity is reduced. Feature selection also improves interpretability and reduces multicollinearity.

- Error metrics (RMSE and MAE) are comparable to Ridge, with no instability hence adding more value without losing out much.

Ridge regression is more stable when all predictors are important and must be kept, but in this case, Lasso provides **better generalisation** without sacrificing accuracy.

Hence, Lasso is the more suitable choice.

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?

Answer : When the five most important predictors from the original Lasso model are unavailable, we must train a new Lasso model excluding those variables.

After refitting the model:

- Lasso will re-optimize coefficients among the remaining predictors.
- Variables that were previously correlated with the removed predictors will gain importance.
- The new top five predictors will be the variables with the highest absolute coefficients in the refitted Lasso model.

According to analysis in the python notebook, the 5 new predictor variables after excluding the top 5 are as follows:

TotRmsAbvGrd 0.099942
GarageArea 0.086679
Fireplaces 0.062066
YearRemodAdd 0.049041
BsmtQual_TA 0.041615

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?

Answer : A model can be made robust and generalisable by the following steps:

- By using regularization (Ridge/Lasso) to control complexity and ensure bias-variance tradeoff is covered.
- Ensuring a minimal train–test performance gap.
- Validating performance using **cross-validation** to ensure there is no overfitting.
- Avoiding data leakage from test data.
- Proper preprocessing and EDA steps with proper business analysis.
- Removing multicollinearity, ensuring noisy/ misleading data is eliminated.
- Evaluating on unseen test data to ensure generalisation.

Implications for accuracy:

- As the model becomes more robust and generalisable, the training accuracy may slightly decrease. Test accuracy generally improves or remains stable. Overall predictive performance of the model is better and reliable.

Why this happens:
A simpler model with controlled complexity does not rely too much on the training data thereby reducing overfitting. Although this introduces some bias, it significantly reduces variance, finding the right balance and hence leading to better real-world performance.

In practice, a slightly lower training accuracy is a worthwhile trade-off for stronger generalisation and robustness.