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?

Ridge Regression : The optimal value for alpha was 2.0 & was derived using the method '*model_cv.best_params_'*
Lasso Regression : The optimal value for alpha was 0.003. We used the method '*model_cv.best_params_'* and also did further analysis by plotting mean abolute error for different values of alpha. With the plot, we identified a small potential range of optimal alpha values and post that we did further analysis on all 3 values to eventually conclude the optimal value of 0.003

The idea of above analysis was to choose the optimal value of Alpha that does not over or under penalise the model and hence accurately eliminated both overfitting & underfitting

Doubling the Alpha value :
Doubling the value for Alpha will penalise the model even more. For Lasso regression, we observed coefficients for 9 more parameters changed to 0 thus reducing the final predictor variable count to 15. Model R-Square value dropped from 0.830 to 0.703.
For Ridge regression, doubling the alpha value to 4.0 would make the model simpler as the penalty increases. However, we did not see a significant change in R-Square for both train & test score with values as 0.937 & 0.905 respectively.

We doubled the *alpha value to 0.006 for Lasso regression* & observed below top variables –

1. GrLivArea
2. FireplaceQu_none
3. Foundation_PConc
4. OverallQual
5. ExterQual_TA
6. GarageType_Attchd
7. KitchenQual_TA
8. MasVnrType_none

We doubled the *alpha value to 4.0 for Ridge regression* & observed below top variables –

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. FireplaceQu_none
5. GarageArea
6. CentralAir_Y
7. Foundation_PConc
8. GarageQual_TA

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?

During the assignment, we got >.80 R-Square value for both Lasso (0.832) & Ridge (0.930) Regression. Both these values are pretty good and acceptable. While Ridge offered a better R-Square on both train & test data but it is also worth noting that -

1. Ridge model is much more complex than Lasso as Ridge does not do feature elimination. This means that with Ridge model, we had 203 predictor variables in our model as compared to only 27 in Lasso regression. In Machine Learning, it is always preferred to have a simpler model as they are more reliable and easier to maintain
2. Difference between train & test R-Square is higher in Ridge than Lasso. This indicates that Lasso model has better learnt the underlying data patterns and is free for any overfitting or underfitting.

Considering all these points, we should consider Lasso model of regression here and consider that our final model for this use case.

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?

Five most important variables in our final Lasso model were - GrLivArea , OverallQual , TotalBsmtSF ,GarageArea &CentralAir_Y

As per the above Question, I eliminated these variables from the train & test data and reran the same steps. This lead to a new Lasso regression model with below 5 new most important predictor variables –

1. 1stFlrSF
2. 2ndFlrSF
3. Property Age
4. SaleType_New
5. GarageType_none

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?

While the quality of data available for training & test plays a key role in creating a robust and generalisable model but there is lot more to creating such models.

1. Choosing the right kind of model is critical and is done by applying exploratory data analysis and relevant business knowledge.
2. It is also crucial to optimise the model to ensure it is robust and generalisable. Same can be done using hyperparameters to find an optimal model with minimum error and maximum accuracy on both train & test data. This keeps the model simple and avoids overfitting and underfitting. There are several techniques to perform this step and we have used Lasso & Ridge regression in our assignment.
3. Keeping the model simple is another aspect to look out for. Simpler model focus more on learning underlying trend of data and tend to perform much better when run on actual production data. Another key thing to simpler model is focussing bias-variance trade off. A simple model would usually have high bias and low variance, whereas a complex model would have low bias and high variance

There are implications to choosing simpler model i.e. their accuracy might seem low initially specially on train dataset as compared to a complex model but they are more robust.